Chapter 4
Semantic Interation, Text Mining, Tools and Technologies

Chandrakant Ekkirala
Cognizant Technologies Limited, India

ABSTRACT

Semantic technologies have gained prominence over the last several years. Semantic technologies are explored in detail and semantic integration of data will be outlined. The various data integration techniques and approaches will also be touched upon. Text Mining, different associated algorithms and the various tools and technologies used in text mining will be enumerated in detail. The chapter will have the following sections – 1. Data Integration Techniques • Data Integration Technique – Extraction, Transformation and Loading (ETL) • Data Integration Technique – Data Federation 2. Data Integration Approaches • Need Based Data Integration • Periodic Data Integration • Continuous Data Integration 3. Semantic Integration 4. Semantic Technologies 5. Semantic Web Technologies 6. Text Mining 7. Text Mining Algorithms 8. Tools and Technologies for Text Mining

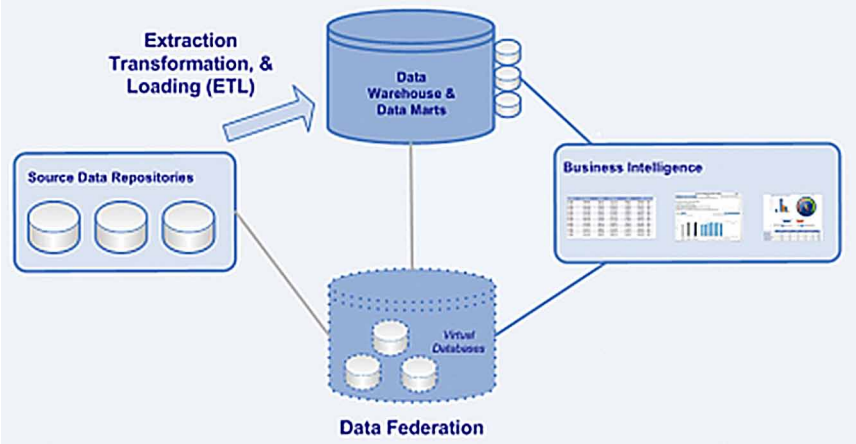
INTRODUCTION

Data Integration Techniques

Data integration is a fundamental, yet deceptively challenging, component of any organization's business intelligence and data warehousing strategy. Data integration involves combining data residing in different data repositories and providing business users with a unified view of this data. In addition, companies face a challenge of ensuring that data being reported is current and up-to-date. Companies are now increasingly incorporating both traditional batch-oriented techniques for query performance and real-time data integration to eliminate the annoyance of out-of-date data. The top batch-oriented technique that companies utilize is known as ETL while one of the popular real-time techniques is known as Data Federation.

DOI: 10.4018/978-1-4666-8726-4.ch004

Figure 1. Data Integration Techniques – ETL and Data Federation



Data Integration Technique: Extraction, Transformation and Loading (ETL)

The term ETL which stands for extraction, transformation, & loading is a batch or scheduled data integration processes that includes extracting data from their operational or external data sources, transforming the data into an appropriate format, and loading the data into a data warehouse repository. ETL enables physical movement of data from source to target data repository. The first step, extraction, is to collect or grab data from its source(s). The second step, transformation, is to convert, reformat, cleanse data into format that can be used be the target database. Finally the last step, loading, is import the transformed data into a target database, data warehouse, or a data mart. A data warehouse holds very detailed information with multiple subject areas and works towards integrating all the date sources. A data mart usually holds more summarized data and often holds only one subject area.

ETL Step 1: Extraction

The extraction step of an ETL process involves connecting to the source systems, and both selecting and collecting the necessary data needed for analytical processing within the data warehouse or data mart. Usually data is consolidated from numerous, disparate source systems that may store the date in a different format. Thus the extraction process must convert the data into a format suitable for transformation processing. The complexity of the extraction process may vary and it depends on the type and amount of source data.

ETL Step 2: Transformation

The transformation step of an ETL process involves execution of a series of rules or functions to the extracted data to convert it to standard format. It includes validation of records and their rejection if they are not acceptable. The amount of manipulation needed for transformation process depends on the data. Good data sources will require little transformation, whereas others may require one or more transformation techniques to to meet the business and technical requirements of the target database or