EGU2020-19280
https://doi.org/10.5194/egusphere-egu2020-19280
EGU General Assembly 2020

Filesystem and object storage for climate data analytics in private clouds with OpenStack

Ezequiel Cimadevilla Alvarez[1], Aida Palacio Hoz[2], Antonio S. Cofiño[1], and Alvaro Lopez Garcia[2]
[1]Santander Meteorology Group, Dep. of Applied Mathematics and Computational Sciences. University of Cantabria, Spain
[2]Advanced Computing Group, Instituto de Física de Cantabria (IFCA - CSIC), Spain

Data analysis in climate science has been traditionally performed in two different environments, local workstations and HPC infrastructures. Local workstations provide a non scalable environment in which data analysis is restricted to small datasets that are previously downloaded. On the other hand, HPC infrastructures provide high computation capabilities by making use of parallel file systems and libraries that allow to scale data analysis. Due to the great increase in the size of the datasets and the need to provide computation environments close to data storage, data providers are evaluating the use of commercial clouds as an alternative for data storage. Examples of commercial clouds are Google Cloud Storage and Amazon S3, although cloud storage is not restricted to commercial clouds since several institutions provide private or hybrid clouds. These providers use systems known as “object storage” in order to provide cloud storage, since they offer great scalability and storage capacity compared to POSIX file systems found in local or HPC infrastructures.

Cloud storage systems, based on object storage, are incompatible with existing libraries and data formats used by climate community to store and analyse data. Legacy libraries and data formats include netCDF and HDF5, which assume the underlying storage is a file system and it’s not an object store. However, new libraries such as Zarr try to solve the problem of storing multidimensional arrays both in file systems and object stores.

In this work we present a private cloud infrastructure built upon OpenStack which provides both file system and object storage. The infrastructure also provides an environment, based on JupyterHub, to perform remote data analysis, close to the data. This has some advantages from users perspective. First, users are no required to deploy the required software and tools for the analysis. Second, it provides a remote environment where users can perform scalable data analytics. And third, there is no constraint to download huge amounts of data, to users local computer, before running the analysis of the data.