Research on Big Data Volume Storage Structure Design and Quick Query Mechanism of Oracle Database

Dingfa Liu
Jiangxi Ahead Software Vocational and Technical College, Nanchang, Jiangxi, 330041, China

Keywords: Big data; data segmentation; indexing; query optimization; distributed

Abstract. Based on a mobile resource management platform, this paper provides a solution for the effective storage and querying of the massive structured data and unstructured data was generated by research and analysis. In the aspect of structured data, this paper first divides the database vertically according to the characteristics of the business, stores the data of the write operation more frequently and stores the data more frequently in the different database, and improves the reading and writing performance of the data. For the basic information of equipment and statistical information related to the existence of large concurrent read operation, the data has been cut again, the two kinds of information on different libraries, and further improve the system to deal with high concurrent read query capabilities.

1. Introduction

The current society has entered a data explosion era, massive data set processing and analysis is called "big data" [1, 2]. Large data is a new research focus in the field of information technology after cloud computing. Ubiquitous sensors, microprocessors and the Internet, forming a huge source of data [3]. Scientific research field, meteorological data, geographic data, biological information data is the traditional mass data set, manufacturing industry, many machines are installed on one or more microprocessors to collect production data; commercial consumption, online shopping transactions , Consumer evaluation and so on data have become a big data problem; national governments of massive statistical data and documents because of the development of computer technology has become urgent need to deal with large data problems. [4, 5] As the traditional relational database in the management of large data encountered difficulties and obstacles, the existing database products and database business model can not meet the large data storage scale, at the same time, not only need mass storage system to store large data, and need New large-scale distributed database management system to manage large data.

The diversity of data types for large data includes structured data and unstructured data [6]. In the past, structured data was mainly stored in text form, and more and more unstructured data generation, new data storage and processing put forward new challenges. In order to meet a large number of users at the same time to submit real-time system requests and high load data query, while the system needs to provide high-speed query response time. This requires that the underlying data storage structure to meet the growing data at the same time can efficiently handle the query request. This is the most significant feature of distinguishing traditional data mining [7, 8]. At the same time, the efficiency of dealing with massive data is the pillar of life.

This paper presents a multi-database parallel algorithm, the use of multi-threaded features, so that the task can be parallel processing, and further improve the speed of statistical queries. Finally, this paper proposes a multi-table paging algorithm with high data volume under the research of multi-table paging process, which effectively solves the problem that the multi-table paging query is slow in a large amount of data. Based on a mobile resource management platform, this paper provides a solution for the effective storage and querying of the massive structured data and unstructured data generated by the research and analysis. In the aspect of structured data, this paper first divides the database vertically according to the characteristics of the business, stores the data of the write operation more frequently and stores the data more frequently in the different database, and improves the reading and writing performance of the data. For the basic information of equipment and

statistical information related to the existence of large concurrent read operation, the data has been cut again, the two kinds of information on different libraries, and further improve the system to deal with high concurrent read query capabilities.

2. Data segmentation technology

2.1 Vertical segmentation diagram

Data segmentation refers to the distribution of data according to a certain cut rules within the specified range, so that when the data query system parallel processing capacity, which reduces the response time of the query to improve the query performance of the database. Data mining can take full advantage of the database system CPU resources and network resources. As the data segmentation smaller, and distributed in multiple databases, when the data query can reduce communication overhead, balance the system load and reduce the amount of calculation, thereby improving system performance.

Data segmentation can be divided into two kinds of segmentation patterns, vertical segmentation and horizontal segmentation according to the different types of rules. Vertical segmentation is a different data table according to the characteristics of the business according to a certain cut rules cut into different databases. Horizontal segmentation is based on the logical relationship between the data in the table, according to a certain cut rules will be the same table of data split into multiple databases. Vertical cut the biggest feature is the cut rules are relatively simple, the implementation is also more convenient. Vertical segmentation for each module that is relatively low degree of mutual impact is relatively small, relatively simple business logic system. This system makes it easy to split tables used by different modules into different databases. According to the different table on the data segmentation, for the application of less impact, cut the rules are relatively simple. Vertical segmentation diagram was shown in Figure 1:

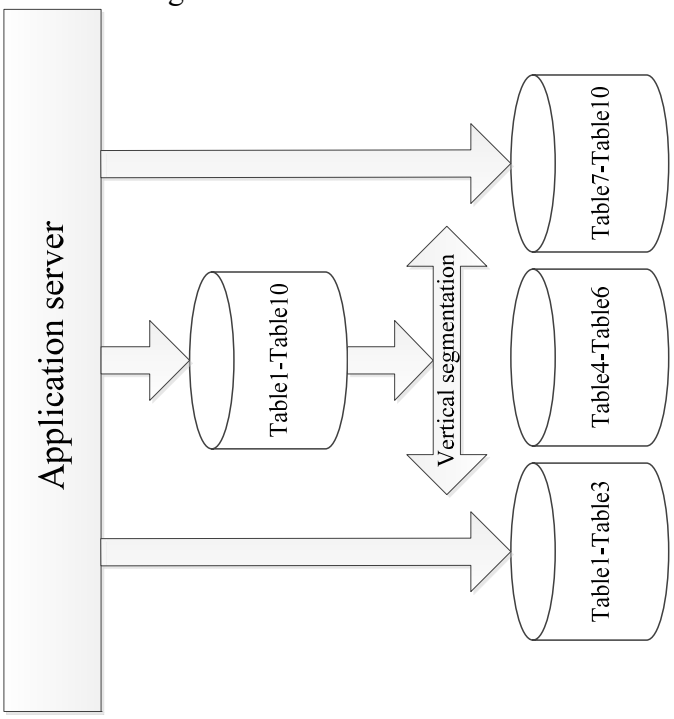

Figure 1. Vertical segmentation diagram

2.2 The impact of big data and its value

With the rapid development of computer and Internet in space and time, in the low-cost high-capacity storage systems and advanced cache technology, driven by massive data rapid search and processing technology is increasingly mature (Figure 2). The large data storage, processing and access without the space and time constraints, making the value of large data is fully realized, mainly include:

(1) A large number of valuable data distributed over the long tail can be easily accessed. The use of simple and efficient algorithm for large data mining, find new needs and technology.

(2) Large data in space and time changes to subvert the traditional 80% -20% of the rules. Traditional data access mode, 20% of the goods to create 80% of the value; and in large data access mode, the value of each commodity creation is almost equal.

(3) Big data created a world of computationally rich resources. Compared to the era of lack of computing resources, the cost of production, storage, and processing of massive data for sustained growth has fallen dramatically, and people have to change their way of thinking.

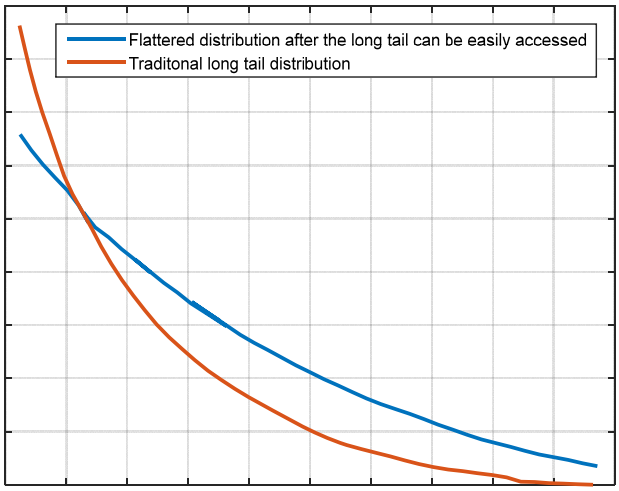

Figure 2. Compare the data access frequency of power-law distribution between traditional data access and big data access patterns

2.3 Advantages and Disadvantages of Line Storage Structure

The row storage structure is a traditional relational database storage structure, and the records are stored in the database relation table in the form of rows. When you add a row, all the columns in each record need to be stored and the records are stored consecutively in the page block of the disk. In the distributed system storage, the table is divided horizontally, and all data in each row is stored in the same HDFS block. Figure 3 shows the distribution of the data structures stored in rows in the HDFS block.

A	B	C	D
101	201	301	401
102	202	302	402
103	203	303	403
104	204	304	404
105	205	305	405

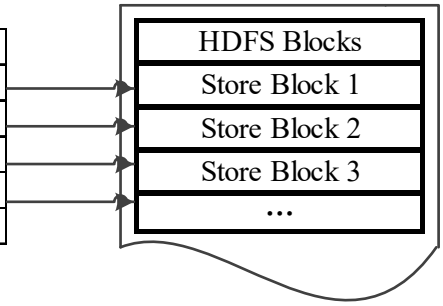

Figure 3. The layout of row store structure among HDFS blocks

When stored in a row structure, all the columns in each row are stored in the same HDFS block. In the distributed file system HDFS, the data in the large table is divided horizontally, and each group of data may be distributed on different datanode nodes.

To estimate the time complexity of accessing large data, read the data consumption time T (r), write data consumption time T (w). p(r) + p (w) = 1, respectively. Set to access large data need to consume time for the E (T), then:

$$E(T) = p(r) \times T(r) + p(w) \times T(w) \tag{1}$$

In the specified data storage mode, by the formula (1) know, access to large data required for the time:

$$E(OP \mid DPS) = \omega_w E(write \mid DPS) + \omega_r E(read \mid DPS) \tag{2}$$

In the data table containing n columns, the combination of all the columns in the lookup table is:

$$C_n^1 + C_n^2 + ... + C_n^n = 2^n - 1 \tag{3}$$

The time taken to process the read operation is:

$$E(\text{read} \mid \text{DPS}) = \sum_{i=1}^{n} \sum_{j=1}^{C_n^1} f(\text{i, j, n})(\frac{S}{B_{local}} \times \frac{1}{\rho} \times \alpha(\text{DPS}) + \lambda(\text{DPS, i, j, n}) \times \frac{S}{B_{networl}} \times \frac{i}{n}) \tag{4}$$

3. Design and Implementation of Data Segmentation Scheme

3.1 Overall structure

Through the discussion, this paper first according to the characteristics of the data on the data were cut off vertically, the data read and write were separated. Then the data of the user's behavior information is large and the situation is increasing, the data is divided into the area by the database, and then the data table is sorted monthly by the database in each region. By increasing the size of the data to reduce the size, thus improving the query performance of massive data. The overall structure of structured data is shown in Figure 4:

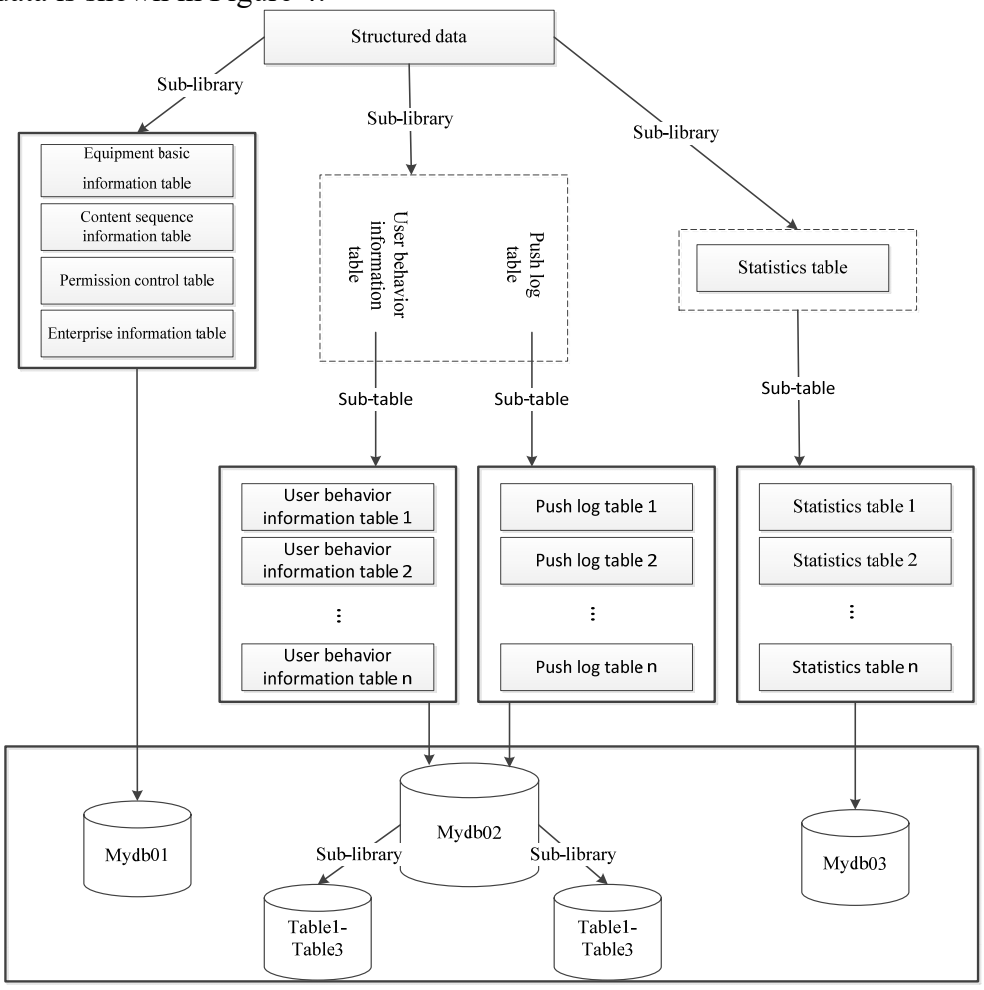

Figure 4. Structured data storage structure

3.2 Experimental testing and analysis

The algorithm for testing the hardware environment for the Dell pc (memory 8QCPU model Intel3.40GHz, 8 nuclear, hard disk 1TB) - Taiwan, the software environment for win7, 64, Oracle11g.

The user behavior information table of the application system is tested as an example. The user behavior information data in the system is divided into monthly basis. Assume that the maximum number of records per month is 25 records. Respectively, the number of test points for the 2 and 3 sub-table data volume of 1 million, 10 million, 50 million times, query a user's behavior information Union multi-table query method and optimized multi-table query algorithm time, test results data are shown in Tables 1 and 2 below.

Table 1 Comparison of efficiency before and after optimization (2 tables)

Number of subscales	Sub-table data volume (millions)	Union method (seconds)	Optimization algorithm (seconds)
2	100	1.655	1.705
2	1000	16.789	9.874
2	5000	113.443	45.675

Table 2 Comparison of efficiency before and after optimization (3 tables)

Number of subscales	Sub-table data volume (millions)	Union method (seconds)	Optimization algorithm (seconds)
3	100	4.652	1.605
3	1000	40.759	12.574
3	5000	288.483	46.676

4. Summary

This paper presents a multi-database parallel algorithm, the use of multi-threaded features, the task can be parallel processing, and further improve the speed of statistical queries. Finally, this paper proposes a multi-table paging algorithm with high data volume under the research of multi-table paging process, which effectively solves the problem that the multi-table paging query is slow in a large amount of data. In the aspect of structured data, this paper first divides the database vertically according to the characteristics of the business, stores the data of the write operation more frequently and stores the data more frequently in the different database, and improves the reading and writing performance of the data. For the basic information of equipment and statistical information related to the existence of large concurrent read operation, the data has been cut again, the two kinds of information on different libraries, and further improve the system to deal with high concurrent read query capabilities.

References

[1] Moniruzzaman A B M, Hossain S A. Nosql database: New era of databases for big data analytics-classification, characteristics and comparison[J]. arXiv preprint arXiv:1307.0191, 2013.

[2] Böhm A, Dittrich J, Mukherjee N, et al. Operational analytics data management systems[J]. Proceedings of the VLDB Endowment, 2016, 9(13): 1601-1604.

[3] Hashem I A T, Yaqoob I, Anuar N B, et al. The rise of “big data” on cloud computing: Review and open research issues[J]. Information Systems, 2015, 47: 98-115.

[4] Bakalash R, Shaked G, Caspi J. Data aggregation module supporting dynamic query responsive aggregation during the servicing of database query requests provided by one or more client machines: U.S. Patent 8,799,209[P]. 2014-8-5.

[5] Storey V C, Song I Y. Big data technologies and management: What conceptual modeling can do[J]. Data & Knowledge Engineering, 2017.

[6] Kumar R, Parashar B B, Gupta S, et al. Apache hadoop, nosql and newsql solutions of big data[J]. International Journal of Advance Foundation and Research in Science & Engineering (IJAFRSE), 2014, 1(6): 28-36.

[7] Bakalash R, Shaked G, Caspi J. Data aggregation module supporting dynamic query responsive aggregation during the servicing of database query requests provided by one or more client machines: U.S. Patent 8,788,453[P]. 2014-7-22.

[8] Velankar S, van Ginkel G, Alhroub Y, et al. PDBe: improved accessibility of macromolecular

structure data from PDB and EMDB[J]. Nucleic acids research, 2016, 44(D1): D385-D395.