International Journal of Database Theory and Application
Vol.6, No.5 (2013), pp.141-152
http://dx.doi.org/10.14257/ijdta.2013.6.5.13

Research on Heterogeneous Data resource Management Model in Cloud Environment

Tao Sun[1, 2] and Xinjun Wang[1]

[1]*School of Computer Science and Technology, Shandong University*
Jinan, Shandong 250101, China
[2]*School of Information, Qilu University of Technology*
Jinan, Shandong 250353, China
suntao0906@163.com,wxj@sdu.edu.cn

Abstract

With the development of the cloud computing, more and more enterprises migrate their application systems to the cloud computing environment, these systems need multiple data resource collaborative work, and integrate existing heterogeneous data storage system. This paper aiming at the massive data processing, proposes a kind of heterogeneous data resource management model. This model implement massive resources storage, massive storage network' generation, update and balance of workload, proposes the security management and monitoring methods. The model proposed by this paper gives a novel solution to the heterogeneous data resource management and application in the cloud.

Keywords: *cloud computing; heterogeneous Data resource Management; distributed storage; dynamic management*

1. Introduction

As cloud computing is a flexible computing mode, more and more enterprises' application systems are migrated to cloud environment, large amount of business data is stored in different data nodes. Data location and organization are transparent to users, the heterogeneous data is distributed in different cloud node, each node only contains part of the information that users need, so it is necessary to effectively manage the heterogeneous cloud data, forms a giant data pool that masked the distribution, heterogeneity and complexity of the data resource, ensures efficient using of data services in cloud environment, realizes data integration and data operation transparency.

There is large variety of heterogeneous data resource in cloud environment, includes traditional relation database that from different software vendors, include different types of NoSQL database, and different types of file, all of the data resource grows rapidly, become a great repository of data [1]. In order to meet the enterprises' requirements for data services, the heterogeneous data resource should be extended and combined dynamically, so the research of heterogeneous data resource management in cloud computing have been presented newer and higher requirements.

The heterogeneous data resource management includes data resource description, dynamic organization, discovery and matching, optimal scheduling, and real-time monitoring and other operations. In cloud computing environment, all of these aspects are not isolate. Data resource dynamic organization will be influenced by heterogeneous data resource description; data resource discovery and matching mechanism should be adjusted to achieve the efficient data

ISSN: 2005-4270 IJDTA

management according to the data resource description and dynamic organization. Meanwhile, it is necessary to develop resource scheduling policy for monitoring and control entire data resources in cloud environment, ensures the entire system efficient operation, security and stability [3].

This paper presents a novel heterogeneous data resource management model, the model consists of all the aspects of heterogeneous data resource management in cloud environment, including: data resource description, physical and logical storage management and resource virtualization *etc*. Meanwhile, this paper presents the main problems of data resource management in cloud environment and key implementation technologies of the model. This paper gives a good solution for heterogeneous data resource management and application in cloud.

2. Related Work

The heterogeneous data resource management in cloud environment is a hot research topic, cloud data is partitioned and replicated to achieve scalability and fault-tolerance. Traditional relational databases could not meet the high scalability, high availability and high performance needed by cloud computing. The software vendors introduced a variety of new type database, these storage systems different from relational database, so they are called "NoSQL" database system. NoSQL database has overcome some shortcoming of relational database, could be deployed on cheap hardware, support for distributed storage, could be transparently extended. Typical NoSQL database adopts key-value pair storage mode, possesses characteristic of pattern free.

Different NoSQL database management systems have emerged, such as HBase and HyperTable are based on BigTable and adopt open source style, they are widely used on Facebook and Twitter. Traditional database vendors, such as Teradata, Oracle, IBM and Microsoft, put forward their own cloud database products, cloud service providers, such as Amazon, Google, Yahoo!, have begun to get involved in cloud database market, and there is some emerging small companies, such as Vertica,LongJump, EnterproseDB, provide their own cloud database products [2]. Major NoSQL database management system is showed in Table 1.

Table 1. Major NoSQL Data Management System

Vendors	Products
Google	BigTable, FusionTable
Amazon	Dynamo, SimpleDB, RDS
Microsoft	Microsoft SQL Azure
Oracle	Oracle Cloud
Yahoo!	PNUTS
Vertica	Analytic Database for the Cloud
EnterpriseDB	Postgres Plus in the Cloud
Open source	Hbase, Hypertable, Cassandra, Redis

Data model for relational databases has special significance, it's very difficult to change when database has been set up, otherwise, enterprise will pay a terrible price. NoSQL databases break the limitation on data model, allows stored data in any structure, add or delete elements in data item would not affect other data unit. But NoSQL database does not support the transaction management, and the free model will bring some unintended consequences. In addition, most of NoSQL databases are based on open source project, compared with the mature relational database, they are not

perfect. Therefore, the traditional relational database and NoSQL database will coexist with a period [4].

Currently cloud data management technology is developing rapidly, with the development of NoSQL, various cloud data processing method comes out one after another. Mainstream cloud database include Google's BigTable, HBase and HyperTable. The latter two cloud database based on BigTable and adopted open source implementation. BigTable runs on GFS, HBase runs on HDFS, and HyperTable runs on KFS, HDFS and GFS. In the calculation mode, BigTable, HBase, Hive and HyperTable are based on the MapReduce computing model [7]. VodeMort, PNUTS and SQL Azure will still be based on relational database computing model [10]. HadoopDB is a hybrid database that combined with relational database and cloud database, could provide SQL interface to access relational database and MapReduce API to access cloud database simultaneously[6]. Any database could not satisfy the CAP (Consistency, Availability, Partition tolerance) theory in three aspects, so the design of cloud data management model could only be considered from two aspects.

3. Problems of Data Resource Management in Cloud Environment

Currently the research and applications of heterogeneous data resource management in cloud computing should solve the following problems [5]:

(1) How to describe the heterogeneous data resource in cloud environment. As there are no uniform standards to describe heterogeneous data resource, resulting in data isolated and scattered. It is not conducive to integrate and use the cloud data. In addition, research on description of heterogeneous data resource lacks semantic information, hard to identify data resource and information extraction.

(2) The problem of discovery and matching on heterogeneous data resource. Currently, there is no mature data resource discovery technology adapted to dynamic changes in cloud environment, as well as lacks resources dynamic matching mechanism that meets the requirement of quality of service.

(3) The issue of heterogeneous data resource dynamic management. Cloud computing could provide variety of different information services for enterprises' users, and the cloud resources are constituted by a number of interconnected sub-resources. The research on sub-resources scheduling lacks considering the user's individual requirements [8].

(4) The problem of real-time monitoring of heterogeneous data resource. Currently, in cloud computing environment still lacks reliable mechanism that could timely detect and rapid diagnose the risks that may exist in cloud environment.

(5) How to access and retrieve heterogeneous data in cloud environment. The cloud data resource differs from traditional relation database systems, so data resource connection, data structures and retrieval methods are different, so it is necessary to design the overall plan of heterogeneous data connection and query solution.

(6) The issue of heterogeneous data resource collaborative management. As each node in cloud may contain partial information, so it is important to consider multiple data resource collaborative operational problems [9]. The first thing should be considered is how to transform the heterogeneous data into a unified format. Second is collaborative management of data resource, each node can join or leave the cooperative system without affecting other node.

(7) The security issue among the data resource in cloud computing. Data integration in cloud computing should have a reliable security mechanisms to sure the normal operation of the entire system.

4. The Heterogeneous Data Resource Management Model in Cloud Environment

This paper presents a novel heterogeneous data resource collaborative management model which absorbed from the multi-resources management in distributed environment research result. The architecture of the model is shown in Figure1, includes physical storage layer, data resource network layer, data conversion layer, data management layer, security management layer, data integration and application layer.

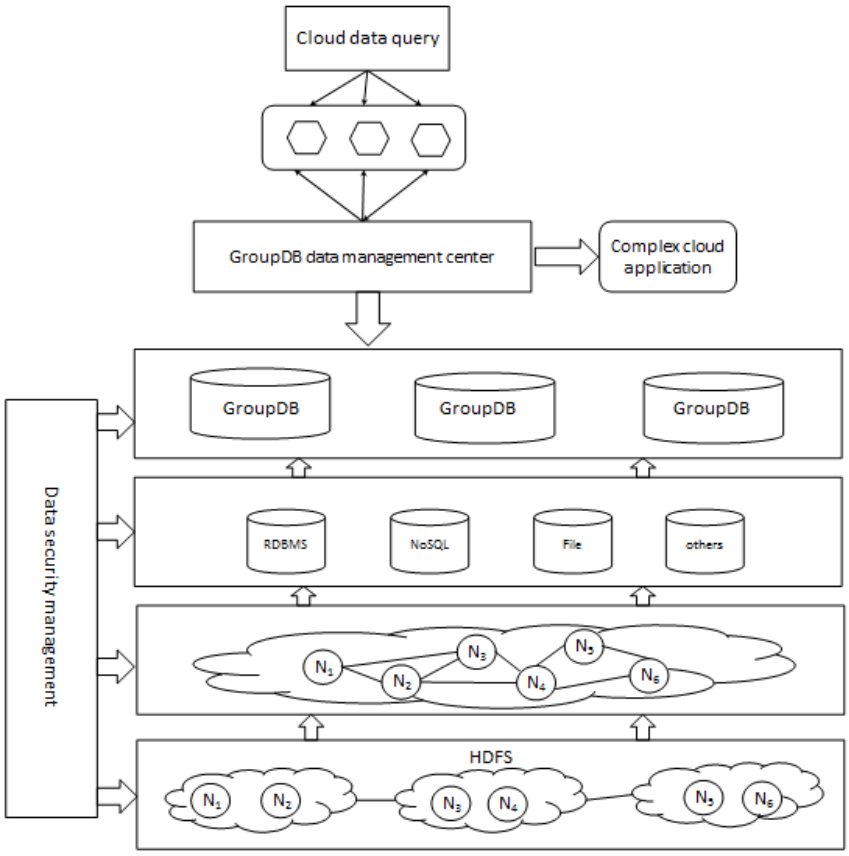

Figure 1. Heterogeneous Data Resource Collaborative Management Model

The heterogeneous data resource collaborative management model includes the following components:

(1) Physical storage layer. This layer is in the bottom of the model, stores all of heterogeneous data in different cloud nodes, constitute a very complex storage status, including traditional relational databases, NoSQL databases and all types file, *etc*.

(2) Data resource network layer. Cloud data is stored in a complex heterogeneous environment. In order to manage and use cloud storage services, data resource network layer

abstract all the physical nodes into logical nodes to form a mass storage network for different cloud nodes, to provide the basis for collaborative management.

(3) Data conversion layer. Data resource network layer includes variety of virtual data resource nodes, each node may store various types of data. In order to unify the data resource format, transform various data format into GroupDB database center's format which eliminated semantic ambiguity.

(4) GroupDB data management layer. GroupDB data management center manages all GroupDB data, includes data mapping, collaborative management, data integration and data fusion, *etc*. It could provide basic services and monitor the changes of cloud data.

(5) Security management layer. This layer responsible for the security of overall architecture, includes trusted monitoring and cloud resources certification center. Through trusted monitoring to ensure the data release, access, use, and storage in a secure environment, and certification center can ensure each participant is credible.

(6) Data integration layer. This layer integrates data from lower layer according to the requirement of application systems, could provide complete data. When acquire the requirements from the users, obtain data from the GroupDB data management center, and then integrates data from different cloud node for users.

(7) The application layer. This layer receives user's uniform query, and then nondestructive decompose the query, form the logical operation for the GroupDB data management center.

5. Implementation of the Model

5.1. Physical Storage Layer

Physical storage system includes data center cluster, enterprise cluster and ordinary servers. The enterprises' business data stores in physical storage system. Currently the mainly distributed file system is Google's GFS and the open source HDFS file system that based on GFS. This paper adopts HDFS as distributed file system, lets massive data distributed in different cloud nodes.

5.2 Data resource network layer

The data resource in cloud are abstracted into node, all the nodes constitute the data resource network. Data nodes increases, failure will cause the data resource network changes. In order to update the huge data source network automatically, we use the following algorithm that shows in Table 2 to maintain the network data sources dynamically.

Table 2. Data Resource Network Generation and Updating Algorithm

Input: new node n; Output: data source network w;
1. Scan “heartbeat” XML, if n==0, and no node failure, go 8; if n!=0, go 2; if node failure, go 3; 2. for(i=n;i>0;i--) locate the position, and find its neighbor, add all the edges, go 4; 3. for(i=n;i>0;i--), locate the position, and find its neighbor, delete all the edges, go 5; 4. Calculate the load of the node, and submit the load of the node to resource migration algorithm to get the node's actual load, go 7. 5. Calculate the failed node's data resources, includes the resource name, quantity, etc., go 6; 6. According to the content of 5 , calculate the resources needed and it’s amount, the results submit to resource migration management and control algorithm; 7. Monitoring the added and failed node, go 1. 8. return w, End

The balance of data resource network should consider the balance of storage in cloud environment, avoid some nodes too busy and other nodes too idle, the network load balance algorithm as following:

Table 3. Data Resource Network Load Balancing Algorithm

Input：Data resource network w; Output: Data source network load n;
1. for(i=0;i<m;i++) //m is the amount of cloud node; Calculate the current node load n_i; 2. Gets the current data source network total load n= $\sum n_i$, and calculate the average load of each node $\overline{n}$; 3.if n $\approx \overline{n}$ the node does not send any signal; 4. If n<< $\overline{n}$ the node is marked as 0 indicates that the node can receive migrated data from other data resources; 5. If n>> $\overline{n}$ the node is marked as 1 indicates that the node does not receive migrated data from other data resources; 6. Weight was 1-5 until the data source network load balancing;

The data resource network load balancing algorithm ignores the small part of the network load imbalance, because if there are several nodes are imbalance, frequent migration of data may bring communication pressure.

5.3. Data Conversion Layer

This layer is mainly converts various data into a unified format and stored in GroupDB, the following is an overview of heterogeneous data conversion algorithm.

Table 4. Data Conversion Algorithm

Input: data resource n_i;
Output：GroupDB d;
1. if (the data resource is relational database)
set the first line of the flag: FLAG= 0, then read the data into GroupDB by row in the table;
2. if (it is a non-relational database)
set the first line of the flag: FLAG=1, then read the data into GroupDB by column in the table;
3. Repeat 1-2 until all the data in data resource network is stored to GroupDB;
4. End;

5.4. GroupDB Data Management

GroupDB is a distributed database in cloud environment, the system is mainly to store and manage data from various cloud nodes. GroupDB is a distributed, sparse, and the ranks of the hybrid storage database system. It could store data by rows or by columns. The data model can be presented as Figure 2: flag (r/w) is the storage flag, when the flag (r/w) = 0, it indicates that GroupDB data management system adopts row storage, and the flag (r/w) = 1 means using column storage.

Flag(r/w)	$Data_1$	$Data_2$	……	$Data_n$	Flag(r/w)

Figure 2. GroupDB Data Storage Mode

5.5. Data Security Management Module

Data security is one of the basic security issues in cloud computing, more and more enterprise migrate their business and users data into cloud. Some analysis shows that data security is one of the biggest obstacles to migrate enterprise applications to the cloud computing. At present, cloud computing security issues have been gotten more and more attention.

The data security management module mainly consists credible monitoring for all resources,ensure that resources are not subject to external attack. Figure 3 shows the detail of the module.

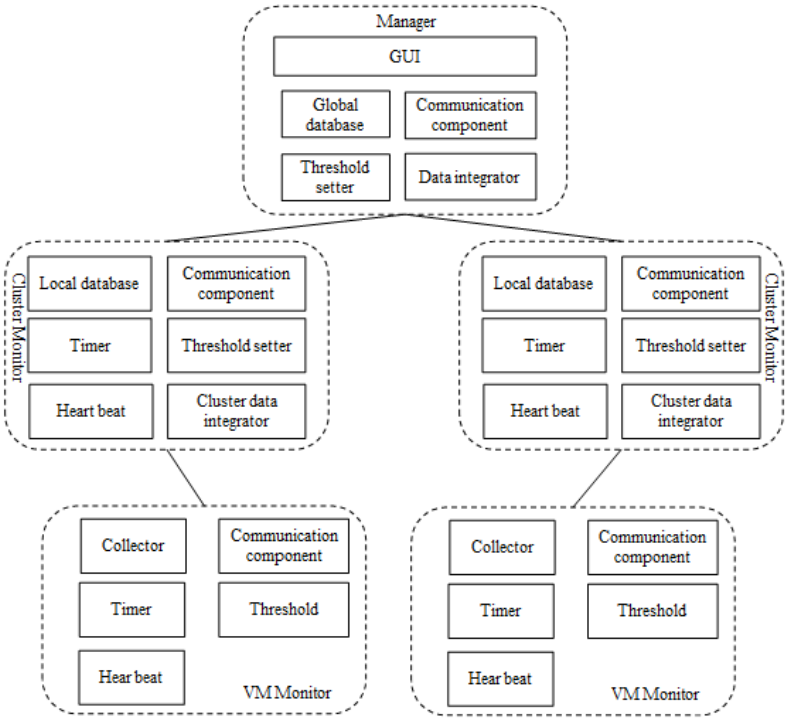

Figure 3. Data Resource Credible Monitoring Module

Figure 3 shows that VM Monitor module is responsible for monitoring the cloud node resource and status information. Cluster Monitor module captures the virtual machine resources and status information through VM Monitor module, and integrates the monitoring information, reports to the Manager modules.

Manager module periodically collects monitoring information of the cloud nodes through the Cluster Monitor module, the pseudo-code algorithm is shown in Table 5. VM Monitor module decides whether to send monitoring information to Cluster Monitor module when the monitoring value exceeds a preset threshold, the pseudo-code algorithm is shown in Table 6.

Table 5. Periodically Collection Monitoring Algorithm

Input: monitoring objects O_i , monitoring objects sets Msets(O_i); // $O_i \in$ Msets(O_i); Output: val(O_i); // Resource monitoring value of each node;

1.for(O_i in every cloud node)
2.If(O_i turn to reports)
3.While(node $\in$ Msets (O_i))
{
MonistorMessage:=val(O_i);
MonistorMessage send to O_i;
}
4.Integrated all the MonitorMessage and send to Manager.
5.Reset the reporting monitoring information cycle of O_i;
6.return val(O_i);
7.End;

Table 6. Monitoring Node Abnormalities Reports Algorithm

Input:D_i,W_i; // D_i presents the threshold of node i, W_i presents filter window size;
Output: Exception information e;

1.$t_s(w) = t_e(w)=0$; // filter window start and end time;
2. if($v_i(t) \geqq T_i$)
3.{
4. record the monitoring data of node i;
5.if(t> $t_e(w)$)
6. {
  $t_s(w)=t$;
  $t_e(w)= t_s(w)+ W_i$;
  e_s=monitoring data of node i;
  return es;
 }
7.else if (t= $t_e(w)$)
{
$t_s(w) = t_e(w)=0$;
if(val(w)>T_i)
{
e_s=monitoring information of node i;
reset w_i;
return e_s;
}
Else reset w_i;
}
}
8.End;

6. Experiment and Performance Analysis

In order to verify the heterogeneous data resource management model in cloud environment, the model was applied to a manufacturing enterprise. The experiment based on historical data about 18.2G, using the Hadoop platform, MapRedeuce distributed computing mode, data block size is 64MB, data copies is 2, build experimental platform. When only one data node running task, the running time is 1788.23s, and when there are eight nodes, only a 204.73s. As can be seen, the cloud computing environment can be large in magnitude increase system capacity for data processing. Figure 2 shows the relationship between nodes and the time has been running.

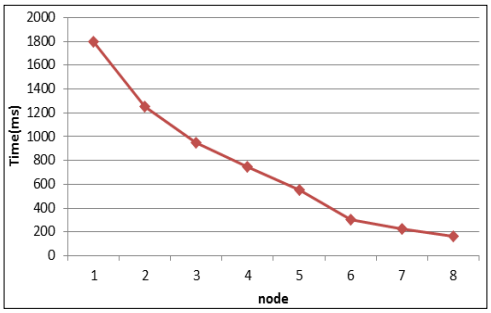

Figure 4. Nodes and Corresponding Time

In addtion, we compare the filtering windows algorithm with the baseline monitoring method, statistics the total number of abnormal within three cloud nodes. Figure 5 shows the collection frequency is 1(data collected once every 1 second), the total number of alarms that generated by monitored node. Figure 6 shows the collection frequency is 5(data collected once every 5 second), the total number of alarms that generated by monitored node.

As Figure 5 and Figure 6 shows that w indicates the window size used by filtering window algorithm. When w=2 and w=4, the periodically collection monitoring algorithm reduces the number of abnormal alarm significantly. From the test results we can find that the value of c far less than baseline method, and with the increase of the value of w, the increasing trend of c become slow. If w is too large may cause failure because could not detect the abnormal. In general, w value setting should be collected considering the size of the frequency and the threshold value, based on the sensitivity needs to be monitored to set the node.

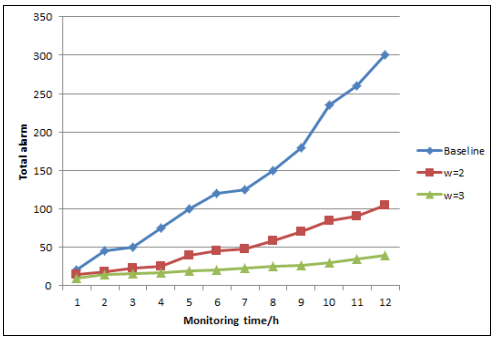

Figure 5. Compare Baseline Monitoring Method with Abnormalities Reports Algorithm

7. Summary

The cloud computing platform has good flexibility characteristics, more and more enterprise management systems are migrated to the cloud platform. This paper proposes a kind of heterogeneous data resource management model. This model implement massive resources storage, massive storage network' generation, update and balance of workload, proposes the security management and monitoring methods. The model proposed by this paper gives a novel solution to the heterogeneous data resource management and application in the cloud. Experiments show that the proposed model has a good performance in massive data resource management. Finally, the regulatory model will be applied to a manufacturing enterprise, this paper has laid a foundation for further research in massive data resource management.

Acknowledgements

This paper supported by the National Key Technologies R&D Program No. 2012BAH54F04 and the Natural Science Foundation of Shandong Province of China under Grant No. ZR2010FM033.

References

[1] J. Shute, M. Oancea and S. Ellner, "F1: The fault tolerant distributed RDBMS supporting google's ad business", Proc of SIGMOD, New York: ACM, **(2012)**, pp. 767-778.
[2] G. DeCandia, D. Hastorun and M. Jampani, "Dynamo: amazon's highly available key value store", Proc of SOSP, New York: ACM, **(2007)**, pp. 205-220.
[3] J. Baker, C. Bond and J. Corbett, "Megastore: Providing Scalable, Highly Available Storage for Interactive Services", Proc of CIDR, **(2011)**, pp. 223-234.
[4] L. Rubao and X. Zhiwei, "Exploiting Stream Request Locality to Improve Query Throughput of a Data Integration System", IEEE Transactions on Computers, vol. 58, no. 10, **(2009)**, pp. 1356-1368.
[5] K. C. Birman and G. van Renesse, "Toward a cloud computing research agenda", ACM SIGACT News, vol. 40, no. 2, **(2009)**, pp. 68-80.
[6] P. Mell and T. Grance, "The NIST Definition of Cloud Computing", National Institute of Standards and Technology, **(2011)**.
[7] J. Dean and S. Ghemawat, "MapReduce: a flexible data processing tool", Commun ACM, vol. 53, no. 1, **(2010)**, pp. 72-77.
[8] T. Hirofuchi, H. Nakada and H. Ogawa, "A live storage migration mechanism over wan and its performance evaluation", VTDC'09. Barcelona, Spain: ACM, **(2009)**, pp. 67-74.
[9] L. Zhen, Y. Fang-Chun and S. Sen, "Fuzzy multi-attribute decision making-based algorithm for semantic web service composition", Journal of Software, vol. 20, no. 3, **(2009)**, pp. 583-596.
[10] K. Shvachko, H. Kuang, S. Radia and R. Chansler, "The Hadoop distributed file system", Proc. of the IEEE 26th Symp. On Mass Storage Systems and Technologies(MSST), Lake Tahoe: IEEE, **(2010)**, pp. 1-10.

Authors

Tao Sun, PhD. Candidate, his current research interests focus on data management for cloud computing, data integration, BPM. Email:suntao0906@163.com.

Xinjun Wang, professor, Ph.D. supervisor, his research interests include database, cloud computing and data integration.