Frederick National Laboratory
for Cancer Research

NCI HPC Data Management Environment (DME)

Aug 3, 2017

DEPARTMENT OF HEALTH AND HUMAN SERVICES • National Institutes of Health • National Cancer Institute
Frederick National Laboratory is a federally funded research and development center operated by Leidos Biomedical Research, Inc., for the National Cancer Institute.

Agenda

- Introduction
- Objectives
- Solution Overview
- API and Deployment Models
- Demo
- Questions

Introduction

- NCI
  - Carl McCabe

    Government Sponsor

    Chief, Informatics and Scientific Computing Services Branch US National Cancer Institute's Center for Bioinformatics & Information Technology

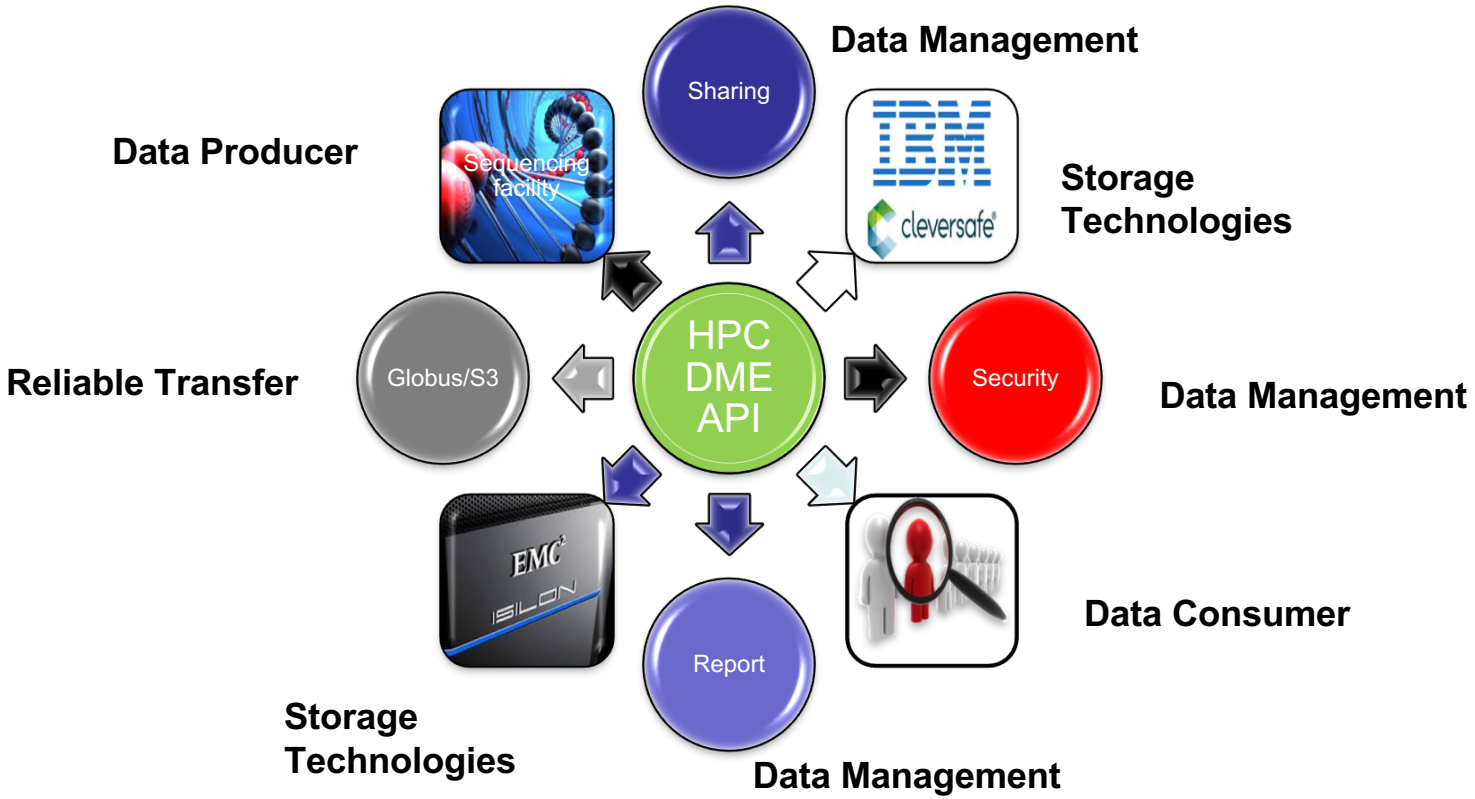
    Phone: 240-276-7366

    Email: Carl.McCabe@nih.gov

Introduction

- Leidos Biomedical Research, Inc. (LBR)
  - Eric A Stahlberg – LBR ISP HPC Strategy Director
    Information System Program Directorate
    High-Performance Computing Strategy
    Phone: 240.276.6729
    stahlbergea@mail.nih.gov
  - Zhengwu Lu – LBR TPM
    Technical Project Manager, Information System Program Directorate
    Phone: 240-276-6487
    Zhengwu.Lu@nih.gov
  - George Zaki – Business Analyst
    High Performance Computing Analyst, Data Science and Information Technology Program
    Phone: 240-276-5171
    george.zaki@nih.gov
- Prasad Konka – Enterprise Architect
  President, SVG, Inc
  Phone: 240-276-5328
  Prasad.konka@nih.gov

Overview (Producer – Consumer)



Objectives / Benefits

- Establish new **highly-reliable** service to support management of mission critical and archival data
- Improve ease of use near-term and long-term
  - Provide stable means to deposit, access and recall managed data from the investigator and analyst perspective
  - Support mechanisms to annotate datasets
  - Technology agnostic interface for users of system
  - Establish services to support the general transfer of large datasets without requiring physical mounting
- Derive greater value from datasets
  - Make metadata searchable
  - Enforce user defined policies
  - Data ownership by end users

Objectives / Benefits

- Be cost effective
  - Reduce unneeded copies of large datasets
  - Enable migration/staging of data to appropriate storage based on utilization
  - Promote centralized core services, tools, documentation and training
  - Adopt open-source based solutions
- Reliable Model
  - Adopt existing established reliable storage model
  - Obtain utilization statistics to assess needs for resource optimization
  - Secure Access
  - Customizable notifications for interested parties

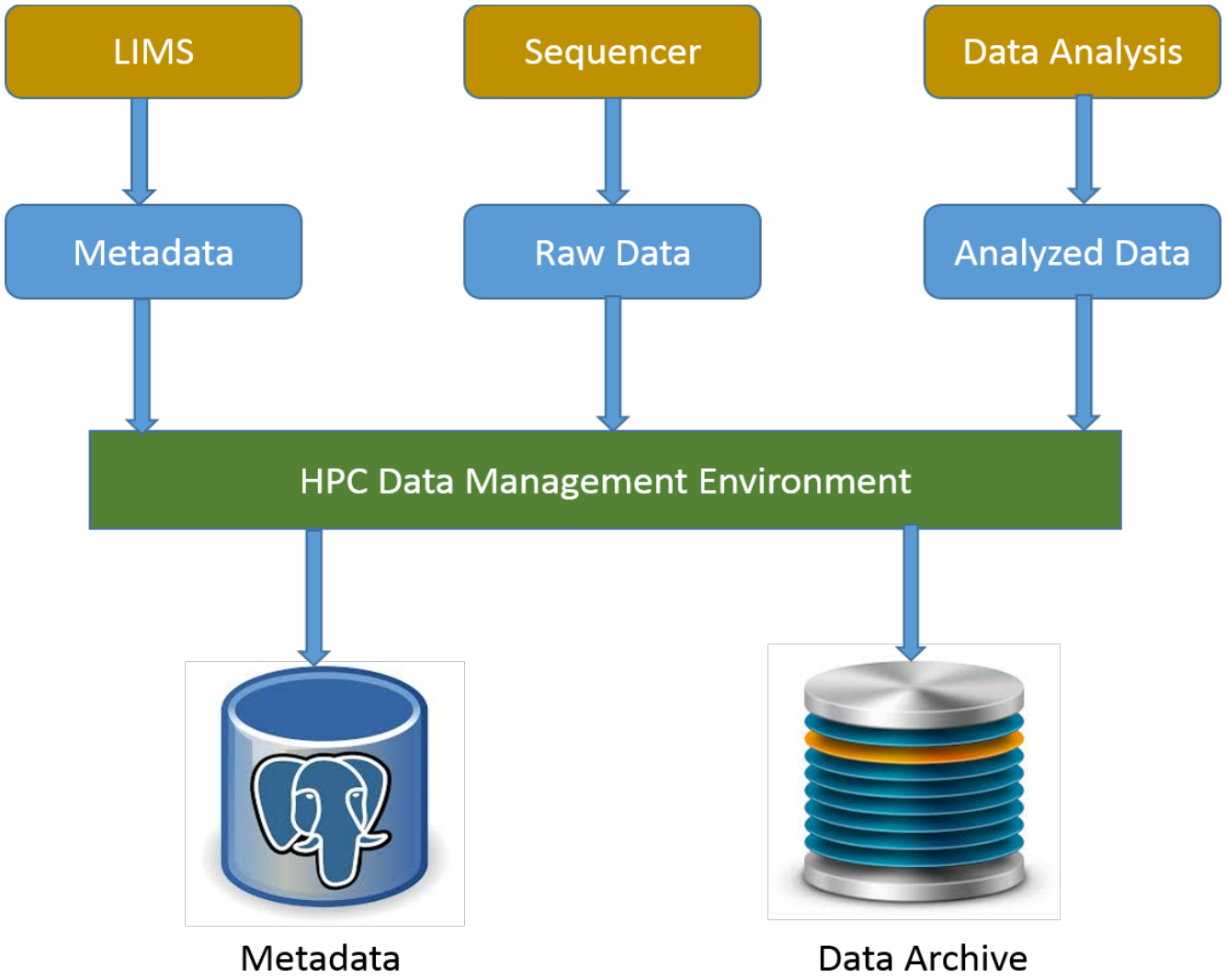
Objectives / Benefits

- Data security/sharing
  - Extending iRODS security model
  - Users, Groups, Role
  - Permissions
- Web UI (in UAT, planned production in May)
  - Search
  - Sharing
  - Upload/Download
  - Update
  - Admin
  - Subscribe to interested tasks/events

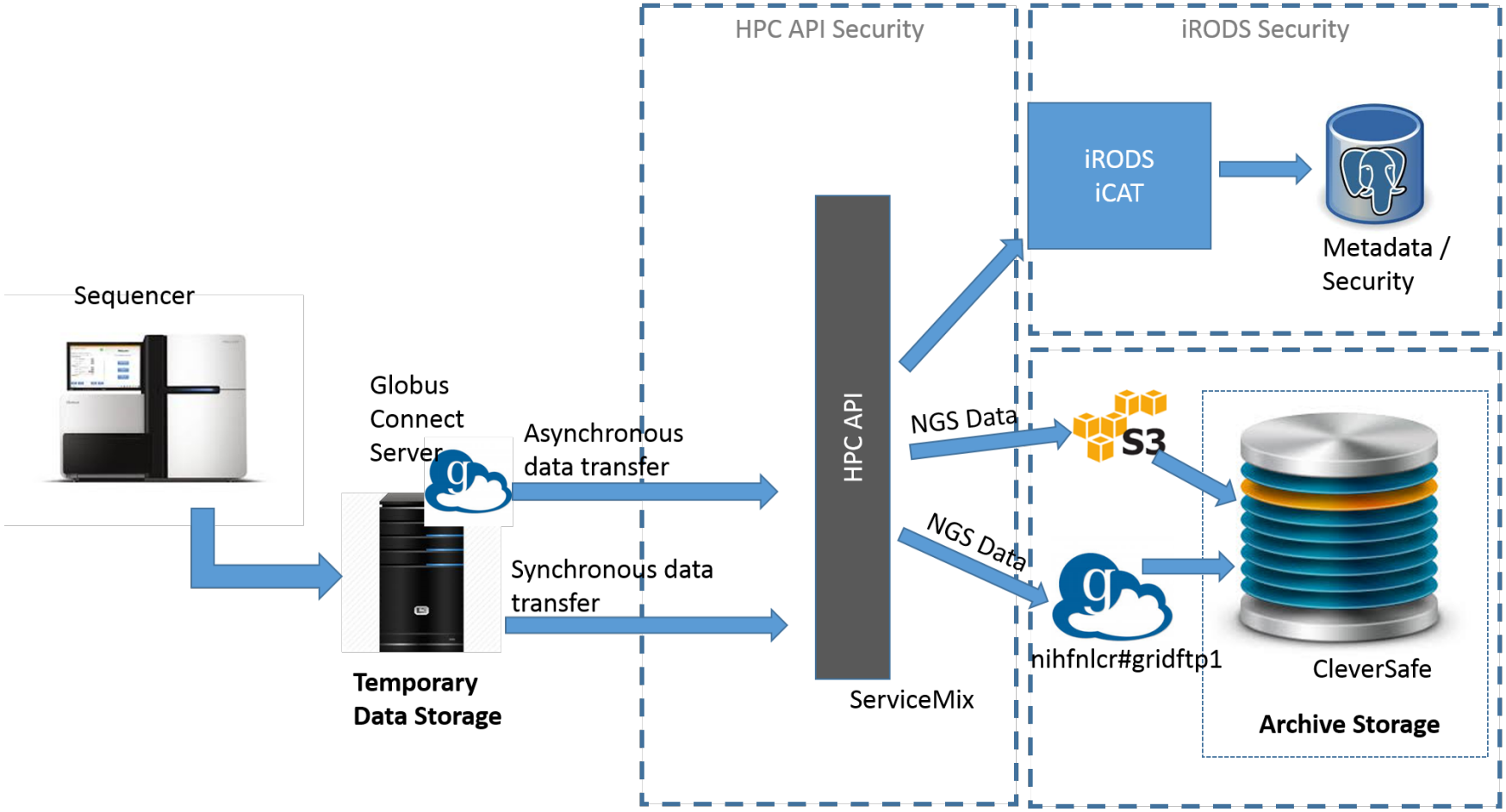
Objectives / Benefits

- Built on top of open source Data Management middleware - **iRODS**
  - NIH, NASA, NOA, Sanger Institute (100 million files, 20 PBS, 15000 users)
  - Data Virtualization
  - Workflow Automation
  - Heterogeneous storage types

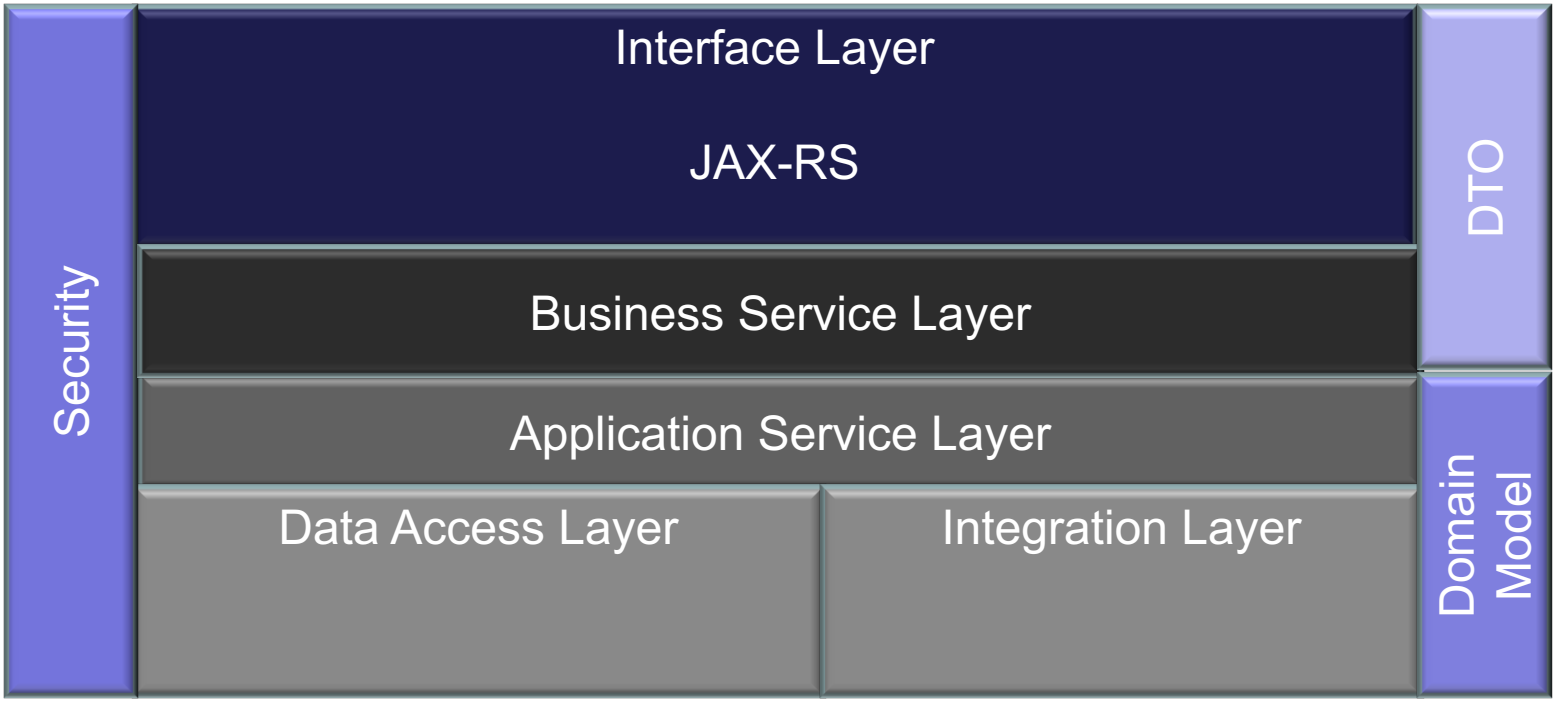
Use case example:



Deployment Model



API Model


Security
Interface Layer
JAX-RS
Business Service Layer
Application Service Layer
Data Access Layer
Integration Layer
DTO
Domain Model

How can an end user provision storage for their usage needs?

- Default Storage is CleverSafe at ITOG
  - 1 Production Vault with 200 TB
  - DOC Based Vault (coming up)
  - Supports S3 based storage devices
- Ability to support heterogenous storage types
  - Data Virtualization through unified namespace
  - Data on different storage devices in different locations can be centrally managed
  - Migration, backup is transparent for end users
  - Workflow automation (validation, replication, backup, etc)
- End users provision their storage needs through system admin/storage groups.

Using your interface, how does the end user store data in the archive, including the metadata?

- Collections, Data files
  - Standardized REST API
  - cmd line utility (Bulk Registration)
  - Bash Script utilities
  - Web UI (coming up)
- Metadata
  - Attribute Name, Value pair in JSON format
  - Policy driven validations
- Data Transfer
  - Synchronous
  - Asynchronous (Globus)
- Demo

How does and end user get a report of a subset or all of their files in the archive?

- Reporting
  - Summarized report
  - Summarized report by dates
  - DOC report
  - DOC report by dates
  - User report
  - User report by Dates
  - Metadata based reporting (coming up)
  - Email notification of summarized reports
- Demo

How can an end user remove data and its metadata from the archive?

- Write once and Read many
- Data security is top priority
- Add metadata to indicate data is obsolete
- Policy to purge obsolete data

How can an end user script a change for a subset of their metadata – perhaps the expiration date, WORM, or access control?

- Collections, Data files
  - Standardized REST API
  - cmd line utility
  - Bash Script utility
  - Web UI
- Demo

How can an end user search through all of the metadata – perhaps for all files of a certain type?

- Collections, Data files
  - Standardized REST API
  - cmd line utility
  - Bash Script utility
  - Web UI
- Demo

How does one set the access controls for data to be shared or not?

- Tiered security levels
- Authentication with NIH AD
- Authorization at API (configurable)
- Authorization through iRODS security model
- Service Account interacts with storage systems

How does one set the access controls for data to be shared or not?

- Access types
  - READ, WRITE, OWN, NONE
- Roles
  - SYSTEM_ADMIN
  - GROUP_ADMIN
  - USER
- Permissions are inherited through the logical hierarchy
- Group, User based permissions
- Data registrar owns the collection / data file
- Demo

Demonstrate a script that will automate moving data from one source to Cleversafe and vice versa?

- Synchronous and Asynchronous
- Web UI or Script
- Demo

How does your solution control the acceptable terms to be entered into the metadata database?

- Policy Driven metadata validations
  - DOC based
  - Required attributes
  - Acceptable values
  - Default value
  - Data Hierarchy
  - System generated attributes

Notifications

- Users can subscribe to notifications
  - Email
  - Text (In future)
- Notifications are generated by user actions
  - File registration status
  - File download status
  - Collection updates
  - Summarized report

Questions

- Contact info:
  - hpc_dme_admin@nih.gov