The Survey of Big Data

Qi Fu[1, a], Jun Tan[2,b] and Yufang Xie [3,c]

[1,3]Modern educational technology center, Jiangxi Science & Technology Normal University, NanChang in China

[2]Jiangxi University of Finance and Economics, NanChang in China

[a]jxsyfq@sina.com, [b]tanjun_jiangxi@163.com, [c]yufang0312@tom.com

Keywords: big data; data processing technology; cloud computing

Abstract. In recent years, the era of researching and applying the big data has already arrived with the continuing increase of data size in Internet applications. The scale effect of big data has a great impact on the traditional technology of data processing. It reforms the existing management model and challenges the original method of data analysis. Therefore, the study of methods and technology on processing the big data is extremely important. Starting from the concept of the big data, this paper summarizes the general process for handing the big data. It mainly focuses on the compare and analysis of the current mainstream processing tools and some key technologies of handing the big data, and finally pointes out the problems and challenges for facing the era of big data.

Introduction

Big data, defined by Gartner Company [1], is the valuable information with high-capacity, high generation rate, wide varieties, which requires new treatments to ensure making judgments, discovering the insight and optimizing the managing. The definition is not only shown a large size of big data, but more important to focus the way to get useful information with a timeliness value from data flow or data block produced dynamically and quickly. However, this data has many types, such as structured, or semi-structured or Non-structured data, which brings a huge challenge to the existing data processing mode. It also reflects the definition of 4V for big data based on 3V. The 4V includes the volume, variety, velocity and value.

The definition of big data by Baidu Encyclopedia is that big data, or a huge amount of data, referring to the amount of data involved is huge, which can not through the current mainstream software tools within a reasonable time to help business to making decision that is more close to the target by managing, processing, and organizing. JohnRauser who is a scientist for big data mentioned a simple definition: Big data is huger amount than processing capabilities of any computer.

No matter what kind of definition,it is clearly that big data is not neither a new product nor a tec hnology. It is only an emerging phenomenon of digital era. Big data gathered by the three major technology trends, which are massive transaction data, massive interactive data and massive processing data.

Processing of Big Data

Although the source of big data is very extensive, data types and applications processing methods all are differ in thousands of ways. But in general, the basic processing of big data is mostly the same. Network and Mobile Data Management Laboratory (WAMDM) in Renmin University of China developed a "Scholar Space" [2], which summed up the general process of big data [3] by collecting from the fields of the relevant computer literature. On this basis, we believe that processing of big data can be divided into four phases which are data acquisition, data processing and integration, data analysis and data interpretation . The whole processing flow of big data is shown in Fig.1.

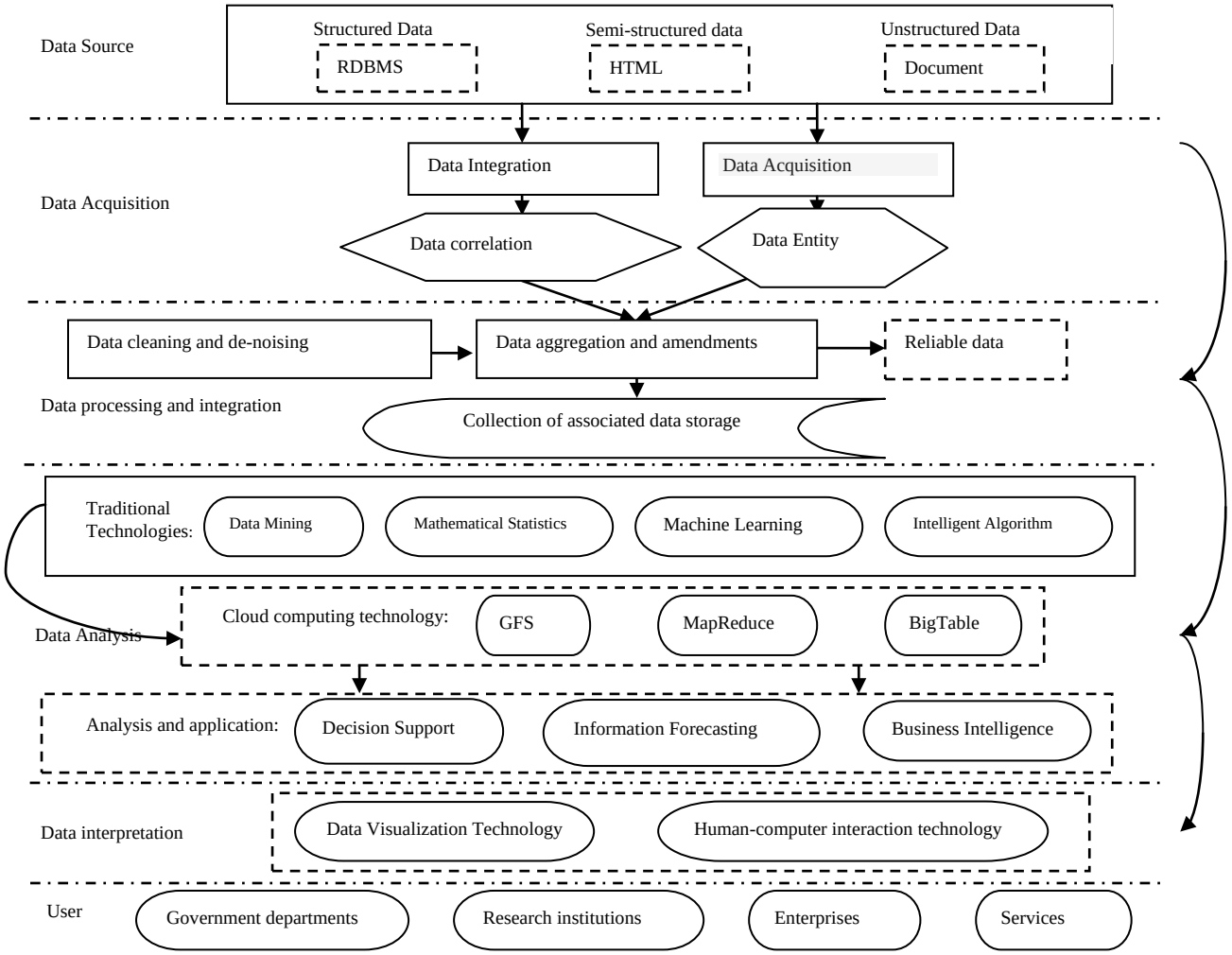

Fig.1 Basic framework of big data processing

Processing Tools and Key Technologies of Big Data

Processing Tools of Big Data

Constantly being updated of big data processing technology also contributes to the emergence of big data processing tools. In this article, we will introduce and compare three big data processing tools.

We are most familiar with block processing platform of Apache's Hadoop. Hadoop is mainly based on programming framework of MapReduce and HDFS [4]. Hadoop appeared relatively early being used more maturely and applied wide range, and it is open source. Using low-cost and large-scale server clusters, data storage and services divided into two levels that are HDFS and HBase, Hadoop maximizes the use of machine resources. That low cost, highly scalable, fault-tolerant, not need to build a predefined pattern, expertise in data processing and analysis of raw data storage, indexing, pattern recognition, recommendation engine and being more used are its advantages. The disadvantage is that the pursuit of high-throughput also brings a delayed batch processing, and data processing of MapReduce has a strong dependence. Large companies using open-source projects must consider technical support and confidentiality.

High Performance Computing Cluster (HPCC) [5] is an open source platform, whose data-intensive processing is distributed. It has the following main components: Thor, Roxie, ECL, ECL IDE, ESP and so on. HPCC fully meets the needs of data-intensive computing, providing a big data flow management services. It is relatively independent between components, processing high-speed and parallel data. It is a language based on data. With good benefits of high reliability and scalability, Amazon has deployed HPCC on its cloud computing platform. The drawback is that,

HPCC has failed in the open source community not to let more large enterprises and developers see the advantages of handling big data. And the development of ecological environment should be improved.

Hadapt is an adaptive analysis platform with high performance. It consists of Hybrid, Storage, Engine, hadoop, HDK and other components. Hadapt combines the advantages of Hadoop with software of relational database management, who both can run in the public and private cloud. Structured data for each node is stored in RDBMS, unstructured data is stored in HDFS, so Hadapt can automatically divide the query tasks between Hadoop layer and relational database layer. HDK allows analysts to create advanced SQL analysis capabilities which have unified structure and flexible model performance, and then can reduce the complexity of the case analysis. The disadvantage is that, Hadapt usually uses the method of expansion-type connector connecting the two different systems, and the result is to bring a certain delay, so this method is very isolated.

The Key Technologies of Big Data

(1) Cloud Computing

As one of the most widely used Internet companies of big data, Google firstly proposed the "cloud computing" in 2006. The so-called "cloud computing", according to the definition in literature is a large-scale distributed model, delivering the data energy, service, storage methods which are abstract, scalable, and easy to manage to the end users through the network. Currently, cloud computing can be considered to include content of three levels: Service (IaaS), Platform (PaaS) and Software (SaaS). Ali Baba and the Cloud Valley's XenSystem in home, as well as Intel and IBM at overseas, who are faithful developers of "cloud computing".

Cloud computing is a core principle of big data analysis and processing technology, which is the basic platform of big data analysis applications [6]. A variety of big data processing technology and application platforms are based on cloud computing in Google. The most typical technology are GFS which is a distributed file system, MapReduce which is batch processing technology, BigTable distributed database which represents big data processing technology, and Hadoop who is the open source data processing platform based on the above of those.

(2) MapReduce

MapReduce is put forward by Google in 2002. As a typical data batch processing technology, it has been widely used in data mining, data analysis, machine learning and other fields. And then, because of its parallel processing way, MapReduce has become the key technology of data processing.MapReduce system mainly consists of two parts: Map and Reduce. The core idea of MapReduce is "divide and rule", that is to say, first data source is divided into several sections, each corresponding to a value of an initial key (Key/Value). And it produces a series of intermediate value of key/Value, after being processed by different mission areas of Map. Shuffle which is the middle process will let all the same key value pass to Reduce after composing a collection. Reduce receives these intermediate value, and combines with the same value, eventually forming a collection of smaller value. MapReduce has simplified the calculation process of lager data, avoiding large amounts of communication expenditures at time of transferring [7]. This has made MapReduce can be applied to a variety of solutions for practical problems. MapReduce has gained a great deal of attentions after being published, and having a wide range of applications in all fields.

(3) GFS

Google File System (GFS) has been developed by Google. GFS has a lot in common with the traditional distributed file systems, such as performance, scalability, availability, and so on. However, according to the effect of application load and the technology environment, the differences between GFS and traditional distributed file systems make it be more widely used in the big data era. GFS uses low-cost hardware and lets some system errors be addressed as a common case, so it has good tolerance of fault. According to the traditional data standards, GFS can handle large files. The large files have usually 100 MB or more, even common reaching to GB, and large files can be effectively managed in the GFS. In addition, GFS mainly takes Master-Slave structure to achieve high-speed storage of massive data through data block and additional update.

(4) BigTable, Megastore, Spanne and F1

With the requirement of development, the second generations of GFS have emerged, which are Colossus, BigTable [8] and Megastore [8]. On the basis of BigTable and Megastore, Spanner [8] has emerged, and its main function is derived from API which has been achieved by GPS and atomic clock. This API can time synchronization accurate to within 10ms between data centers. Based on Spanner servers, Google Institute has put forward the Fault Tolerant Distributed RDBMS (F1) [9] who is the new database in June 2012.

Data as child tables is stored in the child table server in BigTable, and the primary server creates sub-tables. Finally, data as GFS is stored in the GFS file system. At the same time, clients communicate directly with the child table server, which has been monitored by Chubby server [10]. Master server can be viewed whether there is abnormal according to check the status of child tables in Chubby server. If there were abnormal, child server should be terminated and transferred its mission to the rest of the servers. Database has fault tolerance and continuity in BigTable, which can automatically load balancing and be expanded, but it does not provide strong consistency and transaction-level requirements.

Data model in Megastore is similar to the RDBMS. Megastore supports synchronous replication, and data in entities group has a strong consistency of ACID semantics using the Paxos protocol guarantees. However, the throughput is small, and can not meet the application requirements.

Database has the characteristic of temporary multi-version in Spanner, which replaces the version of key-value store in BigTable. The time API in Spanner that can synchronously accurate to within 10ms of time between data centers. And Spanner has advantages of scalable, global distribution, support for external consistent affairs. The disadvantage is that it can not guarantee that all of the nodes are efficient executed in complex SQL queries, and the longer time of delay between data center during data transmission.

Highly scalability of BigTable and functionality of SQL databases are combined in F1. The underlying support by Spanner can provide strong and weak consistency, high availability and high throughput. And transaction committing can be delayed from 50ms to 100ms, reading can be delayed from 5ms to 10ms. The disadvantage of F1 is that parallel query execution, fault recovery, isolation, optimization, migration application and other aspects all face many challenges.

(5) Visualization Technology of Big Data

Data Visualization technology is the theories, methods, or techniques, to convert the data to grap hics or images displayed on screen and run the interactive processing by applying the computer gra phics and images processing technologies. Faced with the emergence of massive data, how to present properly and clearly to the user is an important challenge for big data era. Academic research community and industry have constantly researched for big data visualization.

Challenges

Specifically, big data in the integrative environment of whole life cycle is faced with the following problems:

(1) Collection problems of big data. How does big data become small, as in the case of reducing the size of data without loss the value such as data cleaning, removal, etc? How to handle big data efficiently with similar physical effects? How to extract the high value-added concepts, knowledge, and wisdom from big data?

(2) Storage problems of big data. For structured data, it is inefficient in query, statistics, updating of massive data. For unstructured data, such as pictures, video and other files, it is difficult in storing and retrieve. For semi-structured data, it must be transformed into structured data in storage and analysis. Otherwise, it is more difficult that semi-structured data is stored in unstructured data.

(3) Analysis and processing problems of big data. Distributed and parallel computing can provide effective support, and how to effectively use the existing distributed and parallel processing technology to carry out the analysis of big data needs to be studied.

(4) Timeliness problems of big data processing

With the increasing scale of data, the corresponding time of analysis and processing is getting longer and longer. For big data, the timeliness of information processing becomes increasingly demanding. It requires a simple and effective artificial intelligence algorithms and new solving methods to handle big data.

(5) The problems of information security

With the development of technology, a lot of information spread across organizational boundaries accompanying the problems of information security. It puts forward higher requirements in multiple copies, disaster recovery mechanisms and physical security of storage.

Summary

In this paper, big data and relevant research has been summarized and introduced overall at home and abroad during recent years. We describe the concepts of big data and features of 4V, sum up the general processing of big data, and introduce several processing tools and key technologies of big data in detail. Finally, the challenges of big data during the research have been summarized to provide some think for the future.

Acknowledgements

In this paper, the research was sponsored by Youth Science Foundation of JiangXi Province (Project No. 20122BAB211031).

References

[1] Chang-qing Ji, Yu Li, Wen-ming Qiu, et al. Big Data Processing in Cloud Computing Environments,2012[C]. Algorithms and Networks, IEEE, 2012 International Symposium on Pervasive Systems,2012:17-23.

[2] Lab of Web and Mobile Data Management, WAMDM Homepage[EB/OL]. [2013-07-24]. http://idke.ruc.edu.cn/index.htm

[3] MENG Xiao-feng, CI Xiang. Big data management: concepts, techniques and challenges[J]. Journal of Computer Research and Development, 2013, 50(1):146-169. In Chinese.

[4] HDFS Architecture Guide [EB/OL]. [2012-10-02]. http://hadoop.apache.org/docs/hafs/ r0.22.0/ hdfs_design.html

[5] Seref SAGIROGLU and Duygu SINANC. Big Data: A Review[J].IEEE,2013:42-47.

[6] LUO Jun-zhou, JIN Jia-hui, SONG Ai-bo, et al. Cloud computing: architecture and key technologies[J]. Journal on Communications,2011,32(7):3-21.In Chinese.

[7] LI Cheng-hua, ZHANG Xin-fang, JIN Hai, et al. Map Reduce: A new programming model for distributed parallel computing [J].Computer Engineering And Science, 2011,33(3):129-135.

[8] QIN Xiong-pai,WANG Hui-ju,DU Xiao-yong,et al. Big data analysis-Competition and symbiosis of RDBMS and Map Reduce[J].Journal of Software, 2012, 23(1):32-45.In Chinese.

[9] James C. Corbett, Jeffrey Dean, Michael Epstein et al. Spanner: Google’s Globally Distributed Database,2012[C].Proceedings of OSDI 2012:1-14.

[10]Jeff Shute, Mircea Oancea, Stephan Ellner.F1-The Fault-Tolerant Distributed RDBMS Supporting Google's Ad Business,2012[C].SIGMOD,2012,5.