International Journal of Database Theory and Application
Vol.9, No.11 (2016), pp.257-270
http://dx.doi.org/10.14257/ijdta.2016.9.11.23

A Study of Hybrid Heterogeneous System Based on Big Data Query

Sang Hailing

The Open University of Fujian, No.109, guishan Road, Fuzhou, Fujian, China, 35000
1504857703@qq.com

Abstract

With the development of science and technology as well as the advancement of national strategic deployment, study of hybrid heterogeneous system based on big data query has become a hot topic in internet industry both at home and abroad. Information technology has been widely applied in various fields and information has experienced explosive growth. However, different storage environments, collection systems and implementation platforms of information have hindered the communication and sharing of data between platforms and contributed a lot to deficient utilization of data. Thus the concept of heterogeneity comes into being. Heterogeneity in information system refers to difficulties in data utilization due to various data formats. This paper attempts to discuss the heterogeneous data integration methods based on big data and make an analysis. It explores advantages and convenience of design scheme on the basis of LDAP and offers detailed extracting rules for a better visual understanding of the corresponding model on its application.

Keywords*: Big Data, Hybrid Heterogeneity, LDAP server*

1. Introduction

With the rapid rising and development of high-tech industry in the past many years, people have accumulated increasing data in work and life. [1] Today, the concept of big data has come into people's vision. We can learn the difference between big data and traditional data from the aspects of data volume, speed, diversity and value. In terms of data volume, big data develops from TB to PB above; in speed, big data continuously produces data in real time, and its annual growth rate has exceeded 60%; in diversity, big data is a kind of semi-structured, unstructured and multidimensional data; in value, big data can do data mining and predictive analysis. [2] Compared with traditional scale data engineering, big data shows its significant characteristics of large scale, various types, fast speed, low reliability and low value density which cause great challenges for big data in storage, calculation and analysis. Now, there is no unified conclusion about how much is the big data. [3] Most professionals in this field think PB is the benchmark of big data. Also, some people think that big data is a subjective label which is attached in manpower and technology infrastructure and cannot keep up with the pace of demand. Whereas, storage environments, acquisition ways of existing data and implementation platforms of hard ware are so diversified that many useful data can not be used to the fullest and sharing resources between many platforms can not be achieved. [4] For this, the definition of heterogeneity has been raised. This question is the core of cloud computing and big data and its application and has also become one of the main questions in industry and the internet industry. Some well-renowned periodicals, such as 《Nature》and《Science》have put forward views on big data. [5] Revolving around the big data of hybrid heterogeneous system, this paper introduces the design model of heterogeneous system, integration and application of Lightweight Directory Access Protocol (LDAP) and heterogeneous

ISSN: 2005-4270 IJDTA

database, summarize the technology of analyzing such big data as profound learning and knowledge computing , analyzes and discusses it, at last draws a conclusion [6].

2. Research Status at Home and Abroad

At present, the research effort on big data is still increasing home and abroad. Among these researches, the key contents include big data acquisition technology, pretreatment technology, storage and management technology, analysis and data mining technology and big show and application technology which are badly in need of breakthrough. [7] In addition, the research content s also includes big data computing. While, mentioned in this article, hybrid Heterogeneous system has become a key research task both at home and abroad. All kinds of studies have begun with different ways and technologies. At present, two main methods are adopted in studying the data integration at home and abroad: one is the physical way taking data warehouse as the core and the other is the virtual one with the middle ware as its core. In the first way, all necessary data are stored in data warehouse but system engaged in integration still can operate the original database, which causes serious waste of time and space and shows defects of repeated storage and untimely upgrading. In the virtual one, the middle ware dose not store any actual data and it is only the connector between the receiver of data and source of data and convert the visits of data receiver into query command of data source. When it deals with the query, it needs to visit the data source, which is time consuming and of high cost.

This design scheme is based on the LDAP that is well known by people, so directory service of LDAP will be briefly introduced before establishing schemes. It is known to all that LDAP is the combination of the initials of the lightweight directory protocol or a kind of directory type of service. It is similar to the directory we use in the file system, similar to the telephone directory we use to query number, also similar to other web directory that we use such as the NIS (Network Information Service), DNS(Domain Name Service) and so on [8].

Due to the data of LDAP is shaped like a tree, it is of great flexibility on storing data. We can visit the LDAP directory with increasing number of LDAP client program which is easy to get on any computer platform. And it is easy to customize applications with LDAP support. At the same time, LDAP is deservedly deemed as database, but is much faster than data base query and has great capacity, as well as superiority in aspect of reliability and reducing the complexity of client. Modeling based on many advantages of LDAP enables information retrieval to be more efficient in heterogeneous information environment.

3. Design Model of Heterogeneous System

3.1. Introductions about Models of LDAP

At present, the use of LDAP has reached the generalization abroad, while the research and application is still very few at home. This paper mainly introduces the four models of LDAP corresponding to the four steps of development: information model, naming model, function model and security model respectively.

(1)Information Model

LDAP is based on the directory entry and every directory object on the directory service of LDAP should conclude at least one object class. Object Class stipulates clearly its class. This type determines the attributes that may exist in the directory entry and these attributes which may be included are called the optional and those attributes that must be included are called the mandatory. Object class of directory object can be modified but can not be deleted by users; at the same time the attribute of object class is restricted by server to prevent the change of semantic. Object class defines the objects of object oriented programming (Java) and subclass inherits all optional and mandatory attributes

of super class. Top class is the root of object classes and all other object classes are derived from it directly or indirectly. Mandatory attributes defined in the top class ensure that each directory entry has at least one object class. Please look at the following Table 1.

Table 1. Descriptions of Object Classes in Information Model

Object Class	Optional Attribute	Required Attribute	Remark
person	kn,cn objectclass	user Password telephone Number description,seeAlso	Personal Information Class
account	uid objectclass	description host,i,o,ou seeAlso	Account Class
top	objectclass		Ancestor Class

(2)Naming Model

Naming model illustrates the process during which users organize and quote the data in directory of LDAP. This model defines the organizing and identifying way of entry, and it is of certain flexibility in that it allows users to add entries into the directory. Directory entries are organized and stored in the structure of tree branches.

Naming model provides an independent name for each directory entry and identify them with distinguished names (DN), which can be regarded as path names of file system. They connect and arrange the super directory entries and the independent names of root and construct the names of directory entries of LDAP. In other words, DN is composed of relative distinguished name of directory and their superior distinguished names. If you read names of directory entries from left to right, you can trace the path names of directory entries to directory tree root, with DN, users can visit any entry in the tree at will according to their own need.

(3)Function Model

Function model shows the processing procedure in which information stored in directory server is treated and the query to directory server via reasonable application of LDAP as well as upgrading to strengthen the logic in Application. This model defines all operations of users obtaining and modifying directory, and these LDAP operations are independent of each other. The completion of information in directory server calls for effective utilization of directory to visit interface. Operation types included are shown as Table 2:

Table 2. Function Model

Operation Class	Main Operation Behavior
inquiry	search, compare, bind, *etc.*
update	add, delete, modify, *etc.*
verification	release the connection, the connection, *etc.*
other	extended operations, *etc.*

(4) Security model

The basic goal of security model is to provide a framework which can prevent the data in a directory from unauthorized visit. Security model is focused on unauthorized visit to

resources. To prevent unauthorized users from visiting or damaging resources, LDAP addresses this problem mainly through directory certification and directory visit.

Because all of the user's operation is dependent on the pre-established connection, so it can prevent users' illegal infiltration and visit by setting different access permissions in the connection, that is the directory certification mentioned above. Certification is used to build a dialogue between client sides and directory servers. Authorization, as an resolution, is capitalized on to clinch whether a visit to some resource can be granted.

3.2. Partitions of Model Hierarchies

Whether information data can be retrieved successfully is determined by the coordination of information collection and retrieval. [9] The specific processing procedures are as follows. First it will receive the complex forms input by users and then finalize the query plan according to data model of its own and figures out all the data sources in concord with the user's demand. Next, it will filter all data sources and then deliver query result to the user after complicated information conversion and reasonable combination. Although, this process is very complex but takes little time and is of high efficiency. Data collection adopted in model of this paper is the directory service of LDAP and retrieval is interrelated with management. This model is divided into three layers, which, ranked in the descending order, are USER LAYER, INTERLAYER, DATA LAYER respectively. USER LAYER plays a role of realizing the communicating and interacting of information with users, which means that information input by users and the query results are both achieved in this layer.

INTERLAYER is critical and engaged in decomposing the information input by users and arranging the output information as well as other management services. Its specific processing procedures are like this: Query information input by users is transmitted to the network server, converted into recognizable command of the system, and then transmitted into the server of LDAP. After receiving query command, this server will ascertain the resource center in which information is located and transmit the command to relevant resource center for implementation. There exist two kinds of situations: first, information needed by users can be found among a large quantity of information in the corresponding resource center and query result will be given to users; second, information input by users is not available in resource center and if so, this resource center will provide all detailed concise locations of information sources, dismantle the query command, and then transmit it to related information sources for implementation. Accurate locations of information sources are distributed, which is of great significance to solve the heterogeneity of data.

DATA LAYER is to store assorted data source of the whole information space and data of many formats into the server of LDAP, but the organizing and managing of these data requires professional some database management systems, such as ORACL, BD. It is necessary for LDAP and database management systems to make coordination on managing data before storing data into the server of LDAP to make sure that data has been successfully stored into the server of LDAP. This model is given in Figure 1.

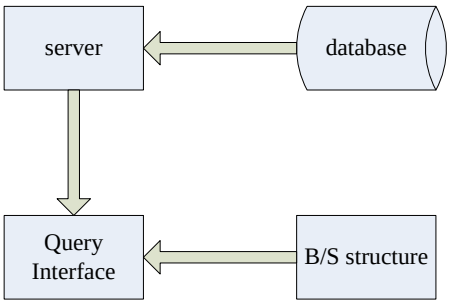

Figure 1. System Model

3.3. Partitions of Details in Model

Three layers of the system model consist of many parts and rely on them to function. The first layer of the system model includes query module composed of five parts-client side, converter, adapter, and LDAP server. The process is like the following: firstly, client side in the query module of this layer receives the input command of users and then transmits it by means of HTTP. Second, input command will be transmitted to the converter Java Served for command processing and identifiable language will be got. Next, adapter will match information and LDAP server will extract input information, transmit it to DSA and then return the result to the user interface. See Figure 2 in detail.

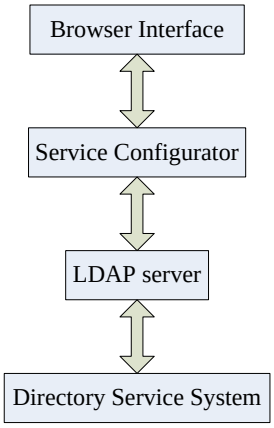

Figure 2. Details of Query Structure

3.4. Design Scheme of LDAP Server

The design of LDAP directory server is the design of directory tree on earth. To be simple, an directory tree is like an organizing method used to stipulate how to store information of different types and sub-tree information in each directory tree is maintained by one LDAP server. Hence, there are many LDAP servers in a heterogeneous information environment and connecting LDAP servers are to combine the distributed sub directory trees into an integrated structure of a tree. It is mentioned above that LDAP servers are distributed hierarchically so each level of server features recommendation between upper and lower levels which enables servers to connect with each other as well as transmit information. For example, when one server receives the query command, it will firstly check the locally stored information and if it can not find the related information, it will transmit command to super LDAP server or sub server; Each server will process the command in compliance with this protocol till the result is found.

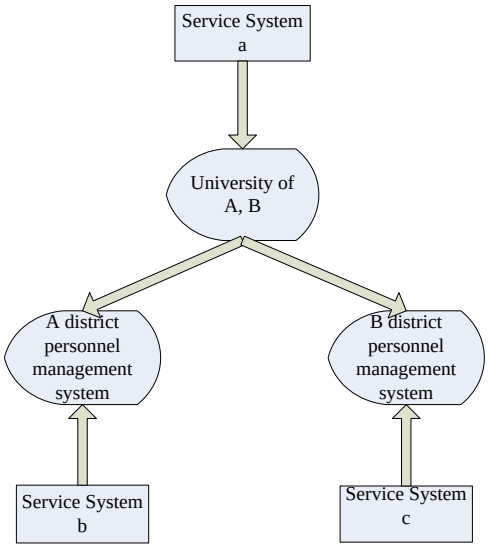

Figure 3. Organization LDAP Server

3.5. Operation Procedures of System

Detailed procedure of heterogeneous data query is like Figure 4:

Query interface of user layer is responsible for receiving the query command of users ,transmitting the command of java server page to web server and then operating data in LDAP through java naming and directory interface after conversion.

Server that has received the command carries out query operation and the query procedure is distributed. Just as the above said, it will search the locally stored information. If information in line with the command of users exists, it will ascertain the location of resource and then return to the user interface. Otherwise, it will probe into other servers existing in space and confirm the location of resource center and then return.

Resource centers will return the query result, deal with it consequently, convert the profile information through web server, return to interface layer with the form of java server page and show it to the person who inputs the command.

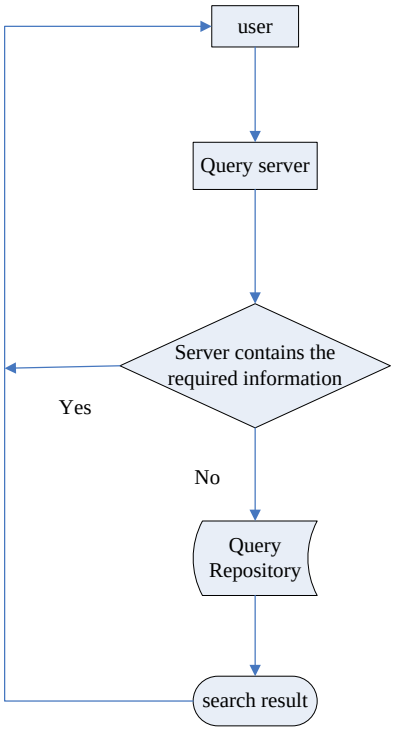

Figure 4. Heterogeneous Data Query Procedure

4. Integration of LDAP Server and Heterogeneous Database

At present, there are three different ways to enable users to retrieve and query information in database only with one retrieval platform, which is guaranteed by the integration of heterogeneous database:

(1) To realize conversion via middle database. It can import source data into the middle database and then transmit it to target database.

(2)Some computing methods that have been developed (distributed).

(3) To process it with middle ware. Query command will be firstly converted into a recognizable language of database and then will be transmitted to all databases existing in space. All query results will be managed and sent to the display interface for users.

(4) System that is capable of containing many databases. This system, which has different manifestations, can contain many databases of heterogeneous forms and is not possessed with one integrated database for managing systems software. However, its merits are very obvious. First, it can raise the efficiency of system; second, it can address the heterogeneity of data source. Because of its demerits, different configurations for transformation are necessary in integrating all databases. Considering that Open LDAP is used this paper and databases are integrated in similar ways, this paper will introduce the integrating method of Open LDAP, taking ORACLE database as an example.

Open LDAP and ORACLE databases

ORACLE database, taken as the background database of Open LDAP, will need ORACLE ODBE when accessing it. ORACLE ODBE makes the ODBE embedded application and accesses the Ibl Oracle and Open LDAP. Data in ORACLE is accessed through ORACLE ODBC that will be regarded as the background database of ORACLE. Similar steps will be adopted just like SQL is taken as the background database:

(1)Set up ORACLE.

(2)Set up ORACLE ODBC and configure.

(3)Configure database source information, during which database in ORACLE should be configured:

lmy_Oraele_dkn]

Driver=ORACLE

Database=my_database

(4)Set up and configure Open LDAP.

(5)Add data table.

This step is involved with conversion between hierarchy database and relational database table. Storage of hierarchy database can be finished through relational data table and this realizes the disguised integration of data. Relational data table stores information of tree structure by mapping. Open LDAP usually accesses data of tree structure and relational data has to be converted in ORACLE for Open LDAP's access. Hence, related knowledge on tree node storage in data structure should be capitalized on to achieve the storage of tree-structured information via data mapping between correlation table. Next, a simple correlation table will be used to realize the storage of tree-structured information supported by Open LDAP. Suppose that information to be stored is like this: uid=Lijun, ou=qufu, dc — school, dc — com.

First, create organization unit table and manning table:

Table 3. Organization Unit Table

id	Name
1	Shanghai

Table 4. Manning Table

id	First name	Last Name
1	Li	Yang

Second, create unit mapping table and attribute mapping table:

unit mapping table ldap_oe_mappings:

Table 5. Unit Mapping Table

id	first name	keytbl	keycol
1	Li	ou	Id

Attribute mapping table ldap_attr_mapping

Table 6. Attribute Mapping Table

id	oc_map_id	name	sel_expl
1	1	om	name
2	2	kn	lastname

Tree-structured information supported by Open LDAP can be accessed according to the following table:

Table 7. Table of Information

id	dn	oc_map_id	Parent	Keyvoal
1	om=Shanghai,dc=school,dc=com	1	0	1
2	uid=LiYang,om=qufu,dc=school,dc=com	2	1	1

System application

The reason why this system is developed is to raise query efficiency of hybrid query platform. This aim can illustrate the importance of query performance. First, test data set will be introduced. In this medical scene, DB2 has three tables and data size in each table is like the following table:

Table 8. Performance Test DB2 Table

Table	Data Size
PATIENT	8000
DOCTOR	100
MEASURE	100

There are 8 collections in newsql database of MongoDB and data size in each document is like the following Table (4-2):

Table 9. Performance Test MongoDB table

Collection	Data Size
PUBLIC_PATIENT_HOSPITAL_IN	1000
PUBLIC_PATIENT_HOSPITAL_OUT	1000

PUBLIC_PATIENT_PROCESS	1000
PUBLIC_PATIENT_LOOKOUT	500
PUBLIC_PATIENT_DOCTORADICE	100
PUBLIC_PATIENT_PRESCRIPTION	100
PUBLIC_PATIENT_SHEET	100
PUBLIC_PATIENT_TEMPERATURE	100

Here SQI collections are divided into the following types Table 10:

Table 10. SQL Query Efficiency Testing Table

Key words	DB2Field	JSON Field	Hybrid Field
SELECT	A	B	C
Query Condition	D	E	F

In the chart above, conditions in different fields are represented by letters from A-F symbolic of key words. This chart shows that both query results and query conditions have three types and there are 9 conditions after permutation and combination. Because of its coverage, three types of designs are chosen in making an example, namely, C-D, C-E, C-F.

However, in order to have a better illustration of this hybrid query platform's query support on JSON data, remaining six conditions will not be omitted. Please look at table 11:

Table 11. SQL Query Efficiency Testing Table

Number	SQL Sentence	Testing Type
1	Select patient_name from patient where JSONVal (Patient.hospital_in ， 'hospitalID')= '0330803_1'	A-E
2	select JSONVal(Patient.hospital_in，'hospitalID') from patient where patient_id='0462589'	B-D
3	select JSONVal(patient.process，'firstProcess') from patient where JSONVal(patient.process，'firstProcess.firstProcessID')='0048152_1_0'	B-E
4	select patient_name，patient_birth， JSONVal(patient.process，"firstProcess.firstProcessID") from Patient join Measure on JSONVal(Patient.Process，"firstProcess.MeasureID")	C-D

	= Measure. Measure_ID where patient_sex= VAL order by cost desc	
5	Select cost.hopital_id， MEDINSURANCETYPE， patient_id，patient_name name，patient birth，patient_sex， JSONVAL(patient.hospital_in ， 'hosTimeIn') ， JSONVAL(patient.hospital_out，'outTime')， JSONVAL(patient.hospital_in ， 'inDays') ， JSONVAL(patient.process ， 'firstProcess.resultCN') ， JSONVAL(patient.hospital_in ， 'season') ， JSONVAL(patient.hospital_in，'ensureResult') from patient	C-E
6	Select patient_name ， measure_name ， measure_score ， jsonval(patient.hospital_in，"doctorName") from patient join measure on jsonval(patient.process，"firstProcess.MeasureID")=measure_id where measure_score= VAL and jsonval(patient.hospital_in，"doctorName") like "张%"	C-F

The index system built above has distinct merits in that it is supportive of both single-column index and hybrid index. Query result should include things shown in Table 12:

Table 12. SQL Query Efficiency Testing Table

Number	SQL Sentence	Testing Type
7	select doctor_name from Doctor where Doctor_account = VAL and Doctor_pwd = VAL	Common hybrid index
8	select patient_name from patient where JSONVal(Patient.hospital_in ， 'hospitalID')= '0398168_1' and JSONVal(Patient.hospital_in，'hosTimeIn') = '20081108'	hybrid index of JASON

Random query of condition numbers from 8 testing examples above is carried out on prepared test data set and comparison result among average query efficiency is shown in Table 13:

Table 13. SQL Query Efficiency Testing Results Comparison Table

Example Number	Time of index-free query	Time of index query	Query efficiency Promotion value	Query efficiency Promotion percentage
1	45	16	30	66.5%
2	280	49	234	82.6%
3	30	16	15	50.8%
4	480	254	232	47.9%
5	90	63	30	35%
6	260	245	24	9.3%
7	15	4	13	86.3%
8	28	17	15	50.5%

Broken line graph can be got after the arrangement of data above.

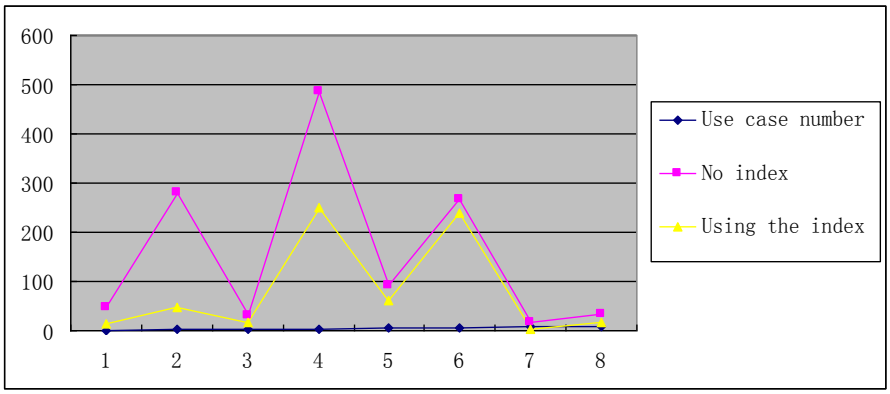

Figure 5. Query Efficiency Test Result

It shows that the query efficiency of this index system is obviously raised.

5. Conclusion

Taking into account the existence of big data and focused on exploring the merits and convenience of the design scheme based on LDAP server at the beginning, this paper put forward the heterogeneous data access system which realizes both the integrated and apparent access to heterogeneous data sources and synchronicity of data. It also gives some examples of corporate personnel and manifests advantages of this system. This kind of research contributes a lot to the increasingly rich knowledge in this area and can boost its development. Since the inner-corporate and Inter-corporate heterogeneous data increases, how to realize fast and effective data inter-operation has become a hot topic. This paper discusses the system framework based on the integration method heterogeneous data, makes a key research on ontology extraction and query decomposition, gives a detailed introduction to extracting rules, imports class-source mapping tables to raise query decomposition efficiency, optimizes query and resolves semantic heterogeneity.

References

[1] M. R. Martnez and N. Roussopoulos, "MOCHA: A self-extensible database middleware system for distributed data sources", In Proceeding ACM SIGMOD Conference, Dallas, Texas, USA, **(2000)**, pp. 213-224.
[2] B. P. Chandra, "Designing performance monitoring tool for NoSQL Cassandra distributed database", 2012 International Conference on Education and e-Learning Innovations, **(2012)**, pp. 1-5.
[3] F. Brezzi and M. Fortin, "Mixed and hybrid finite element methods", Springer-Verlag New York, Inc. New York, NY, USA, **(2011)**, p. 23-45.
[4] http://www.infosecurity.org.cn/article/pki/ldap/23483.htm
[5] C. F. Gomes, "Assessing No SQL Databases for Telecom Applications", 2011. IEEE 13th Conference on Commerce and Enterprise Computing, **(2011)**, pp. 267-270.
[6] O. Cure, R. Hecht and C. L. Due, "Data Integration over No SQL Stores Using Access Path Based Mappings", Database and expert systems applications. part 1, **(2011)**, pp. 481-495.
[7] Q. Tao, "Research on Scalability of Database Design", 2011 2nd International Conference on Data Storage and Data Engineering(DSDE2011), **(2011)**, pp. 114-117.
[8] Y. Z. Wang, X. L. Jin and X. Q. Cheng, "Network big data: Present and future", Chinese Journal of Computers, vol. 36, no. 6, **(2013)**, pp. 1125-1138.
[9] Y. Matias, E. Segal and J. S. Vitter, "Efficient bundle sorting. In: Proc. of the 11th Annual ACM-SIAM Symp. on Discrete Algorithms (2000)", Society for Industrial and Applied Mathematics, **(2000)**, pp. 839-848.
[10] W. Evans, D. Kirkpatrick and G. Townsend, "Right triangular irregular networks", Technical Report, TR97-09, Department of Computer Science, University of Arizona, **(1997)**.
[11] J. Roijackers, "Bridging SQL and No SQL", Eindhoven University of Technology Dissertation, **(2102)**, pp. 15-23.

Author

Sang Hailing, She is a master and works in the Open University of Fujian as a lecturer. The address is No. 109, guishan Road, Fuzhou City, Fujian Province.

