J. Basic. Appl. Sci. Res., 2(1)54-59, 2012

ISSN 2090-4304
Journal of Basic and Applied Scientific Research
www.textroad.com

© 2012, TextRoad Publication

A Service-Based Framework for ETL Process Based on Metadata

HosseinTalebzadeh

Department of Computer, Science and Research Branch, Islamic Azad University, Khouzestan,Iran

ABSTRACT

This article studies the problems of available ETL tools firstly and introduces the service of ETL pattern based on Metadata, and then summarizes different types of Metadata and their application. And shows a comparison between current frameworks of ETL process and service- based framework, based on component distribution.

According to this ETL service pattern, fundamental framework of ETL service is placed higher in the rank and many important services are studied such as Metadata management services, Metadata meaning services, form change of ETL rules, meaning process services, SQL code generation and optimization services, process control services and so on. Finally, meaning-based approach and ETL rules dependent algorithms are studied.

It is proved in practice that this model and the proposed frameworks in the article can increase the effectiveness of ETL largely.

KEY WORDS: ETL, Data Warehouse, Metadata, Service

INTRODUCTION AND REVIEW OF LITERATURE

DW(Data Warehouse) is a big organization database full of OLTP which its data will be selected ,extracted, refined, matched and stored Among OLTP precisely. The main mission of DW is to publish organization data effectively in order to support strategic decision making of organization managers.(6,5)

A wide range of data sources are located in organizations with various implementations. This spends more than 70 percent of DW projects potential in ETL section. The name of this section shows three main tasks of it. This section is responsible for keeping all copies of receptive data from different sources. Naturally, the data entering to this section are first non-integrated and sometimes imprecise. So, their refinement, filtering and equalization is inevitable.

It should be noticed that the application in ETL section is merely a series of executive code which were made by programmers in older projects. Many companies also invest on this section and gradually release some tools to the market which were able to do some applications such as design, modeling and implementation of ETL section. But data changing in this type of tools is not much easy and it takes more time to become an expert in applying their rules and language. ETL designers need to become familiar with data structure, ETL rules and application processes available in the system. So, it seems hard to improve ETL development process effectively.

Moreover, designers should redesign ETL process when business rules or source or target data is subject to change.

In this article, we first analyze currents problems in ETL process, and then a service-based model for ETL according to metadata will be presented. This needs summarizing and classification of different types of required Metadata which shown themselves as several services in this framework such as Metadata management service, Metadata definition service, transform rules in ETL service, service definition process, auto code generation service, optimization service, control service and so on. Finally, research methodology and algorithms related to the ETL process rules in this model need to be defined and analyzed.

METHODOLOGY

i. Designing Model Service for ETL Process

Metadata is a type of data used for describing its involved information. Metadata is in place of various types of information in different environments. In this article not only Metadata

***Corresponding Author:** HosseinTalebzadeh, Department of Computer, Science and Research Branch, Islamic Azad University, Khouzestan,Iran. Email: i@ihossein.com

includes data warehouse, but also it contains Metadata from data sources, changing rules, extraction rules and work flow rules. In ETL model service, Metadata have different approaches which are as follows:

1. Data Warehouse Metadata

In ETL process, this metadata is used mainly for describing data related to the data storage process, especially for data description of stored data. In data warehouse, metadata description format is constant, because contact schema is the baseline for data storage and all the process is based on the communication. It contains data storage pattern, size, variation in the size of data packages(Granularity), communication table, integration restrictions , viewpoints , stored instructions , integration rules.

2. Data Source Metadata

Basically Data source Metadata describes the data in database and should include data format, IP addresses, access ports, database schema, relation table structure, attribute sets, indexes, views, stored procedure, referential integrity constrains and so on. Regarding the average, maximum and minimum value of a field, the total number of records and other statistic information will be defined.

3. Extraction Rules Metadata

Extraction rules metadata mainly contain mapping connection between database and data warehouse and describe source field data, target field data and transform rule data. This can include a large number in comparison to a connection between main source and target source that are able to apply different extraction rules in different situations.

4. Transformation Rules Metadata

The transformation rules metadata involved in these research, become self-defined operations of warehouses by running stored instructions and triggers; and retrieval method of the content of such transform rules , like stored instructions of name, the type of operation description parameter and different types of return rate, are protected by Metadata.

5. Process Rules Metadata

Process rules metadata has been used for describing information details of data extraction process such as process execution succession, expectations and dealing with size. In process execution succession, it is allowed to include other ETL category processes and description method like common nodes that make it possible to reuse whole ETL process.

In this ETL model service, Metadata plays an important role and is the core and the baseline in designing and controlling whole ETL process. All information is managed by Metadata and transformative rules which use those information, so that they will shape executive succession, then they will produce target information for data warehouse. Any Change in ETL process is like addition of an ETL process, an order (instruction) and so on that is answered by transformation in Metadata. Characteristically, ETL processes' succession is divided into two parts of constant and variable, while Metadata manages the variable part. Picture 1 shows the proposal ETL model service. In ETL model service, each ETL service is counted as a logical target which is used for completion of specific tasks of ETL. Sometimes these kinds of targets are placed in a directed graph of G= (V, E) which is summarized with series of logical nodes (V) and logical edges (E). Each logical node can be divided into two categories of independent and compound. Independent logical nodes are related to an ETL task which has high connection and is undividable. Compound logical nodes are a combination of ETL processes with different categories which their recall method is the same with the independent logical node. This is the theoretical basis for sharing ETL process which prevents defects of traditional coding methods. As long as logical edge is connected with logical node, it is used only for describing the logical relation among nodes. If logical target of A is so similar to logical target of B, then B needs little or no transformation. Actually, Metadata can be regarded as an abstract of ETL processes or a range of different data which is always stored by static methods. For example, XML documents used for control process which are run by interpretation method can make a directed graph and finally run the ETL logical target. Picture number 1 shows this model service.

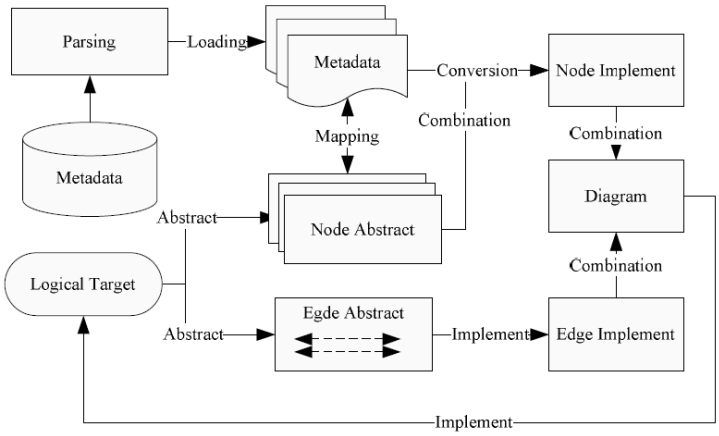

Figure 1- ETL Services Model Based on Metadata

ii. ETL Process Service Framework

Based on model analyze, this research regards Metadata as a core and all operations are ordered as service. So, all ETL services are published in a uniformed format which could be accessible in the same format. The proposed ETL frameworks are shown in picture number 2. The core framework services include Metadata management service, Metadata content service, Metadata transform rules, process definition service, SQL code generation and optimization service, process control service, access service, transform management service and exception control service.

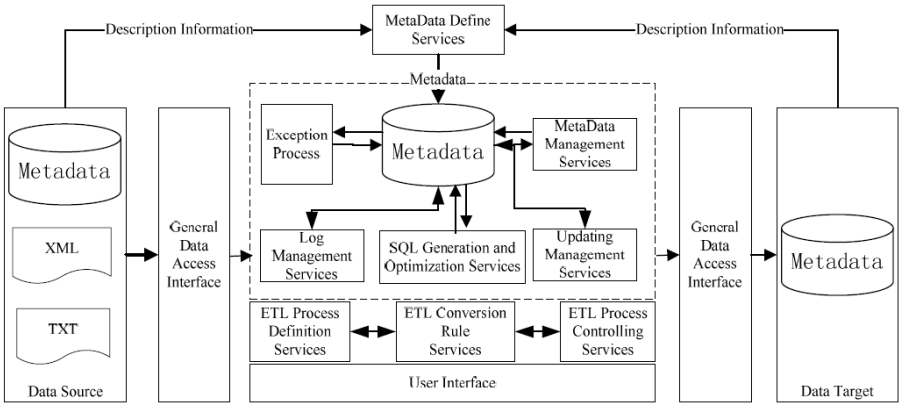

Figure 2-ETL Framework

1. Metadata Management Service

Main operations of Metadata management service are applied for XML documents analysis and management of mass Metadata components. Metadata is stored in XML file. Some of Metadata are universal such as those used for different relational schema. Some of Metadata are related to rules, however. For effective analysis of XML files, Metadata components are loaded according to different Metadata types at the time of initial formatting.

2. Metadata Definition Service

This service provides operation of data content for involved processes and ETL processes data and those XML documents that conclude all contents.

3. ETL Transform Rules Service

It provides data transform service for system based on transform rules for Metadata sources, content service process and control process service.

4. Process Definition Service

It divides each process into series of different ETL categories or ETL services, in fact real ETL processes building is defined as the form of a series of SQL modes. This serial making from

SQL modes develops easily and matches itself by habits. Each ETL process is required to make two processes: ETL final process and ETL growth and development process.

General ETL process will be run only when the information is loaded for the first time, actually when ETL development process is applied in an intermittent update environment. Whereas this service should search each related Metadata and other information at the time of running, object serialization and serialization technology is used. And that is a binary file which is made in its area hard disk after XML file for describing ETL, and is rendered and serialized for the first time. In this method ETL process break performance will be improved largely.

5. SQL Code generation Optimizing Service

This service aggregates with high flexibility by SQL generators so that they make the same interval and it becomes independent among them. These generators are for event table control, size table, size mapping table and access table at the time of first loading and its development update. After generation of SQL mode, it still needs to optimize SQL mode based on the data which will be stored in Metadata warehouse. Metadata warehouse can provide Metadata related to the intervention services including Metadata service access method and parameter data. Information of Metadata database. SQL mode optimization not only can be changed manually, but it could be used by the same format at proportional rank. So, ETL motor has high dynamic developmental ability and follows depth applications. The optimization mechanism could be better in this way.

6. Process Control Service

These services are mostly responsible for process control, and guarantee executive succession according to its content. At the same time, this service can also include improving errors and exceptions that control mechanism. Whenever an error occurs in a node, it will be stored in access table which will restore the tasks done by current node, so the error will be cleared.

7. Data base Access Service

By using a data access mediator, the access process to data bases is directly connected to data bases, and runs the instruction according to Metadata stored in the data base and finally give the result to its client.

8. Log Service

Log service is an effective method to guarantee ETL services quality and can provide the operation of supervision processes and error improvement. The main log service includes functional logs for recording functional data , access errors for recording error nodes , access analysis for recording statistical information, executive succession , and access change for recording variable information of service and so on.

9. Transform Management Service

This service is a confident tool for updating local and distant services. When a service gets updated, all the related services could be set.

10. ETL Rules Service

When an ETL process is run, ETL rules play a crucial role in control and super vision process of ETL. As in the picture number 3 this is a data flowchart in rule definition module and mainly defines the source and target information when it describes ETL rule in client section.

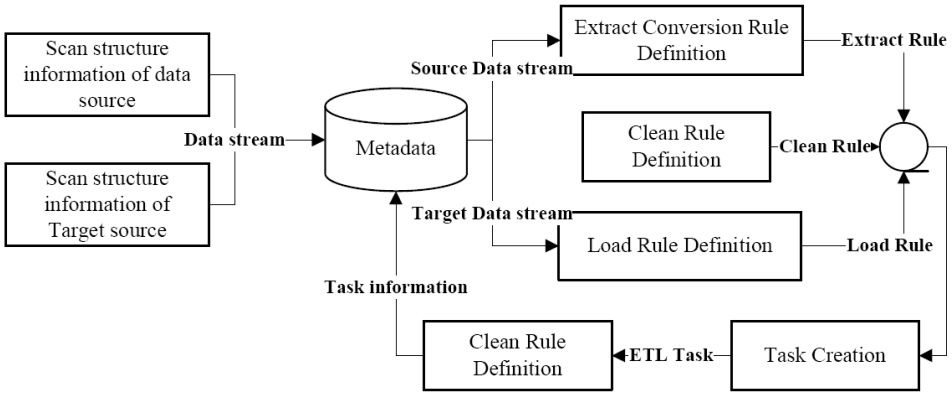

Figure 3-Module of Rule Definition

Before defining ETL rule, the client needs to scan the source and target data base structure which is stored in Metadata database, then rule definition service will acquire Metadata from database. That Metadata can be shown as graphical component, so the client can use it for ETL rule definition and store it in a Metadata database. Detailed algorithm of rule definition is shown in picture number 4.

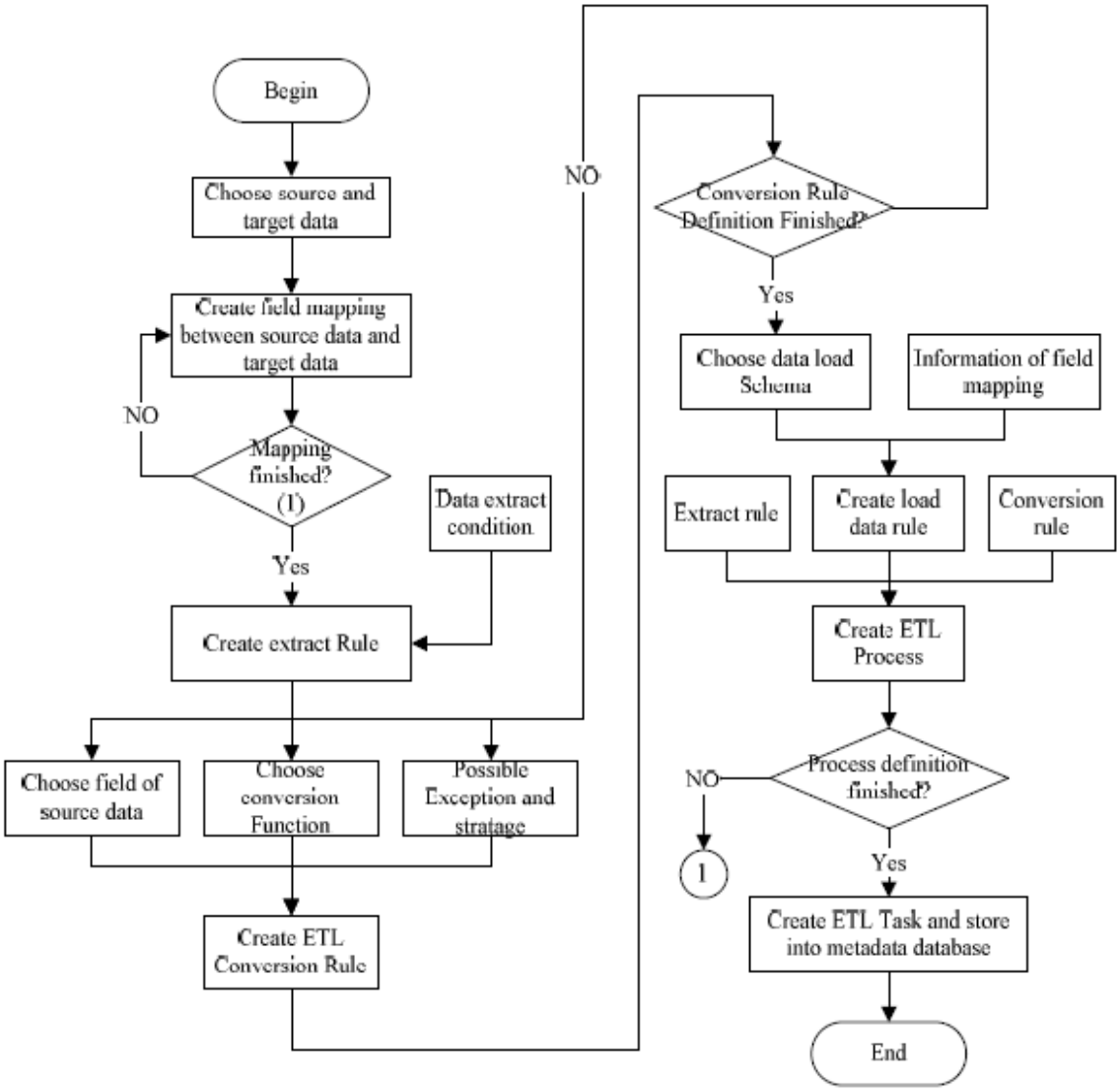

Figure 4-Algorithm of Rule Definition Model

RESULT AND CONCLUSION

According to ETL service framework proposed in this article, an ETL system primary sample is developed based on the service in picture number 5and its executive operation is extracted. In this sample ORACLE 11GR2 is used as a warehouse of Metadata database, and IIS 7.5 is a web server and all ETL operations are published as an integrated service format. So the clients can access those services in a standard mode. The proposed ETL service pattern in this article has a dynamic potential in development and can reach the target of flexible control of ETL process , it also has a good optimization mechanism .at the same time this framework includes all advantages of the research results on communication model and can design ETL process effectively.

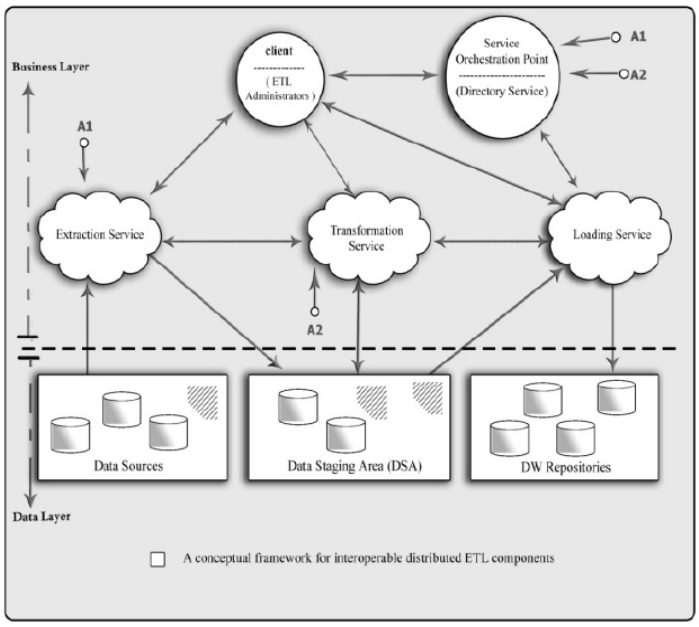

Figure 5- TheImplementation Model

REFERENCES

1. C.J. Date. An Introduction to Database Systems. Addison-Wesley, eighth edition, 2003.
2. Wu Jun, Han Bin, Liu Gang, "XML technology in project developingbased on powermart", Science Technology and Engineering, Vol.l0,no.!, 2010, pp.269-272
3. Wang Yong-zhi, "Design and Implementation of National Oil-GasResource Database Management System Based on ArcGIS and SOA" Journal of Jinlin University(EartScience Edition), Vo1.39, no.5,2009, pp953-958
4. H. Nemati, D. Steiger, L. Iyer, and R. Herschel. Knowledge warehouse: An architectural integration of knowledge management, decision support, artificial intelligence and data warehousing. Decision Support Systems, 33:143–161,
5. J. Shim, M. Warkentin, J. Courtney, D. Power, R. Sharda, and C. Carlsson. Past, present and future of decision support technology. Decision Support Systems, 32(2):111–126, 2002.
6. W. Inmon. Building the Data Warehouse. Wiley, 2002.
7. R. Kimball and M. Ross. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. Wiley, second edition, 2002.
8. "Blueprints and measures for ETL workflows," Lecture notes incomputer science, vol. 3716, p. 385, 2005.
9. E. Newcomer and G. Lomow, Understanding SOA with WebServices (Independent Technology Guides): Addison-WesleyProfessional, 2004.
10. T. Hau, N. Ebert, A. Hochstein, and W. Brenner, "Where to Startwith SOA Criteria for Selecting SOA Projects," Proceedings of the41st Hawaii International Conference on System Sciences,IEEE,2008.
11. Zhang Zhong-ping, "Design of architecture for ETL based onmetadata-driven",Computer Applications and Software ,vo1.26, no.6,2009, pp.61-63.
12. He Cheng-gang "Design and technique research on ETL system",Computer Applications and Software, vo1.26, no.4, 2009, pp.!99-201
13. Yang Ling, Zhang Rong, Zhang Jun, "A novel design of general dataaccess component based on data warehouse", Journal of navaluniversity of engineering, Vo1.2 1 , no.!, 2009, pp59-62
14. P. Vassiliadis, A. Simitsis, M. Terrovitis, and S. Skiadopoulos,"Blueprints and measures for ETL workflows," Lecture notes incomputer science, vol. 3716, p. 385, 2005.