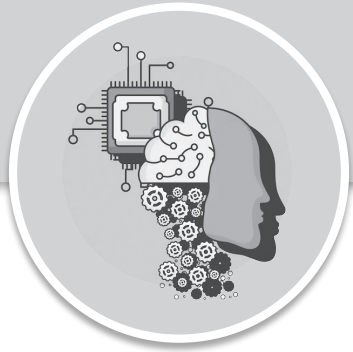
DATA SCIENCE & ARTIFICIAL INTELLIGENCE

Data Warehouse



Comprehensive Theory
with Solved Examples *and* Practice Questions

MADE EASY PUBLICATIONS

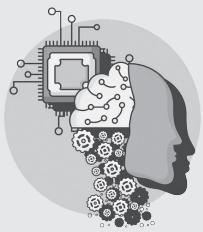
www.madeeasypublications.org

MADE EASY Publications Pvt. Ltd.

Corporate Office: 44-A/4, Kalu Sarai (Near Hauz Khas Metro Station), New Delhi-110016 | **Ph. :** 9021300500

Email : infomep@madeeasy.in | **Web :** www.madeeasypublications.org

Data Warehouse



EDITIONS

First Edition : 2025

Second Edition : 2026

CONTENTS

Data Warehouse

CHAPTER

1

Introduction and Datatypes

1.1 DEFINITION OF DATA

"A collection of facts, usually obtained as the result of experiments, observations, transactions, or processes. Data can exist in various forms, such as numbers, text, images, and signals".

1.2 TYPES OF DATA

1.2.1 Raw Data (Operational Data)

- Data collected directly from operational systems (e.g., transactional databases, ERP, CRM).
- *Example:* Sales transactions, customer orders, inventory records.
- Characteristics:
  - Unprocessed and not yet transformed for analysis.
  - Highly detailed with redundant information.

1.2.2 Processed Data (Transformed Data)

- Data that has undergone cleaning, transformation, and integration before storage.
- *Example:* Data aggregated by product category, customer segment, or region.
- Characteristics:
  - Standardized and formatted for analytical processing.
  - Supports business intelligence (BI) applications.

1.2.3 Summary Data (Aggregated Data)

- Data stored at a higher level of granularity for fast reporting and analysis.
- *Example*: Daily, weekly, or monthly sales summaries rather than individual transactions.
- Characteristics:
  - Faster query performance for decision-making.
  - Often used in OLAP (Online Analytical Processing).

1.2.4 Metadata (Data about Data)

- Information that describes the structure, source, and meaning of data in the warehouse.
- *Example*: Data definitions, column descriptions, data lineage.
- Characteristics:
  - Helps in data governance and understanding relationships.
  - Supports ETL (Extract, Transform, Load) processes.

1.2.5 Historical Data

- Data collected over a long period to support trend analysis and forecasting.
- *Example:* Five years of customer purchase history.
- Characteristics:
  - Used for predictive analytics and data mining.
  - Can be stored in archival systems or historical tables.

1.2.6 Real-Time Data (Streaming Data)

- Data that is updated in near real-time for dynamic reporting.
- *Example:* Live stock market data, IoT sensor data.
- Characteristics:
  - Requires high-performance processing.
  - Used in real-time dashboards and alert systems.

Example 1.1 Sequence data refers to:

(a) Data without order
(b) Ordered data where sequence matters
(c) Data with geographical attributes
(d) Data stored in relational tables

Solution: (b)
Sequence data, such as stock market trends and DNA sequences, depends on the order of events.

Example 1.2 Which of the following is an example of graph data?

(a) Social network connections
(b) Monthly sales reports
(c) Weather forecasts
(d) Product inventories

Solution: (a)
Facebook friendships and LinkedIn connections are classic graph-based datasets.

1.3 DEFINITION OF DATA MINING

The process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the internet, and other information repositories.

Data Mining Process

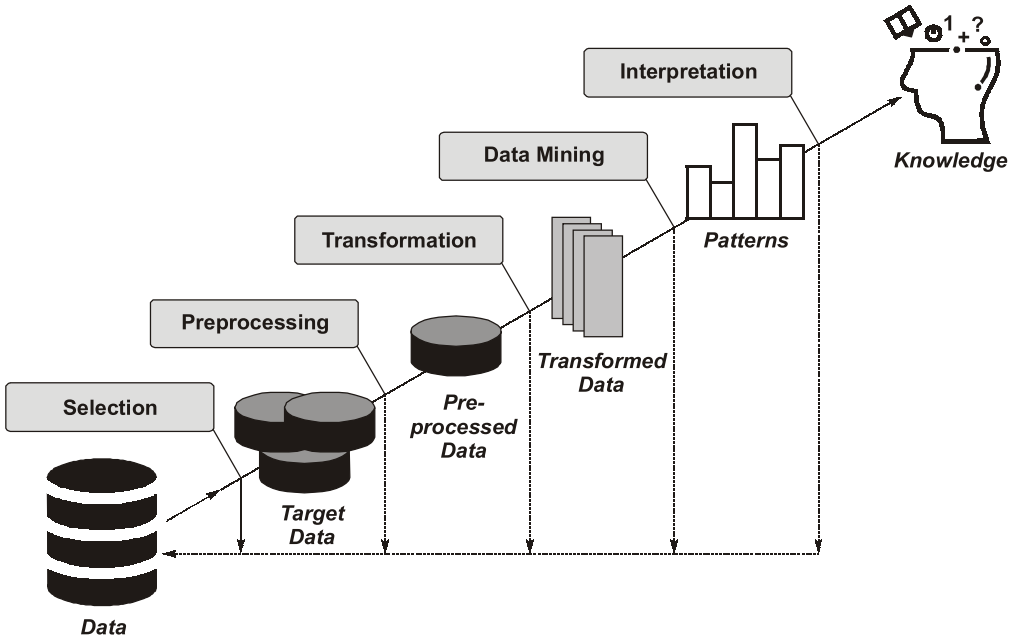


Figure: *Data Mining Process*

1.3.1 Why Need of Data Mining?

The growth of data in various fields has made it essential to develop techniques for analyzing large and complex datasets. Traditional data analysis methods are no longer sufficient due to the size, complexity, and dynamic nature of data. Data mining helps in discovering useful patterns, relationships, and knowledge from large volumes of data.

- **Explosive Growth of Data:** Organizations generate massive data daily.
- **Hidden Patterns and Knowledge Discovery:** Traditional methods can't uncover complex relationships.
- **Decision-Making Support:** Helps businesses and researchers make informed decisions.
- **Automation and Efficiency:** Extracts insights without manual analysis.
- **Applications in Various Domains:** Used in healthcare, finance, marketing, cybersecurity, etc.

1.3.2 Evolution of Data Mining

Data mining has evolved over several decades, driven by advances in databases, machine learning, and computing power. The following stages outline its historical development:

1. **Data Collection and Database Creation (1960s-1980s)**
   (i) Early data storage focused on file systems and traditional databases.
   (ii) Hierarchical and network database models were used for managing structured data.
   (iii) Relational Database Management Systems (RDBMS) emerged, enabling efficient data storage and retrieval using SQL.
2. **Data Management and Query Processing (1980s-1990s):**
   (i) Growth of OLAP (Online Analytical Processing) for multi-dimensional analysis.
   (ii) Development of query optimization techniques for better database performance.
   (iii) Data warehouses became popular for integrating and storing large volumes of structured data.

Student's Assignments

Q.1 What is the primary purpose of a data warehouse?
(a) Storing real-time transactional data
(b) Integrating and analyzing historical data
(c) Managing hardware configurations
(d) None of these

Q.2 Which of the following is not a key feature of a data warehouse?
(a) Stores historical and summarized data
(b) Organizes data by major subjects
(c) Supports frequent updates like transactional databases
(d) Uses a multidimensional model for querying

Q.3 Which process is not involved in data warehouse construction?
(a) Data cleaning
(b) Data transformation
(c) Data fragmentation
(d) Data loading

Q.4 A transactional database is best suited for:
(a) Historical data analysis
(b) Real-time transaction processing
(c) Data visualization
(d) Dimensional modeling

Q.5 In transactional databases, which of the following best describes a market basket analysis?
(a) Analyzing the purchase behavior of customers
(b) Storing customer details
(c) Storing product descriptions
(d) None of these

Q.6 Which of the following is an example of sequence data?
(a) Customer purchase records
(b) DNA sequences
(c) Employee salary details
(d) Product catalogs

Q.7 Which data type ensures globally unique identification?
(a) Enumerated (b) UUID
(c) Floating-point (d) Date-Time

Q.8 Which of the following best describes an advantage of using the UUID data type?
(a) It helps in defining fixed sets of values
(b) It is used for precise arithmetic calculations
(c) It reduces data redundancy
(d) It ensures unique identifiers across distributed databases

Q.9 What is the main advantage of using a floating-point data type in a data warehouse?
(a) It provides exact decimal values
(b) It allows storage of large numbers with precision
(c) It is suitable for categorical attributes
(d) It speeds up text-based queries

Q.10 What is the main difference between OLTP and OLAP systems?
(a) OLTP focuses on analytical processing, while OLAP focuses on transactions
(b) Both are used for real-time processing
(c) OLAP supports complex queries for decision-making, while OLTP handles daily operations
(d) None of these

Q.11 In a data cube, what does each dimension represent?
(a) A single transaction
(b) A set of attributes for analysis
(c) A randomly chosen dataset
(d) A specific row in a database

Q.12 Which is a common challenge in data warehouse design?
(a) Data redundancy
(b) Data cleaning and integration
(c) Lack of transactional support
(d) None of these

Q.13 What is the primary reason for summarizing data in a data warehouse?
(a) To reduce storage space
(b) To improve query performance
(c) Both (a) and (b)
(d) To eliminate irrelevant data