ENTERPRISE INFORMATION INTEGRATION: STATE OF THE ART AND TECHNICAL CHALLENGES

Jingtao Zhou, Mingwei Wang, Han Zhao
The Key Laboratory of Contemporary Design and Integrated Manufacturing Technology, Ministry of Education, Northwestern Polytechnical University, Xi'an, China, 710072. zhou.jingtao@gmail.com

Abstract: To gain insight into increasingly intricate business, and deal with highly complex problem situations, an enterprise needs a more generic, standardized, pervasive and scalable infrastructure to fully leverage the information from different data sources, applications, and environments at both system and semantic level. In this context, we first discuss the state of the art of current approaches and solutions of EII (Enterprise Information Integration). Then, we outline the grand challenges of EII from technical perspective based on the analysis of framework, range, scale and performance.

Key words: Enterprise Information Integration; Scalability; Horizontal Integration; Vertical Integration; Centralized Integration; Semantic Interoperability

1. INTRODUCTION

The global changes in economic, shift of the competitive edge and the advent of new technology continually alter the manufacturing and business environment in which enterprises operate. Manufacturing and business management has been extending outside an enterprise in a distributed form geographically or according to business logic spreading across multiple enterprises.[1] The trend of long- or short-time close collaboration across the whole vertical and horizontal manufacturing industries has been illustrated by the implementation of e-commerce, e-business, or virtual enterprise.[2]

Please use the following format when citing this chapter:
Zhou, Jingtao, Wang, Mingwei, Zhao, Han, 2006, in International Federation for Information Processing (IFIP), Volume 207, Knowledge Enterprise: Intelligent Strategies In Product Design, Manufacturing, and Management, eds. K. Wang, Kovacs G., Wozny M., Fang M., (Boston: Springer), pp. 847-852.

Surviving in such an increasing globalization and flexibility environment requires an extremely flexible, self-adaptive IT infrastructure capable of integrating and coordinating any involved information from any heterogeneous data sources, applications, and environments at both system and semantic level to facilitate interoperation and collaboration over large-scale computer networks.

However, in spite of extensive R&D and successful pilots, traditional enterprise information infrastructure is poorly suited for dealing with the strategic, long-term barriers to efficient information sharing across enterprise internal and external boundaries at both system and semantic or knowledge level. The lack of suitable basic framework leads to many information solutions that have to make overmuch tradeoff between long-term adaptability and short-term applicability, broad interoperability and tailored function for very specific purposes. They can not reap the full potential benefits of information, and ultimately fail to the pursuit of setting and realizing corporate strategic and tactical goals. One underlying problem has remained unsolved yet: data resides in thousands of incompatible formats and cannot be systematically and understandably managed, integrated, and reused. As a result, there is mounting pressure from enterprise itself and outside for a direct move away from disparate information systems operating in parallel towards a more common and fundamental shared architecture for semantic information interoperation.

2. STATE OF THE ART

Enterprises have long recognized the value of data integration. Efforts can be roughly classified into two categories: application centric integration (ACI) and data centric integration (DCI).

ACI, such as point to point integration and enterprise application integration (EAI) integrates relative data by linking applications through custom-coding or integration broker that acts as a hub to route messages between connected applications. Connection at the programmatic level, difficulty of metadata reuse, N square problem at data layer, multiple vendors for multiple systems, lack of common protocols, and tight-coupling of technology and systems in the end, make these solutions difficultly suitable to an open, dynamic information interoperation environment of businesses and operations in or across enterprises. In fact, ACI provides little data integration because it operates at the business-process level rather than data level.

DCI can be implemented by creating either a centralized repository for data access and analysis, such as data warehouses, or a data integrating layer

over a set of distinct and autonomous data sources, such as federated information systems.

Data warehouses are not effective in meeting ad hoc or opportunistic data integration needs. They fall short where an enterprise needs to address one need today and incrementally add additional requirements over time, or where there is a great deal of change in business requirements.[3]

Federated information integration approaches can be divided into two classes, including tightly-coupled and loosely-coupled systems. Tightly-coupled system integrates all data sources by creating logical mappings between them and a single global schema (a single model or ontology) which may be derived from the actual data sources themselves, or the modeling of current and even future business and operation aspects of information. The main drawbacks when using a global schema are the difficulty of creating a single sufficient global schema to represent a large-scale data sources, maintain of changing and evolution caused both from data sources and the global schema itself. Hence, tightly-coupled approach is only adapted to small-scale integration with little independently changing and evolving. By contrast, the loosely-coupled approach coordinates autonomous component data sources without a single global schema, instead, with a set of federated schemas. In the context of real-time information integration, loosely-coupled federation system is more effective and has advantages than other approaches. However, early loosely-coupled federated systems are not broadly applied in real enterprise environment because of the using of private protocol and data model, low performance, laborious process, critical implementation conditions, immature technology and the lack of reliable infrastructure. Several key problems such as discovery of relevant data, semantic conflicts and violation between the autonomy and privacy hinder its further application, too.

With recent advent of new technologies such as web service, XML and SOA, and new drivers behind e-commerce, e-business and e-enterprise, a representative approach for enterprise information integration (EII) is proposed, which intends to integrate any form data in enterprises, and aims to provide a uniform interface for information access, manipulate, and integrate across multiple data sources. Renaissance of data integration begins with the forming of EII industry. Broadly speaking, the architectures underlying most current new EII approaches (e.g. IBM DB2 II,[4] BEA Liquid,[5] MetaMatrix,[6] etc.) are still based on similar principles of loosely-coupled federated systems although they may support broad type data sources, use new XML model, speak with common protocols and publish integrated results with web services. Therefore, some traditional problems are inherited from loosely-coupled federated information integration system,

such as scalable and semantic problem. Furthermore, current EII solutions may encounter their own fierce challenges when they need to integrate all involved information across the whole vertical and horizontal management logic, from both intra-enterprise and inter-enterprise on a highly complex and dynamic networked environment in a timely fashion. The circumstances seem to be too rigorous but are actual in real-world manufacturing and business environment.

3. CHALLENGES

A resent survey[7] has addressed four challenges of EII including scaleup/performance, horizontal or vertical growth, integration with EAI, and metadata management/ semantic heterogeneity from a more general point of view. In this section, we will take a further discussion from a technical perspective.

3.1 Scalability

The framework of most current EII systems is constructed as a hierarchy framework, in general, with (fix) multi-layer, which can be considered as a variation of the traditional five-level federated architecture, [8] and may consist of local schema, export schema (possible wrapped by web service), federated or mediator schema (logic data view), and extern schema (data view for specific application and user, possible wrapped by web service). Generally, approaches relying on a priori creation of federated views do not scale-up efficiently given the complexity involved in constructing and maintaining a shared schema for a large number of, possibly independently managed and evolving, sources.[9] A solution to this problem is to represent federated schemas by a hierarchy of small finely granular schemas, such as business entity schema used in BEA Liquid[5,10] and MetaMatrix[6]. However, this approach needs more endeavors from specialists and relies on their strong intimate knowledge of the desired business entities to be created and deep understanding of the data, underlying schema, and relationships across the various data sources (e.g. BEA Liquid[5,10] and MetaMatrix[6]), which becomes a drawback for scalability when this knowledge grows and changes as more sources join the system and when sources are changing.[9]

3.2 Horizontal vs. vertical integration

From the view of business and manufacturing, data integration requirements come from both the vertical and horizontal logic level. The

framework of most current EII system is more appropriate for horizontal dimension of data integration rather than vertical dimension. Although the using of web service in some systems (e.g. data services in BEA liquid[5,10]) has shifted the focus on vertical data integration by service composition, the connection of vertical and horizontal integration has been artificially dissevered by the isolated definition of business concepts, which are dynamically composed only according to the specific application logic, only involving sources that have direct complementary data described by the corresponding federated schema, and ignoring the nature semantic relationship between themselves. As a consequence, it can not consider the potential relative data in other data sources even though the relationship is implied by other federated schema which is not directly involved in an integration process. It does not consider the probable incompleteness of some sources, either. In fact, these scenarios can be avoided by establishing the network of relationships among federated schemas as well as data sources to connect the horizontal and vertical data integration.

3.3 Centralized integration

Most implementations of EII generally resolve queries from dedicated servers that house federated metadata. Not only do these dedicated servers generally form a bottleneck in terms of scaling performance, but the centralized computing model can be a scalability bottleneck in terms of administration[11]. Federated schema design must be done globally; any changes of the federated schema can be only made by the central administrator. This can be especially challenging when data is owned and managed by numerous heterogeneous groups with different needs, and when integrate data across organizations.

3.4 Semantic

Even for information that is carried by web service and represented using XML, a serious and expensive barrier to dramatic improvement exists: a severe lack of explicit knowledge about associated corporate information. The thorny question of locating and understanding the data to be integrated still remains[7]. The assumption of most of the existing approaches that the relevant data is either already known or identified is not reasonable when integration involves multiple domains, multiple boundaries, or large-scale data sources. Key technologies of data semantics discovering, understanding, and semantic conflicts detecting and resolving are still in their infancy stage.

Achieving an information sharing environment enabling the feasible semantic interoperability becomes really a significant challenge.

4. CONCLUSIONS

Clearly, to overcome the challenges of EII and create a more flexibly and dynamically semantic interoperation environment for enterprise information over a large-scale computer networks, we believe there is a need for a new class of data sharing infrastructure. Such infrastructure is fundamentally distributed and dynamic networked, supports highly flexible sharing relationships, ranging from client-server to peer-to-peer, addresses scalability as well as resource control, and achieves interoperability at not only system level but semantic or knowledge level.

5. REFERENCE

1. Q. Wang , L. Y. Kai, and H. Wai, A hierarchical multi-view modeling for Networked Joint Manufacturing System, *Computers in Industry* **53**(1), 59–73 (2004).
2. Y.-E. Nahm and H. Ishikawa, A hybrid multi-agent system architecture for enterprise integration using computer networks, *Robotics and Computer-Integrated Manufacturing* **21**(3), 217-234(2005).
3. Nimble Technology, Next-Generation Data Integration: Harnessing Data for Business Advantage (14 January, 2002); http://uk.builder.com/whitepapers/0,39026692,60087734p-39001077q,00.htm.
4. B. Paolo, A. Francis, B. Amanda, et al., *Data Federation with IBM DB2 Information Integrator V8.1* (IBM Redbook, October 2003).
5. R. B. Vinayak, Liquid Data for WebLogic: Integrating Enterprise Data and Services, in: *ACM SIGMOD International Conference on Management of Data*, edited by W. Gerhard , C. K. Arnd, D. Stefan (ACM, Paris, 2004), pp.917-918.
6. H. Randall, M. Alex, and C. Rob, Information Intelligence: Metadata for Information Discovery, Access, and Integration, in: *ACM SIGMOD International Conference on Management of Data*, edited by O. Fatma (ACM, Baltimore, 2005), pp. 793-798.
7. H. Y. Alon, A. Naveen, B. Dina, C. Michael , et al. Enterprise Information Integration: Successes, Challenges and Controversies, in: *ACM SIGMOD International Conference on Management of Data*, edited by O. Fatma (ACM, Baltimore, 2005), pp. 778-787.
8. A.P. Sheth and J.A. Larson, Federated database systems for managing distributed, heterogeneous, and autonomous databases, *ACM Computing Surveys* **22**(3), 183-236 (1990).
9. C. Nazli, M. E. Stuart, M. Allen, et al., Information Integration for Counter Terrorism Activities: The Requirement for Context Mediation, Working Paper CISL# 2003-09.
10. BEA, BEA AquaLogic Services Platform: Concepts Guide, BEA, 2005.
11. G. I. Zachary, *Efficient Query Processing for Data Integration*, Phd thesis, university of Washington, 2002.