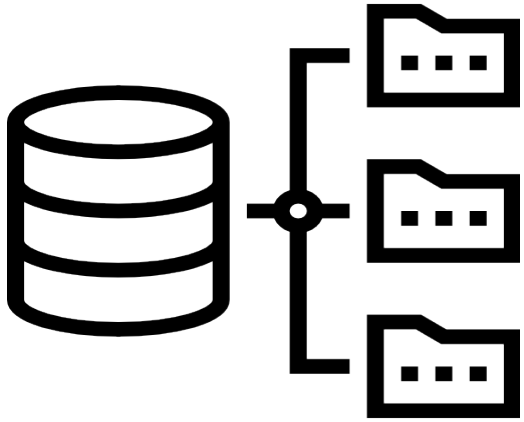
Advanced Database- IS411

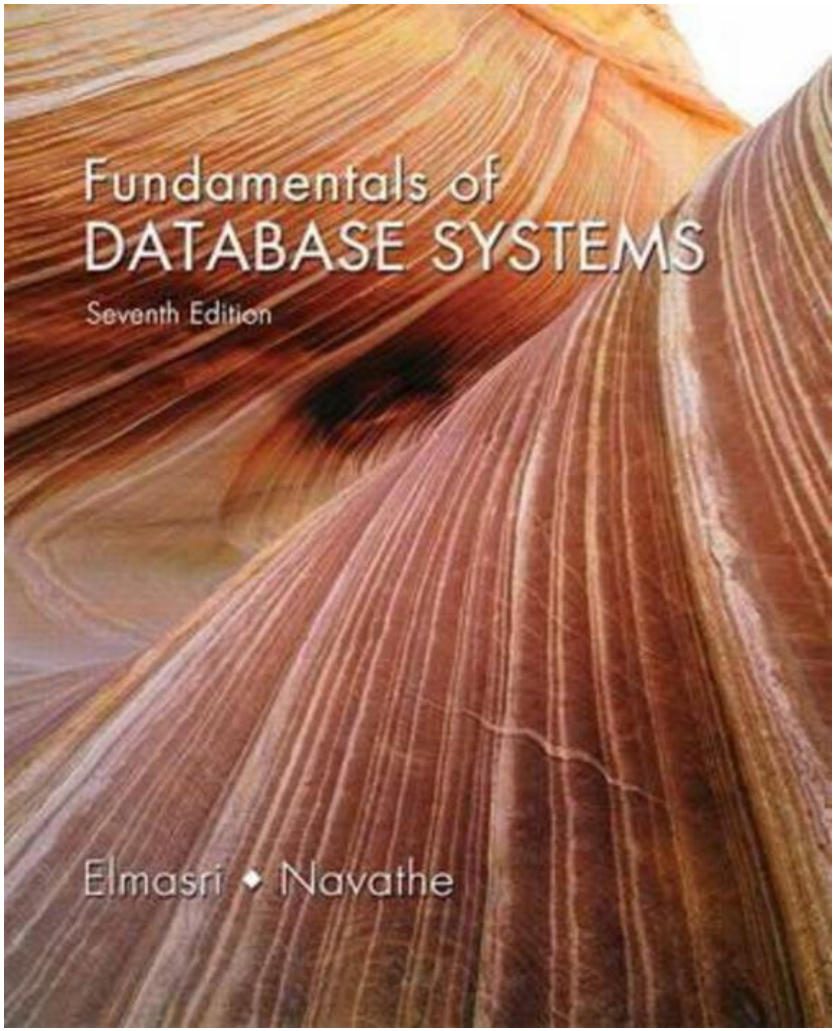


Introduced by

Dr. Ebtsam Adel

Lecturer of Information Systems,
Information Systems department,
Faculty of computers and information,
Damanhour university

Materials



Copyright © 2016 Ramez Elmasri and Shamkant B. Navathe


QUIZ!

- Explain the main differences between the database and IR systems.

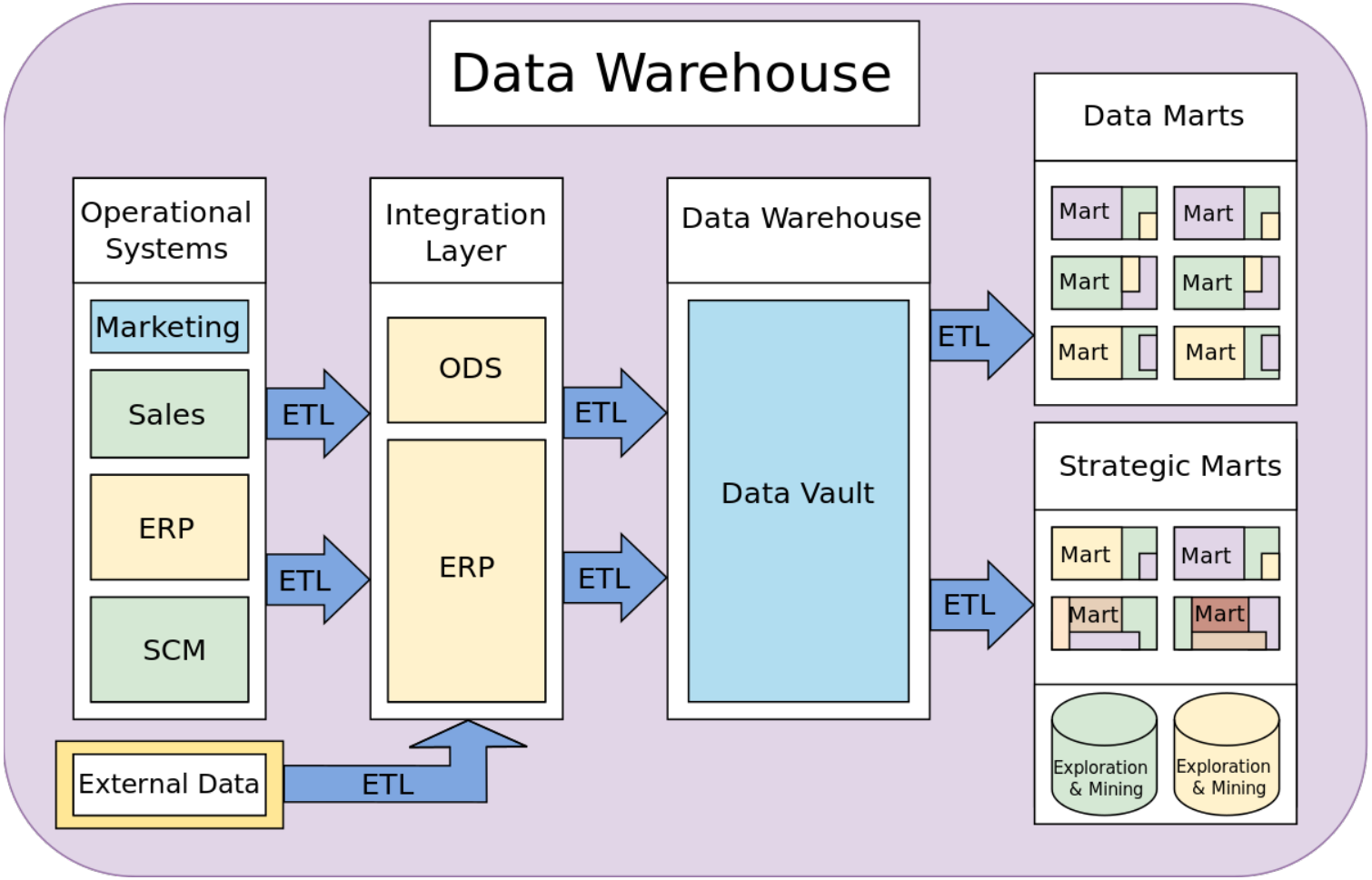
- Describe the main components of the IR system.

Topics

- ✓ **Chapter 20** Introduction to Transaction Processing Concepts and Theory
- ✓ **chapter 24** NOSQL Databases and Big Data Storage Systems
- ✓ **chapter 25** Big Data Technologies Based on MapReduce and Hadoop
- ✓ **chapter 27** Introduction to Information Retrieval and Web Search
- ✓ **chapter 29** **Overview of Data Warehousing and OLAP**
- ✓ **chapter 30** Database Security

Copyright © 2016 Ramez Elmasri and Shamkant B. Navathe

Data Warehousing



Learning Objectives

- ✓ Understand the basic definitions and concepts of data warehouses
- ✓ Learn different types of data warehousing architectures; their comparative advantages and disadvantages
- ✓ Describe the processes used in developing and managing data warehouses
- ✓ Explain data warehousing operations

Learning Objectives

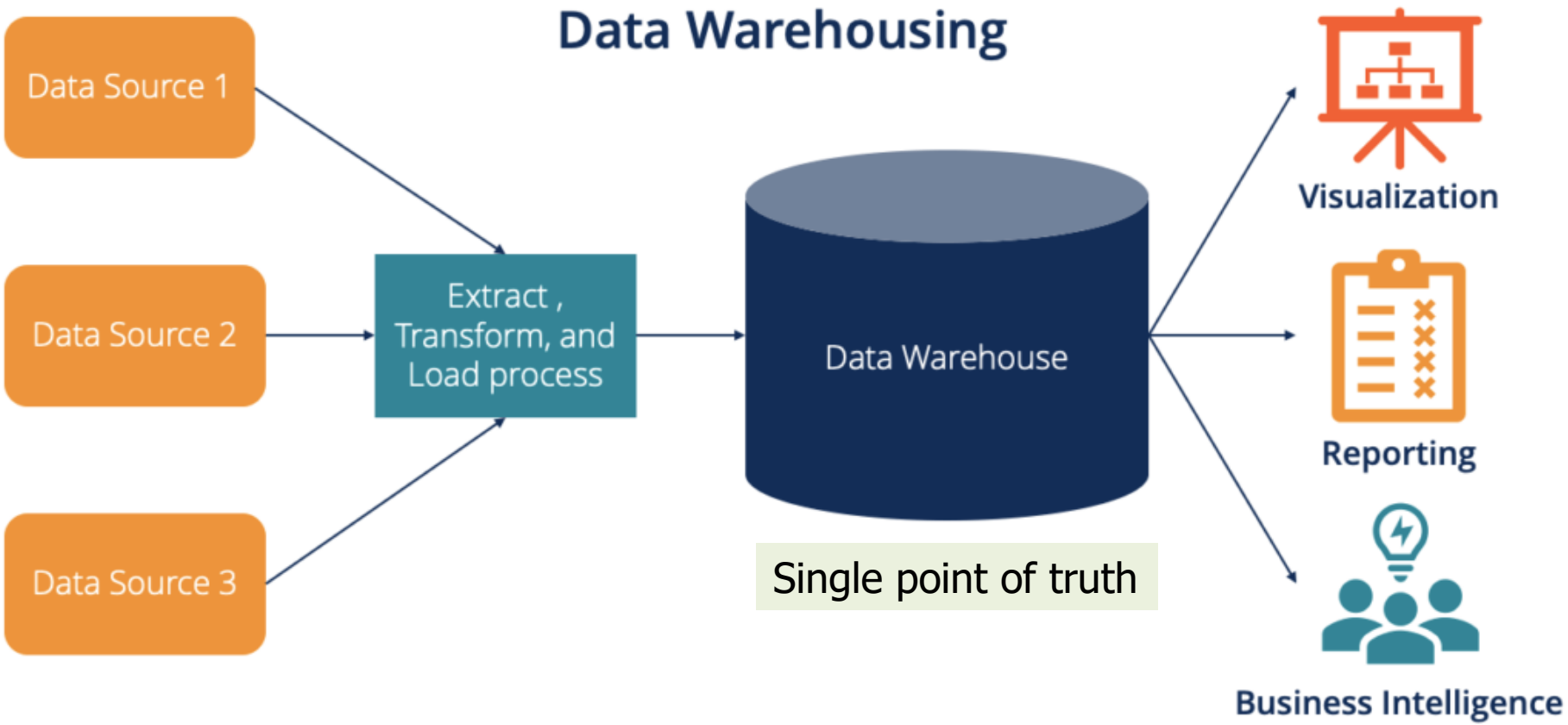
- ✓ Explain the role of data warehouses in decision support
- ✓ Explain data integration and the extraction, transformation, and load (**ETL**) processes
- ✓ Describe real-time (a.k.a. right-time and/or active) data warehousing
- ✓ Understand data warehouse administration and security issues.

Main Data Warehousing Topics

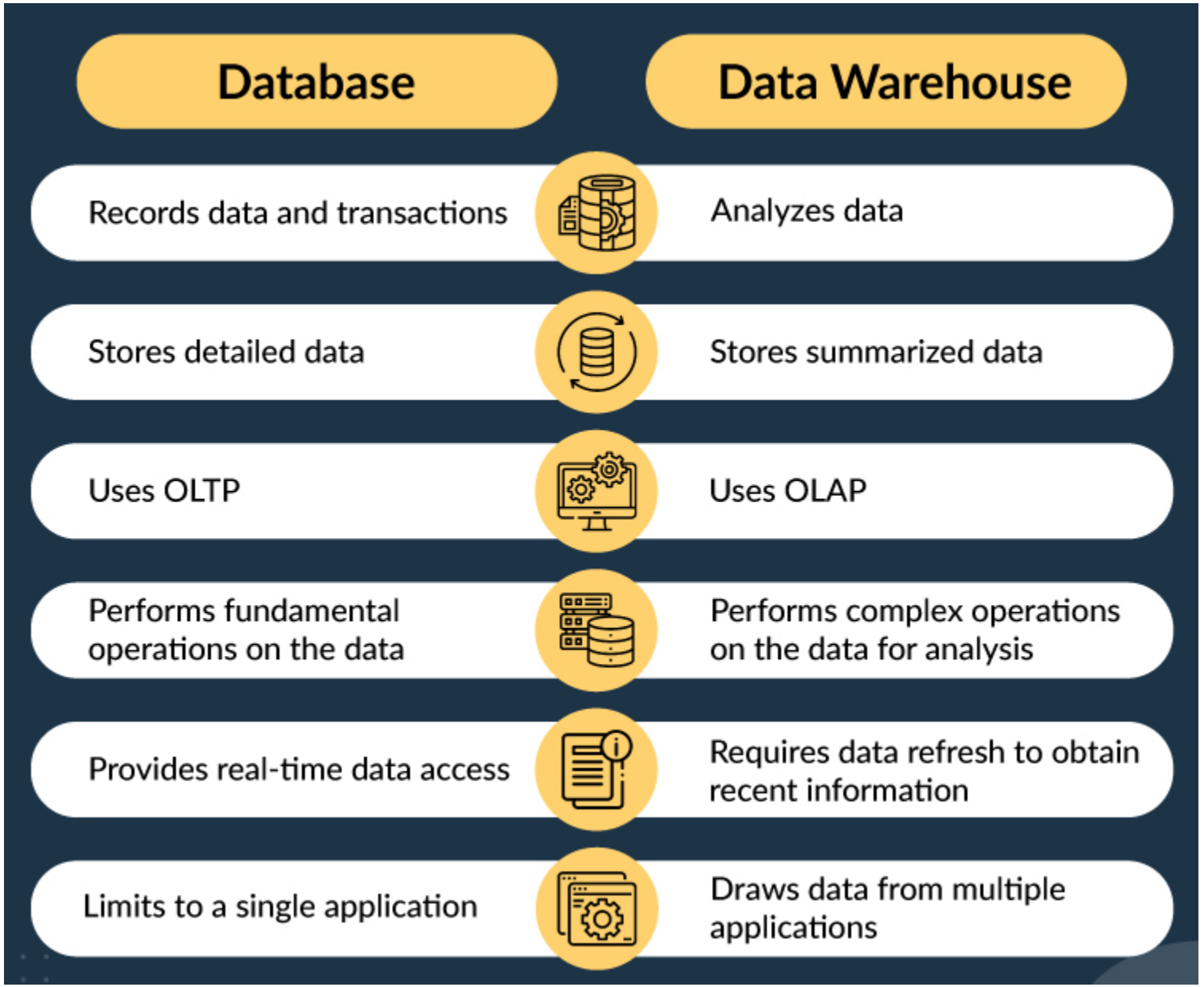
- ✓ DW definition
- ✓ Characteristics of DW
- ✓ Data Marts
- ✓ ODS, EDW, Metadata
- ✓ DW Framework
- ✓ DW Architecture & ETL Process
- ✓ DW Development
- ✓ DW Issues

What Is a Data Warehouse?

What is a Data Warehouse?



What is a Data Warehouse?



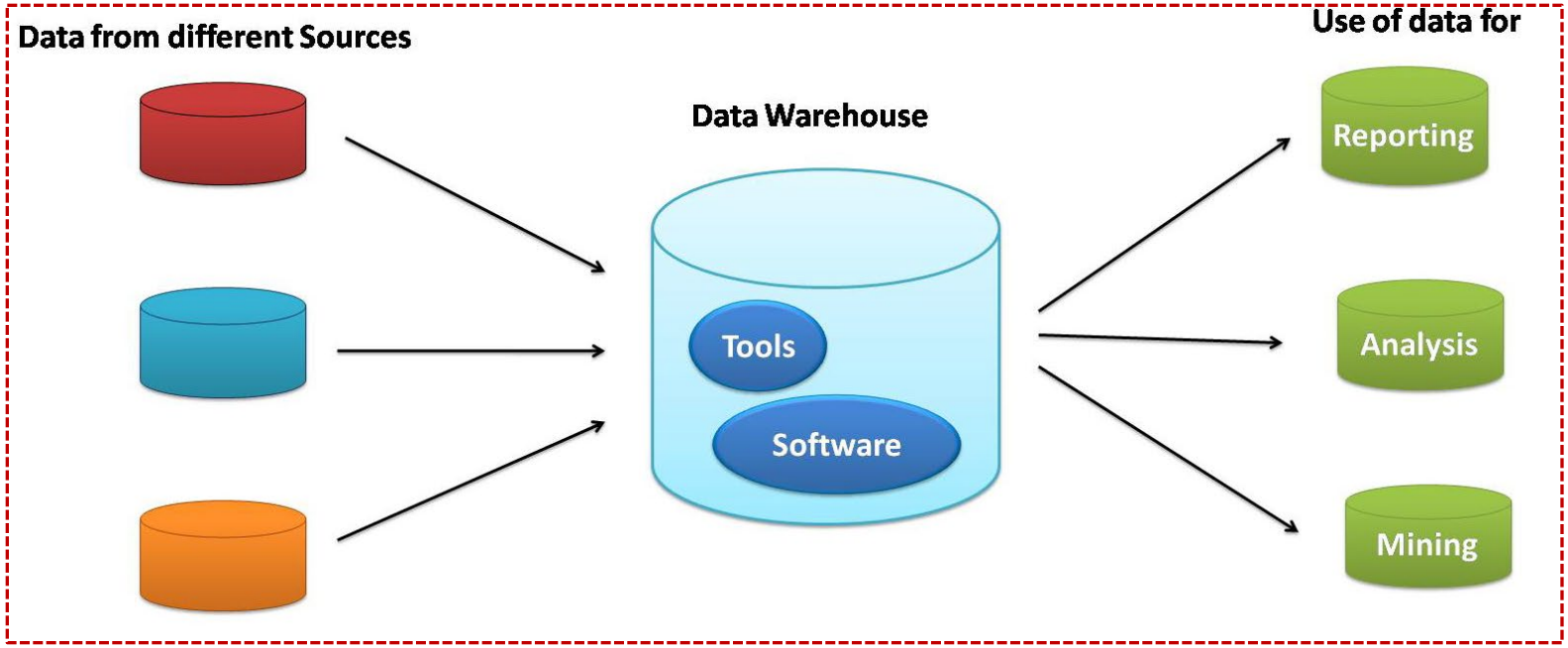
What is a Data Warehouse?

- In simple terms, a **data warehouse** (DW) is a pool of data produced to support decision making; it is also a repository of current and historical data of potential interest to **managers** throughout the organization.

- Data are usually structured to be available in a form ready for analytical processing activities (i.e., online analytical processing [OLAP], data mining, querying, reporting, and other decision support applications).

What is a Data Warehouse?

- The **data warehouse** is a collection of integrated, subject-oriented databases designed to support DSS functions, where each unit of data is non-volatile and relevant to some moment in time.



Characteristics of DWs

- Subject oriented
- Integrated
- Time-variant (time series)
- Nonvolatile
- Summarized
- Not normalized
- Metadata
- Web based, relational/multi-dimensional
- Client/server, real-time/right-time/active...

Characteristics of DWs

- **Subject oriented**. Data are organized by detailed subject, such as sales, products, or customers, containing only information relevant for decision support.

- **Integrated**. Integration is closely related to subject orientation. Data warehouses must place data from different sources into a consistent format.

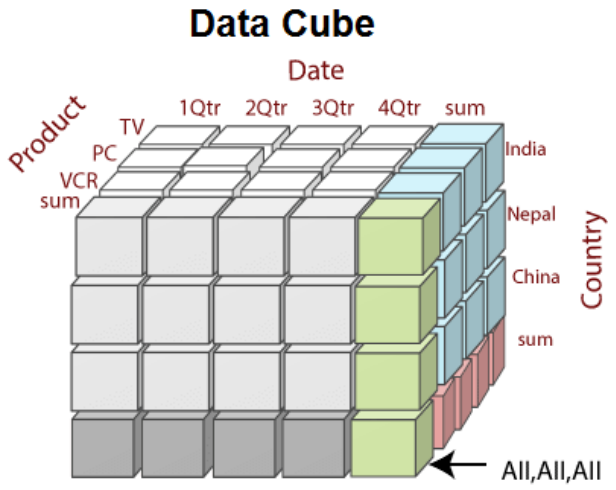
- **Time variant** (time series). A warehouse maintains historical data. The data do not necessarily provide current status (except in real-time systems). They detect trends, deviations, and long-term relationships for forecasting and comparisons, leading to decision making.

Characteristics of DWs

- **Nonvolatile**. After data are entered into a data warehouse, users cannot change or update the data. Obsolete "very old data" data are discarded, and changes are recorded as new data.

- **Web-based**. Data warehouses are typically designed to provide an efficient Computing environment for Web-based applications.

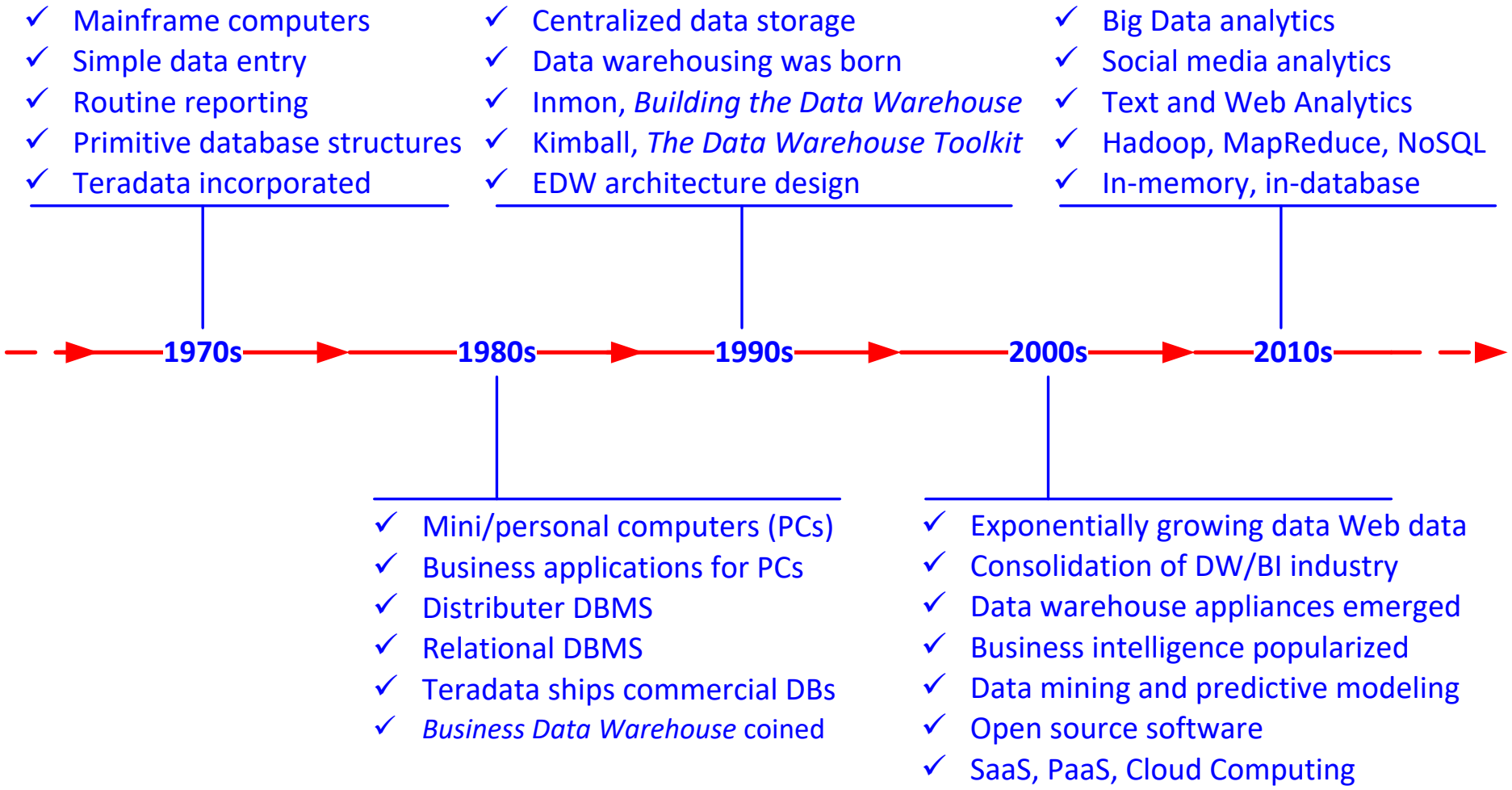
- **Relational/multidimensional.** A data warehouse uses either a relational structure or a multidimensional structure.



Characteristics of DWs

- **Client/server.** A data warehouse uses the client/server architecture to provide easy access for end users.
- **Real time.** Newer data warehouses provide real-time, or active, data-access and analysis capabilities.
- **Include metadata.** A data warehouse contains metadata (data about data) about how the data are organized and how to effectively use them.

A Historical Perspective to Data Warehousing



Data warehouse history

- As data began to proliferate in the 1970s and '80s, organizations needed a way to store and access all their information.
- In this period **Bill Inmon**, the father of data warehousing, began to define the concept data warehouse.
- He published Building the Data Warehouse, lauded as a fundamental source on data warehousing technology.
- in 1992. Inmon's definition of the data warehouse takes a "**top-down**" approach, where a centralized repository is established first and then data marts – which contain specific subsets of data – are created within that repository.

Data warehouse history

- This approach differs in some respects to the "other" father of Data Warehousing, **Ralph Kimball**.
- Ralph Kimball, in 1992 published The Data Warehouse Toolkit.
- It took a slightly different view of the data warehousing concept.
- In his "bottom-up" approach, individual data marts are developed first and integrated together later to create a data warehouse.

Data Marts

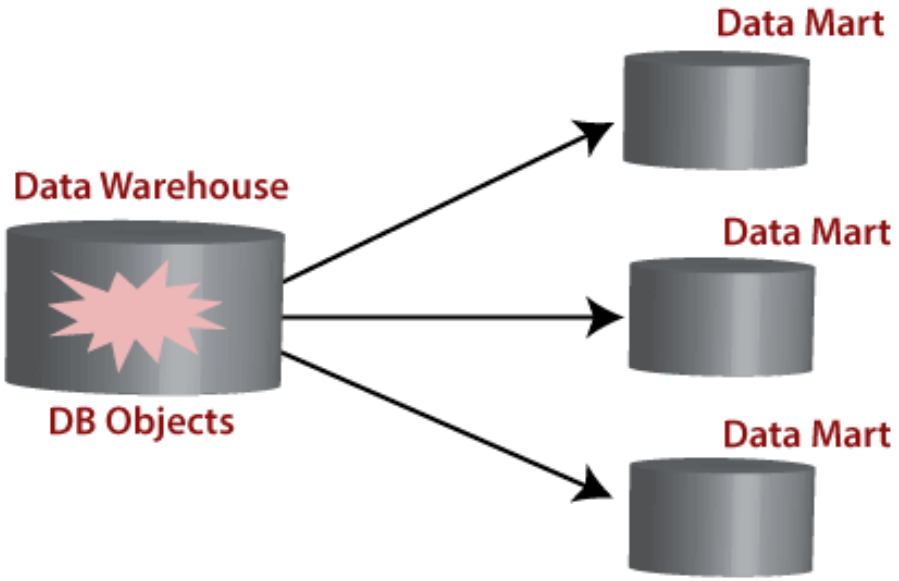
Data Mart

- A departmental small-scale "DW" that stores only limited/relevant data.
  - ❖ **Dependent data mart**

    A subset that is created directly from a datawarehouse.

  - ❖ **Independent data mart**

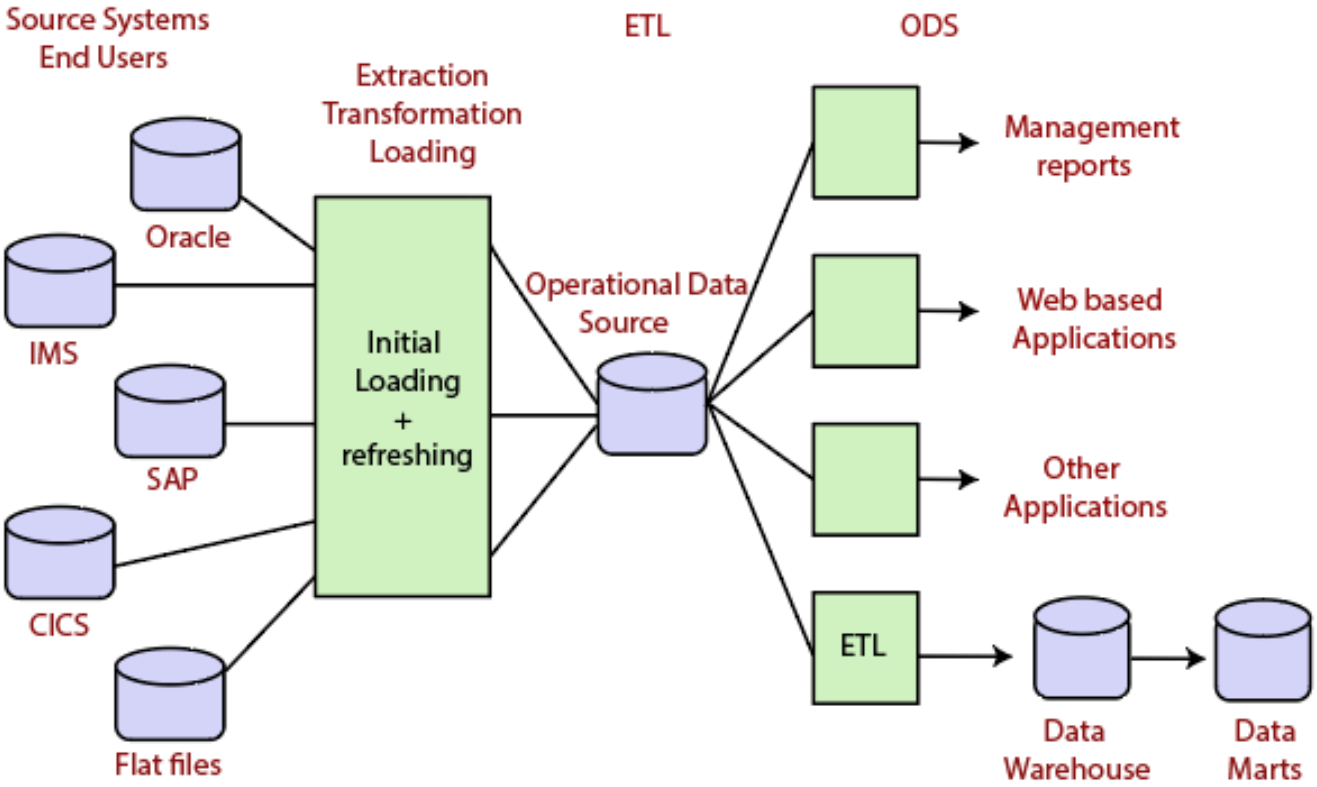
    A small datawarehouse designed for a strategic business unit or a department.



Other DW Components

- Operational data stores (ODS): A type of database often used as an interim "مؤقت" area for a data warehouse.
- Oper marts: an operational data mart.
- Enterprise data warehouse (EDW): A data warehouse for the enterprise.
- Metadata: Data about data: In a data warehouse, metadata describe the contents of a data warehouse and the manner of its acquisition and use.

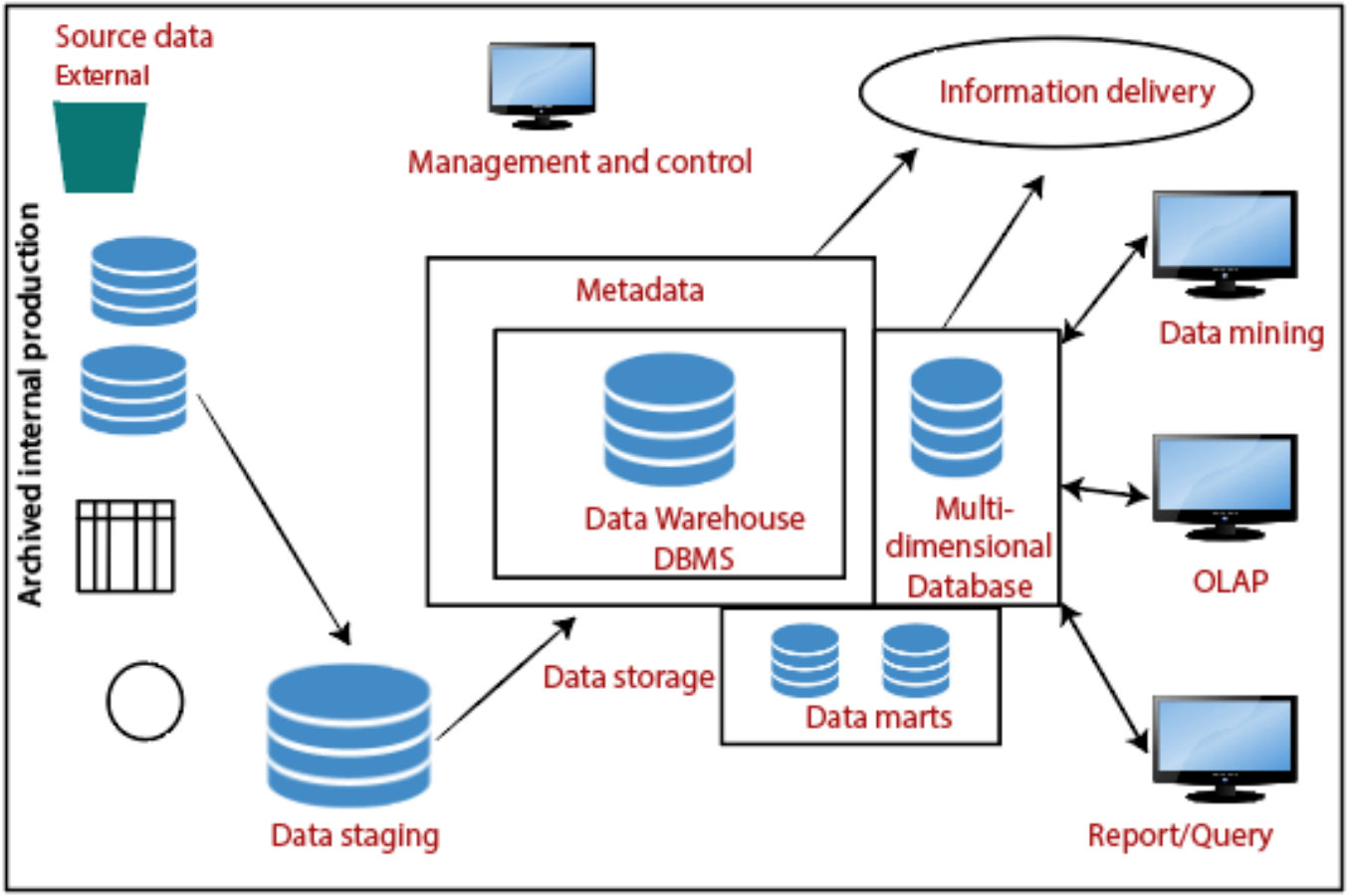
Other DW Components



Operational Data Store Structure

An **operational system** is a method used in data warehousing to refer to a system that is used to process the day-to-day transactions of an organization.

Other DW Components



Components or Building Blocks of Data Warehouse

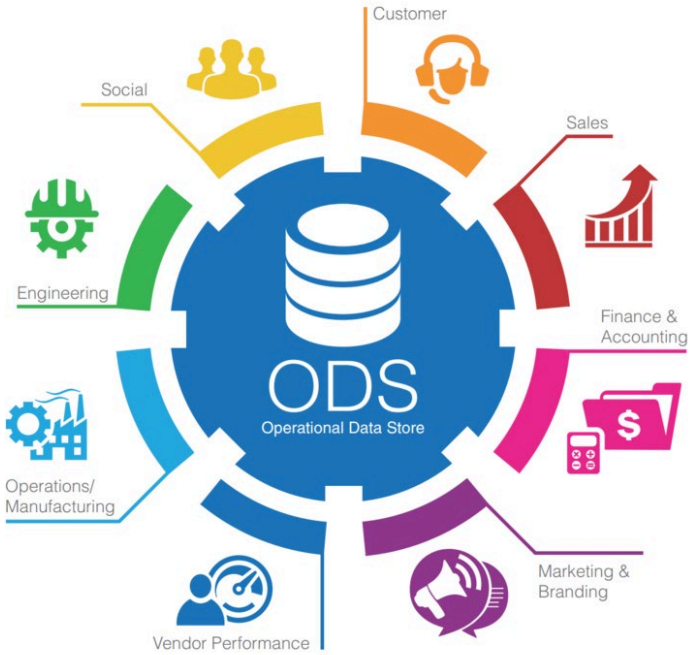
Other DW Components

- **An operational data store** (**ODS**) provides a fairly recent form of customer information file (CIF).
- This type of database is often used as an interim "مؤقت" staging area for a data warehouse.
- Unlike the static contents of a data warehouse, the contents of an ODS are updated throughout the course of business operations.

Other DW Components

- An ODS is used for short-term decisions involving mission-critical applications rather than for the medium and long-term decisions associated with an EDW.

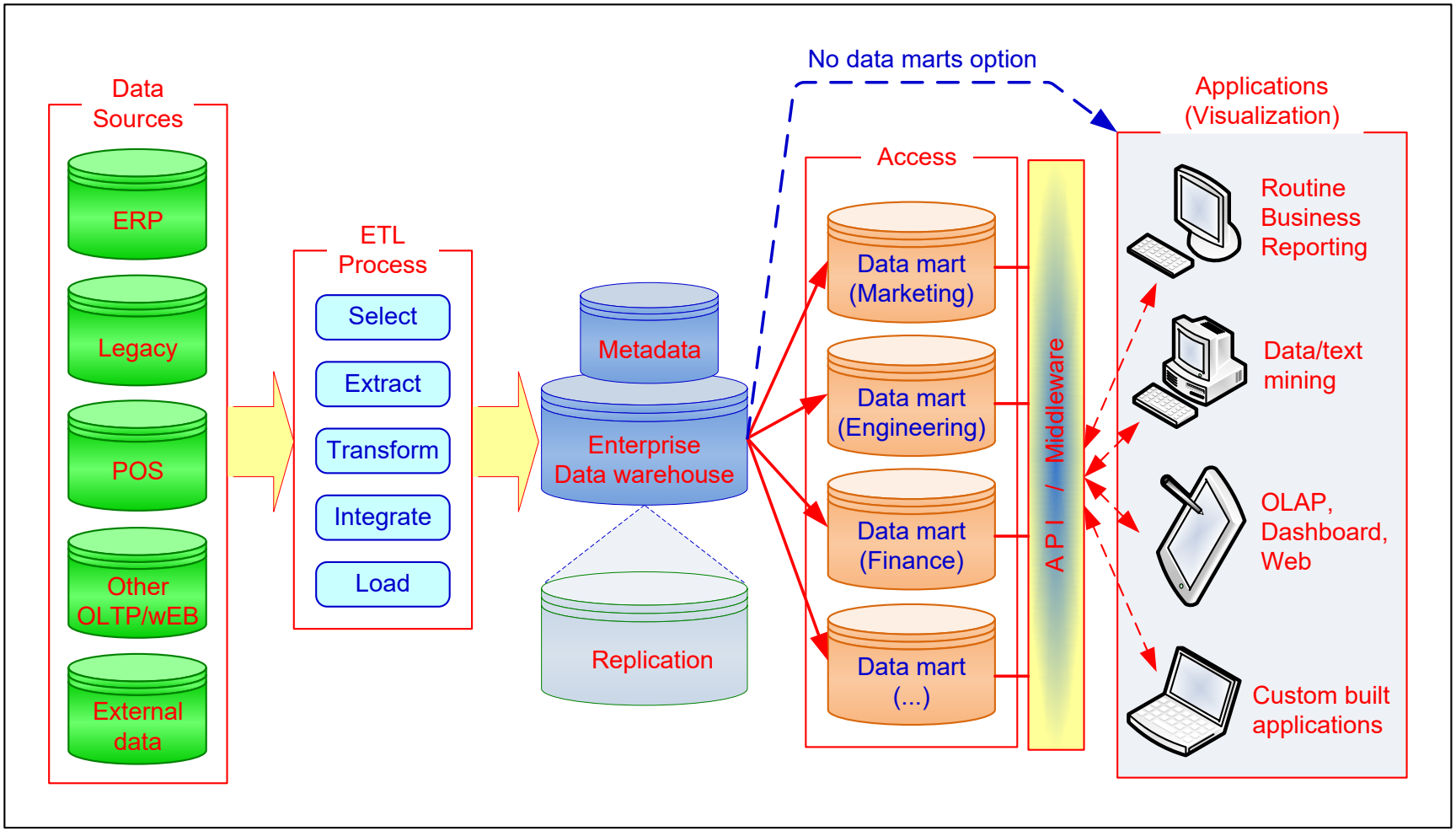
- The exchange, transfer, and load (ETL) processes for an ODS are identical to those for a data warehouse.



Other DW Components

- **oper marts** are created when operational data needs to be analyzed multidimensional.
- **An enterprise data warehouse (EDW)** is a large-scale data warehouse that is used across the enterprise for decision support.
- **Metadata** are data about data. Metadata **describe** the structure of and some meaning about data, thereby contributing to their effective or ineffective use.

A Generic DW Framework



ERP (Enterprise Resource Planning)

Data Warehousing Architectures

DW Architecture

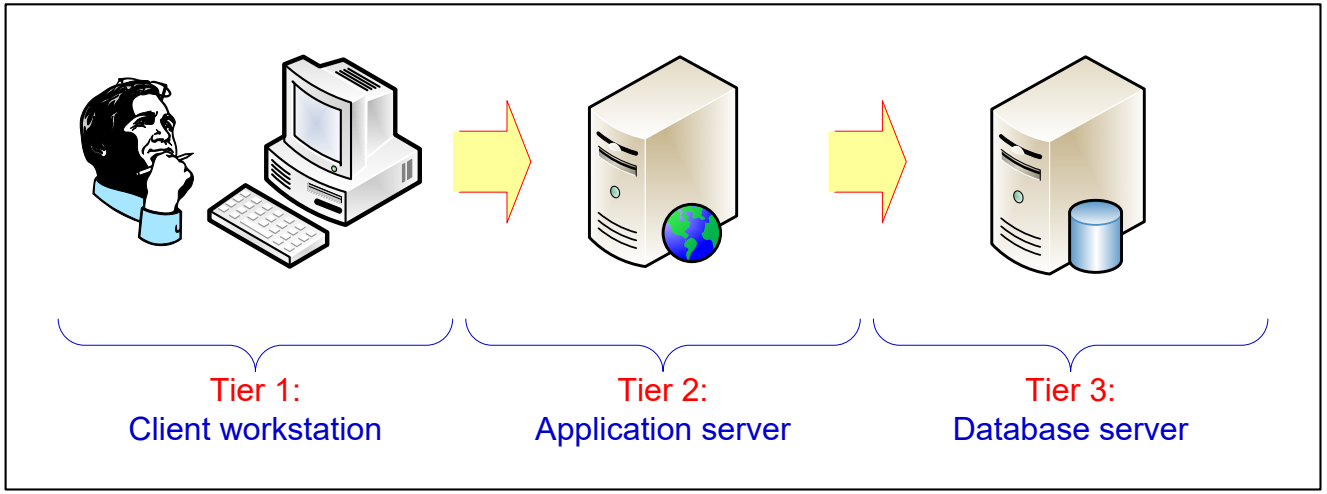
- ✓ Three-tier architecture
  1. Data acquisition software (back-end)
  2. The data warehouse that contains the data & software
  3. Client (front-end) software that allows users to access and analyze data from the warehouse.
- ✓ Two-tier architecture

  First two tiers in three-tier architecture is combined into one

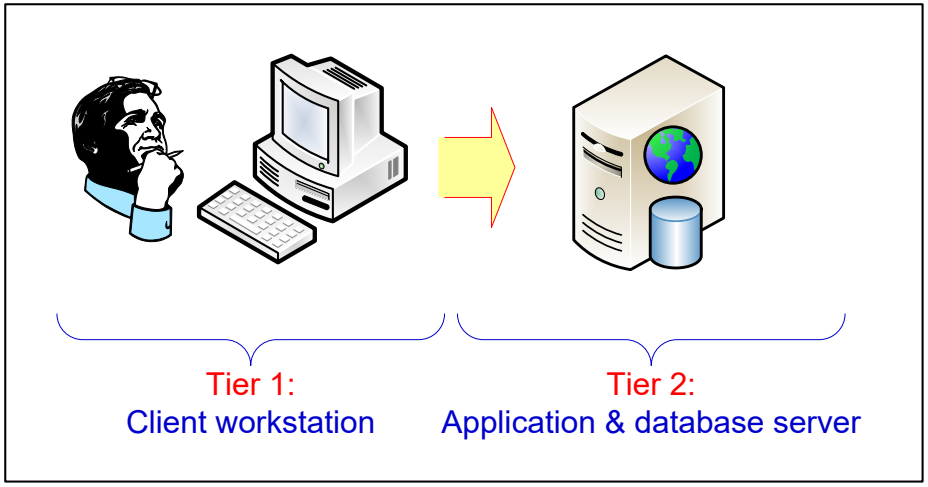
... sometimes there is only one tier?

DW Architectures

3-tier architecture



2-tier architecture



1-tier Architecture ?

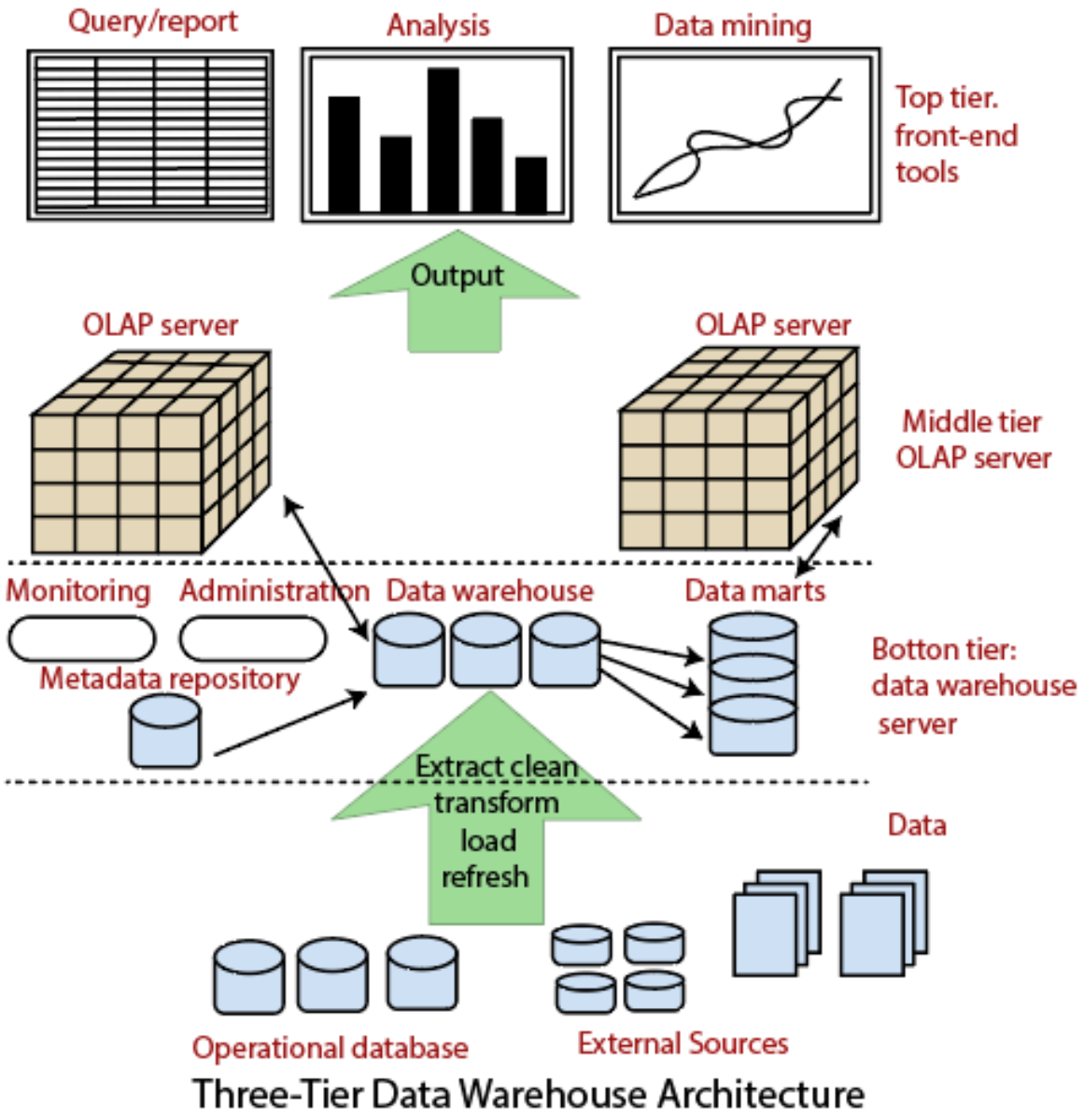
DW Architectures

- In a **three-tier architecture**, operational systems contain the data and the software for data acquisition in one tier (i.e., **the server**), the data warehouse is **another tier**, and **the third tier** includes the DSS/BI/BA engine (i.e., the application server) and the client.

- Data from the warehouse are processed twice and deposited in an additional multidimensional database, organized for easy multidimensional analysis and presentation, or replicated in data marts.

- The advantage of the three-tier architecture is its separation of the functions of the data warehouse, which eliminates resource constraints and makes it possible to easily create data marts.

DW Architectures- three-tier architecture

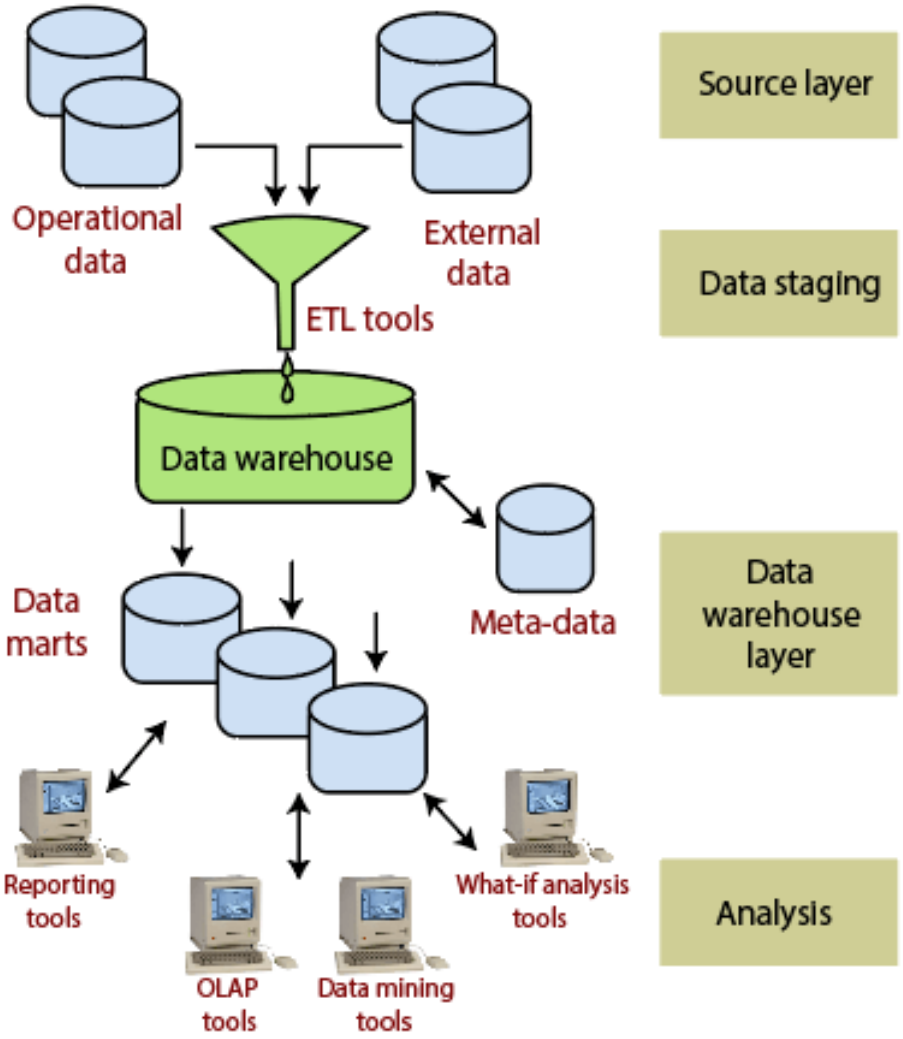


Three-Tier Data Warehouse Architecture

DW Architectures- two-tier architecture

- Although it is typically called **two-layer architecture** to highlight a separation between physically available sources and data warehouses.
- In a **two-tier architecture**, the DSS engine physically runs on the same hardware platform as the data warehouse. Therefore, it is more economical than the three-tier structure.
- The two-tier architecture can have performance problems for large data warehouses that work with data-intensive applications for decision support.

DW Architectures- two-tier architecture



Two-Tier Data Warehouse Architecture

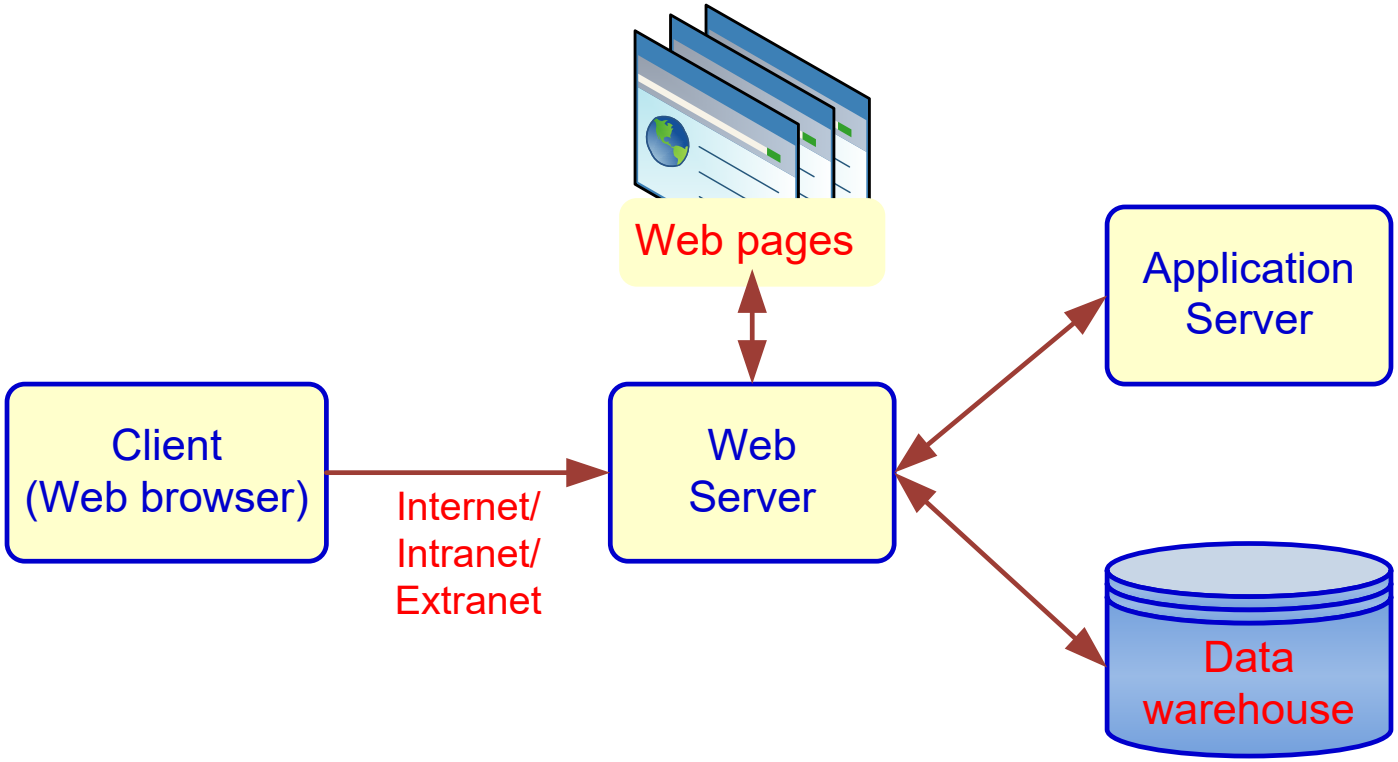
Data Warehousing Architectures Issues

Data Warehousing Architectures

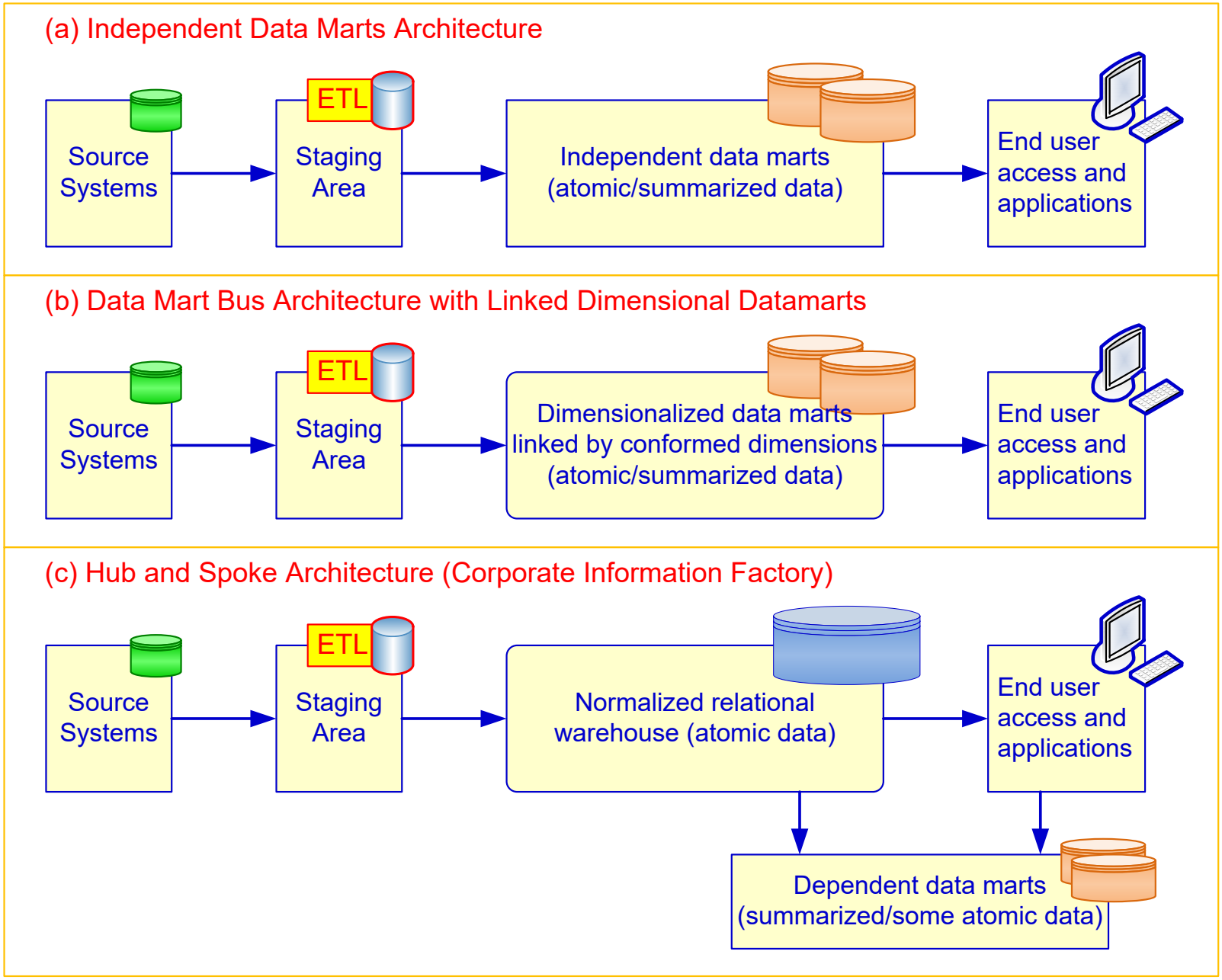
- Issues to consider when deciding which architecture to use:
  - Which database management system (DBMS) should be used?
  - Will parallel processing and/or partitioning be used?
  - Will data migration tools be used to load the data warehouse?
  - What tools will be used to support data retrieval and analysis?

A Web-Based DW Architecture

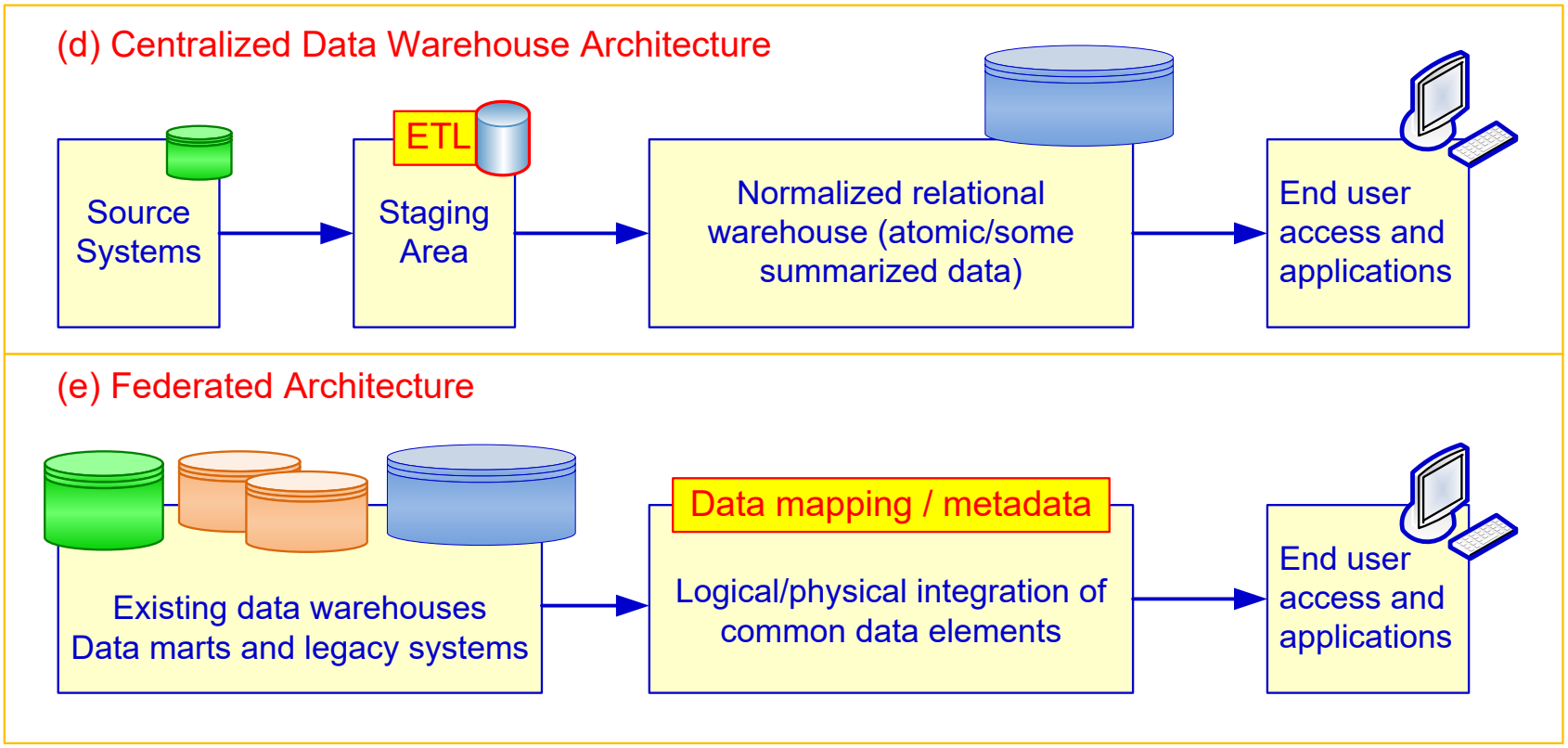
- Data warehousing and the Internet are two key technologies that offer important solutions for managing corporate data. The integration of these two technologies produces Web-based data warehousing.



Alternative DW Architectures



Alternative DW Architectures



- Each architecture has advantages and disadvantages!
- Which architecture is the best?

Alternative DW Architectures

- **Independent data marts**. This is arguably the simplest and the least costly architecture alternative.

- The data marts are developed to operate independently of each another to serve the needs of individual organizational units.

Alternative DW Architectures

- **Data mart bus architecture**. This architecture is a viable alternative to the independent data marts where the individual marts are linked to each other via some kind of middleware.
- Because the data are linked among the individual marts, there is a better chance of maintaining data consistency across the enterprise (at least at the metadata level).

Alternative DW Architectures

- **Hub-and-spoke architecture**. This is perhaps the most famous data warehousing architecture today.

- Here the attention is focused on building a scalable and maintainable infrastructure (often developed in an iterative way, subject area by subject area) that includes a centralized data warehouse and several dependent data marts (each for an organizational unit).

- This architecture allows for easy customization of user interfaces and reports. On the negative side, this architecture lacks the holistic enterprise view, and may lead to data redundancy and data latency.

Alternative DW Architectures

- **Centralized data warehouse.** The centralized data warehouse architecture is similar to the hub-and-spoke architecture except that there are no dependent data marts; instead, there is a gigantic enterprise data warehouse that serves the needs of all organizational units.

Alternative DW Architectures

- **Federated data warehouse**. The federated approach is a concession to the natural forces that undermine the best plans for developing a perfect system.
- It uses all possible means to integrate analytical resources from multiple sources to meet changing needs or business conditions.
- Essentially, the federated approach involves integrating disparate systems.

Ten factors that potentially affect the architecture selection decision

1. Information interdependence between organizational units
2. Upper management's information needs
3. Urgency of need for a data warehouse
4. Nature of end-user tasks
5. Constraints on resources
6. Strategic view of the data warehouse prior to implementation
7. Compatibility with existing systems
8. Perceived ability of the in-house IT staff
9. Technical issues
10. Social/political factors


Thank You!
Any questions?