IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITB8918

Chapter XII

Source Integration for Data Warehousing

Andrea Calí
Università di Roma–La Sapienza, Italy

Domenico Lembo
Università di Roma–La Sapienza, Italy

Maurizio Lenzerini
Università di Roma–La Sapienza, Italy

Riccardo Rosati
Università di Roma–La Sapienza, Italy

ABSTRACT

While the main goal of a data warehouse is to provide support for data analysis and management's decisions, a fundamental aspect in design of a data warehouse system is the process of acquiring the raw data from a set of relevant information sources. We will call source integration system the component of a data warehouse system dealing with this process. The main goal of a source integration system is to deal with the transfer of data from the set of sources constituting the application-oriented operational environment, to the data warehouse. Since sources are typically autonomous, distributed, and heterogeneous, this task has to deal with the problem of cleaning, reconciling,

and integrating data coming from the sources. The design of a source integration system is a very complex task, which comprises several different issues. The purpose of this chapter is to discuss the most important problems arising in the design of a source integration system, with special emphasis on schema integration, processing queries for data integration, and data cleaning and reconciliation.

INTRODUCTION

The typical architecture of a data warehouse system is constituted by two different components, usually called back-end and front-end, respectively. While the latter is intended to provide support for the main task of the system, namely, data analysis and management's decisions, the former is responsible for acquiring the raw data from a set of relevant information sources. One of the basic assumptions in data warehouse architectures is that a correct modularization requires these two components be decoupled. It is the task of the back-end component to free the front-end from the knowledge on where data are, and how data are structured at the sources.

The goal of this chapter is to discuss the most important issues in the design of the back-end component of a data warehouse. We will call such component "source integration system," because its main goal is to deal with the task of transferring data from the set of sources constituting the application-oriented operational environment to the data warehouse. When data passes from one environment to the other, possible inconsistencies and redundancies should be resolved, so that the warehouse is able to provide an integrated and reconciled view of data of the organization (Inmon, 1996).

The constraints that are typical of data warehouse applications restrict the large spectrum of approaches that have been proposed for information integration (Hull & Zhou, 1996; Inmon, 1996; Jarke et al., 1999). In particular, since the data warehouse should reflect the informational needs of the organization, it should be based on a unified, corporate view of data, called *global schema*. Without the definition of a global schema, the risk arises of concentrating on what is in the sources at the operational level, rather than on what is really needed in order to perform the required analysis on data (Devlin, 1997).

The architecture of a source integration system is usually described in terms of two types of modules: wrappers and mediators (Wiederhold, 1994; Ullman, 1997). The goal of a wrapper is to access a source, extract the relevant data, and present such data in a specified format. The role of a mediator is to collect, clean, and combine data produced by different wrappers (or mediators), so as to meet a specific information need of the data warehouse. The specification and the realization of mediators is the core problem in the design of a source integration system.

The design of a source integration system is a very complex task, which comprises several different issues, including the following: