Development and Application Study of Marine Data Managing and Sharing Platform

Zhang Hongxin, Duan Kanghong, Zhang Xiaobo

North China Sea Marine Technical Support Center, State Oceanic Administration, China 266033

Corresponding author's e-mail: bhjszx_zhx@163.com

Received 10 October 2013, www.cmnt.lv

Abstract

The marine scientific data has the characteristics like grid form, multiple dimensions, and geographic information included. There exists strong semantic and grammar heterogeneity among different types of marine data. So this paper studies and develops the technologies in marine data sharing platform to solve above problems. It proposes a managing and sharing scheme for marine data processing based on distributed computing technology. The scheme adopts parallel and distributed computing technology, Linux cluster technology to process the marine data in Hadoop distributed platform. The technologies of HDFS distributed file system; Map/Reduce parallel computing programming model and Hbase are also used. Then we provide the design of key modules by programming and the distributed system managing of the cloud platform, to offers a data managing and sharing platform with high reliability and stability. Finally the tests verify the feasibility and effectiveness of the proposed platform.

Keywords: Marine data; NetCDF; Hbase; parallel loading; Hadoop

1 Introduction

With the development of Web service, ontology, and semantic web technology, there emerges more and more data sharing methods to solve the problems in marine [1,2]. However, due to the specialty of marine application, it is very difficult to realize data sharing of marine field. In order to reach the purpose of resources sharing and extracting the needed information from massive data, the problem of scientific data sharing must be solved first. In recent 20 years, people have greatly studied the schemes in data sharing. They put forward many solutions and systematic structures of data sharing technology [3-5]. Many successful scientific data sharing platforms are established.

From some literatures we find that the cloud computing technology is initially adopted for large-scale data processing. Under the cloud computing environment, the modes of releasing and use of network resources will be unified as service. Any usable information resources in this environment will exist as service and the users make use of these resources with pay on demand [6]. Thus, these autonomous, heterogeneous and distributed information resources need to be managed uniformly. The physical location and detailed information of information resources will be shielded. It is significant to establish the platform of marine surveying data share and develop corresponding decision support system, to utilize advanced cloud computing technology for the construction of scientific data sharing platform, and interoperability management of marine resources database. Besides, it is also necessary to develop, reconstruct decision-making the support system and application software of other businesses and offer basic infrastructure of interoperability service and powerful technology support. For instance, in literature [7], the author uses cloud computing technology to improve the service pattern of current users in college libraries, to improve the service quality and resources utilization of libraries. In literature [8], the author adopts the idea and techniques of cloud computing to establish information resources value-added utilization model; In literature [9], in order to improve the information resources sharing capability of the universities, the author takes cloud computing technology to establish a cloud computing project for university database sharing. In literature [10], the author app-lies cloud computing technology to establish police intelligence data sharing cloud for public security information system. However, at present, there are few related researches of scientific data sharing platform under cloud computing environments. The similar framework technologies for data sharing platform are also rare. In order to Contra posing to these problems, this paper studies and designs a processing cloud platform framework structure for marine data and initially establishes a Hadoop-based large-scale data storage cloud platform. It also provides a constructing project of data processing platform for marine surveying data to offer a parallel data mining result and analysis of marine resources, with massive data storage, analysis, processing, mining, highly providing performance and strong reliability. Therefore, it can perform parallel mining and usable mining massive marine resources data in the cloud platform for marine data processing, to provide full sharing of data in marine resources environment. The rest of paper is organized as follows: section 2 presents the overall structure of marine data sharing and managing platform. Section 3 describes the function design of key modules. Section 4 tests the performance of our system by analysis on the implementation. Section 5 offers the conclusions and directions for future research.

2 Overall Architecture Design

2.1 NETWORK STRUCTURE

The network structure of marine surveying data cloud platform is shown as figure 1. From figure 1, we can see that the bottom layers are different types of users such as the government branches including oceanic bureau, weather bureau, etc. There are some related organizations, institutions and common users. They are all connected to the platform by public network and provide services offered by the platform. The upper layer is cloud storage platform system. Its structure contains the hardware configurations such as database server, interconnection equipment, application server and the software configurations deployed in this system. The hardware configuration, like database server and application server, can be distributed in differrent regions or deployed in the same region. There are two aspects to be considered during the design of our platform: on one hand, the processing capacity of large-scale massive marine scientific data will constantly increase with time; on the other hand, the application requirement of platform system users will cause complicated change. Thus, during the deployment and operation of platform system, it is convenient to transparently extend hardware and software for users, but users will not notice these changing extensions. The platform system also takes into account rational process on the relationship between current relational database system and the system. By rational allocation and process of non-isomerization, they can be integrated together to provide calculation storage [11] service for users, which further extends the data processing function. Furthermore, it can conveniently manage stored data of massive marine surveying data so that these data can exert their application values.

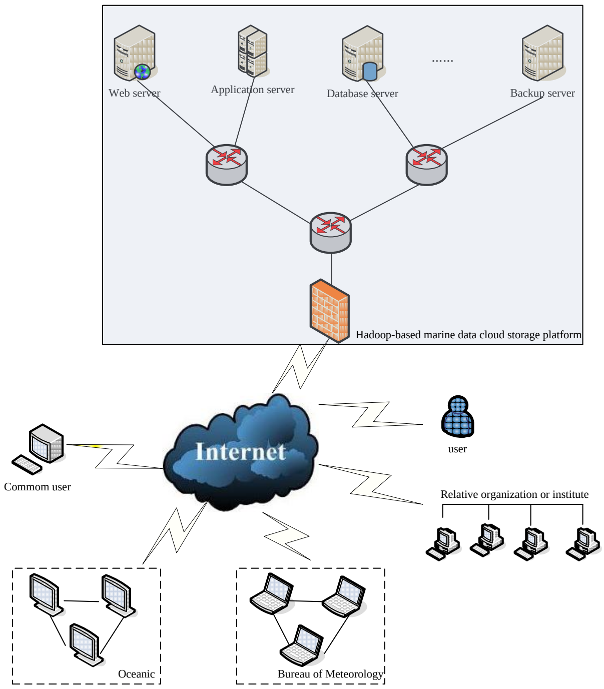


FIGURE 1 The overall topology of system

2.2 FUNCTIONAL FRAMEWORK

The whole system is divided into three layers from the perspective of system function. Its framework is depicted in figure. Data access layer lies in the bottom of the whole functional framework. Its function is the non-isomerization of various data source, shielding different types of data and offering more effective access function to database. When processing massive marine data, since the data complexity cause that the data have different types and structures, it is hard to save and read the data effectively. Only by data transformation and process at this layer, the heterogeneity can be removed when accessing various databases. The platform system can meet the demand for processing and storing of marine massive data by such mode, so the platform system is easier to be managed and deployed and it has stronger expansibility and better completeness.

The data processing layer lies in the second layer of the whole functional framework. This layer adopts parallel distributed database technology, Map/Reduce parallel programming calculation technology, Linux cluster processing technology, etc. The main function is processing in parallel, load and storing marine massive scientific data. The process of this layer is: first, the massive data is parallel processed. Then, the processed data results are efficiently stored in distributed database of this platform system and by database access layer. This layer also provides the management support to guarantee normal operation of the platform system. The service application layer lies at the top of the whole functional framework. It is made up by the algorithm library-based API and users-based GUI interface.

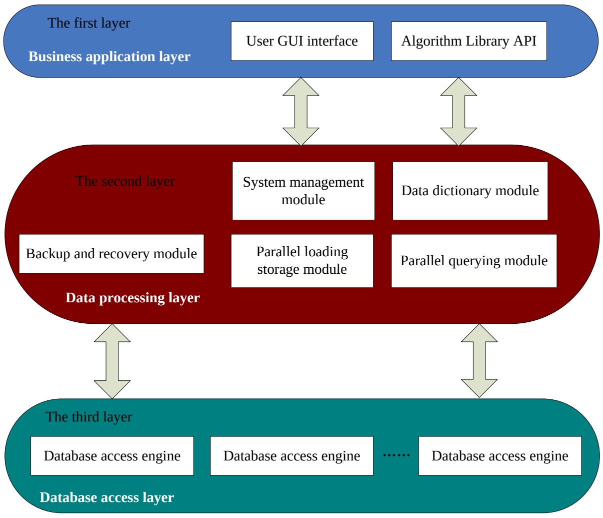


FIGURE 2 The framework of platform system

3 Design and Implementation of Key Functions

According to the functional design of the platform in this paper, the primary part of the system is data processing layer. During the implementation of data processing layer, the parallel loading memory module becomes the core of the whole platform. Since Hadoop distribution technology provides the data storage and processing model and method, we use Hadoop distributed computing model to process the massive source data [12]. Then Hbase distributed database is used to store the processed data, to provide the sharing and managing of massive marine data.

3.1 PARALLER LOADING MEMORY MODULE

One of the important kernel modules of this platform is the parallel loading memory module. It includes the functions like loading, processing and massive amounts of data storing. The flow chart of this module is shown in figure 3:

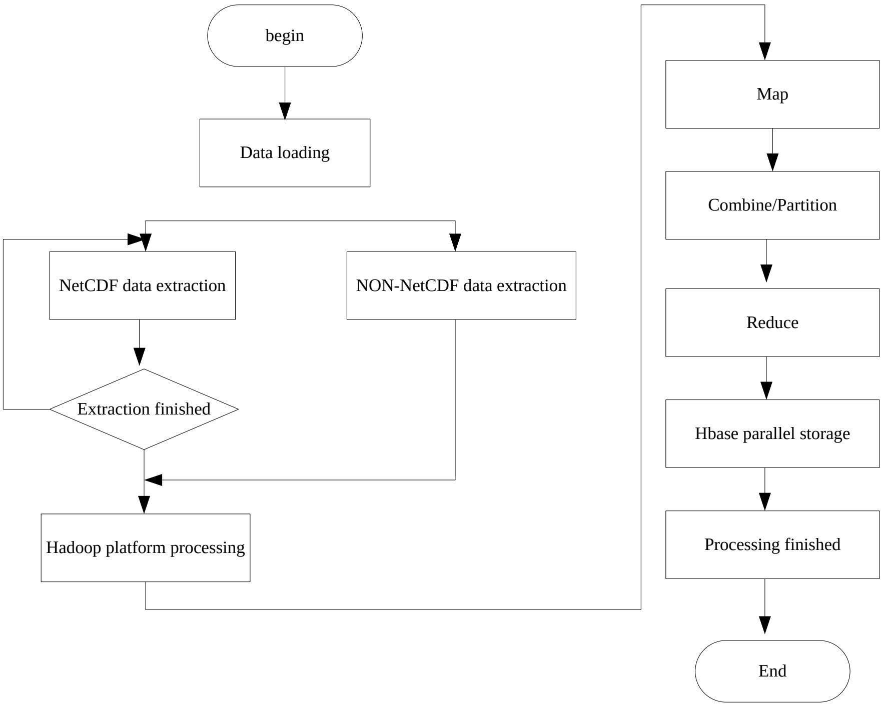


FIGURE 3 flow chart of parallel loading storage module

The NetCDF metadata extraction is the basis and premise of other service. It uses XML grammar structure to describe the NetCDF files and it can extract the elements in NetCDF files such as dimension, scalar quantity and attribute etc. To be transmitted in the network better, NetCDF files need to extract metadata as semi structured texts. When it is processed it will be described as XML-type data, for convenient transmission and usage. The data extraction function of NetCDF is extracting the elements in NetCDF and uses XML language to describe the NetCDF files. So the data with NetCDF format can be transmitted in the environment like SOA and grid. We use NetCDF class to depict one NetCDF file and each NetCDF file contains multiple dimensions, variables and attributes. Each variable has its own variable attributes to describe the content including variable function or units of measurement. The variables and attributes also include the information of storage types. The implementation of NetCDF metadata is depicted in figure 4.

3.2 HBASE ORIENTED ETL COMPONENTS

The Hbase oriented ETL components are used to operate *ETL of Hbsed internal tables. When Hadoop and Hbase* start up, the data in original table will be transformed as given rules. The converted data are saved in new Hbase tables for sequent algorithms and processing. As is shown in figure 5, in the ETL component flow, the users choose a series of parameters at the interface layer, such as data source, target location, data transformation field, incremental field, parallelism, etc. The choices of user are mapped to the internal Hbased oriented ETL components configuration interface as shown in table 1. This configure file denote the choice demand for users. When the program is executed it read the target tables as input of Map. In this paper we need to Reduce because the relationship among the table contents is determined and further integration is also unnecessary.

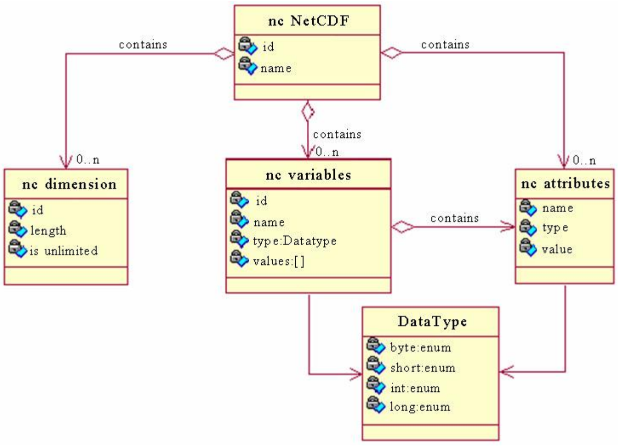


FIGURE 4 Description of implementation structure with NetCDF metadata

TABLE 1 Hbase oriented ETL interface

Parameter function	Name	type	Value range	method
Input	Inputtable	string	Name of input table	run()
	Inputcolumns	string	Row name of Hbase	run()
	Rowrule	string	Generated rules of key and value in each line	run()
output	Outtable	string	Name of output table	run()
	outputcolumns	string	Output row of Hbase	run()

The Hbased oriented ETL components contains DataETL interface, HbaseETL class, HbaseETLMapper class, HbaseETLServer ckass, HbaseETLClient class and ETL Serverclass. HbaseETL class provides DataETL interface and has dependencies with HbaseETLMapper; HbaseETLServer inherits HbaseETLServer and it calls HbaseETL class; HbaseETLClient class calls HbaseETL class.

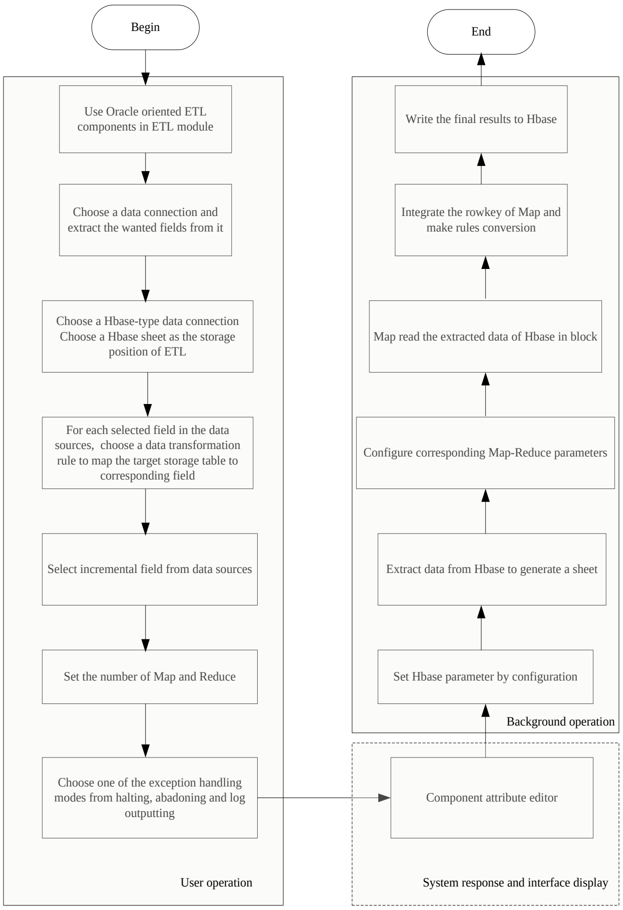

Begin
Use Oracle oriented ETL components in ETL module
Choose a data connection and extract the wanted fields from it
Choose a Hbase-type data connection Choose a Hbase sheet as the storage position of ETL
For each selected field in the data sources, choose a data transformation rule to map the target storage table to corresponding field
Select incremental field from data sources
Set the number of Map and Reduce
Choose one of the exception handling modes from halting, abadoning and log outputting
User operation
End
Write the final results to Hbase
Integrate the rowkey of Map and make rules conversion
Map read the extracted data of Hbase in block
Configure corresponding Map-Reduce parameters
Extract data from Hbase to generate a sheet
Set Hbase parameter by configuration
Background operation
Component attribute editor
System response and interface display

FIGURE 5 Hbase oriented ETL component flow

3.3 INFORMATION RELEASING MANAGEMENT

The marine surveying data platform can release the marine information in time to provide users expedition observations for browse and utilization. The releasing management module is in charge of the data maintenance and management. The components are depicted as figure 6.

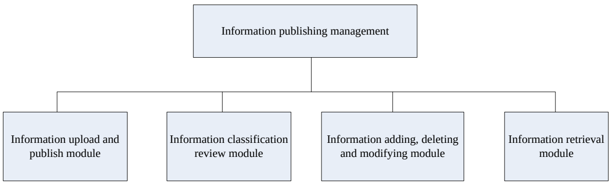
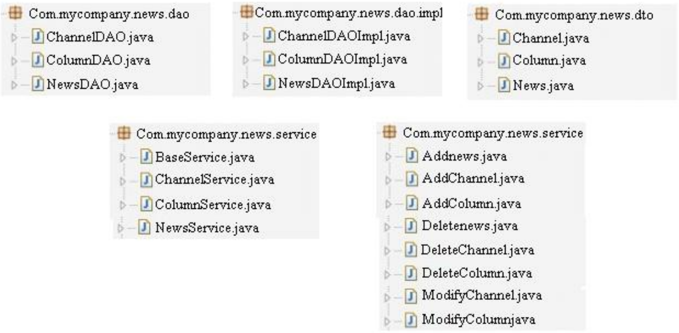

Information publishing management
Information upload and publish module
Information classification review module
Information adding, deleting and modifying module
Information retrieval module

FIGURE 6 Composition of information releasing

Module

In this figure, the information uploading function will update and release the marine data; information audit module will check the information to be released; the adding, deleting and modifying module are in charge of the information management; information retrieval module will retrieve the information according to given conditions. In specific programming the codes are encapsulated which only provide external interface. The packages are separated by class so the function of each packet is definitude. Our module establishes 5 kinds of packets of different use as shown in the following figure:



```
Com.mycompany.news.dao
  ChannelDAO.java
  ColumnDAO.java
  NewsDAO.java

Com.mycompany.news.dao.impl
  ChannelDAOImpl.java
  ColumnDAOImpl.java
  NewsDAOImpl.java

Com.mycompany.news.dto
  Channel.java
  Column.java
  News.java

Com.mycompany.news.service
  BaseService.java
  ChannelService.java
  ColumnService.java
  NewsService.java

Com.mycompany.news.service
  Addnews.java
  AddChannel.java
  AddColumn.java
  Deletenews.java
  DeleteChannel.java
  DeleteColumn.java
  ModifyChannel.java
  ModifyColumnjava
```

FIGURE 7 The package of information release management module

Com.mycompany.news.dao provides the external interface class; *Com.mycompany.news.dao.impl* contains the specific implementation of interface class; *Com.mycompany.news.dto* contains javaBean classes and *Com.mycompany.news.service* contains the service class. They integrate the operation class together; *Com.mycompany.news.servlet* contains specific operation classes. Part of the class interfaces in releasing function module is shown as follows:

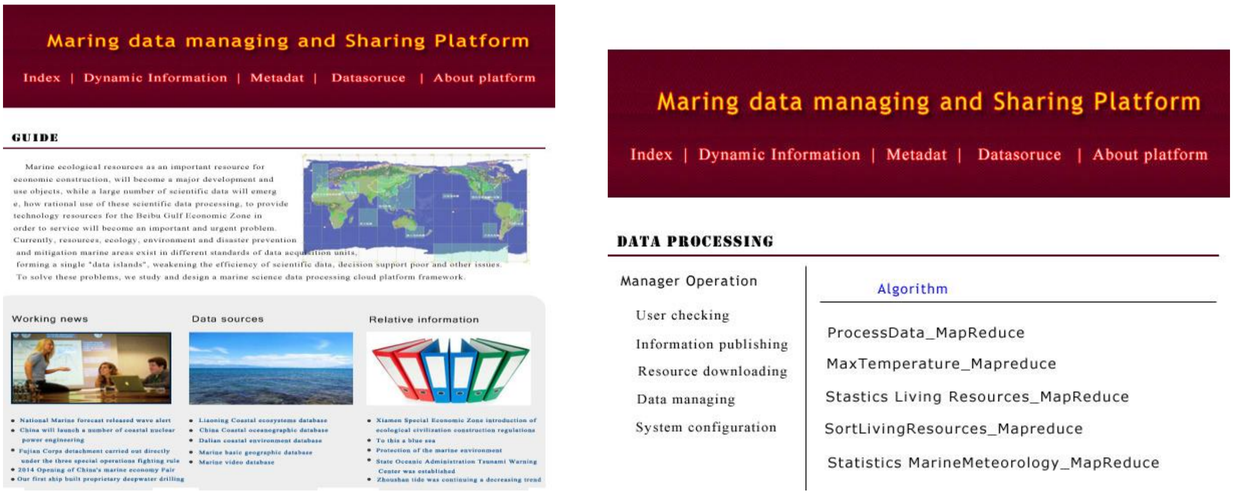
```
    Public void addnews (news news)
throws Exception ;  // Adding news
information
    Public void updatenews (news news)
throws Exception ; // Modifying news
information
    Public voiddeletenews (news news)
throws Exception ; // Deleting news
information
    Public List listAllnews () throws
Exception ; // listing allt he news
information
    Public List listnews (news
newscondition),
    Int curpage, int perpage throws
Execption;  //Combination of
conditions query
```

3.4 BACKUP AND RECOVERY MANAGEMENT

HDFS save each file as a series of data block and the default size is 64MB. We adopt the strategy of rack perception, that is, NameNode can determine each rack of DataNode in every platform. It brings high availability and reliability to the data, improving the efficiency of network width greatly. All the data blocks have duplicates so it can provide fault tolerance for backup and recovery. The strategy of duplications storing assume there are three copy factors. Each file has 3 copies while the rack nodes are divided into three types: the first is the node on local rack; the second is different nodes on the same rack; the third is the node on different racks. There is a file copy for each node of all these three types. Therefore, the whole cluster can read and recover data from another copy, which will not influence the reliability and availability of data.

4 System Test Results

According to the functional analysis and framework design of data on cloud platform of marine surveying data, this paper initially provides the storing and releasing function. It is a reliable and highly stable storage platform. The operational effect of platform is shown as figure 8(a) and the processing interface of platform data is shown as figure 8(b).



Maring data managing and Sharing Platform

Index | Dynamic Information | Metadat | Datasoruce | About platform

GUIDE

Marine ecological resources as an important resource for economic construction, will become a major development and use objects, while a large number of scientific data will emerg e, how rational use of these scientific data processing, to provide technology resources for the Beibu Gulf Economic Zone in order to service will become an important and urgent problem. Currently, resources, ecology, environment and disaster prevention and mitigation marine areas exist in different standards of data acquisition units, forming a single "data islands", weakening the efficiency of scientific data, decision support poor and other issues. To solve these problems, we study and design a marine science data processing cloud platform framework.

Working news

- National Marine forecast released wave alert
- China will launch a number of coastal nuclear power engineering
- Fujian Corps detachment carried out directly under the three special operations fighting rule
- 2014 Opening of China's marine economy Fair
- Our first ship built proprietary deepwater drilling

Data sources

- Liaoning Coastal ecosystems database
- China Coastal oceanographic database
- Dalian coastal environment database
- Marine basic geographic database
- Marine video database

Relative information

- Xiamen Special Economic Zone introduction of ecological civilization construction regulations
- To this a blue sea
- Protection of the marine environment
- State Oceanic Administration Tsunami Warning Center was established
- Zhoushan tide was continuing a decreasing trend

(a) Information releasing platform

Maring data managing and Sharing Platform

Index | Dynamic Information | Metadat | Datasoruce | About platform

DATA PROCESSING

Manager Operation

User checking

Information publishing

Resource downloading

Data managing

System configuration

Algorithm

ProcessData_MapReduce

MaxTemperature_Mapreduce

Stastics Living Resources_MapReduce

SortLivingResources_Mapreduce

Statistics MarineMeteorology_MapReduce

(b) Data processing platform

FIGURE 8 Rendering results of information and data processing

To calculating the highest temperature data of marine within certain time, our using dataset is the meteorological data from national meteorological data center NCDC. The test contents are:

(1) When the data quantity keeps unchanged, the data processing efficiency of platform framework structure project and platform stability in this paper are stable with the change of processing node numbers.

(2) When the processing nodes number keeps unchanged, the data processing efficiency of platform framework structure and platform stability in this paper are tested to testify the effect and feasibility of platform, with changing quantity of data.

The first test is performed with unchanged data quantity and increasing processing node number. The test results are shown as table 1; The second test is performed with unchanged processing node number and increasing data quantity. The test results are shown as table 2.

Table 1 Test results of the first experiment

Group	Data size	Single time-consuming	Hadoop cluster time-consuming	Saved time
1	5G	340s	256s	94s
2	10G	668s	421s	247s
3	15G	1014s	723s	291s
4	20G	1354s	936s	418s
5	40G	2731s	1632s	1069s

Table 2 Test results of the second experiment

Node number	Data quantity	Maps	Reduces	Running time	Stability
4	10G	80	4	8	Normal
4	15 G	120	4	12	Normal
4	20 G	160	4	18	Normal
4	50G	200	4	25	Normal

On the basis of above test results, at first, under the condition that data quantity is defined, the processing efficiency of system platform will be also improved with increasing processing nodes. Meanwhile, it is not abnormal during operating so the system platform has stronger stability. With certain number of nodes and increasing data quantity, the system platform keeps higher processing efficiency. Meanwhile, there is not abnormal during operating so system platform has stronger stability. Therefore, the implementation results and operational test results of platform system module show that our system has strong expansibility and is easy for maintenance. The research design method and technological development route are effective and feasible. Meanwhile, the feasibility and effectiveness of the proposed platform framework are also verified.

Next, when Hadoop cluster is processing the information, under the condition that data size is constantly increasing, the time consumption of Hadoop cluster processing will be much less than by single machine. With constantly increasing data, the time for data process of single machine will basically be in the state of linear increase. However, when the distributed Hadoop cluster is processing larger and larger data, the processing time increase will be relatively slow. When processing massive data such as data in oracle, the advantage of Hadoop cluster will be more obvious than single machine processing. From contrast of above experiment data, we can arrive at this conclusion. Hadoop system which is introduced from massive data processes actually improves the processing efficiency. Meanwhile, the idea of MapReduc is used to extract corresponding *<key,value>*. This processing method accords with features of mass web journals, complicated structure and difficult understanding. With the increasing quantity of data, Hadoop cluster will fully play its part in its distributed calculation and the time consumption is relatively less.

Finally, previous marine scientific data platform uses more technologies like traditional parallel calculation, distributed calculation, etc. Relational database technology is used to process storage data. Therefore, calculation and storage resources are expensively allocated, processing efficiency is inefficient and reliability cannot be guaranteed, etc. Table 3 shows the comparison results between cloud platform system in this paper and current marine scientific data platform.

TABLE 3 Comparison of cloud-based platform in this paper with past marine data storage system

Characteristic classes	Traditional marine data managing platforms	Hadoop-based data managing platform
Design idea	Data sharing and high performance computing	General computing and storage
Component	Advanced computers	Cheap computers
Function	Single	Rich and scalable
Performance	Low efficiency	High efficiency and reliability
Capacity	Variable but limited	On demand
Resource	Non-virtualized	virtualized
Application type	Scientific computation	Data processing

In summary we can see that the cloud platform system in this paper mainly adopts Hadoop distributed technology to process massive marine data and it can reach the object of efficient managing and storing. Meanwhile, for R & D personnel's, the program is easily realized without tedious programming.

5 Conclusion

This paper designs and develops a Hadoop-based platform for massive marine data processing. It adopts Linux cluster, parallel distributed database technology, Hadoop distributed platform and HDFS distributed file system, Map/Reduce parallel computing model and HBase database technology as the major basis. By means of tests on lots of common computers with the construction of the platform in this paper, it is found that this project meets the demand for efficient storage and management for massive marine scientific data. In addition, this system has perfect expansibility and it is easy for maintenance. Therefore, the technical route and design method adopted by us are effective and feasible.

References

[1] YANG Haichao 2009 Research on distributed spatial data sharing model based on metadata **12**(4) 24-7

[2] Wang Juanle, You Songeai 2004 Study on web-oriented geo-data sharing infrastructure and key techniques based on metadata *In Proceedings of International Geoscience and Remote Sensing Symposium* 4448-51

[3] Sarathy Rathindra, Muralidhar Krishnamurty 2006 Secure and useful data sharing *Decision Support Systems* **42**(1) 204-20

[4] Antoniu Gabriel, Bouziane Hinde Lilia, Jan Mathieu 2007 Combining data sharing with the master-worker paradigm in the common component architecture *Cluster Computing* **10**(3)265-76

[5] Gao Feng, He Jingsha, Ma Shunan 2011 Privacy preserving in data sharing applications *Journal of Southeast University* **41**(2) 233-6

[6] DENG Wei, LIU FangMing, JIN Hai 2013 Leveraging Renewable Energy in Cloud Computing Datacenters State of the Art and Future Research *Chinese Journal of Computers* **36**(3)582-7

[7] Samanthula Bharath K, Jiang Wei 2013 Efficient privacy-preserving range queries over encrypted data in cloud computing *In the Proceedings of IEEE International Conference on Cloud Computing* 51-8

[8] XU Xiaolong, YANG Geng, LI Lingjuan 2013 Dynamic data aggregation algorithm for data centers of green cloud computing *Systems Engineering and Electronics* **34**(9) 1923-9

[9] HAO Wei, ZHUO Wei, LI Zhanbo 2013 A novel data exchange architecture based on cloud computing *Computer Engineering & Science* **35**(8)15-9

[10] HU Changjun, YANG Jing, GE Jingjun 2013 Research and Implementation of a Science Data Cloud and Services *Journal of Chinese Computer Systems* **5**(5) 1028-34

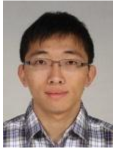
[11] Sadasivam G. Sudha, Selvaraj Dharini 2010 A novel parallel hybrid PSO-GA using MapReduce to schedule jobs in Hadoop data grids *In the Proceedings of 2nd World Congress on Nature and Biologically Inspired Computing* 377-82

[12] CHEN Jirong, Jiaji L E 2013 Reviewing the bid data solution based on Hadoop ecosystem, *Computer Engineering & Science* **35**(10)25-35

Authors

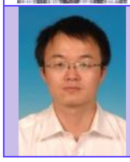
Hongxin Zhang, born in January 1978, Qingdao City, Shandong Province, P.R. China

Current position, grades: Senior Engineer, vice Chief of Center, North China Sea Marine Technical Support Center of State Oceanic Administration
University studies: Graduated from Ocean University of China in 2000, Bachelor of Science in Physical Oceanography
Scientific interest: Physical Oceanography, Marine Information,ship information service system
Publications : More than 10 scientific research projects, More than 20 papers published in various journals. Six patents for an invention
Experience: Graduated from Ocean University of China in 2000
Work as a engineerand Senior Engineer r in North China Sea Marine Technical Support Center of State Oceanic Administration



Kanghong Duan, born on April 1987, Zhucheng City, Shandong Province, P.R. China

Current position, grades: engineer, North China Sea Marine Technical Support Center of State Oceanic Administration
University studies: Graduated from University of Science and Technology, received a master's degree in Computer Science
Scientific interest: Internet of Things, Marine Information, Embedded operating system
Publications : More than 10 papers published in various journals. Three patents for an invention
Experience: Graduated from University of Science and Technology, received a master's degree in Computer Science
Work as a engineer in North China Sea Marine Technical Support Center of State Oceanic Administration



Xiaobo Zhang, born on March 1979, Qingdao City, Shandong Province, P.R. China

Current position, grades: engineer, North China Sea Marine Technical Support Center of State Oceanic Administration
University studies: Graduated from Ocean University of China , received a master's degree in Physical Oceanography
Scientific interest: Marine Information
Publications : More than 5 papers published in various journals. Three patents for an invention
Experience: Graduated from Ocean University of China
Work as a engineer in North China Sea Marine Technical Support Center of State Oceanic Administration