Championing Scientific Data Management: Rucio and the ESCAPE project experience

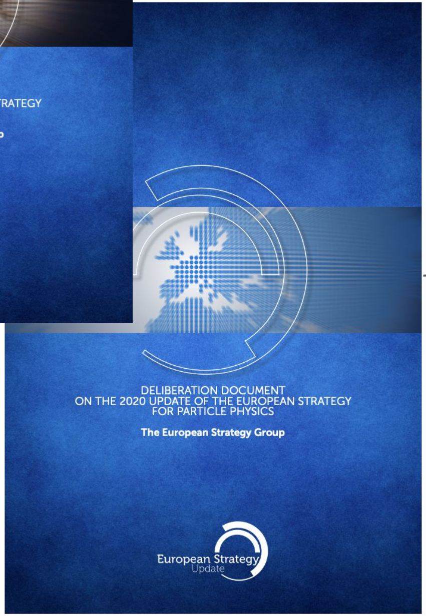
Xavier Espinal (CERN, ESCAPE WP2 lead)

Rucio Workshop, Lancaster, 10-11 November 2022

2020 European Strategy for Particle Physics

4 Other essential scientific activities for particle physics

[Full document]



[Full document]

D. Large-scale data-intensive software and computing infrastructures are an essential ingredient to particle physics research programmes. The community faces major challenges in this area, notably with a view to the HL-LHC. As a result, the software and computing models used in particle physics research must evolve to meet the future needs of the field. ***The community must vigorously pursue common, coordinated R&D efforts in collaboration with other fields of science and industry, to develop software and computing infrastructures that exploit recent advances in information technology and data science. Further development of internal policies on open data and data preservation should be encouraged, and an adequate level of resources invested in their implementation.***

The scientific outcomes of particle physics experiments are made possible by the development of an efficient computing and software infrastructure. Computing and software are profound R&D topics in their own right and are essential to sustain and enhance particle physics research capabilities. There is a need for strong community-wide coordination for computing and software R&D activities, and for the development of common coordinating structures that will promote coherence in these activities, long-term planning and effective means of exploiting synergies with other disciplines and industry. Some recently initiated examples are the HEP Software Foundation addressing the common computing and software challenges related to particle physics, and ESCAPE (European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures) exploring the synergies in the areas of astronomy, astroparticle and accelerator-based particle physics.

Science Projects

ESCAPE

European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructures

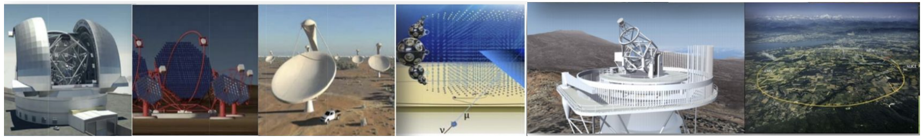
EUROPEAN OPEN SCIENCE CLOUD

Data centres

- Address **RI's upcoming needs** in Data Management, Data Access and User Analysis for Astro-particle, Radio-astronomy, Cosmology, Particle and Nuclear Physics

- Provide a fully working **common data infrastructure** "The ESCAPE Data Lake" to test novel data management tools and models, giving the opportunity to influence and steer its development.

- **Expand collaborations** and foster involvement within diverse Scientific Communities. Maintain and strengthen collaborations with related EC initiatives and projects



The ESCAPE Work Program

Data Lake:
Build a scalable, federated, data infrastructure as the basis of open science for the ESFRI projects within ESCAPE.

Software Repository:
Repository of "scientific software" as a major component of the “data” to be curated in EOSC.

Science Platforms:
Flexible science platforms to enable the open data analysis tailored by and for each facility as well as a global one for transversal workflows.

Citizen Science:
Open gateway for citizen science on ESCAPE data archives and ESFRI community

Virtual Observatory:
Extend the VO *FAIR* standards, methods and to a broader scientific context; prepare the VO to interface the large data volumes of next facilities.

The ESCAPE Data Infrastructure for Open Science

The ESCAPE Scientific-Data Lake is a policy-driven, reliable distributed data infrastructure capable of managing Exabyte-scale data sets. And able to deliver data on-demand at low latency to all types of processing facilities

Services operated by the ESCAPE partner institutes

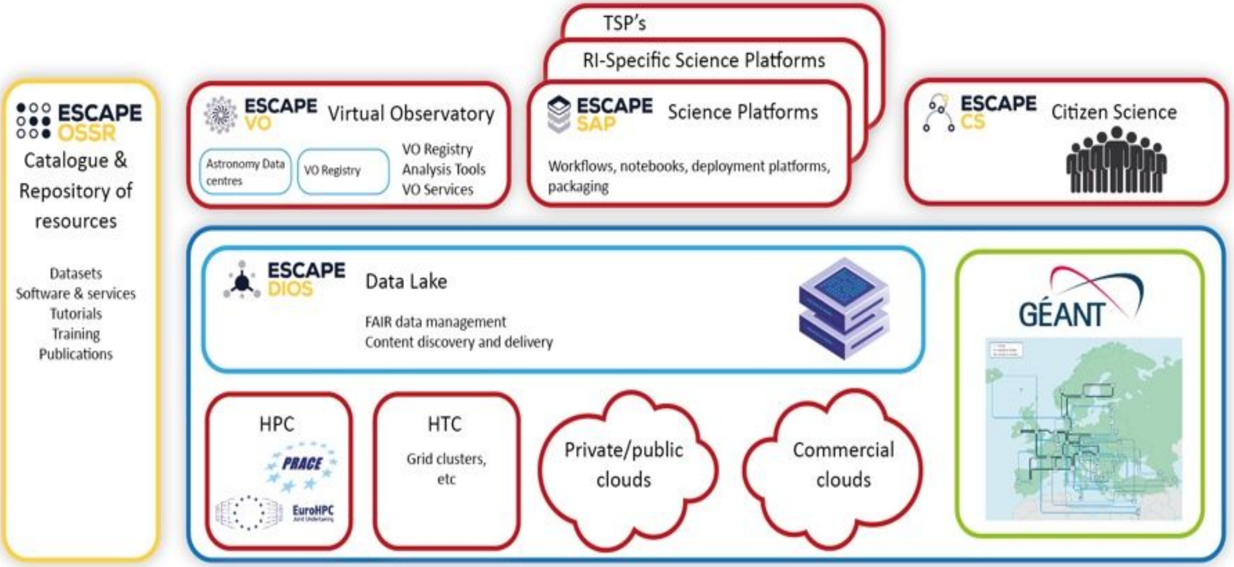
Petabyte scale storage: DESY, SURF-SARA, IN2P3-CC, CERN, IFAE-PIC, LAPP, GSI and INFN (CNAF, ROMA and Napoli)

Data management and storage orchestration (Rucio)

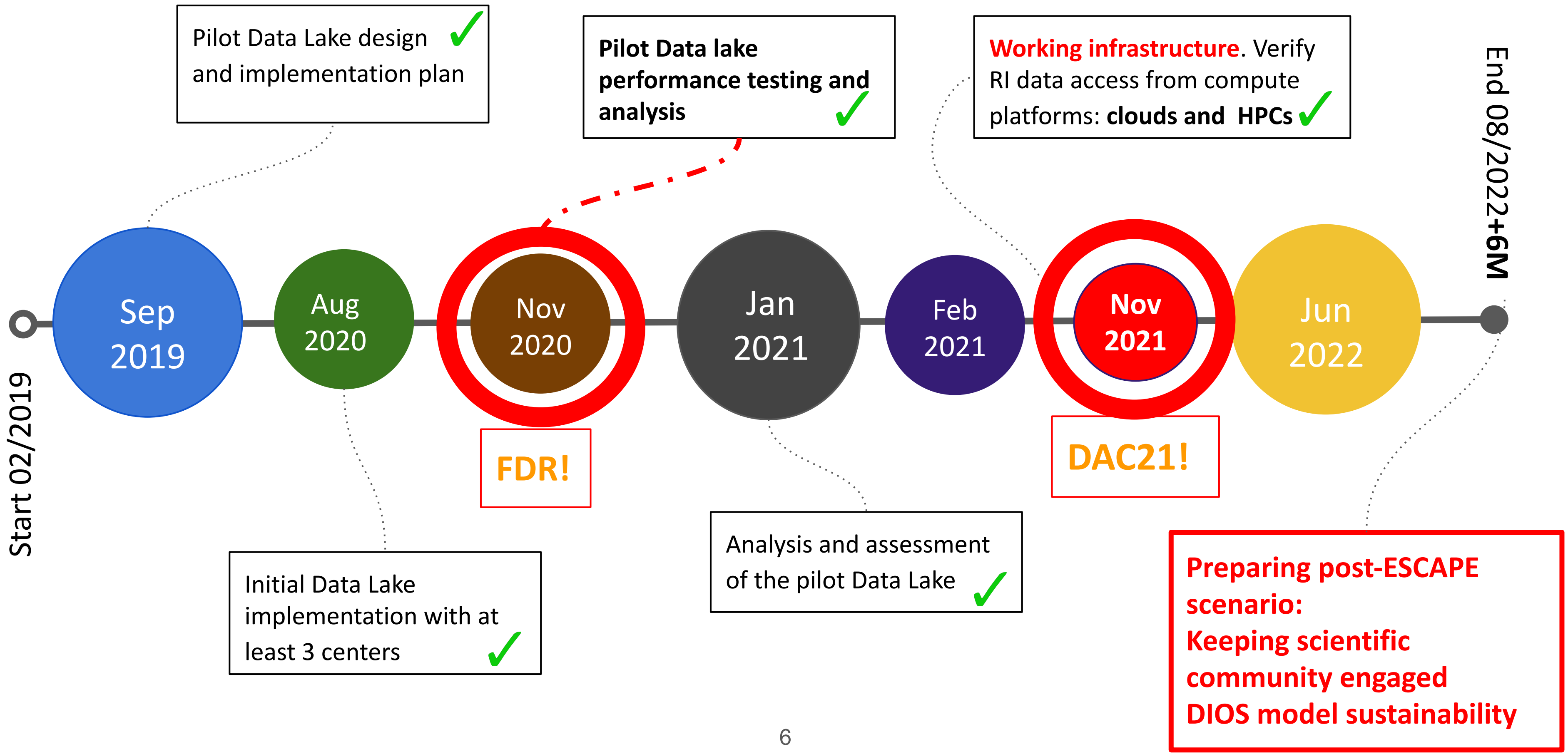
File transfer and data movement services (FTS)

Global Data Lake Information System (CRIC

ESCAPE IAM: common Auth/Authz/IM (AAI)

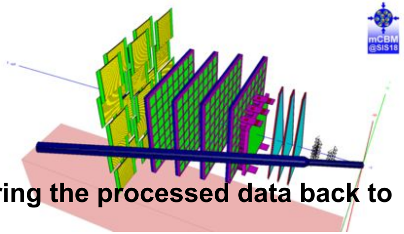


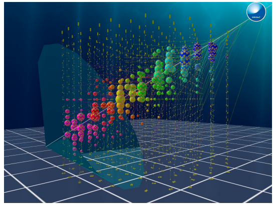
ESCAPE DIOS Roadmap



Putting the system to work (1/4)

FAIR	• **Registration of RAW data acquired by the mCBM detector on FAIR-ROOT** • Ingestion and replication of simulated R3B data • Ingestion and replication of simulated and digitised raw PANDA fallback data • Particle-transport and digitisation of Monte-Carlo events • Live ingestion of simulated data • **Retrieval of stored RAW data from the data-lake, processing of the data and storing the processed data back to the data lake** • Retrieve raw mCBM data from the data lake, run reconstruction on it and store the results • Analyse simulated R3B data stored in the data lake, upload resulting histograms and bitmaps to the DL
KM3NeT	• **Ingestion of raw data from the storage at the KM3Net shore station to the Data Lake, and policy-based data replication across the Data Lake infrastructure** • Download raw data from data lake, calibrate, and ingest into the data lake • Conversion of calibrated and reconstructed data to a format for high level analysis. Data findable and data conversion using DLaaS interface. Data ingest and replication via DLaaS Executed only once per large data set
cta cherenkov telescope array	• Long-haul transfer and replication. CTA-RUCIO @PIC: non-deterministic (La Palma) and deterministic (PIC) RSEs • **Data reprocessing. Primary data stored and findable in the datalake (using the CTA Rucio instance). Data is accessed and processed. New data products stored back in the Data Lake** • Data analysis. Data access via Jupyterhub/mybinder via ESAP. Higher-level analysis products produced





Putting the system to work (2/4)

HiLumi HL-LHC PROJECT ATLAS EXPERIMENT	● Exercises (data production, replication and documentation) before and during the DAC21. Include the creation of datasets for real-kind final user analysis examples using current open-access datasets. ~200*10 = 2000 files uploaded in the Datalake. Two copies of such files (rules) into at least two RSE's ● **User analysis pipeline tests on experimental particle physics by using augmented open data** (http://opendata.atlas.cern/software/). Testing and validating the reading access of the samples via de Jupyter rucio extension, and running multiple analysis pipelines. 
MAGIC Major Atmospheric Gamma Imaging Cerenkov Telescopes	● **Long haul raw data ingestion and replication. Data is successfully transferred from the telescope station and replicated to the Data Lake, file deleted on the telescope storage buffer.** ● Data transfer monitored. Data can be discovered using the CTA-RUCIO instance. ● Data reprocessing. Replication of DL3 file. Validating the reading access of the samples via de Jupyter rucio extension. Multiple analysis using gammapy library. 
LOFAR	● **Ingestion of LOFAR data from a remote site** to the Data Lake. Data transfer and replication into off-site storage, after successful replication delete data at the source ● **Process data in the Data Lake at an external location, combine results with other astronomical data to produce a multiwavelength image.** ● Include a read-only RSE to a location outside the data lake. Get data from there into the DL. ● Extending use cases by using larger files and leveraging several QoS, running all processing in the DLaaS, requiring a the availability of specific LOFAR software in the DLaaS. 

Putting the system to work (3/4)

SKAO

- Data replication. Data in correct place in timely manner.
- **Long haul data replication. SKAO Rucio (Australia and South-Africa to UK RSEs), using the RUCIO SKA instance.**
- End-to-end proof of concept data lifecycle test, AUS/SA to northern hemisphere sites
- Data Analysis. Successfully running SKAO's science data challenge pipeline using data stored in datalake.

HiLumi HL-LHC PROJECT / **CMS**

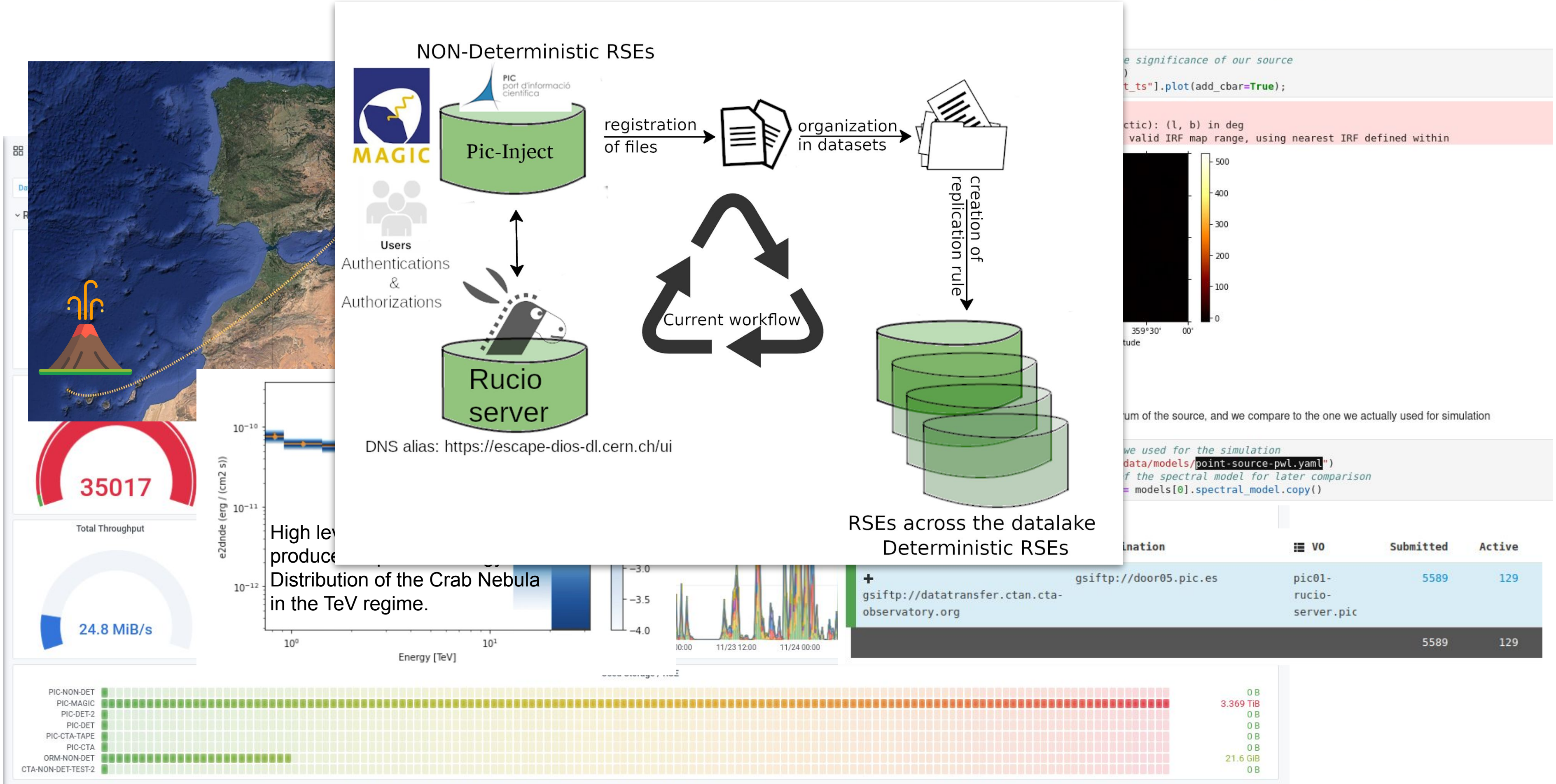
- **Multi purpose Analysis Facility PoC with data access via DASK (workload orchestrator) leveraging computing at Marconi (HPC) and large batch clusters**
- Access control for embargo data, test in CNAF and DESY
- Content delivery and caching: XCache Protocol Translation: xroot internal vs http External for Data Lake data transfer. Performance comparisons for Analysis workflows

DL interface with local and heterogeneous resources, CDN and caching

LSST / **VERA C. RUBIN OBSERVATORY**

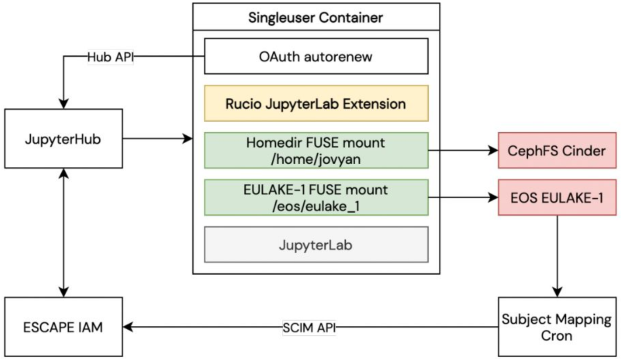
- **Simulate replication of one night's worth of raw images data between two Vera C. Rubin data facilities, perform the exercise several times.** Each iteration is composed of 15TB, 800k files, ideally to be replicated in 12 hours or less
- Incorporate SLAC National Accelerator Laboratory (US) in the data replication chain (postponed)

Putting the system to work (4/4)



Rucio Jupyterlab extension and the ESCAPE DLaaS

- Joint development between ESCAPE CERN team and the Rucio team, via GSoC and Summer Student Program
- Provide the end-user with a Notebook ready-to-be-used and fully integrated with the Data Lake
- OIDC (token based AAI) integration
- Data browsing, file moving, data download and upload
- Storage cache layer interplay (XCache)
- FUSE mount (posix filesystem mount of the local storage), terminal capabilities (RUCIO-cli)
- Further integration with Analysis Platforms, catalogues and data/metadata sources

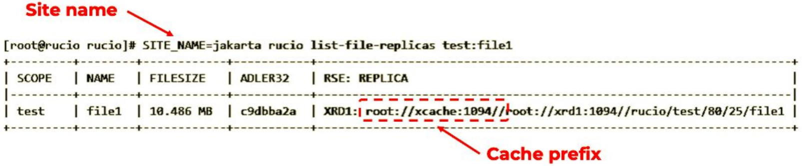


DLaaS demo available here

https://github.com/rucio/jupyterlab-extension

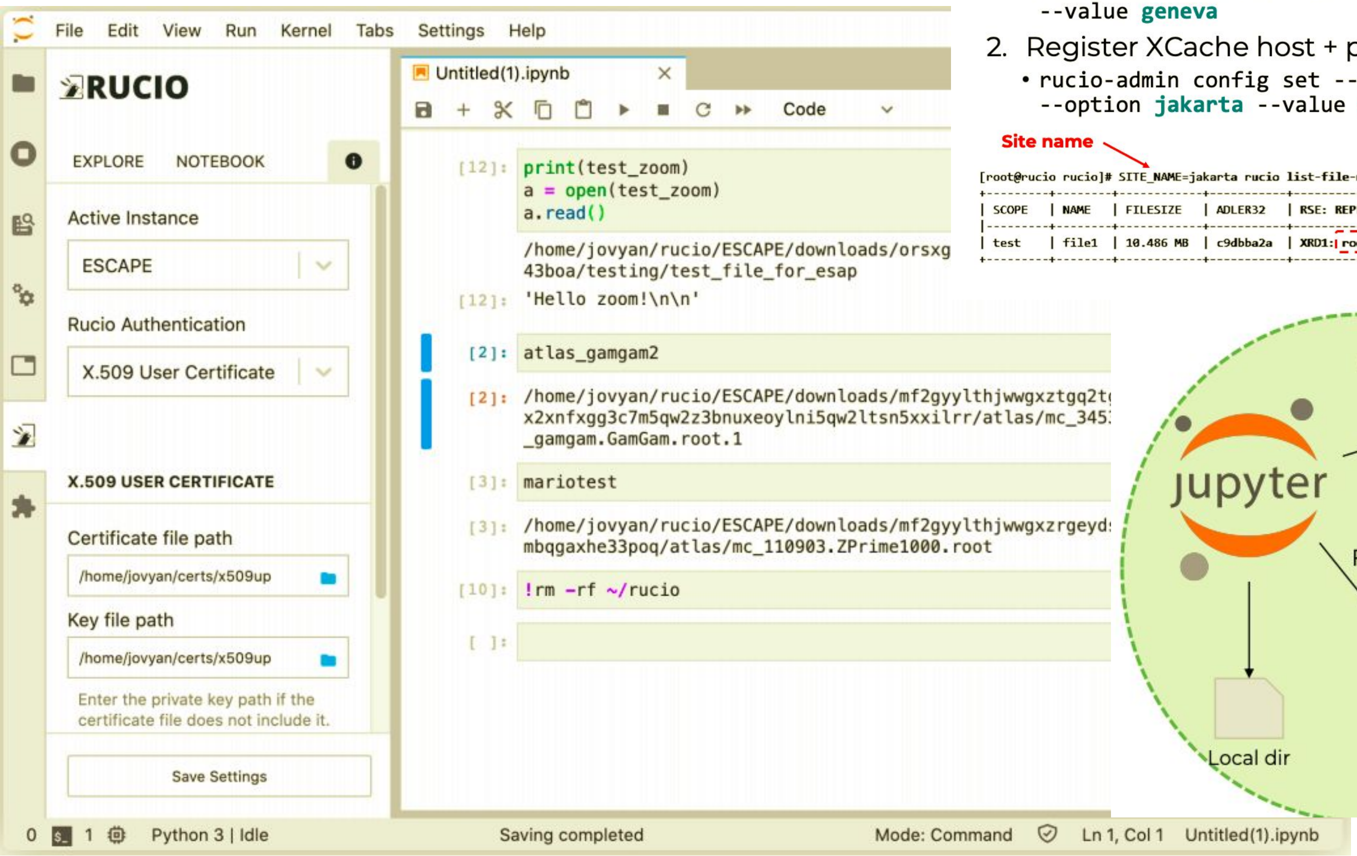
Rucio Jupyterlab extension and the ESCAPE DLaaS

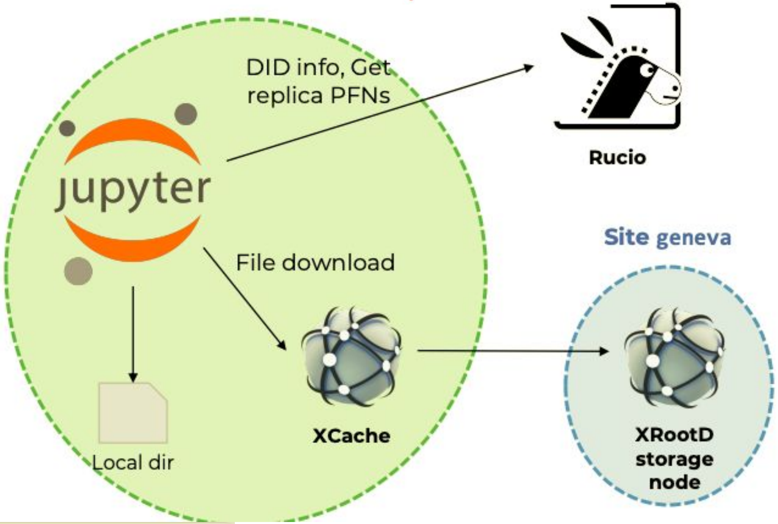
1. Assign a site name to existing RSE
   - `rucio-admin rse set-attribute --rse XRD1 --key site --value geneva`
2. Register XCache host + port to Rucio
   - `rucio-admin config set --section root-proxy-internal --option jakarta --value xcache:1094`



[root@rucio rucio]# SITE_NAME=jakarta rucio list-file-replicas test:file1

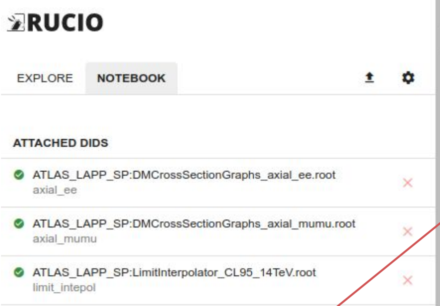
SCOPE	NAME	FILESIZE	ADLER32	RSE: REPLICA
test	file1	10.486 MB	c9dbba2a	XRD1: root://xcache:1094//root://xrd1:1094//rucio/test/80/25/file1

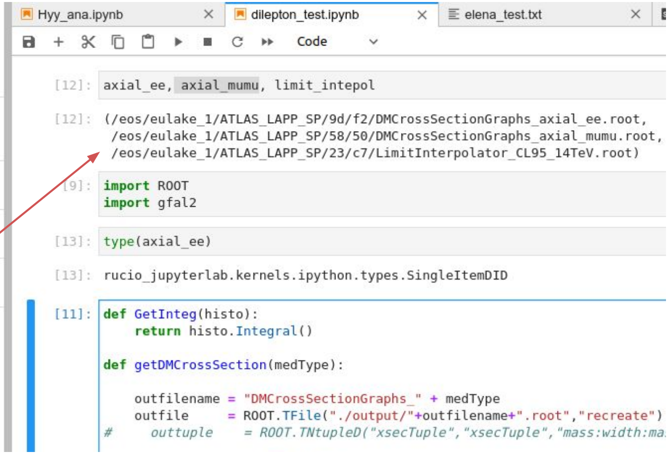




DLaaS (VRE) example: ATLAS Dark Matter Reinterpretation - Dilepton Resonance

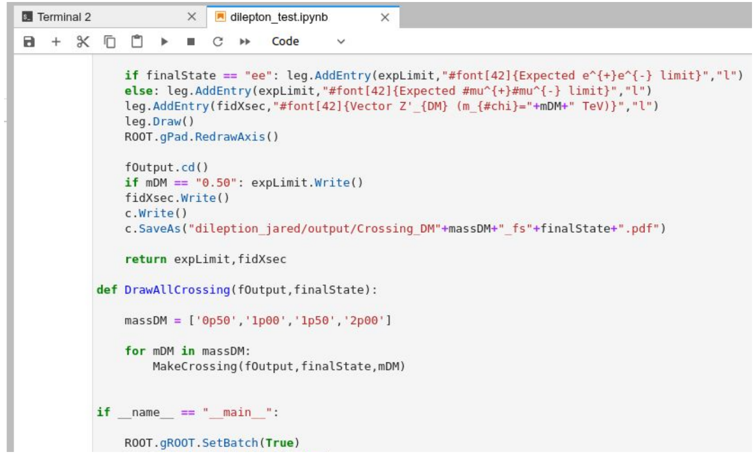
1. Import files from the Data Lake as variables into the notebook:



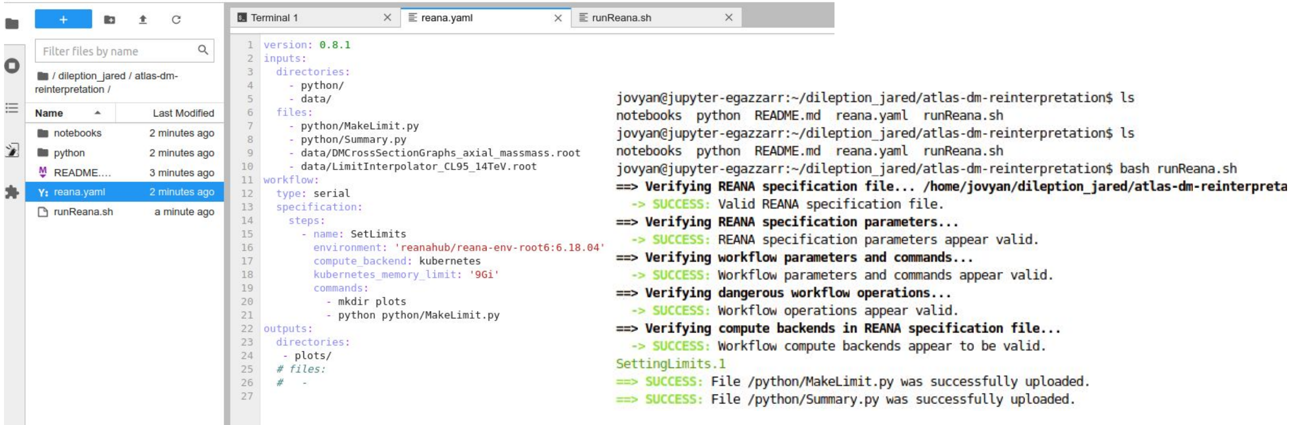


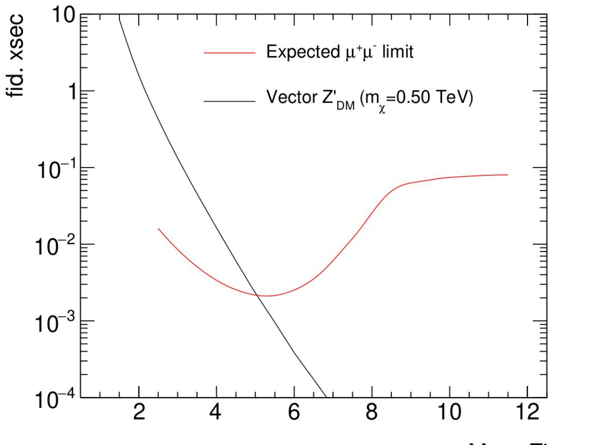
data is imported from the Data Lake as a notebook variable

3. Final results



2. Ensure reusability and reproducibility: REANA workflows





Takeaway 1 - Computing Challenges are global!

Computing scale increasing in all scientific domains, different computing models but **large commonalities** across scientific communities.

Common needs for:

- Large Scale Data Management and Data Transfer Services
- Distributed Storage and Computing integration and orchestration
- Provide services and data processing abilities to large and distributed user communities
- Analysis reproducibility and reusability
- Willingness to promote Open Data and Open Science: outreach and citizen science activities

Takeaway 2 - Technologies and Knowledge sharing

- The ESCAPE project provided a **fully working data infrastructure** where RIs could **test and gain experience** on data management services and tools
- RIs deployed their own **per-experiment** Rucio instances
  - Fostering a realistic “data lake model” across scientific domains: sites serving different experiments. Experiments defining their own “lakes”
- As a result RIs are taking Rucio **onboard** or doing realistic testing for potential future adoption
  - SKAO, CTAO, LSST, KM3Net
  - Several new communities showed interest: ESA? PaN sciences?
- Potential **positive impact** on the Rucio community
  - Enlarging its membership and opening possibilities for cross-cooperation (coding, testing, operation etc.). This is fundamental for sustainability… and to evolve Rucio towards a global scientific data management system

Takeaway 2 - Technologies and Knowledge sharing

e.g. The SKAO Million files injection tests

ls

During this preparatory phase, it was discovered that Rucio's injector daemon was not able to natively cope with handling a replication rule with a million files even if it were given large amounts of resources. Through excellent collaboration from the Rucio development team we were able to deploy a new configuration that more efficiently handled the daemon's memory usage. The daemon attempted to process all the files in a rule all at once thus running into memory issues for rules with large file numbers. The new config works well because it processes the files in a rule in batches, thus allowing the injector daemon to, ideally, deal with any number of files in a rule. This new configuration was already built but had not been tested by the Rucio team - in our Million-file test it was found to be very effective and as a result, this new configuration has been retained.

- Potential **positive impact** on the Rucio community
  - Enlarging its membership and opening possibilities for cross-cooperation (coding, testing, operation etc.). This is fundamental for sustainability… and to evolve Rucio towards a global scientific data management system

Takeaway 3 - New frontiers

- Steady demand for distributed data management tools: simple storage/data management approaches does not work(=scale) anymore
- Some communities confronted with new distributed computing paradigms
  - Integrate different storage on different locations
  - Different types of resources, data replication and data access
  - Fulfill the needs of distributed user communities and sites
  - **Different computing models** wrt HEP, e.g. PaN facilities/community
- These communities feel the knowledge gap to have a hands-on experience with Rucio is high (complex?)
  - Minimal Viable Rucio or uRucio? Providing “simple” LFN/SFN-FTS-Storage relationship for small DM requirements.
  - Non-deterministic namespaces fostering easy catalogue/namespace feel to facilitate onboarding?
  - Containerised version ready for deployment, ready to be used for testing? Some effort have been done within the ESCAPE-CERN towards this direction

Outlook

Rucio is at the forefront on key activities in future distributed computing activities

Consolidate efforts on Rucio R&D instances?

- Joint **operational and maintenance** efforts in the several R&D activities? Common Rucio infrastructure for DOMA, WLCG and EC-projects? (virtualization platform, CD/CI, Terraform, etc.), joint instances?
- To benefit from development on several areas: metadata, tokens, more robust (and wide) testing on new releases
  - ESCAPE ESFRI RIs also participating actively and perhaps interested in common deploy and operational models

Future projects and joint funding opportunities, e.g. the ESCAPE Open Collaboration Agreement

- Signed by the ESCAPE ESFRs is providing a formal statement to foster scientific collaboration in computing, maintaining the community together and following-up on ESCAPE work program
- Involve computing experts from ESCAPE RIs and other scientific communities together with LHC community
- Identify **common needs and initiatives**, common interest will naturally keep activities active and in shape to apply for funding windows

Last but not least

Thanks to the Rucio team

Their openness, patience for understanding new communities and their readiness to collaborate is paramount to feed the community spirit