# New Trend of Data Management

Ying Li, JianJun Peng, Luqiao Fan

[1] *Guangdong Polytechnic of Science and Technology, Zhuhai,Guangdong.519090*

Keywords: Big data;RDBMS; MapReduce; New trend.

Abstract: Data management has experienced 60 years of development since the 1950s, and instead of processing object simply, it has become a basic resource. At present, it has met with problems, like the system's lack of expandability and supporting data types being single. Data management represented by Mapreduce rises as a new force and exploits any techniques and methods that can be used for reference from precious wealth accumulated by other data management technology, solving its performance problems. To adapt to the requirement of data depth analysis, it is seeking for a new mode of data management .

## 1 INTRODUCTION

Big data into our lives, to subvert our ten thousand years of construction of information acquisition, information memory, knowledge of (Qiang D,2012) theory of information storage, data management technology is undergoing major changes, we try to the historical context of technological progress, starting from the application, to explore new trends in data management!

## 2 THE PROCESS OF THE DEVELOPMENT OF DATA MANAGEMENT

### 2.1 Data management from tape drive to relational database

Data management history can be traced back to fifty years before the last century, when the data management is very simple. Through a large number of classification, comparison and machine operation table drawing millions of punched cards for data processing, the results in the paper printed or punched card made new. While the data management is to all these perforation the card for storage and processing of physics. However, in 1951 Remington Rand Co (Remington Rand Inc.) a computer called Univac I launched a tape drive, a second can inputhundreds records, which led to the data management revolution of.Seed database system appeared in 60s. At that time the computer is widely used in data management, put forward higher requirements for data sharing. The traditional file system can not meet the needs of people. To unified management and sharing of the database data management system (DBMS) came into being. The data model is the core and foundation of the database system, DBMS software is based on a data model.

The earliest is DBMS mesh (Aoying Z,2009), Ge Corp in the United States is the successful development of IDS Bachman et al in 1961 (Integrated DataStore).1961 (General ElectricCo). The Ge Corp Charles Bachman successfully developed the world's first DBMS network is the first database management system integrated data storage (Integrated DataStore IDS), laid the foundation of network the database, and has been issued and widely used at the time.

Hierarchical DBMS is followed by the network database appears. The most famous of the most typical hierarchical database system is developed by IBM company in 1968, IMS (Information Management System), a hierarchical database for the host. It is developed by IBM, the first large-scale database program produced products. From the end of 60s. Now IMSV6 has been developed to provide the cluster, N, data sharing, message queue sharing and other advanced features support..

Network database and hierarchical database has been well solved and centralized data sharing problem, but in the data independence and abstraction level is still great. The lack of access to users in the two database, still need to clear the data storage structure, points out the access path and RDBMS later resolved these problems.1970,

351

Fan L., Peng J. and Li Y.
New Trend of Data Management.
DOI: 10.5220/0006450303510357
In *ISME 2016 - Information Science and Management Engineering IV (ISME 2016)*, pages 351-357
ISBN: 978-989-758-208-0

IBM researcher Dr. E.F.Codd in magazine published an article entitled "A Relational Model of Data for Large Shared Data Banks" thesis, put forward the concept of relationship model, laid the theoretical foundation of the relational model. Then Codd has published a number of articles on 12 standards for the paradigm theory and measure system, use mathematical theory laid the foundation of the relational database model. A rigorous mathematical basis and abstract level than High, simple and clear, easy to understand and use. But at that time, some people think that the relational model is the ideal data model, is used to achieve DBMS is not realistic, particularly concerned about the performance of relational database is difficult to accept, more people see it as a serious threat to the network database standardization work is underway in order to promote.

## 2.2 Object oriented database (OODB)

With the rapid development of information technology and market, people found that the relational database system although the technology is mature, but its limitations are obviously: it can well deal with the so-called "form data", but more and more on the technology sector of the complex data types is powerless. After 90s, the technology sector has been looking for a new database system in the study and what is the direction of development. But in the new database system on the issue, the industry was quite puzzled. Influenced by the wave of technology, in quite a long time, people put a lot of effort spent on research of object-oriented database systems (object oriented database) "or the" OO database ". It is worth mentioning that the proposed object-oriented professor Stonebraker relational database theory has once been the industry's favor. However, Several years of development shows that the object relational database system products in the market development situation is not ideal. Theory of perfect and no warm reaction brought the market. The main reason is not successful, the main design idea of this database products is to attempt to use the new database system to replace the existing database system. Accumulated a lot of data of customers has been used for many years and database system, especially for large customers, is unable to withstand the huge workload of the new data conversion between caused and huge expenses. In addition, the object-oriented relational database system to make the query language is extremely complex, which makes both the database development business application or customers are regarded as dangerous and complicated application technology.

So far, the object oriented database (OODB) has experienced three main stages: (1) the laboratory prototype stage, many prototypes completed in the laboratory at the end of 80s, most of them have more new ideas and bold design but the lack of practical application of the test, on behalf of products industry Vbase and Orien so, the academic circles of Gemstone; (2) the initial stage of commercial products, at the end of 80s to early 90s, the launch of the product has the basic features of object-oriented database management system and the actual operation ability, OODB market has played a major role, but there are still many defects; (3) mature product stage in mid 1990s, so far, the commercialization of OODB mature products such as Object Store, Ontos, O2, Jasmin and so on.

# 3 THE ARRIVAL OF THE ERA OF BIG DATA

## 3.1 Large data generation

With the development of the Internet, we collected data of unprecedented set, big data era has come. Big data in the Internet industry refers to such a phenomenon: the Internet Co generated in daily operations, the user network behavior data accumulated. The scale of these data is so large, that can't be measured by G or T.

How big data is? A group called "the Internet Day" data tells us that one day, all the contents of the Internet generated can be engraved with 168 million DVD; messages sent 294 billion letters of (the number is equivalent to the United States two years of printed letters issued); community posts 2 million (the equivalent of "Time magazine for 770 years, the amount of text); mobile phone sold 378 thousand units, higher than the global daily baby number 371 thousand..2012 years, the amount of data from the TB (1024GB=1TB) level jumped to PB (1024TB=1PB), EB (1024PB=1EB) and ZB (1024EB=1ZB).2012 China produced single level the amount of data is up to 0.4ZB, the amount of data and the 2013 Chinese produced is as high as 0.85ZB (equivalent to 800 million TB). The International Data Corporation (IDC) the results showed that the amount of data generated by the 2008 global 0.49ZB, the amount of data for 2009 0.8ZB growth in 2010 For the 1.2ZB, in 2011 the number is as high as 1.82ZB, equivalent to more than 200GB of the global per capita data generated by the world.

This trend will continue, the automation of data generation and data generation speed will also speed up, we need to deal with the data will be expanded rapidly, complicated calculation, large data necessary for analysis (Big Analytics), to data treasure, otherwise, we will be lost in the sea of data.

## 3.2 Data analysis

In order to understand the data changes, continuous optimization and improvement, not only the symptoms but also to cure, make similar problems no longer appear; continuous monitoring and feedback, we find the optimal scheme to solve the problem fundamentally. We must carry out in-depth analysis of the data, rather than just simple to generate these complex statements. Analysis of the model, it is difficult to use SQL to express, collectively referred to as the depth of analysis (deep analysis) (Xiong Pai Q,2012).

We need not only through the data to understand what happens now, more need to predict what will be carried out using the data, in order to make some active preparations in action (Figure 2) shows. For example, by predicting the sales of goods in advance to take action on the goods for timely adjustents.

Here, (Xiong Pai Q,2012) typical OLAP data analysis (data collection, aggregation, slicing etc) are not enough, also need to path analysis, time series analysis, graph analysis, What-if analysis and the statistics of hardware / software and never tried to analysis model, the following is a typical example of such as time series analysis and graph analysis.

Time series analysis (time series analysis): business organization has accumulated a lot of historical transaction information, enterprise management personnel at all levels of the hope from the analysis of this data in order to find business opportunities from some mode, through trend analysis, and even found some advance is emerging opportunities. For example, in the financial services industry, analysts can according to the analysis of the development of software, analysis of time series data, looking for profitable trading mode (profitable trading pattern). After further verification, the operator can actually use the transaction mode of trading profits.

Analysis of large scale map analysis and network (large-scale graph and network analysis): (Ziyu L,2012) (Social Network) social network virtual environment is essentially on the real connectivity description in social network, each independent entity is represented as a node in the graph, the relation between the entity represented as an edge. Through the analysis of the social network, can be found in some useful knowledge, for example, found that some type of entity (a type of entity to each group together, known as the key entity in the network). This information can be used to direct product, organization and individual behavior analysis, the field of potential security threats with growth analysis. The social network, from the perspective of geometry, graph nodes and edges are constantly growing. Using the traditional methods of processing large-scale chart data is insufficient, need urgent and effective means of this kind of number According to the analysis.

Analysis of large scale map analysis and network (large-scale graph and network analysis): (Ziyu L,2012) (Social Network) social network virtual environment is essentially on the real connectivity description in social network, each independent entity is represented as a node in the graph, the relation between the entity represented as an edge. Through the analysis of the social network, can be found in some useful knowledge, for example, found that some type of entity (a type of entity to each group together, known as the key entity in the network). This information can be used to direct product, organization and individual behavior analysis, the field of potential security threats with growth analysis. The social network, from the perspective of geometry, graph nodes and edges are constantly growing. Using the traditional methods of processing large-scale chart data is insufficient, need urgent and effective means of this kind of number According to the analysis.

# 4 MAPREDUCE AS THE REPRESENTATIVE OF THE RISE OF THE DATA MANAGEMENT TECHNOLOGY

## 4.1 Source of MapReduce

The big data era, in order to carry out the processing of large-scale data, whether it is operating or application analysis application, (Yijie W,2012) parallel processing is the only option, the parallel processing is not only across multiple cores, more importantly, it is across nodes, parallel processing depends on the amount of nodes to improve the performance of distributed system, a large number of nodes, even if the cost is not a problem and the selection of high-end, reliable hardware equipment, but because the cluster size is large, reaching thousands of nodes, node failures, network failures become commonplace, fault tolerance guarantee becomes especially important. Large scale horizontal expansion of the system, not for relational database system ready, there is not a relational database system to deploy more than 1000 cluster nodes, and MapReduce technology has reached a 8000 node deployment scale.

According to the theory of CAP (Xiongpai Q,2013) (later proved by Gilbert and Lycnh), large-scale distributed systems, consistency (consistency), the availability of the system (availability) and network partition tolerance (network partition tolerance) of the 3 target, you can only get one of the two characteristics, pursue two goals will harm another goal the 3 goal, it may not be possible (Fig. 3). In other words, if the pursuit of consistency and high availability, network partition tolerance. It is difficult to meet the pursuit of consistency and fault tolerance of highly parallel database system, not obtain availability scalability and good system, and expansion the system is an important prerequisite for big data analysis.

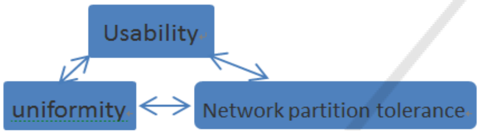

Figure 1: CAP Theory

Proposed by Google company in 2004, MapReduce technology, to solve large-scale unstructured data fast batch processing technology of parallel.MapReduce framework is a simple parallel computing model, which solves the problem of scalability in the system level, fault tolerance and other issues, through the Map and Reduce functions to receive user written, automatic parallel execution in a large cluster scalable, and data processing and analysis of large-scale.MapReduce technology is a typical representative of the non relational data management and analysis technology in Google company, through the large scale cluster and MapReduce software, there are more than 20PB data can be processed every day, more than 400PB. of such a large amount of data analysis and data management each month of treatment, is unable to complete the relationship between traditional data management techniques.

## 4.2 Technical framework of MapReduce

MapReduce framework includes 3 aspects: (1) highly fault-tolerant distributed file system (Google file system); distributed file system running on the cluster, cluster using inexpensive machine construction. Data using key / value pairs (key/value) model for storage (Aoying Z,2009). The entire file system with centralized metadata management, data a distributed storage model, through the data replication (each data backup at least 3) to achieve a high degree of fault tolerance. (2) parallel programming model; parallel programming model of the calculation process is divided into two main stages, namely Map stage and Reduce stage.Map function key/value to deal with, produce a series of intermediate key/value, Reduce function is used to merge with the same key value of intermediate key value, calculate the final results of. (Xiongpai Q,2013) (3) parallel execution engine; the implementation process of MapReduce parallel program execution as shown in Figure 4: firstly, data sources are divided into blocks, and then give it to a Map to perform the task, the Map task Map function, according to the classification of the data of some rules, write the local hard disk; Map phase is complete, enter Reduce Reduce Reduce, the task execution function, with the intermediate results of the same key value, the node from multiple Map mission, were collected together (shuffle) combined treatment, the output written to the local hard disk (into the distributed file system).

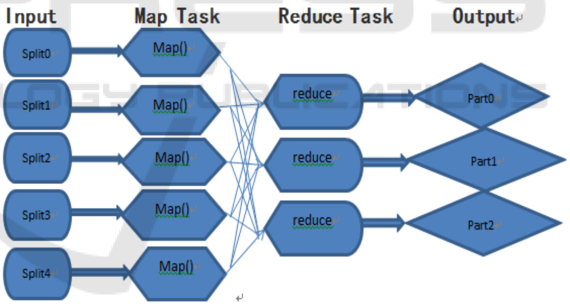

Figure 2: Mapreduce parallel computing process

The birth of each technology, will be questioned and criticized, MapReduce technology is the same. The famous Stonebraker and Dewitt database experts thought that MapReduce technology is a huge setback, and pointed out many disadvantages of MapReduce technology, for example, does not support Schoop, there is no access optimization, relying on brute force (brute force) data processing etc., no innovation at all, and is already out of the database of the technology 25 years ago, is doomed to have no future, two experts view set off a heated debate. The development of Stonebraker (Dengguo F,2014) and Dewitt but a torrent of criticism did not stop with MapReduce technology as the representative of the new data analysis technology.

354

New Trend of Data Management

In contrast with the passage of time, the MapReduce technology has obtained the widespread attention, the researchers conducted in-depth research on MapReduce technology, such as expanding the application field of MapReduce, MapReduce Technology to enhance the performance of MapReduce technology, usability improvement etc.

## 5 DATA MANAGEMENT TECHNIQUES ARE INTEGRATED AND DEVELOPED WITH EACH OTHER

Each kind of data management technology are the advantages and disadvantages of the merger, but has its irreplaceable. After such as MapReduce search related data analysis on widely used, with the continuous expansion of the application fields and improve its performance, quickly became a relational database (RDBMS) of the young competition, technology was for promoting their integration and development.

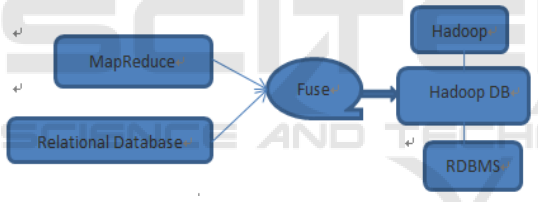

Figure3: Hadoop DB plan

Hadoop DB is a professor at the University of Abadi proposed by Yale, is a (Xiongpai Q,2013) fusion scheme MapReduce and the traditional relational database (Figure 5), to improve the performance of the MapReduce.Hadoop DB optimization MapReduce technology makes full use of fault tolerant RDBMS performance and Hadoop, distribution characteristics. In the Hadoop DB system, clear divided into two layers, the upper layer uses the Hadoop decomposition and scheduling task, lower query and data processing with RDBMS. The work of innovation is to improve scalability and fault tolerance of the system by using the Hadoop task scheduling mechanism, in order to solve the problem of horizontal expansion of data analysis; data storage and query processing using RDBMS, in order to solve the performance problem. In the performance test, the performance of DB Hadoop.Hadoop DB is still lagging behind the integration of MapReduce and traditional relational database relationship database system, is not only a language and Interface level, but also to make a system level improvements, and is being commercialized as Hadapt, allegedly pulled from the 10M to the VC investment. At present, the DB Hadoop system is still in continuous improvement and perfection.

As the two companies emerging big data -- Aster Data and Greenplum methods are similar, is the transformation of (Aoying Z,2009) database using PostgreSQ MapReduce, so it can run in large scale clusters (MPP/Shared nothing), to achieve efficient data processing, both in the integration of MapReduce technology. The two companies sanbanfu are the PostgreSQ database, the core engine Shared nothing architecture and MapReduce technology of.Aster Data and Greenplum database can execute SQL queries, as well as the MapReduce Job execution engine, in the same framework to achieve a unified treatment of SQL/MapReduce Aster and Greenplum Data. The MapReduce technology based on experience and parallel data analysis software package (MapReduce rewrite technology analysis software package), which is focused on the integration of MapReduce technology, the RDBMS technology of MapReduce Study.

The integration of RDBMS and MapReduce technology are in many aspects, including (Xiongpai Q,2013) storage, indexing, query optimization, connection algorithm, application interface, algorithm implementation and other aspects. For example,.RCFile RCFile storage system under the framework of HDFS preserves the scalability and fault tolerance of MapReduce (extended is the primary requirements of large data processing system the massive cluster of highly extended warranty), must pay attention to the fault-tolerant storage structure gives the HDFS data block is similar to the PAX, using RDBMS technology to improve the processing performance of the Hadoop system, which is from the MapReduce camp RDBMS fusion technology and ideas. The work is not the original innovation, the first work is stored in the Hadoop column on the platform, has been widely cited. More important is RCFile was used in Facebook to solve the practical problems is the industrial and academic collaborative innovation paradigm.

355

# 6 RESEARCH DIRECTION OF DATA MANAGEMENT TECHNOLOGY IN THE ERA OF BIG DATA

The era of big data, the data has been increasing in the excitement, data management is facing more challenges in research and development (Mingxuan N,2011), showing an unprecedented bustling scene. The theme of our research is always around the higher performance, more data and more complex analysis.

## 6.1 Large data unified processing platform

With the rise of MapReduce technology, we can see that the ecosystem data analysis (ceo-system) is changing, for example, derived from a new ecological system Facebook and the relational database, the traditional analysis of the ecological system, the ecological system of two purposes, a natural method is two, and the ecosystem is the fusion technology together? We believe that through the theory of academia and industry efforts, a unified data processing framework and ecological system will form.(QiY,2013)but there are still some problems to research further, for example, how to integrate various types of data to a storage layer (data structure); how to build the storage layer more intelligent (storage model, data deduplication, data compression, indexing, query routing, the storage layer performs certain calculation); how to improve the query algorithm debugging, execution The algorithm, make it adapt to multi core, GPU, heterogeneous environment, supercomputers, clusters and other cheap hardware environment, and can run on the cloud platform; how to go beyond the SQL, provide programming and application interface, for all types of users (including ordinary users, advanced users, statisticians, data experts) to provide data analysis environment flexible a variety of tools and how to deal with large data; the results of visualization; how to guide users to conduct exploratory queries on data (Exploratory Query), in order to establish the analysis model, and in-depth analysis.

## 6.2 Graph and social networks

As people become interested in social networks more concentrated, research on social network is more and more attracted attention. Based on social network is analyzed in the figure.how we choose to measure the similarity between nodes on the map (SimRank) as one of our branches? Existing SimRank calculation method has two limitations: first, the computational cost is too large; second, can only be applied to the static map. How to use the new parallel algorithm GPU inherent parallelism and high bandwidth design; how to speed up the big picture (large graph) on the SimRank calculation; secondly, SimRank calculation is essentially a first-order Markov chain problem, based on the observation. What can we design, the coupling degree of the Markov chain to open, to realize the node similarity calculated by scoring what way; in a series of experiments to verify the algorithm in what way the effectiveness and efficiency.

# 7 SUMMARY

This article mainly from the original way of data management, introduced from the relational database to the development process of big data management technology represented by MapReduce, launched a new era of picture data management technology of big data. With the continuous improvement and analysis ability and constantly enhance MapReduce performance and relational database technology and MapReduce technology integration and development, and direction of the current big data research.

# REFERENCE

Qiang D., *Analysis on Quality of Service Provisioning for Communication Services in Network Virtualization*, Journal of Communications,7(2)143-154, 2012

Aoying Z., *Data intensive computing - the challenge of data management technology. China computer society*, communications, 5 (7): 50-53, 2009

Xiong Pai Q, Hui Ji W, Xiaoyong D etc., *Data analysis: RDBMS and MapReduce competition and symbiosis*. Journal of software, 23 (1) 32-45,2012

Ziyu L, Yongxuan L, Chen L, Yi X., *Research on cloud database*. Journal of software,23 (5): 1148-1166,2012

Yijie W, Weidong S, Song Z etc., *Key technologies of distributed storage in the cloud computing environment*. Chinese Journal of software, 23 (4): 962-986,2012

Xiongpai Q, Huiji W , Furong L etc., *The new pattern of data management technology*. Journal of software, 24 (2): 175-197 ,2013

Dengguo F, Min C, Hao L., *Large data security and privacy protection*. Chinese Journal of computers, 37 (1): 50-57, 2014

356

New Trend of Data Management

Mingxuan N, Wuman L., *Technological change in the era of data explosion*. Communication of computer science, 7 (7): 12-20,2011

Qi Y., *Value of big data in "smart cam-pus"*. Netinfo Security. (8)91-93, 2013

357