

ESCAPE WP2 Data Infrastructure for Open Science Milestone 10 (M24)

MS 10: Expanded prototype, more data centres including 3rd party centres, demonstrate integrated data management tools, verify RI data accessibility from compute platforms including commercial clouds.

The ESCAPE Data Lake infrastructure is currently integrating ten storage endpoints provided by the ESCAPE partner institutes across Europe, these are: INFN-CNAF, INFN-ROMA, INFN-Napoli, DESY, SURF-SARA, IN2P3-CC, CERN, IFAE-PIC, LAPP and GSI. This is the first implementation of the conceptual design of building a common storage and data management infrastructure for open science involving different scientific communities.

A 24h Full Dress Rehearsal Exercise was organized on the 17th of November 2020. This allowed us to perform a thorough assessment about the status of the Data Management tools integration with the RIs. This exercise consisted in running multiple experiment data workflows on a single day, focused on data injection, data replication and data access.

During that day the experiments made use of 12 different storage endpoints across the Data Lake and transferred a million files while replicating real production workflows, in some cases ending up with end-user data analysis. The RIs involved were: LOFAR, CTA, MAGIC, FAIR, SKA, LSST, ATLAS, CMS and EGO/VIRGO.

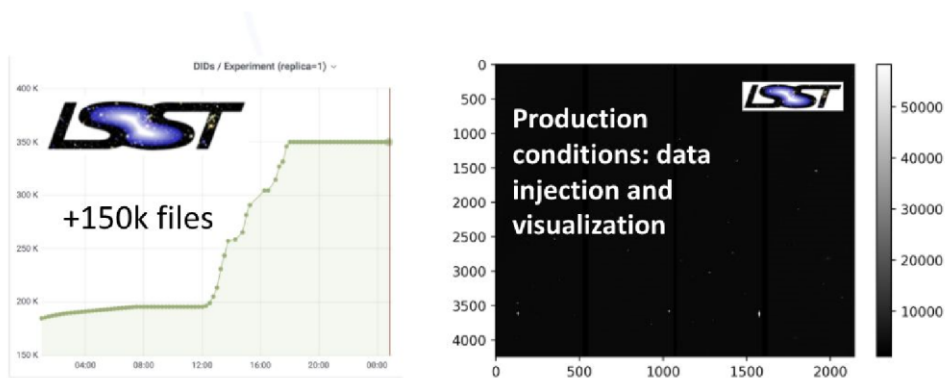
Incorporation of third party centers was proven. As an example, this is the case of the MAGIC telescope where an external storage was integrated in the Data Lake. This external storage is the telescope's data acquisition (DAQ) buffer space, in charge of recording data from the telescope. After the data is store locally at the telescope, then the Data Lake Data Management tools take care of injecting the data in the Data Lake and once the data is consolidated, and with the adequate level of replication, it gets deleted from the DAQ buffer to free space and have room for new data acquisition (fig.1) [1]



MAGIC: Mimics a real MAGIC observation use case. Remote storage (Data Lake aware) **next to the telescope** acts as a buffer for subsequent data injection to the ESCAPE Data Lake (and local deletion after success)

One example to illustrate data accessibility is the LSST experiment, leveraging data injection to the Data Lake and the posterior data access from a compute platform. In this particular case, the data processing platform was based on a Jupyter Notebook infrastructure in CC-

IN2P3. This workflow was mimicking real running conditions for data received from the telescope source (fig.2)



LSST: Simulate production conditions: ingest the HSC RC2 dataset from CC-IN2P3 local storage to the Data Lake, **at a realistic LSST data rate (20TB/24h)**. Then **confirm integrity and accessibility of the data via a notebook**.

→ The image is a reconstruction drawn within a Jupyter Notebook accessing the data used in the Full Dress Rehearsal.

We are currently in the process to commission commercial cloud resources; the procurement phase for a proof of concept started and in the next months the integration of the ESCAPE Data Lake should be finished and ready for a small scale test involving one or more RIs.

[1] Prototype of MAGIC data at PIC: Development and application of an automatic workflow for the replication of scientific data <https://projectescape.eu/news/escape-dios-development-and-application-automatic-workflow-replication-scientific-data>

[2] Initial pilot data lake with at least 3 core data centres <https://projectescape.eu/deliverables-and-reports/milestone-8-initial-pilot-data-lake-least-3-core-data-centres>