Research Overview of Big Data

Naili LIU[1, a]

[1]Department of Information, Linyi University, Linyi, 276000, China

[a]email: lnl1999@163.com

Keywords: Big Data; Big Data Technology; Data Mining

Abstract. Big data are affecting our work, life, even economy and the development of the society. The Big Data era has already come. This paper first sketches the basic concepts of big data, typical 4”V” characteristics as well as related application fields. Next, the paper summarizes the processing technologies of big data. Finally, the paper presents important directions of research of big data for future, points out challenges in the Big Data Era.

Introduction

In recent years, with the rapid development of new generation of information technology, the rapid development of big data become a hot topic in science and business circles, the rise of big data led to the research upsurge of the big data industry. Big data has become a new focus of attention from all walks of life, "big data era is coming". "Nature"[1] and "Science" [2] publish special issue "Big Data" and "Dealing with Data" to explore the opportunities and challenges which have been brought by big data. IT professionals predict the challenges of data processing. We use "Big Data" to describe this problem. In fact, the term "big data" is not new, Americans proposed big data as early as in the last century 80's [3]. In 2008 September, Science magazine published an article "Big Data: Science in the Petabyte Era", the term "big data" began to be widely used. In 2011 May, the world famous consulting company Mckinsey released the report "Big data: The next frontier for innovation, competition and productivity"[4]. The World Economic Forum in 2012 released the report "Big data, big impact: New possibilities for international development[5]". The mining and the use of big data indicate the arrival of the wave of productivity growth and consumer surplus. Behind every successful re-election campaign and America President Obama has a large data mining support, America government thinks big data is the "new oil" in the future, the research of big data will be enhanced state’s will in American, America Department of Energy, America Department of Defense, America Defense Advanced Research Projects Agency, America Geological Exploration Bureau et al. launched a Big Data plan, which purpose is to obtain knowledge and ability to predict from a large number of complex data. In 2013, the Ministry of Science and Technology China officially launched 863 projects "the key technology and the advanced storage for big data [6] " , started 5 big data projects.

Therefore, the research and development of big data has been widespread concern in the world, which has attracted the attention of experts and scholars. We get valuable knowledge from huge data to solve the problems in work and life, this problem has become an important research topic at home and abroad.

The concept of Big Data

What is big data, so far, there is no generally accepted definition, the following definitions of Big Data are given for different enterprises, research institutions and scientists:

Wikipedia’s definition: Big Data refers to the amount of data involved is huge to not use the current mainstream software tools to capture, management, process and help business decision-making within reasonable time.

Gartner’s definition: Big Data refers to the new processing mode to have a stronger decision-making ability, insight discovery and process optimization capabilities of massive, high

rates of growth and diversification of information assets.

McKinsey's definition: Big Data refers to not acquisition, storage, management and analysis by using traditional database software tools in a certain period of time.

Although the description of Big Data is not the same, but there is a general consensus, that is, the key of "Big Data" is quick obtaining valuable information from wide variety and large data. So, big data includes massive transaction data, massive data and massive processing data.

The Characteristics of Big Data

We usually use 4 "V" to summarize the characteristics of big data in Big Data research field, that is, Volume, Variety, Value, Velocity[7-10].

Volume. From TB (1024GB=1TB) level up to PB (1024TB=1PB) level, EB (1024PB=1EB) and ZB (1024EB=1ZB) level. Many websites produce dozens of PB data every day. The data is stored in the form of static on the hard disk, rarely updated, long storage time, can be reused, large quantities of data is not easy to move and backup.

Variety. With the increase of sensor types and the popular of intelligent devices and social network, data types has become more complex, which include traditional relational data types and semi-structured and unstructured data from webpage, audio and video, pictures, location information, these types of data put forward high capacity requirements in data processing.

Velocity. This is the most significant feature of big data to distinguish from the traditional data mining. According to the IDC's "Digital Universe" report, global data will reach 35.2ZB in 2020. In the face of such vast amounts of data, the efficiency of data processing is the life of enterprise. How to solve the problem is very important.

Value. This is the most important characteristics of big data, the amount of data increases exponentially. At the same time, the useful information behind the massive data hasn't corresponding increasing, we get useful information more difficulty. The value of density is proportional to the size of the total data. In case of video data, there has useful data only one or two seconds in continuous monitoring. How to mine useful data from massive data by using strong algorithm is the problem to be solved under the background of big data.

The Applications of Big Data

In order to enhance enterprise's competition and achieve greater efficiency, enterprises need analysis the historical data and process data mining, find out and predict the future trend of development to make right decisions. Data processing is suitable for many kinds of application scenarios, such as networking, Internet, security and public services [11-15], the following examples are introduced.

(1) Electronic Commerce: E-commerce increasingly fierce competition in recent years, there exist a large number of purchase history, product reviews, product Webpage visits and dwell time, these data is very large, we can analyze the user's consumption behavior and recommend related products for customers from these data, so we can improve the quality of the number of customers. For example, Taobao's "Data Cube" mainly provides industry data analysis, store data analysis, which include the ranking list of brand, store and product and the characteristic of purchase of crowd (age, gender, purchase time, region and so on). It can also provide real-time operation data, real-time transaction shops and real time transaction industry operations, which is the right-hand for electronic commerce.

(2) Social Network: Along with people rely WeChat and microblog more and more, we need analyze the large quantities of data to help people refer friend or theme and enhance user experience.

(3) Financial: Big Data also plays an important role in the financial industry. For example, TaoBao launched the "Hua Bei" according to the operation of every business, which can analyze repayment ability of every business and provide different amount of the loan for each business.

(4) Search: Yahoo's special advertising analysis system improves the effect of advertising and

the count of hit by processing relational data. Google predicted winter flu according to the previous frequent query words successfully.

(5) Medical: Jobs was treated by using Big Data and achieved several years' life.

Big data not only is applied to the field of energy, as well as mobile, data analysis, image processing, infrastructure management and other fields. With the people's awareness of the value of data, there will be more fields to mine the value of data in order to support decision-making and discover new insights.

The Processing Technology of Big Data

The key technologies of data processing include big data acquisition, big data preprocessing, big data storage and management, big data analysis and mining and big data visualization and application[16], big data processing flow is shown in Figure 1.

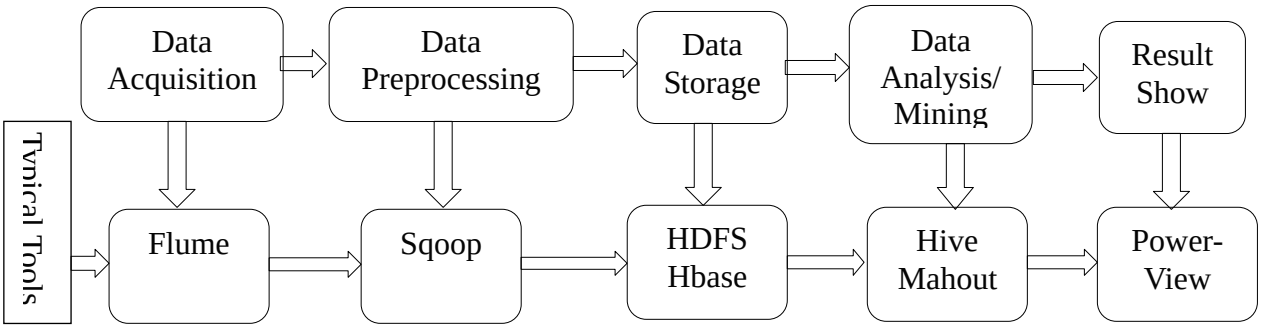

Figure 1 Big data processing flow

(1)Data Acquisition

Data acquisition is the most basic step in the process of data, the data is usually obtained by RFID[17], sensor, network and mobile Internet, structured, semi-structured and unstructured data.

(2)Data Preprocessing

Data processing is mainly to complete operations of data extraction, cleaning. 1) extraction: Because data has different structures and types, data extraction is to convert these complex data into single structure or processing easily structure in order to achieve the purpose of rapid processing. 2) cleaning: we need do cleaning and denoising on accessed data in order to ensure data quality.

(3)Data Storage and Management

How to develop the technology[18] of big data's storage and management which are designed high performance in processing data is a key problem, which is widely used at present in big data storage mechanism based on Hadoop[19-20], which acquired data from different terminal, data has structured, semi-structured and unstructured characteristics. For structured data, although there exist variety of database types, but we still use relational data repository for processing data; Hadoop framework provides a good solution for semi-structured and unstructured data.

Hadoop distributed file system HDFS[21-23] is a big data file system storage reliability in large clusters. Hive[24] and HBase[25] based on HFDS can primely support big data storage. Specifically, Hive can quickly achieve MapReduce[26-27] statistics by SQL statements, which is adapt to statistical analysis for the data warehouse. HBase is a non-relational database based on the distributed column storage, it's query efficiency is very high, which is mainly used to query and display results; Hive is a distributed relational data warehouse, which is mainly used to handle large amounts of data in parallel. We integrate Hive with HBase to process bit data, which can reduce the development process and improve the efficiency of development.

(4)Data Analysis and Mining

According to the huge amount of data, we need adapt parallel data mining algorithms and strategies on cloud computing[28-29] environment, the representative is MapReduce data analysis technology, data processing of MapReduce is shown in Figure 2.

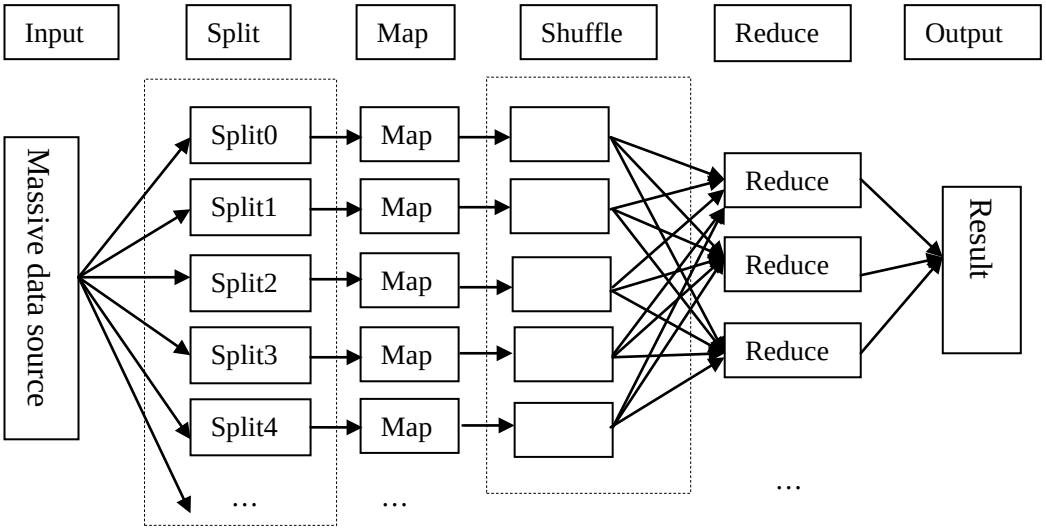

Figure 2 Data processing of MapReduce

As can be seen from the figure, MapReduce is mainly composed of two parts of Map and Reduce, the core idea of MapReduce is to "divide-and-rule". Map mainly processes data of a set of input records, processing way is based on the key/value. Map produce many (key,1) records . Reduce receives middle results and merges the same value, finally forms the set of the final value. Reduce produce many (key, count) records. Many scientists improved map algorithms and reduce algorithms in order to process data efficiently.

(5)Data Show

Big data technology can mine information and knowledge which are hidden in massive data, and provide the basis for social and economic activities of human beings, therefore which improves the operation efficiency of various fields and greatly improves the intensive degree of whole social economy. In our country, big data will be application in the following three fields: business intelligence, decision making, public service. For example: business intelligence technology, government decision-making technology, telecommunications data processing and mining technology, information processing technology and data mining, analysis of meteorological information technology, environmental monitoring, police cloud application system, large scale gene sequence analysis technology, Web information mining technology, parallel processing technology of multimedia, film and television production rendering technology, cloud computing and massive data processing technology.

The Challenges and Research of Big Data

Although the big data era has arrived, the industry also found the value of big data, but the data is still in the initial stage. With the deepening of study, big data faces problems more and more, how to make big data in favor of the whole society development requires comprehensive study on big data, the following is research and challenges of big data.

(1) The exponential growth of data quantity need have efficient storage. Many websites produce dozens of PB data every day, these data have many structures. So solving the storage of big data is a key problem, which is a research hot in big data field.

(2) Processing velocity. With the increasing scale of data, the time of analysis and processing data need more and more. But we need quickly process data in order to get results. Traditional mining algorithms can’t solve the problem. It is a challenge of big data. It is the research hot. So many scientists try to find new method or algorithms to solve the problem.

(3) The variety of data types is the challenge of data mining. From database perspective, the efficiency and scalability of mining algorithm is the key to the realization of data mining, existing algorithms are suitable for memory resident datasets. With the increasing scale of data, the efficiency of algorithm has gradually become the bottleneck of data analysis. This is the research

hot in big data field.

(4) The predictive role of big data has become increasingly prominent. Big data's role is self-evident, many experts summarized the power of big data, big data can change economic and organize. Foresee is the most direct benefit from Big Data. For example, Google predicted winter flu according to the analysis of keyword search. We predicted the value of shares will rise and fall according to the analysis of history records. Many industries make a reasonable decision-making by using big data. We can improve the prediction of unknown reliability and precision by using big data.

Conclusion

The Big Data Era has come. Big data has been involved in various fields of life, More and more enterprises pay attention to the value of big data. This paper is mainly on the concept and the characteristics of big data and the basis of analysis of large data processing technology, Finally, the paper presents important directions of research and development of big data for future, points out challenges in the Big Data Era. The development of big data is still in its initial stage, people need constantly open up a lot of space, how to efficiently process large data and reasonable use big data still need continue to explore.

Acknowledgement

In this paper, the research was sponsored by the Major Projects of Innovation and Transformation of Achievements of Shandong Province (Project No. 2014ZZCX02702).

References

[1] GOLDSTON D.Big data:data wrangling[J/OL].Nature,2008,455:15. http://www.nature.com/

nature/index.html.

[2] REICHMAN O J, MATTHEW B.MARK P P.et al. Challenges and opportunities of open data in ecology[J].Science,2011,311(6018):703-705.

[3] XU Zi-pei. Big Data[M]. Guangxi Normal University Press, 2012: 57.

[4] MANYIK A J,CHUI M,BROWN B,et al.Big data:The next frontier for innovation, competition, and productivity[R/OL].Las Vegas:The McKinsey Global Institute.

[5] World Economic Forum. Big data,big impact:New possibilities for international development.

[6] Tian Jin University.863 Projects "Advanced memory results and key Technology of Big data-oriented.

[7] CNII. Four typical characteristics of large data.

[8] YAN Xiao-feng.ZHANG De-xin.Big data research[J].Computer Technology and Development.2013,23(4):168-172.

[9] CHEN Ru-ming. Challenges, values and countermeasures of the era of big data[J].Mobile Communication,2012(17):14-15.

[10] TAO Xue-jiao, HU Xiao-feng, LIU Yang. Overview of Big Data Research[J]. Journal of System Simulation,2013.8(25):142-146.

[11] Tsourakakis CE. Fast counting of triangles in large real networks without counting: Algorithms and Laws, 2008. 608−617.

[12] Chen Y, Alspaugh S, Katz R. Interactive analytical processing in big data systems: A cross-industry study of MapReduce workloads. Proc. of the VLDB Endowment,

2012,5(12):1802−1813.

[13] Zhou MQ, Zhang R, Xie W, Qian WN, Zhou AY. Security and privacy in cloud computing: A survey. IEEE, 2010. 105−112.

[14] Feblowitz J. Analytics in oil and gas: The big deal about big data. In: Proc. of the SPE Digital Energy Conf. 2013.

[15] Yu H, Wang D. Research and implementation of massive health care data management and analysis based on hadoop. IEEE, 2012. 514−517.

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

[17] LM Ni,YLIU,YC Lau,et al.LANDMARC:Indoor location sensing using active RFID[J].Wireless Networks,2004,10(6):701-710.

[18] CHANG F,DEAN J,CHEMAWAT S,et al.Big Table:A distributed storage system for structured data[J].ACM Transactions on Computer Systems,2008,26(2):4.

[19]YANG Chen-zhu.The research of data mining based on HADOOP[D].Chongqing:Chongqing University,2010.

[20] Bu Y Y, Howe B, Balazinska M. Ernst MD HaLoop: Efficient iterative data processing on large clusters [J]. (S0926-5807), 2010,(1-2).

[21]TheApacheSoftwareFoundation.HDFSArchitecture.http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hadf/HdfsDesign.html.

[22]NoSQL Databases. NoSQL Definition.http://nosql-database.org.

[23]HUANG Xiao-yun.Research of cloud storage service system based on HDFS[D].Dalian:Dalian Maritime University,2010.

[24]The Apache Software Foundation.Hive.http://hive.apache.org/.

[25] The Apache Software Foundation.Hbase.http://hbase.apache.org/.

[26]DEAN J, CHEMAWAT S.Map Reduce: Simplified data processing on large clusters[J].Communications of the ACM51.2008(1):107-113.

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

[28] 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.

[29] CHEN Kang,ZHENG Wei-min.Cloud computing:System instances and current research[J].Journal of Software,2009,20(5)：1337-1348.