Design and Implementation of ETL Tool for Data Warehouse

1st Jingting Wang
College of Engineering and Technology
Xi'an Fanyi University
Xi'an, China
851969149@qq.com

2nd Bao Liu
College of Electronical and Control Engineering
Xi'an University of Science and Technology
Xi'an, China
xiaobei0077@163.com

***Abstract*—This paper takes the current business system of automobile department store chain sales service company as an example, and analyzes the problem that the data from multiple data sources cannot directly be loaded into the data warehouse by the current commercial ETL (Extraction-Transformation-Loading, ETL) tool. Firstly, it designs the logical model of ETL tool based on the metadata, then customizes the data cleaning method in ETL tool. Finally, the system experiments are carried out on ETL tool, and the experimental results and analysis are given. Through the development of the ETL tool, an interface is provided for the user-defined data cleaning function, which makes up for the shortage of the custom data cleaning function by the commercial ETL tool, thereby the use efficiency of ETL tool is improved.**

Keywords—Data Warehouse, Structured Data, Extraction-Transformation-Loading (ETL), Metadata

I. INTRODUCTION

Data warehouse is a subject-oriented, integrated, non-volatile, and time-varying data set that supports management decisions[1-2]. The successful implementation of a data warehouse serving business intelligence reflects the growing demand for new ideas and solutions[3-4]. At present, more and more enterprises or organizations integrate data from various departments to build an information integration platform in order to better help managers make decisions. In the process of information integration, data is derived from different business systems, and heterogeneous, vacant or redundant data is integrated, so the data quality is difficult to guarantee, resulting in reduced reliability of decision support[5]. Enterprise development decisions require credible business data. Therefore, the integration of operational data between different business departments and the generation of consistent data has become an urgent need.

This paper takes the current business system of automobile department store chain sales service company as an example. In the process of data integration, data from multiple data sources cannot directly be loaded into the data warehouse by the current commercial ETL (Extraction-Transformation-Loading, ETL) tool. Therefore, designing ETL tool has become a primary issue.

II. INTRODUCTION TO DATA WAREHOUSE AND ETL TOOL

A. Data Warehouse Architecture

The data warehouse is not only the physical implementation of the data model of the decision support system, but also the information needed for the strategic decision of the enterprise. At the same time, it is more regarded as an architecture that integrates business data from the heterogeneous data sources to support specialized query and analysis reports and decision making.

From the system structure diagram of the data warehouse in Figure 1, it is mainly divided into three parts: data sources, data extraction and storage, and data presentation[6]. The data comes from multiple data sources, including the company's internal data and some external data such as market research. The data extraction part is the entry of data into the data warehouse. The data warehouse is a stand-alone data environment, which needs to process the data from the online transaction processing system, the external data source, and the offline data storage medium through the extraction process, and load it into the data warehouse for the convenience of user analysis. Through the concentration and cleaning of scattered data, it is transformed into clean, consistent, and comprehensive decision-oriented data. Analysts and business managers use a variety of query retrieval tools, such as online analytical processing (OLAP) tool for multidimensional data, and data mining (DM) tool, etc., to implement various requirements of decision support systems.

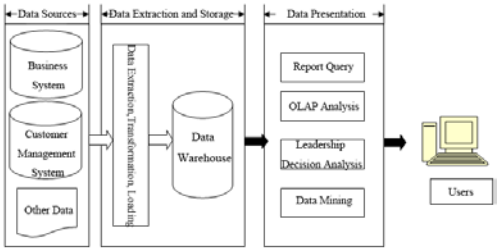

Figure 1. System structure diagram of the Data Warehouse

  DOI: 10.25236/iccpb.2018.020

B. ETL Tool

ETL is an abbreviation for data extraction, data transformation and data loading. The ETL process is to extract the required data from the data sources, through the preprocessing process such as data transformation and cleaning, and finally load the data into data warehouse according to the predetermined data model[7-8]. The ETL conceptual model is shown in Figure 2.

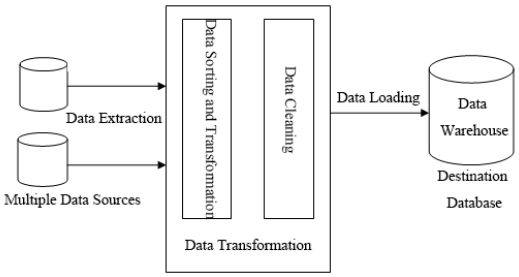

Figure 2. ETL conceptual model

In most cases, data in an operational system cannot be used directly to provide strategic information. The ETL tool transforms relevant data from the source system into useful information for analysis decisions. In order to transform the data in the source system into strategic needed information, firstly the data is captured, and then the extracted data is classified according to the transformation requirements, transformed into information, and loaded into the data warehouse.

The current popular professional ETL vendors include Asential's DataStageXE, Sagent's Solution, and Informatica's products. The overall solution providers and products include Oracle's Warehouse Builder and IBM's Warehouse Manager. After analysis, it is found that such products generally have good support for their own products and can exert maximum efficiency, but the structure is relatively closed, and the support for other manufacturers' products is also limited. In large information integration systems, it is common to deal with databases and data resources of different structures, which makes existing ETL tools unable to meet the requirements, so it is necessary to design ETL tool that meet enterprise customization.

III. Design and Implementation of ETL Tool Based on Metadata

Metadata[9-10] is "descriptive information about the database schema (especially relational database) ". The definition treats metadata as a way of describing information about database properties and attributes (including tables, column names, column attributes, primary keys, and foreign keys). Simply put, metadata is "data about data", which is a description of data resources. Metadata is the application soul of the data warehouse. It can be said that there is no data warehouse without metadata. Data source definitions, mapping rules, transformation rules, loading strategies, etc. in the ETL process are all in the metadata category. How to properly collect, store and manage this information is not only related to the smooth implementation of the ETL process, but also affects the later use and maintenance.

A. Metadata-based ETL logical model design

The metadata-based ETL logic model is divided into three main parts: the maintenance of subjects, the maintenance of external data sources, and the corresponding relationship between them. Its logical model is shown in Figure 3.

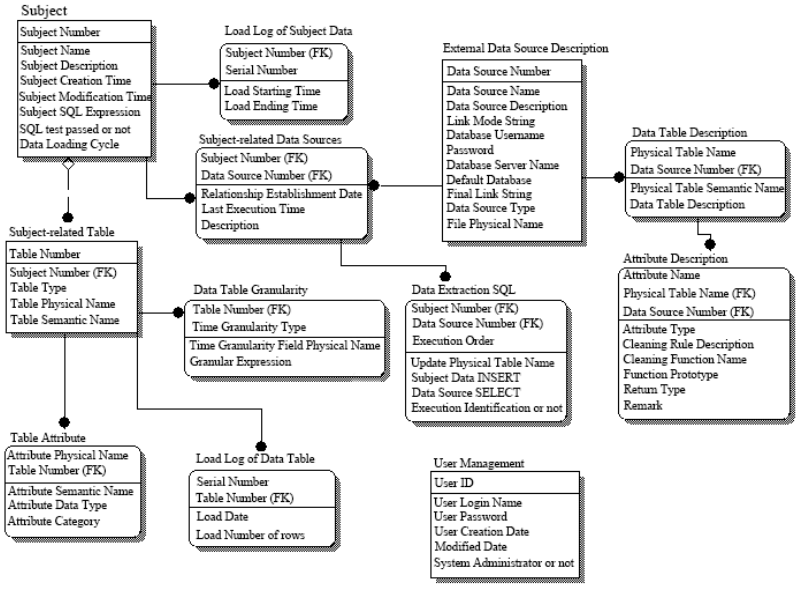

Figure 3. Metadata-based ETL logical model

B. Implementation of custom data cleaning methods in ETL tool

According to actual needs, the user can write a cleaning function to filter illegal data in the specified data item. Taking the catalogue of the automobile department store chain as an example, the following custom function cleanChar was written to realize the cleaning of illegal characters in the commodity name field.

```
CREATE or replace FUNCTION CleanChar(vfx IN varchar2) RETURN varchar2
    IS Result varchar2(200);
BEGIN
    Result:=vfx;
    Result:=replace(Result,'/','');
    Result:=replace(Result,'\','');
    Result:=replace(Result,'"','');
    Result:=replace(Result,'.','');
    Result:=replace(Result,',','');
    Result:=replace(Result,';','');
    Result:=replace(Result,'。','');
    Result:=replace(Result,'%','');
    Result:=replace(Result,'+','');
    Result:=replace(Result,'?','');
    Result:=replace(Result,'~','');
    Result:=replace(Result,'!','');
    Result:=replace(Result,chr(39),'');
    Result:=replace(Result,chr(9),'');
    RETURN (Result);
END;
```

IV. Experimental results and analysis of ETL tool

A. Operating environment

PC Hardware environment: CPU: Inter(R) Core(TM) i5-3470 CPU @ 3.20GHz; Memory: 4GB; Hard disk: 1TB.

PC Software environment: Windows 7 operating system; ORACLE 10g database.

B. Experiment

1) User login

The user enters the correct username and password to access the ETL tool for structured data. The user login interface is shown in Figure 4.

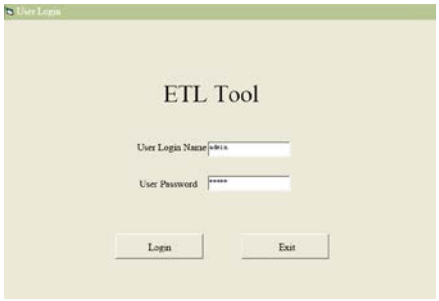

Figure 4. User login

2) Maintenance of Subjects

a) Create subjects. Every subject includes a subject number, a subject name, a subject creation date, a subject modification date, a data loading cycle, a subject SQL expression, and so on. The creation subject interface is shown in Figure 5.

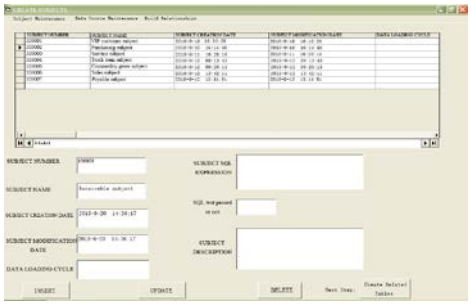

Figure 5. Create subjects

b) Identify the tables related to the subject. When the subject is created, determine the table related to the subject. Maintain the dimension and fact tables related to each subject, and determine the table type, table physical name, and table semantic name item. The interface for determining the subject-related table is shown in Figure 6.

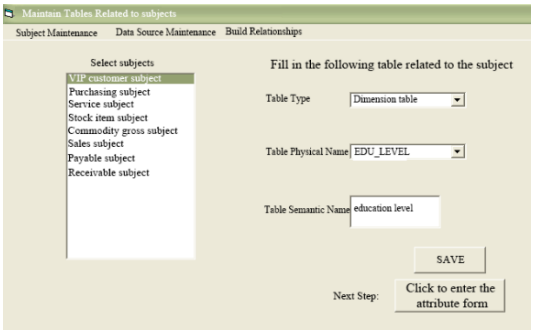

Figure 6. Determine the subject-related table

c) Maintain attributes of the table. Maintain the physical name, attribute data type, and data semantic name items of the attributes. The attributes interface of the maintenance for table is shown in Figure 7.

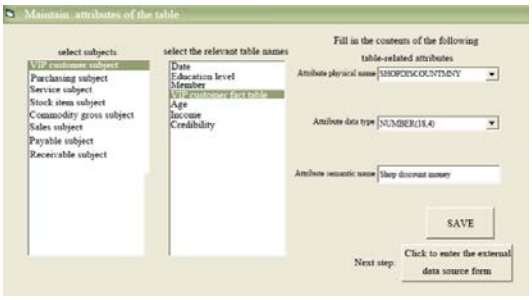

Figure 7. Maintain attributes of the table

3) Maintenance of external data sources

a) Determine the related tables in the external data sources. The semantic name of physical table and table description for the related table in the external data sources are maintained. The maintenance interface of the table in the external data sources is shown in Figure 8.

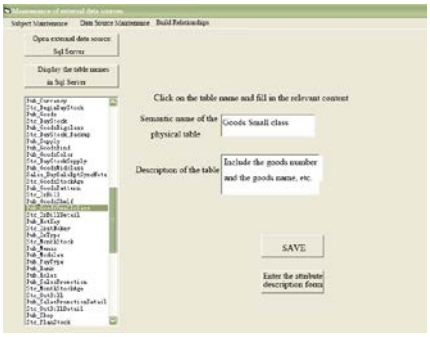

Figure 8. Table maintenance in the external data sources

b) Maintain attributes of the table in the external data sources. Attribute type, attribute name, the description of cleaning rule, and the name of cleaning function and so on are maintained. Figure 9 is an interface of attributes maintenance for the table in the external data sources.

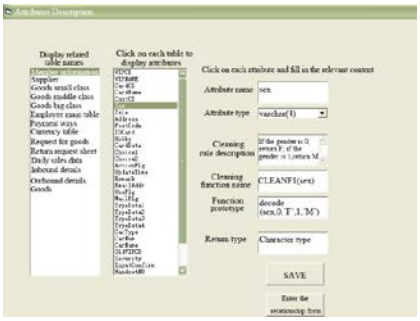

Figure 9. Attribute maintenance in the external data sources

4) Build a corresponding relationship

The relationship between the attributes of the table in the data warehouse and the attributes of the corresponding table in the external data sources is established, and the values are imported into the corresponding table of the data warehouse. Figure 10 is the interface of building a correspondence relationship.

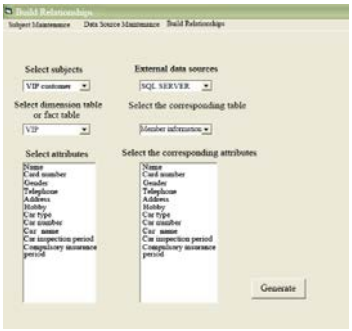

Figure 10. Build a corresponding relationship

C. Results and Analysis

This experiment is carried out in ODS environment. The experiment is based on the analysis of some data tables in SQL SERVER 2000. It can be seen from the experimental results that ETL tool for the structured data can correctly load the operational data into the data warehouse, and the data cleaning results are correct. See Table I for details.

TABLE I. ANALYSIS OF EXPERIMENTAL RESULTS

Data storage platform for data source	Table Name	Original record number	Data storage platform for data warehouse	Table Name	Current record number	Data cleaning rate
SQL SERVER	Pub_Vip	2216	ORACLE	VIP	2000	90.3%
	Pub_Supply	860		Supplier	860	100%
	Pub_GoodsSmallclass	7840		GoodsSmallclass	7700	98.2%
	Pub_GoodsMidclass	1830		GoodsMidclass	1650	90.2%
	Pub_GoodsBigclass	1200		GoodsBigclass	1000	83.3%
	Pub_Users	5500		ServiceEmployee	5000	90.9%
	Pub_Currency	50		Currency	50	100%
	Pub_PayType	20		PaymentType	20	100%

V. SUMMARY

Based on the current business system of the automobile department store chain sales service company, this paper designs and implements a structured ETL tool for the company. Through the development of ETL tool, an interface is provided for the user-defined data cleaning function, which makes up for the shortage of the custom data cleaning function by the commercial ETL tool. In addition, the SQL text of the data extraction is retained in the metadata of supporting the running of the ETL tool, which reduces the time cost caused by recompilation when the same SQL is executed again, thereby the use efficiency of the ETL tool is improved.

ACKNOWLEDGMENT

Foundation: School-level Scientific Research Team (XFU17KYTDB02),Cooperation in Production and Education Program of Ministry of Education (201801193088), Scientific Research Program Funded by Shaanxi Provincial Education Department (Program No.17JK0504 and 18JK1005), Research supported in part by grant for the National Natural Science Foundation of China (61703329), China Postdoctoral Science Foundation (2018M633538),Natural Science Basic Research Priorities Program of Shaanxi Province of China (2018JQ5197), Xi'an University of Science and Technology Scientific Research Foundation for Ph.D. Receiver (2016QDJ039), and Fund for Research Fostering of Xi'an University of Science and Technology (201738).

REFERENCES

[1] M.W.H.Inmon, Data warehouse. Machinery industry press, 2006.

[2] J.Dai Qiaoling,Li Zhen, Research on Construction of Learning Behavior Data Warehouse, 10th ed., vol.17. Software Guide, 2018, pp.187-190.

[3] J.Li Na, The Design and implementation of business intelligence management system based on data warehouse, 15th ed., vol.39. Modern electronic technology, 2016, pp.140-144.

[4] J.Wang Xinbei, Xie Wenge, Wang Zhongquan, Application of data warehouse solution in e-commerce, 4th ed. Digital technology and application, 2015, pp. 101-101.

[5] J.Li Yun, Research on big data strategy based on enhanced ETL process, 34th ed. Computer knowledge and technology, 2014, pp.8081-8082.

[6] J.Zheng Haichuan, Zhang Hao, Design and Implementation of Industrial Enterprise Energy Management System Based on Data Warehouse, 8th ed., vol.39. Process Automation Instrumentation,2018, pp.43-46.

[7] J.Chen Jianyao, Research on Scheduling Optimization Based on Hive Data Warehouse, 8th ed., vol.34. Bulletin of Science and Technology, 2018, pp.113-117.

[8] J.Ding Xiangwu, Xie Shuliang, LI Jiyun, Parallel ETL based on Spark, 9th ed., vol. 38. Computer Engineering and Design,2017, pp.2580-2585.

[9] J.Zhao Xiaofei, Research on Metadata Integration and Management of ETL Tool, 16th ed., vol. 32. Journal of Wuhan University of Technology, 2010, pp.115-118.

[10] J. Su Fang, Shou Yongxi, Su Yila, Research on metadata driven ETL, 6th ed., vol. 48.Computer Engineering and Applications,2012, pp.114-118.