Science For A Better Life

Implementing a Genomic Data Management System using iRODS at Bayer HealthCare

iRODS User Group Meeting 2015

Carsten Jahn – Bayer Business Services GmbH, R&D IT, HealthCare Research
Navya Dabbiru – Innovations Labs, Tata Consultancy Services, Hyderabad, AP India

Improving Data Management for Bioinformatics Users

Collaboration Contract

Patient's Informed Consent

- Genomic data (e.g. DNA, RNA sequencing) is generated with less cost, massive data amounts need to be managed
- data overview - where is the data of a patient who retracted his study consent?
- directory namespace larger than one file system, e.g. archival of data
- multiple systems for data analysis:
  Linux commandline based and GUI applications

iRODS at Bayer

iRODS project started April 2014, productive since December 2014

Addressing requirements for

- data security and user management
- metadata import and search
- data transfer automation – sequencer to iRODS
- stable and safe operations

Implementation is supported by Tata Consultancy Services.

Data Ownership

general Bayer web application

study X

- owner(s)
- members

LDAP group

custom replication

iRODS

iRODS group

study owners are enabled to manage their team and grant access rights

other applications

iRODS admins are no “data admins“

user operating manuals

Users and Metadata

Study / Project Data Owner

"iRODS? Never heard about..."

"Ok, I can tell you the data has *Restriced* access."

Bioinformatician, Data Analyst

"I want to keep track of the tools (and versions) that produced this output."

Metadata

Metadata Entry:

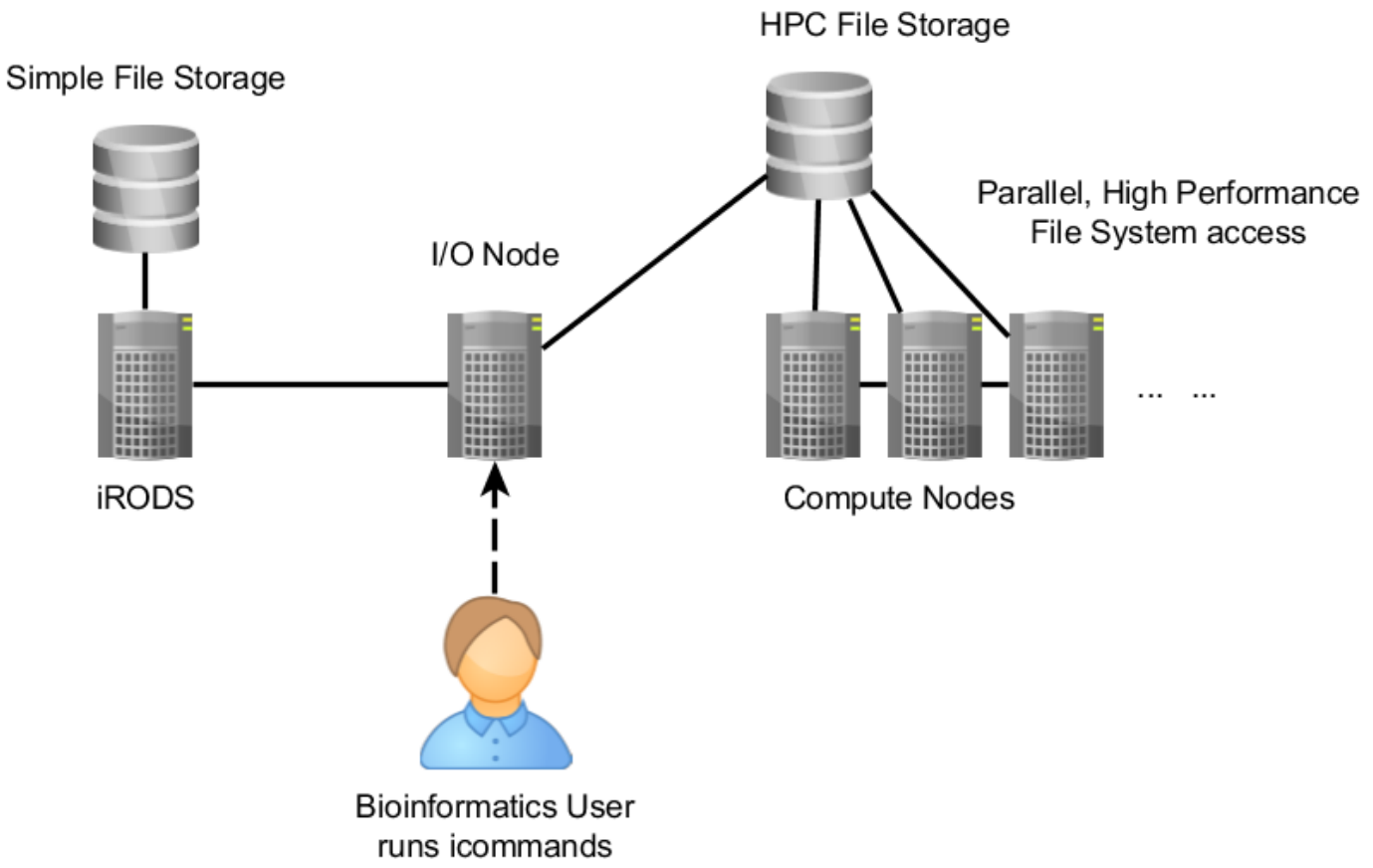
Attribute	Value	Unit	*Message*
study_type	research		
bayer_study_id			mandatory attribute, value is required
contact_person	Karl Heinz		
cross_study_analysis_allowed	within indication		
data_classification	Restricted		
external_reference			
health_status			

Restricted
Secret
Internal

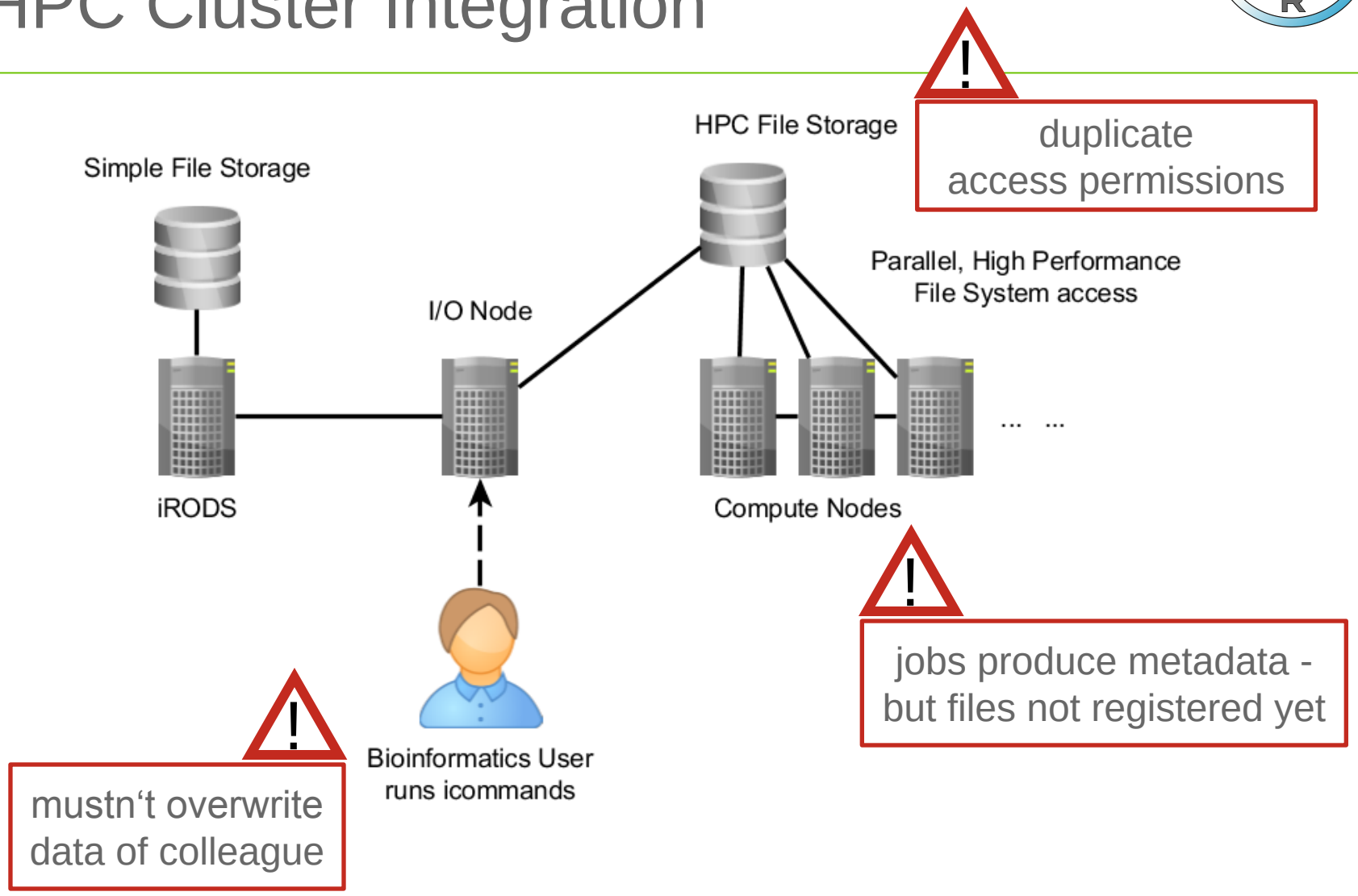
Metadata “Schema“ – Validation in Excel and iRODS:

Attribute Name	mandatory	fixed dictionary	multiple values allowed	intended on study level	intended on project level	intended on file level	definition
bayer_study_id	x			x			study ID within Bayer
contact_person				x	x		acting as backup for the data owner
cross_study_analysis_allow	x			x			statement by data owner, if data can be used for cross stud
data_classification	x	x		x			bayer data classification
data_source					x		"creator" of the data, i.e. external sequencing provider or int
data_type		x				x	i.e. unmapped or mapped sequence reads, variant calls, g
entry_date						x	entry date of data into Bayer

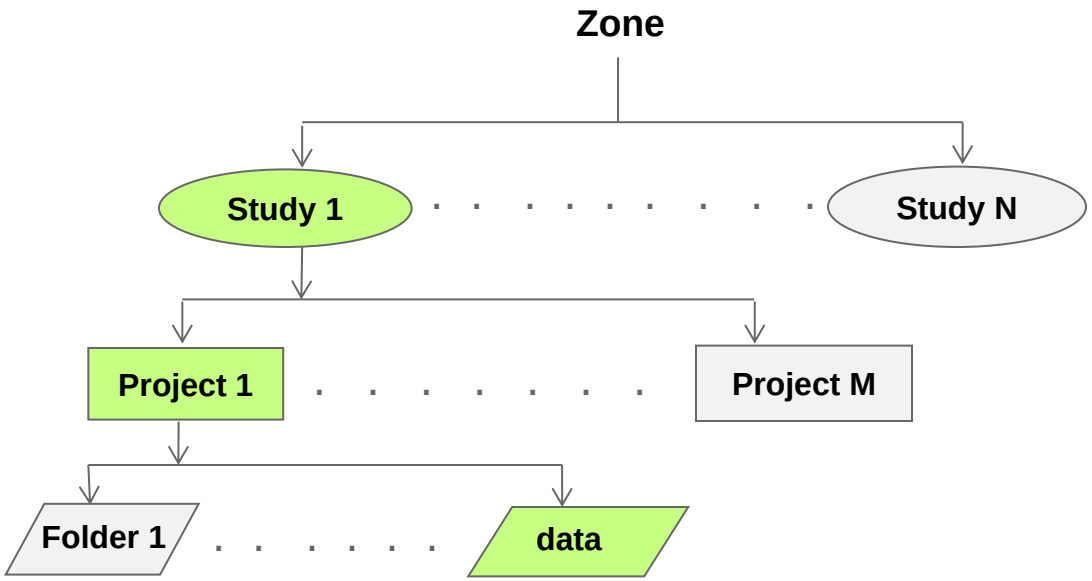
HPC Cluster Integration



HPC Cluster Integration



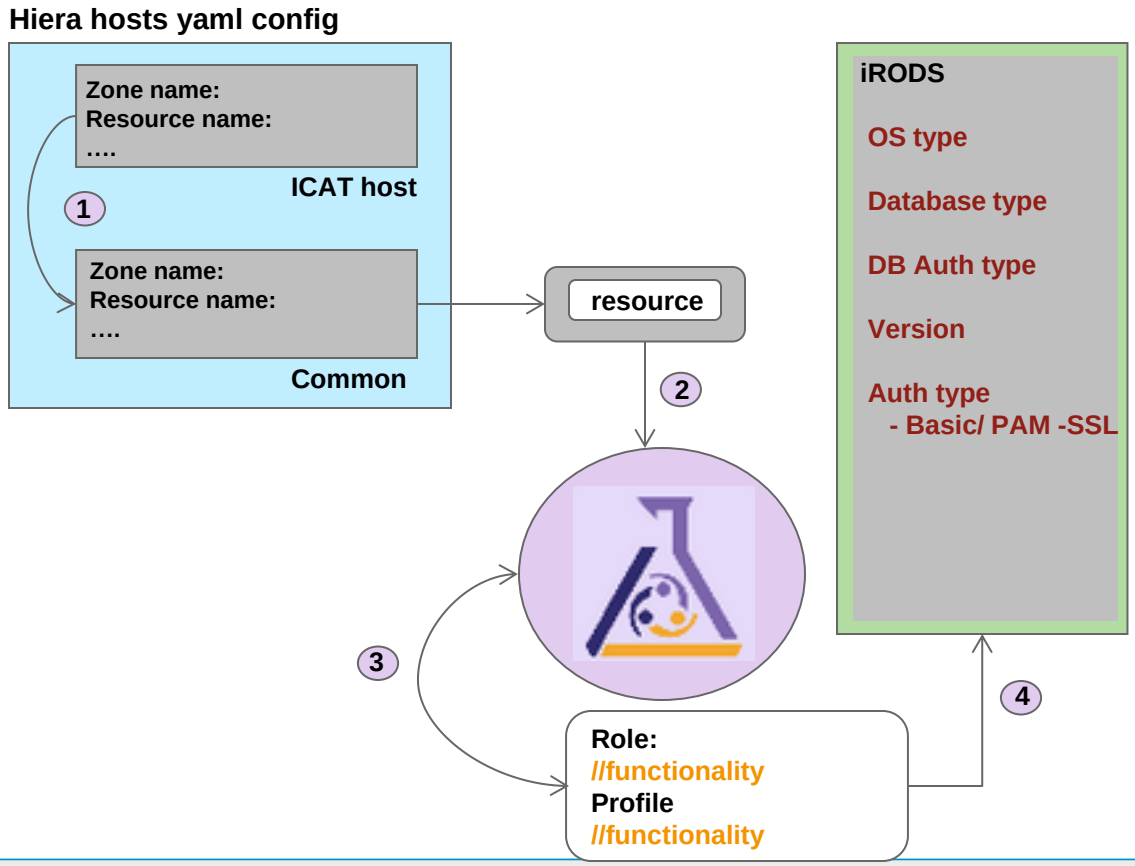
Custom Implementations



- Study based custom ACLs
- Hierarchical Inheritance
- Audit Trails
- Bulk Metadata Operations
- Metadata Validation
- Autom. Checksum Generation
- Data Integrity and Consistency

Installing iRODS with puppet

Puppet is a configuration management system that allows to define the state of the IT infrastructure, automates every step of the software delivery process



- Automated custom installation
- Supports multiple environments
- Easy to discard and rebuild installation on demand.
- Open topics:
  - support for version upgrades
  - security updates, patches

Technical Recommendations for Introducing iRODS

- involve Linux and storage admins early on
- plan effort for testing your installation
  your version of Linux and PostgreSQL, your iRODS rules, your expectations
  - document test procedures
  - deploy three similar environments: “development” / tryout system, acceptance test system, production system
- API selection
  evaluate all use case before deciding for programming platform and API

Organizational Recommendations for Introducing iRODS

- **get involved with iRODS support or the user community**
  save effort by incorporating external knowhow
  iRODS manual and help pages cover many aspects, but not all
- prepare a metadata and data access concept
- Metadata collection is a challenging task...
- user training – iRODS is not a distributed file system
  (e.g. replication – same file name, different content is possible)

? Questions & Answers !

Contact

Carsten Jahn
Bayer Business Services

Phone: +49 – 30 468 12837

E-mail: carsten.jahn@bayer.com

Navya Dabbiru
Tata Consultancy Services

Phone: +49 – 17680892216

E-mail: navyaanantayashasri.dabbiru@bayer.com

Forward-Looking Statements

This presentation may contain forward-looking statements based on current assumptions and forecasts made by Bayer Group or subgroup management.

Various known and unknown risks, uncertainties and other factors could lead to material differences between the actual future results, financial situation, development or performance of the company and the estimates given here. These factors include those discussed in Bayer's public reports which are available on the Bayer website at www.bayer.com.

The company assumes no liability whatsoever to update these forward-looking statements or to conform them to future events or developments.

Image Credits

Slide “Improving Data Management”

Kinghorn Centre for Clinical Genomics, HSXcutout

https://creativecommons.org/licenses/by-nd/2.0/

Uwe Hermann, Organized

https://creativecommons.org/licenses/by-sa/2.0/

Greg Emmerich, Vector DNA

https://creativecommons.org/licenses/by-sa/2.0/

Science For A Better Life

Thank you!