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

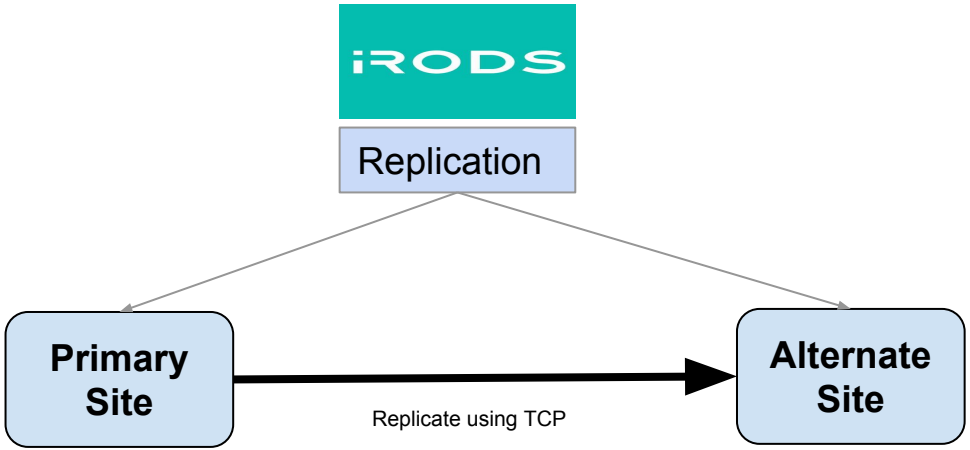
- iRODS as Change Log
- iRODS File System Scanner capability is used to scan the mount path of file system to ingest the system metadata
- To provide the list of all new, updated and deleted files to support for the data replication capability
- iRODS - Data management system can be used to track file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

Scientific Data Archive and Replication

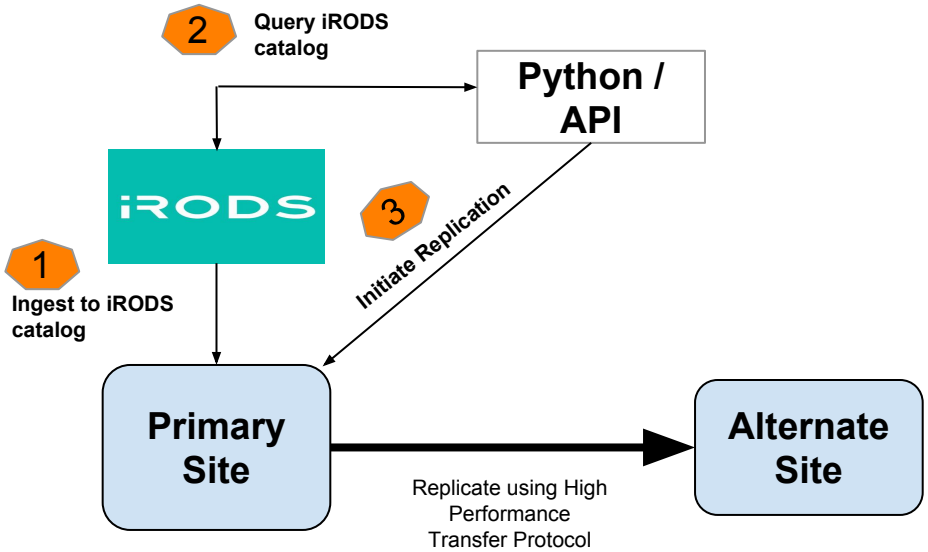
Business requirements to support for Disaster recovery and high availability:

- 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

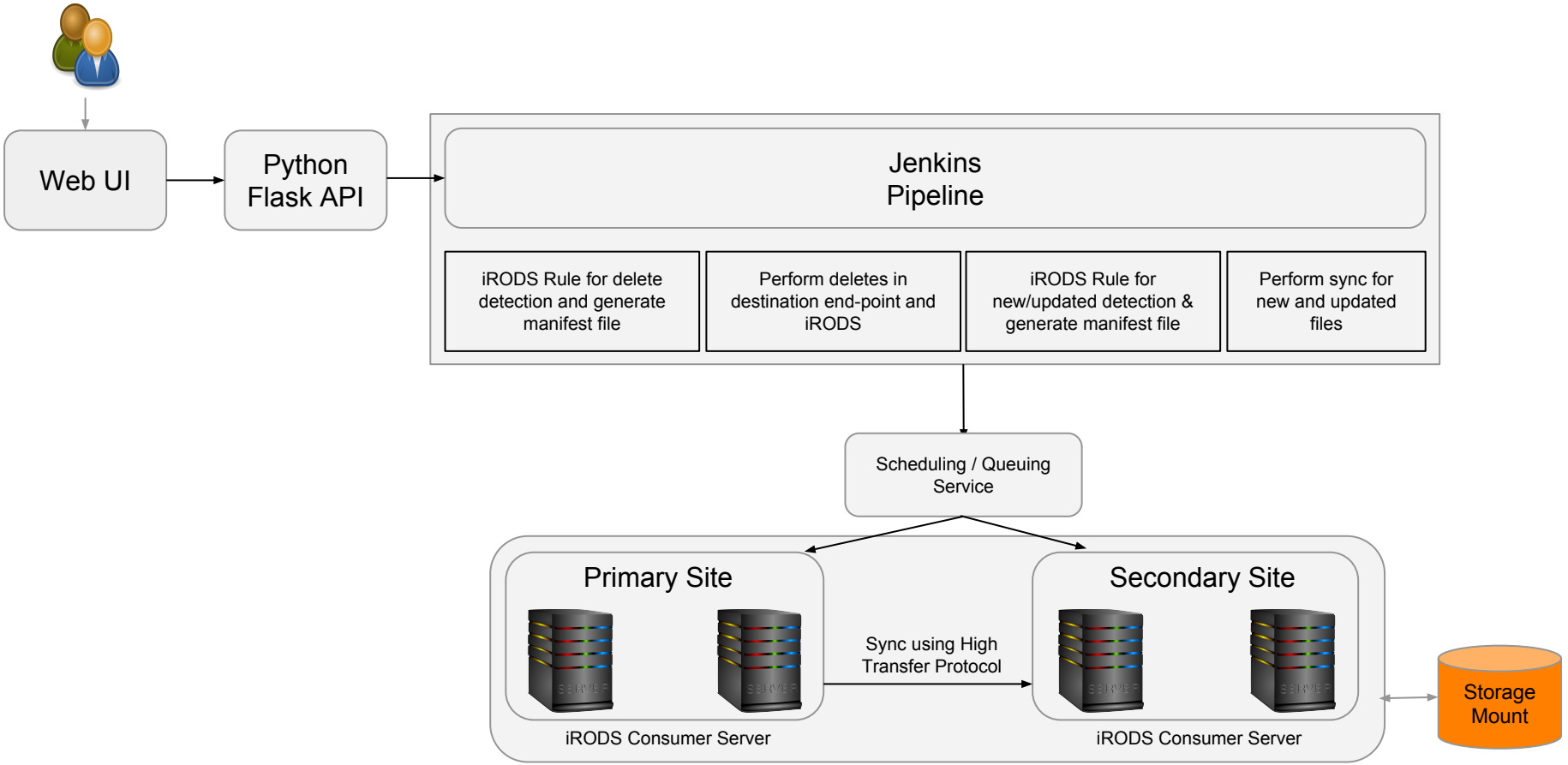
Replication Solution Options



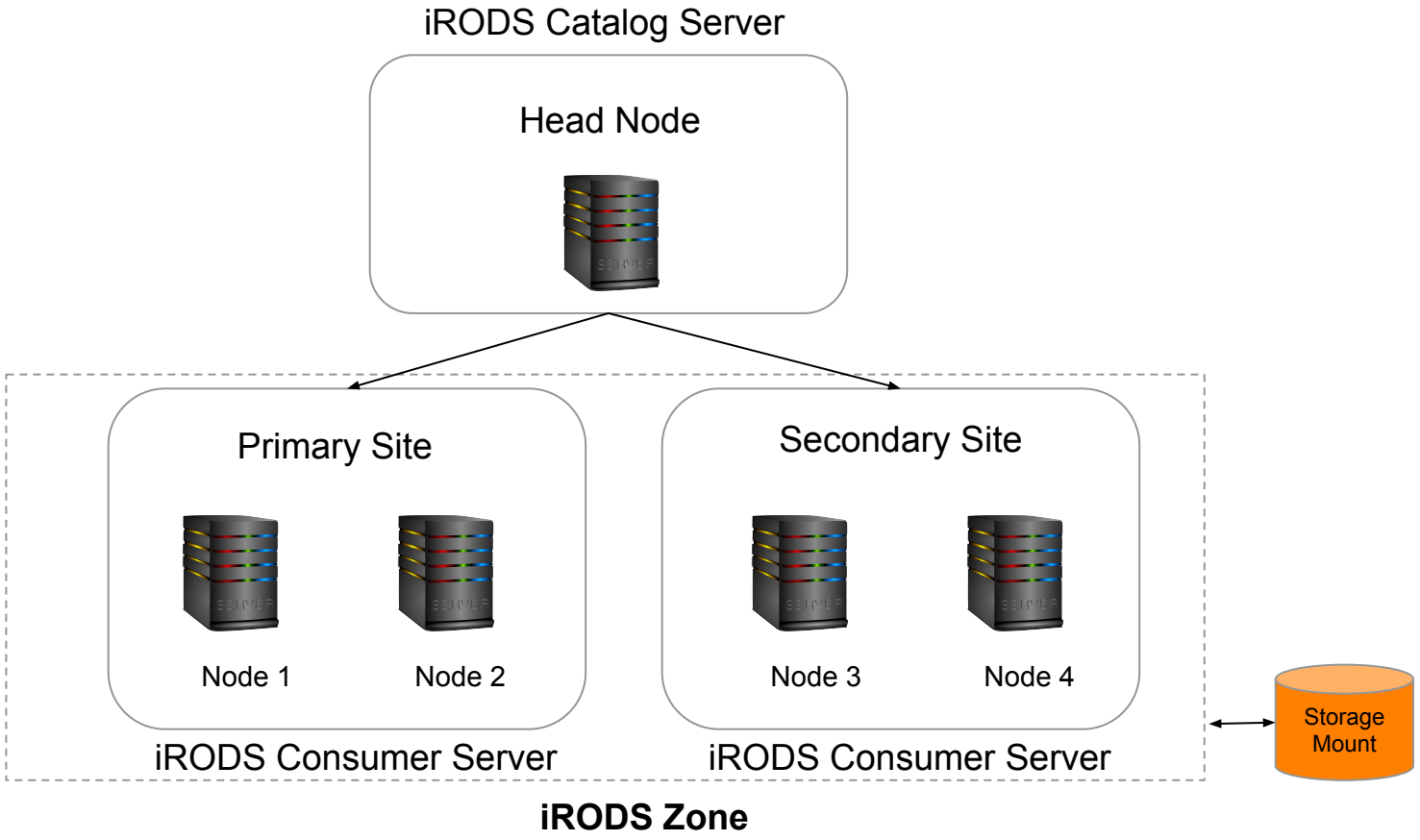
Replication Solution Options



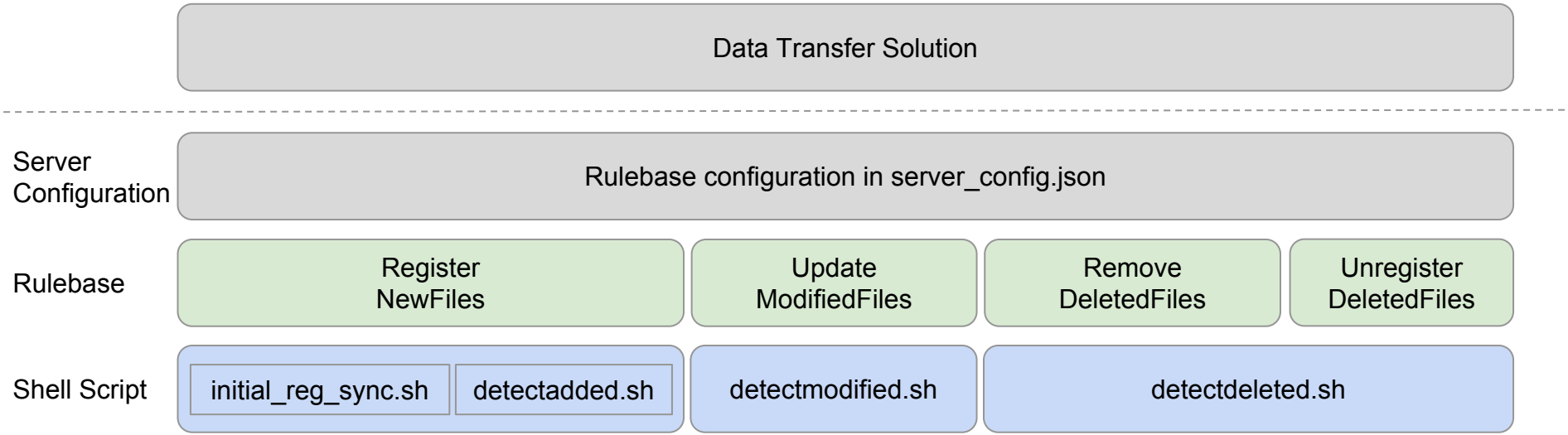
Replication using Data Transfer Solution



iRODS Architecture in Data Transfer Solution



Ingest Metadata using iRODS File System Scanner




```
META_DATA_ATTR_NAME  = filesystem::mtime
META_DATA_ATTR_VALUE = 2018-06-05 13:02:11.914472000
```

```
META_DATA_ATTR_NAME  = filesystem::deleted
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.

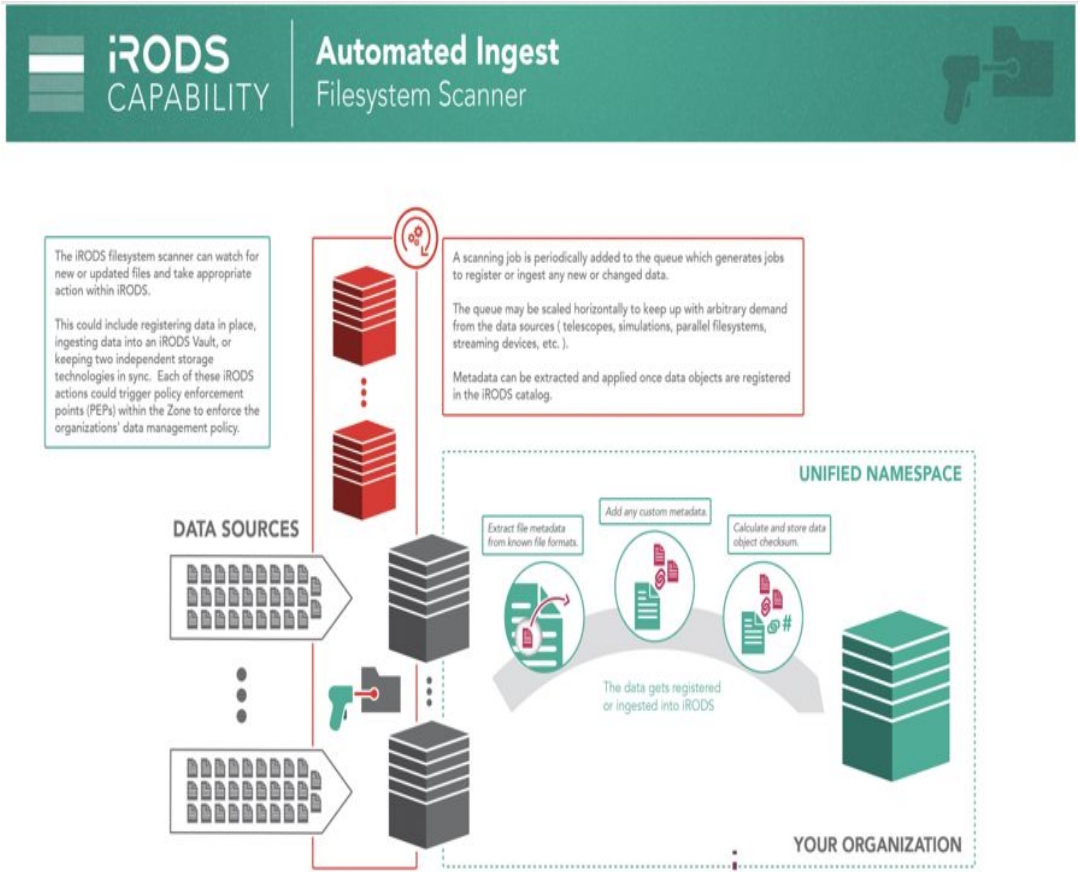
- For the Data Replication use case, iRODS will be used to provide the system metadata of the storage that includes:

    - 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

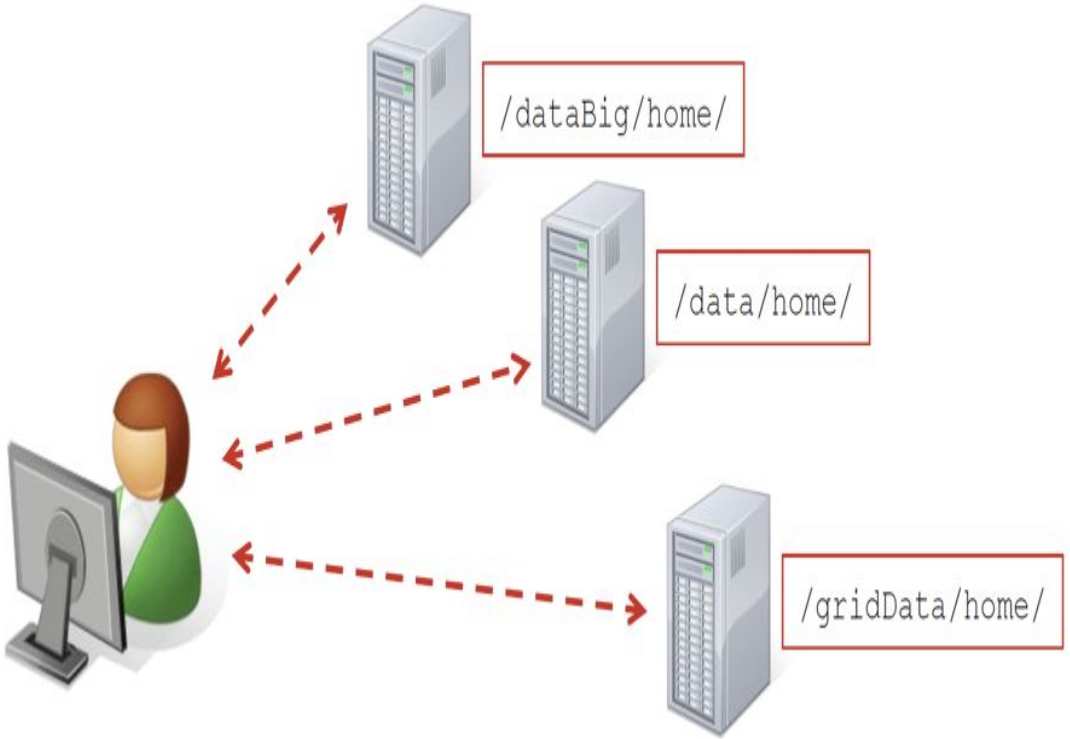
Next Step - iRODS Automated Ingest Framework

- 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

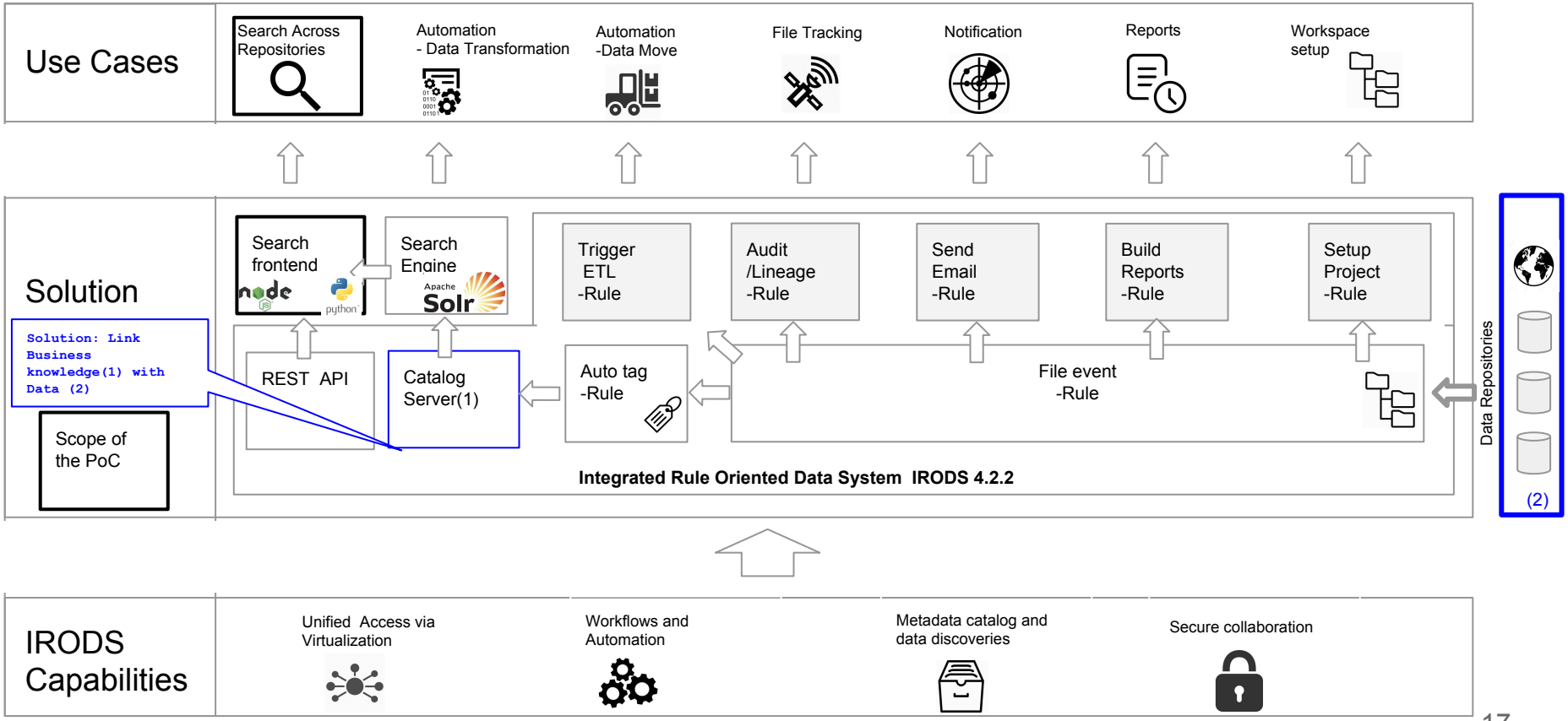


PoC - Data Catalog using iRODS

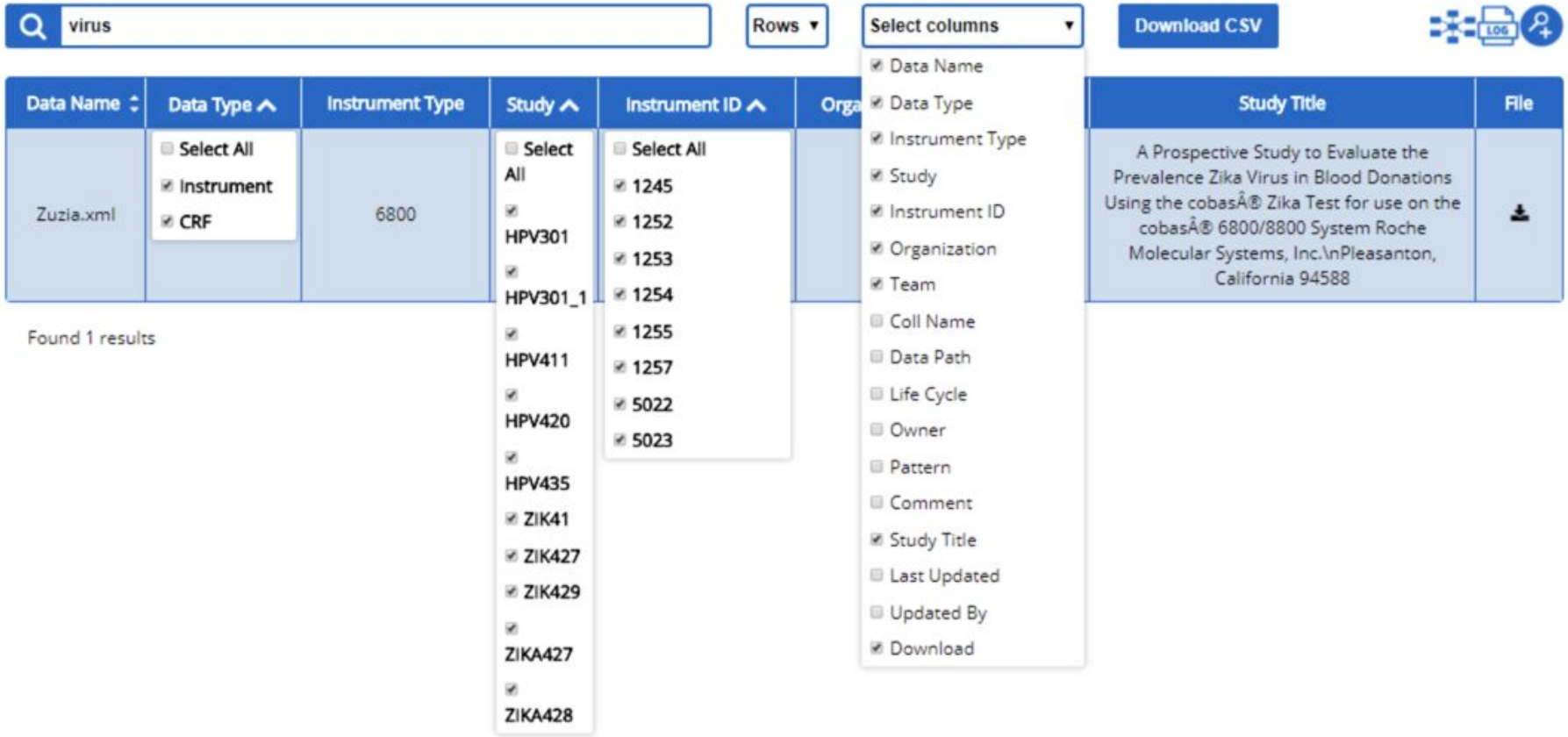
- Enable simplicity of access with one namespace and want to make data locality transparent to the user
- Ability to search and access to data and metadata



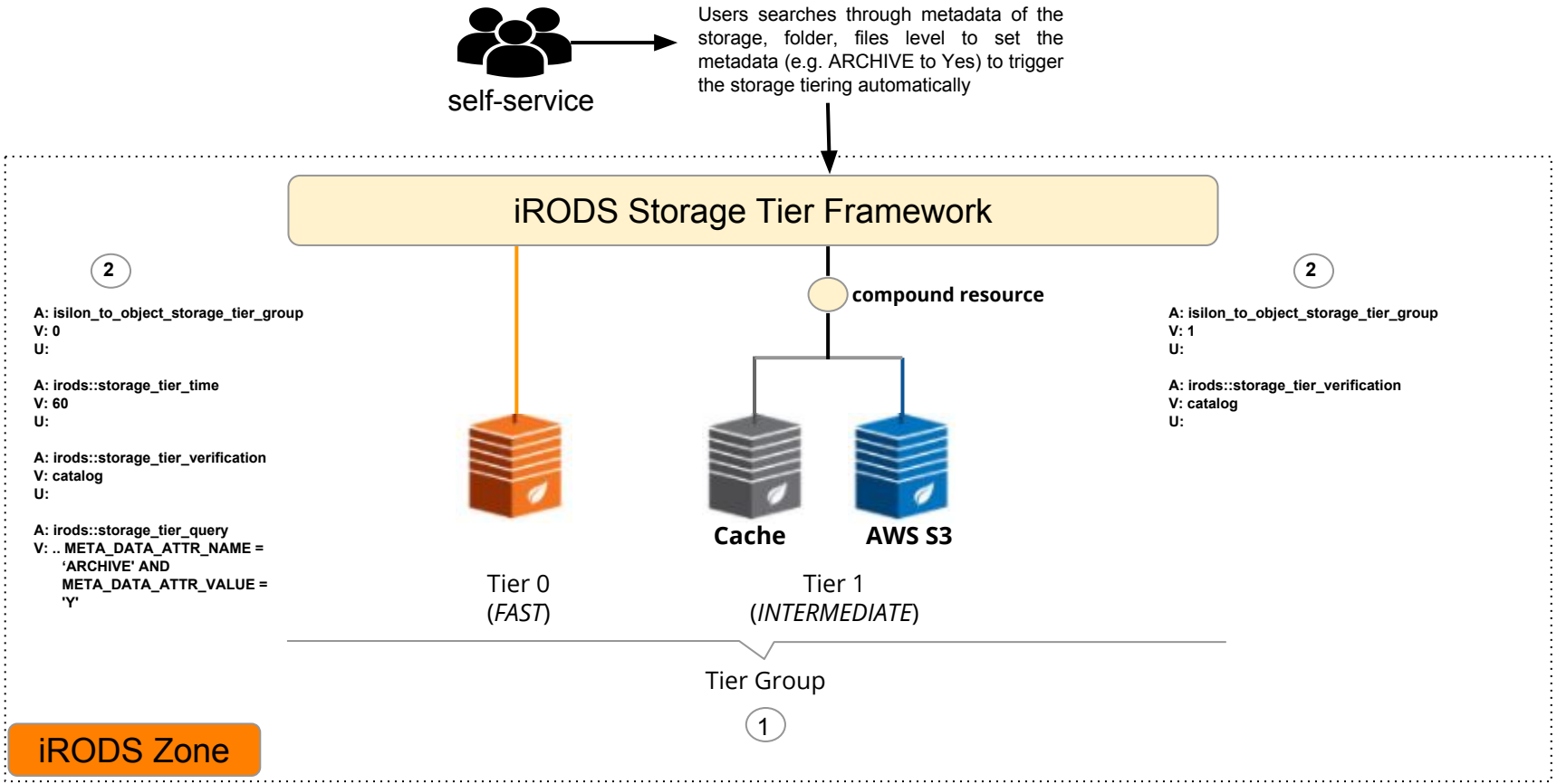
PoC - Data Catalog using iRODS



PoC - Data Catalog using iRODS

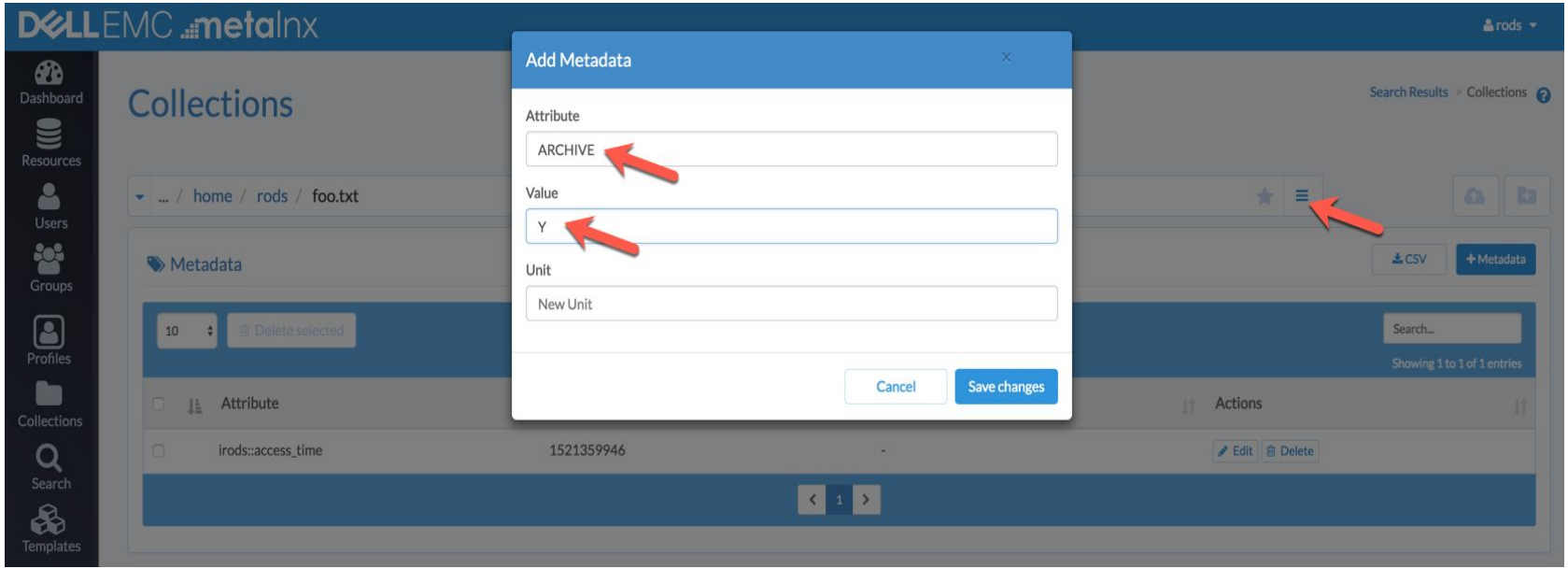


PoC - Enable Intentional Archive



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

```
-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?