Application of data fusion in the main station of power acquisition system

Haifeng Lin [a], Tao Liu [b], Guang Chen [c], Rui Zhao [d]

Beijing Kedong Electric Power Control System co., LTD, Beijing 100192, China

[a]linhaifeng@sgepri.sgcc.com.cn, [b]liutao7@sgepri.sgcc.com.cn, [c]chengguang@163.com, [d]zhaorui @sgepri.sgcc.com.cn

Keywords: multi-database fusion, relational database, non-relational database, the electric power information acquisition system

Abstract. The data of the electric power information acquisition system can generate mass data by the accumulation. It is a question to store the data as the increase of sampling frequency. Facing the data properties with large scale and larger gathering density, it will be less efficient if the past technology is adopted in practical applications. In order to solve the problems of the power information mass data collection storage and transfer in real time, a new multi-database fusion technique is proposed, namely the reach, design and implementation of D3Base based on distributed relational database and NOSQL based on non-relational database. It has been applied in key station of the electric information acquisition system. We comprehensive compare the traditional Oracle database with the improvement multi-database fusion method, and illustrate the availability and the practicability of this new scheme by experiment.

1. Introduction

In recent years, the State Grid Corporation puts forward the project of "strong smart grid". However, currently, the electric energy data acquisition system has the similar job content in accordance with the principle that the electric energy data acquisition system should be constructed according to the administrative area, which has the high cost of operation and maintenance, causing a waste of human and financial resources so that there is an urgent need to provide large and complete information service of electric energy data acquisition station. Therefore, it is necessary to research, design, debug and application of the multi - database of the main station. The fusion technology of multi database, which is the combination of relationship and non-relationship, takes advantage of each of them, promotes the key performance of the system and ensures the storage space of the system.

The relational database model is simple with the features of strict theoretical basis, data independence, flexible and convenient operation, good safety, less demanding of users. However relational data can only handle integers, real numbers and strings and other structured data types, but cannot support complex nested data, so relational database has great limitations in the performance and scaling. In today's Internet, IO intensive type is used in many relational databases. In the application of massive data concurrency, the development of relational database is very tedious, and the challenge of this technology is more and more high.

The non-relational database usually has no fixed table structure, and avoids the use of join operation. The advantage of the non-relational database is that it is more suitable for very high speed concurrent read and write operations, and it has low requirements for the numerical consistency, which is conducive to the data storage and access. But the non-relational database has some limitations in the realization of the data integrity.

According to the characteristics of master data in power information system (timing, sustained growth) and the characteristics of the business needs (in terms of time query and localization and system flexibility) and the problems of the electric energy data acquisition master system encountered, in order to design a new electric energy data acquisition master station system, the key database technology is the integration of relational and non-relational. The design reflects the advanced performance, reasonable structure and reliable utility. Specific performance are completeness, uncertainty and timeliness of data, the combination of real time and timing and unattenuated propagation as well as standard data processing and layered data open access interface.

2. Overall structure of electric energy data acquisition system

Through the needs analysis of electric energy data acquire system master station, using existing open source system results, a test database platform has been integrated combining the characteristics of relational database and non-relational data to research on the technology if the integration. The key content in mining massive data petabytes of storage logic structure diagram is shown in Figure 1.

The new alternative electric information acquisition system designed by the verification platform uses data model. According to strong correlation the weak correlation of the business, it stores the strong correlation data, such as the collection of archives information in the traditional relational database engine and puts the weak related data, such as report system, timing generation time series data generated by smart meters and other equipment in non-relational data engine. Of course, it needs to combine the real business requirements and the latest non-relational database technology to find the appropriate data points. With the integration of relational data engine and non-relational data engine, it developed a unified data access interface to simplify the development of business system under the situation that the background data is stored separately.

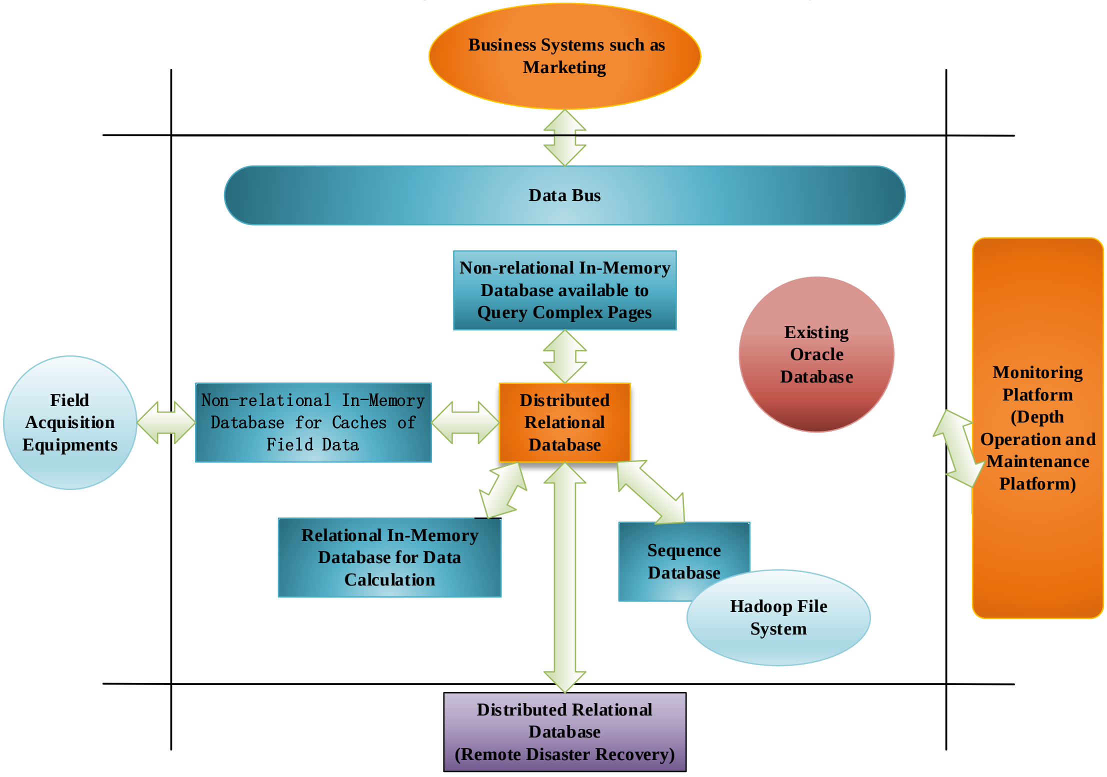
Business Systems such as Marketing
Data Bus
Non-relational In-Memory Database available to Query Complex Pages
Existing Oracle Database
Field Acquisition Equipments
Non-relational In-Memory Database for Caches of Field Data
Distributed Relational Database
Monitoring Platform (Depth Operation and Maintenance Platform)
Relational In-Memory Database for Data Calculation
Sequence Database
Hadoop File System
Distributed Relational Database (Remote Disaster Recovery)

Figure 1. PB level storage logic architecture with massive data

3. Overall structure of electric energy data acquisition system

Multi database integration technology is proposed in this paper to apply into the electricity information collection master system, which relates to the cloud database, cluster technology, memory cache technology, multiple data parallel scheduling, distributed database query etc.

3.1 Cloud Database Technology

At present, Oracle database plays an important role in the data storage and query of electric information collection system. However, the relational database has its limitations in the operation of massive data, which cannot meet the timely and efficient requirements.

It will consume a lot of time to search and calculate the massive collected data generated by the electric information collection system. The operation process has lost its application value, and cannot meet the real time requirements of the system. Therefore it needs to focus on the analysis of each data source storage environment and develops a wide variety of interactive interface among multiple data sources to realize data sharing and exchange operation among the real-time database, cloud storage, and relational database.

Under the precondition of ensuring data security, electric energy data acquisition system takes distributed computing and multi replica technology to achieve massive data real-time storage and fast retrieval based on cloud computing framework, so as to improve the rate of real time concurrent acquisition of massive data storage, real-time and availability.

3.2 Cluster Technology

In practical engineering, high speed network is used to interconnect with each independent computer, and the technology of using software to manage in a single way is the technology of cluster. When the customer and cluster respond to each other, the role of the cluster is the independent server. The participation of the cluster can improve the performance, availability and scalability of the system. In addition, cluster technology, in the case of low operating costs, takes the use of different servers as nodes to achieve high-speed, large amount of computation.

Taking into account the current situation of Chinese power grid, that is, the development of it is very fast and requires the use of electrical information collection system has a good scalability at each level. Therefore, in the important part of electric energy data acquisition system, such as data acquisition, data processing, databases, networks, and other links, they should employ the clustering technique, so that the efficiency, expansion and availability of the system will be greatly improved.

3.3 Memory Cache Technology

The increase of the number of the users, the data items and the frequency of the electric information acquisition system leads to the increasing number of data the system collected, which forms a mass of data. Data statistics, analysis of massive data and the application and expansion of external interface integration will have a serious impact on the performance of the system database. As the memory capacity continues to increase and the hardware price can be accepted, it can make full use of memory cache technology to enhance the performance of the system.

In a computer operating system, different data access is not uniform, namely a few data will be accessed in most of the time period while most of the data is accessed in a few period of time. This is the classical theory, experts pointed out that the majority of data and a few data are for 20% and 80% respectively and most of the time and a few time are for 80% and 20%. After the first visit of a few data, it is stored in the memory, later it can be accessed again directly in memory, at the same time can be marked with different access levels. To do so, it uses the high speed of the memory access, which can improve the system speed. Memory cache uses a new data management method and its architecture, data caching, fast algorithm, parallel operation can be new researched and designed, which has the superiority in data processing speed and precision than non-memory cache method.

In the electric energy data acquisition master system, memory cache technology effectively alleviates the pressure of database reading, writing and calculation and reduces the processing time and it can also provide the reference for the marketing management and analysis of electric power consumption trend.

3.4 Multi Database Parallel Scheduling

In the large-scale computing field, database scheduling is a key technology. For the classical database data fusion technology, the relative spacing between data scheduling command of the same database is relatively close and it exists space and time relativity, which causes the phenomenon of data duplication scheduling or incomplete search. In the main station of electric information acquisition system, the parallel scheduling method of multi database is the complex coupling of cloud fusion. In network environment, after the parallel scheduling of the multi databases, scheduling data takes cloud fusion scheduling method and the parameters take the complex coupling processing to achieve multi database scheduling. Using the method of parallel scheduling in multi database in the information collection system can greatly improve scheduling accuracy and reliability as well as the scheduling scope, time.

3.5 Distributed Database Query

The combination between database system and computer network creates new technology, namely the distributed database system and its advantages are good flexibility and scalability, which can improve its efficiency. But the disadvantage is that the low speed of query processing.

At present, the main research aspect of the distributed database is how to optimize the query database with the short response time and low the cost. The classical algorithms include the half connection algorithm, the direct connection algorithm, the slice copy algorithm and the Hash division algorithm and so on. These algorithms also have limitations, such as low efficiency of data query.

Genetic algorithm is an optimization algorithm that can be used in database query, namely finding the optimal query execution plan from the feasible strategy. The shortcoming of this algorithm is that the query results often fall into the local optimal but non-global optimal. Therefore, this paper proposed a Genetic Algorithm based on distributed database query optimization. It is mainly for the defect of genetic algorithm which is easy to fall into local optimal solution so as to maintain the diversity of population and takes the method of conditional sampling. It optimizes the mutation operator using the Markov chain model to determine the optimal value of mutation operator and then takes crossover and mutation operation to find optimal query execution plan. It can be seen in the database query optimization practice in electric information acquisition system, optimized query algorithm can be in a relatively short period of time to find the optimal query execution plan, speeding up the query speed and improving the query efficiency of the system.

4 The Integration and Realization of the Multi Database

As for the difficulty of huge amount of business data and low operation efficiency of electric energy data acquisition system master station, the test platform combined with the advantages of relational database and relational database to complete the design and implementation of master database of electric energy data acquisition system including the research, design and implementation of distributed relational database D3Base and non-relational database NOSQL.

4.1 Design Principles

The selection principle aims to select the method for different applications. The consistency of key data is always concerned by the designers for the main station of the power consumption information collection system. Handling non-critical data, such as log management and external data integration, designers can consider to use NOSQL. In the design process, it can compare the advantages and disadvantages of the relational and the non-relational database to take the advantage and avoid the short board so as to achieve organic combination to design and implement the application.

4.2 Performance Test

At the experimental stage, the initial performance test of D3Base and Oracle11g was started.

This comparison test follows the following principle:

The key of experiment is the query performance and response time of the system. The experiment environment must be same, namely the same software and hardware. The experimental data is from the system and the Oracle test data is taken as the benchmark.

The server of the Oracle is 3650M IBM (2cpu, 64G memory, 20T hard disk), and the operating system is: Linux Oracle 6.5, Oracle 11g2R DBMS.

The server of the D3BASE is 3650M (2cpu. 64g memory, 20t hard disk), 3 Baode server PR2000R (2cpu, 32g, 5*2T hard disk), the combination for 4 (database plus storage) and 1 lock controller (concurrently). Operating systems is Oracle Linux 6.5, DBMS: D3Base-DFM (Management Framework) plus D3base-Node (database node).

In the experiment, all the used server connect to the server through the Gigabit Ethernet switch, and IP address is in the same network. Test terminal CPU: T8100, 2G memory, Windows 7.

The sample data file is 2402675588 bytes in size, record number is 7152924 (7 million 150 thousand). After the data is loaded, the sample data files are repeatedly inserted into the destination table, which are respectively formed to established index and non-established: 200 million records, 800 million records, 1 billion 600 million records and 2 billion 400 million records. Loading 7 million 150 thousand records, Oracle takes 155 seconds while D3Base takes 45 seconds.

Comparing the simple SQL statement operating performance of the two hundred million, 800 million, 1 billion 600 million and 2 billion 400 million data: D3Base is about 5-10 times faster than the Oracle and D3Base is approximately 20-30 times faster than Oracle in index terms. For the index field to retrieve its maximum value: Oracle and D3Base have the similar speed under the index conditions and Oracle is dominant. For count statistics for the non-indexed fields: D3Base is about 100-1000 times faster than the Oracle under the index condition. For the non-indexed field to get its maximum value: D3Base is about 10-200 times faster than Oracle. Data export performance comparison: D3Base is slightly slower than Oracle in the aspect of export data.

Comparative analysis of test results can come to the following conclusion:

(1) Under the same software and hardware conditions, D3Base is 5-10 times faster than Oracle as for integrated query performance. D3Base is not sensitive to whether it has the index, which can save a lot of time and index space. D3Base can be dynamically deployed on several PC servers to become a distributed cluster database, which is conducive to the full use of hardware resources to achieve the purpose of flexible deployment, expansion on-demand and overall performance enhancement. Compared with the D3Base database with 4 node and single node, the former is about 1 times faster than the latter.

(2) Loading data from the sample file under the same conditions, D3Base is 2-3 times faster than than Oracle. The faster the data loading speed, the more conducive to data migration.

(3) D3Base supports SQL92 standard. Packaging NOSQL massive data access storage technology through the SQL interface can reduce the workload and difficulty of application system transplantation.

(4) D3Base can integrate different grades of PC servers to be a distributed nodes and make full use of existing resources.

D3Base uses a distributed architecture, and its performance increases linearly with the increase of the number of servers. The system is easy to deploy and has high flexibility and good scalability. It can ensure that the search response time is less than 3 seconds when ergodic and conditional searching of Single table with 2 billion records.

The D3Base database platform in this project can run on several x86 PC servers to save maintenance costs and reduce the cost of hardware, software and maintenance services.

Memory database REDIS for complex page query, memory database REDIS for field data cache and memory database Codis for data calculation in this project are all non-relational database.

The performance of relational database and non-relational database are compared through experiments. The selected relational database is MySQL. The experiment content is the changes of physical properties when testing the storage of 3MB physical files in the two kinds of database. With the increase of the number of concurrent, little change occurred in throughput of MySQL relational database while the non-relational database throughput has undergone great changes and increases rapidly. When the concurrent number comes to 200, the throughput reaches the peak and then increasing the number of concurrent, the throughput starts to drop. From the experiment, it is also seen that no matter what stage of the database, compared with the relational database MySQL, non-relational database has obvious advantages. In addition, the advantage of the non-relational database also includes a high scalability and availability, which is suitable for the management of large-scale data especially in the emerging applications, but also faced with challenges.

5 Conclusion

Based on data characteristics and real-time requirements of large power information acquisition system, this paper proposes a new scheme for multi-data fusion making improvement for the relational database. At the same time, the organic combination of the latest technology of non-relational database and the new multi-data fusion scheme can generate a new frame diagram of electric energy data acquisition system which can apply into the electricity information acquisition station test platform on the basis of the theory. The test platform takes a variety of technical innovation and integration like the cloud database technology, group technology, memory cache technology, multi database parallel scheduling, system data cloud storage and real-time database, achieving the realization of the optimization design of relational database D3Base and non-relational database which can effectively solve the various problems encountered with the mass data of power system. Verification system takes replace the Oracle and stores part of the system of historical data in mining system, the time of several longest typical query function reduces from more than 40 minutes to 3 minutes, basically solving the slow query speed and it has fully proved its efficiency and reliability.

References

[1] Yang Zhonghua, Xu Jun. Research on the application of large data processing technology in electric information system based on memory cache [C].2014 National Industrial Control Computer Technology Annual Conference Collection.

[2] Hector Garcia Molina,Kenneth Salem An overview of main memory database system [J] IEEE Transactions on Knowledge and Data Engineering 1992, 4 (6) 509-516.

[3] Fu Jinyu, Chao Yuqiang, Li Jie. Cluster technology application in electric energy data acquisition system [C]. 2012 Power System Automation Committee academic exchange seminar paper set.

[4] Xie Yuhua. The complex coupling system simulation based on the multi database parallel scheduling cloud fusion [J]. Bulletin of science and technology, 2015 (5).

[5] Jia Dongli, Meng Xiaoli. Study on the application of real time database in the electric information acquisition system [J]. Electric power construction, 2012 (1).

[6] Yang Zhenhua. Research on data processing and application of smart grid [D]. Hunan: Hunan University, 2011.