Data Warehousing, Multi-Dimensional Data Models and OLAP

Prasad M. Deshpande
IBM Almaden Research Center, USA

Karthikeyan Ramasamy
Juniper Networks, USA

INTRODUCTION

Since the advent of information technology, businesses have been collecting vast amounts of data about their daily transactions. For example, a company keeps track of data regarding the sales of its various products at different stores over a period of time. Businesses can gain valuable insights by analyzing this data to spot trends and correlations in the data. Data warehousing, multidimensional analysis, and online analytical processing (OLAP) refer to a set of technologies that address the problem of business data analysis. Data warehousing has become the established paradigm for knowledge workers to sift through mountains of historical data in order to extract nuggets of business information. Data analysis tools have been used in various forms historically since the 1960s (Pendse, 2003). Recently, there has been a rapid growth in the industry, with the total worldwide OLAP market estimated at about $3.7 billion in 2003 (Pendse). This has also been an active area of research with many contributions in data warehouse design, storage, view selection, cube computation, indexing, and query evaluation.

BACKGROUND

A data warehouse is a copy of transaction data specifically structured for querying and analysis (Kimball, 1996). Businesses want to separate their operational data from the data used for analysis for various reasons such as performance, maintenance, and security. Operational data is optimized for online transaction processing (OLTP) transactions that include lots of updates and simple queries. Analysis queries, on the other hand, are typically read only and need to aggregate large amounts of data. Data warehousing refers to the process of transforming the operational data, removing the extra fields that are not necessary for analysis, and storing it in a specialized store. In a typical scenario, the data in a data warehouse need not be current and is updated periodically with fresh data from the main store. This flexibility makes it possible to optimize the data warehouse for analytical queries rather than for updates. The data stored in a warehouse is typically multidimensional in nature and consists of a set of dimensions and metrics. In our sales example, the dimensions are products, stores, and time. The data records the sales value for different combinations of these dimensions. “Sales” is referred to as a metric in this case. Analysts might ask queries such as:

Q1. What are the total sales for each product?
Q2. What are the sales for the product “Skis” in the month of January over all stores?

To answer these queries, we would have to aggregate the data in different ways. For example, Q1 requires the addition of sales figures across all stores and over all time for each product. Multidimensional data analysis refers to this process of viewing and analyzing the data along different dimensions. A multidimensional data model is a multidimensional view of the data that identifies the various dimensions and metrics in it. Dimensions often have hierarchies on them along which they can be analyzed. For example, the time dimension has the hierarchy of time, date, month, quarter, and year. We could ask queries at different levels in the hierarchy, such as the monthly sales, quarterly sales, and yearly sales.

OLAP tools provide multidimensional analysis functionality such as aggregation, forecasting, ranking, rolling up, drilling down, etc. Though there is no standard definition of OLAP, it can be summarized in five keywords: fast analysis of shared multidimensional information (Pendse, 2003). The critical challenge for OLAP tools is to make the analysis fast so that it can be interactive. The result of the queries might be small, but the queries are quite expensive to compute since they aggregate a lot of data to get the results. For example, for Q1 we need to aggregate the sales data across all stores and over all time, which can be huge. OLAP systems use a wide variety of techniques such as specialized storage and indexing, pre-computation, and caching to provide interactive analysis.
CURRENT TECHNOLOGIES

Research in data warehousing and OLAP can be classified into the following broad areas.

Data Warehouse Design

A data warehouse is a collection of integrated information that has been designed for analytical querying. The source of this data could be one or more operational OLTP systems or other legacy systems. To get the data in a usable form, the source data needs to be processed to remove the unnecessary attributes, aggregated, and mapped to the schema of the data warehouse. The major issues in building a data warehouse are schema design and maintenance issues such as data cleansing, initial loading, incremental updates, data purging, and metadata management. The warehouse schema is designed by first building a conceptual schema in terms of dimensions and metrics and then mapping it to a logical schema that might be relational or multidimensional. The methodologies proposed for schema integration and multidatabase systems can be applied for data cleansing and mapping the source data to the destination schema. A data warehouse can be considered to be a set of materialized views on the source data. The updates in the source data need to be propagated to the data warehouse. There are various issues to be considered such as consistency requirements, refreshing timing, refreshing frequency, refreshing modes, and refreshing techniques. OLAP systems typically don’t have as stringent requirements as OLTP systems regarding the currency of data. Thus most systems update the data warehouse on a periodic basis. The updates could be done in an online fashion with the warehouse still running or an off-line fashion by shutting down the system during the refresh period. Updating the data also requires updating of the various auxiliary structures such as indices and precomputed aggregates that might be built on it. There has been a lot of research on the problem of view maintenance; we cite only a representative sample (Labio, Yerneni, & Garcia-Molina, 1999; Mumick, Quass, & Mumick, 1997; Quass & Widom, 1997). Metadata management plays a key role in data warehouses for keeping track of what data is available, where it is available, mapping between the different data elements, transformations required, etc. Jarke, Lenzerini, Vassiliou, and Vassiliadis (2003) stress that the quality of a data warehouse can be accessed and improved using enriched quality-focused metadata management and methodologies.

Data Modeling And Storage

The data in a data warehouse is multidimensional in nature. Though this data could be modeled in traditional ways such as ER modeling or relational modeling, it is more intuitive to think of it in terms of dimensions and facts. Facts represent the entity being measured and are a function of the dimension values. In our sales example, Product, Store, and Time are dimensions, and Sales is a metric. Typically the values for a dimension can be grouped in a hierarchical tree structure. For example, the hierarchy for the time dimension is shown in Figure 1. The analyst can view data at different levels along the hierarchy. Different aggregation functions can be applied to the lower-level data in order to get data at higher levels. Some commonly used aggregation functions are sum, count, min, max, and average. For example, the analyst might want to see the total sales figure at a monthly level. In this case, the aggregate function “sum” will be applied to sales values within each month to get the monthly figure. Some dimensions can have multiple hierarchies defined on them. For example, another hierarchy on the time dimension is week, quarter, year.

The multidimensional data can be stored and organized in different ways. There are two contrasting approaches to this called relational OLAP (ROLAP) and multidimensional OLAP (MOLAP). In a ROLAP system, the data is stored in standard relational tables. The analytical engine is built on top of the relational system and uses standard SQL to access data in the tables. The multidimensional data model is most commonly mapped onto a star schema. A star schema consists of a set of fact tables and dimension tables. A fact table has measure attributes that record the facts and dimension attributes that form a foreign key to the dimension tables. The dimension tables store the dimension values, their attributes, and the hierarchies on them. The star schema for the sales example consists of a fact table called Sales and three dimension tables called Product, Time, and Store as shown in Figure 2.

Figure 1. Hierarchy in Time dimension