**NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | Large Survey Data for Cosmology | |
| **Vertical (area)** | | Scientific Research: Cosmic Frontier | |
| **Author/Company/Email** | | Peter Nugent / LBNL / [penugent@lbl.gov](mailto:eszeto@lbl.gov) | |
| **Actors/Stakeholders and their roles and responsibilities** | | Dark Energy Survey, Dark Energy Spectroscopic Instrument, Large Synoptic Survey Telescope. ANL, BNL, FNAL, LBL & SLAC: Create the instruments/telescopes, run the survey and perform the cosmological analysis. | |
| **Goals** | | Provide a way to reduce photometric data in real-time for supernova discovery and follow-up and to handle the large volume of observational data (in conjunction with simulation data) to reduce systematic uncertainties in the measurement of the cosmological parameters via baryon acoustic oscillations, galaxy cluster counting and weak lensing measurements. | |
| **Use Case Description** | | For DES the data are sent from the mountaintop via a microwave link to La Serena, Chile. From there, an optical link forwards them to the NCSA as well as NERSC for storage and "reduction". Subtraction pipelines are run using extant imaging data to find new optical transients through machine learning algorithms. Then galaxies and stars in both the individual and stacked images are identified, catalogued, and finally their properties measured and stored in a database. | |
| **Current**  **Solutions** | **Compute(System)** | | Linux cluster, Oracle RDBMS server, large memory machines, standard Linux interactive hosts. For simulations, HPC resources. |
| **Storage** | | Oracle RDBMS, psql, as well as GPFS and Lustre file systems and tape archives. |
| **Networking** | | Provided by NERSC |
| **Software** | | Standard astrophysics reduction software as well as Perl/Python wrapper scripts, Linux Cluster scheduling and comparison to large amounts of simulation data via techniques like Cholesky decomposition. |
| **Big Data  Characteristics** | **Data Source (distributed/centralized)** | | Distributed. Typically between observation and simulation data. |
| **Volume (size)** | | LSST will generate 60PB of imaging data and 15PB of catalog data and a correspondingly large (or larger) amount of simulation data. Over 20TB of data per night. |
| **Velocity**  **(e.g. real time)** | | 20TB of data will have to be subtracted each night in as near real-time as possible in order to maximize the science for supernovae. |
| **Variety**  **(multiple datasets, mashup)** | | While the imaging data is similar, the analysis for the 4 different types of cosmological measurements and comparisons to simulation data is quite different. |
| **Variability (rate of change)** | | Weather and sky conditions can radically change both the quality and quantity of data. |
| **Big Data Science (collection, curation,**  **analysis,**  **action)** | **Veracity (Robustness Issues)** | | Astrophysical data is a statistician’s nightmare as the both the uncertainties in a given measurement change from night-to-night in addition to the cadence being highly unpredictable. Also, most all of the cosmological measurements are systematically limited, and thus understanding these as best possible is the highest priority for a given survey. |
| **Visualization** | | Interactive speed of web UI on very large data sets is an ongoing challenge. Basic querying and browsing of data to find new transients as well as monitoring the quality of the survey is a must. Ability to download large amounts of data for offline analysis is another requirement of the system. Ability to combine both simulation and observational data is also necessary. |
| **Data Quality** | | Understanding the systematic uncertainties in the observational data is a prerequisite to a successful cosmological measurement. Beating down the uncertainties in the simulation data to under this level is a huge challenge for future surveys. |
| **Data Types** | | Cf. above on “Variety” |
| **Data Analytics** | |  |
| **Big Data Specific Challenges (Gaps)** | | New statistical techniques for understanding the limitations in simulation data would be beneficial. Often it is the case where there is not enough computing time to generate all the simulations one wants and thus there is a reliance on emulators to bridge the gaps. Techniques for handling Cholesky decompostion for thousands of simulations with matricies of order 1M on a side. | |
| **Big Data Specific Challenges in Mobility** | | Performing analysis on both the simulation and observational data simultaneously. | |
| **Security & Privacy**  **Requirements** | | No special challenges. Data is either public or requires standard login with password. | |
| **Highlight issues for generalizing this use case (e.g. for ref. architecture)** | | Parallel databases which could handle imaging data would be an interesting avenue for future research. | |
| **More Information (URLs)** | | <http://www.lsst.org/lsst>, http://desi.lbl.gov, & http://www.darkenergysurvey.org | |
| **Note:** <additional comments> | | | |

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