**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> | | | |