Contribution ID: **90** Type: **Demonstration/Tutorial (30 mins)**

Multi-Cloud Workflow Processing 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 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 [3] 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.

Currently, Onedata is used in European EGI-ACE\cite [4], PRACE-6IP [5], and FINDR [6] project, where it provides a data transparency layer for computation, data processing automation deployed on dynamic hybrid cloud containerised environments.

During the demo, we will showcase the latest features of Onedata, with a special focus on multi-cloud data processing using automation workflows and archive preservation. The demo will be performed on the Onedata services at EGI DataHub, with the intention of easy reproducibility by EGI users.

Keywords: workflows, data preservation, 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] OpenFaaS - Serverless Functions Made Simple. https://www.openfaas.com/.
[3] David Giaretta, CCSDS Group, and CCSDS Panel. Reference model for an Open Archival Information System (OAIS). 06 2012.
[4] EGI-ACE: Advanced Computing for EOSC. https://www.egi.eu/projects/egi-ace/.
[5] Partnership for Advanced Computing in Europe - Sixth Implementation Phase. http://www.prace-ri.eu.
[6] FINDR: Fast and Intuitive Data Retrieval for Earth Observation

Other key topic

Key Topic

Data Spaces

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

Presenter: ORZECHOWSKI, Michal

Session Classification: Demonstrations