iRODS

Building a Federated Research Collaborative

David Fellinger
Data Management Technologist
iRODS Consortium
12 February 2020

Early Data Organization

An early storage medium for research works.

Early Data Discovery

Discovery through the use of metadata

Maybe useful for single instances of data but no federation capability

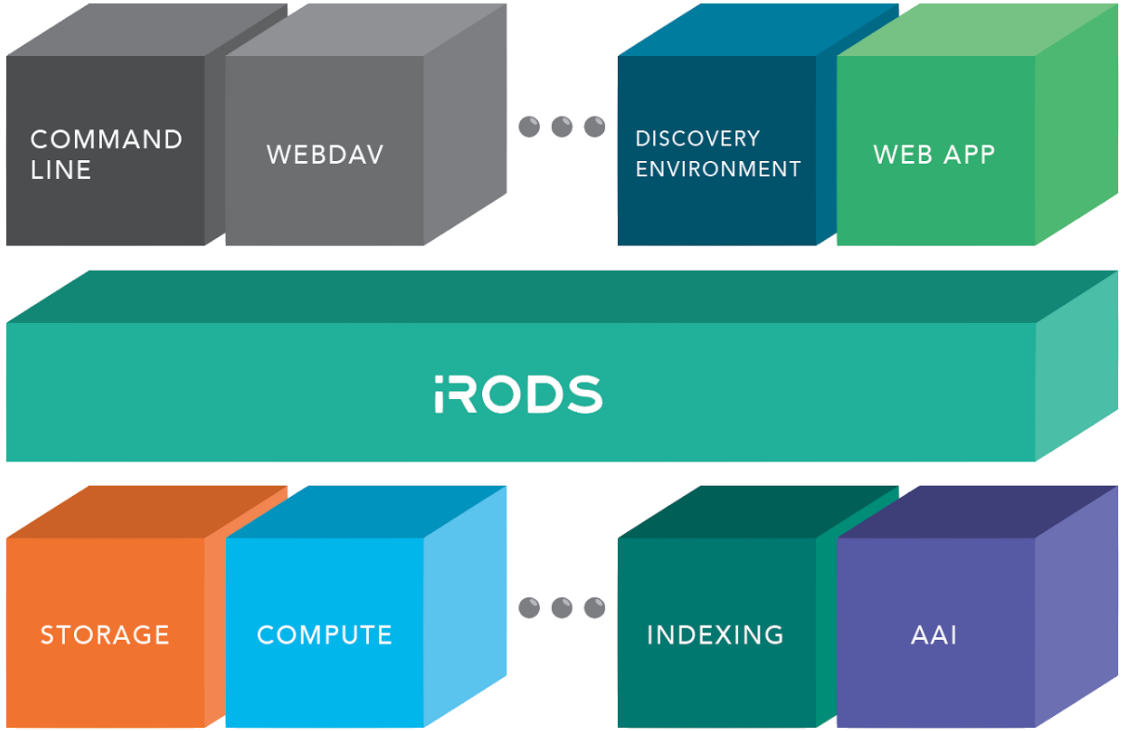
What is iRODS?

- Funded initially by the US Defense Advanced Research Projects Agency (DARPA) in 1995 as the Storage Resource Broker
- The Integrated Rule-Oriented Data System (iRODS) has been designed by the iRODS Consortium with 4 key functionalities;

iRODS is:

- Open Source
- Distributed
- Data Centric
- Metadata Driven

What is iRODS?



iRODS Clients

iRODS provides a layer of abstraction which integrates with your pre-existing infrastructure.

This flexibility allows your infrastructure to continue to change over time.

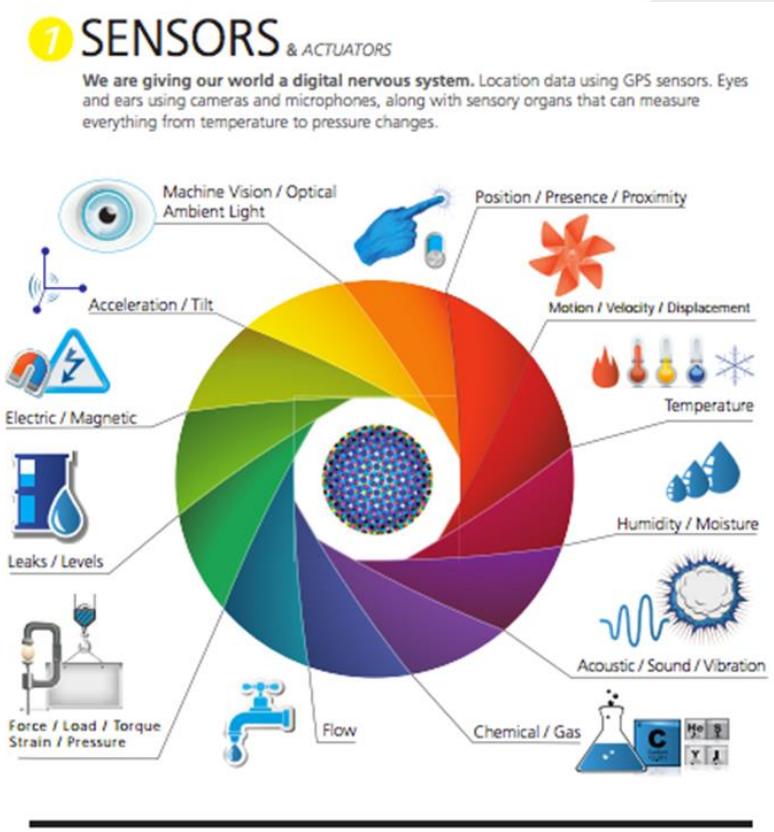
Existing Infrastructure

Metadata Driven

- iRODS moves data based on **User Defined Metadata.**
  - Latitude, longitude, altitude
  - Anomalies in genomic sequences
  - Data collection points
  - Instrumentation details enabling the data collection
  - Specific relevance in a research area
- The use of **Rich Metadata** enables:
  - Discovery
  - Data grouping based on content to enable analysis
  - Data movement to analytic platforms
- iRODS is **Data Centric** and Metadata can be extracted from file headers or actual file content.
  - Metadata extraction is based on set rules to produce a collection
  - Data can be apportioned instantly based on metadata
  - Metadata can include citation instances and can change dynamically

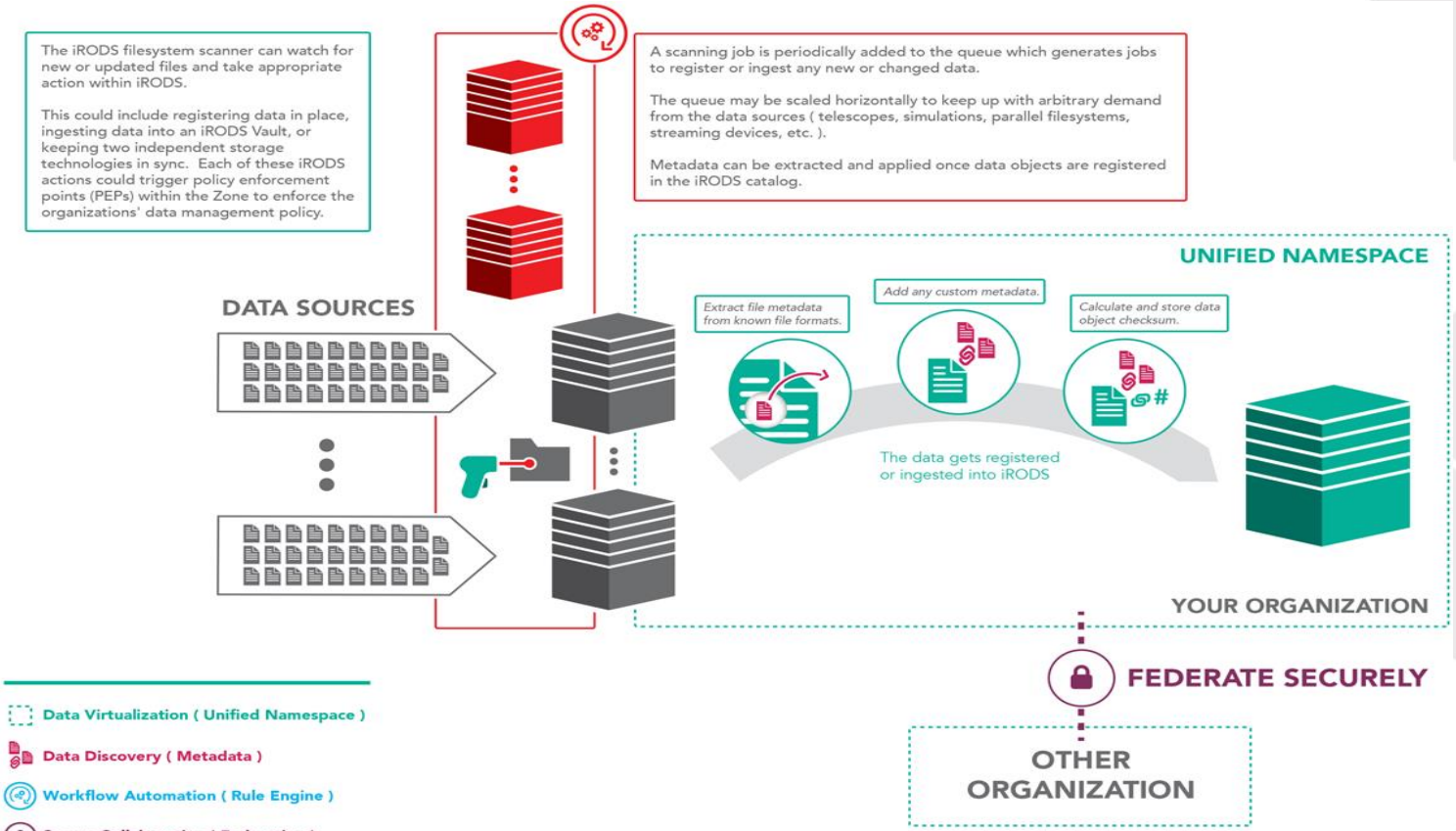
The Rise of Sensor Data and "Big Data"

- It can be argued that sensor data has changed the paradigm of HPC.
  - Huge amounts of data must be collected
  - The data has the characteristics of, Volume, Variety, Velocity and Value
  - The data must be organized for analysis
  - In many instances the data must be moved to a file system close to the analytic element
  - An analytic process must be started only when the full data set is available
- All steps must guarantee provenance of the collected data to assure Veracity.
- Full automation includes moving the results to a data distribution file system.



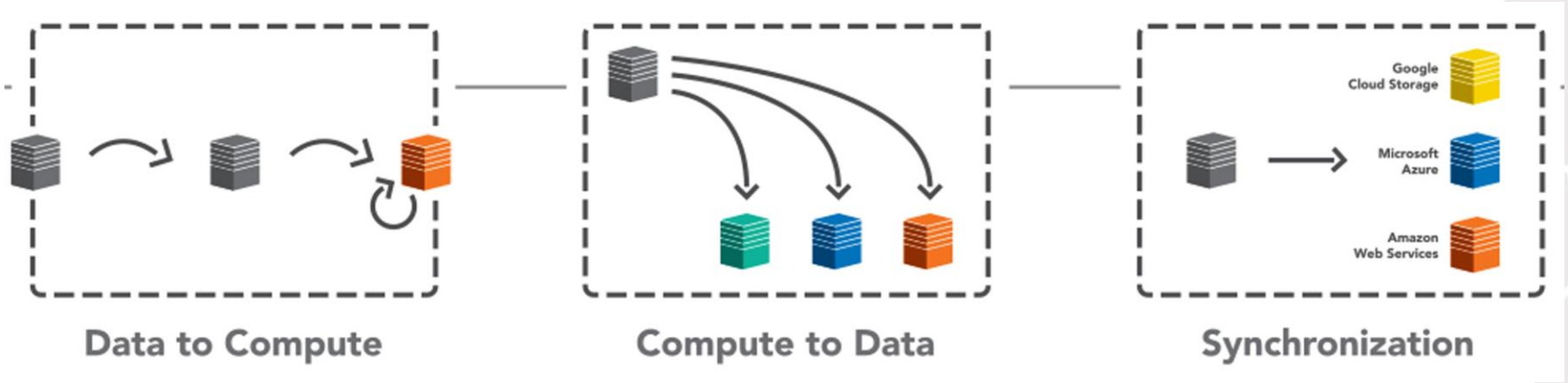
Arcot Rajasekar DataNet Federation Consortium

iRODS Automating, Gathering, and Organizing Data



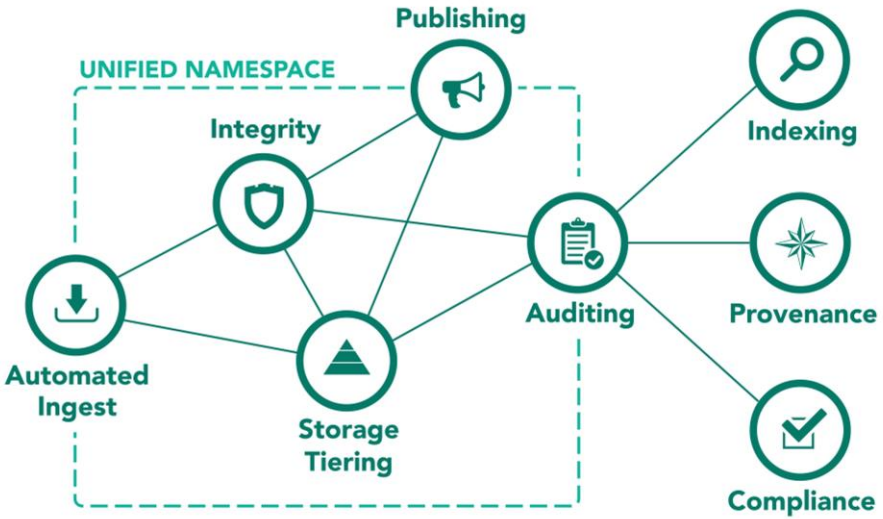
Automated Management Through Synchronization to the Cloud

- iRODS can migrate data to any filesystem maintaining location data in the catalog.
- Data can be synchronized to the cloud or a federated partner.
- Notifications can be provided at each step of the process and an audit report can be generated at any time.

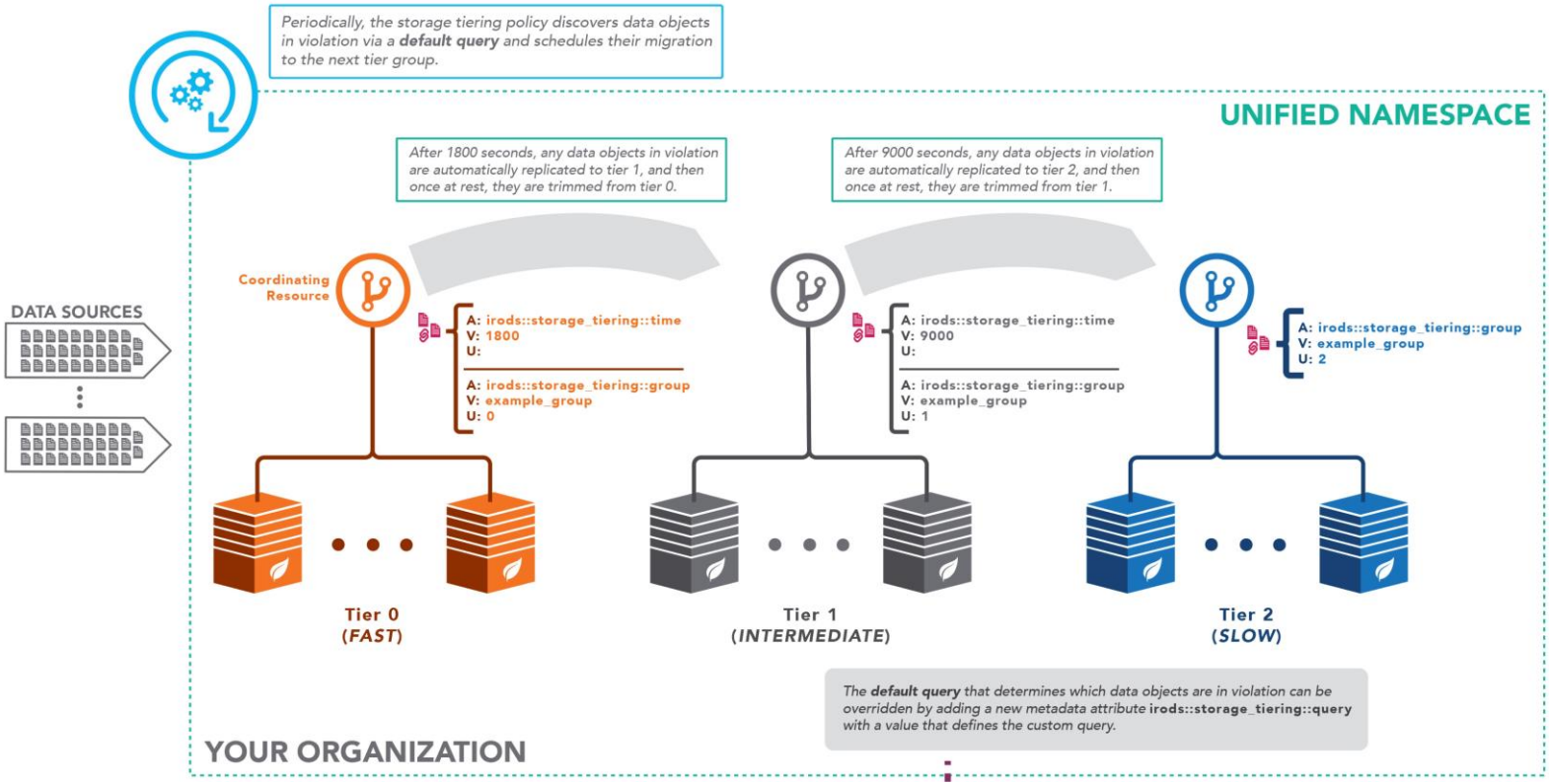


Packaged Capabilities Allow Tracking and Managing Data Through Publication

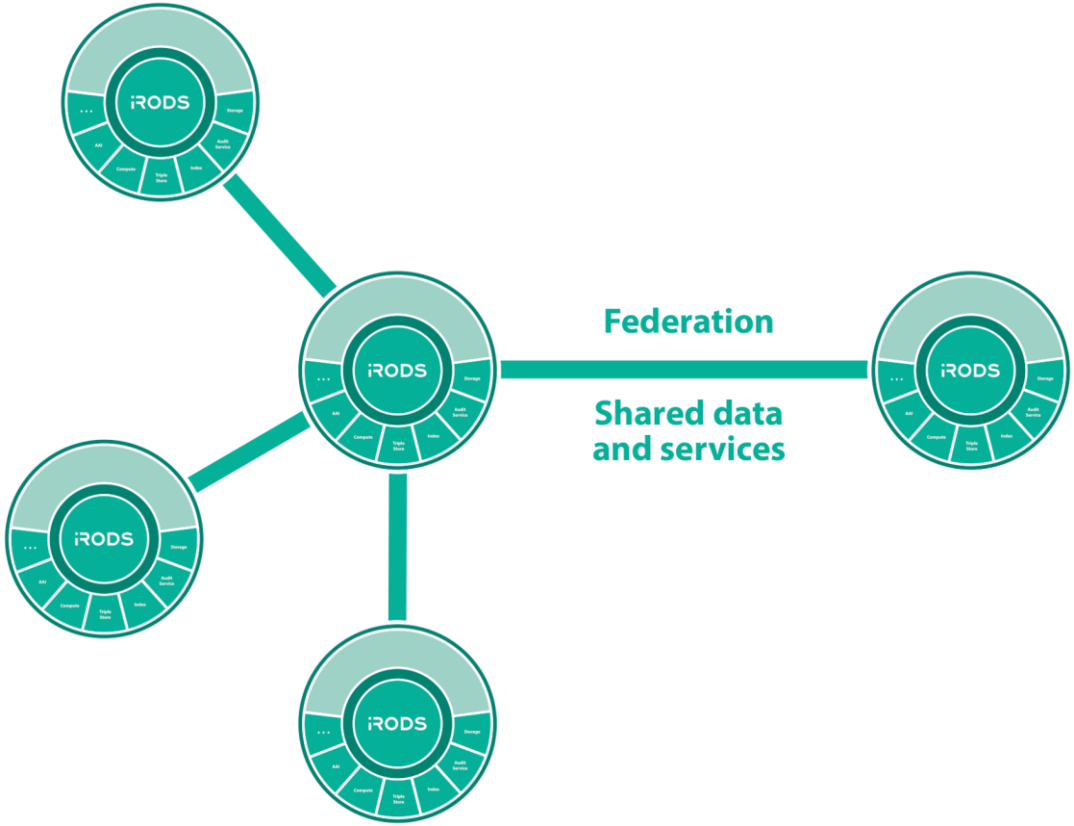
- iRODS provides eight packaged capabilities which can be configured and deployed to serve the needs of the data center.
- Organizations can seamlessly address their immediate needs.
- Additional capabilities can be added or reconfiguration can occur as the need arises.
- A plugin architecture allows customization to address any data migration need.



Automation to Enable the Establishment of an Archive



Secure Federation Enables Geographically Protected Data Archives and Collaboration

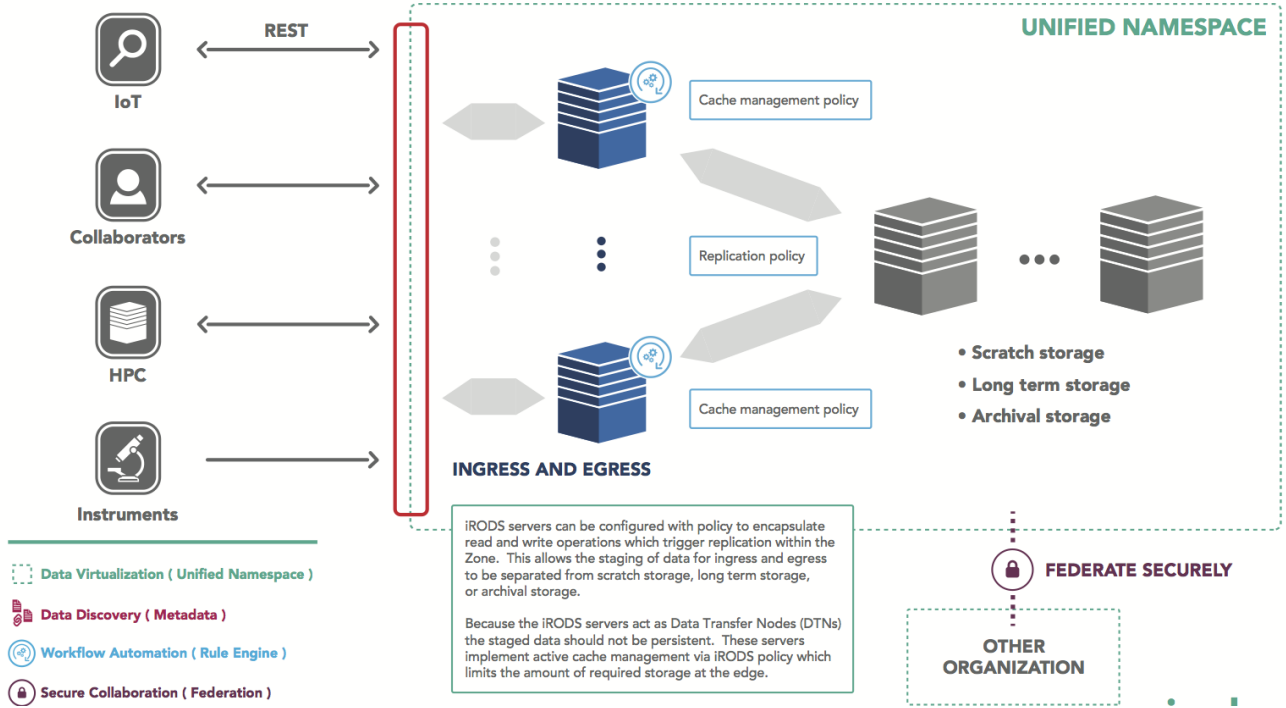


iRODS Data Transfer Nodes

Moving large datasets across organizational boundaries remains a challenge due to the requirement of exposing high performance hardware to the public network. Data Transfer Nodes (DTNs) provide a secure location for ingress and egress of data while avoiding the performance impact of an organizational firewall.

In the following deployment pattern, iRODS satisfies the requirements of a Science DMZ while also providing automated data management.

> The Science DMZ is a portion of the network, built at or near the campus or laboratory's local network perimeter that is designed such that the equipment, configuration, and security policies are optimized for high-performance scientific applications rather than for general-purpose business systems or 'enterprise' computing.
>
> —ESnet



Deployment: CyVerse

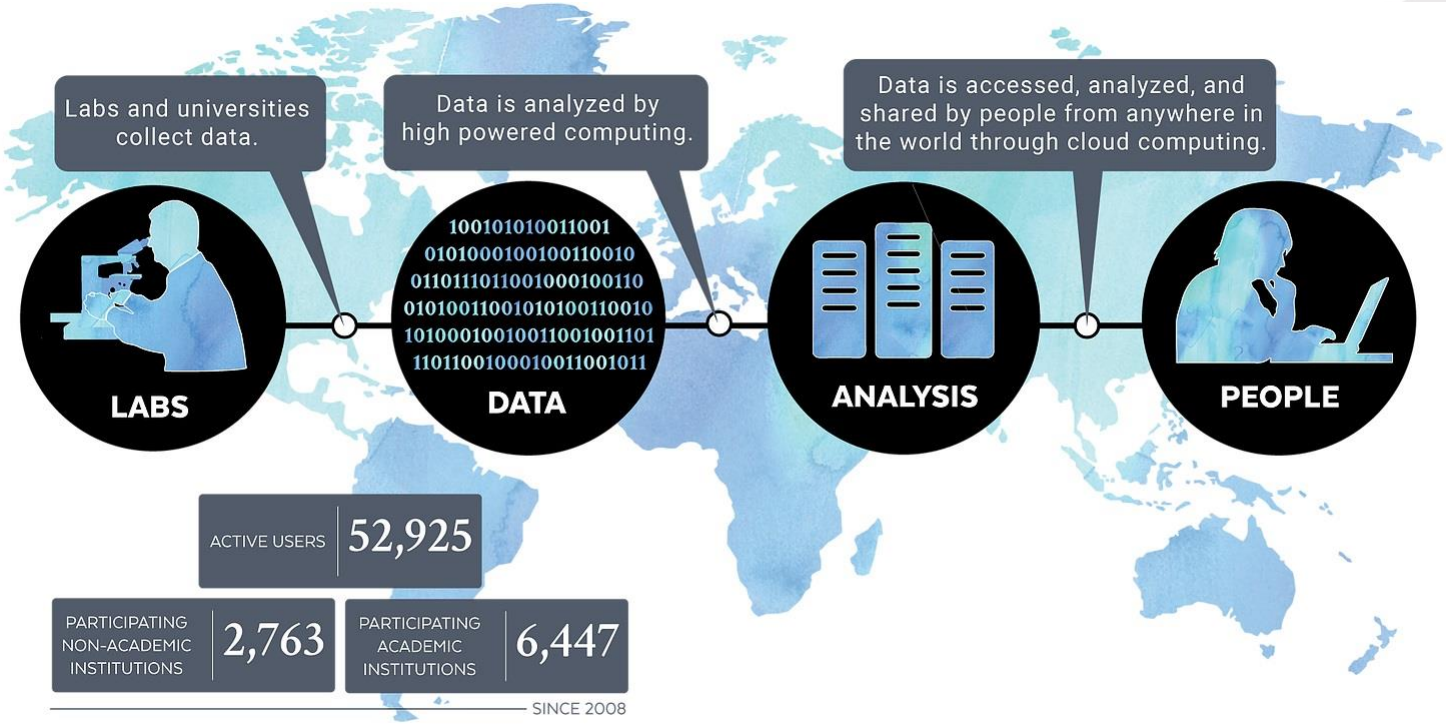


Diagram available from: https://www.cyverse.org/about accessed 25 September 2019

Deployment: EUDAT CDI

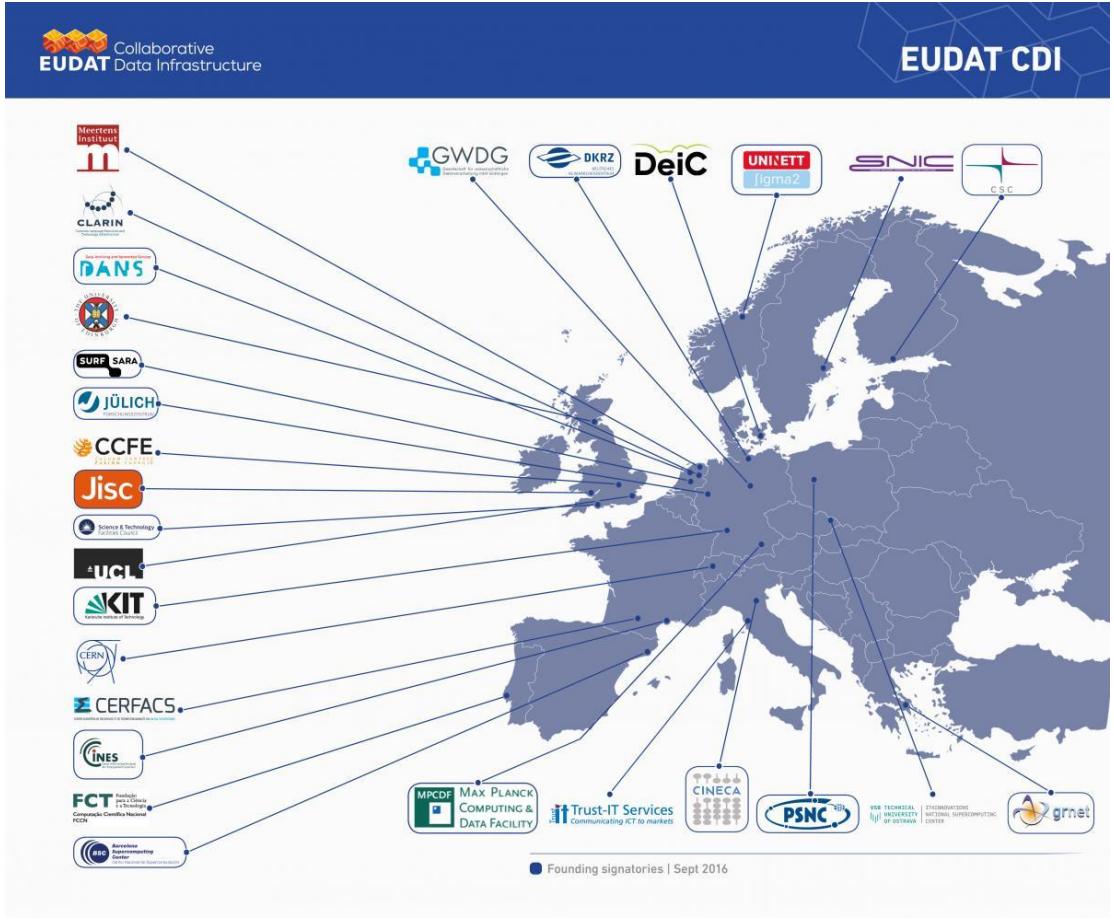


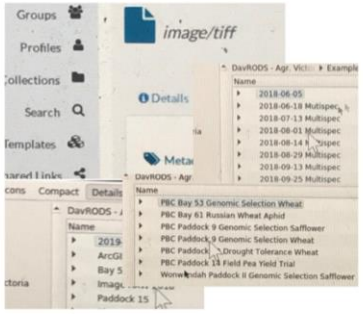
Diagram available from:
https://eudat.eu/eudat-cdi
Accessed 26 September 2019

Deployment: Victoria Department of Agriculture

Completed Use Case. Next iteration endorsed

- Testing data ingest to S3 bucket and open source metadata management application (iRODS).
- A new capability in data discovery and workflow automation.
- Enable data classification and reporting to support rapid assessment of data assets and use.
- Fast track data processing and transfer to defined repositories for management and use.
- Better manage data sovereignty, preservation and reproducibility for researchers.

```
# POLICY CONFIGURATION VARIABLES
SCANNED_RESOURCE = 'example_scanned_resc'
DESTINATION_RESOURCE_ROOT = 'example_dest_resc_root'
LIST_OF_DESTINATION_RESOURCE_LEAVES = ['a','b','c']
```



AGRICULTURE VICTORIA

Diagram courtesy of Kieran Murphy Victoria Department of Agriculture

The Integrated Rule-Oriented Data System (iRODS) is open source data management software used by research organizations and government agencies worldwide

Conclusion

- eResearch has evolved to accommodate sensor and other types of “big data”.
- The use of user defined and extracted metadata improves the disposition of data at every level.
- iRODS enables secure federation simplifying secure, rules based collaboration.
- iRODS can enable complete workflow control, data lifecycle management, and present discoverable data sets with assured traceability and reproducibility.

The iRODS Consortium (iRODS.org)

The iRODS Consortium

- Leads software development and support of iRODS
- Hosts iRODS Events
- Tiered membership model

Research Computing
UNIVERSITY OF COLORADO **BOULDER**

Universiteit Utrecht

Questions?

Thank you!

David Fellinger

davef@renci.org

iRODS.org