**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> | | | |