Data Warehouse and OLAP

• **Data warehouses** generalize and consolidate data in multidimensional space.
  – The construction of data warehouses involves data cleaning, data integration, and data transformation and can be viewed as an important preprocessing step for data mining.

• Data warehouses provide **on-line analytical processing (OLAP)** tools for the interactive analysis of multidimensional data of varied granularities, which facilitates effective data generalization and data mining.
  – Many other data mining functions, such as association, classification, prediction, and clustering, can be integrated with OLAP operations to enhance interactive mining of knowledge at multiple levels of abstraction.
What is a Data Warehouse?

• Data warehousing provides architectures and tools for business executives to systematically organize, understand, and use their data to make strategic decisions.
  – A decision support database that is maintained **separately** from the organization’s operational database
  – Support **information processing** by providing a solid platform of consolidated, historical data for analysis.

• “A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management’s decision-making process.”

• Data warehousing:
  – The process of constructing and using data warehouses
Data Warehouse - Subject-Oriented

• Organized around major subjects, such as customer, product, sales

• Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing

• Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process
Data Warehouse - Integrated

• Constructed by integrating multiple, heterogeneous data sources
  – relational databases, flat files, on-line transaction records

• Data cleaning and data integration techniques are applied.
  – Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources
    • E.g., Hotel price: currency, tax, breakfast covered, etc.
  – When data is moved to the warehouse, it is converted.
Data Warehouse - Time Variant

• The time horizon for the data warehouse is significantly longer than that of operational systems
  – Operational database: current value data
  – Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)

• Every key structure in the data warehouse
  – Contains an element of time, explicitly or implicitly
  – But the key of operational data may or may not contain “time element”
Data Warehouse - Nonvolatile

• A physically separate store of data transformed from the operational environment

• Operational update of data does not occur in the data warehouse environment
  – Does not require transaction processing, recovery, and concurrency control mechanisms
  – Requires only two operations in data accessing:
    • initial loading of data and access of data
Operational Database Systems and Data Warehouses

• The major task of on-line operational database systems is to perform on-line transaction and query processing.
  – These systems are called on-line transaction processing (OLTP) systems.
  – They cover most of the day-to-day operations of an organization, such as purchasing, inventory, banking, payroll, registration, and accounting.

• Data warehouse systems, on the other hand, serve users or knowledge workers in the role of data analysis and decision making.
  – Such systems can organize and present data in various formats in order to accommodate the diverse needs of the different users.
  – These systems are known as on-line analytical processing (OLAP) systems.
OLTP vs. OLAP

- **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.

- **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 historical data, provides facilities for summarization and aggregation.

- **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 snowflake model and a *subject oriented* database design.

- **View**: An OLTP system focuses mainly on the current data. In contrast, an OLAP system often spans multiple versions of a database schema, due to the evolutionary process of an organization.

- **Access patterns**: The access patterns of an OLTP system consist mainly of short, atomic transactions. Such a system requires concurrency control and recovery mechanisms. However, accesses to OLAP systems are mostly read-only operations.
Why a Separate Data Warehouse?

• High performance for both systems
  – DBMS—tuned for OLTP: access methods, indexing, concurrency control, recovery
  – Warehouse—tuned for OLAP: complex OLAP queries, multidimensional view, consolidation

• Different functions and different data:
  – **missing data**: Decision support requires historical data which operational DBs do not typically maintain
  – **data consolidation**: DS requires consolidation (aggregation, summarization) of data from heterogeneous sources
  – **data quality**: different sources typically use inconsistent data representations, codes and formats which have to be reconciled

• Note: There are more and more systems which perform OLAP analysis directly on relational databases
Multidimensional Data Model: Data Cube

- Data warehouses and OLAP tools are based on a multidimensional data model.
- This model views data in the form of 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. For example, a dimension table for item may contain the attributes item name, brand, and type.
  - Facts are numerical measures.
    - Examples of facts for a sales data warehouse include dollars sold units sold
Data Cube: A 2-D Data Cube

- Although we usually think of cubes as 3-D geometric structures, in data warehousing the data cube is n-dimensional.
- A 2-D data cube:
  - **dimensions** time and item, the measure displayed (**fact**) is dollars_sold3.2.
  - In this 2-D representation, the sales for Vancouver are shown with
  - respect to the time dimension (organized in quarters) and the item dimension (organized according to the types of items sold).

<table>
<thead>
<tr>
<th>time (quarter)</th>
<th>location = “Vancouver”</th>
<th>item (type)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>home</strong></td>
<td><strong>computer</strong></td>
</tr>
<tr>
<td></td>
<td><strong>entertainment</strong></td>
<td></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>
Data Cube: A 3-D Data Cube

• A 3-D data cube representation of the data according to the dimensions time, item, and location. The measure displayed is dollars_sold.
Data Cube: A Lattice of Cuboids

- A lattice of cuboids, making up a 4-D data cube for the dimensions time, item, location, and supplier.
  - Each cuboid represents a different degree of summarization.

- The cuboid that holds the lowest level of summarization is called the base cuboid.
  - the base cuboid for the given time, item, location, and supplier dimensions

- A 3-D (nonbase) cuboid for time, item, and location, summarized for all suppliers.

- The 0-D cuboid, which holds the highest level of summarization, is called the apex cuboid.
  - this is the total sales summarized over all four dimensions.
  - The apex cuboid is typically denoted by all.
Conceptual Modeling of Data Warehouses

• 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 on-line transaction processing.

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

• The most popular data model for a data warehouse is a multidimensional model.

• A multidimensional model can exist in the form of a star schema, a snowflake schema, or a fact constellation schema.
Multidimensional Model - 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.
- In the star schema, each dimension is represented by only one table, and each table contains a set of attributes.
Star Schema Example

Sales Fact Table
- time_key
- item_key
- branch_key
- location_key
- units_sold
- dollars_sold
- avg_sales

Measures
- time
  - time_key
day
day_of_the_week
month
quarter
year

branch
- branch_key
- branch_name
- branch_type

item
- item_key
- item_name
- brand
- type
- supplier_type

location
- location_key
- street
- city
- state_or_province
country
Multidimensional Model – 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.
**Snowflake Schema Example**

### Time
- time_key
- day
- day_of_the_week
- month
- quarter
- year

### Branch
- branch_key
- branch_name
- branch_type

### Sales Fact Table
- time_key
- item_key
- branch_key
- location_key
- units_sold
- dollars_sold
- avg_sales

### Measures
- item_key
- item_name
- brand
- type
- supplier_key

### Supplier
- supplier_key
- supplier_type

### Location
- location_key
- street
- city_key

### City
- city_key
- city
- state_or_province
- country
Multidimensional Model – Fact Constellation Schema

- 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 \textit{fact constellation schema}.
Fact Constellation Schema Example

**Fact Constellation Schema**

**Time**
- time_key
- day
- day_of_the_week
- month
- quarter
- year

**Branch**
- branch_key
- branch_name
- branch_type

**Location**
- location_key
- street
- city
- province_or_state
- country

**Sales Fact Table**
- time_key
- item_key
- branch_key
- location_key
- units_sold
- dollars_sold
- avg_sales

**Item**
- item_key
- item_name
- brand
- type
- supplier_type

**Shipping Fact Table**
- time_key
- item_key
- shipper_key
- from_location
- to_location
- dollars_cost
- units_shipped

**Shipper**
- shipper_key
- shipper_name
- location_key
- location
- country
- shipper_type
Concept Hierarchies

• A concept hierarchy defines a sequence of mappings from a set of low-level concepts to higher-level, more general concepts.
• Many concept hierarchies are implicit within the database schema.
• 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 concept hierarchy that is a total or partial order among attributes in a database schema is called a schema hierarchy.
• Concept hierarchies may also be defined by discretizing or grouping values for a given dimension, resulting in a set-grouping hierarchy.
  – A total or partial order can be defined among groups of values.
Concept Hierarchies –
A concept hierarchy for the dimension location

![Diagram of a concept hierarchy for location dimension]

- **location**
  - all
  - country
    - Canada
    - USA
    - "...
  - province_or_state
    - British Columbia
    - Ontario
    - New York
    - Illinois
  - city
    - Vancouver
    - Victoria
    - Toronto
    - Ottawa
    - New York
    - Buffalo
    - Chicago
    - Urbana
Concept Hierarchies: Hierarchical and lattice structures of attributes in warehouse dimensions

a hierarchy for *location*  
a lattice for *time*. 
Concept Hierarchies:
A concept hierarchy for the attribute price
OLAP Operations

• “How are concept hierarchies useful in OLAP?”
• 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 perspectives.
• A number of OLAP data cube operations exist to materialize these different views, allowing interactive querying and analysis of the data at hand.
Typical OLAP Operations

• **Roll up (drill-up):** summarize data
  – by climbing up hierarchy or by dimension reduction

• **Drill down (roll down):** reverse of roll-up
  – from higher level summary to lower level summary or detailed data, or introducing new dimensions

• **Slice and dice:** project and select

• **Pivot (rotate):**
  – reorient the cube, visualization, 3D to series of 2D planes
OLAP Operation: Roll-up

- The **roll-up operation** (also called the drill-up operation) performs aggregation on a data cube either by climbing up a concept hierarchy for a dimension or by dimension reduction.
OLAP Operation: Drill-down

- **Drill-down** is the reverse of roll-up. It navigates from less detailed data to more detailed data.
- Drill-down can be realized by either stepping down a concept hierarchy for a dimension or introducing additional dimensions.
The **slice operation** performs a selection on one dimension of the given cube, resulting in a subcube.

The **dice operation** defines a subcube by performing a selection on two or more dimensions.
OLAP Operation: Pivot

- **Pivot (rotate)** is a visualization operation that rotates the data axes in view in order to provide an alternative presentation of the data.
Data Warehouse Design Process

• **Top-down, bottom-up approaches or a combination** of both
  – **Top-down**: Starts with overall design and planning (mature)
  – **Bottom-up**: Starts with experiments and prototypes (rapid)

• **From software engineering point of view**
  – **Waterfall**: structured and systematic analysis at each step before proceeding to the next
  – **Spiral**: rapid generation of increasingly functional systems, short turn around time, quick turn around

• **Typical data warehouse design process**
  – Choose a business process to model, e.g., orders, invoices, etc.
  – Choose the **grain** (*atomic level of data*) of the business process
  – Choose the dimensions that will apply to each fact table record
  – Choose the measure that will populate each fact table record
A Three-Tier Data Warehouse Architecture

- Data warehouses often adopt a three-tier architecture
Data Warehouse Usage

- Three kinds of data warehouse applications
  - **Information processing**
    - supports querying, basic statistical analysis, and reporting using crosstabs, tables, charts and graphs
  - **Analytical processing**
    - multidimensional analysis of data warehouse data
    - supports basic OLAP operations, slice-dice, drilling, pivoting
  - **Data mining**
    - knowledge discovery from hidden patterns
    - supports associations, constructing analytical models, performing classification and prediction, and presenting the mining results using visualization tools
From On-Line Analytical Processing (OLAP) to On Line Analytical Mining (OLAM)

• Why online analytical mining?
  – High quality of data in data warehouses
    • DW contains integrated, consistent, cleaned data
  – Available information processing structure surrounding data warehouses
    • Web accessing, service facilities, reporting and OLAP tools
  – OLAP-based exploratory data analysis
    • Mining with drilling, dicing, pivoting, etc.
  – On-line selection of data mining functions
    • Integration and swapping of multiple mining functions, algorithms, and tasks
An integrated OLAM and OLAP architecture

- An OLAM server may perform multiple data mining tasks, such as concept description, association, classification, prediction, clustering, time-series analysis, and it usually consists of multiple integrated data mining modules and is more sophisticated than an OLAP server.
Summary

• A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data organized in support of management decision making.
  – Several factors distinguish data warehouses from operational databases.
  – Because the two systems provide quite different functionalities and require different kinds of data, it is necessary to maintain data warehouses separately from operational databases.
• A multidimensional data model is typically used for the design of corporate data warehouses.
  – A multidimensional data model can adopt a star schema, snowflake schema, or fact constellation schema.
  – The core of the multidimensional model is the data cube, which consists of a large set of facts (or measures) and a number of dimensions.
  – Dimensions are the entities or perspectives with respect to which an organization wants to keep records and are hierarchical in nature.
• A data cube consists of a lattice of cuboids, each corresponding to a different degree of summarization of the given multidimensional data.
Summary

• **Concept hierarchies** organize the values of dimensions into gradual levels of abstraction.
• **On-line analytical processing (OLAP)** can be performed in data warehouses using the multidimensional data model.
  – Typical OLAP operations include roll-up, drill-down, slice-and-dice, pivot (rotate), as well as statistical operations such as ranking and computing moving averages and growth rates.
• Data warehouses often adopt a **three-tier architecture**.
  – The bottom tier is a warehouse database server, which is a relational database system.
  – The middle tier is an OLAP server, and
  – The top tier is a client, containing query and reporting tools.
• A data warehouse contains **back-end tools and utilities** for populating and refreshing the warehouse.
  – data extraction, data cleaning, data transformation, loading, refreshing, and warehouse management.
• Data warehouse **metadata** are data defining the warehouse objects.
  – A metadata repository provides details regarding the warehouse structure, data history, the algorithms used for summarization, mappings from the source data to warehouse form.