DATA WAREHOUSE

31.03.2014
(Muscat, Oman)
OUTLINE

Users Problems and necessity for Data Warehouse
BI Definition and Components
Data Warehousing Concepts
DW Goals and Objectives
OLAP and OLTP terms
Data Warehouse vs: Operational DBMS
Datamarts methodology
Explanation of Star Schema and Snowflake Schema
DMQL (Data Mining Query Language)
OLAP Functions
INTRODUCTION

There are 2 types of users: **Operational users** and **Decision-Maker users**

Operational users use **local data** while decision-makers use **historical data**

Database design is changed if data is used for take decision.

Data Warehousing is used for take decisions

Data Warehouses captures data different operational sources

Data Warehouses contain historical data
Operational Data:

- Local data
- Gets frequent updates and queries
- Specific queries are needed

Historical Data:

- "tells" about something
- Very infrequent updates
- Integrated data
- Analytical queries that require huge amounts of aggregation
- Query Performance is crucial
Example OLTP queries:

What is the salary of Mr. Johnson? (point query)

What is address and phone number of Mr. Johnson? (point query)

How many employees have received an 'excellent' credential in the last appraisal?

Example OLAP queries:

Is there a correlation between the geographical location of a company and profit of the company?

How is the age of the employee effect their performance?

Is gender of a staff effect the performance?
Data Problems and necessity for Data Warehouse

Without DW:

Data is everywhere and hard to manage
Same data is exist at different places
Data inconsistency
It is hard to deploy new data
Data is so complex and detailed
Data can not be analysed
There isn’t time series
DECISION SUPPORT SYSTEM and The Origin of BUSINESS INTELLIGENCE

DECISION SUPPORT SYSTEM is a general name of any system to support decision-makers in the decision process.

DSS is the origin of the Business Intelligence system. Business intelligence is designed to support the process of decision-making.
Business intelligence system –
A set of integrated tools, technologies and programmed products used to collect, integrate, analyze, and make data (Koronios & Yeoh, 2010).

The Role of BI in decision making (Olszak & Ziemba, 2007, Figure 2, p.137)
**Business Intelligence Systems:**

Includes all technologies for gathering and analysing data

Provide the input to strategic and tactical decisions at senior managerial levels

Managerial level don’t need daily information, they need historical, strategical data.

Companies invest large amount of money to BI technologies
Query driven Approach and Data Warehouse based

Disadvantages of Query driven Approach:

• Data is up to date, slowly queries because of transactions
• Historical data doesn’t exist
• The Query Driven Approach needs complex integration and filtering processes, aggregation. So it is slow.
• This approach is very inefficient
• This approach is very expensive for frequent queries
• Competes with local processing at sources
Data Warehousing Approach

- This is the approach commonly used in BI systems.
- In this approach; the information from multiple heterogeneous sources is integrated in advance and stored in a warehouse.
- There is another database other than running database.
- The data in the database is stored in a data warehouse in periodically.
- Users don't access database directly, they access Data Warehouse for querying.
- This approach provides high performance.
- Data Warehouse also contains historical data.
BI Components

• Data Warehouse (also called as OLAP systems)

• OLAP Cubes

• Dashboards

• ETL (Extract, Transform, Load)

• Data Mining
Decision-makers use dashboard that contains report, analysis, chart, maps, etc.

- Data Mining
- Take decision
Some Definitions of Data Warehousing

Data warehouse –
A subject oriented, collection of data used to support decision making in organizations (Anderson et al., 2008).

Data warehousing - A systematic approach to collecting relevant business in order to organize and validate the data so that it can be analyzed to support business decision making (Cody et al., 2002).

A Data Warehouse is a subject-oriented, integrated, time-varying, non-volatile collection of data that is used primarily in organizational decision making.”

-- W.H. Inmon, Building the Data Warehouse, 1992
A Data Warehouse is...

Subject-oriented, Organized by subject, not by application

Used for analysis, data mining

Optimized differently from transaction-oriented database

Single repository of information

Data Warehousing involves data cleaning, data integration and data consolidation

Supports analytical reporting, ad-hoc queries and decision making

User interface aimed at executive
Data Marts

A data mart is a subset of the data warehouse. This data is specific to a particular group. Data warehouses are collection of "data marts". Data marts are also seen as small warehouses for OLAP activities. It deals with specific information. For example, although a data warehouse includes the all data of an organization, a data mart includes the data of a department. The data mart is organized for regarding people. So these people don’t need to understand all the data, it is sufficient to understand the regarding data mart. Management and authorization would be easier when using data marts.
Generic Warehouse Architecture

Data Warehouse: A Multi-Tiered Architecture

- Data Sources
- Data Storage
- OLAP Engine
- Front-End Tools

- Other sources
- Operational DBs
- Metadata
- Extract, Transform, Load, Refresh
- Data Warehouse
- Monitor & Integrator
- Data Marts
- OLAP Server
- Serve
- Analysis, Query, Reports, Data mining

Data Sources: Various data sources are integrated into the data warehouse.

Data Storage: Data is stored in the data warehouse, which serves as a central repository.

OLAP Engine: The Online Analytical Processing engine allows for analysis, querying, and reporting on the data.

Front-End Tools: Tools are used to interact with the data warehouse, facilitating analysis and reporting.
Differences Between Operational Systems and Data Warehousing Systems

An **operational database** stores information about the activities of an organization.

The purpose of an operational system is providing online querying and processing, daily routines. Operational systems are also called as **OLTP** (Online Transactional Processing).

**Data Warehouses** help users to data analyze and make decisions.

Warehouse is a Specialized DB. Data Warehousing Systems are also called as **OLAP** (Online Analytical Processing).
## OLAP vs OLTP

<table>
<thead>
<tr>
<th></th>
<th>OLTP</th>
<th>OLAP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>users</strong></td>
<td>clerk, IT professional</td>
<td>knowledge worker</td>
</tr>
<tr>
<td><strong>function</strong></td>
<td>day to day operations</td>
<td>decision support</td>
</tr>
<tr>
<td><strong>DB design</strong></td>
<td>application-oriented (ER)</td>
<td>subject-oriented (star schema)</td>
</tr>
<tr>
<td><strong>data</strong></td>
<td>current, up-to-date</td>
<td>historical, summarized, multidimensional integrated, consolidated</td>
</tr>
<tr>
<td></td>
<td>detailed, flat relational</td>
<td></td>
</tr>
<tr>
<td></td>
<td>isolated</td>
<td></td>
</tr>
<tr>
<td><strong>usage</strong></td>
<td>repetitive</td>
<td>ad-hoc</td>
</tr>
<tr>
<td><strong>access</strong></td>
<td>read/write</td>
<td>lots of scans</td>
</tr>
<tr>
<td></td>
<td>index/hash on prim. key</td>
<td></td>
</tr>
<tr>
<td><strong>unit of work</strong></td>
<td>short, simple transaction</td>
<td>complex query</td>
</tr>
<tr>
<td><strong># records accessed</strong></td>
<td>tens</td>
<td>millions</td>
</tr>
<tr>
<td><strong>#users</strong></td>
<td>thousands</td>
<td>hundreds</td>
</tr>
<tr>
<td><strong>DB size</strong></td>
<td>100MB-GB</td>
<td>100GB-TB</td>
</tr>
<tr>
<td><strong>metric</strong></td>
<td>transaction throughput</td>
<td>query throughput, response</td>
</tr>
</tbody>
</table>
Types of OLAP Servers

• Multidimensional OLAP (MOLAP) : array-based multidimensional storage

• Relational OLAP (ROLAP) : Uses Relational Database. Includes aggregation, additional tools and services.

• Hybrid OLAP (HOLAP) : Both MOLAP and ROLAP

• Specialized SQL Servers
OLAP CUBES

Data cube help us to represent data in multiple dimensions. The data cube is defined by dimensions and facts. The dimensions are the entities with respect to which an enterprise keeps the records.

OLAP cube is the main object of the OLAP. This cube is a multi-dimensional cube. The components of a cube are: fact table, dimension tables and measures.

The cube concept is used to understand multi-dimensional model better. Data cubes are used for people that don’t have advanced database knowledge. The cubes are capable of analyzing data from multiple dimensions. Relational databases are not suitable for very large data. So, OLAP cube is created from these data for an easy analyze. If there are more than 3 dimensions in a cube, it is called hypercube.
OLAP CUBE Example

“For Beef Stew, show me the margin for February”
(Modelling Types) Schema Types in Data Warehousing

Some modelling types:

- Star Schema
- Flat schema
- Terraced Schema
- Snowflake Schema
- Star Cluster Schema
- Fact Constellation Schema, ..

• Data normalization in DW is not as important as OLTP systems
Star Schema
Characteristics of Fact and Dimension Tables

Fact table:
• Contains numeric values that called measurements
• Contains huge size of data
• Expand fast and quickly
• Includes stable, derived, summarized, aggregated data
• Have foreign key relation with dimension tables

Dimension tables:
• Are Reference tables
• Generally includes text type data that users want to see
• Generally includes static data
• Data size is low
Example:

Table 1: 2-D view of sales data according to the dimensions time and item, where location is 'Vancouver'. The measure displayed is dollar in thousand.
Example:

Table 2: 3-D view of sales data according to the dimensions time, item and location. The measure displayed in dollar-sold in thousand.
Example:

3-D data cube representation of the data in table 2
Example:

4-D data cube representation, according to the dimensions time, item, location and supplier
Snowflake Schema

• There is only one fact table
• Some dimension tables are normalized
• Due to normalization,
  redundancy is reduced
  therefore it becomes easy to maintain and save storage space
Snowflake Schema Example
Data Mining Query Language (DMQL)

DMQL is a SQL based language. First, it is developed for data mining process. But, it can be also used for describing data warehouses elements.

SYNTAX FOR CUBE DEFINITION
define cube < cube_name > [ < dimension-list > ]: < measure_list >

SYNTAX FOR DIMENSION DEFINITION
define dimension < dimension_name > as ( < attribute_or_dimension_list > )
DMQL (Data Mining Query Language) Examples

Star Schema of Sales Cube Definitions:

define cube sales star [time, item, branch, location]:

dollars sold = sum(sales in dollars),

define dimension time as (time key, day, month, year)

define dimension item as (item key, item name)

define dimension location as (location key, street, postal code, city)
The Conversion of a Relational Database into a Multidimensional Database

• *Find Subjects and Data Marts*

• *Find all dimensions that exist but are hidden* in a relational database schema

• *Knowledge (measurements)*
OLAP OPERATIONS

• Roll-up : Performs aggregation on a data cube
• Drill-down : reverse of the roll-up
• Slice : Performs selection on a dimension of a cube
• Dice : Performs selection on two or more dimensions
• Pivot (rotate) : It rotates data axes
SUMMARY

• OLAP vs OLTP
• BI components are DW, OLAP, Cubes, Dashboards, ETL and Data Mining
• DW is subject oriented data and used in Decision Support Systems
• DW is also called OLAP
• DW includes time dimension, historical data
• Querying and analysing data is easy on DW system
• Generally managers use DW to query
• Star-Schema structure is used in DW at most (fact, dimensions, measures)
Thank You & Any Question ?