Drilling Material Data Warehouse ETL System Research Based on Crowd-Sourcing

Jie Zhou[1,2]， Ming Fang[1,2] and Xin-ran Cao[1]

[1] Institute of Computer，Xi’an Shiyou University ,Xi’an，China
[2] Engineering Research Centre of Oil&Gas Information System ,Xi’an，China

Abstract. Drilling material data warehouse is an important platform for assisting drilling engineering decision support and data analysis. The construction of high-efficiency and high-quality enterprise-level data warehouse puts high requirements on data quality. The operation targets of oil and gas drilling engineering are buried underground, and there are difficulties in data obtaining. This causes many uncertain data in the oil and gas drilling business database. At the same time, the business database in one area is quite different from those of other areas due to geographical environmental impacts ,so the database selection and integration strategy is uncertain . In order to solve these uncertain problem, this paper proposes building a drilling materials data warehouse ETL system based on crowd-sourcing, and improve the data quality of the data warehouse, thereby improving the efficiency of data warehouse construction.

1 Introduction

Drilling material data warehouse is an important platform for assisting drilling engineering decision support and data analysis. How to extract and convert a large amount of data accumulated from drilling production DB into drilling material data warehouse is the foundation for building a data warehouse. Drilling operations target oil and gas reservoirs are buried in the ground. There are problems in data acquisition. At the same time, the data of oil and gas drilling engineering vary greatly under different geological conditions. When constructing a data warehouse, the mapping metadata to the data warehouse is uncertain. The existence of a variety of uncertainties has made it difficult to ensure the data quality of the drilling materials data warehouse. If the traditional data warehouse ETL method is used to construct the drilling materials data warehouse, a large amount of effective data will be wasted, which will not provide reliable data support for the drilling engineering decision analysis, and will directly affect the efficiency of drilling operations. This paper adopts the theory and method of network crowd-sourcing to combine the traditional drilling materials data warehouse ETL method with crowd-sourcing methods, and proposes a crowd-sourced drilling materials data warehouse ETL processing method, from the data warehouse mapping metadata design and uncertainty. Start with data processing and design and build a crowd-sourcing model to effectively improve the efficiency of traditional ETL processes.

2 Drilling Material Data Warehouse ETL System's Overall Structure Based on Crowd-sourcing

The exploration and practice of drilling material data warehouse construction has lasted for a long time, but it has been affected by poor data quality, and industrial data analysis methods have not been well implemented. The oil drilling project is affected by the industry's own characteristics, with difficulties in obtaining data, incomplete access, and large differences in data between different regions. The uncertainty of the data itself and the uncertainty in mapping metadata from the data source to the data warehouse when building a data warehouse affect the data warehouse construction efficiency. Therefore, it is necessary to introduce more extensive human resources in the traditional ETL process and use human resources to handle uncertain information. Crowd-sourcing was first formally proposed by Howe in 2006. It refers to the practice of a company or organization to outsource its work tasks to employees or contractors in a free and voluntary manner to non-specific public networks or virtual communities.It is an online, a solution to the problem.

DMEC (Drilling Material DW ETL System based on Crowd-sourcing) combines traditional ETL process information processing methods with crowd-sourcing methods to construct a crowdsourced drilling materials data warehouse ETL uncertain data management. The platform (DUDMEC) and the crowdsourced drilling materials data warehouse metadata management platform (DMMMC) are shown in Figure 1 below. Starting from the two aspects of uncertainty data processing and metadata design, human resources are used to deal with uncertain problems in the ETL process of data warehouses. The application of the two platforms runs through the entire process of data warehouse ETL construction. The two platforms will be introduced separately.

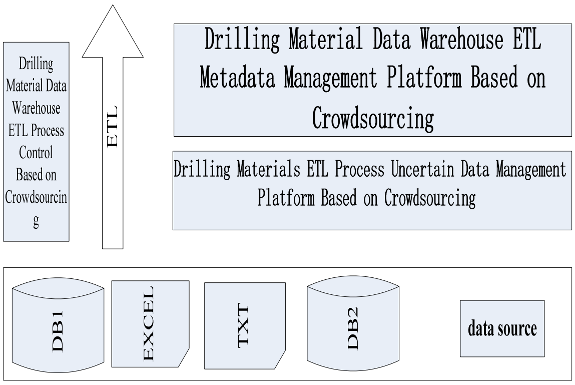
Drilling Material Data Warehouse ETL Process Control Based on Crowdsourcing
ETL
Drilling Material Data Warehouse ETL Metadata Management Platform Based on Crowdsourcing
Drilling Materials ETL Process Uncertain Data Management Platform Based on Crowdsourcing
DB1
EXCEL
TXT
DB2
data source

Figure1. The whole structure

3 Drilling Materials ETL Process Uncertain Data Management Platform Based on Crowd-sourcing

In the ETL process built in the Drilling Material Data Warehouse, basic operations on database data such as query, insert, modify, and delete are performed throughout. The traditional relational database data query is mainly based on exact matching. Based on the “closed world hypothesis”, information not in the database is considered to be false or non-existent[1]. Moreover, the traditional relational database is very extreme. They think that data is entering the data. The warehouse has previously properly cleaned and verified the data and cannot tolerate data query inconsistencies. The oil and gas drilling engineering itself often has incomplete data acquisition, and the relatively poor normative consistency of the early business database management, which makes us often query errors in the oil and gas drilling business database query data or return to the empty. In the process of constructing the data warehouse ETL, the data accumulated by the multi-year oil and gas information system construction cannot be fully utilized. Therefore, a crowdsourcing-based drilling material ETL process uncertain data management platform DUDMEC (Drilling Uncertainty Data Management in ETL Process) is proposed. The structure of System based on Crowd-sourcing is shown in Figure 2 below.

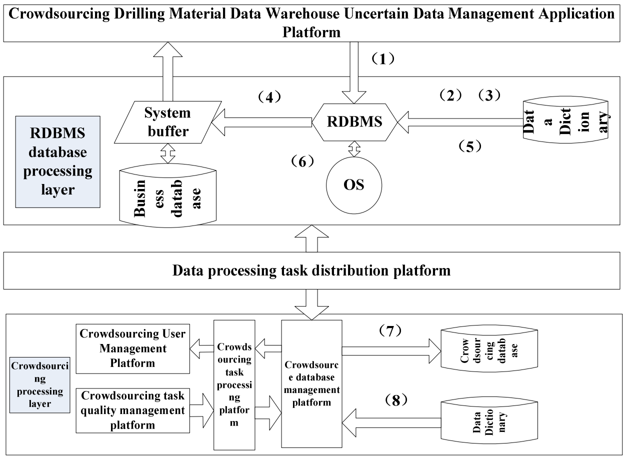
Crowdsourcing Drilling Material Data Warehouse Uncertain Data Management Application Platform
（1）
（4）
（2）（3）
RDBMS database processing layer
System buffer
RDBMS
Data Dictionary
（5）
（6）
Business database
OS
Data processing task distribution platform
Crowdsourcing User Management Platform
Crowdsourcing processing layer
Crowdsourcing task quality management platform
Crowdsourcing task processing platform
Crowdsource database management platform
（7）
Crowdsourcing database
（8）
Data Dictionary

Figure 2 Drilling Material ETL Process Uncertain Data Management Platform Based on Crowd-sourcing

3.1 Traditional relational database query phase

(1) The request is issued. The user sends a query request to the business relation database management system (RDBMS) through the client application, giving the relationship name and the specific query conditions.

(2)Semantic checking and permission checking. The RDBMS first reads the data dictionary to check if there is a relationship given in the client request, the corresponding field, and whether the user has read operation permission. After confirming that the grammar semantics are correct and the access rights are legal, transfer to (3). If the access right is not valid, the request is refused and the situation description is returned to the client. If the semantics are correct and the access rights are legal, but there are no corresponding fields in the data dictionary, call the stored procedure to jump to the crowd-sourcing management platform in the form of a page trigger and enter the crowd-sourcing database query phase.

(3)RDBMS performs query optimization. According to the information in the data dictionary, an appropriate algebraic optimization strategy or access path optimization strategy is selected to convert the user operation request into a series of single-record access operation sequences.

(4)Buffer record search. When a record satisfying the condition is found in the buffer, the RDBMS requests to derive a data format that meets the user's requirements according to the relevant information in the data dictionary and the command sent by the user, and transmits the queried data record to the client. End the query. If no record meeting the conditions can be found in the buffer, go to (5).

(5)Query storage mode. The RDBMS looks up the storage mode of the user command contents in the data dictionary, determines the file reading directory and mode, and the target physical record.

(6)According to the result of (5), a request to read the record is issued to the operating system, operating system can find the corresponding record in the storage area of the database, then the operating system executes the relevant operation according to the instruction. If the corresponding record is sent to the system buffer and enters (4) Process. If the operating system cannot find the relevant records in the storage area of the database, it submits it to the crowdsourcing platform and enters the crowd-sourcing database query phase.

3.2 Crowdsourced database query stage

(7)CrowdSQL language processing. The tables or fields that fail to be queried in the traditional relational database will be entered into the crowd-sourcing platform for query. The specific implementation is as follows: If some fields in a table or the entire table fails to be queried in a relational database query, all the fields of the data table will be queried respectively, and the fields that can be queried will be put in public. The field description type of the package database data dictionary is set to normal. For a field that cannot be queried, the field description type of the crowdsource database data dictionary is set to crown. If all the fields in the query table are of the crowd type, the description type of the table in the crowdsource database data dictionary is set to crown.

(8)Crowdsourcing data management platform task identification. When the crowdsourced data management platform queries the crowdsourced database data dictionary, when the "crowd" keyword is queried, the crowdsourcing task is rationally decomposed and the task is released through a crowd-sourcing platform internal task processing mechanism.

(9) The crowdsourcing management platform selects eligible recipients to answer the tasks, selects the results through the crowdsourcing platform's conclusion evaluation strategy and quality control strategy, and loads the final results to the crowdsourcing database management platform to feedback to the customers[2]. The user therefore queries the target data.

4 Drilling Material Data Warehouse ETL Metadata Management Platform Based on Crowdsourcing

The ultimate goal of managing data warehouse metadata is to obtain high-quality data. In the oil and gas drilling engineering, the BOM data vary greatly under different geological and climatic conditions. In the ETL process, it is of great significance to organically integrate and standardize the data in different databases. Based on this, we designed and built a metadata database management system for ETL process and built a crowdsourcing platform for the mapping metadata processing.

4.1 Metadata Database Management Information System

The design metadata database management system is as follows. The system is mainly composed of three parts. The bottom layer is the metadata acquisition part, the middle layer is the metadata storage part, and the top layer is the metadata application part.

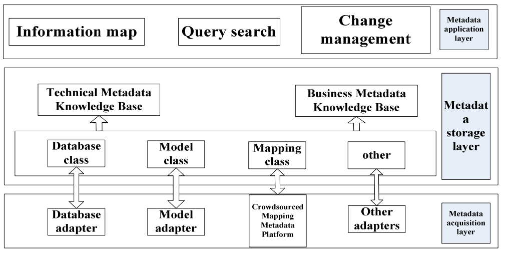
Information map
Query search
Change management
Metadata application layer
Technical Metadata Knowledge Base
Business Metadata Knowledge Base
Metadat a storage layer
Database class
Model class
Mapping class
other
Database adapter
Model adapter
Crowdsourced Mapping Metadata Platform
Other adapters
Metadata acquisition layer

Figure 3 . Metadata Database Management System

The metadata acquisition layer mainly includes database adapters, model adapters, other adapters and a crowdsourced mapping metadata processing platform. The database adapter is mainly used to read the data dictionary information of the designated database table from different business databases at the bottom of the data warehouse, and load the metadata dictionary into the metadata database according to the design requirements of the metadata database[3]. It directly accesses the data warehouse system through the page triggering method through the database stored procedure. After reading the target data, it is automatically loaded into the metadata database management system. The model adapter mainly collects the source files generated using the model tool during the data warehouse construction. For example, using the PowerDesigner tool to parse the model entities and associated relationships, and automatically converted to metadata information is loaded into the metadata database management information system[4]. Other types of adapters are mainly used to automate the conversion of other types of metadata in the process of data warehouse construction and loaded into the metadata database management system.

Different from the above three types of metadata collection methods, the design and acquisition of data mapping metadata is a crucial and difficult part of the metadata design process, and the current technology cannot achieve automatic design acquisition. In order to improve the quality of the data mapping metadata in the ETL process, we propose to build a crowd-sourced data mapping class metadata management platform.

4.2 Crowdsourcing Data Warehouse Mapping Metadata Management Platform

Data warehouse mapping class metadata refers to the mapping between data sources and data warehouse data. That is to say, we need to reflect the class metadata to reflect which specific data source the data item in the data warehouse is extracted from, and finally how to load it into the data warehouse[5]. This process is very complicated. Especially for oil and gas drilling projects, since oil and gas drilling data are greatly affected by environmental impacts, it is crucial to specify data sources. In the mapping metadata design, we use the crowdsourcing method to build a crowdsourced Drilling Material Data Warehouse Mapping metadata management platform based on Crowdsourcing (DMMMC). The platform structure is shown in the figure below.

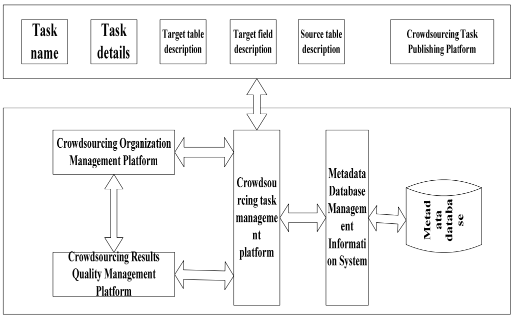
Task name
Task details
Target table description
Target field description
Source table description
Crowdsourcing Task Publishing Platform
Crowdsourcing Organization Management Platform
Crowdsourcing Results Quality Management Platform
Crowdsourcing task management platform
Metadata Database Management Information System
Metadata database

Figure 4. DMMMC

The data warehouse technicians publish the mapping metadata crowdsourcing task on the crowdsourcing task publishing platform, mainly including the task name, the description of the task content and requirements, the description of the target data table, the target field description, and the metadata database management system. The descriptive information is extracted from the database source table. Crowdsourcing task management platform, through its personnel organization platform, selects recipients who meet the conditions to receive tasks. After the tasks are completed, the task completion results, that is, mapping class metadata design results, are submitted to the crowdsourcing result control platform. After the crowdsourcing result control platform's audit finalizes the plan to be adopted, the results are fed back to the crowdsourcing task management platform. At the same time,the results are loaded into the mapping class metadata class library part of the metadata database management system through a database stored procedure in a page-triggered manner.

The process of metadata design is not static and dynamic. It will change with the needs of applications and the environment. Through the design of this platform, the efficiency of dynamic design of metadata is greatly improved.

5 Conclusion

Drilling material data warehouse ETL system based on crowdsourcing design uses a network crowdsourcing approach to solve the data-related uncertainties encountered during the data warehouse construction ETL process, effectively solving the uncertain data in the drilling business database,which provides an effective solution to the effective and unified management of metadata. It is no longer a closed system design but a more open and more systematic design. The data warehouse data quality has been greatly improved, providing good data support for data analysis and decision making in the data warehouse.

References

1.A. Doan, R. Ramakrishnan, A. Halevy.*CACM*,86(2011)

2. J. M. Hellerstein , J. F. Naughton.*SIGMOD*,423(1996)

3. M.-F. Balcan, S. Hanneke, J. W. Vaughan.*Machine Learning,*80(2010)

4. I. Muslea, S. Minton, C. A. Knoblock. *J. Artif. Intell. Res.(JAIR)*,203(2006)

5. J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, M. Anthony.*IEEE Transactions on Information Theory,* 44(1998)