Data Warehousing and OLAP for Decision Support

Surajit Chaudhuri
Microsoft Research, Redmond
surajitc@microsoft.com

Umeshwar Dayal
Hewlett-Packard Laboratories, Palo Alto
dayal@hpl.hp.com

Description

On-Line Analytical Processing (OLAP) and Data Warehousing are *decision support* technologies. Their goal is to enable enterprises to gain competitive advantage by exploiting the ever-growing amount of data that is collected and stored in corporate databases and files for better and faster decision making. Over the past few years, these technologies have experienced explosive growth, both in the number of products and services offered, and in the extent of coverage in the trade press. Vendors, including all database companies, are paying increasing attention to all aspects of decision support.

Decision support places some rather different requirements on database technology as compared to traditional on-line transaction processing (OLTP) applications. OLTP applications typically automate clerical data processing tasks such as order entry and banking transactions that are the bread-and-butter, day-to-day operations of an organization. These tasks are structured and repetitive, and consist of short, atomic, isolated transactions, which require detailed, up-to-date data, and read or update a few (tens of) records. Consistency and recoverability of the database are critical, and maximizing transaction throughput is the key performance metric.

Decision support, in contrast, requires historical, summarized and consolidated data from many sources scattered through the enterprise. Data is extracted from these sources and loaded into a data warehouse, a large, integrated, relatively static, database that is often maintained separately from the organization's operational databases. To facilitate complex analyses and visualization, the data warehouse typically supports a *multidimensional* model of data. Since data warehouses contain consolidated data, perhaps from several operational databases, over potentially long periods of time, they tend to be orders of magnitude larger than operational databases; enterprise data warehouses are projected to be hundreds of gigabytes to terabytes in size. The workloads are query intensive with mostly *ad hoc*, complex queries that can access millions of records. Query throughput and response times are more important than transaction throughput.

Data warehouses might be implemented on standard or extended relational database management systems, called ***Relational OLAP (ROLAP)*** servers. These servers assume that data is stored in relational databases, using special database designs (star and snowflake schemas) to represent the multidimensional data model; special access methods and query processing techniques to efficiently map OLAP operations on the underlying relational database. Alternatively, ***multidimensional OLAP (MOLAP)*** servers may be used. These are specialized servers that directly store multidimensional data in special data structures (e.g., arrays) and implement the OLAP operations over these special data structures.

This tutorial provides a roadmap of data warehousing and OLAP technologies, with an emphasis on their new requirements. We describe back end tools for extracting, cleaning and loading data into a data warehouse; multi-dimensional data models and OLAP operations; front end client tools for querying and data analysis; server extensions for efficient query processing; and tools for metadata management and for managing the warehouse. We survey the state of the art and mention representative products. In a recent overview paper, we have summarized the issues that are discussed in this tutorial [1].

The area opens up interesting research directions, with ties to past work in database systems, but with different assumptions and requirements. Only very recently, however, has the database research community started to address some of these issues. Research in data warehousing so far has focused primarily on query processing and view maintenance issues. There still are many open research problems. We describe some of these briefly.

Outline

1. Introduction
   - definitions, evolution, differences from OLTP, architectures
2. Models and Tools
   - conceptual model for OLAP
   - front-end tools (e.g., multidimensional spreadsheets)
   - database design (e.g., star and snowflake schema)
3. Database Server technologies for Decision Support Queries
   - specialized indexing and query processing techniques

SIGMOD '97 AZ,USA
© 1997 ACM 0-89791-911-4/97/0005...$3.50

- intelligent processing of aggregates
- complex query processing
- extensions to SQL
- ROLAP vs MOLAP

4. Other Services for OLAP/Data warehousing
    - data cleaning, loading and refresh
    - tools for warehouse, system and process management
    - metadata management and the role of repository
5. State of Commercial Practice
6. Research Issues

References

[1] S. Chaudhuri, and U. Dayal, "An Overview of Data Warehousing and OLAP Technology", SIGMOD Record, March 1997.