Contribution ID: **15** Type: **not specified**

EOS as storage back-end for geospatial data analysis

Monday 5 February 2018 15:05 (20 minutes)

The Joint Research Centre (JRC) of the European Commission has set up the JRC Earth Observation Data and Processing Platform (JEODPP) as a pilot infrastructure to enable the knowledge production Units to process and analyze big geospatial data in support to EU policy needs. This platform is built upon commodity hardware and first operational services were made available mid 2016. It currently consists of processing and service nodes with a total of 1,200 cores, and the EOS system as storage back-end with a total gross capacity of 1.9 petabyte. EOS was deployed on the JEODPP with strong support by the CERN EOS team thanks to the CERN-JRC collaboration agreement. The JEODPP EOS instance relies on the EOS FUSE client given that currently there is no XrootD driver for the Geospatial Data Abstraction Library (GDAL) mainly used for reading and writing geospatial data files.

Multiple data processing levels have been implemented in the JEODPP. The batch processing system based on HTCondor is used for running large-scale data processing tasks based on HTCondor Docker or parallel universes and with all application dependent processes running in Docker containers. The web-based remote desktop level provides access to tools and software libraries for fast prototyping in a standard desktop environment. The interactive data processing in Jupyter notebooks allows for on-the-fly advanced data analysis and visualization.

The JEODPP platform is actively used by more than 15 JRC projects for data storage and various types of data processing and analysis. This required an additional monitoring system based on Grafana to better monitor the platform status. In order to better deal with user needs for data transfers and sharing, JRC will test the usage of CERNBox since it provides better integration with EOS than the currently deployed solution based on NextCloud.

The intensified usage of the platform and new data sources made it necessary to head for a major system extension which is currently underway. This will increase the EOS storage to a total gross capacity of 13 petabytes and the processing and service nodes to a total of 1,600 cores. The EOS service is planned to be migrated to the new Citrine release, and the usage of the new metadata management environment is envisaged once available and stable. The RAIN layout will be tested more extensively in 2018 as an alternative to the replica layout. The storage and processing platform in 2018 is going to be opened to JRC projects with new data domains and shall see a more extensive usage of machine learning technology. This way the platform is becoming the main scientific data hub at JRC.

Authors: BURGER, Armin (European Commission - Joint Research Centre); VASILEV, Veselin (European Commission, Joint Research Centre (JRC))

Co-author: SOILLE, Pierre (European Commission)

Presenters: BURGER, Armin (European Commission - Joint Research Centre); VASILEV, Veselin (European Commission, Joint Research Centre (JRC))

Session Classification: Using EOS