# NASA SBIR 2016 Phase I Solicitation

## S5.03 Enabling NASA Science through Large-Scale Data Processing and Analysis

**Lead Center: GSFC**

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

The size of NASA's observational data sets is growing dramatically as new mission data become available. In addition, NASA scientists continue to generate new models that regularly produce data sets of hundreds of terabytes or more. It is growing increasingly difficult for NASA to effectively analyze such large data sets for use within their science projects.Â

The following lists show representative examples of both observational and model generated data sets that are relevant to NASA science projects. This list is not meant to be all-inclusive, but rather to provide examples of data sets and to show the extent of the â€œBig Dataâ€ problems encountered by NASA. Some remote observation examples are the following:Â

- 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.
- 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).
- Landsat and MODIS data sets continue to grow at extremely high rates.
- National Geospatial Agency (NGA) high-resolution imagery data of the Earth.Â

From the NASA climate models, some examples include:Â

- The MERRA2 reanalysis data set is approximately 400 TB.
- Several high-resolution nudged and free running climate simulations have generated Petabytes of data (all publically releasable).Â

This subtopic area seeks innovative, unique, forward-looking, and replicable approaches for using â€œBig Dataâ€ for NASA science programs. The emphasis of this subtopic is on the creation of novel analytics, tools, and infrastructure to enable high performance analytics across large observational and model data sets. *Proposals should be in alignment with existing and/or future NASA science programs*, and the reuse of existing NASA assets is strongly encouraged.

Specifically, innovative proposals are being sought to assist NASA science in the following areas (note that this list is not inclusive and is included to provide guidance for the proposers):Â

- New services and methods for high performance analytics that scale to extremely large data sets â€“ of specific interest are the following:
    - Techniques for data mining, searching, fusion, subsetting, discovery, and visualization
    - Automated derivation of analysis products in large data sets, that can then be utilized into Science models â€“ the following are two representative examples
        - Extraction of features (e.g., volcanic thermal measurement, plume measurement, automated flood mapping, disturbance mapping, change detection, etc.).
        - Geospatial and temporal correlation of climate events (e.g., hurricanes, mesoscale convective systems, atmospheric rivers, etc.).
- Methods to enable in-situ, data proximal, parallel data analytics that will accelerate the access, analysis, and distribution of large Science datasets.
    - Potential use of open source data analytic tools (such as Hadoop, MapReduce, Spark, etc.) to accelerate analytics.
    - Application of these tools to structured, binary, scientific data sets.
    - Performing analytics across both physically collocated and geographically distributed data.
    - High performance file systems and abstractions, such as the use of object storage file systems.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 later plan a potential 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 developed for broad public dissemination or used 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).Â