



Contribution ID: 41

Type: **Presentation**

## Onedata - Managing Data and Metadata in Hybrid Clouds

*Wednesday, 30 January 2019 15:10 (20 minutes)*

Onedata [1] is a transparent, high-performance data management system, which provides transparent 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 enables the creation of complex hybrid-cloud infrastructure deployments, including 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 comprises the following components: Onezone, distributed metadata management and authorisation component

that provides an entry point for users; Oneprovider, which is the main data management component providing the transparent virtual filesystem over distributed heterogeneous storage resources; and Oneclient, which provides a virtual POSIX file system mountpoint on user worker nodes.

Onedata introduces the concept of Space, a virtual volume, owned by one or more users, where they can organize their data under a global namespace. The Spaces are accessible to users via an intuitive web interface, Fuse-based client providing POSIX file system, as well as REST and CDMI standard APIs. Each Space can be supported by a dedicated amount of storage supplied by one or multiple storage providers. Storage providers deploy Oneprovider instance near the storage resources, register it in the selected Onezone instance to become part of a federation and expose those resources to users. By supporting multiple types of storage backends, such as such as POSIX, S3, Ceph, WebDAV, dCache and OpenStack Swift, Onedata can serve as a unified virtual file system for hybrid-cloud environments. Using a comprehensive Onedata REST API it is possible to automatically extract metadata from exposed files and ingest them into Onedata. The data and metadata managed by Onedata are synchronised with any changes made to data directly on the underlying storage. In order to enable an easy way to expose existing data collection, dedicated deployment procedure called Onedatify is available, which provides a command-line wizard that guides a user through \op deployment procedure and exposes existing data from the specified legacy storage.

Currently, Onedata is used in Helix Nebula Science Cloud [2], eXtreme DataCloud [3], PLGrid [4], and European Open Science Cloud Pilot [5], where it provides data transparency layer for computation deployed on hybrid-clouds. Furthermore, in European Open Science Cloud Hub [6] it also serves as the basis of EGI Open Data Platform, supporting various open science use cases such as open data curation (metadata editing), publishing (DOI and PID registration) and discovery (OAI-PMH protocol).

[1] Onedata project website. <http://onedata.org>.

[2] Helix Nebula Science Cloud (Europe's Leading Public-Private Partnership for Cloud). <http://www.helix-nebula.eu>.

[3] eXtreme DataCloud (Developing scalable technologies for federating storage resources). <http://www.extreme-datacloud.eu>.

[4] PL-Grid (Polish Infrastructure for Supporting Computational Science in the European Research Space). <http://projekt.plgrid.pl/en>.

[5] European Open Science Cloud Pilot (The first phase in the development of the European Open Science-Cloud). <https://eoscpilot.eu>.

[6] European Open Science Cloud Hub (Bringing together multiple service providers to create a single contact point for European researchers and innovators.). <https://www.eosc-hub.eu>.

**Primary authors:** Mr ORZECHOWSKI, Michal (AGH University of Science and Technology, Academic Computer Centre Cyfronet AGH, Krakow, Poland); Dr KRYZA, Bartosz (ACC Cyfronet-AGH); Dr DUTKA, Lukasz (ACC Cyfronet-AGH)

**Presenter:** Dr DUTKA, Lukasz (ACC Cyfronet-AGH)

**Session Classification:** Scalable Storage Backends for Cloud, HPC and Global Science

**Track Classification:** Scalable Storage Backends for Cloud, HPC and Global Science