Research of ETL on University Data Exchange Platform

Qian Wang, NaiJia Liu, Chun Yu
Computer and Information management Center
Tsinghua University
Beijing, 100084, China
wq@cic.tsinghua.edu.cn, lnj@cic.tsinghua.edu.cn, yuchun@cic.tsinghua.edu.cn

Abstract—To solve data exchange problems among heterogeneous databases in university data sharing platform, in this paper, the author focused on further research on Extract Transform Load (ETL) technology, designed and executed university data sharing platform based on CloverETL. This module, the core of the whole university data exchange platform, can well achieve data exchange and sharing among heterogeneous databases and run stably with less influence on server and business system, and thus getting a better result.

Keywords: Data Exchange; Data Sharing; ETL; CleverETL

I. INTRODUCTION

With the development and popularity of college information technology and digital campus, information systems, such as educational administration system, admission system, personnel system, one-card campus, scientific research system and etc. are used daily in every college department. However, the relative independence of most of such applications and their usage of different platforms, databases and coding hinder the process of data exchange and information sharing across departments and systems, and thus creating the “information island phenomenon”. It becomes the key in the further development of college information management system how to ensure uniqueness and real-time operation of data.

It is not practical to completely override the original business system to achieve college information integration. To solve the above problems, it is critical to create a data exchange platform which enables processing of bottom data without altering the original system applications

II. KEY TECHNOLOGY

A. Extract Transform Load (ETL)

Extract Transform Load (ETL) is a common terminology used in data warehousing which stands for extracting data from source systems, transforming the data according to the business rules and loading to the target data warehouse [1-3]. ETL can be divided into three modules: data extraction, data transformation and data load. ETL process is in Fig.1:

Data extraction is the process of reading data from a specified source database and extracting a desired subset of data. Transformation is the process of converting the extracted/ acquired data from its previous form into the form it needs to be in so that it can be placed into another database. Transformation occurs by using rules or lookup tables or by combining with other data. Load is the process of writing the data into the target database.

The main current domestic and foreign ETL business tools include Oracle warehouse builder (OWB) of Oracle Corporation, Data Transformation Service (DTS) of Microsoft Corporation, Informatica of Informatica...
Corporation, Enterprise ETL Server of SAS Corporation, Data Stage of IBM Corporation, DataMigrator of iWay Software Corporation, Transformation Server of DataMirror Corporation, and etc. In addition, the current frequently-used Open Source ETL tools include Kettle, Talend, CloverETL, Octopus and etc.

With the help of ETL technology, this paper focused on solving data exchange problems among heterogeneous databases. That is to say, for the heterogeneous database sources with large quantities of data but low requirement for real-time, data exchange is implemented through ETL.

B. CloverETL

CloverETL [4] is a Java based open source framework which can be used to transform structured data, and to some degree free-form data. Its complete platform independence is achieved by the exclusive use of Java technology. It can be used standalone as an application launched from a command line or can be embedded into our own application. CloverETL is based on concept of transformation graph. The transformation graph is in Fig.2:

![CloverETL transformation graph](image)

Graph is divided into units – phases. Graph must have at least one phase. Phase combines several Nodes and Edges into unit which is executed at once, several Phases are executed one by one. Nodes transform (change) data, each Node runs as separate thread knowing about outside world only through Input and Output ports. Edges merely connect two Nodes, every edge connects exactly one Output port of Node A with one Input port of Node B. CloverETL works with data in terms of data records and data fields within records. Clover metadata describes semantics of data record, this is how to create data record from elementary data types.

The reasons why this article executed data exchange through CloverETL mainly lie as follows: Firstly, CloverETL is a Java based open source framework and can be used as a library of Java classes (embedded in our application). Secondly, the bottom interface of CloverETL can fulfil large quantities of data exchange and difference comparison among heterogeneous databases, and thus working stably and efficiently.

III. DATA EXCHANGE PLATFORM OF TSINGHUA UNIVERSITY

A. Design

There are more than one hundred professional information systems in Tsinghua University. After resource integration in recent years, we have centralized management of core data. However, because of the difference of network environment and data storage environment, one single scheme of data exchange and sharing can not meet all needs. Particularly for engineering application, it is not enough only to accomplish exchange and sharing. Furthermore, it is necessary to consider the influence on database, server performance and business system which puts forward higher requirement for data exchange of digital campus. This paper comes up with a solution to synthesize all data exchange demand based on above problems, which is successfully used in data exchange platform of Data Centre, Tsinghua University. The core part is ETL Platform.
Shown in Figure 3 is the flowchart design.

Overall solutions process is divided as follows: it started from analyzing the requirement of data exchange and judged whether the source and the destination are the same type database. If they are the same type database, it would enable privilege management. If it is heterogeneous database, it will go on judging the real-time requirement of data exchange. If the real-time requirement is not high, it can be achieved through ETL platform. If the requirement is high, small quantity of data exchange can be achieved through WebService based CBESB. Otherwise, data exchange task with complicated business and large quantities of data can be implemented through slightly reforming business application, adding corresponding function code and triggering ETL platform by hand. The core module is ETL platform.

B. ETL platform based on CloverETL

The most used tool is ETL in data exchange platform. The exploitation and application of ETL focus on solving data exchange of different databases which include different database types, different database versions and data resource of different storage types. CloverETL is a ETL tool of open source which provides a group of API and defines ETL process with XML. Based on CloverETL, the data centre of Tsinghua University develops task design module, task scheduling module and Log output module of ETL platform independently. This ETL platform can not only support incremental contrast of source data and target data, but also perform synchronous only to changed data, which reduces the performance impact on database server brought by ETL.

The ETL platform based on CloverETL is designed as follows: If the source and destination data can be compared, it will execute incremental contrast. Otherwise, for example, since the ORDER BY between Oracle and Mysql is not accordant, the source and destination data can not be compared under the circumstances. And it will delete destination data and synchronize all data. In the meantime, it tries to package business logic into database, and thus making it easy to manage ETL project.

Shown in Figure 4 is a typical ETL process based on CloverETL. Resource data and target data is read and comparison of increments is performed in this process. CloverETL will put additional data into target database when there’s new data, update corresponding data in target database when there’s change in existing data and delete corresponding data in target database when there’s reduction in the resource database.
The above process corresponding Metadata is as follows:

```xml
<Metadata id="Metadata0">
  <Record name="META0" type="delimited">
    <Field name="id" type="long" />
    <Field name="name" type="string" nullable="yes" />
    <Field name="code" type="string" nullable="yes" />
  </Record>
</Metadata>
```

The above process corresponding Phase is as follows:

```xml
<Phase number="0">
  <Node id="TABLE0" sqlQuery="select ..." type="IN"/>
  <Node id="TABLE1" sqlQuery="select ..." type="IN"/>
  <Node id="INCREMENT0" type="DATA_INCREMENT"/>
  <Node id="WRITER0" type="OUT" sql="insert..."/>
  <Node id="WRITER1" type="OUT" sql="update..."/>
  <Node id="WRITER2" type="OUT" sql="delete..."/>
  <Edge id="Edge1" fromNode="TABLE0:0" metadata="Metadata0" toNode="INCREMENT0:0"/>
  <Edge id="Edge2" fromNode="TABLE1:0" metadata="Metadata0" toNode="INCREMENT0:1"/>
  <Edge id="Edge3" fromNode="INCREMENT0:0" metadata="Metadata0" toNode="WRITER0:0"/>
  <Edge id="Edge4" fromNode="INCREMENT0:1" metadata="Metadata1" toNode="WRITER1:0"/>
  <Edge id="Edge5" fromNode="INCREMENT0:2" metadata="Metadata2" toNode="WRITER2:0"/>
</Phase>
```

IV. CONCLUSION AND FUTURE WORK

The data exchange platform of Tsinghua University makes use of heterogeneous database timing transmission based on ETL, heterogeneous database real-time transmission based on ESB and database internal data sharing based on authorized management. The core module is ETL platform based on ColverETL which can achieve data exchange among incremental contrast heterogeneous database. Through ETL platform, the data exchange platform of Tsinghua University can satisfy most of the data exchange demands and try to accomplish data exchange from the bottom with little impact on production environment performance. It avoids the modification of business system and achieves ideal effect. Based on the independently developed exchange systems of all kinds of open source softwares and middleware platforms, our data exchange platform demonstrates great flexibility and convenience in system update and transformation. And we expect the developed platform to have great impact to data exchange system development of other college information systems.

Though the current developed data exchange platform has satisfied general daily demand of digital campus, the introduction of online decision support system and cloud computing centre will further put forward the capability of the data exchange platform and will be the research direction of next generation data exchange technology in the data center.

REFERENCES