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

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | Climate Studies using the Community Earth System Model at DOE’s NERSC center | |
| **Vertical (area)** | | Research: Climate | |
| **Author/Company/Email** | | PI: Warren Washington, NCAR | |
| **Actors/Stakeholders and their roles and responsibilities** | | Climate scientists, U.S. policy makers | |
| **Goals** | | The goals of the Climate Change Prediction (CCP) group at NCAR are to understand and quantify contributions of natural and anthropogenic-induced patterns of climate variability and change in the 20th and 21st centuries by means of simulations with the Community Earth System Model (CESM). | |
| **Use Case Description** | | With these model simulations, researchers are able to investigate mechanisms of climate variability and change, as well as to detect and attribute past climate changes, and to project and predict future changes. The simulations are motivated by broad community interest and are widely used by the national and international research communities. | |
| **Current**  **Solutions** | **Compute(System)** | | NERSC (24M Hours), DOE LCF (41M), NCAR CSL (17M) |
| **Storage** | | 1.5 PB at NERSC |
| **Networking** | | ESNet |
| **Software** | | NCAR PIO library and utilities NCL and NCO, parall el NetCDF |
| **Big Data  Characteristics** | **Data Source (distributed/centralized)** | | Data is produced at computing centers. The Earth Systems Grid is an open source effort providing a robust, distributed data and computation platform,  enabling world wide access to Peta/Exa-scale scientific data. ESGF manages the first-ever decentralized database for handling climate science data, with multiple petabytes of data at dozens of federated sites worldwide. It is recognized as the leading infrastructure for the management and access of large distributed data volumes for climate change research. It supports the Coupled Model Intercomparison Project (CMIP), whose protocols enable the periodic assessments carried out by the Intergovernmental Panel on Climate Change (IPCC). |
| **Volume (size)** | | 30 PB at NERSC (assuming 15 end-to-end climate change experiments) in 2017; many times more worldwide |
| **Velocity**  **(e.g. real time)** | | 42 GByles/sec are produced by the simulations |
| **Variety**  **(multiple datasets, mashup)** | | Data must be compared among those from from observations, historical reanalysis, and a number of independently produced simulations. The Program for Climate Model Diagnosis and Intercomparison develops methods and tools for the diagnosis and intercomparison of general circulation models (GCMs) that simulate the global climate. The need for innovative analysis of GCM climate simulations is apparent, as increasingly more complex models are developed, while the disagreements among these simulations and relative to climate observations remain significant and poorly understood. The nature and causes of these disagreements must be accounted for in a systematic fashion in order to confidently use GCMs for simulation of putative global climate change. |
| **Variability (rate of change)** | | Data is produced by codes running at supercomputer centers. During runtime, intense periods of data i/O occur regularly, but typically consume only a few percent of the total run time. Runs are carried out routinely, but spike as deadlines for reports approach. |
| **Big Data Science (collection, curation,**  **analysis,**  **action)** | **Veracity (Robustness Issues) and Quality** | | Data produced by climate simulations is plays a large role in informing discussion of climate change simulations. Therefore it must be robust, both from the standpoint of providing a scientifically valid representation of processes that influence climate, but also as that data is stored long term and transferred world-wide to collaborators and other scientists**.** |
| **Visualization** | | Visualization is crucial to understanding a system as complex as the Earth ecosystem. |
| **Data Types** | | Earth system scientists are being inundated by an explosion of data generated by ever-increasing resolution in both global models and remote sensors. |
| **Data Analytics** | | There is a need to provide data reduction and analysis web services through the Earth System Grid (ESG). A pressing need is emerging for data analysis capabilities closely linked to data archives. |
| **Big Data Specific Challenges (Gaps)** | | The rapidly growing size of datasets makes scientific analysis a challenge. The need to write data from simulations is outpacing supercomputers’ ability to accommodate this need. | |
| **Big Data Specific Challenges in Mobility** | | Data from simulations and observations must be shared among a large widely distributed community. | |
| **Security & Privacy**  **Requirements** | |  | |
| **Highlight issues for generalizing this use case (e.g. for ref. architecture)** | | ESGF is in the early stages of being adapted for use in two additional domains: biology (to accelerate drug design and development) and energy (infrastructure for California Energy Systems for the 21st Century (CES21)). | |
| **More Information (URLs)** | | <http://esgf.org/>  <http://www-pcmdi.llnl.gov/>  <http://www.nersc.gov/>  <http://science.energy.gov/ber/research/cesd/>  <http://www2.cisl.ucar.edu/> | |
| **Note:** <additional comments> | | | |

**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> | | | |