Contribution ID: **101** Type: **Presentation**

FaaS Data Processing with Onedata

Wednesday 8 March 2023 14:15 (15 minutes)

Onedata 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 the creation of complex hybrid-cloud infrastructure deployments, with private and commercial cloud resources. It allows users to share, collaborate and publish data and perform high-performance computations on distributed data. Onedata enables users to collaborate, share and perform computations on data using applications relying on POSIX-compliant data access.

Onedata has recently been enhanced with a powerful workflow execution engine powered by OpenFaas. This allows for the creation of complex data processing pipelines that have transparent access to distributed data provisioned by Onedata. The workflow functionality can be used for embedded data processing and includes a library of ready-to-use functionalities such as metadata extraction and format conversion. Custom functions can also be easily added and shared among user groups. The solution has been thoroughly tested on auto-scalable Kubernetes clusters. In addition to the transparent access to distributed data, the use of a Function as a Service (FaaS) platform for data processing offers flexibility and innovation for contemporary data tasks. The use of FaaS allows for the creation of custom, modular functions that can be combined and executed in a variety of ways to meet the specific needs of a given data processing task. This modular approach makes it easy to scale and update individual functions as needed, making FaaS well-suited for the dynamic and constantly evolving nature of modern data processing. The combination of FaaS and Kubernetes further enhances the flexibility and scalability of this data processing solution. By running the FaaS platform on top of Kubernetes, it is possible to easily scale the number of functions being executed and the resources allocated to them based on the needs of the specific task at hand. This allows for efficient and effective use of resources, making it possible to tackle even the most demanding data processing tasks. The use of Kubernetes also enables seamless integration with other tools and technologies, further expanding the capabilities of the FaaS platform. Overall, the application of FaaS on top of Kubernetes makes this data processing solution highly flexible and well-suited for a wide range of contemporary data tasks.

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

Acknowledgements. 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. 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

Authors: KRYZA, Bartosz (ACC Cyfronet-AGH); ORZECHOWSKI, Michał (AGH University of Science and Technology, Academic Computer Centre Cyfronet AGH, Krakow, Poland); DUTKA, Łukasz

Presenter: ORZECHOWSKI, Michał (AGH University of Science and Technology, Academic Computer Centre Cyfronet AGH, Krakow, Poland)

Session Classification: Collaborative Data Science and Visualisation

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