EGU2020-14325
https://doi.org/10.5194/egusphere-egu2020-14325
EGU General Assembly 2020

Parquet Cube to store and process gridded data

elisabeth lambert, Jean-michel Zigna, Thomas Zilio, and Flavien Gouillon
CLS, Data, France (elambert@groupcls.com)

The volume of data in the earth data observation domain grows considerably, especially with the emergence of new generations of satellites providing much more precise measures and thus voluminous data and files. The 'big data' field provides solutions for storing and processing huge amount of data. However, there is no established consensus, neither in the industrial market nor the open source community, on big data solutions adapted to the earth data observation domain. The main difficulty is that these multi-dimensional data are not naturally scalable. CNES and CLS, driven by a CLS business needs, carried out a study to address this difficulty and try to answer it.

Two use cases have been identified, these two being complementary because at different points in the value chain: 1) the development of an altimetric processing chain, storing low level altimetric measurements from multiple satellite missions, and 2) the extraction of oceanographic environmental data along animal and ships tracks. The original data format of these environmental variables is netCDF. We will first show the state of the art of big data technologies that are adapted to this problematic and their limitations. Then, we will describe the prototypes behind both use cases and in particular how the data is split into independent chunks that then can be processed in parallel. The storage format chosen is the Apache parquet and in the first use case, the manipulation of the data is made with the xarray library while all the parallel processes are implemented with the Dask framework. An implementation using Zarr library instead of Parquet has also been developed and results will also be shown. In the second use case, the enrichment of the track with METOC (Meteo/Oceanographic) data is developed using the Spark framework. Finally, results of this second use case, that runs operationally today for the extraction of oceanographic data along tracks, will be shown. This second solution is an alternative to Pangeo solution in the world of industrial and Java development. It extends the traditional THREDDS subsetter, delivered by the Open source Unidata Community, to a bigdata implementation. This Parquet storage and associated service implements a smoothed transition of gridded data in Big Data infrastructures.