International Symposium on Grids & Clouds (ISGC) 2022 Virtual Conference

Contribution ID: **44** Type: **Oral Presentation**

Data Lake as a Service for Open Science

Friday, 25 March 2022 11:00 (30 minutes)

Experiments and scientists, whether in the process of designing and building up a data management system or managing multi-petabyte data historically, gather in the European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures (ESCAPE) project to address computing challenges by developing common solutions in the context of the EOSC. A modular ecosystem of services and tools constitutes the ESCAPE Data Lake, which is exploited by flagship ESFRIs in Astroparticle Physics, Electromagnetic and Gravitational-Wave Astronomy, Particle Physics, and Nuclear Physics to pursue together the FAIR and open-access data principles.

The aforementioned infrastructure fulfils the needs of the ESCAPE community in terms of data organisation, management, and access, and addresses the required functionalities and experiment-specific use cases. As a matter of fact, dedicated assessment exercises during specific testing-focused time windows - the 2020 Full Dress Rehearsal (FDR) and the 2021 Data and Analysis Challenge (DAC) exercises - demonstrated the robustness of the pilot and prototype phases of the various Data Lake components, tools, and services, providing scientists with know-how on their management and utilisation, and evening out differences in knowledge among ESCAPE partners. A variety of challenges and specific use cases boosted ESCAPE to carefully take into account both user and infrastructure perspectives, and contributed to successfully concluding the pilot and prototype phases beyond expectations, embarking on the last stage of the project. As a result, collaborating sciences are choosing their reference implementations of the various technologies among the proposed solutions.

The prototype phase of the project aimed at consolidating the functionalities of the services, e.g. integrating token-based AuthN/Z or deploying a tailored content delivery and caching layer, and at simplifying the user experience. In this respect, a considerable effort has been devoted towards a product named DataLake-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, with the goal to provide the end-user with a Notebook ready-to-be-used and fully Data Lake aware. The project was framed within the CERN-HSF Google Summer of Code (GSoC) programme in 2020, and under the CERN IT-OpenLab programme in 2021. The development of a state-of-the-art “data and analysis portal” as an interface to the Data Lake offers a wide range of possibilities, from very simple I/O access to more complex workflows such as enabling content delivery technologies and integration local storage facilities at the facility hosting the Notebook. The DLaaS project allows end-users to interact with the Data Lake in an easily-understandable and user-friendly way, and it is based on the JupyterLab and JupyterHub software packages. The Rucio-JupyterLab software package that was developed during GSoC2020 is used to integrate the service with the ESCAPE Rucio instance, and the DLaaS is deployed on the same cluster hosting the other Data Lake services and tools. Examples of the 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 content delivery low latency-hiding data access layer based on XRootD-XCache.

ESCAPE milestones achieved during the length of the project represent a fundamental accomplishment under both sociological and computing model aspects for different scientific communities that should address upcoming data management and computing challenges in the next decade.

Primary authors: HILMY, Muhammad Aditya (Institut Teknologi Bandung); DI MARIA, Riccardo (CERN)

Presenter: DI MARIA, Riccardo (CERN)

Session Classification: Data Management & Big Data

Track Classification: Track 6: Data Management & Big Data