**NIST Big Data Public Working Group (NBD-PWG)**

**NBD-PWD-2014/M0325**

**Source: NBD-PWG**

**Status: Draft**

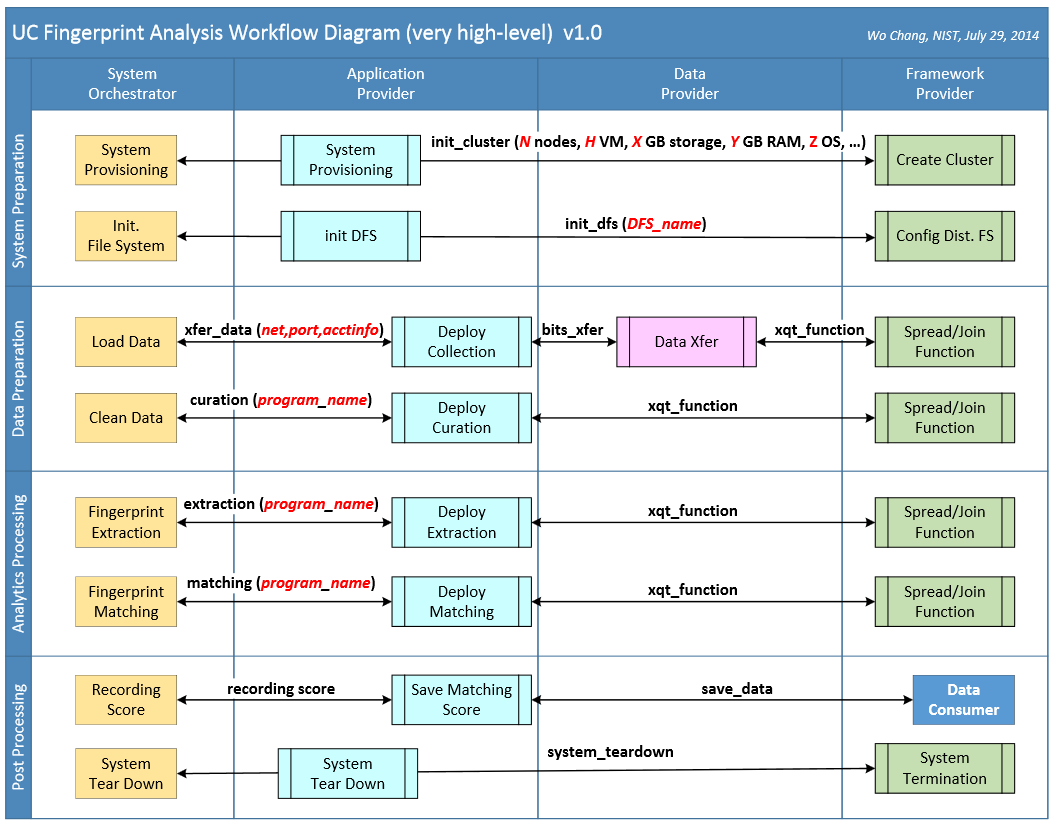
**Title: Proposed Use Cases for Identifying workflow/interaction between NBD-RA Key Components**

**Author: Wo Chang, NIST**

Potential unique scenarios from the 51 general use cases collected from NBD-PWG Use Cases & Requirements Subgroup (<http://bigdatawg.nist.gov/uc_reqs_summary.php>) are: based on the combinations of software used for what platforms (Hadoop, HPC, etc.), smaller size of datasets, and characteristic of scenarios (real-time, batch processing, distributed data sources, etc.).

We should pick use cases with less number of software tools first and then add complexity later.  The list below has not sorted by complexity but just possible good candidates.  We don’t need to use all but as a list of candidates to identify the workflow/interaction between NBD Ref. Arch. Key components.

The exercise is to take each use case and try map the workflow/interaction using the Figure 1 as an example but feel free to expand and extend what’s best for the specific use cases.



**Figure 1: Sample High-level Workflow/Interaction between NBD Reference Architecture**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **3** | [**M0219**](http://bigdatawg.nist.gov/_uploadfiles/M0219_v1_1106458060.docx) **Statistical Survey Response Improvement** | **Approximately 1PB** | **Variable. field data streamed continuously, Census was ~150 million records transmitted.** | **Strings and numerical data** | **Hadoop, Spark, Hive, R, SAS, Mahout, Allegrograph, MySQL, Oracle, Storm, BigMemory, Cassandra, Pig** | **recommendation systems, continued monitoring** | **1. data size approximately one petabyte** | **1. analytics are required for recommendation systems, continued monitoring and general survey improvement.** | **1. software includes Hadoop, Spark, Hive, R, SAS, Mahout, Allegrograph, MySQL, Oracle, Storm, BigMemory, Cassandra, Pig** | **1. data visualization for data review, operational activity and general analysis. It continues to evolve.** | **1. improving recommendation systems that reduce costs and improve quality while providing confidentiality safeguards that are reliable and publically auditable.  2. both confidential and secure all data. All processes must be auditable for security and confidentiality as required by various legal statutes.** | **1. high veracity on data and systems must be very robust. The semantic integrity of conceptual metadata concerning what exactly is measured and the resulting limits of inference remain a challenge** | **1. mobile access** |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **6** | [**M0161**](http://bigdatawg.nist.gov/_uploadfiles/M0161_v1_8712614971.docx) **Mendeley** | **15TB presently, growing about 1 TB/month** | **Currently Hadoop batch jobs are scheduled daily. Future is real-time recommendation** | **PDF documents and log files of social network and client activities** | **Hadoop, Scribe, Hive, Mahout, Python** | **Standard libraries for machine learning and analytics, LDA, custom built reporting tools for aggregating readership and social activities per document** | **1. file-based documents with constant new uploads  2. variety of file types such as PDF, social network log files, client activities images, spreadsheet, presentation files** | **1. standard machine learning and analytics libraries  2. scalable and parallelized efficient way for matching between documents  3. third-party annotation tools or publisher watermarks and cover pages** | **1. EC2 with HDFS (infrastructure)  2. S3 (storage)  3. Hadoop (platform)  4. Scribe, Hive, Mahout, Python (language)  5. moderate storage (15 TB with 1TB/month)  6. needs to batch and real-time processing** | **1. custom built reporting tools  2. visualization tools such as networking graph, scatterplots, etc.** | **1. access controls for who.s reading what content** | **1. metadata management from PDF extraction  2. identify of document duplication  3. persistent identifier  4. metadata correlation between data repositories such as CrossRef, PubMed and Arxiv** | **1. Windows Android and iOS mobile devices for content deliverables from Windows desktops** |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **15** | [**M0215**](http://bigdatawg.nist.gov/_uploadfiles/M0215_v1_1579991796.docx) **Intelligence Data Processing and Analysis** | **10s of Terabytes to 100s of Petabytes. Individual warfighters (first responders) would have at most 1-100s of Gigabytes** | **Much Real Time. Imagery intelligence device gathering petabyte in a few hours** | **Text files, raw media, imagery, video, audio, electronic data, human generated data** | **Hadoop, Accumulo (Big Table), Solr, Natural Language Processing, Puppet (for deployment and security) and Storm. GIS** | **Near Real Time Alerts based on patterns and baseline changes, Link Analysis, Geospatial Analysis, Text Analytics (sentiment, entity extraction, etc.)** | **1. much of Data real-time with processing at worst near real time  2. data currently exists in disparate silos which must be accessible through a semantically integrated data space  3. diverse data includes text files, raw media, imagery, video, audio, electronic data, human generated data.** | **1. analytics include NRT Alerts based on patterns and baseline changes** | **1. tolerance of Unreliable networks to warfighter and remote sensors  2. up to 100.s PB.s data supported by modest to large clusters and clouds  3. software includes Hadoop, Accumulo (Big Table), Solr, NLP (several variants), Puppet (for deployment and security), Storm, Custom applications and visualization tools** | **1. primary visualizations will be Geospatial overlays (GIS) and network diagrams.** | **1. data must be protected against unauthorized access or disclosure and tampering** | **1. data provenance (e.g. tracking of all transfers and transformations) must be tracked over the life of the data.** | **--** |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **17** | [**M0089**](http://bigdatawg.nist.gov/_uploadfiles/M0089_v1_7814086875.docx) **Pathology Imaging** | **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,** | **Once generated, data will not be changed** | **Images** | **MPI for image analysis; MapReduce + Hive with spatial extension** | **Image analysis, spatial queries and analytics, feature clustering and classification** | **1. high resolution spatial digitized pathology images  2. various image quality analysis algorithms  3. various image data formats especially BIGTIFF with structured data for analytical results  4. image analysis, spatial queries and analytics, feature clustering and classification** | **1. high performance image analysis to extract spatial information  2. spatial queries and analytics, and feature clustering and classification  3. analytic processing on huge multi-dimensional large dataset and be able to correlate with other data types such as clinical data, -omic data.** | **1. legacy system and cloud (computing cluster)  2. huge legacy and new storage such as SAN or HDFS (storage)  3. high throughput network link (networking)  4. MPI image analysis, MapReduce, Hive with spatial extension (sw pkgs)** | **1. visualization for validation and training** | **1. security and privacy protection for protected health information** | **1. human annotations for validation** | **1. 3D visualization and rendering on mobile platforms** |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **28** | [**M0160**](http://bigdatawg.nist.gov/_uploadfiles/M0160_v1_6667987957.docx) **Truthy** | **30TB/year compressed data** | **Near real-time data storage, querying & analysis** | **Schema provided by social media data source. Currently using Twitter only. We plan to expand incorporating Google+, Facebook** | **Hadoop IndexedHBase & HDFS. Hadoop, Hive, Redis for data management. Python: SciPy NumPy and MPI for data analysis.** | **Anomaly detection, stream clustering, signal classification and online-learning; Information diffusion, clustering, and dynamic network visualization** | **1. distributed data sources  2. large volume real time streaming  3. raw data in compressed formats  4. fully structured data in JSON, users metadata, geo-locations data  5. multiple data schemas** | **1. various real time data analysis for anomaly detection, stream clustering, signal classification on multi-dimensional time series and online-learning** | **1. Hadoop and HDFS (platform)  2. IndexedHBase, Hive, SciPy, NumPy (software)  3. in-memory database, MPI (platform)  4. high-speed Infiniband network (networking)** | **1. data retrieval and dynamic visualization  2. data driven interactive web interfaces  3. API for data query** | **1. security and privacy policy** | **1. standardized data structured/formats with extremely high data quality** | **1. low-level data storage infrastructure for efficient mobile access to data** |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **37** | [**M0185**](http://bigdatawg.nist.gov/_uploadfiles/M0185_v1_4843821869.docx) **DOE Extreme Data** | **Several petabytes from Dark Energy Survey and Zwicky Transient Factory. Simulations > 10PB** | **Analysis done in batch mode with data from observations and simulations updated daily** | **Image and simulation data** | **MPI, FFTW, viz packages, FFTW, numpy, Boost, OpenMP, ScaLAPCK, PSQL & MySQL databases, Eigen, cfitsio, astrometry.net, and Minuit2** | **New analytics needed to analyze simulation results** | **1. ~1 PB per year becoming 7PB a year observational datal** | **1. interpretation of results from detailed simulations requires advanced analysis and visualization techniques and capabilities** | **1. MPI, OpenMP, C, C++, F90, FFTW, viz packages, python, FFTW, numpy, Boost, OpenMP, ScaLAPCK, PSQL & MySQL databases, Eigen, cfitsio, astrometry.net, and Minuit2  2. supercomputer I/O subsystem limitations must be addressed** | **1. interpretation of results using advanced visualization techniques and capabilities** | **--** | **--** | **--** |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **38** | [**M0209**](http://bigdatawg.nist.gov/_uploadfiles/M0209_v1_4702199454.docx) **Large Survey Data for Cosmology** | **Dark Energy Survey will take PB's of data** | **400 images of one gigabyte in size per night** | **Images** | **Linux cluster, Oracle RDBMS server, Postgres PSQL, large memory machines, standard Linux interactive hosts, GPFS. For simulations, HPC resources. Standard astrophysics reduction software as well as Perl/Python wrapper scripts** | **Machine Learning to find optical transients. Cholesky decompostion for thousands of simulations with matrices of order 1M on a side and parallel image storage** | **1. 20TB data per day** | **1. analysis on both the simulation and observational data simultaneously  2. techniques for handling Cholesky decompostion for thousands of simulations with matricies of order 1M on a side** | **1. standard astrophysics reduction software as well as Perl/Python wrapper scripts  2. Oracle RDBMS, Postgres psql, as well as GPFS and Lustre file systems and tape archives  3. parallel image storage** | **--** | **--** | **1. links between remote telescopes and central analysis sites** | **--** |

dkdkdkd