Web Technologies and Data Warehousing Synergies

John M. Artz
The 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 over the past decade by the simultaneous emergence of Web technologies. However, Web technologies and data warehousing have some natural synergies that are not immediately obvious. First, Web technologies make data warehouse data more easily available to a much wider variety of users. Second, data warehouse technologies can be used to analyze traffic to a Web site in order to gain a much better understanding of the visitors to the Web site. It is this second synergy that is the focus of this article.

DATA WAREHOUSE

A data warehouse is a repository of nonvolatile temporal data used in the analysis and tracking of key business processes. Temporal or time varying is the most important characteristic that distinguishes a data warehouse from a traditional relational database, which represents the state of an organization at a point in time. A relational database is a snapshot of the organization, whereas the data warehouse is a collection of longitudinal data.

One could argue that it should be possible to store longitudinal data in a relational database, and this claim is true. However, relational databases, which model data as entities, create sever limitations in data exploitation. First, although standard SQL does provide a DateTime data type, it is very limited in its handling of dates and times. If an analyst wanted to compare summary data on weekends versus weekdays or holidays versus non-holidays, it would be difficult if not impossible using standard SQL. Second, analysis involving drill down or roll up operations becomes extremely awkward using standard SQL against entities as represented in relational tables.

Data warehousing technology overcomes these deficiencies in the relational model by representing data in a dimensional model. A dimensional model consists of a fact table (see Figure 1) and the associated dimensions. The fact table contains measures of the business process being tracked and the dimensional tables contain information on factors that may influence those measures. More specifically, the fact table contains dependent variables while the dimension tables contain independent variables. Online analytical processing (OLAP) tools provide a means of summarizing the measures in the fact table according to the dimensions provided in the dimension table toward the end of determining what factors influence the business process being modeled. Typically OLAP tools provide a means of easily producing higher levels of summary (roll-up) or greater levels of detail (drill-down).

WEB LOG

A visitor to a Web site requests a page by typing in the address of the page in a Web browser, or by clicking on a link that automatically requests that page. A message is

Figure 1. A dimensional model

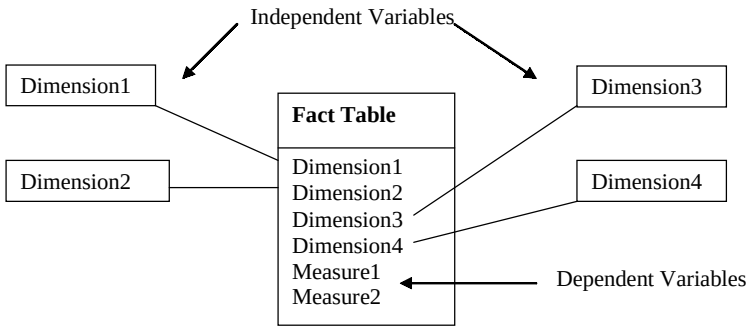
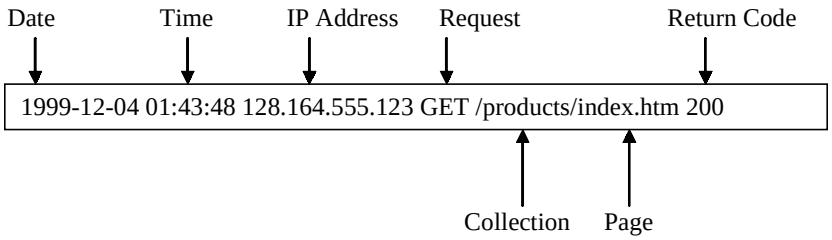


Figure 2. A typical Web log record



sent to the Web server, at that address, and the Web server responds by returning the page. Each request that is processed by the Web server is recorded in a file called the Web log, which contains a record of all activity on the Web site. The record typically contains the date and time, the IP address of the requestor, the page requested and the result. A typical record in standard format is shown in Figure 2.

From this simple log record we can determine quite a bit about the traffic coming to the Web site. For example, we can determine peak times for visitors by date or time and we can determine if Web site usage is cyclical or has other temporal patterns. Further, we can determine which pages or collections are most heavily visited and if their usage also reflects a temporal pattern. Answers to these questions are useful for site management and maintenance and for determining the effectiveness of design decisions or the behavior of the visitors to the site.

SYNERGY

A lot of valuable information can be derived from the Web log. But that is only the beginning. By viewing the Web log as a data source for a data warehouse, it becomes an even richer source of information about the Web site and its visitors. Consider the dimensional model in Figure 3. From this we can derive relationships between visitors and the pages they visit. Web log records can be summarized to produce dwell time (the time a visitor spends viewing a page) and visit length (the time a visitor spends at the site). Further, if the site is used for sales, the IP dimension can be converted to a customer dimension and the page dimension can be converted into a product dimension allowing analysis of customer purchasing behavior. For a more in depth examination of the process of evolving a Web log into a customer data warehouse, see *From Web Log to Data Warehouse: An Evolving Example,* listed in the references.

FUTURE TRENDS

According to information navigators, there are approximately 72 million hosts on the Web. Many of these hosts do not support Web sites, but many others (such as those owned by ISPs) have multiple Web sites. If only 10 million of these Web sites attract as few as 100 visitors a minute, then 1 billion records are generated every minute. This becomes 60 billion records per hour, or 1.4 trillion records per day. This is an enormous amount of information providing great insight into the behavior of consumers and information seekers, among others. This huge volume of temporal data cannot be exploited effectively without data warehouse technology. Hence, the growth of the Web may well push the growth of data warehousing.

At the same time, data warehousing concepts continue to evolve and the technology continues to improve.

Figure 3. A dimensional model based on the Web log