Semantic Model based on Three-layered Metadata for Oil-gas Data Integration

Ma Zhonggui, Wang Chengyao, Wang Zongjie
School of Computer & Communication Engineering, Beijing University of Science and Technology, 100083 Beijing, China
e-mail: mzg1216@126.com, wangchengyao@ustb.edu.cn, wzj@senlang.com
doi : 10.4156/aiss.vol3.issue7.26

Abstract

Data integration is an old problem that reemerged as an active research topic recently, due to the increased requirements for exchange of data in various formats in our increasingly interconnected but inevitably heterogeneous world. On the basis of existing technologies, metadata is proposed to realize heterogeneous data integration. Firstly, three-layered metadata model is introduced, including data source metadata, business metadata and topic metadata. Based on this, four-layered data integration service framework based on metadata has been put forward, and it respectively includes application layer, access layer, data integration layer and data source layer from top to down. The application layer supports C/S and B/S applications by the client SDK, the access layer provides uniform data access interfaces for different users by RESTful Web Services. The data integration layer shields the heterogeneity among all data sources, and provides the transparent access support for users. Finally, multiple sources and heterogeneous information resource access platform is implemented with C#. Metadata can provide a clear standard for scattered and heterogeneous data so as to achieve integration and sharing of data, which makes the users access all sources by uniform data access interface transparently. Meanwhile, metadata can not only keep the data consistency, but also provide better expansibility.

Keywords: Data Integration, Semantic Model, Metadata, Metadata Model

1. Introduction

Data involved in exploration and development of petroleum and natural gas is important resources for the oil-gas field enterprises, which is the basis for exploration, evaluation and development. Because exploration and development involve many disciplines, they will generate large amounts of data. How to conveniently and efficiently use these data resources, increase production of petroleum and natural gas, which is an important issue for asset-like management and application of data involved in exploration and development of the oil-gas fields.

In order to realize coordination among all departments in the oil-gas field enterprise, and timely, accurate and efficient information management, we must consider data exchange and integration. To meet the need, the information resource access platform for heterogeneous data sources must be built. The success of heterogeneous data integration will directly affect the enterprise development in the competitive market.

Data integration is an old problem that reemerged as an active research topic recently, due to the increased requirements for exchange of data in various formats in our increasingly interconnected but inevitably heterogeneous world. More often, however, the necessary data is distributed over multiple sources in the oil-gas fields. Not surprisingly, the sources often use different vocabularies and different data structures, being created, as they are, by different people, at different times, for different purposes. In such cases, data integration is necessary to combine data from these sources [1, 2].

On the basis of the data integration, another important trend is knowledge management and knowledge integration, and intelligent services are provided for the oil-gas field enterprises by the knowledge base and knowledge management systems. InTouch service is an example of knowledge management, which provides a framework upon which to implement effective solutions, and improves business communication efficiency by accelerating acquisition, share and refining of knowledge [3]. WesternGeco, a business segment of Schlumberger, is the world's leading geophysical services...
company, providing comprehensive worldwide reservoir imaging, monitoring, and development services, with the most extensive geophysical survey crews and data processing centers in the industry, as well as the world's largest multi-client data library [4].

The goal of data integration is to provide programmatic and human users with integrated access to multiple, heterogeneous data sources, giving each user the illusion of a single, homogeneous database designed for his or her specific needs [5].

In order to support the heterogeneous data integration, people have already put forward many technologies and solutions. Metadata is used to identify, describe and locate data. In the heterogeneous data integration, metadata can be used to describe variously heterogeneous data sources, which can shield heterogeneity of the integrated environment, and realize the loosely coupling application integration. Therefore, it will be a trend for heterogeneous data integration to use metadata.

2. A survey on data integration

Data integration is to concentrate data from different sources, different formats and different characteristics in logic or physical way, so as to provide comprehensive data share for enterprise, ensure the data consistency, and provide transparent access for the users.

With the development of computer technologies, people summarize various data integration frameworks. At present, the usually used techniques include ETL (extract, transform, and load), multi-database system, data warehouse and data brokering. These techniques provide decision support for the users from different aspects.

ETL approach is offline. The ETL approach involves extracting data from sources, transforming it to fit operational needs, and ultimately loading it into the end target. Multi-database system is a kind of systems which can accept and accommodate more heterogeneous database. It not only presents an integrated structure for the users outside, but allows each of heterogeneous database "autonomy" internally. The data warehouse is subject-oriented, integrated, non-volatile and time-variant. Typically, tools to extract, transform and load data into the repository are used, then other business professionals for data mining, online analytical processing and decision support can be directly executed in the data warehouse. This method directly operates on physical data, so it is of high efficiency. However, often new requirements necessitate gathering, cleaning and integrating new data from "data marts" that are tailoried. Data brokering is an integration modes with centralized management and distributed storage, which accesses heterogeneous database, legacy systems and web resources by globally uniform data schemas. Data brokering is located between application systems and heterogeneous data sources, and it not only coordinates all sources, but provides a uniform data access interface for all application systems. Data brokering is a kind of popular data integration methods, it provides a uniform logical views of data in middle layer to hide data details of all sources, and it allows the users to access all data by same interface. Data brokering offers integrated access to live data. Also, it preserves full information; when done correctly, no data is ever lost in the process.

Recently, metadata is receiving increasing attention from the scientific community, and it is an important ways to realize data standardization and data sharing, exchange and integration. First, all kinds of data resources in the oil-gas fields can be standardized using the metadata, and each professional field establishes its own metadata standards, then each specialty releases data outside according to this standard format. After this, it can improve efficiency and accuracy for data querying by metadata standards. Secondly, these metadata will record about all the context information, and data managers can efficiently manage data, and metadata contains information for data users to easily understand and effectively use. Finally, metadata can further eliminate heterogeneity of different data sources, and achieve data integration and exchange [11, 12].

Metadata records data sources, data targets and the conversion rules thus make the end users can easily understand the whole process of the data generation. Meanwhile, the end-user can easily find the quality problems of data, thereby it increases data reliability. Capturing, representing and processing metadata promises to facilitate the management, consistent use and understanding of data and thus better support the exploitation of available information [13].

Effective data sharing is the key to realize the comprehensive application for data, efficient data processing and data value-added. In this paper, an integrated service framework for heterogeneous
multi-source data based on metadata is put forward according to the detailed analysis about existing systems of the oil-gas fields, and a data resource access platform is built to integrate different systems of the oil-gas fields, so as to improve management and utilization of information resource, realize data sharing.

3. Three-layered metadata model

On the basis of existing data in the oil-gas fields, the establishment and use of three-layered metadata model can make it easy to query and locate data resources, and eliminate heterogeneity of multiple data sources. Meanwhile, it provides users with transparent and quick access to resources services. Three-layered metadata model is shown in figure 1.

As seen from figure 1, the metadata model is divided into three layers from bottom to up, respectively is data source metadata, business metadata and topic metadata.

According to representation of data involved in exploration and development of the oil-gas fields, data sources can be divided into four categories: structured data (relational database, such as SQL Server, Oracle), bulk data, document and graphic data.

3.1. Data source metadata

Data source metadata is designed to describe the existing data resources in the oil-gas fields, including access data interface and exposed data sets. If the data sources or data source structures have changed, data source metadata also changes automatically.

1) Access metadata of data sources

Access metadata of data sources is used for recording access information of data sources, such as the location, type, access mode and authentication information, etc. According to data sources needed to be integrated the access metadata of data sources can be also divided into four categories, including structured data, bulk data, document and graphic data.

2) Data set metadata

Data set metadata describes the structure of data sets, such as the name belonging to the data source, data set name, FTP or buck data file indexes, etc. They can be divided into tables, views, stored
procedures, SQL statements and others.

3) Data item metadata
Data item metadata describes the structure details of data items, such as the name belonging to the data set, data item name, data item type, primary key, foreign key, etc.

4) The category of data sets
The category of data sets is the logical classification of data sets, mainly used for management tools for metadata.

3.2. Business metadata

Business metadata is the collection of basic business objects provided by exploration and development of the oil-gas fields, metadata including business object category, business object, business object property, membership between business object category and business objects, relationship between or among business objects.

Business metadata is built for professional and technical personnel to use the data sources. To the professional and technical personnel being not familiar with the specific database, they can easily find the required data from multiple data sources by business metadata.

1) Business object metadata
Business object metadata records the basic business objects involving in exploration and development activities, includes business object name, corresponding class name, corresponding data set or sets, etc.

2) Property metadata of business object
Property metadata of business object describes the structure details of business objects, such as the name belonging to the business object, property name, corresponding data item or items, whether direct mapping or not, expression, etc.

3) The relationship metadata among business objects
The relationship metadata among business objects describes relations between business objects, which can be used to find associated objects by special business object, and it forms the foundation for the topic application. All relations can be mapped to relationship between primary keys and foreign keys of relational database, such as the one-to-one or one-to-many mapping, format conversion, etc.

4) The category of business objects
The category of business objects is the logical classification of business objects, mainly used for classification for business objects, and facilitates management tools for metadata.

5) The membership between the category of business objects and business objects
The membership between the category of business objects and business objects is used for describing business objects included by the category of business objects.

3.3. Topic metadata

Topic metadata is used to describe part of the business objects which users concerned, consisting of view of business objects and constraints. According to users’ needs, topic metadata is often generated dynamically, but for some typical topic, topic metadata can store persistent objects so as to make easy to query data.

1) Topic metadata
Topic metadata is used to describe what users concerned, is predefined application requirements by the user. Topic metadata is a view of business metadata, presents a series of related business object and their relations.

2) Business objects referred by topic
Business objects referred by topic are used to describe business objects including in the current topic, and it is only reference to real business objects, only records GUID of business objects. Real business objects will be depended when getting data out for users.

3) The referred relationship among business objects including in the current topic
Although the relationship among business objects have already been saved in business metadata layer, yet in order to top-down retrieve data, the relationship among business objects are referred in this layer.

4) Constraints imposed by the clients

There are attribute constraints, sorting constraints, querying conditions in constraints imposed by the clients. In addition to constraints such as “attribute of business object + predicate + attribute value”, external processing functions can also be used to expand by the metadata mapping.

5) Template

Template is used to save some of the business objects including in the topic metadata. How to exclusively define a particular metadata element must be considered in metadata management strategy, so GUID (Globally Unique identifier) is used to identify metadata elements, and it can ensure unique in time and space. At the same time, semantic understandings to each metadata element must be same.

4. An integrated service framework for heterogeneous data based on metadata

4.1. Data integration service framework based on metadata

A data integration service framework based on metadata maintains its functions in four layers: application, access, integration, and data source layer, it is shown in figure 2.

Figure 2. Data integration service framework based on the metadata

The application layer supports C/S and B/S applications by the client SDK, and also supports other data services to access seamlessly. All applications can directly use HTTP protocol to access data.

The access layer accesses the underlying data by uniform data access interfaces based on RESTful Web Services. Uniform data access interface has three functions: to provide a uniform interface for so as to shield the differences of the underlying data, and make the users use easily; to generate queries for data integration services according to the user's call; to get data from data synthesis for users.
Representational State Transfer (REST) is a software architectural style first laid out in the dissertation of Roy Fielding at the University of California Irvine [14]. REST is a set of constraints based on the architectural style for building services, which are based on resource design and uniform interface interaction. Services that follow this style are known as RESTful. RESTful services model the interaction with user agents based on resources. Each resource is represented by a unique URI, and the user agent uses the uniform interface of HTTP to interact with a resource via that URI.

In resource access strategies based on REST, the CRUD of data resources are converted to URL and HTTP flows, then send them to the information integration services, the server can judge access constraints and special operating type by parsing URL and HTTP flows, thus submit the user’s operation request to the query planning and scheduling module.

In addition, URL using the REST has strong legibility, for example, when querying the well business object where WellNo='5', the URL can be written in the following: "http://datasource.service/bo/well/5". Compared to the web service based on SOAP, RESTful web services has more quickly parsing speed.

The data integration layer is located in the middle of the access layer and data source layer, it is used to integrate various heterogeneous data sources, and shield the differences between them and provide the transparent access to the heterogeneous data sources. Data integration layer is composed of two main components: data integration services and data access services.

Using metadata model, the data integration service first parses the request from the uniform data access interface, then sends to query planning and scheduling module which generates sub-query and executes parallel query according to different data sources. Finally, the data integration service provides the query results with corresponding format for the application layer by extracting, processing and combining sub-query results. Metadata has the merit of flexibly customizing queries, accurately positioning data objects, automatically mapping data sources and transparently providing data results in the response process to request.

The data integration service is used to interact with the heterogeneous data sources, because different types of data sources need different adapters. The main functions of adapter are: Firstly, parsing the sub-query from query planning and scheduling module, extracting parameters from the sub-query, and generating corresponding query which data sources can identify; Secondly, returning query results to the data synthesis; Thirdly, providing access modes for different data sources, such as web service, SQL access and object access. In the process of exacting data, all kinds of adapters provided by the data integration services will be used, and corresponding adapters for FTP file, relational database and bulk data are designed respectively, then corresponding data can be gotten by Socket, ORM(Object-Relational Mapping) and ADO.NET. The results can be returned with different data formats, such as Html, binary and c # object.

4.2. Processing flow for the data integration services based on metadata

Processing flow for the data integration services based on metadata is as follows:

Firstly, searching the corresponding topic metadata related with data request, and getting the business objects and relationships including in the topic metadata, and generating directed graph with business objects as nodes and relationships as edge. Secondly, according to directed graph, processing flow for metadata request is automatically generated. Thirdly, calling task queue and executing all sub-tasks in sequence, then getting response results.

In order to be able to automatically call all kinds of services, such as data search, format conversion and reading data, data metadata in three layers metadata model not only is used for describing data content and organizational form, but service metadata is used. For example, using predicate in the RDF to bind services is shown as follows.

```
<mw:Coordinator rdf:about="URL_ROOT/coordinator/GUID1">  
    <!—processing task queue -->  
    <cd:ProcessQueue resource="URL_ROOT/middleware/ GUID11" />  
    <!—generating SQL statements -->  
    <cd:ProcessSQLSelect resource="URL_ROOT/algorithm/ GUID21" />  
    <cd:ProcessSQLInsert resource="URL_ROOT/algorithm/ GUID22" />
```
4.3. Processing flow optimization

Processing flow optimization involves two aspects: generating an optimal flow and making processing quickly. By setting the weights of relationship among business objects, we are looking for a path with smaller weight in the directed graph of the business objects and relationships including in the topic metadata, and generate a optimal processing flow rules. Meanwhile, using high-speed index and cache can reduce repeated query time, and improve the data service efficiency.

5. The implementation of information resource access platform for the oil-gas fields

The information resource access platform for the oil-gas fields is built by .Net of Microsoft on the basis of three-layered metadata model, using REST for resource access strategy, and using SDK for uniform user interface.

Figure 3 shows framework schemes of information resource access platform for the oil-gas fields based on REST. Following the traditional C/S model, the platform is divided into two parts: C# client SDK based on REST and the web server based on REST.

C# client SDK based on REST is used to provide methods for users to access resource objects, generate corresponding REST style request according to selected methods by the user, and return resource objects according the user’s requirements. It is the interface between the information resource access platform and the users.

The client SDK consists of API methods, web request module and web response module. Users for oil and gas exploration and development can call all kinds of API interfaces to manipulate data. The called access operation will be automatically processed by web request module, and generated web request messages are sent to the web server. The web response module will process and eventually return to the user after receiving the response from the web server.
The web server based on REST is mainly used to parse HTTP message which the client sends, extract the corresponding parameter and operating information, and executive database, FTP and bulk data access according to the different request objects and operational information, finally return the required results to the client. It is the core of the information resource access platform.

The web server includes seven sub-modules: parsing and processing module, parsing parameter module, object management module, object filled module, bulk data service module, FTP service module and SQL generating module. The web server processing procedure is as follows: when web requests arriving at the web server, first, parsing and processing module will analyze the messages, then send the request to the corresponding processing module. For bulk data and FTP application service, the platform will forward the request to the corresponding application servers according to the URL. For database services, the platform will call the parsing parameter module to extract the web request parameters, then call the object management module to categorize request, and call the SQL generating module to generate SQL statements according to request parameters and operation, and connect to the application database server to execute, finally invoke the object filled module to fill the objects or object arrays and return to the client.

6. Conclusion

How to effectively use the precious data in legacy systems and realize the transparently access to different application systems, it has become an important issue which we must solve when we develop data integration system. Data integration will provide uniform representation, storage and management for various heterogeneous data, shield the difference between various heterogeneous data, provide users with a uniform interface to access heterogeneous data sources, so users need not consider data model heterogeneity, data extraction and data synthesis, and so on.

Metadata has the merit of flexibly customizing queries, accurately positioning data objects, automatically mapping data sources and transparently providing data results in heterogeneous data integration. In addition, the information resource access platform is of reusability and expansibility according to design thought, design method and the realization technologies. We can dynamically add or delete metadata when data sources and requirements change. The research is not only suitable for oil-gas field enterprise, but also provides ideas and methods for other enterprise to construct information integration platform.

7. Acknowledgements

Financial supports from the National Natural Science Foundation of China (No. 61072039), the Beijing Municipal Natural Science Foundation (No.4102040) and the National High Technology Research and Development Program of China (No. 2009AA062801) are highly appreciated.

8. References

Semantic Model based on Three-layered Metadata for Oil-gas Data Integration
Ma Zhonggui, Wang Chengyao, Wang Zongjie
Advances in Information Sciences and Service Sciences, Volume 3, Number 7, August 2011


