

This research examines the following questions:

- Why integrate multiple data sources?
- What are the benefits of integrating multiple data sources for enterprise-level planning and decision making?

Accessing to the accurate information in a timely manner is a significant challenge facing organizations today. For example, a police officer needs to know if a suspect is wanted in another jurisdiction; A social worker needs to ensure that a welfare applicant is not already receiving benefits elsewhere; A judge needs to see all prior convictions against an offender (McKenna, 1996). These and countless other situations require rapid access to a wide range of complete and accurate information that is often scattered across numerous agencies (McKenna, 1996). However, the problem is that many agencies build information "silos" which are poorly accessible within their own organizations, let alone to the related departments outside the organization. Besides, many agencies seem to have a "genetically encoded political and cultural aversion to information sharing and cooperation, operating instead as isolated fiefdoms or, at best, as grudging partners" (McKenna, 1996).

There has been a spectacular explosion in the quantity of data available in electronic formats in the past few decades. This huge amount of data has been gathered, organized, and stored by a small number of individuals, working for different organizations on varied problems (Subrahmanian, et al., 1996). In light of the ever increasing volume of data, and the expected benefits of integrating the data, a framework for performing integration over multiple data sources is necessary.

What is Data Integration?

Data integration is the process of the standardization of data definitions and data structures by using a common conceptual schema across a collection of data sources (Heimbigner and McLeod, 1985; Litwin, et al., 1990). Integrated data will be consistent and logically compatible in different systems or databases, and can use across time and users (Martin, 1986).

Goodhue et al. (1992, p294) defined data integration as "the use of common field definitions and codes across different parts of an organization". According to Goodhue, et al. (1992), data integration will increase along one or both of two dimensions: (1) the number of fields with common definitions and codes, or (2) the number of systems or databases adhering to these standards. Data integration is an example of a highly formalized language for describing the events occurring in an organization's domain. The scope of data integration is the extent to which that formal language is used across multiple organizations or sub-units of the same organization. The objective of data integration is to bring together data from multiple data sources that have relevant information contributing to the achievement of the users' goals (AFT, 1997).

The Advanced Forest Technologies in Canada (AFT, 1997) identified the following factors which must be addressed to integrate data properly:

- identification of an optimal subset of the available data sources for integration
- estimation of the levels of noise and distortions due to sensory, processing, and environmental conditions when the data are collected
- the spatial resolution, the spectral resolution, and the accuracy of the data
- the formats of the data, the archive systems, and the data storage and retrieval
- the computational efficiency of the integrated data sets to achieve the goals of the users

Benefits of integrating heterogeneous data sources

There are some obvious advantages in integrating information from multiple data sources. Such integration alleviates the burden of duplicating data gathering efforts, and enables the extraction of information that would otherwise be impossible (Subrahmanian, et al., 1996).

Subrahmanian, et al. (1996) gives the following examples of benefits of data integration:

- "... law enforcement agencies such as Interpol benefit from the ability to access databases of various national police forces, to assist their effort in fighting international terrorism, drug trafficking, and other criminal activities. Insurance companies, using data from external sources, including other insurance

company and police records, can identify possible fraudulent claims. Medical researchers and epidemiologists, with access to records across geographical and ethnic boundaries, are in a better position to predict the progression of certain diseases. In each case, the information extracted from the integrated sources is not possible when the data sources are viewed in isolation."

Integrating diverse data source paradigms

Subrahmanian, et al. (1996) established a data source paradigm. There are two important aspects to constructing the data source paradigm: domain integration and semantic integration. Domain integration is the physical linking of data sources and systems, while semantic integration is the coherent extraction and combination of the information provided by the data and reasoning sources, to support a specific purpose (Subrahmanian, et al., 1996).

It is acknowledged that data warehousing is the most effective way to provide the business decision support data (Van Den Hoven, 1998). Under this concept, data is derived from operational systems and external information providers, and subsequently conditioned, integrated, and changed into a read-only database that is optimized for direct access by decision makers. The term 'data warehousing' describes data as an enterprise asset that must be identified, cataloged, and stored to ensure that users will always be able to find the needed information. The data warehouse is generally enterprise-wide in scope, and its purpose is to provide a single, integrated view of the enterprise's data, spanning all enterprise activities (Van Den Hoven, 1998).

This paper identifies and compares the issues, methods, and results of efforts that involve integrating different data sources 1) within one organization, and 2) across multiple organizations.