RESEARCH ARTICLE   OPEN ACCESS

Investigation on Giant Data leaning Data Analyzing and Processing knowledge

Mr. Mehari Mesfin [1], Mr. Tewhasom Aregay Weldu [2], Dr. Narayana Babu [3]

Lecturer [2], Lecturer/HOD [3], Assistant Professor

Department of Computer Science

Adigrat University - Ethiopia

ABSTRACT

At present, enterprises have urgent needs to conduct an effective and stable statistical analysis on big data. With a view to solving the issue of analysis and processing of big data in respect of enterprise business, this essay proposes a hybrid structure mode based on the Map Reduce technology and the parallel database technology, discusses the principle on which the mode is used to realize the analysis and processing of big data and its advantages, and analyzes, expounds and proves the hybrid structure and provides a practical plan on big data processing. It is expected that this study has certain reference value in related researches.

Keywords:- data leaning, processing knowledge, big data

I. INTRODUCTION

With the continuous improvement of informatization construction, most enterprises have completed the deployment of informatization system and the research and development, design and promotion of new products and new services have been greatly improved in respect of the operation efficiency of compared to traditional enterprises. However, although the enterprises have realized a high-efficiency and elaborate management, masses of business data are concurrently accumulated. Furthermore, there are categories of data and the requirement of being real-time is quite high. This is so- called "big data". All of the internet of things, cloud computing, mobile Internet, large-scale e-commerce, mobile phone, tablet computer and the various sensors spread all over each corner of the earth are data sources or the carrying way[1]. At present, as a relatively new concept, big data is not proposed directly as the proper noun to give policy support by the Chinese government. However, in December 8, 2011, the Ministry of Industry and Information Technology of People's Republic of China issued the Internet of things "Twelfth Five-Year Plan", put forward as the information processing technology is the analysis of massive data storage, data mining, image and video intelligent analysis, which are important parts of big data [2]. The other 3

key technical innovation projects including information technology, information technology, information security technology, are closely related with the large data. Large data has four characteristics: first, a huge amount of data, using PB as the unit; second, various data types, including network log, video, pictures, geographic location information and other data; third, the low density of value, with video as an example, the valuable data may be only one or two seconds in the continuous monitoring process; fourth, fast processing speed, this point is essentially different with the traditional data mining. Big data contains masses of valuable mode and information. For example, Wal-Mart, a global retail giant, will mine geographic locations, sales performance and social information of stores in big data to improve customers' understanding [3]. For enterprise organizations, the value of big data is reflected in the two aspects: analysis and use and secondary development. The enterprises need to rely upon the technical platform on informatization system to mine in big data any needed important information and conduct intensive analysis and processing of these big data so as to construct a data warehouse and application and analysis platform based on important data information. By analyzing and processing the masses of the data information of enterprises, it will provide a correct guidance and policy-making for the existence and development of enterprises [4].
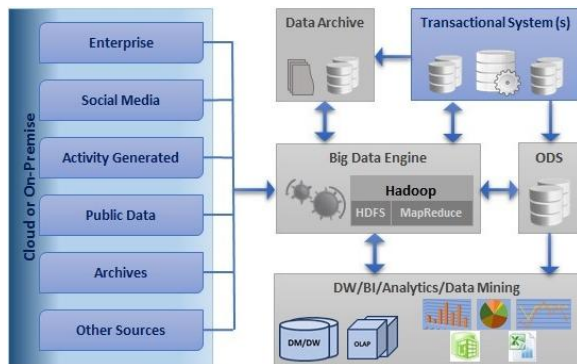
Figure: 1 illustrates big data management system structure.

To design a system platform based on big data analysis, this essay proposes a solution which is big data oriented and integrates storage, management, analysis, processing and application. This proposal is very practical upon being tested and analyzed. Theoretically, there are no limits to the improvement of the processing function of big data. At present, in the practical application on the processing of big data, the most often used and mainstream realization technology is the MapReduce technology, parallel database technology and the hybrid structure technology based on Map Reduce technology and parallel database technology[5].

MapReduce Technology

In 2004, MapReduce was proposed by Google, it is an object-oriented programming model to deal with the large data, primarily used for processing internet data, such as document capture, inverted index construction. But because MapReduce has a simple and powerful data processing interface, and it hides many details on massively parallel execution, fault tolerance and load balance implementation, so the technology has been widely applied in the field of machine learning, data mining, data analysis, text tokenization, indexing research, creation of other kinds of processing of complicated business logics involved in the parallel programming. It realizes complicated the includes two basic operation conceptions: Map(Mapping) and

Reduce(Simplication). The Map technology mainly processes a group of input data record and distributes data to several servers and operation systems. Its means of processing data is a strategy based on the key/value. The Reduce technology mainly occupies itself in summarizing and processing the result after processing the above key/value [6]. Issues and tasks in the real world may be modeled and described by means of this simple means of processing. Programs realized by this means will be distributed cluster. The data processing algorithm based on issues will be distributed to the distributed system formed by ordinary computers and then executed. The system will then solve the problem on the details in relation to the input of big data. Then the algorithm programs crossing in a slide presentation, Google offers the following applications of MapReduce: distributed grep, distributed sort, web link-graph reversal, term-vector per host, web access log stats, inverted index construction, document clustering, machine learning, and statistical machine translation[7].

Parallel Database Technology

During the present phase, the popularizing and applying of relational databases feature a widest scope and they are at a mainstream position in the whole database system field. The original design object of relational database is to realize the application of the large-scale machines based on the "Host Computer – Terminal Computer" mode; however, its application scope is very limited. With the popularity and application of the "Client - Service", the relational database system brings about an application era of "Client - Service" and is widely developed and applied. However, with the popularizing of the Internet technologies, the Internet information resources begin increasingly complicated, and the relational database begins to become unable to apply to complicated Internet application and cannot be used to express and administer each type of complicated document type and multi-media resource information. Therefore, the relational database system is improved and adjusted on this regard, such as adding the support function on the database system that is object oriented, at the same time, adding the function on handling each complicated information data [8].

Database processing technology based on

parallel computing is a technology blended with the parallel computing mode and database processing technology. It originated in the seventies of the 20th century, mainly studies the parallelism of the relational algebra operations and the hardware design for implementation of relation operation, hope to realize some function of relational database operation by hardware. Unfortunately, the study failed. In the late 80s of the 20th century, the research direction of parallel database technology turned gradually to the general parallel machine, and the research was focused on the physical organization of parallel database, operative algorithms, optimization and scheduling policy. From the 90s until now, with the development of the basic techniques for processor technology, storage technology, network technology, the parallel database technology rise to a new level, the focus of the research is also transferred to the time and spatial parallelism of the data operation [9].

In the processing and analysis of the big data, data parallel processing manner is essential. Because the processing strategy of "divide and rule" provides unlimited reverie to extend system performance. With the fast development and application of parallel processing and computing technology, people get to know that the processing of a parallel mode can be realized by means of the conceptions of time or space so that the processing efficiency of system tasks can be improved. After the entry of enterprises' business information data into the big data era, this parallel computing mode can help well solve the data processing issue for large-scale enterprises' business data system. Parallel computing includes two aspects: data parallel processing and task parallel processing. In terms of the data parallel processing means, a large-scale task to be solved can be dissembled into various system sub-tasks with the same scale and then each sub-task will be processed. As such, compared to the whole task, it will be easy to process. Adopting the task-paralleling processing mode might cause the disposal of tasks and coordination of relationships overly complicated. Using the parallel database technology is a means for realizing the parallel processing of data information. Parallel database support standard SQL language, through the SQL to provide data access service, SQ L is widely used because it is simple and easy to apply. But in big data analysis, the SQL interface is facing great challenges. The advantage of SQL comes from packaging the underlying data access, but the packaging affects its openness to a certain extent. User-defined functions which provided by parallel database is mostly based on the design of a single database instance, and therefore they cannot be executed in parallel cluster, it means that the traditional way is not suitable for the processing and analysis of big data[13]. Moreover, the user-defined functions often need to pass complex system interaction in parallel database, and is familiar with the database structure and system calls, so it is difficult to use. Parallel database design is based on high-end hardware, software fault tolerance ability is poor, so the augment ability of parallel database is limited, the traditional data warehouse based on parallel database usually completed data preprocessing and analysis show with the help of external tools, so the data processing and analysis process involves a lot of data migration and calculation, of course, the analysis delay is often higher.

Constructing the Pattern on Big Data Processing

For the analysis and processing of big data, using a system based on the parallel database and data warehouse is not an ideal plan for the analysis and processing of big data. Using the combination of MapReduce and parallel database combines the advantages of the two means. After the analysis and comparison of the two means, the advantages and disadvantages of the two means can be seen. For such reasons, constructing a big data processing pattern based on MapReduce combined parallel data can on the one hand make up for their own disadvantages and can on the other hand improve the reliability of the system [12].

The hybrid structure mode of associating MapReduce with Parallel Database

SQL, as a universal relationship database system scripting language, has been widely used in the field of relational database, and SQL can be applied in the parallel database. MapReduce defines a self-defined interface function for SQL scripting sentence and provides the same grammatical form as common SQL scripting sentence. Within the self-defined function realizes the data processing based on parallel

computing. At the same time, such processing mode based on paralleling is applicable to the enterprise distributed system and can be used to hundreds of servers. These parallel processing machines and their computing of services conceal the realization details to users and are transparent for users, as a result of which big data processing can be realized by using MapReduce modules through SQL scripting sentence[3]. The interface function of such MapReduce can normally operate under the database system circumstance, and its returned result sets remain a usable data table.
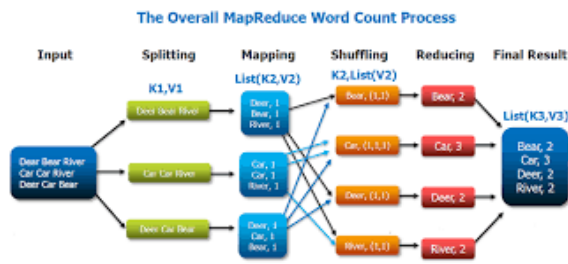

Figure 2 illustrates the pattern on big data processing based on the MapReduce framework.

The loading, analyzing and processing of a large amount of data supported by data warehouse engines based on MapReduce can satisfy the need for the construction and mining needed for enterprise big data and can apply to the need of the performances and functions of the present and future information system [10]. Below is an example on a Hotel Chain as the data analysis of solutions.

System Environment
1. Database Servers
Branch 1: DELL Power Edge R910.
Branch 2: DELL Power Edge R910.
Branch 3: DELL Power Edge R910.
Master Server: Dell 2950.
Data Storage Equipment: 5 MD1000 Direct Connection Storage[13].

2. Operation System
Two branches and the master server are respectively fixed FreeBSD Linux 9.0.

3. Network Environment One switchboard for Gigabit LAN optical network, No Blocking Switch mode and Gigabit LAN optical network.

4. The Service Architecture of the Information System Clusters of a Hotel Chain.

II. PERFORMANCE TEST OF SYSTEM

Statistical Analysis of Data The same business data will be loaded to Greenplum, Oracle and separate SQLSERVER database, the same SQL sentences will be executed by MapReduce. The time for the execution of sentences will be taken notes of the systematic structure based on MapReduce is found to be highly efficient through test comparison[14].
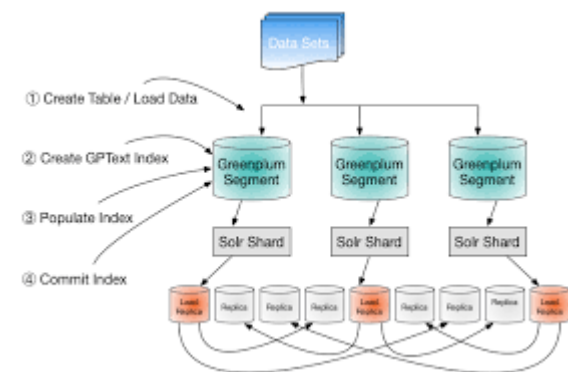

Figure 3: illustrates Data Statistic Analysis

III. CONCLUSION

This essay introduces the mode of big data analysis and processing based on combined and parallel by MapReduce technology into the processing of big data of a hotel chain's information system, draws the loading rate of engine data, and makes comparison as to the execution time for loading the same data as such databases as Oracle and SQL SERVER. The result shows that a combined structure mode based on the combination of MapReduce technology and parallel database technology can improve the disposal efficiency of big data processing.

Simple function integration can't effectively solve the problem of big data processing, so the research on hybrid architecture also need further. There is a distance between the scalability of parallel database and the demand of big data analysis, so it is a very challenging task to improve the scalability of parallel database. Despite the performance of MapReduce is

increasing rapidly, there is still a large promoted space on some hands such as the parallelization of multiple analysis, complex analysis operation display, data compression efficiency.

REFERENCES

[1] Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters/ Proceedings of the 6th Symposium on Operating System Design and Implementation(OSDГ 04) .San Francisco, California, USA, 137-150.

[2] Xu Zipei , The Big Data Revolution. Guilin: Guangxi Normal University Press, 2012.

[3] Peng Hong, and Du Nan, " Research of parallel technology in massive commerce data management system ", Application Research of Computers, Vol. 26, No. 2, 2009, pp. 614-616.

[4] Wang Guiqiang, and Lu Chaojun, "Probing Parallel Technique-Based Statistical Analysis For Enormous Data", Computer Applications and Software, Vol. 28, No. 3, 2011, pp. 162-165.

[5]http://www.asterdata.com/product/mapreduce.php

[6] Wang Min, Zhang Hong, and Yan Peng, "Parallel Technique Analysis for Effective Reducing Test Cost", Computer and Digital Engineering, Vol. 38, No. 9, 2010, pp. 13-15.

[7]http://www.dbms2.com/2008/08/26/known applications of mapreduce/

[8]   Li Huazhi, Database Solution, Beijing: Publishing House of Electronics Industry, 2010.

[9] Das S, Sismanis Y, Beyer K S, Gemulla R, Haas P J, McPherson J, Ricardo: Integrating R and Hadoop.   Proceedings of the SIGMOD. Indianapolis, the United States, 188-190.

[10] Yu Chuli, Xiao Yingyuan, and Yin Bo, "A parallel algorithm for mining frequent item sets on Hadoop",Journal   of   Tianjin   University ofTechnology, Vol. 27, No. 1, 2011, pp. 25-28.

[11] Wang Shan, Wang Huiju, and Qin Xiongpai, "Architecting Big Data: Challenges, Studies and Forecasts", Chinese Journal of Computers , Vol.

34, No. 10, 2011, pp. 1741- 1752.

[12] Zhang Li, SQL Server Database Principle and Application. Beijing: Tsinghua University Press, 2009.

[13]http://www.greenplum.com/technology/mapreduce/