OLAP and OLTP

AMIT KUMAR BINDAL
Associate Professor
Databases

• Databases are developed on the IDEA that DATA is one of the critical materials of the Information Age

• Information, which is created by data, becomes the bases for decision making
What is Data Warehousing?

- Data Warehousing is subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management’s decision-making process.
- A data warehouse is data management and data analysis.
- Data Webhouse is a distributed data warehouse that is implemented over the web with no central data repository.
- Goal: is to integrate enterprise wide corporate data into a single repository from which users can easily run queries.
Data Warehouse

• DSS – friendly data repository for the DSS is the DATA WAREHOUSE

• Definition: Integrated, Subject-Oriented, Time-Variant, Nonvolatile database that provides support for decision making
Integrated

- The data warehouse is a centralized, consolidated database that integrates data derived from the entire organization:
  - Multiple Sources
  - Diverse Sources
  - Diverse Formats
Subject-Oriented

- Data is arranged and optimized to provide answer to questions from diverse functional areas
  - Data is organized and summarized by topic
    - Sales / Marketing / Finance / Distribution / Etc.
Time-Variant

• The Data Warehouse represents the flow of data through time
• Can contain projected data from statistical models
• Data is periodically uploaded then time-dependent data is recomputed
Nonvolatile

• Once data is entered it is NEVER removed
• Represents the company’s entire history
  – Near term history is continually added to it
  – Always growing
  – Must support terabyte databases and multiprocessors
• Read-Only database for data analysis and query processing
A DBMS built for online transaction processing (OLTP) is generally regarded as unsuitable for data warehousing because each system is designed with a differing set of requirements in mind e.g.: OLTP systems are design to maximize the transaction processing capacity, while data warehouses are designed to support ad hoc query processing.
## Comparision of OLTP & Data Warehousing Systems

<table>
<thead>
<tr>
<th>OLTP systems</th>
<th>Data Warehousing Systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hold current data</td>
<td>Holds historical data</td>
</tr>
<tr>
<td>Stores detailed data</td>
<td>Stores detailed, lightly, and highly summarized data</td>
</tr>
<tr>
<td>Data is dynamic</td>
<td>Data is largely static</td>
</tr>
<tr>
<td>Repetitive processing</td>
<td>Ad hoc, unstructured, and heuristic processing</td>
</tr>
<tr>
<td>High level of transaction throughput</td>
<td>Medium to low level of transaction throughput</td>
</tr>
<tr>
<td>Predictable pattern of usage</td>
<td>Unpredictable pattern of usage</td>
</tr>
<tr>
<td>Transaction-driven</td>
<td>Analysis driven</td>
</tr>
<tr>
<td>Application-oriented</td>
<td>Subject-oriented</td>
</tr>
<tr>
<td>Supports day-to-day decisions</td>
<td>supports strategic decisions</td>
</tr>
<tr>
<td>Serves large number of clerical/operation users</td>
<td>Serves relatively number of managerial users</td>
</tr>
</tbody>
</table>
Typical architecture of a data warehouse

- **Operational data source**: Input data from operational systems.
- **Warehouse Manager**: Manages the data flow and transformation.
- **Load Manager**: Loads data into the warehouse.
- **Meta-flow**: Meta-data management and transformation.
- **Operational data store (ODS)**: The central repository for data.
- **Warehouse Manager**: Further processing and summarization.
- **High summarized data**: Data summarized for higher-level analysis.
- **Lightly summarized data**: Less summarized data for detailed analysis.
- **Inflow**: Data entering the warehouse.
- **Outflow**: Data exiting the warehouse for reporting or analysis.
- **Upflow**: Data going from the warehouse to the OLAP tools.
- **Downflow**: Data going from the OLAP tools back to the warehouse.
- **Archive/backup data**: Data stored for future reference.
- **DBMS**: Database Management System.
- **OLAP (online analytical processing)** tools: Tools for data analysis.
- **End user access tools**: Tools for end users to access and interact with the data.
- **Data mining tools**: Tools for extracting patterns from data.
- **Reporting, query, application development, and EIS (executive information system) tools**: Tools for reporting and analysis.

**M M U MULLANA**
The Main Components

- **Operational data sources** for the DW is supplied from mainframe operational data held in first generation hierarchical and network databases, departmental data held in proprietary file systems, private data held on workstations and private servers and external systems such as the Internet, commercially available DB, or DB associated with and organization’s suppliers or customers.

- **Operational datastore (ODS)** is a repository of current and integrated operational data used for analysis. It is often structured and supplied with data in the same way as the data warehouse, but may in fact simply act as a staging area for data to be moved into the warehouse.
The Main Components Contd...

- **Load Manager**—also called the *frontend* component, it performs with all the operations associated with the extraction and loading of data into the warehouse. These operations include simple transformations of the data to prepare the data for entry into the warehouse.

- **Warehouse Manager**—performs all the operations associated with the management of the data in the warehouse. The operations performed by this component include analysis of data to ensure consistency, transformation and merging of source data, creation of indexes and views, generation of denormalizations and aggregations, and archiving and backing-up data.
The Main Components Contd...

- **Query Manager**: also called backend component, it performs all the operations associated with the management of user queries. The operations performed by this component include directing queries to the appropriate tables and scheduling the execution of queries.

- Detailed, lightly and highly summarized data, archive / backup data.

- Meta-Data.

- **End-user access tools**: can be categorized into five main groups: data reporting and query tools, application development tools, executive information system (EIS) tools, online analytical processing (OLAP) tools, and data mining tools.
Data Flows

• **Inflow**- The processes associated with the extraction, cleansing, and loading of the data from the source systems into the data warehouse.

• **Upflow**- The process associated with adding value to the data in the warehouse through summarizing, packaging and distribution of the data.

• **Downflow**- The processes associated with archiving and backing up of data in the warehouse.

• **Outflow**- The process associated with making the data available to the end-users.

• **Meta-flow**- The processes associated with the management of the meta-data.
Information flows of a data warehouse

Operational data source

Warehouse Manager

DBMS

Operational data store (ODS)

Load Manager

Upflow

OLAP (online analytical processing) tools

Query Manager

Downflow

End user access tools

Data mining tools

Reporting, query, application development, and EIS (executive information system) tools

Meta-flow

Operational data source n

Inflow

Operational data store (ods)

High summarized data

Lightly summarized data

Detailed data

Outflow

Archive/backup data

Query

Manage

Operational data store (ODS)

M M U MULLANA
Tools and Technologies

• Critical steps in the construction of a Data Warehouse are:
  a. Extraction
  b. Cleansing
  c. Transformation

• Loading the results into target system can be carried out either by separate products, or by a single, categorie(s):
  • code generators
  • database data replication tools
  • dynamic transformation engines
The benefits of data warehousing

- High returns on investment..
- Substantial competitive advantage..
- Increased productivity of corporate decision-makers..
Decision Support Systems

• Created to facilitate the decision making process

• So much information that it is difficult to extract from a traditional database

• Need for a more comprehensive data storage facility
  – Data Warehouse
Decision Support Systems

• Extract Information from data to use as the basis for decision making
• Used at all levels of the Organization
• Tailored to specific business areas
• Interactive
• Ad Hoc queries to retrieve and display information
• Combines historical operation data with business activities
Components of DSS

• Data Store – The DSS Database
  – Business Data
  – Business Model Data
  – Internal and External Data

• Data Extraction and Filtering
  – Extract and validate data from the operational database and the external data sources
Components of DSS

• End-User Query Tool
  – Create Queries that access either the Operational or the DSS database

• End User Presentation Tools
  – Organize and Present the Data
DSS Database Requirements

• DSS Database Scheme
  – Support Complex and Non-Normalized data
    • Summarized and Aggregate data
    • Multiple Relationships
    • Queries must extract multi-dimensional time slices
    • Redundant Data
DSS Database Requirements

• Data Extraction and Filtering
  – DSS databases are created mainly by extracting data from operational databases combined with data imported from external source
    • Need for advanced data extraction & filtering tools
    • Allow batch / scheduled data extraction
    • Support different types of data sources
    • Check for inconsistent data / data validation rules
    • Support advanced data integration / data formatting conflicts
DSS Database Requirements

• End User Analytical Interface
  – Must support advanced data modeling and data presentation tools
  – Data analysis tools
  – Query generation
  – Must Allow the User to Navigate through the DSS

• Size Requirements
  – VERY Large – Terabytes
  – Advanced Hardware (Multiple processors, multiple disk arrays, etc.)
Data Marts

• Small Data Stores
• More manageable data sets
• Targeted to meet the needs of small groups within the organization

• Small, Single-Subject data warehouse subset that provides decision support to a small group of people
OLAP

• Online Analytical Processing Tools
• DSS tools that use multidimensional data analysis techniques
  – Support for a DSS data store
  – Data extraction and integration filter
  – Specialized presentation interface
OLAP

• Need for More Intensive Decision Support
• 4 Main Characteristics
  – Multidimensional data analysis
  – Advanced Database Support
  – Easy-to-use end-user interfaces
  – Support Client/Server architecture
Multidimensional Data Analysis Techniques

• Advanced Data Presentation Functions
  – 3-D graphics, Pivot Tables, Crosstabs, etc.
  – Compatible with Spreadsheets & Statistical packages
  – Advanced data aggregations, consolidation and classification across time dimensions
  – Advanced computational functions
  – Advanced data modeling functions
Advanced Database Support

• Advanced Data Access Features
  – Access to many kinds of DBMS’s, flat files, and internal and external data sources
  – Access to aggregated data warehouse data
  – Advanced data navigation (drill-downs and roll-ups)
  – Ability to map end-user requests to the appropriate data source
  – Support for Very Large Databases
Easy-to-Use End-User Interface

• Graphical User Interfaces
• Much more useful if access is kept simple
Client/Server Architecture

• Framework for the new systems to be designed, developed and implemented

• Divide the OLAP system into several components that define its architecture
  – Same Computer
  – Distributed among several computer
OLAP Architecture

• 3 Main Modules
  – GUI
  – Analytical Processing Logic
  – Data-processing Logic
OLAP Client / Server Architecture

The OLAP system exhibits ...

- Client/Server architecture
- Easy to use GUI
  - Dimensional presentation
  - Dimensional modeling
  - Dimensional analysis
- Multidimensional data
  - Analysis
  - Manipulation
  - Structure
- Database support
  - Data warehouse
  - Operational DB
  - Relational
  - Multidimensional

Data warehouse
- Integrated
- Subject-oriented
- Time-variant
- Nonvolatile

OLAP System

OLAP GUI
Analytical processing logic
Data-processing logic

Operational data
- Drill-down
- Roll-up
- Detailed

- Dimensional
- Aggregated
- Very large DB

M M U MULLANA
Relational OLAP

• Relational Online Analytical Processing
  – OLAP functionality using relational database and familiar query tools to store and analyze multidimensional data

• Multidimensional data schema support

• Data access language & query performance for multidimensional data

• Support for Very Large Databases
Multidimensional Data Schema Support

• Decision Support Data tends to be
  – Nonnormalized
  – Duplicated
  – Preaggregated

• Star Schema
  – Special Design technique for multidimensional data representations
  – Optimize data query operations instead of data update operations
Star Schemas

- Data Modeling Technique to map multidimensional decision support data into a relational database
- Current Relational modeling techniques do not serve the needs of advanced data requirements
Star Schema

• 4 Components
  – Facts
  – Dimensions
  – Attributes
  – Attribute Hierarchies
Facts

• Numeric measurements (values) that represent a specific business aspect or activity
• Stored in a fact table at the center of the star scheme
• Contains facts that are linked through their dimensions
• Can be computed or derived at run time
• Updated periodically with data from operational databases
Dimensions

• Qualifying characteristics that provide additional perspectives to a given fact
  – In DSS data is almost always viewed in relation to other data

• Dimensions are normally stored in dimension tables
Attributes

• Dimension Tables contain Attributes
• Attributes are used to search, filter, or classify facts
• Dimensions provide descriptive characteristics about the facts through their attributes
• Must define common business attributes that will be used to narrow a search, group information, or describe dimensions. (ex.: Time / Location / Product)
• No mathematical limit to the number of dimensions (3-D makes it easy to model)
Attribute Hierarchies

• Provides a Top-Down data organization
  – Aggregation
  – Drill-down / Roll-Up data analysis

• Attributes from different dimensions can be grouped to form a hierarchy
OLTP vs. OLAP

- **On-Line Transaction Processing (OLTP):**
  - technology used to perform updates on operational or transactional systems (e.g., point of sale systems)

- **On-Line Analytical Processing (OLAP):**
  - technology used to perform complex analysis of the data in a data warehouse

  *OLAP is 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 dimensionality of the enterprise as understood by the user.*

  [source: OLAP Council: www.olapcouncil.org]
## OLTP vs. OLAP

<table>
<thead>
<tr>
<th>User</th>
<th>OLTP</th>
<th>OLAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Function</td>
<td>Clerk, IT Professional</td>
<td>Knowledge worker</td>
</tr>
<tr>
<td></td>
<td>Day to day operations</td>
<td>Decision support</td>
</tr>
<tr>
<td>DB Design</td>
<td>Application-oriented (E-R based)</td>
<td>Subject-oriented (Star, snowflake)</td>
</tr>
<tr>
<td>Data View</td>
<td>Current, Isolated</td>
<td>Historical, Consolidated</td>
</tr>
<tr>
<td>Usage</td>
<td>Detailed, Flat relational</td>
<td>Summarized, Multidimensional</td>
</tr>
<tr>
<td>Unit of work</td>
<td>Structured, Repetitive</td>
<td>Ad hoc</td>
</tr>
<tr>
<td>Access</td>
<td>Short, Simple transaction</td>
<td>Complex query</td>
</tr>
<tr>
<td>Operations</td>
<td>Read/write</td>
<td>Read Mostly</td>
</tr>
<tr>
<td># Records accessed</td>
<td>Index/hash on prim. Key</td>
<td>Lots of Scans</td>
</tr>
<tr>
<td># Users</td>
<td>Tens/Hundreds</td>
<td>Millions</td>
</tr>
<tr>
<td>Db size</td>
<td>100 MB-GB</td>
<td>Hundreds</td>
</tr>
<tr>
<td>Metric</td>
<td>Trans. throughput</td>
<td>100GB-TB</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Query throughput, response</td>
</tr>
</tbody>
</table>

**Metric**

- Trans. throughput
- Query throughput, response

**Unit of work**

- Short, Simple transaction
- Complex query

**Access**

- Read/write
- Lots of Scans

**Usage**

- Detailed, Flat relational
- Historical, Consolidated

**Data View**

- Application-oriented (E-R based)
- Summarized, Multidimensional

**DB Design**

- Structured, Repetitive
- Ad hoc

**Function**

- Clerk, IT Professional
- Day to day operations

**User**

- Application-oriented (E-R based)
- Subject-oriented (Star, snowflake)

**OLTP**

- Current, Isolated
- Detailed, Flat relational
- Structured, Repetitive
- Short, Simple transaction
- Read/write
- Index/hash on prim. Key
- Tens/Hundreds
- 100 MB-GB
- Trans. throughput

**OLAP**

- Knowledge worker
- Decision support
- Subject-oriented (Star, snowflake)
- Historical, Consolidated
- Summarized, Multidimensional
- Ad hoc
- Complex query
- Read Mostly
- Lots of Scans
- Millions
- Hundreds
- 100GB-TB
- Query throughput, response

**MMU Mullana**