# Using Onedata for data caching in hybrid-cloud environments

Michał Orzechowski[1,2],Bartosz Kryza[1], Łukasz Dutka[1],

AGH University of Science and Technology, Academic Computer Centre Cyfronet AGH, Krakow, Poland[1]
AGH University of Science and Technology, Department of Computer Science, Krakow, Poland[2]

**Keywords:** data access, distributed system, file sharing, open data, metadata, replication

**Type:** Presentation

Onedata [1] is a global high-performance data management system, that provides easy and unified access to globally distributed storage resources and supports wide range of use cases from personal data management to data-intensive scientific computations. Onedata enables creation of complex hybrid-cloud deployments, using 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 system consists of zones (*Onezone*) which enable creation of federations of data centers and users, storage providers (*Oneprovider*) which expose storage resources, and clients (*Oneclient*), who can access their data via a virtual POSIX file system. Onedata introduces the concept of *space*, a virtual volume, owned by one or more users, where the data is stored. 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 selected *Onezone* service 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 and OpenStack Swift, and GlusterFS. Onedata can serve as a unified virtual file system for multi-cloud environments.

In large scale hybrid cloud deployments it is often the case that data maintained in the private cloud has to be processed on-demand in the public cloud. While deploying remote jobs is today fairly straightforward, and can be automated using several orchestration platforms, making the data available for processing in the remote cloud is a significant challenge. Onedata makes this easy, by enabling automatic, on-demand, block based data prefetching based on the POSIX requests from user applications and automatically caching the files (up to defined capacity) based on analysis of file popularity. In most cases, prestaging is not necessary at all, as the data blocks are fetched on the fly when requested for reading, however it provides REST API for controlling data replication manually or integrating it with 3rd party services.

Currently Onedata is used in Helix Nebula Science Cloud [2], eXtreme DataCloud [3], PLGrid [4], European Open Science Cloud Hub [5], and European Open Science Cloud Pilot [6], where it procides data transparency layer for computation deployed on hybrid-clouds. Furthermore in European Open Science Cloud Hub [5] 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 registration) and discovery (OAI-PMH protocol).

Presentation will include brief description of Onedata solution and demonstrate caching and file popularity mechanisms for the hybrid cloud use cases.

## REFERENCES

[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`.

Corresponding author: Łukasz Dutka <lukasz.dutka@cyfronet.pl>

[5] 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`.

[6] European Open Science Cloud Pilot (The first phase in the development of the European Open Science Cloud). `https://eoscpilot.eu`.