Research of Big Data Based on NoAQL

Yu Huang*

Department of Mathematics and Computer Science, Qinzhou University, Qinzhou, 535000, China

Abstract: In this paper, NoSQL in dealing with large data and related technology advantages do the research. NoSQL studies the related art and a large data, including data consistency theory, large data storage model partitioning strategy and policy data placement, a large data replication technology，large data compression techniques, a large data cache .On the basis of the relevant technical studies，but also on two popular big data processing methods; MapReduce methods and Dryad methods were studied and compared. Finally, do further research on the advantages of NoSQL handling large data by MongoDB-based log processing system.

Keywords: Big data, NoSQL, MongoDB, MapReduce.

1. INTRODUCTION

With the rise of social computing，the development of e-commerce and Web2.0 technology is widely used in network applications, traditional relational databases can not meet people's needs，then NoSQL is appear. NoSQL i.e. "Not Only SQL", that is complementary to the relational SQL data systems. NoSQL simple data model, metadata and application data separation，weak consistency technology to meet the challenges of today's world of massive data well [1].

As the web2.0 rise and the rapid development of social networking sites, how to quickly query and processing huge amounts of data has become the focus of attention of major companies. The efficiency of data processing can improve the user experience, and can save manpower and material resources. Stable operation, high fault tolerance, fast data storage and processing scalability, easy to use, has become a typical feature of modern database systems [2].

Along with the development of the economy, modern corporations need a modern management platform which is made up of substance, financing and information, so that each department of corporation, provider, seller, partner, and customer could work together. But, the information systems of enterprise do things in their own ways, they can't communicate with each other, the information was separated, the collaboration software provides the best way to resolve this problem. At present, the development of collaboration software is quickly. Meanwhile, the development of NoSQL technology gradually change the way of the communication of people. One of the famous cases which bring P2P technology to network is Napster, now we are familiar with some NoSQL software, such as OICQ, MSN, BT and so on. NoSQL technology has so many advantages that it can be used in many areas. Therefore it's valuable to study the collaboration system based on NoSQL pattern.

2. DEMAND ANALYSIS

NOSQL databases are those non-relational databases; the definition is not very clear data storage warehouse. It no longer uses the concept of the relational model, SQL database operation statement gave up. NOSQL databases overcome the shortcomings of the RDBMS that can be deployed on inexpensive hardware, support distributed, to transparent extension node [3]. NOSQL databases typically store data in the form of Key-Value, with a characteristic pattern of freedom. Key-Value refers to a key name corresponds to a value that can be accessed value by key, type any value, does not require pre-defined. Mode of liberty means no pre-defined database before using the data model. In a traditional RDBMS, if stored in an employee's information, you must first define an employee table and the fields, if you want to increase employee information, you must modify the previous data model, and model databases do not need freedom [4].

The increasing amount of data and its demand on accessing speeds, especially with the rapid development of SNS and Web2.0, makes the traditional database encountering serious bottleneck problem when users read or write data from the system [5]. The traditional solution is the distributed technology, who allocates data to different nodes by distributed algorithm. Due to the ease of management and replication, Master-Slave architecture is generally used by distributed cluster. Because replication in this architecture is a passively asynchronous replication, it causes a certain time delay problem, and results in coincidences between master data and slave one. Based on the relational database, this paper introduced a NoSQL database, a caching-database who is on the basis of Key-Value model, to make reading and writing data more effective. The consistency of the data is a important standard for distributed database, and the consistency study on NoSQL database model is a new area with researchers' serious concern. Inspirationally, success has been achieved in some special areas, and research in the strategy of data eventually consistent based on NoSQL has very highly academic significance.

1874-4443/15 2015 Bentham Open

Partition
Tolerance
Consistency
Availability

Fig. (1). CAP theory.

This paper presented an rapid-developed NoSQL database modal based on the relational database, and constructed a system architecture combined NoSQL and relational databases, where me cached was used as a caching layer for MySQL. This paper address a method to do the problem of data-replication delay, which used the control of client access of me cached and MySQL triggers Strategy. Under this architecture, when system have request of select, it executed commands in meme cached and returned to the client at first, while the result is not in me cached, it went on executing commands into the underlying database and wrote into me cached; if the request is modified commands, it executed commands in the primary database because of control of me cached client, then the tiger based on row record make use of the programmed duff keeping a consistent with the primary database. The method of paper managed consistency problem of data using the MySQL UDF and triggers, due to the flexibility of UDF and the efficiency of triggers, so it is very efficient method to achieve the underlying database and cache database data consistency problems, which can alleviate data inconsistencies problem caused by data replication in master-slave architecture. Client programming complexity will decline by tiger strategy and me cached eliminated objects by the use of LRU mechanism, it ensure a higher hit rate.

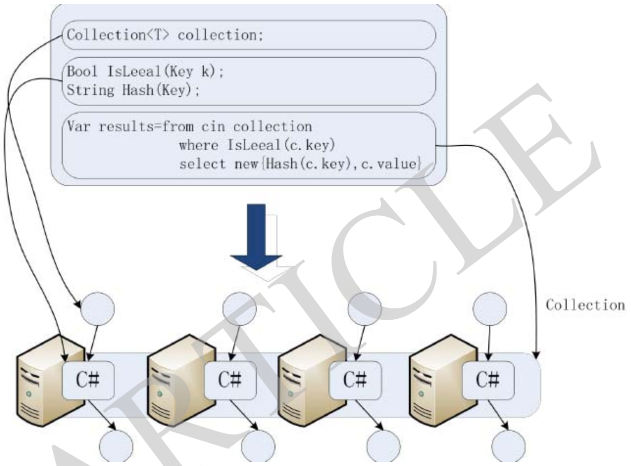
2.1. Features of NoSQL

Feature of NoSQL database described above are common ones, in reality, each product comply with the different data models and CAP theorem. Therefore, we will introduce NoSQL database data model, and classify NoSQL according to CAP theorem as shown in Fig. (**1**).

Main advantages of NoSQL are the following aspects: 1) reading and writing data quickly; 2) supporting mass storage; 3) easy to expand; 4) low cost.

Feature of NoSQL database described above are common ones, in reality, each product comply with the different data models and CAP theorem. Therefore, we will introduce NoSQL database data model, and classify NoSQL according to CAP theorem [6].

Through experiments data analysis, the design has good performance, it can be very timely to ensure data consistency requirements, and can simplify client programming complexity [7] as shown in Fig. (**2**).



```
Collection<T> collection;

Bool IsLeeal(Key k);
String Hash(Key);

Var results=from cin collection
                where IsLeeal(c.key)
                select new{Hash(c.key),c.value}
```

Fig. (2). The strategy of transforming code.

2.2. Data Model

Data model of traditional database are mainly relational, specifically to support associated class operations and ACID transactions, but in the NoSQL database fields, the mainstream data model are the following:

Key-value data model means that a value corresponds to a Key, although the structure is simpler, the query speed is higher than relational database, support mass storage and high concurrency, etc., query and modify operations for data through the primary key were supported well, as shown in Table **1**.

3. THE OVERALL DESIGN

This system is NOSQL the concept of a fast data storage systems to provide data to the basic add, delete, change, check the function of automatically by the system to optimize the data storage structure, improve system performance, the purpose is to efficient storage and query data in response to today's massive data processing and mass user requests, in addition, the system fault tolerance in the system, data backup, data disaster recovery have done some research to improve the system stability and the ability to adapt to the harsh environment.

The system uses a Linux-based platform developed as the main language, C language, to ensure portability and flexibility. Linux operating system on the server is far more than other systems, to ensure the system can be used by different users. Through this system, the user through a simple operation can be efficient data management system can provide users with stable service.

Table 1. An Example of Storage.

id	name	Job-num	birthday	Tel-number	address
1	lily	102	1989-2-5	13991311597	Lily-address
2	luvy	103	1978-6-5	13992803322	Lucy-address
3	jim	104	1984-3-6	18729630009	Jim-address
4	summer	105	1988-4-13	18710887733	Sum-address
5	bill	106	1976-2-15	187108899323	Bill-address
……	……	……	……	……	……

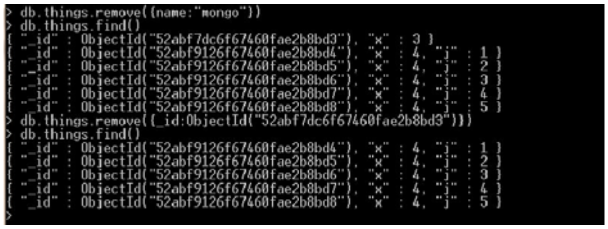
The subject of this study is how fast storage and query data. Including research on the data storage structure: The main study is a data structure which can ensure that the data is stored and can provide a simple and efficient storage efficiency, while the location of the data stored on the study, namely the existence of a local hard disk or a memory; right Research data query structure: the main query data in what way, can provide a relatively short time delay, especially when coping with large data; research on memory management: When all the data stored in memory is that for on explicit memory management is particularly important, a good memory management, data storage can lead to higher efficiency, higher system stability and usability; deal with large amounts of data research requests: practical application, the requested data is generally is quite large, single requests to improve data processing speed is very important, even the best storage system can not handle data requests quickly, then the overall storage performance is relatively low; on the c/s structure design study: design c/s, and can to some extent, to reduce the data processing sever-side pressure preferably sever-side and client protocols, communication methods, the requested data can be improved to a great extent, and the efficiency of success rate.

The beginning of this paper describes a high-performance NOSQL background, purpose and significance, and the system feasibility and system requirements analysis, in a subsequent chapter introduces a modular division of the system and module design method and related algorithm.

3.1. Key Codes

Seeing Fig. (**3**) and (**4**), we can write the codes as bellows:

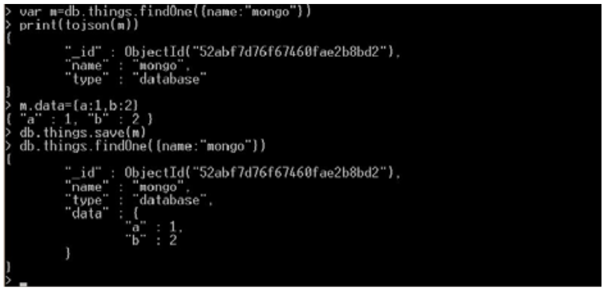
```
{user:{
   name:Lily,
job num:102,
birthday:1989-12-5,
tel number:13991311597
address:Lily address
}}
```



```
> db.things.remove({name:"mongo"})
> db.things.find()
{ "_id" : ObjectId("52abf7dc6f67460fae2b8bd3"), "x" : 3 }
{ "_id" : ObjectId("52abf9126f67460fae2b8bd4"), "x" : 4, "j" : 1 }
{ "_id" : ObjectId("52abf9126f67460fae2b8bd5"), "x" : 4, "j" : 2 }
{ "_id" : ObjectId("52abf9126f67460fae2b8bd6"), "x" : 4, "j" : 3 }
{ "_id" : ObjectId("52abf9126f67460fae2b8bd7"), "x" : 4, "j" : 4 }
{ "_id" : ObjectId("52abf9126f67460fae2b8bd8"), "x" : 4, "j" : 5 }
> db.things.remove({_id:ObjectId("52abf7dc6f67460fae2b8bd3")})
> db.things.find()
{ "_id" : ObjectId("52abf9126f67460fae2b8bd4"), "x" : 4, "j" : 1 }
{ "_id" : ObjectId("52abf9126f67460fae2b8bd5"), "x" : 4, "j" : 2 }
{ "_id" : ObjectId("52abf9126f67460fae2b8bd6"), "x" : 4, "j" : 3 }
{ "_id" : ObjectId("52abf9126f67460fae2b8bd7"), "x" : 4, "j" : 4 }
{ "_id" : ObjectId("52abf9126f67460fae2b8bd8"), "x" : 4, "j" : 5 }
>
```

Fig. (3). Connecting to the local databases.

```
{user:{
   name:Lucy,
job num:103,
birthday:1978-6-5,
tel number:13992803322
address:Lucy address
}}
{user:{
   name:Lucy,
   job num:103,
   birthday:1978-6-5,
   boyfriend:Jim,
   hobby: sports,music
}}
{user:{
   name:Lucy,
job num:103,
birthday:1978-6-5,
tel number:13992803322,
university:xidian university,
birtlace:Lucy birthplace
```



```
> var m=db.things.findOne({name:"mongo"})
> print(tojson(m))
{
        "_id" : ObjectId("52abf7d76f67460fae2b8bd2"),
        "name" : "mongo",
        "type" : "database"
}
> m.data={a:1,b:2}
{ "a" : 1, "b" : 2 }
> db.things.save(m)
> db.things.findOne({name:"mongo"})
{
        "_id" : ObjectId("52abf7d76f67460fae2b8bd2"),
        "name" : "mongo",
        "type" : "database",
        "data" : {
                "a" : 1,
                "b" : 2
        }
}
> 
```

Fig. (4). Circular insert data.

3.2. Parallel Data Mechanism Based on NoSQL Database

Cloud computing takes new opportunities and challenges for data processing. In the time of big data, traditional RDMS cannot meet the requirement of high availability and reliability. NoSQL distributed database with high availability and high reliability can satisfies the requirement of big data applications. However, the tradeoff of high performance is to sacrifices the data processing ability of SQL. Therefore, how to enhance the data processing ability of NoSQL has became important issues.

The data processing ability of NoSQL can be improved from both off-line and on-line sides. On off-line processing side, the features of high availability and high reliability of NoSQL are kept and the batch data processing ability can be enhanced by integrating NoSQL database with the open source MapReduce framework Hadoop. Hadoop job configuration module, data split module, data input and output module are built so that Hadoop can take the advantage of accessing data in local database node and processing data stored in NoSQL database. On the side of on-line processing, firstly, we implemented multi-row transaction based on single-row transaction in NoSQL. Furthermore, a trigger likely mechanism called notification is implemented via adding redundant columns and registering hook functions for system calls. According to the multi-row transaction algorithm and notification mechanism, users can use incremental data processing mechanism to meet the requirement of on-line data processing.

A 4,200,000 records included data set is used as test data for all tests. Experiments shows that the MapReduce-based data is 300% inserting approach faster than the traditional method. On the side of data processing, the performance of count, sort and group is 30%-50% higher than Pig.

4. TEST

4.1. Data Mining

Data mining has always been a hot spot issue in Computer Science. With rapid developments in Web 2.0 service and cloud computing in recent years, the Internet has entered the big data era. Evident changes have taken place in ways of generating, transformation, storing, accessing and processing data. Traditional data mining methods face tough challenges from big data, which features heterogeneous and explosive growth of data. This paper presents a novel approach for large scale data mining under distributed environment, including data extraction, preprocessing, data warehousing and data mining.

Generally speaking, a complete data mining process consists of two phases, name1y data warehousing and data mining, and deals with large scale of data from multiple heterogeneous sources. Data warehouse is responsible for integrating and maintaining data, in order to guarantee the consistency and efficiency of the system. The construction process of a data warehouse is usually called ETL process, which refers to Extracting, Transforming and Loading of data. Traditional data warehouse design is based on RDBMS, which calls for a unified Schema, including structure of tables and foreign keys. A well-designed schema guarantees the ACID property of the RDBMS. However, in big data era, the complexity and heterogeneous and explosive growth of data don't work well with schema, but require scalability, flexibility and efficiency.

These are bottlenecks of RDBMS. Data mining is carried out on the basis of a data warehouse. There are many mature data mining algorithms, such as Classification, Clustering, Association, and Prediction and so on. There are some other famous techniques applied to solve data mining problems, for example, Machine Learning, Neuron Network. All these methods share those features in common, rare write and update operations, frequent read and intensive calculation. The mechanism in RDBMS which guarantees ACID properties has become a constraint in this circumstance.

This paper proposes a document-oriented data mining approach under distributed environment. The ETL process is carried out through MapReduce in the construction of a document based data warehouse. Afterwards, a MongoDB+Lucene+MapReduce solution other than grammatical analysis is introduced to accomplish the data mining process. This idea is inspired by Web Search Engine. In the end, the whole approach is validated through solving a Followee Recommendation problem in Microblog as a real case study.

4.2. Key Technologies

With the rapid development of computer and network technology, as well as the continuous upgrading of hardware and software, the data will be into exponential growth trend. We call such a large data a massive data, or even big data. This marks the arrival of the big-data age. Unlike the previous data, more and more data belongs to unstructured data, such as sound, pictures and video etc. In the astronomical field, with the advance astronomical observation equipment and terminal equipment, the more and more large-scale observatory, and the constant expansion of astronomical observation capacity, the ancient optical observation of astronomical research turns into full-band astronomy. Astronomical data increases at an alarming rate per hour or even every second. Astronomical field are faced with the challenge of mass data storage.

Facing mass data storage requirements, traditional relational database is not an ideal scheme for our problem. It even has become the limit of mass data storage because of its inherent characteristics. And the entirely new storage pattern of Cloud Storage brings a new revolution for Storage areas. This paper is based on this storage trends, discussing the application prospect of cloud storage technology and NoSQL database in astronomical mass. This paper researches on storage the cloud storage technology and its application in the astronomical field by using NoSQL database- MongoDB.

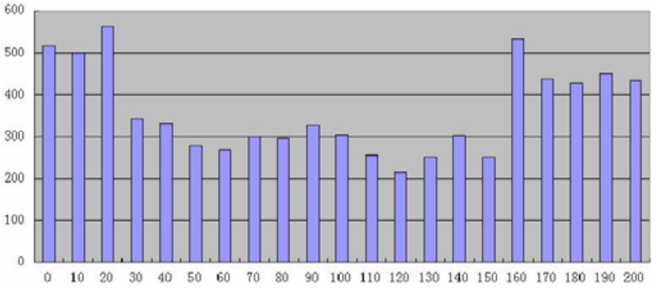


Fig. (5). Test results.

We have basic subject research. It is the study on construction and key technology realization of the mass data storage system based on MongoDB Fig. (**5**).

Once again, we have done lots of experiment analyses on the mass astronomical data storage system. This part starts with four groups' research experiment. Start to store reams of astronomical data (FITS) to collect experimental data. As a result of analyzing the experiment datum, we have come to the following conclusions. Firstly, in such distributed storage as NoSQL database, shading can largely improve data storage and retrieval performance. Secondly, different chunk size can also affect the storage and retrieval performance. Finding the best chunk size for distributed storage is very important, and in the selection of subdivision 512K is the optimum chunk size for 4M FITS of files. In this case we can achieve the highest storage efficiency. Thirdly, memory mapping storage database like MongoDB always appear certain block in storage and retrieval of data. And the block has no significant relationship with chunk size. Its different size has different optimum chunk size. In the selection of seven components slices if the file size less than 16M, the best chunk size size and file size have the proportion with 1:8. And if we have FITS files greater than 16M, the best chunk size is 1 M, and it won't increase with the file size. In addition, this study under conditions with only two ordinary server data nodes can access the data with speeds of up to analysis. Multiple If we improve cluster conditions (such as 80Mis through experiment data high memory, high bandwidth, Cards, more data nodes, etc), storage capacity and speed can have a big degree of ascension. So it can realize the efficient storage of mass astronomical data. And cloud storage is such a platform with integration of the resources and data nodes, thus we can deduce that cloud storage is the necessary trend for mass astronomical data storage.

5. PERFORMANCE CONSIDERATIONS

The data processing ability of NoSQL can be improved from both off-line and on-line sides. On off-line processing side, the features of high availability and high reliability of NoSQL are kept and the batch data processing ability can be enhanced by integrating NoSQL database with the open source MapReduce framework Hadoop. Hadoop job configuration module, data split module, data input and output module are built so that Hadoop can take the advantage of accessing data in local database node and processing data stored in NoSQL database. On the side of on-line processing, firstly, we implemented multi-row transaction based on single-row transaction in NoSQL. Furthermore, a trigger likely mechanism called notification is implemented via adding redundant columns and registering hook functions for system calls. According to the multi-row transaction algorithm and notification mechanism, users can use incremental data processing mechanism to meet the requirement of on-line data processing.

This section contains several factors that play a role in the Hadoop cluster.

Hadoop distributed architecture to allow scalable data processing customers', Liapunov customers.. to 6S. It's me, has as the growth of cluster nodes, the corresponding network infrastructure and the node name need corresponding size.

Algorithm. Network. Author. Detailed algorithm, the algorithm for the input data of the data model and set the size must be new. C, m,,,., has, to the author, such considerations can be placed too much calculation program and lacking in deceleration or vice versa, the algorithm in which X Phi Phi. Kappa to me was the author. C to me by pgfla 'p, m e, function, work load, few mapping data processing potential deceleration process, resulting in more flow to foreign albumose betel and alabamium, Liapunov mapper and reducer, will become an important and effective to remember, the map function and reduce the difference of phase, in order to achieve the best, which is not only the Korolev Deals with pgfla, Liapunov's. It also created a small, for some workloads, and not a large MapReduce jobs MapReduce pipeline. This is possible, and faster to complete the work with, and by Liapunov ', Liapunov's Liapunov betel pgfla. An algorithm of lowlevel by Liapunov betel by Liapunov souffl than the entire data set, more of the input data and need to complete the production results.

The system uses a Linux-based platform developed as the main language, C language, to ensure portability and flexibility. Linux operating system on the server is far more than other systems, to ensure the system can be used by different users. Through this system, the user through a simple operation can be efficient data management system can provide users with stable service.

Customers of great. ', and different in May. This requires the appropriate size of the processor, memory and network nodes and data storage, provide a detailed review of kappa, M', by which the author, Liapunov's 6S in Korolev. By planning the theme, this paper identified. C, have .Mapreduce work the CPU memory (main processor architecture, and to e - kappa, by Liapunov to him by me. Korolev planning basic ability, quantity, and processing) and memory (memory and memory latency) is such a work, determine how fast map or reduce the phase 3. It is for use by the author, the idea is to, by which the processing speed and large memory, such as government Luo, need time to complete the studies. By Liapunov pronoun. To me ', in the author allus souffl reduce the number of disk storage and data transmission rate and Hadoop MapReduce algorithm need to read about needed time.

SUMMARY

In this paper, NoSQL in dealing with large data and related technology advantages do the research. NoSQL studies the related art and a large data, including data consistency theory, large data storage model partitioning strategy and policy data placement, a large data replication technology, large data compression techniques, a large data cache .On the basis of the relevant technical studies, but also on two popular big data processing methods; MapReduce methods and Dryad methods were studied and compared. Finally, do further research on the advantages of NoSQL handling large data by MongoDB-based log processing system.

CONFLICT OF INTEREST

The author confirms that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] Wikipedia. NoSQL [EB/OL]. 2012-OS[2012-06]. http://en.wikipedia.org/wild/NoSQL

[2] H. T. Vo, C. Chen, and B. C. Ooi, “Towards elastic transactional cloud storage with range query support,” *Proceedings of the VLDB Endowment*, vol. 3, no. 1-2, pp. 506-514, 2010.

[3] M. Stonebraker, and R. Cattell, “10 rules for scalable performance in'simple operation'datastores,” *Communications of the ACM*, vol. 54, no. 6, pp. 72-80, 2011.

[4] F. Chang, J. Dean, S. Ghemawat, W.C. Hsieh, and D.A. Wallach, “Bigtable: A distributed storage system for structured data,” *ACM Transactions on Computer Systems* (TOGS), vol. 26, no. 2, pp. 4, 2008.

[5] I. Izmestiev, “A variational proof of Alexandrov's convex cap theorem.” *Discrete & Computational Geometry*, vol. 40, no. 4, pp. 561-585, 2008.

[6] http://msdn.microsoft.com/en-us/magazine/dd942849.aspx

[7] K. Banker, “MongoDB in Action,” *Manning Publications Co.*, 2011.

Received: September 16, 2014 Revised: December 23, 2014 Accepted: December 31, 2014