Pers Ubiquit Comput (2017) 21:55–65
DOI 10.1007/s00779-016-0980-2

ORIGINAL ARTICLE

Big data challenges in ocean observation: a survey

Yingjian Liu[1] · Meng Qiu[1] · Chao Liu[1] · Zhongwen Guo[1]

Received: 1 September 2016 / Accepted: 30 September 2016 / Published online: 12 November 2016

Abstract Ocean observation plays an essential role in ocean exploration. Ocean science is entering into big data era with the exponentially growth of information technology and advances in ocean observatories. Ocean observatories are collections of platforms capable of carrying sensors to sample the ocean over appropriate spatiotemporal scales. Data collected by these platforms help answer a range of fundamental and applied research questions. Many countries are spending considerable amount of resources on ocean observing programs for various purposes. Given the huge volume, diverse types, sustained measurement, and potential uses of ocean observing data, it is a typical kind of big data, namely marine big data. The traditional data-centric infrastructure is insufficient to deal with new challenges arising in ocean science. New distributed, large-scale modern infrastructure backbone is urgently required. This paper discusses some possible strategies to solve marine big data challenges in the phases of data storage, data computing, and analysis. Some applications in physics, chemistry, geology, and biology illustrate the significant uses of marine big data. Finally, we highlight some challenges and key issues in marine big data.

Keywords Ocean observation · Marine big data · Data storage · Data computing · Data analysis

1 Introduction

The ocean covers more than 2/3 of Earth's surface. Phytoplankton in the surface ocean produces half of the oxygen from photosynthesis on Earth. Ninety percent of heat from global warming has been absorbed by the ocean. Ninety percent of international trade travels by ship. No matter where we live, the ocean affects our life. However, 95% of the ocean remains unexplored and under-appreciated by humans [1]. This calls for understanding all facets of the ocean as well as its complex connections with Earth's atmosphere, land, ice, seafloor, and life—including humanity. It is essential not only to advance knowledge about our planet, but also to ensure society's long-term welfare and to help guide human stewardship of the environment.

Oceanography is evolving from a ship-based expeditionary science to a distributed, observatory-based approach, facilitating data collection of long-term time series and providing an interactive capability to conduct experiments using data streaming in real time [2]. For example, the Ocean Observatories Initiative (OOI) [3] manages and integrates data from over 800 instruments deployed among its seven arrays. Instruments are located on a myriad of platforms including gliders, autonomous underwater vehicles (AUVs), surface buoys, profilers, inductive mooring cables, and seafloor junction boxes. Over 200 unique data products are measured or derived from nearly 75 models of specialized instrumentation used in the OOI from the air–sea interface to the seafloor. Multi-source ocean observing data are collected and stored at an unprecedented scale and speed [4]. Based upon Gartner's definition of big data [5], ocean observing data do have the 3Vs (volume, velocity, and variety) characteristics. Therefore, ocean observing data can be regarded as a typical kind of big data, i.e., marine big data.

✉ Yingjian Liu
liuyj@ouc.edu.cn

[1] Department of Computer Science and Technology, Ocean University of China, 238 Songling Road, Qingdao 266100, China

These data must be stored in raw format, parsed, calibrated, and processed for quality control, then analyzed, and further derived into other products such as visualizations [6]. Due to the unique characteristics of marine big data, such as multi-source, long-lasting, uncertainty, and incompleteness, they exceed the processing and analysis capacities of conventional systems. This situation has caused new challenges for the traditional technologies such as relational databases and scale-up infrastructures [7]. Current researches involving big data primarily concern with how to discover and make sense of such high amounts of data more effectively and efficiently [8]. Key issues investigated include infrastructure [9], storage [10], analysis [11], security [12], etc.

The rest of this paper is organized as follows. In Sect. 2, we introduce some important ocean observatories and ocean observing programs for data acquisition. In Sects. 3 and 4, we discuss some possible strategies to resolve key issues in marine big data storage, computing, and analysis, respectively. Applications in Sect. 5 illustrate the significance of marine big data, followed by the conclusion in Sect. 6.

2 Data acquisition

During data acquisition phase, ocean observatories equipped with various sensors are utilized to collect raw data from the ocean. This section will introduce some representative ocean observing platforms and projects for marine data acquisition.

2.1 Ocean observing platforms

The ocean observatories are collections of platforms capable of carrying sensors to collect data over certain spatiotemporal scales. These platforms include ships, satellites, and a range of Eulerian and Lagrangian systems [13].

- *Ships* have been the primary tool for oceanographers for centuries and will remain a central piece of infrastructure in the foreseeable future. Ships available to the ocean observing include both global class vessels and smaller coastal vessels. The capabilities of the ships have improved significantly in the station holding and dynamic positioning, multi-beam and side-scan sonar systems, and more complex sensors and instrumentation becoming routine tools when at sea.
- *Satellites* constitute the most essential oceanographic technology innovation in modern times. They are the new tools for understanding various ocean processes and land–air–sea interactions over decadal time scales. Satellite data, fundamental to weather and ocean state prediction, have revealed new phenomena over critical spatiotemporal scales which were previously inaccessible using only in situ observing data.
- *Seafloor electro-optic cables* with high bandwidth and sustained power offer potential means for providing sustained observation in the ocean. Seafloor cables have been deployed off the coasts of the USA, Canada, Japan, Europe, and China. These cables have successfully been used to study a wide range of topics such as seafloor seismicity, tsunamis, seafloor dynamics, coastal upwelling ecosystem productivity, ocean turbulence, gas hydrates.
- *Drifters and Floats* are passive, battery powered Lagrangian platforms used in creating surface and subsurface maps of ocean currents and ocean properties, respectively. These platforms are relatively inexpensive so that thousands of these platforms can be deployed at sea by regular crews. Measurements are normally made hourly, and the data are transmitted by satellite.
- *Moorings* provide the means to deploy sensors at fixed depths between the seafloor and the sea surface and to deploy packages that profile vertically at one location by moving up and down along the mooring line/cable or by winching themselves up and down from their point of attachment to the mooring. They will continue to be a key element of ocean observing infrastructure that provides high-frequency fixed location subsurface data to supplement the spatial data collected by ships, autonomous underwater vehicles, and satellite remote sensing.
- *Gliders* are a type of autonomous underwater vehicle using buoyancy-based propulsion to convert vertical motion to horizontal motion. Due to very low power consumption, gliders provide data over large spatiotemporal scales, with missions lasting over half a year and over 3500 km of range. They navigate with the help of periodic surface GPS fixes, pressure sensors, tilt sensors, and magnetic compasses.
- *AUVs* provide much-needed flexibility in ocean observations as they allow for the movement of sensors through the water in three dimensions. Unlike gliders, AUVs can move against most currents nominally at 3–5 knots. Therefore, they can systematically and synoptically survey a particular line, area, and/or volume. Like gliders, AUVs relay data and mission information to shore via satellite. The endurance of AUVs depends on the size of the vehicle as well as the power consumption, allowing them to run continuous missions of a day or more with ranges of 70–240 km.

2.2 Ocean observing projects

The dream of long-term observation in the ocean has explored for more than 20 years. Many countries and organizations have given their contributions to establish global, regional, or local ocean observing systems by using various platforms with multiple sensors onboard. Next, we introduce several national or international projects for long-term ocean observation.

- *Argo* [14] is a global array of more than 3000 free-drifting profiling floats that collect high-quality temperature and salinity profiles from the upper 2000 m of the ice-free global ocean and currents from intermediate depths. This allows, for the first time, continuous monitoring of the temperature, salinity, and velocity of the upper ocean. Deployments began in 2000 and national programs need to provide about 800 floats per year to maintain the Argo array. The broad-scale global array has already grown to be a major component of the ocean observing system. It builds on other upper-ocean ocean observing networks, extending their coverage in space and time, their depth range and accuracy, and enhancing them through the addition of salinity and velocity measurements. It is the sole source of global subsurface datasets used in all ocean data assimilation models and reanalyses.
- *Global Ocean Observing System (GOOS)* [15] in its present form was created in 1991, when the UN Intergovernmental Oceanographic Commission (IOC)'s Technical Committee for Ocean Processes and Climate (TC/OPC) agreed that the ocean observing system concept should be broadened to include physical, chemical, and biological coastal ocean monitoring. GOOS is a permanent global system for observations, modeling, and analysis of marine and ocean variables to support operational ocean services worldwide. GOOS provides accurate descriptions of the present state of the oceans, including living resources, continuous forecasts of the future conditions of the sea for as far ahead as possible, and the basis for forecasts of climate change. GOOS is made of many observation platforms: 3000 Argo floats, 1250 drifting buoys, 350 embarked systems on commercial or cruising yachts, 100 research vessels, 200 marigraphs and holographs, 50 commercial ships, 200 moorings in open sea.
- *Ocean Networks Canada (ONC)* [16], an initiative of the University of Victoria, operates the world-leading NEPTUNE and VENUS cabled ocean observatories in the northeast Pacific Ocean off Canada's west coast. In addition, smaller-scale community observatories are located in the Arctic at Cambridge Bay, Nunavut, and Mill Bay, British Columbia, with more installations under development along the BC coast and across Canada. Its goals are to deliver science and information for good ocean management and responsible ocean use for the benefit of Canadians. ONC cabled observatories collect data that help scientists and leaders make informed decisions about coastal earthquakes and tsunamis, climate change, coastal management, conservation, and marine safety.
- *US Integrated Ocean Observing System (US IOOS)* [17] was approved to be founded according to the Integrated Coastal and Ocean Observation System Act of 2009. IOOS is a national–regional partnership working to provide new tools and forecasts to improve safety, enhance the economy and protect environment. It is a vital tool for tracking, predicting, managing, and adapting to changes in the ocean, coastal, and Great Lakes environment. IOOS observing systems consist of sensors that collect marine data and technology that sends the data to a data collection center, often with satellite telemetry. Observing systems come in all sizes, from global scale systems collecting information on climate down to a local system focused on a single estuary. The US GOOS Regional Alliance is the GOOS IOC interface to the US IOOS.
- *Ocean Observatories Initiative (OOI)* [3] was approved by the National Science Board as a potential Major Research Equipment and Facilities Construction project in 2000. This National Science Foundation-funded OOI is an integrated infrastructure project composed of science-driven platforms and sensor systems that measure physical, chemical, geological, and biological properties and processes from the seafloor to the air–sea interface. The OOI has transformed research of the oceans by establishing a network of interactive, globally distributed sensors with near real-time data access, enhancing our capabilities to address critical issues such as climate change, ecosystem variability, ocean acidification, and carbon cycling. The design of the OOI enables multiple scales of marine observations integrated into one observing system. The coastal assets of the OOI expand existing observations off both US coasts, creating focused, configurable observing regions. Cabled observing platforms "wire" a single region in the Northeast Pacific Ocean with a high-speed optical and high power grid. The global component addresses planetary-scale changes via moored open-ocean buoys linked to shore via satellite.

3 Data storage

The collected marine data will be transmitted to a data storage infrastructure for further processing and analysis. Long-term sustained and multi-source data acquisition leads to the rapid expansion and complexity of data. It raises huge challenges in storage and processing of these data [18]. The datasets stored at the data center come from many different sensors hosted on remote sensing or in situ platforms. To optimize the system and considering storage capability, and speed response, the metadata and some types of data are stored in relational databases, and some other types of data are stored in files [19]. Normally, data types with a wide range of parameters but not too much data, such as nutrients, pollutants, and any other sample measurements, are stored in relational databases. However, data types with few parameters but huge volume of data, such as CTD (conductivity, temperature, and depth), ADCP (acoustic Doppler current profiler) and imagery sensors, are stored in binary, ASCII, or image files. Ocean observation is quite complicated and task oriented. Even for acquiring the same oceanographic parameter, e.g., temperature, different observing platforms with different sensors may be chosen according to different task requirements of spatiotemporal scales. Therefore, the acquired data of the same oceanographic parameter may have different data formats and need to be stored in different type of databases.

3.1 Storage foundation

Existing massive storage technologies can be classified as direct-attached storage (DAS), network-attached storage (NAS), and storage area network (SAN). Various hard disks directly connect with servers in DAS, and data management is server centric. However, due to its low scalability, DAS is mainly used in personal computers and small-sized servers. NAS directly connects to a network through a hub or switch through TCP/IP protocols, and data are transmitted in the form of files. SAN is especially designed for data storage with a scalable and bandwidth intensive network. From the organization of a data storage system, DAS, NAS, and SAN all can be divided into three parts: disk array, connection and network sub-systems, and storage management software [20].

File systems, the bottom level in storage mechanisms, are the foundation of the applications at upper levels. Many companies and researchers have their solutions to meet the different demands for storage of big data. For example, Google's GFS is an expandable distributed file system to support large-scale, distributed, data-intensive applications [21]. HDFS [22] and Kosmosfs are derivatives of open source codes of GFS. Microsoft developed Cosmos to support its search and advertisement business [23]. Facebook utilizes Haystack to store the large amount of small-sized photos [24].

3.2 NoSQL databases

Traditional relational databases cannot meet the challenges on categories and scales brought about by marine big data. NoSQL databases are becoming the core technology for big data storage. NoSQL databases feature flexible modes, support for simple and easy copy, simple API, eventual consistency, and support of large volume data [20]. This section will introduce three main NoSQL databases based on different data models: key-value databases, column-oriented databases, and document-oriented databases.

3.2.1 Key-value databases

Key-value databases are constituted on a simple data model, and data are stored corresponding to key-values. Every key is unique, and customers may input queried values according to the keys. Such databases feature a simple structure, and the modern key-value databases have higher expandability and shorter query response time than relational databases. Over the past few years, many key-value databases have appeared as motivated by Amazon's Dynamo system.

- *Dynamo* [25] is a highly available and expandable distributed key-value data storage system. It is used to store and manage the status of some core services, which can be realized with key access, in the Amazon e-Commerce Platform. The public mode of relational databases may generate invalid data and limit data scale and availability. However, Dynamo can resolve these problems with a simple key–object interface constituted by simple reading and writing operations. Dynamo achieves elasticity and availability through the data partition, data copy, and object edition mechanisms.

3.2.2 Column-oriented databases

Column-oriented databases store and process data according to columns other than rows. Both columns and rows are segmented in multiple nodes to realize expandability. Many column-oriented databases are mainly inspired by Google's BigTable.

- *BigTable* [26] is a distributed, structured data storage system, which is designed to process the large-scale (PB class) data among thousands commercial servers. The basic data structure of BigTable is a multi-dimensional sequenced mapping with sparse, distributed, and persistent storage. Indexes of mapping are row key, column key, and timestamps, and every value in mapping is an unanalyzed byte array. Each row key in BigTable is a 64-KB character string. By lexicographical order, rows are stored and continually segmented into Tablets for load balance. The columns are grouped according to the prefixes of keys, and thus forming column families. The timestamps are 64-bit integers to distinguish different editions of cell values.

3.2.3 Document-oriented databases

Compared with key-value storage, document storage can support more complex data forms. Since documents do not follow strict modes, there is no need to conduct mode migration. In addition, key-value pairs can still be saved. MongoDB, SimpleDB, and CouchDB are three important representatives of document storage systems [20].

- *MongoDB* [27] is an open-source and document-oriented database. MongoDB stores documents as Binary JSON (BSON) objects, which is similar to object. Every document has an ID field as the primary key. Query in MongoDB is expressed with syntax similar to JSON. The system allows query on all documents, including embedded objects and arrays. MongoDB supports horizontal expansion with automatic sharing to distribute data among thousands of nodes by automatically balancing load and failover.
- *SimpleDB* [28] is a distributed database and is a Web service of Amazon. Data in SimpleDB are organized into various domains in which data may be stored, acquired, and queried. Domains include different properties and name/value pair sets of projects. This system does not support automatic partition and thus could not be expanded with the change of data volume. SimpleDB allows users to query with SQL. It is worth noting that SimpleDB can assure eventual consistency but does not support to multi-version concurrency control (MVCC).
- *Apache CouchDB* [29] is a document-oriented database written in Erlang. Data in CouchDB are organized into documents consisting of fields named by keys/names and values, which are stored and accessed as JSON objects. Every document is provided with a unique identifier. CouchDB utilizes the optimal copying to obtain scalability without a sharing mechanism. The consistency of CouchDB relies on the copying mechanism. CouchDB supports MVCC with historical Hash records.

3.3 In-memory databases

To optimize the application performance, data centers not only scale their sizes but also change system architectures with a particular focus on storing and retrieving large datasets more quickly. Accessing data stored in secondary devices is time consuming. Therefore, it is highly unlikely that a high-performance application would be able to perform jobs efficiently using disk-based system architectures, such as Hadoop and GFS. Notable trends are the growth of in-memory databases and the widespread adoption of flash SSDs in data centers [30]. In-memory databases primarily rely on DRAM main memory for data storage. They are orders of faster than disk-optimized databases in typical data analytics queries. New databases with simpler data models (often referred to as “NoSQL” or “NewSQL”) become popular for applications that do not require rich RDBMS (relational database management system) functionalities. These systems offer superior scalability and a low response time. Ever increasing main memory capacities have fostered the development of in-memory database systems [31]. For example, CedCom caches data in main memory [32], which combines the power of cache-only memory architecture (COMA) and the structural principle of Hadoop. Stanford University’s RAMClouds aim to build a cluster scale storage system entirely with DRAM [33].

4 Data computing and analysis

Due to the multi-source, massive, heterogeneous, and dynamic characteristics of application data involved in a distributed environment, one of the most important characteristics of big data is to carry out computing on the petabyte (PB), even the exabyte (EB)-level data with a complex computing process [34]. Therefore, utilizing a parallel computing infrastructure to efficiently analyze and mine the distributed data are the critical goals for big data processing. In this section, we introduce some representative computing infrastructures, methods, and tools for big data analysis.

4.1 Computational model

Big data are generally stored in hundreds and even thousands of commercial servers. Thus, the traditional parallel models, such as Message Passing Interface (MPI) and Open Multi-Processing (OpenMP), may not be adequate to support such large-scale parallel programs. Recently, some proposed parallel programming models effectively improve the performance of NoSQL and reduce the performance gap to relational databases. Therefore, these models have become the cornerstone for the analysis of massive data [20].

- *MapReduce* [35] is a simple but powerful programming model for large-scale computing using a large number of clusters of commercial PCs to achieve automatic parallel processing and distribution. In MapReduce, computing model only has two functions, i.e., Map and Reduce, both of which are programmed by users. The Map function processes input key-value pairs and generates intermediate key-value pairs. Then, MapReduce will combine all the intermediate values related to the same key and transmit them to the Reduce function, which further compress the value set into a smaller set. MapReduce has the advantage that it avoids the complicated steps for developing parallel applications, e.g., data scheduling, fault tolerance, and internode communications. The user only needs to program the two functions to develop a parallel application.

- *Dryad* [36] is a general-purpose distributed execution engine for processing parallel applications of coarse-grained data. The operational structure of Dryad is a directed acyclic graph, in which vertexes represent programs and edges represent data channels. Dryad executes operations on the vertexes in clusters and transmits data via data channels, including documents, TCP connections, and shared-memory FIFO. During operation, resources in a logic operation graph are automatically mapped to physical resources. Dryad allows vertexes to use any amount of input and output data, while MapReduce supports only one input and output set.
- *Pregel* [37] facilitates the processing of large-sized graphs, e.g., analysis of network graphs and social networking services. A computational task is expressed by a directed graph constituted by vertexes and directed edges. Every vertex is related to a modifiable and user-defined value, and every directed edge related to a source vertex is constituted by the user-defined value and the identifier of a target vertex. When the graph is built, the program conducts iterative calculations, which is called supersteps among which global synchronization points are set until algorithm completion and output completion.
- *All-Pairs* [38] is a system specially designed for biometrics, bio-informatics, and data mining applications. It focuses on comparing element pairs in two datasets by a given function. All-Pairs can be expressed as three-tuples (Set A, Set B, and Function F), in which Function F is utilized to compare all elements in Set A and Set B. The comparison result is an output matrix M, which is also called the Cartesian product or cross-join of Set A and Set B.

4.2 Data analysis

Data analysis is the final and the most important phase in the value chain of big data, with the purpose of extracting potential useful values and providing suggestions or decisions. However, data analysis is a broad area, which frequently changes and is extremely complex. Many traditional data analysis methods may still be utilized for big data analysis, such as cluster analysis, factor analysis, correlation analysis, regression analysis, A/B testing, statistical analysis, data mining. Some big data analysis methods can be used to speed up the extraction of key information from massive data. At present, the main processing methods of big data include bloom filter, hashing, index, triel, parallel computing, etc.

For marine data analysis applications, data mining is an essential method to extract hidden, unknown, but potentially useful information and knowledge from massive, incomplete, noisy, fuzzy, and random data. In 2006, The IEEE International Conference on Data Mining (ICDM) series identified ten most influential data mining algorithms [39], including C4.5, k-means, SVM, Apriori, EM, PageRank, AdaBoost, kNN, Naive Bayes, and CART. These ten algorithms cover classification, clustering, regression, statistical learning, association analysis, and linking mining, all of which are the most important topics in data mining research and development. To adapt to the multi-source, uncertain, dynamic marine big data, existing data mining methods should be expanded in many ways, including efficiency improvement of single-source knowledge discovery methods [40], mining frequent or interested patterns from uncertain data [41, 42], designing a data mining mechanism from a multi-source perspective [43], as well as the study of dynamic data mining methods and the analysis of stream data [44].

Parallel processing has been the mainstream of designing efficient data-processing platforms so that data could be processed in a distributed and parallel manner, improving the throughput of data processing [45]. MapReduce is the most representative paradigm. Modern research on big data analysis has focused mostly on employing the MapReduce programming paradigm and the Hadoop Ecosystem, giving rise to a number of DBMSs that can be deployed in a distributed cloud-based environment [46], such as Pig [47] and Hive [48].

After algorithm parallelization, the traditional analysis software tools will have the ability of big data processing. Das et al. [49] integrated R, an open-source statistical analysis tool, and Hadoop to improve the weak scalability of traditional analysis tool and poor analysis capabilities of Hadoop. The in-depth integration pushes data computation to parallel processing, which enables powerful deep analysis capabilities for Hadoop. Standard Weka, an open-source machine learning and data mining tool, can only run on a single machine with a limitation of 1-GB memory. Wegener et al. [50] integrated Weka and MapReduce to break through the limitations, taking the advantage of parallel computing to handle more than 100-GB data on MapReduce clusters. In recent years, extracting valuable information and insightful knowledge from big data has become an urgent need in many disciplines. Due to its high impact in many areas, more systems and analytical tools have been developed for big data analysis, such as Apache Mahout, MOA, SAMOA, and Vowpal Wabbit [51].

5 Applications of marine big data

Marine big data are fundamental to a variety of research fields in biology, earth science, and ocean and atmospheric science. This section will give some examples of physical, chemical, geological, and biological applications to demonstrate the potential uses of marine big data.

5.1 Physical application

- *El Niño* is a warming of surface waters in the eastern tropical Pacific Ocean (shown in Fig. 1). Together with, *La Niña*, these make up two of the three states of the constantly changing El Niño/Southern Oscillation (ENSO) that can affect weather patterns around the globe. When ocean and atmospheric conditions in one part of the world change as results of ENSO or any other oscillation, the effects are often felt around the world. The rearrangement of atmospheric pressure, which governs wind patterns, and sea-surface temperature, which affects both atmospheric pressure and precipitation patterns, can drastically rearrange regional weather patterns, occasionally with devastating results. Extreme climate events are often associated with positive and negative ENSO events. Severe storms and flooding have been known to ravage areas of South America and Africa, while intense droughts and fires have occurred in Australia and Indonesia during El Niño events.

5.2 Chemical application

- *Ocean acidification* is an emerging global problem. When carbon dioxide (CO_2) is absorbed by seawater, chemical reactions occur that reduce seawater pH, carbonate ion concentration, and saturation states of biologically important calcium carbonate minerals. These chemical reactions are termed ocean acidification. Since the beginning of the Industrial Revolution, the pH of surface ocean waters has fallen by 0.1 pH units, which represents approximately a 30 percent increase in acidity. Future predictions indicate that the oceans will continue to absorb carbon dioxide and become even more acidic. Estimates of future carbon dioxide levels, based on business as usual emission scenarios, indicate that by the end of this century the surface waters of the ocean could be nearly 150 percent more acidic, resulting in a pH that the oceans have not experienced for more than 20 million years [52].

The Global Ocean Acidification Observing Network (shown in Fig. 2) is a collaborative international approach to document the status and progress of ocean acidification in open-ocean, coastal, and estuarine environments, to understand the drivers and impacts of ocean acidification on marine ecosystems, and to provide spatially and temporally resolved biogeochemical data necessary to optimize modeling for ocean acidification. Since sustained efforts to monitor ocean acidification worldwide are only beginning, it is currently impossible to predict exactly how ocean acidification impacts will cascade throughout the marine food chain and affect the overall structure of marine ecosystems.

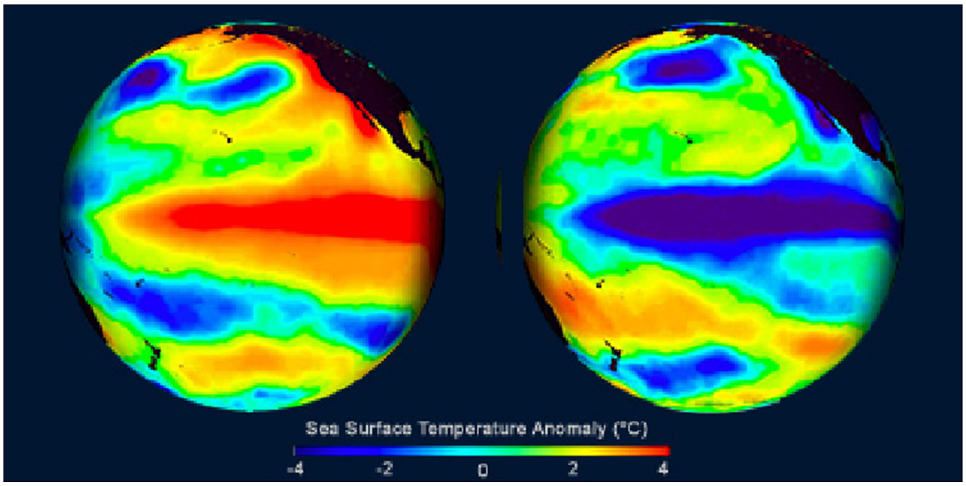

Fig. 1 Global maps of sea surface temperature during El Niño (*left*) and La Niña (*right*) episodes (*source*: http://www.whoi.edu/main/topic/el-nino-other-oscillations)

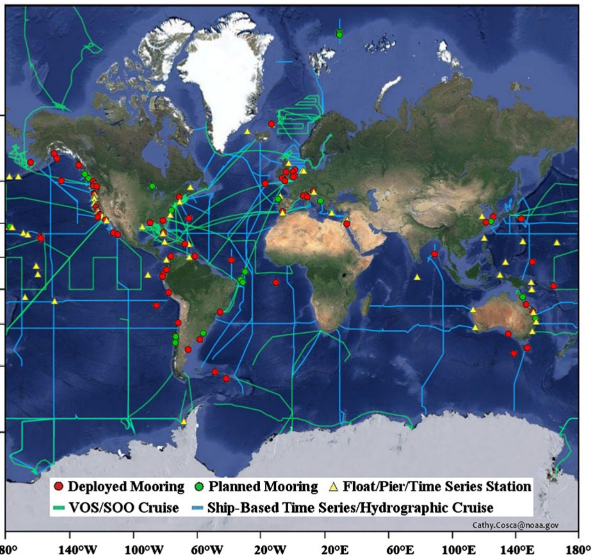

Fig. 2 Map of Global Ocean Acidification Observing Network with ship surveys, moorings, floats, and gliders (*source*: http://www.pmel.noaa.gov/co2/file/GOA_ON_Map)

5.3 Geological application

- *Tsunami* is a massive, fast-moving wave created by an underwater earthquake or landslide. The large volume of water displaced by a sudden movement of the seafloor creates a pulse in the ocean that races out from its source at a speed of up to 500 miles per hour and extends thousands of feet below the surface. Although rare, tsunamis like those that occurred in March 2011 in Japan and December 2004 around the Indian Ocean were tragic reminders of the destructive power of the ocean. As a result, governments of countries surrounding the Pacific and Indian Oceans, with help from scientists from around the world, continuously monitor the ocean bottom for possible tsunami-producing seismic activity and the fast-moving signs of tsunamis in the open ocean. Even a few minutes' warning can mean the difference between wide-scale catastrophe and saving hundreds or thousands of lives.

On April 1 at 4:46:45 PM Pacific Daylight Time (23:46:45 UTC), a magnitude 8.2 earthquake occurred off Chile's Pacific coastline, according to the US Geological Survey. Ocean Networks Canada instrumentation captured both ground shaking and a very small tsunami as they crossed the northeast Pacific (shown in Fig. 3).

5.4 Biological application

- *Biodiversity* refers to the variety of life, encompassing variation at all levels of complexity—genetic, species, ecosystems, and biomes—and including functional diversity and diversity across ecosystems. The maintenance of coastal and marine biodiversity is critical to sustained ecosystem and human health and resilience in a globally changing environment. The condition of marine biodiversity offers a proxy for the status of ocean and coastal ecosystem health and ability to provide ecosystem services such as food, oxygen, socioeconomic benefits that support livelihoods, and a stable climate. Thus, managing our marine resources in a way that conserves existing marine biodiversity would help address other ocean management objectives [53].

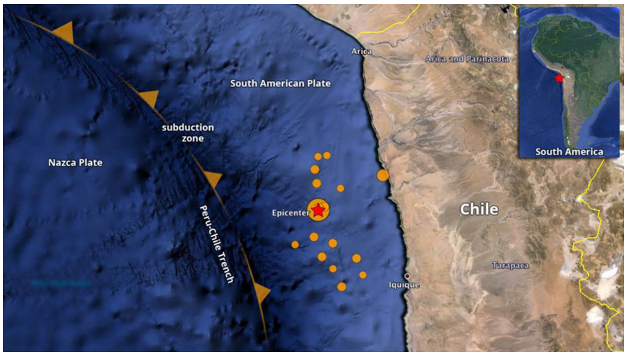

Fig. 3 Map of the epicenter and 16 aftershocks along the subduction zone between the Nazca and South American plates, April 1, 2014 (Source: http://www.oceannetworks.ca/tsunami-alert-follows-82-quake-chile)

A case study is an assessment of reef fish population. This assessment is used to establish reef fish Annual Catch

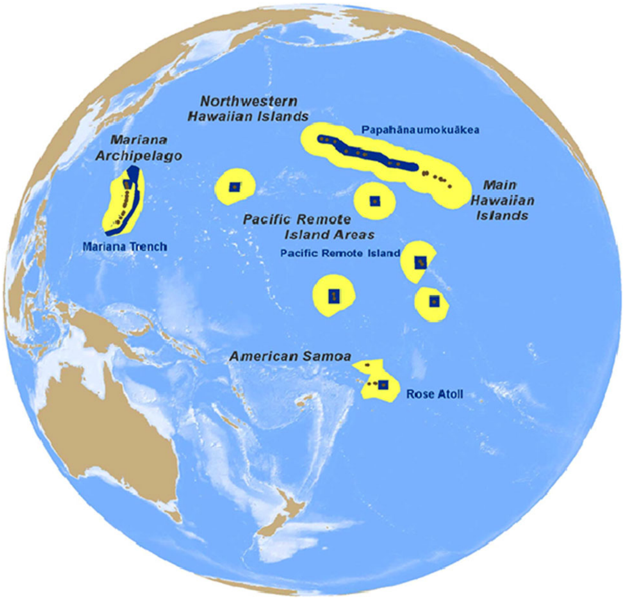

Fig. 4 Locations of the reef areas surveyed by Pacific Islands Fisheries Science Center/Coral Reef Ecosystem Division (Source: http://www.ioos.noaa.gov/biological_observations/welcome.html)

Limits as defined by the Magnuson-Stevens Fishery Conservation and Management Reauthorization Act (MSRA). The biological observations used for this project are species presence/absence/abundance and life history data for reef fishes in the Hawaiian Archipelago and other locations in the Pacific region (shown in Fig. 4).

6 Conclusion

Ocean science is entering into big data era with the exponentially growth of information technology and advances in ocean observatories. However, marine big data are still in its infancy. Many key technical issues, such as big data storage, computing model, analysis method, and application system supporting decision making should be fully investigated. Some challenges need to be resolved in the future work.

- *Infrastructure* Various ocean observatories are collecting and transmitting data continuously. Data quantity reaches to an unprecedented scale that will surpass the storage and processing capacities of existing infrastructures. A traditional data-centric infrastructure, in which a central data management system ingests data and serves them to users on a query basis, is insufficient to accomplish the range of scientific tasks, e.g., collecting real-time data, analyzing data and modeling the ocean on multiple scales and enabling adaptive experimentation within the ocean. The increasingly growing data and its real-time requirement cause problems of how to store and manage such huge heterogeneous datasets with moderate requirements on hardware and software infrastructure.
- *Data transfer* Marine big data are often acquired and stored at different locations. Meanwhile, data volumes are continuously growing. PB- or EB-level data transfer may be involved in data acquisition, transmission, storage, and other spatial transformations. Data transfer usually incurs high costs, which is the bottleneck for big data computing. For example, typical data mining algorithms require all data to be loaded into the main memory. Even if we do have a super large main memory to hold all data for computing, moving such a huge amount of data across different locations is too expensive due to intensive network communication and other I/O costs. Data transfer time is far greater than its computing time. So improving the transfer efficiency is a key issue to improve computing in big data applications.
- *Data quality* Data quality is mainly manifested in its accuracy, completeness, redundancy, and consistency. Uncertainty and incompleteness are defining features for marine big data due to inaccurate data readings and collections. Unexpected transmission or computing errors may also restrict data quality. With the huge volume of generated data, the fast velocity of arriving data, and the large variety of heterogeneous data, the quality of marine big data is far from perfect. Poor quality data can have serious consequences on the results of data analyses, which will influence data utilization, wasting transmission, storage, and computing resources. Although some methods have been adopted to improve data quality, this problem has not been well resolved. Therefore, data quality management remains a challenging research field to detect and repair erroneous data in a scalable and timely manner.
- *Ocean analytics* Ocean analytics is an exciting new way to distill and exploit the vast amount of marine data available from in situ or remote sensing observatories. It can be designed for modeling and forecasting of both short-term high-impact events, such as earthquakes and tsunamis, and long-term large-area events, such as ocean acidification and global warming. It can also be applied to a multitude of different decision support applications that use large amounts of data and require complex calculations. Many existing data mining algorithms do not scale beyond datasets of a few million elements or cannot tolerate the statistical noise and gaps in marine data. New developed analytical algorithms should strengthen scalability, effectiveness, fault tolerance, and parallelization.

The ocean affects our life. In turn, human activities affect the ocean. We need observe, measure, assess, protect, and manage the ocean. Researches on marine big data provide new tools and forecasts of decision making to improve safety, enhance the economy, and protect our environment. Marine big data pave the way for sustainable development and better life.

Acknowledgements This work was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. 61572448, 61379127, 61673357, and by the Shandong Provincial Natural Science Foundation, China under Grant No. ZR2014JL043.

References

1. http://www.whoi.edu/
2. Schofield O, Glenn S, Orcutt J, Arrott M, Meisinger M, Gangopadhyay A, Brown W, Signell R, Moline M, Chao Y, Chien S, Thompson D, Balasuriya A, Lermusiaux P, Oliver M (2010) Automated sensor network to advance ocean science. Eos, Trans Am Geophys Union 91(39):345–346. doi:10.1029/2010EO390001
3. http://oceanobservatories.org/
4. Chave AD, Arrott M, Farcas C, Farcas E, Krueger I, Meisinger M, Orcutt JA, Vernon FL, Peach C, Schofield O, Kleinert JE

(2009) Cyberinfrastructure for the US ocean observatories initiative: enabling interactive observation in the ocean. In: IEEE OCEANS 2009—EUROPE, Bremen, 11–14 May 2009, pp 1–10. doi:10.1109/OCEANSE.2009.5278134

5. Beyer MA, Laney D (2012) The Importance of 'Big Data': a definition. Gartner Inc, Stamford
6. Farcas C, Meisinger M, Stuebe D, Mueller C, Ampe T, Arrott M, Chave A, Farcas E, Graybeal J, Krueger I, Manning M, Orcutt J, Schofield O, Vernon F(2011) Ocean Observatories Initiative Scientific Data Model. In: IEEE OCEANS 2011, Waikoloa, HI, 19–22 Sept. 2011, pp 1–10
7. Park K, Nguyen MC, Won H (2015) Web-based collaborative big data analytics on big data as a service platform. In: IEEE Advanced Communication Technology (ICACT), 2015 17th International Conference on, Seoul, 1–3 July 2015, pp 564–567. doi:10.1109/ICACT.2015.7224859
8. Bellatreche L, Furtado P, Mohania MK (2015) Guest editorial: a special issue in physical design for big data warehousing and mining. Distrib parallel databases 34(3):289–292. doi:10.1007/s10619-015-7182-1
9. Demchenko Y, Laat Cd, Membrey P (2014) Defining architecture components of the Big Data Ecosystem. In: IEEE Collaboration Technologies and Systems (CTS), International Conference on Minneapolis, MN, 19–23 May 2014, pp 104–112. doi:10.1109/CTS.2014.6867550
10. Du Y, Wang Z, Huang D, Yu J (2012) Study of migration model based on the massive marine data hybrid cloud storage. In: IEEE Agro-Geoinformatics (Agro-Geoinformatics), First International Conference on, Shanghai, 2–4 Aug. 2012, pp 1–4. doi:10.1109/Agro-Geoinformatics.2012.6311684
11. Huang D, Zhao D, Wei L, Wang Z, Du Y (2015) Modeling and analysis in marine big data: advances and challenges. Math Probl Eng. doi:10.1155/2015/384742
12. Yang K, Jia X, Ren K, Xie R, Huang L (2014) Enabling efficient access control with dynamic policy updating for big data in the cloud. In: IEEE INFOCOM, 2014 Proceedings IEEE, Toronto, ON, April 27 2014-May 2 2014, pp 2013–2021. doi:10.1109/INFOCOM.2014.6848142
13. Schofield O, Glenn SM, Moline MA, Oliver M, Irwin A, Chao Y, Arrott M (2013) Ocean Observatories and Information: building a global ocean observing network. In: Orcutt J (ed) Earth system monitoring: selected entries from the encyclopedia of sustainability science and technology. Springer, New York, pp 319–336. doi:10.1007/978-1-4614-5684-1_14
14. http://www.argo.net/
15. http://www.ioc-goos.org/
16. http://www.oceannetworks.ca/
17. http://www.ioos.noaa.gov/
18. Siriweera THAS, Paik I, Kumara BTGS, Koswatta KRC (2015) Intelligent Big Data Analysis Architecture Based on Automatic Service Composition. In: IEEE Big Data (BigData Congress), 2015 IEEE International Congress on, New York, NY, June 27 2015–July 2 2015, pp 276–280. doi:10.1109/BigDataCongress.2015.46
19. Antonia C, Andrei N, María-Jesús G (2011) DAMAR: Information management system for marine data. In: OCEANS, 2011 IEEE—Spain, Santander, 6–9 June 2011. IEEE, pp 1–6. doi:10.1109/Oceans-Spain.2011.6003456
20. Chen M, Mao S, Liu Y (2014) Big data: a survey. Mob Netw Appl 19(2):171–209. doi:10.1007/s11036-013-0489-0
21. Ghemawat S, Gobioff H, Leung S-T (2003) The Google file system. SIGOPS Oper Syst Rev 37(5):29–43. doi:10.1145/1165389.945450
22. Shvachko K, Kuang H, Radia S, Chansler R (2010) The Hadoop Distributed File System. In: IEEE Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on, Incline Village, NV, 3–7 May 2010, pp 1–10. doi:10.1109/MSST.2010.5496972
23. Chaiken R, Jenkins B, Larson P-Å, Ramsey B, Shakib D, Weaver S, Zhou J (2008) SCOPE: easy and efficient parallel processing of massive data sets. Proc VLDB Endow 1(2):1265–1276. doi:10.14778/1454159.1454166
24. Beaver D, Kumar S, Li HC, Sobel J, Vajgel P (2010) Finding a needle in Haystack: Facebook's photo storage. Paper presented at the Proceedings of the 9th USENIX conference on Operating systems design and implementation, Vancouver
25. DeCandia G, Hastorun D, Jampani M, Kakulapati G, Lakshman A, Pilchin A, Sivasubramanian S, Vosshall P, Vogels W (2007) Dynamo: amazon's highly available key-value store. SIGOPS Oper Syst Rev 41(6):205–220. doi:10.1145/1323293.1294281
26. Chang F, Dean J, Ghemawat S, Hsieh WC, Wallach DA, Burrows M, Chandra T, Fikes A, Gruber RE (2008) Bigtable: a Distributed Storage System for Structured Data. ACM Trans Comput Syst 26(2):1–26. doi:10.1145/1365815.1365816
27. Chodorow K (2013) MongoDB: the definitive guide, 2nd edn. O'Reilly Media, Sebastopol
28. Murty J (2008) Programming amazon web services: S3, EC2, SQS, FPS, and SimpleDB. O'Reilly Media, Sebastopol
29. Anderson JC, Lehnardt J, Slater N (2010) CouchDB: The Definitive Guide. O'Reilly Media, Sebastopol
30. Cho S (2015) Fast memory and storage architectures for the big data era. In: IEEE Solid-State Circuits Conference (A-SSCC), 2015 IEEE Asian, Xiamen, 9–11 Nov. 2015, pp 1–4. doi:10.1109/ASSCC.2015.7387515
31. Mühlbauer T, Rödiger W, Seilbeck R, Reiser A, Kemper A, Neumann T (2013) Instant loading for main memory databases. Proc VLDB Endow 6(14):1702–1713. doi:10.14778/2556549.2556555
32. Raynaud T, Haque R, Aït-kaci H (2014) CedCom: A high-performance architecture for Big Data applications. In: IEEE Computer Systems and Applications (AICCSA), 2014 IEEE/ACS 11th International Conference on, Doha, 10–13 Nov. 2014, pp 621–632. doi:10.1109/AICCSA.2014.7073257
33. Ousterhout J, Agrawal P, Erickson D, Kozyrakis C, Leverich J, Mazières D, Mitra S, Narayanan A, Parulkar G, Rosenblum M, Rumble SM, Stratmann E, Stutsman R (2010) The case for RAMClouds: scalable high-performance storage entirely in DRAM. SIGOPS Oper Syst Rev 43(4):92–105. doi:10.1145/1713254.1713276
34. Wu X, Zhu X, Wu G-Q, Ding W (2014) Data mining with big data. IEEE Trans Knowl Data Eng 26(1):97–107. doi:10.1109/TKDE.2013.109
35. Dean J, Ghemawat S (2008) MapReduce: simplified data processing on large clusters. Commun ACM 51(1):107–113. doi:10.1145/1327452.1327492
36. Isard M, Budiu M, Yu Y, Birrell A, Fetterly D (2007) Dryad: distributed data-parallel programs from sequential building blocks. Paper presented at the Proceedings of the 2nd ACM SIGOPS/EuroSys European Conference on Computer Systems 2007, Lisbon, Portugal
37. Malewicz G, Austern MH, Bik AJC, Dehnert JC, Horn I, Leiser N, Czajkowski G (2010) Pregel: a system for large-scale graph processing. Paper presented at the Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, Indianapolis, Indiana, USA, 6–11 June 2010
38. Moretti C, Bulosan J, Thain D, Flynn PJ (2008) All-pairs: An abstraction for data-intensive cloud computing. In: IEEE Parallel and Distributed Processing, 2008. IPDPS 2008. IEEE International Symposium on Miami, FL, 14–18 April 2008, pp 1–11. doi:10.1109/IPDPS.2008.4536311
39. Wu XD, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS, Zhou ZH, Steinbach M,

Hand DJ, Steinberg D (2008) Top 10 algorithms in data mining. Knowl Inf Syst 14(1):1–37. doi:10.1007/s10115-007-0114-2

40. Chang EY, Bai H, Zhu K (2009) Parallel algorithms for mining large-scale rich-media data. Paper presented at the Proceedings of the 17th ACM international conference on Multimedia, Beijing, China
41. Leung CK-S, Hayduk Y (2013) Mining frequent patterns from uncertain data with MapReduce for big data analytics. In: 18th International Conference on Database Systems for Advanced Applications, DASFAA 2013, Wuhan, China, 22–25 April 2013. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer Verlag, pp 440–455. doi:10.1007/978-3-642-37487-6_33
42. Leung CKS, MacKinnon RK, Jiang F (2014) Reducing the Search Space for Big Data Mining for Interesting Patterns from Uncertain Data. In: IEEE Big Data (BigData Congress), 2014 IEEE International Congress on, Anchorage, AK, June 27 2014–July 2 2014, pp 315–322. doi:10.1109/BigData.Congress.2014.53
43. Xindong W, Shichao Z (2003) Synthesizing high-frequency rules from different data sources. IEEE Trans Knowl Data Eng 15(2):353–367. doi:10.1109/TKDE.2003.1185839
44. Domingos P, Hulten G (2000) Mining high-speed data streams. Paper presented at the Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, Boston
45. Zhu W, Cui P, Wang Z, Hua G (2015) Multimedia Big Data Computing. IEEE Multimedia 22(3):96-c3. doi:10.1109/MMUL.2015.66
46. Kantere V A (2014) Holistic Framework for Big Scientific Data Management. In: IEEE Big Data (Big Data Congress), 2014 IEEE International Congress on, Anchorage, AK, June 27 2014–July 2 2014, pp 220–226. doi:10.1109/BigData.Congress.2014.39
47. Olston C, Reed B, Srivastava U, Kumar R, Tomkins A (2008) Pig latin: a not-so-foreign language for data processing. Paper presented at the Proceedings of the 2008 ACM SIGMOD international conference on Management of data, Vancouver, Canada
48. Thusoo A, Sarma JS, Jain N, Shao Z, Chakka P, Zhang N, Antony S, Liu H, Murthy R (2010) Hive—a petabyte scale data warehouse using Hadoop. In: IEEE Data Engineering (ICDE), 2010 IEEE 26th International Conference on Long Beach, CA, 1–6 March 2010 , pp 996–1005. doi:10.1109/ICDE.2010.5447738
49. Das S, Sismanis Y, Beyer KS, Gemulla R, Haas PJ, McPherson J (2010) Ricardo: integrating R and Hadoop. Paper presented at the Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, Indianapolis, Indiana, USA
50. Wegener D, Mock M, Adranale D, Wrobel S (2009) Toolkit-Based High-Performance Data Mining of Large Data on MapReduce Clusters. In: IEEE Data Mining Workshops, 2009. ICDMW ‘09. IEEE International Conference on Miami, FL 6 Dec. 2009, pp 296–301. doi:10.1109/ICDMW.2009.34
51. Lin YC, Wu C-W, Tseng VS Mining high utility itemsets in big data. In: 19th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2015, Ho Chi Minh City, Vietnam, May 19–22, 2015 2015. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer Verlag, pp 649-–661. doi:10.1007/978-3-319-18032-8_51
52. http://www.pmel.noaa.gov/co2/
53. Palumbi SR, Sandifer PA, Allan JD, Beck MW, Fautin DG, Fogarty MJ, Halpern BS, Incze LS, Leong J-A, Norse E, Stachowicz JJ, Wall DH (2009) Managing for ocean biodiversity to sustain marine ecosystem services. Front Ecol Environ 7(4):204–211. doi:10.1890/070135