Data Warehousing and OLAP

• Topics
  – Introduction
  – Data modelling in data warehouses
  – Building data warehouses
  – View Maintenance
  – OLAP and data mining

• Reading
  – Lecture Notes
  – Elmasri and Navathe. Chapter 26
  – Ozsu and Valduriez, Chapter 16

Terminology and Definitions

• DW provides access to data for complex analysis, knowledge discovery and decision support
• High performance demands
• OLAP - on line analytical support
• DSS - decision support systems also known as EIS (executive information systems)
• Traditional Databases support on-line transaction processing (OLTP) - updates, deletions and insertions

Data Warehouse characteristics

• Multidimensional conceptual view
• Unlimited dimensions and aggregation levels
• Unrestricted cross-dimensional operations
• Client-server architecture
• Multi user support
• Accessibility
• Transparency
• Intuitive data manipulation
• Flexible reporting
• Scalability

Data Warehouse Types

• Enterprise-wide data warehouses - large projects with massive investment of time and resources
• Virtual data warehouses - provide views of operational Databases that are materialized for efficiency
• Data Marts - target a subset of the organization, such as departments

Typical DW Implementation

Data Warehousing Environments

– In a DW, the data that is subject to analysis is decoupled from the data produced at the source.
– Information in DW can be organised in a form that makes it easy to use for applications. Views: (from simple replication to arbitrarily complex processing).
– Information is available independently at the availability of the source. The views are materialised.
– Information is structured and stored as to optimise processing of queries against the DW.
– A small amount of cooperation is required from the source to keep the warehouse in sync when the sources change.
Data Modeling for DWs

- Multidimensional models – hyper-cubes if more than 3 dimensions,
  - query performance much better than in the relational model.
  - there are typically three dimensions in corporate DW: fiscal periods, products and regions,

Two-dimensional matrix

A data cube

fiscal period

product

region

Multidimensional displays

- Roll-up Display - moves up the hierarchy, grouping into larger units along a dimension
  for example from days to weeks, from weeks to months, from months to quarters, from quarters to years, etc

The roll-up operation

region

Products 1xx
Products 2xx
Products nxx

Multidimensional displays

- Drill-down Display - provides a finer grain view by dis-aggregating some dimensions
  for example from months to weeks, from weeks to days, etc
The drill-down operation

Sub-regions

<table>
<thead>
<tr>
<th></th>
<th>P123</th>
<th>P124</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Tables in multi-dimensional model

- Two types of tables in multi-dimensional model:
  - Dimension table - tuples of attributes of dimension,
  - Fact table - contains measured variable(s) and identifies it (them) with pointers to dimension tables.

Multi-dimensional schemas

Two common multi-dimensional schemas:
- Star schema - consists of a fact table with a single table for each dimension,
- Snowflake schema - a variation on the star schema - dimensional tables are organized into a hierarchy by normalizing them.

A Star Schema

Fact table - business results

Fiscal quarters

Product

Quarter Region

Fact table 1 - business results

Product

Quarter Region

Sales revenue for regions

Region

A Snowflake Schema

A fact constellation

Fact table 1 - business results

Dimension Table Product

Fact table 2 - business forecast

Product
Building a Data Warehouse

- The data must be extracted from multiple, heterogeneous sources
- The data must be formatted for consistency
- The data must be cleaned to ensure the validity
- The data must be fitted to the DW data model
- The data must be loaded into the DW

Important Questions

- How up-to-date must the data be?
- Can the DW go off-line, and for how long?
- What are the data interdependencies?
- What is the storage availability?
- What are the distribution requirements - replication and partitioning?
- What is the loading time - including cleaning, formatting, coping, transmitting, index re-building?

Data Storage Processes

- Storing the data according to the DW data model
- Creating and maintaining required data structures
- Creating and maintaining appropriate access path
- Supporting the updating of VIEWS
- Refreshing the data

Design Considerations

- Usage projection
- The fit of the data model
- Characteristics of available data sources
- Design of the meta data components
- Modular component design
- Design for a change
- Considerations of distributed and parallel architectures

Difficulties with implementation

- Schema conflicts resolutions
- The administration of a DW is an intensive enterprise
- The quality control of the data
- Correct estimations of usage - optimization issues
- Selection of highly specialized team
Query Mediation

- Traditional, virtual tables
- A user query is decomposed into sub-queries that are executed by the data sources
- Answers are based always on current data
- Query performance (at several sites) for large sets of data is a problem

A Query Mediation Architecture

Monitors

- Views are materialised
- Need for a view maintenance
- The question is: WHERE is the view maintenance performed?
- In the monitor architecture, the responsibility rests with the sources. Monitors are installed.
  - A few prototypes/products adopted monitors
  - IBM Starburst
  - ConceptBase system
- Disadvantage is that additional work load is imposed on the data sources

A Monitor Architecture

Data Warehouse Architecture

View Maintenance in DW

- Full re-computation
- DW taken down periodically for scheduled maintenance
- During this period all the views are re-derived from scratch from the data sources
- It is frequently accepted as the simplest and safest policy
- However, for yesterday’s data, it is often wasteful,
- Time limitation for re-computation is often a big problem
Incremental Maintenance

Only parts of DW that change are computed.
1. DW scheduled for maintenance as before but views are incrementally done.
   • All changes made to the data sources during operation hours to be logged.
2. Maintenance can be dynamic, views always reflect fresh data.
   • Problem is efficiency of dynamic maintenance and data quality.

Approaches to Incremental View Maintenance

- Unrestricted base access
- Self-maintainable DW
- Run-time DW self-maintainable

Web Access to DW

Web-Based DW Access

- How do you hone the environment to provide good performance to large numbers of users?
- How can you serve the needs of both web and non-web clients?
- How can you provide Internet access to internal data warehousing applications without opening a security hole?
- How can you help users design queries for efficient processing?
- How do you limit unreasonable queries that would jam the system?

OLAP and Data Mining

Provide a strong decision support environment

- Data Mining:
  - Characterization of patterns inherent in the data,
  - Development of a hypothesis from these patterns,
  - Prediction of future behaviour.
- OLAP
  - The first approach consists of the use of relational technology, suitably adapted and extended. The data is stored using tables, but the analysis operations are carried out efficiently using special data structures. This type of system is normally called ROLAP (Relational OLAP).
  - The second, more radical, approach consists of storing data directly in a multi-dimensional form, using vector data structures. This type of system is called MOLAP (Multi-dimensional OLAP).

Goals of Data Mining

- Prediction - to show a future potential behavior
- Identification - data patterns indicate the existence of an item, an event, or an activity
- Classification - partition the data based on identified combinations of parameters
- Optimization - optimize the use of limited resources such as time, money, space, material
Types of knowledge discovered during Data Mining

- Association rules - correlate the presence of the set of items with another set of values for another set of variables
- Classification hierarchies - to create a hierarchy of classes
- Sequential patterns - a sequence of action or events is sought
- Patterns within time series - similarities detected within positions of the time series

Next Lecture

Introduction to Business Process Technology