Contribution ID: **25** Type: **Extended Presentation (25' + 5' for questions)**

Integrated dataset placement service for scientists

Wednesday, October 12, 2022 4:30 PM (30 minutes)

Within the Helmholtz Association's federated IT platform HIFIS we have observed a trend towards more collaboration between different research groups and communities, both within Germany and the wider European research environment. These geographically distributed groups must manage their data so that members can access the data as needed. Meanwhile, funding agencies and scientific journals increasingly require that research data is FAIR: Findable, Accessible, Interoperable and Reusable. Of these features, Accessibility requires that computational activity can access the data; in particular, visualisation and other low-latency work may require access to data that is stored in a location separate from the compute infrastructure. In both cases, the data has to be transferred to the computing cluster in question.

We furthermore observed that data sets tend to become very large to the extent that transferring it requires specialised tools and systems. In the talk we will outline a setup that is going to be implemented by Helmholtz Federated IT services HIFIS for the research centres of the Helmholtz Association.

The distribution of large data amounts among data centres has been practiced, e.g., in the high-energy physics community by CERN and the worldwide LHC computing grid (WLCG) for years now. More scientific communities have started to discover the need for and advantages of distributed computing. Efficient computing, however, depends on fast data access as to not waste computing time while waiting for slow I/O processes. In order to address easy sharing of data sets between different IT infrastructures, we will propose and present a data analysis setup suited for data exploration and smaller analyses. The setup comprises a Jupyterhub instance with a locally available storage element for user access. The mounted storage element itself is integrated with a mesh of other storage elements at different sites between which data can be transferred.

The mesh integration is achieved by using a combination of Rucio and FTS3, both developed at CERN for the needs of the WLCG. Rucio is integrated with the Jupyterhub instance by means of a plugin which enables searching for data sets and applying replication rules to make them available locally. Transfers started this way are executed by FTS3 asynchronously and within a reasonable time frame such that scientists are able to analyse the data without too much delay or the need to plan ahead. In the Jupyter notebook, all data sets transferred in this manner are available from variables pointing transparently to the corresponding location on the local storage system.

We argue that providing a network of storage elements for storing, replicating and transferring data between sites is a feasible and necessary way enabling researchers to share their data among each other and with the general public. There are many storage and data transfer solutions available commercially but we feel that using a system that has been developed by scientists for scientists, with full control over the data and its availability, should be the sensible option for use in the scientific community.

Primary authors: FUHRMANN, Patrick (DESY); SCHUH, Michael (DESY); WETZEL, Tim (DESY IT (Research and Innovation in Scientific Computing))

Co-authors: SARWAR, Muhammad Aleem (DESY); Dr BEERMANN, Thomas (DESY); MILLAR, Paul (DESY); Dr REPPIN, Johannes (DESY)

Presenter: WETZEL, Tim (DESY IT (Research and Innovation in Scientific Computing))

Session Classification: IBERGRID Contributions

Track Classification: Development of innovative software services