Journal of Applied Science and Engineering, Vol. 18, No. 2, pp. 135–142 (2015) *DOI: 10.6180/jase.2015.18.2.05*

Research on IT Architecture of Heterogeneous Big Data

Yun Liu*, Qi Wang and Hai-Qiang Chen

School of Communication and Information Engineering,
Key Laboratory of Communication and Information Systems,
Beijing Municipal Commission of Education, Beijing JiaoTong University,
Beijing 100044, P.R. China

Abstract

The amount of data has grown exponentially in industry and internet, and we are facing a significant problem of information explosion. The challenge is not only to store and manage the vast volume of data ("big data"), but also to analyze and extract meaningful value from it. This paper analyzed the characters of the future network and the features of data generated by it. Furthermore, we studied key issues and challenges of the multiple heterogeneous data IT infrastructure. The main focuses of this paper are on three aspects of distributed storage, cloud computing, and data fusion. There are several techniques to address these issues. At last, we proposed an architectural solution and key technology program which can be adapted to developing more application functionality and meet future demand for Internet services and new business for big data processing requirements.

***Key Words*:** Cloud Computing, Hadoop, Data Mining, Distributed Storage

1. Introduction

In recent years, the volume of data in Internet has been increasing significantly with the rapid development of Internet. Such kind of data called big data not only has mass and high-speed characteristics, but also the diversity and variability. Big data have many sources including online activities (social networking, social media), telecommunications (mobile computing, call statistics), scientific activities (simulations, experiments, environmental sensors), and the collation of traditional sources (forms, surveys). Large data contains the value, but there are still many difficulties and challenges in the use of big data technologies. The larger the scale of data, the more difficult the processing and storage is. Firstly, data storage is required to achieve the requirements of low cost, high capacity, high reliability, great scalability and adaption of data processing, so before to storage data, preprocessing of data cleaning, classification, filtering and indexing is necessary. Secondly, big data has the characters of diversity and heterogeneity, and diverse heterogeneous data includes not only structured data but also unstructured data. According to the statistics, 90% data through the internet is unstructured data. However, processing unstructured data is much more complicated than structured data. Thirdly, the amount of big data is so huge that the computer memory has no more space and a single computer cannot process it efficiently and in addition to the requirements of real-time processing and the processing of complicated structured data, the complication of big data processing will be considerable.

This paper studies the characteristics of the Internet in the future and the challenges and technical difficulties in the conditions of big data. Then we analyzed some existing technologies and methods to deal with big data and also presented what challenges we were going to face in

*Corresponding author. E-mail: liuyun@bjtu.edu.cn

the future. In order to address these challenges, we proposed an architectural solution and key technology program which can be adapted to developing more application functionality and meet future demand for Internet services and new business for big data processing requirements.

The remainder of this paper is organized as follows: section 2 gives an overview of major approaches for big data processing. Section 3 describes our multiple heterogeneous data IT Infrastructure, including four parts. A detailed discussion of the proposed infrastructure is presented in section 4. Finally, section 5 provides some concluding remarks.

2. Related Work

Google is the advocator and promoter of big data processing, and its famous modules of Bigtable [1], GFS, MapReduce cause the many commercial systems and open source tools birth. Bigtable is a distributed structured data storage system, which is designed to handle massive amounts of data: usually the PB level data which is distributed in the thousands of ordinary servers. GFS [2] is a scalable distributed file system. It is a dedicated file system designed to store massive amounts of search data. GFS is used for large-scale, distributed, large amounts of data access applications and runs on inexpensive commodity hardware, but can provide fault tolerance, and higher overall performance of the services provided to a large number of users. MapReduce [3] is an algorithm model and associated implementation for processing and generating big data sets. Yahoo's outstanding contribution to big data processing is that they created the open source framework Hadoop [4,5] based on the MapReduce model and promoted HBase, Hive, and other peripheral technologies, making Hadoop the base of most current big data processing products. In our country, Baidu, Tencent, Alibaba and other Internet companies to provide services for the network has been faced with large data processing needs, but has also been developing and using large data processing technologies. However, solutions and frameworks purely providing data processing services are very few. So we are far behind foreign at both techniques and services.

Multiple heterogeneous big data have the characteristics of multiple dimensions, multiple sources, structured + semi-structured + unstructured heterogeneous data and the huge amount of data, thus the traditional relational database has been very difficult to store such data. In addition, the demand for the processing of such data is not limited to the mode supported by SQL, but shows the features of the connection and interweaving between the data on the Internet. There are a lot of storage methods for heterogeneous data [6], such as serial method, graph method, tree method, file system method, database field method and object method. Due to the difference between the services of application layer, the characteristics of multi-source heterogeneous data also are different with different services, so our data storage and management must use corresponding methods for differing services, for example, Bigtable [1], the key value of NoSQL and so on.

Data fusion is a research hotspot and has several advantages [7,8]. These advantages mainly involved enhancements in data authenticity or availability. Examples include improved detection, confidence, reliability, as well as reduction in data ambiguity while extending spatial and temporal coverage. There are a lot of conceptualizations of data fusion. Among them, JDL model [9] is the most common and popular. The JDL is originated from the military domain and it considers the fusion process in four increasing level of abstraction, namely object, situation, impact, and process refinement. However, it has many shortcomings, like the too much limitation which has been the subject of several extension researches [10,11] attempting to alleviate them. Dasarathy's framework [12] views the fusion system, from a software engineering perspective, as a data flow characterized by input/output as well as functionalities. One of the most recent fusion frameworks is proposed by Kokar et al. [13].

Web services is an online application service published by service providers in order to solve some specific business requirements. The target of web service is

transferring software to subscription services through Internet and its essence is realizing the online data conversion by XML [14]. Fan et al. [15] proposed enterprise application integration based on web service and Min-Hsiung Hung et al. [16,17] presented a web services based e-diagnostics framework for semiconductor manufacturing industry. Therefore, web service is a significant way to solve heterogeneous data exchange problem in big data. However, the existing frameworks are mainly aimed at special businesses, which hardly to extensive apply in other industries, and there is lack of overall architecture including heterogeneous data storage and exchange module.

3. Multiple Heterogeneous Data IT Infrastructure

Multiple heterogeneous big data have the characteristics of multiple dimensions, multiple sources, structured + semi-structured + unstructured heterogeneous data and the huge amount of data. These features of big data bring us three aspects of challenges at the heterogeneous big data tiered storage and storage management, the integration of heterogeneous data, the off-line and realtime computing architecture and the efficient transmission mode of big data. This technical structure consists of underlying file system, structured data storage system, semi-structured data and unstructured data storage system, the distributed storage of data, the unified data access interface, data indexing and positioning, processing task decomposition and scheduling management, e distributed execution of processing tasks, e service interface of the system and the secondary development interface and system manageability and security and completely defined a data storage and processing platform meeting the specific needs. In the following paper, we presented the four aspects of our big data framework respectively.

3.1 Multiple Heterogeneous Data Tiered Storage and Distributed Storage

In this paper, we used the multiple heterogeneous data tiered storage as our storage method. Big data needs mass storage, so we cannot storage and manage them like the traditional solutions treating them as equally important. However, hierarchical storage is using different storage methods corresponding to the characteristics and values of different data and utilizes the hierarchical storage management software to achieve automatic sorting and automatic storage, which greatly increases the effectiveness and speed of data storage and meets the storage needs of different kinds of data. In our system, there are two kinds of data which are static data and dynamic data respectively. The framework of hierarchical storage is shown in Figure 1.

In order to achieve the storage of big data, especially the heterogeneous data storage of the structured data, semi-structured data and unstructured data, we superimpose a relational database and a distributed non-relational database on a distributed file system to store huge amounts of data which are managed by the master + multi-slave physical structure. Master nodes and multi-node slave nodes are connected via Internet. Applications get access to data through the master host, and each storage node in the network is a separated database without sharing with other storage nodes. Data is exchanged between master node and multi-node storage nodes. Slave nodes are connected via the Internet and they complete the same task, so it is a server system from

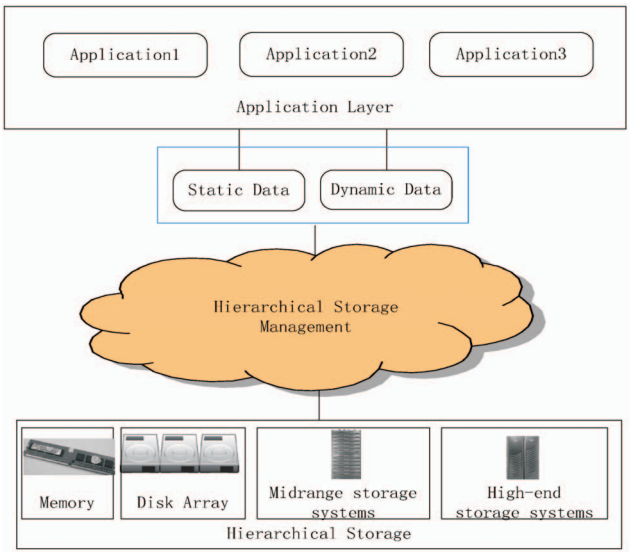

Figure 1. Hierarchical storage diagram.

the user's prospective. In this structure, each node only have right to get access to the local resources, including memory, storage and etc. It is a completely non-sharing structure with great scalability, and theoretically unlimited period of expansion, the current technology can achieve 512 nodes interconnected, thousands of CPU. Each node can run its own database, operating system, but each node cannot access the memory of other nodes, and information exchange between nodes is implemented via the Internet node. This process is known as data reallocation. In this basic data storage mode, according to the specific business needs and the upper data features, the designed specific big data storage architecture is shown as Figure 2.

3.2 Heterogeneous Data Fusion

Heterogeneous data fusion refers to using appropriate processing methods to process the heterogeneous data and producing complete accurate timely and effective integrated information. For data fusion, multi-sensor system is the hardware basic, multi-source heterogeneous information is its processing objects and optimism and integrated processing is its core. Sensor is the source of heterogeneous data, but it is not necessarily physical form. In other words, all the data sources and manual data source can be called sensor.

In our paper, we adopted fusion framework which is based on category theory and is claimed to be sufficiently general to capture all kinds of fusion, including pixel-based fusion, feature fusion and decision fusion. Fusion of the data can be divided into three levels of pixel-based fusion, feature-level fusion and decision fusion. The pixel-based fusion also has a function similar to the data denoising and the cleaning process in addition to the significance of fusion. Feature-level fusion classifies data based on the features and data attributes. Decision fusion uses data to implement the trend assessment and the macro perspective service processing at the highest level. According to the specific services and the specific characteristics of the data, the corresponding fusion algorithms and computational structures are used to treat on different data. Utilizing data fusion technology, we can remove redundant information to reduce the amount of data increase the access efficiency. Moreover, we can extract useful information from a large number of heterogeneous data, providing services and support for the business of the upper services. Figure 3 shows the Functional model of heterogeneous data fusion.

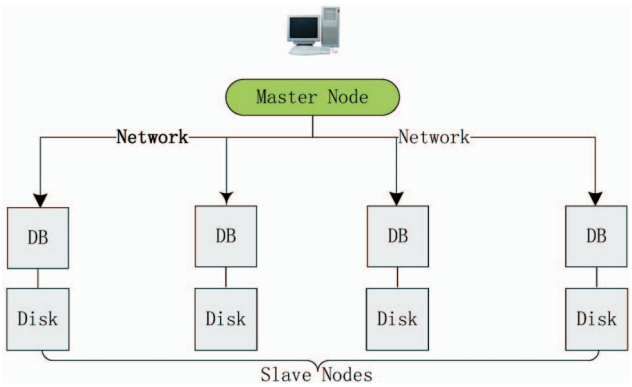

Figure 2. Distributed file system structure.

3.3 Off-line and Online Computation Framework

There are off-line and online computing demands in a large data environment. Off-line computation is more relaxed in the computing time requirement. However, under the conditions of mass data, there are still computing time limit and the balance between recalculation and the incremental calculation becomes an inevitable problem with the increasing of data. Online calculation has a higher requirement on the computing time, which is a quit hard problem for the mass data and needs to a variety of methods to guarantee the calculation real-time.

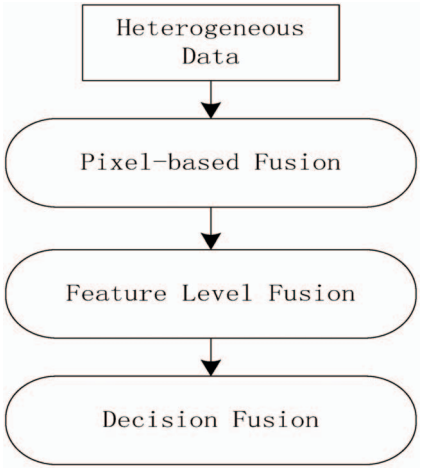

Figure 3. Heterogeneous data fusion diagram.

In this paper, we classed the data into static data and dynamic data. Static data are the historical read-only data, but dynamic data are the read and write data including intermediate result. For static data, when we implement off-line computation, we must design a reasonable storage structure and create effective index to improve the processing efficiency according to the demands of specific services. Hadoop [4] use the idea of MapReduce [3]. Hadoop is a Java implementation of Google's MapReduce. It slices the data to deal with large amount of off-line data and then assign computing tasks to more than one computer, which will greatly improve the speed and efficiency of computation. Online computation needs to use reasonable caching mechanism to solve the massive data processing problem. MapReduce is a simplified distributed programming model that allows the program automatically distributed to ordinary machine consisting of a large cluster of concurrent and be executed. A MapReduce job will usually slice the input data set into separate blocks of data and deal with in a completely parallel way by the Map task. The framework will first sort the output of Map and then distribute the output to Reduce task. Usually the input and output operations will be stored in the file system. The framework takes the task scheduling and monitoring, and re-run the failed tasks. Hadoop MapReduce processing flow is shown in Figure 4. For dynamic data, we will combine data structure designing, index designing and caching mechanism designing together to complete the data processing. Storm is a common stream computing framework which has been widely used for real-time log processing, real-time statistics, real-time risk control. Storm also is used to process the data preliminarily and it can store the data in a distributed database like HBase in order to facilitate the subsequent queries. Storm is a scalable, low-latency, reliable and fault- tolerant distributed computing platform. In addition, we will employ the distributed computing mode for data computation and processing and the assignment, scheming and management of the computing tasks. The off- line and online computation framework is shown in Figure 5.

3.4 Efficient Interactive Transmission Mode of Big Data

Storage and processing of big data necessarily involves the network-based distributed storage and computation, the collection of multi-source data, the remote data

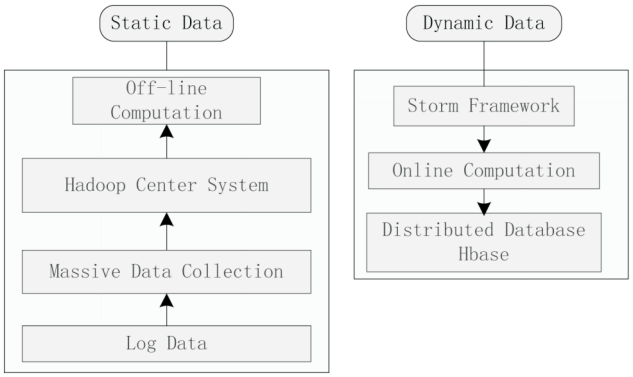

Figure 5. Off-line and online computation.

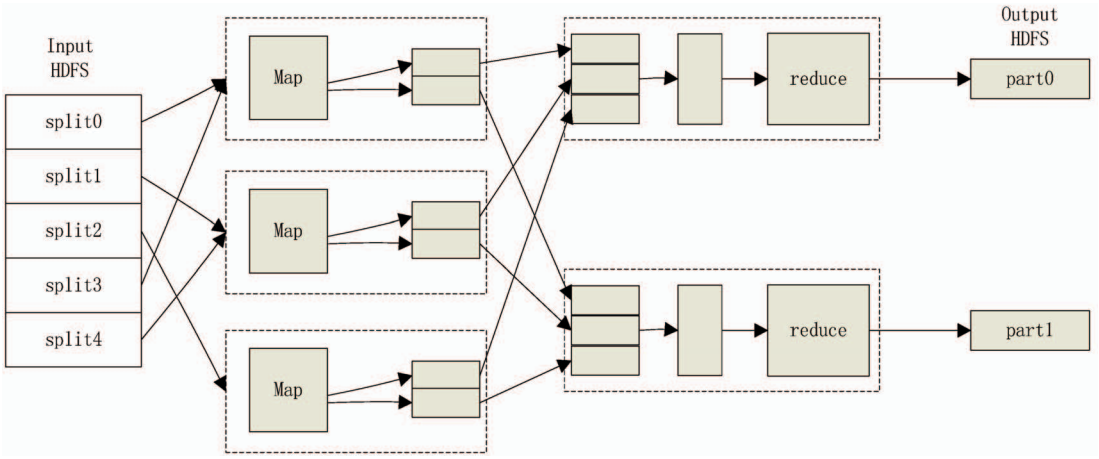

Figure 4. Hadoop MapReduce Processing flow.

management on the basis of network and data sharing and exchange. So data exchange and interactive data service is an indispensable supporting technology.

In this paper, we designed the data exchange framework according to two modes of data service and heterogeneous data exchange. Data service is established data access function for large-scale predefined data transferring and migration and heterogeneous data exchange refers to the data exchange between the different modules within a framework and clients. Data service mode uses connection-oriented channel exchange. However, Heterogeneous data uses connectionless datagram exchange and defines the structure of datagram by XML or JSON.

The framework defines three types of service interfaces for the data interactive interface: the interfaces between the modules within the system, which include two modes of service and API; the external interfaces of the system functions, which appear in the form of API and are used for the developing and data access of the service applications; interactive service interfaces, which have two modes of service and REST API. System interfaces can support each other to form the overall function. The external interfaces in the form of API can be used for the developing the service applications and REST API can be used for extending the service applications of the system and create a more convenient remote data access. In this paper, large-scale data, which is predefined, is transferred through the interface between the system modules; on the contrary, heterogeneous data can be exchanged through the heterogeneous data exchange mode via the external interface. In addition, the datagram structure of heterogeneous data is defined by XML or JSON. Data exchange modes are shown in Figure 6.

4. Discussion

We tested our framework in three aspects of speed-up ratio, efficiency and expandability in Hadoop cluster circumstance using Java programing language to realize the data fusion algorithm based on MapReduce technology. In order to test our framework, we chose two random datasets and two real datasets including Netflix and MovieLens. The results of our framework are proposed in Tables 1 and 2. We tested our framework on three environments which are single-core server, dual-core server and quad-core server. The results of running time are proposed in Table 1 with unit of one thousand seconds.

The speed-up ratio and efficiency of our framework are proposed in Table 2 which shows our framework can actually improve the performance and efficiency of data fusion. Additionally, our framework is suitable for multi-core processing with good expandability. The experiment result shows the parallelization fusion method in our framework has better performance than traditional clustering fusion. Therefore, our framework based on MapReduce is more suitable for large-scale data processing.

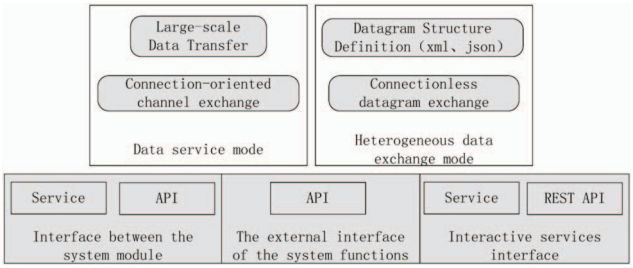

Figure 6. Data exchange module structure.

Table 1. Running time (ks) of four datasets

Dataset	Single-core (traditional method)	Dual-core	Quad-core
RandomData_1	8.64	4.57	2.25
RandomData_2	6.52	3.38	1.68
Netflix	10.86	5.77	2.80
MovieLens	9.44	4.97	2.44

Table 2. Speed-up ratio and efficiency of four datasets

Dataset	Speed-up ratio		Efficiency	
	Dual-core	Quad-core	Dual-core	Quad-core
RandomData_1	1.89	3.84	0.95	0.96
RandomData_2	1.92	3.88	0.96	0.97
Netflix	1.88	3.87	0.94	0.97
MovieLens	1.90	3.87	0.95	0.97

5. Conclusions

In this paper, we described the framework of multiple heterogeneous big data from four aspects: (1) storage management; (2) data fusion; (3) online and offline computation; (4) efficient interactive transmission mode of big data.

Based on these technologies, we proposed a framework of multiple heterogeneous data which can be adapted to developing more application functionality and meet future demand for Internet services and new business for big data processing requirements. Our framework can solve the problems of storage limitations and suit to variety businesses.

Acknowledgements

This work has been supported by the National Natural Science Foundation of China under Grant 61172072, 61271308, the Beijing Natural Science Foundation under Grant 4112045, the Research Fund for the Doctoral Program of Higher Education of China under Grant W11C100030, the Beijing Science and Technology Program under Grant Z121100000312024.

References

[1] Chang, F., Dean, J., Ghemawat, S., et al., "Bigtable: A Distributed Storage System for Structured Data," *ACM Transactions on Computer Systems*, Vol. 26, Issue 2, Article No. 4 (2008). doi: 10.1145/1365815.1365816

[2] Ghemawat, S., Gobioff, H. and Leung, S. T., "The Google File System," *ACM SIGOPS Operating Systems Review. ACM*, Vol. 37, Issue 5, pp. 29–43 (2003). doi: 10.1145/1165389.945450

[3] Dean, J. and Ghemawat, S., "MapReduce: Simplified Data Processing on Large Clusters," *Communications of the ACM*, Vol. 51, Issue 1, pp. 107–113 (2008). doi: 10.1145/1629175.1629198

[4] White, T., Hadoop: The Definitive Guide. 1st, O'Reilly Media (2009).

[5] Information on http://research.yahoo.com/files/ycsb.pdf.

[6] Information on https://www.google.com/patents/US20030051097.

[7] Hall, D. L. and Llinas, J., "An Introduction to Multisensor Fusion," *Proceedings of the IEEE*, Vol. 85, Issue 1, pp. 6–23 (1997). doi: 10.1109/5.554205

[8] Walts, E. L., Data Fusion for C3I: a Tutorial, in: Command, Control, Communications Intelligence (C3I) Handbook, EW Communications Inc., Palo Alto, CA, pp. 217–226 (1986).

[9] White, F. E., Data Fusion Lexicon, Joint Directors of Laboratories, Technical Panel for C3, Data Fusion Sub-Panel, Naval Ocean Systems Center, San Diego (1991).

[10] Steinberg, A. N., Bowman, C. L. and White, F. E., "Revisions to the JDL Data Fusion Model," *in: Proc. of the SPIE Conference on Sensor Fusion: Architectures, Algorithms, and Applications III*, pp. 430–441 (1999). doi: 10.1117/12.341367

[11] Llinas, J., Bowman, C., Rogova, G., Steinberg, A., Waltz, E. and White, F. E., "Revisiting the JDL Data Fusion Model II," *in: Proc. of the International Conference on Information Fusion*, pp. 1218–1230 (2004). doi: 10.1109/ICIF.2005.1591959

[12] Dasarathy, B. V., Decision Fusion, IEEE Computer Society Press, Los Alamitos CA (1994).

[13] Kokar, M. M., Tomasik, J. A. and Weyman, J., "Formalizing Classes of Information Fusion Systems," *Information Fusion*, Vol. 5, No. 3, pp. 189–202 (2004). doi: 10.1016/j.inffus.2003.11.001

[14] Champion, M., Ferris, C., Newcomer, E., et al., Web Services Architecture, W3C Working Draft, (2002-11-14). http://www.w3.org/TR/ws-arch.html.

[15] Huang, S. G., Fan, Y. S., Zhao, D., et al., "Web Service Based Enterprise Application Integration," *Computer Integrated Manufacturing Systems*, Vol. 9, No. 10, pp. 864–867 (2003).

[16] Hung, M.-H., Cheng, F.-T. and Yeh, S.-C., "Development of a Web-Services-based E-diagnostics Framework for Semiconductor Manufacturing Industry," *IEEE Transactions on Semiconductor Manufacturing*,

Vol. 18, No. 1, pp. 122−135 (2005). doi: 10.1109/TSM.2004.836664

[17] Otal, M. and Jelinekl, I., "The Method of Unified Internet-based Communication for Manufacturing Companies," *Lecture Notes in Computer Science, Springer*, Vol. 3190, pp. 133−140 (2004). doi: 10.1007/978-3-540-30103-5_15

Manuscript Received: May 14, 2014
Accepted: May 20, 2015