Volume 2, Issue 7, July 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Data Integration Challenges and Solutions: A Study

Kingshuk Srivastava, P.S.V.S.Sridhar, Ankit Dehwal
Centre for Information Technology
University of Petroleum & Energy Studies
Dehradun, India

Abstract*— The concept of data integration is one of the oldest studies which came into being since the conception of database management system. Basically data integration can be defined as the problem of combining of data from heterogeneous sources to one unified structure, so that user is able to view it as a single entity irrespective of the origination or it's data type. In today's highly competitive market scenario every business decision must be made on strong and reliable data foundation. These data must contain historical, current and sometimes even real time values from different & most likely from heterogeneous sources. Data integration is the technology which enables the system to deliver an infrastructure for the purpose of business intelligence (BI). The biggest challenge of data integration is to make this happen in real time environment. In this paper we list the problems and issues related to data integration in today's information technology from a theoretical perspective. Emphasis would primarily be on modelling, processing of queries, consistent and in consistent data sources and reasoning on queries.

Keywords*— Data Integration, Schema, Business Intelligence, XML, Data warehousing.

I. INTRODUCTION

Data integration is defined as the technique to integrate or collect data from different sources and merge them at one place and finally gives a virtual view to the users. Integration of multiple information systems aims at combining selected systems so that they form a unified new whole and give users the illusion of interacting with one single information system [1]. Integration of information systems seems to be necessary in today's world to meet business and consumers needs. There are two reasons for integration: primarily, in a given set of information systems, an integrated view could be created to enable information access and reuse through a single access point. Secondarily, towards a particular information need, data from different information systems is accumulated to gain a more comprehensive basis towards the required need [10].

There are so many applications that are benefited from integration. In the area of Business Intelligence integrated information can be used for querying and reporting on business activities. In customer relationship management (CRM), integrated information on individual customers, business environment trends, and current sales can be used to improve customer.

Enterprise Information Portals (EIP) present integrated company information as personalized web sites and represent single information access points primarily for employees, but also for customers, business partners, and the public. Lastly, in the area of E-Commerce and E-Business, an integrated information system acts as a facilitator as well as an enabler towards business transactions and services over computer networks [1].

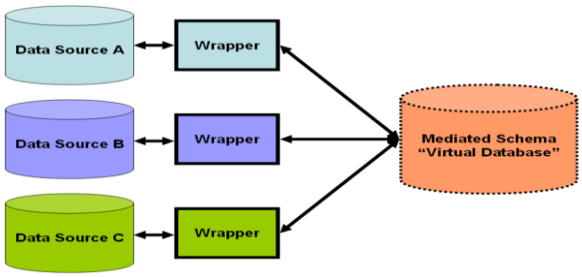

Fig. 1 Data integration chart

II. CHALLENGES OF DATA INTEGRATION

Integrating multiple information systems creates a unified virtual view to the user's imperative of the number of system or location of the actual stored data. The actual users are provided with a homogeneous logical view which is physically distributed over different heterogeneous data sources. For this, all data has to be represented using the same abstraction principles (unified global data model and unified semantics) [5].

Data integration is hard. The evidence is overwhelming. Every company we've talked to about their data has data integration problem. It's not just the IT people that moan about it either, it's IT users too and the company executives. Everywhere data is almost in a constant mess throughout. Today we have a dedicated sector of the industry devoted towards data integration solution; it generates about $3 billion in revenue and its growing space. Aside from that there are probably billions more spent on in house data integration efforts whether they employ the whiz data integration tools or not [12].

This task includes detection and resolution of schema and data conflicts regarding structure and semantics. In general, information systems are not designed for integration. Thus, whenever integrated access to different source systems is desired, the sources and their data that do not fit together have to be coalesced by additional adaptation and reconciliation functionality. Note that there is not the one single integration problem. While the goal is always to provide a homogeneous, unified view on data from different sources, the particular integration task may depend on the following [2]:

- **Architectural view of information system**- The art to express a model or a concept of information for utilization in activities which requires explicit details of complex systems. The most common activities in this sector are library systems, Content Management Systems, Web development, User interactions, Database development, Programming, Technical writing, Enterprise architecture and critical systems software design. As different information systems has different architecture, it becomes very difficult to integrate information from different architecture. [1][2].
- **Content and functionality of the component systems** – Component systems of any architecture have contain different content and have different functionality which made it difficult to integrate [2].
- **Source wrapping** - To access data sources on the Web, a crucial step is wrapping, which translates query responses, rendered in textual HTML, back into their relational form. Traditionally, this problem has been addressed with syntax-based approaches for a single source. However, as online databases multiply, we often need to wrap multiple sources, in particular for domain-based integration.
- **Semantic data** - A chief requirement of data integration systems is that the differences in the syntax and semantics of the underlying data sources should be hidden from the user. However, data sources are often independently created and do not easily interoperate with other sources.
  The rise of XML as a data interchange standard has led to a new era in which research has focused on semi-structured data. Although XML has provided a single interchange format, different users can model the same data in different ways, which can lead to heterogeneities at various levels, including the semantic level. Semantic heterogeneities can arise from entities being perceived differently [13].
- **Streaming data**- Now a day in scientific and industrial environments, the amount of data in form of heterogeneous streams is becoming one of the main sources of information and knowledge acquisition. Advances in wireless communications and sensor technologies have enabled the deployment of networks of interconnected devices capable of ubiquitous data capture, processing and delivery of such streams [5].
- **Large scale automatic schema matching**- In the integration of heterogeneous data sources Schema matching becomes a critical problem. Basically as well as traditionally, the problem of matching multiple schemas has essentially relied on finding pair wise -attribute correspondences in isolation. Schema matching is a basic operation of data integration and several tools for automating it have been proposed and evaluated in the database community. Research in this area reveals that there is no single schema matcher that is guaranteed to succeed in finding a good mapping for all possible domains, and thus an ensemble of schema matchers should be considered. Informally, schema meta-matching stands for computing a "consensus" ranking of alternative mappings between two schemata, given the "individual" graded rankings provided by several schema matchers [3][6].
- **Construction of global schema** - schema here referred to structure data residing at a place has a structure different then other and integration of data from different structure proves to be difficult [6][7][8].
  - Kind of info-this include alphanumeric data, multimedia data; structured, semi-structured, unstructured data.
  - Available resources- time, money, human resources, know-how, etc.
- **Understand Data Needs** – It could be defined as the delivery of the right data to the right application in order to achieve the right business result. Primarily this is the main reason for which we tend to formulate corporate data centers and to ensure data moves to the appropriate location. Every change in data structure of any single unit has impact on the whole architecture of it.
- **Understand Business Timing Needs** –the major contributing factor to any data process is the actual business activity for which it is required. Data is the most important asset owned by any company. It is IT's almost sacred duty to deliver the data where it is needed when it is needed [5].

Data may have many target systems, some need the data in real time others, like the BI system, only require periodic updates. Therefore the integration solution must be able to handle batch and real-time activity.

Integrate Master Data and Governance Rules -Where MDM solutions have been implemented then the MDM becomes the Centre of the hub for particular types of data. e.g., all customer data must be validated against the customer master. In this case customer data must be validated by the customer master before being forwarded to other systems that require the data. Thus data distribution can be a two-step process [5].

Technical- some of the most significant technical challenges of designing an application integration environment involve identifying the technical needs of your solution and determining the combination of protocols and services that will provide for those needs.

Organization issue – an application integration environment is available in multiple departments in an organization [12]. Staff in different department may

choose to deploy application that will need to integrate with your application integration environment.
There may be some kind of heterogeneity [11] which includes differences in-

1. Hardware and operating systems
2. Data management software
3. Data models, data semantics and middleware.

III. APPROACHES TO INTEGRATION

Integration by application
Process of bringing data from multiple application programs into an unified program. It is secured and orchestrated approach towards transferring & sharing of processes or data between different applications within the enterprises. This is applicable for small number of component systems and application become as the number of systems interfaces and data formats integrate [2].
Benefits
Application integration allows the application to introduced into the organization more efficiently for faster work and more accessibility at a lower cost. It allows you to modify the business processes as per the requirement of the organization. Providing so many channels for organization where they interact with each other and get integrated.
Technique
There are two models which increase the efficiency of application integration.

- **Point to point model** - It is decentralized structures in applications communicate directly with each other applications. This is most useful for organizations where they few applications with small number of sources.
- **Integration hub model** - In this an integration hub is placed between the applications at there each application is interact or communicate with it rather than communicating with each other. Application needs only a interface and connection to the integration hub, when a new app are introduced you do not need to rewrite the interface.

Requirement for application integration-
A common interface through which application can communicate with each other. There must be strong connectivity between the platforms to avoid any disturbances [2]. A common set of process and services rules should be used to insure consistency and reuse of integration services. Be capable of reusing the existing transport protocols that already exist in the enterprises.
Common data storage
Data integration can be done by transferring data into a new data storage. Usually it provide fast data access and easily understandable by users. If in some case local data sources are retired or damage, the application used by them is moved to new data storage .a common data storage have been refreshed periodically so that the local data sources remain functional and easily accessible for the user[1]. This type of integration is successful in organization as it increases accessibility among the people working in any organization.
Benefits-
As data is stored is stored at one place so it is easy for data mining. It reduces the cost and time to produce input formats provide the basis for consistent database reuse. It reduces the risk of errors caused by formats conversion failures. Data storage makes it easier to test and compare results of different federations.
Technique
Data warehousing is an approach towards realization of a common data storage and integration. Data is extracted, transformed, and loaded (ETL) into a data warehouse from several mostly heterogeneous operational sources. On this data different analysis tools could be implemented e.g. OLAP, OLTP.
Another common storage example is Operational Data Storage. In it "Warehouses with fresh data" is constructed by "Real Time" updates in local data sources to the data store. This further enables a real time data analysis process for decision support systems. Opposed to data warehouses, data here is neither cleansed nor aggregated.
Uniform data access
It is defined as connectivity and controllability across various data sources in this data from all the sources which are developed in different structures, schemas, and architecture are accumulated at one place which is treated as virtual data. Since it is a time consuming process so data access, homogenization and integration have to be done at the runtime [2]. Without it user face the difficulty to translate various data sources into the format supported by their application to provide a collective single view of data.
Benefits
It is ideal for application developers, software vendors and solution providers. Increase in the consistency in data accessing, accessing of become easier anywhere in the enterprise. Open to each and every person within an organization. Editing can be done to a data without interrupting the normal state of data. It can improve the use of all assets – hardware, software and people.
Techniques
In this process mediated query system is implemented as a solution to a single point for read only querying access to various data sources. These sends sub-queries to local data sources and are then combined at local query results.
P2P integration (peer to peer) –It is a decentralized approach where data is mutually shared and integrated at every peer location. This is entirely dependent on integration functionality available at all the peer location.
FDBMS- Federated database systems (FDBMS) try to achieve a uniform data access solution by integrating data from underlying local DBMS into a logical layer. Federated database systems are fully functional DBMS. They normally have their own data model, global transactions, global queries support as well as global access control. Usually, the five-level reference architecture is employed for building FDBM [15].
Portals are personalized access ways to an uniform data access for information on the internet or intranet where each user is provided with information tailored to his information needs. Usually, web mining is applied.

Manual integration
Here users are directly interacting with relevant information systems and integration can be done by selection of data from various data sources. As data has been developed at different structure and architecture it seems to be easy. We have to deal with different user

interfaces and query languages. Detailed knowledge on location of data is necessary as we have to take specific data for integration. A logical data representation needs to be very accurate while doing it. Logical data length can be 1, 2, 4 or 8 bytes in length. One another important thing Data semantics should be there means connection of database to the real world outside the database or mapping between an object modeled represented and/or stored in an information system.

Benefits

The basic benefit of this type of approach is its accuracy, and adjustability to any type of requirement. In any place where high resolution and accuracy of data is required or the data has an irregular pattern it is the most convenient and appropriate approach for data integration.

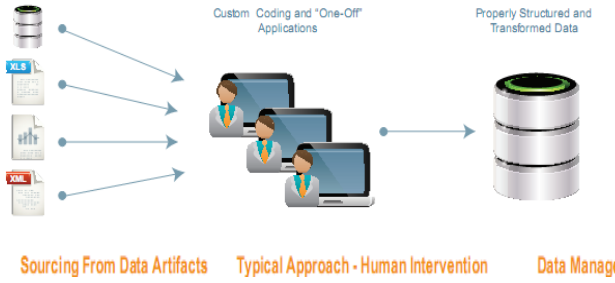

Fig 2 Manual Data Integration Process

Techniques

It requires writing of large numbers of programs/queries, or some existing programs are required to be modified for further customization of the program as the requirement may be. The most labor intensive part of manual integration is data mapping. In data mapping a point to point approach is required, which an operator has to manually find and map different data, before the integration could be possible. All this can lead to increased complexity with the addition of every new data source. On the downside it costly, time consuming, and error prone.

IV. CONCLUSIONS

In this paper we have looked into the most important issues and problems of data integration the industry faces today. Till today there is no one stop solution available in the market which can solve the entire range of problems. Each of the issues of data integration is unique in it and needs different and unique approach to solve it. But if we understand the problem properly and know how to tackle it, we can work on the development of a combining algorithm which can at least be able to solve most of the major issues if not all. This paper is an attempt to identify and bring together the issues, technique, and the benefits of it in one place.

ACKNOWLEDGMENT

We would like to thank to Dr. Ashish Bharadwaj, Chief Information Officer and Dr. Manish Prateek, Head of the Department, Centre for Information Technology, University of Petroleum & Energy Studies, Dehradun for extended support and encouragement for carrying out this work. We wish to special thanks to our department faculty members who help us to carry this work.

REFERENCES

[1] Maurizio Lenzerini, *Data Integration: A Theoretical Perspective Dipartimento di Informaticae Sistemistica*.

[2] Patrick Ziegler, Klaus R.Dittrich, *Data Integration - Problems, Approaches, and Perspectives*, Database Technology Research Group, Department of Informatics, University of Zurich.

[3] Khalid Saleem, Zohra Bellahsene, Large Scale Automatic Schema Matching, Category of submission: Survey Paper, University Montpellier.

[4] Cortney Claiborne, Darren Cunningham, Davythe Dicochea, Erin O'Malley, Philip On, Business Objects - Data Integration: The Key to Effective Decisions.

[5] NesimeTatbul, Streaming Data Integration: Challenges and Opportunities, ETH Zurich, Switzerland.

[6] A Focus on XSLT 2.0: Understanding the Development and Business, Benefits." October 2004.InfoTrends/CAP Ventures Analysis report, (http://www.capv.com/ home/Downloads/ 10. 22.04.pdf)

[7] Aumueller,D.,Do,H.H.,Massmann ,S.,& Rahm, E. (2005). Schema matching with coma++, Acm sigmod (p.906-908).

[8] Batini,C.,Lenzerini,M., Navathe,S.B.(1986). A comparative analysis of, Methodologies for database schema integration, ACM Computing, Surveys, 18(4),323-364.

[9] Bernstein,P.A., Melnik,S.,&Churchill, J.E.(2006). Incremental schema, matching, VLDB.

[10] Gal,A.(2006).Managing uncertainty in schema matching with top-k schema, mappings, JoDS, 90-114.

[11] Williamw.Cohen, Data Integration Using Similarity Joins and,A Word-Based Information Representation, Language, AT&T Labs—Research , Shannon Laboratory.

[12] Cohen, W.W.1998a. Integration of heterogeneous databases without common domains,Using queries based on textual similarit. ProceedingsofACMSIGMOD-98 (Seattle,WA,1998).

[13] Hernandez,m. and Stolfo,s.1995. The merge/purge problem for large databases. Proceedings of the 1995 ACM SIGMOD (May1995).

[14] Ouksel, ArisM.andSheth, AmitP. (1999). Semantic Interoperability in Global Information Systems:ABrief Introduction to the Research Area and the Special Section. SIGMOD, Record, 28(1):5–12.

[15] Sheth, AmitP, Gala, Sunit K., and Navathe, Shamkant B. (1993). On Automatic Reasoning For Schema Integration. International Journal of Intelligent and Cooperative Information Systems,2(1):23–50.

[16] Sheth, Amit P and Larson, James A. (1990). Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases. ACM Computing Surveys, 22(3):183–236.

[17] Arens,Yigal, etal.(1993). Retrieving and Integrating Data from Multiple Information Sources .International Journal of Cooperative Information Systems (IJCIS),2(2):127–158.

[18] J.Linand A.O.Mendelzon, Merging databases under constraints. International Journal of cooperative Information Systems,7(1):55–76,1998

[19] S.Babuand J.Wisdom,"Continuous Queries over Data Streams," ACM SIGMOD Record, vol.30, no.3, September2001.

[20] S.Babuand J.Widom,"Continuous Queries over Data Streams," ACM SIGMOD Record, vol.30,no.3,September2001.

[21] Stream Base, "Stream Base Systems, Inc. "http://www.streambase.com/

[22] Hurson, AliR.and Bright,M.W.(1991). Multi database Systems: An Advanced Concept in Handling Distributed Data. AdvancesinComputers,32:149–200.

[23] S.Abiteboul, D.Quass ,J.McHugh, J.Widom,and J.L.Wiener. The Lore query language for semi structured data. Int.J.onDigitalLibraries,1(1):68–88,1997.

[24] A.Cal`ı,D.Calvanese ,G.DeGiacomo, and M.Lenzerini. On the expressive power of data integration systems.Submittedforpubblication,2002.

