Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 384742, 13 pages
http://dx.doi.org/10.1155/2015/384742

Review Article

Modeling and Analysis in Marine Big Data: Advances and Challenges

Dongmei Huang, Danfeng Zhao, Lifei Wei, Zhenhua Wang, and Yanling Du

College of Information, Shanghai Ocean University, Shanghai 201306, China

Correspondence should be addressed to Dongmei Huang; dmhuang@shou.edu.cn

Received 12 August 2014; Revised 5 September 2014; Accepted 15 September 2014

Academic Editor: L. W. Zhang

It is aware that big data has gathered tremendous attentions from academic research institutes, governments, and enterprises in all aspects of information sciences. With the development of diversity of marine data acquisition techniques, marine data grow exponentially in last decade, which forms *marine big data*. As an innovation, marine big data is a double-edged sword. On the one hand, there are many potential and highly useful values hidden in the huge volume of marine data, which is widely used in marine-related fields, such as tsunami and red-tide warning, prevention, and forecasting, disaster inversion, and visualization modeling after disasters. There is no doubt that the future competitions in marine sciences and technologies will surely converge into the marine data explorations. On the other hand, marine big data also brings about many new challenges in data management, such as the difficulties in data capture, storage, analysis, and applications, as well as data quality control and data security. To highlight theoretical methodologies and practical applications of marine big data, this paper illustrates a broad view about marine big data and its management, makes a survey on key methods and models, introduces an engineering instance that demonstrates the management architecture, and discusses the existing challenges.

1. Introduction

Recently, the data volume all over the world is growing at an overwhelming speed, which is acquired by various devices with regard to Internet of Things and Social Networks. In this context, big data emerges and has been investigated extensively so far. In terms of marine field, countries around the world have launched several observing projects, for example, Argo [1], NEPTUNE-Canada [2], GOOS [3], OOI [4], IOOS [5], and so forth, and numerous marine observation satellites [6, 7]. Acquiring marine data by various observing techniques leads to a sharp increase in data volume. For example, Argo [1] has set up four data centers and deployed up to 10231 buoys all over the world, for real-time acquiring marine data like temperature, salinity, acidity, density, and carbon dioxide. Even one data center alone has to process 21954 profile data with 657 active buoys over the whole of last year [8, 9]. The different data collection devices result in various data as well as their format. We denote the diverse data provisions. A marine observation satellite emitted by NASA, named as *Aquarius* [6], records all the element of ocean circulation, temperature, and ingredient and sea surface height every 7 days. Statistically, the data volume collected by *Aquarius* within every 2 months amounts to that collected by survey ships and buoys in 125 years [6]. By the end of year 2012, the annual data volume had been up to 30 PB (1 PB = 1024 ∗ 1024 GB) maintained by NOAA and over 3.5 billion observational files would be gathered together from satellites, ships, aircrafts, buoys, and other sensors each day [7]. As all-round marine observation systems and multiple observing techniques are widely put into service, data volume sharply increases, data type is greatly diversified, and data value is highly delivered, which forms *marine big data*.

Marine big data contains great values and embodies giant academic appeal, which can be transformed into a rich set of information for people to learn, exploit, and maintain the marine. For example, after analyzing the Argo data, it is found that the earth is seeking an intensification of global hydrological cycle [10]. Communities and species distribution can be determined by analysis of acoustic remote sensing data, which works as powerful scientific supporting

evidence to maintain the marine ecological balance [11]. In addition, researches on forecasting and warning of undersea earthquake and tsunami can be successfully preceding, by analyzing observation data concerning seismic activity, faulting activity and midoceanic ridges acquired by Neptune project [12, 13]. In summary, marine big data supports forecasting and warning potential problems in the field of ecology, climate, and disasters and helps decision making.

In order to maximally exploit the value in marine data, it is of great realistic and theoretical significance to study on the management of marine big data concerning data storage, data analysis, quality control, and data security.

At present, almost all the existing researches concentrate on solving general issues about big data management. As a kind of typical big data, marine big data features massiveness, diverse data provisions, high-dimension besides temporality, and spatiality, which brings exceptional challenges and problems. In terms of data storage, there are problems like weak scalability in storage system and dissatisfaction on timeliness. In terms of data analysis, there are still problems like slow processing speed and failure in real-time response. Furthermore, the data available and data security are two features for the marine big data management. In terms of data available, there are some emerging problems like difference of data quality, diversity of data error, and unfixed schema of quality inspection. Additionally, as data security involves in all the process of marine big data management, security in data storage, data access, data computation, data sharing, and data supervision must be considered all over marine big data management. If the above problems cannot be well solved, the value of marine big data would not be fully exploited.

To our best of knowledge, this paper is the first survey on marine big data management. Our contribution is to study on marine big data management architecture, summarize the related methods and models, introduce a practical application to demonstrate the architecture of marine big data management, discuss the facing challenges, and ultimately prospect the research directions of marine big data management.

Organization of the rest paper is arranged as follows. Section 2 covers the source and informal definition of marine big data and provides an overview of the data characteristics. Related methods and models in marine big data management are summarized in Section 3. The project about marine big data management is presented in Section 4. Section 5 describes the facing challenges of marine big data management. Finally, we draw a conclusion.

2. Marine Big Data Management Architecture

2.1. Marine Big Data. There has been no consensus concerning the definition of marine big data. Given *4V* (*volume*, *variety*, *velocity*, and *value*) characters of big data [14], marine big data is informally described as large amount of data which is collected by satellite, aerial remote sensing, stations, ships, and buoys and serving in the marine-related fields. According to corresponding profiles [15–17], we summarize the significant characteristics of marine big data as follows.

(1) Diverse Data Provisions. Marine big data is acquired from widespread sources, such as satellites, aerial remote sensing, stations, ships, buoys, and undersea sensing. Different data sources take diverse data acquisition technologies to capture marine data; however, varieties in data acquisition technology specification, data format, arguments, and observation region make marine big data reveal its characteristic of data type diversity. Data with different diverse data provisions, as well as the various data types, is a significant characteristic of marine big data.

(2) Temporality and Spatiality. Marine big data features strong timeliness and spatial correlation. Only those marine data who contain specific spatial and temporal information will show significant values. The data storage and the data analysis are based on these two attributes. Without these two features, the marine data will be useless.

(3) High Dimension. The marine science involves several disciplines such as physical oceanography, chemical oceanography, biological oceanography, marine environment, and marine economy. Besides temporality and spatiality, every marine data still contains multiple attributes like water temperature, salinity, acidity, density, and velocity according to the various demands. As a result, it is known as high dimension data.

(4) Huge Volume. Since marine data grows at an overwhelming speed, due to its high dimension and real-time (or periodically) data acquisition by existing marine observation projects all over the world, all of these factors form the huge volume of marine big data.

(5) Data Availability. Marine big data also needs the techniques to keep the data's reliability. Once some illegal data injects in the system, we need some techniques to find out using data sampling technique, data quality inspection technique, and automatic restoration technique.

(6) Data Security. Marine big data involves privileged, confidential and strategic data, like long-cycle meteorology and hydrology data helping disaster evaluation and forecasting, marine fisheries and oil-gas distribution data helping marine resource utilization, large-scale reef data, and off-coast data helping military decision making.

2.2. Marine Big Data Management Architecture. Marine big data comes from various data provisions, and its application requirement and data type differ in each other. By analysis of marine big data, the architecture of marine big data management can be illustrated: data provision, data preprocessing, data storage, data analysis, and data application as well as quality control and data security throughout the whole process, specific as Figure 1.

The management architecture of marine big data involves several parts. Marine big data derives from various sources such as satellite, aerial remote sensing, stations, ships, buoys, and undersea sensing. Due to extra complex data structure

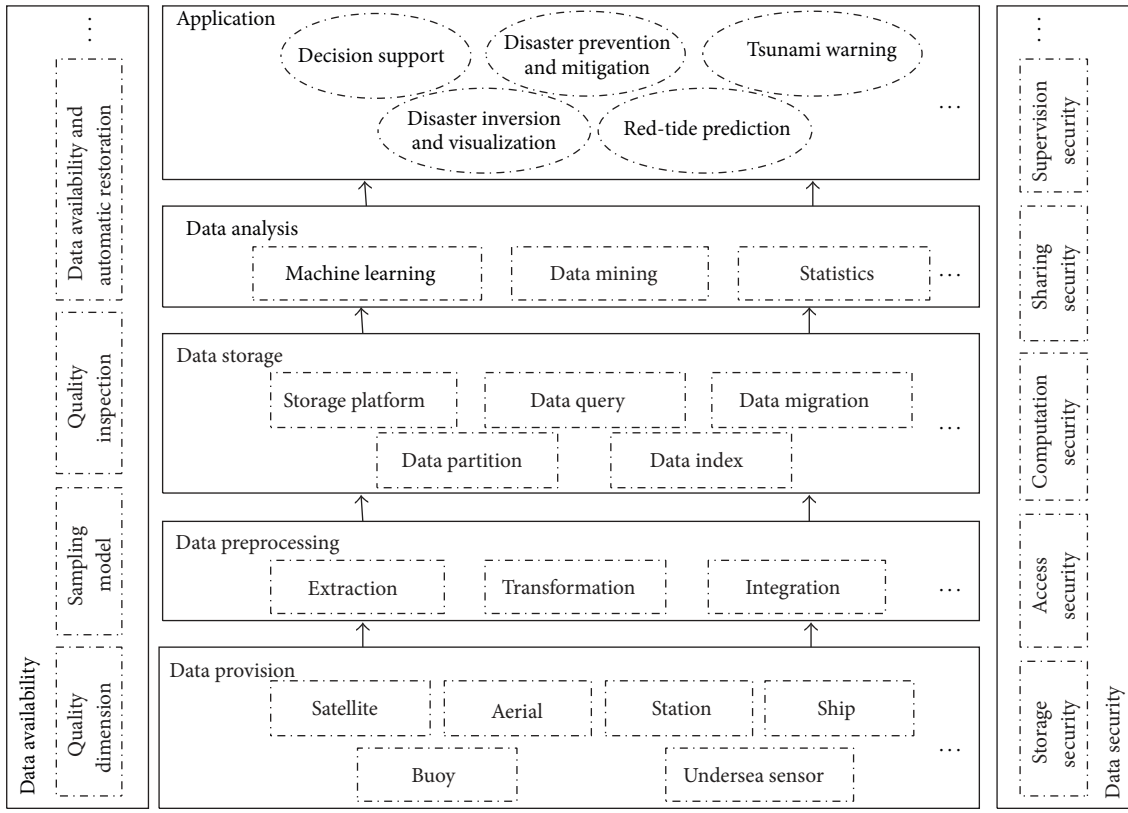

Figure 1: Marine big data management architecture.

and data type characteristics of marine big data, it is essential to perform preprocessing operations, such as data extraction, data transformation, and data integration. At data storage stage, aspects like storage platform, data classification, index building, query, and data migration should be properly taken into account. At data analysis stage, techniques like machine learning, data mining, and statistics are introduced to provide the reliable theoretical basis on applications, including decision support, disaster prevention and mitigation, disaster inversion and visualization modeling, tsunami warning, and red-tide forecasting. Data quality and data security are perceived as the assurance for the whole architecture. Data quality involves quality dimension, sampling model, quality inspection, and data availability and automatic restoration, while data security involves storage security, access security, computation security, sharing security, and supervision security.

3. Methods and Models in Marine Big Data Management

Nowadays, there are a great many of researches on big data management, and a few general technologies have been launched. This section discusses the methods and technologies, with regard to data storage, data analysis, data quality control, and data security in marine big data management.

3.1. Data Storage and Analysis. With the advent and development of cloud computing, new processing frameworks, computing models, and analytical methods emerge as required, which provide technical supports for storage and analysis in big data management. From the view of data storage and data analysis, this section analyzes these key technologies, which are applicable to marine big data with significant characteristics.

3.1.1. Data Storage. Cloud storage is widely applied in big data. Currently there are several cloud storage platforms, including Google Store [18], Amazon S3 [19], Microsoft Azure [20], and IBM Blue Cloud [21, 22]. To make cloud storage play better applicability in sensitive and spatial marine big data, operations like partitioning marine big data by security classification and building suitable index structure should be carried out to raise query efficiency. With the continual accumulation of observing data in data storage system, data should be dynamically migrated, in consideration of characteristics of marine big data. All the above contributes to maximum use of storage system.

Data partition helps to increase execution efficiency of index [23]. In terms of data security, current researches fasten on taking data sensitivity calculation [17], physical isolation [24], and user access restriction [25, 26], to partition data to the corresponding node. Besides, there are some partition methods based on statistical theory, such as clustering-based

data partition [27], sampling-based data partition [28, 29], and adaptive partition based on data distribution [30, 31]. These above methods aims to relieve processing pressure of massive data, avoid data skew, and achieve stable and dynamic data distribution. Data partitioning is a process that a dataset is divided into several fragments according to certain rules and there is no intersection among the various fragments. After the data is divided into a number of data fragments which are stored in clouds, assume that cloud storage is large enough. When a dataset D is uniformly fragmented stored into the n clouds, the information entropy requires

$$I\left(s_1, s_2, \ldots, s_n\right) = -\sum_{i=1}^{n} \frac{s_i}{s} \log_2 \frac{s_i}{s}, \tag{1}$$

where s is the total number of fragments of dataset D and s_i is the number assigned to the Cloud$_i$ fragments. When the value of n is greater, indicating that data is split into more fragments, the greater its entropy.

Index is a powerful technique to improve query efficiency. Cloud storage is a widely accepted distributed storage platform on marine big data. In this case, current researches mainly falls into several classes: hash index [32], tree structure index [33, 34], time-led composite index [35, 36], index dynamically adjusted with data migration [37, 38], and index optimized with parallel processing [39].

To improve query efficiency on cloud storage, it is essential to study on query optimization techniques, so as to relieve computing pressure and improve transmission speed. From the view of algorithm implementation, there are a few improvements, such as sharing history query result as intermediate result [40], adaptively sampling based on data characteristics [41], and extracting representative tuples according to relation compactness [42]. Zadeh introduced the notion of possibility distributions, which acts as a fuzzy restriction on the values that may be assigned to a variable. Given a fuzzy set F and a variable X on U, then the possibility of $X = u$, denoted by $\pi_X(u)$, is defined to be equal to $\mu_F(u)$. The possibility distribution of X on U with respect to F is denoted by

$$\pi_X = \left\{ \frac{\pi_X(u)}{u} \mid u \in U, \pi_X(u) = \mu_F(u) \in [0, 1] \right\}. \tag{2}$$

Additionally, relevant studies still focus on hardware performance improvement, adopting task scheduling [43, 44] to realize efficient parallel processing.

Dynamic data migration on storage platform ensures optimal utilization of storage resource. There are two kinds of traditional data migration methods: one is based on high and low water level method of storage space [44], and the other is based on cache replacement migration algorithm of data access frequency [45, 46]. With the development of storage technology, several different storage patterns have been created. In hierarchical storage, migration model is introduced to support automatic data migration [47]. In multistage storage, CuteMig migration method [48] is involved to realize data migration. In hybrid cloud storage, calculation of data sensitivity and migration function contributes to dynamic data migration [17]. Dremel [49] successes in analyzing massive data in short time and supports data analysis platform over the cloud.

3.1.2. Data Analysis. Considering characteristics like real-time and diversity in data type of marine big data, data analysis should be performed according to data type and analysis target. Hence, adaptive algorithm and model should be taken to ensure the request for real-time data analysis. MapReduce is widely used in numerous big data applications to accelerate the data analysis process. As a result, there is no exception in marine big data application. The paragraph below briefly introduces some representative big data analysis models.

MapReduce is the earliest computing model that Google proposed, which applies to batch processing [50]. MapReduce can be divided into two phase: map phase and reduce phase. Graph is an effective data structure in representing relationships or connections between objects in the real world. Hence, graph computing is a normal computing pattern. Since graph computing involves continuously data updating and numerous message passing, it might impose lots of unnecessary serialization and deserialization overhead using MapReduce. Pregel [51] is another computing model proposed by Google after MapReduce, which is mainly devised to serve graph computing. Its core idea derives from distinguished BSP [52] computing model. Additionally, there exists a PageRank algorithm to reflect the computing quality. The formula is given as follows:

$$\mathrm{PR}(A) = (1 - d) + d\left(\frac{\mathrm{PR}\left(T_1\right)}{C\left(T_1\right)} + \cdots + \frac{\mathrm{PR}\left(T_n\right)}{C\left(T_n\right)} \right), \tag{3}$$

where $T_1, \ldots, T_n$ = Pages that point to page A (citations) and $C(T)$ = number of links going out of T. Dremel [49] successes in analyzing massive data in short time and supports data analysis platform over the cloud, that is, BigQuery [53]. As to its data model, it is based on strongly typed nested records. Its abstract syntax is given by

$$\tau = \mathrm{dom} \mid \left\langle A_1 : \tau\left[* \mid ?\right], \ldots, An : \tau\left[* \mid ?\right] \right\rangle, \tag{4}$$

where τ is an atomic type or a record type. Field i in a record has a name A_i and an optional multiplicity label. Repeated fields $(*)$ may occur multiple times in a record. Optional fields (?) may be missing from the record. Analysis tool, PowerDrill [54], adopts column storage and compress technique to load as much as data into memory. Both PowerDrill and Dremel are big data analysis tools of Google, but they fit into different application scenarios, respectively, and differ in implementation techniques. Dremel is mostly used in analysis of multidatasets, and it can handle up to PB data in several seconds. PoweDrill is mostly applied in analysis of core subset of massive data, and it disposes less data types than Dremel. Since PowerDrill resides data in the memory buffer as much as possible, its processing speed is higher. Microsoft proposed a data analysis model named Dryad [55], which supports applications of Directed

Acycline Graph (DAG), the same as Cascading on Hadoop [56]. The singleton graph is generated from a vertex v as $G = \langle (v), \emptyset, \{v\}, \{v\} \rangle$. A graph can be cloned into a new graph containing k copies of its structure using the $\wedge$ operator where $C = G \wedge k$ is defined as

$$C = \left\langle V_G^1 \oplus \cdots \oplus V_G^K, E_G^1 \cup \cdots \cup E_G^k, I_G^1 \cup \cdots \cup I_G^k, O_G^1 \cup \cdots \cup O_G^K \right\rangle, \tag{5}$$

where $G^n = \langle V_G^n, E_G^n, I_G^n, O_G^n \rangle$ is a "clone" of G containing copies of all of G's vertices and edges, $\oplus$ denotes sequence concatenation, and each cloned vertex inherits the type and parameters of its corresponding vertex in G.

3.2. Data Availability. Facing the quality problems of the uncertainty and inconsistency of marine big data, a scheme of data quality control throughout data management is highly on-demand. So far, academic study on data quality control involves several aspects, including selection of data quality dimensions, design of quality inspection scheme, regulation of quality control standard, and theories of the data usability and data autorestoration.

Quality Dimensions. In essence, data quality is considered as the applicability of data in applications [57] and can be described from five dimensions, including consistency, integrity, timeliness, usability, and credibility [58]. As for spatial data, existing researches put forward five important aspects of data quality evaluation, including spatial accuracy, thematic accuracy, logical consistency, completeness, and lineage [59]. In terms of various quality evaluation methods, spatial data quality is measured as such in ISO/TC211:

$$R = \sum_{i=1}^{k} (C_i \cdot W_i), \tag{6}$$

where R is the result of data quality, $R \in (0.0, 1.0)$; C_i is the accuracy of the ith object, $C_i \in (0.0, 1.0)$; W_i is the weight of the ith object, $W_i \in (0.0, 1.0)$; k is the amount of all kinds of ground objects [60].

Sampling Schemes for Spatial Data. Sampling method is an effective way for processing of massive information, by choosing a small amount of sample to represent the population. The sampling method is efficient with low cost. When spatial samples are not independent, the Bootstrap algorithm introduces two-time sampling technique [61], using the Bag of Little Bootstraps (BLB) functions as follows:

$$s^{-1} \sum_{j=1}^{a} \xi \left(Q_n \left(P_{n,b}^j \right) \right), \tag{7}$$

which has greatly improved the efficiency of data quality evaluation under parallel or distributed computing circumstance. In spatial data sampling, the "Sandwich" sampling model solves the problem of spatial heterogeneity, based on stratified sampling [62, 63] by considering autocorrelation of the spatial objects.

Quality Inspection Schemes for Spatial Data. During the past several years, efforts have been made on quality inspection of marine big data. These studies have put forward an available quality inspection scheme for marine big data, especially for one or a few dimensions. Marine dataset is usually composed of multidimension, multiscale, and multisource. Thus, it is required to propose a quality inspection scheme to inspect the quality of marine big data as a complete, indivisible set [64].

The purpose of quality inspection is to judge whether the data reach the quality levels required by data analysis or data utilization [65]. The principal goal of designing an optimal sampling scheme is to obtain high accuracy of product inspection and to reduce the inspection cost [66]. Current studies have proposed many sampling schemes of quality inspection for industrial product based on statistical theory [67–72]; based on hypergeometric distribution, the accepting probability is calculated as follows:

$$L(p) = \sum_{d=0}^{c} h(d, n, D, N), \tag{8}$$

where d is the actual number of unaccepted data products in the sample, n is the sample size, D is the total number of unaccepted data products in the lot, and N is the lot size.

Thus, the inspection model of marine big data is also brought up:

$$\min_n \varepsilon$$
$$\text{s.t.} \quad \varepsilon = \varepsilon_\varepsilon = \sum_{d=0}^{c} \frac{\binom{N-D}{n-d}\binom{D}{d}}{\binom{N}{n}} - (1 - \alpha), \tag{9}$$
$$(0 < c < n - 1, \varepsilon > 0),$$

where ε is the residual of the accepting probability and α represents the quality demand of data user.

Data Usability. The usability of dataset includes data consistency, data integrity, data accuracy, timeliness, and entity identity [73]. Studies on data consistency are mainly based on description of semantic rules [74] and statistics [75]. The most classic resolution dealing with data integrity is an incomplete data expression system based on conditional table [76]. There are few researches on data accuracy. The most common one is a description method of data accuracy based on possible world semantics. In terms of timeliness, researches mainly fasten on autodetection and autorestoration [77]. Studies on entity identity are based on the detection of entity identity error, including semantic rules and similarity measurement [78].

Data Autodetection and Autorestoration. Studies on data error detection include two aspects, data consistency and entity identity. As for data consistency, studies mainly focus on designing on autodetection algorithm [79] and distributed database detecting method [80]. The purpose of entity identity detection is to maximize the identification accuracy [81] and the recognition efficiency [82]. In terms of studies on

data restoration, traditional functional dependency is used to solve the problem of data inconsistency [83], while data fusion techniques are mostly used for data entity identity issues [84].

3.3. Data Security. According to the challenges in marine big data security, the related researches and development techniques are summarized in the following five aspects as secure data storage, secure data access, secure data computation, secure data sharing, and secure data supervision.

Secure Data Storage. Since the existing data storage security depends on the credibility of the cloud servers, we needs to study the ciphertext-based data storage techniques [85], to resist the administrators of the storage servers and adversary from the server side exposing and tampering data. Besides, it is also necessary to research on the multiauthorities in the access control to reduce the loss due to a single authority compromised by the malicious adversary. In addition, the techniques for data integrity checking [86] and data storage proofing [87] are also essential in the ciphertext-based storage.

Secure Data Access. Marine big data are used for different scenarios and accessed by different users with different roles and different security levels. Traditional access control is no longer suitable for the ciphertext-based storage platform. It is necessary to research the techniques of ciphertext-support data retrieval [88], the fine-grained data access control [89], and supporting the flexible functions such as "and," "or," and "not" logical connectives data access control [90], indexing [91], keyword searching and ranking [92], and similarity searching [93] on the encrypted data to realize the access security.

Secure Data Computation. Since the servers cannot be fully trusted and computation services are often in an outsourcing way, it requires that the input/output should be in an encrypted form for data calculation and data analysis, rather than that the storage ciphertext is decrypted before computation and analysis [94]. In the marine big data computation and analysis, it requires the techniques involving solving the ciphertext-based large scale linear equations [95], analyzing and mining the knowledge from the encrypted data, processing the ciphered images [96], and fully homomorphic encryption/decryption [97] to realize the computation security.

Secure Data Sharing. The marine data sharing security depends on the user's secret key. To keep the data secure sharing and data dissemination in the cloud environment [98], it is inevitable to research the techniques of leakage key tracing like white-box traceability [99, 100] and black-box traceability [101] and access ability revocation [102]. Meanwhile, faced to marine data, it also requires efficient encrypted data sharing and dissemination techniques [103], marine data privacy-preserving techniques [104], and optimized implementation techniques [105] to improve the batch processing ability of marine big data.

Secure Data Supervision. In the data storage, computation, sharing, and dissemination, it needs secure data supervision techniques [106] such as removing illegal data [107], reducing the cost of redundant data [105], checking the completeness of the storage content [87], verifying the correctness of the calculation results [95], and mining the sensitive information and knowledge in the marine data usage. Furthermore, it also requires rules from the government to coordinate the personal privacy preserving and marine big data analyzing [108].

4. Application in Marine Big Data

In terms of the marine big data management architecture, we introduce a practical application of the marine big data—a disaster inversion visualization instance that reproduces a marine disaster happened in Chinese Yellow Sea to show our marine big data.

The disaster results, including latitude and longitude, flow velocity, flow direction, water depth, and height, produced every 10 minutes, involve over 40000 inversion grids. Each monitoring of disaster lasts 5 days, and the disaster data amount alone is up to 4.5 GB. (If we employ more precise data, the data volume will be much huger.) Thus, we choose it as an application since it satisfies all of the characteristics of marine big data. Furthermore, to achieve authenticity and quasireal-time of disaster process, massive data about the geographic locations and continuously rising water level need to be loaded in the disaster visualization, which leads to higher requirements for data transmission, data storage, data analysis, and rendering efficiency.

In this project, we apply hybrid cloud storage architecture, including public cloud and private cloud shown in Figure 2. The project partitions the marine big data in terms of the difference between spatial and temporal attributes. The data with strong timeliness attribute and location related attribute are stored in the private cloud. Public cloud assists to store the rest of the marine data.

Meanwhile, data migration is the key problem in such hybrid cloud storage architecture. Things like data sensitivity, data access frequency, data time length, and data size should be fully considered when performing data migration. To improve query efficiency in the cloud, we use the query optimization technique and improve transmission speed. We take the migration algorithm [17] to help to lower the management cost without sacrificing to slow down the data access speed. The migration function is the key of migration algorithm shown as follows:

$$M(D) = \sum_{i=1}^{n} \frac{1}{T_i} \times \sum_{k=1}^{n} f_k \times \frac{1}{S}, \tag{10}$$

where T_i represents time-length of the ith access of the marine dataset D, f_k represents access frequency of marine dataset D over the period of T_k, and S is the size of marine dataset D.

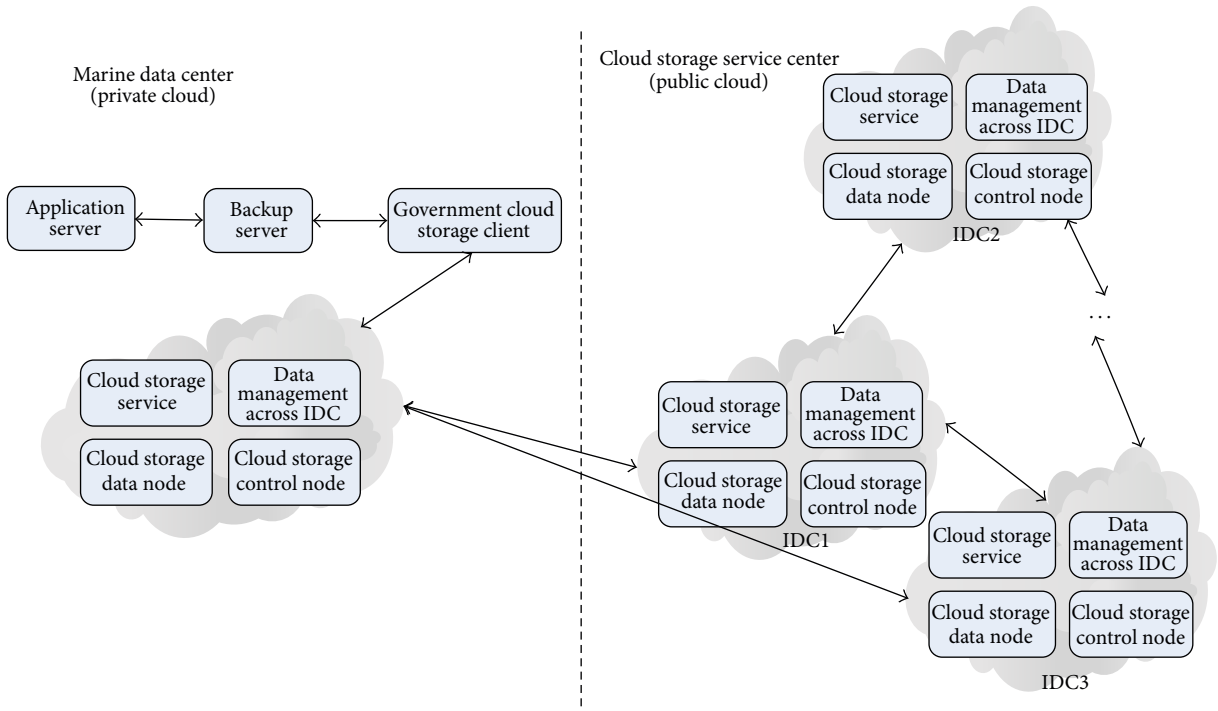

Figure 2: Hybrid cloud storage architecture.

The system performs migration by judging the value of the formula (10).

In this hybrid cloud storage platform, to keep the data availability, we set the data quality inspection model [68] to fit in marine big data and improve the data usage and reliability. The acceptance number c directly affects the inspection result. Given a sample size n, we can obtain c from the formula shown as follows:

$$c = -\frac{1}{2} + \frac{n}{\log\left(p_2/p_1\right)/\log\left(q_1/q_2\right) + 1}, \tag{11}$$

where p_1 represents accepting probability and p_2 is rejecting probability. And $q_1 = 1 - p_1$, $q_2 = 1 - p_2$.

We also use the data security technique to encrypt data in the cloud and keep the data confidential in the private cloud and to distribute the access right for cloud users and provide an effective access to the cloud data.

Along with data storage, data analysis and quality control finishing their works, the loaded disaster data would be cached on cloud, to facilitate demonstration fluency of disaster inversion process. The visualization cases of 3D terrain representation, water level rising process,and detail disaster situation are shown in Figures 3, 4, and 5, respectively.

In belief, the disaster inversion visualization has made significant contributions for marine big data.

(1) Terrain Reconstruction. The project has visualized the disaster of sea terrain in the form of 3-dimension style, which could help to analyze the causes of the disasters based on terrain conditions.

(2) Disaster Reproduction. The project, in a quasireal-time way, has reproduced multiple dataset involved in disaster process, including velocity, flow direction, and water depth, which could further help to fleetly evacuate victims.

(3) Disaster Evaluation. The project has reconstructed the postdisaster scene, which helps to evaluate the economic losses and human victims of the disaster area. (The project (Grant number 20905014-06) is finished by Digital Ocean Institute, College of Information, Shanghai Ocean University in May, 2014.)

5. Challenges in Marine Big Data Management

Prominent characteristics of marine big data have brought about new issues. In this case, this section discusses practical and theoretical challenges in the existence of marine big data management: data storage, data analysis, quality control, and data security.

5.1. Data Storage. Data storage underpins and sustains the efficient application of data. Under storage platform, rational data partition and suitable index building assist to realize efficient data queries. It has to be noted that there are some present situations in traditional storage system, mainly including lack in supporting dynamic scalability, simplified data storage method, relatively fixed data structure, controllable data size, and aware data type.

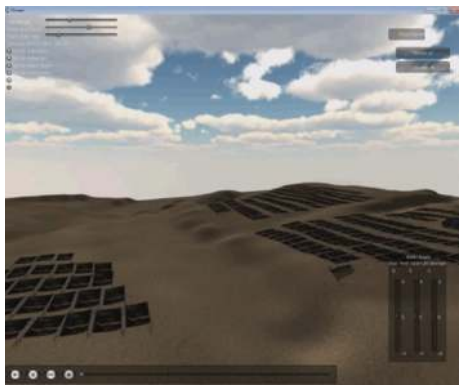

Figure 3: 3D terrain graph.

Figure 4: Water rising graph.

However, characteristics of marine big data, including large-volume, sensitivity, real-time, high dimension, diversity in data provision, and type, pose new challenges for data storage, mainly in two aspects.

(1) Scalability Requirement for Storage Space. Due to huge-volume and real-time characteristics of marine big data, it poses new challenges towards hardware architecture and file system, which requires data storage to be more scalable. Along with real-time acquisition of observing data, data storage should be more flexible.

(2) Diversity Requirement for Storage System and Storage Model. Multisource characteristic of marine big data imposes a great diversity in data type. Marine big data basically falls into three catalogs: structured attribute data (*.MDB, *.dbf, *.bak, *.dmp, etc.), spatial data (*.shp, *.adf, *.tif, *.jpg, etc.), and unstructured data (*.doc, *.xls, *.pdf, *.txt, *.xml, etc.). Diversity in data type puts forward higher request for database consistency, database usability, and partition tolerance.

5.2. Data Analysis. The purpose of data analysis is to find patterns and extract information from complex and vast data, which is the key to effectively exploit the value of marine big data. The object in traditional data analysis tends to be small datasets, which are structured dataset and single objects.

Figure 5: Postdisaster graph.

Data analysis and data mining prefer to build models by manual in advance according to priori knowledge and then analyze based on the selected data model. Diversity in data provision and heterogeneous characteristics of marine big data has raised some new issues, such as huge data amount, nonunified data type, and low data quality. Additionally, traditional analysis techniques like data mining, machine learning, and statistical analysis should be adjusted to make it adaptive to marine big data. Marine big data brings along with some analytical challenges, specific as follows.

(1) Effectiveness Requirement. Marine big data contains its unique characteristics, such as huge data amount, complex data type, and uncertain data distribution. Therefore, adaptive algorithm and model should be selected according to its data type and analysis target, to fleetly process marine data. This further leads to some challenges towards hardware and software, especially on data analysis algorithms.

(2) Efficiency Requirement. The application with marine data requires a higher demand on real-time response. Under such circumstances as Snow Dragon's expedition on extreme conditions in polar, it is essential to make a comprehensive analysis of real-time information on weather, sea ice, seabed, ship, and so forth. However, massive data processing and analyzing in real-time consumes huge computing resources, while traditional computing technologies are insufficient to that. Basically, it performs better in cooperation with cloud computing but proposes new challenges towards the scalability and real-time of its algorithm.

5.3. Data Availability. Quality of marine big data is the foundation of the development of marine Geographic Information Science. Due to the restriction of the acquisition and processing method, there exist a large number of random errors in marine big data, which leads to the unreliability of the marine data products. The existing theory of quality management is mainly used to control the quality of traditional industrial product, which is not quite suitable for the quality control marine big data with characteristics of multisource, massiveness, spatial relativity, and so forth. Therefore, development of quality control theories based on

characteristics of marine big data is one of the key issues in data management. The challenges are combing conventional quality control theory with marine big data management.

(1) Quality Inspection Plan Designing. Considering the characteristics of marine big data, it has become a priority issue to take the required precision into account, to design the optimal the sampling number and the acceptance number.

(2) Spatial Sampling Method Deducing. Due to the spatial autocorrelation characteristic of marine big data, the method of selecting marine data samples is different from the classical sampling method. The distance between data restricts the information redundancy between the sample points. Both considering the spatial autocorrelation of marine big data and achieving the maximum of information under the same inspection cost guarantee the implementation of the quality control of marine big data.

(3) Theory of Usability and Autorestoration. In terms of the quality inspection result, marine big data can be divided into usable data and risk data. Due to various data acquisition methods, most of marine big data are irreversible, which makes it significant to study on the usability of marine data products and data autorestoration.

5.4. Data Security. Compared to the traditional data security, marine data's security and privacy protection appear significantly different and show the typical structure-based characteristics including "one to many" structure (one user stores the data, multiple users access), "many to one" structure, and "many to many" structure. From the data processing perspective, the service of marine big data can be divided into data storage service, data access service, data computation service, data share service, and data supervision service. In short, the challenges in marine big data security can be also summarized as "secure data storage, secure data access, secure data computation, secure data sharing, and secure data supervision."

(1) Secure Data Storage Requirement. From the case of Snowden, people all over the world have realized that the users' privacy as well as the sensitive data will be greatly harmed if the data are not in a properly secure storage. The storage of marine data often relies on the credibility of the servers/nodes, which could not resist the servers' administrators and the inside adversary wiretapping and tampering the data. If the data is not discriminated and used directly, the factual data also cheat the users; in particular forgery or deliberately manufacturing data often leads to the incorrect and incomplete conclusions.

(2) Secure Data Access Requirement. Data access control is an effective way to realize the data sharing. Marine big data are used for different scenarios and accessed by different users with different roles and different security levels. The access control requirements are very prominent since the traditional access control techniques mainly depend on the security of the database and cloud service providers. Once the database administrators and cloud service providers take malicious behaviors, the data are no longer secure in the database and data sharing, which results in violation of the data confidentiality and the users' privacy.

(3) Secure Data Computation Requirement. Data computation such as calculation and analysis of marine data is another important application. Since marine big data service providers cannot be fully trusted and computation services are often in an outsourcing way, it is an important requirement that how to achieve the data confidentiality and realize the data calculation and analysis simultaneously. In addition, it is also important to improve the efficiency of data calculation and analysis as well as ensure the effectiveness of the storage data.

(4) Secure Data Sharing Requirement. In the marine data sharing and dissemination, the data are often shared among the authored users. Thus, the security is based on the users' secret keys. The data will be given away if the user's secret keys were leaked intentionally or unintentionally, which is unable to realize the secure data sharing and dissemination mechanism in the cloud. Furthermore, since the security of modern cryptographic systems depends only on the secret keys, the whole security systems would collapse if there is no technique to trace and revoke leaked secret keys.

(5) Secure Data Supervision Requirement. Data supervision is a guarantee to marine data security. In the processing stages of data storage, computation, sharing, and dissemination, malicious adversaries may insert false data intentionally and unintentional users may insert error data if there is a lack of the techniques of data supervision and monitoring. It is also important to remove illegal information, reduce the redundancy cost, check the content completeness, and verify the correctness of the calculation results in the marine data supervision.

6. Conclusions

As we have entered an era of marine big data, it is of great realistic and theoretical significance to study on the marine big data management. Unfortunately, existing techniques and theories are very limited to solve the real problems completely in the marine big data. To tackle above issues, this paper has analyzed the existing challenges in data storage, data analysis, quality control, and data security, summarized the marine big data models, algorithms, methods, and techniques in field of marine big data management, and finally presented a practical engineering instance that demonstrates the management architecture. There is no doubt that study on marine big data management is still in the initial stage of development; thereby more scientific investments from both academy and industry should be poured into this scientific paradigm to capture huge values from marine big data.

Conflict of Interests

Dongmei Huang, Danfeng Zhao, Lifei Wei, Zhenhua Wang, and Yanling Du declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant no. 61272098 and 61402282) and the National 973 Program (Grant no. 2012CB316206). The authors would like to thank anonymous reviewers who helped us in giving comments to this paper.

References

[1] "The International ARGO Project," http://www.argo.net/.

[2] Ocean Networks Canada, http://www.neptunecanada.ca/.

[3] "Ocean Observatories Initiative," http://oceanobservatories.org/.

[4] The Global Ocean Observing System, http://www.ioc-goos.org/.

[5] Intergrated Ocean Observing System, http://www.ioos.noaa.gov.

[6] National Aeronautics and Space Administration, http://www.nasa.gov/.

[7] National Oceanic and Atmospheric Administration, http://www.noaa.gov/.

[8] China Argo Data Center, http://www.argo.gov.cn/.

[9] "News in brief of Argo," China Argo Real-time Data Center, China, 2014, (Chinese), http://www.argo.org.cn/.

[10] P. J. Durack, S. E. Wijffels, and R. J. Matear, "Ocean salinities reveal strong global water cycle intensification during 1950 to 2000," *Science*, vol. 336, no. 6080, pp. 455–458, 2012.

[11] C. J. Brown, S. J. Smith, P. Lawton, and J. T. Anderson, "Benthic habitat mapping: a review of progress towards improved understanding of the spatial ecology of the seafloor using acoustic techniques," *Estuarine, Coastal and Shelf Science*, vol. 92, no. 3, pp. 502–520, 2011.

[12] G. C. Rogers, R. Meldrum, R. Baldwin et al., "The NEPTUNE Canada seismograph network," *Seismological Research Letters*, vol. 81, no. 2, p. 369, 2009.

[13] A. B. Rabinovich, R. E. Thomson, and I. V. Fine, "The 2010 chilean tsunami off the west coast of Canada and the northwest coast of the United States," *Pure and Applied Geophysics*, vol. 170, no. 9-10, pp. 1529–1565, 2013.

[14] X. Meng and X. Ci, "Big data management: concepts, techniques and challenges," *Computer Research and Development*, vol. 50, no. 1, pp. 146–169, 2013 (Chinese).

[15] National Oceanographic Data Center, http://www.nodc.noaa.gov/access/allproducts.html.

[16] http://en.wikipedia.org/wiki/Oceanography.

[17] D. Huang, Y. Du, and Q. He, "Migration algorithm for big data in hybrid cloud storage," *Journal of Computer Research and Development*, vol. 51, no. 1, pp. 199–205, 2014 (Chinese).

[18] Google Cloud Storage Overview, https://developers.google.com/storage/docs/overview?csw=1.

[19] http://aws.amazon.com/cn/s3/.

[20] "Windows Azure General Availability," http://blogs.microsoft.com/blog/2010/02/01/windows-azure-general-availability/.

[21] "Cloud computing strategy and Blue Cloud of IBM," ftp://ftp.software.ibm.com/software//cn/smsp/4.0/cloudstrategy_IBMbluecloud.pdf.

[22] D. Zhao, *The architecture of artifact-centric business process management system on the cloud computing platform [Ph.D. thesis]*, Yanshan University, Qinhuangdao, China, 2012.

[23] X.-M. Zhou and G.-R. Wang, "Key dimension based high-dimensional data partition strategy," *Journal of Software*, vol. 15, no. 9, pp. 1361–1374, 2004 (Chinese).

[24] P. Ren, W. Liu, and D. Sun, "Partition-based data cube storage and parallel queries for cloud computing," in *Proceedings of the 9th International Conference on Natural Computation (ICNC '13)*, pp. 1183–1187, July 2013.

[25] C. Selvakumar, G. J. Rathanam, and M. R. Sumalatha, "PDDS—improving cloud data storage security using data partitioning technique," in *Proceedings of the 3rd IEEE International Advance Computing Conference (IACC '13)*, pp. 7–11, 2013.

[26] D. Zhao, S. Jin, G. Liu, F. Gao, and N. Wang, "A cryptograph index technology based on query probability in DAS model," *Journal of Yanshan University*, vol. 32, no. 6, pp. 477–482, 2008 (Chinese).

[27] B.-R. Dai and I.-C. Lin, "Efficient map/reduce-based DBSCAN algorithm with optimized data partition," in *Proceedings of the IEEE 5th International Conference on Cloud Computing (CLOUD '12)*, pp. 59–66, June 2012.

[28] L. Han, X. Sun, and Z. Wu, "Optimization study on sample based partition on mapreduce," *Journal of Computer Research and Development*, vol. 50, pp. 77–84, 2013 (Chinese).

[29] Y. Xu, P. Zou, W. Qu, Z. Li, K. Li, and X. Cui, "Sampling-based partitioning in mapreduce for skewed data," in *Proceedings of the 7th ChinaGrid Annual Conference (ChinaGrid '12)*, pp. 1–8, September 2012.

[30] D. Huang, L. Sun, D. Zhao et al., "An efficient hybrid index structure for temporal marine data," in *Proceedings of Conference on Web-Age Information Management*, 2014.

[31] S. Shi and B. Lei, *Theory and Practice on China Digital Ocean*, Ocean Press, Beijing, China, 2011.

[32] A. Fox, C. Eichelberger, J. Hughes, and S. Lyon, "Spatio-temporal indexing in non-relational distributed databases," in *Proceedings of the IEEE International Conference on Big Data*, pp. 291–299, October 2013.

[33] B. Stantic, R. Topor, J. Terry, and A. Sattar, "Advanced indexing technique for temporal data," *Computer Science and Information Systems*, vol. 7, no. 4, pp. 679–703, 2010.

[34] B. Stantic, J. Terry, R. Topor et al., "Indexing temporal data with virtual structure," in *Proceedings of the 14th East European Conference on Advances in Databases and Information Systems*, pp. 591–594, 2010.

[35] T. Emrich, H.-P. Kriegel, N. Mamoulis, M. Renz, and A. Züfle, "Indexing uncertain spatio-temporal data," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM '12)*, pp. 395–404, November 2012.

[36] Y. Zhong, J. Fang, and X. Zhao, "VegaIndexer: a distributed composite index scheme for big spatio-temporal sensor data on cloud," in *Proceedings of the 33rd IEEE International Geoscience and Remote Sensing Symposium (IGARSS '13)*, pp. 1713–1716, July 2013.

[37] S. Chen, B. C. Ooi, K.-L. Tan, and M. A. Nascimento, "ST2B-tree: a self-tunable spatio-temporal B+-tree index for moving objects," in *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD '08)*, pp. 29–42, June 2008.

[38] S. Chen, B. C. Ooi, K.-L. Tan, and M. A. Nascimento, "ST2B-tree: a self-tunable spatio-temporal B+-tree index for moving objects," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 29–42, June 2008.

[39] M. Kaufmann, A. A. Manjili, P. Vagenas et al., "Timeline index: a unified data structure for processing queries on temporal data in SAP HANA," in *Proceedings of the ACM SIGMOD Conference on Management of Data (SIGMOD '13)*, pp. 1173–1184, June 2013.

[40] X. Hu, M. Qiao, and Y. Tao, "Independent range sampling," in *Proceedings of the 33rd ACM Special Interest Group Conference on Management of Data*, pp. 246–255, 2014.

[41] J. Zhang, G. Chen, and X. Tang, "Extracting representative information to enhance flexible data queries," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 6, pp. 928–941, 2012.

[42] S. Thomas and L. Kevin, "MapReduce optimization using regulated dynamic prioritization," in *Proceedings of the 11th international joint conference on Measurement and modeling of computer systems (SIGMETRICS '09)*, pp. 299–310, June 2009.

[43] W. Gharibi and A. Mousa, "Query optimization based on time scheduling approach," in *Proceedings of the 11th IEEE East-West Design & Test Symposium (EWDTS '13)*, pp. 1–7, September 2013.

[44] T. Miller and T. Gibson, "An improved long-term file usage prediction algorithm," http://users.soe.ucsc.edu/~elm/Papers/cmg99.pdf.

[45] J. Jeong and M. Dubois, "Cost-sensitive cache replacement algorithms," in *Proceedings the 9th International Symposium on High-Performance Computer Architecture (HPCA-9 '03)*, vol. 1, pp. 327–337, Anaheim, Calif, USA, 2003.

[46] B. Reed and D. D. E. Long, "Analysis of caching algorithms for distributed file systems," in *Proceedings of the ACM SIGOPS Operating Systems Review*, pp. 12–21, July 1996.

[47] D. He, X. Zhang, D. H. C. Du, and G. Grider, "Coordinating parallel hierarchical storage management in object-based cluster file system," http://wiki.lustre.org/images/f/fc/MSST-2006-paper.pdf.

[48] L. Ao, D. Yu, J. Shu, and W. Xue, "A tiered storage system for massive data: TH-TS," *Journal of Computer Research and Development*, vol. 48, no. 6, pp. 1089–1100, 2011 (Chinese).

[49] S. Melnik, A. Gubarev, J. J. Long et al., "Dremel: interactive analysis of web-scale datasets," *Proceedings of the VLDB Endowment*, vol. 3, no. 1-2, pp. 330–339, 2010.

[50] F. Li, B. C. Ooi, M. T. Ozsu et al., "Distributed data management using mapReduce," *ACM Computing Surveys (CSUR)*, vol. 46, no. 3, pp. 1–41, 2014.

[51] G. Malewicz, M. H. Austern, A. J. C. Bik et al., "Pregel: a system for large-scale graph processing," in *Proceedings of the International Conference on Management of Data (SIGMOD '10)*, pp. 135–146, June 2010.

[52] L. G. Valiant, "Bridging model for parallel computation," *Communications of the ACM*, vol. 33, no. 8, pp. 103–111, 1990.

[53] "Google BigQuery," https://cloud.google.com/products/big-query/.

[54] A. Hall, O. Bachmann, R. Bussow et al., "Processing a trillion cells per mouse click," *PVLDB*, vol. 5, no. 11, pp. 1436–1446, 2012.

[55] M. Isard, M. Budiu, Y. Yu, A. Birrell, and D. Fetterly, "Dryad: distributed data-parallel programs from sequential building blocks," in *Proceedings of the 2nd ACM SIGOPS/EuroSys European Conference on Computer Systems (EuroSys '07)*, vol. 41, pp. 59–72, March 2007.

[56] "Cascading," http://www.cascading.org/.

[57] G. Shank, R. Y. Wang, and Z. Mostapha, "IP-map: representing the manufacture of an information product," in *Proceedings of the Information Quality Conference*, pp. 1–16, 2000.

[58] Y. Wand and R. Y. Wang, "Anchoring data quality dimensions in ontological foundations," *Communications of the ACM*, vol. 39, no. 11, pp. 86–95, 1996.

[59] A. Zargar and R. Devillers, "An operation-based communication of spatial data quality," in *Proceedings of the International Conference on Advanced Geographic Information Systems and Web Services (GEOWS '09)*, pp. 140–145, February 2009.

[60] "ISO 19113: Geographic Information Quality Principles," http://www.statkart.no/isot211/, 2001.

[61] A. Kleiner, A. Talwalkar, P. Sarkar, and M. I. Jordan, "A scalable bootstrap for massive data," *Journal of the Royal Statistical Society B: Statistical Methodology*, vol. 76, no. 4, pp. 795–816, 2014.

[62] J. Wang, J. Liu, D. Zhuan, L. Li, and Y. Ge, "Spatial sampling design for monitoring the area of cultivated land," *International Journal of Remote Sensing*, vol. 23, no. 2, pp. 263–284, 2002.

[63] J. Wang, R. Haining, and Z. Cao, "Sample surveying to estimate the mean of a heterogeneous surface: reducing the error variance through zoning," *International Journal of Geographical Information Science*, vol. 24, no. 4, pp. 523–543, 2010.

[64] Z. Wang, X. N. Zhou, and D. M. Huang, "A sampling model for the quality inspection of uncertain ocean data," to appear in *Computer Science*.

[65] E. G. Schilling and D. V. Neubauer, *Acceptance Sampling and Quality Control*, CRC Press, New York, NY, USA, 2012.

[66] ISO 2859.0, *Sampling Procedures for Inspection by Attributes—Part 0: Introduction to the ISO 2859 Attribute Sampling System*, International Organization for Standardization, 1995.

[67] A. Golub, "Designing single-sampling inspection plans when the sample size is fixed," *Journal of the American Statistical Association*, vol. 48, no. 262, pp. 278–288, 1953.

[68] A. Hald, "The determination of single sampling attribute plans with given producer's and consumer's risk," *Technometrics*, vol. 9, no. 3, pp. 401–415, 1967.

[69] B. P. M. Duarte and P. M. Saraiva, "An optimization-based approach for designing attribute acceptance sampling plans," *International Journal of Quality and Reliability Management*, vol. 25, no. 8, pp. 824–841, 2008.

[70] S. T. A. Niaki and M. S. F. Nezhad, "Designing an optimum acceptance sampling plan using Bayesian inferences and a stochastic dynamic programming approach," *ScientiaIranica Transaction E: Industrial Engineering*, vol. 16, no. 1, pp. 19–25, 2009.

[71] E. B. Jamkhaneh and B. S. Gildeh, "AOQ and ATI for double sampling plan with using fuzzy binomial distribution," in *Proceedings of the International Conference on Intelligent Computing and Cognitive Informatics (ICICCI '10)*, pp. 45–49, June 2010.

[72] C. A. J. Klaassen, "Credit in acceptance sampling on attributes," *Technometrics*, vol. 43, no. 2, pp. 212–222, 2001.

[73] J. Li and X. Liu, "An important aspect of big data: data usability," *Computer Research and Development*, vol. 50, no. 6, pp. 1147–1162, 2012.

[74] W. Fan, F. Geerts, J. Li, and M. Xiong, "Discovering conditional functional dependencies," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 5, pp. 683–698, 2011.

[75] L. Golab, F. Korn, and D. Srivastava, "Efficient and effective analysis of data quality using pattern tableaux," *IEEE on Data Engineering*, vol. 34, no. 3, pp. 26–33, 2011.

[76] G. Grahne, *The Problem of Incomplete Information in Relational Databases*, Springer, Berlin, Germany, 1991.

[77] W. Fan, F. Geerts, and J. Wijsen, "Determining the currency of data," *ACM Transactions on Database Systems*, vol. 37, no. 4, article 25, 2012.

[78] L. W. Ferreira Chaves, E. Buchmann, and K. Böhm, "Finding misplaced items in retail by clustering RFID data," in *Proceedings of the 13th International Conference on Extending Database Technology (EDBT '10)*, pp. 501–512, March 2010.

[79] W. Fan, F. Geerts, X. Jia, and A. Kementsietsidis, "Conditional functional dependencies for capturing data inconsistencies," *ACM Transactions on Database Systems*, vol. 33, no. 2, Article ID 1366103, pp. 1–48, 2008.

[80] W. Fan, F. Geerts, S. Ma, and H. Müller, "Detecting inconsistencies in distributed data," in *Proceedings of the 26th IEEE International Conference on Data Engineering (ICDE '10)*, pp. 64–75, March 2010.

[81] W. Fan, J. Li, N. Tang, and W. Yu, "Incremental detection of inconsistencies in distributed data," in *Proceedings of the IEEE 28th International Conference on Data Engineering (ICDE '12)*, pp. 318–329, April 2012.

[82] S. E. Whang, D. Menestrina, G. Koutrika, M. Theobald, and H. Garcia-Molina, "Entity resolution with iterative blocking," in *Proceedings of the 35th SIGMOD Conference on Managemnt of Data*, pp. 219–231, July 2009.

[83] J. Chomicki and J. Marcinkowski, "Minimal-change integrity maintenance using tuple deletions," *Information and Computation*, vol. 197, no. 1-2, pp. 90–121, 2005.

[84] J. Bleiholder, S. Szott, M. Herschel, F. Kaufer, and F. Naumann, "Subsumption and complementation as data fusion operators," in *Proceedings of the 13th International Conference on Extending Database Technology (EDBT '10)*, pp. 513–524, March 2010.

[85] H. Lin, Z. Cao, X. Liang, and J. Shao, "Secure threshold multi authority attribute based encryption without a central authority," *Information Sciences*, vol. 180, no. 13, pp. 2618–2632, 2010.

[86] K. Yang and X. Jia, "Data storage auditing service in cloud computing: challenges, methods and opportunities," *World Wide Web*, vol. 15, no. 4, pp. 409–428, 2012.

[87] C. Wang, S. S. Chow, Q. Wang, K. Ren, and W. Lou, "Privacy-preserving public auditing for secure cloud storage," *IEEE Transactions on Computers*, vol. 62, no. 2, pp. 362–375, 2013.

[88] M. Li, S. Yu, K. Ren, W. Lou, and Y. Hou, "Toward privacy-assured and searchable cloud data storage services," *IEEE Network*, vol. 27, no. 4, pp. 56–62, 2013.

[89] X. Liang, Z. Cao, H. Lin, and D. Xing, "Provably secure and efficient bounded ciphertext policy attribute based encryption," in *Proceedings of the 4th International Symposium on ACM Symposium on Information, Computer and Communications Security (ASIACCS '09)*, pp. 343–352, March 2009.

[90] K. Yang, X. Jia, K. Ren et al., "Enabling efficient access control with dynamic policy updating for big data in the cloud," in *Proceedings of the IEEE Conference on Computer Communications (INFOCOM '013)*, pp. 2013–2021, 2013.

[91] H. Wang and L. V. S. Lakshmanan, "Efficient secure query evaluation over encrypted XML databases," in *Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB '06)*, pp. 127–138, September 2006.

[92] N. Cao, C. Wang, L. Ming et al., "Privacy-preserving multi-keyword ranked search over encrypted cloud data," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 1, pp. 222–233, 2014.

[93] C. Wang, K. Ren, S. Yu, and K. M. R. Urs, "Achieving usable and privacy-assured similarity search over outsourced cloud data," in *Proceedings of the IEEE Conference on Computer Communications (INFOCOM '12)*, pp. 451–459, March 2012.

[94] E. Shen, E. Shi, and B. Waters, "Predicate privacy in encryption systems," in *Proceedings of the 6th Theory of Cryptography Conference (TCC '09)*, pp. 457–473, San Francisco, Calif, USA, March 2009.

[95] C. Wang, K. Ren, J. Wang, and Q. Wang, "Harnessing the cloud for securely outsourcing large-scale systems of linear equations," *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 6, pp. 1172–1181, 2013.

[96] Z. Xu, C. Wang, K. Ren et al., "Proof-carrying cloud computation: the case of convex optimization," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 11, pp. 1790–1803, 2014.

[97] J. H. Cheon, J. S. Coron, J. Kim et al., "Batch fully homomorphic encryption over the integers," in *Proceedings of the Annual International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT '13)*, pp. 315–335, 2013.

[98] M. Li, S. Yu, Y. Zheng, K. Ren, and W. Lou, "Scalable and secure sharing of personal health records in cloud computing using attribute-based encryption," *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 1, pp. 131–143, 2013.

[99] Z. Liu, Z. Cao, and D. S. Wong, "White-box traceable ciphertext-policy attribute-based encryption supporting any monotone access structures," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 1, pp. 76–88, 2013.

[100] J. Ning, Z. Cao, X. Dong et al., "Large universe ciphertext-policy attribute-based encryption with white-box traceability," in *Proceedings of the European Symposium on Research in Computer Security (ESORICS '14)*, Wroclaw, Poland, September 2014.

[101] Z. Liu, Z. Cao, and D. S. Wong, "Blackbox traceable CP-ABE: how to catch people leaking their keys by selling decryption devices on eBay," in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS '13)*, pp. 475–486, November 2013.

[102] K. Yang, X. Jia, and K. Ren, "Attribute-based fine-grained access control with efficient revocation in cloud storage systems," in *Proceedings of the 8th ACM SIGSAC Symposium on Information, Computer and Communications Security (ASIACCS '13)*, pp. 523–528, May 2013.

[103] Z. Cao, *New Directions of Modern Cryptography*, CRC Press, Boca Raton, Fla, USA, 2012.

[104] K. Yang, X. Jia, K. Ren, B. Zhang, and R. Xie, "DAC-MACS: effective data access control for multiauthority cloud storage systems," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 11, pp. 1790–1801, 2013.

[105] L. Wei, H. Zhu, Z. Cao et al., "Security and privacy for storage and computation in cloud computing," *Information Sciences*, vol. 258, pp. 371–386, 2014.

[106] Q. Wang, K. Ren, and X. Meng, "When cloud meets eBay: towards effective pricing for cloud computing," in *Proceedings of the IEEE Conference on Computer Communications (INFOCOM '12)*, pp. 936–944, March 2012.

[107] D.-G. Feng, M. Zhang, and H. Li, "Big data security and privacy protection," *Chinese Journal of Computers*, vol. 37, no. 1, pp. 1–13, 2014 (Chinese).

[108] G.-H. Kim, S. Trimi, and J.-H. Chung, "Big-data applications in the government sector," *Communications of the ACM*, vol. 57, no. 3, pp. 78–85, 2014.