DaLI: A Data Lifecycle Instrument towards the reproducibility of scientific research

Hakizumwami Birali Runesha
Research Computing Center
University of Chicago
Chicago IL USA
runesha@uchicago.edu

Milson Munakami
Research Computing Center,
University of Chicago
Chicago IL USA
mmunakami@uchicago.edu

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 University of Chicago's Research Computing Center (RCC) designed and deployed a Data Lifecycle Instrument (DaLI) for managing and sharing data from instruments, observations, and simulations. DaLI is a High-Performance Computing (HPC) storage and compute-platform that provides access to high-end hardware and specialized software tools to allow researchers to securely transfer, organize, catalog, preserve and distribute large quantities of data along with the associated metadata. 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. The collection of reproducible data and metadata requires new tools and platforms that facilitate data curation and management. DaLI also serves as an opportunity to prepare and engage the next-generation students and researchers working with instrument data and HPC.

Keywords- Research Instrumentation, Scientific reproducibility, Advanced Cyberinfrastructure, Research data lifecycle management

I. INTRODUCTION

Research instruments, such as telescopes, massively parallel DNA sequencers, high-speed cameras, and biplanar X-ray video systems, and Advanced Photon Source (APS) beamlines, etc. when operating at full capacity, generate data at incredible scales, creating computation and data management challenges. Instruments are often housed in shared research facilities or in geographically distant locations where data is collected and then transferred back to the researcher's home system for analysis. Despite favorable trends in cost-effective storage and capacity, data lifecycle management complexities (acquisition, transfer, analysis, sharing, and publication) result in fragmented data landscapes where data is spread across many instruments and facilities with no cohesive data management strategy. Each scientific breakthrough relies on speed and ease of availability of such large-scale data and metadata [1] [2] [3] that can be discovered, shared, reproduced, and reused. Due to the lack of proper infrastructure and standards, the inability to reproduce published data have raised questions about the reliability and transparency of scientific results. Having research instruments that support data management capabilities is crucial for the reproducibility of scientific research that facilitates reuse and validation of results.

II. THE DATA LIFECYCLE INSTRUMENT

A. *Infrastructure*

The RCC worked with 20 faculty research projects to design, build, and deploy the DaLI[1] platform that enables users to (a) acquire, transfer, process, and store data from experiments and observations 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. *Figure 1* shows the DaLI architecture. DaLI is a secure platform that combines the use of HPC, large-scale storage resources, advanced data management tools. DaLI provides secure and reliable data management capabilities that are ideal for data sharing and collaboration among local and external research groups. In addition to data management needs, there is an urgent need for infrastructures that make data Findable, Accessible, Interoperable, and Reusable (FAIR) to address challenges associated with the discoverability, reproducibility, and reusability of research results [4] [5]. DaLI is a cost-effective and replicable platform to support data management from acquisition to publication. A FAIR instrument for reproducibility of scientific research would capture data along with provenance metadata of how the results are obtained and how they could be reproduced [6] [7] [8].

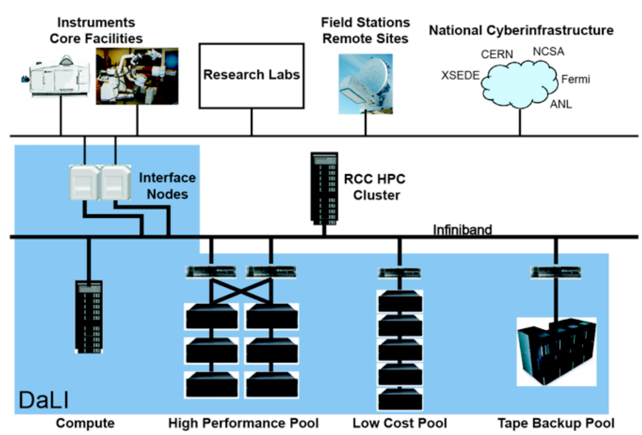

Figure 1: DaLI architecture

B. *The DaLI system*

DaLI is a highly scalable and secured HPC infrastructure that consists of compute resources for pre- and post-processing, data movers, and tiered storage pools that include a high-

[1] https://dali.rcc.uchicago.edu/

performance storage pool, a low-cost storage pool, and a backup storage pool that are integrated by an EDR InfiniBand fast network fabric into a unified and hierarchical storage architecture. It is outfitted with high-performance 3D XPoint non-volatile memory to accelerate the processing of instrument-generated data in-core and relieving I/O bottlenecks.

The easy-to-use data services on DaLI enable researchers to take full advantage of these resources with minimal expertise. DaLI uses the spectrum scale high-performance Global Parallel File System (GPFS), Active File Management (AFM), and Globus Online service. For scalability, DaLI is built to allow researchers to add hardware to its building block, which consists of Intel Xeon Gold compute nodes with 192GB DDR4 RAM each, a data mover node, EDR (100Gb/s) network interconnect, and a 4.2 PB GPFS DDN Gridscaler 14KXE storage.

Some of the unique features of DaLI include the customs tools (e.g. XDM) developed to automatically capture data from the instrument and to make data discoverable from the DaLI storage, the use of microservices, and containers to enable the easy deployment, distribution, and replication of applications.

The seamless integration of DaLI facilitates collaborations and data transport in a fragmented landscape where data is often spread across different locations. With DaLI, you run an experiment and you are ready to sit on your computer to access the data and start your data processing. With its adaptive architecture, DaLI is equipped to scale in a unified fabric and ensure proper utilization and allocation of available resources to users. DaLI is connected to the outside world by an 80Gb/s connection to the ScienceDMZ.

C. DaLI Ecosystem

DaLI includes a large selection of software modules and custom-built data lifecycle management applications such as X-ray Reconstruction of Moving Morphology (XROMM) Data Management (XDM)[2], MyTardis[3] and Globus[4] services to manage data movement to/from instruments and DaLI instrument as shown in *Figure 2*. DaLI integrates its custom Data Management Software to directly link research instruments to HPC resources, allowing a streamlined process and fast data analysis. Supported projects have developed interactive user interfaces and pipelines that connect the data source to the underlying DaLI HPC resources. Through its portals, such as the XDM, DaLI offers a smooth process for pre and post analysis in an integrated environment. The use of microservices framework enables a scalable and distributed system which allows continuous deployment and delivery of the requested services within DaLI. It also provides modular turnkey solutions for portions of the workflow and the flexibility to use HPC resources in running jobs using data collected from the instruments and outside the pipelines.

The inbuilt fine-grained access control mechanism ensures file-level data security and integrity throughout its lifecycle from acquisition to analysis, storage, sharing, and publication.

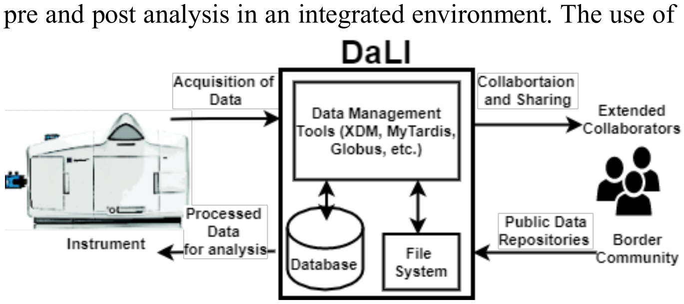

Figure 2: Overview of DaLI to facilitate Data Lifecycle Management

III. RESULTS AND DISCUSSION

We have designed and deployed a scalable platform that helps to capture, store, and share scientific research data along with associated metadata as an integral part of the scientific data lifecycle pipeline. DaLI provides a replicable model to a broad range of disciplines addressing the complexities inherent in managing the full data lifecycle and connecting instruments to computational resources in a seamless scientific workflow. DaLI has enabled more than 20 research projects including 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. There are 609 users and more than 1 PB of data being stored on DaLI space. Data on DaLI is meant to be shared through its data hub and is currently able to mint Data Object Identifiers (DOI). DaLI is also used by a range of classes and workshops to engage students and researchers at UChicago and outreach programs to underprivileged communities. Using microservices APIs, DaLI can enable the integration of data from edge computing resources.

IV. CONCLUSION

We described a data lifecycle instrument that is cost-effective and replicable to address the needs associated with the management and sharing of data across many diverse research domains (e.g. astronomy, ecology, biology, high-energy physics, molecular engineering, linguistics, and neuroscience). DaLI supports pre- and post-processing of data as well as computational and data analysis. The available features of DaLI are facilitating the packaging of data, associated metadata, and artifacts to share and reproduce published results.

Our future work and next steps will focus on developing tools to semi-automatically capture metadata throughout the lifecycle to make reproducible scientific research on HPC systems effortless, to support the reliability and transparency of research findings, and to interoperate with cloud resources. The goal is to avoid waiting to collect data at the end of the research.

ACKNOWLEDGMENT

This research was supported in part by the National Science Foundation (NSF) through grant award 1661523 and resources provided by the RCC at the University of Chicago.

[2] https://xromm.rcc.uchicago.edu:8181/
[3] https://mytardis-uchicago.rcc.uchicago.edu/
[4] https://www.globus.org/

REFERENCES

[1] D. B. Allison et al., Reproducibility of research: Issues and proposed remedies, vol. 115, National Academy of Sciences, 2018, pp. 2561-2562.

[2] A. Brinckman et al., "Computing environments for reproducibility: Capturing the “Whole Tale”," Future Generation Computer Systems, vol. 94, pp. 854-867, 1 5 2019.

[3] H. Meng and D. Thain, "Facilitating the Reproducibility of Scientific Workflows with Execution Environment Specifications," in Procedia Computer Science, 2017.

[4] V. V. Sochat et al., "Enhancing reproducibility in scientific computing: Metrics and registry for Singularity containers," PLoS ONE, vol. 12, no. 11, 1 11 2017.

[5] M. D. Wilkinson et al., "Comment: The FAIR Guiding Principles for scientific data management and stewardship," Scientific Data, vol. 3, 15 3 2016.

[6] J. Freire et al., "Provenance for Computational Tasks: A Survey," 2008.

[7] V. Curcin et al., "Templates as a method for implementing data provenance in decision support systems," Journal of Biomedical Informatics, vol. 65, pp. 1-21, 1 1 2017.

[8] F. Zafar et al., Trustworthy data: A survey, taxonomy and future trends of secure provenance schemes, vol. 94, Academic Press, 2017, pp. 50-68.