Chapter 3
Data Warehouse Concepts

This chapter introduces the basic concepts of data warehouses. A data warehouse is a particular database targeted toward decision support. It takes data from various operational databases and other data sources and transforms it into new structures that fit better for the task of performing business analysis. Data warehouses are based on a multidimensional model, where data are represented as hypercubes, with dimensions corresponding to the various business perspectives and cube cells containing the measures to be analyzed. In Sect. 3.1, we study the multidimensional model and present its main characteristics and components. Section 3.2 gives a detailed description of the most common operations for manipulating data cubes. In Sect. 3.3, we present the main characteristics of data warehouse systems and compare them against operational databases. The architecture of data warehouse systems is described in detail in Sect. 3.4. As we shall see, in addition to the data warehouse itself, data warehouse systems are composed of back-end tools, which extract data from the various sources to populate the warehouse, and front-end tools, which are used to extract the information from the warehouse and present it to users. In Sect. 3.5, we introduce the design methodology we will use throughout the book. We finish by describing in Sect. 3.6 two representative business intelligence suite of tools, SQL Server and Pentaho.

3.1 Multidimensional Model

The importance of data analysis has been steadily increasing from the early 1990s, as organizations in all sectors are being required to improve their decision-making processes in order to maintain their competitive advantage. Traditional database systems like the ones studied in Chap. 2 do not satisfy the requirements of data analysis. They are designed and tuned to support the daily operations of an organization, and their primary concern is to ensure
fast, concurrent access to data. This requires transaction processing and concurrency control capabilities, as well as recovery techniques that guarantee data consistency. These systems are known as operational databases or online transaction processing (OLTP) systems. The OLTP paradigm is focused on transactions. In the Northwind database example, a simple transaction could involve entering a new order, reserving the products ordered, and, if the reorder point has been reached, issuing a purchase order for the required products. Eventually, a user may want to know the status of a given order. If a database is indexed following one of the techniques described in the previous chapter, a typical OLTP query like the above would require accessing only a few records of the database (and normally will return a few tuples). Since OLTP systems must support heavy transaction loads, their design should prevent update anomalies, and thus, OLTP databases are highly normalized using the techniques studied in Chap. 2. Thus, they perform poorly when executing complex queries that need to join many relational tables together or to aggregate large volumes of data. Besides, typical operational databases contain detailed data and do not include historical data.

The above needs called for a new paradigm specifically oriented to analyze the data in organizational databases to support decision making. This paradigm is called online analytical processing (OLAP). This paradigm is focused on queries, in particular, analytical queries. OLAP-oriented databases should support a heavy query load. Typical OLAP queries over the Northwind database would ask, for example, for the total sales amount by product and by customer or for the most ordered products by customer. These kinds of queries involve aggregation, and thus, processing them will require, most of the time, traversing all the records in a database table. Indexing techniques aimed at OLTP are not efficient in this case: new indexing and query optimization techniques are required for OLAP. It is easy to see that normalization is not good for these queries, since it partitions the database into many tables. Reconstructing the data would require a high number of joins.

Therefore, the need for a different database model to support OLAP was clear and led to the notion of data warehouses, which are (usually) large repositories that consolidate data from different sources (internal and external to the organization), are updated off-line (although as we will see, this is not always the case in modern data warehouse systems), and follow the multidimensional data model. Being dedicated analysis databases, data warehouses can be designed and optimized to efficiently support OLAP queries. In addition, data warehouses are also used to support other kinds of analysis tasks, like reporting, data mining, and statistical analysis.

Data warehouses and OLAP systems are based on the multidimensional model, which views data in an $n$-dimensional space, usually called a data cube or a hypercube. A data cube is defined by dimensions and facts. Dimensions are perspectives used to analyze the data. For example, consider