**Current Draft**

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **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 (Identify COTS, open source products** | |  | |  |
| **Big Data  Characteristics** | **Data Source (distributed/centralized)** | |  | |  |
| **Volume (size)** | |  | |  |
| **Velocity**  **(e.g. real time)** | |  | |  |
| **Variety**  **(multiple datasets, mashup, how various)** | |  | |  |
| **Variability (rate of change)** | |  | |  |
| **Big Data Science (collection, curation,**  **analysis,**  **action)** | **Veracity (Robustness Issues)** | |  | |  |
| **Visualization** | |  | |  |
| **Data Quality** | |  | |  |
| **Data Types** | |  | |  |
| **Data Analytics** | |  | |  |
| **Big Data Specific Challenges (Gaps)** | |  | | |  |
| **Big Data Specific Challenges in Mobility** | |  | | |  |
| **Security & Privacy**  **Requirements** | | **ITEM** | | **RESPONSE** | **CLARIFICATION** |
|  | | Investigator provenance | |  | Method used to associate researchers to digital artifacts |
| Sponsor disclosures | |  | Method used to associate researchers to digital artifacts; permissions for redisclosure |
| Investigator interests | |  | Investigators may be required to disclose potential conflicts of interest |
| Institution where performed | |  | Method used to associate responsible institution to digital artifacts; redisclosure rules; point in time considerations |
| Investigator affiliations | |  | Method for associating digital artifacts with investigator affiliations; one to many; point in time |
| Human Subject Data | |  | Yes/No |
| IRB traceability | |  | Institution-specific event(s), digital records, US-specific regulation |
| Data / analytics / meta-result publication rights | |  | Open publisher; traditional publisher; white paper; working paper; IP issues |
| Results repository | |  | Immutable store for data collected, results |
| Reference data | |  | Third party dependency for ref - Census or geospatial data could be basis for independent variables |
| Delegated rights | |  | Distributed delegation:: legal, governance, provenance covenants |
| Intellectual property | |  | Includes COTS, open source EXE, collection artifacts |
| Third party privacy notices | |  | Voluntary or mandated privacy act notices (US - FTC implications) |
| Reidentification risk | |  | Risk assessment by: Data Provider; covenants imposed on Data Consumers |
| Instrumentation and protocols | |  | “Procedure” in some academic paradigms, but considerable domain-specific elaboration may be needed. Sensor provenance, calibration, propagation, audit, aggregation |
| Primary meaning: Digital reproduceability.Secondary: simulation | |  | Full digital forward-construction, backward deconstruction of experiment, data collection, video, other digital artifacts for reproduceability |
|  | | Data Life-cycle | |  | Identify any legal mandates for data “destruction” |
|  | | Disclosure-on-demand | |  | Requirements for mandated or voluntary data disclosure; consumer or data owner;; may be regulation-, court-ordered, veracity motivated |
|  | | Recommended data security / privacy levels | |  | For template, see [HL7 Privacy Segmentation for Privacy](http://wiki.siframework.org/Data+Segmentation+for+Privacy+Paper) |
|  | | Dependency Analytics | |  | Measures in place to assure data integrity and regulatory compliance |
| **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

**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> | | | |