Scalable and High performance infrastructure for Ocean data discovery and visualization

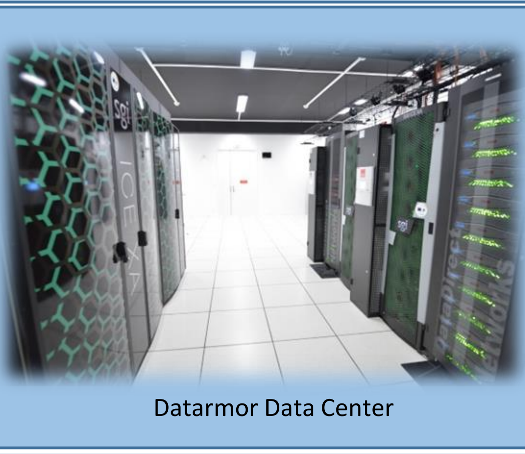
Glenn JUDEAU, Jérôme DETOC, Charlie ANDRE, Léo BRUVRY LAGADEC (IFREMER, French Research Institute for Exploration of the Sea (France))

Context

The **Coriolis "in-situ" dataset** is historically stored in **Oracle** and represents terabytes of data. While the dataset grow, reaching billions of measures, Oracle has shown **limitations** to address **innovative use-cases**. IFREMER has built a **Big Data solution** to face modern challenges. Those challenges include interactive and complex **metadata search-engine**, sub second **data plotting**, robust and high performance **subsetting** and innovative **Copernicus diffusion** with large NetCDF4 files.

Clustered infrastructure

In a **clustered architecture**, data is automatically **replicated** to multiple **nodes** for **fault-tolerance**. Replication across multiple data centers is supported. Failed nodes can be replaced with **no downtime**. One of the great features of clustered architecture is that it's designed from the ground up to be **horizontally scalable**, meaning that adding more nodes to the cluster we are capable to grow the capacity of the cluster.



Datarmor Data Center

Processing and Databases Stack

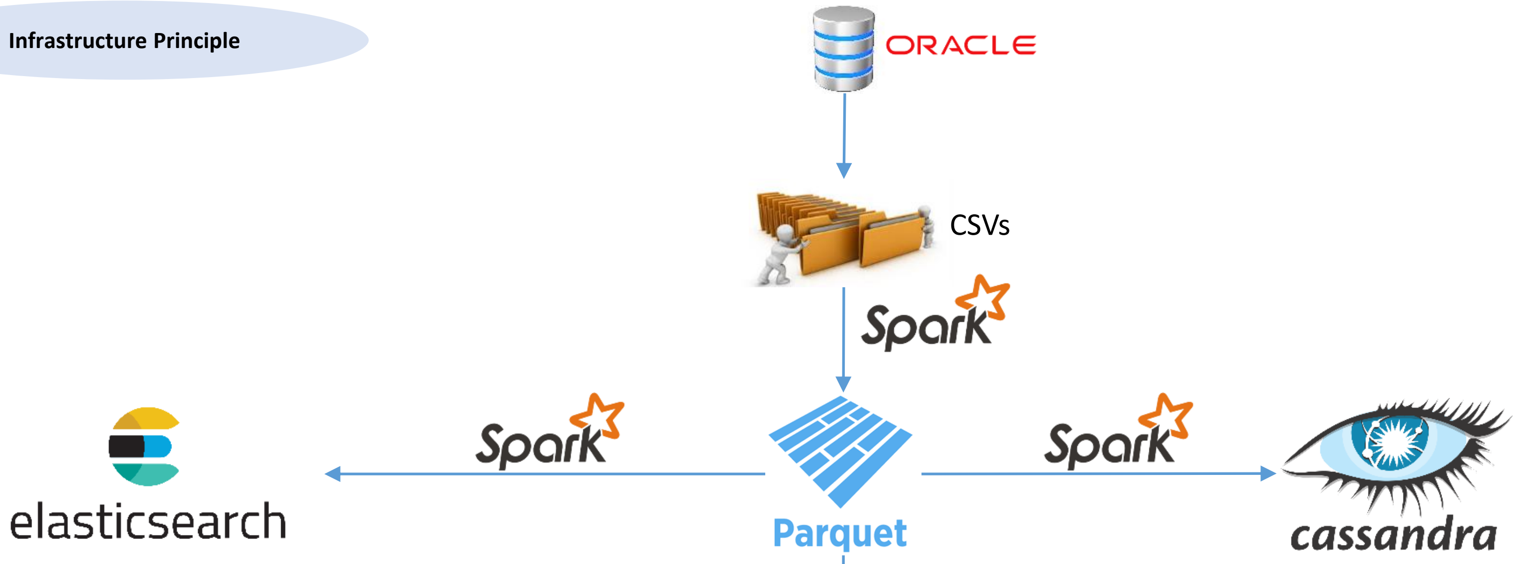
Spark is an open-source unified analytics engine for **large-scale data processing**. Spark provides an interface for programming entire clusters with implicit **data parallelism** and **fault tolrence**.

Parquet is an open source **file format** available to any project in the **Hadoop** ecosystem. Apache Parquet is designed for **efficient** as well as performent flat columnar storage format of data compared to row based files like CSV.

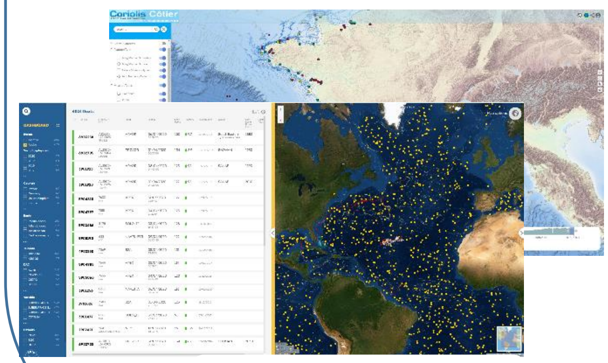
Elasticsearch is a free and open distributed search and analytics engine for any type of data, including **text, numeric, geospatial, structured and unstructured data**. Elastiseach was built on top of Apache Lucene. Elasticsearch is famous for its simple **REST APIs**, **distributed nature**, **speed** and **scalability**.

The **Apache Cassandra** database is the right choice for our needs regarding **scalability** and **high availability** without compromising **performance**. **Linear scalability** and proven **fault-tolerance** on our infrastructure make it the perfect platform for mission-critical data. Cassandra is supporting **replication** across multiple datacenters, providing lower latency for our users.

Infrastructure Principle



DATA DOWNLOAD



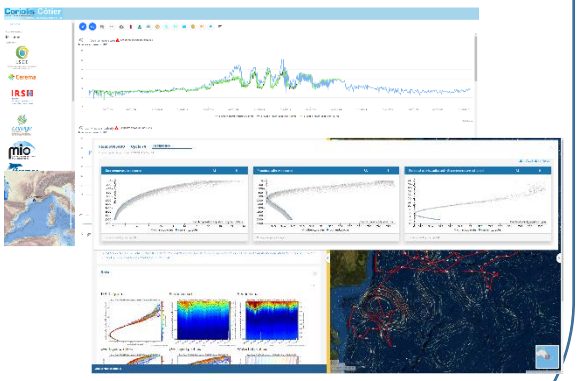
DATA DISCOVERY

2 Thematic Websites

5 Billions Measures

52 000 platforms

5 To Data



DATA PLOTTING

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