heterogeneous information services, such as inconsistent data and descriptions, incompatible data structures, data granularity, etc. Data warehousing systems use a variety of data extraction and cleaning tools, and load and refresh utilities for populating warehouses (*back end* tools).

2. **Organizing** and integrating data consistently into the data warehouse: as defined by [INM96], "a data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process.” *Time-variant* means the data associates with a point or interval of time. *Nonvolatile* means the data doesn't change once it gets into the warehouse; there are only two operations: initial loading of data and access of data.

3. **Accessing** the integrated data in an efficient and flexible fashion: data in the warehouse and data marts is stored and managed by one or more warehouse servers, which present multidimensional views of data to a variety of *front end* tools: query tools, report writers, analysis tools, and data mining tools. There is yet a repository for storing and managing metadata of the system.

Next sections will cover the data cleaning technology, how DM can be improved by multidimensional modeling, and the interaction between OLAP and DM tools.

### 2.1.1 Data cleaning tools

Data warehouse is used for decision making, so it is necessary that the data is consistent. However, since large volumes of data from multiple sources are involved, there is a high probability of errors and anomalies in the data. There are three related, but somewhat different, classes of data cleaning tools [CHA97]. *Data migration* tools are used to help extract data from all types of disparate sources and transform it into consistent formats. *Data scrubbing* tools use domain-specific knowledge (e.g., postal addresses) to clean data. *Data auditing* tools can be specific purpose data mining tools, capable to discover rules and relationships by scanning data, and then look for exceptions, which can represent errors in databases.
Even when DM is not done in the ideal DW scenario, data cleaning technology still can be very useful to improve data quality, and thus the quality of discoveries, as learning techniques follow the “garbage in-garbage out” principle.

2.1.2 Multidimensional modeling support for OLAP and DM

A popular approach for analysis of data warehouses is called On-Line Analytical Processing (OLAP). OLAP tools focus on providing multidimensional data analysis, which is superior to SQL in computing summaries and breakdowns along many dimensions. A model that influences the query engines for OLAP is the multidimensional view of data in the warehouse [CHA97]. In a multidimensional data model, there is a collection of dimensions, which provide a description of a business process upon which numeric measures of the business process may depend. For example, a retail business may wish to see sales figures, a measure, over time, by product, by promotion, or by store. Time, Product, Promotion and Store are the dimensions upon which the sales figures depend. Each dimension is described by a set of attributes. Multidimensional model is done based on analytical needs, providing history of information that could be useful for both OLAP and DM tools.

Building a dimensional data warehouse is a process of matching the needs of the user community to the realities of the available data. [KIM96] describes a methodology (multidimensional modeling) for asking the right questions and eliciting a complete description of the organization’s environment so that a satisfactory design for data warehouse can be completed. The methodology consists of four main steps:

1. Choose a business process to model: a business process is a major operational process in an organization that is supported by some kind of legacy system (or systems) from which data can be collected for the purposes of the data warehouse. Examples of business process are orders, invoices, shipments, inventory and sales.

2. Choose the grain of the business process: the grain is the fundamental, atomic level of data to be represented in the model for this process. Typical grains are individual transactions, individual daily snapshots, or individual monthly snapshots.