Multi-source Heterogeneous Data Integration Technology and Its Development

Yong Wang [a*], *Qi Shi* [a], *Hongtao Song* [a], *Zhigang Li* [a],and *Xue Chen* [a]
[a] College of Computer Science and Technology/Harbin Engineering University, China
[*]Corresponding author: Yong Wang, wangyongcs@hrbeu.edu.cn

Abstract
With the improvement in the level of social information, data has become basic, strategic resources, which represents the advanced productive forces and core competitiveness. How to share heterogeneous data highly and effectively, achieve effective integration of information system, make full use of vast amounts of data resources, reduce duplication of data collection and reduce administrative costs, so that improve the utilization of data is a major challenge faced by data managers. Wherein the effective data integration is one important way to solve the problem. This paper describes the basic concepts of data integration and development proceed, summarizes the principles and methods of the current mainstream data integration technologies, including XML technology, CORBA technology, middleware technology. And discussion and analysis of the key issues in the field of data integration, and future research directions to do a more thorough discussion.
Keywords: *data integration, multi-source heterogeneous, XML technology, CORBA technology, middleware technology*

1 Introduction
With the rapid development of science and technology and information technology constantly promote, the amount of data generated by the human society is growing exponentially. Today's society has entered a veritable "big data era." Data has become basic, strategic resources, which represents the advanced productive forces and core competitiveness[1]. How to achieve effective integration of information system by sharing data highly and effectively, so that people can make full use of large amounts of data resources, reduce duplication of data collection, reduce data management costs, mining its underlying information to improve the utilization of data is the major challenge faced by managers.
However, in the implementation of data sharing, there often appears the following part of the problems:

a) Data sources provided by the users are quite different ways. The content, format and quality of data vary greatly. When conversing different formats, it may not be converted or data loss problems may occur after the conversion.
b) Integration of heterogeneous data sources, apart from the integrated structure of the data, it also needs to integrate semi-structured and unstructured data of different data sources. These data sources are not only different in data model, but in the ability of query, which brings a big problem to the users' use and maintenance[2].
c) In the data sharing process, different data may be present in different platforms, in different forms, to be accessed in different ways, thus causing difficulties in data sharing.
d) When establishing their own applications and data storage, all kinds of units, organizations are lack of unified planning and management, they use different core implementation techniques and storage technology[3]. So the data often become "islands

of information", which not only improves the cost of data holder to maintain, and the data holder is difficult to play the role of distributed data and make the right decisions.

The above situation has seriously hampered the smooth flow of data and efficient sharing between the various types of users and systems, which cannot play the role of data, causing that some areas of information systems cannot be effective integration. Therefore, effective data integration has become an inevitable choice of data managers.

Multi-source heterogeneous data integration technology has been widely applied in different fields, such as geological exploration, land planning, digital mine, etc. Domestic and foreign scholars have carried out a lot of fruitful research, and made a series of remarkable achievements. Purpose of this paper is to introduce the basic concepts and development of data integration to start, and summarizes the principles and methods of the current mainstream data integration technology[4]. Discuss and analyze the key issues in the field of the data integration, and do a more thorough discussion in future research directions.

2 The Concept and Development of Data Integration

2.1 *Definition of Heterogeneous Data*

Heterogeneous data is a meaning rich concept, which refers not only to the data between different database systems are heterogeneous, such as SQL Server database and DB2, MySQL, etc. But also heterogeneity between data of different structures, such as structured data of relational database and semi-structured data of XML.

2.2 Definations and Targets of Data Integration

Data integration is to centralize the data of different sources, formats, features, properties logically, physically and organically, thus providing a comprehensive data sharing. Through the exchange of data between applications to achieve integration, mainly to solve the problem of distribution and heterogeneity of data.

Data integration is used to provide a unified representation, storage and management of all a variety of heterogeneous data. These features are implemented in heterogeneous data integration system[5]. Data integration masks differences between heterogeneous data, and carries out unified operation by heterogeneous data integration system. Therefore, it is uniform and no difference to users after the integration of heterogeneous data. The ideal goal of data integration is to provide a single system image in a distributed environment for users, which means that the interaction between the various data sources must be transparent[6]. Thus, data integration is not the simple accumulation of data or data carriers. Data integration is mainly to solve the data exchange and sharing between the various islands of automation. Thus, data integration is to achieve data conversion, unified data source, data consistency maintenance, data transfer between different application systems in heterogeneous environments, etc[7].

3 Data Integration Technology

3.1 XML-based Data Integration Technology

XML is short for Extensible Markup Language, which is designed by the W3C (World Wide Web Consortium) organization, and released in February 1998. It is a series of markup language, also known as meta-markup language. XML has the features of data reuse, data

separating indicate, scalability, and ease of programming features, which has the advantages of describing those very complex data. Therefore, put XML language into data integration enables heterogeneous data source integration middleware to better for enterprise data integration, convenient to data holders and users of the data reasonable use. In addition, XML language has self-description with content, structured storage, separation of content and representation, scalability, flexibility and cross-platform features[8].

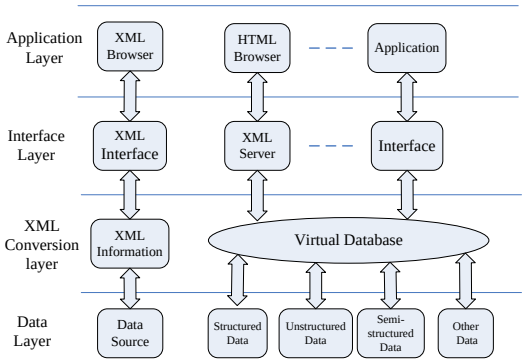

Fig. 1 – XML-based data integration framework

As shown in Fig. 1, use XML language to describe irregular data and use XML as the data integration layer described tools and conversion tools to construct data integration middleware[9]. From different applications to integrate data and put the resulting data into a single XML file and send it to the client, the parsed XML data can be edited or manipulated locally, thus meet the needs of Web development, simplify the implementation of Web data source integration system.

In XML-based data integration technology, user access to information or data is not directly acting on each heterogeneous data sources, but to achieve through XML-Enabled “virtual database”. Through XML, we can integrate and unify heterogeneous data from different data sources, and support different types of terminal equipment.

3.2 CORBA-based Data Integration

CORBA is short for Common Object Request Broker Architecture, which is currently distributed object technology, including one of the three distributed object technologies that OMG (Object Management Group) organization sets. It is widely used and provided a good norm and technical standards for distributed integration[10].

Firstly, the client (UI user interface) access request (standard model data) to the data source. The request is parsed by global server. And then query the catalog summary information in accordance with heterogeneous data sources, query decomposition and conversion through the ORB sub-bus, use a core part of the framework of the package will break down after the query into a particular data source can be identified sub-queries, and sent to the appropriate data source for execution, and finally the data source query results again translated by the wrapper into a standard data model, returned to the client[11].

3.3 Based on Data Warehouse Data Integration

Data warehouse refers to such a storage pool of data from different places. Data from heterogeneous data sources or database store, retrieve and maintain after processing. Traditional database mainly for business processing. And data warehouse for complex data analysis and high-level decision support. The data warehouse provides integrated and historical data from different types of applications for the department or enterprise-wide

strategic decisions and overall long-term trend analysis to provide effective support. Data Warehouse makes it difficult to share the data. The user has the freedom to extract data without interfering with the normal operation of the database business[12].

Based on Data Warehouse Data Integration refers to put data and information extracted from different data sources firstly, and then converts the data into a common data model and existing data warehouse and integrated[13]. When the user queries to the warehouse, the required information is ready. Data conflict, expressing inconsistencies and other issues have been resolved, which makes the decision-making searches easier and more effective.

3.4 Data Integration based on middleware technology

Middleware integration technology is a typical mode of integration technology, which uses global data model[14]. And federal databases, middleware system not only can integrate structured data source information can also be integrated semi-structured or unstructured data source information, such as Web information.

Middleware solves the problem of heterogeneous distribution and distributed components of the system, to achieve the separation of application logic and system services concerns, simplifying application development. But also on the structure and performance of distributed systems impact[15]. Distributed application software share resources with middleware technology between different technology. Large enterprise distributed systems and technology together to achieve large-scale integrated enterprise application software systems[16]. It does this by providing all heterogeneous data sources to integrate their virtual view, data source can be a database, WEB and other data sources and is still stored in the local data source to ensure full autonomy. As shown in Fig.2, the system provides users with a global schema, query submitted by the user is directed to the global mode, so the location of the data source, and the access mode is transparent to the user.

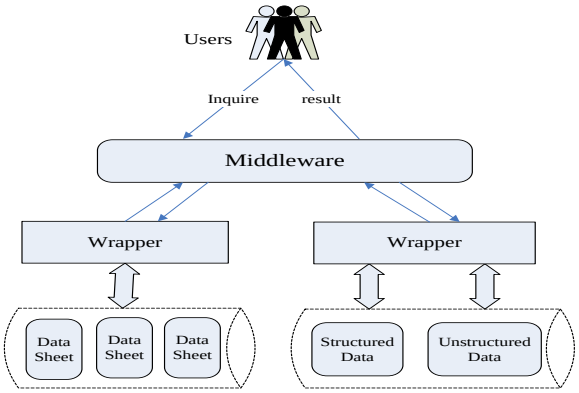

Fig. 2 – Middleware-based data integration framework

3.5 Based on the Number of Federated Database Management System Integrated

Multiple Data Source Data Integration concept was first raised from the mid-1980s heterogeneous database integration. The widely used approach was federated database management system. Federated database management system supports data sharing between multiple heterogeneous databases, interaction and autonomy, and support for multi-database query design but most involve only one library to modify federal affairs. The federal database has any local input or output mode. Enter the model and conceptual model and constitute a federal model, the entire system is not global mode. Use federated database management system can achieve independent heterogeneous database system integration, to achieve the multi-database data sharing[17].

Federated database system integration can be divided into two categories by integration: tightly coupled and loosely coupled federated database system federated database system. Tightly coupled federated database system uses a unified global model, each data source is mapped to the data pattern of global data model to address the heterogeneity of data between sources. This method of integration is high, and less user involvement; drawback is to build a global data model algorithm complexity, poor scalability. Loosely coupled federated database system is rather special, no global model, using the federal model[18]. The method provides a unified query language, a lot of the heterogeneity problem to the users themselves to resolve. The method of the loosely coupled integration of data is not high. But the autonomy of its data source, good dynamic performance, integrated systems need to maintain a global model.

4 Summary and Outlook

Multi-source heterogeneous data integration has been a classic, valuable research. There is a strong comprehensive. The core is to solve the distribution, autonomy and heterogeneity problems of heterogeneous data sources. It involves many aspects of a variety of computer technologies, such as distributed object technology, database technology, information security technology and artificial intelligence technology. In the current big data environment, new problems are emerging, changing the process, prompting the issue is evolving.

Data integration technology is developing rapidly, the quality of different methods for data integration after the establishment of a unified framework and criteria to evaluate the need for further reflection; the arrival of the era of big data, the amount of data is growing exponentially, whether traditional data integration techniques can adapt massive, streaming, and efficiency of high-speed integrated approach is worthy of study; at the same time, artificial intelligence technology in recent years to form a new boom, such as the depth of learning, learning, and intelligent body migration technology, combined with machine learning techniques, will may greatly improve the efficiency and quality of data integration.

Acknowledgements

The research work was supported by The Fundamental Research Funds for the Central Universities under Grant No. HEUCF160604, HEUCF160612, The National Key Technology R&D Program of the Ministry of Science and Technology under Grant No. 2012BAH81F02, and The Youth Foundation of Heilongjiang Province of China under Grant No. QC2016083.

References

1. *A. P.Sheth, J. A. Larson,* Federated databases for managing distributed, heterogeneous, and autonomous databases, J. Computing Surveys, **22**(1990) 183-236.
2. *N. Wang, N. Wang,* Heterogeneous Data Integration System Query Decompositon and Optimization Implementation, J. Journal of Software. **11**(2000) 222-228**.**
3. *M.Lenzerini,* Data integration: A theoretical perspective, C. Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM, 2002: 233-246.
4. *G. C.Begg, S. Y.Griffin W. L., S. Y.O'Reilly, et al,* Geoscience Data Integration: Insights into Mapping Lithospheric Architecture, J. ASEG Extended Abstracts, **1**(2015) 1-2.

5. *C .Daraio, M .Lenzerini, C .Leporelli, et al,* Data integration for research and innovation policy: an Ontology-Based Data Management approach, J. Scientometrics, **2**(2016) 857-871.
6. *P. Papotti, F.Naumann, S.Kruse, et al,* SYSTEMS AND METHODS FOR DATA INTEGRATION: U.S. Patent 20,160,154,830, P. 2016-6-2.
7. *Y. Chen, J. Wang,* Summary of Data Integration, J. Computer Science. **31**(2014) 48-51.
8. *J. Li, M. Zhou, G. Geng, etc,* Application of XML in Heterogeneous Data Integration J. Computer Application, **22**(2002) 10-12.
9. *E. H.Baugh, R.Simmons-Edler, C. L.Mueller, et al,* Robust classification of protein variation using structural modelling and large-scale data integration, J. Nucleic acids research, **6**(2016) 2501-2513.
10. *G.uanyu Li, J. Liu, J. Zhang, etc,* Research and Implementation of Distributed Heterogneous Data Integration System J. Application Research of Computers, **21**(2004) 96-98.
11. *G.Baele, M. A.Suchard, A.Rambaut, et al,* Emerging concepts of data integration in pathogen phylodynamics, J. Systematic Biology, 2016: syw054.
12. *E.Scheurwegs, K. Luyckx, L. Luyten, et al,* Data integration of structured and unstructured sources for assigning clinical codes to patient stays, J. Journal of the American Medical Informatics Association, **23**(2016) 11-19.
13. *K. Hu, S. Xia,* Summary of Data Mining Based on a Large Database J. Journal of Software, **9**(1998) 53-63.
14. *A.Calì, D. Calvanese, De Giacomo G, et al,* Data integration under integrity constraints, M. Seminal Contributions to Information Systems Engineering. Springer Berlin Heidelberg, 2013: 335-352.
15. *X. L.Dong, D.Srivastava,* Big data integration, C. Data Engineering (ICDE), 2013 IEEE 29th International Conference on. IEEE, 2013: 1245-1248.
16. *Y. Zhang, T. Huang, J. Wei, etc,* Evaluation of Structural Components of the System Performance Middleware Container System J. Journal of Software, **17**(2006) 1328-1337.
17. *D.Gomez-Cabrero, I.Abugessaisa, D. Maier, et al,* Data integration in the era of omics: current and future challenges, J. BMC systems biology, **2**(2014) 1.
18. *Y. Jin, N. Wang,* Design and Implementation of Federated Database Management System DBMS J. Journal of Computers. (1993) 431-436.

