An Integration Model of Multi-Source Heterogeneous Audit Data

Li Chunqiang
School of Computer Science and Technology
BIT, Beijing , China
School of Information Management ,BISTU,
Beijing , China

Chai Weiyan
School of Computer Science and Technology
BIT, Beijing , China
Chen Linan
Network Information Center ,BUPT,
Beijing , China

Abstract—**This paper shows the problem of multi-source heterogeneous data integration in the audit data storage and access problems. Based on the analyses of existing data integration model, combining with the service-oriented architecture theory, this paper proposes a data center pattern, which is suitable for multi-source heterogeneous data integration model to audit. Furthermore this paper puts forward an incremental based technology to show the problem of synchronization of business audit data integration model, and expounds the data management mode based on Middleware technology.**

Keywords- heterogeneous database; data integration; middleware; audit

I. INTRODUCTION

In the data integration management, how to achieve flexible control, data storage and access to efficient management; how to combine the data characteristics of the audit industry, select the appropriate data integration model; how to make the integrated system more convenient for the upper application system providing service, have been paying more and more attention.

To the above problems, based on the analysis of the existing data integration models, combining with the service-oriented architecture theory, this paper puts forward a suitable model for multi-source heterogeneous data integration of audit data center; further based on analyses about the hierarchical structure of the model, provides the basis for data integration application stage data display, query, analysis.

II. THE MODEL OF DATA INTEGRATION SYSTEM

Data integration is the deepest and the core of the work in the integration of the information system construction, which integrates all kinds of data together, provides favorable transparent access to user interface[1-2]. Because of the complexity of decentralized data, diverse sources of data location and diverse formats in the integration domain, data integration system needs to support a variety of heterogeneous data access, foreign to shield the difference of various data sources, access interface has nothing to do with the data source. The data is transparent to the user, when the user request data, only concerned with any interface call what data can be, without concern for the interface is how to achieve access to heterogeneous data. The general data integration system model is shown in figure 1.

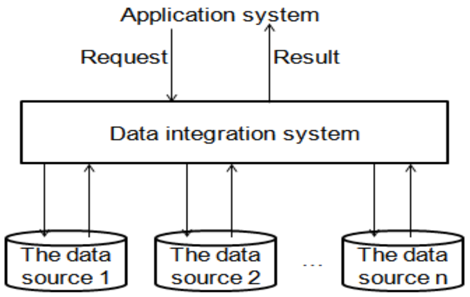

Figure 1. model of data integration system

In order to effectively manag large amounts of data,according to the need of convenient and fast access to the required data, most of the data model store data using rigorous approaches in the database. However, the database has many different types and each with different data models or access interface, which has produced data uniform access problems among heterogeneous databases.

In data storage and access stage, data integration problem to be solved is the main structure of heterogeneous and different database to provide uniform access problems.The main solution to these problems is the database, the federated database technology, database middleware,data warehouse[3-4].

Database technology,centralized management of data,the user according to the client API,use the SQL statement can acquire all kinds of data, but the client driver too,programming efficiency is low;the federal database, between different data sources use the data interfaceto realize the data access,via a data source can access any other data source information,but required data source interface to write too much,heavy workload,poor scalability;database middleware, user query isbased on the intermediate mode,do not need to know the location of the data source,mode or access,good portability,convenient integrationand expansion,but in the database connection sharing,high requirements for middleware stability,concurrent access number bymiddleware connection pool limit; data warehouse,data subject orientedintegration,can keep stable,can show the historical changes,but thedata could not be additions and deletions,mainly used for data query and analysis.

III. DATA INTEGRATION TECHNOLOGY OF DATA CENTER_BASED PATTERN

Audit object and audit work in industry are different,the corresponding data is also rich and varied, multi-source heterogeneous data in a distributed environment, how to conduct effective integration,become an urgent problem to be solved.At the same time, the framework of the application system has become more and more complex,heterogeneous data from multiple databases from processing,to come from a variety of platform functions can work;from the analysis of data,to implement complex business logic in the field; from a single application to adistributed multi server cluster application,all kinds of application more and more complex.On the other hand,component technology matures,Web Service technology becomes more and more popular,thedevelopment with high scalability, and to meet the data management model of program integration and application of heterogeneous data integration is possible.

The data center is a variety of data integration and exchange center,using service oriented architecture system, which is a comprehensive application platform in basic and applied as one of the.The data center to achieve a unified,multi-source heterogeneous data hierarchy management,can achieve coordination of spatial data and non spatial data in a unified framework,build support integrated application solutions, provides basic support for application system.

Data center is a compatibility data warehouse, which can be in the sameframe,but also has the multi-source heterogeneous data distributed management ability.Data center also has a well-defined function warehouse,supporting in a variety of ways (components,plug-ins,process,dynamic database,program fragment,script) to provide functional application, provide the call and execution of these functions in a consistent way [5-6].Data center architecture is shown in figure 2.

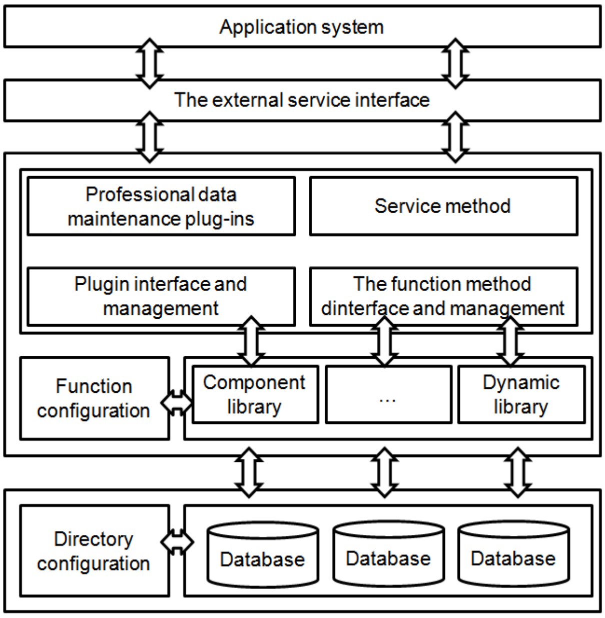
Application system
The external service interface
Professional data maintenance plug-ins
Service method
Plugin interface and management
The function method dinterface and management
Function configuration
Component library
...
Dynamic library
Directory configuration
Database
Database
Database

Figure 2. data center architecture diagram

Data resources through the management of data warehouse, data resources include a variety of GIS data(such as Map GIS6X,Oracle Spatial data),the data in the database(Oracle,SQLSever,Access etc.)and various document data(such as Word,PDF,Excel,Access,image),and supports users to extend or custom data types.

IV. AUDIT DATA INTEGRATION MODEL

The construction of the audit data integration system is to achieve the most core part of data integration technology, system model design is good or bad, is directly related with the data management efficiency and scope. This paper uses SOA architecture theory of service oriented and based on data center pattern data integration technology, according to business characteristics of the auditing profession, proposes audit data integration model as shown in Figure 3.

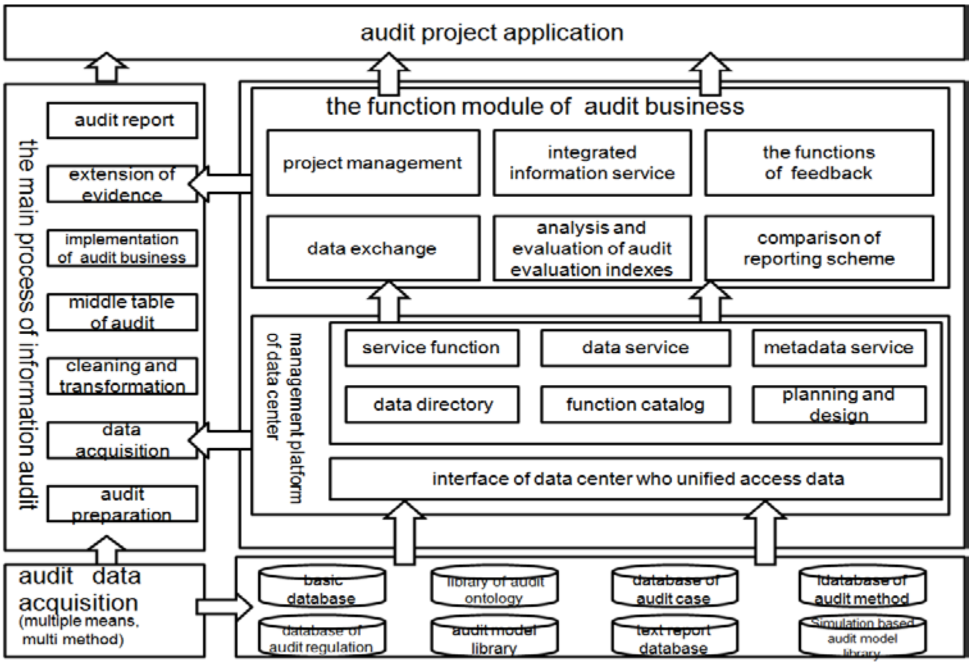
audit project application
the main process of information audit
audit report
extension of evidence
implementation of audit business
middle table of audit
cleaning and transformation
data acquisition
audit preparation
the function module of audit business
project management
integrated information service
the functions of feedback
data exchange
analysis and evaluation of audit evaluation indexes
comparison of reporting scheme
management platform of data center
service function
data service
metadata service
data directory
function catalog
planning and design
interface of data center who unified access data
audit data acquisition (multiple means, multi method)
basic database
library of audit ontology
database of audit case
database of audit method
database of audit regulation
audit model library
text report database
Simulation based audit model library

Figure 3. Audit data integration model

This model uses the C/S and B/S using the hybrid approach to achieve. For users of the application needs, all related to the ordinary Web system data analysis and other complicated difficult to complete the work, achieved by the C/S system, and C/S system analysis results, can be directly by using the B/S system, data dissemination and query.

The module in this model is as follows:

(1) the data source layer. The original data auditing face different industries, different audit object will produce all kinds of data, uniform summary to integrated management system, part of the data according to the work need, direct usage.

(2) data center management platform. The platform is divided into two categories, according to data center management model, a class is a data warehouse, a class is a function of the warehouse. A data warehouse contains all kinds of professional and management data, mainly include: the basic data, the audit body, audit case library data repository data, methods of audit, audit information library data simulation based model library data, text data, audit report base model base and audit law library data etc.. These data can be in the data center based on the unified data access interface for various applications, call. The function of warehouse data center covers the function of various applications, support rules design, build and unified process, in the form of services announced.

(3) audit function service module. This module is mainly aimed at the characteristics of the audit profession and provide service. These modules services include: project management, integrated information service and intelligent feedback, data exchange, audit evaluation index analysis and evaluation and scheme comparison report etc.. These business functions is the core module of audit unit informatization construction, the construction of these modules can greatly improve the audit unit management level.

(4) the main process data information audit. Because of the particularity of the auditing profession, for audit significant audit results must be reported to the state, so for the audit unit not only to the audited units to understand clearly, more important is related to the competent units have requirements for audit results to. This module provides audit of the entire business process, including the audit preparation, data collection, conversion, cleaning validation implementation of the audit business (select key, modeling analysis, generating papers), extending evidence, the audit report process.

V. THE KEY TECHNOLOGY IN AUDIT DATA INTEGRATION MODEL

Data integration is the key point data and heterogeneous. Therefore, in the integrated system should be to reduce the amount of data transmission, the realization of heterogeneous data conversion for the basic principles[7-8].

A. Business synchronization

Because of different audit organ and the data center end physical distance, network transmission delay, and all the historical data, the audit authority data from various audited units to submit regularly, so it is difficult to use real-time updates on the way to ensure the data consistency of data source and data center end, so in this paper, the use of asynchronous mode data update. The data consistency of model using data integration model as shown in figure 4

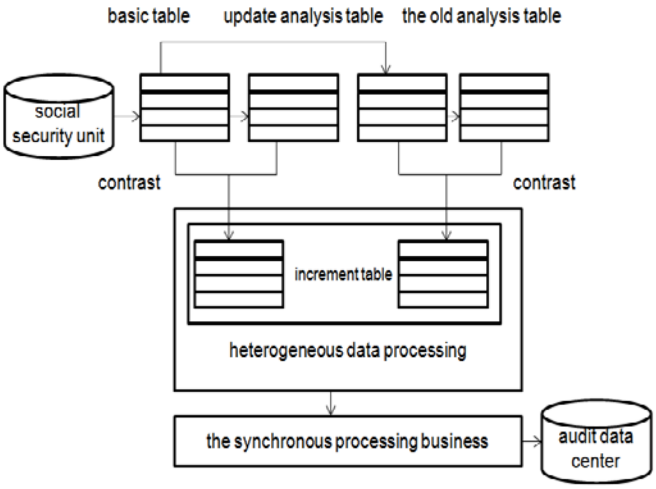

Figure 4. the data consistency model data in integration model

Processing of incremental data including basic data of incremental data acquisition and analysis of data acquisition part two increment, the two part of the process is as follows.

The basic data of incremental data acquisition:

(1) loading unit submitted data to the new base;

(2) to obtain the incremental data through the new and old base library comparison;

(3) the resulting incremental data into basic data incremental database table;

(4) the new foundation library mark the old library, empty the old foundation library data;

(5) repeat steps (1) - (4).

Analysis of the data of incremental data acquisition:

(1) the foundation library every time after loading the data re execute audit methods, the resulting analysis data to insert new analysis library analysis table;

(2) the old and new analysis library comparing incremental data;

(3) the resulting incremental data into the data analysis of incremental database table;

(4) the new analysis library standard for the old library, empty analysis library data;

(5) repeat steps (1) - (4).

B. *Conversion technology of heterogeneous data*

Data center using middleware technology realizes integrated management of all kinds of data shown in figure 5.

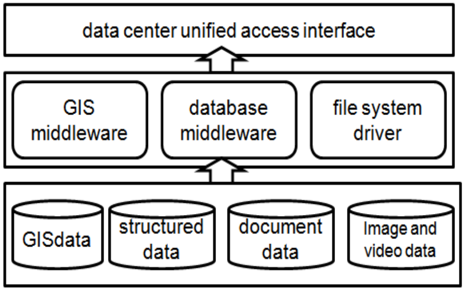

Figure 5. Data center Middleware Technology

Middleware including GIS middleware, database middleware, file drive three kinds, basically covers all the accessible data drive. The intermediate operation mode based on the system, shielding the heterogeneity of various data sources, access interface by using a unified data center data, users need not care about the location of the data, format, driving, can access various data unified, transparent, consistent, achieve seamless integration of data, can achieve the hybrid analysis of heterogeneous data, is the core of the local data center integration.

VI. EXPERIMENT

The data integration system for the national audit will be located in the provinces and cities, the auditing data concentration to audit office data processing center, in this process, through the data conversion module and a synchronous processing module, ensure audit data in audit data warehouse to meet a unified data standard, and has the same business schedule.

A. *The experimental scheme*

In order to verify the effect of the integration of the data integration system, in five provinces and cities in A Province, province of B, C, D City, E City, as the simulation object, the 2011 audit data integration to audit data warehouse. After the statistics of SQL, gets the business 2007 in all regions of the cut-off time see table I.

TABLE I. AROUND 2007 DATA SERVICE CUT-OFF TIME

region name	business deadline
A province	200705
B province	200705
C province	200706
D city	200703
E city	200703

After an audit index data statistics, data conversion and business synchronization of three processes, 15 analysis tables into data warehouse. Select the number of basic old-age insurance support ratio and average amount of dependency ratio analysis table experiment demonstration, as shown in table 2.（H:area,I:yera,J:month,K: receive times,L: payment times,M: the number of dependency ratio,O: average treatment,P: average payment,Q:amount of the dependency ratio）

TABLE II. THE NUMBER OF BASIC OLD-AGE INSURANCE SUPPORT RATIO AND AVERAGE AMOUNT OF DEPENDENCY RATIO ANALYSIS TABLE

H	I	J	K	L	M	O	P	Q
A	2007	1	992,981	3,578,130	0.45	494.63	216.1	2.29
A	2007	2	996,978	2,683,267	0.45	498.46	266.7	1.87
A	2007	3	992,057	2,693,077	0.45	493.23	216.7	1.81
B	2007	1	52,451	50,604	1.03	863.87	106.3	8.21
B	2007	2	52,276	50,367	1.03	493.81	106.7	7.15
B	2007	3	55,875	50,323	1.11	753.68	106.5	7.01
C	2007	1	29,745	27,367	1.65	733.87	916.7	1.38
C	2007	2	35,389	21,481	1.67	1273.67	716.7	1.63
C	2007	3	35,389	133,367	0.55	1263.82	686.7	1.89
D	2007	1	85,383	673,239	0.35	5882.23	1556.7	28.09
D	2007	2	42,369	683,326	0.34	9444.54	1553.4	48.09
D	2007	3	42,383	693,378	0.45	9560.87	1516.7	46.09
E	2007	1	30,372	54,327	4.45	3264.42	1598.2	12.05
E	2007	2	28,283	56,323	4.45	3280.80	1476.9	13.45
E	2007	3	29,382	73,367	3.45	3260.17	1516.7	13.05

From table II, it is not difficult to see that, although around the actual business progress is not consistent, but after storage of data by the end of 2007 March, more than the March statistics did not import data warehouse. Then on the basis of the establishment of multidimensional model.

B. *Experimental results and analysis*

By constructing the number in 2007 than the average amount of maintenance and support more than two multidimensional model, obtains the column chart visually as shown in Figure 6 and 7. Among them, Figure 6 for 2011 5 provinces and cities the number of dependency ratio, Figure 7 is the average amount of dependency ratio in 2011.

Seen from the two picture is not difficult, the two model business time were only 1, 2, March, business progress, not an exact statistical data phenomenon does not exist.

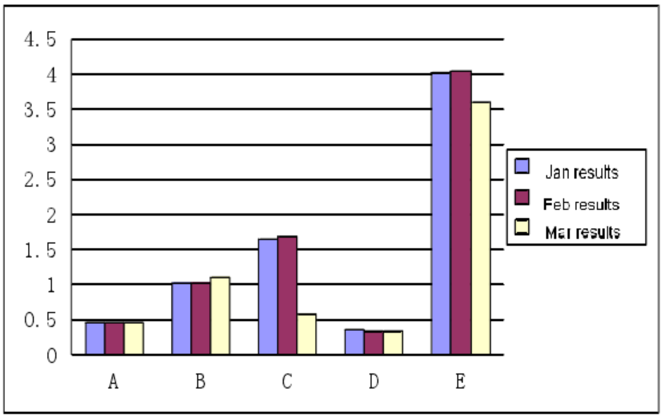

Figure 6. The number of dependency ratio

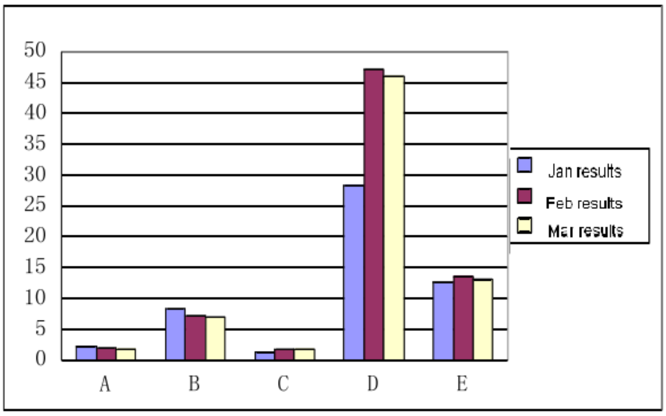

Figure 7. The average amount of dependency ratio

The experimental result shows that this integrated system can not only realize data physical integration, but also can provide business logic synchronization time according to business synchronization module, facilitate various audit of the audit staff.

VII. CONCLUSION

Based on the analysis of existing data integration model, combining with the service-oriented architecture theory, this paper puts forward a suitable model of multi-source heterogeneous data integration data center audit mode, and based on the given business progress, asynchronous data integration process of the corresponding solution, puts forward the implementation process of the core module.

VIII. ACKNOWLEDGEMENT

It is a project supported by National Key Technology Research and Development Program of the Ministry of Science and Technology of China (2012BAH08B02). The corresponding authors are Li Chunqiang and Zhao Gang.

[1] Claudia Wild,Judit Erdös,Marisa Warmuth,Gerda Hinterreiter,Peter Krämer,Patrice Chalon. PLANNED AND ONGOING PROJECTS (POP) DATABASE: DEVELOPMENT AND RESULTS. International Journal of Technology Assessment in Health Care,2015,305.

[2] Wei Xianmin. Heterogeneous Database Integration Middleware Based on Web Services. International Conference on Applied Physics and Industrial Engineering 2012, Pt B. 2012, 24 877-882

[3] Bruno Tomazela,Carmem Satie Hara,Ricardo Rodrigues Ciferri,Cristina Dutra de Aguiar Ciferri. Integration processes with data provenance. Data &amp; Knowledge Engineering,2014,.

[4] Mohania M, Bhide M. New trends in information integration. In Proceedings of the 2nd international conference on Ubiquitous information management and communication. Suwon, Korea ACM, 2008. 74-81

[5] Christoph Quix,Matthias Jarke. Information Integration in Research Information Systems. Procedia Computer Science,2014,33.

[6] Mouhni Naoual,Elkalay Abderrafiaa. Semantic Technologies Applying to Data Warehouses Federation. Journal of Emerging Technologies in Web Intelligence,2014,61.

[7] Mark Scott,Richard P. Boardman,Philippa A. Reed,Simon J. Cox. Managing heterogeneous datasets. Information Systems,2014,44.

[8] Igor Timko,Curtis Dyreson,Torben Bach Pedersen. A probabilistic data model and algebra for location-based data warehouses and their implementation. GeoInformatica,2014,182.