Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 15, No. 3, 2024

Design and implementation of the big data processing system based on cloud computing

Ning Ye

Akamai (Beijing) Technologies Co., Ltd., Beijing 100000, China

Abstract: As a product of the development of information technology, cloud computing technology has outstanding advantages in data processing. Therefore, cloud computing technology has been widely used in the field of big data processing. Building a big data processing system based on cloud computing technology can not only improve the efficiency of data processing, but also effectively reduce the cost of data processing. In this paper, the design and implementation of the big data processing system based on cloud computing are taken as the research object, which first introduces the relevant knowledge of cloud computing and big data processing technology, then analyzes the requirements of the system, and designs the overall architecture and functional module of the system, and finally the system is tested. The results show that the system can effectively handle large data with good stability and scalability.

Keywords: Cloud computing; Big data processing and data processing system.

1. Introduction

With the rapid development of information technology, data has become an important asset. Enterprises and organizations are faced with great challenges when dealing with big data. How to efficiently and accurately process and analyze big data and dig out its value has become an urgent problem to be solved. As an emerging computing mode, cloud computing provides a new solution for big data processing. Cloud computing technology has powerful computing power and storage capacity, which can quickly process and analyze large-scale data. This paper will design and implement an efficient and scalable big data processing system based on cloud computing technology to help enterprises and organizations to better process and analyze big data.

2. Summary of relevant theories

2.1. An Overview of Cloud Computing

Cloud computing is an Internet-based computing method, through which the Internet is the huge computing processing program automatically divided into countless smaller subroutines, and then transferred to a huge system composed of multiple servers after search, calculation and analysis to send the processing results back to users. Cloud computing technology distributes computing tasks over a large number of computers, enabling various application systems to access computing power, storage space, and information services as needed. Cloud computing technology is characterized by virtualization, flexible extension, pay-on-demand, and can provide users with efficient and convenient computing services. Cloud computing technology has a wide range of application scenarios, including big data processing, artificial intelligence, the Internet of Things and other fields. In the field of big data processing, cloud computing technology can provide users with powerful computing power and storage space, and help users to efficiently process and analyze large-scale data. At the same time, cloud computing technology can also provide a variety of data processing tools and services, such as data mining, machine learning, etc., to meet users' needs for data processing and analysis.

2.2. Big data processing technology

Big data processing technology refers to a series of technologies to collect, store, manage and analyze massive data. The goal of big data processing technology is to quickly and accurately extract valuable information from massive data to provide support for decision-making. Big data processing technology includes data collection, data storage, data processing, data analysis and other links. In the data collection process, a large amount of data can be obtained through various sensors, web crawler and other means. In the data storage process, distributed storage systems and other technologies need to be used to store massive data. In the data processing process, it is necessary to use various algorithms and models to clean, transform and aggregate the data. In the data analysis process, data mining, machine learning and other technologies can be used for in-depth analysis of data to extract valuable information.

3. Design of a big data processing system based on cloud computing

3.1. System requirements analysis

3.1.1. Data acquisition requirements

Data acquisition is the first step of big data processing, so the system needs to support a variety of data acquisition methods, including real-time data acquisition and batch data acquisition. Real-time data collection can obtain real-time data through sensors, web crawler and other means, and process and store them. Batch data collection is to import a large amount of historical data into the system for subsequent analysis and processing. In order to meet the needs of different users, the system also needs to support a variety of data formats, such as CSV, JSON, XML, etc.

3.1.2. Data storage requirements

Because of the huge amount of data involved in big data processing, the system needs to have an efficient storage capacity. The system needs to support distributed storage technologies, such as Hadoop HDFS, to ensure data reliability and scalability. At the same time, the system also needs to support data compression and weight removal and other

functions to reduce the occupation of storage space.

3.1.3. Data processing requirements

Data processing is the core link of big data processing, so the system needs to support a variety of data processing algorithms and models, such as data mining, machine learning, etc. Systems need to be able to quickly process and analyze large-scale data to extract valuable information. At the same time, the system also needs to support data cleaning, conversion, aggregation and other operations to meet the needs of different users.

3.1.4. Data analysis requirements

Data analysis is an important part of big data processing, which can help users mine the value in data and provide support for decision making. The system needs to provide various data analysis tools and services, such as data visualization, report generation, etc. At the same time, the system also needs to support two modes of real-time analysis and offline analysis to meet the needs of different users.

3.1.5. System expansibility requirements

Due to the increasing amount of data involved in big data processing, the system needs to have good scalability. The system needs to support horizontal and vertical expansion to ensure that the system can expand with the amount of data. At the same time, the system also needs to support the multi-tenant mode to meet the needs of different users.

3.2. 2.2 Overall system architecture design

The overall architecture design of the big data processing system based on cloud computing is shown in Table 1.

Table 1. System Architecture

Administrative levels	Components / Technology	Functional description
Cloud computing platform	IaaS, PaaS	Provide infrastructure as a service, platform as a service, to provide the underlying support for big data processing
The Big data processing framework	Hadoop, Spark	Realize distributed storage and parallel processing to improve the efficiency of data processing
Data storage layer	HDFS, HBase	Provide distributed file system and high-performance column database for efficient data storage and management
Data processing layer	MapReduce, SQL	Provide data cleaning, conversion, aggregation and other processing functions, and support complex data processing logic
Data application layer	BI tools, and data visualization	Provides data analysis and visualization capabilities to support data-driven decision making

The system mainly includes five parts: data acquisition module, data storage module, data processing module, data analysis module and system management module. Among them, the data acquisition module is responsible for collecting real-time data and batch data and storing them in the data storage module. The data storage module adopts distributed storage technology to ensure data reliability and scalability. The data processing module is responsible for processing and analyzing large-scale data to extract valuable information. The data analysis module provides a variety of data analysis tools and services, such as data visualization, report generation, etc. The system management module is responsible for the management and maintenance of the system, including user management, authority management, system monitoring and other functions.

3.3. System functional module design

3.3.1. Design of the data acquisition module

The data acquisition module is responsible for collecting real-time data and batch data and storing them in the data storage module. The design of data acquisition module requires support of data acquisition methods and data formats. Real-time data collection can obtain real-time data through sensors, web crawler and other means, and process and store them. Batch data collection is to import a large amount of historical data into the system for subsequent analysis and processing. In order to meet the needs of different users, the data acquisition module needs to support multiple data formats, such as CSV, JSON, XML, etc. At the same time, the data acquisition module also needs to support data filtering, data conversion and other functions, in order to ensure the quality and accuracy of the data.

3.3.2. Design of the data storage module

The data storage module adopts distributed storage technology to ensure data reliability and scalability. The design of the data storage module needs to support efficient data reading and writing operation and data compression, weight removal and other functions. In order to improve the reading and writing efficiency of data, the data storage module can adopt distributed file systems such as Hadoop HDFS. At the same time, in order to improve the utilization of storage space, the data storage module also needs to support the functions of data compression and weight. In addition, the data storage module also needs to support the data backup and recovery operations to ensure the security and reliability of the data.

3.3.3. Design of the data processing module

The data processing module is responsible for processing and analyzing large-scale data to extract valuable information. The design of the data processing module needs to support various data processing algorithms and models, such as data mining, machine learning, etc. In order to improve the efficiency of data processing, the data processing modules can adopt technologies such as parallel computing and distributed computing. At the same time, in order to meet the needs of different users, the data processing module also needs to support data cleaning, data conversion, data aggregation and other operations. In addition, the data processing module also needs to support two modes of real-time processing and offline processing to meet the needs of different users.

3.3.4. Design of the data analysis module

The data analysis module provides a variety of data analysis tools and services, such as data visualization, report generation, etc. The design of the data analysis module needs to support a variety of data analysis algorithms and models, such as cluster analysis, classification analysis, regression analysis, etc. In order to improve the efficiency and accuracy of data analysis, the data analysis module can adopt technologies such as parallel computing and distributed computing. At the same time, in order to meet the needs of different users, the data analysis module also needs to support two modes of real-time analysis and offline analysis. In addition, the data analysis module also needs to support data visualization, report generation and other functions to intuitively display the data analysis results.

3.3.5. Design of the system management module

The system management module is responsible for the management and maintenance of the system, including user management, authority management, system monitoring and other functions. The design of the system management module needs to ensure the safety and stability of the system. The user management function can realize the user registration, login, password modification and other operations. The authority management function can realize the role division and authority control to ensure the security of the system. The system monitoring function can realize the real-time monitoring and alarm of the system, as well as logging and analysis functions, to ensure the stability and reliability of the system.

4. Implementation of the big data processing system based on cloud computing

4.1. Implementation of the data acquisition module

The implementation of the data acquisition module mainly includes two parts: the data acquisition interface and the data acquisition engine. The data acquisition interface is responsible for receiving data requests from external data sources and passing them to the data acquisition engine for processing. The data acquisition engine is responsible for implementing various data acquisition algorithms and models, and storing the data in the data storage module. Data acquisition engine can be implemented using a variety of technologies, such as web crawler, API interface call, etc. At the same time, in order to meet the needs of different users, the data acquisition engine also needs to support a variety of data formats, data filtering, conversion and other functions.

4.2. Implementation of the data storage module

The implementation of the data storage module mainly adopts the distributed file system such as Hadoop HDFS. Hadoop HDFS Is a highly reliable distributed file system, which can store data on multiple nodes, and ensure the reliability and security of data through redundant backup and fault-tolerance mechanisms. In the implementation of the data storage module, it is necessary to configure the relevant parameters and node information of Hadoop HDFS and create the corresponding data directory and files. At the same time, in order to realize the functions of data compression and weight removal, it is also necessary to configure the corresponding compression algorithm and weight removal strategy.

4.3. Implementation of the data processing module

The implementation of the data processing module mainly adopts the parallel computing framework such as MapReduce and the machine learning algorithm library. MapReduce Is a framework for programming models and processing large amounts of data that is able to split large-scale datasets into small ones and process them in parallel on multiple nodes. In the implementation of the data processing module, it is necessary to define the MapReduce jobs and write the corresponding Map and Reduce functions to implement the data processing algorithms and models. At the same time, in order to meet the needs of different users, the data processing module also needs to support data cleaning, data conversion, data aggregation and other operations. In addition, in order to improve the efficiency of data processing, distributed computing and other technologies can be adopted to realize parallel computing of data processing module.

4.4. Implementation of the data analysis module

The realization of the data analysis module mainly adopts the data mining algorithm library and the data visualization tools. The data mining algorithm library contains a variety of data mining algorithms and models, such as cluster analysis, classification analysis, regression analysis, etc. In the implementation of the data analysis module, it is necessary to select the appropriate data mining algorithm and models and configure the corresponding parameters and input data. At the same time, in order to realize the visual display of data, data visualization tools also need to use to draw charts and generate reports. Data visualization tools can be implemented using open-source visualization libraries or commercial visualization tools. In addition, in order to meet the needs of different users, the data analysis module also needs to support the two modes of real-time analysis and offline analysis, and generate the corresponding analysis results and reports.

5. System test

5.1. Test environment construction

In order to test the system, the corresponding test environment needs to be built. The test environment includes both hardware environment and software environment. The hardware environment includes hardware devices such as servers and storage devices, while the software environment includes software components such as the operating system, database, middleware, and big data processing framework.

In the hardware environment, high-performance servers and storage devices need to be selected to ensure that the system can handle large-scale data. At the same time, the corresponding network equipment and also need to be configured as security equipment to ensure the network security and data security of the system.

In terms of software environment, suitable operating systems and databases, as well as components such as big data processing framework and middleware. For example, you can choose Hadoop as big data processing framework, MySQL or Oracle as database, Spring framework as middleware, etc. At the same time, it is needs to configure corresponding development tools and environments, such as Java

development tools, Python development tools, etc., for system development and testing.

5.2. Functional test

Functional testing is the test of the various functional modules of the system to ensure that the system can meet the user needs. In functional testing, you need to write corresponding test cases and simulate user operations for testing.

For the data acquisition module, it is necessary to test whether it can correctly collect real-time data and batch data, and support a variety of data formats and data sources. For data storage module, it is necessary to test whether it can efficiently store large-scale data and support functions such as data compression and weight. For the data processing module, it is necessary to test whether it can correctly process and analyze large-scale data, and support a variety of data processing algorithms and models. For the data analysis module, it is necessary to test whether it can provide a variety of data analysis tools and services, and support two modes of real-time analysis and offline analysis. For the system management module, it is necessary to test whether it can realize the functions such as user management, permission management and system monitoring.

Through functional testing, the problems and defects existing in the system can be found, and repaired and improved in time.

5.3. Performance test

Performance testing is a test of the processing power and response time of the system to evaluate the performance and stability of the system. In performance testing, you need to write corresponding test scripts and simulate large-scale data input and concurrent users for testing.

For big data processing systems, performance testing is very important. Because big data processing systems need to process large-scale data, and they need to support high-concurrent user access. If the system performance is poor or unstable, it will lead to inefficient data processing and slow user access.

Through performance testing, we can understand the performance and bottleneck of the system, and take corresponding optimization measures to improve the performance and stability of the system.

5.4. Safety test

Security testing is a test of the security of a system to ensure that it can protect user data and privacy. In the security test, it is necessary to simulate malicious attacks and intrusion behavior, and check the system for security vulnerabilities and hidden dangers.

For big data processing systems, security is very important. Because the big data processing system needs to process a large number of sensitive data, such as user information, transaction records, etc. If the system has security vulnerabilities and hidden dangers, it will lead to data leakage and privacy leakage problems.

Through the security test, we can find out the security vulnerabilities and hidden dangers in the system, and repair and improve them in time.

6. Tag

This paper designs and implements an efficient and scalable big data processing system based on cloud computing technology. The system uses technologies such as distributed storage and parallel computing, which can efficiently process and analyze large-scale data, and provide a variety of data processing and analysis tools and services. At the same time, the system also has good scalability and security, and can meet the needs of different users. Through test verification, the system has good performance and stability.

However, with the continuous development of technology and the changing user needs, the system still needs to be continuously optimized and improved. For example, it can further optimize the performance of the system and improve the efficiency of data processing and analysis; expand the functions of the system and support more data processing and analysis algorithms and models; strengthen the security of the system and protect user data and privacy. I believe that in the continuous efforts and exploration, the system will become more perfect and efficient.

References

[1] Ji, Changqing, et al. "Big data processing in cloud computing environments." 2012 12th international symposium on pervasive systems, algorithms and networks. IEEE, 2012.

[2] Yao, Qin, et al. "Design and development of a medical big data processing system based on Hadoop." Journal of medical systems 39 (2015): 1-11.

[3] Wang, Lizhe, et al. "pipsCloud: High performance cloud computing for remote sensing big data management and processing." Future Generation Computer Systems 78 (2018): 353-368.

[4] Yang, Chao-Tung, et al. "Implementation of a big data accessing and processing platform for medical records in cloud." Journal of medical systems 41 (2017): 1-28.

[5] Abid, Ahlem, Farah Jemili, and Ouajdi Korbaa. "Real-time data fusion for intrusion detection in industrial control systems based on cloud computing and big data techniques." Cluster Computing 27.2 (2024): 2217-2238.