Research on the Bid Data Technique in Electric Power Industry

Wang Chunying, Li Wencui, Wu Lijie, Yang Yi, Shu Xinjian and Zhang Yong

Information & Telecommunication Co. of State Grid Henan Electric Power Company, Zhengzhou, 450052, China

Keywords: big data technique; data acquisition; data storage; data computation

Abstract. With the rapid development of our country electric power industry, as well as the construction of the smart grid, the electric power industry are increasingly demanding data management and analysis, etc；The big data technology is introduced into the electric power industry applications, not only for the electric power enterprise data information analysis and management to provide the necessary technical support, also can promote the electric power enterprises to further improve the service level. This article mainly expounds the technology of data, and the practical application of electric power industry are further analyzed.

Introduction

Big data technology refers to the huge amount of data from various types of fast access to valuable information technology. The core is big data technology to solve the problem of big data. Mainly can be divided into: data collection, data access, infrastructure, data processing and statistical analysis, data mining, model prediction, results of 8 kinds of technology [1]. Big data technology mainly formed the batch, stream processing and interactive analysis of three kinds of calculation model. The big data technology is introduced into the electric power industry applications, not only for the electric power enterprise data information analysis and management to provide the necessary technical support, also can promote the electric power enterprises to further improve the service level.

Big Data Research Basic Theory and Technical Route

Big data is not only refers to the vast amounts of information, more emphasis on human screening and processing of information. Big data processing method has a lot of, big data processing procedure of general application, can be summarized as four steps, are gathering, import and pretreatment, statistics and analysis, finally the data mining.

Gathering. Big data acquisition refers to the use of multiple databases to receive from the client (Web, App or sensor forms, etc.), and the user can through these database for simple queries and processing work. For example, electric chamber of commerce use the traditional relational database MySQL and Oracle to store every transaction data, in addition, NoSQL database such as Redis and MongoDB is often used for data collection.

In large data collection process, its main characteristic and challenge is high concurrency, because at the same time may have tens of thousands of users to access and manipulate, such as train ticketing website and Taobao, they reached millions of concurrent traffic during peak, so need to deploy a large number of database on the acquisition end to support. And how in between these database load balancing and subdivision is the need for in-depth thinking and design.

Import/pretreatment. Collection side, although there will be a lot of the database itself, but to the analysis of these huge amounts of data efficiently, or should the data import from front end to a centralized large distributed database, or distributed storage cluster, and can be based on the import do some simple cleaning and pretreatment. There are some users will use when importing data from Twitter Storm to flow calculation, to meet the demand of real-time computing in the business.

Import and challenges to the characteristics of the pretreatment process is mainly import large amount of data, often correlates to the import amount of every second, even gigabit levels.

Statistics and Analysis. Statistics and analysis the main use of distributed database, or distributed computing cluster to store huge amounts of data in its ordinary analysis and classification summary, etc., in order to satisfy the demands of most common analysis, in this regard, some real-time demand would use the EMC GreenPlum, Oracle Exadata does, as well as the column type based on MySQL storage Infobright, etc., and some of the batch, or demand can use Hadoop based on semi-structured data[2].

Statistics and analysis of the key features of the part and the challenge is to analyze involves large amount of data, the system resources, especially the I/O will be great.

Data Mining. As the previous statistics and analysis, data mining, and laborers typically have little predefined theme, mainly on the existing data calculation, based on all kinds of algorithm to Predict effect, so as to realize some high level data analysis needs. Typical algorithms are used in the Kmeans clustering, used in statistical learning Naivebayes and used for the SVM classification, the main use of tools such as the Hadoop Mahout.

The characteristics of the process and the challenge is mainly used to mining algorithm is very complicated, and the calculation involves the amount of data and the large amount of calculation, commonly used data mining algorithm is given priority to with single thread.

A processing model is now widely accepted Fayyad design multiprocessing stage model, etc., Model is shown in figure 1.

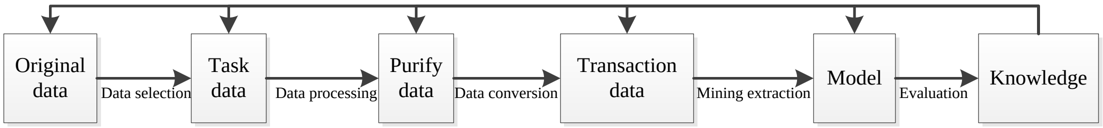

Fig.1 Many processing phase model

At present, big data research mainly as a research method or a tool of discovering new knowledge, rather than the data itself as the research target, it is closely related with the traditional data mining methods have radically different.

Technology Research and Application

The Technology of Large Data Parallel Computing Research. To solve the huge amounts of data, large concurrent and can't meet the demand of system performance under complex computing environment, the objectification parallel computing framework based on the technology of large data system was investigated, the system will be objectification, parallel computing organically fuses in together, technical framework called objectification parallel computing framework, this system based on data cache technology, the business system of various kinds of data cached in memory in the form of objects, provide efficient index of the object and access. Use the objectification parallel computing technology, computing function encapsulation in internal and external services through object interface object, by calling the different objects on the server object interface, the realization of parallel computing. The technique of lateral extension framework supports data, through online extension object server to cache the additional data, provide data based on object computing services. First target the Business data, the object is to replace two-dimensional table, is advantageous for the realization of data organization and management, reduce the development difficulty, improve the development efficiency. Through the parallel computation, the statistical analysis on the large data into small data statistical analysis, statistical analysis can improve the performance [3]. Small data statistical analysis and with the help of the memory is better than disk computing performance, can significantly improve the performance of statistical analysis.

Within the company uses the building cluster, 4 PCS set in grid quality supervision and management system of medium voltage power transmission reliability index statistics as test object, its all provinces and regions 1 years (since 2012-1-1) registered and operating data as a statistical data range. Using big data technology test objectification parallel computing framework system performance, statistical analysis and calculation of performance improvement more than 100 times.

And with the increase of task concurrency, system reliability is 100%, the data quantity increases, the number of servers by increasing the object, basic task execution time remains the same, horizontal extension performance is good.

The system, which is based on grid assets quality supervision and the debugging, deploying in headquarters, during commissioning, system operation is stable and reliable, overall performance improvement more than 100 times, meet the real-time requirements of the system. In this system on the basis of existing achievements in the future will be summarized, formed on the basis of large data technology, support data cache the objectification of the parallel computing framework system, at the same time for other business scenario research supports multiple data sources and framework based on the data cache. The above for the research of this paper work laid a solid technical and practical basis.

The Performance Optimization of Large Data Technology Research. Electricity information acquisition system according to the "full coverage, full collection, full fee charged with" the unity of the request, to the company in 2015 smart meters will amount to 300 million terminal users. As the sampling terminals and a surge in frequency, electricity information collection system involved in the data size will take on explosive growth. Such as Zhejiang province electric power company electricity information collection system user scale will reach 23 million users, more than 3 million terminal size, change and the user more than 400000, annual growth of 12 t, all electricity company system information collection and the data of annual growth of about 300 t.

To solve the huge amounts of electricity information collection and data storage, analysis, statistics and other business demand brought serious challenges to the existing system, using big data technology in Zhejiang electricity information acquisition and data analysis based on cloud computing demonstration of construction, the main construction "1+6" platform environment system, that is, a platform six core application-a platform as electricity information collection system of cloud computing platform, six core applications including the user meter reading data and the power of data storage and query, the user power statistics, area line loss calculation analysis, packet data analysis, low rate calculation and analysis, with complete data file class data ETL.

With eight ordinary PC server (350000) to build cloud platform (1/10 of the total cost for the original system hardware), query performance increase 5 to 18 times, analysis and calculation of performance improvement 5 to 15 times, the overall performance improvement, on average, eight times more. And with the increase of the amount of data, cloud computing performance can be implemented as cheap server nodes increased nearly linear growth, and relational database schema is limited by the structure of the Shared memory limit, cannot achieve effective lateral extension [4]. Oracle will take significantly increased, compared with a cloud platform performance advantages became apparent.

Conclusion

Electricity information acquisition system according to the "full coverage, full collection, full fee charged with" the unity of the request, the sampling terminals and the electricity information collection system involved in the data size will take on explosive growth. To solve the huge amounts of electricity information collection and data storage, analysis, statistics and other business demand brought serious challenges to the existing system, we adopt the big data technology the information acquisition and data analysis based on cloud computing .

References

[1] Yan Ming-liang. Research on Big Data Technologies and Applications in Power Industry[J]. Journal of Nanjing industry professional technology institute, 2015.

[2] Sun Da-wei, Zhang Guang-yan, Zheng Wei-min. Big Data Stream Computing: Technologies and Instances[J]. Journal of software, 2014.

[3] Chen Wei. Opportunities, Challenges and Methods of Electric Data Auditing in Big Data

Environments. Computer science, 2016.

[4] Wang Jiye, Guo Jinghong, Cao Junwei, Gao Lingchao, Hu Ziwei. Review on Information and Communication Key Technologies of Energy Internet[J]. Smart Grid, 2015.