Contribution ID: **28** Type: **Demonstrations & Tutorials**

Running Multi-Cloud Workloads on Distributed Datasets with Onedata

Tuesday, 1 October 2024 18:00 (30 minutes)

Onedata continues to evolve with subsequent releases within the 21.02 line, enhancing its capabilities and solidifying its position as a versatile distributed data management system. Key improvements include the rapid development of the automation workflow engine, the maturation of the S3 interface, and powerful enhancements to the web UI for a smoother user experience and greater control over the distributed data.

Apart from that, a significant focus has been put on enhancing the interoperability of the platform. Onedata can be easily integrated as a back-end storage solution for various scientific tools, data processing and analysis platforms, and domain-specific solutions, providing a unified logical view on otherwise highly distributed datasets. This is achieved thanks to the S3, POSIX, and Pythonic data interfaces and tools that enable effortless inclusion of Onedata as a 3rd party solution in CI/CD pipelines. For example, the "demo mode" makes it straightforward to develop and test arbitrary middleware against a fully functional, zero-configuration Onedata backend. With the ability to integrate with SSO and IAM services and reflect the fine-grained federated VO structures, Onedata can serve as a comprehensive data management solution in federated, multi-cloud, and cross-organizational environments. Currently, it's serving this purpose in the ongoing EuroScienceGateway, EUreka3D, and Dome EU-funded projects.

Automation workflows in Onedata can streamline data processing, transformation, and management tasks by automating repetitive actions and running user-defined logic fitted to their requirements. The integrated automation engine runs containerized jobs on a scalable cluster next to the data provider's storage systems. This allows seamless integration of data management and processing steps, allowing for efficient handling of large-scale datasets across distributed environments.

During our demonstration, we will present a comprehensive use case demonstrating Onedata's capabilities in managing and processing distributed data based on the EGI DataHub environment. It will showcase a pipeline that embraces the user's federated identity and VO entitlements, automated data processing workflows, the wide range of Onedata's tools for data management, and interoperability with scientific tools and middleware — with a special focus on the S3 interface.

Join us for the demo to see how Onedata empowers organizations to manage and process federated and multi-cloud data efficiently, driving collaboration and accelerating scientific discovery.

Topic

Data innovations: Data Management/Integration/Exchange

Primary authors: DUTKA, Lukasz (CYFRONET); OPIOLA, Lukasz (CYFRONET)

Co-authors: KRYZA, Bartosz (CYFRONET); ORZECHOWSKI, Michal (CYFRONET)

Presenter: OPIOLA, Lukasz (CYFRONET)

Session Classification: Demonstrations & Posters