Contribution ID: **36** Type: **Presentation**

Using Workflows for Data Preservation Using Onedata

Thursday, 27 January 2022 10:55 (20 minutes)

Onedata [1] is a distributed, global, high-performance data management system, which provides transparent and unified access to globally distributed storage resources and supports a wide range of use cases from personal data management to data-intensive scientific computations. Due to its fully distributed architecture, Onedata allows for creation of complex hybrid-cloud infrastructure deployments, with private and commercial cloud resources. It allows users to share, collaborate and publish data as well as perform high performance computations on distributed data. Onedata allows users to collaborate, share, and perform computations on data using applications relying on POSIX compliant data access.

Onedata comprises the following services: Onezone - authorisation and distributed metadata management component that provides access to Onedata ecosystem; Oneprovider - provides actual data to the users and exposes storage systems to Onedata and Oneclient - which allows transparent POSIX-compatible data access on user nodes. Oneprovider instances can be deployed, as a single node or an HPC cluster, on top of highperformance parallel storage solutions with the ability to serve petabytes of data with GB/s throughput.

Recently, Onedata was enhanced with a powerful workflow execution engine, powered by OpenFaas [2]. It allows for creation of complex data processing pipelines that can leverage the transparent access to distributed data provisioned by Onedata. In particular the workflow functionality can be used to create a comprehensive, OAIS [3] compliant, data archiving and preservation system, covering all archival requirements including ingestion, validation, curation, storage and publication. The workflow function library contains ready to use functionalities (implemented as Docker images), covering typical archiving actions such as metadata extraction, format conversion, checksum validation, virus checks and others. New custom functions can be easily added and shared among user groups. The solution was thoroughly tested running on auto-scalable Kubernetes clusters.

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

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

Primary authors: Mr ORZECHOWSKI, Michal (ACC Cyfronet-AGH); Dr KRYZA, Bartosz (ACC Cyfronet AGH); Dr DUTKA, Lukasz (ACC Cyfronet AGH)

Presenter: Dr DUTKA, Lukasz (ACC Cyfronet AGH)

Session Classification: User Stories

Track Classification: Main session: User Voice: Novel Applications, Data Science Environments & Open Data