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

Use Case Title		MERRA Analytic Services (MERRA/AS)
Vertical (area)		Earth Science Research and Applications
Author/Company/Email		John L. Schnase & Daniel Q. Duffy / NASA Goddard Space Flight Center / John.L.Schnase@NASA.gov, Daniel.Q.Duffy@NASA.gov
Actors/Stakeholders and their roles and responsibilities		NASA's Modern-Era Retrospective Analysis for Research and Applications (MERRA) integrates observational data with numerical models to produce a global temporally and spatially consistent synthesis of 26 key climate variables. Actors and stakeholders who have an interest in MERRA include the climate research community, science applications community, and a growing number of government and private-sector customers who have a need for the MERRA data in their decision support systems.
Goals		Increase the usability and use of large-scale scientific data collections, such MERRA.
Use Case Description		MERRA Analytic Services enables MapReduce analytics over the MERRA collection. MERRA/AS is an example of cloud-enabled Climate Analytics-as-a-Service, which is an approach to meeting the Big Data challenges of climate science through the combined use of 1) high-performance, data proximal analytics, (2) scalable data management, (3) software appliance virtualization, (4) adaptive analytics, and (5) a domain-harmonized API. The effectiveness of MERRA/AS is being demonstrated in several applications, including data publication to the Earth System Grid Federation (ESGF) in support of Intergovernmental Panel on Climate Change (IPCC) research, the NASA/Department of Interior RECOVER wildland fire decision support system, and data interoperability testbed evaluations between NASA Goddard Space Flight Center and the NASA Langley Atmospheric Data Center.
Current Solutions	**Compute(System)**	NASA Center for Climate Simulation (NCCS)
	Storage	The MERRA Analytic Services Hadoop Filesystem (HDFS) is a 36 node Dell cluster, 576 Intel 2.6 GHz SandyBridge cores, 1300 TB raw storage, 1250 GB RAM, 11.7 TF theoretical peak compute capacity.
	Networking	Cluster nodes are connected by an FDR Infiniband network with peak TCP/IP speeds >20 Gbps.
	Software	Cloudera, iRODS, Amazon AWS
Big Data Characteristics	**Data Source (distributed/centralized)**	MERRA data files are created from the Goddard Earth Observing System version 5 (GEOS-5) model and are stored in HDF-EOS and 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, 7 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)	480TB
	Velocity (e.g. real time)	Realtime or batch, depending on the analysis. We're developing a set of "canonical ops" - early stage, near-data operations common to many analytic workflows. The goal is for the canonical ops to run in near realtime.
	Variety (multiple datasets, mashup)	There is a need in many types of applications to combine MERRA reanalysis data with other reanalyses and observational data. We are 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 provided by data producers, NASA Goddard's Global Modeling and Assimilation Office (GMAO).
	Visualization	There is a growing need for distributed visualization of analytic outputs.
	Data Quality	Quality controls applied by data producers, GMAO.
	Data Types	See above.
	Data Analytics	In our efforts to address the Big Data challenges of climate science, we are moving toward a notion of Climate Analytics-as-a-Service (CAaaS). We focus on analytics, because it is the knowledge gained from our interactions with Big Data that ultimately produce societal benefits. We focus on CAaaS because we believe it provides a useful way of thinking about the problem: a specialization of the concept of business process-as-a-service, which is an evolving extension of IaaS, PaaS, and SaaS enabled by Cloud Computing.
Big Data Specific Challenges (Gaps)	A big question is how to use cloud computing to enable better use of climate science's earthbound compute and data resources. Cloud Computing is providing for us a new tier in the data services stack — a cloud-based layer where agile customization occurs and enterprise-level products are transformed to meet the specialized requirements of applications and consumers. It helps us close the gap between the world of traditional, high-performance computing, which, at least for now, resides in a finely-tuned climate modeling environment at the enterprise level and our new customers, whose expectations and manner of work are increasingly influenced by the smart mobility megatrend.	
Big Data Specific Challenges in Mobility	Most modern smartphones, tablets, etc. actually consist of just the display and user interface components of sophisticated applications that run in cloud data centers. This is a mode of work that CAaaS is intended to accommodate.	
Security & Privacy Requirements	No critical issues identified at this time.	

Highlight issues for generalizing this use case (e.g. for ref. architecture)	MapReduce and iRODS fundamentally make analytics and data aggregation easier; our approach to software appliance virtualization in makes it easier to transfer capabilities to new users and simplifies their ability to build new applications; the social construction of extended capabilities facilitated by the notion of canonical operations enable adaptability; and the Climate Data Services API that we're developing enables ease of mastery. Taken together, we believe that these core technologies behind Climate Analytics-as-a-Service creates a generative context where inputs from diverse people and groups, who may or may not be working in concert, can contribute capabilities that help address the Big Data challenges of climate science.
More Information (URLs)	Please contact the authors for additional information.
Note: <additional comments>	

Note: No proprietary or confidential information should be included