



NASA SBIR 2017 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

Technology Area: TA11 Modeling, Simulation, Information Technology and Processing

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.

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 "Big Data" problems encountered by NASA. Some remote observation examples are the following:

- The HypsIRI 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.

From the NASA climate models, some examples include:

- Reanalysis data sets such as MERRA (200 TB) MERRA2 (400 TB), and emerging reanalysis data sets with chemistry components.
- Several high-resolution nudged and free running climate simulations have generated Petabytes of data (all publically releasable).

This subtopic area seeks innovative, unique, forward-looking, and replicable approaches for using "Big Data" for NASA science programs. The emphasis of this subtopic is on the creation of novel analytics, tools, infrastructure, and/or algorithms to enable high performance analytics across large observational and model data sets.

Proposals MUST be in alignment with existing and/or future NASA programs and address or extend a specific need or question for those programs. It is therefore incumbent upon the proposers to have discussions with NASA scientists and engineers to receive feedback prior to submission and to adequately show the alignment of the

proposed innovation to NASA.

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):

- New services, methods, and/or algorithms for high performance analytics that scale to extremely large data sets – of specific interest are the following:
 - Preference to employ machine and deep learning methods
 - Other techniques to be considered could cover 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 – 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.).
- New services, methods, and/or algorithms to enable in-situ, data proximal, parallel data analytics that will accelerate the access, analysis, and distribution of large Science datasets.
 - Use of open source data analytic tools to accelerate analytics is desired.
 - 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.

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.

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).