Data Integration in Data Warehousing

Diego Calvanese

Dipartimento di Informatica e Sistemistica
Universita` di Roma ``La Sapienza''
Via Salaria 113, 00198 Roma, Italy
http://www.dis.uniroma1.it/~calvanese/

Data integration is a central problem in the design of Data Wareshouses and Decision Support Systems. When data passes from the sources of the application-oriented operational environment to the Data Warehouse, 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.

Generally speaking, a data integration system combines the data residing at different sources, and provides a unified, reconciled view of these data, called global schema, which can be queried by the user. In the design of a data integration system, an important aspect is the way in which the global schema is specified, i.e., which data model is adopted and what kind of constraints on the data can be expressed. Moreover, a basic decision is related to the problem of how to specify the relation between the sources and the global schema. There are basically two approaches for this problem. The first approach, called global-as view (GAV), requires that the global schema is expressed in terms of the data sources. More precisely, to every concept of the global schema, a view over the data sources is associated, so that its meaning is specified in terms of the data residing at the sources. In the second approach, called local-as-view (LAV), the global schema is specified independently from the sources, and the relationships between the global schema and the sources are established by defining every source as a view over the global schema.

The ultimate goal of a data integration system is to answer queries posed by the user in terms of the global schema. Obviously, query processing depends on the form of the data integration system and, specifically, on whether the GAV or LAV approach is adopted and on the form of constraints allowed on the global schema. In the invited talk we illustrate basic techniques for computing the correct answers to a data integration system in various practically significant cases. We then consider the conditions that are typical of Data Warehouse applications, which restrict the large spectrum of approaches that have been proposed for integration. We discuss a data integration architecture specifically developed for this context within the IST European Project ``Foundations of Data Warehouse Quality'' (DWQ). The DWQ integration architecture follows the LAV approach and defines both Data Warehouse tables and source tables in terms of a global schema.