DATA WAREHOUSING

ENGR. MADEHA MUSHTAQ
DEPARTMENT OF COMPUTER SCIENCE
IQRA NATIONAL UNIVERSITY

DATA WAREHOUSE

- A data warehouse (DW or DWH), also known as an enterprise data warehouse (EDW), is a system used for reporting and data analysis.
- DWH is considered a core component of business intelligence.
- DWHs are central repositories of integrated data from one or more disparate sources.

HISTORICAL OVERVIEW

- 1960

Master Files & Reports

- 1965

Lots of Master files!

- 1970

Direct Access Memory & DBMS

- 1975

Online high performance transaction processing

HISTORICAL OVERVIEW

- 1980

PCs and 4GL Technology (MIS/DSS)

- 1985 & 1990

Extract programs, extract processing,

The legacy system’s web.

WHY DATA WAREHOUSE?

- In the 1990s, businesses grew more complex, corporations spread globally, and competition became fiercer.
- The operational computer systems did provide information to run the day-to-day operations, but what the executives needed were different kinds of information that could be readily used to make strategic decisions.
- The operational systems, important as they were, could not provide strategic information.

WHY DATA WAREHOUSE?

- Businesses, therefore, were compelled to turn to new ways of getting strategic information.
- Data warehousing is a new paradigm specifically intended to provide vital strategic information.
- In the 1990s, organizations began to achieve competitive advantage by building data warehouse systems.
- Figure shows a sample of strategic areas where data warehousing is already producing results in different industries.

ORGANIZATION'S USE OF DATA WAREHOUSING

- 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

ESCALATING NEED FOR STRATEGIC INFORMATION

- The executives and managers who are responsible for keeping the enterprise competitive need strategic information to make proper decisions.
- Strategic information is information to formulate the business strategies, establish goals, set objectives, and monitor results.
- Some examples of business objectives are:
  - Retain the present customer base
  - Increase the customer base by 15% over the next 5 years
  - Improve product quality levels in the top five product groups
  - Bring three new products to market in 2 years
  - Increase sales by 15% in the North East Division

ESCALATING NEED FOR STRATEGIC INFORMATION

- For making decisions about these business objectives, executives and managers need in-depth knowledge of their company’s operations, learn about the key business factors and how these affect one another.
- Strategic information is not intended to produce an invoice, make a shipment, settle a claim, or post a withdrawal from a bank account.
- Strategic information is far more important for the continued health and survival of the corporation. Critical business decisions depend on the availability of proper strategic information in an enterprise.

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.

INFORMATION CRISIS

- We are faced with two startling facts:
  - Organizations have lots of data;
  - Information technology resources and systems are not effective at turning all that data into useful strategic information.
- Most companies are faced with an information crisis not because of lack of sufficient data, but because the available data is not readily usable for strategic decision making.
- The reason is the data of an enterprise is spread across many types of incompatible structures and systems.

TECHNOLOGY TRENDS

- Size of Data Sets are going up ↑.
- Cost of data storage is coming down ↓.
  - The amount of data average business collects and stores is doubling every year.
  - Total hardware and software cost to store and manage 1 Mbyte of data
    - 1990: ~ $15
    - 2002: ~ ¢15 (Down 100 times)
    - By 2007: < ¢1 (Down 150 times)

TECHNOLOGY TRENDS

- But size is not everything!
- Businesses demand Intelligence (BI).
  - Complex questions from integrated data.
  - “Intelligent Enterprise” or “Strategic Information”.

DBMS APPROACH

- List of all items that were sold last month?
- List of all items purchased by Tariq Majeed?
- The total sales of the last month grouped by branch?
- How many sales transactions occurred during the month of January?

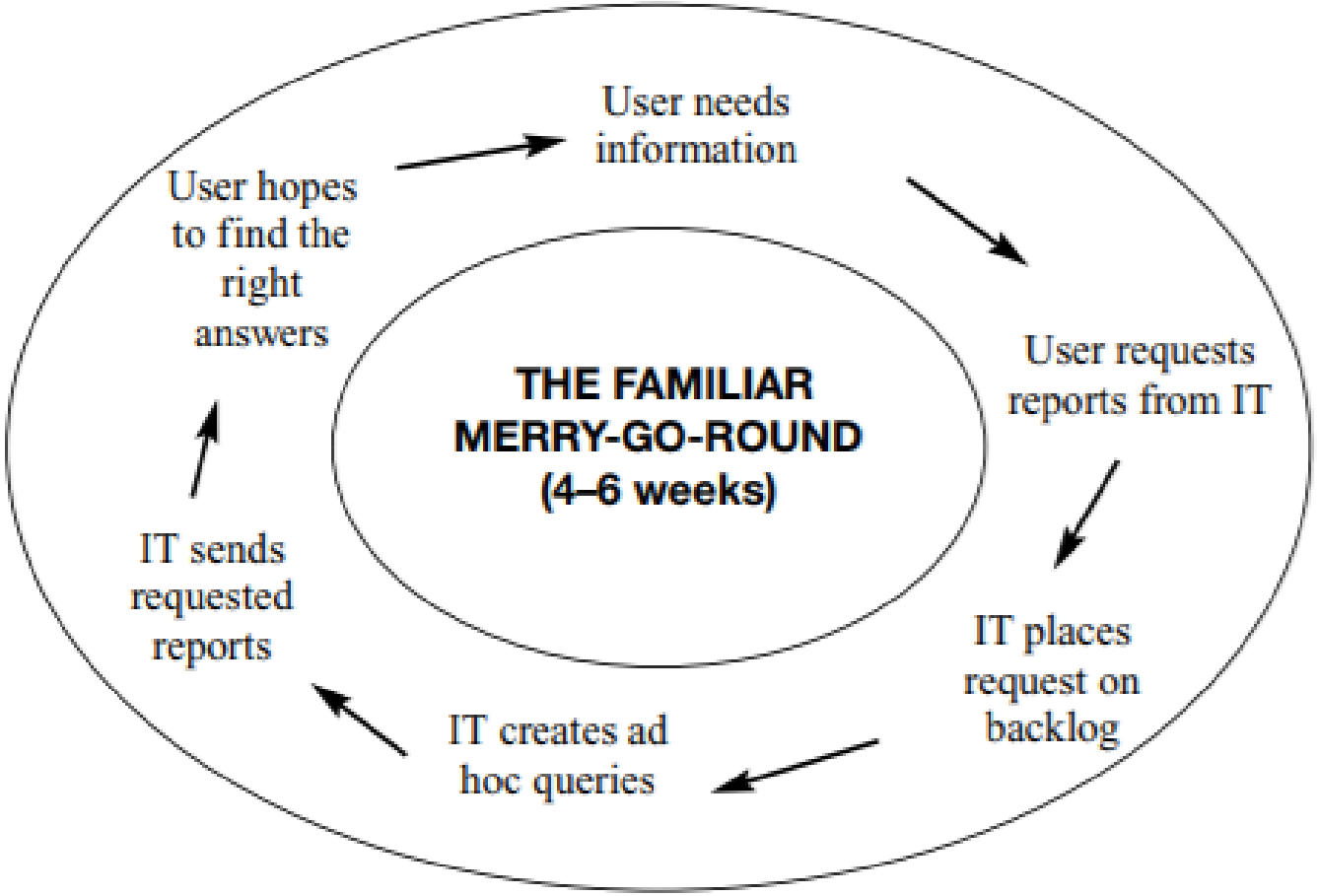
STRATEGIC INFORMATION

- Which items sell together? Which items to stock?
- Where and how to place the items? What discounts to offer?
- How best to target customers to increase sales at a branch?
- Which customers are most likely to respond to my next promotional campaign, and why?

OPERATIONAL VERSUS DECISION-SUPPORT SYSTEMS

- All the past attempts by IT to provide strategic information have been failures.
- This was mainly because IT has been trying to provide strategic information from operational systems.
- These operational systems such as order processing, inventory control, claims processing, outpatient billing, and so on are not designed or intended to provide strategic information.
- Only specially designed decision support systems or informational systems can provide strategic information.

OPERATIONAL VERSUS DECISION-SUPPORT SYSTEMS



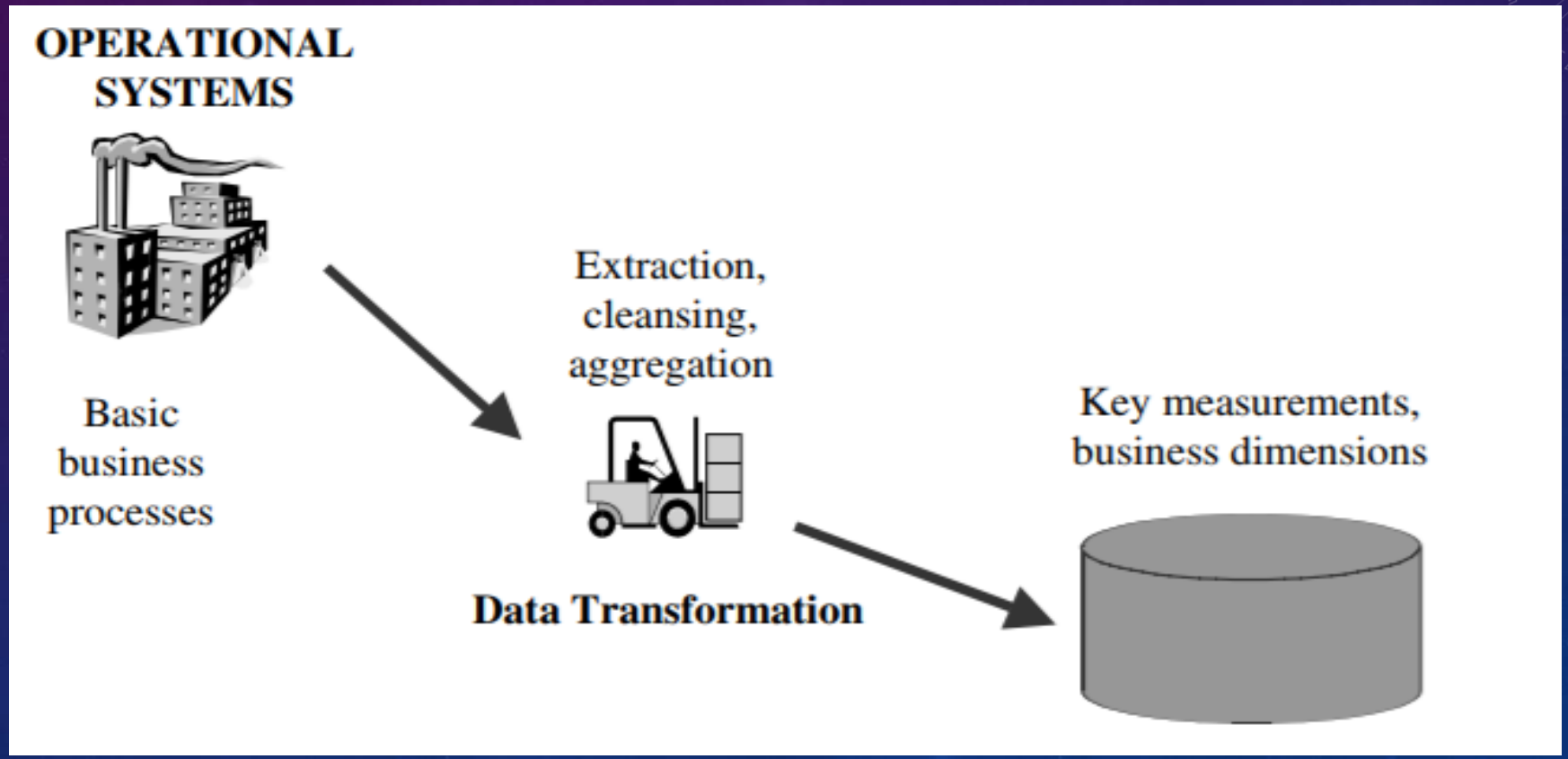
Inadequate attempts by IT to provide strategic information

OPERATIONAL VERSUS DECISION-SUPPORT SYSTEMS

Figure summarizes the differences between the traditional operational systems and the newer informational/Decision Support systems that need to be built.

	OPERATIONAL	INFORMATIONAL
Data Content	Current values	Archived, derived, summarized
Data Structure	Optimized for transactions	Optimized for complex queries
Access Frequency	High	Medium to low
Access Type	Read, update, delete	Read
Usage	Predictable, repetitive	Ad hoc, random, heuristic
Response Time	Sub-seconds	Several seconds to minutes
Users	Large number	Relatively small number

BUSINESS INTELLIGENCE AT THE DATA WAREHOUSE



DATA WAREHOUSE DEFINED

- The Data warehouse is thus the decision support system that can provide strategic information for analysis, discerning trends, and monitoring performance. The DWH

  - Provides an integrated and total view of the enterprise.
  - Makes the enterprise’s current and historical information easily available for decision making.
  - Renders the organization’s information consistent.
  - Presents a flexible and interactive source of strategic information.

PROCESSING REQUIREMENTS IN DATA WAREHOUSE

Most of the processing in DWH for strategic information will have to be analytical. There are four levels of analytical processing requirements:

- Running of simple queries and reports against current and historical data
- Ability to perform “what if” analysis is many different ways 3.
- Ability to query, step back, analyze, and then continue the process to any desired length.
- Spot historical trends and apply them for future results.

A BLEND OF MANY TECHNOLOGIES

Let us re-examine the basic concept of data warehousing. The basic concept of data warehousing is:

- Take all the data from the operational systems
- Where necessary, include relevant data from outside, such as industry benchmark indicators
- Integrate all the data from the various sources
- Remove inconsistencies and transform the data
- Store the data in formats suitable for easy access for decision making
- Although a simple concept, it involves different functions: data extraction, the function of loading the data, transforming the data, storing the data, and providing user interfaces.

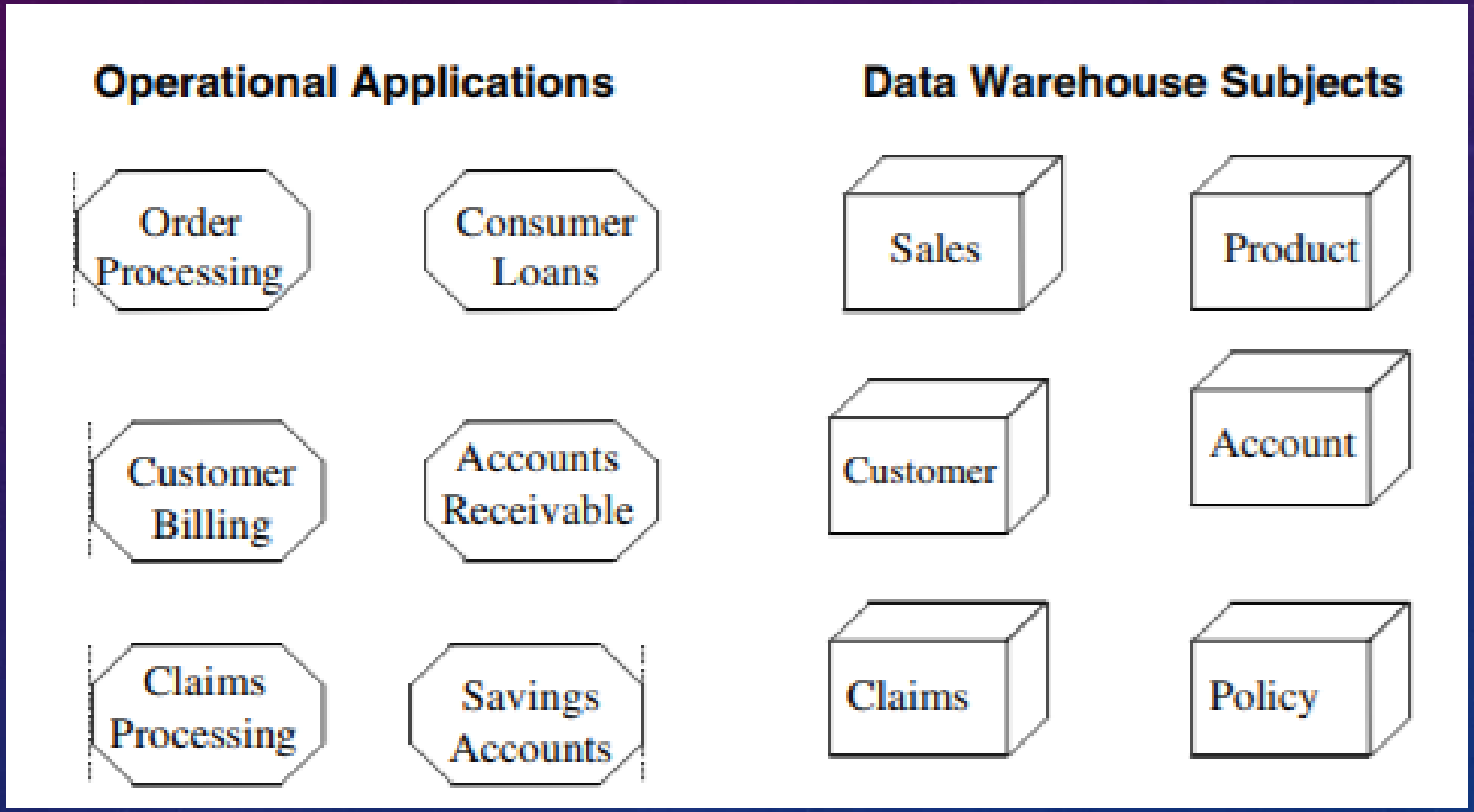
DEFINING FEATURES OF DATA WAREHOUSE

- Subject-Oriented Data
- Integrated Data
- Time-variant Data
- Nonvolatile Data
- Data Granularity

SUBJECT-ORIENTED DATA

- In operational systems, data is stored by individual applications.
- In the data warehouse, data is stored by business subjects, not by applications.
- Business subjects differ from enterprise to enterprise.
- These are the subjects critical for the enterprise.
- For a manufacturing company, sales, shipments, and inventory are critical business subjects.
- For a retail store, sales at the check-out counter is a critical subject.
- Figure distinguishes between how data is stored in operational systems and in the data warehouse.

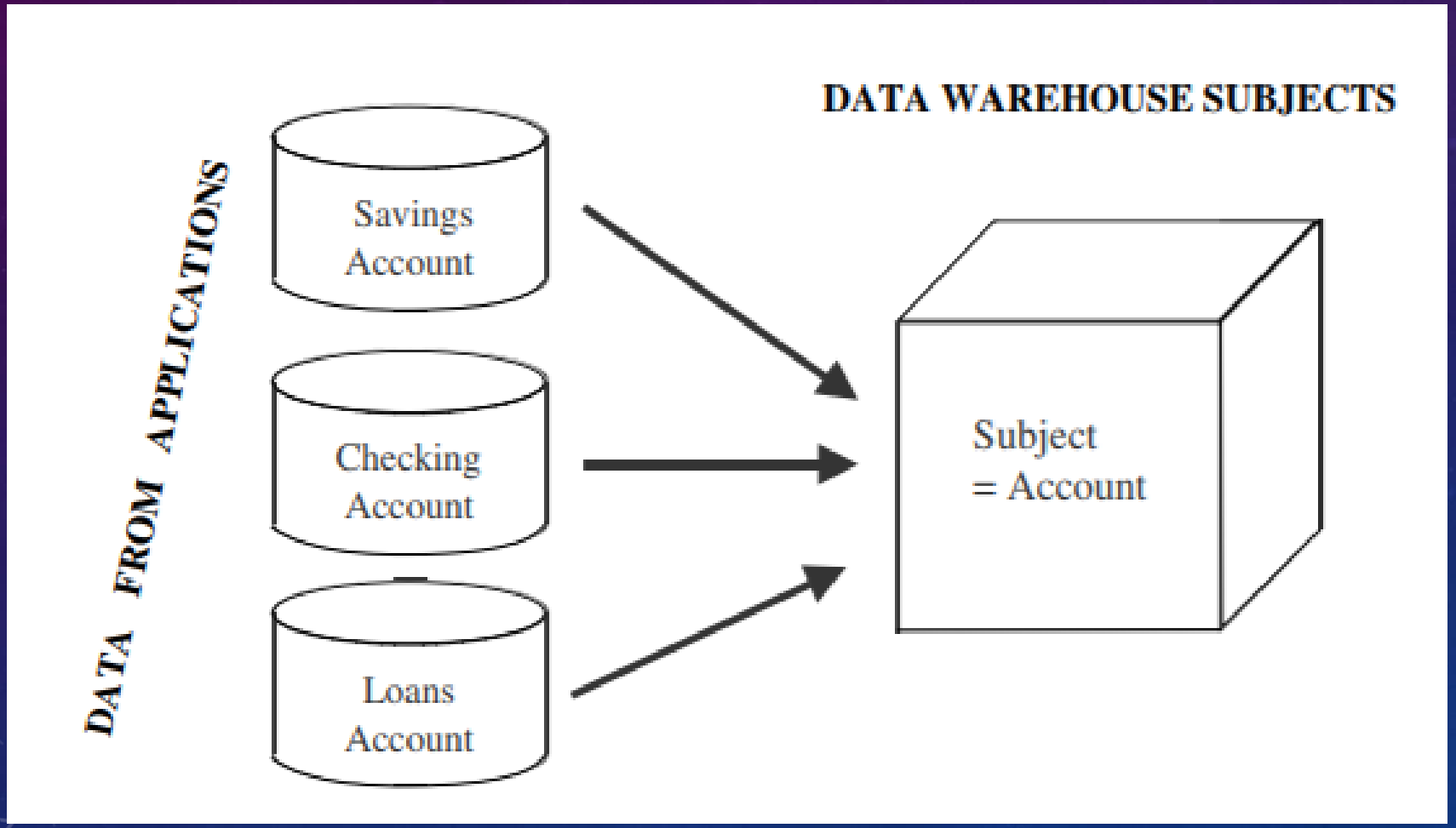
SUBJECT-ORIENTED DATA



INTEGRATED DATA

- The data in the data warehouse comes from several operational systems.
- Source data are in different databases, so the file layouts, character code representations could be different, naming conventions could be different, attributes for data items could be different.
- Before the data from various disparate sources can be usefully stored in a data warehouse, data inconsistencies are removed; data from diverse operational applications is integrated.

INTEGRATED DATA



TIME-VARIANT DATA

- For an operational system, the stored data contains the current values.
- we might store some past transactions in operational systems, but, essentially, operational systems reflect current information because these systems support day-to-day current operations.
- A data warehouse, because of the very nature of its purpose, has to contain historical data, not just current values.
- For example, if a user wants to find out the reason for the drop in sales in the North East division, the user needs all the sales data for that division over a period extending back in time.

TIME-VARIANT DATA

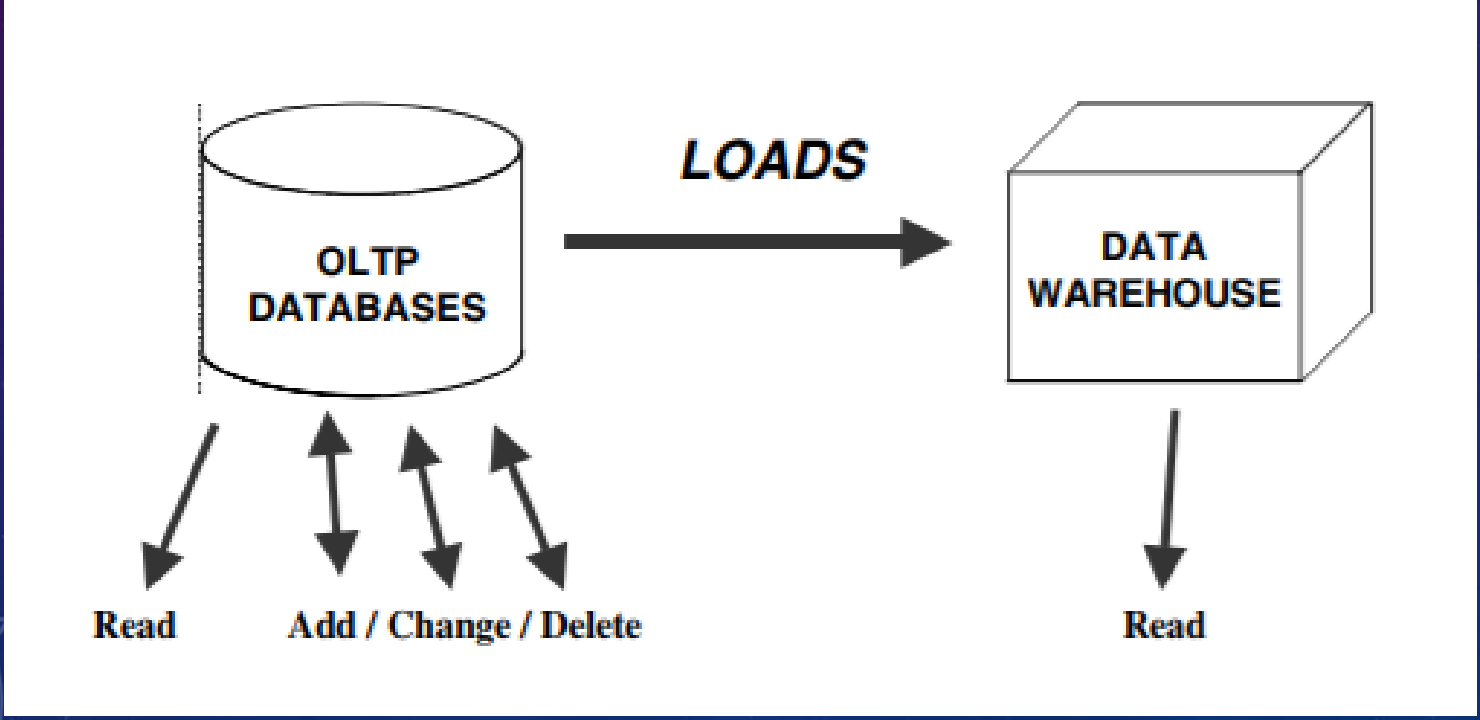
- When an analyst in a grocery chain wants to promote two or more products together, that analyst wants sales of the selected products over a number of past quarters.
- The time-variant nature of the data in a data warehouse:
  - Allows for analysis of the past
  - Relates information to the present
  - Enables forecasts for the future

NONVOLATILE DATA

- Data extracted from the various operational systems and pertinent data obtained from outside sources are transformed, integrated, and stored in the data warehouse.
- The data in the data warehouse is not intended to run the day-to-day business.
- Data from the operational systems are moved into the data warehouse at specific intervals. We do not update the data warehouse every time we process a single order.
- Depending on the requirements of the business, these data movements take place twice a day, once a day, once a week, or once in two weeks.

NONVOLATILE DATA

As illustrated in Figure, every business transaction does not update the data in the data warehouse. The business transactions update the operational system databases in real time.



DATA GRANULARITY

- In an operational system, data is usually kept at the lowest level of detail.
- We do not usually keep summary data in an operational system.
- Frequently, the analysis begins at a high level and moves down to lower levels of detail.
- Data granularity refers to the level of detail. Depending on the requirements, multiple levels of detail may be present. Many data warehouses have at least dual levels of granularity.

DATA GRANULARITY

THREE DATA LEVELS IN A BANKING DATA WAREHOUSE

Daily Detail	Monthly Summary	Quarterly Summary
Account	Account	Account
Activity Date	Month	Month
Amount	Number of transactions	Number of transactions
Deposit/Withdrawal	Withdrawals	Withdrawals
	Deposits	Deposits
	Beginning Balance	Beginning Balance
	Ending Balance	Ending Balance

THREE DATA LEVELS IN A BANKING DATA WAREHOUSE

REFERENCE BOOKS

- *Data Warehousing Fundamentals, 2nd Edition*, Paulraj Ponniah, 2010, John Wiley & Sons Inc., NY.
- *Building the Data Warehouse*, 4th Edition, W. H. Inmon, 2005, John Wiley & Sons Inc., NY.

END OF SLIDES