What is the data warehousing problem? (Are materialized views the answer?)

Ashish Gupta, Junglee Corporation
Inderpal Singh Mumick, AT&T Research

The term *Data Warehousing* is used for database applications with one or more of the following characteristics:

1. Data is integrated from several, possibly heterogeneous, sources into a large data store, called the *data warehouse.*
2. A large data store functions as the database of record, with access to detailed data for operational and/or decision support applications. The database of record is called a *data warehouse.*
3. A *Data Warehouse* summarizes data along several dimensions, and stores the summarized data for aggregate query processing by OLAP and decision support applications. The detailed data may or may not be stored in the warehouse.

A view is a derived relation defined in terms of base (stored) relations. A view can be materialized by storing the tuples of the view in the database. A materialized view provides fast access to data; the speed difference is critical in applications where the query rate is high and the views are complex or over data in remote databases, so that it is not feasible to recompute the view for every query.

Data warehousing has become increasingly visible as a research issue following in the wake of enormous market activity in the past few years. Warehousing is reputed to be the next big corporate information initiative where every database company hopes to make its fortune. Similarly, materialized views are finding increased research activity, with applications in decision support, OLAP, query optimization, and replication, all of which are relevant for data warehousing.

What new database problems are opened up by data warehousing? Clearly, warehouses need database systems to support larger and larger amounts of data, running into hundreds of gigabytes and tens of terabytes. Large parallel database systems need to be developed. However, are there problems other than those associated with building any large database system. What about issues of database integration, heterogenous systems, database loading, batch processing, data snapshots, backups, aggregate query processing, and OLAP query optimization.

Can materialized view technology provide the answer to most or all of these problems? Many people believe so, and claim that warehousing is no more than a new name for caching and materialized views. Many researchers and industry developers have put their time and money behind this belief and are building systems and products based on materialized views.

Can materialized views technology solve the problems encountered in doing data warehousing using database systems? What work needs to be done in materialized views to develop such technology and to make it usable? Are there significant warehousing problems outside materialized views?

Proceedings of the 22nd VLDB Conference
Mumbai(Bombay), India, 1996