Scalable and high performance infrastructure for ocean data discovery and visualization

Glenn JUDEAU, IFREMER (France), Glenn.Judeau@ifremer.fr
Jérôme DETOC, IFREMER (France), Jerome.Detoc@ifremer.fr
Charlie ANDRE, IFREMER (France), Charlie.Andre@ifremer.fr
Léo BRUVRY-LAGADEC, IFREMER (France), Leo.Bruvry.Lagadec@ifremer.fr

This contribution presents a new data management and processing system that can handle massive amounts of oceanographic data as, for instance, held in the Coriolis database. This is achieved by combining established and widely used components in a clever way. Users will also benefit from a wide range of visualization styles. The Coriolis "in-situ" dataset is historically stored in Oracle and represents terabytes of data. While the dataset grows, reaching billions of measures, Oracle has shown limitations to address innovative use-cases.

IFREMER has built a Big Data solution to face modern challenges, ensuring sustainability of the dataset in the future and responding to both use-case workloads: Online Analytical Processing (OLAP) and Online Transaction Processing (OLTP). Those use case include interactive and complex metadata search-engine, sub-second data plotting, robust and high performance sub-setting and innovative Copernicus diffusion with large NetCDF4 files.

The solution implements three innovative data engines *(fig. 1)*:

- Spark, a scalable distributed processing engine
- Cassandra, a scalable high performance storage and access engine for data
- Elasticsearch, a scalable high performance metadata storage and powerful search engine

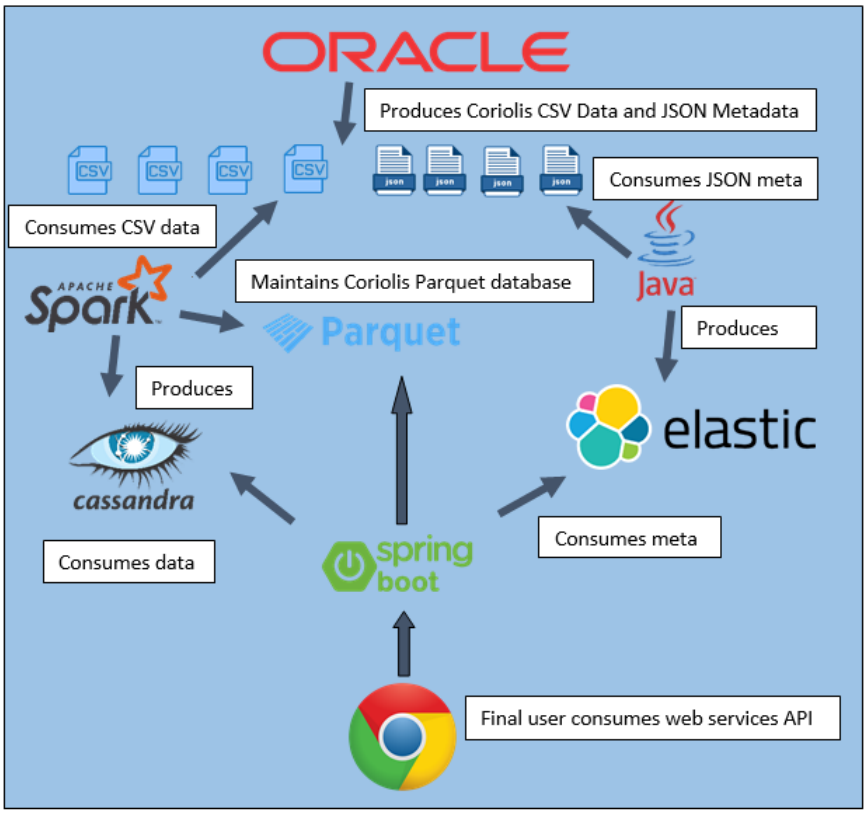
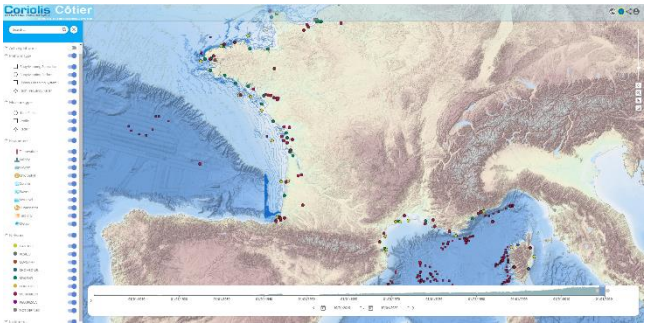


Figure 1: Infrastructure principles

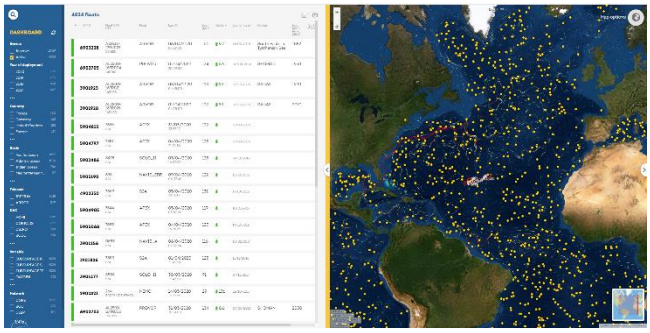
Each data engines deserve its own purpose to give best performances for the user's needs. In order to provide services to different endpoints, *IFREMER* set up Java REST APIs for each:

- Data-Discovery API to research using provided facets through Elasticsearch, returning JSON formatted metadata.
- Data-Plot API to get data-plots for profiles, time series, trajectories. Returning JSON format. Because returning thousands or millions data for visualization is time consuming and not relevant most of the time, a down-sampling algorithm has been implemented to the API: Largest Triangle Three Buckets (LTTB).
- Sub-setting API to download data in CSV or NetCDF4 format. User can be notified by email and data is pushed on *IFREMER*'s shared storage.

The use of API eases the ability to call services through different web portal for different purposes *(fig. 2)*:

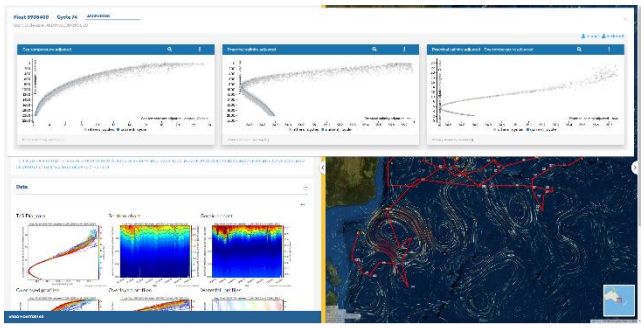


Coriolis-Côtier (data research)



Argo fleet monitoring (data research)

Coriolis-Côtier (data visualization)



Argo fleet monitoring (data visualization)

Figure 2: https://data.coriolis-cotier.org/and https://fleetmonitoring.euro-argo.eu/