We expect other WG to comment on and probably edit the use case proposal that follows.

There are 5 existing use cases (The last 3 of these use cases need minor updates for changed template)

* Web Search
* Remote Sensing of Ice Sheets
* NIST/Genome in a Bottle Consortium
* Particle Physics
* Netflix

We got volunteers to collect use cases

* Yuri Demchenko ( Use case (UvA1): LifeWatch – European Infrastructure for Biodiversity and Ecosystem Research; Use case (UvA2): Humanities and language research infrastructure )
* William Miller (Cargo Shipping)
* Gary Mazzaferro sent template to OOI (Ocean Observatory Initiative)
* Fox will do Astronomy

We need others to contribute

**Current Draft:**

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | Atmospheric Turbulence - Event Discovery and Predictive Analytics | |
| **Vertical (area)** | | 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> | | | |

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

**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 “precisuion@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. 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 lay out 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> | | | |

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | Radar Data Analysis for CReSIS | |
| **Vertical (area)** | | 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> | | | |

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | Genomic Measurements | |
| **Vertical (area)** | | Healthcare | |
| **Author/Company** | | Justin Zook/NIST | |
| **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 |
| **Analytics(Software)** | | Open-source sequencing bioinformatics software from academic groups (UNIX-based) |
| **Big Data  Characteristics** | **Volume (size)** | | 40TB NFS is full, will need >100TB in 1-2 years at NIST; Healthcare community will need many PBs of storage |
| **Velocity** | | DNA sequencers can generate ~300GB compressed data/day. Velocity has increased much faster than Moore’s Law |
| **Variety** | | File formats not well-standardized, though some standards exist. Generally structured data. |
| **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 |
| **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. | |
| **Security & Privacy**  **Requirements** | | Sequencing data in health records or clinical research databases must be kept secure/private. | |
| **More Information (URLs)** | | Genome in a Bottle Consortium: www.genomeinabottle.org | |
| **Note:** <additional comments> | | | |

**Examples using previous draft**

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | Particle Physics: Analysis of LHC (Large Hadron Collider) Data (Discovery of Higgs particle) | |
| **Vertical** | | Fundamental Scientific Research | |
| **Author/Company/email** | | Geoffrey Fox, Indiana University gcf@indiana.edu | |
| **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 Accelerator and Monte Carlo producing events describing particle-apparatus interaction. Processed information defines physics properties of events (lists of particles with type and momenta) | |
| **Current**  **Solutions** | **Compute(System)** | | 200,000 cores running “continuously” arranged in 3 tiers (CERN, “Continents/Countries”. “Universities”). Uses “High Throughput Computing” (Pleasing parallel). |
| **Storage** | | Mainly Distributed cached files |
| **Analytics(Software)** | | 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 to estimate analysis quality |
| **Big Data  Characteristics** | **Volume (size)** | | 15 Petabytes per year from Accelerator and Analysis |
| **Velocity** | | Real time with some long "shut downs" with no data except Monte Carlo |
| **Variety** | | Lots of types of events with from 2- few hundred final particle but all data is collection of particles after initial analysis |
| **Veracity (Robustness Issues)** | | One can lose modest amount of data without much pain as errors proportional to 1/SquareRoot(Events gathered). 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 |
| **Data Quality** | | Huge effort to make certain complex apparatus well understood and "corrections" properly applied to data. Often requires data to be re-analysed |
| **Big Data Specific Challenges (Gaps)** | | Analysis system set up before clouds. Clouds have been shown to be effective for this type of problem. Object databases (Objectivity) were explored for this use case | |
| **Security & Privacy**  **Requirements** | | Not critical although the different experiments keep results confidential until verified and presented. | |
| **More Information (URLs)** | | http://grids.ucs.indiana.edu/ptliupages/publications/ Where%20does%20all%20the%20data%20come%20from%20v7.pdf | |
| **Highlight issues for generalizing this use case (e.g. for ref. architecture)** | | 1. Shall be able to analyze large amount of data in a parallel fashion  2. Shall be able to process huge amount of data in a parallel fashion  3. Shall be able to perform analytic and processing in multi-nodes (200,000 cores) computing cluster  4. Shall be able to convert legacy computing infrastructure into generic big data computing environment | |
| **Note:** <additional comments> | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | Netflix Movie Service | |
| **Vertical** | | 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 with Hadoop and Pig. |
| **Storage** | | Uses Cassandra NoSQL technology with Hive, Teradata |
| **Analytics(Software)** | | 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  Characteristics** | **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** | | Media and Rankings continually updated |
| **Variety** | | Data varies from digital media to user rankings, user profiles and media properties for content-based recommendations |
| **Veracity (Robustness Issues)** | | Success of business requires excellent quality of service |
| **Visualization** | | Streaming media |
| **Data Quality** | | Rankings are intrinsically “rough” data and need robust learning algorithms |
| **Big Data Specific Challenges (Gaps)** | | Analytics needs continued monitoring and improvement. | |
| **Security & Privacy**  **Requirements** | | Need to preserve privacy for users and digital rights for media. | |
| **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> | | | |