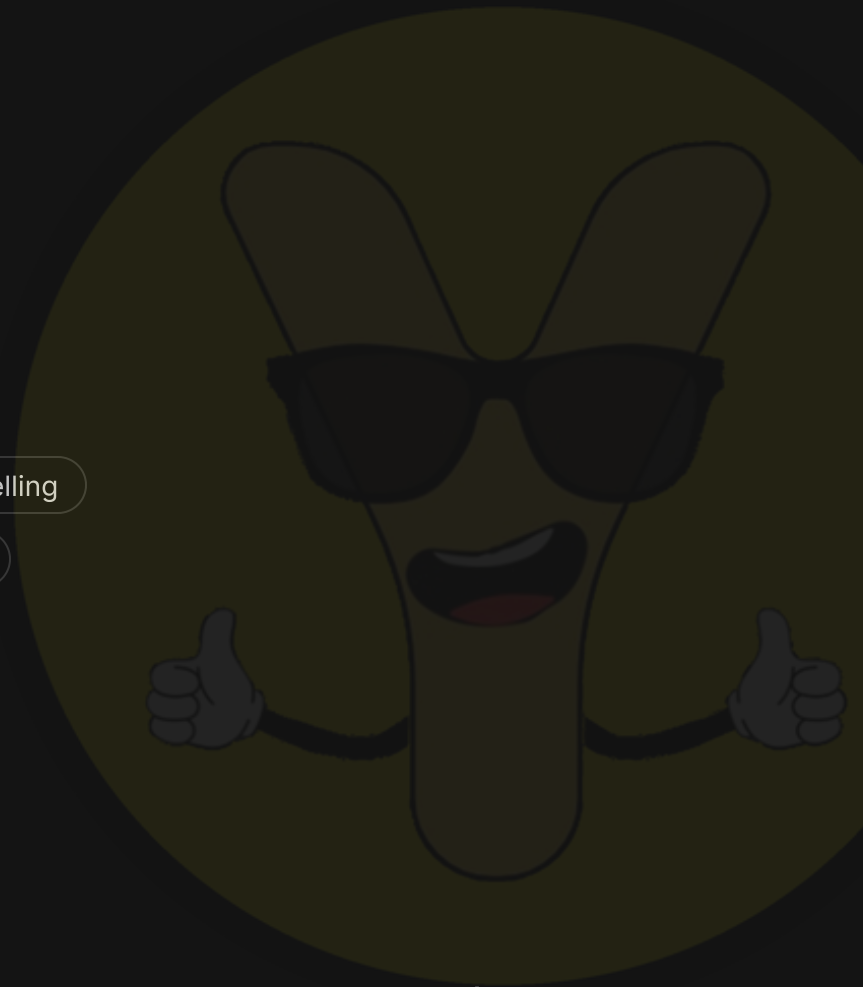
AKTU · 3RD YEAR · SEMESTER 5 · BCDS051

Business Intelligence and Analytics

Unit 2 — BI – Data Mining & Warehousing

- Introduction to Data Mining
- Architecture of Data Mining and How Data Mining Works (Process)
- Functionalities & Classifications of Data Mining
- Representation of Input Data
- Analysis Methodologies
- Introduction to Data Warehousing
- Data Mart
- Online Analytical Processing (OLAP) – Tools
- Data Modelling
- Difference between OLAP and OLTP
- Schema – Star and Snowflake Schemas
- ETL Process – Role of ETL



DECODED BY

Chirag Jain

Founder, Ycotes · Gold Medalist in Artificial Intelligence, IIT Madras · taught 10,000+ students

theycotes.com

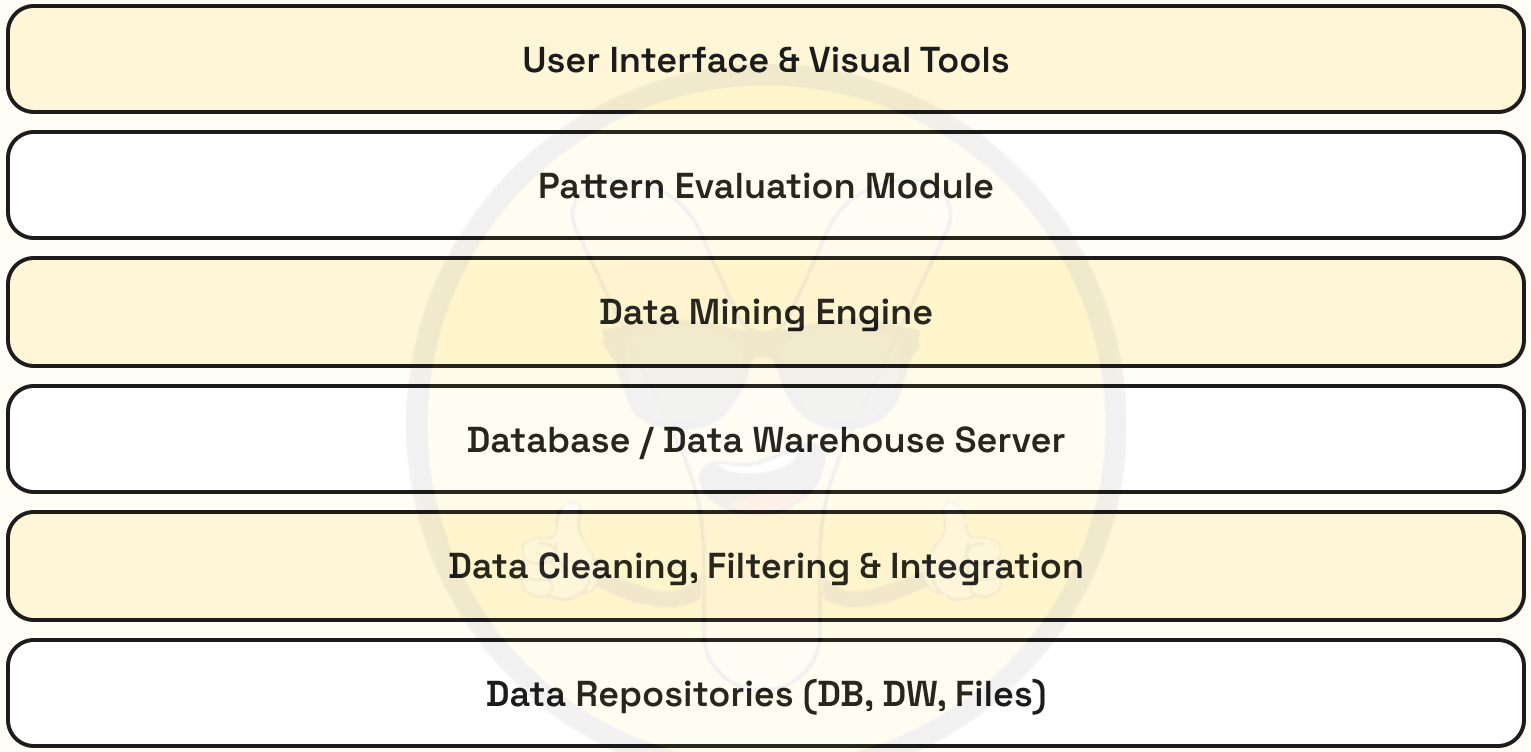
Introduction to Data Mining

DATA MINING

Data Mining is the process of discovering actionable patterns, anomalies, correlations, and knowledge from large repositories of data using statistical, mathematical, and machine learning techniques.

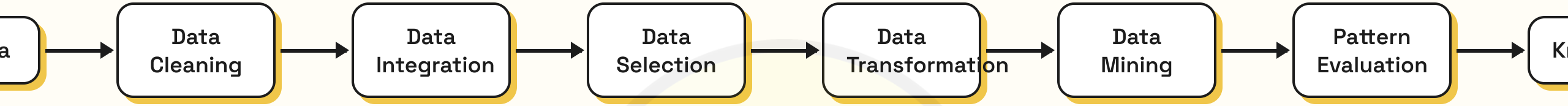
- Also known as Knowledge Discovery in Databases (KDD).
- Transforms raw operational data into strategic decision-making insights.
- Solves real-world problems like churn prediction, fraud detection, and market basket analysis.
- Sits at the intersection of database systems, statistics, and artificial intelligence.

Architecture of Data Mining System



Multi-tier Architecture of a Enterprise Data Mining System

How Data Mining Works: The KDD Process



Step-by-step Knowledge Discovery in Databases (KDD) Pipeline

Functionalities & Classifications of Data Mining

Data mining functionalities define the kinds of patterns to be discovered in data mining tasks.

- Descriptive Tasks
  - Characterization: Summarizing general features of a target class of data.
  - Discrimination: Comparing target class with one or contrasting classes.
  - Association Analysis: Finding frequent itemsets and rules (e.g., $A \Rightarrow B$).
  - Clustering: Grouping objects based on similarity without pre-labeled classes.
- Predictive Tasks
  - Classification: Predicting categorical, discrete class labels (e.g., High/Low risk).
  - Regression: Predicting continuous numerical outcomes (e.g., Sales forecasting).
  - Outlier Analysis: Identifying noise, anomalies, and rare events in datasets.

Representation of Input Data

Input data for data mining must be structured into suitable mathematical representations.

- Data Matrix: An $n \times p$ matrix (n objects across p attributes/features).
- Attribute Types: Nominal (qualitative), Binary (symmetric/asymmetric), Ordinal, and Numeric (Interval/Ratio).
- Transaction Data: Each record contains a transaction ID and a set of items (e.g., retail carts).
- Spatial & Temporal Data: Geographic coordinates, sensor feeds, and time-series sequences.
- Graph & Network Data: Structures containing nodes (entities) and edges (relationships).

Analysis Methodologies in Data Mining

Core computational and statistical methodologies applied during analysis.

- Statistical Analysis: Linear regression, ANOVA, probability distributions, and correlation tests.
- Supervised Machine Learning: Models trained on labeled data (e.g., Decision Trees, Naive Bayes, SVM).
- Unsupervised Machine Learning: Pattern discovery on unlabeled data (e.g., K-Means, Hierarchical Clustering).
- Pattern Mining: Frequent pattern growth (FP-Growth) and Apriori rule generation algorithms.
- Neural Networks: Deep architectures for feature extraction from high-dimensional unstructured data.

Introduction to Data Warehousing

DATA WAREHOUSE (DW)

A Data Warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision-making process (W.H. Inmon).

- Subject-Oriented: Organized around key business subjects like Sales, Customer, or Product.
- Integrated: Consolidates data from disparate, heterogeneous operational sources.
- Time-Variant: Stores historical data over extended horizons (e.g., 5-10 years) for trend analysis.
- Non-Volatile: Read-only analytical data; historical data is never modified or deleted.

Data Mart: Concept & Types

A Data Mart is a subset of an enterprise Data Warehouse tailored to a specific department or business line.

- Dependent Data Mart
  - Sourced directly from a centralized enterprise Data Warehouse.
  - Ensures high data consistency, central control, and unified business logic.
- Independent Data Mart
  - Built directly from operational source systems without a central DW.
  - Faster and cheaper to build, but creates data silos and inconsistency across units.
- Hybrid Data Mart
  - Combines input data from enterprise DWs as well as external operational sources.

Online Analytical Processing (OLAP) Concepts

OLAP enables fast, multi-dimensional analysis of enterprise data stored in data warehouses.

- Data Cube: Multidimensional representation of data organized by Dimensions and Measures.
- Roll-up: Aggregates data by climbing up a concept hierarchy (e.g., City → Country).
- Drill-down: Navigates from summary to detail data (e.g., Year → Quarter → Month).
- Slicing: Selects a single dimension value to produce a sub-cube (e.g., Time = 2024).
- Dicing: Selects a range of values across multiple dimensions.
- Pivot (Rotate): Rotates data axes in display to provide alternative visual perspectives.

OLAP Server Tools Architecture

Architecture	Storage Mechanism	Query Speed	Scalability	Best For
ROLAP (Relational)	Relational RDBMS (Tables)	Moderate (Depends on SQL)	Very High (Terabytes+)	Large volumes of detailed data
MOLAP (Multidimensional)	Proprietary Array Cubes	Fast (Pre-computed)	Medium (Memory limited)	Fast response on aggregated data
HOLAP (Hybrid)	Combines RDBMS & Cubes	High (Optimized)	High	Balanced scale and speed requirement

Data Modelling for Data Warehousing

Dimensional Modelling prioritizes query performance and readability over normalization.

- Fact Table: Central table storing quantitative numerical metrics (Measures) and Foreign Keys.
- Dimension Tables: Surrounding tables holding descriptive context (Attributes) about business entities.
- Surrogate Keys: System-generated integer primary keys replacing natural operational keys.
- Additive Measures: Values that can be summed across all dimensions (e.g., Total Revenue).
- Non-Additive Measures: Values that cannot be summed across dimensions (e.g., Unit Price, Ratios).

Difference between OLTP and OLAP

Feature	OLTP (Operational)	OLAP (Informational)
Primary Focus	Day-to-day transactions	Analytical decision support
Data Source	Operational databases	Data Warehouse / Data Marts
Data Structure	Highly normalized (3NF)	Dimensional (Star / Snowflake)
Operations	INSERT, UPDATE, DELETE	Complex SELECT aggregations
Data Horizon	Current real-time data	Historical data (5 - 10 years)
User Audience	Clerks, Operators, Apps	Analysts, Executives, Managers
Query Response	Milliseconds	Seconds to Minutes

Star Schema Architecture

The simplest dimensional schema structure containing a central fact table and unnormalized dimensions.

- Structure: Single central Fact Table connected directly to multiple Dimension Tables.
- Denormalized Dimensions: Dimension tables are intentionally non-normalized to reduce JOINs.
- Direct Connections: Every dimension table joins directly to the fact table (1-level hierarchy).
- Advantages: Highly simple SQL queries, fast read performance, intuitive for BI tools.
- Disadvantages: High data redundancy leading to potential storage overhead.

Snowflake Schema Architecture

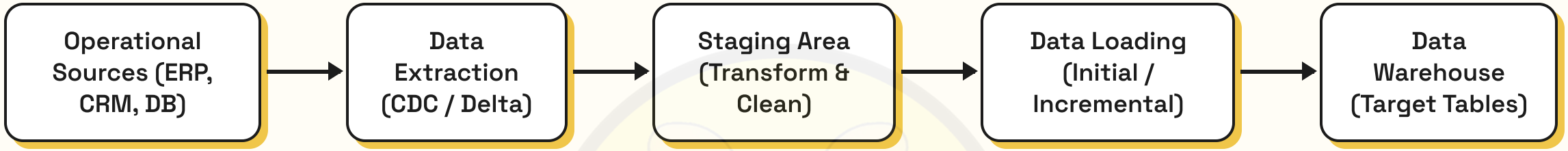
A variant of the star schema where dimension tables are normalized into secondary sub-dimensions.

- Structure: Central Fact Table surrounded by normalized, hierarchical Dimension Tables.
- Normalization: Large dimension tables are broken down (3NF) to remove redundant text.
- Multi-level Joins: Dimensions connect to sub-dimension tables forming a snowflake shape.
- Advantages: Minimizes data redundancy and saves storage space; easier maintenance.
- Disadvantages: Degraded query performance due to multi-table JOIN operations.

Comparison: Star Schema vs. Snowflake Schema

Parameter	Star Schema	Snowflake Schema
Dimension Structure	Denormalized	Normalized (Sub-dimensions)
Join Complexity	Low (1-level JOINs)	High (Multi-level JOINs)
Query Performance	Faster response times	Slower due to extra JOINs
Data Redundancy	High	Minimal to zero
Storage Space	Consumes more disk space	Optimized for minimal space
Ease of Maintenance	Complex dimension updates	Easier updates and maintenance

The ETL (Extract, Transform, Load) Process



End-to-End Data Pipeline Flow in ETL Processing

Role & Functions of the ETL Process

ETL forms the backbone of data warehousing by populating target models with cleaned data.

- Data Extraction
  - Retrieves raw data from diverse files, RDBMS, and API endpoints.
  - Uses Change Data Capture (CDC) to pull incremental delta changes.
- Data Transformation
  - Cleansing: Null handling, deduplication, and data type formatting.
  - Scrubbing: Validation against domain integrity constraints.
  - Enrichment: Business logic application, calculations, and surrogate key assignment.
- Data Loading
  - Initial Load: Bulk populating the warehouse tables for the first time.
  - Incremental Load: Periodically inserting dynamic updates (Handling SCDs).

Worked Example: Mining Association Rule Metrics

In a retail store database of 100 transactions, 30 transactions contain Milk, 40 contain Bread, and 20 contain both Milk and Bread. Compute Support and Confidence for the rule: $\text{Milk} \Rightarrow \text{Bread}$.

1. Step 1: Write formula for Support: $\text{Support}(A \Rightarrow B) = \frac{\text{Count}(A \cup B)}{\text{Total Transactions } N}$
2. Step 2: Calculate Support: \(\text{Support}(\text{Milk} \Rightarrow \text{Bread}) = \frac{20}{100} = 0.20 \text{ (or 20%)} \)
3. Step 3: Write formula for Confidence: $\text{Confidence}(A \Rightarrow B) = \frac{\text{Count}(A \cup B)}{\text{Count}(A)}$
4. Step 4: Calculate Confidence: \(\text{Confidence}(\text{Milk} \Rightarrow \text{Bread}) = \frac{20}{30} = 0.667 \text{ (or 66.7%)} \)

Support = 20%, Confidence = 66.7%. The itemset shows strong association given high confidence.

Exam Corner: High-Frequency AKTU Questions

- **Q1** Explain the KDD process in detail with a neat architectural block diagram. (10 Marks)
- **Q2** Define Data Warehouse. Discuss Inmon's four key characteristics of a Data Warehouse. (7 Marks)
- **Q3** Differentiate between OLTP and OLAP systems across all key parameters. (7 Marks)
- **Q4** Compare Star Schema and Snowflake Schema with architectural diagrams. (10 Marks)
- **Q5** Explain OLAP operations (Roll-up, Drill-down, Slice, Dice, Pivot) with examples. (7 Marks)
- **Q6** Describe the multi-tier architecture of OLAP servers: ROLAP, MOLAP, and HOLAP. (7 Marks)
- **Q7** Detail the ETL architecture and explain the key transformation activities performed in staging. (10 Marks)
- **Q8** Define Data Mining functionalities. Differentiate between Classification and Clustering. (7 Marks)

Unit 2 done. You are exam-ready.

Revise the Exam corner once more, then open Unit 3 on theycotes.com. Every AKTU subject, unit by unit, free. Notes by Chirag Jain, Ycotes.

