A Cloud Computing Workflow for Managing Oceanographic Data

Salma Allam[1], Antonino Galletta[2(✉)], Lorenzo Carnevale[2], Moulay Ali Bekri[1], Rachid El Ouahbi[1], and Massimo Villari[2]

[1] Lab MIASH and Lab MACS, Department of Computer Science and Mathematics, Faulty of Sciences, University Moulay Ismail, Meknes, Morocco
allam.salma@gmail.com, ali.bekri@gmail.com, elouahbi@yahoo.fr
[2] Department of Engineering, University of Messina, Messina, Italy
{angalletta,lcarnevale,mvillari}@unime.it

Abstract. Ocean data management plays an important role in the oceanographic problems, such as ocean acidification. These data, having different physical, biological and chemical nature, are collected from all seas and oceans of the world, generating an international networks for standardizing data formats and facilitating global databases exchange. Cloud computing is therefore the best candidate for oceanographic data migration on a distributed and scalable platform, able to help researchers for performing future predictive analysis. In this paper, we propose a new Cloud based workflow solution for storing oceanographic data and ensuring a good user experience about the geographical data visualization. Experiments prove the goodness of the proposed system in terms of performance.

Keywords: Oceanography · Cloud Computing · Data collection
Data management · Data migration · NoSQL · Big Data

1 Introduction

Ocean Data management is a current challenge because both of ocean specific terminology diversity (physio-chemical parameters, sensor type, units of measures, conditions of measures, etc.) and of huge volume of ocean data collected from several international projects. The last aim to control the ocean acidification phenomena, an emerging global problem related to the seawater CO2 rate [1] that negatively affects the environment. Therefore, scientific community was thinking about software for calculating inorganic seawater carbon in order to track the evolution of CO2 in the oceans.

However, traditional desktop or web application can not provide the functionalities required by similar problem. Indeed, storing and processing a big volumes of data needs availability, reliability and scalability. For this purpose, the best choice for this kind of application is Cloud Computing, which delivers the resources for managing efficiently the collected data.

Z. Á. Mann and V. Stolz (Eds.): ESOCC 2017, CCIS 824, pp. 73–85, 2018.
https://doi.org/10.1007/978-3-319-79090-9_5

The goal of this scientific work follows the previous one [2]. More specifically, in this paper we planned to improve scalability and user experience of the Web Application, used for visualizing oceanographic data, already developed. For this purpose, here we analyzed the oceanographic data contest in order to design a Cloud workflow for migrating data from online databases to a distributed system able to enable future predictive analysis. Thus, we designed a data acquisition and integration workflow through a Cloud Storage approach which use a more recommended NoSQL solution for successfully managing semi-structured data and retrieving them for future seawater's acidification predictive analysis.

The rest of the paper is organized as follows. Related Works are described in the Sect. 2. In Sect. 3, we discussed the material used in this scientific work, from data source up to main oceanographic data issues. The Sect. 4 explains the Cloud approach used in order to migrate oceanographic data from sources to Cloud Storage, whereas in Sect. 5 we discussed the outcomes' experiments. Finally, the Sect. 6 concludes the paper with the lights for the future.

2 Related Work

The most popular oceanographic data visualization software is the Ocean Data View (ODV). According to Schlitzer [3], ODV is a software used for the interactive exploration, analysis and visualization of oceanographic and other georeferenced profile, time-series, and trajectory or sequence data. It displays original data points or gridded fields based on the original data and supports different data formats. ODV displays data on different views representing it in a global map that integrates the gridding algorithms [4] based software, called DIVA [5], in order to grid elements in the map for performing interpolation. Moreover, it allows to select one data source by entering the outer coordinates, considering the result as a separate small collection. In addition, ODV allows to select features for drawing one or more specific diagrams in order to compare these.

A software for 3D visualization has been proposed by Ware et al. [6]. The representation of that requires a user visual stimulation and allows them to compare two or more locations [7].

On the other hand, in recent time, the scientific community has started an investigation about atmospheric and oceanographic research using the Cloud Computing paradigm [8]. In order to proof that, in [9], the author reported a survey for discussing the progress made by Cloud in the oceanographic challenges. These include effective discovery, organization, analysis and visualization of large amounts of data. In [10], the authors reported "*the outcomes of an NSF-funded project that developed a geospatial cyberinfrastructure to support atmospheric research*". Specifically, they provided several modules for covering the aforementioned challenges in order to "*develop an online, collaborative scientific analysis system for atmospheric science*". In [11], instead, the authors described the LiveOcean project, which aims to mitigate "*the financial impact of ocean acidification on the shellfish industry in the Pacific Northwest of the United States*". The authors builded this system on Microsoft Azure Cloud Platform highlighting the modularity as most important theme.

Other important aspect is the management of the sensors designated for gathering row data. Indeed, increasing the number of data also the oceanographic context entries in the Big Data problem and specific solutions, such as [12,13], are useful for be inspired.

Our approach aims to go over the [3–7] solutions, proposing a Cloud Computing scenario in order to provide for Big Data coming into the oceanographic context.

Cloud Computing is a very hot topic in scientific community, it raises challenges in research fields as described in [14–17]. Most of the works are focused on the study and realization of innovative models that allows the collaboration among different Cloud providers focusing on various aspects of the federation. New trends in scientific work aim to adopt the Federation for new interesting scenarios: IoT, Edge, Cloud and Osmotic Computing [18].

3 Material

The following section describes the material used for this scientific work, highlighting the data structure and discussing the main challenges and issues.

3.1 Data Source

Data used in this scientific work came from the Carbon Dioxide Information Analysis Center (CDIAC), which is considered the first information analysis center for the oceanic parameters [19]. It provides an important climate-change database center organized as showed in the Fig. 1.

In particular, with reference to the Fig. 1, it is possible to notice the ODVServer zone, that represents the data sources selected in our study. Data are stored into relational databases, it is possible to export them using the comma separated value (CSV) format. In particular the ODVServer Data System is composed by three databases:

1. the GLobal Ocean Data Analysis Project (GLODAP) which gathers unified dataset for determining the "*global distributions of both natural and anthropogenic inorganic carbon, such as radiocarbon*" [20];
2. the PACIFic ocean Interior CArbon (PACIFICA) database that gathers "*data synthesis of ocean interior carbon and its related parameters in the Pacific Ocean*" [21];
3. the CARbon dioxide IN the Atlantic Ocean (CARINA) database which gathers "*data set of open ocean subsurface measurements for biogeochemical investigations*" [22].

Other data sources (SOCAT, CORILIOS CORA, JGOFS, eWOCE, LDEO and CLIVAR) are out of the scope of this paper and will be treated in future works.

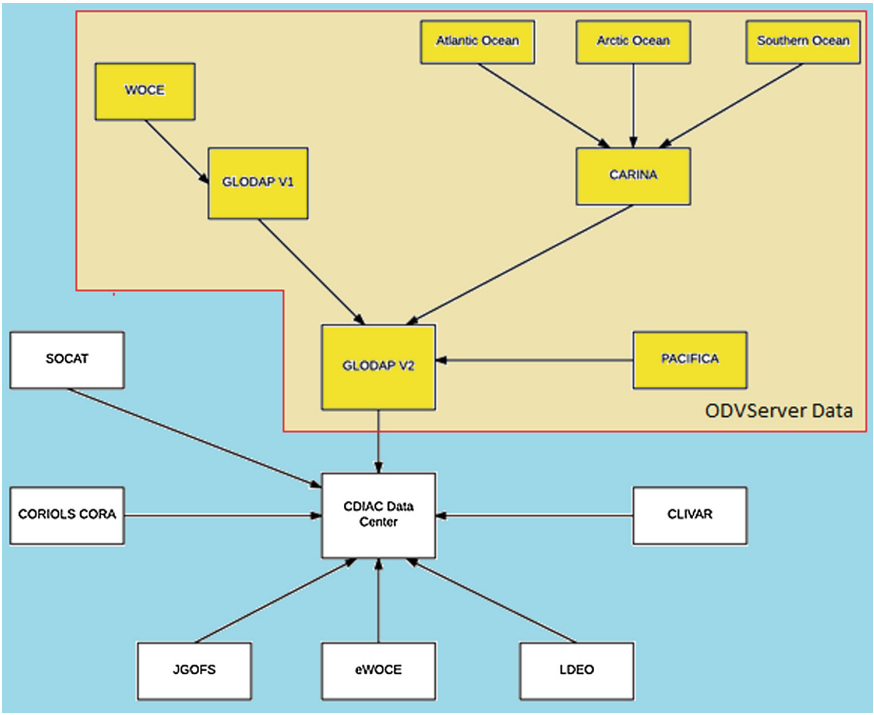

Fig. 1. CDIAC data center

3.2 Data Structure

Data collected in the aforementioned databases have a common structure explained in the following:

- **Time data:** Month, Day, Year;
- **Location data:** Longitude, Latitude, Depth;
- **Physical & Chemical data:** Section, Station, Cruise, BottomDepth, BottleNumber, Cast, Salinity, cdtSalinity, Oxygen, Nitrate, Nitrite, Silicate, Phosphate, CFC11, CFC12, CFC113, TCO2, Alkalinity, pCO2, pHSWS25, pHSWS25_Temp, AnthropogenicCO2, DOC, TOC, DeltaC14, DeltaC13, H3, DeltaH3, He, C14err, H3err, DeltaH3err, He_err, CCl4, SF6, AOU, pCFC11, CFC11Age, pCFC12, pCFC113, pCCl4, pSF6, CFC12Age, PotentialAlkalinity, ConventionalRadiocarbonAge, NaturalC14, bkgc14e, BombC14, BombC14atom, NaturalC14atom, PotTemperature, SigmaTheta, Sigma1, Sigma2, Sigma3, Sigma4, bf, sf, cdtsf, of, no3f, no2f, sif, po4f, cfc11f, cfc12f, cfc113f, tco2f, alkf, pco2f, phsws25f, aco2f, docf, tocf, c14f, c13f, h3f, dh3f, hef, ccl4f, sf6f, aouf, palkf, bkgc14f, bombc14f;
- **Environmental data:** Pressure, Temperature.

Data are collected according to the specific oceanographic region. Specifically, each region is composed by a set of sections that contain several stations

geographically located. For each of them, there are more than one record depending on the depth variation. The Fig. 2 represents a hierarchical overview of the entire dataset just described.

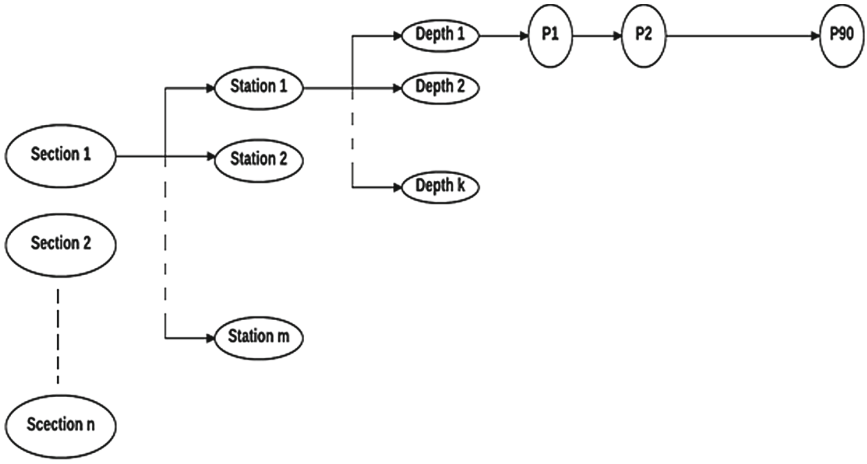

Fig. 2. Hierarchical multivalued attributes. Here P1, P2, ... , P90 denote Physical & chemical data.

3.3 Challenges and Issues of Oceanographic Data

The CDIAC oceanographic data have the 'Cruise/Section-Station-Depth' form. Such data are in relationships with time and space and archive all information about how, when and where data were stored, as well as type and nature of available data. Thus, it provides a conceptual overview of ocean data structure that should be useful in data management.

In ocean data, the dimensions are normally recorded as 'date/time, latitude, longitude, and depth' [23]. By associating latitude and longitude in location, the model will be 'date/time, location, and depth'. Unfortunately, data provided by the GLODAP, PACIFICA and CARINA databases do not have the data/time field in the 'dd/mm/yy hour:minute:second' form, but only in the 'dd/mm/yy' form. Therefore, the model becomes 'data, location, depth'. For this purpose, it is difficult to treat these data as time series.

Referring to the TLZ model [23], there are eight possible relationships, as reported in the Table 1.

We also noted that the geographical distance and depth gauge is not periodic among samples, i.e the Fig. 3 shows that the distance among stations is not the same and the stations depth is also different. More specifically, the data gauge on different depth is not uniform throughout the sample collection process.

Thus, the main challenge was to manage these data, in order to facilitate future predictive analysis and query them in a flexible way.

Table 1. DLH space relationship. Noted Date = 'D', Location = 'L' and Depth = 'H'.

D L H	Relationship
1 1 1	One date, one location, one depth
1 1 n	One date, one location, many depth
1 n 1	One date, many location, one depth
1 n n	One date, many location, many depth
n 1 1	many date, one location, one depth
n 1 n	many date, one location, many depth
n n 1	many date, many location, one depth
n n n	many date, many location, many depth

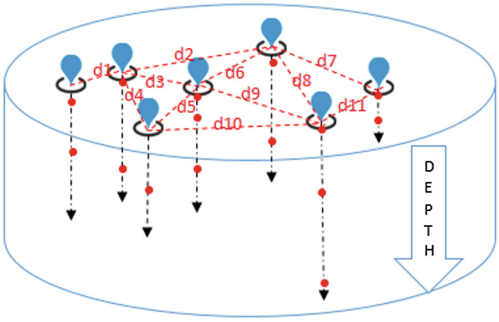

Fig. 3. A representation of data characteristic

4 Methodology

As mentioned in the Sect. 1, this work aimed to improve the previous one [2] through the Cloud Computing utilization. Specifically, the Sun/Oracle's JEE-based cross-platform, called ODVServer, used to store data in a traditional SQL database and to visualize oceanographic data using Google Maps APIs. The Fig. 4 shows an overview of the previous platform.

Unfortunately, the visualization of all sections from one database was not easily distinguishable. Moreover, we were not able to analyze the oceanographic data on the basis of different geographical shapes. For this purpose, we propose the Cloud improvement reported in the following.

4.1 Workflow

Driven by the need to improve the viewing system of the different oceanographic sections, we have planned to replace web data processing, performed with the Google Maps APIs, with the native storage of a GeoJSON, a human and machine-readable format for encoding geographic data structures.

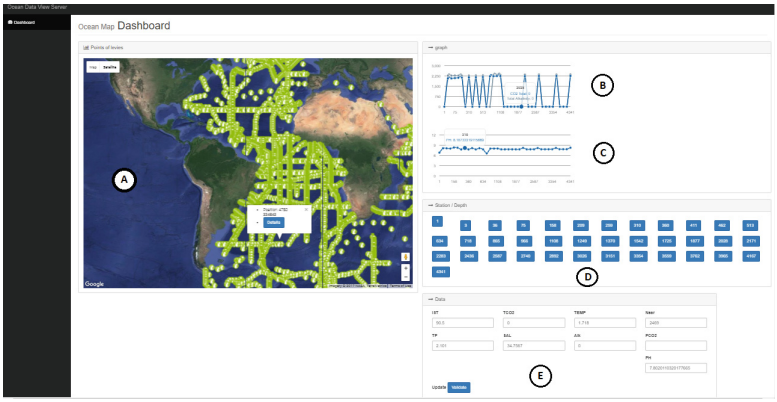

Fig. 4. Overview of the ODVServer platform. The layout was divided into two side. The first one (A) provided a geographical representation of the reading. Indeed, the right side (B, C, D, E) provided the insight view of the selected station.

Therefore, a NoSQL database was required in order to manage semi-structured and non structured data and for storing all the oceanographic databases.

Referring to the Fig. 5, data move from sources (CARINA, PACIFICA and GLODAP) to the Cloud Platform through RESTful APIs. Specifically, the listening microservice receives CSV data in order to implement the transformation into GeoJSON. Thus, this information moves to a sharded and replicated MongoDB distribution, a native JSON NoSQL database. The choice of MongoDB avoids further data transformations. Moreover, in order to scale the microservice workload, it is embedded inside a docker container, ensuring a lightweight and portable service virtualization.

The bottom side of the Fig. 5 shows a HTTP communication between the distributed storage and the front end, in order to view the query and future analyses results. At the same time, Apache Spark has been thought for performing future real-time predictive algorithms on oceanographic data. However, the dotted lines indicate guidelines for future works.

4.2 Oceanographic Data Visualization

Based on the previous description, we looked for two properties: interoperability and flexibility. The first one is guaranteed by the GeoJSON standard. Its fixed structure identifies two parts: **geometry** section contains geospatial information and type of GeoJSON element (see Sect. 2 in the Fig. 6); and **properties** section contains all parameters codified as key-value pairs (see Sect. 3 in the Fig. 6). We remark that Sect. 1 in the Fig. 6 represents the MongoDB's master key.

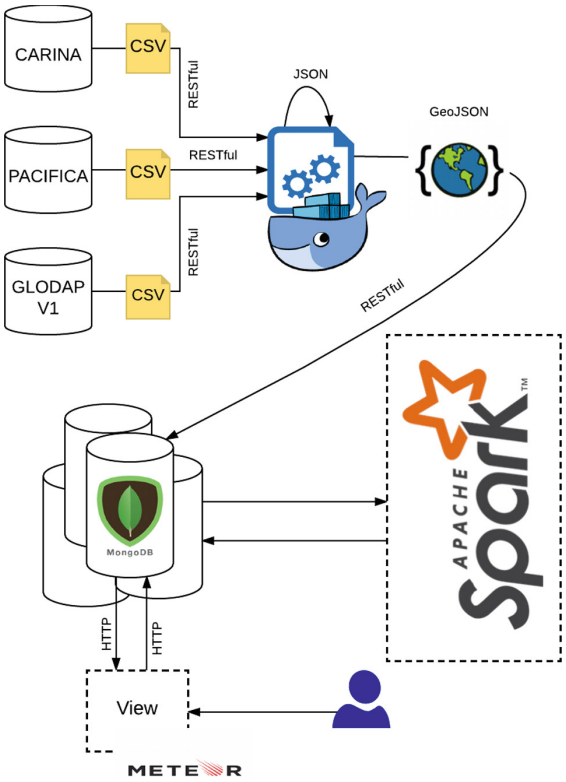

Fig. 5. The workflow includes the data acquisition and integration phases, considering the CSV format databases (CARINA; PACIFICA and GLODAP). A NoSQL solution stores all the GeoJSON information and integrates well with data analysis tool, such as Apache Spark, and MEAN stack web application, such as the Meteor JS framework. The dotted lines indicates guidelines for future works.

On the other hand, the flexibility is guaranteed by data management and visualization dynamism, which allow users to select any representation. Indeed, in our approach, we decided to store all samples in a single MongoDB collection. Thus, we can create virtual representation based on users' demand. For instance, as showed in the Fig. 7, by means of our approach users are able to create virtual representations per each zone of interest, starting from position of samples or other constrains.

Fig. 6. Difference structure of JSON and GeoJSON

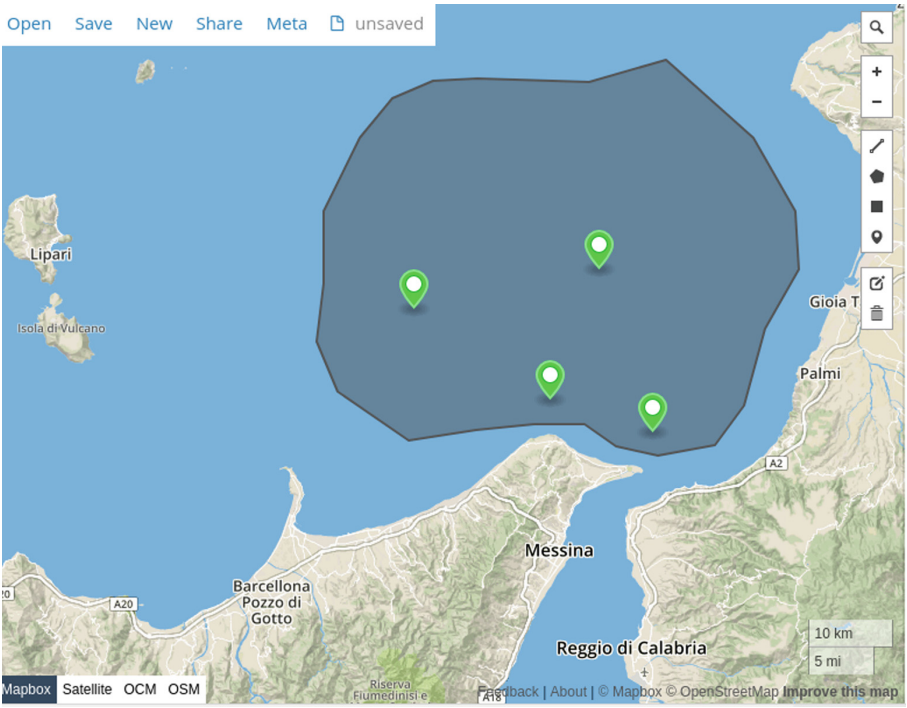

Fig. 7. User demand

5 Performance

In this section we discuss about of the performances of the system from a numerical point of view. In particular, we conducted two different kind of analysis: for populating and retrieve data of our system. Our testbed is composed of two different blades server. More specifically, we have 2 different type of machines one for the computation and the other one for the storage. Computation workstation, in which is running the data conversion module, is equipped by the Intel(R) Core(TM) i7-6700 CPU @ 3.40 GHz, RAM 16 GB, OS: Ubuntu server 16.04 LTS 64 BIT. Storage workstation, in which our database system MongoDB is running on single node, is composed by the Intel(R) Core(TM) i3-6100 CPU @ 3.70 GHz, RAM 32 GB, and Ubuntu server 16.04 LTS 64 BIT. Unfortunately we did not find any solution to compare performances of our system. We made scalability tests in different scenarios for both type of analysis. Experiments were repeated 30 subsequent times in order to consider confidence intervals at 95% and average values.

5.1 Insert Data

Figure 8 shows the performances for parsing CSV data to GeoJSON and storing them into MongoDB. Its behavior is linear with the increasing of the dataset size. On the x-axis we reported the dataset size, whereas on the y-axis, we reported the response time expressed in msec. How we can observe, the response time for 100.000 samples is acceptable less than 10 s.

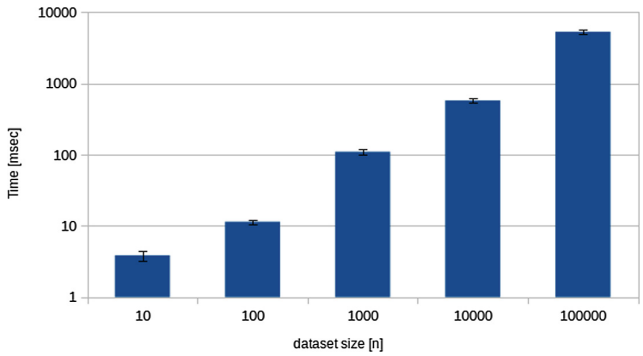

Fig. 8. Data insert performance

5.2 Retrieve Data

Here we consider times for retrieving data from MongoDB in a specific geographic shapes, in order to understand if flexibility features are really implementable. More specifically we considered increasing concentric circles with different radius, starting from 10 m up to 100 km.

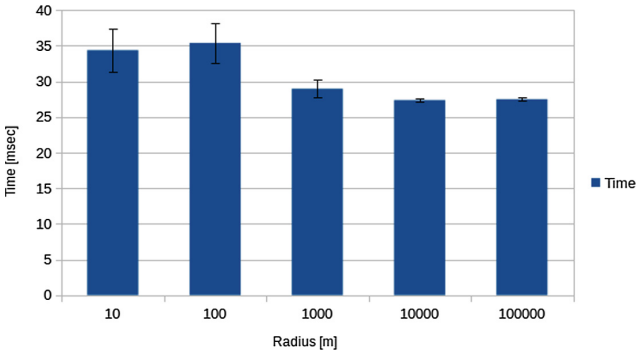

Fig. 9. Data retrieve performance

The behavior, as showed in Fig. 9, is constant around 30 ms, variations are due to networks delay.

6 Conclusions and Future Work

In this scientific work, we investigated the management of oceanographic data through the utilization of a Cloud Computing workflow. First of all, three CSV format databases have been selected as data sources. Therefore, we explained the workflow necessary for migrating these data up to the Cloud Storage. This scientific work is the first initiative adopting Cloud for manage Ocean data, for this reason we did not find any solution to compare performance of our system. However experiments showed that our system response time presents a linear trend. The execution time grows up with the increasing number of considered samples.

On the other hand, the dotted lines in the Fig. 5 shows our idea about the future perspective. In particular, Meteor JS framework will be the technology we

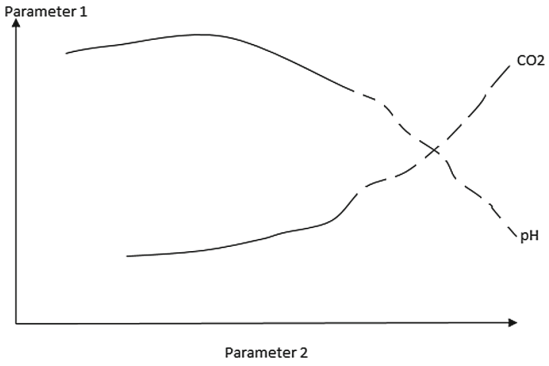

Fig. 10. A goal of the future work can be the acidification prediction.

aim to use for developing the new frontend version. This choice depends on the native MEAN stack adopted, which includes MongoDB as backend database; whereas Apache Spark will be useful for performing predictive oceanographic data analysis, such as the acidification prediction. About that, the Fig. 10 shows a possible future work about the selection of the features that best describe the reported behavior. Other future works are related to adopt the new Osmotic Computing paradigm.

Acknowledgment. This work has been supported by Cloud for Europe grant agreement number FP7-610650 (C4E) Tender: *REALIZATION OF A RESEARCH AND DEVELOPMENT PROJECT (PRE-COMMERCIAL PROCUREMENT) ON "CLOUD FOR EUROPE"*, Italy-Rome: Research and development services and related consultancy services Contract notice: 2014/S 241-424518. Directive: 2004/18/EC. (http://www.cloudforeurope.eu/).

References

1. Doney, S.C., Balch, W.M., Fabry, V.J., Feely, R.A.: Ocean acidification a critical emerging problem. Oceanography **22**(4), 16–25 (2009)
2. Allam, R.E.S., Ouahbi, M.D.E.O.: Adv. Inf. Technol. Theory Appl. **1**, 163–166 (2016). ISSN: 2489–1703
3. Schlitzer, R.: Ocean Data View, pp. 1–11 (2011)
4. Smith, W.H.F., Wessel, P.: Gridding with continuous curvature splines in tension. Geophysics **55**(3), 293–305 (1990). http://library.seg.org/doi/10.1190/1.1442837
5. Started, G.: Ocean Data View, pp. 1–11 (2011)
6. Ware, C., Plumlee, M., Arsenault, R., Mayer, L.A., Smith, S., House, D.: GeoZui3D: data fusion for interpreting oceanographic data. In: Oceans Conference Record (IEEE), vol. 3, pp. 1960–1964 (2001)
7. Plumlee, M., Ware, C.: An evaluation of methods for linking 3D views. In: Proceedings of the Symposium on Interactive 3D Graphics, pp. 193–201 (2003). http://www.scopus.com/inward/record.url?eid=2-s2.0-0038642661&partnerID=tZOtx3y1
8. Butler, K., Merati, N.: Analysis patterns for cloud-centric atmospheric and ocean research. In: Cloud Computing in Ocean and Atmospheric Sciences, pp. 15–34. Elsevier (2016). https://doi.org/10.1016/b978-0-12-803192-6.00002-5
9. Wigton, R.: Forces and patterns in the scientific cloud. In: Cloud Computing in Ocean and Atmospheric Sciences, pp. 35–41. Elsevier (2016). https://doi.org/10.1016/b978-0-12-803192-6.00003-7
10. Li, W., Shao, H., Wang, S., Zhou, X., Wu, S.: A2ci. In: Cloud Computing in Ocean and Atmospheric Sciences, pp. 137–161. Elsevier (2016). https://doi.org/10.1016/b978-0-12-803192-6.00009-8
11. Fatland, R., MacCready, P., Oscar, N.: LiveOcean. In: Cloud Computing in Ocean and Atmospheric Sciences, pp. 277–296. Elsevier (2016). https://doi.org/10.1016/b978-0-12-803192-6.00014-1
12. Fazio, M., Celesti, A., Villari, M., Puliafito, A.: The need of a hybrid storage approach for IoT in PaaS cloud federation. In: 2014 28th International Conference on Advanced Information Networking and Applications Workshops, pp. 779–784 (2014)

13. Celesti, A., Peditto, N., Verboso, F., Villari, M., Puliafito, A., Draco PaaS: a distributed resilient adaptable cloud oriented platform. In: IEEE International Symposium on Parallel Distributed Processing. Workshops and PhD Forum 2013, pp. 1490–1497 (2013)
14. Tusa, F., Celesti, A., Villari, M., Puliafito, A.: How to enhance cloud architectures to enable cross-federation. In: Proceedings of IEEE CLOUD 2010, pp. 337–345. IEEE, July 2010
15. Vernik, G., Shulman-Peleg, A., Dippl, S., Formisano, C., Jaeger, M., Kolodner, E., Villari, M.: Data on-boarding in federated storage clouds. In: 2013 IEEE Sixth International Conference on Cloud Computing (CLOUD), pp. 244–251, June 2013
16. Goiri, I., Guitart, J., Torres, J.: Characterizing cloud federation for enhancing providers' profit. In: 2010 IEEE 3rd International Conference on Cloud Computing (CLOUD), pp. 123–130, July 2010
17. Azodolmolky, S., Wieder, P., Yahyapour, R.: Cloud computing networking: challenges and opportunities for innovations. IEEE Commun. Mag. **51**(7), 54–62 (2013)
18. Villari, M., Fazio, M., Dustdar, S., Rana, O., Ranjan, R.: Osmotic computing: a new paradigm for edge/cloud integration. IEEE Cloud Comput. **3**(6), 76–83 (2016)
19. Catalog of Databases and Reports, May 1999. http://cdiac.ornl.gov/oceans/
20. Key, R., Olsen, A., van Heuven, S., Lauvset, S., Velo, A., Lin, X., Schirnick, C., Kozyr, A., Tanhua, T., Hoppema, M., Jutterström, S., Steinfeldt, R., Jeansson, E., Ishi, M., Perez, F., Suzuki, T.: Global Ocean Data Analysis Project, Version 2 (GLODAPv2). Ornl/Cdiac-162, Ndp-093, vol. 2 (2015)
21. Suzuki, T., Ishii, M., Aoyama, M., Christian, J.R., Enyo, K., Kawano, T., Key, R.M., Kosugi, N., Kozyr, A., Miller, L.A., Murata, A., Nakano, T., Ono, T., Saino, T., Sasaki, K.-I., Sasano, D., Takatani, Y., Wakita, M., Sabine, C.L.: Pacifica Data Synthesis Project. Ornl/Cdiac-159, Ndp-092 (2013)
22. Hoppema, M., Velo, A., van Heuven, S., Tanhua, T., Key, R.M., Lin, X., Bakker, D.C.E., Perez, F.F., Ríos, A.F., Lo Monaco, C., Sabine, C.L., Álvarez, M., Bellerby, R.G.J.: Consistency of cruise data of the CARINA database in the Atlantic sector of the Southern Ocean. Earth Syst. Sci. Data **1**, 63–75 (2009)
23. Hubbs, C.L.: University of Michigan, U.S.A., vol. III, pp. 1–6 (1930)