

# ID43- A BIG DATA ARCHITECTURE DESIGNED FOR OCEAN OBSERVATION DATA MANAGEMENT

PASQUALE ANDRIANI<sup>67</sup>, MASSIMILIANO NIGRELLI<sup>68</sup>, DARIO PELLEGRINO<sup>237</sup>, LEANDRO LOMBARDO<sup>242</sup>, LUCIO BADIALI<sup>94</sup>

The focus of this paper is on the architecture and the early-stage prototyping of a data management platform which is able to collect oceanographic measurements coming from both sensors deployed in different ocean observatories located in European waters (through EGIMs - EMSO Generic Instrument Module) and existing EMSO regional nodes (a European-scale research infrastructure of seafloor observatories).

The EMSODEV project (funded under the Horizon 2020 Programme, H2020-INFRADEV-1-2015-1, Grant Agreement n°676555) will include a Data Management Platform which will simultaneously collect, analyse and make a number of chemical and physical ocean parameters available, addressing needs from a very large scientific user community including biologists, geoscientists, chemists, and engineers.

The platform enables the exploitation of e-Infrastructures capabilities with the final goal of developing flexible and scalable data management services for long-term, high-resolution, (near)-real-time monitoring of environmental processes (e.g. natural hazards, climate change and marine ecosystems)

A big data architecture is a key requirement to guarantee availability, scalability and fault-tolerance for a state-of-the-art data management platform that is able to ensure flexible processing capabilities, pluggable data analysis services and a homogenised data access.

The early-stage data management platform prototype is based on the open source Hortonworks distribution of the Hadoop platform and includes a set of common services, compliant to the ENVRI reference model through the RM-ODP reference model:

- data acquisition and transformation services;
- data curation services (including data storage and partitioning, data quality checking and cataloguing services);
- data processing services (real time and/or batch processing computing capabilities);
- community support services (platform authentication and authorization);
- data access services (APIs for external access).

Sensors data will be coming both in asynchronous/batch and real-time mode, from two different data sources: a Sensor Observation Service (SOS) deployed close to each EGIM and the EMSO regional nodes. Both historical time series and real time data will be accessible and searchable via Application Programming Interfaces (APIs) built on top of NoSQL databases (e.g., HBase, OpenTSDB).

Two main processes happen during the data acquisition and transformation phase:

- Data Scraping extracting parts of a corpus of a sea observation coming/retrieved from the SOS (Sensor Observation Service) server;

- Data Munging/Wrangling converting data from a "raw" format of the observation coming/retrieved from the SOS server into another one that allows data to be more conveniently consumed later on in the data chain.

The data acquisition and transformation phase is implemented with the aid of Apache Nifi. Sensor data can be either retrieved via API exposed by an SOS server ("PULL" mode) or sent to the data management platform before being consolidated on the SOS server itself ("PUSH" mode).

Formatted details are then sent to different systems (an HDFS distributed file system, two different NoSQL databases, a Time series database, a Publish/Subscribe component) via specialised NIFI components performing HTTP POST/PUT submits against those systems. The distributed file system represents the way to store (and make available) long series of historical data. The data flowed into the Time Series DB are stored for further analysis and real time visualisation by customisable dashboards. The data flowed into NoSQL databases are used to feed the data management platform APIs which allow interested scientific communities to access various oceanographic parameters in different time ranges (e.g. the sea water temperature in a specific observatory within a certain time range).

Accurate, consistent, comparable, long-term measurements of ocean parameters are key to address urgent societal and scientific challenges such as climate change, ocean ecosystem disturbance, and marine hazards. Collecting ocean data and analysing them on a day-to-day basis are two big challenges to face in order to continuously monitor environmental processes including natural hazards, climate change, and marine ecosystems.

## REFERENCES

- EMSO - European Multidisciplinary Seafloor and water-column Observatory) - <http://www.emso-eu.org>
- EMSODEV - European Multidisciplinary Seafloor and water-column Observatory DEVELOPMENT - <http://www.emsodev.eu>
- OGC (Open Geospatial Consortium) - <http://www.opengeospatial.org/>
- ENVRI Reference Model - <https://confluence.egi.eu/display/EC/ENVRI+collaboration+and+documentation+space+Home>
- RM-ODP - Reference Model of Open Distributed Processing - <http://www.rm-odp.net/>
- Hortonworks - <http://hortonworks.com/>
- Apache Hadoop - <http://hadoop.apache.org/>
- Sensor Observation Service - <http://www.opengeospatial.org/standards/sos>
- Apache Nifi - <https://nifi.apache.org/>

