Journal of Theoretical and Applied Information Technology
10th August 2013. Vol. 54 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

DATABASE INTEGRATION: IMPORTANCE AND APPROACHES

[1] IBRAHIM ALMAHY, [1]NAOMIE SALIM
1 Faculty of Computer Science and Information Systems,
University Technology Malaysia, 81310, Johor, Malaysia
E-mail: abu_alabbas@hotmail.com

ABSTRACT

Database integration is a multistep process of finding similar entities in two or more databases to create a non-redundant, unified view of all databases. Database integration has become an active area for research due to increasing of information resources with the need of users and applications to integrate data from these different sources. This paper discusses the role of data integration and the fields that require it, problem of integration and the processes of integration. Finally it presents the current methods used in database integration in a form of graph that classified the approaches in clear manner.
Keywords: *Database integration; integration approaches; schema matching*

1. INTRODUCTION

With the growth, widespread and increasing of information resources, users and applications need to integrate data from different sources. These resources are often having a degree of heterogeneity. There is an urgent need to overcome this obstacle and combine the information.

Database integration is a multistep process of finding similar entities in two or more databases to create a non-redundant, unified view of all databases [1].Integrating data has been requested in many fields such as marketing, biological, bioinformatics, medicine, industries and others. But it became inventible in two major areas. Next section argues these two.

Marketing and Enterprises: large enterprises spend a great portions of time and money on "information integration"—combining information from various sources into a unified structure and format. Frequently cited as the biggest and most expensive challenge that information-technology shops face, information integration is thought to consume about 40% of their budget [2].

The need for integration in enterprise information arises clearly in the process of: messages exchanging and merging companies. Merging one company in another includes merging its systems and database which may be completely different and vary widely in their content, formats and access methods.

In Bioinformatics: with the accomplishment of Human Genome Project, a mass of original data are collected by researchers. In order to handle those data better, hundreds of databases of Bioinformatics are formed, such as the three major international databases of the nucleic acid: Genbank[3],the European Molecular Biology Laboratory (EMBL)[4], database and the DNA DataBase of Japan (DDBJ) [5].In general, More than 500 biological databases exist at present [6]. With this great number of database, researchers and experts need to integrate the information. The key point is how to share those heterogeneous databases and make a common query platform for users. [7]The same problem, and the same thing applies to the rest of the areas mentioned above.

2. PROBLEM OF INTEGRATION

The basic problem in data integration is heterogeneity of data in many levels including system-level and semantic level heterogeneity. Integration of heterogeneous data sources mean share information and provide users with a

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

unified data [7]. In other words we should find the similarities between the elements of those schemas. Thus, integrating different databases initially requires integrating their schemas, which may vary in structure, semantics and instance, by finding correspondences between elements and specifying the similarities degrees. Integrating schemas passes through several stages, namely: Schema Matching, Schema mapping and Schema Integration. Next section talks about these phases.

2.1. Schema Matching

Schema matching is the problem of generating correspondences between elements of two schemas. A correspondence refers to relationship between one or more elements of one schema and one or more elements of the other. Schemas have many forms like SQL schema, XML schema, entity-relationship diagram, ontology description and others. Among all processes in integration, schema matching is the major one. it considered as a main bone for all approaches in integration due to existence life in many applications. For example, it appears in object-to-relational mappings, data exchange, data warehouse loading and mediated schemas for data integration. In knowledge-based applications, such as sciences applications and the semantic web, it arises in the alignment of ontologies. For example, it may be used to align gene ontologies or anatomical structures. In health care, it may arise in the alignment of patient records and other medical reports. In web applications, it may be used to align product catalogs. In ecommerce, it may be used to align message formats representing business documents, such as orders and invoices [8].

To match two schemas there are many techniques proposed. The common ones are: Linguistic matching, Instance-based matching, Structure-based matching, Hybrid-matching and others.

Linguistic matching is a technique based on language perspective where it takes the element's name or description and compares the tokens of words to determine the degree of similarity.

We mentioned particularly to linguistic approach because Cupid is depending on it, which considered as a base in the field schema matching according to [9] Keeps track of research, we find that schema matching has been hot area of research for over 20 years .The increase in the volume of available data, heterogeneous databases, have each increased the importance of developing effective schema matching solutions [10]. [8] Is modern Comprehensive survey paper written in this field, it discusses the techniques, strategies, approach and tools in the field of schema matching.

2.2. Schema Integration

Once matching is finished, the relations and correspondences between the various schemas have been identified. Next phase is to create the Global Schema, and this is referred to as schema integration.

2.3. Schema Mapping

After a global schema is defined in previous phase, it is time to determine how the data from the different sources can be mapped to global schema while keeping consistency of data. [12] Speaks in depth about schema integration and mapping.

2.4. Ontology

To find the similarities between deferent elements, one of the schema matching techniques mentioned earlier should be used. This is one direction. Another direction and the newest is ontology. Most recent research of integration field is using ontology based techniques to find correspondences between concepts. The following section discusses this point.

There are many definitions about what an ontology is, but the common thread in these definitions is that an ontology is some formal description of a domain of discourse, intended for sharing among different applications, and expressed in a language that can be used for

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

reasoning [13].according to [14] ontology is an explicit formal specifications of the terms in the domain and relations among them.

Ontologies are, basically, took popularity in the AI field as a mean for constructing formal vocabulary in explicit manner to share between applications. Therefore it is clear that ontology goals are not for integrating heterogeneity [15]. However it has ability to find the degree of similarity between two or more different words, consequently it is used broadly in integration.

[15] Adopted an approach similar to the one that used in ontology translation for the Semantic Web [16]. They use Web-PDDL ontology language to model database schemas, elements, and the relationships (mappings) among them. This is the first ontology approach in integrating relational database as they claim in [15].So many works have been done in using ontology to integration. [17] Is modern survey paper as far as we know.

As we mentioned earlier, schema matching considered as base in every approach in database integration. Next section discusses these approaches.

2.5. Database Integration Approaches

There are many ways to integrate data. The three common approaches are Data Warehouse Approach, Federated Database (or Mediator Approach) and middleware technology [7] and [11].

2.6. Data Warehousing Approach

According to [12] Database integration can be either logical or physical. Data warehouse considered as physical integration because integrated data is materialized.

In logical data integration, data is not materialized. Instead integration is only virtual. with data warehousing [18], data from each data source is extracted, merged, and stored in a centralized repository (warehouse). The warehouse is a database with a global schema that combines the schemas of the sources. Queries on the system are evaluated at the warehouse without accessing the original sources. Client updates to the warehouse are usually not allowed since they are not propagated to the original sources and would make the warehouse inconsistent with the sources. Instead, the warehouse is updated from the data in the sources. There are multiple policies for updating the warehouse from the sources [19].

Putting large amount of data in one place may cause some difficulties to deal. Maintenance also considered a challenge issue. Consequently, it can create problems with queries since queries are only as relevant as the latest updates [20].as a result for the constraints of data warehouse it's better to use it for the creation of highly curated datasets focused on a specific and narrow area of research.[21]

2.7. Federated Database Approach

Federated database systems integrate databases by implementing one-to-one connections between all databases that need to communicate with each other.

Federated database different from data warehouse in that the last one is takes data from various sources and collects it in one place (global schema). Changes in one source may lead to change in the data warehouse. Whereas the federated approach provides a virtual data warehouse without moving the main data. From this perspective, the term “database federation” refers to architecture in which middleware, consisting of a relational database management system, provides uniform access to a number of heterogeneous data sources [22]. The main role of addition layer is to translate queries come from clients and pass them.

Many researchers have been done in this area. Pioneering research projects included TSIMMIS13 and HERMES, 14 which used database concepts to implement “mediators,” special purpose query engines that use nonprocedural specifications to integrate specific data sources. DISCO15 and Pegasus16 were closer in feel to true database federation [22].

Journal of Theoretical and Applied Information Technology
10th August 2013. Vol. 54 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The disadvantage of this federated approach is that components of translation need to be written for each pair of communicating databases. We noted that most approaches that based on middleware are not concern with specific data model (relational, xml, ODB...). They either integrate multi-source heterogeneous data.

Figure (1) presents the classification for integration methods.

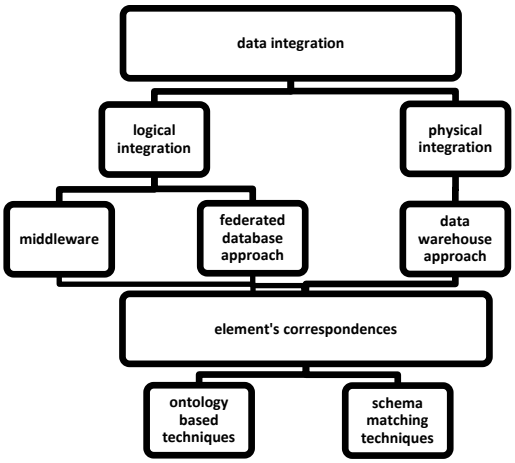

Figure 1: Data Integration Approaches And Techniques.

3. CONCLUSION

Database integration has been an active area because it has become required in many areas and fields. Several approaches adopted to achieve the goal of integration. No single technique is better than the others, and different integration solutions serve different purposes. New trend semantic web integration attracted many researches later [23] [24].many recently works use ontology instead of traditional schema matching techniques.

4. ACKNOWLEDGMENT

This work is supported by Ministry of Higher Education (MOHE) and Research Management Centre (RMC) at the Universiti Teknologi Malaysia (UTM) under Research University Grant Category (VOT Q.J130000.7128.00H72). We also would like to thank MIS-MOHE for sponsoring the first author.

5. REFERENCES:

[1] C. Batini, M. Lenzarini, and S. B. Navathe, A Comparative Analysis of Methodologies for Database Schema Integration, ACM Comput. Surv., vol. 18(4), 1986, pp. 323-364.

[2] Philip A. Bernstein and Laura M. Haas, information integration in enterprices, communications of the acm | september 2008 | vol. 51 | no. 9 [3] Dennis A. Benson, Ilene Karsch-Mizrachi, David J. Lipman,

James Ostell, and David L. Wheeler, "GenBank", Nucleic Acids Res (Database issue), 2005, pp. 34-38.

[4] Kulikova T., Aldebert, P., Althorpe, N., Baker, W., Bates, K.,Browne, P., van den Broek,A., Cochrane, G., Duggan, K.,Eberhardt, R. et al, "The EMBL Nucleotide Sequence Database",Nucleic Acids Res, 2004, 32, pp.27–30.2

[5] Miyazaki S., Sugawara,H., Gojobori, T. and Tateno, Y, "DNA Data Bank of Japan (DDBJ) in XML", Nucleic Acids Res, 2003, 31, pp.13–1656

[6] Jacob Köhler, "Integration of life science databases", DDT: BIOSILICO Vol. 2, No. 2 March 2004

[7] Yuelan Liu, Xiaoming Liu, Lu Yang "Analysis and Design of Heterogeneous Bioinformatics Database Integration System Based on Middleware" IEEE 2010

[8] Philip A. Jayant, Madhavan , Erhard Rahm "Generic Schema Matching, Ten Years Later " , 2011 VLDB

[9] Madhavan, J., P. A. Bernstein, and E. Rahm: Generic Schema Matching with Cupid. Proc. VLDB, 49-58, 2001

[10] AGal , "Why is Schema Matching Tough and What Can We Do About It?", ACM Sigmod Record - 2006 ,

[11] CheoThiamYui, Lim Jun Liang, Wong Jik Soon, and Wahidah Husain "A Survey on Data Integration in Bioinformatics" Springer-Verlag Berlin Heidelberg 2011

[12] Özsu, M. Tamer, Valduriez, Patrick , "Principles of Distributed Database Systems" , 3rd Edition., 2011, XIX, 845 p

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[13] C. Welty. Ontology research. AI Magazine, 24(3), 2003

[14] Thomas R. Gruber, "A Translation Approach to Portable Ontology Specifications", Knowledge Acquisition, 5(2):199-220, 1993.

[15] Dejing Dou "Ontology-based Integration for Relational Databases" SAC'06 April 23-27, 2006, ACM

[16] D. Dou, D. V. McDermott, and P. Qi. Ontology Translation on the Semantic Web. Journal of Data Semantics, 2:35–57, 2005.

[17] PavelShvaiko, and J´erˆomeEuzenat," A Survey of Schema-Based Matching Approaches" , Journal on Data Semantics IV, LNCS 3730, pp. 146–171, 2005

[18] Jennifer Widom , "Research Problems in Data Warehousing", 1995 ACM 0-89791-8124/95/11

[19] HGarcia-Molina ,JD Ullman, "Database system implementation", - 2000 - csd.uoc.gr

[20] Mork P, Halevy A, Tarczy-Hornoch P. A model for data integration systems of biomedical data applied to online genetic databases. AMIA Symp 2001:473–7

[21]Brenton Louie , Peter Mork , Fernando Martin-Sanchez, "Data integration and genomic medicine", 2006 Elsevier

[22] L. M. Haas E. T. Lin M. A. Roth., Data integration through database federation, IBM SYSTEMS JOURNAL, VOL 41, NO 4, 2002

[23] Andreas Langegger,WolframW¨oß, and Martin Bl¨ochl, "A Semantic Web Middleware for Virtual DataIntegration on the Web", Springer-Verlag Berlin Heidelberg 2008

[24] He, B., Patel, M., Zhang, Z., Chang, K.C.-C.:Accessing the deep web. Commun. ACM50(5), 94–101 (2007)