Contribution ID: **81** Type: **Presentation**

Onedata Jupyter Integration Using OnedataFS

Tuesday 28 January 2020 16:40 (20 minutes)

Onedata is a global high-performance data management system that unifies data access across globally distributed environments and multiple types of underlying storages, such as NFS, Lustre, GPFS, Amazon S3, CEPH, as well as other POSIX-compliant file systems. It allows users to share, collaborate and perform computations on their data. 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.

Globally Onedata [1] comprises of Onezones, distributed metadata management and authorisation components that provide entry points for users to access Onedata; and Oneproviders, that expose storage systems to Onedata and provide actual storage to the users. Oneprovider instances can be deployed, as a single node or an HPC cluster, on top of high-performance parallel storage solutions with the ability to serve petabytes of data with GB/s throughput.

Onedata introduces the concept of Space, a virtual directory, owned by one or more users. The Spaces are accessible to users via an intuitive web interface, which allows for Dropbox-like file management and file sharing, Fuse-based client that can be mounted as a virtual POSIX file system, or REST and CDMI standardized APIs. Onedata does not provide users with any physical storage, and each Space has to be supported with a dedicated amount of storage by one or more providers, who are running Oneprovider component. The newly released python library - OnedataFS [2] - allows for even faster access to that data located in Onedata spaces. Thanks to integrating OnedataFS with Jupyter Content Manager API [3], one can not only access the data when using OnedataFS library inside the Notebook but also store the Jupyter Notebooks in Onedata Space.

Currently, Onedata is used in European Open Science Cloud Hub [4], eXtreme DataCloud [5], PRACE-5IP [6], and EOSC Synergy [7], where it provides data transparency layer for computation deployed on hybrid-clouds.

1. Onedata project website. http://onedata.org.
2. OnedataFS - PyFilesystem Interface to Onedata Virtual File System. https://github.com/onedata/fs-onedatafs.
3. Implementation of the Jupyter Content Manager API, for running Jupyter Notebooks on top of Onedataspaces. https://github.com/onedata/onedatafs-jupyter.
4. 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.
5. eXtreme DataCloud (Developing scalable technologies for federating storage resources). http://www.extreme-datacloud.eu.
6. Partnership for Advanced Computing in Europe - Fifth Implementation Phase. http://www.prace-ri.eu.
7. European Open Science Cloud - Expanding Capacities by building Capabilities. https://www.eosc-synergy.eu.

Authors: DUTKA, Lukasz (ACC Cyfronet-AGH); ORZECHOWSKI, Michał (ACC Cyfronet-AGH); KRYZA, Bartosz (ACC Cyfronet-AGH)

Presenter: 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