M1 OCEAN DATA MANAGEMENT: HIGH PERFORMANCE DATA MANAGEMENT SYSTEM WITH FLEXIBLE STRUCTURE TO SUPPORT INTERNATIONAL STANDARDS ON QUALITY ASSURANCE

Xulio Fernández Hermida[2], Sonia Lamas Pose[3], Manuel D. Lago Reguera[4], Darío Lodeiros Vázquez[5]

Abstract- The oceanographic and weather databases, whether historical or the new ones that are being created everyday, have multiple causes for statistical weaknesses: interruptions, errors of measurement, gaps and discontinuities. The interest of the scientific community focus on time series that have the longest and the best quality possible, so they seek to fuse information from different sources and gathered using different methodologies. This further difficults the representativeness and validity of the data. Therefore, there is an international agreement involving major scientific reference entities to establish standards that ensure the quality of ocean-meteorological data. We designed a novel data management system that takes advantage of novel non-relational databases in order to improve the management of the data generated by our stations. This system is designed to provide high-performance while being, at the same time, flexible enough to ensure compliance with future international standards on data quality, easing the integration with third party databases.

Keywords- ITC, Ocean Data, Standardarization, NoSQL databases

In January 2008, the International Oceanographic Data and Information Exchange (IODE) and the Joint Commission for Oceanographic and Marine Meteorology (JCOMM) held a forum about Ocenanographic Data Management and Exchange Standards. The project Ocean Data Standards (ODS) was an outcome of this forum. The main goal of ODS is to reach a broad agreement in order to adopt a set of standards related to the management and exchange of oceanographic data. There are numerous institutions, research centers and international organizations that generate, maintain and manage oceanographic, marine and meteorological databases. Each of these entities uses its own system to determine the quality of stored data and label it with the appropiate flag (e.g. "verified by quality control", "not verified", "absent", "interpolated", etc.). IODE aims to establish an standard methodology for data quality control, and thus ease the task of sharing databases between institutions, and increase the reliability of the available data sets by raising. As providers of ocean-meteorological information, we implemented a project to strengthen and streamline the storage and management of our data in order to ease the validation of our data internationally.

The process of reaching an international agreement on standardization is quite long and complex, as it is trying to set an agreement to the entire international community on how to manage the generation and processing of oceanographic information. Since the start of the project it has reached standardization agreements only with respect to the code of the country of origin of the data, date and time format, and recently for data quality index (the latter published 18 April 2013). These recommendations are available on the ODS website[1].

The document "Recommendation for a Quality Flag Scheme for the Exchange of Oceanographic and Marine Meteorological Data" indicates that the identification system for the quality of the data has two levels of information: a primary level, which refers to the identification of quality level itself, and a secondary level, which reports on the justification of this classification (which is why that data have a certain level of quality).

The quality of the data depends mainly on two issues: sampling protocols, which basically depend on the proper functioning of equipment and measuring instruments and their proper use, and random errors or deviations in measurements by environmental or unexpected causes (signal drift caused by fouling effect on a sensor, for example).

Quality control tests, or QC-tests, are necessary on the functioning of the system (Gap-test: evaluating whether the measure has come off or not, Sintax-text: that evaluates whether the message comes complete with the information chain intact) and tests on the measurements (if they are inside or outside the measuring range of the sensor / equipment / instrument and statistical deviations from average monthly, quarterly, annual, multivariate ... requiring already have values or previous series). In any case, the goal is to achieve a system that labels each data in real-time according to the results of these tests.

Our system for monitoring water quality, Hidroboya, captures and transmit the information (data) obtained in real time, without any prior processing. The final user is who performs the statistical analyses according to the objectives he looks for. We store all met-ocean data collected and maintain a database that can be further processed. If we label our data with the flags that are being agreed by IODE after a standardized statistical quality processed, we will have a more valuable product, and high quality environmental information.

The project that we are undertaking intends to migrate our information management systems to a new database structure that will facilitate the possibilities of computing, information processing, filtering and verification of the quality of each data with speed, flexibility and reliability. The proposal is based on open-source solutions that handle non-relational databases and data-mining software for information processing.

We work on implementing a system that will speed up the realization of the QC test automatically, that will allow the application of different QC tests quickly, and that will put on the web accessible data with the corresponding quality verifiers.

The data warehouse follows a non-relational database (non-sql) structure. It gives the flexibility inherent in the non-sql and can adapt the stored data schema to each particular case, to each client, or to any worldwide standard that is adopted now or in the future. This storage scheme system ensures the traceability of data: its origin, time of capture and measurement, measuring instruments, calibration of these instruments ...

It also allows us to incorporate two important safety features in information systems management: replication and high availability. The labeling for data storage, instead of the rigid structures in relational database tables, provides ease of use of Data Clusters, GridFS (storing files as graphics), and all this with a speed and efficiency much higher than relational databases.

There are open-source software tools, with huge computational possibilities, like MongoD. These tools facilitate the management of high volumes of data from oceanographic stations. In combination with a datamining system such as Apache Hadoop it will allow us to manage large amounts of information, both their own readings, as monitoring logs, communications, equipment states etc ...

Combining this data architecture and software development to extract, filter, combine and process information, we will implement the following functionalities, among others:

- Facilitate the implementation of quality control of data sets prior to being made available to the client, through statistical analysis, detection and / or elimination of outliers, interpolation to complete series, etc. ...

- Apply filtering statistical analysis (QC test) requested by the customer.

- Label each data according to their "level" of quality according to the standards of IODE, or according to any system of labeling required by the customer.

- Provide the customer a personalized treatment that acquires data using an API: the final user will process data with services developed by himself. This system allows a particular user to utilize its own services (web, processing) on the raw data stored (MongoDB) or treated with Apache Hadoop.

1- www.oceandatastandards.org/index.php?option=com_content&task=view&id=36&Itemid=44