Data Warehouse and Data Mining

Lecture No. 07

Data Modeling

Naeem Ahmed
Email: naeemmahoto@gmail.com

Department of Software Engineering
Mehran University of Engineering and Technology Jamshoro
Physical Models

• Based on the physical model used:
  – DOLAP (Desktop OLAP)
  – MOLAP (Multidimensional OLAP)
  – ROLAP (Relational OLAP)
  – HOLAP (Hybrid OLAP)
OLAP

- A decision support system (DSS) that support ad-hoc querying, i.e. enables managers and analysts to interactively manipulate data.
- Analysis of information in a database for the purpose of making management decision.
- The idea is to allow the users to easy and quickly manipulate and visualize the data through multidimensional views (i.e. different perspectives).
- OLAP (OnLine Analytical Processing) analyzes historical data (terabytes) using complex queries.
OLAP

• OLAP Council definition:
  – A category of software technology that enables analysts, managers and executives to gain insight into data through fast, consistent, interactive access to a wide variety of possible views of information that has been transformed from raw data to reflect the real dimensionality of the enterprise as understood by the user.

• OLAP is implemented in a multi-user client/server mode and offers consistently rapid response to queries, regardless of database size and complexity.
OLAP

- OLAP primarily involves aggregating large amounts of diverse data
- OLAP functionality provides dynamic multi-dimensional analysis, supporting analytical and navigational activities
- OLAP functionality is provided by the OLAP Server
- OLAP Council defines OLAP Server as:
  - ‘A high capacity, multi-user data manipulation engine specifically designed to support and operate on multi-dimensional data structures.’
## OLTP vs. OLAP

<table>
<thead>
<tr>
<th>OLTP</th>
<th>OLAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operational processing</td>
<td>Informational processing</td>
</tr>
<tr>
<td>Transaction-oriented</td>
<td>Analysis-oriented</td>
</tr>
<tr>
<td>For operational staffs</td>
<td>For managers, executive &amp; analysts</td>
</tr>
<tr>
<td>Daily operations</td>
<td>Decision support</td>
</tr>
<tr>
<td>Current, up-to-date data</td>
<td>Historical data</td>
</tr>
<tr>
<td>Primitive, highly detailed data</td>
<td>Summarized, consolidated data</td>
</tr>
<tr>
<td>Detailed, flat relational views</td>
<td>Summarized, multi-dimensional views</td>
</tr>
<tr>
<td>Short, simple transactions</td>
<td>Complex aggregate queries</td>
</tr>
<tr>
<td>Read/write</td>
<td>Mostly read only</td>
</tr>
<tr>
<td>Index on keys</td>
<td>Many scans</td>
</tr>
<tr>
<td>Many users</td>
<td>Small number of users</td>
</tr>
<tr>
<td>Large databases</td>
<td>Very large databases</td>
</tr>
</tbody>
</table>
OLTP vs. OLAP

• On-Line Transaction Processing
  – Transfer $100 balance from my saving account to my checking account

• On-Line Analytical Processing
  – What is the average balance of accounts by customer groups, account types, areas, account managers, and their combinations?
Physical Models

• DOLAP
  - Developed as extension to the production system reports
    - It downloads a small hypercube from a central point (data mart or DW)
    - Performs multidimensional analysis while disconnected from the data source
    - The **computation occurs on the client**
  - Requires little investment
  - They lack the ability to manage large data sets
  - All processing work is done in the **desktop**
    - E.g, bring data into Excel and build a pivot table
Physical Models

- **MOLAP**
  - Presentation layer provides the multidimensional view
  - The OLAP server stores data in a multidimensional structure
    - Computation occurs in this layer during the **loading step** (not at query)
  - Advantages
    - Excellent performance
      - Fast data retrieval
      - Optimal for slicing and dicing
      - All calculations are pre-generated when the cube is created
Physical Models

• MOLAP
  – Disadvantages
    • Limited amount of data it can handle
      – Cube can be derived from large amount of data, but only summary level information will be included in the cube
    • Requires additional investment
      – Cube technology are often proprietary
    • Enormous amount of overhead
      – An input file of 200 MB can expand to 5 GB with calculations
  – Products
    • Cognos (IBM), Essbase (Oracle), Microsoft Analysis Service, Palo (open source)
Physical Models

• ROLAP
  – Presentation layer provides the multidimensional view
  – The ROLAP Server generates SQL queries, from the OLAP OLAP requests, to query the RDBMS
  – Data is stored in RDBs
Physical Models

• ROLAP
  – Special schema design: e.g., star, snowflake
  – Special indexes: e.g., bitmap, R-Trees
  – Advantages
    • Proven technology (relational model, DBMS)
    • Can handle large amounts of data (VLDBs)
  – Disadvantages
    • Limited SQL functionalities
  – Products
    • Microsoft Analysis Service, Siebel Analytics (now Oracle BI), Micro Strategy, Mondrian (open source)
# ROLAP vs. MOLAP

- Based on OLAP needs

<table>
<thead>
<tr>
<th>OLAP needs</th>
<th>MOLAP</th>
<th>ROLAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multidimensional View</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Excellent Performance</td>
<td>✓</td>
<td>-</td>
</tr>
<tr>
<td>Analytical Flexibility</td>
<td>✓</td>
<td>-</td>
</tr>
<tr>
<td>Real-Time Data Access</td>
<td>-</td>
<td>✓</td>
</tr>
<tr>
<td>High Data Capacity</td>
<td>-</td>
<td>✓</td>
</tr>
<tr>
<td>Easy Development</td>
<td>✓</td>
<td>-</td>
</tr>
<tr>
<td>Low Structure Maintenance</td>
<td>-</td>
<td>✓</td>
</tr>
<tr>
<td>Low Aggregate Maintenance</td>
<td>✓</td>
<td>-</td>
</tr>
</tbody>
</table>
Physical Models

• HOLAP
  – Best of both models
    • Storing detailed data in RDBs
    • Storing aggregated data in MDBs
  – Different partitioning approaches between MOLAP and ROLAP
    • Vertical
    • Horizontal
HOLAP

• Vertical partitioning
  – Aggregations are stored in MOLAP for **fast query performance**
  – Detailed data in ROLAP to **optimize time of cube processing** (loading the data from the OLTP)

• Horizontal partitioning
  – HOLAP stores some slice of data, usually the more recent one (i.e. sliced by Time dimension) in MOLAP for fast query performance
  – Older data in ROLAP
HOLAP

• Other approaches
  – Store some cubes in MOLAP and others in ROLAP, leveraging the fact that in a large cuboid, there will be dense and sparse sub-regions
Conclusions

• ROLAP
  – RDBMS - star/snowflake schema
  – For detailed and larger volumes of data (TB)

• MOLAP
  – MDBMS - Cube structures, array based storage
  – For summarized and relatively “small” volumes of data (50GB)

• HOLAP is emerging as the OLAP server of choice
Summary

• Logical Model
  – Cubes, Dimensions, Hierarchies, Classification Levels

• Physical Level
  – Relational Implementation through:
    • Star schema: improves query performance for often-used data
      – Less tables and simple structure
      – Efficient query processing with regard to dimensions
      – In some cases, high overhead of redundant data
    • Snowflake schema: reduce the size of the dimension tables
      – However, through dimension normalization - large number of tables
Summary

• Physical Level
  – Array based storage
    • How to perform linearization
    • Problems:
      – Order of dimensions – solution: caching
      – Dense Cubes, Sparse Cubes - solution: 2 level storage
  – MOLAP, ROLAP, HOLAP