HPCS-0006

Archive API Data Management Environment

Annotation and registration of large datasets is inherent for managed datasets to effectively deliver broader scientific impact and advance the mission of the NCI. Consistent with efforts already underway at the NIH within the Big Data To Knowledge (BD2K) program, annotation and registration of datasets will enable managed datasets to be of use to the community of extended and future cancer investigators. The creation and delivery of metadata and tracking utilization of datasets will provide the key insight into scientific impact for each maintained dataset.

This service offering is operational and is currently being evaluated for scalability. We are interested in connecting with you for input and feedback upon utilization and evaluation of the system.

*For support inquiries, please contact us at **nci-cbiit-hpc@list.nih.gov**

Table of Contents

HPC DME 1.0.0 Release Notes

Please find release notes via HPC_DME_General_Training (1).docx

Workflow Diagrams

- **From concept to proposed object archive**

blocked URL

- **HPC Data Management Workflow with CleverSafe**

blocked URL

Overview

One of the most significant challenges to overcome for an effective high performance computing (HPC) support effort is effective data management, i.e., effective tracking, annotation and staging of digital datasets, accompanied with a data life cycle plan/policy for these datasets. While frequently not considered an HPC challenge or opportunity, **an effective solution is needed to contain costs for stored data while increasing the scientific usefulness of data that has been created** in the era of 'big data' where analysis of datasets can take days and total cost to store and maintain large datasets continue to tax personnel and financial resources. Without a reliable managed dataset solution, large datasets are frequently maintained in multiple copies across the physical storage in an isolated fashion, leading to an unnecessary expense as additional storage is required for analysis and storage of new data. **A managed, secured, and high-availability solution will minimize the need for maintaining unnecessarily redundant copies of large datasets**. Even with projected declines in the cost of physical storage, the investment in managing stored data without associated annotation will provide only minimal (at best) long-term scientific usefulness or support to advance the mission of the NCI.

Without an effective data management solution, the HPC effort will struggle with difficulties in staging data for analysis, recovering generated datasets, and inefficiencies created by insufficient physical storage and recomputing results that have once been completed. Therefore, we believe that:

- NCI is in critical need of advancing its core scientific and technological means of data management and services from large, diverse, distributed and heterogeneous datasets
- Large datasets are currently maintained in multiple copies across physical storage in an isolated fashion, leading to an unnecessary expense
- Annotation and registration of datasets is inherent for managed datasets to effectively deliver broader scientific impact and enable the full power of personalized medicine
- Strategically, the absence of an effective data management solution presents a barrier to supporting emerging efforts to leverage the breadth of generated datasets for use in development of computationally and data intensive predictive models as well as efforts to utilize cloud resources for collaboration and analysis.

Vision

- By introducing the core HPC DME APIs, related batch utility toolsets, and Web GUI applications and features, data collected at NCI will help achieve strategic goals.
- By creating and maintaining an object archive repository and associated metadata catalog, value added services or predictive modeling will be tapped and utilized for potential new therapeutic interventions or other competitive advantages
- Improve and enhance community involvement and knowledge sharing via extending the HPC DME services, integrating with an existing Cloud service or exposing through a Cloud based host.

Goals

- Provide core capabilities to get started , but extensible to accommodate future need
- Implement/enhance HPC data management core APIs based on iRODS iCAT and Jargon core APIs – The unified REST APIs will be agnostic of physical storage medium or device being utilized
- Implement HPC DM command and batch utilities with no interference of currently SF adopted Globus workflow
- A highly flexible and reliable storage model for underlying collections and data objects
- Data virtualization with multiple storage types
- Configurable metadata policies for validation
- Secure APIs enforcing authentication and authorization
- Data discovery through descriptive metadata
- Data sharing through REST push and pull methods
- Integration with iRODS, Cleversafe and Globus Transfer API
- Command and batch utilities to register collections, data objects and update permissions
- Configurable security layer extending iRODS security implementation
- Access to Globus, iRODS and Cleversafe through system service account
- Pluggable data transfer implementations (S3, Globus, iRODS)
- Assess technology to focus on adding value for identified use cases
- Re-use and enhance vs invent new tools to address SF needs
- Enhance, extend and collaborate as engagement with other groups goes on.

Objectives

CBIIT, with input from the Office in Scientific Operations (OSO) in Frederick, has identified a requirement to alter the operations and scope of support that are currently offered to NCI for IT services and Informatics support from The National Cancer informatics Program. They recognize the need to: i) establish direct relationships with the Divisions and Centers throughout NCI, ii) gain the trust of those constituents, and iii) re-architect the scientific computing and enterprise IT infrastructure, as well as the IT support services and informatics support offered by CBIIT and the Frederick National Laboratory to better match the requirements of NCI.

It is critical that the prototyped APIs and web front end address the prioritized use cases to meet the user workflow needs.

- Establish GridFTP services to support general transfer of large datasets without requiring physical mounting.
- Establish a pilot dataset registration system to associate a label with a given managed dataset. The System will capture extensible metadata including security and access requirements for the managed dataset. Metadata includes but not limited to: How the dataset was generated, when it was generated, where it was generated, present access method and information to obtain a full copy of the original dataset. System will also be flexible to support either a) export of metadata to a future system; b) development of service APIs to support interrogation by secondary systems; or c) both (a) and (b).
- Establish easy-to-use methods for providing annotation information and creating reports of managed datasets, including utilization of a modest controlled vocabulary related to high-use criteria for searching datasets.
- Establish a high-reliability storage model for underlying datasets included into the dataset registration system
- Obtain utilization statistics for managed datasets.
- Provide required system administration support for data management services
- Provide development and general fixes to existing implementation
- Design and introduce Web-based interfaces for submitting, retrieving and locating data by metadata characteristics.
- Design and implement enhancement features per prioritizations
- Pilot and implement DevOps and required operational modifications to support policies around HPC data management services environment
- Extend implementation to support placement of data objects in cloud (based on requirements to be provided)
- Implement and support shared development and collaborating with NIH/CIT and others in updating and extending the code-base
- Produce updated assessment of comparable/compatible technologies.
- Participate in the evaluation of the CGC pilots, focusing on analyzing and testing of the APIs and web front ends ability to handle workflow needs in handling of big data.

Project Organization

blocked URL

The project organizational structure for the NCI HPC Data Management Environment effort, shown in Figure 1 below, is designed to facilitate collaborative management of the interdependent cross-agency business and technical activities needed to complete this project.

NCI CBIIT, as owner of the High Performance Computing Data Management Environment initiative for which APIs, tools, solutions and related operational support will be targeted and implemented, will collaborate and partner with the Leidos Biomedical Research Inc (LBR) HPC team and its subcontractor in collecting business driven use cases, alignment with NIH BD2K advancement, and agile development and implementation of the object archive with associated infrastructure and capital support.

LBR, specifically DSITP, will be responsible for project management, procuring and managing necessary subcontractor or consultant resource in providing needed expertise, technical review, quality verification and confirmation of the APIs, toolsets or developed features in supporting core use case scenarios by working closely with critical business partners or groups, procurement of necessary storage hardware and software, and coordinating with NCI/NIH CIT group in piloting, collaborating and adopting HPC DME APIs, toolsets, code-base, or solutions.