ATLANTIS PRESS

Advances in Intelligent Systems Research (AISR), volume 145

2017 International Conference on Electronic Industry and Automation (EIA 2017)

The Research of Relational Database Query Processing Based on Cloud Platform

Wei GU*

College of Information Technicology

Shanghai Jianqiao University

Shanghai, 201319, China

E-mail: guwdx@126.com

+* Corresponding author

***Abstract*—Big Data is now increasing rapidly, It need for more servers to handle large amounts of data, resulting in a number of different ways to improve the operation of data processing time, Cloud data processing platform is the most popular way of operation, Non-structured data using a cloud environment, in the form of key-value store, but because many enterprises use relational database structure to store data currently, and therefore cannot directly migrate from these databases to the cloud platform. In this paper, based on the cloud platform relational database, relational database with the new data separation and polymerization techniques, query processing optimization algorithms using MapReduce architecture to build a relational database query processing mechanism under the cloud platform.**

Index Terms*—cloud platform; query processing; relational database

I. INTRODUCTION

Cloud computing is the most popular online services, it makes many different network servers cooperate with each other to provide better service, Cloud platform includes a number of different servers and networks, their stored data access through the network in the world. Cloud computing can run at high speed because the data storage, processing dispersed on different servers to perform. Many cloud platform environment is the same amount of data into the module, dispersed storage on different servers, when reading or processing data, multiple servers simultaneously, thereby increasing the data processing speed.

This massive decentralized data processing techniques, data processing and mainstream way is very differently from relational database systems currently in use. It is not in accordance with the provisions of the data into two-dimensional table, but stored with Key-value. This storage method according to the key, the data is distributed to different servers, so the structure is relatively simple, suitable for distributed processing, so that faster data read and write speeds. But the cloud of data storage lacks correlation features of relational database. Such as it cannot easily get the total number of data, Unable to use the connecting operation, You cannot use SQL language to query data. But relational database is built on a centralized, high stability on the basis, lack of flexibility and capacity expansion for the current cloud environment. Due to most of the enterprises use relational database management systems to manage data currently, this article is going to discuss the content of how to play the advantages of both, use relational database in the cloud platform. That is to discuss how to construct a relational database in the cloud platform, study data segmentation and aggregation technology of relational database and high-performance relational database query processing technology. And consider cloud platform server dynamically increase and decrease of environment issues.

II. THE RELATIONAL DATABASE DATA PARTITIONING ALGORITHM

According to the above, if you want to use a relational database in the cloud platform, you must consider how the data in a relational database is divided into different servers, That is, each table is divided into different storage server, so how to split the data table to improve query performance is very important. According to relational database data relationship between tables, the attributes of each table can be divided into predicate attributes and target attributes [1]. Predicate attributes include join attributes and selection attributes. Join attributes is association attribute between the tables, in relation to the property must be used when making connections between the tables, Therefore split to data queries based on these Join attributes can avoid these problems that require a lot of data transmission times because of query a lot of data that need to connect multiple attributes stored in different servers; The data can be divided according to certain values or specify the range by way Hash method [2]. Selection attributes refers to the condition of the property during the data query, data can be obtained under the conditions corresponding attribute. Target attributes is displayed data in the screen after the user queries[3].

The paper are discussed from five aspects of database segmentation algorithms including in data able partitioning, the table split the order, the relevant attributes and so on.

A. Table Partitioning

When the relationship table is divided, there are two kinds of division method including Range and hash, the

choice of which as the split mode, depending on the partition is a valid way split. First let the join attributes as a basis for the table partitioning, using hash or Range manner to produce different values divided[4], with the same value into a database. For example, there are two tables of Customers and Orders, cid is the same property of the two tables that is their join attribute, now use cid attribute of Customers table to cut up data by hash mode, the same hash value of the data in the two tables are assigned to the same database.

B. Table Partitioning Order

When for multiple relational tables with associated attribute data is divided, the choice of which table join attribute as the starting split basis will greatly affect the split mode and query efficiency. Then the above examples, such as add a Products table and pid attributes of the table as the join attribute[5]. If the Customers table or in cid as a basis for segmentation splitting hash ways, different Orders cid corresponding to the pid have duplicate product data, That is, different consumers can order the same products purchased. Products resulting data will certainly be some duplicate data in different databases. But if at the same time to the two properties cid and pid Orders of hash way split, you can avoid repeat data of Products table. Therefore, how to assess which of the join attribute data as a table of segmentation basis is very important.

C. Selection Attributes and Target Attributes

The relationship table is divided, first according to join attributes as the split criterion, on the other side ,how to assigned attributes that the remaining selection attributes and target attributes to a different server , it will also affect the output of the final result. Therefore, we will consider the two properties put on the same server and on a different server of the situation to see the final result of the case.

D. Outer Join and Inner Join

Data table after being cut, because data is distributed on different servers, if you want to connect more than one tables, using our segmentation algorithm, you can handle the problem of connection property is null through part of the extended process, so that you resolve the problem of outer join and inner join, while not required server communication between excessive costs.

E. Insert Operation

If the data is written to the database, the new data should be placed in which server and whether the new data will lead to a large number of data re-segmentation, the problem should be considered when insert data to database[6].

From the above four aspects analyzed data relational tables segmentation algorithm, follow on a specific example to illustrate specific algorithm. Firstly, relational tables divided manner, the data that has association attributes between the tables are classified based on certain values by way hash function or property function according to the predicate attributes , after the split the association attributes between the tables laterally, the same server can be connected between the operating table and the table of their own. Using the way of hash join, the values of association attributes between tables will be hash, according to different values after the hash generated by dividing the data to a different database server, If property values associated with the two tables of the same data is assigned to the same server, when connecting the two tables the same part of the common property values are assigned to the same database, you can avoid data transmission problem. You can also adjust the number of servers by hash value. Figure 1 shows the results of its hash join.

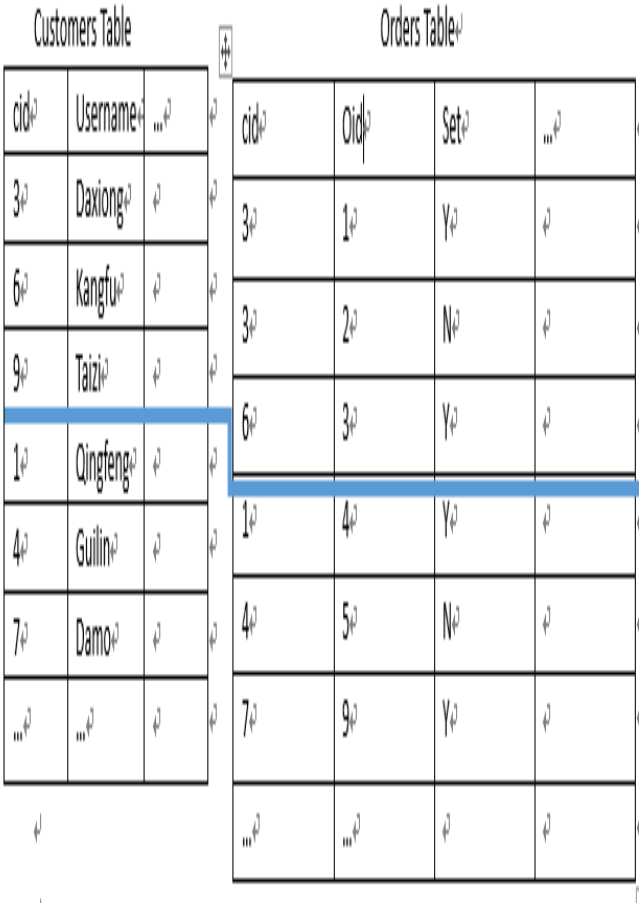

Figure 1. Hash join segmentation results

Another way you can use segmentation associated property to some value as a reference split in the data within the same range of values will be assigned to the same database server, both methods can even be assigned to each of the data in the database, and can achieve the purpose of parallel processing. Figure 2 shows the results of its range join.

When association is divided between multiple relational tables, how to split it? There are Customers, Orders and Products three tables, three tables and two contain two

common attributes cid and pid, during data partitioning, the property sheet from which it split? If you start from the Customers cid, the first 100 records order by cid and Orders table data corresponding cid value of all the records into the first server, which will belong to the relevant customer data before the top 100 are assigned to the same database, you can simplify the query. However, if the split in this way, recording some of the Products table occurs it will be repeated in different servers. Therefore, we must use cost model has been evaluated best cutting table order.

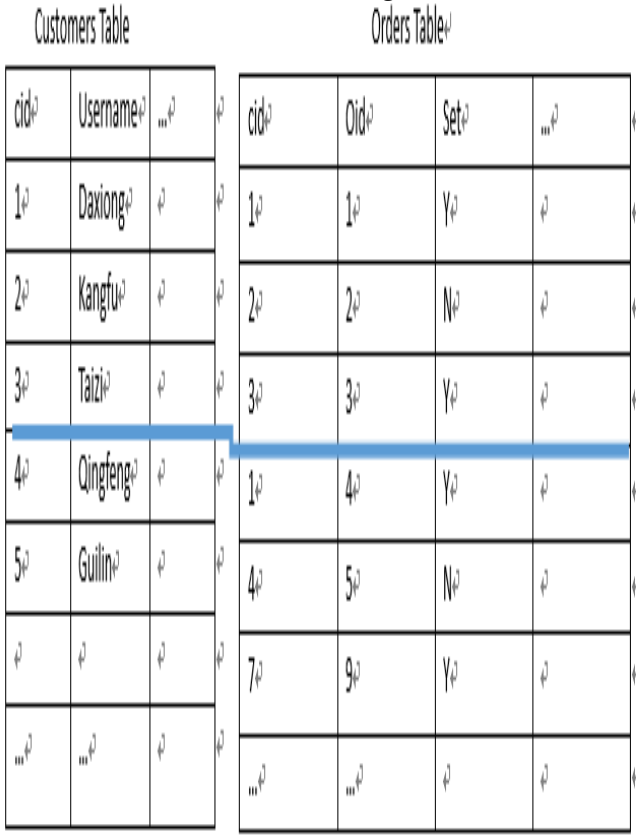

Figure 2. Range join segmentation results

For data distribution predicate attributes of target attribute the need mentioned above, there are two ways: First, the predicate attributes and the target attribute separately to a different server database, so that the benefits can be carried out to improve the efficiency of the data connection, But after the connection is completed, all the tables with id attribute (O_id) need to associate the table with target attribute, this requires increases extra execution time. That is when cut up the table through vertical cutting method; the table of target attributes need to add O_id attribute of original table as key for together other tables, as figure 3 shows. Second, not vertical fragmentation, although it will increase the connection time, but will reduce the integration time. These two methods will be compared in the experiment, it can indicate whether further vertical cutting data table can improve query efficiency[7] [8].

For processing outer join and inner join query methods, the main difference is that associated with the result, for external connection, there may be situations null value, but the two operations can still perform normal way in our division.

For the new data is added to the database, the data can still be attributes associated property relations between tables by way hash function or range predicate according to the same manner into the server database, However, when a large amount of data for the new data may be considered a cost model to re-define the data division process.

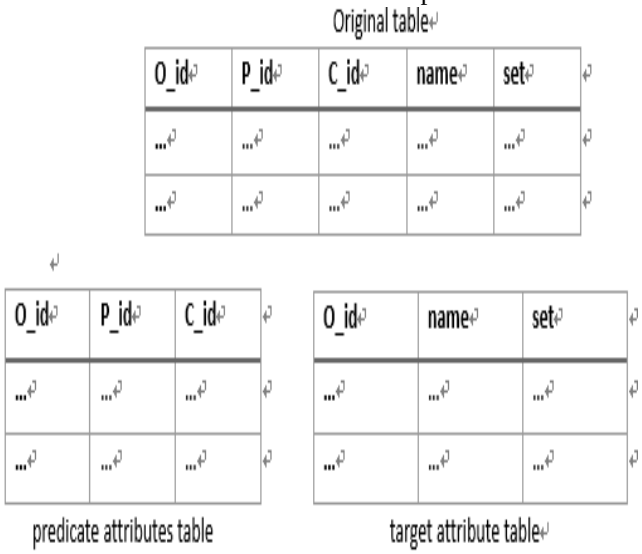

Figure 3. Vertical cutting the table

III. EXPERIMENT AND CONCLUSIONS

The author carried out an experiment of a company work order system in order to verify the above algorithmic analysis. The experimental environment using hadoop to do data analysis and application, use java as a programming environment. Set up a primary node and a plurality of secondary nodes to execute the query processing: When the primary node receives a user query requirements, it will first execute the query conversion action, the query is converted to secondary nodes with the same number of sub-queries, and these queries are assigned to the minor child node to perform data query processing, When each secondary node to complete the work, and then pass the results of the primary node to do the integration and operation of the query results are sorted, and finally converted to the query results users expectations.

This experiment by work order system common type of query is analyzed to understand workload data query and analysis of these data is adjusted by dividing strategy, establish cost model to evaluate the data query performance. Will also in cloud platform, using the proposed segmentation data query processing techniques, and compared with the original query, look through

segmentation query efficiency is better than the divided original query efficiency.

This paper presents a set of cloud-based relational database data segmentation technology platform architecture, By dividing relational tables, relational tables split the order, Selection attribute and Target attributes, insert a new database manipulation and analysis, will make the most appropriate relational tables segmentation strategy, which is conducive to parallel processing of queries, And through experiments to prove that the algorithm can perform efficient data queries, so as to commonly used relational database to provide a reference to the large data conversion.

REFERENCES

[1] X.zhang,J.Ai, " An Efficient Multi-Dimensional Index for Cloud Data Management. Proceedings of the International Workshop on Cloud Data Management", PP: 17-24, Novermber , 2009.

[2] P.xiong, Y.Chi, "Intelligent Management of Virtualized Processed for Database System in Cloud Environment", Proceedings of the International Conferences on Data Engineering, PP: 87-92, April, 2011.

[3] GATTAL A, CHIBANI Y, "Segmentation and recognition strategy of hand written connected digits based on the oriented sliding window", 2012 International Conference on Frontiers in Handwriting Recognition (ICFHR), Bari, PP: 297-301, 2012.

[4] Dong Y, Shen D, Nie T. "Discovering relationships among data resources in data excavate", 6th Web Information Systems and Applications Conference, Xuzhou, PP: 76 - 81, 2009.

[5] John C, Du chi, Lester W, Mackey, Michael, Jordan. "On the Consistency of Ranking Algorithms", Proceedings of the27thIntemational Conference on Machine Learning, PP: l1-14, 2010.

[6] Qiang lin Zhang, "Research of context aware system based on Semantic Web ", Beijing Jiaotong University 2010.

[7] TAL l, MUNTEAN G, "User-oriented cluster-based solution formultimedia content delivery over VANETs", 2012 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB). Seoul, PP: 111 –11 5, 2012.

[8] Song S, Chen L,Yu P S. "On data dependencies in data spaces", EEE 27th International Conference on Data Engineering(ICDE). Hannover, PP: 470 - 481, 2011.