On the Research of Big Data Storage

H.F QIN, Z.M. QIAN
Computer Science department, ChuXiong Normal University, China

ABSTRACT: This paper want to research the storage of big data, the paper from the processing mode of big data start, study big data process, requirements, management and storagea. Analysis currently being studied big data storage technology GFS, Bigtable, HDFS, NoSQL, multi-Replication and disaster recovery technology. The aim is for the storage of big data parepare.
KEYWORD: Big data; big data storage; GFS; Bigtable; HDFS; NoSQL

1 PREFACE

With the advent of the internet of things, cloud computing and big data, data-centric and storage-centric has become an important trend. Data is growing at an unprecedented rate in constant growth and accumulation, big data era has arrived. Academia, industry and even government agencies have begun to pay close attention to the big data, and developed an interest on it. Storage is a key factor affecting the performance of computer systems; storage is an important carrier of the entire chain of data, information, knowledge, wisdom.

Big data is an another technological change after the internet of things, cloud computing. It have a significant impact in national governance, business decision-making, organization, business process and personal life. The core of cloud computing is business model; the essence is the data processing technology. Data is assets, cloud provide the channels of data assets custody, access and location. How to manage data assets and let it severs to the state, businesses, personal .It is the core issue of big data, is the soul for cloud computing and inevitable direction of the research. Research on big data storage technology, mainly focused on cloud computing. At present, the technologies mainly contain google's GFS technology, bigtable technology, HDFS technology, and Nosql technology and multi-copy technology.

2 THE PROCESSING MODE OF BIG DATA

Big data is a broad term for data sets so large or complex that they are difficult to process using traditional data processing applications. Challenges include analysis, capture, curation, search, sharing, storage, transfer, visualization, and information privacy [1]. Big data have some characteristics, enormous volume and modal is various, generate fast, the value huge but the density is low. Big data processing mainly includes the following process: data source, data extraction and aggregation, data analysis, data interpretation and data comprehensive application. The process figure as follows figure1.

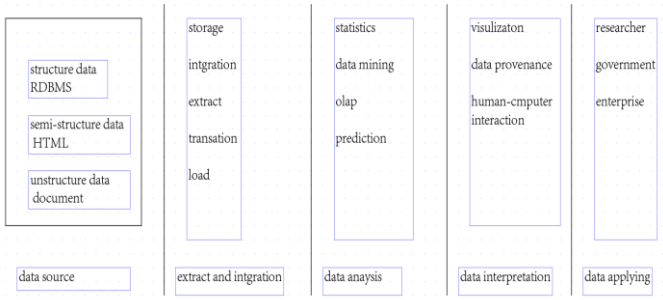

Figure1 big data process architecture

2.1 *Big data processing mode analysis*

Big data processing mode mainly divided into stream processing and batch processing, stream processing is also called straight-through processing, batch processing is store-then-prcess[8].Stream processing, the value of the data will be reduced with the passage of time continuously. And Contrast, batch processing, is store-then-prcess, the value of the data will be increased with the passage of anlysis continuously. So, batch processing is the most popular approach of big data process. As the big data, the first problem to manifest itself has been storage. With an increase in the volume of data, the amount of storage needed to process and store it increases by 1.5 times. You need

the additional 0.5 times storage for intermediate result set processing and storage[2]. Big data is valuable, but to mining the value of big data is similar to the waves in panning. This value not once discovery, it requires a continuous change process, such as video monitoring, producing 24 hours of video every day, most of the data is no value, may be a few seconds to capture shots to a certain physical characteristics of offenders, is precious to the public security sector, for those few seconds, 24 hours all data must be save. This is a typical feature of the big data, low density high value. The store-then-prcess batch processing is more in line with big data processing requirement. Big data will become part of the modern social infrastructure gradually, like roads, railways, ports, utilities and communications networks as indispensable. Big data is not only because of possession and exhausted, but also with the use and dissemination of the continuous enrichment and growth. As you can imagine, the data will be the future of DNA. But having the data volume, quality and application will form a significant segment of data resources and competitive advantage. Therefore, big data will become the countries for the next strategic frontier. With the rise of "smart economy", collecting data, master data, and use data will become the country's core competitiveness of enterprises.

According to the big data storage mode, the data and information always back and forth on the layer of processing and storage converted, get the information from the data, the information can also be obtained from the new data. The converion figure as follows figure2.

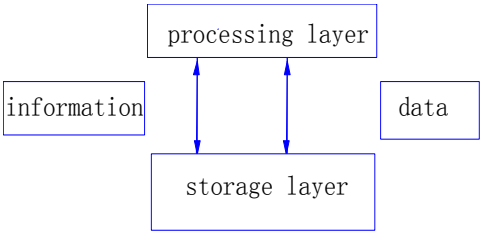

Figure2 Computational units of data processing

2.2 *Big Data processing requirements analysis*

Big Data, compare the data, not only volume, its contains five main characteristics: the data Volume, Velocity, Variety, Ambiguity, Complexity.

According to the characteristics and requirements of big data, we need analysis big data management, find the approach of the big data.

2.3 *Big data management analysis*

The goal of big data management is to ensure a high level of data quality and accessibility for business intelligence and big data analytics applications[5]. At present, the method of big data management is centre in centralized processing and distributed processing. Centralized processing, in this architecture all the data is collected to a single centralized storage area and processed upon completion by a single computer with often very large architectures in terms of memory, processor, and storage [2]. Another method of big data management is distributed processing, in this architecture data and its processing are distributed across geographies or data centers, and processing of data is localized with the federation of the results into a centralized storage. Distributed architectures evolved to overcome the limitations of the centralized processing, where all the data needed to be collected to one central location and results were available in one central location [2].

According to the management mode analysis, we learn the big data have two methods, one is centralized processing, another is distributed processing, these two methods are also solving to the method of storage.

2.4 *Big data storage analysis*

According to the big data management analysis, the big data may be centralized and distributed, also according to the type of storage allocation. At present, there are three types of data storage, the first is traditional storage, traditional storage is mainly contain data, calculations on the same server, the server own high-performance calculator and controller, also carry some harddisk that can store a certain capacity data. The advantage of traditional storage is only running a machine can met the needs of user. The disadvantage is the storage is small, it can be stored a certain volume of data, but can't stored big data. So the traditional storage are in some small and medium-sized companies. The second is the hybrid storage, it need two severs, the server only responsible for the high performance computing and some important data storage, great part of the general data are stored in another server. The advantage of hybrid storage can process more volumes, it requires computing server and storage server work together to meet the needs of all users, and this approach has the advantage of open computer server can meet some part of the customer needs. If you want to meet the needs of all customers, must jointly open when the computing server and storage server. This way of storage for data quantity big or small can meet, this form mainly exists in large scale data, with traditional and new application services, the data is not completely migration are widespread. The third way storage and compute Complete separation, it is by far the most widely applied in the big data, storage is conducted by a dedicated server, computing server, only responsible for scheduling and calculation, not articulated any storage devices, storage all put in storage array, only if the computer servers and storage server open together to meet the needs of any user. As to the big data, the third way is the most.

3 THE KEY TECHNOLOGY OF BIG DATA STORAGE

“How much is memory, how long is the program". In an ideal situation, every programmer will love is infinite, fast and easy to change the content (content that is not lost after power-down) of memory, but you also want it is cheap [3]. Thus, according to the needs of the processing of big data, now, we need a storage system is required to have high availability, high capacity, low cost, fast access, easy to change. The value of big data need multiple technologies together. The file system provide the support at the bottom of the storage capacity

At present, big data storage mainly adopt cloud computing, cloud computing is a recently evolved computing terminology or metaphor based on utility and consumption of computing resources. Cloud computing involves deploying groups of remote servers and software networks that allow centralized data storage and online access to computer services or resources. Clouds can be classified as public, private or hybrid [6] .Cloud computing has some technology in storage such as GFS, Bigtable, HDFS.

3.1 *GFS*

GFS (Google File System), Google is the founder of big data. Google GFS, oriented large-scale data-intensive applications, scalable distributed file system. While GFS running on cheap universal hardware equipment, but it is still the ability to provide disaster redundancy, provides a large number of clients with high availability services. GFS completely satisfy our demand for storage. GFS as storage platform has been widely deployed in inside Google, storing our service production and processing of data, but also for those who need massive data set research and development work. So far, thousands of machine using one of the largest cluster of thousands of hard drives, offers hundreds of terabytes of storage space, for hundreds of client services at the same time. GFS is built on top of a large number of low-cost servers a scalable distributed file system, GFS, as to big files, and read much larger than the written application scenarios, using the master-slave (Master-Slave) structure. Through data block, an additional update (append-only) and other ways to achieve the efficient storage of massive data.

3.2 *Bigtable*

Bigtable is a distributed structured data storage system; it is designed to handle huge amounts of data, usually distributed in thousands of ordinary petabytes of data on the server. Bigtable is google early development of database system, it is a multi-dimensional sparse sorting table, consists of rows and columns, each storage unit has a time stamp, and form a three-dimensional structure. Different time for the same data unit formed multiple operating data by timestamp to distinguish between multiple versions, many projects using google bigtable storage data, including the index of the Web, Google Earth, Google Finance. These applications for bigtable request difference is very big, whether on the amount of data (from the URL to the web to satellite images) or on the response rate (from the backend batch processing to real-time data services). Although it is very big that the application demand difference, but these products on Google, Bigtable provide a flexible, high-performance solution.

3.3 *HDFS*

Hadoop is the most popular big data processing platform. Hadoop is an imitation of one of the first open source cloud computing platform, GFS Mapreduce implemented, and after by Apache company developed into HDFS, hadoop has now developed into a file system (HDFS), database (Hbase, Cassandra), data processing (MapReduce) and other functions of the system, in a sense, Hadoop has become an important tool for big data processing.

HDFS is a highly fault-tolerant, scalable, and distributed file system architected to run on commodity hardware. The HDFS architecture was designed to solve two known problems experienced by the early developers of large-scale data processing. The first problem was the ability to break down the files across multiple systems and process each piece of the file independent of the other pieces and finally consolidate all the outputs in a single result set. The second problem was the fault tolerance both at the file processing level and the overall system level in the distributed data processing systems [2].

From the perspective of cloud computing, the GFS technologies, Bigtable technology, HDFS technology belong to the distributed file system are similar to GFS on implementation. All storage clusters by MasterServer and multiple ChunkServer constituted; more suitable for "Write Once Read Many" model, due to factors massive metadata treat massive small files are relatively weak, but they support online expansion.

Due to the traditional data are stored in the database, people have become accustomed to relational database SQL, accustomed to using a database always want to use the database to solve the problem, therefore therefore developed a NoSQL and other technologies.

4 OTHER TECHNOLOGY

Data storage becomes more and more important in big data. In addition to above, There are some

technologies exist, such as NoSQL, multi-replication and disaster recovery technology.

4.1 *NoSQL Technology*

NoSQL sometimes called Not Only SQL abbreviation, also called non-relational database, is different from a traditional database management system relational database collectively, NoSQL database, is an approach to data management and database design that's useful for very large sets of distributed data[7]. Relational databases cannot handle the scalability requirements of large volumes of transactional data, and often fail when trying to scale up and scale out. The vendors of RDBMS-based technologies have tried hard to address the scalability problem by replication, distributed processing, and many other models, but the relational architecture and the ACID properties of the RDBMS have been a hindrance in accomplishing the performance requirements of applications, such as sensor networks, web applications, trading platforms, and much more. The name NoSQL (not only SQL) database was coined by Eric Evans for the user group meeting to discuss the need for nonrelational and non-SQL-driven databases. This name has become the industry-adopted name for a class of databases that work on similar architectures but are purpose-built for different workloads [2].

4.2 *Multi-Replication Technology*

Due to the emergence of big data, data centralized storage has become impossible. The problems existing in the centralized storage, distributed storage is more and more become the mainstream. As if our data is stored in only one place, if the disk have problem, our data will be lost forever,so multi-replication technology become a key technolgy. Replication is complete copy of the original data. By providing a copy of the files in the system increase various forms, save the file data redundancy, can be very effective to improve the availability of the file, avoid the data loss caused or not accessible due to the network disconnected or dynamic unpredictable factors such as machine failure. Generally, the more the number of replication, the higher the reliability of the file, but if all the files are stored more number of replications, the system will consume a large amount of storage resources, and increase the complexity of the file management. Multi-replication technology improves the data availability and performance. When a replication fail, the system automatically re-distribution of the data, restored to a replication as soon as possible. Multi-replication technology not only on the availability of benefits, but also brings benefits in performance. Although there are many benefits, but there are also a lot of problems, the biggest problem is data redundancy. Therefore, the key to control the quantity and quality is key technology of replication.

4.3 *Disaster Recovery Technology*

Data Disaster Recovery refers to establish a remote data system, which is a real-time replication of local critical application data. Application disaster recovery is based on data disaster, in different places to establish a complete production system with considerable local backup application system (which can be up each other), in the event of a disaster, the remote system quickly take over business operations. Data disaster recovery is guaranteed to withstand disasters, while the application of disaster recovery is the goal of disaster recovery system construction. In disaster recovery technology, the most critical is distributed RAID technology and storage delete redundant technolog.

5 CONCLUSION

Big data storage problems are an important issue in big data processing at present. Big Data is very complex, traditional storage methods can not meet the real needs of big data. Many data storage is now being studied, have their advantages and disadvantages. Analyze existing various storage in the world, to find a better or more conducive to big data storage providing infrastructure and protection.

ACKNOWLEDGEMENT

This work was supported by the Scientific Research Fund of Yunnan Provincial Education Department (No. 2013Y056), and the Program for Innovative Research Team (in Science and Technology) in University of Yunnan Province.

REFERENCES

[1] http://en.wikipedia.org/wiki/Big_data.
[2] Krish Krishnan,Data Warehousing in the Age of Big Data . Elsevier Inc.2013.
[3] Andrew S. Tanenbaum, Albert S. Woodhull, Operating Systems: Design and Implementation Second Edition.
[4] http://en.wikipedia.org/wiki/High_availability.
[5] http://searchdatamanagement.techtarget.com/definition/big-data-management.
[6] http://en.wikipedia.org/wiki/Cloud_computing.
[7] http://searchdatamanagement.techtarget.com/definition/NoSQL-Not-Only-SQL.
[8] Meng Xiaofeng and Ci Xiang, Big Data Management: Concepts, Techniques and Challenges, Journal of Computer Reasearch and Development 50(1):146-169, 2013.