Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

ICSC
Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

ICSC
Centro Nazionale di Ricerca in HPC,
Big Data and Quantum Computing

Data Management Architecture: the testbed

B.Spisso

Mini-Workshop on Data Management
5 July 2024

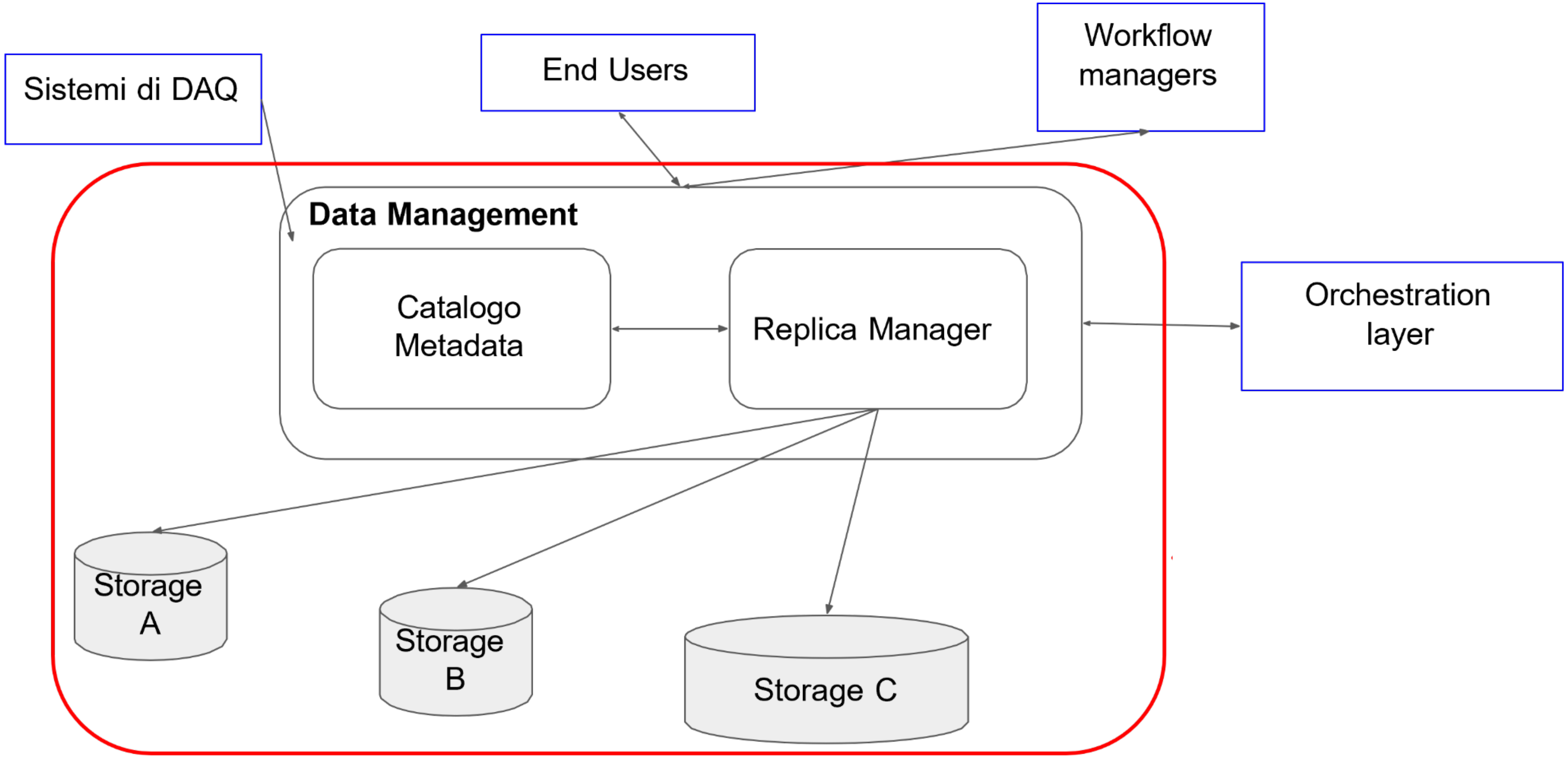
Recap: What is data management

- Data management is the practice of keeping and using data securely, efficiently, and cost-effectively.
- A robust data management solution becomes more necessary as the number of people accessing, generating, and sharing data increases across several sites.

Recap: Why a testbed for a national infrastructure and how should be?

- **Federation of storage with heterogeneous technologies (**ie**.** Both "Grid" and "Cloud“)
- **Abstraction of the "logical" level from that of storage management**
- **A way to implement a data locality strategy**
- **Allows interfaces at various levels** (ie. for end user, or admin)

The Data-Lake model



Storage & Data Magement, M. Sgaravatto and D. Spiga, 2022

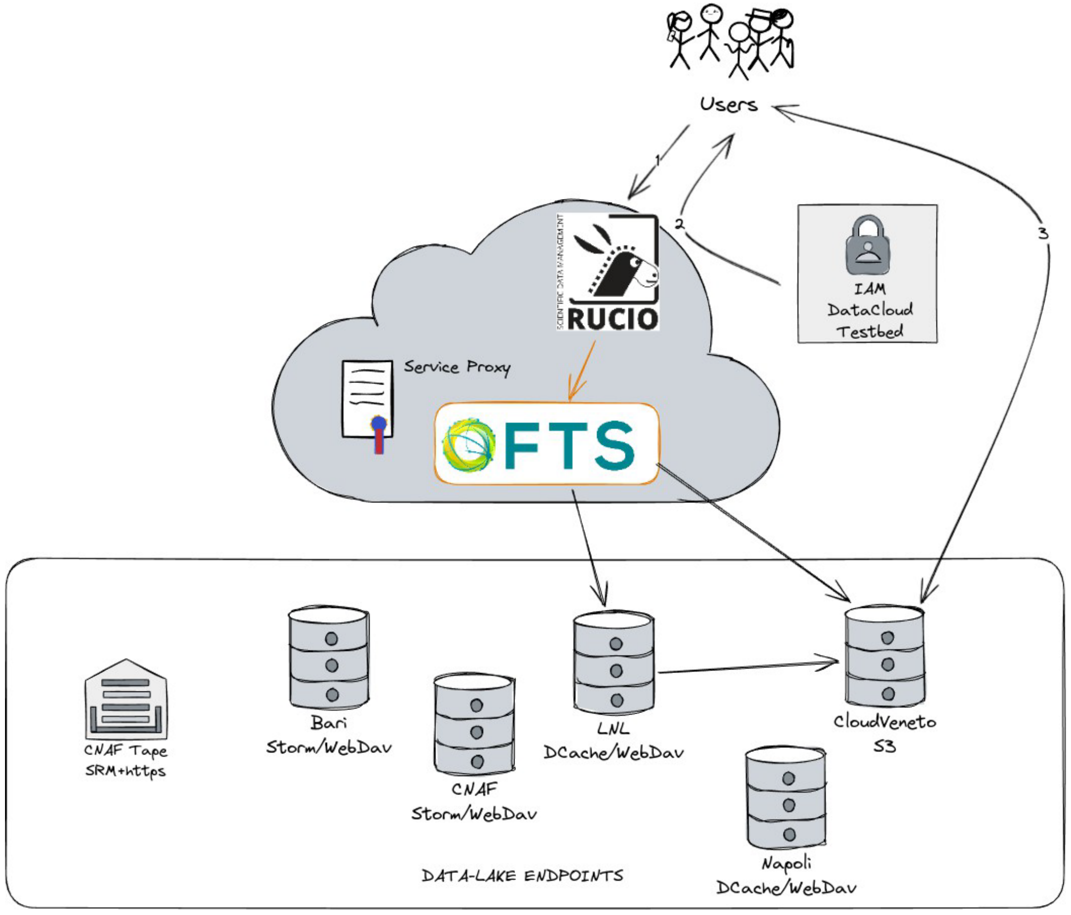
The testbed in DataCloud

We have chosen to start the experimentation by integrating **de-facto** standard tools already in use in scientific realities close to us and that we are familiar with (e.g., LHC):

- **Rucio+FTS** (Data manager)
- **IAM** (AuthN/Z)
- Metadata Catalog: embedded in **Rucio**

6 heterogeneous storage systems of INFN:

- Qos (disk, tape);
- One storage with S3 protocol on ceph @CloudVeneto
- Three storages with WebDav protocol
  - Two based on STORM (CNAF, Bari)
  - Two on dCache (LNL, Naples)
- One tape endpoint @CNAF



Federare lo storage distribuito nazionale, D. Ciangottini, 2023

What is capable of?

- **Storage:** Different sites are federated regardless of their geographical location, their implementation, or their QoS (disk or tape).
- **Users**: can interact with the data in a declarative way: for example, it can declare how many replicas are needed for a certain file, on how many and which storage systems, and for how long they must exist.
- **The Data:** can be organized hierarchically (datasets, containers).
- **Transparency**: the users can upload and read data in the various endpoints without worrying about the protocols used. As simple as Rucio upload/download <MY_FILE> <MY_STORAGE>. There is no need to worry about scp, gridftp, rsync, etc

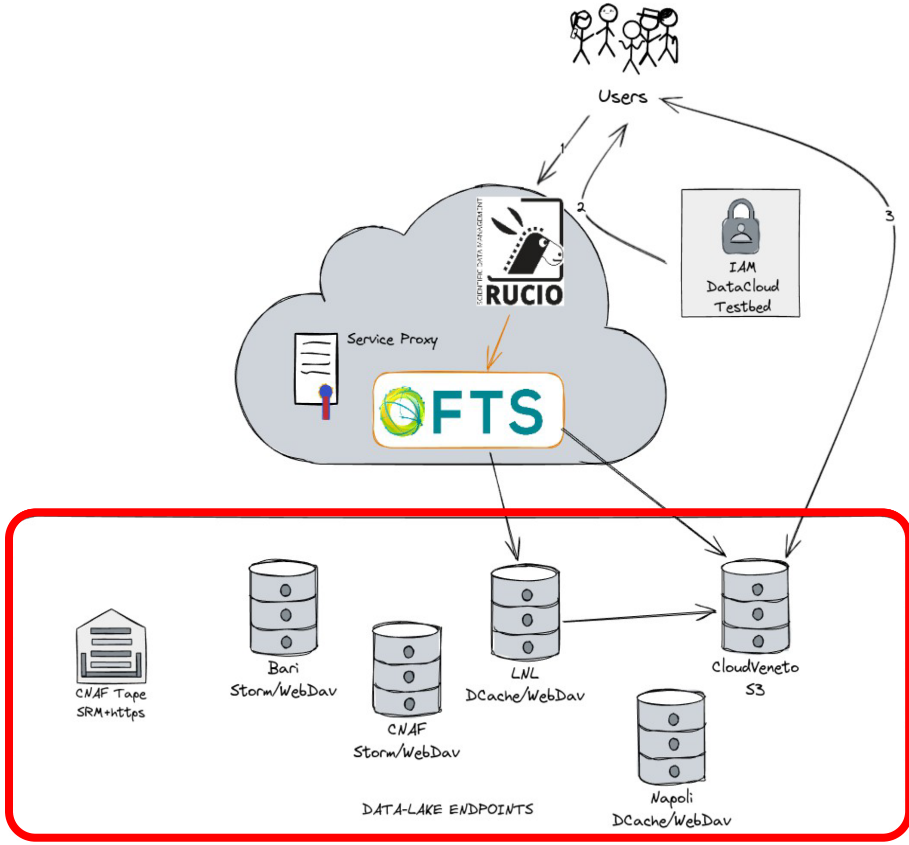
Storage endpoints, the dCache example

- It is where the data is physically stored.
- The raw space is by a local storage manager which typically can aggregate different storage units
- The are different storage manager with different technologies (dCache, Storm, EOS, S3, Minio,…)
- For example, Naples dCache manages 1 PByte of raw storage dived among two storage
- dCache is capable to offer various access protocol (xroot, gsiftp, pnfs, WebDav…)
- For the testbed storage space is used WebDav therefore becomes an Object storage

Object storage is a data storage architecture that stores and manages unstructured data in units called objects.

Can be accessed standalone via CLI using GFAL2 framework

```
$ gfal-copy https://t2-dcache-02.na.infn.it:443/<file>
            https://t2-dcache-02.na.infn.it:443/<file>
```



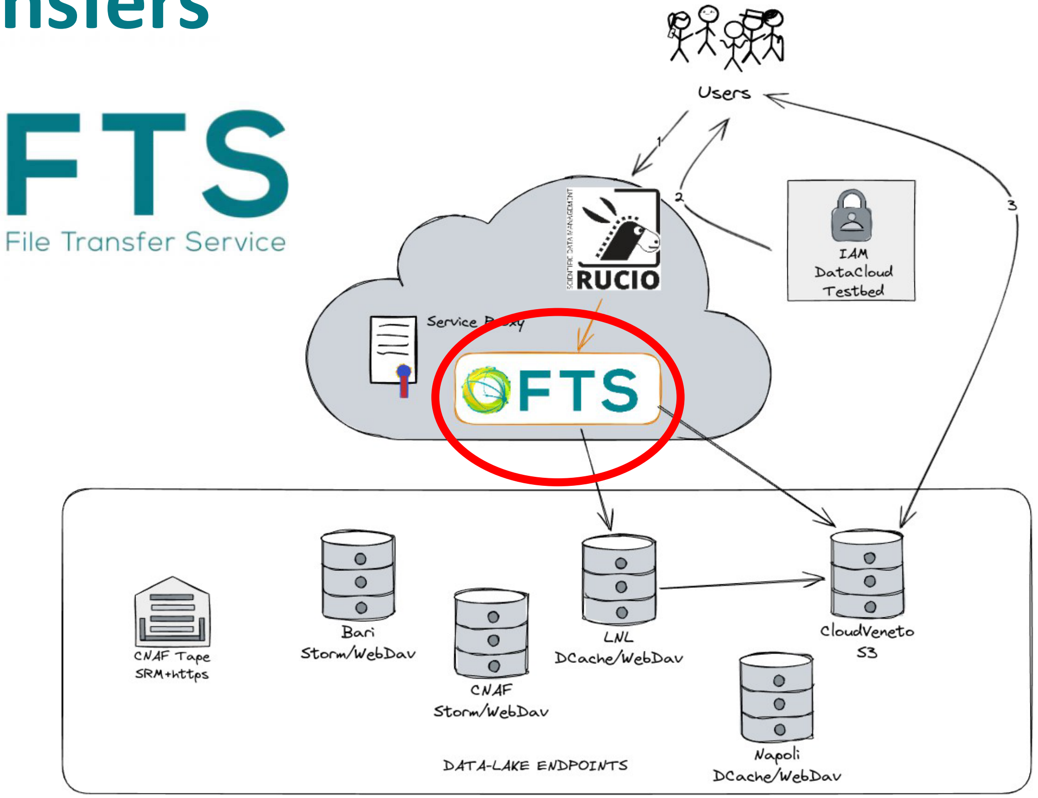
What is FTS?

FTS is an open-source software for large scale queuing and reliable execution of file transfers

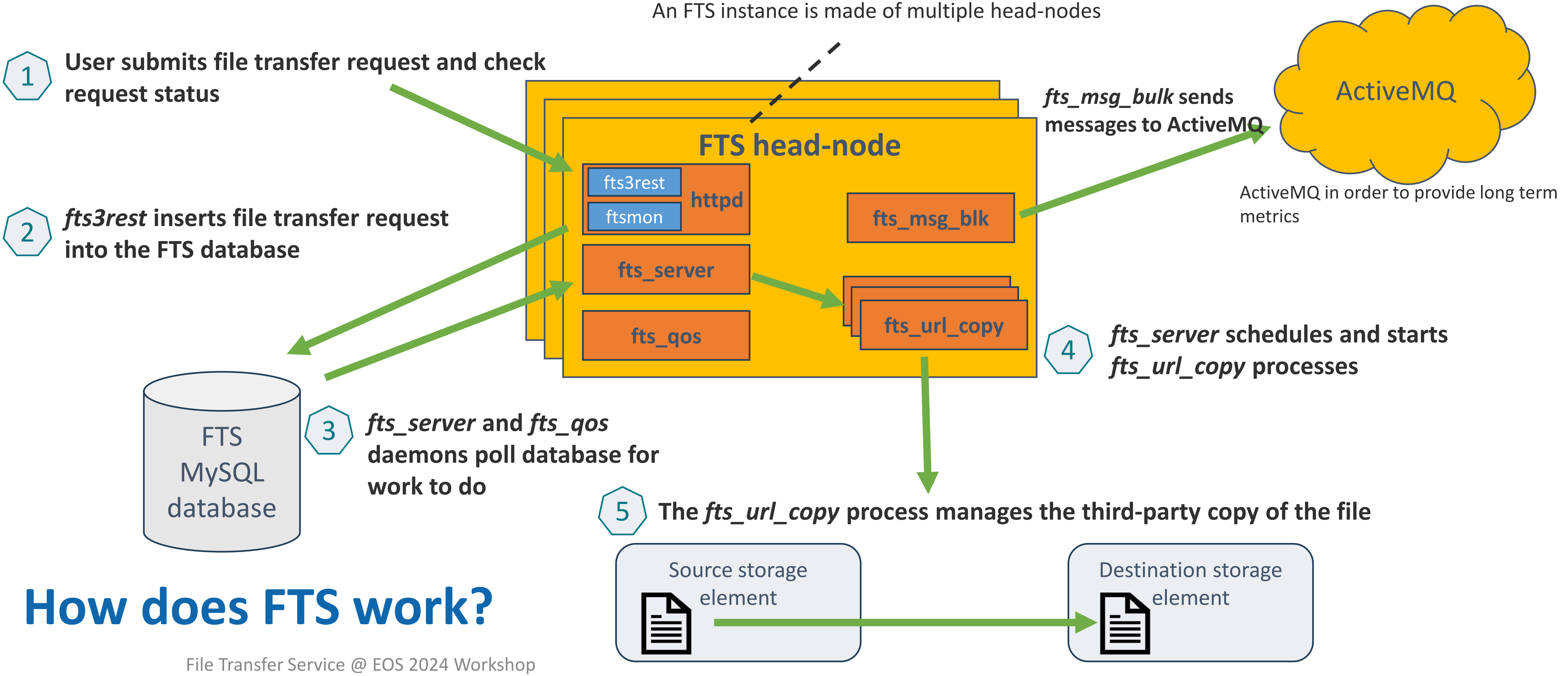
Capabilities:

- Orchestration of Third-Party Copies (TPCs)
- Streams transfers through itself if TPC is not supported
- Tape storage operations via the WLCG HTTP Tape REST API, SRM and XRootD
- Support for Cloud based storage
- Certificate and token authentication

FTS
File Transfer Service



How does FTS work?



An FTS instance is made of multiple head-nodes

1 **User submits file transfer request and check request status**

2 ***fts3rest* inserts file transfer request into the FTS database**

3 ***fts_server* and *fts_qos* daemons poll database for work to do**

4 ***fts_server* schedules and starts *fts_url_copy* processes**

5 **The *fts_url_copy* process manages the third-party copy of the file**

***fts_msg_bulk* sends messages to ActiveMQ**

ActiveMQ in order to provide long term metrics

FTS uses Gfal2

- All FTS ↔ storage interaction is done indirectly via the Gfal2 library
- Gfal2 (Grid File Access Library) provides a common top-level file API

  … but supports multiple protocols behind-the-scenes

- Supported protocols include:
  - HTTP/Webdav
  - Cloud storage (S3, Swift, GCloud)
  - Xrootd
  - SRM
  - GridFTP
  - Local file

How to use FTS?

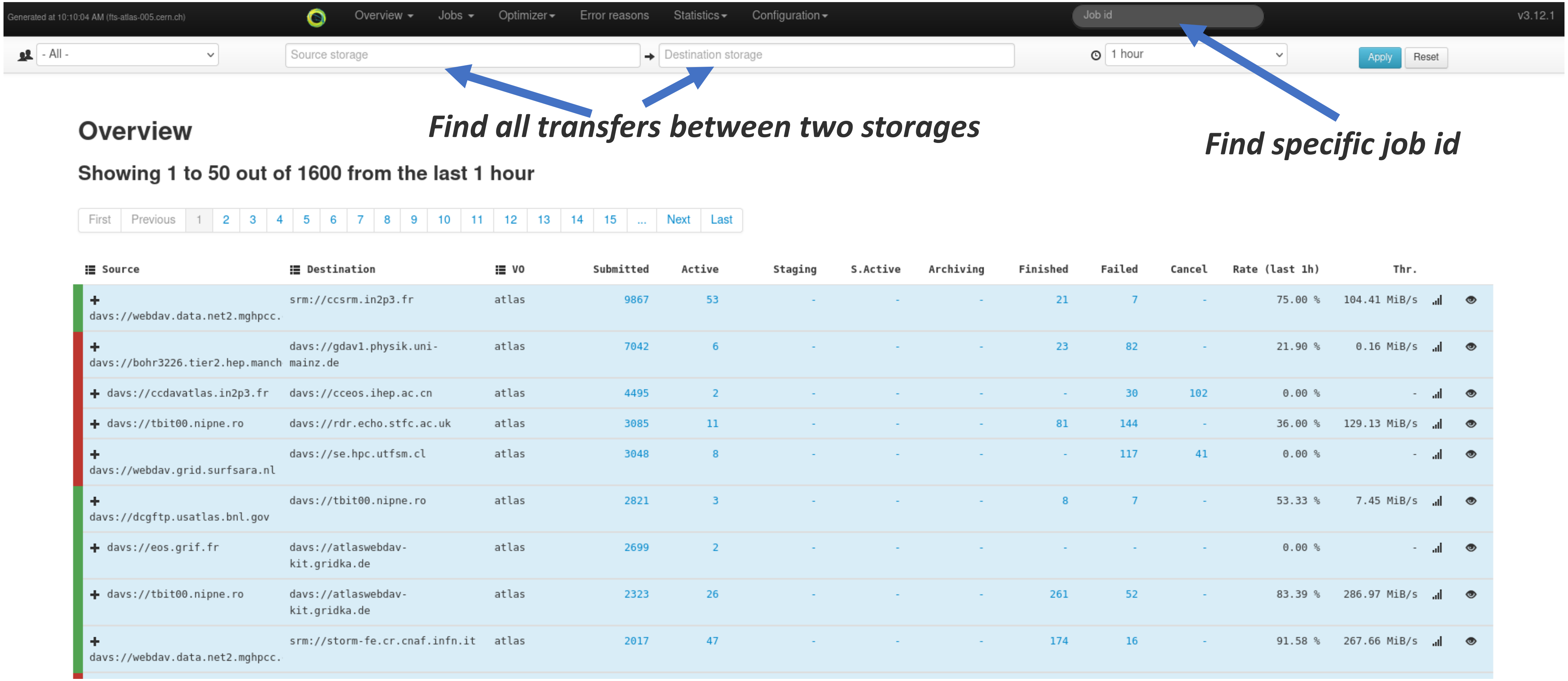
- FTS provides a REST API for transfer submissions and querying its status
- Dedicated CLI clients (`$ fts-rest-transfer-submit`)
- Python 3 bindings (`$ python3 -c ‘import fts3; transfer = fts3.new_job(..)’`)
- Direct JSON submission via `/jobs` endpoint

```
$ fts-rest-transfer-submit -s https://fts3-pilot.cern.ch:8446/
      https://eospublic.cern.ch:443/<path> https://eosatlas.cern.ch:443/<path>

$ fts-rest-transfer-status -s https://fts3-pilot.cern.ch:8446/
      d4e3dc36-f7c2-46f7-8f40-70981d9d539c

$ curl -X POST --cert <cert> --data=submission.json https://fts3-pilot.cern.ch:8446/jobs
```

FTS Web Monitoring:



Find all transfers between two storages

Find specific job id

Overview

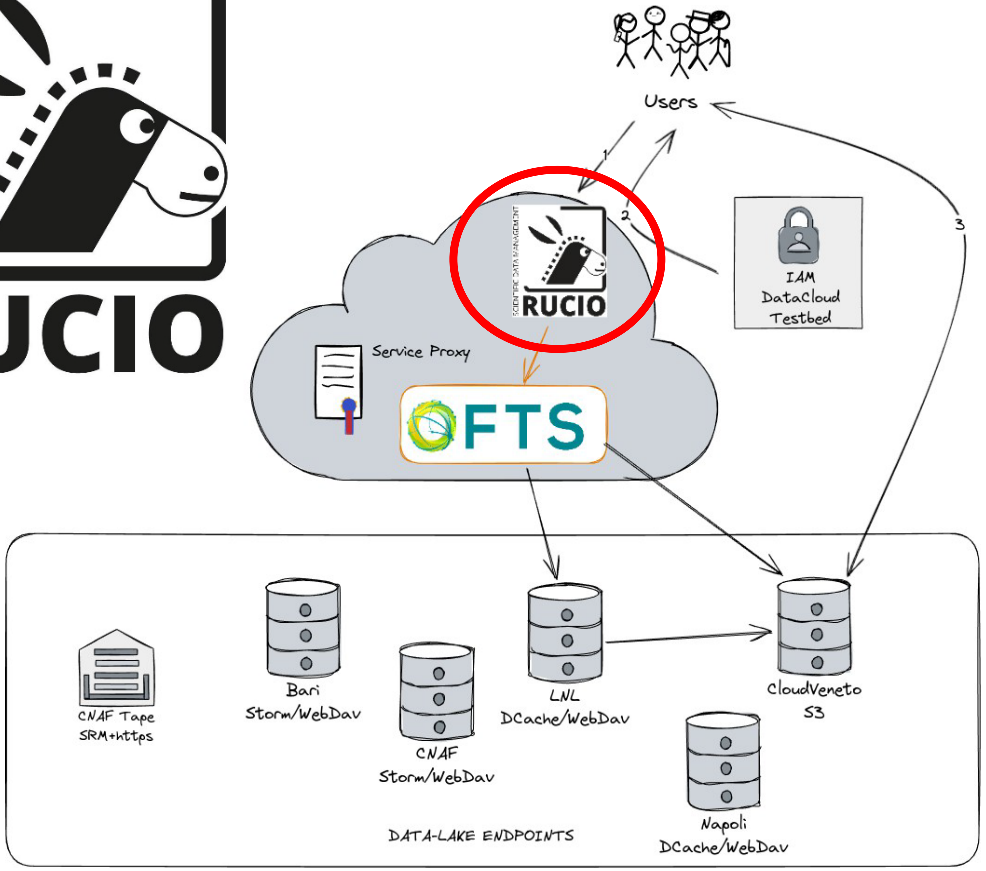
Showing 1 to 50 out of 1600 from the last 1 hour

Source	Destination	VO	Submitted	Active	Staging	S.Active	Archiving	Finished	Failed	Cancel	Rate (last 1h)	Thr.
davs://webdav.data.net2.mghpcc.	srm://ccsrm.in2p3.fr	atlas	9867	53	-	-	-	21	7	-	75.00 %	104.41 MiB/s
davs://bohr3226.tier2.hep.manch	davs://gdav1.physik.uni-mainz.de	atlas	7042	6	-	-	-	23	82	-	21.90 %	0.16 MiB/s
davs://ccdavatlas.in2p3.fr	davs://cceos.ihep.ac.cn	atlas	4495	2	-	-	-	-	30	102	0.00 %	-
davs://tbit00.nipne.ro	davs://rdr.echo.stfc.ac.uk	atlas	3085	11	-	-	-	81	144	-	36.00 %	129.13 MiB/s
davs://webdav.grid.surfsara.nl	davs://se.hpc.utfsm.cl	atlas	3048	8	-	-	-	-	117	41	0.00 %	-
davs://dcgftp.usatlas.bnl.gov	davs://tbit00.nipne.ro	atlas	2821	3	-	-	-	8	7	-	53.33 %	7.45 MiB/s
davs://eos.grif.fr	davs://atlaswebdav-kit.gridka.de	atlas	2699	2	-	-	-	-	-	-	0.00 %	-
davs://tbit00.nipne.ro	davs://atlaswebdav-kit.gridka.de	atlas	2323	26	-	-	-	261	52	-	83.39 %	286.97 MiB/s
davs://webdav.data.net2.mghpcc.	srm://storm-fe.cr.cnaf.infn.it	atlas	2017	47	-	-	-	174	16	-	91.58 %	267.66 MiB/s

What is RUCIO?

- Data management tool
  - Integrates with many storage solutions
  - Data can be stored across multiple sites, with different setups and protocols
  - Data can be anything, images, text….

SCIENTIFIC DATA MANAGEMENT
RUCIO



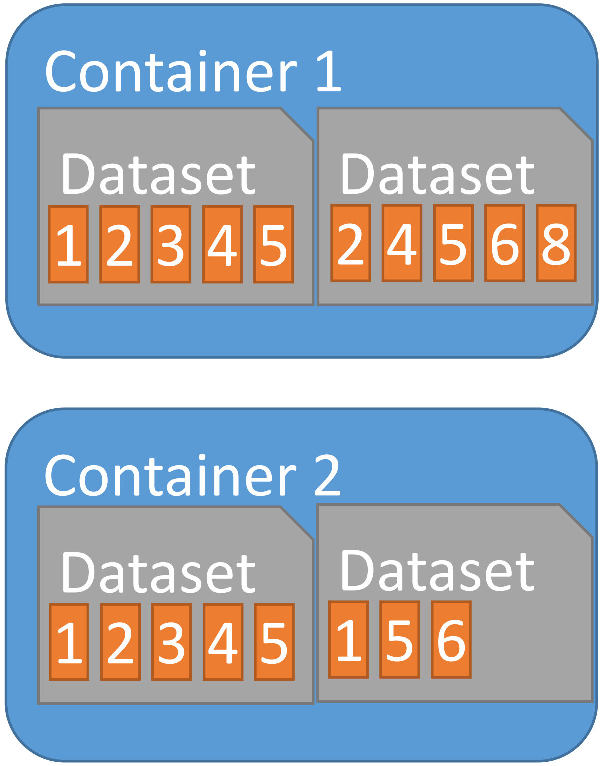
Rucio in a nutshell

- Initially developed by the ATLAS experiment
- Provides services and libraries for scientific collaborations/experiments/communities
    - Designed with more than 10 years of operational experience in data management
    - Full, complete and generic data management service
    - The number of data intensive instruments generating unprecedented data volume is growing

- Store, manage, and process data in a heterogeneous distributed environment
    - Data can be scientific observations, measurements, objects, events, images saved in files
    - Manage transfers, deletions, and storage
    - Connects with workflow management systems
    - Supports both low-level and high-level policies and enforces them
    - A rich set of advanced features and use cases supported

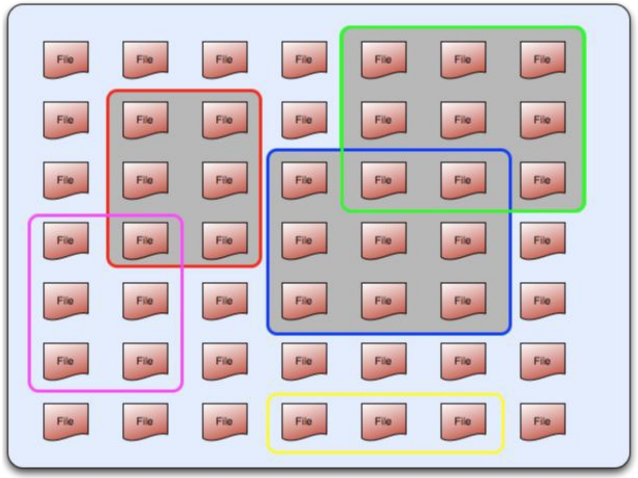
Rucio Files, Datasets and Containers

- **Single files can be replicated using rules**
- **Files are grouped together in datasets**
  - Can belong to multiple datasets
- **Containers are collections of datasets**
- **Containers and datasets can have properties to protect datasets**
  - E.g. Open/closed – can have data added



Namespace handling

- **Data Identifier (DID) is the primary addressable unit**
  - DIDs can be either files, collections (*datasets*), or collections of collections (*containers*)
  - Datasets only hold files, containers only hold datasets
- **DIDs are standalone**
  - Files do not need to be in a dataset
  - Datasets do not need to be in a collection
- **DIDs are globally unique**
  - Files cannot have the same name as collections, and vice versa
  - Cannot reuse names of deleted DIDs
    - Why? Prevents reuse of modified files for consistently repeatable science results
- **Collections can be organised freely**
  - Files can be in multiple datasets, datasets can be in multiple containers



Namespace handling

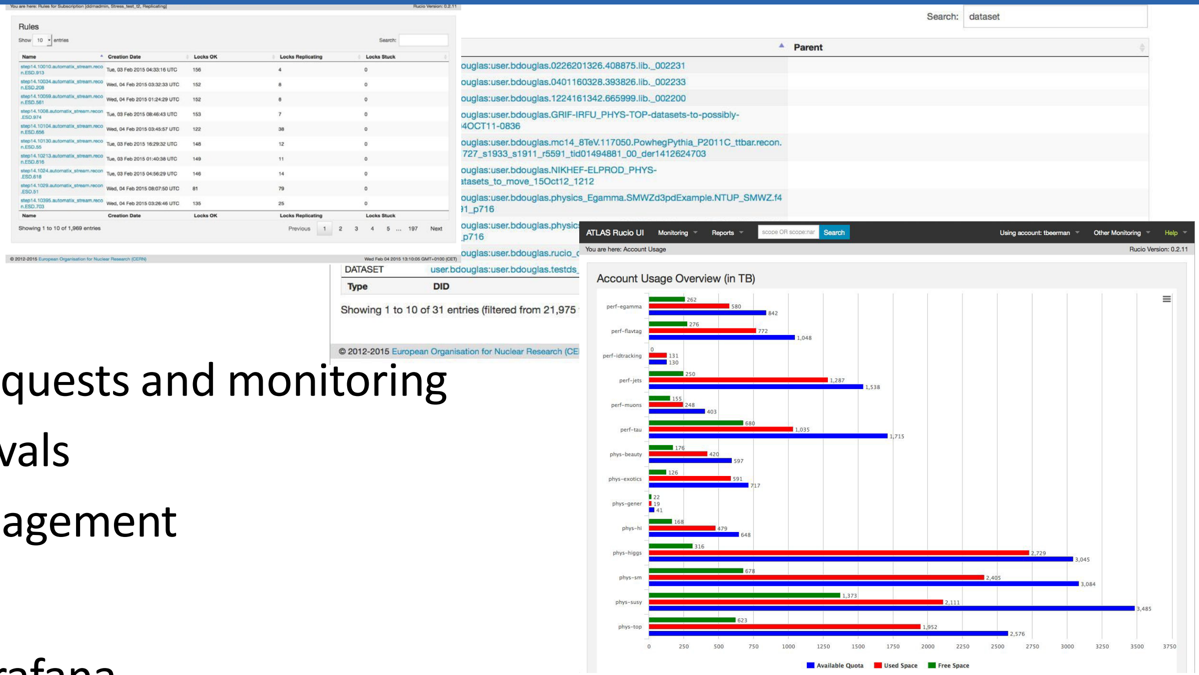
- **The global namespace containing all DIDs can be partitioned (into *scopes*)**
  - At least a single partition must exist (i.e., fallback global)
  - Distinguish different communities, users, groups, or activities (*user.jdoe, group.phys-higgs, …)*
  - Also helps with namespace scalability
- **DIDs are thus always tuples *<scope>:<name>***
  - Cannot have DIDs with *<name>* alone
  - Corollary: Names must be unique inside a scope only, whereas DIDs are globally unique
- **Example**
  - **FILE** *user.jdoe:my-analysis-data-123.tar.gz*
    user.jdoe:susy-analysis-script.py
  - **DATASET** *user.jdoe:run-123* [**contains**: *user.jdoe:my-analysis-data-123.tar.gz, …*]
  - **CONTAINER** *user.jdoe:all-my-runs* [**contains**: *user.jdoe:run-123, …*]
  - **CONTAINER** *group.phys-higgs:all-user-analy* [**contains**: *user.jdoe:run-123, …*]

Storage abstraction

- Rucio Storage Elements (RSEs) are a logical entity of space
  - No software needed to run at the site
  - RSE names are arbitrary (e.g., "CERN-PROD_DATADISK", "AWS_REGION_USEAST", …)
- RSEs collect all necessary metadata for a storage
  - protocols, hostnames, ports, prefixes, paths, implementations, …
  - data access priorities can be set (e.g., to prefer a protocol for LAN access)
- Existing data on storage can be registered into RSEs
- Express what you want with rules
  - *"Three copies of this dataset, distributed evenly across three institutes on different continents, with two copies on DISK and one on TAPE"*
  - Support for different data policies, e.g.

Archive:	difficult/expensive to recreate data
Primary cache:	data that should be readily available, job inputs/outputs, …
Secondary cache:	extra replicas created and deleted based on system usage for performance

Monitoring & Analytics



You are here: Rules for Subscription (ddmadmin, Stress_test_t2, Replicating)   Rucio Version: 0.2.11

Rules

Show 10 entries   Search:

Name	Creation Date	Locks OK	Locks Replicating	Locks Stuck
step14.10010.automatix_stream.recon.ESD.913	Tue, 03 Feb 2015 04:33:16 UTC	156	4	0
step14.10034.automatix_stream.recon.ESD.208	Wed, 04 Feb 2015 03:32:33 UTC	152	8	0
step14.10059.automatix_stream.recon.ESD.561	Wed, 04 Feb 2015 01:24:29 UTC	152	8	0
step14.1008.automatix_stream.recon.ESD.974	Tue, 03 Feb 2015 08:46:43 UTC	153	7	0
step14.10104.automatix_stream.recon.ESD.656	Wed, 04 Feb 2015 03:45:57 UTC	122	38	0
step14.10130.automatix_stream.recon.ESD.55	Tue, 03 Feb 2015 16:29:32 UTC	148	12	0
step14.10213.automatix_stream.recon.ESD.816	Tue, 03 Feb 2015 01:40:38 UTC	149	11	0
step14.1024.automatix_stream.recon.ESD.618	Tue, 03 Feb 2015 04:56:29 UTC	146	14	0
step14.1029.automatix_stream.recon.ESD.51	Wed, 04 Feb 2015 08:07:50 UTC	81	79	0
step14.10395.automatix_stream.recon.ESD.703	Wed, 04 Feb 2015 03:26:46 UTC	135	25	0
Name	Creation Date	Locks OK	Locks Replicating	Locks Stuck

Showing 1 to 10 of 1,969 entries   Previous 1 2 3 4 5 … 197 Next

© 2012-2015 European Organisation for Nuclear Research (CERN)   Wed Feb 04 2015 13:10:05 GMT+0100 (CET)

Search: dataset

	Parent
ouglas:user.bdouglas.0226201326.408875.lib._002231	
ouglas:user.bdouglas.0401160328.393826.lib._002233	
ouglas:user.bdouglas.1224161342.665999.lib._002200	
ouglas:user.bdouglas.GRIF-IRFU_PHYS-TOP-datasets-to-possibly-4OCT11-0836	
ouglas:user.bdouglas.mc14_8TeV.117050.PowhegPythia_P2011C_ttbar.recon.727_s1933_s1911_r5591_tid01494881_00_der1412624703	
ouglas:user.bdouglas.NIKHEF-ELPROD_PHYS-atasets_to_move_15Oct12_1212	
ouglas:user.bdouglas.physics_Egamma.SMWZd3pdExample.NTUP_SMWZ.f41_p716	
ouglas:user.bdouglas.physic_p716	
ouglas:user.bdouglas.rucio_c	

DATASET	user.bdouglas:user.bdouglas.testds_
Type	DID

Showing 1 to 10 of 31 entries (filtered from 21,975

© 2012-2015 European Organisation for Nuclear Research (CE

You are here: Account Usage   Rucio Version: 0.2.11

Account Usage Overview (in TB)

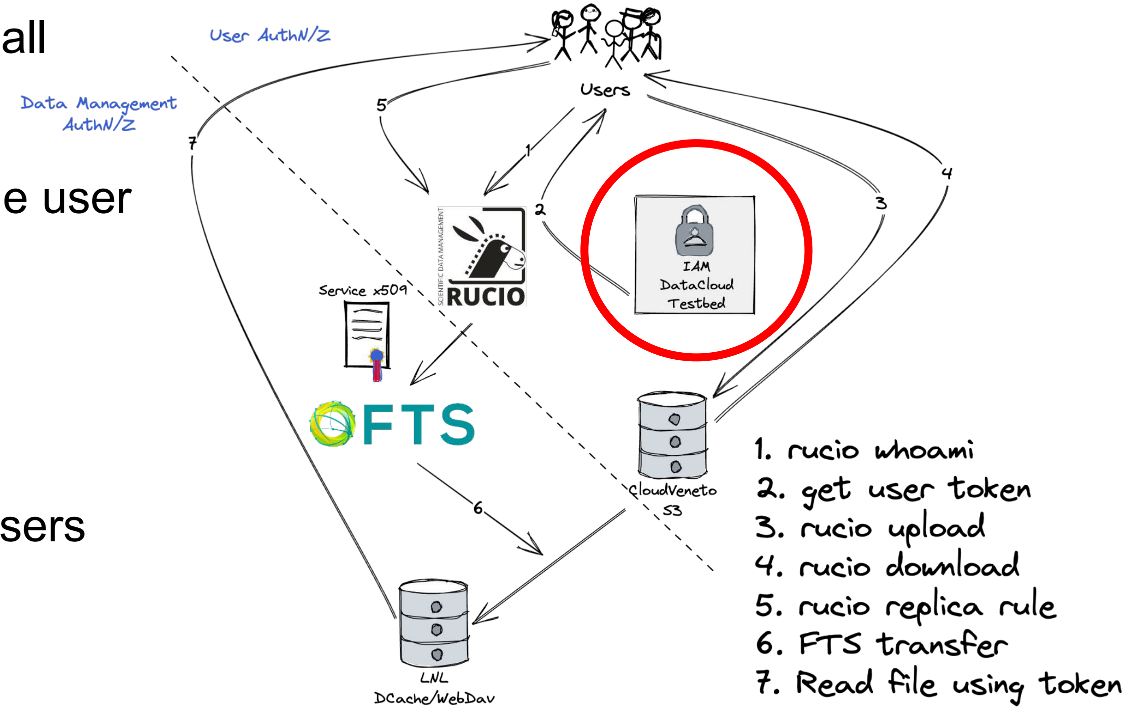
- **RucioUI**
  - Provides several views for different types of users
  - Normal users: Data discovery and details, transfer requests and monitoring
  - Site admins: Quota management and transfer approvals
  - Central administration: Account / Identity / Site management
- **Monitoring**
  - Internal system health monitoring with Graphite / Grafana
  - Transfer / Deletion / … monitoring built on HDFS, ElasticSearch, and Spark
- **Analytics and accounting**
  - E..g, Show which the data is used, where and how space is used
  - Data reports for long-term views
  - Built on Hadoop and Spark

Finanziato dall'Unione europea NextGenerationEU | Ministero dell'Università e della Ricerca | Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA | ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

The testbed AuthN/Z model

- Authentication is split into two logical level
  - **User:** interacts with the RUCIO server via IAM Token for all operation
    - Automatically mapped to a RUCIO account
  - **Data management:** it is RUCIO that acts on behalf of the user with a "service" identity.
    - Delegated to FTS for transfers

- **Autorizzation “Strawman” :**
  - Reading allowed to all users belonging to the data-lake users group via IAM token
  - Writing allowed to:
    - Service x509 proxy (admin only)
    - IAM token issued by RUCIO client
      - Necessary for rucio upload

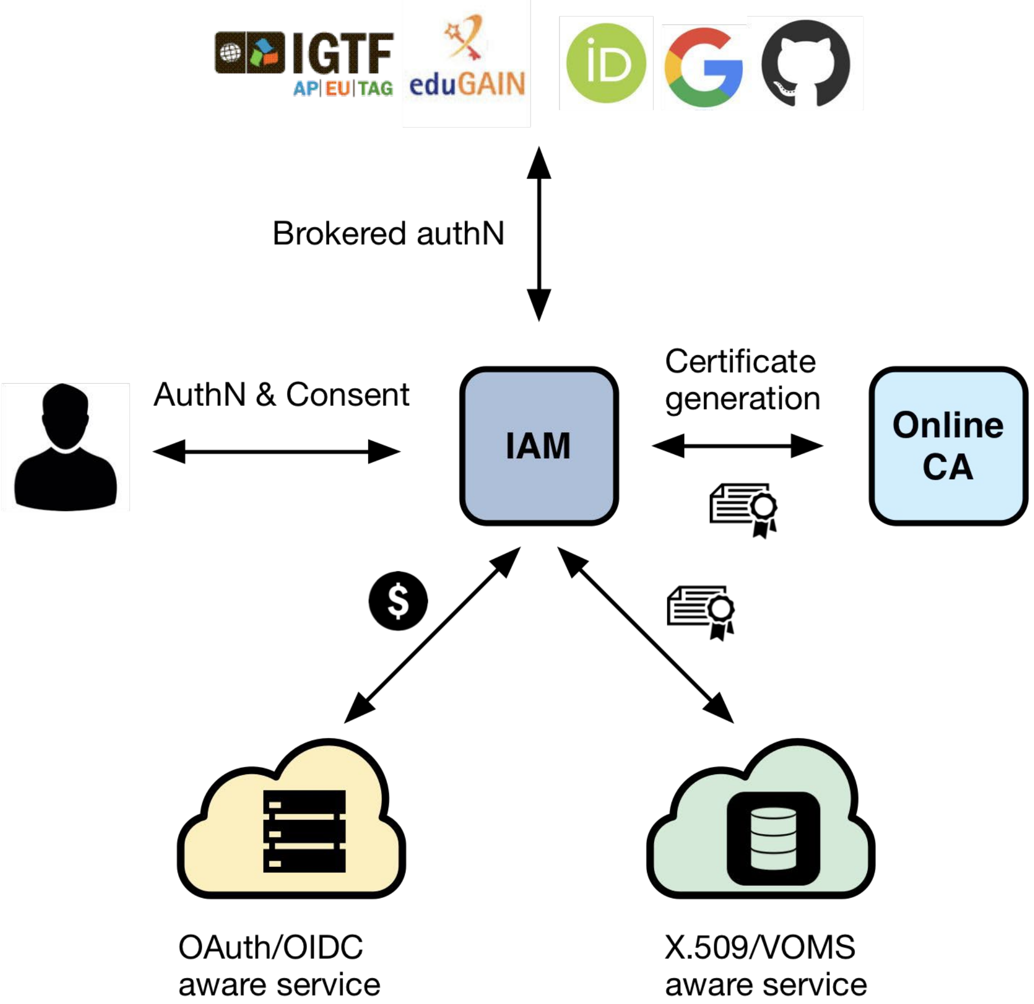


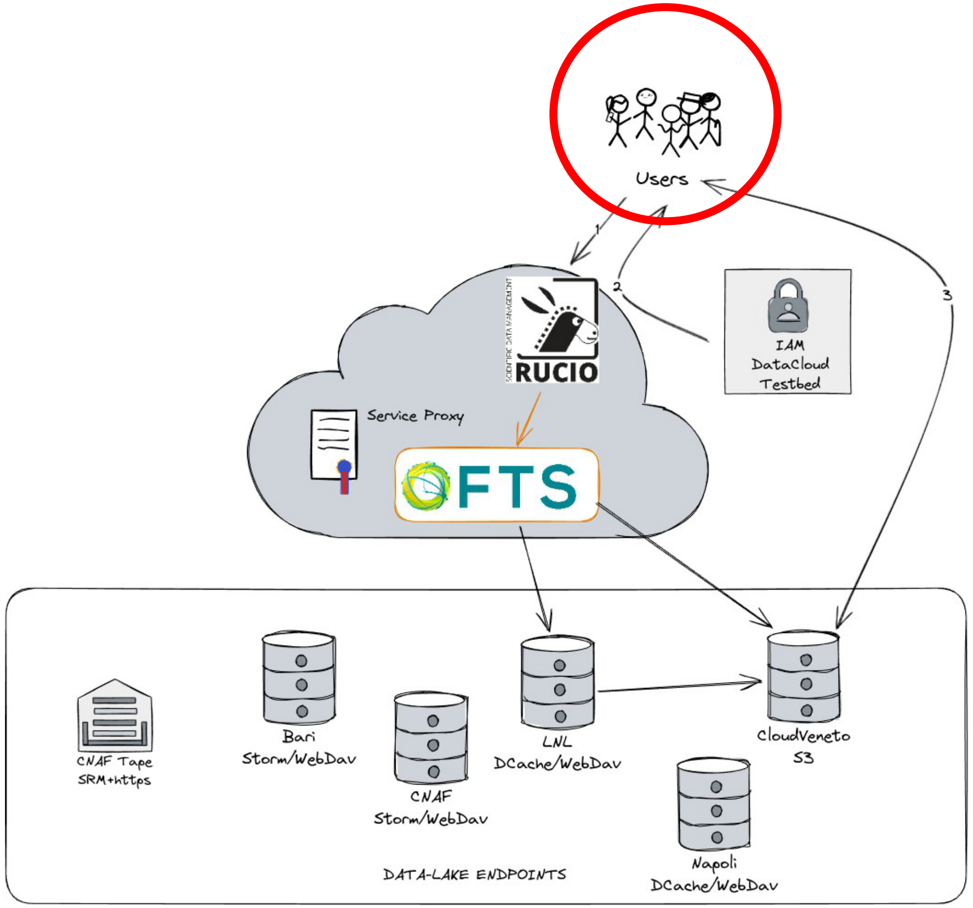
1. rucio whoami
2. get user token
3. rucio upload
4. rucio download
5. rucio replica rule
6. FTS transfer
7. Read file using token

IAM -Identity and Access Management

First developed in the context of the **H2020 INDIGO DataCloud** project (1st release v0.3.0 (2016-07-12))
An authentication and authorization service that:

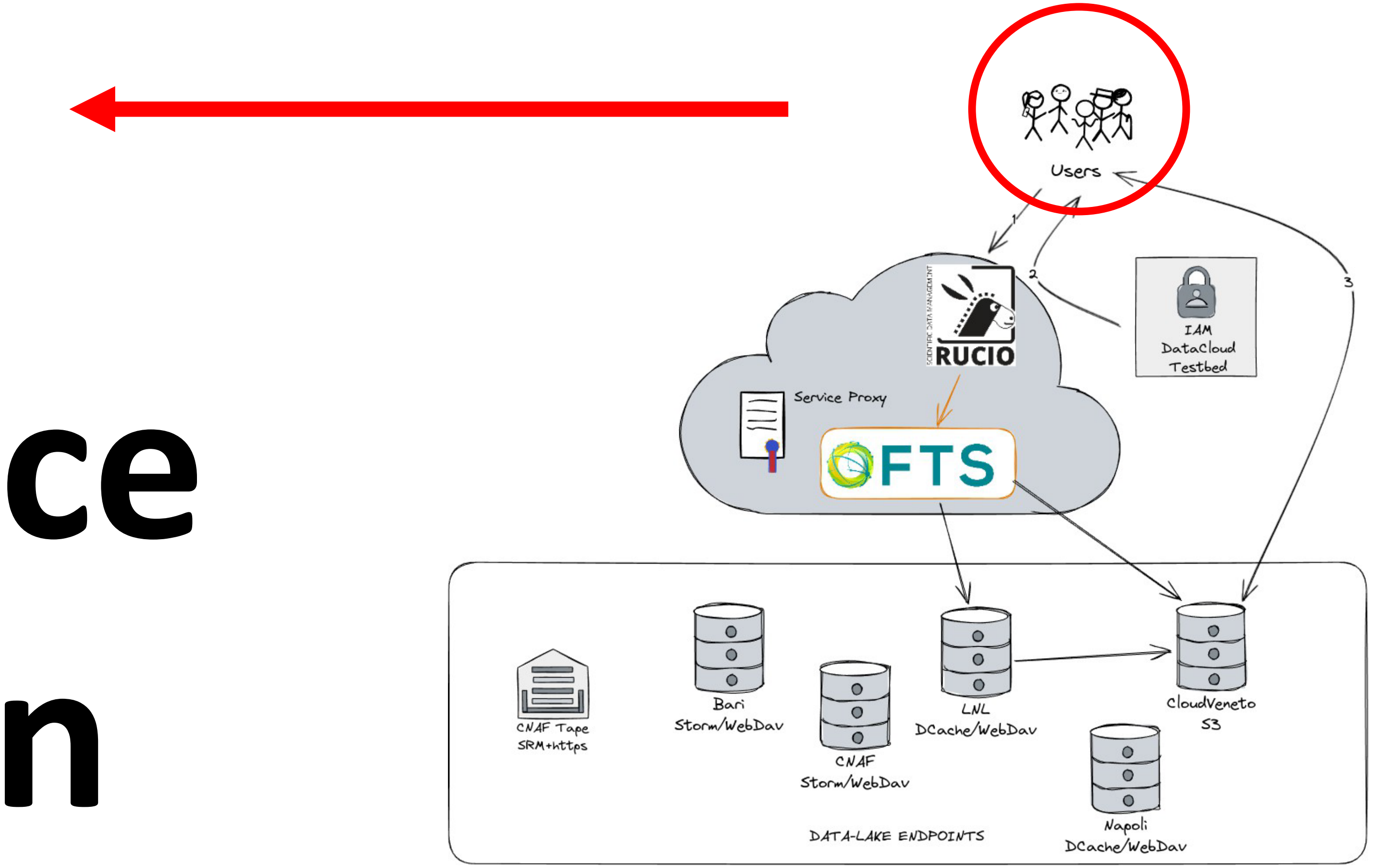
- supports **multiple authentication mechanisms**
- provides users with a **persistent, organization scoped** identifier
- exposes **identity information**, **attributes** and **capabilities** to services via **JWT** tokens and standard **OAuth & OpenID Connect** protocols
- can integrate existing **VOMS**-aware services
- supports **Web** and **non-Web access**, **delegation** and **token renewal**




Users
1
2
3
RUCIO
SCIENTIFIC DATA MANAGEMENT
IAM DataCloud Testbed
Service Proxy
FTS
CNAF Tape SRM+https
Bari Storm/WebDav
CNAF Storm/WebDav
LNL DCache/WebDav
CloudVeneto S3
Napoli DCache/WebDav
DATA-LAKE ENDPOINTS

YOU

Have a nice hands-on


Users
1
2
3
RUCIO
IAM DataCloud Testbed
Service Proxy
FTS
CNAF Tape SRM+https
Bari Storm/WebDav
CNAF Storm/WebDav
LNL DCache/WebDav
CloudVeneto S3
Napoli DCache/WebDav
DATA-LAKE ENDPOINTS