Chapter 1

AN OVERVIEW OF DATA WAREHOUSE, OLAP AND DATA MINING TECHNOLOGY

Anoop Singhal

Abstract: In this chapter, a summary of Data Warehousing, OLAP and Data Mining Technology is provided. The technology to build Data Analysis Application for Network/Web services is also described.

Key words: STAR Schema, Indexing, Association Analysis, Clustering

1. MOTIVATION FOR A DATA WAREHOUSE

Data warehousing (DW) encompasses algorithms and tools for bringing together data from distributed information repositories into a single repository that can be suitable for data analysis [13]. Recent progress in scientific and engineering applications has accumulated huge volumes of data. The fast growing, tremendous amount of data, collected and stored in large databases has far exceeded our human ability to comprehend it without proper tools. It is estimated that the total database size for a retail store chain such as Walmart will exceed 1 Petabyte (1K Terabyte) by 2005. Similarly, the scope, coverage and volume of digital geographic data sets and multidimensional data has grown rapidly in recent years. These data sets include digital data of all sorts created and disseminated by government and private agencies on land use, climate data and vast amounts of data acquired through remote sensing systems and other monitoring devices [16], [18]. It is estimated that multimedia data is growing at about 70% per year. Therefore, there is a critical need of data analysis systems that can automatically
analyze the data, to summarize it and predict future trends. Data warehousing is a necessary technology for collecting information from distributed databases and then performing data analysis [1], [2], [3], and [4].

Data warehousing is an enabling technology for data analysis applications in the area of retail, finance, telecommunication/Web services and bio-informatics. For example, a retail store chain such as Walmart is interested in integrating data from its inventory database, sales database from different stores in different locations, and its promotions from various departments. The store chain executives could then 1) determine how sales trend differ across regions of the country 2) correlate its inventory with current sales and ensure that each store’s inventory is replaced to keep up with the sales 3) analyze which promotions are leading to increases product sales. Data warehousing can also be used in telecommunication/Web services applications for collecting the usage information and then identify usage patterns, catch fraudulent activities, make better use of resources and improve the quality of service. In the area of bio-informatics, the integration of distributed genome databases becomes an important task for systematic and coordinated analysis of DNA databases. Data warehousing techniques will help in integration of genetic data and construction of data warehouses for genetic data analysis. Therefore, analytical processing that involves complex data analysis (usually termed as decision support) is one of the primary uses of data warehouses [14].

The commercial benefit of Data Warehousing is to provide tools for business executives to systematically organize, understand and use the data for strategic decisions. In this paper, we motivate the concept of a data warehouse, provide a general architecture of data warehouse and data mining systems, discuss some of the research issues and provide information on commercial systems and tools that are available in the market.

Some of the key features of a data warehouse (DW) are as follows.

1. Subject Oriented: The data in a data warehouse is organized around major subjects such as customer, supplier and sales. It focuses on modeling data for decision making.
2. Integration: It is constructed by integrating multiple heterogeneous sources such as RDBMS, flat files and OLTP records.
3. Time Variant: Data is stored to provide information from a historical perspective.