Data Lifecycle Instrument (DaLI) towards reproducibility of scientific research

Hakizumwami Birali Runesha and Milson Munakami
Research Computing Center, University of Chicago

ABSTRACT

With the ever-increasing volume, velocity, and variety of data that is created from research lab instruments, comes the challenges of meeting the increased demands for transferring, storing, analyzing, sharing, and publishing these data. To address these challenges in data management and computation, the Research Computing Center (RCC) designed and deployed a Data Lifecycle Instrument (DaLI) for managing and sharing data from instruments, observations, and simulations. DaLI makes it possible to seamlessly capture data from instruments, share data and collect the required artifacts for the reproducibility and preservation of scientific research products. DaLI also serves as an opportunity to prepare and engage the next-generation students and researchers.

Objectives

Develop a cost-effective platform to enable users to (a) acquire, transfer, process, and store data from experiments, observations and simulations in a unified workflow, (b) manage data over their entire life cycle, and (c) share and publish data and (d) enhance outreach and education opportunities. The instrument should serve as a foundational platform to capture data along with provenance metadata of how the results are obtained and reproduced.

Motivation

Challenges

- Research Instruments, operating at full capacity, are generating data at incredible scales, creating computation, storage, and data management challenges.
- The complexity of managing data has resulted in a fragmented data landscape where data is often spread across many devices and core facilities with no cohesive data management strategy.
- Difficulties in intra- and inter-institutions collaboration to complete scientific research.
- Reproducibility in science has raised concerns about the transparency and reuse of published scientific results.

DaLI Solution

- DaLI includes a large selection of software modules. It includes data management and custom-built applications such as the RCC XDM, MyTardis, Globus, etc.
- DaLI ensures ease of use and facilitates data collection and transport along with associated metadata for data reuse.

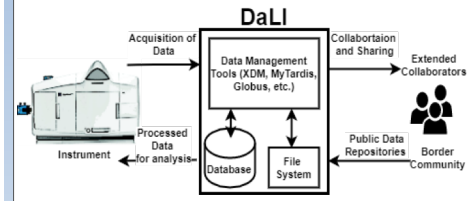

Figure 1: Example of a DaLI workflow

DaLI Platform

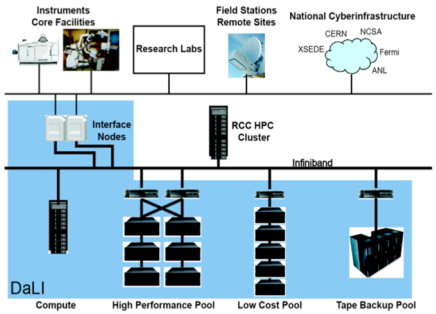

Figure 2: DaLI Architecture

DaLI Hardware

DaLI is a highly scalable and secured HPC infrastructure that

- features compute resources for pre- and post-processing, data movers and tiered storage pools that include an HPC storage pool, a low-cost storage pool, and backup
- has a building block of 1200 compute cores using Intel Xeon Gold compute nodes.
- has data movers
- has 3-tiered storage pools integrated with a fast network fabric to build a unified and hierarchical storage architecture.
- has 4.2 PB DDN Gridscaler 14KXE storage
- uses Spectrum Scale high-performance Global Parallel file system (GPFS) and an Active File Management (AFM).
- runs Slurm workload Manager

Results

Figure 3: Xenon1T dark matter detector in the Gran Sasso Underground Laboratory in Italy. Data is stored and analyzed on Dali at the RCC

- DaLI ensures data security and integrity throughout the data lifecycle.
- DaLI is enabling more than 20 research projects such as the Xenon1T dark matter detector in the Gran Sasso Underground Laboratory in Italy, Light sheet microscopy, Sign and Gesture Archive (SAGA), XROMM core facility, Chicago Electrical Neuroimaging Analytics (CENA), Chicago motion database, etc.
- DaLI has 609 users and more than 1 PB of active data.
- DaLI is also used by a range of courses and workshops to engage students and researchers.
- DaLI runs outreach programs to underprivileged communities.

Future work

- Develop tools to semi-automatically capture metadata throughout the data lifecycle to make reproducible scientific research on HPC systems effortless
- Integration of DaLI with UChicago Institutional Repository
- Interoperability of the DaLI instrument with other repositories and cloud resources

Conclusion

- Designed and deployed a data lifecycle instrument that is cost-effective and replicable to address needs associated with management and sharing of data.
- DaLI's features are facilitating the collection of data, associated metadata and artifacts needed to reproduce published results.
- Using microservices APIs, DaLI can enable the integration of data from edge computing resources.
- Support multiple fields that span astronomy, ecology, biology, high-energy physics, molecular engineering, linguistics, and neuroscience.
- Allows collaboration and sharing of data and promotion of transparency and reproducible scientific research.

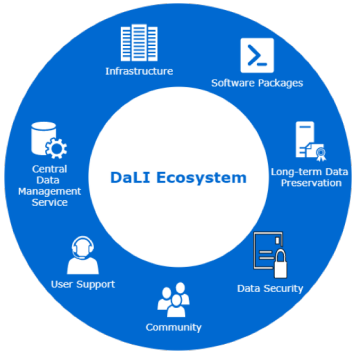

Figure 4: DaLI Ecosystem

Acknowledgments

NSF Award 1661523

THE UNIVERSITY OF CHICAGO | Research, Innovation and National Laboratories
Research Computing Center