Research and Implementation of Heterogeneous Data Integration Based on XML

Hong-jie TANG [1, a]

[1]Department of Vocational Education, Liaoning Police Academy, Dalian, China

[a]thongjie@163.com

Keywords: XML; data integration; data parsing; mapping

Abstract. With the purpose of achieving data integration in heterogeneous environment, this paper proposes a new data sharing scheme based on XML and B/S three-layer architecture. On the basis of this, a new general data integration system is designed and implemented. This paper discusses the framework, workflow of the system, and especially focuses on the main functions of XML data parsing, XML document mapping these two modules.

1. Introduction

With the rapid development of information technology, any independent unit that holds or uses information is likely to become a heterogeneous data source. At the same time, enterprises also want to access all kinds of heterogeneous data, so that they can strengthen the relation of strategic partners, integrate and utilize resources better, make rapid response to market, and improve their own competitiveness [1]. Therefore, it needs a system to support data access of different sources. To solve these problems, this paper brings up a scheme by building a data change center based on XML and B/S three-layer architecture, and implements prototype system by Java technology.

2. Related Technology

2.1 XML Technology

XML is the abbreviation of eXtensible Markup Language, which is a language defined by W3C (World Wide Web Consortium). It is a simplified subset of SGML (Standard Generalized Markup Language). XML document is a kind of text file composed of tag and content. Unlike HTML (HyperText Markup Language), the tags of XML can be defined by users themselves, and the main purpose is to express data structure and data meaning. The purpose that W3C proposes XML is to make data exchange more convenient and make file content clearer on the Internet [2].

As the industry standard of structure data, XML provides many advantages for organizations, software developers, websites, and the end users, such as scalability, self-description, data storage for longer time, which make it very suitable for the application of e-commerce and information exchange.

2.2 Web Services

In order to build a channel between data sender and data receiver, there needs a kind of protocol or component which is cross-platform and cross-language. The protocol or component can describe and package data and content by using XML, so that the data and content from various platforms can be transferred through the same standard. HTTP (HyperText Transfer Protocol) and SOAP (Simple Object Access Protocol) are that kind of protocol (as shown in Figure 1) [3].

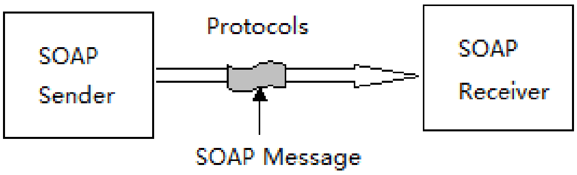

Fig. 1 SOAP Message Processing

3. Enterprise Information Integration Framework Based on XML

3.1 The Solution Scheme

The information integration framework based on XML adopts "data exchange center" structure to solve the coordination problems among different enterprise application systems (as shown in Figure 2). By using a unified data exchange standard, different application systems are connected with data exchange center, and then the system may implement data sharing and routing. Because the data storage layer is isolated from the application layer, top applications are not affected by data structure and storage model.

So enterprises do not need to change original business system or develop current business process again. This kind of connection mode implements seamless integration and sharing access of the data. It not only ensures the effective coordination of various business systems, but also ensures the independence of each application system. Loose coupling enhances the overall performance and safety of the data operation.

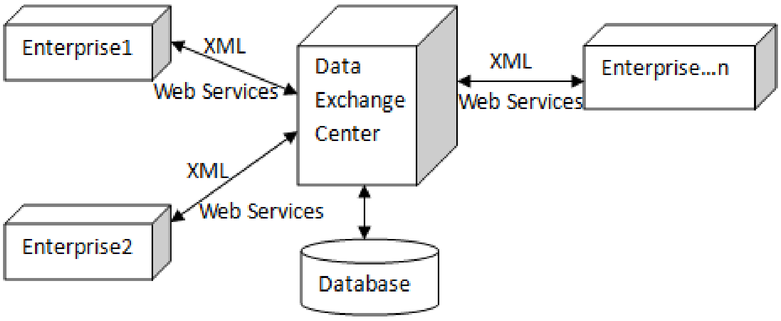

Fig. 2 Topology Structure

Based on above topology structure, heterogeneous data exchange requirements, and existing technology, the system uses data exchange mode based on Web service to implement heterogeneous data integration. This mode is independent from all participants, and the specific data process is determined by enterprise's business logic, therefore real-time and integrity of the data can be realized [4]. More importantly, data is a consequence of loose coupling in this mode, which means the change of one side's business logic or data format wouldn't impact the other side because data sharing method shields external change.

3.2 Framework Structure and Function Modules

Framework structure of data exchange center node and terminal enterprise node is given (as shown in Figure 3), and the function of each module is described here.

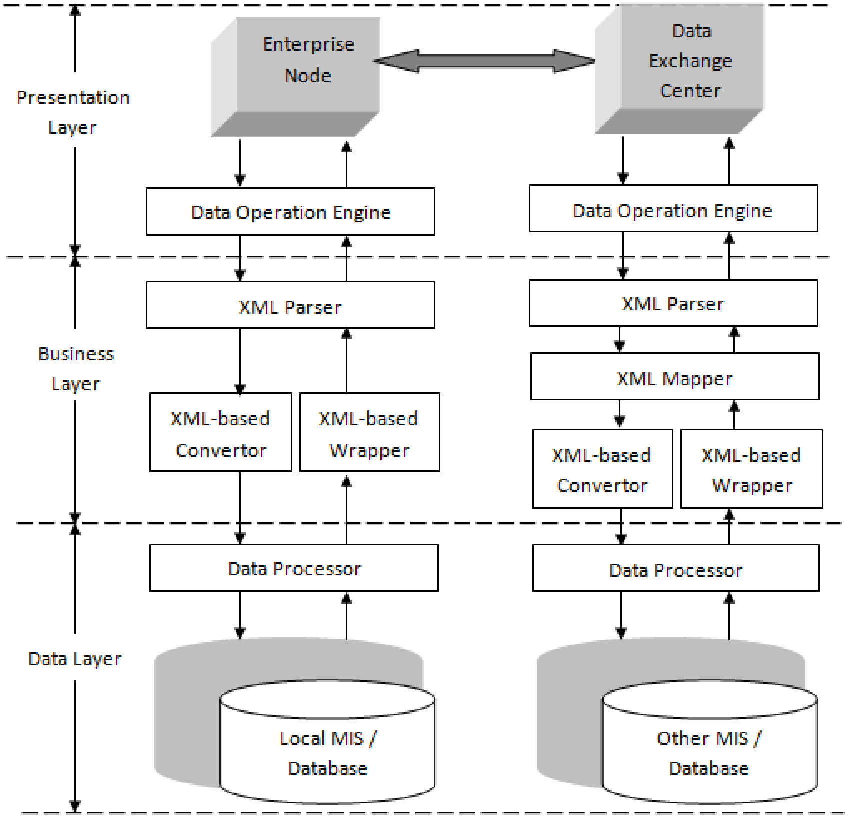

Fig. 3 Data Integration Framework

3.2.1 three-layer Architecture

(1)Presentation Layer: It realizes data expression, interaction, sending and receiving.

(2)Business Layer: It realizes data parsing, data conversion and data processing, which is also known as business logic layer or middle layer. This layer is the core of the whole system structure.

(3)Data Layer: It realizes access control of the data in relational database or in other storage mode.

3.2.2 Function Modules

(1) Data Operation Engine: It is a platform that is responsible for sending and receiving XML data. The sending part can send XML data to specified destination address through a variety of transmission protocols. The receiving part is actually some monitoring components, which can monitor and receive the data from the exchange node.

(2) XML Parser: The main function of this part is schema validation. XML parser tests whether XML document complies with the constraints on it, and analyzes the syntax of XML document to ensure there's no loss or error during the data transmission. It also analyzes whether the schema of enterprise is consistent with the schema of exchange center, and if there's inconsistent the system needs to make data mapping process.

(3) XML Mapper: When the schema of enterprise is inconsistent with the schema of exchange center, XML mapper module would be used [5]. Firstly, it checks whether there is a corresponding XSLT (extensible style sheet language) file. If there it is, XML mapper converts XML document from enterprise to standard XML file directly according to the XSLT rules. If there it isn't, XML mapper extracts data format, data type and field name from enterprise schema, and then map with standard schema according to business rules. Finally, the system produces a new XSLT file as conversion standard.

(4) XML-based Converter: It mainly completes the conversion between XML data model and other data models (relational model, HTML documents, and text documents).

(5) XML-based Wrapper: It converts processed data into XML file according to certain rules, and package the XML file in SOAP format.

(6) Data Processor: It interacts with other database or information system directly.

4. Implementation

The system uses the following development environment according to its characteristics.

The application server is Tomcat which is provided by Apache organization. Tomcat has two functions in the application system. Firstly, it works as a web server of enterprise internal MIS. Secondly, it provides Web services for the system.

development tool is Java programming language. It also needs JWSDP1.6 as extra support of additional development package, because JWSDP1.6 version has integrated web services and client tools.

The development platform is Microsoft Windows operating system. Java language can offer good solutions across multiple platforms, so web services can be easily deployed on other operating systems.

5. Summary

Through the comparison of current data integration strategy, the research on XML technology and web services technology, this paper proposes a heterogeneous data integration framework based on XML, and describes its internal structure in detail. The framework uses a star topology to facilitate unified management of data. Data exchange occurs in the central node. The node receives data from various enterprises, parses the data, converses the data format, and sends the data to destination. It creates a channel which is transparent and safe between the data source and destination by data exchange center.

References

[1] Z.Lu, Research and Realization of Heterogeneous Data Exchange System by XML, Computer Programming Skills & Maintenance, Feb. 2012, pp.32-34

[2] X.Y.Geng, XML Basics Tutorial, Tsinghua University Press, Beijing, 2006

[3] J.S.Wang, Data Exchange System Based on JMS and Web Services, Industrial Control Computer, vol.26, Nov. 2013, p.119

[4] Hass L.M., R.J.Miller, B.Niswonger, Transforming Heterogeneous Data with Database Middleware: Beyond Integration, Data Engineering, Sep. 2002, pp. 31-36

[5] J.L.Yang, G. Zhao, Mapping from XML to Relational Views in XML Based Heterogeneous Data Warehouse, Journal of Yunnan University of Nationalities (Natural Sciences Edition), vol. 22, May 2013, pp.369-372,