Data Management Environment at the National Cancer Institute

iRODS User Group Meeting, July 6, 2022

Sunita Menon
Eran Rosenberg
Yuri Dinh
Zhengwu Lu
Prasad Konka
George Zaki
Udit Sehgal
Sarada Chintala
Ruth Frost
Eric Stahlberg

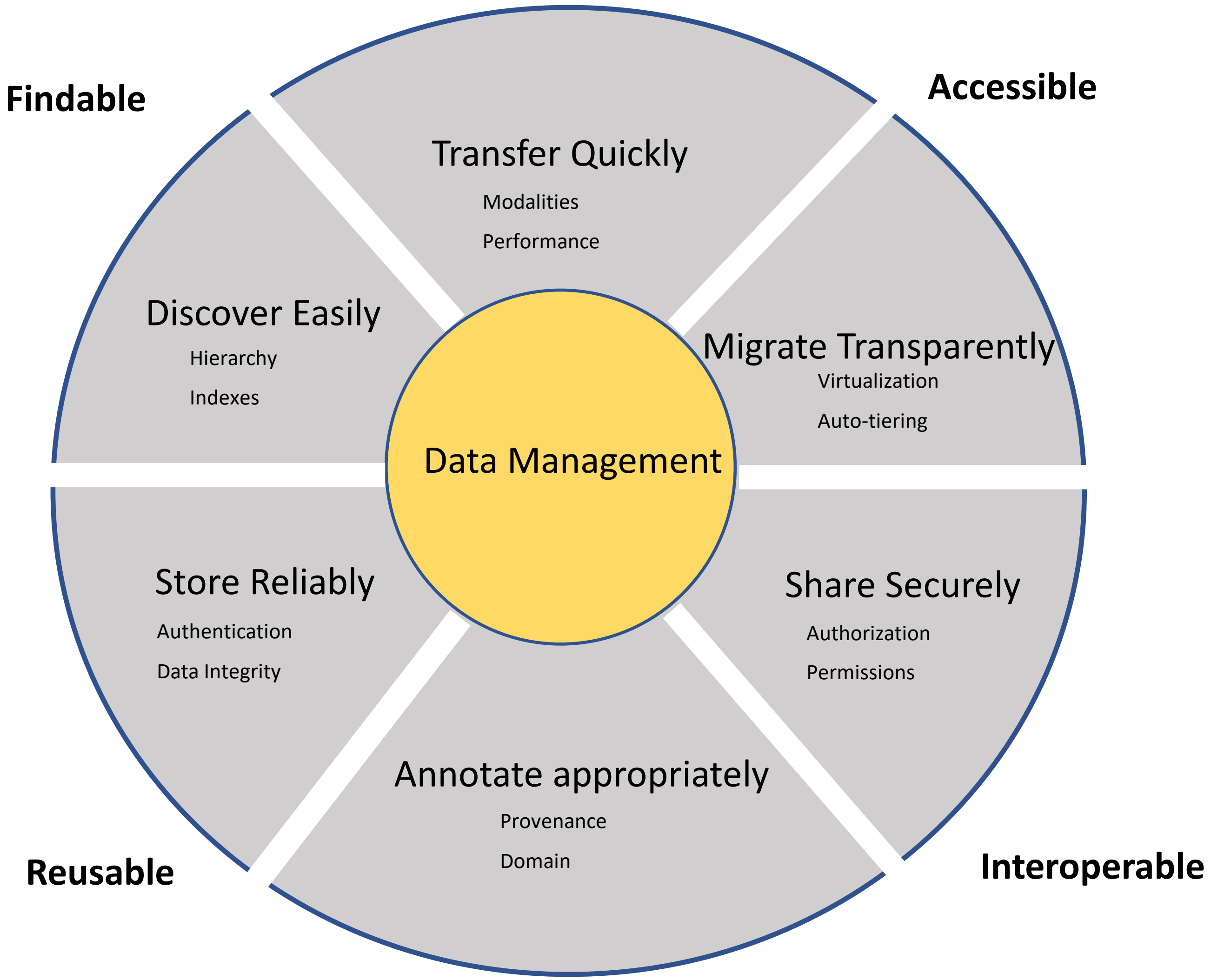
Frederick
National Laboratory
for Cancer Research

Agenda

- Background
- DME Overview
- Architecture
- Data Management
- Permissions Management
- Storage virtualization
- Data Migration and Tiering
- Conclusions
- Q/A

Data Management challenges

- Redundant copies of data in multiple locations
- Missing or insufficient provenance information
- Data integration and data sharing barriers



Data Management System Requirements

DME Overview

- Platform to archive high value, scientific datasets.
- Enables data management, data discovery and data sharing
- Ability to associate user metadata with archived data at any stage in the data lifecycle
- Storage virtualization abstracts the storage technology and storage location

DME Overview

- Piloted in 2017 with Next Generation Sequencing data from the Center for Cancer Research Sequencing Facility
- Production level infrastructure established in 2018
- Automated archival workflow commissioned in 2019
- 23 research research labs and cores leveraging DME as of today
- Over 4 PB of data secured so far

DME Overview

Interfaces – REST API suite, DME web application and command line utilities (CLU)

Transfer endpoints – AWS S3, Globus, Google Cloud, Google Drive, User’s File System

Data Stores – Cleversafe, Cloudian, Amazon S3, Glacier Deep Archive, File System Storage

Logical Architecture

- Modular Layered architecture implemented with Spring Framework
  - REST API invokes underlying business services
    - Business services orchestrate application services to deliver the requested functionality
      - Integration services interface with external subsystems for authentication, data management, and data transfer
      - Data access objects interface with DME tables and materialized views

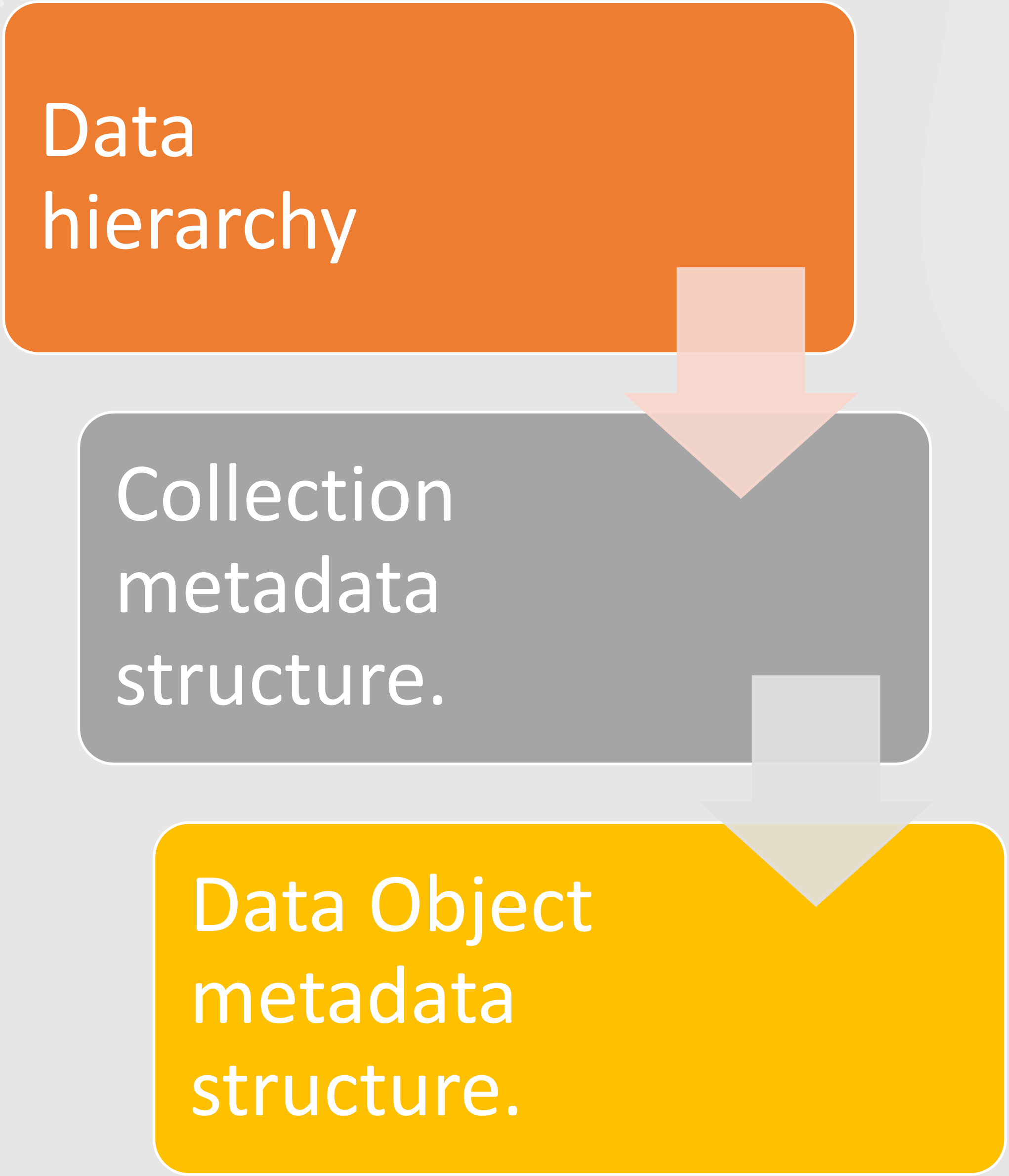
Physical Architecture

- CentOS 7 Linux physical servers and virtual machines
- DME web application hosted on Tomcat 8
- DME API servers on Apache ServiceMix
- iRODS 4.2.9 metadata on Oracle 19c database
- On-premises Cleversafe and Cloudian vaults
- Amazon S3 and S3 Glacier Deep Archive

Data Management with iRODS

- Storage virtualization
- Metadata based data discovery
- Secure collaboration

Data Management with iRODS



Data hierarchy

Collection metadata structure.

Data Object metadata structure.

Data Management with iRODS

System metadata and User metadata

- System metadata is captured automatically when an object is created. Cannot be changed by users.
- User metadata is provided by the user and consists of Provenance and Domain metadata.

Data Management with iRODS

Mandatory or optional user metadata

- Mandatory metadata is validated by the system during data registration.
- Optional metadata can be supplied during registration or anytime later.

Permissions Management

OWN - Data Generator, Data Owner, Lab Manager

WRITE - Researchers

READ - Collaborators

Permissions Management

System Administrator

- DME administrators - data hierarchy setup, data migration and tiering, data management support, Transfer monitoring

Group Administrator

- Lab managers, data generators, bioinformatics analysts - data archival, user management, permissions management

User

- Researchers using the data, collaborators - browse, search and retrieve

Storage Virtualization

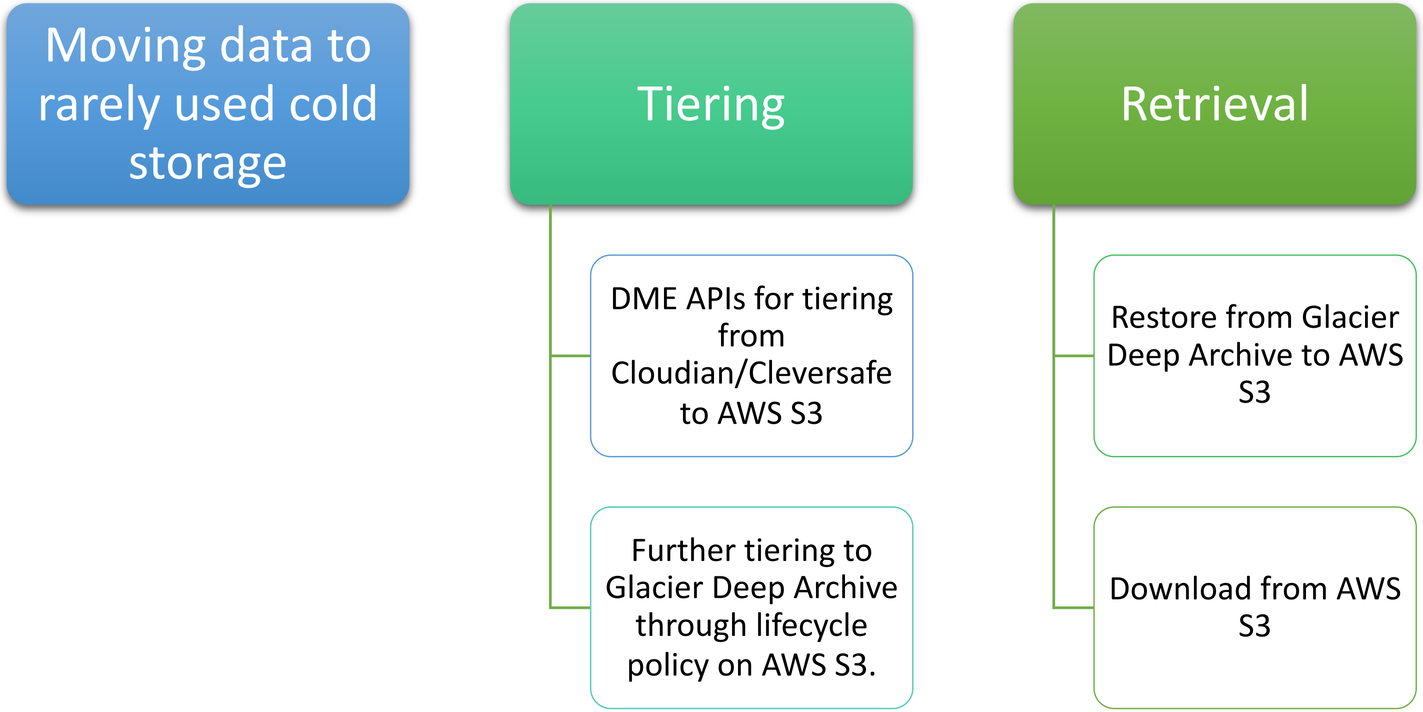
Users view the data through the hierarchy they have defined

- Location and organization of data is transparent to the user.
- Transparent switchover of storage providers (e.g. Cleversafe to Cloudian)

Data Migration

- Required due to end of life or end of support
- New REST API developed to facilitate migration from one S3 storage provider to another
- Transparent to the user – does not impact how they retrieve

Data Tiering



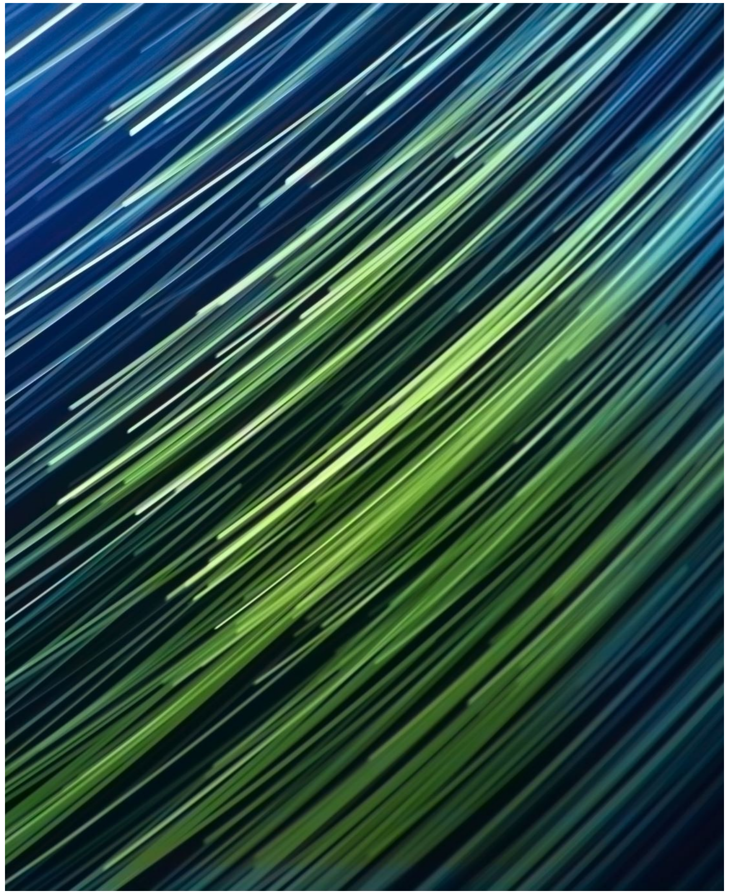
Moving data to rarely used cold storage

Tiering

- DME APIs for tiering from Cloudian/Cleversafe to AWS S3
- Further tiering to Glacier Deep Archive through lifecycle policy on AWS S3.

Retrieval

- Restore from Glacier Deep Archive to AWS S3
- Download from AWS S3

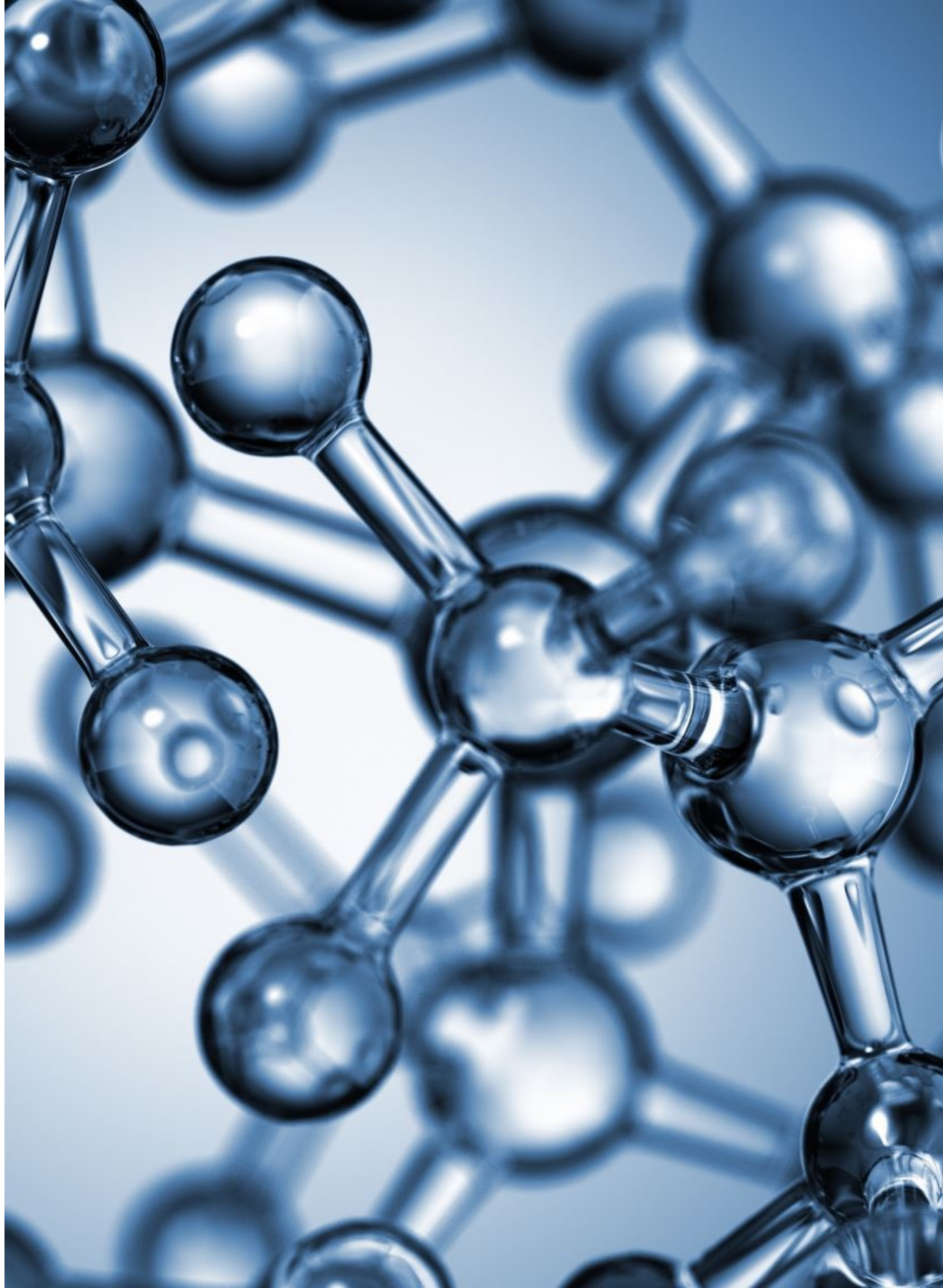


Conclusions

INFRASTRUCTURE SCALE-UP

NEW CAPABILITIES

INTEGRATION WITH ANALYSIS PLATFORMS



Data Management Environment at NCI

Thank You