Construction of Marine Environmental Management Information Sharing Platform Under the Background of Big Data

Dandan Lv(✉)

School of Economics and Management, Dalian University, Dalian, Liaoning, China
591194793@qq.com

Abstract. To realize the comprehensive management of China's oceans, and to make contributions to China's strategy of strengthening the ocean and sustainable development. By combing the current situation of domestic marine environmental information processing problems, combined with big data technology, this paper makes some thoughts on the construction of marine environmental information sharing, and scientifically plans and develops a management information sharing platform that utilizes the ocean and provides data for the development of marine science. This system uses Hadoop, Spark big data, Python web crawler and JavaWeb page development technologies to collect and process data related to the marine environment, and presents efficient, real-time and shared data effects, which adds a strength to China's great dream of realizing a smart ocean, a beautiful ocean and an ecological ocean.

Keywords: marine environment · Hadoop technology · information sharing · big data

1 Introduction

The ocean is a treasure house rich in material resources and energy. China has a vast sea area and tens of thousands of kilometers of coastline. It is of great significance to make full use of marine resources to develop marine economy to improve the living standard of our people and the economic competitiveness of our country. At the same time, the ocean is also a dynamic and time-varying system with huge complexity, interaction of various factors and interaction of various processes. When human beings are faced with a huge and complex ocean system, there are some problems, such as immature mastery of environmental laws, low efficiency of development and utilization, frequent disputes over maritime rights and interests, etc. These problems show that human beings have unclear cognition, improper response and lack of wisdom about the ocean. Therefore, China has implemented the "Digital Ocean" project, and many Internet innovation companies have established a large number of professional marine environmental information management systems. The central and local governments and small and medium-sized marine enterprises have already started the work of marine information management.

Z. Zhan et al. (Eds.): EIMSS 2022, AHCS 7, pp. 437–447, 2023.
https://doi.org/10.2991/978-94-6463-024-4_46

These systematic applications of marine data and information provide strong technical support and information guarantee for the development of China's marine industry. However, the marine environmental information management systems of different enterprises, departments and government units based on "Digital Ocean" project in China are similar internally but independent, and there are some problems, such as redundant construction, low resource utilization rate, high operating cost, lack of unified planning of computing resources, and difficulty in system maintenance. The traditional business model has been difficult to meet the requirements of rapid business deployment. How to realize the sharing of marine ecological environment information is an important issue that needs to be examined to ensure the steady improvement of ecological environment quality [9].

The society is marching towards the era of big data. With the rapid development of information technology, the degree of social intelligence informatization is deepening, and the data resources and application markets are abundant, which leads to the explosive growth of information data. People realize that these fast-growing big data are of great value for the analysis of project realization, so people's demand for a large amount of data processing is increasing day by day, and they are eager to manage a large amount of data reasonably and transform it into network information resources that can be effectively used, thus the big data technology was born. The key technologies of big data R&D are constantly developing in the process of evolution, which makes it easier to collect data. At present, the emerging big data technology puts forward a brand-new resource allocation and service management mode of on-demand distribution, which virtualizes all computing resources into services and provides the functional realization technology of massive data sharing. Therefore, we can use big data technology to build a framework system of marine environmental information management and sharing, and can provide services to end users on demand through a standard mode and effectively integrate various marine business systems to achieve the purpose of marine environmental information and data sharing [8].

According to the above contents, the author of this paper believes that the marine environment information management sharing platform system developed based on Hadoop big data technology, Python web crawler technology and Java Web page technology is a big data platform that shares marine information such as marine resources, climate environment, water quality and ecology, and integrates the functions of marine data query. The platform can effectively solve the problems existing in the construction of other marine information management systems mentioned above, and establish a low-cost operating environment to support marine environmental information services, which will save investment, improve resource utilization, operation and maintenance efficiency, be green and energy-saving, and greatly improve the reusability and sharing of marine resource information, as well as the scalability of application systems. It has great scientific, social and potential economic benefits for realizing the comprehensive management of China's oceans, raising people's awareness of the oceans, and realizing the strategy of strengthening the ocean and sustainable development.

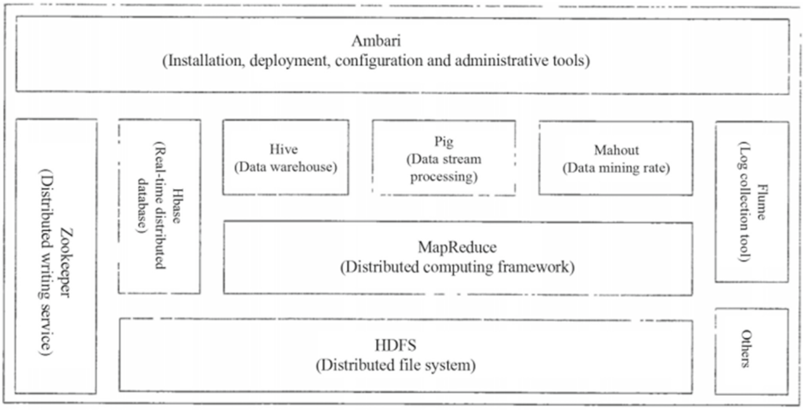

Fig. 1. Hadoop ecosystem hierarchy diagram

2 Key Technology Introduction

2.1 Hadoop Ecosystem

Apache Hadoop software library is a programming framework, which allows simple programming models to be used in distributed clustered computers that process large data sets. It is designed to simply and quickly expand from a single server to thousands of machines, with each machine providing local computing and storage. The fault tolerance of the system does not depend on the high reliability of the hardware, and it is designed to handle the frequent hardware errors and crashes on ordinary hardware servers, and can ensure the high availability of the whole system by automatically detecting and handling faults. Hadoop family is very large, among which Hadoop Common, Hadoop Distributed File System (HDFS) and MapReduce framework constitute the core of Hadoop, and there are a lot of researches related to Hadoop. Such as structured distributed database Hbase, extensible data warehouse Hive, data stream language Pig, Chukwa, which manages data collection of distributed systems, and Zookeeper, which solves the problem of system consistency. A series of projects together constitute the Hadoop ecosystem, and the hierarchical relationship among each project is shown in Fig. 1 [3].

2.2 Spark

Spark is a big data computing framework, and it is expected that all kinds of computing tasks in the field of big data can be solved perfectly by using a technology stack. Apache's official definition of Spark is: a universal big data fast processing engine. Spark successfully solved the most important tasks and problems in the field of big data, such as offline batch processing, interactive query, real-time stream computing, machine learning and graph computing, by using Spark RDD, Spark SQL, Spark Streaming, MLlib and GraphX. In addition to the one-stop feature, the other most important feature of Spark is that it calculates based on memory, so that its speed can reach several times or even dozens of times of MapReduce and Hive. Spark is a general big data computing

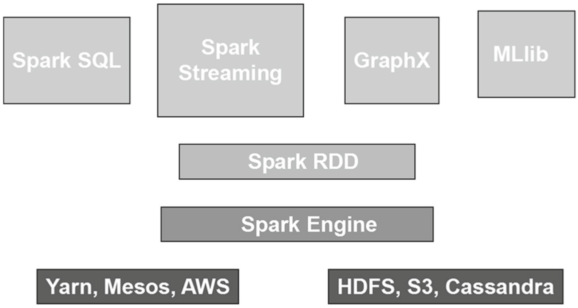

Fig. 2. Spark architecture diagram

framework, just like the MapReduce and Hive engines of Hadoop, the traditional big data technology, and the Storm streaming real-time computing engine. Spark includes various computing frameworks common in the field of big data, such as Spark Core for offline computing, Spark SQL for interactive query, Spark Streaming for real-time streaming computing, Spark MILlib for machine learning, and Spark GraphX for graph computing. Spark is mainly used for the calculation of big data, while Hadoop is mainly used for the storage of big data (such as HDFS, Hive, HBase) and resource scheduling (Yarn). The combination of Spark+Hadoop is the hottest and most promising combination in the future big data field [5] (Fig. 2).

2.3 HDFS Distributed Storage

First of all, HDFS is a file system, which is used to store files and locate files through a unified namespace-directory tree. Secondly, HDFS is distributed, and many servers are combined to realize its functions. The servers in the cluster have their own roles. Important features are as follows: files in HDFS are physically stored in blocks, and the size of blocks can be specified by the configuration parameter (dfs.blocksize). HDFS file system will provide clients with a unified abstract directory tree, and clients can access files through paths. The management of directory structure and file block information and metadata is undertaken by namenode-NameNode is the main node of hdfs cluster, which is responsible for maintaining the directory tree of the whole HDFS file system, and the block information (the id of the block and the datanode server) corresponding to each path (file). The storage management of each block of the file is undertaken by datanode node. HDFS is designed to adapt to write once, read many times, and does not support file modification.

The client sends the path of the file to be read to namenode, and namenode obtains the meta information of the file (mainly the storage location information of the block) and returns it to the client. According to the returned information, the client finds the corresponding datanode to obtain the file blocks one by one, and adds and merges the data locally at the client to obtain the whole file hdfs. In production applications, HDFS is mainly the development of the client. Its core step is to construct an access client object of hdfs from the api provided by HDFS, and then operate (add, delete and check) the files on HDFS through this client object. As shown in Fig. 3.

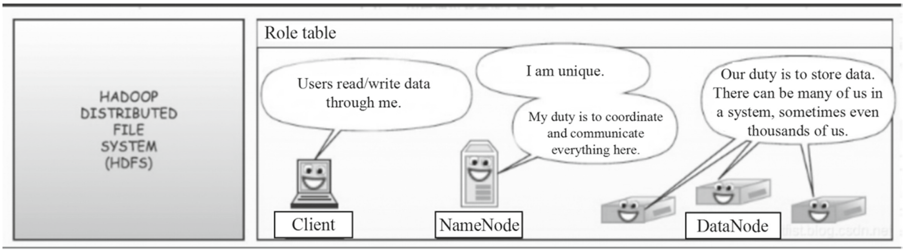

Fig. 3. HDFS file system structure diagram

2.4 Development Environment

The marine environment information sharing platform based on big data is divided into Hadoop-based big data server clusters for data processing and storage in MySQL database. JavaWeb technology is used to call the data and develop the application platform. According to the data volume and overall operation requirements of the system, a total of four cluster servers are used to establish Hadoop cluster, of which one node is Master.Hadoop master node and the other three nodes are slave nodes. Components that Hadoop cluster needs to run include Hive 1.2, Mapreduce, Sqoop-1 and HBase 2.4.0. Hadoop and its functional components are synchronously installed and deployed on four nodes to complete the construction of Hadoop cluster. The cluster will be developed under Linux system. This paper chooses the Linux distribution version of CentOS-8.2, the community enterprise operating system. In addition to Mapreduce, the data calculation engine of Spark is assisted by version 2.4.7 of SPARK. When using web crawler technology to collect data, Scrapy 2.5 framework is used, the development environment is Python 3.8, and the development tool used is VIM. This platform realizes many functions of data processing based on Hadoop 3.3 cluster version.

The Java development tool used by the JavaWeb application of this system is Eclipse Eclipse-SDK-4.5.0-Linux-GTK-x86_64, the development environment is JDK 1.8, the development language is Java, and Apache Tomcat 9.0 is selected for server construction. SpringFramework 5.3.20 is used in the development framework, and MySQL 8.0.28 is used to help manage data. Through the introduction of the above key technical theories, we have determined the overall environment of system development, the configuration of related software and tools, and the technical feasibility of the overall project.

3 Requirement Analysis

3.1 Function Requirement

The users of the marine environmental management information sharing platform based on the background of big data include government departments such as the central and local fisheries departments and environmental protection bureaus, private enterprises involved in water conservancy and fishery industries, and non-profit organizations such as the Natural Resources Conservation Association. As the participants of information sharing in marine environmental management, many marine military drills, fishing, water conservancy projects, sewage treatment, and marine environmental animal protection activities all need a lot of data support, and to a great extent, they all rely on the acquisition of marine environmental management information [7]. According to the user's need to query ocean information, this system develops five functional modules: personal center, real-time update data, historical information data, data query and industry news. In the personal center module, users can manage their own basic information and their favorite information content. This system divides all available marine environment related data into static data such as dynamic data updated in real time and historical data, and according to these data, it develops a real-time updating marine data function module and a historical data summary module. Real-time data include seawater quality monitoring information and marine climate monitoring information, which is helpful for the government, organizations and enterprises to prepare for various marine activities. Climate information and water quality information are important factors to ensure the successful development of many activities. The historical data summary module includes all the collected marine data, including marine chemistry, marine biology, marine geology and other aspects of marine environmental data, and supports the in-depth research of various projects of the government, organizations and enterprises. The data query function module developed by this system can facilitate users to accurately find the specific data they need in a large amount of marine environmental data. The information of the industry news module can help users understand the latest dynamic policies of the industry, and timely modify the decision-making and operation modes of each project according to the market conditions. The marine environment information sharing system establishes a low-cost test and operation environment for various projects and businesses supporting the marine environment, which will save investment, improve business support capacity, improve operation and maintenance efficiency, reduce investment risk and decision risk, and be green and energy-saving. It can also greatly improve the reusability and sharing of marine resources information by various government departments, enterprises, non-profit organizations and other institutions [2].

3.2 Global Design

Under the background of data, the overall design of marine environment information sharing platform is based on the concept of layering, which is conducive to the overall function and technology planning of the system. The source part is mainly the ocean data collected by Python web crawler technology through the information management platform developed by private enterprises, non-profit organizations and government departments at all levels, and stored in the distributed file storage system HDFS. Mapreduce and Apache Spark, the big data processing components, analyze and correlate the ocean data in HDFS, and store the processed data in HIVE data warehouse and HBase, and then transfer it to MySql database through Sqoop [6].

The main application form of this platform is realized by Javaweb technology, and users can perform the functions of subsystems such as data reading, data management, search and query through the Web interface. The search engine function used by users is realized through the process of making a request to the server through the web page, calling the data stored in MySQL database and cleaned by Hadoop, and then returning it to the client (Fig. 4).

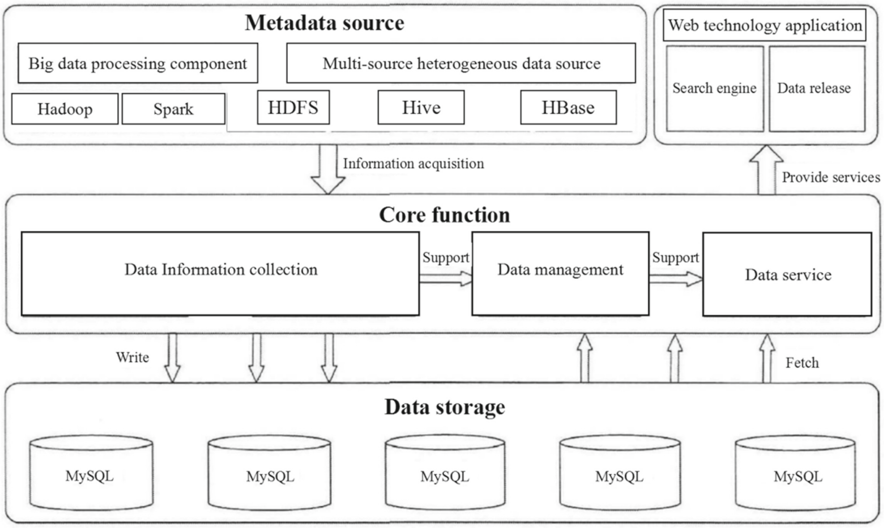

Fig. 4. The overall architecture of the platform

4 Function Implementation

The marine environment management information sharing platform based on the background of big data establishes a system for different needs of users and system administrators. After the account applied by the user passes the password verification, log in to the system. It can be seen that the system consists of personal center basic module and four main functional modules: real-time monitoring data, historical data summary, ocean data query and industry consultation news. The functional realization of the following four main functional modules is introduced in detail.

4.1 Real-Time Monitoring of Data Update

This module is supported by Spark software as a big data computing framework technology. Spark is a programming framework that supports distributed parallel computing, which reduces the landing of data in the iterative process, can update data in real time and quickly, and improves the processing efficiency. This part of data includes monitoring information of seawater quality and marine climate. Among them, the seawater quality monitoring information data comes from the seawater quality monitoring information disclosure system of the National Marine Environment Monitoring Center. According to the "Seawater Quality Standard" (GB 3097–1997), the daily updated data of six indexes of pH, dissolved oxygen, chemical oxygen demand, inorganic nitrogen, active phosphate and petroleum in various sea areas are published. The ocean climate information comes from the National Ocean Forecasting Station and the National Ocean Observation Network. This part of the data includes the real-time update of sea temperature, current direction, current velocity, wave height, swell, tide, salinity and other indicators [10].

4.2 Historical Data Summary

This module summarizes and classifies all kinds of public historical marine information such as central and local government agencies, non-profit organizations and private enterprises in related industries. It is subdivided into seven major modules: marine hydrology, marine meteorology, marine biology, marine chemistry, marine sediment, marine geophysics and submarine topography. Hydrology includes marine hydrological data obtained by observation means such as stations, buoys and survey ships, mainly including temperature, salinity, waves, water level, current and other elements. Marine meteorological data module includes data of sea surface temperature, air pressure, wind direction, wind speed, wind direction, geopotential height, etc. Chemical data cover the Atlantic Ocean, Pacific Ocean and other sea areas, including dissolved oxygen, PH value, alkalinity, nitrate, nitrite, heavy metals, suspended solids and other marine water quality data. Marine biological data include zooplankton, phytoplankton, fish, shellfish and other data. Marine sediment data set includes mineral resources, polymetallic nodules, cobalt-rich crusts, sediments, C14 dating, grain size, oxygen isotope and other elements. Seabed topographic data sets are collected from data published by different organizations abroad, and are raster elevation topographic data covering the whole world's oceans and lands (Fig. 5).

Here, the source code of uploading ocean information data files to HDFS file system under big data platform will be displayed, which is a part of the functional realization code of data collection, as shown in Fig. 6 [4].

4.3 Data Query

In order to make it convenient for users to query all kinds of complicated marine data accurately and quickly, this system has built in a comprehensive search engine. Users can query the data by year, search the provinces, sea areas, ports and fishing grounds in detail, and query specific parameters such as climate and water quality. The normal operation of the search engine function in the system depends on the data processing layer, which provides the query result feedback service. The query results are finally fed back to the user layer by the web server, thus ensuring that the requests made by users are directly converted into query results [1].

	Major institutions	Information sharing responsibilities
Central level	Ministry of Environmental Protection	Undertake the responsibility of monitoring, evaluation and information release of marine economic operation.
	National Marine Data and Information Service	Manage the national marine information resources and coordinate the national marine information business.
	Central Environmental Monitoring Station (Marine Room)	Collect and review the monitoring data of marine environmental quality of coastal Haicheng, rivers into the sea, pollution sources directly discharged into the sea, etc. in the country, and prepare reports on pollutant discharge in coastal waters and rivers into the sea.
	National Marine Environmental Monitoring Center	Establish and manage the national marine environment monitoring database, and audit and manage the monitoring data; Establish the marine environmental monitoring, sea area use management information system (into the national marine integrated information business system), etc.
	National Marine Environmental Forecasting Center	Responsible for the collection, processing, quality control and storage of global marine hydrometeorological data, marine monitoring stations, ships, buoys, radar stations, aviation monitoring, satellite remote sensing and fax data and real-time data of public information network, and the distribution service of real-time data of national marine forecasting system, and the establishment of marine disaster and forecasting data database and information system.
Local level (Take Shandong Province, for an example)	Shandong Provincial Oceanic and Fishery Department	Responsible for issuing the bulletin of Qingdao environmental situation, etc.
	Qingdao Environmental Protection Bureau	Undertake the responsibility of marine economic operation monitoring, evaluation and information release.
	Weihai Ocean and Fisheries Bureau	Responsible for issuing the bulletin of Weihai's environmental status, marine ecological status and marine fishery status, etc. Collect, review and statistically analyze marine data such as offshore marine fishing and freshwater aquaculture.
	Yantai Ocean and Fisheries Bureau	Take the responsibility of monitoring, evaluating and releasing information of marine economic operation, and be responsible for statistics, accounting and information construction of marine and fishery economy.

Fig. 5. Data sources of marine environmental government agencies

4.4 Marine Industry Information and Important News

This part of data is mainly collected from various government information released by the Ministry of Natural Resources of the People's Republic of China. The data is divided into four parts: policies and regulations related to marine environment, planning plans, standards and specifications, and notices of key meetings issued by the Central Committee of the Communist Party of China and the State Council, which help users understand the latest trends of national marine environment.

```
public class HdfsClient {
        FileSystem fs = null;
        @Before
        public void init() throws Exception
                // Construct a configuration parameter object and set a parameter: the URI of hdfs we want to access
                // Thus, the FileSystem.get () method knows that it should construct a client to access the hdfs file system and the access
address of hdfs
                // new Configuration(); , it will load hdfs-default.xml in the jar package
                // And then load hdfs-site.xml under classpath
                Configuration conf = new Configuration();
                conf.set("fs.defaultFS", "hdfs://hdp-node01:9000");
                // Parameter priority: 1. the value set in the client code 2. User-defined configuration file under classpath 3. Then the default
configuration of the server.
                conf.set("dfs.replication", "3");
                // Get an access client of hdfs. According to the parameters, this instance should be an instance of DistributedFileSystem
                fs = FileSystem.get(new URI("hdfs://hdp-node01:9000"), conf, "hadoop");
        }
        //Upload files to hdfs
        @Test
        public void testAddFileToHdfs() throws Exception {
                // The local path of the file to be uploaded
                Path src = new Path("g:/redis-recommend.zip");
                // Destination path to upload to hdfs
                Path dst = new Path("/aaa");
                fs.copyFromLocalFile(src, dst);
                fs.close();
        }//Copy files from hdfs to local file system
```

Fig. 6. Code realization of HDFS uploading data files

5 Conclusions

The marine environment information sharing platform based on big data is conducive to the convenient, orderly cooperation and effective sharing of information among marine-related industries, reducing the blind spots of marine environmental governance, building a good marine ecological environment, enhancing the rational level of government environmental governance, and constraining the opportunistic tendencies of enterprises in various jurisdictions. At present, the marine environment information sharing platform still has some shortcomings in overall planning and sharing security, and it is urgent to develop more advanced modern marine information technology for high-quality marine development and utilization and protection of marine rights and interests. After that, it is expected to better serve the sustainable development of the ocean.

Project Fund. 2021 Social and economic development research project of Liaoning Province.

Study on coordinated management of Marine ecological security in Liaoning Province in ecological Civilization construction 2021lslybkt-090.

References

1. Chen Q (2019) Analysis on the development of marine information technology in the new era of big data. Modern Econ Inf 06
2. Fei S (2014) Research on collaboration between Hadoop big data platform and traditional data warehouse. Donghua Univ 05
3. Jiang W (2009) Data format analysis and data reprocessing of ocean environment information in Ocean University of China. Ocean Univ China
4. Liu Y, Ouyang Y et al (2021) Development and thinking of marine information technology. Ocean Dev Manage 10
5. Lv J, Zhang S (2018) Construction of information sharing mechanism of marine environmental management in China under the background of internet. J Ocean Univ China (Soc Sci) 01
6. Meng L (2016) Development and research of marine environment monitoring information system based on B/S and C/S models. Ship Sci Technol 04
7. Wang J (2013) Research on visualization platform technology of marine environment information based on cloud computing. Ocean Univ China 05
8. Xing H, Lin P (2020) analysis of collaborative disclosure mechanism of marine ecological environment information in Guangdong, Hong Kong, Macao and Dawan District. J Jiangxi Univ Sci Technol 12
9. Yuan Y, Jin J, Li H (2019) Design of distributed architecture of marine environment information based on Hadoop. Marine Geodesy Cartogr 11
10. Zhang H (2020) Design and implementation of ocean information management service platform for smart ocean. Sci Technol Innov Herald 2