Scalable Data Management for National Facilities Using the Modern Research Data Portal

Vas Vasiliadis
Globus, *University of Chicago*
Chicago, U.S.A.
vas@uchicago.edu

Rachana Ananthakrishnan
Globus, *University of Chicago*
Chicago, U.S.A.
ranantha@uchicago.edu

***Abstract*—Managing the growing data volumes generated by national user facilities and shared instruments is becoming untenable without the use of mechanisms that automate the flow of data among resources and investigators. We describe a common use case in user facilities that motivates the need for such mechanisms, and use a data portal, Petrel Data, as an example that illustrates how data management requirements are addressed at the Advanced Photon Source and the Argonne Leadership Computing Facility.**

Keywords—automation, research data management, data portal, scientific instruments, Science DMZ

I. Introduction

High performance computing centers, such as the Leadership Class Facilities at Department of Energy laboratories, and shared instruments such as the Advanced Photon Source (APS) and the Advanced Light Source (ALS), are generating large volumes of data daily. As data volumes grow, the research enterprise is increasingly challenged by what should be mundane tasks: reliably moving data from instruments and computing resources, easily describing data for downstream discovery, and making the data accessible (often with appropriate access controls) to distributed groups of collaborators. The *ad hoc* methods currently employed at many facilities place undue burden on scientists and system administrators alike, and it is clear that some level of automation—with accessible tooling—is required for these tasks.

II. BEYOND SIMPLE WEB APPS

The facilities and instruments described above generate tens of terabytes of data in the form of many millions of files daily. As an example, the X-Ray group beamlines at the APS support two-dimensional and three-dimensional imaging experiments that resulted, on average, in 112 terabytes of data per month during 2016. Instrument upgrades will increase image resolution by 2-3 orders of magnitude in the coming years, with even greater data growth [1].

Tools such as Globus, an established service from the University of Chicago, are widely used for managing research data in national laboratories, campus computing centers, and HPC facilities. The interactive web browser interface provided by such tools addresses simple file transfer and sharing scenarios. Researchers from all backgrounds, and in particular those with limited technical skills, are able to use these simple web apps to effectively manage data by "pointing and clicking".

However, at the scale required by facilities such as the APS, the point-and-click approach becomes untenable, and effective data management requires some form of automation.

III. PORTALS FOR DATA AUTOMATION

A number of facilities have developed custom systems for managing data flows from scientific instruments; excellent examples are the DMagic (https://dmagic.readthedocs.io) and TomoPy (https://tomopy.readthedocs.io) projects in use at the APS. In some instances, these projects have integrated select capabilities—primarily for file transfer and sharing—from the Globus platform, using well-defined REST APIs [2].

Development of more comprehensive, bespoke applications from scratch requires substantial financial investment, which is often out of reach for smaller facilities. It also requires that investigators invest their time in building infrastructure at the expense of conducting research. And, further, it often results in duplicate efforts that recreate existing products.

In response, the Globus team at the University of Chicago created a data portal and science gateway framework that greatly simplifies the process of delivering these capabilities to myriads of researchers with minimal system development and administration effort. It is based on the Modern Research Data Portal [3] design pattern, jointly developed by the ESnet and Globus teams, and leverages capabilities such as the Science DMZ [4] for enhanced performance, and integrates Globus platform services for automated data management at scale.

The framework includes the following components:

A. Authentication and Authorization

A recurring barrier to data access is that of authentication. The portal leverages Globus Auth [5], a foundational service for identity and access management. It enables authentication using existing credentials from over 1,000 federated identity providers, which means that many 100,000s of researchers can simply log in and use the portal without the traditional overhead of account creation, provisioning, and credentials management.

Globus Auth also facilitates implementation of fine-grained authorization policies to control access to data and integrates with Globus Groups to simplify access management for distributed collaborations and larger communities.

B. *Automated Data Ingest*

The portal incorporates tools for automatically retrieving experiment generated data from instrument capture devices and depositing them on a suitable storage system for pre-processing and further analysis. This frees up scarce instrument resources and improves researcher productivity.

C. *Data Description and Discovery*

Metadata and other descriptive attributes may be automatically added to ingested data using the Globus Search service. This service creates indexes which can subsequently be exposed as a faceted search interface within the portal, greatly simplifying data discovery by projects collaborators—and by the broader community, if research products may be published or shared more broadly.

D. *Data and Compute Orchestration*

Integration with the Globus Automate service to enable orchestration of data movement and computation activities, interleaved with manual intervention steps, as required. A researcher can define a flow that automates all steps in the experiment workflow. For example, some APS experiments run thousands of such flows that include data ingest, metadata extraction, pre-processing, image quality review, analysis and reconstruction, indexing, persistent identification, and sharing both intermediate and final data products with collaborators.

IV. Petrel Data

One example, of a portal that leverages these capabilities is Petrel Data (https://petreldata.net) developed by the Argonne Leadership Computing Facility (ALCF) and Globus. It is used by researchers to manage data in diverse fields including materials science, cosmology, machine learning, and serial crystallography.

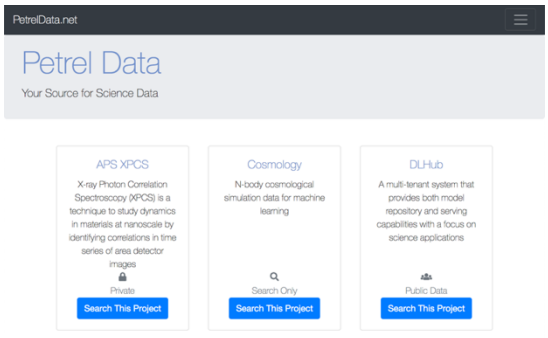

Fig. 1. Petrel Data landing page with links to multiple datasets

The portal facilitates automated ingest of data from APS beamlines and other sources, extraction and addition of metadata for creating search indexes, assignment of persistent identifiers faceted search for rapid data discovery, and point-and-click downloading of datasets by authorized users. As security and privacy are often critical requirements, the portal employs fine-grained permissions that control both visibility of metadata and access to the datasets themselves.

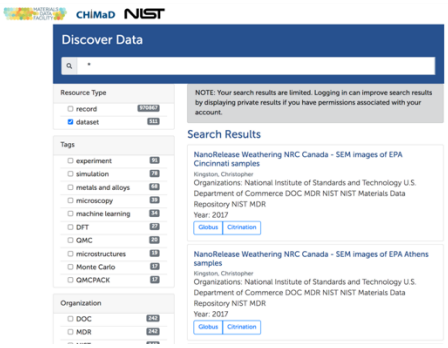

Fig. 2. Materials Data Facility data discovery on the Petrel Data portal

Petrel Data enhances accessibility by aggregating data sets from many different experiments under a single portal. Researchers can log into the portal once, and access one or more data sets if they have the appropriate credentials for each. This ensures a coherent approach to diverse data sharing models, from open access for curated public data sets, to highly restricted individual investigator access for sensitive or protected data.

V. Conclusion

We have described an approach for automating data management for national facilities that incorporates proven systems and approaches, providing a streamlined interface that makes advanced data management capabilities accessible to researchers of all skill levels. The Globus team continues to develop many of the services described and will make additional functionality available to researchers that further streamlines routine tasks as well as more complex research workflows.

References

[1] F. De Carlo, "Globus and X-Ray Imaging at the Advanced Photon Source: From Data Intensive to Data Driven Science," GlobusWorld, April 2016.

[2] Globus API Reference, https://docs.globus.org/api/.

[3] K. Chard, E. Dart, I. Foster, D. Shifflett, S. Tuecke, and J. Williams, "The Modern Research Data Portal: A Design Pattern for Networked, Data Intensive Science," in PeerJ Articles, https://peerj.com/articles/cs-144/, January 2018, https://docs.globus.org/mrdp.

[4] E. Dart, L. Rotman, B. Tierney, M. Hester, and J. Zurawski, "The Science DMZ: A Network Design Pattern for Data Intensive Science," SC '13: Proceedings of the International Conference on High Performance Computing , Networking, Storage Analysis, Denver, CO, 2013, pp. 1-10, doi: 10.1145/2503210.2503245.

[5] S. Tuecke, R. Ananthakrishnan, K. Chard, M. Lidman, B. McCollam, S. Rosen, and I. Foster, "Globus Auth: A Research Identity and Access Management Platform," IEEE 12th International Conference on eScience, 2016.