Contribution ID: **128** Type: **Demonstration/Tutorial (30 mins)**

Demo: Onedata - workflow processing, distributed datasets and archives

Thursday, 22 June 2023 13:35 (20 minutes)

During the demo, we will showcase the latest features of Onedata, introduced in version 21.02.1.

1. Multi-cloud data processing using automation workflows and archive preservation.
2. 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. Onedata version 21.02.1 is being gradually introduced to DataHub since May.

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 Onedata ecosystem comprises several services, including Onezone, which is the authorisation and distributed metadata management component that provides access to the system. Oneprovider delivers actual data to the users and exposes storage systems to Onedata, while Oneclient enables transparent POSIX-compatible data access on user nodes. Oneprovider instances can be deployed as single nodes or HPC clusters, on top of high-performance parallel storage solutions that can serve petabytes of data with GB/s throughput.

The latest Onedata release version, 21.02.1, introduces the integration of a powerful workflow execution engine, which is powered by OpenFaas [2]. This integration enables the creation of complex data processing pipelines that can leverage the transparent access to distributed data provided by Onedata. The workflow functionality is especially useful for creating a comprehensive, OAIS compliant data archiving and preservation system, which covers all archival requirements, including ingestion, validation, curation, storage, and publication. The workflow function library includes ready-to-use functionalities, which are implemented as Docker images and cover typical archiving actions such as metadata extraction, format conversion, and checksum validation. Additionally, new custom functions can be easily added and shared among user groups. The solution underwent thorough testing on auto-scalable Kubernetes clusters, ensuring its reliability and scalability. Since Onedata is a distributed solution, deploying multiple instances of Oneprovider and OpenFaas on separate clouds provides a transparent data-plane layer for multi-cloud workflow processing.

Moreover, version 21.02.1 introduces several new features and improvements that enhance Onedata's 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 [3], PRACE-6IP [4], and FINDR [5] project, where it provides a data transparency layer for managing large, distributed datasets on dynamic hybrid cloud containerised environments.

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).

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

Presenter: DUTKA, Lukasz (CYFRONET)

Session Classification: Demonstrations