ISSN: 2664-9640

Overview of big data storage methods

Zhongkai Zhang, Tong Zhen, Zhihui Li

College of Information Science and Engineering, Henan University of Technology, Zhengzhou 450001, China

*z18739086609@163.com

Abstract

With the rapid development of modern cloud computing, Internet of things and various new social information networks, the era of big data is sweeping, and a large number of unstructured data storage is increasing exponentially, and the form of data is highly complex. At present, the huge amount of data and the complexity of data storage format bring great challenges to the development of our traditional data processing and analysis storage technology. The first difficult problem to be solved is the storage of big data. Because of the particularity of big data different from traditional data, if we use our traditional storage method, we can't realize the unstructured storage of big data. Therefore, we need to further innovate and improve the traditional storage method of big data. This paper mainly introduces the improved data storage method of traditional relational database for big data, as well as non relational database storage and cloud storage.

Keywords

Big data; relational database; non relational database; cloud storage.

1. Concepts and characteristics

1.1. The concept of big data

So far, big data still has no exact and unified concept and definition[1,2], it is a relatively abstract concept. Baidu Encyclopedia basically interprets the concept of big data as: big data refers to a collection of massive data information that cannot be accurately captured, managed and comprehensively processed by traditional or conventional software processing tools within a certain time or range. It is a massive, high-yield growth rate and diversified information comprehensive asset that requires the support of new data processing technology models to achieve stronger analytical decision-making power, insight and discovery, and comprehensive process optimization capabilities.

However, this is not a precise definition, because the scope of commonly used software tools cannot be determined, and the time range is also a rough description. It is generally believed that a more precise definition of big data should be to combine it with our understanding of the four characteristics[3,4] of big data. The four characteristics include volume, velocity, variety and value.

We should also distinguish the difference between big data and massive data: big data is not equal to massive data, massive data only emphasizes the large amount of data, but big data is not only used to clearly describe the large amount of data, but also further clearly points out the complexity of the data The structure, the fast time transfer characteristics of data and the analysis and processing of the characteristics of massive data can finally obtain valuable information.

ISSN: 2664-9640

1.2. The "4V" Characteristics of Big Data

Volume: Scale means that the amount of data in an enterprise is huge. This is also one of the basic attributes of the enterprise processing big data. Although the number and size are not specified, it must be TB or higher, otherwise the personal computer Data processing is completely relaxed, and there is no great academic research and application value.

Varity: The variety of data refers to the variety of data, which may take into account multiple types of unstructured data such as tables, logs, pictures, etc., and may even include various formats such as audio streams. And videos. Data diversity also refers to the complexity and change of big data. Big data itself is a kind of unstructured data, and there is no clear organizational structure and rules. Even some types of local data may obviously have some organizational rules. However, there is still no unified organization rule in general, which is also an important difference between diverse big data and other traditional unstructured data.

Velocity: The so-called high speed refers to the efficiency and speed of obtaining data, and it also requires us to process data quickly, which is also one of the characteristics that are different from traditional mass data. Dynamic data such as passenger booking or flight data at large international airports, online stock trading records of stockholders of listed securities companies, online shopping records of users of large supermarkets... etc. This massive amount of dynamic data is not only a large amount of processed data, but also continuously collects and produces a large amount of new dynamic data. So in other words, when we analyze big data, we must fully consider the changes of dynamic data and other factors, and the efficiency and speed of processing dynamic data must be rapid.

Value: It generally means that the density of data value is low, and it is also one of the important value attributes of unstructured data. The analysis of big data is carried out step by step. It is possible to get some of the most primitive laws first, and then find higher-level laws from these primitive laws... After many steps, we will get valuable information. In order to ensure that there is enough effective information for newly generated applications, it is usually necessary to save all the data at the same time, so that the number of applications of unstructured data will surely increase sharply, and it will definitely make the amount of useful information in unstructured data. The proportion of data decreases, so in the end it will definitely lead to a low density of value in unstructured data.

2. Big data storage method

Because big data is different from traditional mass data in nature, it is impossible to store big data using traditional data storage methods. Because the application and design patterns of traditional relational databases have been restricted, no matter how large or small the amount of stored data is, only a single-machine data storage method needs to be considered, that is, users can use one machine to store all data at the same time. However, because the data storage equipment that the database on each machine can carry at the same time is limited (generally it will not exceed a few terabytes), so after the data volume has soared to a certain extent, the real-time data retrieval and storage speed will be reduced. There may be a sharp drop. For the storage of big data, the following will introduce and compare the changes in relational databases to deal with big data storage, to the rise of non-relational databases, and the application of cloud storage.

2.1. Relational database

Relational data MySQL[5,6] has proposed the MySQL proxy component to deal with the storage problem of big data, as shown in Figure 1.

ISSN: 2664-9640

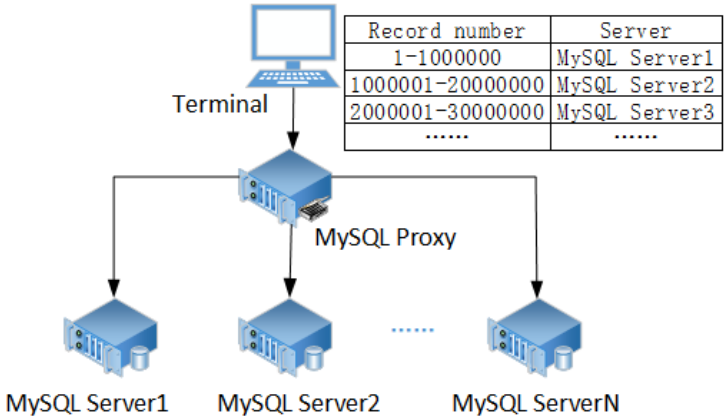

Figure 1 MySQL proxy component

The MySQL proxy component generally exists in the form of an intermediate proxy layer. There are multiple MySQL servers behind the MySQL proxy component. In other words, MySQL proxy is a data connection pool. Its main function and role is to connect the foreground and the client. Application data and requests are directly forwarded to the back-end database, which can directly realize complex data connection load control and data filtering, thereby also realizing read-write load separation and data load balancing. It uses the distributed principle of storage: it intercepts the front-end request data, and then combines distributed storage technology to split all the recorded data in a large data table, and then split it into different queries Query on the node. For each query node, the amount of data is not very large, which also improves the efficiency of the query.

2.2. Non-relational database

Although relational database solves the problem of big data storage to a certain extent, it is only a symptom and not the root cause, but it does not solve this problem fundamentally. In order to effectively deal with the various challenges brought by the large-scale data collection and its application of multiple data types, especially to solve the problems of big data storage applications, non-relational databases (Nosql) were created[7,8]. Moreover, Nosql database has the advantages of strong scalability, high concurrency and flexible data model. In addition, it started late and responds to big data much faster than relational databases.

2.2.1. MongoDB

MongoDB is a non-relational database based on distributed file storage[9], which is very similar to relational databases, and its file storage function is also the most among non-relational databases. MongoDB[10,11] is actually a data type that can be used to store more complex data, mainly because of its good support for loose data structures. High speed is the biggest advantage of MongoDB. It stores hot data in physical memory, so the read and write of hot data is very fast, so its performance is very fast in a moderate amount of memory. Mongo query language is also very powerful, most of the query functions in the single-table query of relational database can be easily implemented, and it can also support data indexing. In addition, it is very scalable.

The features of MongoDB's storage and data query solutions [12] for big data are: first, automatic sharding of nodes during automatic data storage and automatic expansion of node data levels, which can automatically store real-time massive node data; in real-time data indexing In terms of query, MongoDB also supports secondary data indexes, TTL indexes, geospatial indexes, etc., to meet the needs of a variety of actual business and scenario data queries; in an independent sharded data cluster, all shards have corresponding The copy, even if the main shard is damaged, will not affect business use, and has high reliability.

Several business scenarios for MongoDB:

ISSN: 2664-9640

(1) Real-time website data: Since MongoDB is particularly suitable for real-time data update, insertion and data query, it is very suitable for the high scalability of data required for real-time data management and storage for website management.

(2) Caching: Due to the high performance of MongoDB with a moderate amount of memory, it is also a cache management layer suitable for all information infrastructures.

Large-size, low-value data: By using MongoDB to store large-size, low-value data, it can effectively reduce the cost of storing data in traditional relational databases.

(3) For the storage of JSON format data: MongoDB documents are stored in a BSON (binary json) file format. The file format is json, so it is very suitable for storing json file data.

2.2.2. HBase

HBase[13] is also a non-relational database for distributed file storage, which is basically completely open source. Its underlying data storage mode is based on distributed Hadoop[14]. HBase mainly relies on the horizontal expansion of data and the continuous increase of cheap commercial storage servers to enhance and expand the data storage capabilities; it can read and write large amounts of data in real time, A single table can achieve the level of 1 billion-level data volume of millions of columns. The reason why HBase is different from the general distributed database is that it is more suitable for unstructured file data storage; the other is that it is suitable for column-based storage rather than row-based storage.

Comparison of row storage and column storage:

(1) The storage of rows stored on disk is continuous; the storage of columns stored on disk is not continuous;

(2) Compared with the write performance, the fewer write times, the higher the performance. Because for every write to the disk, head scheduling must occur, resulting in seek time. Because row storage is written only once and column storage has to be written multiple times, row storage has more advantages in write performance

(3) Comparison of read performance: a. If the entire table is read, the row storage performance is higher; b. If the specified column is read, the row storage will generate redundant columns, and redundancy The elimination of the columns occurs in memory. The column store will not have redundant columns.

(4) When storing data, if it is based on row storage [12], since the field types of a row of data may be different, frequent data type conversions will occur; but based on column storage, since the types of data in the same column are generally the same, it can be avoided Frequent data type conversions, and some better compression algorithms can be considered to compress a column of data.

HBase stores data in the form of a data list[15]. The table mainly consists of two parts: rows and columns. The structure of HBase's data table is as follows Figure 2:

rowkey	Column cluster:baseinfo	Column cluster:courseinfo
Rk001	Name:zhangsan	Math:93
Rk002	Name :liling Age :30	English :95
Rk003	Name :wanghaitao Age :25 Telephone:13569237659 TS1 Telephone:18965412531 TS2 Telephone:17726893655 TS3	Algorithm :80

Figure 2 The structure of HBase's data table

ISSN: 2664-9640

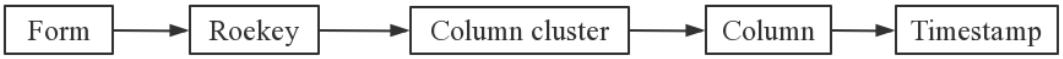

Figure 3 Data checking process

HBase table structure:

Rowkey:dictionary sort

Column cluster:Contains a set of columns, which are specified when inserting data, and column clusters are specified when creating a table.

Column:There can be multiple columns in a column cluster, and they can be different.

Timestamp:The value of each column can store multiple versions of the value, the version number is the timestamp, sorted according to the timestamp from nearest to farthest.

The main characteristics of tables in HBase are:

(1) Large: A single table in HBase can reach hundreds of millions of rows and millions of columns

(2) Multi-column-oriented: HBase uses column-oriented (family) storage and access control through retrieval, and column (family) searches independently.

(3) Sparse: As for the null column, there is no need to occupy any stored data space in the table, so the structure design of the list can also be very sparse.

2.3. Cloud storage

The concept of cloud storage[17,18] is a new technical concept developed from the basic concept of cloud computing. It is a network data storage technology emerging in a new era.

Cloud storage and cloud database comparison:

From the corresponding level:

(1) Cloud storage: It is at the resource layer, the iaas layer of the cloud, which provides storage resource capabilities.

(2) Cloud database: It is at the platform layer, the paas layer of the cloud, which provides middleware service capabilities.

The migration of local databases to the cloud corresponds to cloud databases, while the migration of local hard drives to the cloud can only correspond to cloud storage.

From the service provided:

(1) Cloud storage: Provides powerful storage data management capabilities, facing the scenes are mostly unstructured various types of data, such as files, pictures, videos, etc.

(2) Cloud database: Provides basic database and data object management capabilities, including relational databases such as oracle, Mysql, and sql server, as well as semi-structured databases such as MongoDB and HBase.

3. From the relationship between the two:

Currently, cloud storage is basically packaged based on a distributed file system similar to hdfs, providing storage service capability interfaces. You can also build a layer based on hdfs to form a database, and then expose the database capabilities to form a cloud database, similar to HBase. However, common relational databases can be used as cloud databases, but their underlying cloud storage capabilities are not dependent on them.

2.3.1. Big data storage system architecture

1. Direct Attached Storage (DAS): The storage device is directly connected to the host system

Applicable network environment:

ISSN: 2664-9640

(1) The network geographic center location of the system network management server and its network spatial distribution are very scattered, and it is difficult to interconnect through the SAN or the network between NAS;

(2) The storage system must be directly connected to the application server;

(3) Small network.

Disadvantages: low resource utilization, poor scalability and manageability, and serious isomerization.

2. Network attached storage (NAS): A mode in which special equipment is directly connected to network media to realize data storage. The physical storage device of the NAS must have a dedicated server and a dedicated operating system.

Advantages: (1) Plug and play;

(2) A dedicated operating system can support file sharing between different operating systems of application servers;

(3) The use of an optimized file system on the dedicated file server greatly improves the efficiency of real-time access and management of files;

(4) The server is independent of the application server, even if the application server is damaged, the relevant data can still be read.

Disadvantages: (1) Due to the shared wireless network mode, the bandwidth of the network is another bottleneck to improve the performance of the storage system;

(2) NAS access is subject to file system format conversion, so it can only be accessed at the file level, which is not suitable for block-level applications.

3. Storage area network (SAN): Storage area network storage is a kind of storage based on the storage network connection, networked regional information, connecting multiple storage devices and a storage network server group to each other, thereby forming an independent storage network . It is composed of three parts: storage device, interface and connection corresponding communication control protocol connected to the network.

SAN's main business functions: file management data retrieval and information archiving, recovery and data backup, mobile file management data sharing and migration between storage devices, disk mirroring file management software technology, and file data sharing between mobile wireless network management devices and storage servers, etc. .

According to ISCSI protocol, SAN is divided into FC SAN and IP SAN for easy distinction.

The compatibility of FC SAN is poor and its cost is relatively high.

IP SAN has high scalability, and the verified transmission equipment can guarantee the reliability of operation; and its data concentration can realize remote data replication and damage recovery; the overall cost is not very high.

2.3.2. Advantages of big data cloud storage

1. Cost saving: The cost of software and hardware maintenance is very high, and cloud storage can reduce costs;

2. Data security: When local data is destroyed, cloud data will not be lost; and in cloud storage, if the hard disk is broken, the data can also be migrated to other hard disks;

3. High comprehensive utilization rate of space: Virtualization technology can effectively avoid the serious waste of storage space and improve the comprehensive utilization efficiency of storage space.

3. Summary

With the rapid development of cloud computing, the Internet of Things, and various social mobile networks, the era of big data has come. Big data is different from massive data, so the

storage of big data will inevitably be a difficult point. From relational databases to non-relational databases, when it comes to cloud storage, each has its own characteristics and adaptations. Relational databases and non-relational databases and cloud storage have their own strengths. If they can be effectively combined in the data storage process, instead of being clearly distinguished, the storage efficiency of big data will be greatly improved.For the processing of big data storage, we should make appropriate choices based on needs, instead of blindly pursuing the largest optimization and causing counterproductive waste of resources and cost.

Acknowledgements

The authors acknowledge the National key research and development project (No: 2018YFD0401404).Doctor Fund of Henan University of Technology (2017BS034). Project of Key Laboratory of Grain Information Processing and Control(Henan University of Technology), Ministry of Education.

References:

[1] Jia Yunxiang. Overview of Big Data Research[J]. Economist, 2018,(12):260-261. DOI:10.3969/j.issn.1004-4914.2018.12.134.

[2] Zheng Qiang, Gao Qun. Overview of Big Data Research [J]. Science and Technology Vision, 2018, (30): 179-180. DOI: 10.19694/j.cnki.issn2095-2457.2018.30.078.

[3] Yan Xiaofeng, Zhang Dexin. Research on Big Data[J]. Computer Technology and Development, 2013, (4): 168-172. DOI:10.3969/j.issn.1673-629X.2013.04.041.

[4] Tu Xinli, Liu Bo, Lin Weiwei. Overview of Big Data Research[J]. Computer Application Research, 2014, 31(6): 1612-1616, 1623. DOI:10.3969/j.issn.1001-3695.2014.06.003.

[5] Yan Qing, Miao Zhuang, Lai Xinsheng, et al. Innovation and development of relational database MySQL in the era of big data [J]. Science and Technology Wind, 2020, (20): 75-76. DOI: 10.19392/j.cnki.1671- 7341.202020063.

[6] Yang Guoqing. Research on matrix algorithm of tree structure in relational database[J]. Computer Times, 2020, (3): 50-52, 56. DOI: 10.16644/j.cnki.cn33-1094/tp.2020.03 .014.

[7] Liu Yucheng, Li Gang. Comparison of NoSQL database and relational database[J]. China New Communications, 2018, 20(7): 81. DOI:10.3969/j.issn.1673-4866.2018.07.067.

[8] Zhao Wenshuo. Application research of relational and non-relational databases [D]. North China Electric Power University; North China Electric Power University (Beijing), 2016. DOI:10.7666/d.Y3114918.

[9] Wang Liwei. Application analysis of big data distributed storage technology[J]. Construction Engineering Technology and Design, 2019,(33):194. DOI:10.12159/j.issn.2095-6630.2019.33.0181.

[10] Li Jiwei, Duan Zhongshuai, Wang Shunye. Data storage of unstructured database MongoDB[J]. Computer Knowledge and Technology, 2018, 14(27): 7-9.

[11] Qiu Zehuan. Research on MongoDB database file data storage [J]. Computer Fan, 2017, (25): 31. DOI:10.3969/ j.issn.1672-528X.2017.25.029.

[12] Qi Lan. Research on data storage and query optimization technology based on MongoDB [D]. Jiangsu: Nanjing University of Posts and Telecommunications, 2016.

[13] Wang Jiao. Research and development of big data storage system based on Hbase[D]. Shaanxi: Xi'an University of Technology, 2017. DOI:10.7666/d.D01277252.

[14] Sun Zhaoxu. Research on HBase-based massive data processing under Hadoop platform[D]. Guangxi: Guilin University of Technology, 2014. DOI:10.7666/d.D553498.

[15] Xu Yao. Research on HBase secondary index based on hash [D]. Shandong: Shandong University of Science and Technology, 2018.

[16] Zhang Huanghui. Discuss the development of cloud storage technology in the era of big data [J]. Information Recording Materials, 2020, 21(1): 143-144.

[17] Zhang Xin. Analysis of the development of cloud storage technology in the era of big data[J]. China Security, 2018, (10): 80-83. DOI:10.3969/j.issn.1673-7873.2018.10.018.

[18] Ren Mingfei, Li Xuejun, Cui Mengmeng, et al. Design and development of non-relational database based on MongoDB[J]. Computer Knowledge and Technology, 2019, 15(34): 1-2.

[19] Yin Yan, Zhu Liwei. Talking about the data storage of NoSQL database[J]. Science and Information Technology, 2019, (6): 61, 64.

[20] Wang Hongbo. Exploring relational databases [J]. China Science and Technology Investment, 2019, (6): 276-277. DOI:10.3969/j.issn.1673-5811.2019.06.241.

[21] Di Huiping, Wang Jingning. Talking about non-relational databases[J]. Hebei Agricultural Machinery, 2019, (11): 50.

[22] Zhou Zhenxiong, Gong Haopeng. Research on remote sensing data storage technology based on typical distributed database MongoDB[J]. Dossier, 2019, 9(32):322.

[23] Lang Yunhai. Improvement of NoSQL database security strategy under big data[J]. Communication World, 2019, 26(8): 29-30. DOI:10.3969/j.issn.1006-4222.2019.08.018.

[24] Xia Shujian. Research on Big Data Query Technology of NoSQL Database [C].//Institute of Educational Science, Chinese Academy of Management Science. Proceedings of the 2019 Educational Development Research Planning and Scientific Research Achievement Exchange Conference. 2019:325-325.

[25] Huang Pei. Research on file-type big data storage technology based on non-relational database[J]. Computer Knowledge and Technology, 2019, 15(23): 3-4.

[26] Hu Bo. Research on the unified storage and access interface of non-relational databases [D]. Sichuan: Sichuan Normal University, 2016.

[27] Yang Yeling. Research on data security technology in cloud storage under the background of big data[J]. Science and Information Technology, 2018, (29): 51, 54.

[28] Zhang Huanghui. Discuss the development of cloud storage technology in the era of big data [J]. Information Recording Materials, 2020, 21(1): 143-144.

[29] Fu Xuelei. A preliminary study on cloud storage technology and its application in big data scenarios[J]. Journal of Jiangxi Vocational and Technical College of Electric Power, 2019, 32(8): 22-23, 25.

[30] Yin Lina, Qiu Liyuan. Cloud Storage Technology and Application in Big Data Scenarios[J]. Information and Computer, 2019, (14): 19-20.

[31] Yu Wenyu. Analysis of data storage and processing technology under the background of big data era[J]. Digital User, 2019, 25(42): 80.

[32] Yuan Xiaodong. HBase database schema design guidelines[J]. Microcomputer Applications, 2018,34(10):74-77. DOI:10.3969/j.issn.1007-757X.2018.10.024.

[33] Wang Weichen. Research on retrieval based on non-relational database HBase storage technology [J]. Internet of Things Technology, 2020, 10(1): 103-105. DOI: 10.16667/j.issn.2095-1302. 2020.01.029.