



BSI Standards Publication

Information technology — Big data reference architecture

Part 2: Use cases and derived requirements

National foreword

This Published Document is the UK implementation of ISO/IEC TR 20547-2:2018.

The UK participation in its preparation was entrusted to Technical Committee ICT/-/1/1, Big Data - Foundational Standards.

A list of organizations represented on this committee can be obtained on request to its secretary.

This publication does not purport to include all the necessary provisions of a contract. Users are responsible for its correct application.

© The British Standards Institution 2018
Published by BSI Standards Limited 2018

ISBN 978 0 580 94460 4

ICS 35.020

Compliance with a British Standard cannot confer immunity from legal obligations.

This Published Document was published under the authority of the Standards Policy and Strategy Committee on 28 February 2018.

Amendments/corrigenda issued since publication

Date	Text affected
<hr/>	

**Information technology — Big data
reference architecture —**

**Part 2:
Use cases and derived requirements**

*Technologies de l'information — Architecture de référence des big
data —*

Partie 2: Cas pratiques et exigences dérivées





COPYRIGHT PROTECTED DOCUMENT

© ISO/IEC 2018

All rights reserved. Unless otherwise specified, or required in the context of its implementation, no part of this publication may be reproduced or utilized otherwise in any form or by any means, electronic or mechanical, including photocopying, or posting on the internet or an intranet, without prior written permission. Permission can be requested from either ISO at the address below or ISO's member body in the country of the requester.

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva, Switzerland
Tel. +41 22 749 01 11
Fax +41 22 749 09 47
copyright@iso.org
www.iso.org

Published in Switzerland

Contents

Page

Foreword	v
Introduction	vi
1 Scope	1
2 Normative references	1
3 Terms and definitions	1
3.1 Terms defined elsewhere	1
3.2 Terms defined in this document	1
3.3 Abbreviated terms	1
4 Use case properties for survey	6
4.1 Overall description	6
4.2 Current solution	7
4.3 Big data characteristics	7
4.4 Big data science	7
4.5 Overall big data issues	8
4.6 Big data use case Template	8
5 Use cases summaries	9
5.1 Use case development process	9
5.2 Government operation	10
5.2.1 Use case 1: Census 2010 and 2000 — Title 13 big data	10
5.2.2 Use case 2: NARA Accession, Search, Retrieve, Preservation	10
5.2.3 Use case 3: Statistical survey response improvement	11
5.2.4 Use case 4: Non-Traditional Data in Statistical Survey Response Improvement (Adaptive Design)	11
5.3 Commercial	12
5.3.1 Use case 5: Cloud Eco-System for Financial Industries	12
5.3.2 Use case 6: Mendeley — An International Network of Research	12
5.3.3 Use case 7: Multi-media streaming service	13
5.3.4 Use case 8: Web Search	13
5.3.5 Use case 9: Big data Business Continuity and Disaster Recovery Within a Cloud Eco-System	14
5.3.6 Use case 10: Cargo Shipping	14
5.3.7 Use case 11: Materials Data for Manufacturing	14
5.3.8 Use case 12: Simulation-Driven Materials Genomics	15
5.4 Defense	16
5.4.1 Use case 13: Cloud Large-Scale Geospatial Analysis and Visualization	16
5.4.2 Use case 14: Object Identification and Tracking from Wide-Area Large Format Imagery or Full Motion Video—Persistent Surveillance	16
5.4.3 Use case 15: Intelligence Data Processing and Analysis	17
5.5 Health care and life sciences	17
5.5.1 Use case 16: Electronic Medical Record Data	17
5.5.2 Use case 17: Pathology Imaging/Digital Pathology	18
5.5.3 Use case 18: Computational Bioimaging	18
5.5.4 Use case 19: Genomic Measurements	19
5.5.5 Use case 20: Comparative Analysis for Metagenomes and Genomes	19
5.5.6 Use case 21: Individualized Diabetes Management	19
5.5.7 Use case 22: Statistical Relational Artificial Intelligence for Health Care	20
5.5.8 Use case 23: World Population-Scale Epidemiological Study	20
5.5.9 Use case 24: Social Contagion Modeling for Planning, Public Health, and Disaster Management	21
5.5.10 Use case 25: Biodiversity and LifeWatch	21
5.6 Deep Learning and Social Media	22
5.6.1 Use case 26: Large-Scale Deep Learning	22

5.6.2	Use case 27: Organizing Large-Scale, Unstructured Collections of Consumer Photos	22
5.6.3	Use case 28: Truthy—Information Diffusion Research from Twitter Data	23
5.6.4	Use case 29: Crowd Sourcing in the Humanities as Source for Big and Dynamic Data	23
5.6.5	Use case 30: CINET—Cyberinfrastructure for Network (Graph) Science and Analytics	23
5.6.6	Use case 31: NIST Information Access Division — Analytic Technology Performance Measurements, Evaluations, and Standards	24
5.7	The Ecosystem for research	24
5.7.1	Use case 32: DataNet Federation Consortium	24
5.7.2	Use case 33: The Discinnet Process	25
5.7.3	Use case 34: Semantic Graph Search on Scientific Chemical and Text-Based Data	25
5.7.4	Use case 35: Light Source Beamlines	26
5.8	Astronomy and physics	26
5.8.1	Use case 36: Catalina Real-Time Transient Survey: A Digital, Panoramic, Synoptic Sky Survey	26
5.8.2	Use case 37: DOE Extreme Data from Cosmological Sky Survey and Simulations	27
5.8.3	Use case 38: Large Survey Data for Cosmology	27
5.8.4	Use case 39: Particle Physics—Analysis of Large Hadron Collider Data: Discovery of Higgs Particle	28
5.8.5	Use case 40: Belle II High Energy Physics Experiment	29
5.9	Earth, environmental, and polar science	29
5.9.1	Use case 41: European Incoherent Scatter Scientific Association 3D Incoherent Scatter Radar System	29
5.9.2	Use case 42: Common Operations of Environmental Research Infrastructure	30
5.9.3	Use case 43: Radar Data Analysis for the Center for Remote Sensing of Ice Sheets	31
5.9.4	Use case 44: Unmanned Air Vehicle Synthetic Aperture Radar (UAVSAR) Data Processing, Data Product Delivery, and Data Services	31
5.9.5	Use case 45: NASA Langley Research Center/ Goddard Space Flight Center iRODS Federation Test Bed	32
5.9.6	Use case 46: MERRA Analytic Services (MERRA/AS)	32
5.9.7	Use case 47: Atmospheric Turbulence – Event Discovery and Predictive Analytics	32
5.9.8	Use case 48: Climate Studies Using the Community Earth System Model at the U.S. Department of Energy (DOE) NERSC Center	33
5.9.9	Use case 49: DOE Biological and Environmental Research (BER) Subsurface Biogeochemistry Scientific Focus Area	33
5.9.10	Use case 50: DOE BER AmeriFlux and FLUXNET Networks	34
5.10	Energy	34
5.10.1	Use case 51: Consumption Forecasting in Smart Grids	34
5.10.2	Use case 52: Home Energy Management System	34
6	Use cases derived technical considerations	35
6.1	Use case specific technical considerations	35
6.2	Summary of requirements analysis	35
6.3	Features of use cases	37
	Annex A Submitted use case studies	40
	Annex B Summary of Key Properties	197
	Annex C Use case technical considerations summary	207
	Annex D Use case detail technical considerations	225
	Bibliography	252

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work. In the field of information technology, ISO and IEC have established a joint technical committee, ISO/IEC JTC 1.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular the different approval criteria needed for the different types of document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation on the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see the following URL: www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee ISO/IEC JTC 1, *Information Technology*.

A list of all parts in the ISO/IEC 20547-series can be found on the ISO website.

Introduction

This document is focuses on forming a community of interest from industry, academia, and government, with the goal of developing a consensus list of big data technical considerations across all stakeholders. This included gathering and understanding various examples of use cases from diversified areas (i.e., application domains). To achieve this goal, the following tasks were done:

- gathered input from all stakeholders regarding big data technical considerations;
- analyzed and prioritized a list of challenging use case specific technical considerations that may delay or prevent adoption of big data deployment;
- developed a comprehensive list of generalized big data technical considerations for ISO/IEC 20547-3, *Information technology – Big data reference architecture - Part 3: Reference architecture*; and
- documented the findings in this document.

Information technology — Big data reference architecture —

Part 2: Use cases and derived requirements

1 Scope

This document provides examples of big data use cases with application domains and technical considerations derived from the contributed use cases.

2 Normative references

The following documents, in whole or in part, are normatively referenced in this document and are indispensable for its application. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 20546 *Information technology — Big data — Definition and vocabulary*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 20546 and the following apply.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

- IEC Electropedia: available at <http://www.electropedia.org/>
- ISO Online browsing platform: available at <http://www.iso.org/obp>

3.1 Terms defined elsewhere

None.

3.2 Terms defined in this document

3.2.1

use case

typical application stated at a high level for the purposes of extracting technical considerations or comparing usages across fields

3.3 Abbreviated terms

2D	two-Dimensional
3D	three-Dimensional
6D	six-Dimensional
AOD	Analysis Object Data

API	Application Programming Interface
ASDC	Atmospheric Science Data Center
ASTM	American Society for Testing and Materials
AWS	Amazon Web Services
BC/DR	Business Continuity and Disaster Recovery
BD	Big data
BER	Biological and Environmental Research
BNL	Brookhaven National Laboratory
CAaaS	Climate Analytics as a Service
CADRG	Compressed ARC Digitized Raster Graphics
CBSP	Cloud Brokerage Service Provider
CERES	Clouds and Earth's Radiant Energy System
CERN	European Organization for Nuclear Research
CESM	Community Earth System Model
CFTC	U.S. Commodity Futures Trading Commission
CIA	Confidentiality, Integrity, and Availability
CINET	Cyberinfrastructure for Network (Graph) Science and Analytics
CMIP	Coupled Model Intercomparison Project
CMIP5	Climate Model Intercomparison Project
CMS	Compact Muon Solenoid
COSO	Committee of Sponsoring Organizations
CPU	Central Processing Unit
CRaISIS	Center for Remote Sensing of Ice Sheets
CRTS	Catalina Real-Time Transient Survey
CSP	Cloud Service Provider
CSS	Catalina Sky Survey proper
CV	Controlled Vocabulary
DFC	DataNet Federation Consortium
DHTC	Distributed High Throughput Computing
DNA	DeoxyriboNucleic Acid
DOE	U.S. Department of Energy

DOJ	U.S. Department of Justice
DPO	Data Products Online
EBAF–TOA	Energy Balanced and Filled–Top of Atmosphere
EC2	Elastic Compute Cloud
EDT	Enterprise Data Trust
EHR	Electronic Health Record
EMR	Electronic Medical Record
EMSO	European Multidisciplinary Seafloor and Water Column Observatory
ENVRI	Common Operations of Environmental Research Infrastructures
ENVRI RM	ENVRI Reference Model
EPOS	European Plate Observing System
ESFRI	European Strategy Forum on Research Infrastructures
ESG	Earth System Grid
ESGF	Earth System Grid Federation
FDIC	U.S. Federal Deposit Insurance Corporation
FI	Financial Industries
FLUXNET	Flux Tower Network
FMV	Full Motion Video
FNAL	Fermi National Accelerator Laboratory
GAAP	U.S. Generally Accepted Accounting Principles
GB	Giga Byte
GCM	General Circulation Model
GEOS-5	Goddard Earth Observing System version 5
GeoTiff	Geo Tagged Image File Format
GEWaSC	Genome-Enabled Watershed Simulation Capability
GHG	Green House Gas
GMAO	Global Modeling and Assimilation Office
GPFS	General Parallel File System
GPS	Global Positioning System
GPU	Graphics Processing Unit
GRC	Governance, Risk management, and Compliance

GSFC	Goddard Space Flight Center
HDF5	Hierarchical Data Format
HDFS	Hadoop Distributed File System
HPC	High-Performance Computing
HTC	High-Throughput Computing
HVS	Hosted Virtual Server
I/O	Input Output
IaaS	Infrastructure as a Service
IAGOS	In-service Aircraft for a Global Observing System
ICD	International Classification of Diseases
ICOS	Integrated Carbon Observation System
IMG	Integrated Microbial Genomes
INPC	Indiana Network for Patient Care
IPCC	Intergovernmental Panel on Climate Change
iRODS	Integrated Rule-Oriented Data System
ISACA	International Society of Auditors and Computer Analysts
isc2	International Security Computer and Systems Auditors
ISO	International Organization for Standardization
ITIL	Information Technology Infrastructure Library
JGI	Joint Genome Institute
KML	Keyhole Markup Language
kWh	kilowatt-hour
LaRC	Langley Research Center
LBNL	Lawrence Berkeley National Laboratory
LDA	latent Dirichlet allocation
LHC	Large Hadron Collider
LPL	Lunar and Planetary Laboratory
LSST	Large Synoptic Survey Telescope
MERRA	Modern Era Retrospective Analysis for Research and Applications
MERRA/AS	MERRA Analytic Services
MPI	Message Passing Interface

MRI	Magnetic Resonance Imaging
NARA	National Archives and Records Administration
NARR	North American Regional Reanalysis
NaaS	Network as a Service
NASA	National Aeronautics and Space Administration
NCAR	National Center for Atmospheric Research
NCBI	National Center for Biotechnology Information
NCCS	NASA Center for Climate Simulation
NERSC	National Energy Research Scientific Computing Center
NetCDF	Network Common Data Form
NEX	NASA Earth Exchange
NFS	Network File System
NIKE	NIST Integrated Knowledge Editorial Net
NIST	National Institute of Standards and Technology
NITF	National Imagery Transmission Format
NLP	Natural Language Processing
NRT	Near Real Time
NSF	National Science Foundation
ODP	Open Distributed Processing
OGC	Open Geospatial Consortium
PB	PetaByte
PCA	Principal Component Analysis
PCAOB	Public Company Accounting and Oversight Board
PID	persistent identification
PII	Personally Identifiable Information
PNNL	Pacific Northwest National Laboratory
RDBMS	relational database management system
RDF	Resource Description Framework
RECOVER	Rehabilitation Capability Convergence for Ecosystem Recovery
ROI	return on investment
RPI	Repeat Pass Interferometry

RPO	Recovery Point Objective
RTO	Response Time Objective
SAN	storage area network
SAR	Synthetic Aperture Radar
SDN	software-defined networking
SIOS	Svalbard Integrated Arctic Earth Observing System
SPADE	Support for Provenance Auditing in Distributed Environments
SSH	Secure Shell
SSO	Single Sign-On
TB	TeraByte
tf-idf	term frequency–inverse document frequency
UA	University of Arizona
UAVSAR	Unmanned Air Vehicle Synthetic Aperture Radar
UC	Use Case
UI	User Interface
UPS	United Parcel Service
UQ	Uncertainty Quantification
VASP	Vienna Ab initio Simulation Package
vCDS	virtual Climate Data Server
VO	Virtual Observatory
VOIP	Voice over IP
WALF	Wide Area Large Format Imagery
WLCG	Worldwide LHC Computing Grid
XBRL	eXtensible Business Related Markup Language
XML	Extensible Markup Language
ZTF	Zwicky Transient Factory

4 Use case properties for survey

4.1 Overall description

- **Use case title:** Title provided by the use case author
- **Vertical (area):** Intended to categorize the use cases. However, an ontology was not created prior to the use case submissions so this field was not used in the use case compilation.

- **Author/company/email:** Name, company, and email (if provided) of the person(s) submitting the use case
- **Actors/ stakeholders and their roles and responsibilities:** Description of the players and their roles in the use case
- **Goals:** Objectives of the use case
- **Use case description:** Brief description of the use case

4.2 Current solution

Current solutions describe current approach to processing big data at the hardware and software infrastructure and analytics level.

- **Compute (System):** Computing component of the data analysis system
- **Storage:** Storage component of the data analysis system
- **Networking:** Networking component of the data analysis system
- **Software:** Software component of the data analysis system

4.3 Big data characteristics

Big data Characteristics describe the properties of the (raw) data including the four major 'V's' of big data.

- **Data source:** The origin of data, which could be from instruments, Internet of Things, Web, Surveys, Commercial activity, or from simulations. The source(s) can be distributed, centralized, local, or remote.
- **Data destination:** If data transformed in use case, where the final results end up.
- **Volume:** The characteristic of datasets that is most associated with big data. Volume represents the extensive amount of data available for analysis to extract valuable information. The assumption that you can extract the most value by analysing as much of the volume of data as possible was one of the primary drivers for the creation of the new scaling technologies.
- **Velocity:** The rate of flow at which the data is created, stored, analysed, or visualized. Big data velocity means a large quantity of data needs to be processed in a short amount of time. Dealing with high velocity data is commonly referred to as techniques for streaming data.
- **Variety:** The need to analyse data from a number of domains and a number of data types. The variety of data was handled through transformations or pre-analytics to extract features that would allow integration with other data. The wider range of data formats, logical models, timescales, and semantics, which is desirable to be used in analytics, complicates the integration of the variety of data. Metadata is increasingly used to aid in the integration.
- **Variability:** Changes in data rate, format/structure, semantics, and/or quality that impact the supported application, analytic, or problem. Impacts can include the need to refactor architectures, interfaces, processing/algorithms, integration/fusion, storage, applicability, or use of the data.

4.4 Big data science

Big data science describes the high level aspects of the data analysis process.

- **Veracity and data quality:** This covers the completeness and accuracy of the data with respect to semantic content as well as syntactical quality of data (such as presence of missing fields or incorrect values).

- **Visualization:** Refers to the way data is viewed by an analyst making decisions based on the data. Typically, visualization is the final stage of a technical data analysis pipeline and follows the data analytics stage.
- **Data types:** Refers to the style of data such as structured, unstructured, images (e.g., pixels), text (e.g., characters), gene sequences, and numerical.
- **Metadata:** Comments on quality and richness of metadata.
- **Curation and governance:** Comment on process to ensure good data quality and who is responsible.

NOTE The use case template has a separate item to describe security and privacy issues.

- **Data analytics:** Refers broadly to tools and algorithms used in processing the data at any stage including the data to information or knowledge to wisdom stages, as well as the information to knowledge stage.

4.5 Overall big data issues

- **Other big data issues:** Did we miss something important that your use case highlights? Your chance to address questions which we should have asked.
- **User Interface and mobile access issues:** Refers to issues in accessing or generating big data from clients including smart phones and tablets.
- **List key features and related use cases:** Put use case in context of related use cases. What features generalize and what are idiosyncratic to this use case.
- **Project future:** How do you expect application, and approach (hardware, software, analytics) to change in future?
- **More project information (URLs):** Put a collection of useful links.

4.6 Big data use case Template

This clause provides one blank use case template. The below blank use case template was used for the purpose of capturing use cases to derived technical consideration.

NOTE The terms used in this template may or may not match with ISO/IEC 20546 and other parts of the ISO/IEC 20547-series.

Use case title	
Vertical (area)	
Author/company/email	
Actors/stakeholders and their roles and responsibilities	
Goals	
Use case description	
Current solutions	Compute(System)
	Storage
	Networking
	Software

Big data characteristics	Data source (distributed/centralized)	
	Volume (size)	
	Velocity (e.g. real time)	
	Variety (multiple datasets, mashup)	
	Variability (rate of change)	
Big data science (collection, curation, analysis, action)	Veracity (Robustness Issues, semantics)	
	Visualization	
	Data quality (syntax)	
	Data types	
	Data analytics	
Big data specific challenges (Gaps)		
Big data specific challenges in mobility		
Security and privacy technical considerations		
Highlight issues for generalizing this Use case (e.g. for ref. architecture)		
More information (URLs)		
NOTE <additional comments>		

5 Use cases summaries

5.1 Use case development process

A use case is a typical application stated at a high level for the purposes of extracting technical considerations or comparing usages across fields. In order to develop a consensus list of big data technical considerations across all stakeholders, publicly available information was collected for various big data architectures. After collection of use cases, application domains were identified to better organize the collection of use cases.

NOTE 1 The list of application domains reflects the use cases submitted and is not intended to be exhaustive.

The nine application domains were as follows:

- **Government operation** (4): National Archives and Records Administration, Census Bureau;
- **Commercial** (8): Finance in Cloud, Cloud Backup, Citations, Multi-media streaming, Web Search, Digital Materials, Cargo Shipping;
- **Defense** (3): Sensors, Image Surveillance, Situation Assessment ;
- **Healthcare and life sciences** (10): Medical Records, Graph and Probabilistic Analysis, Pathology, Bioimaging, Genomics, Epidemiology, People Activity Models, Biodiversity;
- **Deep learning and social media** (6) Self-driving cars, Geolocate Images, SNS, Crowd Sourcing, Network Science, Benchmark Datasets;

- **Ecosystem for research** (4): Metadata, Collaboration, Language Translation, Light Source Experiments;
- **Astronomy and physics** (5): Sky Surveys (and comparisons to simulation), LHC at CERN, Belle Accelerator II;
- **Earth, environmental, and polar science** (10): Radar Scattering in Atmosphere, Earthquake, Ocean, Earth Observation, Ice Sheet Radar Scattering, Earth Radar Mapping, Climate Simulation Datasets, Atmospheric Turbulence Identification, Subsurface Biogeochemistry (microbes to watersheds), Gas Sensors;
- **Energy** (2): Smart Grid, Home energy management.

NOTE 2 The template was valuable for gathering consistent information to develop supporting analysis and comparison of the use cases. However, varied levels of detail and quantitative or qualitative information was received for each use case template section. For some application domains, several similar big data use cases are presented, providing a more complete view of big data technical considerations within that application domain.

The examples of use cases are presented in this clause with the information originally submitted. The original content (See [Annex A](#)) has not been modified.

NOTE 3 Specific vendor solutions and technologies are mentioned in the use cases. However, the listing of these solutions and technologies does not constitute endorsement from the JTC 1 WG 9.

The use cases are numbered sequentially to facilitate cross-referencing between the use case summaries presented in this clause, the original use cases ([Annex A](#)), and the use case summary tables ([Annexes B, C, and D](#)).

5.2 Government operation

5.2.1 Use case 1: Census 2010 and 2000 — Title 13 big data

Application:

Census 2010 and 2000—Title 13 data must be preserved for several decades so they can be accessed and analyzed after 75 years. Data must be maintained ‘as-is’ with no access and no data analytics for 75 years, preserved at the bit level, and curated, which may include format transformation. Access and analytics must be provided after 75 years. Title 13 of the U.S. Code authorizes the U.S. Census Bureau to collect and preserve census related data and guarantees that individual and industry-specific data are protected.

Current approach:

The dataset contains 380 TB of scanned documents.

Future:

Future data scenarios and applications were not expressed for this use case.

5.2.2 Use case 2: NARA Accession, Search, Retrieve, Preservation

Application:

This area comprises accession, search, retrieval, and long-term preservation of government data.

Current approach:

The data are currently handled as follows.

- Get physical and legal custody of the data.
- Pre-process data for conducting virus scans, identifying file format, and removing empty files

- Index the data.
- Categorize records (e.g., sensitive, non-sensitive, privacy data).
- Transform old file formats to modern formats.
- Conduct e-discovery.
- Search and retrieve to respond to special requests..
- Search and retrieve public records by public users.

Hundreds of TBs are stored centrally in commercial databases supported by custom software and commercial search products.

Future:

Federal agencies possess many distributed data sources, which currently must be transferred to centralized storage. In the future, those data sources may reside in multiple cloud environments. In this case, physical custody should avoid transferring big data from cloud to cloud or from cloud to data center.

5.2.3 Use case 3: Statistical survey response improvement

Application:

Survey costs are increasing as survey responses decline. The goal of this work is to increase the quality — and reduce the cost — of field surveys by using advanced ‘recommendation system techniques.’ These techniques are open and scientifically objective, using data mashed up from several sources and also historical survey para-data (i.e., administrative data about the survey.)

Current approach:

This use case handles about a PB of data coming from surveys and other government administrative sources. Data can be streamed. During the decennial census, approximately 150 million records transmitted as field data are streamed continuously. All data must be both confidential and secure. All processes must be auditable for security and confidentiality as required by various legal statutes. Data quality should be high and statistically checked for accuracy and reliability throughout the collection process. Solution information is described in [Clause A.1.3](#)

Future:

Improved recommendation systems are needed similar to those used in e-commerce (e.g., similar to the use case [5.3.3](#) that reduce costs and improve quality, while providing confidentiality safeguards that are reliable and publicly auditable. Data visualization is useful for data review, operational activity, and general analysis. The system continues to evolve and incorporate important features such as mobile access.

5.2.4 Use case 4: Non-Traditional Data in Statistical Survey Response Improvement (Adaptive Design)

Application:

Survey costs are increasing as survey response declines. This use case has goals similar to those of the Statistical Survey Response Improvement use case (see [Clause 5.2.3](#)). However, this case involves non-traditional commercial and public data sources from the web, wireless communication, and electronic transactions mashed up analytically with traditional surveys. The purpose of the mashup is to improve statistics for small area geographies and new measures, as well as the timeliness of released statistics.

Current approach:

Data from a range of sources are integrated including survey data, other government administrative data, web scrapped data, wireless data, e-transaction data, possibly social media data, and positioning

data from various sources. Software, visualization, and data characteristics are similar to those in the Statistical Survey Response Improvement use case.

Future:

Analytics need to be developed that give more detailed statistical estimations, on a more near real-time basis, for less cost. The reliability of estimated statistics from such mashed up sources still must be evaluated.

5.3 Commercial

5.3.1 Use case 5: Cloud Eco-System for Financial Industries

Application:

Use of cloud (e.g., big data) technologies needs to be extended in financial industries (i.e., banking, securities and investments, insurance) transacting business within the U.S.

Current approach:

The financial industry is already using big data for fraud detection, risk analysis, assessments, as well as improving their knowledge and understanding of customers. At the same time, the industry is still using traditional client/server/data warehouse/relational database management system (RDBMS) for the handling, processing, storage, and archival of financial data. Real-time data and analysis are important in these applications.

Future:

Security, privacy, and regulation must be addressed. For example, the financial industry must examine SEC-mandated use of XBRL (extensible business-related markup language) and use of other cloud functions.

5.3.2 Use case 6: Mendeley — An International Network of Research

Application:

Mendeley has built a database of research documents and facilitates the creation of shared bibliographies. Mendeley collects and uses the information about research reading patterns and other activities conducted via their software to build more efficient literature discovery and analysis tools. Text mining and classification systems enable automatic recommendation of relevant research, improving research teams' performance and cost-efficiency, particularly those engaged in curation of literature on a particular subject.

Current approach:

Data size is presently 15 TB and growing at a rate of about 1 TB per month. Solution information is described in [Clause A.2.2](#). The database uses standard libraries for machine learning and analytics, latent Dirichlet allocation (LDA, a generative probabilistic model for discrete data collection), and custom-built reporting tools for aggregating readership and social activities for each document.

Future:

Currently big data storage batch jobs are scheduled daily, but work has begun on real-time recommendation. The database contains approximately 400 million documents and roughly 80 million unique documents, and receives 500,000 to 700,000 new uploads on a weekday. Thus a major challenge is clustering matching documents together in a computationally efficient way (i.e., scalable and parallelized) when they are uploaded from different sources and have been slightly modified via third-party annotation tools or publisher watermarks and cover pages.

5.3.3 Use case 7: Multi-media streaming service

Application:

This allows streaming of user-selected movies to satisfy multiple objectives (for different stakeholders)—but with a focus on retaining subscribers. The company needs to find the best possible ordering of a set of videos for a user (e.g., household) within a given context in real time, with the objective of maximizing movie consumption. Recommendation systems and streaming video delivery are core Netflix technologies. Recommendation systems are always personalized and use logistic/linear regression, elastic nets, matrix factorization, clustering, LDA, association rules, gradient-boosted decision trees, and other tools. Digital movies are stored in the cloud with metadata, along with individual user profiles and rankings for small fraction of movies. The current system uses multiple criteria: a content-based recommendation system, a user-based recommendation system, and diversity. Algorithms are continuously refined with A/B testing (i.e., two-variable randomized experiments used in online marketing).

Current approach:

It held a competition for the best collaborative filtering algorithm to predict user ratings for films—the purpose of which was to improve ratings by 10 %. The winning system combined over 100 different algorithms. Solution information is described in [Clause A.2.3](#). Business initiatives have been used to increase viewership.

Future:

Streaming video is a very competitive business. It needs to be aware of other companies and trends in both content (e.g., which movies are popular) and big data technology.

5.3.4 Use case 8: Web Search

Application:

A web search function returns results in $\sim 0,1$ s based on search terms with an average of three words. It is important to maximize quantities such as “precision@10” for the number of highly accurate/appropriate responses in the top 10 ranked results.

Current approach:

The current approach uses the following steps.

- Crawl the web.
- Pre-process data to identify what is searchable (words, positions).
- Form an inverted index, which maps words to their locations in documents
- Rank the relevance of documents using the PageRank algorithm.
- Employ advertising technology, e.g., using reverse engineering to identify ranking models—or preventing reverse engineering.
- Cluster documents into topics (as in Google News).
- Update results efficiently.

Modern clouds and technologies such as Map/Reduce have been heavily influenced by this application, which now comprises ~ 45 billion web pages total.

Future:

Web search is a very competitive field, so continuous innovation is needed. Two important innovation areas are addressing the growing segment of mobile clients, and increasing sophistication of responses

and layout to maximize the total benefit of clients, advertisers, and the search company. The “deep web” (content not indexed by standard search engines, buried behind user interfaces to databases, etc.) and multimedia searches are also of increasing importance. Each day, 500 million photos are uploaded, and each minute, 100 hours of video are uploaded to YouTube.

5.3.5 Use case 9: Big data Business Continuity and Disaster Recovery Within a Cloud Eco-System

Application:

Business Continuity and Disaster Recovery (BC/DR) needs to consider the role that four overlaying and interdependent forces will play in ensuring a workable solution to an entity's business continuity plan and requisite disaster recovery strategy. The four areas are people (i.e., resources), processes (e.g., time/cost/return on investment [ROI]), technology (e.g., various operating systems, platforms, and footprints), and governance (e.g., subject to various and multiple regulatory agencies).

Current approach:

Data replication services are provided through cloud ecosystems, incorporating IaaS and supported by Tier 3 data centers. Replication is different from backup and only moves the changes that took place since the previous replication, including block-level changes. The replication can be done quickly—with a five-second window—while the data are replicated every four hours. This data snapshot is retained for seven business days, or longer if necessary. Replicated data can be moved to a failover center (i.e., a backup system) to satisfy an organization's recovery point objectives (RPO) and recovery time objectives (RTO). Related solution information is described in [Annex A](#). Data sizes range from terabytes to petabytes.

Future:

Migrating from a primary site to either a replication site or a backup site is not yet fully automated. The goal is to enable the user to automatically initiate the failover sequence. Both organizations must know which servers have to be restored and what the dependencies and inter-dependencies are between the primary site servers and replication and/or backup site servers. This knowledge requires continuous monitoring of both.

5.3.6 Use case 10: Cargo Shipping

Application:

Delivery companies need optimal means of monitoring and tracking cargo.

Current approach:

Information is updated only when items are checked with a bar code scanner, which sends data to the central server. An item's location is not currently displayed in real time.

Future:

Tracking items in real time is feasible through the Internet of Things application, in which objects are given unique identifiers and capability to transfer data automatically, i.e., without human interaction. A new aspect will be the item's status condition, including sensor information, global positioning system (GPS) coordinates, and a unique identification schema based upon standards (ISO/IEC 29161:2016 *Information technology – Data structure – Unique identification for the Internet of Things*).

5.3.7 Use case 11: Materials Data for Manufacturing

Application:

Every physical product is made from a material that has been selected for its properties, cost, and availability. This translates into hundreds of billions of dollars of material decisions made every year. However, the adoption of new materials normally takes decades (usually two to three decades) rather

than a small number of years, in part because data on new materials are not easily available. To speed adoption time, accessibility, quality, and usability must be broadened, and proprietary barriers to sharing materials data must be overcome. Sufficiently large repositories of materials data are needed to support discovery.

Current approach:

Decisions about materials usage are currently unnecessarily conservative, are often based on older rather than newer materials research and development data, and do not take advantage of advances in modeling and simulation.

Future:

Materials informatics is an area in which the new tools of data science can have a major impact by predicting the performance of real materials (in gram to ton quantities) starting at the atomistic, nanometer, and/or micrometer levels of description. The following efforts are needed to support this area.

- Establish materials data repositories, beyond the existing ones, that focus on fundamental data.
- Develop internationally accepted data recording standards that can be used by a very diverse materials community, including developers of materials test standards (e.g., American Society for Testing and Materials [ASTM] International and ISO), testing companies, materials producers, and research and development labs.
- Develop tools and procedures to help organizations that need to deposit proprietary materials in data repositories to mask proprietary information while maintaining the data's usability.
- Develop multi-variable materials data visualization tools in which the number of variables can be quite high.

5.3.8 Use case 12: Simulation-Driven Materials Genomics

Application:

Massive simulations spanning wide spaces of possible design lead to innovative battery technologies. Systematic computational studies are being conducted to examine innovation possibilities in photovoltaics. Search and simulation is the basis for rational design of materials. All these require management of simulation results contributing to the materials genome.

Current approach:

Survey results are produced using PyMatGen, FireWorks, Vienna Ab initio Simulation Package (VASP), ABINIT, NWChem, BerkeleyGW, and varied materials community codes running on large supercomputers, such as the Hopper at the National Energy Research Scientific Computing Center (NERSC), a 150,000-core machine that produces high-resolution simulations.

Future:

Large-scale computing and flexible data methods at scale for messy data are needed for simulation science. The advancement of goal-driven thinking in materials design requires machine learning and knowledge systems that integrate data from publications, experiments, and simulations. Other needs include scalable key-value and object store databases; the current 100 TB of data will grow to 500 TB over the next five years.

5.4 Defense

5.4.1 Use case 13: Cloud Large-Scale Geospatial Analysis and Visualization

Application:

Large-scale geospatial data analysis and visualization must be supported. As the number of geospatially aware sensors and geospatially tagged data sources increase, the volume of geospatial data requiring complex analysis and visualization is growing exponentially.

Current approach:

Traditional geographic information systems (GISs) are generally capable of analyzing millions of objects and visualizing thousands. Data types include imagery (various formats such as National Imagery Transmission Format (NITF), Geo Tagged Image File Format (GeoTiff), and Compressed ARC Digitized Raster Graphics (CADRG)) and vector (various formats such as shape files, KML [Keyhole Markup Language], and text streams). Object types include points, lines, areas, polylines, circles, and ellipses. Image registration—transforming various data into one system—requires data and sensor accuracy. Analytics include principal component analysis (PCA) and independent component analysis (ICA) and consider closest point of approach, deviation from route, and point density over time. Software information is described in [A.3.1](#).

Future:

Today's intelligence systems often contain trillions of geospatial objects and must visualize and interact with millions of objects. Critical issues are indexing, retrieval and distributed analysis (note that geospatial data requires unique approaches to indexing and distributed analysis); visualization generation and transmission; and visualization of data at the end of low-bandwidth wireless connections. Data are sensitive and must be completely secure in transit and at rest (particularly on handhelds).

5.4.2 Use case 14: Object Identification and Tracking from Wide-Area Large Format Imagery or Full Motion Video—Persistent Surveillance

Application:

Persistent surveillance sensors can easily collect PB of imagery data in the space of a few hours. The data should be reduced to a set of geospatial objects (e.g., points, tracks) that can be easily integrated with other data to form a common operational picture. Typical processing involves extracting and tracking entities (e.g., vehicles, people, and packages) over time from the raw image data.

Current approach:

It is not feasible for humans to process these data for either alerting or tracking purposes. The data need to be processed close to the sensor, which is likely forward-deployed since it is too large to be easily transmitted. Typical object extraction systems are currently small (e.g., 1 to 20 nodes) graphics processing unit (GPU)-enhanced clusters. There are a wide range of custom software and tools, including traditional RDBMSs and display tools. Real-time data are obtained at Full Motion Video (FMV)—30 to 60 frames per second at full-color 1 080p resolution (i.e., 1 920 × 1 080 pixels, a high-definition progressive scan) or Wide-Area Large Format Imagery (WALF)—1 to 10 frames per second at 10,000 pixels × 10,000 pixels and full-color resolution. Visualization of extracted outputs will typically be as overlays on a geospatial (i.e., GIS) display. Analytics are basic object detection analytics and integration with sophisticated situation awareness tools with data fusion. Significant security issues must be considered; sources and methods cannot be compromised (i.e., “the enemy” should not know what we see).

Future:

A typical problem is integration of this processing into a large GPU cluster capable of processing data from several sensors in parallel and in near real time. Transmission of data from sensor to system is also a major challenge.

5.4.3 Use case 15: Intelligence Data Processing and Analysis

Application:

Intelligence analysts need the following capabilities:

- Identify relationships between entities (e.g., people, organizations, places, equipment).
- Spot trends in sentiment or intent for either the general population or a leadership group such as state and non-state actors.
- Identify the locations and possibly timing of hostile actions including implantation of improvised explosive devices.
- Track the location and actions of potentially hostile actors.
- Reason against and derive knowledge from diverse, disconnected, and frequently unstructured (e.g., text) data sources.
- Process data close to the point of collection, and allow for easy sharing of data to/from individual soldiers, forward-deployed units, and senior leadership in garrisons.

Current approach:

Data size ranges from tens of terabytes to hundreds of petabytes, with imagery intelligence devices gathering a petabyte in a few hours. Dismounted warfighters typically have at most one to hundreds of gigabytes (GBs), which is typically handheld data storage. Software information is described in Annex [A.3.3](#).

Future:

Data currently exist in disparate silos. These data must be accessible through a semantically integrated data space. A wide variety of data types, sources, structures, and quality will span domains and require integrated search and reasoning. Most critical data are either unstructured or maintained as imagery or video, which requires significant processing to extract entities and information. Network quality, provenance, and security are essential.

5.5 Health care and life sciences

5.5.1 Use case 16: Electronic Medical Record Data

Application:

Large national initiatives around health data are emerging. These include developing a digital learning health care system to support increasingly evidence-based clinical decisions with timely, accurate, and up-to-date patient-centered clinical information; using electronic observational clinical data to efficiently and rapidly translate scientific discoveries into effective clinical treatments; and electronically sharing integrated health data to improve healthcare process efficiency and outcomes. These key initiatives all rely on high-quality, large-scale, standardized, and aggregate health data. Advanced methods are needed for normalizing patient, provider, facility, and clinical concept identification within and among separate health care organizations. With these methods in place, feature selection, information retrieval, and enhanced machine learning decision-models can be used to define and extract clinical phenotypes from non-standard, discrete, and free-text clinical data. Clinical phenotype data must be leveraged to support cohort selection, clinical outcomes research, and clinical decision support.

Current approach:

The Indiana Network for Patient Care (INPC), the nation's largest and longest-running health information exchange, houses clinical data from more than 1,100 discrete logical operational healthcare sources. More than 20 TB of raw data, these data describe over 12 million patients and over 4 billion

discrete clinical observations. Between 500,000 and 1,5 million new real-time clinical transactions are added every day.

Future:

Running on an Indiana University supercomputer, Teradata, PostgreSQL, and MongoDB will support information retrieval methods to identify relevant clinical features (e.g., term frequency-inverse document frequency [tf-idf], latent semantic analysis, and mutual information). NLP techniques will extract relevant clinical features. Validated features will be used to parameterize clinical phenotype decision models based on maximum likelihood estimators and Bayesian networks. Decision models will be used to identify a variety of clinical phenotypes such as diabetes, congestive heart failure, and pancreatic cancer.

5.5.2 Use case 17: Pathology Imaging/Digital Pathology

Application:

Digital pathology imaging is an emerging field in which examination of high-resolution images of tissue specimens enables novel and more effective ways to diagnose diseases. Pathology image analysis segments massive spatial objects (e.g., millions of objects per image) such as nuclei and blood vessels, represented with their boundaries, along with many extracted image features from these objects. The derived information is used for many complex queries and analytics to support biomedical research and clinical diagnosis

Current approach:

Each 2D image comprises 1 GB of raw image data and entails 1,5 GB of analytical results. Message Passing Interface (MPI) is used for image analysis. Solution information is described in [A.4.2](#).

Future:

Recently, 3D pathology imaging has been made possible using 3D laser technologies or serially sectioning hundreds of tissue sections onto slides and scanning them into digital images. Segmenting 3D microanatomic objects from registered serial images could produce tens of millions of 3D objects from a single image. This provides a deep 'map' of human tissues for next-generation diagnosis. 3D images can comprise 1 TB of raw image data and entail 1 TB of analytical results. A moderated hospital would generate 1 PB of data per year.

5.5.3 Use case 18: Computational Bioimaging

Application:

Data delivered from bioimaging are increasingly automated, higher resolution, and multi-modal. This has created a data analysis bottleneck that, if resolved, can advance bioscience discovery through big data techniques.

Current approach:

The current piecemeal analysis approach does not scale to situations in which a single scan on emerging machines is 32 TB and medical diagnostic imaging is annually around 70 PB, excluding cardiology. A web-based, one-stop shop is needed for high-performance, high-throughput image processing for producers and consumers of models built on bio-imaging data.

Future:

The goal is to resolve that bottleneck with extreme-scale computing and community-focused science gateways, both of which apply massive data analysis toward massive imaging data sets. Workflow components include data acquisition, storage, enhancement, noise minimization, segmentation of regions of interest, crowd-based selection and extraction of features, and object classification, as well as organization and search. Possible software packages are described [A.4.3](#).

5.5.4 Use case 19: Genomic Measurements

Application:

The NIST Genome in a Bottle Consortium integrates data from multiple sequencing technologies and methods to develop highly confident characterization of whole human genomes as reference materials. The consortium also develops methods to use these reference materials to assess performance of any genome sequencing run.

Current approach:

NIST's approximately 40 TB network file system (NFS) is full. The National Institutes of Health (NIH) and the National Center for Biotechnology Information (NCBI) are also currently storing PBs of data. NIST is also storing data using open-source sequencing bioinformatics software from academic groups (UNIX-based) on a 72-core cluster, supplemented by larger systems at collaborators.

Future:

DNA sequencers can generate ~300 GB of compressed data per day, and this volume has increased much faster than Moore's Law gives for increase in computer processing power. Future data could include other 'omics' (e.g., genomics) measurements, which will be even larger than DNA sequencing. Clouds have been explored as a cost effective scalable approach.

5.5.5 Use case 20: Comparative Analysis for Metagenomes and Genomes

Application:

Given a metagenomic sample this use case aims to do the following:

- determine the community composition in terms of other reference isolate genomes;
- characterize the function of its genes;
- begin to infer possible functional pathways;
- characterize similarity or dissimilarity with other metagenomic samples;
- begin to characterize changes in community composition and function due to changes in environmental pressures;
- isolate subsections of data based on quality measures and community composition.

Current approach:

The current integrated comparative analysis system for metagenomes and genomes is front-ended by an interactive web user interface (UI) with core data. The system involves backend precomputations and batch job computation submission from the UI. The system provides an interface to standard bioinformatics tools (e.g., BLAST, HMMER, multiple alignment and phylogenetic tools, gene callers, sequence feature predictors).

Future:

Management of heterogeneity of biological data is currently performed by a RDBMS (i.e., Oracle). Unfortunately, it does not scale for even the current volume, 50 TB of data. NoSQL solutions aim at providing an alternative, but unfortunately they do not always lend themselves to real-time interactive use or rapid and parallel bulk loading, and sometimes they have issues regarding robustness.

5.5.6 Use case 21: Individualized Diabetes Management

Application:

Diabetes is a growing illness in the world population, affecting both developing and developed countries. Current management strategies do not adequately take into account individual patient profiles, such as co-morbidities and medications, which are common in patients with chronic illnesses. Advanced graph-based data mining techniques must be applied to electronic health records (EHRs), converting them into RDF (Resource Description Framework) graphs. These advanced techniques would facilitate searches for diabetes patients and allow for extraction of their EHR data for outcome evaluation.

Current approach:

Typical patient data records are composed of 100 controlled vocabulary values and 1,000 continuous values. Most values have a timestamp. The traditional paradigm of relational row-column lookup needs to be updated to semantic graph traversal.

Future:

The first step is to compare patient records to identify similar patients from a large EHR database (i.e., an individualized cohort.) Each patient's management outcome should be evaluated to formulate the most appropriate solution for a given patient with diabetes. The time-dependent properties need to be processed before query to allow matching based on derivatives and other derived properties. Software information is described in [A.4.6](#).

5.5.7 Use case 22: Statistical Relational Artificial Intelligence for Health Care

Application:

The goal of the project is to analyze large, multi-modal medical data, including different data types such as imaging, EHR, and genetic and natural language. This approach employs relational probabilistic models that have the capability of handling rich relational data and modeling uncertainty using probability theory. The software learns models from multiple data types, and can possibly integrate information and reason about complex queries. Users can provide a set of descriptions, for instance: magnetic resonance imaging (MRI) images and demographic data about a particular subject. They can then query for the onset of a particular disease (e.g., Alzheimer's), and the system will provide a probability distribution over the possible occurrence of this disease.

Current approach:

A single server can handle a test cohort of a few hundred patients with associated data of hundreds of GBs.

Future:

A cohort of millions of patients can involve PB size datasets. A major issue is the availability of too much data (e.g., images, genetic sequences), which can make the analysis complicated. Sometimes, large amounts of data about a single subject are available, but the number of subjects is not very high (i.e., data imbalance). This can result in learning algorithms picking up random correlations between the multiple data types as important features in analysis. Another challenge lies in aligning the data and merging from multiple sources in a form that will be useful for a combined analysis.

5.5.8 Use case 23: World Population-Scale Epidemiological Study

Application:

There is a need for reliable, real-time prediction and control of pandemics similar to the 2009 H1N1 influenza. Addressing various kinds of contagion diffusion may involve modeling and computing information, diseases, and social unrest. Agent-based models can utilize the underlying interaction network (i.e., a network defined by a model of people, vehicles, and their activities) to study the evolution of the desired phenomena.

Current approach:

There is a two-step approach: (1) build a synthetic global population; and (2) run simulations over the global population to reason about outbreaks and various intervention strategies. The current 100 TB

dataset was generated centrally with an MPI-based simulation system written in Charm++. Parallelism is achieved by exploiting the disease residence time period.

Future:

Large social contagion models can be used to study complex global-scale issues, greatly increasing the size of systems used.

5.5.9 Use case 24: Social Contagion Modeling for Planning, Public Health, and Disaster Management

Application:

Social behavior models are applicable to national security, public health, viral marketing, city planning, and disaster preparedness. In a social unrest application, people take to the streets to voice either unhappiness with or support for government leadership. Models would help quantify the degree to which normal business and activities are disrupted because of fear and anger, the possibility of peaceful demonstrations and/or violent protests, and the potential for government responses ranging from appeasement, to allowing protests, to issuing threats against protestors, to taking actions to thwart protests. Addressing these issues would require fine-resolution models (at the level of individual people, vehicles, and buildings) and datasets.

Current approach:

The social contagion model infrastructure simulates different types of human-to-human interactions (e.g., face-to-face versus online media), and also interactions between people, services (e.g., transportation), and infrastructure (e.g., Internet, electric power). These activity models are generated from averages such as census data.

Future:

One significant concern is data fusion (i.e., how to combine data from different sources and how to deal with missing or incomplete data.). A valid modeling process must take into account heterogeneous features of hundreds of millions or billions of individuals, as well as cultural variations across countries. For such large and complex models, the validation process itself is also a challenge.

5.5.10 Use case 25: Biodiversity and LifeWatch

Application:

Research and monitor different ecosystems, biological species, their dynamics, and their migration with a mix of custom sensors and data access/processing, and a federation with relevant projects in the area. Particular case studies include monitoring alien species, migrating birds, and wetlands. One of many efforts from the consortium titled Common Operations for Environmental Research Infrastructures (ENVRI) is investigating integration of LifeWatch with other environmental e-infrastructures.

Current approach:

At this time, this project is in the preliminary planning phases and, therefore, the current approach is not fully developed.

Future:

The LifeWatch initiative will provide integrated access to a variety of data, analytical, and modeling tools as served by a variety of collaborating initiatives. It will also offer data and tools in selected workflows for specific scientific communities. In addition, LifeWatch will provide opportunities to construct personalized “virtual labs,” allowing participants to enter and access new data and analytical tools. New data will be shared with the data facilities cooperating with LifeWatch, including both the Global Biodiversity Information Facility and the Biodiversity Catalogue, also known as the Biodiversity Science Web Services Registry. Data include ‘omics’, species information, ecological information (e.g.,

biomass, population density), and ecosystem data (e.g., carbon dioxide [CO₂] fluxes, algal blooming, water and soil characteristics.)

5.6 Deep Learning and Social Media

5.6.1 Use case 26: Large-Scale Deep Learning

Application:

There is a need to increase the size of datasets and models that can be tackled with deep learning algorithms. Large models (e.g., neural networks with more neurons and connections) combined with large datasets are increasingly the top performers in benchmark tasks for vision, speech, and NLP. It will be necessary to train a deep neural network from a large (e.g., much greater than 1 TB) corpus of data, which is typically comprised of imagery, video, audio, or text. Such training procedures often require customization of the neural network architecture, learning criteria, and dataset pre-processing. In addition to the computational expense demanded by the learning algorithms, the need for rapid prototyping and ease of development is extremely high.

Current approach:

The largest applications so far are to image recognition and scientific studies of unsupervised learning with 10 million images and up to 11 billion parameters on a 64 GPU HPC Infiniband cluster. Both supervised (i.e., using existing classified images) and unsupervised applications are being investigated.

Future:

Large datasets of 100 TB or more may be necessary to exploit the representational power of the larger models. Training a self-driving car could take 100 million images at megapixel resolution. Deep learning shares many characteristics with the broader field of machine learning. The paramount requirements are high computational throughput for mostly dense linear algebra operations, and extremely high productivity for researcher exploration. High-performance libraries must be integrated with high-level prototyping environments.

5.6.2 Use case 27: Organizing Large-Scale, Unstructured Collections of Consumer Photos

Application:

Collections of millions to billions of consumer images are used to produce 3D reconstructions of scenes—with no a priori knowledge of either the scene structure or the camera positions. The resulting 3D models allow efficient and effective browsing of large-scale photo collections by geographic position. New images can be geolocated by matching them to 3D models, and object recognition can be performed on each image. The 3D reconstruction can be posed as a robust, non-linear, least squares optimization problem: observed or noisy correspondences between images are constraints, and unknowns are six-dimensional (6D) camera poses of each image and 3D positions of each point in the scene.

Current approach:

The current system information is described in [A.5.2](#). Over 505 billion images are currently on SNS, with over 500 million images added to social media sites each day.

Future:

Necessary maintenance and upgrades require many analytics including feature extraction, feature matching, and large-scale probabilistic inference. These analytics appear in many or most computer vision and image processing problems, including recognition, stereo resolution, and image denoising. Other needs are visualizing large-scale, 3D reconstructions and navigating large-scale collections of images that have been aligned to maps.

5.6.3 Use case 28: Truthy—Information Diffusion Research from Twitter Data

Application:

How communication spreads on socio-technical networks must be better understood, and methods are needed to detect potentially harmful information spread at early stages (e.g., deceiving messages, orchestrated campaigns, and untrustworthy information).

Current approach:

Twitter generates a large volume of continuous streaming data—about 30 TB a year, compressed—through circulation of ~100 million messages per day. The increase over time is roughly 500 GB data per day. All these data must be acquired and stored. Additional needs include near real-time analysis of such data for anomaly detection, stream clustering, signal classification, and online-learning; and data retrieval, big data visualization, data-interactive web interfaces, and public application programming interfaces (APIs) for data querying. Software information is described in [A.5.4](#). Information diffusion, clustering, and dynamic network visualization capabilities already exist.

Future:

It plans to expand and so needs to move toward advanced distributed storage programs, and in-memory database, described in [A.5.4](#) for real-time analysis. Solutions will need to incorporate streaming clustering, anomaly detection, and online learning.

5.6.4 Use case 29: Crowd Sourcing in the Humanities as Source for Big and Dynamic Data

Application:

Information is captured from many individuals and their devices using a range of sources: manually entered, recorded multimedia, reaction times, pictures, sensor information. These data are used to characterize wide-ranging individual, social, cultural, and linguistic variations among several dimensions (e.g., space, social space, time).

Current approach:

At this point, typical systems used are Extensible Markup Language (XML) technology and traditional relational databases. Other than pictures, not much multi-media is employed yet.

Future:

Crowd sourcing is beginning to be used on a larger scale. However, the availability of sensors in mobile devices provides a huge potential for collecting large amount of data from numerous individuals. This possibility has not been explored on a large scale so far; existing crowd sourcing projects are usually of a limited scale and web-based. Privacy issues may be involved because of access to individuals' audio-visual files; anonymization may be necessary but not always possible. Data management and curation are critical. With multimedia, the size could be hundreds of terabytes.

5.6.5 Use case 30: CINET—Cyberinfrastructure for Network (Graph) Science and Analytics

Application:

CINET provides a common web-based platform that allows the end user seamless access to the following:

- network and graph analysis tools such as SNAP, NetworkX, and Galib;
- real-world and synthetic networks;
- computing resources;
- data management systems.

Current approach:

CINET uses an InfiniBand-connected HPC cluster with 720 cores to provide HPC as a service. The platform is being used for research and education. CINET is used in classes and to support research by social science and social networking communities

Future:

Rapid repository growth is expected to lead to at least 1,000 to 5,000 networks and methods in about a year. As more fields use graphs of increasing size, parallel algorithms will be important. Two critical challenges are data manipulation and bookkeeping of the derived data, as there are no well-defined and effective models and tools for unified management of various graph data.

5.6.6 Use case 31: NIST Information Access Division — Analytic Technology Performance Measurements, Evaluations, and Standards

Application:

Performance metrics, measurement methods, and community evaluations are needed to ground and accelerate development of advanced analytic technologies in the areas of speech and language processing, video and multimedia processing, biometric image processing, and heterogeneous data processing, as well as the interaction of analytics with users. Typically one of two processing models are employed: (1) push test data out to test participants, and analyze the output of participant systems, and (2) push algorithm test harness interfaces out to participants, bring in their algorithms, and test them on internal computing clusters.

Current approach:

There is a large annotated corpora of unstructured/semi-structured text, audio, video, images, multimedia, and heterogeneous collections of the above, including ground truth annotations for training, developmental testing, and summative evaluations. The test corpora exceed 900 million web pages occupying 30 TB of storage, 100 million tweets, 100 million ground-truthed biometric images, several hundred thousand partially ground-truthed video clips, and terabytes of smaller fully ground-truthed test collections.

Future:

Even larger data collections are being planned for future evaluations of analytics involving multiple data streams and very heterogeneous data. In addition to larger datasets, the future includes testing of streaming algorithms with multiple heterogeneous data. The use of clouds is being explored.

5.7 The Ecosystem for research

5.7.1 Use case 32: DataNet Federation Consortium

Application:

The DataNet Federation Consortium (DFC) promotes collaborative and interdisciplinary research through a federation of data management systems across federal repositories, national academic research initiatives, institutional repositories, and international collaborations. The collaboration environment runs at scale and includes petabytes of data, hundreds of millions of files, hundreds of millions of metadata attributes, tens of thousands of users, and a thousand storage resources.

Current approach:

Currently, 25 science and engineering domains have projects that rely on the iRODS (Integrated Rule-Oriented Data System) policy-based data management system. Active organizations include the National Science Foundation, with major projects such as the Ocean Observatories Initiative (sensor archiving); Temporal Dynamics of Learning Center (cognitive science data grid); iPlant Collaborative (plant genomics); Drexel's engineering digital library; and H. W. Odum Institute for Research in Social

Science (data grid federation with Dataverse). iRODS currently manages PB of data, hundreds of millions of files, hundreds of millions of metadata attributes, tens of thousands of users, and a thousand storage resources. It interoperates with workflow systems (e.g., National Center for Computing Applications' [NCSA's] Cyberintegrator, Kepler, Taverna), cloud, and more traditional storage models, as well as different transport protocols.

Future:

Future data scenarios and applications were not expressed for this use case.

5.7.2 Use case 33: The Discinnet Process

Application:

Discinnet has developed a Web 2.0 collaborative platform and research prototype as a pilot installation, which is now being deployed and tested by researchers from a growing number of diverse research fields. The goal is to reach a wide enough sample of active research fields, represented as clusters (i.e., researchers projected and aggregating within a manifold of mostly shared experimental dimensions) to test general, hence potentially interdisciplinary, epistemological models throughout the present decade.

Current approach:

Currently, 35 clusters have been started, with close to 100 awaiting more resources. There is potential for many more to be created, administered, and animated by research communities. Examples of clusters include optics, cosmology, materials, microalgae, health care, applied math, computation, rubber, and other chemical products/issues.

Future:

Discinnet itself would not be big data but rather will generate metadata when applied to a cluster that involves big data. In interdisciplinary integration of several fields, the process would reconcile metadata from many complexity levels.

5.7.3 Use case 34: Semantic Graph Search on Scientific Chemical and Text-Based Data

Application:

Social media-based infrastructure, terminology and semantic data-graphs are established to annotate and present technology information. The process uses root- and rule-based methods currently associated primarily with certain Indo-European languages, such as Sanskrit and Latin.

Current approach:

Many reports, including a recent one on the Material Genome Project, find that exclusive top-down solutions to facilitate data sharing and integration are not desirable for multi-disciplinary efforts. However, a bottom-up approach can be chaotic. For this reason, there is need for a balanced blend of the two approaches to support easy-to-use techniques to metadata creation, integration, and sharing. This challenge is very similar to the challenge faced by language developers, so a recently developed method is based on these ideas. There are ongoing efforts to extend this method to publications of interest to the Material Genome Initiative, the Open Government movement, and the NIST Integrated Knowledge Editorial Net (NIKE) a NIST-wide publication archive. These efforts are a component of the Research Data Alliance Metadata Standards Directory Working Group.

Future:

A cloud infrastructure should be created for social media of scientific information. Scientists from across the world could use this infrastructure to participate and deposit results of their experiments. Prior to establishing a scientific social medium, some issues must be resolved including the following:

- Minimize challenges related to establishing re-usable, interdisciplinary, scalable, on-demand, use-case, and user-friendly vocabulary.

- Adopt an existing or create new on-demand 'data-graph' to place information in an intuitive way, such that it would easily integrate with existing data-graphs in a federated environment, independently of details of data management.
- Find relevant scientific data without spending too much time on the Internet.

Start with resources such as the Open Government movement, Material Genome Initiative, and Protein Databank. This effort includes many local and networked resources. Developing an infrastructure to automatically integrate information from all these resources using data-graphs is a challenge, but steps are being taken to solve it. Strong database tools and servers for data-graph manipulation are needed.

5.7.4 Use case 35: Light Source Beamlines

Application:

Samples are exposed to X-rays from light sources in a variety of configurations, depending on the experiment. Detectors, essentially high-speed digital cameras, collect the data. The data are then analyzed to reconstruct a view of the sample or process being studied.

Current approach:

A variety of commercial and open source software is used for data analysis. Data transfer is accomplished using physical transport of portable media, which severely limits performance, high-performance GridFTP, managed by Globus Online, or workflow systems such as SPADE (Support for Provenance Auditing in Distributed Environments, an open source software infrastructure).

Future:

Camera resolution is continually increasing. Data transfer to large-scale computing facilities is becoming necessary because of the computational power required to conduct the analysis on timescales useful to the experiment. Because of the large number of beamlines (e.g., 39 at the LBNL Advanced Light Source), aggregate data load is likely to increase significantly over the coming years, as will the need for a generalized infrastructure for analyzing GB per second of data from many beamline detectors at multiple facilities.

5.8 Astronomy and physics

5.8.1 Use case 36: Catalina Real-Time Transient Survey: A Digital, Panoramic, Synoptic Sky Survey

Application:

Catalina Real-Time Transient Survey (CRTS) explores the variable universe in the visible light regime, on timescales ranging from minutes to years, by searching for variable and transient sources. It discovers a broad variety of astrophysical objects and phenomena, including various types of cosmic explosions (e.g., supernovae), variable stars, phenomena associated with accretion to massive black holes (e.g., active galactic nuclei) and their relativistic jets, and high proper motion stars. The data are collected from three telescopes (two in Arizona and one in Australia), with additional ones expected in the near future in Chile.

Current approach:

The survey generates up to approximately 0,1 TB on a clear night with a total of approximately 100 TB in current data holdings. The data are pre-processed at the telescope and then transferred to the University of Arizona and Caltech for further analysis, distribution, and archiving. The data are processed in real time, and detected transient events are published electronically through a variety of dissemination mechanisms, with no proprietary withholding period (CRTS has a completely open data policy). Further data analysis includes classification of the detected transient events, additional observations using other telescopes, scientific interpretation, and publishing. This process makes

heavy use of the archival data (several PBs) from a wide variety of geographically distributed resources connected through the virtual observatory (VO) framework.

Future:

CRTS is a scientific and methodological test bed and precursor of larger surveys to come, notably the Large Synoptic Survey Telescope (LSST), expected to operate in the 2020s and selected as the highest-priority ground-based instrument in the 2010 Astronomy and Astrophysics Decadal Survey. LSST will gather about 30 TB per night.

Survey pipelines from telescopes (on the ground or in space) produce transient event data streams, and the events, along with their observational descriptions, are ingested by one or more depositories, from which the event data can be disseminated electronically to human astronomers or robotic telescopes. Each event is assigned an evolving portfolio of information, which includes all available data on that celestial position. The data are gathered from a wide variety of data archives unified under the Virtual Observatory framework, expert annotations, etc.

Representations of such federated information can be both human-readable and machine-readable. The data are fed into one or more automated event characterization, classification, and prioritization engines that deploy a variety of machine learning tools for these tasks.

The engines' output, which evolves dynamically as new information arrives and is processed, informs the follow-up observations of the selected events, and the resulting data are communicated back to the event portfolios for the next iteration.

Users, either human or robotic, can tap into the system at multiple points, both for information retrieval and to contribute new information, through a standardized set of formats and protocols. This could be done in (near) real-time or in archival (i.e., not time-critical) modes.

5.8.2 Use case 37: DOE Extreme Data from Cosmological Sky Survey and Simulations

Application:

A cosmology discovery tool integrates simulations and observation to clarify the nature of dark matter, dark energy, and inflation—some of the most exciting, perplexing, and challenging questions facing modern physics, including the properties of fundamental particles affecting the early universe. The simulations will generate data sizes comparable to observation.

Current approach:

At this time, this project is in the preliminary planning phases and, therefore, the current approach is not fully developed.

Future:

These systems will use huge amounts of supercomputer time — over 200 million hours. Associated data sizes are as follows:

- Dark Energy Survey (DES): 4 PB per year in 2015;
- Zwicky Transient Factory (ZTF): 1 PB per year in 2015;
- LSST (see CRTS discussion above): 7 PB per year in 2019;
- Simulations: 10 PB per year in 2017.

5.8.3 Use case 38: Large Survey Data for Cosmology

Application:

For DES, the data are sent from the mountaintop, via a microwave link, to La Serena, Chile. From there, an optical link forwards them to the NCSA and to NERSC for storage and 'reduction'.

and stars in both the individual and stacked images are identified and catalogued, and finally their properties are measured and stored in a database.

Current approach:

Subtraction pipelines are run using extant imaging data to find new optical transients through machine learning algorithms. Data technologies and hardware resources are described in [A.7.3](#).

Future:

Techniques are needed for handling Cholesky decomposition for thousands of simulations with matrices of order one million on a side and parallel image storage. LSST will generate 60 PB of imaging data and 15 PB of catalogue data and a correspondingly large (or larger) amount of simulation data. In total, over 20 TB of data will be generated per night.

5.8.4 Use case 39: Particle Physics—Analysis of Large Hadron Collider Data: Discovery of Higgs Particle

Application:

Analysis is conducted on collisions at the European Organization for Nuclear Research (CERN) Large Hadron Collider (LHC) accelerator.

Processed information defines physics properties of events and generates lists of particles with type and momenta. These events are analyzed to find new effects—both new particles (e.g., Higgs), and present evidence that conjectured particles (e.g., Supersymmetry) have not been detected. A few major experiments are being conducted at LHC, including ATLAS and CMS (Compact Muon Solenoid). These experiments have global participants (e.g., CMS has 3,600 participants from 183 institutions in 38 countries), and so the data at all levels are transported and accessed across continents.

Current approach:

The LHC experiments are pioneers of a distributed big data science infrastructure. Several aspects of the LHC experiments' workflow highlight issues that other disciplines will need to solve. These issues include automation of data distribution, high-performance data transfer, and large-scale high-throughput computing. A data grid analysis for Higgs Particle discovery utilised 350,000 cores running near-continuously—over two million jobs per day arranged in three major tiers: CERN, Continents/Countries, and Universities. The analysis uses distributed, high-throughput computing (i.e., pleasing parallel) architecture with facilities integrated across the world by the Worldwide LHC Computing Grid (WLCG) and Open Science Grid in the U.S. Accelerator data and analysis generates 15 PB of data each year for a total of 200 PB. Specifically, in 2012, ATLAS had 8 PB on Tier1 tape and over 10 PB on Tier 1 disk at BNL and 12 PB on disk cache at U.S. Tier 2 centers. CMS has similar data sizes. Over half the resources are used for Monte Carlo simulations as opposed to data analysis.

Future:

In the past, the particle physics community has been able to rely on industry to deliver exponential increases in performance per unit cost over time, as described by Moore's Law. However, the available performance will be much more difficult to exploit in the future since technology limitations, in particular regarding power consumption, have led to profound changes in the architecture of modern central processing unit (CPU) chips.

In the past, software could run unchanged on successive processor generations and achieve performance gains that follow Moore's Law, thanks to the regular increase in clock rate that continued until 2006. The era of scaling sequential applications on an HEP (heterogeneous element processor) is now over. Changes in CPU architectures imply significantly more software parallelism, as well as exploitation of specialized floating point capabilities.

The structure and performance of HEP data processing software need to be changed such that they can continue to be adapted and developed to run efficiently on new hardware. This represents a major paradigm shift in HEP software design and implies large-scale re-engineering of data structures and

algorithms. Parallelism needs to be added simultaneously at all levels: the event level, the algorithm level, and the sub-algorithm level. Components at all levels in the software stack need to interoperate, and therefore the goal is to standardize as much as possible on basic design patterns and on the choice of a concurrency model. This will also help to ensure efficient and balanced use of resources.

5.8.5 Use case 40: Belle II High Energy Physics Experiment

Application:

The Belle experiment is a particle physics experiment with more than 400 physicists and engineers investigating charge parity (CP) violation effects with B meson production at the High Energy Accelerator KEKB e⁺ e⁻ accelerator in Tsukuba, Japan. In particular, numerous decay modes at the Upsilon (4S) resonance are sought to identify new phenomena beyond the standard model of particle physics. This accelerator has the largest intensity of any in the world, but the events are simpler than those from LHC, and so analysis is less complicated, but similar in style to the CERN accelerator analysis.

Current approach:

At this time, this project is in the preliminary planning phases and, therefore, the current approach is not fully developed.

Future:

An upgraded experiment Belle II and accelerator SuperKEKB was starting operation in 2015. Data will increase by a factor of 50, with total integrated raw data of ~120 PB and physics data of ~15 PB and ~100 PB of Monte Carlo samples. The next stage will necessitate a move to a distributed computing model requiring continuous raw data transfer of ~20 GB per second at designed luminosity between Japan and the United States. Required softwares are described in [A.7.5](#).

5.9 Earth, environmental, and polar science

5.9.1 Use case 41: European Incoherent Scatter Scientific Association 3D Incoherent Scatter Radar System

Application:

EISCAT conducts research on the lower, middle, and upper atmosphere and ionosphere using the incoherent scatter radar technique. This technique is the most powerful ground-based tool for these research applications. EISCAT studies instabilities in the ionosphere and investigates the structure and dynamics of the middle atmosphere. EISCAT operates a diagnostic instrument in ionospheric modification experiments with addition of a separate heating facility. Currently, EISCAT operates three of the ten major incoherent radar scattering instruments worldwide; their three systems are located in the Scandinavian sector, north of the Arctic Circle.

Current approach:

The currently running EISCAT radar generates data at rates of terabytes per year. The system does not present special challenges.

Future:

The design of the next-generation radar, EISCAT_3D, will consist of a core site with transmitting and receiving radar arrays and four sites with receiving antenna arrays at some 100 km from the core. The fully operational five-site system will generate several thousand times the number of data of the current EISCAT system, with 40 PB per year in 2022, and is expected to operate for 30 years. EISCAT_3D data e-Infrastructure plans to use high-performance computers for central site data processing and high-throughput computers for mirror site data processing. Downloading the full data is not time-critical, but operations require real-time information about certain pre-defined events, which would be sent from the sites to the operations center, and a real-time link from the operations center to the sites to set the mode of radar operation in real time.

5.9.2 Use case 42: Common Operations of Environmental Research Infrastructure

Application:

ENVRI (Common Operations of Environmental Research Infrastructures) addresses European distributed, long-term, remote-controlled observational networks focused on understanding processes, trends, thresholds, interactions, and feedbacks, as well as increasing the predictive power to address future environmental challenges. The following efforts are part of ENVRI.

- ICOS (Integrated Carbon Observation System) is a European distributed infrastructure dedicated to the monitoring of greenhouse gases (GHGs) through its atmospheric, ecosystem, and ocean networks.
- EURO-Argo is the European contribution to Argo, which is a global ocean observing system.
- EISCAT_3D (described separately) is a European new-generation incoherent scatter research radar system for upper atmospheric science.
- LifeWatch (described separately) is an e-science infrastructure for biodiversity and ecosystem research.
- EPOS (European Plate Observing System) is a European research infrastructure for earthquakes, volcanoes, surface dynamics, and tectonics.
- EMSO (European Multidisciplinary Seafloor and Water Column Observatory) is a European network of seafloor observatories for the long-term monitoring of environmental processes related to ecosystems, climate change, and geo-hazards.
- IAGOS (In-service Aircraft for a Global Observing System) is setting up a network of aircraft for global atmospheric observation.
- SIOS (Svalbard Integrated Arctic Earth Observing System) is establishing an observation system in and around Svalbard that integrates the studies of geophysical, chemical, and biological processes from all research and monitoring platforms.

Current approach:

ENVRI develops a reference model (ENVRI RM) as a common ontological framework and standard for the description and characterization of computational and storage infrastructures. The goal is to achieve seamless interoperability between the heterogeneous resources of different infrastructures. The ENVRI RM serves as a common language for community communication, providing a uniform framework into which the infrastructure's components can be classified and compared. The ENVRI RM also serves to identify common solutions to common problems. Data sizes in a given infrastructure vary from GBs to petabytes per year.

Future:

ENVRI's common environment will empower the users of the collaborating environmental research infrastructures and enable multidisciplinary scientists to access, study, and correlate data from multiple domains for system-level research. Collaboration affects big data requirements coming from interdisciplinary research.

ENVRI analyzed the computational characteristics of the six European Strategy Forum on Research Infrastructures (ESFRI) environmental research infrastructures, and identified five common subsystems. They are defined in the ENVRI RM (<http://www.envri.eu/rm>) and below.

- Data acquisition: Collects raw data from sensor arrays, various instruments, or human observers, and brings the measurements (data streams) into the system.
- Data curation: Facilitates quality control and preservation of scientific data and is typically operated at a data center.

- Data access: Enables discovery and retrieval of data housed in data resources managed by a data curation subsystem.
- Data processing: Aggregates data from various resources and provides computational capabilities and capacities for conducting data analysis and scientific experiments.
- Community support: Manages, controls, and tracks users' activities and supports users in conduct of their community roles.

5.9.3 Use case 43: Radar Data Analysis for the Center for Remote Sensing of Ice Sheets

Application:

The Center for Remote Sensing of Ice Sheets (CReSIS) effort uses custom radar systems to measure ice sheet bed depths and (annual) snow layers at the North and South Poles and mountainous regions.

Resulting data feed into the Intergovernmental Panel on Climate Change (IPCC). The radar systems are typically flown in by aircraft in multiple paths.

Current approach:

The initial analysis uses Matlab signal processing that produces a set of radar images. These cannot be transported from the field over the Internet and are typically copied onsite to a few removable disks that hold a terabyte of data, then flown to a laboratory for detailed analysis. Image features (i.e., layers) are found using image understanding tools with some human oversight. A typical echogram with detected boundaries differentiates the boundary between air and ice layers and boundary between ice and terrain. This information is stored in a database front-ended by a geographical information system. The ice sheet bed depths are used in simulations of glacier flow. Each trip into the field, usually lasting a few weeks, results in 50 to 100 TB of data.

Future:

With improved instrumentation, an order of magnitude more data (a petabyte per mission) is projected. As the increasing field data must be processed in an environment with constrained power access, low-power or low-performance architectures, such as GPU systems, are indicated.

5.9.4 Use case 44: Unmanned Air Vehicle Synthetic Aperture Radar (UAVSAR) Data Processing, Data Product Delivery, and Data Services

Application:

Synthetic aperture radar (SAR) can identify landscape changes caused by seismic activity, landslides, deforestation, vegetation changes, and flooding. This function can be used to support earthquake science as well as disaster management. This use case supports the storage, image processing application, and visualization of geo-located data with angular specification.

Current approach:

Data from planes and satellites are processed on NASA computers before being stored after substantial data communication. The data are made public upon processing. They require significant curation owing to instrumental glitches. The current data size is approximately 150 TB.

Future:

The data size would increase dramatically if Earth Radar Mission launched. Clouds are suitable hosts but are not used today in production.

5.9.5 Use case 45: NASA Langley Research Center/ Goddard Space Flight Center iRODS Federation Test Bed

Application:

NASA Center for Climate Simulation and NASA Atmospheric Science Data Center have complementary data sets, each containing vast amounts of data that are not easily shared and queried. Climate researchers, weather forecasters, instrument teams, and other scientists need to access data from across multiple datasets in order to compare sensor measurements from various instruments, compare sensor measurements to model outputs, calibrate instruments, look for correlations across multiple parameters, and more.

Current approach:

Data are generated from two products: the Modern Era Retrospective Analysis for Research and Applications (MERRA, described separately) and NASA Clouds and Earth's Radiant Energy System (CERES) EBAF-TOA (Energy Balanced and Filled-Top of Atmosphere) product, which accounts for about 420 MB, and the EBAF-Surface product, which accounts for about 690 MB. Data numbers grow with each version update (about every six months). To analyze, visualize, and otherwise process data from heterogeneous datasets is currently a time-consuming effort. Scientists must separately access, search for, and download data from multiple servers, and often the data are duplicated without an understanding of the authoritative source. Often accessing data takes longer than scientific analysis. Current datasets are hosted on modest-sized (144 to 576 cores) InfiniBand clusters.

Future:

Improved access will be enabled through the use of iRODS. These systems support parallel downloads of datasets from selected replica servers, providing users with worldwide access to the geographically dispersed servers. iRODS operation will be enhanced with semantically organized metadata and managed via a highly precise NASA Earth Science ontology. Cloud solutions will also be explored.

5.9.6 Use case 46: MERRA Analytic Services (MERRA/AS)

Application:

This application produces global temporally and spatially consistent syntheses of 26 key climate variables by combining numerical simulations with observational data. Three-dimensional results are produced every six hours extending from 1979 to the present. The data support important applications such as IPCC research and the NASA/Department of Interior RECOVER wildfire decision support system; these applications typically involve integration of MERRA with other datasets.

Current approach:

Map/Reduce is used to process a current total of 480 TB. The current system is hosted on a 36-node InfiniBand cluster.

Future:

Clouds are being investigated. The data is growing by one TB a month.

5.9.7 Use case 47: Atmospheric Turbulence – Event Discovery and Predictive Analytics

Application:

Data mining is built on top of reanalysis products, including MERRA (described separately) and the North American Regional Reanalysis (NARR), a long-term, high-resolution climate data set for the North American domain. The analytics correlate aircraft reports of turbulence (either from pilot reports or from automated aircraft measurements of eddy dissipation rates) with recently completed atmospheric reanalyses. The information is of value to aviation industry and to weather forecasters. There are no standards for reanalysis products, complicating systems for which Map/Reduce is being

investigated. The reanalysis data are hundreds of terabytes, slowly updated, whereas the turbulence dataset is smaller in size and implemented as a streaming service.

Current approach:

The current 200 TB dataset can be analyzed with Map/Reduce or the like using SciDB or another scientific database.

Future:

The dataset will reach 500 TB in five years. The initial turbulence case can be extended to other ocean/atmosphere phenomena, but the analytics would be different in each case.

5.9.8 Use case 48: Climate Studies Using the Community Earth System Model at the U.S. Department of Energy (DOE) NERSC Center

Application:

Simulations with the Community Earth System Model (CESM) can be used to understand and quantify contributions of natural and anthropogenic-induced patterns of climate variability and change in the 20th and 21st centuries. The results of supercomputer simulations across the world should be stored and compared.

Current approach:

The Earth System Grid (ESG) enables global access to climate science data on a massive scale — petascale, or even exascale — with multiple petabytes of data at dozens of federated sites worldwide. The ESG is recognized as the leading infrastructure for the management and access of large distributed data volumes for climate change research. It supports the Coupled Model Intercomparison Project (CMIP), whose protocols enable the periodic assessments carried out by the IPCC.

Future:

Rapid growth of data is expected, with 30 PB produced at NERSC (assuming 15 end-to-end climate change experiments) in 2017 and many times more than this worldwide.

5.9.9 Use case 49: DOE Biological and Environmental Research (BER) Subsurface Biogeochemistry Scientific Focus Area

Application:

A genome-enabled watershed simulation capability (GEWaSC) is needed to provide a predictive framework for understanding the following:

- how genomic information stored in a subsurface microbiome affects biogeochemical watershed functioning;
- how watershed-scale processes affect microbial functioning;
- how these interactions co-evolve.

Current approach:

Current modeling capabilities can represent processes occurring over an impressive range of scales — from a single bacterial cell to that of a contaminant plume. Data cross all scales from genomics of the microbes in the soil to watershed hydro-biogeochemistry. Data are generated by the different research areas and include simulation data, field data (e.g., hydrological, geochemical, geophysical), 'omics' data, and observations from laboratory experiments.

Future:

Little effort to date has been devoted to develop a framework for systematically connecting scales, as is needed to identify key controls and to simulate important feedbacks. GEWaSC will develop a simulation framework that formally scales from genomes to watersheds and will synthesize diverse and disparate field, laboratory, and simulation datasets across different semantic, spatial, and temporal scales.

5.9.10 Use case 50: DOE BER AmeriFlux and FLUXNET Networks

Application:

AmeriFlux and Flux Tower Network (FLUXNET) are U.S. and world collections, respectively, of sensors that observe trace gas fluxes (e.g., CO₂, water vapor) across a broad spectrum of times (e.g., hours, days, seasons, years, and decades) and space. Moreover, such datasets provide the crucial linkages among organisms, ecosystems, and process-scale studies — at climate-relevant scales of landscapes, regions, and continents—for incorporation into biogeochemical and climate models.

Current approach:

Software information is described in [A.8.10](#). There are approximately 150 towers in AmeriFlux and over 500 towers distributed globally collecting flux measurements.

Future:

Field experiment data-taking would be improved by access to existing data and automated entry of new data via mobile devices. Interdisciplinary studies integrating diverse data sources will be expanded.

5.10 Energy

5.10.1 Use case 51: Consumption Forecasting in Smart Grids

Application:

Smart meters support prediction of energy consumption for customers, transformers, substations and the electrical grid service area. Advanced meters provide measurements every 15 min at the granularity of individual consumers within the service area of smart power utilities. Data to be combined include the head end of smart meters (distributed), utility databases (customer information, network topology; centralized), U.S. Census data (distributed), NOAA weather data (distributed), micro-grid building information systems (centralized), and micro-grid sensor networks (distributed). The central theme is real-time, data-driven analytics for time series from cyber physical systems.

Current approach:

Forecasting uses GIS-based visualization. Data amount to around 4 TB per year for a city such as Los Angeles with 1,4 million sensors. There are significant privacy issues requiring anonymization by aggregation. Real-time and historic data are combined with machine learning to predict consumption. Software information is described in [A.9.1](#).

Future:

Advanced grid technologies will be widely deployed. Smart grids will have new analytics integrating diverse data and supporting curtailment requests. New technologies will support mobile applications for client interactions.

5.10.2 Use case 52: Home Energy Management System

Application:

HEMS (Home Energy Management System) is useful system for energy conservation in private homes. In the HEMS, many kinds of sensors and devices are introduced into private homes, such as, smart meter, electric vehicle, solar power panel, light, air conditioner, fuel cell, water heater, storage battery. Energy manager gathers those data generated at private homes and stores them into cloud database named the

large HEMS information platform. Information manager operates the large HEMS information platform and manages data. Privacy and security of users are responsible to Information manager. Servicer analyzes data and provides valuable information to users as a service.

Current approach:

Services provided by servicer is not restricted by monitoring service of power usage. Other examples of useful services are elderly person life watching service, appropriate energy contract plan suggestion, prediction of PV power generation, coupon incentive-based demand response.

Future:

Standardization of Application Programming Interface will be necessary to increase usefulness of HEMS data

6 Use cases derived technical considerations

Technical considerations are the challenges limiting further use of big data. After collection, processing, and review of the use cases, technical considerations within seven characteristic categories were extracted from the individual use cases. These use case specific technical considerations were then aggregated to produce high-level, general technical considerations, within the seven characteristic categories, that are vendor neutral and technology agnostic. It is emphasized that neither the use case nor the requirements lists are exhaustive.

6.1 Use case specific technical considerations

Each use case was evaluated for technical considerations within the following seven categories:

- **Data source** (e.g., data size, file formats, rate of growth, at rest or in motion);
- **Data transformation** (e.g., data fusion, analytics);
- **Capabilities** (e.g., software tools, platform tools, hardware resources such as storage and networking);
- **Data consumer** (e.g., processed results in text, table, visual, and other formats);
- **Security and privacy;**
- **Life cycle management** (e.g., curation, conversion, quality check, pre-analytic processing);
- **Other technical considerations.**

Some use cases contained technical considerations in all seven categories while others only included technical considerations for a few categories. The complete list of specific technical considerations extracted from the use cases is presented in [Annex D](#). These categories informed the eventual selection of the roles specified in ISO/IEC 20547-3.

6.2 Summary of requirements analysis

There were 35 generic technical considerations [1] summarizing 439 specific technical considerations from the 52 use cases. Column 2 of [Table 1](#) gives the number of specific technical considerations driving this generic technical considerations.

Table 1 — Generic technical considerations with count of number of motivating specific technical considerations

#	Count	Generic technical considerations
Data source considerations		
1	28	Needs to support reliable real time, asynchronous, streaming, and batch processing to collect data from centralized, distributed, and cloud data sources, sensors, or instruments.
2	22	Needs to support slow, bursty, and high-throughput data transmission between data sources and computing clusters.
3	28	Needs to support diversified data content ranging from structured and unstructured text, document, graph, web, geospatial, compressed, timed, spatial, multimedia, simulation, and instrumental data.
Transformation considerations		
1	38	Needs to support diversified compute-intensive, analytic processing, and machine learning techniques.
2	7	Needs to support batch and real-time analytic processing.
3	15	Needs to support processing large diversified data content and modeling.
4	6	Needs to support processing data in motion (streaming, fetching new content, tracking, etc.).
Capability considerations		
1	29	Needs to support legacy and advanced software packages (software).
2	17	Needs to support legacy and advanced computing platforms (platform).
3	23	Needs to support legacy and advanced distributed computing clusters, co-processors, input output (I/O) processing (infrastructure).
4	14	Needs to support elastic data transmission (networking).
5	35	Needs to support legacy, large, and advanced distributed data storage (storage).
6	13	Needs to support legacy and advanced executable programming: applications, tools, utilities, and libraries (software).
Data consumer considerations		
1	4	Needs to support fast searches (~0,1 s) from processed data with high relevancy, accuracy, and high recall.
2	16	Needs to support diversified output file formats for visualization, rendering, and reporting.
3	2	Needs to support visual layout for results presentation.
4	11	Needs to support rich user interface for access using browser, visualization tools.
5	20	Needs to support high-resolution multi-dimension layer of data visualization.
6	1	Needs to support streaming results to clients.
Security and privacy considerations		
1	32	Needs to protect and preserve security and privacy on sensitive data.
2	12	Needs to support multi-level policy-driven, sandbox, access control, authentication on protected data.
Lifecycle management considerations		
1	20	Needs to support data quality curation including pre-processing, data clustering, classification, reduction, format transformation.
2	2	Needs to support dynamic updates on data, user profiles, and links.
3	6	Needs to support data lifecycle and long-term preservation policy, including data provenance.
4	4	Needs to support data validation.
5	4	Needs to support human annotation for data validation.
6	3	Needs to support prevention of data loss or corruption.
7	1	Needs to support multi-site archival.
8	2	Needs to support persistent identifier and data traceability.
9	1	Needs to support standardizing, aggregating, and normalizing data from disparate sources.
Other considerations		

Table 1 (continued)

#	Count	Generic technical considerations
1	6	Needs to support rich user interface from mobile platforms to access processed results
2	2	Needs to support performance monitoring on analytic processing from mobile platforms
3	13	Needs to support rich visual content search and rendering from mobile platforms
4	1	Needs to support mobile device data acquisition.
5	1	Needs to support security across mobile devices

6.3 Features of use cases

Table 2 lists the number of use cases tagged by various properties. This analysis from References [2][3][4] was the basis of the use case tags.

Table 2 — Features of use cases

Abbreviation	#	Description
PP	26	Pleasingly Parallel or Map Only
MR	18	Classic MapReduce MR (add MRStat below for full count)
MRStat	7	Simple version of MR where key computations are simple reduction as found in statistical averages, such as histograms and averages
MRIter	23	Iterative MapReduce or MPI
Graph	9	Complex graph data structure needed in analysis
Fusion	11	Integrate diverse data to aid discovery/decision making; could involve sophisticated algorithms or just be a portal
Streaming	41	Some data comes in incrementally and is processed this way
Classify	30	Classification: divide data into categories
S/Q	12	Index, Search and Query
CF	4	Collaborative Filtering for recommender engines
LML	36	Local Machine Learning (Independent for each parallel entity)
GML	23	Global Machine Learning: Deep Learning, Clustering, LDA, PLSI, MDS, Large Scale Optimizations as in Variational Bayes, MCMC, Lifted Belief Propagation, Stochastic Gradient Descent, L-BFGS, Levenberg-Marquardt. Can call EGO or Exascale Global Optimization with scalable parallel algorithm
	51	Workflow: Universal, so no label
GIS	16	Geotagged data often displayed in ESRI, Microsoft Virtual Earth, Google Earth, GeoServer, etc.
HPC	5	Classic large-scale simulation of cosmos, materials, etc., generating (visualization) data
Agent	2	Simulations of models of data-defined macroscopic entities represented as agents

Using this and an extended analysis, this table was expanded [3] to give 50 tags arranged in 4 views given in Tables 3 to 6.

Table 3 — Problem Architecture View Facets of Ogres (Meta or Macro Pattern)

Pleasingly Parallel	Seen in BLAST, Protein docking, some (bio-) imagery including Local Analytics or Local Machine Learning with pleasingly parallel filtering
Classic MapReduce	Search, Index and Query and Classification algorithms like collaborative filtering
Map Collective	Seen in machine learning – especially with linear algebra kernels
Map P2P	Point to Point Communication seen in parallel simulation and graph algorithms
Map Streaming	Combination of (parallel) long running maps accepting streamed data

Table 3 (continued)

Shared Memory	As opposed to distributed data (memory). Corresponds to problem where shared memory implementations are important. Tend to be dynamic asynchronous
SPMD	Single Program Multiple Data, well-known parallel computing style
BSP	Bulk Synchronous Processing: well-defined compute-communication phases
Fusion	Knowledge discovery often involves fusion of multiple methods or sources
Dataflow	Composite structure with multiple components linked by exchanged data
Agents	As used in epidemiology, discrete event simulations, etc. Swarm approaches
Workflow	Many applications often involve orchestration (workflow) of multiple components

Table 4 — Execution Features View Facets of Ogres

Performance metrics	As measured in benchmarks
Flops per byte	Important for performance
Execution Environment	Cloud or HPC; are Core libraries needed such as matrix-matrix/vector algebra, conjugate gradient, reduction, broadcast
Volume	Data size
Velocity	Measures Streaming
Variety	Multiple data sources are often mixed. See Fusion facet
Veracity	Accuracy of data affecting pre-processing needed and reliability of answer
Communication Structure	Interconnect structure? Is communication Synchronous or Asynchronous? In latter case shared memory may be attractive;
Static or Dynamic?	Does application (graph) change during execution?
Regularity	Most applications consist of a set of interconnected entities; is this regular as a set of pixels or is it a complicated irregular graph?
Iterative or not?	Important algorithm characteristic
Data Abstraction	Key-value, pixel, graph, vector, HDF5, Bag of words, etc.
Data Space?	Are data points in metric or non-metric spaces?
Complexity	Is algorithm $O(N^2)$ or $O(N)$ (up to logs) for N points per iteration?

Table 5 — Data Source and Style View Facets of Ogres

SQL, NoSQL or NewSQL	NoSQL includes Document, Column, Key-value, Graph, Triple store
Enterprise data systems	10 examples from NIST [5] integrate SQL/NoSQL
Files or Objects	Files as managed in iRODS and extremely common in scientific research. Objects most common in ABDS
HDFS/Lustre/GPFS	Are data and compute colocated?
Archive/Batched/Streaming	Streaming is Incremental update of datasets with new algorithms to achieve real-time response
Storage system styles	Styles include Shared, Dedicated, Permanent, and Transient
Metadata/Provenance	Define overall features of data and processing
Internet of Things	24 [5] to 50 (Cisco [7][8]) billion devices on the Internet by 2020
HPC generated data	Simulations generate visualization output that often needs to be mined
GIS	Geographical Information Systems provide access to geospatial data

Table 6 — Processing or Run-time View Facets of Ogres

Micro Benchmarks	A simple kernel or mini-app used to measure core system performance
LML	Local Analytics or Local Machine Learning
GML	Global Analytics or Machine Learning requiring iterative runtime

Table 6 (continued)

Base Statistics	Simple statistics seen in Table 2 as MRStat
Recommendations	Collaborative Filtering and other recommender analytics
Search/Query/Index	Rich set of technologies used in Search, Query and Indexing data
Classification	Technologies to label data (SVM, Bayes, deep learning, clustering)
Learning	Training algorithms
Optimization Methodology	Machine Learning, Nonlinear Optimization, Least Squares, Linear/Quadratic Programming, Combinatorial Optimization, expectation maximization, Monte Carlo, Variational Bayes, Global Inference
Streaming	Growing class of fast online $O(N)$ algorithms
Alignment	Variant of Search seen in sequence comparison as in BLAST
Linear Algebra	Many machine learning algorithms build on linear algebra kernels
Graph	Problem set up as a graph as opposed to vector, grid, etc.
Visualization	Important component of many analysis pipelines

Annex A

Submitted use case studies

A.1 Government operation

A.1.1 Use case 1: Big data Archival: Census 2010 and 2000

Use case title	Big data Archival: Census 2010 and 2000 — Title 13 big data	
Vertical (area)	Digital Archives	
Author/company/email	Vivek Navale and Quyen Nguyen (NARA)	
Actors/stakeholders and their roles and responsibilities	NARA's Archivists Public users (after 75 years)	
Goals	Preserve data for a long term in order to provide access and perform analytics after 75 years. Title 13 of U.S. code authorizes the Census Bureau and guarantees that individual and industry specific data is protected.	
Use case description	Maintain data "as-is". No access and no data analytics for 75 years. Preserve the data at the bit-level. Perform curation, which includes format transformation if necessary. Provide access and analytics after nearly 75 years.	
Current solutions	Compute(System)	Linux servers
	Storage	NetApps, Magnetic tapes.
	Networking	
	Software	
Big data characteristics	Data source (distributed/centralized)	Centralized storage.
	Volume (size)	380 TB
	Velocity (e.g. real time)	Static.
	Variety (multiple datasets, mashup)	Scanned documents
	Variability (rate of change)	None
Big data science (collection, curation, analysis, action)	Veracity (Robustness Issues, semantics)	Cannot tolerate data loss.
	Visualization	TBD
	Data quality (syntax)	Unknown.
	Data types	Scanned documents
	Data analytics	Only after 75 years.
Big data specific challenges (Gaps)	Preserve data for a long time scale.	
Big data specific challenges in mobility	TBD	

Security and privacy technical considerations	Title 13 data.
Highlight issues for generalizing this Use case (e.g. for ref. architecture)	
More information (URLs)	

A.1.2 Use case 2: NARA Accession, Search, Retrieve, Preservation

Use case title	National Archives and Records Administration Accession NARA Accession, Search, Retrieve, Preservation	
Vertical (area)	Digital Archives	
Author/company/email	Quyen Nguyen and Vivek Navale (NARA)	
Actors/stakeholders and their roles and responsibilities	Agencies' Records Managers NARA's Records Accessioners NARA's Archivists Public users×	
Goals	Accession, Search, Retrieval, and Long term Preservation of big data.	
Use case description	1) Get physical and legal custody of the data. In the future, if data reside in the cloud, physical custody should avoid transferring big data from Cloud to Cloud or from Cloud to Data Center 2) Pre-process data for virus scan, identifying file format identification, removing empty files 3) Index 4) Categorize records (sensitive, unsensitive, privacy data, etc.) 5) Transform old file formats to modern formats (e.g. WordPerfect to PDF) 6) E-discovery 7) Search and retrieve to respond to special request 8) Search and retrieve of public records by public users	
Current solutions	Compute(System)	Linux servers
	Storage	NetApps, Hitachi, Magnetic tapes.
	Networking	
	Software	Custom software, commercial search products, commercial databases.
Big data characteristics	Data source (distributed/centralized)	Distributed data sources from federal agencies. Current solution requires transfer of those data to a centralized storage. In the future, those data sources may reside in different Cloud environments.
	Volume (size)	Hundreds of Terabytes, and growing.
	Velocity (e.g. real time)	Input rate is relatively low compared to other use cases, but the trend is bursty. That is the data can arrive in batches of size ranging from GB to hundreds of TB.

	Variety (multiple data-sets, mashup)	Variety data types, unstructured and structured data: textual documents, emails, photos, scanned documents, multimedia, social networks, web sites, databases, etc. Variety of application domains, since records come from different agencies. Data come from variety of repositories, some of which can be cloud-based in the future.
	Variability (rate of change)	Rate can change especially if input sources are variable, some having audio, video more, some more text, and other images, etc.
Big data science (collection, curation, analysis, action)	Veracity (Robustness Issues, semantics)	Search results should have high relevancy and high recall. Categorization of records should be highly accurate.
	Visualization	TBD
	Data quality (syntax)	Unknown.
	Data types	Variety data types: textual documents, emails, photos, scanned documents, multimedia, databases, etc.
	Data analytics	Crawl/index; search; ranking; predictive search. Data categorization (sensitive, confidential, etc.) Personally Identifiable Information (PII) data detection and flagging.
Big data specific challenges (Gaps)	Perform pre-processing and manage for long-term of large and varied data. Search huge amount of data. Ensure high relevancy and recall. Data sources may be distributed in different clouds in future.	
Big data specific challenges in mobility	Mobile search must have similar interfaces/results	
Security and privacy technical considerations	Need to be sensitive to data access restrictions.	
Highlight issues for generalizing this Use case (e.g. for ref. architecture)		
More information (URLs)		