Integration Scheme and Strategy Research of Heterogeneous Database

Bin CHEN[1, a], Dawei MA[2]

[1] Graduate School, Chongqing Communication Institute, Chongqing, 400035, China

[2] National 3G Center, Chongqing Communication Institute, Chongqing, 400035, China

[a]email: plato33hd@sina.com

Keywords: Heterogeneous, Database, Integration, Scheme, Strategy

Abstract. The integration of heterogeneous database as the important problem of heterogeneous data integration is catching more and more emphasis by the people. This paper will analyze the character of the heterogeneous database, and propose the scheme and strategy to realize integration of heterogeneous database on this basis.

General Introduction

Data integration is a process to abstract heterogeneous, distributed and conflicting source data, make corresponding transformation of data and structure, and upload to target system. Actually, the process of data integration is to eliminate the difference and conflict between source data and target data, and integrate according to the requirement of a target system, for example, integration of name, data mode and semanteme. The data after integration will use a standard form provided by user, which is good for data applications such as data warehouse and data mining by the user.

The problem of data integration has a long history, with the expanding of computer application field, the spread of data is getting more and more widely, centralized data storing has gradually substituted by distributed data storing. However, the distributed data interacts with each other closely; the data isn't stored on the lonely island without any relationship. The user always has to centralize the distributed data by some kind of requirement in detail applications. For example, the business data of large-scale enterprise is stored in database of different branches respectively, in order to master the sales situation of the whole enterprise, the data in sub-system needs to be centralized together so as to satisfy some particular needs, such as data warehouse, data mining and etc., also to reach standardization and uniform of the data.

The user will face a big difference of the data during data integration, the main reason for generation of data difference is because the conflict between data structure and semanteme. The source data could be relationship type, also could be object type, as well as WEB page type and text type. Therefore, in order to solve the problem of data integration, one big question is how to remove these differences. With the generation of large amount data, the conflict between data structure and semanteme will become more seriously, how to solve kinds of conflicts effectively is becoming a big challenge for data integration.

Heterogeneity of Data

Heterogeneous data is a conception with rich meaning; it refers to the data which belongs to the same type, but different handling method. In terms of content, it not only refers to the data in different database system is heterogeneous, such as the data in Oracle and SQL Server database, but also refers to the heterogeneity among data with different structure, such as data in structured SQL Server database and semi-structured XML data.

In general, the heterogeneity of data could be involved the bellowing three aspects: system heterogeneity, data mode heterogeneity and logic heterogeneity.

System heterogeneity refers to differences of hardware platform, operation system, concurrency control, visit method and etc., the details are following:

1) The difference of computer system structure means the data could be stored respectively in large machine, small machine, work station, PC or embedded system.
2) The difference of operation system means the operation system could be Microsoft Windows

© 2016. The authors - Published by Atlantis Press

NT, kinds version of UNIX, IBMOS/2, Macintosh, etc.

3) The difference of developed language, such as C, C++, Java, Delphi, etc.
4) The difference of network platform, for example, Ethernet, FDDI, ATM, TCP/IP, IPX/SPX, etc.

However, the heterogeneity of data mode refers to the difference of DMBS itself. For instance, data integration system could adopt Oracle, SQL Server which shares the same relationship type database as the data model, also could adopt different type of database to uniform the relationship, layer, network, object or function type of database and so on.

Logic heterogeneous includes name heterogeneity, value heterogeneity, semanteme heterogeneity, mode heterogeneity, etc. Such as, the detail of semanteme heterogeneity lies in same data form shows different meaning, or same meaning is shown by different form of data.

The above mentioned constitute the heterogeneity of data, the technology of data integration is just developed for the heterogeneity of data.

Heterogeneous Database Integration

The problem of heterogeneous database integration as the important application of heterogeneous data integration, is the highlight problem of data resource sharing visit with the development of Internet and WWW, the users need to solve this problem eagerly, which means to realize transparent visit to the data with minimum expense. Use the technology of heterogeneous database integration, could integrate different physic type database, different data type database, database with same data model and different producer, as well as different version of database with same producer, and database products aim to different network environment, etc.

Integration Scheme of Heterogeneous Database

The integration scheme of heterogeneous database could be divided into two forms generally. The first one is to transplant the original data to a new data management system, in order to integrate different type of data, the traditional type of data should be transformed to new type of data. Actually, many suppliers of relationship type database could provide similar service. The drawback of this integration scheme is, with the upgrading of data management system, the relevant application software for origin data, they will be abolished or re-developed so as to adapt new data management system. So, transplant to a new system usually is not a realistic solution.

The second scheme is to use middleware technology to integrate the heterogeneous database, this scheme doesn’t need to change the storing and management way of origin data. Middleware lies in heterogeneous database system and application program, down to coordinate each database system, up to provide uniform data mode for the application of integrated data visit as well as general data visit connection. Each application of database still fulfils their duties and the middleware system is mainly to provide high-level searching service for heterogeneous data source. Obviously, middleware system mode is the ideal solution to realize heterogeneous database integration.

The most common detail database integration method has three kinds:

1) Federated Database

Federated database is the easiest structure of database integration. Its constitution method is to link the component database one on one. It will generate the following problems: if each one in n database needs to realize mutual operation with other n-1, the developer should write n (n-1) pieces of code to support searching and visit between them. Moreover, this mutual operation is partial constraint interaction, which could not realize flexible integration of each database.

2) Data Warehouse

Data warehouse is to store the copy data from several data source in a single database. In this kind of structure, the data is abstracted out from all the data source, combine into a global mode, and store in the data warehouse; it seems no difference with general database system for the users. Data warehouse supports the visit of history data. The user could also search the decisive support through the uniform data connection provided by data warehouse. But this method has many shortcomings: for example, the data in the data warehouse needs screening process before storage, the data warehouse needs periodic updating, the users are not allowed to update the data warehouse

generally, because these updating can't reflect in the basic data source.

3) Mediation

Mediation is a kind of software component, which could support virtual view or back view collection. This integration method is similar to the method of materialization type of integration data source in data warehouse, but it can't store any real data. The effect of mediation is, to translate the user's searching into one or multi searching to the data source after the user submitting a searching requisition. Then, make comprehensive handling for the searching structure of the data source, and return to the user. Actually, this kind of method is middleware system mode which mentioned above.

All in all, no matter which kind of realization scheme, their common point is to keep autonomy of component database, which means to keep still of local definition, local application and local strategy for data exchanging with other database. The integration frame of heterogeneous database is shown in Fig.1.

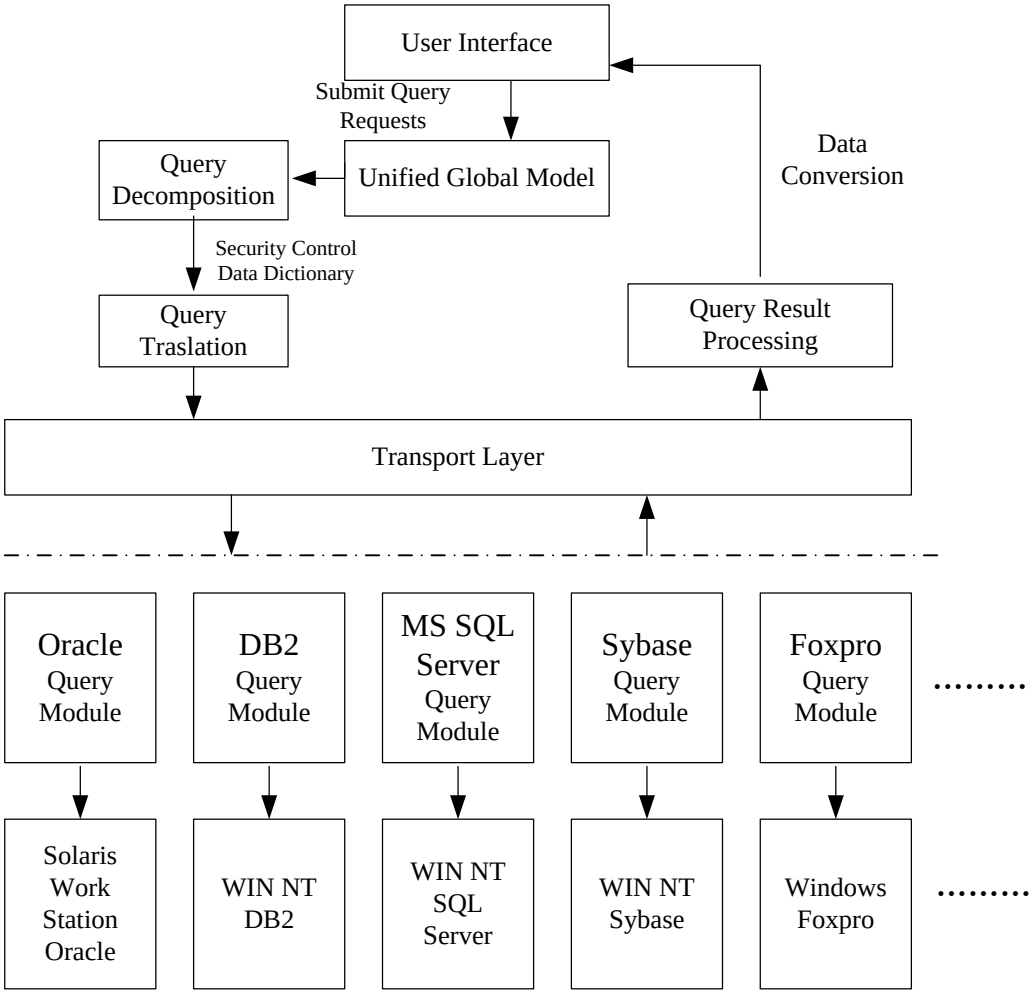

Fig.1. Integration Frame of Heterogeneous Database

Realization Strategy of Heterogeneous Database Integration

In view of the variety of problems in database integration, the problems of transparency of platform and network, transformation of data model and mode should be solved during the integration process. Generally speaking, heterogeneous database integration could be realized through transformation and standardization, there are two kinds of heterogeneous database integration strategies based on different level.

For the database with heterogeneity on DBMS, there are three key solutions: database gateway, common protocol and common programming interface.

Database gateway is a data transformation interface of different DBMS with their own, customer could visit heterogeneous database through it. Most database suppliers like Oracle, Sybase, SQL Server, etc. all have their own gateway products. Such as Sybase Client/Server Interface allows a

variety of customer terminal to visit kinds of data source. Sybase Gateway provides instant integration between Sybase and other SQL type database (eg. Oracle and Informix). Meanwhile, Sybase OmniSQL Gateway provides complete location and product transparency, to integrate different data source in different servers, in order to give a uniform logic entirety.

Adopt the integration method of database gateway could give a satisfied solution to the problem of mutual operation in database. We could link the third-party development tool to self database product through gateway. The user shall use the gateway to link them with new database technology without abolishment of the present application program, in order to protect the existed investment. However, in this complex system with n database, if you want to realize mutual operation between any two databases, n (n-1)/2 gateway should be prepared, in addition, the database gateway is expensive, so, it's difficult for practice. Moreover, data format for some heterogeneous database, grammar or semanteme transformation are unfeasible, utilize database gateway to visit nonlocal heterogeneous database is not easy to attain complete transparency. In a result, the method of database gateway integration is effective for several heterogeneous database integration, once the scope expands, the practicability will be reduced.

Use common protocol to integrate heterogeneous database refers to standardize the communication format and protocol between customer and server, as well as database language, this is comparatively an ideal solution to heterogeneous database integration method. So far, the typical one is distributed type of relationship database system structure (DRDA) which SAG (SQL Access Group) rules IBM.

Common programming interface includes customer application programming interface (CAPI) and server application programming interface (SAPI). CAPI is a group of process base, generally stay in customer's work station by terms of TSR method or DLL method, one CAPI could transship the driving program especially for back terminal, in order to visit different data source. However, SAPI only provides one application programming interface, to control server and customer application requisition, as well as interaction among target database.

For example, ODBC (Open Database Connection) is a well-known public programming interface, which is the database of Microsoft Windows open server architecture, a uniform standard interface for variety of database visit. Actually, ODBC is a database visit base, but only provides one uniform application programming interface. Use ODBC could prevent generating changes by the change of application program. ODBC acquires data dependency through database driving program, and the standard interface which provided by the driving program permits data transmission between application program and data source by the program developer and driving program provider.

ODBC provides the application program with a set of high-level interface rule and support environment for running based on dynamic linkage base. Generally, the often used front-stage tool for database application development, such as Power Builder, Delphi, etc. link variety of database system through ODBC interface. Moreover, most of the database management system like Oracle, Sybase, SQL Server and etc. all provide corresponding ODBC driving program, enables the database system to possess good openness. Fig.2. shows ODBC as the public programming interface to realize linkage among heterogeneous database.

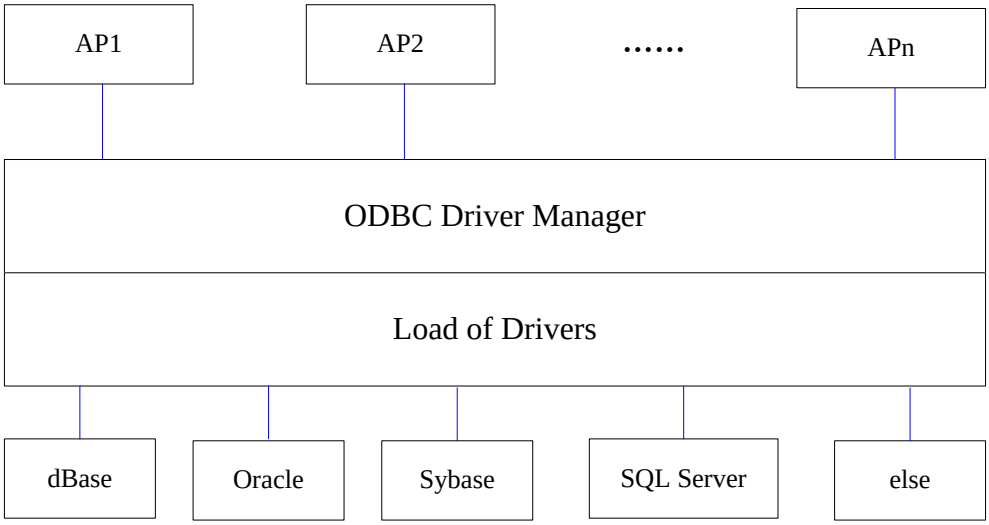

Fig.2. Use ODBC to Realize Interaction among Heterogeneous Database

For semanteme heterogeneous database, the general work at the early stage is to realize integration through defining a conception model or meta-model, which means to transform the heterogeneous data mode to a traditional mode directly, but the process is complicated with little effectiveness. Because during the transformation process, the system may have to abandon some component database eventually, in order to establish common global name, data structure, property value and strategy. All the existing application programs on component database should be transformed, so as to obey new elements after clustering. This transformation process usually pays a lot, but unfeasible perhaps.

In order to improve the tremendous expense made by this method, the forward researches could be divided into two type, close coupled method and loose coupled method.

Close coupled method is established on the basis of database mode integration, this kind of heterogeneous database system has an integrated global mode, for example, Mutibase, Mermaid, Adds, DatapleY, Ingers/Star, Pegasus and WIND system. The advantage of this method is: it permits user to visit heterogeneous, autonomous, distributed heterogeneous database transparently, just like they are a centralized database. Nevertheless, it will cost tremendously to realize this method, and every time, the existing component database mode changes or every new database added into the system, the global mode should be created again. So, it’s very difficult to maintain the global mode. Under a dynamic environment, like Internet, actually, it’s impossible to maintain such a global mode.

Correspondingly, loose coupled method will transmit the integration mission to the user. The mission of the system which adopts this method is to develop a set of tool, make sure the user could use the tool to finish integration work. Such as operation language MDSL of ALCHEMIST and MultiDatabase which face to the object. Although this method is very flexible in updating the component database, it also brings too much burden to the user, who should get familiar with the content and structure of component database. When the quantity of component database is increasing constantly, the user actually still fails to know the relevant information of all component database which needed by searching.

Therefore, under such circumstance, there is an adjustment method developed recently which is based on knowledge base system. It combines the basic elements of close and loose coupled two aspects, the meaning is to use a mediator (mainly for searching) with knowledge base capacity to substitute global mode, the knowledge base system uses rules to identity and handle the type and semanteme of the data, so as to attain good integration and assimilation effectiveness. The mediator defines as: “a kind of knowledge which uses or relates to a database collection or sub-collection through coding, a software module which provides service to high-level application program.” It relieves the burdens of direct interaction with heterogeneous database by the user. The user only needs to interact with an agent, and provides the searching condition in agent’s language. This method has high flexibility and extensibility, because the metadata (the metadata in heterogeneous

database system includes: mode information of each local database, global view information of integrated system and transformation rules of heterogeneous modes) of each component database enters the knowledge agent independently; the typical examples are Carnot and SIMS. In addition, it's used to visit structured and non-structured data based on the method of mediator.

Conclusion

The utilization scope of the above several methods depend on the integrated database's quantity, heterogeneous difference extent, stability extent of database, as well as user's application level. Generally, close coupled method is appropriate to less quantity of database, the metadata doesn't change frequently; loose coupled method is appropriate to high level user or the user who has support of database expert. Under a dynamic environment, one aspect is tremendous quantity of database, the other aspect is the data keeping altering, the user is difficult to maintain the metadata, just that reason, adjustment method based on knowledge base is working.

References

[1] Liu J, Liang S, Ye D, et al. ETL Workflow Analysis and Verification Using Backwards Constraint Propagation[C]//Proceedings of the 21st International Conference on Advanced Information Systems (CAiSE'09). Springer, 455-469.

[2] Zhou C, Chen H, Tao J. GRAPH: a domain ontology-based semantic graph auto extraction system [J]. Applied Mathematics & Information Sciences. 2011.

[3] Halder R, Pal S, Cortesi A. Watermarking Tecjniques for Relational Databases: Syrvey, Classification and Comparsion [J]. The Journal of Universal Computer Science. 2010, 21(16): 3164-3190.

[4] Zhang Wenjiang. Research on security data dictionary based data intergrating technique [J]. Computer Engineering and Design. 2013, 05.