DATA WAREHOUSE AND DATA MINING
IS403

INTRODUCTION TO DATA WAREHOUSE

Alaa Khalaf Hamoud
7-10-2019
1. Introduction
2. Data Warehouse
3. Data Warehousing
4. Operational DS VS DW
5. DW Architecture
6. ETL
7. Data Cube: A Multidimensional Data Model
1. Introduction
1. Introduction

• Data Warehouses **generalize and consolidate** data in multidimensional space.
  – Data cleaning, data integration, and data transformation.

• Data warehouses support **online analytical processing (OLAP)** tools for the interactive analysis of multidimensional data of varied granularities.
1. Introduction

• Data mining functions, such as association rules, classification, prediction, and clustering, can be integrated with OLAP operations.

• Data warehouse has become an increasingly important platform for data analysis and OLAP and will provide an effective platform for data mining.
Contents

1. Introduction
2. Data Warehouse
3. Data Warehousing
4. Operational DS VS DW
5. DW Architecture
6. ETL
7. Data Cube: A Multidimensional Data Model
2. Data Warehouse (DW)

• A data warehouse (DW) is a collection of data marts representing historical data from different operations in the company.
  – Flat, text, numbers and etc.
  – 5 to 10 years of huge amount of data.

• This data is stored in a structure optimized for querying and data analysis as a DW.
2. DW

• But what is data Warehousing?
• A DW refers to a data repository that is maintained separately from an organization’s operational databases.
• Inmon, A DW is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management’s decision making process.
2. **DW-Subject Oriented**

- **DW** is organized around major **subjects** such as **customer, supplier, product, and sales**.
- **Do not** concentrate on the **day-to-day operations and transaction processing of an organization**.
- **Exclude** data that are **not useful in the decision support process**.
2. DW-Integrated

- DW constructed by integrating multiple heterogeneous sources, such as relational databases, flat files, and online transaction records.
  - Data cleaning, ensuring consistency, naming conventions, data structure and data integration.
2. **DW-Time Variant**

- Data are stored to provide information from an **historic perspective** (e.g., the past 5–10 years).

- **Time** element is very **important in the data sources**.
2. DW-Nonvolatile

• DW is always a physically separate store of data transformed from the application data found in the operational environment.

• DW does not require transaction processing, recovery, and concurrency control mechanisms.
  – It requires only two operations: initial loading of data and access of data.
2. DW-Enterprise Warehouse
2. DW-Enterprise Warehouse

• It collects all of the information about subjects spanning the entire organization.

• It provides corporate-wide data integration.

• It contains detailed data as well as summarized data.
2. DW-Enterprise Warehouse

- Size from a few gigabytes to terabytes, or beyond.

- It may be implemented using mainframes, super computer servers, or parallel architecture platforms.
2. **DW-Data Mart**

[Diagram showing a Data Warehouse connected to three Data Marts]

- Data Warehouse
- DB Objects
- Data Mart
- Data Mart
- Data Mart
2. **DW-Data Mart**

- It contains a *subset* of corporate-wide data that is of value to a *specific group of users*.
- It may be implemented using *specific selected subjects*.
- Data marts are usually *implemented on low-cost departmental servers*.
- The data contained in data marts tend to be *summarized*. 
2. DW-Data Mart

- The implementation cycle of a data mart is more likely to be measured in weeks rather than months or years.

- Depending on the source of data, data marts can be categorized as independent or dependent.
2. DW-Data Mart

- Independent data marts are sourced from data captured from one or more operational systems or external information providers, or from data generated locally within a particular department or geographic area.
2. **DW-Data Mart**

- Dependent data marts are sourced directly from enterprise data warehouses.
2. DW-Virtual Warehouse

- A virtual warehouse is a set of views over operational databases.
- For efficient query processing, only some of the possible summary views may be materialized.
- A virtual warehouse is easy to build but requires excess capacity on operational database servers.
Contents

1. Introduction
2. Data Warehouse
3. Data Warehousing
4. Operational DS VS DW
5. DW Architecture
6. ETL
7. Data Cube: A Multidimensional Data Model
3. Data Warehousing
3. Data Warehousing

- A data warehousing is as an architecture constructed by integrating data from multiple heterogeneous sources to support structured and/or ad hoc queries, analytical reporting, and decision making.

- Process of constructing and using data warehouses or process of data warehouse construction.
3. Data Warehousing

- It requires data cleaning, data integration, and data consolidation.
- Warehouse DBMS is used to refer to the management and utilization of data warehouses.
- When a query is posed to a client site, a metadata dictionary is used to translate the query into queries appropriate for the individual heterogeneous sites involved.
3. Data Warehousing

• Approach for queries that requiring aggregations.

• Data warehousing employs an update driven approach rather than query driven approach in order to using information for direct querying and analysis.
Contents

1. Introduction
2. Data Warehouse
3. Data Warehousing
4. Operational DS VS DW
5. DW Architecture
6. ETL
7. Data Cube: A Multidimensional Data Model
4. Operational DB VS DW

• Online operational database systems is to perform online transaction and query processing.

• These systems are called online transaction processing (OLTP) systems.
  – Support day-to-day operations of an organization such as purchasing, banking, payroll and registration.
4. Operational DB VS DW
4. Operational DB VS DW

- DW systems, on the other hand, serve users or knowledge workers in the role of data analysis and decision making.

- These systems are known as online analytical processing (OLAP) systems.
4. Operational DB VS DW

• Users and system orientation.
  – An **OLTP** system is **customer-oriented** and is used for **transaction and query processing** by clerks, clients, and information technology professionals.

  – An **OLAP** system is **market-oriented** and is used for **data analysis** by **knowledge workers**, including **managers**, **executives**, and **analysts**.
4. Operational DB VS DW

• Data contents
  
  – An **OLTP** system manages current data that, typically, are too detailed to be easily used for decision making.
  
  – An **OLAP** system manages large amounts of historic data, provides facilities for summarization and aggregation, and stores and manages information at different levels of granularity.
4. Operational DB VS DW

• Database design
  – An OLTP system usually adopts an entity-relationship (ER) data model and an application-oriented database design.
  – An OLAP system typically adopts either a star or a snowflake model and a subject-oriented database design.
4. Operational DB VS DW

• View

  – An **OLTP** system focuses mainly on the **current data** within an **enterprise or department**, **without** referring to **historic data or data in different organizations**.

  – **OLAP** system often spans **multiple versions** of a database schema, due to the evolutionary process of an organization. Because of their huge volume, **OLAP data** are stored on **multiple storage media**.
4. Operational DB VS DW

• Access patterns

  – The access patterns of an OLTP system consist mainly of short, atomic transactions. Such a system requires concurrency control and recovery mechanisms.

  – Access to OLAP systems are mostly read-only operations (because most data warehouses store historic rather than up-to-date information), although many could be complex queries.
## 4. Operational DB VS DW

<table>
<thead>
<tr>
<th>Feature</th>
<th><strong>OLTP</strong></th>
<th><strong>OLAP</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Characteristic</strong></td>
<td>operational processing</td>
<td>informational processing analysis</td>
</tr>
<tr>
<td><strong>Orientation</strong></td>
<td>transaction</td>
<td>knowledge worker (e.g., manager, executive, analyst)</td>
</tr>
<tr>
<td><strong>User</strong></td>
<td>clerk, DBA, database professional</td>
<td>long-term informational requirements decision support</td>
</tr>
<tr>
<td><strong>Function</strong></td>
<td>day-to-day operations</td>
<td>star/snowflake, subject-oriented</td>
</tr>
<tr>
<td><strong>DB design</strong></td>
<td>ER-based, application-oriented</td>
<td>historic, accuracy maintained over time</td>
</tr>
<tr>
<td><strong>Data</strong></td>
<td>current, guaranteed up-to-date</td>
<td>summarized, consolidated</td>
</tr>
<tr>
<td><strong>Summarization</strong></td>
<td>primitive, highly detailed</td>
<td>summarized, multidimensional</td>
</tr>
<tr>
<td><strong>View</strong></td>
<td>detailed, flat relational</td>
<td>complex query</td>
</tr>
<tr>
<td><strong>Unit of work</strong></td>
<td>short, simple transaction</td>
<td>mostly read</td>
</tr>
<tr>
<td><strong>Access</strong></td>
<td>read/write</td>
<td>information out</td>
</tr>
<tr>
<td><strong>Focus</strong></td>
<td>data in</td>
<td>lots of scans</td>
</tr>
<tr>
<td><strong>Operations</strong></td>
<td>index/hash on primary key</td>
<td></td>
</tr>
<tr>
<td><strong>Number of records accessed</strong></td>
<td>tens</td>
<td>millions</td>
</tr>
<tr>
<td><strong>Number of users</strong></td>
<td>thousands</td>
<td>hundreds</td>
</tr>
<tr>
<td><strong>DB size</strong></td>
<td>GB to high-order GB</td>
<td>≥ TB</td>
</tr>
<tr>
<td><strong>Priority</strong></td>
<td>high performance, high availability</td>
<td>high flexibility, end-user autonomy</td>
</tr>
<tr>
<td><strong>Metric</strong></td>
<td>transaction throughput</td>
<td>query throughput, response time</td>
</tr>
</tbody>
</table>
Contents

1. Introduction
2. Data Warehouse
3. Data Warehousing
4. Operational DS VS DW
5. DW Architecture
6. ETL
7. Data Cube: A Multidimensional Data Model
5. DW Architecture

Top tier: Front-end tools

Middle tier: OLAP server

Bottom tier: Data warehouse server

Data

Operational databases

External sources

Extract Clean
Transform Load Refresh

Data warehouse

OLAP server

Metadata repository

OLAP server

Query/report

Analysis

Data mining

Output

Monitoring

Administration
5. DW Architecture-bottom tier

• The warehouse database server that is almost always a relational database system.

• Back-end tools and utilities are used to feed data into the bottom tier from different data sources.

• These tools and utilities perform data extraction, cleaning, and transformation.
  – (e.g., to merge similar data from different sources into a unified format), as well as load and refresh functions to update the data warehouse.
5. DW Architecture-bottom tier

- The data are extracted using application program interfaces known as gateways.
- Examples of gateways include ODBC (Open Database Connection) and OLEDB (Object Linking and Embedding Database) by Microsoft and JDBC (Java Database Connection).
- This tier also contains a metadata repository, which stores information about the data warehouse and its contents.
5. DW Architecture-bottom tier

• It contains data mart and Meta data repository.

• Meta data is data about data.

• Two categories of meta data: technical and business meta data.

  – Technical Meta data: It contains information about data warehouse data used by warehouse designer, administrator to carry out development and management tasks.
5. DW Architecture-bottom tier

- **Technical Meta data**: examples (table creation time, data types, file size and compression schema).

- **Business Meta data**: It contains info that gives info related to business stored in data warehouse to users.

- Examples *(privacy level, security level and business rules).*
5. DW Architecture-Tech. Meta

- Info about data stores.
- Transformation descriptions.
- Warehouse Object and data structure definitions for target data.
- The rules used to perform clean up, and data enhancement.
- Data mapping operations.
- Access authorization, backup history, archive history, info delivery history, data acquisition history, data access etc.,
5. DW Architecture-Buss. Meta

• Subject areas, and info object type including queries, reports, images, video, audio clips etc.
• Internet home pages.
• Info related to info delivery system.
• Data warehouse operational info such as ownerships, audit trails etc.
5. DW Architecture-Meta Charac.

- It is the **gateway** to the data warehouse environment.
- It supports easy **distribution and replication** of content for **high performance and availability**.
- It should be **searchable** by business oriented key words.
- It should act as a **launch platform** for the end user to access data and analysis tools.
5. DW Architecture-Meta Charac.

- Should support
  - Sharing of info.
  - Scheduling options for request.
  - Providing interface to other applications.
  - End user monitoring of the status of the data warehouse environment.
5. DW Architecture-middle tier

- An **OLAP server** that is typically implemented using either
  - (1) a **relational OLAP (ROLAP)** model (i.e., an extended relational DBMS that maps operations on multidimensional data to standard relational operations).
  - or (2) a **multidimensional OLAP (MOLAP)** model (i.e., a special-purpose server that directly implements multidimensional data and operations).
5. DW Architecture- top tier

- The top tier is a front-end client layer, which contains query and reporting tools, analysis tools, and/or data mining tools (e.g., trend analysis, prediction, and so on).
Contents

1. Introduction
2. Data Warehouse
3. Data Warehousing
4. Operational DS VS DW
5. DW Architecture
6. ETL
7. Data Cube: A Multidimensional Data Model
6. Extract, Transform, and Load (ETL)
6. **ETL**

- Consists of a **work area**, instantiated **data structures**, and a **set of processes**.
- Is everything **between the operational source systems** and the **DW/BI presentation area**.
- **Extraction** is the first step in the **process of getting data into the data warehouse environment**.
6. ETL

- **Data extraction**, which typically gathers data from **multiple, heterogeneous, and external sources**.

- Extracting means **reading and understanding** the source data and **copying** the data needed into the ETL system for further manipulation.

- At this point, the **data belongs to the data warehouse**.
6. ETL

- After the data is extracted to the ETL system, there are numerous potential transformations, such as cleansing the data (correcting misspellings, resolving domain conflicts, dealing with missing elements, or parsing into standard formats), combining data from multiple sources, and de-duplicating data.
6. ETL

- Data cleaning which detects errors in the data and rectifies them when possible.
- Data transformation, which converts data from legacy or host format to warehouse format.
- The ETL system adds value to the data with these cleansing and conforming tasks by changing the data and enhancing it.
6. ETL

• These activities can be architected to create diagnostic metadata to improve data quality in the source systems over time.

• The final step ETL is the physical structuring and loading of data into the presentation area’s target dimensional models.

• Load, which sorts, summarizes, consolidates, computes views, checks integrity, and builds indices and partitions.
6. ETL

- **Refresh**, which propagates the updates from the data sources to the warehouse.

- These subsystems are critical due to the primary mission of the ETL system is to **hand off** the dimension and fact tables.
6. ETL

• Focus on dimension table processing
  – Surrogate key assignments, code lookups to provide appropriate descriptions, splitting, or combining columns to present the appropriate data values, or joining underlying third normal form table structures into flattened de-normalized dimensions.

• In many cases, the ETL is not based on relational technology but instead on a system of flat files.
6. ETL

- After data validation for conformance with 1-1 and M-M business rules, does it need to build 3NF physical database?
- When the data arrives at the doorstep of the ETL system in a 3NF relational format, the ETL system developers may be more comfortable performing the cleansing and transformation tasks using normalized structures.
6. ETL

- The creation of both normalized structures for the ETL and dimensional structures for presentation means that the data is potentially extracted, transformed, and loaded twice, once into the normalized database and then again when you load the dimensional model.

- It is wrong to focus only on constructing normalized structure than developing dimensional presentation.
Contents

1. Introduction
2. Data Warehouse
3. Data Warehousing
4. Operational DS VS DW
5. DW Architecture
6. ETL
7. Data Cube: A Multidimensional Data Model
7. Data Cube: A Multidimensional Data Model

• A data cube allows data to be modeled and viewed in multiple dimensions.

• It is defined by dimensions and facts.

• **Dimensions** are the perspectives or entities with respect to which an organization wants to keep records.
A multidimensional data model is typically organized around a central theme, such as sales.

This theme is represented by a fact table.

Facts are numeric measures.
7. **Data Cube: A Multidimensional Data Model**

- **2-D Cube**

<table>
<thead>
<tr>
<th>time (quarter)</th>
<th>location = “Vancouver”</th>
<th>item (type)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>home entertainment</td>
<td>computer</td>
</tr>
<tr>
<td>Q1</td>
<td>605</td>
<td>825</td>
</tr>
<tr>
<td>Q2</td>
<td>680</td>
<td>952</td>
</tr>
<tr>
<td>Q3</td>
<td>812</td>
<td>1023</td>
</tr>
<tr>
<td>Q4</td>
<td>927</td>
<td>1038</td>
</tr>
</tbody>
</table>

*Note: The sales are from branches located in the city of Vancouver. The measure displayed is dollars_sold (in thousands).*
7. Data Cube: A Multidimensional Data Model

![Data Cube Diagram]

- **Location (Cities)**: Chicago, New York, Toronto, Vancouver
- **Time (Quarters)**: Q1, Q2, Q3, Q4
- **Item (Types)**: Computer, Security, Home, Entertainment, Phone
7. Data Cube: A Multidimensional Data Model

- 4-D Cube
7. Data Cube: A Multidimensional Data Model

- The cuboid that holds the lowest level of summarization is called the base cuboid.

- The 0-D cuboid, which holds the highest level of summarization, is called the apex cuboid.
7. Data Cube: A Multidimensional Data Model
End of Introduction to DW