A Parquet Cube alternative
to store gridded data
for data analytics and modeling

Jean-Michel Zigna, CNES/CLS
jzigna@groupcls.com

CEOS WGISS-51
2021, April 21st
A Parquet Cube to address data analytics and Modeling

Traditional Earth Observations Gridded data in NetCDF files are not suitable to perform large scale processing

How to store them to facilitate ML, Modeling in Big Data infrastructures with existing open source technologies?

Is the Parquet format a good candidate?

➢ To process efficiently data

To evaluate for gridded datasets:
- Time, geographical, variables subsetting
- Enrichment of CSV locations with variables values
- Computing in cloud architecture

➢ To share data for different communities of users

- Scientists (Python oriented)
- Operational processing (Hadoop/Spark/Scala)

➢ To have good performance

Storage
- Compared to ZARR and NetCDF

Processing:
- 3 architectures to stress
- Hadoop/Spark/Scala
- Pangeo/Dask/Python
- THREDDS Services
Typical Scientific data sources with NetCDF files

- Open access hubs
- National hubs
- International hubs
- Copernicus services hubs
NetCDF Metadata is converted into a JSON file for a discovery service or extraction with NetCDF as output.
## Ingestion of Datasets of interest (1 year period)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>NetCDF</th>
<th>THREDDS-S3</th>
<th>ZARR</th>
<th>Parquet</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMEWS Global Analysis Forecast daily</td>
<td>1.22</td>
<td>1.388</td>
<td>0.454</td>
<td>0.720</td>
</tr>
<tr>
<td>CMEWS Global Analysis Forecast hourly</td>
<td>0.578</td>
<td>0.620</td>
<td>0.230</td>
<td>0.129</td>
</tr>
<tr>
<td>CLMS Lake Water Quality</td>
<td>0.068</td>
<td>N/A</td>
<td>0.006</td>
<td>0.164</td>
</tr>
<tr>
<td>CMEWS Sea Level Anomalies</td>
<td>0.070</td>
<td>N/A</td>
<td>N/A</td>
<td>0.074</td>
</tr>
</tbody>
</table>

*In Tb*
100 requests to stress architectures in place and try to understand the incidence of the subsetted period of time, the geographical coverage, the number of variables and their dimensions, the time resolution.

Incidence of geographical and time coverage (in s)
Scenario 1: subsetting

Incidence of the number of variables

Time in s

Number of variables: 1, 4, 11

1 day (dark blue), 1 month (orange), 3 months (grey), 6 months (orange), 1 year (light blue)
Scenario 2: Along Track Enrichment

35 requests to stress architectures in place and try to understand the incidence of:

* the period of time and the geographical coverage of locations (Rotterdam, North Sea, Global),
* the number of variables.

Results for Global geographical coverage (Time in s)
Scenario 1 and 2: Usage of the cloud scalability

Execution time = f(Nb Cores)

Execution time (s)

Number of cores

Execution time= f(Nb Cores)

Time steps X 24

Incidence of the number of time steps

Incidence of the number of cores for Enrichment

Enrichment

Subsetting

Subsetting
Scenario 3: SLA elevation computing

11 s Pangeo/Dask/ZARR in CNES HPC

38 s Hadoop/Spark/Parquet in CLS Cloud architecture

*(the initialization of the Spark context is around 30s)*
Conclusion

This **Parquet Cube** Alternative is a good candidate to face the data analytics and modeling in cloud environment for gridded data:

- **To get good performances in storage and processing**
  - Parquet storage size is around half of the NETCDF 3 (not compressed) size, in the same order than ZARR storage size
  - Generally greater in time extraction than Pangeo in CNES HPC, but faster for long term subsetting in CLS Cloud environment
  - Moving beyond the NetCDF TDS limits for global long term analysis and modeling

- **To share a common storage among communities of users using different development/processing environments**
  - Cloud storage allows efficient R,Scala/Spark,Python/dask computing in memory with Notebooks

- **To provide additional services**
  - To discover the data, describe the information for catalogues
  - To subset data if users want to download data on their premises
  - To enrich locations with environmental variables values and provide the relevant inputs for computing and modelling set up

- **Datasets, tests requests are available if you want to compare your solution/environment with our results (TileDB, COG...)**