



DATA WAREHOUSING

Overview & Concepts

Compelling need for Data Warehousing

- Evolution in data processing
- OLTP : On Line Transaction Processing
- Reporting : Decision Making
- System provides only operational info
- Business grows more complex, globally
- New type of info : Strategic Information

Areas of Strategic Information

◆ Retail

- ◆ Customer Loyalty
- ◆ Market Planning

◆ Financial

- ◆ Risk Management
- ◆ Fraud Detection

◆ Airlines

- ◆ Route Profitability
- ◆ Yield Management

◆ Manufacturing

- ◆ Cost Reduction
- ◆ Logistics Management

◆ Utilities

- ◆ Asset Management
- ◆ Resource Management

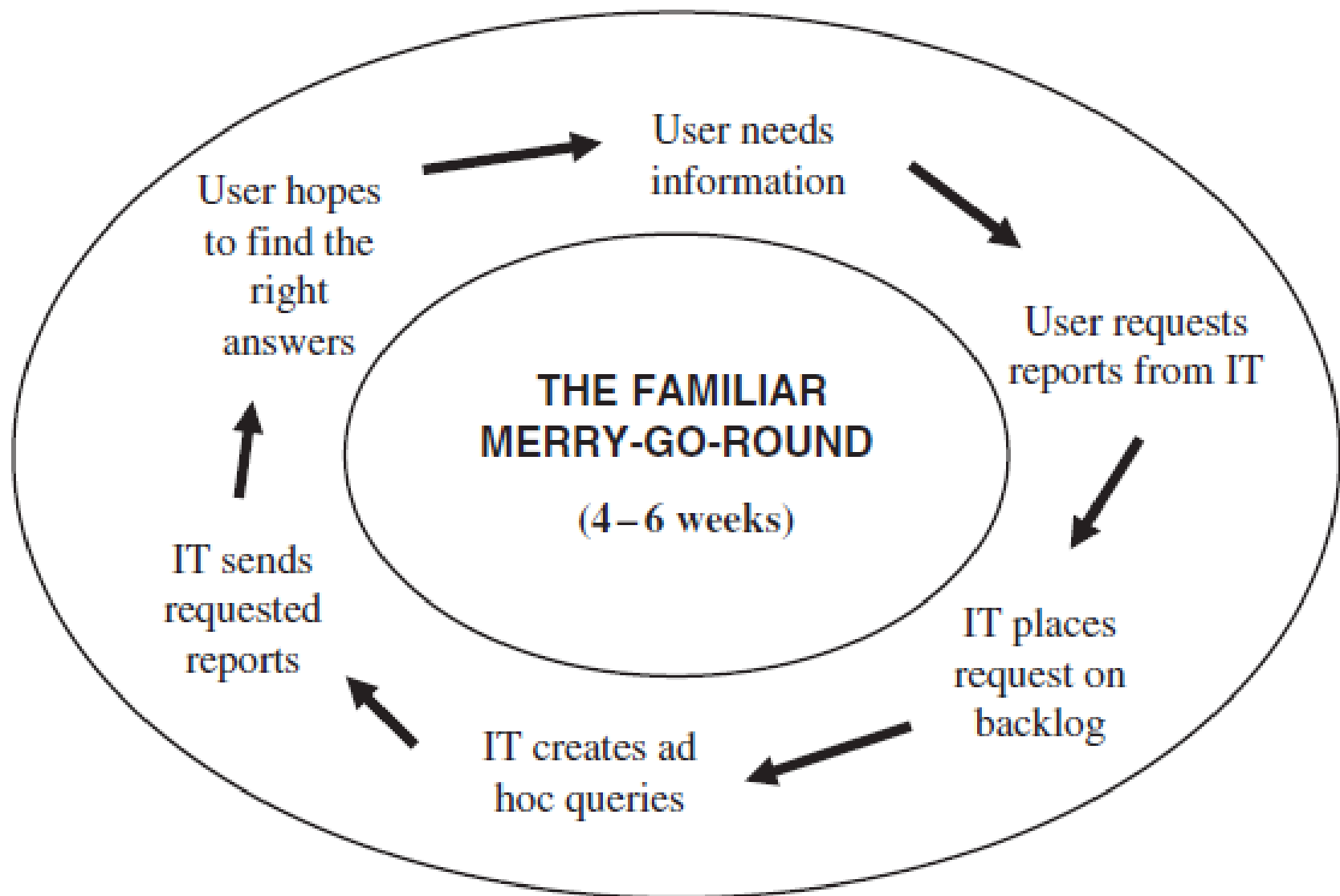
◆ Government

- ◆ Manpower Planning
- ◆ Cost Control

Characteristics of strategic information

- **INTEGRATED**
 - Must have a single, enterprise-wide view.
- **DATA INTEGRITY**
 - Information must be accurate and must conform to business rules.
- **ACCESSIBLE**
 - Easily accessible with intuitive access paths, and responsive for analysis.
- **CREDIBLE**
 - Every business factor must have one and only one value.
- **TIMELY**
 - Information must be available within the stipulated time frame.

FAILURES OF PAST DECISION-SUPPORT SYSTEMS



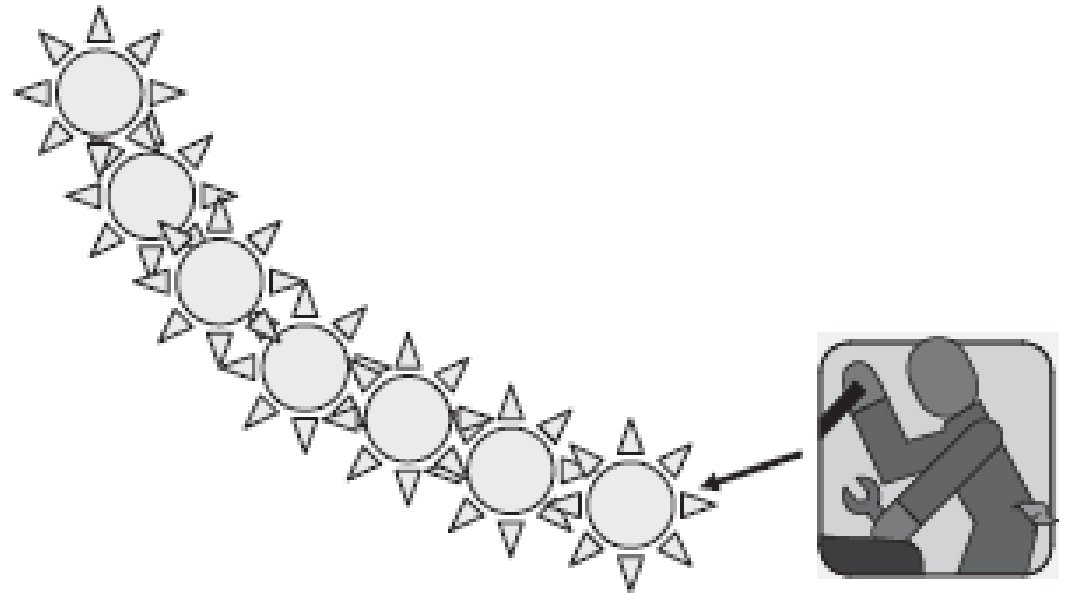
OPERATIONAL VERSUS DECISION-SUPPORT SYSTEMS

Operational System

Get the data in

Making the wheels of business turn

- ◆ Take an order
- ◆ Process a claim
- ◆ Make a shipment
- ◆ Generate an invoice
- ◆ Receive cash
- ◆ Reserve an airline seat



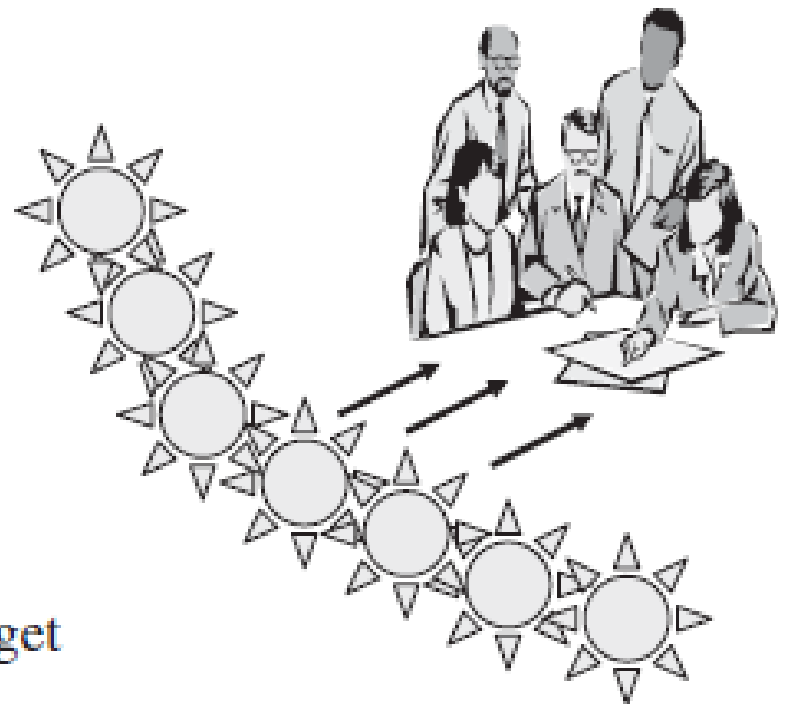
OPERATIONAL VERSUS DECISION-SUPPORT SYSTEMS

Decision Support System

Get the information out

Watching the wheels of business turn

- ◆ Show me the top-selling products
- ◆ Show me the problem regions
- ◆ Tell me why (drill down)
- ◆ Let me see other data (drill across)
- ◆ Show the highest margins
- ◆ Alert me when a district sells below target



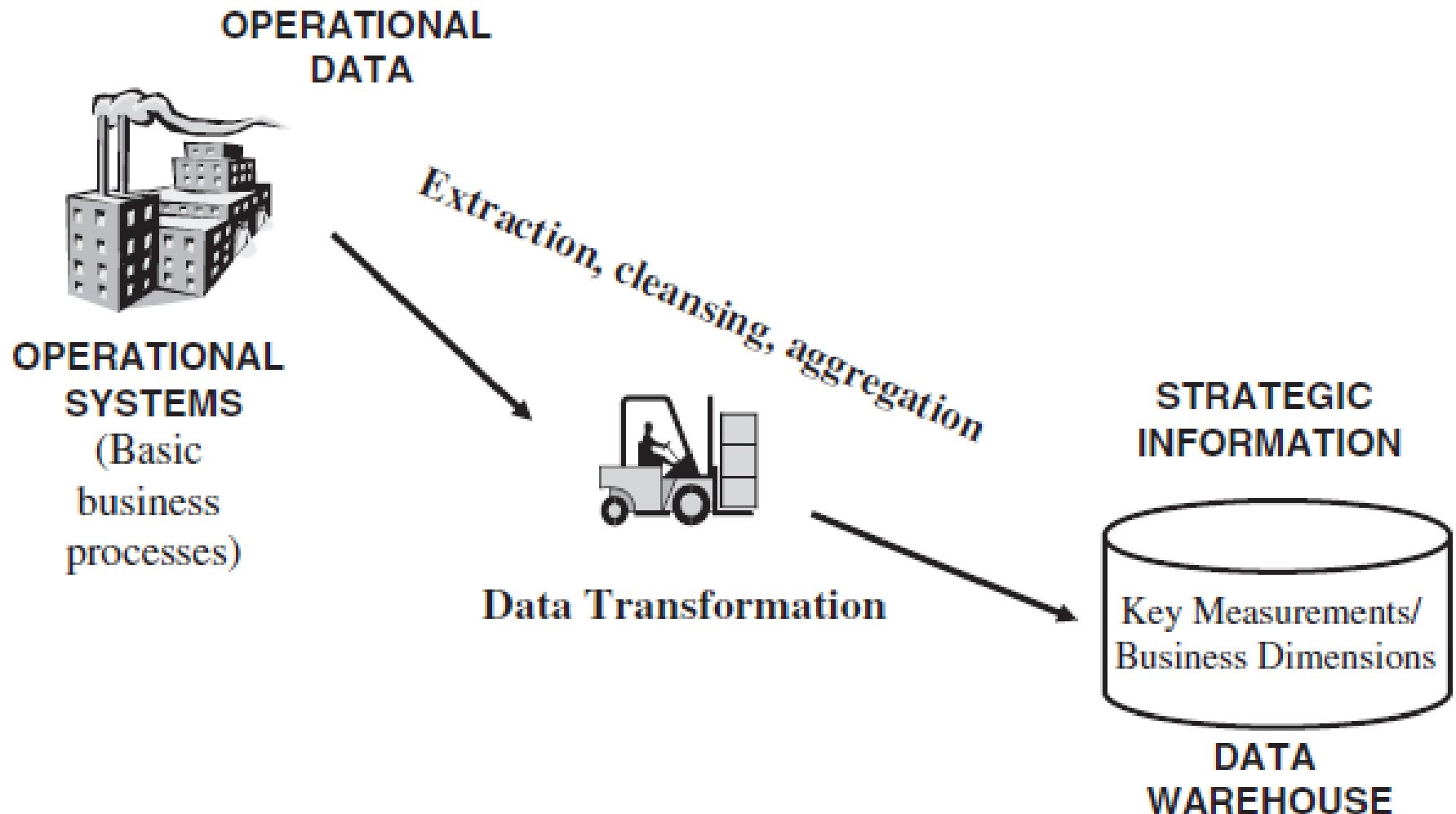
DATA WAREHOUSING

**THE ONLY VIABLE
SOLUTION**

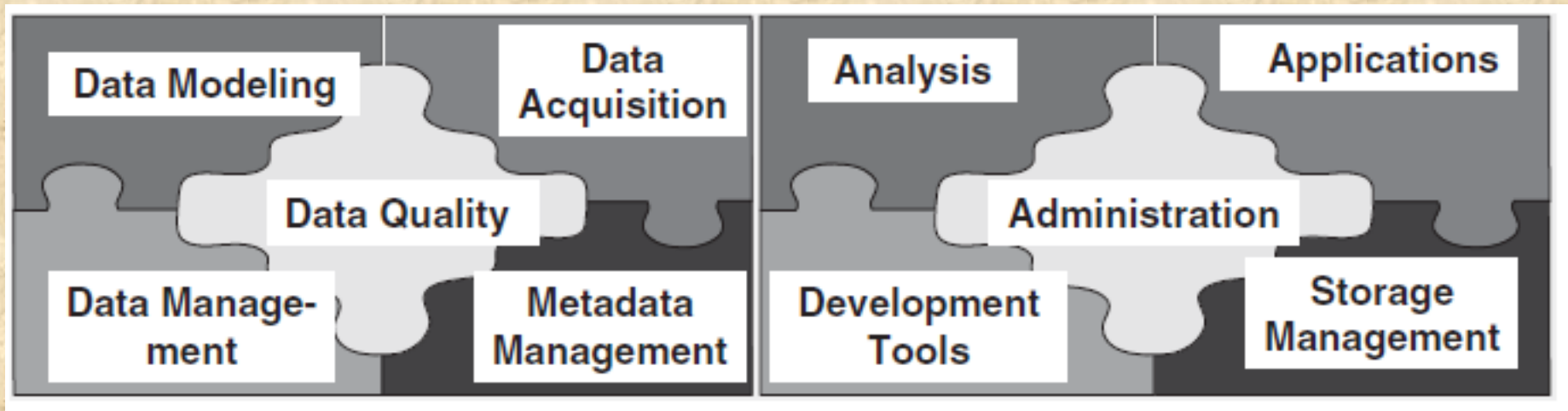
What is Data Warehouse?

- Defined in many different ways, but not rigorously.
 - 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.”—W. H. Inmon
- Data warehousing:
 - The process of constructing and using data warehouses

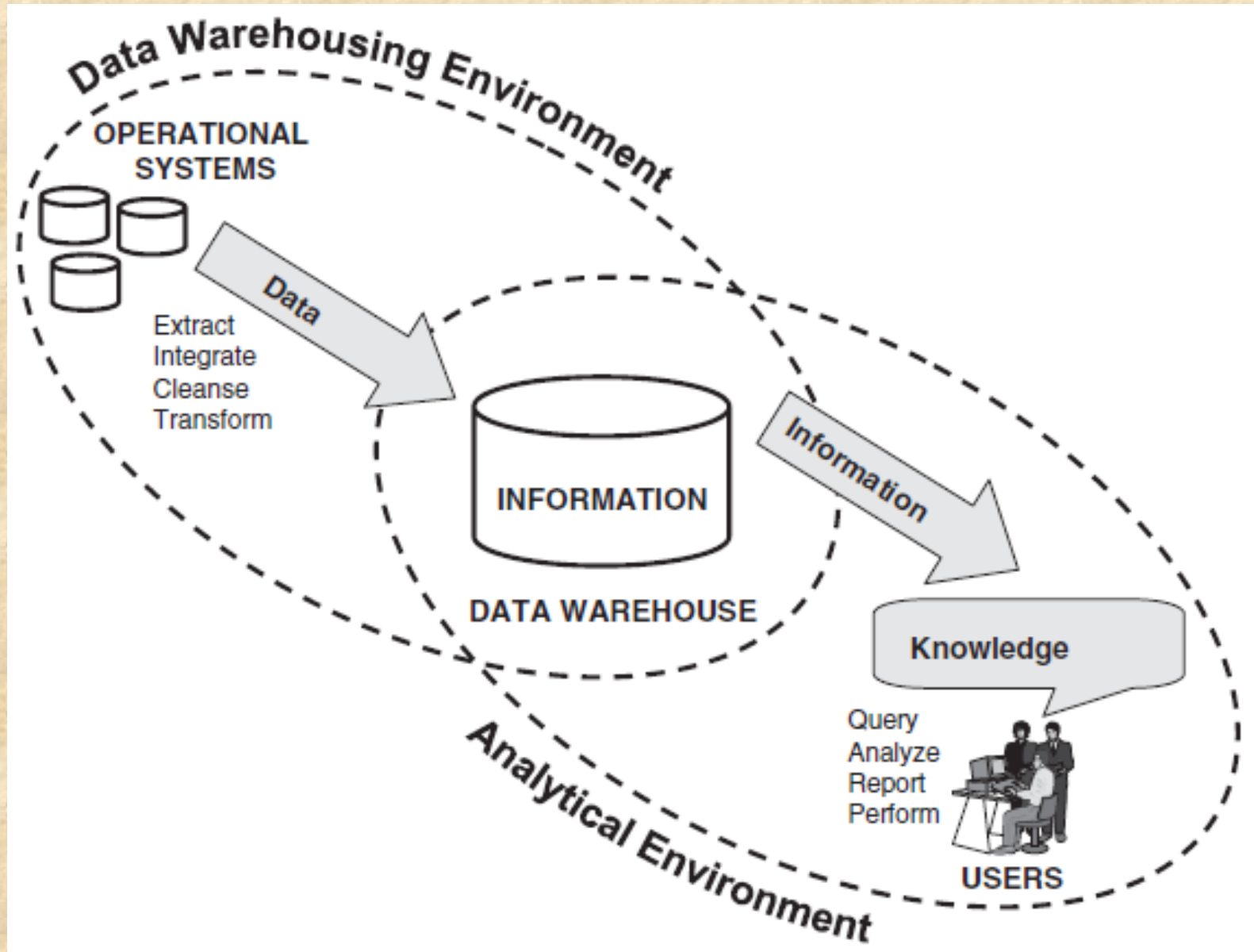
General overview of the data warehouse



The data warehouse: a blend of technologies



Data Warehousing and Analytics



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.

Business Intelligence [BI]

Definition

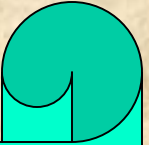
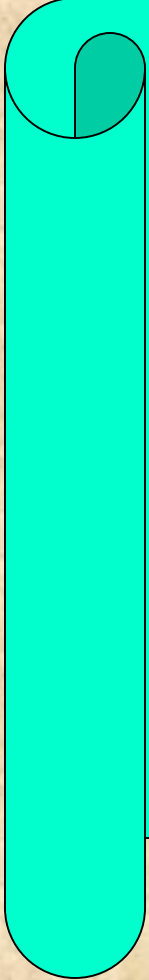
Business Intelligence (BI) is the gathering and analysis of vast amounts of data in order to gain insights that drive strategic and tactical business decisions. It encompasses a broad category of technologies, that allow business users to gather, store, access, and analyze data to improve the business decision-making capabilities.

BI is the collection of the processes and technologies which transform data into information



Comparision of std. DBMS and data warehouse

Std. DBMS	Data warehouse
Raw Data	Summarized / consolidated data
Hold current data	Holds historical data
Stores detailed data	Stores detailed, lightly, and highly summarized data
Data is dynamic	Data is largely static
Repetitive processing	Ad hoc, unstructured, and heuristic processing
High level of transaction throughput	Medium to how level of transaction throughput
Predictable pattern of usage	Unpredictable pattern of usage
Transaction-driven	Analysis driven
Application-orented	Subject-oriented
Supports day-to-day decisions	supports strategic decisions
Database size MBs to GBs	Database size : GBs to TBs
Thousands of users (End users / operators)	Hundreds of users (Managers / decision makers / analysts)



DATA WAREHOUSE

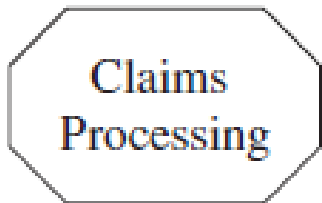
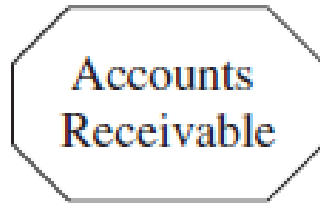
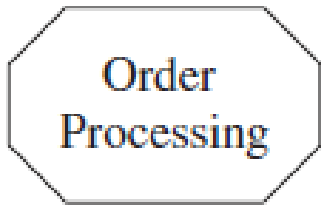
BUILDING BLOCKS

Data Warehouse—Subject-Oriented

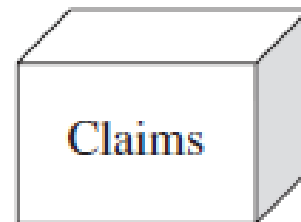
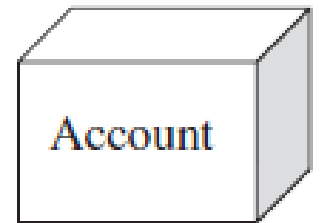
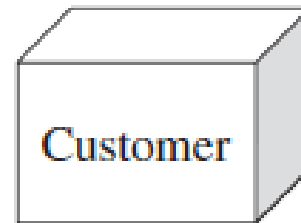
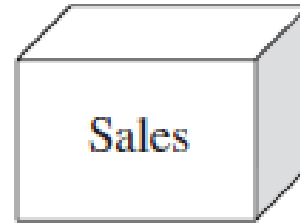
- Organized around major subjects, such as customer, product, sales.
- Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing.
- Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.

Examples

Operational Applications



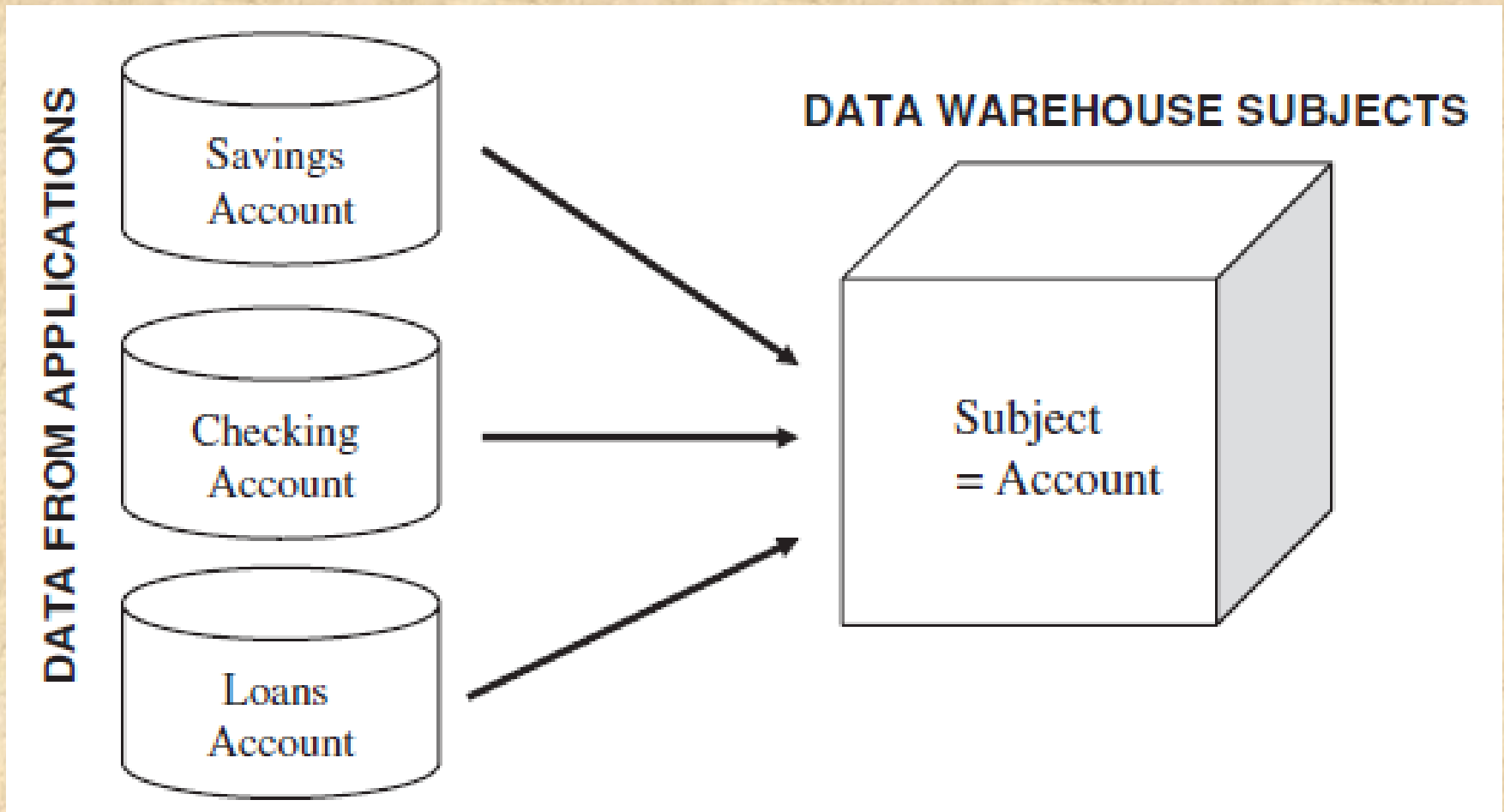
Data Warehouse Subjects



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.

Example



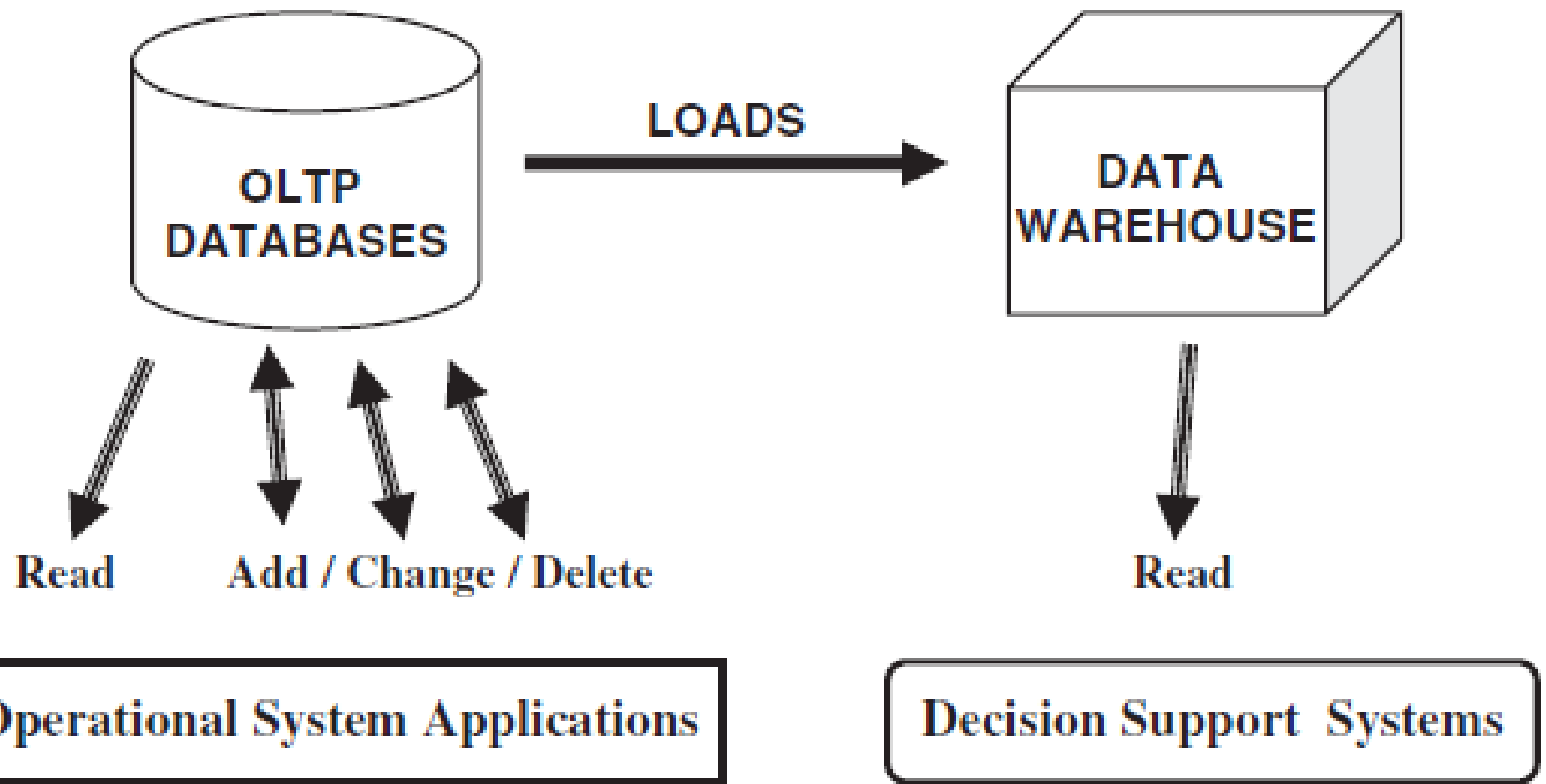
Data Warehouse—Time Variant

- 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.*

Example



Exercise 1.1

**A data warehouse is subject-oriented.
What would be the major critical
business subjects for the following
companies?**

- a)an international manufacturing company**
- b)a local community bank**
- c)a domestic hotel chain**

Data Granularity

- OLTP stores the data at lower level [raw]
- While reporting / Querying, summary of data
- **Data granularity** in a data warehouse refers to the **level of detail**.
- Depending on the requirements, multiple levels of detail may be present
- Lower the level of detail, the finer is the data granularity
- Granularity levels are decided based on the data types and the expected system performance for queries
- Many data warehouses have at least dual levels of granularity.

Example

THREE DATA LEVELS IN A BANKING DATA WAREHOUSE

Daily Detail

Account

Activity Date

Amount

Deposit/Withdrawal

Monthly Summary

Account

Month

Number of transactions

Withdrawals

Deposits

Beginning Balance

Ending Balance

Quarterly Summary

Account

Quarter

Number of transactions

Withdrawals

Deposits

Beginning Balance

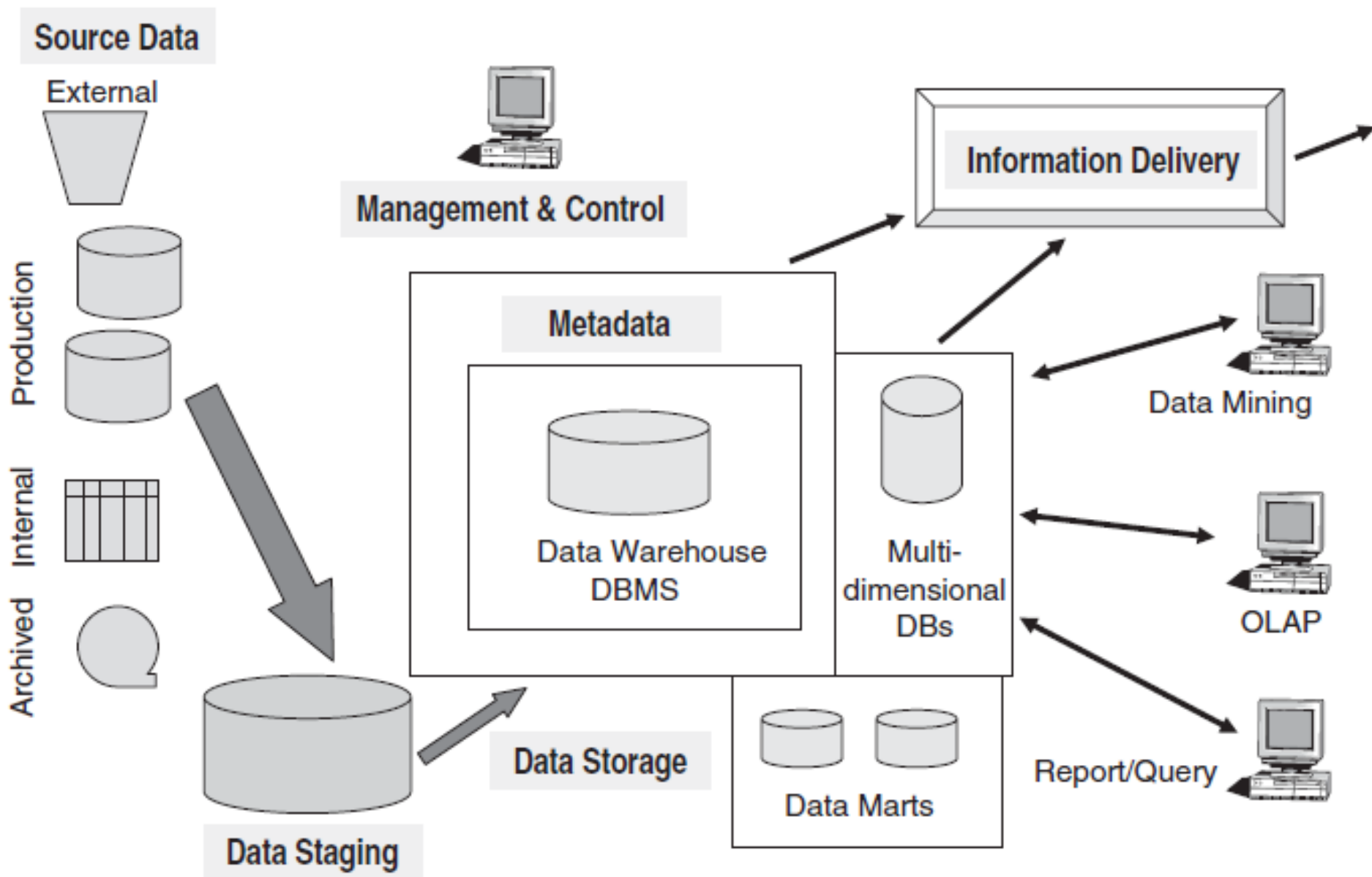
Ending Balance

OLTP

**Fine granularity :
Lower Level**

**Data Warehouse
High level granularity**

Data warehouse: building blocks or components



DATA WAREHOUSES AND DATA MARTS

DATA WAREHOUSE	DATA MART
<ul style="list-style-type: none">• Corporate/Enterprise-wide	<ul style="list-style-type: none">• Departmental
<ul style="list-style-type: none">• Union of all data marts	<ul style="list-style-type: none">• A single business process
<ul style="list-style-type: none">• Data received from staging area	<ul style="list-style-type: none">• STARjoin (facts & dimensions)
<ul style="list-style-type: none">• Queries on presentation resource	<ul style="list-style-type: none">• Technology optimal for data access and analysis
<ul style="list-style-type: none">• Structure for corporate view of data	<ul style="list-style-type: none">• Structure to suit the departmental view of data
<ul style="list-style-type: none">• Organized on E-R model	

Design issues in building DW

- Top-down or bottom-up approach?
- Enterprise-wide or departmental?
- Which first—data warehouse or data mart?
- Build pilot or go with a full-fledged implementation?
- Dependent or independent data marts?

Top-Down Versus Bottom-Up Approach

Top-Down Approach :

- data warehouse as a centralized repository for the entire enterprise
- data in the data warehouse is stored at the lowest level of granularity based on a normalized data model
- “Corporate Information Factory” (CIF) providing the logical framework for delivering business intelligence to the enterprise

Top-Down Versus Bottom-Up Approach

Bottom-Up Approach:

- data warehouse as a collection of conformed data marts.
- data marts are created first to provide analytical and reporting capabilities for specific business subjects based on the dimensional data model
- Data marts contain data at the lowest level of granularity and also as summaries.

Practical (Hybrid) Approach

- Plan and define requirements at the overall corporate level
- Create a surrounding architecture for a complete warehouse
- Conform and standardize the data content
- Implement the data warehouse as a series of supermarts, one at a time

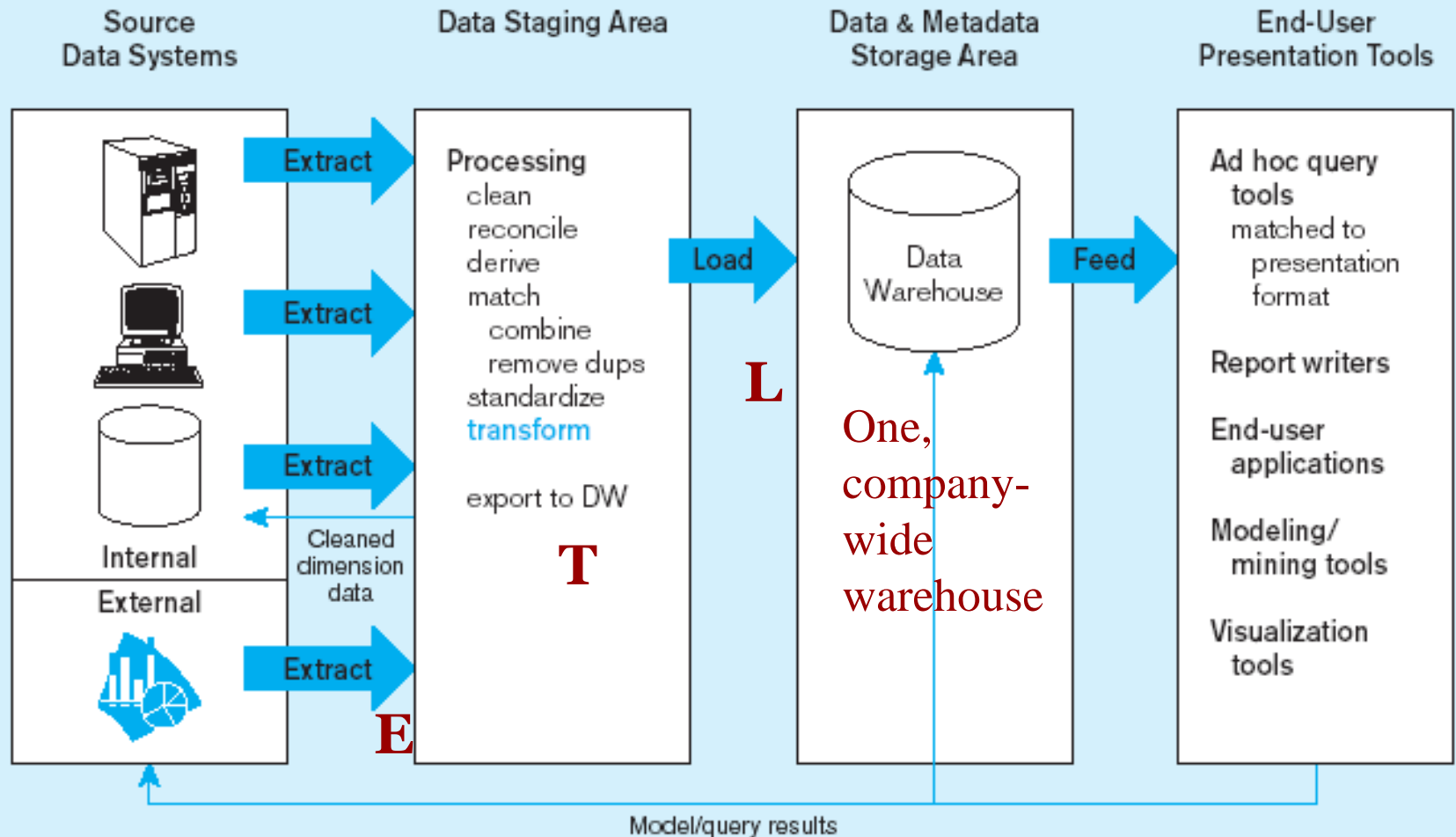
UNDERSTANDING DATA WAREHOUSE ARCHITECTURE

The **structure** that brings all the **components** of a data warehouse together is known as the **architecture**.

- **Generic Two-Level Architecture**
- **Independent Data Mart**
- **Dependent Data Mart and Operational Data Store**
- **Logical Data Mart and Real-Time Data Warehouse**
- **Three-Layer (Multi-Layer) architecture**

All involve some form of **extraction**, **transformation** and **loading** (ETL)

Generic two-level data warehousing architecture

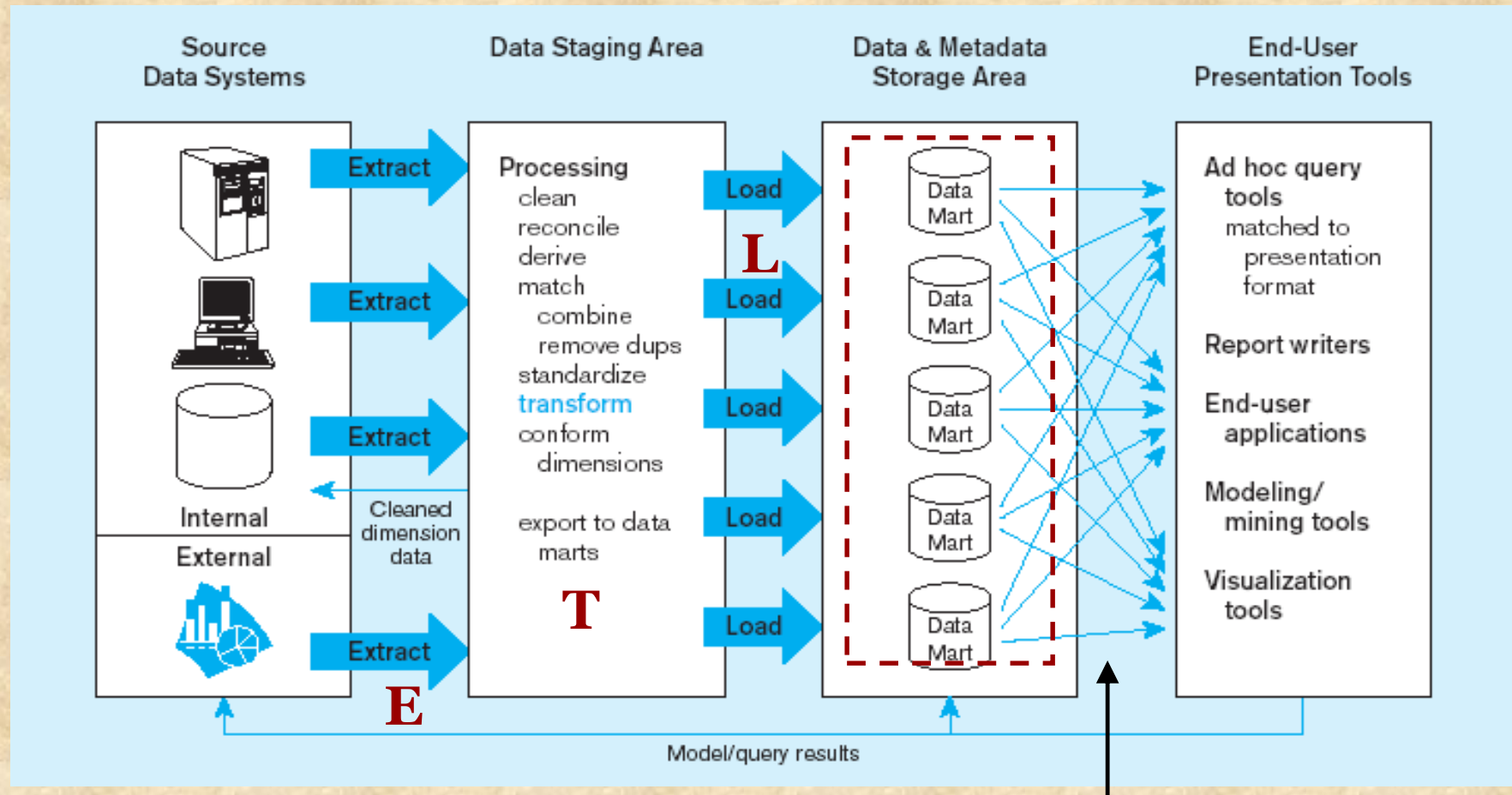


Periodic extraction → data is not completely current in warehouse

Independent data mart data warehousing architecture

Data marts:

Mini-warehouses, limited in scope

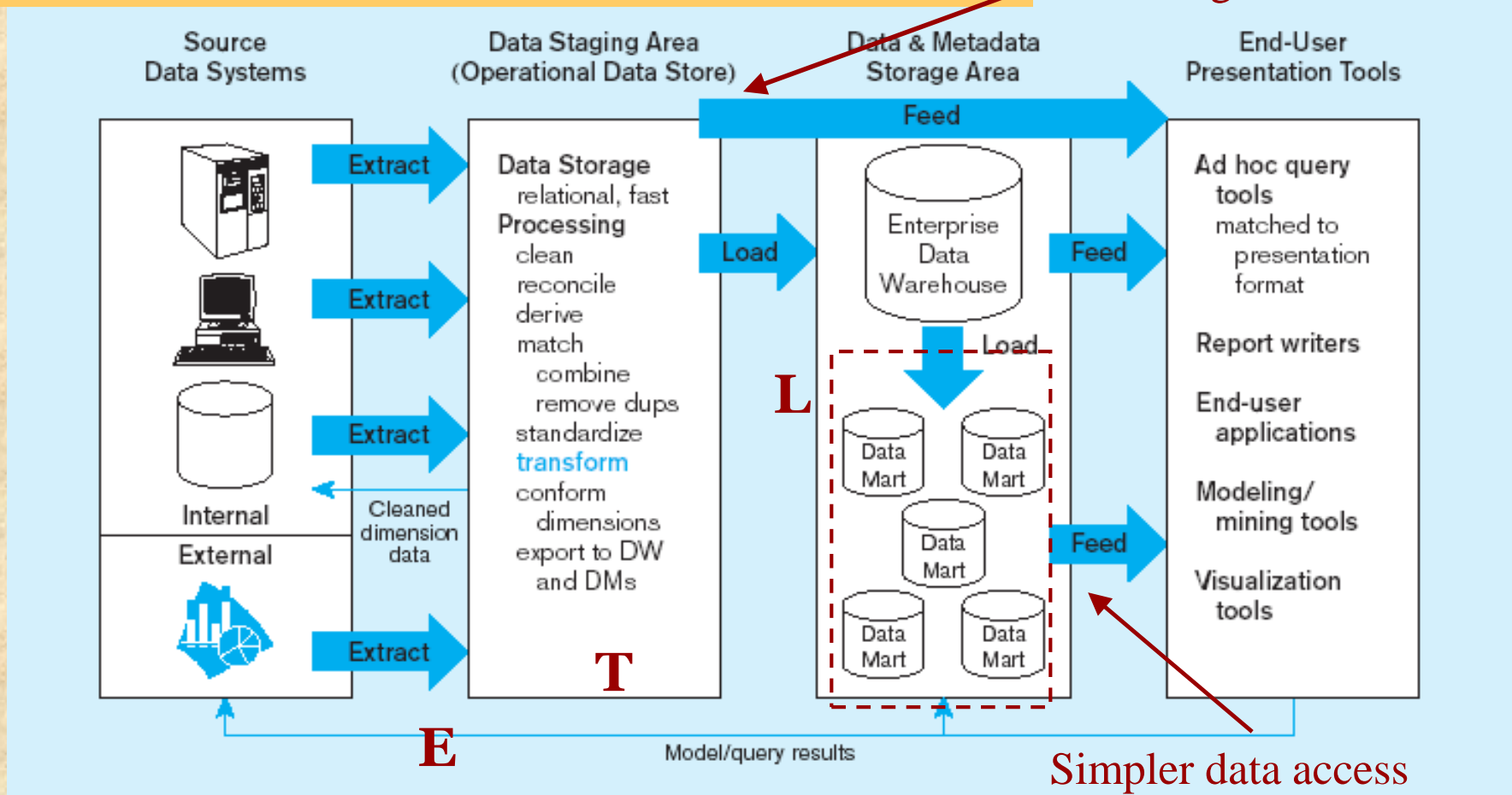


Separate ETL for each *independent* data mart

Data access complexity due to *multiple* data marts

Dependent data mart with operational data store: a three-level architecture

ODS provides option for obtaining *current* data

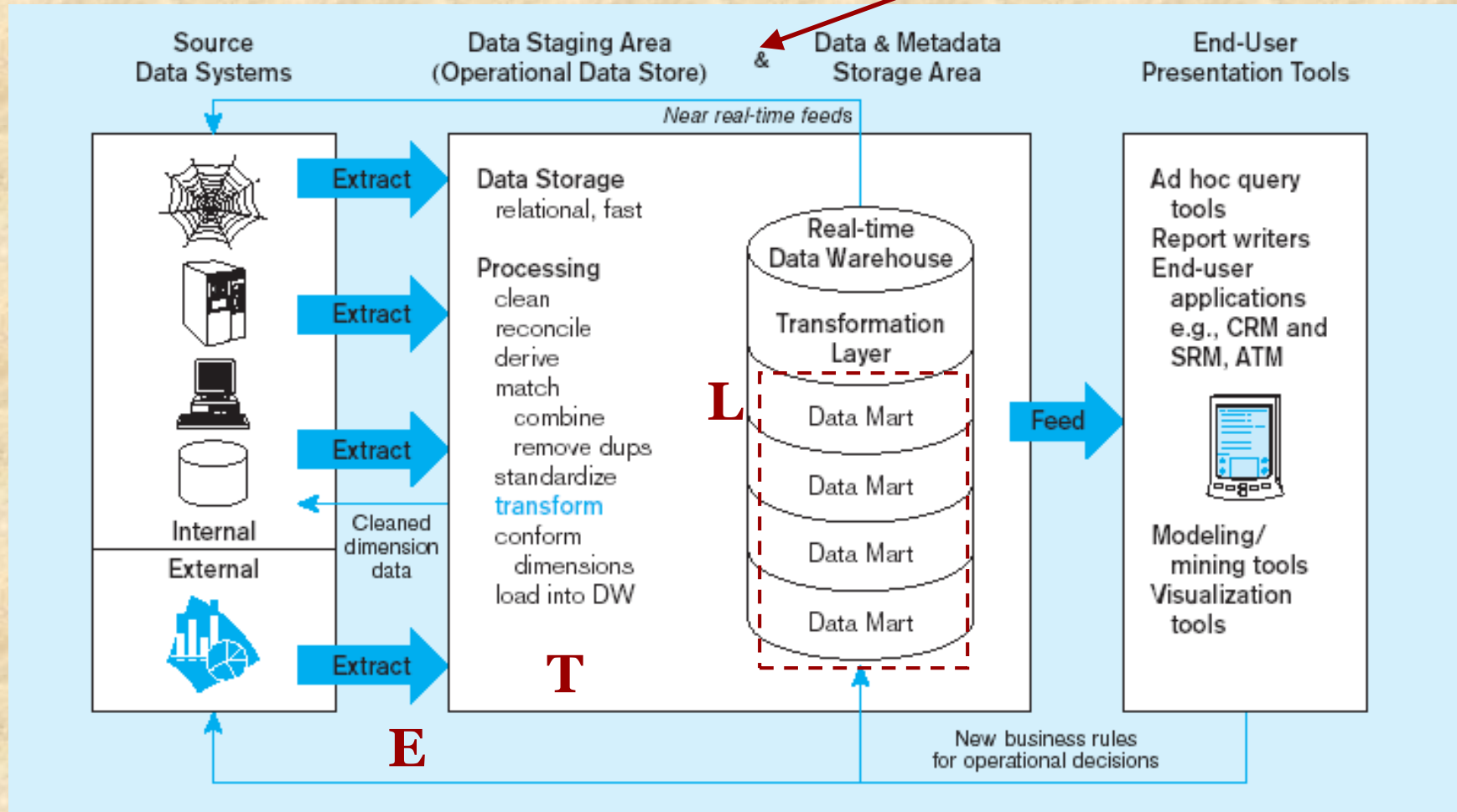


Single ETL for
enterprise data warehouse
(EDW)

Dependent data marts
loaded from EDW

Logical data mart and real time warehouse architecture

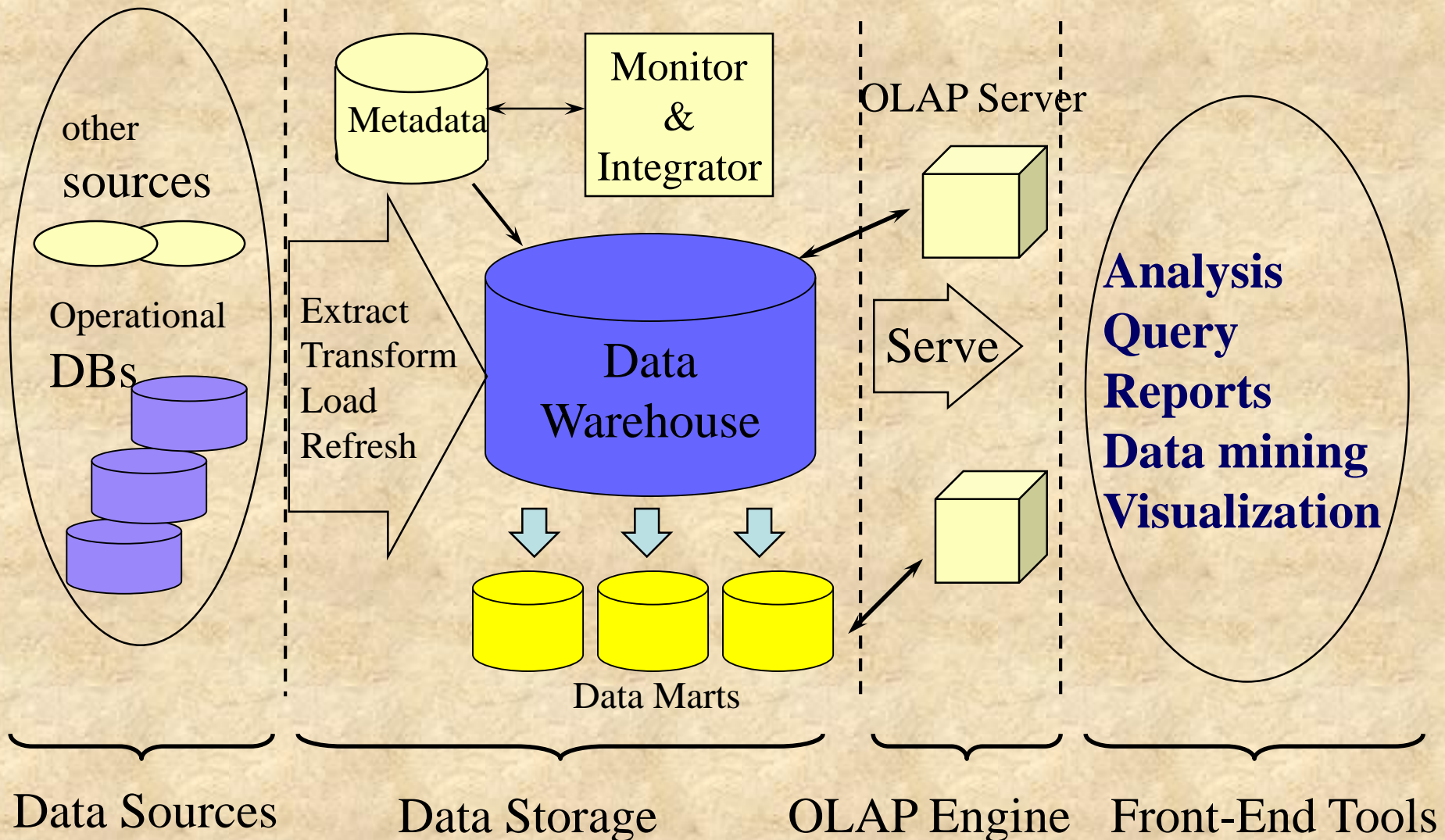
ODS and **data warehouse** are one and the same



Near real-time ETL for
Data Warehouse

Data marts are NOT separate databases,
but logical *views* of the data warehouse
➔ Easier to create new data marts

Multi-Tiered Architecture



Distinguishing Characteristics

- Different Objectives and Scope
- Data Content
- Complex Analysis and Quick Response
- Flexible and Dynamic
- Metadata-Driven

METADATA IN THE DATA WAREHOUSE

- **Operational metadata**
 - information about the operational data sources
 - Different data structures – field length , data types
- **Extraction and transformation metadata**
 - data about the extraction of data from the sources
 - the extraction frequencies, extraction methods, and business rules for the data extraction
- **End-user metadata**
 - navigational map of the data warehouse
 - end-users use their own business terminology to extract data from data warehouse

Exercise 1.2

- You are the data analyst on the project team, building a data warehouse for an insurance company.
- List the possible data sources from which you will bring the data into your data warehouse. State your assumptions.