| | |
|---|---|
| Project Title | European Science Cluster of Astronomy & Particle physics ESF(RI)s research Infrastructure |
| Project Acronym | ESCAPE |
| Grant Agreement No | 824064 |
| Instrument | Research and Innovation Action (RIA) |
| Topic | Connecting ESF(RI)s infrastructures through Cluster projects (INFRA-EOSC-4-2018) |
| Start Date of Project | 01.02.2019 |
| Duration of Project | 42 Months |
| Project Website | www.projectescape.eu |

# ESCAPE DIOS FINAL DELIVERABLE (D2.3)

***"Final assessment and analysis of the full prototypes, outlook for further development and deployment towards full production services within EOSC"***

| | |
|---|---|
| Work Package | WP2 - DIOS |
| Lead Author (Org) | Xavier Espinal (CERN), Rosie Bolton (SKAO) |
| Date | 28/06/2022 |

## Table of Contents



# 1. ESCAPE DIOS, the Data Infrastructure for Open Science

The ESCAPE Data Infrastructure for Open Science (DIOS) is a federated data infrastructure able to cater for the multi-Exabyte needs of current and future ESFRIs (European Strategic Forum for Research Infrastructures) and Research Infrastructures (RIs). It supports FAIR (Findable, Accessible, Interoperable, Reusable) data management principles, serving global and varied scientific communities scalably and effectively. The DIOS Data Lake model integrates a common set of tools to orchestrate resources provided by computing centres to one or many ESF(RI)s . This common set of tools facilitates data management, data transfer, data access, information system and a common identity management framework. The Data Lake model presents resources to the users as a single entity, hiding the inherent complexity and easing the implementation of policies, rules and data life-cycles acting on the Data Lake infrastructure as a whole.

The primary objective of ESCAPE DIOS is to prototype a reliable and scalable federated data infrastructure for the European Open Science Cloud (EOSC[1]), which enables the provisioning of data processing in a flexible and adaptable way, through existing and future European e-infrastructures. The purpose is to enable science communities to build open access data repositories according to the FAIR principles. The requirements and scale is driven by the Exabyte-scale Data Challenges of the ESF(RI)s and major Research Infrastructures which are the major stakeholders of ESCAPE. The Data Lake prototype implemented in ESCAPE DIOS provides a fully operational system addressing several of the main requirements from the participating ESF(RI)s on data management, data access, data analysis and Authentication and Authorization Infrastructures (AAI).

The main focus in ESCAPE DIOS during the second phase of the project was to assess and evolve the pilot Data Lake that was designed and deployed during the first phase. The early pilot Data Lake assessment culminated in a joint exercise labelled as Full Dress Rehearsal (FDR20). The goal was to demonstrate the fulfilment of communities' needs and the addressing of required functionalities and experiment-specific use cases. By November 2020, the Data Lake pilot infrastructure was exercised during a 24-hour production-like window where experiments executed relevant workloads

[1] https://eosc-portal.eu/

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.

covering a wide range of activities, e.g. from experiments' data recording of detectors and sources to data browsing and access via notebooks for user analysis purposes. The FDR20 served to confirm the durability of the DIOS Data Lake and its success in addressing the actual needs of the ESF(RI)s present in ESCAPE. Nevertheless, the FDR20 exercise identified areas for improvement and goals for the following year.

The next major challenge was the Data and Analysis Challenge (DAC21), a 10 day period in November 2021 during which the ESF(RI)s in ESCAPE ran production-like data management, processing and analysis workloads. These included data acquisition activities from data sources, policy-driven data replication and data lifecycle implementation. Data processing was a fundamental target of the DAC21, therefore a big emphasis was put to push different use cases of processing activities including interplay possibilities using large scale resources (batch systems and clouds) and user analysis-oriented platforms (online notebooks and analysis platforms). During DAC21 several ESF(RI)s deployed their own Data Lakes demonstrating their interoperability and integration in a common large data storage infrastructure.

The Data Lake model federates different distributed storage systems via a high level data management layer: Rucio. The different storage endpoints that constitute the global Data Lake system are the Rucio Storage Elements (RSEs), each of them mapping to a storage endpoint offered by the resource provider. Rucio enables file upload and download capabilities mapping the client-server interaction to the RSEs supporting different transfer protocols (HTTP/WebDAV, xroot, GridFTP, S3). The Rucio system is a policy-driven, rule-based Data Management system enabling data lifecycle capabilities for the users and the experiments.

The ESCAPE Data Lake has progressively increased in storage capacity and in complexity, federating 28 Rucio Storage Elements (RSEs) provided by 10 partner sites. These have an aggregate capacity of ~1PB and currently host 4.4M files from the participating ESF(RI)s (Figure 1). The storage endpoints are heterogeneous in terms of size and technology (EOS, DPM, dCache, StoRM, XRootD). Integration of heterogeneous and opportunistic resources, such as HPC elements, and AWS and Openstack-based storage endpoints have also been certified.

The deployment of the Data Lake infrastructure is Kubernetes (k8s) based, providing redundancy and resilience for the key components of the Data Management system. This ability to deploy, customise, and manage services at convenience has allowed ESCAPE ESF(RI)s to start deploying and operating their own Data Lakes. The adoption of the common infrastructure is in line with the goal to allow each ESF(RI) to operate such services and tools according to its specific scientific needs beyond the ESCAPE project term, ensuring project sustainability.

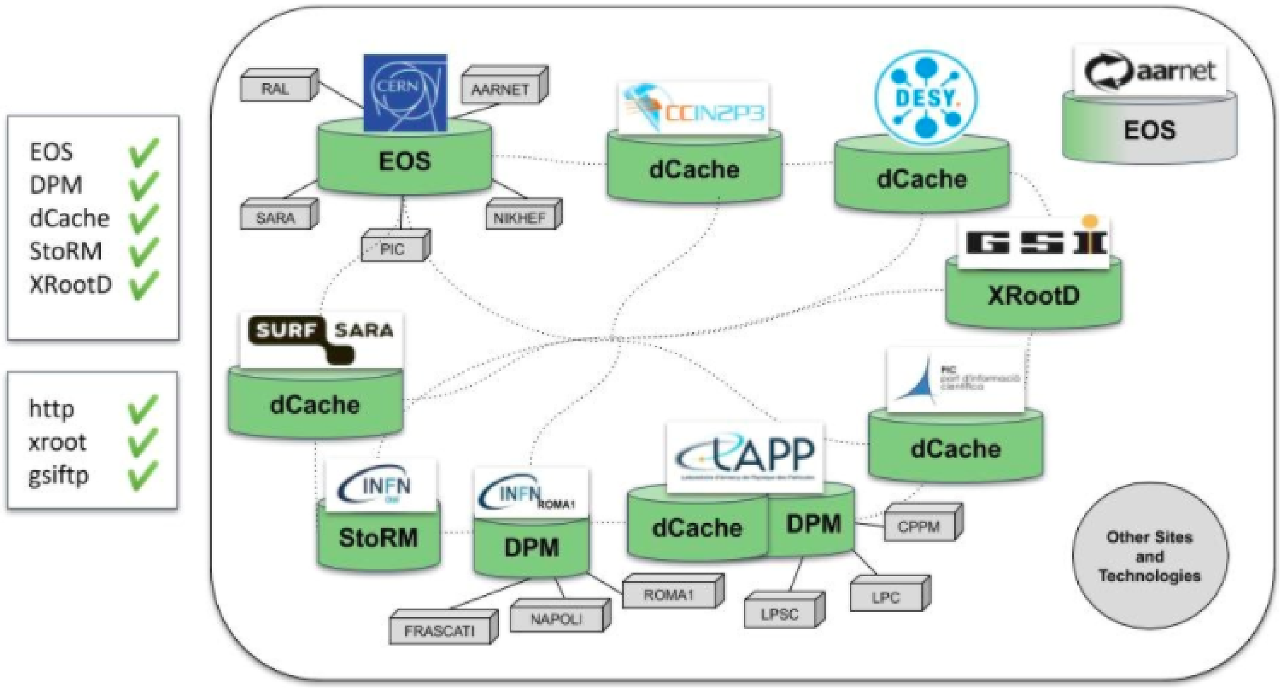


*Figure 1. : Pilot Data Lake overview. 10 storage endpoints from the project partners INFN-CNAF, INFN-ROMA, INFN-Napoli, DESY, SURF-SARA, IN2P3-CC, CERN, IFAE-PIC, LAPP and GSI harnessing different storage technologies dCache, DPM, XRootD, EOS, StoRM and providing access through different protocols: http, GridFTP and xroot.*

A cross work package collaboration has seen DIOS contributing to the ESAP (ESCAPE Science Analysis Platform) with a ready-to-be-used product named "Data Lake-as-a-Service" (DLaaS). The focus was on further integration of the Data Lake, data access, and the related data management capabilities with the activities ongoing in the area of Science Platforms. The work focused on developing a state-of-the-art "data and analysis portal" as an interface to the Data Lake offering a wide range of possibilities, from very simple I/O access to more complex workflows such as enabling content delivery technologies and integration of local storage facilities at the facility hosting the notebook. The DLaaS project allows end-users to interact with the Data Lake in an intuitive and user-friendly way, based on the JupyterLab and JupyterHub software packages. The Rucio JupyterLab software package is used to integrate the service with the ESCAPE Rucio instance. The DLaaS prototype was built on top of a generic JupyterHub infrastructure with the aim to foster further general deployments. As a result of this, several DLaaS portals are in operation serving different ESF(RI) and other deployments are in progress. Examples of the features of the DLaaS include token-based OpenID-Connect authentication, data browser, data download and upload, local storage backend access to enlarge scratch notebook space, multiple environment options, and a low latency content delivery network.

Early implementation of Storage Quality of Service (QoS) has been understood and coded in Rucio. This provided the mapping of QoS classes with individual storage endpoints (Rucio Storage Elements, RSEs), and enabled the experiment's workflows and data life cycles to target specific QoS, independently of the RSE specification. Besides the *standard* disk based full replica layouts, two additional QoS endpoints were introduced in the Data Lake: Tape Storage and Erasure Coding. These different storage QoS possibilities helped the experiments to implement data life cycle workflows, mapping data availability performance and data reliability requirements with storage costs.

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.

The integration of external Cloud resources from both commercial cloud providers (AWS, Google) and private ones (ESCAPE's surfSARA), including both storage and CPU via Swift/S3 protocol, was demonstrated during the DAC21. This exercise successfully achieved the generation of Monte Carlo simulation files for the CTA experiment and the subsequent upload to the Data Lake; it validated the usage of Commercial Cloud as an external RSE for the ESCAPE Data Lake, demonstrating the integration of heterogeneous resources, and the ability of the Data Lake to flexibly add new resources using standard interfaces.

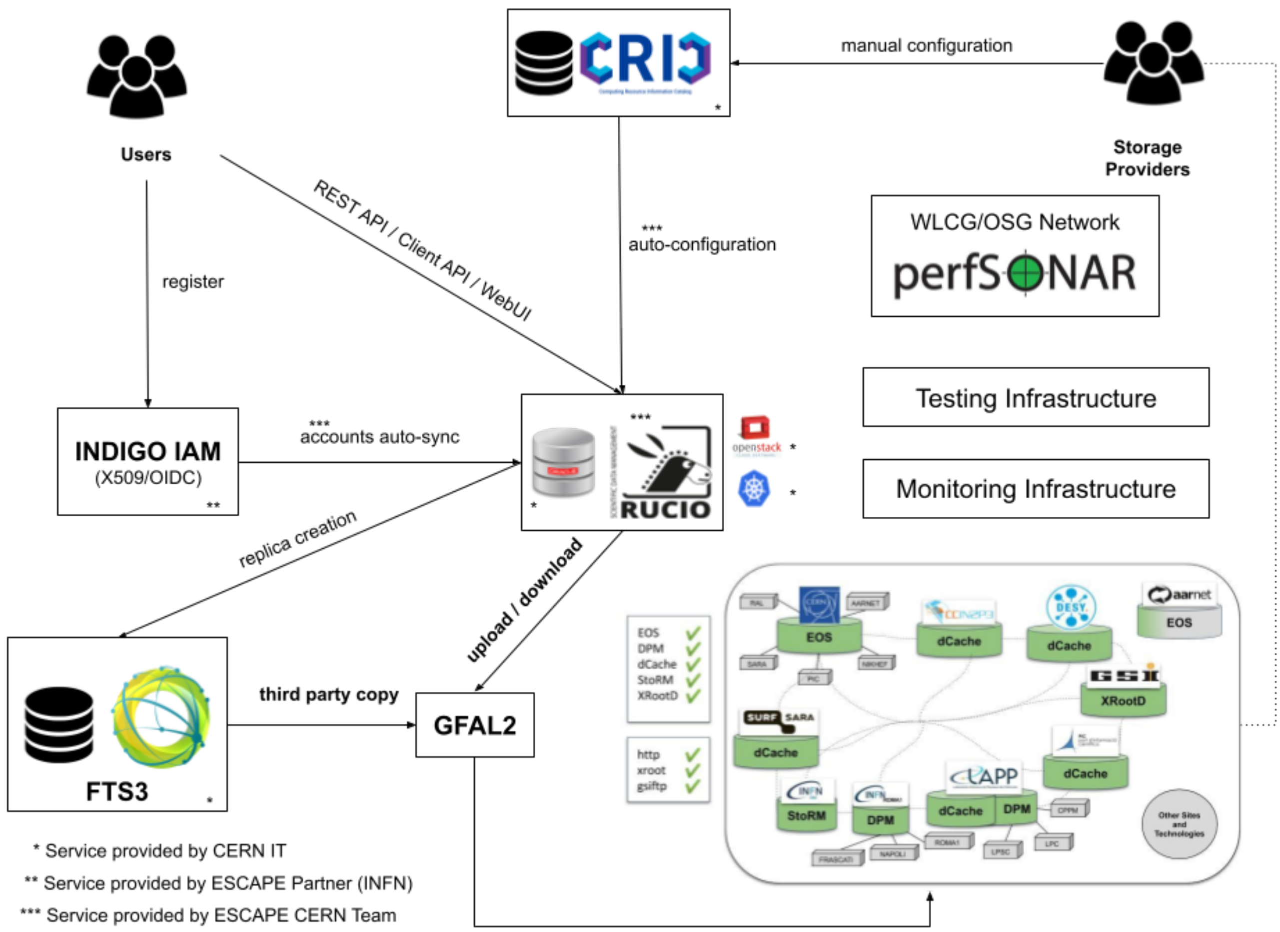


*Figure 2. The ESCAPE pilot Data Lake architecture. All components of the Data Lake and their interactions are portrayed.*

## 2. Global assessment of DIOS objectives: an executive summary

WP2 provided a fully working Data Lake infrastructure allowing the ESCAPE ESF(RI)s to put at work production-like workflows in a realistic manner during the aforementioned Data Challenges. The usability, scalability and flexibility of the Data Lake prototype has been validated. The system caters for integrating external storages: from managing storage buffers at the telescopes in remote locations, to leveraging commercial cloud storage endpoints as reported in the corresponding project milestone (MS11)[2]. The integration of these external storage sources complemented the

[2]https://projectescape.eu/deliverables-and-reports/ms11-extension-data-lake-efficiently-serve-data-external-compute-resources

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.

established storage infrastructure providing well defined global policies, replication rules and data life-cycles for the ESF(RI)s as mentioned below in the “DIOS at work” detailed reports. Data delivery and access has been demonstrated providing the required protocols and integration level, to fulfil the flexibility needed to couple with a) large compute resources: computing farms and HPC centres and b) user-facing portals, notebooks, web-based frameworks and laptops. The data access has been one of the main activities in the last phase of the project as reflected in the following chapters by the ESF(RI)s specific analysis work during the DAC21 exercise.

DIOS provided a centralised monitoring framework to track the overall status and ongoing activities of the infrastructure, providing information about a) Global activity: files being transferred, total storage capacity, realtime throughputs; b) Resources being used at the sites' storage endpoints: volume and files; c) How much data was used by the ESF(RI)s: volumes, files and the geographical location; d) Real-time data management system status: transactions, files requested, in-flight transfers, finished transfers and queues/pending; and e) Status of the network links across the Data Lake sites based on perfSONAR (performance Service-Oriented Network monitoring ARchitecture)[3].

The overarching common AAI layer deployed in ESCAPE ensured the trust among the participants and the services to make use of the infrastructure: all ESCAPE DIOS activities made use of ESCAPE specific credentials to enhance the common trust model for the ESF(RI)s, the participating people and the partner sites.

The DIOS model has been replicated at national and international level, and several ESF(RI)s deployed alternative Data Lake instances to gain first hand experience in deploying, maintaining and operating such an infrastructure.

WP2 very early identified the need to bring together the ESF(RI)s' data management experts, the site operators and the service providers to build up a common team: the Deployment and Operations team, to share expertise and handle operational issues in a centralised manner. This team met regularly and coordinated the two large Data Challenges performed. This paved the way for increased knowledge transfer and community feeling, helping with the projection of the Data Lake model to the ESF(RI)s. As an immediate result MAGIC, CTAO, SKAO and KM3NeT deployed their own services and are evaluating the usage of the DIOS tools for their needs.

The Data Lake model has been presented and introduced to other scientific communities, a common workshop was organised with the two Photon and Neutron science projects PaNOSC and ExPaNDS[4]. Similarly the DIOS model was also presented at the National German Grid Initiative PUNCH4NFDI, we liaised with other EC projects as CS3MESH4EOSC to prototype an EOSC in practice story[5] to link the Data Lake with the Sync and Share services with a scientific objective based on notebook analysis of data in DIOS. Several outreach activities were organised and included specific webinars about DIOS and technical hands-on sessions for AAI and data access in the ESCAPE Data Lake.

The sustainability of the ESCAPE DIOS is ensured in the short to medium-term, in the framework of current fellow projects: EOSC-Future, InterTwin and, in the scope of future collaborations

[3] performance Service-Oriented Network monitoring ARchitecture: https://www.perfsonar.net/index.html

[4] https://www.panosc.eu/events/pan-escape-data-management-workshop/

[5] https://eosc-portal.eu/bringing-big-science-experiment-data-researchers%E2%80%99-fingertips

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.

specifically, the ESCAPE Collaboration Agreement. These activities will rely on and leverage the current DIOS model and the underlying infrastructure, provided to the ESF(RI)s as a system to test and assess their Data Management requirements and needs. This possibility to test ideas and models in a controlled and realistic environment is considered instrumental as many of the participating ESF(RI)s are closer to the start time of their experiments.

# 3. Task-specific assessment of the DIOS objectives

## 3.1 Data lake infrastructure and federation service (Task 2.1)

### Storage Infrastructure

- **Deployment**
  The Data Lake infrastructure provided around a PetaByte (900 TB) of distributed storage. All the participating institutes collaborated to expose storage endpoints to the global infrastructure: CERN, INFN, DESY, GSI, Nikhef, RUG, SURFSara, CC-IN2P3, IFAE-PIC and LAPP. The nature of these endpoints was different and several technologies and backends were successfully integrated as building blocks of the Data Lake: dCache, DPM, EOS, Ceph, XRootD storage systems were accessible by the ESCAPE ESF(RI)s using a variety of protocols: HTTP, GridFTP, WebDAV and S3. The scale of the prototype was sufficient to demonstrate the functionality and the performance of the DIOS model. It provided a single logical view of a distributed storage federation. As a result of this the Data Lake model and its global concept several relevant requirements from the ESF(RI)s could be fulfilled, most notably allowing them to define and exercise Data Lifecycle workflows.

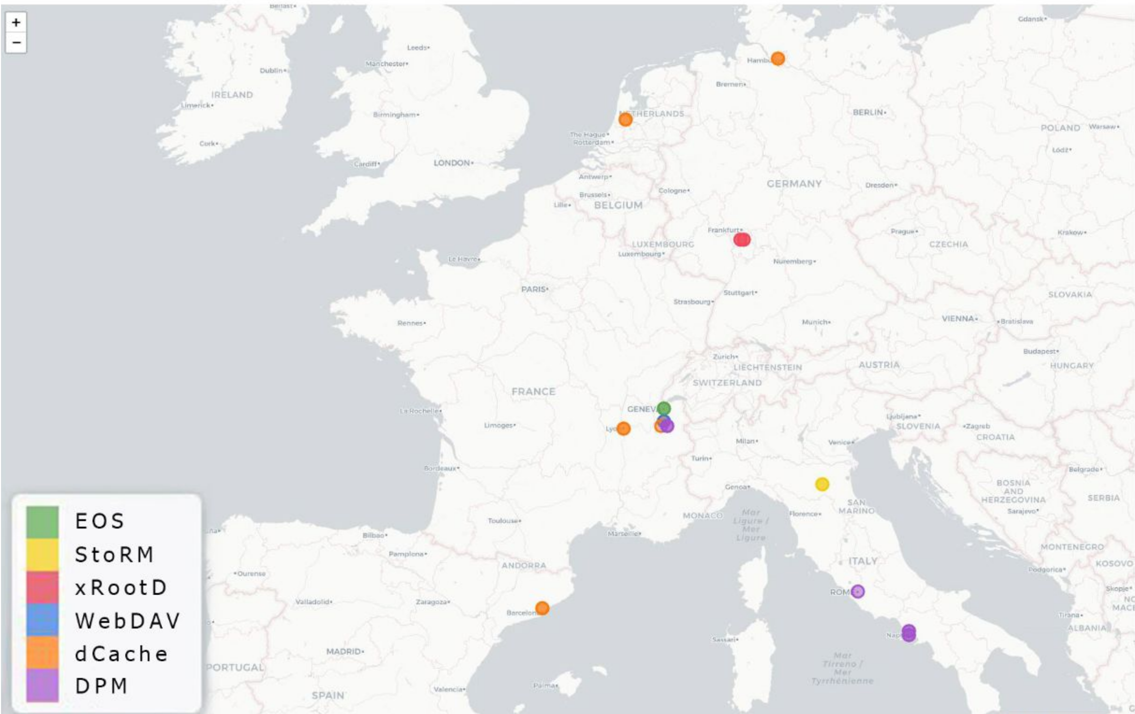


*Figure 3. RSEs that are part of the ESCAPE project identified according to the different storage technologies. This panel was temporarily used to graphically show inbound and outbound traffic to and from the different RSEs.*

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.

- **Delivery**

  Once participating sites had configured their storage to be part of the Data Lake they then committed personnel to act as contact points. Site-representatives were invited to the weekly Deployment and Operations (DepOps) meetings and made aware of issues needing to be addressed through the JIRA ticketing system. An important role for the site reps was to be able to provide production-like levels of support during the data challenges (FDR20 and DAC21) windows, and high responsiveness was achieved.

- **Forward look**
  Large similarities have been identified across the needs of the ESF(RI)s related to Data Management; from data lifecycles to the distributed computing needs down to the data access requirements. ESCAPE pioneered the integration of the several disciplines and proved a common infrastructure is feasible. This integration work and flexibility scope needed to be maintained; in fact, a prototype will be in operation after the ESCAPE project to further evolve the system and cater for ESF(RI)s' needs in some other specific aspects (i.e. data access rights, metadata capabilities)

## Data Management framework (Rucio)

- **Deployment**
  The framework providing the data-driven orchestration capability is Rucio. The operation and management of the primary Rucio service for DIOS was carried out by the CERN team. The CERN Rucio instance was deployed using industry standard technologies - e.g. Docker containers, Kubernetes orchestration technologies (Helm/Flux) and OpenStack for resource provisioning - amounting to a total of 48vCPUs, 90GB RAM and 480GB of local storage. The Data Lake reached a peak of 900 TB of storage across all storage endpoints that participated in the federation. The close collaboration between participating institutes crystallised in several ESF(RI)s operating private Rucio instances, using the same storage infrastructure.

- **Delivery**
  Rucio provided the framework to integrate the different storage endpoints provided by the sites. Some sites provided several RSEs to address specific needs, e.g. to expose different access protocols or storage QoS attributes. The system flexibility was validated in this respect: registration, publication and retirement of storage endpoints was performed and centralised through the DIOS information system. Furthermore the DIOS Data Lake model was stretched with respect to its initial perspective and projected towards a real scenario: several storage endpoints were shared across different Data Lakes (ESFRI private Rucio deployments). This is instrumental to foster a model of coherence between the resource providers (sites) and the stakeholders (ESFRIs and users). This approach has been validated during the DAC21 where three different Data Lakes (SKAO, MAGIC/CTAO and ESCAPE-CERN) operated on the common DIOS storage infrastructure.

- **Forward look**
  The ESF(RI)s should ensure the sustainability of their computing models with respect to the tools and the services which run in the centres providing resources to them. Typically the same service provider (site) provides resources to several ESF(RI)s. As we have experienced

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.

in ESCAPE, the data Data Management and Data Access needs are quite similar across different scientific domains and there is potential for converging in a common ecosystem of tools and services, which would have a beneficial impact on the operation of computing resources at the sites and the associated maintenance costs.

## Data Access and Delivery

- **Deployment**
  Besides supporting the ESF(RI)s' data management heavy-lifting, end-user data specific needs were also addressed. These included enabling data registration and data access. Interfaces were provided to the user community to access data repositories from several and varied compute resources, ranging from laptops (local I/O), web-frameworks (notebooks) or batch systems. In particular a Rucio JupyterLab extension was developed, providing direct access and browsability from notebook to the Data Lake. Enabling notebooks to access the ESCAPE Data Lake triggered the development of the Data Lake-as-a-Service (DLaaS). The focus was on further integration of the DL, data access, and the related data management capabilities with the activities ongoing in the area of Science Platforms. The DLaaS acts as an interface to the DL offering a wide range of possibilities, from very simple I/O access to more complex workflows such as enabling content delivery technologies and integration of local storage facilities at the facility hosting the notebook. DLaaS is based on the JupyterLab and JupyterHub software packages and is built on top of the JupyterHub services running on CERN OpenStack dedicated resources, and deployed on the same k8s cluster hosting the other Data Lake services and tools. Features of the DLaaS include token-based OpenID Connect authentication to ESCAPE IAM, data browser, data download and upload, local storage backend access to enlarge scratch notebook space, multiple environment options, and a low latency content delivery network with latency-hiding mechanisms based on XRootD's XCache[6].
- **Delivery**
  The DLaaS project allowed end-users to interact with the Data Lake in an easily-understandable and user-friendly way, as largely proven during the DAC21 exercise by the majority of the ESFRIs. The DLaaS is being deployed in several partner sites and evaluated as possible Analysis Platforms in support of the ESFRIs for user-analysis.
  A Proof of Concept to deliver data to HPC resources was carried out by CMS, validating the model by bridging the Data Lake with CINECA HPC centre via a Caching layer (more specific information later in the document).

- **Forward look**
  A key aspect in the current scientific computing scenario is to address the globalisation of compute resources. In the Worldwide LHC Computing Grid (WLCG) in particular, whilst Data have, for a long time, been distributed by definition, they have been hosted in quite stable locations. What is relatively new are the opportunistic data processing resources that may appear (and disappear). The data management system needs to have the right level of flexibility to be able to deliver and/or provide easy access to these: Commercial/Private clouds and High Performance Computing (HPC) centres are examples of this new model of operation bridging data and C(G)PU cycles.

[6] http://slateci.io/XCache/

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.

## Data Movement and File Transfer Service

- **Deployment**
  File Transfer Service (FTS)[7]3 is the service responsible for globally distributing the majority of the LHC data across the WLCG infrastructure to the tune of Exabytes per year[8]. FTS3 is a low-level data movement service, responsible for reliable bulk transfer of files from one site to another while allowing participating sites to control the network resource usage. Checksums and retries are provided per transfer, with active network optimization and multiprotocol support. FTS is responsible for handling all DIOS traffic among the Data Lake storage endpoints. The service is deployed in the CERN IT department infrastructure, nevertheless, an extensive integration effort took place in order to have all the functional elements optimised for the ESCAPE use cases. The FTS team participated actively in DIOS and provided guidance towards passing the knowledge on how to deploy and operate an FTS instance[9], i.e. FTS instance deployed at PIC, central FTS service for EOSC-Future (EGI).

- **Delivery**
  As the backbone for the Rucio data management backend for third party transfers, FTS handled the transfers obeying the data policy rules. During DAC21, FTS successfully transferred roughly 70TB (4.43 million files) of data across the storages in a period of 15 days with an average throughput of 1.33 Gb/sec, with secure HTTP being the prevalent protocol used in more than 99% of the cases. Similarly, the FDR20 exercise produced FTS transfer traffic of almost 9TB. In both instances, FTS delivered seamless transfer functionality with no issues, while at the same time optimising the network channels between the remote storage endpoints.

- **Forward look**
  The FTS service keeps evolving in order to anticipate the future needs of the sciences that depend on it. Part of that evolution is offering the flexibility to deliver data to heterogeneous infrastructures such as HPCs and S3/WebDAV cloud storage. Effort is also ongoing in order to refine the token-based authentication capabilities based on the profiles that the experiments are going to use. At the same time, experiments start deploying specific instances (together with the Rucio private deployments) that will allow them to better control the configuration and adjust the service for their scenarios.

## Information System

- **Deployment**
  The Computing Resource Information Catalogue (CRIC[10]) is a high-level information system which provides topology and configuration description of a distributed computing infrastructure. It aggregates information coming from various low-level information sources and complements topology description with experiments specific data structures and settings required to exploit computing resources. CRIC is the information system where

[7] File Transfer Service: https://fts3-docs.web.cern.ch/fts3-docs/
[8] LHC Data Storage: Preparing for the Challenges of Run-3, EPJ Web of Conferences 251, 02023 (2021) [link]
[9] "Why FTS? How to deploy an operate FTS" Presented at the 3rd DIOS workshop [link]
[10] CRIC, the Computing Resource Information Catalogue: https://core-cric-docs.web.cern.ch/core-cric-docs/latest/index.html

storage admins record the configuration of their storage endpoints, allowing other services to query CRIC to pick up appropriate configuration details to enable their interaction with these endpoints. The ESCAPE CRIC instance is deployed and maintained by the CERN IT department.

- **Delivery**
  The storage endpoint administrators are able to access the platform and provide the relevant storage configuration that corresponds to their storage element; this configuration is also accessible via REST API as a JSON response. Based on that API, a periodic synchronisation process takes place from CRIC to Rucio. This includes protocols, endpoint URLs, transfer priorities, transfer matrices, and other information that Rucio needs in order to consider the storage element functional.

- **Forward look**
  CRIC already serves as the central configuration catalogue for many WLCG experiments and continuous development is in progress to meet the experiment needs and specific use cases. ESCAPE CRIC also benefited from these efforts while at the same time pushing new features to the upstream development process in the form of feedback and bug reporting.

## 3.2 Data Lake Orchestration Service (Task 2.2)

### Storage Quality of Service

- **Deployment**
  For FDR20, different storage QoS labels were used, although the underlying storage technology was mostly RAID-6 disk based. Although this allowed FDR20 to validate the abstract storage QoS concept, it was felt that the homogeneity of the underlying storage rendered the demonstration less compelling.

  To address this concern, various sites already contributing to the ESCAPE WP2 testbed were asked to provide access to additional storage. The extra storage sites offered reflected a more diverse collection of storage QoS. This allowed DAC21 to take place with storage that more accurately reflected the stated QoS.

  Magnetic tape based storage capacity was provisioned thanks to efforts of CNAF, DESY, PIC and SURFsara. Magnetic tape is a commonly used storage medium within the HEP community, as it allows a site to store infrequently accessed data with low per-byte and per-day costs. However, magnetic tape has the disadvantage that stored data must first be copied back to a disk buffer before it may be accessed, a process that can take time. Simply labelling disk storage with the desired QoS (as was done with the earlier FDR20 challenge) does not reflect this behaviour.

  In addition to organising tape-based storage, disk-based storage with a non-traditional replication mode was provided with thanks due to CERN. Sites commonly use RAID (typically RAID-6) to provide redundancy and avoid data loss should a single disk fail. The storage provided by CERN is based on Ceph, which is a flexible cluster storage service.

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.

As a distributed block storage system, Ceph allows for other mechanisms to provide fault-tolerance. One popular option is Erasure Coding (EC). In some sense, this spectrum of storage options trades performance when accepting data and recovery time should a disk fail.

- **Delivery**
  The tape-backed storage was made available to Rucio as new RSEs. These new RSEs were configured so that Rucio knew that data was stored on tape, taking advantage of Rucio's existing support for tape-based storage. This additional Rucio configuration was achieved only through CRIC, which demonstrated both the flexibility of CRIC and the practicality of including tape-based storage as part of the storage technology portfolio available to the ESFRI communities.

  CERN were able to provide storage with different resilience levels. The JBOD storage provides no protection against data-loss. EC(4+2) storage is broadly similar to RAID-6, while EC(8+3) trades a reduction in the space-cost overhead of storing data with a modest increase in CPU-cost and networking code when accepting new data.

  LOFAR team's QoS requirements were also addressed: these helped to identify possible cost-saving by provisioning different storage technologies for data ingress and egress. Through the different QoS options these possibilities are available in the testbed.

- **Forward look**
  Work continues in improving the QoS support in Rucio. Experience from ESFRI communities adopting QoS concepts will identify areas where changes in Rucio would help.

  The current QoS model has a separate RSE for each Storage QoS with Rucio initiating QoS transitions by copying data. Some storage systems (e.g., EOS, dCache) support storing data under different QoS classes. Such systems could allow internal QoS transitions, which would allow the storage system to optimise the QoS transition (e.g., avoiding unnecessary data transfers). Work on improving support and standardising interfaces will continue.

## Data Transfer Optimisation

- **Deployment**
  The DIOS team benefits from a close connection with both the development teams of the various underlying technologies and the sites that are provisioning services used within the testbed. This close connection allows for rapid diagnosis of problems and (when feasible) quick development and deployment of "hot" fixes to address problems.

- **Delivery**
  During the DAC21, an LSST use-case initially suffered from poor performance and a large number of failed transfers. Through this close collaboration, the underlying cause of these problems was quickly identified: LSST files are several orders-of-magnitude smaller: O(kB) than the files typically used by WLCG VOs: O(GB). The production-like loads of DAC21 resulted in components being stressed in ways not previously observed; for example, transferring LSST files was often much faster, some taking less than a second to complete.

Such a fast turn-around of transfers placed new strains on storage technologies and the data transfer service.

With the help of the corresponding development teams, bottlenecks and race conditions were identified. Several patches for dCache were quickly developed that resolved several underlying problems by the dCache team and deployed by IN2P3.

The investigation also identified a bottleneck in FTS, in how it schedules new transfers. FTS queries its database periodically, looking for transfers to start. Although the interval between queries is configurable, the impact of the underlying database query places a limit on how often this can happen. When transferring many small files, FTS can struggle to maintain the desired concurrency.

The FTS service that ESCAPE testbed uses was configured to query the database at almost the fastest possible rate, which mitigated the problem. Together with the fixes to dCache, this allowed LSST to achieve a successful demonstration of their use-case of transferring a “night's worth” of data within a few hours (see below).

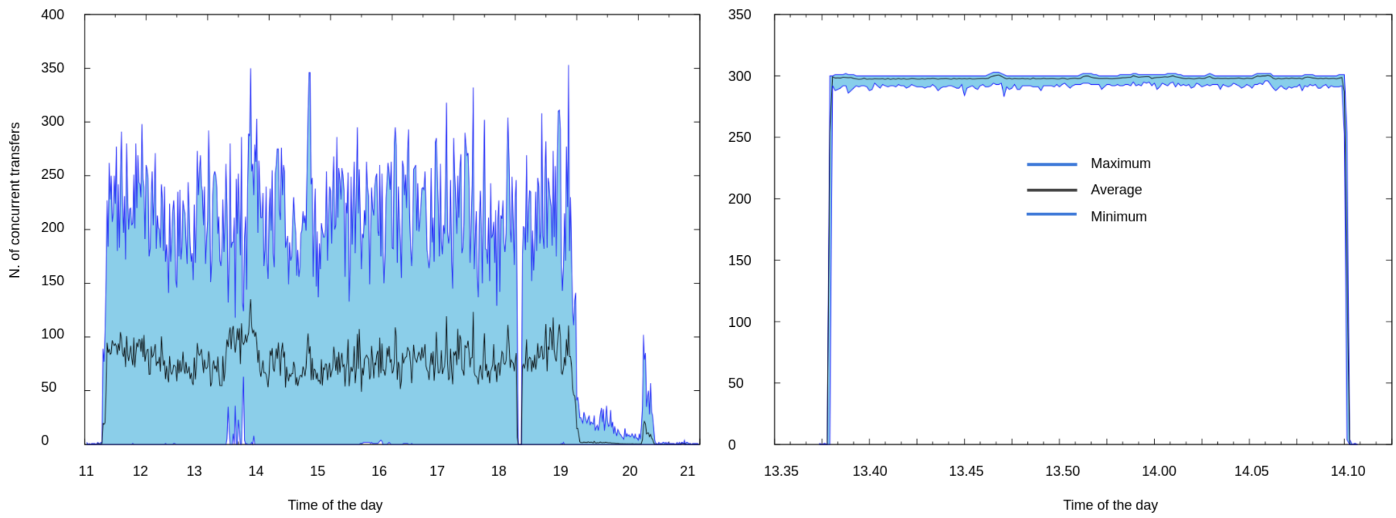


*Figure 4. Improvement of available dCache concurrency. Plot shows the number of concurrent transfers the Data Lake infrastructure was able to sustain. Left: initial observed behaviour during DAC21 (over a 10-hr period) - with an average of around 75 concurrent transfers. Right: service achieving the configured maximum transfer concurrency of 300 (over a 30-minute period).*

- **Forward look**

  Although the changes developed within DAC21 for dCache have already been taken upstream and included in regular dCache releases, we anticipate further improvements may be needed to improve support for LSST use-cases.

  The bottlenecks discovered in FTS have been discussed with the FTS development team and accepted as feedback for future development work. It is anticipated that these limitations will be resolved in the next major version of FTS.

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.

## 3.3 Integration with Compute Services (Task 2.3)

- **Deployment**
  The integration of the Scientific Data Lake with data processing services is related to many activities that took place during the project. As briefly pointed out before in Task 2.1, the first noteworthy activity has been the development of the DLaaS. The DLaaS makes it possible for users who have little knowledge of the underlying Scientific Data Lake system to access data on processing systems that are located near an RSE. Local scientific communities might benefit from deployment of a DLaaS instance together with an RSE nearby, making use of content delivery networks or instrumenting data caching mechanisms. The DLaaS kernel is fully containerised, making it extensible with domain-specific processing software so it can be tailored to the needs of specific scientific communities and their experimental instruments and facilities: accelerators, antennas, and telescopes, which have different needs in software, metadata and control-data (calibrations, alignments etc).

  The DLaaS focus was on further integration of the Data Lake, data access, and the related data management capabilities with the activities ongoing in the area of Science Platforms. The DLaaS acts as an interface to the Data Lake offering a wide range of possibilities, from very simple I/O access to more complex workflows such as enabling content delivery technologies and integration of local storage facilities at the facility hosting the notebook. The fact that the DLaaS is containerised and based on JupyterHub, which is a very common application used in the ESCAPE sciences, means that the DLaaS system can be deployed relatively easily at data centres near RSEs.

- **Delivery**
  Batch-type processing, CTA (use case 2) assessed the integration between the DIRAC workload management system, and Rucio. There is an ongoing effort with interest from both the Rucio and the DIRAC community to integrate both further.

  Another important milestone that has been reached is the fact that the metadata of the ESCAPE Data Lake may be queried from WP5's ESAP. Integration between ESAP and DIOS makes the data in the Data Lake findable for both interactive (e.g. through a DLaaS instance launched from ESAP, or as part of a broader ESAP query for different data source) and batch (e.g. through DIRAC) processing jobs. Finally, the activation of OIDC-token Authentication and Authorisation on both Rucio as well as some of the storage instances inside the Data Lake makes it possible for users to authenticate to the system and access the data by using the trust-relations between academic institutions and let users directly use their institutional credentials instead of having to request X.509 certificates for this purpose. Especially in fields where data is to be accessed by relatively inexperienced users, like is for instance the case for public data in astronomy, having to request an X.509 certificate causes significant overhead

- **Forward look**
  New paradigms to perform analysis are consolidating, the notebooks are addressing a need of the community to simplify access to the computing environments for the scientists, and also for citizen-science and outreach activities. The work initiated in ESCAPE DIOS with the notebook based analysis is being followed-up in the framework of EC-funded activities (see

later EOSC-Future's Virtual Research Environment). These are the first steps towards future Analysis Platforms/Analysis Facilities the ESFRIs might be provisioning to face the user-analysis challenge. One of the key aspects to address in these analysis frameworks is the ability to integrate on one side the *scientific big-data*, the Data Lake, and on the other side the large computing resources. With the goal to offer to the user community and to the ESFRIs data processing managers the ability to scale-out analysis jobs submission: from the notebook running on modest-CPU systems to large batch clusters, clouds and HPCs.

## 3.4 Networking, monitoring and continuous testing (T2.4)

### Long Haul Networking

- **Deployment**
  The ESCAPE project has benefited from an excellent set of highly professional storage providers, who as a community had a good understanding of the WLCG-related technologies in use for the establishment of the ESCAPE Data Lake. Early in the project, sites offering storage into the Data Lake were asked to ensure they had the ability to test network connectivity, by deploying PerfSONAR boxes close to the storage. These boxes run periodic network health tests, its results are aggregated in the perfSONAR global dashboard[11] and funnelled into the ESCAPE monitoring platform.

  In addition to the sites in the main ESCAPE Data Lake instance, SKAO and MAGIC/CTAO each deployed an additional Rucio instance. They made use of some ESCAPE testbed storage services, but also added some storage endpoints of their own - this extended the used network links to the Canary Islands (MAGIC) and South Africa, Australia and the UK (for SKAO).

- **Delivery**
  Since the sites involved in ESCAPE were already quite experienced in providing resources into distributed systems (e.g. through WLCG involvement), there were very few difficulties with the physical network layer - the links themselves were very reliable.

  One parallel activity that contributed greatly in understanding, testing and analysing the network usage was the collaboration that ESCAPE had with the WLCG DOMA project, and more specifically, the active part ESCAPE members played in the DOMA Network Data Challenges[12] exercise that was performed in October 2021. The aim of that exercise was to generate enough traffic among WLCG storage endpoints, in order to achieve a set throughput goal, while at the same time observing the network channels' performance and efficacy. The ESCAPE community provided the expertise and the effort to generate that traffic and build a uniform monitoring framework in order to track and explore the exercise as it was happening. New features were developed that enhanced the analysing/exploring capabilities of the monitoring platform, these features were then also integrated into the ESCAPE specific views.

[11] http://maddash.aglt2.org/maddash-webui/index.cgi?dashboard=ESCAPE%20Mesh%20Config
[12] https://zenodo.org/record/5767913

- **Forward Look**
  The adoption or continued prototyping of aspects of the DIOS stack by major astronomy observatories (CTAO has adopted Rucio and SKAO continues to test functionality in a growing global prototype) will necessitate long-haul data movement test at scale and with some sites, with limited storage capacity, acting as data sources. For example, SKAO will generate an average of up to 1 PByte per day for each of the two telescope sites, but the system sizing (and cost) of the data staging buffer for this will depend not only on the bursty profile of data delivery (which can vary greatly between different astronomy projects) but also on the reliability of the data transfer out of these sites - lower reliability having an immediate knock-on effects of requiring a larger storage buffer and (through reduced average transfer speed) compromising on scientific productivity.

## Continuous Testing and Monitoring

- **Deployment**
  Throughout the ESCAPE project automatic monitoring tests and related dashboards were developed and improved. The main dashboards integrated logs coming from FTS, Rucio, and other sources, including the deployed perfSONAR boxes. These have a set of panels that offer a better understanding of the state of the system, having the possibility to make use of specific filtering to visualise the data relevant to each use case.

- **Delivery**
  Continuous transfer tests have been implemented to allow easy monitoring, linked to a customised Grafana dashboard. These tests span the different layers: direct PerfSONAR tests on the links, storage-to-storage tests with Gfal, FTS-based tests and finally Rucio-level tests triggered by Rucio events. These were implemented in an iterative fashion, with the testing suite developing over time as the communities using the results could see where additional functionality was needed. For example, in order to allow throughput measurements during specific challenges, such as DAC21, these tests were augmented using some ranges of file size for transfers. These dashboards allow access to historical data so as well as showing the current health status of the ESCAPE Data Lake and its components, they are a powerful diagnostic tool when facing failures or spotting repeated failure patterns.

  The main dashboard views are the “Rucio Events”, the “Rucio Stats” and the “FTS Transfers” views.

  The Rucio dashboards give details about events triggered by Rucio (e.g. the queuing of transfers, submission and subsequent completion or failure of a file transfer request, and file deletion activities). Furthermore, they display the overall statistics for data products stored within the Rucio managed federation, reflecting the data according to the physical and logical storages, in relation to the scopes, RSEs, and the different experiments. Statistical data is also displayed according to the protocols or QoS, or any kind of event that happens in Rucio for any of the above.

  The “FTS Transfers” view gives similar information but based on the FTS activity only. It can also show the statistical data according to the RSEs, both origin and destination, or data specified according to the VO or the activity label. This dashboard gives access to links that

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.

point to verbose FTS log files per transfer, allowing this way a very detailed inspection of individual transfer cases.

These dashboards have proven to be powerful tools giving insight not only into the health of the Data Lake but also into the performance of the stack when running customised tests (e.g. in the FDR20 and in DAC21). They allow us to "watch" test events unfolding as patterns and see the impact of failures/fixes as we see the Rucio service working its way through queued up jobs.

The latest updates regarding monitoring were the adjustment of some panels that did not reflect the data correctly, the implementation of new panels with more detail regarding FTS transfers or some Rucio events, and the creation of panels that showed the Rucio statistics in an interconnected way for a greater interaction by the user.

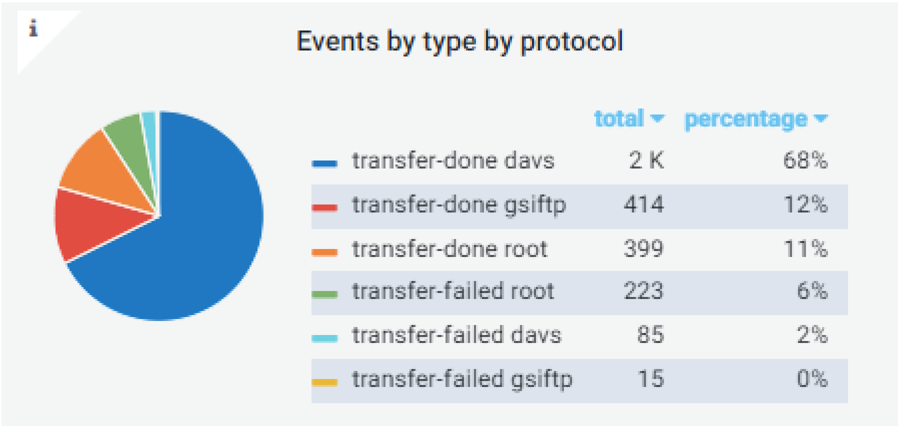


*Figure 5. Example dashboard panel reflecting the Rucio Events by type and by protocol with no filtering.*

- **Forward Look**
  In case there is special interest in adopting monitoring panels similar to the ones described in this section in the future by any system that uses Rucio, the following information should be taken into account. In the case of Rucio Events the Hermes Daemon has been configured in the cluster for the delivering of messages. In addition, the source code of the scripts executed for the Rucio Stats dashboards can be found in the project's GitHub[13]. These dashboards are using the library rucio.client and direct queries to the DB in order to obtain the data used by the different panels.

  Throughout the development of the project, best practices regarding monitoring and continuous testing have been identified. These are the result of feedback from users as well as knowledge sharing between monitoring administrators of the different collaborations. For this reason, they can serve as a basis for other projects using similar infrastructures or services. Examples of this are the EOSC-Future project, which is going to use the monitoring - among many other features - for the preparation of its new cluster and infrastructure, and the dashboard suites adopted by MAGIC/CTAO and SKAO.

[13] https://github.com/ESCAPE-WP2

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.

# 3.5 ESCAPE Authentication and Authorization Infrastructure (Task 2.5 )

## ESCAPE AAI Infrastructure

- **Deployment**
  INDIGO IAM is an authentication and authorization service, initially developed in the context of the H2020 INDIGO DataCloud project and later extended to support the needs of the HEP (High Energy Physics) community. It issues credentials in the form of JSON Web Tokens (JWT) using standard OAuth2/OpenID Connect flows. INDIGO IAM can include a VOMS attribute authority micro-service which issues VOMS credentials compatible with existing WLCG resources, providing a backward-compatible VOMS support. In case a VO already has its legacy VOMS server, a VOMS importer script, currently developed for CERN use cases, can be used to synchronise users, groups and roles between a VOMS server and IAM. INDIGO IAM was selected by the WLCG management board to enable a smooth transition from X.509 to token-based authentication and authorization.

  The ESCAPE INDIGO IAM instance ("ESCAPE IAM" from now on) is the central authentication and authorization service used in the ESCAPE Data Lake, deployed on a Kubernetes cluster at CNAF, which:
  - supports authentication through EduGAIN federation, Google, personal X.509 certificate and local credentials;
  - implements user registration and management;
  - issues JWT bearer tokens following WLCG Common JWT Profiles[14] as a default and AARC guidelines on demand (may be configured per-client);
  - issues VOMS credentials for the escape VO.

- **Delivery**
  The ESCAPE IAM has currently approximately 160 registered users organised in 13 groups (some of them dedicated to the ESF(RI)s) and 4 roles. These users belong to several experiments: ATLAS, CMB/PANDA, CMS, CTA, EGO/VIRGO, FAIR, KM3NET, LOFAR, LSST, MAGIC, R3B, SKAO. The ESCAPE INDIGO IAM instance has been configured to support an admin-moderated enrollment for any incoming user registration request (Figure 6). An automatic process of enrollment, when a new user authenticates itself through a trusted, configurable, Identity Provider, is also supported.

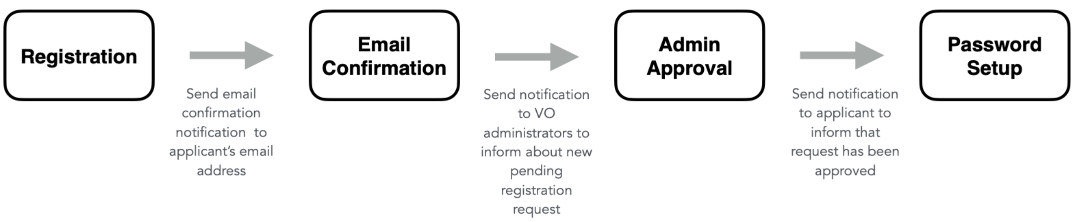


*Figure 6. User enrollment flow in the ESCAPE IAM.*

  The ESCAPE IAM is the first instance with the need to handle multi-VOs. This challenge is temporary solved within the project using a common set of rules, in particular:
  - A VO is expressed in IAM as a group (a subgroup of the *escape* one);

[14] https://doi.org/10.5281/zenodo.3460258

- IAM users have to join to one or more groups, according with the VO they belong to;
- for each VO there is a need to indicate 1-2 reference persons, IAM administrator(s), who manage registration/group requests from users of their VO;
- A common Acceptable Use Policy (AUP[15]) for all VOs (the ESCAPE IAM one).

The group membership information is propagated in the user's credentials (proxy and token) and it is used to apply local authorization policies on the ESCAPE resources. Controlled delegation of privileges among users and services is natively supported by the token exchange[16] OAuth2 specification, which IAM partially implements. The ESCAPE IAM has been successfully integrated and used to login into the ESCAPE web services (Rucio, Jupyter notebooks, etc.) thanks to the use of common OAuth2/OpenID-Connect libraries. Periodic synchronisation routines make sure all new users and their credentials (e.g. X.509 certificates) are getting mapped to their respective Rucio accounts with minimum delay, ensuring this way almost instant access to the Data Lake resources after user approval. During DAC21, the IAM service was running continuously, proving to be capable of serving the ESF(RI)s requirements, enabling the usage of the data management and data access services in the DIOS infrastructure and sustaining the operations load.

At the time of writing, the ESCAPE IAM has been integrated into the EOSC AAI federation in collaboration with GÉANT, in order to allow ESCAPE users to access EOSC resources. This integration can be fully effective as soon as a better compliance with AARC will be satisfied.

- **Forward look:**
  The ESCAPE IAM infrastructure will continue to be supported within the scope of the ESCAPE Collaboration Agreement (CA), providing the required AAI infrastructure to the ESFRIs communities in the ESCAPE CA and associated EC-funded projects and initiatives. The INDIGO IAM release v1.8.1 planned for September 2022 meets the vision of the ESCAPE AAI integration into the EOSC project, in particular foresees more support to the AARC guidelines [17]. The evolution of OAuth2/OpenID Connect standards and all the needs raised by stakeholders will also be tracked. Further in the future, the idea is to support even OpenID Connect federations, which are less than a proof-of-concept at the moment. Moreover, the token exchange implementation in IAM is planned to be more compliant with the specification, such to fully support a controlled delegation during a Rucio download/upload operation. INDIGO IAM software is developed in an open collaboration on GitHub[18].

  In another dimension, the ESCAPE effort of consolidating an authentication and authorization infrastructure goes hand in hand with other WLCG initiatives to drive the HEP community towards a stable token based authentication enabled stack of software and tools. ESCAPE contributions, both on the software level but also on the experience obtained from early adoption of new technologies, are interconnected with future endeavours that WLCG will undertake in order to have industry standard authentication and authorisation mechanisms that serve the scientific communities.

---

[15] https://baltig.infn.it/aceccant/edugain/-/wikis/escape-iam.cloud.cnaf.infn.it-privacy-policy-and-description-en
[16] https://datatracker.ietf.org/doc/html/rfc8693
[17] https://aarc-project.eu/guidelines/
[18] https://github.com/indigo-iam/iam

## DLaaS Authorization policies

- **Deployment**
  Two orthogonal layers of authorization into a federated infrastructure can be used.
    - *identity-based,* where the credentials bring information about attribute ownership (e.g., groups/role membership) and the resources maps these attributes to a local authorization policy
    - *scope-based,* where the credentials bring information about which actions are authorised at a service and the service must honour them. The authorization policy is managed at the VO level

  INDIGO IAM is capable of issuing credentials which support both authorization mechanisms. Multiple groups can be created and the subgroups allow to provide a finer grain of authorization. INDIGO IAM is also able to define more privileged roles, which have to be explicitly requested by the users. The groups/roles membership is encoded as wlcg.group claim in the JWT token, following the requirements agreed within the WLCG community[19]. In order to provide backward compatibility, groups/roles membership are shown as a VOMS attribute in the proxy too. The scope-based access is supported in the JWT token credentials. Definition of particular scopes (storage.*, compute.*) agreed within the WLCG community is also implemented in INDIGO IAM.

- **Delivery**
  The authorization to the DLaaS is determined by people's membership in groups within ESCAPE IAM. Everyone within the ESCAPE project is a member of the *escape* group; local access policies applied to the DLaaS give read-write permission to this group. This scheme is currently focused on supporting open-access data. A test suite[20] which runs periodically in order to verify the proper issuance of JWT tokens and the enforcement of the above authorization policies in ESCAPE storage infrastructure has been developed. As a proof of concept, an *escape/data-manager* role, which consists of a restricted group of expert users with elevated privileges, has been created in ESCAPE IAM. Members of this group have superuser-like access and can make any changes to the ESCAPE Data Lake, such to investigate and potentially resolve VO-specific problems without being a member of that organisation. The authorization based on this group is, at the moment, honoured by few endpoints as a test. The need of having embargoed data for each experiment, for which read-write access is allowed only to VO members, was raised.

- **Forward look**
  The identity-based authorization at the ESCAPE storage infrastructure can become more fine grained by filtering data from different VOs. For instance, by using a different folder for each experiment data, rooted under the ESCAPE root namespace directory, where the authorization is the following:
    - ESCAPE VO members (belonging to the *escape* group) can read any data in the `/escape` namespace, also in experiment sub-folders;

[19] https://doi.org/10.5281/zenodo.3460258
[20] https://github.com/indigo-iam/escape-auth-tests

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.

- write access is granted in an experiment folder to members of the experiment group in ESCAPE IAM (*e.g.* members of *escape/cms* group can write in the `/escape/cms` part of the namespace);
- each experiment branch has an `embargoed` folder where read and write access is only granted to members of the experiment;
- A `data-manager` group grants full access privileges to the whole ESCAPE namespace.

Moreover, an orthogonal scope-based authorization following the use cases of the WLCG community can be applied to the Data Lake (e.g. when a secure service has to automatically renew the credentials without intervention of a user). Future INDIGO IAM releases will include the scope policy management interface, which aims to give/deny access to specific scopes only to certain IAM groups or users, allowing for a more controlled authorization to the Data Lake.

The current authorization to the ESCAPE resources is based on local authorization policies which have to be implemented independently per service. The expression and enforcement of non-local authorization policies is a more difficult subject and will probably be best implemented as a collaboration between INDIGO IAM and domain-specific services, such as Rucio for what concerns data management. Work is underway to implement a workflow specific to file upload and download, leveraging the OAuth2 token-exchange flow between INDIGO IAM and Rucio Server.

# 4. DIOS at Work: ESF(RI)s assessment from Data Challenges

The purpose of the ESCAPE DIOS work was to allow new communities to explore the Data Lake technologies and run tests to see in what ways the DIOS examples can meet their needs. In this section we summarise the context, needs and tests coming from each of our infrastructure communities. Early testing was undertaken in 2020 as part of the pilot data lake assessment (FDR20) and this in most cases led to the development of more tailored or sophisticated tests during the 2021 (DAC21) exercise.

## 4.1 CTA

### Description

The Cherenkov Telescope Array Observatory (CTAO) will be the largest and most sensitive gamma-ray observatory in the world. In order to cover the whole sky CTAO will have sites in both the Northern and Southern hemispheres. The southern array site will be located at the European Southern Observatory site at Paranal (Chile). The northern array site is at the Roque de los Muchachos Observatory on the island of La Palma (Spain). The telescopes will operate nightly, each site producing 3 PB of raw data per year. The raw data will be transferred from the telescope array sites to an off-site archive, distributed on four European data centres, where the data is preserved long-term and processed to higher data levels. Differing data types and categories will have differing priorities for transfer.

### Requirements

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.

CTAO is planned to have seven data levels (DL0 to DL6), two of which are particularly relevant to this work: DL0 corresponds to “raw” data and DL3 to the first level given to science users, called “science ready”. DL0 will be ingested into the Bulk Data Management System (BDMS) and transferred to one of the off-site data centres. There it will be processed to higher data levels, which will be stored on disk, while the DL0 data will be stored on tape. The DL0 data will be replicated to a second data centre where it will also be stored on tape. During the next transfer period, typically the following day, the plan is to initially transfer the DL0 data to the second data centre and then replicate it to the first. This load sharing will help equalise the use of storage and computing resources across data centres.

The Bulk Data Management System (BDMS) is a subsystem of the Data Processing and Preservation System (DPPS) within CTAO. The DPPS is a suite of software systems responsible for processing, monitoring, storing and verifying the science-ready (DL3) data products. It will be responsible for storing all data products for the lifetime of the experiment and ensuring their quality. DPPS will also generate simulated data and periodically re-process raw data. In this way it is the most applicable part of CTAO when considering ESCAPE’s DIOS work package.

Selected applicable DPPS requirements include:

- The DPPS shall be capable of reprocessing the equivalent of 1 year of DL0 data products from both CTAO array sites within one month.
- The DPPS shall be able to perform a full reprocessing of all preserved data products, up to DL3, annually.
- The DPPS shall ensure that, after data transfer to off-site, at least two copies of all data products are maintained on independent storage systems at physically distinct locations separated by at least 300 km, to minimise the risk of data loss
- The DPPS must check the integrity of Data Products prior to and following data transfer off-site and between data centres.
- The DPPS must support and provide unique identifiers for all data products that are independent of the storage location or number of copies.
- The DPPS must support the automatic removal of files at the origin after their successful transfer.
- The DPPS must support the use of tape storage for long term storage of raw data.

## Assessment: CTA Use Cases, Tests and Results

Based on the experimental setup described above and the selected CTAO requirements, the following three use cases were developed for testing.

- **Use case 1: Long-haul transfers**
  *“Ingestion of CTA data from a remote site (RSE at the CTA Array Site, ‘on-site’) to the data lake, transfer and replication in off-site RSEs and after replication deletion of the data at origin”*

  Data was firstly ingested into an RSE on La Palma, simulating what would happen during operations. It was automatically transferred to the Port d’Informació Científica (PIC) data centre, to then be further transferred to another RSE and copied onto tape at both locations.

The on-site data was deleted once the transfers were secured and validated, and it was finally discoverable and available.

A set of simulated raw CTAO data was used, and a CTA specific Rucio instance was installed with RSEs at PIC and in La Palma. The PIC team was in charge of developing the CTA Rucio instance as well as for the MAGIC experiment. The interested reader is invited to read the MAGIC section for more details on the setup.

The use case was split into several tests, each built upon the previous. The amount of data transferred during each test increased, while each file size corresponded to a constant value of around 1.5GB. Two protocols were tested, xroot and GridFTP.

The results and details of each test are summarised in Table 1. Five of the seven tests were successful. Test 4's aim was an ambitious 40 TB transfer, which was not achieved within the allotted time slot. Test 6, which was aimed at replicating the data on the CNAF RSE, also failed due to certificate permissions.

| Test name | Protocol | Estimated #Files, data volume | STORAGE | Replicas/RSEs | Results | Observations |
|---|---|---|---|---|---|---|
| Test 1 | GSIFTP | 300 GB | Disk | CTA-RUCIO: non-deterministic and deterministic RSEs | Completed successfully. | |
| Test 2 | XROOTD | 300 GB | Disk | | Completed successfully. | |
| Test 2.1 | XROOTD | 10 TB | Disk | | Completed successfully. | Initially not planned, is a mix of Tests 2 and 3 |
| Test 3 | GSIFTP | 10 TB | Disk | | Completed successfully. | |
| Test 4 | GSIFTP | 40 TB | Disk | | Failed. Completed in more time than stated | Problems with the Rucio server side transferring large number of files. |
| Test 5 | GSIFTP | 10 TB | Tape | | Completed successfully. | |
| Tes 6 | GSIFTP | 10 TB | Disk | | Failed. Completed internally. | We couldn't run the test on production environment since we are experiencing problems to connect from PIC to CNAF due to an update of Grid certificates in both sides, changes in the CA |

*Table 1. Tests conducted for CTA Use Case 1m including setup and results*

- **Use case 2: Data Reprocessing**
  *"The ability to reprocess all raw data (DL0) to higher (DL3) level"*

  As CTAO's raw data will be stored on tape, the initial step of this use case is data identification. The data would then be staged on disk, in order to be processed from raw (DL0) to science-ready (DL3) data. The processing is done via the DIRAC Workload Management System (WMS), and the resulting data products ingested into the Data Lake. Whilst CTAO has not yet settled on a solution for the Bulk Data Management System, it was decided that DIRAC would form the WMS framework for scheduling and executing batch processing, which were fulfilled thanks to the help of the chief CTA DIRAC developer and the use of the CTA-DIRAC test cluster.

  The use case aimed to take advantage of the Rucio plugin developed by Belle II, which has been already successfully deployed. The plugin was designed to enable users to work with both DIRAC and Rucio. Its adoption was, however, non-trivial due to some setup and workflow specifics in the Belle II workflow, and several DIRAC functions needed to be added to the Rucio-Belle II plugin.

Data was successfully ingested into the Data Lake using DIRAC, and later accessed by processes launched via DIRAC. However, it was not possible to ingest data using Rucio directly and then access it via DIRAC. Rucio's upload API did not create the hierarchy as expected by DIRAC. Despite these issues, the CTA analysis scripts were successfully launched and data processed from raw to the first data level. At the time of testing the full pipeline was not available, however generating DL1 data is an excellent proof of concept for the tested workflow.

In order to fully test this use case, a further CTA-Rucio instance was created which is connected to the tape storage. The data was then staged/replicated to the original CTA-Rucio RSE using the Rucio API.

- **Use case 3: Data Analysis**
  *"Analysis of science data to higher level data via an analysis platform."*

  The use case's aim was to enable a user to find data from the Data Lake, select software from the Open-source scientific Software and Service Repository (OSSR), and deploy an analysis on an interactive analysis platform. Several steps were required to initiate this use case, which tested two different analysis platforms: the ESCAPE Science Analysis Platform (ESAP) and the DLaaS.

  For the former, science-ready data was uploaded to the ESCAPE instance of the Data Lake using the Docker image of the Rucio client. As the gammapy analysis package had not yet been onboarded to the OSSR, a hard linked version was added to the development version of the ESAP. From the ESAP it was then possible to find the CTA data files and select the gammapy software. A MyBinder instance already containing the environment and a reference to the selected data could then be launched from there. However, it was not possible to download the data, as the Rucio REST API did not have a download function.

  The second method attempted was to use the DLaaS. From the Rucio Jupyterlab extension it was very easy to locate and replicate the data to the local system. The gammapy software however was required to be installed manually.

  Both methods were able to produce the required higher-level data products with slight work arounds. Work on resolving these issues will continue within ESAP and the Virtual Research Environment (VRE) group within the EOSC-Future project.

**Assessment summary**

The ability to automatically transfer on the order of 10 TB of data from one of the future array sites using such a prototype was an extremely important data point for CTAO. The addition of tape end-point storage was also a requirement and an important demonstration made during these tests. Integration of monitoring tools such as Grafana was also key.

The work undertaken showed the importance for CTAO of employing a bulk storage solution able to interface with DIRAC. Work is still ongoing in order to fully-support the second use case: CTAO would be interested in collaborating with other ESF(RI)s with similar requirements in order to finish this work.

The DLaaS was an important and interesting development. A combination of the two methods outlined - i.e. being able to bring both data and software to the analysis system - would represent the ideal system. An important but untested component was also the ability to upload the results back to the data lake.

Use case 1 confirmed that deletion of the file at source is possible along with a replication to a second RSE. As outlined, we are continuing the work begun to unite the CTA-DIRAC and Rucio instances, including testing, defining the scope of work needed and missing methods. CTAO is also interested in how Rucio and DIRAC will work together using tokens, post X.509 authentication.

Future investigations CTAO wishes to carry out include work on metadata, further longer-haul transfers (from Japan), the ability to prioritise transfers of certain data types (different data levels imply a range of potential file sizes), fine-grained authorization, the use of tokens and testing the transfer of files of differing size. As CTAO will operate as an observatory, storage points will need to be able to handle embargoed data, which is another objective of future testing. Last but not least, a self-contained DIOS as a docker container to ease the setup of an RSE or instance would be of interest. CTAO will investigate the adoption of certain ESCAPE technologies for its next Science Data Challenge.

## 4.2 SKAO

### Description

The Square Kilometre Array Observatory is a 2-billion euro project to build and run the world's largest Radio Observatory - using two separate interferometric telescope arrays, one in South Africa and one in Australia, with the HQ in the UK. SKAO's telescopes will enable a vast range of radio-astronomy science projects to be undertaken, with studies encompassing the early (and distant) Universe (when stars first started to burn) through to our own galaxy. The technological advances that unlock new scientific capabilities with SKA also lead to vast increases in output data rate and the size of data products - SKA will be able to produce maps of the radio sky at such high spatial and frequency resolution that these can be 1 PB in size each. Whilst this will be exceptional rather than typical, an average data rate of about 2 PB per day is estimated to be outputted; this will need to be replicated around the globe to ensure it can be accessed by users.

### Requirements

SKA Observatory's fundamental use case related to WP2 is the distribution and tracking of data products from the Observatory sites into and between externally-managed SKA Regional Centres (SRCs). These SRCs will provide science users with tools and capabilities to analyse data products that they choose, but must be able to respect the data access policies. Staff at the Observatory and in the global network of regional centres need the ability to see which sites have replicas of data products, and to control replica placement.
SKAO has identified four key user stories to derive appropriate WP2 tests.

- **User Story 1**
  *As a SKA staff member I want to be able to manage data storage use, supporting exabyte-scale data collections across globally distributed locations and with heterogeneous hardware and middleware configurations so that I can make use of externally-managed resources.*

- **User Story 2**
  *As a SKA staff member I want to be able to pre-define replication rules for data products from particular telescope projects so that I can understand resource implications at SRCs for the anticipated data product lifecycle and plan accordingly, with methods that accommodate flexible observing[21].*

- **User Story 3**
  *As a science user I want to interactively analyse data products managed in the Data Lake*

- **User Story 4**
  *As a science user I want to be able to access SKA's data products and appropriate SRC resources according to the collaborations I am involved in, with minimal administrative overhead when verifying my identity, and using technologies (like OIDC) that I am already familiar with.*

As a result of the successful tests undertaken by the SKA team during the FDR20 tests in 2020 (reported in D2.2), and using the support from experienced colleagues in WP2, SKAO deployed (and continues to maintain) its own Rucio instance. Several storage end points supporting the SKAO VO are connected to the instance. These include both sites in Europe (UK sites and WP2 site at DESY), and sites located in South Africa and Australia, close to the SKA telescope. This global reach allows to simulate more representative long-haul transfers.

In setting up the SKAO prototype Rucio instance, resources and experience from the ESCAPE Data Lake were used as a model: CERN's pilot FTS instance, functional testing and Grafana dashboards for analysing the Data Lake health. The Rucio developer community was very helpful in providing guidance along the way. Two instances were deployed (development and production instances), along with documentation being written along the way[22]. There have been improvements to the documentation and installation process as a result of new non-HEP communities deploying Rucio.

Continuous automated testing was running across the instance and was used for most DAC21 dedicated testing. The functionality provided by the "rucio-analysis" testing suite developed by the SKA team was expanded to enable the DAC21 tests[23].

[21] i.e. I will not know when a particular (weather or resource dependent) project will have new data products released.
[22] https://gitlab.com/ska-telescope/src/ska-rucio-prototype
[23] https://github.com/ESCAPE-WP2/rucio-analysis

## Assessment

Various tests were performed to address the four user stories described above:

- **User Story 1**
  - **Test 1**: **End-to-end data lifecycle**
    Regular data products (100x10MB files once every 4 hours) were uploaded for a day at each of the two "telescope" RSEs (IDIA in Cape Town and AARNET in Perth), incorporating the following aspects to mimic telescope data product placement into a staging area:
    - Uploading the data to a non-deterministic source RSE and registering the data in place in Rucio;
    - "Subscription"-based movement (movement based on rules set before data exists);
    - Long haul transfers;
    - Life-time based QoS transitions (including start and end dates);
    - Short Rucio lifetime (1hr) at source so that files are deletable at source storage as soon as successfully transferred.

    This test was fully successful.

  - **Test2: Long-haul transfer**
    The test was specific to data transfers from IDIA and AARNET to Manchester and Lancaster (both UK sites). The impact of different file sizes on the overall throughput was investigated: file sizes of 50 MB, 500 MB and 5 GB (transferring between 3-7 TB of data in total for each test) were tested.

| Source site | Protocol | 50 MB file size average throughput | 500 MB file size average throughput | 5 GB file size average throughput |
|---|---|---|---|---|
| **IDIA (South Africa)** | GridFTP | 10.2 MB/s | 11.3 MB/s | 15.0 MB/s |
| **AARNET (Australia)** | xroot | 3.2 MB/s | 8.3 MB/s | 11.0 MB/s |

*Table 2. SKAO's long-haul transfer test results for User Story 1.*

As shown in Table 2, all transfers show a positive correlation between file size and throughput, with the transfers from AARNET (via the xroot protocol) being a factor of 3 slower at 50 MB than at 5 GB file size.

- **User Story 2**
  Supporting pre-planning of data product location per project required the use of a subscription service triggered by the addition of new data products into the Rucio instance. The data products were associated with appropriate metadata matching the undertaken

astronomy project. This was necessary to ensure that data collection could be built based on metadata and made available at a single SRC site in order to easily co-analyse data.

- **Test 1: subscriptions**
  During the DAC21 subscriptions to add data products with the Rucio Digital Identifier (DID) column “project” metadata field matching a predetermined value were created. The subscription would employ the Rucio “transmogrifier” daemon to create a metadata-based replication rule, which would then pick up further data meeting this criteria. Some temperamentality in the “transmogrifier” daemon during testing was observed, though for the DAC21 it worked as expected. Every hour data products were uploaded into the non-deterministic storage at Lancaster, registered to Rucio with the metadata field included, and replication followed asynchronously. It should be noted that the subscription only had to be created once, after which the “transmogrifier” created the rule, which would apply to subsequently registered data. The only issue which was encountered was Rucio’s "conveyor-submitter" daemon (responsible for submitting transfer requests) crashing for unrelated reasons; the daemon however caught up with the backlog upon restarting the instance after deploying 2 replicas of the daemon running with more threads. The test was resolved and continued smoothly for the remainder of the day. Rucio was quite resilient to daemon down-time, as its architecture relies on another daemon running, checking if all rules are satisfied and triggering changes if not.

- **User story 3**
  - **Test 1: running interactive analysis on the ESCAPE DLaaS**
    A scaled-down SKAO workflow running on simulated SKA science data (the software was already onboarded into the ESCAPE WP3 OSSR) was run. During the DAC21 week a JupyterHub environment was successfully created, and added to the CERN JupyterHub deployment. The container image was based on the "datalake-singleuser" image[24], and additionally contained the SKAO ("Science Data Challenge 1") software[25]. The DLaaS system allowed users to search for data products (based on scope and filename) and replicate them to the local CERN storage - which is mounted within the notebook environment - giving read-only access to the file. Ideally, access to an instance with more RAM available would have enabled a full-scale workflow; however, using a smaller test image was sufficient as a proof-of-concept (the full scale workflow ran on SKAO’s JupyterHub instance, and the full-size images were already replicated in the Data Lake).

    The upload of results from the DLaaS back into the Data Lake using the user interface failed. The success of this test has, however, been demonstrated since (for CNAF-STORM, as an example site that accepts OIDC authentication tokens). During DAC21, we worked around this issue by using the Rucio CLI to upload the generated data products back into the Data Lake.

- **User story 4**
  - **Test 1: permissions with tokens**

[24] https://gitlab.cern.ch/escape-wp2/docker-images/-/tree/master/datalake-singleuser
[25] SKAO software demo: https://drive.google.com/file/d/1nDDtbXipQeZBiDtN7GVp71NlfVqxYhga/view

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.

The use case was aimed at ensuring that appropriate access permissions, recognising group membership, were given to users. The test also investigated the scalability of the service outside the HEP environment. A key part of the development was moving to OAuth2 JSON web tokens, allowing X.509 certificates to be deprecated. Several explorations were carried out (both theoretical and actual), taking into account which authentication flow was needed and what was either already supported, or theoretically supportable, by the WP2 technologies.

The SKA development of the Rucio instance was moved over to token-based authentication. The procedure was not straightforward, and it involved a lot of iteration to establish the correct configuration. Although this had been proven to work in the past, the code had been refactored since then, thus preventing some functionalities from working as intended. These issues had to be expected and were solved by pull requests to the Rucio software. This is still an active area of development and deployment, as configuration changes to be tested on SKA's deployment were identified.

The use of X.509 authentication in the SKAO ecosystem was also a problem from a resource provider perspective, since site admins were not familiar with all the details of grid storage, how to install, configure, and maintain it. To understand how tokens might help with this, a token-based StoRM-WebDAV endpoint was deployed with the help of ESCAPE colleagues at INAF. The endpoint serves as an RSE in the SKA Rucio instance with https protocol enabled; apart from the wider current issues with the token-based deployment, the technology behaved as expected. Documentation was relatively easy to follow[26] and the process was documented[27]. The latter was a very useful outcome as a simple guide was written for site admins to integrate POSIX file storage as an RSE and participate in the future SKAO prototyping efforts.

**Future outlook/summary:**

Overall, SKAO has been positively impressed with the data management functionality offered by the DIOS stack, and is especially keen to continue testing Rucio and FTS, and the integration with AAI mechanisms going forward. SKAO is considering the option of using Rucio as a means for getting new data products from the generation sites and distributing them to selected SKA Regional Centre sites globally for users to access.
On the other hand, SKAO identified a number of areas where further development work is needed to support SKAO data access policies, which are summarised in Table 3.

| Feature required | Ease of fixing / notes |
|---|---|
| INDIGO IAM groups synchronised with Rucio account attributes | Sync script currently exists, would need to create a definition on how groups map to Rucio account attributes (and thus permissions in the policy package). |
| Ability to control view permissions | This is essential for the protection of embargoed data. |

[26] https://github.com/italiangrid/storm-webdav
[27] https://gitlab.com/ska-telescope/src/ska-rucio-prototype/-/blob/master/notes/setup-storm-webdav-rse.md

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.

| | |
|---|---|
| | Rucio has been written with open access in mind, and although the current policy package implementation can restrict access to functions e.g. list_dids, it cannot restrict the returned results of these functions, i.e. all users can view all data. |
| **Feature required** | **Ease of fixing / notes** |
| Method APIs have necessary information to verify request comes from authorised user | Some functions within the policy package do not have the necessary parameters to restrict permission based on e.g. scope and account attributes, these would need to be included in the arguments wherever the corresponding permission checking functions are called from. |
| Embed SKAO data access policies into Rucio policy package | Need to write a custom Rucio permission policy package that restricts functionality according to requesting account attributes (and correspondingly, INDIGO IAM groups). |

*Table 3. Future work areas identified to support SKAO Rucio use cases.*

SKAO has a clearer picture of the development effort required to achieve some of the embargoed data use cases and would like to keep working with the Rucio development team on this, and we want to take a closer look along with the Rucio development team to understand the impact and feasibility of these changes. We also want to develop a better understanding of our use cases and SKAO is interested in assessing the option of using a translation layer (like a science platform) that has the responsibility of managing user permissions as an alternative to investing further development effort in Rucio.

Over the next 18 months, until the end of 2023, SKAO's intention is to develop the SKA Rucio instance by improving the token-based support and onboarding new sites and new prototyping team members. The experiment aims to be in the position to run further data delivery and dissemination tests (with challenging data volumes) by the end of 2022, ready for assessment against the early global SRC Architecture designs, and high-level requirements, in 2023.

## 4.3 MAGIC

### Description

The MAGIC (Major Atmospheric Gamma Imaging Cherenkov) Telescope system consists of two Imaging Atmospheric Cherenkov telescopes located at the Roque de los Muchachos Observatory on La Palma, one of the Canary Islands. It is designed to detect gamma rays tens of billions to tens of trillions of times more energetic than visible light.

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.

## Requirements

The MAGIC data centre is located at PIC in Barcelona and receives stereo observed data (RAW and Reduced datasets) on a daily basis from the on-site IT infrastructure. Reduced datasets are provided by the on-site pipelines which are in charge of computing the data and registering in the on-site database. Thus, MAGIC's participation in the ESCAPE project was instrumental in assessing the implementation of Rucio as the main transfer manager, and particularly in testing each of the above points on a real case scenario.

Currently, the MAGIC data transfering system is carried out through a set of scripts interfacing FTS that are becoming obsolete. Therefore, the requirements for the orchestration of the MAGIC data transfer from the origin to PIC must be to:

- Explore the MAGIC onsite namespace and create the destination namespace which introduce some differences
- Detect new data produced from MAGIC observations
- Regular data transfers: O(1TB/day)
- Transfer different file sizes, from MB to GB
- Implement a mechanism to delete already transferred data
- Run in an automatic way.

For some time, an effort was devoted to the design and development of tests to imitate MAGIC's data production system. So, by the time the FDR20 arrived, the PIC was able to execute running tests on the ESCAPE Data Lake using mock data in a way that mimicked a first approximation of the workflow data product for MAGIC. By means of the Rucio software, the tests consisted of recording all data that appeared on the source endpoint, defined as a non-deterministic RSE at PIC (PIC-INJECT), which simulates la Palma on-site resources, and replicating to the destination endpoint defined as a deterministic RSE, also located in the PIC (PIC-DCACHE). On the PIC-INJECT RSE, 24 files were created every hour. At the end of the 24 hour test period 240 files each 600 MB were created. By setting up a cron job, the Rucio client script was initialised to search for candidate files to be transferred on the source RSE (PIC-INJECT). Once identified, they were registered with the Rucio database, and the relevant replication rules were created for the destination RSE (PIC-DCACHE). Once transferred, the same replication rule allowed the deletion of the files at the source. This test demonstrated that the configuration for RSEs, file transfers and their automatic deletion was successfully managed by the ESCAPE instance.

FDR20's results encouraged PIC to adopt Rucio and manage a Data Lake outside the ESCAPE-managed infrastructure. Hence, following on from the experiments carried out on the FDR20, and in the months leading up to the DAC21, a Rucio instance (version 1.26 LTS) for MAGIC experiments was deployed on top of a kubernetes platform at PIC. FTS (v3.10) was used as a transport layer, while Elasticsearch (v7.12.1) and Prometheus (v2.11.1) were employed for logging and metrics monitoring. The storage services consisted of an instance of dCache (v6.2.33), with exportations using GridFTP, and xroot protocols. On the database side, the instance on the Rucio PIC used a PostgreSQL database (v13.3) along with Ceph storage.

At the same time, PIC has been working on further developing tests to mimic future MAGIC data movements, and stress tests for specific aspects of the Rucio functionality. That is, two RSEs were set up, one in La Palma On-site, which was the non-deterministic algorithm where data would be generated continuously from the MAGIC experiment, and the other with a deterministic path creation algorithm at PIC. In this phase, some steps were changed to be more transparent in the migration from the old transfer system and the Rucio transfer script. Specifically, the discovery of files was done from a query to an onsite database, giving the path of the files to be transferred and the status control of each transferred file. Symbolic links were then created at the source and it is through these links that the transfer took place. The advantage of doing the transfer and deletion from the symbolic links was the possibility of using real MAGIC data without worrying about compromising it.

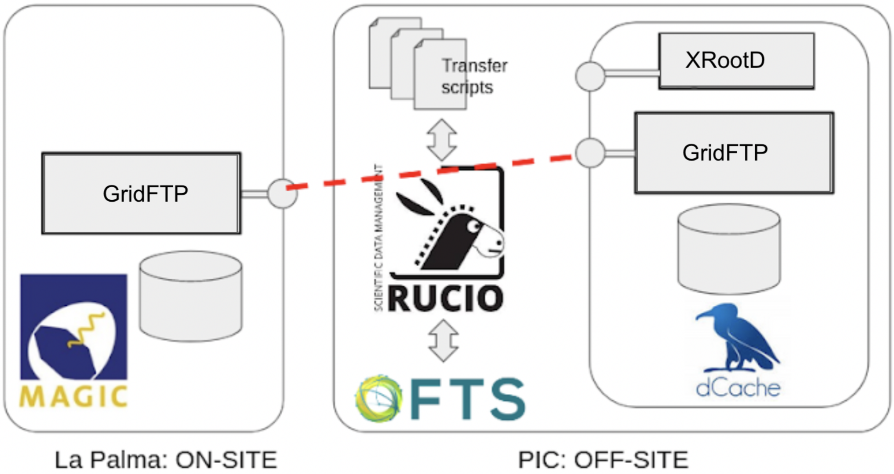


*Figure 7. Data appearing at the storage buffer next to the MAGIC telescope gets automatically replicated to the destination endpoint. The destination endpoint is part of the ESCAPE Data Lake where related policies and replications rules defined on this data apply. Source files at the telescope storage buffer get deleted and storage freed for new data to come.*

## Assessment

For the DAC21, The following two use cases were proposed for testing based on the experimental setup described above and the selected MAGIC requirements.

- **MAGIC - Long haul ingestion and replication**
  Testing the workflow of MAGIC automatic detection of new appearing data, addition of metadata, and data transfer from the Observatorio del Roque de los Muchachos (ORM) in the canary island of La Palma to the PIC Data Centre in Barcelona.
- **MAGIC - Data reprocessing**
  Testing the integration of the ESCAPE Data Lake with Jupyter Notebooks to search and instantiate files from remote when running a generic type of astroparticle physics analysis.

Table 4 outlines the results and details of each test. Both the tests and the use cases showed long term success.

| Use Case | Test name | Context | Estimated #Files, data volume | STORAGE | Replicas/RSEs | Results |
|---|---|---|---|---|---|---|
| MAGIC01 | Test 1 | CTA/PIC-RUCIO | 631 GB | Disk | non-deterministic and deterministic RSEs | Completed successfully |

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.

| | | | | | | |
|---|---|---|---|---|---|---|
| | Test 2 | CTA/PIC-RUCIO | 3155 GB | Disk | | Completed successfully |
| MAGIC02 | Test 1 | ESCAPE-RUCIO | 675MB | | | Completed successfully |

*Table 4. MAGIC's test results during the DAC21.*

Overall, the work done for the MAGIC experiments in WP2 has evolved from simple workflows of dummy data transfer between two endpoints within the PIC exercised during the FDR20, to a duty cycle very close to the real MAGIC implementation, as seen during DAC21. The exercise has seen a wide range of data lifecycle stages (i.e. simulation, injection, ingestion, and analysis) and underlying platforms that have shown a great advantage for MAGIC, such as Rucio and Jupyter. In particular, Rucio could efficiently orchestrate the data transfer from la Palma (onsite) to PIC (offsite) endpoint for long haul distances. The orchestration also worked automatically to discover and transfer new files.

## 4.4 LOFAR

### Description

The Low-Frequency Array (LOFAR) is the currently largest radio telescope that operates at the lowest frequencies that are observable from earth. LOFAR is a multi-purpose sensor network of antennas spread over Europe, grouped in stations. The core of the array is built in the north of the Netherlands (near the village of Exloo). LOFAR has been operational since 2010, and currently a major upgrade of the system (referred to as LOFAR 2.0) is on the way. Apart from being an operational telescope, LOFAR is also a prototype for the SKA.

The data collected at the antennae is first combined at the station by local hardware. By doing this, the station is essentially 'pointed' at one or multiple direction(s) in the sky, depending on the observation mode. The data is then streamed over dedicated 10Gbps network connections (each station produces 3 Gbps of data) to the central correlator, hosted by the University of Groningen (in Groningen). The correlator combines the data from the different antenna stations to a single signal. The combined data is then further processed offline and sent off to the archive.

### Requirements

Further processing could consist of anything between averaging the data in time and frequency, calibration, to generating images. Currently the bulk of the data is ingested with a low level of post-processing, but work is on its way to store calibrated data products in the archive, and images from large projects are also being released.

The main requirements the Lofar Long-Term Archive (LTA) has on a data management system have to do with life cycle and data access requirements. LOFAR data is acquired based on scientific proposals. The data will then need to be made available for further processing by the Principal Investigator (PI) or any group of people appointed by them (i.e. the science team) for further processing. In the first period (in the current LTA a week), the data should be directly downloadable by the science team, so it should be guaranteed to be on a storage system (or QoS level) that guarantees direct download access (e.g. a hard disk pool). In parallel the data is also copied to a slow cheap tier of storage (e.g. a tape library). After this week, it is expected that the science team

has downloaded all the relevant data, and proceed with processing and the data on the direct access tier can be removed (unless the team requests for the data to be staged again). Further access to the raw data may be needed on an irregular basis, but then this involves the extra step of first moving the data from the cheap to the direct access tier (referred to as staging).
After a set period in time (currently one year), the data in the archive is made publicly available. From that moment on, everybody should be allowed to stage data and retrieve it.

**The LOFAR Long-Term Archive**
The LOFAR Long-Term Archive (LTA) consists of a distributed set of storage locations. Currently the three locations are Jülich (Germany), Poznań (Poland), and Amsterdam (The Netherlands). The metadata is kept in a central database which is hosted in Groningen. From the Central Processing, the data is ingested into a disk pool at one of the archive locations. The data is then pinned to stay on disk for at least one week. Next to that, the data is put on a tape system on which it is stored for the longer term. During the first year after the observation, the data is exclusively accessible for the team that submitted the proposal. After this year the restrictions are lifted and data becomes publicly available. Since the data is stored on tape, it needs to be first copied to a disk pool. Currently LOFAR requires the systems that provide storage to the LTA to support the SRM protocol in its backend. To prevent users from having to install srm clients, and to obtain X.509 certificates for incidental data access, , a custom-built service has been set up, that wraps the staging commands in an API. In the current setup, access is limited to users known to our system, to prevent abuse, log usage for statistical purposes and traceability in case of unintended usage (e.g. users trying to stage petabytes of data). It is foreseen that with the migration from local AAI to Federated AAI, it will be easier for new users to gain access to our public data.

## Assessment

The data management architecture under investigation in the DIOS work package fitted the model of a distributed archive in which decisions should be made on where data will be stored, based on rules. Three use cases, each describing how current data life cycle being mapped on the Data Lake could be envisioned, were investigated.

- **Use case 1: Data ingestion and life cycle**
  The goal of this use case was to demonstrate how data could be ingested from the central processing into a Data Lake, defining the life cycle in advance. In general, LOFAR datasets can be in the form of large files, which may interfere with the way data transfer tooling handles checksum computations, causing network timeouts to occur because the checksum is still computing while the data is already transferred (leading to an overall transfer failure). The use case therefore required the transfer with tooling external to Rucio (i.e. gfal), the registration of the dataset and the application of life cycle rules to happen. In practice, the implementation of the use cases consisted of the following steps:

    1. Data was copied to a webdav endpoint, that is part of a non-deterministic RSE, using *rclone.*
    2. The data was registered with Rucio using the *python library*.
    3. A rule was added to Rucio to store the data on a QoS level representing the currently used disk pool (using QoS level *CHEAP-ANALYSIS*), for a set time (*3 hours*), using the *python library*.

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.

4. Another rule was added to store the data on a QoS level representing the currently used tape library (using QoS level *SAFE*), for a set time limit (*10 hours*)

   It should be noted that step 4 was time-limited because of the nature of the ESCAPE Data Lake as a test instance with limited storage. In a real case scenario, the rule would not be time limited. The use case was considered successful if the data was ultimately accessible in the correct QoS level for the configured times.

- **Use case 2: Data organisation and imaging**
This use case modelled the use of external compute infrastructure to take uncalibrated data out of the LTA, image it and ingest the image data product in the LTA. This also involved grouping files to observations. The data from a single LOFAR observation consists of hundreds of directories, each containing files representing a single wavelength band. Each of those directories was archived into a single file and stored in the LTA.

  To create an image, the data from a pointing towards the source of interest (referred to as the target) needed to be downloaded, together with data from a calibrator, which is a well-known source observed at a time and location which is close to the observation. The calibration solutions were then applied to the target observation, and an image (in FITS format) was generated based on the target observation. Finally, the image could be ingested back into the Data Lake to make it accessible to scientists. The upload of data products to the ESCAPE Data Lake instance assumed the products to be at a QoS level that made it possible to download them directly. The steps taken to achieve this were:

  1. The data products were split into three groups. Each of those three groups of frequency products was added to a dataset in Rucio. Those three groups could represent two *target* observations and a *calibrator*.
  2. The group defined to be the *calibrator* was combined with each of the two simulated *target* observations to form two *containers*. This meant there were three datasets and two containers (which we refer to as an *observation package*).
  3. Each of the two *observation package*s was downloaded on a processing node.
  4. On this node both observation packages were imaged and combined to a single image.
  5. The resulting image (in FITS format) was uploaded back to the Data Lake.

  Step 4 in the list above is technically out of scope for the assessment of the data management infrastructure. This use case was successfully completed if the *data package*s were correctly downloaded and the final image was uploaded to the Data Lake.

- **Use case 3: Multiwavelength imaging**
The third and last use case consisted in taking the image that was created in Use Case 2 from the Data Lake and combining it with an astronomical image from another instrument. This use case integrated the work from the ESCAPE WP5 (ESAP), from WP4's Virtual Observatory and from WP2's DLaaS. The goal was to create a composite image in the Jupyter notebook and upload it back into the Data Lake. The following steps were considered:

  1. Finding the image created in use case 2 using the ESAP framework.
  2. Finding an image from a different telescope using the Virtual Observatory protocols.

3. Launching the DLaaS and obtaining the data from the ESAP shopping basket in the notebook.
4. Combining the data from both sources into a new image (in FITS format).
5. Uploading the image to the Data Lake.

The success criterion of this use case was having a composite image uploaded to the Data Lake from the notebook.

| Use Case | Estimated #Files, data volume | STORAGE | Replicas/RSEs | Results |
|---|---|---|---|---|
| **Use case 1** | Around 10 files (~2TB) | Disk, Tape | non-deterministic and deterministic RSEs | **Completed successfully.** Bandwidth to the original non-deterministic RSE was low. After switching to another one this use case ran through well. |
| **use case 2** | 31 files, 140 GB | Disk | deterministic RSE | **Completed successfully.** All went as expected |
| **use case 3** | Image from UC2 + image from the VO. Fe 100 MB max. | Disk | deterministic RSE | **Completed successfully.** All went as expected |

*Table 5 : Results of LOFAR use cases.*

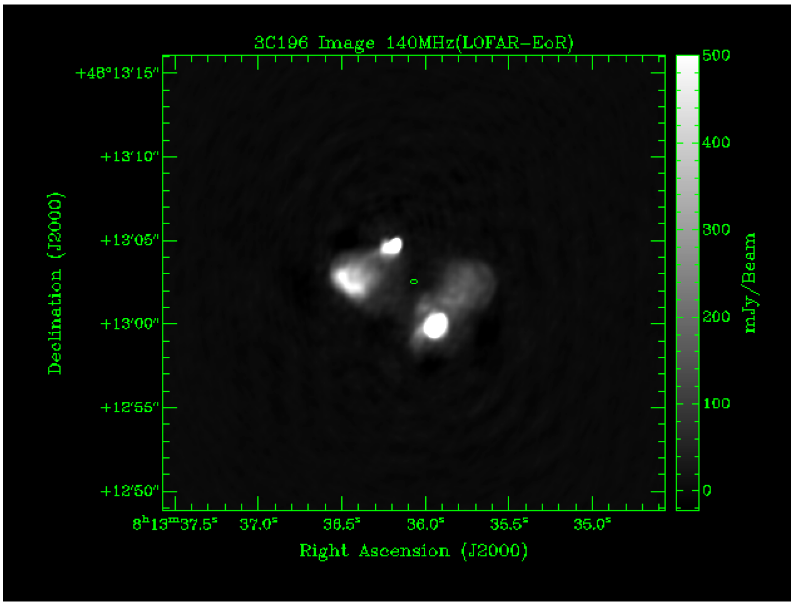


*Figure 8. LOFAR image created using Use Case 2 of 3C196.*

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.

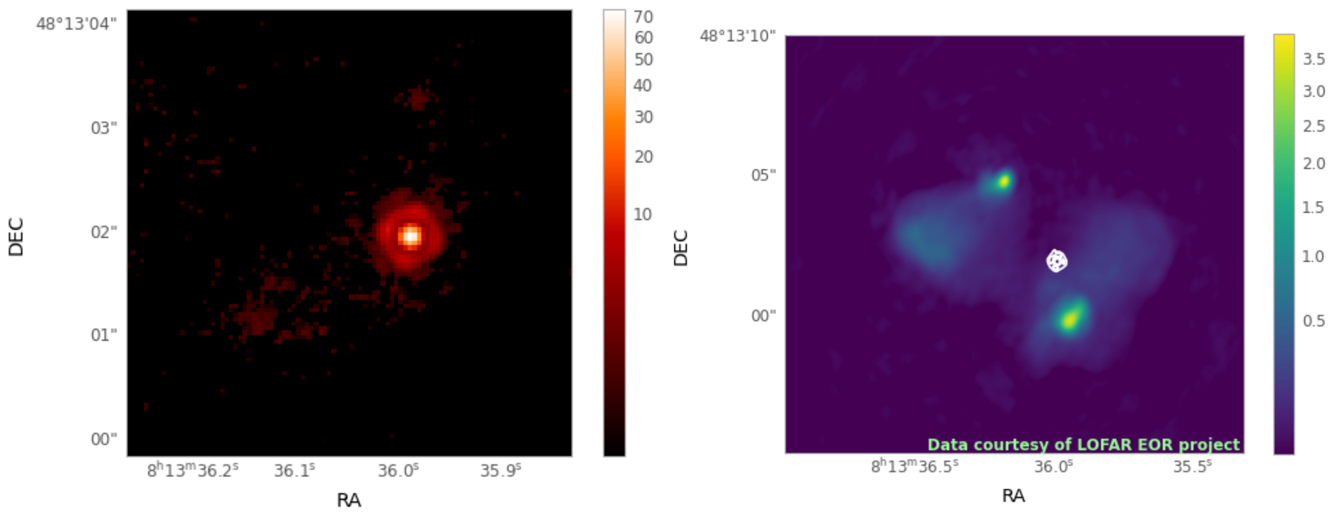


*Figure 9. Left: Hubble Space Telescope (HST) image of 3C196. Right: Composite image of the LOFAR image of LOFAR from 3C196 with in white the contours from the HST. Both images were made in the DLaaS notebook following Use Case 3.*

## 4.5 KM3NeT

### Description

KM3NeT is a distributed infrastructure and consists of a network of deep-sea neutrino detectors in the Mediterranean Sea with user ports for Earth and Sea sciences. The detectors will eventually consist of 345 detection units in total and will have a total data output rate of 100 Gbps. The detector is currently under construction: every detection unit whose construction is completed is deployed and is then operational. At the time of the ESCAPE DIOS DAC21 exercise, about 10% of the KM3NeT infrastructure was operational.

### Requirements

The general data processing concept of KM3NeT consists of a tiered processing structure. The data collection encompasses data products at different abstraction levels. For the DIOS DAC21 exercise, KM3NeT presented three use cases, one at each of the three layers in the computing model. These use cases also demonstrate the data lifecycles and the data handling at the different abstraction levels. The use cases were performed on real data from only one of the detector sites, which corresponded at that time to about 5% of the complete infrastructure.

At the first level, located at the Tier 0, a computing cluster filters all raw data from the deep sea in real time, where signal patterns are selected that are consistent with scientifically relevant information. These computing clusters are located in the shore station at each detector site. The main requirement is the reduction of the overall data rate by about three orders of magnitude. The output is temporarily stored on a persistent medium and distributed with fixed latency (typically less than a few hours) to at least two computing centres.

### Assessment

- **Use case 1: Data ingestion and replication at Tier 0 level**
  The corresponding DAC21 use case at the Tier 0 level was the ingestion into the Data Lake of data output by the shore station filters, and its replication. The data ingestion took place

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.

once every day during the DAC21, and included all data collected on the persistent medium in the shore station during the past 24 hours. The data corresponded to data level 1. The ingestion was to safe storage, as the data needed to be stored permanently. After the ingestion, two replication rules were applied: the data was replicated to a safe storage at another site, in order to have a backup available of all level 1 data. The second replication rule replicated the data to disk storage, from where it could be easily accessed for Tier 1 data processing at a later stage. The DIOS exercise automated the process of copying level 1 data to different sites and storage types, avoiding manual operations. The storing of metadata, also implemented within the DIOS data management tool, facilitated data processing at Tier 1 level. The combination of metadata storage and data organisation into datasets and containers greatly contributed to data findability and ultimately facilitated Tier 1 data processing. This use case was finalised by a checksum to verify the transmission and storage, after which level 1 data was deleted from the storage elements at the shore station. This was necessary to have disk space available for newer data as the detector operates 24/7.

- **Use case 2: Data processing at Tier 1 level**
  At the second layer (Tier 1), the filtered data was processed. Various models for hit patterns were fit to the data, usually referred to as reconstruction. In order to achieve an optimal pointing resolution of the neutrino detector, the reconstruction step required a detector calibration using auxiliary data to determine the time offsets, efficiencies, positions, and orientation of the detection elements (photomultiplier tubes) in the sea. The results of the calibration and quality identifiers were stored in a central database system. The typical update frequencies of the photomultiplier tube positions and time offsets were $10^{-2}$ Hz and $10^{-5}$ Hz respectively. The calibration and reconstruction procedures operated with a fixed latency of a few hours or less. The fully calibrated reconstruction results were level 2 data products, and were stored for further analysis. The corresponding DAC21 use case at the Tier 1 level performed only part of the complete data processing chain. This use case performed the calibration step where the time offsets and efficiencies of the photomultiplier tubes were determined. In this, the level 1 data that was ingested into the Data Lake in the first use case was retrieved from the fast access storage, and was calibrated. The resulting product was ingested into the Data Lake for further Tier 1 data processing. The possibility of metadata storage and the organisation of data into containers and datasets were used to easily find level 1 data, and to label and organise level 2 data for further data processing. During the period of DAC21, this use case was performed every day, on all the data that was ingested into the Data Lake from the shore station just a few hours before. A lifetime was set to the replica of the level 1 data on disk, as these were only intermediate products and were no longer needed after they had been processed (two replicas of the level 1 data were still available on safe storage).

  The calibrated and reconstructed data produced from Tier 1 data processing was initially stored on storage labelled with "Fast" QoS. The data would then be accessible by KM3NeT collaborators and would be used for different scientific analyses. Reconstructed data analyses were processed at the local computing clusters of the partner institutes (Tier 2).

- **Use case 3: Data analysis at Tier 2 level**
  For the DAC21 use case at the Tier 2 layer, KM3NeT used the DLaaS framework. A fully reconstructed dataset, covering about a week of data, was previously ingested into the Data Lake. The DLaaS was used to find this data, and the integrated development environment (as

an online notebook) was used to convert this data into a reduced dataset for a high level data analysis. Consequently, the data product was ingested back into the Data Lake. This use case was performed only once during the DAC21 period, as it encompassed the last step for a higher level analysis, covering about a week of data. The DLaaS framework offered a user-friendly interface to the Data Lake, as well as an integrated development platform for straightforward data operations.

**Future plans**
In addition to data handling and data processing needs, the KM3NeT experiment requires dedicated simulations to assess the detector efficiency and systematics. It is expected that the data volume of the simulations exceeds the data volume of the observatory by a factor of about 10. As the simulated data is processed in the same way as the observational data, it was assessed that the DIOS infrastructure can be used for the data management of both simulated and observational data.

KM3NeT will need to grow into a production setting and scale to the full research infrastructure; in fact, the experiment is currently transitioning to distributed data processing and a distributed data infrastructure. At the moment, KM3NeT is investigating the implementation of a DIOS-like infrastructure, and has already started to set up an instance of the DIOS data management tool, Rucio, for its data management pipelines.

## 4.6 ATLAS

### Description
ATLAS is a general-purpose particle physics experiment at the Large Hadron Collider (LHC) at CERN. It is designed to exploit the full discovery potential of the LHC, pushing the frontiers of scientific knowledge. ATLAS' exploration uses precision measurement to push the frontiers of knowledge by seeking answers to fundamental questions. ATLAS is the largest detector ever constructed for a particle collider: 46 metres long and 25 metres in diameter. Its construction pushed the limits of existing technology. ATLAS is designed to record the high-energy particle collisions of the LHC, which take place at a rate of over a billion interactions per second in the centre of the detector. More than 100 million sensitive electronics channels are used to record the particles produced by the collisions, which are then analysed by ATLAS scientists.

### Requirements
In 2021 ATLAS implemented a complete strategy for the usage of the ATLAS Open Data[28] (AOD) project as a playground to design and perform multiple tests regarding the use, prototyping, evaluation and experimental particle physics analysis execution on infrastructure like the the Data Lake, multiple JupyterLab standalone instances developed at LAPP and ATLAS at CERN, The RUCIO CLI at multiple instances: CERN LXPLUS, CERN OpenStack self-installation, LAPP computer facilities, personal computers using containers, and the Data Lake as a Service (DLaaS).

The current ATLAS Open Access resources are dataset with a compendium of several hundreds of millions of real (10/fb) and simulated (~150/fb) events of proton-proton 13 TeV collisions at the LHC

[28] http://opendata.atlas.cern/software/

and a set of analysis examples written in C++ (12 analyses) from Standard Model (SM) to Beyond SM searches; as well as more than 20 Jupyter notebooks with multiple examples analysis in Python.

## Assessment

Several activities were designed to run during the DAC21 that can be summarised in two main categories:

- **Upload, bookkeeping, download and direct access of ATLAS Open Access datasets (ROOT files) to and from the Datalake**
  Artificial datasets were generated; multiple versions of the same data were created, simulating a data-augmentation process (from a few hundred GBs of data to single-digit TBs of data). This procedure allowed the creation of up to six TBs of artificial data uploaded and retrieved to/from the Data Lake in several months, with a higher intensity during the DAC21. The exercise allowed to test not only the system, but also the Rucio client and the Rucio Jupyter extension, the scope and other ways to bookkeep and retrieve samples, and the usage of monitoring tools to learn (and document) how users interact with such tools at their disposal.
- **Usage of the Data Lake infrastructure to perform realistic final particle physics analysis with the Open Data samples**
  The onboarding of the AOD C++ analysis framework in the WP3's OSSR was completed. This step was relevant to all WP2 members as, in order to have a fully reproducible analysis, the datasets, software, pipelines and workflow need to be preserved. As analyses can take between ~5min to 3 hours in a commodity computer, in order to mimic the activity related to the final physics analysis with derived data and triggered by a single user or group, the same analysis with augmented samples was run on different platforms (VMs, DLaaS, laptops). "Heavier" download and replications of samples between sites (or RSEs) and from/into the Data Lake was also tested.

All the mentioned activities and tests allowed the documentation and training of newer members on the usage of the different tools and the overall pipeline design, extending the ATLAS LAPP activities to other members outside the ATLAS collaboration that were part of ESCAPE (and EOSC).

The exercises tested the data production, replication and documentation before and during the DAC21. 2000 open-access files were uploaded to the Data Lake. The required redundancy levels of the data is ensured by Rucio, in this particular case two copies of the files were transferred to two different RSEs in the Data Lake.

- **Use case 1: Data multiplication**

  *Objective: using the current AOD datasets to create more data samples artificially.*

  For this use case, ROOT files could be added to the Data Lake when they shared the same internal structure (i.e. same trees inside). The data augmentation was achieved by artificially multiplying the datasets up to any arbitrary value. The physics results of artificially created data are meaningless. Large ROOT files (~2.5GB) were generated, as well as smaller ROOT files (~4MB), eventually accounting for ~10 TBs of artificial data. Data from the Datalake was then requested at a higher rate than the analysis, and performance tests on the possible differences to retrieve large-but-few files versus many-but-small files were run. This could also be done in/from different RSEs. The exercise included the analysis of data stored in the Data Lake and the upload of the results (small files of ~100's kB size each) back into the Data Lake. The analyses were submitted making use of the Rucio client on the JupyterLab UI. A series of scripts to automatically augment data upload and replication was tested. The test was successful: the data was successfully stored and transferred/replicated among several RSEs. Basic metadata was stored. Users were able to discover data using the ESCAPE Rucio instance via the Rucio client and the Jupyter Rucio extension. The cleaning procedure of the data from the Data Lake was also tested.

- **Use case 2: Data analysis**

  *Objective: using different AOD analysis examples and any other open-access code and examples[29] that can run over the mentioned samples.*

  The use case included checking the simulation of data calibration and data augmentation and writing the results to the Data Lake. The execution of analysis examples over the multiplied data to simulate longer analysis pipelines was tested. The analysis was then modified to write outputs (ROOT files) and store them back into the Data Lake. The analysis examples were correctly executed in several instances. The access of the data samples via the Jupyter Rucio extension was tested and validated. Multiple analysis pipelines lasting from a few minutes to a few hours (<4 hours) ran successfully. The localisation, retrieval and usage of the data in several types of user analysis was proven successfully. Several storage QoS transitions were executed to mimic the changing popularity of data for user analysis. Transitions from FAST to SAFE storage and from SAFE to CHEAP-ANALYSIS storage were executed.

## 4.7 CMS

### Description

The Compact Muon Solenoid (CMS) is a general-purpose detector at the Large Hadron Collider (LHC). It has a broad physics programme ranging from studying the Standard Model (including the Higgs boson) to searching for extra dimensions and particles that could make up dark matter. Although it has the same scientific goals as the ATLAS experiment, it uses different technical solutions and a different magnet-system design.

### Requirements

The CMS experiment activity has been about setting up and testing several scenarios intended to demonstrate the capability of the ESCAPE infrastructure to be integrated with the experiment workflows, and also to enable the data access on an heterogeneous set of workloads and resources. The first target was to deploy an infrastructure accessible via INDIGO-IAM-issued token authentication that would serve as a single entrypoint for analyzers, with everything as transparent as possible for them. The second one was to demonstrate that an actual analysis reading data from the lake (eventually enabled through a cache layer) could actually work in different authentication

[29] https://hepsoftwarefoundation.org/

and workflow execution patterns. Finally, to prove the analysis capability to access data also in scenarios where the resource scale out was performed on the Marconi A2 HPC hosted at CINECA. Moreover, there was a “desiderata” for having the possibility to manage CMS “embargoed” data in the Data Lake, thus allowing a subset of ESCAPE users to have exclusive access to a certain storage namespace.

In this respect, the activities of the first period were focused on setting up all the needed pieces for an infrastructure (kubernetes based) at CNAF capable of providing users with on-demand interactive notebooks, with all the required CMS software for analysis and accessible to all ESCAPE users via INDIGO IAM authentication. The data access phase was addressed via the instantiation of a dedicated cache server at CNAF, capable of authenticating with the ESCAPE Data Lake and forwarding the data to clients on the analysis infrastructure network. In this regard, a set of configuration files was made available to reproduce the different working patterns tested. The first one was a simple authentication proxy where the server authenticated with a service account to the Data Lake, while adopting a different authentication method for the client incoming requests. The second one introduced the cache-to-disk capability, together with a first evaluation of protocol translation, meaning clients may request data via WebDAV from the cache while the cache was taking data from an xroot endpoint, and vice versa.

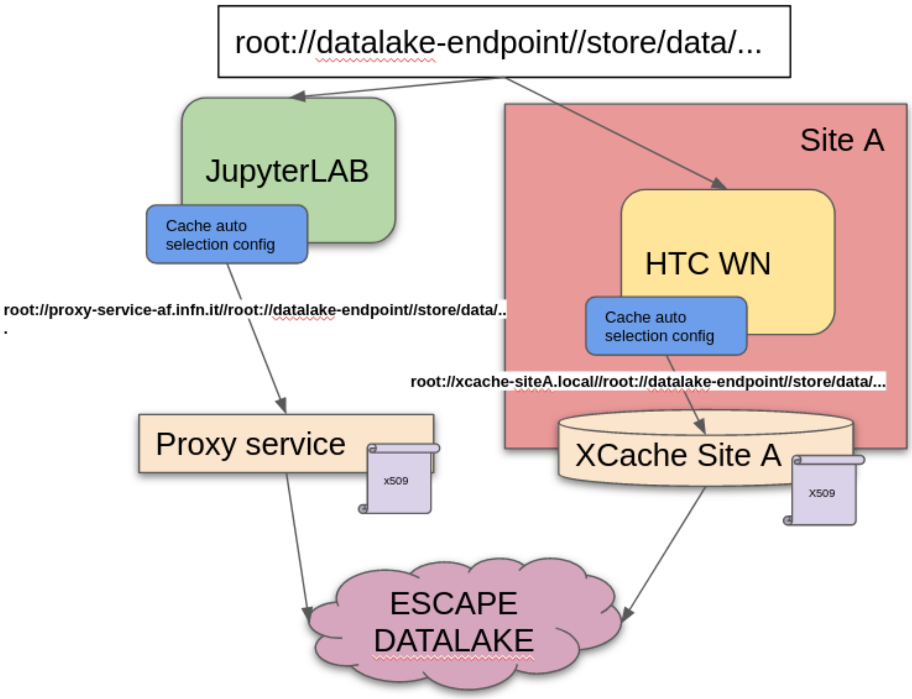


*Figure 10. Extended CMS analysis connection to the ESCAPE Data Lake, with the Inclusion of an HTCondor batch system instrumented with a caching service (XCache).*

In the same period the capability of such of an infrastructure to spawn notebooks directly on a HPC node (CINECA Marconi A2) via SLURM has been demonstrated[30]; data-access, thus the fan-out

30 https://projectescape.eu/news/escape-dios-supporting-cms-experiment-obtaining-maximum-value-its-data-through-high-performance

connectivity, in this scenario was enabled by another dedicated cache instance at CNAF. All the authentication at this stage was performed via X.509.

### Assessment

During DAC21 exercises an end to end workflow fully based on authentication via tokens was an important target milestone. As a first step a CMS Open Data dataset was injected into the ESCAPE Data Lake as the input for the subsequent activities. The infrastructure for the CMS analysis has been then extended with the inclusion of an HTCondor batch system, also in this case the authentication was using the ESCAPE IAM and the cluster hosted at CNAF. At this point, everything was in place to proceed with several functionality tests; the capability to discover input data and register analysis output back to the Data Lake while interacting with Rucio via token authentication has been demonstrated from a single user notebook.

```
[7]: file_list = []
     for repl in cli.list_replicas([{"scope":"CMS_INFN_DCIANGOT", "name": "ESCAPE-CMS-Opendata"}], schemes=["davs"]):
         file_list.append(list(repl["pfns"].keys())[0])

     print(file_list)

     ['davs://dcache-door-doma01.desy.de:2880//escape/wp2_rucio_testbed/desy_dcache/CMS_INFN_DCIANGOT/72/27/Run2012B_Doub
     ucio_testbed/desy_dcache/CMS_INFN_DCIANGOT/df/47/Run2012B_DoubleMuParked.root', 'davs://dcache-door-doma01.desy.de:2
     12C_DoubleElectron.root', 'davs://dcache-door-doma01.desy.de:2880//escape/wp2_rucio_testbed/desy_dcache/CMS_INFN_DCI
     esy.de:2880//escape/wp2_rucio_testbed/desy_dcache/CMS_INFN_DCIANGOT/64/e4/SMHiggsToZZTo4L.root', 'davs://dcache-door
     ANGOT/0b/5f/ZZTo2e2mu.root', 'davs://dcache-door-doma01.desy.de:2880//escape/wp2_rucio_testbed/desy_dcache/CMS_INFN_
     scape/wp2_rucio_testbed/desy_dcache/CMS_INFN_DCIANGOT/42/5d/ZZTo4mu.root']
```

*Figure 11. Notebook interaction on the DLaaS with CMS Open Data. Data Lake files are imported in the notebooks and real paths to storage (i.e. DESY-DCACHE) endpoints are automatically provided.*

Finally, a first evaluation of the capability of the Data Lake to manage CMS "embargoed data" has been tested on STORM-CNAF endpoint with success, allowing access to a subset of data only to users belonging to the ESCAPE CMS group and collecting a sample configuration available for other interested sites/experiments. Exclusive access to reserved files in the Rucio scope CMS_EMBARGOED_DATA was achieved, targeting only the members of the IAM escape/cms group. The tested authorization workflows demonstrated the functionality for both X.509 and token based authentication protocols.

For all the reasons reported above, the CMS experience has been very fruitful with a particular mention of the outstanding feedback received from other people working on similar topics; this synergy has been fundamental to achieve all the objectives.

## 4.8 R$^3$B (Reactions with Relativistic Radioactive beams)

### Description

The University of Groningen (RUG) is one of the top research institutes in the Netherlands and is involved in many large-scale and data-intensive international projects like FAIR-GSI, CERN, LOFAR and KM3NeT. For the ESCAPE collaboration, the university has chosen the R3B (Reactions with Relativistic Radioactive Beams) experiment because of its large data volume and data analysis complexity to challenge the ESCAPE WP2 infrastructure. The R3B is a multi-purpose experimental facility at FAIR-GSI used to study nuclear structure properties of short-lived isotopes through inverse kinematic reactions. Each type of outgoing particle is detected by a specific type of detector. The neutrons produced at the target, which are generally also very forward-boosted, are detected by Neu-LAND (Neu Large-Area Neutron Detector). The problem of finding the shower head among all the scintillator signals in Neu-LAND is challenging. Especially in the situation where multiple neutrons have to be detected in coincidence, solutions are far from trivial because of two reasons: 1) it is not (always) known a priori how many neutrons were detected and 2) showers from distinct neutrons tend to overlap.

## Requirements

Two analysis approaches have been developed to analyse the data of the NueLAND: 1) Technical Design Report (TDR) and 2) Deep Neural Network (DNN). The R3B setup is under construction and, therefore, these two methods are applied to the Monte Carlo data from the specially designed simulation and data analysis framework of this experiment (R3BRoot). The data analysis workflow of R3B is almost the same for both simulation and real data (Figure 12).

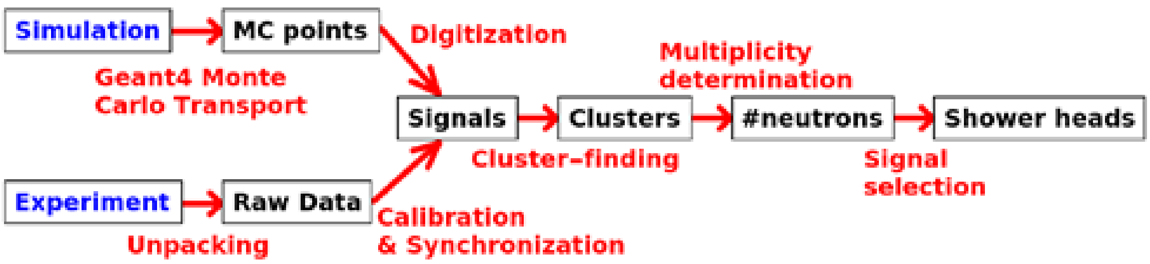


*Figure 12. The analysis workflow of the R3B-NeuLAND detector.*

## Assessment

For DAC21, the R3B simulation data generated using R3BRoot (simulation and data analysis framework of R3B) was used. Three use cases were explored: data ingestion, data replication, and data analysis.

- **Use case 1: data ingestion and replication**

  For data ingestion into the Data Lake, the Rucio client's Docker image, configured to access the Data Lake through the ESCAPE-IAM, was used. X.509 certificates were used to obtain the authorization to upload/download data to/from the Data Lake.

  The Python API of Rucio was used to manage the ingestion of the 2.6 TB of data into the Data Lake. All the data was successfully uploaded to the different RSEs during a 24 hours ingestion campaign. To protect the data, two copies of the data were placed on different RSEs using the Quality of Storage=CHEAP-ANALYSIS label. Both ingestion and replication went smoothly and successfully during the DAC21. All the operations were monitored through the Grafana dashboard, which made tracking the status of the tasks more convenient. A summary of the file transfer is presented in Table 7.

| | Volume of Data (TB) | Type of data | Number of files | Method(s) | RSEs or QoS |
|---|---|---|---|---|---|
| **Ingestion** | 2.6 | .root | ~ 1800 | Rucio Python &<br>Rucio CLI | FAIR<br>ROOT, GSI<br>ROOT, EULAKE 1 |
| **Replication** | 5.2 | .root | ~ 3600 | Rucio CLI | *CHEAP ANALYSIS* |

*Table 7. A summary of the R3B data transfer to the ESCAPE Data Lake during DAC21.*

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.

- **Use case 2: data analysis**
  For the analysis of the R3B data, Python-Root scripts were developed in the format of Jupyter Notebook. To perform the analysis, the DLaaS was used to interact with the Data Lake, perform the analysis, and send the results back to the Data Lake. The analysis was based on reconstructing the invariant mass spectrum of the 4-neutron multiplicity for different analysis methods.

  Despite some minor issues, the interactions with ESCAPE DIOS services have been promising. Most of the issues with different services were resolved with a common effort since the beginning of the project. R3B is interested in using DIOS services in future collaborations with international experiments.

## 4.9 CBM/PANDA

### Description

The antiProton ANihilation at DArmstadt experiment (PANDA) at FAIR will use antiproton beams of up to 15 GeV leading to antiproton-proton collisions of maximal 5.5 GeV in the centre of mass system and a data stream of 200 GB/s into the Data Center (GSI's Green IT Cube). The average interaction rate of 800 kHz is reduced by an online filter by a factor 100 to an 8 kHz event rate which is stored for offline reconstruction.

PANDA will generate 1 PB/year of raw data which has to be securely stored in a cold storage, mirrored on different computing centres, and needs to be accessible with a low latency for reprocessing in a hot storage. In addition, up to 2 PB/year of high-level data is generated which is used by the users to perform physics analysis. This data is complemented by 3 PB/year of simulated data which is needed for the analysis. Both the high-level data as well as the simulated data has to be available on a time scale of 10 years.

The Compressed Baryonic Matter experiment (CBM) will run at an interaction rate of up to 10 MHz with typically Au+Au collisions of up to 11 GeV leading to a peak raw data flow of 1 TB/sec into GSI's Green IT Cube. The First-level Event Selector (FLES) processor farm corresponding to about 60,000 cores will do full event reconstruction and data reduction by event selection at a rate of 10 MHz.

### Requirements

CBM will generate 18 PB/year of raw data. The current computing model foresees that RAW data are delivered to on-site storage media (QoS label "RAW_HOT") and copied to long-term archive (QoS label "RAW_COLD"). Two copies of RAW data have to be archived in two distinct data centres, one being FAIR/GSI. From the RAW data about 4 PB of AOD data are created annually which are transferred to the participating data centres, serving a regional CBM user community. Additionally, 9 PB of SIM data are generated in correspondence with the experiment settings, conditions and statistics. The AOD data will be kept for 5 years while the SIM data will be available up to 3 years after production.

### Assessment

CBM addressed the following use cases during the DAC21 exercise.

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.

- **Use Case 1: mCBM data ingestion**

  The first use case was to inject existing mCBM data into the Data Lake without copying it over to the RSEs. To do this, the existing replica on the shared file system needed to be registered with the Rucio instance. The corresponding DAC21 test was run with a script polling the directory where mCBM data was replayed to, and registering files using the Rucio Python bindings. The test was successful, registering 100 replicas, a total of 400 GB.

- **Use case 2: CBM simulation**

  This use case described the process of simulating the transport of particles through the CBM experiment in its electron setup. Monte Carlo data was initially uploaded to the Data Lake, and then extracted to be used in simulations and files containing results. Afterwards, transport data was extracted and digitalisation data generated. The corresponding DAC21 tests were sporadically successful due to a client side error preventing one of the protocols being used.

- **Use case 3: mCBM reconstruction**

  This use case covered the reconstruction of recent mCBM (July 2021) data, getting it from the Data Lake. This test was partially successful during DAC21. In the prepared demo environment, the reconstruction run was run 100 times. Again, due to the client being misconfigured, the WebDAV protocol could not be used.

PANDA addressed the following use cases during the DAC21 exercise.

- **Use Case 1: PANDA data ingestion**

  The DAC21 test of ingestion for PANDA involved 125 parallel batch jobs sending fallback data (37500 files, ~2.3 TB) and from the simulation.

- **Use case 2: PANDA upload data from simulation activities**

  As PANDA simulation jobs were run on the cluster nodes, live data (37500 fles, ~2.3 TB) was uploaded directly from the cluster nodes, with 125 jobs running at the same time.

- **Use case 3: PANDA data reconstruction**

  The PANDA reconstruction use case was to have batch jobs read from the Data Lake and upload the processed data. The DAC21 tests succeeded only for small sets of test data, as the larger sets of test data could not be used due to the data being stuck in REPLICATING state.

| Use Case | Estimated #Files and data volume | replicas/RSEs | Completed |
|---|---|---|---|
| mCBM Ingestion | Zero copy ingestion of 400GB | FAIR-ROOT | Completed successfully |
| CBM simulation | Fallback input data: 1 run - ~ 3 MB x 1 replica<br>Output size: ~5 GB/run, 84 runs<br>Lifetime of output data: 48h<br>Total: 122 GB peak | FAIR-ROOT | Completed successfully |
| mCBM reconstruction | Fallback input data: 10 files x 10 GB x 1 replica<br>Output size: ~1 GB/run, 84 runs<br>Lifetime of output data: 48h<br>Total: 125 GB peak | FAIR-ROOT | Completed successfully |
| PANDA ingestion | Fallback input data: 500 GB x 1 replica | FAIR-ROOT, QOS=SAFE, QOS=FAST | Completed, but many files stuck in REPLICATING state |
| PANDA simulation | Output size: 1150 TB per run, 3 output replicas<br>Total: 3.95 TB | FAIR-ROOT, QOS=SAFE, QOS=FAST | Completed, but many files stuck in REPLICATING state |
| PANDA reconstruction | Fallback input data: 500 GB x 1 replica<br>Output size: 30 MB per run, 84 runs<br>Total: 505 GB | FAIR-ROOT, QOS=OPPORTUNISTIC | Completed, but many files stuck in REPLICATING state |

*Table 8. Assessment table for mCBM and PANDA use cases.*

The tests showed that overall, the Data Lake could be used for real workflows, including different stages of ingestion, simulation and replication, for both CBM and PANDA. The Rucio instance was used for larger amounts of data, using different QoS rules for the datasets. While issues arose during DAC21, they were minor and would not persist in a production environment. The assessment summary can be found in Table 8.

### Future Plans

CBM/PANDA will continue their investigations of integrating the Data Lake into their workflows. Moreover, the experiments are also interested in furthering several topics, including metadata, long-haul transfers, fine grained authorization and the use of tokens.

## 4.10 LSST

### Description

The Vera C. Rubin Observatory is being prepared to carry out the most ambitious astronomical survey attempted to date, the Legacy Survey of Space and Time (LSST). The main instrument is a large-aperture, wide-field, reflecting, ground-based telescope that will repeatedly survey the southern sky every few nights in six optical bands. Its imaging device is a digital camera with a resolution of 3.2 Gigapixels designed to take about 2000 exposures per observing night, for a total nightly raw data volume of about 20 TB. Its wide-field camera and its agile supporting structure will allow for covering the entire sky visible from the Chilean Andes mountains in less than 4 nights.

### Requirements

The ultimate science-enabling deliverable of LSST will be the fully reduced data. Every observing night, alerts to objects that have changed in brightness or position will be broadcast. In addition, every year a new data release will be produced and published: a set of immutable reduced data, images and catalogues of celestial objects and their physical properties, the starting point for most of the scientific discovery that LSST will allow for. Those releases will include well calibrated and characterised images, measurements of positions, fluxes, shapes and orbital parameters of automatically detected celestial objects.

During the 10 years of the survey scheduled to start in 2024, a dataset of 0.5 million TB of raw and reduced data will be progressively collected, transported and analysed using highly automated tools. This dataset, composed of both images and catalogues, will allow scientists to explore a wide range of astrophysical questions, from studies of the Solar System and the Milky Way, to explorations of the changing sky, to examining the nature of dark energy and dark matter.

The data processing campaigns for composing the annual data releases will be performed at 3 different locations. The data archive centre located at SLAC National Laboratory (CA, USA) will promptly receive images of the sky from the acquisition site in a mountain top in the Chilean Andes and distribute them to 2 sites in Europe: IN2P3 computing centre (CC-IN2P3) in Lyon (France), where an integral copy of the dataset will be permanently stored, and the LSST:UK consortium (UK), where a partial copy of the raw data will be stored. Annually, each of those 3 sites will locally perform a share of the processing of the raw data collected to date which aggregated will compose the annual data release.

Reliable replication of data among those long term storage and processing facilities is a key requirement of LSST's distributed data processing infrastructure. In the framework of the ESCAPE project, CC-IN2P3 took the initiative to explore how Rucio and FTS could help to satisfy the long distance data replication needs of LSST. In the following paragraphs activities performed and the lessons learnt are going to be introduced.

### Assessment

Prior to, and during, FDR20. In the initial phase of ESCAPE, educational instances of Rucio and FTS were deployed locally at CC-IN2P3. Connected to production-level local storage elements (dCache in this particular case), that testbed was a key initial step to get familiar with the data model supported by Rucio and to understand how to map that model with the specifics of the LSST data management assumptions.

Building on that experience, during FDR20 repeated ingestion exercises were performed using a dataset of 45k files for an aggregated volume of 800GB. The Rucio command line interface, packaged in the form of a Singularity container, was used to ingest the dataset to the storage element, first using compute nodes in the local batch farm and then a dedicated set of hosts. Several ingestion runs were performed until 0% error rate was reached and an elapsed time of 68 minutes, for a movement of 150k files and 2 TB of data.

**DAC21**

The goal for LSST for the DAC21 was to perform a realistic inter-site data replication exercise of approximately one night's worth of LSST data, repeatedly over 5 consecutive days. The time budget set for each replication campaign was 12 hours, which corresponds to the average duration of one night in the location of the telescope.

The dataset selected for this exercise was composed of 4000 exposures recorded in the form of 800k files in FITS format, for a total of 15 TB, roughly equivalent to 1 night of raw images in terms of data volume and 2 nights in terms of number of files. The average file size was 18MB, which is a representative individual file size of how raw data will be recorded.

Prior to starting the exercise the dataset was placed at a CERN storage element connected to the Data Lake. This storage element played the role of the replication source, while CC-IN2P3 played the role of replication destination. The network latency between those sites (about 4ms) was not representative of the conditions under which data replication for LSST will be performed. Pre-placement of data was performed using the Rucio CLI against ESCAPE's Rucio instance shared with other science projects. Rucio was daily instructed via time-limited replication rules to perform the replication of the entire dataset from CERN to CC-IN2P3.

This exercise was very instructive and a fantastic learning experience. Suspiciously high transfer error rates were observed, even if FTS correctly recovered them. This was traced to some issues in the way dCache, CC-IN2P3's storage element, handles some specific conditions. A total of 4 hot fixes were delivered by the dCache development team during the exercise and promptly applied by dCache experts at CC-IN2P3. On day 5 Rucio- and FTS-driven error-free inter-site replication of about 800k files in less than 8 hours (set time budget) were demonstrated.

**Next steps**

In all the previous exercises X.509 grid proxy certificates were used for authentication purposes. The experiment intends to explore authentication based on bearer tokens over the whole chain of services: IAM, Rucio, FTS, dCache. Although independent token tests of some of these components have been successful, authentication over the whole chain has not yet been achieved. LSST's aim is to achieve the aforementioned work before the end of ESCAPE lifetime.

**Beyond ESCAPE**

LSST has started investigating in realistic conditions the relatively small individual size of the files composing the dataset (18 MB on average), suspected to have a significant impact in the replication throughput. Instances of Rucio and FTS were deployed by the Rubin project at SLAC National Laboratory. The Rucio instance was used to upload the same dataset used in DAC21 to SLAC and the same replication exercise has started, this time using the high-latency trans-atlantic network link from SLAC to CC-IN2P3 (150 ms). Improving the replication throughput by tuning the parameters of FTS has also been ongoing work.

## 4.11 EGO/VIRGO

**Description**

VIRGO is a Michelson Interferometer with 2km-long arms, built to detect gravitational waves, located at the European Gravitational Observatory (EGO) in Italy. To date VIRGO has completed three observing campaigns alongside other gravitational wave detectors, over several years. The next observing run will not be until at least June 2022.

### Requirements

VIRGO's data moving use case is interesting because it requires some very low latencies, in particular the need to perform online processing quickly and share any events with the LIGO detector in the USA - if both detectors see evidence for a gravitational wave at the same time (appearing in the Event Aggregation Database), this will lead to an astronomical alert being triggered and subscribing (electromagnetic) telescopes can move to observe the region of interest.

This need for low latency data movement has led to a complex data distribution diagram (Figure 13), which has some inefficiencies in data transfer since scientists at some European sites would still access EGO/VIRGO data through Caltech. An alternative option could be as seen in Figure 14, where dedicated transfers still manage low-latency data movement if required, but a Data Lake concept makes data products available for offline computing at the compute centres at CNAF and IN2P3 and also to users across Europe and in the LIGO consortium.

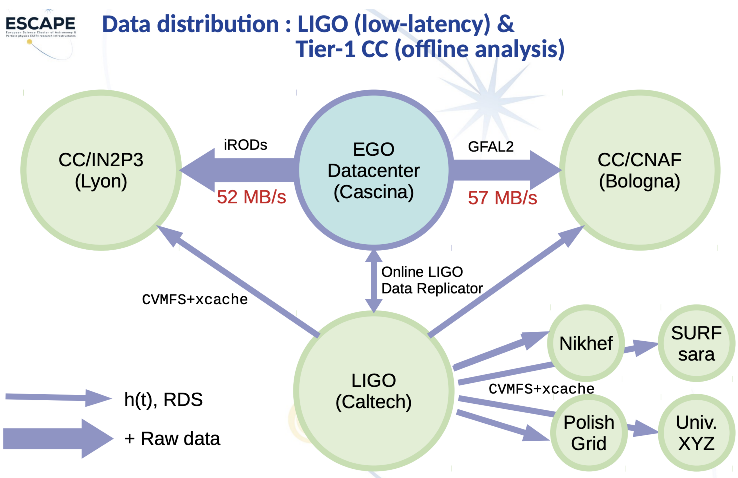


*Figure 13. EGO1. Data distribution diagram.*

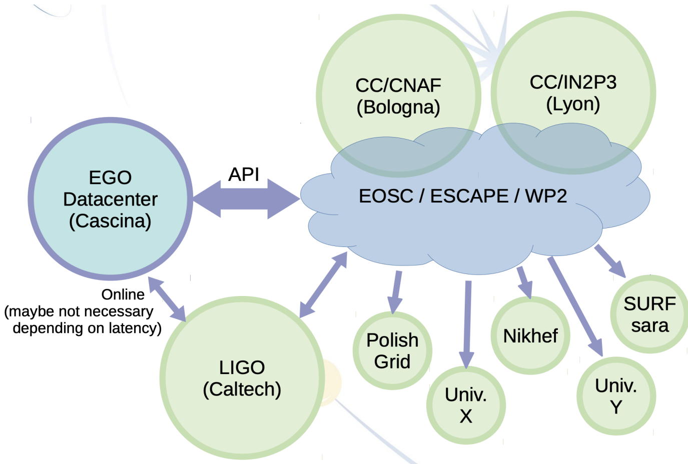


*Figure 14. EGO2. EGO1 has been improved via the Data Lake concept making data products available for offline computing at the CNAF and IN2P3 centres.*

EGO Raw data has a lifecycle - initially "Hot" and needed for low-latency processing and online analysis (at EGO), it then would move to "Cold" when it is used about once per month for detector characterisation. This would be data younger than ~2 years. After that, data would be moved into the archive, from where access might be needed only 2-3 times per year to help calibrate the signals. Thus data placement on different QoS classes (e.g. onto slow Tape or fast disk) could be determined by rules depending on the access frequency for each dataset - this is handled by Rucio using the rucio-c3po daemon.

### Assessment

The goal of the test performed during the FDR20 was to exercise Virgo's data real-time transfer needs to DIOS. This was done by injecting public Virgo data into the Data Lake, and downloading it afterwards for data verification. The scope was to assess the ability of the infrastructure to make available the real-time strain data to pipelines and tools assessing the data quality. A 4 hour dataset was uploaded to the Data Lake containing 14,400 one second data chunks. The test's progress on Nov 17th can be seen in the form of a corresponding ~14.4 k increase in the number of DIDs associated with the VIRGO experiment. All chunks were timely transferred, consolidated in the Data Lake infrastructure and verified after a subsequent download at the CNAF facilities.

## 5. Established relationships and outlook towards joint initiatives

The encouragement to pursue extra activities, identified as potentially relevant extensions of WP2/DIOS's original goals, was gathered during the first Project Review and via fundamental feedback in several workshops and meetings. The Data Lake orchestration layer and the tools rapidly proved to give a solid response during the initial phase of the project and the experiments managed to port workflows in the early prototype as demonstrated during the two Data Challenges. This resulted in the adoption of the core Data Management services by several ESF(RI)s. The identified relevant extensions were geared towards widening the Data Lake flexibility to demonstrate its usability and benefits for varied scientific communities. Some examples include extra metadata functionalities in Rucio for Astronomy, Virtual Research Environments, or Photon and Neutron (PaN) sciences use cases, which promoted end-user analyses, Open Data, Citizen Science, Outreach, Teaching and Training initiatives.

**Rucio**

The Rucio project and its dissemination was strongly accelerated by ESCAPE. This is especially visible by the expansion from high energy physics, where Rucio was already well established, into the astronomy domain. ESCAPE and several ESF(RI)s are now an integral part of the Rucio community as well as the Rucio development team. The 2021 Rucio community workshop was co-organised by ESCAPE members and featured a dedicated session on Astronomy and metadata use cases. Numerous ESF(RI)s also presented ideas at the workshop. Two ESCAPE members took formal responsibility in the Rucio development team and are currently leading the development of the metadata and authentication components within Rucio. These contributions prompted a significant evolution of functionality of Rucio. Besides these focused developments, ESCAPE WP2's engagement with Rucio, through the continuous testing, reporting of operational issues and bugs and participation in the weekly Rucio meeting, has been responsible for crucial improvements of the Rucio software stack and its documentation. Future initiatives, already identified, will be focused on metadata evolution as well as token-based authentication. The metadata development will aim to integrate metadata storage and query capabilities needed by ESF(RI)s to Rucio. The token-based authentication aims to evolve the current support for OIDC tokens to fully capability-based tokens. This will enable the Rucio software to truly enforce data embargos through the entire data infrastructure, a feature crucially needed by several ESCAPE communities.

**Photon and Neutron Sciences**

PaN user communities are very diverse and wide-ranging in their requirements, the commonality is to have an increasing amount of data and a need for distributed computing. The ESCAPE Data Lake model and vision, with its Data Management and File Transfer service, was shared with these communities. There is an opportunity to use the Data Lake as a simple service, as the complexity can be much reduced depending on the use case.

A joint workshop[31] to understand the scientific data management needs of the PaN community was organised with PANOSC and EXpands. Next steps should be geared towards prototyping a real use case to: store, distribute and access data from PaN source.

**FENIX HPC project: integrating heterogeneous computing resources:**

Exploitation of heterogeneous resources, CPU cycles and various storage backends has been addressed and successfully integrated with the Data Lake (see the preamble for AWS/Google cloud resources integration). A collaboration with the FENIX project[32] composed by six European supercomputing centres (BSC (Spain), CEA (France), CINECA (Italy), CSC (Finland), CSCS (Switzerland) and JSC (Germany)) was established, which agreed to align their services to facilitate the creation of a common Infrastructure. A collaboration between FENIX and ESCAPE WP2 was also initiated. The goal is to provide a framework of common tools to enable sciences to make use of HPC resources with Data Management capabilities to be able to access data and steer/move data to/from HPC facilities. To achieve this an AAI infrastructure synchronised with the internal HPC storage and with the RSEs is needed. Throughout the last months, FTS has been implemented to have Swift/S3 capabilities ready. A successful integration of the ESCAPE IAM as a *community* with the FENIX AAI

---

[31] PaN ESCAPE Data Management Workshop, 12th January 2021.

[32] https://fenix-ri.eu/about-fenix

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.

federation has been achieved, using an OIDC-to-SAML proxy (a SATOSA[33] proxy has been chosen to satisfy this use-case). This is also a good demonstrator towards an EOSC ecosystem for AAI federations. This activity will be certainly pursued after the ESCAPE project in the framework of EOSC-Future.

**CS3MESH4EOSC: sync&share platforms integration**

Concerning the CS3MESH4EOSC project, a collaboration started to integrate data transfer services based on RUCIO/FTS and the ESCAPE IAM into "ScienceMesh", the new federated infrastructure being built within CS3MESH4EOSC based on storage sync&share platforms. To this end, FTS-driven HTTP Third-Party transfers were demonstrated between nodes of ScienceMesh, and in the coming months it is foreseen to provide a full integration between ESCAPE and ScienceMesh sites, such that managed data transfers can seamlessly take place across them.

**EOSC-Future project**

The successful ESCAPE model is already being integrated into an early stage development of the European Open Science Cloud “EOSC-Future” project. The synergies between the two European projects, both targeting scientific collaboration across diverse platforms, are being exploited in various ways.

The ESCAPE Data Lake infrastructure’s functionalities, namely data injection, replication, processing and monitoring, are being extended to serve the Astrophysics scientific community as well as the High Energy Physics one, providing a reliable AAI framework and the federated storage services with the required Data Management needs. The EOSC Future Project is aimed at bringing together diverse analysis workflows from both communities, demonstrating how a common platform could support the goals of the Dark Matter Science Project and the Extreme Universe Science Project in the respect of FAIR data policies. Such a platform is being implemented as a Virtual Research Environment (VRE)[34], an ecosystem with different complexity levels enhanced by a seamless Web UI interface, where both scientists and new onboarding users will be able to run and reuse analysis workflows. This will be achieved fetching the data from the Rucio Data Lake, using the updated software version preserved on Zenodo, and retrieving the computational environment by leveraging the power of Reana[35], a reproducible research analysis platform initially developed at CERN, and of RECAST[36], the tool which enables the automation of signal passing through an analysis at the development time.

**The Virtual Research Environment, Analysis Platforms and Analysis Facilities**

The Virtual Research Environment (VRE) aims to be an environment that integrates all of the ESCAPE tools from WP2, 3, 4 and 5, in order to provide a framework within which to deploy the EOSC-Future Science Projects (SPs) and their sub-projects. The VRE will expose all the digital content related to a scientific result, making it easily findable, accessible and reusable; it will thus be an access point where interested researchers (coming from both EOSC-Future Science Projects) will be

[33] SATOSA is a proxy to translate between different authentication protocols (SAML2, OpenID Connect and OAuth2)
[34] https://escape2020.pages.in2p3.fr/virtual-environment/home/
[35] https://docs.reana.io/
[36] https://iris-hep.org/projects/recast.html

able to share data, software, interact with specific steps of a scientific analysis, and run workflows on external resource.

The ESCAPE Data Lake-as-a-Service model will be utilised as a springboard to spawn flexible re-analyses with the possibility to incorporate distributed hybrid workflows thanks to the synergies between file system handling specific to Reana and FTS. The K8s cluster's functionality on which the DLaaS relies on will be expanded to include a software repository (Zenodo[37]) plug-in enabling researchers to directly select the required computational environments from Docker images. The integration of Reana inside the VRE is already achieved on a client level; later implementation will connect a Rucio instance with a locally deployed Reana instance. This will include a common AAI infrastructure to ease the onboarding of new sciences on the platform and will allow Reana specification files to reference datasets from the Data Lake and software from Zenodo repositories. The VRE will then provide the common hub to run workflows on external resources being provided by EOSC-Future by taking advantage of the flexibility of containerized applications and of Reana's support of various computing backends (K8s, HTCondor, Slurm).

The VRE project, bringing together data and software access, workflow reproducibility and enhanced user interface, is onboarding sciences from various experiments (ATLAS, Km3Net, Fermi-LAT, EGO, LOFAR) and is gaining interest from the Analysis Facilities community: its progress was presented at the JENAS Symposium[38] and at the CERN's Analysis Facilities Forum[39]. The nuance of the VRE is that it will be completely open source, easily reproducible on different clusters, and easily accessible by anyone having an EOSC account; moreover, the target audience and workflows are not only HEP sciences, but also smaller experiments who would hugely benefit from the provisioning of shared computing resources.

## 6. Summary and future plans

The following are some of the reasons to keep the ESCAPE activities and legacy alive after the project end date:

- ESCAPE and EOSC-Future co-existence, synergies and transition: the ESCAPE Data Lake services, infrastructure and tools will be the basis for the EOSC-Future Science Projects;
- Consolidation: several experiments already deployed Data Lake orchestration tools (Rucio and FTS) for potential addressing their upcoming Data Management needs;
- ESCAPE WP2 is leading the activity for commissioning a new AAI scenario in HEP (X.509 to token-based auth);
- ESCAPE WP2 is leading the activity to address the file metadata needs of non-HEP sciences, this is reflected in the involvement with the Rucio development team and the joint interest to pursue a system that can deliver the required needs for a broad range of sciences;
- The ability demonstrated by a Data Lake architecture is of special relevance for the usage in heterogeneous resources, the ongoing collaboration with the FENIX/HPC project will be persisted after ESCAPE and within EOSC-Future;
- Large room for collaboration with sciences with well established data infrastructures, historically separated from HEP but with bigger commonalities than initially supposed.

[37] https://zenodo.org/
[38] JENAS Symposium, Madrid, May 2022 https://indico.cern.ch/event/1040535/
[39] Analysis Facilities Forum https://indico.cern.ch/event/1151054/

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.

Synergies should be further explored and pursued in the next months with the astronomy community: Virtual Observatory community and ESA in particular. A common technical workshop needs to be organised before the end of the project.

The ESCAPE Data Lake prototype and its overarching orchestration and identity management layer is providing the foundations of what could be seen as an early EOSC cell. These proto-cells are environments where data services co-exist with analysis tools to provide researchers with the required experiment specific framework. These frameworks would contain the full experiment computing environment, including data browsability and data access, software and code repos.
The integration of Data Lake tools on standard notebook technologies and analysis platforms opens the door to connect the scientific big data coming from the experiments with the diverse scientific community:

- Researchers log into these experiment specific EOSC cells and find the environment to start working with minimal configuration overheads.
- WP2 DLaaS toolkit bridges experiments related to Big Data with local site CPU and storage resources through capabilities for data browsing, download, upload and content delivery (caches) orchestration.

The Data Lake model and tools should enable experiments to implement their Data Management plans. These include:

- Providing the ability to humans and computers to **look for data and metadata** is at the core of Rucio, the Data Lake's Data Management System. Rucio connects the physical location of files at the sites (storage) with Logical File Names and associated metadata to the files and/or file collections (datasets);
- Extended support for **metadata** is an ongoing activity in WP2 together with scientific communities and science clusters. The possibilities to extend further browsability and searchability through metadata are huge for radio-astronomy, astronomy and cosmology;
- **Data accessibility** in the Data Lake is guaranteed via a common authentication and authorization method for users (ESCAPE IAM) and making use of well defined standard protocols for data access (http and S3/SWIFT, and legacy protocols like xroot, gfal);
- The commonality of these protocols for data access (upload, download, list, search) ensures a common **interoperability** from the client side on different services and platforms used for: a) data management (including raw data recording), b) data analysis and workflows implementation and c) data processing in heterogeneous resources (batch systems, commercial clouds and HPCs);
- ESCAPE WP2 activities and data challenges pursued all of the above across experiments and sites providing resources as a vision for further consolidation of the Data Lake model and the services and tools being prototyped. Perhaps one of the immediate consequences is being developed together with EOSC-Future project which, besides the three principles mentioned before, will pursue the idea of analysis **reproducibility** by integrating workflows in a well defined framework for data **reusability** making use of existing technologies like REANA and RECAST and making use of well defined repositories like Zenodo.

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. 

# 7. Selected presentations, publications and communications

| Title | Authors | Date | Weblink |
|---|---|---|---|
| Webinar, *"Steps forward in detection and identification of anomalous atmospheric events"* | PHIDIAS in collaboration with ESCAPE (Pascal Prunet (SPASCIA), Nicolas Pascal (ICARE), Dominique Jolivet (HYGEOS), Xavier Espinal (ESCAPE, CERN) | 13/10/2020 | https://projectescape.eu/events/phidias-steps-forward-detection-and-identification-anomalous-atmospheric-events |
| 24th International Conference on Computing in High Energy and Nuclear Physics (CHEP 2019) *"ESCAPE prototypes a data infrastructure for open science"* | Rosie Bolton (SKAO), Simone Campana (CERN), Andrea Ceccanti (INFN, CNAF), Xavier Espinal (CERN), Aristeidis Fkiaras (CERN) et al. | 16/11/2020 | *EPJ Web Conf. 245 (2020) 04019,* https://doi.org/10.1051/epjconf/202024504019 |
| *"Data Lakes in the PUNCH sciences"* | Xavier Espinal (CERN) and Christoph Wissing (DESY) | 11/02/2021 | https://www.punch4nfdi.de/sites/sites_custom/site_punch4nfdi/content/e112863/e128704/20210211.PUNCHLunch.v1.pdf |
| "PaN ESCAPE Data Management Workshop" | Xavier Espinal (CERN), Patrick Fuhrmann (DESY) et al. | 12/01/2021 | https://www.panosc.eu/events/pan-escape-data-management-workshop/ |
| *"Integrating services with the ESCAPE AAI"* | Andrea Ceccanti (INFN) | 19/01/2021 | https://indico.in2p3.fr/event/22812/contributions/89052/ |
| *"OIDC support in Rucio"* | Rizart Dona (CERN) | 19/01/2021 | https://indico.in2p3.fr/event/22812/contributions/89071/ |
| "New directions in Distributed Computing: DOMA and ESCAPE", [ESCAPE news] | Xavier Espinal (CERN) | | https://projectescape.eu/news/new-directions-distributed-computing-doma-and-escape |

| Title | Authors | Date | Weblink |
|---|---|---|---|
| World-wide LHC Computing Grid - Grid Deployment Board, "ESCAPE Full Dress Rehearsal Exercise results" | Riccardo Di Maria (CERN) | 13/01/2021 | https://indico.cern.ch/event/876772/#3-escape-full-dress-rehearsal |
| International Symposium on Grids & Clouds 2021 (ISGC 2021), *"ESCAPE, next generation management of exabytes of cross discipline scientific data"* | Riccardo Di Maria (CERN) | 24/03/2021 | https://indico4.twgrid.org/event/14/contributions/369/ |
| International Conference on Computing in High-Energy and Nuclear Physics (vCHEP 2021) *"The ESCAPE Data Lake: The machinery behind testing, monitoring and supporting a unified federated storage infrastructure of the exabyte-scale"* | Rizart Dona (CERN), Riccardo Di Maria (CERN) | 20/05/2021 | https://indico.cern.ch/event/948465/contributions/4323972/ |
| Second WP5 Workshop, *"Integration with WP2/DIOS - DLaaS Demo"* | Yan Grange (ASTRON) | 5/8/2021 | https://indico.in2p3.fr/event/24716/#8-integration-with-wp2dios |
| 4th Rucio Community Workshop, *"ESCAPE Data Lake as a Service"*, September 29th, 2021 | Riccardo Di Maria (CERN) | 29/09/2021 | https://indico.cern.ch/event/1037922/contributions/4536323/ |
| 4th Rucio Community Workshop, "SKA Rucio deployment and metadata/ data lifecycle use case" | Rohini Joshi (SKAO) | 29/09/2021 | https://indico.cern.ch/event/1037922/contributions/4488603/ |

| Title | Authors | Date | Weblink |
|---|---|---|---|
| 4th Rucio Community Workshop, "MAGIC telescope Rucio Deployment" | Agustin Bruzzese (PIC) | 29/09/2021 | https://indico.cern.ch/event/1037922/contributions/4488600/ |
| 4th Rucio Community Workshop, "CTA rucio use case and development with cloud storage" | Frederic Gillardo (CNRS) | 29/09/2021 | https://indico.cern.ch/event/1037922/contributions/4488602/ |
| EGI Conference 2021, *"The ESCAPE Data Lake as the bridgehead for the EOSC"*, October 19th, 2021 | Riccardo Di Maria (CERN) | 19/10/2021 | https://indico.egi.eu/event/5464/contributions/15676/ |
| Title | Authors | Date | Weblink |
| *Astronomical data organization, management and access in Scientific Data Lakes* to appear in the proceedings of Astronomical Data Analysis Software and Systems XXXI published by ASP | Y.G. Grange, V.N. Pandey, X. Espinal, R. Di Maria, and A.P. Millar (on behalf of ESCAPE WP2), 2022, | 02/2022 | https://doi.org/10.48550/arXiv.2202.01828 |

*Table 9. Selection of ESCAPE WP2 presentations and publications.*