ESCAPE DIOS
Developments & Achievements

Rizart Dona and Riccardo Di Maria on behalf of the ESCAPE project

CERN

March 31st, 2022 - Webinar: ESCAPE DIOS | A federated data-lake to connect large data volumes from different communities - benefits & use cases

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

ESCAPE
European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructures

Science Projects

HiLumi HL-LHC PROJECT
SKA SQUARE KILOMETRE ARRAY
VIRGO
FAIR
KM3NeT
JIVE Joint Institute for VLBI ERIC
cta cherenkov telescope array
LSST
EST european solar telescope
ESO
CERN

ESCAPE

European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructures

EUROPEAN OPEN SCIENCE CLOUD

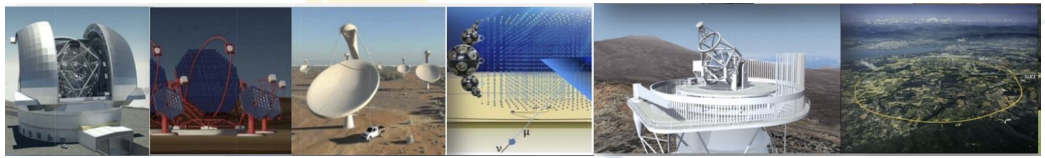
Horizon2020
European Union Funding
for Research & Innovation

Partners

CERN
INFN
GSI
DESY
Nikhef
SURF SARA
rijksuniversiteit groningen
CCIN2P3
cnrs
PIC port d'informació científica
LAPP Laboratoire d'Annecy de Physique des Particules
INAF ISTITUTO NAZIONALE DI ASTROFISICA NATIONAL INSTITUTE FOR ASTROPHYSICS

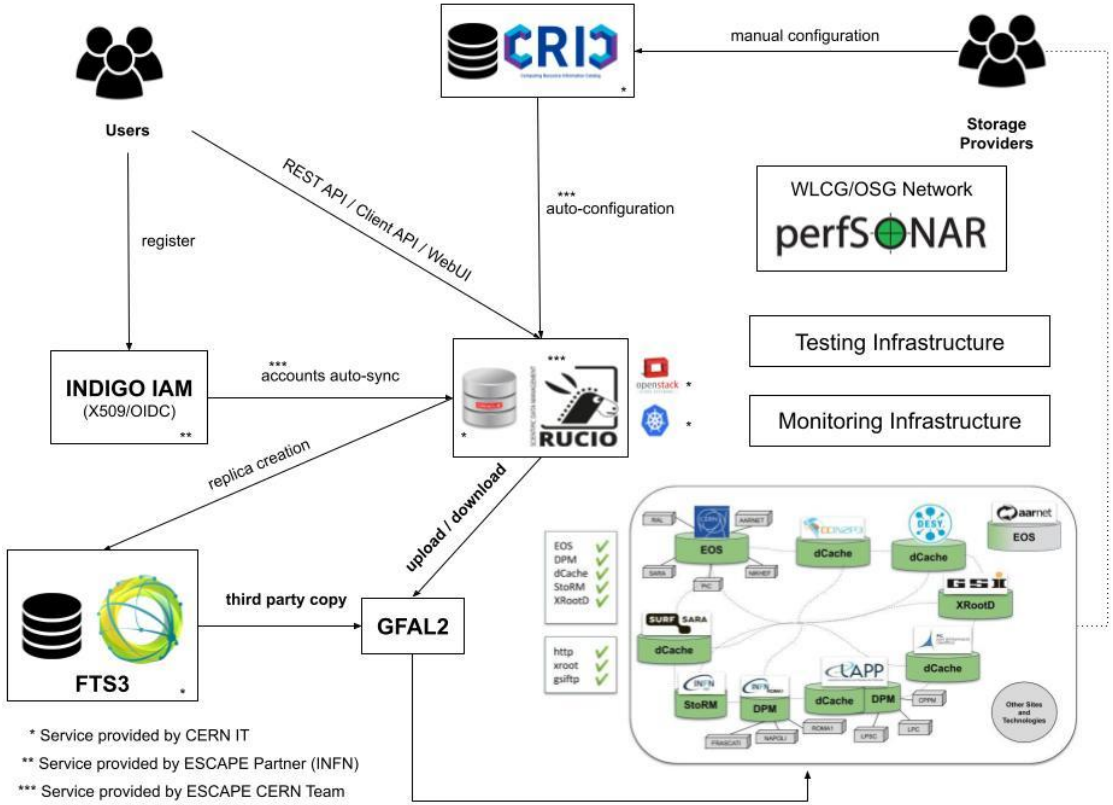
Project Goals

- Prototype an infrastructure adapted to exabyte-scale **future needs** of large science projects
- Ensure sciences drive the development of EOSC
- Address FAIR data management principles



rizart.dona@cern.ch
Funded by the European Union's Horizon 2020 - Grant N° 824064

The Data Lake

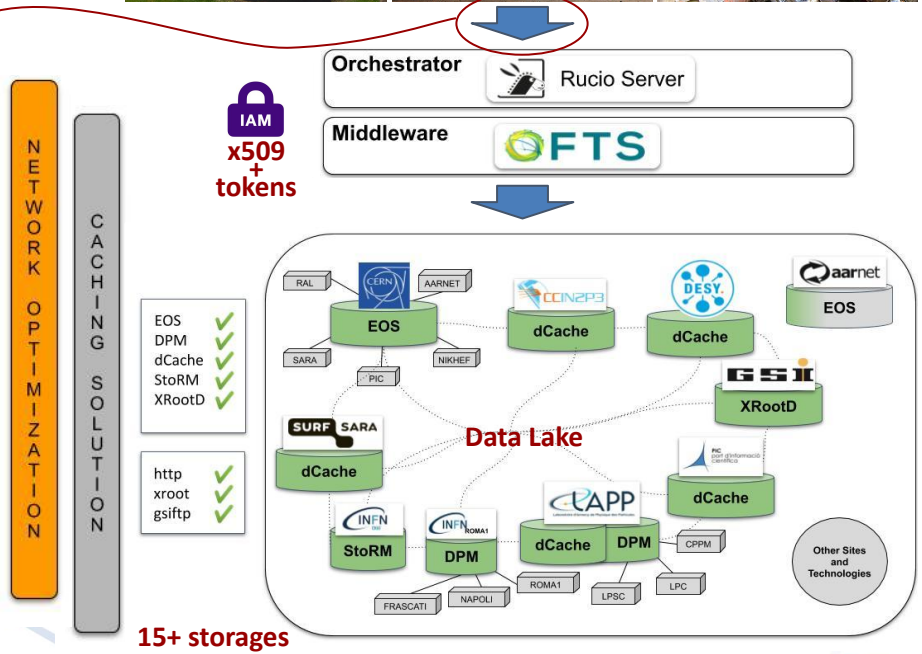


- Federated data infrastructure
  - Multiple **storage/protocol** technologies
  - Combination of HEP-specific services and industry standards
  - **QoS** and file transitions, **distributed redundancy** and **data policies**
- Data Transfer Stack
  - **Rucio** → Data orchestrator, policy-driven data management
  - **FTS3** → Middleware, reliable large scale file transfer service
  - **GFAL2** → Grid file access library, multi-protocol access
- Identity and Access Management
- CRIC Information Catalogue
- perfSONAR boxes deployed in the OSG/WLCG network
- **Testing** infrastructure
- **Monitoring** infrastructure

Funded by the European Union's Horizon 2020 - Grant N° 824064

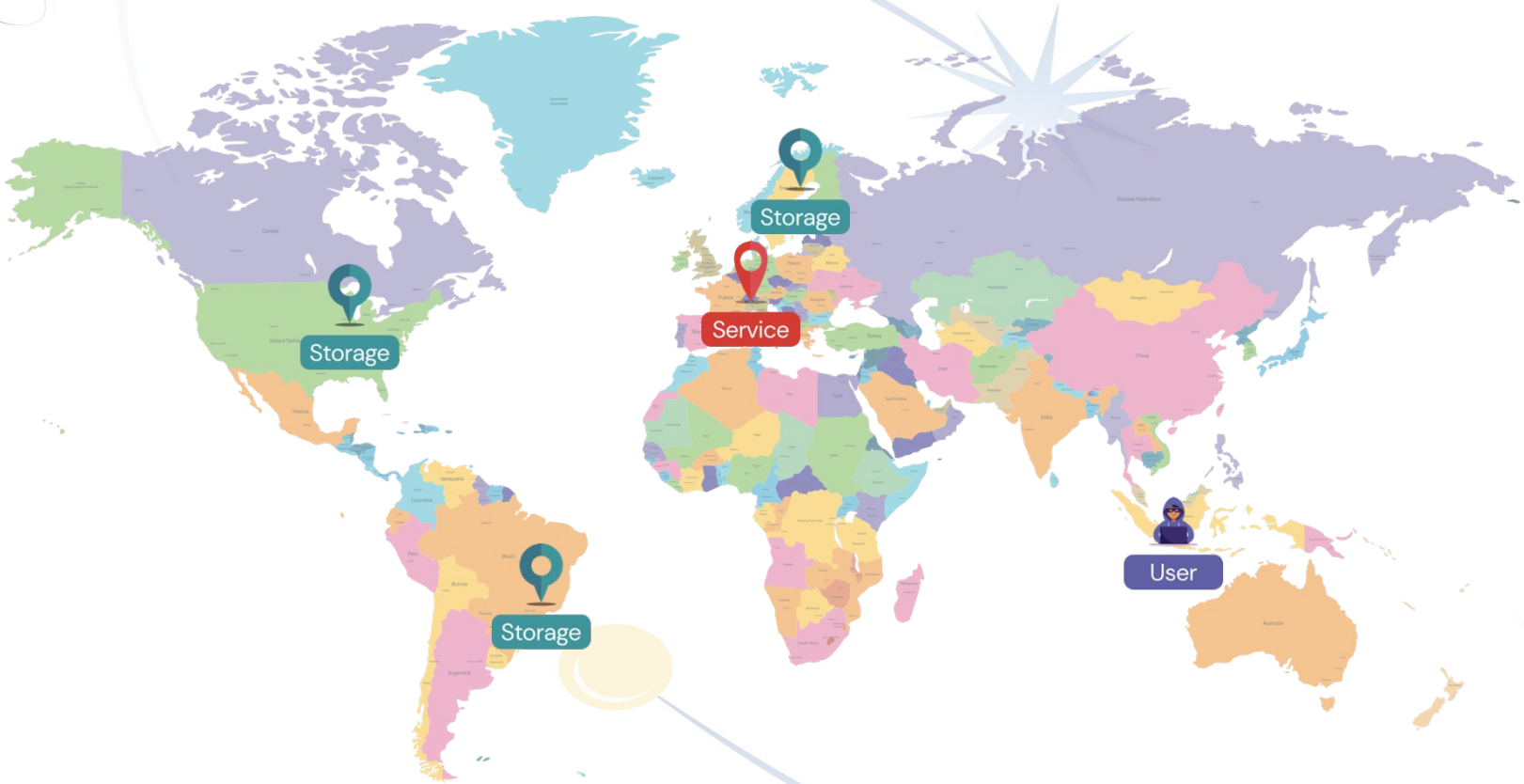
DLaaS for Open Science

- 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**



Riccardo.Di.Maria@cern.ch
Funded by the European Union's Horizon 2020 - Grant N° 824064

DLaaS - Overview



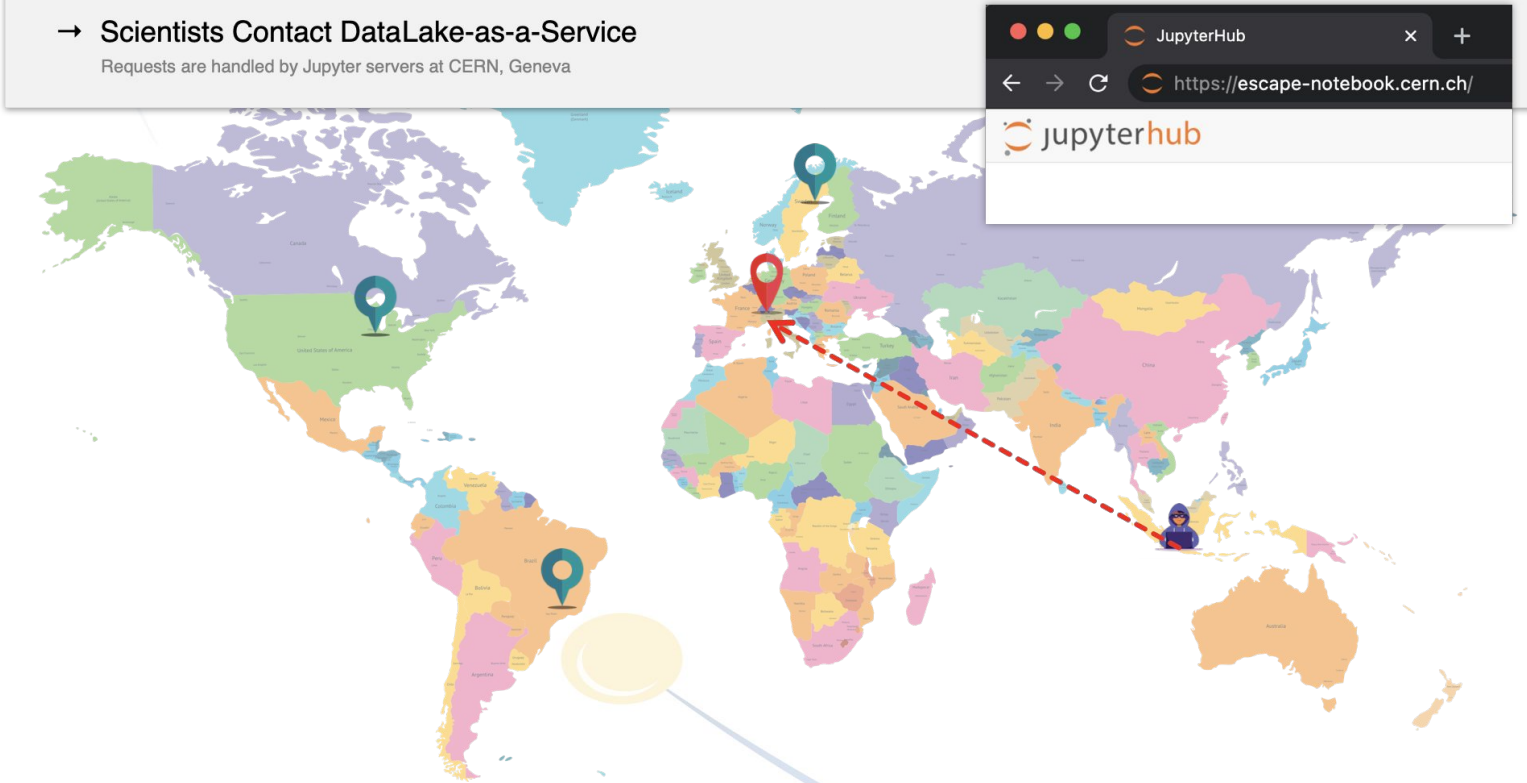
Riccardo.Di.Maria@cern.ch

Funded by the European Union's Horizon 2020 - Grant N° 824064

DLaaS - IAM (OpenID Connect) and Experiment Environments

→ Scientists Contact DataLake-as-a-Service

Requests are handled by Jupyter servers at CERN, Geneva

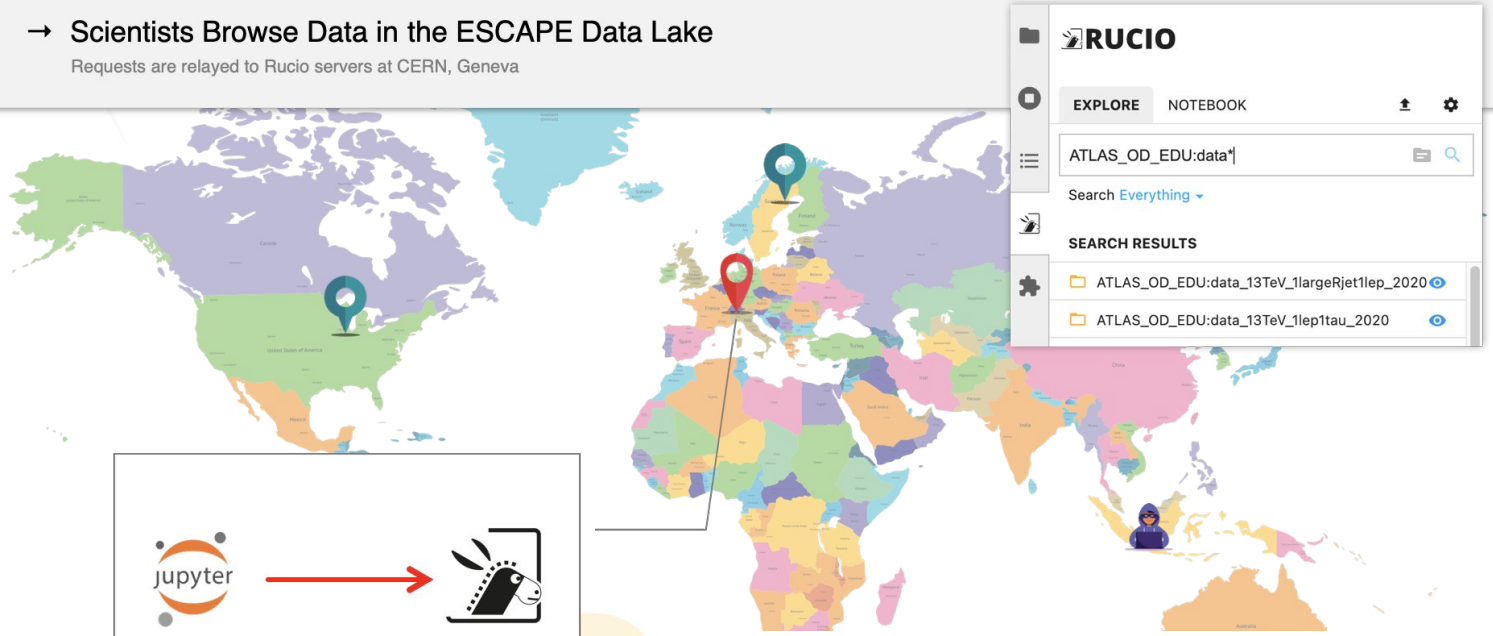


Funded by the European Union's Horizon 2020 - Grant N° 824064

DLaaS - Data Browsing/Discovery & Access

→ Scientists Browse Data in the ESCAPE Data Lake

Requests are relayed to Rucio servers at CERN, Geneva




```
[3]: for item in hyy_20:
         print(item.pfn)
```

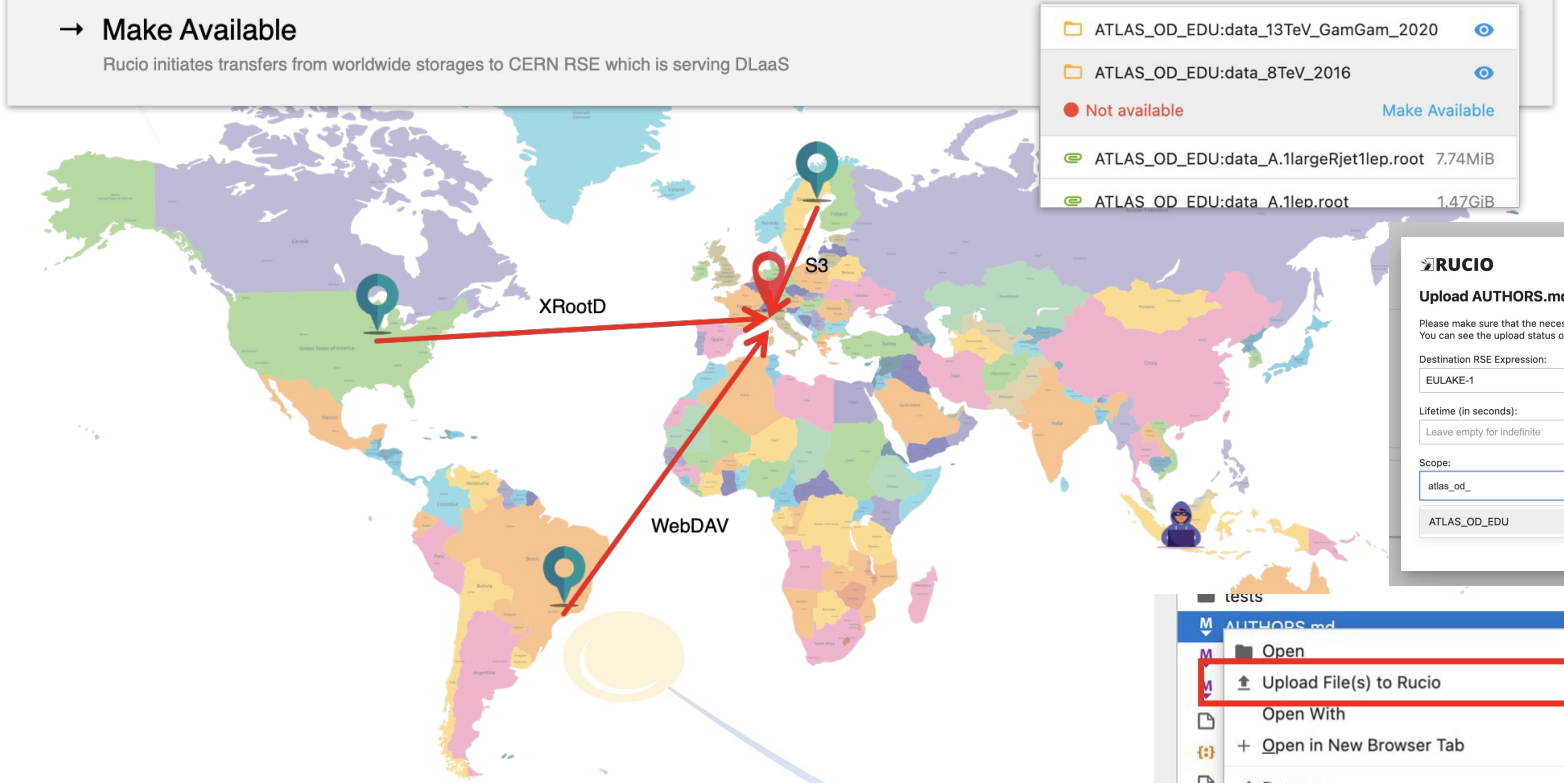
```
root://eoseulake.cern.ch:1094//eos/eulake/tests/rucio_test/eulake_1/ATLAS_LAPP_JEZEQUEL/6f/98/data_A.GamGam.root
root://eoseulake.cern.ch:1094//eos/eulake/tests/rucio_test/eulake_1/ATLAS_LAPP_JEZEQUEL/f1/3a/data_B.GamGam.root
root://eoseulake.cern.ch:1094//eos/eulake/tests/rucio_test/eulake_1/ATLAS_LAPP_JEZEQUEL/45/95/data_C.GamGam.root
root://eoseulake.cern.ch:1094//eos/eulake/tests/rucio_test/eulake_1/ATLAS_LAPP_JEZEQUEL/73/e3/data_D.GamGam.root
root://eoseulake.cern.ch:1094//eos/eulake/tests/rucio_test/eulake_1/ATLAS_LAPP_JEZEQUEL/6d/aa/mc_341081.ttH125_gamgam.GamGam.root.1
root://eoseulake.cern.ch:1094//eos/eulake/tests/rucio_test/eulake_1/ATLAS_LAPP_JEZEQUEL/1b/95/mc_343981.ggH125_gamgam.GamGam.root.1
root://eoseulake.cern.ch:1094//eos/eulake/tests/rucio_test/eulake_1/ATLAS_LAPP_JEZEQUEL/ff/c7/mc_345041.VBFH125_gamgam.GamGam.root.1
root://eoseulake.cern.ch:1094//eos/eulake/tests/rucio_test/eulake_1/ATLAS_LAPP_JEZEQUEL/13/b8/mc_345318.WpH125J_Wincl_gamgam.GamGam.root.1
root://eoseulake.cern.ch:1094//eos/eulake/tests/rucio_test/eulake_1/ATLAS_LAPP_JEZEQUEL/76/fd/mc_345319.ZH125J_Zincl_gamgam.GamGam.root
```


Funded by the European Union's Horizon 2020 - Grant N° 824064

DLaaS - Make Data Available in User Space, Data Preparation & Preservation

→ **Make Available**

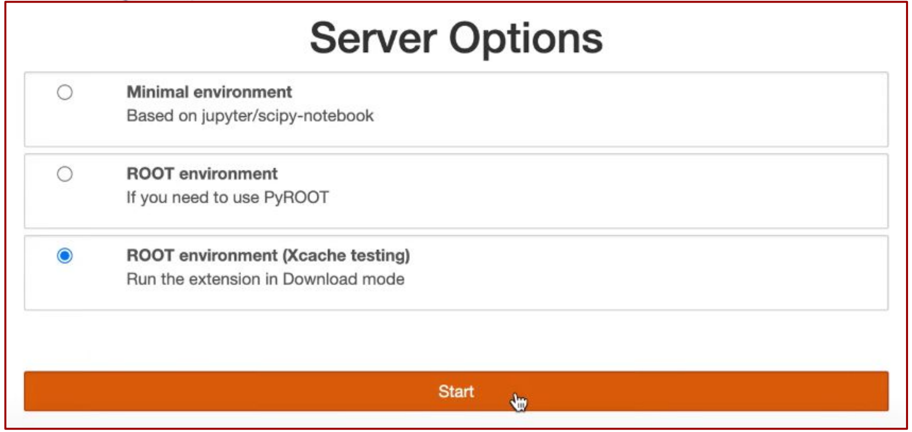
Rucio initiates transfers from worldwide storages to CERN RSE which is serving DLaaS

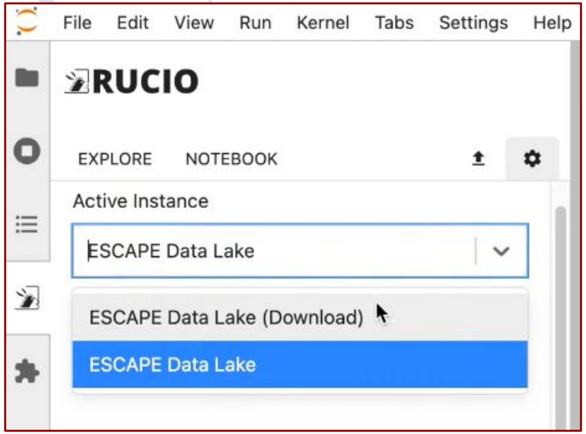


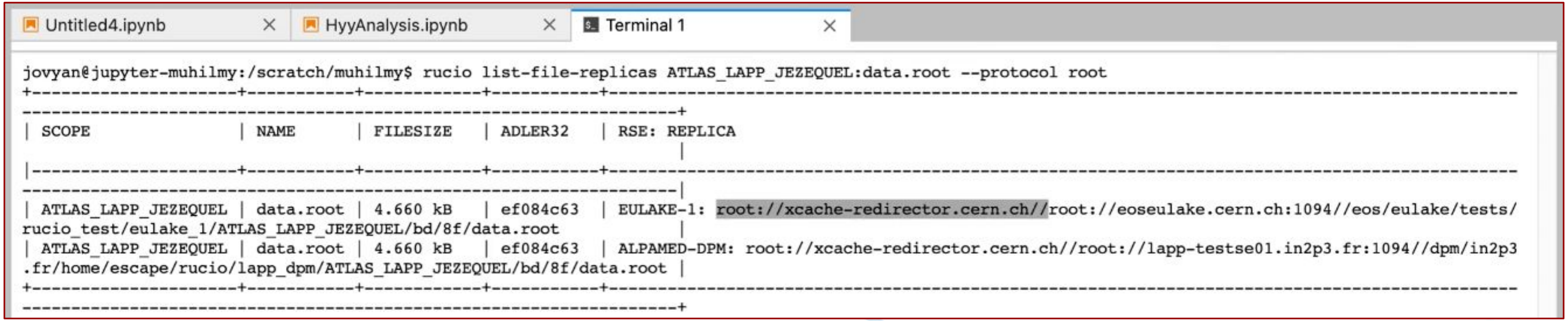
Riccardo.Di.Maria@cern.ch

Funded by the European Union's Horizon 2020 - Grant N° 824064

DLaaS - Content Delivery & Latency Hiding







Riccardo.Di.Maria@cern.ch
Funded by the European Union's
Horizon 2020 - Grant N° 824064

Achievements

- ESCAPE managed to **pilot** and **prototype** a **Data Lake** infrastructure fulfilling functional data management needs of flagship ESFRIs from several scientific disciplines
  - successful assessments of the Data Lake in 2020 and 2021
    - sensible technologies choice from WLCG environment and LHC experiments
  - community engagement & contribution
    - Astro-particle Physics, Electromagnetic and Gravitational-Wave Astronomy, Particle Physics, and Nuclear Physics pursuing together **FAIR** and **open-access** data principles
    - collaboration with other communities, e.g. PaNOSC, ExPaNDS, CS3MESH4EOSC, EOSC-Future
  - **DataLake-as-a-Service** hides the complexities of the Data Lake from the end users

- ESCAPE end in 2022 ⟶ addressing long term sustainability

Funded by the European Union's Horizon 2020 - Grant N° 824064

Thank you!

Questions/Comments?

Backup Slides →

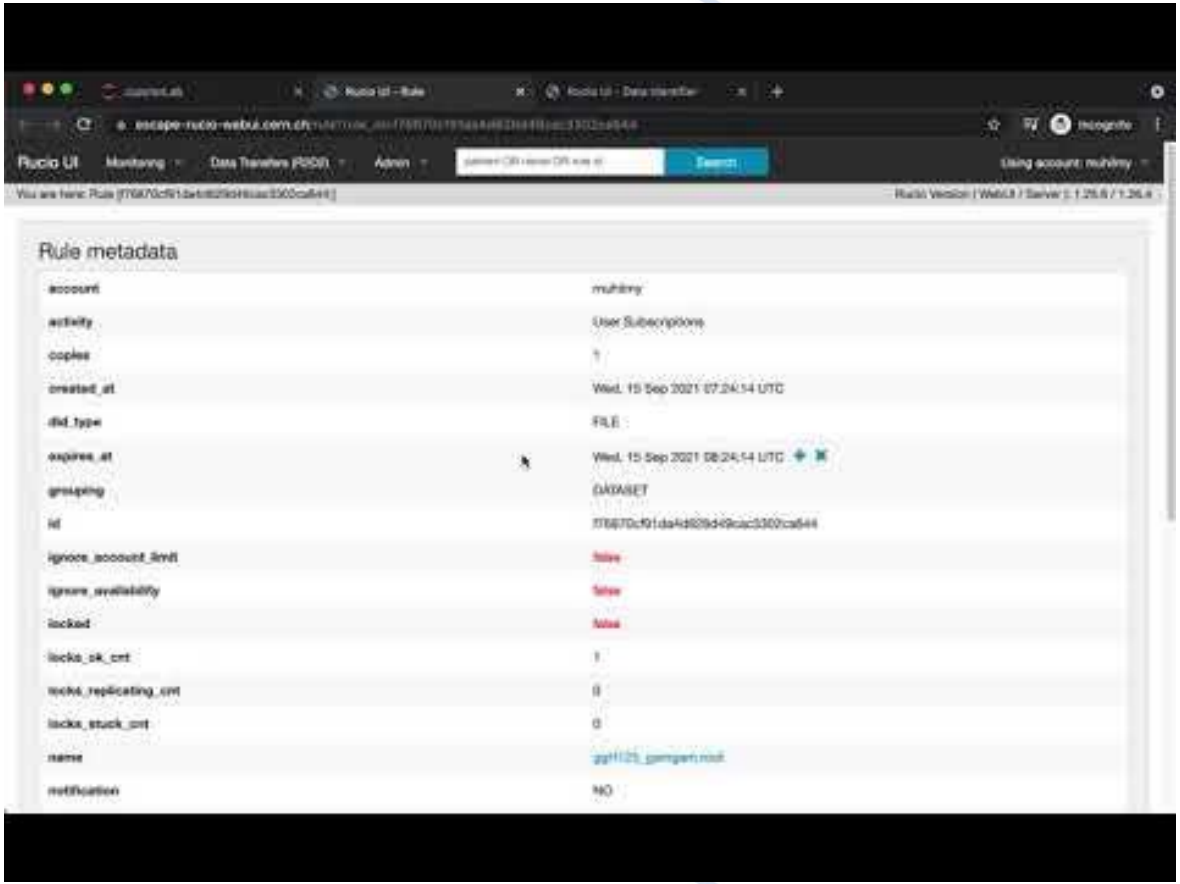
rizart.dona@cern.ch
Funded by the European Union's Horizon 2020 - Grant N° 824064

References

- ESCAPE Project, https://projectescape.eu/
- ESCAPE Data Lake Wiki, https://wiki.escape2020.de/index.php/WP2_-_DIOS
- ESCAPE Data Lake: Next-generation management of cross-discipline Exabyte-scale scientific data, CHEP 2021
- The ESCAPE Data Lake: The machinery behind testing, monitoring and supporting a unified federated storage infrastructure of the exabyte-scale, CHEP 2021
- Data Lake as a Service for Open Science, ISGC 2022

rizart.dona@cern.ch
Funded by the European Union's Horizon 2020 - Grant N° 824064

DataLake-as-a-Service for Open Science



Riccardo.Di.Maria@cern.ch

Funded by the European Union's Horizon 2020 - Grant N° 824064

DLaaS - As It All Started

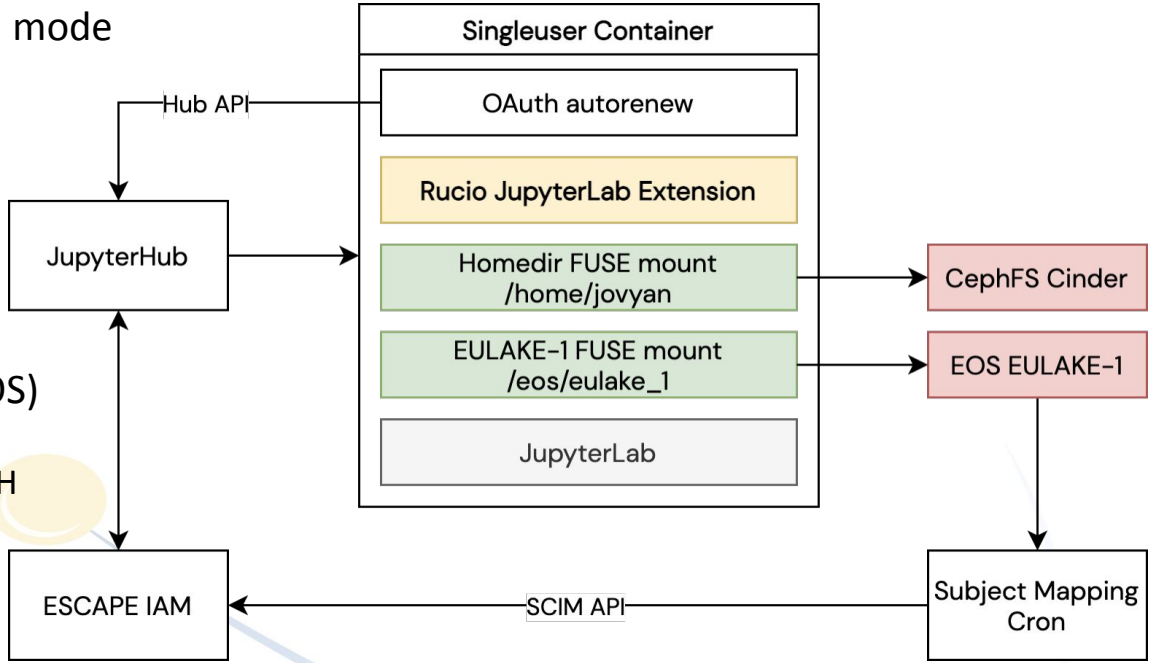
- An idea presented at CS3 2020 by the Rucio team [1]
- Development of a “Rucio JupyterLab Extension” as part of GSoC 2020 [2,3]
- A long time has passed, many things have happened…
  - CERN OpenLab Summer Student to concretise the effort in 2021
    - deployment, docker-images, documentation
  - DataLake-as-a-Service (DLaaS) in production-like phase
    - extensively exploited during ESCAPE “Data and Analysis Challenge” in November 2021 by SKA, MAGIC, CTA, ATLAS, KM3NET, LOFAR, FAIR
      → 3rd ESCAPE DIOS Workshop
    - EU projects e.g. EOSC-Future and CS3Mesh4EOSC/ScienceMesh
    - other communities e.g. EGI

Funded by the European Union’s Horizon 2020 - Grant N° 824064

DLaaS Implementation

RSE = Rucio Storage Element

- 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
  - download mode still possible (if configured)
  - connected to ESCAPE Data Lake
  - automatically pre-configured to use OpenID Connect
  - 2 FUSE mounts to EULAKE-1 (EOS)
    - ESCAPE RSE in r-only
    - additional RSE in r/w: SCRATCH
  - making files available *aka* creating a replication rule to move files to EULAKE-1



Riccardo.Di.Maria@cern.ch
Funded by the European Union's Horizon 2020 - Grant N° 824064