NASA SBIR 2015 Phase I Solicitation

S5.03 Algorithms and Tools for Science Data Processing, Discovery and Analysis, in State-of-the-Art Data Environments

Lead Center: GSFC

Participating Center(s): ARC, JPL, KSC, LaRC, MSFC, SSC

The size of NASA's observational data sets is growing dramatically as new missions come on line. In addition, NASA scientists continue to generate new models that regularly produce data sets of hundreds of terabytes or more. It is growing ever increasingly difficult to manage all of the data through its full lifecycle, as well as provide effective data analytical methods to analyze the large amount of data.Â

Using remote observation examples, the HyspIRI mission is expected to produce an average science data rate of 800 million bits per second (Mbps), JPSS-1 will be 300 Mbps and NPP is already producing 300 Mbps, compared to 150 Mbps for the EOS-Terra, Aqua and Aura missions. Other examples are SDO with a rate of 150 Mbps and 16.4 Gigabits for a single image from the HiRise camera on the Mars Reconnaissance Orbiter (MRO). From the NASA climate models, the MERRA reanalysis data set is approximately 200 TB, and MERRA2 will start generating even more data late in 2014.

This subtopic area seeks innovation and unique approaches to solve issues associated around the use of "Big Data" within NASA. The emphasis of this subtopic is on tools that leverage existing systems, interfaces, and infrastructure, where it exists and where appropriate. Reuse of existing NASA assets is strongly encouraged.Â

Specifically, innovations are being sought in the following areas:Â

- *Parallel Processing for Data Analytics* - Open source tools like the Hadoop Distributed File Systems (HDFS) have shown promise for use in simple MapReduce operations to analyze model and observation data. In addition to HDFS, there is a rapid emergence of an ecosystem of tools associated with high performance data analytics using cloud software packages, such as Hive, Impala, Spark, etc. The goal is to accelerate these types of open source tools for use with binary structured data from observations and model output using MapReduce or a similar paradigm.
- *High Performance File System Abstractions* - NASA scientists currently use a large number of existing applications for data analysis, such as GrADS, python scripts, and more, that are not compatible with an object storage environment. If data were stored within an object storage environment, these applications would not be able to access the data. Many of these applications would require a substantial amount of investment to enable them to use object storage file systems. Therefore, a file system abstraction, such as FUSE (file system in user space) is needed to facilitate the use of existing data analysis applications with an object storage environment. The goal is to make a FUSE-like file system abstraction robust, reliable, and highly performing for use with large NASA data sets.
- *Data Management of Large-Scale Scientific Repositories* - With increasing size of scientific repositories comes an increasing demand for using the data in ways that may never have been imagined when the

repository was conceived. The goal is to provide capabilities for the flexible repurposing of scientific data, including large-scale data integration, aggregation, representation, and distribution to emerging user communities and applications.

- *Server Side Data Processing* - Large data repositories make it necessary for analytical codes to migrate to where the data are stored. In a densely networked world of geographically distributed repositories, tiered intermediation is needed. The goal is to provide support for migratable codes and analytical outputs as first class objects within a provenance-oriented data management cyberinfrastructure.
- *Techniques for Data Analysis and Visualization* - New methods for data analytics that scale to extremely large and geographically distributed data sets are necessary for data mining, searching, fusion, subsetting, discovery, visualization, and more. In addition, new algorithms and methods are needed to look for unknown correlations across large, distributed scientific data sets. The goal is to increase the scientific value of model and observation data by making analysis easier and higher performing. Among others, some of the topics of interest are:
    - Techniques for automated derivation of analysis products such as machine learning for extraction of features in large image datasets (e.g., volcanic thermal measurement, plume measurement, automated flood mapping, disturbance mapping, change detection, etc.).
    - Workflows for automated data processing, interpretation, and distribution.Â

Research proposed to this subtopic should demonstrate technical feasibility during Phase I, and in partnership with scientists, show a path toward a Phase II prototype demonstration, with significant communication with missions and programs to ensure a successful Phase III infusion. It is highly desirable that the proposed projects lead to software that is infused into NASA programs and projects.Â

Tools and products developed under this subtopic may be used for broad public dissemination or within a narrow scientific community. These tools can be plug-ins or enhancements to existing software, on-line data/computing services, or new stand-alone applications or web services, provided that they promote interoperability and use standard protocols, file formats and Application Programming Interfaces (APIs) or prevalent applications.Â