An Overview of Data Warehousing and OLAP Technology

CS743 Paper Presentation

S. Chaudhuri and U. Dayal
ACM SIGMOD 1997

Presenter: Jasnoor Singh Mann
Outline

- Background
- Data Warehousing and OLAP
  - Architecture
  - Back End
  - Conceptual Model
  - Front End
  - Design
  - Metadata and Management
  - Warehouse Efficiency
- Concluding Remarks
Background

- **Operational (traditional) databases**
  - On-Line Transactional Processing (OLTP) systems
  - Focused on transactions and transactional efficiency
  - Designed to be write-optimized, and managing transactions

- **Data Warehousing**
  - On-Line Analytical Processing (OLAP) systems
  - Geared at analytics; emphasis on throughput and response time
  - Optimized for reading data, and answering questions critical in business environment
  - Not designed for “inserting” data in a traditional sense
Operational (traditional) databases

- Data *preferably* stored in normalized form
- Schema usually modeled after E-R diagrams
- Typical tasks include fetching or storing records

Data Warehousing

- Data not usually stored in normalized form
- Schema not typically modeled after E-R diagrams
- Involves querying over historical, summarized and consolidated data
A Data Warehouse is indeed a database.

The data stored is used for decision support.

Sourced from operational databases, usually multiple, and external sources.

The data spans hundreds of gigabytes to terabytes in size,

From the users’ viewpoint: read-only.

Typically used in businesses, including manufacturing, retail, financial services, transportation, telecommunications, and health care.
Data Warehousing (cont.)

- Traditional (OLTP) systems not used for this purpose, as their performance is unacceptable.
- Operational databases missing some data needed for decision support.
- OLAP systems store historical and consolidated data, which is essential.
Architecture

Figure 1. A typical Data Warehousing Architecture
 Variety of tools for data extraction, cleaning, loading and refreshing the data warehouse.

**Data extraction** from “foreign” sources done though standardized interfaces (ODBC, Oracle Open Connect).

**Data cleaning** needed as data loaded from different sources, possibly containing anomalies.

- *Data migration*: Basic cleaning, data transformation rules such as string replacement.
- *Data scrubbing*: Intermediate cleaning, using domain specific knowledge (e.g., postal addresses).
- *Data auditing*: Advanced cleaning, audit based on rules and relationships. (e.g., suspicious patterns based on statistics).
**Data Loading** is addition of data to the warehouse from the foreign sources. Load utilities are used.

- Pre-processing of the data may be required, batch load utilities used typically.
- System admin must be able to monitor status, cancel, suspend and resume, and restart the process.
- Larger volumes of data involved; pipelining and parallel processing used.
Data Refreshing is propagating the changes made on the source data to the warehouse.

- Usually refreshed periodically; some queries require current data warranting propagation of every update.

- Replication support in modern database systems:
  - Data Shipping, or
  - Transaction Shipping
A popular conceptual model for OLAP is *multi-dimensional view* of data stored in a warehouse.

- A set of numeric *measures* are the objects of analysis. E.g., Sales.
- Each depending on a set of *dimensions*, such as the City, Product Name and the Date of Sale.

Each *measure* is a value in multi-dimensional space.
A **dimension** has a set of attributes associated with it. E.g., Product may have category, industry, year of introduction, and average profit margin as attributes.

The attributes of a dimension may be related in a hierarchy of relationships.

Hierarchies are significant in OLAP systems, more ahead.

Dimensions: Product, City, Date

Hierarchical summarization paths

- Industry
  - Category
    - Product
  - Country
    - State
      - City
    - Month
      - Week
      - Date
  - Year
    - Quarter
      - Month
      - Week
      - Date
Front-End

Tools used to interact with the data warehouse.

- **Multi-dimensional spreadsheet applications**: One of the most compelling front-ends, and spreadsheets used traditionally by analysts.

- **Managed query environments**: They use stored procedures and predefined complex queries to provide packaged analysis tools.

- **Data Mining tools**: Some data mining tools are also used directly to discover correlations in data.
Front-End: Operations

- Different operations supported in spreadsheet environment.
- Aggregation in OLAP is one of the most used operations. Aggregation of measures by one or more dimensions, such as Sales by City and Year.
- Spreadsheet Operations:
  - Pivoting,
  - Roll Up,
  - Drill Down,
  - Slice and Dice, and
  - Others – Ranking, Selections, Pre-computing.
Front-End: Operations (cont.)

- Pivoting:
- **Roll-Up:**

<table>
<thead>
<tr>
<th>Locations (countries)</th>
<th>Time (Quarter)</th>
<th>Mobile Modem Phone Security item (types)</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>Q1</td>
<td>1000</td>
</tr>
<tr>
<td>Canada</td>
<td>Q2</td>
<td>2000</td>
</tr>
<tr>
<td></td>
<td>Q3</td>
<td>1000</td>
</tr>
<tr>
<td></td>
<td>Q4</td>
<td>2000</td>
</tr>
</tbody>
</table>

- Roll-up on location (from cities to countries):

<table>
<thead>
<tr>
<th>Locations (Cities)</th>
<th>Time (Quarter)</th>
<th>Mobile Modem Phone Security item (types)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Q1</td>
<td>605</td>
</tr>
<tr>
<td></td>
<td>Q2</td>
<td>825</td>
</tr>
<tr>
<td></td>
<td>Q3</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Q4</td>
<td>400</td>
</tr>
</tbody>
</table>

Jasnoor Singh Mann  An Overview of Data Warehousing and OLAP Technology
Drill-Down:

- Mobile Modem Phone Security item (types)
- Locations (countries)
- Time (quarters)
- Time (months)
- Drill down on time (from quarters to month)
Front-End: Operations (cont.)

- **Slice:**

  ![Slice Diagram]

- **Dice:**

  ![Dice Diagram]
The multi-dimensional model is directly implemented by MOLAP servers.

However, relational ROLAP servers can be used. The multi-dimensional model and operations are mapped into relations and SQL queries.

Some commonly used relational schemas:
- *Star schema*,
- *Snowflake schema, and*
- *Fact constellation schema.*
Database Design (cont.)

Star Schema:

Order
- OrderNo
- OrderDate

Customer
- CustomerNo
- CustomerName
- CustomerAddress
- City

Salesperson
- SalespersonID
- SalespersonName
- City
- Quota

Fact table
- OrderNo
- SalespersonID
- CustomerNo
- ProdNo
- Date
- DateKey
- CityName
- Quantity
- TotalPrice

ProdNo
- ProdName
- ProdDescr
- Category
- CategoryDescr
- UnitPrice
- QOH

Figure 4. A Star Schema
Database Design (cont.)

Snowflake Schema:

- **Order**
  - OrderNo
  - OrderDate

- **Customer**
  - CustomerNo
  - CustomerName
  - CustomerAddress
  - City

- **Salesperson**
  - SalespersonID
  - SalespersonName
  - City
  - Quota

- **Fact table**
  - OrderNo
  - SalespersonID
  - CustomerNo
  - DateKey
  - CityName
  - ProdNo
  - Quantity
  - TotalPrice

- **ProdNo**
  - ProdName
  - ProdDescr
  - Category
  - UnitPrice
  - QOH

- **Category**
  - CategoryName
  - CategoryDescr

- **Date**
  - DateKey
  - Date
  - Month

- **Month**
  - Month
  - Year

- **City**
  - CityName
  - State

- **State**

Figure 5. A Snowflake Schema
- **Fact Constellations** are more complex schemas where multiple fact tables share dimension tables.

- Pre-aggregated data stored also stored along with, either in summary tables or in existing dimension tables and fact table.

- The latter technique may lead to more operational errors; additional interpretation and distinction of pre-aggregated data required.
An essential part of data warehousing is metadata management. Different kinds of metadata serve distinct purposes.

- **Administrative metadata** contains information for setting up and using the warehouse, descriptions of different elements.
- **Business metadata** includes business terms and definitions, ownership of data and policies.
- **Operational metadata** includes monitoring information like usage statistics, and error reports.
Data Warehousing involves massive amounts of data, answering queries requires a highly efficient system.

Additional access structures such as indices, (materialized) views, join indices used.

Complex queries need to be optimized.

Some queries may need sequential scans, which need to be optimized as well.

Parallelism needs to be exploited for query execution.

Different server architectures available.

SQL extensions for analytics proposed.
Concluding Remarks

- Warehouse could be distributed for scalability and higher availability.
- Designing and rolling out is a complex process, and involves careful planning and execution of different activities.
- Data Warehouse management is also not a straightforward task. Different tools are available for the management.

- Credit: tutorialspoint.com for graphical representations of OLAP operations.
Summary

- Data Warehousing contains large amounts of data which is used for decision support in business intelligence.
- It is based on the OLAP model, as opposed to traditional systems.
- It involves different tools at the back-end to handle the data, which are crucial for the operation of a Data Warehouse.
- Conceptually, the data is represented using a multi-dimensional model.
- Data Warehousing requires an efficient system, and makes use of different technologies to ensure that.