

METHOD FOR EFFICIENTLY PROCESSING MULTI-DIMENSIONAL DATA BASED ON GROUP MANAGEMENT

TECHNICAL FIELD

[0001] The present invention relates to the field of medical management, and in particular, to a method for efficiently processing multi-dimensional data based on group management.

BACKGROUND

[0002] Interconnection of service ports, collection of multi-source data and data sharing operations are generally involved in an existing clustered information management system. The multi-source data involved at the service ports will be distributed in different service systems and different types of databases. During calling and actual use of data, this will lead to differences in terms of a structure, format and standardization level of the data due to different sources of the data, making it impossible to better perform structurization and standardization processing of the collected data. In an existing data processing technology, a scheme of data collection as processing is generally adopted, and unified data extraction and calling are performed after data processing is completed. However, this processing mode is inefficient for processing batch data in a plurality of databases, which seriously affects operational efficiency of an information management system with bulk data update.

[0003] The Chinese patent with Publication No. CN102650995A provides a system and method for generating a multi-dimensional data analysis model. Although the patent also relates to design of a method for processing multi-dimensional data, data processing is performed by establishing a reusable multi-dimensional data analysis model in the patent. Processing of the multi-dimensional data in the patent is mainly used for improving accuracy of the processing of the multi-dimensional data. However, in the patent, data processing is performed separately by establishing a plurality of multi-dimensional data models and then association and combination of the models are performed, making it impossible to further improve efficiency of the processing of the multi-dimensional data, and thus impossible to further solve the problem of optimizing operational efficiency of an information management system with bulk data update. The Chinese patent with Publication No. CN103488475B provides a multi-dimensional data analysis system and a multi-dimensional data analysis method. The patent relates to a method for analyzing multi-dimensional data and is designed

to improve ease of use in terms of user operations. In the patent, however, received data is directly classified into path information in different dimensions, and separate transmission and analysis of data are performed based on the path information.

[0004] Therefore, a method for efficiently processing multi-dimensional data based on group management is provided in the present invention for problems present during existing efficient processing of the multi-dimensional data.

SUMMARY

[0005] For the above problems, a method for efficiently processing multi-dimensional data based on group management is provided in the present invention, specifically including: establishing a method for parallel collection of multi-dimensional management data under a plurality of databases, performing data integration of multi-source database tables on the collected multi-dimensional management data, performing data standardization processing on an integrated database table, calling data subject to the standardization processing in a database of a service system, and performing data calling and query according to a service rule in the service system.

[0006] Preferably, the data calling and query includes data alert, data notification, and standardized quantized data display.

[0007] Preferably, the plurality of databases include MySQL, Oracle and Excel databases; and in the method for parallel collection, a cache area of the multi-source database tables is established, and the data integration and the data standardization processing are performed on the collected multi-dimensional management data in the cache area of the multi-source database tables.

[0008] Preferably, a python-based pandas tool is used in the parallel collection to set connection attributes with the plurality of databases; and the connection attributes include a database table address, a user-defined login name and login password, and a flag bit determination.

[0009] Preferably, the data standardization processing includes the following specific steps: collecting metadata of the multi-dimensional management data under the plurality of databases, in which case the data standardization processing is not performed on the metadata; and performing unified standardization processing of data after the data integration based on the multi-source database tables is completed.

[0010] Preferably, the data integration of the multi-source database tables includes the

following specific steps:

[0011] A1. performing labeled classification on database tables under the plurality of databases, and performing clustering processing on the database tables based on a label in the labeled classification;

[0012] A2. dividing the clustering processing into parallel label clustering and tree-structured label clustering on the basis of A1;

[0013] A3. when performing the parallel collection of the multi-dimensional management data under the plurality of databases, performing classification processing on the parallel label clustering and the tree-structured label clustering, respectively, based on the label by loop traversal; and

[0014] A4. on the basis of A3, performing merging of the database tables under the plurality of databases, and performing the data standardization processing on a merged database table, where the data standardization processing includes data cleaning, data deduplication, data format unification, and abnormal data value rejection.

[0015] Preferably, in the parallel label clustering and the tree-structured label clustering, labels in the database tables are graded according to the service rule and labels at a same level fall into a same set in the database tables under the plurality of databases.

[0016] Preferably, in the tree-structured label clustering, attribute relationships of labels at all levels fall into a set.

[0017] Preferably, in the data standardization processing, a unified calling logic based on required data in the service system is established and the unified calling logic is used for data application in the service system.

[0018] Preferably, in the data standardization processing, processed data is stored in a cache area of the multi-source database tables, and data in the cache area of the multi-source database tables is called by a PostgreSQL database.

[0019] Compared with the prior art, the present invention has the following beneficial effects:

[0020] (1) In the method for efficiently processing multi-dimensional data based on group management according to the present invention, a python-based cross-database data extraction and data standardization processing method is established to further improve efficiency of multi-source data collection and data sharing at a plurality of service ports under the service system in a group management system.

[0021] (2) On the basis of (1), the service system according to the present invention adopts a MySQL database to avoid introduction of noisy data for reasons such as data migration by another database and a history bug of the system and thus operational interference to a system

for efficiently processing multi-dimensional data based on group management, a requirement to add an extra specialized data cleaning process and impact on efficiency of data processing.

[0022] (3) On the basis of (2), the present invention requires that data from the three types of data sources be extracted in a unified manner by using a python script and stored in the PostgreSQL database, and that a data cleaning operation is performed on standardized data according to the service rule, so as to perform data screening and data repair operations and to ensure that the data is displayed accurately in the service system.

[0023] (4) On the basis of (3), after the unified data cleaning work is completed, the present invention performs in-depth processing and handling with the pandas tool based on an actual service requirement of group management service, and ultimately forms a data structure that is in line with the service requirement and easy to query, to be stored in various new data tables. Finally, the service system calls data in the database. A low-cost data analysis and query system is implemented. The system can be loaded with more than three hundred database tables of a group and implement processing and query of real service data to the greatest extent with high efficiency and low cost.

BRIEF DESCRIPTION OF THE DRAWINGS

[0024] FIG. 1 is a flowchart of a method for efficiently processing multi-dimensional data based on group management.

DETAILED DESCRIPTION OF EMBODIMENTS

[0025] Embodiment 1:

[0026] In the method for efficiently processing multi-dimensional data based on group management in this embodiment, as shown in FIG. 1, a method for parallel collection of multi-dimensional management data under a plurality of databases is established, data integration of multi-source database tables is performed on the collected multi-dimensional management data, data standardization processing is performed on an integrated database table, data subject to the standardization processing is called in a database of a service system, and data calling and query is performed according to a service rule in the service system.

[0027] A python-based pandas tool is used in the parallel collection to set connection attributes with the plurality of databases; and the connection attributes include a database table address, a user-defined login name and login password, and flag bit determination.

[0028] Specifically, data extraction includes the following specific processes:

[0029] S1: establishing a cache area of multi-source database tables, defining a python-based communication interface under a plurality of data sources, configuring a python-based connection attribute with Oracle, selecting a configuration port of the Oracle, obtaining a name and cursor information of a database table in the Oracle on this basis, call required data information in the Oracle and then close a cursor, and store the called data information in the cache area of the multi-source database tables;

[0030] S2: configuring a MySQL-based connection attribute, performing positioning and batch import operations of specified data in MySQL by using the pandas tool in python after successful connection to the MySQL, and specifying a data format of data to be imported; and

[0031] S3: on the basis of S2, the data format includes data in an Excel format, where the data in the Excel format is imported by specifying an Excel sheet type, obtaining quantities of row data and column data in the specified sheet type, establishing an index list based on the row data and the column data, and reading corresponding data in an Excel sheet one by one through the index list.

[0032] A unified calling logic based on required data in the service system is established in the data standardization processing, and the unified calling logic is used for data application in the service system.

[0033] In a constraint of the service rule, data that is standardized according to the service rule is subject to unified data cleaning. During the data cleaning, if a total amount of an order is inconsistent with a summary of details of a payment method, corroboration is performed based on details of charging items, and attempts are made to repair data. If the data cannot be repaired, such data is discarded and a list is submitted to the Business Department for verification and correction. If the data is corrected, update is completed in the service system and the data is updated to a unified database through the extraction logic.

CLAIMS

WHAT IS CLAIMED IS:

1. A method for efficiently processing multi-dimensional data based on group management, specifically comprising: establishing a method for parallel collection of multi-dimensional management data under a plurality of databases, collecting metadata of the multi-dimensional management data under the plurality of databases, performing data integration of multi-source database tables on the collected multi-dimensional management data, performing data standardization processing on an integrated database table, calling data subject to the standardization processing in a database of a service system, and performing data calling and query according to a service rule in the service system;

wherein the data integration of the multi-source database tables comprises the following specific steps:

A1: performing labeled classification on database tables under the plurality of databases, and performing clustering processing on the database tables based on a label in the labeled classification;

A2: dividing the clustering processing into parallel label clustering and tree-structured label clustering on the basis of A1;

A3: when performing the parallel collection of the multi-dimensional management data under the plurality of databases, performing classification processing on the parallel label clustering and the tree-structured label clustering, respectively, based on the label by loop traversal; and

A4: on the basis of A3, performing merging of the database tables under the plurality of databases, and performing the data standardization processing on a merged database table, wherein the data standardization processing comprises data cleaning, data deduplication, data format unification, and abnormal data value rejection.

2. The method for efficiently processing multi-dimensional data based on group management according to claim 1, wherein the data calling and query comprises data alert, data notification, and standardized quantized data display.

3. The method for efficiently processing multi-dimensional data based on group management according to claim 1, wherein the plurality of databases comprise MySQL, Oracle and Excel databases; and in the method for parallel collection, a cache area of the multi-source database tables is established, and the data integration and the data standardization processing are performed on the collected multi-dimensional management data in the cache area of the multi-source database tables.

4. The method for efficiently processing multi-dimensional data based on group management according to claim 3, wherein a python-based pandas tool is used in the parallel collection to set connection attributes with the plurality of databases; and the connection attributes comprise a database table address, a user-defined login name and login password, and flag bit determination.

5. The method for efficiently processing multi-dimensional data based on group management according to claim 1, wherein in the parallel label clustering and the tree-structured label clustering, labels in the database tables are graded according to the service rule and labels at a same level fall into a same set in the database tables under the plurality of databases.

6. The method for efficiently processing multi-dimensional data based on group management according to claim 5, wherein in the tree-structured label clustering, attribute relationships of labels at all levels fall into a set.

7. The method for efficiently processing multi-dimensional data based on group management according to claim 1, wherein in the data standardization processing, a unified calling logic based on required data in the service system is established, and the unified calling logic is used for data application in the service system.

8. The method for efficiently processing multi-dimensional data based on group management according to claim 1, wherein in the data standardization processing, processed data is stored in a cache area of the multi-source database tables, and data in the cache area of the multi-source database tables is called by a PostgreSQL database.