Project Title	European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructure
Project Acronym	ESCAPE
Grant Agreement No	824064
Instrument	Research and Innovation Action (RIA)
Topic	Connecting ESFRI 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

D2.1 - IMPLEMENTATION PLAN AND DESIGN OF PILOT

Work Package	WP2, DIOS
Lead Author (Org)	Simone Campana (CERN)
Contributing Author(s) (Org)	Rosie Bolton (SKAO), Andrea Ceccanti (INFN), Xavier Espinal (CERN), Patrick Fuhrmann (DESY), Yan Grange (ASTRON),
Due Date	30.09.2019, M8
Date	27.09.2019
Version	1.0

Dissemination Level

X	PU: Public
	PP: Restricted to other programme participants (including the Commission)
	RE: Restricted to a group specified by the consortium (including the Commission)
	CO: Confidential, only for members of the consortium (including the Commission)

Versioning and contribution history

Version	Date	Authors	Notes
0.1	23.09.2019	Simone Campana (CERN)	Architecture document
1.0	27.09.2019	Jayesh Wagh (CNRS)	Final version

Disclaimer

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.

Table of Contents

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.

Table of Figures

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.

Project Summary

ESCAPE (European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures) addresses the Open Science challenges shared by ESFRI facilities (SKA, CTA, KM3Net, EST, ELT, HL-LHC, FAIR) as well as other pan-European research infrastructures (CERN, ESO, JIVE) in astronomy and particle physics. ESCAPE actions are focused on developing solutions for the large data sets handled by the ESFRI facilities. These solutions shall: i) connect ESFRI projects to EOSC ensuring integration of data and tools; ii) foster common approaches to implement open-data stewardship; iii) establish interoperability within EOSC as an integrated multi-messenger facility for fundamental science. To accomplish these objectives ESCAPE aims to unite astrophysics and particle physics communities with proven expertise in computing and data management by setting up a data infrastructure beyond the current state-of-the-art in support of the FAIR principles. These joint efforts are expected result into a data-lake infrastructure as cloud open-science analysis facility linked with the EOSC. ESCAPE supports already existing infrastructure such as astronomy Virtual Observatory to connect with the EOSC. With the commitment from various ESFRI projects in the cluster, ESCAPE will develop and integrate the EOSC catalogue with a dedicated catalogue of open source analysis software. This catalogue will provide researchers across the disciplines with new software tools and services developed by astronomy and particle physics community. Through this catalogue ESCAPE will strive to cater researchers with consistent access to an integrated open-science platform for data-analysis workflows. As a result, a large community “foundation” approach for cross-fertilization and continuous development will be strengthened. ESCAPE has the ambition to be a flagship for scientific and societal impact that the EOSC can deliver.

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.

Executive summary

This report describes the program of work to implement a prototype of the ESCAPE datalake, a data infrastructure supporting open science for the ESFRIs being part of the project.

The first part of the report explains the overall architecture of the datalake and the various functional elements. For each of the functional elements we analysed the available existing implementations, considered their adaptability to the needs of the scientific communities involved in ESCAPE and identified the most suitable options. The datalake should be seen as an ecosystem of services and tools rather than a monolithic infrastructure. In the context of our work package we will therefore identify and demonstrate a reference implementation, while the science projects involved in ESCAPE will be free to adopt only part of the service sets that we plan to integrate. In identifying which components to use for the functional elements we focused on large open source community supported products, leveraging as much as possible other EU-funded initiatives. Many if not all components will need some R&D work to cover all our use cases. Such R&D areas are explained and planned in this document, together with the effort needed for the integration of the different components.

The second part of the document gives a timeline of the project evolution which can be divided in two main phases. In the first, pilot phase the focus will be demonstrating at the functional level the choices we made in terms of architecture and reference implementation. We will therefore deploy a small scale pilot, spanning at least three of our partner institutes and demonstrate the core functionalities in terms of data transfer, data access and storage. We will also demonstrate the authentication and authorization model. The second, prototype phase will enlarge the pilot to all partner sites in our work package and expose all functionalities to our science projects. Scalability will also be demonstrated using real scenarios and applications from the ESFRIs.

At the end of the prototype phase we will be able to demonstrate that our reference implementation for the datalake can serve as an infrastructure for FAIR data, supporting the policies and use cases of our science projects.

1. Introduction

The goal of the ESCAPE Work Package 2 (WP2 DIOS - Data Infrastructure for Open Science) is to build a cloud of data services, often referred as datalake. The datalake should serve as core infrastructure to support open data and enable the FAIR principles, by providing a flexible and scalable infrastructure to store and access scientific data, while optimizing the total cost of ownership. Such infrastructure should be seen as a modular ecosystem of services. The WP2 will build and assess a pilot by the end of the Project Month 24, before finally developing a prototype choosing a reference implementation. The prototype will be available by the end of the ESCAPE project. The different science projects will be free to adopt such an implementation or to consider individual building blocks and replace others according to their needs, while continuing to adhere to the overall architecture. From the early stages of the WP2 work, it was agreed to limit the number of newly-developed components, while considering and evaluating first, state of the art solutions from different Research Infrastructures and European Projects.

The preparatory work for this document started therefore with the evaluation of such solutions. We held a series of fortnightly meetings from March to July focusing on available technologies and the role they could play in the datalake architecture. In early July we held the first WP2 workshop (Milestone 3.1 of the ESCAPE project) and during the first day we concluded the review of the different technologies. The second part of the workshop focused on the different ESFRIs' needs and use cases. The workshop was partly of joint nature and co-organized with ESCAPE WP5, given the many synergies in different areas of work. We also had presentations from WP3 and WP4 experts respectively on the Software and Service repository and the integration between the datalake and the Virtual Observatory catalogue. Those have been in fact identified as the areas of interface between WP2 and the other two work packages.

In the last session of the workshop we conceived a datalake architecture, which would cover the needs of the ESFRIs and be largely based on the technologies we evaluated. The Use Case document produced by WP5 as Deliverable 5.1 was also taken as an input to this process. That WP5 document, based on a survey, while focusing on science analysis, contained important information for WP2 as well, such as the data sizes and volumes, data organization and data access pattern of the different ESFRIs.

The next section will describe the datalake architecture and define the functional elements that will be part of the reference implementation. The subsequent sections will dive into the different components and, for each one, define a roadmap. The document will conclude with the integration plan of the pilot and the final datalake prototype.

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.

2. Datalake architecture

The architecture of the datalake in terms of functional elements is described in Fig.1. The core of the infrastructure consists of the storage services at the different facilities. These facilities differ in size and expertise and will be able to provide different classes of service: large national centres operate archive storage (today based on tape media), as well as smaller disk-based systems to serve compute intensive data processing. Other facilities operate disk-based solutions only, and in some cases such storage is volatile. The storage can be deployed as a distributed service, spanning multiple physical facilities but offering a single entry point for the users. The storage services rely on different software technologies, which need to interoperate by following the defined interfaces for file access, file transfer and storage management.

Data are replicated across the different storage services asynchronously through transfer scheduling service. Such service is able to trigger third party copy operations through supported storage protocols for many files in parallel, between different endpoints. It throttles the activity to optimize transfer bandwidth and network usage, while avoiding congestion. The data centres constituting the backbone of the datalake infrastructure are connected by fast networks, i.e. with at least multiple 10Gb/s links. These network links need to be monitored and the information fed back to the transfer service to be considered in its decision-making engine.

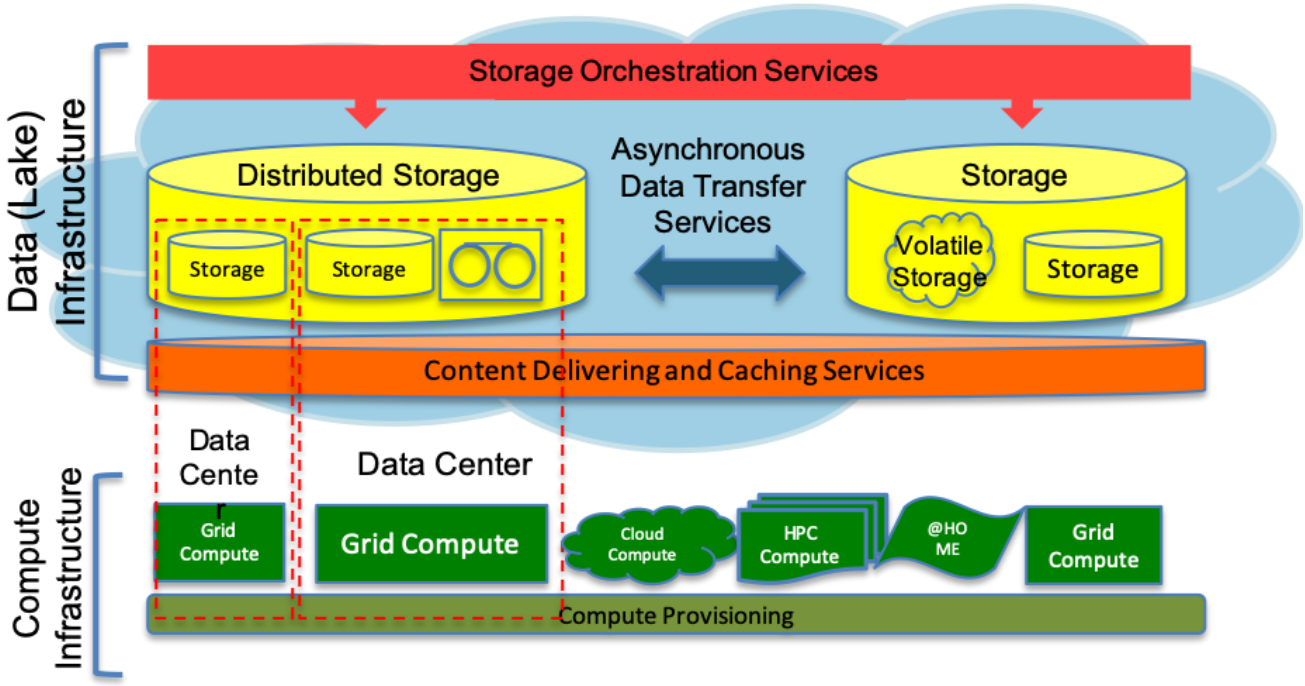

Figure 1: diagram showing the building blocks of the datalake infrastructure and the connection with compute services

A higher-level service organizes the storage and orchestrates the data inside the datalake. Such service offers functionalities for cataloguing the data files with some metadata, organizes the files in higher-level structures such as datasets, and stores file locations in the datalake. It also allows the definition of policies for data replication and data deletion, and enforces them through a set of

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.

asynchronous agents. It presents the end user with a set of APIs to manage and access the data in the datalake, hiding the complexity of its distributed nature.

In this model, the resources to perform compute on the data could be co-located with the data, i.e. at the same data centre where the data are stored, but in general one should expect that not to be the case. Accessing remote data comes with challenges and in general has an impact on the efficiency of the application. For this reason, a content delivery service is needed to hide the effects due to network latency in data access. At the same time, the service should be able to cache the most-accessed data close to the computing resources, to optimize data access performance in case of re-use. The content delivery and caching layer should be able to serve data to a heterogeneous set of resources, ranging from sites accessible through Grid interfaces, to Cloud resources, to High Performance Computers. For completeness, the Computing Provisioning layer, orchestrating workflows and pipelines is also shown in Fig. 1, though it is not part of the WP2 scope.

Finally, given the heterogeneous and distributed nature of the system, the Authentication, Authorization and Identity management (AAI) play a crucial role in the architecture. The system must accommodate the needs of the ESFRIs in terms of policies and allow open access outside the data embargo periods. At the same time, the mechanisms for authentication and authorization need to scale for exa-scale data management, reducing the overhead, while enforcing an adequate level of security.

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. Datalake components and reference implementation

Different components that will be used to build the datalake reference implementation are described in the following. These will be also accessible to the ESFRIs as stand-alone services. Some elements will necessitate R&D work to reach the level required within ESCAPE. This activity will be part of the WP2 implementation plan.

3.1. Storage Technologies

We plan to build the datalake leveraging a heterogeneous set of storage solutions: dCache [1], DPM [2], EOS [3], StoRM [4] and xrootd [5] at the minimum. Such technologies have been deployed for many years in the Worldwide LHC Computing Grid (WLCG) [6] infrastructure, and have demonstrated their capability to operate at the hundred-petabyte scale. Some storage solutions such as dCache, EOS and DPM are used also in federating geographically distributed data centres under a unique storage namespace. Such solution will also be included in the ESCAPE datalake prototype.

The data access will be enabled through a set of protocols. The xrootd protocol will be supported particularly for the High Energy Physics (HEP) use case, where files are stored in a ROOT [7] format. The HTTP protocol will be supported to cover the use case of most ESCAPE science projects as it is the most widely adopted within the open source community. Both xrootd and HTTP offer functionalities to upload/download and stream files, trigger third party transfers (more on this in the File Transfer Service Section), perform basic storage management operations and delete files. The gridFTP [8] protocol will be available for legacy reasons even though its usage will be highly discouraged given its uncertain future and its support model. The SRM [9] interface could be offered by some storage solutions, while not mandatory (with the exception of access to tape storage, for which SRM will remain the reference solution). Storages should be accessible through x509 credentials, but should also support token-based authentication, which will be the reference solution in ESCAPE, see section 3.5.

For the future optimization of storage cost and performance, Quality of Service (QoS) will be a key element. So far disk and tape have been used as storage classes to implement online and archive storage, respectively, but in ESCAPE we intend to consider a richer set of QoS metrics and a corresponding set of capabilities at the storage level. We are focusing on three basic metrics: Reliability, Performance and Cost, and we intend to enable access to storage resources corresponding to combinations of such capabilities. R&D work is needed also in the definition of the API for which a QoS request is passed to the storage. In addition, the storage needs to be able to map the request in a proper class of hardware of a configuration that can meet it. Different storage technologies have reached different levels of maturity in this respect and some development is needed to ensure all storages in ESCAPE are able to offer QoS functionality.

We plan to implement event-driven data management, in the sense that events at the level of the storage would trigger an action in some upstream service. Such events could be the appearance of a new file, the change of some file metadata, or the loss of some file, for example. The storage will therefore need to be able to dispatch events to a reliable messaging system, for the upstream services to consume. This will require some R&D work in WP2 as it will be described section 3.6.

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.

3.2. Asynchronous Data Transfer Services and Networking

The WLCG File Transfer Service - FTS [10] has been used in production for LHC experiments for more than a decade. It went through a major refactoring a few years ago to take advantage of modern technologies and improve scalability and performance. It has demonstrated the capability of efficiently performing file replication in a distributed infrastructure at the level of many millions of files per day, consisting of multiple petabytes. In ESCAPE we plan to use FTS as reference implementation for an asynchronous data transfer service across sites of the datalake. However, we expect R&D work in this area to meet the ESCAPE goals. The plan is to replace the gridFTP protocol, in use for many years for third party copy in WLCG, with alternatives better supported as open source software. Two viable alternatives have been identified in the xrootd and HTTP protocols. In ESCAPE we plan to complement the already ongoing effort in commissioning those protocols for third party copy, ensuring they are well supported by the various storage solutions and integrated in FTS. The commissioning should be seen both at the level of functionality, including verification of file integrity and metadata handling, and performance, through a set of stress tests and data challenges. From the WP5 survey it is clear that file sizes will vary by many orders of magnitude depending on the ESFRI and the use case. In commissioning third party copy we will need to verify the efficiency of the system in all scenarios. Particularly, treating very large files upwards of tens of gigabytes will likely require tuning of the transfer parameters. Dealing with very small files is also a challenge as performance is affected by overheads. Therefore, an important work in this area of WP2 will consist in integrating token-based authentication for asynchronous data movement to reduce the authentication overhead that, in the case of X509, is substantial.

We plan also to use FTS as an engine to trigger data staging from tape to disk buffers in an organized way. Its capability in this respect has been already demonstrated by the LHC experiments. However, modern storage solutions might not offer the SRM interface for file staging, and WP2 will ensure that FTS is able to handle the situation when such an interface is not available.

We plan to use the FTS monitoring and the capabilities of its data analytics backend for most of the commissioning activity. We will also expose the monitoring data through a set of APIs for the different ESCAPE ESFRIs to consume and possibly build specific dashboards. At the same time, we will instrument the infrastructure with a network-level monitoring system. For this, we agreed to use the perfSONAR [11] technology as it is already deployed at most of the data centres participating in the ESCAPE datalake. The plan is to ensure that each storage site is instrumented with a perfSONAR instance close, in network proximity, to the storage service. We will schedule a set of regular tests through perfSONAR, probing packet loss and throughput between all sites. The results will be displayed in an ESCAPE perfSONAR dashboard and accessible again through an API. At the same time, we intend to leverage the work in the GÉANT community in instrumenting key network waypoints with Data Transfer Nodes, allowing us to test network segments through a variety of protocols including gridFTP and HTTP. We see this information as complementary to the one offered by perfSONAR, and the integration of these two metrics set will be an important component, particularly to commission data transfers over long distances.

We finally intend to participate to the NOTED [12] R&D effort, setting up a mechanism to tailor network paths based on the load as measured by high level services such as 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. Storage Orchestration Service

The orchestration service plays a central role in the architecture of the datalake, as explained in the previous section. We agreed to use Rucio [13] as a reference implementation of the orchestration service in ESCAPE. The Rucio architecture and the main components are shown in Fig.2. Rucio is an open-source data management system for scientific computing. It has been initially developed in the scope of the ATLAS [14] experiment and adopted by other HEP experiments such as CMS [15] and DUNE [16]. It was positively evaluated by many ESCAPE science projects. Rucio is able to manage data at a file granularity level and to organize them in collections such as archives, datasets and containers. “File” and “collection” definitions can be enriched with some user-defined metadata. The physical management of data in Rucio is driven by “rules”. A rule is a user-defined policy declaring the expectations in terms of data replication for a particular file or dataset: number of replicas and geolocation are examples of rule attributes. Rucio leverages FTS for file transfers and the GFAL [17] library for file deletion. Both replication and deletion rely on the gridFTP/xrootd/HTTP protocols, consistent therefore with the plans described in the previous two sections. In addition, Rucio offers an extensive set of monitoring tools. Information are collected in a BigData infrastructure and can be used to generate customizable dashboards as well as to perform detailed performance studies and operational debugging. In ESCAPE, we intend to rely on the CERN Agile Infrastructure [18] to store, organize and expose such monitoring data.

While Rucio offers the core functionalities necessary for an ESCAPE orchestration service, R&D is required, as well as the commissioning and improvement of several components, to meet the needs of different science projects. Rucio allows storing and retrieving *some* metadata, whereas the full support for user-defined metadata is at a prototype stage. In ESCAPE, we intend to evaluate the current prototype and improve it to accommodate the use cases of the ESCAPE ESFRIs, or to propose an alternative implementation if needed. The concept of QoS in storage, described in section 3.1, needs to be integrated into the orchestration service. Therefore, Rucio needs to evolve to acquire the knowledge of QoS classes and consider them as part of the storage system attributes when defining new storage endpoints. Special attention needs to be paid in transitions between classes as Rucio will need to handle them. Finally, Rucio needs to support the transition between X509 authentication and token-based authentication as will be explained in more details in section 3.5. An important development contribution in this area is being carried out in the context of the XDC [19] European funded project, and WP2 must complement this effort at the level of integration and commissioning. Task 2.3 in ESCAPE focuses in data orchestration and therefore will in general contribute to the aforementioned areas of development and integration.

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.

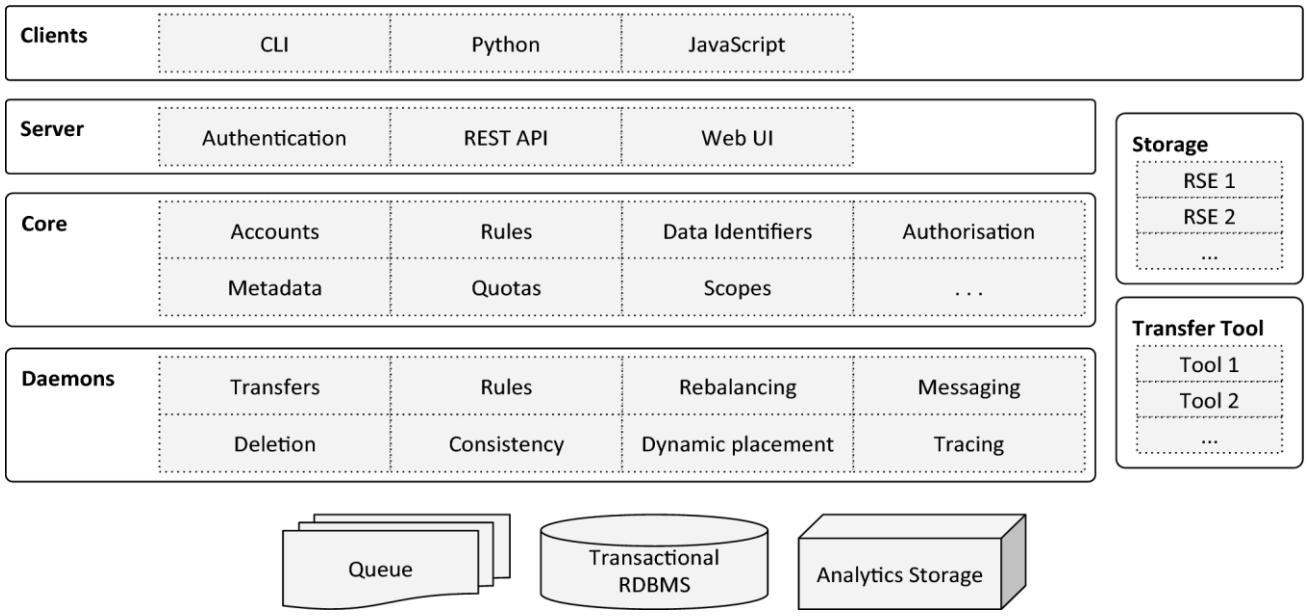

Figure 2: diagram of the Rucio functional elements.

3.4. Content Delivery and Caching

In the datalake model, the processing resources might not be necessarily co-located with the data. Different models can be foreseen to deliver data content to processing units. In fact, from the discussion with the ESCAPE science projects it was clear that we need to support two scenarios, as detailed in the following.

In Fig.3 (left), the application running in a batch cluster or interactively needs to access remote data. The protocol connects to a content delivery layer, which redirects to a storage location containing the file. The content delivery layer fetches the data and streams it to the client. The ability of the content delivery layer to buffer data allows the client to process the data as if it was local, provided that the application is CPU bound and not I/O bound. The latency due to the network distance will be perceived by the client only at the beginning of the process, when the first part of the data needs to be buffered from the storage to the content delivery. Afterwards, the latency is not perceived by the application.

In Fig.3 (right), the application running in a batch cluster or interactively tries to download data from the content delivery layer. Such system, enabled with some storage capacity, serves the data directly to the client. In case the data is not present, the content delivery layer will fetch it from the storage location containing the file, cache it locally, and serve it to the client. The content of the cache will be kept for a period of time depending on the cache size and the number of requests. The cache system will delete the data depending on a set of conditions, for example the least-recently used data first.

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.

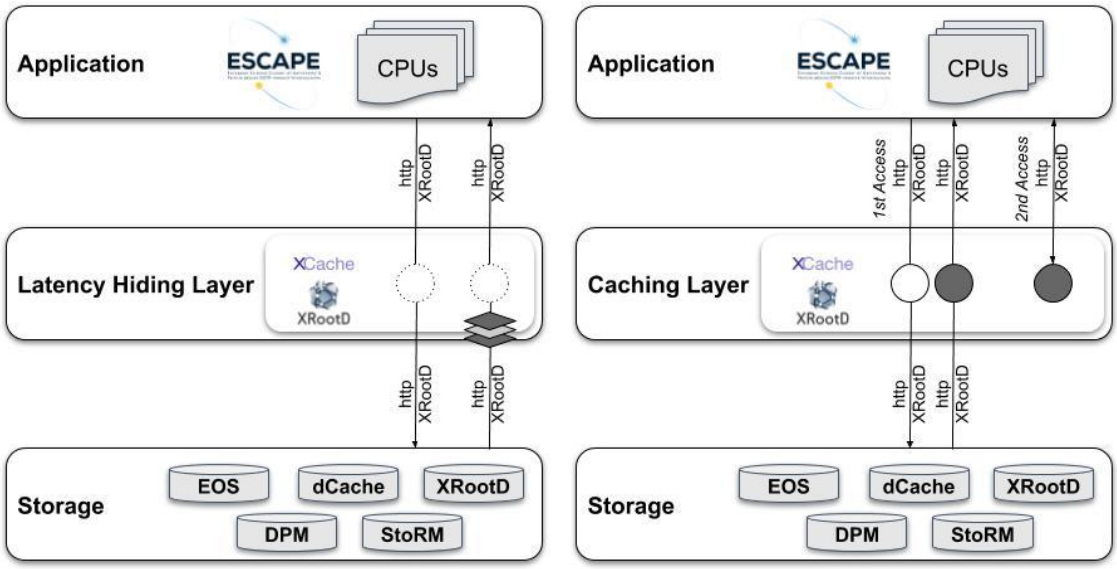

Figure 3: diagrams illustrating the latency hiding (left) and caching (right) use cases.

In both scenarios, data need to be accessible both via the xrootd and HTTP protocols. Xrootd proves to be particularly performant in accessing files in ROOT-compatible format and it is broadly adopted in HEP, particularly for streaming files. On the other hand, HTTP is the most widely-adopted mechanism of downloading data among the ESCAPE science projects. Authentication and authorization aspects are particularly important: the cache needs to communicate with the storage and therefore be authorized by it. At the same time, the cache should authenticate the client and serve only the data that it is entitled to read in order to prevent the cache being used as a bypass to access data under embargo. The caching and latency hiding solution therefore needs to support X509 for legacy reasons, as well as token-based authentication.

We identified the xCache technology [5] as the most promising option fulfilling the requirements of ESCAPE. We will therefore adopt it as reference implementation for the content delivery layer in the datalake. Some studies and R&D will be needed in this area. Both scenarios in Fig.3 should be tested with real applications from the ESCAPE science projects. In particular, the impact on the application efficiency should be measured with respect to the scenario of local data access. The xCache service itself needs to be instrumented with a dedicated service monitoring. Moreover, we intend to integrate this with the MonIT [18] platform to aggregate and organize such information. As it is intended to be a lightweight storage solution, the operational effort to deploy and maintain xCache will be evaluated by the sites and reported in the following ESCAPE deliverables.

3.5. Authentication, Authorization and Identity services

A common and flexible Authentication and Authorization Infrastructure (AAI) is a key requirement to enable secure and controlled data access in the datalake. The ESCAPE project will not invent new authentication and authorization mechanisms but will build upon existing work, leveraging the 15-

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.

years experience of WLCG in building a global AAI, and the recent results of the INDIGO-Datacloud [20] and AARC [21] projects.

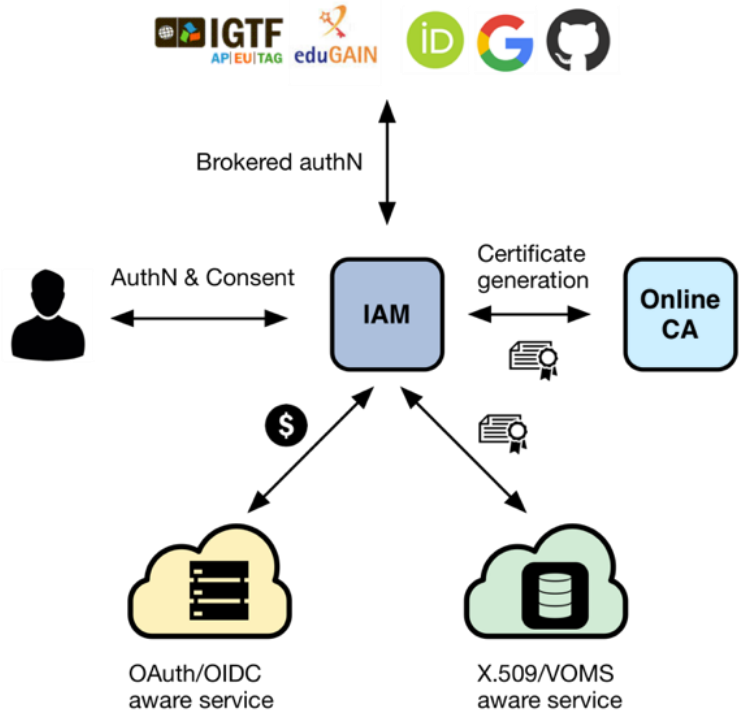

Figure 4: the Indigo Identity and Access Management service functional elements

The INDIGO Identity and Access Management (IAM) [19] service (see Fig.4) will be the central enabling technology for the ESCAPE AAI. IAM is an integrated identity solution developed at INFN as the evolution of VOMS [22] and has been selected as the building block for the next-generation AAI for WLCG. In this next-generation AAI, services expose functionality through OAuth [23] protected APIs: only agents presenting a valid and trusted access token are granted access. Access tokens, which are signed Json Web Tokens (JWTs), can be obtained by client applications (browsers, command line interfaces or other services) from the central IAM service. Access tokens, depending on the IAM configuration and community requirements, can provide identity information (e.g., an opaque user identifier, groups, and other attributes) and other authorization information (e.g., capabilities). Authorization is then performed at services based on the token contents, after a token verification step that assesses token integrity and validity. Some services may require the exchange of the science project issued access token with a local, service-specific one issued by the service itself that is then used to drive authorization decisions. In both scenarios, the agent is authorized based on the information asserted by the central IAM service in the initial access token.
Legacy services are integrated via token translation, i.e. without requiring changes to their codebase and authentication/authorization logic.
In ESCAPE, IAM will act as a community-scoped authorization service responsible for:

- user authentication, supporting multiple authentication mechanisms (identity federations like EduGAIN [24], X509 certificates, social logins)

- providing registered users with a persistent, community-scoped identifier that can be used across the infrastructure for traceability and accounting purposes
- acting as a trusted source for identity and authorization information (e.g., attributes and capabilities) and exposing this information to relying services via standard protocols (i.e. OAuth)
- integrating with an online Certification Authority to provide X509 certificates on-demand in support of integration use-cases and legacy services that require X509 authentication

IAM will also provide a registration service that implements a moderated enrolment flow similar to the one used in production by WLCG, with support for periodic Acceptable Usage Policy enforcement.

3.6. Event-driven data management and data processing

In order to implement rule-based, event-driven data handling throughout the system, we propose a multi-level architecture with interchangeable components designed to maximize the flexibility and to allow changes to the architecture if required. The application of rule sets to storage systems is already realized as a feature of the widely-used Rucio data management system. This will, consequently, be used to apply static, time-independent rules. At Rucio's side, the Data Lifecycle Orchestrator DLCO (a new component that will need to be developed) is intended to apply time-dependent rules. Examples of those would trigger on files residing on a given storage system for more than a specified amount of time, older than a given amount of time or used throughout a specified time span.

The actions determined by those two systems are then made available to the proposed Storage Actions Translation Service, a service that can translate between the requests for state modifications coming in from the rule engines and the various storage systems in the backend. For the execution of the actual file transfers in the system, the architecture uses another widely-used and production-proven system, namely FTS.

3.7. Information and configuration system

The ESCAPE datalake integrates many distributed services fulfilling different roles and those services need to interoperate. The topology of the services and the information about how to access their interfaces needs to be stored in an information system and exposed through a well-defined API. For this functional element we agreed to use the CRIC [25] technology as reference implementation. CRIC is a central catalogue containing services information. A core module aggregates generic information about storage and processing services. This information can be complemented and decorated with community-specific configuration through a set of plugins. A REST API exposes the information to users and other services and also the changing of some of the parameters and creation of new structures. CRIC has been adopted as the WLCG information system and serves as configuration system for Rucio.

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.

3.8. Monitoring

Logs of the services running on the infrastructure will be collected, stored and visualized using industry standard solutions like the Elastic Stack [26]. Events being produced by the Rucio Server will also be stored on the same solution. For Third Party Transfers in particular we will also use existing FTS monitoring dashboards to enrich our view.

The caching layer will be integrated with dedicated monitoring relying on different metrics. In particular, the efficiency of the data transfer as a function of the number of client requests is one of the major aspects to be taken into account for the quality of the global service. Moreover, an effort to provide the MonIT platform with this monitoring information will be carried out for an exhaustive and complete global monitoring bookkeeping.

The combination of those solutions will allow us to have a real-time view of the performance and load of the datalake as well as spot and investigate potential issues or downtime.

3.9. Functional and Stress Testing

We intend to use the Hammercloud [27] system for measurements of the datalake performance in terms of content delivery. Hammercloud is a framework for testing workflows in a distributed computing environment. It allows definition of testing templates and their submission to Grid sites, clouds infrastructures or local batch resources. The management of the jobs is handled by different possible backends depending on the experiment providing the application. Hammercloud has been used successfully by the CERN experiments for many years and it can easily integrate workflows from different science projects. Hammercloud relies on the CERN MonIT infrastructure, mentioned above, to collect metrics and can therefore be flexibly adapted to the needs of different science projects. A diagram of the Hammercloud functional elements con be found in Fig.5.

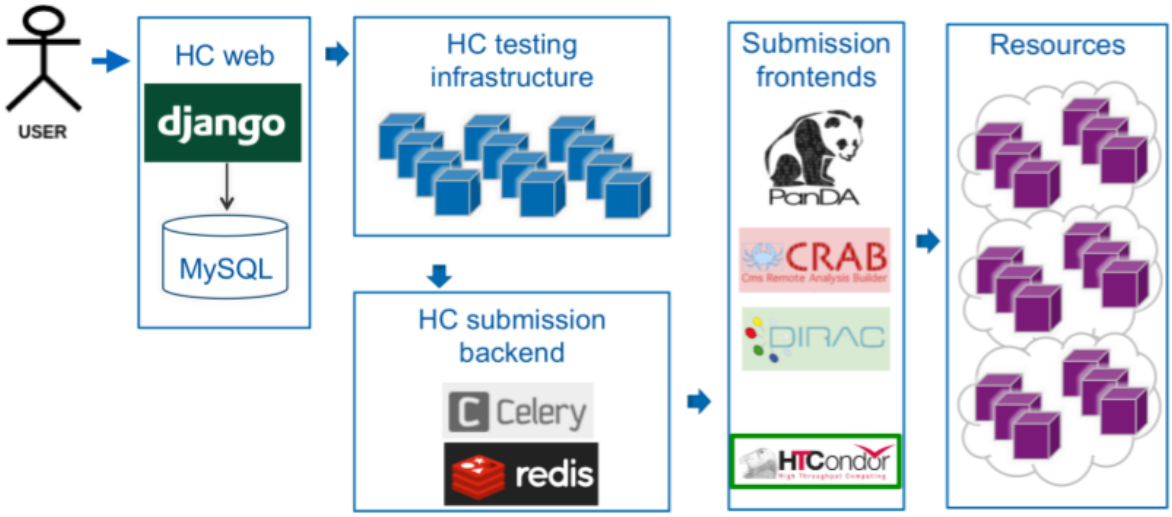

Figure 5: diagram of the Hammercloud functional elements

4. Datalake integration and deployment

The WP2 implementation workplan consists of two phases, as defined in the ESCAPE proposal. The pilot phase, lasting 18 months will focus on demonstrating the data lake model, by providing a small scale but functional system, integrating the technologies identified in the preparation phase. A diagram explaining the elements of the data lake and the agreed reference implementations can be found in Fig.6. The prototype phase will focus on deploying a full-scale system, allowing functional tests as well as stress tests of all the capabilities needed by the ESCAPE science projects for FAIR data management. The next sections will describe in more details the workplan of each of the two phases. A tabular summary and a diagram of the different steps toward the completion of the pilot phase (M24) and prototype phase (M42) can be found in Appendix A.

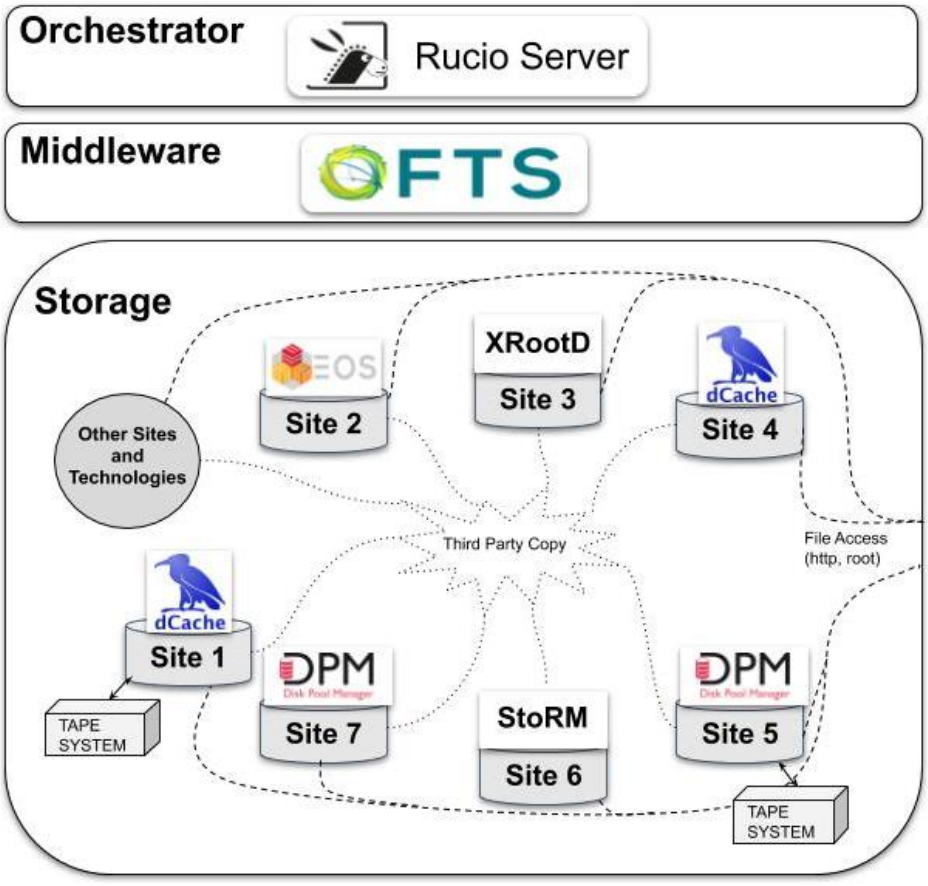

Figure 6: functional elements and reference implementation of the datalake components

4.1. Pilot phase

The pilot phase of WP2 starts at Project Month 9 and concludes at Project Month 24, with the Deliverable 2.2 assessing the performance of the pilot data lake implementation. We will base the datalake pilot on an ecosystem of services providing the ability to federate and harness the different

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.

storage systems at the ESCAPE WP2 partner sites. Such storages are based on different technologies such as dCache, DPM, EOS, STORM and xrootd.

The pilot will demonstrate that a common storage orchestration service can be used across different storage technologies. We will therefore require at least three participating sites, with different storage solutions to provide a dedicated ESCAPE area at the level of a few Terabytes. We will set up an ESCAPE-specific Rucio instance and attach the aforementioned storage services as Rucio endpoints, thanks to the CRIC configuration system.

The authentication and authorization in the pilot will be still based on X509 certificates and proxies, therefore we will create an ESCAPE Virtual Organization and provide a VOMS server for that VO. At the same time however we will prepare for token-based authentication. We will therefore provide also an IAM ESCAPE instance and in the timescale of the pilot also demonstrate token-based storage access for at least one of the participating partner sites.

The pilot will demonstrate the capability to organize and distribute data across the sites, based on high level policies. For this to happen, we will provide an FTS server connecting the different sites and accepting transfers for the ESCAPE VO and initiated by Rucio. We will commission the different transfer links based on the parameters specific to each ESFRI, such as average file sizes and number of files. We also intend to demonstrate the possibility to define different storage QoS for the same endpoint and integrate such definition in Rucio for data organization. We will therefore implement such functionality for at least one of the storages in the pilot.

In terms of data access, the plan is to conclude the pilot phase mastering the xCache technology both in terms of service and integration with the ESFRIs workflows. For this to happen, we will initially focus on installation and configuration of xCache at one site. We will instrument the needed monitoring to measure load, cache reuse and I/O activity. We intend to conclude the pilot phase by providing a software suite able to measure the impact of the cache for the various science projects in ESCAPE.

Finally, several monitoring components need to be implemented or deployed to have an overview of the activity in the pilot. We intend to deploy perfSONAR in the participating sites to the pilot and define a simple dashboard exposing the results of latency and performance network measurements. We will also adopt the FTS transfer monitoring dashboard and provide an ESCAPE-specific view. We will finally provide a first prototype of the datalake dashboard based on the CERN MonIT infrastructure. We will ensure that the information is exposed through a REST API to be consumed by upstream services.

4.2. Prototype phase

The prototype phase will generally focus on integrating more partner sites into the datalake, consolidate the functionalities that were pioneered in the pilot phase and scale up the performance of the different datalake services. By the end of this phase it should be demonstrated that the datalake ecosystem suits the needs of the ESFRI partner projects, in terms of functionality, performance and usability.

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 prototype will demonstrate that a cloud-like storage federation, composed of different institutes and different storage providers, can be orchestrated. The aim is to have sizeable storage resources (100TB) from 5 partner institutes in the project leveraging more than three storage technologies. Data injection and distribution through FTS and Rucio should be enabled at the Gigabyte per second scale across sites, targeting a Terabyte per day of net data flow from production to distribution. We should demonstrate the ability to provide at least two different Quality of Service at the orchestrator level and to transition across different Quality of Service layouts.

We will provide a token-based AAI infrastructure to be used for general data movement. This implies integrating token support across the main components: the storage systems, the storage orchestrator and the File Transfer Service. The membership database and updating mechanisms for new members will be in place.

By the end of the prototype we will provide a common set of monitoring tools offering a high level view of the status (e.g. storage availability, online files), usage (e.g. number of file and capacity) and activity (e.g. number of data transfers and number of clients accessing data).

We will define representative workloads for at least two of the partner projects and integrate them in Hammercloud. Those workloads will be used to assess the impact of latency and bandwidth in remote data access, when compute resources outside the boundary of the datalake are used. We plan to integrate, as a demonstrator, one HPC facility for data processing. Hammercloud will allow us to run these workloads periodically and collect statistics about data access performance. In particular, we intend to measure the impact of network latency with and without the use of xCache and the performance of different QoS classes.

Finally, by the end of the prototype we intend to have a first demonstration of event-based data processing. At least one of the storage endpoints would provide file-level information through a messaging system, such as new files being created or existing files being lost.

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.

5. Conclusions

In the first 8 months of the project we went through the process of collecting requirements from the involved science projects. We also went through a series of technical meetings highlighting the available technologies that could be candidates building blocks of the ESCAPE datalake. We defined an architecture of such system and elaborated an implementation plan that would deliver in steps of increasing complexity all functionalities needed by our science communities. The milestones for this plan would be the ones described in the ESCAPE proposal document, together with the next deliverables at the end of the pilot and prototype phase.

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.

6. References

[1] http://dcache.org
[2] http://lcgdm.web.cern.ch/dpm
[3] https://eos.web.cern.ch/
[4] https://italiangrid.github.io/storm/index.html
[5] http://xrootd.og
[6] http://wlcg.web.cern.ch/
[7] https://root.cern.ch/
[8] https://www.ogf.org/documents/GFD.47.pdf
[9] https://sdm.lbl.gov/srm-wg/
[10] https://fts.web.cern.ch/
[11] https://www.perfsonar.net/
[12] http://cern.ch/go/lkr7
[13] https://rucio.cern.ch/
[14] http://atlas.cern
[15] http://cms.cern
[16] https://www.dunescience.org
[17] https://dmc.web.cern.ch/projects/gfal-2/home
[18] https://doi.org/10.1051/epjconf/201921408031
[19] http://www.extreme-datacloud.eu
[20] https://www.indigo-datacloud.eu
[21] https://aarc-project.eu
[22] https://italiangrid.github.io/voms/
[23] https://oauth.net/2/
[24] https://edugain.org
[25] http://cms-cric.cern.ch
[26] https://www.elastic.co
[27] http://hammercloud.cern.ch/hc/

Appendix A

Due date/ Theme	AAI	Storage Orchestration	Networks and asynchronous data transfer	Content delivery and Caching	Configuration, Monitoring, and Accounting
Month 8	**Project planning and hiring stage complete**				
Month 12	- X509 authentication enabled on all partner sites	- Enable ESCAPE VO on partner storage sites - Test RUCIO instance federating storage from 3 partner institutes with different technologies	- perfSONAR deployed at three partner storage sites	- Installation and configuration of XCache at one site	- CRIC information system deployed in ESCAPE datalake
Month 18	- Enable IAM on ESCAPE rucio instance	- ESCAPE Rucio instance deployed. - Ability to define simple file QoS rules	- Network health dashboard developed for ESCAPE	- Instrumentation of XCache monitoring	- Adoption of FTS monitoring dashboards for ESCAPE datalake
Month 24	- Demonstration of token-based storage access	- FTS instance ready for bulk data transfers among the pilot partner sites	- Data Transfer and access commissioning	- Installable software suite measuring the impact of a cache	- Prototype of ESCAPE datalake dashboard
Month 24	**Pilot phase testing complete, report delivered**				
Month 30		- Quality of Service parameter development and tuning for Reliability, Performance and Cost	- Third party transfers enabled		- Instrument workload testing on the ESCAPE datalake (Hammercloud)
Month 36 Data lake development close to complete to enable final assessment phase	- Token based data management deployed on ESCAPE datalake	- Event-driven data management tested	- Network route optimisation for intelligent transfers	- Interactions with science analysis methods within ESCAPE WP5	- Final ESCAPE datalake dashboard
Month 40	**Final deliverable assessing data lake performance in all themes.**				

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.


Month 8
Month 12
Month 18
Month 24
Month 30
Month 36
Month 40
Pilot Phase
Full ESCAPE prototype phase
AAI
Storage Orchestration
Networks and asynchronous data transfer
Content delivery and caching
Configuration, Monitoring and Accounting
Project planning and hiring stage complete
X509 authentication enabled on all partner sites
Enable IAM on ESCAPE rucio instance
Demonstration of token-based storage access
Token-based data management deployed ob ESCAPE datalake
Enable ESCAPE VO on partner storage sites
Test RUCIO instance federating storage from 3 partner institutes with different technolodies
ESCAPE Rucio instance deployed. Ability to define simple file QoS rules
FTS instance ready for bulk data transfers among the pilot partner sites
Quality of Service parameter development and tuning for Reliability, Performance and Cost
Event-driven data management tested
perfSONAR deployed at three partner storage sites
Network health dashboard developed for ESCAPE
Data Transfer and access commissioning
Third party transfers enabled
Network route optimisation for intelligent transfers
CRIC information system deployed in ESCAPE datalake
Instrumentation of XCache monitoring
Installable SW suite measuring the impact of a cache
Interactions with science analysis methods within ESCAPE WP5
Installation and configuration of Cache at one site
Adoption of FTS monitoring dashboard for ESCAPE datalake
Prototype of ESCAPE datalake dashboard
Instrument workload testing on the ESCAPE datalake (Hammercloud)
Final ESCAPE data lake dashboard
Pilot phase testing complete, report delivered
Final deliverable assessing data lake performance in all themes.