**Use Cases from NBD(NIST Big Data) Requirements WG V1.0**

<http://bigdatawg.nist.gov/home.php>

**Contents**

1. Blank Template

**Government Operation**

1. Census 2010 and 2000 – Title 13 Big Data; Vivek Navale & Quyen Nguyen, NARA
2. National Archives and Records Administration Accession NARA, Search, Retrieve, Preservation; Vivek Navale & Quyen Nguyen, NARA
3. Statistical Survey Response Improvement (Adaptive Design); Cavan Capps, U.S. Census Bureau
4. Non-Traditional Data in Statistical Survey Response Improvement (Adaptive Design); Cavan Capps, U.S. Census Bureau

**Commercial**

1. Cloud Eco-System, for Financial Industries (Banking, Securities & Investments, Insurance) transacting business within the United States; Pw Carey, Compliance Partners, LLC
2. Mendeley – An International Network of Research; William Gunn , Mendeley
3. Netflix Movie Service; Geoffrey Fox, Indiana University
4. Web Search; Geoffrey Fox, Indiana University
5. IaaS (Infrastructure as a Service) Big Data Business Continuity & Disaster Recovery (BC/DR) Within A Cloud Eco-System; Pw Carey, Compliance Partners, LLC
6. Cargo Shipping; William Miller, MaCT USA
7. Materials Data for Manufacturing; John Rumble, R&R Data Services
8. Simulation driven Materials Genomics; David Skinner, LBNL

**Defense**

1. Large Scale Geospatial Analysis and Visualization; David Boyd, Data Tactics
2. Object identification and tracking from Wide Area Large Format Imagery (WALF) Imagery or Full Motion Video (FMV) – Persistent Surveillance; David Boyd, Data Tactics
3. Intelligence Data Processing and Analysis; David Boyd, Data Tactics

**Healthcare and Life Sciences**

1. Electronic Medical Record (EMR) Data; Shaun Grannis, Indiana University
2. Pathology Imaging/digital pathology; Fusheng Wang, Emory University
3. Computational Bioimaging; David Skinner, Joaquin Correa, Daniela Ushizima, Joerg Meyer, LBNL
4. Genomic Measurements; Justin Zook, NIST
5. Comparative analysis for metagenomes and genomes; Ernest Szeto, LBNL (Joint Genome Institute)
6. Individualized Diabetes Management; Ying Ding , Indiana University
7. Statistical Relational Artificial Intelligence for Health Care; Sriraam Natarajan, Indiana University
8. World Population Scale Epidemiological Study; Madhav Marathe, Stephen Eubank or Chris Barrett, Virginia Tech
9. Social Contagion Modeling for Planning, Public Health and Disaster Management; Madhav Marathe or Chris Kuhlman, Virginia Tech
10. Biodiversity and LifeWatch; Wouter Los, Yuri Demchenko, University of Amsterdam

**Deep Learning and Social Media**

1. Large-scale Deep Learning; Adam Coates , Stanford University
2. Organizing large-scale, unstructured collections of consumer photos; David Crandall, Indiana University
3. Truthy: Information diffusion research from Twitter Data; Filippo Menczer, Alessandro Flammini, Emilio Ferrara, Indiana University
4. Crowd Sourcing in the Humanities as Source for Big and Dynamic Data; Sebastian Drude, Max-Planck-Institute for Psycholinguistics, Nijmegen The Netherlands
5. CINET: Cyberinfrastructure for Network (Graph) Science and Analytics; Madhav Marathe or Keith Bisset, Virginia Tech
6. NIST Information Access Division analytic technology performance measurement, evaluations, and standards; John Garofolo, NIST

**The Ecosystem for Research**

1. DataNet Federation Consortium DFC; Reagan Moore, University of North Carolina at Chapel Hill
2. The ‘Discinnet process’, metadata <-> big data global experiment; P. Journeau, Discinnet Labs
3. Semantic Graph-search on Scientific Chemical and Text-based Data; Talapady Bhat, NIST
4. Light source beamlines; Eli Dart, LBNL

**Astronomy and Physics**

1. Catalina Real-Time Transient Survey (CRTS): a digital, panoramic, synoptic sky survey; S. G. Djorgovski, Caltech
2. DOE Extreme Data from Cosmological Sky Survey and Simulations; Salman Habib, Argonne National Laboratory; Andrew Connolly, University of Washington
3. Large Survey Data for Cosmology; Peter Nugent LBNL
4. Particle Physics: Analysis of LHC Large Hadron Collider Data: Discovery of Higgs particle; Michael Ernst BNL, Lothar Bauerdick FNAL, Geoffrey Fox, Indiana University; Eli Dart, LBNL
5. Belle II High Energy Physics Experiment; David Asner & Malachi Schram, PNNL

**Earth, Environmental and Polar Science**

1. EISCAT 3D incoherent scatter radar system; Yin Chen, Cardiff University; Ingemar Häggström, Ingrid Mann, Craig Heinselman, EISCAT Science Association
2. ENVRI, Common Operations of Environmental Research Infrastructure; Yin Chen, Cardiff University
3. Radar Data Analysis for CReSIS Remote Sensing of Ice Sheets; Geoffrey Fox, Indiana University
4. UAVSAR Data Processing, Data Product Delivery, and Data Services; Andrea Donnellan and Jay Parker, NASA JPL
5. NASA LARC/GSFC iRODS Federation Testbed; Brandi Quam, NASA Langley Research Center
6. MERRA Analytic Services MERRA/AS; John L. Schnase & Daniel Q. Duffy , NASA Goddard Space Flight Center
7. Atmospheric Turbulence - Event Discovery and Predictive Analytics; Michael Seablom, NASA HQ
8. Climate Studies using the Community Earth System Model at DOE’s NERSC center; Warren Washington, NCAR
9. DOE-BER Subsurface Biogeochemistry Scientific Focus Area; Deb Agarwal, LBNL
10. DOE-BER AmeriFlux and FLUXNET Networks; Deb Agarwal, LBNL

**Energy**

1. Consumption forecasting in Smart Grids; Yogesh Simmhan, University of Southern California **NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

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

**Note: No proprietary or confidential information should be included**

**ADD picture of operation or data architecture of application below table**

**Government Operation  
NBD(NIST Big Data) Requirements WG Use Case Template**

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | Big Data Archival: Census 2010 and 2000 – Title 13 Big Data | |
| **Vertical (area)** | | Digital Archives | |
| **Author/Company/Email** | | Vivek Navale & 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** | | 1. Maintain data “as-is”. No access and no data analytics for 75 years. 2. Preserve the data at the bit-level. 3. Perform curation, which includes format transformation if necessary. 4. 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 & Privacy**  **Requirements** | | Title 13 data. | |
| **Highlight issues for generalizing this use case (e.g. for ref. architecture)** | | . | |
| **More Information (URLs)** | |  | |

**Government Operation  
NBD(NIST Big Data) Requirements WG Use Case Template**

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | National Archives and Records Administration Accession NARA Accession, Search, Retrieve, Preservation | |
| **Vertical (area)** | | Digital Archives | |
| **Author/Company/Email** | | Quyen Nguyen & 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)** | | Hundred 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. |
| **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 pre-processing and manage for long-term of large and varied data.  Search huge amount of data.  Ensure high relevancy and recall.  Data sources may be distributed in different clouds in future. | |
| **Big Data Specific Challenges in Mobility** | | Mobile search must have similar interfaces/results | |
| **Security & 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)** | |  | |
| **Note:** <additional comments> | | | |

**Government Operation**

**NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **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 publically auditable. | |
| **Big Data Specific Challenges in Mobility** | | Mobile access is important. | |
| **Security & 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)** | | Recommender systems have features in common to e-commerce like Amazon, Netflix, UPS etc. | |
| **More Information (URLs)** | |  | |
| **Note:** <additional comments> | | | |

**Government Operation**

**NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **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. |
| **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 publically auditable. | |
| **Big Data Specific Challenges in Mobility** | | Mobile access is important. | |
| **Security & 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)** | |  | |
| **Note:** <additional comments> | | | |

**Commercial  
Draft, Ver. 0.1\_Aug. 24Th, 2013: NBD (NIST Big Data) Finance Industries (FI) Taxonomy/Requirements WG Use Case**

|  |  |  |  |
| --- | --- | --- | --- |
| **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 & Investments**, such as; Retail Brokerage, Private Banking/Wealth Management, Institutional Brokerages, Investment Banking, Trust Banking, Asset Management, Custody & Clearing Services  **Insurance**, including; Personal and Group Life, Personal and Group Property/Casualty, Fixed & 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 GRC/CIA (Governance, Risk & Compliance/Confidentiality, Integrity & Availability) 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 & Exchange Commission), FDIC (Federal Deposit Insurance Corporation), CFTC (Commodity Futures Trading Commission), US Treasury, PCAOB (Public Corporation Accounting & Oversight Board), COSO, CobiT, reporting supply chains & stakeholders, investment community, share holders, 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 (Governance, Risk & Compliance) and CIA (Confidentiality, Integrity & Availability) 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 & Controls 3. Business Impact Analysis 4. Design, Development & Testing of the Business Continuity Strategies 5. Emergency Response & Operations (aka; Disaster Recovery) 6. Developing & Implementing Business Continuity Plans 7. Awareness & Training Programs 8. Maintaining & 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, BigData has evolved into a robust data analysis and storage tool that is still under going development. However, in the end, BigData is still being developed as an adjunct to the current IT client/server/big iron data warehouse architectures which is better at somethings, 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 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/RDBM (Relational Database Management ) systems are use for the handling, processing, 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 (Security and Exchange Commission), CFTC (Commodity Futures Trading Commission), FDIC (Federal Deposit Insurance Corporation), DOJ (Dept. 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 (Security and Exchange Commission), CFTC (Commodity Futures Trading Commission), FDIC (Federal Deposit Insurance Corporation), US Treasury Dept., DOJ (Dept. of Justice), 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, MapReduce, 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 (Security and Exchange Commission), CFTC (Commodity Futures Trading Commission), US Treasury, FDIC (Federal Deposit Insurance Corporation), DOJ (Dept. of Justice), and my favorite the PCAOB (Public Company Accounting and Oversight Board). |
| **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 NAS (Network as A Service), DC2--->, VMWare/KVM (Clusters w/Virtual Firewalls), DataLink (Vmware Link, Vmotion Link, Network Link), Multiple PB of NAS (Network as A Service), (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 (Security and Exchange Commission), CFTC (Commodity Futures Trading Commission), US Treasury, FDIC (Federal Deposit Insurance Corporation), DOJ (Dept. of Justice), 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 data sets, 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.  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 (Confidentiality, Integrity & Availability) and GRC (Governance, Risk & Compliance) 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 (Confidentiality, Integrity & Availability) and GRC (Governance, Risk & Compliance) 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 (Confidentiality, Integrity & Availability) and GRC (Governance, Risk & Compliance) requirements. |
| **Data Types** | | **Please Note:** Based upon legislative and regulatory concerns, **data types** is 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 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) |
| **Data Analytics** | | **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. |
| **Big Data Specific Challenges (Gaps)** | | 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 R&D (Research & Development) SDLC/HDLC (Software Development Life Cycle/Hardware Development Life Cycle) sausage makers of technology. Please stay tuned for future developments in this regard | |
| **Big Data Specific Challenges in Mobility** | | Mobility is a continuously growing layer of technical complexity, however, not all Big Data mobility solutions are technical in nature. There are to 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. * **Durable**- The results of a committed transaction survive failures.   When each of these data categories are 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 & 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's 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 organization's:   * ISACA (International Society of Auditors & Computer Analysts) * isc2 (International Security Computer & 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 are 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.  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 (Generally Accepted Accounting Practices) or IFRS (…..):  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). | |
| **re 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, www.xbrl-eu.org 3. IEEE Big Data conference http://www.ischool.drexel.edu/bigdata/bigdata2013/topics.htm 4. MapReduce [http://www.mapreduce.org](http://www.mapreduce.org/). 5. PCAOB [http://www.pcaob.org](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](http://www.cftc.org/) 9. SEC [http://www.sec.gov](http://www.sec.gov/) 10. FDIC [http://www.fdic.gov](http://www.fdic.gov/) 11. COSO [http://www.coso.org](http://www.coso.org/) 12. isc2 International Information Systems Security Certification Consortium, Inc.: [http://www.isc2.org](http://www.isc2.org/) 13. ISACA Information Systems Audit and Control Association: [http://www.isca.org](http://www.isca.org/) 14. IFARS [http://www.ifars.org](http://www.ifars.org/) 15. Apache [http://www.opengroup.org](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, www.coso.org. 20. ITIL Information Technology Infrastructure Library, Copyright© 2007-13 APM Group Ltd. All rights reserved, Registered in England No. 2861902, www.itil-officialsite.com. 21. CobiT, Ver. 5.0, 2013, ISACA, Information Systems Audit and Control Association, (a framework for IT Governance and Controls), www.isaca.org. 22. TOGAF, Ver. 9.1, The Open Group Architecture Framework (a framework for IT architecture), www.opengroup.org. 23. ISO/IEC 27000:2012 Info. Security Mgt., International Organization for Standardization and the International Electrotechnical Commission, www.standards.iso.org/ | |
| **Note:** <additional comments> 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  
NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **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 |
| **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 & 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> | |
| **Note:** <additional comments> | | | |

**Commercial  
NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **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. |
| **Big Data Specific Challenges (Gaps)** | | Analytics needs continued monitoring and improvement. | |
| **Big Data Specific Challenges in Mobility** | | Mobile access important | |
| **Security & 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/> | |
| **Note:** <additional comments> | | | |

**Commercial  
NBD(NIST Big Data) Requirements WG Use Case Template**

|  |  |  |  |
| --- | --- | --- | --- |
| **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** | | MapReduce + 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 & Privacy**  **Requirements** | | Need to be sensitive to crawling restrictions. Avoid Spam results | |
| **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/ | |
| **Note:** <additional comments> | | | |

**Commercial  
NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | IaaS (Infrastructure as a Service) Big Data Business Continuity & Disaster Recovery (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 business continuity and disaster recovery (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 (Business Continuity/Disaster Recovery), 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 & Controls 3. Business Impact Analysis 4. Design, Development & Testing of the Business Continuity Strategies 5. Emergency Response & Operations (aka; Disaster Recovery 6. Developing & Implementing Business Continuity Plans 7. Awareness & Training Programs 8. Maintaining & Exercising Business Continuity Plans, (aka: Maintaining Currency) 9. Public Relations (PR) & 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, BigData has evolved into a robust data analysis and storage tool that is still under going development. However, in the end, BigData is still being developed as an adjunct to the current IT client/server/big iron data warehouse architectures which is better at somethings, 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 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 (Response Time Objectives and Recovery Point Objectives. 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 (Recovery Time Objectives) |
| **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, MapReduce, 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 (Hosted Virtual Servers), such as the following: DC1---> VMWare/KVM (Clusters, w/Virtual Firewalls), Data link-Vmware Link-Vmotion Link-Network Link, Multiple PB of NAS (Network as A Service), DC2--->, VMWare/KVM (Clusters w/Virtual Firewalls), DataLink (Vmware Link, Vmotion Link, Network Link), Multiple PB of NAS (Network as A Service), (Requires Fail-Over Virtualization) |
| **Volume (size)** | | Terabytes up to Petabytes |
| **Velocity**  **(e.g. real time)** | | Tier 3 Data Centers with Secure Fault Tolerant (Power) for Security, Power, 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. This replications can be performed within a 5 second window, and this Snap Shot will be kept for 7 business days, or longer if necessary to a Fail-Over Center.....at the RPO and RTO.... |
| **Variety**  **(multiple data sets, 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 (Confidentiality, Integrity & Availability) and GRC (Governance, Risk & Compliance) 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 (Confidentiality, Integrity & Availability) and GRC (Governance, Risk & Compliance) 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 (Confidentiality, Integrity & Availability) and GRC (Governance, Risk & Compliance) 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 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 & 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 who 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 & 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. www.disasterrecovery.org/, (March, 2013). 2. BC\_DR From the Cloud, Avoid IT Disasters EN POINTE Technologies and dinCloud, Webinar Presenter Barry Weber, www.dincloud.com. 3. COSO, The Committee of Sponsoring Organizations of the Treadway Commission (COSO), Copyright© 2013, www.coso.org. 4. ITIL Information Technology Infrastructure Library, Copyright© 2007-13 APM Group Ltd. All rights reserved, Registered in England No. 2861902, www.itil-officialsite.com. 5. CobiT, Ver. 5.0, 2013, ISACA, Information Systems Audit and Control Association, (a framework for IT Governance and Controls), www.isaca.org. 6. TOGAF, Ver. 9.1, The Open Group Architecture Framework (a framework for IT architecture), www.opengroup.org. 7. ISO/IEC 27000:2012 Info. Security Mgt., International Organization for Standardization and the International Electrotechnical Commission, www.standards.iso.org/. 8. PCAOB, Public Company Accounting and Oversight Board, 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  
NBD (NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **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. | |
| **Big Data Specific Challenges in Mobility** | | Currently conditions are not monitored on-board trucks, ships, and aircraft | |
| **Security & 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)** | |  | |
| **Note:** <additional comments> | | | |

****

**Commercial**

**NBD(NIST Big Data) Requirements WG Use Case Template Aug 22 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **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 R&D 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 is 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** |
| **Variety**  **(multiple datasets, mashup)** | | **Many data sets 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 data sets.** |
| **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 R&D 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 & 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)** | |  | |
| **Note:** <additional comments> | | | |

**Commercial**

**NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | Simulation driven Materials Genomics | |
| **Vertical (area)** | | Scientific Research: Materials Science | |
| **Author/Company/Email** | | David [Skinner/LBNL/deskinner@lbl.gov](mailto: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 fileformats |
| **Data Analytics** | | MapReduce 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. | |
| **Security & 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 | |
| **Note:** <additional comments> | | | |

**Defense**

**NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **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](mailto:%20dboyd@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 a 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 Gigabytes 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 data sets (millions of points) on small devices (handhelds) at the end of low bandwidth networks. |
| **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 & 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:** The 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 still remain with visualization. | | | |

**Defense**

**NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **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 RDBM’s 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 – 30-60 frames per/sec at full color 1080P resolution.  WALF – 1-10 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). |
| **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    1. is the line at the bread store, the butcher, or the ice cream store,    2. are more trucks traveling north with goods than trucks going south    3. 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 & 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 ident/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/> | |
| **Note:** <additional comments> | | | |

**Defense**

**NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **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    1. Relationships between entities (people, organizations, places, equipment)    2. Trends in sentiment or intent for either general population or leadership group (state, non-state actors).    3. Location of and possibly timing of hostile actions (including implantation of IEDs).    4. 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:    1. Search    2. Reasoning    3. 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 Gigabytes (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. |
| **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. | |
| **Security & 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/> | |
| **Note:** <additional comments> | | | |

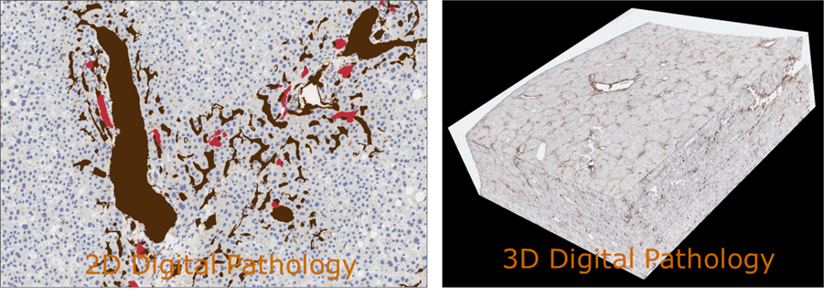
**Healthcare and Life Sciences**

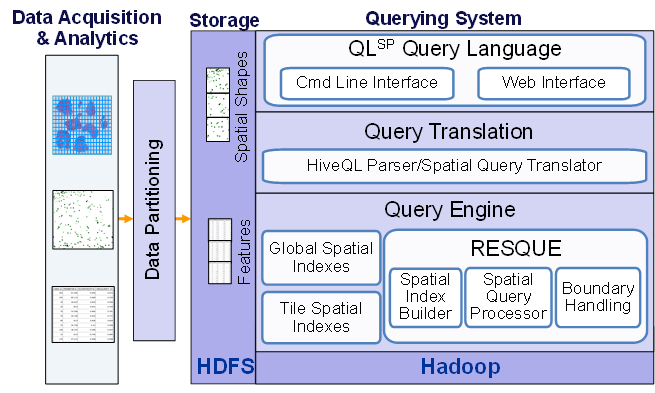
**NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **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 electronic medical record 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 electronic medical record 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 drives 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. |
| **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 electronic medical record 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). |
| **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 & 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](http://www.regenstrief.org)); Logical observation identifiers names and codes (http://[www.loinc.org](http://www.loinc.org)); Indiana Health Information Exchange (http://[www.ihie.org](http://www.ihie.org)); Institute of Medicine Learning Healthcare System (http://www.iom.edu/Activities/Quality/LearningHealthcare.aspx) | |
| **Note:** <additional comments> | | | |

**Healthcare and Life Sciences  
NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **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; MapReduce + 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 pre-processing 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 |
| **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 & 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> | |
| **Note:** <additional comments> | | | |

**Figure 1: Examples of 2-D and 3-D pathology images.**

****

**Figure 2: Architecture of Hadoop-GIS, a spatial data warehousing system over MapReduce to support spatial analytics for analytical pathology imaging.**

**Healthcare and Life Sciences  
NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | Computational Bioimaging | |
| **Vertical (area)** | | Scientific Research: Biological Science | |
| **Author/Company/Email** | | David Skinner**1**, [deskinner@lbl.gov](mailto:deskinner@lbl.gov)  Joaquin Correa**1**, [JoaquinCorrea@lbl.gov](mailto:JoaquinCorrea@lbl.gov)  Daniela Ushizima**2**, [dushizima@lbl.gov](mailto:dushizima@lbl.gov)  Joerg Meyer**2**, [joergmeyer@lbl.gov](mailto: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 data sets. 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. |
| **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 & 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)** | |  | |
| **Note:** <additional comments> | | | |

**Healthcare and Life Sciences  
NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **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 | |
| **Security & 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: www.genomeinabottle.org | |
| **Note:** <additional comments> | | | |

**Healthcare and Life Sciences   
NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | Comparative analysis for metagenomes and genomes | |
| **Vertical (area)** | | Scientific Research: Genomics | |
| **Author/Company/Email** | | Ernest Szeto / LBNL / [eszeto@lbl.gov](mailto: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). |
| **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 is better defined, but still an open research area. |
| **Visualization** | | Interactive speed of web UI on very large data sets 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 relational database management system (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 & 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> | |
| **Note:** <additional comments> | | | |

**Healthcare and Life Sciences**

**NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **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 |
| **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 & 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)** | |  | |
| **Note:** <additional comments> | | | |

**Healthcare and Life Science  
NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **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 – 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 internet. |
| **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 data sets. 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)** | |  |
| **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 & 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)** | |  | |
| **Note:** <additional comments> | | | |

**Healthcare and Life Sciences  
NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **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](mailto:mmarathe@vbi.vt.edu), [seubank@vbi.vt.edu](mailto: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 | |
| **Security & 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)** | |  | |
| **Note:** <additional comments> | | | |

**Healthcare and Life Sciences  
NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **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](mailto: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,**  **analysis,**  **action)** | **Veracity (Robustness Issues, semantics)** | | Failover of soft realtime analyses. |
| **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 & 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)** | |  | |
| **Note:** <additional comments> | | | |

**Healthcare and Life Sciences**

**NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **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  Datacenter: 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 data sets/sources  Collected amount of data TBD |
| **Velocity**  **(e.g. real time)** | | Data analysed 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 |
| **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 are 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 visualisation facilities, visualisation 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.  Visualisation, distributed sensor networks.  Data storage and archiving, data exchange and integration; data linkage: from the initial observation data to processed data and reported/visualised 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 & 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 lifecycle 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/> | |
| **Note:** <additional comments>  **Variety of data used in Biodiversity research**  Genetic (genomic) diversity   * DNA sequences & 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 * CO2 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 & 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 & satellite observation spectra * Field \* Laboratory experimentation * Radar & LiDAR * Fisheries & agricultural data * Deceases and epidemics | | | |

**Deep Learning and Social Media   
NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | Large-scale Deep Learning | |
| **Vertical (area)** | | Machine Learning/AI | |
| **Author/Company/Email** | | Adam Coates / Stanford University / [acoates@cs.stanford.edu](mailto: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 pre-processing. In addition to the computational expense demanded by the learning algorithms, the need for rapid prototyping and ease of development is extremely high. | |
| **Current**  **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 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. |
| **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 pre-processing; 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 & Privacy**  **Requirements** | | None. | |
| **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/> | |
| **Note:** <additional comments> | | | |

**Deep Learning and Social Media**

**NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **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), |
| **Variability (rate of change)** | | Rate of photos varies significantly, e.g. roughly 10x photos to Facebook on New Years 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 & 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 | |
| **Note:** <additional comments> | | | |

## Deep Learning and Social Media NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013

|  |  |  |  |
| --- | --- | --- | --- |
| **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](mailto:fil@indiana.edu);  Alessandro Flammini, Indiana University, [aflammin@indiana.edu](mailto:aflammin@indiana.edu);  Emilio Ferrara, Indiana University, [ferrarae@indiana.edu](mailto:ferrara@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 & HDFS distributed storage. Redis as a 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 & 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. |
| **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 & 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> | |
| **Note:** <additional comments> | | | |

**Deep Learning and Social Media  
NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **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 (MPI) | |
| **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 gigabytes (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 data sets; 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 lifecycle.  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) |
| **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  Digitising 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 & 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  
NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

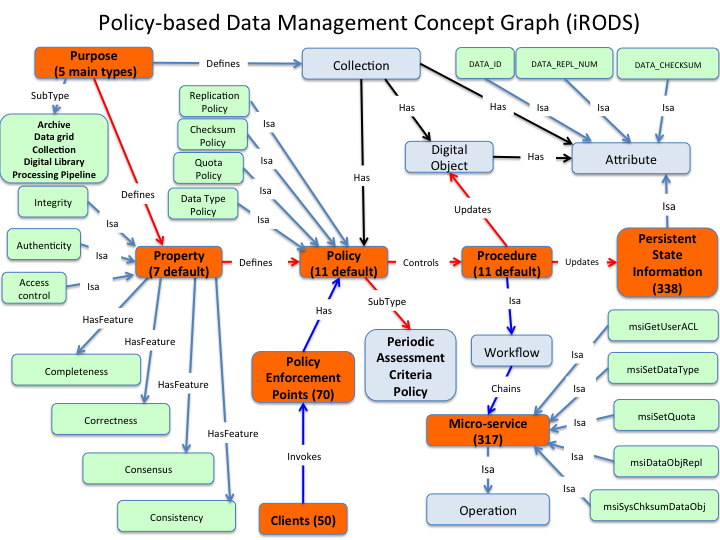
|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | CINET: Cyberinfrastructure for Network (Graph) Science and Analytics | |
| **Vertical (area)** | | Network Science | |
| **Author/Company/Email** | | Team lead by Virginia Tech and comprising of researchers from Indiana University, University 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 |
| **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)** | | Data sets 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 & Privacy**  **Requirements** | |  | |
| **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/ | |
| **Note:** <additional comments> | | | |

**Deep Learning and Social Media   
NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **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 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 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 & 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)** | | www.nist.gov/itl/iad/ | |
| **Note:** <additional comments> | | | |

**The Ecosystem for Research  
NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **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 |
| **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 & 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:** <additional comments>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. | | | |



**The Ecosystem for Research**

**NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **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](mailto: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 R&D epistemology, Discinnet Labs applied in web 2.0 [www.discinnet.org](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 R&D 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 1st 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,**  **action)** | **Veracity (Robustness Issues, semantics)** | | Methods to detect overall consistency, holes, errors, misstatements, known but mostly to be implemented |
| **Visualization** | | Multidimensional (hypercube) |
| **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 & 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 [www.discinnet.org](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 [www.discinnet.org](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:** <additional comments>: 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. | | | |

**The Ecosystem for Research  
NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **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**   + 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**   + 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 |
| **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 & 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:** <additional comments> | | | |

**The Ecosystem for Research   
NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **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 scientists analysis. | |
| **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 Gigabyte 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.** |
| **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 & 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 gigabytes 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> | |
| **Note:** <additional comments> | | | |

**Astronomy and Physics**

**NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **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 data set 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. |
| **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. | |
| **Security & 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 (~ 20-30 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. | | | |

**Figure:** 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 unidied 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   
NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **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 & 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 and Analytics** | | Interpretation of results from detailed simulations requires advanced analysis and visualization techniques 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. |
| **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 & 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> | |
| **Note:** <additional comments> | | | |

**Astronomy and Physics**

**NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | Large Survey Data for Cosmology | |
| **Vertical (area)** | | Scientific Research: Cosmic Frontier | |
| **Author/Company/Email** | | Peter Nugent / LBNL / [penugent@lbl.gov](mailto:eszeto@lbl.gov) | |
| **Actors/Stakeholders and their roles and responsibilities** | | Dark Energy Survey, Dark Energy Spectroscopic Instrument, Large Synoptic Survey Telescope. ANL, BNL, FNAL, LBL & 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 60PB of imaging data and 15PB of catalog data and a correspondingly large (or larger) amount of simulation data. Over 20TB 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. |
| **Visualization** | | Interactive speed of web UI on very large data sets 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 decompostion for thousands of simulations with matricies of order 1M on a side. | |
| **Big Data Specific Challenges in Mobility** | | Performing analysis on both the simulation and observational data simultaneously. | |
| **Security & 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, & http://www.darkenergysurvey.org | |
| **Note:** <additional comments> | | | |

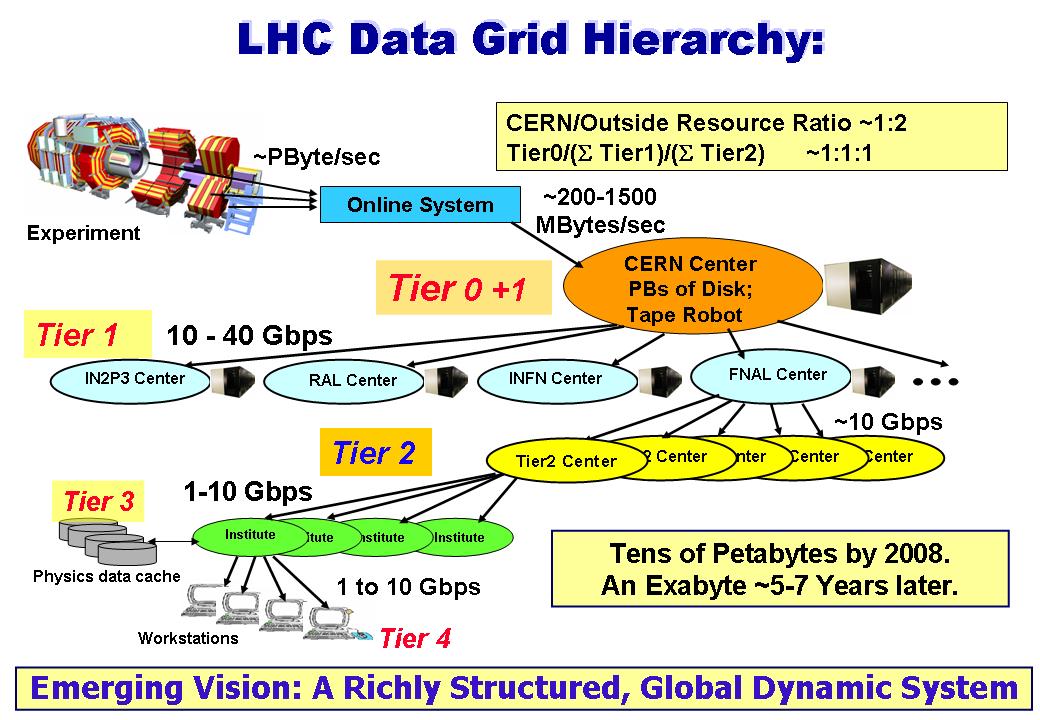
**Astronomy and Physics   
NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **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](mailto:mernst@bnl.gov), Lothar Bauerdick <bauerdick@fnal.gov> based on an initial version written by Geoffrey Fox, Indiana University [gcf@indiana.edu](mailto:gcf@indiana.edu), Eli Dart, LBNL [eddart@lbl.gov](mailto: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 * 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**  **(multiple datasets, mashup)** | | Lots of types of events with from 2- few hundred final particle but all data is collection of particles after initial 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 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—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 & 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 lifecycle. 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:** <additional comments> | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Use Case Stages** | **Data Sources** | **Data Usage** | **Transformations  (Data Analytics)** | **Infrastructure** | **Security & 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 109 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 data sets the classic program is Root from CERN that reads multiple event (AOD, NTUP) files from selected data sets 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 |

****

**Figure 1: The LHC Collider location at CERN**



**Figure 2: The Multi-tier LHC computing infrastructure**

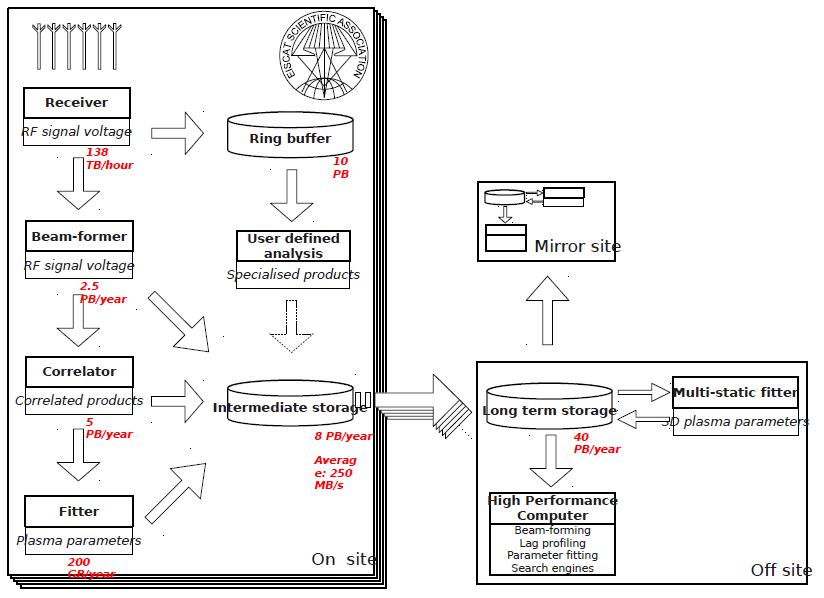
**Astronomy and Physics**

**NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | Belle II Experiment | |
| **Vertical (area)** | | Scientific Research: High Energy Physics | |
| **Author/Company/Email** | | David Asner & Malachi Schram, PNNL, [david.asner@pnnl.gov](mailto:david.asner@pnnl.gov) & [malachi.schram@pnnl.gov](mailto: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 & Privacy**  **Requirements** | | No special challenges. Data is accessed using grid authentication. | |
| **Highlight issues for generalizing this use case (e.g. for ref. architecture)** | |  | |
| **More Information (URLs)** | | http://belle2.kek.jp | |
| **Note:** <additional comments> | | | |

**Earth, Environmental and Polar Science  
NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

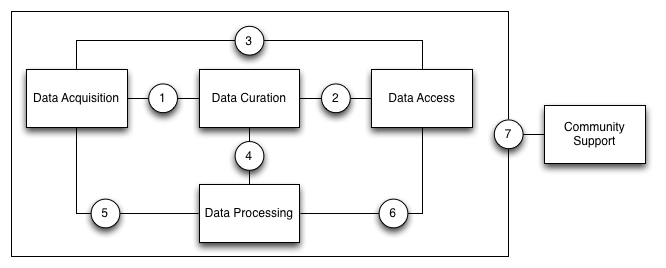
|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | EISCAT 3D incoherent scatter radar system | |
| **Vertical (area)** | | Environmental Science | |
| **Author/Company/Email** | | Yin Chen /Cardiff University/ [chenY58@cardiff.ac.uk](mailto:chenY58@cardiff.ac.uk)  Ingemar Häggström, Ingrid Mann, Craig Heinselman/  EISCAT Science Association/ {[Ingemar.Haggstrom, Ingrid.mann, Craig.Heinselman}@eiscat.se](mailto:Ingemar.Haggstrom,%20Ingrid.mann,%20Craig.Heinselman%7d@eiscat.se) | |
| **Actors/Stakeholders and their roles and responsibilities** | | The EISCAT Scientific Association is an international research organisation 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. | |
| **Goals** | | 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. | |
| **Use Case Description** | | 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. | |
| **Current**  **Solutions** | **Compute(System)** | | 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 |
| **Storage** | | 32TB |
| **Networking** | | 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. |
| **Software** | | * 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   + Control & monitoring tools including, system configuration, quick-look, fault reporting, etc.   + Data dissemination utilities   + User software e.g., for cyclic buffer, data cleaning, RFI detection and excision, auto-correlation, data integration, data analysis, event identification, discovery & retrieval, calculation of value-added data products, ingestion/extraction, plot   + User-oriented computing   + APIs into standard software environments   + 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 visualisation of analysed 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) visualisation of the physical parameters of interest, e.g.,   + by standard plots,   + using three-dimensional block to show to spatial variation (in the user selected cuts),   + using animations to show the temporal variation,   + allow the visualisation 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 data set. * **(Interactive) Visualisation**. E.g., to allow users to combine the information on several spectral features, e.g., by using colour coding, and to provide real-time visualisation facility to allow the users to link or plug in tailor-made data visualisation 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 Visualisation 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 & 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/> | |
| **Note:** <additional comments> | | | |



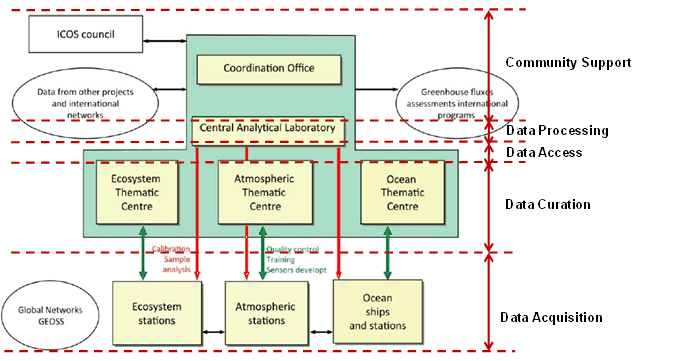
**Earth, Environmental and Polar Science**

**NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

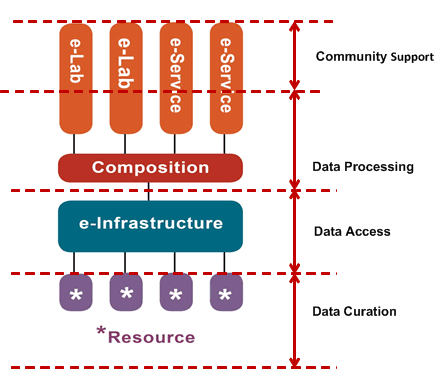
|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | ENVRI, Common Operations of Environmental Research Infrastructure | |
| **Vertical (area)** | | Environmental Science | |
| **Author/Company/Email** | | Yin Chen/ Cardiff University/ [ChenY58@cardiff.ac.uk](mailto:ChenY58@cardiff.ac.uk) | |
| **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 Reference Model (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 harmonised 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, visualisation 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 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](http://confluence.envri.eu:8090/download/attachments/327687/D3.3%20Analysis%20of%20Requirements%20V1.0.pdf?version=1&modificationDate=1366965933706&api=v2), the ENVRI Reference Model ([www.envri.eu/rm](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 (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 data set. * 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 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 & 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 modelers 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 & Privacy**  **Requirements** | | Most of the projects follow the open data sharing policy. E.g.,   * 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: [www.envri.eu](http://www.envri.eu) * ENVRI Reference Model [www.envri.eu/rm](http://www.envri.eu/rm) * [ENVRI deliverable D3.2](http://confluence.envri.eu:8090/download/attachments/327687/D3.3%20Analysis%20of%20Requirements%20V1.0.pdf?version=1&modificationDate=1366965933706&api=v2) : 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/> | |
| **Note:** <additional comments> | | | |



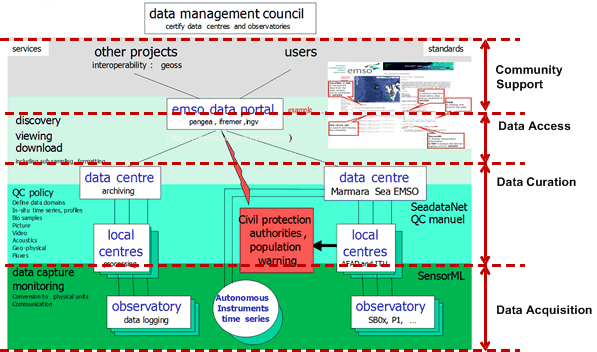
**Figure 1**: ENVRI Common Subsystems



1. ICOS Architecture



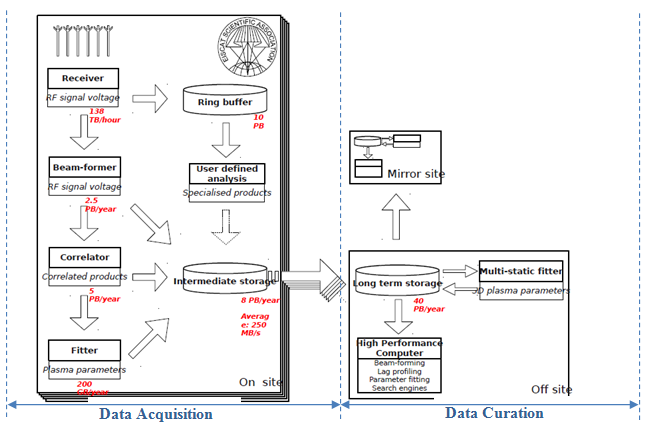
1. LifeWatch Architecture



1. EMSO Architecture



1. Eura-Argo Architecture



1. EISCAT 3D Architecture

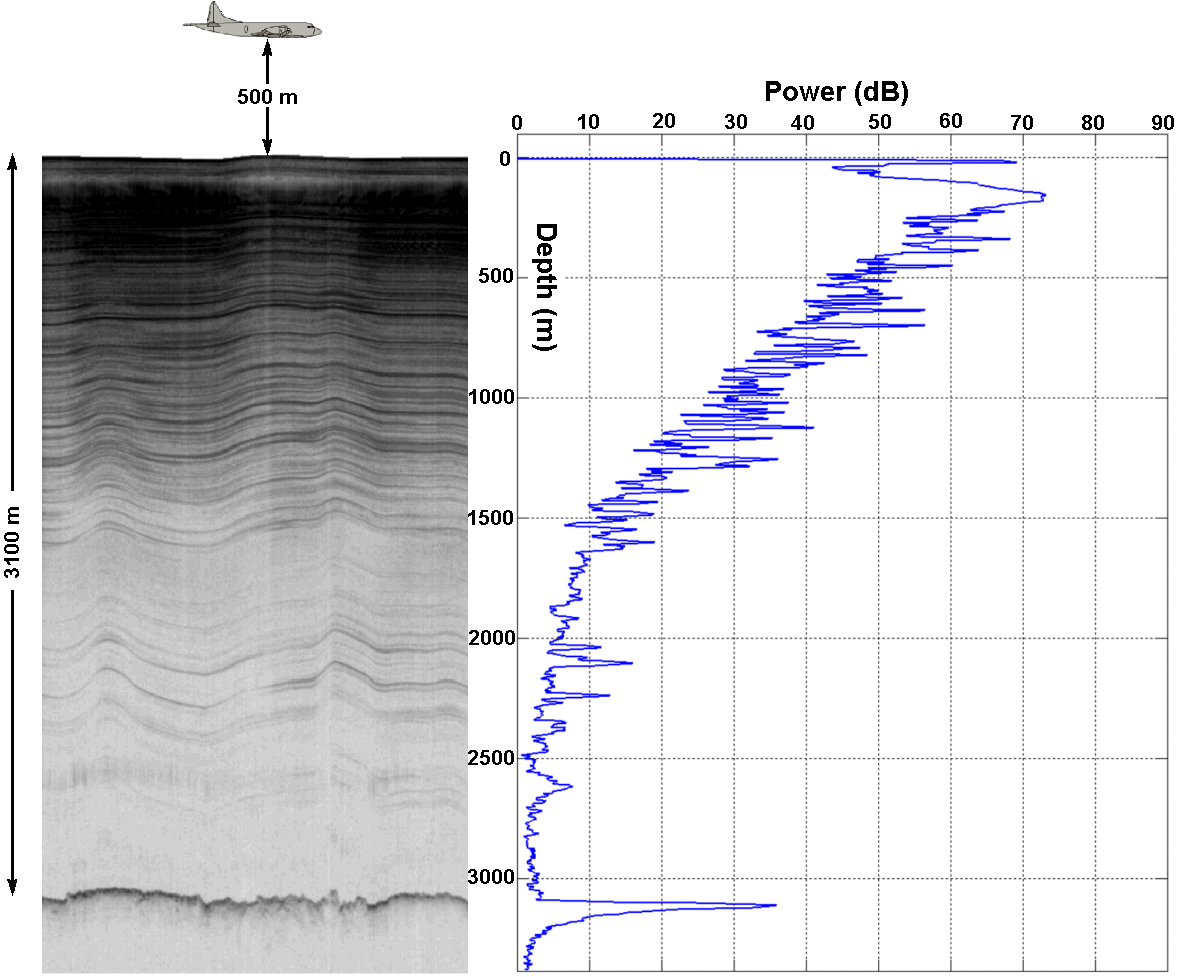
**Figure 2:** Architectures of the ESFRI Environmental Research Infrastructures

**Earth, Environmental and Polar Science**

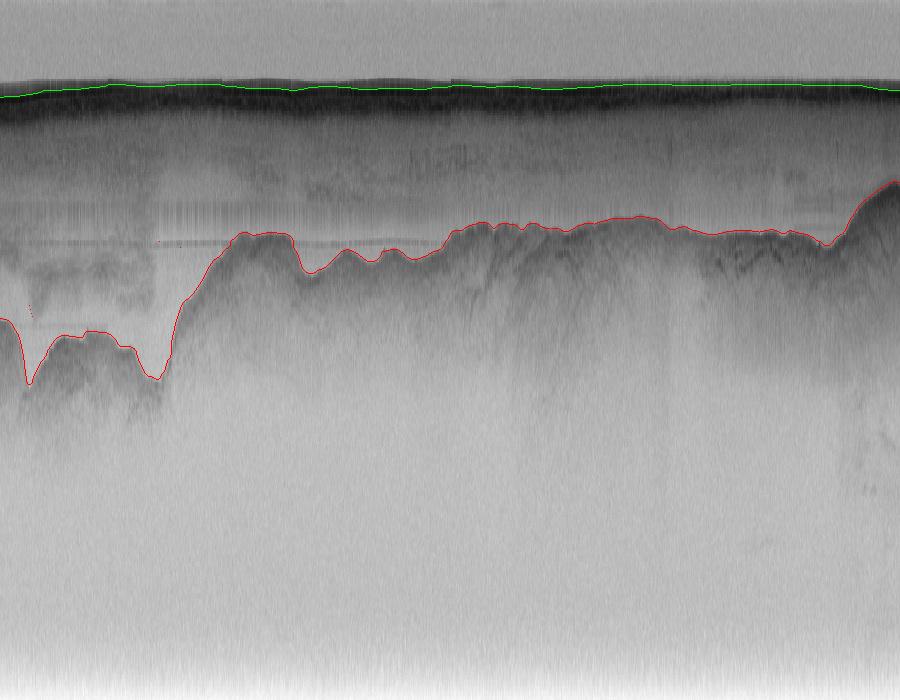
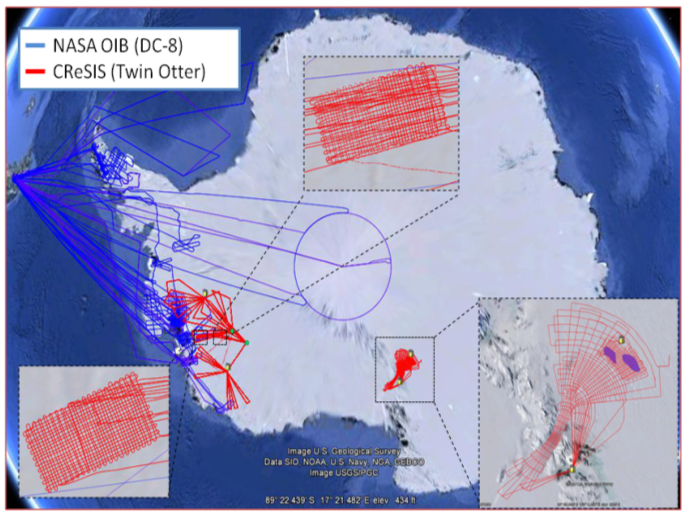
**NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **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 MapReduce 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 & Privacy**  **Requirements** | | Himalaya studies fraught with political issues and require UAV. Data itself open after initial study | |
| **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:** <additional comments> | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Use Case Stages** | **Data Sources** | **Data Usage** | **Transformations  (Data Analytics)** | **Infrastructure** | **Security & 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 & 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. |

****

**Figure 1: Typical Radar Data after analysis**

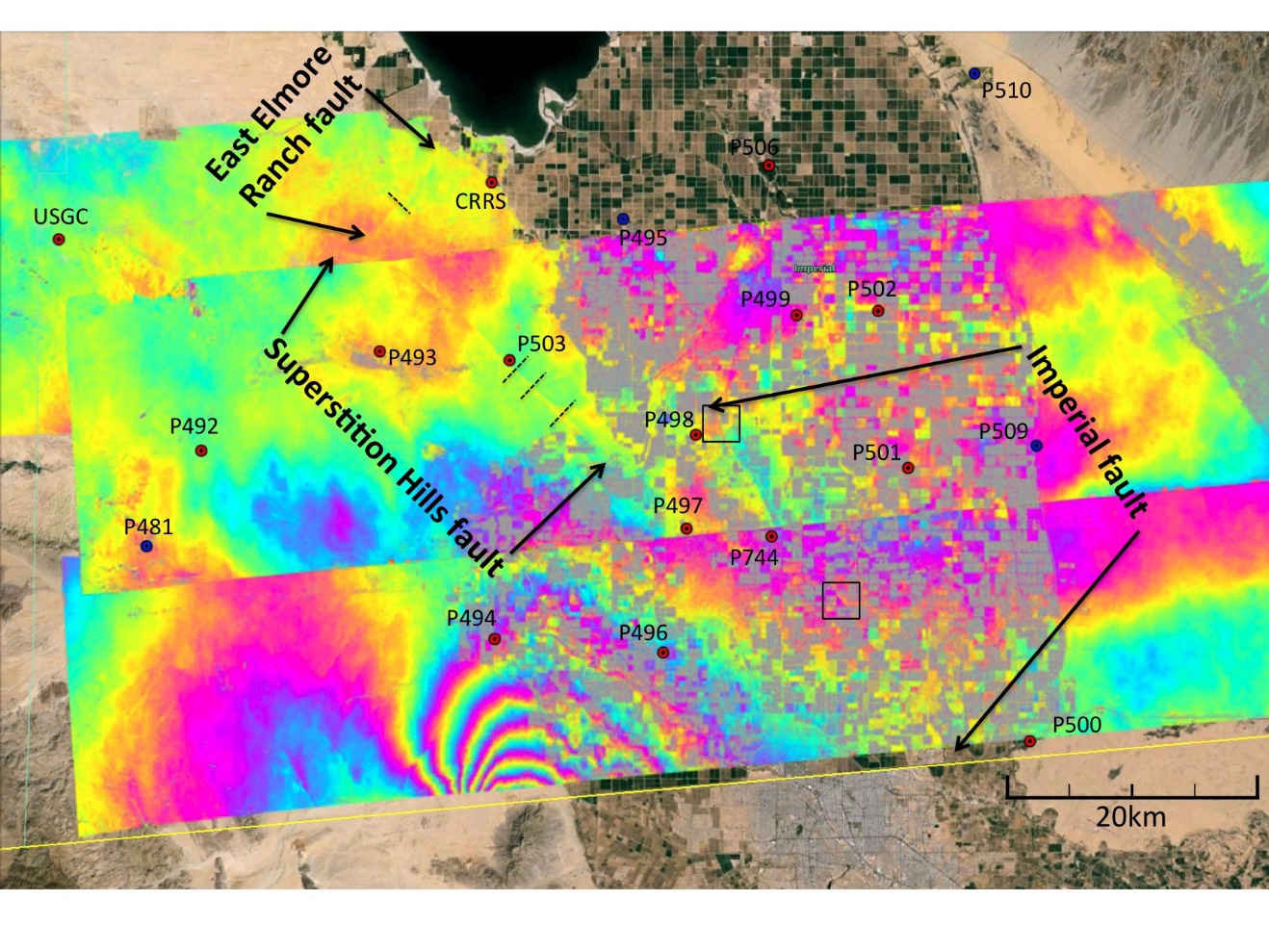
**Figure 2: Typical flight paths of data gathering in survey region**

**Figure 3. Typical echogram with Detected Boundaries. The upper (green) boundary is between air and ice layer while the lower (red) boundary is between ice and terrain**

**Earth, Environmental and Polar Science**

**NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **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 data set is unwieldy to transfer. This is a problem to downstream groups like QuakeSim who want to reformat and add value to data sets. |
| **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. |
| **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 unforseen 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 data sets to understandable images or statistics. | |
| **Security & 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 | |
| **Note:** <additional comments> | | | |



**Figure 1: UAVSAR Data Processing, Data Product Delivery, and Data Services (Section 2.9.4)** 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**

**NBD(NIST Big Data) Requirements WG Use Case Aug 15 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **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](mailto:M.M.Little@NASA.gov), [Roger.A.Dubois@nasa.gov](mailto:Roger.A.Dubois@nasa.gov), Brandi.M.Quam@NASA.gov, [Tiffany.J.Mathews@NASA.gov](mailto:Tiffany.J.Mathews@NASA.gov), & Andrei.A.Vakhnin@NASA.gov  John Schnase,,Daniel Duffy, Glenn Tamkin, Scott Sinno, John Thompson, & Mark McInerney / NASA Goddard Space Flight Center (GSFC) / John.L.Schnase@NASA.gov, [Daniel.Q.Duffy@NASA.gov](mailto:Daniel.Q.Duffy@NASA.gov), [Glenn.S.Tamkin@nasa.gov](mailto:Glenn.S.Tamkin@nasa.gov). [Scott.S.Sinno@nasa.gov](mailto:Scott.S.Sinno@nasa.gov), [John.H.Thompson@nasa.gov](mailto:John.H.Thompson@nasa.gov), & 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 data sets, 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**  **Solutions** | **Compute (System)** | | NASA Center for Climate Simulation (NCCS) and 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 Global 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-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 data sets. |
| **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 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 & 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 R&D 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. | |
| **Note:** <additional comments> | | | |

**Earth, Environmental and Polar Science**

**NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | MERRA Analytic Services (MERRA/AS) | |
| **Vertical (area)** | | Scientific Research: Earth Science | |
| **Author/Company/Email** | | John L. Schnase & 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 MapReduce analytics over the MERRA collection. MERRA/AS is an example of cloud-enabled Climate Analytics-as-a-Service, 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 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 data sets. |
| **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 (CAaaS). 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 & Privacy**  **Requirements** | | No critical issues identified at this time. | |
| **Highlight issues for generalizing this use case (e.g. for ref. architecture)** | | MapReduce and iRODS fundamentally make analytics and data aggregation easier; 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 Climate Analytics-as-a-Service 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. | |
| **Note:** <additional comments> | | | |

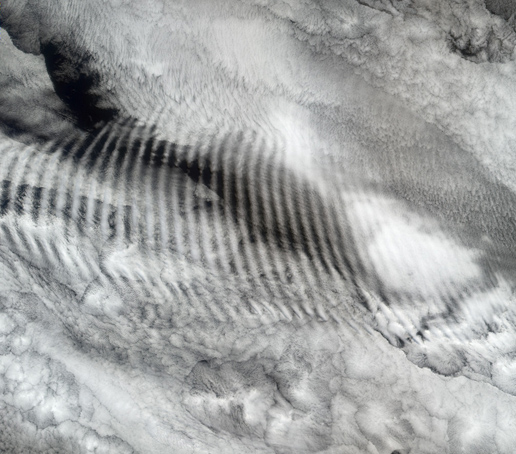
**C:\Users\Geoffrey Fox\Downloads\Panoply_LoRes.tiff**

**Figure: Typical MERRA/AS Output**

**Earth, Environmental and Polar Science**

**NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **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** | | MapReduce 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. | |
| **Big Data Specific Challenges in Mobility** | | Development for mobile platforms not essential at this time. | |
| **Security & 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/ | |
| **Note:** <additional comments> | | | |

**Figure: Typical NASA image of turbulent waves**

**Earth, Environmental and Polar Science**

**NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **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, parall el 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 GBytes/sec are produced by the simulations |
| **Variety**  **(multiple datasets, mashup)** | | Data must be compared among those from 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 intercomparison 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 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 & 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/> | |
| **Note:** <additional comments> | | | |

**Earth, Environmental and Polar Science**

**NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **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 soruces 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. |
| **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 & 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 | |
| **Note:** <additional comments> | | | |

**Earth, Environmental and Polar Science**

**NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **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 & Privacy**  **Requirements** | |  | |
| **Highlight issues for generalizing this use case (e.g. for ref. architecture)** | |  | |
| **More Information (URLs)** | | Ameriflux.lbl.gov  www.fluxdata.org | |
| **Note:** <additional comments> | | | |

**Energy  
NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **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 & Storage*: time-series data from (potentially) millions of smart meters in near-realtime, features on consumers, facilities and regions, weather forecasts, archival of data for training, testing and validating models; *2) Data Cleaning & 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 |
| **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 & 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 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> | |
| **Note:** <additional comments> | | | |