NIST Special Publication 1500-3r1

NIST Big Data Interoperability Framework:

Volume 3, Use Cases and General Requirements

NIST Big Data Public Working Group
Use Cases and Requirements Subgroup

Version 2
June 2018

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.1500-3r1

National Institute of
Standards and Technology
U.S. Department of Commerce

NIST Special Publication 1500-3r1

NIST Big Data Interoperability Framework: Volume 3, Use Cases and General Requirements

Version 2

NIST Big Data Public Working Group (NBD-PWG)
Use Cases and Requirements Subgroup
Information Technology Laboratory
National Institute of Standards and Technology
Gaithersburg, MD 20899

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.1500-3r1

June 2018

U.S. Department of Commerce
Wilbur L. Ross, Jr., Secretary

National Institute of Standards and Technology
Walter Copan, NIST Director and Undersecretary of Commerce for Standards and Technology

National Institute of Standards and Technology (NIST) Special Publication 1500-3r1
Natl. Inst. Stand. Technol. Spec. Publ. 1500-3r1, 334 pages (June 2018) CODEN: NSPUE2

This publication is available free of charge from: https://doi.org/10.6028/NIST.SP.1500-3r1

NIST Special Publication series 1500 is intended to capture external perspectives related to NIST standards, measurement, and testing-related efforts. These external perspectives can come from industry, academia, government, and others. These reports are intended to document external perspectives and do not represent official NIST positions.

Comments on this publication may be submitted to Wo Chang

National Institute of Standards and Technology
Attn: Wo Chang, Information Technology Laboratory
100 Bureau Drive (Mail Stop 8900) Gaithersburg, MD 20899-8930
Email: SP1500comments@nist.gov

Reports on Computer Systems Technology

The Information Technology Laboratory (ITL) at NIST promotes the U.S. economy and public welfare by providing technical leadership for the Nation's measurement and standards infrastructure. ITL develops tests, test methods, reference data, proof of concept implementations, and technical analyses to advance the development and productive use of information technology. ITL's responsibilities include the development of management, administrative, technical, and physical standards and guidelines for the cost-effective security and privacy of other than national security-related information in federal information systems. This document reports on ITL's research, guidance, and outreach efforts in Information Technology and its collaborative activities with industry, government, and academic organizations.

Abstract

Big Data is a term used to describe the large amount of data in the networked, digitized, sensor-laden, information-driven world. While opportunities exist with Big Data, the data can overwhelm traditional technical approaches and the growth of data is outpacing scientific and technological advances in data analytics. To advance progress in Big Data, the NIST Big Data Public Working Group (NBD-PWG) is working to develop consensus on important fundamental concepts related to Big Data. The results are reported in the *NIST Big Data Interoperability Framework* series of volumes. This volume, Volume 3, contains the original 51 Version 1 use cases gathered by the NBD-PWG Use Cases and Requirements Subgroup and the requirements generated from those use cases. The use cases are presented in their original and summarized form. Requirements, or challenges, were extracted from each use case, and then summarized over all the use cases. These generalized requirements were used in the development of the NIST Big Data Reference Architecture (NBDRA), which is presented in Volume 6. Currently, the subgroup is accepting additional use case submissions using the more detailed Use Case Template 2. The Use Case Template 2 and the two Version 2 use cases collected to date are presented and summarized in this volume.

Keywords

Big Data; Big Data Application Provider; Big Data characteristics; Big Data Framework Provider; Big Data taxonomy; Data Consumer; Data Provider; data science; Management Fabric; reference architecture; Security and Privacy Fabric; System Orchestrator; use cases.

Acknowledgements

This document reflects the contributions and discussions by the membership of the NBD-PWG, co-chaired by Wo Chang (NIST ITL), Bob Marcus (ET-Strategies), and Chaitan Baru (San Diego Supercomputer Center; National Science Foundation). For all versions, the Subgroups were led by the following people: Nancy Grady (SAIC), Natasha Balac (San Diego Supercomputer Center), and Eugene Luster (R2AD) for the Definitions and Taxonomies Subgroup; Geoffrey Fox (Indiana University) and Tsegereda Beyene (Cisco Systems) for the Use Cases and Requirements Subgroup; Arnab Roy (Fujitsu), Mark Underwood (Krypton Brothers; Synchrony Financial), and Akhil Manchanda (GE) for the Security and Privacy Subgroup; David Boyd (InCadence Strategic Solutions), Orit Levin (Microsoft), Don Krapohl (Augmented Intelligence), and James Ketner (AT&T) for the Reference Architecture Subgroup; and Russell Reinsch (Center for Government Interoperability), David Boyd (InCadence Strategic Solutions), Carl Buffington (Vistronix), and Dan McClary (Oracle), for the Standards Roadmap Subgroup.

The editors for this document were the following:

- ***Version 1***: Geoffrey Fox (Indiana University) and Wo Chang (NIST)
- ***Version 2***: Geoffrey Fox (Indiana University) and Wo Chang (NIST)

Laurie Aldape (Energetics Incorporated) and Elizabeth Lennon (NIST) provided editorial assistance across all NBDIF volumes.

NIST SP1500-3, Version 2 has been collaboratively authored by the NBD-PWG. As of the date of this publication, there are over six hundred NBD-PWG participants from industry, academia, and government. Federal agency participants include the National Archives and Records Administration (NARA), National Aeronautics and Space Administration (NASA), National Science Foundation (NSF), and the U.S. Departments of Agriculture, Commerce, Defense, Energy, Health and Human Services, Homeland Security, Transportation, Treasury, and Veterans Affairs.

NIST would like to acknowledge the specific contributions[1] to this volume, during Version 1 and/or Version 2 activities, by the following NBD-PWG members:

Tsegereda Beyene
Cisco Systems

Deborah Blackstock
MITRE Corporation

David Boyd
InCadence Strategic Services

Scott Brim
Internet2

Pw Carey
Compliance Partners, LLC

Wo Chang
NIST

Marge Cole
SGT, Inc.

Nancy Grady
SAIC

Jay Greenberg
The Boeing Company

Karen Guertler
Consultant

Keith Hare
JCC Consulting, Inc.

Babak Jahromi
Microsoft

Pavithra Kenjige
PK Technologies

Donald Krapohl
Augmented Intelligence

Ashok Malhotra
Oracle Corporation

Robert Marcus
ET-Strategies

Gary Mazzaferro
AlloyCloud, Inc.

William Miller
MaCT USA

Sanjay Mishra
Verizon

Doug Scrimager
Slalom Consulting

Cherry Tom
IEEE-SA

[1] "Contributors" are members of the NIST Big Data Public Working Group who dedicated great effort to prepare and gave substantial time on a regular basis to research and development in support of this document.

Yuri Demchenko
University of Amsterdam

Safia Djennane
Cloud-Age-IT

Geoffrey Fox
Indiana University

Luca Lepori
Data Hold

Orit Levin
Microsoft

Eugene Luster
DISA/R2AD

Wilco van Ginkel
Verizon

Timothy Zimmerlin
Consultant

Alicia Zuniga-Alvarado
Consultant

TABLE OF CONTENTS

Figures

Tables

EXECUTIVE SUMMARY

The *NIST Big Data Interoperability Framework* consists of nine volumes, each of which addresses a specific key topic, resulting from the work of the NBD-PWG. The nine volumes are:

- Volume 1, Definitions [1]
- Volume 2, Taxonomies [2]
- Volume 3, Use Cases and General Requirements (this volume)
- Volume 4, Security and Privacy [3]
- Volume 5, Architectures White Paper Survey [4]
- Volume 6, Reference Architecture [5]
- Volume 7, Standards Roadmap [6]
- Volume 8: Reference Architecture Implementation [7]
- Volume 9: Adoption and Modernization [8]

The *NIST Big Data Interoperability Framework* will be released in three versions, which correspond to the three development stages of the NBD-PWG work. The three stages aim to achieve the following with respect to the NIST Big Data Reference Architecture (NBDRA).

Stage 1: Identify the high-level Big Data reference architecture key components, which are technology-, infrastructure-, and vendor-agnostic;
Stage 2: Define general interfaces between the NBDRA components; and
Stage 3: Validate the NBDRA by building Big Data general applications through the general interfaces.

The *NIST Big Data Interoperability Framework: Volume 3, Use Cases and General Requirements* document was prepared by the NIST Big Data Public Working Group (NBD-PWG) Use Cases and Requirements Subgroup to document the collection of use cases and extraction of requirements. The Subgroup developed the first use case template with 26 fields that were completed by 51 users in the following broad areas:

- Government Operations (4)
- Commercial (8)
- Defense (3)
- Healthcare and Life Sciences (10)
- Deep Learning and Social Media (6)
- The Ecosystem for Research (4)
- Astronomy and Physics (5)
- Earth, Environmental and Polar Science (10)
- Energy (1)

The use cases are, of course, only representative, and do not encompass the entire spectrum of Big Data usage. All the use cases were openly submitted and no significant editing was performed. While there are differences between the use cases in scope and interpretation, the benefits of free and open submission outweighed those of greater uniformity. The Use Cases and Requirements Subgroup examined the use cases, extracted specific and general requirements, and provided input to the other subgroups to inform their work as documented in the other NBDIF Volumes.

During the development of version 2 of the NBDIF, the Use Cases and Requirements Subgroup and the Security and Privacy Subgroup identified the need for additional use cases to strengthen the future work of the NBD-PWG. These two subgroups collaboratively created the Use Case Template 2, which is currently being used to collect additional use cases. The first two Version 2 use cases are presented in this document and belong to the "Earth, Environmental and Polar Science" application domain. To submit a use case, please fill out the PDF form (https://bigdatawg.nist.gov/_uploadfiles/M0621_v2_7345181325.pdf) and email it to Wo Chang (wchang@nist.gov). Use cases will be evaluated as they are submitted and will be accepted until the end of Phase 3 work.

This volume documents the process used by the Subgroup to collect the 51 use cases and extract requirements to form the NIST Big Data Reference Architecture (NBDRA). Included in this document are summaries of the 51 Version 1 use cases, extracted requirements, the original, unedited 51 Version 1 use cases, the questions contained in Use Case Template 2 and the two Version 2 use cases submitted to date. Potential areas of future work for the Subgroup during stage 3 are highlighted in Section 1.5 of this volume. The current effort documented in this volume reflects concepts developed within the rapidly evolving field of Big Data.

1 INTRODUCTION

1.1 BACKGROUND

There is broad agreement among commercial, academic, and government leaders about the remarkable potential of Big Data to spark innovation, fuel commerce, and drive progress. Big Data is the common term used to describe the deluge of data in today's networked, digitized, sensor-laden, and information-driven world. The availability of vast data resources carries the potential to answer questions previously out of reach, including the following:

- How can a potential pandemic reliably be detected early enough to intervene?
- Can new materials with advanced properties be predicted before these materials have ever been synthesized?
- How can the current advantage of the attacker over the defender in guarding against cyber-security threats be reversed?

There is also broad agreement on the ability of Big Data to overwhelm traditional approaches. The growth rates for data volumes, speeds, and complexity are outpacing scientific and technological advances in data analytics, management, transport, and data user spheres.

Despite widespread agreement on the inherent opportunities and current limitations of Big Data, a lack of consensus on some important fundamental questions continues to confuse potential users and stymie progress. These questions include the following:

- How is Big Data defined?
- What attributes define Big Data solutions?
- What is new in Big Data?
- What is the difference between Big Data and *bigger data* that has been collected for years?
- How is Big Data different from traditional data environments and related applications?
- What are the essential characteristics of Big Data environments?
- How do these environments integrate with currently deployed architectures?
- What are the central scientific, technological, and standardization challenges that need to be addressed to accelerate the deployment of robust, secure Big Data solutions?

Within this context, on March 29, 2012, the White House announced the Big Data Research and Development Initiative. [9] The initiative's goals include helping to accelerate the pace of discovery in science and engineering, strengthening national security, and transforming teaching and learning by improving analysts' ability to extract knowledge and insights from large and complex collections of digital data.

Six federal departments and their agencies announced more than $200 million in commitments spread across more than 80 projects, which aim to significantly improve the tools and techniques needed to access, organize, and draw conclusions from huge volumes of digital data. The initiative also challenged industry, research universities, and nonprofits to join with the federal government to make the most of the opportunities created by Big Data.

Motivated by the White House initiative and public suggestions, the National Institute of Standards and Technology (NIST) has accepted the challenge to stimulate collaboration among industry professionals to further the secure and effective adoption of Big Data. As one result of NIST's Cloud and Big Data Forum held on January 15–17, 2013, there was strong encouragement for NIST to create a public working group for the development of a Big Data Standards Roadmap. Forum participants noted that this roadmap should define and prioritize Big Data requirements, including interoperability, portability, reusability, extensibility, data usage, analytics, and technology infrastructure. In doing so, the roadmap would accelerate the adoption of the most secure and effective Big Data techniques and technology.

On June 19, 2013, the NIST Big Data Public Working Group (NBD-PWG) was launched with extensive participation by industry, academia, and government from across the nation. The scope of the NBD-PWG involves forming a community of interests from all sectors—including industry, academia, and government—with the goal of developing consensus on definitions, taxonomies, secure reference architectures, security and privacy, and, from these, a standards roadmap. Such a consensus would create a vendor-neutral, technology- and infrastructure-independent framework that would enable Big Data stakeholders to identify and use the best analytics tools for their processing and visualization requirements on the most suitable computing platform and cluster, while also allowing added value from Big Data service providers.

The *NIST Big Data Interoperability Framework* (NBDIF) will be released in three versions, which correspond to the three stages of the NBD-PWG work. The three stages aim to achieve the following with respect to the NIST Big Data Reference Architecture (NBDRA).

Stage 1: Identify the high-level Big Data reference architecture key components, which are technology, infrastructure, and vendor agnostic;
Stage 2: Define general interfaces between the NBDRA components; and
Stage 3: Validate the NBDRA by building Big Data general applications through the general interfaces.

On September 16, 2015, seven NBDIF Version 1 volumes were published (http://bigdatawg.nist.gov/V1_output_docs.php), each of which addresses a specific key topic, resulting from the work of the NBD-PWG. The seven volumes are as follows:

- Volume 1, Definitions [1]
- Volume 2, Taxonomies [2]
- Volume 3, Use Cases and General Requirements (this volume)
- Volume 4, Security and Privacy [3]
- Volume 5, Architectures White Paper Survey [4]
- Volume 6, Reference Architecture [5]
- Volume 7, Standards Roadmap [6]

Currently, the NBD-PWG is working on Stage 2 with the goals to enhance the Version 1 content, define general interfaces between the NBDRA components by aggregating low-level interactions into high-level general interfaces, and demonstrate how the NBDRA can be used. As a result of the Stage 2 work, the following two additional NBDIF volumes have been developed.

- Volume 8, Reference Architecture Interfaces [7]
- Volume 9, Adoption and Modernization [8]

Version 2 of the NBDIF volumes, resulting from Stage 2 work, can be downloaded from the NBD-PWG website (https://bigdatawg.nist.gov/V2_output_docs.php). Potential areas of future work for each volume during Stage 3 are highlighted in Section 1.5 of each volume. The current effort documented in this volume reflects concepts developed within the rapidly evolving field of Big Data.

1.2 SCOPE AND OBJECTIVES OF THE USE CASES AND REQUIREMENTS SUBGROUP

This volume was prepared by the NBD-PWG Use Cases and Requirements Subgroup. The effort focused on forming a community of interest from industry, academia, and government, with the goal of developing a consensus list of Big Data requirements across all stakeholders. This included gathering and understanding various use cases from nine diversified areas (i.e., application domains.) To achieve this goal, the Subgroup completed the following tasks:

- Gathered input from all stakeholders regarding Big Data requirements;
- Analyzed and prioritized a list of challenging use case specific requirements that may delay or prevent adoption of Big Data deployment;
- Developed a comprehensive list of generalized Big Data requirements;
- Collaborated with the NBD-PWG Reference Architecture Subgroup to provide input for the NBDRA;
- Collaborated with the NBD-PWG Security and Privacy Subgroup to produce the Use Case Template 2, which will help gather valuable input to strengthen future work of the NBD-PWG; and
- Documented the findings in this report.

1.3 REPORT PRODUCTION

Version 1 of this report was produced using an open collaborative process involving weekly telephone conversations and information exchange using the NIST document system. The 51 Version 1 use cases, included herein, came from Subgroup members participating in the calls and from other interested parties informed of the opportunity to contribute.

The outputs from the use case process are presented in this report and online at the following locations:

- Index to all use cases: https://bigdatawg.nist.gov/usecases.php
- List of specific requirements versus use case: https://bigdatawg.nist.gov/uc_reqs_summary.php
- List of general requirements versus architecture component: https://bigdatawg.nist.gov/uc_reqs_gen.php
- List of general requirements versus architecture component with record of use cases giving requirements: https://bigdatawg.nist.gov/uc_reqs_gen_ref.php
- List of architecture components and specific requirements plus use case constraining the components: https://bigdatawg.nist.gov/uc_reqs_gen_detail.php
- General requirements: https://bigdatawg.nist.gov/uc_reqs_gen.php.

During development of version 2 of this report, the subgroup focused on preparing the revised Use Case Template 2 (an outline of which is provided in Appendix E) and collaborating with other subgroups on content development for the other NBDIF volumes.

To achieve technical and high-quality document content, this document will go through a public comments period along with NIST internal review.

1.4 REPORT STRUCTURE

Following this introductory section, the remainder of this document is organized as follows:

- Section 2 presents the original (version 1) 51 use cases and 2 new use cases gotten with updated version 2 summary.

  - Section 2.1 discusses the process that led to their production. of the use cases.
  - Sections 2.2 through 2.10 provide summaries of the 53 use cases; each summary has three subsections: Application, Current Approach, and Future. The use cases are organized into the nine broad areas (application domains) listed below, with the number of associated use cases in parentheses:
    - Government Operation (4)
    - Commercial (8)
    - Defense (3)
    - Healthcare and Life Sciences (10)
    - Deep Learning and Social Media (6)
    - The Ecosystem for Research (4)
    - Astronomy and Physics (5)
    - Earth, Environmental, and Polar Science (10) plus 2 additional version 2 use cases (12 total)
    - Energy (1)
- Section 3 presents a more detailed analysis of requirements across use cases.
- Section 4 introduces the version 2 use cases.
- Appendix A contains the original, unedited use cases.
- Appendix B summarizes key properties of each use case.
- Appendix C presents a summary of use case requirements.
- Appendix D provides the requirements extracted from each use case and aggregated general requirements grouped by characterization category.
- Appendix E presents the structure of the revised Use Case Template 2. The fillable pdf can be downloaded from https://bigdatawg.nist.gov/_uploadfiles/M0621_v2_7345181325.pdf.
- Appendix F contains the Version 2 use cases.
- Appendix G contains acronyms and abbreviations used in this document.
- Appendix H supplies the document references.

1.5 FUTURE WORK ON THIS VOLUME

The revised Use Case Template 2, developed during phase 2, contains enhanced, comprehensive coverage of various topics, which aim to increase the depth of insight gained from submitted use cases. Use cases will be accepted by the NBD-PWG on a continuous basis until the end of Phase 3. To submit a use case, please fill out the PDF form (https://bigdatawg.nist.gov/_uploadfiles/M0621_v2_7345181325.pdf) and email it to Wo Chang (wchang@nist.gov). The NBD-PWG will evaluate additional use cases as they are submitted, to extract information that will strengthen and shape the content of version 3 of NBDIF documents.

2 USE CASE SUMMARIES

2.1 USE CASE DEVELOPMENT PROCESS

A *use case* is a typical application stated at a high level for the purposes of extracting requirements or comparing usages across fields. In order to develop a consensus list of Big Data requirements across all stakeholders, the Subgroup began by collecting use cases. Publicly available information was collected for various Big Data architecture examples with special attention given to some areas including Healthcare and Government. After collection of 51 use cases, nine broad areas (i.e., application domains) were identified by the Subgroup members to better organize the collection of use cases. The list of application domains reflects the use cases submitted and is not intended to be exhaustive. If other application domains are proposed, they will be considered. Each example of Big Data architecture constituted one use case. The nine application domains were as follows:

- Government Operation;
- Commercial;
- Defense;
- Healthcare and Life Sciences;
- Deep Learning and Social Media;
- The Ecosystem for Research;
- Astronomy and Physics;
- Earth, Environmental, and Polar Science; and
- Energy.

As noted above, participants in the NBD-PWG Use Cases and Requirements Subgroup and other interested parties supplied the information for the use cases. The template used to collect use case information and provided at the front of Appendix A, was valuable for gathering consistent information that enabled the Subgroup to develop supporting analysis and comparison of the use cases. However, varied levels of detail and quantitative or qualitative information were received for each use case template section. The original, unedited use cases are also included in Appendix A and may be downloaded from the NIST document library (http://bigdatawg.nist.gov/usecases.php).

Beginning with Section 2.2 below, the following information is presented for each Big Data use case:

- Application: a high-level description of the use case;
- Current approach: the current manifestation of the use case; and
- Future: desired computational environment, if submitted.

For some application domains, several similar Big Data use cases are presented, providing a more complete view of Big Data requirements within that application domain.

The use cases are presented in this section with the information originally submitted. The original content has not been modified. 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 NBD-PWG. The front matter (page ii) contains a general disclaimer. The use cases are numbered sequentially to facilitate cross-referencing between the use case summaries presented in this section, the original use cases (Appendix A), and the use case summary tables (Appendices B, C, and D).

2.2 GOVERNMENT OPERATION

2.2.1 USE CASE 1: CENSUS 2010 AND 2000—TITLE 13 BIG DATA

Submitted by Vivek Navale and Quyen Nguyen, National Archives and Records Administration (NARA)

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 terabytes (TB) of scanned documents.

FUTURE

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

2.2.2 USE CASE 2: NARA ACCESSION, SEARCH, RETRIEVE, PRESERVATION

Submitted by Vivek Navale and Quyen Nguyen, NARA

APPLICATION

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

CURRENT APPROACH

The data are currently handled as follows:

1. Get physical and legal custody of the data
2. Pre-process data for conducting virus scans, identifying file format identifications, and removing empty files
3. Index the data
4. Categorize records (e.g., sensitive, non-sensitive, privacy data)
5. Transform old file formats to modern formats (e.g., WordPerfect to PDF)
6. Conduct e-discovery
7. Search and retrieve to respond to special requests
8. Search and retrieve public records by public users

Currently 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.

2.2.3 USE CASE 3: STATISTICAL SURVEY RESPONSE IMPROVEMENT

Submitted by Cavan Capps, U.S. Census Bureau

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 petabyte (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. Software used includes Hadoop, Spark, Hive, R, SAS, Mahout, Allegrograph, MySQL, Oracle, Storm, BigMemory, Cassandra, and Pig.

FUTURE

Improved recommendation systems are needed similar to those used in e-commerce (e.g., similar to the Netflix use case) 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.

2.2.4 USE CASE 4: NON-TRADITIONAL DATA IN STATISTICAL SURVEY RESPONSE IMPROVEMENT (ADAPTIVE DESIGN)

Submitted by Cavan Capps, U.S. Census Bureau

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. 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.

2.3 COMMERCIAL

2.3.1 USE CASE 5: CLOUD ECO-SYSTEM FOR FINANCIAL INDUSTRIES

Submitted by Pw Carey, Compliance Partners, LLC

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 and Hadoop 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.

2.3.2 USE CASE 6: MENDELEY—AN INTERNATIONAL NETWORK OF RESEARCH

Submitted by William Gunn, Mendeley

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. Processing takes place on Amazon Web Services (AWS) using the following software: Hadoop, Scribe, Hive, Mahout, and Python. 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 Hadoop 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.

RESOURCES

- Mendeley. http://mendeley.com. Accessed March 3, 2015.
- Mendeley. http://dev.mendeley.com. Accessed March 3, 2015.

2.3.3 USE CASE 7: NETFLIX MOVIE SERVICE

Submitted by Geoffrey Fox, Indiana University

APPLICATION

Netflix 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

Netflix 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. Netflix systems use SQL, NoSQL, and Map/Reduce on AWS. Netflix recommendation systems have features in common with e-commerce systems such as Amazon.com. Streaming video has features in common with other content-providing services such as iTunes, Google Play, Pandora, and Last.fm. Business initiatives such as Netflix-sponsored content have been used to increase viewership.

FUTURE

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

RESOURCES

- Building Large-scale Real-world Recommender Systems - Recsys2012 tutorial. http://www.slideshare.net/xamat/building-largescale-realworld-recommender-systems-recsys2012-tutorial. Accessed March 3, 2015.
- RAD – Outlier Detection on Big Data. http://techblog.netflix.com/. Accessed March 3, 2015.

2.3.4 USE CASE 8: WEB SEARCH

Submitted by Geoffrey Fox, Indiana University

APPLICATION

A web search function returns results in ≈0.1 seconds 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:

1. Crawl the web
2. Pre-process data to identify what is searchable (words, positions)
3. Form an inverted index, which maps words to their locations in documents
4. Rank the relevance of documents using the PageRank algorithm
5. Employ advertising technology, e.g., using reverse engineering to identify ranking models—or preventing reverse engineering
6. Cluster documents into topics (as in Google News)
7. 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.

RESOURCES

- Internet Trends D11 Conference. http://www.slideshare.net/kleinerperkins/kpcb-Internet-trends-2013. Accessed March 3, 2015.
- Introduction to Search Engine Technology. http://webcourse.cs.technion.ac.il/236621/Winter2011-2012/en/ho_Lectures.html. Accessed March 3, 2015.
- Lecture "Information Retrieval and Web Search Engines" (SS 2011). http://www.ifis.cs.tu-bs.de/teaching/ss-11/irws. Accessed March 3, 2015.
- Recommender Systems Tutorial (Part 1) –Introduction. http://www.slideshare.net/beechung/recommender-systems-tutorialpart1intro. Accessed March 3, 2015.
- The size of the World Wide Web (The Internet). http://www.worldwidewebsize.com/. Accessed March 3, 2015.

2.3.5 USE CASE 9: BIG DATA BUSINESS CONTINUITY AND DISASTER RECOVERY WITHIN A CLOUD ECO-SYSTEM

Submitted by Pw Carey, Compliance Partners, LLC

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). There are some relevant technologies from VMware, NetApps, Oracle, IBM, and Brocade. 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.

RESOURCES

- Disaster Recovery. http://www.disasterrecovery.org/. Accessed March 3, 2015.

2.3.6 USE CASE 10: CARGO SHIPPING

Submitted by William Miller, MaCT USA

APPLICATION

Delivery companies such as Federal Express, United Parcel Service (UPS), and DHL 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. Figure 1 provides an architectural diagram.

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 under development (specifically International Organization for Standardization [ISO] standard 29161) from the ISO Joint Technical Committee 1, Subcommittee 31, Working Group 2, which develops technical standards for data structures used for automatic identification applications.

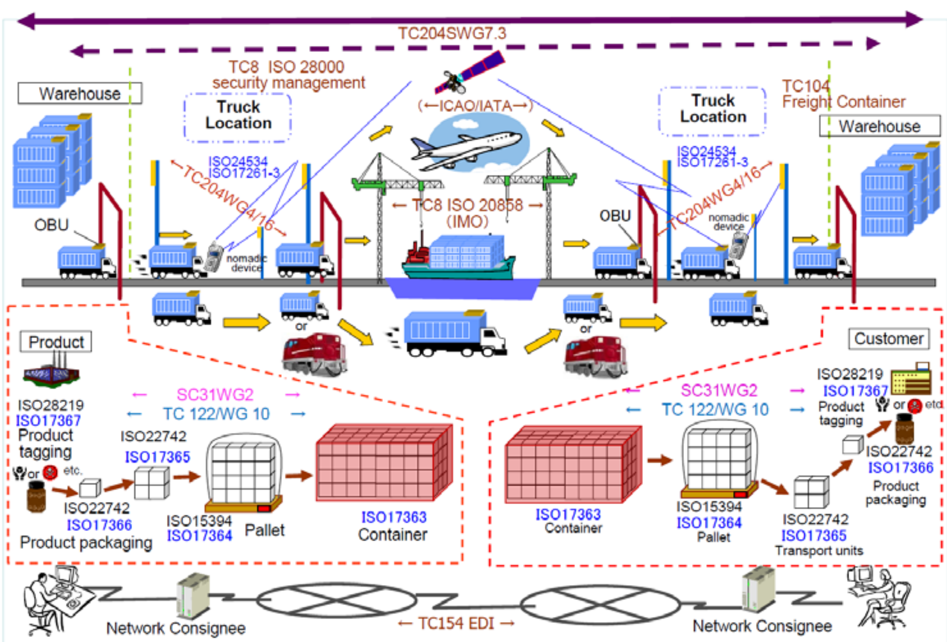

Figure 1: Cargo Shipping Scenario

2.3.7 USE CASE 11: MATERIALS DATA FOR MANUFACTURING

Submitted by John Rumble, R&R Data Services

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., 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.

RESOURCES

- The Materials Project. http://www.materialsproject.org. Accessed March 3, 2015.

2.3.8 USE CASE 12: SIMULATION-DRIVEN MATERIALS GENOMICS

Submitted by David Skinner, Lawrence Berkeley National Laboratory (LBNL)

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, 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.

RESOURCES

- The Materials Project. http://www.materialsproject.org. Accessed March 3, 2015.

2.4 DEFENSE

2.4.1 USE CASE 13: CLOUD LARGE-SCALE GEOSPATIAL ANALYSIS AND VISUALIZATION

Submitted by David Boyd, Data Tactics

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 NITF, GeoTiff, and 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 includes a server with a geospatially enabled RDBMS, geospatial server/analysis software (ESRI ArcServer or Geoserver), and visualization (either browser-based or using the ArcMap application).

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).

RESOURCES

- OGC® Standards and Supporting Documents. http://www.opengeospatial.org/standards. Accessed March 3, 2015.
- GeoJSON. http://geojson.org/. Accessed March 3, 2015.
- Compressed ARC Digitized Raster Graphics (CADRG). http://earth-info.nga.mil/publications/specs/printed/CADRG/cadrg.html. Accessed March 3, 2015.

2.4.2 USE CASE 14: OBJECT IDENTIFICATION AND TRACKING FROM WIDE-AREA LARGE FORMAT IMAGERY OR FULL MOTION VIDEO—PERSISTENT SURVEILLANCE

Submitted by David Boyd, Data Tactics

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, 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 1080p resolution (i.e., 1920 x 1080 pixels, a high-definition progressive scan) or Wide-Area Large Format Imagery (WALF)—1 to 10 frames per second at 10,000 pixels x 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.

RESOURCES

- Persistent surveillance relies on extracting relevant data points and connecting the dots. http://www.militaryaerospace.com/topics/m/video/79088650/persistent-surveillance-relies-on-extracting-relevant-data-points-and-connecting-the-dots.htm. Accessed March 3, 2015.
- Wide Area Persistent Surveillance Revolutionizes Tactical ISR. http://www.defencetalk.com/wide-area-persistent-surveillance-revolutionizes-tactical-isr-45745/. Accessed March 3, 2015.

2.4.3 USE CASE 15: INTELLIGENCE DATA PROCESSING AND ANALYSIS

Submitted by David Boyd, Data Tactics

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

Software includes Hadoop, Accumulo (Big Table), Solr, natural language processing (NLP), Puppet (for deployment and security), and Storm running on medium-size clusters. 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.

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.

RESOURCES

- Program Overview: AFCEA Aberdeen Chapter Luncheon March 14th, 2012. http://www.afcea-aberdeen.org/files/presentations/AFCEAAberdeen_DCGSA_COLWells_PS.pdf. Accessed March 3, 2015.
- Horizontal Integration of Warfighter Intelligence Data: A Shared Semantic Resource for the Intelligence Community. http://stids.c4i.gmu.edu/papers/STIDSPapers/STIDS2012_T14_SmithEtAl_HorizontalIntegrationOfWarfighterIntel.pdf. Accessed March 3, 2015.
- Integration of Intelligence Data through Semantic Enhancement. http://stids.c4i.gmu.edu/STIDS2011/papers/STIDS2011_CR_T1_SalmenEtAl.pdf. Accessed March 3, 2015.
- DCGSA Standard Cloud. http://www.youtube.com/watch?v=l4Qii7T8zeg. Accessed March 3, 2015.
- Distributed Common Ground System – Army. http://dcgsa.apg.army.mil/. Accessed March 3, 2015.

2.5 HEALTH CARE AND LIFE SCIENCES

2.5.1 USE CASE 16: ELECTRONIC MEDICAL RECORD DATA

Submitted by Shaun Grannis, Indiana University

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, 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.

RESOURCES

- A universal code system for tests, measurements, and observations. http://loinc.org/. Accessed March 3, 2015.

2.5.2 USE CASE 17: PATHOLOGY IMAGING/DIGITAL PATHOLOGY

Submitted by Fusheng Wang, Emory University

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. Figure 2 presents examples of two- and three-dimensional (2D and 3D) pathology images.

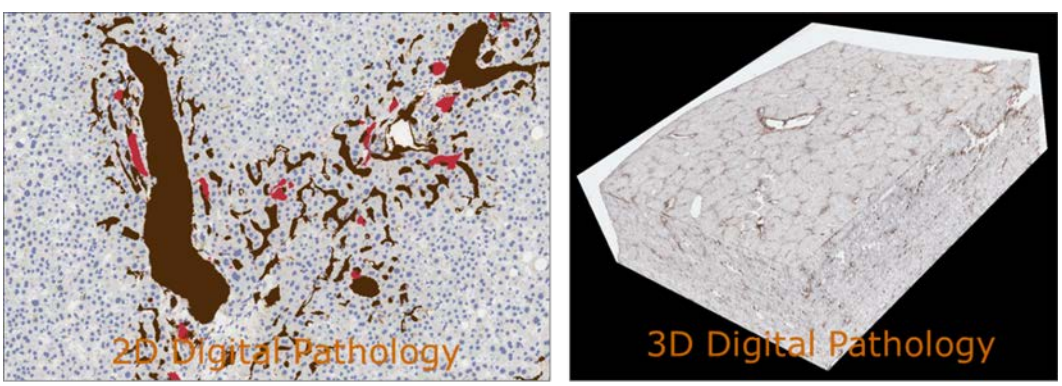

Figure 2: Pathology Imaging/Digital Pathology—Examples of 2-D and 3-D Pathology Images

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. Data processing happens with Map/Reduce (a data processing program) and Hive (to abstract the Map/Reduce program and support data warehouse interactions), along with spatial extension on supercomputers and clouds. GPUs are used effectively for image creation. Figure 3 shows the architecture of Hadoop-GIS, a spatial data warehousing system, over Map/Reduce to support spatial analytics for analytical pathology imaging.

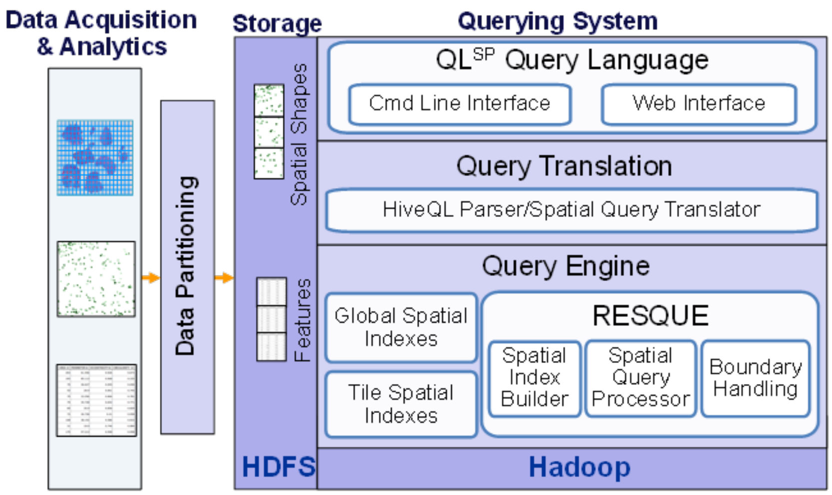

Figure 3: Pathology Imaging/Digital Pathology

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.

RESOURCES

- Pathology Analytical Imaging Standards. http://openpais.org. Accessed March 3, 2015.
- Hadoop-GIS: Spatial Big Data Solutions. http://hadoopgis.org/. Accessed March 3, 2015.

2.5.3 USE CASE 18: COMPUTATIONAL BIOIMAGING

Submitted by David Skinner, Joaquin Correa, Daniela Ushizima, and Joerg Meyer, LBNL

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 datasets. 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. Suggested software packages are ImageJ, OMERO, VolRover, and advanced segmentation and feature detection software.

2.5.4 USE CASE 19: GENOMIC MEASUREMENTS

Submitted by Justin Zook, National Institute of Standards and Technology

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.

RESOURCES

- Genome in a Bottle Consortium. http://www.genomeinabottle.org. Accessed March 3, 2015.

2.5.5 USE CASE 20: COMPARATIVE ANALYSIS FOR METAGENOMES AND GENOMES

Submitted by Ernest Szeto, LBNL, Joint Genome Institute

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; and
- 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.

RESOURCES

- IMG Data Management. http://img.jgi.doe.gov. Accessed March 3, 2015.

2.5.6 USE CASE 21: INDIVIDUALIZED DIABETES MANAGEMENT

Submitted by Ying Ding, Indiana University

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 process would use efficient parallel retrieval algorithms, suitable for cloud or high-performance computing (HPC), using the open source Hbase database with both indexed and custom search capability to identify patients of possible interest. The Semantic Linking for Property Values method would be used to convert an existing data warehouse at Mayo Clinic, called the Enterprise Data Trust (EDT), into RDF triples that enable one to find similar patients through linking of both vocabulary-based and continuous values. The time-dependent properties need to be processed before query to allow matching based on derivatives and other derived properties.

2.5.7 USE CASE 22: STATISTICAL RELATIONAL ARTIFICIAL INTELLIGENCE FOR HEALTH CARE

Submitted by Sriram Natarajan, Indiana University

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.

2.5.8 USE CASE 23: WORLD POPULATION-SCALE EPIDEMIOLOGICAL STUDY

Submitted by Madhav Marathe, Stephen Eubank, and Chris Barrett, Virginia Tech

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.

2.5.9 USE CASE 24: SOCIAL CONTAGION MODELING FOR PLANNING, PUBLIC HEALTH, AND DISASTER MANAGEMENT

Submitted by Madhav Marathe and Chris Kuhlman, Virginia Tech

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.

2.5.10 USE CASE 25: BIODIVERSITY AND LIFEWATCH

Submitted by Wouter Los and Yuri Demchenko, University of Amsterdam

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 [CO_2] fluxes, algal blooming, water and soil characteristics.)

2.6 DEEP LEARNING AND SOCIAL MEDIA

2.6.1 USE CASE 26: LARGE-SCALE DEEP LEARNING

Submitted by Adam Coates, Stanford University

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 preprocessing. 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 (e.g., Python) prototyping environments.

RESOURCES

- Scientists See Promise in Deep-Learning Programs. http://www.nytimes.com/2012/11/24/science/scientists-see-advances-in-deep-learning-a-part-of-artificial-intelligence.html. Accessed March 3, 2015.
- How Many Computers to Identify a Cat? 16,000. http://www.nytimes.com/2012/06/26/technology/in-a-big-network-of-computers-evidence-of-machine-learning.html. Accessed March 3, 2015.
- Now You Can Build Google's $1M Artificial Brain on the Cheap. http://www.wired.com/wiredenterprise/2013/06/andrew_ng/. Accessed March 3, 2015.
- Coates, A., Huval, B., Wang, T., Wu, D. J., Ng, A., Catanzaro, B. "Deep learning with COTS HPC systems." *Proceedings of the 30th International Conference on Machine Learning*, Atlanta, Georgia, USA, 2013. JMLR: W&CP Volume 28. http://www.cs.stanford.edu/~acoates/papers/CoatesHuvalWangWuNgCatanzaro_icml2013.pdf . Accessed March 3, 2015.
- Unsupervised Feature Learning and Deep Learning. http://ufldl.stanford.edu/wiki/index.php/Main_Page. Accessed March 3, 2015.
- Welcome to Deep Learning. http://deeplearning.net/. Accessed March 3, 2015.

2.6.2 USE CASE 27: ORGANIZING LARGE-SCALE, UNSTRUCTURED COLLECTIONS OF CONSUMER PHOTOS

Submitted by David Crandall, Indiana University

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 is a Hadoop cluster with 480 cores processing data of initial applications. Over 500 billion images are currently on Facebook, and over 5 billion are on Flickr, 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.

RESOURCES

- Discrete-continuous optimization for large-scale structure from motion. http://vision.soic.indiana.edu/disco. Accessed March 3, 2015.

2.6.3 USE CASE 28: TRUTHY—INFORMATION DIFFUSION RESEARCH FROM TWITTER DATA

Submitted by Filippo Menczer, Alessandro Flammini, and Emilio Ferrara, Indiana University

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, 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 packages for data analysis include Python/ SciPy/ NumPy/ MPI. Information diffusion, clustering, and dynamic network visualization capabilities already exist.

FUTURE

Truthy plans to expand, incorporating Google+ and Facebook, and so needs to move toward advanced distributed storage programs, such as Hadoop/Indexed HBase and Hadoop Distributed File System (HDFS). Redis should be used as an in-memory database to be a buffer for real-time analysis. Solutions will need to incorporate streaming clustering, anomaly detection, and online learning.

RESOURCES

- Truthy: Information diffusion research at Indiana University. http://truthy.indiana.edu/. Accessed March 3, 2015.
- Truthy: Information Diffusion in Online Social Networks. http://cnets.indiana.edu/groups/nan/truthy. Accessed March 3, 2015.
- Detecting Early Signature of Persuasion in Information Cascades (DESPIC). http://cnets.indiana.edu/groups/nan/despic. Accessed March 3, 2015.

2.6.4 USE CASE 29: CROWD SOURCING IN THE HUMANITIES AS SOURCE FOR BIG AND DYNAMIC DATA

Submitted by Sebastian Drude, Max-Planck-Institute for Psycholinguistics, Nijmegen, the Netherlands

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' audiovisual files; anonymization may be necessary but not always possible. Data management and curation are critical. With multimedia, the size could be hundreds of terabytes.

2.6.5 USE CASE 30: CINET—CYBERINFRASTRUCTURE FOR NETWORK (GRAPH) SCIENCE AND ANALYTICS

Submitted by Madhav Marathe and Keith Bisset, Virginia Tech

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; and
- 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.

RESOURCES

- Computational Network Sciences (CINET) GRANITE system. http://cinet.vbi.vt.edu/ . Accessed March 3, 2015.

2.6.6 USE CASE 31: NIST INFORMATION ACCESS DIVISION—ANALYTIC TECHNOLOGY PERFORMANCE MEASUREMENTS, EVALUATIONS, AND STANDARDS

Submitted by John Garofolo, NIST

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.

RESOURCES

- Information Access Division. http://www.nist.gov/itl/iad/. Accessed March 3, 2015.

2.7 THE ECOSYSTEM FOR RESEARCH

2.7.1 USE CASE 32: DATANET FEDERATION CONSORTIUM

Submitted by Reagan Moore, University of North Carolina at Chapel Hill

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. Figure 4 presents a diagram of the iRODS architecture.

FUTURE

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

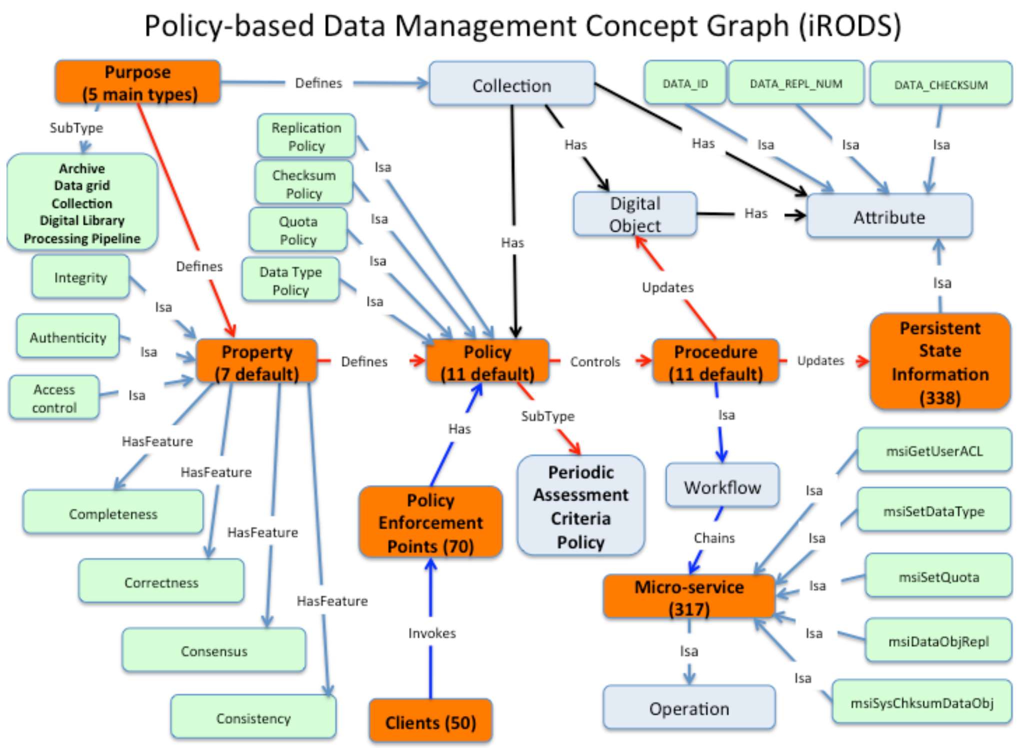

Figure 4: DFC—iRODS Architecture

RESOURCES

- DataNet Federation Consortium. http://renci.org/research/datanet-federation-consortium/. Accessed March 3, 2015.

2.7.2 USE CASE 33: THE DISCINNET PROCESS

Submitted by P. Journeau, Discinnet Labs

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.

RESOURCES

- DiscInNet: Interdisciplinary Networking. http://www.discinnet.org. Accessed March 3, 2015.

2.7.3 USE CASE 34: SEMANTIC GRAPH SEARCH ON SCIENTIFIC CHEMICAL AND TEXT-BASED DATA

Submitted by Talapady Bhat, NIST

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 [10], the Open Government movement [11], and the NIST Integrated Knowledge Editorial Net (NIKE) [12], a NIST-wide publication archive. These efforts are a component of the Research Data Alliance Metadata Standards Directory Working Group. [13]

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.

RESOURCES

- Facebook for molecules. http://www.eurekalert.org/pub_releases/2013-07/aiop-ffm071813.php. Accessed March 3, 2015.

- Chem-BLAST. http://xpdb.nist.gov/chemblast/pdb.pl. Accessed March 3, 2015.

2.7.4 Use Case 35: Light Source Beamlines

Submitted by Eli Dart, LBNL

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. For example, Octopus is used for tomographic reconstruction, and Avizo (http://vsg3d.com) and FIJI (a distribution of ImageJ) are used for visualization and 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 coming years, as will the need for a generalized infrastructure for analyzing GB per second of data from many beamline detectors at multiple facilities.

Resources

- Advanced Light Source. http://www-als.lbl.gov/. Accessed March 3, 2015.
- Advanced Photon Source. http://www.aps.anl.gov/. Accessed March 3, 2015.

2.8 Astronomy and Physics

2.8.1 Use Case 36: Catalina Real-Time Transient Survey: A Digital, Panoramic, Synoptic Sky Survey

Submitted by S. G. Djorgovski, Caltech

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 preprocessed 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. Figure 5 illustrates the schematic architecture for a cyber infrastructure for time domain astronomy.

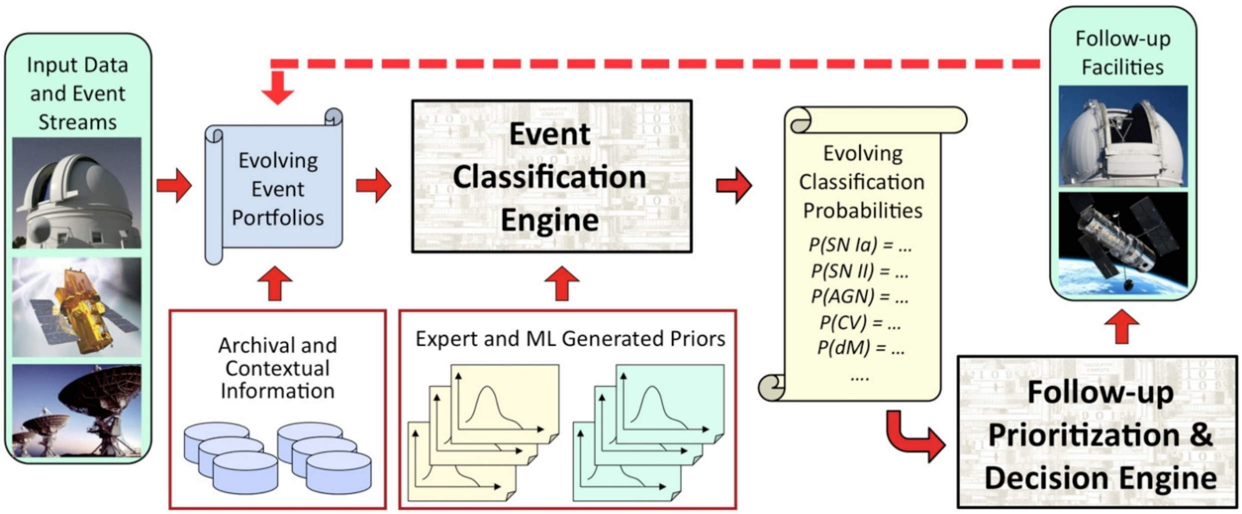

Figure 5: Catalina CRTS: A Digital, Panoramic, Synoptic Sky Survey

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.

RESOURCES

- Flashes in a Star Stream: Automated Classification of Astronomical Transient Events. http://arxiv.org/abs/1209.1681. Accessed March 3, 2015.

2.8.2 USE CASE 37: DOE EXTREME DATA FROM COSMOLOGICAL SKY SURVEY AND SIMULATIONS

Submitted by Salman Habib, Argonne National Laboratory; Andrew Connolly, University of Washington

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

RESOURCES

- The New Sky. http://www.lsst.org/lsst/. Accessed March 3, 2015.
- National Energy Research Scientific Computing Center. http://www.nersc.gov/. Accessed March 3, 2015.
- Basic Research: Non-Accelerator Physics. http://science.energy.gov/hep/research/basic-research/non-accelerator-physics/. Accessed March 3, 2015.
- Present and Future Computing Requirements for Computational Cosmology. http://www.nersc.gov/assets/Uploads/HabibcosmosimV2.pdf. Accessed March 3, 2015.

2.8.3 USE CASE 38: LARGE SURVEY DATA FOR COSMOLOGY

Submitted by Peter Nugent, LBNL

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.’ Here, galaxies 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 are Linux cluster, Oracle RDBMS server, Postgres PSQL, large memory machines, standard Linux interactive hosts, and the General Parallel File System (GPFS). HPC resources are needed for simulations. Software needs include standard astrophysics reduction software as well as Perl/Python wrapper scripts and Linux Cluster scheduling.

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 catalog data and a correspondingly large (or larger) amount of simulation data. In total, over 20 TB of data will be generated per night.

Resources

- Dark Energy Spectroscopic Instrument (DESI). http://desi.lbl.gov. Accessed March 3, 2015.
- Why is the universe speeding up? http://www.darkenergysurvey.org. Accessed March 3, 2015.

2.8.4 Use Case 39: Particle Physics—Analysis of Large Hadron Collider Data: Discovery of Higgs Particle

Submitted by Michael Ernst, Brookhaven National Laboratory (BNL); Lothar Bauerdick, Fermi National Accelerator Laboratory (FNAL); Geoffrey Fox, Indiana University; Eli Dart, LBNL

Application

Analysis is conducted on collisions at the European Organization for Nuclear Research (CERN) Large Hadron Collider (LHC) accelerator (Figure 6) and Monte Carlo producing events describing particle-apparatus interaction.

Figure 6: Particle Physics: Analysis of LHC Data: Discovery of Higgs Particle—CERN LHC Location

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. Figure 7 shows grid analysis with 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.

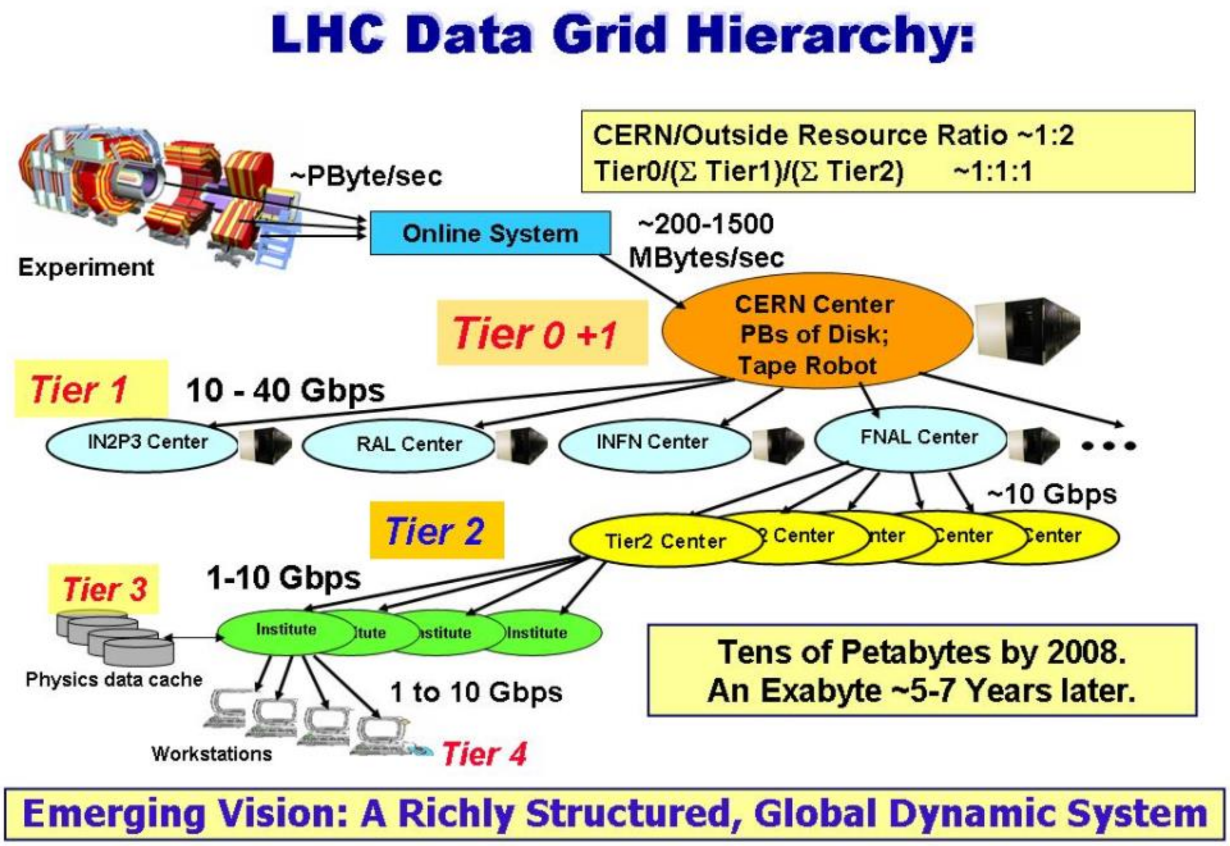

Figure 7: Particle Physics: Analysis of LHC Data: Discovery of Higgs Particle—The Multi-tier LHC Computing Infrastructure

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.

RESOURCES

- Where does all the data come from? http://grids.ucs.indiana.edu/ptliupages/publications/Where%20does%20all%20the%20data%20come%20from%20v7.pdf. Accessed March 3, 2015.
- Enabling high throughput in widely distributed data management and analysis systems: Lessons from the LHC. http://www.es.net/assets/pubs_presos/High-throughput-lessons-from-the-LHC-experience.Johnston.TNC2013.pdf. Accessed March 3, 2015.

2.8.5 USE CASE 40: BELLE II HIGH ENERGY PHYSICS EXPERIMENT

Submitted by David Asner and Malachi Schram, Pacific Northwest National Laboratory (PNNL)

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 will start 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. Open Science Grid, Geant4, DIRAC, FTS, and Belle II framework software will be needed.

RESOURCES

- Belle II Collaboration. http://belle2.kek.jp. Accessed March 3, 2015.

2.9 EARTH, ENVIRONMENTAL, AND POLAR SCIENCE

2.9.1 USE CASE 41: EUROPEAN INCOHERENT SCATTER SCIENTIFIC ASSOCIATION 3D INCOHERENT SCATTER RADAR SYSTEM

Submitted by Yin Chen, Cardiff University; Ingemar Häggström, Ingrid Mann, and Craig Heinselman, European Incoherent Scatter Scientific Association (EISCAT)

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 kilometers 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. See Figure 8.

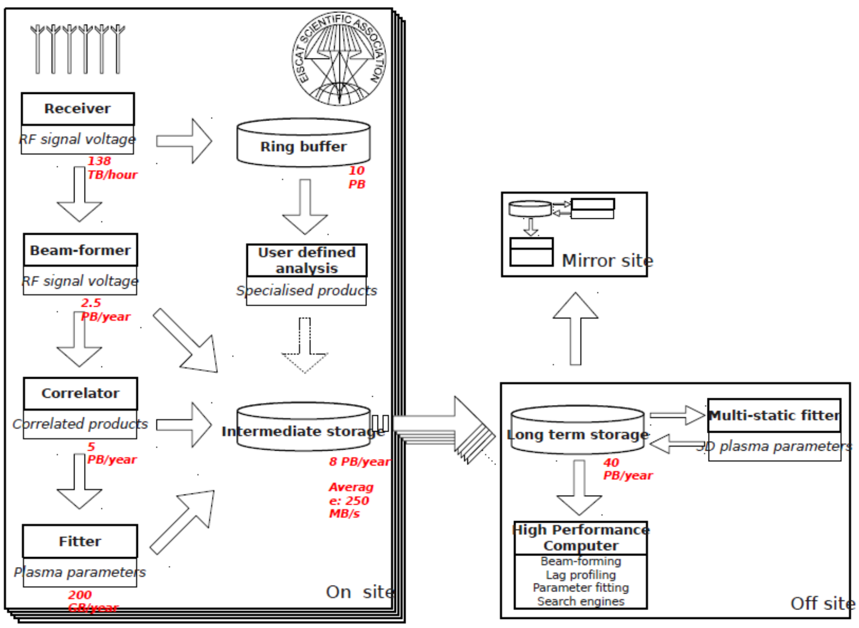

Figure 8: EISCAT 3D Incoherent Scatter Radar System – System Architecture

RESOURCES

- EISCAT 3D. https://www.eiscat3d.se/. Accessed March 3, 2015.

2.9.2 USE CASE 42: COMMON OPERATIONS OF ENVIRONMENTAL RESEARCH INFRASTRUCTURE

Submitted by Yin Chen, Cardiff University

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 (Figure 9). 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.

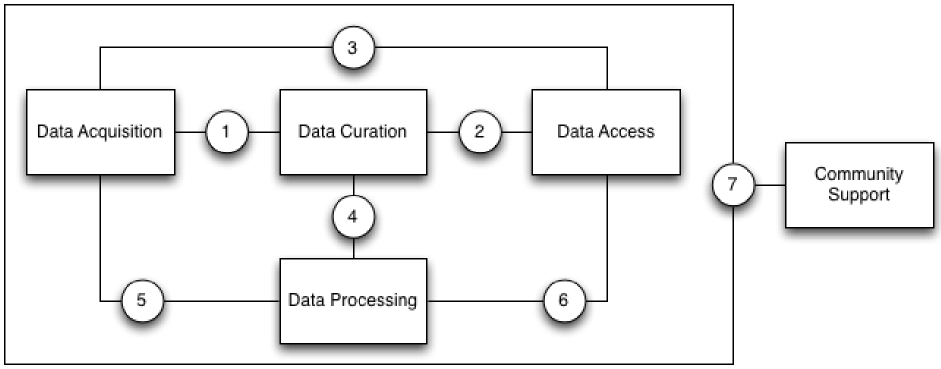

Figure 9: ENVRI Common Architecture

Figures 10(a) through 10(e) illustrate how well the five subsystems map to the architectures of the ESFRI environmental research infrastructures.

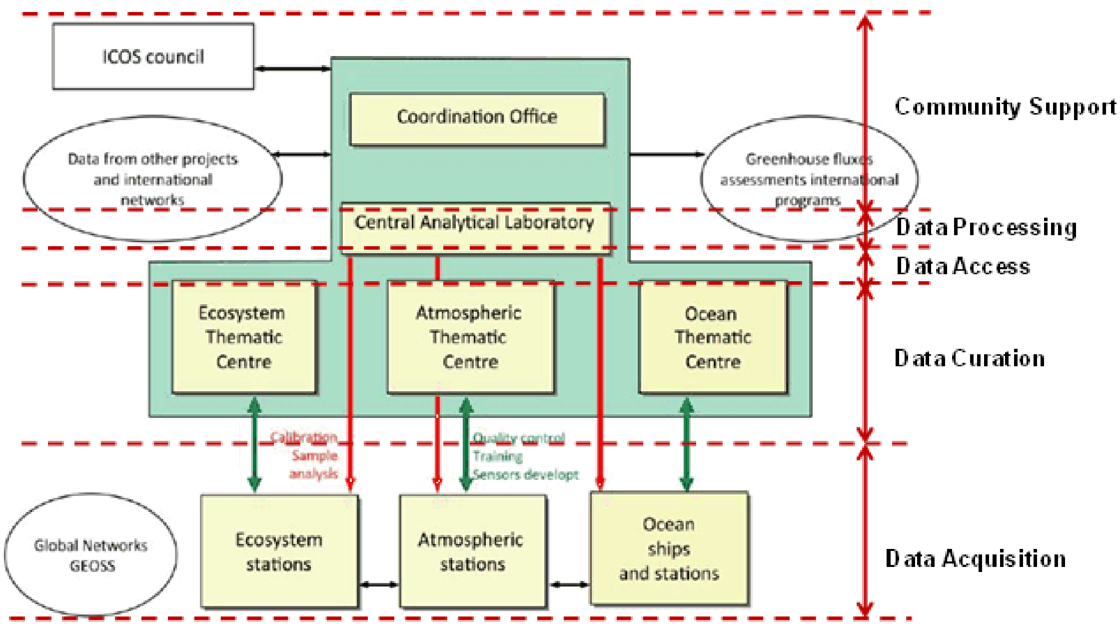

Figure 10(a): ICOS Architecture

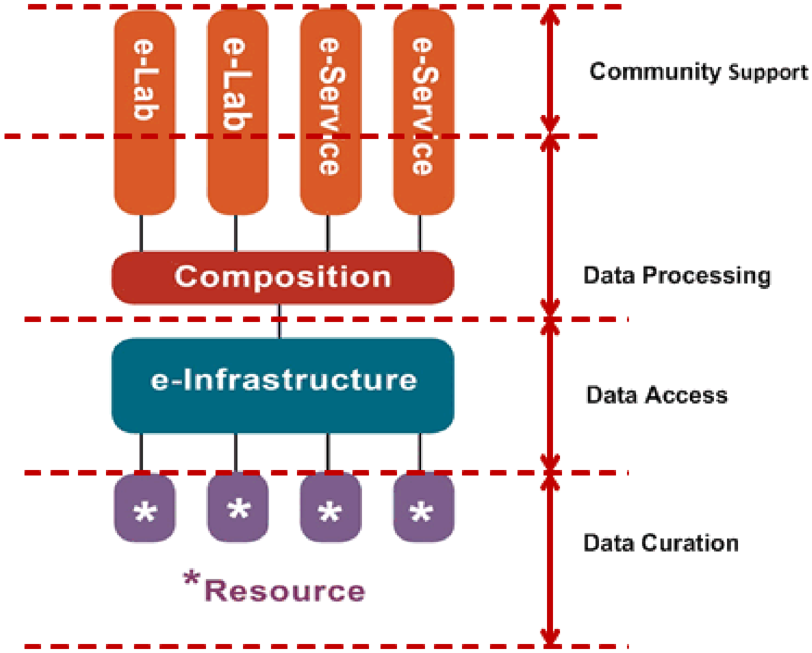

Figure 10(b): LifeWatch Architecture

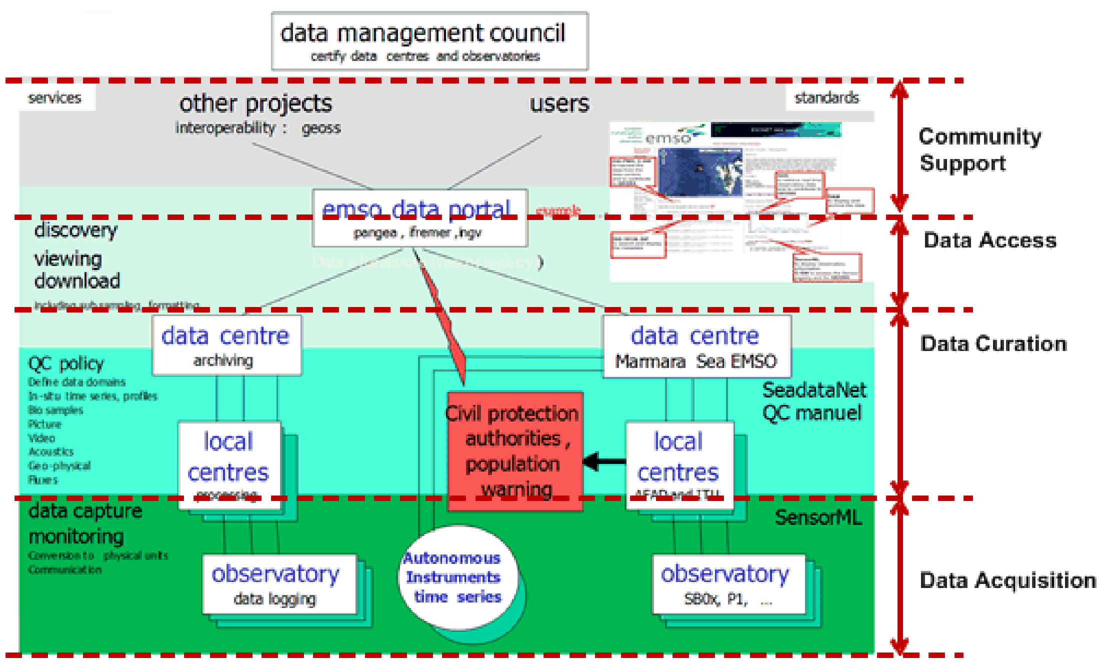

Figure 10(c): EMSO Architecture

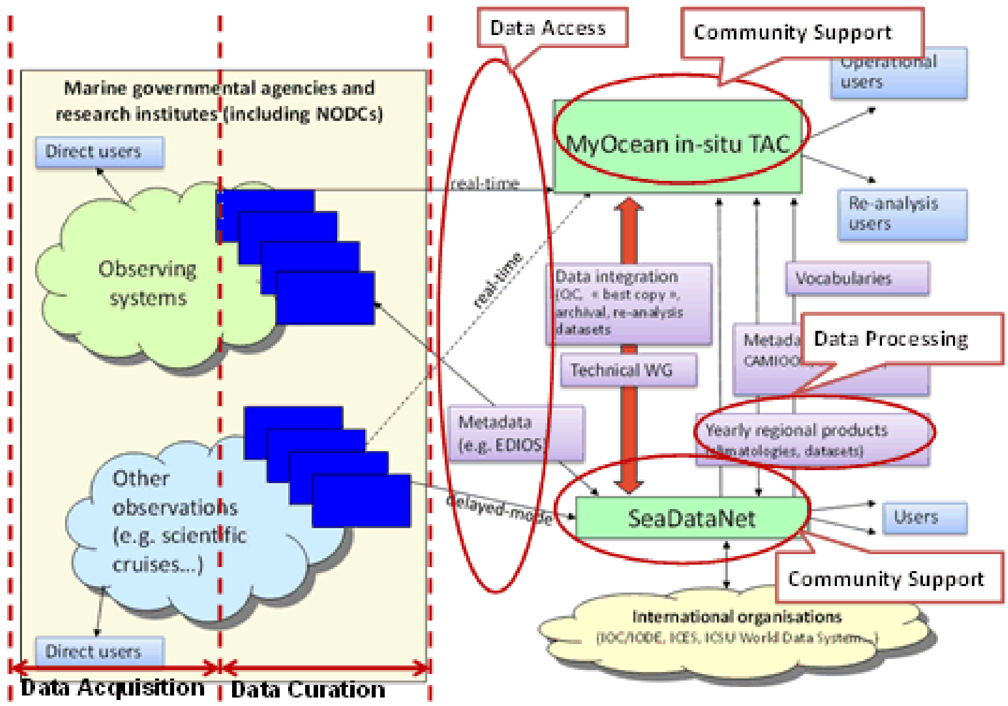

Figure 10(d): EURO-Argo Architecture

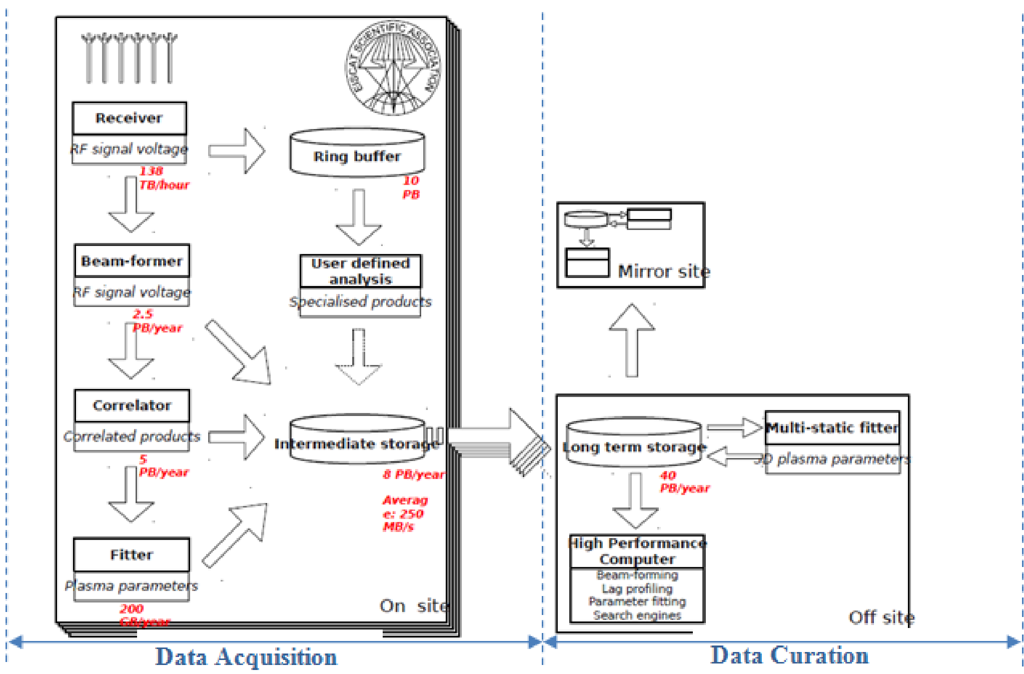

Figure 10(e): EISCAT 3D Architecture

RESOURCES

- Analysis of Common Requirements for Environmental Science Research Infrastructures. http://pos.sissa.it/archive/conferences/179/032/ISGC%202013_032.pdf. Accessed March 3, 2015.
- Euro-Argo RI. http://www.euro-argo.eu/. Accessed March 3, 2015.
- EISCAT 3D. https://www.eiscat3d.se/. Accessed March 3, 2015.

- LifeWatch. http://www.lifewatch.com/. Accessed March 3, 2015.
- European Multidisciplinary Seafloor & Water Column Observatory (EMSO). http://www.emso-eu.org/. Accessed March 3, 2015.

2.9.3 USE CASE 43: RADAR DATA ANALYSIS FOR THE CENTER FOR REMOTE SENSING OF ICE SHEETS

Submitted by Geoffrey Fox, Indiana University

APPLICATION

As illustrated in Figure 11, 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.

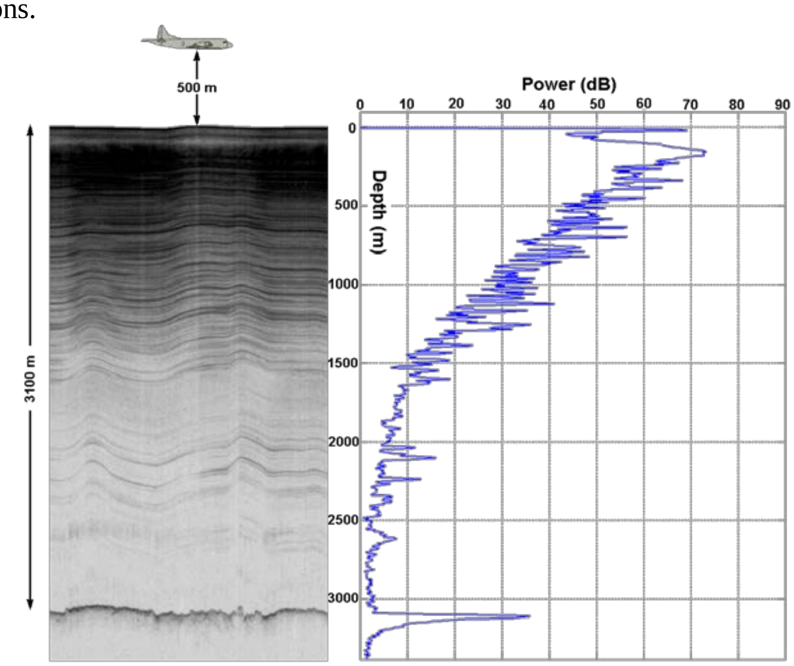

Figure 11: Typical CReSIS Radar Data After Analysis

Resulting data feed into the Intergovernmental Panel on Climate Change (IPCC). The radar systems are typically flown in by aircraft in multiple paths, as illustrated by Figure 12.

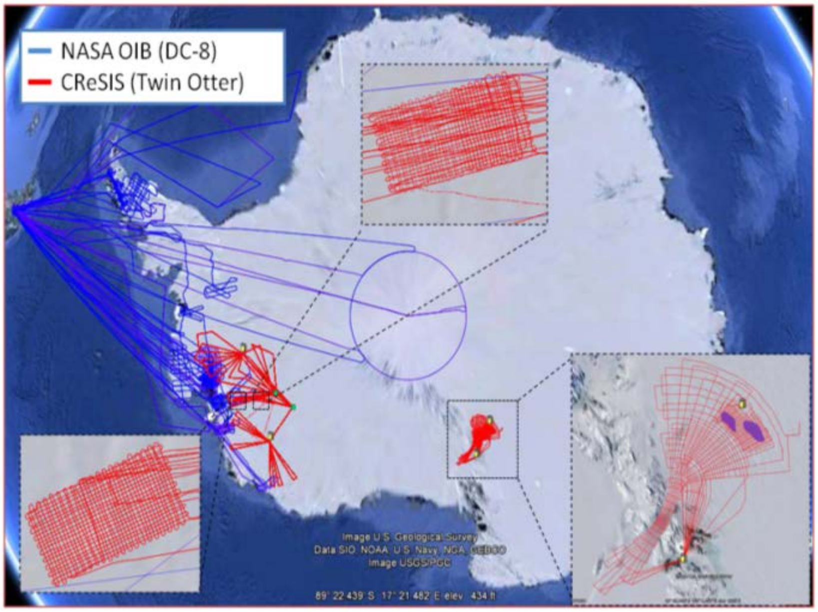

Figure 12: Radar Data Analysis for CReSIS Remote Sensing of Ice Sheets– Typical Flight Paths of Data Gathering in Survey Region

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. Figure 13 illustrates image features (i.e., layers) found using image understanding tools with some human oversight. Figure 13 is a typical echogram with detected boundaries. The upper (green) boundary is between air and ice layers, while the lower (red) boundary is 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 TB to 100 TB of data.

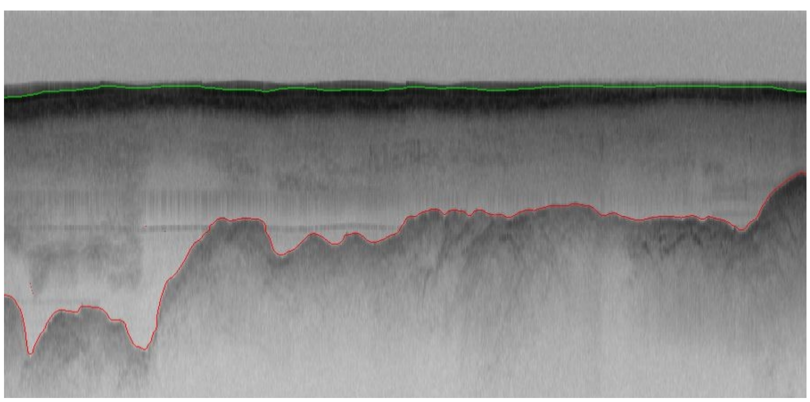

Figure 13: Typical echogram with detected boundaries

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.

RESOURCES

- CReSIS. https://www.cresis.ku.edu. Accessed March 3, 2015.
- Polar Grid Multimedia Gallery, Indiana University. http://polargrid.org/gallery.html . Accessed March 3, 2015.

2.9.4 USE CASE 44: UNMANNED AIR VEHICLE SYNTHETIC APERTURE RADAR (UAVSAR) DATA PROCESSING, DATA PRODUCT DELIVERY, AND DATA SERVICES

Submitted by Andrea Donnellan and Jay Parker, National Aeronautics and Space Administration (NASA) Jet Propulsion Laboratory

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 shown in Figure 14, as well as disaster management. Figure 14 shows the combined unwrapped coseismic interferograms for flight lines 26501, 26505, and 08508 for the October 2009 to April 2010 time period. End points where slip can be seen on the Imperial, Superstition Hills, and Elmore Ranch faults are noted. GPS stations are marked by dots and are labeled. This use case supports the storage, image processing application, and visualization of geo-located data with angular specification.

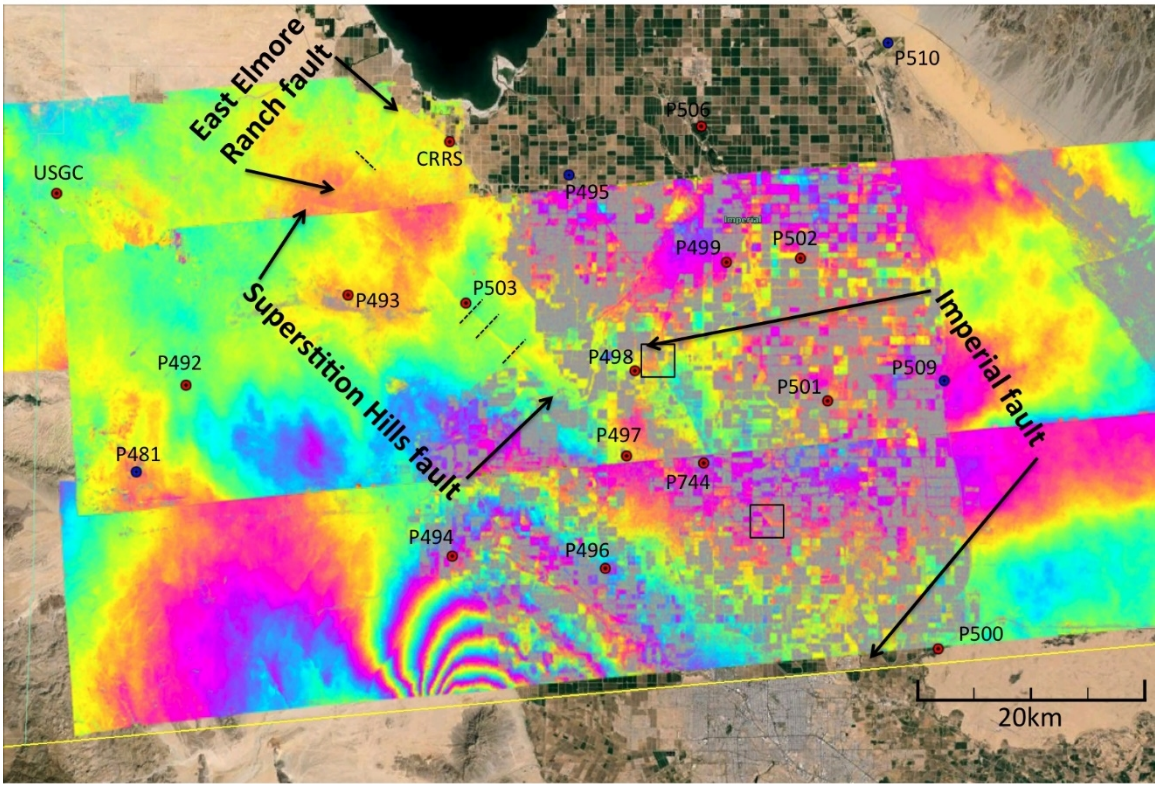

Figure 14: Combined Unwrapped Coseismic Interferograms

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.

RESOURCES

- Uninhabited Aerial Vehicle Synthetic Aperture Radar. http://uavsar.jpl.nasa.gov/. Accessed March 3, 2015.
- Alaska Satellite Facility. http://www.asf.alaska.edu/program/sdc. Accessed March 3, 2015.
- QuakeSim: Understanding Earthquake Processes. http://quakesim.org. Accessed March 3, 2015.

2.9.5 USE CASE 45: NASA LANGLEY RESEARCH CENTER/ GODDARD SPACE FLIGHT CENTER IRODS FEDERATION TEST BED

Submitted by Brandi Quam, NASA Langley Research Center

APPLICATION

NASA Center for Climate Simulation and NASA Atmospheric Science Data Center have complementary datasets, 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.

2.9.6 USE CASE 46: MERRA ANALYTIC SERVICES (MERRA/AS)

Submitted by John L. Schnase and Daniel Q. Duffy, NASA Goddard Space Flight Center

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. Figure 15 shows a typical MERRA/AS output.

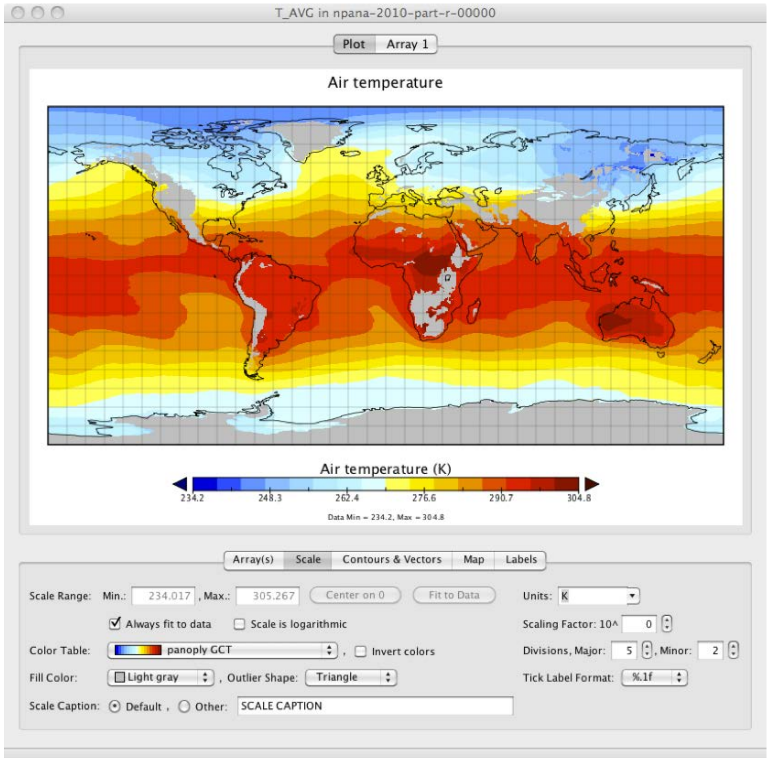

Figure 15: Typical MERRA/AS Output

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.

2.9.7 USE CASE 47: ATMOSPHERIC TURBULENCE – EVENT DISCOVERY AND PREDICTIVE ANALYTICS

Submitted by Michael Seablom, NASA headquarters

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 dataset 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. Figure 16 shows a typical turbulent wave image.

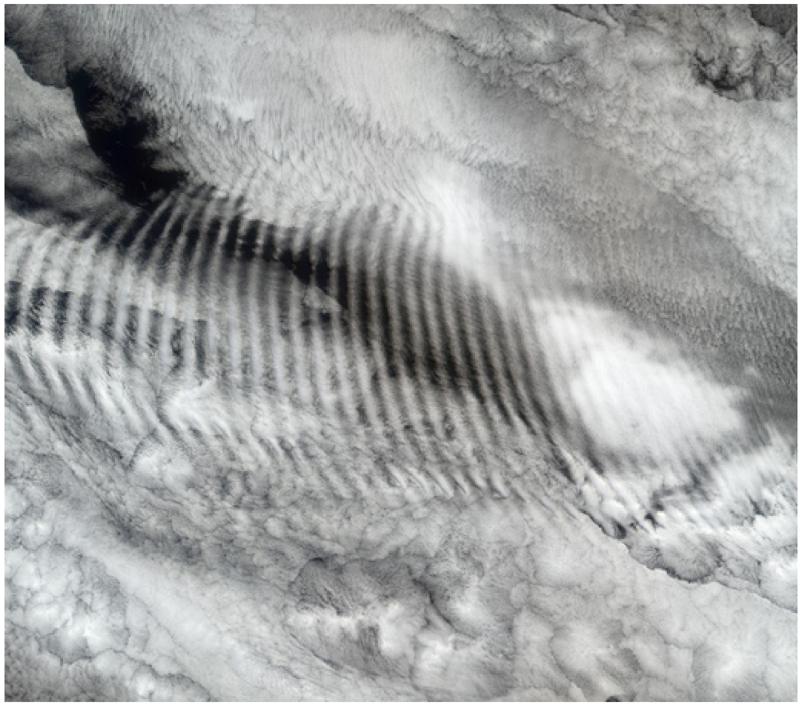

Figure 16: Typical NASA Image of Turbulent Waves

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.

RESOURCES
- El Niño Teleconnections. http://oceanworld.tamu.edu/resources/oceanography-book/teleconnections.htm. Accessed March 3, 2015.
- Meet The Scientists Mining Big Data To Predict The Weather. http://www.forbes.com/sites/toddwoody/2012/03/21/meet-the-scientists-mining-big-data-to-predict-the-weather/. Accessed March 3, 2015.

2.9.8 USE CASE 48: CLIMATE STUDIES USING THE COMMUNITY EARTH SYSTEM MODEL AT THE U.S. DEPARTMENT OF ENERGY (DOE) NERSC CENTER

Submitted by Warren Washington, National Center for Atmospheric Research

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.

RESOURCES

- Earth System Grid (ESG) Gateway at the National Center for Atmospheric Research. http://www.earthsystemgrid.org. Accessed March 3, 2015.
- Welcome to PCMDI! http://www-pcmdi.llnl.gov/. Accessed March 3, 2015.
- National Energy Research Scientific Computing Center. http://www.nersc.gov/. Accessed March 3, 2015.
- Research: Climate and Environmental Sciences Division (CESD). http://science.energy.gov/ber/research/cesd/. Accessed March 3, 2015.
- Computational & Information Systems Lab (CISL). http://www2.cisl.ucar.edu/. Accessed March 3, 2015.

2.9.9 USE CASE 49: DOE BIOLOGICAL AND ENVIRONMENTAL RESEARCH (BER) SUBSURFACE BIOGEOCHEMISTRY SCIENTIFIC FOCUS AREA

Submitted by Deb Agarwal, LBNL

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 developing 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.

2.9.10 USE CASE 50: DOE BER AMERIFLUX AND FLUXNET NETWORKS

Submitted by Deb Agarwal, LBNL

APPLICATION

AmeriFlux and Flux Tower Network (FLUXNET) are U.S. and world collections, respectively, of sensors that observe trace gas fluxes (e.g., CO_2, 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 includes EddyPro, custom analysis software, R, Python, neural networks, and MATLAB. 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.

RESOURCES

- AmeriFlux. http://Ameriflux.lbl.gov. Accessed March 3, 2015.
- Welcome to the Fluxdata.org web site. http://www.fluxdata.org. Accessed March 3, 2015.

2.9.11 USE CASE 2-1: NASA EARTH OBSERVING SYSTEM DATA AND INFORMATION SYSTEM (EOSDIS)

Submitted by Christopher Lynnes

APPLICATION

The Earth Observing System Data and Information System (EOSDIS) is the main system maintained by NASA for the archive and dissemination of Earth Observation data. The system comprises 12 discipline-oriented data systems spread across the United States. This network is linked together using interoperability frameworks such as the Common Metadata Repository, a file-level database that supports one-stop searching across EOSDIS. The data consist of satellite, aircraft, field campaign, and in situ data over a variety of disciplines related to Earth science, covering the Atmosphere, Hydrosphere, Cryosphere, Lithosphere, Biosphere, and Anthroposphere. Data are distributed to a diverse community ranging from Earth science researchers to applications to citizen science and educational users.

EOSDIS faces major challenges in both Volume and Variety. As of early 2017, the cumulative archive data volume is over 20 Petabytes. Higher-resolution space-borne instruments are expected to increase that volume by two orders of magnitude (~200 PB) over the next 7 years. More importantly, the data distribution to users is equally high. In a given year, EOSDIS distributes a volume that is comparable to the overall cumulative archive volume.

Detailed topics include the following:

- Data Archiving: storing NASA's Earth Observation data;
- Data Distribution: disseminating data to end users in Research, Applications (e.g., water resource management) and Education;
- Data Discovery: search and access to Earth Observation data;
- Data Visualization: static browse images and dynamically constructed visualizations;

- Data Customization: subsetting, reformatting, regridding, mosaicking, and quality screening on behalf of end users;
- Data Processing: routine production of standard scientific datasets, converting raw data to geophysical variables; and
- Data Analytics: end-user analysis of large datasets, such as time-averaged maps and area-averaged time series.

CURRENT APPROACH

Standard data processing converts raw data to geophysical parameters. Though much of this is heritage custom Fortran or C code running, current prototypes are using cloud computing to scale up to rapid reprocessing campaigns.

EOSDIS support of end-user analysis currently uses high-performance software, such as the netCDF Command Operators. However, current prototypes are using cloud computing and data-parallel algorithms (e.g., Spark) to achieve an order of magnitude speed-up.

FUTURE

EOSDIS is beginning to migrate data archiving to the cloud to enable end users to bring algorithms to the data. We also expect to reorganize certain high-value datasets into forms that lend themselves to cloud data-parallel computing. Prototypes are under way to prove out storage schemes that are optimized for cloud analytics, such as space-time tiles stored in cloud databases and cloud file systems.

RESOURCES

- Global Web-Enabled Landsat Data, Geospatial Sciences Center of Excellence (GSCE), South Dakota State University: http://globalmonitoring.sdstate.edu/projects/weldglobal/gweld.html
- Global Web-Enabled Landsat Data, U.S. Geological Survey: http://globalweld.cr.usgs.gov/
- NASA Earth Exchange (NEX): https://nex.nasa.gov
- NASA High-End Computing Capability: http://www.nas.nasa.gov/hecc/resources/pleiades.html
- NASA Earth Data, Global Imagery Browse Services (GIBS): https://earthdata.nasa.gov/about/science-system-description/eosdis-components/global-imagery-browse-services-gibs
- NASA Earthdata, Worldview: https://worldview.earthdata.nasa.gov/

2.9.12 USE CASE 2-2: WEB-ENABLED LANDSAT DATA (WELD) PROCESSING

Submitted by Andrew Michaelis

APPLICATION

The use case shown in Figure17 is specific to the part of the project where data is available on the HPC platform and processed through the science workflow. It is a 32-stage processing pipeline of images from the Landsat 4, 5, and 7 satellites that includes two separate science products (Top-of-the-Atmosphere [TOA] reflectances and surface reflectances) as well as QA and visualization components which forms a dataset of science products of use to the land surface science community that is made freely available by NASA.

CURRENT APPROACH

This uses the High Performance Computing (HPC) system Pleiades at NASA Ames Research Center with storage in NASA Earth Exchange (NEX) NFS storage system for read-only data storage (2.5PB), Lustre for read-write access during processing (1PB), tape for near-line storage (50PB). The networking is InfiniBand partial hypercube internal interconnect within the HPC system; 1G to 10G connection to external data providers. The software is the NEX science platform for data management, workflow processing, provenance capture; the WELD science processing algorithms from South Dakota State University for visualization and time-series; the Global Imagery Browse Service (GIBS) data visualization platform; and the USGS data distribution platform. This is a custom-built application and libraries built on top of open-source libraries.

FUTURE

Processing will be improved with newer and updated algorithms. This process may also be applied to future datasets and processing systems (Landsat 8 and Sentinel-2 satellites, for example).

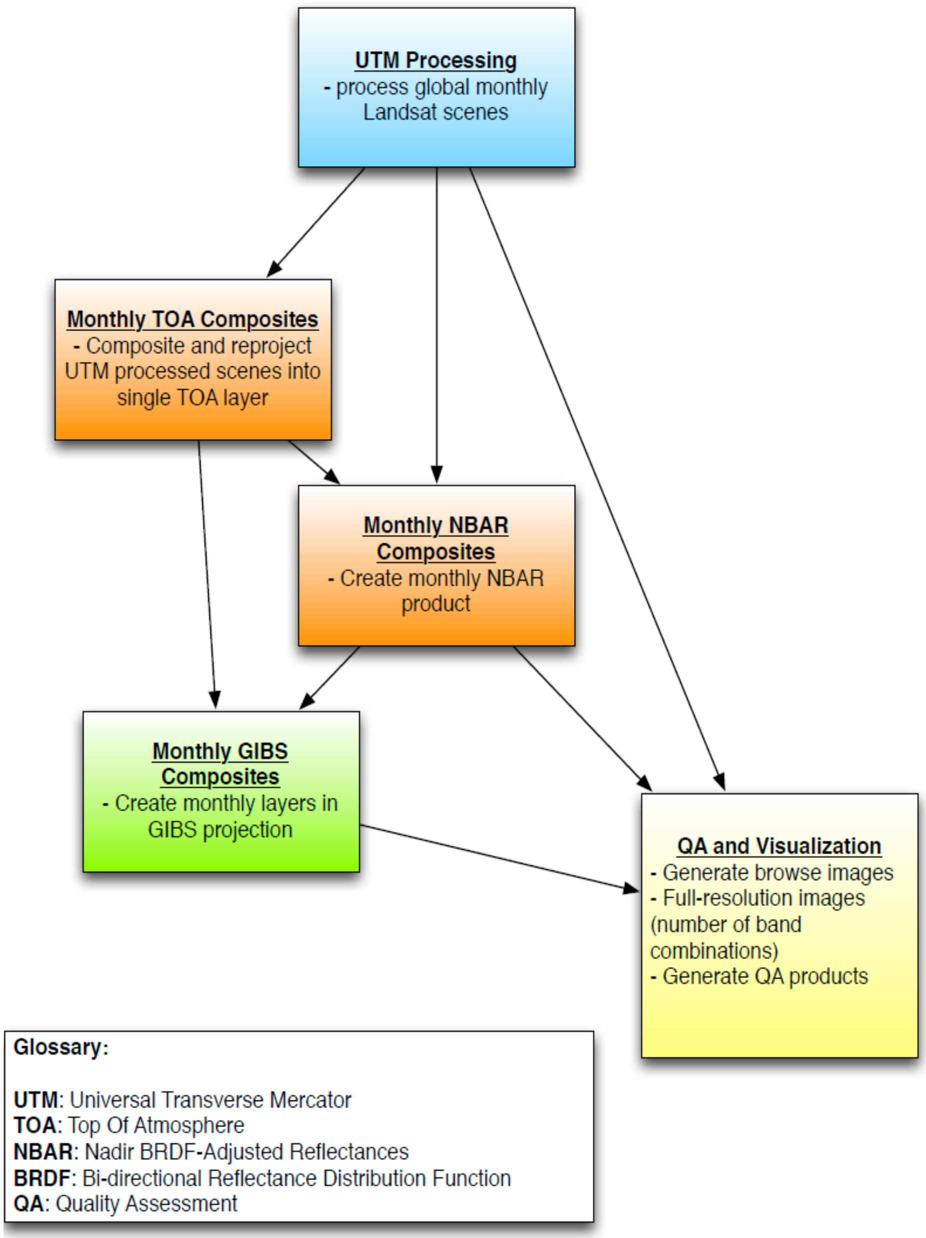

Figure 17: NASA NEX WELD/GIBS Processing Workflow

RESOURCES

- NASA, Earthdata: https://earthdata.nasa.gov/

2.10 ENERGY

2.10.1 USE CASE 51: CONSUMPTION FORECASTING IN SMART GRIDS

Submitted by Yogesh Simmhan, University of Southern California

APPLICATION

Smart meters support prediction of energy consumption for customers, transformers, substations and the electrical grid service area. Advanced meters provide measurements every 15 minutes 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. The process uses R/Matlab, Weka, and Hadoop software. There are significant privacy issues requiring anonymization by aggregation. Real-time and historic data are combined with machine learning to predict consumption.

FUTURE

Advanced grid technologies will have wide-spread deployment. Smart grids will have new analytics integrating diverse data and supporting curtailment requests. New technologies will support mobile applications for client interactions.

RESOURCES

- USC Smart Grid. http://smartgrid.usc.edu. Accessed March 3, 2015.
- Smart Grid. http://ganges.usc.edu/wiki/Smart_Grid. Accessed March 3, 2015.
- Smart Grid L.A. https://www.ladwp.com/ladwp/faces/ladwp/aboutus/a-power/a-p-smartgridla. Accessed March 3, 2015.
- Cloud-Based Software Platform for Big Data Analytics in Smart Grids. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6475927. Accessed March 3, 2015.

3 USE CASE REQUIREMENTS

Requirements are the challenges limiting further use of Big Data. After collection, processing, and review of the use cases, requirements within seven characteristic categories were extracted from the individual use cases. These use case specific requirements were then aggregated to produce high-level, general requirements, within the seven characteristic categories, that are vendor-neutral and technology-agnostic. Neither the use case nor the requirements lists are exhaustive.

3.1 USE CASE SPECIFIC REQUIREMENTS

Each use case was evaluated for requirements within the following seven categories. These categories were derived from Subgroup discussions and motivated by components of the evolving reference architecture at the time. The process involved several Subgroup members extracting requirements and iterating back their suggestions for modifying the categories.

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

Some use cases contained requirements in all seven categories while others included only requirements for a few categories. The complete list of specific requirements extracted from the use cases is presented in Appendix D. Section 2.1 of the *NIST Big Data Interoperability Framework: Volume 6 Reference Architecture* maps these seven categories to terms used in the reference architecture. The categories map in a one-to-one fashion but have slightly different terminology as the use case requirements analysis was performed before the reference architecture was finalized.

3.2 GENERAL REQUIREMENTS

Aggregation of the use case-specific requirements allowed formation of more generalized requirements under the seven categories. These generalized requirements are listed below by category.

DATA SOURCE REQUIREMENTS (DSR)

- DSR-1: Needs to support reliable real-time, asynchronous, streaming, and batch processing to collect data from centralized, distributed, and cloud data sources, sensors, or instruments.
- DSR-2: Needs to support slow, bursty, and high-throughput data transmission between data sources and computing clusters.
- DSR-3: 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 PROVIDER REQUIREMENTS (TPR)

- TPR-1: Needs to support diversified compute-intensive, statistical and graph analytic processing, and machine learning techniques.
- TPR-2: Needs to support batch and real-time analytic processing.
- TPR-3: Needs to support processing large diversified data content and modeling.
- TPR-4: Needs to support processing data in motion (streaming, fetching new content, tracking, etc.).

CAPABILITY PROVIDER REQUIREMENTS (CPR)

- CPR-1: Needs to support legacy and advanced software packages (software).
- CPR-2: Needs to support legacy and advanced computing platforms (platform).
- CPR-3: Needs to support legacy and advanced distributed computing clusters, co-processors, input output (I/O) processing (infrastructure).
- CPR-4: Needs to support elastic data transmission (networking).
- CPR-5: Needs to support legacy, large, and advanced distributed data storage (storage).
- CPR-6: Needs to support legacy and advanced executable programming: applications, tools, utilities, and libraries (software).

DATA CONSUMER REQUIREMENTS (DCR)

- DCR-1: Needs to support fast searches from processed data with high relevancy, accuracy, and recall.
- DCR-2: Needs to support diversified output file formats for visualization, rendering, and reporting.
- DCR-3: Needs to support visual layout for results presentation.
- DCR-4: Needs to support rich user interface for access using browser, visualization tools.
- DCR-5: Needs to support high-resolution, multidimension layer of data visualization.
- DCR-6: Needs to support streaming results to clients.

SECURITY AND PRIVACY REQUIREMENTS (SPR)

- SPR-1: Needs to protect and preserve security and privacy of sensitive data.
- SPR-2: Needs to support sandbox, access control, and multilevel, policy-driven authentication on protected data.

LIFE CYCLE MANAGEMENT REQUIREMENTS (LMR)

- LMR-1: Needs to support data quality curation including preprocessing, data clustering, classification, reduction, and format transformation.
- LMR-2: Needs to support dynamic updates on data, user profiles, and links.
- LMR-3: Needs to support data life cycle and long-term preservation policy, including data provenance.
- LMR-4: Needs to support data validation.
- LMR-5: Needs to support human annotation for data validation.
- LMR-6: Needs to support prevention of data loss or corruption.
- LMR-7: Needs to support multisite archives.
- LMR-8: Needs to support persistent identifier and data traceability.
- `LMR-9: Needs to support standardizing, aggregating, and normalizing data from disparate sources.

OTHER REQUIREMENTS (OR)

- OR-1: Needs to support rich user interface from mobile platforms to access processed results.
- OR-2: Needs to support performance monitoring on analytic processing from mobile platforms.
- OR-3: Needs to support rich visual content search and rendering from mobile platforms.
- OR-4: Needs to support mobile device data acquisition.
- OR-5: Needs to support security across mobile devices.

4 ADDITIONAL USE CASE CONTRIBUTIONS

During the development of version 2 of the NBDIF, the Use Cases and Requirements Subgroup and the Security and Privacy Subgroup identified the need for additional use cases to strengthen the future work of the NBD-PWG. These two subgroups collaboratively created the Use Case Template 2 with the aim of collecting specific and standardized information for each use case. In addition to questions from the original use case template, the Use Case Template 2 contains questions that will provide a comprehensive view of security, privacy, and other topics for each use case.

The NBD-PWG invites the public to submit new use cases through the Use Case Template 2. To submit a use case, please fill out the PDF form (https://bigdatawg.nist.gov/_uploadfiles/M0621_v2_7345181325.pdf) and email it to Wo Chang (wchang@nist.gov). Use cases will be accepted until the end of Phase 3 work and will be evaluated as they are submitted.

Appendix A: Use Case Study Source Materials

Appendix A contains one blank use case template and the original completed use cases. The Use Case Studies Template 1 included in this Appendix is no longer being used to collect use case information. To submit a new use case, refer to Appendix E for the current Use Case Template 2.

These use cases were the source material for the use case summaries presented in Section 2 and the use case requirements presented in Section 3 of this document. The completed use cases have not been edited and contain the original text as submitted by the author(s). The use cases are as follows:

NBD-PWG USE CASE STUDIES TEMPLATE 1

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 Requirements		
Highlight issues for generalizing this use case (e.g. for ref. architecture)		
More Information (URLs)		
Note: <additional comments>		

Notes: No proprietary or confidential information should be included.
ADD picture of operation or data architecture of application below table.

Comments on fields

The following descriptions of fields in the template are provided to help with the understanding of both document intention and meaning of the 26 fields and also to indicate ways that they can be improved.

- **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:** Describes the players and their roles in the use case.
- **Goals:** Objectives of the use case.
- **Use Case Description:** Brief description of the use case.
- **Current Solutions:** Describes current approach to processing Big Data at the hardware and software infrastructure 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.
- **Big Data Characteristics:** Describes the properties of the (raw) data including the four major 'V's' of Big Data described in *NIST Big Data Interoperability Framework: Volume 1, Big Data Definition* of this report series.
    - **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.
    - **Volume:** The characteristic of data at rest that is most associated with Big Data. The size of data varied drastically between use cases from terabytes to petabytes for science research (100 petabytes was the largest science use case for LHC data analysis), or up to exabytes in a commercial use case.
    - **Velocity:** Refers to the rate of flow at which the data is created, stored, analyzed, and visualized. For example, big velocity means that a large quantity of data is being processed in a short amount of time.
    - **Variety:** Refers to data from multiple repositories, domains, or types.
    - **Variability:** Refers to changes in rate and nature of data gathered by use case.
- **Big Data Science:** Describes the high-level aspects of the data analysis process
    - **Veracity:** Refers to the completeness and accuracy of the data with respect to semantic content. *NIST Big Data Interoperability Framework: Volume 1, Big Data Definition* discusses veracity in more detail.
    - **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 Quality:** This refers to syntactical quality of data. In retrospect, this template field could have been included in the Veracity field.
    - **Data Types:** Refers to the style of data such as structured, unstructured, images (e.g., pixels), text (e.g., characters), gene sequences, and numerical.
    - **Data Analytics:** Defined in *NIST Big Data Interoperability Framework: Volume 1, Big Data Definition* as "the synthesis of knowledge from information". In the context of these use cases, 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.

- **Big Data Specific Challenges (Gaps):** Allows for explanation of special difficulties for processing Big Data in the use case and gaps where new approaches/technologies are used.
- **Big Data Specific Challenges in Mobility:** Refers to issues in accessing or generating Big Data from Smart Phones and tablets.
- **Security and Privacy Requirements:** Allows for explanation of security and privacy issues or needs related to this use case.
- **Highlight issues for generalizing this use case:** Allows for documentation of issues that could be common across multiple use-cases and could lead to reference architecture constraints.
- **More Information (URLs):** Resources that provide more information on the use case.
- **Note: <additional comments>:** Includes pictures of use-case in action but was not otherwise used.

SUBMITTED USE CASE STUDIES

Government Operation> 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 Terabytes.
	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)**	Cannot tolerate data loss.
	Visualization	TBD
	Data Quality	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 Requirements	Title 13 data.	
Highlight issues for generalizing this use case (e.g. for ref. architecture)		
More Information (URLs)		

Government Operation> 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 datasets, 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.

Government Operation> Use Case 2: NARA Accession, Search, Retrieve, Preservation

Use Case Title	National Archives and Records Administration Accession NARA Accession, Search, Retrieve, Preservation	
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues)**	Search results should have high relevancy and high recall. Categorization of records should be highly accurate.
	Visualization	TBD
	Data Quality	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 preprocessing 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 Requirements	Need to be sensitive to data access restrictions.	
Highlight issues for generalizing this use case (e.g. for ref. architecture)		
More Information (URLs)		

Government Operation> Use Case 3: Statistical Survey Response Improvement

Use Case Title	Statistical Survey Response Improvement (Adaptive Design)	
Vertical (area)	Government Statistical Logistics	
Author/Company/Email	Cavan Capps: U.S. Census Bureau/cavan.paul.capps@census.gov	
Actors/Stakeholders and their roles and responsibilities	U.S. statistical agencies are charged to be the leading authoritative sources about the nation's people and economy, while honoring privacy and rigorously protecting confidentiality. This is done by working with states, local governments and other government agencies.	
Goals	To use advanced methods, that are open and scientifically objective, the statistical agencies endeavor to improve the quality, the specificity and the timeliness of statistics provided while reducing operational costs and maintaining the confidentiality of those measured.	
Use Case Description	Survey costs are increasing as survey response declines. The goal of this work is to use advanced "recommendation system techniques" using data mashed up from several sources and historical survey para-data to drive operational processes in an effort to increase quality and reduce the cost of field surveys.	
Current Solutions	**Compute(System)**	Linux systems
	Storage	SAN and Direct Storage
	Networking	Fiber, 10 gigabit Ethernet, Infiniband 40 gigabit.
	Software	Hadoop, Spark, Hive, R, SAS, Mahout, Allegrograph, MySQL, Oracle, Storm, BigMemory, Cassandra, Pig
Big Data Characteristics	**Data Source (distributed/centralized)**	Survey data, other government administrative data, geographical positioning data from various sources.
	Volume (size)	For this particular class of operational problem approximately one petabyte.
	Velocity (e.g. real time)	Varies, paradata from field data streamed continuously, during the decennial census approximately 150 million records transmitted.
	Variety (multiple datasets, mashup)	Data is typically defined strings and numerical fields. Data can be from multiple datasets mashed together for analytical use.
	Variability (rate of change)	Varies depending on surveys in the field at a given time. High rate of velocity during a decennial census.
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues, semantics)**	Data must have high veracity and systems must be very robust. The semantic integrity of conceptual metadata concerning what exactly is measured and the resulting limits of inference remain a challenge
	Visualization	Data visualization is useful for data review, operational activity and general analysis. It continues to evolve.
	Data Quality (syntax)	Data quality should be high and statistically checked for accuracy and reliability throughout the collection process.
	Data Types	Pre-defined ASCII strings and numerical data
	Data Analytics	Analytics are required for recommendation systems, continued monitoring and general survey improvement.
Big Data Specific Challenges (Gaps)	Improving recommendation systems that reduce costs and improve quality while providing confidentiality safeguards that are reliable and publicly auditable.	
Big Data Specific Challenges in Mobility	Mobile access is important.	
Security and Privacy	All data must be both confidential and secure. All processes must be auditable for	

Government Operation> Use Case 3: Statistical Survey Response Improvement

Requirements	security and confidentiality as required by various legal statutes.
Highlight issues for generalizing this use case (e.g. for ref. architecture)	Recommender systems have features in common to e-commerce like Amazon, Netflix, UPS etc.
More Information (URLs)	

Government Operation> Use Case 4: Non-Traditional Data in Statistical Survey

Use Case Title	Non-Traditional Data in Statistical Survey Response Improvement (Adaptive Design)	
Vertical (area)	Government Statistical Logistics	
Author/Company/Email	Cavan Capps: U.S. Census Bureau / cavan.paul.capps@census.gov	
Actors/Stakeholders and their roles and responsibilities	U.S. statistical agencies are charged to be the leading authoritative sources about the nation's people and economy, while honoring privacy and rigorously protecting confidentiality. This is done by working with states, local governments and other government agencies.	
Goals	To use advanced methods, that are open and scientifically objective, the statistical agencies endeavor to improve the quality, the specificity and the timeliness of statistics provided while reducing operational costs and maintaining the confidentiality of those measured.	
Use Case Description	Survey costs are increasing as survey response declines. The potential of using non-traditional commercial and public data sources from the web, wireless communication, electronic transactions mashed up analytically with traditional surveys to improve statistics for small area geographies, new measures and to improve the timeliness of released statistics.	
Current Solutions	**Compute(System)**	Linux systems
	Storage	SAN and Direct Storage
	Networking	Fiber, 10 gigabit Ethernet, Infiniband 40 gigabit.
	Software	Hadoop, Spark, Hive, R, SAS, Mahout, Allegrograph, MySQL, Oracle, Storm, BigMemory, Cassandra, Pig
Big Data Characteristics	**Data Source (distributed/centralized)**	Survey data, other government administrative data, web scrapped data, wireless data, e-transaction data, potentially social media data and positioning data from various sources.
	Volume (size)	TBD
	Velocity (e.g. real time)	TBD
	Variety (multiple datasets, mashup)	Textual data as well as the traditionally defined strings and numerical fields. Data can be from multiple datasets mashed together for analytical use.
	Variability (rate of change)	TBD.
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues, semantics)**	Data must have high veracity and systems must be very robust. The semantic integrity of conceptual metadata concerning what exactly is measured and the resulting limits of inference remain a challenge
	Visualization	Data visualization is useful for data review, operational activity and general analysis. It continues to evolve.
	Data Quality (syntax)	Data quality should be high and statistically checked for accuracy and reliability throughout the collection process.
	Data Types	Textual data, pre-defined ASCII strings and numerical data
	Data Analytics	Analytics are required to create reliable estimates using data from traditional survey sources, government administrative data sources and non-traditional sources from the digital economy.

Government Operation> Use Case 4: Non-Traditional Data in Statistical Survey

Big Data Specific Challenges (Gaps)	Improving analytic and modeling systems that provide reliable and robust statistical estimated using data from multiple sources that are scientifically transparent and while providing confidentiality safeguards that are reliable and publicly auditable.
Big Data Specific Challenges in Mobility	Mobile access is important.
Security and Privacy Requirements	All data must be both confidential and secure. All processes must be auditable for security and confidentiality as required by various legal statutes.
Highlight issues for generalizing this use case (e.g. for ref. architecture)	Statistical estimation that provide more detail, on a more near real time basis for less cost. The reliability of estimated statistics from such “mashed up” sources still must be evaluated.
More Information (URLs)	

Commercial> Use Case 5: Cloud Computing in Financial Industries

Use Case Title	This use case represents one approach to implementing a BD (Big Data) strategy, within a Cloud Eco-System, for FI (Financial Industries) transacting business within the United States.
Vertical (area)	The following lines of business (LOB) include: **Banking**, including: Commercial, Retail, Credit Cards, Consumer Finance, Corporate Banking, Transaction Banking, Trade Finance, and Global Payments. **Securities and Investments**, such as; Retail Brokerage, Private Banking/Wealth Management, Institutional Brokerages, Investment Banking, Trust Banking, Asset Management, Custody and Clearing Services **Insurance**, including; Personal and Group Life, Personal and Group Property/Casualty, Fixed and Variable Annuities, and Other Investments **Please Note:** Any Public/Private entity, providing financial services within the regulatory and jurisdictional risk and compliance purview of the United States, are required to satisfy a complex multilayer number of regulatory governance, risk management, and compliance (GRC)/ confidentiality, integrity, and availability (CIA) requirements, as overseen by various jurisdictions and agencies, including; Fed., State, Local and cross-border.
Author/Company/Email	Pw Carey, Compliance Partners, LLC, pwc.pwcarey@email.com
Actors/Stakeholders and their roles and responsibilities	Regulatory and advisory organizations and agencies including the; SEC (Securities and Exchange Commission), FDIC (Federal Deposit Insurance Corporation), CFTC (Commodity Futures Trading Commission), US Treasury, PCAOB (Public Company Accounting and Oversight Board), COSO, CobiT, reporting supply chains and stakeholders, investment community, shareholders, pension funds, executive management, data custodians, and employees. At each level of a financial services organization, an inter-related and inter-dependent mix of duties, obligations and responsibilities are in-place, which are directly responsible for the performance, preparation and transmittal of financial data, thereby satisfying both the regulatory GRC and CIA of their organizations financial data. This same information is directly tied to the continuing reputation, trust and survivability of an organization's business.
Goals	The following represents one approach to developing a workable BD/FI strategy within the financial services industry. Prior to initiation and switch-over, an organization must perform the following baseline methodology for utilizing BD/FI within a Cloud Eco-system for both public and private financial entities offering financial services within the regulatory confines of the United States; Federal, State, Local and/or cross-border such as the UK, EU and China. Each financial services organization must approach the following disciplines supporting their BD/FI initiative, with an understanding and appreciation for the impact each of the following four overlaying and inter-dependent forces will play in a workable implementation. These four areas are: 1. People (resources), 2. Processes (time/cost/ROI), 3. Technology (various operating systems, platforms and footprints) and 4. Regulatory Governance (subject to various and multiple regulatory agencies). In addition, these four areas must work through the process of being; identified, analyzed, evaluated, addressed, tested, and reviewed in preparation for attending to the following implementation phases: 1. Project Initiation and Management Buy-in 2. Risk Evaluations and Controls 3. Business Impact Analysis

Commercial> Use Case 5: Cloud Computing in Financial Industries

	4. Design, Development and Testing of the Business Continuity Strategies 5. Emergency Response and Operations (aka; Disaster Recovery) 6. Developing and Implementing Business Continuity Plans 7. Awareness and Training Programs 8. Maintaining and Exercising Business Continuity, (aka: Maintaining Regulatory Currency) **Please Note:** Whenever appropriate, these eight areas should be tailored and modified to fit the requirements of each organizations unique and specific corporate culture and line of financial services.	
Use Case Description	Big Data as developed by Google was intended to serve as an Internet Web site indexing tool to help them sort, shuffle, categorize and label the Internet. At the outset, it was not viewed as a replacement for legacy IT data infrastructures. With the spin-off development within OpenGroup and Hadoop, Big Data has evolved into a robust data analysis and storage tool that is still undergoing development. However, in the end, Big Data is still being developed as an adjunct to the current IT client/server/ big iron data warehouse architectures which is better at some things, than these same data warehouse environments, but not others. Currently within FI, BD/Hadoop is used for fraud detection, risk analysis and assessments as well as improving the organizations knowledge and understanding of the customers via a strategy known as....'know your customer', pretty clever, eh? However, this strategy still must be following a well thought out taxonomy that satisfies the entities unique, and individual requirements. One such strategy is the following formal methodology which address two fundamental yet paramount questions; “What are we doing”? and “Why are we doing it”? 1). Policy Statement/Project Charter (Goal of the Plan, Reasons and Resources....define each), 2). Business Impact Analysis (how does effort improve our business services), 3). Identify System-wide Policies, Procedures and Requirements, 4). Identify Best Practices for Implementation (including Change Management/ Configuration Management) and/or Future Enhancements, 5). Plan B-Recovery Strategies (how and what will need to be recovered, if necessary), 6). Plan Development (Write the Plan and Implement the Plan Elements), 7). Plan buy-in and Testing (important everyone Knows the Plan, and Knows What to Do), and 8). Implement the Plan (then identify and fix gaps during first 3 months, 6 months, and annually after initial implementation) 9). Maintenance (Continuous monitoring and updates to reflect the current enterprise environment) 10). Lastly, System Retirement	
Current Solutions	**Compute(System)**	Currently, Big Data/Hadoop within a Cloud Eco-system within the FI is operating as part of a hybrid system, with BD being utilized as a useful tool for conducting risk and fraud analysis, in addition to assisting in organizations in the process of ('know your customer'). These are three areas where BD has proven to be good at; 1. detecting fraud, 2. associated risks and a 3. 'know your customer' strategy. At the same time, the traditional client/server/data warehouse/RDBMS are used for the handling, processing,

Commercial> Use Case 5: Cloud Computing in Financial Industries

		storage and archival of the entities financial data. Recently the SEC has approved the initiative for requiring the FI to submit financial statements via the XBRL (extensible Business-Related Markup Language), as of May 13th, 2013.
	Storage	The same Federal, State, Local and cross-border legislative and regulatory requirements can impact any and all geographical locations, including; VMware, NetApps, Oracle, IBM, Brocade, et cetera. **Please Note:** Based upon legislative and regulatory concerns, these storage solutions for FI data must ensure this same data conforms to US regulatory compliance for GRC/CIA, at this point in time. For confirmation, please visit the following agencies web sites: SEC (U.S. Security and Exchange Commission), CFTC (U.S. Commodity Futures Trading Commission), FDIC (U.S. Federal Deposit Insurance Corporation), DOJ (U.S. Department of Justice), and my favorite the PCAOB (Public Company Accounting and Oversight Board).
	Networking	**Please Note:** The same Federal, State, Local and cross-border legislative and regulatory requirements can impact any and all geographical locations of HW/SW, including but not limited to; WANs, LANs, MANs WiFi, fiber optics, Internet Access, via Public, Private, Community and Hybrid Cloud environments, with or without VPNs. Based upon legislative and regulatory concerns, these networking solutions for FI data must ensure this same data conforms to US regulatory compliance for GRC/CIA, such as the US Treasury Dept., at this point in time. For confirmation, please visit the following agencies web sites: SEC, CFTC, FDIC, US Treasury Dept., DOJ, and my favorite the PCAOB (Public Company Accounting and Oversight Board).
	Software	**Please Note:** The same legislative and regulatory obligations impacting the geographical location of HW/SW, also restricts the location for; Hadoop, Map/Reduce, Open-source, and/or Vendor Proprietary such as AWS (Amazon Web Services), Google Cloud Services, and Microsoft Based upon legislative and regulatory concerns, these software solutions incorporating both SOAP (Simple Object Access Protocol), for Web development and OLAP (online analytical processing) software language for databases, specifically in this case for FI data, both must ensure this same data conforms to US regulatory compliance for GRC/CIA, at this point in time. For confirmation, please visit the following agencies web sites: SEC, CFTC, U.S. Treasury, FDIC, DOJ, and my favorite the PCAOB (Public Company Accounting and Oversight Board).

Commercial> Use Case 5: Cloud Computing in Financial Industries

Big Data Characteristics	**Data Source (distributed/ centralized)**	**Please Note:** The same legislative and regulatory obligations impacting the geographical location of HW/SW, also impacts the location for; both distributed/centralized data sources flowing into HA/DR Environment and HVSs (Hosted Virtual Servers), such as the following constructs: DC1---> VMWare/KVM (Clusters, w/Virtual Firewalls), Data link-Vmware Link-Vmotion Link-Network Link, Multiple PB of NaaS (Network as a Service), DC2--->, VMWare/KVM (Clusters w/Virtual Firewalls), DataLink (Vmware Link, Vmotion Link, Network Link), Multiple PB of NaaS, (Requires Fail-Over Virtualization), among other considerations. Based upon legislative and regulatory concerns, these data source solutions, either distributed and/or centralized for FI data, must ensure this same data conforms to US regulatory compliance for GRC/CIA, at this point in time. For confirmation, please visit the following agencies web sites: SEC, CFTC, US Treasury, FDIC, DOJ, and my favorite the PCAOB (Public Company Accounting and Oversight Board).
	Volume (size)	Tera-bytes up to Peta-bytes**.** **Please Note: This is a 'Floppy Free Zone'.**
	Velocity (e.g. real time)	Velocity is more important for fraud detection, risk assessments and the 'know your customer' initiative within the BD FI. **Please Note:** However, based upon legislative and regulatory concerns, **velocity** is not at issue regarding BD solutions for FI data, except for fraud detection, risk analysis and customer analysis. Based upon legislative and regulatory restrictions, **velocity** is not at issue, rather the primary concern for FI data, is that it must satisfy all US regulatory compliance obligations for GRC/CIA, at this point in time.
	Variety (multiple datasets, mash-up)	Multiple virtual environments either operating within a batch processing architecture or a hot-swappable parallel architecture supporting fraud detection, risk assessments and customer service solutions. **Please Note:** Based upon legislative and regulatory concerns, **variety** is not at issue regarding BD solutions for FI data within a Cloud Eco-system, except for fraud detection, risk analysis and customer analysis. Based upon legislative and regulatory restrictions, **variety** is not at issue, rather the primary concern for FI data, is that it must satisfy all US regulatory compliance obligations for GRC/CIA, at this point in time.
	Variability (rate of change)	**Please Note:** Based upon legislative and regulatory concerns, **variability** is not at issue regarding BD solutions for FI data within a Cloud Eco-system, except for fraud detection, risk analysis and customer analysis.

Commercial> Use Case 5: Cloud Computing in Financial Industries

		Based upon legislative and regulatory restrictions, **variability** is not at issue, rather the primary concern for FI data, is that it must satisfy all US regulatory compliance obligations for GRC/CIA, at this point in time. Variability with BD FI within a Cloud Eco-System will depending upon the strength and completeness of the SLA agreements, the costs associated with (CapEx), and depending upon the requirements of the business.
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues)**	**Please Note:** Based upon legislative and regulatory concerns, **veracity** is not at issue regarding BD solutions for FI data within a Cloud Eco-system, except for fraud detection, risk analysis and customer analysis. Based upon legislative and regulatory restrictions, **veracity** is not at issue, rather the primary concern for FI data, is that it must satisfy all US regulatory compliance obligations for GRC/CIA, at this point in time. Within a Big Data Cloud Eco-System, data integrity is important over the entire life cycle of the organization due to regulatory and compliance issues related to individual data privacy and security, in the areas of CIA and GRC requirements.
	Visualization	**Please Note:** Based upon legislative and regulatory concerns, **visualization** is not at issue regarding BD solutions for FI data, except for fraud detection, risk analysis and customer analysis, FI data is handled by traditional client/server/data warehouse big iron servers. Based upon legislative and regulatory restrictions, **visualization** is not at issue, rather the primary concern for FI data, is that it must satisfy all US regulatory compliance obligations for GRC/CIA, at this point in time. Data integrity within BD is critical and essential over the entire life-cycle of the organization due to regulatory and compliance issues related to CIA and GRC requirements.
	Data Quality	**Please Note:** Based upon legislative and regulatory concerns, **data quality** will always be an issue, regardless of the industry or platform. Based upon legislative and regulatory restrictions, **data quality** is at the core of data integrity, and is the primary concern for FI data, in that it must satisfy all US regulatory compliance obligations for GRC/CIA, at this point in time. For BD/FI data, data integrity is critical and essential over the entire life-cycle of the organization due to regulatory and compliance issues related to CIA and GRC requirements.
	Data Types	**Please Note:** Based upon legislative and regulatory concerns, **data types** are important in that it must have a degree of consistency and especially survivability during audits and digital forensic investigations where the data format deterioration can negatively impact both an audit

Commercial> Use Case 5: Cloud Computing in Financial Industries

<table>
<tr><td rowspan="2"></td><td></td><td>and a forensic investigation when passed through multiple cycles.
For BD/FI data, multiple data types and formats, include but is not limited to; flat files, .txt, .pdf, android application files, .wav, .jpg and VOIP (Voice over IP)</td></tr>
<tr><td>Data Analytics</td><td>Please Note: Based upon legislative and regulatory concerns, data analytics is an issue regarding BD solutions for FI data, especially in regards to fraud detection, risk analysis and customer analysis.
However, data analytics for FI data is currently handled by traditional client/server/data warehouse big iron servers which must ensure they comply with and satisfy all United States GRC/CIA requirements, at this point in time.
For BD/FI data analytics must be maintained in a format that is non-destructive during search and analysis processing and procedures.</td></tr>
<tr><td>Big Data Specific Challenges (Gaps)</td><td colspan="2">Currently, the areas of concern associated with BD/FI with a Cloud Eco-system, include the aggregating and storing of data (sensitive, toxic and otherwise) from multiple sources which can and does create administrative and management problems related to the following:
Access control
Management/Administration
Data entitlement and
Data ownership
However, based upon current analysis, these concerns and issues are widely known and are being addressed at this point in time, via the Research and Development SDLC/HDLC (Software Development Life Cycle/Hardware Development Life Cycle) sausage makers of technology. Please stay tuned for future developments in this regard</td></tr>
<tr><td>Big Data Specific Challenges in Mobility</td><td colspan="2">Mobility is a continuously growing layer of technical complexity; however, not all Big Data mobility solutions are technical in nature. There are two interrelated and co-dependent parties who required to work together to find a workable and maintainable solution, the FI business side and IT. When both are in agreement sharing a, common lexicon, taxonomy and appreciation and understand for the requirements each is obligated to satisfy, these technical issues can be addressed.
Both sides in this collaborative effort will encounter the following current and on-going FI data considerations:
Inconsistent category assignments
Changes to classification systems over time
Use of multiple overlapping or
Different categorization schemes
In addition, each of these changing and evolving inconsistencies, are required to satisfy the following data characteristics associated with ACID:
Atomic- All of the work in a transaction completes (commit) or none of it completes
Consistent- A transmittal transforms the database from one consistent state to another consistent state. Consistency is defined in terms of constraints.
Isolated- The results of any changes made during a transaction are not visible until the transaction has committed.</td></tr>
</table>

Commercial> Use Case 5: Cloud Computing in Financial Industries

	• **Durable**- The results of a committed transaction survive failures. When each of these data categories is satisfied, well, it's a glorious thing. Unfortunately, sometimes glory is not in the room, however, that does not mean we give up the effort to resolve these issues.
Security and Privacy Requirements	No amount of security and privacy due diligence will make up for the innate deficiencies associated with human nature that creep into any program and/or strategy. Currently, the BD/FI must contend with a growing number of risk buckets, such as: • AML-Anti-Money Laundering • CDD- Client Due Diligence • Watch-lists • FCPA – Foreign Corrupt Practices Act ...to name a few. For a reality check, please consider Mr. Harry M. Markopolos' nine-year effort to get the SEC among other agencies to do their job and shut down Mr. Bernard Madoff's billion dollar Ponzi scheme. However, that aside, identifying and addressing the privacy/security requirements of the FI, providing services within a BD/Cloud Eco-system, via continuous improvements in: 1. technology, 2. processes, 3. procedures, 4. people and 5. regulatory jurisdictions ...is a far better choice for both the individual and the organization, especially when considering the alternative. Utilizing a layered approach, this strategy can be broken down into the following sub categories: 1. Maintaining operational resilience 2. Protecting valuable assets 3. Controlling system accounts 4. Managing security services effectively, and 5. Maintaining operational resilience For additional background security and privacy solutions addressing both security and privacy, we'll refer you to the two following organizations: • ISACA (International Society of Auditors and Computer Analysts) • isc2 (International Security Computer and Systems Auditors)
Highlight issues for generalizing this use case (e.g. for ref. architecture)	Areas of concern include the aggregating and storing data from multiple sources can create problems related to the following: • Access control • Management/Administration • Data entitlement and • Data ownership Each of these areas is being improved upon, yet they still must be considered and addressed, via access control solutions, and SIEM (Security Incident/Event Management) tools. I don't believe we're there yet, based upon current security concerns mentioned whenever Big Data/Hadoop within a Cloud Eco-system is brought up in polite conversation.

Commercial> Use Case 5: Cloud Computing in Financial Industries

	Current and on-going challenges to implementing BD Finance within a Cloud Eco, as well as traditional client/server data warehouse architectures, include the following areas of Financial Accounting under both US GAAP (U.S. Generally Accepted Accounting Practices) or IFRS (International Financial Reporting Standards): XBRL (extensible Business-Related Markup Language) Consistency (terminology, formatting, technologies, regulatory gaps) SEC mandated use of XBRL (extensible Business-Related Markup Language) for regulatory financial reporting. SEC, GAAP/IFRS and the yet to be fully resolved new financial legislation impacting reporting requirements are changing and point to trying to improve the implementation, testing, training, reporting and communication best practices required of an independent auditor, regarding: Auditing, Auditor's reports, Control self-assessments, Financial audits, GAAS / ISAs, Internal audits, and the Sarbanes–Oxley Act of 2002 (SOX).
More Information (URLs)	1. Cloud Security Alliance Big Data Working Group, “Top 10 Challenges in Big Data Security and Privacy”, 2012. 2. The IFRS, Securities and Markets Working Group, http://www.xbrl-eu.org 3. IEEE Big Data conference http://www.ischool.drexel.edu/bigdata/bigdata2013/topics.htm 4. Map/Reduce http://www.mapreduce.org. 5. PCAOB http://www.pcaob.org 6. http://www.ey.com/GL/en/Industries/Financial-Services/Insurance 7. http://www.treasury.gov/resource-center/fin-mkts/Pages/default.aspx 8. CFTC http://www.cftc.org 9. SEC http://www.sec.gov 10. FDIC http://www.fdic.gov 11. COSO http://www.coso.org 12. isc2 International Information Systems Security Certification Consortium, Inc.: http://www.isc2.org 13. ISACA Information Systems Audit and Control Association: http://www.isca.org 14. IFARS http://www.ifars.org 15. Apache http://www.opengroup.org 16. http://www.computerworld.com/s/article/print/9221652/IT_must_prepare_for_Hadoop_security_issues?tax ... 17. "No One Would Listen: A True Financial Thriller" (hard-cover book). Hoboken, NJ: John Wiley & Sons. March 2010. Retrieved April 30, 2010. ISBN 978-0-470-55373-2 18. Assessing the Madoff Ponzi Scheme and Regulatory Failures (Archive of: Subcommittee on Capital Markets, Insurance, and Government Sponsored Enterprises Hearing) (http://financialserv.edgeboss.net/wmedia/financialserv/hearing020409.wvx) (Windows Media). U.S. House Financial Services Committee. February 4, 2009. Retrieved June 29, 2009. 19. COSO, The Committee of Sponsoring Organizations of the Treadway Commission (COSO), Copyright© 2013, http://www.coso.org. 20. (ITIL) Information Technology Infrastructure Library, Copyright© 2007-13 APM Group Ltd. All rights reserved, Registered in England No. 2861902, http://www.itil-officialsite.com. 21. CobiT, Ver. 5.0, 2013, ISACA, Information Systems Audit and Control Association, (a framework for IT Governance and Controls), http://www.isaca.org. 22. TOGAF, Ver. 9.1, The Open Group Architecture Framework (a framework for IT architecture), http://www.opengroup.org.

Commercial> Use Case 5: Cloud Computing in Financial Industries

	23. ISO/IEC 27000:2012 Info. Security Mgt., International Organization for Standardization and the International Electrotechnical Commission, http://www.standards.iso.org/
Note: Please feel free to improve our **INITIAL DRAFT, Ver. 0.1, August 25th, 2013**....as we do not consider our efforts to be pearls, at this point in time......Respectfully yours, Pw Carey, Compliance Partners, LLC_pwc.pwcarey@gmail.com	

Commercial> Use Case 6: Mendeley—An International Network of Research

Use Case Title	Mendeley – An International Network of Research	
Vertical (area)	Commercial Cloud Consumer Services	
Author/Company/Email	William Gunn / Mendeley / william.gunn@mendeley.com	
Actors/Stakeholders and their roles and responsibilities	Researchers, librarians, publishers, and funding organizations.	
Goals	To promote more rapid advancement in scientific research by enabling researchers to efficiently collaborate, librarians to understand researcher needs, publishers to distribute research findings more quickly and broadly, and funding organizations to better understand the impact of the projects they fund.	
Use Case Description	Mendeley has built a database of research documents and facilitates the creation of shared bibliographies. Mendeley uses the information collected about research reading patterns and other activities conducted via the software to build more efficient literature discovery and analysis tools. Text mining and classification systems enables automatic recommendation of relevant research, improving the cost and performance of research teams, particularly those engaged in curation of literature on a particular subject, such as the Mouse Genome Informatics group at Jackson Labs, which has a large team of manual curators who scan the literature. Other use cases include enabling publishers to more rapidly disseminate publications, facilitating research institutions and librarians with data management plan compliance, and enabling funders to better understand the impact of the work they fund via real-time data on the access and use of funded research.	
Current Solutions	**Compute(System)**	Amazon EC2
	Storage	HDFS Amazon S3
	Networking	Client-server connections between Mendeley and end user machines, connections between Mendeley offices and Amazon services.
	Software	Hadoop, Scribe, Hive, Mahout, Python
Big Data Characteristics	**Data Source (distributed/centralized)**	Distributed and centralized
	Volume (size)	15TB presently, growing about 1 TB/month
	Velocity (e.g. real time)	Currently Hadoop batch jobs are scheduled daily, but work has begun on real-time recommendation
	Variety (multiple datasets, mashup)	PDF documents and log files of social network and client activities
	Variability (rate of change)	Currently a high rate of growth as more researchers sign up for the service, highly fluctuating activity over the course of the year
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues)**	Metadata extraction from PDFs is variable, it's challenging to identify duplicates, there's no universal identifier system for documents or authors (though ORCID proposes to be this)
	Visualization	Network visualization via Gephi, scatterplots of readership vs. citation rate, etc.
	Data Quality	90% correct metadata extraction according to comparison with Crossref, Pubmed, and Arxiv
	Data Types	Mostly PDFs, some image, spreadsheet, and presentation files

Commercial> Use Case 6: Mendeley—An International Network of Research

	Data Analytics	Standard libraries for machine learning and analytics, LDA, custom built reporting tools for aggregating readership and social activities per document
Big Data Specific Challenges (Gaps)	The database contains ≈400M documents, roughly 80M unique documents, and receives 5-700k new uploads on a weekday. Thus, a major challenge is clustering matching documents together in a computationally efficient way (scalable and parallelized) when they're uploaded from different sources and have been slightly modified via third-part annotation tools or publisher watermarks and cover pages	
Big Data Specific Challenges in Mobility	Delivering content and services to various computing platforms from Windows desktops to Android and iOS mobile devices	
Security and Privacy Requirements	Researchers often want to keep what they're reading private, especially industry researchers, so the data about who's reading what has access controls.	
Highlight issues for generalizing this use case (e.g. for ref. architecture)	This use case could be generalized to providing content-based recommendations to various scenarios of information consumption	
More Information (URLs)	http://mendeley.com http://dev.mendeley.com	

Commercial> Use Case 7: Netflix Movie Service

Use Case Title	Netflix Movie Service	
Vertical (area)	Commercial Cloud Consumer Services	
Author/Company/Email	Geoffrey Fox, Indiana University gcf@indiana.edu	
Actors/Stakeholders and their roles and responsibilities	Netflix Company (Grow sustainable Business), Cloud Provider (Support streaming and data analysis), Client user (Identify and watch good movies on demand)	
Goals	Allow streaming of user selected movies to satisfy multiple objectives (for different stakeholders) -- especially retaining subscribers. Find best possible ordering of a set of videos for a user (household) within a given context in real time; maximize movie consumption.	
Use Case Description	Digital movies stored in cloud with metadata; user profiles and rankings for small fraction of movies for each user. Use multiple criteria – content based recommender system; user-based recommender system; diversity. Refine algorithms continuously with A/B testing.	
Current Solutions	**Compute(System)**	Amazon Web Services AWS
	Storage	Uses Cassandra NoSQL technology with Hive, Teradata
	Networking	Need Content Delivery System to support effective streaming video
	Software	Hadoop and Pig; Cassandra; Teradata
Big Data Characteristics	**Data Source (distributed/centralized)**	Add movies institutionally. Collect user rankings and profiles in a distributed fashion
	Volume (size)	Summer 2012. 25 million subscribers; 4 million ratings per day; 3 million searches per day; 1 billion hours streamed in June 2012. Cloud storage 2 petabytes (June 2013)
	Velocity (e.g. real time)	Media (video and properties) and Rankings continually updated
	Variety (multiple datasets, mashup)	Data varies from digital media to user rankings, user profiles and media properties for content-based recommendations
	Variability (rate of change)	Very competitive business. Need to aware of other companies and trends in both content (which Movies are hot) and technology. Need to investigate new business initiatives such as Netflix sponsored content
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues)**	Success of business requires excellent quality of service
	Visualization	Streaming media and quality user-experience to allow choice of content
	Data Quality	Rankings are intrinsically “rough” data and need robust learning algorithms
	Data Types	Media content, user profiles, “bag” of user rankings
	Data Analytics	Recommender systems and streaming video delivery. Recommender systems are always personalized and use logistic/linear regression, elastic nets, matrix factorization, clustering, latent Dirichlet allocation, association rules, gradient boosted decision trees and others. Winner of Netflix competition (to improve ratings by 10%) combined over 100 different algorithms.

Commercial> Use Case 7: Netflix Movie Service

Big Data Specific Challenges (Gaps)	Analytics needs continued monitoring and improvement.
Big Data Specific Challenges in Mobility	Mobile access important
Security and Privacy Requirements	Need to preserve privacy for users and digital rights for media.
Highlight issues for generalizing this use case (e.g. for ref. architecture)	Recommender systems have features in common to e-commerce like Amazon. Streaming video has features in common with other content providing services like iTunes, Google Play, Pandora and Last.fm
More Information (URLs)	http://www.slideshare.net/xamat/building-largescale-realworld-recommender-systems-recsys2012-tutorial by Xavier Amatriain http://techblog.netflix.com/

Commercial> Use Case 8: Web Search

Use Case Title	Web Search (Bing, Google, Yahoo...)	
Vertical (area)	Commercial Cloud Consumer Services	
Author/Company/Email	Geoffrey Fox, Indiana University gcf@indiana.edu	
Actors/Stakeholders and their roles and responsibilities	Owners of web information being searched; search engine companies; advertisers; users	
Goals	Return in ≈0.1 seconds, the results of a search based on average of 3 words; important to maximize "precision@10"; number of great responses in top 10 ranked results	
Use Case Description	1) Crawl the web; 2) Pre-process data to get searchable things (words, positions); 3) Form Inverted Index mapping words to documents; 4) Rank relevance of documents: PageRank; 5) Lots of technology for advertising, "reverse engineering ranking" "preventing reverse engineering"; 6) Clustering of documents into topics (as in Google News) 7) Update results efficiently	
Current Solutions	**Compute(System)**	Large Clouds
	Storage	Inverted Index not huge; crawled documents are petabytes of text – rich media much more
	Networking	Need excellent external network links; most operations pleasingly parallel and I/O sensitive. High performance internal network not needed
	Software	Map/Reduce + Bigtable; Dryad + Cosmos. PageRank. Final step essentially a recommender engine
Big Data Characteristics	**Data Source (distributed/centralized)**	Distributed web sites
	Volume (size)	45B web pages total, 500M photos uploaded each day, 100 hours of video uploaded to YouTube each minute
	Velocity (e.g. real time)	Data continually updated
	Variety (multiple datasets, mashup)	Rich set of functions. After processing, data similar for each page (except for media types)
	Variability (rate of change)	Average page has life of a few months
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues)**	Exact results not essential but important to get main hubs and authorities for search query
	Visualization	Not important although page layout critical
	Data Quality	A lot of duplication and spam
	Data Types	Mainly text but more interest in rapidly growing image and video
	Data Analytics	Crawling; searching including topic based search; ranking; recommending
Big Data Specific Challenges (Gaps)	Search of "deep web" (information behind query front ends) Ranking of responses sensitive to intrinsic value (as in Pagerank) as well as advertising value Link to user profiles and social network data	
Big Data Specific Challenges in Mobility	Mobile search must have similar interfaces/results	
Security and Privacy Requirements	Need to be sensitive to crawling restrictions. Avoid Spam results	

Commercial> Use Case 8: Web Search

Highlight issues for generalizing this use case (e.g. for ref. architecture)	Relation to Information retrieval such as search of scholarly works.
More Information (URLs)	http://www.slideshare.net/kleinerperkins/kpcb-Internet-trends-2013 http://webcourse.cs.technion.ac.il/236621/Winter2011-2012/en/ho_Lectures.html http://www.ifis.cs.tu-bs.de/teaching/ss-11/irws http://www.slideshare.net/beechung/recommender-systems-tutorialpart1intro http://www.worldwidewebsize.com/

Commercial> Use Case 9: Cloud-based Continuity and Disaster Recovery

Use Case Title	IaaS (Infrastructure as a Service) Big Data BC/DR Within a Cloud Eco-System provided by Cloud Service Providers (CSPs) and Cloud Brokerage Service Providers (CBSPs)
Vertical (area)	Large Scale Reliable Data Storage
Author/Company/Email	Pw Carey, Compliance Partners, LLC, pwc.pwcarey@email.com
Actors/Stakeholders and their roles and responsibilities	Executive Management, Data Custodians, and Employees responsible for the integrity, protection, privacy, confidentiality, availability, safety, security and survivability of a business by ensuring the 3-As of data accessibility to an organizations services are satisfied; anytime, anyplace and on any device.
Goals	The following represents one approach to developing a workable BC/DR strategy. Prior to outsourcing an organizations BC/DR onto the backs/shoulders of a CSP or CBSP, the organization must perform the following Use Case, which will provide each organization with a baseline methodology for BC/DR best practices, within a Cloud Eco-system for both Public and Private organizations. Each organization must approach the ten disciplines supporting BC/DR, with an understanding and appreciation for the impact each of the following four overlaying and inter-dependent 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 (resources), processes (time/cost/ROI), technology (various operating systems, platforms and footprints) and governance (subject to various and multiple regulatory agencies). These four concerns must be; identified, analyzed, evaluated, addressed, tested, reviewed, addressed during the following ten phases: 1. Project Initiation and Management Buy-in 2. Risk Evaluations and Controls 3. Business Impact Analysis 4. Design, Development and Testing of the Business Continuity Strategies 5. Emergency Response and Operations (aka; Disaster Recovery 6. Developing and Implementing Business Continuity Plans 7. Awareness and Training Programs 8. Maintaining and Exercising Business Continuity Plans, (aka: Maintaining Currency) 9. Public Relations (PR) and Crises Management Plans 10. Coordination with Public Agencies Please Note: When appropriate, these ten areas can be tailored to fit the requirements of the organization.
Use Case Description	Big Data as developed by Google was intended to serve as an Internet Web site indexing tool to help them sort, shuffle, categorize and label the Internet. At the outset, it was not viewed as a replacement for legacy IT data infrastructures. With the spin-off development within OpenGroup and Hadoop, Big Data has evolved into a robust data analysis and storage tool that is still undergoing development. However, in the end, Big Data is still being developed as an adjunct to the current IT client/server/big iron data warehouse architectures which is better at some things, than these same data warehouse environments, but not others. As a result, it is necessary, within this business continuity/disaster recovery use case, we ask good questions, such as; why are we doing this and what are we trying to accomplish? What are our dependencies upon manual practices and when can we leverage them? What systems have been and remain outsourced to other organizations, such as our Telephony and what are their DR/BC business functions, if any? Lastly, we must recognize the functions that can be simplified and what are the

Commercial> Use Case 9: Cloud-based Continuity and Disaster Recovery

	preventative steps we can take that do not have a high cost associated with them such as simplifying business practices. We must identify what are the critical business functions that need to be recovered, 1st, 2nd, 3rd in priority, or at a later time/date, and what is the Model of a Disaster we're trying to resolve, what are the types of disasters more likely to occur realizing that we don't need to resolve all types of disasters. When backing up data within a Cloud Eco-system is a good solution, this will shorten the fail-over time and satisfy the requirements of RTO/RPO. In addition, there must be 'Buy-in', as this is not just an IT problem; it is a business services problem as well, requiring the testing of the Disaster Plan via formal walk-throughs, et cetera. There should be a formal methodology for developing a BC/DR Plan, including: 1). Policy Statement (Goal of the Plan, Reasons and Resources....define each), 2). Business Impact Analysis (how does a shutdown impact the business financially and otherwise), 3). Identify Preventive Steps (can a disaster be avoided by taking prudent steps), 4). Recovery Strategies (how and what you will need to recover), 5). Plan Development (Write the Plan and Implement the Plan Elements), 6). Plan buy-in and Testing (very important so that everyone knows the Plan and knows what to do during its execution), and 7). Maintenance (Continuous changes to reflect the current enterprise environment)	
Current Solutions	**Compute(System)**	Cloud Eco-systems, incorporating IaaS (Infrastructure as a Service), supported by Tier 3 Data Centers....Secure Fault Tolerant (Power).... for Security, Power, Air Conditioning et cetera...geographically off-site data recovery centers...providing data replication services, Note: Replication is different from Backup. Replication only moves the changes since the last time a replication, including block level changes. The replication can be done quickly, with a five second window, while the data is replicated every four hours. This data snap shot is retained for seven business days, or longer if necessary. Replicated data can be moved to a Fail-over Center to satisfy the organizations RPO (Recovery Point Objectives) and RTO
	Storage	VMware, NetApps, Oracle, IBM, Brocade,
	Networking	WANs, LANs, WiFi, Internet Access, via Public, Private, Community and Hybrid Cloud environments, with or without VPNs.
	Software	Hadoop, Map/Reduce, Open-source, and/or Vendor Proprietary such as AWS (Amazon Web Services), Google Cloud Services, and Microsoft
Big Data Characteristics	**Data Source (distributed /centralized)**	Both distributed/centralized data sources flowing into HA/DR Environment and HVSs, such as the following: DC1---> VMWare/KVM (Clusters, w/Virtual Firewalls), Data link-VMware Link-Vmotion Link-Network Link, Multiple PB of NaaS, DC2--->, VMWare/KVM (Clusters w/Virtual Firewalls), DataLink (VMware Link, Motion Link, Network Link), Multiple PB of NaaS, (Requires Fail-Over Virtualization)
	Volume (size)	Terabytes up to Petabytes

Commercial> Use Case 9: Cloud-based Continuity and Disaster Recovery

	Velocity (e.g. real time)	Tier 3 Data Centers with Secure Fault Tolerant (Power) for Security, Power, and Air Conditioning. IaaS (Infrastructure as a Service) in this example, based upon NetApps. Replication is different from Backup; replication requires only moving the CHANGES since the last time a REPLICATION was performed, including the block level changes. The Replication can be done quickly as the data is Replicated every four hours. These replications can be performed within a 5 second window, and this Snap Shot will be kept for seven business days, or longer if necessary to a Fail-Over Center.....at the RPO and RTO....
	Variety (multiple datasets, mash-up)	Multiple virtual environments either operating within a batch processing architecture or a hot-swappable parallel architecture.
	Variability (rate of change)	Depending upon the SLA agreement, the costs (CapEx) increases, depending upon the RTO/RPO and the requirements of the business.
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues)**	Data integrity is critical and essential over the entire life-cycle of the organization due to regulatory and compliance issues related to data CIA and GRC data requirements.
	Visualization	Data integrity is critical and essential over the entire life-cycle of the organization due to regulatory and compliance issues related to data CIA and GRC data requirements.
	Data Quality	Data integrity is critical and essential over the entire life-cycle of the organization due to regulatory and compliance issues related to data CIA and GRC data requirements.
	Data Types	Multiple data types and formats, including but not limited to; flat files, .txt, .pdf, android application files, .wav, .jpg and VOIP (Voice over IP)
	Data Analytics	Must be maintained in a format that is non-destructive during search and analysis processing and procedures.
Big Data Specific Challenges (Gaps)	The complexities associated with migrating from a Primary Site to either a Replication Site or a Backup Site is not fully automated at this point in time. The goal is to enable the user to automatically initiate the Fail Over Sequence, moving Data Hosted within Cloud requires a well-defined and continuously monitored server configuration management. In addition, both organizations must know which servers have to be restored and what are the dependencies and inter-dependencies between the Primary Site servers and Replication and/or Backup Site servers. This requires a continuous monitoring of both, since there are two solutions involved with this process, either dealing with servers housing stored images or servers running hot all the time, as in running parallel systems with hot-swappable functionality, all of which requires accurate and up-to-date information from the client.	
Big Data Specific Challenges in Mobility	Mobility is a continuously growing layer of technical complexity; however, not all DR/BC solutions are technical in nature, as there are two sides required to work together to find a solution, the business side and the IT side. When they are in agreement, these technical issues must be addressed by the BC/DR strategy	

Commercial> Use Case 9: Cloud-based Continuity and Disaster Recovery

	implemented and maintained by the entire organization. One area, which is not limited to mobility challenges, concerns a fundamental issue impacting most BC/DR solutions. If your Primary Servers (A, B, C) understand X, Y, Z....but your Secondary Virtual Replication/Backup Servers (a, b, c) over the passage of time, are not properly maintained (configuration management) and become out of sync with your Primary Servers, and only understand X, and Y, when called upon to perform a Replication or Back-up, well "Houston, we have a problem...." Please Note: Over time all systems can and will suffer from sync-creep, some more than others, when relying upon manual processes to ensure system stability.
Security and Privacy Requirements	Dependent upon the nature and requirements of the organization's industry verticals, such as; Finance, Insurance, and Life Sciences including both public and/or private entities, and the restrictions placed upon them by; regulatory, compliance and legal jurisdictions.
Highlight issues for generalizing this use case (e.g. for ref. architecture)	Challenges to Implement BC/DR, include the following: 1) Recognition, a). Management Vision, b). Assuming the issue is an IT issue, when it is not just an IT issue, 2). People: a). Staffing levels - Many SMBs are understaffed in IT for their current workload, b). Vision - (Driven from the Top Down) Can the business and IT resources see the whole problem and craft a strategy such a 'Call List' in case of a Disaster, c). Skills - Are there resources that can architect, implement and test a BC/DR Solution, d). Time - Do Resources have the time and does the business have the Windows of Time for constructing and testing a DR/BC Solution as DR/BC is an additional Add-On Project the organization needs the time and resources. 3). Money - This can be turned in to an OpEx Solution rather than a CapEx Solution which and can be controlled by varying RPO/RTO, a). Capital is always a constrained resource, b). BC Solutions need to start with "what is the Risk" and "how does cost constrain the solution"? 4). Disruption - Build BC/DR into the standard "Cloud" infrastructure (IaaS) of the SMB, a). Planning for BC/DR is disruptive to business resources, b). Testing BC is also disruptive.....
More Information (URLs)	1. http://www.disasterrecovery.org/, (March 2013). 2. BC_DR From the Cloud, Avoid IT Disasters EN POINTE Technologies and dinCloud, Webinar Presenter Barry Weber, http://www.dincloud.com. 3. COSO, The Committee of Sponsoring Organizations of the Treadway Commission (COSO), Copyright© 2013, http://www.coso.org. 4. ITIL Information Technology Infrastructure Library, Copyright© 2007-13 APM Group Ltd. All rights reserved, Registered in England No. 2861902, http://www.itil-officialsite.com. 5. CobiT, Ver. 5.0, 2013, ISACA, Information Systems Audit and Control Association, (a framework for IT Governance and Controls), http://www.isaca.org. 6. TOGAF, Ver. 9.1, The Open Group Architecture Framework (a framework for IT architecture), http://www.opengroup.org. 7. ISO/IEC 27000:2012 Info. Security Mgt., International Organization for Standardization and the International Electrotechnical Commission, http://www.standards.iso.org/. 8. PCAOB, Public Company Accounting and Oversight Board, http://www.pcaobus.org.
Note: Please feel free to improve our INITIAL DRAFT, Ver. 0.1, August 10th, 2013....as we do not consider our efforts to be pearls, at this point in time......Respectfully yours, Pw Carey, Compliance Partners, LLC_pwc.pwcarey@gmail.com	

Commercial> Use Case 10: Cargo Shipping

Use Case Title	Cargo Shipping	
Vertical (area)	Industry	
Author/Company/Email	William Miller/MaCT USA/mact-usa@att.net	
Actors/Stakeholders and their roles and responsibilities	End-users (Sender/Recipients) Transport Handlers (Truck/Ship/Plane) Telecom Providers (Cellular/SATCOM) Shippers (Shipping and Receiving)	
Goals	Retention and analysis of items (Things) in transport	
Use Case Description	The following use case defines the overview of a Big Data application related to the shipping industry (i.e., FedEx, UPS, DHL, etc.). The shipping industry represents possible the largest potential use case of Big Data that is in common use today. It relates to the identification, transport, and handling of item (Things) in the supply chain. The identification of an item begins with the sender to the recipients and for all those in between with a need to know the location and time of arrive of the items while in transport. A new aspect will be status condition of the items which will include sensor information, GPS coordinates, and a unique identification schema based upon a new ISO 29161 standards under development within ISO JTC1 SC31 WG2. The data is in near real time being updated when a truck arrives at a depot or upon delivery of the item to the recipient. Intermediate conditions are not currently known; the location is not updated in real time, items lost in a warehouse or while in shipment represent a problem potentially for homeland security. The records are retained in an archive and can be accessed for xx days.	
Current Solutions	**Compute(System)**	Unknown
	Storage	Unknown
	Networking	LAN/T1/Internet Web Pages
	Software	Unknown
Big Data Characteristics	**Data Source (distributed/centralized)**	Centralized today
	Volume (size)	Large
	Velocity (e.g. real time)	The system is not currently real time.
	Variety (multiple datasets, mashup)	Updated when the driver arrives at the depot and download the time and date the items were picked up. This is currently not real time.
	Variability (rate of change)	Today the information is updated only when the items that were checked with a bar code scanner are sent to the central server. The location is not currently displayed in real time.
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues)**	
	Visualization	NONE
	Data Quality	YES
	Data Types	Not Available
	Data Analytics	YES
Big Data Specific Challenges (Gaps)	Provide more rapid assessment of the identity, location, and conditions of the shipments, provide detailed analytics and location of problems in the system in real time.	

Commercial> Use Case 10: Cargo Shipping

Big Data Specific Challenges in Mobility	Currently conditions are not monitored on-board trucks, ships, and aircraft
Security and Privacy Requirements	Security need to be more robust
Highlight issues for generalizing this use case (e.g. for ref. architecture)	This use case includes local data bases as well as the requirement to synchronize with the central server. This operation would eventually extend to mobile device and on-board systems which can track the location of the items and provide real-time update of the information including the status of the conditions, logging, and alerts to individuals who have a need to know.
More Information (URLs)	

See Figure 1: Cargo Shipping – Scenario.

Commercial> Use Case 11: Materials Data

Use Case Title	Materials Data	
Vertical (area)	Manufacturing, Materials Research	
Author/Company/Email	John Rumble, R&R Data Services; jumbleusa@earthlink.net	
Actors/Stakeholders and their roles and responsibilities	Product Designers (Inputters of materials data in CAE) Materials Researchers (Generators of materials data; users in some cases) Materials Testers (Generators of materials data; standards developers) Data distributors (Providers of access to materials, often for profit)	
Goals	Broaden accessibility, quality, and usability; Overcome proprietary barriers to sharing materials data; Create sufficiently large repositories of materials data to support discovery	
Use Case Description	Every physical product is made from a material that has been selected for its properties, cost, and availability. This translates into hundreds of billion dollars of material decisions made every year. In addition, as the Materials Genome Initiative has so effectively pointed out, the adoption of new materials normally takes decades (two to three) rather than a small number of years, in part because data on new materials is not easily available. All actors within the materials life cycle today have access to very limited quantities of materials data, thereby resulting in materials-related decision that are non-optimal, inefficient, and costly. While the Materials Genome Initiative is addressing one major and important aspect of the issue, namely the fundamental materials data necessary to design and test materials computationally, the issues related to physical measurements on physical materials (from basic structural and thermal properties to complex performance properties to properties of novel (nanoscale materials) are not being addressed systematically, broadly (cross-discipline and internationally), or effectively (virtually no materials data meetings, standards groups, or dedicated funded programs). One of the greatest challenges that Big Data approaches can address is predicting the performance of real materials (gram to ton quantities) starting at the atomistic, nanometer, and/or micrometer level of description. As a result of the above considerations, decisions about materials usage are unnecessarily conservative, often based on older rather than newer materials research and development data, and not taking advantage of advances in modeling and simulations. Materials informatics is an area in which the new tools of data science can have major impact.	
Current Solutions	**Compute(System)**	None
	Storage	Widely dispersed with many barriers to access
	Networking	Virtually none
	Software	Narrow approaches based on national programs (Japan, Korea, and China), applications (EU Nuclear program), proprietary solutions (Granta, etc.)
Big Data Characteristics	**Data Source (distributed/centralized)**	Extremely distributed with data repositories existing only for a very few fundamental properties
	Volume (size)	It has been estimated (in the 1980s) that there were over 500,000 commercial materials made in the last fifty years. The last three decades has seen large growth in that number.
	Velocity (e.g. real time)	Computer-designed and theoretically design materials (e.g., nanomaterials) are growing over time

Commercial> Use Case 11: Materials Data

	Variety (multiple datasets, mashup)	Many datasets and virtually no standards for mashups
	Variability (rate of change)	Materials are changing all the time, and new materials data are constantly being generated to describe the new materials
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues)**	More complex material properties can require many (100s?) of independent variables to describe accurately. Virtually no activity no exists that is trying to identify and systematize the collection of these variables to create robust datasets.
	Visualization	Important for materials discovery. Potentially important to understand the dependency of properties on the many independent variables. Virtually unaddressed.
	Data Quality	Except for fundamental data on the structural and thermal properties, data quality is poor or unknown. See Munro's NIST Standard Practice Guide.
	Data Types	Numbers, graphical, images
	Data Analytics	Empirical and narrow in scope
Big Data Specific Challenges (Gaps)	1. Establishing materials data repositories beyond the existing ones that focus on fundamental data 2. Developing internationally-accepted data recording standards that can be used by a very diverse materials community, including developers materials test standards (such as ASTM and ISO), testing companies, materials producers, and research and development labs 3. Tools and procedures to help organizations wishing to deposit proprietary materials in data repositories to mask proprietary information, yet to maintain the usability of data 4. Multi-variable materials data visualization tools, in which the number of variables can be quite high	
Big Data Specific Challenges in Mobility	Not important at this time	
Security and Privacy Requirements	Proprietary nature of many data very sensitive.	
Highlight issues for generalizing this use case (e.g. for ref. architecture)	Development of standards; development of large scale repositories; involving industrial users; integration with CAE (don't underestimate the difficulty of this – materials people are generally not as computer savvy as chemists, bioinformatics people, and engineers)	
More Information (URLs)		

Commercial> Use Case 12: Simulation Driven Materials Genomics

Use Case Title	Simulation driven Materials Genomics	
Vertical (area)	Scientific Research: Materials Science	
Author/Company/Email	David Skinner/LBNL/deskinner@lbl.gov	
Actors/Stakeholders and their roles and responsibilities	Capability providers: National labs and energy hubs provide advanced materials genomics capabilities using computing and data as instruments of discovery. User Community: DOE, industry and academic researchers as a user community seeking capabilities for rapid innovation in materials.	
Goals	Speed the discovery of advanced materials through informatically driven simulation surveys.	
Use Case Description	Innovation of battery technologies through massive simulations spanning wide spaces of possible design. Systematic computational studies of innovation possibilities in photovoltaics. Rational design of materials based on search and simulation.	
Current Solutions	**Compute(System)**	Hopper.nersc.gov (150K cores), omics-like data analytics hardware resources.
	Storage	GPFS, MongoDB
	Networking	10Gb
	Software	PyMatGen, FireWorks, VASP, ABINIT, NWChem, BerkeleyGW, varied community codes
Big Data Characteristics	**Data Source (distributed/centralized)**	Gateway-like. Data streams from simulation surveys driven on centralized peta/exascale systems. Widely distributed web of dataflows from central gateway to users.
	Volume (size)	100TB (current), 500TB within 5 years. Scalable key-value and object store databases needed.
	Velocity (e.g. real time)	High throughput computing (HTC), fine-grained tasking and queuing. Rapid start/stop for ensembles of tasks. Real-time data analysis for web-like responsiveness.
	Variety (multiple datasets, mashup)	Mashup of simulation outputs across codes and levels of theory. Formatting, registration and integration of datasets. Mashups of data across simulation scales.
	Variability (rate of change)	The targets for materials design will become more search and crowd-driven. The computational backend must flexibly adapt to new targets.
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues, semantics)**	Validation and UQ of simulation with experimental data of varied quality. Error checking and bounds estimation from simulation inter-comparison.
	Visualization	Materials browsers as data from search grows. Visual design of materials.
	Data Quality (syntax)	UQ in results based on multiple datasets. Propagation of error in knowledge systems.
	Data Types	Key value pairs, JSON, materials file formats
	Data Analytics	Map/Reduce and search that join simulation and experimental data.
Big Data Specific Challenges (Gaps)	HTC at scale for simulation science. Flexible data methods at scale for messy data. Machine learning and knowledge systems that integrate data from publications, experiments, and simulations to advance goal-driven thinking in materials design.	
Big Data Specific Challenges in Mobility	Potential exists for widespread delivery of actionable knowledge in materials science. Many materials genomics “apps” are amenable to a mobile platform.	

Commercial> Use Case 12: Simulation Driven Materials Genomics

Security and Privacy Requirements	Ability to "sandbox" or create independent working areas between data stakeholders. Policy-driven federation of datasets.
Highlight issues for generalizing this use case (e.g. for ref. architecture)	An OSTP blueprint toward broader materials genomics goals was made available in May 2013.
More Information (URLs)	http://www.materialsproject.org

Defense> Use Case 13: Large Scale Geospatial Analysis and Visualization

Use Case Title	Large Scale Geospatial Analysis and Visualization	
Vertical (area)	Defense – but applicable to many others	
Author/Company/Email	David Boyd/Data Tactics/ dboyd@data-tactics.com	
Actors/Stakeholders and their roles and responsibilities	Geospatial Analysts Decision Makers Policy Makers	
Goals	Support large scale geospatial data analysis and visualization.	
Use Case Description	As the number of geospatially aware sensors increase and the number of geospatially tagged data sources increases the volume geospatial data requiring complex analysis and visualization is growing exponentially. Traditional GIS systems are generally capable of analyzing millions of objects and easily visualizing thousands. Today's intelligence systems often contain trillions of geospatial objects and need to be able to visualize and interact with millions of objects.	
Current Solutions	**Compute(System)**	Compute and Storage systems - Laptops to Large servers (see notes about clusters) Visualization systems - handhelds to laptops
	Storage	Compute and Storage - local disk or SAN Visualization - local disk, flash ram
	Networking	Compute and Storage - Gigabit or better LAN connection Visualization - Gigabit wired connections, Wireless including WiFi (802.11), Cellular (3g/4g), or Radio Relay
	Software	Compute and Storage – generally Linux or Win Server with Geospatially enabled RDBMS, Geospatial server/analysis software – ESRI ArcServer, Geoserver Visualization – Windows, Android, IOS – browser based visualization. Some laptops may have local ArcMap.
Big Data Characteristics	**Data Source (distributed/centralized)**	Very distributed.
	Volume (size)	Imagery – 100s of Terabytes Vector Data – 10s of GBs but billions of points
	Velocity (e.g. real time)	Some sensors delivery vector data in NRT. Visualization of changes should be NRT.
	Variety (multiple datasets, mashup)	Imagery (various formats NITF, GeoTiff, CADRG) Vector (various formats shape files, kml, text streams: Object types include points, lines, areas, polylines, circles, ellipses.
	Variability (rate of change)	Moderate to high
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues)**	Data accuracy is critical and is controlled generally by three factors: 1. Sensor accuracy is a big issue. 2. datum/spheroid. 3. Image registration accuracy
	Visualization	Displaying in a meaningful way large datasets (millions of points) on small devices (handhelds) at the end of low bandwidth networks.

Defense> Use Case 13: Large Scale Geospatial Analysis and Visualization

	Data Quality	The typical problem is visualization implying quality/accuracy not available in the original data. All data should include metadata for accuracy or circular error probability.
	Data Types	Imagery (various formats NITF, GeoTiff, CADRG) Vector (various formats shape files, kml, text streams: Object types include points, lines, areas, polylines, circles, ellipses.
	Data Analytics	Closest point of approach, deviation from route, point density over time, PCA and ICA
Big Data Specific Challenges (Gaps)	Indexing, retrieval and distributed analysis Visualization generation and transmission	
Big Data Specific Challenges in Mobility	Visualization of data at the end of low bandwidth wireless connections.	
Security and Privacy Requirements	Data is sensitive and must be completely secure in transit and at rest (particularly on handhelds)	
Highlight issues for generalizing this use case (e.g. for ref. architecture)	Geospatial data requires unique approaches to indexing and distributed analysis.	
More Information (URLs)	Applicable Standards: http://www.opengeospatial.org/standards http://geojson.org/ http://earth-info.nga.mil/publications/specs/printed/CADRG/cadrg.html Geospatial Indexing: Quad Trees, Space Filling Curves (Hilbert Curves) – You can google these for lots of references.	
Note: There has been some work with in DoD related to this problem set. Specifically, the DCGS-A standard cloud (DSC) stores, indexes, and analyzes some Big Data sources. However, many issues remain with visualization.		

Defense> Use Case 14: Object Identification and Tracking – Persistent Surveillance

Use Case Title	Object identification and tracking from Wide Area Large Format Imagery (WALF) Imagery or Full Motion Video (FMV) – Persistent Surveillance	
Vertical (area)	Defense (Intelligence)	
Author/Company/Email	David Boyd/Data Tactics/dboyd@data-tactics.com	
Actors/Stakeholders and their roles and responsibilities	1. Civilian Military decision makers 2. Intelligence Analysts 3. Warfighters	
Goals	To be able to process and extract/track entities (vehicles, people, packages) over time from the raw image data. Specifically, the idea is to reduce the petabytes of data generated by persistent surveillance down to a manageable size (e.g. vector tracks)	
Use Case Description	Persistent surveillance sensors can easily collect petabytes of imagery data in the space of a few hours. It is unfeasible for this data to be processed by humans for either alerting or tracking purposes. The data needs to be processed close to the sensor which is likely forward deployed since it is too large to be easily transmitted. The data should be reduced to a set of geospatial object (points, tracks, etc.) which can easily be integrated with other data to form a common operational picture.	
Current Solutions	**Compute(System)**	Various – they range from simple storage capabilities mounted on the sensor, to simple display and storage, to limited object extraction. Typical object extraction systems are currently small (1-20 node) GPU enhanced clusters.
	Storage	Currently flat files persisted on disk in most cases. Sometimes RDBMS indexes pointing to files or portions of files based on metadata/telemetry data.
	Networking	Sensor comms tend to be Line of Sight or Satellite based.
	Software	A wide range custom software and tools including traditional RDBMS and display tools.
Big Data Characteristics	**Data Source (distributed/centralized)**	Sensors include airframe mounted and fixed position optical, IR, and SAR images.
	Volume (size)	FMV – 30 to 60 frames per/sec at full color 1080P resolution. WALF – 1 to 10 frames per/sec at 10Kx10K full color resolution.
	Velocity (e.g. real time)	**Real Time**
	Variety (multiple datasets, mashup)	Data Typically exists in one or more standard imagery or video formats.
	Variability (rate of change)	**Little**
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues)**	The veracity of extracted objects is critical. If the system fails or generates false positives people are put at risk.
	Visualization	Visualization of extracted outputs will typically be as overlays on a geospatial display. Overlay objects should be links back to the originating image/video segment.
	Data Quality	Data quality is generally driven by a combination of sensor characteristics and weather (both obscuring factors - dust/moisture and stability factors – wind).

Defense > Use Case 14: Object Identification and Tracking – Persistent Surveillance

	Data Types	Standard imagery and video formats are input. Output should be in the form of OGC compliant web features or standard geospatial files (shape files, KML).
	Data Analytics	1. Object identification (type, size, color) and tracking. 2. Pattern analysis of object (did the truck observed every Weds. afternoon take a different route today or is there a standard route this person takes every day). 3. Crowd behavior/dynamics (is there a small group attempting to incite a riot. Is this person out of place in the crowd or behaving differently? 4. Economic activity a. is the line at the bread store, the butcher, or the ice cream store, b. are more trucks traveling north with goods than trucks going south c. Has activity at or the size of stores in this market place increased or decreased over the past year. 5. Fusion of data with other data to improve quality and confidence.
Big Data Specific Challenges (Gaps)	Processing the volume of data in NRT to support alerting and situational awareness.	
Big Data Specific Challenges in Mobility	Getting data from mobile sensor to processing	
Security and Privacy Requirements	Significant – sources and methods cannot be compromised the enemy should not be able to know what we see.	
Highlight issues for generalizing this use case (e.g. for ref. architecture)	Typically this type of processing fits well into massively parallel computing such as provided by GPUs. Typical problem is integration of this processing into a larger cluster capable of processing data from several sensors in parallel and in NRT. Transmission of data from sensor to system is also a large challenge.	
More Information (URLs)	Motion Imagery Standards - http://www.gwg.nga.mil/misb/ Some of many papers on object identity/tracking: http://www.dabi.temple.edu/~hbling/publication/SPIE12_Dismount_Formatted_v2_BW.pdf http://csce.uark.edu/~jgauch/library/Tracking/Orten.2005.pdf http://www.sciencedirect.com/science/article/pii/S0031320305004863 General Articles on the need: http://www.militaryaerospace.com/topics/m/video/79088650/persistent-surveillance-relies-on-extracting-relevant-data-points-and-connecting-the-dots.htm http://www.defencetalk.com/wide-area-persistent-surveillance-revolutionizes-tactical-isr-45745/ http://www.defencetalk.com/wide-area-persistent-surveillance-revolutionizes-tactical-isr-45745/	

Defense> Use Case 15: Intelligence Data Processing and Analysis

Use Case Title	Intelligence Data Processing and Analysis	
Vertical (area)	Defense (Intelligence)	
Author/ Company/Email	David Boyd/Data Tactics/dboyd@data-tactics.com	
Actors/Stakeholders and their roles and responsibilities	Senior Civilian/Military Leadership Field Commanders Intelligence Analysts Warfighters	
Goals	1. Provide automated alerts to Analysts, Warfighters, Commanders, and Leadership based on incoming intelligence data. 2. Allow Intelligence Analysts to identify in Intelligence data a. Relationships between entities (people, organizations, places, equipment) b. Trends in sentiment or intent for either general population or leadership group (state, non-state actors). c. Location of and possibly timing of hostile actions (including implantation of IEDs). d. Track the location and actions of (potentially) hostile actors 3. Ability to reason against and derive knowledge from diverse, disconnected, and frequently unstructured (e.g. text) data sources. 4. Ability to process data close to the point of collection and allow data to be shared easily to/from individual soldiers, forward deployed units, and senior leadership in garrison.	
Use Case Description	1. Ingest/accept data from a wide range of sensors and sources across intelligence disciplines (IMINT, MASINT, GEOINT, HUMINT, SIGINT, OSINT, etc.) 2. Process, transform, or align date from disparate sources in disparate formats into a unified data space to permit: a. Search b. Reasoning c. Comparison 3. Provide alerts to users of significant changes in the state of monitored entities or significant activity within an area. 4. Provide connectivity to the edge for the Warfighter (in this case the edge would go as far as a single soldier on dismounted patrol)	
Current Solutions	**Compute(System)**	Fixed and deployed computing clusters ranging from 1000s of nodes to 10s of nodes.
	Storage	10s of Terabytes to 100s of Petabytes for edge and fixed site clusters. Dismounted soldiers would have at most 1-100s of GBs (mostly single digit handheld data storage sizes).
	Networking	Networking with-in and between in garrison fixed sites is robust. Connectivity to forward edge is limited and often characterized by high latency and packet loss. Remote comms might be Satellite based (high latency) or even limited to RF Line of sight radio.

Defense> Use Case 15: Intelligence Data Processing and Analysis

	Software	Currently baseline leverages: 1. Hadoop 2. Accumulo (Big Table) 3. Solr 4. NLP (several variants) 5. Puppet (for deployment and security) 6. Storm 7. Custom applications and visualization tools
Big Data Characteristics	**Data Source (distributed/centralized)**	Very distributed
	Volume (size)	Some IMINT sensors can produce over a petabyte of data in the space of hours. Other data is as small as infrequent sensor activations or text messages.
	Velocity (e.g. real time)	Much sensor data is real time (Full motion video, SIGINT) other is less real time. The critical aspect is to be able ingest, process, and disseminate alerts in NRT.
	Variety (multiple datasets, mashup)	Everything from text files, raw media, imagery, video, audio, electronic data, human generated data.
	Variability (rate of change)	While sensor interface formats tend to be stable, most other data is uncontrolled and may be in any format. Much of the data is unstructured.
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues, semantics)**	Data provenance (e.g. tracking of all transfers and transformations) must be tracked over the life of the data. Determining the veracity of “soft” data sources (generally human generated) is a critical requirement.
	Visualization	Primary visualizations will be Geospatial overlays and network diagrams. Volume amounts might be millions of points on the map and thousands of nodes in the network diagram.
	Data Quality (syntax)	Data Quality for sensor generated data is generally known (image quality, sig/noise) and good. Unstructured or “captured” data quality varies significantly and frequently cannot be controlled.
	Data Types	Imagery, Video, Text, Digital documents of all types, Audio, Digital signal data.
	Data Analytics	1. NRT Alerts based on patterns and baseline changes. 2. Link Analysis 3. Geospatial Analysis 4. Text Analytics (sentiment, entity extraction, etc.)
Big Data Specific Challenges (Gaps)	1. Big (or even moderate size data) over tactical networks 2. Data currently exists in disparate silos which must be accessible through a semantically integrated data space. 3. Most critical data is either unstructured or imagery/video which requires significant processing to extract entities and information.	
Big Data Specific Challenges in Mobility	The outputs of this analysis and information must be transmitted to or accessed by the dismounted forward soldier.	

Defense> Use Case 15: Intelligence Data Processing and Analysis

Security and Privacy Requirements	Foremost. Data must be protected against: 1. Unauthorized access or disclosure 2. Tampering
Highlight issues for generalizing this use case (e.g. for ref. architecture)	Wide variety of data types, sources, structures, and quality which will span domains and requires integrated search and reasoning.
More Information (URLs)	http://www.afcea-aberdeen.org/files/presentations/AFCEAAberdeen_DCGSA_COLWells_PS.pdf http://stids.c4i.gmu.edu/papers/STIDSPapers/STIDS2012_T14_SmithEtAl_HorizontalIntegrationOfWarfighterIntel.pdf http://stids.c4i.gmu.edu/STIDS2011/papers/STIDS2011_CR_T1_SalmenEtAl.pdf http://www.youtube.com/watch?v=l4Qii7T8zeg http://dcgsa.apg.army.mil/

Healthcare and Life Sciences> Use Case 16: Electronic Medical Record Data

Use Case Title	Electronic Medical Record (EMR) Data	
Vertical (area)	Healthcare	
Author/Company/Email	Shaun Grannis/Indiana University/sgrannis@regenstrief.org	
Actors/Stakeholders and their roles and responsibilities	Biomedical informatics research scientists (implement and evaluate enhanced methods for seamlessly integrating, standardizing, analyzing, and operationalizing highly heterogeneous, high-volume clinical data streams); Health services researchers (leverage integrated and standardized EMR data to derive knowledge that supports implementation and evaluation of translational, comparative effectiveness, patient-centered outcomes research); Healthcare providers – physicians, nurses, public health officials (leverage information and knowledge derived from integrated and standardized EMR data to support direct patient care and population health)	
Goals	Use advanced methods for normalizing patient, provider, facility and clinical concept identification within and among separate health care organizations to enhance models for defining and extracting clinical phenotypes from non-standard discrete and free-text clinical data using feature selection, information retrieval and machine learning decision-models. Leverage clinical phenotype data to support cohort selection, clinical outcomes research, and clinical decision support.	
Use Case Description	As health care systems increasingly gather and consume EMR data, large national initiatives aiming to leverage such data are emerging, and 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. Despite the promise that increasingly prevalent and ubiquitous EMR data hold, enhanced methods for integrating and rationalizing these data are needed for a variety of reasons. Data from clinical systems evolve over time. This is because the concept space in healthcare is constantly evolving: new scientific discoveries lead to new disease entities, new diagnostic modalities, and new disease management approaches. These in turn lead to new clinical concepts, which drive the evolution of health concept ontologies. Using heterogeneous data from the Indiana Network for Patient Care (INPC), the nation's largest and longest-running health information exchange, which includes more than 4 billion discrete coded clinical observations from more than 100 hospitals for more than 12 million patients, we will use information retrieval techniques to identify highly relevant clinical features from electronic observational data. We will deploy information retrieval and natural language processing techniques to extract clinical features. Validated features will be used to parameterize clinical phenotype decision models based on maximum likelihood estimators and Bayesian networks. Using these decision models we will identify a variety of clinical phenotypes such as diabetes, congestive heart failure, and pancreatic cancer.	
Current Solutions	**Compute(System)**	Big Red II, a new Cray supercomputer at I.U.
	Storage	Teradata, PostgreSQL, MongoDB
	Networking	Various. Significant I/O intensive processing needed.
	Software	Hadoop, Hive, R. Unix-based.

Healthcare and Life Sciences> Use Case 16: Electronic Medical Record Data

Big Data Characteristics	**Data Source (distributed/centralized)**	Clinical data from more than 1,100 discrete logical, operational healthcare sources in the Indiana Network for Patient Care (INPC) the nation's largest and longest-running health information exchange.
	Volume (size)	More than 12 million patients, more than 4 billion discrete clinical observations. > 20 TB raw data.
	Velocity (e.g. real time)	Between 500,000 and 1.5 million new real-time clinical transactions added per day.
	Variety (multiple datasets, mashup)	We integrate a broad variety of clinical datasets from multiple sources: free text provider notes; inpatient, outpatient, laboratory, and emergency department encounters; chromosome and molecular pathology; chemistry studies; cardiology studies; hematology studies; microbiology studies; neurology studies; provider notes; referral labs; serology studies; surgical pathology and cytology, blood bank, and toxicology studies.
	Variability (rate of change)	Data from clinical systems evolve over time because the clinical and biological concept space is constantly evolving: new scientific discoveries lead to new disease entities, new diagnostic modalities, and new disease management approaches. These in turn lead to new clinical concepts, which drive the evolution of health concept ontologies, encoded in highly variable fashion.
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues, semantics)**	Data from each clinical source are commonly gathered using different methods and representations, yielding substantial heterogeneity. This leads to systematic errors and bias requiring robust methods for creating semantic interoperability.
	Visualization	Inbound data volume, accuracy, and completeness must be monitored on a routine basis using focus visualization methods. Intrinsic informational characteristics of data sources must be visualized to identify unexpected trends.
	Data Quality (syntax)	A central barrier to leveraging EMR data is the highly variable and unique local names and codes for the same clinical test or measurement performed at different institutions. When integrating many data sources, mapping local terms to a common standardized concept using a combination of probabilistic and heuristic classification methods is necessary.
	Data Types	Wide variety of clinical data types including numeric, structured numeric, free-text, structured text, discrete nominal, discrete ordinal, discrete structured, binary large blobs (images and video).

Healthcare and Life Sciences> Use Case 16: Electronic Medical Record Data

	Data Analytics	Information retrieval methods to identify relevant clinical features (tf-idf, latent semantic analysis, mutual information). Natural Language Processing techniques to 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.
Big Data Specific Challenges (Gaps)	Overcoming the systematic errors and bias in large-scale, heterogeneous clinical data to support decision-making in research, patient care, and administrative use-cases requires complex multistage processing and analytics that demands substantial computing power. Further, the optimal techniques for accurately and effectively deriving knowledge from observational clinical data are nascent.	
Big Data Specific Challenges in Mobility	Biological and clinical data are needed in a variety of contexts throughout the healthcare ecosystem. Effectively delivering clinical data and knowledge across the healthcare ecosystem will be facilitated by mobile platform such as mHealth.	
Security and Privacy Requirements	Privacy and confidentiality of individuals must be preserved in compliance with federal and state requirements including HIPAA. Developing analytic models using comprehensive, integrated clinical data requires aggregation and subsequent de-identification prior to applying complex analytics.	
Highlight issues for generalizing this use case (e.g. for ref. architecture)	Patients increasingly receive health care in a variety of clinical settings. The subsequent EMR data is fragmented and heterogeneous. In order to realize the promise of a Learning Health Care system as advocated by the National Academy of Science and the Institute of Medicine, EMR data must be rationalized and integrated. The methods we propose in this use-case support integrating and rationalizing clinical data to support decision-making at multiple levels.	
More Information (URLs)	Regenstrief Institute (http://www.regenstrief.org); Logical observation identifiers names and codes (http://www.loinc.org); Indiana Health Information Exchange (http://www.ihie.org); Institute of Medicine Learning Healthcare System (http://www.iom.edu/Activities/Quality/LearningHealthcare.aspx)	

Healthcare and Life Sciences> Use Case 17: Pathology Imaging/Digital Pathology

Use Case Title	Pathology Imaging/digital pathology	
Vertical (area)	Healthcare	
Author/Company/Email	Fusheng Wang/Emory University/fusheng.wang@emory.edu	
Actors/Stakeholders and their roles and responsibilities	Biomedical researchers on translational research; hospital clinicians on imaging guided diagnosis	
Goals	Develop high performance image analysis algorithms to extract spatial information from images; provide efficient spatial queries and analytics, and feature clustering and classification	
Use Case Description	Digital pathology imaging is an emerging field where examination of high resolution images of tissue specimens enables novel and more effective ways for disease diagnosis. Pathology image analysis segments massive (millions per image) spatial objects 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. Recently, 3D pathology imaging is made possible through 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.	
Current Solutions	**Compute(System)**	Supercomputers; Cloud
	Storage	SAN or HDFS
	Networking	Need excellent external network link
	Software	MPI for image analysis; Map/Reduce + Hive with spatial extension
Big Data Characteristics	**Data Source (distributed/centralized)**	Digitized pathology images from human tissues
	Volume (size)	1GB raw image data + 1.5GB analytical results per 2D image; 1TB raw image data + 1TB analytical results per 3D image. 1PB data per moderated hospital per year
	Velocity (e.g. real time)	Once generated, data will not be changed
	Variety (multiple datasets, mashup)	Image characteristics and analytics depend on disease types
	Variability (rate of change)	No change
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues)**	High quality results validated with human annotations are essential
	Visualization	Needed for validation and training
	Data Quality	Depend on preprocessing of tissue slides such as chemical staining and quality of image analysis algorithms
	Data Types	Raw images are whole slide images (mostly based on BIGTIFF), and analytical results are structured data (spatial boundaries and features)
	Data Analytics	Image analysis, spatial queries and analytics, feature clustering and classification

Healthcare and Life Sciences> Use Case 17: Pathology Imaging/Digital Pathology

Big Data Specific Challenges (Gaps)	Extreme large size; multi-dimensional; disease specific analytics; correlation with other data types (clinical data, -omic data)
Big Data Specific Challenges in Mobility	3D visualization of 3D pathology images is not likely in mobile platforms
Security and Privacy Requirements	Protected health information has to be protected; public data have to be de-identified
Highlight issues for generalizing this use case (e.g. for ref. architecture)	Imaging data; multi-dimensional spatial data analytics
More Information (URLs)	https://web.cci.emory.edu/confluence/display/PAIS https://web.cci.emory.edu/confluence/display/HadoopGIS

See Figure 2: Pathology Imaging/Digital Pathology – Examples of 2-D and 3-D pathology images.

See Figure 3: Pathology Imaging/Digital Pathology – Architecture of Hadoop-GIS, a spatial data warehousing system, over MapReduce to support spatial analytics for analytical pathology imaging.

Healthcare and Life Sciences> Use Case 18: Computational Bioimaging

Use Case Title	Computational Bioimaging	
Vertical (area)	Scientific Research: Biological Science	
Author/Company/Email	David Skinner[1], deskinner@lbl.gov Joaquin Correa[1], JoaquinCorrea@lbl.gov Daniela Ushizima[2], dushizima@lbl.gov Joerg Meyer[2], joergmeyer@lbl.gov [1]National Energy Scientific Computing Center (NERSC), Lawrence Berkeley National Laboratory, USA [2]Computational Research Division, Lawrence Berkeley National Laboratory, USA	
Actors/Stakeholders and their roles and responsibilities	Capability providers: Bioimaging instrument operators, microscope developers, imaging facilities, applied mathematicians, and data stewards. User Community: DOE, industry and academic researchers seeking to collaboratively build models from imaging data.	
Goals	Data delivered from bioimaging is increasingly automated, higher resolution, and multi-modal. This has created a data analysis bottleneck that, if resolved, can advance the biosciences discovery through Big Data techniques. Our goal is to solve that bottleneck with extreme scale computing. Meeting that goal will require more than computing. It will require building communities around data resources and providing advanced algorithms for massive image analysis. High-performance computational solutions can be harnessed by community-focused science gateways to guide the application of massive data analysis toward massive imaging datasets. Workflow components include data acquisition, storage, enhancement, minimizing noise, segmentation of regions of interest, crowd-based selection and extraction of features, and object classification, and organization, and search.	
Use Case Description	Web-based one-stop-shop for high performance, high throughput image processing for producers and consumers of models built on bio-imaging data.	
Current Solutions	**Compute(System)**	Hopper.nersc.gov (150K cores)
	Storage	Database and image collections
	Networking	10Gb, could use 100Gb and advanced networking (SDN)
	Software	ImageJ, OMERO, VolRover, advanced segmentation and feature detection methods from applied math researchers
Big Data Characteristics	**Data Source (distributed/centralized)**	Distributed experimental sources of bioimages (instruments). Scheduled high volume flows from automated high-resolution optical and electron microscopes.
	Volume (size)	Growing very fast. Scalable key-value and object store databases needed. In-database processing and analytics. 50TB here now, but currently over a petabyte overall. A single scan on emerging machines is 32TB
	Velocity (e.g. real time)	High throughput computing (HTC), responsive analysis
	Variety (multiple datasets, mashup)	Multi-modal imaging essentially must mash-up disparate channels of data with attention to registration and dataset formats.
	Variability (rate of change)	Biological samples are highly variable and their analysis workflows must cope with wide variation.

Healthcare and Life Sciences> Use Case 18: Computational Bioimaging

Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues, semantics)**	Data is messy overall as is training classifiers.
	Visualization	Heavy use of 3D structural models.
	Data Quality (syntax)	
	Data Types	Imaging file formats
	Data Analytics	Machine learning (SVM and RF) for classification and recommendation services.
Big Data Specific Challenges (Gaps)	HTC at scale for simulation science. Flexible data methods at scale for messy data. Machine learning and knowledge systems that drive pixel based data toward biological objects and models.	
Big Data Specific Challenges in Mobility		
Security and Privacy Requirements		
Highlight issues for generalizing this use case (e.g. for ref. architecture)	There is potential in generalizing concepts of search in the context of bioimaging.	
More Information (URLs)		

Healthcare and Life Sciences> Use Case 19: Genomic Measurements

Use Case Title	Genomic Measurements	
Vertical (area)	Healthcare	
Author/Company/Email	Justin Zook/NIST/jzook@nist.gov	
Actors/Stakeholders and their roles and responsibilities	NIST/Genome in a Bottle Consortium – public/private/academic partnership	
Goals	Develop well-characterized Reference Materials, Reference Data, and Reference Methods needed to assess performance of genome sequencing	
Use Case Description	Integrate data from multiple sequencing technologies and methods to develop highly confident characterization of whole human genomes as Reference Materials, and develop methods to use these Reference Materials to assess performance of any genome sequencing run	
Current Solutions	**Compute(System)**	72-core cluster for our NIST group, collaboration with >1000 core clusters at FDA, some groups are using cloud
	Storage	≈40TB NFS at NIST, PBs of genomics data at NIH/NCBI
	Networking	Varies. Significant I/O intensive processing needed
	Software	Open-source sequencing bioinformatics software from academic groups (UNIX-based)
Big Data Characteristics	**Data Source (distributed/centralized)**	Sequencers are distributed across many laboratories, though some core facilities exist.
	Volume (size)	40TB NFS is full, will need >100TB in 1-2 years at NIST; Healthcare community will need many PBs of storage
	Velocity (e.g. real time)	DNA sequencers can generate ≈300GB compressed data/day. Velocity has increased much faster than Moore's Law
	Variety (multiple datasets, mashup)	File formats not well-standardized, though some standards exist. Generally structured data.
	Variability (rate of change)	Sequencing technologies have evolved very rapidly, and new technologies are on the horizon.
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues)**	All sequencing technologies have significant systematic errors and biases, which require complex analysis methods and combining multiple technologies to understand, often with machine learning
	Visualization	"Genome browsers" have been developed to visualize processed data
	Data Quality	Sequencing technologies and bioinformatics methods have significant systematic errors and biases
	Data Types	Mainly structured text
	Data Analytics	Processing of raw data to produce variant calls. Also, clinical interpretation of variants, which is now very challenging.
Big Data Specific Challenges (Gaps)	Processing data requires significant computing power, which poses challenges especially to clinical laboratories as they are starting to perform large-scale sequencing. Long-term storage of clinical sequencing data could be expensive. Analysis methods are quickly evolving. Many parts of the genome are challenging to analyze, and systematic errors are difficult to characterize.	
Big Data Specific Challenges in Mobility	Physicians may need access to genomic data on mobile platforms	

Healthcare and Life Sciences> Use Case 19: Genomic Measurements

Security and Privacy Requirements	Sequencing data in health records or clinical research databases must be kept secure/private, though our Consortium data is public.
Highlight issues for generalizing this use case (e.g. for ref. architecture)	I have some generalizations to medical genome sequencing above, but focus on NIST/Genome in a Bottle Consortium work. Currently, labs doing sequencing range from small to very large. Future data could include other 'omics' measurements, which could be even larger than DNA sequencing
More Information (URLs)	Genome in a Bottle Consortium: http://www.genomeinabottle.org

Healthcare and Life Sciences> Use Case 20: Comparative Analysis for (meta) Genomes

Use Case Title	Comparative analysis for metagenomes and genomes	
Vertical (area)	Scientific Research: Genomics	
Author/Company/Email	Ernest Szeto / LBNL / eszeto@lbl.gov	
Actors/Stakeholders and their roles and responsibilities	Joint Genome Institute (JGI) Integrated Microbial Genomes (IMG) project. Heads: Victor M. Markowitz, and Nikos C. Kyrpides. User community: JGI, bioinformaticians and biologists worldwide.	
Goals	Provide an integrated comparative analysis system for metagenomes and genomes. This includes interactive Web UI with core data, backend precomputations, batch job computation submission from the UI.	
Use Case Description	Given a metagenomic sample, (1) determine the community composition in terms of other reference isolate genomes, (2) characterize the function of its genes, (3) begin to infer possible functional pathways, (4) characterize similarity or dissimilarity with other metagenomic samples, (5) begin to characterize changes in community composition and function due to changes in environmental pressures, (6) isolate sub-sections of data based on quality measures and community composition.	
Current Solutions	**Compute(System)**	Linux cluster, Oracle RDBMS server, large memory machines, standard Linux interactive hosts
	Storage	Oracle RDBMS, SQLite files, flat text files, Lucy (a version of Lucene) for keyword searches, BLAST databases, USEARCH databases
	Networking	Provided by NERSC
	Software	Standard bioinformatics tools (BLAST, HMMER, multiple alignment and phylogenetic tools, gene callers, sequence feature predictors…), Perl/Python wrapper scripts, Linux Cluster scheduling
Big Data Characteristics	**Data Source (distributed/centralized)**	Centralized.
	Volume (size)	50tb
	Velocity (e.g. real time)	Front end web UI must be real time interactive. Back end data loading processing must keep up with exponential growth of sequence data due to the rapid drop in cost of sequencing technology.
	Variety (multiple datasets, mashup)	Biological data is inherently heterogeneous, complex, structural, and hierarchical. One begins with sequences, followed by features on sequences, such as genes, motifs, regulatory regions, followed by organization of genes in neighborhoods (operons), to proteins and their structural features, to coordination and expression of genes in pathways. Besides core genomic data, new types of “Omics” data such as transcriptomics, methylomics, and proteomics describing gene expression under a variety of conditions must be incorporated into the comparative analysis system.
	Variability (rate of change)	The sizes of metagenomic samples can vary by several orders of magnitude, such as several hundred thousand genes to a billion genes (e.g., latter in a complex soil sample).

Healthcare and Life Sciences> Use Case 20: Comparative Analysis for (meta) Genomes

Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues)**	Metagenomic sampling science is currently preliminary and exploratory. Procedures for evaluating assembly of highly fragmented data in raw reads are better defined, but still an open research area.
	Visualization	Interactive speed of web UI on very large datasets is an ongoing challenge. Web UI's still seem to be the preferred interface for most biologists. It is use for basic querying and browsing of data. More specialized tools may be launched from them, e.g. for viewing multiple alignments. Ability to download large amounts of data for offline analysis is another requirement of the system.
	Data Quality	Improving quality of metagenomic assembly is still a fundamental challenge. Improving the quality of reference isolate genomes, both in terms of the coverage in the phylogenetic tree, improved gene calling and functional annotation is a more mature process, but an ongoing project.
	Data Types	Cf. above on "Variety"
	Data Analytics	Descriptive statistics, statistical significance in hypothesis testing, discovering new relationships, data clustering and classification is a standard part of the analytics. The less quantitative part includes the ability to visualize structural details at different levels of resolution. Data reduction, removing redundancies through clustering, more abstract representations such as representing a group of highly similar genomes in a pangenome are all strategies for both data management as well as analytics.
Big Data Specific Challenges (Gaps)	The biggest friend for dealing with the heterogeneity of biological data is still the RDBMS. Unfortunately, it does not scale for the current volume of data. NoSQL solutions aim at providing an alternative. Unfortunately, NoSQL solutions do not always lend themselves to real time interactive use, rapid and parallel bulk loading, and sometimes have issues regarding robustness. Our current approach is currently ad hoc, custom, relying mainly on the Linux cluster and the file system to supplement the Oracle RDBMS. The custom solution oftentimes rely in knowledge of the peculiarities of the data allowing us to devise horizontal partitioning schemes as well as inversion of data organization when applicable.	
Big Data Specific Challenges in Mobility	No special challenges. Just world wide web access.	
Security and Privacy Requirements	No special challenges. Data is either public or requires standard login with password.	
Highlight issues for generalizing this use case (e.g. for ref. architecture)	A replacement for the RDBMS in Big Data would be of benefit to everyone. Many NoSQL solutions attempt to fill this role, but have their limitations.	
More Information (URLs)	http://img.jgi.doe.gov	

Healthcare and Life Sciences> Use Case 21: Individualized Diabetes Management

Use Case Title	Individualized Diabetes Management	
Vertical (area)	Healthcare	
Author/Company/Email	Peter Li, Ying Ding, Philip Yu, Geoffrey Fox, David Wild at Mayo Clinic, Indiana University, UIC; dingying@indiana.edu	
Actors/Stakeholders and their roles and responsibilities	Mayo Clinic + IU/semantic integration of EHR data UIC/semantic graph mining of EHR data IU cloud and parallel computing	
Goals	Develop advanced graph-based data mining techniques applied to EHR to search for these cohorts and extract their EHR data for outcome evaluation. These methods will push the boundaries of scalability and data mining technologies and advance knowledge and practice in these areas as well as clinical management of complex diseases.	
Use Case Description	Diabetes is a growing illness in world population, affecting both developing and developed countries. Current management strategies do not adequately take into account of individual patient profiles, such as co-morbidities and medications, which are common in patients with chronic illnesses. We propose to approach this shortcoming by identifying similar patients from a large Electronic Health Record (EHR) database, i.e., an individualized cohort, and evaluate their respective management outcomes to formulate one best solution suited for a given patient with diabetes. Project under development as below **Stage 1**: Use the Semantic Linking for Property Values method to convert an existing data warehouse at Mayo Clinic, called the Enterprise Data Trust (EDT), into RDF triples that enables us to find similar patients much more efficiently through linking of both vocabulary-based and continuous values, **Stage 2**: Needs efficient parallel retrieval algorithms, suitable for cloud or HPC, using open source Hbase with both indexed and custom search to identify patients of possible interest. **Stage 3**: The EHR, as an RDF graph, provides a very rich environment for graph pattern mining. Needs new distributed graph mining algorithms to perform pattern analysis and graph indexing technique for pattern searching on RDF triple graphs. **Stage 4**: Given the size and complexity of graphs, mining subgraph patterns could generate numerous false positives and miss numerous false negatives. Needs robust statistical analysis tools to manage false discovery rate and determine true subgraph significance and validate these through several clinical use cases.	
Current Solutions	**Compute(System)**	supercomputers; cloud
	Storage	HDFS
	Networking	Varies. Significant I/O intensive processing needed
	Software	Mayo internal data warehouse called Enterprise Data Trust (EDT)
Big Data Characteristics	**Data Source (distributed/centralized)**	distributed EHR data
	Volume (size)	The Mayo Clinic EHR dataset is a very large dataset containing over 5 million patients with thousands of properties each and many more that are derived from primary values.
	Velocity (e.g. real time)	not real time but updated periodically

Healthcare and Life Sciences> Use Case 21: Individualized Diabetes Management

	Variety (multiple datasets, mashup)	Structured data, a patient has controlled vocabulary (CV) property values (demographics, diagnostic codes, medications, procedures, etc.) and continuous property values (lab tests, medication amounts, vitals, etc.). The number of property values could range from less than 100 (new patient) to more than 100,000 (long term patient) with typical patients composed of 100 CV values and 1000 continuous values. Most values are time based, i.e., a timestamp is recorded with the value at the time of observation.
	Variability (rate of change)	Data will be updated or added during each patient visit.
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues)**	Data are annotated based on domain ontologies or taxonomies. Semantics of data can vary from labs to labs.
	Visualization	no visualization
	Data Quality	Provenance is important to trace the origins of the data and data quality
	Data Types	text, and Continuous Numerical values
	Data Analytics	Integrating data into semantic graph, using graph traverse to replace SQL join. Developing semantic graph mining algorithms to identify graph patterns, index graph, and search graph. Indexed Hbase. Custom code to develop new patient properties from stored data.
Big Data Specific Challenges (Gaps)	For individualized cohort, we will effectively be building a datamart for each patient since the critical properties and indices will be specific to each patient. Due to the number of patients, this becomes an impractical approach. Fundamentally, the paradigm changes from relational row-column lookup to semantic graph traversal.	
Big Data Specific Challenges in Mobility	Physicians and patient may need access to this data on mobile platforms	
Security and Privacy Requirements	Health records or clinical research databases must be kept secure/private.	
Highlight issues for generalizing this use case (e.g. for ref. architecture)	Data integration: continuous values, ontological annotation, taxonomy Graph Search: indexing and searching graph Validation: Statistical validation	
More Information (URLs)		

Healthcare and Life Sciences> Use Case 22: Statistical Relational AI for Health Care

Use Case Title	Statistical Relational AI for Health Care	
Vertical (area)	Healthcare	
Author/Company/Email	Sriraam Natarajan / Indiana University /natarasr@indiana.edu	
Actors/Stakeholders and their roles and responsibilities	Researchers in Informatics, medicine and practitioners in medicine.	
Goals	The goal of the project is to analyze large, multi-modal, longitudinal data. Analyzing different data types such as imaging, EHR, genetic and natural language data requires a rich representation. This approach employs the 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 the information and reason about complex queries.	
Use Case Description	Users can provide a set of descriptions – say for instance, MRI images and demographic data about a particular subject. They can then query for the onset of a particular disease (say Alzheimer's) and the system will then provide a probability distribution over the possible occurrence of this disease.	
Current Solutions	**Compute(System)**	A high performance computer (48 GB RAM) is needed to run the code for a few hundred patients. Clusters for large datasets
	Storage	A 200 GB to 1 TB hard drive typically stores the test data. The relevant data is retrieved to main memory to run the algorithms. Backend data in database or NoSQL stores
	Networking	Intranet.
	Software	Mainly Java based, in house tools are used to process the data.
Big Data Characteristics	**Data Source (distributed/centralized)**	All the data about the users reside in a single disk file. Sometimes, resources such as published text need to be pulled from IInternet.
	Volume (size)	Variable due to the different amount of data collected. Typically can be in 100s of GBs for a single cohort of a few hundred people. When dealing with millions of patients, this can be in the order of 1 petabyte.
	Velocity (e.g. real time)	Varied. In some cases, EHRs are constantly being updated. In other controlled studies, the data often comes in batches in regular intervals.
	Variety (multiple datasets, mashup)	This is the key property in medical datasets. That data is typically in multiple tables and need to be merged in order to perform the analysis.
	Variability (rate of change)	The arrival of data is unpredictable in many cases as they arrive in real time.
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues, semantics)**	Challenging due to different modalities of the data, human errors in data collection and validation.
	Visualization	The visualization of the entire input data is nearly impossible. But typically, partially visualizable. The models built can be visualized under some reasonable assumptions.
	Data Quality (syntax)	

Healthcare and Life Sciences> Use Case 22: Statistical Relational AI for Health Care

	Data Types	EHRs, imaging, genetic data that are stored in multiple databases.
	Data Analytics	
Big Data Specific Challenges (Gaps)	Data is in abundance in many cases of medicine. The key issue is that there can possibly be too much data (as images, genetic sequences etc.) that can make the analysis complicated. The real challenge lies in aligning the data and merging from multiple sources in a form that can be made useful for a combined analysis. The other issue is that sometimes, large amount of data is available about a single subject but the number of subjects themselves 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. Hence, robust learning methods that can faithfully model the data are of paramount importance. Another aspect of data imbalance is the occurrence of positive examples (i.e., cases). The incidence of certain diseases may be rare making the ratio of cases to controls extremely skewed making it possible for the learning algorithms to model noise instead of examples.	
Big Data Specific Challenges in Mobility		
Security and Privacy Requirements	Secure handling and processing of data is of crucial importance in medical domains.	
Highlight issues for generalizing this use case (e.g. for ref. architecture)	Models learned from one set of populations cannot be easily generalized across other populations with diverse characteristics. This requires that the learned models can be generalized and refined according to the change in the population characteristics.	
More Information (URLs)		

Healthcare and Life Sciences> Use Case 23: World Population Scale Epidemiology

Use Case Title	World Population Scale Epidemiological Study	
Vertical (area)	Epidemiology, Simulation Social Science, Computational Social Science	
Author/Company/Email	Madhav Marathe Stephen Eubank or Chris Barrett/ Virginia Bioinformatics Institute, Virginia Tech, mmarathe@vbi.vt.edu, seubank@vbi.vt.edu or cbarrett@vbi.vt.edu	
Actors/Stakeholders and their roles and responsibilities	Government and non-profit institutions involved in health, public policy, and disaster mitigation. Social Scientist who wants to study the interplay between behavior and contagion.	
Goals	(a) Build a synthetic global population. (b) Run simulations over the global population to reason about outbreaks and various intervention strategies.	
Use Case Description	Prediction and control of pandemic similar to the 2009 H1N1 influenza.	
Current Solutions	**Compute(System)**	Distributed (MPI) based simulation system written in Charm++. Parallelism is achieved by exploiting the disease residence time period.
	Storage	Network file system. Exploring database driven techniques.
	Networking	Infiniband. High bandwidth 3D Torus.
	Software	Charm++, MPI
Big Data Characteristics	**Data Source (distributed/centralized)**	Generated from synthetic population generator. Currently centralized. However, could be made distributed as part of post-processing.
	Volume (size)	100TB
	Velocity (e.g. real time)	Interactions with experts and visualization routines generate large amount of real time data. Data feeding into the simulation is small but data generated by simulation is massive.
	Variety (multiple datasets, mashup)	Variety depends upon the complexity of the model over which the simulation is being performed. Can be very complex if other aspects of the world population such as type of activity, geographical, socio-economic, cultural variations are taken into account.
	Variability (rate of change)	Depends upon the evolution of the model and corresponding changes in the code. This is complex and time intensive. Hence low rate of change.
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues, semantics)**	Robustness of the simulation is dependent upon the quality of the model. However, robustness of the computation itself, although non-trivial, is tractable.
	Visualization	Would require very large amount of movement of data to enable visualization.
	Data Quality (syntax)	Consistent due to generation from a model
	Data Types	Primarily network data.
	Data Analytics	Summary of various runs and replicates of a simulation
Big Data Specific Challenges (Gaps)	Computation of the simulation is both compute intensive and data intensive. Moreover, due to unstructured and irregular nature of graph processing the problem is not easily decomposable. Therefore it is also bandwidth intensive. Hence, a supercomputer is applicable than cloud type clusters.	
Big Data Specific Challenges in Mobility	None	

Healthcare and Life Sciences> Use Case 23: World Population Scale Epidemiology

Security and Privacy Requirements	Several issues at the synthetic population-modeling phase (see social contagion model).
Highlight issues for generalizing this use case (e.g. for ref. architecture)	In general contagion diffusion of various kinds: information, diseases, social unrest can be modeled and computed. All of them are agent-based model that utilize the underlying interaction network to study the evolution of the desired phenomena.
More Information (URLs)	

Healthcare and Life Sciences> Use Case 24: Social Contagion Modeling

Use Case Title	Social Contagion Modeling	
Vertical (area)	Social behavior (including national security, public health, viral marketing, city planning, disaster preparedness)	
Author/Company/Email	Madhav Marathe or Chris Kuhlman /Virginia Bioinformatics Institute, Virginia Tech mmarathe@vbi.vt.edu or ckuhlman@vbi.vt.edu	
/Actors/Stakeholders and their roles and responsibilities		
Goals	Provide a computing infrastructure that models social contagion processes. The infrastructure enables different types of human-to-human interactions (e.g., face-to-face versus online media; mother-daughter relationships versus mother-coworker relationships) to be simulated. It takes not only human-to-human interactions into account, but also interactions among people, services (e.g., transportation), and infrastructure (e.g., Internet, electric power).	
Use Case Description	Social unrest. People take to the streets to voice unhappiness with government leadership. There are citizens that both support and oppose government. Quantify the degrees to which normal business and activities are disrupted owing to fear and anger. Quantify the possibility of peaceful demonstrations, violent protests. Quantify the potential for government responses ranging from appeasement, to allowing protests, to issuing threats against protestors, to actions to thwart protests. To address these issues, must have fine-resolution models and datasets.	
Current Solutions	**Compute(System)**	Distributed processing software running on commodity clusters and newer architectures and systems (e.g., clouds).
	Storage	File servers (including archives), databases.
	Networking	Ethernet, Infiniband, and similar.
	Software	Specialized simulators, open source software, and proprietary modeling environments. Databases.
Big Data Characteristics	**Data Source (distributed/centralized)**	Many data sources: populations, work locations, travel patterns, utilities (e.g., power grid) and other man-made infrastructures, online (social) media.
	Volume (size)	Easily 10s of TB per year of new data.
	Velocity (e.g. real time)	During social unrest events, human interactions and mobility key to understanding system dynamics. Rapid changes in data; e.g., who follows whom in Twitter.
	Variety (multiple datasets, mashup)	Variety of data seen in wide range of data sources. Temporal data. Data fusion. Data fusion a big issue. How to combine data from different sources and how to deal with missing or incomplete data? Multiple simultaneous contagion processes.
	Variability (rate of change)	Because of stochastic nature of events, multiple instances of models and inputs must be run to ranges in outcomes.
Big Data Science (collection, curation,	**Veracity (Robustness Issues, semantics)**	Failover of soft real-time analyses.

Healthcare and Life Sciences> Use Case 24: Social Contagion Modeling

analysis, action)	**Visualization**	Large datasets; time evolution; multiple contagion processes over multiple network representations. Levels of detail (e.g., individual, neighborhood, city, state, country-level).
	Data Quality (syntax)	Checks for ensuring data consistency, corruption. Preprocessing of raw data for use in models.
	Data Types	Wide-ranging data, from human characteristics to utilities and transportation systems, and interactions among them.
	Data Analytics	Models of behavior of humans and hard infrastructures, and their interactions. Visualization of results.
Big Data Specific Challenges (Gaps)	How to take into account heterogeneous features of 100s of millions or billions of individuals, models of cultural variations across countries that are assigned to individual agents? How to validate these large models? Different types of models (e.g., multiple contagions): disease, emotions, behaviors. Modeling of different urban infrastructure systems in which humans act. With multiple replicates required to assess stochasticity, large amounts of output data are produced; storage requirements.	
Big Data Specific Challenges in Mobility	How and where to perform these computations? Combinations of cloud computing and clusters. How to realize most efficient computations; move data to compute resources?	
Security and Privacy Requirements	Two dimensions. First, privacy and anonymity issues for individuals used in modeling (e.g., Twitter and Facebook users). Second, securing data and computing platforms for computation.	
Highlight issues for generalizing this use case (e.g. for ref. architecture)	Fusion of different data types. Different datasets must be combined depending on the particular problem. How to quickly develop, verify, and validate new models for new applications. What is appropriate level of granularity to capture phenomena of interest while generating results sufficiently quickly; i.e., how to achieve a scalable solution. Data visualization and extraction at different levels of granularity.	
More Information (URLs)		

Healthcare and Life Sciences> Use Case 25: LifeWatch Biodiversity

Use Case Title	LifeWatch – E-Science European Infrastructure for Biodiversity and Ecosystem Research	
Vertical (area)	Scientific Research: Life Science	
Author/Company/Email	Wouter Los, Yuri Demchenko (y.demchenko@uva.nl), University of Amsterdam	
Actors/Stakeholders and their roles and responsibilities	End-users (biologists, ecologists, field researchers) Data analysts, data archive managers, e-Science Infrastructure managers, EU states national representatives	
Goals	Research and monitor different ecosystems, biological species, their dynamics and migration.	
Use Case Description	LifeWatch project and initiative intends to provide integrated access to a variety of data, analytical and modeling tools as served by a variety of collaborating initiatives. Another service is offered with data and tools in selected workflows for specific scientific communities. In addition, LifeWatch will provide opportunities to construct personalized 'virtual labs', also allowing to enter new data and analytical tools. New data will be shared with the data facilities cooperating with LifeWatch. Particular case studies: Monitoring alien species, monitoring migrating birds, wetlands LifeWatch operates Global Biodiversity Information facility and Biodiversity Catalogue that is Biodiversity Science Web Services Catalogue	
Current Solutions	**Compute(System)**	Field facilities TBD Data center: General Grid and cloud based resources provided by national e-Science centers
	Storage	Distributed, historical and trends data archiving
	Networking	May require special dedicated or overlay sensor network.
	Software	Web Services based, Grid based services, relational databases
Big Data Characteristics	**Data Source (distributed/centralized)**	Ecological information from numerous observation and monitoring facilities and sensor network, satellite images/information, climate and weather, all recorded information. Information from field researchers
	Volume (size)	Involves many existing datasets/sources Collected amount of data TBD
	Velocity (e.g. real time)	Data analyzed incrementally, processes dynamics corresponds to dynamics of biological and ecological processes. However may require real-time processing and analysis in case of the natural or industrial disaster. May require data streaming processing.
	Variety (multiple datasets, mashup)	Variety and number of involved databases and observation data is currently limited by available tools; in principle, unlimited with the growing ability to process data for identifying ecological changes, factors/reasons, species evolution and trends. See below in additional information.
	Variability (rate of change)	Structure of the datasets and models may change depending on the data processing stage and tasks

Healthcare and Life Sciences> Use Case 25: LifeWatch Biodiversity

Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues)**	In normal monitoring mode are data are statistically processed to achieve robustness. Some biodiversity research is critical to data veracity (reliability/trustworthiness). In case of natural and technogenic disasters data veracity is critical.
	Visualization	Requires advanced and rich visualization, high definition visualization facilities, visualization data • 4D visualization • Visualizing effects of parameter change in (computational) models • Comparing model outcomes with actual observations (multi-dimensional)
	Data Quality	Depends on and ensued by initial observation data. Quality of analytical data depends on used mode and algorithms that are constantly improved. Repeating data analytics should be possible to re-evaluate initial observation data. Actionable data are human aided.
	Data Types	Multi-type. Relational data, key-value, complex semantically rich data
	Data Analytics	Parallel data streams and streaming analytics
Big Data Specific Challenges (Gaps)	Variety, multi-type data: SQL and no-SQL, distributed multi-source data. Visualization, distributed sensor networks. Data storage and archiving, data exchange and integration; data linkage: from the initial observation data to processed data and reported/visualized data. • Historical unique data • Curated (authorized) reference data (i.e., species names lists), algorithms, software code, workflows • Processed (secondary) data serving as input for other researchers • Provenance (and persistent identification (PID)) control of data, algorithms, and workflows	
Big Data Specific Challenges in Mobility	Require supporting mobile sensors (e.g. birds migration) and mobile researchers (both for information feed and catalogue search) • Instrumented field vehicles, Ships, Planes, Submarines, floating buoys, sensor tagging on organisms • Photos, video, sound recording	
Security and Privacy Requirements	Data integrity, referral integrity of the datasets. Federated identity management for mobile researchers and mobile sensors Confidentiality, access control and accounting for information on protected species, ecological information, space images, climate information.	
Highlight issues for generalizing this use case (e.g. for ref. architecture)	• Support of distributed sensor network • Multi-type data combination and linkage; potentially unlimited data variety • Data life cycle management: data provenance, referral integrity and identification • Access and integration of multiple distributed databases	
More Information (URLs)	http://www.lifewatch.eu/web/guest/home https://www.biodiversitycatalogue.org/	

Healthcare and Life Sciences> Use Case 25: LifeWatch Biodiversity

Note:
Variety of data used in Biodiversity research
Genetic (genomic) diversity

- DNA sequences and barcodes
- Metabolomics functions

Species information

- species names
- occurrence data (in time and place)
- species traits and life history data
- host-parasite relations
- collection specimen data

Ecological information

- biomass, trunk/root diameter and other physical characteristics
- population density etc.
- habitat structures
- C/N/P etc. molecular cycles

Ecosystem data

- species composition and community dynamics
- remote and earth observation data
- CO_2 fluxes
- Soil characteristics
- Algal blooming
- Marine temperature, salinity, pH, currents, etc.

Ecosystem services

- productivity (i.e.., biomass production/time)
- fresh water dynamics
- erosion
- climate buffering
- genetic pools

Data concepts

- conceptual framework of each data
- ontologies
- provenance data

Algorithms and workflows

- software code and provenance
- tested workflows

Multiple sources of data and information

- Specimen collection data
- Observations (human interpretations)
- Sensors and sensor networks (terrestrial, marine, soil organisms), bird etc. tagging
- Aerial and satellite observation spectra
- Field * Laboratory experimentation
- Radar and LiDAR
- Fisheries and agricultural data
- Deceases and epidemics

Deep Learning and Social Media> Use Case 26: Large-scale Deep Learning

Use Case Title	Large-scale Deep Learning	
Vertical (area)	Machine Learning/AI	
Author/Company/Email	Adam Coates / Stanford University / acoates@cs.stanford.edu	
Actors/Stakeholders and their roles and responsibilities	Machine learning researchers and practitioners faced with large quantities of data and complex prediction tasks. Supports state-of-the-art development in computer vision as in automatic car driving, speech recognition, and natural language processing in both academic and industry systems.	
Goals	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.	
Use Case Description	A research scientist or machine learning practitioner wants to train a deep neural network from a large (>>1TB) corpus of data (typically imagery, video, audio, or text). Such training procedures often require customization of the neural network architecture, learning criteria, and dataset preprocessing. In addition to the computational expense demanded by the learning algorithms, the need for rapid prototyping and ease of development is extremely high.	
Current Solutions	**Compute(System)**	GPU cluster with high-speed interconnects (e.g., Infiniband, 40gE)
	Storage	100TB Lustre filesystem
	Networking	Infiniband within HPC cluster; 1G ethernet to outside infrastructure (e.g., Web, Lustre).
	Software	In-house GPU kernels and MPI-based communication developed by Stanford CS. C++/Python source.
Big Data Characteristics	**Data Source (distributed/centralized)**	Centralized filesystem with a single large training dataset. Dataset may be updated with new training examples as they become available.
	Volume (size)	Current datasets typically 1 TB to 10 TB. With increases in computation that enable much larger models, datasets of 100TB or more may be necessary in order to exploit the representational power of the larger models. Training a self-driving car could take 100 million images.
	Velocity (e.g. real time)	Much faster than real-time processing is required. Current computer vision applications involve processing hundreds of image frames per second in order to ensure reasonable training times. For demanding applications (e.g., autonomous driving) we envision the need to process many thousand high-resolution (6 megapixels or more) images per second.
	Variety (multiple datasets, mashup)	Individual applications may involve a wide variety of data. Current research involves neural networks that actively learn from heterogeneous tasks (e.g., learning to perform tagging, chunking and parsing for text, or learning to read lips from combinations of video and audio).
	Variability (rate of change)	Low variability. Most data is streamed in at a consistent pace from a shared source. Due to high computational requirements, server loads can introduce burstiness into data transfers.

Deep Learning and Social Media> Use Case 26: Large-scale Deep Learning

Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues, semantics)**	Datasets for ML applications are often hand-labeled and verified. Extremely large datasets involve crowd-sourced labeling and invite ambiguous situations where a label is not clear. Automated labeling systems still require human sanity-checks. Clever techniques for large dataset construction is an active area of research.
	Visualization	Visualization of learned networks is an open area of research, though partly as a debugging technique. Some visual applications involve visualization predictions on test imagery.
	Data Quality (syntax)	Some collected data (e.g., compressed video or audio) may involve unknown formats, codecs, or may be corrupted. Automatic filtering of original source data removes these.
	Data Types	Images, video, audio, text. (In practice: almost anything.)
	Data Analytics	Small degree of batch statistical preprocessing; all other data analysis is performed by the learning algorithm itself.
Big Data Specific Challenges (Gaps)	Processing requirements for even modest quantities of data are extreme. Though the trained representations can make use of many terabytes of data, the primary challenge is in processing all of the data during training. Current state-of-the-art deep learning systems are capable of using neural networks with more than 10 billion free parameters (akin to synapses in the brain), and necessitate trillions of floating point operations per training example. Distributing these computations over high-performance infrastructure is a major challenge for which we currently use a largely custom software system.	
Big Data Specific Challenges in Mobility	After training of large neural networks is completed, the learned network may be copied to other devices with dramatically lower computational capabilities for use in making predictions in real time. (E.g., in autonomous driving, the training procedure is performed using a HPC cluster with 64 GPUs. The result of training, however, is a neural network that encodes the necessary knowledge for making decisions about steering and obstacle avoidance. This network can be copied to embedded hardware in vehicles or sensors.)	
Security and Privacy Requirements	None.	

Deep Learning and Social Media> Use Case 26: Large-scale Deep Learning

Highlight issues for generalizing this use case (e.g. for ref. architecture)	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. Most deep learning systems require a substantial degree of tuning on the target application for best performance and thus necessitate a large number of experiments with designer intervention in between. As a result, minimizing the turn-around time of experiments and accelerating development is crucial. These two requirements (high throughput and high productivity) are dramatically in contention. HPC systems are available to accelerate experiments, but current HPC software infrastructure is difficult to use which lengthens development and debugging time and, in many cases, makes otherwise computationally tractable applications infeasible. The major components needed for these applications (which are currently in-house custom software) involve dense linear algebra on distributed-memory HPC systems. While libraries for single-machine or single-GPU computation are available (e.g., BLAS, CuBLAS, MAGMA, etc.), distributed computation of dense BLAS-like or LAPACK-like operations on GPUs remains poorly developed. Existing solutions (e.g., ScaLapack for CPUs) are not well-integrated with higher level languages and require low-level programming which lengthens experiment and development time.
More Information (URLs)	Recent popular press coverage of deep learning technology: http://www.nytimes.com/2012/11/24/science/scientists-see-advances-in-deep-learning-a-part-of-artificial-intelligence.html http://www.nytimes.com/2012/06/26/technology/in-a-big-network-of-computers-evidence-of-machine-learning.html http://www.wired.com/wiredenterprise/2013/06/andrew_ng/ A recent research paper on HPC for Deep Learning: http://www.stanford.edu/~acoates/papers/CoatesHuvalWangWuNgCatanzaro_icml2013.pdf Widely-used tutorials and references for Deep Learning: http://ufldl.stanford.edu/wiki/index.php/Main_Page http://deeplearning.net/

Deep Learning and Social Media> Use Case 27: Large Scale Consumer Photos Organization

Use Case Title	Organizing large-scale, unstructured collections of consumer photos	
Vertical (area)	(Scientific Research: Artificial Intelligence)	
Author/Company/Email	David Crandall, Indiana University, djcran@indiana.edu	
Actors/Stakeholders and their roles and responsibilities	Computer vision researchers (to push forward state of art), media and social network companies (to help organize large-scale photo collections), consumers (browsing both personal and public photo collections), researchers and others interested in producing cheap 3d models (archaeologists, architects, urban planners, interior designers…)	
Goals	Produce 3d reconstructions of scenes using collections of millions to billions of consumer images, where neither the scene structure nor the camera positions are known a priori. Use resulting 3d models to allow efficient and effective browsing of large-scale photo collections by geographic position. Geolocate new images by matching to 3d models. Perform object recognition on each image.	
Use Case Description	3d reconstruction is typically posed as a robust non-linear least squares optimization problem in which observed (noisy) correspondences between images are constraints and unknowns are 6-d camera pose of each image and 3-d position of each point in the scene. Sparsity and large degree of noise in constraints typically makes naïve techniques fall into local minima that are not close to actual scene structure. Typical specific steps are: (1) extracting features from images, (2) matching images to find pairs with common scene structures, (3) estimating an initial solution that is close to scene structure and/or camera parameters, (4) optimizing non-linear objective function directly. Of these, (1) is embarrassingly parallel. (2) is an all-pairs matching problem, usually with heuristics to reject unlikely matches early on. We solve (3) using discrete optimization using probabilistic inference on a graph (Markov Random Field) followed by robust Levenberg-Marquardt in continuous space. Others solve (3) by solving (4) for a small number of images and then incrementally adding new images, using output of last round as initialization for next round. (4) is typically solved with Bundle Adjustment, which is a non-linear least squares solver that is optimized for the particular constraint structure that occurs in 3d reconstruction problems. Image recognition problems are typically embarrassingly parallel, although learning object models involves learning a classifier (e.g. a Support Vector Machine), a process that is often hard to parallelize.	
Current Solutions	**Compute(System)**	Hadoop cluster (about 60 nodes, 480 core)
	Storage	Hadoop DFS and flat files
	Networking	Simple Unix
	Software	Hadoop Map-reduce, simple hand-written multithreaded tools (ssh and sockets for communication)
Big Data Characteristics	**Data Source (distributed/centralized)**	Publicly-available photo collections, e.g. on Flickr, Panoramio, etc.
	Volume (size)	500+ billion photos on Facebook, 5+ billion photos on Flickr.
	Velocity (e.g. real time)	100+ million new photos added to Facebook per day.
	Variety (multiple datasets, mashup)	Images and metadata including EXIF tags (focal distance, camera type, etc.),

Deep Learning and Social Media> Use Case 27: Large Scale Consumer Photos Organization

	Variability (rate of change)	Rate of photos varies significantly, e.g. roughly 10x photos to Facebook on New Year's versus other days. Geographic distribution of photos follows long-tailed distribution, with 1000 landmarks (totaling only about 100 square km) accounting for over 20% of photos on Flickr.
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues)**	Important to make as accurate as possible, subject to limitations of computer vision technology.
	Visualization	Visualize large-scale 3-d reconstructions, and navigate large-scale collections of images that have been aligned to maps.
	Data Quality	Features observed in images are quite noisy due both to imperfect feature extraction and to non-ideal properties of specific images (lens distortions, sensor noise, image effects added by user, etc.)
	Data Types	Images, metadata
	Data Analytics	
Big Data Specific Challenges (Gaps)	Analytics needs continued monitoring and improvement.	
Big Data Specific Challenges in Mobility	Many/most images are captured by mobile devices; eventual goal is to push reconstruction and organization to phone to allow real-time interaction with the user.	
Security and Privacy Requirements	Need to preserve privacy for users and digital rights for media.	
Highlight issues for generalizing this use case (e.g. for ref. architecture)	Components of this use case including feature extraction, feature matching, and large-scale probabilistic inference appear in many or most computer vision and image processing problems, including recognition, stereo resolution, image denoising, etc.	
More Information (URLs)	http://vision.soic.indiana.edu/disco	

Deep Learning and Social Media> Use Case 28: Truthy Twitter Data Analysis

Use Case Title	Truthy: Information diffusion research from Twitter Data	
Vertical (area)	Scientific Research: Complex Networks and Systems research	
Author/Company/Email	Filippo Menczer, Indiana University, fil@indiana.edu; Alessandro Flammini, Indiana University, aflammin@indiana.edu; Emilio Ferrara, Indiana University, ferrarae@indiana.edu;	
Actors/Stakeholders and their roles and responsibilities	Research funded by NFS, DARPA, and McDonnel Foundation.	
Goals	Understanding how communication spreads on socio-technical networks. Detecting potentially harmful information spread at the early stage (e.g., deceiving messages, orchestrated campaigns, untrustworthy information, etc.)	
Use Case Description	(1) Acquisition and storage of a large volume of continuous streaming data from Twitter (≈100 million messages per day, ≈500GB data/day increasing over time); (2) near real-time analysis of such data, for anomaly detection, stream clustering, signal classification and online-learning; (3) data retrieval, Big Data visualization, data-interactive Web interfaces, public API for data querying.	
Current Solutions	**Compute(System)**	Current: in-house cluster hosted by Indiana University. Critical requirement: large cluster for data storage, manipulation, querying and analysis.
	Storage	Current: Raw data stored in large compressed flat files, since August 2010. Need to move towards Hadoop/IndexedHBase and HDFS distributed storage. Redis as an in-memory database as a buffer for real-time analysis.
	Networking	10GB/Infiniband required.
	Software	Hadoop, Hive, Redis for data management. Python/SciPy/NumPy/MPI for data analysis.
Big Data Characteristics	**Data Source (distributed/centralized)**	Distributed – with replication/redundancy
	Volume (size)	≈30TB/year compressed data
	Velocity (e.g. real time)	Near real-time data storage, querying and analysis
	Variety (multiple datasets, mashup)	Data schema provided by social media data source. Currently using Twitter only. We plan to expand incorporating Google+, Facebook
	Variability (rate of change)	Continuous real-time data stream incoming from each source.
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues, semantics)**	99.99% uptime required for real-time data acquisition. Service outages might corrupt data integrity and significance.
	Visualization	Information diffusion, clustering, and dynamic network visualization capabilities already exist.
	Data Quality (syntax)	Data structured in standardized formats, the overall quality is extremely high. We generate aggregated statistics; expand the features set, etc., generating high-quality derived data.
	Data Types	Fully-structured data (JSON format) enriched with users' meta-data, geo-locations, etc.

Deep Learning and Social Media> Use Case 28: Truthy Twitter Data Analysis

	Data Analytics	**Stream clustering**: data are aggregated according to topics, meta-data and additional features, using ad hoc online clustering algorithms. **Classification**: using multi-dimensional time series to generate, network features, users, geographical, content features, etc., we classify information produced on the platform. **Anomaly detection**: real-time identification of anomalous events (e.g., induced by exogenous factors). **Online learning**: applying machine learning/deep learning methods to real-time information diffusion patterns analysis, users profiling, etc.
Big Data Specific Challenges (Gaps)	Dealing with real-time analysis of large volume of data. Providing a scalable infrastructure to allocate resources, storage space, etc. on-demand if required by increasing data volume over time.	
Big Data Specific Challenges in Mobility	Implementing low-level data storage infrastructure features to guarantee efficient, mobile access to data.	
Security and Privacy Requirements	Twitter publicly releases data collected by our platform. Although, data-sources incorporate user meta-data (in general, not sufficient to uniquely identify individuals) therefore some policy for data storage security and privacy protection must be implemented.	
Highlight issues for generalizing this use case (e.g. for ref. architecture)	Definition of high-level data schema to incorporate multiple data-sources providing similarly structured data.	
More Information (URLs)	http://truthy.indiana.edu/ http://cnets.indiana.edu/groups/nan/truthy http://cnets.indiana.edu/groups/nan/despic	

Deep Learning and Social Media> Use Case 29: Crowd Sourcing in the Humanities

Use Case Title	Crowd Sourcing in the Humanities as Source for Big and Dynamic Data	
Vertical (area)	Humanities, Social Sciences	
Author/Company/Email	Sebastian Drude <Sebastian.Drude@mpi.nl>, Max Planck Institute for Psycholinguistics	
Actors/Stakeholders and their roles and responsibilities	Scientists (Sociologists, Psychologists, Linguists, Politic Scientists, Historians, etc.), data managers and analysts, data archives The general public as data providers and participants	
Goals	Capture information (manually entered, recorded multimedia, reaction times, pictures, sensor information) from many individuals and their devices. Thus capture wide ranging individual, social, cultural and linguistic variation among several dimensions (space, social space, time).	
Use Case Description	Many different possible use cases: get recordings of language usage (words, sentences, meaning descriptions, etc.), answers to surveys, info on cultural facts, transcriptions of pictures and texts -- correlate these with other phenomena, detect new cultural practices, behavior, values and believes, discover individual variation	
Current Solutions	**Compute(System)**	Individual systems for manual data collection (mostly Websites)
	Storage	Traditional servers
	Networking	barely used other than for data entry via web
	Software	XML technology, traditional relational databases for storing pictures, not much multi-media yet.
Big Data Characteristics	**Data Source (distributed/centralized)**	Distributed, individual contributors via webpages and mobile devices
	Volume (size)	Depends dramatically, from hundreds to millions of data records. Depending on data-type: from GBs (text, surveys, experiment values) to hundreds of terabytes (multimedia)
	Velocity (e.g. real time)	Depends very much on project: dozens to thousands of new data records per day Data has to be analyzed incrementally.
	Variety (multiple datasets, mashup)	so far mostly homogeneous small datasets; expected large distributed heterogeneous datasets which have to be archived as primary data
	Variability (rate of change)	Data structure and content of collections are changing during data life cycle. There is no critical variation of data producing speed, or runtime characteristics variations.
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues)**	Noisy data is possible, unreliable metadata, identification and pre-selection of appropriate data
	Visualization	important for interpretation, no special visualization techniques
	Data Quality	validation is necessary; quality of recordings, quality of content, spam
	Data Types	individual data records (survey answers, reaction times); text (e.g., comments, transcriptions,…); multi-media (pictures, audio, video)

Deep Learning and Social Media> Use Case 29: Crowd Sourcing in the Humanities

	Data Analytics	pattern recognition of all kind (e.g., speech recognition, automatic A&V analysis, cultural patterns), identification of structures (lexical units, linguistic rules, etc.)
Big Data Specific Challenges (Gaps)	Data management (metadata, provenance info, data identification with PIDs) Data curation Digitizing existing audio-video, photo and documents archives	
Big Data Specific Challenges in Mobility	Include data from sensors of mobile devices (position, etc.); Data collection from expeditions and field research.	
Security and Privacy Requirements	Privacy issues may be involved (A/V from individuals), anonymization may be necessary but not always possible (A/V analysis, small speech communities) Archive and metadata integrity, long term preservation	
Highlight issues for generalizing this use case (e.g. for ref. architecture)	Many individual data entries from many individuals, constant flux of data entry, metadata assignment, etc. Offline vs. online use, to be synchronized later with central database. Giving significant feedback to contributors.	
More Information (URLs)	---	
Note: Crowd sourcing has been barely started to be used on a larger scale. With the availability of mobile devices, now there is a huge potential for collecting much data from many individuals, also making use of sensors in mobile devices. This has not been explored on a large scale so far; existing projects of crowd sourcing are usually of a limited scale and web-based.		

Deep Learning and Social Media> Use Case 30: CINET Network Science Cyberinfrastructure

Use Case Title	CINET: Cyberinfrastructure for Network (Graph) Science and Analytics	
Vertical (area)	Network Science	
Author/Company/Email	Team lead by Virginia Tech and comprising of researchers from Indiana University, University at Albany, North Carolina AT, Jackson State University, University at Houston Downtown, Argonne National Laboratory Point of Contact: Madhav Marathe or Keith Bisset, Network Dynamics and Simulation Science Laboratory, Virginia Bio-informatics Institute Virginia Tech, mmarathe@vbi.vt.edu / kbisset@vbi.vt.edu	
Actors/Stakeholders and their roles and responsibilities	Researchers, practitioners, educators and students interested in the study of networks.	
Goals	CINET cyberinfrastructure middleware to support network science. This middleware will give researchers, practitioners, teachers and students access to a computational and analytic environment for research, education and training. The user interface provides lists of available networks and network analysis modules (implemented algorithms for network analysis). A user, who can be a researcher in network science area, can select one or more networks and analysis them with the available network analysis tools and modules. A user can also generate random networks following various random graph models. Teachers and students can use CINET for classroom use to demonstrate various graph theoretic properties and behaviors of various algorithms. A user is also able to add a network or network analysis module to the system. This feature of CINET allows it to grow easily and remain up-to-date with the latest algorithms. The goal is to provide a common web-based platform for accessing various (i) network and graph analysis tools such as SNAP, NetworkX, Galib, etc. (ii) real-world and synthetic networks, (iii) computing resources and (iv) data management systems to the end-user in a seamless manner.	
Use Case Description	Users can run one or more structural or dynamic analysis on a set of selected networks. The domain specific language allows users to develop flexible high level workflows to define more complex network analysis.	
Current Solutions	**Compute(System)**	A high performance computing cluster (DELL C6100), named Shadowfax, of 60 compute nodes and 12 processors (Intel Xeon X5670 2.93GHz) per compute node with a total of 720 processors and 4GB main memory per processor. Shared memory systems ; EC2 based clouds are also used Some of the codes and networks can utilize single node systems and thus are being currently mapped to Open Science Grid
	Storage	628 TB GPFS
	Networking	Internet, infiniband. A loose collection of supercomputing resources.
	Software	Graph libraries: Galib, NetworkX. Distributed Workflow Management: Simfrastructure, databases, semantic web tools

Deep Learning and Social Media> Use Case 30: CINET Network Science Cyberinfrastructure

Big Data Characteristics	**Data Source (distributed/centralized)**	A single network remains in a single disk file accessible by multiple processors. However, during the execution of a parallel algorithm, the network can be partitioned and the partitions are loaded in the main memory of multiple processors.
	Volume (size)	Can be hundreds of GB for a single network.
	Velocity (e.g. real time)	Two types of changes: (i) the networks are very dynamic and (ii) as the repository grows, we expect at least a rapid growth to lead to over 1000-5000 networks and methods in about a year
	Variety (multiple datasets, mashup)	Datasets are varied: (i) directed as well as undirected networks, (ii) static and dynamic networks, (iii) labeled, (iv) can have dynamics over these networks,
	Variability (rate of change)	The rate of graph-based data is growing at increasing rate. Moreover, increasingly other life sciences domains are using graph-based techniques to address problems. Hence, we expect the data and the computation to grow at a significant pace.
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues, semantics)**	Challenging due to asynchronous distributed computation. Current systems are designed for real-time synchronous response.
	Visualization	As the input graph size grows the visualization system on client side is stressed heavily both in terms of data and compute.
	Data Quality (syntax)	
	Data Types	
	Data Analytics	
Big Data Specific Challenges (Gaps)	Parallel algorithms are necessary to analyze massive networks. Unlike many structured data, network data is difficult to partition. The main difficulty in partitioning a network is that different algorithms require different partitioning schemes for efficient operation. Moreover, most of the network measures are global in nature and require either i) huge duplicate data in the partitions or ii) very large communication overhead resulted from the required movement of data. These issues become significant challenges for big networks. Computing dynamics over networks is harder since the network structure often interacts with the dynamical process being studied. CINET enables large class of operations across wide variety, both in terms of structure and size, of graphs. Unlike other compute + data intensive systems, such as parallel databases or CFD, performance on graph computation is sensitive to underlying architecture. Hence, a unique challenge in CINET is manage the mapping between workload (graph type + operation) to a machine whose architecture and runtime is conducive to the system. Data manipulation and bookkeeping of the derived for users is another big challenge since unlike enterprise data there is no well-defined and effective models and tools for management of various graph data in a unified fashion.	
Big Data Specific Challenges in Mobility		
Security and Privacy Requirements		

Deep Learning and Social Media> Use Case 30: CINET Network Science Cyberinfrastructure

Highlight issues for generalizing this use case (e.g. for ref. architecture)	HPC as a service. As data volume grows increasingly large number of applications such as biological sciences need to use HPC systems. CINET can be used to deliver the compute resource necessary for such domains.
More Information (URLs)	http://cinet.vbi.vt.edu/cinet_new/

Deep Learning and Social Media> Use Case 31: NIST Analytic Technology Measurement and Evaluations

Use Case Title	NIST Information Access Division analytic technology performance measurement, evaluations, and standards	
Vertical (area)	Analytic technology performance measurement and standards for government, industry, and academic stakeholders	
Author/Company/Email	John Garofolo (john.garofolo@nist.gov)	
Actors/Stakeholders and their roles and responsibilities	NIST developers of measurement methods, data contributors, analytic algorithm developers, users of analytic technologies for unstructured, semi-structured data, and heterogeneous data across all sectors.	
Goals	Accelerate the development of advanced analytic technologies for unstructured, semi-structured, and heterogeneous data through performance measurement and standards. Focus communities of interest on analytic technology challenges of importance, create consensus-driven measurement metrics and methods for performance evaluation, evaluate the performance of the performance metrics and methods via community-wide evaluations which foster knowledge exchange and accelerate progress, and build consensus towards widely-accepted standards for performance measurement.	
Use Case Description	Develop performance metrics, measurement methods, and community evaluations to ground and accelerate the 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 employ one of two processing models: 1) Push test data out to test participants and analyze the output of participant systems, 2) Push algorithm test harness interfaces out to participants and bring in their algorithms and test them on internal computing clusters. Developing approaches to support scalable Cloud-based developmental testing. Also perform usability and utility testing on systems with users in the loop.	
Current Solutions	**Compute (System)**	Linux and OS-10 clusters; distributed computing with stakeholder collaborations; specialized image processing architectures.
	Storage	RAID arrays, and distribute data on 1-2TB drives, and occasionally FTP. Distributed data distribution with stakeholder collaborations.
	Networking	Fiber channel disk storage, Gigabit Ethernet for system-system communication, general intra- and Internet resources within NIST and shared networking resources with its stakeholders.
	Software	PERL, Python, C/C++, Matlab, R development tools. Create ground-up test and measurement applications.
Big Data Characteristics	**Data Source (distributed/centralized)**	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.
	Volume (size)	The test corpora exceed 900M Web pages occupying 30 TB of storage, 100M tweets, 100M ground-truthed biometric images, several hundred thousand partially ground-truthed video clips, and terabytes of smaller fully ground-truthed test collections. Even larger data collections are being planned for future evaluations of

Deep Learning and Social Media> Use Case 31: NIST Analytic Technology Measurement and Evaluations

		analytics involving multiple data streams and very heterogeneous data.
	Velocity (e.g. real time)	Most legacy evaluations are focused on retrospective analytics. Newer evaluations are focusing on simulations of real-time analytic challenges from multiple data streams.
	Variety (multiple datasets, mashup)	The test collections span a wide variety of analytic application types including textual search/extraction, machine translation, speech recognition, image and voice biometrics, object and person recognition and tracking, document analysis, human-computer dialogue, and multimedia search/extraction. Future test collections will include mixed type data and applications.
	Variability (rate of change)	Evaluation of tradeoffs between accuracy and data rates as well as variable numbers of data streams and variable stream quality.
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues, semantics)**	The creation and measurement of the uncertainty associated with the ground-truthing process – especially when humans are involved – is challenging. The manual ground-truthing processes that have been used in the past are not scalable. Performance measurement of complex analytics must include measurement of intrinsic uncertainty as well as ground truthing error to be useful.
	Visualization	Visualization of analytic technology performance results and diagnostics including significance and various forms of uncertainty. Evaluation of analytic presentation methods to users for usability, utility, efficiency, and accuracy.
	Data Quality (syntax)	The performance of analytic technologies is highly impacted by the quality of the data they are employed against with regard to a variety of domain- and application-specific variables. Quantifying these variables is a challenging research task in itself. Mixed sources of data and performance measurement of analytic flows pose even greater challenges with regard to data quality.
	Data Types	Unstructured and semi-structured text, still images, video, audio, multimedia (audio+video).
	Data Analytics	Information extraction, filtering, search, and summarization; image and voice biometrics; speech recognition and understanding; machine translation; video person/object detection and tracking; event detection; imagery/document matching; novelty detection; a variety of structural/semantic/temporal analytics and many subtypes of the above.
Big Data Specific Challenges (Gaps)	Scaling ground-truthing to larger data, intrinsic and annotation uncertainty measurement, performance measurement for incompletely annotated data, measuring analytic performance for heterogeneous data and analytic flows involving	

Deep Learning and Social Media> Use Case 31: NIST Analytic Technology Measurement and Evaluations

	users.
Big Data Specific Challenges in Mobility	Moving training, development, and test data to evaluation participants or moving evaluation participants' analytic algorithms to computational testbeds for performance assessment. Providing developmental tools and data. Supporting agile developmental testing approaches.
Security and Privacy Requirements	Analytic algorithms working with written language, speech, human imagery, etc. must generally be tested against real or realistic data. It's extremely challenging to engineer artificial data that sufficiently captures the variability of real data involving humans. Engineered data may provide artificial challenges that may be directly or indirectly modeled by analytic algorithms and result in overstated performance. The advancement of analytic technologies themselves is increasing privacy sensitivities. Future performance testing methods will need to isolate analytic technology algorithms from the data the algorithms are tested against. Advanced architectures are needed to support security requirements for protecting sensitive data while enabling meaningful developmental performance evaluation. Shared evaluation testbeds must protect the intellectual property of analytic algorithm developers.
Highlight issues for generalizing this use case (e.g. for ref. architecture)	Scalability of analytic technology performance testing methods, source data creation, and ground truthing; approaches and architectures supporting developmental testing; protecting intellectual property of analytic algorithms and PII and other personal information in test data; measurement of uncertainty using partially-annotated data; composing test data with regard to qualities impacting performance and estimating test set difficulty; evaluating complex analytic flows involving multiple analytics, data types, and user interactions; multiple heterogeneous data streams and massive numbers of streams; mixtures of structured, semi-structured, and unstructured data sources; agile scalable developmental testing approaches and mechanisms.
More Information (URLs)	http://www.nist.gov/itl/iad/

The Ecosystem for Research> Use Case 32: DataNet Federation Consortium (DFC)

Use Case Title	DataNet Federation Consortium (DFC)	
Vertical (area)	Collaboration Environments	
Author/Company/Email	Reagan Moore / University of North Carolina at Chapel Hill / rwmoore@renci.org	
Actors/Stakeholders and their roles and responsibilities	National Science Foundation research projects: Ocean Observatories Initiative (sensor archiving); Temporal Dynamics of Learning Center (Cognitive science data grid); the iPlant Collaborative (plant genomics); Drexel engineering digital library; Odum Institute for social science research (data grid federation with Dataverse).	
Goals	Provide national infrastructure (collaboration environments) that enables researchers to collaborate through shared collections and shared workflows. Provide policy-based data management systems that enable the formation of collections, data grid, digital libraries, archives, and processing pipelines. Provide interoperability mechanisms that federate existing data repositories, information catalogs, and web services with collaboration environments.	
Use Case Description	Promote collaborative and interdisciplinary research through federation of data management systems across federal repositories, national academic research initiatives, institutional repositories, and international collaborations. The collaboration environment runs at scale: 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 Solutions	**Compute(System)**	Interoperability with workflow systems (NCSA Cyberintegrator, Kepler, Taverna)
	Storage	Interoperability across file systems, tape archives, cloud storage, object-based storage
	Networking	Interoperability across TCP/IP, parallel TCP/IP, RBUDP, HTTP
	Software	Integrated Rule Oriented Data System (iRODS)
Big Data Characteristics	**Data Source (distributed/centralized)**	Manage internationally distributed data
	Volume (size)	Petabytes, hundreds of millions of files
	Velocity (e.g. real time)	Support sensor data streams, satellite imagery, simulation output, observational data, experimental data
	Variety (multiple datasets, mashup)	Support logical collections that span administrative domains, data aggregation in containers, metadata, and workflows as objects
	Variability (rate of change)	Support active collections (mutable data), versioning of data, and persistent identifiers
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues)**	Provide reliable data transfer, audit trails, event tracking, periodic validation of assessment criteria (integrity, authenticity), distributed debugging
	Visualization	Support execution of external visualization systems through automated workflows (GRASS)
	Data Quality	Provide mechanisms to verify quality through automated workflow procedures
	Data Types	Support parsing of selected formats (NetCDF, HDF5, Dicom), and provide mechanisms to invoke other data manipulation methods

The Ecosystem for Research> Use Case 32: DataNet Federation Consortium (DFC)

	Data Analytics	Provide support for invoking analysis workflows, tracking workflow provenance, sharing of workflows, and re-execution of workflows
Big Data Specific Challenges (Gaps)	Provide standard policy sets that enable a new community to build upon data management plans that address federal agency requirements	
Big Data Specific Challenges in Mobility	Capture knowledge required for data manipulation, and apply resulting procedures at either the storage location, or a computer server.	
Security and Privacy Requirements	Federate across existing authentication environments through Generic Security Service API and Pluggable Authentication Modules (GSI, Kerberos, InCommon, Shibboleth). Manage access controls on files independently of the storage location.	
Highlight issues for generalizing this use case (e.g. for ref. architecture)	Currently 25 science and engineering domains have projects that rely on the iRODS policy-based data management system:	
	Astrophysics	Auger supernova search
	Atmospheric science	NASA Langley Atmospheric Sciences Center
	Biology	Phylogenetics at CC IN2P3
	Climate	NOAA National Climatic Data Center
	Cognitive Science	Temporal Dynamics of Learning Center
	Computer Science	GENI experimental network
	Cosmic Ray	AMS experiment on the International Space Station
	Dark Matter Physics	Edelweiss II
	Earth Science	NASA Center for Climate Simulations
	Ecology	CEED Caveat Emptor Ecological Data
	Engineering	CIBER-U
	High Energy Physics	BaBar
	Hydrology	Institute for the Environment, UNC-CH; Hydroshare
	Genomics	Broad Institute, Wellcome Trust Sanger Institute
	Medicine	Sick Kids Hospital
	Neuroscience	International Neuroinformatics Coordinating Facility
	Neutrino Physics	T2K and dChooz neutrino experiments
	Oceanography	Ocean Observatories Initiative
	Optical Astronomy	National Optical Astronomy Observatory
	Particle Physics	Indra
	Plant genetics	the iPlant Collaborative
	Quantum Chromodynamics	IN2P3
	Radio Astronomy	Cyber Square Kilometer Array, TREND, BAOradio
	Seismology	Southern California Earthquake Center
	Social Science	Odum Institute for Social Science Research, TerraPop
More Information (URLs)	The DataNet Federation Consortium: http://www.datafed.org iRODS: http://www.irods.org	
Note: A major challenge is the ability to capture knowledge needed to interact with the data products of a research domain. In policy-based data management systems, this is done by encapsulating the knowledge in procedures that are controlled through policies. The procedures can automate retrieval of data from external repositories, or execute processing workflows, or enforce management policies on the resulting data products. A standard application is the enforcement of data management plans and the verification that the plan has been successfully applied.		

See Figure 4: DataNet Federation Consortium DFC – iRODS architecture.

The Ecosystem for Research> Use Case 33: The 'Discinnet Process'

Use Case Title	The 'Discinnet process', metadata <-> Big Data global experiment	
Vertical (area)	Scientific Research: Interdisciplinary Collaboration	
Author/Company/Email	P. Journeau / Discinnet Labs / phjourneau@discinnet.org	
Actors/Stakeholders and their roles and responsibilities	Actors Richeact, Discinnet Labs and I4OpenResearch fund France/Europe. American equivalent pending. Richeact is fundamental research and development epistemology, Discinnet Labs applied in web 2.0 http://www.discinnet.org, I4 non-profit warrant.	
Goals	Richeact scientific goal is to reach predictive interdisciplinary model of research fields' behavior (with related meta-grammar). Experimentation through global sharing of now multidisciplinary, later interdisciplinary Discinnet process/web mapping and new scientific collaborative communication and publication system. Expected sharp impact to reducing uncertainty and time between theoretical, applied, technology research and development steps.	
Use Case Description	Currently 35 clusters started, close to 100 awaiting more resources and potentially much more open for creation, administration and animation by research communities. Examples range from optics, cosmology, materials, microalgae, health to applied maths, computation, rubber and other chemical products/issues. How does a typical case currently work: - A researcher or group wants to see how a research field is faring and in a minute defines the field on Discinnet as a 'cluster' - Then it takes another 5 to 10 mn to parameter the first/main dimensions, mainly measurement units and categories, but possibly later on some variable limited time for more dimensions - Cluster then may be filled either by doctoral students or reviewing researchers and/or communities/researchers for projects/progress Already significant value but now needs to be disseminated and advertised although maximal value to come from interdisciplinary/projective next version. Value is to detect quickly a paper/project of interest for its results and next step is trajectory of the field under types of interactions from diverse levels of oracles (subjects/objects) + from interdisciplinary context.	
Current Solutions	**Compute(System)**	Currently on OVH (Hosting company http://www.ovh.co.uk/) servers (mix shared + dedicated)
	Storage	OVH
	Networking	To be implemented with desired integration with others
	Software	Current version with Symfony-PHP, Linux, MySQL
Big Data Characteristics	**Data Source (distributed/centralized)**	Currently centralized, soon distributed per country and even per hosting institution interested by own platform
	Volume (size)	Not significant : this is a metadata base, not Big Data
	Velocity (e.g. real time)	Real time
	Variety (multiple datasets, mashup)	Link to Big data still to be established in a Meta<->Big relationship not yet implemented (with experimental databases and already 1^{st} level related metadata)
	Variability (rate of change)	Currently real time, for further multiple locations and distributed architectures, periodic (such as nightly)
Big Data Science (collection, curation, analysis,	**Veracity (Robustness Issues, semantics)**	Methods to detect overall consistency, holes, errors, misstatements, known but mostly to be implemented
	Visualization	Multidimensional (hypercube)

The Ecosystem for Research> Use Case 33: The 'Discinnet Process'

action)	**Data Quality (syntax)**	A priori correct (directly human captured) with sets of checking + evaluation processes partly implemented
	Data Types	'cluster displays' (image), vectors, categories, PDFs
	Data Analytics	
Big Data Specific Challenges (Gaps)	Our goal is to contribute to Big 2 Metadata challenge by systematic reconciling between metadata from many complexity levels with ongoing input from researchers from ongoing research process. Current relationship with Richeact is to reach the interdisciplinary model, using meta-grammar itself to be experimented and its extent fully proven to bridge efficiently the gap between as remote complexity levels as semantic and most elementary (big) signals. Example with cosmological models versus many levels of intermediary models (particles, gases, galactic, nuclear, geometries). Others with computational versus semantic levels.	
Big Data Specific Challenges in Mobility	Appropriate graphic interface power	
Security and Privacy Requirements	Several levels already available and others planned, up to physical access keys and isolated servers. Optional anonymity, usual protected exchanges	
Highlight issues for generalizing this use case (e.g. for ref. architecture)	Through 2011-2013, we have shown on http://www.discinnet.org that all kinds of research fields could easily get into Discinnet type of mapping, yet developing and filling a cluster requires time and/or dedicated workers.	
More Information (URLs)	On http://www.discinnet.org the already started or starting clusters can be watched in one click on 'cluster' (field) title and even more detail is available through free registration (more resource available when registering as researcher (publications) or pending (doctoral student) Maximum level of detail is free for contributing researchers in order to protect communities but available to external observers for symbolic fee: all suggestions for improvements and better sharing welcome. We are particularly open to provide and support experimental appropriation by doctoral schools to build and study the past and future behavior of clusters in Earth sciences, Cosmology, Water, Health, Computation, Energy/Batteries, Climate models, Space, etc..	
Note: We are open to facilitate wide appropriation of both global, regional and local versions of the platform (for instance by research institutions, publishers, networks with desirable maximal data sharing for the greatest benefit of advancement of science.		

This publication is available free of charge from: https://doi.org/10.6028/NIST.SP.1500-3r1

The Ecosystem for Research> Use Case 34: Graph Search on Scientific Data

Use Case Title	Enabling Face-Book like Semantic Graph-search on Scientific Chemical and Text-based Data	
Vertical (area)	Management of Information from Research Articles	
Author/Company/Email	Talapady Bhat, bhat@nist.gov	
Actors/Stakeholders and their roles and responsibilities	Chemical structures, Protein Data Bank, Material Genome Project, Open-GOV initiative, Semantic Web, Integrated Data-graphs, Scientific social media	
Goals	Establish infrastructure, terminology and semantic data-graphs to annotate and present technology information using 'root' and rule-based methods used primarily by some Indo-European languages like Sanskrit and Latin.	
Use Case Description	• Social media hype o Internet and social media play a significant role in modern information exchange. Every day most of us use social-media both to distribute and receive information. Two of the special features of many social media like Face-Book are ▪ the community is both data-providers and data-users ▪ they store information in a pre-defined 'data-shelf' of a data-graph ▪ Their core infrastructure for managing information is reasonably language free • What this has to do with managing scientific information? During the last few decades science has truly evolved to become a community activity involving every country and almost every household. We routinely 'tune-in' to Internet resources to share and seek scientific information. What are the challenges in creating social media for science o Creating a social media of scientific information needs an infrastructure where many scientists from various parts of the world can participate and deposit results of their experiment. Some of the issues that one has to resolve prior to establishing a scientific social media are: ▪ How to minimize challenges related to local language and its grammar? ▪ How to determining the 'data-graph' to place an information in an intuitive way without knowing too much about the data management? ▪ How to find relevant scientific data without spending too much time on the Internet? **Approach:** Most languages and more so Sanskrit and Latin use a novel 'root'-based method to facilitate the creation of on-demand, discriminating words to define concepts. Some such examples from English are Bio-logy, Bio-chemistry. Youga, Yogi, Yogendra, Yogesh are examples from Sanskrit. Genocide is an example from Latin. These words are created on-demand based on best-practice terms and their capability to serve as node in a discriminating data-graph with self-explained meaning.	
Current Solutions	**Compute(System)**	Cloud for the participation of community
	Storage	Requires expandable on-demand based resource that is suitable for global users location and requirements
	Networking	Needs good network for the community participation
	Software	Good database tools and servers for data-graph manipulation are needed
Big Data Characteristics	**Data Source (distributed/centralized)**	Distributed resource with a limited centralized capability
	Volume (size)	Undetermined. May be few terabytes at the beginning

The Ecosystem for Research> Use Case 34: Graph Search on Scientific Data

	Velocity (e.g. real time)	Evolving with time to accommodate new best-practices
	Variety (multiple datasets, mashup)	Wildly varying depending on the types available technological information
	Variability (rate of change)	Data-graphs are likely to change in time based on customer preferences and best-practices
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues)**	Technological information is likely to be stable and robust
	Visualization	Efficient data-graph based visualization is needed
	Data Quality	Expected to be good
	Data Types	All data types, image to text, structures to protein sequence
	Data Analytics	Data-graphs is expected to provide robust data-analysis methods
Big Data Specific Challenges (Gaps)	This is a community effort similar to many social media. Providing a robust, scalable, on-demand infrastructures in a manner that is use-case and user-friendly is a real challenge by any existing conventional methods	
Big Data Specific Challenges in Mobility	A community access is required for the data and thus it has to be media and location independent and thus requires high mobility too.	
Security and Privacy Requirements	None since the effort is initially focused on publicly accessible data provided by open-platform projects like open-gov, MGI and protein data bank.	
Highlight issues for generalizing this use case (e.g. for ref. architecture)	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 that we are trying to solve.	
More Information (URLs)	http://www.eurekalert.org/pub_releases/2013-07/aiop-ffm071813.php http://xpdb.nist.gov/chemblast/pdb.pl http://xpdb.nist.gov/chemblast/pdb.pl	

Note: Many reports, including a recent one on Material Genome Project finds that exclusive top-down solutions to facilitate data sharing and integration are not desirable for federated 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 developer at the beginning. One of the successful effort used by many prominent languages is that of 'roots' and rules that form the framework for creating on-demand words for communication. In this approach a top-down method is used to establish a limited number of highly re-usable words called 'roots' by surveying the existing best practices in building terminology. These 'roots' are combined using few 'rules' to create terms on-demand by a bottom-up step.
Y(uj) (join), O (creator, God, brain), Ga (motion, initiation) –leads to 'Yoga' in Sanskrit, English
Geno (genos)-cide–race based killing – Latin, English
Bio-technology –English, Latin
Red-light, red-laser-light –English.
A press release by the American Institute of Physics on this approach is at
http://www.eurekalert.org/pub_releases/2013-07/aiop-ffm071813.php
Our efforts to develop automated and rule and root-based methods (Chem-BLAST -.
http://xpdb.nist.gov/chemblast/pdb.pl) to identify and use best-practice, discriminating terms in generating semantic data-graphs for science started almost a decade back with a chemical structure database. This database has millions of structures obtained from the Protein Data Bank and the PubChem used world-wide. Subsequently we extended our efforts to build root-based terms to text-based data of cell-images. In this work

The Ecosystem for Research> Use Case 34: Graph Search on Scientific Data

we use few simple rules to define and extend terms based on best-practice as decided by weaning through millions of popular use-cases chosen from over hundred biological ontologies. Currently we are working on extending this method to publications of interest to Material Genome, Open-Gov and NIST-wide publication archive - NIKE. - http://xpdb.nist.gov/nike/term.pl. These efforts are a component of Research Data Alliance Working Group on Metadata https://www.rd-alliance.org/filedepot_download/694/160 and https://rd-alliance.org/poster-session-rda-2nd-plenary-meeting.html

The Ecosystem for Research> Use Case 35: Light Source Beamlines

Use Case Title	Light source beamlines	
Vertical (area)	Research (Biology, Chemistry, Geophysics, Materials Science, others)	
Author/Company/Email	Eli Dart, LBNL (eddart@lbl.gov)	
Actors/Stakeholders and their roles and responsibilities	Research groups from a variety of scientific disciplines (see above)	
Goals	Use of a variety of experimental techniques to determine structure, composition, behavior, or other attributes of a sample relevant to scientific enquiry.	
Use Case Description	Samples are exposed to X-rays 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. The reconstructed images are used by scientist's analyses.	
Current Solutions	**Compute(System)**	Computation ranges from single analysis hosts to high-throughput computing systems at computational facilities
	Storage	Local storage on the order of 1-40TB on Windows or Linux data servers at facility for temporary storage, over 60TB on disk at NERSC, over 300TB on tape at NERSC
	Networking	10Gbps Ethernet at facility, 100Gbps to NERSC
	Software	A variety of commercial and open source software is used for data analysis – examples include: • Octopus (http://www.inct.be/en/software/octopus) for Tomographic Reconstruction • Avizo (http://vsg3d.com) and FIJI (a distribution of ImageJ; http://fiji.sc) for Visualization and Analysis Data transfer is accomplished using physical transport of portable media (severely limits performance) or using high-performance GridFTP, managed by Globus Online or workflow systems such as SPADE.
Big Data Characteristics	**Data Source (distributed/centralized)**	Centralized (high resolution camera at facility). Multiple beamlines per facility with high-speed detectors.
	Volume (size)	3GB to 30GB per sample – up to 15 samples/day
	Velocity (e.g. real time)	Near real-time analysis needed for verifying experimental parameters (lower resolution OK). Automation of analysis would dramatically improve scientific productivity.
	Variety (multiple datasets, mashup)	Many detectors produce similar types of data (e.g. TIFF files), but experimental context varies widely
	Variability (rate of change)	Detector capabilities are increasing rapidly. Growth is essentially Moore's Law. Detector area is increasing exponentially (1k x 1k, 2k x 2k, 4k x 4k, …) and readout is increasing exponentially (1Hz, 10Hz, 100Hz, 1kHz, …). Single detector data rates are expected to reach 1 GB per second within 2 years.
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues)**	Near real-time analysis required to verify experimental parameters. In many cases, early analysis can dramatically improve experiment productivity by providing early feedback. This implies high-throughput computing, high-performance data transfer, and high-speed storage are routinely available.

The Ecosystem for Research> Use Case 35: Light Source Beamlines

	Visualization	Visualization is key to a wide variety of experiments at all light source facilities
	Data Quality	Data quality and precision are critical (especially since beam time is scarce, and re-running an experiment is often impossible).
	Data Types	Many beamlines generate image data (e.g. TIFF files)
	Data Analytics	Volume reconstruction, feature identification, others
Big Data Specific Challenges (Gaps)	Rapid increase in camera capabilities, need for automation of data transfer and near-real-time analysis.	
Big Data Specific Challenges in Mobility	Data transfer to large-scale computing facilities is becoming necessary because of the computational power required to conduct the analysis on time scales useful to the experiment. Large number of beamlines (e.g. 39 at LBNL ALS) means that aggregate data load is likely to increase significantly over the coming years.	
Security and Privacy Requirements	Varies with project.	
Highlight issues for generalizing this use case (e.g. for ref. architecture)	There will be significant need for a generalized infrastructure for analyzing GBs per second of data from many beamline detectors at multiple facilities. Prototypes exist now, but routine deployment will require additional resources.	
More Information (URLs)	http://www-als.lbl.gov/ http://www.aps.anl.gov/ https://portal.slac.stanford.edu/sites/lcls_public/Pages/Default.aspx	

Astronomy and Physics> Use Case 36: Catalina Digital Sky Survey for Transients

Use Case Title	Catalina Real-Time Transient Survey (CRTS): a digital, panoramic, synoptic sky survey	
Vertical (area)	Scientific Research: Astronomy	
Author/Company/Email	S. G. Djorgovski / Caltech / george@astro.caltech.edu	
Actors/Stakeholders and their roles and responsibilities	The survey team: data processing, quality control, analysis and interpretation, publishing, and archiving. Collaborators: a number of research groups world-wide: further work on data analysis and interpretation, follow-up observations, and publishing. User community: all of the above, plus the astronomical community world-wide: further work on data analysis and interpretation, follow-up observations, and publishing.	
Goals	The survey explores the variable universe in the visible light regime, on time scales 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 (active galactic nuclei) and their relativistic jets, high proper motion stars, etc.	
Use Case Description	The data are collected from 3 telescopes (2 in Arizona and 1 in Australia), with additional ones expected in the near future (in Chile). The original motivation is a search for near-Earth (NEO) and potential planetary hazard (PHO) asteroids, funded by NASA, and conducted by a group at the Lunar and Planetary Laboratory (LPL) at the Univ. of Arizona (UA); that is the Catalina Sky Survey proper (CSS). The data stream is shared by the CRTS for the purposes for exploration of the variable universe, beyond the Solar system, led by the Caltech group. Approximately 83% of the entire sky is being surveyed through multiple passes (crowded regions near the Galactic plane, and small areas near the celestial poles are excluded). The data are preprocessed at the telescope, and transferred to LPL/UA, and hence to 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 period (CRTS has a completely open data policy). Further data analysis includes automated and semi-automated classification of the detected transient events, additional observations using other telescopes, scientific interpretation, and publishing. In this process, it makes a heavy use of the archival data from a wide variety of geographically distributed resources connected through the Virtual Observatory (VO) framework. Light curves (flux histories) are accumulated for ≈ 500 million sources detected in the survey, each with a few hundred data points on average, spanning up to 8 years, and growing. These are served to the community from the archives at Caltech, and shortly from IUCAA, India. This is an unprecedented dataset for the exploration of time domain in astronomy, in terms of the temporal and area coverage and depth. CRTS is a scientific and methodological testbed and precursor of the grander surveys to come, notably the Large Synoptic Survey Telescope (LSST), expected to operate in 2020's.	
Current Solutions	**Compute(System)**	Instrument and data processing computers: a number of desktop and small server class machines, although more powerful machinery is needed for some data analysis tasks. This is not so much a computationally-intensive project, but rather a data-handling-intensive one.

Astronomy and Physics> Use Case 36: Catalina Digital Sky Survey for Transients

	Storage	Several multi-TB / tens of TB servers.
	Networking	Standard inter-university Internet connections.
	Software	Custom data processing pipeline and data analysis software, operating under Linux. Some archives on Windows machines, running a MS SQL server databases.
Big Data Characteristics	**Data Source (distributed/centralized)**	Distributed: 1. Survey data from 3 (soon more?) telescopes 2. Archival data from a variety of resources connected through the VO framework 3. Follow-up observations from separate telescopes
	Volume (size)	The survey generates up to ≈ 0.1 TB per clear night; ≈ 100 TB in current data holdings. Follow-up observational data amount to no more than a few % of that. Archival data in external (VO-connected) archives are in PBs, but only a minor fraction is used.
	Velocity (e.g. real time)	Up to ≈ 0.1 TB / night of the raw survey data.
	Variety (multiple datasets, mashup)	The primary survey data in the form of images, processed to catalogs of sources (db tables), and time series for individual objects (light curves). Follow-up observations consist of images and spectra. Archival data from the VO data grid include all of the above, from a wide variety of sources and different wavelengths.
	Variability (rate of change)	Daily data traffic fluctuates from ≈ 0.01 to ≈ 0.1 TB / day, not including major data transfers between the principal archives (Caltech, UA, and IUCAA).
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues, semantics)**	A variety of automated and human inspection quality control mechanisms is implemented at all stages of the process.
	Visualization	Standard image display and data plotting packages are used. We are exploring visualization mechanisms for highly dimensional data parameter spaces.
	Data Quality (syntax)	It varies, depending on the observing conditions, and it is evaluated automatically: error bars are estimated for all relevant quantities.
	Data Types	Images, spectra, time series, catalogs.
	Data Analytics	A wide variety of the existing astronomical data analysis tools, plus a large amount of custom developed tools and software, some of it a research project in itself.
Big Data Specific Challenges (Gaps)	Development of machine learning tools for data exploration, and in particular for an automated, real-time classification of transient events, given the data sparsity and heterogeneity. Effective visualization of hyper-dimensional parameter spaces is a major challenge for all of us.	
Big Data Specific Challenges in Mobility	Not a significant limitation at this time.	

Astronomy and Physics> Use Case 36: Catalina Digital Sky Survey for Transients

Security and Privacy Requirements	None.
Highlight issues for generalizing this use case (e.g. for ref. architecture)	• Real-time processing and analysis of massive data streams from a distributed sensor network (in this case telescopes), with a need to identify, characterize, and respond to the transient events of interest in (near) real time. • Use of highly distributed archival data resources (in this case VO-connected archives) for data analysis and interpretation. • Automated classification given the very sparse and heterogeneous data, dynamically evolving in time as more data come in, and follow-up decision making given limited and sparse resources (in this case follow-up observations with other telescopes).
More Information (URLs)	CRTS survey: http://crts.caltech.edu CSS survey: http://www.lpl.arizona.edu/css For an overview of the classification challenges, see, e.g., http://arxiv.org/abs/1209.1681 For a broader context of sky surveys, past, present, and future, see, e.g., the review http://arxiv.org/abs/1209.1681
Note: CRTS can be seen as a good precursor to the astronomy's flagship project, the Large Synoptic Sky Survey (LSST; http://www.lsst.org), now under development. Their anticipated data rates (≈ 20TB to 30 TB per clear night, tens of PB over the duration of the survey) are directly on the Moore's law scaling from the current CRTS data rates and volumes, and many technical and methodological issues are very similar. It is also a good case for real-time data mining and knowledge discovery in massive data streams, with distributed data sources and computational resources.	

See Figure 5: Catalina CRTS: A Digital, Panoramic, Synoptic Sky Survey

The figure shows one possible schematic architecture for a cyber-infrastructure for time domain astronomy. Transient event data streams are produced by survey pipelines from the telescopes on the ground or in space, and the events with their observational descriptions are ingested by one or more depositories, from which they can be disseminated electronically to human astronomers or robotic telescopes. Each event is assigned an evolving portfolio of information, which would include all of the available data on that celestial position, 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. They are fed into one or more automated event characterization, classification, and prioritization engines that deploy a variety of machine learning tools for these tasks. Their 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 (human or robotic) can tap into the system at multiple points, both for an information retrieval, and to contribute new information, through a standardized set of formats and protocols. This could be done in a (near) real time, or in an archival (not time critical) modes.

Astronomy and Physics> Use Case 37: Cosmological Sky Survey and Simulations

Use Case Title	DOE Extreme Data from Cosmological Sky Survey and Simulations	
Vertical (area)	Scientific Research: Astrophysics	
Author/Company/Email	PIs: Salman Habib, Argonne National Laboratory; Andrew Connolly, University of Washington	
Actors/Stakeholders and their roles and responsibilities	Researchers studying dark matter, dark energy, and the structure of the early universe.	
Goals	Clarify the nature of dark matter, dark energy, and inflation, some of the most exciting, perplexing, and challenging questions facing modern physics. Emerging, unanticipated measurements are pointing toward a need for physics beyond the successful Standard Model of particle physics.	
Use Case Description	This investigation requires an intimate interplay between Big Data from experiment and simulation as well as massive computation. The melding of all will 1) Provide the direct means for cosmological discoveries that require a strong connection between theory and observations ('precision cosmology'); 2) Create an essential 'tool of discovery' in dealing with large datasets generated by complex instruments; and, 3) Generate and share results from high-fidelity simulations that are necessary to understand and control systematics, especially astrophysical systematics.	
Current Solutions	**Compute(System)**	Hours: 24M (NERSC / Berkeley Lab), 190M (ALCF / Argonne), 10M (OLCF / Oak Ridge)
	Storage	180 TB (NERSC / Berkeley Lab)
	Networking	ESNet connectivity to the national labs is adequate today.
	Software	MPI, OpenMP, C, C++, F90, FFTW, viz packages, python, FFTW, numpy, Boost, OpenMP, ScaLAPCK, PSQL and MySQL databases, Eigen, cfitsio, astrometry.net, and Minuit2
Big Data Characteristics	**Data Source (distributed/centralized)**	Observational data will be generated by the Dark Energy Survey (DES) and the Zwicky Transient Factory in 2015 and by the Large Synoptic Sky Survey starting in 2019. Simulated data will generated at DOE supercomputing centers.
	Volume (size)	DES: 4 PB, ZTF 1 PB/year, LSST 7 PB/year, Simulations > 10 PB in 2017
	Velocity (e.g. real time)	LSST: 20 TB/day
	Variety (multiple datasets, mashup)	1) Raw Data from sky surveys 2) Processed Image data 3) Simulation data
	Variability (rate of change)	Observations are taken nightly; supporting simulations are run throughout the year, but data can be produced sporadically depending on access to resources
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues)**	
	Visualization	Interpretation of results from detailed simulations requires advanced analysis and visualization techniques

Astronomy and Physics> Use Case 37: Cosmological Sky Survey and Simulations

		and capabilities. Supercomputer I/O subsystem limitations are forcing researchers to explore "in-situ" analysis to replace post-processing methods.
	Data Quality	
	Data Types	Image data from observations must be reduced and compared with physical quantities derived from simulations. Simulated sky maps must be produced to match observational formats.
	Data Analytics	
Big Data Specific Challenges (Gaps)	Storage, sharing, and analysis of 10s of PBs of observational and simulated data.	
Big Data Specific Challenges in Mobility	LSST will produce 20 TB of data per day. This must be archived and made available to researchers world-wide.	
Security and Privacy Requirements		
Highlight issues for generalizing this use case (e.g. for ref. architecture)		
More Information (URLs)	http://www.lsst.org/lsst/ http://www.nersc.gov/ http://science.energy.gov/hep/research/non-accelerator-physics/ http://www.nersc.gov/assets/Uploads/HabibcosmosimV2.pdf	

Astronomy and Physics> Use Case 38: Large Survey Data for Cosmology

Use Case Title	Large Survey Data for Cosmology	
Vertical (area)	Scientific Research: Cosmic Frontier	
Author/Company/Email	Peter Nugent / LBNL / penugent@lbl.gov	
Actors/Stakeholders and their roles and responsibilities	Dark Energy Survey, Dark Energy Spectroscopic Instrument, Large Synoptic Survey Telescope. ANL, BNL, FNAL, LBL and SLAC: Create the instruments/telescopes, run the survey and perform the cosmological analysis.	
Goals	Provide a way to reduce photometric data in real time for supernova discovery and follow-up and to handle the large volume of observational data (in conjunction with simulation data) to reduce systematic uncertainties in the measurement of the cosmological parameters via baryon acoustic oscillations, galaxy cluster counting and weak lensing measurements.	
Use Case Description	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 as well as NERSC for storage and "reduction". Subtraction pipelines are run using extant imaging data to find new optical transients through machine learning algorithms. Then galaxies and stars in both the individual and stacked images are identified, catalogued, and finally their properties measured and stored in a database.	
Current Solutions	**Compute(System)**	Linux cluster, Oracle RDBMS server, large memory machines, standard Linux interactive hosts. For simulations, HPC resources.
	Storage	Oracle RDBMS, Postgres psql, as well as GPFS and Lustre file systems and tape archives.
	Networking	Provided by NERSC
	Software	Standard astrophysics reduction software as well as Perl/Python wrapper scripts, Linux Cluster scheduling and comparison to large amounts of simulation data via techniques like Cholesky decomposition.
Big Data Characteristics	**Data Source (distributed/centralized)**	Distributed. Typically between observation and simulation data.
	Volume (size)	LSST will generate 60 PB of imaging data and 15 PB of catalog data and a correspondingly large (or larger) amount of simulation data. Over 20 TB of data per night.
	Velocity (e.g. real time)	20TB of data will have to be subtracted each night in as near real time as possible in order to maximize the science for supernovae.
	Variety (multiple datasets, mashup)	While the imaging data is similar, the analysis for the 4 different types of cosmological measurements and comparisons to simulation data is quite different.
	Variability (rate of change)	Weather and sky conditions can radically change both the quality and quantity of data.
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues)**	Astrophysical data is a statistician's nightmare as the both the uncertainties in a given measurement change from night-to-night in addition to the cadence being highly unpredictable. Also, most all of the cosmological measurements are systematically limited, and thus understanding these as best possible is the highest priority for a given survey.

Astronomy and Physics> Use Case 38: Large Survey Data for Cosmology

	Visualization	Interactive speed of web UI on very large datasets is an ongoing challenge. Basic querying and browsing of data to find new transients as well as monitoring the quality of the survey is a must. Ability to download large amounts of data for offline analysis is another requirement of the system. Ability to combine both simulation and observational data is also necessary.
	Data Quality	Understanding the systematic uncertainties in the observational data is a prerequisite to a successful cosmological measurement. Beating down the uncertainties in the simulation data to under this level is a huge challenge for future surveys.
	Data Types	Cf. above on "Variety"
	Data Analytics	
Big Data Specific Challenges (Gaps)	New statistical techniques for understanding the limitations in simulation data would be beneficial. Often it is the case where there is not enough computing time to generate all the simulations one wants and thus there is a reliance on emulators to bridge the gaps. Techniques for handling Cholesky decomposition for thousands of simulations with matrices of order 1M on a side.	
Big Data Specific Challenges in Mobility	Performing analysis on both the simulation and observational data simultaneously.	
Security and Privacy Requirements	No special challenges. Data is either public or requires standard login with password.	
Highlight issues for generalizing this use case (e.g. for ref. architecture)	Parallel databases which could handle imaging data would be an interesting avenue for future research.	
More Information (URLs)	http://www.lsst.org/lsst, http://desi.lbl.gov, and http://www.darkenergysurvey.org	

Astronomy and Physics> Use Case 39: Analysis of LHC (Large Hadron Collider) Data

Use Case Title	Particle Physics: Analysis of LHC (Large Hadron Collider) Data (Discovery of Higgs particle)	
Vertical (area)	Scientific Research: Physics	
Author/Company/Email	Michael Ernst mernst@bnl.gov, Lothar Bauerdick bauerdick@fnal.gov based on an initial version written by Geoffrey Fox, Indiana University gcf@indiana.edu, Eli Dart, LBNL eddart@lbl.gov,	
Actors/Stakeholders and their roles and responsibilities	Physicists(Design and Identify need for Experiment, Analyze Data) Systems Staff (Design, Build and Support distributed Computing Grid), Accelerator Physicists (Design, Build and Run Accelerator), Government (funding based on long term importance of discoveries in field))	
Goals	Understanding properties of fundamental particles	
Use Case Description	CERN LHC Detectors and Monte Carlo producing events describing particle-apparatus interaction. Processed information defines physics properties of events (lists of particles with type and momenta). These events are analyzed to find new effects; both new particles (Higgs) and present evidence that conjectured particles (Supersymmetry) not seen.	
Current Solutions	**Compute(System)**	WLCG and Open Science Grid in the US integrate computer centers worldwide that provide computing and storage resources into a single infrastructure accessible by all LHC physicists. 350,000 cores running “continuously” arranged in 3 tiers (CERN, “Continents/Countries”. “Universities”). Uses “Distributed High Throughput Computing (DHTC)”; 200PB storage, >2million jobs/day.
	Storage	ATLAS: • Brookhaven National Laboratory Tier1 tape: 10PB ATLAS data on tape managed by HPSS (incl. RHIC/NP the total data volume is 35PB) • Brookhaven National Laboratory Tier1 disk: 11PB; using dCache to virtualize a set of ≈60 heterogeneous storage servers with high-density disk backend systems • US Tier2 centers, disk cache: 16PB CMS: • Fermilab US Tier1, reconstructed, tape/cache: 20.4PB • US Tier2 centers, disk cache: 7PB • US Tier3 sites, disk cache: 1.04PB
	Networking	• As experiments have global participants (CMS has 3600 participants from 183 institutions in 38 countries), the data at all levels is transported and accessed across continents. • Large scale automated data transfers occur over science networks across the globe. LHCOPN and LHCONE network overlay provide dedicated network allocations and traffic isolation for LHC data traffic

Astronomy and Physics> Use Case 39: Analysis of LHC (Large Hadron Collider) Data

		• ATLAS Tier1 data center at BNL has 160Gbps internal paths (often fully loaded). 70Gbps WAN connectivity provided by ESnet. • CMS Tier1 data center at FNAL has 90Gbps WAN connectivity provided by ESnet • Aggregate wide area network traffic for LHC experiments is about 25Gbps steady state worldwide
	Software	The scalable ATLAS workload/workflow management system PanDA manages ≈1 million production and user analysis jobs on globally distributed computing resources (≈100 sites) per day. The new ATLAS distributed data management system Rucio is the core component keeping track of an inventory of currently ≈130PB of data distributed across grid resources and to orchestrate data movement between sites. The data volume is expected to grow to exascale size in the next few years. Based on the xrootd system ATLAS has developed FAX, a federated storage system that allows remote data access. Similarly, CMS is using the OSG glideinWMS infrastructure to manage its workflows for production and data analysis the PhEDEx system to orchestrate data movements, and the AAA/xrootd system to allow remote data access. Experiment-specific physics software including simulation packages, data processing, advanced statistic packages, etc.
Big Data Characteristics	**Data Source (distributed/centralized)**	High speed detectors produce large data volumes: • ATLAS detector at CERN: Originally 1 PB/sec raw data rate, reduced to 300MB/sec by multi-stage trigger. • CMS detector at CERN: similar Data distributed to Tier1 centers globally, which serve as data sources for Tier2 and Tier3 analysis centers
	Volume (size)	15 Petabytes per year from Detectors and Analysis
	Velocity (e.g. real time)	• Real time with some long LHC "shut downs" (to improve accelerator and detectors) with no data except Monte Carlo. • Besides using programmatically and dynamically replicated datasets, real-time remote I/O (using XrootD) is increasingly used by analysis which requires reliable high-performance networking capabilities to reduce file copy and storage system overhead
	Variety	Lots of types of events with from 2- few hundred final particle but all data is collection of particles after initial

Astronomy and Physics> Use Case 39: Analysis of LHC (Large Hadron Collider) Data

	(multiple datasets, mashup)	analysis. Events are grouped into datasets; real detector data is segmented into ≈20 datasets (with partial overlap) on the basis of event characteristics determined through real-time trigger system, while different simulated datasets are characterized by the physics process being simulated.
	Variability (rate of change)	Data accumulates and does not change character. What you look for may change based on physics insight. As understanding of detectors increases, large scale data reprocessing tasks are undertaken.
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues)**	One can lose modest amount of data without much pain as errors proportional to 1/SquareRoot(Events gathered), but such data loss must be carefully accounted. Importance that accelerator and experimental apparatus work both well and in understood fashion. Otherwise data too "dirty" / "uncorrectable".
	Visualization	Modest use of visualization outside histograms and model fits. Nice event displays but discovery requires lots of events so this type of visualization of secondary importance
	Data Quality	Huge effort to make certain complex apparatus well understood (proper calibrations) and "corrections" properly applied to data. Often requires data to be re-analyzed
	Data Types	Raw experimental data in various binary forms with conceptually a name: value syntax for name spanning “chamber readout” to “particle momentum”. Reconstructed data is processed to produce dense data formats optimized for analysis
	Data Analytics	Initial analysis is processing of experimental data specific to each experiment (ALICE, ATLAS, CMS, LHCb) producing summary information. Second step in analysis uses “exploration” (histograms, scatter-plots) with model fits. Substantial Monte-Carlo computations are necessary to estimate analysis quality. A large fraction (≈60%) of the available CPU resources available to the ATLAS collaboration at the Tier-1 and the Tier-2 centers is used for simulated event production. The ATLAS simulation requirements are completely driven by the physics community in terms of analysis needs and corresponding physics goals. The current physics analyses are looking at real data samples of roughly 2 billion (B) events taken in 2011 and 3B events taken in 2012 (this represents ≈5 PB of experimental data), and ATLAS has roughly 3.5B MC events for 2011 data, and 2.5B MC events for 2012 (this represents ≈6 PB of simulated data). Given the resource requirements to fully simulate an event using the GEANT

Astronomy and Physics> Use Case 39: Analysis of LHC (Large Hadron Collider) Data

		4 package, ATLAS can currently produce about 4 million events per day using the entire processing capacity available to production worldwide. Due to its high CPU cost, the outputs of full Geant4 simulation (HITS) are stored in one custodial tape copy on Tier1 tapes to be re-used in several Monte-Carlo re-processings. The HITS from faster simulation flavors will be only of transient nature in LHC Run 2.
Big Data Specific Challenges (Gaps)	The translation of scientific results into new knowledge, solutions, policies and decisions is foundational to the science mission associated with LHC data analysis and HEP in general. However, while advances in experimental and computational technologies have led to an exponential growth in the volume, velocity, and variety of data available for scientific discovery, advances in technologies to convert this data into actionable knowledge have fallen far short of what the HEP community needs to deliver timely and immediately impacting outcomes. Acceleration of the scientific knowledge discovery process is essential if DOE scientists are to continue making major contributions in HEP. Today's worldwide analysis engine, serving several thousand scientists, will have to be commensurately extended in the cleverness of its algorithms, the automation of the processes, and the reach (discovery) of the computing, to enable scientific understanding of the detailed nature of the Higgs boson. E.g. the approximately forty different analysis methods used to investigate the detailed characteristics of the Higgs boson (many using machine learning techniques) must be combined in a mathematically rigorous fashion to have an agreed upon publishable result. *Specific challenges: Federated semantic discovery:* Interfaces, protocols and environments that support access to, use of, and interoperation across federated sets of resources governed and managed by a mix of different policies and controls that interoperate across streaming and "at rest" data sources. These include: models, algorithms, libraries, and reference implementations for a distributed non-hierarchical discovery service; semantics, methods, interfaces for life-cycle management (subscription, capture, provenance, assessment, validation, rejection) of heterogeneous sets of distributed tools, services and resources; a global environment that is robust in the face of failures and outages; and flexible high-performance data stores (going beyond schema driven) that scale and are friendly to interactive analytics *Resource description and understanding:* Distributed methods and implementations that allow resources (people, software, computing incl. data) to publish varying state and function for use by diverse clients. Mechanisms to handle arbitrary entity types in a uniform and common framework – including complex types such as heterogeneous data, incomplete and evolving information, and rapidly changing availability of computing, storage and other computational resources. Abstract data streaming and file-based data movement over the WAN/LAN and on exascale architectures to allow for real-time, collaborative decision making for scientific processes.	
Big Data Specific Challenges in Mobility	The agility to use any appropriate available resources and to ensure that all data needed is dynamically available at that resource is fundamental to future discoveries in HEP. In this context "resource" has a broad meaning and includes data and people as well as computing and other non-computer based entities: thus, any kind of data—	

Astronomy and Physics> Use Case 39: Analysis of LHC (Large Hadron Collider) Data

	raw data, information, knowledge, etc., and any type of resource—people, computers, storage systems, scientific instruments, software, resource, service, etc. In order to make effective use of such resources, a wide range of management capabilities must be provided in an efficient, secure, and reliable manner, encompassing for example collection, discovery, allocation, movement, access, use, release, and reassignment. These capabilities must span and control large ensembles of data and other resources that are constantly changing and evolving, and will often be in-deterministic and fuzzy in many aspects. *Specific Challenges: Globally optimized dynamic allocation of resources:* These need to take account of the lack of strong consistency in knowledge across the entire system. *Minimization of time-to-delivery of data and services:* Not only to reduce the time to delivery of the data or service but also allow for a predictive capability, so physicists working on data analysis can deal with uncertainties in the real-time decision making processes.
Security and Privacy Requirements	While HEP data itself is not proprietary unintended alteration and/or cyber-security related facility service compromises could potentially be very disruptive to the analysis process. Besides the need of having personal credentials and the related virtual organization credential management systems to maintain access rights to a certain set of resources, a fair amount of attention needs to be devoted to the development and operation of the many software components the community needs to conduct computing in this vastly distributed environment. The majority of software and systems development for LHC data analysis is carried out inside the HEP community or by adopting software components from other parties which involves numerous assumptions and design decisions from the early design stages throughout its life cycle. Software systems make a number of assumptions about their environment - how they are deployed, configured, who runs it, what sort of network is it on, is its input or output sensitive, can it trust its input, does it preserve privacy, etc.? When multiple software components are interconnected, for example in the deep software stacks used in DHTC, without clear understanding of their security assumptions, the security of the resulting system becomes an unknown. A trust framework is a possible way of addressing this problem. A DHTC trust framework, by describing what software, systems and organizations provide and expect of their environment regarding policy enforcement, security and privacy, allows for a system to be analyzed for gaps in trust, fragility and fault tolerance.
Highlight issues for generalizing this use case (e.g. for ref. architecture)	Large scale example of an event based analysis with core statistics needed. Also highlights importance of virtual organizations as seen in global collaboration. The LHC experiments are pioneers of distributed Big Data science infrastructure, and several aspects of the LHC experiments' workflow highlight issues that other disciplines will need to solve. These include automation of data distribution, high performance data transfer, and large-scale high-throughput computing.
More Information (URLs)	http://grids.ucs.indiana.edu/ptliupages/publications/Where%20does%20all%20the%20data%20come%20from%20v7.pdf http://www.es.net/assets/pubs_presos/High-throughput-lessons-from-the-LHC-experience.Johnston.TNC2013.pdf
Note:	

Use Case Stages	Data Sources	Data Usage	Transformations (Data Analytics)	Infrastructure	Security and Privacy
Particle Physics: Analysis of LHC Large Hadron Collider Data, Discovery of Higgs particle (Scientific Research: Physics)					
Record Raw Data	CERN LHC Accelerator	This data is staged at CERN and then distributed across the globe for next stage in processing	LHC has 10^9 collisions per second; the hardware + software trigger selects "interesting events". Other utilities distribute data across the globe with fast transport	Accelerator and sophisticated data selection (trigger process) that uses ≈7000 cores at CERN to record ≈100-500 events each second (≈1 megabyte each)	N/A
Process Raw Data to Information	Disk Files of Raw Data	Iterative calibration and checking of analysis which has for example "heuristic" track finding algorithms. Produce "large" full physics files and stripped down Analysis Object Data (AOD) files that are ≈10% original size	Full analysis code that builds in complete understanding of complex experimental detector. Also Monte Carlo codes to produce simulated data to evaluate efficiency of experimental detection.	≈300,000 cores arranged in 3 tiers. Tier 0: CERN Tier 1: "Major Countries" Tier 2: Universities and laboratories. Note processing is compute and data intensive	N/A
Physics Analysis Information to Knowledge/Discovery	Disk Files of Information including accelerator and Monte Carlo data. Include wisdom from lots of physicists (papers) in analysis choices	Use simple statistical techniques (like histogramming, multi-variate analysis methods and other data analysis techniques and model fits to discover new effects (particles) and put limits on effects not seen	Data reduction and processing steps with advanced physics algorithms to identify event properties, particle hypothesis etc. For interactive data analysis of those reduced and selected datasets the classic program is Root from CERN that reads multiple event (AOD, NTUP) files from selected datasets and use physicist generated C++ code to calculate new quantities such as implied mass of an unstable (new) particle	While the bulk of data processing is done at Tier 1 and Tier 2 resources, the end stage analysis is usually done by users at a local Tier 3 facility. The scale of computing resources at Tier 3 sites range from workstations to small clusters. ROOT is the most common software stack used to analyze compact data formats generated on distributed computing resources. Data transfer is done using ATLAS and CMS DDM tools, which mostly rely on gridFTP middleware. XROOTD based direct data access is also gaining importance wherever high network bandwidth is available.	Physics discoveries and results are confidential until certified by group and presented at meeting/journal. Data preserved so results reproducible

See Figure 6: Particle Physics: Analysis of LHC Data: Discovery of Higgs Particle – CERN LHC location.

See Figure 7: Particle Physics: Analysis of LHC Data: Discovery of Higgs Particle – The multi-tier LHC computing infrastructure.

Astronomy and Physics> Use Case 40: Belle II Experiment

Use Case Title	Belle II Experiment	
Vertical (area)	Scientific Research: High Energy Physics	
Author/Company/Email	David Asner and Malachi Schram, PNNL, david.asner@pnnl.gov and malachi.schram@pnnl.gov	
Actors/Stakeholders and their roles and responsibilities	David Asner is the Chief Scientist for the US Belle II Project Malachi Schram is Belle II network and data transfer coordinator and the PNNL Belle II computing center manager	
Goals	Perform precision measurements to search for new phenomena beyond the Standard Model of Particle Physics	
Use Case Description	Study numerous decay modes at the Upsilon(4S) resonance to search for new phenomena beyond the Standard Model of Particle Physics	
Current Solutions	**Compute(System)**	Distributed (Grid computing using DIRAC)
	Storage	Distributed (various technologies)
	Networking	Continuous RAW data transfer of ≈20Gbps at designed luminosity between Japan and US Additional transfer rates are currently being investigated
	Software	Open Science Grid, Geant4, DIRAC, FTS, Belle II framework
Big Data Characteristics	**Data Source (distributed/centralized)**	Distributed data centers Primary data centers are in Japan (KEK) and US (PNNL)
	Volume (size)	Total integrated RAW data ≈120PB and physics data ≈15PB and ≈100PB MC samples
	Velocity (e.g. real time)	Data will be re-calibrated and analyzed incrementally Data rates will increase based on the accelerator luminosity
	Variety (multiple datasets, mashup)	Data will be re-calibrated and distributed incrementally.
	Variability (rate of change)	Collisions will progressively increase until the designed luminosity is reached (3000 BB pairs per sec). Expected event size is ≈300kB per events.
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues)**	Validation will be performed using known reference physics processes
	Visualization	N/A
	Data Quality	Output data will be re-calibrated and validated incrementally
	Data Types	Tuple based output
	Data Analytics	Data clustering and classification is an integral part of the computing model. Individual scientists define event level analytics.
Big Data Specific Challenges (Gaps)	Data movement and bookkeeping (file and event level meta-data).	
Big Data Specific Challenges in Mobility	Network infrastructure required for continuous data transfer between Japan (KEK) and US (PNNL).	
Security and Privacy Requirements	No special challenges. Data is accessed using grid authentication.	
Highlight issues for generalizing this use case (e.g. for ref. architecture)		

Astronomy and Physics> Use Case 40: Belle II Experiment

More Information (URLs)	http://belle2.kek.jp

Earth, Environmental and Polar Science> Use Case 41: EISCAT 3D Incoherent Scatter Radar System

<table>
<tr><td>Use Case Title</td><td colspan="2">EISCAT 3D incoherent scatter radar system</td></tr>
<tr><td>Vertical (area)</td><td colspan="2">Environmental Science</td></tr>
<tr><td>Author/Company/Email</td><td colspan="2">Yin Chen /Cardiff University/ chenY58@cardiff.ac.uk
Ingemar Häggström, Ingrid Mann, Craig Heinselman/
EISCAT Science Association/{Ingemar.Haggstrom, Ingrid.mann, Craig.Heinselman}@eiscat.se</td></tr>
<tr><td>Actors/Stakeholders and their roles and responsibilities</td><td colspan="2">The EISCAT Scientific Association is an international research organization operating incoherent scatter radar systems in Northern Europe. It is funded and operated by research councils of Norway, Sweden, Finland, Japan, China and the United Kingdom (collectively, the EISCAT Associates). In addition to the incoherent scatter radars, EISCAT also operates an Ionospheric Heater facility, as well as two Dynasondes.</td></tr>
<tr><td>Goals</td><td colspan="2">EISCAT, the European Incoherent Scatter Scientific Association, is established to conduct 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 is also being used as a coherent scatter radar for studying instabilities in the ionosphere, as well as for investigating the structure and dynamics of the middle atmosphere and as a diagnostic instrument in ionospheric modification experiments with the Heating facility.</td></tr>
<tr><td>Use Case Description</td><td colspan="2">The design of the next generation incoherent scatter radar system, EISCAT_3D, opens up opportunities for physicists to explore many new research fields. On the other hand, it also introduces significant challenges in handling large-scale experimental data which will be massively generated at great speeds and volumes. This challenge is typically referred to as a Big Data problem and requires solutions from beyond the capabilities of conventional database technologies.</td></tr>
<tr><td rowspan="4">Current Solutions</td><td>Compute(System)</td><td>EISCAT 3D data e-Infrastructure plans to use the high performance computers for central site data processing and high throughput computers for mirror sites data processing</td></tr>
<tr><td>Storage</td><td>32TB</td></tr>
<tr><td>Networking</td><td>The estimated data rates in local networks at the active site run from 1 GB/s to 10 GB/s. Similar capacity is needed to connect the sites through dedicated high-speed network links. Downloading the full data is not time critical, but operations require real-time information about certain pre-defined events to be sent from the sites to the operation centre and a real-time link from the operation centre to the sites to set the mode of radar operation on with immediate action.</td></tr>
<tr><td>Software</td><td>• Mainstream operating systems, e.g., Windows, Linux, Solaris, HP/UX, or FreeBSD
• Simple, flat file storage with required capabilities e.g., compression, file striping and file journaling
• Self-developed software
o Control and monitoring tools including, system configuration, quick-look, fault reporting, etc.
o Data dissemination utilities
o User software e.g., for cyclic buffer, data cleaning, RFI detection and excision, auto-correlation, data integration, data analysis,</td></tr>
</table>

Earth, Environmental and Polar Science> Use Case 41: EISCAT 3D Incoherent Scatter Radar System

		event identification, discovery and retrieval, calculation of value-added data products, ingestion/extraction, plot o User-oriented computing o APIs into standard software environments o Data processing chains and workflow
Big Data Characteristics	**Data Source (distributed/centralized)**	EISCAT_3D will consist of a core site with a transmitting and receiving radar arrays and four sites with receiving antenna arrays at some 100 km from the core.
	Volume (size)	• The fully operational 5-site system will generate 40 PB/year in 2022. • It is expected to operate for 30 years, and data products to be stored at less 10 years
	Velocity (e.g. real time)	At each of 5-receiver-site: • each antenna generates 30 Msamples/s (120MB/s); • each antenna group (consists of 100 antennas) to form beams at speed of 2 Gbit/s/group; • these data are temporary stored in a ringbuffer: 160 groups ->125 TB/h.
	Variety (multiple datasets, mashup)	• Measurements: different versions, formats, replicas, external sources ... • System information: configuration, monitoring, logs/provenance ... • Users' metadata/data: experiments, analysis, sharing, communications …
	Variability (rate of change)	In time, instantly, a few ms. Along the radar beams, 100ns.
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues)**	• Running 24/7, EISCAT_3D have very high demands on robustness. • Data and performance assurance is vital for the ring-buffer and archive systems. These systems must be able to guarantee to meet minimum data rate acceptance at all times or scientific data will be lost. • Similarly, the systems must guarantee that data held is not volatile or corrupt. This latter requirement is particularly vital at the permanent archive where data is most likely to be accessed by scientific users and least easy to check; data corruption here has a significant possibility of being non-recoverable and of poisoning the scientific literature.
	Visualization	• Real-time visualization of analyzed data, e.g., with a figure of updating panels showing electron density, temperatures and ion velocity to those data for each beam. • Non-real-time (post-experiment) visualization of the physical parameters of interest, e.g., o by standard plots,

Earth, Environmental and Polar Science> Use Case 41: EISCAT 3D Incoherent Scatter Radar System

		o using three-dimensional block to show to spatial variation (in the user selected cuts), o using animations to show the temporal variation, o allow the visualization of 5 or higher dimensional data, e.g., using the 'cut up and stack' technique to reduce the dimensionality, that is take one or more independent coordinates as discrete; or volume rendering technique to display a 2D projection of a 3D discretely sampled dataset. • **(Interactive) Visualization**. E.g., to allow users to combine the information on several spectral features, e.g., by using color coding, and to provide real-time visualization facility to allow the users to link or plug in tailor-made data visualization functions, and more importantly functions to signal for special observational conditions.
	Data Quality	• Monitoring software will be provided which allows The Operator to see incoming data via the Visualization system in real-time and react appropriately to scientifically interesting events. • Control software will be developed to time-integrate the signals and reduce the noise variance and the total data throughput of the system that reached the data archive.
	Data Types	HDF-5
	Data Analytics	Pattern recognition, demanding correlation routines, high level parameter extraction
Big Data Specific Challenges (Gaps)	• High throughput of data for reduction into higher levels. • Discovery of meaningful insights from low-value-density data needs new approaches to the deep, complex analysis e.g., using machine learning, statistical modelling, graph algorithms etc. which go beyond traditional approaches to the space physics.	
Big Data Specific Challenges in Mobility	Is not likely in mobile platforms	
Security and Privacy Requirements	Lower level of data has restrictions for 1 year within the associate countries. All data open after 3 years.	
Highlight issues for generalizing this use case (e.g. for ref. architecture)	EISCAT 3D data e-Infrastructure shares similar architectural characteristics with other ISR radars, and many existing Big Data systems, such as LOFAR, LHC, and SKA	
More Information (URLs)	https://www.eiscat3d.se/	

See Figure 8: EISCAT 3D Incoherent Scatter Radar System – System architecture.

Earth, Environmental and Polar Science> Use Case 42: Common Environmental Research Infrastructure

Use Case Title	ENVRI (Common Operations of Environmental Research Infrastructure)
Vertical (area)	Environmental Science
Author/Company/Email	Yin Chen/ Cardiff University / ChenY58@cardiff.ac.uk
Actors/Stakeholders and their roles and responsibilities	The ENVRI project is a collaboration conducted within the European Strategy Forum on Research Infrastructures (ESFRI) Environmental Cluster. The ESFRI Environmental research infrastructures involved in ENVRI including: • ICOS is a European distributed infrastructure dedicated to the monitoring of greenhouse gases (GHG) through its atmospheric, ecosystem and ocean networks. • EURO-Argo is the European contribution to Argo, which is a global ocean observing system. • EISCAT-3D is a European new-generation incoherent-scatter research radar for upper atmospheric science. • LifeWatch is an e-science Infrastructure for biodiversity and ecosystem research. • EPOS is a European Research Infrastructure on earthquakes, volcanoes, surface dynamics and tectonics. • EMSO is a European network of seafloor observatories for the long-term monitoring of environmental processes related to ecosystems, climate change and geo-hazards. ENVRI also maintains close contact with the other not-directly involved ESFRI Environmental research infrastructures by inviting them for joint meetings. These projects are: • IAGOS Aircraft for global observing system • SIOS Svalbard arctic Earth observing system ENVRI IT community provides common policies and technical solutions for the research infrastructures, which involves a number of organization partners including, Cardiff University, CNR-ISTI, CNRS (Centre National de la Recherche Scientifique), CSC, EAA (Umweltbundesamt Gmbh), EGI, ESA-ESRIN, University of Amsterdam, and University of Edinburgh.
Goals	The ENVRI project gathers 6 EU ESFRI environmental science infra-structures (ICOS, EURO-Argo, EISCAT-3D, LifeWatch, EPOS, and EMSO) in order to develop common data and software services. The results will accelerate the construction of these infrastructures and improve interoperability among them. The primary goal of ENVRI is to agree on a reference model for joint operations. The ENVRI RM is a common ontological framework and standard for the description and characterisation of computational and storage infrastructures in order 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, also serving to identify common solutions to common problems. This may enable reuse, share of resources and experiences, and avoid duplication of efforts.
Use Case Description	ENVRI project implements harmonized solutions and draws up guidelines for the common needs of the environmental ESFRI projects, with a special focus on issues as architectures, metadata frameworks, data discovery in scattered repositories, visualization and data curation. This 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. ENVRI investigates a collection of representative research infrastructures for

Earth, Environmental and Polar Science> Use Case 42: Common Environmental Research Infrastructure

	environmental sciences, and provides a projection of Europe-wide requirements they have; identifying in particular, requirements they have in common. Based on the analysis evidence, the ENVRI Reference Model (http://www.envri.eu/rm) is developed using ISO standard Open Distributed Processing. Fundamentally the model serves to provide a universal reference framework for discussing many common technical challenges facing all of the ESFRI-environmental research infrastructures. By drawing analogies between the reference components of the model and the actual elements of the infrastructures (or their proposed designs) as they exist now, various gaps and points of overlap can be identified.	
Current Solutions	**Compute(System)**	
	Storage	File systems and relational databases
	Networking	
	Software	Own
Big Data Characteristics	**Data Source (distributed/centralized)**	Most of the ENVRI Research Infrastructures (ENV RIs) are *distributed, long-term, remote controlled observational networks* focused on understanding processes, trends, thresholds, interactions and feedbacks and increasing the predictive power to address future environmental challenges. They are spanning from the Arctic areas to the European Southernmost areas and from Atlantic on west to the Black Sea on east. More precisely: • ***EMSO,*** network of fixed-point, deep-seafloor and water column observatories, is geographically distributed in key sites of European waters, presently consisting of thirteen sites. • ***EPOS*** aims at integrating the existing European facilities in solid Earth science into one coherent multidisciplinary RI, and to increase the accessibility and usability of multidisciplinary data from seismic and geodetic monitoring networks, volcano observatories, laboratory experiments and computational simulations enhancing worldwide interoperability in Earth Science. • ***ICOS*** dedicates to the monitoring of greenhouse gases (GHG) through its atmospheric, ecosystem and ocean networks. The ICOS network includes more than 30 atmospheric and more than 30 ecosystem primary long term sites located across Europe, and additional secondary sites. It also includes three Thematic Centres to process the data from all the stations from each network, and provide access to these data. • ***LifeWatch*** is a "virtual" infrastructure for biodiversity and ecosystem research with services mainly provided through the Internet. Its Common Facilities is coordinated and managed at a central European level; and the *LifeWatch Centres* serve as specialized facilities from member countries

Earth, Environmental and Polar Science> Use Case 42: Common Environmental Research Infrastructure

		(regional partner facilities) or research communities. • ***Euro-Argo*** provides, deploys and operates an array of around 800 floats contributing to the global array (3,000 floats) and thus provide enhanced coverage in the European regional seas. • ***EISCAT- 3D***, makes continuous measurements of the geospace environment and its coupling to the Earth's atmosphere from its location in the auroral zone at the southern edge of the northern polar vortex, and is a distributed infrastructure.
	Volume (size)	Variable data size. e.g., • The amount of data within the ***EMSO*** is depending on the instrumentation and configuration of the observatory between several MBs to several GB per dataset. • Within ***EPOS***, the EIDA network is currently providing access to continuous raw data coming from approximately more than 1000 stations recording about 40GB per day, so over 15 TB per year. EMSC stores a Database of 1.85 GB of earthquake parameters, which is constantly growing and updated with refined information. - 222705 – events - 632327 – origins - 642555 – magnitudes • Within ***EISCAT 3D*** raw voltage data will reach 40PB/year in 2023.
	Velocity (e.g. real time)	Real-time data handling is a common request of the environmental research infrastructures
	Variety (multiple datasets, mashup)	Highly complex and heterogeneous
	Variability (rate of change)	Relative low rate of change
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues, semantics)**	Normal
	Visualization	Most of the projects have not yet developed the visualization technique to be fully operational. • ***EMSO*** is not yet fully operational, currently only simple graph plotting tools. • Visualization techniques are not yet defined for ***EPOS***. • Within ***ICOS*** Level-1.b data products such as near real time GHG measurements are available to users via ATC web portal. Based on Google Chart Tools, an interactive time series line chart with optional annotations allows user to scroll and zoom inside a time series of CO2 or CH4 measurement at an ICOS

Earth, Environmental and Polar Science> Use Case 42: Common Environmental Research Infrastructure

		Atmospheric station. The chart is rendered within the browser using Flash. Some Level-2 products are also available to ensure instrument monitoring to PIs. It is mainly instrumental and comparison data plots automatically generated (R language and Python Matplotlib 2D plotting library) and daily pushed on ICOS web server. Level-3 data products such as gridded GHG fluxes derived from ICOS observations increase the scientific impact of ICOS. For this purpose ICOS supports its community of users. The Carbon portal is expected to act as a platform that will offer visualization of the flux products that incorporate ICOS data. Example of candidate Level-3 products from future ICOS GHG concentration data are for instance maps of European high-resolution CO2 or CH4 fluxes obtained by atmospheric inversion modellers in Europe. Visual tools for comparisons between products will be developed by the Carbon Portal. Contributions will be open to any product of high scientific quality. • ***LifeWatch*** will provide common visualization techniques, such as the plotting of species on maps. New techniques will allow visualizing the effect of changing data and/or parameters in models.
	Data Quality (syntax)	Highly important
	Data Types	• Measurements (often in file formats), • Metadata, • Ontology, • Annotations
	Data Analytics	• Data assimilation, • (Statistical) analysis, • Data mining, • Data extraction, • Scientific modeling and simulation, • Scientific workflow
Big Data Specific Challenges (Gaps)	• Real-time handling of extreme high volume of data • Data staging to mirror archives • Integrated Data access and discovery • Data processing and analysis	
Big Data Specific Challenges in Mobility	The need for efficient and high performance mobile detectors and instrumentation is common: • In **ICOS**, various mobile instruments are used to collect data from marine observations, atmospheric observations, and ecosystem monitoring. • In **Euro-Argo**, thousands of submersible robots to obtain observations of all of the oceans • In **Lifewatch**, biologists use mobile instruments for observations and measurements.	
Security and Privacy	Most of the projects follow the open data sharing policy. E.g.,	

Earth, Environmental and Polar Science> Use Case 42: Common Environmental Research Infrastructure

Requirements	• The vision of **EMSO** is to allow scientists all over the world to access observatories data following an open access model. • Within **EPOS**, EIDA data and Earthquake parameters are generally open and free to use. Few restrictions are applied on few seismic networks and the access is regulated depending on email based authentication/authorization. • The **ICOS** data will be accessible through a license with full and open access. No particular restriction in the access and eventual use of the data is anticipated, expected the inability to redistribute the data. Acknowledgement of ICOS and traceability of the data will be sought in a specific, way (e.g. DOI of dataset). A large part of relevant data and resources are generated using public funding from national and international sources. • **LifeWatch** is following the appropriate European policies, such as: the European Research Council (ERC) requirement; the European Commission's open access pilot mandate in 2008. For publications, initiatives such as Dryad instigated by publishers and the Open Access Infrastructure for Research in Europe (OpenAIRE). The private sector may deploy their data in the LifeWatch infrastructure. A special company will be established to manage such commercial contracts. • In **EISCAT 3D**, lower level of data has restrictions for 1 year within the associate countries. All data open after 3 years.
Highlight issues for generalizing this use case (e.g. for ref. architecture)	Different research infrastructures are designed for different purposes and evolve over time. The designers describe their approaches from different points of view, in different levels of detail and using different typologies. The documentation provided is often incomplete and inconsistent. What is needed is a uniform platform for interpretation and discussion, which helps to unify understanding. In ENVRI, we choose to use a standard model, Open Distributed Processing (ODP), to interpret the design of the research infrastructures, and place their requirements into the ODP framework for further analysis and comparison.
More Information (URLs)	• ENVRI Project website: http://www.envri.eu • ENVRI Reference Model http://www.envri.eu/rm • ENVRI deliverable D3.2: Analysis of common requirements of Environmental Research Infrastructures • ICOS: http://www.icos-infrastructure.eu/ • Euro-Argo: http://www.euro-argo.eu/ • EISCAT 3D: http://www.eiscat3d.se/ • LifeWatch: http://www.lifewatch.com/ • EPOS: http://www.epos-eu.org/ • EMSO http://www.emso-eu.org/management/

See Figure 9: ENVRI, Common Operations of Environmental Research Infrastructure – ENVRI common architecture.

See Figure 10(a): ICOS architecture

See Figure 10(b): LifeWatch architecture

See Figure 10(c): EMSO architecture

See Figure 10(d): EURO-Argo architecture

See Figure 10(e): EISCAT 3D architecture

Earth, Environmental and Polar Science> Use Case 43: Radar Data Analysis for CReSIS

Use Case Title	Radar Data Analysis for CReSIS	
Vertical (area)	Scientific Research: Polar Science and Remote Sensing of Ice Sheets	
Author/Company/Email	Geoffrey Fox, Indiana University gcf@indiana.edu	
Actors/Stakeholders and their roles and responsibilities	Research funded by NSF and NASA with relevance to near and long term climate change. Engineers designing novel radar with "field expeditions" for 1-2 months to remote sites. Results used by scientists building models and theories involving Ice Sheets	
Goals	Determine the depths of glaciers and snow layers to be fed into higher level scientific analyses	
Use Case Description	Build radar; build UAV or use piloted aircraft; overfly remote sites (Arctic, Antarctic, Himalayas). Check in field that experiments configured correctly with detailed analysis later. Transport data by air-shipping disk as poor Internet connection. Use image processing to find ice/snow sheet depths. Use depths in scientific discovery of melting ice caps etc.	
Current Solutions	**Compute(System)**	Field is a low power cluster of rugged laptops plus classic 2-4 CPU servers with ≈40 TB removable disk array. Off line is about 2500 cores
	Storage	Removable disk in field. (Disks suffer in field so 2 copies made) Lustre or equivalent for offline
	Networking	Terrible Internet linking field sites to continental USA.
	Software	Radar signal processing in Matlab. Image analysis is Map/Reduce or MPI plus C/Java. User Interface is a Geographical Information System
Big Data Characteristics	**Data Source (distributed/centralized)**	Aircraft flying over ice sheets in carefully planned paths with data downloaded to disks.
	Volume (size)	≈0.5 Petabytes per year raw data
	Velocity (e.g. real time)	All data gathered in real time but analyzed incrementally and stored with a GIS interface
	Variety (multiple datasets, mashup)	Lots of different datasets – each needing custom signal processing but all similar in structure. This data needs to be used with wide variety of other polar data.
	Variability (rate of change)	Data accumulated in ≈100 TB chunks for each expedition
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues)**	Essential to monitor field data and correct instrumental problems. Implies must analyze fully portion of data in field
	Visualization	Rich user interface for layers and glacier simulations
	Data Quality	Main engineering issue is to ensure instrument gives quality data
	Data Types	Radar Images
	Data Analytics	Sophisticated signal processing; novel new image processing to find layers (can be 100's one per year)
Big Data Specific Challenges (Gaps)	Data volumes increasing. Shipping disks clumsy but no other obvious solution. Image processing algorithms still very active research	
Big Data Specific Challenges in Mobility	Smart phone interfaces not essential but LOW power technology essential in field	
Security and Privacy Requirements	Himalaya studies fraught with political issues and require UAV. Data itself open after initial study	

Earth, Environmental and Polar Science> Use Case 43: Radar Data Analysis for CReSIS

Highlight issues for generalizing this use case (e.g. for ref. architecture)	Loosely coupled clusters for signal processing. Must support Matlab.
More Information (URLs)	http://polargrid.org/polargrid https://www.cresis.ku.edu/ See movie at http://polargrid.org/polargrid/gallery
Note:	

Use Case Stages	Data Sources	Data Usage	Transformations (Data Analytics)	Infrastructure	Security and Privacy
Radar Data Analysis for CReSIS (Scientific Research: Polar Science and Remote Sensing of Ice Sheets)					
Raw Data: Field Trip	Raw Data from Radar instrument on Plane/Vehicle	Capture Data on Disks for L1B. Check Data to monitor instruments.	Robust Data Copying Utilities. Version of Full Analysis to check data.	Rugged Laptops with small server (≈2 CPU with ≈40TB removable disk system)	N/A
Information: Offline Analysis L1B	Transported Disks copied to (LUSTRE) File System	Produce processed data as radar images	Matlab Analysis code running in parallel and independently on each data sample	≈2500 cores running standard cluster tools	N/A except results checked before release on CReSIS web site
Information: L2/L3 Geolocation and Layer Finding	Radar Images from L1B	Input to Science as database with GIS frontend	GIS and Metadata Tools Environment to support automatic and/or manual layer determination	GIS (Geographical Information System). Cluster for Image Processing.	As above
Knowledge, Wisdom, Discovery: Science	GIS interface to L2/L3 data	Polar Science Research integrating multiple data sources e.g. for Climate change. Glacier bed data used in simulations of glacier flow		Exploration on a cloud style GIS supporting access to data. Simulation is 3D partial differential equation solver on large cluster.	Varies according to science use. Typically results open after research complete.

See Figure 11: Radar Data Analysis for CReSIS Remote Sensing of Ice Sheets– Typical CReSIS radar data after analysis.

See Figure 12: Radar Data Analysis for CReSIS Remote Sensing of Ice Sheets– Typical flight paths of data gathering in survey region.

See Figure 13: Radar Data Analysis for CReSIS Remote Sensing of Ice Sheets – Typical echogram with detected boundaries. The upper (green) boundary is between air and ice layers, while the lower (red) boundary is between ice and terrain.

Earth, Environmental and Polar Science > Use Case 44: UAVSAR Data Processing

Use Case Title	UAVSAR Data Processing, Data Product Delivery, and Data Services	
Vertical (area)	Scientific Research: Earth Science	
Author/Company/Email	Andrea Donnellan, NASA JPL, andrea.donnellan@jpl.nasa.gov; Jay Parker, NASA JPL, jay.w.parker@jpl.nasa.gov	
Actors/Stakeholders and their roles and responsibilities	NASA UAVSAR team, NASA QuakeSim team, ASF (NASA SAR DAAC), USGS, CA Geological Survey	
Goals	Use of Synthetic Aperture Radar (SAR) to identify landscape changes caused by seismic activity, landslides, deforestation, vegetation changes, flooding, etc.; increase its usability and accessibility by scientists.	
Use Case Description	A scientist who wants to study the after effects of an earthquake examines multiple standard SAR products made available by NASA. The scientist may find it useful to interact with services provided by intermediate projects that add value to the official data product archive.	
Current Solutions	**Compute(System)**	Raw data processing at NASA AMES Pleiades, Endeavour. Commercial clouds for storage and service front ends have been explored.
	Storage	File based.
	Networking	Data require one time transfers between instrument and JPL, JPL and other NASA computing centers (AMES), and JPL and ASF. Individual data files are not too large for individual users to download, but entire dataset is unwieldy to transfer. This is a problem to downstream groups like QuakeSim who want to reformat and add value to datasets.
	Software	ROI_PAC, GeoServer, GDAL, GeoTIFF-supporting tools.
Big Data Characteristics	**Data Source (distributed/centralized)**	Data initially acquired by unmanned aircraft. Initially processed at NASA JPL. Archive is centralized at ASF (NASA DAAC). QuakeSim team maintains separate downstream products (GeoTIFF conversions).
	Volume (size)	Repeat Pass Interferometry (RPI) Data: ≈ 3 TB. Increasing about 1-2 TB/year. Polarimetric Data: ≈40 TB (processed) Raw Data: 110 TB Proposed satellite missions (Earth Radar Mission, formerly DESDynI) could dramatically increase data volumes (TBs per day).
	Velocity (e.g. real time)	RPI Data: 1-2 TB/year. Polarimetric data is faster.
	Variety (multiple datasets, mashup)	Two main types: Polarimetric and RPI. Each RPI product is a collection of files (annotation file, unwrapped, etc.). Polarimetric products also consist of several files each.
	Variability (rate of change)	Data products change slowly. Data occasionally get reprocessed: new processing methods or parameters. There may be additional quality assurance and quality control issues.

Earth, Environmental and Polar Science > Use Case 44: UAVSAR Data Processing

Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues, semantics)**	Provenance issues need to be considered. This provenance has not been transparent to downstream consumers in the past. Versioning used now; versions described in the UAVSAR web page in notes.
	Visualization	Uses Geospatial Information System tools, services, standards.
	Data Quality (syntax)	Many frames and collections are found to be unusable due to unforeseen flight conditions.
	Data Types	GeoTIFF and related imagery data
	Data Analytics	Done by downstream consumers (such as edge detections): research issues.
Big Data Specific Challenges (Gaps)	Data processing pipeline requires human inspection and intervention. Limited downstream data pipelines for custom users. Cloud architectures for distributing entire data product collections to downstream consumers should be investigated, adopted.	
Big Data Specific Challenges in Mobility	Some users examine data in the field on mobile devices, requiring interactive reduction of large datasets to understandable images or statistics.	
Security and Privacy Requirements	Data is made immediately public after processing (no embargo period).	
Highlight issues for generalizing this use case (e.g. for ref. architecture)	Data is geolocated, and may be angularly specified. Categories: GIS; standard instrument data processing pipeline to produce standard data products.	
More Information (URLs)	http://uavsar.jpl.nasa.gov/, http://www.asf.alaska.edu/program/sdc, http://quakesim.org	

See Figure 14: UAVSAR Data Processing, Data Product Delivery, and Data Services – Combined unwrapped coseismic interferograms for flight lines 26501, 26505, and 08508 for the October 2009–April 2010 time period. End points where slip can be seen on the Imperial, Superstition Hills, and Elmore Ranch faults are noted. GPS stations are marked by dots and are labeled.

Earth, Environmental and Polar Science> Use Case 45: NASA LARC/GSFC iRODS Federation Testbed

Use Case Title	NASA LARC/GSFC iRODS Federation Testbed	
Vertical (area)	Earth Science Research and Applications	
Author/Company/Email	Michael Little, Roger Dubois, Brandi Quam, Tiffany Mathews, Andrei Vakhnin, Beth Huffer, Christian Johnson / NASA Langley Research Center (LaRC) / M.M.Little@NASA.gov, Roger.A.Dubois@nasa.gov, Brandi.M.Quam@NASA.gov, Tiffany.J.Mathews@NASA.gov, and Andrei.A.Vakhnin@NASA.gov John Schnase, Daniel Duffy, Glenn Tamkin, Scott Sinno, John Thompson, and Mark McInerney / NASA Goddard Space Flight Center (GSFC) / John.L.Schnase@NASA.gov, Daniel.Q.Duffy@NASA.gov, Glenn.S.Tamkin@nasa.gov. Scott.S.Sinno@nasa.gov, John.H.Thompson@nasa.gov, and Mark.Mcinerney@nasa.gov	
Actors/Stakeholders and their roles and responsibilities	NASA's Atmospheric Science Data Center (ASDC) at Langley Research Center (LaRC) in Hampton, Virginia, and the Center for Climate Simulation (NCCS) at Goddard Space Flight Center (GSFC) both ingest, archive, and distribute data that is essential to stakeholders including the climate research community, science applications community, and a growing community of government and private-sector customers who have a need for atmospheric and climatic data.	
Goals	To implement a data federation ability to improve and automate the discovery of heterogeneous data, decrease data transfer latency, and meet customizable criteria based on data content, data quality, metadata, and production. To support/enable applications and customers that require the integration of multiple heterogeneous data collections.	
Use Case Description	ASDC and NCCS have complementary datasets, each containing vast amounts of data that is 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, etc. To analyze, visualize and otherwise process data from heterogeneous datasets is currently a time consuming effort that requires scientists to separately access, search for, and download data from multiple servers and often the data is duplicated without an understanding of the authoritative source. Many scientists report spending more time in accessing data than in conducting research. Data consumers need mechanisms for retrieving heterogeneous data from a single point-of-access. This can be enabled through the use of iRODS, a Data grid software system that enables parallel downloads of datasets from selected replica servers that can be geographically dispersed, but still accessible by users worldwide. Using iRODS in conjunction with semantically enhanced metadata, managed via a highly precise Earth Science ontology, the ASDC's Data Products Online (DPO) will be federated with the data at the NASA Center for Climate Simulation (NCCS) at Goddard Space Flight Center (GSFC). The heterogeneous data products at these two NASA facilities are being semantically annotated using common concepts from the NASA Earth Science ontology. The semantic annotations will enable the iRODS system to identify complementary datasets and aggregate data from these disparate sources, facilitating data sharing between climate modelers, forecasters, Earth scientists, and scientists from other disciplines that need Earth science data. The iRODS data federation system will also support cloud-based data processing services in the Amazon Web Services (AWS) cloud.	
Current	**Compute (System)**	NASA Center for Climate Simulation (NCCS) and

Earth, Environmental and Polar Science> Use Case 45: NASA LARC/GSFC iRODS Federation Testbed

Solutions		NASA Atmospheric Science Data Center (ASDC): Two GPFS systems
	Storage	The ASDC's Data Products Online (DPO) GPFS File system consists of 12 x IBM DC4800 and 6 x IBM DCS3700 Storage subsystems, 144 Intel 2.4 GHz cores, 1,400 TB usable storage. NCCS data is stored in the NCCS MERRA cluster, which is a 36 node Dell cluster, 576 Intel 2.6 GHz SandyBridge cores, 1,300 TB raw storage, 1,250 GB RAM, 11.7 TF theoretical peak compute capacity.
	Networking	A combination of Fibre Channel SAN and 10GB LAN. The NCCS cluster nodes are connected by an FDR Infiniband network with peak TCP/IP speeds >20 Gbps.
	Software	SGE Univa Grid Engine Version 8.1, iRODS version 3.2 and/or 3.3, IBM General Parallel File System (GPFS) version 3.4, Cloudera version 4.5.2-1.
Big Data Characteristics	**Data Source (distributed/centralized)**	iRODS will be leveraged to share data collected from CERES Level 3B data products including: CERES EBAF-TOA and CERES-Surface products. Surface fluxes in EBAF-Surface are derived from two CERES data products: 1) CERES SYN1deg-Month Ed3 - which provides computed surface fluxes to be adjusted and 2) CERES EBAFTOA Ed2.7 – which uses observations to provide CERES-derived TOA flux constraints. Access to these products will enable the NCCS at GSFC to run data from the products in a simulation model in order to produce an assimilated flux. The NCCS will introduce Modern-Era Retrospective Analysis for Research and Applications (MERRA) data to the iRODS federation. MERRA integrates observational data with numerical models to produce a global temporally and spatially consistent synthesis of 26 key climate variables. MERRA data files are created from the Goddard Earth Observing System version 5 (GEOS-5) model and are stored in HDF-EOS and (Network Common Data Form) NetCDF formats. Spatial resolution is 1/2° latitude × 2/3° longitude × 72 vertical levels extending through the stratosphere. Temporal resolution is 6-hours for three-dimensional, full spatial resolution, extending from 1979-present, nearly the entire satellite era. Each file contains a single grid with multiple 2D and 3D variables. All data are stored on a longitude-latitude grid with a vertical dimension applicable for all 3D variables. The GEOS-5 MERRA products are divided into 25 collections: 18 standard products, chemistry products. The collections comprise monthly means files and daily files at six-hour intervals running from 1979 – 2012. MERRA data are typically packaged as multi-

Earth, Environmental and Polar Science> Use Case 45: NASA LARC/GSFC iRODS Federation Testbed

		dimensional binary data within a self-describing NetCDF file format. Hierarchical metadata in the NetCDF header contain the representation information that allows NetCDF- aware software to work with the data. It also contains arbitrary preservation description and policy information that can be used to bring the data into use-specific compliance.
	Volume (size)	Currently, Data from the EBAF-TOA Product is about 420MB and Data from the EBAF-Surface Product is about 690MB. Data grows with each version update (about every six months). The MERRA collection represents about 160 TB of total data (uncompressed); compressed is ≈80 TB.
	Velocity (e.g. real time)	Periodic since updates are performed with each new version update.
	Variety (multiple datasets, mashup)	There is a need in many types of applications to combine MERRA reanalysis data with other reanalyses and observational data such as CERES. The NCCS is using the Climate Model Intercomparison Project (CMIP5) Reference standard for ontological alignment across multiple, disparate datasets.
	Variability (rate of change)	The MERRA reanalysis grows by approximately one TB per month.
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues)**	Validation and testing of semantic metadata, and of federated data products will be provided by data producers at NASA Langley Research Center and at Goddard through regular testing. Regression testing will be implemented to ensure that updates and changes to the iRODS system, newly added data sources, or newly added metadata do not introduce errors to federated data products. MERRA validation is provided by the data producers, NASA Goddard's Global Modeling and Assimilation Office (GMAO).
	Visualization	There is a growing need in the scientific community for data management and visualization services that can aggregate data from multiple sources and display it in a single graphical display. Currently, such capabilities are hindered by the challenge of finding and downloading comparable data from multiple servers, and then transforming each heterogeneous dataset to make it usable by the visualization software. Federation of NASA datasets using iRODS will enable scientists to quickly find and aggregate comparable datasets for use with visualization software.
	Data Quality	For MERRA, quality controls are applied by the data producers, GMAO.
	Data Types	See above.
	Data Analytics	Pursuant to the first goal of increasing accessibility and discoverability through innovative technologies, the

Earth, Environmental and Polar Science> Use Case 45: NASA LARC/GSFC iRODS Federation Testbed

		ASDC and NCCS are exploring a capability to improve data access capabilities. Using iRODS, the ASDC's Data Products Online (DPO) can be federated with data at GSFC's NCCS creating a data access system that can serve a much broader customer base than is currently being served. Federating and sharing information will enable the ASDC and NCCS to fully utilize multi-year and multi-instrument data and will improve and automate the discovery of heterogeneous data, increase data transfer latency, and meet customizable criteria based on data content, data quality, metadata, and production.
Big Data Specific Challenges (Gaps)		
Big Data Specific Challenges in Mobility	A major challenge includes defining an enterprise architecture that can deliver real-time analytics via communication with multiple APIs and cloud computing systems. By keeping the computation resources on cloud systems, the challenge with mobility resides in not overpowering mobile devices with displaying CPU intensive visualizations that may hinder the performance or usability of the data being presented to the user.	
Security and Privacy Requirements		
Highlight issues for generalizing this use case (e.g. for ref. architecture)	This federation builds on several years of iRODS research and development performed at the NCCS. During this time, the NCCS vetted the iRODS features while extending its core functions with domain-specific extensions. For example, the NCCS created and installed Python-based scientific kits within iRODS that automatically harvest metadata when the associated data collection is registered. One of these scientific kits was developed for the MERRA collection. This kit in conjunction with iRODS bolsters the strength of the LaRC/GSFC federation by providing advanced search capabilities. LaRC is working through the establishment of an advanced architecture that leverages multiple technology pilots and tools (access, discovery, and analysis) designed to integrate capabilities across the earth science community – the research and development completed by both data centers is complementary and only further enhances this use case. Other scientific kits that have been developed include: NetCDF, Intergovernmental Panel on Climate Change (IPCC), and Ocean Modeling and Data Assimilation (ODAS). The combination of iRODS and these scientific kits has culminated in a configurable technology stack called the virtual Climate Data Server (vCDS), meaning that this runtime environment can be deployed to multiple destinations (e.g., bare metal, virtual servers, cloud) to support various scientific needs. The vCDS, which can be viewed as a reference architecture for easing the federation of disparate data repositories, is leveraged by but not limited to LaRC and GSFC.	
More Information (URLs)	Please contact the authors for additional information.	

Earth, Environmental and Polar Science> Use Case 46: MERRA Analytic Services

Use Case Title	MERRA Analytic Services (MERRA/AS)	
Vertical (area)	Scientific Research: Earth Science	
Author/Company/Email	John L. Schnase and Daniel Q. Duffy / NASA Goddard Space Flight Center John.L.Schnase@NASA.gov, Daniel.Q.Duffy@NASA.gov	
Actors/Stakeholders and their roles and responsibilities	NASA's Modern-Era Retrospective Analysis for Research and Applications (MERRA) integrates observational data with numerical models to produce a global temporally and spatially consistent synthesis of 26 key climate variables. Actors and stakeholders who have an interest in MERRA include the climate research community, science applications community, and a growing number of government and private-sector customers who have a need for the MERRA data in their decision support systems.	
Goals	Increase the usability and use of large-scale scientific data collections, such as MERRA.	
Use Case Description	MERRA Analytic Services enables Map/Reduce analytics over the MERRA collection. MERRA/AS is an example of cloud-enabled climate analytics as a service (CAaaS), which is an approach to meeting the Big Data challenges of climate science through the combined use of 1) high performance, data proximal analytics, (2) scalable data management, (3) software appliance virtualization, (4) adaptive analytics, and (5) a domain-harmonized API. The effectiveness of MERRA/AS is being demonstrated in several applications, including data publication to the Earth System Grid Federation (ESGF) in support of Intergovernmental Panel on Climate Change (IPCC) research, the NASA/Department of Interior RECOVER wild land fire decision support system, and data interoperability testbed evaluations between NASA Goddard Space Flight Center and the NASA Langley Atmospheric Data Center.	
Current Solutions	**Compute(System)**	NASA Center for Climate Simulation (NCCS)
	Storage	The MERRA Analytic Services Hadoop Filesystem (HDFS) is a 36 node Dell cluster, 576 Intel 2.6 GHz SandyBridge cores, 1300 TB raw storage, 1250 GB RAM, 11.7 TF theoretical peak compute capacity.
	Networking	Cluster nodes are connected by an FDR Infiniband network with peak TCP/IP speeds >20 Gbps.
	Software	Cloudera, iRODS, Amazon AWS
Big Data Characteristics	**Data Source (distributed/centralized)**	MERRA data files are created from the Goddard Earth Observing System version 5 (GEOS-5) model and are stored in HDF-EOS and NetCDF formats. Spatial resolution is 1/2 °latitude ×2/3 °longitude × 72 vertical levels extending through the stratosphere. Temporal resolution is 6-hours for three-dimensional, full spatial resolution, extending from 1979-present, nearly the entire satellite era. Each file contains a single grid with multiple 2D and 3D variables. All data are stored on a longitude latitude grid with a vertical dimension applicable for all 3D variables. The GEOS-5 MERRA products are divided into 25 collections: 18 standard products, 7 chemistry products. The collections comprise monthly means files and daily files at six-hour intervals running from 1979–2012. MERRA data are typically packaged as multi-dimensional binary data within a self-describing NetCDF file format. Hierarchical

Earth, Environmental and Polar Science> Use Case 46: MERRA Analytic Services

		metadata in the NetCDF header contain the representation information that allows NetCDF aware software to work with the data. It also contains arbitrary preservation description and policy information that can be used to bring the data into use-specific compliance.
	Volume (size)	480TB
	Velocity (e.g. real time)	Real-time or batch, depending on the analysis. We're developing a set of "canonical ops" -early stage, near-data operations common to many analytic workflows. The goal is for the canonical ops to run in near real-time.
	Variety (multiple datasets, mashup)	There is a need in many types of applications to combine MERRA reanalysis data with other re-analyses and observational data. We are using the Climate Model Inter-comparison Project (CMIP5) Reference standard for ontological alignment across multiple, disparate datasets.
	Variability (rate of change)	The MERRA reanalysis grows by approximately one TB per month.
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues, semantics)**	Validation provided by data producers, NASA Goddard's Global Modeling and Assimilation Office (GMAO).
	Visualization	There is a growing need for distributed visualization of analytic outputs.
	Data Quality (syntax)	Quality controls applied by data producers, GMAO.
	Data Types	See above.
	Data Analytics	In our efforts to address the Big Data challenges of climate science, we are moving toward a notion of climate analytics-as-a-service. We focus on analytics, because it is the knowledge gained from our interactions with Big Data that ultimately produce societal benefits. We focus on CAaaS because we believe it provides a useful way of thinking about the problem: a specialization of the concept of business process-as-a-service, which is an evolving extension of IaaS, PaaS, and SaaS enabled by Cloud Computing.
Big Data Specific Challenges (Gaps)	A big question is how to use cloud computing to enable better use of climate science's earthbound compute and data resources. Cloud Computing is providing for us a new tier in the data services stack —a cloud-based layer where agile customization occurs and enterprise-level products are transformed to meet the specialized requirements of applications and consumers. It helps us close the gap between the world of traditional, high-performance computing, which, at least for now, resides in a finely-tuned climate modeling environment at the enterprise level and our new customers, whose expectations and manner of work are increasingly influenced by the smart mobility megatrend.	
Big Data Specific Challenges in Mobility	Most modern smartphones, tablets, etc. actually consist of just the display and user interface components of sophisticated applications that run in cloud data centers. This is a mode of work that CAaaS is intended to accommodate.	
Security and Privacy Requirements	No critical issues identified at this time.	
Highlight issues for	Map/Reduce and iRODS fundamentally make analytics and data aggregation easier;	

Earth, Environmental and Polar Science> Use Case 46: MERRA Analytic Services

generalizing this use case (e.g. for ref. architecture)	our approach to software appliance virtualization in makes it easier to transfer capabilities to new users and simplifies their ability to build new applications; the social construction of extended capabilities facilitated by the notion of canonical operations enable adaptability; and the Climate Data Services API that we're developing enables ease of mastery. Taken together, we believe that these core technologies behind CAaaS creates a generative context where inputs from diverse people and groups, who may or may not be working in concert, can contribute capabilities that help address the Big Data challenges of climate science.
More Information (URLs)	Please contact the authors for additional information.

See Figure 15: MERRA Analytic Services MERRA/AS – Typical MERRA/AS output.

Earth, Environmental and Polar Science> Use Case 47: Atmospheric Turbulence—Event Discovery

Use Case Title	Atmospheric Turbulence - Event Discovery and Predictive Analytics	
Vertical (area)	Scientific Research: Earth Science	
Author/Company/Email	Michael Seablom, NASA Headquarters, michael.s.seablom@nasa.gov	
Actors/Stakeholders and their roles and responsibilities	Researchers with NASA or NSF grants, weather forecasters, aviation interests (for the generalized case, any researcher who has a role in studying phenomena-based events).	
Goals	Enable the discovery of high-impact phenomena contained within voluminous Earth Science data stores and which are difficult to characterize using traditional numerical methods (e.g., turbulence). Correlate such phenomena with global atmospheric re-analysis products to enhance predictive capabilities.	
Use Case Description	Correlate aircraft reports of turbulence (either from pilot reports or from automated aircraft measurements of eddy dissipation rates) with recently completed atmospheric re-analyses of the entire satellite-observing era. Reanalysis products include the North American Regional Reanalysis (NARR) and the Modern-Era Retrospective-Analysis for Research (MERRA) from NASA.	
Current Solutions	**Compute(System)**	NASA Earth Exchange (NEX) - Pleiades supercomputer.
	Storage	Re-analysis products are on the order of 100TB each; turbulence data are negligible in size.
	Networking	Re-analysis datasets are likely to be too large to relocate to the supercomputer of choice (in this case NEX), therefore the fastest networking possible would be needed.
	Software	Map/Reduce or the like; SciDB or other scientific database.
Big Data Characteristics	**Data Source (distributed/centralized)**	Distributed
	Volume (size)	200TB (current), 500TB within 5 years
	Velocity (e.g. real time)	Data analyzed incrementally
	Variety (multiple datasets, mashup)	Re-analysis datasets are inconsistent in format, resolution, semantics, and metadata. Likely each of these input streams will have to be interpreted/analyzed into a common product.
	Variability (rate of change)	Turbulence observations would be updated continuously; re-analysis products are released about once every five years.
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues)**	Validation would be necessary for the output product (correlations).
	Visualization	Useful for interpretation of results.
	Data Quality	Input streams would have already been subject to quality control.
	Data Types	Gridded output from atmospheric data assimilation systems and textual data from turbulence observations.
	Data Analytics	Event-specification language needed to perform data mining / event searches.
Big Data Specific Challenges (Gaps)	Semantics (interpretation of multiple reanalysis products); data movement; database(s) with optimal structuring for 4-dimensional data mining.	

Earth, Environmental and Polar Science > Use Case 47: Atmospheric Turbulence—Event Discovery

Big Data Specific Challenges in Mobility	Development for mobile platforms not essential at this time.
Security and Privacy Requirements	No critical issues identified.
Highlight issues for generalizing this use case (e.g. for ref. architecture)	Atmospheric turbulence is only one of many phenomena-based events that could be useful for understanding anomalies in the atmosphere or the ocean that are connected over long distances in space and time. However the process has limits to extensibility, i.e., each phenomena may require very different processes for data mining and predictive analysis.
More Information (URLs)	http://oceanworld.tamu.edu/resources/oceanography-book/teleconnections.htm http://www.forbes.com/sites/toddwoody/2012/03/21/meet-the-scientists-mining-big-data-to-predict-the-weather/

See Figure 16: Atmospheric Turbulence – Event Discovery and Predictive Analytics (Section 2.9.7) – Typical NASA image of turbulent waves

Earth, Environmental and Polar Science> Use Case 48: Climate Studies using the Community Earth System Model

Use Case Title	Climate Studies using the Community Earth System Model at DOE's NERSC center	
Vertical (area)	Research: Climate	
Author/Company/Email	PI: Warren Washington, NCAR	
Actors/Stakeholders and their roles and responsibilities	Climate scientists, U.S. policy makers	
Goals	The goals of the Climate Change Prediction (CCP) group at NCAR are to understand and quantify contributions of natural and anthropogenic-induced patterns of climate variability and change in the 20th and 21st centuries by means of simulations with the Community Earth System Model (CESM).	
Use Case Description	With these model simulations, researchers are able to investigate mechanisms of climate variability and change, as well as to detect and attribute past climate changes, and to project and predict future changes. The simulations are motivated by broad community interest and are widely used by the national and international research communities.	
Current Solutions	**Compute(System)**	NERSC (24M Hours), DOE LCF (41M), NCAR CSL (17M)
	Storage	1.5 PB at NERSC
	Networking	ESNet
	Software	NCAR PIO library and utilities NCL and NCO, parallel NetCDF
Big Data Characteristics	**Data Source (distributed/centralized)**	Data is produced at computing centers. The Earth Systems Grid is an open source effort providing a robust, distributed data and computation platform, enabling world wide access to Peta/Exa-scale scientific data. ESGF manages the first-ever decentralized database for handling climate science data, with multiple petabytes of data at dozens of federated sites worldwide. It 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 Intergovernmental Panel on Climate Change (IPCC).
	Volume (size)	30 PB at NERSC (assuming 15 end-to-end climate change experiments) in 2017; many times more worldwide
	Velocity (e.g. real time)	42 GB/s are produced by the simulations
	Variety (multiple datasets, mashup)	Data must be compared among those from observations, historical reanalysis, and a number of independently produced simulations. The Program for Climate Model Diagnosis and Intercomparison develops methods and tools for the diagnosis and inter-comparison of general circulation models (GCMs) that simulate the global climate. The need for innovative analysis of GCM climate simulations is apparent, as increasingly more complex models are developed, while the disagreements among these simulations and relative to climate observations remain significant and poorly

Earth, Environmental and Polar Science> Use Case 48: Climate Studies using the Community Earth System Model

		understood. The nature and causes of these disagreements must be accounted for in a systematic fashion in order to confidently use GCMs for simulation of putative global climate change.
	Variability (rate of change)	Data is produced by codes running at supercomputer centers. During runtime, intense periods of data i/O occur regularly, but typically consume only a few percent of the total run time. Runs are carried out routinely, but spike as deadlines for reports approach.
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues) and Quality**	Data produced by climate simulations is plays a large role in informing discussion of climate change simulations. Therefore, it must be robust, both from the standpoint of providing a scientifically valid representation of processes that influence climate, but also as that data is stored long term and transferred world-wide to collaborators and other scientists.
	Visualization	Visualization is crucial to understanding a system as complex as the Earth ecosystem.
	Data Types	Earth system scientists are being inundated by an explosion of data generated by ever-increasing resolution in both global models and remote sensors.
	Data Analytics	There is a need to provide data reduction and analysis web services through the Earth System Grid (ESG). A pressing need is emerging for data analysis capabilities closely linked to data archives.
Big Data Specific Challenges (Gaps)	The rapidly growing size of datasets makes scientific analysis a challenge. The need to write data from simulations is outpacing supercomputers' ability to accommodate this need.	
Big Data Specific Challenges in Mobility	Data from simulations and observations must be shared among a large widely distributed community.	
Security and Privacy Requirements		
Highlight issues for generalizing this use case (e.g. for ref. architecture)	ESGF is in the early stages of being adapted for use in two additional domains: biology (to accelerate drug design and development) and energy (infrastructure for California Energy Systems for the 21st Century (CES21)).	
More Information (URLs)	http://esgf.org/ http://www-pcmdi.llnl.gov/ http://www.nersc.gov/ http://science.energy.gov/ber/research/cesd/ http://www2.cisl.ucar.edu/	

Earth, Environmental and Polar Science> Use Case 49: Subsurface Biogeochemistry

Use Case Title	DOE-BER Subsurface Biogeochemistry Scientific Focus Area	
Vertical (area)	Research: Earth Science	
Author/Company/Email	Deb Agarwal, Lawrence Berkeley Lab. daagarwal@lbl.gov	
Actors/Stakeholders and their roles and responsibilities	LBNL Sustainable Systems SFA 2.0, Subsurface Scientists, Hydrologists, Geophysicists, Genomics Experts, JGI, Climate scientists, and DOE SBR.	
Goals	The Sustainable Systems Scientific Focus Area 2.0 Science Plan ("SFA 2.0") has been developed to advance predictive understanding of complex and multiscale terrestrial environments relevant to the DOE mission through specifically considering the scientific gaps defined above.	
Use Case Description	Development of a **G**enome-**E**nabled **Wa**tershed **S**imulation **C**apability (**GEWaSC)** that will provide a predictive framework for understanding how genomic information stored in a subsurface microbiome affects biogeochemical watershed functioning, how watershed-scale processes affect microbial functioning, and how these interactions co-evolve. While modeling capabilities developed by our team and others in the community have represented processes occurring over an impressive range of scales (ranging from a single bacterial cell to that of a contaminant plume), to date little effort has been devoted to developing a framework for systematically connecting scales, as is needed to identify key controls and to simulate important feedbacks. A simulation framework that formally scales from genomes to watersheds is the primary focus of this GEWaSC deliverable.	
Current Solutions	**Compute(System)**	NERSC
	Storage	NERSC
	Networking	ESNet
	Software	PFLOWTran, postgres, HDF5, Akuna, NEWT, etc.
Big Data Characteristics	**Data Source (distributed/centralized)**	Terabase-scale sequencing data from JGI, subsurface and surface hydrological and biogeochemical data from a variety of sensors (including dense geophysical datasets) experimental data from field and lab analysis
	Volume (size)	
	Velocity (e.g. real time)	
	Variety (multiple datasets, mashup)	Data crosses all scales from genomics of the microbes in the soil to watershed hydro-biogeochemistry. The SFA requires the synthesis of diverse and disparate field, laboratory, and simulation datasets across different semantic, spatial, and temporal scales through GEWaSC. Such datasets will be generated by the different research areas and include simulation data, field data (hydrological, geochemical, geophysical), 'omics data, and data from laboratory experiments.
	Variability (rate of change)	Simulations and experiments
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues) and Quality**	Each of the sources samples different properties with different footprints – extremely heterogeneous. Each of the sources has different levels of uncertainty and precision associated with it. In addition, the translation across scales and domains introduces uncertainty as does the data mining. Data quality is critical.

Earth, Environmental and Polar Science> Use Case 49: Subsurface Biogeochemistry

	Visualization	Visualization is crucial to understanding the data.
	Data Types	Described in “Variety” above.
	Data Analytics	Data mining, data quality assessment, cross-correlation across datasets, reduced model development, statistics, quality assessment, data fusion, etc.
Big Data Specific Challenges (Gaps)	Translation across diverse and large datasets that cross domains and scales.	
Big Data Specific Challenges in Mobility	Field experiment data taking would be improved by access to existing data and automated entry of new data via mobile devices.	
Security and Privacy Requirements		
Highlight issues for generalizing this use case (e.g. for ref. architecture)	A wide array of programs in the earth sciences are working on challenges that cross the same domains as this project.	
More Information (URLs)	Under development	

Earth, Environmental and Polar Science> Use Case 50: AmeriFlux and FLUXNET

Use Case Title	DOE-BER AmeriFlux and FLUXNET Networks	
Vertical (area)	Research: Earth Science	
Author/Company/Email	Deb Agarwal, Lawrence Berkeley Lab. daagarwal@lbl.gov	
Actors/Stakeholders and their roles and responsibilities	AmeriFlux scientists, Data Management Team, ICOS, DOE TES, USDA, NSF, and Climate modelers.	
Goals	AmeriFlux Network and FLUXNET measurements provide the crucial linkage between organisms, ecosystems, and process-scale studies at climate-relevant scales of landscapes, regions, and continents, which can be incorporated into biogeochemical and climate models. Results from individual flux sites provide the foundation for a growing body of synthesis and modeling analyses.	
Use Case Description	AmeriFlux network observations enable scaling of trace gas fluxes (CO2, water vapor) across a broad spectrum of times (hours, days, seasons, years, and decades) and space. Moreover, AmeriFlux and FLUXNET 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 Solutions	**Compute(System)**	NERSC
	Storage	NERSC
	Networking	ESNet
	Software	EddyPro, Custom analysis software, R, python, neural networks, Matlab.
Big Data Characteristics	**Data Source (distributed/centralized)**	≈150 towers in AmeriFlux and over 500 towers distributed globally collecting flux measurements.
	Volume (size)	
	Velocity (e.g. real time)	
	Variety (multiple datasets, mashup)	The flux data is relatively uniform, however, the biological, disturbance, and other ancillary data needed to process and to interpret the data is extensive and varies widely. Merging this data with the flux data is challenging in today's systems.
	Variability (rate of change)	
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues) and Quality**	Each site has unique measurement and data processing techniques. The network brings this data together and performs a common processing, gap-filling, and quality assessment. Thousands of users
	Visualization	Graphs and 3D surfaces are used to visualize the data.
	Data Types	Described in "Variety" above.
	Data Analytics	Data mining, data quality assessment, cross-correlation across datasets, data assimilation, data interpolation, statistics, quality assessment, data fusion, etc.
Big Data Specific Challenges (Gaps)	Translation across diverse datasets that cross domains and scales.	
Big Data Specific Challenges in Mobility	Field experiment data taking would be improved by access to existing data and automated entry of new data via mobile devices.	
Security and Privacy Requirements		

Earth, Environmental and Polar Science> Use Case 50: AmeriFlux and FLUXNET

Highlight issues for generalizing this use case (e.g. for ref. architecture)	
More Information (URLs)	http://Ameriflux.lbl.gov http://www.fluxdata.org

Energy> Use Case 51: Consumption Forecasting in Smart Grids

Use Case Title	Consumption forecasting in Smart Grids	
Vertical (area)	Energy Informatics	
Author/Company/Email	Yogesh Simmhan, University of Southern California, simmhan@usc.edu	
Actors/Stakeholders and their roles and responsibilities	Electric Utilities, Campus MicroGrids, Building Managers, Power Consumers, Energy Markets	
Goals	Develop scalable and accurate forecasting models to predict the energy consumption (kWh) within the utility service area under different spatial and temporal granularities to help improve grid reliability and efficiency.	
Use Case Description	Deployment of smart meters are making available near-realtime energy usage data (kWh) every 15-mins at the granularity individual consumers within the service area of smart power utilities. This unprecedented and growing access to fine-grained energy consumption information allows novel analytics capabilities to be developed for predicting energy consumption for customers, transformers, sub-stations and the utility service area. Near-term forecast can be used by utilities and microgrid managers to take preventive action before consumption spikes cause brown/blackouts through demand-response optimization by engaging consumers, bringing peaker units online, or purchasing power from the energy markets. These form an OODA feedback loop. Customers can also use them for energy use planning and budgeting. Medium- to long-term predictions can help utilities and building managers plan generation capacity, renewable portfolio, energy purchasing contracts and sustainable building improvements. Steps involved include *1) Data Collection and Storage*: time-series data from (potentially) millions of smart meters in near real time, features on consumers, facilities and regions, weather forecasts, archival of data for training, testing and validating models; *2) Data Cleaning and Normalization*: Spatio-temporal normalization, gap filling/Interpolation, outlier detection, semantic annotation; *3) Training Forecast Models*: Using univariate timeseries models like ARIMA, and data-driven machine learning models like regression tree, ANN, for different spatial (consumer, transformer) and temporal (15-min, 24-hour) granularities; *4) Prediction*: Predict consumption for different spatio-temporal granularities and prediction horizons using near-realtime and historic data fed to the forecast model with thresholds on prediction latencies.	
Current Solutions	**Compute(System)**	Many-core servers, Commodity Cluster, Workstations
	Storage	SQL Databases, CSV Files, HDFS, Meter Data Management
	Networking	Gigabit Ethernet
	Software	R/Matlab, Weka, Hadoop
Big Data Characteristics	**Data Source (distributed/centralized)**	Head-end of smart meters (distributed), Utility databases (Customer Information, Network topology; centralized), US Census data (distributed), NOAA weather data (distributed), Microgrid building information system (centralized), Microgrid sensor network (distributed)
	Volume (size)	10 GB/day; 4 TB/year *(City scale)*
	Velocity (e.g. real time)	Los Angeles: Once every 15-mins (≈100k streams); Once every 8-hours (≈1.4M streams) with finer grain data aggregated to 8-hour interval

Energy> Use Case 51: Consumption Forecasting in Smart Grids

	Variety (multiple datasets, mashup)	Tuple-based: Timeseries, database rows; Graph-based: Network topology, customer connectivity; Some semantic data for normalization.
	Variability (rate of change)	Meter and weather data change, and are collected/used, on hourly basis. Customer/building/grid topology information is slow changing on a weekly basis
Big Data Science (collection, curation, analysis, action)	**Veracity (Robustness Issues, semantics)**	Versioning and reproducibility is necessary to validate/compare past and current models. Resilience of storage and analytics is important for operational needs. Semantic normalization can help with inter-disciplinary analysis (e.g. utility operators, building managers, power engineers, behavioral scientists)
	Visualization	Map-based visualization of grid service topology, stress; Energy heat-maps; Plots of demand forecasts vs. capacity, what-if analysis; Realtime information display; Apps with push notification of alerts
	Data Quality (syntax)	Gaps in smart meters and weather data; Quality issues in sensor data; Rigorous checks done for “billing quality” meter data;
	Data Types	Timeseries (CSV, SQL tuples), Static information (RDF, XML), topology (shape files)
	Data Analytics	Forecasting models, machine learning models, time series analysis, clustering, motif detection, complex event processing, visual network analysis,
Big Data Specific Challenges (Gaps)	Scalable realtime analytics over large data streams Low-latency analytics for operational needs Federated analytics at utility and microgrid levels Robust time series analytics over millions of customer consumption data Customer behavior modeling, targeted curtailment requests	
Big Data Specific Challenges in Mobility	Apps for engaging with customers: Data collection from customers/premises for behavior modeling, feature extraction; Notification of curtailment requests by utility/building managers; Suggestions on energy efficiency; Geo-localized display of energy footprint.	
Security and Privacy Requirements	Personally identifiable customer data requires careful handling. Customer energy usage data can reveal behavior patterns. Anonymization of information. Data aggregation to avoid customer identification. Data sharing restrictions by federal and state energy regulators. Surveys by behavioral scientists may have IRB (Institutional Review Board) restrictions.	
Highlight issues for generalizing this use case (e.g. for ref. architecture)	Realtime data-driven analytics for cyber-physical systems	
More Information (URLs)	http://smartgrid.usc.edu http://ganges.usc.edu/wiki/Smart_Grid https://www.ladwp.com/ladwp/faces/ladwp/aboutus/a-power/a-p-smartgridla http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6475927	

Appendix B: Summary of Key Properties

Information related to five key properties was extracted from each use case. The five key properties were three Big Data characteristics (volume, velocity, and variety), software related information, and associated analytics. The extracted information is presented in Table B-1.

Table B-1: Use Case Specific Information by Key Properties

	Use Case	Volume	Velocity	Variety	Software	Analytics
1	**M0147** **Census 2000 and 2010**	380 TB	Static for 75 years	Scanned documents	Robust archival storage	None for 75 years
2	**M0148** **NARA: Search, Retrieve, Preservation**	Hundreds of terabytes, and growing	Data loaded in batches, so bursty	Unstructured and structured data: textual documents, emails, photos, scanned documents, multimedia, social networks, web sites, databases, etc.	Custom software, commercial search products, commercial databases	Crawl/index, search, ranking, predictive search; data categorization (sensitive, confidential, etc.); personally identifiable information (PII) detection and flagging
3	**M0219** **Statistical Survey Response Improvement**	Approximately 1 PB	Variable, field data streamed continuously, Census was ≈150 million records transmitted	Strings and numerical data	Hadoop, Spark, Hive, R, SAS, Mahout, Allegrograph, MySQL, Oracle, Storm, BigMemory, Cassandra, Pig	Recommendation systems, continued monitoring
4	**M0222** **Non-Traditional Data in Statistical Survey Response Improvement**	—	—	Survey data, other government administrative data, web-scraped data, wireless data, e-transaction data, (potentially) social media data and positioning data from various sources	Hadoop, Spark, Hive, R, SAS, Mahout, Allegrograph, MySQL, Oracle, Storm, BigMemory, Cassandra, Pig	New analytics to create reliable information from non-traditional disparate sources
5	**M0175** **Cloud Eco-**	—	Real time	—	Hadoop RDBMS XBRL	Fraud detection

	Use Case	Volume	Velocity	Variety	Software	Analytics
	System for Finance					
6	**M0161 Mendeley**	15 TB presently, growing about 1 TB per month	Currently Hadoop batch jobs scheduled daily, real-time recommended in future	PDF documents and log files of social network and client activities	Hadoop, Scribe, Hive, Mahout, Python	Standard libraries for machine learning and analytics, LDA, custom-built reporting tools for aggregating readership and social activities per document
7	**M0164 Netflix Movie Service**	Summer 2012 – 25 million subscribers, 4 million ratings per day, 3 million searches per day, 1 billion hours streamed in June 2012; Cloud storage – 2 petabytes in June 2013	Media (video and properties) and rankings continually updated	Data vary from digital media to user rankings, user profiles, and media properties for content-based recommendations	Hadoop and Pig; Cassandra; Teradata	Personalized recommender systems using logistic/linear regression, elastic nets, matrix factorization, clustering, LDA, association rules, gradient-boosted decision trees, and others; streaming video delivery
8	**M0165 Web Search**	45 billion web pages total, 500 million photos uploaded each day, 100 hours of video uploaded to YouTube each minute	Real-time updating and real-time responses to queries	Multiple media	Map/Reduce + Bigtable; Dryad + Cosmos; PageRank; final step essentially a recommender engine	Crawling; searching, including topic-based searches; ranking; recommending
9	**M0137 Business Continuity and Disaster Recovery Within a Cloud Eco-System**	Terabytes up to petabytes	Can be real time for recent changes	Must work for all data	Hadoop, Map/Reduce, open source, and/or vendor proprietary such as AWS, Google Cloud Services, and Microsoft	Robust backup

	Use Case	Volume	Velocity	Variety	Software	Analytics
10	**M0103** **Cargo Shipping**	—	Needs to become real time, currently updated at events	Event-based	—	Distributed event analysis identifying problems
11	**M0162** **Materials Data for Manufacturing**	500,000 material types in 1980s, much growth since then	Ongoing increase in new materials	Many datasets with no standards	National programs (Japan, Korea, and China), application areas (EU nuclear program), proprietary systems (Granta, etc.)	No broadly applicable analytics
12	**M0176** **Simulation-Driven Materials Genomics**	100 TB (current), 500 TB within five years, scalable key-value and object store databases needed	Regular data added from simulations	Varied data and simulation results	MongoDB, GPFS, PyMatGen, FireWorks, VASP, ABINIT, NWChem, BerkeleyGW, varied community codes	Map/Reduce and search that join simulation and experimental data
13	**M0213** **Large-Scale Geospatial Analysis and Visualization**	Imagery – hundreds of terabytes; vector data – tens of GBs but billions of points	Vectors transmitted in near real time	Imagery, vector (various formats such as shape files, KML, text streams) and many object structures	Geospatially enabled RDBMS, Esri ArcServer, Geoserver	Closest point of approach, deviation from route, point density over time, PCA and ICA
14	**M0214** **Object Identification and Tracking**	FMV – 30 to 60 frames per second at full-color 1080P resolution; WALF – 1 to 10 frames per second at 10,000 x 10,000 full-color resolution	Real time	A few standard imagery or video formats	Custom software and tools including traditional RDBMS and display tools	Visualization as overlays on a GIS, basic object detection analytics and integration with sophisticated situation awareness tools with data fusion
15	**M0215** **Intelligence Data Processing and Analysis**	Tens of terabytes to hundreds of petabytes, individual warfighters (first responders) would have at most one to hundreds of GBs	Much real-time, imagery intelligence devices that gather a petabyte of data in a few hours	Text files, raw media, imagery, video, audio, electronic data, human-generated data	Hadoop, Accumulo (BigTable), Solr, NLP, Puppet (for deployment and security) and Storm; GIS	Near real-time alerts based on patterns and baseline changes, link analysis, geospatial analysis, text analytics (sentiment, entity extraction, etc.)

	Use Case	Volume	Velocity	Variety	Software	Analytics
16	**M0177 EMR Data**	12 million patients, more than 4 billion discrete clinical observations, > 20 TB raw data	0.5 to 1.5 million new real-time clinical transactions added per day	Broad variety of data from doctors, nurses, laboratories and instruments	Teradata, PostgreSQL, MongoDB, Hadoop, Hive, R	Information retrieval methods (tf-idf), NLP, maximum likelihood estimators, Bayesian networks
17	**M0089 Pathology Imaging**	1 GB raw image data + 1.5 GB analytical results per 2D image, 1 TB raw image data + 1 TB analytical results per 3D image, 1 PB data per moderated hospital per year	Once generated, data will not be changed	Images	MPI for image analysis, Map/Reduce + Hive with spatial extension	Image analysis, spatial queries and analytics, feature clustering and classification
18	**M0191 Computational Bioimaging**	Medical diagnostic imaging around 70 PB annually, 32 TB on emerging machines for a single scan	Volume of data acquisition requires HPC back end	Multi-modal imaging with disparate channels of data	Scalable key-value and object store databases; ImageJ, OMERO, VolRover, advanced segmentation and feature detection methods	Machine learning (support vector machine [SVM] and random forest [RF]) for classification and recommendation services
19	**M0078 Genomic Measurements**	>100 TB in 1 to 2 years at NIST, many PBs in healthcare community	≈300 GB of compressed data/day generated by DNA sequencers	File formats not well-standardized, though some standards exist; generally structured data	Open-source sequencing bioinformatics software from academic groups	Processing of raw data to produce variant calls, clinical interpretation of variants
20	**M0188 Comparative Analysis for Metagenomes and Genomes**	50 TB	New sequencers stream in data at growing rate	Biological data that are inherently heterogeneous, complex, structural, and hierarchical; besides core genomic data, new types of omics data such as transcriptomics, methylomics, and proteomics	Standard bioinformatics tools (BLAST, HMMER, multiple alignment and phylogenetic tools, gene callers, sequence feature predictors), Perl/Python wrapper scripts	Descriptive statistics, statistical significance in hypothesis testing, data clustering and classification

	Use Case	Volume	Velocity	Variety	Software	Analytics
21	**M0140 Individualized Diabetes Management**	5 million patients	Not real time but updated periodically	100 controlled vocabulary values and 1,000 continuous values per patient, mostly time-stamped values	HDFS supplementing Mayo internal data warehouse (EDT)	Integration of data into semantic graphs, using graph traverse to replace SQL join; development of semantic graph-mining algorithms to identify graph patterns, index graph, and search graph; indexed Hbase; custom code to develop new patient properties from stored data
22	**M0174 Statistical Relational Artificial Intelligence for Health Care**	Hundreds of GBs for a single cohort of a few hundred people; possibly on the order of 1 PB when dealing with millions of patients	Constant updates to EHRs; in other controlled studies, data often in batches at regular intervals	Critical feature – data typically in multiple tables, need to be merged to perform analysis	Mainly Java-based, in-house tools to process the data	Relational probabilistic models (Statistical Relational AI) learned from multiple data types
23	**M0172 World Population-Scale Epidemiological Study**	100 TB	Low number of data feeding into the simulation, massive amounts of real-time data generated by simulation	Can be rich with various population activities, geographical, socio-economic, cultural variations	Charm++, MPI	Simulations on a synthetic population
24	**M0173 Social Contagion Modeling for Planning**	Tens of terabytes per year	During social unrest events, human interactions and mobility leads to rapid changes in data; e.g., who follows whom in Twitter	Big issues – data fusion, combining data from different sources, dealing with missing or incomplete data	Specialized simulators, open source software, proprietary modeling environments; databases	Models of behavior of humans and hard infrastructures, models of their interactions, visualization of results

	Use Case	Volume	Velocity	Variety	Software	Analytics
25	M0141 **Biodiversity and LifeWatch**	N/A	Real-time processing and analysis in case of natural or industrial disaster	Rich variety and number of involved databases and observation data	RDBMS	Requires advanced and rich visualization
26	M0136 **Large-Scale Deep Learning**	Current datasets typically 1 TB to 10 TB, possibly 100 million images to train a self-driving car	Much faster than real-time processing; for autonomous driving, need to process thousands of high-resolution (six megapixels or more) images per second	Neural net very heterogeneous as it learns many different features	In-house GPU kernels and MPI-based communication developed by Stanford, C++/Python source	Small degree of batch statistical preprocessing, all other data analysis performed by the learning algorithm itself
27	M0171 **Organizing Large-Scale Unstructured Collections of Consumer Photos**	500+ billion photos on Facebook, 5+ billion photos on Flickr	Over 500 million images uploaded to Facebook each day	Images and metadata including EXIF (Exchangeable Image File) tags (focal distance, camera type, etc.)	Hadoop Map/Reduce, simple hand-written multi-threaded tools (Secure Shell [SSH] and sockets for communication)	Robust non-linear least squares optimization problem, SVM
28	M0160 **Truthy Twitter Data**	30 TB/year compressed data	Near real-time data storage, querying and analysis	Schema provided by social media data source; currently using Twitter only; plans to expand, incorporating Google+ and Facebook	Hadoop IndexedHBase and HDFS; Hadoop, Hive, Redis for data management; Python: SciPy NumPy and MPI for data analysis	Anomaly detection, stream clustering, signal classification, online learning; information diffusion, clustering, dynamic network visualization
29	M0211 **Crowd Sourcing in Humanities**	GBs (text, surveys, experiment values) to hundreds of terabytes (multimedia)	Data continuously updated and analyzed incrementally	So far mostly homogeneous small datasets; expected large distributed heterogeneous datasets	XML technology, traditional relational databases	Pattern recognition (e.g., speech recognition, automatic audio-visual analysis, cultural patterns), identification of structures (lexical units, linguistic rules, etc.)

	Use Case	Volume	Velocity	Variety	Software	Analytics
30	**M0158** **CINET for Network Science**	Can be hundreds of GBs for a single network, 1,000 to 5,000 networks and methods	Dynamic networks, network collection growing	Many types of networks	Graph libraries (Galib, NetworkX); distributed workflow management (Simfrastructure, databases, semantic web tools)	Network visualization
31	**M0190** **NIST Information Access Division**	>900 million web pages occupying 30 TB of storage, 100 million tweets, 100 million ground-truthed biometric images, hundreds of thousands of partially ground-truthed video clips, terabytes of smaller fully ground-truthed test collections	Legacy evaluations mostly focused on retrospective analytics, newer evaluations focused on simulations of real-time analytic challenges from multiple data streams	Wide variety of data types including textual search/extraction, machine translation, speech recognition, image and voice biometrics, object and person recognition and tracking, document analysis, human-computer dialogue, multimedia search/extraction	PERL, Python, C/C++, Matlab, R development tools; create ground-up test and measurement applications	Information extraction, filtering, search, and summarization; image and voice biometrics; speech recognition and understanding; machine translation; video person/object detection and tracking; event detection; imagery/document matching; novelty detection; structural semantic temporal analytics
32	**M0130** **DataNet (iRODS)**	Petabytes, hundreds of millions of files	Real time and batch	Rich	iRODS	Supports general analysis workflows
33	**M0163** **The Discinnet Process**	Small as metadata to Big Data	Real time	Can tackle arbitrary Big Data	Symfony-PHP, Linux, MySQL	--
34	**M0131** **Semantic Graph-Search**	A few terabytes	Evolving in time	Rich	Database	Data graph processing
35	**M0189** **Light Source Beamlines**	50 to 400 GB per day, total ≈400 TB	Continuous stream of data, but analysis need not be real time	Images	Octopus for Tomographic Reconstruction, Avizo (http://vsg3d.com) and FIJI (a distribution of ImageJ)	Volume reconstruction, feature identification, etc.

	Use Case	Volume	Velocity	Variety	Software	Analytics
36	**M0170** **Catalina Real-Time Transient Survey**	≈100 TB total increasing by 0.1 TB a night accessing PBs of base astronomy data, 30 TB a night from successor LSST in 2020s	Nightly update runs processes in real time	Images, spectra, time series, catalogs	Custom data processing pipeline and data analysis software	Detection of rare events and relation to existing diverse data
37	**M0185** **DOE Extreme Data from Cosmological Sky Survey**	Several petabytes from Dark Energy Survey and Zwicky Transient Factory, simulations > 10 PB	Analysis done in batch mode with data from observations and simulations updated daily	Image and simulation data	MPI, FFTW, viz packages, numpy, Boost, OpenMP, ScaLAPCK, PSQL and MySQL databases, Eigen, cfitsio, astrometry.net, and Minuit2	New analytics needed to analyze simulation results
38	**M0209** **Large Survey Data for Cosmology**	Petabytes of data from Dark Energy Survey	400 images of 1 GB in size per night	Images	Linux cluster, Oracle RDBMS server, Postgres PSQL, large memory machines, standard Linux interactive hosts, GPFS; for simulations, HPC resources; standard astrophysics reduction software as well as Perl/Python wrapper scripts	Machine learning to find optical transients, Cholesky decomposition for thousands of simulations with matrices of order 1 million on a side and parallel image storage
39	**M0166** **Particle Physics at LHC**	15 PB of data (experiment and Monte Carlo combined) per year	Data updated continuously with sophisticated real-time selection and test analysis but all analyzed "properly" offline	Different format for each stage in analysis but data uniform within each stage	Grid-based environment with over 350,000 cores running simultaneously	Sophisticated specialized data analysis code followed by basic exploratory statistics (histogram) with complex detector efficiency corrections
40	**M0210** **Belle II High Energy Physics Experiment**	Eventually 120 PB of Monte Carlo and observational data	Data updated continuously with sophisticated real-time selection and test analysis but all	Different format for each stage in analysis but data uniform within each stage	DIRAC Grid software	Sophisticated specialized data analysis code followed by basic exploratory statistics (histogram) with

	Use Case	Volume	Velocity	Variety	Software	Analytics
			analyzed "properly" offline			complex detector efficiency corrections
41	**M0155 EISCAT 3D incoherent scatter radar system**	Terabytes/year (current), 40 PB/year starting ≈2022	Data updated continuously with real-time test analysis and batch full analysis	Big data uniform	Custom analysis based on flat file data storage	Pattern recognition, demanding correlation routines, high-level parameter extraction
42	**M0157 ENVRI Environmental Research Infrastructure**	Low volume (apart from EISCAT 3D given above), one system EPOS ≈15 TB/year	Mainly real-time data streams	Six separate projects with common architecture for infrastructure, data very diverse across projects	R and Python (Matplotlib) for visualization, custom software for processing	Data assimilation, (statistical) analysis, data mining, data extraction, scientific modeling and simulation, scientific workflow
43	**M0167 CReSIS Remote Sensing**	Around 1 PB (current) increasing by 50 to 100 TB per mission, future expedition ≈1 PB each	Data taken in ≈two-month missions including test analysis and then later batch processing	Raw data, images with final layer data used for science	Matlab for custom raw data processing, custom image processing software, GIS as user interface	Custom signal processing to produce radar images that are analyzed by image processing to find layers
44	**M0127 UAVSAR Data Processing**	110 TB raw data and 40 TB processed, plus smaller samples	Data come from aircraft and so incrementally added, data occasionally get reprocessed: new processing methods or parameters	Image and annotation files	ROI_PAC, GeoServer, GDAL, GeoTIFF-supporting tools; moving to clouds	Process raw data to get images that are run through image processing tools and accessed from GIS
45	**M0182 NASA LARC/GSFC iRODS**	MERRA collection (below) represents most of total data, other smaller collections	Periodic updates every six months	Many applications to combine MERRA reanalysis data with other reanalyses and observational data such as CERES	SGE Univa Grid Engine Version 8.1, iRODS Version 3.2 and/or 3.3, IBM GPFS Version 3.4, Cloudera Version 4.5.2-1	Federation software

	Use Case	Volume	Velocity	Variety	Software	Analytics
46	M0129 **MERRA Analytic Services**	480 TB from MERRA	Increases at ≈1 TB/month	Applications to combine MERRA reanalysis data with other re-analyses and observational data	Cloudera, iRODS, Amazon AWS	CAaaS
47	M0090 **Atmospheric Turbulence**	200 TB (current), 500 TB within 5 years	Data analyzed incrementally	Re-analysis datasets are inconsistent in format, resolution, semantics, and metadata; interpretation/analysis of each of these input streams into a common product	Map/Reduce or the like, SciDB or other scientific database	Data mining customized for specific event types
48	M0186 **Climate Studies**	Up to 30 PB/year from 15 end-to-end simulations at NERSC, more at other HPC centers	42 GB/second from simulations	Variety across simulation groups and between observation and simulation	National Center for Atmospheric Research (NCAR) PIO library and utilities NCL and NCO, parallel NetCDF	Need analytics next to data storage
49	M0183 **DOE-BER Subsurface Biogeochemistry**	—	—	From omics of the microbes in the soil to watershed hydro-biogeochemistry, from observation to simulation	PFLOWTran, postgres, HDF5, Akuna, NEWT, etc.	Data mining, data quality assessment, cross-correlation across datasets, reduced model development, statistics, quality assessment, data fusion
50	M0184 **DOE-BER AmeriFlux and FLUXNET Networks**	—	Streaming data from ≈150 towers in AmeriFlux and over 500 towers distributed globally collecting flux measurements	Flux data merged with biological, disturbance, and other ancillary data	EddyPro, custom analysis software, R, Python, neural networks, Matlab	Data mining, data quality assessment, cross-correlation across datasets, data assimilation, data interpolation, statistics, quality assessment, data fusion
51	M0223 **Consumption forecasting in Smart Grids**	4 TB/year for a city with 1.4 million sensors, such as Los Angeles	Streaming data from millions of sensors	Tuple-based: timeseries, database rows; graph-based: network topology, customer connectivity; some semantic data for normalization	R/Matlab, Weka, Hadoop; GIS-based visualization	Forecasting models, machine learning models, time series analysis, clustering, motif detection,

Use Case	Volume	Velocity	Variety	Software	Analytics
					complex event processing, visual network analysis
M0633 NASA Earth Observing System Data and Information System (EOSDIS)	Data size is 22PB corresponding to Total Earth Observation Data managed by NASA EOSDIS accumulated since 1994. Higher resolution spaceborne instruments are expected to increase that volume by two orders of magnitude (~200 PB) over the next 7 years. In a given year, EOSDIS distributes a volume that is comparable to the overall cumulative archive volume.	This is now an archive of 23 years data but is continually increasing in both gathered and distributed data. In a given year, EOSDIS distributes a volume that is comparable to the overall cumulative archive volume.	EOSDIS's Common Metadata Repository includes over 6400 EOSDIS data collections as of June 2017, providing significant challenges in data discovery. CMR and other interoperability frameworks (metrics, browse imagery, governance) knit together 12 different archives, each with a different implementation. Nearly all Earth science disciplines are represented in EOSDIS.	EOSDIS uses high-performance software, such as the netCDF Command Operators. However, current prototypes are using cloud computing and data-parallel algorithms (e.g., Spark) to achieve an order of magnitude speed-up. Cloud storage and database schemes are being investigated. Python, Fortran, C languages. Visualization through tools such as Giovanni.	Analytics used includes: (1) computing statistical measures of Earth Observation data across a variety of dimensions (2) examining covariance and correlation of a variety of Earth observations (3) assimilating multiple data variables into a model using Kalman filtering (4) analyzing time series.
M0634 **Web-Enabled Landsat Data (WELD) Processing**	The data represent the operational time period of 1984 to 2011 for the Landsat 4, 5, and 7 satellites and corresponds to 30PB of processed data through the pipeline (1PB inputs, 10PB intermediate, 6PB outputs)	Data was collected over a period of 27 years and is being processed over a period of 5 years. Based on programmatic goals of processing several iterations of the final product over the span of the project,	None. This use case basically deals with a single dataset.	NEX science platform – data management, workflow processing, provenance capture; WELD science processing algorithms from South Dakota State University (SDSU), browse visualization, and time-series code; Global Imagery Browse Service (GIBS) data visualization	There are number of analytics processes throughout the processing pipeline. The key analytics is identifying best available pixels for spatio-temporal composition and spatial aggregation processes as a part of the overall QA. The analytics

Use Case	Volume	Velocity	Variety	Software	Analytics
		150TB/day is processed per day during processing time periods.		platform; USGS data distribution platform. Custom-built application and libraries built on top of open-source libraries.	algorithms are custom developed for this use case.

Appendix C: Use Case Requirements Summary

Requirements were extracted from each version 1 use case (the version 2 use cases were not included) within seven characteristic categories introduced in Section 3.1. The number of requirements within each category varied for each use case. Table C-1 contains the use case specific requirements.

Table C-1: Use Case Specific Requirements

	Use Case	Data Sources	Data Transformation	Capabilities	Data Consumer	Security and Privacy	Life Cycle Management	Other
1	M0147 Census 2010 and 2000	1. Large document format from centralized storage	--	1. Large centralized storage (storage)	--	1. Title 13 data	1. Long-term preservation of data as-is for 75 years 2. Long-term preservation at the bit level 3. Curation process including format transformation 4. Access and analytics processing after 75 years 5. No data loss	--
2	M0148 NARA: Search, Retrieve, Preservation	1. Distributed data sources 2. Large data storage 3. Bursty data ranging from GBs to hundreds of terabytes 4. Wide variety of data formats including unstructured and	1. Crawl and index from distributed data sources 2. Various analytics processing including ranking, data categorization, detection of PII data 3. Data preprocessing	1. Large data storage 2. Various storage systems such as NetApps, Hitachi, magnetic tapes	1. High relevancy and high recall from search 2. High accuracy from categorization of records 3. Various storage systems such as NetApps,	1. Security policy	1. Pre-process for virus scan 2. File format identification 3. Indexing 4. Records categorization	1. Mobile search with similar interfaces/ results from desktop

	Use Case	Data Sources	Data Transformation	Capabilities	Data Consumer	Security and Privacy	Life Cycle Management	Other
		structured data 5. Distributed data sources in different clouds	4. Long-term preservation management of large varied datasets 5. Huge numbers of data with high relevancy and recall		Hitachi, magnetic tapes			
3	**M0219** Statistical Survey Response Improvement	1. Data size of approximately one petabyte	1. Analytics for recommendation systems, continued monitoring, and general survey improvement	1. Hadoop, Spark, Hive, R, SAS, Mahout, Allegrograph, MySQL, Oracle, Storm, BigMemory, Cassandra, Pig	1. Data visualization for data review, operational activity, and general analysis; continual evolution	1. Improved recommendation systems that reduce costs and improve quality while providing confidentiality safeguards that are reliable and publicly auditable 2. Confidential and secure data; processes that are auditable for security and confidentiality as required by various legal statutes	1. High veracity on data and very robust systems (challenges: semantic integrity of conceptual metadata concerning what exactly is measured and the resulting limits of inference)	1. Mobile access
4	**M0222** Non-Traditional Data in	--	1. Analytics to create reliable estimates using data from	1. Hadoop, Spark, Hive, R, SAS, Mahout, Allegrograph,	1. Data visualization for data review,	1. Confidential and secure data; processes that are	1. High veracity on data and very robust systems (challenges: semantic integrity of	--

	Use Case	Data Sources	Data Transformation	Capabilities	Data Consumer	Security and Privacy	Life Cycle Management	Other
	Statistical Survey Response Improvement		traditional survey sources, government administrative data sources, and non-traditional sources from the digital economy	MySQL, Oracle, Storm, BigMemory, Cassandra, Pig	operational activity, and general analysis; continual evolution	auditable for security and confidentiality as required by various legal statutes	conceptual metadata concerning what exactly is measured and the resulting limits of inference)	
5	M0175 Cloud Eco-System for Finance	1. Real-time ingestion of data	1. Real-time analytics	--	--	1. Strong security and privacy constraints	--	1. Mobile access
6	M0161 Mendeley	1. File-based documents with constant new uploads 2. Variety of file types such as PDFs, social network log files, client activities images, spreadsheet, presentation files	1. Standard machine learning and analytics libraries 2. Efficient scalable and parallelized way to match between documents 3. Third-party annotation tools or publisher watermarks and cover pages	1. Amazon Elastic Compute Cloud (EC2) with HDFS (infrastructure) 2. S3 (storage) 3. Hadoop (platform) 4. Scribe, Hive, Mahout, Python (language) 5. Moderate storage (15 TB with 1 TB/ month) 6. Batch and real-time processing	1. Custom-built reporting tools 2. Visualization tools such as networking graph, scatterplots, etc.	1. Access controls for who reads what content	1. Metadata management from PDF extraction 2. Identification of document duplication 3. Persistent identifier 4. Metadata correlation between data repositories such as CrossRef, PubMed, and Arxiv	1. Windows Android and iOS mobile devices for content deliverables from Windows desktops
7	M0164 Netflix Movie Service	1. User profiles and ranking information	1. Streaming video contents to multiple clients 2. Analytic processing for matching client interest in movie selection	1. Hadoop (platform) 2. Pig (language) 3. Cassandra and Hive 4. Huge numbers of subscribers, ratings, and	1. Streaming and rendering media	1. Preservation of users, privacy and digital rights for media	1. Continued ranking and updating based on user profile and analytic results	1. Smart interface accessing movie content on mobile platforms

	Use Case	Data Sources	Data Transformation	Capabilities	Data Consumer	Security and Privacy	Life Cycle Management	Other
			3. Various analytic processing techniques for consumer personalization 4. Robust learning algorithms 5. Continued analytic processing based on monitoring and performance results	searches per day (DB) 5. Huge amounts of storage (2 PB) 6. I/O intensive processing				
8	M0165 Web Search	1. Distributed data sources 2. Streaming data 3. Multimedia content	1. Dynamic fetching content over the network 2. Linking of user profiles and social network data	1. Petabytes of text and rich media (storage)	1. Search time of ≈0.1 seconds 2. Top 10 ranked results 3. Page layout (visual)	1. Access control 2. Protection of sensitive content	1. Data purge after certain time interval (a few months) 2. Data cleaning	1. Mobile search and rendering
9	M0137 Business Continuity and Disaster Recovery Within a Cloud Eco-System	--	1. Robust backup algorithm 2. Replication of recent changes	1. Hadoop 2. Commercial cloud services	--	1. Strong security for many applications	--	--
10	M0103 Cargo Shipping	1. Centralized and real-time distributed sites/sensors	1. Tracking items based on the unique identification with its sensor information, GPS coordinates	1. Internet connectivity	--	1. Security policy	--	--

	Use Case	Data Sources	Data Transformation	Capabilities	Data Consumer	Security and Privacy	Life Cycle Management	Other
			2. Real-time updates on tracking items					
11	M0162 Materials Data for Manufacturing	1. Distributed data repositories for more than 500,000 commercial materials 2. Many varieties of datasets 3. Text, graphics, and images	1. Hundreds of independent variables need to be collected to create robust datasets	--	1. Visualization for materials discovery from many independent variables 2. Visualization tools for multi-variable materials	1. Protection of proprietary sensitive data 2. Tools to mask proprietary information	1. Handle data quality (currently poor or no process)	--
12	M0176 Simulation-Driven Materials Genomics	1. Data streams from peta/exascale centralized simulation systems 2. Distributed web dataflows from central gateway to users	1. High-throughput computing real-time data analysis for web-like responsiveness 2. Mashup of simulation outputs across codes 3. Search and crowd-driven with computation backend, flexibility for new targets 4. Map/Reduce and search to join simulation and experimental data	1. Massive (150,000 cores) legacy infrastructure (infrastructure) 2. GPFS (storage) 3. MonogDB systems (platform) 4. 10 GB networking 5. Various analytic tools such as PyMatGen, FireWorks, VASP, ABINIT, NWChem, BerkeleyGW, varied community codes 6. Large storage	1. Browser-based search for growing materials data	1. Sandbox as independent working areas between different data stakeholders 2. Policy-driven federation of datasets	1. Validation and uncertainty quantification (UQ) of simulation with experimental data 2. UQ in results from multiple datasets	1. Mobile applications (apps) to access materials genomics information

	Use Case	Data Sources	Data Transformation	Capabilities	Data Consumer	Security and Privacy	Life Cycle Management	Other
				(storage) 7. Scalable key-value and object store (platform) 8. Data streams from peta/exascale centralized simulation systems				
13	M0213 Large-Scale Geospatial Analysis and Visualization	1. Unique approaches to indexing and distributed analysis required for geospatial data	1. Analytics: closest point of approach, deviation from route, point density over time, PCA and ICA 2. Unique approaches to indexing and distributed analysis required for geospatial data	1. Geospatially enabled RDBMS, geospatial server/analysis software, e.g., ESRI ArcServer, Geoserver	1. Visualization with GIS at high and low network bandwidths and on dedicated facilities and handhelds	1. Complete security of sensitive data in transit and at rest (particularly on handhelds)	--	--
14	M0214 Object Identificatio n and Tracking	1. Real-time data FMV (30 to 60 frames/ second at full-color 1080P resolution) and WALF (1 to 10 frames/ second at 10,000 x 10,000 full-color resolution)	1. Rich analytics with object identification, pattern recognition, crowd behavior, economic activity, and data fusion	1. Wide range of custom software and tools including traditional RDBMSs and display tools 2. Several network requirements 3. GPU usage important	1. Visualization of extracted outputs as overlays on a geospatial display; links back to the originating image/video segment as overlay objects	1. Significant security and privacy issues; sources and methods never compromised	1. Veracity of extracted objects	--

	Use Case	Data Sources	Data Transformation	Capabilities	Data Consumer	Security and Privacy	Life Cycle Management	Other
					2. Output the form of Open Geospatial Consortium (OGC)-compliant web features or standard geospatial files (shape files, KML)			
15	**M0215** Intelligence Data Processing and Analysis	1. Much real-time data with processing at near-real time (at worst) 2. Data in disparate silos, must be accessible through a semantically integrated data space 3. Diverse data: text files, raw media, imagery, video, audio, electronic data, human-generated data	1. Analytics: Near Real Time (NRT) alerts based on patterns and baseline changes	1. Tolerance of unreliable networks to warfighter and remote sensors 2. Up to hundreds of petabytes of data supported by modest to large clusters and clouds 3. Hadoop, Accumulo (Big Table), Solr, NLP (several variants), Puppet (for deployment and security), Storm, custom applications, visualization tools	1. Geospatial overlays (GIS) and network diagrams (primary visualizations)	1. Protection of data against unauthorized access or disclosure and tampering	1. Data provenance (e.g. tracking of all transfers and transformations) over the life of the data	--
16	**M0177** EMR Data	1. Heterogeneous, high-volume,	1. A comprehensive and consistent	1. Hadoop, Hive, R. Unix-based 2. Cray	1. Results of analytics provided for	1. Data consumer direct access to	1. Standardize, aggregate, and normalize data from	1. Security across mobile devices

Use Case	Data Sources	Data Transformation	Capabilities	Data Consumer	Security and Privacy	Life Cycle Management	Other
	diverse data sources 2. Volume: > 12 million entities (patients), > 4 billion records or data points (discrete clinical observations), aggregate of > 20 TB raw data 3. Velocity: 500,000 to 1.5 million new transactions per day 4. Variety: formats include numeric, structured numeric, free-text, structured text, discrete nominal, discrete ordinal, discrete structured, binary large blobs (images and video) 5. Data evolve over time in a highly variable fashion	view of data across sources and over time 2. Analytic techniques: information retrieval, NLP, machine learning decision models, maximum likelihood estimators, Bayesian networks	supercomputer 3. Teradata, PostgreSQL, MongoDB 4. Various, with significant I/O intensive processing	use by data consumers/ stakeholders, i.e., those who did not actually perform the analysis; specific visualization techniques	data as well as to the results of analytics performed by informatics research scientists and health service researchers 2. Protection of all health data in compliance with governmental regulations 3. Protection of data in accordance with data providers, policies. 4. Security and privacy policies unique to a data subset 5. Robust security to prevent data breaches	disparate sources 2. Reduce errors and bias 3. Common nomenclature and classification of content across disparate sources—particularly challenging in the health IT space, as the taxonomies continue to evolve—SNOMED, International Classification of Diseases (ICD) 9 and future ICD 10, etc.	

	Use Case	Data Sources	Data Transformation	Capabilities	Data Consumer	Security and Privacy	Life Cycle Management	Other
17	M0089 Pathology Imaging	1. High-resolution spatial digitized pathology images 2. Various image quality analyses algorithms 3. Various image data formats, especially BigTIFF with structured data for analytical results 4. Image analysis, spatial queries and analytics, feature clustering, and classification	1. High-performance image analysis to extract spatial information 2. Spatial queries and analytics, feature clustering and classification 3. Analytic processing on huge multi-dimensional large dataset; correlation with other data types such as clinical data, omic data	1. Legacy system and cloud (computing cluster) 2. Huge legacy and new storage such as storage area network (SAN) or HDFS (storage) 3. High-throughput network link (networking) 4. MPI image analysis, Map/Reduce, Hive with spatial extension (software packages)	1. Visualization for validation and training	1. Security and privacy protection for protected health information	1. Human annotations for validation	1. 3D visualization and rendering on mobile platforms
18	M0191 Computational Bioimaging	1. Distributed multi-modal high-resolution experimental sources of bioimages (instruments) 2. 50 TB of data in formats that include images	1. High-throughput computing with responsive analysis 2. Segmentation of regions of interest; crowd-based selection and extraction of features; object classification, and organization; and search 3. Advanced biosciences	1. ImageJ, OMERO, VolRover, advanced segmentation and feature detection methods from applied math researchers; scalable key-value and object store databases needed 2. NERSC's Hopper	1. 3D structural modeling	1. Significant but optional security and privacy including secure servers and anonymization	1. Workflow components including data acquisition, storage, enhancement, minimizing noise	--

	Use Case	Data Sources	Data Transformation	Capabilities	Data Consumer	Security and Privacy	Life Cycle Management	Other
			discovery through Big Data techniques / extreme-scale computing; in-database processing and analytics; machine learning (SVM and RF) for classification and recommendation services; advanced algorithms for massive image analysis; high-performance computational solutions 4. Massive data analysis toward massive imaging datasets.	infrastructure 3. database and image collections 4. 10 GB and future 100 GB and advanced networking (software defined networking [SDN])				
19	M0078 Genomic Measurements	1. High-throughput compressed data (300 GB/day) from various DNA sequencers 2. Distributed data source (sequencers) 3. Various file formats with both	1. Processing raw data in variant calls 2. Challenge: characterizing machine learning for complex analysis on systematic errors from sequencing technologies	1. Legacy computing cluster and other PaaS and IaaS (computing cluster) 2. Huge data storage in PB range (storage) 3. Unix-based legacy sequencing bioinformatics	1. Data format for genome browsers	1. Security and privacy protection of health records and clinical research databases	--	1. Mobile platforms for physicians accessing genomic data (mobile device)

	Use Case	Data Sources	Data Transformation	Capabilities	Data Consumer	Security and Privacy	Life Cycle Management	Other
		structured and unstructured data		software (software package)				
20	M0188 Comparative Analysis for Metagenomes and Genomes	1. Multiple centralized data sources 2. Proteins and their structural features, core genomic data, new types of omics data such as transcriptomics, methylomics, and proteomics describing gene expression 3. Front real-time web UI interactive; backend data loading processing that keeps up with exponential growth of sequence data due to the rapid drop in cost of sequencing technology 4. Heterogeneous, complex,	2. Scalable RDBMS for heterogeneous biological data 2. Real-time rapid and parallel bulk loading 3. Oracle RDBMS, SQLite files, flat text files, Lucy (a version of Lucene) for keyword searches, BLAST databases, USEARCH databases 4. Linux cluster, Oracle RDBMS server, large memory machines, standard Linux interactive hosts 5. Sequencing and comparative analysis techniques for highly complex data 6. Descriptive statistics	1. Huge data storage	1. Real-time interactive parallel bulk loading capability 2. Interactive Web UI, backend pre-computations, batch job computation submission from the UI. 3. Download of assembled and annotated datasets for offline analysis 4. Ability to query and browse data via interactive web UI 5. Visualize data structure at different levels of resolution; ability to view abstract representation	1. Login security: username and password 2. Creation of user account to submit and access dataset to system via web interface 3. Single sign-on capability (SSO)	1. Methods to improve data quality 2. Data clustering, classification, reduction 3. Integration of new data/content into the system's data store and data annotation	--

	Use Case	Data Sources	Data Transformation	Capabilities	Data Consumer	Security and Privacy	Life Cycle Management	Other
		structural, and hierarchical biological data 5. Metagenomic samples that can vary by several orders of magnitude, such as several hundred thousand genes to a billion genes			s of highly similar data			
21	M0140 Individualized Diabetes Management	1. Distributed EHR data 2. Over 5 million patients with thousands of properties each and many more derived from primary values 3. Each record: a range of 100 to 100,000 data property values, average of 100 controlled vocabulary values, and average of 1,000 continuous values 4. No real-time, but data updated periodically; data timestamped with	1. Data integration using ontological annotation and taxonomies 2. Parallel retrieval algorithms for both indexed and custom searches; identification of data of interest; patient cohorts, patients' meeting certain criteria, patients sharing similar characteristics 3. Distributed graph mining algorithms, pattern analysis and graph indexing, pattern searching on RDF triple graphs	1. data warehouse, open source indexed Hbase 2. supercomputers, cloud and parallel computing 3. I/O intensive processing 4. HDFS storage 5. custom code to develop new properties from stored data.	1. Efficient data graph-based visualization needed	1. Protection of health data in accordance with privacy policies and legal requirements, e.g., HIPAA. 2. Security policies for different user roles	1. Data annotated based on domain ontologies or taxonomies 2. Traceability of data from origin (initial point of collection) through use 3. Data conversion from existing data warehouse into RDF triples	1. Mobile access

	Use Case	Data Sources	Data Transformation	Capabilities	Data Consumer	Security and Privacy	Life Cycle Management	Other
		the time of observation (time the value is recorded) 5. Two main categories of structured data about a patient: data with controlled vocabulary (CV) property values and data with continuous property values (recorded/ captured more frequently) 6. Data consist of text and continuous numerical values	4. Robust statistical analysis tools to manage false discovery rates, determine true sub-graph significance, validate results, eliminate false positive/false negative results 5. Semantic graph mining algorithms to identify graph patterns, index and search graph 6. Semantic graph traversal					
22	M0174 Statistical Relational Artificial Intelligence for Health Care	1. Centralized data, with some data retrieved from Internet sources 2. Range from hundreds of GBs for a sample size to 1 PB for very large studies 3. Both constant updates/additions (to data subsets)	1. Relational probabilistic models/ probability theory; software that learns models from multiple data types and can possibly integrate the information and reason about complex queries 2. Robust and	1. Java, some in house tools, [relational] database and NoSQL stores 2. Cloud and parallel computing 3. High-performance computer, 48 GB RAM (to perform analysis for a	1. Visualization of very large data subsets	1. Secure handling and processing of data	1. Merging multiple tables before analysis 2. Methods to validate data to minimize errors	--

	Use Case	Data Sources	Data Transformation	Capabilities	Data Consumer	Security and Privacy	Life Cycle Management	Other
		and scheduled batch inputs 4. Large, multi-modal, longitudinal data 5. Rich relational data comprising multiple tables, different data types such as imaging, EHR, demographic, genetic, and natural language data requiring rich representation 6. Unpredictable arrival rates, often real time	accurate learning methods to account for data imbalance (where large numbers of data are available for a small number of subjects) 3. Learning algorithms to identify skews in data, so as to not to (incorrectly) model noise 4. Generalized and refined learned models for application to diverse sets of data 5. Challenge: acceptance of data in different modalities (and from disparate sources)	moderate sample size) 4. Dlusters for large datasets 5. 200 GB–1 TB hard drive for test data				
23	M0172 World Population Scale Epidemiological Study	1. File-based synthetic population, either centralized or distributed sites 2. Large volume of real-time output data 3. Variety of output datasets	1. Compute-intensive and data-intensive computation, like supercomputer performance 2. Unstructured and irregular nature of graph processing	1. Movement of very large volume of data for visualization (networking) 2. Distributed MPI-based simulation system (platform) 3. Charm++ on	1. Visualization	1. Protection of PII on individuals used in modeling 2. Data protection and secure platform for computation	1. Data quality, ability to capture the traceability of quality from computation	--

	Use Case	Data Sources	Data Transformation	Capabilities	Data Consumer	Security and Privacy	Life Cycle Management	Other
		depending on the model's complexity	3. Summary of various runs of simulation	multi-nodes (software) 4. Network file system (storage) 5. Infiniband network (networking)				
24	M0173 Social Contagion Modeling for Planning	1. Traditional and new architecture for dynamic distributed processing on commodity clusters 2. Fine-resolution models and datasets to support Twitter network traffic 3. Huge data storage supporting annual data growth	1. Large-scale modeling for various events (disease, emotions, behaviors, etc.) 2. Scalable fusion between combined datasets 3. Multilevel analysis while generating sufficient results quickly	1. Computing infrastructure that can capture human-to-human interactions on various social events via the Internet (infrastructure) 2. File servers and databases (platform) 3. Ethernet and Infiniband networking (networking) 4. Specialized simulators, open source software, and proprietary modeling (application) 5. Huge user accounts across country boundaries (networking)	1. Multilevel detailed network representations 2. Visualization with interactions	1. Protection of PII of individuals used in modeling 2. Data protection and secure platform for computation	1. Data fusion from variety of data sources (i.e., Stata data files) 2. Data consistency and no corruption 3. Preprocessing of raw data	1. Efficient method of moving data

	Use Case	Data Sources	Data Transformation	Capabilities	Data Consumer	Security and Privacy	Life Cycle Management	Other
25	M0141 Biodiversity and LifeWatch	1. Special dedicated or overlay sensor network 2. Storage: distributed, historical, and trends data archiving 3. Distributed data sources, including observation and monitoring facilities, sensor network, and satellites 4. Wide variety of data: satellite images/ information, climate and weather data, photos, video, sound recordings, etc. 5. Multi-type data combination and linkage, potentially unlimited data variety 6. Data streaming	1. Web-based services, grid-based services, relational databases, NoSQL 2. Personalized virtual labs 3. Grid- and cloud-based resources 4. Data analyzed incrementally and/or in real time at varying rates owing to variations in source processes 5. A variety of data and analytical and modeling tools to support analytics for diverse scientific communities 6. Parallel data streams and streaming analytics 7. Access and integration of multiple distributed databases	1. Expandable on-demand-based storage resource for global users 2. Cloud community resource required	1. Access by mobile users 2. Advanced/ rich/high-definition visualization 3. 4D visualization computational models	1. Federated identity management for mobile researchers and mobile sensors 2. Access control and accounting	1. Data storage and archiving, data exchange and integration 2. Data life cycle management: data provenance, referral integrity and identification traceability back to initial observational data 3. Processed (secondary) data storage (in addition to original source data) for future uses 4. Provenance (and persistent identification [PID]) control of data, algorithms, and workflows 5. Curated (authorized) reference data (e.g. species name lists), algorithms, software code, workflows	--
26	M0136 Large-Scale	--	--	1. GPU 2. High-	--	--	--	--

	Use Case	Data Sources	Data Transformation	Capabilities	Data Consumer	Security and Privacy	Life Cycle Management	Other
	Deep Learning			performance MPI and HPC Infiniband cluster 3. Libraries for single-machine or single-GPU computation – available (e.g., BLAS, CuBLAS, MAGMA, etc.); distributed computation of dense BLAS-like or LAPACK-like operations on GPUs – poorly developed; existing solutions (e.g., ScaLapack for CPUs) – not well-integrated with higher-level languages and require low-level programming, lengthening experiment and development time				
27	**M0171** Organizing Large-Scale Unstructured Collections	1. Over 500 million images uploaded to social media sites each day	1. Classifier (e.g. an SVM), a process that is often hard to parallelize 2. Features seen in many large-scale	1. Hadoop or enhanced Map/Reduce	1. Visualize large-scale 3D reconstructions; navigate large-scale collections of	1. Preserve privacy for users and digital rights for media	--	--

	Use Case	Data Sources	Data Transformation	Capabilities	Data Consumer	Security and Privacy	Life Cycle Management	Other
	of Consumer Photos		image processing problems		images that have been aligned to maps			
28	M0160 Truthy Twitter Data	1. Distributed data sources 2. Large volume of real-time streaming data 3. Raw data in compressed formats 4. Fully structured data in JSON, user metadata, geo-location data 5. Multiple data schemas	1. Various real-time data analysis for anomaly detection, stream clustering, signal classification on multi-dimensional time series, online learning	1. Hadoop and HDFS (platform) 2. IndexedHBase, Hive, SciPy, NumPy (software) 3. In-memory database, MPI (platform) 4. High-speed Infiniband network (networking)	1. Data retrieval and dynamic visualization 2. Data-driven interactive web interfaces 3. API for data query	1. Security and privacy policy	1. Standardized data structures/ formats with extremely high data quality	1. Low-level data storage infrastructure for efficient mobile access to data
29	M0211 Crowd Sourcing in Humanities	--	1. Digitize existing audio-video, photo, and documents archives 2. Analytics: pattern recognition of all kinds (e.g., speech recognition, automatic A&V analysis, cultural patterns), identification of structures (lexical units, linguistic rules, etc.)	--	--	1. Privacy issues in preserving anonymity of responses in spite of computer recording of access ID and reverse engineering of unusual user responses	--	--

	Use Case	Data Sources	Data Transformation	Capabilities	Data Consumer	Security and Privacy	Life Cycle Management	Other
30	M0158 CINET for Network Science	1. A set of network topologies files to study graph theoretic properties and behaviors of various algorithms 2. Asynchronous and real-time synchronous distributed computing	1. Environments to run various network and graph analysis tools 2. Dynamic growth of the networks 3. Asynchronous and real-time synchronous distributed computing 4. Different parallel algorithms for different partitioning schemes for efficient operation	1. Large file system (storage) 2. Various network connectivity (networking) 3. Existing computing cluster 4. EC2 computing cluster 5. Various graph libraries, management tools, databases, semantic web tools	1. Client-side visualization	--	--	--
31	M0190 NIST Information Access Division	1. Large amounts of semi-annotated web pages, tweets, images, video 2. Scaling ground-truthing to larger data, intrinsic and annotation uncertainty measurement, performance measurement for incompletely annotated data, measuring analytic performance for	1. Test analytic algorithms working with written language, speech, human imagery, etc. against real or realistic data; challenge: engineering artificial data that sufficiently captures the variability of real data involving humans	1. PERL, Python, C/C++, Matlab, R development tools; creation of ground-up test and measurement applications	1. Analytic flows involving users	1. Security requirements for protecting sensitive data while enabling meaningful developmental performance evaluation; shared evaluation testbeds that protect the intellectual property of analytic algorithm developers	--	--

	Use Case	Data Sources	Data Transformation	Capabilities	Data Consumer	Security and Privacy	Life Cycle Management	Other
		heterogeneous data and analytic flows involving users						
32	**M0130** DataNet (iRODS)	1. Process key format types NetCDF, HDF5, Dicom 2. Real-time and batch data	1. Provision of general analytics workflows needed	1. iRODS data management software 2. interoperability across storage and network protocol types	1. General visualization workflows	1. Federate across existing authentication environments through Generic Security Service API and pluggable authentication modules (GSI, Kerberos, InCommon, Shibboleth) 2. Access controls on files independent of the storage location	--	--
33	**M0163** The Discinnet Process	1. Integration of metadata approaches across disciplines	--	1. Software: Symfony-PHP, Linux, MySQL	--	1. Significant but optional security and privacy including secure servers and anonymization	1. Integration of metadata approaches across disciplines	--
34	**M0131** Semantic Graph-Search	1. All data types, image to text, structures to protein sequence	1. Data graph processing 2. RDBMS	1. Cloud community resource required	1. Efficient data-graph-based	--	--	--

	Use Case	Data Sources	Data Transformation	Capabilities	Data Consumer	Security and Privacy	Life Cycle Management	Other
					visualization needed			
35	**M0189** Light source beamlines	1. Multiple streams of real-time data to be stored and analyzed later 2. Sample data to be analyzed in real time	1. Standard bioinformatics tools (BLAST, HMMER, multiple alignment and phylogenetic tools, gene callers, sequence feature predictors, etc.), Perl/Python wrapper scripts, Linux Cluster scheduling	1. High-volume data transfer to remote batch processing resource	--	1. Multiple security and privacy requirements to be satisfied	--	--
36	**M0170** Catalina Real-Time Transient Survey	1. ≈0.1 TB per day at present, will increase by factor of 100	1. A wide variety of the existing astronomical data analysis tools, plus a large number of custom developed tools and software programs, some research projects in and of themselves 2. Automated classification with machine learning tools given the very sparse and heterogeneous data, dynamically evolving in time as more data come in,	--	1. Visualization mechanisms for highly dimensional data parameter spaces	--	--	--

	Use Case	Data Sources	Data Transformation	Capabilities	Data Consumer	Security and Privacy	Life Cycle Management	Other
			with follow-up decision making reflecting limited follow-up resources					
37	**M0185** DOE Extreme Data from Cosmological Sky Survey	1. ≈1 PB/year becoming 7 PB/year of observational data	1. Advanced analysis and visualization techniques and capabilities to support interpretation of results from detailed simulations	1. MPI, OpenMP, C, C++, F90, FFTW, viz packages, Python, FFTW, numpy, Boost, OpenMP, ScaLAPCK, PSQL and MySQL databases, Eigen, cfitsio, astrometry.net, and Minuit2 2. Methods/ tools to address supercomputer I/O subsystem limitations	1. Interpretation of results using advanced visualization techniques and capabilities	--	--	--
38	**M0209** Large Survey Data for Cosmology	1. 20 TB of data/day	1. Analysis on both the simulation and observational data simultaneously 2. Techniques for handling Cholesky decomposition for thousands of simulations with matrices of order 1 million on a side	1. Standard astrophysics reduction software as well as Perl/Python wrapper scripts 2. Oracle RDBMS, Postgres psql, GPFS and Lustre file systems and tape archives 3. Parallel image storage	--	--	1. Links between remote telescopes and central analysis sites	--

	Use Case	Data Sources	Data Transformation	Capabilities	Data Consumer	Security and Privacy	Life Cycle Management	Other
39	M0166 Particle Physics at LHC	1. Real-time data from accelerator and analysis instruments 2. Asynchronization data collection 3. Calibration of instruments	1. Experimental data from ALICE, ATLAS, CMS, LHB 2. Histograms, scatter-plots with model fits 3. Monte-Carlo computations	1. Legacy computing infrastructure (computing nodes) 2. Distributed cached files (storage) 3. Object databases (software package)	1. Histograms and model fits (visual)	1. Data protection	1. Data quality on complex apparatus	--
40	M0210 Belle II High-Energy Physics Experiment	1. 120 PB of raw data	--	1. 120 PB raw data 2. International distributed computing model to augment that at accelerator (Japan) 3. Data transfer of ≈20 GB/ second at designed luminosity between Japan and United States 4. Software from Open Science Grid, Geant4, DIRAC, FTS, Belle II framework	--	1. Standard grid authentication	--	--
41	M0155 EISCAT 3D Incoherent Scatter	1. Remote sites generating 40 PB data/year by 2022 2. Hierarchical	1. Queen Bea architecture with mix of distributed on-sensor and	1. Architecture compatible with ENVRI	1. Support needed for visualization of high-	--	1. Preservation of data and avoidance of lost data due to	1. Support needed for real-time monitoring of

	Use Case	Data Sources	Data Transformation	Capabilities	Data Consumer	Security and Privacy	Life Cycle Management	Other
	Radar System	Data Format (HDF5) 3. Visualization of high-dimensional (≥5) data	central processing for 5 distributed sites 2. Real-time monitoring of equipment by partial streaming analysis 3. Hosting needed for rich set of radar image processing services using machine learning, statistical modelling, and graph algorithms		dimensional (≥5) data		instrument malfunction	equipment by partial streaming analysis
42	**M0157** ENVRI Environmental Research Infrastructure	1. Huge volume of data from real-time distributed data sources 2. Variety of instrumentation datasets and metadata	1. Diversified analytics tools	1. Variety of computing infrastructures and architectures (infrastructure) 2. Scattered repositories (storage)	1. Graph plotting tools 2. Time series interactive tools 3. Brower-based flash playback 4. Earth high-resolution map display 5. Visual tools for quality comparisons	1. Open data policy with minor restrictions	1. High data quality 2. Mirror archives 3. Various metadata frameworks 4. Scattered repositories and data curation	1. Various kinds of mobile sensor devices for data acquisition
43	**M0167** CReSIS Remote Sensing	1. Provision of reliable data transmission from aircraft sensors/ instruments or	1. Legacy software (Matlab) and language (C/Java) binding for processing	1. ≈0.5 PB/year of raw data 2. Transfer content from removable disk to	1. GIS user interface 2. Rich user interface for simulations	1. Security and privacy on sensitive political issues 2. Dynamic	1. Data quality assurance	1. Monitoring data collection instruments/ sensors

	Use Case	Data Sources	Data Transformation	Capabilities	Data Consumer	Security and Privacy	Life Cycle Management	Other
		removable disks from remote sites 2. Data gathering in real time 3. Varieties of datasets	2. Signal processing and advanced image processing to find layers needed	computing cluster for parallel processing 3. Map/Reduce or MPI plus language binding for C/Java		security and privacy policy mechanisms		
44	**M0127** UAVSAR Data Processing	1. Angular and spatial data 2. Compatibility with other NASA radar systems and repositories (Alaska Satellite Facility)	1. Geolocated data that require GIS integration of data as custom overlays 2. Significant human intervention in data processing pipeline 3. Hosting of rich set of radar image processing services 4. ROI_PAC, GeoServer, GDAL, GeoTIFF-supporting tools	1. Support for interoperable Cloud-HPC architecture 2. Hosting of rich set of radar image processing services 3. ROI_PAC, GeoServer, GDAL, GeoTIFF-supporting tools 4. Compatibility with other NASA radar systems and repositories (Alaska Satellite Facility)	1. Support for field expedition users with phone/tablet interface and low-resolution downloads	--	1. Significant human intervention in data processing pipeline 2. Rich robust provenance defining complex machine/human processing	1. Support for field expedition users with phone/tablet interface and low-resolution downloads
45	**M0182** NASA LARC/ GSFC iRODS	1. Federate distributed heterogeneous datasets	1. CAaaS on clouds	1. Support virtual climate data server (vCDS) 2. GPFS parallel file system integrated with Hadoop 3. iRODS	1. Support needed to visualize distributed heterogeneous data	--	--	--
46	**M0129** MERRA	1. Integrate simulation output and observational	1. CAaaS on clouds	1. NetCDF aware software 2. Map/Reduce	1. High-end distributed visualization	--	--	1. Smart phone and tablet access

	Use Case	Data Sources	Data Transformation	Capabilities	Data Consumer	Security and Privacy	Life Cycle Management	Other
	Analytic Services	data, NetCDF files 2. Real-time and batch mode needed 3. Interoperable use of AWS and local clusters 4. iRODS data management		3. Interoperable use of AWS and local clusters				required 2. iRODS data management
47	M0090 Atmospheric Turbulence	1. Real-time distributed datasets 2. Various formats, resolution, semantics, and metadata	1. Map/Reduce, SciDB, and other scientific databases 2. Continuous computing for updates 3. Event specification language for data mining and event searching 4. Semantics interpretation and optimal structuring for 4D data mining and predictive analysis	1. Other legacy computing systems (e.g. supercomputer) 2. high throughput data transmission over the network	1. Visualization to interpret results	--	1. Validation for output products (correlations)	--
48	M0186 Climate Studies	1. ≈100 PB data in 2017 streaming at high data rates from large supercomputers across the world 2. Integration of large-scale distributed data	1. Data analytics close to data storage	1. Extension of architecture to several other fields	1. Worldwide climate data sharing 2. High-end distributed visualization	--	--	1. Phone-based input and access

	Use Case	Data Sources	Data Transformation	Capabilities	Data Consumer	Security and Privacy	Life Cycle Management	Other
		from simulations with diverse observations 3. Linking of diverse data to novel HPC simulation						
49	**M0183** DOE-BER Subsurface Biogeochem istry	1. Heterogeneous diverse data with different domains and scales, translation across diverse datasets that cross domains and scales 2. Synthesis of diverse and disparate field, laboratory, omic, and simulation datasets across different semantic, spatial, and temporal scales 3. Linking of diverse data to novel HPC simulation	--	1. Postgres, HDF5 data technologies, and many custom software systems	1. Phone-based input and access	--	--	1. Phone-based input and access
50	**M0184** DOE-BER AmeriFlux and	1. Heterogeneous diverse data with different domains and scales, translation across	1. Custom software such as EddyPro, and custom analysis software, such as	1. Custom software, such as EddyPro, and custom analysis software, such as	1. Phone-based input and access	--	--	1. Phone-based input and access

	Use Case	Data Sources	Data Transformation	Capabilities	Data Consumer	Security and Privacy	Life Cycle Management	Other
	FLUXNET Networks	diverse datasets that cross domains and scales 2. Link to many other environment and biology datasets 3. Link to HPC climate and other simulations 4. Link to European data sources and projects 5. Access to data from 500 distributed sources	R, Python, neural networks, Matlab	R, Python, neural networks, Matlab 2. Analytics including data mining, data quality assessment, cross-correlation across datasets, data assimilation, data interpolation, statistics, quality assessment, data fusion, etc.				
51	**M0223** Consumption Forecasting in Smart Grids	1. Diverse data from smart grid sensors, city planning, weather, utilities 2. Data updated every 15 minutes	1. New machine learning analytics to predict consumption	1. SQL databases, CVS files, HDFS (platform) 2. R/Matlab, Weka, Hadoop (platform)	--	1. Privacy and anonymization by aggregation	--	1. Mobile access for clients

Appendix D: Use Case Detail Requirements

This appendix contains the version 1 use case specific requirements and the aggregated general requirements within each of the following seven characteristic categories:

- Data sources
- Data transformation
- Capabilities
- Data consumer
- Security and privacy
- Life cycle management
- Other

Within each characteristic category, the general requirements are listed with the use cases to which that requirement applies. The use case IDs, in the form of MNNNN, contain links to the use case documents in the NIST document library (http://bigdatawg.nist.gov/usecases.php).

After the general requirements, the use case specific requirements for the characterization category are listed by use case. If requirements were not extracted from a use case for a particular characterization category, the use case will not be in this section of the table.

TABLE D-1: DATA SOURCES REQUIREMENTS	
GENERAL REQUIREMENTS	
Needs to support reliable real time, asynchronous, streaming, and batch processing to collect data from centralized, distributed, and cloud data sources, sensors, or instruments.	Applies to 28 use cases: M0078, M0090, M0103, M0127, M0129, M0140, M0141, M0147, M0148, M0157, M0160, M0160, M0162, M0165, M0166, M0166, M0167, M0172, M0173, M0174, M0176, M0177, M0183, M0184, M0186, M0188, M0191, M0215
Needs to support slow, bursty, and high-throughput data transmission between data sources and computing clusters.	Applies to 22 use cases: M0078, M0148, M0155, M0157, M0162, M0165, M0167, M0170, M0171, M0172, M0174, M0176, M0177, M0184, M0185, M0186, M0188, M0191, M0209, M0210, M0219, M0223
Needs to support diversified data content: structured and unstructured text, document, graph, web, geospatial, compressed, timed, spatial, multimedia, simulation, instrumental data.	Applies to 28 use cases: M0089, M0090, M0140, M0141, M0147, M0148, M0155, M0158, M0160, M0161, M0162, M0165, M0166, M0167, M0171, M0172, M0173, M0177, M0183, M0184, M0186, M0188, M0190, M0191, M0213, M0214, M0215, M0223
USE CASE SPECIFIC REQUIREMENTS FOR DATA SOURCES	
1	M0147 Census 2010 and 2000 • Needs to support large document format from a centralized storage.

TABLE D-1: DATA SOURCES REQUIREMENTS	
2	M0148 NARA: Search, Retrieve, Preservation • Needs to support distributed data sources. • Needs to support large data storage. • Needs to support bursty data ranging from a GB to hundreds of terabytes. • Needs to support a wide variety of data formats including unstructured and structured data. • Needs to support distributed data sources in different clouds.
3	M0219 Statistical Survey Response Improvement • Needs to support data size of approximately one petabyte.
5	M0175 Cloud Eco-System for Finance • Needs to support real-time ingestion of data.
6	M0161 Mendeley • Needs to support file-based documents with constant new uploads. • Needs to support a variety of file types such as PDFs, social network log files, client activities images, spreadsheets, presentation files.
7	M0164 Netflix Movie Service • Needs to support user profiles and ranking information.
8	M0165 Web Search • Needs to support distributed data sources • Needs to support streaming data. • Needs to support multimedia content.
10	M0103 Cargo Shipping • Needs to support centralized and real-time distributed sites/sensors.
11	M0162 Materials Data for Manufacturing • Needs to support distributed data repositories for more than 500,000 commercial materials. • Needs to support many varieties of datasets. • Needs to support text, graphics, and images.
12	M0176 Simulation-Driven Materials Genomics • Needs to support data streams from peta/exascale centralized simulation systems. • Needs to support distributed web dataflows from central gateway to users.
13	M0213 Large-Scale Geospatial Analysis and Visualization • Needs to support geospatial data that require unique approaches to indexing and distributed analysis.
14	M0214 Object identification and tracking • Needs to support real-time data FMV (30 to 60 frames per second at full-color 1080P resolution) and WALF (1 to 10 frames per second at 10,000 x 10,000 full-color resolution).
15	M0215 Intelligence Data Processing and Analysis • Needs to support real-time data with processing at (at worst) near-real time. • Needs to support data that currently exist in disparate silos that must be accessible through a semantically integrated data space. • Needs to support diverse data: text files, raw media, imagery, video, audio, electronic data, human-generated data.

TABLE D-1: DATA SOURCES REQUIREMENTS

16	M0177 EMR Data • Needs to support heterogeneous, high-volume, diverse data sources. • Needs to support volume of > 12 million entities (patients), > 4 billion records or data points (discrete clinical observations), aggregate of > 20 TB of raw data. • Needs to support velocity: 500,000 to 1.5 million new transactions per day. • Needs to support variety: formats include numeric, structured numeric, free-text, structured text, discrete nominal, discrete ordinal, discrete structured, binary large blobs (images and video). • Needs to support data that evolve in a highly variable fashion. • Needs to support a comprehensive and consistent view of data across sources and over time.
17	M0089 Pathology Imaging • Needs to support high-resolution spatial digitized pathology images. • Needs to support various image quality analysis algorithms. • Needs to support various image data formats, especially BigTIFF, with structured data for analytical results. • Needs to support image analysis, spatial queries and analytics, feature clustering, and classification.
18	M0191 Computational Bioimaging • Needs to support distributed multi-modal high-resolution experimental sources of bioimages (instruments). • Needs to support 50 TB of data in formats that include images.
19	M0078 Genomic Measurements • Needs to support high-throughput compressed data (300 GB per day) from various DNA sequencers. • Needs to support distributed data source (sequencers). • Needs to support various file formats for both structured and unstructured data.
20	M0188 Comparative Analysis for Metagenomes and Genomes • Needs to support multiple centralized data sources. • Needs to support proteins and their structural features, core genomic data, and new types of omics data such as transcriptomics, methylomics, and proteomics describing gene expression. • Needs to support front real-time web UI interactive. Backend data loading processing must keep up with the exponential growth of sequence data due to the rapid drop in cost of sequencing technology. • Needs to support heterogeneous, complex, structural, and hierarchical biological data. • Needs to support metagenomic samples that can vary by several orders of magnitude, such as several hundred thousand genes to a billion genes.
21	M0140 Individualized Diabetes Management • Needs to support distributed EHR data. • Needs to support over 5 million patients with thousands of properties each and many more that are derived from primary values. • Needs to support each record, a range of 100 to 100,000 data property values, an average of 100 controlled vocabulary values, and an average of 1,000 continuous values. • Needs to support data that are updated periodically (not real time). Data are timestamped with the time of observation (the time that the value is recorded). • Needs to support structured data about patients. The data fall into two main categories: data with controlled vocabulary (CV) property values and data with continuous property values (which are recorded/captured more frequently). • Needs to support data that consist of text and continuous numerical values.

TABLE D-1: DATA SOURCES REQUIREMENTS	
22	M0174 Statistical Relational Artificial Intelligence for Health Care • Needs to support centralized data, with some data retrieved from Internet sources. • Needs to support data ranging from hundreds of GBs for a sample size to one petabyte for very large studies. • Needs to support both constant updates/additions (to data subsets) and scheduled batch inputs. • Needs to support large, multi-modal, longitudinal data. • Needs to support rich relational data comprising multiple tables, as well as different data types such as imaging, EHR, demographic, genetic and natural language data requiring rich representation. • Needs to support unpredictable arrival rates; in many cases, data arrive in real-time.
23	M0172 World Population-Scale Epidemiological Study • Needs to support file-based synthetic populations on either centralized or distributed sites. • Needs to support a large volume of real-time output data. • Needs to support a variety of output datasets, depending on the complexity of the model.
24	M0173 Social Contagion Modeling for Planning • Needs to support traditional and new architecture for dynamic distributed processing on commodity clusters. • Needs to support fine-resolution models and datasets to support Twitter network traffic. • Needs to support huge data storage per year.
25	M0141 Biodiversity and LifeWatch • Needs to support special dedicated or overlay sensor network. • Needs to support storage for distributed, historical, and trends data archiving. • Needs to support distributed data sources and include observation and monitoring facilities, sensor network, and satellites. • Needs to support a wide variety of data, including satellite images/information, climate and weather data, photos, video, sound recordings, etc. • Needs to support multi-type data combinations and linkages with potentially unlimited data variety. • Needs to support data streaming.
27	M0171 Organizing Large-Scale Unstructured Collections of Consumer Photos • Needs to support over 500 million images uploaded to social media sites each day.
28	M0160 Truthy Twitter Data • Needs to support distributed data sources. • Needs to support large data volumes and real-time streaming. • Needs to support raw data in compressed formats. • Needs to support fully structured data in JSON, user metadata, and geo-location data. • Needs to support multiple data schemas.
30	M0158 CINET for Network Science • Needs to support a set of network topologies files to study graph theoretic properties and behaviors of various algorithms. • Needs to support asynchronous and real-time synchronous distributed computing.
31	M0190 NIST Information Access Division • Needs to support large amounts of semi-annotated web pages, tweets, images, and video. • Needs to support scaling of ground-truthing to larger data, intrinsic and annotation uncertainty measurement, performance measurement for incompletely annotated data, measurement of analytic performance for heterogeneous data, and analytic flows involving users.
32	M0130 DataNet (iRODS) • Needs to support process key format types: NetCDF, HDF5, Dicom. • Needs to support real-time and batch data.
33	M0163 The Discinnet Process • Needs to support integration of metadata approaches across disciplines.

TABLE D-1: DATA SOURCES REQUIREMENTS	
34	M0131 Semantic Graph-Search • Needs to support all data types, image to text, structures to protein sequence.
35	M0189 Light Source Beamlines • Needs to support multiple streams of real-time data to be stored and analyzed later. • Needs to support sample data to be analyzed in real time.
36	M0170 Catalina Real-Time Transient Survey • Needs to support ≈0.1 TB per day at present; the volume will increase by a factor of 100.
37	M0185 DOE Extreme Data from Cosmological Sky Survey • Needs to support ≈1 PB per year, becoming 7 PB per year, of observational data.
38	M0209 Large Survey Data for Cosmology • Needs to support 20 TB of data per day.
39	M0166 Particle Physics at LHC • Needs to support real-time data from accelerator and analysis instruments. • Needs to support asynchronization data collection. • Needs to support calibration of instruments.
40	M0210 Belle II High Energy Physics Experiment • Needs to support 120 PB of raw data.
41	M0155 EISCAT 3D Incoherent Scatter Radar System • Needs to support remote sites generating 40 PB of data per year by 2022. • Needs to support HDF5 data format. • Needs to support visualization of high-dimensional (≥5) data.
42	M0157 ENVRI Environmental Research Infrastructure • Needs to support a huge volume of data from real-time distributed data sources. • Needs to support a variety of instrumentation datasets and metadata.
43	M0167 CReSIS Remote Sensing • Needs to provide reliable data transmission from aircraft sensors/instruments or removable disks from remote sites. • Needs to support data gathering in real time. • Needs to support varieties of datasets.
44	M0127 UAVSAR Data Processing • Needs to support angular and spatial data. • Needs to support compatibility with other NASA radar systems and repositories (Alaska Satellite Facility).
45	M0182 NASA LARC/GSFC iRODS • Needs to support federated distributed heterogeneous datasets.
46	M0129 MERRA Analytic Services • Needs to support integration of simulation output and observational data, NetCDF files. • Needs to support real-time and batch mode. • Needs to support interoperable use of AWS and local clusters. • Needs to support iRODS data management.
47	M0090 Atmospheric Turbulence • Needs to support real-time distributed datasets. • Needs to support various formats, resolution, semantics, and metadata.

TABLE D-1: DATA SOURCES REQUIREMENTS	
48	M0186 Climate Studies • Needs to support ≈100 PB of data (in 2017) streaming at high data rates from large supercomputers across the world. • Needs to support integration of large-scale distributed data from simulations with diverse observations. • Needs to link diverse data to novel HPC simulation.
49	M0183 DOE-BER Subsurface Biogeochemistry • Needs to support heterogeneous diverse data with different domains and scales, and translation across diverse datasets that cross domains and scales. • Needs to support synthesis of diverse and disparate field, laboratory, omic, and simulation datasets across different semantic, spatial, and temporal scales. • Needs to link diverse data to novel HPC simulation.
50	M0184 DOE-BER AmeriFlux and FLUXNET Networks • Needs to support heterogeneous diverse data with different domains and scales, and translation across diverse datasets that cross domains and scales. • Needs to support links to many other environment and biology datasets. • Needs to support links to HPC for climate and other simulations. • Needs to support links to European data sources and projects. • Needs to support access to data from 500 distributed sources.
51	M0223 Consumption Forecasting in Smart Grids • Needs to support diverse data from smart grid sensors, city planning, weather, and utilities. • Needs to support data from updates every 15 minutes.

TABLE D-2: DATA TRANSFORMATION	
GENERAL REQUIREMENTS	
1. Needs to support diversified compute-intensive, analytic processing, and machine learning techniques.	Applies to 38 use cases: M0078, M0089, M0103, M0127, M0129, M0140, M0141, M0148, M0155, M0157, M0158, M0160, M0161, M0164, M0164, M0166, M0166, M0167, M0170, M0171, M0172, M0173, M0174, M0176, M0177, M0182, M0185, M0186, M0190, M0191, M0209, M0211, M0213, M0214, M0215, M0219, M0222, M0223
2. Needs to support batch and real-time analytic processing.	Applies to 7 use cases: M0090, M0103, M0141, M0155, M0164, M0165, M0188
3. Needs to support processing of large diversified data content and modeling.	Applies to 15 use cases: M0078, M0089, M0127, M0140, M0158, M0162, M0165, M0166, M0166, M0167, M0171, M0172, M0173, M0176, M0213
4, Needs to support processing of data in motion (streaming, fetching new content, tracking, etc.)	Applies to 6 use cases: M0078, M0090, M0103, M0164, M0165, M0166
USE CASE SPECIFIC REQUIREMENTS FOR DATA TRANSFORMATION	
1.	M0148 NARA: Search, Retrieve, Preservation **Transformation Requirements:** • Needs to support crawl and index from distributed data sources. • Needs to support various analytics processing including ranking, data categorization, and PII data detection. • Needs to support preprocessing of data. • Needs to support long-term preservation management of large varied datasets. Needs to support a huge amount of data with high relevancy and recall.
2.	M0219 Statistical Survey Response Improvement **Transformation Requirements:**

TABLE D-2: DATA TRANSFORMATION	
	• Needs to support analytics that are required for recommendation systems, continued monitoring, and general survey improvement.
3.	M0222 Non-Traditional Data in Statistical Survey Response Improvement **Transformation Requirements:** • Needs to support analytics to create reliable estimates using data from traditional survey sources, government administrative data sources, and non-traditional sources from the digital economy.
4.	M0175 Cloud Eco-System for Finance **Transformation Requirements:** • Needs to support real-time analytics.
5.	M0161 Mendeley **Transformation Requirements:** • Needs to support standard machine learning and analytics libraries. • Needs to support efficient scalable and parallelized ways of matching between documents. • Needs to support third-party annotation tools or publisher watermarks and cover pages.
6.	M0164 Netflix Movie Service **Transformation Requirements:** • Needs to support streaming video contents to multiple clients. • Needs to support analytic processing for matching client interest in movie selection. • Needs to support various analytic processing techniques for consumer personalization. • Needs to support robust learning algorithms. • Needs to support continued analytic processing based on the monitoring and performance results.
7.	M0165 Web Search **Transformation Requirements:** • Needs to support dynamic fetching content over the network. • Needs to link user profiles and social network data.
8.	M0137 Business Continuity and Disaster Recovery within a Cloud Eco-System **Transformation Requirements:** • Needs to support a robust backup algorithm. • Needs to replicate recent changes.
9.	M0103 Cargo Shipping **Transformation Requirements:** • Needs to support item tracking based on unique identification using an item's sensor information and GPS coordinates. • Needs to support real-time updates on tracking items.
10.	M0162 Materials Data for Manufacturing **Transformation Requirements:** • Needs to support hundreds of independent variables by collecting these variables to create robust datasets.
11.	M0176 Simulation-Driven Materials Genomics **Transformation Requirements:** • Needs to support high-throughput computing real-time data analysis for web-like responsiveness. • Needs to support mashup of simulation outputs across codes. • Needs to support search and crowd-driven functions with computation backend flexibility for new targets. • Needs to support Map/Reduce and search functions to join simulation and experimental data.
12.	M0213 Large-Scale Geospatial Analysis and Visualization **Transformation Requirements:** • Needs to support analytics including closest point of approach, deviation from route, point density over time, PCA, and ICA. • Needs to support geospatial data that require unique approaches to indexing and distributed analysis.
13.	M0214 Object Identification and Tracking **Transformation Requirements:** • Needs to support rich analytics with object identification, pattern recognition, crowd behavior, economic activity, and data fusion.
14.	M0215 Intelligence Data Processing and Analysis **Transformation Requirements:** • Needs to support analytics including NRT alerts based on patterns and baseline changes.
15.	M0177 EMR Data **Transformation Requirements:**

TABLE D-2: DATA TRANSFORMATION

	• Needs to support a comprehensive and consistent view of data across sources and over time. • Needs to support analytic techniques: information retrieval, natural language processing, machine learning decision models, maximum likelihood estimators, and Bayesian networks.
16.	M0089 Pathology Imaging **Transformation Requirements:** • Needs to support high-performance image analysis to extract spatial information. • Needs to support spatial queries and analytics, and feature clustering and classification. • Needs to support analytic processing on a huge multi-dimensional dataset and be able to correlate with other data types such as clinical data and omic data.
17.	M0191 Computational Bioimaging **Transformation Requirements:** • Needs to support high-throughput computing with responsive analysis. • Needs to support segmentation of regions of interest; crowd-based selection and extraction of features; and object classification, organization, and search. • Needs to support advanced biosciences discovery through Big Data techniques/extreme-scale computing, in-database processing and analytics, machine learning (SVM and RF) for classification and recommendation services, advanced algorithms for massive image analysis, and high-performance computational solutions. • Needs to support massive data analysis toward massive imaging datasets.
18.	M0078 Genomic Measurements **Transformation Requirements:** • Needs to support processing of raw data in variant calls. • Needs to support machine learning for complex analysis on systematic errors from sequencing technologies, which are hard to characterize.
19.	M0188 Comparative Analysis for Metagenomes and Genomes **Transformation Requirements:** • Needs to support sequencing and comparative analysis techniques for highly complex data. • Needs to support descriptive statistics.
20.	M0140 Individualized Diabetes Management **Transformation Requirements:** • Needs to support data integration using ontological annotation and taxonomies. • Needs to support parallel retrieval algorithms for both indexed and custom searches and the ability to identify data of interest. Potential results include patient cohorts, patients meeting certain criteria, and patients sharing similar characteristics. • Needs to support distributed graph mining algorithms, pattern analysis and graph indexing, and pattern searching on RDF triple graphs. • Needs to support robust statistical analysis tools to manage false discovery rates, determine true sub-graph significance, validate results, and eliminate false positive/false negative results. • Needs to support semantic graph mining algorithms to identify graph patterns, index, and search graphs. • Needs to support semantic graph traversal.
21.	M0174 Statistical Relational Artificial Intelligence for Health Care **Transformation Requirements:** • Needs to support relational probabilistic models/probability theory. The software learns models from multiple data types and can possibly integrate the information and reason about complex queries. • Needs to support robust and accurate learning methods to account for data imbalance, i.e., situations in which large amounts of data are available for a small number of subjects. • Needs to support learning algorithms to identify skews in data, so as to not—incorrectly—model noise. • Needs to support learned models that can be generalized and refined to be applied to diverse sets of data. • Needs to support acceptance of data in different modalities and from disparate sources.
22.	M0172 World Population-Scale Epidemiological Study **Transformation Requirements:** • Needs to support compute-intensive and data-intensive computation, like a supercomputer's performance. • Needs to support the unstructured and irregular nature of graph processing. • Needs to support summaries of various runs of simulation.

TABLE D-2: DATA TRANSFORMATION	
23.	M0173 Social Contagion Modeling for Planning **Transformation Requirements:** • Needs to support large-scale modeling for various events (disease, emotions, behaviors, etc.). • Needs to support scalable fusion between combined datasets. • Needs to support multilevels analysis while generating sufficient results quickly.
24.	M0141 Biodiversity and LifeWatch **Transformation Requirements:** • Needs to support incremental and/or real-time data analysis; rates vary because of variations in source processes. • Needs to support a variety of data, analytical, and modeling tools to support analytics for diverse scientific communities. • Needs to support parallel data streams and streaming analytics. • Needs to support access and integration of multiple distributed databases.
25.	M0171 Large-Scale Deep Learning **Transformation Requirements:** • Needs to support classifier (e.g., an SVM), a process that is often hard to parallelize. • Needs to support features seen in many large-scale image processing problems.
26.	M0160 Truthy Twitter Data **Transformation Requirements:** • Needs to support various real-time data analyses for anomaly detection, stream clustering, signal classification on multi-dimensional time series, and online learning.
27.	M0211 Crowd Sourcing in Humanities **Transformation Requirements:** • Needs to support digitization of existing audio-video, photo, and document archives. • Needs to support analytics including pattern recognition of all kinds (e.g., speech recognition, automatic A&V analysis, cultural patterns) and identification of structures (lexical units, linguistics rules, etc.).
28.	M0158 CINET for Network Science **Transformation Requirements:** • Needs to support environments to run various network and graph analysis tools. • Needs to support dynamic growth of the networks. • Needs to support asynchronous and real-time synchronous distributed computing. • Needs to support different parallel algorithms for different partitioning schemes for efficient operation.
29.	M0190 NIST Information Access Division **Transformation Requirements:** • Needs to support analytic algorithms working with written language, speech, human imagery, etc. The algorithms generally need to be tested against real or realistic data. It is extremely challenging to engineer artificial data that sufficiently capture the variability of real data involving humans.
30.	M0130 DataNet (iRODS) **Transformation Requirements:** • Needs to provide general analytics workflows.
31.	M0131 Semantic Graph-Search **Transformation Requirements:** • Needs to support data graph processing. • Needs to support RDBMS.
32.	M0189 Light Source Beamlines **Transformation Requirements:** • Needs to support standard bioinformatics tools (BLAST, HMMER, multiple alignment and phylogenetic tools, gene callers, sequence feature predictors, etc.), Perl/Python wrapper scripts, and Linux Cluster scheduling.
33.	M0170 Catalina Real-Time Transient Survey **Transformation Requirements:** • Needs to support a wide variety of the existing astronomical data analysis tools, plus a large number of custom-developed tools and software programs, some of which are research projects in and of themselves. • Needs to support automated classification with machine learning tools given very sparse and heterogeneous data, dynamically evolving as more data are generated, with follow-up decision making reflecting limited follow up resources.
34.	M0185 DOE Extreme Data from Cosmological Sky Survey **Transformation Requirements:**

TABLE D-2: DATA TRANSFORMATION	
	• Needs to support interpretation of results from detailed simulations. Interpretation requires advanced analysis and visualization techniques and capabilities.
35.	M0209 Large Survey Data for Cosmology **Transformation Requirements:** • Needs to support analysis on both the simulation and observational data simultaneously. • Needs to support techniques for handling Cholesky decomposition for thousands of simulations with matrices of order 1 million on a side.
36.	M0166 Particle Physics at LHC **Transformation Requirements:** • Needs to support experimental data from ALICE, ATLAS, CMS, and LHb. • Needs to support histograms and scatter-plots with model fits. • Needs to support Monte Carlo computations.
37.	M0155 EISCAT 3D Incoherent Scatter Radar System **Transformation Requirements:** • Needs to support Queen Bea architecture with mix of distributed on-sensor and central processing for 5 distributed sites. • Needs to support real-time monitoring of equipment by partial streaming analysis. • Needs to host rich set of radar image processing services using machine learning, statistical modelling, and graph algorithms.
38.	M0157 ENVRI Environmental Research Infrastructure **Transformation Requirements:** • Needs to support diversified analytics tools.
39.	M0167 CReSIS Remote Sensing **Transformation Requirements:** • Needs to support legacy software (Matlab) and language (C/Java) binding for processing. • Needs signal processing and advanced image processing to find layers.
40.	M0127 UAVSAR Data Processing **Transformation Requirements:** • Needs to support geolocated data that require GIS integration of data as custom overlays. • Needs to support significant human intervention in data-processing pipeline. • Needs to host rich sets of radar image processing services. • Needs to support ROI_PAC, GeoServer, GDAL, and GeoTIFF-supporting tools.
41.	M0182 NASA LARC/GSFC iRODS **Transformation Requirements:** • Needs to support CAaaS on clouds.
42.	M0129 MERRA Analytic Services **Transformation Requirements:** • Needs to support CAaaS on clouds.
43.	M0090 Atmospheric Turbulence **Transformation Requirements:** • Needs to support Map/Reduce, SciDB, and other scientific databases. • Needs to support continuous computing for updates. • Needs to support event specification language for data mining and event searching. • Needs to support semantics interpretation and optimal structuring for 4D data mining and predictive analysis.
44.	M0186 Climate Studies **Transformation Requirements:** • Needs to support data analytics close to data storage.
45.	M0184 DOE-BER AmeriFlux and FLUXNET Networks **Transformation Requirements:** • Needs to support custom software, such as EddyPro, and custom analysis software, such as R, python, neural networks, Matlab.
46.	M0223 Consumption Forecasting in Smart Grids Transformation Requirements: • Needs to support new machine learning analytics to predict consumption.

TABLE D-3: CAPABILITIES
GENERAL REQUIREMENTS

TABLE D-3: CAPABILITIES

1. Needs to support legacy and advanced software packages (subcomponent: SaaS).	Applies to 30 use cases: M0078, M0089, M0127, M0136, M0140, M0141, M0158, M0160, M0161, M0164, M0164, M0166, M0167, M0172, M0173, M0174, M0176, M0177, M0183, M0188, M0191, M0209, M0210, M0212, M0213, M0214, M0215, M0219, M0219, M0223
2. Needs to support legacy and advanced computing platforms (subcomponent: PaaS).	Applies to 17 use cases: M0078, M0089, M0127, M0158, M0160, M0161, M0164, M0164, M0171, M0172, M0173, M0177, M0182, M0188, M0191, M0209, M0223
3. Needs to support legacy and advanced distributed computing clusters, co-processors, and I/O processing (subcomponent: IaaS).	Applies to 24 use cases: M0015, M0078, M0089, M0090, M0129, M0136, M0140, M0141, M0155, M0158, M0161, M0164, M0164, M0166, M0167, M0173, M0174, M0176, M0177, M0185, M0186, M0191, M0214, M0215
4. Needs to support elastic data transmission (subcomponent: networking).	Applies to 4 use cases: M0089, M0090, M0103, M0136, M0141, M0158, M0160, M0172, M0173, M0176, M0191, M0210, M0214, M0215
5. Needs to support legacy, large, and advanced distributed data storage (subcomponent: storage).	Applies to 35 use cases: M0078, M0089, M0127, M0140, M0147, M0147, M0148, M0148, M0155, M0157, M0157, M0158, M0160, M0161, M0164, M0164, M0165, M0166, M0167, M0c170, M0171, M0172, M0173, M0174, M0176, M0176, M0182, M0185, M0188, M0209, M0209, M0210, M0210, M0215, M0219
6. Needs to support legacy and advanced executable programming: applications, tools, utilities, and libraries.	Applies to 13 use cases: M0078, M0089, M0140, M0164, M0c166, M0167, M0174, M0176, M0184, M0185, M0190, M0214, M0215

USE CASE SPECIFIC REQUIREMENTS FOR CAPABILITIES

1. M0147 Census 2010 and 2000 **Capability Requirements:**
   - Needs to support large centralized storage.
2. M0148 NARA: Search, Retrieve, Preservation **Capability Requirements:**
   - Needs to support large data storage.
   - Needs to support various storages such as NetApps, Hitachi, and magnetic tapes.
3. M0219 Statistical Survey Response Improvement **Capability Requirements:**
   - Needs to support the following software: Hadoop, Spark, Hive, R, SAS, Mahout, Allegrograph, MySQL, Oracle, Storm, BigMemory, Cassandra, and Pig.
4. M0222 Non-Traditional Data in Statistical Survey Response Improvement **Capability Requirements:**
   - Needs to support the following software: Hadoop, Spark, Hive, R, SAS, Mahout, Allegrograph, MySQL, Oracle, Storm, BigMemory, Cassandra, and Pig.
5. M0161 Mendeley **Capability Requirements:**
   - Needs to support EC2 with HDFS (infrastructure).
   - Needs to support S3 (storage).
   - Needs to support Hadoop (platform).
   - Needs to support Scribe, Hive, Mahout, and Python (language).
   - Needs to support moderate storage (15 TB with 1 TB/month).
   - Needs to support batch and real-time processing.

TABLE D-3: CAPABILITIES	
6.	M0164 Netflix Movie Service **Capability Requirements:** • Needs to support Hadoop (platform). • Needs to support Pig (language). • Needs to support Cassandra and Hive. • Needs to support a huge volume of subscribers, ratings, and searches per day (DB). • Needs to support huge storage (2 PB). • Needs to support I/O-intensive processing.
7.	M0165 Web Search **Capability Requirements:** • Needs to support petabytes of text and rich media (storage).
8.	M0137 Business Continuity and Disaster Recovery within a Cloud Eco-System **Capability Requirements:** • Needs to support Hadoop. • Needs to support commercial cloud services.
9.	M0103 Cargo Shipping **Capability Requirements:** • Needs to support Internet connectivity.
10.	M0176 Simulation-Driven Materials Genomics **Capability Requirements:** • Needs to support massive (150,000 cores) of legacy infrastructure (infrastructure). • Needs to support GPFS (storage). • Needs to support MonogDB systems (platform). • Needs to support 10 GB of networking data. • Needs to support various analytic tools such as PyMatGen, FireWorks, VASP, ABINIT, NWChem, BerkeleyGW, and varied community codes. • Needs to support large storage (storage). • Needs to support scalable key-value and object store (platform). • Needs to support data streams from peta/exascale centralized simulation systems.
11.	M0213 Large-Scale Geospatial Analysis and Visualization **Capability Requirements:** • Needs to support geospatially enabled RDBMS and geospatial server/analysis software (ESRI ArcServer, Geoserver).
12.	M0214 Object Identification and Tracking **Capability Requirements:** • Needs to support a wide range of custom software and tools including traditional RDBMS and display tools. • Needs to support several network capability requirements. • Needs to support GPU usage.
13.	M0215 Intelligence Data Processing and Analysis **Capability Requirements:** • Needs to support tolerance of unreliable networks to warfighter and remote sensors. • Needs to support up to hundreds of petabytes of data supported by modest to large clusters and clouds. • Needs to support the following software: Hadoop, Accumulo (Big Table), Solr, NLP (several variants), Puppet (for deployment and security), Storm, and custom applications and visualization tools.
14.	M0177 EMR Data **Capability Requirements:** • Needs to support Hadoop, Hive, and R Unix-based. • Needs to support a Cray supercomputer. • Needs to support teradata, PostgreSQL, MongoDB. • Needs to support various capabilities with significant I/O-intensive processing.
15.	M0089 Pathology Imaging **Capability Requirements:** • Needs to support legacy systems and clouds (computing cluster). • Needs to support huge legacy and new storage such as SAN or HDFS (storage). • Needs to support high-throughput network links (networking). • Needs to support MPI image analysis, Map/Reduce, and Hive with spatial extension (software packages).

TABLE D-3: CAPABILITIES	
16.	M0191 Computational Bioimaging **Capability Requirements:** • Needs to support ImageJ, OMERO, VolRover, advanced segmentation, and feature detection methods from applied math researchers. Scalable key-value and object store databases are needed. • Needs to support NERSC's Hopper infrastructure • Needs to support database and image collections. • Needs to support 10 GB and future 100 GB and advanced networking (SDN).
17.	M0078 Genomic Measurements **Capability Requirements:** • Needs to support legacy computing cluster and other PaaS and IaaS (computing cluster). • Needs to support huge data storage in the petabyte range (storage). • Needs to support Unix-based legacy sequencing bioinformatics software (software package).
18.	M0188 Comparative Analysis for Metagenomes and Genomes **Capability Requirements:** • Needs to support huge data storage. • Needs to support scalable RDBMS for heterogeneous biological data. • Needs to support real-time rapid and parallel bulk loading. • Needs to support Oracle RDBMS, SQLite files, flat text files, Lucy (a version of Lucene) for keyword searches, BLAST databases, and USEARCH databases. • Needs to support Linux cluster, Oracle RDBMS server, large memory machines, and standard Linux interactive hosts.
19.	M0140 Individualized Diabetes Management **Capability Requirements:** • Needs to support a data warehouse, specifically open source indexed Hbase. • Needs to support supercomputers with cloud and parallel computing. • Needs to support I/O-intensive processing. • Needs to support HDFS storage. • Needs to support custom code to develop new properties from stored data.
20.	M0174 Statistical Relational Artificial Intelligence for Health Care **Capability Requirements:** • Needs to support Java, some in-house tools, a relational database, and NoSQL stores. • Needs to support cloud and parallel computing. • Needs to support a high-performance computer with 48 GB RAM (to perform analysis for a moderate sample size). • Needs to support clusters for large datasets. • Needs to support 200 GB to 1 TB hard drive for test data.
21.	M0172 World Population-Scale Epidemiological Study **Capability Requirements:** • Needs to support movement of very large numbers of data for visualization (networking). • Needs to support distributed an MPI-based simulation system (platform). • Needs to support Charm++ on multi-nodes (software). • Needs to support a network file system (storage). • Needs to support an Infiniband network (networking).
22.	M0173 Social Contagion Modeling for Planning **Capability Requirements:** • Needs to support a computing infrastructure that can capture human-to-human interactions on various social events via the Internet (infrastructure). • Needs to support file servers and databases (platform). • Needs to support Ethernet and Infiniband networking (networking). • Needs to support specialized simulators, open source software, and proprietary modeling (application). • Needs to support huge user accounts across country boundaries (networking).
23.	M0141 Biodiversity and LifeWatch **Capability Requirements:** • Needs to support expandable on-demand-based storage resources for global users. • Needs to support cloud community resources.

TABLE D-3: CAPABILITIES

24.	M0136 Large-scale Deep Learning **Capability Requirements:** • Needs to support GPU usage. • Needs to support a high-performance MPI and HPC Infiniband cluster. • Needs to support libraries for single-machine or single-GPU computation (e.g., BLAS, CuBLAS, MAGMA, etc.). • Needs to support distributed computation of dense BLAS-like or LAPACK-like operations on GPUs, which remains poorly developed. Existing solutions (e.g., ScaLapack for CPUs) are not well integrated with higher-level languages and require low-level programming, which lengthens experiment and development time.
25.	M0171 Organizing Large-Scale Unstructured Collections of Consumer Photos **Capability Requirements:** • Needs to support Hadoop or enhanced Map/Reduce.
26.	M0160 Truthy Twitter Data **Capability Requirements:** • Needs to support Hadoop and HDFS (platform). • Needs to support IndexedHBase, Hive, SciPy, and NumPy (software). • Needs to support in-memory database and MPI (platform). • Needs to support high-speed Infiniband network (networking).
27.	M0158 CINET for Network Science **Capability Requirements:** • Needs to support a large file system (storage). • Needs to support various network connectivity (networking). • Needs to support an existing computing cluster. • Needs to support an EC2 computing cluster. • Needs to support various graph libraries, management tools, databases, and semantic web tools.
28.	M0190 NIST Information Access Division **Capability Requirements:** • Needs to support PERL, Python, C/C++, Matlab, and R development tools. • Needs to support creation of a ground-up test and measurement applications.
29.	M0130 DataNet (iRODS) **Capability Requirements:** • Needs to support iRODS data management software. • Needs to support interoperability across storage and network protocol types.
30.	M0163 The Discinnet Process **Capability Requirements:** • Needs to support the following software: Symfony-PHP, Linux, and MySQL.
31.	M0131 Semantic Graph-Search **Capability Requirements:** • Needs to support a cloud community resource.
32.	M0189 Light Source Beamlines **Capability Requirements:** • Needs to support high-volume data transfer to a remote batch processing resource.
33.	M0185 DOE Extreme Data from Cosmological Sky Survey **Capability Requirements:** • Needs to support MPI, OpenMP, C, C++, F90, FFTW, viz packages, Python, FFTW, numpy, Boost, OpenMP, ScaLAPCK, PSQL and MySQL databases, Eigen, cfitsio, astrometry.net, and Minuit2. • Needs to address limitations of supercomputer I/O subsystem.
34.	M0209 Large Survey Data for Cosmology **Capability Requirements:** • Needs to support standard astrophysics reduction software as well as Perl/Python wrapper scripts. • Needs to support Oracle RDBMS and Postgres psql, as well as GPFS and Lustre file systems and tape archives. • Needs to support parallel image storage.
35.	M0166 Particle Physics at LHC **Capability Requirements:** • Needs to support legacy computing infrastructure (computing nodes). • Needs to support distributed cached files (storage). • Needs to support object databases (software package).

TABLE D-3: CAPABILITIES	
36.	M0210 Belle II High Energy Physics Experiment **Capability Requirements:** • Needs to support 120 PB of raw data. • Needs to support an international distributed computing model to augment that at the accelerator in Japan. • Needs to support data transfer of ≈20 BG per second at designed luminosity between Japan and the United States. • Needs to support software from Open Science Grid, Geant4, DIRAC, FTS, and the Belle II framework.
37.	M0155 EISCAT 3D Incoherent Scatter Radar System **Capability Requirements:** • Needs to support architecture compatible with the ENVRI collaboration.
38.	M0157 ENVRI Environmental Research Infrastructure **Capability Requirements:** • Needs to support a variety of computing infrastructures and architectures (infrastructure). • Needs to support scattered repositories (storage).
39.	M0167 CReSIS Remote Sensing **Capability Requirements:** • Needs to support ≈0.5 PB per year of raw data. • Needs to support transfer of content from removable disk to computing cluster for parallel processing. • Needs to support Map/Reduce or MPI plus language binding for C/Java.
40.	M0127 UAVSAR Data Processing **Capability Requirements:** • Needs to support an interoperable cloud–HPC architecture. • Needs to host rich sets of radar image processing services. • Needs to support ROI_PAC, GeoServer, GDAL, and GeoTIFF-supporting tools. • Needs to support compatibility with other NASA radar systems and repositories (Alaska Satellite Facility).
41.	M0182 NASA LARC/GSFC iRODS **Capability Requirements:** • Needs to support vCDS. • Needs to support a GPFS integrated with Hadoop. • Needs to support iRODS.
42.	M0129 MERRA Analytic Services **Capability Requirements:** • Needs to support NetCDF aware software. • Needs to support Map/Reduce. • Needs to support interoperable use of AWS and local clusters.
43.	M0090 Atmospheric Turbulence **Capability Requirements:** • Needs to support other legacy computing systems (e.g., a supercomputer). • Needs to support high-throughput data transmission over the network.
44.	M0186 Climate Studies **Capability Requirements:** • Needs to support extension of architecture to several other fields.
45.	M0183 DOE-BER Subsurface Biogeochemistry **Capability Requirements:** • Needs to support Postgres, HDF5 data technologies, and many custom software systems.
46.	M0184 DOE-BER AmeriFlux and FLUXNET Networks **Capability Requirements:** • Needs to support custom software, such as EddyPro, and analysis software, such as R, Python, neural networks, and Matlab. • Needs to support analytics: data mining, data quality assessment, cross-correlation across datasets, data assimilation, data interpolation, statistics, quality assessment, data fusion, etc.
47.	M0223 Consumption Forecasting in Smart Grids **Capability Requirements:** • Needs to support SQL databases, CVS files, and HDFS (platform). • Needs to support R/Matlab, Weka, and Hadoop (platform).

TABLE D-4: DATA CONSUMER	
GENERAL REQUIREMENTS	
1. Needs to support fast searches from processed data with high relevancy, accuracy, and high recall.	Applies to 4 use cases: M0148, M0160, M0165, M0176
2. Needs to support diversified output file formats for visualization, rendering, and reporting.	Applies to 16 use cases: M0078, M0089, M0090, M0157, M0c161, M0164, M0164, M0165, M0166, M0166, M0167, M0167, M0174, M0177, M0213, M0214
3. Needs to support visual layouts for results presentation.	Applies to 2 use cases: M0165, M0167
4. Needs to support rich user interfaces for access using browsers, visualization tools.	Applies to 1 use cases: M0089, M0127, M0157, M0160, M0162, M0167, M0167, M0183, M0184, M0188, M0190
5. Needs to support a high-resolution multi-dimension layer of data visualization.	Applies to 21 use cases: M0129, M0155, M0155, M0158, M0161, M0162, M0171, M0172, M0173, M0177, M0179, M0182, M0185, M018c6, M0188, M0191, M0213, M0214, M02c15, M0219, M0222
6. Needs to support streaming results to clients.	Applies to 1 use case: M0164

USE CASE SPECIFIC REQUIREMENTS FOR DATA CONSUMERS

1. M0148 NARA: Search, Retrieve, Preservation **Data Consumer Requirements:**
   - Needs to support high relevancy and high recall from search.
   - Needs to support high accuracy from categorization of records.
   - Needs to support various storages such as NetApps, Hitachi, and magnetic tapes.
2. M0219 Statistical Survey Response Improvement **Data Consumer Requirements:**
   - Needs to support evolving data visualization for data review, operational activity, and general analysis.
3. M0222 Non-Traditional Data in Statistical Survey Response Improvement **Data Consumer Requirements:**
   - Needs to support evolving data visualization for data review, operational activity, and general analysis.
4. M0161 Mendeley **Data Consumer Requirements:**
   - Needs to support custom-built reporting tools.
   - Needs to support visualization tools such as networking graphs, scatterplots, etc.
5. M0164 Netflix Movie Service **Data Consumer Requirements:**
   - Needs to support streaming and rendering media
6. M0165 Web Search **Data Consumer Requirements:**
   - Needs to support search times of ≈0.1 seconds.
   - Needs to support top 10 ranked results.
   - Needs to support appropriate page layout (visual).
7. M0162 Materials Data for Manufacturing **Data Consumer Requirements:**
   - Needs to support visualization for materials discovery from many independent variables.
   - Needs to support visualization tools for multi-variable materials.
8. M0176 Simulation-Driven Materials Genomics **Data Consumer Requirements:**
   - Needs to support browser-based searches for growing material data.
9. M0213 Large-Scale Geospatial Analysis and Visualization **Data Consumer Requirements:**
   - Needs to support visualization with GIS at high and low network bandwidths and on dedicated facilities and handhelds.

TABLE D-4: DATA CONSUMER	
10.	M0214 Object Identification and Tracking **Data Consumer Requirements:** • Needs to support visualization of extracted outputs. These will typically be overlays on a geospatial display. Overlay objects should be links back to the originating image/video segment. • Needs to output the form of OGC-compliant web features or standard geospatial files (shape files, KML).
11.	M0215 Intelligence Data Processing and Analysis **Data Consumer Requirements:** • Needs to support primary visualizations, i.e., geospatial overlays (GIS) and network diagrams.
12.	M0177 EMR Data **Data Consumer Requirements:** • Needs to provide results of analytics for use by data consumers/stakeholders, i.e., those who did not actually perform the analysis. • Needs to support specific visualization techniques.
13.	M0089 Pathology Imaging **Data Consumer Requirements:** • Needs to support visualization for validation and training.
14.	M0191 Computational Bioimaging **Data Consumer Requirements:** • Needs to support 3D structural modeling.
15.	M0078 Genomic Measurements **Data Consumer Requirements:** • Needs to support data format for genome browsers.
16.	M0188 Comparative Analysis for Metagenomes and Genomes **Data Consumer Requirements:** • Needs to support real-time interactive parallel bulk loading capability. • Needs to support interactive web UI, backend pre-computations, and batch job computation submission from the UI. • Needs to support download assembled and annotated datasets for offline analysis. • Needs to support ability to query and browse data via interactive web UI. • Needs to support visualized data structure at different levels of resolution, as well as the ability to view abstract representations of highly similar data.
17.	M0174 Statistical Relational Artificial Intelligence for Health Care **Data Consumer Requirements:** • Needs to support visualization of subsets of very large data.
18.	M0172 World Population-Scale Epidemiological Study **Data Consumer Requirements:** • Needs to support visualization.
19.	M0173 Social Contagion Modeling for Planning **Data Consumer Requirements:** • 1. Needs to support multilevel detail network representations. • Needs to support visualization with interactions.
20.	M0141 Biodiversity and LifeWatch **Data Consumer Requirements:** • Needs to support advanced/rich/high-definition visualization. • Needs to support 4D visualization.
21.	M0171 Organizing Large-Scale Unstructured Collections of Consumer Photos **Data Consumer Requirements:** • Needs to support visualization of large-scale 3D reconstructions and navigation of large-scale collections of images that have been aligned to maps.
22.	M0160 Truthy Twitter Data **Data Consumer Requirements:** • Needs to support data retrieval and dynamic visualization. • Needs to support data-driven interactive web interfaces. • Needs to support API for data query.
23.	M0158 CINET for Network Science **Data Consumer Requirements:** • Needs to support client-side visualization.
24.	M0190 NIST Information Access Division **Data Consumer Requirements:** • Needs to support analytic flows involving users.

TABLE D-4: DATA CONSUMER

25. M0130 DataNet (iRODS) **Data Consumer Requirements:**
    - Needs to support general visualization workflows.
26. M0131 Semantic Graph-Search **Data Consumer Requirements:**
    - Needs to support efficient data-graph-based visualization.
27. M0170 Catalina Real-Time Transient Survey **Data Consumer Requirements:**
    - Needs to support visualization mechanisms for highly dimensional data parameter spaces.
28. M0185 DOE Extreme Data from Cosmological Sky Survey **Data Consumer Requirements:**
    - Needs to support interpretation of results using advanced visualization techniques and capabilities.
29. M0166 Particle Physics at LHC **Data Consumer Requirements:**
    - Needs to support histograms and model fits (visual).
30. M0155 EISCAT 3D Incoherent Scatter Radar System **Data Consumer Requirements:**
    - Needs to support visualization of high-dimensional (≥5) data.
31. M0157 ENVRI Environmental Research Infrastructure **Data Consumer Requirements:**
    - Needs to support graph-plotting tools.
    - Needs to support time series interactive tools.
    - Needs to support browser-based flash playback.
    - Needs to support earth high-resolution map displays.
    - Needs to support visual tools for quality comparisons.
32. M0167 CReSIS Remote Sensing **Data Consumer Requirements:**
    - Needs to support GIS user interface.
    - Needs to support rich user interface for simulations.
33. M0127 UAVSAR Data Processing **Data Consumer Requirements:**
    - Needs to support field expedition users with phone/tablet interface and low-resolution downloads.
34. M0182 NASA LARC/GSFC iRODS **Data Consumer Requirements:**
    - Needs to support visualization of distributed heterogeneous data.
35. M0129 MERRA Analytic Services **Data Consumer Requirements:**
    - Needs to support high-end distributed visualization.
36. M0090 Atmospheric Turbulence **Data Consumer Requirements:**
    - Needs to support visualization to interpret results.
37. M0186 Climate Studies **Data Consumer Requirements:**
    - Needs to support worldwide climate data sharing.
    - Needs to support high-end distributed visualization.
38. M0183 DOE-BER Subsurface Biogeochemistry **Data Consumer Requirements:**
    - Needs to support phone-based input and access.
39. M0184 DOE-BER AmeriFlux and FLUXNET Networks **Data Consumer Requirements:**
    - Needs to support phone-based input and access.

TABLE D-5: SECURITY AND PRIVACY

GENERAL REQUIREMENTS	
1. Needs to protect and preserve security and privacy for sensitive data.	Applies to 32 use cases: M0078, M0089, M0103, M0140, M0141, M0147, M0148, M0157, M0160, M0162, M0164, M0165, M0166, M0166, M0167, M0167, M0171, M0172, M0173, M0174, M0176, M0177, M0190, M0191, M0210, M0211, M0213, M0214, M0215, M0219, M0222, M0223

TABLE D-5: SECURITY AND PRIVACY		
2. Needs to support sandbox, access control, and multilevel policy-driven authentication on protected data.		Applies to 13 use cases: M0006, M0078, M0089, M0103, M0140, M0161, M0165, M0167, M0176, M0177, M0188, M0210, M0211
USE CASE SPECIFIC REQUIREMENTS FOR SECURITY AND PRIVACY		
1.	M0147 Census 2010 and 2000 **Security and Privacy Requirements:** • Needs to support Title 13 data.	
2.	M0148 NARA: Search, Retrieve, Preservation **Security and Privacy Requirements:** • Needs to support security policy.	
3.	M0219 Statistical Survey Response Improvement **Security and Privacy Requirements:** • Needs to support improved recommendation systems that reduce costs and improve quality while providing confidentiality safeguards that are reliable and publicly auditable. • Needs to support confidential and secure data. All processes must be auditable for security and confidentiality as required by various legal statutes.	
4.	M0222 Non-Traditional Data in Statistical Survey Response Improvement **Security and Privacy Requirements:** • Needs to support confidential and secure data. All processes must be auditable for security and confidentiality as required by various legal statutes.	
5.	M0175 Cloud Eco-System for Finance **Security and Privacy Requirements:** • Needs to support strong security and privacy constraints.	
6.	M0161 Mendeley **Security and Privacy Requirements:** • Needs to support access controls for who is reading what content.	
7.	M0164 Netflix Movie Service **Security and Privacy Requirements:** • Needs to support preservation of users' privacy and digital rights for media.	
8.	M0165 Web Search **Security and Privacy Requirements:** • Needs to support access control. • Needs to protect sensitive content.	
9.	M0137 Business Continuity and Disaster Recovery within a Cloud Eco-System **Security and Privacy Requirements:** • Needs to support strong security for many applications.	
10.	M0103 Cargo Shipping **Security and Privacy Requirements:** • Needs to support security policy.	
11.	M0162 Materials Data for Manufacturing **Security and Privacy Requirements:** • Needs to support protection of proprietary sensitive data. • Needs to support tools to mask proprietary information.	
12.	M0176 Simulation-Driven Materials Genomics **Security and Privacy Requirements:** • Needs to support sandbox as independent working areas between different data stakeholders. 2. Needs to support policy-driven federation of datasets.	
13.	M0213 Large-Scale Geospatial Analysis and Visualization **Security and Privacy Requirements:** • Needs to support complete security of sensitive data in transit and at rest (particularly on handhelds).	
14.	M0214 Object Identification and Tracking **Security and Privacy Requirements:** • Needs to support significant security and privacy; sources and methods cannot be compromised. The enemy should not be able to know what the user sees.	
15.	M0215 Intelligence Data Processing and Analysis **Security and Privacy Requirements:** • Needs to support protection of data against unauthorized access or disclosure and tampering.	

TABLE D-5: SECURITY AND PRIVACY	
16.	M0177 EMR Data **Security and Privacy Requirements:** • Needs to support direct consumer access to data, as well as referral to results of analytics performed by informatics research scientists and health service researchers. • Needs to support protection of all health data in compliance with government regulations. • Needs to support protection of data in accordance with data providers' policies. • Needs to support security and privacy policies, which may be unique to a subset of the data. • Needs to support robust security to prevent data breaches.
17.	M0089 Pathology Imaging **Security and Privacy Requirements:** • Needs to support security and privacy protection for protected health information.
18.	M0191 Computational Bioimaging **Security and Privacy Requirements:** • Needs to support significant but optional security and privacy, including secure servers and anonymization.
19.	M0078 Genomic Measurements **Security and Privacy Requirements:** • Needs to support security and privacy protection of health records and clinical research databases.
20.	M0188 Comparative Analysis for Metagenomes and Genomes **Security and Privacy Requirements:** • Needs to support login security, i.e., usernames and passwords. • Needs to support creation of user accounts to access datasets, and submit datasets to systems, via a web interface. • Needs to support single sign-on (SSO) capability.
21.	M0140 Individualized Diabetes Management **Security and Privacy Requirements:** • Needs to support protection of health data in accordance with privacy policies and legal security and privacy requirements, e.g., HIPAA. • Needs to support security policies for different user roles.
22.	M0174 Statistical Relational Artificial Intelligence for Health Care **Security and Privacy Requirements:** • Needs to support secure handling and processing of data.
23.	M0172 World Population-Scale Epidemiological Study **Security and Privacy Requirements:** • Needs to support protection of PII on individuals used in modeling. • Needs to support data protection and a secure platform for computation.
24.	M0173 Social Contagion Modeling for Planning **Security and Privacy Requirements:** • Needs to support protection of PII on individuals used in modeling. • Needs to support data protection and a secure platform for computation.
25.	M0141 Biodiversity and LifeWatch **Security and Privacy Requirements:** • Needs to support federated identity management for mobile researchers and mobile sensors. • Needs to support access control and accounting.
26.	M0171 Organizing Large-Scale Unstructured Collections of Consumer Photos **Security and Privacy Requirements:** • Needs to preserve privacy for users and digital rights for media.
27.	M0160 Truthy Twitter Data **Security and Privacy Requirements:** • Needs to support security and privacy policy.
28.	M0211 Crowd Sourcing in Humanities **Security and Privacy Requirements:** • Needs to support privacy issues in preserving anonymity of responses in spite of computer recording of access ID and reverse engineering of unusual user responses.
29.	M0190 NIST Information Access Division **Security and Privacy Requirements:** • Needs to support security and privacy requirements for protecting sensitive data while enabling meaningful developmental performance evaluation. Shared evaluation testbeds must protect the intellectual property of analytic algorithm developers.

TABLE D-5: SECURITY AND PRIVACY	
30.	M0130 DataNet (iRODS) **Security and Privacy Requirements:** • Needs to support federation across existing authentication environments through Generic Security Service API and pluggable authentication modules (GSI, Kerberos, InCommon, Shibboleth). • Needs to support access controls on files independent of the storage location.
31.	M0163 The Discinnet Process **Security and Privacy Requirements:** • Needs to support significant but optional security and privacy, including secure servers and anonymization.
32.	M0189 Light Source Beamlines **Security and Privacy Requirements:** • Needs to support multiple security and privacy requirements.
33.	M0166 Particle Physics at LHC **Security and Privacy Requirements:** • Needs to support data protection.
34.	M0210 Belle II High Energy Physics Experiment **Security and Privacy Requirements:** • Needs to support standard grid authentication.
35.	M0157 ENVRI Environmental Research Infrastructure **Security and Privacy Requirements:** • Needs to support an open data policy with minor restrictions.
36.	M0167 CReSIS Remote Sensing Security **and Privacy Requirements:** • Needs to support security and privacy on sensitive political issues. • Needs to support dynamic security and privacy policy mechanisms.
37.	M0223 Consumption Forecasting in Smart Grids **Security and Privacy Requirements:** • Needs to support privacy and anonymization by aggregation.

TABLE D-6: LIFE CYCLE MANAGEMENT	
GENERAL REQUIREMENTS	
1. Needs to support data quality curation including preprocessing, data clustering, classification, reduction, and format transformation.	Applies to 20 use cases: M0141, M0147, M0148, M0157, M0160, M0161, M0162, M0165, M0166, M0167, M0172, M0173, M0174, M0177, M0188, M0191, M0214, M0215, M0219, M0222)
2. Needs to support dynamic updates on data, user profiles, and links.	Applies to 2 use cases: M0164, M0209)
3. Needs to support data life cycle and long-term preservation policy, including data provenance.	Applies to 6 use cases: M0141, M0c147, M0155, M0163, M0164, M0165
4. Needs to support data validation.	Applies to 4 use cases: M0090, M0161, M0174, M0175
5. Needs to support human annotation for data validation.	Applies to 4 use cases: M0089, M01c27, M0140, M0188
6. Needs to support prevention of data loss or corruption.	Applies to 3 use cases: M0147, M0155, M0173)
7. Needs to support multisites archival.	Applies to 1 use case: M0157
8. Needs to support persistent identifier and data traceability.	Applies to 2 use cases: M0140, M0161)
9. Needs to standardize, aggregate, and normalize data from disparate sources.	Applies to 1 use case: M0177)
USE CASE SPECIFIC REQUIREMENTS FOR LIFE CYCLE MANAGEMENT	

TABLE D-6: LIFE CYCLE MANAGEMENT	
1.	M0147 Census 2010 and 2000 **Life Cycle Requirements:** • Needs to support long-term preservation of data as-is for 75 years. • Needs to support long-term preservation at the bit level. • Needs to support the curation process, including format transformation. • Needs to support access and analytics processing after 75 years. • Needs to ensure there is no data loss.
2.	M0148 NARA: Search, Retrieve, Preservation **Life Cycle Requirements:** • Needs to support pre-process for virus scans. • Needs to support file format identification. • Needs to support indexing. • Needs to support record categorization.
3.	M0219 Statistical Survey Response Improvement **Life Cycle Requirements:** • Needs to support high veracity of data, and systems must be very robust. The semantic integrity of conceptual metadata concerning what exactly is measured and the resulting limits of inference remain a challenge.
4.	M0222 Non-Traditional Data in Statistical Survey Response Improvement **Life Cycle Requirements:** • Needs to support high veracity of data, and systems must be very robust. The semantic integrity of conceptual metadata concerning what exactly is measured and the resulting limits of inference remain a challenge.
5.	M0161 Mendeley **Life Cycle Requirements:** • Needs to support metadata management from PDF extraction. • Needs to support identify of document duplication. • Needs to support persistent identifiers. • Needs to support metadata correlation between data repositories such as CrossRef, PubMed and Arxiv.
6.	M0164 Netflix Movie Service **Life Cycle Requirements:** • Needs to support continued ranking and updating based on user profiles and analytic results.
7.	M0165 Web Search **Life Cycle Requirements:** • Needs to support purge data after a certain time interval (a few months). • Needs to support data cleaning.
8.	M0162 Materials Data for Manufacturing **Life Cycle Requirements:** • Needs to support data quality handling; current process is poor or unknown.
9.	M0176 Simulation-Driven Materials Genomics **Life Cycle Requirements:** • Needs to support validation and UQ of simulation with experimental data. • Needs to support UQ in results from multiple datasets.
10.	M0214 Object Identification and Tracking **Life Cycle Requirements:** • Needs to support veracity of extracted objects.
11.	M0215 Intelligence Data Processing and Analysis **Life Cycle Requirements:** • Needs to support data provenance (e.g., tracking of all transfers and transformations) over the life of the data.
12.	M0177 EMR Data **Life Cycle Requirements:** • Needs to standardize, aggregate, and normalize data from disparate sources. • Needs to reduce errors and bias. • Needs to support common nomenclature and classification of content across disparate sources.
13.	M0089 Pathology Imaging **Life Cycle Requirements:** • Needs to support human annotations for validation.

TABLE D-6: LIFE CYCLE MANAGEMENT	
14.	M0191 Computational Bioimaging **Life Cycle Requirements:** • Needs to support workflow components include data acquisition, storage, enhancement, and noise minimization.
15.	M0188 Comparative Analysis for Metagenomes and Genomes **Life Cycle Requirements:** • Needs to support methods to improve data quality. • Needs to support data clustering, classification, and reduction. • Needs to support integration of new data/content into the system's data store and annotate data.
16.	M0140 Individualized Diabetes Management **Life Cycle Requirements:** • Needs to support data annotation based on domain ontologies or taxonomies. • Needs to ensure traceability of data from origin (initial point of collection) through use. • Needs to support data conversion from existing data warehouse into RDF triples.
17.	M0174 Statistical Relational Artificial Intelligence for Health Care **Life Cycle Requirements:** • Needs to support merging multiple tables before analysis. • Needs to support methods to validate data to minimize errors.
18.	M0172 World Population-Scale Epidemiological Study **Life Cycle Requirements:** • Needs to support data quality and capture traceability of quality from computation.
19.	M0173 Social Contagion Modeling for Planning **Life Cycle Requirements:** • Needs to support data fusion from variety of data sources. • Needs to support data consistency and prevent corruption. • Needs to support preprocessing of raw data.
20.	M0141 Biodiversity and LifeWatch **Life Cycle Requirements:** • Needs to support data storage and archiving, data exchange, and integration. • Needs to support data life cycle management: data provenance, referral integrity, and identification traceability back to initial observational data. • Needs to support processed (secondary) data (in addition to original source data) that may be stored for future uses. • Needs to support provenance (and PID) control of data, algorithms, and workflows. • Needs to support curated (authorized) reference data (i.e., species name lists), algorithms, software code, and workflows.
21.	M0160 Truthy Twitter Data **Life Cycle Requirements:** • Needs to support standardized data structures/formats with extremely high data quality.
22.	M0163 The Discinnet Process **Life Cycle Requirements:** • Needs to support integration of metadata approaches across disciplines.
23.	M0209 Large Survey Data for Cosmology **Life Cycle Requirements:** • Needs to support links between remote telescopes and central analysis sites.
24.	M0166 Particle Physics at LHC **Life Cycle Requirements:** • Needs to support data quality on complex apparatus.
25.	M0155 EISCAT 3D Incoherent Scatter Radar System **Life Cycle Requirements:** • Needs to support preservation of data and avoid data loss due to instrument malfunction.
26.	M0157 ENVRI Environmental Research Infrastructure **Life Cycle Requirements:** • Needs to support high data quality. • Needs to support mirror archives. • Needs to support various metadata frameworks. • Needs to support scattered repositories and data curation.
27.	M0167 CReSIS Remote Sensing **Life Cycle Requirements:** • Needs to support data quality assurance.

TABLE D-6: LIFE CYCLE MANAGEMENT

28.	M0127 UAVSAR Data Processing **Life Cycle Requirements:** • Needs to support significant human intervention in data processing pipeline. • Needs to support rich robust provenance defining complex machine/human processing.
29.	M0090 Atmospheric Turbulence **Life Cycle Requirements:** • Needs to support validation for output products (correlations).

TABLE D-7: OTHERS

GENERAL REQUIREMENTS

1. Needs to support rich user interfaces from mobile platforms to access processed results.	Applies to 6 use cases: M0078, M0127, M0129, M0148, M0160, M0164
2. Needs to support performance monitoring on analytic processing from mobile platforms.	Applies to 2 use cases: M0155, M0167
3. Needs to support rich visual content search and rendering from mobile platforms.	Applies to 13 use cases: M0078, M0089, M0161, M0164, M0165, M0166, M0176, M0177, M0183, M0184, M0186, M0219, M0223
4. Needs to support mobile device data acquisition.	Applies to 1 use case: M0157
5. Needs to support security across mobile devices.	Applies to 1 use case: M0177

USE CASE SPECIFIC REQUIREMENTS FOR OTHERS

1.	M0148 NARA: Search, Retrieve, Preservation **Other Requirements**: • Needs to support mobile search with similar interfaces/results from a desktop.
2.	M0219 Statistical Survey Response Improvement **Other Requirements**: • Needs to support mobile access.
3.	M0175 Cloud Eco-System for Finance **Other Requirements**: • Needs to support mobile access.
4.	M0161 Mendeley **Other Requirements**: • Needs to support Windows Android and iOS mobile devices for content deliverables from Windows desktops.
5.	M0164 Netflix Movie Service **Other Requirements**: • Needs to support smart interfaces for accessing movie content on mobile platforms.
6.	M0165 Web Search **Other Requirements**: • Needs to support mobile search and rendering.
7.	M0176 Simulation-Driven Materials Genomics **Other Requirements**: • Needs to support mobile apps to access materials genomics information.
8.	M0177 EMR Data **Other Requirements:** • Needs to support security across mobile devices.
9.	M0089 Pathology Imaging **Other Requirements**: • Needs to support 3D visualization and rendering on mobile platforms.
10.	M0078 Genomic Measurements **Other Requirements**: • Needs to support mobile platforms for physicians accessing genomic data (mobile device).
11.	M0140 Individualized Diabetes Management **Other Requirements**: • Needs to support mobile access.
12.	M0173 Social Contagion Modeling for Planning **Other Requirements**: • Needs to support an efficient method of moving data.

TABLE D-7: OTHERS	
13.	M0141 Biodiversity and LifeWatch **Other Requirements**: • Needs to support access by mobile users.
14.	M0160 Truthy Twitter Data **Other Requirements**: • Needs to support a low-level data storage infrastructure for efficient mobile access to data.
15.	M0155 EISCAT 3D Incoherent Scatter Radar System **Other Requirements**: • Needs to support real-time monitoring of equipment by partial streaming analysis.
16.	M0157 ENVRI Environmental Research Infrastructure **Other Requirements**: • Needs to support various kinds of mobile sensor devices for data acquisition.
17.	M0167 CReSIS Remote Sensing **Other Requirements**: • Needs to support monitoring of data collection instruments/sensors.
18.	M0127 UAVSAR Data Processing **Other Requirements**: • Needs to support field expedition users with phone/tablet interface and low-resolution downloads.
19.	M0129 MERRA Analytic Services **Other Requirements**: • Needs to support smart phone and tablet access. • Needs to support iRODS data management.
20.	M0186 Climate Studies **Other Requirements**: • Needs to support phone-based input and access.
21.	M0183 DOE-BER Subsurface Biogeochemistry **Other Requirements**: • Needs to support phone-based input and access.
22.	M0184 DOE-BER AmeriFlux and FLUXNET Networks **Other Requirements**: • Needs to support phone-based input and access.
23.	M0223 Consumption Forecasting in Smart Grids **Other Requirements**: • Needs to support mobile access for clients.

Appendix E: Use Case Template 2

Use Case Template 2 is currently being used to gather information on additional use cases, which will be incorporated into future work of the NBDIF. Appendix E contains an outline of the questions in the Use Case Template 2 and is provided for the readers' reference. To submit a new use case, please use the fillable PDF form that can be downloaded from the NBD-PWG website at https://bigdatawg.nist.gov/_uploadfiles/M0621_v2_7345181325.pdf.

BIG DATA USE CASE TEMPLATE 2

NIST Big Data Public Working Group

This template was designed by the NIST Big Data Public Working Group (NBD-PWG) to gather Big Data use cases. The use case information you provide in this template will greatly help the NBD-PWG in the next phase of developing the NIST Big Data Interoperability Framework. We sincerely appreciate your effort and realize it is nontrivial.

The template can also be completed in the Google Form for Use Case Template 2: http://bit.ly/1ff7iM9.

More information about the NBD-PWG and the NIST Big Data Interoperability Framework can be found at http://bigdatawg.nist.gov.

TEMPLATE OUTLINE

General Instructions:

Brief instructions are provided with each question requesting an answer in a text field. For the questions offering check boxes, please check any that apply to the use case. .

No fields are required to be filled in. Please fill in the fields that you are comfortable answering. The fields that are particularly important to the work of the NBD-PWG are marked with *.

Please email the completed template to Wo Chang at wchang@nist.gov.

NOTE: No proprietary or confidential information should be included.

1 OVERALL PROJECT DESCRIPTION

1.1 USE CASE TITLE *

Please limit to one line. A description field is provided below for a longer description.

1.2 USE CASE DESCRIPTION *

Summarize all aspects of use case focusing on application issues (later questions will highlight technology).

1.3 USE CASE CONTACTS *

Add names, phone number, and email of key people associated with this use case. Please designate who is authorized to edit this use case.

Name	Phone	Email	PI / Author	Edit rights?	Primary

1.4 DOMAIN ("VERTICAL") *

What application area applies? There is no fixed ontology. Examples: Health Care, Social Networking, Financial, Energy, etc.

1.5 APPLICATION *

Summarize the use case applications.

1.6 CURRENT DATA ANALYSIS APPROACH *

Describe the analytics, software, hardware approach used today. This section can be qualitative with details given in Section 3.6.

1.7 FUTURE OF APPLICATION AND APPROACH *

Describe the analytics, software, hardware, and application future plans, with possible increase in data sizes/velocity.

1.8 ACTORS / STAKEHOLDERS

Please describe the players and their roles in the use case. Identify relevant stakeholder roles and responsibilities. Note: Security and privacy roles are discussed in a separate part of this template.

1.9 PROJECT GOALS OR OBJECTIVES

Please describe the objectives of the use case.

1.10 USE CASE URL(S)

Include any URLs associated with the use case. Please separate with semicolon (;).

1.11 PICTURES AND DIAGRAMS?

Please email any pictures or diagrams with this template.

2 BIG DATA CHARACTERISTICS

Big Data Characteristics describe the properties of the (raw) data including the four major ‘V’s’ of Big Data described in NIST Big Data Interoperability Framework: Volume 1, Big Data Definition.

2.1 DATA SOURCE

Describe 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.

2.2 DATA DESTINATION

If the data is transformed in the use case, describe where the final results end up. This has similar characteristics to data source.

2.3 VOLUME

Size	
Units	
Time Period	
Proviso	

Size: Quantitative volume of data handled in the use case
Units: What is measured such as "Tweets per year", Total LHC data in petabytes, etc.?
Time Period: Time corresponding to specified size.
Proviso: The criterion (e.g. data gathered by a particular organization) used to get size with units in time period in three fields above

2.4 VELOCITY

Enter if real-time or streaming data is important. Be quantitative: this number qualified by 3 fields below: units, time period, proviso. Refers to the rate of flow at which the data is created, stored, analyzed, and visualized. For example, big velocity means that a large quantity of data is being processed in a short amount of time.

Unit of measure	
Time Period	
Proviso	

Unit of Measure: Units of Velocity size given above. What is measured such as "New Tweets gathered per second", etc.?

Time Period: Time described and interval such as September 2015; items per minute

Proviso: The criterion (e.g., data gathered by a particular organization) used to get Velocity measure with units in time period in three fields above

2.5 VARIETY

Variety refers to data from multiple repositories, domains, or types. Please indicate if the data is from multiple datasets, mashups, etc.

2.6 VARIABILITY

Variability refers to changes in rate and nature of data gathered by use case. It captures a broader range of changes than Velocity which is just change in size. Please describe the use case data variability.

3 BIG DATA SCIENCE

3.1 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 (e.g., presence of missing fields or incorrect values).

3.2 VISUALIZATION

Describe the way the 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.

3.3 DATA TYPES

Refers to the style of data, such as structured, unstructured, images (e.g., pixels), text (e.g., characters), gene sequences, and numerical.

3.4 METADATA

Please comment on quality and richness of metadata.

3.5 CURATION AND GOVERNANCE

Note that we have a separate section for security and privacy. Comment on process to ensure good data quality and who is responsible.

3.6 DATA ANALYTICS

In the context of these use cases, 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. This section should be reasonably precise so quantitative comparisons with other use cases can be made. Section 1.6 is qualitative discussion of this feature.

4 GENERAL SECURITY AND PRIVACY

The following questions are intended to cover general security and privacy topics. Security and privacy topics are explored in more detail in Section 8. For the questions with checkboxes, please select the item(s) that apply to the use case.

4.1 CLASSIFIED DATA, CODE OR PROTOCOLS

- ☐ Intellectual property protections
- ☐ Military classifications, e.g., FOUO, or Controlled Classified
- ☐ Not applicable
- ☐ Creative commons/ open source
- ☐ Other:

4.2 DOES THE SYSTEM MAINTAIN PERSONALLY IDENTIFIABLE INFORMATION (PII)? *

- ☐ Yes, PII is part of this Big Data system
- ☐ No, and none can be inferred from 3rd party sources
- ☐ No, but it is possible that individuals could be identified via third party databases
- ☐ Other:

4.3 PUBLICATION RIGHTS

Open publisher; traditional publisher; white paper; working paper

- ☐ Open publication
- ☐ Proprietary
- ☐ Traditional publisher rights (e.g., Springer, Elsevier, IEEE)
- ☐ "Big Science" tools in use
- ☐ Other:

4.4 IS THERE AN EXPLICIT DATA GOVERNANCE PLAN OR FRAMEWORK FOR THE EFFORT?

Data governance refers to the overall management of the availability, usability, integrity, and security of the data employed in an enterprise.

- ☐ Explicit data governance plan
- ☐ No data governance plan, but could use one
- ☐ Data governance does not appear to be necessary
- ☐ Other:

4.5 DO YOU FORESEE ANY POTENTIAL RISKS FROM PUBLIC OR PRIVATE OPEN DATA PROJECTS?

Transparency and data sharing initiatives can release into public use datasets that can be used to undermine privacy (and, indirectly, security.)

- ☐ Risks are known.
- ☐ Currently no known risks, but it is conceivable.
- ☐ Not sure
- ☐ Unlikely that this will ever be an issue (e.g., no PII, human-agent related data or subsystems.)
- ☐ Other:

4.6 CURRENT AUDIT NEEDS *

- ☐ We have third party registrar or other audits, such as for ISO 9001
- ☐ We have internal enterprise audit requirements
- ☐ Audit is only for system health or other management requirements
- ☐ No audit, not needed or does not apply
- ☐ Other:

4.7 UNDER WHAT CONDITIONS DO YOU GIVE PEOPLE ACCESS TO YOUR DATA?

4.8 UNDER WHAT CONDITIONS DO YOU GIVE PEOPLE ACCESS TO YOUR SOFTWARE?

5 CLASSIFY USE CASES WITH TAGS

The questions below will generate tags that can be used to classify submitted use cases. See http://dsc.soic.indiana.edu/publications/OgrePaperv11.pdf (Towards an Understanding of Facets and Exemplars of Big Data Applications) for an example of how tags were used in the initial 51 use cases. Check any number of items from each of the questions.

5.1 DATA: APPLICATION STYLE AND DATA SHARING AND ACQUISITION

- [ ] Uses Geographical Information Systems?
- [ ] Use case involves Internet of Things?
- [ ] Data comes from HPC or other simulations?
- [ ] Data Fusion important?
- [ ] Data is Real time Streaming?
- [ ] Data is Batched Streaming (e.g. collected remotely and uploaded every so often)?
- [ ] Important Data is in a Permanent Repository (Not streamed)?
- [ ] Transient Data important?
- [ ] Permanent Data Important?
- [ ] Data shared between different applications/users?
- [ ] Data largely dedicated to only this use case?

5.2 DATA: MANAGEMENT AND STORAGE

- [ ] Application data system based on Files?
- [ ] Application data system based on Objects?
- [ ] Uses HDFS style File System?
- [ ] Uses Wide area File System like Lustre?
- [ ] Uses HPC parallel file system like GPFS?
- [ ] Uses SQL?
- [ ] Uses NoSQL?
- [ ] Uses NewSQL?
- [ ] Uses Graph Database?

5.3 DATA: DESCRIBE OTHER DATA ACQUISITION/ ACCESS/ SHARING/ MANAGEMENT/ STORAGE ISSUES

5.4 ANALYTICS: DATA FORMAT AND NATURE OF ALGORITHM USED IN ANALYTICS

	Data regular?
	Data dynamic?
	Algorithm O(N^2)?
	Basic statistics (regression, moments) used?
	Search/Query/Index of application data Important?
	Classification of data Important?
	Recommender Engine Used?
	Clustering algorithms used?
	Alignment algorithms used?
	(Deep) Learning algorithms used?
	Graph Analytics Used?

5.5 ANALYTICS: DESCRIBE OTHER DATA ANALYTICS USED

Examples include learning styles (supervised) or libraries (Mahout).

5.6 PROGRAMMING MODEL

	Pleasingly parallel Structure? Parallel execution over independent data. Called Many Task or high throughput computing. MapReduce with only Map and no Reduce of this type
	Use case NOT Pleasingly Parallel -- Parallelism involves linkage between tasks. MapReduce (with Map and Reduce) of this type
	Uses Classic MapReduce? such as Hadoop
	Uses Apache Spark or similar Iterative MapReduce?
	Uses Graph processing as in Apache Giraph?
	Uses MPI (HPC Communication) and/or Bulk Synchronous Processing BSP?
	Dataflow Programming Model used?
	Workflow or Orchestration software used?
	Python or Scripting front ends used? Maybe used for orchestration
	Shared memory architectures important?
	Event-based Programming Model used?
	Agent-based Programming Model used?
	Use case I/O dominated? I/O time > or >> Compute time
	Use case involves little I/O? Compute >> I/O

5.7 OTHER PROGRAMMING MODEL TAGS

Provide other programming style tags not included in the list above.

5.8 PLEASE ESTIMATE RATIO I/O BYTES/FLOPS

Specify in text box with units.

5.9 DESCRIBE MEMORY SIZE OR ACCESS ISSUES

Specify in text box with any quantitative detail on memory access/compute/I/O ratios.

6 OVERALL BIG DATA ISSUES

6.1 OTHER BIG DATA ISSUES

Please list other important aspects that the use case highlights. This question provides a chance to address questions which should have been asked.

6.2 USER INTERFACE AND MOBILE ACCESS ISSUES

Describe issues in accessing or generating Big Data from clients, including Smart Phones and tablets.

6.3 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?

7 WORKFLOW PROCESSES

Please answer this question if the use case contains multiple steps where Big Data characteristics, recorded in this template, vary across steps. If possible, flesh out workflow in the separate set of questions. Only use this section if your use case has multiple stages where Big Data issues differ significantly between stages.

7.1 PLEASE COMMENT ON WORKFLOW PROCESSES

Please record any overall comments on the use case workflow.

7.2 WORKFLOW DETAILS FOR EACH STAGE *

Description of table fields below:
Data Source(s): 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. Often data source at one stage is destination of previous stage with raw data driving first stage.
Nature of Data: What items are in the data?
Software Used: List software packages used
Data Analytics: List algorithms and analytics libraries/packages used
Infrastructure: Compute, Network and Storage used. Note sizes infrastructure -- especially if "big".
Percentage of Use Case Effort: Explain units. Could be clock time elapsed or fraction of compute cycles

Other Comments: Include comments here on items like veracity and variety present in upper level but omitted in summary.

7.2.1 WORKFLOW DETAILS FOR STAGE 1

Stage 1 Name	
Data Source(s)	
Nature of Data	
Software Used	
Data Analytics	
Infrastructure	
Percentage of Use Case Effort	
Other Comments	

7.2.2 WORKFLOW DETAILS FOR STAGE 2

Stage 2 Name	
Data Source(s)	
Nature of Data	
Software Used	
Data Analytics	
Infrastructure	
Percentage of Use Case Effort	
Other Comments	

7.2.3 WORKFLOW DETAILS FOR STAGE 3

Stage 3 Name	
Data Source(s)	
Nature of Data	
Software Used	
Data Analytics	
Infrastructure	
Percentage of Use Case Effort	
Other Comments	

7.2.4 WORKFLOW DETAILS FOR STAGE 4

Stage 4 Name	
Data Source(s)	
Nature of Data	
Software Used	
Data Analytics	
Infrastructure	
Percentage of Use Case Effort	
Other Comments	

7.2.5 WORKFLOW DETAILS FOR STAGES 5 AND ANY FURTHER STAGES

If you have more than five stages, please put stages 5 and higher here.

Stage 5 Name	
Data Source(s)	
Nature of Data	
Software Used	
Data Analytics	
Infrastructure	
Percentage of Use Case Effort	
Other Comments	

8 DETAILED SECURITY AND PRIVACY

Questions in this section are designed to gather a comprehensive image of security and privacy aspects (e.g., security, privacy, provenance, governance, curation, and system health) of the use case. Other sections contain aspects of curation, provenance, and governance that are not strictly speaking only security and privacy considerations. The answers will be very beneficial to the NBD-PWG in understanding your use case. However, if you are unable to answer the questions in this section, the NBD-PWG would still be interested in the information gathered in the rest of the template. The security and privacy questions are grouped as follows:

- Roles
- Personally Identifiable Information
- Covenants and Liability
- Ownership, Distribution, Publication
- Risk Mitigation
- Audit and Traceability
- Data Life Cycle
- Dependencies
- Framework provider S&P
- Application Provider S&P
- Information Assurance | System Health
- Permitted Use Cases

8.1 ROLES

Roles may be associated with multiple functions within a big data ecosystem.

8.1.1 IDENTIFYING ROLE

Identify the role (e.g., Investigator, Lead Analyst, Lead Scientists, Project Leader, Manager of Product Development, VP Engineering) associated with identifying the use case need, requirements, and deployment.

8.1.2 INVESTIGATOR AFFILIATIONS

This can be time-dependent and can include past affiliations in some domains.

8.1.3 SPONSORS

Include disclosure requirements mandated by sponsors, funders, etc.

8.1.4 DECLARATIONS OF POTENTIAL CONFLICTS OF INTEREST

8.1.5 INSTITUTIONAL S/P DUTIES

List and describe roles assigned by the institution, such as via an IRB (Institutional Review Board).

8.1.6 CURATION

List and describe roles associated with data quality and curation, independent of any specific Big Data component. Example: Role responsible for identifying U.S. government data as FOUO or Controlled Unclassified Information, etc.

8.1.7 CLASSIFIED DATA, CODE OR PROTOCOLS

- [ ] Intellectual property protections
- [ ] Military classifications, e.g., FOUO, or Controlled Classified
- [ ] Not applicable
- [ ] Creative commons/ open source
- [ ] Other:

8.1.8 MULTIPLE INVESTIGATORS | PROJECT LEADS *

- [ ] Only one investigator | project lead | developer
- [ ] Multiple team members, but in the same organization
- [ ] Multiple leads across legal organizational boundaries
- [ ] Multinational investigators | project leads
- [ ] Other:

8.1.9 LEAST PRIVILEGE ROLE-BASED ACCESS

Least privilege requires that a user receives no more permissions than necessary to perform the user's duties.

- [ ] Yes, roles are segregated and least privilege is enforced
- [ ] We do have least privilege and role separation but the admin role(s) may be too all-inclusion
- [ ] Handled at application provider level
- [ ] Handled at framework provider level
- [ ] There is no need for this feature in our application
- [ ] Could be applicable in production or future versions of our work
- [ ] Other:

8.1.10 ROLE-BASED ACCESS TO DATA *

Please describe the level at which access to data is limited in your system.

- [ ] Dataset
- [ ] Data record / row
- [ ] Data element / field
- [ ] Handled at application provider level
- [ ] Handled at framework provider level
- [ ] Other:

8.2 PERSONALLY IDENTIFIABLE INFORMATION (PII)

8.2.1 DOES THE SYSTEM MAINTAIN PII? *

- [ ] Yes, PII is part of this Big Data system.
- [ ] No, and none can be inferred from third-party sources.
- [ ] No, but it is possible that individuals could be identified via third-party databases.
- [ ] Other:

8.2.2 DESCRIBE THE PII, IF APPLICABLE

Describe how PII is collected, anonymized, etc. Also list disclosures to human subjects, interviewees, or web visitors.

8.2.3 ADDITIONAL FORMAL OR INFORMAL PROTECTIONS FOR PII

8.2.4 ALGORITHMIC / STATISTICAL SEGMENTATION OF HUMAN POPULATIONS

- [ ] Yes, doing segmentation, possible discrimination issues if abused. Please also answer the next question.
- [ ] Yes, doing segmentation, but no foreseeable discrimination issues.
- [ ] Does not apply to this use case at all (e.g., no human subject data).
- [ ] Other:

8.2.5 PROTECTIONS AFFORDED STATISTICAL / DEEP LEARNING DISCRIMINATION

Identify what measures are in place to address this concern regarding human populations, if it applies. Refer to the previous question.

8.3 COVENANTS, LIABILITY, ETC.

8.3.1 IDENTIFY ANY ADDITIONAL SECURITY, COMPLIANCE, REGULATORY REQUIREMENTS *

Refer to 45 CFR 46: http://1.usa.gov/1bg6JQ2

- [ ] FTC regulations apply
- [ ] HHS 45 CFR 46
- [ ] HIPAA
- [ ] EU General Data Protection (Reference: http://bit.ly/1Ta8S1C)
- [ ] COPPA
- [ ] Other Transborder issues
- [ ] Fair Credit Reporting Act (Reference: http://bit.ly/1Ta8XSN)
- [ ] Family Educational Rights and Protection (FERPA)
- [ ] None apply
- [ ] Other:

8.3.2 CUSTOMER PRIVACY PROMISES

Select all that apply,e.g., RadioShack promise that is subject of this DOJ ruling: http://bit.ly/1f0MW9t

- [ ] Yes, we're making privacy promises to customers or subjects.
- [ ] We are using a notice-and-consent model.
- [ ] Not applicable
- [ ] Other:

8.4 OWNERSHIP, IDENTITY AND DISTRIBUTION

8.4.1 PUBLICATION RIGHTS

Open publisher; traditional publisher; white paper; working paper

- [ ] Open publication
- [ ] Proprietary
- [ ] Traditional publisher rights (e.g., Springer, Elsevier, IEEE)
- [ ] "Big Science" tools in use
- [ ] Other:

8.4.2 CHAIN OF TRUST

Identify any chain-of-trust mechanisms in place (e.g., ONC Data Provenance Initiative). Potentially very domain-dependent; see the ONC event grid, for instance. Reference: http://bit.ly/1f0PGDL

8.4.3 DELEGATED RIGHTS

Example of one approach: “Delegation Logic: A Logic-based Approach to Distributed Authorization”, Li, N., Grosof, B.N., Feigenbaum, J.(2003) https://www.cs.purdue.edu/homes/ninghui/papers/thesis.pdf

8.4.4 SOFTWARE LICENSE RESTRICTIONS

Identify proprietary software used in the use case Big Data system which could restrict use, reproducibility, results, or distribution.

8.4.5 RESULTS REPOSITORY

Identify any public or private / federated consortia maintaining a shared repository.

8.4.6 RESTRICTIONS ON DISCOVERY

Describe restrictions or protocols imposed on discoverable end points.

8.4.7 PRIVACY NOTICES

Indicate any privacy notices required / associated with data collected for redistribution to others,

- [ ] Privacy notices apply
- [ ] Privacy notices do not apply
- [ ] Other:

8.4.8 KEY MANAGEMENT

- [ ] A key management scheme is part of our system.
- [ ] We are using public key infrastructure.
- [ ] We do not use key management, but it could have been useful.
- [ ] No readily identifiable use for key management.
- [ ] Other:

8.4.9 DESCRIBE THE KEY MANAGEMENT PRACTICES

8.4.10 IS AN IDENTITY FRAMEWORK USED?

- [ ] A framework is in place. (See next question.)
- [ ] Not currently using a framework.
- [ ] There is no perceived need for an identity framework.
- [ ] Other:

8.4.11 CAC / ECA CARDS OR OTHER ENTERPRISE-WIDE FRAMEWORK

- [ ] Using an externally maintained enterprise-wide identity framework.
- [ ] Could be used, but none are available.
- [ ] Not applicable

8.4.12 DESCRIBE THE IDENTITY FRAMEWORK.

8.4.13 HOW IS INTELLECTUAL PROPERTY PROTECTED?

- [ ] Login screens advising of IP issues
- [ ] Employee or team training
- [ ] Official guidelines limiting access or distribution
- [ ] Required to track all access to, distribution of digital assets
- [ ] Does not apply to this effort (e.g., public effort)
- [ ] Other:

8.5 RISK MITIGATION

8.5.1 ARE MEASURES IN PLACE TO DETER RE-IDENTIFICATION? *

- [ ] Yes, in place
- [ ] Not in place, but such measures do apply
- [ ] Not applicable
- [ ] Other:

8.5.2 PLEASE DESCRIBE ANY RE-IDENTIFICATION DETERRENTS IN PLACE

8.5.3 ARE DATA SEGMENTATION PRACTICES BEING USED?

Data segmentation for privacy has been suggested as one strategy to enhance privacy protections. Reference: http://bit.ly/1P3h12Y

- [ ] Yes, being used
- [ ] Not in use, but does apply
- [ ] Not applicable
- [ ] Other:

8.5.4 IS THERE AN EXPLICIT DATA GOVERNANCE PLAN OR FRAMEWORK FOR THE EFFORT?

Data governance refers to the overall management of the availability, usability, integrity, and security of the data employed in an enterprise.

- [ ] Explicit data governance plan
- [ ] No data governance plan, but could use one
- [ ] Data governance does not appear to be necessary
- [ ] Other:

8.5.5 PRIVACY-PRESERVING PRACTICES

Identify any privacy-preserving measures that are in place.

8.5.6 DO YOU FORESEE ANY POTENTIAL RISKS FROM PUBLIC OR PRIVATE OPEN DATA PROJECTS?

Transparency and data sharing initiatives can release into public use datasets that can be used to undermine privacy (and, indirectly, security).

- [ ] Risks are known.
- [ ] Currently no known risks, but it is conceivable.
- [ ] Not sure
- [ ] Unlikely that this will ever be an issue (e.g., no PII, human-agent related data or subsystems).
- [ ] Other:

8.6 PROVENANCE (OWNERSHIP)

Provenance viewed from a security or privacy perspective. The primary meaning for some domains is digital reproducibility, but it could apply in simulation scenarios as well.

8.6.1 DESCRIBE YOUR METADATA MANAGEMENT PRACTICES

- ☐ Yes, we have a metadata management system.
- ☐ There is no need for a metadata management system in this use case.
- ☐ It is applicable but we do not currently have one.
- ☐ Other:

8.6.2 IF A METADATA MANAGEMENT SYSTEM IS PRESENT, WHAT MEASURES ARE IN PLACE TO VERIFY AND PROTECT ITS INTEGRITY?

8.6.3 DESCRIBE PROVENANCE AS RELATED TO INSTRUMENTATION, SENSORS OR OTHER DEVICES.

- ☐ We have potential machine-to-machine traffic provenance concerns.
- ☐ Endpoint sensors or instruments have signatures periodically updated.
- ☐ Using hardware or software methods, we detect and remediate outlier signatures.
- ☐ Endpoint signature detection and upstream flow are built into system processing.
- ☐ We rely on third-party vendors to manage endpoint integrity.
- ☐ We use a sampling method to verify endpoint integrity.
- ☐ Not a concern at this time.
- ☐ Other:

8.7 DATA LIFE CYCLE

8.7.1 DESCRIBE ARCHIVE PROCESSES

- ☐ Our application has no separate "archive" process.
- ☐ We offload data using certain criteria to removable media which are taken offline.
- ☐ We use a multi-stage, tiered archive process.
- ☐ We allow for "forgetting" of individual PII on request.
- ☐ Have ability to track individual data elements across all stages of processing, including archive.
- ☐ Additional protections, such as separate encryption, are applied to archival data.
- ☐ Archived data is saved for potential later use by applications or analytics yet to be built.
- ☐ Does not apply to our application.
- ☐ Other:

8.7.2 DESCRIBE POINT IN TIME AND OTHER DEPENDENCY ISSUES

- [ ] Some data is valid only within a point in time,
- [ ] Some data is only valid with other, related data is available or applicable, such as the existence of a building, the presence of a weather event, or the active use of a vehicle.
- [ ] There are specific events in the application that render certain data obsolete or unusable.
- [ ] Point and Time and related dependencies do not apply.
- [ ] Other:

8.7.3 COMPLIANCE WITH SECURE DATA DISPOSAL REQUIREMENTS

Per NCSL: "at least 29 states have enacted laws that require entities to destroy, dispose. . ."
http://www.ncsl.org/research/telecommunications-and-information-technology/privacy-and-security.aspx

- [ ] We are required to destroy or otherwise dispose of data.
- [ ] Does not apply to us.
- [ ] Not sure
- [ ] Other:

8.8 AUDIT AND TRACEABILITY

Big Data use case: SEC Rule 613 initiative

8.8.1 CURRENT AUDIT NEEDS *

- [ ] We have third-party registrar or other audits, such as for ISO 9001.
- [ ] We have internal enterprise audit requirements.
- [ ] Audit is only for system health or other management requirements.
- [ ] No audit, not needed or does not apply.
- [ ] Other:

8.8.2 AUDITING VERSUS MONITORING

- [ ] We rely on third-party or O.S. tools to audit, e.g., Windows or Linux auditing.
- [ ] There are built-in tools for monitoring or logging that are only used for system or application health monitoring.
- [ ] Monitoring services include logging of role-based access to assets such as PII or other resources.
- [ ] The same individual(s) in the enterprise are responsible for auditing as for monitoring.
- [ ] This aspect of our application is still in flux.
- [ ] Does not apply to our setting.
- [ ] Other:

8.8.3 SYSTEM HEALTH TOOLS

- [ ] We rely on system-wide tools for health monitoring.
- [ ] We built application health tools specifically to address integrity, performance monitoring, and related concerns.
- [ ] There is no need in our setting.
- [ ] Other:

8.8.4 WHAT EVENTS ARE CURRENTLY AUDITED? *

- [ ] All data access must be audited.
- [ ] Only selected / protected data must be audited.
- [ ] Maintenance on user roles must be audited (new users, disabled user, updated roles or permissions).
- [ ] Purge and archive events.
- [ ] Domain-dependent events (e.g., adding a new sensor).
- [ ] REST or SOAP events
- [ ] Changes in system configuration
- [ ] Organizational changes
- [ ] External project ownership / management changes
- [ ] Requirements are externally set, e.g., by PCI compliance.
- [ ] Domain-specific events (patient death in a drug trial)
- [ ] Other:

8.9 APPLICATION PROVIDER SECURITY

8.9.1 DESCRIBE APPLICATION PROVIDER SECURITY *

One example of application layer security is the SAP ERP application.

- [ ] There is a security mechanism implemented at the application level.
- [ ] The app provider level is aware of PII or privacy data elements.
- [ ] The app provider implements audit and logging.
- [ ] The app provider security relies on framework-level security for its operation.
- [ ] Does not apply to our application.
- [ ] Other:

8.10 FRAMEWORK PROVIDER SECURITY

One example is Microsoft Active Directory as applied across LANs to Azure, or LDAP mapped to Hadoop. Reference: http://bit.ly/1f0VDR3

8.10.1 DESCRIBE THE FRAMEWORK PROVIDER SECURITY *

- [ ] Security is implemented at the framework level.
- [ ] Roles can be defined at the framework level.
- [ ] The framework level is aware of PII or related sensitive data.
- [ ] Does not apply in our setting.
- [ ] Is provided by the Big Data tool.
- [ ] Other:

8.11 SYSTEM HEALTH

Also included in this grouping: Availability, Resilience, Information Assurance

8.11.1 MEASURES TO ENSURE AVAILABILITY *

- [ ] Deterrents to man-in-the-middle attacks
- [ ] Deterrents to denial of service attacks
- [ ] Replication, redundancy or other resilience measures
- [ ] Deterrents to data corruption, drops or other critical big data components
- [ ] Other:

8.12 PERMITTED USE CASES

Beyond the scope of S&P considerations presented thus far, please identify particular domain-specific limitations

8.12.1 DESCRIBE DOMAIN-SPECIFIC LIMITATIONS ON USE

8.12.2 PAYWALL

- [ ] A paywall is in use at some stage in the workflow.
- [ ] Not applicable

Description of NIST Public Working Group on Big Data

NIST is leading the development of a Big Data Technology Roadmap. This roadmap will define and prioritize requirements for interoperability, portability, reusability, and extendibility for Big Data analytic techniques and technology infrastructure in order to support secure and effective adoption of Big Data. To help develop the ideas in the Big Data Technology Roadmap, NIST created the Public Working Group for Big Data.

Scope: The focus of the NBD-PWG is to form a community of interest from industry, academia, and government, with the goal of developing consensus definitions, taxonomies, secure reference architectures, and a technology roadmap. The aim is to create vendor-neutral, technology- and infrastructure-agnostic deliverables to enable Big Data stakeholders to pick and choose best analytics tools for their processing and visualization requirements on the most suitable computing platforms and clusters while allowing value-added from Big Data service providers and flow of data between the stakeholders in a cohesive and secure manner.

For more, refer to the website at http://bigdatawg.nist.gov.

Appendix F: Version 2 Raw Use Case Data

This appendix contains the raw data from the two Template 2 use cases that have been submitted to date. Summaries of these use cases are included in Section 2

F.1 Use Case 2-1: Web-Enabled Landsat Data (WELD) Processing

1 Overall Project Description Use Case 2-1	
1.1 Use Case Title *	NASA Earth Observing System Data and Information System (EOSDIS)
1.2 Use Case Description *	The Earth Observing System Data and Information System (EOSDIS) is the main system maintained by NASA for the archive and dissemination of Earth Observation data. The system comprises 12 discipline-oriented data systems spread across the United States. This network is linked together using interoperability frameworks such as the Common Metadata Repository, a file-level database that supports one-stop searching across EOSDIS. The data consist of satellite, aircraft, field campaign and in situ data over a variety of disciplines related to Earth science, covering the Atmosphere, Hydrosphere, Cryosphere, Lithosphere, Biosphere, and Anthroposphere. Data are distributed to a diverse community ranging from Earth science researchers to applications to citizen science and educational users. EOSDIS faces major challenges in both Volume and Variety. As of early 2017, the cumulative archive data volume is over 20 Petabytes. Higher resolution spaceborne instruments are expected to increase that volume by two orders of magnitude (~200 PB) over the next 7 years. More importantly, the data distribution to users is equally high. In a given year, EOSDIS distributes a volume that is comparable to the overall cumulative archive volume.
1.3 Use Case Contacts *	
Name	Christopher Lynnes
PI or Author	Author
Edit Privileges?	Yes
Primary author?	Yes

1 Overall Project Description

Use Case 2-1

1.4 Domain ("Vertical") *	Earth Science
1.5 Application *	Data Archiving: storing NASA's Earth Observation dataData Distribution: disseminating data to end users in Research, Applications (e.g., water resource management) and EducationData Discovery: search and access to Earth Observation dataData Visualization: static browse images and dynamically constructed visualizationsData Customization: subsetting, reformatting, regridding, mosaicking, and quality screening on behalf of end usersData Processing: routine production of standard scientific datasets, converting raw data to geophysical variables.Data Analytics: end-user analysis of large datasets, such as time-averaged maps and area-averaged time series
1.6 Current Data Analysis Approach *	Standard data processing converts raw data to geophysical parameters. Though much of this is heritage custom Fortran or C code running, current prototypes are using cloud computing to scale up to rapid reprocessing campaigns.EOSDIS support of end-user analysis currently uses high-performance software, such as the netCDF Command Operators. However, current prototypes are using cloud computing and data-parallel algorithms (e.g., Spark) to achieve an order of magnitude speed-up.
1.7 Future of Application and Approach *	EOSDIS is beginning to migrate data archiving to the cloud in order to enable end users to bring algorithms to the data. We also expect to reorganize certain high-value datasets into forms that lend themselves to cloud data-parallel computing. Prototypes are underway to prove out storage schemes that are optimized for cloud analytics, such as space-time tiles stored in cloud databases and cloud filesystems.
1.8 Actors / Stakeholders	Science Research Users consume the data and apply their analysis techniques to derive knowledge of Earth System Science.Applications users consume the data for real-world practical use, such as hazard mitigation or resource management.Educational users and citizen scientists consume the data in order to understand more about the world in which they live.Satellite project and science teams use EOSDIS as a data archive and dissemination agent.
1.9 Project Goals or Objectives	The objectives are to distribute useful and usable science data and information relating to Earth system science to a diverse community.
1.10 Use Case URL(s)	https://earthdata.nasa.gov

2 Big Data Characteristics **Use Case 2-1**	
2.1 Data Source	The two most voluminous sources are:1. high spatial resolution satellite-borne instruments; and 2. long-time-series models assimilating data from satellites and instruments. Most of the Variety comes from the many field campaigns that are run to validate satellite data and explore questions that cannot be answered by spaceborne instruments alone.
2.2 Data Destination	Final results most often end up in science research papers. Data consumed by Applications users may end up in Decision Support Systems, systems that Applications users employ to properly digest and infer information from the data.
2.3 Volume	
Size	22 PB
Units	Total Earth Observation Data managed by NASA EOSDIS
Time Period	Accumulated since 1994
Proviso	
2.4 Velocity	
Unit of measure	
Time Period	
Proviso	
2.5 Variety	EOSDIS's Common Metadata Repository includes over 6400 EOSDIS data collections as of June 2017, providing significant challenges in data discovery. CMR and other interoperability frameworks (metrics, browse imagery, governance) knit together 12 different archives, each with a different implementation. Nearly all Earth science disciplines are represented in EOSDIS.
2.6 Variability	Data latency varies from Near Real Time (within 3-5 hours) to research-scale times (days to weeks time lag). Datasets also vary widely in size from small to multi-terabyte size. (Future radar data will be petabyte-scale.)

3 Big Data Science
Use Case 2-1

3.1 Veracity and Data Quality	Satellite data typically undergo extensive validation with data from aircraft, in situ, and other satellite data. In addition, the processing algorithms usually specify a quality flag for each data point, indicating a relative estimate of quality.
3.2 Visualization	Many datasets are represented in EOSDIS's Global Imagery Browse System, which supports highly interactive exploration through the Worldview imagery browser (https://worldview.earthdata.nasa.gov). In addition, dynamic, customized visualization of many data types is available through tools such as Giovanni (https://giovanni.gsfc.nasa.gov/)
3.3 Data Types	Datatypes include raster images, vector data, ASCII tables, geospatial grids of floating point values, and floating point values in satellite coordinates.
3.4 Metadata	Metadata about the data collections and their constituent files are maintained in EOSDIS Common Metadata Repository. Also, the standard data formats include self-describing formats such as Hierarchical Data Format (HDF) and network Common Data Form (netCDF), which include detailed metadata for individual variables inside the data files, such as units, standard name, fill value, scale and offset.
3.5 Curation and Governance	EOSDIS maintains an active metadata curation team that coordinates the activities of the data centers to help ensure completeness and consistency of metadata population. EOSDIS also maintains an EOSDIS Standards Office (ESO) to vet standards on data format and metadata. In addition, the 12 discipline data archives are coordinated through the Earth Science Data and Information Systems project at NASA, which oversees interoperability efforts.
3.6 Data Analytics	Analytics sometimes consists of:(1) computing statistical measures of Earth Observation data across a variety of dimensions(2) examining covariance and correlation of a variety of Earth observations(3) assimilating multiple data variables into a model using Kalman filtering(4) analyzing time series.

4 Security and Privacy
Use Case 2-1

4.1 Roles	
4.1.1 Identifying Role	System Architect
4.1.2 Investigator Affiliations	NASA

4 Security and Privacy

Use Case 2-1

4.1.3 Sponsors	NASA Program Executive for Earth Science Data Systems
4.1.4 Declarations of Potential Conflicts of Interest	
4.1.5 Institutional S/P duties	
4.1.6 Curation	Distributed Active Archive Center Manager
4.1.7 Classified Data, Code or Protocols	
Intellectual property protections	Yes
Military classifications, e.g., FOUO, or Controlled Classified	Yes
Not applicable	
Other:	
Other text	
4.1.8 Multiple Investigators | Project Leads *	
Only one investigator | project lead | developer	
Multiple team members, but in the same organization	
Multiple leads across legal organizational boundaries	Yes
Multinational investigators | project leads	
Other:	
Other text	
4.1.9 Least Privilege Role-based Access	
Yes, roles are segregated and least privilege is enforced	Yes
We do have least privilege and role separation but the admin role(s) may be too all-inclusion	
Handled at application provider level	
Handled at framework provider level	
There is no need for this feature in our application	
Could be applicable in production or future versions of our work	
Other:	
Other text	

4 Security and Privacy **Use Case 2-1**	
4.1.10 Role-based Access to Data *	
Dataset	Yes
Data record / row	
Data element / field	
Handled at application provider level	
Handled at framework provider level	
Other:	
Other text	
4.2 Personally Identifiable Information (PII)	
4.2.1 Does the System Maintain PII? *	
Yes, PII is part of this Big Data system	
No, and none can be inferred from 3rd party sources	Yes
No, but it is possible that individuals could be identified via third party databases	
Other:	
Other text	
4.2.2 Describe the PII, if applicable	
4.2.3 Additional Formal or Informal Protections for PII	
4.2.4 Algorithmic / Statistical Segmentation of Human Populations	
Yes, doing segmentation, possible discrimination issues if abused. Please also answer the next question.	
Yes, doing segmentation, but no foreseeable discrimination issues.	
Does not apply to this use case at all (e.g., no human subject data)	Yes
Other:	
Other text	

4 Security and Privacy **Use Case 2-1**	
4.2.5 Protections afforded statistical / deep learning discrimination	
4.3 Covenants, Liability, Etc.	
4.3.1 Identify any Additional Security, Compliance, Regulatory Requirements *	
FTC regulations apply	
HHS 45 CFR 46	
HIPAA	
EU General Data Protection (Reference: http://bit.ly/1Ta8S1C)	
COPPA	
Other Transborder issues	
Fair Credit Reporting Act (Reference: http://bit.ly/1Ta8XSN)	
Family Educational Rights and Protection (FERPA)	
None apply	
Other:	Yes
Other text	HSPD-12
4.3.2 Customer Privacy Promises	
Yes, we're making privacy promises to customers or subjects	
We are using a notice-and-consent model	Yes
Not applicable	
Other:	
Other text	
4.4 Ownership, Identity and Distribution	
4.4.1 Publication rights	
Open publication	Yes
Proprietary	
Traditional publisher rights (e.g., Springer, Elsevier, IEEE)	
"Big Science" tools in use	

4 Security and Privacy **Use Case 2-1**	
Other:	
Other text	
4.4.2 Chain of Trust	
4.4.3 Delegated Rights	
4.4.4 Software License Restrictions	Patents are applicable in some cases. Off-the-shelf commercial analysis packages are also used. Software which has not passed through NASA Software Release process is not eligible for public distribution.
4.4.5 Results Repository	PubMed Central (PMC)
4.4.6 Restrictions on Discovery	
4.4.7 Privacy Notices	
Privacy notices apply	
Privacy notices do not apply	Yes
Other:	
Other text	
4.4.8 Key Management	
A key management scheme is part of our system	
We are using public key infrastructure.	Yes
We do not use key management, but it could have been useful	
No readily identifiable use for key management	
Other:	
Other text	
4.4.9 Describe and Key Management Practices	
4.4.10 Is an identity framework used?	
A framework is in place. (See next question.)	Yes
Not currently using a framework.	
There is no perceived need for an identity framework.	
Other:	
Other text	
4.4.11 CAC / ECA Cards or Other Enterprise-wide Framework	

4 Security and Privacy

Use Case 2-1

Using an externally maintained enterprise-wide identity framework	Yes
Could be used, but none are available	
Not applicable	
4.4.12 Describe the Identity Framework.	
4.4.13 How is intellectual property protected?	
Login screens advising of IP issues	
Employee or team training	
Official guidelines limiting access or distribution	
Required to track all access to, distribution of digital assets	
Does not apply to this effort (e.g., public effort)	Yes
Other:	
Other text	
4.5 Risk Mitigation	
4.5.1 Are measures in place to deter re-identification? *	
Yes, in place	
Not in place, but such measures do apply	
Not applicable	Yes
Other:	
Other text	
4.5.2 Please describe any re-identification deterrents in place	
4.5.3 Are data segmentation practices being used?	
Yes, being used	
Not in use, but does apply	
Not applicable	Yes
Other:	
Other text	

4 Security and Privacy **Use Case 2-1**	
4.5.4 Is there an explicit governance plan or framework for the effort?	
Explicit governance plan	Yes
No governance plan, but could use one	
I don't think governance contributes anything to this project	
Other:	
Other text	
4.5.5 Privacy-Preserving Practices	A privacy assessment is performed for each new publicly accessible NASA system and tracked in a NASA-wide database.
4.5.6 Do you foresee any potential risks from public or private open data projects?	
Risks are known.	
Currently no known risks, but it is conceivable.	
Not sure	
Unlikely that this will ever be an issue (e.g., no PII, human-agent related data or subsystems.)	Yes
Other:	
Other text	
4.6 Provenance (Ownership)	
4.6.1 Describe your metadata management practices	
Yes, we have a metadata management system.	Yes
There is no need for a metadata management system in this use case	
It is applicable but we do not currently have one.	
Other:	
Other text	
4.6.2 If a metadata management system is present, what measures are in place to verify and protect its integrity?	

4 Security and Privacy **Use Case 2-1**	
4.6.3 Describe provenance as related to instrumentation, sensors or other devices.	
We have potential machine-to-machine traffic provenance concerns.	
Endpoint sensors or instruments have signatures periodically updated	
Using hardware or software methods, we detect and remediate outlier signatures	
Endpoint signature detection and upstream flow are built into system processing	
We rely on third party vendors to manage endpoint integrity	
We use a sampling method to verify endpoint integrity	
Not a concern at this time	Yes
Other:	
Other text	
4.7 Data Life Cycle	
4.7.1 Describe Archive Processes	
Our application has no separate "archive" process	
We offload data using certain criteria to removable media which are taken offline	
we use a multi-stage, tiered archive process	Yes
We allow for "forgetting" of individual PII on request	
Have ability to track individual data elements across all stages of processing, including archive	Yes
Additional protections, such as separate encryption, are applied to archival data	
Archived data is saved for potential later use by applications or analytics yet to be built	
Does not apply to our application	
Other:	
Other text	

4 Security and Privacy **Use Case 2-1**	
4.7.2 Describe Point in Time and Other Dependency Issues	
Some data is valid only within a point in time,	
Some data is only valid with other, related data is available or applicable, such as the existence of a building, the presence of a weather event, or the active use of a vehicle	
There are specific events in the application that render certain data obsolete or unusable	
Point and Time and related dependencies do not apply	Yes
Other:	
Other text	
4.7.3 Compliance with Secure Data Disposal Requirements	
We are required to destroy or otherwise dispose of data	
Does not apply to us	Yes
Not sure	
Other:	
Other text	
4.8 Audit and Traceability	
4.8.1 Current audit needs*	
We have third party registrar or other audits, such as for ISO 9001	Yes
We have internal enterprise audit requirements	Yes
Audit is only for system health or other management requirements	
No audit, not needed or does not apply	
Other:	
Other text	
4.8.2 Auditing versus Monitoring	
We rely on third-party or O.S. tools to audit, e.g., Windows or Linux auditing	Yes

4 Security and Privacy

Use Case 2-1

There are built-in tools for monitoring or logging that are only used for system or application health monitoring	Yes
Monitoring services include logging of role-based access to assets such as PII or other resources	
The same individual(s) in the enterprise are responsible for auditing as for monitoring	
This aspect of our application is still in flux	
Does not apply to our setting	
Other:	
Other text	
4.8.3 System Health Tools	
We rely on system-wide tools for health monitoring	Yes
We built application health tools specifically to address integrity, performance monitoring and related concerns	Yes
There is no need in our setting	
Other:	
Other text	
4.8.4 What events are currently audited? *	
All data access must be audited	
Only selected / protected data must be audited	Yes
Maintenance on user roles must be audited (new users, disabled user, updated roles or permissions)	Yes
Purge and archive events	
Domain-dependent events (e.g., adding a new sensor)	
REST or SOAP events	
Changes in system configuration	Yes
Organizational changes	
External project ownership / management changes	
Requirements are externally set, e.g., by PCI compliance	

4 Security and Privacy **Use Case 2-1**	
Domain-specific events (patient death in a drug trial)	
Other:	
Other text	
4.9 Application Provider Security	
4.9.1 Describe Application Provider Security *	
There is a security mechanism implemented at the application level	
The app provider level is aware of PII or privacy data elements	
The app provider implements audit and logging	
The app provider security relies on framework-level security for its operation	
Does not apply to our application	Yes
Other:	
Other text	
4.10 Framework Provider Security	
4.10.1 Describe the framework provider security *	
Security is implemented at the framework level	
Roles can be defined at the framework level	
The framework level is aware of PII or related sensitive data	
Does not apply in our setting	Yes
Is provided by the Big Data tool	
Other:	
Other text	
4.11 System Health	
4.11.1 Measures to Ensure Availability *	
Deterrents to man-in-the-middle attacks	
Deterrents to denial of service attacks	

4 Security and Privacy **Use Case 2-1**	
Replication, redundancy or other resilience measures	
Deterrents to data corruption, drops or other critical big data components	
Other:	
Other text	
4.12 Permitted Use Cases	
4.12.1 Describe Domain-specific Limitations on Use	
4.12.2 Paywall	
A paywall is in use at some stage in the workflow	
Not applicable	

5 Classify Use Cases with Tags **Use Case 2-1**	
5.1 DATA: Application Style and Data sharing and acquisition	
Uses Geographical Information Systems?	Yes
Use case involves Internet of Things?	
Data comes from HPC or other simulations?	Yes
Data Fusion important?	Yes
Data is Real time Streaming?	
Data is Batched Streaming (e.g. collected remotely and uploaded every so often)?	Yes
Important Data is in a Permanent Repository (Not streamed)?	Yes
Transient Data important?	Yes
Permanent Data Important?	Yes
Data shared between different applications/users?	Yes
Data largely dedicated to only this use case?	
5.2 DATA: Management and Storage	

5 Classify Use Cases with Tags **Use Case 2-1**	
Application data system based on Files?	Yes
Application data system based on Objects?	
Uses HDFS style File System?	
Uses Wide area File System like Lustre?	
Uses HPC parallel file system like GPFS?	
Uses SQL?	Yes
Uses NoSQL?	Yes
Uses NewSQL?	
Uses Graph Database?	
5.3 DATA: Describe Other Data Acquisition/ Access/ Sharing/ Management/ Storage Issues	
5.4 ANALYTICS: Data Format and Nature of Algorithm used in Analytics	
Data regular?	Yes
Data dynamic?	
Algorithm O(N^2) ?	
Basic statistics (regression, moments) used?	Yes
Search/Query/Index of application data Important?	
Classification of data Important?	Yes
Recommender Engine Used?	
Clustering algorithms used?	Yes
Alignment algorithms used?	
(Deep) Learning algorithms used?	
Graph Analytics Used?	
5.5 ANALYTICS: Describe Other Data Analytics Used	
5.6 PROGRAMMING MODEL	

5 Classify Use Cases with Tags

Use Case 2-1

Pleasingly parallel Structure? Parallel execution over independent data. Called Many Task or high throughput computing. MapReduce with only Map and no Reduce of this type	Yes
Use case NOT Pleasingly Parallel -- Parallelism involves linkage between tasks. MapReduce (with Map and Reduce) of this type	
Uses Classic MapReduce? such as Hadoop	
Uses Apache Spark or similar Iterative MapReduce?	Yes
Uses Graph processing as in Apache Giraph?	
Uses MPI (HPC Communication) and/or Bulk Synchronous Processing BSP?	
Dataflow Programming Model used?	
Workflow or Orchestration software used?	Yes
Python or Scripting front ends used? Maybe used for orchestration	Yes
Shared memory architectures important?	
Event-based Programming Model used?	
Agent-based Programming Model used?	
Use case I/O dominated? I/O time > or >> Compute time	Yes
Use case involves little I/O? Compute >> I/O	
5.7 Other Programming Model Tags	
5.8 Please Estimate Ratio I/O Bytes/Flops	
5.9 Describe Memory Size or Access issues	

6 Overall Big Data Issues

Use Case 2-1

6.1 Other Big Data Issues	Currently, the Variety in Big Data is producing a set of data discovery issues for the end users. Searching for datasets turns out to be different from searching for documents in a variety of subtle, but important, ways.
6.2 User Interface and Mobile Access Issues	
6.3 List Key Features and Related Use Cases	
6.4 Project Future	More data will be stored in the cloud, likely with copies in some cases of reorganized data in order to make them more tractable to data-parallel algorithms. More analysis support will also be offered to users that want to run analyses of data n the cloud.

7 Workflow Processes

Use Case 2-1

7.1 Please comment on workflow processes	Satellite Data Processing commonly goes through the following processing steps: Level 0 - raw data in files, de-duplicatedLevel 1 - calibrated data with geolocation Level 2 - inferred geophysical measurements, in sensor coordinates Level 3 - geophysical measurements Level 4 - model output (usually done outside EOSDIS)The characteristics of the data, especially their geolocations vary significantly from L0 to L1, and from L2 to L3. The usability to various audiences crosses a significant border between L1 and L2.
7.2 Workflow details for each stage *	
7.2.1 Workflow Details for Stage 1	
Stage 1 Name	Level 0 Processing
Data Source(s)	Satellite downlink station
Nature of Data	Packets of raw data
Software Used	Custom software
Data Analytics	Reordering of packets into time order, deduplication
Infrastructure	Local servers
Percentage of Use Case Effort	
Other Comments	
7.2.2 Workflow Details for Stage 2	
Stage 2 Name	Level 1b Processing

7 Workflow Processes

Use Case 2-1

Data Source(s)	EOS Data Operations System (Level 0 processor)
Nature of Data	Files of cleaned-up raw data
Software Used	Instrument-specific calibration codes
Data Analytics	Geolocation and calibration of raw data
Infrastructure	Multiple local servers
Percentage of Use Case Effort	
Other Comments	
7.2.3 Workflow Details for Stage 3	
Stage 3 Name	Level 2 Processing
Data Source(s)	Level 1B processing system
Nature of Data	Level 1B geolocated, calibrated data
Software Used	Scientist-authored physical retrieval code
Data Analytics	Transform calibrated data (radiances, waveforms, ...) into geophysical measurements
Infrastructure	Large compute clusters
Percentage of Use Case Effort	
Other Comments	
7.2.4 Workflow Details for Stage 4	
Stage 4 Name	Level 3 Processing
Data Source(s)	Level 2 Processor
Nature of Data	Geophysical variables in sensor coordinates
Software Used	Scientist-authored gridding code
Data Analytics	Data projection and aggregation over space and/or time
Infrastructure	Compute clusters with large amounts of disk space
Percentage of Use Case Effort	
Other Comments	
7.2.5 Workflow Details for Stages 5 and any further stages	
Stage 5 Name	
Data Source(s)	
Nature of Data	
Software Used	
Data Analytics	
Infrastructure	
Percentage of Use Case Effort	
Other Comments	

F.2 Use Case 2-2: NASA Earth Observing System Data and Information System (EOSDIS)

1 Overall Project Description Use Case 2-2	
1.1 Use Case Title *	Web-Enabled Landsat Data (WELD) Processing
1.2 Use Case Description *	The use case is specific to the part of the project where data is available on the HPC platform and processed through the science workflow. It is a 32-stage processing pipeline that includes two separate science products (Top-of-the-Atmosphere (TOA) reflectances and surface reflectances) as well as QA and visualization components.
1.3 Use Case Contacts *	
	Andrew Michaelis
	Author
	Yes
	Yes
1.4 Domain ("Vertical") *	Land use science: image processing
1.5 Application *	The product of this use case is a dataset of science products of use to the land surface science community that is made freely available by NASA. The dataset is produced through processing of images from the Landsat 4, 5, and 7 satellites.
1.6 Current Data Analysis Approach *	>> Compute System: Shared High Performance Computing (HPC) system at NASA Ames Research Center (Pleiades) >> Storage: NASA Earth Exchange (NEX) NFS storage system for read-only data storage (2.5PB), Lustre for read-write access during processing (1PB), tape for near-line storage (50PB) >> Networking: InfiniBand partial hypercube internal interconnect within the HPC system; 1G to 10G connection to external data providers >> Software: NEX science platform – data management, workflow processing, provenance capture; WELD science processing algorithms from South Dakota State University (SDSU), browse visualization, and time-series code; Global Imagery Browse Service (GIBS) data visualization platform; USGS data distribution platform. Custom-built application and libraries built on top of open-source libraries.
1.7 Future of Application and Approach *	Processing will be improved with newer and updated algorithms. This process may also be applied to future datasets and processing systems (Landsat 8 and Sentinel-2 satellites, for example)

1 Overall Project Description Use Case 2-2	
1.8 Actors / Stakeholders	South Dakota State University – science, algorithm development, QA, data browse visualization and distribution framework; NASA Advanced Supercomputing Division at NASA Ames Research Center – data processing at scale; USGS – data source and data distribution; NASA GIBS – native resolution data visualization; NASA HQ and NASA EOSDIS – sponsor.
1.9 Project Goals or Objectives	The WELD products are developed specifically to provide consistent data that can be used to derive land cover as well as geophysical and biophysical products for assessment of surface dynamics and to study Earth system functioning. The WELD products are free and are available via the Internet. The WELD products are processed so that users do not need to apply the equations and spectral calibration coefficients and solar information to convert the Landsat digital numbers to reflectance and brightness temperature, and successive products are defined in the same coordinate system and align precisely, making them simple to use for multi-temporal applications.
1.10 Use Case URL(s)	http://globalmonitoring.sdstate.edu/projects/weldglobal/gweld.html http://globalweld.cr.usgs.gov/ https://nex.nasa.gov http://www.nas.nasa.gov/hecc/resources/pleiades.html https://earthdata.nasa.gov/about/science-system-description/eosdis-components/global-imagery-browse-services-gibs https://worldview.earthdata.nasa.gov/

2 Big Data Characteristics Use Case 2-2	
2.1 Data Source	Satellite Earth observation data from Landsat 4, 5, and 7 missions. The data source is remote and centralized – distributed from USGS EROS Center.
2.2 Data Destination	The final data is distributed by USGS EROS Center – a remote centralized data system. It is also available on the NEX platform for further analysis and product development.
2.3 Volume	
Size	30PB of processed data through the pipeline (1PB inputs, 10PB intermediate, 6PB outputs)
Units	Petabytes of data that flow through the processing pipeline
Time Period	Data was collected over a period of 27 years and is being processed over a period of 5 years

2 Big Data Characteristics
Use Case 2-2

Proviso	The data represent the operational time period of 1984 to 2011 for the Landsat 4, 5, and 7 satellites
2.4 Velocity	
Unit of measure	Terabytes processed per day during processing time periods: 150 TB/day
Time Period	24 hours
Proviso	Based on programmatic goals of processing several iterations of the final product over the span of the project. Observed run-time and volumes during processing
2.5 Variety	This use case basically deals with a single dataset.
2.6 Variability	Not clear what the difference is between variability and variety. This use case basically deals with a single dataset.

3 Big Data Science
Use Case 2-2

3.1 Veracity and Data Quality	This data dealt with in this use case are a high-quality, curated dataset.
3.2 Visualization	Visualization is not used in this use case per se, but visualization is important in QA processes conducted outside of the use case as well as in the ultimate use by scientists of the product datasets that result from this use case
3.3 Data Types	structured image data
3.4 Metadata	Metadata adhere to accepted metadata standards widely used in the earth science imagery field.
3.5 Curation and Governance	Data is governed by NASA data release policy; data is referred to by the DOI and the algorithms have been peer-reviewed. The data distribution center and the PI are responsible for science data support.
3.6 Data Analytics	There are number of analytics processes throughout the processing pipeline. The key analytics is identifying best available pixels for spatio-temporal composition and spatial aggregation processes as a part of the overall QA. The analytics algorithms are custom developed for this use case.

4 Security and Privacy
Use Case 2-2

4.1 Roles	
4.1.1 Identifying Role	PI; Project sponsor (NASA EOSDIS program)
4.1.2 Investigator Affiliations	Andrew Michaelis, NASA, NEX Processing Pipeline Development and Operations David Roy, South Dakota State University, Project PI Hankui Zhang, South Dakota State University, Science Algorithm Development Adam Dosch, South Dakota State University, SDSU operations/data management Lisa Johnson, USGS, Data Distribution Matthew Cechini, Ryan Boller, Kevin Murphy, NASA, GIBS project
4.1.3 Sponsors	NASA EOSDIS project
4.1.4 Declarations of Potential Conflicts of Interest	None
4.1.5 Institutional S/P duties	None
4.1.6 Curation	Joint responsibility of NASA, USGS, and Principal Investigator
4.1.7 Classified Data, Code or Protocols	
Intellectual property protections	Off
Military classifications, e.g., FOUO, or Controlled Classified	Off
Not applicable	Yes
Other:	Off
Other text	
4.1.8 Multiple Investigators | Project Leads*	
Only one investigator | project lead | developer	Off
Multiple team members, but in the same organization	Off
Multiple leads across legal organizational boundaries	Yes
Multinational investigators | project leads	Off
Other:	Off
Other text	
4.1.9 Least Privilege Role-based Access	

4 Security and Privacy **Use Case 2-2**	
Yes, roles are segregated and least privilege is enforced	Off
We do have least privilege and role separation but the admin role(s) may be too all-inclusion	Off
Handled at application provider level	Off
Handled at framework provider level	Off
There is no need for this feature in our application	Off
Could be applicable in production or future versions of our work	Off
Other:	Yes
Other text	Not used
4.1.10 Role-based Access to Data *	
Dataset	Yes
Data record / row	Off
Data element / field	Off
Handled at application provider level	Off
Handled at framework provider level	Off
Other:	Off
Other text	
4.2 Personally Identifiable Information (PII)	
4.2.1 Does the System Maintain PII? *	
Yes, PII is part of this Big Data system	Off
No, and none can be inferred from 3rd party sources	Yes
No, but it is possible that individuals could be identified via third party databases	Off
Other:	Off
Other text	
4.2.2 Describe the PII, if applicable	

4 Security and Privacy **Use Case 2-2**	
4.2.3 Additional Formal or Informal Protections for PII	
4.2.4 Algorithmic / Statistical Segmentation of Human Populations	
Yes, doing segmentation, possible discrimination issues if abused. Please also answer the next question.	Off
Yes, doing segmentation, but no foreseeable discrimination issues.	Off
Does not apply to this use case at all (e.g., no human subject data)	Yes
Other:	Off
Other text	
4.2.5 Protections afforded statistical / deep learning discrimination	Not applicable to this use case.
4.3 Covenants, Liability, Etc.	
4.3.1 Identify any Additional Security, Compliance, Regulatory Requirements *	
FTC regulations apply	Off
HHS 45 CFR 46	Off
HIPAA	Off
EU General Data Protection (Reference: http://bit.ly/1Ta8S1C)	Off
COPPA	Off
Other Transborder issues	Off
Fair Credit Reporting Act (Reference: http://bit.ly/1Ta8XSN)	Off
Family Educational Rights and Protection (FERPA)	Off
None apply	Yes
Other:	Off
Other text	

4 Security and Privacy **Use Case 2-2**	
4.3.2 Customer Privacy Promises	
Yes, we're making privacy promises to customers or subjects	Off
We are using a notice-and-consent model	Off
Not applicable	Yes
Other:	Off
Other text	
4.4 Ownership, Identity and Distribution	
4.4.1 Publication rights	
Open publication	Off
Proprietary	Off
Traditional publisher rights (e.g., Springer, Elsevier, IEEE)	Off
"Big Science" tools in use	Off
Other:	Yes
Other text	Datasets produced are available to the public with a requirement for appropriate citation when used.
4.4.2 Chain of Trust	None
4.4.3 Delegated Rights	None
4.4.4 Software License Restrictions	None
4.4.5 Results Repository	The datasets produced from this dataset are distributed to the public from repositories at the USGS EROS Center and the NASA EOSDIS program.
4.4.6 Restrictions on Discovery	None
4.4.7 Privacy Notices	
Privacy notices apply	Off
Privacy notices do not apply	Yes
Other:	Off
Other text	
4.4.8 Key Management	
A key management scheme is part of our system	Off
We are using public key infrastructure.	Off

4 Security and Privacy **Use Case 2-2**	
We do not use key management, but it could have been useful	Off
No readily identifiable use for key management	Yes
Other:	Off
Other text	
4.4.9 Describe and Key Management Practices	
4.4.10 Is an identity framework used?	
A framework is in place. (See next question.)	Off
Not currently using a framework.	Off
There is no perceived need for an identity framework.	Yes
Other:	Off
Other text	
4.4.11 CAC / ECA Cards or Other Enterprise-wide Framework	
Using an externally maintained enterprise-wide identity framework	Off
Could be used, but none are available	Off
Not applicable	Yes
4.4.12 Describe the Identity Framework.	
4.4.13 How is intellectual property protected?	
Login screens advising of IP issues	Off
Employee or team training	Off
Official guidelines limiting access or distribution	Off
Required to track all access to, distribution of digital assets	Off
Does not apply to this effort (e.g., public effort)	Off
Other:	Yes

4 Security and Privacy
Use Case 2-2

Other text	Believe there are standards for citation of datasets that apply to use of the datasets from the USGS or NASA repositories.
4.5 Risk Mitigation	
4.5.1 Are measures in place to deter re-identification? *	
Yes, in place	Off
Not in place, but such measures do apply	Off
Not applicable	Yes
Other:	Off
Other text	
4.5.2 Please describe any re-identification deterrents in place	
4.5.3 Are data segmentation practices being used?	
Yes, being used	Off
Not in use, but does apply	Off
Not applicable	Yes
Other:	Off
Other text	
4.5.4 Is there an explicit governance plan or framework for the effort?	
Explicit governance plan	Off
No governance plan, but could use one	Off
I don't think governance contributes anything to this project	Off
Other:	Yes
Other text	Resulting datasets are governed by the data access policies of the USGS and NASA.
4.5.5 Privacy-Preserving Practices	None
4.5.6 Do you foresee any potential risks from public or private open data projects?	
Risks are known.	Off

4 Security and Privacy **Use Case 2-2**	
Currently no known risks, but it is conceivable.	Off
Not sure	Yes
Unlikely that this will ever be an issue (e.g., no PII, human-agent related data or subsystems.)	Off
Other:	Off
Other text	
4.6 Provenance (Ownership)	
4.6.1 Describe your metadata management practices	
Yes, we have a metadata management system.	Off
There is no need for a metadata management system in this use case	Off
It is applicable but we do not currently have one.	Off
Other:	Yes
Other text	There is no metadata management system within this use case, but the resultant datasets' metadata is managed as NASA EOSDIS datasets.
4.6.2 If a metadata management system is present, what measures are in place to verify and protect its integrity?	
4.6.3 Describe provenance as related to instrumentation, sensors or other devices.	
We have potential machine-to-machine traffic provenance concerns.	Off
Endpoint sensors or instruments have signatures periodically updated	Off
Using hardware or software methods, we detect and remediate outlier signatures	Off
Endpoint signature detection and upstream flow are built into system processing	Off

4 Security and Privacy **Use Case 2-2**	
We rely on third party vendors to manage endpoint integrity	Off
We use a sampling method to verify endpoint integrity	Off
Not a concern at this time	Off
Other:	Off
Other text	
4.7 Data Life Cycle	
4.7.1 Describe Archive Processes	
Our application has no separate "archive" process	Off
We offload data using certain criteria to removable media which are taken offline	Off
we use a multi-stage, tiered archive process	Off
We allow for "forgetting" of individual PII on request	Off
Have ability to track individual data elements across all stages of processing, including archive	Off
Additional protections, such as separate encryption, are applied to archival data	Off
Archived data is saved for potential later use by applications or analytics yet to be built	Off
Does not apply to our application	Off
Other:	Yes
Other text	Resultant datasets are not archived per se, but the repositories do have a stewardship responsibility.
4.7.2 Describe Point in Time and Other Dependency Issues	
Some data is valid only within a point in time,	Off
Some data is only valid with other, related data is available or applicable, such as the existence of a building, the presence of a weather event, or the active use of a vehicle	Off

4 Security and Privacy
Use Case 2-2

There are specific events in the application that render certain data obsolete or unusable	Off
Point and Time and related dependencies do not apply	Off
Other:	Yes
Other text	Data are relevant and valid independent of when accessed/used, but all data have a specific date/time/location reference that is part of the metadata.
4.7.3 Compliance with Secure Data Disposal Requirements	
We are required to destroy or otherwise dispose of data	Off
Does not apply to us	Yes
Not sure	Off
Other:	Off
Other text	
4.8 Audit and Traceability	
4.8.1 Current audit needs *	
We have third party registrar or other audits, such as for ISO 9001	Off
We have internal enterprise audit requirements	Off
Audit is only for system health or other management requirements	Off
No audit, not needed or does not apply	Yes
Other:	Off
Other text	
4.8.2 Auditing versus Monitoring	
We rely on third party or O.S. tools to audit, e.g., Windows or Linux auditing	Off
There are built-in tools for monitoring or logging that are only used for system or application health monitoring	Off

4 Security and Privacy **Use Case 2-2**	
Monitoring services include logging of role-based access to assets such as PII or other resources	Off
The same individual(s) in the enterprise are responsible for auditing as for monitoring	Off
This aspect of our application is still in flux	Off
Does not apply to our setting	Yes
Other:	Off
Other text	
4.8.3 System Health Tools	
We rely on system-wide tools for health monitoring	Off
We built application health tools specifically to address integrity, performance monitoring and related concerns	Off
There is no need in our setting	Off
Other:	Yes
Other text	Systems employed in the use case are operated and maintained by the NASA Advanced Supercomputing Division and the use case staff do not have to deal with system health. Repositories for the resultant data are operated and maintained under the auspices of NASA and the USGS.
4.8.4 What events are currently audited? *	
All data access must be audited	Off
Only selected / protected data must be audited	Off
Maintenance on user roles must be audited (new users, disabled user, updated roles or permissions)	Off
Purge and archive events	Off
Domain-dependent events (e.g., adding a new sensor)	Off
REST or SOAP events	Off
Changes in system configuration	Off
Organizational changes	Off
External project ownership / management changes	Off

4 Security and Privacy **Use Case 2-2**	
Requirements are externally set, e.g., by PCI compliance	Off
Domain-specific events (patient death in a drug trial)	Off
Other:	Yes
Other text	None
4.9 Application Provider Security	
4.9.1 Describe Application Provider Security *	
There is a security mechanism implemented at the application level	Off
The app provider level is aware of PII or privacy data elements	Off
The app provider implements audit and logging	Off
The app provider security relies on framework-level security for its operation	Off
Does not apply to our application	Yes
Other:	Off
Other text	
4.10 Framework Provider Security	
4.10.1 Describe the framework provider security *	
Security is implemented at the framework level	Off
Roles can be defined at the framework level	Off
The framework level is aware of PII or related sensitive data	Off
Does not apply in our setting	Yes
Is provided by the Big Data tool	Off
Other:	Off
Other text	
4.11 System Health	
4.11.1 Measures to Ensure Availability *	

4 Security and Privacy
Use Case 2-2

Deterrents to man-in-the-middle attacks	Off
Deterrents to denial of service attacks	Off
Replication, redundancy or other resilience measures	Off
Deterrents to data corruption, drops or other critical big data components	Off
Other:	Yes
Other text	System resources are provided by the NASA Advanced Supercomputing Division (NAS) for the use case; NAS has responsibility for system availability.
4.12 Permitted Use Cases	
4.12.1 Describe Domain-specific Limitations on Use	None
4.12.2 Paywall	
A paywall is in use at some stage in the workflow	Off
Not applicable	Yes

5 Classify Use Cases with Tags
Use Case 2-2

5.1 DATA: Application Style and Data sharing and acquisition

Uses Geographical Information Systems?	Off
Use case involves Internet of Things?	Off
Data comes from HPC or other simulations?	Off
Data Fusion important?	Off
Data is Real time Streaming?	Off
Data is Batched Streaming (e.g. collected remotely and uploaded every so often)?	Yes
Important Data is in a Permanent Repository (Not streamed)?	Off
Transient Data important?	Off
Permanent Data Important?	Yes

5 Classify Use Cases with Tags **Use Case 2-2**	
Data shared between different applications/users?	Yes
Data largely dedicated to only this use case?	Off
5.2 DATA: Management and Storage	
Application data system based on Files?	Yes
Application data system based on Objects?	Off
Uses HDFS style File System?	Off
Uses Wide area File System like Lustre?	Yes
Uses HPC parallel file system like GPFS?	Off
Uses SQL?	Off
Uses NoSQL?	Off
Uses NewSQL?	Off
Uses Graph Database?	Off
5.3 DATA: Describe Other Data Acquisition/ Access/ Sharing/ Management/ Storage Issues	
5.4 ANALYTICS: Data Format and Nature of Algorithm used in Analytics	
Data regular?	Yes
Data dynamic?	Off
Algorithm O(N^2) ?	Off
Basic statistics (regression, moments) used?	Off
Search/Query/Index of application data Important?	Off
Classification of data Important?	Yes
Recommender Engine Used?	Off
Clustering algorithms used?	Off
Alignment algorithms used?	Off
(Deep) Learning algorithms used?	Off
Graph Analytics Used?	Off
5.5 ANALYTICS: Describe Other Data Analytics Used	None
5.6 PROGRAMMING MODEL	

5 Classify Use Cases with Tags
Use Case 2-2

Pleasingly parallel Structure? Parallel execution over independent data. Called Many Task or high throughput computing. MapReduce with only Map and no Reduce of this type	Off
Use case NOT Pleasingly Parallel -- Parallelism involves linkage between tasks. MapReduce (with Map and Reduce) of this type	Off
Uses Classic MapReduce? such as Hadoop	Off
Uses Apache Spark or similar Iterative MapReduce?	Off
Uses Graph processing as in Apache Giraph?	Off
Uses MPI (HPC Communication) and/or Bulk Synchronous Processing BSP?	Off
Dataflow Programming Model used?	Off
Workflow or Orchestration software used?	Off
Python or Scripting front ends used? Maybe used for orchestration	Off
Shared memory architectures important?	Off
Event-based Programming Model used?	Off
Agent-based Programming Model used?	Off
Use case I/O dominated? I/O time > or >> Compute time	Off
Use case involves little I/O? Compute >> I/O	Off
5.7 Other Programming Model Tags	
5.8 Please Estimate Ratio I/O Bytes/Flops	Do not have the data to develop this ratio.
5.9 Describe Memory Size or Access issues	None

6 Overall Big Data Issues
Use Case 2-2

6.1 Other Big Data Issues	

6.2 User Interface and Mobile Access Issues	No mobile access is applicable to this use case.
6.3 List Key Features and Related Use Cases	
6.4 Project Future	Processing will be improved with newer and updated algorithms. This process may also be applied to future datasets and processing systems (Landsat 8 and Sentinel-2 satellites, for example).

7 Workflow Processes
Use Case 2-2

7.1 Please comment on workflow processes	The processing for this use case is a 32-stage pipeline. The WELD-Overview diagram presents a five-stage high-level workflow. Workflow details are not available at this time, but may be provided in the future if time allows. A top-level workflow diagram is being emailed separately.
7.2 Workflow details for each stage *	
7.2.1 Workflow Details for Stage 1	
Stage 1 Name	
Data Source(s)	
Nature of Data	
Software Used	
Data Analytics	
Infrastructure	
Percentage of Use Case Effort	
Other Comments	
7.2.2 Workflow Details for Stage 2	
Stage 2 Name	
Data Source(s)	
Nature of Data	
Software Used	
Data Analytics	
Infrastructure	
Percentage of Use Case Effort	
Other Comments	
7.2.3 Workflow Details for Stage 3	
Stage 3 Name	
Data Source(s)	

Nature of Data
Software Used
Data Analytics
Infrastructure
Percentage of Use Case Effort
Other Comments
7.2.4 Workflow Details for Stage 4
Stage 4 Name
Data Source(s)
Nature of Data
Software Used
Data Analytics
Infrastructure
Percentage of Use Case Effort
Other Comments
7.2.5 Workflow Details for Stages 5 and any further stages
Stage 5 Name
Data Source(s)
Nature of Data
Software Used
Data Analytics
Infrastructure
Percentage of Use Case Effort
Other Comments

Appendix G: Acronyms

2D and 3D	two- and three-dimensional
6D	six-dimensional
AOD	Analysis Object Data
API	application programming interface
ASDC	Atmospheric Science Data Center
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
CBSP	Cloud Brokerage Service Provider
CCP	Climate Change Prediction
CERES	Clouds and Earth's Radiant Energy System
CERN	European Organization for Nuclear Research
CES21	California Energy Systems for the 21st Century
CESM	Community Earth System Model
CFTC	U.S. Commodity Futures Trading Commission
CIA	confidentiality, integrity, and availability
CMIP	Coupled Model Intercomparison Project
CMIP5	Climate Model Intercomparison Project
CMS	Compact Muon Solenoid
CNRS	Centre National de la Recherche Scientifique
COSO	Committee of Sponsoring Organizations
CP	charge parity
CPR	Capability Provider Requirements
CPU	central processing unit
CReSIS	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
DCR	Data Consumer Requirements
DES	Dark Energy Survey
DFC	DataNet Federation Consortium
DHTC	Distributed High Throughput Computing
DOE	U.S. Department of Energy
DOJ	U.S. Department of Justice
DPO	Data Products Online
DSR	Data Source Requirements
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
ERC	European Research Council
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	AmeriFlux and Flux Tower Network
FMV	full motion video
FNAL	Fermi National Accelerator Laboratory
GAAP	U.S. Generally Accepted Accounting Practices
GB	gigabyte
GCM	general circulation model
GEOS-5	Goddard Earth Observing System version 5
GEWaSC	Genome-Enabled Watershed Simulation Capability
GHG	greenhouse gas
GISs	geographic information systems
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
ICA	independent component analysis
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
ITL	Information Technology Laboratory
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
LMR	Life cycle Management Requirements
LOB	lines of business
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
NBD-PWG	NIST Big Data Public Working Group
NBDRA.	NIST Big Data Reference Architecture
NCAR	National Center for Atmospheric Research
NCBI	National Center for Biotechnology Information
NCCS	NASA Center for Climate Simulation
NEO	near-Earth
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
NLP	natural language processing
NRT	Near Real Time
NSF	National Science Foundation
ODAS	Ocean Modeling and Data Assimilation
ODP	Open Distributed Processing
OGC	Open Geospatial Consortium
OLAP	online analytical processing
OpenAIRE	Open Access Infrastructure for Research in Europe
OR	Other Requirements
PB	petabyte
PCA	principal component analysis
PCAOB	Public Company Accounting and Oversight Board
PHO	planetary hazard
PID	persistent identification
PII	Personally Identifiable Information
PNNL	Pacific Northwest National Laboratory
PR	Public Relations
RDBMS	relational database management system
RDF	Resource Description Framework
ROI	return on investment
RPI	Repeat Pass Interferometry
RPO	Recovery Point Objective
RTO	Response Time Objective
SAN	storage area network
SAR	Synthetic aperture radar

SAR	Synthetic Aperture Radar
SDLC/HDLC	Software Development Life Cycle/Hardware Development Life Cycle
SDN	software-defined networking
SEC	U.S. Securities and Exchange Commission
SFA 2.0	Scientific Focus Area 2.0 Science Plan
SIEM	Security Incident/Event Management
SIOS	Svalbard Integrated Arctic Earth Observing System
SOAP	Simple Object Access Protocol
SOX	Sarbanes–Oxley Act of 2002
SPADE	Support for Provenance Auditing in Distributed Environments
SPR	Security and Privacy Requirements
SSH	Secure Shell
SSO	Single sign-on capability
tf-idf	term frequency–inverse document frequency
TPR	Transformation Provider Requirements
UA	University of Arizona
UAVSAR	Unmanned Air Vehicle Synthetic Aperture Radar
UI	user interface
UPS	United Parcel Service
UQ	uncertainty quantification
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

Appendix H: References

[1] W. Chang and NIST Big Data Public Working Group, "NIST Big Data Interoperability Framework: Volume 1, Definitions (SP1500-1)," 2015.

[2] W. Chang and NIST Big Data Public Working Group, "NIST Big Data Interoperability Framework: Volume 2, Big Data Taxonomies (SP1500-2)," 2015.

[3] W. Chang and NIST Big Data Public Working Group, "NIST Big Data Interoperability Framework: Volume 4, Security and Privacy (SP1500-4)," 2015.

[4] W. Chang and NIST Big Data Public Working Group, "NIST Big Data Interoperability Framework: Volume 5, Architectures White Paper Survey (SP1500-5)," 2015.

[5] W. Chang and NIST Big Data Public Working Group, "NIST Big Data Interoperability Framework: Volume 6, Reference Architecture (SP1500-6)," 2015.

[6] W. Chang and NIST Big Data Public Working Group, "NIST Big Data Interoperability Framework: Volume 7, Standards Roadmap (SP1500-7)," 2015.

[7] W. Chang and NIST Big Data Public Working Group, "NIST Big Data Interoperability Framework: Volume 8, Reference Architecture Interface (SP1500-9)," 2017.

[8] W. Chang and NIST Big Data Public Working Group, "NIST Big Data Interoperability Framework: Volume 9, Adoption and Modernization (SP1500-10)," 2017.

[9] T. White House Office of Science and Technology Policy, "Big Data is a Big Deal," *OSTP Blog*, 2012. [Online]. Available: http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal. [Accessed: 21-Feb-2014].

[10] Multiple Federal Agencies, "Materials Genome Initiative," *The White House, President Barak Obama*, 2011. [Online]. Available: https://www.mgi.gov/. [Accessed: 15-Dec-2014].

[11] Multiple Federal Agencies, "Open Government Initiative," *The White House, President Barak Obama*, 2013. [Online]. Available: https://www.whitehouse.gov/open.

[12] N. Allmang and J. A. Remshard, "NIKE: Integrating workflow, digital library, and online catalog systems," in *Proceedings of the ACM IEEE International Conference on Digital Libraries, JCDL 2004*, 2004, p. 399.

[13] J. Greenberg, K. Jeffery, R. Koskela, and A. Ball, "Metadata Standards Directory WG," *Research Data Alliance*. [Online]. Available: https://rd-alliance.org/group/metadata-standards-directory-working-group.html. [Accessed: 28-Sep-2014].