IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Hershey PA 17033-1117, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

#ITB7888

CHAPTER TWELVE

From Web Log to Data Warehouse: An Evolving Example

John M. Artz
George Washington University, USA

INTRODUCTION

Data warehousing is an emerging technology that greatly extends the capabilities of relational databases specifically in the analysis of very large sets of time-oriented data. The emergence of data warehousing has been somewhat eclipsed by the simultaneous emergence of Web technologies. However, Web technologies and data warehousing have some natural synergies that are just now being recognized. First, Web technologies make data warehouse data more easily available to a much wider variety of users both internally and externally. Since the value of data is directly related to its availability for exploitation, Internets and intranets help increase the value of the data in the warehouse. Second, data warehouse technologies can be used to analyze traffic to a Web site in a wide variety of ways in order to make the Web site more effective. This chapter will focus on the latter of these synergies and show, through an evolving example, how a simple data set from the Web log can be enhanced, in a step-wise fashion, into a full-fledged market research data warehouse.

BACKGROUND

For people who are not intimately familiar with data warehousing technologies, the first questions that arise are—what is a data warehouse and how does it differ from a traditional relational database? The standard definition for a data warehouse is "A data warehouse is a subject-oriented, integrated, nonvolatile and time variant collection of data in support of management decisions" (Inmon, 1996). This definition provides a good starting point, and closer inspection of the compo-

This chapter appears in the book, *Managing Internet and Intranet Technologies in Organizations: Challenges and Opportunities* by Subhasish Dasgupta. Copyright 2001, Idea Group Inc.

nents of this definition does provide some insight into the nature of a data warehouse. First, a subject-oriented database is a database that is organized around subjects of interest to facilitate information exploitation rather than processing. This means that, by nature, the purpose of the data warehouse is to deliver information rather than to support processing. Integrated means that all data is accessible through the same interface and that underlying data sets can be linked using common keys. To say that the data warehouse is integrated is to say that users of the data warehouse do not need to be aware of the technologies used to manage the underlying data nor do they need to be aware of peculiarities in data set design when performing analysis across data sets.

Time varying is the most important characteristic that distinguishes a data warehouse from a traditional relational database. A traditional relational database represents the state of an organization at a point in time. Kimball (1986) calls this a "twinkling database" because records come and go but do not reveal a pattern over time. A data warehouse, on the other hand, represents that organization as it changes over time. To put this more precisely, a relational database is a snapshot of the organization at a point in time whereas the data warehouse is a collection of longitudinal data.

Since the data is time oriented it is also nonvolatile. The facts that were true last month remain true for last month. If they change this month then they are new facts for the current month. The only reason you would change historical data is if you were to discover that a fact from a previous point in time was recorded incorrectly. So the data warehouse accepts new data on some periodic basis, but existing data should not change. Finally, the data warehouse is used primarily for organizational decision making which means that it is not operational data, nor is it management information. It is data that is used primarily for forecasting and decision support. This means that the typical usage of the data will be at a much higher level of summary than transaction-level data.

Having given the broad view of the data warehouse, it is useful to take a few steps to further refine this perspective. The unit of decomposition in a relational database is the entity, which is generally defined as a thing of interest to the organization. Entities typically include customers, products, employees and the like. The purpose of a relational database is to maintain the current state of the entities—how many are there, and what facts are true about them at the current moment. Typical questions that one might ask of a relational database are—how many employees are there in the marketing department, which department has the most employees or which department has the highest average salary. These questions all address the state of the organization at a point in time.

The unit of decomposition for the data warehouse is a fact that represents a measure of a key business process. Units sold, dollars sold and gross margin are typical measures for some business processes although numbers of visitors, or time spent viewing a Web page are also useful measures for others. While the relational database tracks the status of entities as indicated by the values of attributes, the data