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

Michael E. Papka, ALCF Director - papka@anl.gov
Contributors: David Martin, Hal Finkel, William Allcock, Vitali A. Morozov, William Scullin
(dem, hfinkel, allcock, morozov, wscullin)@alcf.anl.gov
Submitted on behalf of the
Argonne Leadership Computing Facility
Argonne National Laboratory, Argonne, IL 60439

The Argonne Leadership Computing Facility (ALCF) is one of two DOE Leadership Computing Facility (LCF) centers in the nation dedicated to open science. Supported by the DOE's Advanced Scientific Computing Research program, the LCF operates centers at Argonne National Laboratory and Oak Ridge National Laboratory, and deploys two diverse high-performance computer architectures that are 10 to 100 times more powerful than systems typically available for open scientific research. The LCF provides world-class computational capabilities to the scientific and engineering community to advance fundamental discovery and understanding in a broad range of disciplines. ALCF supports over 1,100 users and roughly 250 projects. ALCF users have generated more than 5 petabytes (PB) of data this year and ALCF manages over 13 PB of user data on disk and 14 PB on tape. For more information about the ALCF, visit alcf.anl.gov.

What are the gaps that are not addressed in the Visions and Priority Actions document?

Hierarchical and Distributed Data Management. A comprehensive management system to optimize data movement and placement between the various levels and locations of the storage hierarchy would both accelerate data-intensive applications and reduce or eliminate the need for programmers to develop custom data management schemes. Development of such a system will require the cooperation of storage vendors, operating systems developers, and applications programmers.

Co-Managing Computing and Long-Term Storage. Scientific datasets are growing ever-larger, and moving them around for simulation or analysis has become increasingly impractical. Instead, middleware and algorithms should manage the optimal placement of data in order to carry out the calculations requested. Research and development of new co-management techniques where compute, network, and storage are dynamically managed to provide the overall optimization, including predictive analysis of overall systems behavior, is essential. Dedicated facilities for the long-term storage of carefully curated data should also be established.

Trusted Data. Investment is required to ensure end-to-end data integrity, validation, recoverability, access control, and security are part of the Big Data ecosystem. Expecting each application development team to independently integrate these technologies is unrealistic. Instead, these capabilities should be provided by underlying software libraries and integrated into the hardware infrastructure. Likewise, significant work is needed to develop end-to-end security that does not inhibit scientific data analysis or sharing, but still provides the required level of security and access control.

From an interagency perspective, what do you think are the most high impact ideas at the frontiers of big data research and development?

Metadata. Metadata describes the content and structure of a particular dataset, and is critical to making that data available beyond the scientific group that created it. As experimental, sensor, and simulation datasets become larger, more complex, and potentially useful to a broader set of scientists, rich and well-structured metadata is essential. Domain scientists must be encouraged to invest time and effort into making their datasets easier for others to use, which vastly increases the value of the data and the potential for interagency collaboration.

Portals. Because the movement of large datasets is expensive and time consuming, co-locating accessible compute capability with data storage increases the potential of their scientific usefulness. However, most potential data users are not programmers skilled in the use of supercomputers or complex datasets. Web-based data portals are the best way to make these datasets widely usable. Computing facilities like the ALCF must invest in the infrastructure and support needed to host and integrate these interfaces.

What new research, education, and/or infrastructure investments do you think will be game-changing for the big data innovation ecosystem?

Investment in Data Facilities. DOE National User Facilities generate massive experimental, observational, and simulation datasets that are critical to increasing scientific understanding in a variety of fields. Investments at facilities like the ALCF are required to make this data widely available through structured data management and the availability of experts to help in data retrieval and analysis.

How can the federal government most effectively enable new partnerships, particularly those that cross sectors or domains?

Data User Facilities. There is the need for federally funded national user facilities for data that parallel the current national facilities for computation. These data user facilities would be dedicated to the curation, storage, and dissemination of important data sets. Such facilities must have the computing capabilities to provide analysis and visualization without requiring excessive data movement. In addition, these facilities will ensure the integrity and security of the data, and make it available to a broad range of scientific and industrial projects. The federal government should also establish clear rules for data ownership and requirements for data sharing.

Summary. Facilities like ALCF are at the forefront of applying advanced computing, storage, and networking technology to achieve breakthroughs in science and engineering. By enhancing the ability of facilities like ALCF to provide open access to important datasets, along with the computing time and the human expertise needed to make use of these datasets, the federal government will usher in a new era of scientific discovery that leverages existing data in novel ways.