Research on the Flows of Big Data Processing Technology

L.P. ZHU, B. HU, J. DING & A.H. ZHOU
State Grid Smart Grid Research Institute, Beijing, China

ABSTRACT: Comparing with the conventional data, big data has characteristics of Volume, Variety, Velocity and Value. The widespread use of big data poses a challenge to data processing, inducing more concerns and studies in big data processing technology. This paper researches on the basic flows of big data processing technology, which are divided into data acquisition, data extraction and integration, data analysis, data interpretation. At last, we conduct that combine traditional way of information organization with big data becomes a new research area.

KEYWORD: Big data; data process; data extraction

1 BACKGROUND

The data exploration in the 21st century has brought about the “Big Data era”. Big data[1] has characteristics of Volume, Variety, Velocity and Value, comparing with the conventional data. Volume and Velocity have been studied and discussed, Variety and Value are constantly appear problems in current data processing. The issue would raises and inevitable as the Smart City, Smart Earth become a reality.

The Big Data era is the inevitable result of the development of digital device computing power and deployment index. Hence, the provenance of big data becomes the key to solve the big data research problems. Big Data originates from the scale effect, which pose a great challenge to data storage, regulation and analysis. Thus the storage and operational program should take into account under big data scale effect. Traditional technique that vertical development rely on single-device processing capacity could no longer meet the demand of big data storage and process. Head companies like Google have resolved the issue by transverse distributed storage, distributed process and distributed data analysis[2].

The big data is regard as a huge business opportunity among technological enterprises, including IBM, Google, Amazon and Microsoft as well as start-ups. For big data analysis, Hadoop[3] is preferred in commercial services program for its lower costs, high scalability and flexibility. Many well-known companies provide their own commercial big data solutions based on the Hadoop, and Oracle, IBM, Microsoft are the main supporters for this technique.

From the general data process, massive data storage and massive data computation can be seen as the mainly techniques under the big data. Data resources are abundant and manifold under the big data, thus the efficiency and feasibility are valued for data processing. The sources of traditional data acquisition are single and relatively small in data storage, regulation and analysis. Most of them are handled by relational database and parallel data warehouse. However, according to the CAP theory[4], consistency and fault-tolerance are not ensured by traditional data processing. Compared with big data processing, which need the data-centric processing, the traditional data processing’s processor-centric processing is not adaptable under the big data. Thus the widespread of big data pose a challenge to data processing, inducing more concerns and studies in big data issues. This paper focuses on the basic flows of big data processing.

2 THE BIG DATA PROCESSING FLOWS

The basic flows are consistent in big data processing despite its various sources, application requirements and data types. Big data processing differs from the traditional data processing in that it deals with large, unstructured data by MapReduce[5] and other methods. The processing flows are shown in Figure 1:

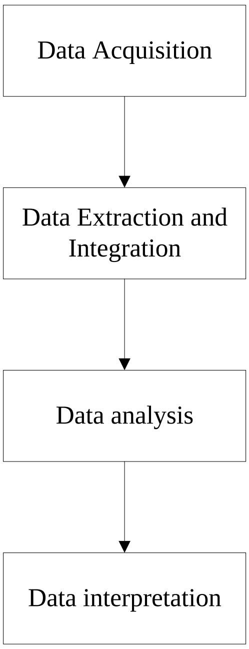

Figure 1. The big data processing flows

Big data processing can be defined as under the aid of suitable tools, extraction and integration on a wide range of heterogeneous data, results unified storage in accordance with certain criteria. Extraction useful knowledge from appropriate analysis data and presentation in a proper way to the end-user. Specifically, divided into data acquisition, data extraction and integration, data analysis, data interpretation.

Novel methods in big data acquisition including System log collection method, Network data collection method and other data collection method.

System log collection method: Many Internet companies have their own tools for massive data collection, which used for system log collection. Such as Hadoop's Chukwa, Cloudera's Flume, Facebook's Scribe[6], all of them use a distributed structure to meet the acquisition and transmission demands of hundreds of MB of log data per second.

Network data collection method: collection from unstructured data. The Network data collection is getting data by Web Crawler[7], or through a public web site API ,etc. This method enables the extraction of unstructured data from web pages, then in a structured way to store the data, which as the unified local data files. It supports the collection of images, audios, videos and accessories, which can automatically associate with the text. Besides, bandwidth management technology of DPI or DFI resolve the collection of network traffic[8].

Other data collection method: for data of high confidentiality requirements such as operating data in enterprise, disciplinary research data and so forth, the data collection could be achieved by applying specific system interface through cooperation with companies or research institutions.

Data extraction and integration is not a new technology, more mature research has been studied in traditional database field. With the development of the new data source, the techniques of data integration are constantly evolving. From the data integration model, the Data extraction and integration methods can be classified as following: materialization or ETL engine, federation engine or mediator, stream engine and search engine.

Data analysis is the core of big data processing, in terms of the value of big data generated in the process of data analysis. The extraction and integration from heterogeneous data sources constituted the raw data of data analysis. For different application demands, these data could be analyzed in whole or in part. Traditional analysis technology like data mining, machine learning, statistical analysis, and so on, which should be adjusted for the big data era. Big data analysis has been applied to various of fields, especially in recommended system, business intelligence, decision support.

Data interpretation is of the most concerns to users. If the results are correct but not interpret appropriately, they might confuse the users, even misleading. Traditional methods of data interpretation is text output or display the results directly on computer. It is a good way in the face of a small amount of data. However, consider the large amount and complexity of data analysis results, the method is no longer feasible under big data era. To enhance the data interpretation ability, two following aspects can take into account. 1)Introduce the visualization technology such as tag cloud, history flow, spatial information flow. 2)Allow the user to understand and participate in the data analysis to a certain extent. Human-computer interaction techniques use the interactive analysis process to guide the user, afford the user to obtain the result as well as better understand in the origin of analysis results. The other method is Data origin technology, which helps the user by tracing the data analysis process.

3 THE PROSPECT OF BIG DATA PROCESSING

The extensive use of big data processing attest for its bright prospects. Open source software provides more opportunities for the big data market .Revolutionary methods would appear in big data analysis. Just like the development of computer and the Internet, big data could be a new technological revolution. And breakthroughs in algorithms and theories become very possibility. Further, the development of big data is inseparable with cloud computing, which supports with the Infrastructure environment as well as efficient mode data services. While the big data gives the cloud computing new business value, so that a .perfect combination for them is a certain. The same as the

Internet of things, Mobile Internet, necessitating the importance of big data processing.

4 CONCLUSIONS

The big data is at a start stage, in which rely on the Hadoop architecture. However, the big data differs from Hadoop in that it has irreplaceable advantages from cloud storage and computing, and the traditional relational database technology. Hence, combine traditional way of information organization with big data becomes a new research topic.

REFERENCES

[1] Sagiroglu S, Sinanc D. Big data: A review. Collaboration Technologies and Systems (CTS), 2013 International Conference on, 2013. 2013: 42-47.

[2] Sivaraman E, Manickachezian R. High Performance and Fault Tolerant Distributed File System for Big Data Storage and Processing Using Hadoop. Intelligent Computing Applications (ICICA), 2014 International Conference on, 2014. 2014: 32-36.

[3] Kala Karun A, Chitharanjan K. A review on hadoop -- HDFS infrastructure extensions. Information & Communication Technologies (ICT), 2013 IEEE Conference on, 2013. 2013: 132-137.

[4] Ramakrishnan R. CAP and Cloud Data Management. Computer. 2012, 45(2): 43-49.

[5] Pandey S, Tokekar V. Prominence of MapReduce in Big Data Processing. Communication Systems and Network Technologies (CSNT), 2014 Fourth International Conference on, 2014. 2014: 555-560.

[6] Jose A S, Binu A. Automatic Detection and Rectification of DNS Reflection Amplification Attacks with Hadoop MapReduce and Chukwa. Advances in Computing and Communications (ICACC), 2014 Fourth International Conference on, 2014. 2014: 195-198.

[7] Gupta P, Johari K. Implementation of Web Crawler. Emerging Trends in Engineering and Technology (ICETET), 2009 2nd International Conference on, 2009. 2009: 838-843.

[8]Kim N, Choi G, Choi J. A Scalable Carrier-Grade DPI System Architecture Using Synchronization of Flow Information. Selected Areas in Communications, IEEE Journal on. 2014, 32(10): 1834-1848.