Data Mining
Data warehousing

Hamid Beigy
Sharif University of Technology
Fall 1396
1 Introduction

2 Data warehousing concepts

3 Schemas for multidimensional data models

4 OLAP server architectures

5 Reading
Table of contents

1 Introduction

2 Data warehousing concepts

3 Schemas for multidimensional data models

4 OLAP server architectures

5 Reading
Data warehouses generalize and consolidate data in multidimensional space.

Construction of data warehouses involves data cleaning, data integration, and data transformation.

Data warehouses provide online analytical processing (OLAP) tools for interactive analysis of multidimensional data of varied granularities, which facilitates effective data mining.

Data mining functions such as clustering, classification, and associative rule mining can be integrated with OLAP functions to enhance interactive data mining.

As a conclusion, data warehousing form an essential step in knowledge discovery process.
<table>
<thead>
<tr>
<th></th>
<th>Introduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td><strong>Data warehousing concepts</strong></td>
</tr>
<tr>
<td>3</td>
<td>Schemas for multidimensional data models</td>
</tr>
<tr>
<td>4</td>
<td>OLAP server architectures</td>
</tr>
<tr>
<td>5</td>
<td>Reading</td>
</tr>
</tbody>
</table>
What is a data warehouse?

- A data warehouse is a **subject-oriented, integrated, time-variant, and nonvolatile** collection of data in support of management decision making process. (William H. Inmon)

The following **keywords** distinguish data warehouse from other data repository systems such as relational database systems.

**Subject-oriented**  A data warehouse is organized around major subjects such as customer, supplier, product, and sales.

**Integrated**  A data warehouse is usually constructed by integrating multiple heterogeneous sources, such as relational databases, flat files, and online transaction records.

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

**Nonvolatile**  A data warehouse does not require transaction processing, recovery, and concurrency control mechanisms. It usually requires only two operations in data accessing: initial loading of data and access of data.
### Differences between operational databases and data warehouses

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

Data warehouses often adopt a three-tier architecture, as presented below.
Data warehouse models

- From the architecture point of view, there are three data warehouse models

  **Enterprise warehouse**  An enterprise warehouse collects all of the information about subjects spanning the entire organization.

  **Data mart**  Data mart contains a subset of corporate-wide data that is of value to a specific group of users. The scope is confined to specific selected subjects.

  **Virtual warehouse**  A virtual warehouse is a set of views over operational databases.

- There are two approaches for constructing data warehouse: top-down and bottom-up approaches.

- What are the pros and cons of the top-down and bottom-up approaches to data warehouse development?

  - The top-down development of an enterprise warehouse serves as a systematic solution and minimizes integration problems. However, it is expensive, takes a long time to develop, and lacks flexibility due to the difficulty in achieving consistency and consensus for a common data model for the entire organization.

  - The bottom-up approach to the design, development, and deployment of independent data marts provides flexibility, low cost, and rapid return of investment. It, however, can lead to problems when integrating various disparate data marts into a consistent enterprise data warehouse.
Data warehouse systems use back-end tools and utilities to populate and refresh their data. These tools and utilities include the following functions:

- **Data extraction** This typically gathers data from multiple, heterogeneous, and external sources.
- **Data cleaning** This detects errors in the data and rectifies them when possible.
- **Data transformation** This converts data from legacy or host format to warehouse format.
- **Load** This sorts, summarizes, consolidates, computes views, checks integrity, and builds indices and partitions.
- **Refresh** This propagates the updates from the data sources to the warehouse.

Besides the above functions, data warehouse systems usually provide a good set of data warehouse management tools.
Metadata repository

- Metadata are data about data. When used in a data warehouse, metadata are the data that define warehouse objects.

- A metadata repository should contain the following:
  - **A description of the data warehouse structure** including the warehouse schema, view, dimensions, hierarchies, and derived data definitions, as well as data mart locations and contents.
  - **Operational metadata** such as history of migrated data and the sequence of transformations applied to it and monitoring information (warehouse usage statistics, error reports, and audit trails).
  - **The algorithms used for summarization** including measure and dimension definition algorithms, data on granularity, partitions, subject areas, aggregation, summarization, and predefined queries and reports.
  - **Mapping from the operational environment to the data warehouse** including source databases and their contents, gateway descriptions, data partitions, data extraction, cleaning, transformation rules and defaults, data refresh and purging rules, and user authorization and access control.
  - **Data related to system performance** including indices and profiles that improve data access and retrieval performance, in addition to rules for the timing and scheduling of refresh, update, and replication cycles.
  - **Business metadata** including business terms and definitions, data ownership information, and charging policies.
Data warehouse modeling

- Data warehouses and OLAP tools are based on a multidimensional data model.
- This model views data in the form of a data cube.
- What is a data cube?
  - 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.
  - Each dimension may have a table associated with it, called a dimension table, which further describes the dimension.
  - Dimension tables can be specified by users or experts, or automatically generated and adjusted based on data distributions.
  - A multidimensional data model is typically organized around a central theme represented by a fact table. Facts are numeric measures.
  - Fact table contains the names of the facts, or measures, as well as keys to each of the related dimension tables.
1.3 What Kinds of Data Can Be Mined?

1.3.1 Database Data

A database is a large set of related data organized for storage and retrieval. A database stores information, including a unique customer identity number (cust_ID), name, address, age, occupation, annual_income, credit_information, category, ...).

- **item** (item_ID, brand, category, type, price, place_made, supplier, cost, ...)
- **employee** (empl_ID, name, category, group, salary, commission, ...)
- **branch** (branch_ID, name, address, ...)
- **purchases** (trans_ID, cust_ID, empl_ID, date, time, method_paid, amount)
- **items_sold** (trans_ID, item_ID, qty)
- **works_at** (empl_ID, branch_ID)

2-D view of sales data

<table>
<thead>
<tr>
<th>time (quarter)</th>
<th>home entertainment</th>
<th>computer</th>
<th>phone</th>
<th>security</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>605</td>
<td>825</td>
<td>14</td>
<td>400</td>
</tr>
<tr>
<td>Q2</td>
<td>680</td>
<td>952</td>
<td>31</td>
<td>512</td>
</tr>
<tr>
<td>Q3</td>
<td>812</td>
<td>1023</td>
<td>30</td>
<td>501</td>
</tr>
<tr>
<td>Q4</td>
<td>927</td>
<td>1038</td>
<td>38</td>
<td>580</td>
</tr>
</tbody>
</table>

location = “Vancouver”
### 3-D view of sales data

<table>
<thead>
<tr>
<th>location = “Chicago”</th>
<th>location = “New York”</th>
<th>location = “Toronto”</th>
<th>location = “Vancouver”</th>
</tr>
</thead>
<tbody>
<tr>
<td>item</td>
<td>item</td>
<td>item</td>
<td>item</td>
</tr>
<tr>
<td>home</td>
<td>home</td>
<td>home</td>
<td>home</td>
</tr>
<tr>
<td>time</td>
<td>time</td>
<td>time</td>
<td>time</td>
</tr>
<tr>
<td>ent.</td>
<td>ent.</td>
<td>ent.</td>
<td>ent.</td>
</tr>
<tr>
<td>comp.</td>
<td>comp.</td>
<td>comp.</td>
<td>comp.</td>
</tr>
<tr>
<td>phone</td>
<td>phone</td>
<td>phone</td>
<td>phone</td>
</tr>
<tr>
<td>sec.</td>
<td>sec.</td>
<td>sec.</td>
<td>sec.</td>
</tr>
<tr>
<td>Q1</td>
<td>854 882 89 623</td>
<td>1087 968 38 872</td>
<td>818 746 43 591</td>
</tr>
<tr>
<td>Q2</td>
<td>943 890 64 698</td>
<td>1130 1024 41 925</td>
<td>894 769 52 682</td>
</tr>
<tr>
<td>Q3</td>
<td>1032 924 59 789</td>
<td>1034 1048 45 1002</td>
<td>940 795 58 728</td>
</tr>
<tr>
<td>Q4</td>
<td>1129 992 63 870</td>
<td>1142 1091 54 984</td>
<td>978 864 59 784</td>
</tr>
</tbody>
</table>

Note: The measure displayed is dollars sold (in thousands).
A 3-D data cube representation of the data (previous slide)

Figure 4.3

A 3-D data cube representation of the data in Table 4.3, according to time, item, and location. The measure displayed is dollars sold (in thousands).

Figure 4.4

A 4-D data cube representation of sales data, according to time, item, location, and supplier. The measure displayed is dollars sold (in thousands). For improved readability, only some of the cube values are shown.

and location, summarized for all suppliers. The 0-D cuboid, which holds the highest level of summarization, is called the apex cuboid. In our example, this is the total sales, or dollars sold, summarized over all four dimensions. The apex cuboid is typically denoted by all.
A 4-D data cube representation of the data

- Figure 4.3: A 3-D data cube representation of the data in Table 4.3, according to time, item, and location. The measure displayed is dollars sold (in thousands).
- Figure 4.4: A 4-D data cube representation of sales data, according to time, item, location, and supplier. The measure displayed is dollars sold (in thousands). For improved readability, only some of the cube values are shown.

For example, consider the sales data for Q1 in Chicago for home entertainment. The measure displayed is dollars sold (in thousands).
In the data warehousing, a data cube like those shown in (previous slides) is often referred to as a **cuboid**.

Given a set of dimensions, we can generate a cuboid for each of the possible subsets of the given dimensions.

The result would form a **lattice of cuboids**, each showing the data at a **different level of summarization**, or group-by.

The **lattice of cuboids** is then referred to as a **data cube**.

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**.
4.2 Data Warehouse Modeling: Data Cube and OLAP

The entity-relationship data model is commonly used in the design of relational databases, where a database schema consists of a set of entities and the relationships between them. Such a data model is appropriate for online transaction processing. A data warehouse, however, requires a concise, subject-oriented schema that facilitates online data analysis.

The most popular data model for a data warehouse is a multidimensional model, which can exist in the form of a star schema, a snowflake schema, or a fact constellation schema. Let's look at each of these.

Star schema:

The most common modeling paradigm is the star schema, in which the data warehouse contains (1) a large central table (fact table) containing the bulk of the data, with no redundancy, and (2) a set of smaller attendant tables (dimension tables), one for each dimension. The schema graph resembles a starburst, with the dimension tables displayed in a radial pattern around the central fact table.

Example 4.1 Star schema. A star schema for AllElectronics sales is shown in Figure 4.6. Sales are considered along four dimensions: time, item, branch, and location. The schema contains a central fact table for sales that contains keys to each of the four dimensions, along with the data. The schema is designed to allow efficient querying and analysis of the sales data.
1 Introduction

2 Data warehousing concepts

3 Schemas for multidimensional data models

4 OLAP server architectures

5 Reading
The entity-relationship data model is commonly used in the design of relational databases, where a database schema consists of a set of entities and the relationships between them.

The entity-relationship data model is appropriate for online transaction processing.

A data warehouse, however, requires a concise, subject-oriented schema that facilitates online data analysis.

The most popular data model for a data warehouse is a multidimensional model, which can exist in the form of a

- Star schema
- Snowflake schema
- Galaxy schema
The most common modeling paradigm is the **star schema**, in which the data warehouse contains:

1. A large central table (fact table) containing the bulk of the data, with no redundancy.
2. A set of smaller attendant tables (dimension tables), one for each dimension.

The schema graph resembles a starburst, with the dimension tables displayed in a radial pattern around the central fact table.

![Star schema diagram]

---

**Figure 4.6** Star schema of sales data warehouse.

With two measures: dollars sold and units sold. To minimize the size of the fact table, dimension identifiers (e.g., time_key and item_key) are system-generated identifiers.

Notice that in the star schema, each dimension is represented by only one table, and each table contains a set of attributes. For example, the location dimension table contains the attribute set `{location_key, street, city, province_or_state, country}`. This constraint may introduce some redundancy. For example, "Urbana" and "Chicago" are both cities in the state of Illinois, USA. Entries for such cities in the location dimension table will create redundancy among the attributes province_or_state and country; that is, `..., Urbana, IL, USA` and `..., Chicago, IL, USA`.

Moreover, the attributes within a dimension table may form either a hierarchy (total order) or a lattice (partial order).

**Snowflake schema:**

The snowflake schema is a variant of the star schema model, where some dimension tables are normalized, thereby further splitting the data into additional tables. The resulting schema graph forms a shape similar to a snowflake.

The major difference between the snowflake and star schema models is that the dimension tables of the snowflake model may be kept in normalized form to reduce redundancies. Such a table is easy to maintain and saves storage space. However, this space savings is negligible in comparison to the typical magnitude of the fact table. Furthermore, the snowflake structure can reduce the effectiveness of browsing, since more joins will be needed to execute a query. Consequently, the system performance may be adversely impacted. Hence, although the snowflake schema reduces redundancy, it is not as popular as the star schema in data warehouse design.
In star schema, each dimension is represented by only one table, and each table contains a set of attributes.

This constraint may introduce some redundancy.

The attributes within a dimension table may form either a hierarchy (total order) or a lattice (partial order).

The snowflake schema is a variant of the star schema model, where some dimension tables are normalized, thereby further splitting the data into additional tables.

The resulting schema graph forms a shape similar to a snowflake.

The major difference between snowflake and star schema models is that the dimension tables of the snowflake model may be kept in normalized form to reduce redundancies.

Tables in snowflake schema are easy to maintain and save storage space.

This space savings is negligible in comparison to the typical magnitude of the fact table.

The snowflake structure can reduce the effectiveness of browsing, since more joins will be needed to execute a query.

The snowflake schema reduces redundancy, it is not as popular as the star schema in data warehouse design.
An example of snowflake scheme

- time
  - Dimension table
  - time_key
  - day
  - day_of_week
  - month
  - quarter
  - year

- sales
  - Fact table
  - time_key
  - item_key
  - branch_key
  - location_key
  - dollars_sold
  - units_sold

- item
  - Dimension table
  - item_key
  - item_name
  - brand
  - type
  - supplier_key

- supplier
  - Dimension table
  - supplier_key
  - supplier_type

- branch
  - Dimension table
  - branch_key
  - branch_name
  - branch_type

- location
  - Dimension table
  - location_key
  - street
  - city_key

- city
  - Dimension table
  - city_key
  - city
  - province_or_state
  - country

Figure 4.7
Snowflake schema of a sales data warehouse.

Example 4.2
Snowflake schema.
A snowflake schema for allElectronics sales is given in Figure 4.7. Here, the sales fact table is identical to that of the star schema in Figure 4.6. The main difference between the two schemas is in the definition of dimension tables. The single dimension table for item in the star schema is normalized in the snowflake schema, resulting in new item and supplier tables. For example, the item dimension table now contains the attributes item key, item name, brand, type, and supplier key, where supplier key is linked to the supplier dimension table, containing supplier key and supplier type information. Similarly, the single dimension table for location in the star schema can be normalized into two new tables: location and city. The city key in the new location table links to the city dimension. Notice that, when desirable, further normalization can be performed on province or state and country in the snowflake schema shown in Figure 4.7.

Fact constellation:
Sophisticated applications may require multiple fact tables to share dimension tables. This kind of schema can be viewed as a collection of stars, and hence is called a galaxy schema or a fact constellation.

Example 4.3
Fact constellation.
A fact constellation schema is shown in Figure 4.8. This schema specifies two fact tables, sales and shipping. The sales table definition is identical to that of the star schema (Figure 4.6). The shipping table has five dimensions, or keys—item key, time key, shipper key, from location, and to location—and two measures—dollars cost.
- Sophisticated applications may require **multiple fact tables** to share dimension tables.
- This kind of schema can be viewed as a collection of stars, and hence is called a **galaxy schema** or a **fact constellation**.
A concept hierarchy defines a sequence of mappings from a set of low-level concepts to higher-level, more general concepts.
Concept hierarchies may also be defined by discretizing or grouping (following figure) values for a given dimension or attribute.

There may be more than one concept hierarchy for a given attribute or dimension, based on different user viewpoints.

Concept hierarchies may be provided manually by system users, domain experts, or knowledge engineers, or may be automatically generated based on statistical analysis of the data distribution.
A data cube measure is a numeric function that can be evaluated at each point in the data cube space.

A measure value is computed for a given point by aggregating the data corresponding to the respective dimension value pairs defining the given point.

Measures can be organized into three categories based on the kind of aggregate functions.

1. **Distributive** An aggregate function is distributive if it can be computed in a distributed manner. For example, `count()`, `min()`, and `max()` are distributive aggregate functions. (appropriate for large data cube)

2. **Algebraic** An aggregate function is algebraic if it can be computed by an algebraic function with M arguments (where M is a bounded positive integer), each of which is obtained by applying a distributive aggregate function. For example, `avg()` can be computed by `sum()/count()`, where both `sum()` and `count()` are distributive aggregate functions. (appropriate for large data cube)

3. **Holistic** An aggregate function is holistic if there is no constant bound on the storage size needed to describe a subaggregate. That is, there does not exist an algebraic function with M arguments (where M is a constant) that characterizes the computation. For example, `median()`, `mode()`, and `rank()`. A measure is holistic if it is obtained by applying a holistic aggregate function. (There are some techniques to approximate the computation of some holistic measures)
In the multidimensional model, data are organized into **multiple dimensions**, and each dimension contains multiple levels of abstraction defined by concept hierarchies.

This organization provides users with the flexibility to view data from **different views**.

A number of **OLAP operations** exist to materialize these different views.

- **Roll-up**  The roll-up operation performs aggregation on a data cube.
A number of OLAP operations exist to materialize these different views.

- **Drill-down** Drill-down is the reverse of roll-up.
A number of OLAP operations exist to materialize these different views.

- **Slice** The slice operation performs a selection on one dimension of the given cube, resulting in a subcube.
A number of **OLAP operations** exist to materialize these different views.

- **Dice** The dice operation defines a subcube by performing a selection on two or more dimensions.
A number of **OLAP operations** exist to materialize these different views.

- **Pivot (rotate)** Pivot is a visualization operation that rotates the data axes in view to provide an alternative data presentation.

---

**OLAP specific operation**
Introduction

Data warehousing concepts

Schemas for multidimensional data models

OLAP server architectures

Reading
Implementations of a warehouse server for OLAP processing include the following:

1. **Relational OLAP (ROLAP) servers**  These are the intermediate servers that stand in between a relational back-end server and client front-end tools. They use a relational or extended-relational DBMS to store and manage warehouse data, and OLAP middleware to support missing pieces.
   
   **Example:** DSS server (Microstrategy) and MetaCube (Informix)

2. **Multidimensional OLAP (MOLAP) servers**  These servers support multidimensional data views through array-based multidimensional storage engines. They map multi-dimensional views directly to data cube array structures.
   
   **Example:** Oracle Database OLAP Option and Microsoft Analysis Services

3. **Hybrid OLAP (HOLAP) servers**  The hybrid OLAP approach combines ROLAP and MOLAP technology, benefiting from the greater scalability of ROLAP and the faster computation of MOLAP.
   
   **Example:** Microsoft SQL Server 2000

4. **Specialized SQL servers**  Some database system vendors implement specialized SQL servers that provide advanced query language and query processing support for SQL queries over star and snowflake schemas in a read-only environment.
Read chapter 4 of the following book
J. Han, M. Kamber, and Jian Pei, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2012.