Contribution ID: **225** Type: **Presentation**

Onedata Distributed Storage for HPC Galaxy Workflows

Friday 21 March 2025 11:45 (15 minutes)

Onedata [1] is a high-performance, distributed data management system designed for global infrastructures. It provides seamless access to heterogeneous storage resources and supports diverse use cases ranging from personal data management to large-scale scientific computations. Leveraging a fully distributed architecture, Onedata facilitates the creation of hybrid cloud environments that integrate private and public cloud resources. The system enables users to collaborate, share, and publish data while supporting high-performance computations on distributed datasets via various interfaces, including POSIX-compliant native mounts, pyfs (Python filesystem) plugins, REST/CDMI APIs, and an S3 protocol (currently in beta).

Recent advancements in Onedata include the development of the *fs.onedatarestfs* Python library, a lightweight *pyfilesystem* client built upon the *OnedataFileRESTClient* library. Within the scope of the EuroScienceGateway project [2], these libraries have been instrumental in integrating Onedata with the Galaxy Project [3] , an open-source platform for data analysis workflows predominantly used in the life sciences. This integration has resulted in a new File Source Plugin and an Object Store for Galaxy. The File Source Plugin enables users to import and export datasets between Onedata and Galaxy, while the Object Store integration allows Onedata to function as a backend storage system for Galaxy datasets. This implementation takes advantage of Onedata' s distributed architecture, creating a synergy with Galaxy's distributed network of Pulsar endpoints (workflow execution services). By tracking data distribution, it opens the door to locality-aware, smart workflow scheduling, which can reduce data transfer costs, processing delays, and energy usage.

Onedata is currently deployed in several European projects, including EUreka3D [4], EuroScienceGateway [2], DOME [5], and InterTwin [6]. In these projects, Onedata provides a data transparency layer for managing large, distributed datasets in dynamic, hybrid cloud environments with containerized deployments.

Acknowledgements. This work is co-financed by the Polish Ministry of Education and Science under the program entitled International Co-financed Projects (projects no. 5398/DIGITAL/2023/2 and 5399/DIGITAL/2023/2)

References:

1. Onedata. https://onedata.org.
2. EuroScienceGateway Project: Open Infrastructure for Data-Driven Research. https://galaxyproject.org/projects/esg/.
3. The Galaxy Project. https://galaxyproject.org/.
4. EUreka3D: European Union's REKonstructed in 3D. https://eureka3d.eu.
5. DOME: A Distributed Open Marketplace for Europe Cloud and Edge Services. https://dome-marketplace.eu.
6. InterTwin: Interdisciplinary Digital Twin Engine for Science. https://intertwin.eu.

Authors: Dr ORZECHOWSKI, Michał (ACK Cyfronet AGH); Mr OPIOŁA, Łukasz (ACK Cyfronet AGH); Dr DUTKA, Łukasz (ACK Cyfronet AGH)

Presenter: Dr ORZECHOWSKI, Michał (ACK Cyfronet AGH)

Session Classification: HPC data access and integration

Track Classification: Main sessions: Scalable Storage Backends and Integration with Data Processing