Binary Journal of Data Mining & Networking 4 (2014) 06-09

http://www.arjournals.org/index.php/bjdmn/index

Research Article

ISSN: 2229 -7170

A Survey on Data Integration in Data Warehouse

Geetanjli Khambra[1] , Pankaj Richhariya[1]

*Corresponding author:

Geetanjli Khambra

[1]Computer Science and Engineering, BITS, Bhopal, India

Abstract

Data warehousing embraces technology of integrating data from multiple distributed data sources and using that at an in annotated and aggregated form to support business decision-making and enterprise management. Although many techniques have been revisited or newly developed in the context of data warehouses, such as view maintenance and OLAP, little attention has been paid to data mining techniques for supporting the most important and costly tasks of data integration for data warehouse design.

Keywords: Data warehousing; business intelligence; data integration; Redundancy; online-transaction processing.

Introduction

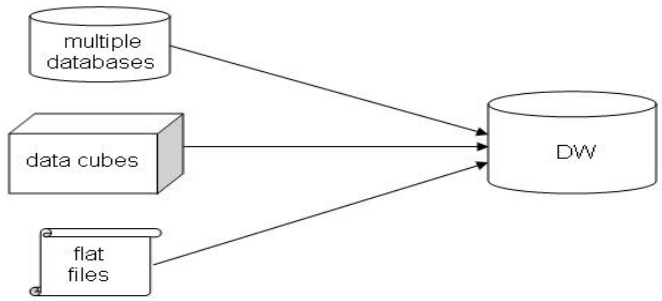
Since the past decade data warehouses have been gaining enormous ground in the business intelligence (BI) domain. A corporate data warehouse was on every organization's priority list. Companies began to rely more and more on these BI systems. Critical business decisions were based on the current and historical data available in the data warehouse.

Data warehouse (DW) is a system that extracts, cleans, conforms, and delivers source data into a dimensional data store and then supports and implements querying and analysis for the purpose of decision making. For business leaders a corporate data warehouse and then consequently data mining seemed to the long term strategy. Sophisticated OLAP tools, which facilitate multidimensional analysis, were used. Business trends are identified using data mining (DM) tools and applying complex business models. As businesses grow from local to global, the complexities and parameters involved in decision making and analysis became more complex. The most visible part of a data warehouse project is the data access portion—usually in the form of products and some attention is brought to the dimensional model. But by spotlighting only those portions, a gaping hole is left out of the data warehouse lifecycle. When it comes time to make the data warehouse a reality, the data access tool can be in place, and the dimensional model can be created, but then it takes many months from that point until the data warehouse is actually usable because the ETL process (extraction, transformation, loading) still needs to be completed. Data warehousing is the process of taking data from legacy and transaction database systems and transforming it into organized information in a user-friendly format to encourage data analysis and support fact-based business decision making. The process that involves transforming data from its original format to a dimensional data store accounts for at least 70 percent of the time, effort, and expense of most data warehouse projects. As it is very costly and critical part of a data warehouse implementation there is a variety of data extraction and data cleaning tools, and load and refresh utilities for DW. Here we present another point of view to this problem—using data mining techniques to facilitate the integration of data DW.

Data integration involves combining data residing in different sources and providing users with a unified view of these data. This process becomes significant in a variety of situations both commercial (when two similar companies need to merge their databases) and scientific (combining research results from different bioinformatics repositories, for example). Data integration appears with increasing frequency as the volume and the need to share existing data explodes. It has become the focus of extensive theoretical work, and numerous open problems remain unsolved. In management circles, people frequently refer to data integration as "Enterprise Information Integration" (EII).

Needs of Data Integration

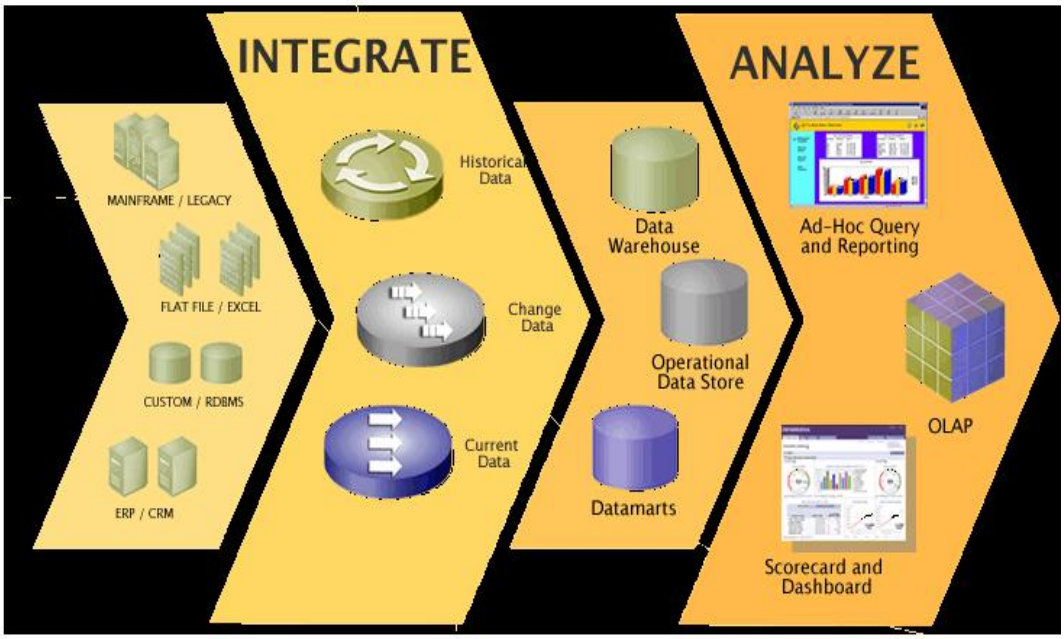
The data integration is a useful preprocessing step in which we include data from multiple sources in our analysis. This would involve integrating multiple databases, data cubes, or files. Yet some attributes representing a given concept may have different databases, causing inconsistencies and redundancies. The basic process views of data integration are:



Data Integration Process

The data integration is one of the most important characteristic of the data warehouse. Data is fed from multiple disparate sources into the data warehouse. As the data is fed it is converted, reformatted, summarized, and so forth. The result is that data—once it resides in the data warehouse—has a single physical corporate image. Many problems arise in this process. Designers of different applications made up their decisions over the years in different ways. In the past, when application designers built an application, they never considered that the data they were operating on would ever have to be integrated with other data. Such a consideration was only a wild theory. Consequently, across multiple applications there is no application consistency in encoding, naming conventions, physical attributes, measurement of attributes, and so forth. Each application designer has had free rein to make his or her own design decisions. The result is that any application is very different from any other application.

One simple example of lack of integration is data that is not encoded consistently, as shown by the encoding of gender. In one application, gender is encoded as m or f. In another, it is encoded as 0 or 1. As data passes to the data warehouse, the applications' different values must be correctly deciphered and recoded with the proper value. This consideration of consistency applies to all application design issues, such as naming conventions, key structure, measurement of attributes, and physical characteristics of data. Some of the same data exists in various places with different names, some data is labeled the same way in different places, some data is all in the same place with the same name but reflects a different measurement, and so on. Whatever integration architecture we choose, there are different problems that come up when trying to integrate data from various sources (see figure).



Core Data Integration is the use of data integration technology for a significant, centrally planned and managed IT initiative within a company. Examples of core data integration initiatives could include:

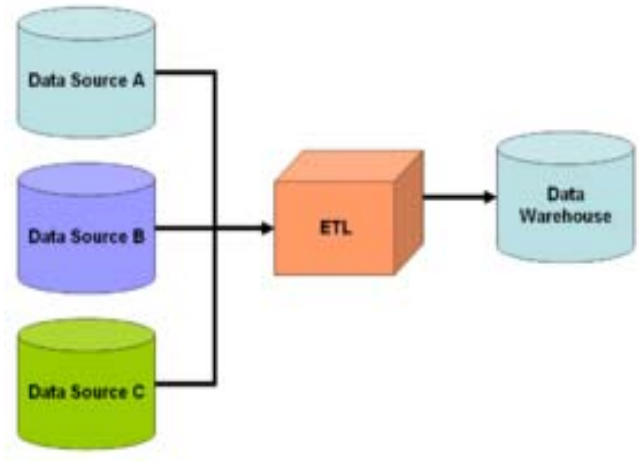
ETL (Extract, transform, load) implementations.
EAI (Enterprise Application Integration) implementations.
SOA (Service-Oriented Architecture) implementations.
ESB (Enterprise Service Bus) implementations.
Core data integrations are often designed to be enterprise-wide integration solutions. They may be designed to provide a data abstraction layer, which in turn will be used by individual core data integration implementations, such as ETL servers or applications integrated through EAI.

Because it is difficult to promptly roll out a centrally managed data integration solution that anticipates and meets all data integration requirements across an organization, IT engineers and even business users create edge data integration, using technology that may be incompatible with that used at the core. In contrast to core data integration, edge data integration is not centrally planned and is generally completed with a smaller budget and a tighter deadline.

Issues in Data Integration

Data analysis task will involve data integration, which combines data from multiple sources into a coherent data store as in data warehousing. These sources may include multiple databases, data cubes, or flat files. Issues with combining heterogeneous data sources under a single query interface have existed for some time. The rapid adoption of databases after the 1960s naturally led to the

need to share or to merge existing repositories. This merging can take place at several levels in the database architecture. One popular solution involves data warehousing (see figure).



The warehouse system extracts, transforms, and loads data from heterogeneous sources into a single common queriable schema so data becomes compatible with each other. The ETL process extracts information from the source databases, transforms it and then loads it into the data warehouse. There are number of issues to consider during data integration:

Schema Integration

The most important issue in data integration is the Schema integration. How can equivalent real-world entities from multiple data sources be matched up? This is referred to as entity identification process. Terms may be given different interpretations at different sources. For example, how can be data analyst be sure that customer_id in one database and cust_number in another refer the same entity? Data mining algorithms can be used to discover the implicit information about the semantics of the data structures of the information sources. Often, the exact meaning of an attribute cannot be deduced from its name and data type. The task of reconstructing the meaning of attributes would be optimally supported by dependency modeling using data mining techniques and mapping this model against expert knowledge, e.g., business models. Association rules are suited for this purpose. Other data mining techniques, e.g., classification tree and rule induction, and statistical methods, e.g., multivariate regression, probabilistic networks, can also produce useful hypotheses in this context. Many attribute values are (numerically) encoded. Identifying inter-field dependencies helps to build hypotheses about encoding schemes when the semantics of some fields are known. Also encoding schemes change over. Data mining algorithms are useful to identify changes in encoding schemes, the time when they took place, and the part of the code that is affected. Methods which use data sets to train a “normal” behavior can be adapted to the task. The model learned can be used to evaluate significant changes. A further approach would be to partition the data set, to build models on these partitions applying the same data mining algorithms, and to compare the differences between these models. Data mining and statistical methods can be used to induce integrity constraint candidates from the data. These include, for example, visualization methods to identify distributions for finding domains of attributes or methods for dependency modeling. Other data mining methods can find intervals of attribute values, which are rather compact and cover a high percentage of the existing values. Once each single data source is understood, content and structural integration follows. This step involves resolving different kinds of structural and semantic conflicts. To a certain degree, data mining methods can be used to identify and resolve these conflicts. Data mining methods can discover functional relationships between different databases when they are not too complex. A linear regression method would discover the corresponding conversion factors. If the type of functional dependency (linear, quadratic, exponential etc.) is a priori not known, model search instead of parameter search has to be applied.

Data Redundancy

Redundancy is another important issue. An attributes may be redundant if it can be derived from another attribute or set of attributes. Inconsistencies in attribute or dimension naming can also cause redundancies in the resulting data set. In non-database systems each application has its own private files. This can often lead to redundancy in stored data, with resultant waste in storage space. In a database the data is integrated. The database may be thought of as a unification of several otherwise distinct data files, with any redundancy among those files partially or wholly eliminated. Data integration is generally regarded as an important characteristic of a database. The avoidance of redundancy should be an aim, however, the vigor with which this aim should be pursued is open to question.

Redundancy is

Direct if a value is a copy of another.

Indirect if the value can be derived from other values:

simplifies retrieval but complicates update

conversely integration makes retrieval slow and updates easier

Data redundancy can lead to inconsistency in the database unless controlled.

The system should be aware of any data duplication - the system is responsible for ensuring updates are carried out correctly.

A database with uncontrolled redundancy can be in an inconsistent state - it can supply incorrect or conflicting information.

A given fact represented by a single entry cannot result in inconsistency - few systems are capable of propagating updates i.e. most systems do not support controlled redundancy.

Detection

A third major issue in data integration is the detection and resolution of data value conflicts. For example, for the same real-world entity, attribute value from different sources may differ. This may be due to differences in representation, scaling, or encoding.

For instance, a weight attribute may be stored in metric units in one system and British imperial units in another. An attributes in one system may be recorded at a lower level of abstraction than the same attributes in another.

OLTP & OLAP

The major task of online operational database system is to perform online transaction and query processing; these systems are called online-transaction processing (OLTP) systems. They cover most of the day-to-day operations of an organization such as purchasing, inventory, manufacturing, banking, payroll, registration and accounting. Data warehouse systems, on the other hand serve users or knowledge workers in the role of data analysis and decision making, such systems can organize and present data in various formats in order to accommodate the diverse needs of the different users; these systems are known as online-analytical processing (OLAP) systems. These both systems play important role in data integration process for covering in huge area of data mining fields.

Database technology is at the center of many types of information systems. Among these information systems are decision support systems and executive information systems. Online Transaction Processing (OLTP) environments use database technology to transact and query against data, and support the daily operational needs of the business enterprise. Online Analytical Processing (OLAP) environments use database technology to support analysis and mining of long data horizons, and to provide decision makers with a platform from which to generate decision making information. Decision support systems and business intelligence tools, as well as data mining analytics, generally utilize a different data structure to do their work. One of the main disadvantages to the OLAP environment is the concurrency of the data. Because the process by which data is extracted, transformed, and loaded into the OLAP environment can be relatively slow by transactional data standards, the ability to achieve "real-time" data analysis is lost. The importance of generating real-time business intelligence is that it is a building block to achieve better business process management and true business process optimization. OLTP database structures are characterized by storage of "atomic" values, meaning individual fields cannot store multi-values. They are also transaction oriented as the name implies. Alternatively, OLAP database structures are generally aggregated and summarized. There is an analytical orientation to the nature of the data, and the values represent a historical view of the entity.

Data Integration In The Life Sciences

Large-scale questions in science, such as global warming, invasive species spread, and resource depletion, are increasingly requiring the collection of disparate data sets for meta-analysis. This type of data integration is especially challenging for ecological and environmental data because metadata standards are not agreed upon and there are many different data types produced in these fields. National Science Foundation initiatives such as Data-Net are intended to make data integration easier for scientists by providing cyber infrastructure and setting standards. The two funded Data-Net initiatives are Data One and the Data Conservancy.

Conclusion

Data integration has attracted many diverse and diverging contributions. The purpose, and the main intended contribution of this paper is to provide a clear picture of what are the process of data integration, needs of data integration, core data integration and the major issues of data integration and also using in online database systems and life science.

In this short survey, several important problems remain to be investigated; examples are integration of complex objects, integration of integrity constraints and methods, integration of heterogeneous data. Theoretical work is needed to assess integration rules and their behavior. It is therefore important that the effort to survey on data integration and solve data integration issues be continued and that proposed newly concepts and methodologies be evaluated through experiments with real applications.

References

[1]. http://www.nbu.bg/PUBLIC/IMAGES/File/departamenti/informatika/17.pdf

[2]. http://academic.regis.edu/cias/Library/pid54211.pdf

[3]. http://db.grussell.org/section002.html

[4]. http://www.infoalchemy.com/images/diagram.gif

[5]. Patrick Ziegler and Klaus R. Dittrich (2004). "Three Decades of Data Integration - All Problems Solved?". WCC 2004. pp. 3–12.

[6]. Maurizio Lenzerini (2002). "Data Integration: A Theoretical Perspective". PODS 2002. 233–246.

[7]. Motro A. Superviews: Virtual integration of multiple databases. IEEE Trans. Softw. Eng. 1978;13, 7:785–798.

[8]. Sheth A., Larson J. Federated database systems for managing distributed, heterogeneous, and autonomous databases. ACM Comput. Sur. 1990;22, 3:183–236.

[9]. Dupont Y. Resolving Fragmentation Conflicts in Schema Integration, In Entity-Relationship Approach - ER'94. P. Loucopoulos Ed. LNCS 881, Springer-Verlag, Germany, 1994, 513–532.

[10]. Y. Wilks. Information extraction as a core language technology. In M-T. Pazienza, editor, *Information Extraction*. Springer, Berlin, 1997.

[11]. A. Spoerri. *InfoCrystal: A Visual Tool for Information Retrieval*. PhD thesis,

Massachusetts Institute of Technology, Cambridge, MA, 1995.

[12]. D. G. Roussinov and H. Chen. Information navigation on the web by clustering and summarizing query results. *Information Processing & Management*, 37(6):789–816, 2001.

[13]. Zamir, O., Etzioni, O., Madani, O., and Karp, R. in Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, 1997.

[14]. Ramesh, "Knowledge Mining of Test Case System," *International Journal on Computer Science and Engineering* 2009;2(1),69-73.

[15]. Mark Last and Menahem Friedman."The Data Mining approach to automated software testing.".

[16]. Data Mining: Adriaans & Zantinge: Pearson education.

[17]. Mastering Data Mining: Berry Linoff: Wiley.

[18]. Data Mining: Dunham : Pearson education.

[19]. Dr. Gary Parker, vol 7, 2004, Data Mining: Modules in emerging fields, CD-ROM.

[20]. Crisp-DM 1.0 Step by step Data Mining guide from http://www.crisp-dm.org/CRISPWP-0800.pdf.

[21]. Jiawei Han and Micheline Kamber (2006), Data Mining Concepts and Techniques, published by Morgan Kauffman, 2nd ed.