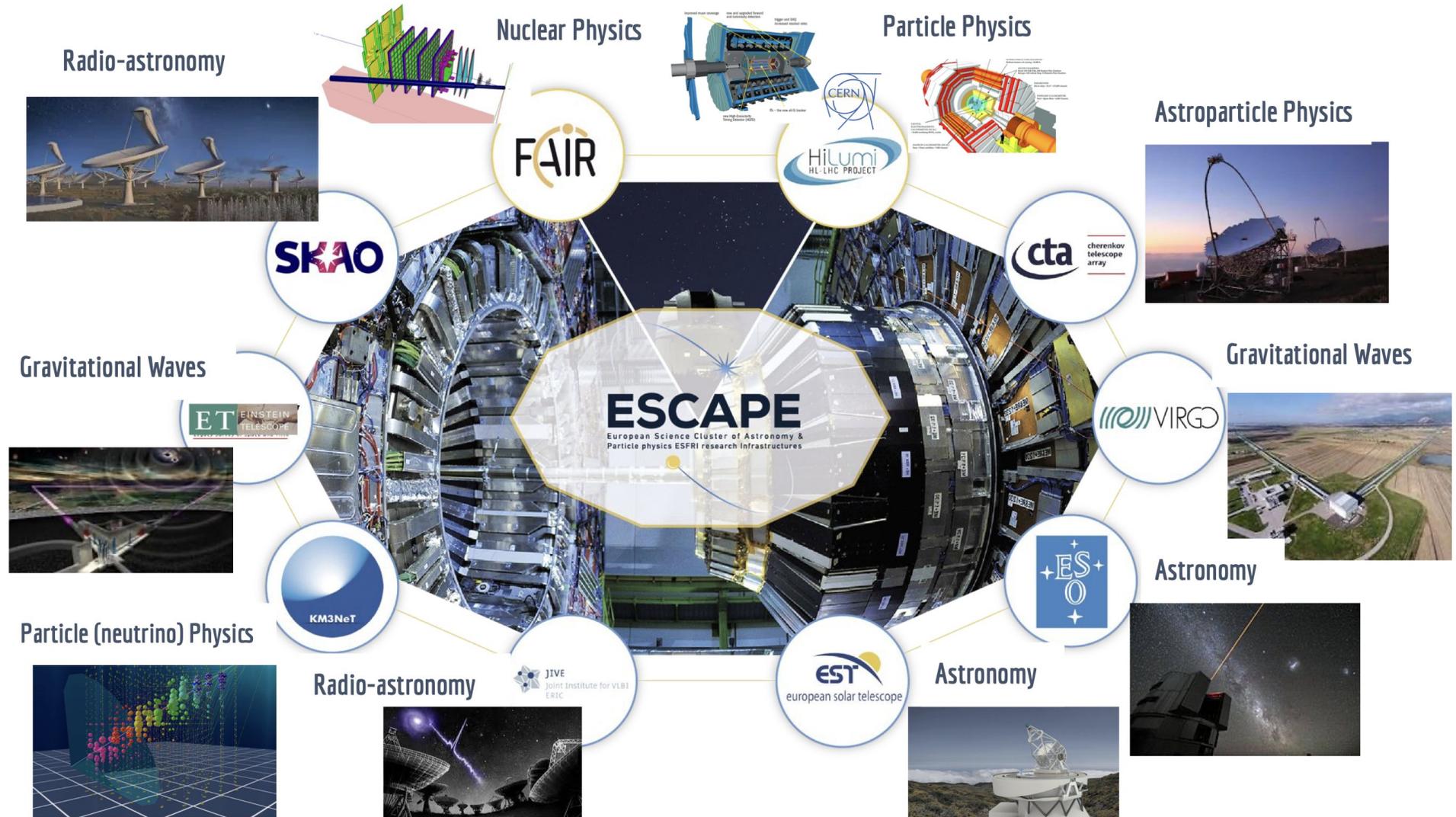# Federating Sciences: The ESCAPE Data Infrastructure for Open Science

G. Guerrieri, CERN

A. Iess, CNRS LAPP

The ESCAPE Open Collaboration

The ESCAPE Open Collaboration

- ESCAPE Science Cluster for next ESFRIs astronomy, astroparticle and particle physics established in 2019 as result of H2020 call.

- Open Science and FAIR data and software practices, provide a link to EOSC (EOSC Future project, OSCARS)

- Collaboration Agreement signed by 10 RIs in 2023.

- Letter of Intent to align national initiatives with the ESCAPE work programme.
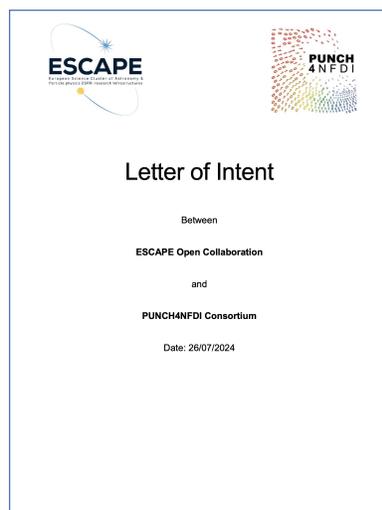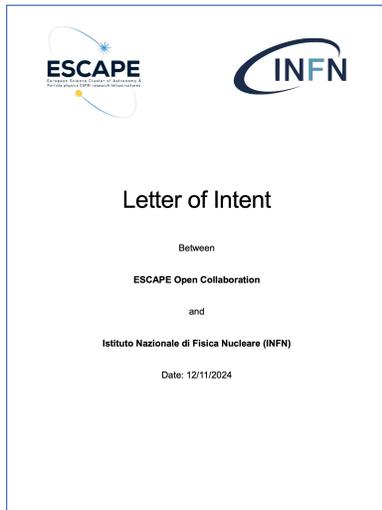
- ESCAPE Data Infrastructure for Open Science (DIOS) WG contributed to JENA Computing initiative White Paper on Federated Computing, outlining a strategic vision on federated data management, virtual research environments and FAIR/open data.

The ESCAPE Work Programme

ESCAPE is not participating as a Node in EOSC Federation build-up phase, but closely following developments.

🟡 Open Science Clusters' Action for Research & Society ([OSCARS](#))

🟡 European Virtual Institute for Research Software Excellence ([EVERSE](#))

ESCAPE Open Collaboration Work Programme

Entities

ESCAPE CC
Operating the community-based "Competence Center" for EOSC-alignment, train and support, extended outreach, financial model for services and networking with other SCL-CCs

ESCAPE VRE | Virtual Research Environment
A web-based highly-composable research-enabling interoperable data platform

ESCAPE EVSI
R&I for an "European Virtual Institute for Research Software" for advanced technologies

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

Working Groups

ESCAPE DIOS | Data Infrastructure for Open Science
Access physical & e-infrastructures
Processing & Analysis
Security & Operations

ESCAPE OSSR | Open-source Scientific Software and Service Repository
Aggregator & Integrators
Sharing and Discover
Training & Support

ESCAPE ESAP | ESFRI Science Analysis Platform
Processing & Analysis
Sharing and Discovery
Training & Support

ESCAPE CS | Citizen Science
Sharing and Discovery

ESCAPE VO | Virtual Observatory
Processing & Analysis
Sharing and Discovery
Training & Support

Programmes

ESCAPE COSO
Challenging "Open Science Objectives" by RI commitments in Open Science Projects (OSP) as well as Cross-Cluster Open Science Projects (COSP)

ESCAPE TECH
Bring the FAIRness within technology, R&D and innovation projects as well as explore new "close-to-sensors" low-latency open-data science

ESCAPE CARS
Career development and rewarding for researcher committing in Open Science. Planning, tracking, and assessing scientific knowledge production

ESCAPE SDSS
Building synergies on "Sector Data Spaces" for Society: Green deal, Health, Manufacturing, Education and Skills
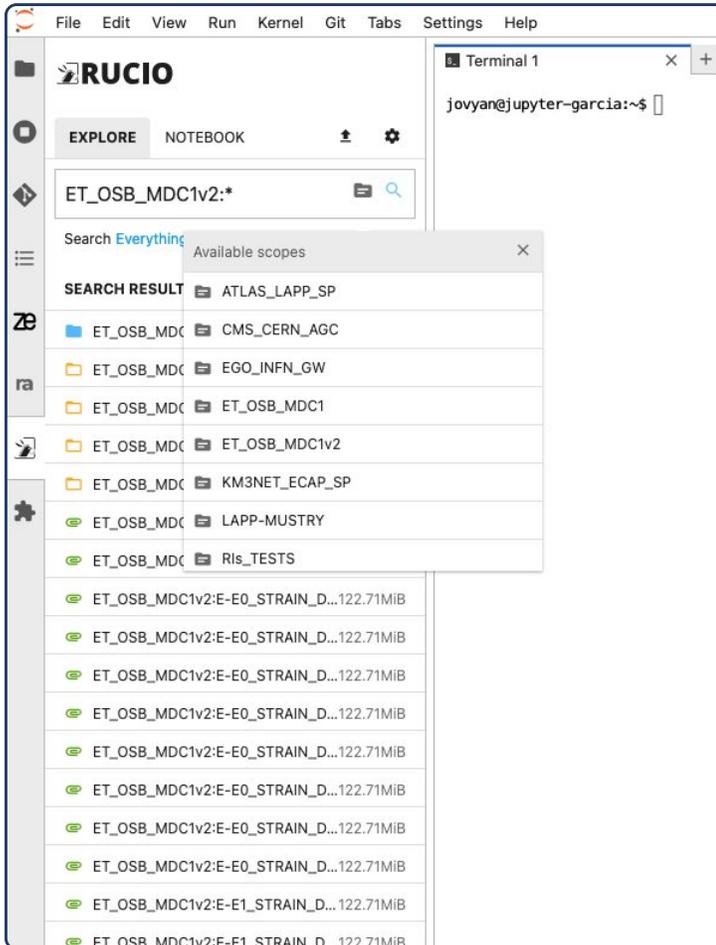
ESCAPE Virtual Research Environment

- **AAI**: A federated and reliable **Authentication and Authorization** layer

- **The Rucio Data Lake**: A **federated distributed storage** solution, providing functionalities for data injection and replication through a Data Management framework (Rucio)

- **REANA**: A **computing** cluster supplying the processing power to run full analyses with REANA, a re-analysis service

- **CVMFS**: A **read-only file system** designed to distribute software, and more.

- **JupyterHub**: A **notebook interface** with containerised environments to hide the infrastructure's complexity from the user.
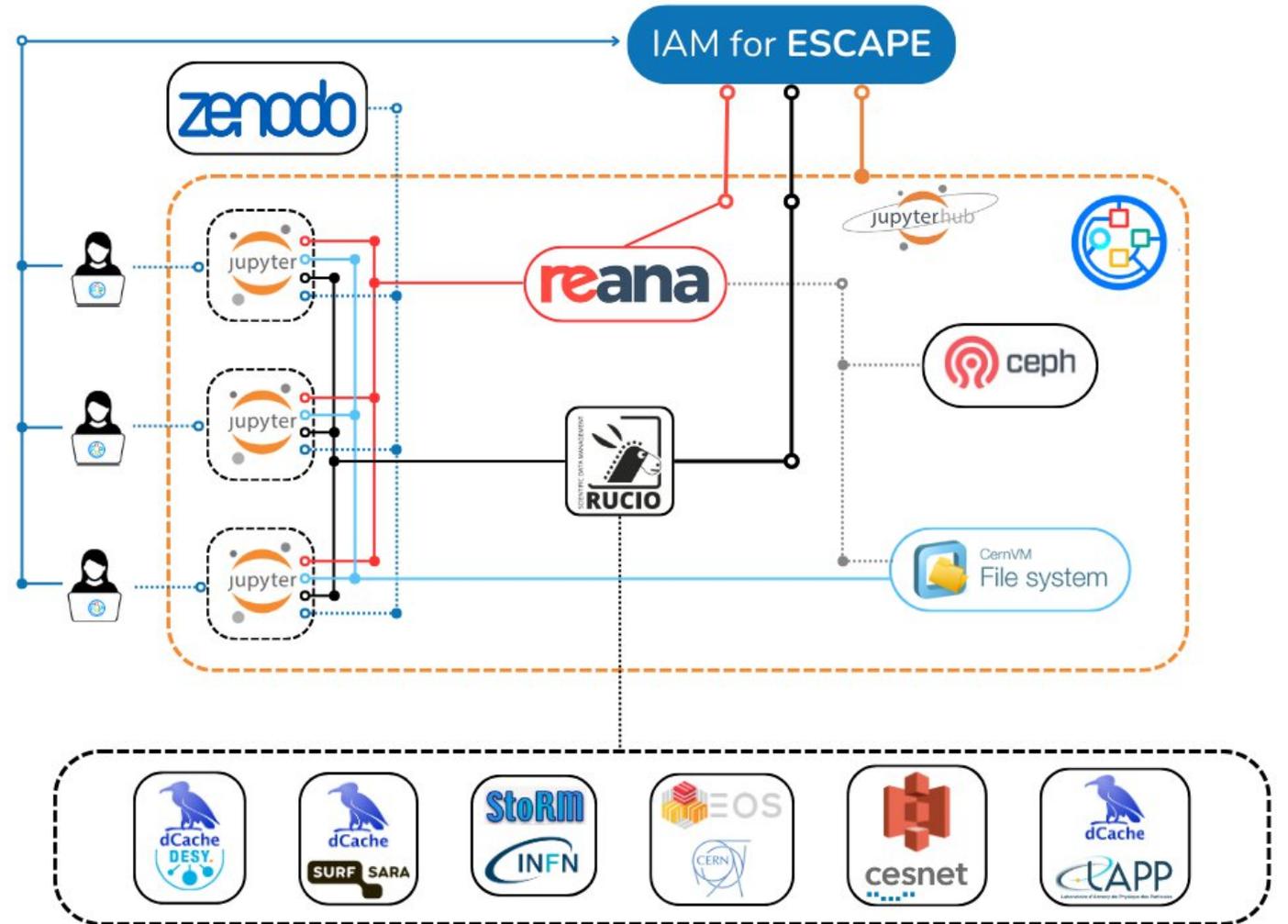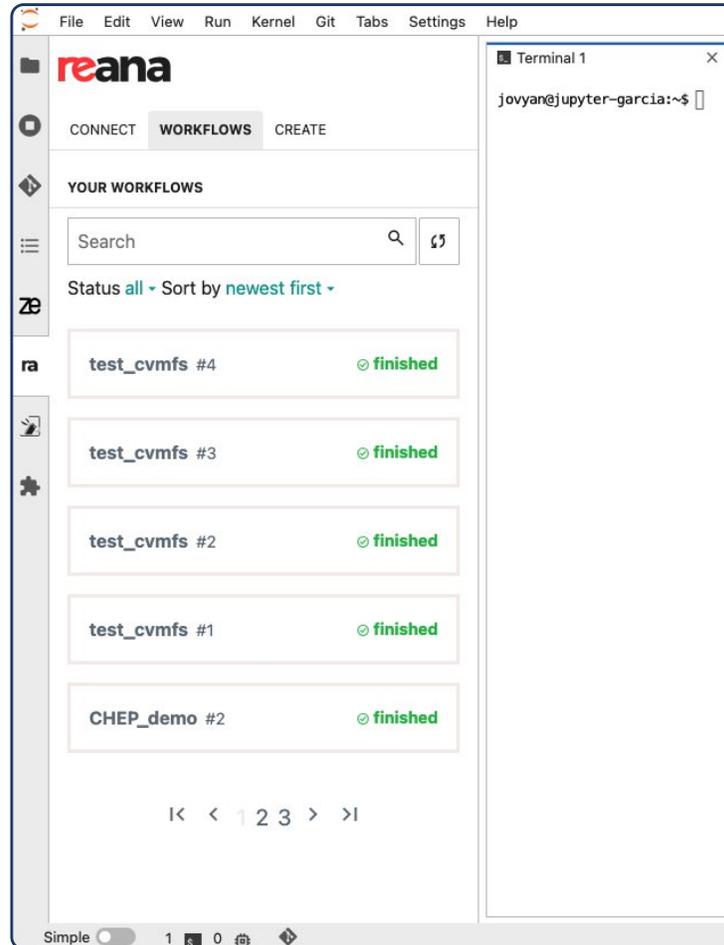
ESCAPE Virtual Research Environment
What makes the VRE, the VRE

ESCAPE Virtual Research Environment
A multi-wavelength analysis in the VRE (3C196 quasar)

1. Data **injected** to the Data Lake from different radio sources from **external** locations

2. Download of data, **process** and **storage** of results in the VRE

3. Combined **optical** data located via the VO plugin available in the **VRE**

4. Optical and **radio** data combined**.** Results uploaded back to the Data Lake.
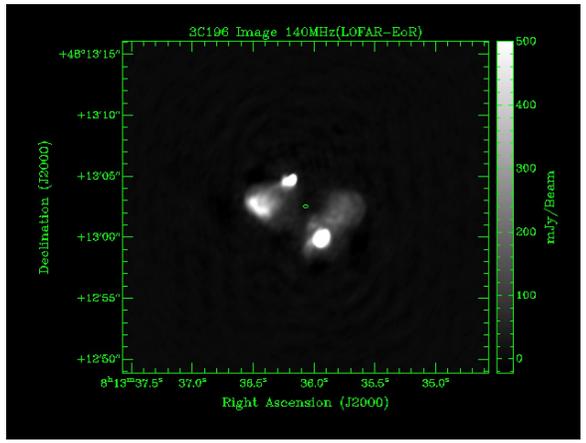
Radio image
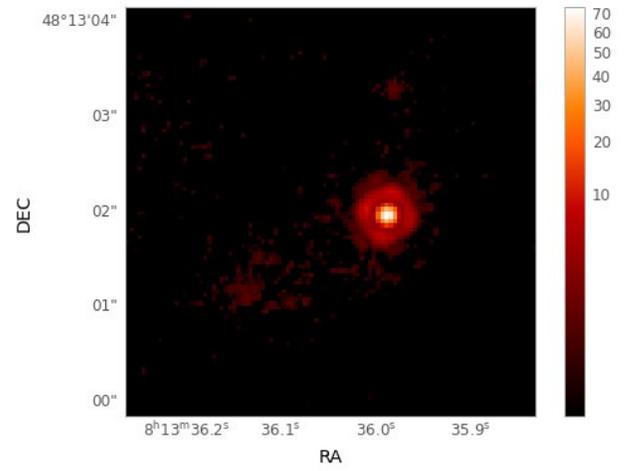
Optical image
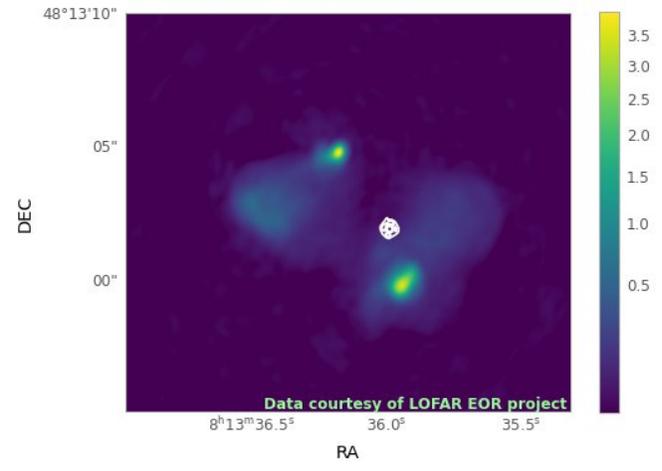
Combined image

ESCAPE Data Infrastructure for Open Science

🟡 How can we *support* the ESCAPE communities and possibly be of use to others as well?

1. Provide a clear and accessible way to **test and later deploy** distributed data management solutions

2. Offer a forum where communities can **exchange views** on shared technologies, both among themselves and with experts

3. Facilitate the **operation** of both the data and analysis infrastructure of individual RIs

4. Foster the co-design and co-development of technologies useful **across** RIs: economies of scale!

Example 1: building connections

ESCAPE Data Infrastructure for Open Science
The ESCAPE VRE as part of the EOSC Federation

- We're describing the result delivered by **CERN** as stakeholder of the ESCAPE Collaboration

  - This doesn't mean that ESCAPE is part of the Federation

- CERN onboarded (among other services) both the VRE and REANA into the EOSC EU Node

  - To see what this means, you can look at the [Federation Handbook](#), and [Jose's presentation](#)

- Currently working on building a **pilot Rucio Data Lake** to address data management

- To an extent this effort is decoupled w.r.t. [Giuseppe's presentation](#) about Sync&Share

  - We are not talking about users' data, but scientific data from RIs

  - There could (should) be an interaction point between Rucio and OCM (see [memo](#) about OCM)

eosc | Tripartite

Building the EOSC Federation

The vision for EOSC is to put in place a system in Europe to find and access data and services for research and innovation. This is to help researchers store, share, process, analyse and reuse FAIR research outputs within and across disciplines and borders.

The deployment of a network between data repositories and services will be instrumental for Open Science to progress in Europe. For this, the EOSC Federation of nodes is being created.

The ongoing build-up phase is the first phase of development of an operational EOSC Federation.

Organisations participating in the workshop for the kick-off of the build-up phase of the EOSC Federation
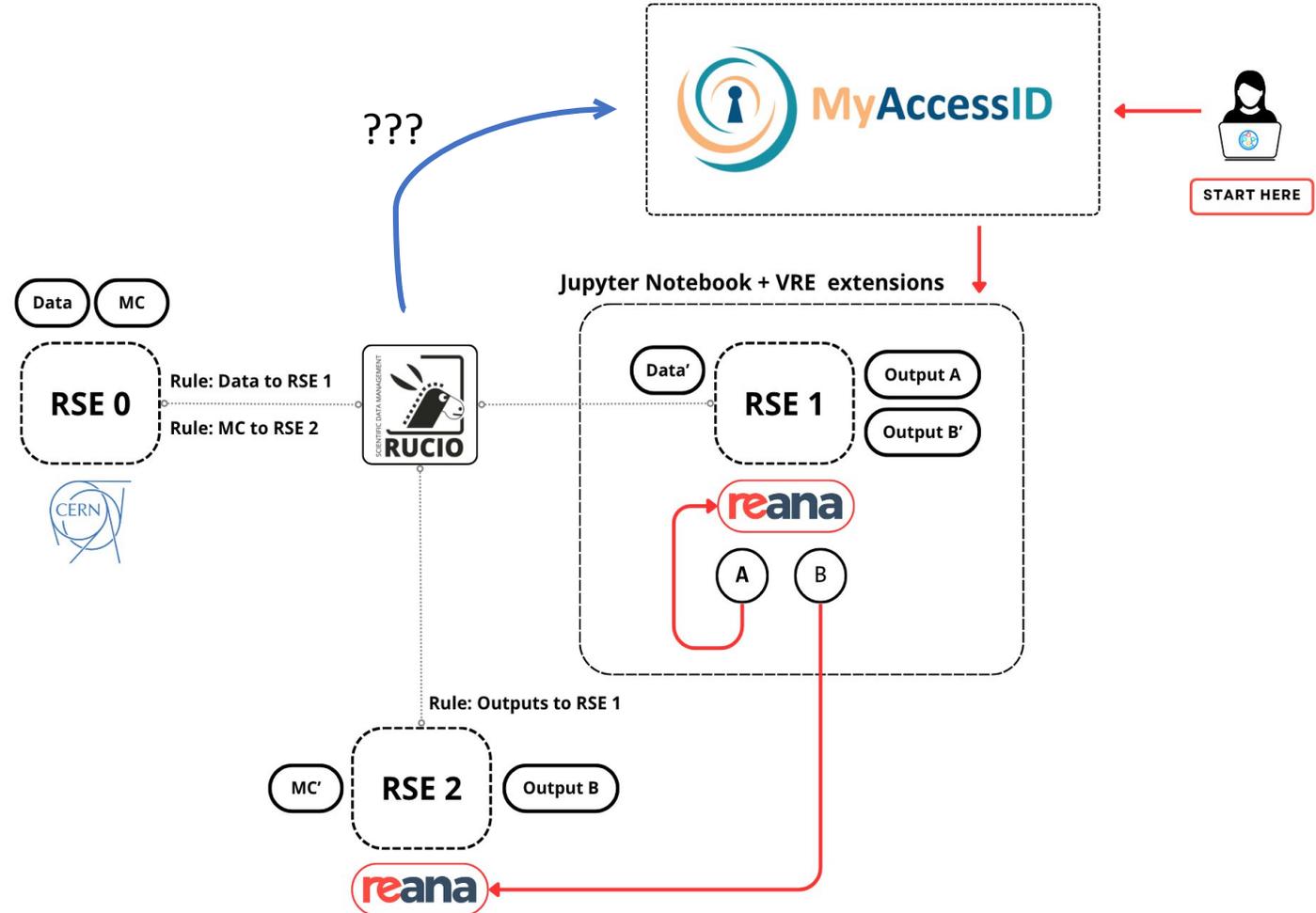
ESCAPE Data Infrastructure for Open Science
The Rucio data lake

- Currently working on building a **pilot Rucio data Lake** to address data management

  - Registered storage element in **BARI** (ICSC)

  - Looking forward to **CNAF** (ICSC)

  - Negotiation with:

    - (ongoing) **EOSC-PL**

    - (planned) **SURF**

    - (need to revamp) **PaNOSC** (CSC / DESY?)

    - (need to revamp) **EEN**

- How to set up a storage element in Rucio (node admin POV):

  - Grant r/w access to a service certificate

  - If there are Virtual Organizations involved, add necessary VO to the storage mapping

  - FYI: Rucio (or rather, FTS) supports XrootD servers, StoRM / WebDav endpoints, DCache, CEPH, S3, ...

Certificates are ok, but tokens are what EOSC uses.

Currently Rucio only accepts token from INDIGO IAM (with wlcg profile)

Will the proxied token introspection be enough, or should we think about native support for aarc payloads?

IAM wlcg token payload

```
{
  "wlcg.ver": "1.0",
  "sub": "09d42bee-c7c9-44a1-a5f1-7a24e326708e",
  "aud": "rucio",
  "act": {
    "sub": "75fee7e9-a022-425e-9540-7bb97e1fe618"
  },
  "nbf": 1761908262,
  "scope": "openid profile",
  "iss": "https://iam-escape.cloud.cnaf.infn.it/",
  "exp": 1761915462,
  "iat": 1761908262,
  "jti": "b0e06f3d-2f2d-47d1-aae8-b59c7bef1b53",
  "client_id": "75fee7e9-a022-425e-9540-7bb97e1fe618"
}
```

vs.

EOSC aarc token payload

```
{
  "aud": [
    "APP-xx-5571-4FD1-8A4C-E9CE97095B29"
  ],
  "auth_time": 1761907829,
  "client_id": "APP-xxx-5571-4FD1-8A4C-E9CE97095B29",
  "exp": 1761908131,
  "iat": 1761907831,
  "iss": "https://proxy.testing.eosc-federation.eu",
  "jti": "6bad6axxxxxx1f08cdf0242ac100004",
  "nonce": "fizxxxxxf2V1aw9WTtT2wHBs8",
  "scope": [
    "openid",
    "profile",
    "email",
    "aarc",
    "entitlements",
    "eunode_projects",
    "voperson_external_affiliation"
  ],
  "sid": "xxx",
  "sub": "12e61b6e-pizza-mamma-mia-8e230c6e8c76@acc.myaccessid.org",
  "voperson_id": [
    "12e61b6e-pizza-mamma-mia-8e230c6e8c76@acc.myaccessid.org"
  ]
}
```

Example 2: consolidating existing efforts
And supporting new ones

Benchmarking our infrastructure is hard.

How do I make sure that I can sustain the rate of data I expect?

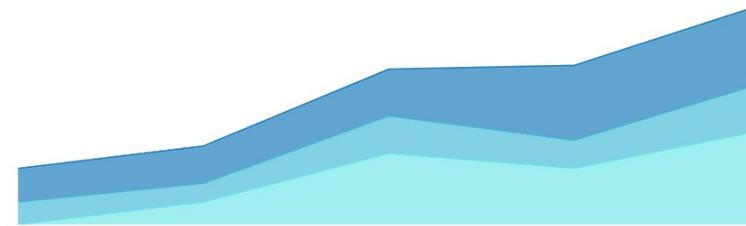Which metric do I use to evaluate the performance of my data lake?

How do I ensure my users can actually use the data, and how?

*Bonus interpretation: how do I make sure that I'm **not** creating infrastructure and tools that I think are very nice, but no one will use?*

We setup Data and Analysis challenges.

ESCAPE Data Infrastructure for Open Science
Data and Analysis challenges: what are they, and how can they be useful

A Data Challenge is an initiative that focuses on **stress-testing** various data transfer tools and methodologies, optimising network configurations, and investigating potential limitations in our RIs' infrastructure, which is partially shared in some sites.

The analysis part aims at **benchmarking** the VRE in terms of tools, compute, and workflows

The aim we have for these initiatives is to:
1. Prove that a distributed computing model à la HEP can be adopted, used, and partially **shared** (site-wise) between our RIs.

2. Facilitate RIs in testing **their own infrastructure**, with their own computing objectives; we help with the shared tooling and support.

3. Allow RIs to benefit from the **collective preparative effort**, and possibly the momentum in setting up the infrastructure and the tools.