

Lecture 3-4



Learning Objectives

- Understand the basic definitions and concepts of data warehouses
- Understand data warehousing architectures
- Describe the processes used in developing and managing data warehouses
- Explain data warehousing operations
- Explain the role of data warehouses in decision support

Data Warehousing

Definitions and Concepts

- **Data warehouse**

A physical repository where relational data are specially organized to provide enterprise-wide, cleansed data in a standardized format

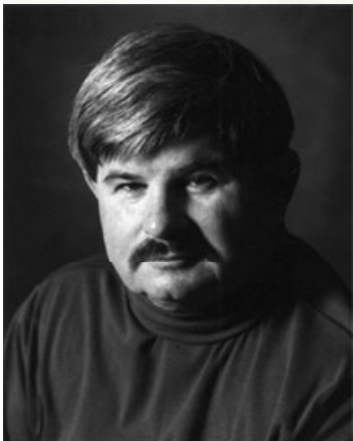
Data Warehousing

Definitions and Concepts

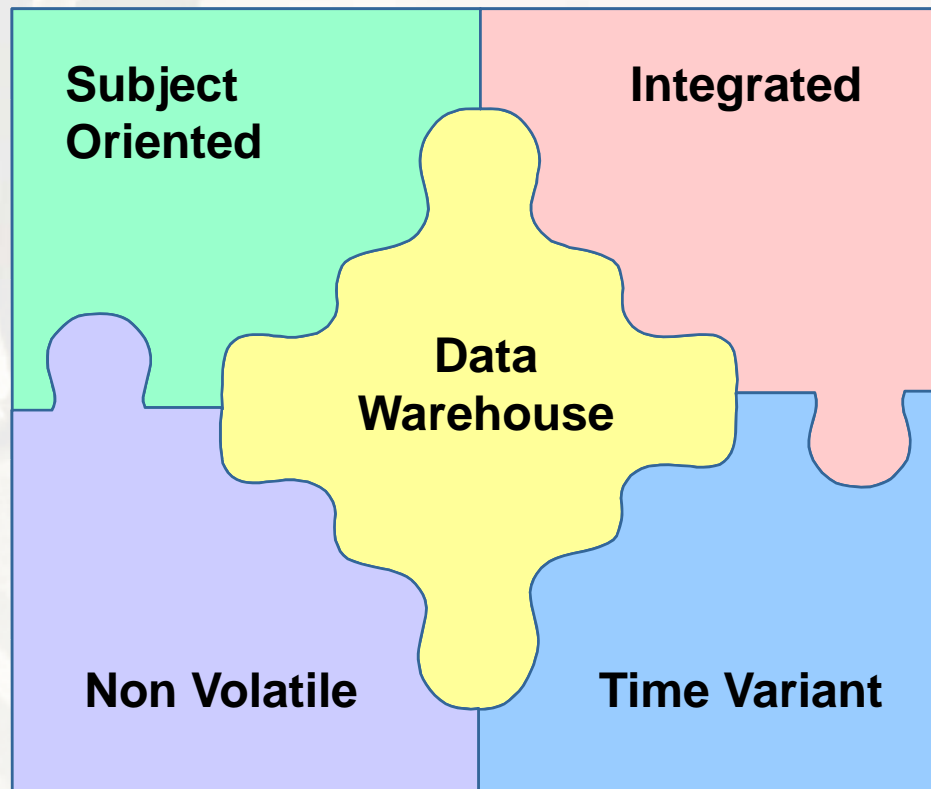
- Characteristics of data warehousing
 - Subject oriented
 - Integrated
 - Time variant (time series)
 - Nonvolatile
 - Web based
 - Relational/multidimensional
 - Client/server
 - Real-time
 - Include metadata

What is Data Warehouse?

- “A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management’s decision-making process.”
—W. H. Inmon, the father of the term ‘data warehouse’

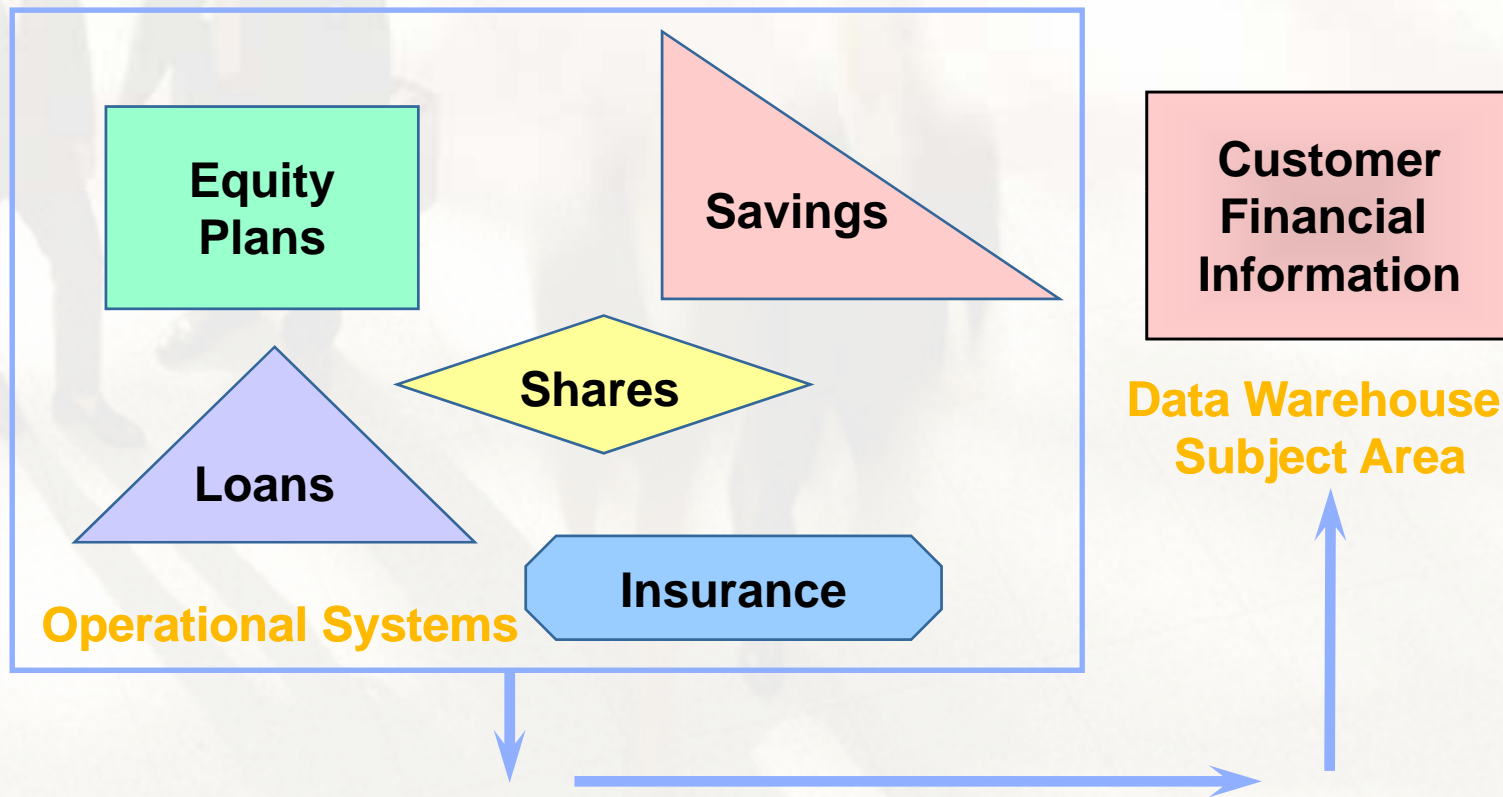


What is Data Warehouse?



DW: Subject-Oriented

Data is categorized and stored by business subject rather than by application

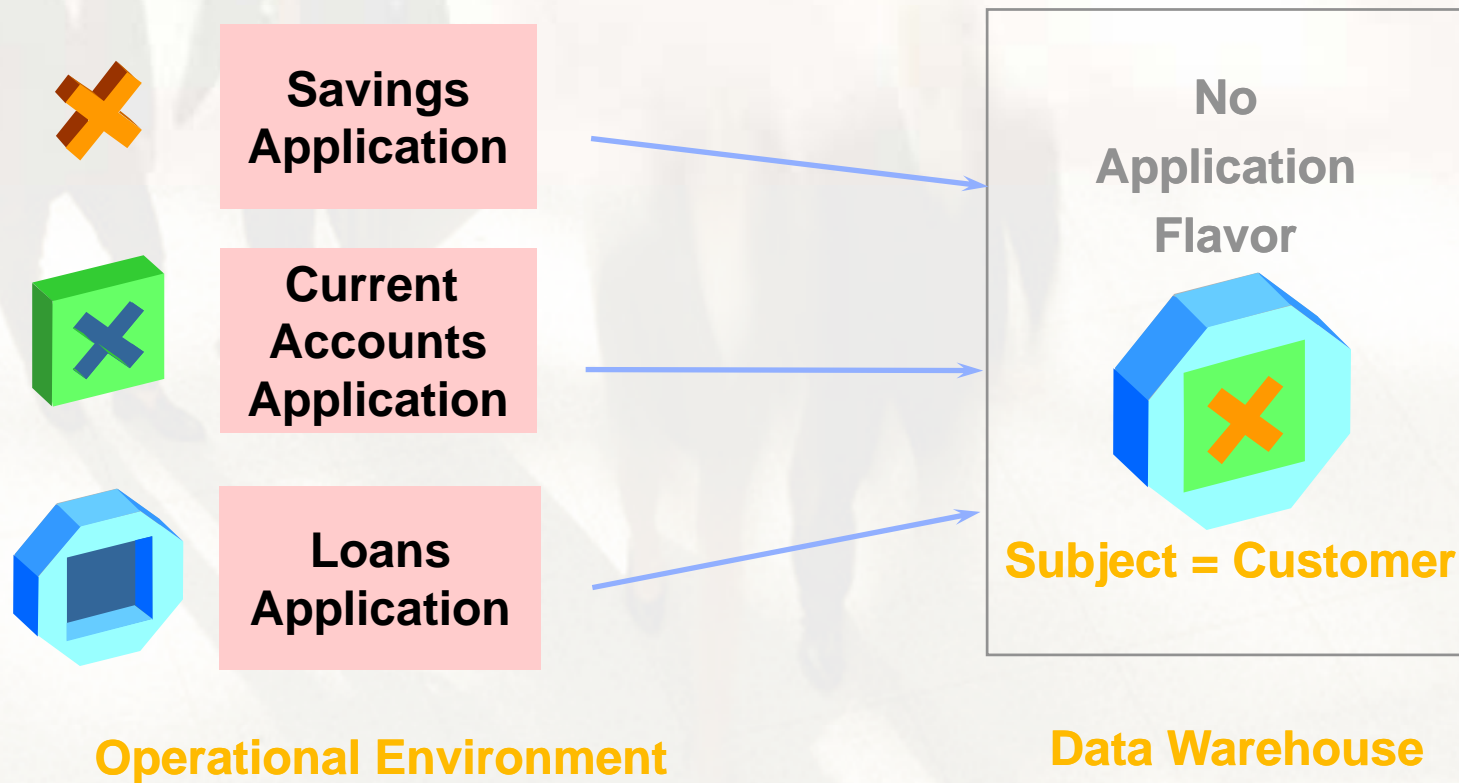


DW: Subject-Oriented

- 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**.

DW: Integrated

Data on a given subject is defined and stored once



DW: Integrated

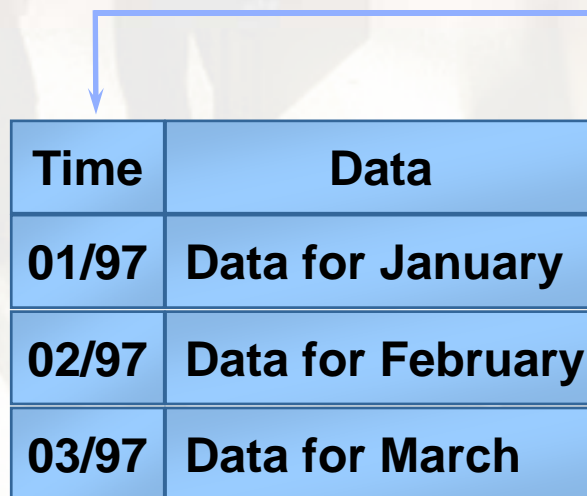
- Constructed by integrating multiple, heterogeneous data sources
 - relational databases, flat files, on-line transaction records
- One set of consistent, accurate, quality information
- Standardization
 - Naming conventions
 - Coding structures
 - Data attributes
 - Measures

DW: Integrated

- Data cleaning and data integration techniques are applied.
 - Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources
 - When data is moved to the warehouse, it is converted

DW: Time Variant

Data is stored as a series of snapshots, each representing a period of time



Time	Data
01/97	Data for January
02/97	Data for February
03/97	Data for March

**Data
Warehouse**

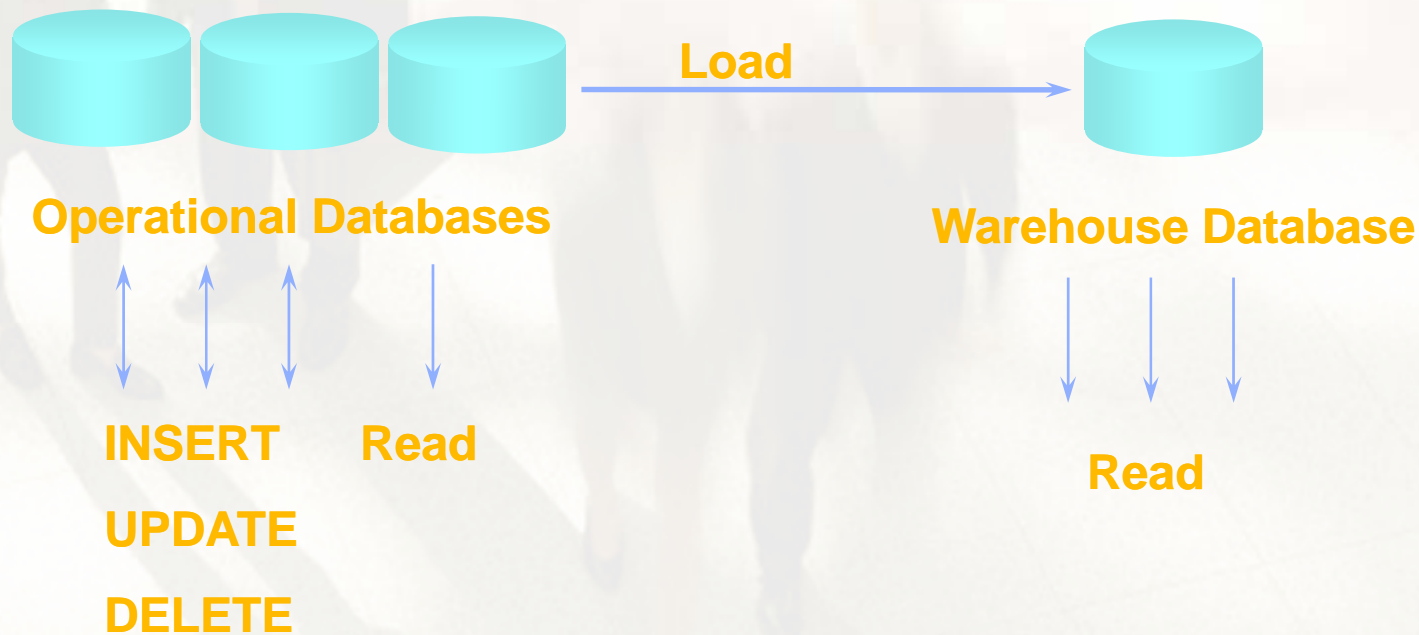


DW: Time Variant

- 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”

DW: Non-Volatile

Typically data in the data warehouse is **not** deleted



DW: 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*

Other definitions

- A decision-support database, which separately maintained from the operational database of the organization
 - S. Chaudhuri, U. Dayal, VLDB'96 tutorial
- A database specifically modeled and fine-tuned for analysis and decision making
- A single, integrated store of corporate data
- A body of data, extracted from a variety of sources
- . . . a strategic collection of **all types of data** in support of the decision-making process at **all levels of the enterprise**
 - Oracle datawarehouse

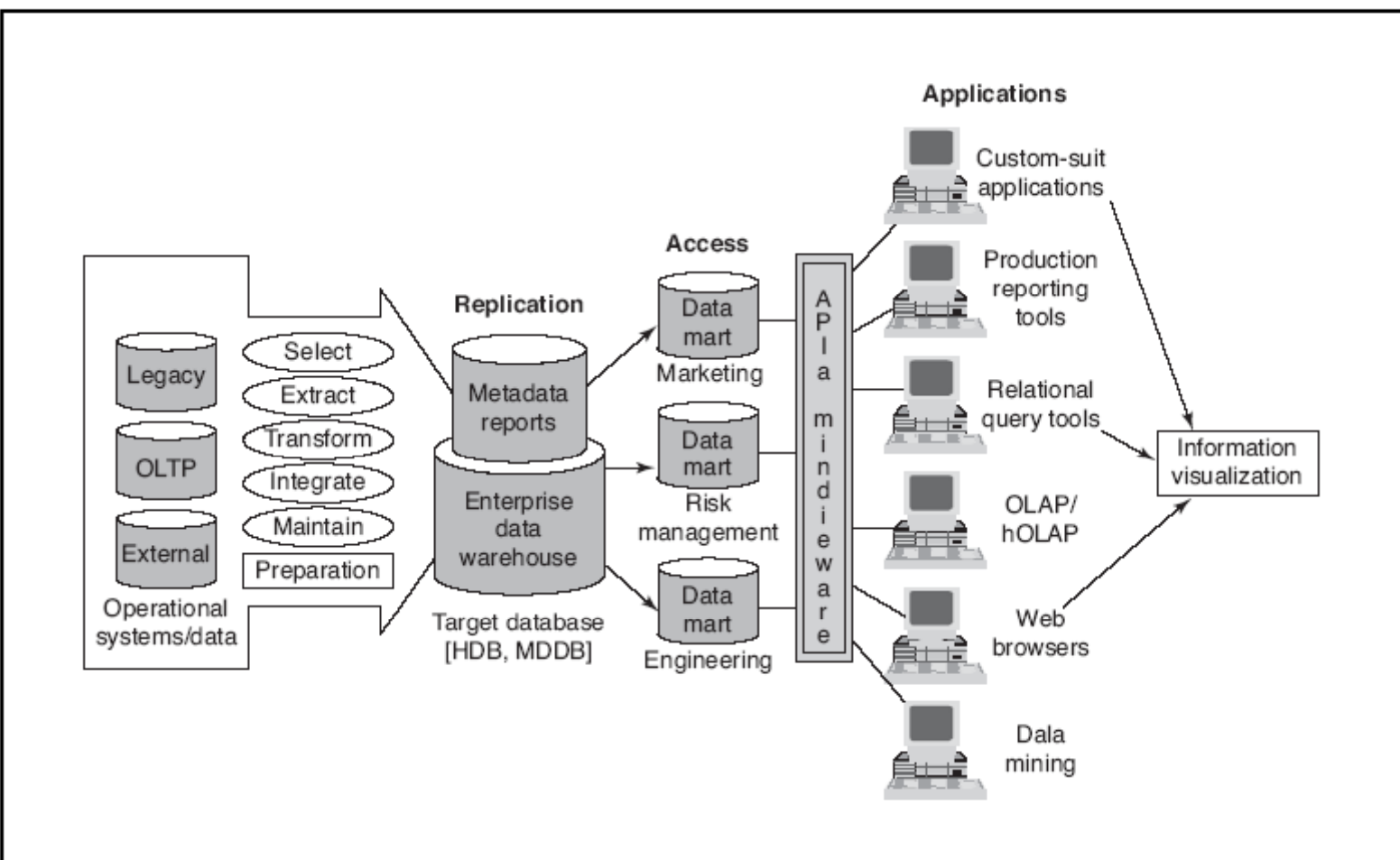
Other definitions

“A data warehouse is simply a single, complete, and consistent store of data obtained from a variety of sources and made available to end users in a way they can understand and use it in a business context.”

-- Barry Devlin, *IBM Consultant*

Data Warehousing Process Overview

FIGURE 5.1 Data Warehouse Framework and Views



Data Warehousing

Definitions and Concepts

- **Data mart**

A departmental data warehouse that stores only relevant data

- **Dependent data mart**

A subset that is created directly from a data warehouse

- **Independent data mart**

A small data warehouse designed for a strategic business unit or a department

Data Warehousing

Definitions and Concepts

- **Operational data stores (ODS)**

A type of database often used as an interim area for a data warehouse, especially for customer information files

- **Oper marts**

An operational data mart. An oper mart is a small-scale data mart typically used by a single department or functional area in an organization

Data Warehousing

Definitions and Concepts

- **Enterprise data warehouse (EDW)**

A technology that provides a vehicle for pushing data from source systems into a data warehouse

- **Metadata**

Data about data. In a data warehouse, metadata describe the contents of a data warehouse and the manner of its use

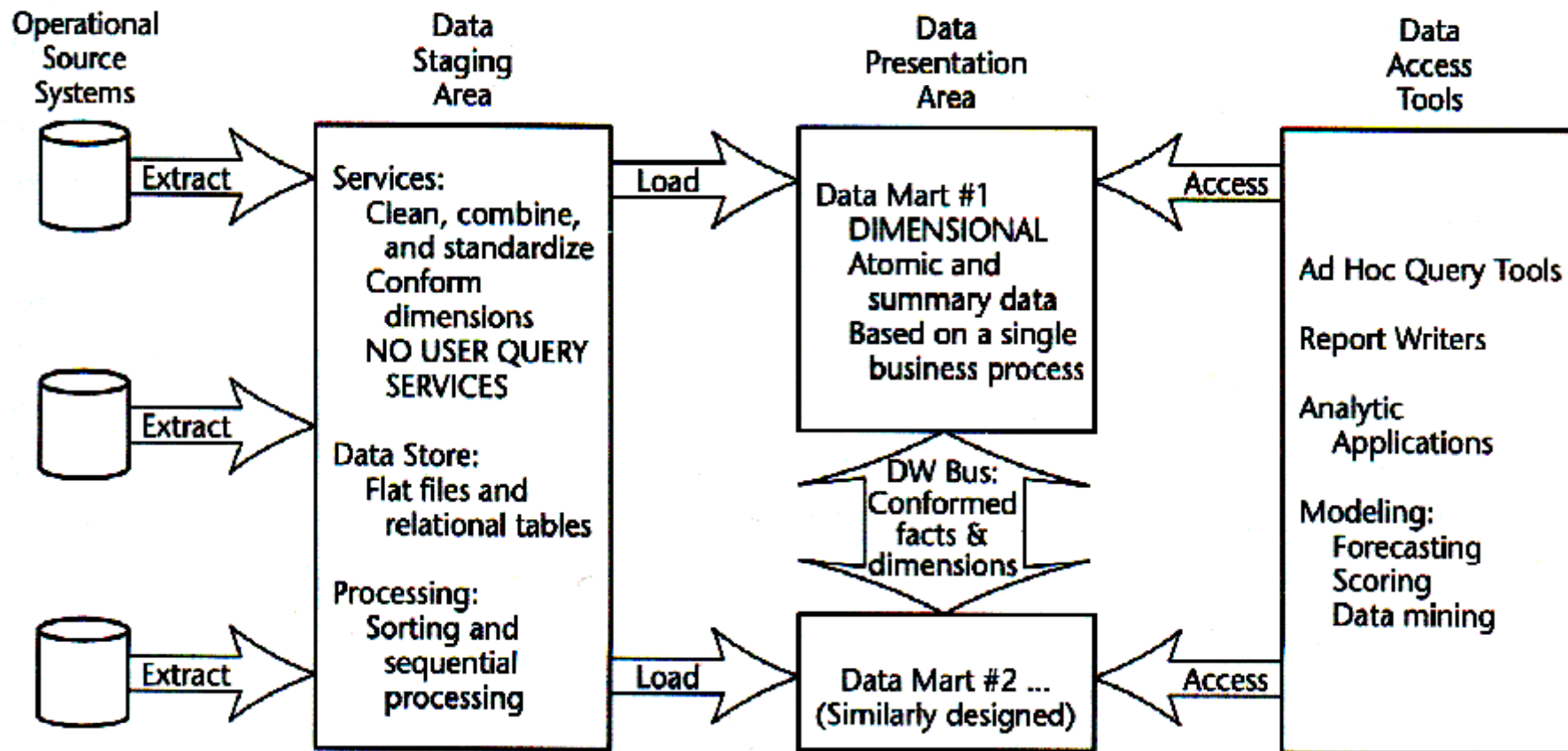
Data Warehousing Process Overview

- Organizations continuously collect data, information, and knowledge at an increasingly accelerated rate and store them in computerized systems
- The number of users needing to access the information continues to increase as a result of improved reliability and availability of network access, especially the Internet

Data Warehousing Process Overview

- The major components of a data warehousing process
 - Data sources
 - Data extraction
 - Data loading
 - Comprehensive database
 - Metadata
 - Middleware tools

DW components (according to Kimball)



Operational Source Systems

- Each source system is a stovepipe application, little investment to sharing common data (e.g., product, customer)
- Reengineering with consistent view would be great
 - Enterprise Application Integration (EAI) effort will make the pass to DW more easy

Data staging area

- It is off-limits to the business users
- Does *not* provide query and presentation services
- Normalization is not the end goal
 - Normalized databases are excluded from the presentation area, so no need to normalize in data staging area

Data presentation

- Where the data are organized, stored, made available for querying
- This *is* the data warehouse for business community (remember, they can't see data staging area)
- Series of integrated data marts
- A data mart presents the data from a single business process
 - Business processes cross the boundaries of organizational functions

Data presentation

- Data must be stored and accessed in *dimensional schemas*
 - No normalization (3NF) should be used
 - Dimensional schemas are simple and intuitive for business users; Normalized schemas are difficult to grasp by them
- Data must be atomic (at lower granularity)
 - Not only summarized – they don't allow for arbitrary, complex queries
- Data marts must be build on *dimensions and facts* that are conformed
 - Otherwise, data marts are stovepipes
 - Conformation leads to bus architecture – data marts can cooperate

Data access tools

- Ad hoc, complex queries are targeted to small percentage of business users
- 80-90% of the potential users will be served by 'canned' applications
 - Canned: pre-build parameter-driven analytic applications

Again on Metadata and ODS

- Metadata repository
 - Management of metadata
- Operational Data Store (ODS)
 - Management of almost real-time data

Metadata Repository

- Meta data is the data defining warehouse objects. It has the following kinds
 - Description of the structure of the warehouse
 - schema, view, dimensions, hierarchies, derived data defn, data mart locations and contents
 - Operational meta-data
 - data lineage (history of migrated data and transformation path), currency of data (active, archived, or purged), monitoring information (warehouse usage statistics, error reports, audit trails)
 - The algorithms used for summarization (measures, gran, etc)
 - The mapping from operational environment to the data warehouse
 - Data related to system performance
 - warehouse schema, view and derived data definitions
 - Business data
 - business terms and definitions, ownership of data, charging policies₁

Importance of metadata repository

- Ultimate goal of metadata repositories: to corral, catalog, integrate, and leverage the disparate varieties of metadata (like the resources of a library)
- The task looms, but we can't ignore it
- Need to develop an overall metadata plan

Standards for metadata

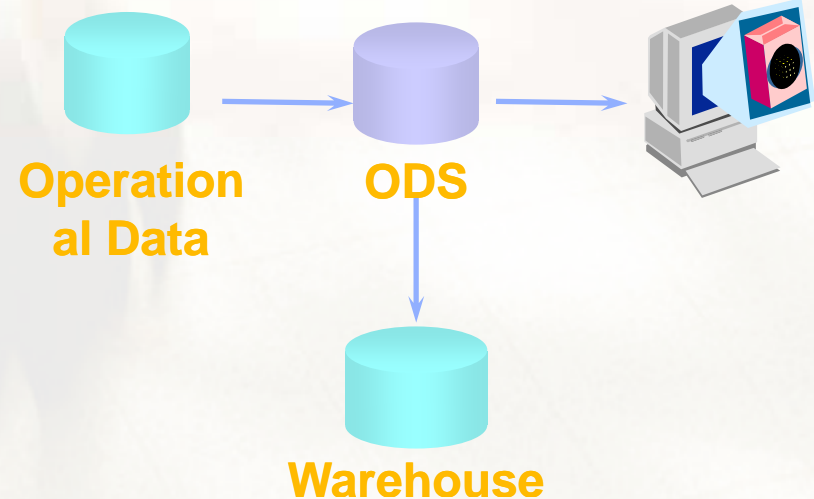
- Metadata Coalition proposed:
 - MetaData Interchange Specification (MDIS)
 - Open Information Model (OIM)
- OMG
 - Common Warehouse Model (CWM)
- Microsoft Repository (in the Office suit)
 - Includes some simple Information Models for data warehousing

Operational Data Store (ODS)

- What is it?
- Do we need it?
- If yes, when we need it?
- How is it usually implemented?

What is ODS?

- There is no single definition for the ODS, if and where to belong to a DW
- ODS is a frequently updated, *somewhat* integrated copies of operational data
- It stands between operational databases and the rest of DW
- The frequency of update and degree varies on application requirements



When we need an ODS?

- ODS is implemented to deliver operational reporting, especially when the OLTP operational databases do not provide any reporting capabilities
- The reports address the organization's tactical requirements, especially those that need the most current information
 - No aggregation abilities, much simpler than the OLAP analysis

How is ODS usually implemented?

- In CRM (Customer Relationship Management), especially in the case of e-commerce, need near-real-time data
- In such cases, they are implemented within the DW
 - ODS then feeds the DW
- Alternatively, ODS is a third, physically separated system
- Conclusion: Get an ODS when OLTP operational databases and the DW cannot answer immediate operational questions (if you need them)

Data Warehousing Architectures

- Three parts of the data warehouse
 - The data warehouse that contains the data and associated software
 - Data acquisition (back-end) software that extracts data from legacy systems and external sources, consolidates and summarizes them, and loads them into the data warehouse
 - Client (front-end) software that allows users to access and analyze data from the warehouse

Data Warehousing Process Overview

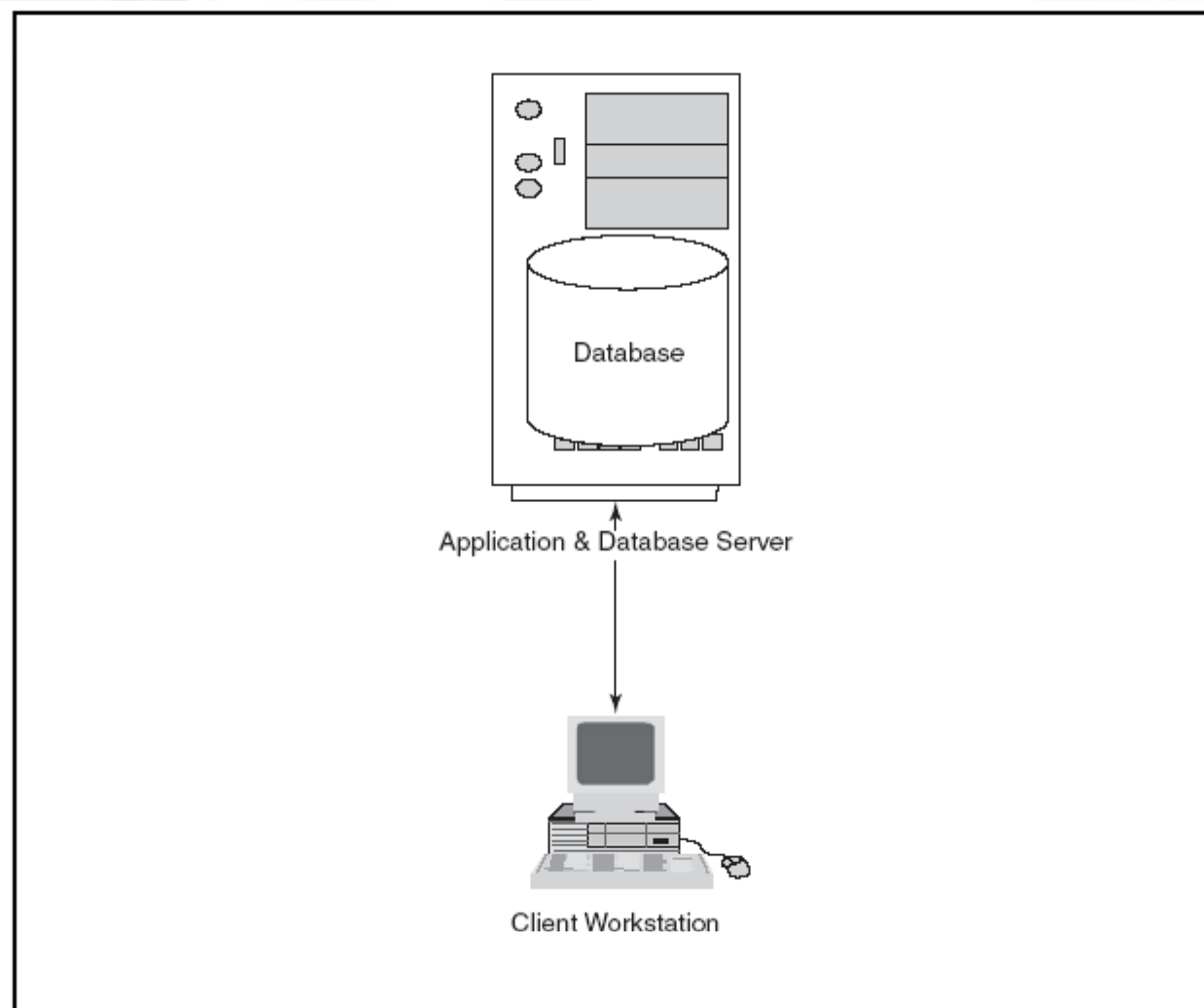


FIGURE 5.3 Architecture of a Two-Tier Data Warehouse

Data Warehousing Process Overview

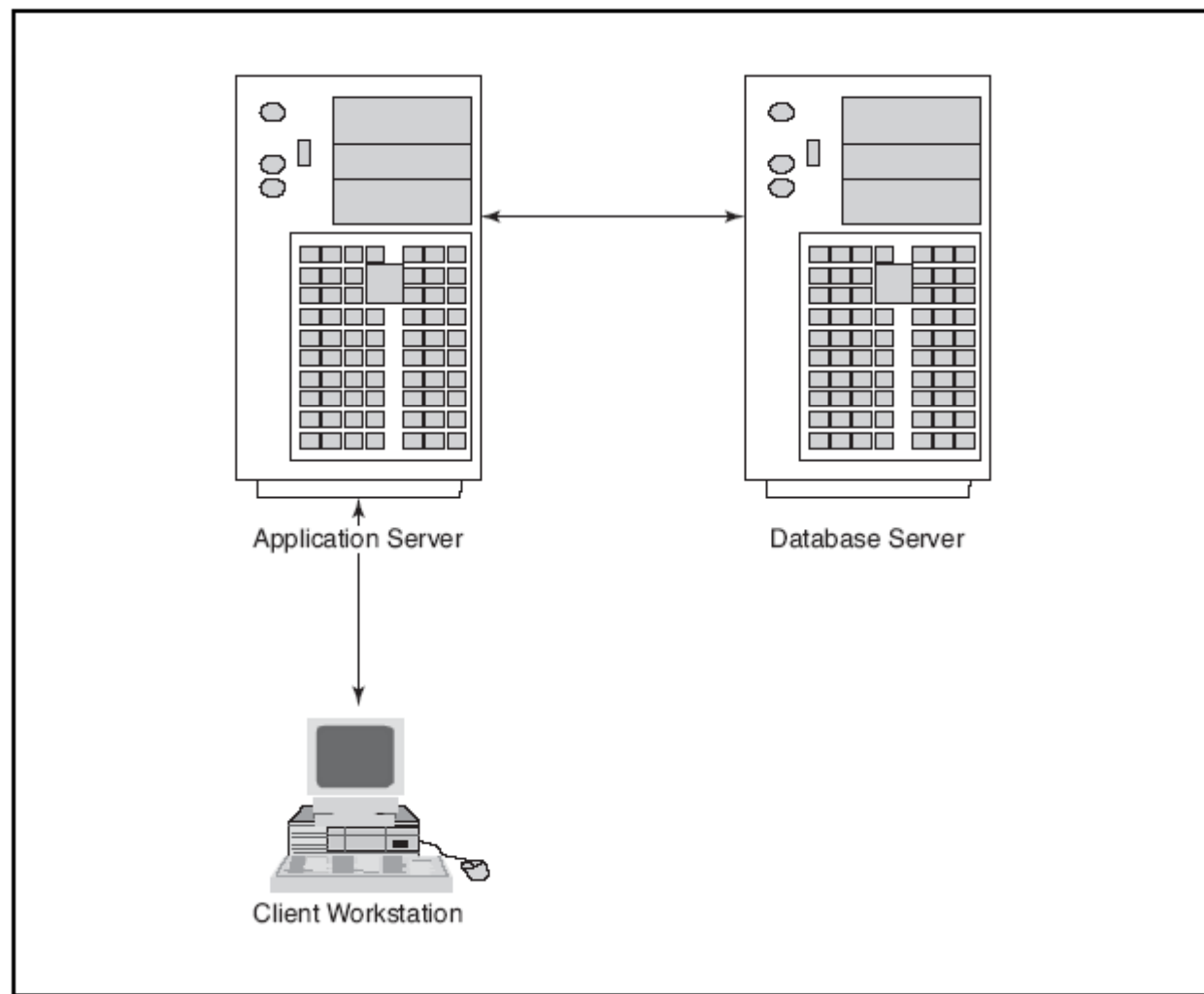
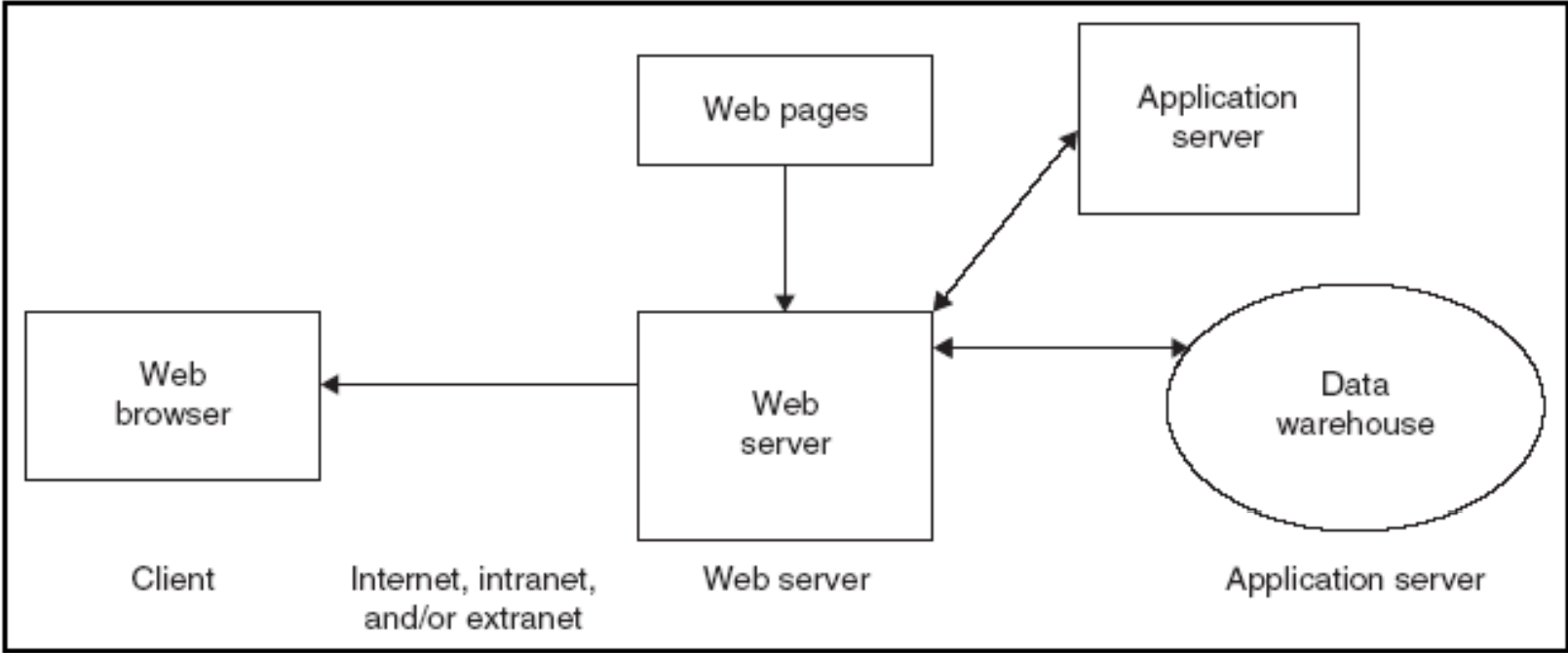


FIGURE 5.2 Architecture of a Three-Tier Data Warehouse

Data Warehousing Process Overview

FIGURE 5.4 Architecture of Web-Based Data Warehousing



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?*

Data Warehousing Process Overview

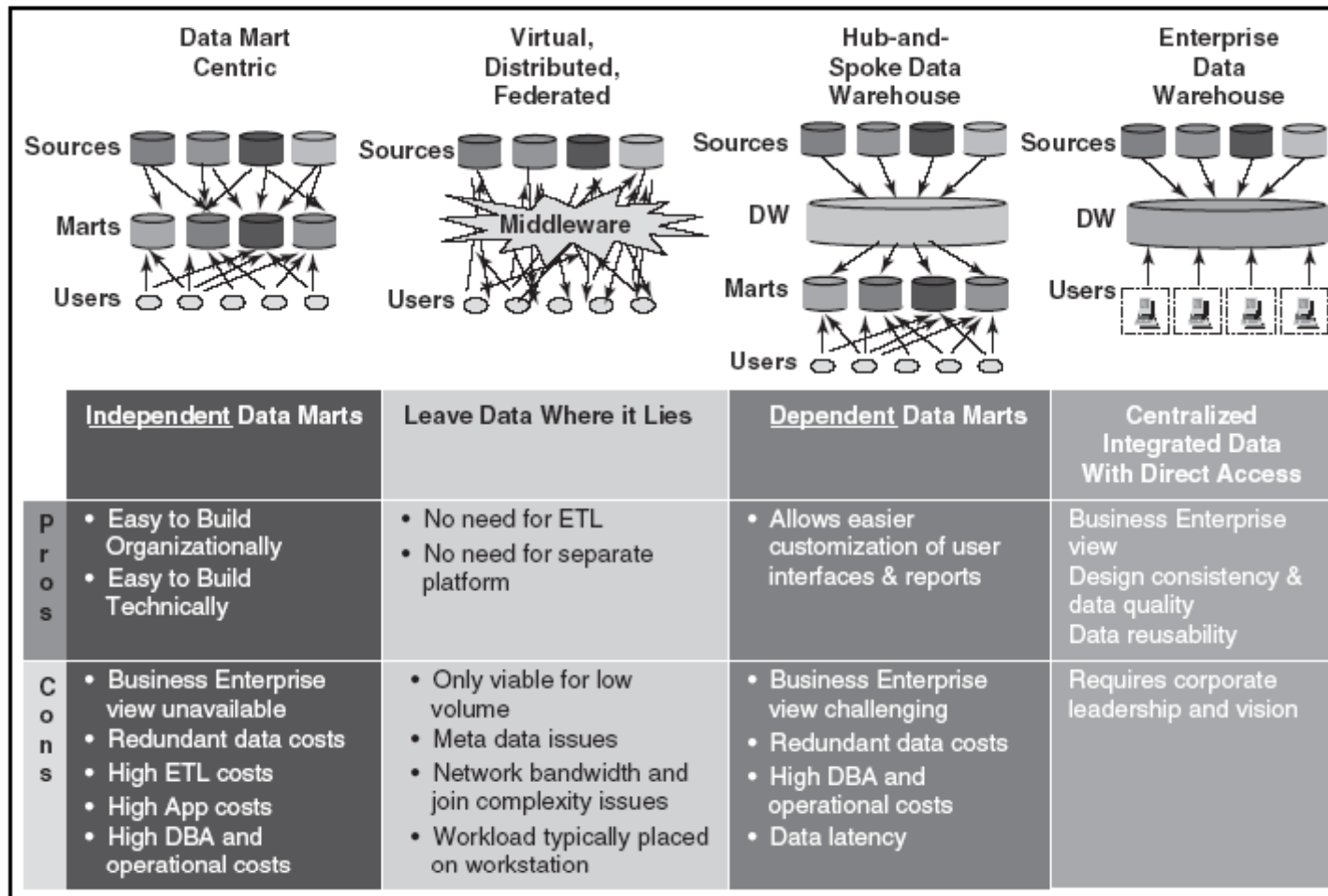


FIGURE 5.6 Alternative Architectures for Data Warehousing Efforts

Data Warehousing Process Overview

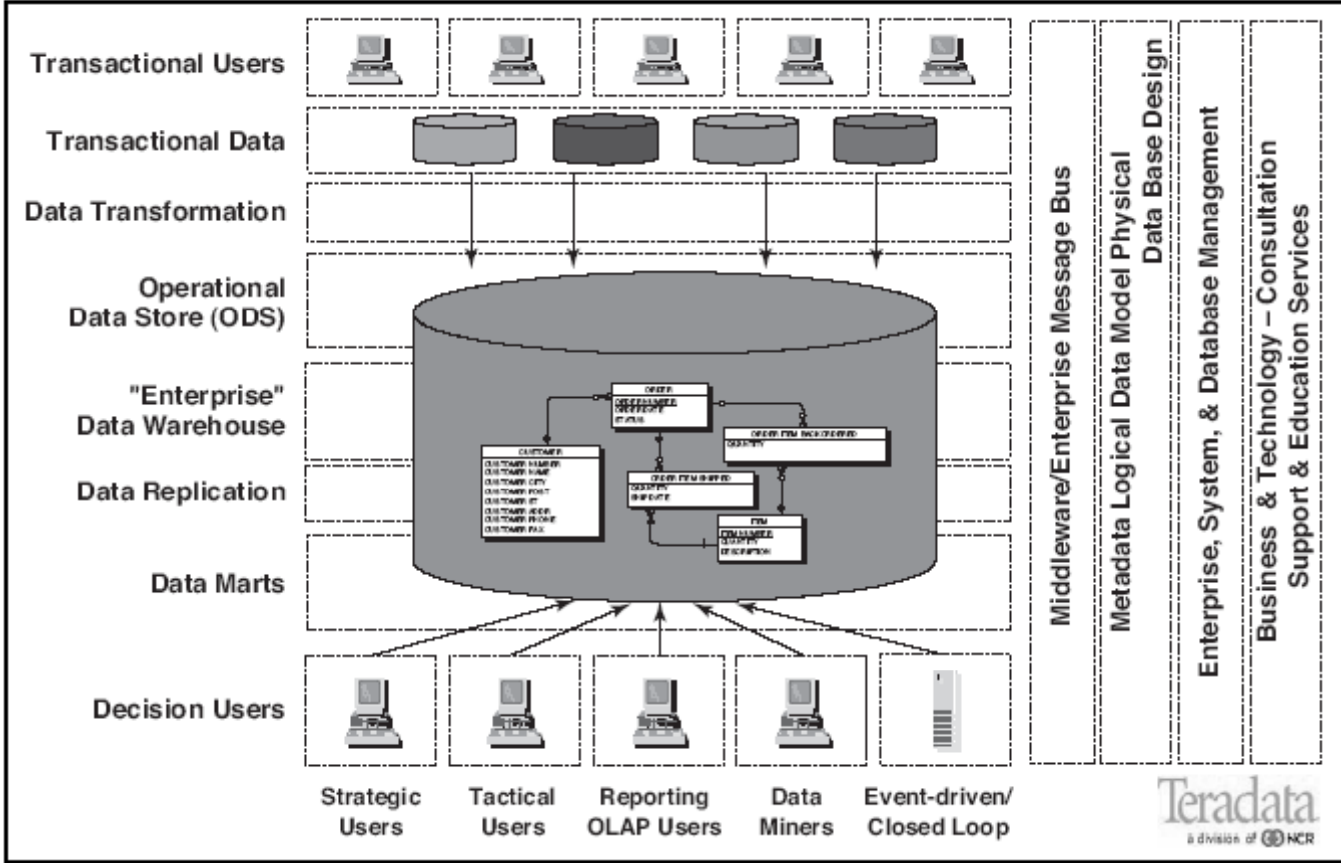


FIGURE 5.7 Teradata Corp.'s Enterprise Data Warehouse

Case study from Teradata

- http://searchdatamanagement.techtarget.com/news/article/0,289142,sid91_gci1137386,00.html

Distributed DW

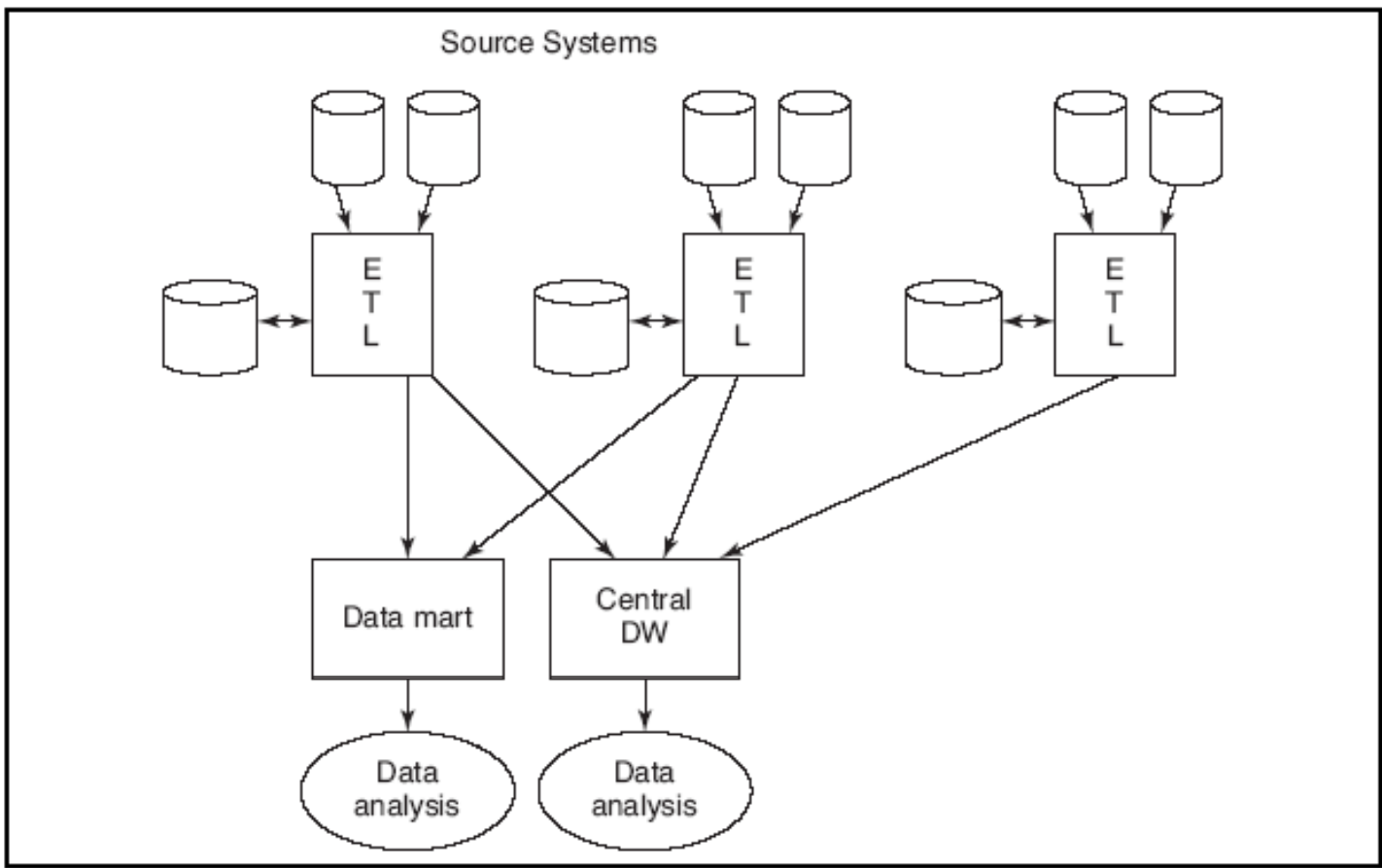


FIGURE 5.5e Distributed Data Warehouse Architecture

Data Warehousing Architectures

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

Controversial issues

- http://en.wikipedia.org/wiki/Data_warehouse