iRODS®

A Survey of iRODS Rules to Enforce Site Policies and Enable Automated Workflows

David Fellinger
Data Management Technologist
iRODS Consortium
11 February 2021

Factors Defining Data Site Policies

- How is authentication and authorization defined?
  - Organizational constraints
  - Constraints set by funding agencies
  - Local and national government rules
  - International collaboration rules
  - Adherence to rules pertaining to patient health data
- Is the site hosting multiple collections?
  - Sites may host everything from genomic to climate data
  - Quantities of storage must be budgeted and apportioned
  - Collaboration must be considered
- Is the site responsible for data processing as well?
  - How are processing priorities handled?
  - How is provenance of data products guaranteed?
  - How are data products stored and distributed?

Factors Defining Data Site Policies

- Who owns the data?
  - Is the data owned by the researcher?
  - Are ownership rights dictated by governing bodies?
  - What contracts are in place with funding organizations?
- How long is data retained?
  - Retention must be based on factors beyond file date
  - Funding agencies may dictate data retention
  - Determinations must be made regarding storage tiering
- How is policy adherence guaranteed?
  - Audit trails and reporting to those responsible are essential
  - Storage usage tracking may be critical in budget determinations

iRODS addresses these challenges and is used by data sites across the world to manage policy adherence

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) was developed starting in 2006 by a university-based group
- The Integrated Rule-Oriented Data System (iRODS) has been designed by the iRODS Consortium with 4 key functionalities:

iRODS is:

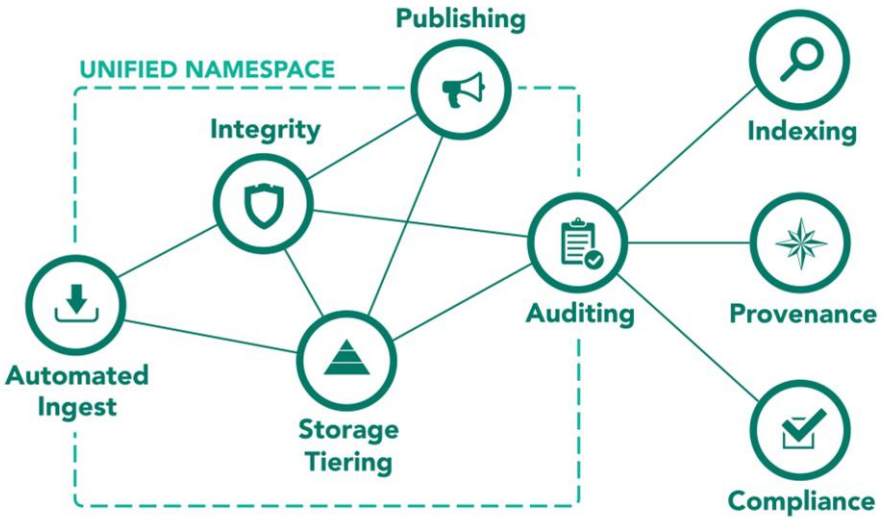
- Open Source
- Distributed
- Data Centric
- Metadata Driven

Metadata Driven to Enable Policy Adherence

- iRODS builds a catalog 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:
  - Management of data retention and placement
  - Placement based on content enabling enhanced security
  - 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

Packaged Capabilities Allow Managing Data Through Publication

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



Initial Goals

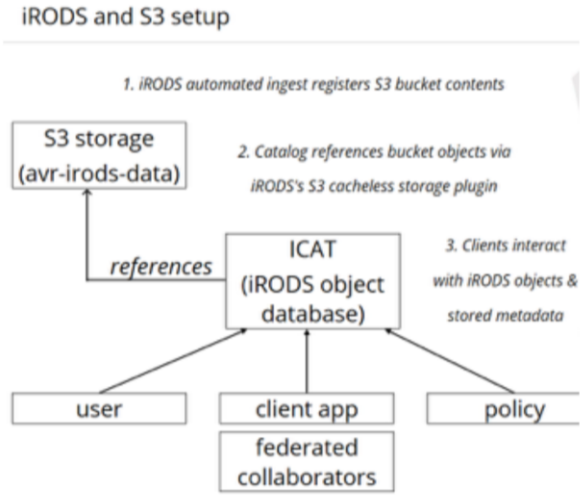
1. Upload existing AVR data as example content into S3 bucket **avr-irods-data**
2. Get S3 files / folders registered to iRODS catalogue
3. Extract salient metadata – e.g. EXIF tags in TIF files
4. Tag Data Objects and Collections to make them Actionable and Discoverable

AGRICULTURE VICTORIA

Presentation available from: https://irods.org/uploads/2020/Murphy-AgVic-SmartFarm_Data_Management-slides.pdf accessed 14 January 2021

The Content

- **Ingest policy registers object in place then extracts metadata**
- **Apply metadata to the object in the catalogue**
  - Metadata headers available in the files
  - Contextual metadata : LZ directory, instrument, etc
- **Demonstrate**
  - Ingest
  - Discovery
  - Data egress
  - Graphical presentation
  - File system presentation : WebDAV & emerging new front ends.



AGRICULTURE VICTORIA

Presentation available from: https://irods.org/uploads/2020/Murphy-AgVic-SmartFarm_Data_Management-slides.pdf accessed 14 January 2021

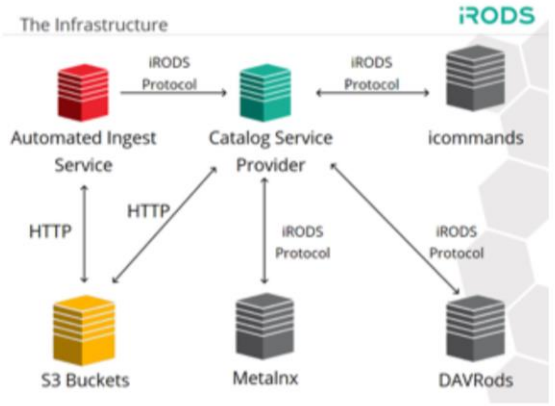
Data Discovery with Metalnx

Automated ingest has provided metadata for data discovery

The metadata can be directly inspected in Metalnx

The query builder can be used to identify data sets of interest via Attribute, Value, Unit matches

Queries to the system metadata may also be performed, searching on values such as file name, collection path, user, etc.



AGRICULTURE VICTORIA

Presentation available from: https://irods.org/uploads/2020/Murphy-AgVic-SmartFarm_Data_Management-slides.pdf accessed 14 January 2021

Emerging SmartFarm Data Infrastructure

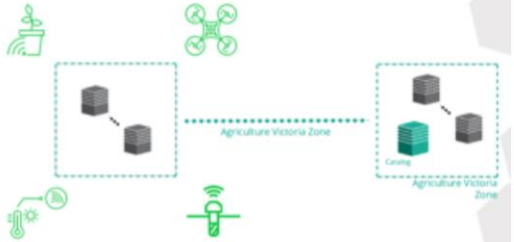
Each SmartFarm may host their own application (iRODS) to manage metadata description and catalogue for each UAV trial.

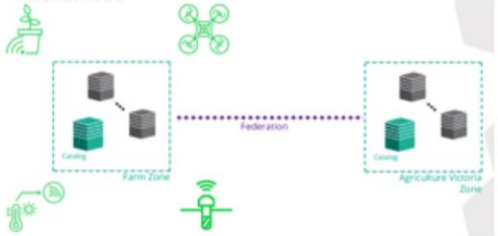
Data is gathered from the UAV over the protocol of choice.

Data is periodically synchronised to Agriculture Victoria Research servers (S3 / Hybrid)

SmartFarm hosts Agriculture Victoria Research servers (S3 / Hybrid)

Data is periodically **replicated** to Agriculture Victoria Research Servers (BASC)



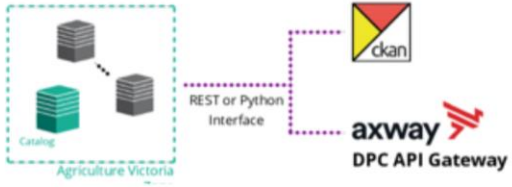


Once data is at rest in the Agriculture Victoria Research namespace i.e. Horsham_UAV_AVR_Plot1

Data may be replicated to HPC storage for analytics.

Data may be published to CKAN or made accessible via the API gateway

Data may be shared over an IRODS interface : WebDAV, MetaInx, NFS, Command Line.



Presentation available from: https://irods.org/uploads/2020/Murphy-AgVic-SmartFarm_Data_Management-slides.pdf accessed 14 January 2021

Deployment: CyVerse

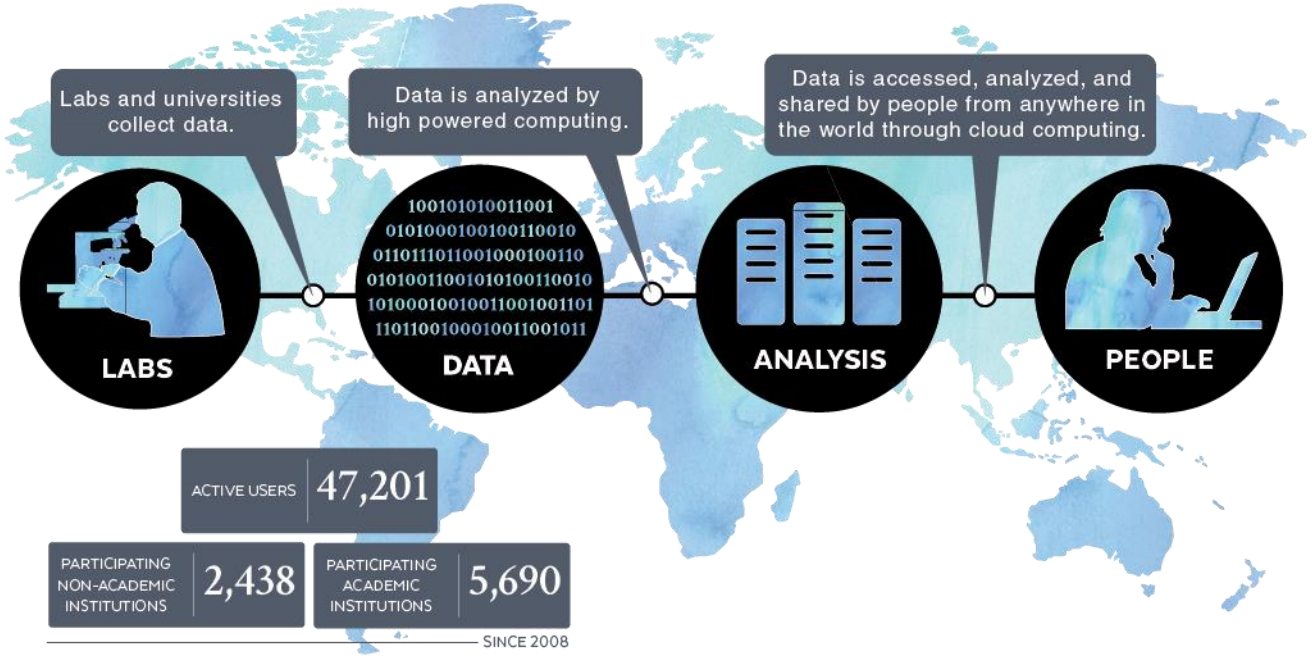
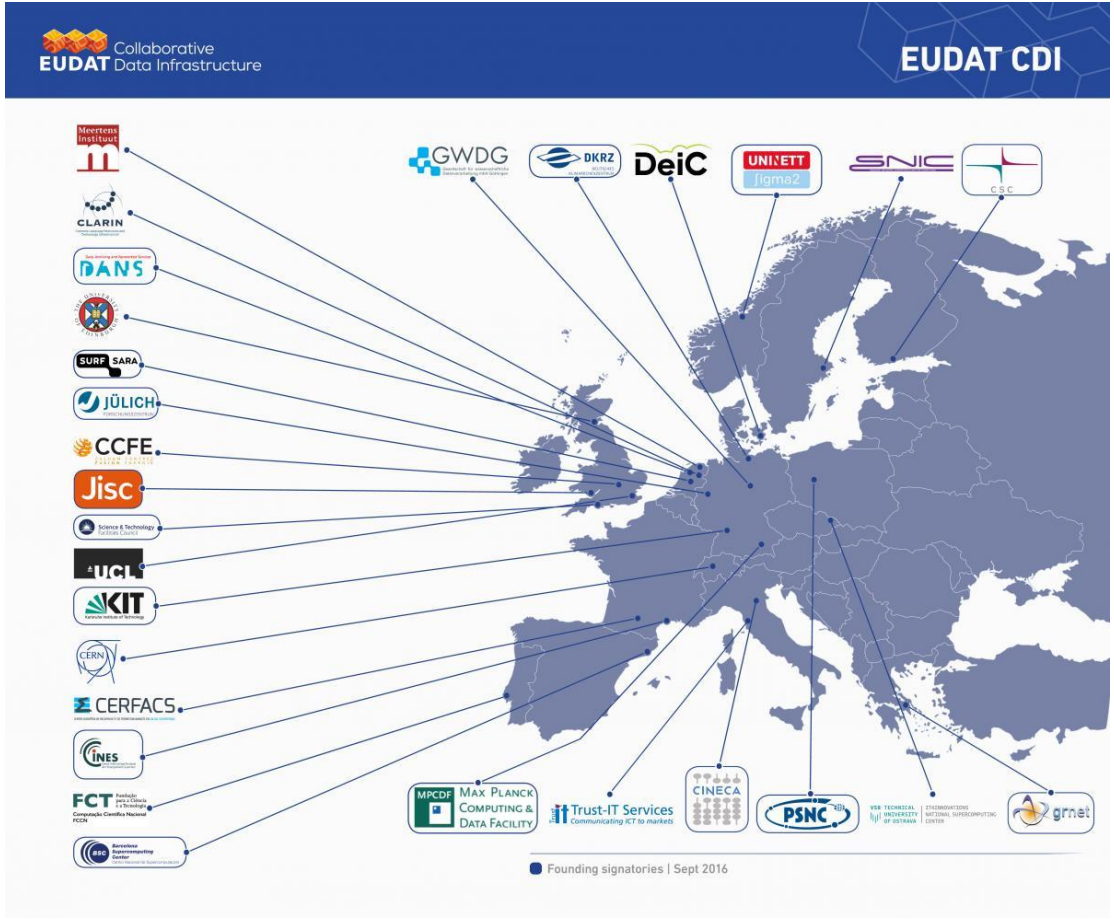


Diagram available from: https://www.cyverse.org/about accessed 14 January 2021

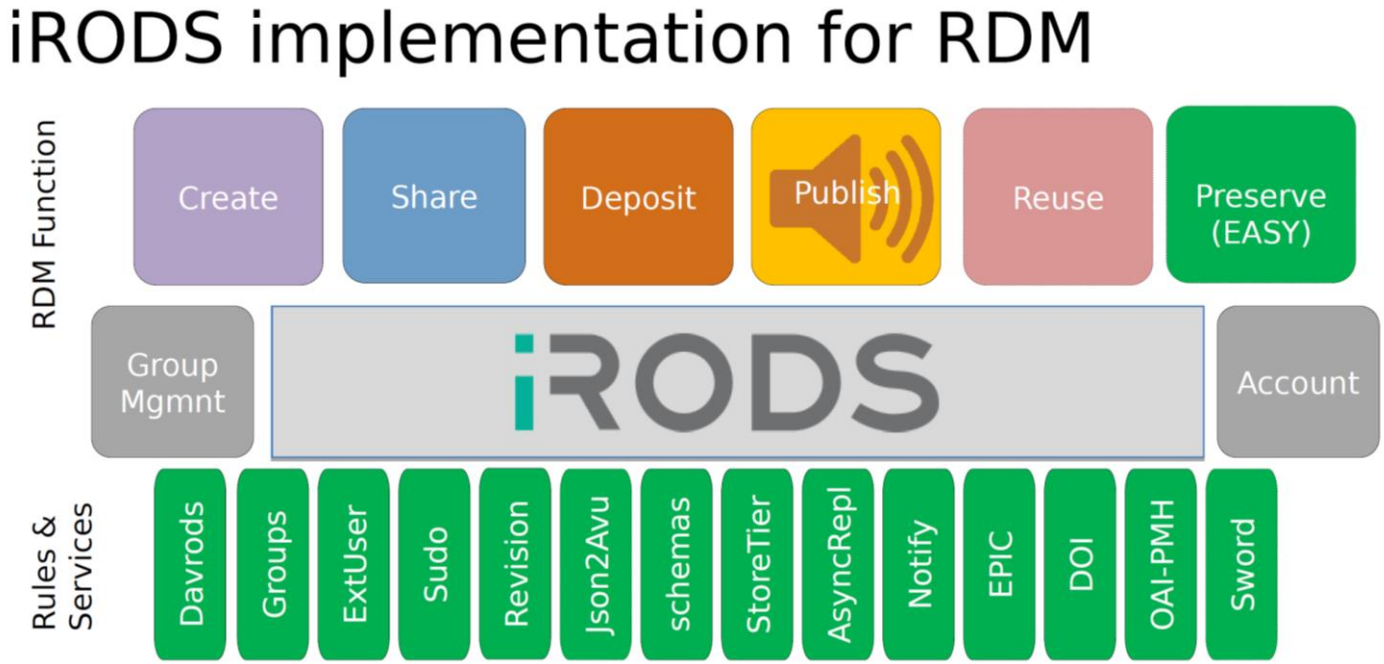
Deployment: EUDAT CDI



28 members
spanning the EU

Diagram available from:
https://eudat.eu/eudat-cdi/members
Accessed 14 January 2021

Deployment: Utrecht University

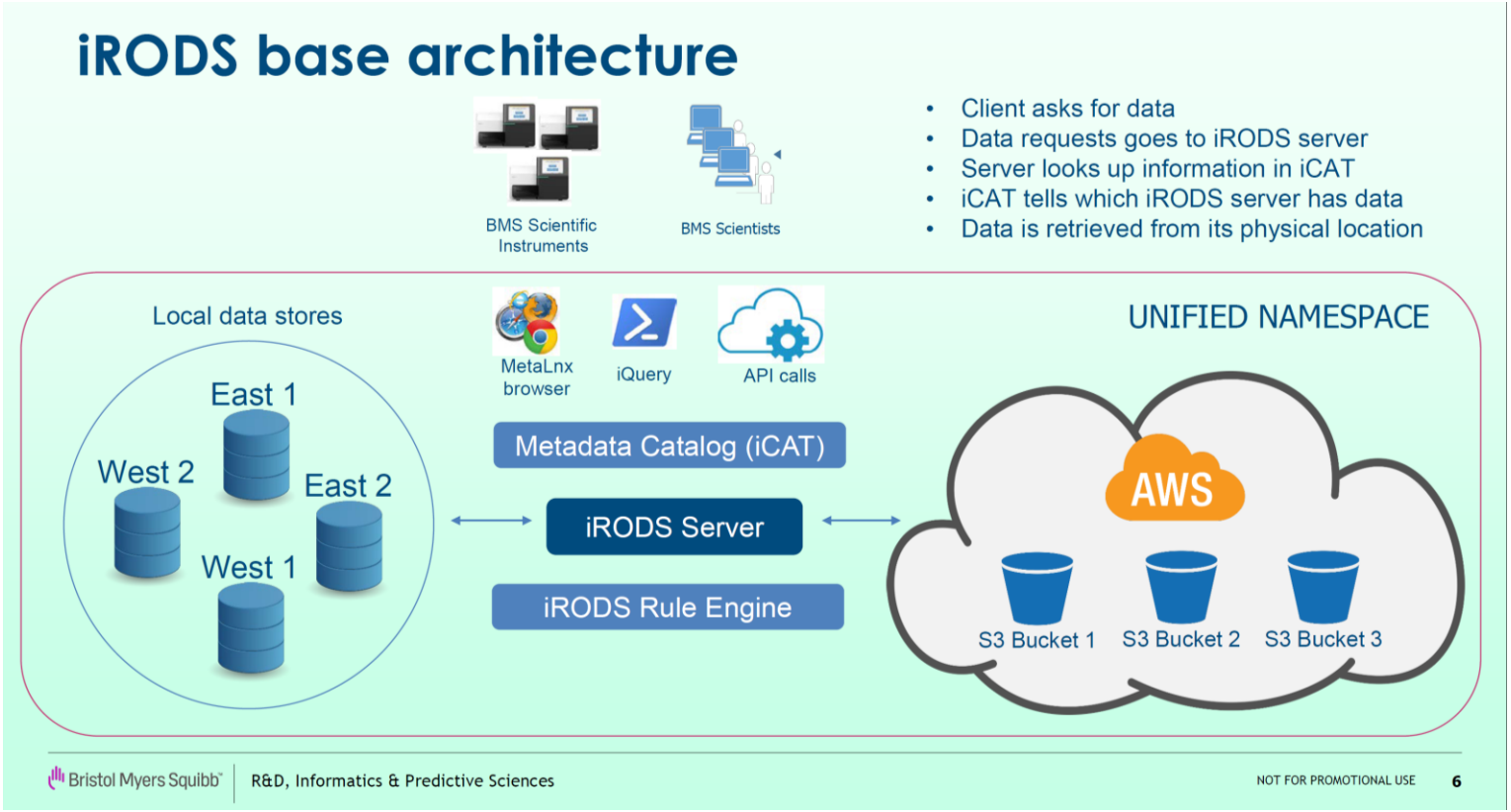


Utrecht University

https://github.com/UtrechtUniversity/yoda

Presentation available from: https://irods.org/uploads/2020/Westerhof-Smeele-UtrechtUni-Yoda_and_iRODS_Python_rule_engine_plugin-slides.pdf accessed 14 January 2021

Deployment: Bristol Meyers Squibb



Presentation available from: https://irods.org/uploads/2020/Shaikh-BMS-iRODS_for_scientific_applications_in_AWS_Cloud-slides.pdf accessed 14 January 2021

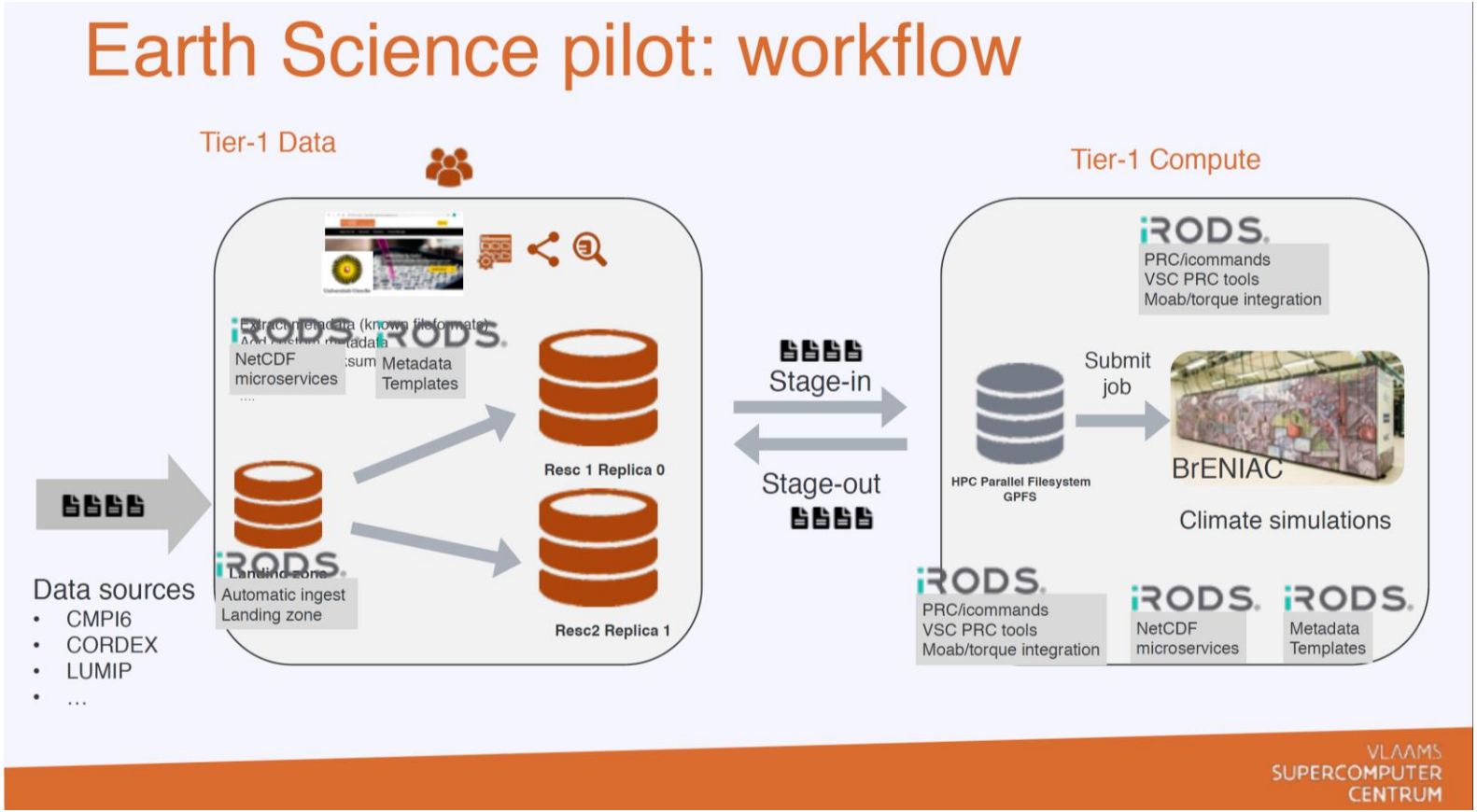
Processing Data at Scale

Using iRODS for managing petabytes of data in hundreds of millions of files on distributed storage resources spread across the country.

- Number of S3 buckets: **200+**
- Number of objects in S3: **800+ millions**
- Size of dataset: **10+ PB**
- Processing rate (regular data ingest): **5 millions objects per hour**

Presentation available from: https://irods.org/uploads/2020/Shaikh-BMS-iRODS_for_scientific_applications_in_AWS_Cloud-slides.pdf accessed 14 January 2021

Deployment: KU Leuven



Presentation available from: https://irods.org/uploads/2020/Barcena-KULeuven-VSC-iRODS_Data_Management_Platform-slides.pdf accessed 2 October 2021

Conclusion

- iRODS has been built to enable the most stringent requirements of data site policy requirements.
- Use cases span research sites and disciplines worldwide.
- iRODS enables secure federation simplifying secure, rules-based collaboration across sites.
- All iRODS data actions are auditable to enable proof of adherence to site policies and assurance of end-to-end data provenance.
- 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



Questions?

Additional use cases can be found in the proceedings of the 2020 iRODS User Group Meeting:
https://irods.org/ugm2020/

Thank you!
David Fellinger

davef@renci.org