Contribution ID: **91** Type: **Demonstration/Tutorial (30 mins)**

Distributed Dataset Lifecycle Management with Onedata

Onedata [1] is a high-performance data management system with a distributed, global infrastructure that enables users to access storage resources worldwide. It supports various use cases ranging from personal data management to data-intensive scientific computations. Onedata has a fully distributed architecture that facilitates the creation of a hybrid-cloud infrastructure with private and commercial cloud resources. Users can collaborate, share, and publish data, as well as perform high-performance computations on distributed data using POSIX-compliant data access applications.

The latest Onedata release version, 21.02.1, introduces several new features and improvements that enhance its capabilities in managing distributed datasets throughout their lifecycle. The software allows users to establish a hierarchical structure of datasets, control multi-site replication and distribution using Quality-of-Service rules, and keep track of the dataset size statistics over time. In addition, it also supports the annotation of datasets with metadata, which is crucial for organizing and searching for specific data. The platform also includes robust protection mechanisms that prevent data and metadata modification, ensuring the integrity of the dataset in its final stage of preparation. Another key feature of Onedata is its ability to archive datasets for long-term preservation, enabling organizations to retain critical data for future use. This is especially useful in fields such as scientific research, where datasets are often used for extended periods or cited in academic papers. Finally, Onedata supports data sharing mechanisms aligned with the idea of Open Data, such as the OAI-PMH protocol and the newly introduced Space Marketplace. These features enable users to easily share their datasets with others, either openly or through controlled access.

Currently, Onedata is used in European EGI-ACE [2], PRACE-6IP [3], and FINDR [4] project, where it provides a data transparency layer for managing large, distributed datasets on dynamic hybrid cloud containerised environments.

During the demo, we will showcase the capabilities of Onedata in terms of managing the overall lifecycle of distributed datasets, from preparation, through annotation and dissemination, to archiving for long-term preservation. The demo will be performed on the Onedata services at EGI DataHub, with the intention of easy reproducibility by EGI users.

Keywords: distributed dataset, data lifecycle, data access, distributed systems, file sharing.

Acknowledgments: This work was supported in part by 2018-2020's research funds in the scope of the co-financed international projects framework (project no. 5145/H2020/2020/2).

References:
[1] Onedata project website. https://onedata.org.
[2] EGI-ACE: Advanced Computing for EOSC. https://www.egi.eu/projects/egi-ace/.
[3] Partnership for Advanced Computing in Europe - Sixth Implementation Phase. http://www.prace-ri.eu.
[4] FINDR: Fast and Intuitive Data Retrieval for Earth Observation

Other key topic

Key Topic

Data Spaces

Primary authors: OPIOLA, Lukasz (CYFRONET); ORZECHOWSKI, Michal; KRYZA, Bartosz; DUTKA, Lukasz

Presenter: DUTKA, Lukasz

Session Classification: Demonstrations