Send Orders for Reprints to reprints@benthamscience.ae

The Open Automation and Control Systems Journal, 2015, *7*, 1759-1767

Research of Distributed Query and Optimization Method Based on Metadata

Huaiyuan Wang*

College of Electronics and Information Engineering, QiongZhou University, Sanya, Hainan, 572022, China

Abstract: A method of distributed query based on metadata, which uses metadata to define and manage the virtual table containing key information of the data source, has been studied in this paper. Then, in view of the different data level, designed two different data solutions based on query and optimization, for applying to common data and huge data respectively. In common data query, using the virtual table, the syntax analysis tree and memory database was realized by; copying, moving, and dividing the branch from virtual SQL query syntax tree to make the query optimized. In terms of huge amounts of data query, Pig, Hadoop, Python is used to implement data query; by optimizing the Pig code, using multiple processes, processing file merging and file uploading or downloading in HDFS, making index on high frequency business and so on to achieve optimization of big data.

Keywords: Distributed, federated query, Hadoop, memory database, metadata.

1. INTRODUCTION

As the data and business became complicated, query data which meet the conditions has also become more and more complex. When you want to query a message from distributed data, programmers need to know all kinds of information about data such as data storage location, and storage structure. Programmers need to call a lot of interface to obtain the relevant data. This process takes much energy on programming and it requires that a programmer has high familiarity with the data interface. If providing uniform data programming interface to programmers is impossible, it will be shielding the backend access details, and then greatly improve the programming efficiency of programmers [1-3].

Entering the 21st century, the life sciences have developed rapidly. Scientists have tapped hundreds of types of biological databases to reserve research results for effective and efficient research and analysis. In the course of actualizing and accomplishing the Human Genome Project, biological data show a rapid increase of exponential series, thus the integration of the query for distributed heterogeneous autonomous databases becomes a major problem [4].

2. METHODS

We propose an integral scheme of data resource based on metadata in order to solve the current problem pertaining to query of biologic data. In our integral scheme, the metadata distributed in all databases will be integrated to one metadata database according to a multimodal standard. The integral scheme annotates the metadata using the ontology and actualizes the integrating query for biological databases through mapping the relationship among metadata, the definition of ontology, and the term of ontology. The final goal of the scheme is to solve the common problem of integrating data resources by building up a sharable integrating data platform, to form a database of a virtual center that faces the fixed topic, the focused metadata, and distributed basic data, and thus to support related researches in various fields as shown in Fig. (**1**).

Retrospect of the development and actuality of distributed heterogeneous spatial data integration technology, points out that the direction is seamless integration and advanced fusion, deeply studding on the widely used concerned technologies, such as geographic information sharing and interoperability, maps joint, map conflation, and spatial data fusion [5].

Anglicizing the usage scope of each technology, and any of them is half-baked for seamless integration. Considering the characteristics of distributed heterogeneous spatial data, the research work of seamless GIS, seamless spatial database, distributed GIS, and distributed multi-spatial database at home and abroad are introduced and discussed. It also summarizes the main kinds of existing spatial data center based spatial data integration patterns at home and abroad, pointing out the feasibility and necessity of these pattern to integrate spatial data seamlessly.

As shown in Fig. (**2**), studying various viewpoints of this problem, and the connection between these viewpoints, proposing the multi-layered semantic seamless integration model, points out that the seamless integration of spatial data is multi-level, multi-angle "stand-alone single-layer single-map"-liked transparent access and transparent application of distributed heterogeneous spatial data, and it's a systematic and comprehensive platform. Research on various spatial data seam problems is conducted, especially the seams came

*Address correspondence to this author at the College of Electronics and Information Engineering, QiongZhou University, Sanya, Hainan, 572022, China; Tel: +8613278695656; E-mail: wanghy@163.com

1874-4443/15

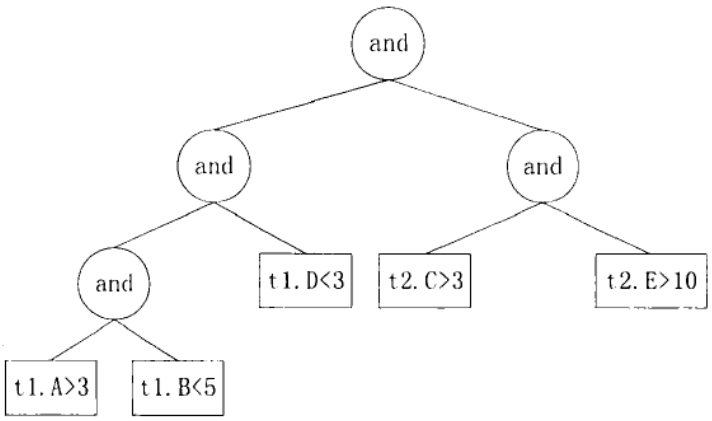

Fig. (1). Language tree.

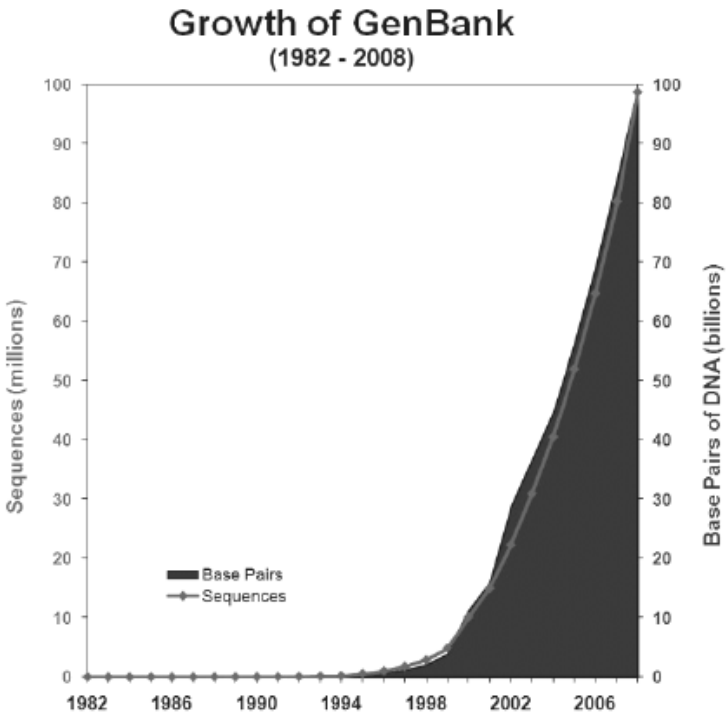

Fig. (2). Genbank increasing data.

from distributed heterogeneous environment. Anglicizing the cognition and modeling process of people to the real world geographic objects, points out that the continuity cracks and categorizes of information in the transformation from the real world to the geo-feature set world is the reason of the seam, and thus the idea of reverting the continuity and consistency back into the information has been proposed (Fig. **3**). Focusing on the top-level user view of seamless integration, the concept of a seamless integrated system model, data location and description of resources, and data containers seamless integration, these works provide the foundation of the research and design of the integrated framework and prototype system.

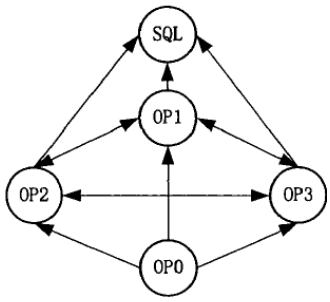

Fig. (3). Querying route.

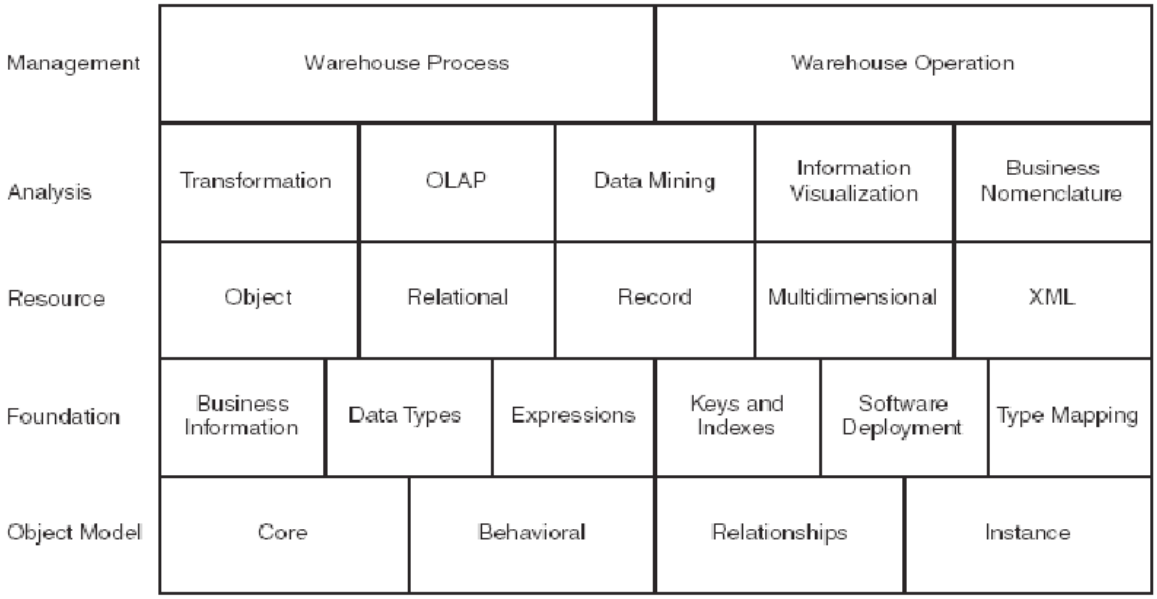

Fig. (4). CWM architecture.

The study deeply anglicizes the spatial data center architecture and its general patterns, [6] and discusses the feasibility to realize seamless integration framework based on the patterns of spatial data center (Fig. **4**). It researches and designs out a seamless integration system architecture, focusing on the key components, such as the framework, global catalogue system, global spatial index, multi-layer intelligent middleware, and so on.

It is proved that based on this mechanism a variety of seamless spatial queries can be done using the set difference operation in a high-performance manner [7]. The seamless integration archive mechanism is proposed to improve the performance. Distributed heterogeneous spatial data register mechanism is studied and designed. The data cleaning preprocessing method is used to set up the initial seamless version of the data. Based on the data's initial seamless version, the maintenance of topological relationship and seamless characteristic in the process of data editing and updating can be done with the consistency maintenance measures, to ensure the data is always in validate status.

The general process data query is studied firstly, and then elaborated in detail on the implementation steps of concrete data query types. Anglicizing the key influencing factors in seamless query processing cost model, to provide research guidance for seamless data query optimization. It focuses on the research of some important optimization methods, such as query processing compaction, the use of equivalent effectiveness mechanisms to reduce the volume of seamless computing, the use of data cache, and so on (Fig. **5**).

The key components of spatial data center related with seamless integration are extended and enhanced, and the basic functional needs are achieved through the data center function building mechanism. Through the practical application of the prototype system in the foundation geographic information database integration project, the feasibility and effectiveness of the technologies are verified, seeing Table **1**.

3. DATA MINING AND DISTRIBUTED APATIAL DATA

After many years of accumulation and development, the national spatial information infrastructure (NSII) of our country has reached a great scale, and provided massive spatial information resources for the national construction, social development and public services. However, with the wide development of network-based applications and the continuous improvement of its requirement, the application of these spatial information resources also encountered some issues requiring urgent solutions, the key performance of these issues is that the integrated use of these spatial information resources is given a low priority. At the same time, spatial data interoperability technology has improved; the cooperative processing technology of distributed GIS has also been studied deeply; spatial data mining and information fusion technologies are becoming mature day by day. All of these provide new approach to solve the problems mentioned above. At present, a large number of geographically distributed data collection stations can share their own spatial data resources "in its own position"; spatial information between different level of government departments, enterprises and institutions are exchanged frequently; a wide range of new system patterns have been proposed to deal with distributed heterogeneous spatial data in an integrated way.- In this situation, although the "information island" has been broken up in a certain extent, but the integrated use of distributed heterogeneous spatial data is still far from "seamless", the reason is that, for the seamless integration of spatial data, there is no clear concept defined, technology is not blameless, and there is no prototype system, and so on.

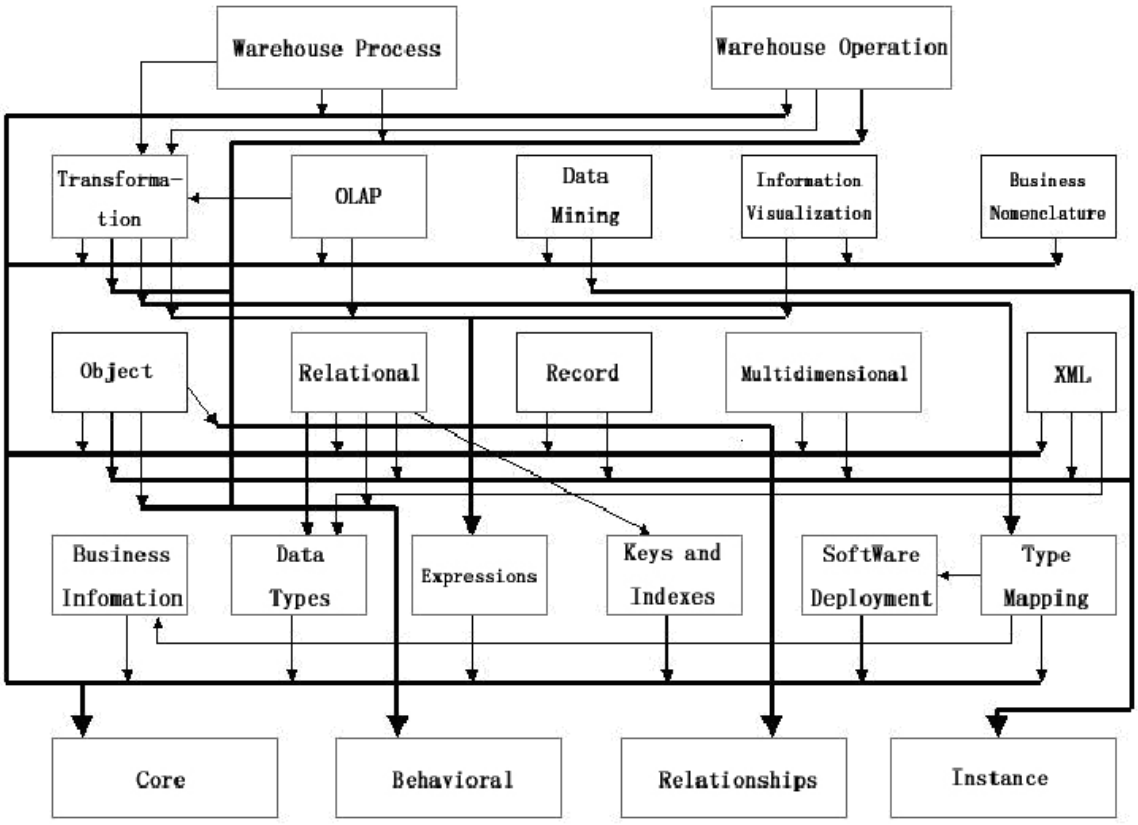

Fig. (5). Relationship between CWM.

Table 1. Information.

t		Ip: db:table:field:type:folsg		
A	192.168.1.220:efonn:infos:A:int:mysql	true		
B	192.168.1.220:eforni:infos:B:int:mysql 192.168.1.89:eform:infot:A:int:mmdb			MMDB
C	192.168.1.89:efonn: infot:C:int:mmdb			MMDB
D	192.168.1.220:efonn:infos:D:int:mysql			
E	192.168.1.89:efonn: infot:E:int:mmdb			MMDB
F	192.168.1.89:efornl:infot:F:int:mmdb			MMDB

On the other hand, in various application areas of GIS some spatial data center systems have appeared to manage spatial information resources. These systems have some similar patterns, which are: (1) able to manage distributed heterogeneous spatial data resources in a unified and concentrated portal form; (2) widely used for multi-level catalogue resource manager as the means of resource management; spatial data, non-spatial data and other resources can be grouped with a variety of themes and their order; (3) be able to organize into multi-level data centers cluster with the connection ability to the higher and lower level of data centers, and data can be exchanged between all levels; (4) it's important to accumulate and reuse GIS functions, and sometimes has the abilities to reorganize the application system as soon as the requirement is changed. However, these spatial data center systems almost simply put all kinds of spatial data and non-spatial data together to simplify the management process, and are not concerned with integrating the data resource in a seamless and consistent way, that is all how data resources can be geo-referenced using a unique spatial reference. Through the extension of data center model by introducing the concept, theory and technical methods of spatial data seamless integration, data center pattern will be upgrated into a feasibly distributed heterogeneous spatial data seamless integration pattern, and then the existing spatial data center system can be prepared as prototype system.

Key Codes

```
'%extra_arl;umcnt { Args*tt}//Parscr() }=f}
"/<}token-prefix SS_
'}}tokeytype I'enninal;
`%default_ type {"token}
'%lcft OR.
0f0left AND
0/left DOT.
%type in {Token}
%type select {Token J
%type oneselect{Token;
%type collist {Token)
%type seltable {Token
%token destructor{ token destructor(&$$);}
0loparse accept
%syntax error
cmd ::=in
sclp::=.
fieldname::=fieldname fieldnames.
fieldnames::=COMMA fieldname.
fieldname::=expr.
nm:: NAME
%type from{Terminal)
%type tablename{Terminal}
from::“FROM(Y) tablename(X).
n℃m::=
tablenarne(A):: NAME(X).
where op::
where_ op:: WHERE(Y) expr{X).
%type exprExpr*}
expr(A)::=expr(X) LT(Z) expr(Y).
expr(A)::= expr(X) GT(Z) expr(Y).
expr(A)::=expr(X) GE(Z) expr(Y)
expr(A)::=expr(X) I_L(Z) expr(Y).
expr(A)::=LP(Y) expr(X) RP(Z).
expr(A)::=expr(X) I:Q(Z) expr(Y).
expr(A)::=expr(X) NE(Z) expr(Y).
expr(A)::=expr(X) AND(Z) expr(Y).
expr(A)::=expr(X) OR(Z) expr(Y).
expr(A)::=expr(X) MINUS(Z) expyY).
expr(A)::=expr(X) PL.US(Z) expr(Y).
expr(A)::=expr(X) TIMI:S(Z) expr(Y).
expr(A)::=expr(X) DIVIDE(Z) expr(Y).
expr(A)::=MINUS(Z) expr(X).[NOT]
```

Big data querying see in the Table **2**.

4. METADATA INTEGRATION

The coming massive data era caused that the size of data sharply grew and searching space expanded. It was the new challenge to the data mining, as the demand for mass data of data mining became stronger and stronger. In order to solve the effectiveness and poor efficiency problem of traditional data mining techniques in the massive and high-dimensional data sets, the existed data mining algorithms have been improved to adapt to the actual conditions of the massive and high-dimensional data, (Fig. **6**) as well as the algorithm's execution efficiency has been improved, targeting the quality of the mining results of the massive data.

In the full understanding of the principle of the vertical frequent pattern mining and the problem of the massive data, a distributed vertical frequent pattern mining algorithm based on metadata integration was proposed in this paper, including a loading balance strategy for distributed vertical frequent pattern mining. The algorithm included three parts. Firstly, we sampled a small amount of data samples and calculated attribute correlation based on the rules generated for samples. Secondly, we finished the data partition that is we can divide the data into several independent data blocks according to the attribute correlation. Finally, we built the frequent pattern tree for every data block, mining the vertical frequent pattern tree to generate rules. Because the model of vertical frequent storage structure can ensure that the mining results were global results, therefore there was no need to combine the local mining results. At the same time, a load balancing strategy for distributed vertical frequent pattern mining is needed to differentiate the status of every site which was firstly calculated by the local site processing capacity and network capacity, and then assigned the task according to the different status of the site. Finally, through the experimental analysis to verify the proposed distributed vertical frequent pattern mining algorithm based on metadata integration; the results show that the proposed algorithm was more efficient under the large-scale data set compared with the traditional association rule mining algorithm. The load balancing strategy used to improve the efficiency of algorithm showed that it was better than the traditional load balancing strategy in terms of efficiency and load balancing degree.

Table 2. Big data querying.

t		Video Contentinfo-yyyymmdd.cfg	
	Video Base Data		**Drea info_yyyymmd**
Count: int	Count: int	content idaong	Area:
qqaong	qqaong	content name:chararray	area_ city:chararray
oid: int	oid:	length type:chararray,	area_type: int
Cld	cid:	pac — ontent: int	area_ country:int
sid:int	SICJ	pac content name:chararra Y	areaprov:
timestampaong	timestampaong		area_ cap_ tag:mt
gender:int	gender:int		area_ level:chararray
Age:in	Age:in		country code::int
scene:int	scene:int		
country code:int	country code:int		
Province-code:int	Province-code:int		

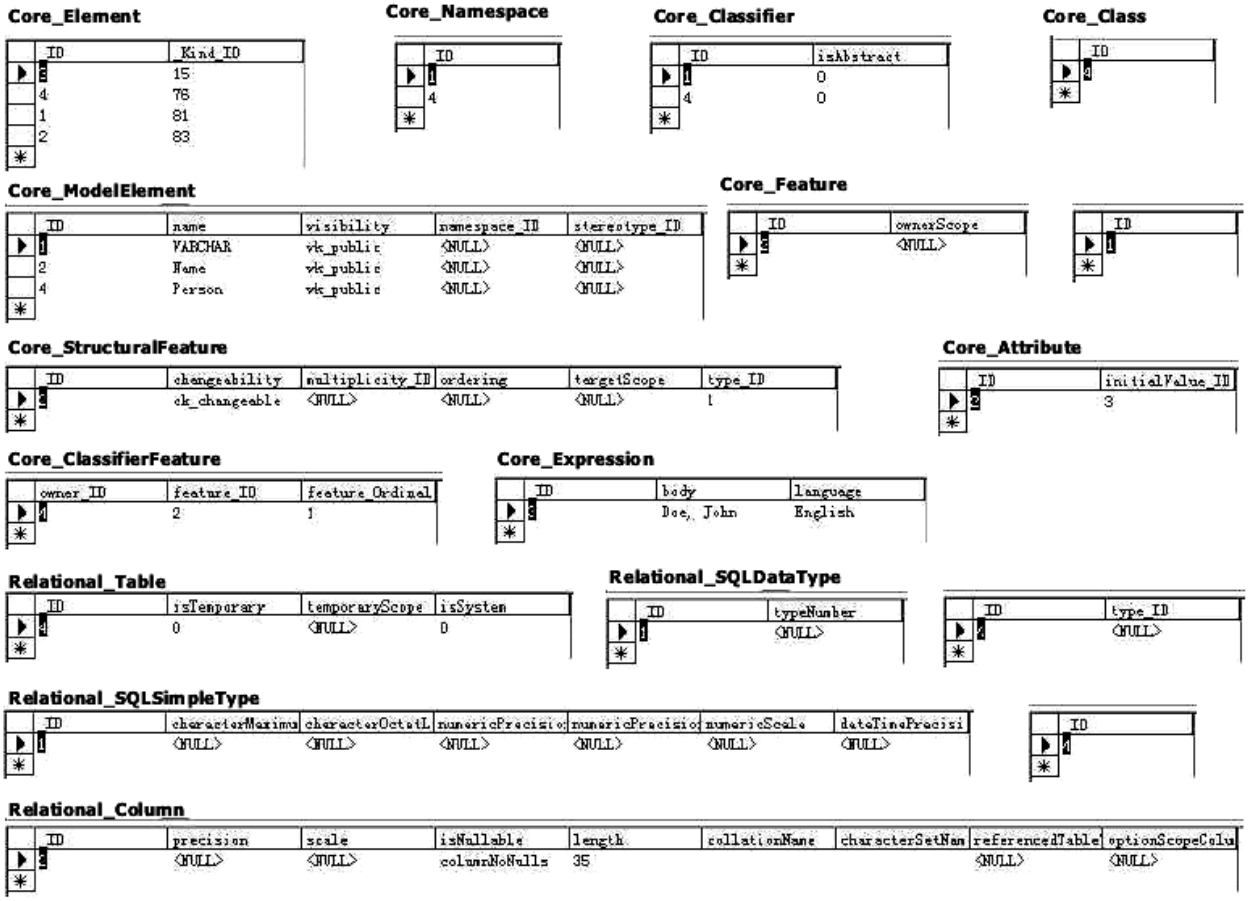

Fig. (6). Structure and storage part of the element of information in the database.

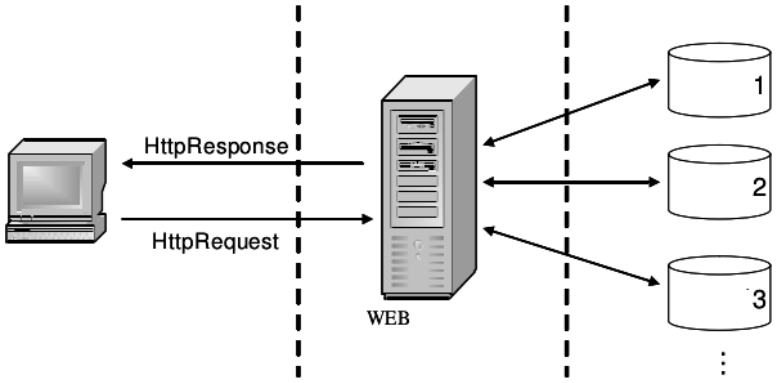

Fig. (7). B/S Structure.

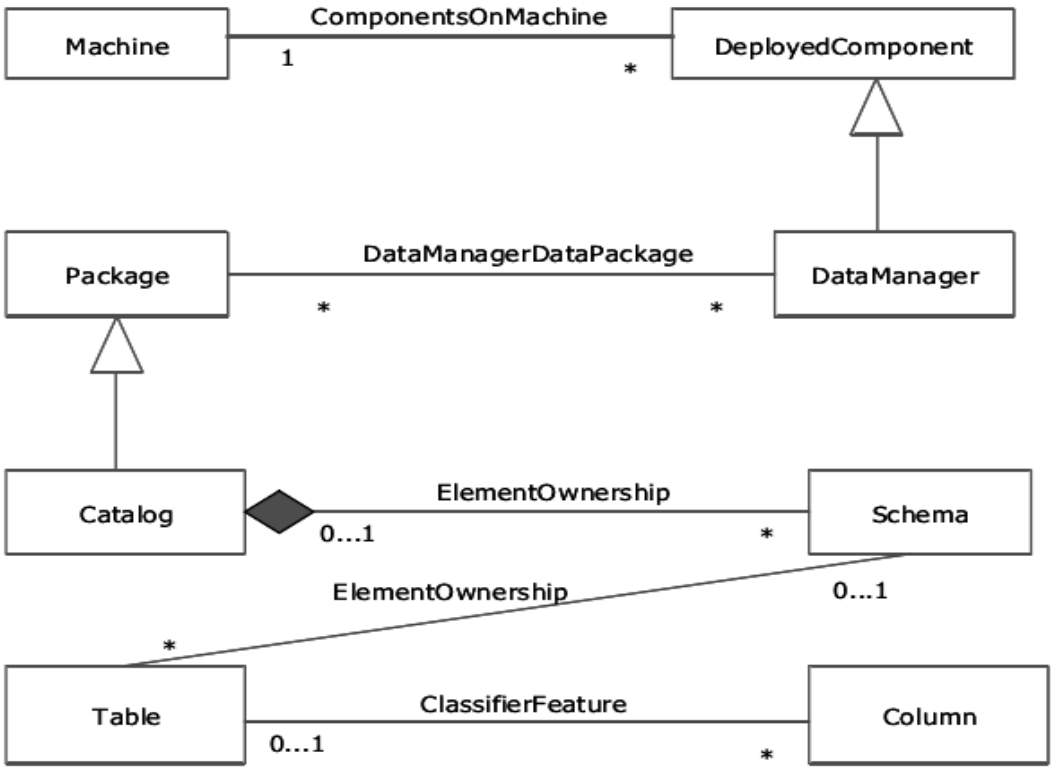

Fig. (8). Parts of CMW entity class.

In conclusion, this paper included the following three parts: first, analyzed the research background and significance, reviewed the pertinent literature; and then put forward a distributed vertical frequent pattern mining method based on metadata integration and the load balancing strategy for distributed vertical frequent pattern mining was presented in order to further improve the algorithm performance. Finally, the experimental analysis was performed for the proposed method, summary and forecasting (Fig. **7**).

Real-time database system (RTDBS) integrates real-time system and database technology. Currently, the real-time database system is widely applied to military, telecommunication, electricity, aviation, industry control and some other fields, and since the most of these fields are distributed, therefore, the research of distributed real-time database system (DRTDB) has caught the attention of experts in real-time field, database field and distributed field. DRTDB becomes a new research hot point (Fig. **8**).

Relative to the FRD method, up down reduction strategy can also reach the full reduction state, but can further reduce the data transmission cost of the down phase, especially in relation to many connect attributes; thereby reducing the query response time (Fig. **9**).

Real-time query must match its time requirement, or disaster may occur. The study researches the fault-tolerance

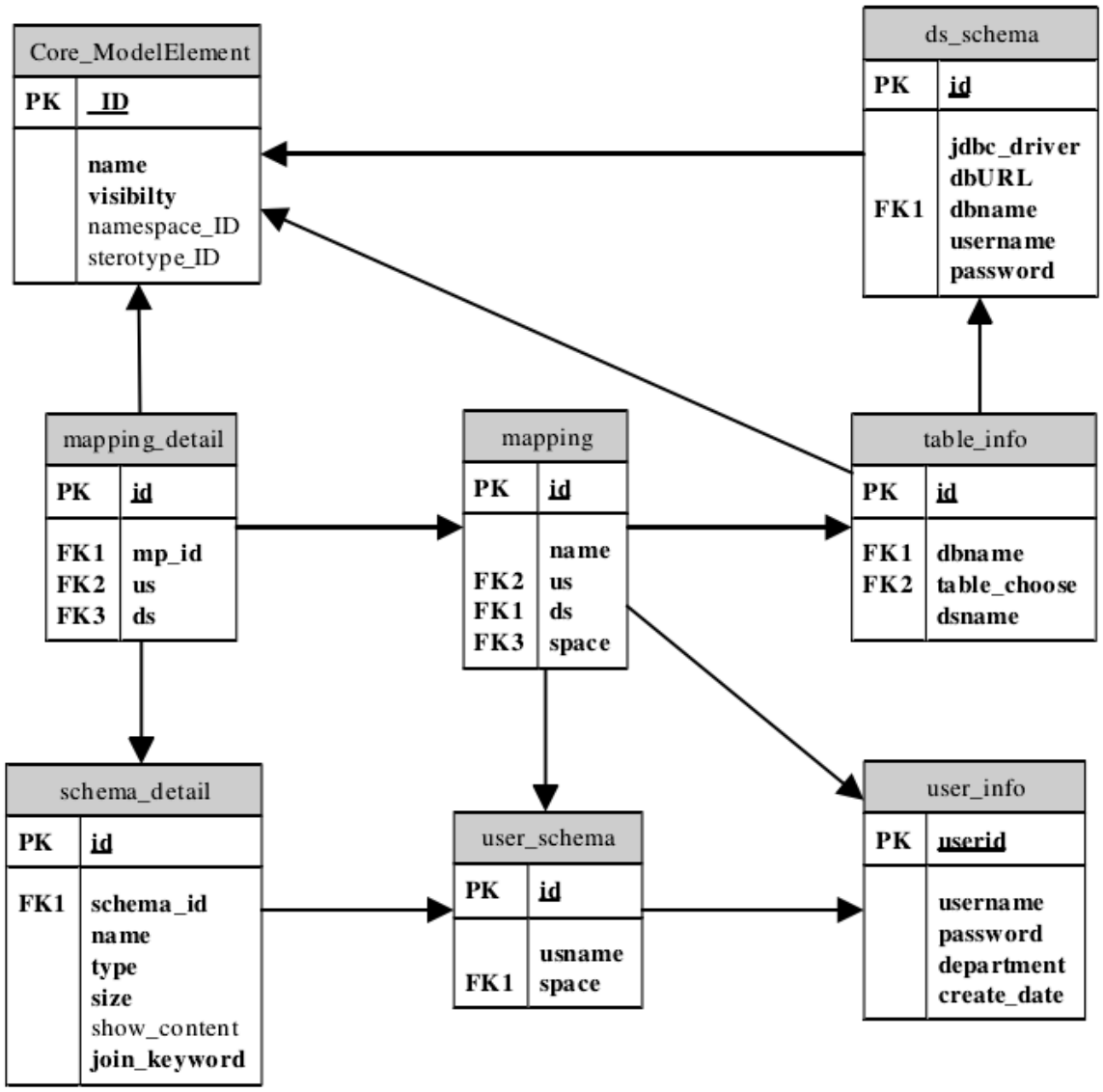

Fig. (9). Structure model in relational database and the relationship between them.

scheduling algorithm of the periodicity real-time transactions, and proposes a preemptive estimate algorithm, which can reduce the preemption to the low priority task that will finish soon so as to decrease the waste of the CPU resources. The new algorithm also uses a load balancing strategy to further optimize the task scheduling. By the experiment, the new algorithm could decrease the CPU resources waste and improved the success rate of system's real-time task.

The data management subsystem of marine monitoring system has the character of distributed and local real-time databases transmitting real-time data and commanding through wireless network. For the restriction of transmission rate of wireless network, it is required to develop an effective query processing method to reduce the query response time, so as to match the system's real-time proficiency. After analyzing traditional query optimization strategy, this study points out the merit and the drawback of the traditional strategies and proposes a new distributed query method, *i.e.* the up down semi-join strategy based on three query. The new query method proposes the concept of strong semi-join, in which multi relations' restriction is used to improve the reduction of one relation, the method also proposed an up down reduce strategy, which can lead the relations in a tree to a full reduction state. Therefore, the useless records are eliminated before transmission and the query's response time is reduced. The study also analyzes the full reduction strategy when there are two or more join attributes between two relations and the full reduction strategy of cyclic query, proposes the conception of default join restriction, points out the inherent difference between tree query and cyclic query, and finally proposes a new two-way method based on additional attribute to convert a cyclic query into tree query.

CONCLUSION

Use of metadata information is suggested to build a virtual table that can implement a unified query of distributed data sources. LEMON grinner parser is used to parse and check SQL statement on virtue table which the users submitted. In terms of common data query, the syntax tree is used for semantic optimization; and memory database is used to merge multiple source results. For big data query, Pig generate script is used to submit tasks; and Hadoop is used for distributed computing and query. Through multiple processes for processing IIDFS small file merging and file uploading or downloading to reduce the load of the NameNode node, the speed of uploads and downloads can be improved; making index on high frequency business, can find the data quickly and decrease the 'message program loaded'. Those solutions not only realized the data query optimization, but also achieved the goal of optimization.

Research methods in this article blocked the complex details of distributed data query, and provided a unified, simple SQL query interface to user. It makes the combination of distributed data query more convenient, and effectively improves the efficiency of the federated query execution.

CONFLICT OF INTEREST

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

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] K.Y. Lam, V.C.S. Lee, and S.L. Hung, "Priority assignment in distributed real-time database using optimistic concurrency control," *IEEE Proceeding on Computer and Digital Technology*, vol. 144, no. 5, pp. 324-330, 2014.

[2] M. J. Carey, L. M. Haas, P. M. Schwarz, M. Avy, W.F. Cody, R. Fagin, M. Flickner, A.W. Luniewski, W. Niblack, D. Petkoniv, J. Thomas, J.H. Wiliams, E.L. Wimmers, "Towards heterogeneous multimedia information systems: the Garlic approach", In: *Proceeding of the 5th International Workshop on Research Issues in Data Engineering-Distributed Object Management*," IEEE Computer Society Press, pp. 124-131, 2013.

[3] A B. Dennis, K.M. Hene, J. L. David, O. James, and L. David, "Wheeler. Gen Bank," *Nucleic Acids Research*, vol. 31. no. 1, pp. 23-27, 2013.

[4] W. John, F. Zukang, C. Li, Y. Huanwang, and M. B. Helen, "The protein Data Bank and structural genomics," *Nucleic Acids Research,* vol. 31, no. 1, pp. 489-491, 2013.

[5] http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html NCBI GenBank Statistics.

[6] Gene Ontology WWW Resources [EB/OL]. http:// www. geneontologty.org

[7] M. Kanehisa, S. Goto, S. Kawashima, and A. Nakaya, "The KEGG database at GenomeNet," *Nucleic Acids Research*, vol. 30, pp. 42-46, 2012.

Received: June 16, 2015 Revised: August 23, 2015 Accepted: September 11, 2015