ФЕДЕРАЛЬНОЕ АГЕНТСТВО
ПО ТЕХНИЧЕСКОМУ РЕГУЛИРОВАНИЮ И МЕТРОЛОГИИ

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

ГОСТ Р 59926— 2021/ ISO/IEC TR 20547-2:2018

Информационные технологии

ЭТАЛОННАЯ АРХИТЕКТУРА БОЛЬШИХ ДАННЫХ

Часть 2

Варианты использования и производные требования

(ISO/IEC TR 20547-2:2018, IDT)

Издание официальное

Москва
Российский институт стандартизации
2022

Предисловие

1 ПОДГОТОВЛЕН Федеральным государственным бюджетным образовательным учреждением высшего образования «Московский государственный университет имени М.В. Ломоносова» (МГУ имени М.В. Ломоносова) в лице Научно-образовательного центра компетенций в области цифровой экономики МГУ и Автономной некоммерческой организацией «Институт развития информационного общества» (ИРИО) на основе собственного перевода на русский язык англоязычной версии документа, указанного в пункте 4

2 ВНЕСЕН Техническим комитетом по стандартизации ТК 164 «Искусственный интеллект»

3 УТВЕРЖДЕН И ВВЕДЕН В ДЕЙСТВИЕ Приказом Федерального агентства по техническому регулированию и метрологии от 2 декабря 2021 г. № 1685-ст

4 Настоящий стандарт идентичен международному документу ISO/IEC TR 20547-2:2018 «Информационные технологии. Эталонная архитектура больших данных. Часть 2. Варианты использования и производные требования» (ISO/IEC TR 20547-2:2018 «Information technology — Big data reference architecture — Part 2: Use cases and derived requirements», IDT).

При применении настоящего стандарта рекомендуется использовать вместо ссылочных международных стандартов соответствующие им национальные стандарты, сведения о которых приведены в дополнительном приложении ДА.

Дополнительные сноски в тексте стандарта, выделенные курсивом, приведены для пояснения текста стандарта

5 ВВЕДЕН ВПЕРВЫЕ

Правила применения настоящего стандарта установлены в статье 26 Федерального закона от 29 июня 2015 г. № 162-ФЗ «О стандартизации в Российской Федерации». Информация об изменениях к настоящему стандарту публикуется в ежегодном (по состоянию на 1 января текущего года) информационном указателе «Национальные стандарты», а официальный текст изменений и поправок — в ежемесячном информационном указателе «Национальные стандарты». В случае пересмотра (замены) или отмены настоящего стандарта соответствующее уведомление будет опубликовано в ближайшем выпуске ежемесячного информационного указателя «Национальные стандарты». Соответствующая информация, уведомление и тексты размещаются также в информационной системе общего пользования — на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет (www.rst.gov.ru)

Содержание

Введение

Международная организация по стандартизации (ИСО) и Международная электротехническая комиссия (МЭК) вместе образуют специализированную систему всемирной стандартизации. Национальные органы по стандартизации, являющиеся членами ИСО или МЭК, принимают участие в разработке международных стандартов через технические комитеты, созданные соответствующей организацией для рассмотрения вопросов, касающихся конкретных областей технической деятельности. Технические комитеты ИСО и МЭК сотрудничают в областях, представляющих взаимный интерес. Другие международные правительственные и неправительственные организации в сотрудничестве с ИСО и МЭК также принимают участие в этой работе. В области информационных технологий ИСО и МЭК создали Совместный технический комитет ИСО/МЭК СТК1 (ISO/IEC JTC1).

Процедуры, как использованные при подготовке настоящего стандарта, так и те, что будут применяться для его последующей поддержки, описаны в части 1 Директив ИСО/МЭК. Следует в первую очередь обратить внимание на отличающиеся критерии утверждения для различных типов документов. Данный стандарт был подготовлен в соответствии с правилами редактирования, установленными частью 2 Директив ИСО/МЭК (см. www.iso.org/directives).

Следует иметь в виду возможность того, что некоторые элементы данного стандарта могут подпадать под действие патентного права. ИСО и МЭК не несут ответственности за идентификацию соответствующих патентных прав. Детальные сведения о патентных правах, выявленных в ходе разработки настоящего стандарта, будут содержаться во введении и/или в публикуемом ИСО списке полученных патентных деклараций (см. www.iso.org/patents).

Любые торговые марки, использованные в данном стандарте, представляют собой информацию, приводимую для удобства пользователей, и их упоминание не является формой поддержки или одобрения.

Разъяснение добровольного характера стандартов, объяснение смысла специфических терминов и выражений ИСО, связанных с оценкой соответствия, а также сведения о приверженности ИСО принципам Всемирной торговой организации (ВТО) в отношении технических барьеров в торговле (ТВТ), см. www.iso.org/iso/foreword.html.

Настоящий стандарт был подготовлен Совместным техническим комитетом ИСО/МЭК СТК1 «Информационные технологии».

Список всех частей стандарта ИСО/МЭК 20547 можно найти на веб-сайте ИСО.

Данный документ направлен на формирование сообщества, объединяющего интересы представителей промышленности, академических кругов и правительства, с целью подготовки согласованного перечня технических аспектов в области больших данных всех заинтересованных сторон. Эта работа включала сбор и изучение вариантов использования в различных областях (то есть областях применения). Для достижения этой цели были решены следующие задачи:

- собраны материалы, связанные с техническими аспектами работы с большими данными всех заинтересованных сторон;
- проанализирован и приоритизирован перечень технических проблем, возникающих в сложных вариантах использования, которые могут привести к задержке или помешать внедрению технологий больших данных;
- подготовлен всеобъемлющий перечень обобщенных технических аспектов в области работы с большими данными для стандарта ИСО/МЭК 20547-3 «Информационные технологии. Эталонная архитектура больших данных. Часть 3. Эталонная архитектура» (Information technology — Big data reference architecture — Part 3: Reference architecture);
- полученные результаты зафиксированы в настоящем стандарте.

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

Информационные технологии

ЭТАЛОННАЯ АРХИТЕКТУРА БОЛЬШИХ ДАННЫХ

Часть 2

Варианты использования и производные требования

Information technology. Big data reference architecture. Part 2. Use cases and derived requirements

Дата введения — 2022—03—01

1 Область применения

Настоящий стандарт содержит анализ вариантов использования больших данных в различных областях применения, а также выводы, сделанные на основе этого анализа.

2 Нормативные ссылки

В настоящем стандарте использованы нормативные ссылки на следующие стандарты [для датированных ссылок применяют только указанное издание ссылочного стандарта, для недатированных — последнее издание (включая все изменения)]:

ISO/IEC 20546, Information technology — Big data — Definition and vocabulary (Информационные технологии. Большие данные. Обзор и словарь).

3 Термины и определения

В настоящем стандарте применены термины и определения, представленные в ИСО/МЭК 20546 и приведенные ниже.

Терминологические базы данных для использования в стандартизации поддерживаются ИСО и МЭК по следующим адресам:

- Электропедия МЭК доступна по адресу http://www.electropedia.org/;
- платформа онлайн-просмотра ИСО: доступна по https://www.iso.org/obp/.

3.1 Термины, определенные в других источниках

Отсутствуют.

3.2 Термины, определенные в настоящем стандарте

3.2.1 **вариант использования** (use case): Типичное применение, сформулированное на высоком уровне для выделения технических особенностей или сравнения практики использования в различных областях.

Издание официальное

3.3 Сокращения

2D — двумерный;

3D — трехмерный;

6D — шестимерный;

AOD — данные по объекту анализа (Analysis Object Data);

API — интерфейс программирования приложений (Application Programming Interface);

ASDC — центр обработки атмосферных данных[1)] (Atmospheric Science Data Center);

ASTM — Американское общество испытаний и материалов (American Society for Testing and Materials);

AWS — платформа облачных сервисов компании Амазон (Amazon Web Services);

BC/DR — непрерывность деятельности и восстановление после чрезвычайных ситуаций (Business Continuity and Disaster Recovery);

BD — большие данные (Big data);

BER — Управление биологических и экологических исследований Министерства энергетики США (Biological and Environmental Research);

BNL — Брукхейвенская национальная лаборатория, США (Brookhaven National Laboratory);

CAaaS — аналитика климата как сервис (Climate Analytics as a Service);

CADRG — формат для оцифрованных растровых изображений с ARC-сжатием (ARC Digitized Raster Graphic (ADRG);

CBSP — провайдер облачного брокерского сервера (CBSP Cloud Brokerage Service Provider);

CERES — проект НАСА «Система для изучения облачности и излучения Земли» (Clouds and Earth's Radiant Energy System);

CERN — Европейский центр ядерных исследований (The European Organization for Nuclear Research), ЦЕРН;

CESM — модель климатической системы Земли (Community Earth System Model);

CFTC — Комиссия по торговле товарными фьючерсами (Commodity Futures Trading Commission), США;

CIA — конфиденциальность, целостность и доступность (Confidentiality, Integrity and Availability);

CINET — цифровая инфраструктура для исследований и анализа сетей и графов (Cyberinfrastructure for Network (Graph) Science and Analytics);

CMIP — проект сопоставления связанных климатических моделей (Coupled Model Intercomparison Project);

CMIP5 — пятая фаза проекта сопоставления связанных комплексных климатических моделей (Coupled Model Intercomparison Project 5);

CMS — компактный мюонный соленоид (Compact Muon Solenoid);

COSO — Комитет спонсорских организаций Комиссии Тредвея (Committee of Sponsoring Organizations of the Treadway Commission), США);

CPU — центральный процессор (Central Processing Unit);

CReSIS — Центр дистанционного зондирования ледяного покрова Университета Канзаса (Center for Remote Sensing of Ice Sheets), США;

CRTS — каталинский обзор оптических переходных процессов в режиме реального времени (Catalina Real-Time Transient Survey);

CSP — провайдер облачного сервиса (Cloud Service Provider);

[1)] Подразделение научно-исследовательского центра НАСА в Лэнгли, США.

CSS — каталинский обзор неба (Catalina Sky Survey);

CV — контролируемый словарь (Controlled Vocabulary);

DFC — Консорциум федеративных сетей данных (DataNet Federation Consortium);

DHTC — распределенные вычисления с высокой пропускной способностью (Distributed High Throughput Computing);

DNA — дезоксирибонуклеиновая кислота; ДНК (DeoxyriboNucleic Acid);

DOE — Министерство энергетики США;

DOJ — Министерство юстиции США;

DPO — онлайновые инструменты работы с данными Центра обработки атмосферных данных (Data Products Online);

EBAF — TOA — баланс и накопление энергии верхних слоев атмосферы (Energy Balanced and Filled-Top of Atmosphere). Средство генерации данных проекта НАСА «Система для изучения облачности и излучения Земли»;

EC2 — эластичное вычислительное облако (Elastic Compute Cloud);

EDT — хранилище данных в Клинике Мейо (Enterprise Data Trust), США;

EHR — электронные данные (карта) здоровья (Electronic Health Record);

EMR — электронная медицинская карта (Electronic Medical Record);

EMSO — европейская междисциплинарная обсерватория исследования морского дна и слоев воды (European Multidisciplinary Seafloor and Water Column Observatory);

ENVRI — совместная деятельность европейских сетевых инфраструктур в области экологических исследований (Common Operations of Environmental Research Infrastructures);

ENVRI RM — эталонная модель ENVRI (ENVRI Reference Model);

EPOS — европейская исследовательская инфраструктура для слежения за [геологическими] плитами (European Plate Observing System);

ESFRI — европейский стратегический форум по исследовательским инфраструктурам (European Strategy Forum on Research Infrastructures);

ESG — грид-система обработки данных о Земле (Earth System Grid);

ESGF — федеративная грид-система обработки данных о Земле (Earth System Grid Federation);

FDIC — Федеральная корпорация страхования депозитов (U.S. Federal Deposit Insurance Corporation), США;

FI — финансовый сектор (Financial Industries);

FLUXNET — сеть вышек для наблюдения за газовыми потоками (Flux Tower Network);

FMV — формат высококачественного видео (Full Motion Video);

FNAL — Национальная ускорительная лаборатория имени Ферми Fermi National Accelerator Laboratory, Fermilab), США;

GAAP — общепринятые принципы бухгалтерского учета США (U.S. Generally Accepted Accounting Principles);

GB — Гигабайт;

GCM — модель общей циркуляции (General Circulation Model);

GEOS-5 — годдардовская система наблюдения Земли, 5-я версия (Goddard Earth Observing System version 5);

GeoTiff — Tiff-формат изображения с указанием местоположения (Geo Tagged Image File Format);

GEWaSC — проект моделирования водоразделов с использованием генома (Genome-Enabled Watershed Simulation Capability);

GHG — парниковый газ (Green House Gas);

GMAO — Отдел глобального моделирования и ассимиляции Центра управления полетами имени Годдарда, НАСА (Global Modeling and Assimilation Office);

GPFS — общая параллельная файловая система (General Parallel File System);

GPS — глобальная навигационная система (Global Positioning System);

GPU — графический процессор (Graphics Processing Unit);

GRC — стратегическое управление, управление рисками и соблюдение требований (Governance, Risk management, and Compliance);

GSFC — Центр управления полетами имени Годдарда, США (Goddard Space Flight Center);

HDF5 — иерархический формат данных, 5-я версия (Hierarchical Data Format);

HDFS — распределенная файловая система Hadoop (Hadoop Distributed File System);

HPC — высокопроизводительные вычисления (High-Performance Computing);

HTC — вычисления с высокой пропускной способностью (High-Throughput Computing);

HVS — хостинговый виртуальный сервер (Hosted Virtual Server);

I/O — ввод-вывод (Input Output);

IaaS — инфраструктура как сервис (Infrastructure as a Service);

IAGOS — использование самолетов в глобальной системе наблюдений (In-service Aircraft for a Global Observing System);

ICD — международная классификация болезней (International Classification of Diseases);

ICOS — интегрированная система наблюдения за выбросами углерода (Integrated Carbon Observation System);

IMG — проект «Интегрированные микробные геномы» Объединенного института генома Министерства энергетики США (Integrated Microbial Genomes);

INPC — инфраструктура клинических данных по уходу за пациентами штата Индиана (Indiana Network for Patient Care), США;

IPCC — Межправительственная группа экспертов по изменению климата (Intergovernmental Panel on Climate Change);

iRODS — интегрированная система управления данными, основанная на использовании правил (integrated Rule-Oriented Data System);

ISACA — Международная ассоциация аудита и контроля информационных систем (Information Systems Audit and Control Association);

isc2 — Международный консорциум по сертификации в области безопасности информационных систем (International Security Computer and Systems Auditors);

ISO — Международная организация по стандартизации (International Organization for Standardization);

ITIL — библиотека инфраструктуры информационных технологий (Information Technology Infrastructure Library);

JGI — объединенный институт генома Министерства энергетики США (Joint Genome Institute);

KML — язык разметки Keyhole (Keyhole Markup Language);

kWh — киловатт-час;

LaRC — Исследовательский центр в Ленгли, НАСА (Langley Research Center);

LBNL — Национальная лаборатория имени Лоуренса в Беркли (Lawrence Berkeley National Laboratory), США;

LDA — латентное размещение Дирихле (latent Dirichlet allocation)

LHC — большой адронный коллайдер (Large Hadron Collider);

LPL — Лаборатория изучения Луны и планет в Университете Аризоны (Lunar and Planetary Laboratory), США;

LSST — большой синоптический обзорный телескоп в Обсерватории имени Веры Рубин (Large Synoptic Survey Telescope), Чили;

MERRA — система для ретроспективного анализа современной эры для исследований и приложений (Modern Era Retrospective Analysis for Research and Applications);

MERRA/AS — аналитические сервисы MERRA (MERRA Analytic Services);

MPI — интерфейс передачи сообщений (Message Passing Interface);

MRI — магнитно-резонансная томография (Magnetic Resonance Imaging);

NARA — Национальные архивы США (National Archives and Records Administration);

NARR — реанализ метеорологических данных для региона Северной Америки (North American Regional Reanalysis);

NaaS — сеть как сервис (Network as a Service);

NASA — Национальное управление по аэронавтике и исследованию космического пространства (National Aeronautics and Space Administration), США;

NCAR — Национальный центр атмосферных исследований (National Center for Atmospheric Research), США;

NCBI — Национальный центр биотехнологической информации (National Center for Biotechnology Information);

NCCS — Центр моделирования климата НАСА (Center for Climate Simulation);

NERSC — Национальный научно-исследовательский вычислительный центр энергетических исследований Министерства энергетики США (National Energy Research Scientific Computing Center);

NetCDF — NetCDF-формат представления данных (Network Common Data Form);

NEX — платформа НАСА для обмена данными о Земле (NASA Earth Exchange);

NFS — сетевая файловая система (Network File System);

NIKE — интегрированная сеть управления знаниями Национального института стандартов и технологий США (NIST Integrated Knowledge Editorial Net);

NIST — Национальный институт стандартов и технологий США (National Institute of Standards and Technology);

NITF — национальный формат передачи изображений (National Imagery Transmission Format);

NLP — обработка естественного языка (Natural Language Processing);

NRT — почти в режиме реального времени (Near Real Time);

NSF — Национальный научный фонд (National Science Foundation), США;

ODP — открытая распределенная обработка (Open Distributed Processing);

OGC — Открытый геопространственный консорциум (Open Geospatial Consortium);

PB — петабайт;

PCA — метод главных компонентов (Principal Component Analysis);

PCAOB — Некоммерческая организация по надзору за отчетностью публичных компаний (Public Company Accounting and Oversight Board), США ;

PID — присвоение постоянного идентификатора (persistent identificator);

PII — персональные данные (Personally Identifiable Information);

PNNL — Тихоокеанская северо-западная национальная лаборатория (Pacific Northwest National Laboratory), США;

RDBMS — система управления реляционными базами данных (relational database management system);

RDF — среда описания ресурсов (Resource Description Framework);
RECOVER — система поддержки принятия решений по восстановлению экосистем (Rehabilitation Capability Convergence for Ecosystem Recovery);
ROI — возврат инвестиций (return on investment);
RPI — интерферометрия повторного хода (Repeat Pass Interferometry);
RPO — заданная точка восстановления (Recovery Point Objective);
RTO — заданное время восстановления (Recovery Time Objective);
SAN — сеть хранения данных (Storage Area Network);
SAR — радар с синтезируемой апертурой (Synthetic Aperture Radar);
SDN — программно-конфигурируемая сеть [передачи данных] (software-defined networking);
SIOS — интегрированная система наблюдений за Арктикой на Шпицбергене (Svalbard Integrated Arctic Earth Observing System);
SPADE — поддержка аудита происхождения в распределенных средах (Support for Provenance Auditing in Distributed Environments);
SSH — защищенная командная среда (Secure Shell);
SSO — технология единого входа (Single Sign-On);
TB — терабайт;
tf-idf — частота встречаемости термина в документе — обратная величина частоты документов с данным термином (term frequency-inverse document frequency);
UA — Университет Аризоны (University of Arizona), США;
UAVSAR — радар с синтезируемой апертурой для беспилотного летательного аппарата (Unmanned Air Vehicle Synthetic Aperture Radar);
UC — вариант использования (Use Case);
UI — пользовательский интерфейс (User Interface);
UPS — транснациональная компания, специализирующаяся на экспресс-доставке и логистике, США (United Parcel Service);
UQ — количественная оценка неопределенности (Uncertainty Quantification);
VASP — венский пакет для «ab initio» моделирования материалов на атомарном уровне (Vienna Ab initio Simulation Package);
vCDS — виртуальный сервер климатических данных (virtual Climate Data Server);
VO — виртуальная обсерватория (Virtual Observatory);
VOIP — передача голоса с использованием IP-протокола (Voice over IP);
WALF — WALF-формат видео с высоким разрешением (Wide Area Large Format Imagery);
WLCG — глобальная грид-инфраструктура Большого адронного коллайдера (Worldwide LHC Computing Grid);
XBRL — расширяемый язык разметки для деловой отчетности (Extensible Business Reporting Language);
XML — расширяемый язык разметки (Extensible Markup Language);
ZTF — обзор «Фабрика транзиентов Цвики» (Zwicky Transient Factory);

4 Характеристики варианта использования для проведения обследования

4.1 Общие характеристики

Предметная область: поле предназначено для классификации вариантов использования. Не заполнялось, поскольку до представления вариантов использования онтология не была создана.

Автор / организация / адрес электронной почты: имя и фамилия, название организации и адрес электронной почты (если предоставлен) лица (лиц), представившего(их) вариант использования.

Акторы/заинтересованные лица, их роли и ответственность: описание участников и их ролей в варианте использования.

Цели: поле для описания цели варианта использования.

Описание варианта использования: краткое описание варианта использования.

4.2 Текущие решения

В разделе описывается используемый подход к обработке больших данных на уровнях программно-аппаратной инфраструктуры и аналитики, включая следующие процессы:

- **вычислительная система:** вычислительный компонент системы анализа данных;
- **хранилище данных:** компонент хранения системы анализа данных;
- **сеть связи:** сетевой компонент системы анализа данных;
- **программное обеспечение:** программный компонент системы анализа данных.

4.3 Характеристики больших данных

Характеристики больших данных, которые описывают свойства (исходных, необработанных) данных, включая четыре основные V-характеристики больших данных.

Источник данных: происхождение данных, которые могут быть получены из интернета вещей, Всемирной паутины, в ходе опросов, коммерческой деятельности, моделирования или от измерительных приборов. Источник (источники) может быть распределенным, централизованным, локальным или удаленным.

Место назначения данных[1)]**:** если в варианте использования данные преобразуются, в поле указывают, куда поступают окончательные результаты.

Объем: характеристика массивов данных, которая преимущественно ассоциируется с большими данными. Объем определяет значительное количество данных, доступных для анализа с целью извлечения ценной информации. Представление о том, что большую ценность можно получить при анализе большего объема данных, было одним из стимулов создания новых технологий масштабирования.

Скорость обработки: скорость потока, с которой данные создаются, передаются, сохраняются, анализируются или визуализируются. Скорость обработки больших данных означает, что большие массивы данных должны быть обработаны за короткий промежуток времени. При высоких скоростях обработки данных обычно имеют дело с методами обработки потоковых данных.

Разнообразие: характеризует необходимость анализа данных из нескольких предметных областей и/или нескольких типов данных. Разнообразные массивы данных преобразовывались или предварительно анализировались для определения характеристик, которые позволили бы интегрировать их с другими данными. Широкий диапазон форматов данных, логических моделей, шкал времени и семантик, который желательно применять в аналитике, усложняет интеграцию разнообразных данных. Возрастает необходимость использования метаданных для интеграции.

Вариативность: изменения в скорости передачи, формате или структуре, семантике или качестве массива данных, которые оказывают воздействие на поддерживаемое приложение, аналитику или решение задачи. Результаты воздействия могут приводить к необходимости изменений в архитектурах, интерфейсах, процессах/алгоритмах, а также способах интеграции/слияния, хранения, применения и использования данных.

4.4 Наука о больших данных

Наука о больших данных описывает высокоуровневые аспекты процесса анализа данных.

Достоверность и качество данных: полнота и точность данных с точки зрения семантического содержания, а также качества синтаксиса данных (например, наличия пропущенных полей или неправильных значений).

Визуализация: способ представления данных для аналитика, принимающего решения на их основе. Как правило, визуализация следует за этапом анализа данных и является заключительным этапом процесса технического анализа данных.

Качество данных (синтаксис) [Data quality (syntax)]: синтаксическое качество данных (например, наличие пропущенных полей или неправильных значений).

[1)] *В шаблоне данное поле не использовалось.*

Типы данных: характеристики данных, такие как структурированные или неструктурированные данные, изображения (например, пиксельные), текст (например, последовательности символов), последовательности генов, числовое значение.

Метаданные: характеристики качества и полноты используемых метаданных.

Курирование и управление: характеризует процесс, обеспечивающий высокое качество данных, и ответственное лицо.

Примечание — Форма представления варианта использования включает отдельное поле для описания проблем безопасности и защиты персональных данных.

Аналитика данных: характеристики, в обобщенном виде инструменты и алгоритмы, используемые при обработке данных на любой стадии, включая преобразование данных в информацию, информации — в знания, а знания — в мудрость.

4.5 Общие проблемы больших данных

В заключительных полях формы опроса содержатся следующие вопросы:

- **Иные проблемы больших данных:** упустили ли мы нечто важное, демонстрирующее Ваш вариант использования? Это Ваш шанс ответить на вопросы, которые мы должны были бы задать.
- **Проблемы пользовательского интерфейса и мобильного доступа:** описание проблем доступа или генерации больших данных клиентами, включая смартфоны и планшеты.
- **Технические проблемы обеспечения безопасности и защиты персональных данных:** укажите проблемы обеспечения информационной безопасности и особенно защиты персональных данных, возникающие в результате ужесточения требований законодательства.
- **Перечислите основные характеристики и связанные варианты использования:** поместите вариант использования в контекст подобных ему вариантов. Опишите характеристики, которые допускают обобщение или специфичны для данного варианта.
- **Будущее проекта:** какие в будущем ожидаются изменения в применении и/или подходе (оборудование, программное обеспечение, аналитика)?
- **Дополнительная информация о проекте (URLs)**: приведите полезные гипертекстовые ссылки.

4.6 Шаблон описания варианта использования больших данных

Данный раздел содержит незаполненную форму для представления варианта использования. Эта форма использовалась для сбора данных о вариантах использования для определения технических требований (проблем).

Примечание — Термины, используемые в этом шаблоне, могут не совпадать с терминами стандарта ИСО/МЭК 20546 и других частей серии ИСО/МЭК 20547.

Название		
Предметная область		
Автор/организация/эл.почта		
Акторы/заинтересованные лица, их роли и ответственность		
Цели		
Описание варианта использования		
Текущие решения	**Вычислительная система**	
	Хранилище данных	
	Сеть связи	
	Программное обеспечение	

Характеристики больших данных	**Источник данных (распределенный/ централизованный)**	
	Объем (количество)	
	Скорость обработки (например, в реальном времени)	
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	
	Вариативность (темпы изменения)	
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	
	Визуализация	
	Качество данных (синтаксис)	
	Типы данных	
	Аналитика данных	
Иные проблемы больших данных		
Проблемы пользовательского интерфейса и мобильного доступа		
Технические проблемы обеспечения безопасности и защиты персональных данных		
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)		
Дополнительная информация (гиперссылки)		

5 Обзор вариантов использования

5.1 Процесс подготовки вариантов использования

Вариант использования — типичное применение, сформулированное на высоком уровне для выделения технических особенностей или сравнения практики использования в различных областях. Формирование списка содержательных вопросов и проблем, с которыми сталкиваются заинтересованные стороны, осуществлено на основе собранной общедоступной информации о различных архитектурах больших данных. В целях структуризации этого списка описания вариантов использования сгруппированы по предметным областям.

Примечание 1 — Список областей применения отражает представленные варианты использования и не является исчерпывающим.

Были выделены следующие девять предметных областей.

Деятельность государственных органов (4): Национальные архивы США, Бюро переписей населения США.

Коммерческая деятельность (8): облачная экосистема бизнеса, включая финансовую отрасль, облачное резервное копирование, интеллектуальный тематический поиск научной литературы, потоко-

вая передача мультимедийного контента, веб-поиск, цифровое материаловедение и геномика материалов, грузоперевозки.

Оборона (3): анализ показаний датчиков, идентификация и отслеживание объектов по данным фотосъемки и видеонаблюдения, оценка ситуации.

Здравоохранение и медико-биологические науки (10): электронные медицинские документы, анализ графов и вероятностный анализ, цифровая патология, анализ биоизображений, геномика, эпидемиология, моделирования распространения социального влияния, биологическое разнообразие.

Глубокое обучение и социальные сети (6): беспилотные автомобили, географическая привязка фотографий, распространение информации в социальных сетях, краудсорсинг, аналитика сетей и графов, эталонные наборы данных.

Экосистема для исследований (4): коллективная работа с метаданными, анализ текстов на естественном языке, эксперименты на синхротронах.

Астрономия и физика (5): обзоры неба (и сравнение данных наблюдений с результатами моделирования), Большой адронный коллайдер в ЦЕРН, эксперимент в области физики элементарных частиц Belle Accelerator II.

Науки о Земле, экологические науки и полярные исследования (10): некогерентное рассеяние радиоволн в атмосфере, исследования землетрясений, океана, наблюдения Земли, радиолокационное зондирование ледяного покрова, радиолокационное картографирование Земли, массивы данных для моделирования климата, изучение турбулентности атмосферы, подповерхностная биогеохимия (микробы в водоразделах), датчики газовых потоков.

Энергетика (2): интеллектуальные энергосети, управление энергопотреблением домашнего хозяйства.

Примечание 2 — Шаблон описания варианта использования был полезен при сборе обобщенной информации с целью проведения вспомогательного и сопоставительного анализа вариантов использования. В то же время в содержании каждого раздела заполненной формы описания наблюдались различия в степени детализации количественной и качественной информации. Для некоторых областей применения были представлены схожие варианты использования анализа больших данных, что позволило получить более полное представление о технических особенностях и проблемах применения анализа больших данных в этих областях.

Примеры вариантов использования описаны в этом разделе на основе первоначально представленной информации. Исходный контент (см. приложение А) изменен не был.

Примечание 3 — В описаниях вариантов использования упоминаются конкретные решения и технологии коммерческих поставщиков, однако перечисление этих решений и технологий не означает их одобрения рабочей группой РГ9 Совместного технического комитета ИСО/МЭК СТК1.

Варианты использования пронумерованы последовательно для облегчения перекрестных ссылок между их краткими описаниями, представленными в данном разделе, исходными описаниями (приложение А) и сводными таблицами по вариантам использования (приложения В, С и Д).

5.2 Деятельность государственных органов

5.2.1 Вариант использования 1: Большие данные переписи населения в США, проведенной в 2010 и 2000 годах на основании части 13 Свода законов США

Применение

Данные переписи населения в США, проведенной в 2010 и 2000 гг. в соответствии с разделом 13 «Переписи населения» Свода законов США, в течение нескольких десятилетий должны сохраняться таким образом, чтобы обеспечить их доступность и возможность анализа через 75 лет, по истечении ограничительного периода.

В течение ограничительного периода в 75 лет данные должны храниться «как есть», без возможности доступа и анализа, с обеспечением сохранности на уровне битов. Данные курируются, что может включать преобразование формата. Доступ и аналитика должны быть обеспечены через 75 лет.

Часть 13 Свода законов США уполномочивает Бюро переписи населения США (U.S. Census Bureau) собирать и сохранять относящиеся к переписи данные и гарантирует защиту персональных и отраслевых данных.

Текущий подход

Набор данных содержит отсканированные документы общим объемом 380 терабайт.

Планы на будущее

Для данного варианта использования будущие сценарии использования и приложения данных описаны не были.

5.2.2 Вариант использования 2: Прием Национальными архивами США (NARA) государственных данных на хранение, поиск, извлечение и обеспечение долговременной сохранности

Применение

Прием государственных данных на хранение, поиск, извлечение и обеспечение их долговременной сохранности.

Текущий подход

Данные в настоящее время обрабатываются следующим образом:

- передача данных под физический контроль Национальных архивов и переход к Национальным архивам юридической ответственности за их сохранность;
- предварительная обработка данных, включающая проверки на наличие вирусов, определение файловых форматов и удаления пустых файлов;
- индексирование данных;
- категоризация документов (выделяются, например, чувствительные конфиденциальные, неконфиденциальные, персональные данные);
- преобразование устаревших файловых форматов в современные;
- проведение электронного раскрытия;
- поиск и извлечение данных в рамках исполнения специальных запросов;
- поиск и извлечение государственных документов представителями общественности.

Сотни терабайт информации хранятся централизованно в коммерческих базах данных, поддерживаемых кастомизированным программным обеспечением и коммерческими поисковыми продуктами.

Планы на будущее

Федеральные органы исполнительной власти США располагают многочисленными распределенными источниками данных, которые в настоящее время должны быть переданы в централизованное хранилище. В будущем эти источники данных могут находиться в ряде облачных сред. В этом случае в рамках передачи Национальным Архивам ответственности за физическую сохранность желательно избегать перемещения больших данных из одного облака в другое либо из облака в центр обработки данных.

5.2.3 Вариант использования 3: Повышение активности респондентов в статистических обследованиях

Применение

Затраты на проведение статистических обследований растут, в то время как активность респондентов падает. Целью текущей работы является повышение качества, включая и сокращение затрат на проведение обследований посредством применения усовершенствованных «методов рекомендательных систем» (recommendation system techniques). Эти методы являются открытыми и научно обоснованными; они предусматривают использование комбинации данных из нескольких источников, а также вспомогательных данных исторических обследований (т. е. административные данные об обследованиях).

Текущий подход

В настоящем варианте использования речь идет о массиве данных, полученных в ходе опросов, а также из других государственных административных источников. Объем этих данных составляет около петабайта. Данные могут передаваться в потоковом режиме. Во время последней всеобщей переписи населения, проводимой раз в 10 лет в США, осуществлялась непрерывная потоковая передача полученных на местах данных, содержащих около 150 млн документов. Необходимо было обеспечить безопасность и конфиденциальность всех данных. Согласно требованиям законодательства следовало обеспечить возможность аудита всех процессов на предмет безопасности и конфиденциальности. Качество данных должно было быть высоким и статистически проверяться на точность и надежность на протяжении всего процесса сбора данных. Информация о решении приведена в А.1.3.

Планы на будущее

Необходимы улучшенные рекомендательные системы, аналогичные тем, которые используются в электронной коммерции (например, аналогичные системе, упоминаемой в варианте использования 5.3.3), позволяющие снизить затраты и повысить качество, обеспечить одновременно надежные и публично проверяемые меры защиты конфиденциальности. Визуализация полезна для проверки данных,

оперативной деятельности и общего анализа. Система продолжает развиваться, и в нее включаются такие важные функциональные возможности, как поддержка мобильного доступа.

5.2.4 Вариант использования 4: Использование нетрадиционных данных для повышения активности респондентов в статистических обследованиях (адаптивная схема)

Применение

Затраты на проведение статистических обследований растут, в то время как активность респондентов падает. Цели данного варианта использования близки к целям варианта «Повышение активности респондентов в статистических обследованиях» (см. 5.2.3). Однако данный вариант использования охватывает коммерческие и публичные источники данных из интернета, сетей беспроводной связи и систем электронных транзакций, которые для целей аналитических исследований объединяются с данными традиционных статистических обследований. Цель такого комбинирования данных — повысить качество статистики для небольших регионов и новых показателей, а также обеспечить своевременность публикуемой статистики.

Текущий подход

Интегрируются данные из ряда источников, включая данные статистических обследований, иные государственные административные данные, данные из интернета, систем беспроводной связи, данные электронных транзакций, возможно, данные из социальных сетей, а также геопространственные данные из различных источников. Характеристики программного обеспечения, визуализации и данных аналогичны соответствующим характеристикам варианта использования «Повышение активности респондентов в статистических обследованиях».

Планы на будущее

Требуется разработать инструменты аналитики, позволяющие дать более детальные статистические оценки почти в режиме реального времени и с меньшими затратами. Надежность статистических оценок, полученных на основе комбинирования данных из подобных смешанных источников, пока еще предстоит определить.

5.3 Коммерческая деятельность

5.3.1 Вариант использования 5: Облачная экосистема для финансовой отрасли

Применение

Необходимо расширить использование облачных технологий (например, больших данных) в деятельности секторов финансовой отрасли (т. е. в банковском деле, операциях с ценными бумагами и управлении инвестициями, страховании), осуществляющих операции в США.

Текущий подход

Финансовая отрасль уже использует большие данные для выявления мошенничества, анализа и оценки рисков, а также расширения знаний и понимания клиентов. В то же время в отрасли все еще используются традиционные системы типа клиент/сервер/хранилище данных/ система управления реляционными базами данных (RDBMS) для управления, обработки, хранения и архивирования финансовых данных. В этой области важны обработка и анализ данных в реальном времени.

Планы на будущее

Необходимо решить задачи обеспечения безопасности, неприкосновенности персональных данных и исполнения законодательно-нормативных требований. Например, в финансовой отрасли необходимо рассмотреть вопрос о требуемом Федеральной комиссией по ценным бумагам и биржам (Securities and Exchange Commission, SEC) применении языка XBRL (расширяемый язык разметки для деловой отчетности) и использовании иных облачных функций.

5.3.2 Вариант использования 6: Международная исследовательская сеть Mendeley

Применение

Международная сеть «Менделей» (Mendeley) позволила сформировать базу данных научно-исследовательских материалов, которая облегчает создание коллективно используемых библиографий. Mendeley дает возможность собирать и использовать информацию о закономерностях чтения материалов исследований, а также о других видах деятельности, осуществляемых с помощью ее программного обеспечения и с целью создания более эффективных инструментов для поиска и анализа научной литературы. Системы интеллектуального анализа и классификации текста позволяют автоматически рекомендовать взаимосвязанные исследования, повышая производительность и экономическую эффективность исследовательских групп, в особенности тех, которые занимаются мониторингом литературы по конкретной теме.

Текущий подход

Объем данных в настоящее время составляет 15 терабайт и увеличивается со скоростью около 1 терабайта в месяц. Информация о решении приведена в А.2.2. База данных использует стандартные библиотеки для проведения машинного обучения и аналитики, выполнения латентного размещения Дирихле (Latent Dirichlet Allocation, LDA, порождающая вероятностная модель для сбора дискретных данных), а также специально разработанные инструменты для составления отчетности и визуализации данных, агрегирования сведений о читательской и социальной активности, связанной с каждым документом.

Планы на будущее

В настоящее время пакетные задания по сохранению больших данных планируются раз в день, но началась работа над рекомендациями по выполнению работ в реальном времени. База данных содержит примерно 400 млн документов, в том числе около 80 млн уникальных документов, принимая в рабочие дни от 500 до 700 тыс. новых загрузок. Таким образом, основная проблема заключается в группировке соответствующих друг другу документов вычислительно эффективным (т. е. масштабируемым и распараллеливаемым) способом, когда они загружаются из разных источников и могут быть слегка модифицированы инструментами аннотирования третьих сторон или же путем присоединения титульных страниц либо наложения «водяных знаков» издателя.

5.3.3 Вариант использования 7: Сервис потоковой передачи мультимедийного контента

Применение

Сервис Netflix обеспечивает потоковую передачу выбранных пользователем фильмов, решая одновременно несколько задач (в интересах различных заинтересованных сторон), но с акцентом на удержание подписчиков. Компании нужно в режиме реального времени определить наилучшую возможную подборку видеоматериалов для пользователя (например, домохозяйства) в заданном контексте с целью максимизации потребления фильмов. Основными технологиями Netflix являются рекомендательные системы и доставка потокового видео. Рекомендательные системы всегда персонализированы и используют логистическую/линейную регрессию, эластичные сети, факторизацию матриц, кластеризацию, разведочный анализ данных (exploratory data analysis, EDA), ассоциативные правила, градиентный бустинг деревьев решений и другие инструменты. Цифровые фильмы хранятся в облаке вместе с метаданными, а также с индивидуальными профилями пользователей и рейтингами для небольшой части фильмов. В настоящее время в системе используется несколько критериев: рекомендательная система на основе контента, рекомендательная система на основе данных пользователей и разнообразие. Алгоритмы постоянно совершенствуются с помощью A/B-тестирования (т. е. используемого в онлайн-маркетинге метода рандомизированных экспериментов с двумя переменными).

Текущий подход

Компания Netflix провела конкурс на лучший алгоритм совместного фильтрования для прогнозирования пользовательских рейтингов фильмов, целью которых было повышение точности прогнозирования на 10 %. Победившая система объединила более 100 различных алгоритмов. Информация о решении описана в А.2.3. Были организованы бизнес-инициативы с целью увеличения зрительской аудитории.

Планы на будущее

Потоковое видео — очень конкурентный бизнес. Необходимо знать о других компаниях, а также о тенденциях, связанных как с контентом (например, какие фильмы популярны), так и с технологиями больших данных.

5.3.4 Вариант использования 8: Веб-поиск

Применение

Функция веб-поиска через ~ 0,1 секунды возвращает результаты поисковых запросов, включающих в среднем три слова. Важно максимизировать такие метрики, как «точность 10 наилучших результатов» (precision@10), отражающие количество высокоточных, соответствующих запросу ответов в первой десятке лучших ранжированных результатов.

Текущий подход

Текущий подход использует следующие шаги:

- сканирование интернета;
- предварительная обработка данных с целью выделения элементов, по которым можно вести поиск (слова, позиции);
- формирование инвертированного индекса, который связывает слова с их местоположением в документах;

- ранжирование релевантности документов с использованием алгоритма PageRank;
- использование маркетинговых технологий (например, обратного проектирования — reverse engineering) для определения моделей ранжирования либо создание препятствий для использования обратного проектирования;
- кластеризация документов по темам (как в Google News);
- эффективное обновление результатов.

Данный вариант использования, в настоящее время охватывающий около 45 млрд веб-страниц, значительно повлиял на развитие современных облачных решений и появление таких технологий, как Map/Reduce.

Планы на будущее

Поиск в интернете — очень конкурентная сфера деятельности, поэтому здесь необходимы постоянные инновации. Двумя важными областями для внедрения инноваций являются удовлетворение потребностей растущего сегмента мобильных клиентов, а также растущая изощренность возвращаемых результатов поиска и схем размещения информации с целью максимизации общей выгоды клиентов, рекламодателей и поисковой компании. Все большее значение также приобретают «глубокий интернет» (deep web-контент, не индексируемый стандартными поисковыми системами, скрытый за пользовательскими интерфейсами доступа к базам данных и т. д.) и поиск по мультимедийным материалам. Ежедневно загружается 500 млн фотографий, и ежеминутно на YouTube закачивается 100 часов видеоматериалов.

5.3.5 Вариант использования 9: Обеспечение непрерывности деловой деятельности и восстановления после катастроф для больших данных в облачной экосистеме

Применение

При обеспечении непрерывности деловой деятельности и ее восстановления после катастроф (Business Continuity and Disaster Recovery, BC/DR) необходимо учесть роль, которую четыре перекрывающихся и взаимозависимых фактора будут играть в обеспечении реализации стратегического плана организации. Этими четырьмя факторами являются люди (как ресурсы), процессы [например, время/затраты/возврат инвестиций (ROI)], технологии (например, различные операционные системы, платформы, а также зоны влияния/масштабы воздействия технологий) и стратегическое управление (зависит от многочисленных различных регулирующих органов).

Текущий подход

Сервисы репликации данных предоставляются через облачные экосистемы, включающие предоставление инфраструктуры как сервиса (IaaS) и поддерживаемые центрами обработки данных уровня Tier 3. Репликация отличается от резервного копирования тем, что воспроизводятся только те изменения, которые произошли после предыдущей репликации, включая изменения на уровне блоков. Репликация может быть выполнена быстро — в рамках пятисекундного «окна», при этом репликация данных может проводиться каждые четыре часа. Соответствующий «снимок» данных сохраняется в течение семи рабочих дней или дольше, если это необходимо. Реплицированные данные могут быть перемещены в запасной центр (т. е. в резервную систему) для удовлетворения требований организации в отношении заданной точки восстановления (recovery point objective, RPO) и заданного времени восстановления (recovery time objective, RTO). Соответствующая информация о решении приведена в приложении А. Объемы данных варьируются от терабайтов до петабайтов.

Планы на будущее

Переключение с основного сайта на сайт репликации или резервный сайт еще не полностью автоматизировано. Цель заключается в том, чтобы дать пользователю возможность автоматически инициировать последовательность действий по переходу на резервную систему. Обе организации должны знать, какие серверы должны быть восстановлены и какие существуют зависимости и взаимозависимости между серверами основного сайта и сайта репликации и/или резервного сайта. С этой целью необходим постоянный мониторинг обоих сайтов.

5.3.6 Вариант использования 10: Грузоперевозки

Применение

Компаниям, занимающимся доставкой грузов, нужны оптимальные средства мониторинга и отслеживания груза.

Текущий подход

Информация обновляется только тогда, когда сведения с маркировки объекта считываются сканером штрихкода, который отправляет данные на центральный сервер. В настоящее время местоположение объекта в реальном времени не отображается.

Планы на будущее

Отслеживание объектов в режиме реального времени возможно с помощью приложения «интернета вещей», в котором объектам присваиваются уникальные идентификаторы и которое способно автоматически передавать данные, то есть без участия человека.

Новым аспектом станут сведения о статусе и состоянии объекта, включая информацию с датчиков и получаемые от глобальной системы позиционирования (GPS) координаты, а также уникальная схема идентификации, основанная на международном стандарте ИСО/МЭК 29161:2016 «Информационные технологии. Структура данных. Уникальная идентификация для интернета вещей»[1)].

5.3.7 Вариант использования 11: Данные об используемых в производстве материалах

Применение

Каждый физический продукт изготовлен из материалов, которые были выбраны исходя из их свойств, стоимости и доступности. Каждый год принимаются связанные с выбором материалов решения на общие суммы, исчисляемые сотнями миллиардов долларов. Однако внедрение новых материалов обычно занимает два-три десятилетия, а не несколько лет, отчасти из-за того, что сведения о новых материалах не являются легкодоступными. Чтобы ускорить процесс внедрения, необходимо улучшить доступность, качество и удобство использования данных о материалах, а также преодолеть проприетарные барьеры для обмена такими данными. Необходимы достаточно крупные хранилища данных о материалах, способствующие поиску и раскрытию этой информации.

Текущий подход

Решения об использовании материалов в настоящее время излишне консервативны, часто основываются на более старых, а не последних данных соответствующих исследований и разработок, и не используют достижения в области построения моделей и моделирования.

Планы на будущее

Информатика материалов (materials informatics) — это область, в которой новые инструменты науки о данных могут оказывать существенное влияние, позволяя предсказывать поведение и характеристики реальных материалов (в количествах от грамма до тонны), начиная с описаний на атомном, нано- и/или микрометровом уровнях. Для поддержки этого необходимы следующие усилия:

- создание хранилищ данных о материалах помимо существующих, которые ориентированы на хранение лишь базовых данных;
- разработка международных стандартов регистрации данных, которые могут использоваться многочисленными специалистами по материалам, включая разработчиков стандартов испытаний материалов (таких как ассоциация ASTM International и Международная организация по стандартизации ИСО), занимающимися испытаниями материалов компании, производителями материалов, а также научно-исследовательскими и опытно-конструкторскими лабораториями;
- разработка инструментов и процедур, помогающих организациям, которым требуется депонировать в хранилищах данных сведения о проприетарных материалах, маскировать проприетарную информацию, сохраняя при этом пригодность данных к использованию;
- разработка многопараметрических инструментов визуализации данных о материалах, позволяющих работать с достаточно большим количеством переменных.

5.3.8 Вариант использования 12: «Геномика» материалов на основе результатов моделирования

Применение

Широкое применение моделирования, охватывающее большое количество возможных проектных решений, приведет к появлению инновационных технологий для электрических батарей и аккумуляторов. Проводятся систематические вычислительные исследования для изучения инновационных возможностей фотоэлектрических устройств. Поиск и моделирование являются фундаментом рационального проектирования материалов. Для всего этого требуется менеджмент результатов моделирования, используемых в интересах «генома материалов».

Текущий подход

Результаты создаются с использованием программного обеспечения PyMatGen, FireWorks, VASP, ABINIT, NWChem, BerkeleyGW, а также различных программ, разрабатываемых при участии специалистов по материаловедению. Программы исполняются на больших суперкомпьютерах, таких как состоящая из 150 тыс. процессоров вычислительная система Hopper в Национальном научно-исследователь-

---

[1)] *ISO/IEC 29161:2016, Information technology — Data structure — Unique identification for the Internet of Things, https://www.iso.org/standard/45240.html.*

ском вычислительном центре энергетических исследований Министерства энергетики США (NERSC), которые позволяют проводить моделирование с высоким разрешением.

Планы на будущее

Для моделирования необходимы крупномасштабные вычисления и гибкие методы обработки данных, подходящие для обработки неупорядоченных данных. Развитие направленного на результат мышления при проектировании материалов требует машинного обучения и систем управления знаниями, объединяющих данные из публикаций, результаты экспериментов и моделирования. В числе прочих потребностей можно назвать масштабируемые базы данных для данных типа «ключ-значение» и библиотек объектов. В течение следующих пяти лет ожидается рост объемов данных со 100 терабайт в настоящее время до 500 терабайт.

5.4 Оборона

5.4.1 Вариант использования 13: Облачный крупномасштабный анализ и визуализация геопространственных данных

Применение

Необходимо обеспечить крупномасштабный анализ и визуализацию геопространственных данных. По мере того, как увеличивается количество датчиков и источников данных с географической привязкой, объемы требующих сложного анализа и визуализации геопространственных данных увеличиваются в геометрической прогрессии.

Текущий подход

Традиционные географические информационные (геоинформационные) системы (ГИС) обычно способны анализировать миллионы и визуализировать тысячи объектов.

Типы данных включают растровые графические образы и изображения в различных форматах, таких как национальный формат передачи изображений (National Imagery Transmission Format, NITF), Tiff-формат изображения с указанием местоположения (GeoTiff) и формат для оцифрованных растровых изображений с ARC-сжатием (Compressed ARC Digitized Raster Graphics, CADRG), а также векторную графику в различных формах, таких как формат Shapefile, язык разметки Keyhole (Keyhole Markup Language, KML) и текстовые потоки. Типы объектов включают точки, линии, области, ломаные линии (polylines), окружности и эллипсы.

Регистрация изображений — преобразование различных данных в единую систему — требует точности данных и датчика. Аналитика включает в себя метод главных компонентов (principal component analysis, PCA) и анализ независимых компонентов (independent component analysis, ICA), ближайшую точку подхода, отклонение от маршрута и плотность точек во времени. Информация о программном обеспечении приведена в А.3.1.

Планы на будущее

Современные интеллектуальные системы часто содержат триллионы геопространственных объектов и должны визуализировать и взаимодействовать с миллионами объектов. Критически важными проблемами являются индексирование, поиск/извлечение и распределенный анализ (обратите внимание, что геопространственные данные требуют уникальных подходов к индексации и проведению распределенного анализа), формирование и передача визуализации, а также визуализация данных в конечной точке беспроводных соединений с низкой пропускной способностью. Данные являются чувствительными, и должна быть обеспечена их полная безопасность при передаче и хранении (особенно на портативных устройствах).

5.4.2 Вариант использования 14: Идентификация и отслеживание объектов по данным широкоформатной фотосъемки территории или полнокадрового видео. Постоянное наблюдение

Применение

Датчики постоянного наблюдения легко могут за считанные часы собирать петабайты фото- и видеоданных. Данные должны быть редуцированы к набору геопространственных объектов (например, точек, путей), которые можно легко интегрировать с другими данными для формирования общей оперативной картины. Типичная обработка включает выделение из первичных необработанных фото/видеоданных объектов (например, транспортных средств, людей и грузов) и их отслеживание во времени.

Текущий подход

Человек не способен обработать такие объемы данных в целях предупреждения о событиях или отслеживания. Обработка данных должна осуществляться рядом с датчиком, который, вероятно, развернут на передовой, поскольку объемы данных слишком велики для того, чтобы их можно было легко

передать. Типичные системы выделения объектов в настоящее время представляют собой небольшие (от 1 до 20 узлов) кластеры расширенных за счет использования графических процессоров (GPU) компьютерных систем.

Существует широкий спектр специализированного программного обеспечения и инструментов, включая, в том числе, традиционные реляционные СУБД и средства отображения.

Данные в режиме реального времени захватываются в FMV-формате высококачественного видео — от 30 до 60 кадров в секунду при полноцветном разрешении 1080 пикселей (т. е. с размером кадра 1920 на 1080 пикселей, построчная развертка высокой четкости) или в WALF-формат видео с высоким разрешением (WALF) — от 1 до 10 кадров в секунду при полноцветном разрешении 10 тысяч на 10 тысяч пикселей.

Извлеченные результаты обычно визуализируются путем наложения на отображение геопространственных данных. Аналитика включает базовую аналитику обнаружения объектов и интеграцию со сложными инструментами информирования о ситуации посредством объединения данных. Необходимо принимать во внимание серьезные проблемы безопасности; нельзя допустить компрометацию источников данных и методов их обработки (т. е. «враг» не должен знать, что именно мы видим).

Планы на будущее

Типичной проблемой является интеграция обработки такого рода в большой кластер графических процессоров, способный параллельно обрабатывать данные от нескольких датчиков в масштабе времени, близком к реальному. Передача данных от датчика к системе также является серьезной проблемой.

5.4.3 Вариант использования 15: Обработка и анализ разведывательных данных

Применение

Работающим с разведданными аналитикам требуются следующие возможности:

- идентифицировать взаимосвязи между объектами (например, людьми, организациями, местами, оборудованием);
- выявлять тенденции в настроениях или намерениях как населения в целом, так и групп лидеров, таких как государственные деятели и представители негосударственных структур;
- выявлять с упреждением случаи злонамеренного использования искусственного интеллекта;
- определять место и, по возможности, время проведения враждебных действий, включая установку самодельных взрывных устройств;
- отслеживать местоположение и действия потенциально враждебных действующих лиц;
- осмысливать и извлекать знания из многообразных, разрозненных и часто неструктурированных (например, текстовых) источников данных;
- обрабатывать данные вблизи точки сбора и обеспечивать легкий обмен данными с/между отдельными солдатами, подразделениями, отрядами передового базирования и высшим руководством гарнизонов.

Текущий подход

Объем данных варьируется в диапазоне от десятков терабайт до сотен петабайт, причем устройства сбора фото/видеоданных собирают петабайт данных за несколько часов. У пехотинцев обычно имеется от одного до сотен гигабайт данных, хранящихся в портативном/карманном устройстве. Сведения о программном обеспечении приведены в А.3.3.

Планы на будущее

Данные в настоящее время существуют в изолированных хранилищах. Эти данные должны быть доступны через семантически интегрированное пространство данных. Широкий спектр типов, источников, структур данных различного качества будет охватывать рад предметных областей и требует интегрированного поиска и анализа. Большинство ключевых по важности данных либо являются неструктурированными, либо хранятся в виде графических образов или видеоматериалов, что требует значительной обработки для выделения объектов и извлечения информации. Качество сети, происхождение данных и безопасность имеют важнейшее значение.

5.5 Здравоохранение и медико-биологические науки

5.5.1 Вариант использования 16: Данные электронной медицинской документации[1)]

Применение

В настоящее время появляются крупные национальные инициативы, касающиеся данных о здоровье. К ним относятся:

- разработка информационной системы в сфере здравоохранения с использованием технологии машинного обучения, поддерживающей принятие клинических решений, все больше основанных на фактических данных, посредством предоставления своевременной, точной и актуальной клинической информации, ориентированной на пациента;
- использование электронных данных клинических наблюдений для эффективного и быстрого преобразования научных открытий в эффективные клинические методы лечения;
- электронный обмен интегрированными данными о здоровье в интересах повышения эффективности и результативности процесса оказания медицинских услуг.

Все эти ключевые инициативы опираются на высококачественные, крупномасштабные, стандартизированные и агрегированные данные о здоровье. Требуются развитые методы для стандартизации выделения понятий (concept identification), связанных с пациентом, поставщиком, учреждением и клинической деятельностью, осуществляемой внутри отдельных организаций сферы здравоохранения и между ними. В случае применения этих методов при определении и извлечении клинических фенотипов (проявлений болезни) из нестандартных, дискретных и представленных в виде свободного текста клинических данных могут выделяться признаки, извлекаться информация и расширяться модели принятия решений на основе машинного обучения. Данные клинического фенотипа должны быть использованы для поддержки объединения пациентов в группы (cohort selection), изучения результатов лечения и принятия клинических решений.

Текущий подход

Инфраструктура клинических данных по уходу за пациентами штата Индиана, США (INPC) является крупнейшей и старейшей в США системой обмена медицинской информацией, которая хранит клинические данные из более чем 1100 отдельных оперативных медицинских источников. Это более 20 терабайт первичных данных, которые описывают более 12 млн пациентов и более 4 млрд отдельных клинических наблюдений. Ежедневно добавляется от 500 тыс. до 1,5 млн новых клинических транзакций в режиме реального времени.

Планы на будущее

Исполняемое на суперкомпьютере Университета Индианы программное обеспечение Teradata, PostgreSQL и MongoDB будет поддерживать методы извлечения информации с целью выявления соответствующих клинических признаков (это такие методы, как статистическая мера TF-IDF (от term frequency — inverse document frequency), латентно-семантический анализ (latent semantic analysis, LSA) и статистическая функция «взаимная информация» (mutual information)). Методы обработки естественного языка (natural language processing, NLP) позволят извлечь релевантные клинические признаки. Проверенные признаки будут использоваться для параметризации моделей принятия решений по клиническим фенотипам на основе метода оценки максимального правдоподобия и Байесовских сетей. Модели принятия решений будут использоваться для выявления ряда клинических фенотипов, таких как диабет, хроническая сердечная недостаточность и рак поджелудочной железы.

5.5.2 Вариант использования 17: Анализ графических образов в патологической анатомии/ Цифровая патологическая анатомия

Применение

Анализ цифровых графических образов в патологической анатомии (digital pathology imaging) является нарождающейся областью, в которой изучение сделанных с высоким разрешением изображений образцов тканей позволяет создавать новые и более эффективные способы диагностики заболеваний.

[1)] *В российской литературе термин «Electronic Medical Record», переведенный в настоящем стандарте как «электронная медицинская документация», иногда переводится как «электронная медицинская карта» (ЭМК), «электронная история болезни», «электронный учет здоровья» (ЭУЗ) — см. ГОСТ Р ИСО/ТС 18308—2008; «электронный медицинский учет» (ЭМУ) — см. ГОСТ Р ИСО/ТО 20514—2009. В данном случае речь идет не о некоем едином документе («карте», «истории болезни») и не о виде деятельности («учет»), а о совокупности, относящейся к конкретному пациенту разнообразной документированной информации, содержащейся обычно в ряде независимых источников и представленной в разнообразных форматах.*

В рамках патологического анализа графических изображений выделяется огромное количество пространственных объектов (например, миллионы объектов на изображение), таких как ядра клеток и кровеносные сосуды, представленные их границами, наряду со многими извлеченными по изображению признаками этих объектов. Полученная информация используется для многих сложных запросов и аналитики, поддерживающих биомедицинские исследования и клиническую диагностику.

Текущий подход

Каждое двумерное изображение содержит 1 гигабайт первичных данных изображения, и на его основе производится 1,5 гигабайта аналитических результатов. Для анализа изображений используется интерфейс передачи информации MPI (Message Passing Interface). Информация о решении приведена в А.4.2.

Планы на будущее

Недавно стал возможен патологический анализ трехмерных изображений на основе использования трехмерных лазерных технологий либо последовательного размещения сотен срезов тканей на предметные стекла и их сканирования в цифровые изображения. Выделение трехмерных гистологических объектов на основе серий зафиксированных изображений может породить десятки миллионов трехмерных объектов по одному трехмерному изображению. В результате формируется глубокая «карта» тканей человека для использования в методах диагностики следующего поколения. Трехмерное изображение может содержать 1 терабайт первичных данных изображения, и на его основе производится 1 терабайт аналитических результатов. Средняя по размерам больница будет генерировать 1 петабайт данных в год.

5.5.3 Вариант использования 18: Вычислительный анализ биоизображений (Computational Bioimaging)

Применение

Данные биоизображений все более автоматизировано создаются с более высоким разрешением и являются более мультимодальными. В результате возникает узкое место в анализе данных, устранение которого может способствовать новым открытиям в биологических науках посредством применения технологий больших данных.

Текущий подход

Ныне используемый фрагментарный подход к проведению анализа не масштабируется на ситуации, в которых объем данных в результате одного сканирования на появляющихся устройствах составляет 32 терабайта, а годовой объем медицинских диагностических изображений — около 70 петабайт, не считая данные кардиологии. Для высокопроизводительной, с высокой пропускной способностью обработки изображений в интересах создателей и потребителей моделей, построенных на основе данных биоизображений, необходима единая онлайн-точка обслуживания.

Планы на будущее

Цель заключается в том, чтобы устранить данное узкое место (единую онлайн-точку обслуживания) с помощью экстремально масштабных вычислений и ориентированных на обслуживание сообщества научных порталов, которые применяют средства анализа больших объемов данных к большим наборам данных изображений. Компоненты потока рабочих процессов включают сбор, хранение, улучшение качества данных, минимизацию шума, сегментацию представляющих интерес областей, групповой отбор и извлечение признаков, классификацию объектов, а также организацию и поиск. Возможные пакеты программного обеспечения описаны в А.4.3.

5.5.4 Вариант использования 19: Геномные измерения

Применение

Поддерживаемое американским Национальным институтом стандартов и технологий (NIST) государственно-частно-академическое партнерство «Консорциум «Геном в бутылке»» (Genome in a Bottle Consortium, https://www.nist.gov/programs-projects/genome-bottle) занимается объединением данных, полученных в результате применения различных технологий и методов секвенирования (определения первичной структуры макромолекул) с целью создания высоконадежных описаний полных геномов человека в качестве эталонных материалов. Консорциум также разрабатывает методы использования этих эталонных материалов для оценки эффективности алгоритмов секвенирования генома.

Текущий подход

Используемая NIST сетевая файловая система (network file system, NFS) емкостью примерно 40 терабайт заполнена. «Национальные учреждения здравоохранения» (National Institutes of Health, NIH) и Национальный центр биотехнологической информации (National Center for Biotechnology Information, NCBI) в настоящее время хранят петабайты данных. NIST также хранит данные с использованием про-

граммного обеспечения с открытым исходным кодом для секвенирования в биоинформатике, разработанного академическими группами (на основе UNIX) на 72-ядерном кластере, дополненном более крупными системами участников коллективной работы.

Планы на будущее

Секвенсоры ДНК способны генерировать порядка ~300 гигабайт сжатых данных в день, и эти объемы росли намного быстрее предсказанного законом Мура роста вычислительной мощности компьютеров. В будущем в состав данных могут войти результаты измерений, сделанных в рамках других направлений биологической науки — «омиков» (omics — например, геномика), объем которых будет даже больше, чем объем результатов секвенирования ДНК. В качестве экономически эффективного масштабируемого подхода изучалась возможность использования облачных решений.

5.5.5 Вариант использования 20: Сравнительный анализ метагеномов и геномов

Применение

Использование данного варианта при изучении образцов в метагеномике преследует следующие цели:

- определить состав изучаемой колонии/сообщества с точки зрения присутствия других эталонных изолированных геномов;
- охарактеризовать функции его генов;
- начать выявление возможных функциональных путей (functional pathways);
- охарактеризовать сходство или различие по сравнению с другими метагеномными образцами;
- начать характеризацию изменений в составе и функциях сообщества в связи с изменениями воздействием факторов окружающей среды;
- выделить подразделы данных на основе показателей качества и состава сообщества.

Текущий подход

Современная интегрированная система сравнительного анализа метагеномов и геномов снабжена интерактивным пользовательским веб-интерфейсом. Система включает в себя предварительные вычисления на сервере (backend precomputations) и отправку пакетных заданий из пользовательского интерфейса. Система предоставляет интерфейсы к стандартным инструментам биоинформатики (таким как BLAST, HMMER, инструменты множественного выравнивания последовательностей и филогенетики, программы поиска/предсказания генов и генных структур (gene callers), программы предсказания свойств по результатам секвенирования (sequence feature predictors)).

Планы на будущее

Управление разнородными биологическими данными в настоящее время осуществляется с помощью СУБД (например, Oracle). К сожалению, оно не масштабируется даже для текущего объема в 50 терабайт данных. Решения класса NoSQL (СУБД, существенно отличающиеся от традиционных реляционных) должны были обеспечить альтернативу, но, к сожалению, они не всегда пригодны для интерактивного использования в реальном времени или же для быстрой параллельной массовой загрузки, и иногда у них возникают проблемы с надежностью.

5.5.6 Вариант использования 21: Индивидуальное управление лечением диабета

Применение

Диабет — это болезнь, которая становится все более распространенной среди населения Земли, затрагивая как развивающиеся, так и развитые страны. Современные стратегии управления лечением не учитывают должным образом индивидуальные профили пациентов, в том числе наличие сопутствующих заболеваний и прием соответствующих лекарств — обычное явление у пациентов с хроническими заболеваниями. Для обработки данных в электронных медицинских документах и записях (EHR) следует применять передовые методы интеллектуального анализа данных на основе графов, преобразуя данные в графы RDF (Resource Description Framework)[1)]. Эти передовые методы облегчат поиск пациентов с диабетом и позволят извлечь их медицинские данные для оценки результатов лечения.

Текущий подход

Типичные данные о пациенте включают порядка сотни терминов из контролируемых словарей и тысячу непрерывных числовых величин. Большинство значений свойств снабжены отметками времени.

[1)] *Resource Description Framework (RDF) — среда описания ресурсов, разработанная Консорциумом Всемирной паутины модель для представления данных и особенно метаданных. RDF представляет утверждения о ресурсах в виде, пригодном для машинной обработки. Множество RDF-утверждений образует ориентированный граф, в котором вершинами являются субъекты и объекты, а ребра отображают отношения.* — Википедия, *https://ru.wikipedia.org/wiki/Resource_Description_Framework*

Традиционную парадигму поиска в таблицах реляционной базы данных следует обновить, сменив ее на обход семантического графа.

Планы на будущее

Первым шагом является сопоставление документов пациентов для выявления схожих пациентов в большой базе данных медицинской документации (т. е. формирование индивидуализированной демографической когорты). Необходимо оценить результаты лечения каждого пациента с тем, чтобы выбрать наиболее подходящее решение для конкретного больного диабетом. Зависящие от времени свойства должны быть обработаны перед выполнением запроса для того, чтобы сделать возможным сопоставление на основе производных и других выводимых свойств. Информация о программном обеспечении описана в А.4.6.

5.5.7 Вариант использования 22: Статистический реляционный искусственный интеллект для здравоохранения

Применение

Целью проекта является анализ больших мультимодальных медицинских данных, включая данные различных типов, такие как изображения, электронные медицинские документы и записи (EHR), генетические данные и данные на естественном языке. В рамках этого подхода используются реляционные вероятностные модели, способные работать с богатыми реляционными данными и моделирующие неопределенности на основе теории вероятности.

Программное обеспечение обучает модели на основе различных массивов данных и, возможно, позволит интегрировать информацию и логические рассуждения о сложных запросах. Пользователи могут представить набор сведений, например результаты магнитно-резонансной томографии (МРТ) и демографические данные о конкретном субъекте. Затем они могут сделать запрос о начале конкретного заболевания (например, болезни Альцгеймера), и система выдаст распределение вероятностей для возможного возникновения этого заболевания.

Текущий подход

Один сервер может обрабатывать тестовую когорту из нескольких сотен пациентов, при этом объем соответствующих данных составит сотни гигабайт.

Планы на будущее

В случае когорты из миллионов пациентов придется иметь дело с базами данных петабайтного объема. Основной проблемой является наличие слишком большого количества данных (например, изображений, генетических последовательностей), что может усложнить анализ. Иногда доступны большие объемы данных об одном субъекте, но число субъектов при этом не очень велико (то есть имеется дисбаланс данных). Это может привести к тому, что в ходе анализа алгоритмы обучения расценят случайные корреляции между данными нескольких типов как важные свойства. Еще одна проблема заключается в согласовании и слиянии данных из нескольких источников в форме, полезной для их совместного анализа.

5.5.8 Вариант использования 23: Эпидемиологическое исследование в масштабе всего населения Земли

Применение

Существует потребность в надежном, в режиме реального времени, прогнозировании и контроле над пандемиями, аналогичными пандемии гриппа H1N1 в 2009 г. и COVID19. Борьба с различными видами распространения инфекции может включать моделирование и расчеты, касающиеся распространения информации, болезней и социальных волнений. Модели на основе действующих лиц-агентов могут использовать базовую сеть взаимодействий (т. е. сеть, определяемую моделью людей, транспортных средств и их деятельности) для изучения эволюции рассматриваемых явлений.

Текущий подход

Используется двухэтапный подход: (1) сформировать синтетическую глобальную популяцию; и (2) провести моделирование в масштабе глобальной популяции с тем, чтобы сделать выводы о вспышках заболеваемости и различных стратегиях вмешательства. Текущий набор данных объемом 100 терабайт был сгенерирован централизованно с помощью написанной на Charm ++ системы моделирования, использующей интерфейс передачи сообщений MPI (Message Passing Interface). Параллелизм достигается за счет использования меры «время присутствия болезни» (disease residence time period).

Планы на будущее

Для изучения сложных проблем глобального масштаба могут быть использованы большие модели распространения социального влияния (social contagion models), что значительно увеличит размер используемых систем.

5.5.9 Вариант использования 24: Применение моделирования распространения социального влияния в планировании, здравоохранении и менеджменте катастроф

Применение

Модели социального поведения применимы в сферах национальной безопасности, здравоохранения, вирусного маркетинга, городского планирования и обеспечения готовности к чрезвычайным ситуациям и катастрофам.

В случае социальной напряженности и волнений люди выходят на улицы, чтобы выразить свое недовольство либо поддержку руководству государства. Модели могли бы помочь количественно определить степень, в которой деловая деятельность и активность населения нарушаются из-за страха и гнева; вероятность мирных демонстраций и/или насильственных протестов; а также диапазон возможных ответных мер правительства, начиная от умиротворения, разрешения протестов и до угроз в адрес протестующих и действий по срыву протестов. Для решения таких задач потребуются модели и наборы данных с высоким разрешением (на уровне отдельных лиц, транспортных средств и зданий).

Текущий подход

Инфраструктура модели распространения социального влияния (social contagion model) представляет различные типы взаимодействия между людьми (например, лицом к лицу, через социальные сети), а также между людьми и сервисами (например, транспорт) либо инфраструктурой (например, интернет, электроснабжение). Эти модели деятельности генерируются на основе усредненных данных, таких как данные переписи населения.

Планы на будущее

Одной из важных проблем является объединение данных (data fusion —как комбинировать данные из разных источников и что делать в случае отсутствия или неполноты данных). Правильно организованный процесс моделирования должен учитывать разнородные особенности сотен миллионов или миллиардов людей, а также культурные различия в разных странах. Для таких больших и сложных моделей сам по себе процесс их валидации также представляет собой проблему.

5.5.10 Вариант использования 25: Биологическое разнообразие и европейская электронная научно-исследовательская инфраструктура LifeWatch

Применение

Мониторинг и исследование различных экосистем, биологических видов, их динамики и миграции с помощью набора специализированных датчиков и доступа/обработки получаемых данных, а также посредством кооперации с соответствующими проектами в данной области. В числе конкретных тематических исследований можно назвать мониторинг чужеродных видов, мигрирующих птиц и водно-болотных угодий.

Одно из направлений деятельности консорциума под названием «Совместная деятельность европейских сетевых инфраструктур в области экологических исследований» (ENVRI) заключается в изучении интеграции инфраструктуры LifeWatch с другими электронными инфраструктурами экологических исследований.

Текущий подход

В настоящее время данный проект находится на стадии предварительного планирования и, соответственно, текущий подход не полностью проработан.

Планы на будущее

Проект LifeWatch обеспечит интегрированный доступ к различным данным, инструментам аналитики и моделирования, предоставленными другими проектами. Он также будет предлагать данные и инструменты в составе отдельных рабочих процессов конкретным научным сообществам. Помимо этого LifeWatch предоставит возможности для создания персонализированных «виртуальных лабораторий», позволяя участникам вводить и получать доступ к новым данным и аналитическим инструментам.

Новые данные будут коллективно использоваться сотрудничающими с LifeWatch центрами обработки данных, включая Всемирную систему информации о биоразнообразии (Global Biodiversity Information Facility) и Каталог биоразнообразия (Biodiversity Catalogue), известный как Реестр вебсервисов науки о биоразнообразии (Biodiversity Science Web Services Registry). В состав данных входят данные других направлений биологической науки — «омиков» (omics), сведения о биологических видах, экологическая информация (например, сведения о биомассе, плотности населения) и данные об экосистеме (например, о потоках диоксида углерода CO_2, о цветении водорослей, характеристики воды и почвы).

5.6 Глубокое обучение (Deep Learning) и социальные сети

5.6.1 Вариант использования 26: Крупномасштабное глубокое обучение

Применение

Существует потребность в увеличении объема массивов данных и размера моделей, с которыми способны работать алгоритмы глубокого обучения. Большие модели (например, нейронные сети с большим количеством нейронов и соединений) в сочетании с большими массивами данных все чаще показывают наилучшие результаты при выполнении эталонных задач в области зрения, речи и обработки естественного языка. Необходимо будет обучать глубокую нейронную сеть на большом (например, намного более 1 терабайта) массиве данных, обычно состоящем из изображений, видео-, аудиоматериалов или текста. Такие процедуры обучения часто требуют специфической настройки архитектуры нейронной сети, критериев обучения и предварительной обработки данных. Помимо вычислительных затрат, которых требуют алгоритмы обучения, чрезвычайно высока потребность в быстрой разработке прототипа и удобстве разработки.

Текущий подход

На сегодняшний день наиболее крупными приложениями являются распознавание изображений и научные исследования в области обучения без учителя, проводимые на высокопроизводительном кластере из 64 графических процессоров с коммутационной сетью Infiniband, в которых используется 10 млн изображений и до 11 млрд параметров. Изучаются как машинное обучение с учителем (т. е. использующее существующие классифицированные изображения), так и обучение без учителя.

Планы на будущее

Массивы данных объемом 100 терабайт и более могут стать необходимыми для использования репрезентативной способности более крупных моделей. Для обучения беспилотного автомобиля могут потребоваться 100 млн изображений в мегапиксельном разрешении. Глубокое обучение имеет много общих черт с более широкой областью машинного обучения. Первостепенными требованиями являются высокая вычислительная пропускная способность (computational throughput) главным образом для операций линейной алгебры с плотными матрицами, а также чрезвычайно высокая эффективность научного труда. Высокопроизводительные библиотеки должны быть интегрированы с высокоуровневыми средами разработки прототипов.

5.6.2 Вариант использования 27: Организация крупномасштабных, неструктурированных коллекций сделанных потребителями фотографий

Применение

Коллекции, содержащие от миллионов до миллиардов сделанных потребителями фотографий, используются для создания трехмерных реконструкций сцен при отсутствии априорных знаний как о структуре сцены, так и положениях камеры. Получающиеся в результате трехмерные модели позволяют эффективно и результативно организовать просмотр больших коллекций фотографий по географическому положению. Географическая привязка новых изображений может осуществляться путем сопоставления их с трехмерными моделями, и для каждого изображения может быть выполнено распознавание объектов. Задачу трехмерной реконструкции можно сформулировать как задачу робастной нелинейной оптимизации с использованием метода наименьших квадратов; наблюдаемые (зашумленные) соответствия между изображениями являются ограничениями, а в число неизвестных входят 6-мерные координаты, задающие положение камеры для каждого изображения, и 3-мерные координаты положения каждой точки сцены.

Текущий подход

Текущая информация о системе приведена в А.5.2. В социальных сетях в настоящее время размещено более 505 млрд изображений, и каждый день на сайты социальных сетей добавляется более 500 млн изображений.

Планы на будущее

В рамках технического обслуживания и обновлений необходимо добавить большое число инструментов аналитики, включая инструменты извлечения признаков, сопоставления признаков и крупномасштабную машину вероятностных логических выводов. Эти инструменты аналитики используются при решении многих или даже большинства проблем компьютерного зрения и обработки изображений, включая распознавание, разделение по глубине (stereo resolution) и устранение шума в изображениях. В числе иных потребностей можно назвать визуализацию крупномасштабных трехмерных реконструкций и навигацию по крупномасштабным коллекциям изображений, которые были согласованы с картами.

5.6.3 Вариант использования 28: Truthy — Исследование распространения информации на основе данных Твиттера

Применение

Необходимо лучше понимать, как информация распространяется по социально-техническим сетям, и требуются методы для обнаружения потенциально опасной информации (например, вводящих в заблуждение сообщений, скоординированных кампаний и недостоверной информации) на ранних стадиях ее распространения.

Текущий подход

Твиттер генерирует непрерывный поток данных большого объема — около 30 терабайт в год в сжатом виде — посредством распространения примерно 100 млн сообщений в день. Темпы роста объемов данных составляют примерно 500 гигабайт в день. Все эти данные должны быть собраны и сохранены. Дополнительные потребности включают анализ таких данных в режиме времени, близком к реальному, с целью выявления аномалий, кластеризации потока, классификации сигналов и онлайн-обучения; а также поиск данных, визуализацию больших данных, интерактивные веб-интерфейсы и общедоступные программные интерфейсы (API) для запросов к данным. Сведения о программном обеспечении приведены в А.5.4. Возможности для анализа процесса распространения информации, для кластеризации и динамической визуализации уже существуют.

Планы на будущее

Планируется расширение проекта, поэтому необходимо двигаться в сторону описанных в А.5.4 развитых программ распределенного хранения и базы данных, располагающейся в оперативной памяти компьютера, с целью обеспечения анализа в реальном времени. Решения должны включать кластеризацию потока, обнаружение аномалий и онлайн-обучение.

5.6.4 Вариант использования 29: Краудсорсинг в гуманитарных науках как источник больших и динамических данных

Применение

Информация собирается у многих людей и с их устройств с использованием ряда источников данных: ручного ввода, записанных мультимедийных материалов, времени реагирования, изображений, информации с датчиков. Эти данные используются для характеризации широкого спектра индивидуальных, социальных, культурных и лингвистических вариаций в нескольких измерениях (например, в пространстве, социальном пространстве, во времени).

Текущий подход

На данный момент типичным является использование расширяемого языка разметки (XML) и традиционных реляционных баз данных. Пока что помимо изображений используется не очень много мультимедийных материалов.

Планы на будущее

Краудсорсинг начинает использоваться в более широком масштабе. Наличие датчиков в мобильных устройствах создает огромный потенциал для сбора большого количества данных от многочисленных физических лиц. Эта возможность до настоящего времени в широком масштабе не опробовалась; существующие краудсорсинговые проекты обычно имеют ограниченный масштаб и основаны на веб-технологиях. Могут возникнуть проблемы с обеспечением защиты персональных данных в связи с доступом к аудиовизуальным файлам физических лиц; анонимизация может быть необходима, но она не всегда возможна. Важное значение имеют управление данными и их курирование. В случае обработки мультимедийных материалов объем данных может составлять сотни терабайт.

5.6.5 Вариант использования 30: Цифровая инфраструктура для исследований и анализа сетей и графов (CINET)

Применение

CINET предоставляет общую веб-платформу, обеспечивающую конечному пользователю бесперебойный доступ:

- к инструментам анализа сетей и графов, таким как SNAP, NetworkX и Galib;
- к созданным для решения реальных задач и синтезированным сетям;
- к вычислительным ресурсам;
- к системе управления данными.

Текущий подход

CINET используют как сервис высокопроизводительного вычислительного кластера с 720 ядрами и соединениями на основе InfiniBand. Платформа используется для научных исследований и в обра-

зовательных целях. CINET используется специалистами в области общественных наук и социального взаимодействия на занятиях и для поддержки исследований.

Планы на будущее

Ожидается быстрое расширение хранилища, в котором примерно через год будет храниться как минимум от одной до 5 тыс. сетей и методов. Поскольку все больше дисциплин используют графы увеличивающегося размера, будут важны параллельные алгоритмы. Двумя ключевыми проблемами являются манипулирование данными и учет производных данных, поскольку отсутствуют четко определенные и эффективные модели и инструменты для унифицированного управления различными данными графов.

5.6.6 Вариант использования 31: Измерения, оценки и стандарты эффективности аналитических технологий в отделе доступа к информации NIST

Применение

Для создания основ и ускорения дальнейшего развития передовых аналитических технологий в областях обработки речи и языка, видеозаписей и мультимедийных материалов, биометрических изображений и неоднородных данных необходимы метрики эффективности, методы измерения и проведение оценок сообществом, а также взаимодействие аналитиков с пользователями.

Обычно применяется одна из двух моделей обработки:

1) предоставить участникам тестирования тестовые данные и проанализировать выходные данные систем — участников, и

2) предоставить участникам интерфейсы к тестовой обвязке для алгоритмов, взять их алгоритмы и провести тестирование алгоритмов на внутренних вычислительных кластерах.

Текущий подход

Для целей обучения, испытаний в ходе разработки и итоговых оценок имеются большие аннотированные совокупности неструктурированного/полуструктурированного текста, аудио- и видеозаписей, изображений, мультимедийных материалов и разнородные коллекции вышеперечисленного, включая аннотации о точности и достоверности (ground truth). В составе этой совокупности более 900 млн веб-страниц общим объемом 30 терабайт, 100 млн твиттов, 100 млн проверенных биометрических изображений, несколько сотен тысяч частично проверенных видеоклипов и терабайты более мелких полностью проверенных тестовых коллекций.

Планы на будущее

Для будущих оценок аналитики планируется собрать еще большие коллекции данных с использованием нескольких потоков данных, включая очень неоднородные данные. В дополнение к более крупным массивам данных в будущем предполагается тестирование потоковых алгоритмов на различных неоднородных данных. Изучается возможность использования облаков.

5.7 Экосистема для исследований

5.7.1 Вариант использования 32: Консорциум федеративных сетей данных (DFC)

Применение

Консорциум федеративных сетей данных (DFC) содействует совместным и междисциплинарным исследованиям посредством объединения на федеративных началах систем управления данными, используемых федеральными органами и учреждениями США, национальными академическими научно-исследовательскими инициативами, хранилищами учреждений и участниками международного сотрудничества. Эта масштабная среда совместной работы включает петабайты данных, сотни миллионов файлов, сотни миллионов атрибутов метаданных, десятки тысяч пользователей и тысячу ресурсов хранения.

Текущий подход

В настоящее время в 25 областях науки и техники имеются проекты, полагающиеся на интегрированную систему управления данными, основанную на использовании правил (iRODS). В числе активных пользователей можно назвать:

- Национальный научный фонд США, со следующими крупными проектами:

1) «Инициатива океанических наблюдательных станций» (Ocean Observatories Initiative) — архивация показаний датчиков;

2) «Динамика во времени учебного центра» (Temporal Dynamics of Learning Center) — грид-система управления данными для науки о процессах познания;

3) проект создания киберинфраструктуры для ботаники (iPlant Collaborative) — геномика растений;

- проект электронной инженерной библиотеки Университета им. Дрекселя (Drexel University);

- Институт социальных наук им. Говарда Одума (H.W.Odum Institute for Research in Social Science) при Университете Северной Каролины в Чапел-Хилл — объединение грид-системы управления данными с открытым программным обеспечением для управления научно-исследовательскими данными Dataverse.

В настоящее время iRODS управляет петабайтами данных, сотнями миллионов файлов, сотнями миллионов атрибутов метаданных, десятками тысяч пользователей и тысячей ресурсов хранения. iRODS взаимодействует с системами управления потоками рабочих процессов [такими как решение Cyberintegrator Национального центра компьютерных приложений (National Center for Computing Applications, NCSA), Kepler, Taverna], совместим с облачными и более традиционными моделями хранения, а также поддерживает различные транспортные протоколы.

Планы на будущее

Будущие сценарии использования и приложения данных не были представлены для этого варианта использования.

5.7.2 Вариант использования 33: «Discinnet-процесс»

Применение

Компания Discinnet Labs разработала прототип «Веб 2.0» — платформы для совместной работы, которая, в качестве пилотной системы, в настоящее время развертывается и тестируется исследователями из растущего числа различных областей науки.

Цель заключается в том, чтобы набрать достаточно большую выборку активных областей исследований, представленных в виде кластеров (то есть исследователей, отображенных и агрегируемых в рамках множества главным образом коллективных экспериментальных измерений), с тем чтобы проверить общие, а следовательно, потенциально междисциплинарные, эпистемологические модели в течение текущего десятилетия.

Текущий подход

В настоящее время активировано 35 кластеров, и еще около 100 ждут, пока будут выделены дополнительные ресурсы. Существует потенциал для сознания исследовательскими сообществами управления и модерирования многих других кластеров. Примеры кластеров включают в себя оптику, космологию, материаловедение, микроводоросли, здравоохранение, прикладную математику, вычисления, резину и другие химические продукты/проблемы.

Планы на будущее

Сам по себе «Discinnet-процесс» не является большими данными. Скорее, он будет генерировать метаданные при применении к кластеру, который включает большие данные. При междисциплинарной интеграции нескольких предметных областей процесс будет согласовывать метаданные многих уровней сложности.

5.7.3 Вариант использования 34: Поиск по семантическому графу для текстовых научных данных по химии

Применение

Для аннотирования и представления информации о технологиях создаются инфраструктура на основе социальных сетей, терминология и семантические графы данных. В этом процессе используются методы, основанные на корневых морфемах (root-based) и правилах (rule-based), которые в настоящее время главным образом ориентированы на определенные индоевропейские языки, такие как санскрит и латынь.

Текущий подход

Во многих отчетах, в том числе в недавнем отчете по проекту «Геном материала» (Materials Genome Initiative), отмечается, что исключительно нисходящие решения, облегчающие обмен данными и интеграцию, нежелательны в случае междисциплинарных усилий. В то же время подход «снизу вверх» может быть хаотичным. По этой причине существует потребность в сбалансированном сочетании двух подходов с целью поддержки простых в использовании методов создания, интеграции и обмена метаданными. Эта проблема очень похожа на проблему, с которой сталкиваются разработчики языка, поэтому недавно разработанный метод основан на этих идеях. В настоящее время предпринимаются усилия по распространению этого метода на публикации, представляющие интерес для инициативы «Геном материала», движения «Открытое правительство», а также для «Сети интегрированных знаний NIST — EditorialNet» (NIKE) — архива публикаций американского Национального института стандартов и технологий (NIST). Эти усилия являются частью деятельности рабочей группы «Справочник стандартов метаданных» (Metadata Standards Directory) Альянса научных данных (Research Data Alliance).

Планы на будущее

Должна быть создана облачная инфраструктура для социальных сетей научной информации. Ученые всего мира смогут использовать эту инфраструктуру для участия и размещения результатов своих экспериментов. Перед созданием научной социальной сети необходимо решить некоторые вопросы, включая следующие:

- минимизировать проблемы, связанные с созданием повторно используемого, междисциплинарного, масштабируемого по требованию, дружественного по отношению к варианту использования и пользователю словаря;
- использовать существующий или создать новый индивидуализированный граф данных для размещения информации интуитивно понятным способом, таким образом, чтобы он легко интегрировался с существующими графами данных в объединенной среде независимо от специфики управления данными;
- найти адекватные научные данные, не проводя чересчур много времени в интернете.

Начать предполагается с таких ресурсов, как движение «Открытое правительство», инициатива «Геном материала» и «Банк данных белковых структур» (Protein Data Bank, PDB). Эти усилия охватят множество локальных и сетевых ресурсов. Разработка инфраструктуры для автоматической интеграции информации из всех этих ресурсов с использованием графов данных является сложной задачей, однако предпринимаются шаги для ее решения. Необходимы мощные инструменты базы данных и серверы для манипулирования графами данных.

5.7.4 Вариант использования 35: Анализ больших объемов данных, получаемых в экспериментах на синхротроне

Применение

Образцы подвергаются воздействию рентгеновского излучения от источников излучения в различных конфигурациях, в зависимости от эксперимента. Данные собираются детекторами, которые фактически представляют собой высокоскоростные цифровые фотокамеры. Затем данные анализируются с целью восстановления вида исследуемого образца или процесса.

Текущий подход

Для анализа данных используется различное программное обеспечение, как коммерческое, так и с открытым исходным кодом. Передача данных осуществляется посредством физического перемещения портативных носителей информации (что сильно ограничивает производительность); либо с использованием высокопроизводительного протокола GridFTP в реализации компании Globus Online и систем управления потоками рабочих процессов, таких как программная инфраструктура с открытым исходным кодом (Support for Provenance Auditing in Distributed Environments — «Поддержка аудита происхождения в распределенных средах»).

Планы на будущее

Разрешение фотокамер постоянно увеличивается. Становится необходимой передача данных в крупномасштабные вычислительные центры из-за вычислительной мощности, необходимой для проведения анализа в разумные, с точки зрения эксперимента, сроки. Из-за большого количества каналов отвода излучения к экспериментальным установкам (их, например, 39 у синхротрона Advanced Light Source (ALS) Национальной лаборатории имени Лоуренса в Беркли, США (LBNL), совокупное производство данных, вероятно, значительно возрастет в ближайшие годы, равно как и потребность в обобщенной инфраструктуре для анализа гигабайт данных в секунду, поступающих от множества детекторов на ряде экспериментальных установок.

5.8 Астрономия и физика

5.8.1 Вариант использования 36: Каталинский обзор оптических переходных процессов в режиме реального времени (CRTS) — цифровой, панорамный, синоптический обзор неба

Применение

В рамках проекта «Каталинский обзор оптических переходных процессов в режиме реального времени» (CRTS) проводятся исследования меняющейся Вселенной в диапазоне видимого света, в масштабах времени, варьирующихся от минут до лет, путем поиска переменных и транзиентных (непостоянных, преходящих) источников.

Такие исследования позволяют выявить широкий спектр астрофизических объектов и явлений, включая различные типы космических взрывов (например, сверхновых), переменные звезды, явления,

связанные с аккрецией на массивные черные дыры (примером служат активные галактические ядра) и их релятивистские потоки частиц и энергий, и звезды с большим собственным движением.

Данные поступают с трех телескопов (два в Аризоне, США и один в Австралии), и в ближайшем будущем ожидается подключение дополнительных телескопов в Чили.

Текущий подход

В ходе обзора создается примерно до 0,1 терабайта данных в ясную ночь, а суммарный объем фондов данных составляет в настоящее время около 100 терабайт. Данные предварительно обрабатываются на телескопе, а затем передаются в Университет Аризоны и Калифорнийский технологический институт (Caltech) для дальнейшего анализа, распространения и архивирования.

Данные обрабатываются в режиме реального времени, а обнаруженные транзиентные события публикуются с использованием различных электронных механизмов распространения, без использования проприетарного периода отсрочки до широкого распространения данных (CRTS использует политику полностью открытых данных).

Дальнейший анализ данных включает классификацию обнаруженных транзиентных событий, дополнительные наблюдения с использованием других телескопов, научную интерпретацию и публикацию. В этом процессе интенсивно используются архивные данные (несколько петабайт) из широкого спектра географически распределенных ресурсов, объединенных структурой Виртуальной обсерватории.

Планы на будущее

Проект CRTS является научным и методологическим испытательным стендом и предшественником предстоящих крупных обзоров, которые будут проводиться, в частности, Большим синоптическим обзорным телескопом в Обсерватории имени Веры Рубин, Чили (LSST). Этот телескоп, который, как ожидается, войдет в эксплуатацию в 2020-х гг., в «Астрономическом и астрофизическом ежедекадном обозрении» (Astronomy and Astrophysics Decadal Survey) 2010 г. признан наиболее приоритетным наземным инструментом. Телескоп LSST будет собирать около 30 терабайт данных за ночь.

Потоки данных обзора от телескопов (размещенных на земле или в космосе) формируют потоки данных о транзиентных событиях. Данные о событиях вместе с их качественными описаниями поступают на хранение в одно или несколько хранилищ, которые могут распространять их в электронном виде для астрономов или роботизированных телескопов. С каждым событием ассоциируется пополняющийся портфель информации, который включает в себя все доступные данные о конкретной небесной позиции. Данные собираются из разнообразных архивов, объединенных в структуре Виртуальной обсерватории, из аннотаций экспертов и т. д.

Представления такой объединенной информации могут быть как человекочитаемыми, так и машиночитаемыми. Данные поступают в один или несколько автоматических механизмов определения характеристик, классификации и приоритизации, которые используют различные инструменты машинного обучения для выполнения этих задач.

Выходные данные этих механизмов, которые динамически эволюционируют по мере поступления и обработки новой информации, учитываются при планировании последующих наблюдениях за избранными событиями, а полученные в ходе таких наблюдений данные передаются обратно в портфели событий для следующей итерации.

Пользователи, как люди, так и автоматы, могут подключаться к системе во многих точках для поиска и извлечения информации и для предоставления новой информации посредством использования стандартизированного набора форматов и протоколов. Это может быть сделано в режиме почти реального времени либо в «архивном» режиме (когда время не является критическим фактором).

5.8.2 Вариант использования 37: Проект Министерства энергетики США анализа экстремально больших данных космологических обзоров неба и моделирования

Применение

Инструмент выявления космологических явлений объединяет моделирование и данные наблюдений с тем, чтобы прояснить природу темной материи, темной энергии и инфляции, — это вопросы, которые относятся к числу самых волнующих, озадачивающих и проблемных, которые стоят перед современной физикой, включая вопрос о влиянии свойств элементарных частиц на раннюю Вселенную. В ходе моделирования будут создаваться данные в объемах, сопоставимых с объемами данных наблюдений.

Текущий подход

В настоящее время данный проект находится на стадии предварительного планирования и, соответственно, текущий подход не полностью разработан.

Планы на будущее

Такого рода системы будут использовать колоссальное количество суперкомпьютерного времени — более 200 млн часов. Соответствующие объемы данных следующие:

- обзор «Темная энергия» (Dark Energy Survey, DES): 4 петабайта в год в 2015 г.;
- обзор Zwicky Transient Factory (ZTF): 1 петабайт в год в 2015 г.;
- большой синоптический обзорный телескоп в Обсерватории имени Веры Рубин, Чили (LSST) — 7 петабайт в год в 2019 г. (см. описание проекта CRTS в 5.8.1);
- моделирование: 10 петабайт в год в 2017 г.

5.8.3 Вариант использования 38: Большие данные космологических обзоров

Применение

При выполнении обзора «Темная энергия» (Dark Energy Survey, DES) данные с вершины горы передаются по микроволновой связи в чилийский город Ла Серена (La Serena). Оттуда по оптическим каналам связи они поступают в американский Национальный центр компьютерных приложений (National Center for Computing Applications, NCSA) и Национальный научно-исследовательский вычислительный центр энергетических исследований Министерства энергетики США (NERSC) для хранения и «редуцирования». Здесь проводится идентификация и каталогизация галактик и звезд как на отдельных изображениях, так и на сериях изображений, и, наконец, их характеристики измеряются и сохраняются в базе данных.

Текущий подход

Работают конвейеры «вычитания» с использованием существующих изображений с целью найти новые оптические транзиенты при помощи алгоритмов машинного обучения. Технологии работы с данными и аппаратные ресурсы описаны в А.7.3.

Планы на будущее

Необходимы методы для выполнения разложения Холецкого для тысяч моделирований с матрицами порядка миллиона по каждой стороне и параллельное хранение изображений. Телескоп LSST создаст 60 петабайт графических данных и 15 петабайт данных каталога, и будет создан соответственно большой (или даже больший) объем данных моделирования. В общей сложности за ночь будет создаваться более 20 терабайт данных.

5.8.4 Вариант использования 39: Физика элементарных частиц — Анализ данных «Большого адронного коллайдера»: открытие бозона Хиггса

Применение

Проводится анализ соударений на ускорителе «Большого адронного коллайдера» (БАК — Large Hadron Collider, LHC) Европейского центра ядерных исследований ЦЕРН (CERN).

Обработанная информация описывает физические свойства событий, и на ее основе создаются списки частиц с указанием их типа и импульса. Эти события анализируются с целью обнаружения новые явлений, как новых частиц (например, бозона Хиггса), так и сбора доказательств того, что предполагаемые частицы (предсказываемые, например, теорией суперсимметрии) не были обнаружены. На Большом адронном коллайдере проводится несколько крупных экспериментов, включая «Тороидальный детектор БАК» ATLAS (A Toroidal LHC ApparatuS) и «Компактный мюонный соленоид» (Compact Muon Solenoid, CMS). В этих экспериментах принимают участие представители глобального научного сообщества (например, в эксперименте CMS насчитывается 3600 участников из 183 учреждений 38 стран), поэтому данные на всех уровнях передаются и являются доступными на всех континентах.

Текущий подход

Эксперименты на Большом адронном коллайдере являются пионерами в области распределенной инфраструктуры больших данных. Ряд аспектов потока рабочих процессов этих экспериментов высвечивают задачи, которые в рамках других дисциплин тоже нужно будет решить. В числе этих задач — автоматизация распределения данных, высокопроизводительная передача данных и крупномасштабные вычисления с большой пропускной способностью.

В рамках анализа на гриде данных, проводившегося для обнаружения бозона Хиггса, использовались 350 тысяч ядер, работавших почти непрерывно, выполняя в день более двух миллионов заданий, распределенных по трем основным уровням: ЦЕРН, континенты/страны и университеты.

Для анализа используется распределенная архитектура для вычислений с высокой пропускной способностью (т. е. комфортабельно-параллельная), в рамках которой участвующие вычислительные центры объединены в мировом масштабе с помощью «Всемирного вычислительного грида Большого адронного коллайдера» (Worldwide LHC Computing Grid, WLCG) и, в США, «Грида открытой науки» (Open Science Grid).

В общей сложности в ходе экспериментов на ускорителе и при анализе их результатов ежегодно создается 15 петабайт данных, а суммарный объем данных составляет 200 петабайт. В частности, в 2012 г. эксперимент ATLAS хранил 8 петабайт на магнитной ленте для обеспечения первого уровня хранения Tier-1 и более 10 петабайт на диске уровня Tier-1 в Брукхейвенской национальной лаборатории (BNL), и 12 петабайт в кэш памяти на дисках в американских центрах уровня Tier-2. В рамках эксперимента CMS объемы данных аналогичны. Более половины ресурсов используется для моделирования по методу Монте-Карло, а не для анализа данных.

Планы на будущее

В прошлом сообщество специалистов в области физики элементарных частиц могло рассчитывать на то, что промышленность обеспечит во времени экспоненциальный рост производительности в расчете на единицу затрат в соответствии с законом Мура. Однако в будущем доступную производительность будет гораздо сложнее использовать, поскольку технологические ограничения, связанные, в частности, с энергопотреблением, привели к глубоким изменениям в архитектуре современных микросхем центральных процессоров (CPU).

В прошлом программное обеспечение могло использоваться без изменений на последовательных поколениях процессоров и достигать соответствующего закону Мура прироста производительности благодаря регулярному повышению тактовой частоты процессоров, которое продолжалось до 2006 г. Эра масштабирования последовательных приложений на процессорах, построенных на неоднородных элементах (heterogeneous element processor, HEP), теперь уже закончилась. Изменения в архитектуре центральных процессоров предполагают значительно больший параллелизм программного обеспечения, а также использование специализированных возможностей для вычислений с плавающей запятой.

Структура и производительность программного обеспечения для обработки данных физики высоких энергий должны быть изменены таким образом, чтобы его можно было продолжать адаптировать и развивать, обеспечивая его эффективную работу на новом оборудовании. Это означает серьезную смену парадигмы в разработке программного обеспечения для физики высоких энергий и подразумевает крупномасштабную реорганизацию структур данных и алгоритмов. Параллелизм необходимо добавлять одновременно на всех уровнях: на уровне событий, на уровне алгоритма и на суб-алгоритмическом уровне. Компоненты на всех уровнях стека программного обеспечения должны быть способны взаимодействовать, поэтому цель заключается в том, чтобы максимально стандартизировать типовые проектировочные решения и выбор модели параллелизма. Это также поможет обеспечить эффективное и сбалансированное использование ресурсов.

5.8.5 Вариант использования 40: Эксперимент Belle II в области физики высоких энергий

Применение

«Belle» — это эксперимент в области физики элементарных частиц, в рамках которого более 400 физиков и инженеров исследуют эффекты нарушения зарядовой четности (CP-инвариантности) при получении B-мезонов на ускорителе высоких энергий — электронно-позитронном коллайдере KEKB, находящемся в Цукубе, Япония. В частности, идет поиск различных мод распада в мезонном резонансе Y(4S) с целью обнаружения новых явлений, выходящих за рамки стандартной модели физики элементарных частиц.

Данный ускоритель имеет наибольшую интенсивность из всех существующих в мире, но события проще, чем те, что наблюдаются на «Большом андронном коллайдере» (LHC), и поэтому анализ менее сложен, но по стилю похож на анализ данных ускорителя LHC в ЦЕРН.

Текущий подход

В настоящее время данный проект находится на стадии предварительного планирования и, соответственно, текущий подход не полностью разработан.

Планы на будущее

Модернизированный эксперимент Belle II и ускоритель SuperKEKB начали работу в 2015 г. Объем данных увеличится в 50 раз, при этом суммарный объем интегрированных первичных данных составил около 120 петабайт, физических данных — около 15 петабайт, данных моделирования по методу Монте-Карло — около 100 петабайт.

На новом этапе потребуется переход к модели распределенных вычислений, требующей непрерывной передачи необработанных данных со скоростью ~20 гигабит в секунду между Японией и США при проектной яркости ускорителя. Необходимое программное обеспечение описано в А.7.5.

5.9 Науки о Земле, экологические науки и полярные исследования

5.9.1 Вариант использования 41: Радарная система некогерентного рассеяния EISCAT-3D Европейской научной ассоциации по некогерентному рассеянию радиоволн

Применение

Европейская научная ассоциация по некогерентному рассеянию радиоволн (European Incoherent Scatter Scientific Association, EISCAT) проводит исследования нижней, средней и верхней атмосферы и ионосферы с использованием радарных систем некогерентного рассеяния. Эти установки являются наиболее мощными наземными инструментами, используемыми в такого рода исследованиях. EISCAT изучает нестабильности в ионосфере и исследует структуру и динамику средней атмосферы. В экспериментах по искусственной модификации ионосферы EISCAT использует измерительный комплекс в сочетании с отдельным нагревным стендом. В настоящее время EISCAT эксплуатирует три из десяти основных радарных систем некогерентного рассеяния в мире. Эти три системы расположены в скандинавском секторе к северу от полярного круга.

Текущий подход

Эксплуатируемая в настоящее время радарная система EISCAT производит данные со скоростью несколько терабайт в год. Каких-либо особых проблем у системы нет.

Планы на будущее

Конструктивно радарная система следующего поколения EISCAT-3D будет состоять из центрального радиолокационного поста с приемными и передающими антенными решетками, и четырех приемных постов с приемными антенными решетками на расстоянии около 100 км от центрального поста.

Полностью функциональная система из пяти постов будет производить в несколько тысяч раз большие объемы данных по сравнению с ныне используемой системой EISCAT, на уровне 40 петабайт в год в 2022 г. и, как ожидается, будет эксплуатироваться в течение 30 лет.

В электронной инфраструктуре данных эксперимента EISCAT-3D планируется использовать высокопроизводительные компьютеры для обработки данных в основном центре и компьютеры с высокой пропускной способностью в зеркальных центрах обработки данных. Операция скачивания всего массива данных не является критичной ко времени, однако для оперативного управления требуется информация в реальном времени о некоторых заранее определенных событиях, которая будет поступать с постов в центр управления, а также связь в реальном времени центра управления с постами для установления в реальном времени режима работы радара.

5.9.2 Вариант использования 42: «Совместная деятельность европейских сетевых инфраструктур в области экологических исследований» (ENVRI)

Применение

Предметом проекта «Совместная деятельность европейских сетевых инфраструктур в области экологических исследований» (ENVRI) являются европейские распределенные, рассчитанные на длительную перспективу, дистанционно управляемые сети наблюдений, ориентированные на понимание процессов, тенденций, порогов, взаимодействий и обратных связей, а также на повышение предсказательной способности в интересах разрешения будущих экологических проблем. Следующие усилия являются частью проекта ENVRI:

- «Интегрированная система наблюдения за выбросами углерода» ICOS (Integrated Carbon Observation System) — европейская распределенная инфраструктура, предназначенная для мониторинга парниковых газов через ее атмосферные, экосистемные и океанские сети наблюдений;
- EURO-Argo — европейский вклад в международную систему наблюдений за океаном Argo;
- проект EISCAT-3D (описан в отдельном варианте применения № 41) — европейская исследовательская радарная система некогерентного рассеяния нового поколения для исследований верхней атмосферы;
- проект LifeWatch (описан в отдельном варианте применения № 25) — европейская электронная инфраструктура для исследований в области экологии и биологического разнообразия;
- «Европейская исследовательская инфраструктура для слежения за [геологическими] плитами» EPOS (European Plate Observing System) — это европейская инфраструктура для исследования землетрясений, вулканов, динамики поверхности и тектоники;
- «Европейская междисциплинарная обсерватория исследования морского дна и слоев воды» (EMSO) — европейская сеть наблюдательных станций морского дна, предназначенная для мониторинга в долгосрочном масштабе времени экологических процессов, связанных с экосистемами, изменением климата и геологическими опасностями;

- проект «Использование самолетов в глобальной системе наблюдений» (IAGOS) организует сеть самолетов для глобального наблюдения за атмосферой;

- проект «Интегрированная система наблюдений за Арктикой на Шпицбергене» (SIOS) создает систему наблюдений на Шпицбергене и вокруг него, которая объединяет исследования геофизических, химических и биологических процессов, проводимые на всех платформах исследований и мониторинга.

Текущий подход

В рамках проекта ENVRI разрабатывается эталонная модель (ENVRI RM) в качестве общей онтологической структуры и стандарта для описания и характеризации вычислительной инфраструктуры и инфраструктуры хранения. Цель состоит в том, чтобы обеспечить бесперебойную интероперабельность между неоднородными ресурсами различных инфраструктур. Модель ENVRI RM служит языком общения, обеспечивая единую концепцию, на основе которой можно классифицировать и сравнивать компоненты инфраструктуры. Модель ENVRI RM также используется для выявления типовых решений общих проблем. Темпы производства данных в инфраструктурах варьируются от нескольких гигабайт до нескольких петабайт в год.

Планы на будущее

Общая среда ENVRI откроет новые возможности для пользователей взаимодействующих инфраструктур экологических исследований и обеспечит участникам междисциплинарных исследований возможность получать, изучать и сопоставлять данные из нескольких областей знаний в интересах исследований системного уровня. Сотрудничество влияет на требования к большим данным, образующиеся в результате междисциплинарных исследований.

ENVRI проанализировала вычислительные характеристики шести инфраструктур экологических исследований «Европейского стратегического форума по исследовательским инфраструктурам» (European Strategy Forum on Research Infrastructures, ESFRI) и выделила пять общих подсистем. Они описаны в эталонной модели ENVRI RM (см. https://confluence.egi.eu/display/EC/Download+of+ENVRI+Reference+Model) и перечислены ниже:

- подсистема сбора данных: собирает первичные данные от групп датчиков, различных приборов или наблюдателей-людей, направляет потоки данных измерений в систему;

- подсистема курирования данных: облегчает контроль качества и обеспечение долговременной сохранности научных данных и обычно размещается в центре обработки данных;

- подсистема доступ к данным: обеспечивает поиск и извлечение данных, размещенных в ресурсах данных, управляемых подсистемой курирования данных;

- подсистема обработки данных: объединяет данные из различных ресурсов и предоставляет вычислительные ресурсы и возможности для проведения анализа данных и научных экспериментов;

- подсистема поддержки сообщества: управляет, контролирует и отслеживает действия пользователей и поддерживает пользователей при выполнении ими их ролей в сообществе.

5.9.3 Вариант использования 43: Анализ радиолокационных данных для Центра дистанционного зондирования ледяного покрова (CReSIS)

Применение

Центр дистанционного зондирования ледяного покрова университета Канзаса, США (CReSIS) использует специализированные радиолокационные системы для измерения толщины слоя ледяного покрова и (ежегодно) толщины слоя снега на Северном и Южном полюсах и в горных районах.

Полученные данные передаются в Межправительственную группу экспертов по изменению климата (IPCC). Радарные системы, как правило, устанавливаются на самолетах, летающих по нескольким траекториям.

Текущий подход

Первоначальный анализ предусматривает использование инструментов обработки сигналов пакета Matlab, в результате которой выдается набор радиолокационных изображений. Эти изображения не могут быть переданы с места исследований через интернет, поэтому они, как правило, копируются на месте на несколько съемных жестких дисков терабайтного объема, а затем доставляются по воздуху в лабораторию для подробного анализа.

Элементы изображения (слои) выявляются с использованием инструментов понимания изображений при некотором контроле со стороны человека. Типичная эхограмма с выявленными границами позволяет различать границы между слоями воздуха и льда, между льдом и рельефом местности. Эта информация хранится в базе данных, доступ к которой осуществляется через географическую информационную систему. Данные о толщине слоя ледяного покрова используются при моделировании дви-

жения ледников. В ходе каждой полевой экспедиции, длящейся, как правило, несколько недель, производится от 50 до 100 терабайт данных.

Планы на будущее

Прогнозируется, что при использовании улучшенных инструментов объемы данных вырастут на порядок величины (до петабайта за экспедицию). Поскольку увеличивающиеся в объеме первичные данные должны обрабатываться в среде с ограниченным доступом к энергии, в качестве предпочтительных рассматриваются архитектуры с низким энергопотреблением или с низкой производительностью, такие как системы на основе графических процессоров.

5.9.4 Вариант использования 44: Обработка данных, доставка результатов и сервисы данных проекта «Радар с синтезированной апертурой для беспилотного летательного аппарата» (UAVSAR)

Применение

Радар с синтезированной апертурой (SAR) способен выявлять изменения ландшафта, вызванные сейсмической активностью, оползнями, обезлесением, изменениями растительности и наводнениями. Эта функциональная возможность может быть использована в интересах науки о землетрясениях, а также менеджмента стихийных бедствий. Данный вариант использования охватывает хранение данных, приложение для обработки изображений и визуализацию данных с географической привязкой.

Текущий подход

После передачи существенных объемов данные с самолетов и спутников перед сохранением обрабатываются на компьютерах Национального управления по аэронавтике и исследованию космического пространства США (NASA). Данные раскрываются для общественности после обработки и требуют значительного курирования из-за сбоев измерительного оборудования. Текущий объем данных составляет примерно 150 терабайт.

Планы на будущее

Размер данных резко увеличится в случае запуска программы НАСА спутникового радиолокационного зондирования Земли (Earth Radar Mission). Облачные системы хранения являются подходящими для хранения данных, однако в настоящее время не используются.

5.9.5 Вариант использования 45: Объединенный испытательный стенд iRODS Исследовательского центра в Ленгли НАСА и Центра управления полетами имени Годдарда

Применение

Центр моделирования климата НАСА (NCCS) и Центр обработки атмосферных данных (ASDC) Национального управления по аэронавтике и исследованию космического пространства США (NASA) имеют в своем распоряжении взаимодополняющие друг друга наборы данных огромного объема, ввиду чего по этим данным трудно выполнять запросы и ими сложно обмениваться.

Исследователям климата, специалистам по прогнозированию погоды, группам разработки и обслуживания измерительной аппаратуры и другим специалистам нужен доступ к данным из нескольких массивов данных с тем, чтобы сравнивать показания датчиков различных измерительных инструментов, сопоставлять показания датчиков с результатами моделирования, калибровать приборы, искать корреляции между несколькими параметрами и т. д.

Текущий подход

Данные создаются на основе двух продуктов: «Система для ретроспективного анализа современной эры для исследований и приложений» (MERRA), описывается отдельно в варианте применения № 46, и проекта НАСА «Система для изучения облачности и излучения Земли» (CERES):

- база данных «Баланс и накопление энергии верхних слоев атмосферы» EBAF-TOA (Energy Balanced and Filled-Top of Atmosphere) объемом около 420 мегабайт;
- продукт «Баланс и накопление энергии — Поверхность» EBAF-Surface, объемом около 690 мегабайт.

Количество данных увеличивается с каждым обновлением версии, которое происходит примерно раз в полгода. В настоящее время усилия по анализу, визуализации и обработке данных из неоднородных массивов данных требуют много времени. Ученым приходится отдельно получать доступ, искать и загружать данные с каждого из нескольких серверов. Данные часто дублируются, при этом непонятно, какой источник считать авторитетным. Нередко получение доступа к данным отнимает больше времени, чем научный анализ. Текущие массивы данных размещаются на кластерах InfiniBand умеренного размера (от 144 до 576 ядер).

Планы на будущее

Улучшенный доступ будет обеспечиваться благодаря использованию интегрированной системы управления данными, основанной на использовании правил» (iRODS). Эти системы поддерживают параллельную загрузку массивов данных с выбранных серверов копий (replica servers), обеспечивая пользователям всемирный доступ к географически рассредоточенным серверам. Работе iRODS будут способствовать семантически организованные метаданные, управление которыми осуществляется на основе высокоточной онтологии НАСА для наук о Земле. Также будет рассмотрен вопрос о возможности использования облачных решений.

5.9.6 Вариант использования 46: Аналитические сервисы MERRA (MERRA/AS)

Применение

Данное приложение «Система для ретроспективного анализа современной эры для исследований и приложений» (MERRA) осуществляет глобальный, согласованный во времени и пространстве синтез значений 26 ключевых климатических параметров путем объединения результатов численного моделирования с данными наблюдений.

Пространственные результаты выдаются каждые шесть часов начиная с 1979 г. и по настоящее время. Эти данные поддерживают такие важные приложения, как исследования Межправительственной группы экспертов по изменению климата (IPCC) и системы поддержки принятия решений по восстановлению экосистем (RECOVER) и борьбы НАСА и Министерства внутренних дел США с природными пожарами. В этих приложениях данные MERRA обычно интегрируются с данными из других массивов данных.

Текущий подход

Для обработки текущего объема данных в 480 терабайт используется Map/Reduce. Существующая система размещена в кластере InfiniBand с 36 узлами.

Планы на будущее

Изучается вопрос об использовании облачных вычислений. Прирост объема данных составляет один терабайт в месяц.

5.9.7 Вариант использования 47: Атмосферная турбулентность — Обнаружение событий и прогностическая аналитика

Применение

Интеллектуальный анализ данных на основе продуктов ретроспективного анализа, таких как массивы данных проектов «Система для ретроспективного анализа современной эры для исследований и приложений» (MERRA), который описывается отдельно в варианте использования № 46, и «Реанализ метеорологических данных для региона Северной Америки» (NARR), который представляет собой набор климатических данных высокого разрешения за длительный период времени для Северной Америки.

В ходе анализа сопоставляются сведения о турбулентности, полученные от летательных аппаратов (либо из отчетов пилотов, либо из автоматических измерений на летательных аппаратах скорости диссипации вихрей), с данными недавно завершенного ретроспективного анализа.

Получаемая информация представляет ценность для авиационной промышленности и специалистов по прогнозу погоды. В настоящее время стандартов для продуктов ретроспективного анализа нет, что приводит к усложнению систем, для которых изучаются возможности использования инструмента Map/Reduce. Объем медленно обновляемых данных реанализа составляет сотни терабайт, в то время как набор данных турбулентности меньше по размеру и реализован как потоковый сервис.

Текущий подход

Текущий массив данных объемом 200 терабайт может быть проанализирован с помощью Map/Reduce или аналогичного инструмента с использованием SciDB или иной научной СУБД.

Планы на будущее

Через пять лет объем массива данных достигнет 500 терабайт. Исходная тематика турбулентности может быть расширена за счет других океанических/атмосферных явлений, однако аналитика в каждом случае будет отличаться.

5.9.8 Вариант использования 48: Исследования климата с использованием модели климатической системы Земли (CESM) в Национальном научно-исследовательском вычислительном центре энергетических исследований Министерства энергетики США (NERSC)

Применение

Моделирование с использованием модели климатической системы Земли (CESM) может быть использовано для понимания и количественного определения вклада естественных и антропогенно-об-

условленных типовых сценариев изменчивости и изменения климата в 20-м и 21-м столетиях. Результаты проводимого по всему миру суперкомпьютерного моделирования должны сохраняться и сравниваться.

Текущий подход

Грид-система обработки данных о Земле (ESG) обеспечивает глобальный доступ к климатическим данным в огромных масштабах — в пета или даже в экза-масштабе, храня многие петабайты данных в десятках центрах по всему миру, объединенных в грид. Инфраструктура ESG считается ведущей в плане управления и обеспечения доступа к большим распределенным объемам данных, используемых в исследованиях в области изменения климата. Она поддерживает «Проект сопоставления связанных климатических моделей» (CMIP), протоколы которого обеспечиваются периодическими оценками, выполняемыми «Межправительственной группой экспертов по изменению климата» (IPCC).

Планы на будущее

Ожидается быстрый рост объемов данных: в 2017 г. только в Национальном научно-исследовательском вычислительном центре энергетических исследований Министерства энергетики США (NERSC) будет произведено 30 петабайт данных (при условии выполнения 15 сквозных экспериментов по теме изменения климата) и во много раз больше в мире в целом.

5.9.9 Вариант использования 49: Фокус-область подповерхностных биогеохимических исследований Управления биологических и экологических исследований Министерства энергетики США (BER)

Применение

Обеспечиваемые проектом моделирования водоразделов с использованием генома (Genome — Enabled Watershed Simulation Capability, GEWaSC) возможности необходимы для создания прогнозирующей структуры для понимания следующего:

- как геномная информация, хранящаяся в подповерхностном микробиоме, влияет на функционирование биогеохимического водораздела;
- как процессы в масштабе водораздела влияют на функционирование микробов;
- как эти взаимодействия сосуществуют.

Текущий подход

Текущие средства моделирования позволяют воспроизводить процессы, происходящие во внушительном диапазоне масштабов — от отдельной бактериальной клетки до шлейфа загрязнения. Данные охватывают все масштабы от геномики микробов в почве до гидробиогеохимии водораздела. Данные производятся различными областями исследований и включают данные моделирования, полевых измерений (например, гидрологические, геохимические, геофизические), биологических наук — «омиков» и наблюдений в ходе лабораторных экспериментов.

Планы на будущее

До сегодняшнего дня недостаточно внимания уделялось разработке концепции для систематического соединения явлений различных масштабов, что необходимо для выявления ключевых элементов контроля и управления и моделирования существенных обратных связей. В рамках проекта GEWaSC будет разработана концепция моделирования, которая охватит широкий диапазон данных — от геномов до водоразделов. Она позволит объединять разнообразные и разрозненные массивы данных полевых, лабораторных измерений и моделирования, включая различные семантические, пространственные и временные измерения.

5.9.10 Вариант использования 50: Сеть AmeriFlux Управления биологических и экологических исследований Министерства энергетики США и сеть FLUXNET

Применение

Сети AmeriFlux и FLUXNET представляют собой, соответственно, американскую и мировую коллекции датчиков, которые отслеживают потоки малых газовых составляющих (таких как CO_2, водяной пар) в широком временном (часы, дни, времена года, годы и десятилетия) и пространственном диапазонах. Кроме того, формируемые наборы данных содержат информацию о важнейших взаимосвязях между организмами, экосистемами и исследованиями на уровне процессов — в адекватных для изучения климата масштабах ландшафтов, регионов и континентов, которые следует учитывать в биогеохимических и климатических моделях.

Текущий подход

Сведения о программном обеспечении приведены в А.8.10. Имеется около 150 измерительных вышек в составе сети AmeriFlux и более 500 распределенных по всему миру вышек для сбора измерений газовых потоков.

Планы на будущее

Сбор данных полевых экспериментов будет улучшен благодаря доступу к существующим данным и автоматическому вводу новых данных через мобильные устройства. Будут расширены междисциплинарные исследования, объединяющие различные источники данных.

5.10 Энергетика

5.10.1 Вариант использования 51: Прогнозирование потребления электроэнергии в интеллектуальных энергосетях

Применение

«Умные» счетчики поддерживают прогнозирование потребления энергии для потребителей, трансформаторов, подстанций и зон обслуживания электросетей. Передовые счетчики выдают показания каждые 15 минут, обеспечивая детализацию на уровне отдельных потребителей в зоне обслуживания интеллектуальных электросетей.

В состав объединяемых данных входят данные умных счетчиков (распределенные), служебные базы данных энергетических компаний (информация о клиентах, топология сети — централизованные), данные всеобщей переписи населения США (распределенные), метеорологические данные Национального управления океанических и атмосферных исследований США (National Oceanic and Atmospheric Administration, NOAA) (распределенные), данные информационных систем для построения микроэнергосетей (централизованные) и сенсорных сетей микроэнергосетей (распределенные). Центральной темой является выполняемый в реальном времени, управляемый данными анализ временных рядов из киберфизических систем.

Текущий подход

Прогнозирование использует визуализацию на основе геоинформационных систем (ГИС). Темпы производства данных составляют около 4 терабайт в год для такого города, как Лос-Анджелес, где имеется 1,4 млн датчиков. Существуют серьезные проблемы в плане обеспечения защиты персональных данных, требующие анонимизации путем агрегирования данных. Данные в реальном времени и исторические данные в сочетании с машинным обучением используются для прогнозирования потребления. Информация о программном обеспечении приведена в А.9.1.

Планы на будущее

Будут широко развернуты передовые технологии энергосетей. В интеллектуальных сетях появятся новые инструменты аналитики, объединяющие разнородные данные и поддерживающие выдачу требований к крупным потребителям о сокращении энергопотребления в пиковые периоды (curtailment request). Новые технологии будут поддерживать мобильные приложения для взаимодействия с клиентами.

5.10.2 Вариант использования 52: Система управления энергией домашнего хозяйства HEMS

Применение

Система управления энергией домашнего хозяйства (HEMS) является полезной для энергосбережения в частных домах. В рамках системы HEMS в частных домах устанавливается различного вида датчики и устройства, такие как «умный» счетчик, электромобиль, панель солнечных батарей, осветительные приборы, кондиционер, топливный элемент, водонагреватель, аккумуляторная батарея. «Менеджер энергопотребления» собирает произведенные в частных домах данные и сохраняет их в облачной базе данных, называемой «большой информационной платформой HEMS». «Информационный менеджер» управляет большой информационной платформой HEMS и осуществляет менеджмент данных. Он также отвечает за обеспечение неприкосновенности частной жизни и безопасность пользователей. «Сервисный агент» анализирует данные и предоставляет пользователям ценную информацию в качестве услуги.

Текущий подход

Услуги, предоставляемые «сервисным агентом», не ограничиваются мониторингом энергопотребления. Другими примерами полезных услуг являются услуги по наблюдению за состоянием пожилых людей, помощь с выбором оптимального тарифного плана для электроэнергии, прогнозирование выработки электроэнергии фотоэлектрической системой, управление спросом на электроэнергию посредством стимулирования купонами (coupon incentive-based demand response, CIDR).

Планы на будущее

Для повышения полезности данных HEMS необходима будет стандартизация API-интерфейса программирования приложений.

6 Технические проблемы, выявленные в результате анализа вариантов использования

Технические проблемы — это проблемы и препятствия, ограничивающие дальнейшее использование больших данных. После сбора, обработки и анализа вариантов использования из отдельных описаний были выделены упомянутые в них технические проблемы и сгруппированы на основе семи характерных признаков. Затем эти специфические проблемы были обобщены с целью выделения, в рамках семи характерных категорий, высокоуровневых требований, которые не зависят от производителя и от технологии. При этом следует отметить, что ни списки вариантов использования, ни списки требований не являются исчерпывающими.

6.1 Технические проблемы в конкретных вариантах использования

Каждый вариант использования был оценен на предмет наличия технических проблем по семи критериям, определяемым следующими ключевыми факторами:

- **источник данных** [например, объемы данных, файловые форматы, темпы увеличения объемов, нахождение данных в покое (неактивные данные) или движении (данные в процессе передачи либо обработки)];
- **преобразование данных** (например, объединение данных, аналитика);
- **возможности обработки** (например, программные инструменты, инструменты платформ, аппаратные ресурсы, такие как ресурсы хранения и сетевые);
- **потребитель данных** (например, представление обработанных результатов в текстовом, табличном, визуальном и иных форматах);
- **безопасность и защита персональных данных**;
- **управление жизненным циклом** (например, курирование, конверсия (конвертация), проверка качества, предварительная обработка перед проведением анализа);
- **иные технические проблемы**.

В описаниях некоторых вариантов использования присутствовали все технические проблемы, в то время как в описаниях других вариантов назывались лишь несколько проблем. Полный список специфических проблем, извлеченных из описаний вариантов использования, приведен в приложении D. Данные признаки были приняты во внимание при окончательном отборе ролей, описанных в ИСО/МЭК 20547-3.

6.2 Сводные итоги анализа требований

Были выделены 35 общих требований [1] на основе анализа и обобщения 439 специфических технических проблем, извлеченных из 52 вариантов использования. В графе 2 таблицы 1 указано количество специфических технических проблем, послуживших основой для выделения соответствующего общего требования.

Таблица 1 — Общие технические требования, сформулированные на основе специфических технических проблем

#	Количество вариантов	Требование
Требования к поставщику данных		
1	26[1)]	Необходимо поддерживать надежную, в реальном времени и/или асинхронную, потоковую и/или пакетную обработку с целью сбора данных из централизованных, распределенных и/или облачных источников, от датчиков и/или приборов
2	22	Необходимо поддерживать передачу данных — медленную и/или неравномерную с периодическими пиковыми нагрузками и/или с высокой пропускной способностью — между источниками данных и вычислительными кластерами
3	28	Необходимо поддерживать данные разнообразных типов и видов, включая структурированные и неструктурированные тексты, документы, графы, веб-материалы, геопространственные данные, сжатые, с привязкой ко времени, пространственные, мультимедийные данные, данные моделирования и показания измерительных инструментов
Требования к провайдеру сервиса преобразования данных		
1	36[1)]	Необходимо поддерживать разнообразные вычислительно-интенсивные методы аналитической обработки и методы машинного обучения
2	7	Необходимо поддерживать аналитическую обработку в реальном времени и/или пакетную
3	14[1)]	Необходимо поддерживать обработку большого объема разнородных данных и данных моделирования
4	6	Необходимо поддерживать обработку данных в движении (потоковая передача, доставка нового контента, отслеживание и т. д.)
Требования к провайдеру вычислительных возможностей		
1	27[1)]	Необходимо поддерживать как унаследованные, так и продвинутые пакеты программ (ПО)
2	16[1)]	Необходимо поддерживать как унаследованные, так и продвинутые вычислительные платформы (платформа)
3	23[1)]	Необходимо поддерживать как унаследованные, так и продвинутые распределенные вычислительные кластеры, сопроцессоры, обработку ввода-вывода (инфраструктура)
4	14	Необходимо поддерживать гибкую передачу данных (сети)
5	28[1)]	Необходимо поддерживать унаследованные, крупномасштабные и продвинутые распределенные хранилища данных (хранение)
6	13	Необходимо поддерживать как унаследованные, так и продвинутые исполняемые программы: приложения, инструменты, утилиты и библиотеки (ПО)
Требования к потребителю данных		
1	4	Необходимо поддерживать быстрый поиск по обработанным данным — с высокой релевантностью, точностью и полнотой результатов поиска
2	13[1)]	Необходимо поддерживать различные форматы выходных файлов для визуализации, рендеринга и создания отчетов
3	2	Необходимо поддерживать визуальную разметку для представления результатов
4	9[1)]	Необходимо поддерживать пользовательский интерфейс с широкими функциональными возможностями для доступа с помощью браузера и средства визуализации

[1)] *Исправлена неверная цифра, здесь и в приложении D.*

Окончание таблицы 1

#	Количество вариантов	Требование
5	20	Необходимо поддерживать инструменты многомерной, с высоким разрешением визуализации данных
6	1	Необходимо поддерживать потоковую передачу результатов клиентам
Требования по обеспечению безопасности и защиты персональных данных		
1	30[1)]	Необходимо обеспечить безопасность и конфиденциальность чувствительных данных
2	12	Необходимо поддерживать изолированную среду («песочницу»), обеспечивать контроль доступа и многоуровневую аутентификацию на основе политик в отношении подлежащих защите данных
Требования к управлению жизненным циклом		
1	20	Необходимо поддерживать курирование качества данных, включая предварительную обработку, кластеризацию, классификацию, редуцирование (преобразование к физическим величинам) и преобразование форматов
2	2	Необходимо поддерживать динамическое обновление данных, профилей пользователей и ссылок
3	6	Необходимо поддерживать жизненный цикл данных и политику обеспечения долговременной сохранности, включая отслеживание происхождения данных
4	4	Необходимо поддерживать валидацию данных
5	4	Необходимо поддерживать аннотирование данных человеком для их валидации
6	3	Необходимо принимать меры для предотвращения утраты или порчи данных
7	1	Необходимо поддерживать географически распределенные (multi-site) архивы
8	2	Необходимо поддерживать постоянные идентификаторы и прослеживаемость данных
9	1	Необходимо поддерживать стандартизацию, агрегирование и нормализацию данных из разнородных источников
Иные требования		
1	6	Необходимо поддерживать пользовательский интерфейс с широкими возможностями для мобильных платформ с целью обеспечения доступа к обработанным результатам
2	2	Необходимо поддерживать мониторинг с использованием мобильных платформ и учетом производительности аналитической обработки
3	13	Необходимо поддерживать визуальный поиск по контенту с широкими функциональными возможностями и отображение контента на мобильных платформах
4	1	Необходимо поддерживать сбор данных с использованием мобильных устройств
5	1	Необходимо обеспечивать безопасность на мобильных устройствах

[1)] *Исправлена неверная цифра, здесь и в приложении D.*

6.3 Признаки вариантов использования

В таблице 2 указано количество вариантов использования, обладавших определенными признаками. Выбор этих признаков был сделан на основе анализа, описанного в публикациях [2], [3] и [4].

Таблица 2 — Признаки вариантов использования

Аббревиатура	#	Описание
PP	26	Хорошо распараллеливаемая задача или задача Map-Only в парадигме Map/Reduce
MR	18	Классический Map/Reduce (добавьте данные по MRStat ниже для полного подсчета)
MRStat	7	Простая версия Map/Reduce, в которой ключевые вычисления представляют собой простое редуцирование, подобное вычислению статистических средних величин, таких как гистограммы и средние значения
MRIter	23	Итеративный Map/Reduce
Graph	9	Для анализа необходима сложная структура данных в виде графа
Fusion	11	Интеграция разнообразных данных в интересах выявления/принятия решений; может включать сложные алгоритмы или быть просто порталом
Streaming	41	Некоторые данные поступают порциями и таким же образом обрабатываются
Classify	30	Классификация: разделение данных по категориям
S/Q	12	Индексирование, поиск и выполнение запросов
CF	4	Использование совместной фильтрации рекомендательной системой
LML	36	Локальное машинное обучение (независимое для каждой параллельной сущности)
GML	23	Глобальное машинное обучение: глубокое обучение, кластеризация, LDA, PLSI, MDS, оптимизация большой размерности, как в вариационном байесовском методе, MCMC, алгоритм с распространением доверия «с подъемом» (Lifted Belief Propagation), стохастический градиентный спуск, L-BFGS, алгоритм Левенберга-Марквардта. Может вызывать алгоритм эффективной глобальной оптимизации (Efficient Global Optimization, EGO) или оптимизация сверхбольшой размерности (Exascale Global Optimization) вместе с масштабируемым параллельным алгоритмом.
	51	Управление потоками рабочих процессов — универсальное свойство, поэтому без идентификатора
GIS	16	Данные с геопривязкой часто отображаются с использованием ESRI, Microsoft Virtual Earth, Google Earth, GeoServer и т. д.
HPC	5	Классическое крупномасштабное моделирование космоса, материалов и т. д., производящее данные (например, для визуализации)
Agent	2	Моделирование с использованием моделей управляемыми данными макрообъектов, представленных в виде агентов

С учетом этого дополнительного анализа данная таблица была расширена [3]. В итоге были выделены 50 свойств, сгруппированных в четыре представления, приведенные в таблицах 3—6.

Таблица 3 — Фасеты ракурса «архитектуры проблемы» концепции Ogres (мета/макрошаблон)

Pleasingly Parallel, PP (хорошее распараллеливание)	Можно найти в BLAST, в моделировании белково-белковых взаимодействий (белковом докинге), в некоторых вариантах обработки (био) изображений, включая локальную аналитику или локальное машинное обучение с хорошо распараллеливаемой фильтрацией
Classic Map/Reduce, MR (классический Map/Reduce)	Алгоритмы индексирования, поиска, выполнения запросов и классификации, такие как совместная фильтрация («вычислительные задачи-гиганты»: G1 для MRStat в таблице 2, G7)
Map Collective	Встречается в машинном обучении — особенно в случае ядра на основе линейной алгебры
Map P2P	Прямая связь между узлами (Point to Point Communication), наблюдаемая в параллельном моделировании и графовых алгоритмах
Map Streaming (архитектура работы с потоковыми данными)	Комбинация (параллельных) длительно выполняемых процессов отображения (картирования — maps), принимающих потоковые данные
Shared Memory	Общая память — в отличие от распределенных данных (памяти). Используется в задачах, где важна реализация совместно используемой памяти. Имеет тенденцию быть динамически асинхронной
SPMD	Хорошо известный метод распараллеливания «Единая программа, множество данных» (Single Program Multiple Data)
BSP	Массовая синхронная обработка (Bulk Synchronous Processing, также расшифровывается как Bulk Synchronous Parallel model — массовая синхронная параллельная модель): четко определенные этапы вычислений/обмена информацией
Fusion (объединение)	Процесс выявления знаний часто включает в себя объединение ряда методов или источников данных
Dataflow (потоки данных)	Составная структура, в рамках которой ряд компонентов связан друг с другом посредством обмена данными
Agents (агенты)	Используется в эпидемиологии, при моделировании дискретных событий и т. д. «Роевые» подходы
Workflow (потоки рабочих процессов)	Во многих приложениях часто используется «аранжировка» (orchestration) / управление потоками рабочих процессов многих компонентов

Таблица 4 — Фасеты ракурса «Особенности исполнения» концепции Ogres

Метрики производительности (эффективности)	Измеряются в рамках сопоставительного анализ на основе эталонных показателей
Отношение флоп/байт	Важно для производительности
Среда исполнения	Облако или среда высокопроизводительных вычислений; нужны ли базовые библиотеки, такие как библиотеки матричной / векторной алгебры, метода сопряженного градиента, редукции, трансляции и т. д.? (Задача «гигант» G4)
Объем	Обширность данных, доступных для анализа с целью извлечения ценной информации
Скорость обработки	Скорость потока, с которой данные создаются, передаются, хранятся, анализируются или визуализируются
Разнообразие	Разнородность массива данных, полученных из нескольких предметных областей и/или объединяющих несколько их типов. См. также фасет «объединение» (fusion)

Окончание таблицы 4

Достоверность	Полнота и точность данных, влияющие на процесс необходимой предварительной обработки и надежность результатов
Структура информационного обмена	Какова структура соединений? Является ли информационный обмен синхронным или асинхронным? В последнем случае может оказаться привлекательным использование общей памяти
Статическое или динамическое?	Изменяется ли приложение (граф) во время исполнения?
Регулярность	Большинство приложений состоит из набора взаимосвязанных объектов; является ли этот набор регулярным, как набор пикселей, или же представляет собой сложный нерегулярный граф?
Алгоритм итеративный или нет?	Важная характеристика алгоритма
Абстрактная модель данных	Пары «ключ-значение», пиксели, графы, вектора, файлы формата HDF5, «мешок слов» и т. д.
Является ли пространство данных метрическим?	Находятся ли точки данных в метрическом или неметрическом пространстве? (Задача «гигант» G2)
Сложность	Является ли сложность алгоритма порядка $O(N^2)$ или $O(N)$ включая $\log(N)$, для N элементов, обрабатываемых за итерацию? (Задача «гигант» G2)

Таблица 5 — Фасеты ракурса «Источник данных и стиль обработки данных» концепции Ogres

SQL, NoSQL или NewSQL	NoSQL включают в себя хранилища документов, столбцы, пары «ключ-значение», графы, Triplestore (хранилище триплетов, или RDF-хранилище)
Корпоративные системы управления данными	В 10 вариантах использования из публикации NIST [1] интегрируются SQL/NoSQL-решения
Файлы и объекты	Файлы в том виде, в каком они управляются в iRODS, чрезвычайно распространены в научных исследованиях. Объекты наиболее часто встречаются в стеке программного обеспечения для обработки больших данных Apache Big Data Stack (ABDS)
HDFS / Luster / GPFS	Располагаются ли данные и вычисления в одном месте?
Архивация / пакетная обработка / потоковая обработка	Потоковая обработка представляет собой процесс постепенного обновления наборов данных, при этом внедряются новые алгоритмы для достижения отклика в реальном времени (Задача «гигант» G7)
Виды систем хранения	Виды включают «коллективное использование» (shared), «выделение» (dedicated), «постоянное хранение» (permanent) и «временное хранение» (transient)
Метаданные / Происхождение данных	Описывают общие характеристики данных, историю и особенности их обработки
Интернет вещей	К 2020 г. Интернет вещей будет охватывать от 24 (см. [6][1)]) до 50 млрд устройств (см. [7], [8])
Данные, создаваемые в ходе высокопроизводительных вычислений	В результате математического моделирования генерируется визуализация, для формирования которой часто требуется проводить интеллектуальный анализ данных моделирования
Геоинформационные системы (ГИС)	Географические информационные системы обеспечивают доступ к геопространственным данным

[1)] *Исправлена ошибочная ссылка.*

Таблица 6 — Фасеты ракурса «Обработка / реальное время» концепции Ogres

Микро-рейтинги (micro benchmarks)	Простое ядро или мини-приложение, используемое для измерения производительности базовой системы
LML	Локальная аналитика или локальное машинное обучение
GML	Глобальная аналитика или машинное обучение, требующее итеративной среды выполнения (задачи «гиганты» G5, G6)
Базовая статистика	Простая статистика, представленная в таблице 2 как MRStat
Рекомендации	Совместное фильтрование и другие аналитические методы, используемые в рекомендательных системах
Индексирование, поиск и выполнение запросов	Богатый набор технологий, используемых для индексирования данных, поиска и выполнения запросов к данным
Классификация	Технологии для маркировки/тегирования данных (SVM, Байес, глубокое обучение, кластеризация)
Обучение	Обучение алгоритмов
Методы оптимизации	Машинное обучение, нелинейная оптимизация, метод наименьших квадратов, линейное / квадратичное программирование, комбинаторная оптимизация, EM-алгоритм, метод Монте-Карло, вариационный байесовский анализ, глобальный вывод
Потоковая обработка	Расширяющийся класс быстрых онлайн-алгоритмов сложности O(N)
Согласование (alignment)	Вариант поиска, используемый при сопоставлении последовательностей (как, например, в BLAST)
Линейная алгебра	Многие алгоритмы машинного обучения основаны на ядрах вычислений линейной алгебры
Граф	Задача представлена в виде графа, а не вектора, сетки и т.д. (задача «гигант» G3)
Визуализация	Важный компонент многих конвейеров аналитической обработки

Приложение А
(справочное)

Представленные описания вариантов использования

А.1 Деятельность государственных органов

А.1.1 Вариант использования № 1: Архивное хранение больших данных переписи населения, проведенной в США в 2010 и 2000 гг.

Название	Архивное хранение больших данных: Большие данные переписи населения, проведенной в США в 2010 и 2000 гг. на основании части 13 Свода законов США	
Предметная область	Электронные архивы	
Автор/организация/эл.почта	Вивек Наваль (Vivek Navale) и Куин Нгуен (Quyen Nguyen), Национальные архивы США (NARA)	
Акторы/заинтересованные лица, их роли и ответственность	Архивисты Национальных архивов США, представители общественности (после 75 лет)	
Цели	Обеспечить долговременную сохранность данных с целью предоставления к ним доступа и проведения аналитики по истечении 75-летнего ограничительного периода. Часть 13 Свода законов США уполномочивает Бюро переписи населения США (U.S. Census Bureau) собирать и сохранять данные, относящиеся к переписи, и гарантирует защиту персональных и отраслевых данных	
Описание варианта использования	В течение ограничительного периода в 75 лет данные должны храниться «как есть», без возможности доступа и анализа, с обеспечением сохранности на уровне битов. Данные курируются, что может включать преобразование формата. Доступ и аналитика должны быть обеспечены через 75 лет	
Текущие решения	**Вычислительная система**	Сервера под ОС Linux
	Хранилище данных	Облачные сервисы NetApp, магнитные ленты
	Сеть связи	
	Программное обеспечение	
Характеристики больших данных	**Источник данных (распределенный/централизованный)**	Централизованное хранение
	Объем (количество)	380 терабайт
	Скорость обработки (например, в реальном времени)	Данные статичны
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Отсканированные документы
	Вариативность (темпы изменения)	Нет

Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Утрата данных недопустима
	Визуализация	Будет определена в будущем
	Качество данных (синтаксис)	Неизвестно
	Типы данных	Отсканированные документы
	Аналитика данных	Только по истечении 75 лет
Иные проблемы больших данных	Обеспечение долговременной сохранности данных	
Проблемы пользовательского интерфейса и мобильного доступа	Будут определены в будущем	
Технические проблемы обеспечения безопасности и защиты персональных данных	Подпадают под положения части 13 Свода законов США	
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)		
Дополнительная информация (гиперссылки)		

А.1.2 Вариант использования № 2: Прием Национальными архивами США государственных данных на хранение, поиск, извлечение и обеспечение долговременной сохранности

Название	Прием Национальными архивами США (NARA) государственных данных на хранение, поиск, извлечение и обеспечение долговременной сохранности
Предметная область	Электронные архивы
Автор/организация/эл.почта	Куин Нгуен (Quyen Nguyen) и Вивек Наваль (Vivek Navale), Национальные архивы США (NARA)
Акторы/заинтересованные лица, их роли и ответственность	Специалисты по управлению документами федеральных органов исполнительной власти США. Специалисты по комплектованию фондов Национальных архивов США. Архивисты Национальных архивов США. Пользователи архивов — представители общественности
Цели	Прием на архивное хранение, поиск, извлечение и обеспечение долговременной сохранности больших данных
Описание варианта использования	1) Передача данных под физический контроль Национальных архивов и переход к Национальным архивам юридической ответственности за их сохранность. В будущем, если данные хранятся в облаке, при передаче Национальным архивам ответственности за физическую сохранность желательно избегать перемещения больших данных из одного облака в другое либо из облака в центр обработки данных. 2) Предварительная обработка данных, включающая проверки на наличие вирусов, определение файловых форматов и удаления пустых файлов. 3) Индексирование данных. 4) Категоризация документов (чувствительные конфиденциальные, неконфиденциальные, персональные данные и т. д.). 5) Преобразование устаревших файловых форматов в современные (например, WordPerfect в PDF). 6) Электронное раскрытие. 7) Поиск и извлечение данных в рамках исполнения специальных запросов. 8) Поиск и извлечение государственных документов представителями общественности

Текущие решения	**Вычислительная система**	Сервера под ОС Linux
	Хранилище данных	Облачные сервисы NetApp, система хранения Hitachi, магнитные ленты
	Сеть связи	
	Программное обеспечение	Кастомизированное ПО, коммерческие поисковые продукты, коммерческие базы данных
Характеристики больших данных	**Источник данных (распределенный/централизованный)**	Распределенные источники данных федеральных органов исполнительной власти США. Используемый в настоящее время подход требует передачи этих данных в централизованное хранилище. В будущем эти источники данных могут находиться в различных облачных средах
	Объем (количество)	Сотни терабайт, постоянно увеличивается
	Скорость обработки (например, в реальном времени)	Скорость поступления данных относительно низкая по сравнению с другими вариантами использования, однако случаются всплески, т. е. данные могут поступать партиями размером от гигабайта до сотен терабайт
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Разнообразные типы данных, неструктурированные и структурированные: текстовые документы, электронная почта, фотографии, отсканированные документы, мультимедийные материалы, материалы из социальных сетей, веб-сайты, базы данных и т. д. Разнообразие прикладных областей, поскольку документы поступают от различных государственных органов. Данные поступают из различных хранилищ, некоторые из которых в будущем могут стать облачными
	Вариативность (темпы изменения)	Темпы могут варьироваться, особенно если источники данных неоднородны: в некоторых больше представлены аудио- и видеоматериалы, в других преобладают текстовые материалы, в-третьих — графические образы и т. д.
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Результаты поиска должны иметь высокую релевантность и полноту поиска. Требуется высокая точность категоризации документов
	Визуализация	Будет определена в будущем
	Качество данных (синтаксис)	Неизвестно
	Типы данных	Разнообразные типы данных: текстовые документы, электронная почта, фотографии, отсканированные документы, мультимедийные материалы, базы данных и т. д.
	Аналитика данных	Сканирование/индексирование; поиск; ранжирование; прогностический поиск. Категоризация данных (чувствительные, конфиденциальные и т. д.).Выявление и маркировка персональных данных (Personally Identifiable Information, PII)

Иные проблемы больших данных	Выполнение предварительной обработки и дальнейшее долговременное управление объемными и разнообразными данными. Проведение поиска по огромному объему данных. Обеспечение высокой релевантности и полноты результатов поиска. В будущем источники данных могут быть распределены по различным облакам
Проблемы пользовательского интерфейса и мобильного доступа	Мобильный поиск должен иметь похожий интерфейс и выдавать похожие результаты
Технические проблемы обеспечения безопасности и защиты персональных данных	Необходимо внимательно относиться к имеющимся ограничениям на доступ к данным
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	
Дополнительная информация (гиперссылки)	

А.1.3 Вариант использования № 3: Повышение активности респондентов в статистических обследованиях

Название	Повышение активности респондентов в статистических обследованиях (адаптивная схема)
Предметная область	Логистическое обеспечение сбора государственной статистики
Автор/организация/эл.почта	Каван Каппс (Cavan Capps), Бюро переписей населения США (U.S. Census Bureau), cavan.paul.capps@census.gov
Акторы/заинтересованные лица, их роли и ответственность	Задача органов государственной статистики США — быть ведущими авторитетными источниками информации о населении и экономике страны, уважая при этом неприкосновенность персональных данных и строго защищая их конфиденциальность. Эту задачу они решают, взаимодействуя со штатами, местными органами власти и другими федеральными органами исполнительной власти
Цели	Используя открытые и научно объективные передовые методы, органы статистики стремятся повысить качество, конкретность и своевременность выдаваемых статистических данных при одновременном снижении эксплуатационных расходов и обеспечении конфиденциальности респондентов
Описание варианта использования	Затраты на проведение статистических обследований растут, в то время как активность респондентов падает. Целью данной работы является применение усовершенствованных «методов рекомендательных систем», использующих комбинацию данных из нескольких источников, а также вспомогательные данные исторических обследований, — для поддержки процессов оперативной деятельности, направленных на повышение качества и снижение расходов проводимых «на местах» статистических обследований

Текущие решения	**Вычислительная система**	Системы под ОС Linux
	Хранилище данных	В SAN-сети систем хранения данных (Storage Area Network) и на непосредственно подключаемых к серверу устройствах (Direct Storage)
	Сеть связи	Оптоволоконный кабель, 10-гигабитный Ethernet, 40-гигабитный Infiniband
	Программное обеспечение	Hadoop, Spark, Hive, R, SAS, Mahout, Allegrograph, MySQL, Oracle, Storm, BigMemory, Cassandra, Pig
Характеристики больших данных	**Источник данных (распределенный/централизованный)**	Данные обследований, другие государственные административные данные, геопространственные данные из различных источников
	Объем (количество)	Для данного конкретного вида проблем оперативной деятельности, примерно один петабайт
	Скорость обработки (например, в реальном времени)	Варьируется, данных с мест о ходе проведения обследования передаются непрерывно в потоковом режиме. Во время последней всеобщей переписи населения в потоковом режиме были переданы 150 млн документов
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Данные обычно представляют собой заданные текстовые и числовые поля. Данные могут происходить из разных наборов данных, объединенных для достижения целей аналитики
	Вариативность (темпы изменения)	Варьируется в зависимости от обследований, проводимых на местах в данный момент. Высокие темпы поступления во время всеобщей переписи населения
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Данные должны иметь высокую степень достоверности, а системы должны быть очень надежными. Остается проблемой семантическая целостность концептуальных метаданных, содержащих описание объекта измерения и вытекающие из этого описания ограничений достоверности

Наука о больших данных (сбор, курирование, анализ, операции)	**Визуализация**	Визуализация полезна для проверки данных, оперативной деятельности и общего анализа. Продолжает развиваться
	Качество данных (синтаксис)	Качество данных должно быть высоким и статистически проверяться на точность и надежность на протяжении всего процесса сбора данных
	Типы данных	Предопределенные ASCII — строки и числовые данные
	Аналитика данных	Аналитика необходима для рекомендательных систем, постоянного мониторинга и для общего совершенствования процесса проведения обследования
Иные проблемы больших данных	Совершенствование рекомендательных систем, позволяющих снизить затраты и повысить качество, обеспечивая одновременно надежные и публично проверяемые меры защиты конфиденциальности	
Проблемы пользовательского интерфейса и мобильного доступа	Мобильный доступ важен	
Технические проблемы обеспечения безопасности и защиты персональных данных	Необходимо обеспечить безопасность и конфиденциальность всех данных. Согласно требованиям законодательства должна быть обеспечена возможность аудита всех процессов на предмет обеспечения безопасности и конфиденциальности	
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Рекомендательные системы имеют общие функциональные возможности с системами, используемыми в электронной коммерции такими фирмами, как Amazon, Netflix, UPS и др	
Дополнительная информация (гиперссылки)		

А.1.4 Вариант использования № 4: Использование нетрадиционных данных для повышения активности респондентов в статистических обследованиях

Название	Использование нетрадиционных данных для повышения активности респондентов в статистических обследованиях (адаптивная схема)
Предметная область	Логистическое обеспечение сбора государственной статистики
Автор/организация/эл.почта	Каван Каппс (Cavan Capps), Бюро переписи населения США (U.S. Census Bureau), cavan.paul.capps@census.gov
Акторы/заинтересованные лица, их роли и ответственность	Задача органов государственной статистики США — быть ведущими авторитетными источниками информации о населении и экономике страны, уважая при этом неприкосновенность персональных данных и строго защищая их конфиденциальность. Эту задачу они решают, взаимодействуя со штатами, местными органами власти и другими федеральными органами исполнительной власти

Цели	Используя открытые и научно объективные передовые методы, органы статистики стремятся повысить качество, конкретность и своевременность выдаваемых статистических данных при одновременном снижении эксплуатационных расходов и обеспечении конфиденциальности респондентов	
Описание варианта использования	Затраты на проведение статистических обследований растут, в то время как активность респондентов падает. В данной работе изучается потенциал использования нетрадиционных коммерческих и публичных источников данных из интернета, беспроводной связи и электронных транзакций, которые в рамках аналитических исследований объединяются с данными традиционных статистических обследований с целью повысить качество статистики для небольших регионов и новых показателей, а также обеспечить своевременность публикуемой статистики	
Текущие решения	**Вычислительная система**	Системы под ОС Linux
	Хранилище данных	В SAN-сети систем хранения данных (Storage Area Network) и на непосредственно подключаемых к серверу устройствах (Direct Storage)
	Сеть связи	Оптоволоконный кабель, 10 — гигабитный Ethernet, 40 — гигабитный Infiniband
	Программное обеспечение	Hadoop, Spark, Hive, R, SAS, Mahout, Allegrograph, MySQL, Oracle, Storm, BigMemory, Cassandra, Pig
Характеристики больших данных	**Источник данных (распределенный/централизованный)**	Данные обследований, другие государственные административные данные, данные из интернета, систем беспроводной связи, данные электронных транзакций, возможно, данные из социальных сетей, а также геопространственные данные из различных источников
	Объем (количество)	Будет определен в будущем
	Скорость обработки (например, в реальном времени)	Будет определена в будущем
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Текстовые данные, а также традиционным образом определенные текстовые строки и числовые поля. Данные могут происходить из нескольких наборов данных, объединенных для целей аналитики
	Вариативность (темпы изменения)	Будет определена в будущем

Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Данные должны иметь высокую степень достоверности, а системы должны быть очень надежными. Остается проблемой семантическая целостность концептуальных метаданных, описывающих, что именно измеряется, и вытекающие из этого пределы точности выводов
	Визуализация	Визуализация полезна для проверки данных, оперативной деятельности и общего анализа. Продолжает развиваться
	Качество данных (синтаксис)	Качество данных должно быть высоким и статистически проверяться на точность и надежность на протяжении всего процесса сбора данных
	Типы данных	Текстовые данные, предопределенные ASCII — строки и числовые данные
	Аналитика данных	Аналитика необходима для получения надежных оценок на основе совместного использования данных традиционных обследований, государственных административных данных и данных из нетрадиционных источников сферы цифровой экономики
Иные проблемы больших данных	Совершенствование систем аналитики и моделирования, выдающих надежные и устойчивые статистические оценки с использованием данных из ряда источников и являющихся научно прозрачными, которые одновременно обеспечивают надежные и публично проверяемые меры защиты конфиденциальности	
Проблемы пользовательского интерфейса и мобильного доступа	Мобильный доступ важен	
Технические проблемы обеспечения безопасности и защиты персональных данных	Необходимо обеспечить безопасность и конфиденциальность всех данных. Согласно требованиям законодательства должна быть обеспечена возможность аудита всех процессов на предмет обеспечения безопасности и конфиденциальности	
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Процесс получения статистических оценок, способный дать более детальные оценки в режиме почти реального времени и с меньшими затратами. Надежность статистических оценок, полученных на основе комбинирования данных из подобных смешанных источников, пока еще предстоит определить	
Дополнительная информация (гиперссылки)		

А.2 Коммерческая деятельность

А.2.1 Вариант использования № 5: Облачные вычисления в секторах финансовой отрасли

Название	Данный вариант использования представляет собой один из подходов к внедрению стратегии больших данных в рамках облачной экосистемы для секторов финансовой отрасли, осуществляющих операции в Соединенных Штатах
Предметная область	Включает следующие направления основной деловой деятельности: **Банковское дело**, в том числе: обслуживание юридических лиц, обслуживание физических лиц, кредитные карты, потребительское кредитование, обслуживание корпоративных клиентов, операционное обслуживание, финансирование торговых операций и глобальные платежи. **Ценные бумаги и инвестиции**, включая: брокерское обслуживание физических лиц, банковское обслуживание состоятельных физических лиц / управление частным капиталом, брокерское обслуживании институциональных инвесторов, инвестиционно-банковские услуги, трастовые банковские услуги, управление активами, депозитарные и клиринговые услуги **Страхование**, в том числе: персональное и групповое страхование жизни, персональное и групповое страхование имущества / несчастных случаев, фиксированный и переменный аннуитет и другие виды инвестиций. **Для сведения**: Любая государственная/частная организация, предоставляющая финансовые услуги и подпадающая под действие законодательства США в плане нормативно-правового риска и обязанности исполнять нормативно-правовые требования, обязана соответствовать сложной многослойной системе стратегического управления, управления рисками и соблюдения требований (GRC), а также конфиденциальности, целостности и доступности (confidentiality, integrity, and availability, CIA), надзор над исполнением которых осуществляется различными юрисдикциями и органами, в том числе федеральными, штатов, местными и трансграничными
Автор/организация/ эл.почта	П.Кэри (Pw Carey), Compliance Partners LLC, pwc.pwcarey@email.com
Акторы/ заинтересованные лица, их роли и ответственность	Регулирующие и консультативные организации и органы, в том числе Федеральная комиссия по ценным бумагам и биржам (Securities and Exchange Commission, SEC), Федеральная корпорация страхования депозитов (FDIC), Комиссия по торговле товарными фьючерсами (Commodity Futures Trading Commission, CFTC), Казначейство США, Некоммерческая организация по надзору за отчетностью публичных компаний, США (PCAOB), Комитет спонсорских организаций (COSO), CobiT, лица и организации, участвующие в подготовке отчетности, заинтересованные стороны, инвестиционное сообщество, акционеры, пенсионные фонды, высшее руководство организаций, хранители данных и иные сотрудники. На каждом уровне организации финансовых услуг существует взаимосвязанное и взаимозависимое сочетание обязанностей, обязательств и ответственности тех, кто непосредственно несет ответственность за использование, подготовку и передачу финансовых данных, тем самым соответствуя требованиям стратегического управления, управления рисками и соблюдения требований (GRC), (GRC), так и конфиденциальности, целостности и доступности (CIA) финансовых данных их организаций. Эта же информация напрямую связана с поддержанием репутации, доверия и жизнеспособностью бизнеса организации
Цели	В данном варианте использования представлен один из подходов к разработке работоспособной стратегии внедрения больших данных в сфере финансовых услуг. До начала внедрения и переключения на новые технологии организация должна выполнить ряд действий, следуя базовой методологии использования больших данных в рамках облачной экосистемы, адресованной как государственным, так и частным финансовым учреждениям, предлагающим финансовые услуги в рамках федеральной юрисдикции США, юрисдикции штатов и местных органов власти и/или в иных юрисдикциях, таких как Великобритания, Евросоюз и Китай. Каждая предоставляющая финансовые услуги организация должна подходить к введению последующих мер, поддерживающих их инициативу в области больших данных, с пониманием и осознанием того воздействия, которое каждый из накладывающихся друг на друга и взаимозависимых факторов будет оказывать в реализации.

Цели	Эти четыре фактора следующие: 1) люди (как ресурсы), 2) процессы (время/расходы/возврат на инвестиции), 3) технологии (различные операционные системы, платформы, а также зоны влияния/масштабы воздействия технологий), и 4) регуляторное управление (зависит от многочисленных различных регулирующих органов). Кроме того, эти четыре фактора должны быть выявлены, проанализированы, оценены, должны быть приняты соответствующие меры, проведены тестирование и анализ результатов в ходе подготовки к переходу на следующие этапы внедрения: 1) инициирование проекта и получение поддержки со стороны руководства, 2) оценка рисков и выбор мер контроля и управления, 3) анализ влияния на деловую активность, 4) проектирование, разработка и тестирование стратегий обеспечения непрерывности деловой активности, 5) реагирование и деятельность в условиях чрезвычайных ситуаций (известное также как «Восстановление после катастроф»), 6) разработка и внедрение планов обеспечения непрерывности деловой активности, 7) программы ознакомления и обучения, 8) реализация мер по обеспечению непрерывности деловой активности (известное также как Maintaining Regulatory Currency — поддержание доверия со стороны регуляторов). Примечание — Где уместно, эти восемь направлений деятельности должны быть адаптированы и модифицированы в соответствии с потребностями каждой организации, ее уникальной корпоративной культурой и видами оказываемых финансовых услуг
Описание варианта использования	Разработанная Google технология больших данных предназначалась для использования в качестве инструмента индексирования веб-сайтов в интернете, помогая компании сортировать, перемешивать, классифицировать и маркировать интернет. Первоначально она не рассматривалась как замена для устаревших ИТ-инфраструктур данных. Благодаря побочным разработкам в рамках OpenGroup и Hadoop, большие данные превратились в надежный инструмент анализа и хранения данных, который все еще продолжает развиваться. В итоге, однако, технологии больших данных по-прежнему разрабатываются в качестве дополнения к существующим ИТ-архитектурам хранилищ данных типа клиент/сервер/суперкомпьютер, что в некоторых отношениях лучше, чем эти самые среды хранилищ данных, но не во всех. В настоящее время в финансовой отрасли большие данные/Hadoop используются для выявления мошенничества, анализа и оценки рисков, а также для расширения своих знаний и понимания клиентов в рамках стратегии, известной как «знай своего клиента? Однако эта стратегия по-прежнему должна следовать хорошо продуманной таксономии, которая удовлетворяет уникальные и индивидуальные потребности субъектов. Одной из таких стратегий является следующая формальная методология, которая дает ответ на два простейших, но крайне важных вопроса: «Что мы делаем?» и «Почему мы это делаем?». 1) Заявление о политике/устав проекта (цель плана, причины и ресурсы — все это следует определить). 2) Анализ воздействия на деловую деятельность (как приложенные усилия улучшают наши деловые услуги). 3) Определение общесистемных политик, процедур и требований. 4) Определение наилучшей практики внедрения (включая управление изменениями / управление конфигурацией) и/или будущих доработок. 5) План «Б» — стратегии восстановления (как и что нужно будет восстанавливать, если это потребуется). 6) Разработка плана (пишется план и определяются его элементы). 7) Обеспечение поддержки плана в организации и его тестирование (важно, чтобы все знали план и знали, что делать).

Описание варианта использования	8) Реализация плана (затем выявляются и устраняются недостатки — после первых 3 мес, после 6 мес и ежегодно с момента первоначальной реализации). 9) Актуализация (постоянный мониторинг и внесение изменений, отражающих текущее состояние корпоративной среды). 10) Наконец, вывод системы из эксплуатации	
Текущие решения	**Вычислительная система**	В настоящее время большие данные/Hadoop в облачной экосистеме в рамках финансовой отрасли работают как часть гибридной системы, причем технология больших данных используется в качестве полезного инструмента для проведения анализа рисков и выявления мошенничества, а также помогает организациям в процессе «знай своего клиента». Три области, в которых большие данные хорошо себя проявили, следующие: 1) выявление мошенничества, 2) взаимосвязанные риски, и 3) стратегия «знай своего клиента». В то же время традиционные клиент/сервер/хранилище данных/СУБД используются для управления, обработки, хранения и архивирования финансовых данных субъектов. Недавно SEC одобрила инициативу, согласно которой с 13 мая 2013 г. учреждения финансовой отрасли должны будут представлять документы финансовой отчетности в формате XBRL
	Хранилище данных	Одни и те же федеральные, штатов, местные и трансграничные законодательно-нормативные требования могут оказывать влияние в любых географических точках, затрагивая решения VMware, NetApps, Oracle, IBM, Brocade и т. д. **Для сведения** Исходя из требований законодательства эти решения для хранения данных финансовой отрасли должны обеспечивать исполнение существующих на данный момент времени американских законодательно-нормативных требований стратегического управления, управления рисками и соблюдения требований (GRC) и конфиденциальности, целостности и доступности (CIA). Чтобы убедиться в этом, необходимо посетить веб-сайты следующих федеральных органов: Федеральной комиссии по ценным бумагам и биржам (Securities and Exchange Commission, SEC), Комиссии по торговле товарными фьючерсами (Commodity Futures Trading Commission, CFTC), Федеральной корпорации страхования депозитов (FDIC), Министерства юстиции США (U.S. Department of Justice), и Некоммерческой организации по надзору за отчетностью публичных компаний, США (PCAOB)
	Сеть связи	**Для сведения** Одни и те же федеральные, штатов, местные и трансграничные законодательно-нормативные требования могут оказывать влияние в любых географических точках расположения оборудования и программного обеспечения, включая, но не ограничиваясь системами типа WAN, LAN, MAN, WiFi, оптоволокно, доступ в интернет, через публичные, частные, кооперативные и гибридные облачные среды, с VPN или без него.

	Сеть связи	Исходя из требований законодательства эти сетевые решения для данных финансовой отрасли должны обеспечивать исполнение существующих на данный момент времени американских законодательно-нормативных требований стратегического управления, управления рисками и соблюдения требований (GRC) и конфиденциальности, целостности и доступности (CIA), таких как требования Казначейства США. Чтобы убедиться в этом, пожалуйста, посетите веб-сайты следующих федеральных органов: SEC, CFTC, FDIC, Казначейства США, Министерства юстиции США и Некоммерческой организации по надзору за отчетностью публичных компаний, США (PCAOB)
	Программное обеспечение	**Для сведения** Те же федеральные, штатов, местные и трансграничные законодательно-нормативные требования, что оказывают влияние в местах расположения оборудования и программного обеспечения, также ограничивают возможное местоположение для решений с открытым исходным кодом Hadoop, Map/Reduce и проприетарных решений поставщиков, таких как AWS (Amazon Web Services), Google Cloud Services и Microsoft. Исходя из требований законодательства эти программные решения, включающие как протокол SOAP (Simple Object Access Protocol) для веб-разработки, так и программный язык OLAP (online analytical processing) для баз данных, особенно в случае обработки данных финансовой отрасли, должны обеспечивать соответствие этих данных существующим на данный момент времени американским законодательно-нормативным требованиям стратегического управления, управления рисками и соблюдения требований (GRC) и конфиденциальности, целостности и доступности (CIA). Чтобы убедиться в этом, пожалуйста, посетите веб-сайты следующих федеральных органов: SEC, CFTC, Казначейства США, FDIC, Министерства юстиции США и Некоммерческой организации по надзору за отчетностью публичных компаний, США (PCAOB)
Характеристики больших данных	**Источник данных (распределенный/централизованный)**	**Для сведения** Те же федеральные, штатов, местные и трансграничные законодательно-нормативные требования, что оказывают влияние в местах расположения оборудования и программного обеспечения, также оказывают свое влияние в местах расположения распределенных/централизованных источников данных, поступающих в среду высокой доступности с обеспечением восстановления после катастроф (HA/DR Environment) и в хостинговый виртуальный сервер (HVS), например, в следующих конфигурациях: DC1 — > VMWare/KVM (кластеры, с виртуальными брандмауэрами), Data link — VMWare Link — Vmotion Link — Network Link, несколько мостовых соединений с поставщиком (PB) в рамках NaaS (сеть как сервис), DC2 — > VMWare/KVM (кластеры с виртуальными брандмауэрами), DataLink (Vmware Link, Vmotion Link, Network Link), несколько мостовых соединений с поставщиком в рамках NaaS (требуется отказоустойчивая виртуализация), среди прочих соображений.

Характеристики больших данных	**Источник данных (распределенный/ централизованный)**	Исходя из требований законодательства эти решения для источников данных, как распределенных, так и/или централизованных, в случае обработки данных финансовой отрасли должны обеспечивать соответствие этих данных существующим на данный момент времени американским законодательно-нормативным требованиям стратегического управления, управления рисками и соблюдения требований (GRC) и конфиденциальности, целостности и доступности (CIA). Чтобы убедиться в этом, пожалуйста, посетите веб-сайты следующих федеральных органов: SEC, CFTC, Казначейства США, FDIC, Министерства юстиции США и Некоммерческой организации по надзору за отчетностью публичных компаний, США (PCAOB)
	Объем (количество)	От нескольких терабайт до нескольких петабайт. **Для сведения** Это зона, свободная от флоппи-дисков
	Скорость обработки (например, в реальном времени)	При использовании больших данных в финансовой отрасли скорость обработки более важна для выявления мошенничества, оценки риска и в рамках процесса «знай своего клиента». **Для сведения** Однако исходя из требований законодательства, скорость обработки не является проблемой для решений на основе технологии больших данных, используемых для обработки данных финансовой отрасли, за исключением задач выявления мошенничества, анализа рисков и анализа клиентов. Исходя из установленных законодательством ограничений, **скорость обработки** не является проблемой; скорее, главной проблемой при обработке данных финансовой отрасли является необходимость обеспечивать соответствие существующим на данный момент времени американским законодательно-нормативным требованиям стратегического управления, управления рисками и соблюдения требований (GRC) и конфиденциальности, целостности и доступности (CIA)
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Различные виртуальные среды, работающие в рамках архитектуры пакетной обработки или параллельной архитектуры с «горячей» заменой (hot-swappable parallel architecture), поддерживающие выявление мошенничества, оценку риска и решений по обслуживанию клиентов. **Для сведения** Исходя из требований законодательства, разнообразие не является проблемой для решений на основе технологии больших данных, используемых для обработки данных финансовой отрасли в рамках облачной экосистемы, за исключением задач выявления мошенничества, анализа рисков и анализа клиентов. Исходя из установленных законодательством ограничений, **разнообразие** не является проблемой; скорее, главной проблемой при обработке данных финансовой отрасли является необходимость обеспечивать соответствие существующим на данный момент времени американским законодательно-нормативным требованиям стратегического управления, управления рисками и соблюдения требований (GRC) и конфиденциальности, целостности и доступности (CIA)

	Вариативность (темпы изменения)	**Для сведения** Исходя из требований законодательства, **вариативность** не является проблемой для решений на основе технологии больших данных, используемых для обработки данных финансовой отрасли в рамках облачной экосистемы, за исключением задач выявления мошенничества, анализа рисков и анализа клиентов. Исходя из установленных законодательством ограничений, **вариативность** не является проблемой; скорее, главной проблемой при обработке данных финансовой отрасли является необходимость обеспечивать соответствие существующим на данный момент времени американским законодательно-нормативным требованиям стратегического управления, управления рисками и соблюдения требований (GRC) и конфиденциальности, целостности и доступности (CIA). Вариативность больших данных финансовой отрасли в облачной экосистеме будет зависеть от силы и полноты соглашений об уровне обслуживания (SLA), от связанных с деловой активностью и зависящих от ее потребностей капитальных затрат (CapEx)
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	**Для сведения** Исходя из требований законодательства, **достоверность** не является проблемой для решений на основе технологии больших данных, используемых для обработки данных финансовой отрасли в рамках облачной экосистемы, за исключением задач выявления мошенничества, анализа рисков и анализа клиентов. Исходя из установленных законодательством ограничений, **достоверность** не является проблемой; скорее, главной проблемой при обработке данных финансовой отрасли является необходимость обеспечивать соответствие существующим на данный момент времени американским законодательно-нормативным требованиям стратегического управления, управления рисками и соблюдения требований (GRC) и конфиденциальности, целостности и доступности (CIA). В облачной экосистеме больших данных целостность данных важна на протяжении всего жизненного цикла организации, связанных с защитой персональных данных и обеспечением безопасности и законодательно-нормативных требований стратегического управления, управления рисками и соблюдения требований (GRC) и конфиденциальности, целостности и доступности (CIA)
	Визуализация	**Для сведения** Исходя из требований законодательства, **визуализация** не является проблемой для решений на основе технологии больших данных, используемых для обработки данных финансовой отрасли, за исключением задач выявления мошенничества, анализа рисков и анализа клиентов; данные обрабатываются традиционными клиент/сервер/хранилище данных — серверами на базе суперкомпьютеров.

	Визуализация	Исходя из установленных законодательством ограничений, **визуализация** не является проблемой; скорее, главной проблемой при обработке данных финансовой отрасли является необходимость обеспечивать соответствие существующим на данный момент времени американским законодательно-нормативным требованиям стратегического управления, управления рисками и соблюдения требований (GRC) и конфиденциальности, целостности и доступности (CIA). Целостность данных в рамках больших данных играет критически важную роль на протяжении всего жизненного цикла организации ввиду законодательно-нормативных требований стратегического управления, управления рисками и соблюдения требований (GRC) и конфиденциальности, целостности и доступности (CIA)
	Качество данных (синтаксис)	**Для сведения** Исходя из требований законодательства, **качество данных** всегда будет серьезным вопросом, вне зависимости от отрасли или платформы. Исходя из установленных законодательством ограничений, **качество данных** является ключевым для целостности данных; и оно представляет собой главную проблему при обработке данных финансовой отрасли в связи с необходимостью обеспечивать их соответствие существующим на данный момент времени американским законодательно-нормативным требованиям стратегического управления, управления рисками и соблюдения требований (GRC) и конфиденциальности, целостности и доступности (CIA). Для больших данных финансовой отрасли целостность данных играет критически важную, ключевую роль на протяжении всего жизненного цикла организации ввиду, целостности и доступности (CIA)
	Типы данных	**Для сведения** Исходя из требований законодательства, **типы данных** важны ввиду того, что они должны обладать определенной степенью согласованности и особенно жизнеспособности во время аудитов и цифровой криминалистической экспертизы, когда деградация формата данных может негативно повлиять как на аудит, так и на криминалистическую экспертизу, когда те проходят через несколько циклов. Многочисленные типы данных и форматов в составе больших данных финансовой отрасли включают (но не ограничиваются ими): плоские файлы, txt, .pdf, файлы приложений для Android, .wav, .jpg и VOIP (передача голоса с использованием протокола IP)
	Аналитика данных	**Для сведения** Исходя из требований законодательства **аналитика данных** является серьезным вопросом для решений на основе технологии больших данных, используемых для обработки данных финансовой отрасли, особенно в плане задач выявления мошенничества, анализа рисков и анализа клиентов. В то же время задачи аналитики данных для данных финансовой отрасли в настоящее время обрабатываются традиционными клиент/сервер/хранилище данных — серверами на базе суперкомпьютеров, которые должны обеспечивать соответствие существующим на данный момент времени американским законодательно-нормативным требованиям стратегического управления, управления рисками и соблюдения требований (GRC) и конфиденциальности, целостности и доступности (CIA).

	Аналитика данных	Для целей аналитики данных на основе больших данных финансовой отрасли, данные должны поддерживаться в формате, исключающем деградацию во время обработки и процедур поиска и анализа
Иные проблемы больших данных	В настоящее время проблемные области, связанные с большими данными финансовой отрасли в рамках облачной экосистемы, включают в себя агрегирование и хранение данных (чувствительных, токсичных и иных) из нескольких источников, что может создавать и создает административные и управленческие проблемы, связанные со следующими аспектами: - контроль доступа, - управление / администрирование, - право на получение данных, и - права собственности на данные. Тем не менее, как показывает текущий анализ, эти вопросы и проблемы широко известны и решаются в данный момент времени поставщиками технологий с помощью методологий управления жизненным циклом программного обеспечения и оборудования (Software Development Life Cycle/Hardware Development Life Cycle, SDLC/HDLC) на стадиях исследований и разработки	
Проблемы пользовательского интерфейса и мобильного доступа	Обеспечение мобильного доступа — это постоянно растущий слой технической сложности, однако не все решения для мобильного использования больших данных носят технический характер. Есть две взаимосвязанные и взаимозависимые стороны, которые должны работать вместе над тем, чтобы найти работоспособное и жизнеспособное решение — это представители основной деятельности финансовой отрасли и ИТ. Технические проблемы решаемы, если обе эти стороны согласны использовать общую лексику и таксономию и уважают, и понимают требования, которые каждая из них обязана удовлетворить. Обе стороны в рамках этих совместных усилий столкнутся со следующими существующими и длящимися проблемными вопросами, связанными с данными финансовой отрасли: - несогласованность при отнесении к категориям, - изменения с течением времени в системах классификации, - использование нескольких перекрывающихся или различающихся схем категоризации. Помимо решения задачи, связанной с этими изменяющимися и эволюционирующими несоответствиями, необходимо также обеспечить следующие характеристики данных, связанные с принципом ACID: - **атомарность** (Atomic) — либо будут полностью выполнены все подоперации в рамках транзакции, либо не будет выполнена ни одна из них. - **согласованность** (Consistent) — в результате выполнения транзакции база данных переходит из одного согласованного состояния в другое согласованное состояние. Согласованность определяется с точки зрения выполнения ограничений. - **изолированность** (Isolated) — результаты любых изменений, внесенных в ходе транзакции, не видны до тех пор, пока транзакция не будет полностью завершена. - **стойкость** (Durable) — изменения, внесенные успешно совершенной транзакцией, должны сохраниться в случае сбоев и отказов системы	
Технические проблемы обеспечения безопасности и защиты персональных данных	Никакое количество должной предусмотрительности, проявленной в вопросах безопасности и защиты персональных данных, не способно компенсировать врожденные недостатки, связанные с природой человека и проникающие в любую программу и/или стратегию. В настоящее время при внедрении технологии больших данных в финансовой отрасли приходится иметь дело с растущим числом групп риска, среди которых, в частности, можно назвать такие, как: - борьба с легализацией (отмыванием) незаконных доходов (Anti-Money Laundering), - надлежащая проверка клиентов (Client Due Diligence), - списки наблюдения (Watch lists), - федеральный закон США о борьбе с коррупцией в международной деятельности (Foreign Corrupt Practices Act, FCPA).	

Технические проблемы обеспечения безопасности и защиты персональных данных	Для того чтобы получить представление о реальном положении дел, посмотрите на девятилетние усилия Гарри Маркополоса (Harry M. Markopolos), направленные на то, чтобы заставить SEC, среди других федеральных органов исполнительной власти США, выполнить свою работу и закрыть финансовую пирамиду Бернарда Мэдоффа (Bernard Madoff) на сумму в миллиард долларов. Помимо этого, выявление и удовлетворение требований по защите неприкосновенности частной жизни и безопасности для организаций финансовой отрасли, предоставляющих услуги в рамках экосистемы больших данных/облака, благодаря постоянному совершенствованию: 1) технологий, 2) процессов, 3) процедур, 4) кадров и 5) нормативного регулирования, — это гораздо лучший выбор как для отдельного человека, так и для организации, особенно если сравнить с альтернативами. Используя многоуровневый подход, данную стратегию можно разбить на следующие подкатегории: 1) поддержание устойчивости операционной деятельности, 2) защита ценных активов, 3) контроль над учетными записями в системе, 4) эффективное управление сервисами безопасности, и 5) поддержание устойчивости операционной деятельности. За дополнительной информацией о базовых решениях задач безопасности и защиты персональных данных рекомендуется обращаться к двум организациям: - Международная ассоциация аудита и контроля информационных систем (ISACA); - Международный консорциум по сертификации в области безопасности информационных систем (isc2)
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Проблемные области включают в себя агрегирование и хранение данных из нескольких источников, где могут возникнуть проблемы, связанные с: - контролем доступа, - управлением / администрированием, - правом на получение данных, и - правами собственности на данные. Каждая из этих областей совершенствуется, но на них, тем не менее, следует обратить внимание и принять меры, используя решения для контроля доступа и инструменты управления информацией о безопасности и событиями безопасности SIEM (Security information and event management). Эта задача пока что не решена, принимая во внимание проблемы безопасности, которые упоминаются всякий раз, когда речь заходит о больших данных/Hadoop в рамках облачной экосистемы. Текущие и длящиеся проблемы внедрения больших данных для финансовой отрасли в рамках облачной экосистемы, а также традиционных архитектур типа клиент/сервер/хранилище данных, включают следующие области финансового учета в соответствии как с общепринятыми принципами бухгалтерского учета США (U.S. Generally Accepted Accounting Practices, US GAAP), так и Международными стандартами финансовой отчетности, МСФО (International Financial Reporting Standards, IFRS): - использование расширяемого языка разметки для деловой отчетности (XBRL); - согласованность (терминологии, форматирования, технологий, нормативного регулирования); - предписание Федеральной комиссии по ценным бумагам и биржам SEC использовать XBRL для финансовой отчетности перед регулятором; - меняются требования SEC, принципы бухгалтерской отчетности GAAP/IFRS и еще не полностью завершенное новое финансовое законодательство, влияющее на требования к отчетности, — и эти изменения указывают на попытки усовершенствовать наилучшие практики внедрения, тестирования, обучения, отчетности и информационного обмена, требуемые от независимого аудитора в отношении аудита, аудиторских отчетов, самооценки мер контроля и управления, финансовых аудитов, внутренних аудитов, соблюдения Общепринятых стандартов аудита (Generally Accepted Auditing Standards, GAAS) / Международных стандартов аудита (International Standard on Auditing, ISA), а также Закона Сарбейнса-Оксли 2002 г. (Sarbanes-Oxley Act of 2002, SOX)

Дополнительная информация (гиперссылки)	1) «10 главных проблем обеспечения безопасности и неприкосновенности частной жизни при использовании технологии больших данных» (Top 10 Challenges in big data Security and Privacy), Рабочая группа по большим данным Альянса облачной безопасности (Cloud Security Alliance), 2012, https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Big_Data_Top_Ten_v1.pdf 2) Рабочая группа «Международная финансовая отчетность, ценные бумаги и рынки» (IFRS, Securities and Markets Working Group) на сайте сообщества XBRL Europe (https://www.xbrleurope.org/), продвигающего использование языка XBLR в Европе, см. https://www.xbrleurope.org/?page_id=357 3) Конференция по большим данным Международной ассоциации специалистов по электротехнике и радиоэлектронике IEEE (IEEE International Conference on Big Data), см. http://bigdataieee.org/ 4) Сайт по технологии Map/Reduce, http://www.mapreduce.org (ссылка неработающая) 5) Некоммерческая организация по надзору за отчетностью публичных компаний, США (PCAOB), https://pcaobus.org/ 6) Аналитика по вопросам страхования на сайте фирмы «Эрнст и Янг» (Ernst & Young), см. https://www.ey.com/en_gl/insurance 7) Ресурсы по теме финансовых рынков и финансовых институтов на сайте Казначейства США, см. https://www.treasury.gov/resource-center/fin-mkts/Pages/default.aspx 8) Комиссия по торговле товарными фьючерсами (Commodity Futures Trading Commission, CFTC), см. https://www.cftc.gov/ 9) Федеральная комиссия по ценным бумагам и биржам (Securities and Exchange Commission, SEC), см. https://www.sec.gov/ 10) Федеральная корпорация страхования депозитов (FDIC), см. https://www.fdic.gov/ 11) Комитет спонсорских организаций (COSO), см. https://www.coso.org/ 12) Международный консорциум по сертификации в области безопасности информационных систем (isc2), см. https://www.isc2.org/ 13) Международная ассоциация аудита и контроля информационных систем (ISACA), см. https://www.isaca.org/ 14) Фонд IFRS — разработчик Международных стандартов финансовой отчетности, МСФО (International Financial Reporting Standards, IFRS), см. https://www.ifrs.org/ 15) Сайт консорциума Open Group, https://www.opengroup.org/ 16) Джейкумар Виджаян (Jaikumar Vijayan) «ИТ должно подготовиться к проблемам безопасности в Hadoop» (IT must prepare for Hadoop security issues), Computerworld, 9 ноября 2011 года, см. https://www.computerworld.com/article/2498601/it-must-prepare-for-hadoop-security-issues.html 17) Гарри Маркополос «Финансовая пирамида Бернарда Мэдоффа. Расследование самой грандиозной аферы в истории», изд-во Диалектика, 2012, ISBN: 978-5-8459-1686-0, 978-0-470-55373-2 18) «Оценка финансовой пирамиды Мэдоффа и провалов в работе регуляторов» (Assessing the Madoff Ponzi Scheme and Regulatory Failures), слушания подкомитета по рынкам капитала, страхованию и спонсируемым государством предприятиям (Subcommittee on Capital Markets, Insurance, and Government Sponsored Enterprises), 2009 год, https://www.gpo.gov/fdsys/pkg/CHRG-111hhrg48673/pdf/CHRG-111hhrg48673.pdf 19) Сайт ITIL (Библиотека инфраструктуры информационных технологий), см. https://www.axelos.com/best-practice-solutions/itil 20) Стандарт COBIT® 2019 CobiT (от Control Objectives for Information and Related Technology — «Цели управления информационными и смежными технологиями) на сайте Международной ассоциации аудита и контроля информационных систем (ISACA), см. https://www.isaca.org/resources/cobit 21) Концепция архитектуры «Открытой группы» (The Open Group Architecture Framework, TOGAF) версии 9.2, http://www.opengroup.org/togaf/ 22) Международный стандарт ИСО/МЭК 27000:2018 «Информационная технология. Методы и средства обеспечения безопасности. Системы менеджмента информационной безопасности. Общий обзор и терминология» (Information technology — Security techniques — Information security management systems — Overview and vocabulary), https://www.iso.org/standard/73906.html , свободно доступен по адресу https://standards.iso.org/ittf/PubliclyAvailableStandards/c073906_ISO_IEC_27000_2018_E.zip[1)]

[1)] *В России стандарт адаптирован (в более ранней редакции) как ГОСТ Р ИСО/МЭК 27000—2012, см. http://protect.gost.ru/v.aspx?control=8&baseC=6&id=175549*

А.2.2 Вариант использования № 6: Международная исследовательская сеть Mendeley

Название	Международная исследовательская сеть Mendeley	
Предметная область	Коммерческие облачные услуги для клиентов	
Автор/организация/ эл.почта	Уильям Ган (William Gunn) / Mendeley / william.gunn@mendeley.com	
Акторы/ заинтересованные лица, их роли и ответственность	Исследователи, библиотекари, издатели и финансирующие организации	
Цели	Содействие более быстрому прогрессу в научных исследованиях, обеспечивая возможность исследователям эффективно сотрудничать, библиотекарям — понимать потребности исследователей, издателям — быстрее и шире распространять результаты исследований, а финансирующим организациям — лучше понимать воздействие финансируемых ими проектов	
Описание варианта использования	Международная сеть «Менделей» (Mendeley) создала базу данных научно-исследовательских материалов, которая облегчает создание коллективно используемых библиографий. Менделей использует собранную информацию о закономерностях чтения материалов об исследованиях, а также о других видах деятельности, осуществляемых с помощью программного обеспечения, с целью создания более эффективных инструментов для поиска и анализа научной литературы. Системы интеллектуального анализа и классификации текста позволяют автоматически рекомендовать взаимосвязанные исследования, повышая производительность и экономическую эффективность исследовательских групп, в особенности тех, которые занимаются мониторингом литературы по конкретной теме, таких как группа «Информатика генома мышей» (Mouse Genome Informatics) в некоммерческом научно-исследовательском институте Jackson Laboratory, в которой большая группа специалистов занимается просмотром литературы «вручную». Другие варианты использования включают поддержку более быстрого распространения публикаций издателями, содействие научно-исследовательским учреждениям и библиотекарям в исполнении планов менеджмента данных, а также предоставление спонсорам возможности лучше понять воздействие финансируемой ими работы благодаря доступным в реальном времени данным о доступе и использовании финансируемых исследований	
Текущие решения	**Вычислительная система**	Amazon EC2
	Хранилище данных	HDFS Amazon S3
	Сеть связи	Клиент — серверные соединения между компьютерами Mendeley и конечных пользователей, соединения между офисами Mendeley и сервисами Amazon
	Программное обеспечение	Hadoop, Scribe, Hive, Mahout, Python
Характеристики больших данных	**Источник данных (распределенный/ централизованный)**	Распределенные и централизованные
	Объем (количество)	В настоящее время 15 терабайт, с темпом прироста около 1 терабайта в месяц
	Скорость обработки (например, в реальном времени)	В настоящее время пакетные задания Hadoop планируются раз в день, но началась работа над рекомендациями по выполнению работ в реальном времени
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	PDF-документы, лог-файлы социальной сети и активности клиентов
	Вариативность (темпы изменения)	В настоящее время темпы роста высокие, поскольку все больше исследователей подписываются на данную услугу; темпы роста сильно колеблются в течение года

Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Извлечение метаданных из PDF-файлов осуществляется в различной степени; выявление дубликатов является проблемой; нет универсальной системы идентификаторов для документов и авторов (хотя «Открытый идентификатор исследователя и участника» ORCID (Open Researcher and Contributor ID) обещает решить эту задачу)
	Визуализация	Визуализация сети с использованием программного обеспечения Gephi, диаграммы рассеяния (scatterplots) в плоскости читательская аудитория — цитируемость, и т. д.
	Качество данных (синтаксис)	На основе сопоставления со сведениями в базах данных Crossref, PubMed и arXiv, корректность извлечения метаданных оценивается в 90 %
	Типы данных	В основном PDF-файлы, а также некоторое количество графических образов, электронных таблиц и презентаций
	Аналитика данных	Стандартные библиотеки для проведения машинного обучения и аналитики, выполнения латентного размещения Дирихле (LDA), а также специально разработанные инструменты составления отчетности и визуализации данных для агрегирования сведений о читательской и социальной активности, связанной с каждым документом
Иные проблемы больших данных	База данных содержит примерно 400 миллионов документов, в том числе около 80 миллионов уникальных документов, принимая в рабочие дни от 500 до 700 тысяч новых загрузок. Таким образом, основная проблема заключается в группировке соответствующих друг другу документов вычислительно эффективным (т. е. масштабируемым и распараллеливаемым) способом, когда они загружаются из разных источников и могут быть слегка модифицированы инструментами аннотирования третьих сторон или же путем присоединения титульных страниц или наложения «водяных знаков» издателя	
Проблемы пользовательского интерфейса и мобильного доступа	Доставка контента и услуг на различные вычислительные платформы, от настольных компьютеров под Windows до мобильных устройств под ОС Android и iOS	
Технические проблемы обеспечения безопасности и защиты персональных данных	Исследователи, особенно отраслевые, часто хотят, чтобы сведения о том, что они читают, оставались конфиденциальными, поэтому доступ к данным о том, кто что читает, контролируется	
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Данный вариант использования может быть обобщен как предоставление основанных на контенте рекомендаций для различных сценариев потребления информации	
Дополнительная информация (гиперссылки)	Сайт Mendeley, https://www.mendeley.com/ Портал Mendeley для разработчиков, https://dev.mendeley.com/	

А.2.3 Вариант использования № 7: Сервис кинофильмов Netflix

Название	Сервис кинофильмов Netflix	
Предметная область	Коммерческие облачные услуги для клиентов	
Автор/организация/ эл.почта	Джоффри Фокс (Geoffrey Fox), университет штата Индиана (США), gcf@indiana.edu	
Акторы/ заинтересованные лица, их роли и ответственность	Компания Netflix (устойчивое развитие бизнеса), провайдер облачных услуг (поддержка потоковой передачи и анализа данных), пользователь-клиент (отбор и просмотр хороших фильмов по требованию)	
Цели	Обеспечение потоковой передачи выбранных пользователем фильмов с целью достижения нескольких целей (в интересах различных заинтересованных сторон), — в первую очередь, с целью удержания подписчиков. Определение наилучшей возможной подборки видеоматериалов для пользователя (домохозяйства) в заданном контексте, в режиме реального времени; максимизация потребления фильмов	
Описание варианта использования	Цифровые фильмы хранятся в облаке вместе с метаданными, а также с индивидуальными профилями пользователей и рейтингами для небольшой части фильмов. Используется несколько критериев: рекомендательная система на основе контента, рекомендательная система на основе данных пользователей и разнообразие. Алгоритмы постоянно совершенствуются с помощью A/B — тестирования	
Текущие решения	**Вычислительная система**	Amazon Web Services (AWS)
	Хранилище данных	Используется технология Cassandra NoSQL вместе с Hive, Teradata
	Сеть связи	Требуется система доставки контента для поддержки эффективного потокового видео
	Программное обеспечение	Hadoop и Pig, Cassandra, Teradata
Характеристики больших данных	**Источник данных (распределенный/ централизованный)**	Фильмы добавляются в сервис на основе соглашений с производителями контента. Распределенным образом собираются пользовательские рейтинги и профили
	Объем (количество)	По состоянию на лето 2012 г.: 25 млн подписчиков; 4 млн оценок в день; 3 млн поисковых запросов в день; 1 млрд часов потокового видео в июне 2012 г. Объем облачного хранения 2 петабайта (июнь 2013 г.)
	Скорость обработки (например, в реальном времени)	Контент (видео и характеристики) и рейтинги постоянно обновляются
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Данные варьируются от цифровых мультимедийных материалов до пользовательских рейтингов, профилей пользователей и параметров мультимедиа, используемых для основанных на контенте рекомендаций
	Вариативность (темпы изменения)	Потоковое видео — очень конкурентный бизнес. Необходимо знать о других компаниях, а также о тенденциях, связанных как с контентом (какие фильмы популярны), так и с технологиями. Нужно изучать новые деловые инициативы, такие, как спонсируемый Netflix контент

Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Для успешности бизнеса требуется отличное качество обслуживания
	Визуализация	Потоковое мультимедиа и качественный пользовательский опыт, позволяющий выбирать контент
	Качество данных (синтаксис)	Рейтинги по своей природе являются «непричесанными» данными, и для их обработки требуются надежные и устойчивые алгоритмы обучения
	Типы данных	Мультимедийный контент, профили пользователей, набор пользовательских рейтингов
	Аналитика данных	Рекомендательные системы и доставка потокового видео. Рекомендательные системы всегда персонализированы и используют логистическую/линейную регрессию, эластичные сети, факторизацию матриц, кластеризацию, латентное размещение Дирихле (LDA), ассоциативные правила, градиентный бустинг деревьев решений и другие инструменты. Победитель конкурса Netflix, в котором ставилась задача повышения рейтинга на 10 %, использовал комбинацию более 100 различных алгоритмов
Иные проблемы больших данных	Аналитика требует постоянного мониторинга и совершенствования	
Проблемы пользовательского интерфейса и мобильного доступа	Мобильный доступ важен	
Технические проблемы обеспечения безопасности и защиты персональных данных	Необходимо защитить неприкосновенность частной жизни пользователей и цифровые права на контент.	
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Рекомендательные системы имеют общие черты с системами электронной коммерции типа Amazon. Потоковое видео имеет общие черты с другими сервисами доставки контента, такими как iTunes, Google Play, Pandora и Last.fm	
Дополнительная информация (гиперссылки)	Ксавьер Аматрян (Xavier Amatriain) «Создание реальных крупномасштабных рекомендательных систем — Обучающий курс Recsys — 2012» (Building Large — scale Real — world Recommender Systems — Recsys — 2012 Tutorial), конференция по рекомендательным системам 2012 г. Recsys-2012 в Дублине, Ирландия, https://www.slideshare.net/xamat/building-largescale-realworld-recommender-systems-recsys2012-tutorial «Алгоритм надежного выявления аномалий (Robust Anomaly Detection, RAD) — Выявление аномалий в больших данных» (RAD — Outlier Detection on big data), блог Netflix по техническим вопросам, https://netflixtechblog.com/rad-outlier-detection-on-big-data-d6b0494371cc	

А.2.4 Вариант использования № 8: Веб-поиск

Название	Веб-поиск (Bing, Google, Yahoo и др.)	
Предметная область	Коммерческие облачные услуги для клиентов	
Автор/организация/ эл.почта	Джоффри Фокс (Geoffrey Fox), университет штата Индиана (США), gcf@indiana.edu	
Акторы/ заинтересованные лица, их роли и ответственность	Владельцы выложенной в Интернете информации, по которой проводится поиск; компании — поставщики поисковых систем; рекламодатели; пользователи	
Цели	Выдать примерно через ~ 0,1 секунды результаты поиска по запросу, включающему в среднем три слова. Важно максимизировать такие метрики, как «точность 10 наилучших результатов» (precision@10), отражающие количество высокоточных / соответствующих запросу ответов в первой десятке лучших ранжированных результатов.	
Описание варианта использования	1) Провести сканирование Интернета; 2) провести предварительную обработку данных с целью выделения элементов, по которым можно вести поиск (слова, позиции); 3) сформировать инвертированный индекс, связывающий слова с их местоположением в документах; 4) ранжирование документов по релевантности с использованием алгоритма PageRank; 5) использовать разнообразные рекламно-маркетинговые технологии, обратное проектирование определения моделей ранжирования либо блокирование обратного проектирования; 6) провести кластеризацию документов по темам (как в Google News); 7) обеспечить эффективное обновление результатов	
Текущие решения	**Вычислительная система**	Крупные облачные системы
	Хранилище данных	Инвертированный индекс не является огромным; в то же время собранные в ходе сканирования Интернета материалы представляют собой петабайты текста, а мультимедийные материалы по объемам еще намного больше
	Сеть связи	В плане сетевой инфраструктуры, необходимы отличные внешние сетевые соединения; большинство операций хорошо распараллеливаются и требовательны к скорости ввода/вывода (I/O sensitive). Высокая производительность внутренней сети не требуется
	Программное обеспечение	Map/Reduce + Bigtable; Dryad + Cosmos. PageRank. Последний этап по сути представляет собой рекомендательную систему
Характеристики больших данных	**Источник данных (распределенный/ централизованный)**	Распределенные веб-сайты
	Объем (количество)	В общей сложности около 45 млрд веб-страниц; ежедневно загружается 500 млн фотографий; и ежеминутно на YouTube закачивается 100 часов видеоматериалов
	Скорость обработки (например, в реальном времени)	Данные постоянно обновляются
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Богатый набор функций. После обработки данные для каждой страницы (кроме мультимедийных объектов) аналогичны
	Вариативность (темпы изменения)	В среднем срок существования веб-страницы составляет несколько месяцев

Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Абсолютная точность результатов не является жизненно необходимой, однако важно, чтобы были найдены соответствующие поисковому запросу основные центры компетенций и авторитетные источники
	Визуализация	Не важна, однако схема расположения выдаваемых результатов (page layout) имеет ключевое по важности значение
	Качество данных (синтаксис)	Огромное количество дублирования и спама
	Типы данных	В основном текст, но растет интерес к быстро растущим объемам графических образов и видеоконтента
	Аналитика данных	Веб-сканирование, поиск (в том числе по тематике), ранжирование, рекомендации
Иные проблемы больших данных	Поиск по «глубинному интернету» (deep web-контент, не индексируемый стандартными поисковыми системами, скрытый за пользовательскими интерфейсами к базам данных и т. д.). Ранжирование результатов, способное учитывать как внутреннюю ценность материалов (как в алгоритме PageRank), так и ценность для маркетинга. Связывание профилей пользователей с данными из социальных сетей	
Проблемы пользовательского интерфейса и мобильного доступа	Мобильный поиск должен иметь похожие интерфейсы и выдавать похожие результаты	
Технические проблемы обеспечения безопасности и защиты персональных данных	Следует принимать во внимание ограничения на веб-сканирование; избегать спама в результатах поиска	
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Взаимосвязь с направлением поиска и извлечения информации (information retrieval), например с поиском научных работ	
Дополнительная информация (гиперссылки)	Мэри Миикер (Mary Meeker) и Лиан Ву (Liang Wu) из фирмы Kleiner Perkins Caufield & Byers (KPCB), «Тенденции развития интернета» (Internet Trends — D11 Conference), 29 мая 2013 года, https://www.slideshare.net/betobetico/kpcb-internet-trends-2013-mary-meeker План учебного курса «Введение в технологию поисковых систем» (236621 Introduction to Search Engine Technology), Израильский технологический институт «Технион», 2011–2012, https://webcourse.cs.technion.ac.il/236621/Winter2011-2012/comp/WCFiles/syllabus3p-2011-12.pdf План учебного курса SS 2011 «Поиск и извлечение информации и системы поиска в интернете» (Information Retrieval and Web Search Engines) и Институте информационных систем Технического университета Брауншвейга, Германия, http://www.ifis.cs.tu-bs.de/teaching/ss-11/irws Дипак Агарваль (Deepak Agarwal) и Бичун Чень (Bee-Chung Chen), «Учебный курс ICML'11: Проблемы рекомендационных систем для веб-приложений. Часть 1: Введение» (ICML'11 Tutorial: Recommender Problems for Web Applications. Part 1: Introduction), Международная конференция по машинному обучению (International Conference on Machine Learning, ICML) 2011 года, https://www.slideshare.net/beechung/recommender-systems-tutorialpart1intro Сайт «Объем всемирной паутины» (The size of the World Wide Web (The Internet)), https://www.worldwidewebsize.com/	

А.2.5 Вариант использования № 9: Обеспечение непрерывности деловой деятельности и восстановления после катастроф для больших данных в облачной экосистеме

Название	Обеспечение непрерывности деловой деятельности и восстановления после катастроф по модели IaaS (инфраструктура как сервис) для больших данных в облачной экосистеме, осуществляемое провайдерами облачного сервиса (CSP) и провайдерами облачного брокерского сервера (CBSP)
Предметная область	Крупномасштабное надежное хранение данных
Автор/ организация/ эл. почта	П.Кэри (Pw Carey), Compliance Partners LLC, pwc.pwcarey@email.com
Акторы/ заинтересованные лица, их роли и ответственность	Высшее руководство, хранители данных и сотрудники, ответственные за целостность, защиту, обеспечение неприкосновенности частной жизни, конфиденциальности, доступности, безопасности, защищенности и живучести деловой деятельности посредством обеспечения трех аспектов доступности данных для сервисов организации: в любое время, в любом месте и на любом устройстве
Цели	Ниже представлен один из подходов к разработке работоспособной стратегии обеспечения непрерывности деловой деятельности и восстановления после катастроф (BC/DR). Прежде чем отдать данную стратегию организации на аутсорсинг, переложив ее на плечи провайдера облачного сервиса (CSP) или провайдера облачного брокерского сервера (CBSP), организация должна выполнить следующий комплекс работ, обеспечивающий любой организации, как государственной, так и частной, разработку базовой методологии для реализации наилучших практик BC/DR в рамках облачной экосистемы. Каждая организация должна рассмотреть десять сфер деятельности, поддерживающим обеспечение непрерывности деловой деятельности и восстановление после катастроф, с тем, чтобы понять и оценить то влияние, которое каждый из следующих четырех перекрывающихся и взаимозависимых факторов может оказать на обеспечении работоспособности решения по реализации BC/DR — плана организации. Этими четырьмя факторами являются люди (как ресурсы), процессы (например, время / затраты / возврат инвестиций (ROI)), технологии (например, различные операционные системы, платформы, а также зоны влияния/масштабы воздействия технологий) и стратегическое управление (зависит от многочисленных различных регулирующих органов). Данные четыре фактора должны быть выявлены, проанализированы, оценены, должны быть приняты соответствующие меры, проведены тестирование и анализ результатов. Данные факторы должны быть приняты во внимание на следующих десяти этапах: 1) инициирование проекта и получение поддержки со стороны руководства, 2) оценка рисков и выбор мер контроля и управления, 3) анализ влияния на деловую деятельность, 4) проектирование, разработка и тестирование стратегий обеспечения непрерывности деловой деятельности, 5) реагирование и деятельность в условиях чрезвычайных ситуаций (известное также как «Восстановление после катастроф»), 6) разработка и внедрение планов обеспечения непрерывности деловой деятельности, 7) программы ознакомления и обучения, 8) реализация мер по обеспечению непрерывности деловой деятельности (известное также как Maintaining Regulatory Currency — поддержание доверия со стороны регуляторов). 9) подготовка планов взаимодействия с общественностью (Public Relations, PR) и кризисного управления, 10) координация с государственными органами. Примечание — Где это уместно, эти десять направлений деятельности могут быть адаптированы к потребностям организации

<table>
<tr><td>Описание варианта использования</td><td colspan="2">Разработанная Google технология больших данных предназначались для использования в качестве инструмента индексирования веб-сайтов в Интернете, помогая компании сортировать, перемешивать, классифицировать и маркировать Интернет. Первоначально она не рассматривалась как замена для устаревших ИТ-инфраструктур данных. Благодаря побочным разработкам в рамках OpenGroup и Hadoop, большие данные превратились в надежный инструмент анализа и хранения данных, который все еще продолжает развиваться. В итоге, однако, технологии больших данных по-прежнему разрабатываются в качестве дополнения к существующим ИТ-архитектурам хранилищ данных типа клиент/сервер/суперкомпьютер, что в некоторых отношениях лучше, чем эти самые среды хранилищ данных, но не во всех.
В результате, в рамках настоящего варианта использования, связанного с обеспечением непрерывности деловой деятельности и восстановления после катастроф, необходимо задать правильные вопросы, такие как: Почему мы это делаем и чего мы пытаемся достичь? В чем мы зависим от «ручных» практик, и когда мы можем их использовать? Какие системы (как, например, телефонная связь) были и остаются переданными на аутсорсинг другим организациям, и каковы их функции в плане обеспечения непрерывности деловой деятельности (если есть)? Наконец, мы должны определить функции, которые можно упростить, и понять, какие профилактические меры, не требующие больших затрат, мы можем предпринять, такие как упрощение деловой практики.
Мы должны определить, какие деловые функции являются критически важными и требующими восстановления в соответствии с приоритетом в первую, вторую, третью очередь или в более позднее время; какова модель чрезвычайных ситуаций, с которыми мы намерены бороться; каковы типы наиболее вероятных чрезвычайных ситуаций — исходя из понимания того, что нам не нужно рассматривать все возможные виды катастроф. Если резервное копирование данных в облачной экосистеме является хорошим решением, это сократит время восстановления после сбоя и удовлетворит требования к RTO/RPO. Кроме того, должны быть понимание и поддержка усилий по обеспечению непрерывности деловой деятельности в организации, поскольку это не проблема одной лишь службы ИТ; это также проблема оказания деловых услуг, требующая тестирования Плана действий в случае чрезвычайных ситуаций посредством плановой пошаговой проработки и т. д.
Должна быть формальная методология разработки плана BC/DR, включающая:
1) заявление о политике (цель плана, обоснование и ресурсы и т. д. — каждый такой элемент следует определить),
2) анализ воздействия на деловую деятельность (как остановка повлияет на деловую деятельность в финансовом и в иных отношениях),
3) определение превентивных мер (можно ли избежать катастрофы, приняв разумные меры предосторожности),
4) стратегии восстановления (как и что нужно будет восстановить),
5) разработка плана (напишите план и реализуйте его элементы),
6) обеспечение поддержки плана в организации и его тестирование (важно, чтобы все знали план и знали, что делать в случае введения его в действие),
7) актуализация (регулярное внесение изменений, отражающих текущее состояние корпоративной среды)</td></tr>
<tr><td>Текущие решения</td><td>Вычислительная система</td><td>Облачные экосистемы, включающие предоставление инфраструктуры как сервиса (IaaS), поддерживаемые центрами обработки данных уровня Tier3 — защищенными, отказоустойчивыми в случае сбоев питания, отказов системы кондиционирования воздуха и т. д. Географически удаленные центры восстановления данных, обеспечивающие услуги репликации данных.

Примечание — Репликация отличается от резервного копирования тем, что воспроизводятся только те изменения, которые произошли после предыдущей репликации, включая изменения на уровне блоков. Репликация может быть выполнена быстро — в рамках пятисекундного «окна», при этом репликация данных может проводиться каждые четыре часа. Соответствующий «снимок» данных сохраняется в течение семи рабочих дней или дольше, если это необходимо. Реплицированные данные могут быть перемещены в запасной центр (т. е. в резервную систему) для удовлетворения требований организации в отношении заданной точка восстановления (recovery point objective, RPO) и заданного времени восстановления (recovery time objective, RTO)</td></tr>
</table>

Текущие решения	**Хранилище данных**	VMware, NetApps, Oracle, IBM, Brocade
	Сеть связи	Сети WAN, LAN, MAN, WiFi, доступ в Интернет, через публичные, частные, кооперативные и гибридные облачные среды, с VPN или без него
	Программное обеспечение	Hadoop, Map/Reduce, Open-source и/или проприетарные решения поставщиков, таких как AWS (Amazon Web Services), Google Cloud Services и Microsoft
Характеристики больших данных	**Источник данных (распределенный/централизованный)**	Как распределенные, так и централизованные источники данных, поступающих в среду высокой доступности с обеспечением восстановления после катастроф (HA/DR Environment) и в хостинговый виртуальный сервер (HVS), например, в следующих конфигурациях: DC1 — > VMWare/KVM (кластеры, с виртуальными брандмауэрами), Data link — VMWare Link — Vmotion Link — Network Link, несколько мостовых соединений с поставщиком (PB) в рамках NaaS (сеть как сервис), DC2 — > VMWare/KVM (кластеры с виртуальными брандмауэрами), DataLink (Vmware Link, Vmotion Link, Network Link), несколько мостовых соединений с поставщиком в рамках NaaS (требуется отказоустойчивая виртуализация)
	Объем (количество)	От нескольких терабайт до нескольких петабайт
	Скорость обработки (например, в реальном времени)	Центры обработки данных уровня Tier3 — защищенные, отказоустойчивые в случае сбоев питания, отказов системы кондиционирования воздуха и т. д. В данном случае инфраструктура как сервис (IaaS) предоставляется на основе NetApps. Репликация отличается от резервного копирования тем, что воспроизводятся только те изменения, которые произошли после предыдущей репликации, включая изменения на уровне блоков. Репликация может быть выполнена быстро — в рамках пятисекундного «окна», при этом репликация данных может проводиться каждые четыре часа. Соответствующий «снимок» данных сохраняется в течение семи рабочих дней или дольше, если это необходимо. Реплицированные данные могут быть перемещены в запасной центр для удовлетворения требований организации в отношении RPO/RTO
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Различные виртуальные среды, работающие в рамках архитектуры пакетной обработки или параллельной архитектуры с «горячей» заменой (hot-swappable parallel architecture)
	Вариативность (темпы изменения)	Капитальные затраты (CapEx) увеличиваются в зависимости от соглашений об уровне обслуживания (SLA), от требований RTO/RPO и от потребностей деловой деятельности
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Целостность данных играет критически важную роль на протяжении всего жизненного цикла организации ввиду законодательно-нормативных требований стратегического управления, управления рисками и соблюдения требований (GRC) и конфиденциальности, целостности и доступности (CIA)

Наука о больших данных (сбор, курирование, анализ, операции)	**Визуализация**	Целостность данных играет критически важную роль на протяжении всего жизненного цикла организации ввиду законодательно-нормативных требований стратегического управления, управления рисками и соблюдения требований (GRC) и конфиденциальности, целостности и доступности (CIA)
	Качество данных (синтаксис)	Целостность данных играет критически важную роль на протяжении всего жизненного цикла организации ввиду законодательно-нормативных требований стратегического управления, управления рисками и соблюдения требований (GRC) и конфиденциальности, целостности и доступности (CIA)
	Типы данных	Многочисленные типы данных и форматов включают (но не ограничиваются ими): плоские файлы, txt, .pdf, файлы приложений для Android, .wav, .jpg и VOIP (передача голоса с использованием протокола IP)
	Аналитика данных	Данные должны поддерживаться в формате, неподверженном деградации во время обработки и процедур поиска и анализа
Иные проблемы больших данных	Сложные операции, связанные с переключением с основного сайта на сайт репликации или на резервный сайт, в настоящее время еще не полностью автоматизированы. Цель заключается в том, чтобы дать пользователю возможность автоматически инициировать последовательность действий по переходу на резервную систему. Перемещение размещенных в облаке данных требует четко определенного и подвергающегося постоянному мониторингу управления конфигурацией сервера. Кроме того, обе организации должны знать, какие серверы должны быть восстановлены, и каковы зависимости и взаимозависимости между серверами основного сайта и серверами репликации и / или резервного сайта. С этой целью необходим постоянный мониторинг обоих сайтов, поскольку в этом процессе задействованы два решения, имеющие дело либо с серверами, на которых хранятся образы, либо с постоянно работающими «боевыми» серверами, как это имеет место в параллельных системах с «горячей» заменой (hot-swappable parallel architecture), — и все это требует точной и актуальной информации от клиента	
Проблемы пользовательского интерфейса и мобильного доступа	Обеспечение мобильного доступа — это постоянно растущий слой технической сложности, однако не все решения для обеспечения непрерывности деловой деятельности / восстановления после катастроф (BC/DR) носят технический характер. Есть две взаимосвязанные и взаимозависимые стороны, которые должны работать вместе над тем, чтобы найти работоспособное и жизнеспособное решение — это представители основной деловой деятельности и ИТ. Если обе эти стороны приходят к согласию, эти технические вопросы решаются в стратегии BC/DR, внедрение и поддержание которой обеспечивает вся организация. Один из вопросов, который не сводится к проблемам мобильности, касается фундаментальной проблемы, влияющей на большинство решений BC/DR. Если Ваши основные серверы (A, B, C) понимают X, Y, Z .... , но Ваши вторичные виртуальные серверы репликации / резерва (a, b, c) с течением времени не поддерживаются должным образом (не обеспечивается надлежащее управление конфигурацией) и происходит их рассинхронизация с основными серверами, так что они понимают только X и Y, — когда поступает команда на выполнение репликации или резервного копирования, то … «Хьюстон, у нас проблемы ...» Обратите внимание: с течением времени все системы могут и будут страдать от ползучей потери синхронизации — и некоторые больше, чем другие, если они полагаются на ручные процессы для обеспечения стабильности системы	

Технические проблемы обеспечения безопасности и защиты персональных данных	В зависимости от природы и требований отраслевых вертикалей, таких как финансовая деятельность, страхование и медико-биологические науки (Life Sciences), охватывающих как государственные, так и частные учреждения и организации; и от ограничений, налагаемых законодательно-нормативными требованиями стратегического управления, управления рисками и соблюдения требований (GRC) и конфиденциальности
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	В число проблем обеспечения непрерывности деловой деятельности / восстановления после катастроф (BC/DR) входят следующие: **1) Признание** a) Видение менеджмента BC/DR b) Предполагается, что данная проблема является проблемой ИТ, когда на само деле это не так **2) Люди** a) Степень укомплектованности персоналом — многие малые и средние предприятия недоукомплектованы ИТ-персоналом в сравнении с их текущей рабочей нагрузкой b) Видение (руководствуясь подходом сверху — вниз) — Способны ли деловые и ИТ-подразделения увидеть проблему в целом и выработать стратегию типа «Списка вызовов» для использования в случае чрезвычайной ситуации? c) Навыки — Есть ли специалисты, способные спроектировать, внедрить и протестировать BC/DR — решение? d) Время — Есть ли время у специалистов, и есть ли в деловой деятельности «окно» времени для создания и тестирования DR/BC решения, поскольку подобное решение является дополнительным проектом, на который требуется время и ресурсы? **3) Деньги** Затраты можно перевести в категорию операционных расходов (OpEx), а не в капитальные затраты (CapEx), варьируя требования RPO/RTO a) Капитал всегда является ограниченным ресурсом b) Решения BC должны начинаться с вопросов «В чем риск?» и «Как затраты ограничивают решение?» **4) Нарушение привычного порядка** Встроить BC/DR в стандартную «облачную» инфраструктуру (IaaS) малых и средних предприятий a) Планирование BC/DR «съедает» деловые ресурсы b) Тестирование BC также нарушает обычный ход деловой деятельности
Дополнительная информация (гиперссылки)	1. Сайт независимой консультационной организации «Восстановление после катастроф» (DisasterRecovery.org), https://www.disasterrecovery.org/ 2. Вебинар компании dinCloud «Как пережить катастрофы, используя облако» (Surviving Disasters by Leveraging the Cloud), https://www.dincloud.com/webinars/surviving-disasters-by-leveraging-the-cloud 3. Комитет спонсорских организаций (COSO), см. https://www.coso.org/ 4. Сайт ITIL (Библиотека инфраструктуры информационных технологий), см. https://www.axelos.com/best-practice-solutions/itil 5. Стандарт COBIT® 2019 CobiT (от Control Objectives for Information and Related Technology — «Цели управления информационными и смежными технологиями) на сайте Международной ассоциации аудита и контроля информационных систем (ISACA), см. https://www.isaca.org/resources/cobit 6. Концепция архитектуры «Открытой группы» (The Open Group Architecture Framework, TOGAF) версии 9.2, http://www.opengroup.org/togaf/ 7. Международный стандарт ИСО/МЭК 27000:2018 «Информационная технология. Методы и средства обеспечения безопасности. Системы менеджмента информационной безопасности. Общий обзор и терминология» (Information technology — Security techniques — Information security management systems — Overview and vocabulary), https://www.iso.org/standard/73906.html, свободно доступен по адресу https://standards.iso.org/ittf/PubliclyAvailableStandards/c073906_ISO_IEC_27000_2018_E.zip. В России стандарт адаптирован (в более ранней редакции) как ГОСТ Р ИСО/МЭК 27000—2012, см. http://protect.gost.ru/v.aspx?control=8&baseC=6&id=175549 8. Некоммерческая организация по надзору за отчетностью публичных компаний, США (PCAOB), https://pcaobus.org/

А.2.6 Вариант использования № 10: Грузоперевозки

Название	Грузоперевозки	
Предметная область	Отрасль грузоперевозок	
Автор/организация/ эл.почта	Уильям Миллер (William Miller)/компания MaCT USA/mact-usa@att.net	
Акторы/ заинтересованные лица, их роли и ответственность	Конечные пользователи (отправители/получатели). Лица, обслуживающие транспортные средства (грузовик/корабль/самолет). Операторы связи (сотовая связь/спутниковая связь). Грузоотправители (отправка и получение)	
Цели	Хранение и анализ объектов («вещей») в процессе перевозки	
Описание варианта использования	В настоящем варианте использования дается общее представление о приложения «больших данных» для отрасли грузоперевозок, в которой работают такие компании, как FedEx, UPS, DHL и т. д. Отрасль грузоперевозок, вероятно, является самым крупным из широко распространенных сегодня потенциальных вариантов использования больших данных. Он охватывает идентификацию, транспортировку и обработку грузов («вещей») в цепочке поставок. Идентификация груза начинается с отправителя, и используется получателями и всеми стоящими между ними посредниками, которым необходимо знать место и время прибытия транспортируемых грузов. Новым аспектом станут сведения о статусе и состоянии объекта, включая информацию с датчиков и получаемые от глобальной системы позиционирования (GPS) координаты, а также уникальная схема идентификации, основанная на международном стандарте ИСО/МЭК 29161:2016 «Информационные технологии. Структура данных. Уникальная идентификация для Интернета вещей», разработанном подкомитетом SC31 Объединенного технического комитета ИСО/МЭК СТК1. Данные обновляются в масштабе времени, близком к реальному, когда грузовик прибывает на склад или при доставке товара получателю. Промежуточные состояния в настоящее время неизвестны; данные о местоположении в реальном времени не обновляются; а товары, утерянные на складе или во время транспортировки, могут представлять собой потенциальную проблему для безопасности страны. Сведения хранятся в архиве и остаются доступными в течение xx дней	
Текущие решения	**Вычислительная система**	Неизвестно
	Хранилище данных	Неизвестно
	Сеть связи	LAN/T1/веб-страницы интернета
	Программное обеспечение	Неизвестно
Характеристики больших данных	**Источник данных (распределенный/ централизованный)**	В настоящее время — централизованный
	Объем (количество)	Большой
	Скорость обработки (например, в реальном времени)	В настоящее время система в реальном масштабе времени не работает
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Данные обновляются, когда водитель прибывает на склад и загружает время и дату принятия груза. Это в настоящее время осуществляется не в режиме реального времени
	Вариативность (темпы изменения)	Сейчас информация обновляется только после сканирования объектов с помощью сканера штрих-кода, который отправляет данные на центральный сервер. В настоящее время местоположение объекта в реальном времени не отображается

Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	
	Визуализация	Нет
	Качество данных (синтаксис)	Да
	Типы данных	Нет сведений
	Аналитика данных	Да
Иные проблемы больших данных	Обеспечение более быстрой оценки идентичности, местоположения и состояния грузов, предоставление подробной аналитики и локализация проблем в системе в режиме реального времени	
Проблемы пользовательского интерфейса и мобильного доступа	В настоящее время мониторинг условий на борту грузовиков, кораблей и самолетов не осуществляется	
Технические проблемы обеспечения безопасности и защиты персональных данных	Безопасность должна быть более надежной	
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	В данном варианте использования применяются локальные базы данных, а также существует требование синхронизации с центральным сервером. Эти операции в итоге будут распространены на мобильные устройства и бортовые системы, способные отслеживать местонахождение грузов и обеспечивать обновление информации в режиме реального времени, включая передачу сведений об условиях, протоколирование событий и рассылку оповещений лицам, которым соответствующая информация необходима	
Дополнительная информация (гиперссылки)		

А.2.7 Вариант использования № 11: Данные о материалах

Название	Данные о материалах
Предметная область	Производство, исследования в области материаловедения
Автор/организация/ эл.почта	Джон Рамбл (John Rumble) / компания R&R Data Services / jumbleusa@earthlink.net
Акторы/ заинтересованные лица, их роли и ответственность	Разработчики продуктов (вводят данные о материалах в системы автоматизированного проектирования). Исследователи свойств материалов (производят данные о материалах; в некоторых случаях являются пользователями таких данных). Испытатели материалов (производят данные о материалах; разработчики стандартов). Распространители данных (поставщики доступа к данным материалам, часто на коммерческой основе)
Цели	Улучшить доступность, качество и удобство использования данных о материалах, а также преодолеть проприетарные барьеры для обмена такими данными. Создать достаточно крупные хранилища данных о материалах, способствующие поиску и раскрытию этой информации

Описание варианта использования	Каждый физический продукт изготовлен из материалов, которые были выбраны исходя из их свойств, стоимости и доступности. Каждый год принимаются связанные с выбором материалов решения на общие суммы, исчисляемые сотнями миллиардов долларов. Помимо того, как столь убедительно показала инициатива «Геном материала» (Materials Genome Initiative), внедрение новых материалов обычно занимает два-три десятилетия, а не несколько лет, отчасти из-за того, что сведения о новых материалах не являются легкодоступными. Все действующие лица в рамках жизненного цикла материалов сегодня имеют доступ к очень ограниченным объемам данных о материалах, что приводит к принятию неоптимальных, неэффективных и затратных решений, связанных с материалами. В то время, как в рамках инициативы «Геном материала» рассматривается один важный существенный аспект проблемы, а именно, базовые данные о материалах, необходимые для компьютерного проектирования и испытания материалов, — вопросы, связанные с физическими измерениями на физических материалах (от базовых структурных и термических свойств до сложных эксплуатационных свойств и свойства новых наноразмерных материалов) не рассматриваются систематически, широко (междисциплинарно и на международном уровне) или же эффективно (практически отсутствуют встречи по тематике данных о материалах, группы по разработке стандартов и целевые финансируемые программы). Одной из наиболее сложных проблем, которые способны решить методы «больших данных», является предсказание поведения и характеристик реальных материалов (в количествах от грамма до тонны), начиная с описаний на атомном, нано- и/или микрометровом уровнях. По перечисленным выше причинам решения об использовании материалов в настоящее время излишне консервативны, часто основываясь на более старых, а не на последних данных соответствующих исследований и разработок, и не используют достижения в области построения моделей и моделирования. Информатика материалов (materials informatics) — это та область, в которой новые инструменты науки о данных могут оказать существенное влияние	
Текущие решения	**Вычислительная система**	Нет
	Хранилище данных	Широко рассеянное, существует множество препятствий для доступа
	Сеть связи	Практически отсутствует
	Программное обеспечение	Узкие подходы в рамках национальных программ (Япония, Южная Корея и Китай), прикладных программ (ядерная программа Евросоюза); проприетарные решения (Granta, и др.)
Характеристики больших данных	**Источник данных (распределенный/централизованный)**	Чрезвычайно сильно распределенный, при этом хранилища данных обеспечивают хранение сведений лишь об очень немногих базовых свойствах
	Объем (количество)	Согласно оценке, сделанной в 1980-х годах, за последние пятьдесят лет появилось более 500 тыс. коммерческих материалов. В последние три десятилетия этот показатель значительно вырос
	Скорость обработки (например, в реальном времени)	С течением времени растет количество материалов, спроектированных с использованием компьютерных средств и разработанных теоретически (примером являются наноматериалы)

Характеристики больших данных	**Разнообразие (множество наборов данных, комбинация данных из различных источников)**	Много наборов данных при практическом отсутствии стандартов, поддерживающих комбинирование этих данных
	Вариативность (темпы изменения)	Материалы постоянно изменяются, и постоянно создаются новые данные, описывающие новые материалы
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Для точного описания более сложных свойств материалов может потребоваться множество (сотни?) независимых переменных. В настоящее время не предпринимается практически никаких усилий, направленных на выявление этих переменных и систематизацию сбора их значений с целью создания надежных наборов данных
	Визуализация	Важна для отыскания подходящих материалов. Потенциально важна для понимания зависимости свойств материалов от множества независимых переменных. Практически остается без внимания
	Качество данных (синтаксис)	За исключением базовых данных о структурных и тепловых свойствах качество данных является низким или непонятным. См. написанное Мунро (Munro) Руководство NIST по рекомендуемой практике
	Типы данных	Числовая информация, графики, графические образы
	Аналитика данных	Эмпирическая и узкая по сфере охвата
Иные проблемы больших данных	1) Создание хранилищ данных о материалах, помимо существующих, которые ориентированы на хранение лишь базовых данных. 2) Разработка международных стандартов регистрации данных, которые могут использоваться очень многообразным сообществом специалистов по материалам, включающим разработчиков стандартов испытаний материалов (таких, как ассоциация ASTM International и Международная организация по стандартизации ИСО), занимающиеся испытаниями материалов компании, производителей материалов, а также научно-исследовательские и опытно-конструкторские лаборатории. 3) Разработка инструментов и процедур, помогающих организациям, которым требуется депонировать в хранилищах данных сведения о проприетарных материалах, маскировать проприетарную информацию, сохраняя при этом пригодность данных к использованию. 4) Разработка многопараметрических инструментов визуализации данных о материалах, способных работать с достаточно большим количеством переменных	
Проблемы пользовательского интерфейса и мобильного доступа	В настоящее время не являются существенными	
Технические проблемы обеспечения безопасности и защиты персональных данных	По своей природе многие проприетарные данные являются весьма конфиденциальными и «чувствительными»	

Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Разработка стандартов; создание крупномасштабных хранилищ данных; привлечение отраслевых пользователей; интеграция с системами автоматизированного проектирования (не стоит недооценивать сложность этой работы — специалисты в области материаловедения обычно не столь хорошо разбираются в компьютерах, как химики, специалисты по биоинформатике и инженеры)
Дополнительная информация (гиперссылки)	

А.2.8 Вариант использования № 12: Геномика материалов на основе результатов моделирования

Название	Геномика материалов на основе результатов моделирования	
Предметная область	Научные исследования, материаловедение	
Автор/организация/ эл.почта	Дэвид Скиннер (David Skinner) / Национальная лаборатория имени Лоуренса в Беркли (LBNL), deskinner@lbl.gov	
Акторы/ заинтересованные лица, их роли и ответственность	**Поставщики ресурсов** В обязанности Национальных лабораторий и энергетических центров входит предоставление расширенных возможностей для работ по геномике материалов, с использованием в качестве инструментов вычислений и данных. **Сообщество пользователей** Министерство энергетики США, отраслевые и академические исследователи являются сообществом пользователей, ищущих ресурсы и возможности для быстрых инноваций в материалах	
Цели	Ускорение разработки материалов с улучшенными свойствами с помощью проектов моделирования, управление которыми осуществляется с использованием искусственного интеллекта	
Описание варианта использования	Осуществление инноваций в технологиях электрических батарей и аккумуляторов посредством масштабных проектов моделирования, охватывающих большое количество возможных проектных решений. Систематические вычислительные исследования с целью поиска возможностей для инноваций в фотовольтаике (фотоэлектрических технологиях). Рациональное проектирование материалов на основе поиска и моделирования	
Текущие решения	**Вычислительная система**	Суперкомпьютер Cray XE6 «Hopper» (150 тысяч процессоров); аппаратные ресурсы для аналитики данных аналогичные тем, что используются «омиками» (omics — направлениями биологической науки, такими как геномика, протеомика, метаболомика и др.)
	Хранилище данных	GPFS, MongoDB
	Сеть связи	10 гигабит/с
	Программное обеспечение	PyMatGen, FireWorks, VASP, ABINIT, NWChem, BerkeleyGW; различное ПО, разработанное сообществом

Характеристики больших данных	**Источник данных (распределенный/ централизованный)**	Потоки данных поступают от проектов моделирования, выполняемых на централизованных пета/ экзафлопсных вычислительных системах. Сильно распределенная сеть потоков данных от центрального шлюза до пользователей
	Объем (количество)	100 терабайт (текущий), 500 терабайт через 5 лет. Требуются масштабируемые базы данных для данных типа «ключ-значение» и для библиотек объектов
	Скорость обработки (например, в реальном времени)	Высокопроизводительные вычисления (HTC), детальное управление задачами и очередями. Быстрый старт/остановка для группы задач. Анализ данных в режиме реального времени для оперативного реагирования
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Комбинирование результатов моделирования от разных программ и на различном теоретическом уровне. Форматирование, регистрация и интеграция наборов данных. Комбинирование данных, полученных при различных масштабах моделирования
	Вариативность (темпы изменения)	Цели при проектировании материалов будут в большей степени поисковыми и ориентированными на потребности потребителей. Вычислительная база должна гибко адаптироваться к новым целям
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Проверка и оценка неопределенностей результатов моделирования путем сопоставления с экспериментальными данными различного качества. Проверка на наличие ошибок и оценка границ путем сопоставления разных результатов моделирования
	Визуализация	Использование программ просмотра данных о материалах ввиду роста объемов, выдаваемых в ходе поиска данных. Визуальное проектирование материалов
	Качество данных (синтаксис)	Количественная оценка неопределенности в результатах на основе нескольких наборов данных. Распространение ошибок в системах знаний
	Типы данных	Пары ключ-значение, JSON, файловые форматы данных о материалах
	Аналитика данных	Технологии Map/Reduce и поиска, позволяющие комбинировать данные моделирования и экспериментальные данные
Иные проблемы больших данных	Масштабное применение высокопроизводительных вычислений для выполнения проектов моделирования. Гибкие методы обработки данных в масштабе для неупорядоченных данных. Системы машинного обучения и управления знаниями, объединяющие данные из публикаций, результаты экспериментов и моделирования для развитие направленного на результат мышления при проектировании материалов	
Проблемы пользовательского интерфейса и мобильного доступа	Существует потенциал для широкого распространения практически применимых знаний в области материаловедения. Многие программные приложения геномики материалов могут быть перенесены на мобильную платформу	

Технические проблемы обеспечения безопасности и защиты персональных данных	Возможность работать в изолированной зоне — «песочнице» или же создавать независимые рабочие зоны для заинтересованных в данных сторонах. Объединение наборов данных на основе политик
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Проект Управления администрации президента США по вопросам науки и технической политики (Office of Science and Technology Policy, OSTP) по достижению более масштабных целей в области геномики материалов был опубликован в мае 2013 г.
Дополнительная информация (гиперссылки)	Сайт поддерживаемого Министерством энергетики США проекта «Материалы» (The Materials Project), https://www.materialsproject.org/

А.3 Оборона

А.3.1 Вариант использования № 13: Облачный крупномасштабный анализ и визуализация геопространственных данных

Название	Облачный крупномасштабный анализ и визуализация геопространственных данных	
Предметная область	Оборона (но вариант также применим во многих других областях)	
Автор/организация/ эл.почта	Дэвид Бойд (David Boyd) / компания Data Tactics / dboyd@data-tactics.com	
Акторы/ заинтересованные лица, их роли и ответственность	Аналитики геопространственных данных Принимающие решения лица Лица, определяющие политику	
Цели	Поддержка крупномасштабного анализа и визуализации геопространственных данных	
Описание варианта использования	По мере того, как увеличивается количество датчиков и источников данных с географической привязкой, объемы требующих сложного анализа и визуализации геопространственных данных увеличиваются в геометрической прогрессии. Традиционные географические информационные (геоинформационные) системы (ГИС) обычно способны анализировать миллионы и легко визуализируют тысячи объектов. Современные интеллектуальные системы часто содержат триллионы геопространственных объектов и должны быть способны визуализировать и взаимодействовать с миллионами объектов	
Текущие решения	**Вычислительная система**	Системы вычислений и хранения — от ноутбуков до больших серверов (см. примечание о кластерах). Системы визуализации — от карманных устройств до ноутбуков
	Хранилище данных	Системы вычислений и хранения — локальный жесткий диск или сеть хранения данных (SAN). Системы визуализации — локальный жесткий диск, оперативная флеш-память
	Сеть связи	Системы вычислений и хранения — гигабитное или более скоростное сетевое соединение по локальной сети. Системы визуализации — гигабитные беспроводные соединения, беспроводная связь включая WiFi (802.11), сотовую связь (3G/4G) и радиорелейную связь

Текущие решения	**Программное обеспечение**	Системы вычислений и хранения — обычно Linux или Windows Server с реляционной СУБД с геопространственной поддержкой; геопространственный сервер/программное обеспечение для анализа — ESRI ArcServer, Geoserver. Системы визуализации — Windows, Android, iOS — браузерная визуализация. На некоторых ноутбуках может быть установлена локальная версия ArcMap
Характеристики больших данных	**Источник данных (распределенный/централизованный)**	Сильно распределенный
	Объем (количество)	Растровая графика — сотни терабайт; векторные данные — десятки гигабайт, но при этом миллиарды точек
	Скорость обработки (например, в реальном времени)	Некоторые датчики передают векторные данные в масштабе времени, близком к реальному. Визуализация изменений должна быть в масштабе времени, близком к реальному
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Растровые изображения (различные форматы: NITF, GeoTiff, CADRG). Векторная графика (различные форматы: формат Shapefile, язык разметки Keyhole (Keyhole Markup Language, KML) и текстовые потоки. Типы объектов включают точки, линии, области, ломаные линии (polylines), окружности и эллипсы)
	Вариативность (темпы изменения)	От умеренной до высокой
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Точность данных имеет критически важное значение и обычно контролируется на основе трех факторов: 1) точность датчика — является большой проблемой, 2) датум (система геодезических координат)/сфероид (двухосного эллипсоид), 3) точность регистрации изображений
	Визуализация	Отображение осмысленным образом больших наборов данных (миллионы точек) на небольших устройствах (карманных устройствах), являющихся оконечными точками сетей с низкой пропускной способностью
	Качество данных (синтаксис)	Типичной проблемой является визуализация в том случае, когда отсутствуют сведения о качестве/точности первичных данных. Все данные должны включать метаданные, указывающие точность или круговое вероятное отклонение
	Типы данных	Растровые изображения (различные форматы: NITF, GeoTiff, CADRG). Векторная графика (различные форматы: формат Shapefile, язык разметки Keyhole (Keyhole Markup Language, KML) и текстовые потоки. Типы объектов включают точки, линии, области, ломаные линии (polylines), окружности и эллипсы)

Наука о больших данных (сбор, курирование, анализ, операции)	**Аналитика данных**	Ближайшая точка подхода, отклонение от маршрута, плотность точек во времени, метод главных компонентов (principal component analysis, PCA) и метод анализа независимых компонентов (independent component analysis, ICA)
Иные проблемы больших данных	Индексация, поиск/извлечение и распределенный анализ. Формирование и передача визуализации	
Проблемы пользовательского интерфейса и мобильного доступа	Визуализация данных на устройствах, являющихся оконечными точками беспроводных сетей с низкой пропускной способностью	
Технические проблемы обеспечения безопасности и защиты персональных данных	Данные являются чувствительными, и должна быть обеспечена их полная безопасность при передаче и при хранении (особенно на портативных/карманных устройствах)	
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Геопространственные данные требуют уникальных подходов к индексации и проведению распределенного анализа	
Дополнительная информация (гиперссылки)	Применимые стандарты: - стандарты «Открытого геопространственного консорциума» (Open Geospatial Consortium, OGC), https://www.ogc.org/standards - спецификации формата GeoJSON, https://geojson.org/ - спецификации формата Compressed ARC Digitized Raster Graphics (CADRG), https://earth-info.nga.mil/publications/specs/printed/CADRG/cadrg.html Индексирование геопространственных данных: Quad-деревья; заполняющие пространство кривые (кривые Гильберта) — многочисленные источники можно найти в Интернете	
Примечание — В Министерстве обороны США проводилась определенная работа, связанная с этим набором проблем. В частности, стандартное облако (DSC, DCGS-A Standard Cloud) для унифицированной армейской наземной станция с распределенными терминалами (DCGS-A, Distributed Common Ground System — Army) хранит, индексирует и анализирует некоторые источники больших данных. Однако все еще остается много проблем с визуализацией.		

А.3.2 Вариант использования № 14: Идентификация и отслеживание объектов — Постоянное наблюдение

Название	Идентификация и отслеживание объектов по данным WALF-формат видео с высоким разрешением (WALF) или FMV-формат высококачественного видео — Постоянное наблюдение
Предметная область	Оборона (разведка)
Автор/организация/ эл.почта	Дэвид Бойд (David Boyd) / компания Data Tactics / dboyd@data-tactics.com
Акторы/ заинтересованные лица, их роли и ответственность	1) Гражданские и военные лица, принимающие решения. 2) Специалисты по анализу разведданных. 3) Участники боевых действий
Цели	Способность обрабатывать первичные фото/видеоданные и выделять из них/ отслеживать во времени объекты (транспортные средства, люди, грузы). В частности, идея заключается в том, чтобы редуцировать петабайты собранных в ходе непрерывного наблюдения данных к разумному размеру (например, векторным путям)

Описание варианта использования	Датчики постоянного наблюдения легко могут за считанные часы собирать петабайты фото- и видеоданных. Человек не способен обработать такие объемы данных в целях предупреждения о событиях или отслеживания. Обработка данных должна осуществляться рядом с датчиком, который, вероятно, развернут на передовой, поскольку объемы данных слишком велики для того, чтобы их можно было легко передать. Данные должны быть редуцированы к набору геопространственных объектов (например, точек, путей), которые можно легко интегрировать с другими данными для формирования общей оперативной картины	
Текущие решения	**Вычислительная система**	Различные, варьируются от простых устройств хранения, соединенных с датчиком, и простых средств отображения и хранения до систем, поддерживающих ограниченное выделение объектов. Типичные системы выделения объектов в настоящее время представляют собой небольшие (от 1 до 20 узлов) кластеры расширенных за счет использования графических процессоров (GPU) компьютерных систем
	Хранилище данных	В настоящее время — плоские файлы, хранимые в большинстве случаев на жестком диске. Иногда индексы реляционных СУБД указывают на файлы или части файлов на основе метаданных / данных телеметрии
	Сеть связи	Обмен информацией с датчиками, как правило, осуществляется или в пределах прямой видимости, или с использованием спутниковой связи
	Программное обеспечение	Широкий спектр специализированного программного обеспечения и инструментов, включая, в том числе, традиционные реляционные СУБД и средства отображения
Характеристики больших данных	**Источник данных (распределенный/ централизованный)**	В число передающих фото/видеоданные датчиков входят стационарные и установленные на летательных аппаратах оптические и инфракрасные датчики, а также радары с синтезированной апертурой (SAR)
	Объем (количество)	FMV — от 30 до 60 кадров в секунду при полноцветном разрешении 1080 пикселей WALF — от 1 до 10 кадров в секунду при полноцветном разрешении 10 тысяч на 10 тысяч пикселей
	Скорость обработки (например, в реальном времени)	В реальном времени
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Данные обычно представлены в одном или нескольких стандартных форматах для графических изображений или видео
	Вариативность (темпы изменения)	Небольшая

Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Достоверность извлеченных объектов имеет жизненно важное значение. Если система дает сбой или генерирует ложные срабатывания, подвергаются риску жизни людей
	Визуализация	Извлеченные результаты обычно визуализируются путем наложения на отображение геопространственных данных. Наложенные объекты должны отсылать к соответствующему сегменту исходного изображения/видеопотока
	Качество данных (синтаксис)	Качество данных, как правило, определяется сочетанием характеристик датчиков и погодных условий (маскирующим фактором является пыль/влажность, а фактором стабильности — ветер)
	Типы данных	Исходные данные представлены в стандартных форматах для графических изображений и видео. Выходные данные должны быть в форме веб-функций, соответствующих стандартам «Открытого геопространственного консорциума» (Open Geospatial Consortium, OGC), либо в виде стандартных геопространственных файлов [Shapefile, язык разметки Keyhole (Keyhole Markup Language, KML)]
	Аналитика данных	1) Идентификация объекта (тип, размер, цвет) и его отслеживание. 2) Анализ закономерностей поведения объекта (проходил ли сегодня днем грузовик, который ездит каждую среду после полудня, по иному маршруту; есть ли стандартный маршрут, которому каждый день следует конкретный человек). 3) Групповое поведение/динамика (есть ли небольшая группа, пытающаяся спровоцировать бунт; выделяется ли данный человек в толпе, ведет ли он себя не так, как все?) 4) Хозяйственная деятельность: a) Есть ли очередь в хлебном магазине, мясной лавке или за мороженым? b) Больше ли грузовиков движется с товарами на север, чем на юг? c) Увеличилась или уменьшилась на данном рынке активность лавок и/или их размер за последний год? 5) Объединение (слияние) данных
Иные проблемы больших данных	Обработка больших объемов данных почти в режиме реального времени (NRT) для поддержки оповещения о событиях и осведомленности о ситуации	
Проблемы пользовательского интерфейса и мобильного доступа	Доставка данных с мобильного датчика на обработку	

Технические проблемы обеспечения безопасности и защиты персональных данных	Значительные — нельзя допустить компрометацию источников данных и методов их обработки; враг не должен знать, что именно мы видим
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Как правило, данный тип обработки хорошо вписывается в массово-параллельные вычисления, поддерживаемые, например, графическими процессорами. Типичной проблемой является интеграция этой обработки в более крупный кластер, способный параллельно обрабатывать данные от нескольких датчиков и в масштабе времени, близком к реальному. Передача данных с датчика в систему также является большой проблемой
Дополнительная информация (гиперссылки)	Стандарты по вопросам управления видеоматериалами: Страница Бюро стандартов управления видеоматериалами (Motion Imagery Standards Board, MISB) на сайте Национального агентства геопространственной разведки США (National Geospatial — Intelligence Agency, NGA), https://gwg.nga.mil/misb/index.html Некоторые из многочисленных статей по теме выделения/отслеживания объектов: Erik Blasch, Haibin Ling, Yi Wu, Guna Seetharaman, Mike Talbert, Li Bai, Genshe Chen “Dismount Tracking and Identification from Electro-Optical Imagery”, http://www.dabi.temple.edu/~hbling/publication/SPIE12_Dismount_Formatted_v2_BW.pdf Fang-Hsuan Cheng, Yu-Liang Chen “Real time multiple objects tracking and identification based on discrete wavelet transform”, https://www.sciencedirect.com/science/article/abs/pii/S0031320305004863 Статьи о потребностях общего характера: John Keller “Persistent surveillance relies on extracting relevant data points and connecting the dots”, 2012, https://www.militaryaerospace.com/computers/article/16719589/persistent-surveillance-relies-on-extracting-relevant-data-points-and-connecting-the-dots “Wide Area Persistent Surveillance Revolutionizes Tactical ISR”, by Lexington institute, 2012, https://www.defencetalk.com/wide-area-persistent-surveillance-revolutionizes-tactical-isr-45745/

А.3.3 Вариант использования № 15: Обработка и анализ разведывательных данных

Название	Обработка и анализ разведывательных данных
Предметная область	Оборона (разведка)
Автор/организация/эл.почта	Дэвид Бойд (David Boyd) / компания Data Tactics / dboyd@data- tactics.com
Акторы/заинтересованные лица, их роли и ответственность	Высшее гражданское и военное руководство. Полевые командиры. Специалисты по анализу разведданных. Участники боевых действий
Цели	1) Выдача автоматических оповещений аналитикам, участникам боевых действий, командирам и высшему руководству на основе поступающих разведданных. 2) Предоставление аналитикам разведанных возможностей для выявления по этим данным: a) взаимосвязей между объектами (например, людьми, организациями, местами, оборудованием), b) тенденции в настроениях или намерениях как населения в целом, так и групп лидеров, таких как государственные деятели и представители негосударственных структур, c) определить место и, по возможности, время проведения враждебных действий, включая установку самодельных взрывных устройств, d) отслеживать местоположение и действия (потенциально) враждебных действующих лиц.

Цели	3) Способность осмысливать и извлекать знания из многообразных, разрозненных и часто неструктурированных (например, текстовых) источников данных. 4) Способность обрабатывать данные вблизи точки сбора и обеспечивать легкий обмен данными с/между отдельными солдатами, подразделениями, отрядами передового базирования и высшим руководством гарнизонов	
Описание варианта использования	1) Ввод/прием данные от широкого спектра датчиков и источников, принадлежащих к различным направлениям разведывательной деятельности, таким, как сбор и анализ изображений, полученных фотографической, оптико — электронной или радиолокационной аппаратурой (imagery intelligence, IMINT), разведка физических полей (measurement and signatures intelligence, MASINT), геопространственная разведка (geospatial intelligence, GEOINT), сбор информации людьми и от людей (human intelligence, HUMINT), радиоэлектронная разведка (signals intelligence, SIGINT), разведка на основе открытых источников (open source intelligence, OSINT) и т. д. 2) Обработка, преобразование или согласование данных из различных источников в разных форматах в единое пространство данных с целью поддержки: a) поиска; b) осмысления; c) сопоставления. 3) Оповещение пользователей о существенных изменениях в состоянии контролируемых объектов или о существенной активности в определенной области. 4) Обеспечение связи с периферией для участников боевых действий (в этом случае понятие периферии будет охватывать даже отдельного солдата в пешем патруле)	
Текущие решения	**Вычислительная система**	Стационарные и мобильные вычислительные кластеры с количеством узлов в диапазоне от 10 до 1000
	Хранилище данных	От десятков терабайт до сотен петабайт в случае периферийных и стационарных кластеров. У пехотинцев, как правило, имеется от одного до сотен гигабайт данных (обычно на портативном/карманном устройстве с объемом памяти менее 10 гигабайт)
	Сеть связи	Сеть связи внутри и между стационарными гарнизонами является надежной. Связь с передним краем ограничена и часто отличается большими задержками и потерей пакетов. Дистанционная связь может быть спутниковой (с большой задержкой) или даже ограничена радиосвязью на линии прямой видимости
	Программное обеспечение	Основными в настоящее время являются: 1) Hadoop 2) Accumulo (с системой хранения данных BigTable) 3) Solr 4) NLP (несколько вариантов) 5) Puppet (управление жизненным циклом ИТ, обеспечение безопасности) 6) Storm 7) Специализированные приложения и инструменты визуализации

Характеристики больших данных	**Источник данных (распределенный/централизованный)**	Сильно распределенный
	Объем (количество)	Некоторые передающие графические изображения / видео (IMINT) датчики способны генерировать более петабайта данных в течение нескольких часов. Другие данные столь же малы, как результаты нечастых срабатываний датчиков или текстовые сообщения
	Скорость обработки (например, в реальном времени)	Большая часть данных с датчиков поступает в реальном времени (полнокадровое видео, данные радиоэлектронной разведки), остальные — в режиме «менее реального» времени. Критически важным аспектом является возможность принимать, обрабатывать и распространять оповещения в масштабе времени, близком к реальному
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Все, что угодно, включая текстовые файлы, первичные данные с датчиков (raw media), графические образы, видео, аудио, электронные данные и данные, созданные человеком
	Вариативность (темпы изменения)	Хотя форматы интерфейсов с датчиками имеют тенденцию быть стабильными, большинство других данных не контролируется, и они могут быть в любом формате. Большая часть данных не структурирована
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Происхождение данных (включая, например, отслеживание всех передач и преобразований) должно контролироваться в течение жизненного цикла данных. Установление достоверности «мягких» источников данных (как правило, данных, созданных человеком) является критически важным требованием
	Визуализация	Основными видами визуализации будут наложения на геопространственную картину и сетевые графики (network diagrams). Данные могут включать миллионы точек на карте и тысячи узлов на сетевом графике
	Качество данных (синтаксис)	Качество генерируемых датчиком обычно известное (качество изображения, соотношение сигнал/шум) и хорошее. Качество неструктурированных или «захваченных» данных существенно варьируется и зачастую не поддается контролю

Наука о больших данных (сбор, курирование, анализ, операции)	**Типы данных**	Графические изображения, видео, текст, электронные документы всех типов, аудио, цифровые сигналы
	Аналитика данных	1) Оповещения в масштабе времени, близком к реальному, основанные на закономерностях и изменениях основных параметров, 2) Анализ взаимосвязей, 3) Геопространственный анализ, 4) Аналитика текстов (определение настроений, выделение сущностей и т. д.)
Иные проблемы больших данных	1) Передача больших данных (или даже данных умеренного размера) по тактическим сетям. 2) Данные, которые в настоящее время существуют в разрозненных хранилищах, должны быть доступны через семантически интегрированное пространство данных. 3) Большинство ключевых по важности данных либо являются неструктурированными, либо хранятся в виде графических образов или видеоматериалов, что требует значительной обработки для выделения объектов и извлечения информации	
Проблемы пользовательского интерфейса и мобильного доступа	Результаты этого анализа и информация должны передаваться или быть доступными для пехотинцев передовых отрядов	
Технические проблемы обеспечения безопасности и защиты персональных данных	Имеют первостепенную важность. Данные должны быть защищены от: 1) несанкционированного доступа или раскрытия, 2) несанкционированного вмешательства	
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Широкий спектр типов, источников, структур данных различного качества будет охватывать рад предметных областей и требует интегрированного поиска и анализа	
Дополнительная информация (гиперссылки)	Чарльз Уэллс (Col. Charles A. Wells) «Обзор программы унифицированной армейской наземной станция с распределенными терминалами» (DCGS-A, Distributed Common Ground System — Army Program Overview), 2012, http://aberdeen.afceachapter.org/files/presentations/AFCEAAberdeen_DCGSA_COLWells_PS.pdf Barry Smith, Tatiana Malyuta, William S. Mandrick, Chia Fu, Kesny Parent, Milan Patel «Horizontal Integration of Warfighter Intelligence Data- A Shared Semantic Resource for the Intelligence Community», 2012, http://stids.c4i.gmu.edu/papers/STIDSPapers/STIDS2012_T14_SmithEtAl_HorizontalIntegrationOfWarfighterIntel.pdf Salmen David, Malyuta Tatiana, Hansen Alan, Cronen Shaun, Smith Barry «Integration of Intelligence Data through Semantic Enhancement», 2011, http://stids.c4i.gmu.edu/STIDS2011/papers/STIDS2011_CR_T1_SalmenEtAl.pdf	

А.4 Здравоохранение и медико-биологические науки

А.4.1 Вариант использования № 16: Электронная медицинская карта (EMR)

Название	Электронная медицинская карта (EMR)
Предметная область	Здравоохранение
Автор/организация/эл.почта	Шон Грэнис (Shaun Grannis) / Университет Индианы, США / sgrannis@regenstrief.org
Акторы/заинтересованные лица, их роли и ответственность	**Ученые-исследователи в области биомедицинской информатики** (внедряют и оценивают усовершенствованные методы для бесшовной интеграции, стандартизации, анализа и практического использования сильно неоднородных, высокообъемных потоков клинических данных); **Исследователи в службах здравоохранения** (используют интегрированные и стандартизированные данные электронной медицинской документации для получения знаний, поддерживающих внедрение и оценку трансляционных (ориентированных на практическое использование), сопоставительных (comparative effectiveness), ориентированных на интересы пациента исследований результатов деятельности систем здравоохранения); **Поставщики медицинских услуг — врачи, медсестры, сотрудники государственных органов здравоохранения** (используют информацию и знания, извлеченные из интегрированных и стандартизированных данных электронной медицинской документации, для поддержки непосредственного ухода за пациентами и обеспечения здоровья населения)
Цели	Применение развитых методов для стандартизации выделения понятий (concept identification), связанных с пациентом, поставщиком, учреждением и клинической деятельностью, осуществляемого внутри отдельных организаций сферы здравоохранения и между ними, с целью развития моделей, используемых для определения и извлечения клинических фенотипов (проявлений болезни) из нестандартных, дискретных и представленных в виде свободного текста клинических данных с использованием методов выделения признаков, извлечения информации и моделей принятия решений на основе машинного обучения. Данные клинического фенотипа должны быть использованы для поддержки отбора пациентов в группы (cohort selection), изучения результатов лечения и поддержки принятия клинических решений
Описание варианта использования	По мере того, как системы здравоохранения все в большей степени собирают и потребляют данные электронной медицинской документации, появляются крупные национальные инициативы, направленные на эффективное использование таких данных. В их числе разработка электронной медицинской системы с использованием технологий машинного обучения, поддерживающей принятие клинических решений, все больше основанных на фактических данных, посредством предоставления своевременной, точной и актуальной клинической информации, ориентированной на пациента; использование электронных данных клинических наблюдений для эффективного и быстрого преобразования научных открытий в эффективные клинические методы лечения; и электронный обмен интегрированными данными о здоровье в интересах повышения эффективности и результативности процесса оказания медицинских услуг. Все эти ключевые инициативы опираются на высококачественные, крупномасштабные, стандартизированные и агрегированные данные о здоровье. Несмотря на надежды и обещания, связанные с все более распространенными и вездесущими данными электронной медицинской документации, существует потребность, по целому ряду причин, в развитых методах для интеграции и рационализации этих данных. Данные в клинических системах с течением времени эволюционируют. Это связано с тем, что концептуальное пространство в здравоохранении постоянно развивается: новые научные открытия приводят к выделению новых заболеваний, появлению новых методов диагностики и новых подходов к лечению заболеваний. Это, в свою очередь, приводит к появлению новых клинических понятий, которые являются движущей силой эволюции онтологий для понятий в сфере здравоохранения.

Описание варианта использования	Используя неоднородные данные инфраструктуры клинических данных по уходу за пациентами штата Индиана, США (INPC), крупнейшей и старейшей в США системы обмена медицинской информацией, хранящей свыше 4 млрд дискретных закодированных клинических наблюдений данных из более чем 100 больниц для более чем 12 млн пациентов, мы будем использовать методы извлечения информации для выявления высокорелевантных клинических признаков из электронных данных наблюдений. Для извлечения клинических признаков мы будем использовать методы извлечения информации и обработки естественного языка. Проверенные признаки будут использоваться для параметризации моделей принятия решений по клиническим фенотипам на основе метода оценки максимального правдоподобия и Байесовских сетей. Используя эти модели принятия решений, мы намерены выявить ряд клинических фенотипов, таких как диабет, хроническая сердечная недостаточность и рак поджелудочной железы	
Текущие решения	**Вычислительная система**	Новый суперкомпьютер Cray «Big Red II» в Университете Индианы
	Хранилище данных	Teradata, PostgreSQL, MongoDB
	Сеть связи	Разное. Требуется интенсивная обработка ввода/вывода.
	Программное обеспечение	Hadoop, Hive, R. На основе Unix
Характеристики больших данных	**Источник данных (распределенный/централизованный)**	Клинические данные из более чем 1100 отдельных оперативных медицинских источников в составе инфраструктуры клинических данных по уходу за пациентами штата Индиана, США (INPC), которая является крупнейшей и старейшей в США системой обмена медицинской информацией
	Объем (количество)	Свыше 12 млн пациентов, более 4 млрд отдельных клинических наблюдений, более 20 терабайт первичных данных
	Скорость обработки (например, в реальном времени)	Ежедневно добавляется от 500 тыс. до 1,5 млн новых клинических транзакций в режиме реального времени
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Мы интегрируем широкий спектр клинических наборов данных из ряда источников: записи поставщиков медицинских услуг в виде свободного текста; сведения о лечении в стационаре, амбулаторном лечении, о лечении в отделении интенсивной терапии, о лабораторных исследованиях; данные хромосомной и молекулярной патологии, химических анализов, кардиологических, гематологических, микробиологических и неврологических исследований, записи поставщиков медицинских услуг, данные специализированных лабораторий (referral labs), серологических исследований, хирургической патологии и цитологии, банков крови и токсикологических исследований

Характеристики больших данных	**Вариативность (темпы изменения)**	Данные в клинических системах с течением времени эволюционируют, потому что клиническое и биологическое концептуальные пространства постоянно развиваются: новые научные открытия приводят к выделению новых заболеваний, появлению новых методов диагностики и новых подходов к лечению заболеваний. Это, в свою очередь, приводит к появлению новых клинических понятий, которые являются движущей силой эволюции онтологий для понятий в сфере здравоохранения, которые кодируются самыми разнообразными способами
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Данные каждого клинического источника обычно собираются с использованием отличающихся методов и представлений, что приводит к существенной неоднородности. Это приводит к систематическим ошибкам и отклонениям, требующих применения надежных методов обеспечения семантической совместимости
	Визуализация	Объем, точность и полнота входящих данных должны контролироваться на регулярной основе с использованием нацеленных на это методов визуализации. Внутренне присущие информационные характеристики источников данных должны визуализироваться для выявления неожиданных тенденций
	Качество данных (синтаксис)	Главным препятствием для эффективного использования данных электронной медицинской документации являются сильно различающиеся и уникальные местные названия и коды для одного и того же клинического теста или измерения при выполнении их в разных учреждениях. При интеграции многочисленных источников данных необходимо проводить сопоставление локальных терминов с общей стандартизированной концепцией, с применением, при необходимости, комбинации вероятностных и эвристических методов классификации
	Типы данных	Типы клинических данных весьма разнообразны, включая числовые и структурированные числовые данные, тексты в свободном формате, структурированные тексты, дискретные номинальные данные, дискретные порядковые данные, дискретные структурированные данные, большие двоичные объекты (изображения и видео)
Наука о больших данных (сбор, курирование, анализ, операции)	**Аналитика данных**	Методы извлечения информации с целью выявления соответствующих клинических признаков (статистическая мера TF–IDF, латентно-семантический анализ и статистическая функция «взаимная информация» (mutual information)). Методы обработки естественного языка (natural language processing, NLP) для извлечения релевантных клинических признаков. Проверенные признаки будут использоваться для параметризации моделей принятия решений по клиническим фенотипам на основе метода оценки максимального правдоподобия и Байесовских сетей. Модели принятия решений будут использоваться для выявления ряда клинических фенотипов, таких как диабет, хроническая сердечная недостаточность и рак поджелудочной железы

Иные проблемы больших данных	Устранение систематических ошибок и отклонений в крупномасштабных неоднородных клинических данных в интересах поддержки принятия решений в отношении проведения исследований, ухода за пациентами и в сфере административного управления требует сложной многоэтапной обработки и аналитики, для чего необходимы значительные вычислительные мощности. Кроме того, появляются оптимальные методы для точного и эффективного вывода знаний из данных клинических наблюдений
Проблемы пользовательского интерфейса и мобильного доступа	В рамках всей экосистемы здравоохранения в целом биологические и клинические данные требуются в различных контекстах. Эффективной доставке клинических данных и знаний в рамках экосистемы здравоохранения будет способствовать мобильная платформа, такая как mHealth
Технические проблемы обеспечения безопасности и защиты персональных данных	Неприкосновенность частной жизни и конфиденциальность данных физических лиц должны быть обеспечены в соответствии с требованиями федерального законодательства и законодательства штатов, в том числе Закона США о переносимости и подотчетности медицинского страхования (Health Insurance Portability and Accountability Act, HIPAA) 1996 г. Разработка аналитических моделей с использованием всесторонних интегрированных клинических данных требует агрегирования и последующей деидентификации (обезличивания) перед применением методов сложной аналитики
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Пациенты все чаще получают медицинские услуги в различных местах. Как следствие, данные электронной медицинской документации оказываются фрагментированными и неоднородными. Для того, чтобы реализовать идею самообучающейся медицинской системы (Learning Health Care system), которую продвигает Национальная академия наук и Институт медицины США, данные электронной медицинской документации должны быть рационализированы и интегрированы. Методы, которые мы предлагаем в этом варианте использования, поддерживают интеграцию и рационализацию клинических данных в интересах поддержки принятия решений на различных уровнях
Дополнительная информация (гиперссылки)	Сайт Института Регенстриф (Regenstrief Institute), https://www.regenstrief.org/ Сайт программного обеспечения LOINC (Logical observation identifiers names and codes — «Логичные идентификаторы, имена и коды наблюдений»), https://loinc.org/ Сайт Центра обмена медицинской информацией Индианы (Indiana Health Information Exchange), https://www.ihie.org/ «Самообучающаяся медицинская система. Итоги семинара» (The Learning Healthcare System — Workshop Summary), Круглый стол Института доказательной медицины (медицины, основывающейся на фактах) (IOM roundtable on evidencebased medicine), 2007 год, 375 стр., https://www.nap.edu/catalog/11903/the-learning-healthcare-system-workshop-summary (возможно бесплатное скачивание)

А.4.2 Вариант использования № 17: Анализ графических образов в патологии / Цифровая патология

Название	Анализ графических образов в патологии / Цифровая патология
Предметная область	Здравоохранение
Автор/организация/эл.почта	Ван Фушен (Fusheng Wang) / Университет Эмори (Emory University) / fusheng.wang@emory.edu
Акторы/заинтересованные лица, их роли и ответственность	Исследователи в сфере биомедицины, занимающиеся трансляционными исследованиями; врачи больниц, занимающиеся диагностикой на основе изображений
Цели	Разработка высокоэффективных алгоритмов анализа изображений для извлечения из них пространственной информации; поддержка эффективных пространственных запросов и аналитики, а также кластеризации и классификации признаков

Описание варианта использования	Анализ цифровых графических образов в патологии (digital pathology imaging) является нарождающейся областью, в которой изучение сделанных с высоким разрешением изображений образцов тканей позволяет создавать новые и более эффективные способы диагностики заболеваний. В рамках патологического анализа изображений выделяется огромное (миллионы на изображение) количество пространственных объектов, таких как ядра клеток и кровеносные сосуды, представленные их границами, наряду со многими извлеченными по изображению признаками этих объектов. Полученная информация используется для многих сложных запросов и аналитики, поддерживающих биомедицинские исследования и клиническую диагностику. Недавно стал возможен патологический анализ трехмерных изображений, на основе использования трехмерных лазерных технологий либо последовательного размещения сотен срезов тканей на предметные стекла и их сканирования в цифровые изображения. Выделение трехмерных гистологических объектов на основе серий зафиксированных изображений может породить десятки миллионов трехмерных объектов по одному трехмерному изображению. В результате формируется глубокая «карта» тканей человека для использования в методах диагностики следующего поколения	
Текущие решения	**Вычислительная система**	Суперкомпьютеры; облако
	Хранилище данных	SAN или HDFS
	Сеть связи	Требуется отличное внешнее сетевое соединение
	Программное обеспечение	MPI для анализа изображений; Map/Reduce + Hive с пространственным расширением
Характеристики больших данных	**Источник данных (распределенный/централизованный)**	Оцифрованные графические образы образцов человеческих тканей для целей патологического анализа
	Объем (количество)	1 гигабайт первичных данных + 1,5 гигабайта аналитических результатов на двумерное изображение; 1 терабайт первичных данных + 1 терабайт аналитических результатов на трехмерное изображение. 1 петабайт данных в год в средней больнице
	Скорость обработки (например, в реальном времени)	После создания данные не подвергаются изменениям
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Характеристики изображений и виды аналитики зависят от типа заболевания
	Вариативность (темпы изменения)	Изменений нет
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Важнейшее значение имеет высокое качество результатов, подтвержденное сделанными человеком аннотациями
	Визуализация	Необходима для проверки и обучения
	Качество данных (синтаксис)	Зависит от предварительной обработки предметных стекол, такой, как химическое окрашивание, и от качества алгоритмов анализа изображений

Наука о больших данных (сбор, курирование, анализ, операции)	**Типы данных**	Первичные изображения представляют собой полные графические образы предметных стекол (в основном на основе BIGTIFF), а аналитические результаты представляют собой структурированные данные (пространственные границы и признаки)
	Аналитика данных	Анализ изображений, пространственные запросы и аналитика, кластеризация и классификация признаков
Иные проблемы больших данных	Экстремально большие объемы; многомерность; аналитика является специфической для конкретных заболеваний; корреляция с данными других типов (клинические данные, данные «омиков» (omics) — таких направлений биологической науки, как геномика, протеомика, метаболомика и др.)	
Проблемы пользовательского интерфейса и мобильного доступа	Трехмерная визуализация трехмерных патологических изображений маловероятна на мобильных платформах	
Технические проблемы обеспечения безопасности и защиты персональных данных	Следует обеспечить защиту защищаемой информации о здоровье (protected health information); общедоступные данные должны быть деидентифицированы	
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Данные графических изображений; многомерная пространственная аналитика данных	
Дополнительная информация (гиперссылки)	“Digital Pathology: Data-Intensive Frontier in Medical Imaging”, Proceedings of the IEEE, Volume 100, Number 4, 2012, https://open.library.emory.edu/publications/emory:tzzn8/ Fusheng Wang et al. “A data model and database for high-resolution pathology analytical image informatics”, J.Pathol.Inform., 2011; 2:32, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3153692/ Fusheng Wang “Hadoop-GIS: A high performance query system for analytical medical imaging with MapReduce”, 2011, https://www.researchgate.net/publication/291559237_Hadoop-gis_A_high_performance_query_system_for_analytical_medical_imaging_with_mapreduce Ablimit Aji, Fusheng Wang, Hoang Vo, Rubao Lee, Qiaoling Liu, Xiaodong Zhang, Joel Saltz «Hadoop-GIS: A High Performance Spatial Data Warehousing System over MapReduce», Proceedings of the VLDB Endowment, Volume 6, Number 11, 2013, https://open.library.emory.edu/publications/emory:v0fvn/	

А.4.3 Вариант использования № 18: Вычислительный анализ биоизображений

Название	Вычислительный анализ биоизображений (Computational Bioimaging)	
Предметная область	Научные исследования, биология	
Автор/организация/эл.почта	Дэвид Скиннер (David Skinner), deskinner@lbl.gov, и Хоакин Корреа (Joaquin Correa), JoaquinCorrea@lbl.gov , — оба из Национального научно-исследовательского вычислительного центра энергетических исследований Министерства энергетики США (NERSC) при Национальной лаборатории имени Лоуренса в Беркли, США (LBNL), Дэниэла Ушидзима (Daniela Ushizima), dushizima@lbl.gov, и Йорг Мейер (Joerg Meyer), joergmeyer@lbl.gov, оба из Отделения вычислительных исследований (Computational Research Division) Национальной лаборатории имени Лоуренса в Беркли, США	
Акторы/заинтересованные лица, их роли и ответственность	**Поставщики возможностей и ресурсов:** операторы оборудования для работы с биоизображениями, разработчики микроскопов, организации и подразделения по обработке графических образов, специалисты в области прикладной математики и кураторы данных. **Сообщество пользователей:** Министерство энергетики США, представители теоретической и отраслевой науки, стремящиеся совместными усилиями создавать модели на основе данных, содержащихся в графических образах	
Цели	Данные биоизображений все более автоматизированно создаются с более высоким разрешением и являются более мультимодальными. В результате возникает узкое место в анализе данных, устранение которого может способствовать новым открытиям в биологических науках посредством применения технологий больших данных. Цель заключается в том, чтобы устранить данное узкое место с помощью экстремально масштабных вычислений. Достижение этой цели потребует не только вычислений. Потребуется создать сообщества вокруг ресурсов данных и разработать продвинутые алгоритмы для массового анализа изображений. Высокопроизводительные вычислительные решения могут использоваться ориентированными на эти сообщества научными шлюзами с целью направлять применение массового анализа данных к огромным наборам данных, полученных из изображений. Компоненты потока рабочих процессов включают сбор, хранение, улучшение качества данных, минимизацию шума, сегментацию представляющих интерес областей, групповой отбор и извлечение признаков, классификацию объектов, а также организацию и поиск	
Описание варианта использования	Интернет-точка обслуживания по принципу одного окна, обеспечивающая высокопроизводительную, с высокой пропускной способностью обработку изображений в интересах создателей и потребителей моделей, построенных на основе данных биоизображений	
Текущие решения	**Вычислительная система**	Суперкомпьютер Hopper (150 тысяч процессоров) в Национальном научно-исследовательском вычислительном центре энергетических исследований Министерства энергетики США (NERSC)
	Хранилище данных	База данных и коллекции изображений
	Сеть связи	10 гигабит/с, желательны 100 гигабит/с и расширенные сетевые возможности (программно-конфигурируемая сеть [передачи данных] SDN)
	Программное обеспечение	ImageJ, OMERO, VolRover, разработанные прикладными математиками продвинутые методы сегментации и выявления признаков

Характеристики больших данных	**Источник данных (распределенный/ централизованный)**	Распределенные экспериментальные источники биоизображений (приборы). Запланированные потоки большого объема от автоматизированных оптических и электронных микроскопов высокого разрешения
	Объем (количество)	Объемы данных растут очень быстро. Необходимы масштабируемые базы данных для данных типа «ключ-значение» и для библиотек объектов. Является актуальной обработка данных и аналитика непосредственно в базах данных. Проект в настоящее время работает с 50 терабайтами, однако в целом объем таких данных превышает петабайт. Объем данных в результате одного сканирования на появляющихся установках составляет 32 терабайта
	Скорость обработки (например, в реальном времени)	Высокопроизводительные вычисления (high throughput computing, HTC), гибкий анализ
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Мультимодальный сбор и анализ изображений (multimodal imaging), по сути, должен обеспечить комбинирование поступающих по разрозненным каналам данных, с акцентом на регистрацию и форматы наборов данных
	Вариативность (темпы изменения)	Биологические образцы сильно различаются, и рабочие процессы их анализа должны с этим справляться
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Данные в целом неупорядоченные, как и обучение классификаторов
	Визуализация	Очень интенсивно используются трехмерные структурные модели
	Качество данных (синтаксис)	
	Типы данных	Файловые форматы изображений
	Аналитика данных	Машинное обучение (метод опорных векторов (Support Vector Machine, SVM) и алгоритм «случайный лес» (random forest, RF) для сервисов классификации и рекомендательных сервисов
Иные проблемы больших данных	Масштабные высокопроизводительные вычисления для программ моделирования. Гибкие методы массовой обработки неупорядоченных данных. Системы машинного обучения и знаний, которые извлекают из данных растровой графики информацию, связанную с биологическими объектами и моделями	
Проблемы пользовательского интерфейса и мобильного доступа		
Технические проблемы обеспечения безопасности и защиты персональных данных		

Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Существует потенциал для обобщения концепций поиска в контексте обработки биоизображений
Дополнительная информация (гиперссылки)	

А.4.4 Вариант использования № 19: Геномные измерения

Название	Геномные измерения	
Предметная область	Здравоохранение	
Автор/организация/эл.почта	Джастин Зук (Justin Zook) / Национальный институт стандартов и технологий (NIST) / jzook@nist.gov	
Акторы/заинтересованные лица, их роли и ответственность	Поддерживаемое американским Национальным институтом стандартов и технологий (NIST) государственно — частно — академическое партнерство «Консорциум «Геном в бутылке»» (Genome in a Bottle Consortium, https://www.nist.gov/programs-projects/genome- bottle)	
Цели	Разработка надежных и хорошо изученных эталонных материалов, данных и методов, необходимых для оценки эффективности секвенирования генома	
Описание варианта использования	Объединение данных, полученных в результате применения различных технологий и методов секвенирования с целью создания высоконадежных описаний полных геномов человека в качестве эталонных материалов; а также разработка методов использования этих эталонных материалов для оценки эффективности алгоритмов секвенирования генома	
Текущие решения	**Вычислительная система**	72-ядерный кластер нашей группы в NIST, взаимодействие с ~1000-ядерными кластерами Управления по контролю за качеством пищевых продуктов и медикаментов (Food and Drug Administration, FDA). Некоторые группы используют облако
	Хранилище данных	Около 40 терабайт в файловой системе NFS в NIST, петабайты геномных данных в Национальных учреждениях здравоохранения (NIH) / Национальном центре биотехнологической информации (NCBI)
	Сеть связи	Разное. Требуется интенсивная обработка ввода/вывода
	Программное обеспечение	Программное обеспечение с открытым исходным кодом для секвенирования в биоинформатике, разработанное академическими группами (на основе UNIX)

Характеристики больших данных	**Источник данных (распределенный/ централизованный)**	Секвенсоры распределены по многим лабораториям, хотя существует ряд ключевых центров
	Объем (количество)	40-терабайтная файловая система NFS в NIST заполнена. В течение года-двух в NIST потребуется >100 терабайт. Сообществу здравоохранения в целом потребуется много петабайт для хранения данных
	Скорость обработки (например, в реальном времени)	Секвенсоры ДНК способны генерировать порядка ~300 гигабайт сжатых данных в день; рост объемов данных идет намного быстрее закона Мура
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Файловые форматы недостаточно хорошо стандартизированы, хотя некоторые стандарты существуют. Как правило, структурированные данные
	Вариативность (темпы изменения)	Технологии секвенирования развиваются очень быстро, и новые технологии уже появились на горизонте
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	У всех технологий секвенирования имеются значительные систематические ошибки и погрешности, для выявления которых требуются сложные методы анализа и совместное применение ряда технологий, часто с использованием машинного обучения
	Визуализация	Для визуализации обработанных данных были разработаны «браузеры генома»
	Качество данных (синтаксис)	У технологий секвенирования и методов биоинформатики имеются значительные систематические ошибки и погрешности
	Типы данных	В основном, структурированный текст
	Аналитика данных	Обработка первичных данных с целью выделения вариаций (variant calls), а также клиническая интерпретация вариаций, которая в настоящее время является серьезной проблемой
Иные проблемы больших данных	Обработка данных требует значительных вычислительных мощностей, что создает проблемы — особенно для клинических лабораторий, по мере того они начинают проводить широкомасштабное секвенирование. Долговременное хранение данных клинического секвенирования может быть дорогостоящим. Методы анализа быстро эволюционируют. Многие части генома сложно анализировать, а систематические ошибки трудно выявлять	
Проблемы пользовательского интерфейса и мобильного доступа	Врачам может понадобиться доступ к геномным данным на мобильных платформах	

Технические проблемы обеспечения безопасности и защиты персональных данных	Необходимо обеспечить безопасность и защиту неприкосновенности частной жизни в отношении данных секвенирования, хранимых в составе медицинской документации или в базах данных клинических исследований. В то же время данные Консорциума являются общедоступными
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	У нас есть ряд идей по обобщению описанных выше работ по секвенированию генома в медицине; однако основное внимание мы уделяем работе в рамках деятельности NIST/Консорциума «Геном в бутылке». В настоящее время наша лаборатории занимается секвенированием разного масштаба, от малого до очень большого. В будущем в состав данных могут входить результаты измерений, сделанных в рамках других направлений биологической науки — «омиков» (omics, например, геномика), объем которых будет даже больше, чем объемы результатов секвенирования ДНК
Дополнительная информация (гиперссылки)	Сайт Консорциума «Геном в бутылке» (Genome in a Bottle Consortium), https://www.nist.gov/programs-projects/genome-bottle

А.4.5 Вариант использования № 20: Сравнительный анализ (мета) геномов

Название	Сравнительный анализ метагеномов и геномов	
Предметная область	Научные исследования, геномика	
Автор/организация/эл.почта	Эрнест Сзето (Ernest Szeto) / Национальная лаборатория имени Лоуренса в Беркли, США (LBNL), eszeto@lbl.gov	
Акторы/заинтересованные лица, их роли и ответственность	Проект «Интегрированные микробные геномы» (IMG) Объединенного института генома (JGI) Министерства энергетики США; руководители Виктор Маркович (Victor M. Markowitz) и Никос Кипридес (Nikos C. Kyrpides). **Сообщество пользователей** JGI, биологи и специалисты по биоинформатике различных стран	
Цели	Создание интегрированной системы сравнительного анализа метагеномов и геномов. Сюда входит разработка интерактивного пользовательского веб-интерфейса к основным данным, предварительные вычисления на сервере (backend precomputations) и отправка пакетных заданий из пользовательского интерфейса	
Описание варианта использования	Для метагеномных образцов: (1) Определить состав изучаемой колонии/сообщества с точки зрения присутствия других эталонных изолированных геномов; (2) Охарактеризовать функции его генов; (3) Начать выявление возможных функциональных путей (functional pathways); (4) Охарактеризовать сходство или различие по сравнению с другими метагеномными образцами; (5) Начать характеризацию изменений в составе и функциях сообщества в связи с изменениями воздействием факторов окружающей среды; (6) Выделить подразделы данных на основе показателей качества и состава сообщества	
Текущие решения	**Вычислительная система**	Linux-кластер, сервер реляционной СУБД Oracle, большие системы хранения данных, стандартные интерактивные хосты Linux
	Хранилище данных	Реляционная СУБД Oracle, файлы SQLite, плоские текстовые файлы, Lucy (версия Lucene) для поиска по ключевым словам, базы данных BLAST, базы данных USEARCH

Текущие решения	**Сеть связи**	Обеспечивается Национальным научно-исследовательским вычислительным центром энергетических исследований Министерства энергетики США (NERSC)
	Программное обеспечение	Стандартные инструменты биоинформатики (BLAST, HMMER, инструменты множественного выравнивания последовательностей и филогенетики, программы поиска/предсказания генов и генных структур (gene callers), программы предсказания свойств по результатам секвенирования (sequence feature predictors) и т. д.), скрипты Perl / Python, планировщик задач Linux-кластера
Характеристики больших данных	**Источник данных (распределенный/ централизованный)**	Централизованный
	Объем (количество)	50 терабайт
	Скорость обработки (например, в реальном времени)	Веб-интерфейс пользователя должен быть интерактивным в реальном времени. Возможности обработки загружаемых данных на сервере должны соответствовать экспоненциальному росту объемов данных секвенирования из-за быстрого снижения стоимости технологии секвенирования
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Биологические данные по своей природе неоднородны, сложны, структурны и иерархичны — начинается с последовательностей, за которыми следуют свойства последовательностей, таких как гены, мотивы, регуляторные области; далее следует организация находящихся по соседству генов (опероны); и так вплоть до белков и их структурных особенностей; координации и экспрессии генов в путях. Помимо базовых геномных данных, в систему сравнительного анализа должны быть включены новые типы данных таких направлений биологической науки — «омиков» (omics), как транскриптомика, метиломика (methylomics) и протеомика, описывающих экспрессии генов в различных условиях
	Вариативность (темпы изменения)	Размеры метагеномных образцов могут варьироваться на несколько порядков величины — от нескольких сотен тысяч до миллиарда генов (как, например, в сложном образце почвы)

Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Методы отбора и анализа метагеномных проб в настоящее время являются предварительными и экспериментальными. Процедуры оценки набора сильно фрагментированных данных первичных измерений проработаны лучше, но все еще остаются открытой областью исследований
	Визуализация	Проблемой остается быстрота интерактивного пользовательского веб-интерфейса при работе с очень большими наборами данных. Пользовательский веб-интерфейс, судя по всему, по-прежнему является предпочтительным для большинства биологов. Он используется для базовых запросов и просмотра данных. Из него могут быть запущены более специализированные инструменты, например, для просмотра множественных выравниваний. Еще одним требованием к системе является возможность загружать большие объемы данных для анализа в автономном (offline) режиме
	Качество данных (синтаксис)	Улучшение качества метагеномной «сборки» (metagenomic assembly) по-прежнему является ключевой проблемой. Улучшение качества эталонных изолированных геномов, с точки зрения как охвата филогенетического дерева, так и улучшенного поиска/предсказания генов и генных структур и функциональной аннотации — более зрелый процесс, который, однако, постоянно продолжается
	Типы данных	См. выше раздел «Разнообразие»
	Аналитика данных	Описательная статистика, статистическая значимость при проверке гипотез, выявление новых взаимосвязей, кластеризация и классификация данных являются стандартными элементами аналитики. Менее «количественная» часть включает в себя возможность визуализации структурных элементов на разных уровнях разрешения. Редукция данных, устранение избыточности посредством кластеризации, более абстрактные представления, такие как представление группы очень похожих геномов в виде пангенома, — все это стратегии, предназначенные как для управления данными, так и для аналитики

Иные проблемы больших данных	Главным другом и союзником в деле управления неоднородными биологическими данными по-прежнему является реляционная СУБД. К сожалению, она не масштабируется на ныне имеющиеся объемы данных. Решения класса NoSQL (СУБД, существенно отличающиеся от традиционных реляционных) должны были обеспечить альтернативу, но, к сожалению, они не всегда пригодны для интерактивного использования в реальном времени или же для быстрой параллельной массовой загрузки; и иногда у них возникают проблемы с надежностью. Наш текущий подход в настоящее время является нестандартным, специфическим для нашей ситуации, и мы опираемся главным образом на Linux — кластер и файловую систему в качестве дополнения к реляционной СУБД Oracle. Наше решение часто полагается на знание особенностей данных, что позволяет нам разрабатывать схемы горизонтального секционирования, а также осуществлять, когда это уместно, реорганизацию данных
Проблемы пользовательского интерфейса и мобильного доступа	Каких-то особых проблем нет. Требуется лишь доступ в интернет
Технические проблемы обеспечения безопасности и защиты персональных данных	Каких-то особых проблем нет. Данные либо являются общедоступными, либо для доступа к ним требуются обычные логин и пароль
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Для всех принесло бы пользу появление альтернативы реляционным СУБД в сфере больших данных. Многие решения типа NoSQL пытаются выполнить эту роль, однако у них есть свои ограничения
Дополнительная информация (гиперссылки)	Страница проекта «Интегрированные микробные геномы и микробиомы» (Integrated Microbial Genomes and Microbioms, IMG/M) на сайте Объединенного института генома (JGI), https://img.jgi.doe.gov/

А.4.6 Вариант использования № 21: Индивидуальное управление лечением диабета

Название	Индивидуальное управление лечением диабета
Предметная область	Здравоохранение
Автор/организация/эл.почта	Питер Ли (Peter Li), Йин Дин (Ying Ding), Филип Юи (Philip Yu), Джоффри Фокс (Geoffrey Fox), Дэвид Уальд (David Wild) / Клиника Мейо (Mayo Clinic), университет Индианы (IU), университет Иллинойса в Чикаго (UIC) / dingying@indiana.edu
Акторы/заинтересованные лица, их роли и ответственность	Клиника Мейо + университет Индианы — семантическая интеграция данных из электронных медицинских документов. Университет Иллинойса в Чикаго — интеллектуальный анализ семантических данных из электронных медицинских документов. Университет Индианы — облачные и параллельные вычисления
Цели	Разработка передовых методов интеллектуального анализа данных, представленных в виде графов, и их применение в отношении электронной медицинской документации, с целью выявления демографических когорт и извлечения из электронных медицинских документов соответствующих данных для оценки результатов лечения. Эти методы расширят границы масштабируемости и технологий интеллектуального анализа данных; будут способствовать развитию знаний и практики в этих областях, а также клиническому управлению лечением сложных заболеваний

Описание варианта использования	Диабет — это болезнь, которая становится все более распространенной среди населения Земли, затрагивая как развивающиеся, так и развитые страны. Современные стратегии управления лечением не учитывают должным образом индивидуальные профили пациентов, в том числе наличие сопутствующих заболеваний и прием соответствующих лекарств — обычное явление у пациентов с хроническими заболеваниями. Мы предлагаем устранить этот недостаток путем выявления похожих пациентов из большой базы данных электронной медицинской документации (т. е. путем формирования индивидуализированной демографической когорты), и оценки результатов их лечения с тем, чтобы выбрать наилучшее решение, подходящее для конкретного больного диабетом. Ниже описаны этапы выполнения проекта: **Этап 1:** Применение «метода семантического связывания для значений свойств» (Semantic Linking for Property Values) для преобразования данных из хранилище данных в Клинике Мейо, США (EDT), в триплеты RDF, что дает нам возможность гораздо эффективнее выявлять похожих пациентов за счет связывания как словарных, так и числовых значений. **Этап 2:** Требуются эффективные параллельные алгоритмы поиска и извлечения, подходящие для облачных и/или высокопроизводительных вычислений. Нереляционная СУБД Hbase с открытым исходным кодом используется для поиска по индексу и настраиваемого поиска с целью выявления потенциально представляющих интерес пациентов. **Этап 3:** Данные из электронных медицинских документов, представленные в виде RDF-графа, предоставляют собой богатую среду для интеллектуального анализа закономерностей в графе. Требуются новые алгоритмы распределенного интеллектуального анализа графов с целью выполнения анализа закономерностей и применения метода индексации графов в интересах поиска закономерностей в графах на основе триплетов RDF. **Этап 4:** Учитывая размер и сложность графов, интеллектуальный анализ закономерностей в подграфах может сгенерировать множество ложноположительных и ложноотрицательных результатов. Требуются надежные инструменты статистического анализа для контроля частоты ложных срабатываний, определения истинной значимости подграфа и проверки результатов в рамках нескольких клинических вариантов использования	
Текущие решения	**Вычислительная система**	Суперкомпьютеры, облако
	Хранилище данных	Распределенная файловая система HDFS (Hadoop distributed file system)
	Сеть связи	Разное. Требуется интенсивная обработка ввода/вывода
	Программное обеспечение	Внутреннее хранилище данных в Клинике Мейо, США (EDT)
Характеристики больших данных	**Источник данных (распределенный/централизованный)**	Распределенные данные электронной медицинской документации
	Объем (количество)	База данных электронных медицинских документов Клини Мейо (Clinic Mayo) представляет собой очень большой набор данных, охватывающий более 5 млн пациентов с тысячами свойств по каждому, и многие другие сведения, полученные из первичных данных
	Скорость обработки (например, в реальном времени)	Не в режиме реального времени, но данные периодически обновляются

Характеристики больших данных	**Разнообразие (множество наборов данных, комбинация данных из различных источников)**	Структурированные данные. Значения свойств пациента могут быть как из контролируемых словарей (демография, диагностические коды, лекарства, процедуры и т. д.), так и непрерывные числовые величины (лабораторные анализы, количество лекарств, показатели жизненно важных функций и т. д.). Число значений свойств может варьироваться от менее 100 (новый пациент) до более чем 100 тысяч (длительно наблюдаемый пациент), при этом типичным для пациента является около 100 значений свойств из контролируемых словарей и 1000 непрерывных числовых величин. Большинство значений привязаны ко времени, т. е. отметка времени фиксируется вместе со значением в момент наблюдения
	Вариативность (темпы изменения)	Данные обновляются или добавляются при каждом визите пациента
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Данные аннотируются на основе онтологий или таксономий предметной области. Семантика данных может варьироваться от лаборатории к лаборатории
	Визуализация	Отсутствует
	Качество данных (синтаксис)	Сведения о происхождение (provenance) имеют важное значение для отслеживания происхождения данных и оценки их качества
	Типы данных	Текстовые данные, непрерывные числовые величины
	Аналитика данных	Интеграция данных в семантический граф, использование обхода графа взамен операции join в SQL. Разработка алгоритмов интеллектуального анализа семантических графов с целью выявления закономерностей в графе, индексирования графа и поиска по нему. СУБД Hbase с индексированием. Специализированная программа для выявления новых свойств пациента на основе хранящихся данных
Иные проблемы больших данных	В рамках индивидуализированной демографической когорты, мы по существу создадим информационное табло (datamart) для каждого пациента, поскольку важнейшие свойства и показатели будут индивидуальными для каждого пациента. Из-за количества пациентов создание таких табло в индивидуальном порядке становится непрактичным. По сути, парадигма меняется от поиска строки — столбца в таблицах реляционной базы данных на обход семантического графа	
Проблемы пользовательского интерфейса и мобильного доступа	Врачам и пациентам может понадобиться доступ к этим данным на мобильных платформах	

Технические проблемы обеспечения безопасности и защиты персональных данных	Должны быть обеспечены безопасность и защита персональных данных в медицинских документах и клинических базах данных
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Интеграция данных: непрерывные числовые величины, онтологическая аннотация, таксономия. Поиск по графу: индексирование графа и поиск по нему. Валидация: статистическая валидация
Дополнительная информация (гиперссылки)	

А.4.7 Вариант использования № 22: Статистический реляционный искусственный интеллект для здравоохранения

Название	Статистический реляционный искусственный интеллект для здравоохранения	
Предметная область	Здравоохранение	
Автор/организация/эл.почта	Шрирам Натараджан (Sriraam Natarajan) / Университет Индианы (Indiana University) / natarasr@indiana.edu	
Акторы/заинтересованные лица, их роли и ответственность	Исследователи в области информатики и медицины, практики в области медицины	
Цели	Целью проекта является анализ больших, мультимодальных данных длительного наблюдения (longitudinal data). Анализ различных типов, таких, как изображения, электронные данные (карта) здоровья (EHR), генетические данные и данные на естественном языке, требует богатых средств представления (rich representation). В рамках данного подхода используются реляционные вероятностные модели, способные работать с богатыми реляционными данными и моделирующие неопределенности на основе теории вероятности. Программное обеспечение обучает модели на основе ряда типов данных, и, возможно, сможет интегрировать информацию и логические рассуждения о сложных запросах	
Описание варианта использования	Пользователи могут представить набор сведений, например образы магнитно-резонансной томографии (МРТ) и демографические данные о конкретном субъекте. Затем они могут сделать запрос о начале конкретного заболевания (например, болезни Альцгеймера), и система выдаст распределение вероятностей для возможного возникновения этого заболевания	
Текущие решения	**Вычислительная система**	Для исполнения программы обработки данных нескольких сотен пациентов необходим высокопроизводительный компьютер (48 ГБ ОЗУ). Кластеры нужны в случае обработки больших наборов данных
	Хранилище данных	Обычно тестовые данные хранятся на жестком диске емкостью от 200 гигабайт до 1 терабайта. При выполнении алгоритмов соответствующие данные извлекаются в основную память. Данные на сервере хранятся в базе данных или в хранилищах типа NoSQL
	Сеть связи	Интранет

Текущие решения	**Программное обеспечение**	В основном на основе Java, для обработки данных используются инструменты собственной разработки
Характеристики больших данных	**Источник данных (распределенный/ централизованный)**	Все данные о пользователях хранятся в одном файле на диске. Иногда должны быть извлечены из Интернета такие ресурсы, как опубликованные тексты
	Объем (количество)	Объем может варьироваться из-за разного количества собранных данных. Типичный объем измеряется сотнями гигабайт для одной когорты из нескольких сотен человек. Когда речь идет о миллионах пациентов, объем данных может быть порядка 1 петабайта
	Скорость обработки (например, в реальном времени)	Различная. В некоторых случаях электронные медицинские документы постоянно обновляются. В других контролируемых исследованиях данные часто поступают партиями через равные промежутки времени
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Это ключевое свойство наборов медицинских данных. Такие данные обычно содержатся в ряде таблиц, которые необходимо объединить для выполнения анализа
	Вариативность (темпы изменения)	Поступление данных во многих случаях непредсказуемо, поскольку они поступают в режиме реального времени
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Проблематична, ввиду различной модальности данных, человеческих ошибок при сборе и проверке данных
	Визуализация	Визуализация всей совокупности исходных данных практически невозможна. Обычно данные визуализируются частично. Построенные модели могут быть визуализированы при определенных разумных допущениях
	Качество данных (синтаксис)	
	Типы данных	Электронные медицинские документы, графические изображения, генетические данные, которые хранятся в нескольких базах данных
	Аналитика данных	

Иные проблемы больших данных	Во многих направлениях медицины данные имеются в изобилии. Ключевой вопрос заключается в том, что данных может быть слишком много (таких как изображения, генетические последовательности и т. д.), что может усложнить анализ. Реальной проблемой является согласование данных и слияние данных из нескольких источников в форме, полезной для их совместного анализа. Еще одна проблема заключается в том, что иногда доступны большие объемы данных об одном субъекте, но число субъектов пи этом не очень велико (то есть имеется дисбаланс данных). Это может привести к тому, что в ходе анализа алгоритмы обучения расценят случайные корреляции между данными нескольких типов как важные свойства. Ввиду этого имеют первостепенное значение робастные методы обучения, способные верно моделировать данные. Еще одним аспектом дисбаланса данных является частота позитивных примеров (случаев). Некоторые заболевания могут встречаться редко, что делает отношение позитивных примеров к «контролям» крайне искаженным, и в этом случае алгоритмы обучения могут моделировать шум вместо примеров
Технические проблемы обеспечения безопасности и защиты персональных данных	Обеспечение безопасности при подготовке и обработке данных имеет критически важное значение в медицинских областях
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Модели, обученные на одной группе населения, трудно обобщить на другие группы населения с отличающимися характеристиками. Для этого необходимо, чтобы обученные модели можно было обобщать и уточнять в соответствии с изменением характеристик населения
Дополнительная информация (гиперссылки)	

А.4.8 Вариант использования № 23: Эпидемиологическое исследование в масштабе всего населения Земли

Название	Эпидемиологическое исследование в масштабе всего населения Земли
Предметная область	Эпидемиология, моделирование в социальных (общественных) науках, вычислительные социальные науки
Автор/организация/эл.почта	Мадхав Марате (Madhav Marathe, mmarathe@vbi.vt.edu), Стивен Юбанк (Stephen Eubank, seubank@vbi.vt.edu) и Крис Барретт (Chris Barrett, cbarrett@vbi.vt.edu) / Институт биосложности (Biocomplexity Institute, ранее Институт биоинформатики) Политехнического университета/университета штата Вирджиния (Virginia Tech)
Акторы/заинтересованные лица, их роли и ответственность	Государственные и некоммерческие учреждения, занимающиеся вопросами здравоохранения, государственной политики и смягчения последствий стихийных бедствий и катастроф. Социологи, желающие изучить взаимодействие между поведением и распространением инфекции
Цели	(a) Сформировать синтетическую глобальную популяцию; и (b) Провести моделирование в масштабе глобальной популяции с тем, чтобы сделать выводы о вспышках заболеваемости и различных стратегиях вмешательства
Описание варианта использования	Прогнозирование и контроль над пандемиями, аналогичными пандемии гриппа H1N1 в 2009 г.

Текущие решения	**Вычислительная система**	Распределенная на основе использования интерфейса передачи сообщений MPI (Message Passing Interface) система моделирования, написанная на Charm++. Параллелизм достигается за счет использования меры «время присутствия болезни» (disease residence time period)
	Хранилище данных	Сетевая файловая система NFS (Network file system). Изучаются методы на основе баз данных
	Сеть связи	Высокоскоростная коммутируемая компьютерная сеть Infiniband. Топология трехмерного тора с высокой пропускной способностью
	Программное обеспечение	Charm++, MPI
Характеристики больших данных	**Источник данных (распределенный/централизованный)**	Данные генерируются с помощью генератора синтетической популяции, в настоящее время — централизованно. Тем не менее генерация может быть сделана распределенной как часть постобработки
	Объем (количество)	100 терабайт
	Скорость обработки (например, в реальном времени)	Взаимодействие с экспертами и процедуры визуализации производят большие объемы данных в реальном времени. Подача данных в программу моделирования мала, однако в ходе моделирования создаются огромные объемы данных
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Разнообразие зависит от сложности используемой в моделировании модели. Модель может быть очень сложной, если принять во внимание иные аспекты мировой популяции, такие как тип деятельности, географические, социально-экономические и культурные различия
	Вариативность (темпы изменения)	Зависит от эволюции модели и соответствующих изменений в программе. Это сложная работа, требующая много времени, — отсюда и низкая скорость изменения
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Стабильность результатов моделирования зависит от качества модели. В то же время стабильность собственно вычислений — вопрос решаемый, хотя и нетривиальный
	Визуализация	Для подключения визуализации потребуется пересылать очень большие объемы данных
	Качество данных (синтаксис)	Данные согласованы благодаря генерации на основе модели
	Типы данных	В основном сетевые данные

Наука о больших данных (сбор, курирование, анализ, операции)	Аналитика данных	Сводка по различным прогонам и повторам моделирования
Иные проблемы больших данных	Вычисления в процессе моделировании требуют как значительных вычислительных ресурсов, так и обработки больших объемов данных. Более того, из-за неструктурированного и нерегулярного характера обработки графов, проблему сложно решать по частям. По этой причине также требуется широкая полоса пропускания. Следовательно, суперкомпьютер подходит больше, чем кластеры облачного типа	
Проблемы пользовательского интерфейса и мобильного доступа	Нет	
Технические проблемы обеспечения безопасности и защиты персональных данных	Есть ряд проблем на этапе моделирования синтетической популяции (см. модель распространения социального влияния)	
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	В общем случае можно моделировать и вычислять распространение явлений различного рода: информации, болезней, социальных волнений. Во всех этих случаях применяется модель на основе лиц-агентов (agent-based model), которая использует базовую сеть взаимодействий для изучения эволюции рассматриваемых явлений	
Дополнительная информация (гиперссылки)		

А.4.9 Вариант использования № 24: Моделирование распространения социального влияния

Название	Моделирование распространения социального влияния
Предметная область	Социальное поведение (включая вопросы национальной безопасности, здравоохранения; вирусный маркетинг, городское планирование, готовность к чрезвычайным ситуациям и катастрофам)
Автор/организация/эл.почта	Мадхав Марате (Madhav Marathe, mmarathe@vbi.vt.edu) и Крис Кульман (Chris Kuhlman, ckuhlman@vbi.vt.edu) / Институт биосложности (Biocomplexity Institute, ранее Институт биоинформатики) Политехнического университета/университета штата Вирджиния (Virginia Tech)
Акторы/заинтересованные лица, их роли и ответственность	
Цели	Создать вычислительную инфраструктуру, которая моделирует процессы распространения социального влияния. Эта инфраструктура позволяет моделировать различные типы взаимодействия между людьми (например, лицом к лицу либо через социальные сети; отношения мать — дочь в сравнении с отношениями мать — коллега). Учитываются не только взаимоотношения между людьми, но и взаимоотношения между людьми и сервисами (например, транспорт) либо инфраструктурой (например, Интернет, электроснабжение)
Описание варианта использования	Социальные волнения. Люди выходят на улицы, чтобы выразить свое недовольство либо поддержку руководству государства. Среди граждан есть как те, кто поддерживает правительство, так и те, кто ему противостоит. Ставятся задачи количественно определить степень, в которой нормальная деловая деятельность и активность населения нарушаются из-за страха и гнева; количественно определить вероятность мирных демонстраций и/или насильственных протестов; определить диапазон возможных ответных мер правительства, начиная от умиротворения, разрешения протестов и до угроз в адрес протестующих и действий по срыву протестов. Для решения таких вопросов потребуются модели и наборы данных с высоким разрешением

Текущие решения	**Вычислительная система**	Программное обеспечение для распределенной обработки, исполняемое на коммерческих кластерах и в более новых архитектурах и системах (например, в облаке)
	Хранилище данных	Файловые серверы (включая архивы), базы данных
	Сеть связи	Ethernet, Infiniband и аналогичные им решения
	Программное обеспечение	Специализированные программы моделирования, программное обеспечение с открытым исходным кодом и проприетарные среды моделирования. Базы данных
Характеристики больших данных	**Источник данных (распределенный/ централизованный)**	Множество источников данных: сведения о населении, местах работы, типичных маршрутах поездок, коммунальных услугах (например, электросети) и иных созданных человеком инфраструктурах, онлайн-источниках информации и социальных сетях
	Объем (количество)	Десятки терабайт новых данных ежегодно
	Скорость обработки (например, в реальном времени)	Во время социальных волнений взаимодействие между людьми и мобильность являются ключом к пониманию динамики системы. Быстрые изменения в данных, например о том, кто на кого подписан в Твиттере
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Разнообразие данных проявляется в широком диапазоне источников данных. Данные, изменяющиеся с течением времени. Объединение данных. Одной из важных проблем является объединение данных (data fusion). Как комбинировать данные из разных источников и что делать в случае отсутствия или неполноты данных? Многочисленные одновременно протекающие процессы распространения социального влияния
	Вариативность (темпы изменения)	Ввиду стохастической природы событий необходимо выполнить ряд запусков моделирования при различных параметрах модели и исходных данных, чтобы оценить диапазоны разброса результатов
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	В качестве меры обеспечения достоверности результатов проводится анализ поступающих данных в «мягком» реальном времени

Наука о больших данных (сбор, курирование, анализ, операции)	**Визуализация**	Большие наборы данных; эволюция во времени; моделирование нескольких процессов распространения социального воздействия на нескольких представлениях сети. Различные уровни детализации (например, уровень отдельного человека, микрорайона, города, штата, страны)
	Качество данных (синтаксис)	Проверки с целью обеспечения согласованности данных, на наличие порчи данных. Предварительная обработка первичных данных для использования их в моделях
	Типы данных	Очень разнообразные данные, от характеристик человека до данных о коммунальных и транспортных системах и взаимодействии между ними
	Аналитика данных	Модели поведения людей и физических инфраструктур, а также взаимодействия между ними. Визуализация результатов
Иные проблемы больших данных	Как учесть разнородные особенности сотен миллионов или миллиардов людей и модели культурных различий между странами, которые приписаны отдельным агентам? Как проверить эти большие модели? Различные типы моделей (например, с несколькими процессами распространения социального влияния): болезни, эмоции, поведение. Моделирование различных систем городской инфраструктуры, в условиях которой действуют люди. Поскольку для оценки стохастичности требуется повторное моделирование, создаются большие объемы выходных данных; соответственно, требования к их хранению	
Проблемы пользовательского интерфейса и мобильного доступа	Где и как выполнять эти вычисления? Комбинации облачных вычислений и кластеров. Как добиться максимальной эффективности вычислений — переместить данные к вычислительным ресурсам?	
Технические проблемы обеспечения безопасности и защиты персональных данных	В данном вопросе есть два аспекта. Во-первых, обеспечение неприкосновенности частной жизни и анонимности людей, сведения о которых использованы при моделировании (это, например, данные о пользователях Twitter и Facebook). Во-вторых, обеспечении защиты данных и вычислительных платформ	
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Объединение данных различных типов. В зависимости от конкретной рассматриваемой проблемы необходимо комбинировать разные наборы данных. Встает вопрос о том, каким образом обеспечить быструю разработку, проверку и валидацию новых моделей для новых приложений. Проблема выбора надлежащего уровня детализации, позволяющего схватить изучаемое явление, обеспечивая в то же время достаточно быстрое получение результатов, — то есть это вопрос о том, как сделать решение масштабируемым. Визуализация и извлечение данных с разной степенью детализации	
Дополнительная информация (гиперссылки)		

А.4.10 Вариант использования № 25: Биологическое разнообразие и европейская электронная научно-исследовательская инфраструктура LifeWatch

Название	LifeWatch — европейская электронная инфраструктура для исследований в области экологии и биологического разнообразия	
Предметная область	Научные исследования, медико-биологические науки	
Автор/организация/эл.почта	Ваутер Лос (Wouter Los), Юрий Демченко (Yuri Demchenko, y.demchenko@uva.nl), университет Амстердама	
Акторы/заинтересованные лица, их роли и ответственность	Конечные пользователи (биологи, экологи, полевые исследователи) Аналитики данных, менеджеры архивов данных, менеджеры инфраструктуры электронной науки, национальные представители стран — членов Евросоюза	
Цели	Мониторинг и изучение различных экосистем, биологических видов, их динамики и миграции	
Описание варианта использования	Целью проекта LifeWatch является обеспечение интегрированного доступ к различным данным, инструментам аналитики и моделирования, предоставленным рядом сотрудничающих с ним проектов. Он также будет предлагать данные и инструменты в составе отдельных рабочих процессов конкретным научным сообществам. Помимо этого, LifeWatch предоставит возможности для создания персонализированных «виртуальных лабораторий», также позволяя вводить/подключать новые данные и аналитические инструменты. Новые данные будут коллективно использоваться сотрудничающими с LifeWatch центрами обработки данных. Конкретные тематические исследования: мониторинг чужеродных видов, мигрирующих птиц и водно-болотных угодий	
Текущие решения	**Вычислительная система**	Полевые объекты: будут определены позднее Центр обработки данных: Типичные ресурсы сетевых параллельных вычислений и облачные ресурсы, предоставляемые национальными центрами электронной науки
	Хранилище данных	Распределенное; архивируются исторические данные и данные о тенденциях
	Сеть связи	Может потребоваться специальная выделенная или оверлейная (наложенная) сенсорная сеть
	Программное обеспечение	Веб-сервисы, грид-сервисы, реляционные базы данных
Характеристики больших данных	**Источник данных (распределенный/централизованный)**	Экологическая информация от многочисленных пунктов наблюдения и мониторинга и сенсорной сети, спутниковые изображения/информация, данные о климате и погоде, вся зарегистрированная информация. Информация от полевых исследователей
	Объем (количество)	Охватывает множество существующих наборов данных/источников. Суммарный объем данных предстоит определить

Характеристики больших данных	**Скорость обработки (например, в реальном времени)**	Данные анализируются поэтапно, динамика обработки соответствует динамике биологических и экологических процессов. Может, однако, потребоваться обработка и анализ в реальном времени в случае стихийных бедствий или техногенных катастроф. Может потребоваться обработка потоковых данных
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Разнообразие и количество задействованных баз данных и данных наблюдений в настоящее время ограничено возможностями доступных инструментов. В принципе оно является неограниченным, с учетом растущих возможностей для обработки данных с целью выявления экологических изменений, факторов/причин, эволюции видов и тенденций. См. ниже в разделе дополнительной информации
	Вариативность (темпы изменения)	Структура наборов данных и моделей может изменяться в зависимости от этапа обработки данных и поставленных задач
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	В обычном режиме мониторинга данные статистически обрабатываются для достижения надежности результатов. Для некоторых исследований в области биоразнообразия достоверность данных (их надежность и возможность им доверять) имеет критически важное значение. В случае стихийных бедствий и техногенных катастроф достоверность данных имеет критически важное значение
	Визуализация	Требуются развитая и богатая визуализация, средства визуализации высокой четкости, данные визуализации, поддерживающие: - 4D-визуализацию; - визуализацию влияния изменения параметров в (вычислительных) моделях; - сравнение полученных по модели результатов с реальными наблюдениями (многомерное)
	Качество данных (синтаксис)	Качество зависит и является следствием качества исходных данных наблюдений. Качество аналитических результатов зависит от используемых моделей и алгоритмов, которые постоянно совершенствуются. Нужна возможность повторного анализа данных с целью переоценки исходных данных наблюдений. Данные, на основе которых должны приниматься решения, контролируются человеком
	Типы данных	Данные многих типов. Реляционные данные, пары ключ–значение, сложные данные с развитой семантикой

Наука о больших данных (сбор, курирование, анализ, операции)	**Аналитика данных**	Аналитика параллельных потоков данных и аналитика данных, поступающих в потоковом режиме
Иные проблемы больших данных	Хранение и архивация данных, обмен данными и их интеграция; связь данных: от исходных данных наблюдений до обработанных данных и данных отчетности/визуализированных данных: - уникальные исторические данные; - курированные (авторизованные) эталонные данные (т. е. списки названий видов), алгоритмы, программные коды, рабочие процессы; - обработанные (вторичные) данные, являющиеся исходным материалом для других исследователей; - контроль происхождения с присвоением постоянного идентификатора (PID) данных, алгоритмов и рабочих процессов	
Проблемы пользовательского интерфейса и мобильного доступа	Требуется поддержка мобильных датчиков (например, при изучении миграции птиц) и мобильной работы исследователей (как в плане передачи информации, так и в плане поиска в каталоге) - Оснащенные инструментами полевые транспортные средства, корабли, самолеты, подводные лодки, плавучие буи; сенсорные бирки на особях - Фотографии, видео- и звукозаписи	
Технические проблемы обеспечения безопасности и защиты персональных данных	Целостность данных, ссылочная целостность наборов данных. Объединенное управление идентификацией для мобильных исследователей и мобильных датчиков Обеспечение конфиденциальности, контроля доступа и учета информации об охраняемых видах, экологической информации, космических снимков, климатической информации	
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Поддержка распределенной сенсорной сети Комбинирование и установление связей между данными различных типов; потенциально неограниченное разнообразие данных. Управление жизненным циклом данных: происхождение данных, ссылочная целостность и идентификация. Доступ и интеграция нескольких распределенных баз данных	
Дополнительная информация (гиперссылки)	Сайт европейского проекта LifeWatch-ERIC, https://www.lifewatch.eu/web/guest/home Сайт реестра веб-сервисов в области биоразнообразия BiodiversityCatalogue, https://www.biodiversitycatalogue.org/	

Примечание — Разнообразие данных, используемых в исследованиях по биоразнообразию:
Генетическое (геномное) разнообразие:
- последовательности ДНК и ДНК — баркодирование;
- метаболомические функции.

Информация о биологических видах:
- названия видов;
- сведения о наблюдениях (по времени и месту);
- отличительные признаки вида и данные об истории его развития;
- взаимоотношения хозяин–паразит;
- данных об образцах в коллекции.

Экологическая информация:
- биомасса, диаметр ствола/корня и другие физические характеристики;
- плотность населения и т. п.;
- структуры среды обитания;
- геохимические циклы углерода, азота, фосфора и т. д.

Данные об экосистеме:
- видовой состав и динамика сообщества;
- данные дистанционного и наземного наблюдения;
- потоки CO_2;
- характеристики почвы;
- цветение водорослей;
- температура, соленость, кислотность морской среды, течения и т. д.

Эксплуатация экосистемы:
- продуктивность (т. е. производство биомассы в единицу времени);
- динамика пресной воды;
- эрозия;
- буферизация тепла и влажности;
- генетические пулы.

Концепции данных:
- концептуальная основа каждого вида данных;
- онтологии;
- данные о происхождении.

Алгоритмы и потоки рабочих процессов:
- программный код и происхождение;
- протестированные рабочие процессы.

Многочисленные источники данных и информации:
- данные сбора образцов;
- наблюдения (в сделанной человеком интерпретации);
- датчики и сенсорные сети (наземные, морские, почвенных организмов), кольцевание птиц и т. д.;
- спектры воздушного и спутникового наблюдения;
- полевые и лабораторные эксперименты;
- данные радаров и лазерных радаров (лидаров);
- данные о рыболовстве и сельском хозяйстве;
- данные о заболеваниях и эпидемиях.

А.5 Облачные вычисления

А.5.1 Вариант использования № 26: Крупномасштабное глубокое обучение

Название	Крупномасштабное глубокое обучение
Предметная область	Машинное обучение, искусственный интеллект
Автор/организация/эл.почта	Адам Коутс (Adam Coates) / Стэнфордский университет (Stanford University) / acoates@cs.stanford.edu
Акторы/заинтересованные лица, их роли и ответственность	Исследователям и практикам в области машинного обучения приходится иметь дело с большими объемами данных и сложными задачами прогнозирования. Данный вариант использования поддерживает новейшие разработки в области компьютерного зрения, управления беспилотным автомобилем, распознавания речи и обработки естественного языка в научно-исследовательских и отраслевых системах
Цели	Увеличение объема наборов данных и размера моделей, с которыми способны работать алгоритмы глубокого обучения. Большие модели (например, нейронные сети с большим количеством нейронов и соединений) в сочетании с большими наборами данных все чаще показывают наилучшие результаты при выполнении эталонных задач в области зрения, речи и обработки естественного языка
Описание варианта использования	Научный сотрудник или специалист-практик в области машинного обучения хочет обучать глубокую нейронную сеть на большом (намного более 1 терабайта) массиве данных, обычно состоящем из изображений, видео-, аудиоматериалов и/или текста. Такие процедуры обучения часто требуют специфической настройки архитектуры нейронной сети, критериев обучения и предварительной обработки набора данных. Помимо вычислительных затрат, которых требуют алгоритмы обучения, чрезвычайно высока потребность в быстрой разработке прототипа и удобстве разработки

Текущие решения	**Вычислительная система**	Кластер графических процессоров с высокоскоростными соединениями (например, Infiniband, 40 гигабит в секунду)
	Хранилище данных	Файловая система Lustre объемом 100 терабайт
	Сеть связи	В кластере высокопроизводительных вычислений — Infiniband; 1-гигабитный Ethernet для сетевых соединений с внешней инфраструктурой (такой как интернет, файловая система Lustre)
	Программное обеспечение	Программное обеспечение для информационного обмена между ядрами графических профессоров и для взаимодействия на основе MPI, разработанное на факультете вычислительных наук Стэнфордского университета. Исходный код на языках C ++ / Python
Характеристики больших данных	**Источник данных (распределенный/ централизованный)**	Централизованная файловая система, содержащая один большой обучающий набор данных. Набор данных может обновляться путем включения новых учебных примеров по мере их появления
	Объем (количество)	Типичный объем наборов данных обычно составляет от 1 до 10 терабайт. С ростом вычислительных мощностей, позволяющим использовать модели гораздо большего размера, могут потребоваться наборы данных объемом 100 терабайт и более для использования в полной мере репрезентативной способности более крупных моделей. Для обучения беспилотного автомобиля могут потребоваться 100 млн изображений
	Скорость обработки (например, в реальном ремени)	Требуется намного более быстрая обработка, чем в реальном времени. Современные приложения компьютерного зрения включают обработку сотен кадров в секунду с тем, чтобы обеспечить разумное время обучения. Для требовательных приложений (таких, как управление беспилотным автомобилем) мы предвидим потребность в обработке многих тысяч изображений с высоким разрешением (6 мегапикселей и более) в секунду
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Отдельные приложения могут использовать широкий спектр данных. В настоящее время изучаются, в частности, нейронные сети, которые активно учатся на разнородных задачах, таких как обучение выполнению тегирование, разбивки на фрагменты и разбора текста, или обучение чтению по губам с использованием комбинации видео и аудиозаписей

Характеристики больших данных	**Вариативность (темпы изменения)**	Вариативность низкая. Большая часть данных поступает в постоянном темпе в потоковом режиме из общего источника. Из-за высоких вычислительных требований нагрузка на сервер может сделать передачу данных неравномерной
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Наборы данных для приложений машинного обучения часто размечаются и проверяются вручную. При подготовке чрезвычайно больших наборов данных разметка может выполняться с использованием краудсорсинга, тем самым возникает риск неоднозначных ситуаций, когда метка непонятна. Автоматизированные системы разметки по-прежнему требуют проведения человеком проверки результатов на соответствие здравому смыслу. Активной областью исследований являются умные методы построения больших наборов данных
	Визуализация	Визуализация обученных сетей является открытой областью исследований, хотя отчасти рассматривается как метод отладки. Некоторые визуальные приложения включают использование визуализации для прогнозирования (visualization predictions) на основе тестовых изображений
	Качество данных (синтаксис)	Некоторые из собранных данных (например, сжатое видео или аудио) могут быть представлены в неизвестные форматах, использовать неизвестные кодеки или оказаться поврежденными. Автоматическая фильтрация исходных данных удаляет такие данные
	Типы данных	Изображения, видео, аудио, текст (на практике, почти любые)
	Аналитика данных	В небольшой степени выполняется пакетная статистическая предварительная обработка; весь остальной анализ данных выполняется самим алгоритмом обучения
Иные проблемы больших данных	Требования к обработке даже для скромных объемов данных являются чрезвычайно высокими. Хотя обученные представления могут использовать много терабайт данных, основная проблема заключается в обработке всех данных во время обучения. Современные системы глубокого обучения способны использовать нейронные сети с более чем 10 млрд свободных параметров (аналогичных синапсам мозга), что требует триллионов операций с плавающей запятой для каждого учебного примера. Распределение этих вычислений по высокопроизводительной инфраструктуре является серьезной проблемой, для решения которой в настоящее время мы в основном используем специализированную программную систему	

Проблемы пользовательского интерфейса и мобильного доступа	После завершения обучения больших нейронных сетей, обученная сеть может быть скопирована на другие устройства с кардинально меньшими вычислительными возможностями для использования в прогнозировании в реальном времени. (Например, при управлении беспилотными автомобилями, процедура обучения выполняется с использованием высокопроизводительного кластера с 64 графическими процессорами. Результатом обучения является нейронная сеть, которая кодирует необходимые знания для принятия решений о пилотировании и обходе препятствий. Эта сеть может быть скопирована во встроенное в транспортные средства оборудование или в датчики.)
Технические проблемы обеспечения безопасности и защиты персональных данных	Нет
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Глубокое обучение имеет много общих черт с более широкой областью машинного обучения. Первостепенными требованиями являются высокая вычислительная пропускная способность (computational throughput), главным образом, для операций линейной алгебры с плотными матрицами, а также чрезвычайно высокая продуктивность. Для обеспечения лучшей производительности большинство систем глубокого обучения требуют значительных усилий по настройке на целевое приложение и, следовательно, требуют большого количества экспериментов, с вмешательством проектировщика между экспериментами. В результате ключевое значение имеет минимизация времени на проведение эксперимента и ускорение процесса разработки. Эти два требования — высокая вычислительная пропускная способность и высокая продуктивность — резко противоречат друг другу. Существуют системы высокопроизводительных вычислений (HPC), которые можно использовать для ускорения экспериментов, однако текущую программную HPC-инфраструктуру сложно использовать, что удлиняет время разработки и отладки, а во многих случаях, делает невозможными в остальном посильные в вычислительном плане приложения. В число основных компонент, необходимых для этих приложений (которые в настоящее время являются программами нашей собственной разработки), входят операции линейной алгебры над плотными матрицами, выполняемые в высокопроизводительных вычислительных системах с распределенной памятью. Если библиотеки для вычислений на одной машине или на одном графическом процессоре доступны (например, BLAS, CuBLAS, MAGMA и др.), то распределенные вычисления с плотными матрицами на графических процессорах, подобные тем, что поддерживаются BLAS или LAPACK, остаются слабо развитыми. Существующие решения (например, ScaLapack для центральных процессоров) не очень хорошо интегрированы с языками высокого уровня и требуют низкоуровневого программирования, что удлиняет время эксперимента и процесса разработки
Дополнительная информация (гиперссылки)	Недавние популярные публикации в прессе о технологии глубокого обучения: Джон Марков (John Markoff) «Ученые видят потенциал у программ глубокого обучения» (Scientists See Promise in Deep-Learning Programs), «Нью-Йорк таймс», 23 ноября 2012 г., https://www.nytimes.com/2012/11/24/science/scientists-see-advances-in-deep-learning-a-part-of-artificial-intelligence.html Джон Марков (John Markoff) «Сколько компьютеров нужно, чтобы идентифицировать кошку? 16 тысяч» (How Many Computers to Identify a Cat? 16,000), «Нью-Йорк таймс», 25 июня 2012 г., https://www.nytimes.com/2012/06/26/technology/in-a-big-network-of-computers-evidence-of-machine-learning.html Даниэла Эрнандес (Daniela Hernandez) «Теперь Вы можете задешево создать искусственный мозг, который Google обошелся в миллион долларов» (Now You Can Build Google's $1M Artificial Brain on the Cheap), Wired, 17 июня 2013 г., https://www.wired.com/2013/06/andrew-ng/

Дополнительная информация (гиперссылки)	Недавняя научная статья по использованию высокопроизводительных вычислений при глубоком обучении: Adam Coates, Brody Huval, Tao Wang, David J. Wu, Andrew Y. Ng, Bryan Catanzaro "Deep learning with COTS HPC systems", Proceedings of the 30-th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013, http://proceedings.mlr.press/v28/coates13.pdf Широко используемые учебные пособия и ссылки на ресурсы по глубокому обучению: «Руководство по машинному обучению без учителя и глубокому обучению» (Unsupervised Feature Learning and Deep Learning (UFLDL) Tutorial), Стэнфордский университет, http://deeplearning.stanford.edu/tutorial/ Сайт сообщества специалистов по глубокому обучению (архивный), http://deeplearning.net/

А.5.2 Вариант использования № 27: Организация крупномасштабных, неструктурированных коллекций сделанных потребителями фотографий

Название	Организация крупномасштабных, неструктурированных коллекций сделанных потребителями фотографий
Предметная область	Научные исследования, искусственный интеллект
Автор/организация/эл.почта	Дэвид Крендал (David Crandall), Университет Индианы, djcran@indiana.edu
Акторы/заинтересованные лица, их роли и ответственность	Исследователи в области компьютерного зрения (развитие данного направления), новостные агентства и компании — владельцы социальных сетей (способствование организации крупномасштабных коллекций фотографий), потребители (просмотр как личных, так и публичных коллекций фотографий), исследователи и другие специалисты, заинтересованные в создании дешевых трехмерных моделей (археологи, архитекторы, градостроители, дизайнеры интерьеров и т. д.)
Цели	Создание трехмерных реконструкций сцен с использованием коллекций, содержащих от миллионов до миллиардов сделанных потребителями фотографий, когда ни структура сцены, ни положение камеры заранее не известны. Использование полученных трехмерных моделей для поддержки эффективного и результативного просмотра крупномасштабных коллекций фотографий по географическому положению. Географическая привязка новых изображений осуществляется путем сопоставления с трехмерными моделями. Для каждого изображения может быть выполнено распознавание объектов
Описание варианта использования	Задача трехмерной реконструкции обычно формулируется как задача робастной нелинейной оптимизации с использованием метода наименьших квадратов, в рамках которой наблюдаемые (зашумленные) соответствия между изображениями являются ограничениями, а в число неизвестных входят 6-мерные координаты, задающие положение камеры для каждого изображения и 3-мерные координаты положения каждой точки сцены. Разреженность и большая степень шума в ограничениях обычно приводят к тому, что базовые методы оптимизации сходятся в локальные минимумы, которые далеки от реальной структуры сцены. Типичные конкретные шаги включают: (1) извлечение признаков из изображений, (2) сопоставление изображений для выявления пар с общими структурами сцены, (3) оценку первоначального решения, которое близко к структуре сцены и/или параметрам камеры, (4) непосредственную оптимизация нелинейной целевой функции. Можно отметить, что операции на шаге (1) прекрасно распараллеливаются; шаг (2) — это проблема сопоставления всех пар, обычно с использованием эвристик, которые на ранней стадии отбрасывает маловероятные пары.

Описание варианта использования	Шаг (3) выполняется нами путем дискретной оптимизации, использующей вероятностный вывод в графе (марковское случайное поле), после чего применяется робастный алгоритм Левенберга–Марквардта в непрерывном пространстве. Другие выполняют шаг (3), решая задачу шага (4) для небольшого числа изображений, а затем постепенно добавляя новые изображения и используя выходные данные последнего этапа расчетов в качестве начальных условий очередного этапа. Шаг (4) обычно выполняется с помощью алгоритма уравнивания по связкам (bundle adjustment), который является реализацией нелинейного метода наименьших квадратов, оптимизированного под конкретные структуры ограничений, возникающих в задачах трехмерной реконструкции. Решение задачи распознавания образов обычно хорошо распараллеливается, хотя обучения моделей объектов включают в себя обучение классификатора (например, метода опорных векторов) — процесс, который зачастую трудно распараллелить	
Текущие решения	**Вычислительная система**	Кластер Hadoop (около 60 узлов, 480 ядер)
	Хранилище данных	Hadoop DFS и плоские файлы
	Сеть связи	Простой Unix
	Программное обеспечение	Написанные вручную простые многопоточные инструменты (ssh и сокеты для обмена информацией)
Характеристики больших данных	**Источник данных (распределенный/централизованный)**	Общедоступные коллекции фотографий, например, на Flickr, Panoramio и др.
	Объем (количество)	Более 500 млрд фотографий на Facebook, более 5 млрд фотографий на Flickr
	Скорость обработки (например, в реальном времени)	Ежедневно в Facebook добавляется более 100 миллионов новых фотографий
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Изображения и метаданные, включая теги EXIF (фокусное расстояние, тип камеры и т. д.).
	Вариативность (темпы изменения)	Темпы поступления фотографий значительно варьируются. Например, на Facebook в Новый год выкладывается примерно в 10 раз больше фотографий, чем в другие дни. Географическое распределение фотографий подчиняется распределению «с длинным хвостом», при этом с 1000 примечательных объектов на местности (общей площадью всего около 100 кв. км) связаны более 20 % фотографий на сайте Flickr
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Важна максимальная точность, с учетом ограничений технологии компьютерного зрения
	Визуализация	Визуализация крупномасштабных трехмерных реконструкций и навигация по крупномасштабным коллекциям изображений, которые были согласованы с картами

Наука о больших данных (сбор, курирование, анализ, операции)	**Качество данных (синтаксис)**	Наблюдаемые в изображениях признаки достаточно сильно зашумлены как из-за несовершенного извлечения признаков, так и из-за неидеальных свойств конкретных изображений (дисторсия объектива, шум сенсора, добавленные пользователем к изображению эффекты и т. д.)
	Типы данных	Изображения, метаданные
	Аналитика данных	
Иные проблемы больших данных	Аналитика нуждается в постоянном мониторинге и совершенствовании	
Проблемы пользовательского интерфейса и мобильного доступа	Многие / большинство изображений захватываются мобильными устройствами. Конечная цель заключается в том, чтобы приблизить процессы реконструкции и организации коллекции к телефону и сделать возможным взаимодействие с пользователем в реальном времени	
Технические проблемы обеспечения безопасности и защиты персональных данных	Необходимо обеспечить неприкосновенность частной жизни для пользователей и цифровые права для средств массовой информации	
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Компоненты этого варианта использования, включая извлечение признаков, сопоставление признаков и крупномасштабную машину вероятностных логических выводов, появляются при решении многих или даже большинства проблем компьютерного зрения и обработки изображений, включая распознавание, разделение по глубине (stereo resolution), устранение шума в изображениях и т. д.	
Дополнительная информация (гиперссылки)	Сайт лаборатории компьютерного зрения (Computer Vision Lab) Университета Индианы, http://vision.soic.indiana.edu/projects/disco/	

А.5.3 Вариант использования № 28: Truthy — Анализ данных Твиттера

Название	Truthy — Исследование распространения информации на основе данных Твиттера
Предметная область	Научные исследования: Изучение сложных сетей и систем
Автор/организация/эл.почта	Филиппо Менцер (Filippo Menczer), Университет Индианы, fil@indiana.edu Алессандро Фламмини (Alessandro Flammini), Университет Индианы, aflammin@indiana.edu Эмилио Феррара (Emilio Ferrara), Университет Индианы, ferrarae@indiana.edu
Акторы/заинтересованные лица, их роли и ответственность	Исследования финансируются Национальным научным фондом США (National Science Foundation, NSF), Агентством по передовым военным научно-техническим проектам (Defense Advanced Research Projects Agency, DARPA), фондом Макдоннела (James S. McDonnell Foundation)
Цели	Понять, как информация распространяется по социально-техническим сетям. Обнаружение потенциально опасной информации (например, вводящих в заблуждение сообщений, скоординированных кампаний и недостоверной информации и т. п.) на ранних стадиях ее распространения
Описание варианта использования	(1) Сбор и хранение большого объема данных, поступающих непрерывным потоком от Твиттера (≈100 млн сообщений в день, темпы роста объемов данных ≈500 гигабайт данных в день); (2) Анализ таких данных в режиме времени, близком к реальному, с целью выявления аномалий, кластеризации потока, классификации сигналов и онлайн-обучения; (3) Поиск и извлечение данных, визуализация больших данных, интерактивные веб-интерфейсы к данным и общедоступные программные интерфейсы (API) для запросов к данным

Текущие решения	**Вычислительная система**	В настоящее время: собственный кластер, поддерживаемый Университетом Индианы. Критическое требование: большой кластер для хранения данных, манипулирования ими, выполнения запросов и анализа
	Хранилище данных	В настоящее время: первичные данные (с августа 2010 г.), хранящиеся в больших сжатых плоских файлах. Требуется переход на Hadoop/Indexed HBase и распределенное хранение в файловой системе HDFS. База данных в оперативной памяти под СУБД Redis как буфер для анализа в реальном времени
	Сеть связи	Требуется 10-гигабитный Infiniband
	Программное обеспечение	Hadoop, Hive, Redis для управления данными; Python/SciPy/NumPy/MPI для анализа данных
Характеристики больших данных	**Источник данных (распределенный/централизованный)**	Распределенный — с репликацией / избыточностью
	Объем (количество)	≈30 терабайт в год сжатых данных
	Скорость обработки (например, в реальном времени)	Хранение данных, выполнение запросов и анализ в масштабе времени, близком к реальному
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Схема данных предоставлена социальной сетью — источником данных. В настоящее время используются только данные Твиттера. Мы планируем расширять проект, охватив Google+ и Facebook
	Вариативность (темпы изменения)	Непрерывный поток данных в реальном времени, поступающий из каждого источника
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Для получения данных в реальном времени требуется, чтобы система 99,99 % времени находилась в рабочем состоянии. Перебои в работе могут нарушить целостность данных и уменьшить их значимость
	Визуализация	Уже существуют возможности для визуализации распространения информации, кластеризации и для динамической визуализации сети
	Качество данных (синтаксис)	Данные структурированы в стандартизированных форматах, общее качество данных чрезвычайно высокое. Мы генерируем агрегированную статистику; расширяем набор признаков и т. д., производя высококачественные производные данные
	Типы данных	Полностью структурированные данные (формат JSON), обогащенные пользовательскими метаданными данными геолокации и т. д.

Наука о больших данных (сбор, курирование, анализ, операции)	**Аналитика данных**	**Кластеризация потока:** данные агрегируются по темам, метаданным и дополнительным признакам с использованием специализированных онлайн-алгоритмов кластеризации. **Классификация:** используя многомерные временные ряды для генерации сетевых признаков, признаков пользователей, географических, контента и т. д., мы классифицируем производимую на платформе информацию. **Обнаружение аномалий:** идентификация аномальных событий в реальном времени (например, вызванных внешними факторами). **Онлайн-обучение:** применение методов машинного обучения / глубокого обучения для анализа в режиме реального времени закономерностей распространения информации, профилирования пользователей и т. д.
Иные проблемы больших данных	Обеспечение анализа в реальном времени большого объема данных. Обеспечение масштабируемой инфраструктуры для выделения по требованию ресурсов, пространства хранения и т. д., если это потребуется ввиду увеличения с течением времени объема данных	
Проблемы пользовательского интерфейса и мобильного доступа	Реализация низкоуровневых функциональных возможностей инфраструктуры хранения данных с целью обеспечения эффективного мобильного доступа к данным	
Технические проблемы обеспечения безопасности и защиты персональных данных	Твиттер раскрывает в открытом доступе данные, собранные нашей платформой. Поскольку источники данных включают в себя пользовательские метаданные (которых, как правило, недостаточно для однозначной идентификации физических лиц), необходимо реализовать определенную политику обеспечения безопасности хранения данных и защиты неприкосновенности частной жизни	
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Определение высокоуровневой схемы данных для подключения нескольких источников данных, предоставляющих аналогично структурированные данные	
Дополнительная информация (гиперссылки)	Сайт проекта Truthy Университета Индианы, http://truthy.indiana.edu/ Страница проекта Truthy на сайте Центра исследований сложных сетей и систем (Center for Complex Network and System Research, CNetS) Университета Индианы, https://cnets.indiana.edu/groups/nan/truthy/ Страница проекта «Выявление ранних признаков подстрекательства в информационных каскадах» (Detecting Early Signature of Persuasion in Information Cascades, DESPIC) на сайте Центра исследований сложных сетей и систем (Center for Complex Network and System Research, CNetS) Университета Индианы, https://cnets.indiana.edu/groups/nan/despic/	

А.5.4 Вариант использования № 29: Краудсорсинг в гуманитарных науках

Название	Краудсорсинг в гуманитарных науках как источник больших и динамических данных	
Предметная область	Гуманитарные науки, социальные науки	
Автор/организация/эл.почта	Себастьян Друде (Sebastian Drude) / Институт психолингвистики общества Макса Планка (Max Planck Institute for Psycholinguistics, Неймеген, Нидерланды) / Sebastian.Drude@mpi.nl	
Акторы/заинтересованные лица, их роли и ответственность	Ученые (социологи, психологи, лингвисты, политологи, историки и т. д.), специалисты по управлению данными и аналитики, архивы данных. Представители широкой общественности как поставщики данных и участники	
Цели	Сбор информации (введенные вручную данные, записанные мультимедийные материалы, время реакции, изображения, информация от датчиков) у многих людей и с их устройств. Это позволяет, охватить многообразные индивидуальные, социальные, культурные и лингвистические различия в нескольких измерениях (пространство, социальное пространство, время)	
Описание варианта использования	Множество различных возможных вариантов использования: собрав записи, отражающие использование языка (слов, предложений, описаний значений и т. д.), ответы на опросы, информацию о фактах культуры, описания изображений и тексты — соотнести их с другими явлениями, выявить новые культурные практики, поведение, ценности и убеждения, определить индивидуальные вариации	
Текущие решения	**Вычислительная система**	Индивидуальные системы, в которых проводится ручной сбор данных (в основном, веб-сайты)
	Хранилище данных	Традиционные сервера
	Сеть связи	Помимо ввода данных через интернет используется мало
	Программное обеспечение	Язык XML, традиционные реляционные базы данных для хранения изображений. Мультимедийных материалов (соответственно, программного обеспечения для работы с ними) пока еще немного
Характеристики больших данных	**Источник данных (распределенный/централизованный)**	Распределенный, отдельные участники передают данные через веб-страницы и мобильные устройства
	Объем (количество)	Варьируется в очень больших масштабах, от сотен до миллионов записей данных. В зависимости от типа данных, объем может варьироваться от нескольких гигабайт (текст, опросы, экспериментальные значения) до сотен терабайт (мультимедиа)
	Скорость обработки (например, в реальном времени)	Очень сильно зависит от проекта: от десятков до тысяч новых записей данных в день. Данные должны анализироваться инкрементально
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	До настоящего времени — в основном однородные небольшие наборы данных; ожидаются большие распределенные неоднородные наборы данных, которые должны быть заархивированы как первичные данные

Характеристики больших данных	**Вариативность (темпы изменения)**	Структура данных и содержание коллекций меняются на протяжении жизненного цикла данных. Изменения скорости производства данных или их характеристик в процессе сбора не являются критическими
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Возможны зашумленность данных, ненадежные метаданные, проведение выявления и предварительного отбора соответствующих данных
	Визуализация	Важна для интерпретации; какие-либо специальные методы визуализации не применяются
	Качество данных (синтаксис)	Необходима валидация. Вопросы качества записей, качества контента, спама
	Типы данных	Индивидуальные записи данных (ответы на опросы, время реакции); тексты (например, комментарии, транскрипции и т. п.); мультимедиа (изображения, аудио, видео)
	Аналитика данных	Все виды распознавания закономерностей (например, распознавание речи, автоматический анализ аудиовизуальных материалов, культурные закономерности); выявление структур (лексические единицы, лингвистические правила и т. д.)
Иные проблемы больших данных	Управление данными — метаданные, сведения о происхождении, присвоение постоянного идентификатора (PID). Курирование данных. Оцифровка существующих аудиовизуальных, фото- и документальных архивов	
Проблемы пользовательского интерфейса и мобильного доступа	Включение данных с датчиков мобильных устройств (геолокации и т. д.); Сбор данных в ходе экспедиций и полевых исследований	
Технические проблемы обеспечения безопасности и защиты персональных данных	Могут возникнуть вопросы защиты неприкосновенности частной жизни (аудиовидеозаписи, поступившие от отдельных лиц); анонимность может быть необходима, но не всегда возможна (анализ аудиовидеозаписей, небольшие речевые сообщества). Целостность архива и метаданных, обеспечение долговременной сохранности	
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Множество отдельных записей данных, поступающих от многих людей, постоянный поток вводимых данных, присвоение метаданных и т. д. Автономное использование (в сравнении онлайн-использованием), последующая синхронизация с центральной базой данных. Обеспечение авторам существенной обратной связи	
Дополнительная информация (гиперссылки)		
Примечание — Краудсорсинг только начал использоваться в более широком масштабе. С появлением мобильных устройств появился огромный потенциал для сбора большого количества данных от многочисленных физических лиц, а также для использования датчиков, имеющихся в мобильных устройствах. Эта возможность до настоящего времени в широком масштабе не опробовалась; существующие краудсорсинговые проекты обычно имеют ограниченный масштаб и основаны на веб-технологиях.		

А.5.5 Вариант использования № 30: Цифровая инфраструктура для исследований и анализа сетей и графов (CINET)

Название	Цифровая инфраструктура для исследований и анализа сетей и графов (CINET)
Предметная область	Теория и методы анализа сетей (network science)
Автор/организация/эл.почта	Группа, возглавляемая Политехническим университетом/университетом штата Вирджиния (Virginia Tech) и включающая исследователей из Университета Индианы, Университета штата Нью-Йорк в Олбани (Albany), сельскохозяйственного и технического университета штата Северная Каролина (North Carolina Agricultural and Technical State University), Университета штата в г. Джексон (штат Миссисипи), Университета центрального Хьюстона (штат Техас) и Аргоннской национальной лаборатории Министерства энергетики США. Контактные лица: Мадхав Марате (Madhav Marathe, mmarathe@vbi.vt.edu) и Кит Биссет (Keith Bisset, kbisset@vbi.vt.edu) из научной лаборатории сетевой динамики и моделирования (Network Dynamics and Simulation Science Laboratory) Института биосложности (Biocomplexity Institute, ранее Институт биоинформатики) Политехнического университета/университета штата Вирджиния (Virginia Tech)
Акторы/заинтересованные лица, их роли и ответственность	Исследователи, практики, преподаватели и студенты, интересующиеся изучением сетей
Цели	Промежуточное программное обеспечение цифровой инфраструктуры для исследований и анализа сетей и графов (CINET) предназначено для поддержки исследований и аналитики сетей. Это промежуточное ПО обеспечит исследователям, практикам, преподавателям и студентам доступ к вычислительно — аналитической среде для проведения исследований, в образовательных целях и в целях обучения. Пользовательский интерфейс предоставляет списки доступных сетей и модулей анализа сетей (реализующих алгоритмы анализа сетей). Пользователь, которым может быть исследователь в области теории сетей и ее приложений, может выбрать одну или несколько сетей и проанализировать их с помощью доступных инструментов и модулей анализа. Пользователь также может генерировать случайные сети, следуя различным моделям случайных графов. Преподаватели и студенты могут использовать CINET в ходе учебных занятий для демонстрации различных теоретических свойств графов и поведения различных алгоритмов. Пользователь также может добавить в систему сеть или модуль анализа сети. Эта функциональная возможность CINET позволяет платформе легко расти, сохраняя актуальность инструментов анализа благодаря добавлению новейших алгоритмов. Цель заключается в том, чтобы предоставить общую веб-платформу, обеспечивающую конечному пользователю бесперебойный доступ: - к различным инструментам анализа сетей и графов, таким как SNAP, NetworkX, Galib и др.; - к созданным для решения реальных задач и к синтезированным сетям; - к вычислительным ресурсам; - к системе управления данными.
Описание варианта использования	Пользователи могут запустить один или несколько вариантов структурного или динамического анализа на наборе выбранных ими сетей. Специальный предметно-ориентированный язык дает пользователям возможность проектировать гибкие высокоуровневые потоки рабочих процессов для организации более сложного анализа сетей

Текущие решения	**Вычислительная система**	Высокопроизводительный вычислительный кластер Shadowfax (DELL C6100), состоящий из 60 вычислительных узлов с 12 процессорами (Intel Xeon X5670 2,93 ГГц) в каждом узле, — в общей сложности 720 процессоров с 4 гигабайтами оперативной памяти у каждого процессора. Система с общей памятью; также используются облачные вычисления на основе Amazon Elastic Compute Cloud (Amazon EC2). Некоторые из программ и сетей могут использовать системы с одним узлом, и ввиду этого в настоящее время отображаются на грид-инфраструктуру Open Science Grid («Открытый научный грид», США, http://www.opensciencegrid.org/)
	Хранение	Общая параллельная файловая система GPFS (ныне IBM Spectrum Scale) фирмы IBM, емкостью 628 терабайт
	Сеть связи	Интернет, Infiniband. Довольно пестрая коллекция суперкомпьютерных ресурсов
	Программное обеспечение	Библиотеки для работы с графами: Galib, NetworkX. Управление распределенными потоками рабочих процессов: Simfrastructure, Базы данных, семантические веб-инструменты
Характеристики больших данных	**Источник данных (распределенный/централизованный)**	Сеть хранится в одном файле на диске, доступном для нескольких процессоров. Однако во время выполнения параллельного алгоритма сеть может быть разделена, и ее части загружаются в основную память нескольких процессоров
	Объем (количество)	Может составлять сотни гигабайт для одной сети
	Скорость обработки (например, в реальном времени)	Два типа изменений: (i) сети очень динамичны; и (ii) мы ожидаем быстрое расширение хранилища, в котором примерно через год будет храниться как минимум от тысячи до 5 тыс. сетей и методов
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Наборы данных различны: - ориентированные и неориентированные сети; - статические и динамические сети, - помеченные сети, - могут иметь динамику на этих сетях
	Вариативность (темпы изменения)	Объемы связанных с графами данных увеличиваются возрастающими темпами. Кроме того, в различных областях медико-биологических наук методы на основе графов все чаще используются для решения проблем. В этой связи мы ожидаем, что объемы данных и вычислений будут расти значительными темпами

Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Есть проблемы, связанные с асинхронными распределенными вычислениями. Современные системы спроектированы в расчете на синхронный отклик в реальном времени
	Визуализация	По мере увеличения размера исходного графа нагрузка на систему визуализации на стороне клиента сильно возрастает как с точки зрения данных, так и с точки зрения вычислений
	Качество данных (синтаксис)	
	Типы данных	
	Аналитика данных	
Иные проблемы больших данных	Для анализа больших сетей необходимы параллельные алгоритмы. В отличие от многих структурированных данных сетевые данные трудно разделять на части. Основная сложность при разделении сети заключается в том, что для эффективной работы различных алгоритмов требуются разные схемы разделения. Более того, большинство сетевых метрик имеют глобальный характер и требуют либо: i) огромного дублирования данных в разделах, либо ii) очень больших издержек на пересылку в результате требуемого перемещения данных. Для больших сетей эти трудности перерастают в серьезные проблемы. Вычислять динамику на сетях сложнее, поскольку структура сети часто взаимодействует с изучаемым динамическим процессом. CINET поддерживает большой класс операций для самых разных по структуре и размеру графов. В отличие от других систем, требующих интенсивных вычислений и работы с данными, таких, как параллельные базы данных или методы вычислительной гидродинамики, производительность вычислений на графах чувствительна к базовой архитектуре. Таким образом, уникальной задачей CINET является управление отображением рабочей нагрузки (тип графа + операция) на машину, чья архитектура и время выполнения благоприятны для системы. Манипулирование данными и ведение учета производных данных для пользователей является еще одной большой проблемой, поскольку, в отличие от корпоративных данных, отсутствуют четко определенные и эффективные модели и инструменты для унифицированного управления различными данными графов	
Проблемы пользовательского интерфейса и мобильного доступа		
Технические проблемы обеспечения безопасности и защиты персональных данных		
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Высокопроизводительные вычисления как услуга. По мере роста объемов данных, все в большем числе приложений, таких, как приложения биологических наук, приходится использовать высокопроизводительные системы. CINET может использоваться для предоставления необходимых для таких областей вычислительных ресурсов	
Дополнительная информация (гиперссылки)	Шериф Абдельхамид (Sherif Abdelhamid) и др. «CINET 2.0: Цифровая инфраструктура для исследований и анализа сетей и графов» (CINET 2.0: A CyberInfrastructure for Network Science), 2014, http://grids.ucs.indiana.edu/ptliupages/publications/CINETv2.pdf	

А.5.6 Вариант использования № 31: Измерения и оценки эффективности аналитических технологий в Национальном институте стандартов и технологий (NIST)

Название	Измерения, оценки и стандарты эффективности аналитических технологий в отделе доступа к информации NIST	
Предметная область	Измерения и стандарты эффективности аналитических технологий для заинтересованных сторон из государственного сектора, промышленности и научных кругов	
Автор/организация/эл.почта	Джон Гэрофоло (John Garofolo), Национальный институт стандартов и технологий (NIST), john.garofolo@nist.gov	
Акторы/заинтересованные лица, их роли и ответственность	Разработчики методов измерений в NIST, поставщики данных, разработчики аналитических алгоритмов, пользователи аналитических технологий для неструктурированных, полуструктурированных и разнородных данных изо всех секторов	
Цели	Ускорение разработки передовых аналитических технологий для неструктурированных, полуструктурированных и разнородных данных с помощью измерения и стандартов эффективности. Привлечение внимания сообществ по интересам к важным проблемам, стоящим перед аналитическими технологиями, создание на основе консенсуса метрик и методов измерения для оценки эффективности, определение эффективности этих метрик и методов посредством проведения их оценки в масштабах сообщества, способствующей обмену знаниями и ускоряющей прогресс, а также формирование консенсуса в отношении широкого используемых стандартов для измерения эффективности	
Описание варианта использования	Разработка, с целью создания основ и ускорения дальнейшего развития передовых аналитических технологий в областях обработки речи и языка, видеозаписей и мультимедийных материалов, биометрических изображений и неоднородных данных метрик эффективности, методов измерения и проведение оценок сообществом, а также взаимодействие аналитиков с пользователями. Обычно применяется одна из двух моделей обработки: (1) предоставить участникам тестирования тестовые данные и проанализировать выходные данные систем-участников, и (2) предоставить участникам интерфейсы к тестовой обвязке для алгоритмов, взять их алгоритмы и провести тестирование алгоритмов на внутренних вычислительных кластерах. Разработка подходов для поддержки масштабируемого тестирования на основе облачных вычислений, а также выполнение тестирования на удобство использования и полезность в системах с пользователями в контуре	
Текущие решения	**Вычислительная система**	Кластеры под Linux и OS-10; распределенные вычисления с участием заинтересованных сторон; специализированные архитектуры обработки изображений
	Хранилище данных	RAID-массивы, размещение данных на жестких дисках емкостью 1–2 терабайта, а иногда на FTP-серверах. Распределенное распространение данных с участием заинтересованных сторон
	Сеть связи	Подключение жестких дисков по волоконно-оптическому каналу; гигабитный Ethernet для межсистемного информационного обмена; общие интранет- и интернет-ресурсы NIST и сетевые ресурсы, используемые совместно с заинтересованными сторонами

Текущие решения	**Программное обеспечение**	Средства разработки PERL, Python, C/C++, Matlab, R. Разработка по принципу «снизу вверх» тестовых и измерительных приложений
Характеристики больших данных	**Источник данных (распределенный/ централизованный)**	Для целей обучения, испытаний в ходе разработки и итоговых оценок имеются большие аннотированные совокупности неструктурированного/полуструктурированного текста, аудио и видеозаписей, изображений, мультимедийных материалов и разнородные коллекции вышеперечисленного, включая аннотации о точности и достоверности
	Объем (количество)	В составе совокупности тестовых данных более 900 млн веб-страниц общим объемом 30 терабайт, 100 млн твиттов, 100 млн проверенных биометрических изображений, несколько сотен тысяч частично проверенных видеоклипов и терабайты более мелких полностью проверенных тестовых коллекций. Для будущих оценок аналитики планируются еще более крупные коллекции данных, с использованием нескольких потоков данных и сильно неоднородных данных
	Скорость обработки (например, в реальном времени)	Большинство старых методов оценки было основано на ретроспективной аналитике. В новых методах оценки основное внимание уделяется моделированию проблем анализа в реальном времени на основании данных из нескольких потоков
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Тестовые коллекции охватывают широкий спектр типов аналитических приложений, включая текстовый поиск / извлечение, машинный перевод, распознавание речи, биометрию изображений и голоса, распознавание и отслеживание объектов и людей, анализ документов, диалог между человеком и компьютером и поиск / извлечение мультимедиа. Будущие тестовые коллекции будут включать данные и приложения смешанных типов
	Вариативность (темпы изменения)	Оценка компромиссов между точностью и скоростью передачи данных, а также между числом потоков данных и их качеством
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Создание и измерение неопределенности, связанной с процессом проверки правильности данных (ground-truthing), особенно когда речь идет о людях, является сложной задачей. Использовавшиеся в прошлом ручные процессы проверки не масштабируются. Измерение эффективности комплексной аналитики, чтобы быть полезным, должно включать измерение внутренней неопределенности, а также погрешности проверки

Наука о больших данных (сбор, курирование, анализ, операции)	**Визуализация**	Визуализация результатов оценки эффективности и диагностики аналитических технологий, включая значимость и различные формы неопределенности. Оценка методов представления результатов аналитики пользователям на предмет удобства использования, полезности, эффективности и точности
	Качество данных (синтаксис)	На эффективность аналитических технологий сильное влияние оказывает качество данных, с которыми они работают, в отношении множества параметров, специфичных для предметной области и приложения. Количественная оценка этих параметров сама по себе является сложной исследовательской задачей. Смешанные источники данных и измерение эффективности аналитических потоков предъявляют еще большие требования к качеству данных
	Типы данных	Неструктурированный и полуструктурированный текст, неподвижные изображения, видео, аудио, мультимедиа (аудио + видео)
	Аналитика данных	Извлечение информации, фильтрация, поиск и резюмирование; биометрия изображения и голоса; распознавание и понимание речи; машинный перевод; обнаружение и отслеживание людей и объектов в видеозаписях; детектирование событий; сопоставление изображений и документов; обнаружение новизны в данных; разнообразная структурная / семантическая / временная аналитика и множество подтипов вышеперечисленного
Иные проблемы больших данных	Масштабирование процесса проверки на большие объемы данных, измерение внутренней неопределенности и неопределенности аннотаций, измерение эффективности для не полностью аннотированных данных, измерение эффективности аналитики для разнородных данных и аналитических потоков с участием пользователей	
Проблемы пользовательского интерфейса и мобильного доступа	Перемещение обучения, разработки и тестовых данных на сторону участников оценки либо перемещение аналитических алгоритмов участников оценки в вычислительные испытательные стенды для проведения оценки эффективности. Предоставление инструментов разработки и данных. Поддержка гибких подходов к тестированию в процессе разработки	
Технические проблемы обеспечения безопасности и защиты персональных данных	Аналитические алгоритмы, работающие с письменным языком, речью, изображениями людей и т. д., как правило, должны тестироваться на реальных или реалистичных данных. Крайне проблематично создание искусственных данных, которые бы в достаточной степени отражали вариативность реальных данных, связанных с людьми. Искусственно сформированные данные могут создавать искусственные проблемы, которые могут быть прямо или косвенно смоделированы аналитическими алгоритмами, что может приводить к завышенным показателям эффективности.	

Технические проблемы обеспечения безопасности и защиты персональных данных	Развитие самих аналитических технологий увеличивает риски, связанные с обеспечением неприкосновенности частной жизни. Будущие методы тестирования эффективности должны будут изолировать алгоритмы аналитических технологий от данных, на которых алгоритмы тестируются. Необходимы усовершенствованные архитектуры для поддержки требований по безопасности в отношении защиты чувствительных данных, обеспечивающие при этом возможность проведения содержательной оценки эффективности разработок. Совместно используемые испытательные стенды должны обеспечивать защиту интеллектуальной собственности разработчиков аналитических алгоритмов
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Масштабируемость методов тестирования эффективности аналитических технологий, подготовка исходных данных и проведение их проверки; методы и архитектуры, поддерживающие тестирование разработок; защита интеллектуальной собственности в аналитических алгоритмах, персональных данных и иной персональной информации в тестовых данных; измерение неопределенности с использованием частично аннотированных данных; формирование тестовых данных с учетом качеств, влияющих на эффективность, и оценка сложности тестового набора; оценка сложных аналитических потоков с участием ряда видов аналитики, типов данных и взаимодействия с пользователем; многочисленные неоднородные потоки данных и огромное число потоков; смеси структурированных, полуструктурированных и неструктурированных источников данных; гибкие (agile) масштабируемые подходы и механизмы тестирования разработок
Дополнительная информация (гиперссылки)	Страница отдела доступа к информации на сайте NIST, https://www.nist.gov/itl/iad

А.6 Экосистема для исследований

А.6.1 Вариант использования № 32: Консорциум федеративных сетей данных (DFC)

Название	Консорциум федеративных сетей данных (DFC)
Предметная область	Среды совместной работы
Автор/организация/эл.почта	Рейган Мур (Reagan Moore) / Университет Северной Каролины в Чапел-Хилл (University of North Carolina at Chapel Hill) / rwmoore@renci.org
Акторы/заинтересованные лица, их роли и ответственность	Научно-исследовательские проекты Национального научного фонда США: «Инициатива океанических наблюдательных станций» (архивация показаний датчиков); «Динамика во времени учебного центра» (грид-система управления данными для науки о процессах познания); проект iPlant Collaborative (геномика растений); проект электронной инженерной библиотеки Университета им. Дрекселя; и проект Института социальных наук им. Говарда Одума при Университете Северной Каролины в Чапел-Хилл (объединение грид-системы управления данными с открытым программным обеспечением для управления научно-исследовательскими данными Dataverse)
Цели	Организовать национальную инфраструктуру (среду совместной работы), которая позволит исследователям сотрудничать посредством коллективно используемых коллекций данных и общих рабочих процессов. Предоставить основанные на политике системы управления данными, поддерживающие формирование коллекций, грид-систему управления данными, электронные библиотеки, архивы и конвейеры обработки. Обеспечить механизмы интероперабельности, объединяющие существующие хранилища данных, информационные каталоги и веб-сервисы со средами совместной работы

Описание варианта использования	Содействовать совместным и междисциплинарным исследованиям посредством объединения систем управления данными, используемых федеральными органами и учреждениями США, национальными академическими научно-исследовательскими инициативами, хранилищами учреждений и участниками международного сотрудничества. Эта масштабная среда совместной работы включает петабайты данных, сотни миллионов файлов, сотни миллионов атрибутов метаданных, десятки тысяч пользователей и тысяча ресурсов хранения	
Текущие решения	**Вычислительная система**	Интероперабельность с workflow — системами управления потоками рабочих процессов (NCSA Cyberintegrator, Kepler, Taverna)
	Хранилище данных	Интероперабельность файловых систем, ленточных архивов, облачного хранения, объектно-ориентированного хранения
	Сеть связи	Совместимость с протоколами TCP/IP, параллельный TCP/IP, RBUDP, HTTP
	Программное обеспечение	Интегрированная система управления данными, основанная на использовании правил (iRODS)
Характеристики больших данных	**Источник данных (распределенный/централизованный)**	Управление данными, распределенными в международном масштабе
	Объем (количество)	Петабайты данных, сотни миллионов файлов
	Скорость обработки (например, в реальном времени)	Поддержка работы с потоками данных от датчиков, управления спутниковыми изображениями, результатами моделирования, данными наблюдений, экспериментальными данными
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Поддержка логических коллекций, пересекающих границы стран и организаций, агрегирование данных в контейнерах, метаданные и рабочие процессы как объекты
	Вариативность (темпы изменения)	Поддержка активных коллекций (изменяемые данные), управление версиями данных и использование постоянных идентификаторов
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Обеспечение надежной передачи данных, журналы аудита, отслеживание событий, периодическая проверка соответствия критериям оценки (целостность, подлинность), распределенная отладка
	Визуализация	Поддержка работы внешних систем визуализации посредством автоматизированных рабочих процессов (GRASS)
	Качество данных (синтаксис)	Обеспечение механизмов проверки качества с помощью автоматизированных процедур

Наука о больших данных (сбор, курирование, анализ, операции)	**Типы данных**	Поддержка синтаксического анализа избранных форматов (NetCDF, HDF5, Dicom) и предоставление механизмов для вызова других методов обработки данных
	Аналитика данных	Поддержка запуска рабочих процессов (workflow) анализа, отслеживания происхождения рабочих процессов, совместное использование рабочих процессов и их повторного выполнение
Иные проблемы больших данных	Предоставление стандартных наборов политик, позволяющих новому сообществу воспользоваться и развивать дальше планы управления данными, отвечающие требованиям федеральных органов исполнительной власти США	
Проблемы пользовательского интерфейса и мобильного доступа	Сбор знаний, необходимых для манипулирования данными, и применение созданных в результате процедур либо в месте хранения, либо на компьютерном сервере	
Технические проблемы обеспечения безопасности и защиты персональных данных	Объединение существующих сред аутентификации с помощью «Типового API-интерфейса программирования приложений служб защиты данных» (Generic Security Service, интерфейс GSS-API) и подключаемых модулей аутентификации (GSI, Kerberos, InCommon, Shibboleth). Менеджмент мер и средств управления доступом к файлам независимо от места хранения	
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	В настоящее время в 25 областях науки и техники имеются проекты, полагающиеся на интегрированную систему управления данными, основанную на использовании правил (iRODS): - астрофизика: проект поиска сверхновых «Аугер» (Auger); - изучение атмосферы: Научно-исследовательский центр по атмосферным наукам НАСА в Лэнгли (NASA Langley Atmospheric Sciences Center); - биология: проект филогенетики в Вычислительном центре французского Национальный институт ядерной физики и физики элементарных частиц (L'Institut national de physique nucléaire et de physique des particules, IN2P3); - климат: Национальный центр климатических данных США (National Climatic Data Center) Национального управления океанических и атмосферных исследований (National Oceanic and Atmospheric Administration, NOAA); - наука о процессах познания: «Динамика во времени учебного центра» Национального научного фонда (США); - компьютерные науки: виртуальная лаборатория для исследований в области компьютерных сетей и распределенных систем GENI (Global Environment for Network Innovations — «Глобальная среда для сетевых инноваций»); - исследование космического излучения: эксперименты на магнитном альфа-спектрометре (Alpha Magnetic Spectrometer, AMS) на Международной космической станции; - физика темной материи: проект EDELWEISS II (Experience pour DEtecter Les Wimps En Site Souterrain) французской «Подземной лаборатории в Модане» (Laboratoire Souterrain de Modane); - геологические науки: Центр моделирования климата (Center for Climate Simulations) Национального управления по аэронавтике и исследованию космического пространства (NASA); - экология: проект CEED ('Caveat Emptor' Ecological Data Repository — Хранилище экологических данных «Предостережение покупателю») Университета штата Калифорния в Сан-Диего (San Diego State University) - инженерное дело: совместный проект группы американских университетов CIBER-U (Cyber — Infrastructure — Based Engineering Repositories for Undergraduates — «Инженерные хранилища данных на основе киберинфраструктуры для студентов»);	

Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	- физика высоких энергий: проект BaBar Стенфордского центра линейных ускорителей (SLAC); - гидрология: Институт окружающей среды им. Вудсов (Institute for the Environment) Стенфордского университета, Университет Северной Каролины в Чапел-Хилл; проект Hydroshare Консорциума Университетов по развитию Гидрологических Наук (Consortium of Universities for the Advancement of Hydrologic Science, CUAHSI); - геномика: Институт Броада (Broad Institute), Институт Сенгера (Wellcome Trust Sanger Institute); - медицина: Госпиталь для больных детей (Sick Kids Hospital), г. Торонто (Канада) - нейробиология: Международная организация по координации научных исследований в области нейроинформатики (International Neuroinformatics Coordinating Facility, INCF); - физика нейтрино: эксперименты по изучению нейтрино T2K и dChooz; - океанография: «Инициатива океанических наблюдательных станций» Национального научного фонда (США); - оптическая астрономия: Национальная обсерватория оптической астрономии (National Optical Astronomy Observatory, NOAO) в США; - физика элементарных частиц: проект INDRA (Identification de Noyaux et Détection avec Résolutions Accrues — «Идентификация ядер и детектирование с повышенным разрешением») французского центра GANIL (Grand Accélérateur National d'Ions Lourds — «Большой национальный ускоритель тяжелых ионов»); - фитогенетика: проект iPlant Collaborative Национального научного фонда (США) - квантовая хромодинамика: французский Национальный институт ядерной физики и физики элементарных частиц (L'Institut national de physique nucléaire et de physique des particules, IN2P3); - радиоастрономия: проект киберинфраструктуры для радиоастрономии Cyber Square Kilometer Array (CyberSKA), проекты TREND, BAOradio; - сейсмология: Центр землетрясений Южной Калифорнии (Southern California Earthquake Center); - социальные науки: Институт социальных наук им. Говарда Одума (Odum Institute for Social Science Research), проект IPUMS Terra (ранее TerraPop)
Дополнительная информация (гиперссылки)	Сайт консорциума DataNet Federation Consortium, http://datafed.org/ Сайт системы управления данными на основе политик iRODS, https://irods.org/
Примечание — Основной проблемой является сбор знаний, необходимых для взаимодействия с результатами обработки данных предметной области. В системах управления данными на основе политик это достигается путем включения знаний в процедуры, которые контролируются с помощью политик. Эти процедуры могут автоматизировать извлечение данных из внешних хранилищ, или же выполнять рабочие процессы обработки, или же обеспечивать исполнение политик управления применительно к полученным результатам обработки данных. Типовым приложением является обеспечение выполнения планов управления данными и проверка того, что план был успешно применен.	

А.6.2 Вариант использования № 33: Discinnet-процесс

Название	Discinnet-процесс; глобальный эксперимент метаданные — большие данные	
Предметная область	Научные исследования; междисциплинарное сотрудничество	
Автор/организация/эл.почта	Филипп Журно (Philippe Journeau) / компания Discinnet Labs, Франция / phjourneau@discinnet.org	
Акторы/заинтересованные лица, их роли и ответственность	Участники: французские компании Richeact и Discinnet Labs, а также некоммерческий фонд I4OpenResearch. Ожидается создание аналогичных американских структур. Компания Richeact занимается вопросами эпистемологии фундаментальных научных исследований и опытно-конструкторских разработок; компания Discinnet Labs работает в области «Веб 2.0»	
Цели	Научная цель компании Richeact заключается в разработке прогнозной междисциплинарной модели поведения областей исследований (с соответствующей метаграмматикой). Проводится экспериментирование посредством глобального распространения в настоящее время многодисциплинарного, а позднее междисциплинарного Discinnet-процесса с помощью веб-инструментов, и новой системы для совместного научного общения и публикации. Ожидается сильное влияние на сокращение неопределенности и временных задержек между теоретическими, прикладными, технологическими исследованиями и разработками	
Описание варианта использования	В настоящее время активировано 35 кластеров; около 100 ждут, пока будут выделены дополнительные ресурсы; и потенциально еще больше кластеров открыто для сознания, управления и модерирования исследовательскими сообществами. Примеры кластеров варьируются от оптики, космологии, материаловедения, микроводорослей, здравоохранения до прикладной математики, вычислений, резины и других химические продуктов/проблем. Типичный вариант применения работает в настоящее время следующим образом: - исследователь или группа исследователей интересуется тем, как обстоят дела в определенной области исследований, и в течение минуты определяет данную область в Discinnet как «кластер»; - требуется еще от 5 до 10 минут для параметризации первых/основных измерений, в основном посредством указания единиц измерения и категорий (возможно, позднее будет выделено некоторое переменное ограниченное время для большего количества измерений). - кластер затем может быть заполнен сведениями о проектах / прогрессе либо аспирантами, либо занимающимися рецензированием специалистами и/или сообществами/исследователями. Такое решение уже имеет существенную ценность. Теперь его необходимо распространять и рекламировать, хотя максимальная ценность, как ожидается, будет исходить из междисциплинарной/проецирующей следующей версии. Полезность заключается в возможности быстро обнаружить представляющий интерес документ/проект по его результатам, и следующим шагом является построение «траектории» области исследований путем взаимодействия с различного уровня оракулами (субъектами/объектами) + из междисциплинарного контекста	
Текущие решения	**Вычислительная система**	В настоящее время на серверах хостинговой компании OVH (https://www.ovh.co.uk/) — смесь коллективно используемых и выделенных ресурсов
	Хранилище данных	На серверах хостинговой компании OVH
	Сеть связи	Должно быть реализовано в рамках желаемой интеграции с другими участниками

Текущие решения	**Программное обеспечение**	Текущая версия использует Symfony PHP, Linux, MySQL
Характеристики больших данных	**Источник данных (распределенный/ централизованный)**	В настоящее время централизованный, вскоре будет распределен по странам и даже по предоставляющим хостинг учреждениям, заинтересованным иметь собственные платформы
	Объем (количество)	Не имеет значения: это база метаданных, а не больших данных
	Скорость обработки (например, в реальном времени)	В реальном времени
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Связь с большими данными еще предстоит установить через взаимоотношения метаданные < — > большие данные, которые пока еще не реализованы (экспериментальные базы данных уже связаны с метаданными 1-го уровня)
	Вариативность (темпы изменения)	В настоящее время — в режиме реального времени; в будущем для других местоположений и распределенных архитектур — периодическая (например, в ночное время)
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Методы обнаружения общей согласованности, «дыр», ошибок, неверных утверждений известны, но их еще в основном предстоит реализовать
	Визуализация	Многомерная (гиперкуб)
	Качество данных (синтаксис)	Данные априори предполагаются правильными (прямой ввод человеком), частично реализован ряд процессов проверки и оценки
	Типы данных	«Кластерные дисплеи» (изображения), векторы, категории, PDF-файлы
	Аналитика данных	
Иные проблемы больших данных	Наша цель заключается в том, чтобы внести свой вклад в проблему генерации метаданных на основе больших данных, путем систематического согласования метаданных на многих уровнях сложности с постоянно поступающими от исследователей данными о продолжающихся процессах исследований. В настоящее время партнерство с компанией Richeact направлено на то, чтобы создать междисциплинарную модель, используя саму метаграмматику для экспериментирования и подтверждения того, что ее степень охвата эффективно преодолевает разрыв между столь сильно отличающимися уровнями сложности, как семантический и уровень самых элементарных сигналов. Пример с космологическими моделями в сравнении с промежуточными моделями различных уровней (частицы, газы, галактики, ядерный уровень, геометрия). Другие примеры с сопоставлением вычислительного и семантического уровней	
Проблемы пользовательского интерфейса и мобильного доступа	Соответствующая мощность графического интерфейса	

Технические проблемы обеспечения безопасности и защиты персональных данных	Уже доступно несколько уровней, другие запланированы, вплоть до ключей для физического доступа и изолированных серверов. Опциональная анонимность, обычные защищенные соединения
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	В течение 2011—2013 гг. мы показали на http://www.discinnet.org , что все виды областей исследования легко поддаются отображению типа Discinnet, однако для разработки и заполнения кластера требуются время и/или выделенные сотрудники
Дополнительная информация (гиперссылки)	На сайте http://www.discinnet.org уже созданные или создаваемые кластеры можно просмотреть одним щелчком мыши по названию кластера (полю), и еще больше сведений доступно в случае прохождения бесплатной регистрации [зарегистрированным в качестве исследователей или аспирантов пользователям доступно больше ресурсов (публикации)]. Максимальный уровень детализации является бесплатным для участвующих исследователей в интересах защиты сообществ, но для внешних наблюдателей он доступен за символическую плату: приветствуются все предложения по совершенствованию и улучшению обмена. Мы особенно открыты для поддержки экспериментального использования платформы аспирантурами в целях создания и изучения прошлого и будущего поведения кластеров в области геологических наук, космологии, гидрологии, здравоохранения, вычислений, энергии / аккумуляторов, моделей климата, изучения космоса и т. д.
Примечание — Мы открыты для того, чтобы способствовать широкому использованию как глобальной, так и региональной и локальной версий платформы (например, исследовательскими институтами, издателями, сетями) в интересах максимально широкого обмена данными с целью извлечения наибольшей пользы для развития науки.	

А.6.3 Вариант использования № 34: Поиск по графу для научных данных

Название	Обеспечение поиска по семантическому графу в отношении текстовых научных данных по химии, аналогичного поиску в Facebook
Предметная область	Управление информацией из научных статей
Автор/организация/эл.почта	Талапади Бхат (Talapady Bhat), Национальный институт стандартов и технологий (NIST), bhat@nist.gov
Акторы/заинтересованные лица, их роли и ответственность	Химические структуры, «Банк данных белковых структур» (Protein Data Bank, PDB), инициатива «Геном материала» (Materials Genome Initiative), инициатива «Открытое правительство», семантическая паутина, интегрированные графы данных, научные социальные сети
Цели	Создать инфраструктуру, терминологию и семантические графы данных для аннотирования и представления информации о технологиях, используя методы, основанные на корневых морфемах (root-based) и на правилах (rule-based), которые применяются главным образом в отношении индоевропейских языков, таких как санскрит и латынь
Описание варианта использования	**Шумиха вокруг социальных сетей** Интернет и социальные сети играют важную роль в современном обмене информацией. Каждый день большинство из нас используют социальные сети и для распространения, и для получения информации. Тремя специфическими особенностями многих социальных сетей, таких как Facebook, являются: - члены сообщества одновременно и поставщики данных, и их пользователи; - социальные сети хранят информацию на предопределенной «полке данных» графа данных; - основная инфраструктура социальных сетей для управления информацией в разумной степени независима от языка.

Описание варианта использования	**Какое это имеет отношение к управлению научной информацией?** За последние несколько десятилетий наука действительно эволюционировала, превратившись в общественную деятельность, охватывающую каждую страну и почти каждую семью. Мы регулярно «настраиваемся» на интернет-ресурсы для того, чтобы поделиться и найти научную информацию. **Каковы проблемы создания социальных сетей для науки?** Создание социальных сетей научной информации требует инфраструктуры, в рамках которой многие ученые из разных частей мира могут принимать участие и размещать результаты своих экспериментов. Перед созданием научной социальной сети необходимо решить некоторые вопросы, включая следующие: - Как минимизировать проблемы, связанные с местным языком и его грамматикой? - Как, не слишком много зная об управлении данными, определить «граф данных» так, чтобы размещать информацию интуитивно понятным способом? - Как найти адекватные научные данные, не проводя чересчур много времени в Интернете? **Метод** При работе с большинством языков, и особенно с санскритом и латынью, используется новый метод на основе корневых морфем для упрощения создания, когда в этом возникает потребность, хорошо выделяющихся слов для определения понятий. Некоторыми примерами такого рода из английского языка являются «био-логия» (bio-logy), «био-химия» (bio-chemistry). Примерами из санскрита являются Youga, Yogi, Yogendra, Yogesh. Примером на латыни может служить «геноцид» (genocide). Эти слова создаются по требованию на основе ставших «хорошей практикой» терминов и их способности служить узлом с самоочевидным значением в дискриминирующем графе данных	
Текущие решения	**Вычислительная система**	Облако для участия членов сообщества
	Хранилище данных	Требуется расширяемый по требованию ресурс, подходящий с учетом местоположения и требований глобальных пользователей
	Сеть связи	Нужна хорошая сеть для участия членов сообщества
	Программное обеспечение	Нужны хорошие инструменты базы данных и серверы для манипулирования графами данных
Характеристики больших данных	**Источник данных (распределенный/ централизованный)**	Распределенный ресурс с ограниченными централизованными возможностями
	Объем (количество)	Не определен. Первоначально может составлять несколько терабайт
	Скорость обработки (например, в реальном времени)	Со временем эволюционирует, чтобы соответствовать новым наилучшим практикам
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Очень сильно варьируется в зависимости от типов доступной информации о технологиях
	Вариативность (темпы изменения)	Вероятно, графы данных будут изменяться со временем в зависимости от предпочтений клиентов и наилучших практик

Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Информация о технологиях, вероятно, будет стабильной и надежной
	Визуализация	Требуется эффективная визуализация на основе графа данных
	Качество данных (синтаксис)	Ожидается, что будет хорошим
	Типы данных	Любые типы данных, от изображений до текстов, от структуры до белковых последовательностей
	Аналитика данных	Ожидается, что графы данных будут способствовать появлению надежных методов анализа данных
Иные проблемы больших данных	Эта деятельность сообщества похожая на многие социальные сети. Обеспечение устойчивых, масштабируемых, предоставляемых по требованию инфраструктур таким образом, который был бы дружественным и варианту использования, и пользователю, является реальной проблемой для любых существующих традиционных методов	
Проблемы пользовательского интерфейса и мобильного доступа	Сообществу необходим доступ к данным, поэтому доступ должен быть независимым от носителя и местоположения, и, следовательно, также требует высокой мобильности	
Технические проблемы обеспечения безопасности и защиты персональных данных	Нет, поскольку изначально усилия были сфокусированы на общедоступных данных, предоставляемых проектами с открытой платформой, такими, как инициатива «Открытое правительство», инициатива «Геном материала» и «Банк данных белковых структур»	
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Данные усилия охватывают множество локальных и сетевых ресурсов. Разработка инфраструктуры для автоматической интеграции информации из всех этих ресурсов с использованием графов данных является сложной задачей, которую мы стараемся решить	
Дополнительная информация (гиперссылки)	Пресс-релиз «Фейсбук для молекул» (Facebook for molecules) Американского института физики (American Institute of Physics), 18 июля 2013 г., https://www.eurekalert.org/pub_releases/2013-07/aiop-ffm071813.php Страница поиска по Банку данных белковых структур и веб-сервиса поиска и визуализации химических структур Chem-BLAST на сайте Национального института стандартов и технологий (США), https://randr.nist.gov/chemblast/default.aspx	

Примечание — Во многих отчетах, в том числе в недавнем отчете по проекту «Геном материала» (Materials Genome Initiative), отмечается, что исключительно нисходящие решения, облегчающие обмен данными и интеграцию, нежелательны в случае междисциплинарных усилий. В то же время подход «снизу вверх» может быть хаотичным. По этой причине существует потребность в сбалансированном сочетании двух подходов с целью поддержки простых в использовании методов создания, интеграции и обмена метаданными.

Эта проблема очень похожа на проблему, с которой сталкиваются разработчики языка на начальной стадии. Одними из успешных подходов, используемых во многих известных языках, являются методы на основе корневых морфем и на основе правил, которые формируют основу для создания, когда это требуется, новых слов для общения. В этом подходе метод «сверху вниз» используется для выделения ограниченного числа многократно используемых слов, называемых «корневыми морфемами», путем изучения существующих передовых практик построения терминологии. Затем корневые морфемы комбинируются с использованием нескольких «правил» для создания новых терминов, на этапе, выполняемом снизу вверх.

Y (uj) («присоединяться»), O («создатель», «Бог», «мозг»), Ga («движение», «посвящение») — ведет к формированию слова «йога», используемого в санскрите, и английском языке.

Geno («род» на греческом) — cide (от латинского occidendum — «убийство») = genocide («геноцид», убийство по расовым мотивам).

Bio-technology («биотехнология») — английский, латынь.

Red-light, red-laser-light — английский.

Пресс-релиз Американского института физики об этом подходе см. по адресу https://www.eurekalert.org/pub_releases/2013-07/aiop-ffm071813.php

Наши усилия по разработке автоматизированных методов, сочетающих подходы на основе корневых морфем и на основе правил (проект Chem-BLAST, см. https://randr.nist.gov/chemblast/default.aspx) для выявления и использования лучших практик, различающих термины при создании семантических графов данных для науки, начались почти десять лет тому назад с базы данных химических структур. Эта база данных содержит миллионы структур, полученных из используемых во всем мире «Банка данных белковых структур» и базы данных химических соединений и смесей PubChem, используемых по всему миру. Впоследствии мы расширили наши усилия и занялись созданием на основе корневых морфем терминов для текстовых данных, связанных с изображениями клеток. В данной работе мы используем несколько простых правил для определения и расширения терминов, основанных на хорошей практике, идентифицируемой путем изучения миллионов популярных вариантов использования, выбранных из более чем сотни биологических онтологий.

В настоящее время мы работаем над распространением этого метода на публикации, представляющие интерес для инициативы «Геном материала», движения «Открытое правительство», а также для «Сети интегрированных знаний NIST — EditorialNet» (NIKE) — архива публикаций американского Национального института стандартов и технологий (NIST). Эти усилия являются частью деятельности рабочей группы «Справочник стандартов метаданных» (Metadata Standards Directory) Альянса научных данных (Research Data Alliance), см. www.rd-alliance.org/filedepot_download/694/160 и https://www.rd-alliance.org/plenary-meetings/second-plenary/poster-session-rda-2nd-plenary-meeting.html

А.6.4 Вариант использования № 35: Анализ больших объемов данных, получаемых в экспериментах на синхротроне

Название	Анализ больших объемов данных, получаемых в экспериментах на синхротроне	
Предметная область	Научные исследования (биология, химия, геофизика, материаловедение и др.)	
Автор/организация/эл.почта	Эли Дарт (Eli Dart) / Национальная лаборатория имени Лоуренса в Беркли, США (LBNL), eddart@lbl.gov	
Акторы/заинтересованные лица, их роли и ответственность	Научно-исследовательские группы из различных научных дисциплин (см. выше)	
Цели	Использование различных экспериментальных методов для определения структуры, состава, поведения и других характеристик образца, имеющих отношение к соответствующему научному исследованию	
Описание варианта использования	Образцы подвергаются воздействию рентгеновского излучения в различных конфигурациях, в зависимости от эксперимента. Данные собираются детекторами, которые фактически представляют собой высокоскоростные цифровые фотокамеры. Затем данные анализируются с целью восстановления вида исследуемого образца или процесса. Реконструированные изображения используются учеными для анализа	
Текущие решения	**Вычислительная система**	Диапазон вычислений варьируется от отдельных компьютеров для анализа до вычислительных систем с высокой пропускной способностью в вычислительных центрах
	Хранилище данных	Локальное временное хранение на объекте от одного до 40 терабайт данных на серверах данных под Windows или Linux; более 60 терабайт на жестком диске и более 300 терабайт на ленте в Национальном научно-исследовательском вычислительном центре энергетических исследований Министерства энергетики США (NERSC)

Текущие решения	**Сеть связи**	Ethernet 10 гигабит/с на объекте, 100 гигабит/с связь с NERSC
	Программное обеспечение	Для анализа данных используется различное программное обеспечение, как коммерческое, так и с открытым исходным кодом, например: - Octopus (см. https://octopusimaging.eu/) для томографической реконструкции; - Avizo и FIJI (дистрибутив открытого программного обеспечения ImageJ, см. http://fiji.sc/) для визуализации и анализа. Передача данных осуществляется посредством физического перемещения портативных носителей информации (что сильно ограничивает производительность); либо с использованием высокопроизводительного протокола GridFTP в реализации компании Globus Online, и систем управления потоками рабочих процессов, таких как программная инфраструктура с открытым исходным кодом SPADE (Support for Provenance Auditing in Distributed Environments — «Поддержка аудита происхождения в распределенных средах»)
Характеристики больших данных	**Источник данных (распределенный/ централизованный)**	Централизованный (фотокамера высокого разрешения на объекте). На объекте имеется несколько каналов отвода излучения к экспериментальным установкам с высокоскоростными детекторами
	Объем (количество)	От 3 до 30 гигабайт на образец, до 15 образцов в день
	Скорость обработки (например, в реальном времени)	Анализ в почти реальном времени необходим для проверки параметров эксперимента (для этого может использоваться низкое разрешение). Автоматизация анализа могла бы резко повысить продуктивность научных исследований
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Многие детекторы выдают однотипные данные (например, файлы формата TIFF), но контекст эксперимента сильно варьируется
	Вариативность (темпы изменения)	Возможности детекторов быстро растут, практически подчиняясь закону Мура. Площадь детектора экспоненциально увеличивается (1000 x 1000, 2000 x 2000, 4000 x 4000,…), а частота снятия показаний экспоненциально растет (1 Гц, 10 Гц, 100 Гц, 1 кГц,…). Ожидается, что в течение двух лет скорость передачи данных с одного детектора достигнет 1 гигабайта в секунду

Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Анализ в почти реальном времени необходим для проверки параметров эксперимента. Во многих случаях раннее проведение анализа может резко повысить продуктивность эксперимента, обеспечивая раннюю обратную связь. Это подразумевает повседневную доступность вычислений с высокой пропускной способностью, высокопроизводительную передачу данных и высокоскоростное хранилище
	Визуализация	Визуализация является ключом к широкому спектру экспериментов на всех экспериментальных объектах — генераторах излучения
	Качество данных (синтаксис)	Качество и точность данных имеют решающее значение (особенно в связи с тем, что время работы генератора излучения ограничено, а повторный эксперимент часто невозможен)
	Типы данных	Многие экспериментальные установки производят графические данные (например, файлы формата TIFF)
	Аналитика данных	Объемная реконструкция, идентификация характеристик и т. д.
Иные проблемы больших данных	Быстрое увеличение возможностей фотокамер, необходимость автоматизации передачи данных и анализа в почти реальном времени	
Проблемы пользовательского интерфейса и мобильного доступа	Становится необходимой передача данных в крупномасштабные вычислительные центры из-за вычислительной мощности, необходимой для проведения анализа в разумные, с точки зрения эксперимента, сроки. Из-за большого количества каналов отвода излучения к экспериментальным установкам, например, 39 у синхротрона Advanced Light Source (ALS) Национальной лаборатории имени Лоуренса в Беркли, США (LBNL), совокупное производство данных, вероятно, значительно возрастет в ближайшие годы	
Технические проблемы обеспечения безопасности и защиты персональных данных	Варьируются в зависимости от проекта	
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Ожидается значительная потребность в обобщенной инфраструктуре для анализа гигабайт данных в секунду, поступающих от множества детекторов на ряде экспериментальных установок. В настоящее время существуют прототипы, однако развертывание для целей промышленной эксплуатации потребует дополнительных ресурсов	
Дополнительная информация (гиперссылки)	Сайт синхротрона ALS (Advanced Light Source) Национальной лаборатории имени Лоуренса в Беркли, США (LBNL), https://als.lbl.gov/ Сайт синхротрона APS (Advanced Photon Source) Аргоннской национальной лаборатории (Argonne National Laboratory), США, https://www.aps.anl.gov/ Сайт рентгеновского лазера на свободных электронах LCLS (Linac Coherent Light Source) в Национальной ускорительной лаборатории SLAC (SLAC National Accelerator Laboratory) Стэнфордского университета, США, https://portal.slac.stanford.edu/sites/lcls_public/Pages/Default.aspx (исторический), https://lcls.slac.stanford.edu/ (действующий)	

А.7 Астрономия и физика

А.7.1 Вариант использования № 36: «Каталинский обзор оптических переходных процессов в режиме реального времени» (CRTS)

Название	«Каталинский обзор оптических переходных процессов в режиме реального времени» (CRTS) — цифровой, панорамный, синоптический обзор неба
Предметная область	Научные исследования: астрономия
Автор/организация/эл.почта	Станислав Джорговский (Stanislav G. Djorgovski) / Калифорнийский технологический институт (Caltech) / george@astro.caltech.edu
Акторы/заинтересованные лица, их роли и ответственность	Исследовательская группа обзора: обработка данных, контроль качества, анализ и интерпретация, публикация и архивирование. Участники сотрудничества — ряд научно-исследовательских групп по всему миру: дальнейшая работа по анализу и интерпретации данных, дополнительные наблюдения и публикационная деятельность. Сообщество пользователей: все вышеперечисленное. Мировое астрономическое сообщество: дальнейшая работа по анализу и интерпретации данных, дополнительные наблюдения и публикационная деятельность
Цели	В рамках обзора проводятся исследования меняющейся Вселенной в диапазоне видимого света, в масштабах времени, варьирующихся от минут до лет, путем поиска переменных и транзиентных (непостоянных, преходящих) источников. Обзор позволяют выявить широкий спектр астрофизических объектов и явлений, включая различные типы космических взрывов (например, сверхновых), переменные звезды, явления, связанные с аккрецией на массивные черные дыры (примером служат активные галактические ядра) и их релятивистские потоки частиц и энергий, звезды с большим собственным движением и т. д.
Описание варианта использования	Данные поступают с трех телескопов (два в Аризоне, США и один в Австралии), и в ближайшем будущем ожидается подключение дополнительных телескопов (в Чили). Первоначальной мотивацией проекта являлся поиск околоземных и потенциально представляющих для Земли угрозу астероидов, финансируемый Национальным управлением по аэронавтике и исследованию космического пространства США (NASA) и проводимый группой из Лаборатории изучения Луны и планет в Университете Аризоны, США (LPL) — это был базовый проект «Каталинский обзор неба» (CSS). CRTS делится данными в целях изучения меняющейся Вселенной за пределами Солнечной системы, эту работу возглавляет группа из Калифорнийского технологического института. С использованием нескольких проходов обозревается приблизительно 83 % всего неба (исключены переполненные области вблизи плоскости Галактики и небольшие области вблизи небесных полюсов). Данные предварительно обрабатываются на телескопе, а затем передаются в Лабораторию изучения Луны и планет в Университете Аризоны, США (LPL) и Калифорнийский технологический институт (Caltech) для дальнейшего анализа, распространения и архивирования. Данные обрабатываются в режиме реального времени, а обнаруженные транзиентные события публикуются с использованием различных электронных механизмов распространения, без использования проприетарного периода отсрочки до широкого распространения данных (CRTS использует политику полностью открытых данных). Дальнейший анализ данных включает автоматическую и полуавтоматическую классификацию обнаруженных транзиентных событий, дополнительные наблюдения с использованием других телескопов, научную интерпретацию и публикацию. В этом процессе интенсивно используются архивные данные из широкого спектра географически распределенных ресурсов, объединенных структурой Виртуальной обсерватории (VO).

Описание варианта использования	Кривые блеска (истории потоков) накапливаются для ≈ 500 миллионов источников, выявленных в ходе обзора. Для каждого из них в среднем имеется несколько сотен точек данных, охватывающих период до 8 лет, и их объемы продолжают расти. Эти данные предоставляются сообществу из архивов Калифорнийского технологического института, и вскоре — также из архивов Межвузовского центра астрономии и астрофизики (Inter-University Centre for Astronomy and Astrophysics, IUCAA), Индия. Это беспрецедентный по своим масштабам набор данных для исследования измерения времени в астрономии, с точки зрения периода наблюдений, покрытия неба и глубины. Проект CRTS служит научным и методологическим испытательным стендом и является предшественником предстоящих более крупных обзоров, которые будут проводиться, в особенности, Большим синоптическим обзорным телескопом в Обсерватории имени Веры Рубин, Чили (LSST), который, как ожидается, войдет в эксплуатацию в 2020-х гг.	
Текущие решения	**Вычислительная система**	Оборудование и компьютеры для обработки данных: несколько настольных компьютеров и небольших компьютеров серверного класса, хотя для некоторых задач анализа данных требуется более мощное оборудование Данный проект не столько требователен к вычислительным ресурсам, сколько к процессу обработки данных
	Хранилище данных	Несколько многотерабайтных и десятки терабайтных серверов
	Сеть связи	Стандартные интернет-соединения между университетами
	Программное обеспечение	Специализированные «конвейер» обработки данных и программное обеспечение для анализа данных, работающее под ОС Linux. Некоторые архивы располагаются на машинах под ОС Windows, на которых используется СУБД MS SQL
Характеристики больших данных	**Источник данных (распределенный/централизованный)**	Распределенными являются: 1) данные обзора, поступающие с трех (впоследствии — с большего числа) телескопов; 2) архивные данные из различных ресурсов, объединенных структурой Виртуальной обсерватории; 3) данные последующих наблюдений с отдельных телескопов
	Объем (количество)	В ходе обзора создается примерно до 0,1 терабайта данных в ясную ночь, а суммарный объем фондов данных составляет в настоящее время около 100 терабайт. Данные последующих дополнительных наблюдений составляют не более нескольких процентов от этого объема. Объем архивных данных во внешних (подключенных к структуре Виртуальной обсерватории) архивах измеряется петабайтами, но используется только небольшая их часть

Характеристики больших данных	**Скорость обработки (например, в реальном времени)**	До ≈ 0,1 терабайта за ночь первичных данных обзора
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Первичные данные обзора представлены в виде изображений, которые обрабатываются с целью каталогизации источников (представлены в таблицах баз данных) и построения временных рядов для отдельных объектов (кривые блеска). Данные последующих дополнительных наблюдений представлены в виде изображений и спектров. Архивные данные из грида данных Виртуальной обсерватории включают все вышеперечисленное из широкого спектра источников, полученное в различных диапазонах длин волн
	Вариативность (темпы изменения)	Ежедневный трафик данных колеблется в диапазоне от ≈ 0,01 до ≈ 0,1 терабайт в день, не включая крупномасштабную передачу данных между основными архивами (Caltech, Университет Аризоны и Межвузовский центр астрономии и астрофизики IUCAA в Индия).
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	На всех этапах процесса реализованы различные механизмы контроля качества, включающие автоматизированные средства и инспектирование человеком
	Визуализация	Используются стандартные пакеты визуального отображения и построения графиков. Мы исследуем механизмы визуализации для пространств параметров данных высокой размерности
	Качество данных (синтаксис)	Качество варьируется в зависимости от условий наблюдений, и оценивается автоматически: оценки погрешности делаются для всех соответствующих величин
	Типы данных	Изображения, спектры, временные ряды, каталоги
	Аналитика данных	Существует большое количество разнообразных инструментов анализа астрономических данных, а также большое количество специализированных инструментов и программного обеспечения, часть которых является самостоятельными исследовательскими проектами
Иные проблемы больших данных	Разработка инструментов машинного обучения для изучения данных, и в частности для автоматической классификации транзиентных событий в режиме реального времени, с учетом немногочисленности и неоднородности данных. Эффективная визуализация многомерных пространств параметров является для всех нас серьезной проблемой	
Проблемы пользовательского интерфейса и мобильного доступа	В настоящее время не является существенным ограничением	

Технические проблемы обеспечения безопасности и защиты персональных данных	Нет
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Обработка и анализ в реальном времени больших потоков данных, поступающих из распределенной сенсорной сети (в данном случае, с телескопов), когда требуется выявить, охарактеризовать и отреагировать на представляющие интерес транзиентные события в (почти) реальном времени. Использование сильно распределенных архивных ресурсов данных (в данном случае, архивов, объединенных в рамках Виртуальной обсерватории) для анализа и интерпретации данных. Автоматическая классификация с учетом немногочисленности и разнородных данных, динамически эволюционирующая во времени по мере поступления большего количества данных; и принятия решений о проведении дополнительных исследований в условиях немногочисленности и ограниченности ресурсов (в данном случае, проведение последующих наблюдений с использованием других телескопов)
Дополнительная информация (гиперссылки)	Страница проекта CRTS на сайте Калифорнийского технологического института, http://crts.caltech.edu/ Страница проекта CSS на сайте Лаборатории изучения Луны и планет в Университете Аризоны, США (LPL), https://catalina.lpl.arizona.edu/ Более подробные сведения об обзорах неба, их прошлом, настоящем и будущем, а также обзор проблем классификации см., например, в статье S.G. Djorgovski et al «Flashes in a Star Stream: Automated Classification of Astronomical Transient Events», IEEE eScience 2012 conference, October 2012, IEEE Press, https://arxiv.org/abs/1209.1681
Примечание — Проект CRTS можно рассматривать как хорошего предшественника для флагманского проекта астрономии, Большого синоптического обзора неба (Large Synoptic Sky Survey) с использованием Большого синоптического обзорного телескопа в Обсерватории имени Веры Рубин, Чили (LSST), https://www.lsst.org/ , который сейчас строится. Его ожидаемые объемы передачи данных (от 20 до 30 терабайт в ясную ночь, десятки петабайт за время проведения обзора в целом) соответствуют росту по закону Мура от текущих скоростей и объемов данных проекта CRTS, и многие технические и методологические проблемы очень похожи. Это также хороший вариант применения для интеллектуального анализа данных в реальном времени и выделения знаний в больших потоках данных, в условиях распределенности источников данных и вычислительных ресурсов	

А.7.2 Вариант использования № 37: Космологический обзор неба и моделирование

Название	Проект Министерства энергетики США анализа экстремально больших данных космологических обзоров неба и моделирования
Предметная область	Научные исследования: астрофизика
Автор/организация/эл.почта	Салман Хабиб (Salman Habib), Аргоннская национальная лаборатория (Argonne National Laboratory); Эндрю Конноли (Andrew Connolly), Университет Вашингтона, США
Акторы/заинтересованные лица, их роли и ответственность	Ученые, изучающие темную материю, темную энергию и структуру ранней Вселенной
Цели	Прояснить природу темной материи, темной энергии и инфляции, дав ответ на некоторые из самых волнующих, озадачивающих и проблемных вопросов из тех, что стоят перед современной физикой. Появляющиеся неожиданные результаты измерений указывают на потребность в физике, выходящей за рамки успешной «стандартной модели» физики элементарных частиц

Описание варианта использования	Данное исследование требует тесного взаимодействия между «большими данными» из экспериментов и моделирования, а также огромных объемов вычислений. Сплав всего этого позволит: 1) предоставить прямые методы и средства для космологических открытий, требующие тесной связи между теорией и наблюдениями («прецизионная космология»); 2) создать ключевой по важности «инструмент выявления» для работы с большими наборами данных, генерируемыми сложными инструментами; 3) производить и обмениваться результатами высокоточного моделирования, которые необходимы для понимания и контроля системы классификации (systematics), особенно астрофизической	
Текущие решения	**Вычислительная система**	Время вычислений: 24 млн часов (NERSC / Berkeley Lab), 190 млн часов (ALCF / Argonne), 10 млн часов (OLCF / Oak Ridge)
	Хранилище данных	180 терабайт (NERSC / Berkeley Lab)
	Сеть связи	На данный момент соединения с национальными лабораториями по высокоскоростной сети ESnet (Energy Sciences Network) Министерства энергетики США являются адекватными
	Программное обеспечение	MPI, OpenMP, C, C++, F90, FFTW, пакеты визуализации, Python, FFTW, Numpy, Boost, OpenMP, ScaLAPCK, СУБД PSQL и MySQL, Eigen, Cfitsio, http://astrometry.net/ и Minuit2
Характеристики больших данных	**Источник данных (распределенный/ централизованный)**	Данные наблюдений будут получены в ходе обзоров «Темная энергия» (Dark Energy Survey, DES) и Zwicky Transient Factory в 2015 г.; «Большой синоптический обзор неба» (Large Synoptic Sky Survey) начнется с 2019 г. Данные моделирования будут создаваться в суперкомпьютерных центрах Министерства энергетики США
	Объем (количество)	Обзоры DES: 4 петабайта/год, ZTF: 1 петабайт/год, LSST: 7 петабайт/год. Моделирование — более 10 петабайт в 2017 г.
	Скорость обработки (например, в реальном времени)	Обзор LSST: 20 терабайт в день
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	1) Первичные данные обзоров неба. 2) Обработанные данные изображений. 3) Данные моделирования
	Вариативность (темпы изменения)	Наблюдения проводятся по ночам; вспомогательное моделирование проводится в течение года, однако данные могут поступать спорадически в зависимости от доступности ресурсов

Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	
	Визуализация	Интерпретация результатов детального моделирования требует развитых методов и средств анализа и визуализации. Ограничения подсистемы ввода/вывода суперкомпьютера вынуждают исследователей изучать идею анализа «по месту» взамен методов постобработки
	Качество данных (синтаксис)	
	Типы данных	Данные наблюдений в виде изображений должны быть обработаны и полученные результаты сопоставлены с физическими величинами, полученными по итогам моделирования. Должны быть составлены смоделированные карты неба, соответствующие форматам наблюдений
	Аналитика данных	
Иные проблемы больших данных	Хранение, коллективное использование и анализ петабайт данных наблюдений и моделирования	
Проблемы пользовательского интерфейса и мобильного доступа	Обзор LSST будет производить 20 терабайт данных в день. Эти данные должны быть заархивированы и сделаны доступными исследователям во всем мире	
Технические проблемы обеспечения безопасности и защиты персональных данных		
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)		
Дополнительная информация (гиперссылки)	Страница, Большого синоптического обзорного телескопа в Обсерватории имени Веры Рубин, Чили (LSST), https://www.lsst.org/lsst Сайт Национального научно-исследовательского вычислительного центра энергетических исследований Министерства энергетики США (NERSC), https://www.nersc.gov/ Презентация к докладу Салмана Хабиба (Salman Habib, Аргоннская национальная лаборатория) на тему «Текущие и будущие вычислительные потребности вычислительной космологии» (Present and Future Computing Requirements for Computational Cosmology), 27—28 ноября 2012 г., https://www.nersc.gov/assets/Uploads/HabibcosmosimV2.pdf Страница программ в области физики высоких энергий сайта Управления науки Министерства энергетики США, https://www.energy.gov/science/hep/high-energy-physics	

А.7.3 Вариант использования № 38: Большие данные космологических обзоров неба

Название	Большие данные космологических обзоров неба	
Предметная область	Научные исследования: Границы космоса	
Автор/организация/эл.почта	Питер Ньюджент (Peter Nugent) / Национальная лаборатория имени Лоуренса в Беркли, США (LBNL), penugent@lbl.gov	
Акторы/заинтересованные лица, их роли и ответственность	Обзор неба «Темная энергия» (Dark Energy Survey, DES), «Спектроскопическая установка для исследования темной энергии» (Dark Energy Spectroscopic Instrument, DESI), Большой синоптический обзорный телескоп в Обсерватории имени Веры Рубин, Чили (LSST), Аргоннская национальная лаборатория (Argonne National Laboratory, ANL), Брукхейвенская национальная лаборатория (BNL), Национальная ускорительная лаборатория имени Ферми, США (FNAL/Fermilab), Национальная лаборатория имени Лоуренса в Беркли, США (LBNL), Национальная ускорительная лаборатории SLAC (SLAC National Accelerator Laboratory) Стэнфордского университета: — Создание установок/телескопов, проведение обзора и выполнение космологического анализа	
Цели	Обеспечить возможность обработки фотометрических данных в режиме реального времени для обнаружения и дальнейшего наблюдения сверхновых звезд, а также обработки больших объемов данных наблюдений (совместно с данными моделирования) с целью уменьшения систематических погрешностей в измерении космологических параметров посредством изучения барионных акустических осцилляций, подсчета галактических кластеров и измерений методом слабого гравитационного линзирования	
Описание варианта использования	При выполнении обзора «Темная энергия» (Dark Energy Survey, DES), данные с вершины горы передаются по микроволновой связи в чилийский город Ла Серена (La Serena). Оттуда по оптическим каналам связи они поступают в американский Национальный центр компьютерных приложений (National Center for Computing Applications, NCSA) и Национальный научно-исследовательский вычислительный центр энергетических исследований Министерства энергетики США (NERSC) для хранения и «редуцирования». Применяются конвейеры «вычитания» с использованием существующих изображений, с целью найти новые оптические транзиенты при помощи алгоритмов машинного обучения. Затем проводится идентификация и каталогизация галактик и звезд как на отдельных изображениях, так и на сериях изображений; и, наконец, их характеристики измеряются и сохраняются в базе данных	
Текущие решения	**Вычислительная система**	Linux-кластер, сервер реляционной СУБД Oracle, большие машины памяти, стандартные интерактивные хосты Linux. Для моделирования — ресурсы высокопроизводительных вычислений
	Хранилище данных	Реляционная СУБД Oracle, терминальный клиент psql (PostgreSQL interactive terminal) для работы с объектно-реляционной СУБД PostgreSQL, а также файловые системы GPFS и Luster и ленточные архивы
	Сеть связи	Предоставляется Национальным научно-исследовательским вычислительным центром энергетических исследований Министерства энергетики США (NERSC)

Текущие решения	**Программное обеспечение**	Стандартное астрофизическое программное обеспечение для обработки («редуцирования») данных, а также сценарии-обертки (wrapper scripts) Perl / Python, планирование Linux Cluster; и сопоставление с большими объемами данных моделирования с помощью таких методов, как разложение Холецкого
Характеристики больших данных	**Источник данных (распределенный/централизованный)**	Распределенный, обычно данные делятся на данные наблюдений и результаты моделирования
	Объем (количество)	Телескоп LSST создаст 60 петабайт графических данных и 15 петабайт данных каталога; и также будет создан соответственно большой (или даже больший) объем данных моделирования. В общей сложности за ночь будет создаваться более 20 терабайт данных
	Скорость обработки (например, в реальном времени)	Каждую ночь необходимо будет обрабатывать 20 терабайт данных в режиме, как можно более близком к реальному времени, чтобы максимизировать количество научных данных о сверхновых звездах
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Хотя данные в виде изображений схожи, анализ, выполняемый в интересах четырех различных типов космологических измерений и для сопоставления с данными моделирования, сильно различается
	Вариативность (темпы изменения)	Погодные условия и облачность могут кардинально изменить как качество, так и количество данных
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Астрофизические данные — это кошмар для статистиков, поскольку погрешности при выполнении конкретных измерений варьируются от ночи к ночи, в дополнение к крайней непредсказуемости частоты наблюдаемых явлений. Кроме того, возможности проведения практически всех космологических измерений ограничены, и, как следствие, как можно лучшее понимание собранных данных имеет наивысший приоритет в рамках каждого обзора неба
	Визуализация	Интерактивная скорость пользовательского веб-интерфейса при работе с большими наборами данных остается проблемой. Обязательной является возможность выполнять основные виды запросов и просмотр данных с целью поиска новых транзиентов, а также для мониторинга качества обзора. Возможность скачивать большие объемы данных для автономного анализа является еще одним требованием к системе. Также необходима способность комбинировать результаты моделирования и данные наблюдений

Наука о больших данных (сбор, курирование, анализ, операции)	**Качество данных (синтаксис)**	Понимание систематических погрешностей в данных наблюдений является необходимым условием успешности космологических измерений. Для будущих обзоров огромной проблемой является уменьшение погрешностей в результатах моделирования ниже этого уровня
	Типы данных	См. выше подпункт «Разнообразие»
	Аналитика данных	
Иные проблемы больших данных	Для понимания ограничений в данных моделирования будут полезны новые статистические методы. Часто случается, что не хватает компьютерного времени для выполнения желаемого количества объемов моделирования, и для закрытия пробелов приходится полагаться на эмуляторы. Необходимы методы для выполнения разложения Холецкого для тысяч моделирований с матрицами порядка миллиона по каждой стороне	
Проблемы пользовательского интерфейса и мобильного доступа	Одновременное выполнение анализа как данных моделирования, так и данных наблюдений	
Технические проблемы обеспечения безопасности и защиты персональных данных	Никаких особых проблем нет. Данные либо являются общедоступными, либо для доступа к ним требуется стандартный вход с паролем	
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Интересным направлением будущих исследований могут стать параллельные базы данных, способные работать с данными изображений	
Дополнительная информация (гиперссылки)	Страница Большого синоптического обзорного телескопа в Обсерватории имени Веры Рубин, Чили (LSST), https://www.lsst.org/lsst Сайт «Спектроскопической установки для исследования темной энергии» (Dark Energy Spectroscopic Instrument, DESI) Министерства энергетики США, https://www.desi.lbl.gov/ Сайт обзора неба «Темная энергия» (Dark Energy Survey, DES), https://www.darkenergysurvey.org/	

А.7.4 Вариант использования № 39: Анализ данных Большого адронного коллайдера

Название	Физика элементарных частиц — Анализ данных Большого адронного коллайдера: открытие бозона Хиггса
Предметная область	Научные исследования: физика
Автор/организация/эл.почта	Майкл Эрнст (Michael Ernst, mernst@bnl.gov) из Брукхейвенской национальной лаборатории (BNL) и Лотар Бауэрдик (Lothar Bauerdick, bauerdick@fnal.gov) из Национальной ускорительной лаборатории именио Ферми, на основе первоначальной версии, написанной Джеффри Фоксом (Geoffrey Fox, gcf@indiana.edu) из Университета Индианы и Эли Дартом (Eli Dart, eddart@lbl.gov) из Национальной лаборатории им. Лоуренса в Беркли, США (LBNL)
Акторы/заинтересованные лица, их роли и ответственность	Физики (проектирование и выявление потребностей в экспериментах, анализ данных). Персонал систем (проектирование, создание и поддержка распределенных вычислительных грид-сетей). Специалисты в области физики ускорителей (проектирование, создание и эксплуатация ускорителя). Правительство (финансирование на основе долгосрочной важности открытий в данной области)

Цели	Понимание свойств элементарных частиц	
Описание варианта использования	Детекторы Большого адронного коллайдера в ЦЕРН и моделирование по методу Монте-Карло «выдают» события, отражающие взаимодействие частиц с приборами. Обработанная информация описывает физические свойства событий, и на ее основе создаются списки частиц с указанием их типа и импульса. Эти события анализируются с целью обнаружения новых явлений — как новых частиц (например, бозона Хиггса), так и сбора доказательств того, что предполагаемые частицы (предсказываемые, например, теорией суперсимметрии) не были обнаружены	
Текущие решения	**Вычислительная система**	«Глобальная грид-инфраструктура Большого адронного коллайдера» (WLCG) и, в США, «Грид открытой науки» (Open Science Grid) объединяют во всемирном масштабе предоставляющие вычислительные ресурсы и ресурсы хранения компьютерные центры в единую инфраструктуру, доступную для всех физиков, работающих с данными Большого адронного коллайдера. 350 тысяч ядер, работающих почти непрерывно, организованы в три уровня (сам ЦЕРН, континенты/страны и университеты). Используются распределенные компьютерные вычисления высокой пропускной способности (Distributed High Throughput Computing, DHTC). Объемы хранения данных — 200 петабайт; в день выполняется более двух миллионов заданий
	Хранилище данных	Эксперимент ATLAS: - Tier1-хранение на лентах в Брукхейвенской национальной лаборатории (BNL) — 10 петабайт данных проекта ATLAS на лентах под управлением высокопроизводительной системы хранения (High Performance Storage System, HPSS). С учетом данных эксперимента по ядерной физике на установке BNL «Релятивистский коллайдер тяжелых ионов» (Relativistic Heavy Ion Collider, RHIC), которые хранятся в том же вычислительном центре, общий объем данных составляет 35 петабайт; - Tier1-хранение на дисках в Брукхейвенской национальной лаборатории: 11 петабайт; система dCache используется для виртуализации набора из ~60 разнородных серверов хранения с дисковыми системами хранения высокой плотности; - Tier2-центры в США: объем дисковой кэш-памяти 16 петабайт. Эксперимент CMS: - Tier1 — хранение в Национальной ускорительной лаборатории им. Энрико Ферми (Fermilab), США: лента/кэш, 20,4 петабайта; - Tier2 — центры в США, объем дисковой кэш-памяти 7 петабайт; - Tier3 — центры в США, объем дисковой кэш - памяти 1,04 петабайта.

Текущие решения	**Сеть связи**	Поскольку участие в экспериментах глобальное (в эксперименте CMS насчитывается 3600 участников из 183 учреждений 38 стран), то данные на всех уровнях передаются и являются доступными на всех континентах. По научным сетям идет масштабная автоматизированная передача данных по всему миру. Комбинация сетевых инфраструктур LHCOPN (соединяет ЦЕРН с центрами Tier1) и LHCONE (связь с центрами Tier2) обеспечивает целевое выделение сетевых ресурсов и изоляцию трафика для данных Большого адронного коллайдера. Tier1 — центр хранения данных эксперимента ATLAS в BNL имеет внутренние пути с пропускной способностью 160 гигабит/с (часто полностью загруженные). Внешние соединения на скорости 70 гигабит/с обеспечивается высокоскоростной сетью ESnet (Energy Sciences Network) Министерства энергетики США. Tier1 — центр хранения данных эксперимента CMS в Национальной ускорительной лаборатория имени Ферми (FNAL/Fermilab) располагает внешними подключениями на скорости 90 гигабит/с, обеспечиваемыми сетью ESnet. Стабильный совокупный трафик данных экспериментов на Большом адронном коллайдере по глобальной сети составляет около 25 гигабит/с
	Программное обеспечение	Масштабируемая система управления рабочей нагрузкой/рабочими процессами эксперимента ATLAS PanDA управляет ≈1 миллионом заданий в сутки на производство данных и их анализ пользователями на глобально распределенных вычислительных ресурсах (≈100 сайтов). Новая распределенная система управления данными эксперимента ATLAS Rucio является основным компонентом, ведущим учет и отслеживающим, в настоящее время, ≈130 петабайт данных, распределенных по грид-ресурсам. Эта система также используется для организации перемещения данных между сайтами. Ожидается, что объем данных в ближайшие несколько лет увеличится до масштаба экзабайтов. На основе системы xrootd, для эксперимента ATLAS была разработала система федеративного хранения FAX, которая обеспечивает удаленный доступ к данным.

Текущие решения	**Программное обеспечение**	Аналогичным образом, в эксперименте CMS используются инфраструктура OSG GlideinWMS для управления потоками рабочих процессов производства и анализа данных, система PhEDEx для управления перемещениями данных, и система AAA/ xrootd для обеспечения удаленного доступа к данным. Специфическое для экспериментов физическое программное обеспечение включает пакеты моделирования, средства обработки данных, развитые пакеты статистической обработки и т. д.
Характеристики больших данных	**Источник данных (распределенный/ централизованный)**	Высокоскоростные детекторы производят большие объемы данных: - детектор ATLAS в ЦЕРНе: первоначальная скорость составляла 1 петабайт/с первичных данных, затем была снижена до 300 мегабайт/с благодаря использованию многоступенчатого триггера. - детектор CMS в ЦЕРНе: аналогично. Данные распространяются глобально Tier1 — центрам, которые выступают в роли источников данных для центров анализа уровней Tier2 и Tier3
	Объем (количество)	15 петабайт в год данных от детекторов и результатов анализа
	Скорость обработки (например, в реальном времени)	В режиме реального времени. Иногда случаются длительные остановки Большого адронного коллайдера (для улучшения ускорителя и детекторов), когда поступают только данные моделирования по методу Монте-Карло. Помимо использования программно- и динамически реплицируемых наборов данных, все чаще при проведении анализа используется удаленный ввод-вывод в режиме реального времени (с использованием XrootD), который требует надежных высокопроизводительных сетевых средств для снижения накладных расходов на копирование файлов и на использование системы хранения
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Множество типов событий, в которых участвуют от двух до нескольких сот конечных элементарных частиц, но все данные — это сведения о частицах после первоначального анализа. События сгруппированы в наборы данных; реальные данные детектора сегментируются на ≈20 наборов данных (с частичным перекрытием) на основе особенностей событий, определенных с помощью работающей в реальном времени триггер-системы; в то время, как различные наборы данных моделирования характеризуются моделируемым физическим процессом

Характеристики больших данных	**Вариативность (темпы изменения)**	Данные накапливаются и не меняют свой характер. В зависимости физической интуиции может меняться предмет Вашего поиска. По мере того как растет понимание функционирования детекторов, выполняются масштабные задачи по повторной обработке данных
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Потеря небольшого количества данных не влечет за собой особых неприятностей, поскольку ошибки пропорциональны обратной величине квадратного корня от числа зафиксированных событий, однако такую потерю данных следует тщательно учитывать. Важно, чтобы ускоритель и экспериментальная установка работали хорошо и понятным образом, в противном случае данные будут слишком «грязными»/«неисправимыми»
	Визуализация	Умеренное использование визуализации, помимо гистограмм и подгонок модели. Имеется отличная визуализация отдельных событий, но для обнаружения явлений требуется много событий, поэтому такой тип визуализации имеет второстепенное значение
	Качество данных (синтаксис)	Огромные усилия прилагаются для того, чтобы сделать поведение определенных сложных экспериментальных установок вполне понятными (правильные калибровки) и надлежащим образом корректировать систематические ошибки в данных. Часто требует повторного анализа данных
	Типы данных	Необработанные первичные экспериментальные данные в различном двоичном представлении с концептуальным синтаксисом «имя: значение», где «имя» может изменяться в диапазоне от «данных сенсора фотокамеры» до «импульса частицы». Данные выделенных из первичных данных физических сигналов (reconstructed data) обрабатываются для получения оптимизированных для анализа представлений в «плотных» форматах
	Аналитика данных	Первоначальный анализ включает обработку специфических экспериментальных данных каждого эксперимента (ALICE, ATLAS, CMS, LHCb), в результате которой выдается сводная информация. На втором этапе анализа проводится «предварительная разведка» (гистограммы, диаграммы рассеяния) с подбором моделей. Существенные объемы моделирования по методу Монте-Карло необходимы для оценки качества анализа.

Наука о больших данных (сбор, курирование, анализ, операции)	**Аналитика данных**	Большая часть (≈60 %) ресурсов центральных процессоров, доступных для совместной работы в рамках проекта ATLAS на уровнях Tier1 и Tier2, используется для моделирования событий. Требования ATLAS к моделированию полностью определяются физическим сообществом с точки зрения потребностей анализа и соответствующих целей в области физики. В настоящее время в рамках физического анализа рассматриваются реальные данные примерно 2 млрд событий, собранные в 2011 г., и данные 3 млрд событий, собранные в 2012 г. (это составляет ≈5 петабайт экспериментальных данных). В рамках ATLAS также было произведено примерно 3,5 млрд смоделированных событий в 2011 г., и 2,5 млрд таких событий в 2012 г. (это составляет ≈ 6 петабайт данных моделирования). Учитывая требования к ресурсам для проведения полного моделирования события с использованием пакета Geant4, ATLAS в настоящее время может производить около 4 млн событий в день в случае использования всех вычислительных мощностей, доступных эксперименту для этой цели в мировом масштабе. Из-за высокой стоимости времени центрального процессора результаты полного моделирования с использованием Geant4 («хитов») хранятся на одной курируемой ленте, среди Tier1-лент, для повторного использования в нескольких повторных пусках моделирования по методу Монте-Карло. «Хиты» от более быстрых программ моделирования будут лишь временно храниться в наборе данных второго периода работы коллайдера
Иные проблемы больших данных	Преобразование научных результатов в новые знания, технические решения, политики и политические решения является основой той научной миссии, которую выполняет как физика высоких энергий в целом, так и, в частности, анализ данных Большого адронного коллайдера. Однако если достижения в области экспериментальных и вычислительных технологий привели к экспоненциальному росту объемов, скорости производства и разнообразия доступных для научных исследований данных, то достижения в технологиях, позволяющих преобразовать эти данные в полезные знания, далеко не соответствуют потребностям сообщества специалистов в области физики высоких энергий обеспечить своевременные и дающие немедленную отдачу результаты. Ускорение процесса отыскания научных знаний абсолютно необходимо, если ученые Министерства энергетики США собираются и впредь вносить большой вклад в развитие физики высоких энергий.	

Иные проблемы больших данных	Ныне существующий всемирный механизм анализа, обслуживающий несколько тысяч ученых, должен быть соразмерно расширен в плане «умности» своих алгоритмов, автоматизации процессов и сферы охвата вычислений, с тем, чтобы сделать возможным научное осмысление детальной природы бозона Хиггса. Так, например, результаты приблизительно сорока различных методов анализа (многие из которых применяют методы машинного обучения), используемых для изучения подробных характеристик бозона Хиггса, должны быть скомбинированы математически строгим образом, чтобы получить согласованный результат, который можно было бы опубликовать. **Специфические проблемы** Объединенный (федеративный) семантический поиск: интерфейсы, протоколы и среды, поддерживающие доступ, использование и интероперабельность между объединяемыми наборами ресурсов, — управляемые на стратегическом и оперативном уровне с использованием сочетания различных политик и мер и инструментов контроля и управления, взаимодействующих с потоковыми и «стационарными» источниками данных. К числу таких мер относятся: - модели, алгоритмы, библиотеки и эталонные реализации распределенной неиерархической службы поиска и выявления; - семантика, методы, интерфейсы для управления жизненным циклом (подписка, захват, происхождение, оценка, проверка, отклонение) неоднородного набора распределенных инструментов, сервисов и ресурсов; - глобальная среда, устойчивая к отказам и сбоям; и - гибкие высокопроизводительные хранилища данных (выходящие за рамки управляемых на основе схемы данных), которые масштабируются и являются дружественными к интерактивной аналитике. **Описание и понимание ресурсов** Распределенные методы и реализации, дающие возможность ресурсам (людям, программному обеспечению, вычислениям, включая данные) публиковать различные состояния и функции для использования разнообразными клиентами. Механизмы для обработки произвольных типов объектов в рамках общей единообразной концепции (включая сложные типы, такие как неоднородные данные, неполная и изменяющаяся информация), — а также быстро меняющаяся доступность вычислительных ресурсов, хранилищ и других ресурсов. Потоковая передача абстрактных данных и перемещение данных на основе файлов по глобальной/локальной сетям и на архитектурах экзабайтного масштаба, поддерживая тем самым возможность совместного принятия в реальном времени решений, касающихся научных процессов
Проблемы пользовательского интерфейса и мобильного доступа	Возможность гибко использовать любые соответствующие доступные ресурсы и обеспечить динамическую доступность всех необходимых данных на этих ресурсах имеет основополагающее значение для будущих открытий в области физики высоких энергий. В данном контексте понятие «ресурс» имеет широкое значение и включает в себя данные и людей, а также вычислительные и некомпьютерные объекты: таким образом, оно охватывает данные любого рода — необработанные данные, информация, знания и т. д.; и ресурсы любого типа — люди, компьютеры, системы хранения, научные инструменты, программное обеспечение, ресурсы, сервисы и т. д. Чтобы эффективно использовать такие ресурсы, необходимо эффективным, безопасным и надежным образом предоставить широкий спектр мер управления, охватывающих, например, сбор, выявление, распределение, перемещение, доступ, использование, выпуск и переназначение. Эти меры управления должны охватывать и контролировать большие ансамбли данных и других ресурсов, которые постоянно меняются и эволюционируют, и часто во многих своих аспектах будут недетерминированными и нечеткими.

Проблемы пользовательского интерфейса и мобильного доступа	**Специфические проблемы** Глобально оптимизированное динамическое распределение ресурсов. Здесь необходимо учитывать отсутствие строгой согласованности знаний в масштабах всей системы. **Минимизация времени доставки данных и услуг** Это нужно не только для сокращения времени на доставку данных или услуг, но и для поддержки возможности прогнозирования, с тем, чтобы занимающиеся анализом данных физики могли адекватно учитывать погрешности в процессах принятии решений в режиме реального времени
Технические проблемы обеспечения безопасности и защиты персональных данных	Хотя сами по себе данные физики высоких энергий не являются проприетарными, внесение в них неумышленных изменений и/или связанная с проблемами кибербезопасности компрометация сервисов центра обработки потенциально могут быть весьма разрушительными для процесса анализа. Помимо необходимости наличия средств удостоверения личности и соответствующих виртуальных систем управления идентификационными данными в организациях для управления правами доступа к конкретным наборам ресурсов, — немало внимания необходимо уделить разработке и эксплуатации многих программных компонент, которые нужны сообществу для проведения вычислений в этой чрезвычайно распределенной среде. Основная часть разработки программного обеспечения и систем для анализа данных Большого андронного коллайдера выполняется внутри сообщества специалистов в области физики высоких энергий или посредством адаптации компонент программного обеспечения, разработанного другими сторонами, что предполагает многочисленные допущения и проектные решения начиная с ранних этапов проектирования и далее на протяжении всего жизненного цикла программного обеспечения. Программные системы основаны на ряде предположений относительно своей среды применения — как они развертываются, конфигурируются, кто их эксплуатирует, в какой сети они находятся, являются ли их входные или выходные данные чувствительными и конфиденциальными, могут ли они доверять своим входным данным, обеспечивают ли защиту неприкосновенности частной жизни и т. д. В случае, когда несколько программных компонентов взаимосвязаны друг с другом, как, например, в стеках «глубинного» программного обеспечения, используемых в компьютерных вычислениях высокой пропускной способности, — без четкого понимания их предположений о безопасности, общая безопасность получаемой системы становится непонятной. Возможным способом решения этой проблемы является создание доверенной среды (trust framework) для компьютерных вычислений высокой пропускной способности. Такая доверенная среда, посредством описания того, что программное обеспечение, системы и организации предоставляют и чего ожидают от своей среды в отношении обеспечения исполнения политик, безопасности и неприкосновенности частной жизни, — позволяет проанализировать систему на наличие пробелов в отношения доверии, робастности и отказоустойчивости
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Крупномасштабный пример анализа на основе событий, с необходимой базовой статистикой. Этот пример также подчеркивает важность виртуальных организаций с точки зрения глобального сотрудничества. Эксперименты на Большом адронном коллайдере являются пионерами в области распределенной инфраструктуры больших данных. Ряд аспектов потока рабочего процессов этих экспериментов высвечивают проблемы, которые другие дисциплины тоже нужно будет решить. В числе этих проблем автоматизацию распределения данных, высокопроизводительная передача данных и крупномасштабные вычисления с большой пропускной способностью
Дополнительная информация (гиперссылки)	Geoffrey Fox, Tony Hey, Anne Trefethen "Where does all the data come from?", 2011, http://cgl.soic.indiana.edu/publications/Where%20does%20all%20the%20data%20come%20from%20v7.pdf William E. Johnston, Eli Dart, Michael Ernst, Brian Tierney “Enabling high throughput in widely distributed data management and analysis systems: Lessons from the LHC”, TNC2013 Conference, 2013, https://www.es.net/assets/pubs_presos/High-throughput-lessons-from-the-LHC-experience.Johnston.TNC2013.pdf

Примечание —

Стадии варианта использования	Источники данных	Использование данных	Трансформации (аналитика данных)	Инфраструктура	Безопасность и приватность
Физика элементарных частиц: Анализ данных Большого адронного коллайдера (LHC), открытие бозона Хиггса (Научные исследования: физика)					
Регистрация первичных данных	Ускоритель LHC, ЦЕРН	Эти данные размещаются в ЦЕРН и затем распространяются по всему миру для следующего этапа обработки	LHC детектирует 109 столкновений в секунду; аппаратно-программный триггер отбирает «интересные события». Другие утилиты распространяют данные по всему миру по скоростным линиям связи	Ускоритель и сложный процесс отбора данных, использующий ≈7000 ядер в ЦЕРН для регистрации ≈100–500 событий в секунду (≈1 мегабайт каждое)	Нет
Обработка первичных данных в информацию	Файлы с первичными данными на диске	Итеративная калибровка и проверка анализа, включающая, например, «эвристические» алгоритмы поиска траекторий. Производит «большие» полные файлы с физическими параметрами и урезанные файлы данных по объекту анализа (AOD), размер которых составляет ≈10 % от исходного	Программа полного анализа, формирующая всестороннее понимание сложного экспериментального детектора. Также программы моделирования методом Монте-Карло для получения смоделированных данных, используемых для оценки эффективности экспериментального детектирования	≈300 тысяч ядер, организованных в 3 уровня: Tier0: ЦЕРН; Tier1: «Основные страны»; Tier2: Университеты и лаборатории. Для сведения: обработка требовательна в плане вычислений и объемов данных	Нет
Физический анализ информации. Извлечение знаний/явлений	Дисковые файлы с информацией, включая данные ускорителя и моделирования методом Монте-Карло	Используются простые статистические методы (такие как гистограммы), методы многомерного анализа и другие методы анализа данных	Определение на основе первичных данных физических параметров и обработка данных с использованием развитых физических алгоритмов для	В то время, как большая часть обработки данных выполняется на ресурсах уровня Tier1 и Tier2, заключительная стадия анализа обычно выполняется пользователями	Физические открытия и результаты являются конфиденциальными до тех пор, пока они не будут подтверждены группой и

Стадии варианта использования	Источники данных	Использование данных	Трансформации (аналитика данных)	Инфраструктура	Безопасность и приватность
	Учитывает знания многих физиков (публикации) при выборе в ходе анализа	и подбора моделей, для обнаружения новых эффектов (частиц) и установления ограничений на еще не наблюдавшиеся эффекты	определения свойств событий, проверки гипотез в отношении элементарных частиц и т. д. Классической программой для интерактивного анализа отобранных наборов обработанных данных является Root (ЦЕРН). Программа считывает файлы ряда событий (AOD, NTUP) из выбранных наборов данных и использует созданный физиком C ++ — код для вычисления новых параметров, таких, как предполагаемая масса нестабильной (новой) частицы	на локальных объектах уровня Tier3. Масштаб вычислительных ресурсов на узлах уровня Tier3 варьируется от рабочих станций до небольших кластеров. Наиболее распространенным программным стеком, применяемым для анализа компактных форматов данных, генерируемых на распределенных вычислительных ресурсах, является ROOT. Передача данных выполняется с использованием инструментов распределенного управления данными экспериментов ATLAS и CMS, которые в основном полагаются на промежуточное программное обеспечение gridFTP. Прямой доступ к данным на основе XROOTD также приобретает большое значение там, где доступна высокая пропускная способность сети	представлены на встрече/в журнале. Данные сохраняются, поэтому результаты воспроизводимы

А.7.5 Вариант использования № 40: Эксперимент Belle II

Название	Эксперимент Belle II
Предметная область	Научные исследования: физика высоких энергий
Автор/организация/эл.почта	Дэвид Аснер (David Asner, david.asner@pnnl.gov) и Малакай Шрам (Malachi Schram, malachi.schram@pnnl.gov), Тихоокеанская северо-западная национальная лаборатория, США (PNNL)
Акторы/заинтересованные лица, их роли и ответственность	Дэвид Аснер (David Asner) — научный руководитель американского проекта Belle II. Малакай Шрам (Malachi Schram) — координатор сети и передачи данных в проекте Belle II, а также руководитель вычислительного центра Belle II в Тихоокеанской северо-западной национальной лаборатории, США (PNNL)
Цели	Выполнять точные измерения для поиска новых явлений, выходящих за рамки стандартной модели физики элементарных частиц

Описание варианта использования	Изучение многочисленных мод распада в мезонном резонансе Y(4S) с целью обнаружения новых явлений, выходящих за рамки стандартной модели физики элементарных частиц	
Текущие решения	**Вычислительная система**	Распределенные (грид-вычисления на базе инфраструктуры DIRAC (Distributed Infrastructure with Remote Agent Control))
	Хранилище данных	Распределенное (различные технологии)
	Сеть связи	Непрерывная передача первичных данных со скоростью ~20 гигабит/с между Японией и США при проектной яркости ускорителя. Дополнительные скорости передачи в настоящее время изучаются
	Программное обеспечение	«Грид Открытой науки» (Open Science Grid), Geant4, DIRAC, FTS, инфраструктура Belle II
Характеристики больших данных	**Источник данных (распределенный/централизованный)**	Распределенные центры обработки данных. Центры хранения первичных данных находятся в Японии («Организация по изучению высокоэнергетических ускорителей» KEK) и США (PNNL)
	Объем (количество)	Объем интегрированных первичных данных составит около 120 петабайт, физических данных — около 15 петабайт, данных моделирования по методу Монте-Карло — около 100 петабайт
	Скорость обработки (например, в реальном времени)	Данные будут перекалибровываться и анализироваться постепенно. Скорость передачи данных будет увеличиваться в зависимости от яркости ускорителя
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Данные будут перекалибровываться и распределяться постепенно
	Вариативность (темпы изменения)	Количество столкновений будет постепенно увеличиваться до тех пор, пока не будет достигнута расчетная яркость (3000 B-B пар в секунду). Ожидаемый объем данных о каждом событии ~300 килобайт
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Валидация будет выполняться с использованием известных эталонных физических процессов
	Визуализация	Нет
	Качество данных (синтаксис)	Выходные данные будут перекалибровываться и проверяться постепенно

Наука о больших данных (сбор, курирование, анализ, операции)	**Типы данных**	Вывод на основе кортежа
	Аналитика данных	Кластеризация и классификация данных является неотъемлемой частью вычислительной модели. Отдельные ученые определяют, как проводится анализ на уровне событий
Иные проблемы больших данных	Перемещение и учет данных (метаданные на уровне файлов и событий)	
Проблемы пользовательского интерфейса и мобильного доступа	Сетевая инфраструктура, необходимая для непрерывной передачи данных между Японией (KEK) и США (PNNL)	
Технические проблемы обеспечения безопасности и защиты персональных данных	Никаких особых проблем нет. Доступ к данным осуществляется с использованием аутентификации в грид-системе	
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)		
Дополнительная информация (гиперссылки)	Сайт проекта Belle II, https://www.belle2.org/	

А.8 Науки о Земле, экологические науки и полярные исследования

А.8.1 Вариант использования № 41: Радарная система некогерентного рассеяния EISCAT-3D

Название	Радарная система некогерентного рассеяния EISCAT-3D
Предметная область	Экологические науки
Автор/организация/эл.почта	Инь Чен (Yin Chen) / Кардиффский университет (Cardiff University), Великобритания / chenY58@cardiff.ac.uk Ингемар Хегстрем (Ingemar Häggström, Ingemar.Haggstrom@eiscat.se), Ингрид Манн (Ingrid Mann, Ingrid.mann@eiscat.se), Крейг Хайнсельман (Craig Heinselman, Craig.Heinselman@eiscat.se) / Европейская научная ассоциация но некогерентному рассеянию радиоволн EISCAT
Акторы/заинтересованные лица, их роли и ответственность	Научная ассоциация EISCAT является международной научно-исследовательской организацией, эксплуатирующей радиолокационные системы некогерентного рассеяния в Северной Европе. Она финансируется и управляется научно-исследовательскими советами Норвегии, Швеции, Финляндии, Японии, Китая и Великобритании (коллективно именуемыми «партнеры EISCAT»). Помимо радаров некогерентного рассеяния, EISCAT также эксплуатирует ионосферный нагревной стенд, а также два динамических цифровых ионозонда Dynasonde
Цели	Европейская научная ассоциация но некогерентному рассеянию радиоволн EISCAT (European Incoherent Scatter Scientific Association) была создана для проведения исследований нижней, средней и верхней атмосферы и ионосферы с использованием радарных систем некогерентного рассеяния. Эти установки являются наиболее мощными наземными инструментами, используемыми в такого рода исследованиях. EISCAT также использует радар некогерентного рассеяния для изучения нестабильностей в ионосфере; для исследования структуры и динамики средней атмосферы; и в качестве измерительно-диагностического инструмента в экспериментах по модификации ионосферы с использованием нагревного стенда EISCAT/Heating.

Описание варианта использования	Конструкция радарной системы некогерентного рассеяния следующего поколения EISCAT-3D открывает перед физиками возможности для проведения исследований во многих новых областях. С другой стороны, возникают значительные проблемы, связанные с обработкой больших объемов экспериментальных данных, которые будут производиться массово и с высокими темпами. Данная проблема, о которой обычно говорят как о проблеме «больших данных», требует решений, выходящих за рамки возможностей традиционных технологий баз данных	
Текущие решения	**Вычислительная система**	В электронной инфраструктуре данных эксперимента EISCAT-3D планируется использовать высокопроизводительные компьютеры для обработки данных в основном центре и компьютеры с высокой пропускной способностью в зеркальных центрах обработки данных
	Хранилище данных	32 терабайта
	Сеть связи	Согласно оценкам, скорости передачи данных в локальных сетях на центральном посте составляют от 1 до 10 гигабит/с. Аналогичная скорость требуется при подключении постов по выделенным высокоскоростным сетевым соединениям. Операция скачивания всего массива данных не является критичной ко времени, однако для оперативного управления требуется информация в реальном времени о некоторых заранее определенных событиях, которая будет поступать с постов в центр управления; а также связь в реальном времени центра управления с постами для установления в реальном времени режима работы радара
	Программное обеспечение	Распространенные операционные системы, такие как Windows, Linux, Solaris, HP/UX и FreeBSD Простое одноуровневое хранение файлов с поддержкой необходимых функциональных возможностей, таких как сжатие, страйпинг и журналирование файлов Самостоятельно разработанное программное обеспечение: - инструменты управления и мониторинга, включая конфигурирование системы; быстрый просмотр, отчеты об отказах и т. д.; - утилиты распространения данных; - пользовательское программное обеспечение, например, для циклического буфера, очистки данных, обнаружения и удаления радиочастотных помех, автокорреляции, интеграции данных, анализа данных, выявления событий, поиска и извлечения, производства вторичных полезных продуктов данных, приема / извлечения, построения графиков; - ориентированные на пользователя вычисления; - API-интерфейсы к стандартным программным средам; - цепочки и потоки рабочих процессов обработки данных

Характеристики больших данных	**Источник данных (распределенный/ централизованный)**	Комплекс EISCAT-3D будет состоять из центрального поста с передающими и приемными антенными решетками и четырех постов с приемными антенными решетками на расстоянии около 100 км от центрального поста
	Объем (количество)	Полностью функциональная система из пяти постов будет производить 40 петабайт в год в 2022 г. Ожидается, что комплекс будет эксплуатироваться в течение 30 лет, а результаты обработки данных будут храниться не менее 10 лет
	Скорость обработки (например, в реальном времени)	На каждом из пяти постов: - каждая антенна выполняет 30 миллионов измерений в секунду (120 мегабайт/ с); - каждая группа из 100 антенн формирует поток данных мощностью 2 гигабайта в секунду - эти данные временно хранятся в кольцевом буфере: 160 групп — > 125 терабайт в час
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Измерения: различные версии, форматы, реплики, внешние источники ... Системная информация: конфигурация, мониторинг, журналы аудита. Пользовательские метаданные/данные: эксперименты, анализ, коллективное использование
	Вариативность (темпы изменения)	Во времени: мгновенно или несколько миллисекунд. Вдоль радиолокационных лучей — 100 наносекунд
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Круглосуточно эксплуатируемый в режиме 24/7 комплекс EISCAT-3D предъявляет очень высокие требования к надежности и отказоустойчивости. Обеспечение надежности данных и стабильной производительности играют важнейшую роль для систем кольцевого буфера и архивного хранения. Эти системы должны обеспечить соответствие требованиям к минимальной скорости приема данных в любое время, иначе научные данные будут потеряны. Аналогичным образом, эти системы должны гарантировать неизменность хранимых данных и отсутствие в них искажений. Последнее требование особенно важно для постоянного архива, в котором научные данные, скорее всего, будут доступны для исследователей, и где их труднее всего проверить; искажение данных в архиве с большой вероятностью окажется неисправимым и способно негативно повлиять на научную литературу

Наука о больших данных (сбор, курирование, анализ, операции)	**Визуализация**	Визуализация анализируемых данных в режиме реального времени, например, в виде обновляемой диаграммы, на которой показаны концентрация электронов, показания температуры и скорость ионов на основе данным для каждого луча. Не в режиме реального времени (после эксперимента) визуализируются представляющие интерес физические параметры, например: - стандартные графики, используемые в экспериментах EISCAT; - отображение данных нескольких лучей в виде трехмерного «блока» для демонстрации пространственных изменений (в выбранных пользователем разрезах); - использование анимации для отображения изменений во времени; - поддержка визуализации 5-мерных (и более) данных, например, с использования метода «разрезания и складывания стопкой» (cut up and stack) для понижения размерности, когда одна или несколько независимых координат изменяются дискретно; или метод объемного рендеринга для отображения двумерной проекции трехмерного дискретного набора данных. Интерактивная визуализация. Дает пользователям возможность объединять информацию о нескольких спектральных особенностях (используя, в том числе, цветовое кодирование). Предоставляет пользователям возможность в реальном времени связывать или подключать специализированные функции визуализации данных, и, что более важно, функции для сигнализации о специфических условиях наблюдения
	Качество данных (синтаксис)	Будет предоставлено программное обеспечение для мониторинга, которое позволит оператору «видеть» поступающие данные через систему визуализации в режиме реального времени и соответствующим образом реагировать на интересные с научной точки зрения события. Будет разработано управляющее программное обеспечение для временной интеграции сигналов, уменьшения дисперсии шума и увеличения общей пропускной способности системы по передаче данных в архив данных
	Типы данных	HDF-5
	Аналитика данных	Распознавание образов, требовательные процедуры корреляции, извлечение высокоуровневых параметров

Иные проблемы больших данных	Высокая пропускная способность преобразования данных в данные более высокого уровня Для извлечения существенных знаний из малополезных в их исходном виде данных (low value — density data) необходимы новые методы глубокого, сложного анализа, например, с использованием машинного обучения, статистического моделирования, алгоритмов поиска на графе и т. д., выходящих за рамки традиционных подходов к физике космоса
Проблемы пользовательского интерфейса и мобильного доступа	Использование на мобильных платформах маловероятно
Технические проблемы обеспечения безопасности и защиты персональных данных	Доступ к данным нижнего уровня ограничен в странах-партнерах на один год. Все данные раскрываются через три года
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Электронная инфраструктура данных проекта EISCAT- 3D имеет сходные архитектурные характеристики с другими радарами разведки и наблюдения, и со многими существующими системами, производящими и анализирующими большие данные, такими, как радиоинтерферометр LOFAR голландского института ASTRON, «Большой адронный коллайдер» ЦЕРН и международный проект радиоинтерферометра SKA (Square Kilometer Array)
Дополнительная информация (гиперссылки)	Веб-сайт проекта EISCAT-3D, https://eiscat.se/eiscat3d/

А.8.2 Вариант использования № 42: Совместная деятельность европейских сетевых инфраструктур в области экологических исследований (ENVRI)

Название	Совместная деятельность европейских сетевых инфраструктур в области экологических исследований (ENVRI)
Предметная область	Экологические науки
Автор/организация/эл.почта	Инь Чен (Yin Chen) / Кардиффский университет (Cardiff University), Великобритания / chenY58@cardiff.ac.uk
Акторы/заинтересованные лица, их роли и ответственность	ENVRI — это проект сотрудничества, выполняемый в рамках экологического кластера «Европейского стратегического форума по исследовательским инфраструктурам» (European Strategy Forum on Research Infrastructures, ESFRI). В число участвующих в проекте ENVRI инфраструктур экологических исследований ESFRI входят: - «Интегрированная система наблюдения за выбросами углерода» ICOS (Integrated Carbon Observation System) — европейская распределенная инфраструктура, предназначенная для мониторинга парниковых газов через ее атмосферные, экосистемные и океанские сети наблюдений; - EURO-Argo — европейский вклад в международную систему наблюдений за океаном Argo; - EISCAT-3D (описан в отдельном варианте применения № 41) — европейская исследовательская радарная система некогерентного рассеяния нового поколения для исследований верхней атмосферы; - LifeWatch (описан в отдельном варианте применения № 25) — европейская электронная инфраструктура для исследований в области экологии и биологического разнообразия; - «Европейская исследовательская инфраструктура для слежения за [геологическими] плитами» (EPOS) — это европейская инфраструктура для исследования землетрясений, вулканов, динамики поверхности и тектоники; - «Европейская междисциплинарная обсерватория исследования морского дна и слоев воды» (EMSO) — европейская сеть наблюдательных станций морского дна, предназначенная для мониторинга в долгосрочном масштабе времени экологических процессов, связанных с экосистемами, изменением климата и геологическими опасностями.

Акторы/заинтересованные лица, их роли и ответственность	Проект ENVRI также поддерживает тесные контакты с другими, непосредственно не участвующими в деятельности форума ESFRI, инфраструктурами экологических исследований, приглашая их представителей на совместные обсуждения. Это проекты: - «Использование самолетов в глобальной системе наблюдений» (IAGOS) организует сеть самолетов для глобального наблюдения за атмосферой; - «Интегрированная система наблюдений за Арктикой на Шпицбергене» (SIOS) создает систему наблюдений на Шпицбергене и вокруг него, которая объединяет исследования геофизических, химических и биологических процессов, проводимые на всех платформах исследований и мониторинга. ИТ-сообщество проекта ENVRI разрабатывает общую политику и технические решения для научно-исследовательских инфраструктур, привлекая к этой работе ряд организаций-партнеров, включая Кардиффский Университет (Cardiff University, Великобритания), Институт информационных наук и технологий им. Алессандро Фаедо (Istituto di scienza e tecnologie dell'informazione «Alessandro Faedo», ISTI) итальянского Национального совета по научным исследованиям (Consiglio Nazionale delle Ricerche, CNR), Национальный центр научных исследований Франции (Centre National de la Recherche Scientifique, CNRS), финскую ИТ-компанию CSC- координатор европейского проекта EUDAT, агентство по защите окружающей среды Австрии (Umweltbundesamt), федерацию европейских грид-инфраструктур (European Grid Infrastructure, EGI), Европейский институт космических исследований (European Space Research Institute, ESRIN) Европейского космического агентства (European Space Agency, ESA), Амстердамский и Эдинбургский университеты
Цели	Проект ENVRI объединяет усилия шести инфраструктур «Европейского стратегического форума по исследовательским инфраструктурам» ESFRI (ICOS, EURO-Argo, EISCAT-3D, LifeWatch, EPOS и EMSO) по разработке общих сервисов данных и программного обеспечения. Результаты проекта ускорят создание этих инфраструктур и улучшат интероперабельность между ними. Основной целью ENVRI является согласование эталонной модели для целей совместной деятельности. Эталонная модель ENVRI RM служит общей онтологической структурой и стандартом для описания и характеризации вычислительной инфраструктуры и инфраструктуры хранения, с целью обеспечения бесперебойной интероперабельности между неоднородными ресурсами различных инфраструктур. Модель ENVRI RM также служит общим языком общения в сообществе, обеспечивая единую концепцию, на основе которой можно классифицировать и сравнивать компоненты инфраструктуры. Модель ENVRI RM также используется для выявления типовых решений общих проблем. Все это позволяет обеспечить повторное использование, совместное использование ресурсов и обмен опытом, а также избежать дублирования усилий
Описание варианта использования	Проект ENVRI реализует гармонизированные решения и разрабатывает руководства и рекомендации в отношении общих потребностей экологических проектов ESFRI, уделяя особое внимание таким вопросам, как архитектура, схемы метаданных, поиск данных в разбросанных хранилищах, визуализация и курирование данных. Это откроет новые возможности для пользователей сотрудничающих инфраструктур экологических исследований и обеспечит участникам междисциплинарных исследований возможность получать, изучать и сопоставлять данные из нескольких областей знаний в интересах исследований системного уровня. В проекте ENVRI изучается репрезентативная выборка научно-исследовательских инфраструктур для экологических исследований, выдавая на выходе прогноз общеевропейских требований, которые эти инфраструктуры предъявляют. В частности, выявляются общие для них требования. На основании данных анализа и с использованием международного стандарта ИСО «Открытая распределенная обработка», разработана эталонная модель ENVRI Reference Model (https://confluence.egi.eu/display/EC/Download+of+ENVRI+Reference+Model).

Описание варианта использования	По сути, эта модель выполняет роль универсальной эталонной концепции при обсуждении многих общих технических проблем, стоящих перед всеми инфраструктурами ESFRI для экологических исследований. Сопоставляя эталонные компоненты модели и фактические компоненты инфраструктур (или их предлагаемые проекты) в том виде, в котором они существуют в настоящее время, можно выявить различные пробелы и области перекрытия	
Текущие решения	**Вычислительная система**	
	Хранилище данных	Файловые системы и реляционные базы данных
	Сеть связи	
	Программное обеспечение	Собственное
Характеристики больших данных	**Источник данных (распределенный/ централизованный)**	Большинство исследовательских инфраструктур ENVRI представляют собой распределенные, рассчитанные на длительную перспективу, дистанционно управляемые сети наблюдений, ориентированные на понимание процессов, тенденций, порогов, взаимодействий и обратных связей, а также на повышение предсказательной способности в интересах решения будущих экологических проблем. Они простираются от арктических районов до самых южных европейских областей и от Атлантики на западе до Черного моря на востоке. Более конкретно: - **EMSO**, сеть стационарных наблюдательных станций для изучения морского дна и слоев воды, географически распределена по ключевым участкам европейских вод, и в настоящее время состоит из тринадцати станций. - Европейская инфраструктура для исследования землетрясений, вулканов, динамики поверхности и тектоники **EPOS** ставит своей целью интеграцию существующих европейских центров геологических исследований в единую междисциплинарную исследовательскую сеть, а также повышение доступности и удобства использования междисциплинарных данных из сетей сейсмического и геодезического мониторинга, наблюдений за вулканической активностью, лабораторных экспериментов и компьютерного моделирования, повышая в мировом масштабе интероперабельность исследований в области наук о Земле.

Характеристики больших данных	**Источник данных (распределенный/ централизованный)**	- Проект **ICOS** занимается мониторингом парниковых газов через свои сети атмосферных, экосистемных и океанских наблюдений. Сеть ICOS включает в себя более 30 атмосферных и более 30 экосистемных первичных долговременных станций наблюдения, расположенных по всей Европе, и дополнительные вторичные станции. Она также включает три «тематических центра» (Thematic Centres), занимающиеся обработкой данных всех станций каждой сети, и обеспечивающих доступ к этим данным. - **LifeWatch** — это «виртуальная» европейская инфраструктура для исследований в области экосистем и биологического разнообразия, услуги которой предоставляются в основном через Интернет. Ее центры общего пользования (Common Facilities) координируются и управляются на центрально-европейском уровне; а узлами сети LifeWatch являются специализированные центры стран-участниц (региональные партнерские центры) или исследовательских сообществ. - Проект **Eurio-Argo** предоставляет, развертывает и эксплуатирует примерно 800 буев, внося свой вклад в глобальные усилия (3000 буев) и, таким образом, обеспечивает расширенное покрытие европейских региональных морей. - Европейская исследовательская радарная система некогерентного рассеяния нового поколения **EISCAT-3D** проводит непрерывные измерения геопространственной среды и ее взаимосвязи с атмосферой Земли, в точке, расположенной в зоне полярных сияний на южном крае северного полярного вихря, и представляет собой распределенную инфраструктуру
	Объем (количество)	Объемы данных различные, например: - в проекте EMSO, в зависимости от инструментария и конфигурации наблюдательной станции, объем набора данных варьируется от нескольких мегабайт до нескольких гигабайт;

Характеристики больших данных	**Объем (количество)**	- в рамках проекта EPOS сеть EIDA в настоящее время предоставляет доступ к потоку первичных данных, непрерывно поступающему с более чем 1000 станций, регистрирующих в день около 40 гигабайт, т. е. более 15 терабайт в год. EMSC хранит базу данных объемом 1,85 гигабайта с параметрами землетрясений, которая постоянно растет и пополняется уточненной информацией: - событий — 222 705; - мест — 632 327; - магнитуд — 642 555; - в рамках проекта EISCAT-3D, темпы производства первичных данных достигнут 49 петабайт в год в 2023 г.
	Скорость обработки (например, в реальном времени)	Обработка данных в режиме реального времени является распространенным требованием инфраструктур экологических исследований
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Данные очень сложные и неоднородные
	Вариативность (темпы изменения)	Скорость изменений сравнительно низкая
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Нормальная
	Визуализация	В большинстве проектов методы визуализации еще не доведены до уровня полной работоспособности. В проекте **EMSO** визуализация не полностью работоспособна; в настоящее время используются только простые инструменты построения графиков. В проекте **EPOS** методы визуализации еще не определены. В проекте **ICOS** результаты обработки данных «уровня 1.b», такие, как измерения концентрации парниковых газов в почти реальном времени, доступны пользователям через веб-портал «Атмосферного тематического центра» (Atmospheric Thematic Centre, ATC). Интерактивная линейная временная (на базе Google Chart Tools) для временной последовательности с опциональными аннотациями позволяет пользователю прокручивать и менять увеличение временных последовательностей измерений CO_2 или CH_4 на атмосферной станции ICOS. Диаграмма отображается в браузере с использованием Flash.

Наука о больших данных (сбор, курирование, анализ, операции)	**Визуализация**	Также ведущим исследователям станций (principle investigators) доступны для обеспечения мониторинга инструментов некоторые результаты обработки данных «уровня 2». В основном это автоматически создаваемые графики показаний инструментов и сравнительные графики (используются язык R и библиотека построенние графиков Python Matplotlib 2D), которые ежедневно помещаются на веб-сервер ICOS. Результаты обработки данных «уровня 3», такие как данные о потоках парниковых газов с географической привязкой, произведенные на основе наблюдений ICOS, способствуют росту научного влияния ICOS. Для этого ICOS поддерживает сообщество пользователей. Ожидается, что портал Carbon станет платформой, которая будет поддерживать визуализацию данных о потоках парниковых газов, включающих данные ICOS. Примером возможных будущих результатов обработки данных ICOS о концентрации парниковых газов «уровня 3» могут, например, служить карты европейских потоков CO_2 или CH_4 с высоким разрешением, полученные европейскими специалистами по моделированию атмосферной инверсии. Визуальные инструменты для сравнения данных будут разработаны порталом Carbon. Приветствуются любые продукты высокого научного качества. Проект **LifeWatch** предоставит общие методы визуализации, такие как нанесение данных о видах на карты. Новые методы позволят визуализировать эффект изменения данных и/или параметров моделей
	Качество данных (синтаксис)	Очень важно
	Типы данных	Измерения (часто сохраненные в файловых форматах). Метаданные. Онтология. Аннотации
	Аналитика данных	Ассимиляция данных. (Статистический) анализ. Интеллектуальный анализ данных. Извлечение данных. Построение научных моделей и моделирование. Управление потоками научных рабочих процессов

Иные проблемы больших данных	Обработка в реальном времени экстремально больших объемов данных. Резервирование данных в реальном времени в зеркальные архивы. Интегрированный доступ к данным и поиск данных. Обработка и анализ данных
Проблемы пользовательского интерфейса и мобильного доступа	Общей является потребность в эффективных и высокопроизводительных мобильных детекторах, и контрольно — измерительных приборах: - в проекте ICOS различные мобильные инструменты используются для сбора данных океанических и атмосферных наблюдений, и данных мониторинга экосистем; - в проекте Euro-Argo используются тысячи подводных роботов для наблюдения за всеми океанами; - в проекте Lifewatch биологи используют мобильные инструменты для наблюдений и измерений
Технические проблемы обеспечения безопасности и защиты персональных данных	Большинство проектов придерживаются политики открытых данных и их коллективного использования. Например: - Видение проекта EMSO — предоставить ученым всего мира доступ к данным наблюдений по модели открытого доступа. - В проекте EPOS данные в сети EIDA и параметры землетрясений, как правило, открыты и могут свободно использоваться. Некоторые ограничения существуют в отношении отдельных сейсмических сетей, и доступ регулируется в зависимости от аутентификации/авторизации на основе электронной почты. - Данные проекта ICOS будут доступны с лицензией на полный и открытый доступ. Каких-либо ограничений на доступ и возможное использование данных не предвидится; ожидается, что данные будет невозможно распространить дальше. Будут приниматься меры по обеспечению признания происхождения данных от ICOS и их прослеживаемости с использованием специальных мер (например, DOI набора данных). Большая часть соответствующих данных и ресурсов создается с использованием государственного финансирования из национальных и международных источников. - Проект LifeWatch следует соответствующим европейским политикам, таким, как: требования Европейского совета по научным исследованиям (European Research Council, ERC); пилотный проект Европейской комиссии по открытому доступу 2008 года. В отношении публикаций, такие инициативы, как инициированные издателями проект «Дриада» (Dryad) и «Инфраструктура открытого доступа к европейским исследованиям» (Open Access Infrastructure for Research in Europe, OpenAIRE). Частный сектор может размещать свои данные в инфраструктуре LifeWatch. Для управления такими коммерческими контрактами будет создана специальная компания. - В проекте EISCAT-3D доступ к данным более низкого уровня ограничивается на один год в странах-партнерах. Все данные раскрываются через три года
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Различные научно-исследовательские инфраструктуры предназначены для разных целей и эволюционируют с течением времени. Проектировщики описывают свои подходы с различных точек зрения, на разных уровнях детализации и с использованием разных типологий. Предоставленная документация часто является неполной и непоследовательной. Необходима единая платформа для интерпретации и обсуждения, которая помогла бы обеспечить единое понимание. В проекте ENVRI мы решили использовать стандартную модель Открытой распределенной обработки (Open Distributed Processing, ODP) для интерпретации проектов и структур исследовательских инфраструктур и для помещения их требования в структуру ODP для дальнейшего анализа и сопоставления

Дополнительная информация (гиперссылки)	Сайт проекта ENVRI: https://envri.eu/ Эталонная модель ENVRI (ENVRI Reference Model), https://confluence.egi.eu/display/EC/Download+of+ENVRI+Reference+Model Инь Чен (Yin Chen) и др. «Анализ общих требований к научно-исследовательским инфраструктурам экологических исследований» (Analysis of Common Requirements for Environmental Science Research Infrastructures), Международный симпозиум по гридам и облакам 2013 (International Symposium on Grids and Clouds, ISGC), 17–22 марта 2013 г., см. https://pos.sissa.it/179/032/pdf Сайт проекта ICOS: http://www.icos-infrastructure.eu/ Сайт проекта Euro-Argo: https://www.euro-argo.eu/ Сайт проекта EISCAT-3D: https://eiscat.se/eiscat3d/ Сайт проекта LifeWatch-ERIC: https://www.lifewatch.eu/web/guest/home Сайт проекта EPOS: https://www.epos-eu.org/ Сайт проекта EMSO-ERIC: http://emso.eu/

А.8.3 Вариант использования № 43: Анализ радиолокационных данных для Центра дистанционного зондирования ледяного покрова CReSIS

Название	Анализ радиолокационных данных для Центра дистанционного зондирования ледяного покрова (CReSIS)	
Предметная область	Научные исследования: исследования полярных регионов и дистанционное зондирование ледяного покрова	
Автор/организация/эл.почта	Джоффри Фокс (Geoffrey Fox), университет штата Индиана (США), gcf@indiana.edu	
Акторы/заинтересованные лица, их роли и ответственность	Исследования, финансируемые Национальным научным фондом (National Science Foundation, NSF) и Национальным управлением по аэронавтике и исследованию космического пространства (NASA), имеют отношение к изменениям климата в краткосрочной и длительной перспективе. Инженеры проектируют новый радар, который будет отправляться в «полевые экспедиции» длительностью 1—2 мес в отдаленные места. Результаты, используются учеными для создания моделей и теорий, учитывающих ледяной покров	
Цели	Определение толщины слоя ледяного покрова и слоев снега, с целью использования этих данных в научном анализе более высокого уровня	
Описание варианта использования	Создание радара; создание беспилотного летательного аппарата (БПЛА) или использование пилотируемого самолета; облеты отдаленных районов (в Арктике, Антарктиде, Гималаях). Проверка на месте правильности настройки эксперимента, и проведение подробного анализа данных в более позднее время. Транспортировка данных по воздуху — доставка жесткого диска, ввиду плохого интернет-соединения. Использование обработки изображений для определения толщины льда/снежного покрова. Использование полученных данных в научных исследованиях процессов таяния ледяных шапок и т. д.	
Текущие решения	**Вычислительная система**	В поле: кластер с низким энергопотреблением из прочных ноутбуков плюс классические серверы с 2—4 процессорами и съемным дисковым массивом емкостью ≈40 терабайт. Автономная обработка: кластер из 2500 ядер
	Хранилище данных	В поле: съемный жесткий диск. Диски в полевых условиях подвергаются неблагоприятным воздействиям, поэтому делаются две копии. Автономная обработка: Lustre или эквивалентная система хранения

Текущие решения	**Сеть связи**	Ужасного качества интернет, связывающий полевые станции с континентальными США
	Программное обеспечение	Обработка радиолокационных сигналов в пакете Matlab. Анализ изображений с использованием Map/Reduce или MPI, плюс C/Java. Пользовательский интерфейс — географическая информационная система
Характеристики больших данных	**Источник данных (распределенный/централизованный)**	Самолеты пролетают над ледяными полями по тщательно спланированным маршрутам. Собранные данные записываются на жесткие диски
	Объем (количество)	≈0,5 петабайт в год необработанных данных
	Скорость обработки (например, в реальном времени)	Все данные собираются в режиме реального времени, однако анализируются постепенно и хранятся базе данных, интерфейс к которой обеспечивает географическая информационная система
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Множество различных, похожих по своей структуре наборов данных, каждый из которых требует индивидуализированной обработки сигналов. Эти данные необходимо использовать с большим количеством других данных исследований полярных регионов
	Вариативность (темпы изменения)	Данные каждой экспедиции накапливаются блоками объемом примерно по 100 терабайт
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Крайне важна для мониторинга полевых данных и корректировки проблем с измерительными инструментами. Это означает, что часть данных должна полностью анализироваться в полевых условиях
	Визуализация	Богатый пользовательский интерфейс для моделирования слоев снежного и ледяного покрова и движения ледников
	Качество данных (синтаксис)	Обеспечение получения от измерительного оборудования качественных данных является основным инженерным вопросом
	Типы данных	Радиолокационные изображения
	Аналитика данных	Сложная обработка сигналов; новые методы обработки изображений с целью выделения слоев (которых могут быть сотни — один слой в год)
Иные проблемы больших данных	Объемы данных увеличиваются. Доставка жестких дисков выглядит неуклюже, но другого очевидного решения нет. Алгоритмы обработки изображений по-прежнему являются очень активной областью исследований	
Проблемы пользовательского интерфейса и мобильного доступа	Интерфейсы для смартфонов существенного значения не имеют, в то время как технологии с низким энергопотреблением имеют крайне важное значение для полевых исследований	
Технические проблемы обеспечения безопасности и защиты персональных данных	Исследования в Гималаях осложняются из-за местных политических проблем, поэтому требуются беспилотные летательные аппараты. Сами данные являются открытыми после первоначального изучения	

Перечислите основные характеристики и связанные варианты использования (в интересах эталонной архитектуры)	Слабосвязанные кластеры для обработки сигналов. Необходима поддержка Matlab
Дополнительная информация (гиперссылки)	Сайт проекта Polar Grid, поддерживаемый Университетом Индианы, http://polargrid.org/about.html Сайт проекта CReSIS, https://cresis.ku.edu/ Видеоролик об исследованиях ледяного покрова на сайте проекта Polar Grid, http://polargrid.org/gallery.html#

Примечание —

Стадии варианта использования	**Источники данных**	**Использование данных**	**Трансформации (аналитика данных)**	**Инфраструктура**	**Безопасность и приватность**
Анализ радиолокационных данных для Центра дистанционного зондирования ледяного покрова CReSIS (Научные исследования: исследования полярных регионов и дистанционное зондирование ледяного покрова)					
Первичные данные: Полевая экспедиция	Первичные данные с радарной системы на самолете/транспортном средстве	Запись данных на жесткие диски для этапа обработки L1B. Проверка данных для контроля состояния инструментов	Надежные утилиты копирования данных. Вариант полного анализа для проверки данных	Прочные ноутбуки и небольшой сервер (~2 ЦП с системой съемных жестких дисков емкостью ~40 Тб)	Нет
Информация: Автономный анализ L1B	Данные с доставленных жестких дисков копируются в файловую систему (Lustre)	Создание в результате обработки данных радиолокационных изображений	Для каждой выборки данных запускается свой экземпляр программа анализа в Matlab, который работает параллельно и независимо от других экземпляров	~2500 ядер, на которых исполняются стандартные инструменты кластера	Нет, за исключением проверки результатов перед раскрытием на вебсайте CReSIS
Информация: L2/L3 Геолокация и выделение слоев	Радарные изображения с этапа обработки L1B	Вклад в науку — база данных с ГИС-интерфейсом	Средства ГИС и инструменты работы с метаданными. Среда, поддерживающая автоматическое и/или ручное выделение слоев	ГИС (географическая информационная система). Кластер для обработки изображений	См. выше
Знание, мудрость, открытия: Наука	ГИС-интерфейс для данных с этапов обработки L2/L3	Полярные научные исследования, объединяющие многие источники данных, например в интересах изучения изменения климата. Данные о дне ледника, используемые при моделировании течения ледника		Исследования в ГИС облачного стиля, поддерживающей доступ к данным. Моделирование — программа решения трехмерных дифференциальных уравнений в частных производных на большом кластере	Зависит от вида научного использования. Обычно результаты раскрываются после завершения исследования

А.8.4 Вариант использования № 44: Обработка данных проекта UAVSAR

Название	Обработка данных, доставка результатов и сервисы данных проекта UAVSAR	
Предметная область	Научные исследования: науки о Земле	
Автор/организация/эл.почта	Андреа Доннеллан (Andrea Donnellan, andrea.donnellan@jpl.nasa.gov), Джей Паркер (Jay Parker, jay.w.parker@jpl.nasa.gov), Лаборатория реактивного движения (Jet Propulsion Laboratory, JPL) Национального управления по аэронавтике и исследованию космического пространства США (НАСА)	
Акторы/заинтересованные лица, их роли и ответственность	Группа проекта НАСА UAVSAR, группа проекта НАСА QuakeSim, Распределенный активный архивный центр (Distributed Active Archive Center, DAAC) Спутникового центра НАСА на Аляске (Alaska Satellite Facility, ASF), Геологическая служба США (United States Geological Survey, USGS), Геологическая служба штата Калифорния (California Geological Survey)	
Цели	Использование радиолокатора с синтезированной апертурой для выявления изменений ландшафта, вызванных сейсмической активностью, оползнями, обезлесением, изменениями растительности, наводнениями и т. д.; повышение удобства его использования и доступности для ученых	
Описание варианта использования	Ученый, желающий исследовать последствия землетрясения, изучает несколько стандартных результатов обработки данных радиолокатора с синтезированной апертурой, предоставляемых НАСА. Для ученого может оказаться полезным взаимодействие с сервисами, предоставляемыми проектами-посредниками, которые повышают ценность материалов официального архива результатов обработки данных	
Текущие решения	**Вычислительная система**	Обработка первичных данных в Научно-исследовательском центре НАСА им. Эймса (Ames Research Center, ARC) на суперкомпьютерах Pleiades и Endeavour. Изучается вопрос об использовании коммерческих облачных систем для хранения и в качестве пользовательского интерфейса
	Хранилище данных	На основе файлов
	Сеть связи	Требуется однократная передача данных от измерительного прибора в Лабораторию реактивного движения (Jet Propulsion Laboratory, JPL), из JPL в другие вычислительные центры НАСА (ARC), а также из JPL в Спутниковый центр НАСА на Аляске (ASF). Отдельные файлы данных не слишком велики и могут быть скачаны индивидуальными пользователями, однако передача всего набора данных является неподъемной задачей. Это является проблемой для других проектов, работающих с полученными данными, таких как QuakeSim, которые хотят переформатировать наборы данных и повысить их ценность
	Программное обеспечение	Инструменты ROI_PAC, GeoServer, GDAL, а также инструменты, поддерживающие стандарт метаданных GeoTIFF

Характеристики больших данных	**Источник данных (распределенный/ централизованный)**	Данные изначально должны собираться беспилотными летательными аппаратами. Первоначальная обработка данных проводится в Лаборатории реактивного движения НАСА (JPL). Архив данных централизованно хранится в Распределенном активном архивном центре Спутникового центра НАСА на Аляске (DAAC ASF), Группа проекта QuakeSim поддерживает отдельные результаты дальнейшей обработки данных (конверсия в GeoTIFF)
	Объем (количество)	Данные многопроходной интерферометрической съемки (Repeat Pass Interferometry, RPI): ~ 3 терабайт. Темпы прироста около 1—2 терабайт в год. Поляриметрические данные: первичные — 100 терабайт, обработанные ~40 терабайт. Предлагаемые спутниковые программы радиолокационного зондирования Земли (Earth Radar Mission, ранее DESDynI) способны значительно увеличить объемы производства данных (до нескольких терабайт в день)
	Скорость обработки (например, в реальном времени)	Данные многопроходной интерферометрической съемки: 1—2 терабайта в год. Темпы производства поляриметрических данных еще выше
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Два основных типа: поляриметрические данные и данные RPI. Каждый продукт данных RPI представляет собой набор файлов (файл аннотации, распакованные фалы и т.д.). Поляриметрические продукты данных также состоят из нескольких файлов каждый
	Вариативность (темпы изменения)	Продукты данных изменяются медленно. Данные время от времени обрабатываются повторно: используются новые методы обработки или иные параметры. Возможно использование дополнительных мер обеспечения и контроля качества
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Следует принимать во внимание вопросы происхождения данных. В прошлом происхождение не было прозрачным для нижестоящих потребителей данных. В настоящее время используется управление версиями; версии описываются на вебстранице UAVSAR в примечаниях
	Визуализация	Используются инструменты, сервисы и стандарта геопространственной информационной системы
	Качество данных (синтаксис)	Многие кадры и коллекции оказались непригодными для использования из-за непредвиденных полетных условий

Наука о больших данных (сбор, курирование, анализ, операции)	**Типы данных**	GeoTIFF и взаимосвязанные данные изображений
	Аналитика данных	Анализ (такой, как выделение контуров) выполняется нижестоящими потребителями данных: это проблемы исследования
Иные проблемы больших данных	Конвейер обработки данных требует контроля и вмешательства человека. Имеются ограниченные нисходящие конвейеры специализированной обработки данных для отдельных пользователей. Следует изучить и внедрить облачные архитектуры для распространения целых коллекций продуктов данных среди потребителей	
Проблемы пользовательского интерфейса и мобильного доступа	Некоторые пользователи изучают данные в полевых условиях на мобильных устройствах, что требует интерактивной переработки больших наборов данных в понятные изображения или статистику	
Технические проблемы обеспечения безопасности и защиты персональных данных	Данные делаются публично доступными сразу после обработки (без периода эмбарго)	
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Данные привязываются к географическим координатам и могут быть указаны в угловых координатах. Категории: ГИС; конвейер обработки данных стандартными средствами для производства стандартных продуктов данных	
Дополнительная информация (гиперссылки)	Сайт проекта UAVSAR, https://uavsar.jpl.nasa.gov/ Сайт Спутникового центра НАСА на Аляске (Alaska Satellite Facility, ASF), https://asf.alaska.edu/ Страница Википедии о проекте QuakeSim, https://en.wikipedia.org/wiki/QuakeSim	

А.8.5 Вариант использования № 45: Объединенный испытательный стенд iRODS центра НАСА в Лэнгли и Центра космических полетов имени Годдарда

Название	Объединенный испытательный стенд iRODS центра НАСА в Лэнгли и Центра космических полетов имени Годдарда
Предметная область	Исследования и приложения наук о Земле
Автор/организация/эл.почта	Майкл Литтл (Michael Little, M.M.Little@NASA.gov), Роджер Дюбуа (Roger Dubois, Roger.A.Dubois@nasa.gov), Бренди Квам (Brandi Quam, Brandi.M.Quam@NASA.gov), Тиффани Мэтьюз (Tiffany Mathews, Tiffany.J.Mathews@NASA.gov), Андрей Вахнин (Andrei Vakhnin, Andrei.A.Vakhnin@NASA.gov), Бет Хаффер (Beth Huffer), Кристиан Джонсон (Christian Johnson) / Исследовательский центр в Ленгли, НАСА (LaRC) Джон Шнас (John Schnase, John.L.Schnase@NASA.gov), Даниэль Даффи (Daniel Duffy, Daniel.Q.Duffy@NASA.gov), Гленн Тамкин (Glenn Tamkin, Glenn.S.Tamkin@nasa.gov), Скотт Синно (Scott Sinno, Scott.S.Sinno@nasa.gov), Джон Томпсон (John Thompson, John.H.Thompson@nasa.gov) и Марк Макинерни (Mark McInerney, Mark.Mcinerney@nasa.gov) / Центр управления полетами имени Годдарда, США (GSFC)
Акторы/заинтересованные лица, их роли и ответственность	Центр обработки атмосферных данных (ASDC) в Исследовательском центре в Ленгли, НАСА (LaRC) в г. Хэмптоне (Hampton), штат Вирджиния и Центр моделирования климата НАСА (NCCS) в Центре управления полетами имени Годдарда, США (GSFC) принимают, архивируют и распространяют данные, являющиеся ключевыми по важности для заинтересованных сторон, в число которых входят сообщество по исследованиям климата, сообщество по прикладному применению науки и растущее сообщество клиентов из государственного и частного секторов, которым необходимы данные об атмосфере и климате

Цели	Реализация возможности объединения данных с целью совершенствования и автоматизации поиска по неоднородным данным, уменьшения задержек при передаче данных и соответствия индивидуализируемым критериям, касающимся контента данных, их качества, метаданных и производства данных. Поддержка / создание возможностей для приложений и клиентов, которым требуется интеграция ряда неоднородных наборов данных	
Описание варианта использования	Центр моделирования климата НАСА (NCCS) и Центр обработки атмосферных данных (ASDC) Национального управления по аэронавтике и исследованию космического пространства, США (NASA) имеют в своем распоряжении взаимодополняющие друг друга наборы данных огромного объема, ввиду чего по этим данным трудно выполнять запросы, и ими сложно обмениваться. Исследователям климата, специалистам по прогнозированию погоды, группам разработки и обслуживания измерительной аппаратуры и другим специалистам нужен доступ к данным из нескольких наборов данных с тем, чтобы сравнивать показания датчиков различных измерительных инструментов, сопоставлять показания датчиков с результатами моделирования, калибровать приборы, искать корреляции между несколькими параметрами и т. д. В настоящее время усилия по анализу, визуализации и иной обработке данных из неоднородных наборов данных требуют много времени. Ученым приходится отдельно получать доступ, искать и загружать данные с каждого из нескольких серверов. Данные часто дублируются, при этом непонятно, какой источник считать авторитетным. Многие ученые отмечают, что они тратят больше времени на получение доступа к данным, чем научные исследования. Потребителям данных нужны механизмы для извлечения неоднородных данных с использованием единой точки доступа. Такую возможность можно реализовать с помощью iRODS — программного обеспечения грида данных, поддерживающего параллельную загрузку наборов данных с выбранных серверов копий (replica servers), которые могут быть географически распределены, но при этом доступны пользователям со всего мира. Используя iRODS в сочетании с семантически обогащенными метаданными, управляемыми на основе высокоточной онтологии НАСА для наук о Земле, данные системы «Онлайновые инструменты работы с данными» () центра ASDC будут объединены с данными NCCS в Центре управления полетами имени Годдарда, США (GSFC). Неоднородные продукты данных этих двух центров НАСА семантически аннотируются на основе общих концепций онтологии НАСА для наук о Земле. Семантические аннотации позволят системе iRODS идентифицировать взаимодополняющие наборы данных и агрегировать данные из этих разрозненных источников, облегчая обмен данными между специалистами по моделированию климата, синоптиками, представителями наук о Земле и учеными из других дисциплин, нуждающимися в данных науки о Земле. Система объединения данных iRODS также будет поддерживать облачные сервисы обработки данных в облаке Amazon Web Services (AWS)	
Текущие решения	**Вычислительная система**	Центр моделирования климата НАСА (NCCS) и Центр обработки атмосферных данных (ASDC): две файловые системы GPFS

Текущие решения	**Хранилище данных**	Файловая система GPFS решения «Онлайновые инструменты работы с данными» (DPO) центра ASDC состоит из 12 подсистем хранения IBM DC4800 и 6 подсистем IBM DCS3700, 144 ядер Intel 2,4 ГГц и 1400 терабайт располагаемой памяти на жестких дисках. Данные центра NCCS хранятся в кластере NCCS MERRA, который представляет собой кластер Dell с 36 узлами, 576 ядрами Intel 2.6 ГГц SandyBridge, 1300 терабайт дисковой памяти, 1250 гигабайт оперативной памяти, с теоретической пиковой вычислительной мощностью 11,7 терафлопс
	Сеть связи	Комбинация оптоволоконной связи в SAN-сети систем хранения данных (Storage Area Network) и 10 гигабит/с в локальной сети. Узлы кластера NCCS связаны сетью Infiniband FDR с максимальной скоростью TCP/IP соединений более 20 гигабит/с
	Программное обеспечение	SGE Univa Grid Engine версии 8.1, iRODS версии 3.2 и/или 3.3, файловая система IBM General Parallel File System (GPFS) версии 3.4, Cloudera версии 4.5.2-1
Характеристики больших данных	**Источник данных (распределенный/централизованный)**	iRODS будет использоваться для обмена данными, собранными из продуктов данных проекта CERES уровня 3B, включая продукты CERES EBAF-TOA и CERES-Surface. Поверхностные потоки в EBAF-Surface получены на основе двух продуктов данных CERES: 1) CERES SYN1deg- Month Ed.3, который предоставляет рассчитанные поверхностные потоки, подлежащие корректировке, и 2) CERES EBAF-TOA Ed.2.7, который использует данные наблюдений, чтобы обеспечить полученные из CERES ограничения потока в верхних слоях атмосферы (TOA). Доступ к этим продуктам позволит центру NCCS использовать данные из продуктов в моделировании для получения «ассимилированного» потока. NCCS представит объединенному стенду iRODS данные из приложения «Система для ретроспективного анализа современной эры для исследований и приложений» (MERRA). MERRA объединяет данные наблюдений с данными численного моделирования для получения глобального, согласованного во времени и в пространстве синтез значений 26 ключевых климатических параметров. Файлы данных MERRA создаются на основе модели «Годдардовская система наблюдения Земли, 5-я версия» (GEOS-5) и хранятся в форматах HDF-EOS и NetCDF (Network Common Data Form).

Характеристики больших данных	**Источник данных (распределенный/ централизованный)**	Пространственное разрешение составляет 1/2° широты × 2/3° долготы × 72 вертикальных уровня, охватывающих стратосферу. Временное разрешение составляет 6 часов для трехмерного, полного пространственного разрешения, начиная с 1979 года по настоящее время, т. е. охватывая почти всю спутниковую эпоху. Каждый файл содержит одну сетку с несколькими 2D и 3D-переменными. Все данные хранятся в сетке долгота — широта, с вертикальным измерением, применимым ко всем 3D-переменным. Продукты MERRA на основе модели GEOS-5 разделены на 25 коллекций: 18 стандартных продуктов, химические продукты. Коллекции включают файлы со среднемесячными величинами и ежедневные файлы с шестичасовыми интервалами, начиная с 1979 по 2012 год. Данные MERRA обычно упаковываются как многомерные двоичные данные в самодокументированный файловый формат NetCDF. Иерархические метаданные в заголовке NetCDF содержат информацию о представлении, которая позволяет работать с данными программному обеспечению с поддержкой NetCDF. Они также содержат произвольную информацию о мерах обеспечения долговременной сохранности и о политике, которую можно использовать для обеспечения соответствия установленным требованиям с учетом специфики использования данных
	Объем (количество)	В настоящее время объем данных в продукте EBAF-TOA составляют около 420 мегабайт, а в продукте EBAF-Surface — около 690 мегабайт. Объемы данных увеличиваются с каждым обновлением версии (примерно каждые полгода). Коллекция MERRA в общей сложности содержит около 160 терабайт данных в несжатом виде; в сжатом виде объем составляет ~80 терабайт
	Скорость обработки (например, в реальном времени)	Периодически, поскольку обновления выполняются при выпуске каждой новой версии
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Существует потребность во многих типах приложений для объединения данных реанализа MERRA с другими данными повторного анализа и данными наблюдений, такими как данные CERES. NCCS использует эталонный стандарт CMIP5 «Проекта сопоставления связанных климатических моделей» для обеспечения онтологической согласованности нескольких разнородным наборов данных
	Вариативность (темпы изменения)	Объем данных реанализа MERRA увеличивается примерно на 1 терабайт в месяц

Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Валидация и тестирование семантических метаданных и объединенных продуктов данных будут осуществляться производителями данных в Научно-исследовательском центре НАСА в Лэнгли и в Центре космических полетов имени Годдарда посредством регулярного тестирования. Для того чтобы убедиться, что обновления и изменения в системе iRODS, добавленные новые источники данных и/или добавленные новые метаданные не привносят ошибок в объединенные продукты данных, будет реализовано регрессионное тестирование. Валидация данных MERRA обеспечивается производителем данных — Отделом глобального моделирования и ассимиляции Центра управления полетами имени Годдарда, НАСА (GMAO)
	Визуализация	В научном сообществе растет потребность в сервисах управления данными и их визуализации, способных агрегировать данные из нескольких источников и отображать их в рамках единого графического представления. В настоящее время развитию таких возможностей препятствует проблема поиска и скачивания сопоставимых данных с нескольких серверов, а затем преобразования каждого из разнородных наборов данных таким образом, чтобы сделать его пригодным для использования программным обеспечением для визуализации. Объединение наборов данных НАСА посредством системы iRODS даст ученым возможность быстро находить и агрегировать сопоставимые наборы данных для их использования совместно с программным обеспечением для визуализации
	Качество данных (синтаксис)	Для данных MERRA, контроль качества обеспечивается производителем данных — Отделом глобального моделирования и ассимиляции Центра управления полетами имени Годдарда, НАСА (GMAO)
	Типы данных	См. выше
	Аналитика данных	В соответствии с первой целью повышения доступности и удобства выявления данных посредством применения инновационных технологий, Центр обработки атмосферных данных (ASDC) и Центр моделирования климата НАСА (NCCS) изучают пути совершенствования способов доступа к данным. Используя iRODS, данные системы «Онлайновые инструменты работы с данными» (DPO) центра ASDC могут объединяться с данными NCCS в Центре управления полетами имени Годдарда, США (GSFC), формируя систему доступа к данным, способную обслуживать гораздо более широкую клиентскую базу, чем в настоящее время.

Наука о больших данных (сбор, курирование, анализ, операции)	**Аналитика данных**	Объединение и обмен информацией позволят центрам ASDC и NCCS в полной мере использовать мультиинструментальные данные многолетних наблюдений, а также улучшат и автоматизируют процесс поиска и выявления среди разнородных данных, уменьшат задержки при передаче данных и обеспечат соответствие индивидуализируемым критериям, касающимся контента данных, их качества, метаданных и производства данных
Иные проблемы больших данных		
Проблемы пользовательского интерфейса и мобильного доступа	Главная проблема заключается в определении корпоративной архитектуры, способной обеспечить аналитику в режиме реального времени посредством информационного обмена с несколькими API-интерфейсами и системами облачных вычислений. Если исходить из того, что вычислительные ресурсы будут располагаться в облачных системах, то проблема мобильности сводится к тому, чтобы не перегружать мобильные устройства отображением визуализации, требующей интенсивной загрузки центрального процессора, что может понизить эффективность или удобство использования представляемых пользователю данных	
Технические проблемы обеспечения безопасности и защиты персональных данных		
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Это объединение данных опирается на опыт нескольких лет исследований и разработок, связанных с решением iRODS, которые выполнялись в центре NCCS. За это время NCCS проверил функциональные возможности iRODS, одновременно расширяя состав базовых функциональных возможностей решения посредством добавления возможностей, специфических для данной области применения. Например, NCCS создал и установил в iRODS научные инструментальные наборы на основе Python, которые автоматически собирают метаданные при регистрации соответствующей коллекции данных. Один из этих инструментальных наборов был разработан для коллекции MERRA. В сочетании с iRODS он повышает отдачу от объединения данных LaRC / GSFC, предоставляя расширенные возможности поиска. Центр в Ленгли работает над созданием усовершенствованной архитектуры, использующей несколько пилотных технологий и инструментов (для доступа, обнаружения и анализа), и предназначенной для интеграции возможностей всего сообщества специалистов в области наук о Земле. Исследования и разработки, выполненные двумя центрами обработки данных, дополняют друг друга и лишь дальше улучшают данный вариант использования. В число других разработанных научных инструментальных наборов входят наборы для поддержки формата NetCDF и для взаимодействия с «Межправительственной группой экспертов по изменению климата» (IPCC) и проектом «Моделирование динамики океана с усвоением данных наблюдений» (Ocean Modeling and Data Assimilation, ODAS). Комбинация iRODS и этих научных инструментальных наборов позволила создать конфигурируемый технологический стек, так называемый «виртуальный сервер климатических данных» (virtual Climate Data Server, vCDS), — это означает, что данная среда реального времени может быть развернута в различных условиях (например, на «голом железе», виртуальных серверах, в облаке) для удовлетворения различных научных потребностей. Виртуальный сервер vCDS, который можно рассматривать как эталонную архитектуру, способствующую объединению разрозненных хранилищ данных, используется в центрах LaRC и GSFC (и не только в них)	
Дополнительная информация (гиперссылки)	Для получения дополнительной информации, пожалуйста, свяжитесь с авторами	

А.8.6 Вариант использования № 46: Аналитические сервисы MERRA

Название	Аналитические сервисы MERRA (MERRA/AS)	
Предметная область	Научные исследования: науки о Земле	
Автор/организация/эл.почта	Джон Шнас (John L. Schnase, John.L.Schnase@NASA.gov) и Дэниел Даффи (Daniel Q. Duffy, Daniel.Q.Duffy@NASA.gov), Центр управления полетами имени Годдарда, США (GSFC), США	
Акторы/заинтересованные лица, их роли и ответственность	В рамках проекта Национального управления по аэронавтике и исследованию космического пространства США (НАСА) «Система для ретроспективного анализа современной эры для исследований и приложений» (MERRA) осуществляется глобальный, согласованный во времени и в пространстве синтез значений 26 ключевых климатических параметров путем объединения результатов численного моделирования с данными наблюдений. К числу действующих лиц и заинтересованных сторон проекта MERRA относятся сообщество по исследованиям климата, сообщество по прикладному применению науки и растущее сообщество клиентов из государственного и частного секторов, которым необходимы данные MERRA для их систем поддержки принятия решений	
Цели	Повысить удобство и степень использования крупномасштабных коллекций научных данных, таких как MERRA	
Описание варианта использования	Аналитические сервисы MERRA (MERRA/AS) дают возможность использовать средства аналитики Map/Reduce для обработки данных коллекции MERRA. MERRA/AS является примером опирающегося на облачные технологии аналитики климата как сервиса (CAaaS), который представляет собой подход к решению связанных с большими данными проблем в области климатологии посредством совместного использования: 1) высокопроизводительной аналитики, осуществляемой близко к месту хранения данных, 2) масштабируемого управления данными, 3) виртуализации программных инструментов, 4) адаптивной аналитики, и 5) API-интерфейсов, гармонизированных с потребностями прикладной области. Эффективность MERRA/AS демонстрируется в ряде приложений, включая публикацию данных в «Федеративной грид-системе обработки данных о Земле» (ESGF) с целью поддержки исследований Межправительственной группы экспертов по изменению климата (IPCC), систем поддержки принятия решений по восстановлению экосистем (RECOVER), поддержки принятия решений НАСА и Министерства внутренних дел США, связанных с предотвращением ущерба от природных пожаров, и работ по оценке тестового стенда для обеспечения интероперабельности данных Центра космических полетов имени Годдарда и Центр обработки атмосферных данных (ASDC) в Лэнгли	
Текущие решения	**Вычислительная система**	Центр моделирования климата НАСА (NCCS) в Центре управления полетами имени Годдарда, США (GSFC)
	Хранилище данных	Файловая система Hadoop (HDFS) «Аналитических сервисов MERRA» (MERRA/AS) (HDFS) представляет собой кластер Dell с 36 узлами, 576 ядрами Intel 2.6 ГГц SandyBridge, 1300 терабайт дисковой памяти, 1250 гигабайт оперативной памяти, с теоретической пиковой вычислительной мощностью 11,7 терафлопс
	Сеть связи	Узлы кластера соединены сетью Infiniband FDR с максимальной скоростью TCP/IP соединений более 20 гигабит/с
	Программное обеспечение	Cloudera, iRODS, Amazon AWS

Характеристики больших данных	**Источник данных (распределенный/ централизованный)**	Файлы данных MERRA создаются на основе модели «Годдардовская система наблюдения Земли, 5-я версия» (GEOS-5) и хранятся в форматах HDF-EOS и NetCDF (Network Common Data Form). Пространственное разрешение составляет 1/2° широты × 2/3° долготы × 72 вертикальных уровня, охватывающих стратосферу. Временное разрешение составляет 6 часов для трехмерного, полного пространственного разрешения с 1979 г. по настоящее время, т. е. охватывая почти всю спутниковую эпоху. Каждый файл содержит одну сетку с несколькими 2D и 3D — переменными. Все данные хранятся в сетке долгота — широта, с вертикальным измерением, применимым ко всем 3D-переменным. Продукты MERRA на основе модели GEOS5 разделены на 25 коллекций: 18 стандартных продуктов, химические продукты. Коллекции включают файлы со среднемесячными величинами и ежедневные файлы с шестичасовыми интервалами, начиная с 1979 по 2012 год. Данные MERRA обычно упаковываются как многомерные двоичные данные в самодокументированный файловый формат NetCDF. Иерархические метаданные в заголовке NetCDF содержат информацию о представлении, которая позволяет работать с данными программному обеспечению с поддержкой NetCDF. Они также содержат произвольную информацию о мерах обеспечения долговременной сохранности и о политике, которую можно использовать для обеспечения соответствия установленным требованиям с учетом специфики использования данных
	Объем (количество)	480 терабайт
	Скорость обработки (например, в реальном времени)	Обработка в режиме реального времени либо пакетная, в зависимости от вида анализа. Мы разрабатываем набор «канонических операций», выполняемых на ранней стадии обработки, близко к месту нахождения данных, общих для многих потоков рабочих процессов аналитики. Цель заключается в том, чтобы канонические операции выполнялись в почти реальном времени
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Во многих типах приложений существует потребность для объединения данных реанализа MERRA с другими данными повторного анализа и данными наблюдений, мы используем эталонный стандарт CMIP5 «Проекта сопоставления связанных климатических моделей» для обеспечения онтологической согласованности нескольких разнородным наборов данных
	Вариативность (темпы изменения)	Объем данных реанализа MERRA увеличивается примерно на 1 терабайт в месяц

Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Валидация обеспечивается производителем данных — Отделом глобального моделирования и ассимиляции Центра управления полетами имени Годдарда, НАСА (GMAO)
	Визуализация	Существует растущая потребность в распределенной визуализации результатов аналитики
	Качество данных (синтаксис)	Контроль качества обеспечивается производителем данных — Отделом глобального моделирования и ассимиляции Центра управления полетами имени Годдарда, НАСА (GMAO)
	Типы данных	См. выше
	Аналитика данных	В наших усилиях по решению проблем больших данных в климатологии, мы движемся в сторону концепции аналитики климата как сервиса (CAaaS). Мы концентрируем внимание на аналитике, потому что в итоге приносят пользу обществу именно знания, полученные в результате нашего взаимодействия с большими данными. Мы ориентируемся на CAaaS, поскольку считаем, что эта концепция является полезным способом осмысления проблемы: это специализация концепции бизнес-процесса как услуги, представляющая собой эволюционирующее расширение поддерживаемых облачными вычислениями моделей IaaS, PaaS и SaaS
Иные проблемы больших данных	Большой вопрос заключается в том, как использовать облачные вычисления таким образом, чтобы обеспечить лучшее использование наземных вычислительных ресурсов и ресурсов данных в области климатологии. Облачные вычисления предоставляют нам новый уровень в стеке услуг обработки данных — облачный слой, в котором осуществляется гибкая настройка, и продукты данных корпоративного уровня преобразуются с тем, чтобы удовлетворить специфические потребности приложений и потребителей. Это помогает нам преодолевать разрыв между миром традиционных высокопроизводительных вычислений, который, по крайней мере на данный момент времени, обитает в тонко настроенной среде моделирования климата на корпоративном уровне, — и нашими новыми клиентами, на чьи ожидания и методы работы все сильнее влияет мега-тенденция «умной мобильности»	
Проблемы пользовательского интерфейса и мобильного доступа	Большинство современных смартфонов, планшетов и иных устройств фактически состоят лишь из дисплея и компонентов пользовательского интерфейса к сложным приложениям, выполняемым в облачных центрах обработки данных. Концепция CAaaS призвана обеспечить поддержку именно этого стиля работы	
Технические проблемы обеспечения безопасности и защиты персональных данных	На данный момент времени каких-либо критичных проблем не выявлено	

Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Инструменты Map/Reduce и iRODS кардинально упрощают проведение анализ и агрегирование данных. Наш подход к виртуализации программных инструментов упрощает предоставление соответствующих возможностей новым пользователям и упрощает для них создание новых приложений. Социальное конструирование расширенных возможностей, поддерживаемое концепцией «канонических операций», обеспечивает адаптивность; а API-интерфейсы к сервисам климатических данных (Climate Data Services), который мы разрабатываем, обеспечивают простоту освоения. Мы считаем, что в совокупности эти базовые технологии, лежащие в основе концепции CAaaS, создают подходящие условия, при которых благодаря вкладу различных людей и групп (действующих как сообща, так и самостоятельно) могут появиться возможности, помогающие решить проблемы больших данных в климатологии
Дополнительная информация (гиперссылки)	Для получения дополнительной информации, пожалуйста, свяжитесь с авторами

А.8.7 Вариант использования № 47: Атмосферная турбулентность — Обнаружение событий

Название	Атмосферная турбулентность — Обнаружение событий и прогностическая аналитика	
Предметная область	Научные исследования — науки о Земле	
Автор/организация/эл.почта	Майкл Сиблом (Michael Seablom), штаб-квартира Национального управления по аэронавтике и исследованию космического пространства США (НАСА), michael.s.seablom@nasa.gov	
Акторы/заинтересованные лица, их роли и ответственность	Исследователи, получившие гранты НАСА или Национального научного фонда (National Science Foundation, NSF), синоптики, представители авиационной отрасли (в общем случае — любой исследователь, который участвует в изучении событий, связанных с погодными явлениями)	
Цели	Создать возможности для обнаружения оказывающих сильное воздействие погодных явлений, данные о которых могут быть найдены в объемных хранилищах данных наук о Земле, и которые трудно охарактеризовать с использованием традиционных численных методов (например, турбулентность). Сопоставить такие явления с продуктами глобального ретроспективного анализа атмосферных данных, с целью совершенствования возможностей прогнозирования	
Описание варианта использования	Отчеты о турбулентности с летательных аппаратов (из отчетов пилотов, либо из автоматических измерений на летательных аппаратах скорости диссипации вихрей) сопоставляются с недавно завершенным ретроспективным анализом атмосферных данных всей эпохи спутниковых наблюдений. Продукты включают данные Реанализа метеорологических данных для региона Северной Америки (NARR) и «Системы для ретроспективного анализа современной эры для исследований и приложений» (MERRA) — проектов Национального управления по аэронавтике и исследованию космического пространства США (НАСА)	
Текущие решения	**Вычислительная система**	Платформа НАСА для обмена данными о Земле (NEX), в т. ч. суперкомпьютер Pleiades
	Хранилище данных	Объем каждого из продуктов ретроспективного анализа составляет порядка 100 терабайт; по сравнению с этим объем данных о турбулентности незначителен

Текущие решения	**Сеть связи**	Наборы данных ретроспективного анализа, вероятно, будут слишком большими, чтобы переместить их на предпочтительный суперкомпьютер (в данном случае, NEX), поэтому потребуется максимально быстрая сеть
	Программное обеспечение	Инструмент Map/Reduce или аналогичный; SciDB или другая научная СУБД
Характеристики больших данных	**Источник данных (распределенный/централизованный)**	Распределенный
	Объем (количество)	Текущий объем 200 терабайт, через 5 лет — 500 терабайт
	Скорость обработки (например, в реальном времени)	Данные анализируются по частям
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Массивы данных ретроспективного анализа не согласованы по формату, разрешению, семантике и метаданным. Скорее всего, придется интерпретировать / анализировать каждый из этих входных потоков для включения в общий продукт
	Вариативность (темпы изменения)	Данные наблюдений за турбулентностью будут постоянно обновляться; продукты ретроспективного анализа выпускаются примерно раз в пять лет
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Валидация будет необходима для итогового продукта (корреляции)
	Визуализация	Полезна для интерпретации результатов
	Качество данных (синтаксис)	Входные потоки уже должны были пройти контроль качества
	Типы данных	Вывод в виде данных на сетке из систем ассимиляции атмосферных данных, и текстовые данные из наблюдений турбулентности
	Аналитика данных	Язык спецификации событий необходим для интеллектуального анализа данных/поиска событий
Иные проблемы больших данных	Семантика (интерпретация множества продуктов ретроспективного анализа); перемещение данных; базы данных с оптимальной структурой для 4-мерного интеллектуального анализа данных	
Проблемы пользовательского интерфейса и мобильного доступа	Разработки для мобильных платформ в настоящее время не являются остро необходимыми	
Технические проблемы обеспечения безопасности и защиты персональных данных	Критичных проблем не выявлено	

Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Атмосферная турбулентность является лишь одним из многих погодных явлений, которые могут быть полезны для понимания аномалий в атмосфере или океане, взаимосвязанных друг с другом на больших расстояниях в пространстве и во времени. Однако описанный процесс имеет пределы расширяемости, т. е. для каждого явления могут потребоваться совершенно разные процессы для интеллектуального анализа данных и прогнозирования
Дополнительная информация (гиперссылки)	Роберт Стюарт (Robert Stewart) «Океанография в 21-м веке — онлайн-учебник» (Oceanography in the 21-st Century — An Online Textbook), раздел «Телевзаимосвязи Эль-Ниньо» (El Niño Teleconnections), см. http://pro.unibz.it/staff2/fzavatti/corso/teleconnections.htm Тодд Вуди (Todd Woody) «Познакомьтесь с учеными, проводящими интеллектуальный анализ больших данных с целью предсказания погоды» (Meet The Scientists Mining Big Data To Predict The Weather), журнал «Форбс», 21 марта 2012 года, https://www.forbes.com/sites/toddwoody/2012/03/21/meet-the-scientists-mining-big-data-to-predict-the-weather/

А.8.8 Вариант использования № 48: Исследования климата с использованием модели климатической системы Земли (CESM)

Название	Исследования климата с использованием модели климатической системы Земли (CESM) в Национальном научно-исследовательском вычислительном центре энергетических исследований Министерства энергетики США (NERSC)	
Предметная область	Научные исследования: климат	
Автор/организация/эл.почта	Научный руководитель проекта: Уоррен Вашингтон (Warren Washington), Национальный центр атмосферных исследований (National Center for Atmospheric Research, NCAR), США	
Акторы/заинтересованные лица, их роли и ответственность	Климатологи, принимающие решения лица в США	
Цели	Цели группы по прогнозированию изменения климата (Climate Change Prediction, CCP) в Национальном центре атмосферных исследований США (National Center for Atmospheric Research, NCAR) заключаются в том, чтобы понять и количественно оценить вклад естественных и антропогенно-обусловленных типовых сценариев изменчивости и изменения климата в 20-м и 21-м столетиях посредством моделирования с использованием модели климатической системы Земли (CESM)	
Описание варианта использования	С помощью моделирования исследователи могут изучать механизмы изменчивости и изменения климата, выявлять и атрибутировать изменения климата в прошлом, а также прогнозировать и предсказывать будущие изменения. Проведение моделирования мотивировано интересом со стороны общества, и оно широко используется национальными и международными научно-исследовательскими сообществами	
Текущие решения	**Вычислительная система**	Вычисления в центре NERSC (24 млн часов), в ведущих вычислительных центрах (Leadership Computing Facility, LCF) Министерства энергетики США (41 млн часов), в Лаборатории моделирования климата (Climate Simulation Laboratory, CSL) Национального центра атмосферных исследований США (National Center for Atmospheric Research, NCAR) (17 млн часов)
	Хранилище данных	1,5 петабайта в центре NERSC
	Сеть связи	ESNet

Текущие решения	**Программное обеспечение**	Разработанные центром NCAR библиотека параллельного ввода-вывода и утилиты «NCAR Командный язык» (NCAR Command Language, NCL) и «NetCDF-операторы» (NetCDF Operators, NCO); параллельные библиотеки NetCDF
Характеристики больших данных	**Источник данных (распределенный/централизованный)**	Данные производятся в вычислительных центрах. Грид-система обработки данных о Земле (ESG) — это проект с открытым исходным кодом, обеспечивающий надежную, распределенную платформу для вычислений и хранения данных, а также всемирный доступ к научным данным в масштабе пета / экза ESGF управляет первой в мире децентрализованной базой данных для работы с климатологическими данными, содержащей многие петабайты данных в десятках объединенных в грид-центров по всему миру. Эта инфраструктура считается ведущей в плане управления и обеспечения доступа к большим распределенным объемам данных, используемых в исследованиях в области изменения климата. Она поддерживает «Проект сопоставления связанных климатических моделей» (CMIP), протоколы которого обеспечиваются периодическими оценками, выполняемых «Межправительственной группой экспертов по изменению климата» (IPCC)
	Объем (количество)	30 петабайт в центре NERSC (при условии проведения 15 сквозных экспериментов по теме изменения климата) к 2017 г.; во много раз больше по всему миру
	Скорость обработки (например, в реальном времени)	В ходе моделирований производится 42 гигабайта/с
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Необходимо сопоставлять друг с другом данные наблюдений, данные исторического ретроспективного анализа и результаты ряда независимо выполненных моделирований. В рамках «Программы диагностики и взаимного сопоставления климатических моделей» (Program for Climate Model Diagnosis and Intercomparison, PCMDI) разрабатываются методы и инструменты для диагностики и взаимного сравнения моделей общей циркуляции (general circulation model, GCM), которые моделируют глобальный климат. Потребность в инновационном анализе результатов моделирования климата на основе GCM-моделей очевидна, так как разрабатываются все более сложные модели, в то время, как расхождения между соответствующими результатами моделирования, и этих результатов — с климатическими наблюдениями остаются значительными и плохо

Характеристики больших данных	**Разнообразие (множество наборов данных, комбинация данных из различных источников)**	понятыми. Характер и причины этих расхождений должны учитываться систематическим образом для того, чтобы уверенно использовать GCM-модели для моделирования предполагаемых глобальных изменений климата
	Вариативность (темпы изменения)	Данные производятся с помощью программ, исполняемых в суперкомпьютерных центрах. Во время исполнения регулярно бывают периоды интенсивного ввода/вывода данных, однако обычно они составляют всего несколько процентов от общего времени вычислений. Прогоны моделирования выполняются регулярно, но их количество подскакивает при приближении окончательных сроков сдачи отчетов
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Данные, полученные в результате выполнения моделирования климата, существенно влияют на ход дискуссий по вопросам моделирования изменения климата. Следовательно, эти данные должны быть стабильными и надежными как с точки зрения обеспечения научно обоснованного представления влияющих на климат процессов, так и с учетом того, что эти данные хранятся в течение длительного времени и передаются по всему миру партнерам и другим ученым
	Визуализация	Визуализация имеет решающее значение для понимания столь сложной системы, как экосистема Земли
	Качество данных (синтаксис)	См. пункт «Достоверность» выше
	Типы данных	Специалисты по земной системе тонут в стремительно возрастающих объемах данных, рост производства которых вызван постоянно увеличивающимся разрешением как глобальных моделей, так и дистанционных датчиков
	Аналитика данных	Существует потребность в предоставлении через грид-систему обработки данных о Земле (ESG веб-сервисов по редуцированию данных (пересчету в физические величины) и их анализу возникает острая необходимость в развитии средств анализа данных, тесно взаимосвязанных с архивами данных
Иные проблемы больших данных	Быстро растущие объемы наборов данных затрудняют проведение научного анализа. Потребность в регистрации данных моделирования опережает способность суперкомпьютеров удовлетворить эту потребность	
Проблемы пользовательского интерфейса и мобильного доступа	Данные, полученные в результате моделирования и наблюдений, должны быть распространены среди большого, широко распределенного сообщества	

Технические проблемы обеспечения безопасности и защиты персональных данных	
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Проект «Федеративная грид-система обработки данных о Земле» (ESGF) находится на ранней стадии адаптации для использования в двух дополнительных областях: биологии (для ускорения проектирования и разработки лекарств) и энергетики (инфраструктура для проекта «Энергетические системы Калифорнии для 21-го века» (California Energy Systems for the 21st Century (CES-21)).
Дополнительная информация (гиперссылки)	Сайт проекта «Объединенный грид данных земной системы» (Earth System Grid Federation, ESGF), https://esgf.llnl.gov/ Сайт «Программы диагностики и взаимного сопоставления климатических моделей» (Program for Climate Model Diagnosis and Intercomparison, PCMDI), https://pcmdi.llnl.gov/index.html Сайт Национального научно-исследовательского вычислительного центра энергетических исследований Министерства энергетики США (NERSC), https://www.nersc.gov/ Раздел биологических и экологических исследований на сайте Министерства энергетики США, https://www.energy.gov/science/ber/biological-and-environmental-research и страница отделения климатологии и экологических наук (Climate and Environmental Sciences Division, CESD) на том же сайте http://science.energy.gov/ber/research/cesd/ Страница лаборатории вычислительных и информационных систем (Computational and Information Systems Laboratory, CISL) на сайте Университетской корпорации атмосферных исследований (University Corporation for Atmospheric Research, UCAR), https://www2.cisl.ucar.edu/

А.8.9 Вариант использования № 49: Подповерхностные биогеохимические исследования

Название	Фокус-область подповерхностных биогеохимических исследований Управления биологических и экологических исследований Министерства энергетики США (BER)
Предметная область	Научные исследования: науки о Земле
Автор/организация/эл.почта	Деб Агарвал (Deb Agarwal) / Национальная лаборатория имени Лоуренса в Беркли (LBNL), daagarwal@lbl.gov
Акторы/заинтересованные лица, их роли и ответственность	Проект Национальной лаборатория им. Лоуренса в Беркли «Природосберегающие системы — научная фокус-область 2.0» (Sustainable Systems Scientific Focus Area 2.0); специалисты в области подповерхностной биогеохимии, гидрологи, геофизики, эксперты по геномике, Объединенный институт генома (JGI) Министерства энергетики США; климатологи и «Программа подповерхностных биогеохимических исследований» (Subsurface Biogeochemical Research, SBR) Министерства энергетики США
Цели	Научный план проекта «Природосберегающие системы — научная фокус-область 2.0» (Sustainable Systems Scientific Focus Area 2.0 (SFA 2.0)) был разработан с целью совершенствования понимания и прогнозирования сложных многомасштабных наземных сред, имеющих отношение к миссии Министерства энергетики США, посредством целевого рассмотрения существующих пробелов в научном знании
Описание варианта использования	Выполнение проекта «Моделирование водораздела с использованием генома» (GEWaSC), который обеспечит прогнозирующую структуры для понимания того, как геномная информация, хранящаяся в подповерхностном микробиоме, влияет на функционирование биогеохимического водораздела; как процессы в масштабе водораздела влияют на функционирование микробов, и как эти взаимодействия совместно эволюционируют. Хотя имеющиеся средства моделирования, разработанные нашей группой и другими членами сообщества, позволяют воспроизводить процессы, происходящие во внушительном диапазоне масштабов (от отдельной

Описание варианта использования	бактериальной клетки до шлейфа загрязнения), до настоящего времени недостаточно усилий уделялось разработке концепции для систематического соединения явлений различных масштабов, что необходимо для выявления ключевых элементов контроля и управления и для моделирования существенных обратных связей. В рамках проекта GEWaSC основное внимание будет уделено разработке концепция моделирования, которая формально охватит масштабы от геномов до водоразделов	
Текущие решения	**Вычислительная система**	Обеспечивает Национальный научно-исследовательский вычислительный центр энергетических исследований Министерства энергетики США (NERSC)
	Хранилище данных	Обеспечивает центр NERSC
	Сеть связи	ESNet
	Программное обеспечение	PFLOWTran, Postgres, HDF5, Akuna, NEWT и др.
Характеристики больших данных	**Источник данных (распределенный/централизованный)**	Терабайтных объемов данные секвенирования (определения первичной структуры макромолекул) из Объединенного института генома (JGI), подповерхностные и поверхностные гидрологические и биогеохимические данные из различных датчиков (включая «плотные» геофизические наборы данных), экспериментальные данные полевых измерений и лабораторного анализа
	Объем (количество)	
	Скорость обработки (например, в реальном времени)	
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Данные охватывают все масштабы, от геномики микробов в почве до гидробиогеохимии водораздела. Для проекта SFA требуется объединение разнообразных и разрозненных наборов данных полевых, лабораторных измерений и моделирования, охватывая различные семантические, пространственные и временные масштабы посредством GEWaSC. Такие наборы данных будут производиться различными областями исследований и будут включать данные моделирования, данные полевых измерений (гидрологических, геохимических, геофизических), данные биологических наук — «омиков» и данные лабораторных экспериментов
	Вариативность (темпы изменения)	Моделирование и эксперименты

Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Каждый из источников данных изучает разные свойства, оказывающие различное влияние, ввиду чего данные крайне неоднородные. Каждый из источников данных характеризуется различными связанные с ним уровнями неопределенности и точности. Кроме того, перемещение данных по различным масштабам и областям вносит неопределенность, как и интеллектуальный анализ данных. Качество данных имеет решающее значение
	Визуализация	Визуализация крайне важна для понимания данных
	Качество данных (синтаксис)	Качество данных имеет решающее значение
	Типы данных	См. пункт «Разнообразие» выше
	Аналитика данных	Интеллектуальный анализ данных, оценка качества данных, взаимная корреляция между наборами данных, ускорение процесса разработки моделей, статистика, оценка качества, слияние данных и т. д.
Иные проблемы больших данных	Перемещение данных между разнообразными и большими наборами данных, которые охватывают различные предметные области и масштабы	
Проблемы пользовательского интерфейса и мобильного доступа	Сбор данных полевых экспериментов будет улучшен за счет доступа к уже имеющимся данным и автоматического ввода новых данных через мобильные устройства	
Технические проблемы обеспечения безопасности и защиты персональных данных		
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Широкий спектр проектов и программ в области наук о Земле работает над проблемами, которые охватывают те же области, что и настоящий проект	
Дополнительная информация (гиперссылки)		

А.8.10 Вариант использования № 50: Сети AmeriFlux и FLUXNET

Название	Сеть AmeriFlux управления биологических и экологических исследований Министерства энергетики США и сеть FLUXNET
Предметная область	Научные исследования: науки о Земле
Автор/организация/эл.почта	Деб Агарвал (Deb Agarwal) / Национальная лаборатория имени Лоуренса в Беркли, США (LBNL), daagarwal@lbl.gov
Акторы/заинтересованные лица, их роли и ответственность	Участники сети AmeriFlux, группа менеджмента данных (Data Management Team) проекта AmeriFlux, европейская «Интегрированная система наблюдения за выбросами углерода» ICOS (Integrated Carbon Observation System), программа «Исследования экосистемы суши» (Terrestrial Ecosystem Science, TES) Министерства энергетики США, Министерство сельского хозяйства США (U.S. Department of Agriculture, USDA), Национальный научный фонд США (National Science Foundation, NSF) и специалисты по моделированию климата

Цели	Измерения, выполняемые сетями AmeriFlux Network и FLUXNET предоставляют информацию о ключевых по важности взаимосвязях между организмами, экосистемами и исследованиями на уровне процессов, в адекватных для изучения климата масштабах ландшафтов, регионов и континентов, которые могут быть учтены в биогеохимических и климатических моделях. Поступающие от отдельных станций измерения газовых потоков данные являются основой для расширяющихся усилий по обобщающему анализу и анализу на основе моделей	
Описание варианта использования	Наблюдения сети AmeriFlux позволяют укрупненно изучать потоки малых газовых составляющих (следовых газов — CO2, водяной пар) в широком временном (часы, дни, времена года, годы и десятилетия) и пространственном диапазоне. Кроме того, наборы данных AmeriFlux и FLUXNET содержат информацию о важнейших взаимосвязях между организмами, экосистемами и исследованиями на уровне процессов в адекватных для изучения климата масштабах ландшафтов, регионов и континентов, которые следует учитывать в биогеохимических и климатических моделях	
Текущие решения	**Вычислительная система**	Обеспечивает Национальный научно-исследовательский вычислительный центр энергетических исследований Министерства энергетики США (NERSC)
	Хранилище данных	Обеспечивает центр NERSC
	Сеть связи	ESNet
	Программное обеспечение	EddyPro, специализированное аналитическое программное обеспечение, R, Python, нейронные сети, Matlab
Характеристики больших данных	**Источник данных (распределенный/централизованный)**	Около ~150 вышек для измерения газовых потоков в сети AmeriFlux, и более 500 вышек распределены по всему миру
	Объем (количество)	
	Скорость обработки (например, в реальном времени)	
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	Данные о газовых потоках относительно однородны, однако биологические данные, данные об атмосферных возмущениях и другие вспомогательные данные, необходимые для обработки и интерпретации основных данных, обширны и сильно варьируются. Объединение этих данных с данными о потоках является сложной задачей для современных систем
	Вариативность (темпы изменения)	
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Каждая измерительная станция использует уникальные методы измерения и обработки данных. Сеть объединяет эти данные и выполняет типовую обработку, заполнение пробелов и оценку качества. В сети тысячи пользователей
	Визуализация	Для визуализации данных используются графики и трехмерные поверхности

Наука о больших данных (сбор, курирование, анализ, операции)	**Качество данных (синтаксис)**	Сеть выполняет оценку качества данных
	Типы данных	Описаны в пункте «Разнообразие» выше
	Аналитика данных	Интеллектуальный анализ данных, оценка качества данных, взаимная корреляция между наборами данных, ассимиляция данных, интерполяция данных, статистика, оценка качества, слияние данных и т. д.
Иные проблемы больших данных	Перемещение данных между разнообразными и большими наборами данных, которые охватывают различные предметные области и масштабы	
Проблемы пользовательского интерфейса и мобильного доступа	Сбор данных полевых экспериментов будет улучшен за счет доступа к уже имеющимся данным и автоматического ввода новых данных через мобильные устройства	
Технические проблемы обеспечения безопасности и защиты персональных данных		
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)		
Дополнительная информация (гиперссылки)	Сайт сети AmeriFlux, https://ameriflux.lbl.gov/ Сайт портала данных Fluxdata, обслуживающего сообщество FLUXNET, https://fluxnet.fluxdata.org/	

А.9 Энергетика

А.9.1 Вариант использования № 51: Прогнозирование потребления электроэнергии в интеллектуальных энергосетях

Название	Прогнозирование потребления электроэнергии в интеллектуальных энергосетях
Предметная область	Информатизация энергетической отрасли
Автор/организация/эл.почта	Йогеш Симмхан (Yogesh Simmhan), Университет Южной Калифорнии, США, simmhan@usc.edu
Акторы/заинтересованные лица, их роли и ответственность	Электроэнергетические компании, микроэнергосети университетских кампусов, менеджеры зданий, потребители электроэнергии, рынки электроэнергии
Цели	Разработка масштабируемых и точных моделей прогнозирования, предсказывающих, с различной пространственной и временной детализацией, потребление электроэнергии (в киловатт-часах) в пределах зоны обслуживания электросети с тем, чтобы помочь повысить надежность и эффективность энергосистемы
Описание варианта использования	Развертывание умных счетчиков позволяет получать данные об энергопотреблении (в киловатт-часах) в почти реальном времени — показания поступают каждые 15 минут с точностью до отдельного потребителя в пределах зоны обслуживания интеллектуальных энергосетей. Подобный беспрецедентный и расширяющийся доступ к детальной информации о потреблении энергии позволяет разрабатывать новые аналитические инструменты для прогнозирования энергопотребления для индивидуальных потребителей, трансформаторов, подстанций и зоны обслуживания энергосети

Описание варианта использования	Краткосрочные прогнозы могут использоваться энергетическими компаниями и менеджерами микросетей для принятия предупредительных мер до того, как пики потребления приведут к падению напряжения/ отключению электроэнергии, — за счет оптимизации управления спросом путем взаимодействия с потребителями, подключения резервных генераторов или закупки мощностей на энергетических рынках. Эти меры обратной связи образуют цикл «наблюдение, ориентация, решение, действие» (observe–orient–decide–act, OODA). Клиенты также могут использовать такие прогнозы для планирования энергопотребления и составления своего бюджета. Среднесрочные и долгосрочные прогнозы могут помочь энергетическим компаниям и менеджерам зданий планировать генерирующие мощности, портфели возобновляемых источников энергии, контракты на закупку электроэнергии и меры по повышению энергоэффективности зданий. Этапы работы включают: 1) сбор и хранение данных: данные временных рядов от (потенциально) миллионов умных счетчиков в режиме, близком к реальному времени; сведения о потребителях, объектах и регионах; прогнозы погоды. Архивация данных для целей обучения, тестирования и валидации моделей; 2) очистку и нормализацию данных: пространственно-временная нормализация, заполнение пробелов / интерполяция, обнаружение выбросов, семантическое аннотирование; 3) обучение моделей прогнозирования: использование одномерных моделей временных рядов, таких как «Авторегрессионное интегрированное скользящее среднее» (autoregressive integrated moving average, ARIMA), и управляемых данными моделей машинного обучения, таких как дерево регрессии, искусственная нейронная сеть ANN (Artificial Neural Network), с различной пространственной (потребитель, трансформатор) и временной (15-минутной, 24-часовой) разрешающей способностью; 4) прогнозирование: Прогнозирование потребления в различных пространственно-временных разрезах при разных горизонтах прогнозирования, с использованием данных, поступающих в почти реальном времени, и исторических данных, которые подаются на вход прогнозной модели, вместе с ограничениями на время запаздывания прогноза	
Текущие решения	**Вычислительная система**	Многоядерные сервера, коммерческий кластер, рабочие станции
	Хранилище данных	СУБД SQL, CSV-файлы, HDFS, управлении показаниями счетчиков
	Сеть связи	Гигабитный Ethernet
	Программное обеспечение	R/Matlab, Weka, Hadoop
Характеристики больших данных	**Источник данных (распределенный/ централизованный)**	Данные умных счетчиков (распределенные), служебные базы данных энергетических компаний (информация о клиентах, топология сети — централизованные), данные всеобщей переписи населения США (распределенные), метеорологические данные Национального управления океанических и атмосферных исследований США (National Oceanic and Atmospheric Administration, NOAA) (распределенные), данные информационных систем для построения микроэнергосетей (централизованные) и сенсорных сетей микроэнергосетей (распределенные)
	Объем (количество)	10 гигабайт в день; 4 терабайта в год (масштабы города)

Характеристики больших данных	**Скорость обработки (например, в реальном времени)**	Лос-Анджелес: передача показаний раз в 15 минут (~100 тысяч потоков); либо раз в 8 часов (~1,4 миллиона потоков). Более детальные данные агрегируются в 8-часовые интервалы
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	На основе кортежей: временные ряды, строки баз данных. На основе графов: топология сети, подключение клиентов. Некоторые семантические данные используются для нормализации
	Вариативность (темпы изменения)	Показания счетчиков и данные о погоде изменяются и собираются/используются ежечасно. Информация о клиентах/зданиях/топологии сети изменяется медленно, на еженедельной основе
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	Необходимы управление версиями и обеспечение воспроизводимости для целей валидации / сопоставления прошлых и текущих моделей. Устойчивость систем хранения и инструментов аналитики важна для целей оперативного управления. Семантическая нормализация может помочь при выполнении междисциплинарного анализа (например, операторами энергосетей, менеджерами зданий, инженерами-энергетиками, специалистами по поведению)
	Визуализация	Визуализация в привязке к географическим координатам (построение картограмм) топологии энергосети, нагрузки; «тепловые карты» потребления энергии; графики прогнозируемого спроса в сопоставлении с располагаемой мощностью», анализ «что, если»; отображение информации в реальном времени; приложения с push-уведомлениями о предупреждениях
	Качество данных (синтаксис)	Пробелы в данных от умных счетчиков и в погодных данных; проблемы с качеством данных датчиков; проводятся тщательные проверки данных счетчиков, используемых для выставления счетов
	Типы данных	Временные ряды (CSV, кортежи SQL), статическая информация (RDF, XML), топология (файлы форм)
	Аналитика данных	Модели прогнозирования, модели машинного обучения, анализ временных рядов, кластеризация, выявление закономерностей, обработка сложных событий, визуальный анализ сети

Иные проблемы больших данных	Масштабируемая аналитика в реальном времени над большими потоками данных. Аналитика с низкой задержкой для оперативных нужд. Объединенная аналитика на уровне энергетической компании и микроэнергосетей. Надежная аналитика временных рядов по данным об энергопотреблении миллионов клиентов. Моделирование поведения клиентов, целевая выдача требований к крупным потребителям о временном сокращении энергопотребления
Проблемы пользовательского интерфейса и мобильного доступа	Приложения для взаимодействия с клиентами: сбор данных от клиентов/зданий для моделирования поведения, извлечение признаков; уведомление о требованиях сократить энергопотребления со стороны энергетической компании/менеджеров зданий; предложения по повышению энергоэффективности; отображение объема энергопотребления с географической привязкой
Технические проблемы обеспечения безопасности и защиты персональных данных	Персональные данные клиента требуют осторожного обращения. Данные о потреблении энергии клиентом могут раскрыть закономерности поведения. Анонимизация информации. Агрегирование данных, чтобы избежать идентификации клиентов. Ограничения на распространение данных, установленные федеральными регулирующими органами в области энергетики и регулирующими органами штатов. Обследования, проведенные специалистами по изучению поведения, могут подпадать под ограничения, установленные органами внутреннего контроля этичности научных исследований (в США это «Институциональные наблюдательные советы» — Institutional Review Board, IRB)
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Управляемая данных аналитика в реальном времени для киберфизических систем
Дополнительная информация (гиперссылки)	Страница проекта «Интеллектуальная энергосеть» (Smart Grid) на сайте Университета Южной Калифорнии, -https://sites.usc.edu/dslab/projects/smart-grid/ Статья «Умные сети электроснабжения» в Википедии, https://ru.wikipedia.org/wiki/Умные_сети_электроснабжения Страница интеллектуальной энергосети Лос-Анджелеса, https://www.ladwp.com/ladwp/faces/ladwp/aboutus/a-power/a-p-smartgridla Йогеш Симмхан (Yogesh Simmhan) и др. «Облачная программная платформа для аналитики больших данных в интеллектуальных энергосетях» (Cloud-Based Software Platform for Big Data Analytics in Smart Grids), Computing in Science & Engineering, том 15, вып.4, июль-август 2013, https://www.researchgate.net/publication/260585847_Cloud-Based_Software_Platform_for_Big_Data_Analytics_in_Smart_Grids

А.9.2 Вариант использования № 52: Система управления энергией домашнего хозяйства HEMS

Название	Система управления энергией домашнего хозяйства HEMS
Предметная область	Деловая деятельность — оказание услуг
Автор/организация/эл.почта	Японский национальный орган по стандартизации
Акторы/заинтересованные лица, их роли и ответственность	Пользователи являются обычные люди, проживающие в частных домах. «Менеджер энергопотребления» (Energy manager) — это компания, устанавливающая в частные дома ряд датчиков и устройств. Роль «менеджера энергопотребления» может выполнять жилищно-строительная фирма или энергетическая компания. «Информационный менеджер» — это компания, собирающая данные из частных домов и отвечающая за неприкосновенность частной жизни и безопасность пользователей. «Сервисный агент» — это компания, которая анализирует собранные данные и предоставляет пользователям ценную информацию в качестве услуги

Цели	Предоставление пользователям полезных информационных услуг, посредством комбинирования данных об энергопотреблении с другими доступными данными и их анализа	
Описание варианта использования	Система управления энергией домашнего хозяйства (Home Energy Management System, HEMS является полезной для энергосбережения в частных домах. В рамках системы HEMS в частных домах устанавливаются различного вида датчики и устройства, такие как «умный» счетчик, электромобиль, панель солнечных батарей, осветительные приборы, кондиционер, топливный элемент, водонагреватель, аккумуляторная батарея. «Менеджер энергопотребления» собирает произведенные в частных домах данные и сохраняет их в облачной базе данных, называемой «большой информационной платформой HEMS». «Информационный менеджер» управляет большой информационной платформой HEMS и осуществляет менеджмент данных. Он также отвечает за обеспечение неприкосновенности частной жизни и безопасность пользователей. «Сервисный агент» анализирует данные и предоставляет пользователям ценную информацию в качестве услуги	
Текущие решения	**Вычислительная система**	—
	Хранилище данных	—
	Сеть связи	—
	Программное обеспечение	—
Характеристики больших данных	**Источник данных (распределенный/централизованный)**	Источники данных распределены по отдельным частным домам
	Объем (количество)	Около 14 тысяч домохозяйств. Об объемах данных сведений нет
	Скорость обработки (например, в реальном времени)	В режиме реального времени, потоковая передача данных с датчиков
	Разнообразие (множество наборов данных, комбинация данных из различных источников)	«Умный» счетчик, электромобиль, панель солнечных батарей, осветительные приборы, кондиционер, топливный элемент, водонагреватель, аккумуляторная батарея
	Вариативность (темпы изменения)	Данные могут изменяться в течение секунд, минут или часов
Наука о больших данных (сбор, курирование, анализ, операции)	**Достоверность (вопросы надежности, семантика)**	
	Визуализация	Визуализация показаний датчиков является базовой услугой
	Качество данных (синтаксис)	Качество данных напрямую влияет на качество услуг
	Типы данных	Временные ряды
	Аналитика данных	Прогнозирование, анализ временных рядов
Иные проблемы больших данных		
Проблемы пользовательского интерфейса и мобильного доступа		

Технические проблемы обеспечения безопасности и защиты персональных данных	Персональные данные должны тщательно обрабатываться для обеспечения защиты неприкосновенности частной жизни пользователя. «Сервисный агент» обязан проинформировать пользователей о том, как будут обрабатываться данные и какие услуги могут быть предоставлены. Пользователи должны выбрать услуги, которые они хотят получить, принимая во внимание сведения, сообщенные «сервисным агентом»
Перечислите основные характеристики и связанные варианты использования (например, в интересах эталонной архитектуры)	Несколько игроков участвуют в цепочке поставок для потока больших данных. Игрок А получает данные из источника данных (от поставщика данных). Игрок А предоставляет данные другому игроку В. Оба игрока, А и В, анализируют данные. Таким образом, данные могут поступать, проходя цепочку из нескольких игроков
Дополнительная информация (гиперссылки)	«Участие Министерства экономики, торговли и промышленности Японии в крупномасштабном проекте по развитию информационной инфраструктуры HEMS», пресс релиз Японской телеграфно-телефонной компании NTT–Восток на сайте компании, 28 августа 2014 года, https://www.ntt-east.co.jp/release/detail/20140828_01.html (на японском языке)

Приложение В
(справочное)

Сводка ключевых характеристик

Из описания каждого варианта использования были извлечены сведения, связанные с пятью ключевыми характеристиками, в число которых вошли три характеристики больших данных (объем, скорость обработки и разнообразие), сведения о программном обеспечении и соответствующей аналитике. Данные сведения представлены в таблице В.1.

Таблица В.1 — Специфическая для варианта использования информация о ключевых характеристиках

Вариант использования	Объем	Скорость обработки	Разнообразие	Программное обеспечение	Аналитика
А.1.1 Вариант использования № 1: Архивное хранение больших данных переписей населения в США 2010 и 2000 годов	380 терабайт	Данные статичны в течение 75 лет	Отсканированные документы	Надежное архивное хранение	Только по истечении 75 лет
А.1.2 Вариант использования № 2: Прием Национальными архивами США государственных данных на архивное хранение, поиск, извлечение и обеспечение долговременной сохранности	Сотни терабайт, постоянно увеличивается	Скорость поступления данных относительно низкая по сравнению с другими вариантами использования, однако случаются всплески, т. е. данные могут поступать партиями размером от гигабайта до сотен терабайт	Неструктурированные и структурированные данные: текстовые документы, электронная почта, фотографии, отсканированные документы, мультимедийные материалы, материалы из социальных сетей, веб-сайты, базы данных и т. д.	Кастомизированное ПО, коммерческие поисковые продукты, коммерческие базы данных	Сканирование/индексирование; поиск; ранжирование; прогностический поиск; категоризация данных (чувствительные, конфиденциальные и т. д.). Выявление и маркировка персональных данных (Personally Identifiable Information, PII).
А.1.3 Вариант использования № 3: Повышение активности респондентов в статистических обследованиях	Примерно один петабайт	Варьируется, данные с мест о ходе проведения обследования передаются непрерывно в потоковом режиме. Во время последней переписи были переданы 150 миллионов документов	Данные обычно представляют собой заданные текстовые и числовые поля	Hadoop, Spark, Hive, R, SAS, Mahout, Allegrograph, MySQL, Oracle, Storm, BigMemory, Cassandra, Pig	Рекомендательные системы, постоянный мониторинг

Продолжение таблицы В.1

Вариант использования	Объем	Скорость обработки	Разнообразие	Программное обеспечение	Аналитика
А.1.4 Вариант использования № 4: Использование нетрадиционных данных для повышения активности респондентов в статистических обследованиях	Будет определен в будущем	Будет определена в будущем	Данные обследований, другие государственные административные данные, геопространственные данные из различных источников	Hadoop, Spark, Hive, R, SAS, Mahout, Allegrograph, MySQL, Oracle, Storm, BigMemory, Cassandra, Pig	Новая аналитика необходима для получения надежных оценок на основе нетрадиционных разнородных источников
А.2.1 Вариант использования № 5: Облачные вычисления в секторах финансовой отрасли	От нескольких терабайт до нескольких петабайт	В реальном времени	Различные виртуальные среды, работающие в рамках архитектуры пакетной обработки или параллельной архитектуры с «горячей» заменой	Hadoop, RDBMS, XBRL	Выявление мошенничества
А.2.2 Вариант использования № 6: Международная исследовательская сеть Mendeley	В настоящее время 15 терабайт с темпом прироста около 1 терабайта в месяц	В настоящее время пакетные задания Hadoop планируются раз в день, но началась работа над рекомендациями по выполнению работ в реальном времени	PDF-документы, лог-файлы социальной сети и активности клиентов	Hadoop, Scribe, Hive, Mahout, Python	Стандартные библиотеки для проведения машинного обучения и аналитики, выполнения латентного размещения Дирихле (LDA), а также специально разработанные инструменты составления отчетности и визуализации данных для агрегирования сведений о читательской и социальной активности, связанной с каждым документом

Продолжение таблицы В.1

Вариант использования	Объем	Скорость обработки	Разнообразие	Программное обеспечение	Аналитика
А.2.3 Вариант использования № 7: Сервис кинофильмов Netflix	По состоянию на лето 2012 г.: 25 миллионов подписчиков; 4 млн оценок в день; 3 млн поисковых запросов в день; 1 млрд часов потокового видео в июне 2012 г. Объем облачного хранения 2 петабайта (июнь 2013 г.)	Контент (видео и характеристики) и рейтинги постоянно обновляются	Данные варьируются от цифровых мультимедийных материалов до пользовательских рейтингов, профилей пользователей и параметров мультимедиа, используемых для основанных на контенте рекомендаций	Hadoop и Pig, Cassandra, Teradata	Персонализированные рекомендательные системы, использующие логистическую / линейную регрессию, эластичные сети, факторизацию матриц, кластеризацию, латентное размещение Дирихле (LDA), ассоциативные правила, градиентный бустинг деревьев решений и другие инструменты; доставка потокового видео
А.2.4 Вариант использования № 8: Веб-поиск	В общей сложности около 45 млрд веб-страниц; ежедневно загружается 500 млн фотографий; и ежеминутно на YouTube закачивается 100 часов видеоматериалов	Данные обновляются и ответы на запросы выдаются в реальном времени	Различные мультимедийные форматы	Map/Reduce + Bigtable; Dryad + Cosmos. PageRank. Последний этап по сути представляет собой рекомендательную систему	Веб-сканирование, поиск (в том числе по тематике), ранжирование, рекомендации
А.2.5 Вариант использования № 9: Обеспечение непрерывности деловой деятельности и восстановления после катастроф для больших данных в облачной экосистеме	От нескольких терабайт до нескольких петабайт	Возможна обработка в реальном времени для последних изменений	Решение должно работать с любыми данными	Hadoop, Map/Reduce, Open-source и/или проприетарные решения поставщиков, таких как AWS (Amazon Web Services), Google Cloud Services и Microsoft	Надежное резервное копирование
А.2.6 Вариант использования № 10: Грузоперевозки	Большой	Должна заработать в реальном времени; в настоящее время обновления идут по наступлении событий	По наступлении событий	Неизвестно	Распределенный анализ событий с целью выявления проблем

Продолжение таблицы В.1

Вариант использования	Объем	Скорость обработки	Разнообразие	Программное обеспечение	Аналитика
А.2.7 Вариант использования № 11: Данные о материалах	Более 500 тысяч видов материалов в 1980-х годах, значительный рост с того времени	Постоянное нарастание количества новых материалов	Много наборов данных при практическом отсутствии стандартов	Национальные программы (Япония, Южная Корея и Китай), прикладные программы (ядерная программа Евросоюза); проприетарные решения (Granta, и др.)	Широко применяемой аналитики нет
А.2.8 Вариант использования № 12: Геномика материалов на основе результатов моделирования	100 терабайт (текущий), 500 терабайт через 5 лет. Требуются масштабируемые базы данных для данных типа «ключ — значение» и для библиотек объектов	Регулярное добавление результатов моделирования	Разнообразные данные и результаты моделирования	MongoDB, GPFS, PyMatGen, FireWorks, VASP, ABINIT, NWChem, BerkeleyGW; различное ПО, разработанное сообществом	Технологии Map/Reduce и поиска, позволяющие комбинировать данные моделирования и экспериментальные данные
А.3.1 Вариант использования № 13: Облачный крупномасштабный анализ и визуализация геопространственных данных	Растровая графика — сотни терабайт; векторные данные — десятки гигабайт, но при этом миллиарды точек	Некоторые датчики передают векторные данные в масштабе времени, близком к реальному	Растровые изображения, векторная графика (различные форматы: формат Shapefile, язык разметки Keyhole (Keyhole Markup Language, KML) и текстовые потоки), различные структуры из объектов	Реляционная СУБД с геопространственной поддержкой; ESRI ArcServer, Geoserver	Ближайшая точка подхода, отклонение от маршрута, плотность точек во времени, метод главных компонентов (principal component analysis, PCA) и метод анализа независимых компонентов (independent component analysis, ICA)
А.3.2 Вариант использования № 14: Идентификация и отслеживание объектов — Постоянное наблюдение	FMV — от 30 до 60 кадров в секунду при полноцветном разрешении 1080 пикселей; WALF — от 1 до 10 кадров в секунду при полноцветном разрешении 10 тысяч * 10 тысяч пикселей	В реальном времени	Данные обычно представлены в одном или нескольких стандартных форматах для графических изображений или видео	Широкий спектр специализированного программного обеспечения и инструментов, включая, в том числе, традиционные реляционные СУБД и средства отображения	Визуализация путем наложения на отображение геопространственных данных; базовая аналитика для выявления объектов и интеграция с развитыми средствами оценки ситуации на основе объединения данных

Продолжение таблицы В.1

Вариант использования	Объем	Скорость обработки	Разнообразие	Программное обеспечение	Аналитика
А.3.3 Вариант использования № 15: Обработка и анализ разведывательных данных	От десятков терабайт до сотен петабайт в случае периферийных и стационарных кластеров. У пехотинцев, как правило, имеется от одного до сотен гигабайт данных (обычно на портативном/карманном устройстве с объемом памяти менее 10 гигабайт)	Многие устройства сбора фото/видеоданных собирают петабайт данных за несколько часов	Текстовые файлы, первичные данные с датчиков (raw media), графические образы, видео, аудио, электронные данные и данные, созданные человеком	Hadoop, Accumulo (BigTable), Solr, NLP, Puppet (управление жизненным циклом ИТ, обеспечение безопасности) и Storm; ГИС	Оповещения в масштабе времени, близком к реальному, основанные на закономерностях и изменениях основных параметров; анализ взаимосвязей; Геопространственный анализ; Аналитика текстов (определение настроений, выделение сущностей и т. д.)
А.4.1 Вариант использования № 16: Данные электронной медицинской документации	Свыше 12 млн пациентов, более 4 млрд отдельных клинических наблюдений, более 20 терабайт первичных данных	Ежедневно добавляется от 500 тыс. до 1,5 млн новых клинических транзакций в режиме реального времени	Широкий спектр данных, поступающих от врачей, медсестер, лабораторий и измерительных инструментов	Teradata, PostgreSQL, MongoDB, Hadoop, Hive, R	Методы извлечения информации (статистическая мера TF–IDF, латентно-семантический анализ и статистическая функция «взаимная информация»). Методы обработки естественного языка (NLP), метода оценки максимального правдоподобия и Байесовских сетей
А.4.2 Вариант использования № 17: Анализ графических образов в патологии/Цифровая патология	1 гигабайт первичных данных + 1,5 гигабайта аналитических результатов на двумерное изображение; 1 терабайт первичных данных + 1 терабайт аналитических результатов на трехмерное изображение. 1 петабайт данных в год в средней больнице	После создания данные не подвергаются изменениям	Характеристики изображений и виды аналитики зависят от типа заболевания	MPI для анализа изображений; Map/Reduce + Hive с пространственным расширением	Анализ изображений, пространственные запросы и аналитика, кластеризация и классификация признаков

Продолжение таблицы В.1

Вариант использования	Объем	Скорость обработки	Разнообразие	Программное обеспечение	Аналитика
А.4.3 Вариант использования № 18: Вычислительный анализ биоизображений	Объем данных в результате одного сканирования на появляющихся установках составляет 32 терабайта, а годовой объем медицинских диагностических изображений — около 70 петабайт	Объемы собираемых данных требуют использования высокопроизводительных вычислений	Мультимодальный сбор и анализ изображений (multimodal imaging), поступающих по разрозненным каналам данных	Масштабируемые базы данных для данных типа «ключ — значение» и для библиотек объектов. ImageJ, OMERO, из новых продуктов — VolRover, продвинутые методы сегментации и выявления признаков	Машинное обучение (метод опорных векторов (Support Vector Machine, SVM) и алгоритм «случайный лес» (random forest, RF) для сервисов классификации и рекомендательных сервисов
А.4.4 Вариант использования № 19: Геномные измерения	В течение года — двух в NIST потребуется >100 терабайт. Сообществу здравоохранения в целом потребуется много петабайт для хранения данных	Секвенсоры ДНК способны генерировать порядка ~300 гигабайт сжатых данных в день	Файловые форматы недостаточно хорошо стандартизированы, хотя некоторые стандарты существуют. Как правило, структурированные данные	Программное обеспечение с открытым исходным кодом для секвенирования в биоинформатике, разработанное академическими группами	Обработка первичных данных с целью выделения вариаций (variant calls) , а также клиническая интерпретация вариаций
А.4.5 Вариант использования № 20: Сравнительный анализ (мета) геномов	50 терабайт	Новые секвенсоры выдают потоки данных, скорость которых растет	Биологические данные по своей природе неоднородны, сложны, структурны и иерархичны. Помимо базовых геномных данных, новые типы данных таких направлений биологической науки — «омиков» (omics), как транскриптомика, метиломика (methylomics) и протеомика	Стандартные инструменты биоинформатики (BLAST, HMMER, инструменты множественного выравнивания последовательностей и филогенетики, программы поиска/предсказания генов и генных структур (gene callers), программы предсказания свойств по результатам секвенирования (sequence feature predictors) и т. д.), скрипты Perl / Python	Описательная статистика, статистическая значимость при проверке гипотез, кластеризация и классификация

Продолжение таблицы В.1

Вариант использования	Объем	Скорость обработки	Разнообразие	Программное обеспечение	Аналитика
А.4.6 Вариант использования № 21: Индивидуальное управление лечением диабета	5 млн пациентов	Не в режиме реального времени, но данные периодически обновляются	Типичным для пациента является около 100 значений свойств из контролируемых словарей и 1000 непрерывных числовых величин. Большинство значений привязаны ко времени	Внутреннее хранилище данных в Клинике Мейо, США (EDT), дополняемой HDFS	Интеграция данных в семантический граф, использование обхода графа взамен операции join в SQL. Разработка алгоритмов интеллектуального анализа семантических графов с целью выявления закономерностей в графе, индексирования графа и поиска по нему. СУБД Hbase с индексированием. Специализированная программа для выявления новых свойств пациента на основе хранящихся данных
А.4.7 Вариант использования № 22: Статистический реляционный искусственный интеллект для здравоохранения	Сотни гигабайт для одной когорты из нескольких сотен человек. Когда речь идет о миллионах пациентов, объем данных может быть порядка 1 петабайта	Электронные медицинские документы постоянно обновляются. В других контролируемых исследованиях данные часто поступают партиями с регулярными интервалами	Ключевая особенность — данные обычно содержатся в ряде таблиц, которые необходимо объединить для выполнения анализа	В основном на основе Java, для обработки данных используются инструменты собственной разработки	Реляционные вероятностные модели (статистический реляционный искусственный интеллект), обучающиеся на данных различных типов
А.4.8 Вариант использования № 23: Эпидемиологическое исследование в масштабе всего населения Земли	100 терабайт	Подача данных в программу моделирования мала, однако в ходе моделирования создаются огромные объемы данных	Возможно большое разнообразие, если принять во внимание различные аспекты мировой популяции, географические, социально-экономические и культурные различия	Charm++, MPI	Моделирование на основе синтетической глобальной популяции

Продолжение таблицы В.1

Вариант использования	Объем	Скорость обработки	Разнообразие	Программное обеспечение	Аналитика
А.4.9 Вариант использования № 24: Моделирование распространения социального влияния	Десятки терабайт новых данных ежегодно	Во время социальных волнений взаимодействие между людьми и мобильность являются ключом к пониманию динамики системы. Быстрые изменения в данных: например, о том, кто на кого подписан в Твиттере	Серьезные проблемы: объединение данных (data fusion), комбинирование данные из разных источников, проблема отсутствующих или неполных данных	Специализированные программы моделирования, программное обеспечение с открытым исходным кодом и проприетарные среды моделирования. Базы данных	Модели поведения людей и физических инфраструктур, а также взаимодействия между ними. Визуализация результатов
А.4.10 Вариант использования № 25 Биологическое разнообразие и европейская электронная научно-исследовательская инфраструктура LifeWatch	Суммарный объем данных предстоит определить	Обработка и анализ в реальном времени в случае стихийных бедствий или техногенных катастроф	Большое разнообразие и количество задействованных баз данных и данных наблюдений	Веб-сервисы, грид-сервисы, реляционные базы данных	Требуются развитая и богатая визуализация
А.5.1 Вариант использования № 26: Крупномасштабное глубокое обучение	Типичный объем наборов данных обычно составляет от 1 до 10 терабайт. Для обучения беспилотного автомобиля могут потребоваться 100 миллионов изображений	Требуется намного более быстрая обработка, чем в реальном времени. Для управления беспилотным автомобилем необходимо обрабатывать многие тысячи изображений с высоким разрешением (6 мегапикселей и более) в секунду	Нейронная сеть очень неоднородна, поскольку она изучает множество различных признаков	Программное обеспечение для информационного обмена между ядрами графических профессоров и для взаимодействия на основе MPI, разработанное на факультете вычислительных наук Стэндфордского университета. Исходный код на языках C ++ / Python	В небольшой степени выполняется пакетная статистическая предварительная обработка; весь остальной анализ данных выполняется самим алгоритмом обучения
А.5.2 Вариант использования № 27: Организация крупномасштабных, неструктурированных коллекций сделанных потребителями фотографий	Более 500 млрд фотографий на Facebook, более 5 млн фотографий на Flickr	Ежедневно в Facebook добавляется более 100 миллионов новых фотографий	Изображения и метаданные, включая теги EXIF (фокусное расстояние, тип камеры и т. д.)	Hadoop Map/Reduce, написанные вручную простые многопоточные инструменты (ssh и сокеты для обмена информацией)	Надежное решение задачи оптимизации с использованием нелинейного метода наименьших квадратов, метод опорных векторов SVM

Продолжение таблицы В.1

Вариант использования	Объем	Скорость обработки	Разнообразие	Программное обеспечение	Аналитика
А.5.3 Вариант использования № 28: Truthy — Анализ данных Твиттера	≈30 терабайт в год сжатых данных	Хранение, выполнение запросов и анализ в масштабе времени, близком к реальному	Непрерывный поток данных в реальном времени, поступающий из каждого источника	Hadoop / HBase с индексированием и HDFS; Hadoop, Hive, Redis для управления данными; Python/SciPy/NumPy/MPI для анализа данных	Выявление аномалий, кластеризация потока, классификация сигналов и онлайн-обучение; анализ процесса распространения информации и кластеризации, динамическая визуализация сети
А.5.4 Вариант использования № 29: Краудсорсинг в гуманитарных науках	От нескольких гигабайт (текст, опросы, экспериментальные значения) до сотен терабайт (мультимедиа)	Данные постоянно обновляются и анализируются инкрементально	До настоящего времени — в основном однородные небольшие наборы данных; ожидаются большие распределенные неоднородные наборы данных	Язык XML, традиционные реляционные базы данных	Все виды распознавания закономерностей (например, распознавание речи, автоматический анализ аудиовизуальных материалов, культурные закономерности); выявление структур (лексические единицы, лингвистические правила и т. д.)
А.5.5 Вариант использования № 30: Цифровая инфраструктура для исследований и анализа сетей и графов» (CINET)	Может составлять сотни гигабайт для одной сети; 1000—5000 сетей и методов	Сети очень динамичны; быстрое расширение коллекции сетей	Многочисленные типы сетей	Библиотеки для работы с графами: Galib, NetworkX. Управление распределенными потоками рабочих процессов: Simfrastructure, Базы данных, семантические веб-инструменты	Визуализация сетей

Продолжение таблицы В.1

Вариант использования	Объем	Скорость обработки	Разнообразие	Программное обеспечение	Аналитика
А.5.6 Вариант использования № 31: Измерения и оценки эффективности аналитических технологий в Национальном институте стандартов и технологий (NIST)	Более 900 млн веб-страниц общим объемом 30 терабайт, 100 млн твитов, 100 млн проверенных биометрических изображений, несколько сотен тысяч частично проверенных видеоклипов и терабайты более мелких полностью проверенных тестовых коллекций	Большинство старых методов оценки было основано на ретроспективной аналитике. В новых методах оценки основное внимание уделяется моделированию проблем анализа в реальном времени на основании данных из нескольких потоков	Широкий спектр типов данных, включая текстовый поиск/извлечение, машинный перевод, распознавание речи, биометрию изображений и голоса, распознавание и отслеживание объектов и людей, анализ документов, диалог между человеком и компьютером и поиск/извлечение мультимедиа	PERL, Python, C/C++, Matlab, R. Разработка по принципу «снизу вверх» тестовых и измерительных приложений	Извлечение информации, фильтрация, поиск и резюмирование; биометрия изображения и голоса; распознавание и понимание речи; машинный перевод; обнаружение и отслеживание людей и объектов в видеозаписях; детектирование событий; сопоставление изображений и документов; обнаружение новизны в данных; разнообразная структурная/семантическая/временная аналитика
А.6.1 Вариант использования № 32: Консорциум федеративных сетей данных (DFC)	Петабайты данных, сотни миллионов файлов	Обработка в реальном времени и пакетная	Большое	Интегрированная система управления данными, основанная на использовании правил (iRODS)	Поддержка общих рабочих процессов анализа
А.6.2 Вариант использования № 33: Discinnet-процесс	Не имеет значения: это база метаданных, а не больших данных	В реальном времени	Способность работать с произвольными большими данными	Symfony PHP, Linux, MySQL	—
А.6.3 Вариант использования № 34: Поиск по графу для научных данных	Несколько терабайт	Со временем эволюционирует	Большое	СУБД	Обработка графов данных
А.6.4 Вариант использования № 35: Анализ больших объемов данных, получаемых в экспериментах на синхротроне	От 50 до 400 гигабайт в день, в общей сложности ~400 терабайт	Непрерывный поток данных, однако анализ не обязательно проводить в реальном времени	Изображения	Octopus для томографической реконструкции, Avizo (http://vsg3d.com/) и FIJI (дистрибутив открытого программного обеспечения ImageJ)	Объемная реконструкция, идентификация характеристик и т. д.

Продолжение таблицы В.1

Вариант использования	Объем	Скорость обработки	Разнообразие	Программное обеспечение	Аналитика
А.7.1 Вариант использования № 36: Каталинский цифровой обзор неба в поисках транзиентов	Увеличение 0,1 терабайта за ночь, суммарный объем в настоящее время около 100 терабайт. Доступ к петабайтам базовых астрономических данных. Новый телескоп LSST будет собирать 30 терабайт данных за ночь	Обновление каждую ночь, процессы выполняются в реальном времени	Изображения, спектры, временные ряды для отдельных объектов (кривые блеска), каталоги	Специализированные «конвейер» обработки данных и программное обеспечение для анализа данных	Детектирование редких событий и установление связей с разнообразными существующими данными
А.7.2 Вариант использования № 37: Космологический обзор неба и моделирование	Несколько петабайт данных обзоров DES и ZTF; данные моделирования — более 10 петабайт	Анализ выполняется в пакетном режиме; данные наблюдений и моделирования пополняются ежедневно	Изображения и данные моделирования	MPI, FFTW, пакеты визуализации, Numpy, Boost, OpenMP, ScaLAPCK, СУБД PSQL и MySQL, Eigen, Cfitsio, http://astrometry.net/ и Minuit2	Требуются новые средства аналитики для анализа результатов моделирования
А.7.3 Вариант использования № 38: Большие данные космологических обзоров неба	Петабайты данных обзора «Темная энергия» (DES)	За ночь 400 изображений объемом 1 гигабайт каждое	Изображения	Linux-кластер, сервер реляционной СУБД Oracle, Postgres PSQL, большие машины памяти, стандартные интерактивные хосты Linux, GPFS; для моделирования — ресурсы высокопроизводительных вычислений; стандартное астрофизическое программное обеспечение для редуцирования данных, а также сценарии — обертки Perl / Python	Поиск новых оптических транзиентов при помощи алгоритмов машинного обучения, разложения Холецкого для тысяч моделирований с матрицами порядка миллиона по каждой стороне; и параллельное хранение изображений

Продолжение таблицы В.1

Вариант использования	Объем	Скорость обработки	Разнообразие	Программное обеспечение	Аналитика
А.7.4 Вариант использования № 39: Анализ данных Большого адронного коллайдера	15 петабайт в год данных от детекторов и результатов анализа	Данные поступают непрерывно, проходя при этом сложную процедуру отбора в реальном времени и тестовый анализ; однако полноценный анализ всех данных выполняется в автономном режиме	На каждой стадии анализа используются свои форматы, однако в рамках каждой стадии данные однородны	Грид-среда, содержащая 350 тысяч одновременно работающих ядер	Используются сложные специализированные программы анализа данных, а после этого — базовые статистические инструменты «предварительной разведки» (гистограммы); сложные поправки для устранения систематических погрешностей
А.7.5 Вариант использования № 40: Эксперимент Belle II	В конечном итоге объем данных наблюдений и моделирования по методу Монте-Карло составит около 120 петабайт	Данные поступают непрерывно, проходя при этом сложную процедуру отбора в реальном времени и тестовый анализ; однако полноценный анализ всех данных выполняется в автономном режиме	На каждой стадии анализа используются свои форматы, однако в рамках каждой стадии данные однородны	Программное обеспечение грида DIRAC	Используются сложные специализированные программы анализа данных, а после этого — базовые статистические инструменты «предварительной разведки» (гистограммы); сложные поправки для устранения систематических погрешностей
А.8.1 Вариант использования № 41: Радарная система некогерентного рассеяния EISCAT-3D	В настоящее время — несколько терабайт в год; 40 петабайт в год начиная с ~2022 года	Данные поступают непрерывно. Проводится тестовый анализ в реальном времени и полный анализ в пакетном режиме	Однородные большие данные	Специализированное программное обеспечение для анализа, на основе простого одноуровневого хранения файлов	Распознавание образов, требовательные процедуры корреляции, извлечение высокоуровневых параметров
А.8.2 Вариант использования № 42: Совместная деятельность европейских сетевых инфраструктур в области экологических исследований (ENVRI)	Объемы данных невелики (за исключением описанной выше системы EISCAT-3D), одна EPOS-система производит ~15 терабайт в год	В основном обработка потоков данных в реальном времени	6 отдельных проектов с единой архитектурой инфраструктуры. Данные очень разные от проекта к проекту	Для визуализации R и Python (Matplotlib); для обработки — специальное программное обеспечение	Ассимиляция данных, (статистический) анализ, интеллектуальный анализ данных, извлечение данных, построение научных моделей и моделирование, управление потоками научных рабочих процессов

Продолжение таблицы В.1

Вариант использования	Объем	Скорость обработки	Разнообразие	Программное обеспечение	Аналитика
А.8.3 Вариант использования № 43: Анализ радиолокационных данных для Центра дистанционного зондирования ледяного покрова CReSIS	Около 1 петабайта в настоящее время; рост на 50–100 терабайт за экспедицию. В будущем в ходе каждой экспедиции будет создаваться ~1 петабайт данных	Данные собираются в ходе двухмесячных экспедиций (включая результаты тест — анализов) и впоследствии обрабатываются в пакетном режиме	Первичные данные; изображения. Результаты последнего этапа обработки используются в научных исследованиях	Пакет Matlab для специализированной обработки первичных данных; специализированное ПО для обработки изображений; геоинформационная система как пользовательский интерфейс	Специализированная обработка сигналов для получения радиолокационных изображений, которые затем анализируются с помощью средств обработки изображений с целью выделения слоев
А.8.4 Вариант использования № 44: Обработка данных проекта UAVSAR	110 терабайт первичных данных и ~40 терабайт обработанных; плюс наборы данных меньшего размера	Данные поступают от инструмента, установленного на самолете, и добавляются порциями. Время от времени проводится повторная обработка ввиду появления новых методов или изменения параметров.	Изображения и файлы аннотаций	ROI_PAC, GeoServer, GDAL, а также инструменты, поддерживающие стандарт метаданных GeoTIFF; переход в облако	Проводится обработка первичных данных для получения изображений, которые пропускаются через инструменты обработки изображений; доступ через ГИС
А.8.5 Вариант использования № 45: Объединенный испытательный стенд iRODS центра НАСА в Лэнгли и Центра космических полетов имени Годдарда	Основная часть данных содержится в коллекции MERRA (описана ниже); остальные коллекции данных меньшего размера	Периодические обновления раз в полгода	Для многих приложений необходимо объединять данные реанализа MERRA с другими данными повторного анализа и с данными наблюдений, таких как CERES	SGE Univa Grid Engine версии 8.1, iRODS версии 3.2 и/или 3.3, файловая система IBM General Parallel File System (GPFS) версии 3.4, Cloudera версии 4.5.2–1	Программное обеспечение для объединения данных
А.8.6 Вариант использования № 46: Аналитические сервисы MERRA	480 терабайт в коллекции MERRA	Рост объема данных ~1 терабайт в месяц	Для многих приложений необходимо объединять данные реанализа MERRA с другими данными повторного анализа и с данными наблюдений	Cloudera, iRODS, Amazon AWS	Аналитика климата как сервис (CAaaS)

Продолжение таблицы В.1

Вариант использования	Объем	Скорость обработки	Разнообразие	Программное обеспечение	Аналитика
А.8.7 Вариант использования № 47: Атмосферная турбулентность — Обнаружение событий	Текущий объем 200 терабайт, через 5 лет — 500 терабайт	Данные анализируются по частям	Наборы данных ретроспективного анализа несогласованны по формату, разрешению, семантике и метаданным. Интерпретация/анализ каждого из входных потоков для включения в общий продукт	Инструмент Map/Reduce или аналогичный; SciDB или другая научная СУБД	Интеллектуальный анализ данных, ориентированный на поиск событий конкретных типов
А.8.8 Вариант использования № 48: Исследования климата с использованием модели климатической системы Земли (CESM)	До 30 петабайт в год (при условии проведения 15 сквозных моделирований в NERSC); еще больше в прочих центрах высокопроизводительных вычислений	В ходе моделирований производится 42 гигабайта/с	Существенные различия между данными моделирования различных групп и между данными наблюдений и результатами моделирования	Разработанные центром NCAR библиотека параллельного ввода-вывода и утилиты «NCAR Командный язык» (NCAR Command Language, NCL) и «NetCDF-операторы» (NetCDF Operators, NCO); параллельные библиотеки NetCDF	Необходима возможность анализа рядом с местом хранения данных
А.8.9 Вариант использования № 49: Подповерхностные биогеохимические исследования	—	—	От данных биологических наук — «омиков», от геномики микробов в почве до гидробиогеохимии водораздела; от данных наблюдений до результатов экспериментов	PFLOWTran, Postgres, HDF5, Akuna, NEWT и др.	Интеллектуальный анализ данных, оценка качества данных, взаимная корреляция между наборами данных, ускорение процесса разработки моделей, статистика, оценка качества, слияние данных

Окончание таблицы В.1

Вариант использования	Объем	Скорость обработки	Разнообразие	Программное обеспечение	Аналитика
А.8.10 Вариант использования № 50: Сети AmeriFlux и FLUXNET	—	Данные измерений газовых потоков поступают от ~150 вышек в сети AmeriFlux, и более 500 вышек, распределенных по всему миру	Данные о газовых потоках объединяются с биологическими данными, данными об атмосферных возмущениях и другими вспомогательными данными	EddyPro, специализированное аналитическое программное обеспечение, R, Python, нейронные сети, Matlab	Интеллектуальный анализ данных, оценка качества данных, взаимная корреляция между наборами данных, ассимиляция данных, интерполяция данных, статистика, оценка качества, слияние данных
А.9.1 Вариант использования № 51: Прогнозирование потребления электроэнергии в интеллектуальных энергосетях	4 терабайта в год для города с 1,4 млн датчиков, такого, как Лос-Анджелес	Поточная передача данных с миллионов датчиков	На основе кортежей: временные ряды, строки баз данных; На основе графов: топология сети, подключение клиентов; Некоторые семантические данные используются для нормализации	R/Matlab, Weka, Hadoop, визуализация на основе ГИС	Модели прогнозирования, модели машинного обучения, анализ временных рядов, кластеризация, выявление закономерностей, обработка сложных событий, визуальный анализ сети
А.9.2 Вариант использования № 52: Система управления энергией домашнего хозяйства HEMS	Около 14 тысяч домохозяйств. Об объемах данных сведений нет	В режиме реального времени потоковая передача данных с датчиков	«Умный» счетчик, электромобиль, панель солнечных батарей, осветительные приборы, кондиционер, топливный элемент, водонагреватель, аккумуляторная батарея	—	Сервис мониторинга энергопотребления, услуги по наблюдению за состоянием пожилых людей, помощь с выбором оптимального тарифного плана для электроэнергии, прогнозирование выработки электроэнергии фотоэлектрической системой, управление спросом на электроэнергию посредством стимулирования купонами (coupon incentive-based demand response, CIDR)

Приложение С
(справочное)

Сводка технических проблем вариантов использования

Из описаний вариантов использования были извлечены сведения о технических проблемах в семи категориях, описанных в разделе 6.1. Количество технических проблем по каждой категории варьировалось от варианта к варианту. Таблица С.1 содержит сведения о специфических для вариантов использования технических проблемах.

Таблица С.1 — Технические проблемы, специфические для вариантов использования

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
А.1.1 Вариант использования № 1: Архивное хранение больших данных переписи населения в США 2010 и 2000 годов	Большие объемы документов в центральном хранилище	—	Большое централизованное хранилище	—	Исполнение положений части 13 Свода законов США	1. Долговременная сохранность данных «как есть» в течение 75 лет. 2. Долговременная сохранность на уровне битов. 3. Курирование, включая преобразование формата. 4. Доступ и анализ после 75 лет. 5. Отсутствие утраты данных	—

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
А.1.2 Вариант использования № 2: Прием Национальными архивами США государственных данных на архивное хранение, поиск, извлечение и обеспечение долговременной сохранности	1. Распределенные источники данных. 2. Хранение больших объемов данных. 3. Неравномерно поступающие данные, партии от гигабайт до сотен терабайт. 4. Много разнообразных форматов в т. ч. для неструктурированных и структурированных данных. 5. Распределенные источники данных в различных облачных решениях	1. Сканирование и индексирование распределенных источников данных. 2. Различные методы аналитики, вкл. ранжирование, категоризацию данных и выявление ПДн. 3. Предварительная обработка данных. 4. Долговременная сохранность больших разнообразных наборов данных. 5. Поиск по огромному количеству данных с высокой релевантностью и полнотой результатов	1. Большое хранилище данных. 2. Различные системы хранения: NetApps, Hitachi, магнитные ленты	1. Высокая релевантность и полнота результатов поиска. 2. Высокая точность классификации документов. 3. Различные системы хранения: NetApps, Hitachi, магнитные ленты	Политика в области безопасности	1. Предварительная обработка, в т. ч. сканирование на вирусы. 2. Идентификация файлового формата. 3. Индексация. 4. Классификация документов	Мобильный поиск, имеющий интерфейс похожий на интерфейс стационарных компьютеров, и выдающий похожие результаты

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
А.1.3 Вариант использования № 3: Повышение активности респондентов в статистических обследованиях	Объем данных примерно один петабайт	Аналитика для рекомендательных систем, постоянного мониторинга и для общего совершенствования процесса обследования	Hadoop, Spark, Hive, R, SAS, Mahout, Allegrograph, MySQL, Oracle, Storm, BigMemory, Cassandra и Pig	Визуализация для проверки данных, оперативной деятельности и общего анализа; непрерывная эволюция	1. Улучшенные рекомендательные системы, позволяющие снизить затраты и повысить качество, обеспечивая одновременно надежные и публично проверяемые меры защиты конфиденциальности. 2. Безопасность и конфиденциальность данных. Возможность аудита процессов на предмет обеспечения безопасности и конфиденциальности	Высокая достоверность данных и надежность систем (проблемы: семантическая целостность концептуальных метаданных, описывающих, что именно измеряется, и вытекающие из этого пределы точности выводов)	Мобильный доступ
А.1.4 Вариант использования № 4: Использование нетрадиционных данных для повышения активности респондентов в статистических обследованиях	—	Аналитика для получения надежных оценок на базе данных традиционных обследований, государственных административных данных и данных из нетрадиционных источников из сферы цифровой экономики	Hadoop, Spark, Hive, R, SAS, Mahout, Allegrograph, MySQL, Oracle, Storm, BigMemory, Cassandra и Pig	Визуализация для проверки данных, оперативной деятельности и общего анализа; постоянная эволюция	Безопасность и конфиденциальность данных. Возможность аудита всех процессов на предмет безопасности и конфиденциальности согласно законодательству	Высокая достоверность данных, и надежность систем (проблемы: семантическая целостность концептуальных метаданных, описывающих, что именно измеряется, и вытекающие из этого пределы точности выводов)	—

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
А.2.1 Вариант использования № 5: Облачные вычисления в секторах финансовой отрасли	Ввод данных в реальном времени	Аналитика в реальном времени	—	—	Исполнение строгих требований к обеспечению безопасности и неприкосновенности частной жизни	—	Мобильный доступ
А.2.2 Вариант использования № 6: Международная исследовательская сеть Mendeley	1. Документы в виде файлов. В систему постоянно загружаются новые документы. 2. Различные типы файлов: PDF-файлы, лог-файлы социальных сети и активность клиентов, изображения, электронные таблицы, файлы презентаций	1. Стандартные библиотеки для машинного обучения и аналитики. 2. Эффективные масштабируемые и распараллеленные способы сопоставления документов, группировки похожих (вкл. те, что были модифицированы инструментами аннотирования третьих сторон или путем добавления титульных страниц или водяных знаков издателя)	1. Amazon EC2 с HDFS (инфраструктура). 2. S3 (хранение). 3. Hadoop (платформа). 4. Scribe, Hive, Mahout и Python (язык). 5. Хранилище умеренного объема (15 терабайт, с приростом 1 терабайт в месяц). 6. Обработка в реальном времени и пакетная	1. Специализированные инструменты создания отчетов. 2. Визуализация графа сети с помощью Gephi; диаграммы рассеяния и т. д.	Контроль доступа: кто и к какому контенту получает доступ	1. Управление метаданными, извлеченными из PDF-файлов. 2. Выявление дублирования документов. 3. Постоянные идентификаторы. 4. Сопоставление метаданных со сведениями в базах данных Crossref, PubMed и arXiv	Доставка контента и услуг на различные вычислительные платформы, от настольных компьютеров под Windows до мобильных устройств под ОС Android и iOS

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
А.2.3 Вариант использования № 7: Сервис кинофильмов Netflix	Профили пользователей и рейтинговая информация	1. Передача потокового видео многочисленным клиентам. 2. Аналитика для подбора фильмов, соответствующих интересам клиента. 3. Различные методы аналитики для персонализации услуг. 4. Надежные алгоритмы обучения. 5. Непрерывная аналитическая обработка на основе результатов мониторинга и оценки эффективности	1. Hadoop (платформа). 2. Pig (язык). 3. Cassandra и Hive. 4. Огромное количество подписчиков, рейтингов и поисков в сутки (база данных). 5. Огромное хранилище (2 петабайта). 6. Обработка с интенсивным вводом-выводом	Потоковая передача и представление видеоматериалов	Неприкосновенность частной жизни пользователей и соблюдение цифровых прав на видеоконтент	Постоянное вычисление рейтингов и их обновление на основе профилей пользователей и результатов аналитики	Интеллектуальные интерфейсы для доступа к киноконтенту на мобильных платформах
А.2.4 Вариант использования № 8: Веб-поиск	1. Распределенные источники данных. 2. Потоковые данные. 3. Мультимедийный контент	1. Динамическая доставка контента по сети. 2. Связывание профилей пользователей и данных из социальных сетей	Петабайты текстовых и мультимедийных данных (хранение)	1. Время поиска ≈0,1 секунды. 2. Максимизация метрики «точность 10 наилучших результатов». 3. Адекватный макет страницы выдачи результатов (визуализация)	1. Контроль доступа. 2. Защита чувствительного контента	1. Уничтожение данных по истечении определенного времени (несколько месяцев). 2. Чистка данных	Мобильный поиск и отображение

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
А.2.5 Вариант использования № 9: Обеспечение непрерывности деловой деятельности и восстановления после катастроф в облачной экосистеме	—	1. Надежный алгоритм резервного копирования. 2. Репликация последних изменений	1. Hadoop. 2. Коммерческие облачные сервисы	—	Высокий уровень безопасности во многих приложениях	—	—
А.2.6 Вариант использования № 10: Грузоперевозки	Централизованные и распределенные источники информации/ датчики, в реальном времени	1. Отслеживание объекта на основе уникальной идентификации с использованием установленного на нем датчика и координат GPS. 2. Обновление в реальном времени сведений об отслеживаемых объектах	Подключение к Интернету	—	Политика в области безопасности	—	—
А.2.7 Вариант использования № 11: Данные о материалах	1. Распределенные хранилища данных о более чем 500 тысячах коммерческих материалов. 2. Множество видов наборов данных. 3. Тексты, графики и изображения	Описания свойств материалов, содержание сотни независимых переменных. Сбор значений этих переменных для создания надежных наборов данных	—	1. Визуализация для отыскания подходящих материалов, свойства которых зависят от множества независимых переменных. 2. Многопараметрические инструменты визуализации	1. Защита чувствительных проприетарных данных. 2. Средства маскирования проприетарной информации	Управление качеством данных (сейчас низкое или непонятное)	—

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
А.2.8 Вариант использования № 12: Геномика материалов на основе результатов моделирования	1. Потоки данных от пета/экзафлопсных централизованных систем моделирования. 2. Распределенные веб-потоки данных от центрального шлюза к пользователям	1. Анализ данных в режиме реального времени с использованием вычислений с высокой пропускной способностью для оперативного реагирования. 2. Комбинирование результатов моделирования с использованием различных программ. 3. Поисковые исследования, ориентированные на потребности потребителей; вычислительная база должна гибко адаптироваться к новым целям. 4. Map/Reduce и поиск, для комбинирования данные моделирования и экспериментальных данных	1. Массивная (суперкомпьютер Cray XE6 «Hopper», 150 тыс. процессоров) унаследованная инфраструктура (инфраструктура). 2. GPFS (хранение). 3. MonogDB (платформа). 4. Сеть 10 гигабит/с. 5. Различные аналитические инструменты: PyMatGen, FireWorks, VASP, ABINIT, NWChem, BerkeleyGW и различное ПО сообщества. 6. Большое хранилище (хранение). 7. Масштабируемые базы данных для данных «ключ-значение» и объектов (платформа). 8. Потоки данных от пета/экзафлопсных централизованных систем моделирования	Программы просмотра данных о материалах, необходимые ввиду растущих объемов выдаваемых в ходе поиска данных	1. Возможность работать в изолированной зоне — песочнице и создавать независимые рабочие зоны для заинтересованных сторон. 2. Объединение (федерацию) наборов данных на основе политик	1. Валидация и оценка неопределенности результатов моделирования путем сопоставления с экспериментом. 2. Количественная оценка неопределенности на основе нескольких наборов данных	Мобильные приложения для доступа к информации по геномике материалов

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
А.3.1 Вариант использования № 13: Облачный крупномасштабный анализ и визуализация геопространственных данных	Уникальные подходы для индексирования и распределенного анализа геопространственных данных	1. Аналитика: ближайшая точка подхода, отклонение от маршрута, плотность точек во времени, метод главных компонентов (PCA) и метод анализа независимых компонентов (ICA). 2. Уникальные подходы для индексирования и распределенного анализа геопространственных данных	Реляционная СУБД с геопространственной поддержкой; геопространственный сервер/ПО для анализа — ESRI ArcServer, Geoserver	Визуализация посредством ГИС как при высокой, так при низкой пропускной способности сети, а также на выделенных устройствах и на портативных устройствах	Безопасность чувствительных данных при передаче и при хранении (особенно на портативных/карманных устройствах)	—	—

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
А.3.2 Вариант использования № 14: Идентификация и отслеживание объектов — Постоянное наблюдение	Поступающие в реальном времени данные FMV-формата высококачественного видео (от 30 до 60 кадров в секунду при полноцветном разрешении 1080 пикселей) и WALF-формат видео с высоким разрешением (WALF) — от 1 до 10 кадров в секунду при полноцветном разрешении 10 тысяч * 10 тысяч пикселей	Расширенная аналитика: средства идентификации объекта, анализа закономерностей поведения объекта, анализа группового поведения/динамики и хозяйственной деятельности, а также для объединения (слияния) данных	1. Широкий спектр специализированного ПО и инструментов, включая реляционные СУБД и средства отображения. 2. Несколько каналов сетевого взаимодействия. 3. Кластеры графических процессоров (GPU)	1. Визуализация извлеченных результатов путем наложения на отображение геопространственных данных. Обратные ссылки на соответствующие сегменты исходного изображения/видеопотока. 2. Выходные данные в форме веб-функций, соответствующих стандартам «Открытого геопространственного консорциума» (OGC), либо в виде стандартных геопространственных файлов (Shapefile, KML)	Высокий уровень безопасности и конфиденциальности; нельзя допустить компрометацию источников данных и методов их обработки	Достоверность извлеченных объектов	—

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
А.3.3 Вариант использования № 15: Обработка и анализ разведывательных данных	1. Данные, поступающие в реальном времени с их обработкой (в худшем случае) в масштабе времени, близком к реальному. 2. Данные в разрозненных хранилищах должны быть доступны через семантически интегрированное пространство данных. 3. Разнообразные данные: текстовые файлы, первичные данные с датчиков, изображения, видео, аудио, электронные данные и данные, созданные человеком	Аналитика: оповещения в масштабе времени, близком к реальному, основанные на закономерностях и изменениях основных параметров	1. Стабильность системы в случае ненадежной связи с солдатами и удаленными датчиками. 2. До сотен петабайт, хранимых средними и крупными кластерами и облачными системами. 3: Hadoop, Accumulo (с системой хранения данных BigTable), Solr, NLP (несколько вариантов), Puppet (управление жизненным циклом ИТ, обеспечение безопасности), Storm, а также специализированные приложения и инструменты визуализации	Визуализация: наложения на геопространственную картину и сетевые графики (network diagrams)	Защита данных от несанкционированного доступа или раскрытия и от несанкционированного вмешательства	Происхождение данных (включая, например, отслеживание всех передач и преобразований) в течение жизненного цикла данных	—

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
А.4.1 Вариант использования № 16: Данные электронной медицинской документации	1. Неоднородные, большого объема, разнообразные источники данных. 2. Объем: > 12 млн пациентов, > 4 млрд отдельных клинических наблюдений, всего более 20 терабайт первичных данных. 3. Скорость: от 500 тыс. до 1,5 млн новых клинических транзакций в день. 4. Разнообразие форматов: числовые и структурированные числовые данные, тексты в свободном формате, структурированные тексты, дискретные номинальные данные, дискретные порядковые данные, дискретные структурированные данные, большие двоичные объекты (изображения и видео). 5. Данные с течением времени эволюционируют	1. Всестороннее и согласованное представление данных из разных источников во времени. 2. Методы аналитики: методы извлечения информации с целью выявления клинических признаков; обработка естественного языка; машинное обучение моделей принятия решений; методы оценки максимального правдоподобия и Байесовских сетей	1. Hadoop, Hive и R на основе Unix. 2. Суперкомпьютер Cray. 3. Teradata, PostgreSQL, MongoDB. 4. Различные сетевые возможности с учетом значительных объемов обработки с интенсивным вводом — выводом	Предоставление результатов аналитики для использования потребителями данных/заинтересованными сторонами, то есть теми, кто сам анализ не проводил	Прямой доступ потребителей к данным, а также ссылки на результаты аналитики, выполненной специалистами в области информатики и исследователями системы здравоохранения. 2. Защита всех данных о здоровье в соответствии с действующим законодательством. 3. Защита данных в соответствии с политиками поставщиков данных. 4. Политики безопасности и защиты ПДн, уникальные для конкретных подмножеств данных. 5. Надежная безопасность для предотвращения утечек данных	1. Стандартизация, агрегирование и нормализация данных из разнородных источников. 2. Уменьшение количества ошибок и устранение систематических погрешностей. 3. Общая номенклатура и классификация контента из разных источников	Обеспечение безопасности на мобильных устройствах

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
А.4.2 Вариант использования № 17: Анализ графических образов в патологии/Цифровая патология	1. Пространственные цифровые графические образы высокого разрешения в патологии. 2. Различные алгоритмы анализа качества изображений. 3. Различные форматы графических данных, особенно BigTIFF; и результаты анализа в структурированном виде. 4. Анализ изображений, пространственные запросы и аналитика, кластеризация и классификация признаков	1. Высокопроизводительный анализ изображений с целью извлечения пространственной информации. 2. Пространственные запросы и аналитика, кластеризация и классификация признаков. 3. Аналитическая обработка огромного многомерного набора данных, возможность корреляции с данными других типов, такими, как клинические данные и данные биологических наук — «омиков»	1. Унаследованные системы и облачные решения (вычислительный кластер). 2. Огромные объемы данных в унаследованных и новых системах хранения, таких как SAN и HDFS (хранение). 3. Сетевые соединения с высокой пропускной способностью (сети). 4. Анализ изображений с использованием MPI, Map/Reduce и Hive с пространственным расширением (пакеты программ).	Визуализация для целей проверки и обучения	Обеспечение безопасности и защита ПДн для защищаемой медицинской информации	Аннотирование материалов человеком для использования при валидации	Трехмерная визуализация и отображение на мобильных платформах

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
А.4.3 Вариант использования № 18: Вычислительный анализ биоизображений	1. Распределенные мультимодальные экспериментальные источники (инструменты) биологических изображений высокого разрешения. 2. 50 терабайт данных в различных форматах, включая графические	1. Высокопроизводительные вычисления и управление анализом полученных результатов. 2. Сегментация представляющих интерес областей; групповой отбор и извлечение признаков, классификация объектов, организация и поиск. 3. Расширенное выявление представляющих интерес для биологических наук новых явлений, с помощью методов больших данных/экстремальных вычислений, обработки и анализа данных непосредственно в базе данных, машинного обучения (SVM и RF) для сервисов классификации и рекомендательных сервисов, продвинутых алгоритмов для массового анализа изображений	1. ImageJ, OMERO, VolRover, разработанные прикладными математиками продвинутые методы сегментации и выявления признаков. Необходимы масштабируемые базы данных для данных типа «ключ-значение» и для библиотек объектов. 2. Инфраструктура суперкомпьютера Hopper в NERSC. 3. Базы данных и коллекций изображений. 4. 10-гигабитные, в будущем 100-гигабитные сети и расширенные сетевые возможности (SDN)	Работа с трехмерными структурными моделями	Достаточно высокий, но не являющийся обязательным уровень безопасности и защиты ПДн, включая использование защищенных серверов и анонимизацию	Компоненты потока рабочих процессов, вкл. сбор, хранение, улучшение качества данных и минимизацию шума	—

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
		и высокопроизводительных вычислительных решений. 4. Массовый анализ данных применительно к масштабным наборам данных изображений					
А.4.4 Вариант использования № 19: Геномные измерения	1. Поступающие с высокой скоростью сжатые данные (~300 гигабайт в день) от различных секвенсоров ДНК. 2. Распределенные источники данных (секвенсоры). 3. Различные файловые форматы как для структурированных, так и для неструктурированных данных	1. Обработка первичных данных с целью выделения вариаций. 2. Машинное обучение для комплексного анализа систематических ошибок технологий секвенирования, которые сложно охарактеризовать	1. Унаследованный вычислительный кластер и другие PaaS и IaaS-решения (вычислительный кластер). 2. Огромное хранилище данных петабайтного масштаба (хранение). 3. Унаследованное ПО с открытым исходным кодом для секвенирования в биоинформатике на основе UNIX (пакет программ)	Формат данных, используемый браузерами генома	Обеспечение безопасности и защита персональных данных для медицинских документов и баз данных клинических исследований	—	Обеспечение врачам доступа к геномным данным на мобильных платформах

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
А.4.5 Вариант использования № 20: Сравнительный анализ (мета) геномов	1. Многочисленные централизованные источники данных. 2. От сведений о последовательностях аминокислот до данных о белках и их структурных особенностях (базовые геномные данные), а также данные биологических наук — «омиков», таких как транскриптомика, метиломика и протеомика, описывающих экспрессии генов в различных условиях. 3. Интерактивный пользовательский веб-интерфейс в реальном времени. Возможности обработки загружаемых данных на сервере должны соответствовать экспоненциальному росту объемов данных секвенирования из-за быстрого снижения стоимости технологии секвенирования	1. Методы сравнительного анализа очень сложных данных. 2. Описательная статистика	1. Огромное хранилище данных. 2. Масштабируемая реляционная СУБД для разнородных биологических данных. 3. Быстрая и параллельная массовая загрузка в реальном времени. 4. Реляционная СУБД Oracle, файлы SQLite, плоские текстовые файлы, Lucy (версия Lucene) для поиска по ключевым словам, базы данных BLAST, базы данных USEARCH. 5. Linux-кластер, сервер реляционной СУБД Oracle, большие системы хранения данных, стандартные интерактивные хосты Linux	1. Параллельная массовая загрузка в реальном времени. 2. Интерактивный пользовательский веб-интерфейс к основным данным, предварительные вычисления на сервере и отправка пакетных заданий из пользовательского интерфейса. 3. Скачивание сформированных и аннотированных наборов данных для анализа в автономном режиме. 4. Возможность запрашивать и просматривать данные через интерактивный пользовательский веб-интерфейс. 5. Визуализация структурных элементов на разных уровнях разрешения; возможность представления группы очень похожих геномов в виде пангенома	1. Безопасность учетных данных для входа в систему, т. е. логинов и паролей. 2. Создание учетных записей пользователей для доступа к наборам данных и представления наборов данных в систему через веб-интерфейс. 3. Технология единого входа (SSO)	1. Методы повышения качества данных. 2. Кластеризация, классификация и редуцирование данных. 3. Интеграция новых данных/контента в системное хранилище данных и аннотирование данных	—

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
	4. Разнородные, сложные, структурные и иерархические биологические данные. 5. Метагеномные образцы, размеры которых могут варьироваться на несколько порядков величины — от нескольких сотен тысяч до миллиарда генов						
А.4.6 Вариант использования № 21: Индивидуальное управление лечением диабета	1. Распределенные данные электронных медицинских документов. 2. Более 5 млн пациентов с тысячами свойств по каждому, и производные данные на основе первичных. 3. По каждому пациенту число значений свойств от ~100 до более чем 100 тыс.; в среднем ~100 значений свойств из контролируемых словарей и 1000 числовых величин.	1. Интеграция данных с использованием аннотаций на основе онтологий и таксономий. 2. Алгоритмы параллельного поиска и извлечения как для поиска по индексу, так и для настраиваемого поиска; способность выделять представляющие интерес данные: когорты пациентов, пациентов, удовлетворяющих определенным критериям, и пациентов,	1. Хранилища данных, в т. ч. нереляционная СУБД Hbase с открытым исходным кодом. 2. Суперкомпьютеры, облачные и параллельные вычисления. 3. Обработка с интенсивным вводом-выводом. 4. Распределенная файловая система HDFS. 5. Специализированное ПО для выявления новых признаков на основе хранимых данных	Эффективная визуализация данных на основе графов	1. Защита медицинских данных в соответствии с политиками защиты ПДн и законодательно-нормативными требованиями к безопасности и защите персональных данных, например, американского закона HIPAA. 2. Политики безопасности для разных пользовательских ролей	1. Аннотирование данных на основе онтологий и таксономий. 2. Прослеживаемость данных от источника (начальной точки сбора) и далее на протяжении периода работы с ними. 3. Преобразование данных из существующего хранилища данных в триплеты RDF	Мобильный доступ

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
	4. Данные периодически обновляются (не в режиме реального времени). Данные снабжаются отметками времени наблюдения (времени записи значения). 5. Две основные категории данных: со значениями из контролируемого словаря и числовыми значениями (которые документируются/ регистрируются чаще). 6. Данные состоят из текста и числовых значений	имеющих сходные характеристики. 3. Алгоритмы распределенного интеллектуального анализа закономерностей в графе, индексации графов, а также поиска закономерностей в графах на основе триплетов RDF. 4. Надежные инструменты статистического анализа для контроля частоты ложных срабатываний, определения значимости подграфа и исключения ложных позитивных и ложных негативных результатов. 5. Алгоритмы интеллектуального анализа закономерностей в графах, их индексации и поиска по графам. 6. Обход семантического графа					

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
А.4.7 Вариант использования № 22: Статистический реляционный искусственный интеллект для здравоохранения	1. Централизованные данные, некоторые данные — из интернет-источников. 2. Данные в диапазоне от сотен Гб для одной когорты из нескольких сотен человек, и до одного Пб в очень масштабных исследованиях. 3. Как постоянно обновляемые/пополняемые данные о пациентах, так и данные, поступающие партиями по графику. 4. Большие, мультимодальные данные длительного наблюдения. 5. Богатые реляционные данные, состоящие из многочисленных таблиц, а также различные типы данных, такие как изображения, электронныемедицинские документы, демографические,	1. Реляционные вероятностные модели, моделирующие неопределенности на основе теории вероятности. ПО обучает модели на основе ряда типов данных, потенциально может интегрировать информацию и логические рассуждения о сложных запросах. 2. Надежные и точные методы обучения для учета дисбаланса данных, когда большие объемы данных доступны для небольшого числа субъектов. 3. Алгоритмы обучения для определения перекосов в данных, чтобы избежать ошибочного моделирования «шума». 4. Обобщенные и уточненные обученные модели для применения	1. Java, некоторые инструменты собственной разработки, реляционную базу данных и хранилища NoSQL. 2. Облачные и параллельные вычисления. 3. Высокопроизводительный компьютер с 48 гигабайт ОЗУ (для анализа при умеренном размере выборки). 4. Вычислительные кластеры для обработки больших наборов данных. 5. Жесткий диск объемом от 200 гигабайт до 1 терабайта для тестовых данных	Визуализация подмножеств очень больших наборов данных	Защищенная обработка данных	1. Объединение нескольких таблиц перед выполнением анализа. 2. Методы валидации данных с целью минимизации ошибок	—

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
	генетические данные и данные на естественном языке, требующие богатых средств представления. 6. Непредсказуемые темпы поступления данных, которые во многих случаях поступают в режиме реального времени	к другим наборам данных. 5. Принятие данных в разных формах и из разрозненных источников					
А.4.8 Вариант использования № 23: Эпидемиологическое исследование в масштабе всего населения Земли	1. Синтетическая глобальная популяция, на централизованных либо распределенных ресурсах. 2. Большие объемы выходных данных, поступающих в режиме реального времени. 3. Различные выходные наборы данных в зависимости от сложности модели	1. Вычисления, требующие как значительных вычислительных ресурсов, так и обработки больших объемов данных, соответствуют характеристикам суперкомпьютеров. 2. Алгоритмы, учитывающие неструктурированный и нерегулярный характер обработки графов. 3. Получение сводок по различным прогонам и повторам моделирования	1. Перемещение очень больших объемов данных для визуализации (сети). 2. Распределенная система моделирования на основе MPI (платформа). 3. Charm++ на нескольких узлах (ПО). 4. Сетевая файловая система (хранение). 5. Сеть Infiniband (сети)	Визуализация	1. Защита используемых в моделировании персональных данных физических лиц. 2. Защита данных и защищенная платформа для вычислений	Качество данных и отслеживание происхождения данных в ходе вычислений	—

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
А.4.9 Вариант использования № 24: Моделирование распространения социального влияния	1. Динамическая распределенная обработка с использованием как традиционной архитектуры коммерческих кластеров, так и более новых (например, облачной). 2. Модели с высокой детализацией; и наборы данных, поддерживающие сетевой трафик Twitter. 3. Хранение огромных объемов данных	1. Крупномасштабное моделирование различных событий (болезни, эмоции, поведение и т. д.). 2. Масштабируемое объединение наборов данных. 3. Многоуровневый анализ, одновременно обеспечивая быстрое получение достаточных результатов	1. Вычислительная инфраструктура, позволяющая моделировать различные типы взаимодействия между людьми через интернет в связи с различными социальными событиями (инфраструктура). 2. Файловые сервера и базы данных (платформа). 3. Сети Ethernet и Infiniband (сети) 4. Специализированные программы моделирования, ПО с открытым исходным кодом и проприетарные среды моделирования. (приложения). 5. Обработка огромного количества учетных записей пользователей социальных сетей из различных стран (сети)	1. Многоуровневые детальные представления в виде сетей. 2. Визуализация с возможностью интерактивного взаимодействия	1. Защита используемых в моделировании персональных данных физических лиц. 2. Защита данных и защищенная платформа для вычислений	1. Объединение данных из различных источников данных. 2. Согласованность данных и предотвращение их порчи. 3. Предварительная обработка первичных данных	Перемещение данных ближе к вычислительным ресурсам с целью повышения эффективности

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
А.4.10 Вариант использования № 25 Биологическое разнообразие и европейская электронная научно-исследовательская инфраструктура LifeWatch	1. Специальные выделенные или оверлейные (наложенные) сенсорные сети. 2. Распределенное хранение, в том числе архивирование и сохранение исторических данных и данных о тенденциях. 3. Распределенные источники данных, в том числе многочисленные пункты наблюдения и мониторинга, сети датчиков и спутники. 4. Широкий спектр данных, включая спутниковые изображения/информацию, данные о климате и погоде, фотографии, видео и звукозаписи и т. д. 5. Комбинации данных различных типов, и связи с потенциально неограниченными в своем разнообразии данными. 6. Потоковая передача данных	1. Поэтапный анализ и/или анализ данных в реальном времени; темпы поступления данных варьируются в зависимости от исходных биологических и экологических процессов. 2. Разнообразие данных, аналитических инструментов и инструментов моделирования для поддержки аналитики в интересах различных научных сообществ. 3. Аналитика параллельных потоков данных и аналитика данных, поступающих в потоковом режиме. 4. Доступ и интеграция нескольких распределенных баз данных	1. Расширяемые и предоставляемые по требованию ресурсы хранения для глобальных пользователей. 2. Облачные ресурсы сообщества. 3. Веб-сервисы, грид-сервисы, реляционные базы данных. 4. Персонализированные «виртуальные лаборатории». 5. Грид-ресурсы и облачные ресурсы	1. Доступ мобильных пользователей. 2. Развитая и богатая визуализация, средства визуализации высокой четкости. 3. 4D-визуализация	1. Объединенное (федеративное) управление идентификацией для мобильных исследователей и мобильных датчиков. 2. Управление доступом и контроль над ним	1. Хранение и архивация данных, обмен данными и их интеграция. 2. Управление жизненным циклом данных, включая происхождение данных, ссылочную целостность и идентификацию, прослеживаемость до первоначальных данных наблюдений. 3. Обработанные (вторичные) данные (в дополнение к исходным данным), для использования в будущем 4. Контроль происхождения с присвоением постоянного идентификатора (PID) данных, алгоритмов и рабочих процессов. 5. Курированные (авторизованные) эталонные данные (т. е. списки названий видов), алгоритмы, программные коды и рабочие процессы	—

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
А.5.1 Вариант использования № 26: Крупномасштабное глубокое обучение	—	—	1. Графические процессоры. 2. Высокопроизводительный кластер с внутренними соединениями на основе MPI и Infiniband. 3. Библиотеки для вычислений на одной машине или на одном графическом процессоре (например, BLAS, CuBLAS, MAGMA и др.). 4. Распределенные вычисления с плотными матрицами на графических процессорах, подобно BLAS или LAPACK, которые пока слабо развиты. Существующие решения (например, ScaLapack для центральных процессоров) не очень хорошо интегрированы с языками высокого уровня и требуют низкоуровневого программирования, что удлиняет время эксперимента и процесса разработки	—	—	—	—

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
А.5.2 Вариант использования № 27: Организация крупномасштабных, неструктурированных коллекций, сделанных потребителями фотографий	Более 500 миллионов изображений, загружаемых ежедневно на сайты социальных сетей	1. Классификатор (например, SVM) — процесс, который часто трудно распараллелить. 2. Функциональные возможности, применяемые во многих крупномасштабных задачах обработки изображений	Hadoop или усовершенствованный Map/Reduce	Визуализация крупномасштабных трехмерных реконструкций и навигация по крупномасштабным коллекциям изображений, которые были согласованы с картами	Требуется обеспечивать защиту ПДн пользователей и защиту цифровых прав на контент	—	—
А.5.3 Вариант использования № 28: Truthy — Анализ данных Твиттера	1. Распределенные источники данных. 2. Большие объемы данных и потоковая передача в реальном времени. 3. Первичные данные в сжатых форматах. 4. Полностью структурированные данные в формате JSON, пользовательские метаданные и данные геолокации. 5. Несколько схем данных	Различные методы анализа данных в реальном времени для выявления аномалий, кластеризации потока, классификации сигналов на основе многомерных временных рядов и онлайн-обучения	1. Hadoop и HDFS (платформа). 2. Indexed HBase, Hive, SciPy и NumPy (ПО). 3. Базы данных в памяти и MPI (платформа). 4. Высокоскоростная сеть Infiniband (сети)	1. Поиск/извлечение данных и их динамическая визуализация. 2. Управляемые данными интерактивные веб-интерфейсы. 3. API-интерфейсы программирования приложений для запросов к данным	Политика в области безопасности и защиты неприкосновенности частной жизни	Стандартизированные структуры данных/форматы и исключительно высокое качество данных	Низкоуровневые функциональные возможности инфраструктуры хранения данных с целью обеспечения эффективного мобильного доступа к данным

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
А.5.4 Вариант использования № 29: Краудсорсинг в гуманитарных науках	—	1. Оцифровка существующих архивов документов и аудио-, видео- и фотоматериалов. 2. Аналитика, включая все виды распознавания закономерностей (например, распознавание речи, автоматический анализ аудиовизуальных материалов, культурные закономерности) и выявления структур (лексические единицы, лингвистические правила и т. д.)	—	—	Требуется решать вопросы обеспечения неприкосновенности частной жизни, сохраняя анонимность авторов полученных материалов	—	—

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
А.5.5 Вариант использования № 30: Цифровая инфраструктура для исследований и анализа сетей и графов» (CINET)	1. Набор файлов сетевых топологий для изучения теоретических свойств графов и поведения различных алгоритмов. 2. Асинхронные и синхронные распределенные вычисления в реальном времени	1. Среды для запуска различных инструментов анализа сетей и графов. 2. Динамический рост сетей. 3. Асинхронные и синхронные, выполняемые в реальном времени распределенные вычисления. 4. Различные параллельные алгоритмы для разных схем разделения, используемых для повышения эффективности вычислений	1. Высокопроизводительная кластеризованная файловая система (хранение). 2. Различные сетевые подключения (сети). 3. Существующий вычислительный кластер. 4. Вычислительный кластер Amazon EC2. 5. Различные библиотеки для работы с графами, инструменты управления потоками процессов, СУБД и семантические веб-инструменты	Визуализация на стороне клиента	—	—	—

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
А.5.6 Вариант использования № 31: Измерения и оценки эффективности аналитических технологий в Национальном институте стандартов и технологий (NIST)	1. Большое количество частично аннотированных веб-страниц, твитов, изображений и видеозаписей. 2. Масштабирование процесса проверки на большие объемы данных; измерение внутренней неопределенности и неопределенности аннотаций, измерение эффективности для не полностью аннотированных данных, измерение эффективности аналитики для разнородных данных и аналитических потоков с участием пользователей	Аналитические алгоритмы, работающих с письменным языком, речью, изображениями людей и т. д. Алгоритмы, как правило, следует тестировать на реальных или реалистичных данных. Проблематично создание искусственных данных, в достаточной степени отражающих вариативность реальных данных, связанных с людьми	1. Средства разработки PERL, Python, C/C++, Matlab, R. 2. Разработка по принципу «снизу вверх» тестовых и измерительных приложений	Потоки работ аналитики с участием пользователей	Исполнение требований по безопасности и защите ПДн в отношении защиты чувствительных данных, обеспечивая при этом возможность проведения содержательной оценки эффективности разработок. Совместно используемые испытательные стенды должны обеспечивать защиту интеллектуальной собственности разработчиков аналитических алгоритмов	—	—

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
А.6.1 Вариант использования № 32: Консорциум федеративных сетей данных (DFC)	1. Обработка ключевых файловых форматов: NetCDF, HDF5, Dicom. 2. Обработка данных в режиме реального времени и пакетная обработка	Типовые потоки рабочих процессов аналитики	1. ПО для управления данными iRODS. 2. Интероперабельность между различными типами протоколов хранения и сетевого взаимодействия	Типовые потоки рабочих процессов визуализации	1. Объединение (федерация) существующих сред аутентификации с помощью «Типового API-интерфейса служб защиты данных» (Generic Security Service API) и подключаемых модулей аутентификации (GSI, Kerberos, InCommon, Shibboleth). 2. Управление доступом к файлам независимо от места хранения	—	—
А.6.2 Вариант использования № 33: Discinnet-процесс	Интеграция методов работы с метаданными различных дисциплин	—	Программное обеспечение: Symfony-PHP, Linux и MySQL	—	Достаточно высокий, но необязательный уровень безопасности и защиты ПДн, включая использование защищенных серверов и анонимизацию	Интеграция методов работы с метаданными различных дисциплин	—
А.6.3 Вариант использования № 34: Поиск по графу для научных данных	Любые типы данных, от изображений до текстов, от структур до белковых последовательностей	1. Обработка графа данных. 2. Реляционная СУБД	Облачные ресурсы сообщества	Эффективная визуализация на основе графа данных	—	—	—

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
А.6.4 Вариант использования № 35: Анализ больших объемов данных, получаемых в экспериментах на синхротроне	1. Многочисленные потоки данных в реальном времени, сохранение данных для последующего анализа. 2. Анализ в режиме реального времени выборок данных	Стандартные инструменты биоинформатики (BLAST, HMMER, инструменты множественного выравнивания последовательностей и филогенетики, программы поиска/предсказания генов и генных структур, программы предсказания свойств по результатам секвенирования и т. д.), скрипты Perl / Python и планировщик задач Linux — кластера	Передача больших объемов данных на удаленный ресурс для пакетной обработки	—	Исполнение многочисленных требований к безопасности и защите неприкосновенности частной жизни	—	—
А.7.1 Вариант использования № 36: Каталинский цифровой обзор неба в поисках транзиентов	Обработка поступающих за ночь ≈ 0.1 Тб первичных данных обзора; в будущем темпы производства данных могут возрасти в 100 раз	1. Большое количество разнообразных инструментов анализа астрономических данных, а также большое количество специализированных инструментов и ПО, часть которых является	—	Механизмы визуализации для пространств параметров данных высокой размерности	—	—	—

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
		самостоятельными исследовательскими проектами. 2. Автоматизированная классификация с помощью инструментов машинного обучения, учитывающая немногочисленность и разнородность данных, которая динамически эволюционирует во времени по мере поступления большего количества данных; и принятия решений о проведении дополнительных исследований в условиях ограниченности выделяемых для этого ресурсов					

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
А.7.2 Вариант использования № 37: Космологический обзор неба и моделирование	Обработка ≈1 петабайта данных наблюдений в год. В будущем темпы производства данных вырастут до 7 петабайт в год	Интерпретация результатов детального моделирования, которая требует развитых методов и средств анализа и визуализации	1. MPI, OpenMP, C, C++, F90, FFTW, пакеты визуализации, Python, FFTW, Numpy, Boost, OpenMP, ScaLAPCK, СУБД PSQL и MySQL, Eigen, Cfitsio, http://astrometry.net/ и Minuit2. 2. Разработка новых методов анализа ввиду ограничений подсистемы ввода/вывода суперкомпьютера	Интерпретация результатов с использованием передовых методов и средств визуализации	—	—	—
А.7.3 Вариант использования № 38: Большие данные космологических обзоров неба	Обработку ~20 терабайт данных в день	1. Анализ как результатов моделирования, так и данных наблюдений. 2. Методы для выполнения разложения Холецкого для тысяч моделирований с матрицами порядка миллиона по каждой стороне	1. Стандартное астрофизическое ПО для обработки («редуцирования») данных, а также сценарии — обертки Perl/ Python. 2. Реляционная СУБД Oracle, терминальный клиент psql, файловые системы GPFS и Luster и ленточные архивы. 3. Параллельные базы данных для хранения изображений	—	—	Связи между удаленными телескопами и центрами аналитической обработки	—

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
А.7.4 Вариант использования № 39: Анализ данных Большого адронного коллайдера	1. Обработка данных, поступающих в реальном времени от ускорителей и инструментов анализа. 2. Асинхронизация сбора данных. 3. Калибровка экспериментальных установок	1. Экспериментальные данные проектов ALICE, ATLAS, CMS и LHC. 2. Гистограммы, диаграммы рассеяния, подбор моделей 3. Вычисления по методу Монте-Карло	1. Унаследованная вычислительная инфраструктура (вычислительные узлы). 2. Распределенное хранение файлов (хранение) 3. Объектно-ориентированные базы данных (ПО)	Построение гистограмм, диаграмм рассеяния с подбором моделей (визуализация)	Защита данных	Качество данных на сложных установках	—
А.7.5 Вариант использования № 40: Эксперимент Belle II	120 петабайт первичных данных	—	1. Хранение 120 петабайт первичных данных. 2. Модель международных распределенных вычислений, для расширения имеющихся возможностей на ускорителе (в Японии). 3. Передача первичных данных со скоростью ~20 гигабит/с между Японией и США (при проектной яркости ускорителя). 4. Программное обеспечение: «Грид Открытой науки» (Open Science Grid), Geant4, DIRAC, FTS, инфраструктура Belle II	—	Стандартная аутентификация в грид-системе	—	—

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
А.8.1 Вариант использования № 41: Радарная система некогерентного рассеяния EISCAT-3D	1. Систему из пяти постов, которая будет производить 40 петабайт данных в год в 2022 году. 2. Формат данных Hierarchical Data Format (HDF5). 3. Визуализация многомерных (≥5) данных	1. Архитектура «пчелиной матки» (Queen Bee), в которой централизованная обработка сочетается с распределенной обработкой на измерительных устройствах для данных с 5 распределенных постов. 2. Мониторинг оборудования в режиме реального времени путем частичного анализа потока данных. 3. Богатый набор сервисов обработки радиолокационных изображений с использованием машинного обучения, статистического моделирования и алгоритмов поиска на графе	Архитектура, позволяющая принимать участие в сотрудничестве в рамках проекта ENVRI	Визуализация многомерных (≥5) данных	—	Долговременная сохранность данных и предотвращение утраты данных в случае сбоев в работе измерительного комплекса	Требуется поддержка мониторинга оборудования в режиме реального времени, посредством частичного анализа потока данных

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
А.8.2 Вариант использования № 42: Совместная деятельность европейских сетевых инфраструктур в области экологических исследований (ENVRI)	1. Огромный объем данных, поступающих в реальном времени из распределенных источников. 2. Разнообразные наборы данных и метаданных, поступающих с измерительных инструментов	Разнообразные аналитические инструменты	1. Взаимодействие с различными вычислительными инфраструктурами и архитектурами (инфраструктура). 2. Взаимодействие с разрозненными хранилищами (хранение)	1. Инструменты построения графиков. 2. Инструменты интерактивной линейной временной визуализации (на базе Google Chart Tools) для временных рядов. 3. Отображение диаграмм в браузере с использованием технологии Flash. 4. Визуализация данных с высоким разрешением с привязкой к картам Земли. 5. Визуальные инструменты для сравнения качества моделей	Политика открытых данных с небольшими ограничениями	1. Высокое качество данных. 2. Зеркальные архивы. 3. Различные схемы метаданных. 4. Разрозненные хранилища и курирование данных	Мобильные датчики и измерительные устройства различных типов с целью сбора данных

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
А.8.3 Вариант использования № 43: Анализ радиолокационных данных для Центра дистанционного зондирования ледяного покрова CReSIS	1. Надежная передача данных с установленных на самолете датчиков/приборов либо со съемных жестких дисков, доставленных с удаленных объектов. 2. Сбор данных в режиме реального времени. 3. Различные наборы данных	1. Унаследованное ПО (Matlab) и языки (C/Java) для обработки данных. 2. Обработка сигналов и методы обработки изображений с целью выделения слоев	1. ≈0,5 петабайт первичных данных в год. 2. Передача материалов со съемного жесткого диска в вычислительный кластер для параллельной обработки. 3. Map/Reduce или MPI, плюс C/Java	1. ГИС как пользовательский интерфейс. 2. Богатый пользовательский интерфейс для моделирования	Обеспечение безопасности и неприкосновенности частной жизни, в том числе с учетом деликатности политической ситуации в зоне проведения исследований. Требуется поддерживать динамичные механизмы политик в области безопасности и неприкосновенности частной жизни	Обеспечение уверенности в качестве данных	Мониторинг собирающих данные устройств и датчиков
А.8.4 Вариант использования № 44: Обработка данных проекта UAVSAR	1. Пространственные данные и данные в угловых координатах. 2. Совместимость с другими радиолокационными системами и хранилищами данных НАСА, например, Спутникового центра НАСА на Аляске (Alaska Satellite Facility, ASF)	1. Данные с географической привязкой, требующие интеграции в ГИС в качестве дополнительных наложений (оверлеев). 2. Значительное вмешательство человека в конвейер обработки данных. 3. Поддержка богатого набора сервисов обработки радиолокационных изображений.4. Инструменты ROI_PAC,	1. Архитектура, обеспечивающая интероперабельность системы высокопроизводительных вычислений с облачными решениями. 2. Поддержка богатого набора сервисов обработки радиолокационных изображений. 3. Инструменты ROI_PAC, GeoServer, GDAL, а также поддерживающие стандарт	Поддержка пользователей в полевых экспедициях посредством предоставления интерфейса для смартфонов/планшетов и поддержки скачивания данных с низким разрешением		1. Значительное вмешательство человека в конвейер обработки данных. 2. Подробные и надежные сведения о происхождении, описывающие сложный процесс обработки компьютером/человеком	Поддержка работающих в полевых условиях пользователей посредством предоставления интерфейсов к смартфонам/планшетам и возможности скачивания данных в низком разрешении

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
		GeoServer, GDAL, а также поддерживающие стандарт метаданных GeoTIFF	метаданных GeoTIFF. 4. Совместимость с другими радиолокационными системами и хранилищами данных НАСА				
А.8.5 Вариант использования № 45: Объединенный испытательный стенд iRODS центра НАСА в Лэнгли и Центра космических полетов им. Годдарда	Федеративные распределенные неоднородные наборы данных	Облачная аналитика климата как сервис (CAaaS)	1. «Виртуальный сервер климатических данных» vCDS. 2. Файловая система GPFS, интегрированная с Hadoop. 3. iRODS	Визуализация распределенных разнородных данных	—	—	—
А.8.6 Вариант использования № 46: Аналитические сервисы MERRA	1. Интеграция результатов моделирования и данных наблюдений, файлы формата NetCDF. 2. Обработка в режиме реального времени и в пакетном режиме. 3. Интероперабельность между облачным решением AWS и локальными кластерами. 4. Управление данными с помощью iRODS	Облачная аналитика климата как сервис (CAaaS)	1. Программное обеспечение, способное работать с форматом NetCDF. 2. Map/Reduce. 3. Интероперабельное использование Amazon AWS и локальных кластеров	Высокопроизводительная распределенная визуализация	—	—	1. Требуется поддержка доступа со смартфонов и планшетов. 2. Управление данными посредством iRODS

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
А.8.7 Вариант использования № 47: Атмосферная турбулентность — Обнаружение событий	1. Распределенные наборы данных, полученные данные в реальном времени. 2. Различные форматы, разрешения, семантики и метаданные	1. Инструмент Map/Reduce или аналогичный; SciDB или другая научная СУБД. 2. Непрерывные вычисления по мере поступления новых данных. 3. Язык спецификации событий для интеллектуального анализа данных/поиска событий. 4. Интерпретации семантики, а также базы данных с оптимальной структурой для 4-мерного интеллектуального анализа данных и прогнозного анализа	1. Унаследованные вычислительные системы (например, суперкомпьютер). 2. Передача данных по сети с высокой пропускной способностью	Визуализация для помощи в интерпретации результатов	—	Валидация для выходных продуктов (корреляции)	—

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
А.8.8 Вариант использования № 48: Исследования климата с использованием модели климатической системы Земли (CESM)	1. Потоковая передача (до ~100 петабайт в 2017 году), при высокой скорости передачи данных от крупных суперкомпьютеров, расположенных по всему миру. 2. Интеграция крупномасштабных распределенных данных моделирования с результатами различных наблюдений. 3. Сопоставление разнообразных существующих данных с новыми данными моделирования в среде высокопроизводительных вычислений	Выполнение анализа данных вблизи места их хранения	Расширение архитектуры с целью охватить данные ряда других областей науки	1. Коллективное использование климатических данных в глобальном масштабе. 2. Высокопроизводительная распределенная визуализация	—	—	Ввод данных и доступ со смартфонов

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
А.8.9 Вариант использования № 49: Подповерхностные биогеохимические исследования	1. Разнородные разнообразные данные различных областей и разного масштаба, а также их перемещение по различным масштабам и областям. 2. Объединение разнообразных и разрозненных наборов данных полевых, лабораторных измерений, биологических наук и моделирования, охватывая различные семантические, пространственные и временные масштабы. 3. Сопоставление разнообразных существующих данных с новыми данными моделирования в среде высокопроизводительных вычислений	—	Postgres, HDF5 и различные специализированные программные системы	Доступ к данным и ввод данных со смартфона	—	—	Ввод данных и доступ со смартфонов

Продолжение таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
А.8.10 Вариант использования № 50: Сети AmeriFlux и FLUXNET	1. Разнородные разнообразные данные различных областей и разного масштаба, а также их перемещение по различным масштабам и областям. 2. Ссылки на многие другие экологические и биологические наборы данных. 3. Ссылки на данные моделирования климата и иные результаты моделирования в среде высокопроизводительных вычислений. Требуется поддерживать ссылки на европейские источники данных и проекты. Требуется поддерживать доступ к данным из 500 распределенных источников	Специализированное ПО, такое как EddyPro, и специальное ПО для анализа, такое как R, Python, нейронные сети, Matlab	1.Специализированное ПО, такое, как EddyPro; и специализированное ПО для анализа, такое, как R, Python, нейронные сети, Matlab. 2. Методы аналитики: интеллектуальный анализ данных, оценка качества данных, взаимная корреляция между наборами данных, ассимиляция данных, интерполяция данных, статистика, оценка качества, слияние данных и т. д.	Доступ к данным и ввод данных со смартфона	—	—	Ввод данных и доступ со смартфонов

Окончание таблицы С.1

Вариант использования	Источник данных	Преобразование данных	Возможности обработки	Потребитель данных	Безопасность и защита ПДн	Управление жизненным циклом	Иные технические проблемы
А.9.1 Вариант использования № 51: Прогнозирование потребления электроэнергии в интеллектуальных энергосетях	1. Разнообразные данные: показания датчиков интеллектуальной энергосети, данные городского планирования, метеорологические данные и служебные базы данных энергетических компаний. 2. Обновление данных каждые 15 минут	Новые виды аналитики на основе машинного обучения для прогнозирования энергопотребления	1. СУБД SQL, CSV-файлы, HDFS (платформа). 2. R/Matlab, Weka и Hadoop (платформа)	—	Защита персональных данных посредством анонимизации и агрегирования данных	—	Мобильный доступ для клиентов
А.9.2 Вариант использования № 52: Система управления энергией домашнего хозяйства HEMS	Источники данных распределены по отдельным частным домам	—	Большое централизованное хранилище (хранение)	Потребители услуг, предоставляемых HEMS, это, как правило, люди, проживающие в частных домах	Обработка персональных данных должна производиться ответственно и осмотрительно, с целью обеспечить неприкосновенность частной жизни пользователей	Данные будут полностью уничтожены, если пользователи расторгнут договор	Несколько игроков участвуют в цепочке поставок для потока больших данных

Приложение D
(справочное)

Детальное описание специфических для вариантов использования технических проблем

В данном приложении описаны специфические для конкретных вариантов использования технические проблемы, а также обобщенные технические проблемы в каждой из следующих семи категорий:

- источник данных (таблица D.1);
- преобразование данных (таблица D.2);
- возможности обработки (таблица D.3);
- потребитель данных (таблица D.4);
- безопасность и неприкосновенность частной жизни (таблица D.5);
- управление жизненным циклом (таблица D.6);
- иные технические проблемы (таблица D.7).

В каждой категории перечислены общие технические проблемы с указанием примеров использования, к которым применяется соответствующее требование.

Вслед за общими техническими проблемами перечислены специфические технические проблемы данной категории для каждого варианта использования. Если из описания конкретного варианта использования не удалось извлечь требований для определенной категории характеристик, соответствующий вариант не будет указан в данном пункте таблицы.

Таблица D.1 — Технические проблемы в категории «Источник данных»

Общие технические требования	
1. Необходимо поддерживать надежную, в реальном времени и/или асинхронную; потоковую и/или пакетную обработку с целью сбора данных из централизованных, распределенных и/или облачных источников данных, от датчиков и/или приборов	Применимо к 26 вариантам использования: № 1—2, 8, 10—12, 15—16, 18—25, 28, 39, 42—44, 46—50
2. Необходимо поддерживать передачу данных — медленную и/или неравномерную с периодическими пиковыми нагрузками и/или с высокой пропускной способностью — между источниками данных и вычислительными кластерами	Применимо к 22 вариантам использования: № 2—3, 8, 11—12, 16, 18—20, 22—23, 27, 36—38, 40—43, 48, 50—51
3. Необходимо поддерживать данные разнообразных типов и видов, включая структурированные и неструктурированные тексты, документы, графы, веб-материалы, геопространственные данные, сжатые, с привязкой ко времени, пространственные, мультимедийные данные, данные моделирования и показания измерительных инструментов	Применимо к 28 вариантам использования: № 1—2, 6, 8, 11, 13—18, 20—21, 23—25, 27—28, 30—31, 39, 41, 43, 47—51
Специфические для варианта применения технические требования к источнику данных	
1[1)]	А.1.1 Вариант использования № 1: Архивное хранение больших данных переписей населения в США 2010 и 2000 годов Требуется обеспечить долговременную сохранность больших объемов документов в центральном хранилище
2	А.1.2 Вариант использования № 2: Прием Национальными архивами США государственных данных на архивное хранение, поиск, извлечение и обеспечение долговременной сохранности Требуется поддержка распределенных источников данных. Требуется обеспечить хранение больших объемов данных. Требуется обеспечить обработку неравномерно поступающих данных, когда объем партии документов может варьироваться от гигабайт до сотен терабайт. Требуется поддерживать большое количество разнообразных форматов данных, в том числе для неструктурированных и для структурированных данных. Требуется поддержка распределенных источников данных в различных облачных решениях
3	А.1.3 Вариант использования № 3: Повышение активности респондентов в статистических обследованиях Требуется поддержка данных объемом примерно один петабайт

[1)] *Сохранена использованная в данной таблице нумерация с пропусками.*

Продолжение таблицы D.1

Общие технические требования	
5	A.2.1 Вариант использования № 5: Облачные вычисления в секторах финансовой отрасли Требуется поддержка ввода данных в реальном времени
6	A.2.2 Вариант использования № 6: Международная исследовательская сеть Mendeley Требуется поддержка представленных в виде файлов документов. В систему постоянно загружаются новые документы. Требуется поддержка различных типов файлов, таких как PDF-файлы, лог-файлы социальных сети и активности клиентов, изображения, электронные таблицы, файлы презентаций
7	A.2.3 Вариант использования № 7: Сервис кинофильмов Netflix Необходима поддержка профилей пользователей и рейтинговой информации.
8	A.2.4 Вариант использования № 8: Веб-поиск Требуется поддерживать распределенные источники данных. Требуется поддерживать потоковые данные. Требуется поддерживать мультимедийный контент
10	A.2.6 Вариант использования № 10: Грузоперевозки Требуется поддержка в реальном времени централизованных и распределенных источников информации/датчиков
11	A.2.7 Вариант использования № 11: Данные о материалах Требуется поддерживать распределенные хранилища данных о более чем 500 тысячах коммерческих материалов. Требуется поддерживать множество видов наборов данных. Требуется поддержка текста, графики и изображений
12	A.2.8 Вариант использования № 12: Геномика материалов на основе результатов моделирования Требуется поддерживать потоки данных от пета/экзафлопсных централизованных систем моделирования. Требуется поддерживать распределенные веб-потоки данных от центрального шлюза к пользователям
13	A.3.1 Вариант использования № 13: Облачный крупномасштабный анализ и визуализация геопространственных данных Нужна поддержка геопространственных данных, требующих уникальных подходов для индексирования и распределенного анализа
14	A.3.2 Вариант использования № 14: Идентификация и отслеживание объектов — Постоянное наблюдение Требуется поддерживать поступающие в реальном времени данные FMV-формата высококачественного видео (от 30 до 60 кадров в секунду при полноцветном разрешении 1080 пикселей) и WALF-формат видео с высоким разрешением (WALF) — от 1 до 10 кадров в секунду при полноцветном разрешении 10 тысяч * 10 тысяч пикселей
15	A.3.3 Вариант использования № 15: Обработка и анализ разведывательных данных Требуется поддерживать данные, поступающие в реальном времени с их обработкой (в худшем случае) в масштабе времени, близком к реальному. Требуется поддерживать данные, которые в настоящее время существуют в разрозненных хранилищах, и которые должны быть доступны через семантически интегрированное пространство данных. Требуется поддерживать разнообразные данные: текстовые файлы, первичные данные с датчиков, графические образы, видео, аудио, электронные данные и данные, созданные человеком

Продолжение таблицы D.1

	Общие технические требования
16	A.4.1 Вариант использования № 16: Данные электронной медицинской документации Требуется поддерживать неоднородные, большого объема, разнообразные источники данных. Требуется поддерживать данные на более чем 12 млн пациентов, содержащие более 4 млрд отдельных клинических наблюдений, суммарный объем которых превышает 20 терабайт первичных данных. Требуется поддерживать обработку данных, поступающих со скоростью от 500 тыс. до 1,5 млн новых клинических транзакций в день. Требуется поддерживать разнообразные данные: числовые и структурированные числовые данные, тексты в свободном формате, структурированные тексты, дискретные номинальные данные, дискретные порядковые данные, дискретные структурированные данные, большие двоичные объекты (изображения и видео). Требуется поддерживать данные, которые с течением времени эволюционируют. Требуется поддерживать во времени всестороннее и согласованное представление данных из разных источников
17	A.4.2 Вариант использования № 17: Анализ графических образов в патологии/Цифровая патология Требуется поддерживать пространственные цифровые графические образы высокого разрешения в патологии. Требуется поддерживать различные алгоритмы анализа качества изображений. Требуется поддерживать различные форматы графических данных, особенно BigTIFF, и результаты анализа, представленные в виде структурированных данных. Требуется поддерживать анализ изображений, пространственные запросы и аналитику, кластеризацию и классификацию признаков
18	A.4.3 Вариант использования № 18: Вычислительный анализ биоизображений Требуется поддерживать распределенные мультимодальные экспериментальные источники (инструменты) биологических изображений высокого разрешения. Требуется поддерживать 50 терабайт данных в различных форматах, включая графические.
19	A.4.4 Вариант использования № 19: Геномные измерения Требуется поддерживать поступающие с высокой скоростью сжатые данные (~300 гигабайт в день) от различных секвенсоров ДНК. Требуется поддерживать распределенные источники данных (секвенсоры). Требуется поддерживать различные файловые форматы, как для структурированных, так и для неструктурированных данных
20	A.4.5 Вариант использования № 20: Сравнительный анализ (мета) геномов Требуется поддерживать многочисленные централизованные источники данных. Требуется поддерживать разнообразные данные, от сведений о последовательностях аминокислот до данных о белках и их структурных особенностях (базовые геномные данные), а также новые типы данных таких направлений биологической науки — «омиков», как транскриптомика, метиломика и протеомика, описывающих экспрессии генов в различных условиях. Требуется поддерживать интерактивный пользовательский веб-интерфейс в реальном времени. Возможности обработки загружаемых данных на сервере должны соответствовать экспоненциальному росту объемов данных секвенирования из-за быстрого снижения стоимости технологии секвенирования. Требуется поддерживать разнородные, сложные, структурные и иерархические биологические данные. Требуется поддерживать метагеномные образцы, размеры, которые могут варьироваться на несколько порядков величины — от нескольких сотен тысяч до миллиарда генов
21	A.4.6 Вариант использования № 21: Индивидуальное управление лечением диабета Требуется поддерживать распределенные данные электронных медицинских документов. Требуется поддерживать данные более 5 миллионов пациентов с тысячами свойств по каждому, а также многие другие сведения, полученные из первичных данных. Требуется поддерживать данные по каждому пациенту, при этом число значений свойств может варьироваться от менее 100 до более чем 100 тысяч; типичным для пациента является около 100 значений свойств из контролируемых словарей и 1000 непрерывных числовых величин. Требуется поддерживать данные, которые периодически обновляются (не в режиме реального времени). Данные снабжаются отметками времени наблюдения (времени записи значения). Требуется поддерживать структурированные данные о пациентах. Данные делятся на две основные категории: данные со значениями свойств из контролируемого словаря и данные со значениями свойств, являющимися непрерывными числовыми величинами (которые документируются/регистрируются чаще). Требуется поддерживать данные, которые состоят из текста и непрерывных числовых значений

Продолжение таблицы D.1

Общие технические требования	
22	A.4.7 Вариант использования № 22: Статистический реляционный искусственный интеллект для здравоохранения Требуется поддержка централизованных данных, при этом некоторые данные извлекаются из интернет-источников. Требуется поддержка данных в диапазоне от сотен гигабайт для одной когорты из нескольких сотен человек и до одного петабайта в очень масштабных исследованиях, охватывающих миллионы пациентов. Требуется поддерживать как постоянно обновляемые/пополняемые данные о пациентах, так и данные, поступающие партиями по графику. Требуется поддерживать большие, мультимодальные данные длительного наблюдения. Требуется поддержка богатых реляционных данных, состоящих из многочисленных таблиц, а также различные типы данных, такие как изображения, электронные медицинские документы, демографические, генетические данные и данные на естественном языке, требующие богатых средств представления. Требуется поддерживать непредсказуемые темпы поступления данных, которые во многих случаях поступают в режиме реального времени
23	A.4.8 Вариант использования № 23: Эпидемиологическое исследование в масштабе всего населения Земли Требуется поддерживать синтетическую глобальную популяцию на централизованных либо распределенных ресурсах. Требуется поддерживать большие объемы выходных данных, поступающих в режиме реального времени. Требуется поддержка различных выходных наборов данных, в зависимости от сложности модели
24	A.4.9 Вариант использования № 24: Моделирование распространения социального влияния Требуется поддерживать динамическую распределенную обработку с использованием как традиционной архитектуры коммерческих кластеров, так и более новых архитектур (например, облачной). Требуется поддержка моделей с высокой детализацией и наборов данных, поддерживающих сетевой трафик Twitter. Требуется поддерживать хранение данных, поступающих каждый год в огромном объеме
25	A.4.10 Вариант использования № 25: Биологическое разнообразие и европейская электронная научно-исследовательская инфраструктура LifeWatch Требуется поддерживать специальные выделенные или оверлейные (наложенные) сенсорные сети. Требуется поддерживать распределенное хранение, в том числе архивирования и сохранение исторических данных и данных о тенденциях. Требуется поддерживать распределенные источники данных, в том числе многочисленные пункты наблюдения и мониторинга, сети датчиков и спутники. Требуется поддерживать широкий спектр данных, включая спутниковые изображения/информацию, данные о климате и погоде, фотографии, видео и звукозаписи и т. д. Требуется поддерживать комбинации данных различных типов и связи с потенциально неограниченными в своем разнообразии данными. Требуется поддерживать потоковую передачу данных
27	A.5.2 Вариант использования № 27: Организация крупномасштабных, неструктурированных коллекций сделанных потребителями фотографий Требуется поддерживать более 500 миллионов изображений, загружаемых ежедневно на сайты социальных сетей
28	A.5.3 Вариант использования № 28: Truthy — Анализ данных Твиттера Требуется поддерживать распределенные источники данных. Требуется поддерживать большие объемы данных и потоковую передачу в реальном времени. Требуется поддерживать первичные данные в сжатых форматах. Требуется поддерживать полностью структурированные данные в формате JSON, пользовательские метаданные и данные геолокации. Требуется поддерживать несколько схем данных
30	A.5.5 Вариант использования № 30: Цифровая инфраструктура для исследований и анализа сетей и графов (CINET) Требуется поддерживать набор файлов сетевых топологий для изучения теоретических свойств графов и поведения различных алгоритмов. Требуется поддерживать асинхронные и синхронные распределенные вычисления в реальном времени

Продолжение таблицы D.1

Общие технические требования	
31	A.5.6 Вариант использования № 31: Измерения и оценки эффективности аналитических технологий в Национальном институте стандартов и технологий (NIST) Требуется поддерживать большое количество частично аннотированных веб-страниц, твиттов, изображений и видеозаписей. Требуется поддерживать масштабирование процесса проверки на большие объемы данных, измерение внутренней неопределенности и неопределенности аннотаций, измерение эффективности для не полностью аннотированных данных, измерение эффективности аналитики для разнородных данных и аналитических потоков с участием пользователей
32	A.6.1 Вариант использования № 32: Консорциум федеративных сетей данных (DFC) Требуется поддерживать обработку ключевых файловых форматов: NetCDF, HDF5, Dicom. Треуется поддерживать обработку данных в режиме реального времени и пакетную обработку
33	A.6.2 Вариант использования № 33: Discinnet-процесс Требуется поддерживать интеграцию методов работы с метаданными различных дисциплин
34	A.6.3 Вариант использования № 34: Поиск по графу для научных данных Необходимо поддерживать любые типы данных, от изображений до текстов, от структур до белковых последовательностей
35	A.6.4 Вариант использования № 35: Анализ больших объемов данных, получаемых в экспериментах на синхротроне Требуется поддерживать многочисленные потоки данных в реальном времени, сохраняя данные для последующего анализа. Требуется поддерживать анализ в режиме реального времени выборок данных
36	A.7.1 Вариант использования № 36: Каталинский цифровой обзор неба в поисках транзиентов Необходимо поддерживать обработку поступающих за ночь ≈ 0.1 терабайта первичных данных обзора; в будущем темпы производства данных могут возрасти в 100 раз
37	A.7.2 Вариант использования № 37: Космологический обзор неба и моделирование Требуется поддерживать обработку ≈1 петабайта данных наблюдений в год. В будущем темпы производства данных вырастут до 7 петабайт в год
38	A.7.3 Вариант использования № 38: Большие данные космологических обзоров неба Требуется поддерживать обработку 20 терабайт данных в день
39	A.7.4 Вариант использования № 39: Анализ данных Большого адронного коллайдера Требуется поддерживать обработку данных, поступающих в реальном времени от ускорителей и инструментов анализа. Требуется поддерживать асинхронизацию сбора данных. Требуется поддерживать калибровку экспериментальных установок
40	A.7.5 Вариант использования № 40: Эксперимент Belle II Требуется поддерживать 120 петабайт первичных данных
41	A.8.1 Вариант использования № 41: Радарная система некогерентного рассеяния EISCAT-3D Требуется поддерживать систему из пяти постов, которая будет производить 40 петабайт данных в год в 2022 г. Требуется поддерживать формат данных HDF5. Требуется поддерживать визуализацию многомерных (≥5) данных
42	A.8.2 Вариант использования № 42: Совместная деятельность европейских сетевых инфраструктур в области экологических исследований (ENVRI) Требуется поддерживать огромный объем данных, поступающих в реальном времени из распределенных источников. Требуется поддерживать разнообразные наборы данных и метаданных, поступающих с измерительных инструментов

Окончание таблицы D.1

	Общие технические требования
43	A.8.3 Вариант использования № 43: Анализ радиолокационных данных для Центра дистанционного зондирования ледяного покрова CReSIS Требуется обеспечить надежную передачу данных с установленных на самолете датчиков/приборов либо со съемных жестких дисков, доставленных с удаленных объектов. Требуется поддерживать сбор данных в режиме реального времени. Требуется поддерживать различные наборы данных
44	A.8.4 Вариант использования № 44: Обработка данных проекта UAVSAR Требуется поддерживать пространственные данные и данные в угловых координатах. Требуется поддерживать совместимость с другими радиолокационными системами и хранилищами данных НАСА, например Спутникового центра НАСА на Аляске (Alaska Satellite Facility, ASF)
45	A.8.5 Вариант использования № 45: Объединенный испытательный стенд iRODS центра НАСА в Лэнгли и Центра космических полетов имени Годдарда Необходимо поддерживать федеративные распределенные неоднородные наборы данных
46	A.8.6 Вариант использования № 46: Аналитические сервисы MERRA Требуется поддерживать интеграцию результатов моделирования и данных наблюдений, файлы формата NetCDF. Требуется поддерживать обработку в режиме реального времени и в пакетном режиме. Требуется обеспечить интероперабельность между облачным решением AWS и локальными кластерами. Требуется поддерживать управление данными с помощью iRODS
47	A.8.7 Вариант использования № 47: Атмосферная турбулентность — Обнаружение событий Требуется поддержка распределенных наборов данных, получающих данные в реальном времени. Требуется поддерживать различные форматы, разрешения, семантику и метаданные
48	A.8.8 Вариант использования № 48: Исследования климата с использованием модели климатической системы Земли (CESM) Требуется поддерживать потоковую передачу (до 100 петабайт в 2017 г.), обеспечивая высокую скорость передачи данных от крупных суперкомпьютеров, расположенных по всему миру. Требуется поддерживать интеграцию крупномасштабных распределенных данных моделирования с результатами различных наблюдений. Требуется сопоставлять разнообразные существующие данные с новыми данными моделирования в среде высокопроизводительных вычислений
49	A.8.9 Вариант использования № 49: Подповерхностные биогеохимические исследования Требуется поддерживать разнородные разнообразные данные различных областей и разного масштаба, а также их перемещение по различным масштабам и областям. Требуется поддерживать объединение разнообразных и разрозненных наборов данных полевых, лабораторных измерений, медико-биологических наук и моделирования, охватывая различные семантические, пространственные и временные масштабы. Требуется сопоставлять разнообразные существующие данные с новыми данными моделирования в среде высокопроизводительных вычислений
50	A.8.10 Вариант использования № 50: Сети AmeriFlux и FLUXNET Требуется поддерживать разнородные разнообразные данные различных областей и разного масштаба, а также их перемещение по различным масштабам и областям. Требуется поддерживать ссылки на многие другие экологические и биологические наборы данных. Требуется поддерживать ссылки на данные моделирования климата и иные результаты моделирования в среде высокопроизводительных вычислений. Требуется поддерживать ссылки на европейские источники данных и проекты. Требуется поддерживать доступ к данным из 500 распределенных источников
51	A.9.1 Вариант использования № 51: Прогнозирование потребления электроэнергии в интеллектуальных энергосетях Требуется поддерживать разнообразные данные: показания датчиков интеллектуальной энергосети, данные городского планирования, метеорологические данные и служебные базы данных энергетических компаний. Требуется поддерживать обновление данных каждые 15 минут

Таблица D.2 — Технические проблемы в категории «Преобразование данных»

Общие технические требования	
1. Необходимо поддерживать разнообразные вычислительно-интенсивные методы аналитической обработки и методы машинного обучения	Применимо к 36 вариантам использования: № 2—4, 6—7, 10, 12—19, 21—25, 27—31, 36—39, 41—46, 48, 51
2. Необходимо поддерживать аналитическую обработку в реальном времени и/или пакетную	Применимо к 7 вариантам использования: № 7—8, 10, 20, 25, 41, 47
3. Необходимо поддерживать обработку большого объема разнородных данных и данных моделирования	Применимо к 14 вариантам использования: № 8, 11—13, 17, 19, 21, 23—24, 27, 30, 39, 43—44
4. Необходимо поддерживать обработку данных в движении (потоковая передача, доставка нового контента, отслеживание и т. д.)	Применимо к 6 вариантам использования: № 7—8, 10, 19, 39, 47
Специфические для варианта применения технические требования к поставщику услуг преобразования данных	
1	A.1.2 Вариант использования № 2: Прием Национальными архивами США государственных данных на архивное хранение, поиск, извлечение и обеспечение долговременной сохранности Требуется поддерживать сканирование и индексирование распределенных источников данных. Требуется поддерживать различные методы аналитической обработки, включая ранжирование, категоризацию данных и выявление персональных данных. Требуется поддерживать предварительную обработку данных. Требуется поддерживать управление обеспечением долговременной сохранности больших разнообразных наборов данных. Требуется поддерживать поиск по огромному количеству данных с высокой релевантностью и полнотой результатов
2	A.1.3 Вариант использования № 3: Повышение активности респондентов в статистических обследованиях Требуется поддерживать аналитику, которая необходима для рекомендательных систем, постоянного мониторинга и для общего совершенствования процесса проведения обследования
3	A.1.4 Вариант использования № 4: Использование нетрадиционных данных для повышения активности респондентов в статистических обследованиях Требуется поддерживать аналитику, позволяющую получать надежные оценки с использованием данных традиционных обследований, государственных административных данных и данных из нетрадиционных источников из сферы цифровой экономики
4	A.2.1 Вариант использования № 5: Облачные вычисления в секторах финансовой отрасли Требуется поддерживать аналитику в реальном времени
5	A.2.2 Вариант использования № 6: Международная исследовательская сеть Mendeley Требуется поддерживать стандартные библиотеки для проведения машинного обучения и аналитики. Требуется поддерживать эффективные масштабируемые и распараллеленные способы сопоставления документов, группировки похожих документов (включая те, что были слегка модифицированы инструментами аннотирования третьих сторон или же путем присоединения титульных страниц или наложения «водяных знаков» издателя)
6	A.2.3 Вариант использования № 7: Сервис кинофильмов Netflix Требуется поддерживать потоковое видео для многочисленных клиентов. Требуется поддерживать аналитическую обработку с целью подбора фильмов, соответствующих интересам клиента. Требуется поддерживать различные методы аналитической обработки с целью персонализации оказываемых клиенту услуг. Требуется поддерживать надежные алгоритмы обучения. Требуется поддерживать непрерывную аналитическую обработку на основе результатов мониторинга и оценки эффективности

Продолжение таблицы D.2

Общие технические требования	
7	A.2.4 Вариант использования № 8: Веб-поиск Требуется поддерживать динамическую доставку контента по сети. Необходимо обеспечить связывание профилей пользователей с данными из социальных сетей
8	A.2.5 Вариант использования № 9: Обеспечение непрерывности деловой деятельности и восстановления после катастроф для больших данных в облачной экосистеме Требуется поддерживать надежный алгоритм резервного копирования. Необходимо реплицировать последние изменения
9	A.2.6 Вариант использования № 10: Грузоперевозки Требуется поддерживать отслеживания объекта на основе уникальной идентификации с использованием закрепленного на объекте датчика и получаемых от глобальной системы позиционирования (GPS) координат. Требуется поддерживать обновление в реальном времени сведений об отслеживаемых объектах
10	A.2.7 Вариант использования № 11: Данные о материалах Требуется поддерживать описания свойств материалов, содержание сотни независимых переменных, и сбор значений этих переменных, с конечной целью создания надежных наборов данных
11	A.2.8 Вариант использования № 12: Геномика материалов на основе результатов моделирования Требуется поддерживать анализ данных в режиме реального времени с использованием вычислений с высокой пропускной способностью для оперативного реагирования. Требуется поддерживать комбинирование результатов моделирования, полученных с использованием различных программ. Требуется поддерживать поисковые исследования, ориентированные на потребности потребителей; вычислительная база должна гибко адаптироваться к новым целям. Требуется поддерживать технологии Map/Reduce и поиска, позволяющие комбинировать данные моделирования и экспериментальные данные
12	A.3.1 Вариант использования № 13: Облачный крупномасштабный анализ и визуализация геопространственных данных Требуется поддерживать методы аналитики, включая ближайшую точку подхода, отклонение от маршрута, плотность точек во времени, метод главных компонентов (PCA) и метод анализа независимых компонентов (ICA). Требуется поддерживать геопространственные данные, требующие уникальных подходов к индексации и проведению распределенного анализа
13	A.3.2 Вариант использования № 14: Идентификация и отслеживание объектов — Постоянное наблюдение Требуется поддерживать расширенную аналитику, включающую возможности для идентификации объекта, анализа закономерностей поведения объекта, анализа группового поведения/динамики и хозяйственной деятельности, а также для объединения (слияния) данных
14	A.3.3 Вариант использования № 15: Обработка и анализ разведывательных данных Требуется поддерживать аналитику, включая оповещения в масштабе времени, близком к реальному, основанные на закономерностях и изменениях основных параметров
15	A.4.1 Вариант использования № 16: Данные электронной медицинской документации Требуется поддерживать во времени всестороннее и согласованное представление данных из разных источников. Требуется поддерживать аналитических методы: методы извлечения информации с целью выявления соответствующих клинических признаков; обработка естественного языка; машинное обучение моделей принятия решений; методы оценки максимального правдоподобия и Байесовских сетей

Продолжение таблицы D.2

Общие технические требования	
16	A.4.2 Вариант использования № 17: Анализ графических образов в патологии / Цифровая патология Требуется поддерживать высокопроизводительный анализ изображений с целью извлечения пространственной информации. Требуется поддерживать пространственные запросы и аналитику, а также кластеризацию и классификацию признаков. Требуется поддерживать аналитическую обработку огромного многомерного набора данных, и обеспечивать возможность корреляции с данными других типов, такими, как клинические данные и данные других направлений биологической науки — «омиков»
17	A.4.3 Вариант использования № 18: Вычислительный анализ биоизображений Требуется поддерживать высокопроизводительные вычисления и управление анализом полученных результатов. Требуется поддерживать сегментацию представляющих интерес областей; групповой отбор и извлечение признаков, классификацию объектов, а также организацию и поиск. Требуется поддерживать расширенное выявление новых фактов и явлений, представляющих интерес для биологических наук, с помощью методов больших данных / экстремальных вычислений, обработки и анализа данных непосредственно в базе данных, машинного обучения (SVM и RF) для сервисов классификации и рекомендательных сервисов, продвинутых алгоритмов для массового анализа изображений и высокопроизводительных вычислительных решений. Требуется поддерживать массовый анализ данных применительно к масштабным наборам данных изображений
18	A.4.4 Вариант использования № 19: Геномные измерения Требуется поддерживать обработку первичных данных с целью выделения вариаций. Требуется поддерживать машинное обучение для комплексного анализа систематических ошибок технологий секвенирования, которые сложно охарактеризовать
19	A.4.5 Вариант использования № 20: Сравнительный анализ (мета) геномов Требуется поддерживать методы сравнительного анализа очень сложных данных. Требуется поддерживать описательную статистику
20	A.4.6 Вариант использования № 21: Индивидуальное управление лечением диабета Требуется поддерживать интеграцию данных с использованием аннотаций на основе онтологий и таксономий. Требуется поддерживать алгоритмы параллельного поиска и извлечения как для поиска по индексу, так и для настраиваемого поиска; а также способность выделять представляющие интерес данные. Потенциальные результаты включают когорты пациентов, группы пациентов, удовлетворяющих определенным критериям, и группы пациентов, имеющих сходные характеристики. Требуется поддерживать алгоритмы распределенного интеллектуального анализа закономерностей в графе, анализа закономерностей и индексации графов, а также поиска закономерностей в графах на основе триплетов RDF. Требуется поддерживать надежные инструменты статистического анализа для контроля частоты ложных срабатываний, определения истинной значимости подграфа и исключения ложных позитивных и ложных негативных результатов. Требуется поддерживать алгоритмы интеллектуального анализа закономерностей в графах с целью выявления закономерностей в графах, их индексации и поиска по графам. Требуется поддерживать обход семантического графа
21	A.4.7 Вариант использования № 22: Статистический реляционный искусственный интеллект для здравоохранения Требуется поддерживать реляционные вероятностные модели, моделирующие неопределенности на основе теории вероятности. Программное обеспечение обучает модели на основе ряда типов данных, и, возможно, сможет интегрировать информацию и логические рассуждения о сложных запросах. Требуется поддерживать надежных и точных методов обучения для учета дисбаланса данных, то есть ситуаций, в которых большие объемы данных доступны для небольшого числа субъектов. Требуется поддерживать алгоритмы обучения для определения перекосов в данных, чтобы избежать ошибочного моделирования «шума». Требуется поддерживать обученные модели, которые могут быть обобщены и уточнены для применения к другим наборам данных. Требуется поддерживать принятие данных в разных формах и из разрозненных источников

Продолжение таблицы D.2

Общие технические требования	
22	A.4.8 Вариант использования № 23: Эпидемиологическое исследование в масштабе всего населения Земли Требуется поддерживать вычисления, требующие как значительных вычислительных ресурсов, так и обработки больших объемов данных, что больше всего соответствует характеристикам суперкомпьютеров. Требуется поддерживать алгоритмы, учитывающие неструктурированный и нерегулярный характер обработки графов. Требуется поддерживать получение сводок по различным прогонам и повторам моделирования
23	A.4.9 Вариант использования № 24: Моделирование распространения социального влияния Требуется поддерживать крупномасштабное моделирование различных событий (болезни, эмоции, поведение и т. д.). Требуется поддерживать масштабируемое объединение наборов данных. Требуется поддерживать многоуровневый анализ, одновременно обеспечивая быстрое получение достаточных результатов
24	A.4.10 Вариант использования № 25 Биологическое разнообразие и европейская электронная научно-исследовательская инфраструктура LifeWatch Требуется поддерживать поэтапный анализ и/или анализ данных в реальном времени; темпы поступления данных варьируются в зависимости от исходных биологических и экологических процессов. Требуется поддерживать разнообразие данных, аналитических инструментов и инструментов моделирования для поддержки аналитики в интересах различных научных сообществ. Требуется поддерживать аналитику параллельных потоков данных и аналитику данных, поступающих в потоковом режиме. Требуется поддерживать доступ и интеграцию нескольких распределенных баз данных
25	A.5.2 Вариант использования № 27: Организация крупномасштабных, неструктурированных коллекций сделанных потребителями фотографий Требуется поддерживать классификатор (например, SVM) — процесс, который часто трудно распараллелить. Требуется поддерживать функциональные возможности, применяемые во многих крупномасштабных задачах обработки изображений
26	A.5.3 Вариант использования № 28: Truthy — Анализ данных Твиттера Требуется поддерживать различные методы анализа данных в реальном времени для выявления аномалий, кластеризации потока, классификации сигналов на основе многомерных временных рядов и онлайн-обучения
27	A.5.4 Вариант использования № 29: Краудсорсинг в гуманитарных науках Требуется поддерживать оцифровку существующих архивов документов и аудио-, видео- и фотоматериалов. Требуется поддерживать аналитику, включая все виды распознавания закономерностей (например, распознавание речи, автоматический анализ аудиовизуальных материалов, культурные закономерности) и выявления структур (лексические единицы, лингвистические правила и т. д.)
28	A.5.5 Вариант использования № 30: Цифровая инфраструктура для исследований и анализа сетей и графов» (CINET) Требуется поддерживать среды для запуска различных инструментов анализа сетей и графов. Требуется поддерживать динамический рост сетей. Требуется поддерживать асинхронные и синхронные, выполняемые в реальном времени распределенные вычисления. Требуется поддерживать различные параллельные алгоритмы для разных схем разделения, используемых для повышения эффективности вычислений
29	A.5.6 Вариант использования № 31: Измерения и оценки эффективности аналитических технологий в Национальном институте стандартов и технологий (NIST) Требуется поддерживать аналитические алгоритмы работающих с письменным языком, речью, изображениями людей и т. д. Алгоритмы, как правило, следует тестировать на реальных или реалистичных данных. Крайне проблематично создание искусственных данных, которые бы в достаточной степени отражали вариативность реальных данных, связанных с людьми
30	A.6.1 Вариант использования № 32: Консорциум федеративных сетей данных (DFC) Необходимо обеспечить типовые потоки рабочих процессов аналитики

Продолжение таблицы D.2

Общие технические требования	
31	A.6.3 Вариант использования № 34: Поиск по графу для научных данных Требуется поддерживать обработку графа данных. Требуется поддерживать реляционную СУБД
32	A.6.4 Вариант использования № 35: Анализ больших объемов данных, получаемых в экспериментах на синхротроне Требуется поддерживать стандартные инструменты биоинформатики (BLAST, HMMER, инструменты множественного выравнивания последовательностей и филогенетики, программы поиска/предсказания генов и генных структур (gene callers), программы предсказания свойств по результатам секвенирования (sequence feature predictors) и т. д.), скрипты Perl / Python и планировщик задач Linux-кластера
33	A.7.1 Вариант использования № 36: Каталинский цифровой обзор неба в поисках транзиентов Требуется поддерживать большое количество разнообразных инструментов анализа астрономических данных, а также большое количество специализированных инструментов и программного обеспечения, часть которых является самостоятельными исследовательскими проектами. Требуется поддерживать автоматизированную классификацию с помощью инструментов машинного обучения, учитывающую немногочисленность и разнородность данных, которая динамически эволюционирует во времени по мере поступления большего количества данных; и принятия решений о проведении дополнительных исследований в условиях немногочисленности и ограниченности выделяемых для этого ресурсов
34	A.7.2 Вариант использования № 37: Космологический обзор неба и моделирование Требуется поддерживать интерпретацию результатов детального моделирования, которая требует развитых методов и средств анализа и визуализации
35	A.7.3 Вариант использования № 38: Большие данные космологических обзоров неба Требуется одновременно поддерживать анализ результатов моделирования и данных наблюдений. Требуется поддерживать методы для выполнения разложения Холецкого для тысяч моделирований с матрицами порядка миллиона по каждой стороне
36	A.7.4 Вариант использования № 39: Анализ данных Большого адронного коллайдера Требуется поддерживать экспериментальные данные проектов ALICE, ATLAS, CMS и LHC. Требуется поддерживать гистограммы, диаграммы рассеяния, подбор моделей. Требуется поддерживать вычисления по методу Монте-Карло
37	A.8.1 Вариант использования № 41: Радарная система некогерентного рассеяния EISCAT-3D Требуется поддерживать архитектуру «пчелиной матки» (Queen Bee), в рамках которой централизованная обработка сочетается с распределенной обработкой на измерительных устройствах для данных с 5 распределенных постов. Требуется поддерживать мониторинг оборудования в режиме реального времени путем частичного анализа потока данных. Требуется поддерживать богатый набор сервисов обработки радиолокационных изображений с использованием машинного обучения, статистического моделирования и алгоритмов поиска на графе
38	A.8.2 Вариант использования № 42: Совместная деятельность европейских сетевых инфраструктур в области экологических исследований (ENVRI) Требуется поддерживать разнообразные аналитические инструменты.
39	A.8.3 Вариант использования № 43: Центра дистанционного зондирования ледяного покрова CReSIS Требуется поддерживать унаследованное программное обеспечение (Matlab) и языки (C/Java) для обработки данных. Требуется поддерживать обработку сигналов и методы обработки изображений с целью выделения слоев
40	A.8.4 Вариант использования № 44: Обработка данных проекта UAVSAR Требуется поддержка данных с географической привязкой, которые требуют интеграции данных в ГИС в качестве дополнительных наложений (оверлеев). Требуется поддерживать значительное вмешательство человека в конвейер обработки данных. Необходимо обеспечить богатый набор сервисов обработки радиолокационных изображений. Требуется поддерживать инструменты ROI_PAC, GeoServer, GDAL, а также инструменты, поддерживающие стандарт метаданных GeoTIFF
41	A.8.5 Вариант использования № 45: Объединенный испытательный стенд iRODS центра НАСА в Лэнгли и Центра космических полетов имени Годдарда Требуется поддерживать облачную аналитику климата как сервис (CAaaS)

Окончание таблицы D.2

Общие технические требования	
42	A.8.6 Вариант использования № 46: Аналитические сервисы MERRA Требуется поддерживать облачную аналитику климата как сервис (CAaaS)
43	A.8.7 Вариант использования № 47: Атмосферная турбулентность — Обнаружение событий Требуется поддерживать инструмент Map/Reduce или аналогичный; SciDB или другую научную СУБД. Требуется поддерживать непрерывные вычисления по мере поступления новых данных. Требуется поддерживать язык спецификации событий для интеллектуального анализа данных/поиска событий. Требуется поддерживать интерпретации семантики, а также базы данных с оптимальной структурой для четырехмерного интеллектуального анализа данных и прогнозного анализа
44	A.8.8 Вариант использования № 48: Исследования климата с использованием модели климатической системы Земли (CESM) Требуется поддерживать выполнение анализа данных вблизи места их хранения
45	A.8.10 Вариант использования № 50: Сети AmeriFlux и FLUXNET Требуется поддерживать специализированное программное обеспечение, такое как EddyPro, и специальное программное обеспечение для анализа, такого как R, Python, нейронные сети, Matlab
46	A.9.1 Вариант использования № 51: Прогнозирование потребления электроэнергии в интеллектуальных энергосетях Требуется поддерживать новые виды аналитики на основе машинного обучения для прогнозирования энергопотребления

Таблица D.3 — Технические проблемы в категории «Возможности обработки»

Общие технические требования	
1. Необходимо поддерживать как унаследованные, так и продвинутые пакеты программ (субкомпонент: SaaS)	Применимо к 27 вариантам использования: № 3, 6—7, 12—26, 28, 30, 38—40, 43—44, 49, 51
2. Необходимо поддерживать как унаследованные, так и продвинутые вычислительные платформы (субкомпонент: PaaS)	Применимо к 16 вариантам использования: № 6—7, 16—20, 23—24, 27—28, 30, 38, 44—45, 51
3. Необходимо поддерживать как унаследованные, так и продвинутые распределенные вычислительные кластеры, сопроцессоры, обработку ввода-вывода (субкомпонент: IaaS)	Применимо к 23 вариантам использования: № 6 — 7, 12, 14 — 19, 21 — 26, 30, 37, 39, 41, 43, 46 — 48
4. Необходимо поддерживать гибкую передачу данных (субкомпонент: сети)	Применимо к 14 вариантам использования: № 10, 12, 14—15, 17—18, 23—26, 28, 30, 40, 47
5. Необходимо поддерживать унаследованные, крупномасштабные и продвинутые распределенные хранилища данных (субкомпонент: хранение)	Применимо к 28 вариантам использования: № 1—3, 6—8, 12, 15, 17, 19—24, 27—28, 30, 36—45
6. Необходимо поддерживать как унаследованные, так и продвинутые исполняемые программы: приложения, инструменты, утилиты и библиотеки (субкомпонент: программное обеспечение)	Применимо к 13 вариантам использования: № 7, 12, 14—15, 17, 19, 21—22, 31, 37, 39, 43, 50
Специфические для варианта применения технические требования к поставщику вычислительных возможностей	
1	A.1.1 Вариант использования № 1: Архивное хранение больших данных переписи населения в США 2010 и 2000 годов Требуется поддерживать большое централизованное хранилище
2	A.1.2 Вариант использования № 2: Прием Национальными архивами США государственных данных на архивное хранение, поиск, извлечение и обеспечение долговременной сохранности Требуется поддерживать большое хранилище данных. Требуется поддерживать различные системы хранения, такие как NetApps, Hitachi и магнитные ленты

Продолжение таблицы D.3

Общие технические требования	
3	A.1.3 Вариант использования № 3: Повышение активности респондентов в статистических обследованиях Требуется поддерживать следующее программное обеспечение: Hadoop, Spark, Hive, R, SAS, Mahout, Allegrograph, MySQL, Oracle, Storm, BigMemory, Cassandra и Pig
4	A.1.4 Вариант использования № 4: Использование нетрадиционных данных для повышения активности респондентов в статистических обследованиях Требуется поддерживать следующее программное обеспечение: Hadoop, Spark, Hive, R, SAS, Mahout, Allegrograph, MySQL, Oracle, Storm, BigMemory, Cassandra и Pig
5	A.2.2 Вариант использования № 6: Международная исследовательская сеть Mendeley Требуется поддерживать Amazon EC2 с HDFS (инфраструктура). Требуется поддерживать S3 (хранение). Требуется поддерживать Hadoop (платформа). Требуется поддерживать Scribe, Hive, Mahout и Python (язык). Требуется поддерживать хранилище умеренного объема (15 терабайт, с приростом 1 терабайт в месяц). Требуется поддерживать пакетную обработку и обработку в реальном времени
6	A.2.3 Вариант использования № 7: Сервис кинофильмов Netflix Требуется поддерживать Hadoop (платформа). Требуется поддерживать Pig (язык). Требуется поддерживать Cassandra и Hive. Требуется поддерживать огромный объем подписчиков, рейтингов и поисков в сутки (база данных). Требуется поддерживать огромное хранилище (2 петабайта). Требуется поддерживать обработку с интенсивным вводом-выводом
7	A.2.4 Вариант использования № 8: Веб-поиск Требуется поддерживать петабайты текстовых и мультимедийных данных (хранение)
8	A.2.5 Вариант использования № 9: Обеспечение непрерывности деловой деятельности и восстановления после катастроф для больших данных в облачной экосистеме Требуется поддерживать Hadoop. Требуется поддерживать использование коммерческих облачных сервисов
9	A.2.6 Вариант использования № 10: Грузоперевозки Требуется поддерживать подключение к Интернету
10	A.2.8 Вариант использования № 12: Геномика материалов на основе результатов моделирования Требуется поддерживать массивной (суперкомпьютер Cray XE6 «Hopper», 150 тысяч процессоров) унаследованной инфраструктуры (инфраструктура). Требуется поддерживать GPFS (хранение). Требуется поддерживать систему MonogDB (платформа). Требуется поддерживать сетевое подключение 10 гигабит/с. Требуется поддерживать различные аналитические инструменты, такие как PyMatGen, FireWorks, VASP, ABINIT, NWChem, BerkeleyGW и различное ПО, разработанное сообществом. Требуется поддерживать большое хранилище (хранение). Требуется поддерживать масштабируемые базы данных для данных типа «ключ-значение» и для библиотек объектов (платформа). Требуется поддерживать потоки данных моделирования из централизованных пета/экзафлопсных вычислительных систем
11	A.3.1 Вариант использования № 13: Облачный крупномасштабный анализ и визуализация геопространственных данных Требутся поддерживать реляционную СУБД с геопространственной поддержкой; а также геопространственный сервер / программное обеспечение для анализа — ESRI ArcServer, Geoserver
12	A.3.2 Вариант использования № 14: Идентификация и отслеживание объектов — Постоянное наблюдение Требуется поддерживать широкий спектр специализированного программного обеспечения и инструментов, включая традиционные реляционные СУБД и средства отображения. Требуется поддерживать несколько каналов сетевого взаимодействия. Требуется поддерживать кластеры расширенных за счет использования графических процессоров (GPU) компьютерных систем

Продолжение таблицы D.3

Общие технические требования	
13	A.3.3 Вариант использования № 15: Обработка и анализ разведывательных данных Требуется обеспечивать стабильность и жизнеспособность системы в случае ненадежной связи с солдатами и удаленными датчиками. Требуется поддерживать объемы данных до сотен петабайт, хранимые средними и крупными кластерами и облачными системами. Требуется поддерживать следующее программное обеспечение: Hadoop, Accumulo (с системой хранения данных BigTable), Solr, NLP (несколько вариантов), Puppet (управление жизненным циклом ИТ, обеспечение безопасности), Storm, а также специализированные приложения и инструменты визуализации
14	A.4.1 Вариант использования № 16: Данные электронной медицинской документации Требуется поддерживать Hadoop, Hive и R на основе Unix. Требуется поддерживать суперкомпьютер Cray. Требуется поддерживать Teradata, PostgreSQL, MongoDB. Требуется поддерживать различные сетевые возможности с учетом значительных объемов обработки с интенсивным вводом-выводом
15	A.4.2 Вариант использования № 17: Анализ графических образов в патологии / Цифровая патология Требуется поддержка унаследованных систем и облачных решений (вычислительный кластер). Требуется поддерживать огромные объемы данных в унаследованных и новых системах хранения, таких как SAN и HDFS (хранение). Требуется поддерживать сетевые соединения с высокой пропускной способностью (сети). Требуется поддерживать анализ изображений с использованием MPI, Map/Reduce и Hive с пространственным расширением (пакеты программ)
16	A.4.3 Вариант использования № 18: Вычислительный анализ биоизображений Требуется поддерживать ImageJ, OMERO, VolRover, разработанные прикладными математиками продвинутые методы сегментации и выявления признаков. Необходимы масштабируемые базы данных для данных типа «ключ-значение» и для библиотек объектов. Требуется поддерживать инфраструктуру суперкомпьютера Hopper в Национальном научно-исследовательском вычислительном центре энергетических исследований Министерства энергетики США (NERSC). Требуется поддерживать базы данных и коллекций изображений. Требуется поддерживать 10-гигабитные, в будущем 100-гигабитные сети и расширенные сетевые возможности (SDN)
17	A.4.4 Вариант использования № 19: Геномные измерения Требуется поддерживать унаследованный вычислительный кластер и другие PaaS и IaaS-решения (вычислительный кластер). Требуется поддерживать огромное хранилище данных петабайтного масштаба (хранение). Требуется поддерживать унаследованное программное обеспечение с открытым исходным кодом для секвенирования в биоинформатике на основе UNIX (пакет программ)
18	A.4.5 Вариант использования № 20: Сравнительный анализ (мета) геномов Требуется поддерживать огромное хранилище данных. Требуется поддерживать масштабируемую реляционную СУБД для разнородных биологических данных. Требуется поддерживать быструю и параллельную массовую загрузку в реальном времени. Требуется поддерживать реляционную СУБД Oracle, файлы SQLite, плоские текстовые файлы, Lucy (версия Lucene) для поиска по ключевым словам, базы данных BLAST, базы данных USEARCH. Требуется поддерживать Linux-кластер, сервер реляционной СУБД Oracle, большие системы хранения данных, стандартные интерактивные хосты Linux
19	A.4.6 Вариант использования № 21: Индивидуальное управление лечением диабета Требуется поддерживать хранилища данных, в частности нереляционную СУБД Hbase с открытым исходным кодом. Требуется поддерживать использование суперкомпьютеров в рамках облачных и параллельных вычислений. Требуется поддерживать обработку с интенсивным вводом-выводом. Требуется поддерживать распределенную файловую систему HDFS. Требуется поддерживать специализированное программное обеспечение для выявления новых признаков на основе хранимых данных

Продолжение таблицы D.3

	Общие технические требования
20	A.4.7 Вариант использования № 22: Статистический реляционный искусственный интеллект для здравоохранения Требуется поддерживать Java, некоторые инструменты собственной разработки, реляционную базу данных и хранилища NoSQL. Требуется поддерживать облачные и параллельные вычисления. Требуется поддерживать высокопроизводительный компьютер с 48 гигабайт ОЗУ (для анализа при умеренном размере выборки). Требуется поддерживать вычислительные кластеры для обработки больших наборов данных. Требуется поддерживать жесткий диск объемом от 200 гигабайт до 1 терабайта для тестовых данных
21	A.4.8 Вариант использования № 23: Эпидемиологическое исследование в масштабе всего населения Земли Требуется поддерживать перемещение очень больших объемов данных для визуализации (сети). Требуется поддерживать распределенную систему моделирования на основе MPI (платформа). Требуется поддерживать Charm++ на нескольких узлах (программное обеспечение). Требуется поддерживать сетевую файловую систему (хранение). Требуется поддерживать сеть Infiniband (сети)
22	A.4.9 Вариант использования № 24: Моделирование распространения социального влияния Требуется поддерживать вычислительную инфраструктуру, позволяющую моделировать различные типы взаимодействия между людьми через интернет в связи с различными социальными событиями (инфраструктура). Требуется поддерживать файловые серверы и базы данных (платформа). Требуется поддерживать сети Ethernet и Infiniband (сети). Требуется поддерживать специализированные программы моделирования, программное обеспечение с открытым исходным кодом и проприетарные среды моделирования (приложения). Требуется поддерживать обработку огромного количества учетных записей пользователей социальных сетей из различных стран (сети)
23	A.4.10 Вариант использования № 25 Биологическое разнообразие и европейская электронная научно-исследовательская инфраструктура LifeWatch Требуется поддерживать расширяемые и предоставляемые по требованию ресурсы хранения для глобальных пользователей. Требуется поддерживать облачные ресурсы сообщества
24	A.5.1 Вариант использования № 26: Крупномасштабное глубокое обучение Требуется поддерживать использование графических процессоров. Требуется поддерживать высокопроизводительный кластер с внутренними соединениями на основе MPI и Infiniband. Требуется поддерживать библиотеки для вычислений на одной машине или на одном графическом процессоре (например, BLAS, CuBLAS, MAGMA и др.). Требуется поддерживать распределенные вычисления с плотными матрицами на графических процессорах, подобно BLAS или LAPACK, которые остаются слабо развитыми. Существующие решения (например, ScaLapack для центральных процессоров) не очень хорошо интегрированы с языками высокого уровня и требуют низкоуровневого программирования, что удлиняет время эксперимента и процесса разработки
25	A.5.2 Вариант использования № 27: Организация крупномасштабных, неструктурированных коллекций сделанных потребителями фотографий Требуется поддерживать Hadoop или усовершенствованный Map/Reduce
26	A.5.3 Вариант использования № 28: Truthy — Анализ данных Твиттера Требуется поддерживать Hadoop и HDFS (платформа). Требуется поддерживать IndexedHBase, Hive, SciPy и NumPy (программное обеспечение). Требуется поддерживать базы данных в памяти и MPI (платформа). Требуется поддерживать высокоскоростную сеть Infiniband (сети)

Продолжение таблицы D.3

Общие технические требования	
27	A.5.5 Вариант использования № 30: Цифровая инфраструктура для исследований и анализа сетей и графов» (CINET) Требуется поддерживать высокопроизводительную кластеризованную файловую систему (хранение). Требуется поддерживать различные сетевые подключения (сети). Требуется поддерживать существующий вычислительный кластер. Требуется поддерживать вычислительный кластер Amazon EC2. Требуется поддерживать различные библиотеки для работы с графами, инструменты управления потоками процессов, СУБД и семантические веб-инструменты
28	A.5.6 Вариант использования № 31: Измерения и оценки эффективности аналитических технологий в Национальном институте стандартов и технологий (NIST) Требуется поддерживать средства разработки PERL, Python, C/C++, Matlab, R. Требуется поддерживать разработку по принципу «снизу вверх» тестовых и измерительных приложений
29	A.6.1 Вариант использования № 32: Консорциум федеративных сетей данных (DFC) Требуется поддерживать программное обеспечение для управления данными iRODS. Требуется поддерживать интероперабельность между различными типами протоколов хранения и сетевого взаимодействия
30	A.6.2 Вариант использования № 33: Discinnet-процесс Требуется поддерживать следующее программное обеспечение: Symfony-PHP, Linux и MySQL
31	A.6.3 Вариант использования № 34: Поиск по графу для научных данных Требуется поддерживать облачные ресурсы сообщества
32	A.6.4 Вариант использования № 35: Анализ больших объемов данных, получаемых в экспериментах на синхротроне Требуется поддерживать передачу больших объемов данных на удаленный ресурс для пакетной обработки
33	A.7.2 Вариант использования № 37: Космологический обзор неба и моделирование Требуется поддерживать программное обеспечение MPI, OpenMP, C, C++, F90, FFTW, пакеты визуализации, Python, FFTW, Numpy, Boost, OpenMP, ScaLAPCK, СУБД PSQL и MySQL, Eigen, Cfitsio, http://astrometry.net/ и Minuit2. Требуется поддерживать разработку новых методов анализа ввиду ограничений подсистемы ввода/вывода суперкомпьютера
34	A.7.3 Вариант использования № 38: Большие данные космологических обзоров неба Требуется поддерживать стандартное астрофизическое программное обеспечение для обработки («редуцирования») данных, а также сценарии-обертки Perl / Python. Требуется поддерживать реляционную СУБД Oracle, терминальный клиент psql (PostgreSQL interactive terminal) для работы с объектно-реляционной СУБД PostgreSQL, а также файловые системы GPFS и Luster и ленточные архивы. Требуется поддерживать параллельные базы данных для хранения изображений
35	A.7.4 Вариант использования № 39: Анализ данных Большого адронного коллайдера Требуется поддерживать унаследованную вычислительную инфраструктуру (вычислительные узлы). Требуется поддерживать распределенное хранение файлов (хранение). Требуется поддерживать объектно-ориентированные базы данных (программное обеспечение)
36	A.7.5 Вариант использования № 40: Эксперимент Belle II Требуется поддерживать хранение 120 петабайт первичных данных. Требуется поддерживать модель международных распределенных вычислений, для расширения имеющихся возможностей на ускорителе в Японии. Требуется поддерживать передачу первичных данных со скоростью ~20 гигабит/с между Японией и США (при проектной яркости ускорителя). Требуется поддерживать программное обеспечение: «Грид Открытой науки» (Open Science Grid), Geant4, DIRAC, FTS, инфраструктуру Belle II
37	A.8.1 Вариант использования № 41: Радарная система некогерентного рассеяния EISCAT-3D Требуется поддерживать архитектуру, позволяющую принимать участие в сотрудничестве в рамках проекта ENVRI

Окончание таблицы D.3

Общие технические требования	
38	A.8.2 Вариант использования № 42: Совместная деятельность европейских сетевых инфраструктур в области экологических исследований (ENVRI) Требуется поддерживать взаимодействие с различными вычислительными инфраструктурами и архитектурами (инфраструктура). Требуется поддерживать взаимодействие с разрозненными хранилищами (хранение)
39	A.8.3 Вариант использования № 43: Центра дистанционного зондирования ледяного покрова CReSIS Требуется поддерживать хранение необработанных данных, объемы которых увеличиваются на ≈0,5 петабайт в год. Требуется поддерживать передачу материалов со съемного жесткого диска в вычислительный кластер для параллельной обработки. Требуется поддерживать Map/Reduce или MPI, плюс C/Java
40	A.8.4 Вариант использования № 44: Обработка данных проекта UAVSAR Требуется поддерживать архитектуру, обеспечивающую интероперабельность системы высокопроизводительных вычислений с облачными решениями. Требуется поддерживать богатый набор сервисов обработки радиолокационных изображений. Требуется поддерживать инструменты ROI_PAC, GeoServer, GDAL, а также инструменты, поддерживающие стандарт метаданных GeoTIFF. Требуется поддерживать совместимость с другими радиолокационными системами и хранилищами данных НАСА, например, Спутникового центра НАСА на Аляске (Alaska Satellite Facility, ASF)
41	A.8.5 Вариант использования № 45: Объединенный испытательный стенд iRODS центра НАСА в Лэнгли и Центра космических полетов имени Годдарда Требуется поддерживать «виртуальный сервер климатических данных» vCDS. Требуется поддерживать файловую систему GPFS интегрированную с Hadoop. Требуется поддерживать iRODS
42	A.8.6 Вариант использования № 46: Аналитические сервисы MERRA Требуется поддерживать программное обеспечение, способное работать с форматом NetCDF. Требуется поддерживать Map/Reduce. Требуется поддерживать интероперабельное использование Amazon AWS и локальных кластеров
43	A.8.7 Вариант использования № 47: Атмосферная турбулентность — Обнаружение событий Требуется поддерживать другие унаследованные вычислительные системы (например, суперкомпьютер). Требуется поддерживать передачу данных по сети с высокой пропускной способностью
44	A.8.8 Вариант использования № 48: Исследования климата с использованием модели климатической системы Земли (CESM) Требуется поддерживать расширение архитектуры с тем, чтобы охватить данные ряда других областей науки
45	A.8.9 Вариант использования № 49: Подповерхностные биогеохимические исследования Требуется поддерживать Postgres, HDF5 и различные специализированные программные системы
46	A.8.10 Вариант использования № 50: Сети AmeriFlux и FLUXNET Требуется поддерживать специализированное программное обеспечение, такое как EddyPro; и специализированное аналитическое программное обеспечение, такое как R, Python, нейронные сети, Matlab. Требуется поддерживать методы аналитики: интеллектуальный анализ данных, оценка качества данных, взаимная корреляция между наборами данных, ассимиляция данных, интерполяция данных, статистика, оценка качества, слияние данных и т. д.
47	A.9.1 Вариант использования № 51: Прогнозирование потребления электроэнергии в интеллектуальных энергосетях Требуется поддерживать СУБД SQL, CSV-файлы, HDFS (платформа). Требуется поддерживать R/Matlab, Weka и Hadoop (платформа)

Таблица D.4 — Технические проблемы в категории «Потребитель данных»

Общие технические требования	
1. Необходимо поддерживать быстрый поиск по обработанным данным, с высокой релевантностью, точностью и полнотой результатов поиска	Применимо к четырем вариантам использования: № 2, 8, 12, 28
2. Необходимо поддерживать различные форматы выходных файлов для визуализации, рендеринга и создания отчетов	Применимо к 13 вариантам использования: № 6—8, 13—14, 16—17, 19, 22, 39, 42, 43, 47
3. Необходимо поддерживать визуальную разметку для представления результатов	Применимо к двум вариантам использования: № 8, 43
4. Необходимо поддерживать пользовательский интерфейс с широкими функциональными возможностями для доступа с помощью браузера, средства визуализации	Применимо к девяти вариантам использования: № 11, 20, 28, 31, 42—44, 49—50
5. Необходимо поддерживать инструменты многомерной, с высоким разрешением визуализации данных	Применимо к 20 вариантам использования: № 3—4, 6, 11, 13—14, 16, 18, 20, 23—24, 27, 12, 30, 37, 41, 45—46, 48, 15[1)]
6. Необходимо поддерживать потоковую передачу результатов клиентам	Применимо к одному варианту использования: № 7

Специфические для варианта применения технические требования к потребителю данных	
1	А.1.2 Вариант использования № 2: Прием Национальными архивами США государственных данных на архивное хранение, поиск, извлечение и обеспечение долговременной сохранности Требуется поддерживать высокую релевантность и полноту результатов поиска. Требуется поддерживать высокую точность классификации документов. Требуется поддерживать различные системы хранения, такие как облачные сервисы NetApp, система хранения Hitachi, магнитные ленты
2	А.1.3 Вариант использования № 3: Повышение активности респондентов в статистических обследованиях Требуется поддерживать развивающуюся визуализацию для проверки данных, оперативной деятельности и общего анализа
3	А.1.4 Вариант использования № 4: Использование нетрадиционных данных для повышения активности респондентов в статистических обследованиях Требуется поддерживать развивающуюся визуализацию для проверки данных, оперативной деятельности и общего анализа
4	А.2.2 Вариант использования № 6: Международная исследовательская сеть Mendeley Требуется поддерживать специализированные инструменты создания отчетов. Требуется поддерживать инструменты визуализации, такие как визуализация сети с использованием программного обеспечения Gephi, диаграммы рассеяния (scatterplots) и т. д.
5	А.2.3 Вариант использования № 7: Сервис кинофильмов Netflix Требуется поддерживать потоковую передачу и представление видеоматериалов
6	А.2.4 Вариант использования № 8: Веб-поиск Требуется поддерживать время поиска ≈0,1 секунды. Требуется максимизировать такую метрику, как «точность 10 наилучших результатов». Требуется поддерживать адекватный макет страницы выдачи результатов (визуализация)
7	А.2.7 Вариант использования № 11: Данные о материалах Требуется поддерживать инструменты визуализации, способствующие отысканию подходящих материалов и пониманию зависимости свойств материалов от множества независимых переменных. Требуется поддерживать многопараметрические инструменты визуализации данных о материалах, способные работать с достаточно большим количеством переменных
8	А.2.8 Вариант использования № 12: Геномика материалов на основе результатов моделирования Требуется поддерживать программы просмотра данных о материалах, необходимые ввиду растущих объемов выдаваемых в ходе поиска данных
9	А.3.1 Вариант использования № 13: Облачный крупномасштабный анализ и визуализация геопространственных данных Требуется поддерживать визуализацию посредством ГИС как при высокой, так при низкой пропускной способности сети, а также на выделенных устройствах и на портативных устройствах

[1)] *Исправлены неверные ссылки на варианты использования.*

Продолжение таблицы D.4

Общие технические требования	
10	A.3.2 Вариант использования № 14: Идентификация и отслеживание объектов — Постоянное наблюдение Требуется поддерживать визуализацию извлеченных результатов путем наложения на отображение геопространственных данных. Наложенные объекты должны отсылать к соответствующему сегменту исходного изображения/видеопотока. Требуется поддерживать выходные данные в форме веб-функций, соответствующих стандартам «Открытого геопространственного консорциума» (Open Geospatial Consortium, OGC), либо в виде стандартных геопространственных файлов (Shapefile, язык разметки Keyhole (Keyhole Markup Language, KML))
11	A.3.3 Вариант использования № 15: Обработка и анализ разведывательных данных Требуется поддерживать такие основные виды визуализации, как наложения на геопространственную картину и сетевые графики (network diagrams)
12	A.4.1 Вариант использования № 16: Данные электронной медицинской документации Требуется обеспечить предоставление результатов аналитики для использования потребителями данных/заинтересованными сторонами, то есть теми, кто сам анализ не проводил. Требуется поддерживать специализированные методы визуализации
13	A.4.2 Вариант использования № 17: Анализ графических образов в патологии Цифровая патология Требуется поддерживать визуализацию для целей проверки и обучения
14	A.4.3 Вариант использования № 18: Вычислительный анализ биоизображений Требуется поддерживать работу с трехмерными структурными моделями
15	A.4.4 Вариант использования № 19: Геномные измерения Требуется поддерживать формат данных, используемый браузерами генома
16	A.4.5 Вариант использования № 20: Сравнительный анализ (мета) геномов Требуется поддерживать параллельную массовую загрузку в реальном времени. Требуется поддерживать интерактивный пользовательский веб-интерфейс к основным данным, предварительные вычисления на сервере и отправку пакетных заданий из пользовательского интерфейса. Требуется поддерживать скачивание сформированных и аннотированных наборов данных для анализа в автономном режиме. Требуется поддерживать возможность запрашивать и просматривать данные через интерактивный пользовательский веб-интерфейс. Требуется поддерживать визуализацию структурных элементов на разных уровнях разрешения, а также возможность представления группы очень похожих геномов в виде пангенома
17	A.4.7 Вариант использования № 22: Статистический реляционный искусственный интеллект для здравоохранения Требуется поддерживать визуализацию подмножеств очень больших наборов данных
18	A.4.8 Вариант использования № 23: Эпидемиологическое исследование в масштабе всего населения Земли Требуется поддерживать визуализацию
19	A.4.9 Вариант использования № 24: Моделирование распространения социального влияния Требуется поддерживать многоуровневые детальные представления в виде сетей. Требуется поддерживать визуализацию с возможностью интерактивного взаимодействия
20	A.4.10 Вариант использования № 25 Биологическое разнообразие и европейская электронная научно-исследовательская инфраструктура LifeWatch Требуется поддерживать развитую и богатую визуализацию, средства визуализации высокой четкости. Требуется поддерживать 4D-визуализацию
21	A.5.2 Вариант использования № 27: Организация крупномасштабных, неструктурированных коллекций сделанных потребителями фотографий Требуется поддерживать визуализацию крупномасштабных трехмерных реконструкций и навигацию по крупномасштабным коллекциям изображений, которые были согласованы с картами
22	A.5.3 Вариант использования № 28: Truthy — Анализ данных Твиттера Требуется поддерживать поиск/извлечение данных и их динамическую визуализацию. Требуется поддерживать управляемые данными интерактивные веб-интерфейсы. Требуется поддерживать API-интерфейсы программирования приложений для запросов к данным
23	A.5.5 Вариант использования № 30: Цифровая инфраструктура для исследований и анализа сетей и графов» (CINET) Требуется поддерживать визуализацию на стороне клиента

Окончание таблицы D.4

Общие технические требования	
24	A.5.6 Вариант использования № 31: Измерения и оценки эффективности аналитических технологий в Национальном институте стандартов и технологий (NIST) Требуется поддерживать потоки работ аналитики с участием пользователей
25	A.6.1 Вариант использования № 32: Консорциум федеративных сетей данных (DFC) Требуется поддерживать типовые потоки рабочих процессов визуализации
26	A.6.3 Вариант использования № 34: Поиск по графу для научных данных Требуется поддерживать эффективную визуализацию на основе графа данных
27	A.7.1 Вариант использования № 36: Каталинский цифровой обзор неба в поисках транзиентов Требуется поддерживать механизмы визуализации для пространств параметров данных высокой размерности
28	A.7.2 Вариант использования № 37: Космологический обзор неба и моделирование Требуется поддерживать интерпретацию результатов с использованием передовых методов и средств визуализации
29	A.7.4 Вариант использования № 39: Анализ данных Большого адронного коллайдера Требуется поддерживать построение гистограмм, диаграмм рассеяния с подбором моделей (визуализация)
30	A.8.1 Вариант использования № 41: Радарная система некогерентного рассеяния EISCAT-3D Требуется поддерживать визуализацию многомерных (≥5) данных
31	A.8.2 Вариант использования № 42: Совместная деятельность европейских сетевых инфраструктур в области экологических исследований (ENVRI) Требуется поддерживать инструменты построения графиков. Требуется поддерживать инструменты интерактивной линейной временной визуализации (на базе Google Chart Tools) для временных рядов. Требуется поддерживать отображение диаграмм в браузере с использованием технологии Flash. Требуется поддерживать визуализацию данных с высоким разрешением с привязкой к картам. Требуется поддерживать визуальные инструменты для сравнения качества моделей
32	A.8.3 Вариант использования № 43: Центра дистанционного зондирования ледяного покрова CReSIS Требуется поддерживать ГИС как пользовательский интерфейс. Требуется поддерживать богатый пользовательский интерфейс для моделирования
33	A.8.4 Вариант использования № 44: Обработка данных проекта UAVSAR Требуется поддерживать пользователей в полевых экспедициях посредством предоставления интерфейса для смартфонов/планшетов и поддержки скачивания данных с низким разрешением
34	A.8.5 Вариант использования № 45: Объединенный испытательный стенд iRODS центра НАСА в Лэнгли и Центра космических полетов имени Годдарда Требуется поддерживать визуализацию распределенных разнородных данных
35	A.8.6 Вариант использования № 46: Аналитические сервисы MERRA Требуется поддерживать высокопроизводительную распределенную визуализацию
36	A.8.7 Вариант использования № 47: Атмосферная турбулентность — Обнаружение событий Требуется поддерживать визуализацию для помощи в интерпретации результатов
37	A.8.8 Вариант использования № 48: Исследования климата с использованием модели климатической системы Земли (CESM) Требуется поддерживать коллективное использование климатических данных в глобальном масштабе. Требуется поддерживать высокопроизводительную распределенную визуализацию
38	A.8.9 Вариант использования № 49: Подповерхностные биогеохимические исследования Требуется поддерживать доступ к данным и ввод данных со смартфона
39	A.8.10 Вариант использования № 50: Сети AmeriFlux и FLUXNET Требуется поддерживать доступ к данным и ввод данных со смартфона

Таблица D.5 — Технические проблемы в категории «Безопасность и неприкосновенность частной жизни (защита персональных данных)»

Общие технические требования	
1. Необходимо обеспечить безопасность и конфиденциальность чувствительных данных.	Применимо к 30 вариантам использования: № 1–4, 7–8, 10–19, 21–25, 27–29, 31, 39–40, 42–43, 51
2. Необходимо поддерживать изолированную среду («песочницу»), обеспечивать контроль доступа и многоуровневую аутентификацию на основе политик в отношении подлежащих защите данных.	Применимо к 13 вариантам использования: № M0006[1)], 6, 8, 10, 12, 16–17, 19–21, 29, 40, 43
Специфические для варианта применения технические требования по обеспечению безопасности и неприкосновенности частной жизни (защите персональных данных)	
1	A.1.1 Вариант использования № 1: Архивное хранение больших данных переписи населения в США 2010 и 2000 годов Требуется поддерживать исполнение положений части 13 Свода законов США
2	A.1.2 Вариант использования № 2: Прием Национальными архивами США государственных данных на архивное хранение, поиск, извлечение и обеспечение долговременной сохранности Требуется поддерживать политику в области безопасности
3	A.1.3 Вариант использования № 3: Повышение активности респондентов в статистических обследованиях Требуется поддерживать более совершенные рекомендательные системы, позволяющих снизить затраты и повысить качество, обеспечивая одновременно надежные и публично проверяемые меры защиты конфиденциальности. Требуется обеспечивать безопасность и конфиденциальность всех данных. Согласно требованиям законодательства, должна быть обеспечена возможность аудита всех процессов на предмет обеспечения безопасности и конфиденциальности
4	A.1.4 Вариант использования № 4: Использование нетрадиционных данных для повышения активности респондентов в статистических обследованиях Требуется обеспечивать безопасность и конфиденциальность всех данных. Согласно требованиям законодательства, должна быть обеспечена возможность аудита всех процессов на предмет обеспечения безопасности и конфиденциальности
5	A.2.1 Вариант использования № 5: Облачные вычисления в секторах финансовой отрасли Требуется поддерживать исполнение строгих требований к обеспечению безопасности и неприкосновенности частной жизни
6	A.2.2 Вариант использования № 6: Международная исследовательская сеть Mendeley Требуется поддерживать меры контроля доступа, в частности, отслеживать, кто и к какому контенту получает доступ
7	A.2.3 Вариант использования № 7: Сервис кинофильмов Netflix Требуется обеспечивать неприкосновенность частной жизни пользователей и соблюдение цифровых прав на видеоконтент
8	A.2.4 Вариант использования № 8: Веб-поиск Требуется поддерживать контроль доступа. Требуется обеспечивать защиту чувствительного контента
9	A.2.5 Вариант использования № 9: Обеспечение непрерывности деловой деятельности и восстановления после катастроф для больших данных в облачной экосистеме Требуется обеспечивать высокий уровень безопасности во многих приложениях
10	A.2.6 Вариант использования № 10: Грузоперевозки Требуется поддерживать политику в области безопасности
11	A.2.7 Вариант использования № 11: Данные о материалах Требуется обеспечивать защиту чувствительных проприетарных данных. Требуется поддерживать инструменты для маскирования проприетарной информации

[1)] *Ссылка на несуществующий вариант использования.*

Продолжение таблицы D.5

Общие технические требования	
12	A.2.8 Вариант использования № 12: Геномика материалов на основе результатов моделирования Требуется поддерживать возможность работать в изолированной зоне-«песочнице» или же создавать независимые рабочие зоны для заинтересованных в работе с данными сторонах. Требуется поддерживать объединение (федерацию) наборов данных на основе политик
13	A.3.1 Вариант использования № 13: Облачный крупномасштабный анализ и визуализация геопространственных данных Требуется обеспечивать полную безопасность чувствительных данных при передаче и при хранении (особенно на портативных/карманных устройствах)
14	A.3.2 Вариант использования № 14: Идентификация и отслеживание объектов — Постоянное наблюдение Требуется обеспечивать высокий уровень безопасности и конфиденциальности; нельзя допустить компрометацию источников данных и методов их обработки; враг не должен знать, что именно мы видим
15	A.3.3 Вариант использования № 15: Обработка и анализ разведывательных данных Требуется обеспечивать защиту данных от несанкционированного доступа или раскрытия и от несанкционированного вмешательства
16	A.4.1 Вариант использования № 16: Данные электронной медицинской документации Требуется поддерживать прямой доступ потребителей к данным, а также ссылки на результаты аналитики, выполненной специалистами в области информатики и исследователями системы здравоохранения. Требуется обеспечивать защиту всех данных о здоровье в соответствии с действующим законодательством. Требуется обеспечивать защиту данных в соответствии с политиками поставщиков данных. Требуется поддерживать политики безопасности и обеспечения неприкосновенности частной жизни, которые могут быть уникальными для конкретных подмножеств данных. Требуется обеспечивать надежную безопасность для предотвращения утечек данных
17	A.4.2 Вариант использования № 17: Анализ графических образов в патологии/Цифровая патология Требуется обеспечивать безопасность и защиту неприкосновенности частной жизни в отношении подлежащей защите медицинской информации
18	A.4.3 Вариант использования № 18: Вычислительный анализ биоизображений Требуется обеспечивать достаточно высокий, но не являющийся обязательным уровень безопасности и защиты неприкосновенности частной жизни, включая использование защищенных серверов и анонимизацию
19	A.4.4 Вариант использования № 19: Геномные измерения Требуется обеспечивать безопасность и защиту персональных данных для медицинских документов и баз данных клинических исследований
20	A.4.5 Вариант использования № 20: Сравнительный анализ (мета) геномов Требуется обеспечивать безопасность учетных данных для входа в систему, т. е. логинов и паролей Требуется поддерживать создания учетных записей пользователей для доступа к наборам данных и представления наборов данных в систему через веб-интерфейс. Требуется поддерживать технологию единого входа (SSO).
21	A.4.6 Вариант использования № 21: Индивидуальное управление лечением диабета Требуется обеспечивать защиту медицинских данных в соответствии с политиками защиты неприкосновенности частной жизни и законодательно-нормативными требованиями к безопасности и защите персональных данных, например, имеющимися в американском законе HIPAA. Требуется поддерживать политики безопасности для разных пользовательских ролей
22	A.4.7 Вариант использования № 22: Статистический реляционный искусственный интеллект для здравоохранения Требуется поддерживать защищенную обработку данных
23	A.4.8 Вариант использования № 23: Эпидемиологическое исследование в масштабе всего населения Земли Требуется обеспечивать защиту используемых в моделировании персональных данных физических лиц. Необходимо поддерживать защиту данных и защищенную платформу для вычислений
24	A.4.9 Вариант использования № 24: Моделирование распространения социального влияния Требуется обеспечивать защиту используемых в моделировании персональных данных физических лиц. Необходимо поддерживать защиту данных и защищенную платформу для вычислений

Окончание таблицы D.5

Общие технические требования	
25	A.4.10 Вариант использования № 25 Биологическое разнообразие и европейская электронная научно-исследовательская инфраструктура LifeWatch Требуется поддерживать объединенное (федеративное) управление идентификацией для мобильных исследователей и мобильных датчиков. Требуется поддерживать управление доступом и контроль над ним
26	A.5.2 Вариант использования № 27: Организация крупномасштабных, неструктурированных коллекций сделанных потребителями фотографий Требуется обеспечивать защиту неприкосновенности частной жизни для пользователей и защиту цифровых прав на контент
27	A.5.3 Вариант использования № 28: Truthy — Анализ данных Твиттера Требуется обеспечивать политику в области безопасности и защиты неприкосновенности частной жизни
28	A.5.4 Вариант использования № 29: Краудсорсинг в гуманитарных науках Требуется решать вопросы обеспечения неприкосновенности частной жизни, сохраняя анонимность авторов полученных материалов
29	A.5.6 Вариант использования № 31: Измерения и оценки эффективности аналитических технологий в Национальном институте стандартов и технологий (NIST) Требуется обеспечивать исполнение требований по безопасности и защите персональных данных в отношении защиты чувствительных данных, обеспечивая при этом возможность проведения содержательной оценки эффективности разработок. Совместно используемые испытательные стенды должны обеспечивать защиту интеллектуальной собственности разработчиков аналитических алгоритмов
30	A.6.1 Вариант использования № 32: Консорциум федеративных сетей данных (DFC) Требуется поддерживать объединение (федерацию) существующих сред аутентификации с помощью «Типового API-интерфейса программирования приложений служб защиты данных» (Generic Security Service API) и подключаемых модулей аутентификации (GSI, Kerberos, InCommon, Shibboleth). Требуется поддерживать управление доступом к файлам независимо от места хранения
31	A.6.2 Вариант использования № 33: Discinnet-процесс Требуется обеспечивать достаточно высокий, но необязательный уровень безопасности и защиты персональных данных, включая использование защищенных серверов и анонимизацию
32	A.6.4 Вариант использования № 35: Анализ больших объемов данных, получаемых в экспериментах на синхротроне Требуется обеспечивать исполнение многочисленных требований к безопасности и защите неприкосновенности частной жизни
33	A.7.4 Вариант использования № 39: Анализ данных Большого адронного коллайдера Необходимо обеспечить защиту данных
34	A.7.5 Вариант использования № 40: Эксперимент Belle II Требуется поддерживать стандартную аутентификацию в грид-системе
35	A.8.2 Вариант использования № 42: Совместная деятельность европейских сетевых инфраструктур в области экологических исследований (ENVRI) Необходимо поддерживать политику открытых данных с небольшими ограничениями
36	A.8.3 Вариант использования № 43: Центра дистанционного зондирования ледяного покрова CReSIS Требуется обеспечивать безопасность и неприкосновенность частной жизни, в том числе с учетом деликатности политической ситуации в зоне проведения исследований. Требуется поддерживать динамичные механизмы политик в области безопасности и неприкосновенности частной жизни
37	A.9.1 Вариант использования № 51: Прогнозирование потребления электроэнергии в интеллектуальных энергосетях Требуется обеспечивать защиту персональных данных посредством анонимизации и агрегирования данных

Таблица D.6 — Технические проблемы в категории «Управление жизненным циклом»

Общие технические требования	
1. Необходимо поддерживать курирование качества данных, включая предварительную обработку, кластеризацию данных, классификацию, редуцирование (преобразование к физическим величинам) и преобразование форматов	Применимо к 20 вариантам использования: № 1—4, 6, 8, 11, 14—16, 18, 20, 22—25, 28, 39, 42—43
2. Необходимо поддерживать динамическое обновление данных, профилей пользователей и ссылок	Применимо к двум вариантам использования: № 7, 38
3. Необходимо поддерживать жизненный цикл данных и политику обеспечения долговременной сохранности, включая отслеживание происхождения данных	Применимо к шести вариантам использования: № 1, 7—8, 25, 33, 41
4. Необходимо поддерживать валидацию данных	Применимо к четырем вариантам использования: № 5—6, 22, 47
5. Необходимо поддерживать аннотирование данных человеком для целей их валидации	Применимо к 4 вариантам использования: № 17, 20—21, 44
6. Необходимо принимать меры для предотвращения утраты или порчи данных	Применимо к трем вариантам использования: № 1, 24, 41
7. Необходимо поддерживать географически распределенные (multi-site) архивы	Применимо к одному варианту использования: № 42
8. Необходимо поддерживать постоянные идентификаторы и прослеживаемость данных	Применимо к двум вариантам использования: № 6, 21
9. Необходимо поддерживать стандартизацию, агрегирование и нормализацию данных из разнородных источников	Применимо к одному варианту использования: № 16
Специфические для варианта применения технические требования к управлению жизненным циклом	
1	A.1.1 Вариант использования № 1: Архивное хранение больших данных переписи населения в США 2010 и 2000 годов Требуется поддерживать обеспечение долговременной сохранности данных «как есть» в течение 75-летнего ограничительного периода. Требуется поддерживать обеспечение долговременной сохранности на уровне битов. Требуется поддерживать процесс курирования, включая преобразование формата (конверсию). Требуется обеспечивать доступ и аналитическую обработку по истечении 75-летнего ограничительного периода. Требуется обеспечить отсутствие утраты данных
2	A.1.2 Вариант использования № 2: Прием Национальными архивами США государственных данных на архивное хранение, поиск, извлечение и обеспечение долговременной сохранности Требуется поддерживать предварительную обработку, в т. ч. сканирование на вирусы. Требуется поддерживать идентификацию файлового формата. Требуется поддерживать индексацию. Требуется поддерживать классификацию документов
3	A.1.3 Вариант использования № 3: Повышение активности респондентов в статистических обследованиях Требуется обеспечивать достоверность данных, и системы должны быть очень надежными. Остаются проблемой семантическая целостность концептуальных метаданных, описывающих, что именно измеряется, и вытекающие из этого пределы точности выводов
4	A.1.4 Вариант использования № 4: Использование нетрадиционных данных для повышения активности респондентов в статистических обследованиях Требуется обеспечивать достоверность данных, и системы должны быть очень надежными. Остаются проблемой семантическая целостность концептуальных метаданных, описывающих, что именно измеряется, и вытекающие из этого пределы точности выводов

Продолжение таблицы D.6

Общие технические требования	
5	A.2.2 Вариант использования № 6: Международная исследовательская сеть Mendeley Требуется поддерживать управление метаданными, извлеченными из PDF-файлов. Требуется поддерживать выявление дублирования документов. Требуется поддерживать постоянные идентификаторы. Требуется поддерживать сопоставление метаданных со сведениями в базах данных Crossref, PubMed и arXiv
6	A.2.3 Вариант использования № 7: Сервис кинофильмов Netflix Требуется поддерживать постоянное вычисление рейтингов и их обновление на основе профилей пользователей и результатов аналитики
7	A.2.4 Вариант использования № 8: Веб-поиск Требуется поддерживать безвозвратное уничтожение данных по истечении определенного интервала времени (несколько месяцев). Требуется поддерживать чистку данных
8	A.2.7 Вариант использования № 11: Данные о материалах Требуется поддерживать качество данных, которое, за исключением базовых данных о структурных и тепловых свойствах, является низким или непонятным
9	A.2.8 Вариант использования № 12: Геномика материалов на основе результатов моделирования Требуется поддерживать валидацию и количественную оценку неопределенности результатов моделирования путем сопоставления с экспериментальными данными. Требуется поддерживать количественную оценку неопределенности в результатах на основе нескольких наборов данных
10	A.3.2 Вариант использования № 14: Идентификация и отслеживание объектов — Постоянное наблюдение Требуется обеспечивать достоверность извлеченных объектов
11	A.3.3 Вариант использования № 15: Обработка и анализ разведывательных данных Требуется контролировать происхождение данных (включая, например, отслеживание всех передач и преобразований) в течение жизненного цикла данных
12	A.4.1 Вариант использования № 16: Данные электронной медицинской документации Требуется стандартизировать, агрегировать и нормализовать данные из разнородных источников. Требуется уменьшать количество ошибок и устранять систематические погрешности. Требуется поддерживать общую номенклатуру и классификацию контента из разных источников
13	A.4.2 Вариант использования № 17: Анализ графических образов в патологии/Цифровая патология Необходимо поддерживать аннотирование материалов человеком для использования при валидации
14	A.4.3 Вариант использования № 18: Вычислительный анализ биоизображений Требуется поддерживать компоненты потока рабочих процессов, включающие сбор, хранение, улучшение качества данных и минимизацию шума
15	A.4.5 Вариант использования № 20: Сравнительный анализ (мета) геномов Требуется поддерживать методы повышения качества данных. Требуется поддерживать кластеризацию, классификацию и редуцирование данных. Требуется поддерживать интеграцию новых данных/контента в системное хранилище данных и аннотирование данных
16	A.4.6 Вариант использования № 21: Индивидуальное управление лечением диабета Требуется поддерживать аннотирование данных на основе онтологий и таксономий. Требуется обеспечивать прослеживаемость данных от источника (начальной точки сбора) и далее на протяжении периода работы с ними. Требуется поддерживать преобразование данных из существующего хранилища данных в триплеты RDF
17	A.4.7 Вариант использования № 22: Статистический реляционный искусственный интеллект для здравоохранения Требуется поддерживать объединение нескольких таблиц перед выполнением анализа. Требуется поддерживать методы валидации данных с целью минимизации ошибок

Окончание таблицы D.6

Общие технические требования	
18	A.4.8 Вариант использования № 23: Эпидемиологическое исследование в масштабе всего населения Земли Требуется обеспечивать качество данных и отслеживание происхождения данных в ходе вычислений
19	A.4.9 Вариант использования № 24: Моделирование распространения социального влияния Требуется поддерживать объединение данных из различных источников данных. Требуется поддерживать согласованность данных и предотвращать их порчу. Требуется поддерживать предварительную обработку первичных данных
20	A.4.10 Вариант использования № 25 Биологическое разнообразие и европейская электронная научно-исследовательская инфраструктура LifeWatch Требуется поддерживать хранение и архивацию данных, обмен данными и их интеграцию. Требуется поддерживать управление жизненным циклом данных, включая происхождение данных, ссылочную целостность и идентификацию, прослеживаемость до первоначальных данных наблюдений. Требуется поддерживать обработанные (вторичные) данных (в дополнение к оригинальным исходным данным), которые могут быть сохранены для использования в будущем. Требуется контролировать происхождение с присвоением постоянного идентификатора (PID) данных, алгоритмов и рабочих процессов. Требуется поддерживать курированные (авторизованные) эталонные данные (т. е. списки названий видов), алгоритмы, программные коды и рабочие процессы
21	A.5.3 Вариант использования № 28: Truthy — Анализ данных Твиттера Требуется поддерживать стандартизированные структуры данных/форматы и исключительно высокое качество данных
22	A.6.2 Вариант использования № 33: Discinnet-процесс Требуется поддерживать интеграцию методов работы с метаданными различных дисциплин
23	A.7.3 Вариант использования № 38: Большие данные космологических обзоров неба Требуется поддерживать связи между удаленными телескопами и центрами аналитической обработки
24	A.7.4 Вариант использования № 39: Анализ данных Большого адронного коллайдера Требуется поддерживать качество данных на сложных установках
25	A.8.1 Вариант использования № 41: Радарная система некогерентного рассеяния EISCAT-3D Требуется обеспечивать долговременную сохранность данных и предотвращать утрату данных в случае сбоев в работе измерительного комплекса
26	A.8.2 Вариант использования № 42: Совместная деятельность европейских сетевых инфраструктур в области экологических исследований (ENVRI) Требуется поддерживать высокое качество данных. Требуется поддерживать зеркальные архивы. Требуется поддерживать различные схемы метаданных. Требуется поддерживать разрозненные хранилища и курирование данных
27	A.8.3 Вариант использования № 43: Центра дистанционного зондирования ледяного покрова CReSIS Требуется поддерживать уверенность в качестве данных
28	A.8.4 Вариант использования № 44: Обработка данных проекта UAVSAR Требуется поддерживать значительное вмешательство человека в конвейер обработки данных. Требуется поддерживать подробные и надежные сведения о происхождении, описывающие сложный процесс обработки компьютером/человеком
29	A.8.7 Вариант использования № 47: Атмосферная турбулентность — Обнаружение событий Требуется поддерживать валидацию для выходных продуктов (корреляции)

Таблица D.7 — Технические проблемы в категории «Иные технические проблемы»

Общие технические требования	
1. Необходимо поддерживать пользовательский интерфейс с широкими возможностями для мобильных платформ, с целью обеспечения доступа к обработанным результатам	Применимо к шести вариантам использования: № 2, 7, 19, 28, 44, 46
2. Необходимо поддерживать мониторинг, с использованием мобильных платформ, производительности аналитической обработки	Применимо к двум вариантам использования: № 41, 43
3. Необходимо поддерживать визуальный поиск по контенту, с широкими функциональными возможностями, и отображение контента на мобильных платформах	Применимо к 13 вариантам использования: № 3, 6—8, 12, 16—17, 19, 39, 48—51
4. Необходимо поддерживать сбор данных с использованием мобильных устройств	Применимо к одному варианту использования: № 42
5. Необходимо обеспечивать безопасность на мобильных устройствах	Применимо к одному варианту использования: № 16

Специфические для варианта применения иные технические требования	
1	A.1.2 Вариант использования № 2: Прием Национальными архивами США государственных данных на архивное хранение, поиск, извлечение и обеспечение долговременной сохранности Требуется поддержка мобильного поиска, который должен иметь похожий интерфейс и выдавать похожие результаты
2	A.1.3 Вариант использования № 3: Повышение активности респондентов в статистических обследованиях Требуется поддержка мобильного доступа
3	A.2.1 Вариант использования № 5: Облачные вычисления в секторах финансовой отрасли Требуется поддержка мобильного доступа
4	A.2.2 Вариант использования № 6: Международная исследовательская сеть Mendeley Требуется поддержка доставки контента и услуг на различные вычислительные платформы, от настольных компьютеров под Windows до мобильных устройств под ОС Android и iOS
5	A.2.3 Вариант использования № 7: Сервис кинофильмов Netflix Требуется поддержка интеллектуальных интерфейсов для доступа к киноконтенту на мобильных платформах
6	A.2.4 Вариант использования № 8: Веб-поиск Требуется поддержка мобильного поиска и отображения
7	A.2.8 Вариант использования № 12: Геномика материалов на основе результатов моделирования Требуется поддержка мобильных приложений для доступа к информации по геномике материалов
8	A.4.1 Вариант использования № 16: Данные электронной медицинской документации Требуется обеспечение безопасности на мобильных устройствах
9	A.4.2 Вариант использования № 17: Анализ графических образов в патологии/Цифровая патология Требуется поддержка трехмерной визуализации и отображения на мобильных платформах
10	A.4.4 Вариант использования № 19: Геномные измерения Требуется обеспечить доступ врачам к геномным данным на мобильных платформах
11	A.4.6 Вариант использования № 21: Индивидуальное управление лечением диабета Требуется обеспечить поддержку мобильного доступа к данным
12	A.4.9 Вариант использования № 24: Моделирование распространения социального влияния Требуется перемещение данных ближе к вычислительным ресурсам с целью повышения эффективности
13	A.4.10 Вариант использования № 25 Биологическое разнообразие и европейская электронная научно-исследовательская инфраструктура LifeWatch Требуется поддержка доступа для мобильных пользователей

Окончание таблицы D.7

	Общие технические требования
14	A.5.3 Вариант использования № 28: Truthy — Анализ данных Твиттера Требуется поддержка низкоуровневых функциональных возможностей инфраструктуры хранения данных с целью обеспечения эффективного мобильного доступа к данным
15	A.8.1 Вариант использования № 41: Радарная система некогерентного рассеяния EISCAT-3D Требуется поддержка мониторинга оборудования в режиме реального времени, посредством частичного анализа потока данных
16	A.8.2 Вариант использования № 42: Совместная деятельность европейских сетевых инфраструктур в области экологических исследований (ENVRI) Требуется поддержка мобильных датчиков и измерительных устройств различных типов с целью сбора данных
17	A.8.3 Вариант использования № 43: Центра дистанционного зондирования ледяного покрова CReSIS Требуется поддержка мониторинга собирающих данные устройств и датчиков
18	A.8.4 Вариант использования № 44: Обработка данных проекта UAVSARъ Требуется поддержка работающих в полевых условиях пользователей посредством предоставления интерфейсов к смартфонам/планшетам и возможности скачивания данных в низком разрешении
19	A.8.6 Вариант использования № 46: Аналитические сервисы MERRA Требуется поддержка доступа со смартфонов и планшетов. Требуется поддержка управления данными посредством iRODS
20	A.8.8 Вариант использования № 48: Исследования климата с использованием модели климатической системы Земли (CESM) Требуется поддержка ввода данных и доступа со смартфонов
21	A.8.9 Вариант использования № 49: Подповерхностные биогеохимические исследования Требуется поддержка ввода данных и доступа со смартфонов
22	A.8.10 Вариант использования № 50: Сети AmeriFlux и FLUXNET Требуется поддержка ввода данных и доступа со смартфонов
23	A.9.1 Вариант использования № 51: Прогнозирование потребления электроэнергии в интеллектуальных энергосетях Требуется поддержка мобильного доступа для клиентов

Приложение ДА
(справочное)

Сведения о соответствии ссылочных международных стандартов национальным стандартам

Таблица ДА.1

Обозначение ссылочного международного стандарта	Степень соответствия	Обозначение и наименование соответствующего национального стандарта
ISO/IEC 20546	IDT	ГОСТ Р ИСО/МЭК 20546—2021 «Информационные технологии. Большие данные. Обзор и словарь»
Примечание — В настоящей таблице использовано следующее условное обозначение степени соответствия стандартов: - IDT — идентичные стандарты.		

Библиография

[1] NIST. NIST Big data Public Working Group (NBD-PWG) List of general requirements versus architecture component. 2013 [accessed 2014 August 8]; Available from: http://bigdatawg.nist.gov/uc_reqs_gen.php

[2] Fox G.C., Jha S., Qiu J., Luckow A. Ogres: A Systematic Approach to Big data Benchmarks, in Big data and Extreme-scale Computing (BDEC) January 29-30, 2015. Barcelona. http://www.exascale.org/bdec/sites/www.exascale.org.bdec/files/whitepapers/OgreFacets.pdf

[3] Geoffrey C. FOX, Shantenu JHA, Judy QIU, Saliya EKANAYAKE, and Andre LUCKOW, Towards a Comprehensive Set of Big data Benchmarks, Chapter in Big data and High Performance Computing, Lucio Grandinetti and Gerhard Joubert, Editors. 2015, IOS. http://grids.ucs.indiana.edu/ptliupages/publications/OgreFacetsv9.pdf

[4] Fox G.C., Jha S., Qiu J., Luckow A. Towards an Understanding of Facets and Exemplars of Big data Applications, in 20 Years of Beowulf: Workshop to Honor Thomas Sterling's 65th Birthday April 13, 2015. Annapolis. http://dsc.soic.indiana.edu/publications/OgrePaperv11.pdf

[5] NIST. Big data Working Group Reports from V1. 2013 [accessed 2014 March 26]; Report at http://bigdatawg.nist.gov/V1_output_docs.php Available from: http://bigdatawg.nist.gov/home.php

[6] Malik O. Internet of things will have 24 billion devices by 2020 from GSMA, the global mobile industry trade group. 2011 [accessed 2014 July 19]; Available from: http://gigaom.com/2011/10/13/internet-of-things-will-have-24-billion-devices-by-2020/

[7] Cisco V.N.I. Forecast and Methodology, 2012—2017. 2013 May 29 [accessed 2013 August 14]; Available from: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-481360_ns827_Networking_Solutions_White_Paper.html

[8] Cisco Internet Business Solutions Group (IBSG). (Dave Evans), The Internet of Things: How the Next Evolution of the Internet Is Changing Everything. 2011 April [accessed 2013 August 14]; Available from: http://www.cisco.com/web/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf

[9] ISO/IEC 20547-3, Information technology — Big data reference architecture — Part 3: Reference architecture

[10] ISO/IEC 27000:2016, Information technology — Security techniques — Information security management systems — Overview and vocabulary

[11] ISO/IEC 29161:2016, Information technology — Data structure — Unique identification for the Internet of Things

УДК 004.01:006.354 ОКС 35.020

Ключевые слова: информационные технологии, ИТ, данные, большие данные, аналитика данных, база данных, модель данных, массив данных, разнообразие данных, скорость обработки данных, достоверность данных, объем данных, распределенная обработка данных, неструктурированные данные, частично структурированные данные, потоковые данные

Редактор *Л.В. Коретникова*
Технический редактор *В.Н. Прусакова*
Корректор *С.В. Смирнова*
Компьютерная верстка *Г.Д. Мухиной*

Сдано в набор 08.12.2021. Подписано в печать 10.01.2022. Формат 60×84⅛. Гарнитура Ариал.
Усл. печ. л. 33,95. Уч.-изд. л. 30,72.

Подготовлено на основе электронной версии, предоставленной разработчиком стандарта

Создано в единичном исполнении в ФГБУ «РСТ»
для комплектования Федерального информационного фонда стандартов,
117418 Москва, Нахимовский пр-т, д. 31, к. 2.
www.gostinfo.ru info@gostinfo.ru