Hadoop Application of Technical Analysis of Large Data

JIANG Li
Conghua, Guangzhou Guang Hua from the 13th Avenue, Institute of Software Engineering, Guangzhou University

ABSTRACT: Big data analysis is one of the main applications of big data technology. This paper introduces the basic methods and principles of big data analysis, current academic and industry-wide adoption HDFS distributed file system and MapReduce programming model for building big data analysis techniques. Based on the analysis hadoop technology in mainstream non-relational data, based on comparative advantage in HBase, Hive, HadoopDB other mainstream non-relational database systems and traditional database, and describes the relevant algorithms. Using Hadoop core technology provides a platform for new big data analysis applications, presents a big data analytics in the enterprise Solution, a good solution to big data analytics in the enterprise application advantages.
KEY WORDS: big data analysis; visualization; MapReduce; Hadoop technology

1 INTRODUCTION

With the development of information technology, Chinese enterprises are gradually aware of various types of business value and commercial value of big data, where the market share of small and medium enterprises in China is large, this will be an extremely large data impetus. Traditional business intelligence through increased data types and sources, improve analysis speed, response to a growing number of data sets, and gradually evolved into big data analysis. Big Data Analysis [1] have gone through three different phases, the first batch analysis stage, mainly from the company's internal structured data (such as enterprise relational database) based. The second phase of near real-time analysis, data analysis from the traditional structured data and gradually evolved into a structured, unstructured (audio, video, SNS, etc.) [2], and semi-structured data (Linux system log, customer information). The third phase, real-time analysis, data sources and types of richer, not only from within the enterprise production data, user data, and social networking sites, but also included data from a third party, such as real-time monitoring of competition, the target user groups purchasing behavior monitoring [2] [3].

2 THE ANALYSIS OF BIG DATA AND CHARACTERISTICS

Data analysis is collected, processed data, and the process of obtaining information [4]. Big Data Analytics is a huge-scale data analysis, can help companies better adapt to change, to make more informed decisions. It is built on the basis of traditional data analysis, including the following aspects:

Large scale data analysis is a huge data analysis, looking for patterns, correlations and other useful information in the study process large amounts of data that can help those who need to better adapt to change, to make more informed decisions.

2. The analysis of big data and characteristics

Data analysis is collected, processed data, and the process of obtaining information [4]. Big Data Analytics is a huge-scale data analysis, can help companies better adapt to change, to make more informed decisions. It is built on the basis of traditional data analysis, including the following aspects:

Large scale data analysis is a huge data analysis, looking for patterns, correlations and other useful information in the study process large amounts of data that can help those who need to better adapt to change, to make more informed decisions.

(1) visual analysis: analysis of large data users and analysts are divided into ordinary users, their big data analysis is the basic requirement of visual

analysis, as visual analysis can visually presents the characteristics of big data, the data show, users see the results of big data analysis.

(2) Data Mining: Data mining theory as the core theory as big data analytics, and its correlation algorithm based on different data types and formats can be more scientific data showing a Bunsen characteristics, can process and analyze data faster. Based on data mining algorithms like C4.5 [4], K-Means algorithm [5] in the efficiency is very high, these algorithms for large data analysis provides a very good foundation. Data mining is given, visual analysis of machine memory is available to the user decision analysis. In data mining algorithm can handle large data mass data, while a certain extent to meet the speed requirements of big data.

(3) Predictive Analytics: Predictive analysis enables analysts to predict judgments based on the results of visual analysis and data mining.

(4) Semantic Engine: Characteristics of heterogeneous unstructured data with data diversity, when conducting the data analysis requires a series of tools to parse, extract, analyze the data [6]. Intelligent semantic engine is designed to extract information from documents, from the characteristics of large data mining, scientific modeling and enter new data, forecast data for convenience.

(5) Data quality and data management: big data analysis we consider data quality and data management from two aspects, high-quality data and effective management to ensure the authenticity and valuable analytical results.

3 ANALYSIS OF THE HADOOP BIG DATA TECHNOLOGY

Implemented to achieve high data consider several factors, based on open source systems, robustness and fault tolerance, scalability, scalability, easy maintenance and so on. Based on these requirements, Hadoop technology produced. Hadoop is an open source Apache Foundation supports framework, Hadoop consists of two parts HDFS, MapReduce, the former is a distributed file system, which is a distributed computing system [7]. Hadoop using the Java implementation that can run on commodity hardware and different operating systems such as Linux, Mac OSX, Windows and Solaris.

HDFS is a distributed file management system, was a master / slave architecture, a master node becomes the name of the node, the rest of the computer is a slave node, the data can also be called a node. The role of the master node is to manage metadata from a node to store and manage application data. In order to support large data on HDFS built NoSQL database systems, NoSQL is a non-relational database system. HDFS through the interface language pig to build a column store database system, in addition, Hadoop provides a data warehouse on HBase / data mining software Hivi [8], and in order to meet the requirements of machine learning, but also provides a machine packages learning Mahout, to meet the requirements of large data management and analysis.

Hadoop core algorithm is Map-Reduce, called mapping - Statute of the algorithm, the algorithm ideas and simple, it is a popular task, and a computer can not do, Map-Reduce algorithm [9], it can be mapped into many subtasks, and give many machines do, each machine is done, the final results to the user in a statute, which is the Map-Reduce algorithm is the core idea.

4 BASED ON THE HADOOP TECHNOLOGY ENTERPRISE SOLUTIONS APPLICATIONS

4.1 *Analysis of demand enterprise applications*

The use of big data analytics to build enterprise-class applications for feature big data, enterprise using Hadoop technology to build solutions need to consider these requirements:

(1) integrate and manage different kinds and different flow rates, and data;

(2) analysis is applied to the high-level information, and does not change the format of the original information;

(3) all available information visualization, for the use of ad hoc analysis;

(4) the establishment of a development environment for the new analytic applications; optimizing workload and schedule progress;

(5) To ensure the security and governance.

This solution provides enterprises using Hadoop core technology provides a platform for new big data analysis applications, can help users achieve the above requirements.

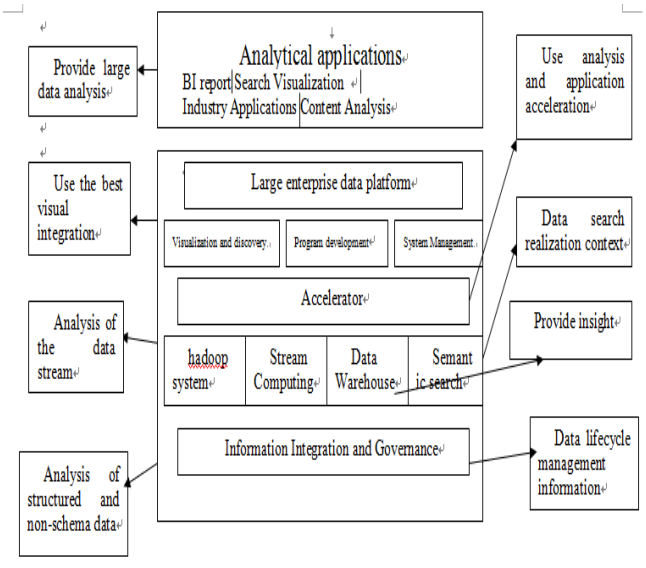

Figure 1 Hadoop system solutions for large enterprise data platform in

4.2 *Analysis of Enterprise Hadoop Solutions Module*

4.2.1 *Hadoop MapReduce framework system*

The system uses Hadoop MapReduce framework, based on a cost-effective way to analyze structured and unstructured information PB level. In applications such as enterprise software architecture level data management and analysis of Internet volume semi-structured and unstructured data, simple and reliable, this program implementation framework based on the open source Apache Hadoop [11]. An increase in the MapReduce framework related to algorithm design, unique technical advantages for businesses, such as workflow management, security management, and integrated into the business in the actual product unique and leading data analysis, machine learning techniques, and text data analysis of mining. All of these enhancements MapReduce framework makes the program can be applied to complex analysis of massive data.

Hive is not an alternative to the data warehouse, it is a complement and extension of the traditional data warehouse, the overall configuration of a broader Internet-level massive data warehouse.

4.2.2 *Big Data platform ilk calculation module*

Flow calculation module, Streams calculation software, which is a breakthrough mobile data analysis platform. Stream computing dynamic collection of multiple data streams, uses advanced algorithms to provide near-instantaneous analysis. In traditional data analysis strategy, data is collected in a database, and is search or query answer, stream computing subvert this strategy can be used to make decisions that require immediate complex dynamic situations, such as the spread of the epidemic forecasting or monitoring preterm child's condition changes.

4.2.3 *Large Data Warehouse module*

Large data warehouse module has a data analysis method has been patented and proven, minimize data movement, while the physical speed to process it, enabling massively parallel processing. To facilitate use of the device for processing a data warehouse. Fast processing speed, low cost. And allows customers to run previously impossible or impractical BI and advanced analytics [12].

Enterprise Data Warehouse Appliance for high-speed analysis and the establishment of its powerful features do not come from the most expensive and most powerful IT components, but how to assemble the right components and performance of the play to the extreme.

Massively parallel processing (MPP) stream multi-core CPU and FPGA acceleration Netezza's unique streaming technology engine (FASTTM) combined so as to provide even a very expensive system that can not even come close to matching the performance. And, as a very simple device to use, only requires the user to input commands, the system will be able to speed the results surprising direct feedback to the user, without the need for indexing or make any adjustments to the system and optimization. Simplicity of the device makes it easy to develop applications that rapid innovation and performance analysis capabilities for a wide range of users and processes to provide services.

5 CONCLUSION

Based on the analysis of big data, Hadoop technology is proposed for large data analysis and presentation of the relevant calculation framework and algorithms, and use Hadoop to build enterprise-class big data analytics solution, this solution can meet the business needs function. Follow-up work in this paper focuses on the following aspects: a large data analysis algorithms based on the MapReduce framework of the research; 2.Hadoop achieve high performance data analysis problems, mainly for HadoopDB, load performance Hbase; and three, the business needs to improve. Hadoop technology and a good meeting point.

REFERENCES

[1] YuanXiaoChao as the era of big data visualization and visual analysis of the opportunities and challenges. 2013.http://www.chinacloud.cn/upload/2013-12/131228145655172.pdf.

[2] Lei study any interactive technology information visualization [PhD thesis]. Beijing: Institute of Software, 2009.

[3] REN Lei, Wang prestige, Zhou Mingjun, Teng Dongxing, Ma Cuixia, Dai Guozhong, Wang Hongan a driven interactive information visualization software development methodology Journal, 2008,19 (8): 1947-1964. http://www.jos.org.cn/1000-9825/19/1947.htm [doi:10.3724/SP.J.1001.2008.01947].

[4] Li Bohu, Zhang Lin, Ren Lei, Chai Xudong, Tao Fei, Wang Yongzhi, shoots Chao, Huang Pei, Zhao Xinpei, CHONG typical characteristics of cloud manufacturing, key technology and application of computer integrated manufacturing system and, 2011,18 (7): 1345-1356.

[5] Cui W, Liu S, Tan L, Shi C, Song Y, Gao Z, Qu H, Tong X.TextFlow: Towerads better understanding of evolving topics in text. IEEE Trans.on Visualization and Com-puter Graphics, 2011, 17 (12): 2412-2412. [doi: 10,1109 / TVCG.2011.239].

[6] Halevi G, Moded H. The evolution of big data as a research and scientific top-ic: Overview of the literature. Research Trends, 2012,30 (1): 3-6.

[7] Meng Xiaofeng, kindly Big Data management: concepts, techniques and challenges Computer Research and Development, 2013 (1): 146-164.

[8] Zhu Zhu Hadoop-based model of massive data processing and application of. Beijing: Beijing University of Posts and Telecommunications, 2008.
[9] Dong West into .Hadoop Inside Ding Xuefeng Beijing: People's Posts and Telecommunications Press
[10] Hadoop Quick Start [EB / OL]. [2013.09.12] http :: //hadoop.apche.org
[11] Li Renyi data mining research and application of cluster analysis algorithm. Journal of Computer Applications. 2009, 29 (1), 293-296.
[12] Qian Yan Jiang large-scale data clustering technology research and implementation of Chengdu: University of Electronic Science and Technology, 2009.