



ESCAPE

European Science Cluster of Astronomy &
Particle physics ESFRI research Infrastructures

Data Lake as a Service for Open Science

Riccardo Di Maria

CERN

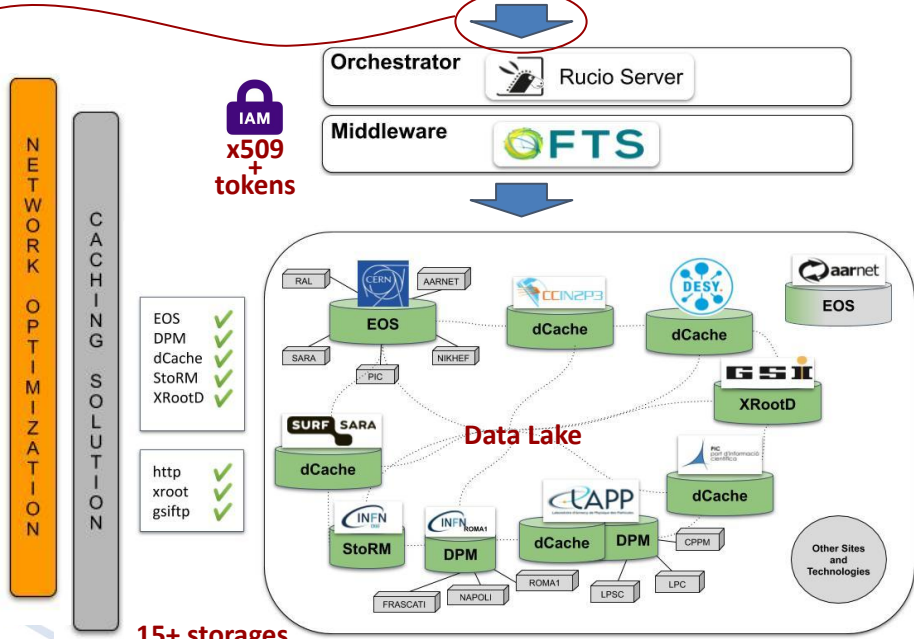
March 23rd, 2022 - 3rd ESCAPE DIOS Workshop



DLaaS for Open Science

- **Goal:** make **end-user comfortable** in embarking on a Data Lake experience
 - abstract the complexities of the Data Lake from the scientists
→ focus on doing science instead of data procurement
- An ever-increasing number of experiments are looking at Rucio Data Management system
 - **DLaaS** potentially interesting for both **aficionados** and **newcomers**

sciences



Further info: [https://wiki.escape2020.de/index.php/WP2 - DIOS#Datalake Status](https://wiki.escape2020.de/index.php/WP2_-_DIOS#Datalake_Status)



As It All Started

- An idea presented at CS3 2020 by the Rucio team [1]
- Development of a “Rucio JupyterLab Extension” as part of GSoC 2020 [2,3]

- A long time has passed, many things have happened...
 - CERN Summer Student to concretise the effort in 2021
 - [deployment](#), [docker-images](#), [docs](#)
 - DataLake-as-a-Service (DLaaS) in production-like phase
 - extensively exploited in DAC21 → SKA, MAGIC, CTA, ATLAS, KM3NET, LOFAR, FAIR
 - EOSC-Future (VRE), CS3Mesh4EOSC/ScienceMesh, EGI



DLaaS Use Cases

- Data discovery and access
- Submitting jobs to external services (remote computing)
 - conveniently browse data in Rucio through the extension
 - access file PFN directly from the Notebook
- Data preparation and processing
 - prepare/process data and upload back to the Data Lake
- Data preservation
 - produce data and upload to the Data Lake
- ++



Possible Future Developments

- Additional kernel compatibility
 - currently, only Python supported
- Token-support for direct download and upload
 - OIDC integration ongoing to all remaining ESCAPE RSEs
- Integration with content delivery and caching layer
 - XCache can be integrated to allow faster file download
 - completely transparent from the user PoV
 - successfully tested at small scale
- Multi-VO or off-site (CERN) deployment
- Deployment and distribution model for sciences → may spend countless hours on this...



Prospectives

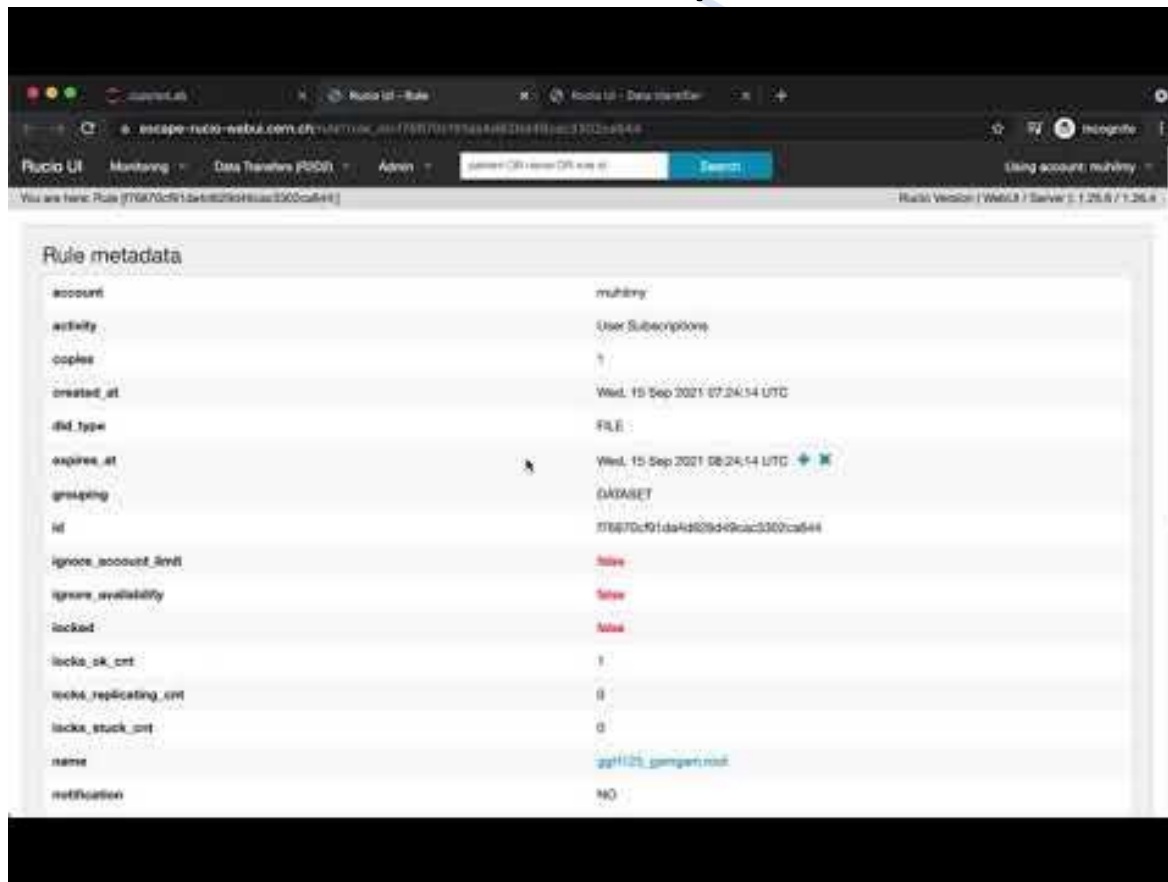
- ESCAPE end in 2022 → addressing long term sustainability
- **DLaaS** interesting for both **aficionados** and **newcomers** of **Rucio**
 - community-driven “development and operation”
 - needs and requirements of different experiments and sciences
- Proposal to establish a **Special Interest Group**
 - keep project alive, along with knowledge and know-how



Backup Slides



DataLake-as-a-Service for Open Science



The screenshot shows the Rucio UI interface. At the top, there's a navigation bar with 'Rucio UI', 'Monitoring', 'Data Transfers (PDS)', and 'Admin'. A search bar contains 'dataset:cms' and a 'Search' button. Below the navigation, the page title is 'Rule metadata' and the URL is 'rucio.webui.com/dataset/76670c91da4d829d49ca2300ca64'. The main content area displays a table of rule metadata:

account	ruhby
activity	User Subscriptions
copies	1
created_at	Wed, 15 Sep 2021 07:24:14 UTC
file_type	FILE
expires_at	Wed, 15 Sep 2021 08:24:14 UTC
grouping	DATASET
id	76670c91da4d829d49ca2300ca64
ignore_account_limit	false
ignore_availability	false
locked	True
locks_ok_cnt	1
locks_replicating_cnt	0
locks_stuck_cnt	0
name	ggH125_gamgarn.root
notification	NO



DLaaS Implementation

- Deployed in Kubernetes, using Zero-to-JupyterHub Helm chart → <https://escape-notebook.cern.ch>
- OAuth authentication using ESCAPE IAM (X509 still supported)
- [Rucio JupyterLab Extension](#) in Replica mode (i.e. TPC to local storage) used
 - connected to ESCAPE Data Lake (escape-rucio.cern.ch)
 - automatically pre-configured to use OIDC Auth
 - FUSE mount to EULAKE-1 RSE (EOS) *aka* creating a replication rule to move files to EULAKE-1
 - download mode still possible (if configured)

