Study on Cloud Storage based on the MapReduce for Big Data

Huang Yi
Key Laboratory of Machine Vision and Intelligent Information System
Chongqing University of Arts and Sciences
Chongqing, P.R China
e-mail: cqhy@21cn.com

Zhang Yongdan
College of Technology
Guizhou University
Guiyang, P.R China
e-mail: 312865700@qq.com

Ma Xinqiang*
Key Laboratory of Machine Vision and Intelligent Information System
Chongqing University of Arts and Sciences
Chongqing, P.R China
*Corresponding author e-mail: xinqma@163.com

Liu Youyuan
Key Laboratory of Machine Vision and Intelligent Information System
Chongqing University of Arts and Sciences
Chongqing, P.R China
e-mail: 39541385@qq.com

Abstract—Big Data has been one of the current and future research frontiers. Data sets grow in size because they are increasingly being gathered by ubiquitous information-sensing mobile devices, aerial sensory technologies, remote sensing, software logs, cameras, microphones, radio-frequency identification readers, wireless sensor networks, and so on. Big Data has changed the way we capture and store data, including data storage device, data storage architecture, data access mechanism. Cloud computing moves the application software and databases to large data centers, where the management of the data and services may not be fully trustworthy. Storage is one of the major issues which hamper the growth of cloud. To deal with Big Data storage problems we discuss cloud storage platform based on the MapReduce for Big Data.

Keywords- Big Data; cloud storage; MapReduce; Hadoop; security

I. INTRODUCTION

In this age, Big Data applications are increasingly becoming the main focus of attention because of the enormous increment of data generation and storage that has taken place in the last years. Many real-world areas such as telecommunications, health care, pharmaceutical or financial businesses generate massive amounts of data. Gaining critical business insights by querying and analyzing such massive amounts of data is becoming the need of the hour [1] and has become a challenge to the standard data storage approaches. Traditionally, data warehouses have been used to manage large amounts of data. However, for the management of massive data that grow day after day, these are not able to provide reasonable response times. Big Data can be defined as data that exceeds the processing capacity of conventional systems [2]. This initiative will also lay the groundwork for complementary Big Data activities, such as dig data infrastructure projects, platforms development, and techniques in settling complex, data-driven problems in sciences and engineering. Finally, they will be put into practice and benefit society [3].There are 2.5 quintillion bytes of data created every day, and this number keeps increasing exponentially. The world's technological capacity to store information has roughly doubled about every 3 years since the 1980s. In many fields, like financial and medical data often be deleted just because there is no enough space to store these data. These valuable data are created and captured at high cost, but ignored finally [4]. The bulk storage requirements for experimental data bases, array storage for large-scale scientific computations, and large output files are reviewed in [5].

One of the most popular paradigms nowadays for addressing Big Data is MapReduce[6], a new distributed programming model that organizes the computation into two key operations: the map function that is responsible for splitting the original dataset and processing each sub-problem independently, and the reduce function that collects and aggregates the results from the map function.

In the last few years, cloud computing has grown from being a promising business concept to one of the fast growing segments of the IT industry. But as more and more information on individuals and companies are placed in the cloud, concerns are beginning to grow about just how safe an environment it is. Security issues in cloud computing has played a major role in slowing down its acceptance, in fact security ranked first as the greatest challenge issue of cloud computing [7].

In this work, we present an analysis of several techniques to deal with data storage. Specifically, the techniques evaluated in this study are techniques that have been proved as useful for cloud storage and which we have adapted for Big Data following a MapReduce scheme. Finally, in this paper, we discuss the trusted cloud computing platform (TCCP) based on the MapReduce for

Big Data that can deal with Big Data trusted storage problems.

II. BIG DATA AND DATA STORAGE

Big Data is a broad term for data sets so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, curation, search, sharing, storage, transfer, visualization, and information. The term often refers simply to the use of predictive analytics or other certain advanced methods to extract value from data, and seldom to a particular size of data set [8].

A. Big Data

Big Data [8] usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time [9]. Big Data "size" is a constantly moving target, as of 2012 ranging from a few dozen terabytes to many petabytes of data. Big Data is a set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale[10].

In a 2001 research report [11]and related lectures, META Group (now Gartner) analyst Doug Laney defined data growth challenges and opportunities as being three-dimensional, i.e. increasing volume (amount of data), velocity (speed of data in and out), and variety (range of data types and sources). Gartner, and now much of the industry, continue to use this "3Vs" model for describing Big Data[12].In 2012, Gartner updated its definition as follows: "Big Data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization"[13] . Additionally, a new V "Veracity" is added by some organizations to describe it.

If Gartner's definition (the 3Vs) is still widely used, the growing maturity of the concept fosters a more sound difference between Big Data and Business Intelligence, regarding data and their use:

- Business Intelligence uses descriptive statistics with data with high information density to measure things, detect trends etc.;
- Big Data uses inductive statistics and concepts from nonlinear system identification to infer laws (regressions, nonlinear relationships, and causal effects) from large sets of data with low information density to reveal relationships, dependencies and perform predictions of outcomes and behaviors.

A more recent, consensual definition states that "Big Data represents the Information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value"[14].

Big Data can be described by the following characteristics [8]:

- Volume - The quantity of data that is generated is very important in this context. It is the size of the data which determines the value and potential of the data under consideration and whether it can actually be considered Big Data or not. The name 'Big Data' itself contains a term which is related to size and hence the characteristic.
- Variety - The next aspect of Big Data is its variety. This means that the category to which Big Data belongs to is also a very essential fact that needs to be known by the data analysts. This helps the people, who are closely analyzing the data and are associated with it, to effectively use the data to their advantage and thus upholding the importance of the Big Data.
- Velocity - The term 'velocity' in the context refers to the speed of generation of data or how fast the data is generated and processed to meet the demands and the challenges which lie ahead in the path of growth and development.
- Variability - This is a factor which can be a problem for those who analyze the data. This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.
- Veracity - The quality of the data being captured can vary greatly. Accuracy of analysis depends on the veracity of the source data.
- Complexity - Data management can become a very complex process, especially when large volumes of data come from multiple sources. These data need to be linked, connected and correlated in order to be able to grasp the information that is supposed to be conveyed by these data. This situation, is therefore, termed as the 'complexity' of Big Data.

Big Data analytics consists of 6 Cs in the integrated industry 4.0 and Cyber Physical Systems environment. 6C system, that is, consist of connection (sensor and networks), Cloud (computing and data on demand), Cyber (model and memory), content/context (meaning and correlation), community (sharing and collaboration), and customization (personalization and value). In this scenario and in order to provide useful insight to the factory management and gain correct content, data has to be processed with advanced tools (analytics and algorithms) to generate meaningful information. Considering the presence of visible and invisible issues in an industrial factory, the information generation algorithm has to capable of detecting and addressing invisible issues such as machine degradation, component wear, etc. in the factory floor[15,16].

B. Data Storage

Big Data has changed the way we capture and store data [4], including data storage device, data storage architecture, data access mechanism. As we require more storage mediums and higher I/O speed to meet the challenges, there is no doubt that we need great innovations. Firstly, the accessibility of Big Data is on the top priority of the knowledge discovery process. Big Data should be accessed easily and promptly for further analysis, fully or partially break the restraint: CPU-heavy but I/O-poor. In addition, the under-developing storage technologies, such as solid-state drive (SSD) and phase-change memory (PCM), may help us alleviate the

difficulties, but they are far from enough. One significant shift is also underway, that is the transformative change of the traditional I/O subsystems. In the past decades, the persistent data were stored by using hard disk drives (HDDs). As we known, HDDs had much slower random I/O performance than sequential I/O performance, and data processing engines formatted their data and designed their query processing methods to work around this limitation. But, HDDs are increasingly being replaced by SSDs today, and other technologies such as PCM are also around the corner. These current storage technologies cannot possess the same high performance for both the sequential and random I/O simultaneously, which requires us to rethink how to design storage subsystems for Big Data processing systems [4].

Direct-attached storage (DAS), network-attached storage (NAS), and storage area network (SAN) are the enterprise storage architectures that were commonly used. However, all these existing storage architectures have severe drawbacks and limitations when it comes to large-scale distributed systems. Aggressive concurrency and per server throughput are the essential requirements for the applications on highly scalable computing clusters, and today's storage systems lack the both. Optimizing data access is a popular way to improve the performance of data-intensive computing [17], these techniques include data replication, migration, distribution, and access parallelism. In [18], the performance, reliability and scalability in data-access platforms were discussed. Data-access platforms, such as CASTOR, dCache, GPFS and Scalla/Xrootd, are employed to demonstrate the large scale validation and performance measurement. Data storage and search schemes also lead to high overhead and latency [19], distributed data-centric storage is a good approach in large-scale wireless sensor networks (WSNs). Shen, Zhao and Li proposed a distributed spatial–temporal similarity data storage scheme to provide efficient spatial–temporal and similarity data searching service in WSNs. The collective behaviors of individuals that cooperate in a swarm provide approach to achieve self-organization in distributed systems [20].

III. HADOOP AND MAPREDUCE

A. Hadoop

One of the most famous and powerful batch process-based Big Data tools is Apache Hadoop. It provides infrastructures and platforms for other specific Big Data applications. A number of specified Big Data systems (Table 1) are built on Hadoop, and have special usages in different domains, for example, data mining and machine learning used in business and commerce [4].

Apache Hadoop is one of the most well-established software platforms that support data-intensive distributed applications. It implements the computational paradigm named MapReduce. Apache Hadoop (see Fig. 1 and Fig. 2) platform consists of the Hadoop kernel, MapReduce and Hadoop distributed file system (HDFS), as well as a number of related projects, including Apache Hive, Apache HBase, and so on[4].

With the addition of MapReduce, Hadoop works as a powerful software framework for easily writing applications which process vast quantities of data in-parallel on large clusters (perhaps thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.

TABLE I. BIG DATA TOOLS BASED ON BATCH PROCESSING

Big Data tools	Big Data systems are built on Hadoop		
	Name	***Specified Use***	***Advantage***
	Apache Hadoop	Infrastructure and platform	High scalability, reliability, completeness
	Dryad	Infrastructure and platform	High performance distributed execution engine, good programmability
	Apache Mahout	Machine learning algorithms in business	Good maturity
	Jaspersoft BI Suite	Business intelligence software	Cost-effective, self-service BI at scale
	Pentaho Business	Analytics Business analytics platform	Robustness, scalability, flexibility in knowledge discovery
	Skytree Server	Machine learning and advanced analytics	Process massive datasets accurately at high speeds
	Tableau	Data visualization, Business analytics	Faster, smart, fit, beautiful and ease of use dashboards
	Karmasphere Studio and Analyst	Big Data Workspace	Collaborative and standards-based unconstrained analytics and self service
	Talend Open Studio	Data management and application integration	Easy-to-use, eclipse-based graphical environment

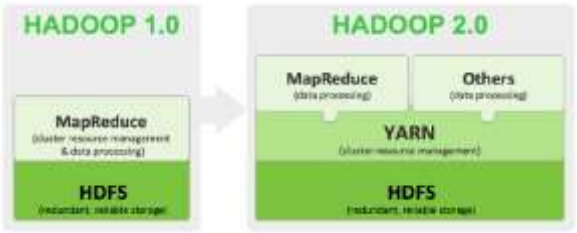

Figure 1. Hadoop system version.

Figure 2. Hadoop system architecture.

B. MapReduce programming model

The MapReduce programming model was introduced by Google in 2004[6]. It is a distributed programming model for writing massive, scalable and fault tolerant data applications that was developed for processing large datasets over a cluster of machines.

The MapReduce programming model abstracts the calculation process in two phases: Map and Reduce. In the Map phase, the master node splits the input dataset into independent sub-problems and distributes them to worker

nodes. Then, the worker nodes process in a parallel way the smaller problems and pass the answer back to its master node. Finally, in the Reduce phase, the master node takes the answers to all the sub-problems and combines them in a way to form the output. The users in this paradigm only have to define what should be computed in the Map and Reduce functions while the system automatically distributes the data processing over a highly distributed cluster of machines.

In the MapReduce model all the computation is organized around (key, value) pairs. In the first stage, the Map function, takes a single (key, value) pair as input and produces a list of intermediate (key, value) pairs as output. It could be represented as:

$$\text{map}(\text{key1}, \text{value1}) \rightarrow \text{list}(\text{key2}, \text{value2}) \quad (1)$$

Then, the system merges and groups by keys these intermediate pairs and passes them to the Reduce function. Finally, the Reduce function takes a key and an associated value list as input and generates a new list of values as output, which can be represented as follows:

$$\text{reduce}(\text{key2}, \text{list}(\text{value2})) \rightarrow (\text{key2}, \text{value3}) \quad (2)$$

Fig. 3 shows the overall flow of a MapReduce operation in our implementation. When the user program calls the MapReduce function, the following sequence of actions occurs [6].

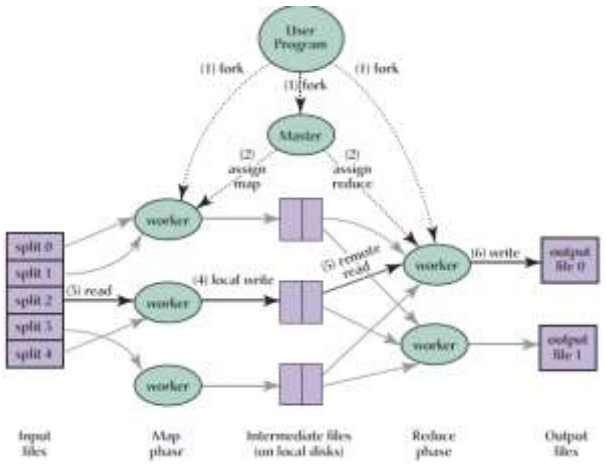

Figure 3. Execution overview.

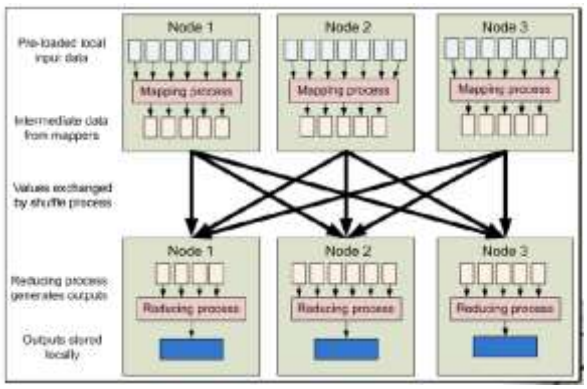

Figure 4. MapReduce program steps.

As illustrated in Fig. 4, a typical MapReduce program with its *map* and *reduce* steps. The master node is in charge of job scheduling and task distribution for the slaves. The slaves implement the tasks exactly as assigned by the master. As long as the systems start to run, the master node keeps monitoring all the data nodes. If there is a data nodes failed to execute the related tasks, the master node will ask the data node or another data node to re-execute the failed tasks. In practice, applications specify the input files and output locations, and submit their Map and Reduce functions via interactions of client interfaces.

IV. CLOUD COMPUTING AND CLOUD STORAGE.

A. *Cloud computing*

As illustrated in Fig. 5, cloud computing not only delivers applications and services over the Internet, it also has been extended to infrastructure as a service, for example, Amazon EC2, and platform as a service, such as Google AppEngine and Microsoft Azure. Infrastructure vendors provide hardware and a software stack including operating system, database, middleware and perhaps single instance of a conventional application. Therefore, it shows out illusion of infinite resources without up-front cost and fine-grained billing. It leads to the utility computing, i.e., pay-as-you-go computing [4].

Figure 5. Cloud computing logical diagram.

B. *Cloud storage*

Surprisingly, the cloud computing options available today are already well matched to the major themes of need, though some of us might not see it. Big Data forms a framework for discussing cloud computing options. Depending on special need, users can go into the marketplace and buy infrastructure services from providers like Google and Amazon, Software as a Service (SaaS) from a whole crew of companies starting at Salesforce and proceeding through NetSuite, Cloud9, Jobscience and Zuora-a list that is almost never ending. Another bonus brought by cloud environments is cloud storage which provides a possible tool for storing Big Data. Cloud storage have good extensibility and scalability in storing information as demonstrated in Fig. 6.

Cloud computing is a highly feasible technology and attract a large number of researchers to develop it and try to apply to Big Data problems [4]. Usually, we need to combine the distributed MapReduce and cloud computing

to get an effective answer for providing petabyte-scale computing [21]. CloudView [22] is a framework for storage, processing and analysis of massive machine maintenance data in a cloud computing environment, which is formulated using the MapReduce model and reaches real-time response. In [23], the authors extended MapReduce's filtering aggregation programming model in cloud environment and boosts the performance of complex analysis queries.

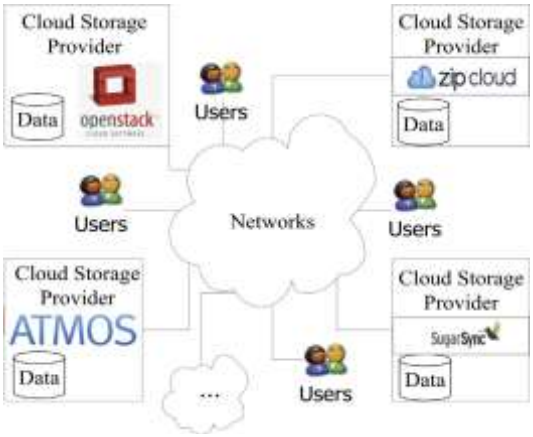

Figure 6. Cloud storage.

We will discuss the trusted cloud computing platform in the future. Such as, in a computer and service systems, access is a specific type of interaction between a subject and an object that results in the flow of information from one to the other [24].

ACKNOWLEDGMENT

This work is supported by the Natural Science Foundation Project of CQ CSTC (cstc2014jcyjA40056 and cstc2013jcyjA40053) and the Scientific and Technological Research Program of Chongqing Municipal Education Commission (KJ1401112 and KJ1401126) and the Natural Science Foundation Project of Ycstc (2013nb8001, 2014bf2001 and 2013ad2002) and the Natural Science Foundation Project of Guizhou University for Young Teachers (201240).

REFERENCES

[1] R. Gupta, H. Gupta, M. Mohania, Cloud computing and big data analytics: what is new from databases perspective? in: Proceedings of the 1st International Conference on Big Data Analytics (BDA 2012), vol. 7678 of Lecture Notes on Computer Science, 2012, pp. 42-61.

[2] M. Minelli, M. Chambers, A. Dhiraj, Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses, John Wiley &Sons, 2013.

[3] Sara del Río, Victoria López, José Manuel Benítez and Francisco Herrera, On the use of MapReduce for imbalanced big data using Random Forest. Information Sciences 285 (2014) 112–137.

[4] C.L. Philip Chen and Chun-Yang Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, Information Sciences 275 (2014) 314–347.

[5] William J. Worlton, Bulk storage requirements in large-scale scientific calculations, IEEE Trans. Magn. 7 (4) (1971) 830-833.

[6] Jeffrey Deam, Sanjay Ghemawat, Mapreduce: simplified data processing on large clusters, Commun. ACM 51 (1) (2008) 107–113.

[7] Huang Yi, Ma Xinqiang, Liu Youyuan and Li Danning. Research of the Issues based on Trusted Cloud Security. Advanced Materials Research, Vols. 989-994(2014), pp5000-5003.

[8] http://en.wikipedia.org/wiki/Big_data#cite_note-Editorial-13,2015

[9] Snijders, C.; Matzat, U. and Reips, U.-D.. "'Big Data': Big gaps of knowledge in the field of Internet". International Journal of Internet Science 7 (2012): 1–5.

[10] Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani and Samee Ullah Khan, The rise of "big data" on cloud computing: Review and open research issues, Information Systems, Volume 47, January 2015, Pages 98-115, ISSN 0306-4379,http://dx.doi.org/10.1016/j.is.2014.07.006

[11] Laney, Douglas. "3D Data Management: Controlling Data Volume, Velocity and Variety". Gartner. Retrieved 6 February 2001.

[12] Beyer, Mark. "Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data". Gartner. Archived from the original on 10 July 2011. Retrieved 13 July 2011.

[13] Laney, Douglas. "The Importance of 'Big Data': A Definition". Gartner. Retrieved 21 June 2012.

[14] De Mauro, Andrea; Greco, Marco; Grimaldi, Michele. "What is big data? A consensual definition and a review of key research topics". AIP Conference Proceedings 1644(2015): 97–104.

[15] Lee, Jay; Bagheri, Behrad; Kao and Hung-An. "Recent Advances and Trends of Cyber-Physical Systems and Big Data Analytics in Industrial Informatics". IEEE Int. Conference on Industrial Informatics (INDIN) 2014.

[16] Lee, Jay; Lapira, Edzel; Bagheri, Behrad; Kao and Hung-an."Recent advances and trends in predictive manufacturing systems in big data environment". Manufacturing Letters 1 (1): 38–41.doi:10.1016/j.mfglet.2013.09.005.

[17] Renato Porfirio Ishii, Rodrigo Fernandes de Mello, An online data access prediction and optimization approach for distributed systems, IEEE Trans.Parallel Distrib. Syst. 23 (6) (2012) 1017-1029.

[18] M. Bencivenni, F. Bonifazi, A. Carbone, A. Chierici, A. D' Apice, D. De Girolamo, L. dell' Agnello, M. Donatelli, G. Donvito, A. Fella, F. Furano, D. Galli, A.Ghiselli, A. Italiano, G. Lo Re, U. Marconi, B. Martelli, M. Mazzucato, M. Onofri, P.P. Ricci, F. Rosso, D. Salomoni, V. Sapunenko, V. Vagnoni, R. Veraldi,M.C. Vistoli, D. Vitlacil and S. Zani, A comparison of data-access platforms for the computing of large hadron collider experiments, IEEE Trans. Nucl. Sci. 55(3) (2008) 1621-1630.

[19] Haiying Shen, Lianyu Zhao, Ze Li, A distributed spatial-temporal similarity data storage scheme in wireless sensor networks, IEEE Trans. Mobile Comput. 10 (7) (2011) 982-996.

[20] Hannes Muhleisen and Kathrin Dentler, Large-scale storage and reasoning for semantic data using swarms, IEEE Comput. Intell. Mag. 7 (2) (2012) 32-44.

[21] Steve Loughran, Jose Alcaraz Calero, Andrew Farrell, Johannes Kirschnick, Julio Guijarro, Dynamic cloud deployment of a mapreduce architecture,IEEE Internet Comput. 16 (6) (2012) 40-50.

[22] Arshdeep Bahga, Vijay K. Madisetti, Analyzing massive machine maintenance data in a computing cloud, IEEE Trans Parallel Distrib. Syst. 23 (10) (2012) 1831-1843.

[23] Dawei Jiang, Anthony K.H. Tung, Gang Chen, Map-join-reduce: toward scalable and efficient data analysis on large clusters, IEEE Trans. Knowl. Data Eng. 23 (9) (2011) 1299-1311.

[24] Yi. Huang and Xinqiang Ma. An access control model based on Trusted Computing (In Chinese), Journal of Chongqing University of Arts and Sciences, (2010), 29(3): p. 54-57.