Agenda

- Introduction
- Challenges
- Data Transfer Solution
- iRODS use in Data Transfer Solution
- iRODS Proof-of-Concept
- Q & A
**Introduction**

- **Genentech / Roche**
  - Biotech Company
  - Fortune’s “100 Best Companies to Work For” List
- **Integration Services**
  - Application Integration
  - Partner Integration
  - Data Integration
- **Data Virtualization**
  - Enterprise Information Integration
Challenges

The some of challenges faced by business with respect to data movement are:

- Bottlenecks in Hardware infrastructure and Network
- Data Transfer is too slow
- No Automated or Scheduled transfers
- No user-friendly GUI
- Custom developed scripts for every type of data transfer job
- Manually executing data transfer jobs
- Lack of visibility and traceability of data transfer jobs
- No Metadata managed related to transfer process
Data Transfer Solution

Data Transfer Platform system designed to support and manage high speed transfer of scientific data that includes capabilities such as:

- Optimized high-speed protocols
- API driven interface to monitor and manage transfers
- Metadata management related to transfer process
- Ability to automate the transfers
- Post-transfer workflows
- Store, search, and manage data and transfer metadata in the data management system
- Implement solution for first use case - data replication.
Data Transfer Solution includes multiple components:

- Hardware
- Infrastructure Management
- Software
  - File Transfer Solution
  - *Data Management (iRODS)*
  - Pipeline Management
- User Interfaces
- Security
iRODS use in Data Transfer Solution

● iRODS as Change Log

● iRODS File System Scanner capability is used to scan the mount path of file system to ingest the system metadata

● To provide the list of all new, updated and deleted files to support for the data replication capability

● iRODS - Data management system can be used to track file lifecycle and provenance
Scientific Data Archive and Replication

Business requirements to support for Disaster recovery and high availability:

- High Performance Transfer
- Storage agnostic solution
- Scalability to support large number of files
- Detecting the changes in the file system
- Preserving Unix, Windows permission and timestamp for file creation and modification
Replication Solution Options

Sync Tool

- Primary Site
- Alternate Site
- Replicate using High Performance Transfer Protocol

Replication

- Primary Site
- Alternate Site
- Replicate using TCP

iRODS
Replication Solution Options

1. Ingest to iRODS catalog
2. Query iRODS catalog
3. Initiate Replication

Primary Site

Alternate Site

Replicate using High Performance Transfer Protocol
Replication using Data Transfer Solution

Web UI → Python Flask API → Jenkins Pipeline

- iRODS Rule for delete detection and generate manifest file
- Perform deletes in destination end-point and iRODS
- iRODS Rule for new/updated detection & generate manifest file
- Perform sync for new and updated files

Scheduling / Queuing Service

Primary Site: iRODS Consumer Server → Secondary Site: iRODS Consumer Server

Sync using High Transfer Protocol

Storage Mount
iRODS Architecture in Data Transfer Solution

- **Head Node**
- **Primary Site**
  - Node 1
  - Node 2
- **Secondary Site**
  - Node 3
  - Node 4

- **iRODS Catalog Server**
- **iRODS Consumer Server**

**iRODS Zone**

**Storage Mount**
Ingest Metadata using iRODS File System Scanner

Data Transfer Solution

Server Configuration

Rulebase configuration in server_config.json

Rulebase

Register NewFiles
Update ModifiedFiles
Remove DeletedFiles
Unregister DeletedFiles

Shell Script

initial_reg_sync.sh detectadded.sh detectmodified.sh detectdeleted.sh

META_DATA_ATTR_NAME = filesystem::mtime
META_DATA_ATTR_NAME = filesystem::deleted
META_DATA_ATTR_VALUE = 2018-06-05 13:02:11.914472000
META_DATA_ATTR_VALUE = Y
Ingestion using iRODS in DTP

- As part of the data transfer in DTP, iRODS will be used for the data management component to track file lifecycle and provenance.

- For the Data Replication use case, iRODS will be used to provide the system metadata of the storage that includes:
  - New files added since last ingest of metadata
  - Updated files since last ingest of metadata
  - Deletes files since last ingest of metadata

- The system metadata can be queried using iRODS CLI or Python iRODS Client
Next Step - iRODS Automated Ingest Framework

- We are planning to implement this new framework for ingest of new and updated files metadata

- It is required sync wrapper and some additional changes for our use case

- This framework will help to simplify ingestion of metadata and also improves the performance
PoC - Data Catalog using iRODS

- Enable simplicity of access with one namespace and want to make data locality transparent to the user
- Ability to search and access to data and metadata
PoC - Data Catalog using iRODS

Use Cases
- Search Across Repositories
- Automation - Data Transformation
- Automation - Data Move
- File Tracking
- Notification
- Reports
- Workspace setup

Solution
- Search frontend
- Search Engine
- Trigger ETL - Rule
- Audit / Lineage - Rule
- Send Email - Rule
- Build Reports - Rule
- Setup Project - Rule
- REST API
- Catalog Server(1)
- Auto tag - Rule
- File event - Rule

Integrated Rule Oriented Data System IRODS 4.2.2

IRODS Capabilities
- Unified Access via Virtualization
- Workflows and Automation
- Metadata catalog and data discoveries
- Secure collaboration

Solution: Link Business knowledge(1) with Data (2)
PoC - Data Catalog using iRODS

<table>
<thead>
<tr>
<th>Data Name</th>
<th>Data Type</th>
<th>Instrument Type</th>
<th>Study</th>
<th>Instrument ID</th>
<th>Organ</th>
<th>Study Title</th>
<th>File</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zuzia.xml</td>
<td></td>
<td>6800</td>
<td></td>
<td></td>
<td></td>
<td>A Prospective Study to Evaluate the Prevalence Zika Virus in Blood Donations Using the cobas® Zika Test for use on the cobas® 6800/8800 System Roche Molecular Systems, Inc., in Pleasanton, California 94588</td>
<td></td>
</tr>
</tbody>
</table>
PoC - Enable Intentional Archive

Users searches through metadata of the storage, folder, files level to set the metadata (e.g. ARCHIVE to Yes) to trigger the storage tiering automatically.
PoC - Enable Intentional Archive

- To enable self-service for users to set the flag at folder or file level and then iRODS will automatically apply the tiering storage for the set flag files or folders
PoC - Enable Intentional Archive

- After the metadata is set to trigger the tiered storage framework, the file moved from Tier 1 to Tier 2 (AWS S3) automatically.

- When the file is accessed / read, the file will be moved automatically from Tier 2 (AWS S3) to Tier 1.
Thanks! Questions?