ASCR Workshop January 2022

Basic Research Needs for **Management and Storage of Scientific Data**

U.S. DEPARTMENT OF ENERGY | Office of Science

Scientific discoveries rely heavily on efficient access, search, and management of massive data sets.

Data management technologies have, for decades, provided foundational capabilities for scientific computing. Just as storage, input/output (I/O), and data management have been fundamental to simulation-based science for many years, so too are capable data-management technologies key to the success of today's scientific workflows utilizing data intensive and machine learning (ML) techniques. The Department of Energy, Office of Science, Advanced Scientific Computing Research (ASCR) program has invested broadly in data-management research focused on high-performance computing (HPC) systems, from parallel file systems that store data to application software that makes these systems more productive. Still, advances in technology combined with growing diversity of supported science strongly motivate continued investment in this area.

In January 2022, ASCR convened a workshop to identify priority research directions in the area of data management for high-performance and scientific computing. Attendees were challenged to identify promising approaches that would support the breadth of the DOE mission, including the explosion of artificial intelligence (AI) uses and the growing needs of experimental and observational science. Technological and science drivers were identified and considered as they relate to key aspects of data management such as interfaces, architectural design, and FAIR principles (Findable, Accessible, Interoperable, and Reusable). The thoughts of the workshop participants were distilled into a set of four priority research directions with the potential for high impact on DOE science. These research directions are summarized in the following pages.

Full details can be found in the workshop report at **doi.org/10.2172/1845707**

WHAT DO OUR SCIENTISTS AND FACILITIES NEED?

EXAMPLES

- Experimental/observational data science
- Workflows, data processing pipelines, ML/AI applications
- FAIR principles, introspection, provenance

WHAT NEW TECHNOLOGIES MIGHT BE BENEFICIAL?

EXAMPLES

- In situ and in transit data analysis
- New storage, memory, networking technologies
- Disaggregated, dynamically provisioned resources

Priority Research Directions

1. High-productivity interfaces for accessing scientific data efficiently

Key Questions:

1. How can application developers search and access important information seamlessly in massive amounts of scientific data?
2. What changes are needed to existing I/O application programming interfaces (APIs) to enable complex AI workflows?
3. What are effective interfaces and abstractions for capturing user intent for optimizing data management?

A redesign of data access interfaces is critically important to locate and to access data in deep memory and storage hierarchies and across systems (e.g., memory, file systems, archives, online repositories, edge devices, and cloud storage). New interfaces are needed for enabling data management in complex AI workflows. Interfaces are also needed to capture user intent (e.g., metadata and provenance, data usage pattern, etc.) for optimizing workflows, performing automated data movement, and extracting important information from datasets.

2. Understanding the behavior of complex data management systems in DOE science

Key Questions:

1. How can disparate information from multiple sources regarding data management activities be fused into useful knowledge?
2. In what ways can people and software leverage this knowledge to improve the reliability and performance of data management systems?

Understanding the behavior of complex data management systems, including user behavior, underlying hardware behavior, and associated compute and networking activities, is key to maximizing the reliability and performance of these systems. Through improved understanding we can eliminate application bottlenecks and unlock the potential of AI to enable the next generation of self-tuning data management services.

3. Rich metadata and provenance collection, management, search, and access

Key Questions:

1. What metadata and provenance are needed to support FAIR principles?
2. How do we support collection, storage, and search of rich metadata and provenance?
3. How can we use rich metadata and provenance for optimizing data management?

Metadata and provenance are critical for supporting the FAIR principles for reproducible science. R&D efforts are needed to enable management of the voluminous metadata inherent in modern science, to identify metadata and provenance that are effective for supporting FAIR principles, and to understand how to best collect and use metadata and provenance for improving data management systems and scientific discovery as a whole.

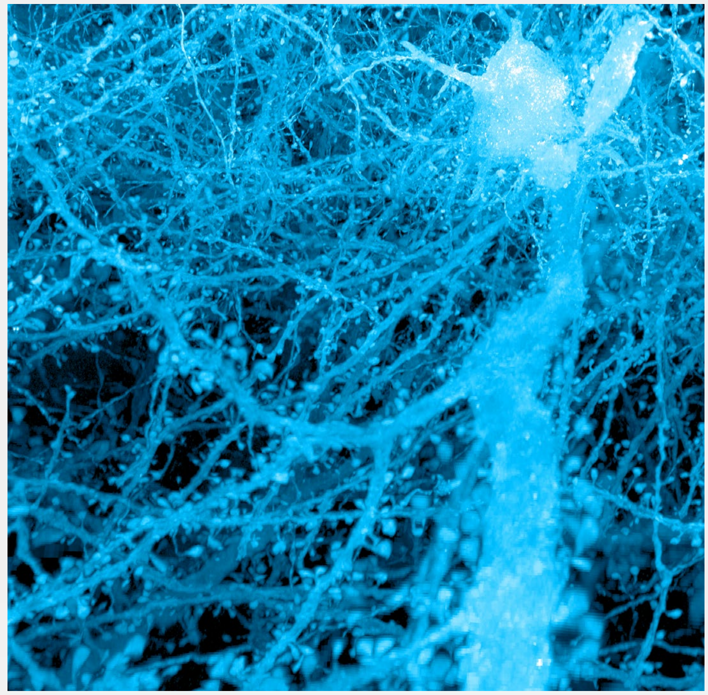
4. Reinventing data services for new applications, devices, and architectures

Key Questions:

1. Using a co-design approach, how do we create specialized data services leveraging emerging device technologies to enable revolutionary breakthroughs across the breadth of DOE science?

New science endeavors and approaches require specialization of how data are accessed, organized, and retained. New networking and storage devices, including ones with computational capabilities, merit revisiting data service design to maximally exploit these technologies. New architectures, including scenarios in which data lives across sites, across administrative domains, or are generated at the edge, similarly place new requirements on data services. Co-design of these services with scientists, hardware architects, and facility operators is needed to unlock the potential of data in these unique environments and ease porting to new ones.

Opportunities from AI



The use of AI technologies to reveal insights in massive scientific datasets has been increasing. R&D efforts are needed to support AI applications as well as to use AI for optimizing data-management systems. These efforts cut across the four priority research directions identified above. We need to rethink I/O APIs to enable complex AI workflows on modern HPC and edge systems that have diverse memory/storage hierarchies and heterogeneous accelerators. Better schedulers and data-movement tools are needed for using computation, memory, and storage resources efficiently to perform training and inference. On the other hand, AI analysis methods are useful for understanding behavior of data movement and for optimizing data management services and architectures. Research is also needed on AI methods for analysis of metadata and provenance information to identify or recommend relevant datasets and information to scientists. Achieving these goals also requires development of representative benchmarks for complex AI workflows.

Summary

In the future, scientific activities will encompass an increasingly broad range of domains and span both HPC resources and advanced scientific instruments. Scientists and facility operators working together to co-design data management architectures will ensure we have the most capable and robust tools for managing these troves of valuable scientific results. Improvements in how we describe and structure these data will enable greater sharing of data than ever before, and facilitate automation of science with artificial intelligence.

COVER PHOTO: Eye of a developing zebrafish embryo imaged using adaptive optical lattice light-sheet microscopy. The cell boundaries were computationally segmented and cells separated in 3D. Cells are colored based on their 3D position. The size of the data collected to produce the image was 6.2 TB. The latest microscopes in this research produce 1 TB per hour and require analyzing data that is stored in multiple locations, e.g., HPC centers, cloud, and near microscopes. Advances in data management and storage techniques are critical for supporting data intensive workflows like this example, which are an important step towards the goal of enabling real-time control of microscope configurations and other self-driving experiments. (Liu, T., Upadhyayula, S., et al. Science 360, 284, 2018, DOI: 10.1126/science.aaq1392)

BACK PHOTO: Forest of sparsely labeled neurons in the somatosensory cortex of a mouse brain imaged using expansion microscopy and lattice light-sheet microscopy. [Size of the data collected to produce the image: 3.4 TB] (Gao, R., Asana, S., Upadhyayula, S., et al. Science 363, 245, 2019, DOI: 10.1126/science.aau8302)

Images provided by Gokul Upadhyayula, UC Berkeley and LBNL.