Chapter 6

Data warehousing and data mining techniques

Chapter outline

- Introduction to Data warehousing
- Why Data warehousing
- Data warehousing and online transaction processing
- Introduction to data mining
- Data mining techniques

Introduction to Data warehousing

- A data warehouse as a storehouse, is a **repository of data** collected from multiple data sources (often heterogeneous) and is intended to be used as a whole under the same unified schema.
- **Data Warehousing** is the process of **collecting, storing, and managing large volumes of historical data** from multiple sources to support **analysis, reporting, and decision making**.
- It gives the option to analyze data from different sources under the same roof.
- Data warehousing involves data cleaning, data integration, and data consolidations.
- Generally A data warehouse is a **centralized repository** of integrated data from one or more different, separate sources.

Cont..

A **database** stores **current operational data** needed to run daily business activities.

Example:

University database storing:

- ✓ Student registration
- ✓ Course enrollment
- ✓ Exam results

✓ This data change frequently

A **data warehouse** stores **historical data** collected from multiple databases for **analysis and reporting**.

Example:

University data warehouse used to analyze:

- ✓ Student performance over 5 years
- ✓ Enrollment trends by department
- ✓ Graduation rates

✓ This data is mostly **read-only**.

How Data warehouse works?

- ✓ A Data Warehouse works as a central repository where information arrives from one or more data sources.
- ✓ Data flows into a data warehouse from the transactional system and other relational databases.
- ✓ With in each database data organized into row and columns.
- ✓ A data warehouse merges information coming from different sources into one comprehensive database tables can be organized inside schema which are think as folder.
- ✓ By merging all of this information in one place, an organization can analyze its customers more holistically. This helps to ensure that it has considered all the information available.

Benefits of Data Warehouse

- **Centralized Data Repository:-** Combines data from **multiple sources** into one place.
- To facilitate reporting as well as analysis
- Informed decision making:- Provides **accurate, integrated, and historical data**
- Historical Data Analysis :- Enables **time-based analysis** (monthly, yearly trends) Stores data over long periods (years).
- **Faster Query Performance:-** Optimized for complex analytical queries

Characteristics of Data warehouse

- ✓ **Subject-Oriented:-** Data warehouse contains data organized by topics. E.g. Sales, marketing, finance, etc.
- ✓ **Time variant:-** Data warehouse contains data that reflect what happened last week, last month, past five years, and so on. Time based reporting .
- ✓ **Integrated** :- centralized, consolidated database that integrates data derived from the entire organization.
- ✓ **Non volatile:-** Once data enter the data warehouse, they are never removed.

  Because the data in the warehouse represent the company‘s entire history.

Data warehousing and online transaction processing

- ✓ Two minor systems that are useful in managing this data include Data Warehousing (DWH) as well as Online Transaction Processing (OLTP).
- ✓ Data Warehousing is a technique that gathers or collects data from different sources into a central repository, or, in other words, a single, complete, and consistent store of data that is obtained from different sources.
- ✓ It is a powerful database model that enhances the user's ability to analyze huge, multidimensional datasets.

Online-Transaction Processing /OLTP/

- It is a technique used for detailed day-to-day transactions of data which continuously chain on an everyday-basis.
- We can describe OLTP **support daily operational activities** of an organization by processing a **large number of short, fast, and concurrent transactions** in real time.
- It is featured by a large number of short on-line transactions (INSERT, UPDATE, and DELETE).
- OLTP or Online Transaction Processing is a type of data processing that consists of executing a number of transactions occurring concurrently—online banking, shopping, order entry, or sending text messages, for example.
- The primary significance of OLTP operations is put on very rapid query processing, maintaining record integrity in multi-access environments, and effectiveness consistent by the number of transactions per second.

eg

Example 1: Banking System (ATM / Mobile Banking)

OLTP Activities

- ✓ Deposit money
- ✓ Withdraw money
- ✓ Transfer funds

Example 2: University Student Registration System

OLTP Activities

- ✓ Register new students
- ✓ Add or drop courses
- ✓ Enter grades
- ✓ Update student profiles

eg

Example 1: Bank Data Warehouse

Data Stored

- ✓ 10 years of transaction history
- ✓ Customer demographics
- ✓ Loan and credit data

Example 2: University Data Warehouse

Data Stored

- ✓ Student performance over many years
- ✓ Enrollment trends
- ✓ Graduation and dropout data

Feature	OLTP	Data Warehouse
Purpose	Daily operations	Analysis & decision making
Data type	Current, detailed	Historical, summarized
Transactions	Many, short	Few, complex
Operations	Insert, Update, Delete	Read-only (SELECT)
Design	Highly normalized	Denormalized
Users	Clerks, customers	Managers, analysts
Performance focus	Speed & consistency	Query efficiency
Time dimension	Current data	Time-variant
Example	ATM system	Business intelligence system

Introduction to data mining

- ✓ Data mining refers to extracting or mining knowledge from large amounts of data.
- ✓ It is a process of **extracting and discovering patterns in large data sets and** get patterns or knowledge from huge amount of data.
- ✓ The main goals of data mining is to **discover meaningful patterns, relationships, trends, and knowledge from large volumes of data**
- ✓ Data warehouse stores data → data mining analyzes data

Cont..

Data Mining Process (KDD Model)

1. Data Cleaning

- Remove **noise**, errors, and missing values
- Handle duplicate records and inconsistencies

Example:

Removing incomplete student records from a university database.

2. Data Integration

Combine data from **multiple sources** into one dataset

Example:

Merging student data from **admission**, **exam**, and **finance** systems.

3. Data Selection

Select **relevant data** for the mining task

Example:

Selecting only GPA, attendance, and exam scores to analyze student performance.

4. Data Transformation

Convert data into a **suitable format**

Normalize, aggregate, or encode data

Example:

Converting raw marks into grade categories (A, B, C).

5. Data Mining (Core Step)

Apply algorithms to extract patterns

6 Pattern Evaluation

Identify **interesting and useful patterns**

Remove irrelevant or redundant results

Example:

Keeping only rules with high confidence like

“Students with attendance > 80% usually pass.”

7. Knowledge Presentation

Present results in a **human-understandable form**

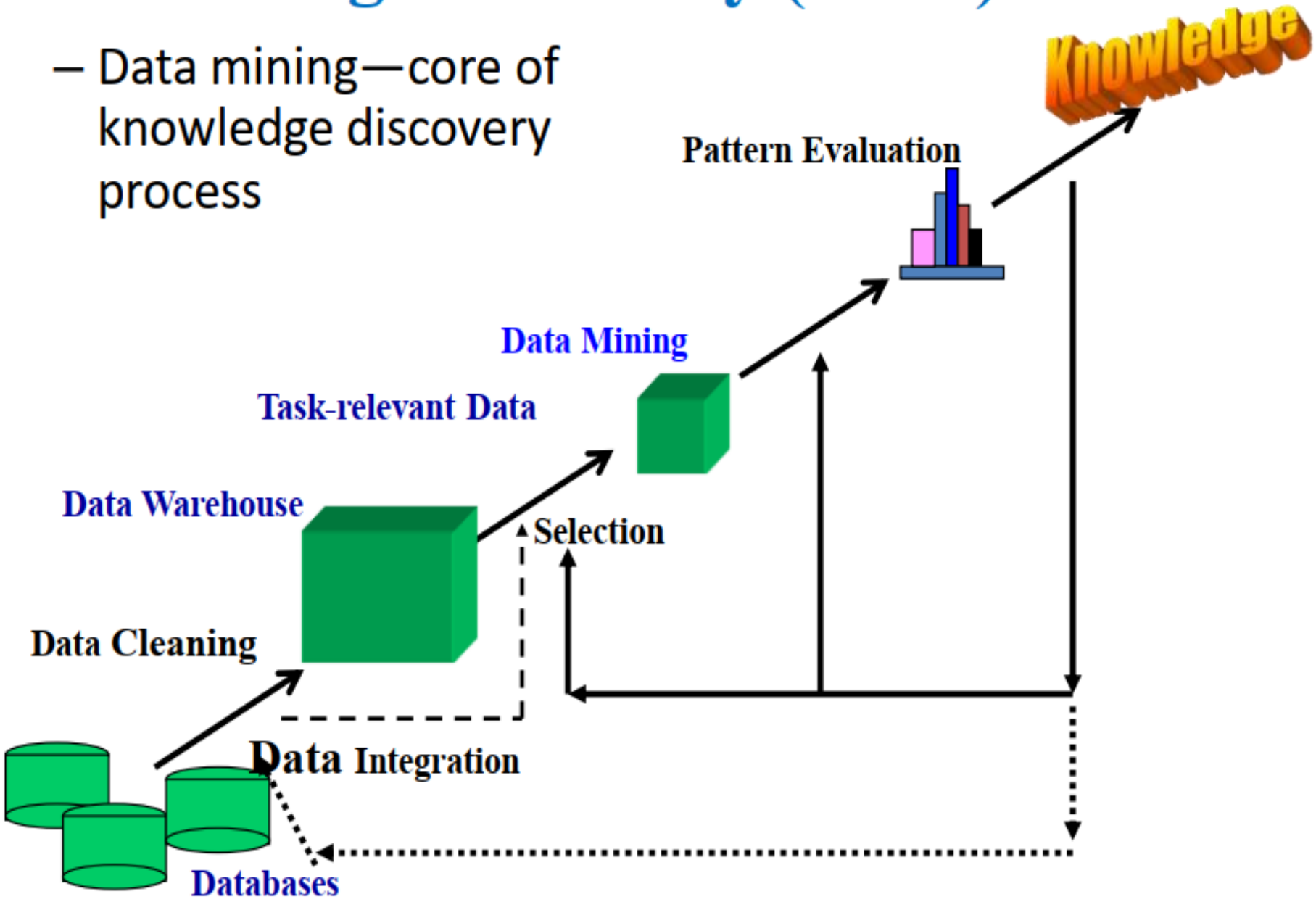
Use graphs, reports, dashboards

Example:

Charts showing student performance trends.

Knowledge Discovery (KDD) Process

– Data mining—core of knowledge discovery process



Data Mining Techniques

Classification

- ✓ Assigns data to predefined classes
- ✓ Uses labeled data (supervised learning)

Examples:

  - Pass / Fail students
  - Spam / Not Spam emails

- **Techniques:** Decision Trees, Naïve Bayes, Neural Networks

Clustering

- ✓ Groups similar data items
- ✓ No predefined classes (unsupervised learning)

Cont..

Examples:

Group customers by buying behavior

Group students by performance

Techniques: K-means, Hierarchical clustering

Association Rule Mining

Finds relationships between items

Uses **IF–THEN** rules

Example:

If a customer buys bread → they also buy butter

Algorithm: Apriori

Cont..

Regression

Predicts numerical values

Finds relationship between variables

Examples:

- Predict salary based on experience
- Predict sales amount
- Predicting student GPA

Types: Linear regression, Multiple regression

Anomaly (Outlier) Detection

Identifies unusual or abnormal data

Examples:

- Fraud detection
- Network intrusion detection

Cont..

Example (University)

- **Data Warehouse:**

  Stores student records, grades, attendance for many years

- **Data Mining:**

  - Predict students at risk of failure
  - Group students by performance
  - Find factors affecting GPA

Data Mining Applications :

Here is the list of areas where data mining is widely used

1. **Financial Data Analysis**
2. **Retail Industry**
3. **Telecommunication Industry**
4. **Biological Data Analysis**
5. **Other Scientific Applications**
6. **Intrusion Detection**

Conclusion

- Data mining is the task of discovering interesting patterns from large amounts of data, where the data can be stored in databases, data warehouses, or other information repositories.
- It is a young interdisciplinary field, drawing from areas such as database systems, data warehousing, statistics, machine learning, data visualization, information retrieval, and high-performance computing

?