# Data Mining

Associate Professor Dr. Raed Ibraheem Hamed

University of Human Development,
College of Science and Technology

2016 – 2017

# Points to Cover

- Why Do We Need Data Warehouses?
- Operational System
- What is Data Warehouse?
- Data Warehouse—Subject-Oriented
- Data Warehouse—Integrated
- Data Warehouse—Time Variant
- Data Warehouse—Non-Volatile
- Data Warehouse vs. Operational DBMS
- OLTP vs. OLAP
- Why Separate Data Warehouse?
- Data Warehouse Architecture: Basic

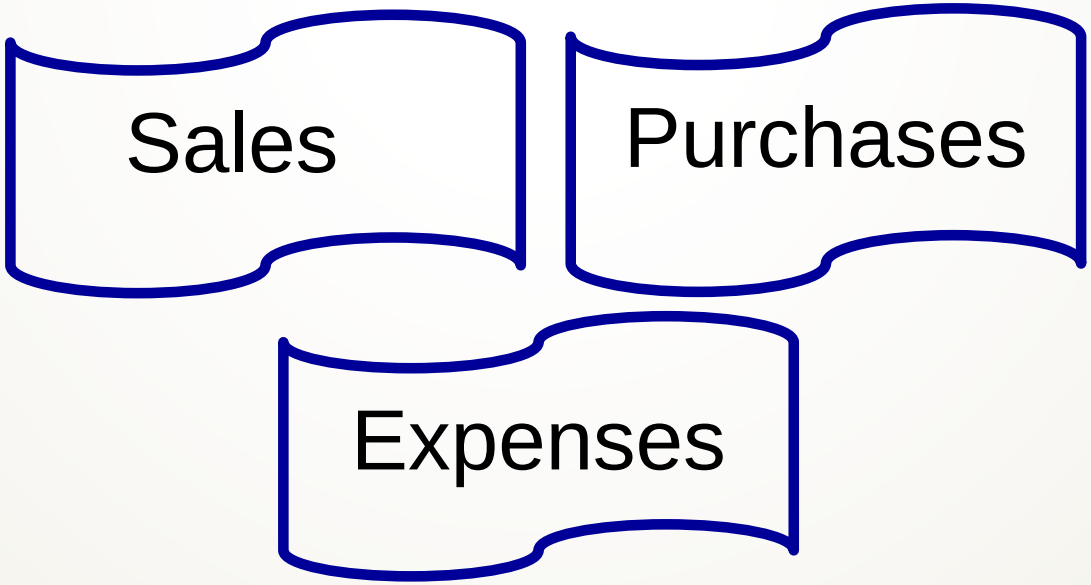
# Why Do We Need Data Warehouses?

1. **Unification of information resources**. Improved query performance “Separate research and decision support functions from the operational systems.

2. **The data stored in the warehouse is uploaded from the operational systems**. The data may pass through an operational data store for additional operations before it is used in the DW for reporting.

# Operational System

An **operational system** is a term used in data warehousing to refer to a **system** that is used to process the day-to-day transactions of an organization. These **systems** are designed in a manner that processing of day-to-day transactions is performed efficiently and the integrity of the transactional data is preserved.



# What is Data Warehouse?

**Defined in many different ways, but not rigorously:-**

1. A decision support database that is maintained separately from the organization's operational database.

2. "A data warehouse is a **subject-oriented**, **integrated**, **time-variant**, and **nonvolatile** collection of data in support of management's decision-making process." by William H. Inmon

# 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

1. **Constructed by integrating multiple, heterogeneous data sources**
   - relational databases, flat files, on-line transaction records

2. **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.

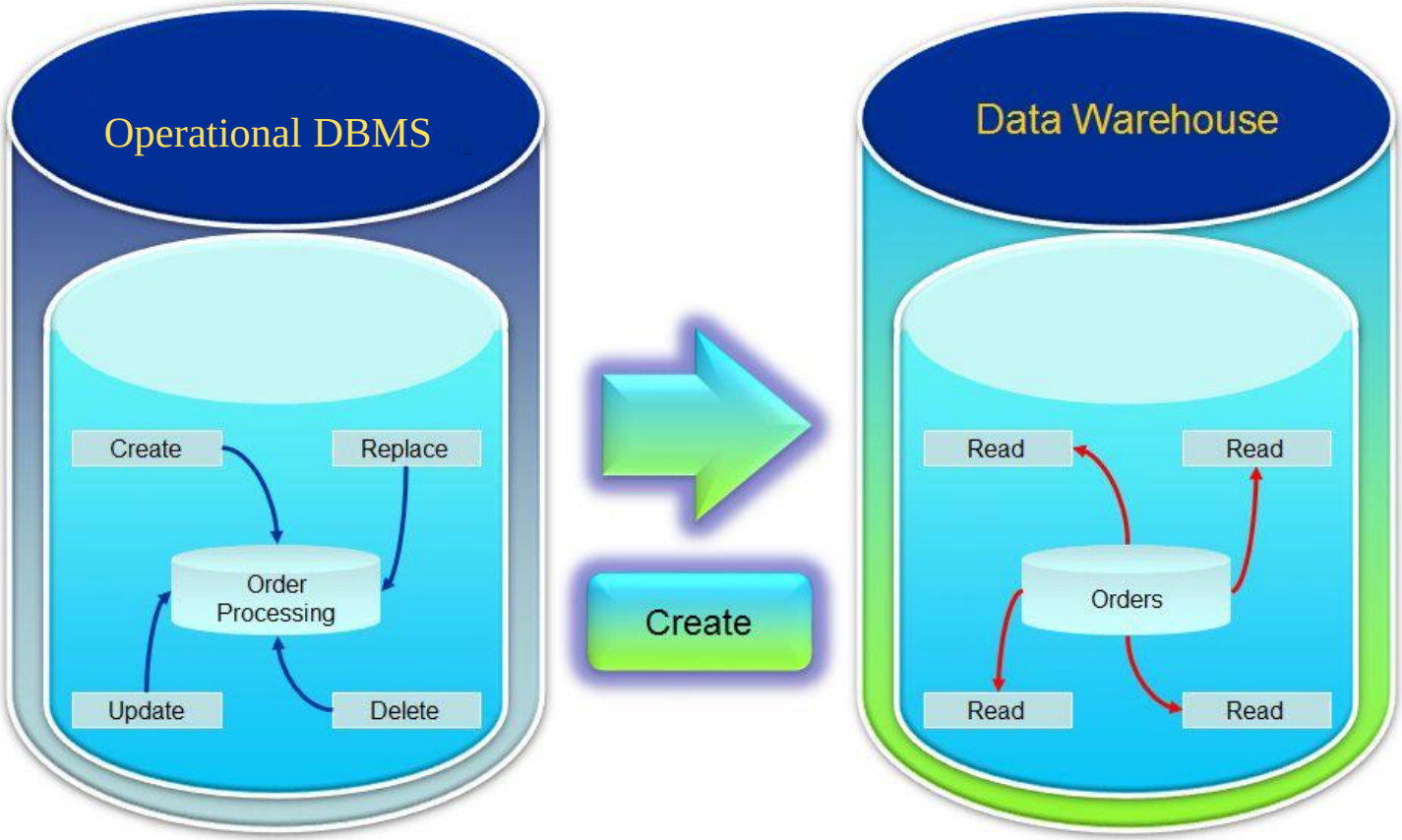
# Data Warehouse—Time Variant

- **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—Non-Volatile

- 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*.

# Data Warehouse Versus Operational DBMS



# Data Warehouse Versus Operational DBMS

- **OLTP (on-line transaction processing)**

1. Major task of traditional relational DBMS
2. Day-to-day operations: purchasing, inventory, banking, manufacturing, payroll, registration, accounting, etc.

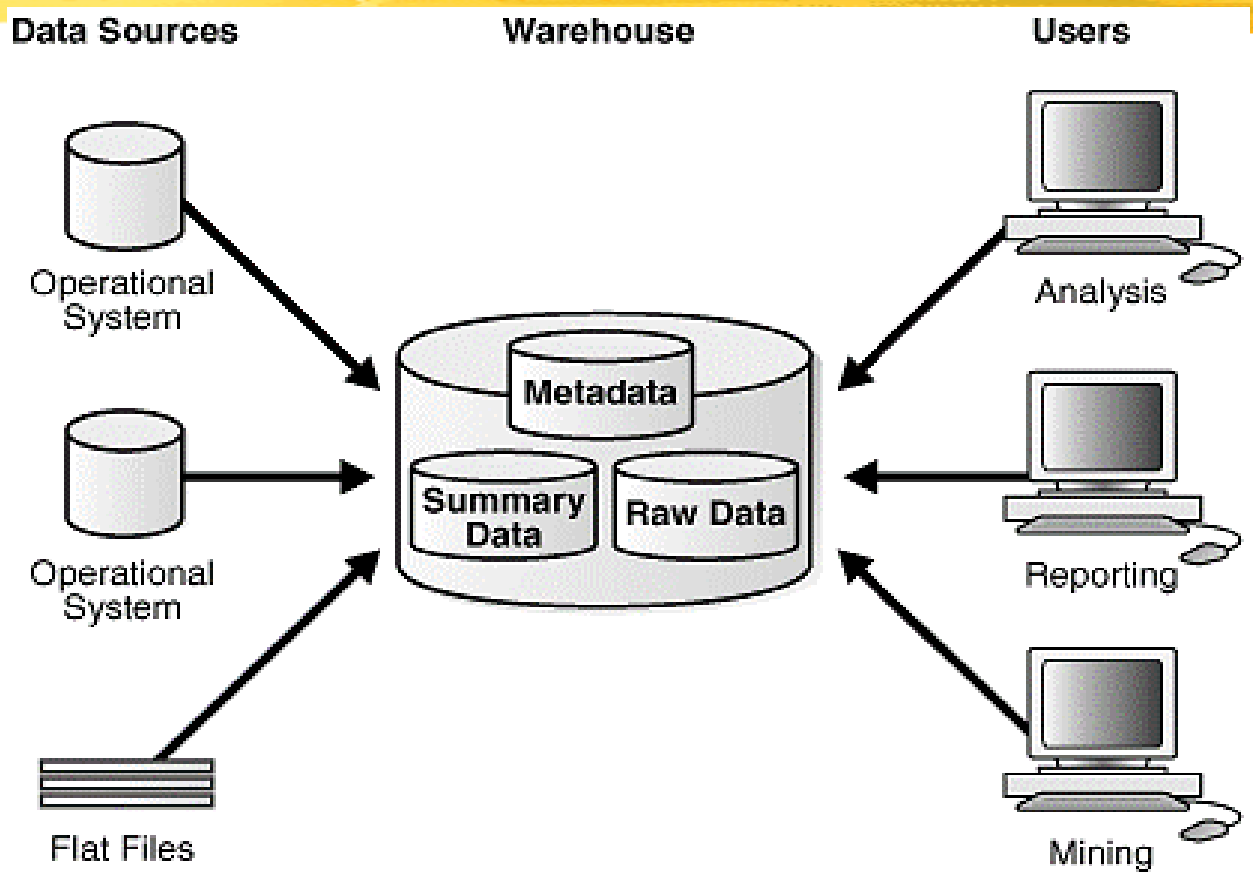
- **OLAP (on-line analytical processing)**

1. Major task of data warehouse system
2. Data analysis and decision making

# Difference Between OLTP and OLAP

| | **OLTP** | **OLAP** |
|---|---|---|
| **users** | Writer, IT professional | knowledge worker |
| **function** | day to day operations | decision support |
| **DB design** | application-oriented | subject-oriented |
| **data** | current, up-to-date<br>detailed, flat relational<br>isolated | historical,<br>summarized, multidimensional<br>integrated, consolidated |
| **access** | read/write<br>index/hash on primary key | lots of scans |
| **unit of work** | short, simple transaction | complex query |
| **# records accessed** | tens | millions |
| **#users** | thousands | hundreds |
| **DB size** | 100MB-GB | 100GB-TB |

# Data Warehouse Architecture: Basic




shows a simple architecture for a data warehouse. End users directly access data derived from several source systems through the data warehouse.


Thank
you