**NBD(NIST Big Data) Requirements WG Use Case Aug 15 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | NASA LARC/GSFC iRODS Federation Testbed | |
| **Vertical (area)** | | Earth Science Research and Applications | |
| **Author/Company/Email** | | Michael Little, Roger Dubois, Brandi Quam, Tiffany Mathews, Andrei Vakhnin, Beth Huffer, Christian Johnson / NASA Langley Research Center (LaRC) / [M.M.Little@NASA.gov](mailto:M.M.Little@NASA.gov), [Roger.A.Dubois@nasa.gov](mailto:Roger.A.Dubois@nasa.gov), Brandi.M.Quam@NASA.gov, [Tiffany.J.Mathews@NASA.gov](mailto:Tiffany.J.Mathews@NASA.gov), & Andrei.A.Vakhnin@NASA.gov  John Schnase,,Daniel Duffy, Glenn Tamkin, Scott Sinno, John Thompson, & Mark McInerney / NASA Goddard Space Flight Center (GSFC) / John.L.Schnase@NASA.gov, [Daniel.Q.Duffy@NASA.gov](mailto:Daniel.Q.Duffy@NASA.gov), [Glenn.S.Tamkin@nasa.gov](mailto:Glenn.S.Tamkin@nasa.gov). [Scott.S.Sinno@nasa.gov](mailto:Scott.S.Sinno@nasa.gov), [John.H.Thompson@nasa.gov](mailto:John.H.Thompson@nasa.gov), & Mark.Mcinerney@nasa.gov | |
| **Actors/Stakeholders and their roles and responsibilities** | | NASA’s Atmospheric Science Data Center (ASDC) at Langley Research Center (LaRC) in Hampton, Virginia, and the Center for Climate Simulation (NCCS) at Goddard Space Flight Center (GSFC) both ingest, archive, and distribute data that is essential to stakeholders including the climate research community, science applications community, and a growing community of government and private-sector customers who have a need for atmospheric and climatic data. | |
| **Goals** | | To implement a data federation ability to improve and automate the discovery of heterogeneous data, decrease data transfer latency, and meet customizable criteria based on data content, data quality, metadata, and production.  To support/enable applications and customers that require the integration of multiple heterogeneous data collections. | |
| **Use Case Description** | | ASDC and NCCS have complementary data sets, each containing vast amounts of data that is not easily shared and queried. Climate researchers, weather forecasters, instrument teams, and other scientists need to access data from across multiple datasets in order to compare sensor measurements from various instruments, compare sensor measurements to model outputs, calibrate instruments, look for correlations across multiple parameters, etc. To analyze, visualize and otherwise process data from heterogeneous datasets is currently a time consuming effort that requires scientists to separately access, search for, and download data from multiple servers and often the data is duplicated without an understanding of the authoritative source. Many scientists report spending more time in accessing data than in conducting research. Data consumers need mechanisms for retrieving heterogeneous data from a single point-of-access. This can be enabled through the use of iRODS, a Data grid software system that enables parallel downloads of datasets from selected replica servers that can be geographically dispersed, but still accessible by users worldwide. Using iRODS in conjunction with semantically enhanced metadata, managed via a highly precise Earth Science ontology, the ASDC’s Data Products Online (DPO) will be federated with the data at the NASA Center for Climate Simulation (NCCS) at Goddard Space Flight Center (GSFC). The heterogeneous data products at these two NASA facilities are being semantically annotated using common concepts from the NASA Earth Science ontology. The semantic annotations will enable the iRODS system to identify complementary datasets and aggregate data from these disparate sources, facilitating data sharing between climate modelers, forecasters, Earth scientists, and scientists from other disciplines that need Earth science data. The iRODS data federation system will also support cloud-based data processing services in the Amazon Web Services (AWS) cloud. | |
| **Current**  **Solutions** | **Compute (System)** | | NASA Center for Climate Simulation (NCCS) and NASA Atmospheric Science Data Center (ASDC): Two GPFS systems |
| **Storage** | | The ASDC’s Data Products Online (DPO) GPFS File system consists of 12 x IBM DC4800 and 6 x IBM DCS3700 Storage subsystems, 144 Intel 2.4 GHz cores, 1,400 TB usable storage. NCCS data is stored in the NCCS MERRA cluster, which is a 36 node Dell cluster, 576 Intel 2.6 GHz SandyBridge cores, 1,300 TB raw storage, 1,250 GB RAM, 11.7 TF theoretical peak compute capacity. |
| **Networking** | | A combination of Fibre Channel SAN and 10GB LAN. The NCCS cluster nodes are connected by an FDR Infiniband network with peak TCP/IP speeds >20 Gbps. |
| **Software** | | SGE Univa Grid Engine Version 8.1, iRODS version 3.2 and/or 3.3, IBM Global Parallel File System (GPFS) version 3.4, Cloudera version 4.5.2-1. |
| **Big Data  Characteristics** | **Data Source (distributed/centralized)** | | iRODS will be leveraged to share data collected from CERES Level 3B data products including: CERES EBAF-TOA and CERES-Surface products.  Surface fluxes in EBAF-Surface are derived from two CERES data products: 1) CERES SYN1deg-Month Ed3 - which provides computed surface fluxes to be adjusted and 2) CERES EBAFTOA Ed2.7 – which uses observations to provide CERES-derived TOA flux constraints. Access to these products will enable the NCCS at GSFC to run data from the products in a simulation model in order to produce an assimilated flux.  The NCCS will introduce Modern-Era Retrospective Analysis for Research and Applications (MERRA) data to the iRODS federation. MERRA integrates observational data with numerical models to produce a global temporally and spatially consistent synthesis of 26 key climate variables. MERRA data files are created from the Goddard  Earth Observing System version 5 (GEOS-5) model and are stored in HDF-EOS and (Network Common Data Form) NetCDF formats.  Spatial resolution is 1/2 ̊ latitude × 2/3 ̊ longitude ×  72 vertical levels extending through the stratosphere. Temporal resolution is 6-hours for three-dimensional, full spatial resolution, extending from 1979-present, nearly the entire satellite era.  Each file contains a single grid with multiple 2D and  3D variables. All data are stored on a longitude-latitude grid with a vertical dimension applicable for all 3D variables. The GEOS-5 MERRA products are divided into 25 collections: 18 standard products, chemistry products. The collections comprise monthly means files and daily files at six-hour intervals running from 1979 – 2012. MERRA data are typically packaged as multi-dimensional binary data within a self-describing NetCDF file format. Hierarchical metadata in the NetCDF header contain the representation information that allows NetCDF- aware software to work with the data. It also contains arbitrary preservation description and policy information that can be used to bring the data into use-specific compliance. |
| **Volume (size)** | | Currently, Data from the EBAF-TOA Product is about 420MB and Data from the EBAF-Surface Product is about 690MB. Data grows with each version update (about every six months). The MERRA collection represents about 160 TB of total data (uncompressed); compressed is ~80 TB. |
| **Velocity**  **(e.g. real time)** | | Periodic since updates are performed with each new version update. |
| **Variety**  **(multiple datasets, mashup)** | | There is a need in many types of applications to combine MERRA reanalysis data with other reanalyses and observational data such as CERES. The NCCS is using the Climate Model Intercomparison Project (CMIP5) Reference standard for ontological alignment across multiple, disparate data sets. |
| **Variability (rate of change)** | | The MERRA reanalysis grows by approximately one TB per month. |
| **Big Data Science (collection, curation,**  **analysis,**  **action)** | **Veracity (Robustness Issues)** | | Validation and testing of semantic metadata, and of federated data products will be provided by data producers at NASA Langley Research Center and at Goddard through regular testing. Regression testing will be implemented to ensure that updates and changes to the iRODS system, newly added data sources, or newly added metadata do not introduce errors to federated data products. MERRA validation is provided by the data producers, NASA Goddard's Global Modeling and Assimilation Office (GMAO). |
| **Visualization** | | There is a growing need in the scientific community for data management and visualization services that can aggregate data from multiple sources and display it in a single graphical display. Currently, such capabilities are hindered by the challenge of finding and downloading comparable data from multiple servers, and then transforming each heterogeneous dataset to make it usable by the visualization software. Federation of NASA datasets using iRODS will enable scientists to quickly find and aggregate comparable datasets for use with visualization software. |
| **Data Quality** | | For MERRA, quality controls are applied by the data producers, GMAO. |
| **Data Types** | | See above. |
| **Data Analytics** | | Pursuant to the first goal of increasing accessibility and discoverability through innovative technologies, the ASDC and NCCS are exploring a capability to improve data access capabilities. Using iRODS, the ASDC’s Data Products Online (DPO) can be federated with data at GSFC’s NCCS creating a data access system that can serve a much broader customer base than is currently being served. Federating and sharing information will enable the ASDC and NCCS to fully utilize multi-year and multi-instrument data and will improve and automate the discovery of heterogeneous data, increase data transfer latency, and meet customizable criteria based on data content, data quality, metadata, and production. |
| **Big Data Specific Challenges (Gaps)** | |  | |
| **Big Data Specific Challenges in Mobility** | | A major challenge includes defining an enterprise architecture that can deliver real-time analytics via communication with multiple APIs and cloud computing systems. By keeping the computation resources on cloud systems, the challenge with mobility resides in not overpowering mobile devices with displaying CPU intensive visualizations that may hinder the performance or usability of the data being presented to the user. | |
| **Security & Privacy**  **Requirements** | |  | |
| **Highlight issues for generalizing this use case (e.g. for ref. architecture)** | | This federation builds on several years of iRODS research and development performed at the NCCS. During this time, the NCCS vetted the iRODS features while extending its core functions with domain-specific extensions. For example, the NCCS created and installed Python-based scientific kits within iRODS that automatically harvest metadata when the associated data collection is registered. One of these scientific kits was developed for the MERRA collection. This kit in conjunction with iRODS bolsters the strength of the LaRC/GSFC federation by providing advanced search capabilities. LaRC is working through the establishment of an advanced architecture that leverages multiple technology pilots and tools (access, discovery, and analysis) designed to integrate capabilities across the earth science community – the R&D completed by both data centers is complementary and only further enhances this use case.  Other scientific kits that have been developed include: NetCDF, Intergovernmental Panel on Climate Change (IPCC), and Ocean Modeling and Data Assimilation (ODAS). The combination of iRODS and these scientific kits has culminated in a configurable technology stack called the virtual Climate Data Server (vCDS), meaning that this runtime environment can be deployed to multiple destinations (e.g., bare metal, virtual servers, cloud) to support various scientific needs. The vCDS, which can be viewed as a reference architecture for easing the federation of disparate data repositories, is leveraged by but not limited to LaRC and GSFC. | |
| **More Information (URLs)** | | Please contact the authors for additional information. | |
| **Note:** <additional comments> | | | |