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

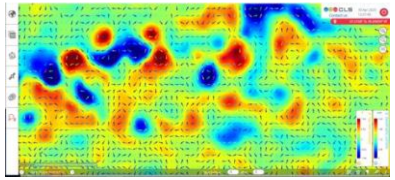
A Parquet Cube to address data analytics and Modeling

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 ?

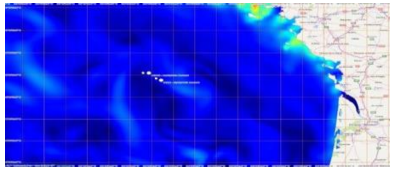
Is the Parquet format a good candidate ? Parquet

- **To process efficiently data** : 3 typical scenarii to evaluate for gridded datasets

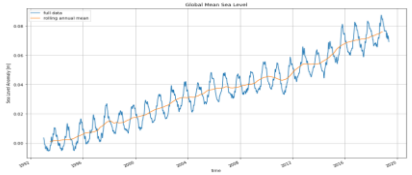
Time, geographical, variables subsetting



Enrichment of CSV locations with variables values



Computing in cloud architecture



- **To share data for different communities of users**

Scientists (Python oriented)

Operational processing (Haddop/Spark/Scala)

- **To have good performance**

Storage

- Compared to ZARR and NetCDF

Processing :

- 3 architectures to stress

Hadoop/Spark/Scala

hadoop Apache Spark Scala

Pangeo/Dasks/Python

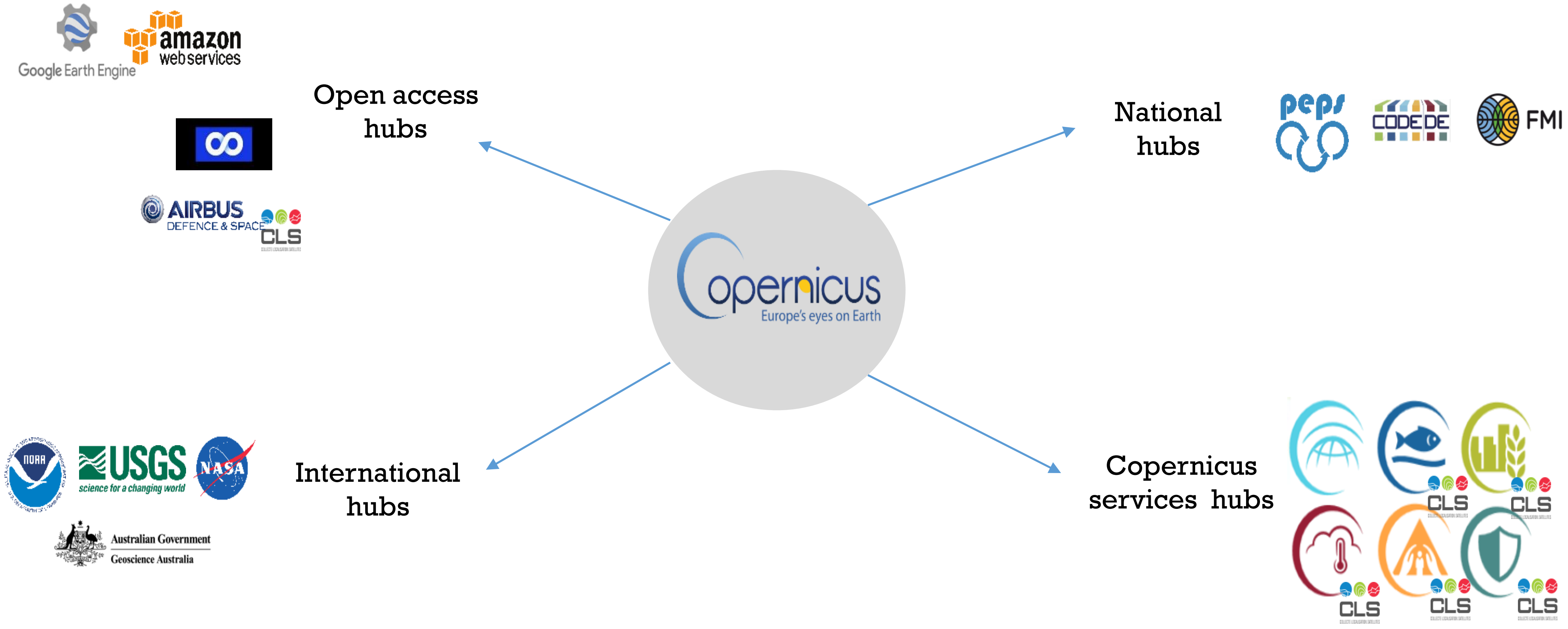
PANGEO
A community platform for Big Data geoscience

THREDDS Services

UCAR COMMUNITY PROGRAMS unidata
THREDDS Data Server (TDS)

CLS
COLLECTE LOCALISATION SATELLITES

Typical Scientific data sources with NetCDF files



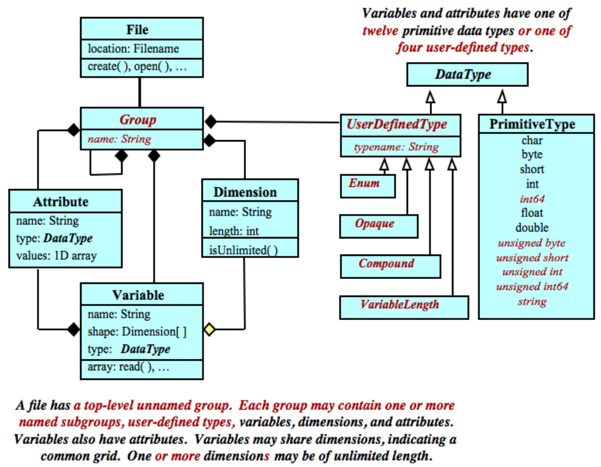
Open access hubs

National hubs

International hubs

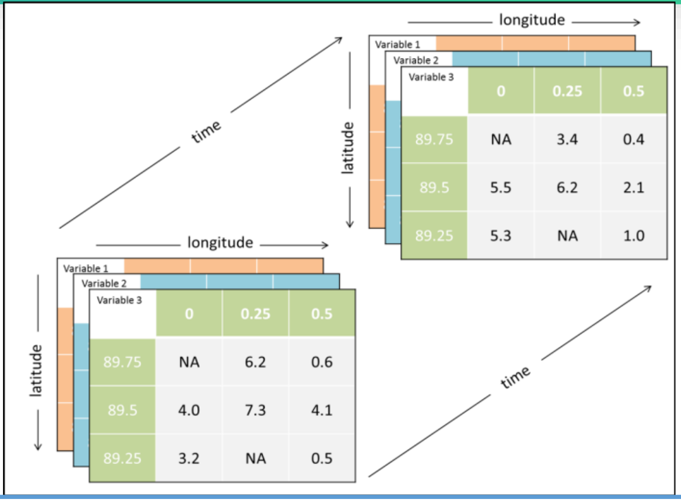
Copernicus services hubs

Please, do not change values during ingestion in Cubes



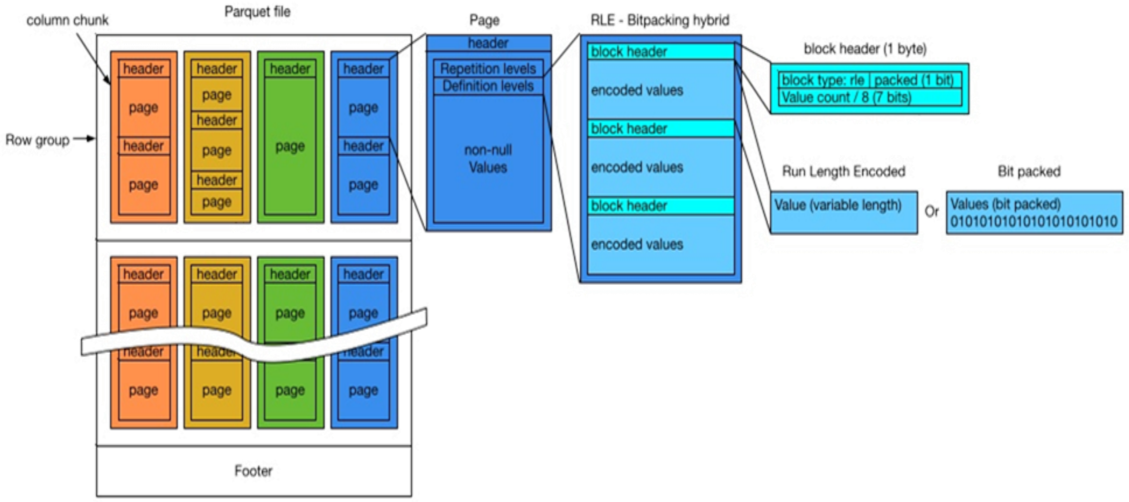
Variables and attributes have one of twelve primitive data types or one of four user-defined types.

A file has a top-level unnamed group. Each group may contain one or more named subgroups, user-defined types, variables, dimensions, and attributes. Variables also have attributes. Variables may share dimensions, indicating a common grid. One or more dimensions may be of unlimited length.



	0	0.25	0.5
89.75	NA	3.4	0.4
89.5	5.5	6.2	2.1
89.25	5.3	NA	1.0

	0	0.25	0.5
89.75	NA	6.2	0.6
89.5	4.0	7.3	4.1
89.25	3.2	NA	0.5



Dimensions				Variables		
latitude	longitude	time	depth	var1	….	varn

```
data
|_ index
|  |_ alti8-h
|  |  |_ latitude
|  |  |_ longitude
|  |  |_ time
|  |  |_ depth
|  |_ alti8-uv
|     |_ latitude
|     |_ longitude
|     |_ time
|     |_ depth
|_ data
   |_ alti8-h
   |  |_ tspartday=2017032
   |  |_ tspartday=2017033
   |  |_ tspartday=2017034
   |  |_ tspartday=2017035
   |_ alti8-uv
      |_ tspartday=2017032
      |_ tspartday=2017033
      |_ tspartday=2017034
      |_ tspartday=2017035
```

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)

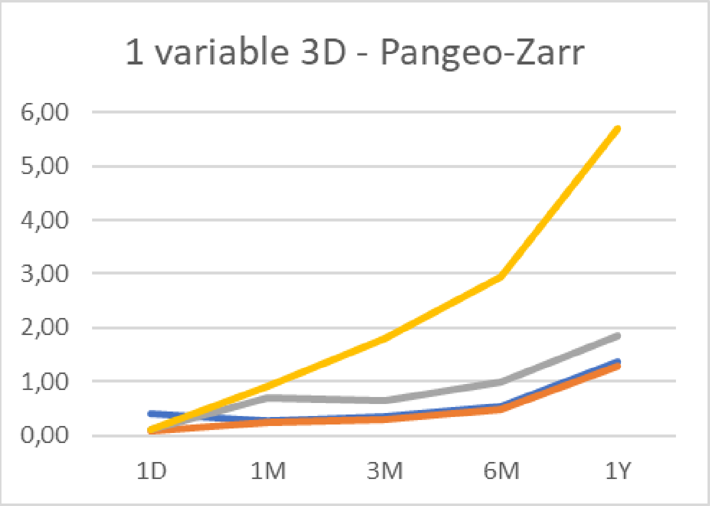
Dataset	NetCDF	THREDDS-S3	ZARR	Parquet
CMEMS Global Analysis Forecast daily	1.22	1.388	0.454	0.720
CMEMS Global Analysis Forecast hourly	0.578	0.620	0.230	0.129
CLMS Lake Water Quality	0.068	N/A	0.006	0.164
CMEMS Sea Level Anomalies	0.070	N/A	N/A	0.074

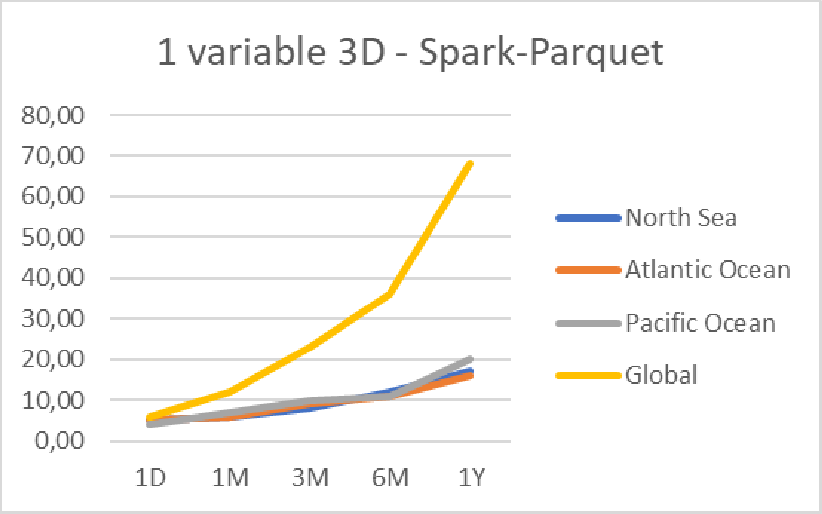
In Tb

Scenario 1 : subsetting

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)





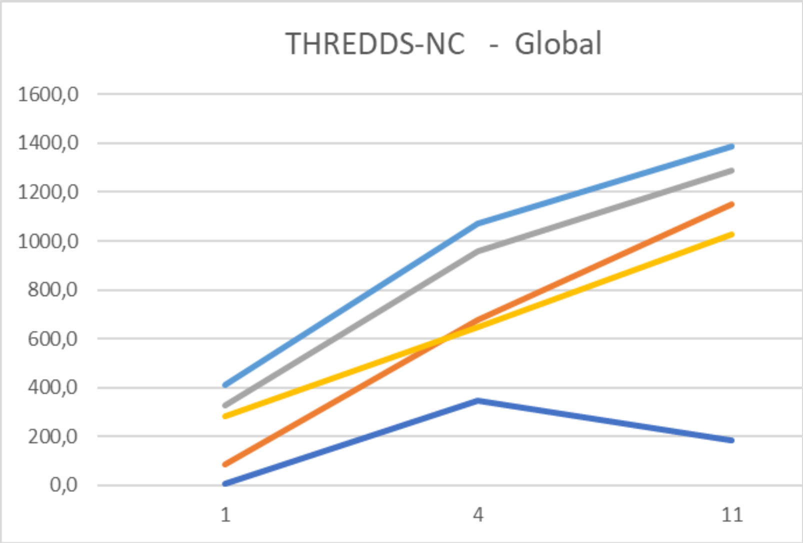
Scenario 1 : subsetting

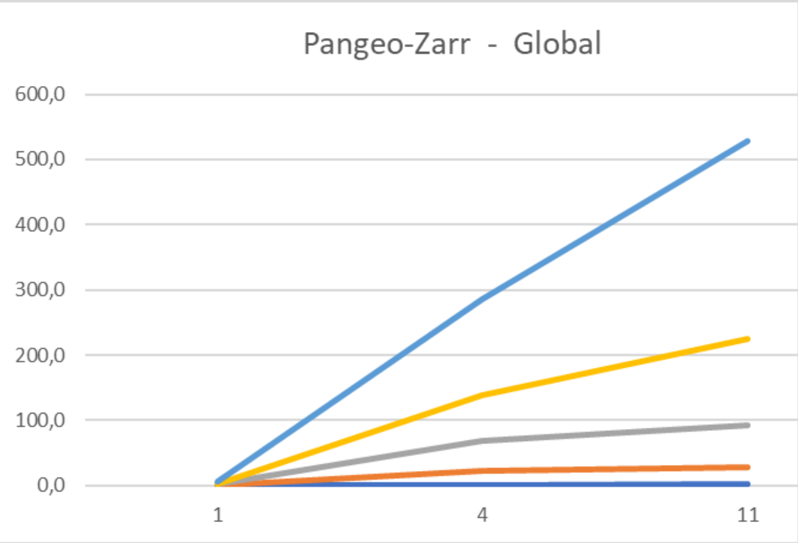
Incidence of the number of variables

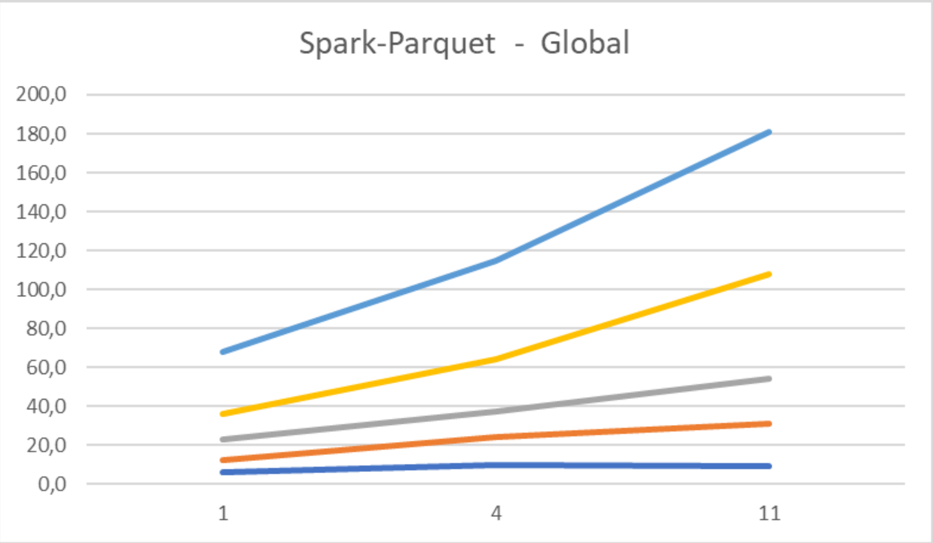
Time in s

Number of variables : 1, 4, 11

1 day (dark blue), 1 month (orange), 3 months (grey), 6 months (orange), 1 year (light blue)







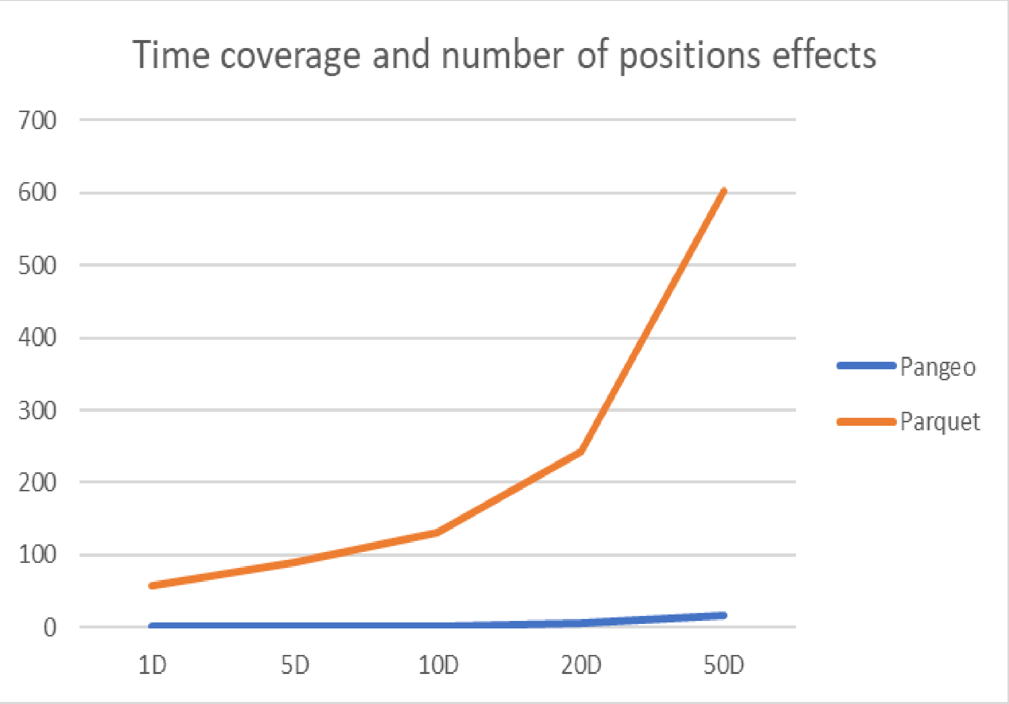
Scenario 2: Along Track Enrichment

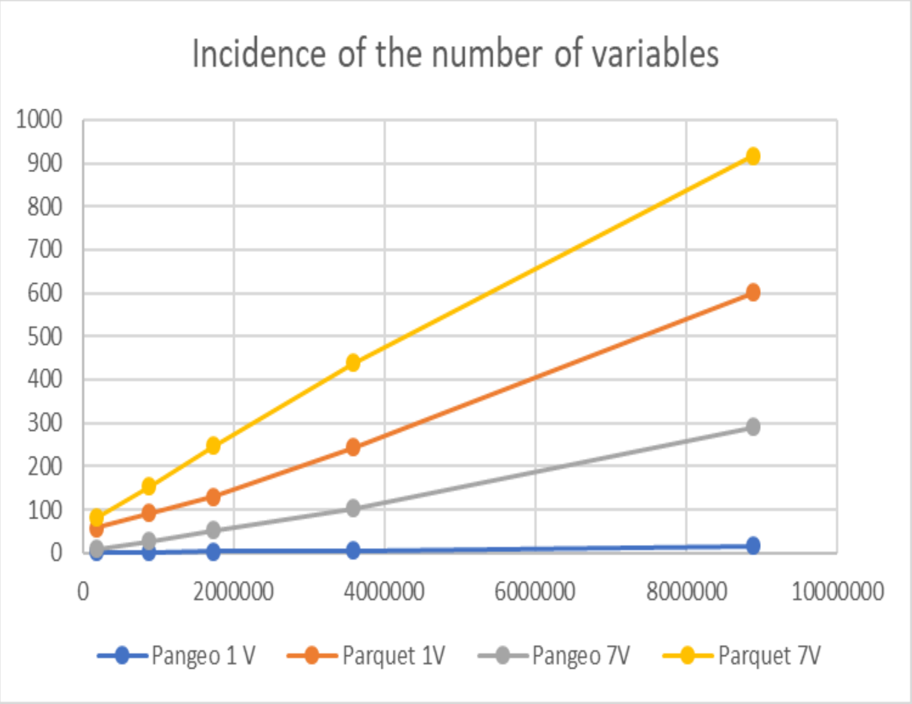
35 requests to stress architectures in place and try to understand the incidence of:

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

* the number of variables.

time	nb points
1D	176 735
5D	876 941
10D	1 736 849
20D	3 588 894
50D	8 897 349

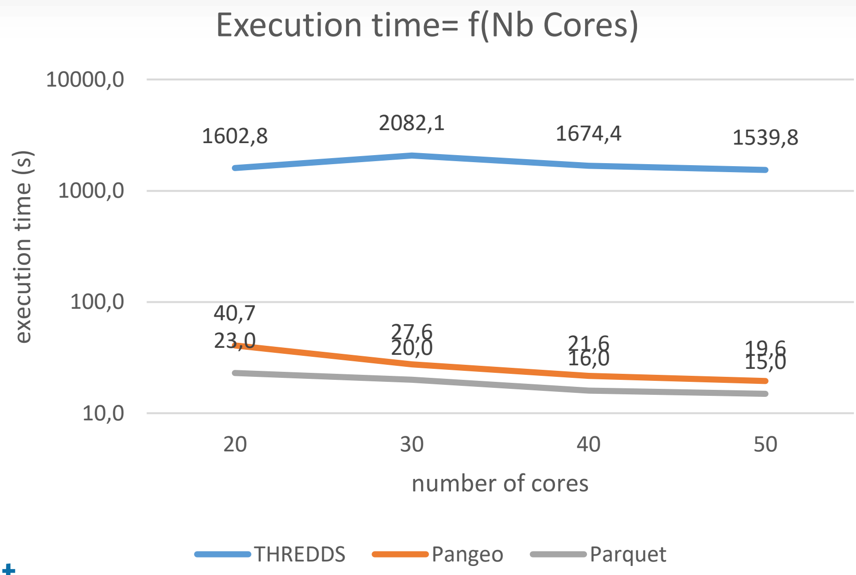




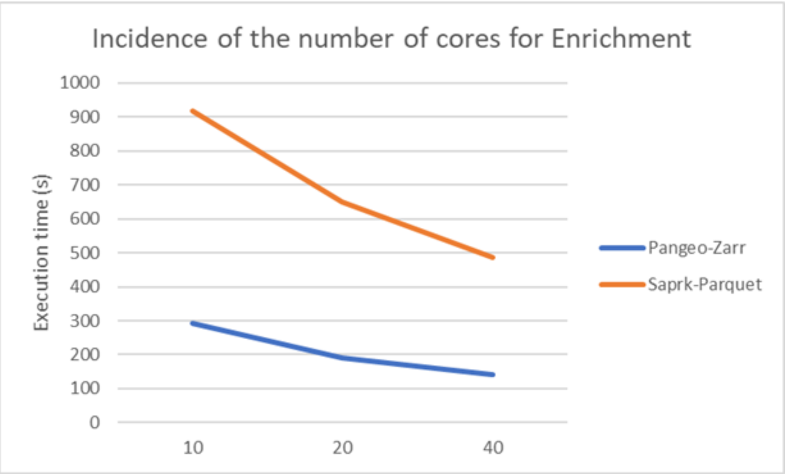
Results for Global geographical coverage (Time in s)

Scenario 1 and 2 : Usage of the cloud scalability

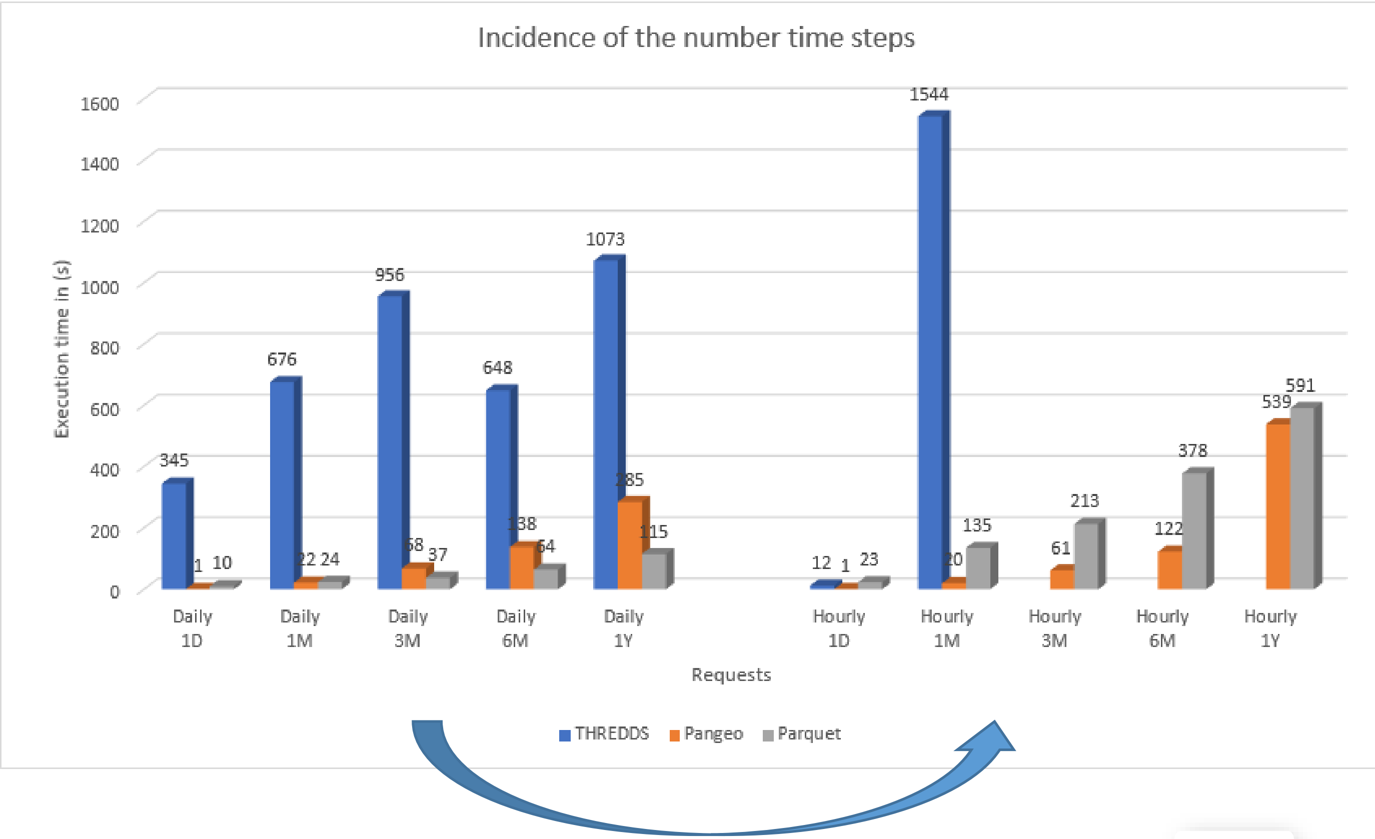
Subsetting



Enrichment



Subsetting

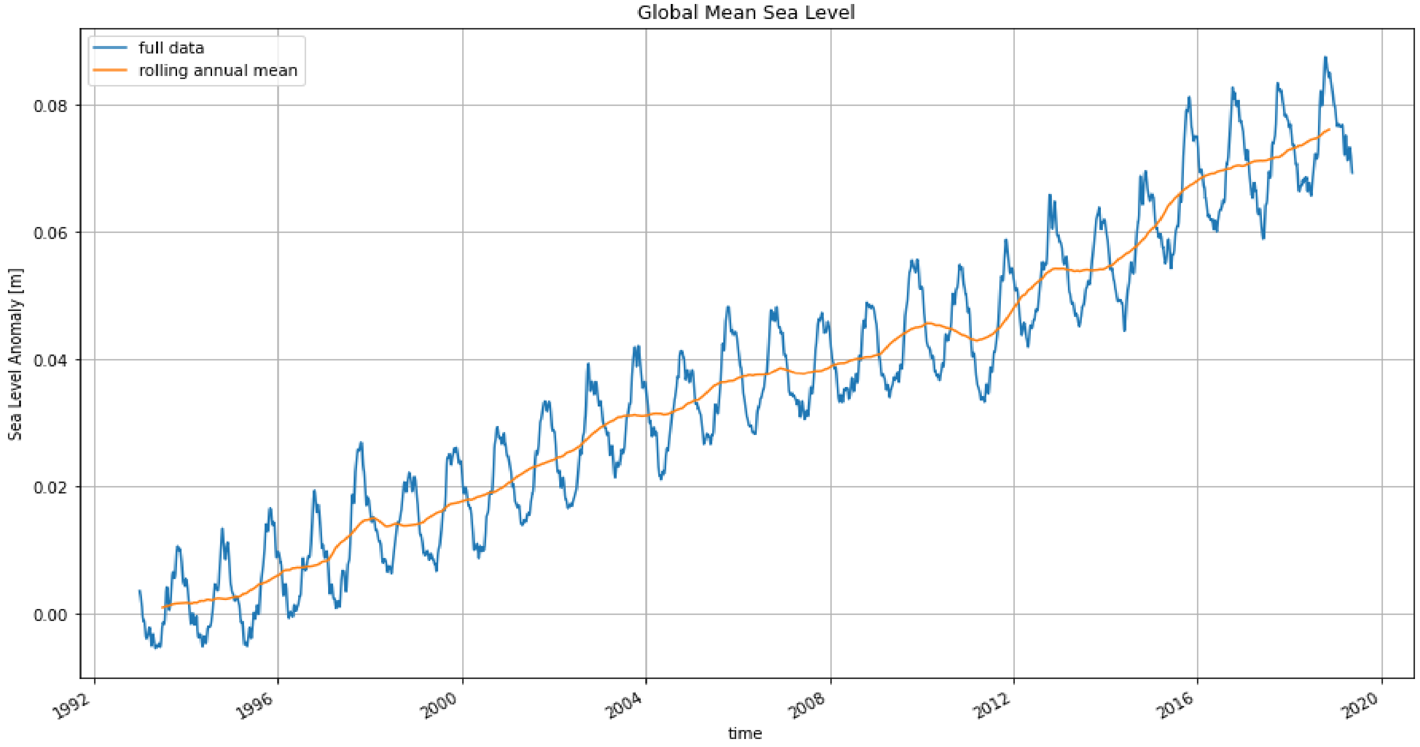


Time steps X 24

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 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…)