TELKOMNIKA, Vol.12, No.4, December 2014, pp. 1039~1044
ISSN: 1693-6930, accredited A by DIKTI, Decree No: 58/DIKTI/Kep/2013
DOI: 10.12928/TELKOMNIKA.v12i4.436

Optimization Research of the OLAP Query Technology Based on P2P

Chunfeng Wang
Modern Education Technology Center, Yancheng Institute of Technology,
Yancheng 224051, China
e-mail: wcf@ycit.cn

Abstract

With the increasing data of the application system, the fast and efficient access to the information of support decision-making analysis has become more and more difficult. At the same time, analysis of the data is no longer on a single server or a single enterprise data, but on multiple servers, multiple departments or multiple enterprise data. So, the original OLAP technologies have also revealed many shortcomings. Although we can use index technology optimization method to improve the performance in a certain extent, but for the continually expanding amount of information in data warehouse, its performance is still the problem needed to be solved. Using the method of P2P network technology and OLAP storage query and query method, the paper has constructed a distributed P2P-OLAP network model and the model distributed stores the data and centralized manages the node. Next, the paper has put forward the storage and sharing scheme of multidimensional data, OLAP query scheme based on collaboration support. The idea of the scheme is that the query analysis is done by the coordination and cooperation of OLAP nodes. Finally, the paper has shown that the scheme can effectively improve the performance of decision analysis by the experiment.

Keywords: *P2P, OLAP, query optimization, multidimensional data set*

1. Introduction

With the increasingly fierce competition of market, the information plays a more and more important role for the survival and development of enterprises. At the same time, along with the extensive application of database and data warehouse technologies, the enterprise information system with the accumulation of time will produce a large amount of data [1]. That how to get useful decision information from the complicated data environment and how to make the right analysis and decision-making have become a crucial link for the survival and development of enterprises. Online Analytical Processing (OLAP) system can help users to analyze the dimensional structure of commercial information efficiently and easily. It is fast software technology of accessing and analyzing the specific on-line data for specific issues. It tries to convert mass data in data warehouse to useful decision information, so as to realize the data analysis and decision, then to help enterprises to achieve the decision.

In recent years, with the further research and application of OLAP technology, OLAP technology has made considerable development. The decision analysis of the traditional client/server mode and the widely used browser/server mode decision can provide effective support for the quick decision analysis and trend analysis of massive data [2].

However, with the expansion of network scale and the enterprise data, the existing methods have been shown some deficiencies. Analysis of the data is no longer on a single server or a single enterprise data, but on multiple servers, multiple departments or multiple enterprise data. Especially in the current P2P network technology continues, the decision analysis by data coordination and cooperation stored in a multi node has become possible. Different from the traditional C/S mode, P2P technology can organize structure by the way of network node in the application layer, which will weaken the server role or even cancel the server [3]. The node in the P2P system is both client and server. In an ideal P2P system, each node can acquire the same rights and obligations, equality exchange data and provide services. The communication of each node is effective without the control of server. Relatively speaking, the biggest advantage of P2P structure is that the services are distributed to each peer of the network [4]. So, the P2P network can provide effective services even when one node is failure or abnormal in the whole network. Distributed storage based on P2P is one of the most

Received September 12, 2014; Revised November 3, 2014; Accepted November 20, 2014

important application models and is very suitable for cube storage [5]. Therefore, in the P2P network environment, how to respond to OLAP query and the establishment of multidimensional data more perfectly has become a research focus of many scholars and experts.

Based on the above technology development and the P2P distributed storage construction, this paper has constructed an OLAP network model and formulated the related query scheme. The scheme can set data sharing by using the multidimensional data of OLAP nodes, complete query analysis by coordination and cooperation, complete the dynamic join and exit of the OLAP node.

2. Definitions of Related Concepts

2.1 OLAP

The OLAP technology is designed to support the complex analysis operation and the emphasis is query analysis demand for user, and then helps them to quickly and accurately grasp the overall situation, market demand and development trend in their respective areas, so as to make the right choices [6].

OLAP can help users to observe information from multiple angles and aspects in the usual way of thinking. OLAP can high efficiency deep uses the historical data for services. Its core concept of OLAP is "dimension" [7]. OLAP can meet the analysis demand of multidimensional environmental reports and queries and be called the multidimensional analysis tools.

2.2 P2P

P2P has a simple definition: "P2P is a kind of application. It makes use of the storage space, execution cycle and the content resources idled in the network". Between words, P2P is a distributed system located in application layer and each node can communicate directly by routing protocol in the P2P layer [8]. Each node with an object database (such as file, MP3, MPEG etc.) can query the object in the other nodes by logical connection of P2P layer.

2.3 Data Warehouse

The definition of data warehouse has many kinds. W.H.Inmon, the father of the data warehouse proposed that data warehouse is data set oriented to support management decision-making process, subject, integrated, changes with time in the "Building the Data Warehouse". Data warehouse allow the integration of various application systems and provide a unified support database for the analysis of historical data.

Data warehouse is a semantically consistent data collection and stored the information needed for decision making [9]. It has great significance to improve the efficiency of data storage and data processing capability. Users can be more flexible in analysis of data and information and can find the valuable information, then will bring huge benefits to the enterprise.

3. OLAP Query Technology Problems in the Environment of P2P

When users were analyzed when using the OLAP system will inevitably involves the fact table and dimension table join and aggregation number. A large number of join and aggregation operations of fact table and dimension table will be inevitably involved, when user analyzes the data using OLAP system. These actions are always the operations of very consuming system resources [10]. So when user sends a query operation, performance of OLAP system is not up to the expected user response requirement. Although we can use index technology optimization method to improve the performance in a certain extent, but for the continually expanding amount of information in data warehouse, its performance is still the problem needed to be solved.

In the environment of P2P, the basic of OLAP query is multidimensional data set. In the multidimensional data set, the storage is achieved by the multi-dimensional and multi-level way, and the aggregation records number of multiple granularities will occupy GB, PB space, the computing time is also very long [11]. Therefore, we must firstly solve the storage efficiency of multidimensional data set, so as to improve the analysis efficiency of OLAP. While the existing

P2P query analysis algorithm is completed by the OLAP server, the load of OLAP server is not been effectively reduce and the efficiency of the algorithm is greatly reduced.

The research of this paper will take all query service from the original server to each node in P2P-OLAP network for processing, so as to realize the cyber source sharing, network load balancing, and put forward the distributed query algorithm based on the multidimensional data sets. The research can improve the decision analysis capability of the whole system.

4. Optimization Design of the OLAP Query Scheme

4.1 Build of the Network Environment

Each node in the P2P network is both the service provider and service enjoy. The main objective of building P2P-OLAP network model is to achieve the sharing of multidimensional data sets based on semantic level, reduce the load of OLAP server, and rapidly complete the OLAP query request analysis of each node [12]. The deployment and operation flow chart of P2P-OLAP network is shown in Figure 1.

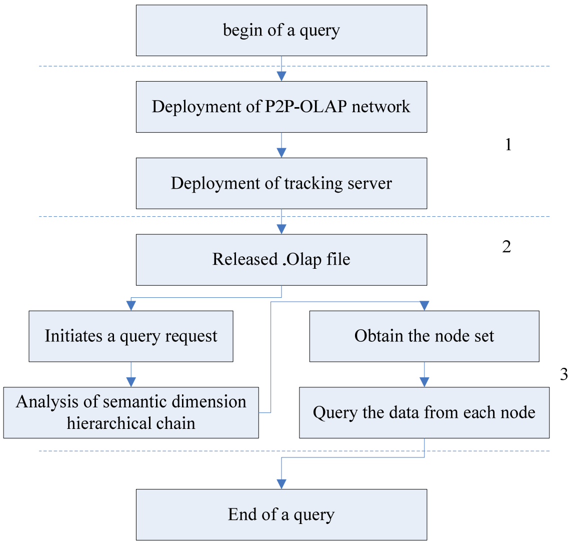
begin of a query
Deployment of P2P-OLAP network
Deployment of tracking server
1
Released .Olap file
2
Initiates a query request
Obtain the node set
Analysis of semantic dimension hierarchical chain
Query the data from each node
3
End of a query

Figure 1. Deployment and operation flow chart of P2P-OLAP network

The following is the three indispensable steps of building P2P-OLAP network:

(1) Deployment of tracking server

The process of sharing data needs a tracking server. It is mainly to assist a node in the network to access the information of other nodes and coordinate the information between the different nodes at the same time.

The interaction between server and the node is by the HTTP protocol. The node registers the information of the multidimensional data sets, IP address and port to the server. The connection between the nodes is established according to the registration information, and the tracking server will tell other nodes about the registered information.

(2) Released Olap file

The original node will create .Olap file to share the related information of multidimensional data sets and the address of tracking server. The system can upload the file to the Web servers. The all nodes in network have the same role and collaborating data analysis and decision making request in P2P-OLAP network.

(3) Sharing query between multidimensional data sets

Node gets .Olap file from the Web server, obtains the address of the tracking server address from .Olap file, and registers information to the tracking server according to the address. Firstly, node obtains registration information from the tracking server stored in the other nodes. Secondly, node can build the connection to other nodes according to the address location information and port information from other nodes. Finally, node can complete the query analysis of multidimensional data sets.

4.2 Distributed Query of Multidimensional Data Sets

This paper has used the CSMD-Tree structure as the storage mode of multidimensional data sets in the P2P-OLAP network node based on semantic dimension hierarchical chain. The multidimensional analysis of P2P-OLAP network makes the service processing to be distributed to each node, which requires a service method corresponding to complete the query and analysis, so as to realize the share of multidimensional data sets based on the semantic level in the P2P-OLAP mode. Query optimization scheme is shown in Figure 2.

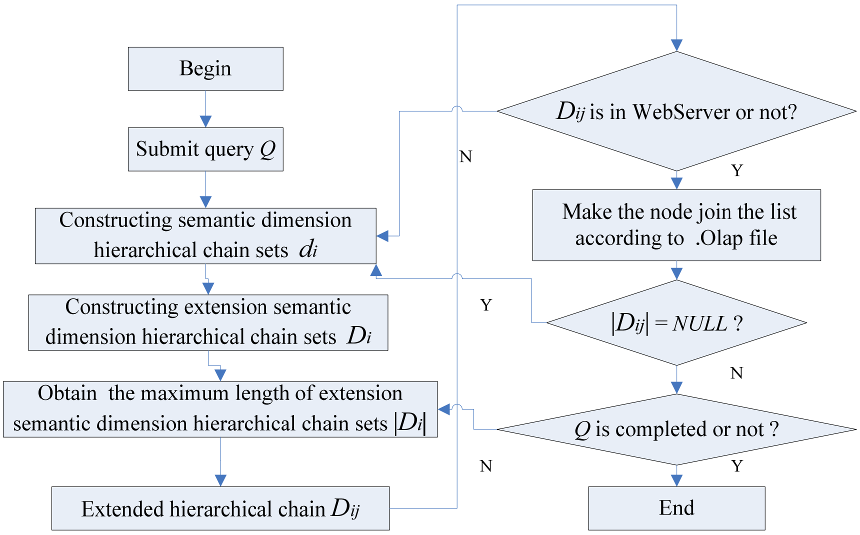
Begin
Submit query Q
Constructing semantic dimension hierarchical chain sets d_i
Constructing extension semantic dimension hierarchical chain sets D_i
Obtain the maximum length of extension semantic dimension hierarchical chain sets $|D_i|$
Extended hierarchical chain D_{ij}
D_{ij} is in WebServer or not?
N
Y
Make the node join the list according to .Olap file
Y
$|D_{ij}| = NULL$?
N
Q is completed or not ?
N
Y
End

Figure 2. Flow chart of query optimization scheme

The pseudo code of query optimization scheme as follows:

(1) A node (P) input query analysis statements, then construct semantic hierarchical chain sets (*di*) and extended semantic hierarchical chain sets (*Di*) .
(2) According to the semantic hierarchical chain sets, obtain the lengths sets of the semantic extended hierarchical chain | *Di* |.
(3) Calculation of the all lengths is assigned to |*Dij*|, the value expresses the maximum length of semantic extended hierarchical chain sets.
(4) According to |*Dij*| and *Di*, obtains the extended hierarchical chain *Dij*.
(5) Search the hierarchical chain matching with *Dij* from the every .Olap file on the Web Server.
(6) If successful, the node from the .Olap file will be added to the list.
(7) If |*Dij*| is NULL, remove |*Dij*| and *Dij* from set | *Di* | and set *Di*, the program will jump to (2). If |*Dij*| is not NULL and the node (P) query results will be turned into "not finish" state, the scheme will change *Di* to *di.*
(8) If the query is successful, the scheme can output the query results, and then exit the program.

From the above, the query optimization scheme in this paper can realize the match query according to the extended semantic dimension hierarchy chain, make the node shared with the form of multidimensional data set in the network and cooperated with other nodes to complete the OLAP query analysis.

5. Experimental Analysis of OLAP Query Optimization Scheme

The following test analyzes the performance of OLAP query optimization from two aspects. One is the time complexity; other is the query analysis rate. The experiment build a Web server, a tracking server and the P2P-OLAP network stored two multidimensional data sets. The following test will compare two schemes: OLAP query optimization algorithm (denoted as S) and P2P pattern matching algorithm (denoted as D).

(1) Multidimensional data sets search the corresponding .Olap file from the Web server and download data when the number of nodes is up to n, which all need a certain cost of time. The experiment assumes that the number of nodes is in a certain range. The execution time of two algorithms is shown in Figure 3.

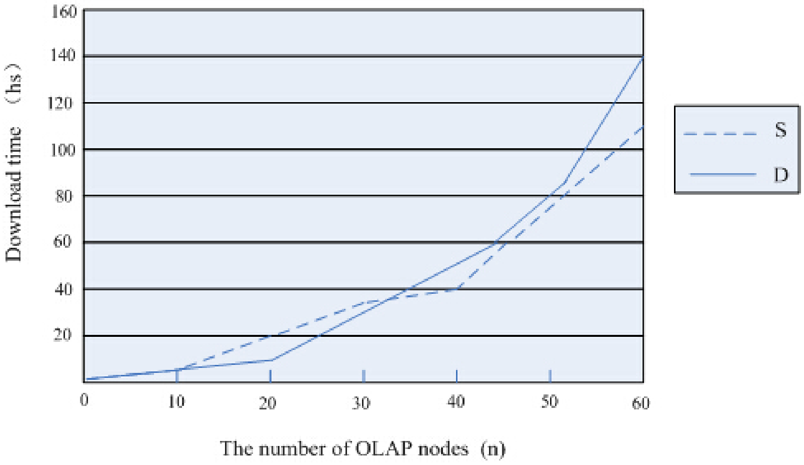

Figure 3. The time load comparison chart of two algorithms

From the above figure, we can know that the data tuple is the share unit in P2P-OLAP network and physicochemical treatment of multidimensional data set can greatly improve the access speed of network node.

(2) The query analysis rate of two algorithms is shown in Figure 4.

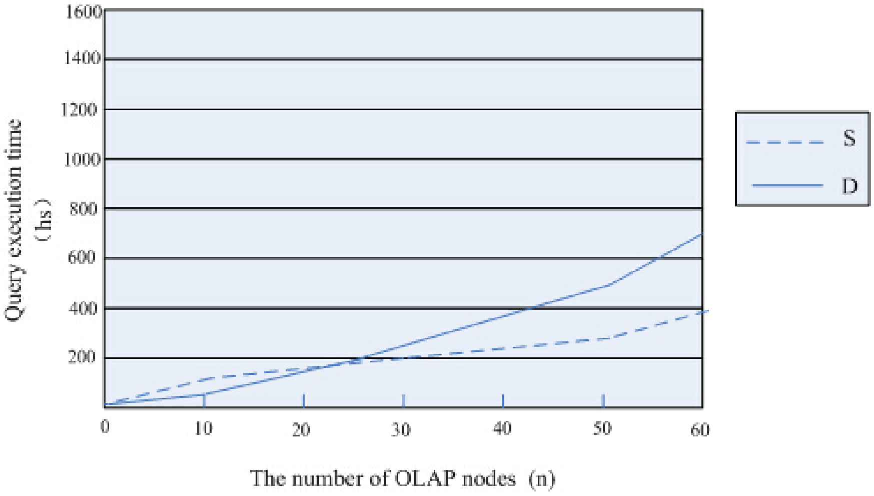

Figure 4. The OLAP query time Comparison chart of two algorithms

The analysis rate of S is not much difference with the D algorithm when the multidimensional data sets is only include some server nodes or there is fewer nodes in the network.

But when the number of nodes increases, the performance of S algorithm is much better than D algorithm. Of course, it is not difficult to understand that this is because the implementation of D algorithm is completed on the server, when the number of nodes increases, its performance is limited to the increase load of server.

6. Conclusions

OLAP has become a research focus of decision support. In recent years, with the study of P2P network technology in-depth, P2P-OLAP model is more suitable for the deployment and implementation of OLAP query system. How to improve the OLAP analysis query efficiency in the P2P network environment has become a critical problem. The main innovation of this paper lies in the following aspects:

(1) This paper has constructed the OLAP network model based on the P2P environment. The model distributed stores the data and centralized manages the node, which can obviously improve the efficiency of management.

(2) This paper has put forward the distributed query algorithm of multidimensional data set in P2P-OLAP network. The arithmetic idea is that the query analysis is done by the coordination and cooperation of OLAP nodes, which can obviously improve the query efficiency.

Acknowledgments

This work was supported by the 12th Five Year Planning Foundation of Jiangsu Province (Grant No.B-b/2013/01/012). It was also supported by the Modern Education Technology Foundation of Jiangsu Province (Grant No.2013-R24773).

References

[1] Golfarelli Matteo, Mandreoli Federica, Penzo Wilma, Rizzi Stefano, Turricchia Elisa. *A query reformulation framework for P2P OLAP.* Proceedings of the 20th Italian Symposium on Advanced Database Systems, SEBD. 2012: 147-154.

[2] Park, Nam Hun, Joo Kil Hong. Query processing on OLAP system with cloud computing environment. *International Journal of Multimedia and Ubiquitous Engineering*. 2014; 9(5): 169-174.

[3] Wang, Yi Fei. The web foreign language teaching research based on P2P technology. *Applied Mechanics and Materials*. 2014; 9(20): 132-136.

[4] Zeng, Degui, Geng, Yishuang. Content distribution mechanism in mobile P2P network. *Journal of Networks*. 2014; 9(5): 1229-1236.

[5] Gómez, Leticia Irene, Gómez, Silvia Alicia, Vaisman, Alejandro. Modeling and querying continuous fields with OLAP cubes. *International Journal of Data Warehousing and Mining*. 2013; 9(3): 22-45.

[6] Kumar Arvind, Singh Deepti, Sharma Vineet. *Achieving query optimization using sparsity management in OLAP system*. Proceedings of the 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques, ICICT 2014. 2014: 797-801.

[7] Weidner Martin, Dees Jonathan, Sanders Peter. *Fast OLAP query execution in main memory on large data in a cluster*. Proceedings - 2013 IEEE International Conference on Big Data. 2013: 518-524.

[8] Ciferri Cristina, Ciferri Ricardo, Gómez Leticia, Schneider Markus, Vaisman Alejandro, Zimányi Esteban. Cube algebra: A generic user-centric model and query language for OLAP cubes. *International Journal of Data Warehousing and Mining*. 2013; 9(2): 39-65.

[9] McCarthy Mitzi, He Zhen. Efficient updates for OLAP range queries on flash memory. *Computer Journal*. 2011; 54(11): 1773-1789.

[10] Zhu Yue-An, Zhang Yan-Song, Zhou Xuan, Wang Shan. Column-oriented query execution engine for OLAP based on triplet. *Ruan Jian Xue Bao/Journal of Software*. 2014; 25(4): 753-767.

[11] Cuzzocrea Alfredo, Gunopulos Dimitrios. A decomposition framework for computing and querying multidimensional OLAP data cubes over probabilistic relational data. *Fundamenta Informaticae*. 2014; 132(2): 239-266.

[12] Molina Carlos, Prados-Suárez Belen, De Reyes, Miguel Prados, Peña Yañez Carmen. Improving the understandability of OLAP queries by semantic interpretations. *Lecture Notes in Computer Science*. 2013; 8132: 176-185.