Design and Research of Big Data Collection and Analysis Platform Based on Cloud Computing

Xuan Pei*, Xiaoying Ren
Shandong University of Science and Technology, Qingdao 266590, China;
*Corresponding author Email:7996387@qq.com

Abstract. More new technologies and products emerged in the era of big data. In the era of big data, the paper proposes a big data collection and analysis platform based on cloud computing. The paper first designed the software architecture and network architecture of the big data analysis platform. Secondly, comparing the data platform with the traditional data analysis platform, the platform can quickly process high-correlation and complex data, and reduce the response time of data analysis and processing. The platform has good practical value in big data service processing.

1. Introduction
With the advent of the information age, the development of big data technology has gradually matured, and various network data has also exploded. In the face of fiercely competitive markets, how to integrate this huge amount of data has become a concern of major operators. These data include traffic information, life information, etc., all from the network or other channels. The data information is complex and lengthy. Only by establishing a management platform that integrates analysis and storage, can we make better use of these data and tap into self-interest. Helpful data to increase the market competitiveness of operators. The emergence of big data has facilitated various operators. It is another major technological change after the development of the Internet of Things, cloud computing, and IT industry. Big data has a great impact on business processes, organizations, and corporate decision-making. Some telecom operators use big data technology to classify and store massive amounts of data, and provide users with big data broadband processing by shortening the processing data path, thereby improving the speed of data analysis and providing more services to major operators. Value of business information, thereby enhancing the market competitiveness of operators.

2. Problems to be solved with big data
The data size in the era of big data is very large, it is impossible to accurately express it with numbers, and there is still correlation between the data, and the data can only be analyzed by feeling. The distinguishing feature of big data is that data has an interaction relationship, and the data is deeply mined according to the relationship between the data [1]. If there is no such relationship between the data, then simply putting the data together does not make any sense, and the data does not affect the effect of life. In addition, the data structure also has a relationship, and is a very complex relationship. The attributes of different data of the database are also classified differently. Therefore, there are subsystems and total systems, and useful information can be obtained according to the complexity characteristics of the data structure.
2.1. Characteristics of big data
(1) Has a huge data scale: the size of the data cannot be determined by specific numbers, and the relative comparison is made with a vague, approximate personal feeling; (2) The data structure has a high complexity: the complexity of the data structure can pass more and more information; (3) having higher data relevance: the degree of data excavation is determined by the degree of data relevance, and the high degree of data relevance is one of the most important characteristics of big data. Without this feature, even if the amount of data is large, the structure of the data is complicated, and it cannot be called big data [2].

2.2. Problems to be solved by using big data
(1) Preprocessing is done by scripting language. Because the language is too simple, it can't parse too complicated data structures; (2) Relational database can't meet the requirements of big data well; (3) How to broaden its Commercial database optimization space; (4) How to control the quality of the obtained data; (5) The phenomenon that the work needs to depend on the computing power of the data is increasing.

3. Big Data Analysis Platform Design Strategy

3.1. Platform design
The big data platform based on cloud computing technology can provide users with highly scalable and cost-effective hardware supporting PB-level systems, and even semi-structured, massively structured, unstructured ZB-level data storage. Moreover, the establishment of a unified big data analysis platform can speed up the data mining speed and mine the data support, bringing greater economic value to users. (1) Build a unified data computing platform in the enterprise; (2) Enterprise owners can directly control their data instances; (3) Provide enterprise-level data access functions directly through entity integration [3]; (4) Flexible expansion and configuration reduces the average risk of investment.

In the era of cloud technology, a big data platform with a high cost performance and high scalability hardware system can store massive amounts of PB and ZB data, which can be structured or semi-structured. Or even unstructured. At the same time, we want to make the enterprise obtain the ideal profit, create big value with big data, and must realize the high-speed mining of data value [4].

This paper proposes a platform for unified analysis of big data through cloud computing, and builds an efficient platform for analyzing big data through MapReduce structure and database storage, thereby realizing the organization of data, semi-structured data and unstructured data. Analysis, based
on this platform, customers with big data can convert costs into profits and use data to drive business growth.

3.2. Unified analysis of the software structure of the platform of big data

(1) Software structure. The Segment host has multiple nodes that connect the Segment host, Master host, and related databases over the Internet. In the network, each storage node has no information exchange, and is independent of each other. In order to realize the exchange of information between them, it must pass the master host, and each application program realizes data access through the master host [5]. The task of each node's segment server is the same, and they are connected through the Internet, thus forming a server system. The nodes of the server system are independent of each other and share-nothing. They can only access their own local resources. It is because of this feature that the server has excellent expansion performance. In theory, there is no limit to the expansion, but as far as the current technology is concerned, there are thousands of connected CPUs and 512 nodes. Each node exchanges data through the Internet connection, and each node cannot access the memory of other nodes. This process is named data redistribution.

(2) Designed solutions with high availability. The master host of the solution adopts a master-slave mode. The GE network is used to connect the segment host to the master host. The master-slave scheme ensures that each segment host has two data: the primary network segment data and the backup network. Segment data makes the entire server system very highly available.
(3) Existing shared structure. A system with only one server is usually a "full share" mode, which is typically a relatively expensive SMP server. The Disk Sharing architecture typically has multiple servers and other shared storage devices that are connected to the server. There is usually a narrow data pipeline within the system that filters the information of the I/O device and stores it in the shared disk subsystem [6]. Whether it is "full sharing" or "disk sharing," their performance and scalability are limited in terms of their structure. In particular, disk sharing is difficult to handle such huge data as terabytes, and the system has the disadvantages of being fragile and complicated.

(4) Structures that are not shared at all. In the share-nothing architecture, a common database, host SAN/shared disk, and disk SAN/FC network are designed for online processing functions, which is used for smaller data queries. In the share-nothing system, the items that the user wants to query can be divided into multiple parts and run in parallel in the entire cluster. All the communication data to be used is completed on an interconnected high-bandwidth system. The structure of this system is not only simplified, but also has a separate high-speed channel between each node and the local disk, which can well complete the parallel processing problem [7].

Fig. 4 Different network architecture

3.3. Big Data Unified Platform Network Architecture

(1) Big data platform architecture sharing solution. The big data platform that realizes "complete sharing" is mainly limited by a single server. Currently, the server that can meet the needs of data sharing is SMP. This kind of server is expensive, and most enterprises will not consider this because of cost. In order to realize the information sharing function of the big data platform, enterprises generally establish a "disk sharing" system, which constitutes a system of multiple servers, and connects these servers and the SAN to realize the function of storing and sharing data. The system requires a narrow data pipeline to filter all I/O information and then store it on a shared disk. From a structural point of view, the "disk sharing" and "full sharing" systems are insufficient in performance and scalability, and the common shared disk system is relatively fragile and complicated, and it is impossible to accurately and timely process tens of thousands of bytes of data.

(2) Characteristics of the unified platform scheme for big data.

1) Node mirroring - data protection. In the big data platform, the master is responsible for storing the system metadata, and the host segment node stores the user-related data. Under the mirror image, the mirror data can be stored on different segment hosts. By analyzing this image configuration, it is concluded that if the Segment host is down, it can also find relevant data in other Segment hosts and store it in the Segment host database.

2) Load external table high speed data. This part has the advantages of: data flow parallel engine technology, using SQL directly to the external table operation and full parallel loading, the speed can reach 4.5TB / hour.

3) SQL&MapReduce integration environment. Compared with the traditional RDBMS system, the big data programming environment is a system environment that is a combination of SQL statements.
and MapReduce. The analyzed cloud computing platform is a hardware system built with X86 open architecture server PC. It has the functions of large-scale data computing and distributed storage of data. It can solve the problems of I/O, has high security, and expands performance comparison. Well, various data resources can be developed as needed. As shown in Figure 5.

![SQL & MapReduce environment](image-url)

**Fig. 5 SQL & MapReduce environment**

### 3.4. Platform Advantage

Cloud computing is the processing and settlement of scalable, dynamic, and virtualized data in the Internet environment. It also provides users with data interaction and storage functions. Big data is the data with more information, mainly the general term for all kinds of data in this information network era. By mining data from different structures, types and frequencies, it brings more economic value information to users, big data. It is a product that emerged in the development of the Internet era. Data mining technology is to find valuable information from a large amount of data, analyze the information implicit in the data, and provide users with decision-making basis to serve various fields. Data mining technology has the function of describing tasks and predicting tasks. The description task is to summarize the hidden relationships between some data. The forecasting tasks are based on the attributes of the data and the value of the target data. Big data has the advantages of massive information, high value of use, and fast information processing. Cloud computing is a big data running platform that can provide better storage space for big data.

### 4. Conclusion

At present, China's network information technology is developing very fast, big data technology has also been rapidly developed, and related new technologies and methods are constantly emerging, and many related products of big data are also increasing. In this environment, the article designed a unified analysis platform for big data based on cloud computing. This platform can handle data with high data relevance and complex data structure. The platform designed in this way also supports PB
level data. The purpose of reducing the response time of data analysis can be well achieved. This platform has good practical value in the development of big data future business and technology.

References
[1] Wang Tianyu, Sun Yanqiu, Yan. Application of Cloud Computing in Regional Medical Informatization in the Age of Big Data. China Medical Devices, Vol.6 (2015) No.30, p.72-74.
[2] Zhang Xingang, Yu Bo, Cheng Xindang, et al. Research on Collaborative Protection of Personal Information Security in Big Data and Cloud Computing Environments. Computer Knowledge and Technology: Academic Exchange, Vol.1 (2016) No.12, p.53-55.
[3] Jiang Suping. Research on the Impact of Big Data and Cloud Computing Technology on Auditing. Finance and Accounting, Vol.16 (2017) No.17, p. 160-165.
[4] Huang Lei. Data Security in Big Data Cloud Computing Environment. Electronic Production.Vol.20 (2017) No.18, p.57-58.
[5] Li Hongyan. Discussion on Data Security in Big Data Cloud Computing Environment. Information and Computer (Theoretical Edition), Vol.3 (2017) No.11, p.201-202.
[6] Meng Cunxi, MengCunxi. Application Requirements of Big Data and Cloud Computing in Rail Transit Engineering. Civil Engineering and Information Technology, Vol.5 (2015) No.28, p. 103-105.
[7] Liu Congbing. Accounting Information Construction Based on Big Data and Cloud Computing. Science and Technology Innovation and Application, Vol.3 (2015) No.17, p.173-174.