
iRODS for Data Management and Archiving

UGM 2018

Masilamani Subramanyam



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

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

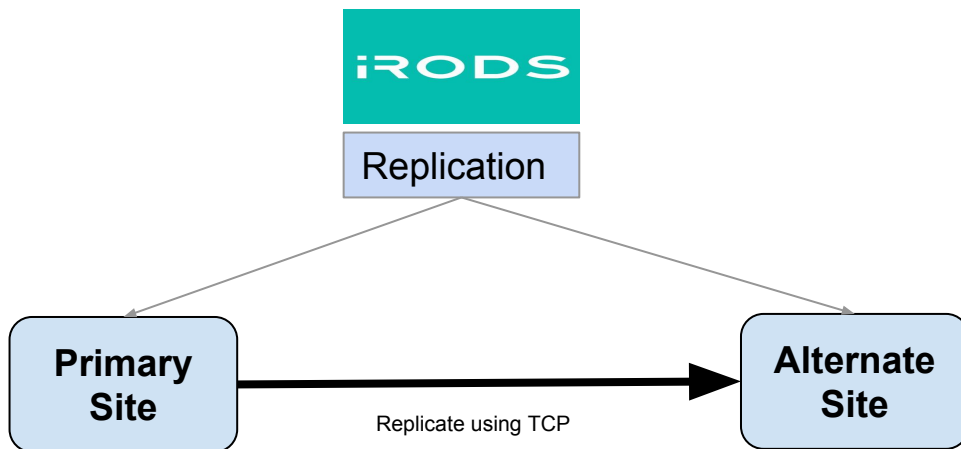
- iRODS as Change Log
- iRODS File System Scanner capability is used to scan the mount path of file system to ingest the system metadata
- To provide the list of all new, updated and deleted files to support for the data replication capability
- iRODS - Data management system can be used to track file lifecycle and provenance

Scientific Data Archive and Replication

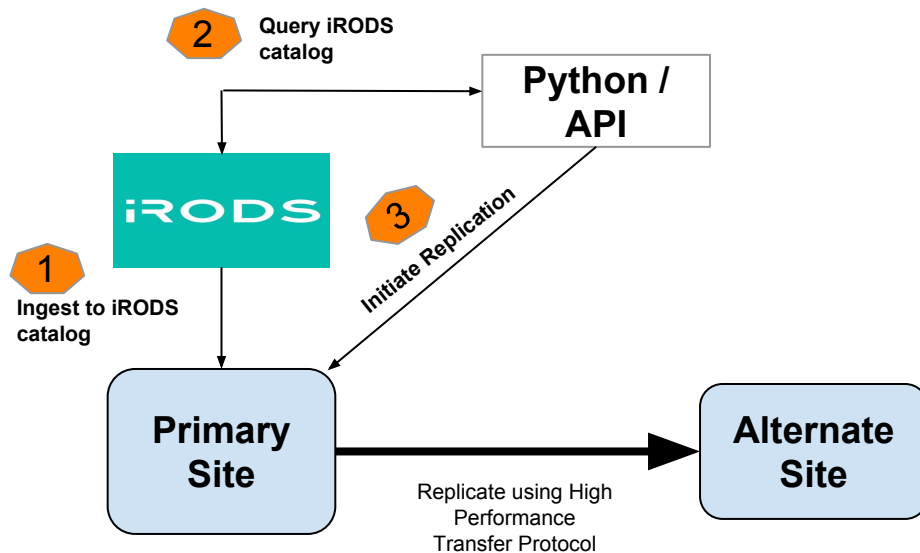
Business requirements to support for Disaster recovery and high availability:

- 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

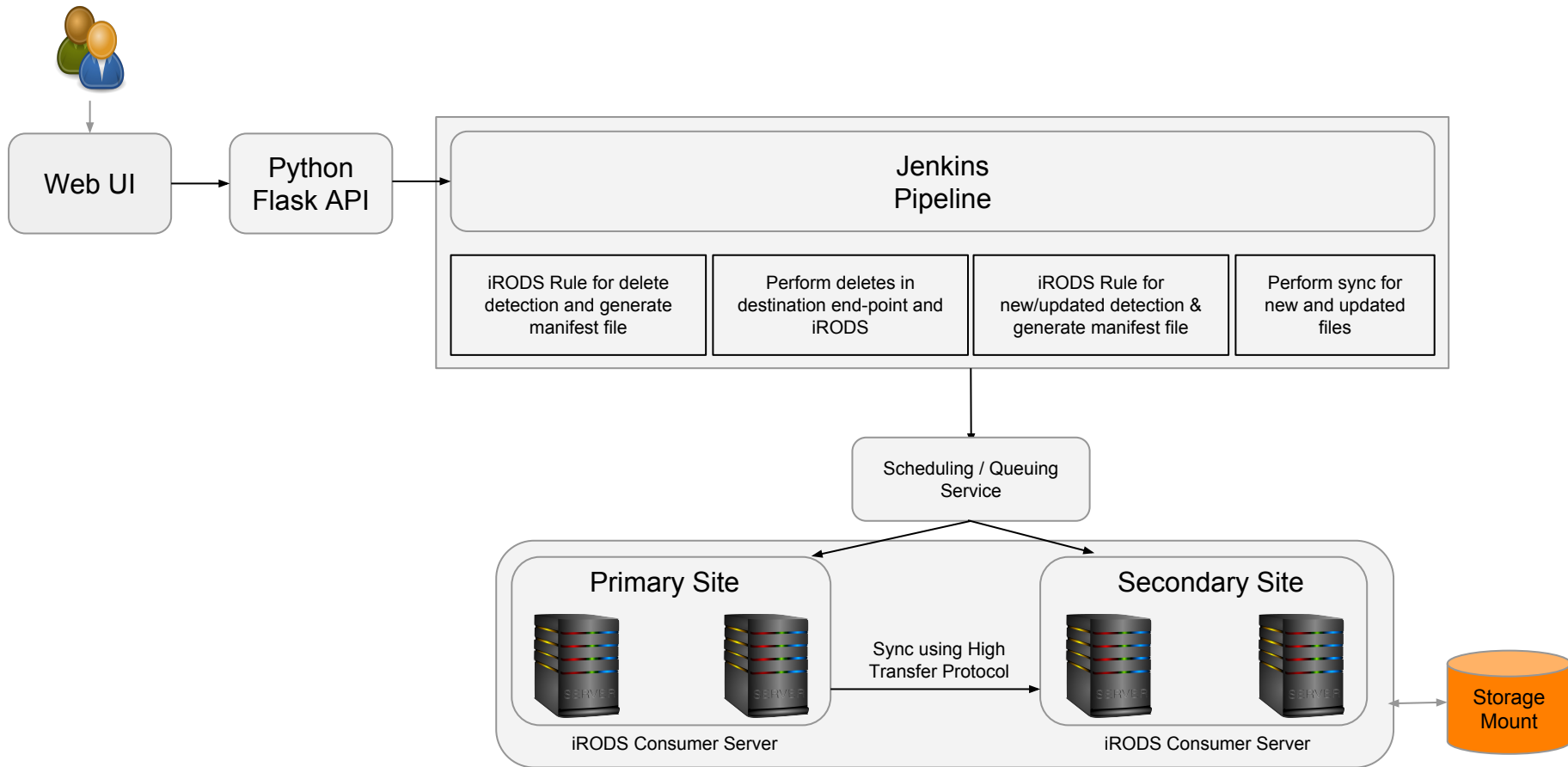
Replication Solution Options



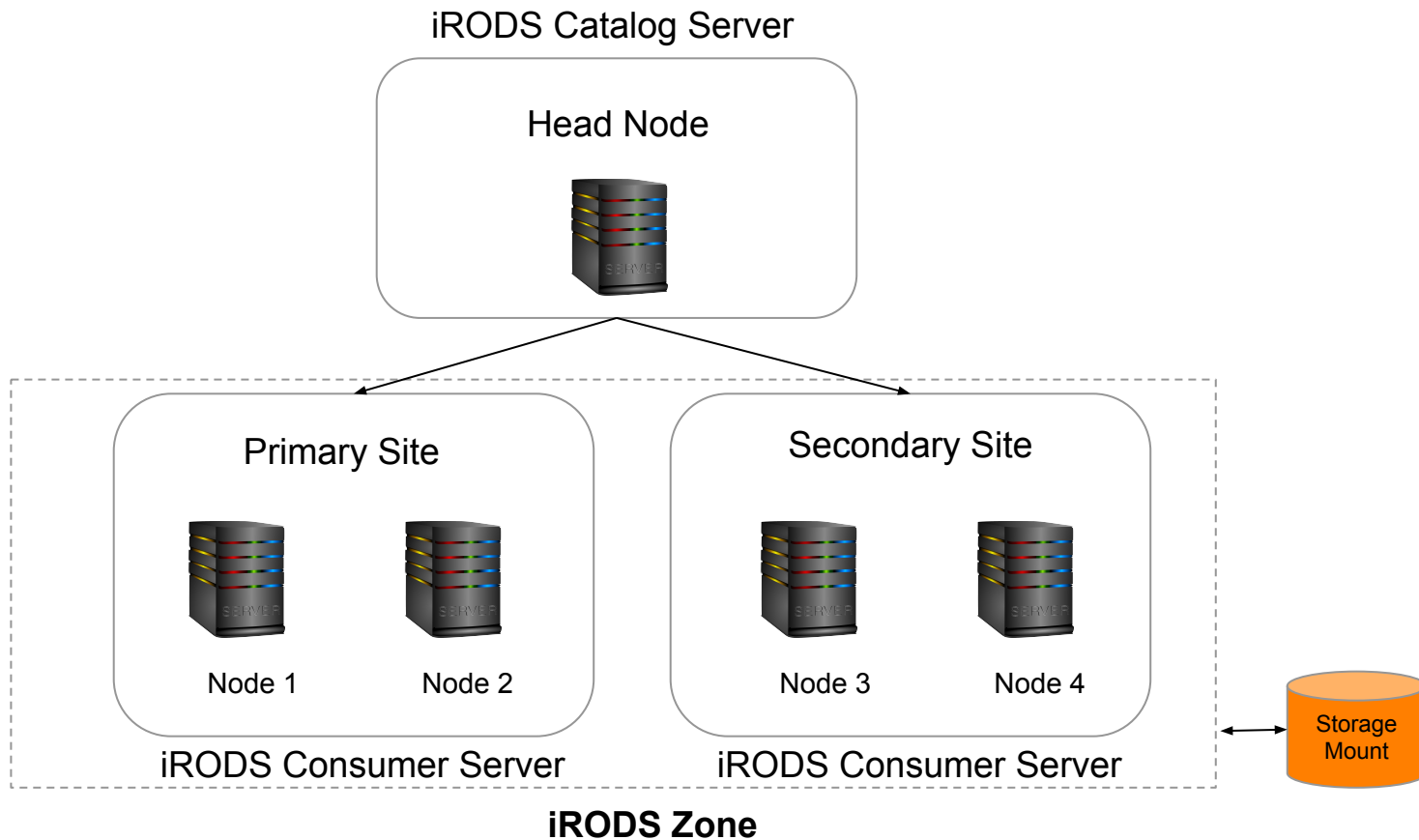
Replication Solution Options



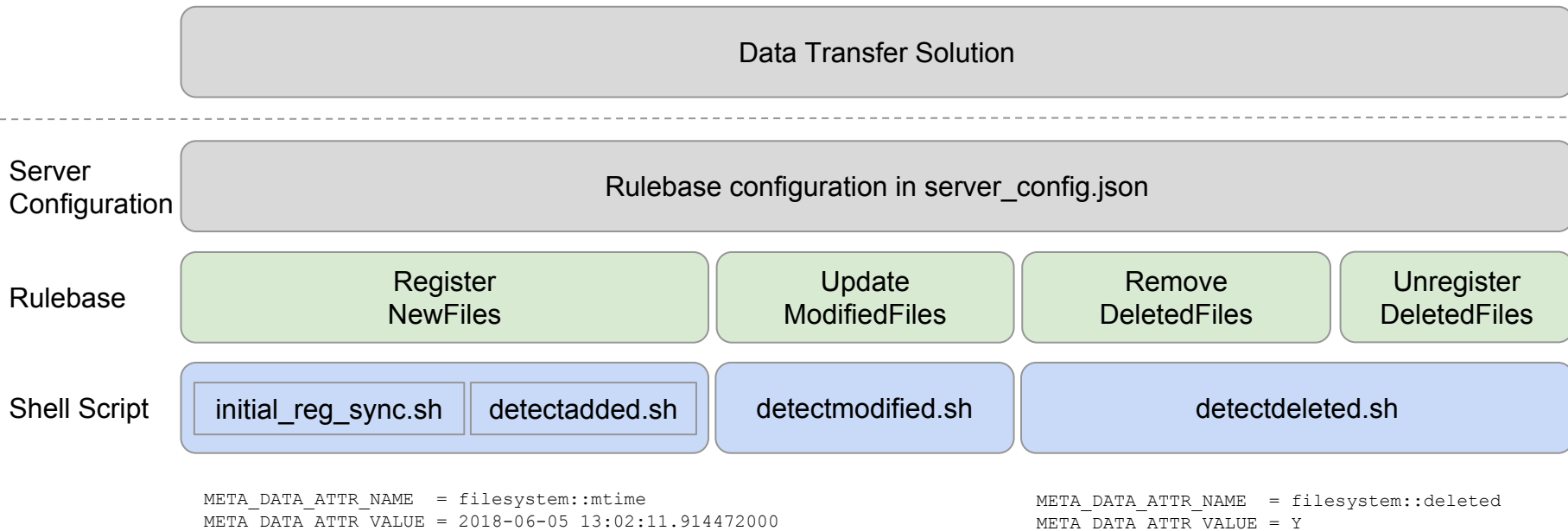
Replication using Data Transfer Solution



iRODS Architecture in Data Transfer Solution



Ingest Metadata using iRODS File System Scanner

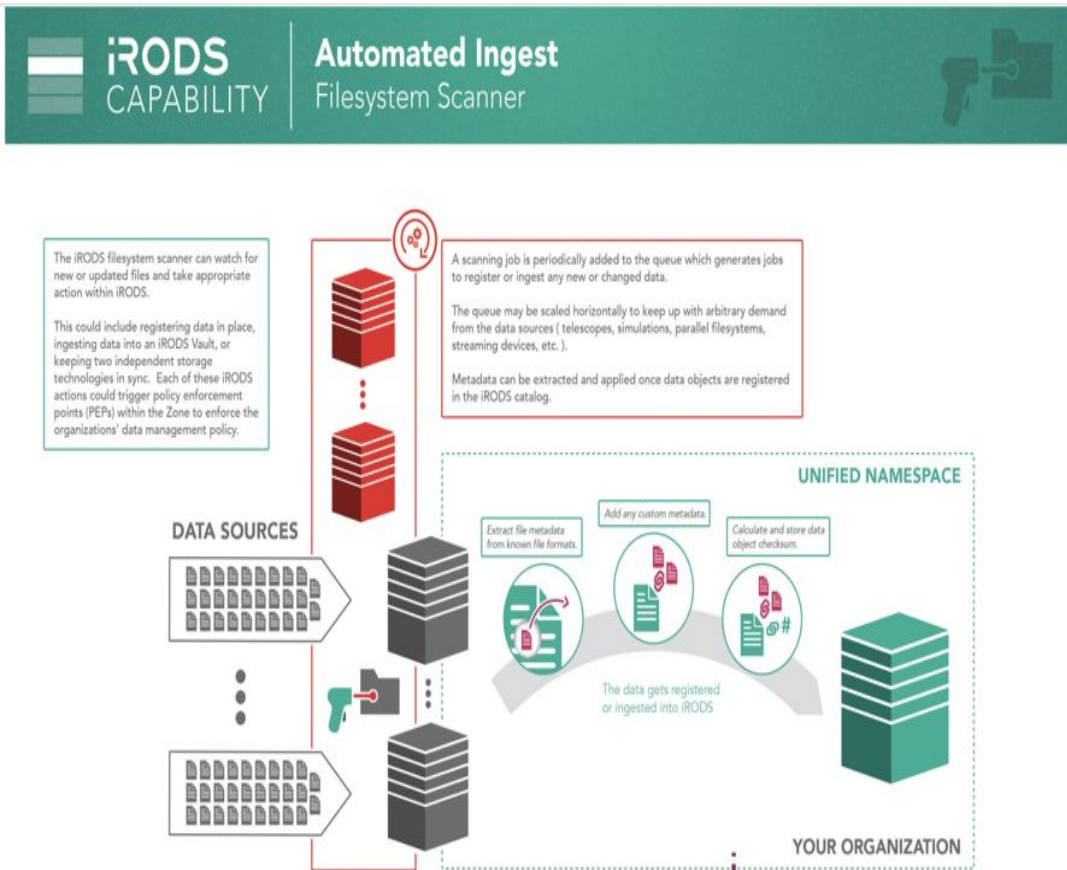


Ingestion using iRODS in DTP

- 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

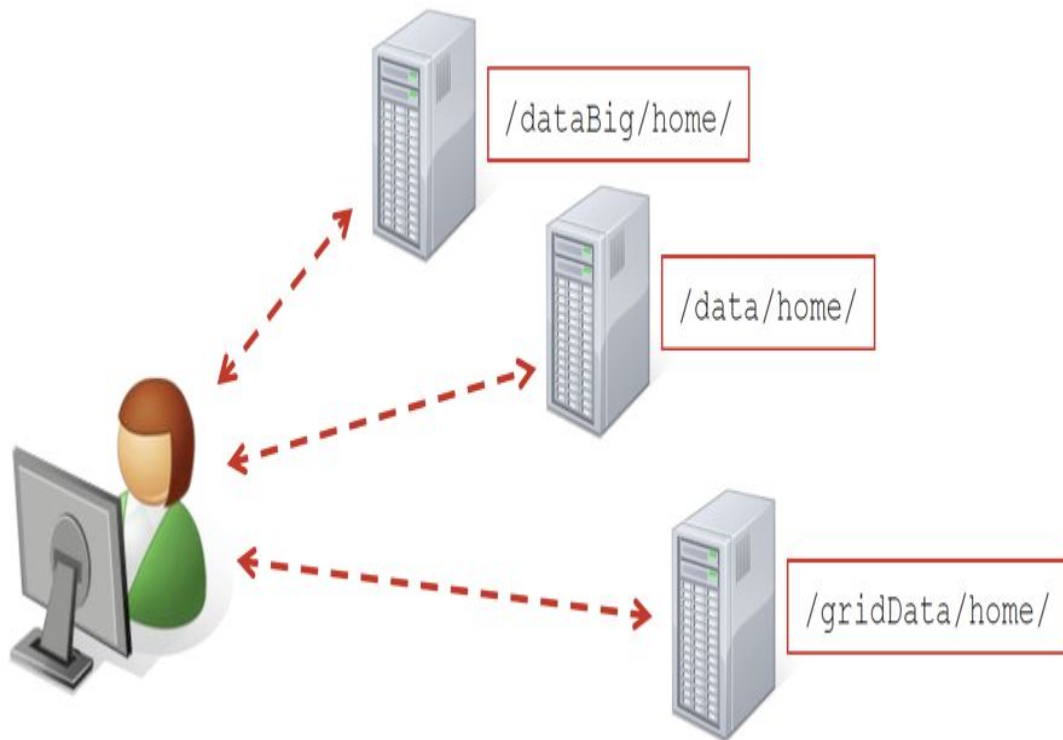
Next Step - iRODS Automated Ingest Framework

- 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

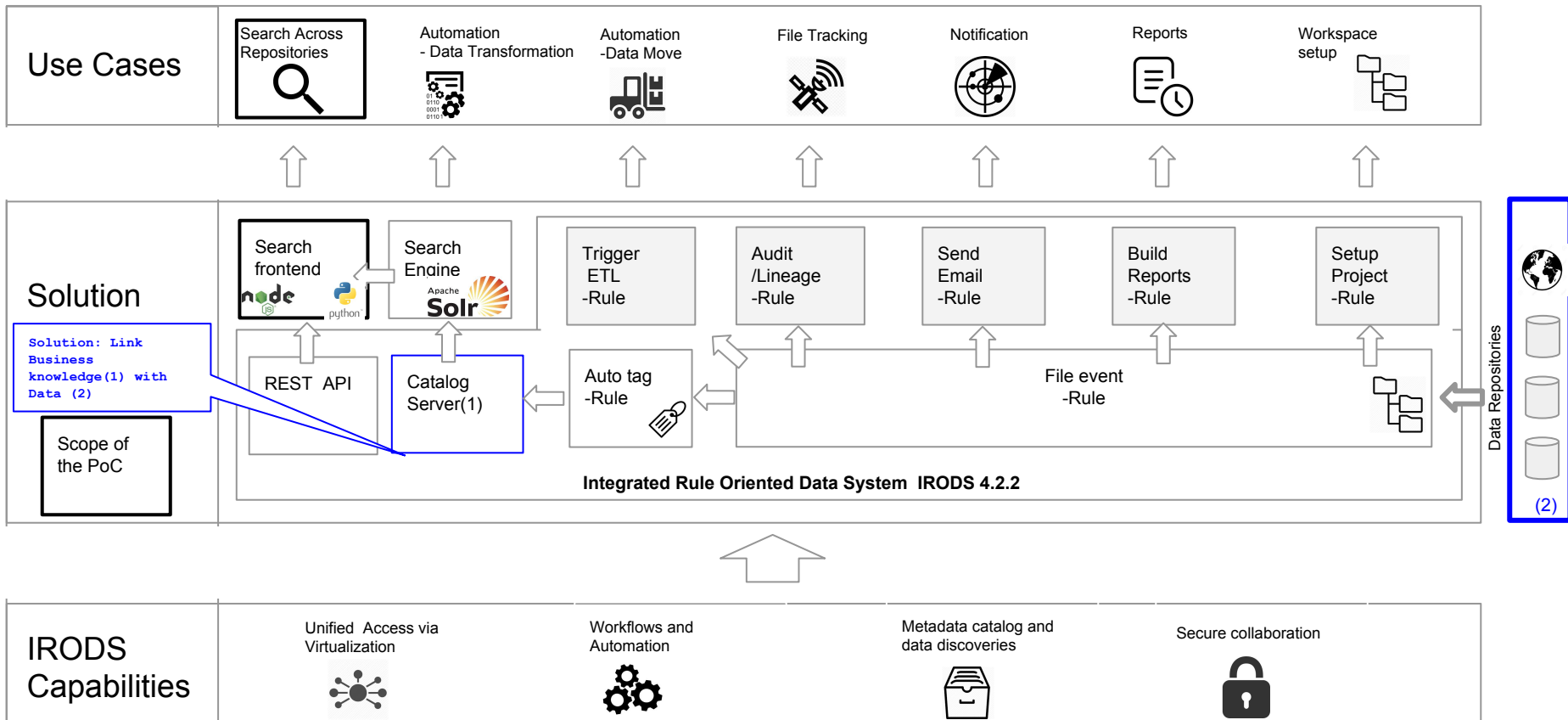


PoC - Data Catalog using iRODS

- 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



PoC - Data Catalog using iRODS

Q

virus

Rows ▾

Select columns ▾

Download CSV

LOG

Data Name ▾	Data Type ▲	Instrument Type	Study ▲	Instrument ID ▲	Organization	Study Title	File
Zuzia.xml	<div><div>Select All</div><div><div><input checked="" type="checkbox"/> Instrument</div><div><input checked="" type="checkbox"/> CRF</div></div></div>	6800	<div><div>Select All</div><div><div><input checked="" type="checkbox"/> HPV301</div><div><input checked="" type="checkbox"/> HPV301_1</div><div><input checked="" type="checkbox"/> HPV411</div><div><input checked="" type="checkbox"/> HPV420</div><div><input checked="" type="checkbox"/> HPV435</div><div><input checked="" type="checkbox"/> ZIK41</div><div><input checked="" type="checkbox"/> ZIK427</div><div><input checked="" type="checkbox"/> ZIK429</div><div><input checked="" type="checkbox"/> ZIKA27</div><div><input checked="" type="checkbox"/> ZIKA428</div></div></div>				

Found 1 results

☒ Data Name

☒ Data Type

☒ Instrument Type

☒ Study

☒ Instrument ID

☒ Organization

☒ Team

☐ Coll Name

☐ Data Path

☐ Life Cycle

☐ Owner

☐ Pattern

☐ Comment

☒ Study Title

☐ Last Updated

☐ Updated By

☒ Download

PoC - Enable Intentional Archive



self-service

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

iRODS Storage Tier Framework

2

A: isilon_to_object_storage_tier_group
V: 0
U:

A: irods::storage_tier_time
V: 60
U:

A: irods::storage_tier_verification
V: catalog
U:

A: irods::storage_tier_query
V: .. META_DATA_ATTR_NAME =
'ARCHIVE' AND
META_DATA_ATTR_VALUE =
'Y'



Tier 0
(FAST)



compound resource



Cache



AWS S3

Tier 1
(INTERMEDIATE)

Tier Group

1

2

A: isilon_to_object_storage_tier_group
V: 1
U:

A: irods::storage_tier_verification
V: catalog
U:

iRODS Zone


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

The screenshot displays the Dell EMC metalnx web interface. A modal dialog titled "Add Metadata" is open in the center. It contains three input fields: "Attribute" with the text "ARCHIVE", "Value" with the text "Y", and "Unit" with the text "New Unit". At the bottom of the dialog are "Cancel" and "Save changes" buttons. In the background, the "Collections" page is visible, showing a breadcrumb path "... / home / rods / foo.txt" and a table with metadata for "irods::access_time". A red arrow points to the "ARCHIVE" text in the dialog, another red arrow points to the "Y" value, and a third red arrow points to the "+ Metadata" button in the top right of the background interface.

PoC - Enable Intentional Archive

```
-bash-4.2$ imeta ls -R fastResc  
  
AVUs defined for resource fastResc:  
attribute: irods::storage_tier_query  
value: SELECT DATA_NAME, COLL_NAME WHERE DATA_RESC_ID IN ('123527') AND META_DATA_ATTR_NAME = 'ARCHIVE' AND META_DATA_ATTR_VALUE = 'Y'  
units:  
----  
attribute: irods::storage_tier_group  
value: example_group  
units: 0  
----  
attribute: irods::storage_tier_verification  
value: catalog  
units:  
----  
attribute: irods::storage_tier_time  
value: 60  
units:
```



- 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?