OSCARS
Open Science Clusters' Action for Research & Society

OSCARS Consolidation and Terminology Workshop: Services and Data Sources Portfolio @ ESCAPE

Giovanni Guerrieri, Frederic Gillardo, Marion Pierre
for the ESCAPE cluster

30-09-2024

Funded by the European Union

ESCAPE: European Science Cluster of Astronomy and Particle Physics

OSCARS

Consortium of 31 members, including:

- 10 ESFRI projects & landmarks: **CTA, EST, FAIR, HL-LHC, KM3NeT, SKA, LSST, VIRGO, ESO, JIVE**
- 2 pan-European International Organizations: **CERN and ESO**
- 2 European Research Infrastructures: **EGO and JIV-ERIC**
- 4 supporting European consortia: **APPEC, ASTRONET, ECFA and NuPECC**

Budget: **15.98 M€**

Duration: **48 months** (1/2/2019 -31/1/2023)

Portfolio status

Services and Data Sources

https://projectescape.eu/services

Distributed Data Management:

- Rucio is a high level data management system catering for the needs of modern scientific experiments.
  - Exabyte-scale data, multi-billion file namespaces, diverse storage software providers across hundreds of sites.

- FTS3 is the service responsible for globally distributing exabytes of scientific data across the WLCG sites every year.
  - Low level data movement third-party orchestrator, responsible for reliable bulk transfer of files across sites.
  - Able to interact with diverse storage system solutions.

Services and Data Sources

https://projectescape.eu/services

Distributed Data Management (ctd):

- **Content Delivery and Latency Hiding mechanisms** are software-managed disk storage caching layers
  - Enhance file reusability and hide latency, allowing efficient CPU use by streaming data from the local cache.
  - Bridge scientific data present in the data lake with non-standard resources (e.g., via Edge Services, commercial clouds, HPCs, opportunistic resources).

- HiPS is an IVOA standard for the description, storage and access of large sky survey data across multiple international nodes.

Services and Data Sources

https://projectescape.eu/services

Analysis Frameworks for Scientific Computing:

- The ESFRI Science Analysis Platform (ESAP) is a platform-service for data analysis
  - Find data, services, and resources across heterogeneous infrastructures.

- The Virtual Research Environment (VRE) is an analysis platform aiming to facilitate the development of end-to-end physics workflows,
  - Manage, access and preserve data and analyses in compliance with FAIR principles.

Services and Data Sources

https://projectescape.eu/services

Interoperability Frameworks for Data and Services:

- The Virtual Observatory is an international astronomical community-based initiative.
  - Global electronic access to the available astronomical data archives of space and ground-based observatories and other sky survey databases.

Services and Data Sources

https://projectescape.eu/services

Preservation and re-interpretation of analysis workflows and results:

- Zenodo is a general-purpose open repository
  - Deposit research papers, data sets, research software, and any other digital artefacts, and provides them with a persistent identifier (PID) such as a Digital Object Identifier (DOI).

- The Open-source Scientific Software and Service Repository (OSSR) is an open-access repository for collaborative software development, uptake, and reuse.
  - Share and find software and services developed during the ESCAPE project.

- REANA is a reusable and reproducible research data analysis platform.
  - Structure their input data, analysis code, containerised environments and computational workflows so that the analysis can be instantiated, run and preserved on remote compute clouds.

Services and Data Sources

https://projectescape.eu/services

Widening scientific participation:

- The ESCAPE Citizen Science (CS) is an astronomy and astroparticle physics programme of crowdsourced data mining.
  - Train and educate both the scientific community and the wider science-inclined public.

Composability

Composability

Composability is a system design principle that deals with the inter-relationships of components. (source)

A highly composable system provides components that can be selected and assembled in various combinations to satisfy specific user requirements.

Analysis Facility: the infrastructure and services that provide ***integrated*** data, software and computational resources to execute one or more elements of an analysis workflow. (source)

Our use case for composable services:

Composability

An analysis lifecycle

Composability

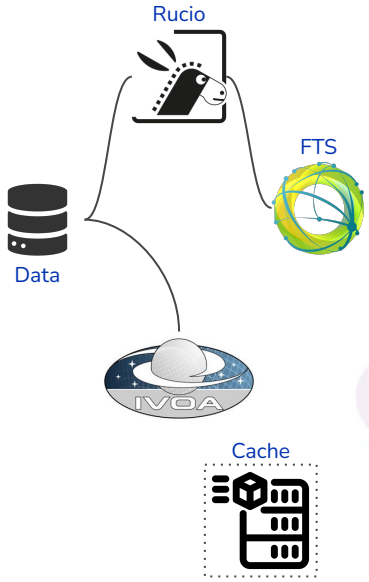
An analysis lifecycle

Data

Composability

An analysis lifecycle

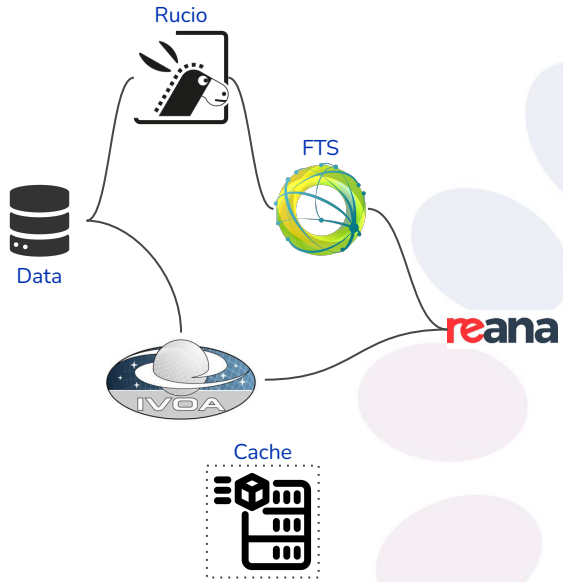
Interactions



Data gathering → Data registration and access → Data processing → Data analysis → Results publication → Analysis and data preservation → time

Composability

An analysis lifecycle

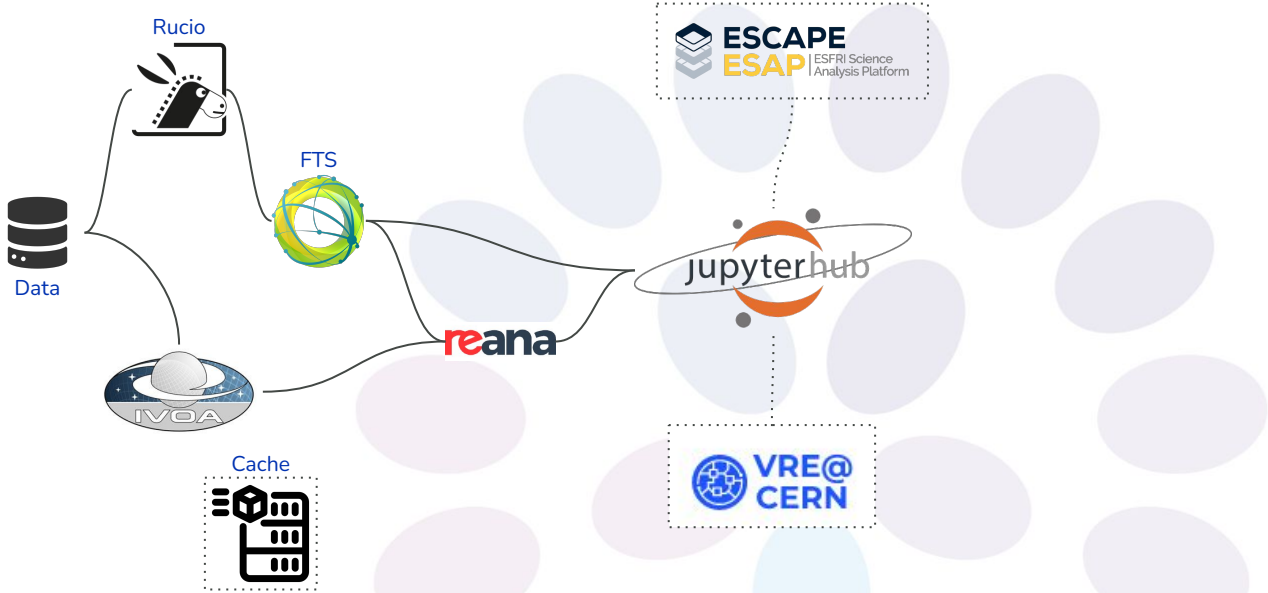


Interactions

Data gathering → Data registration and access → Data processing → Data analysis → Results publication → Analysis and data preservation → time

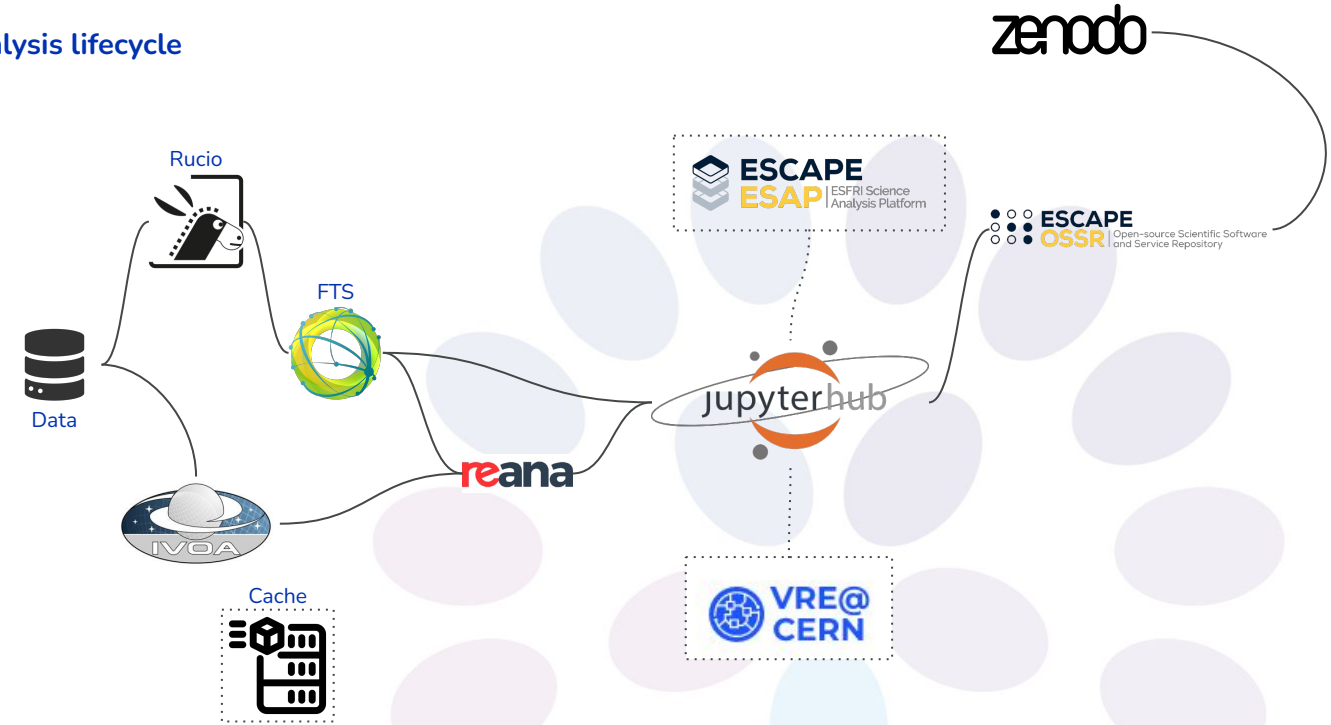
Composability

An analysis lifecycle



Composability

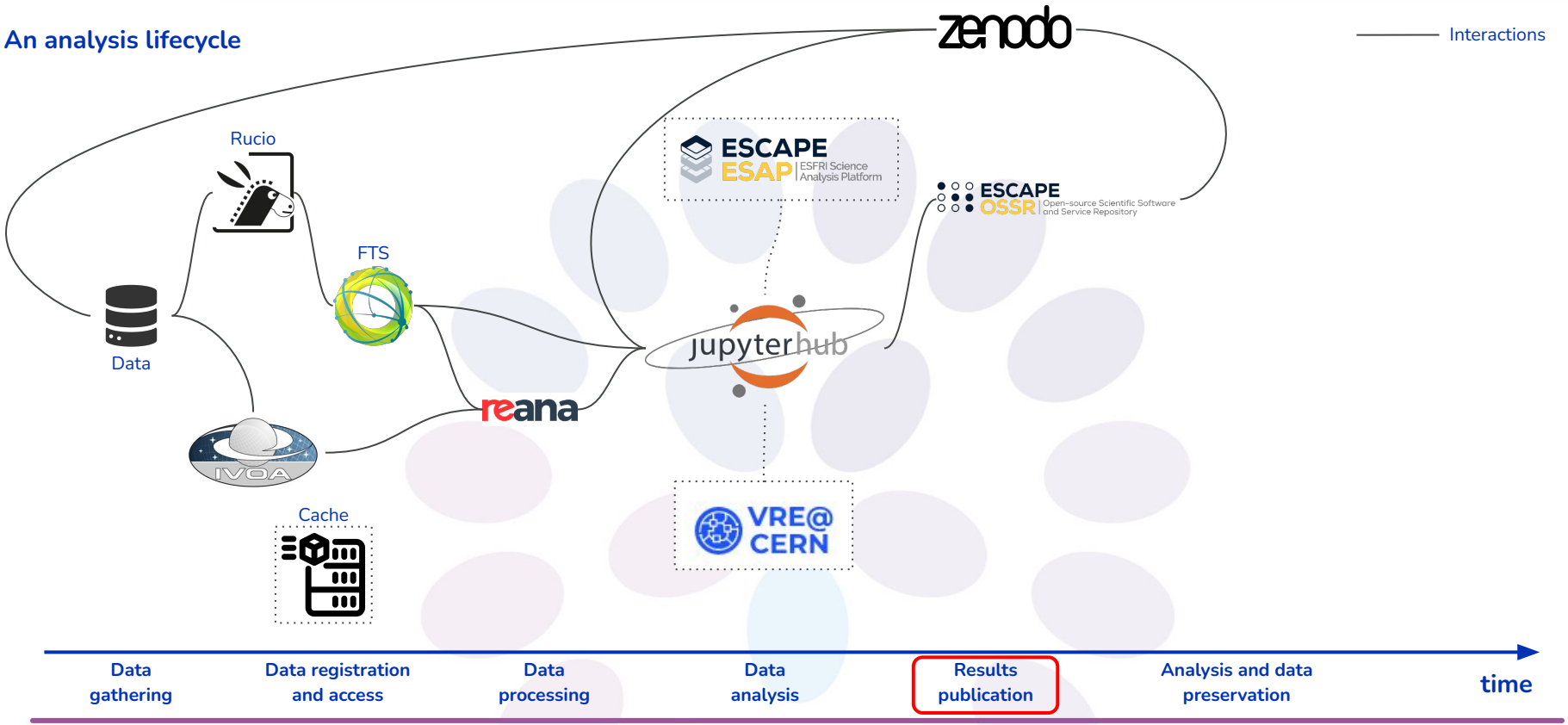
An analysis lifecycle



Data gathering | Data registration and access | Data processing | Data analysis | Results publication | Analysis and data preservation → time

Composability

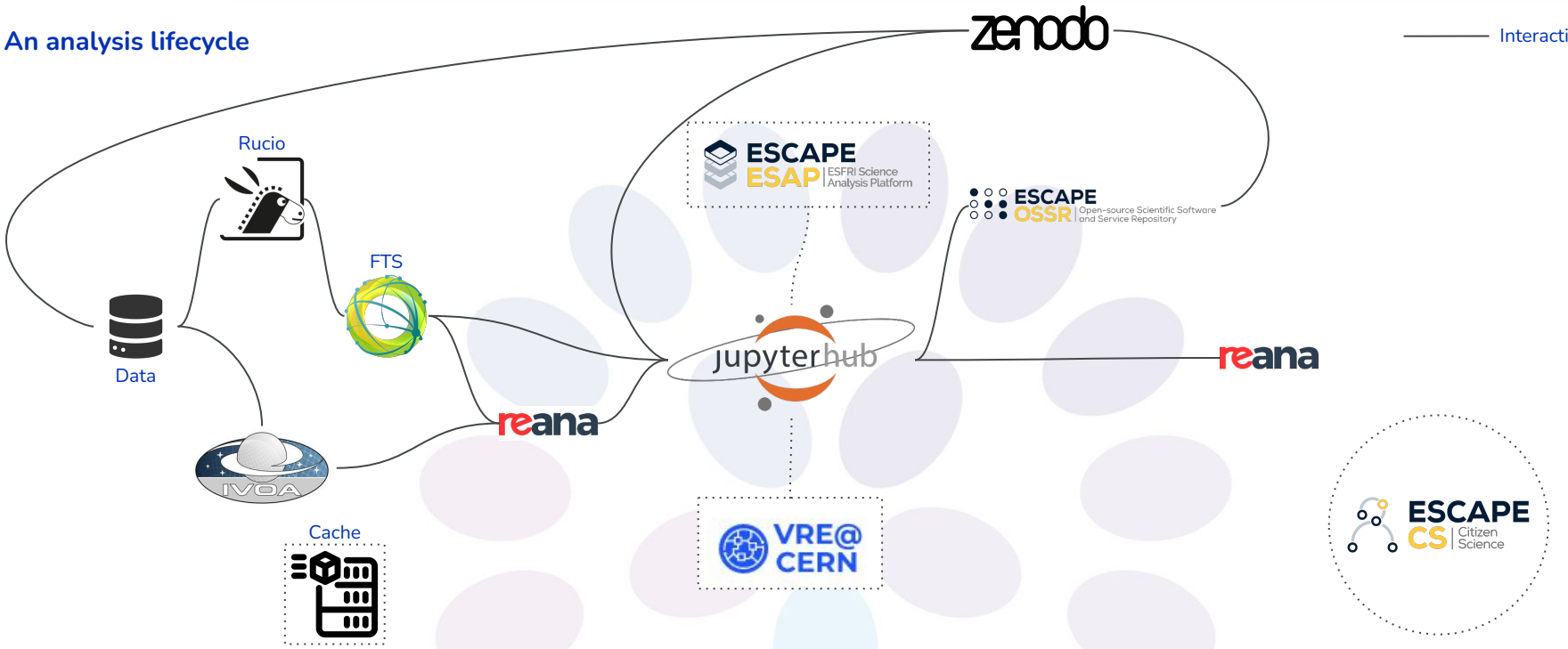
An analysis lifecycle



Composability

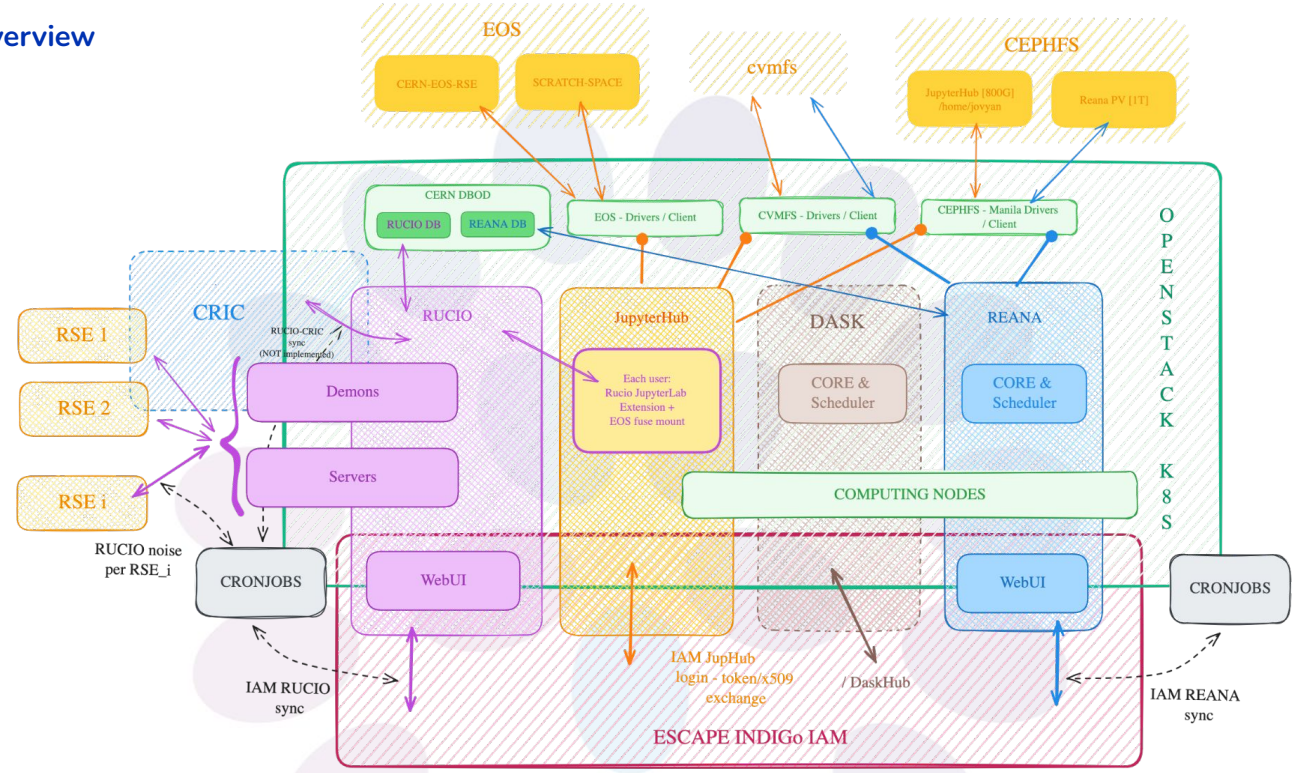
OSCARS

An analysis lifecycle



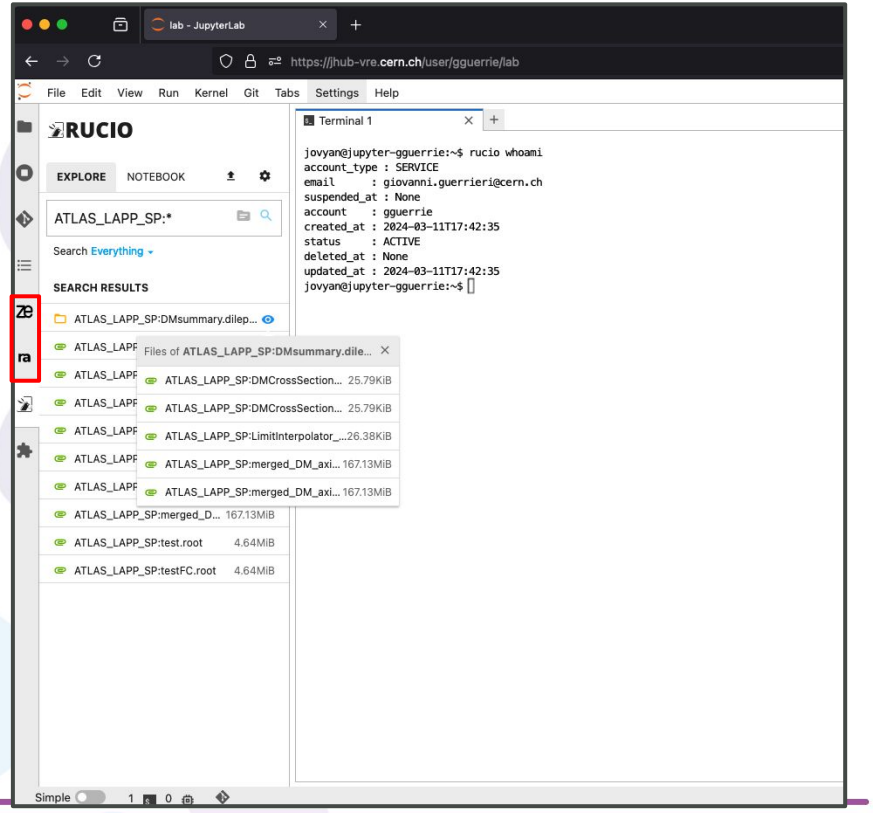
Demonstrators

The CERN VRE: an overview



The CERN VRE: an overview

- Modularity
  - Integrates software, tools and packages.
  - Can be configured to connect to remote storage and computing resources.
- Flexibility
  - Ad-hoc workflows can be created via easily editable declarative files.
  - Can be installed on different machines independent of CERN restrictions.
- Reproducibility
  - Deployment is kept simple and documented to be used as a blueprint for other research infrastructures.
  - Allows analysis preservation.
- Features recently implemented
  - Reana extension to manage and create workflows.
  - Zenodo extension to download and upload data, software, publications.



Summary

- The participating ESFRIs in the ESCAPE cluster commissioned the usage of common tools for Data Management and Data Analysis in joint Data Challenges at scale
  - Demonstrated adequacy and integration for different sciences: from raw data recording to standard analysis workflows.
  - Composability is achievable (not only) in Jupyter-based analysis facilities.
- The standing ESCAPE collaboration is supporting integration of new sciences through dedicated working groups, providing demonstrators platforms and pilot projects together with first hand expertise
  - Developed and tested throughout the ESCAPE project.
  - Expanding reach towards ESCAPE-external use cases (e.g. HEP Open Data, interTwin).

Challenges and future steps

- Unprecedented Data Management requirements of new ESFRIs are hinting to a need of paradigm shift in the way computing is organized across a large diversity of scientific activities
  - Favoring economies of scale: personpower, knowledge transfer and resources usage. Towards an scalable and unified systems.
  - Capable of seamlessly cater with multi-**exabyte** scale data lifecycle needs: data and metadata management, data access and analysis at large, identity management and access policies.
- Analysis frameworks need flexibility
  - Facilitate streamlined access to data and software.
  - Accelerate the research process.
  - Need to build a shared community of developers and operators.
- Different users need the same authentication
  - Secure and seamless access to shared resources.
  - Ensure the integrity of identities across diverse platforms.