General Atomics Response to the Request for Input (RFI)-National Big Data R&D Initiative

Robert Murphy is the Big Data Program Manager for General Atomics Energy and Advanced Concepts. Previously, Robert was responsible for High Performance Data and Analytics in IBM's Software Defined Environments organization. Before IBM, Bob held positions with increasing levels of responsibility at Hewlett Packard, Silicon Graphics, Oracle, and Sun Microsystems. He has a Biomedical Engineering Degree from Purdue University.

Robert Murphy
Big Data Program Manager, Energy and Advanced Concepts
General Atomics
3550 General Atomics Court, San Diego, CA 92121
robert.murphy@ga.com

Comments and suggestions from General Atomics

1) Data from large-scale experiments and extreme-scale computing is expensive to produce and may be used for high-consequence applications. However, it is not the mere existence of data that is important, but our ability to make use of it. Experience has shown that as we make the associated metadata better organized and more complete, the more useful the underlying data becomes. In line with the "trustworthiness of data" and "ensuring the long-term sustainability…of high value data," generous provisioning of metadata, including data provenance and data relations, is critical to enhance data sharing, to allow data to retain its usefulness over extended periods of time, and to allow traceability of results. There is an unmet need to better document workflows that create, transform, or disseminate data and to capture (and later present) data provenance, enabling scientists to answer the questions "who, what, when, how and why" for each data element; provide information about the connections and dependencies between the data elements; and allow human or automatic annotation for any data element.

2) In addition to "schema on read" (where data is applied to a plan or schema as it is pulled out of a stored location) data analysis and handling tools like Hadoop and MapReduce, NITRD should explicitly incorporate "schema on write" (where data is mapped to a plan or schema when it is written) metadata centric data analysis and handling tools in their VISION STATEMENT. Metadata centric "schema on write" techniques are critical to meeting NITRD goals of ensuring data consistency and trustworthiness along with being the only techniques capable of meeting NITRD member agency concerns such as data identification, access, reproducibility, provenance, curation, unique referencing, and future data availability.

3) Incorporate best practices in Big Data pioneered at NITRD agencies such as the DOE (SLAC BaBar) and NASA (HST and Kepler), that due to the nature of their work, have already encountered and overcame many of the Big Data problems that will be met more broadly by NITRD agencies and incorporated in the NITRD VISION STATEMENT.

4) Where possible, NITRD should encourage the use of existing, cost effective, supported, commercially available tools that meet NITRD member agency requirements and discourage expending NITRD agency resources on developing redundant internal tools that are difficult to sustain and leverage across the multiple NITRD agencies.

5) The ability to orchestrate global data intensive NITRD agency workflows and access and manage data on multiple storage devices, anywhere in the world is required. NITRD agency data needs to be easily and securely shared among globally distributed teams within NITRD member agencies. Data needs to move automatically to various workflow resources, based on policies, so data is always available at the right place, at the right time, and at the right cost — while keeping an audit trail as data is ingested, transformed, accessed, and archived through its complete lifecycle.

6) Like a needle in a haystack, high value data stored in NITRD agency storage systems can be effectively lost over time – stranding this data and losing its value forever. Metadata-based tagging and tracking of valuable information is needed so NITRD agency data can be found and analyzed even if it resides on very different, incompatible platforms anywhere in the world.

7) It's essential to maintain NITRD member agency data provenance, audit, security, and access control in order to track data within workflows, through all transformations, analyses, and interpretations. NITRD agency data needs to be optimally managed, shared, and reused with verified provenance of the data and the conditions under which it was generated – so results are reproducible and analyzable for defects.

8) It's imperative for NITRD member agencies to restrain storage costs by incorporating tools and procedures to:
   - Prevent worthless data from entering NITRD agency workflows and being stored
   - Migrate data to lower cost storage tiers using workflow policies where possible
   - Remove data that's no longer valuable
   - Consolidate and automate complete data lifecycle management across multiple NITRD agency resources, avoiding costly individual resource management and administration