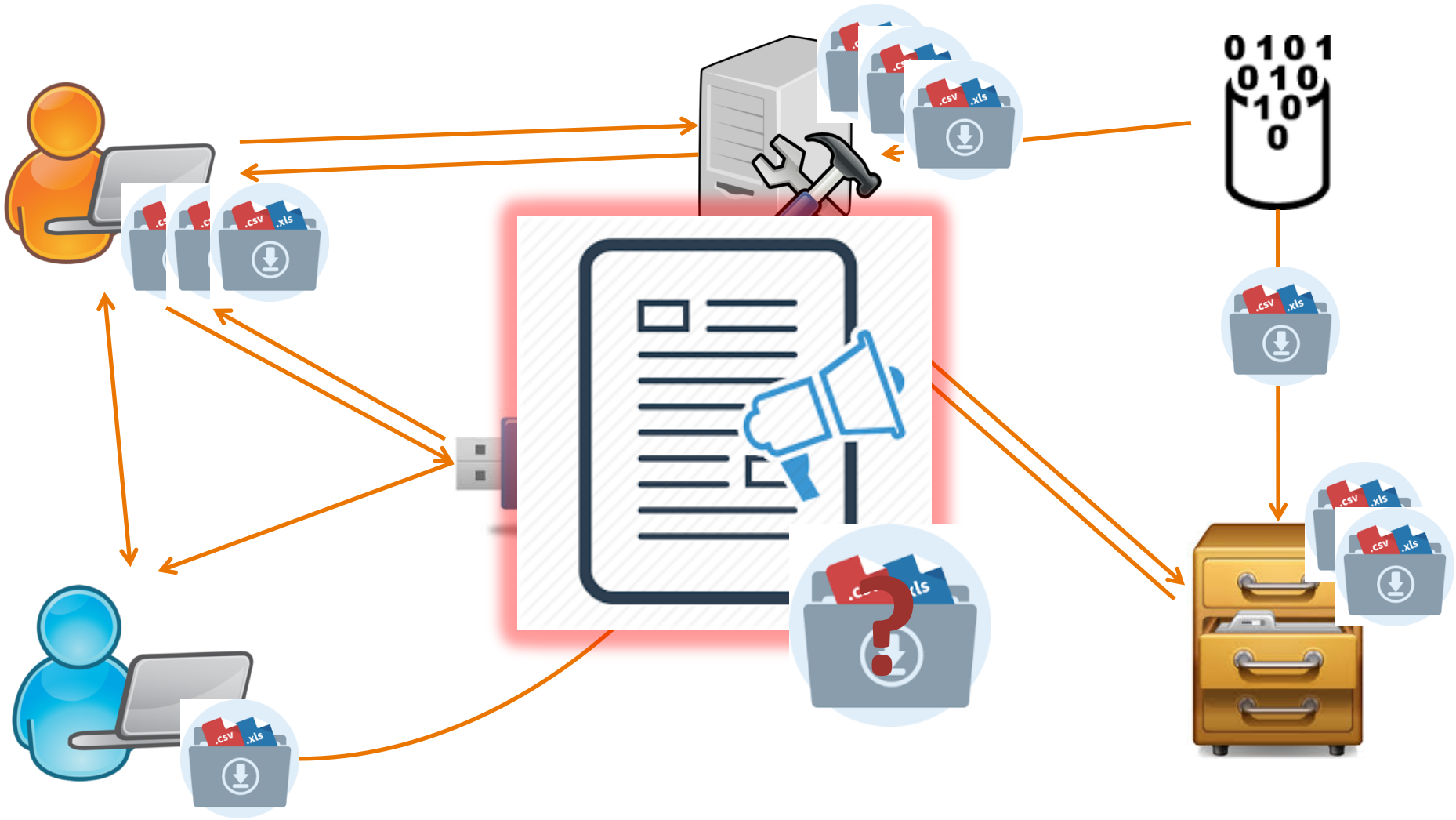
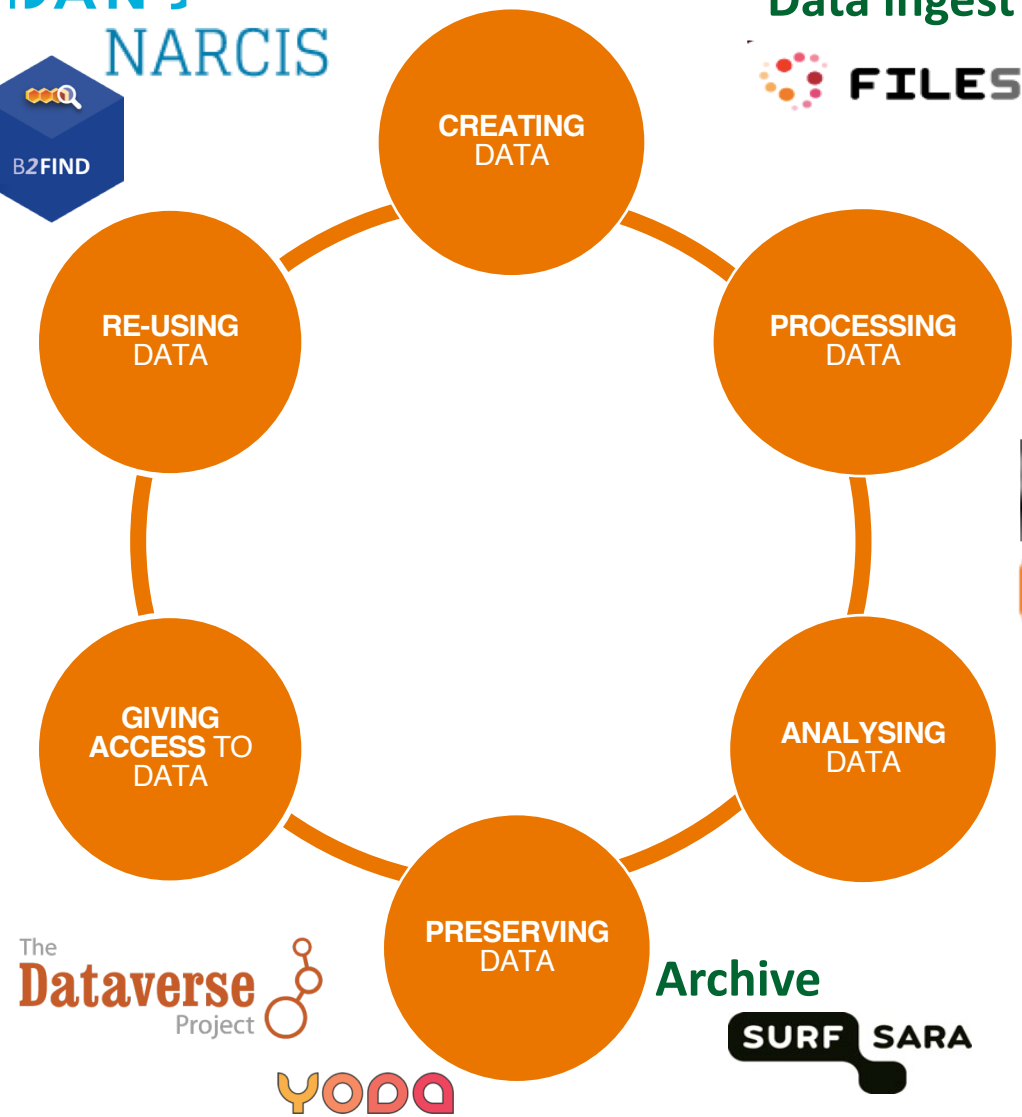
DATA MANAGEMENT AND COMPUTE WORKFLOWS

SURF SARA

Data – where is the problem?



Supporting the Data Life Cycle



Data ingest service

FILESENDER

SURF RESEARCHDRIVE
SURF DRIVE

Archive

SURF SARA

SURF RESEARCHDRIVE
SURF DRIVE

FILESENDER

The Dataverse Project

YODA

SURF SARA

The current challenge for users

Offer:

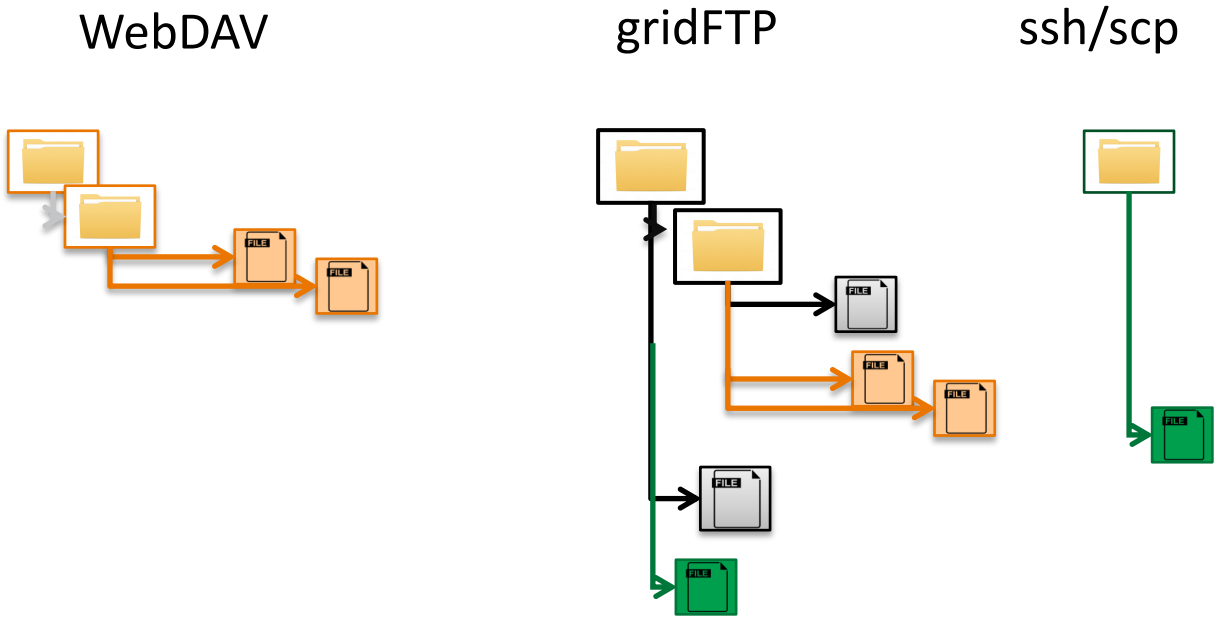
- Different storage media
- Different protocols to steer data transfers
- Different clients and user interfaces
- Different services

→ Accommodating various needs and use cases

Requires users to:

- Proper book-keeping
- Knowledge about storage infrastructure

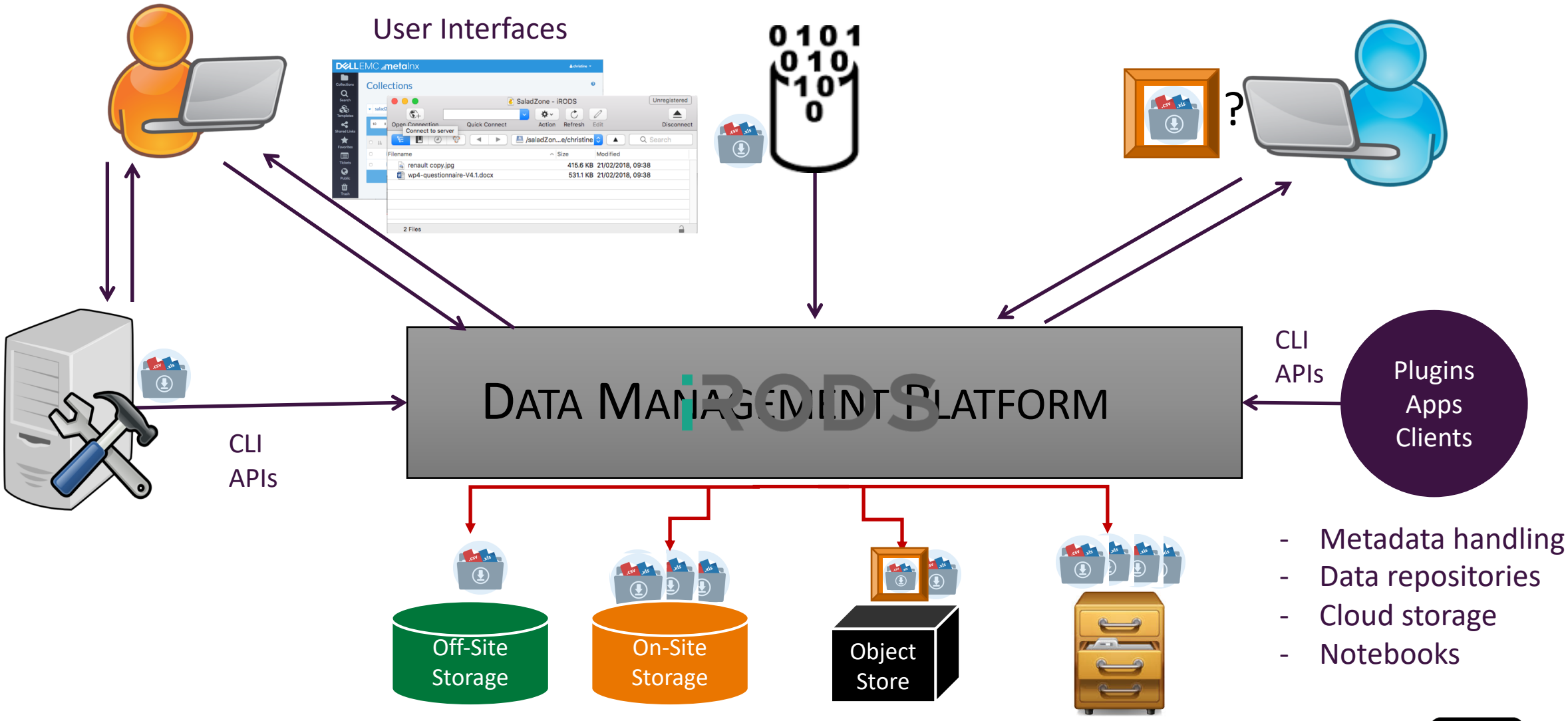
When to use what?



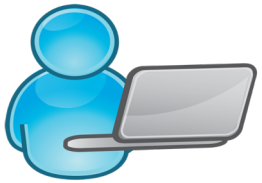
Goal:

- One entrance point to data
- Good interface to compute services
- Configurable data policies per use case or community

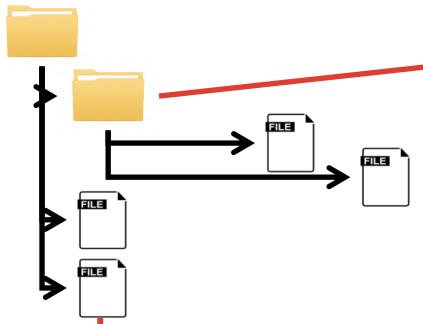
Data Management Platform and related Services



The iCAT – iRODS catalogue



User information:
Name
Groups
Access rights
Roles

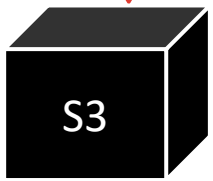


+ Extra information:
attribute: distance
value: 12
units: miles
---
attribute: author
value: Alice
units:

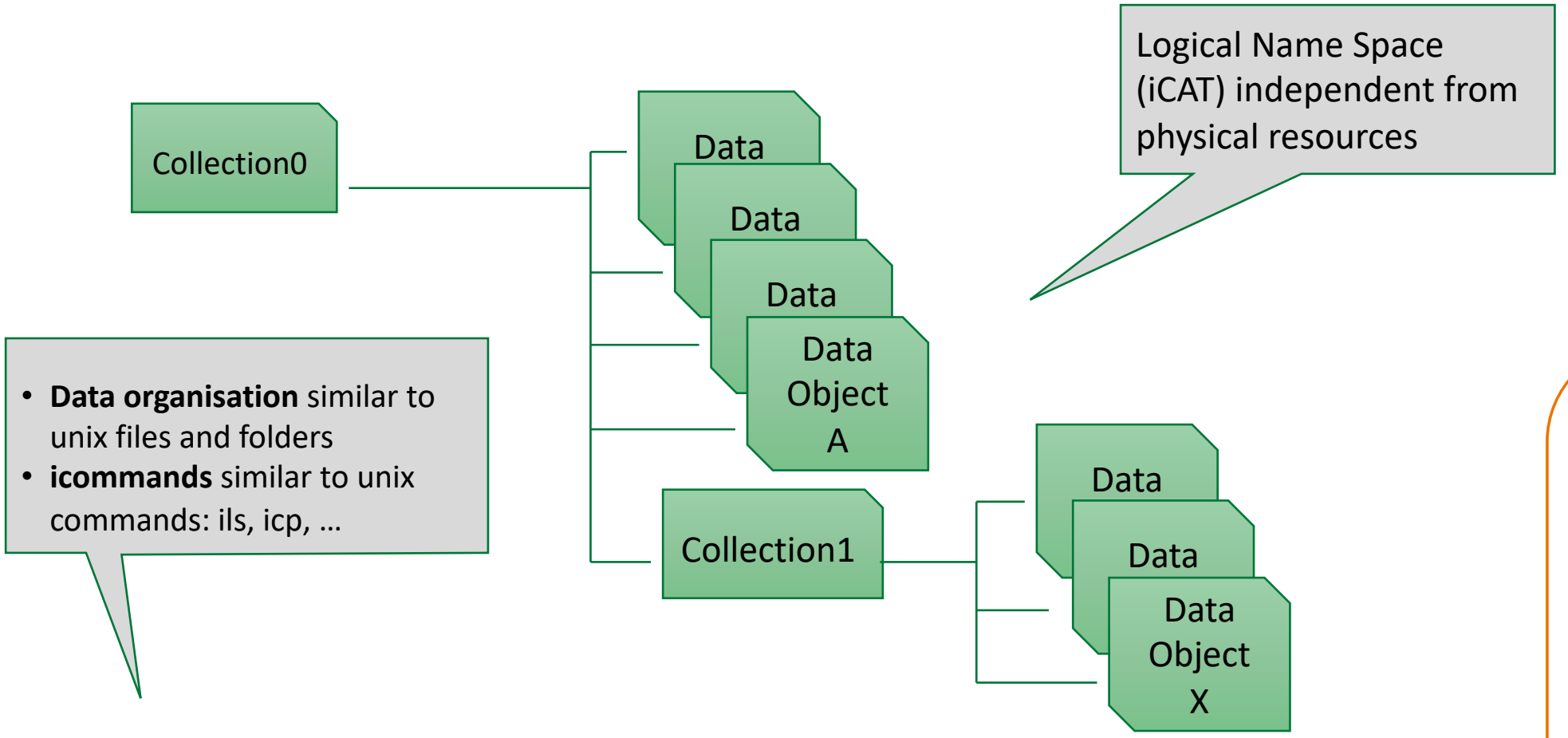
Abstraction layer:
Mapping from logical to physical namespace

iCAT (iRODS metadata catalogue):
/irodsZone/home/user/Collection0/testfile.txt →
/irodsVault/home/rods/testfile.txt

Storage layer:
Different storage media
Different protocols to steer data



The users' view



/zone/home/user/Collection0/
…/Collection0/DataObject A
…/Collection0/Collection1/
…/Collection0/Collection1/DataObject X

iRODS is ..

- Not meant for ad-hoc data management
- Overkill when only one storage resource is needed
- Not a monitoring system (although it provides sufficient logging information)
- Not an accounting system (although it maps storage usage to users and groups)
- Not a synchronisation or data publishing tool

  (although it can be used for such implementations)

Use iRODS for …

- Data management across different administrator domains

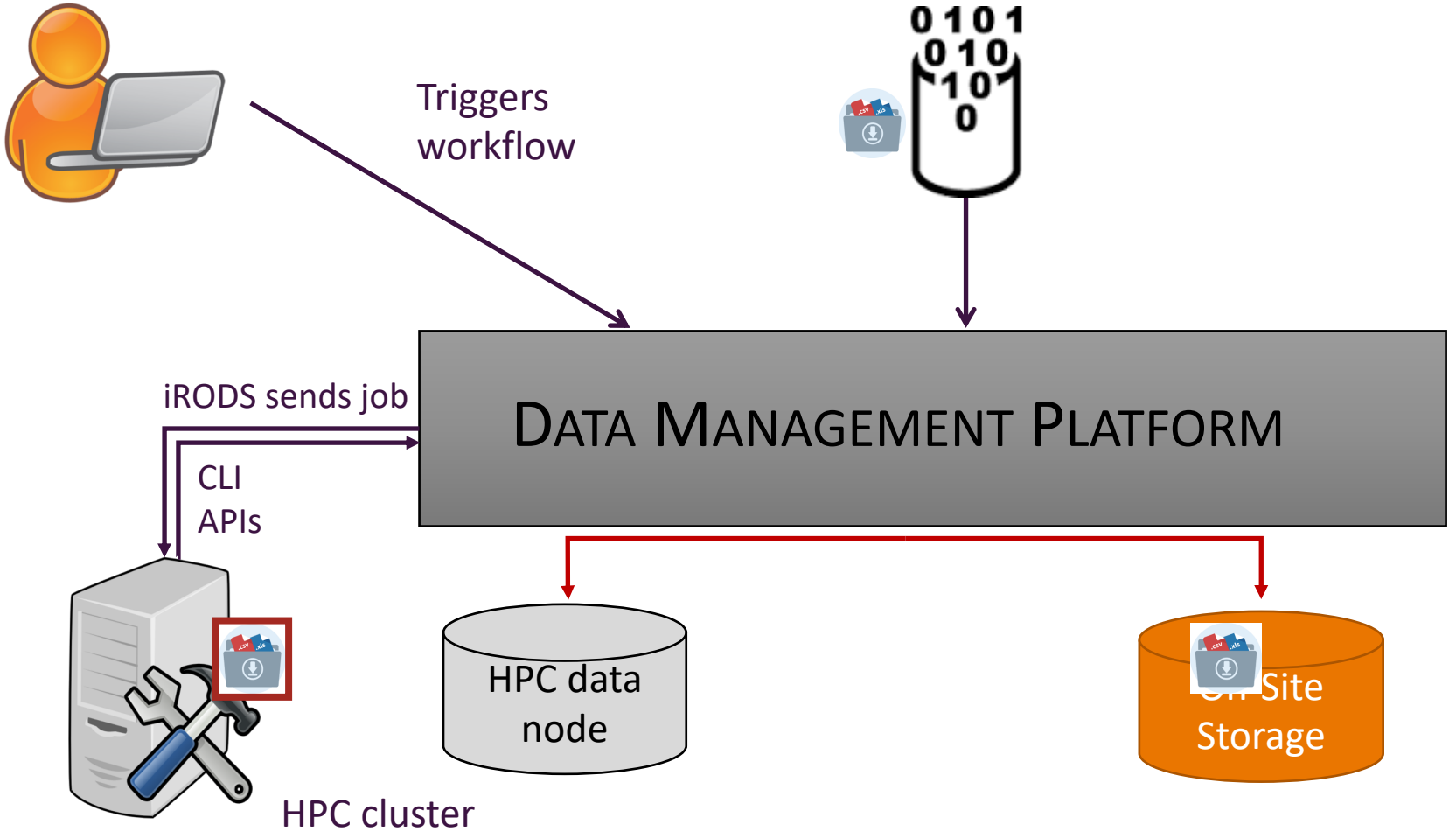
- Integration of different data services, modelling transition policies
- Combinaton of several storage systems that you want your users to access and steer in a uniform way

  → spare users to employ tons of different protocols

- Scaling storage out easily → simply plug in a new storage system

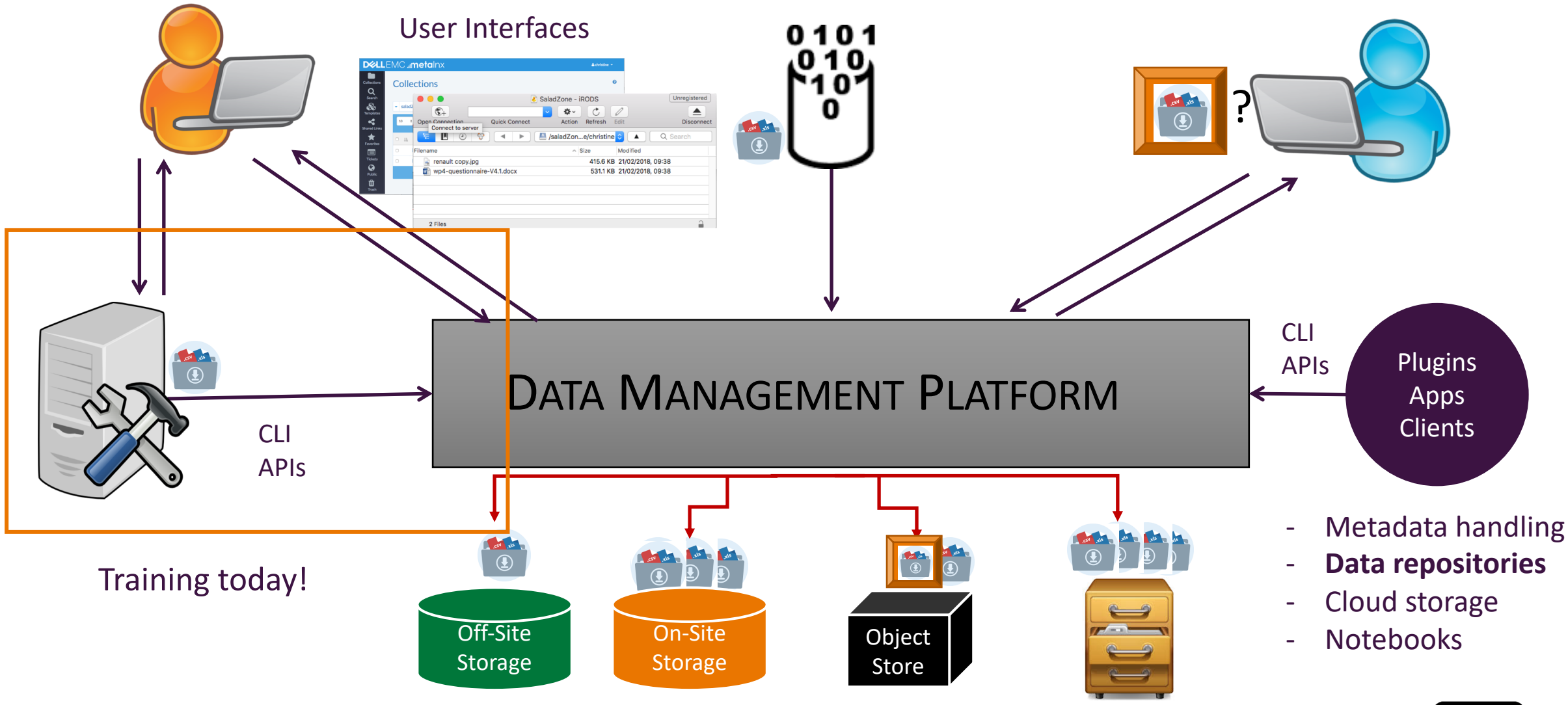
- Automatisation of data management workflows, e.g. metadata extraction, automatisation of recurrent actions

Compute integration



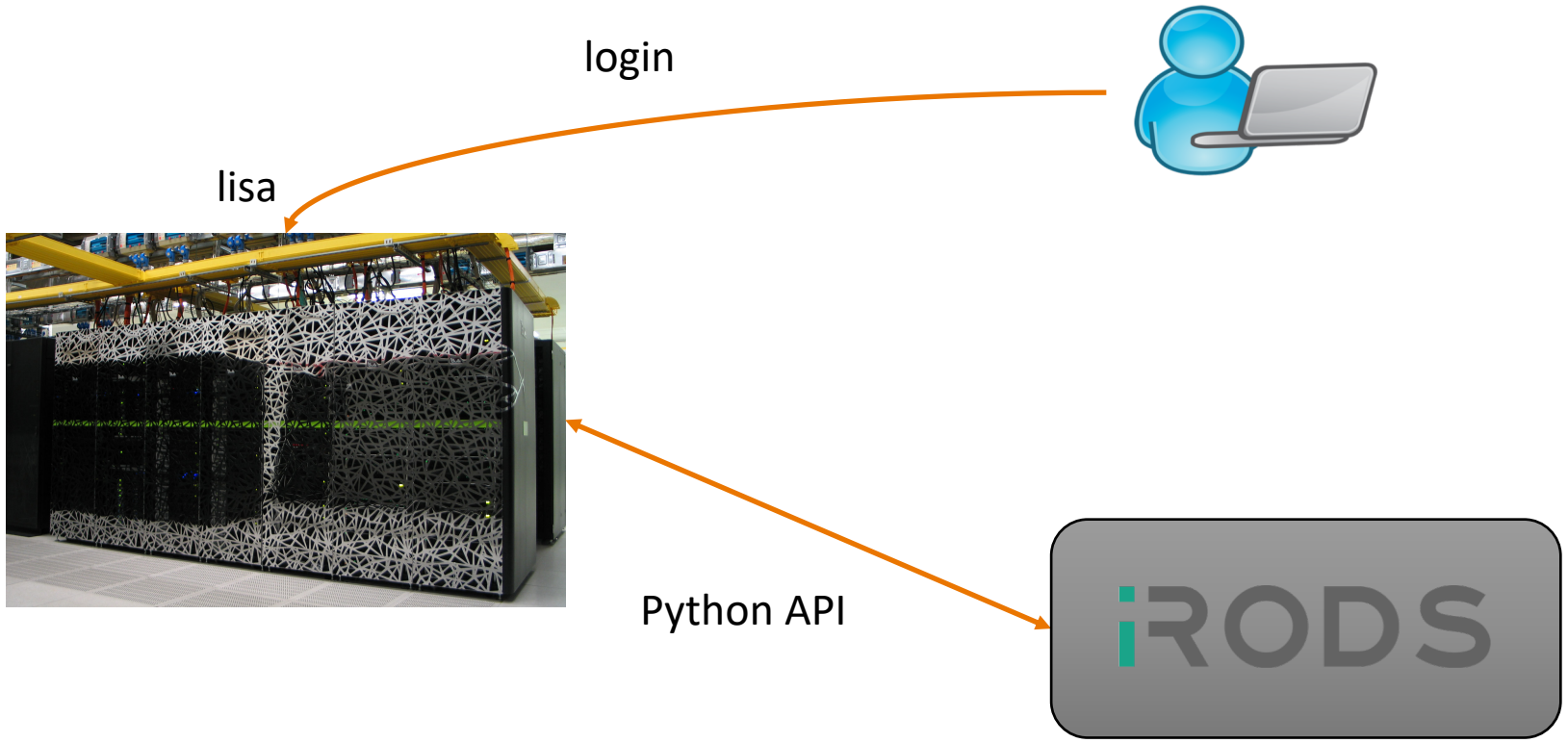
- Data node needs to be managed by iRODS as resource server
- Future work
  - iRODS consortium is working on Lustre plugin
  - Test plugin with the Dutch national HPC cluster

Compute integration



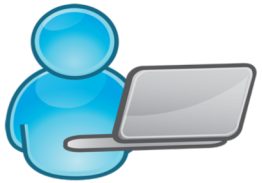
- Metadata handling
- **Data repositories**
- Cloud storage
- Notebooks

Training setup – A simple workflow



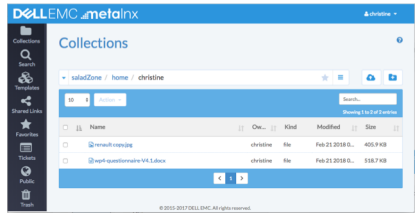
1. Search for data in iRODS
2. Download data to lisa
3. Some calculations
4. Upload results to iRODS
5. Annotate results

Training setup – Remote workflow execution

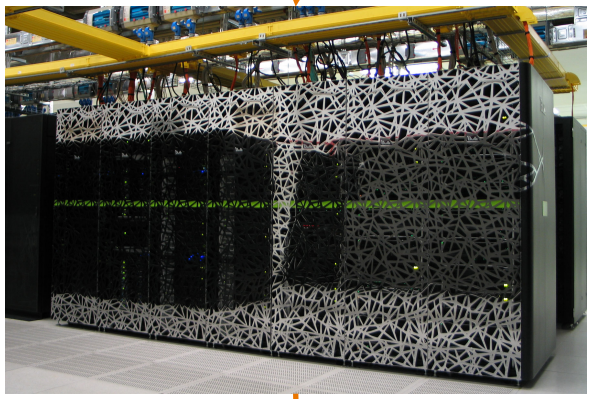


Log in

Return results to user

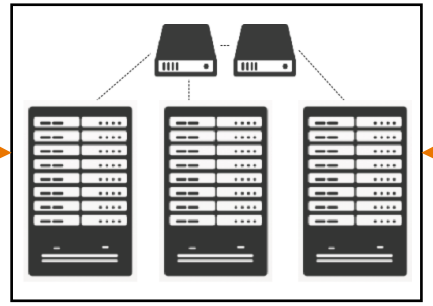


Lisa



iRODS

Compute nodes



Start jobs via PBS

Python script

1. Search for data in iRODS
2. Download data to lisa
3. Some calculations
4. Upload results to iRODS
5. Annotate results

SURF SARA

Data management services:

Hylke Koers

AAI:
Fatih Turkmen
Gerben Venekamp
Harry Kodden

DM:
Arthur Newton
Christine Staiger
Sharif Islam
Stefan Wolfsheimer

Slides based on PRACE iRODS training:

Zheng Meyer-Zhao, Arthur Newton, Christine Staiger
Data Management with iRODS, Sept 2017
https://events.prace-ri.eu/event/638/

Arthur Newton, Christine Staiger (SURFsara)
Arthur.newton(at)surfsara.nl, Christine.staiger(at)surfsara.nl

SURF SARA