Data Warehousing and Data Mining Detialed Notes

Accroding Syllabus

Unit – 1

1. Need for a Data Warehouse

Operational databases (used by applications) are built for day-to-day transactions, not for long, complex analytical queries. When data grows and comes from many systems (billing, CRM, sales, web logs), managers need one **integrated, clean, historical store** to analyse trends and support decisions – this is the data warehouse.geeksforgeeks+2

Key needs:

- Combine data from multiple heterogeneous sources.
- Store many years of history for trend analysis.
- Provide fast, consistent queries for reports and BI dashboards.sap+1

Hindi video – "Introduction to Datawarehouse in Hindi | Data warehouse and data mining Lectures"
Link: https://www.youtube.com/watch?v=KgjUsie50WQyoutube

2. OLTP vs OLAP

OLTP (Online Transaction Processing):

- Handles day-to-day operations like insert/update/delete for orders, payments, registrations.
- Data is highly normalised for consistency; transactions are small, frequent and need very fast response.interworks+1

OLAP (Online Analytical Processing):

- Supports complex analytical queries over large amounts of historical data (e.g., "total sales per region for last 3 years").
- Data is denormalised into star/snowflake schemas for fast reading; workloads are read-heavy and can run longer.aerospike+1

OLTP feeds fresh data into the warehouse; OLAP queries that warehouse for insights.

Hindi video – "Data Warehouse Part-1 Explained in Hindi | OLTP vs OLAP intro"
Link: https://www.youtube.com/watch?v=7f6lFEKEY4kyoutube

3. Basic Data Warehouse Architecture

A **data warehouse** is a subject-oriented, integrated, time-variant, non-volatile collection of data for decision making. Typical high-level architecture:ibm+1

- **Data sources:** OLTP databases, files, external feeds.
- **ETL layer:** Extracts, transforms (cleans, integrates) and loads data into warehouse storage.
- **Data warehouse storage:** Central repository optimised for OLAP queries.
- **Front-end tools:** Reporting, dashboards, data-mining tools that business users see.geeksforgeeks+1

Hindi video – "Lec-1: Introduction to Data Warehouse with Examples"
Link: https://www.youtube.com/watch?v=1JEdPd05QrAyoutube

4. Data Marts and Metadata

- **Data mart:** A smaller, subject-specific subset of the warehouse (e.g., sales mart, finance mart) created to serve a particular department quickly and simply. Data marts can be dependent (sourced from enterprise warehouse) or independent (directly from operational systems).naanmudhalvan+1
- **Metadata:** "Data about data" – describes source systems, table/column meanings, transformation rules, refresh schedules, and security rules. It lives in a **metadata repository** and helps ETL developers, admins and BI tools understand and manage the warehouse.ijecs+1

Hindi video – "Datawarehouse full explained in hindi || types, design approach and tier"
(Contains clear explanation of data marts and basic metadata idea.)
Link: https://www.youtube.com/watch?v=1JEdPd05QrAyoutube

5. Three-Tier Architecture of Data Warehouse

Most warehouses use a **3-tier architecture**:geeksforgeeks+1

1. **Bottom tier – Data sources & warehouse database:**
    a. Contains the warehouse database server that stores integrated, cleaned data, fed from OLTP systems, files, etc.
2. **Middle tier – OLAP server:**
    a. Implements OLAP engine (ROLAP/MOLAP/HOLAP) that builds cubes/aggregations and answers analytical queries quickly.scribd+1
3. **Top tier – Front-end / client tools:**

a. Dashboards, ad-hoc query tools and reporting/BI applications used by managers and analysts to explore data.geeksforgeeks+1

This separation improves performance, scalability and manageability.

Hindi video – "Datawarehouse in Hindi | 3-Tier Architecture & Components"
Link: https://www.youtube.com/watch?v=KgjUsie50WQ (watch the middle section where tiers are drawn and explained)youtube

Unit – 2

1. Schema Types: Star, Snowflake, Fact Constellation

Star schema:

- Central **fact table** surrounded by multiple **dimension tables**; looks like a star.geeksforgeeks+1
- Fact table holds numeric measures (sales amount, quantity) with foreign keys to dimensions; dimension tables are denormalised, wide and descriptive (customer, product, time, location).naanmudhalvan

Snowflake schema:

- Extension of star where one or more dimensions are **normalised into sub-tables** (e.g., Product → Product, Category, Brand) to reduce redundancy.naanmudhalvan
- More complex joins but can save space and improve manageability for large, detailed dimensions.

Fact constellation (Galaxy schema):

- Multiple fact tables share common dimension tables (e.g., Sales fact and Returns fact both using Time, Product, Store dimensions).naanmudhalvan
- Used in enterprise data warehouses where many related business processes are modelled.

Hindi video – "Datawarehouse full explained in hindi || types, design approach and tier || datawarehousing"
(Explains star, snowflake and fact-constellation with diagrams.)
Link: https://www.youtube.com/watch?v=1JEdPd05QrAyoutube

2. Fact Tables

Fact tables store the **measures** of business processes and connect to dimensions through foreign keys.geeksforgeeks+1

Key points:

- Each row represents a business event at a chosen **grain** (e.g., "one line item of one invoice on one day in one store").
- Columns: foreign keys (to dimensions) plus numeric measures (sales amount, discount, units sold, cost).naanmudhalvan
- Types:
  - **Transactional fact** (most detailed, one row per event).
  - **Periodic snapshot** (summary per period, e.g., monthly account balance).
  - **Accumulating snapshot** (tracks lifecycle with multiple date keys, e.g., order from placed to delivered).

Hindi video – "Data Warehouse Part-1 Explained in Hindi | Facts and Dimensions intro"
Link: https://www.youtube.com/watch?v=7f6lFEKEY4kyoutube

3. Dimension Tables

Dimension tables provide the **context** for facts – who, what, when, where, how.geeksforgeeks

Characteristics:

- Contain descriptive, textual attributes such as customer name, city, product category, time hierarchy (day, month, quarter, year).ibm+1
- Usually denormalised to make queries simpler and faster (fewer joins).
- Support **hierarchies** for OLAP (e.g., Day → Month → Quarter → Year, City → State → Country).

Hindi video – "Introduction to Datawarehouse in hindi | Fact and Dimension tables"
Link: https://www.youtube.com/watch?v=KgjUsie50WQ (watch segment where fact & dimension tables are drawn)youtube

4. Surrogate Keys

A **surrogate key** is an artificial, numeric key (often integer) created in the warehouse to uniquely identify each dimension row, instead of using messy business keys from source systems.naanmudhalvan

Why needed:

- Business keys may be long strings or composite (e.g., country-code + customer-number) and can change over time; surrogate keys stay stable.naanmudhalvan
- Support **slowly changing dimensions (SCDs)**: when a customer's attribute (like city) changes, a new row with a new surrogate key is inserted to preserve history while keeping old facts linked to the old row.

Surrogate keys are used as foreign keys in fact tables for performance and consistency.

Hindi video – "Data Warehouse Design in Hindi | Star Schema, Dimension & Surrogate Keys"
Link: https://www.youtube.com/watch?v=J326LIUrZM8 (covers warehouse design concepts including surrogate keys)youtube

Unit-3

1. ETL Process: Extract, Transform, Load

ETL stands for **Extract, Transform, Load** – the standard pipeline used to move data from multiple sources into a data warehouse.geeksforgeeks+1

- **Extract:** Copy data from source systems (OLTP databases, CSV files, APIs, logs) into a staging area without changing meaning.
- **Transform:** Clean, standardise and reshape the extracted data (fix errors, join tables, convert formats, derive new fields) so it is consistent and analysis-ready.
- **Load:** Insert the transformed data into warehouse tables (full initial load, then periodic incremental loads).bigdataframework+1

Hindi video – ETL overview (direct link)

- ETL (Extract Transform Load) process fully explained in Hindi | Datawarehouse https://www.youtube.com/watch?v=CPFd0Q0xecgyoutube

2. Data Cleaning

Data cleaning improves data quality by handling missing, noisy or inconsistent values before loading into the warehouse.mimuw+1

Typical tasks:

- Fill or handle **missing values** (mean/median, previous value, special "unknown" code).
- Remove **duplicates** and resolve conflicting records from different sources.

- Smooth **noisy data** (e.g., using binning or averaging) and correct obvious typos or out-of-range values.learninglabb+1

Hindi video – Data cleaning as part of preprocessing

- Data Preprocessing Techniques in Detail [Hindi] – covers data cleaning, missing values and noisy data.
  https://www.youtube.com/watch?v=3imSHVySLRcyoutube

3. Data Integration

Data integration combines data from multiple sources (different databases, files, applications) into a single, coherent dataset.bigdataframework+1

Key issues:

- **Schema integration:** Matching different field names and structures that represent the same concept (e.g., `cust_id` vs `customerNo`).
- **Entity resolution:** Detecting when records from two sources refer to the same real-world object.
- **Value conflicts:** Resolving different coding schemes or units (e.g., currency conversion, date formats).learninglabb+1

Hindi video – Integration inside ETL

- DataBasics Concepts: ETL vs ELT data warehouse with example in Hindi – shows extraction from multiple sources and integration before loading.
  https://www.youtube.com/watch?v=Dwz2D4DQ_Scyoutube

4. Data Transformation

Data transformation converts data into formats suitable for analysis and warehouse schema.ijmsm+1

Common operations:

- **Standardisation / Normalisation of formats:** unify date formats, currencies, codes, etc.
- **Aggregation:** summarise detailed records (daily → monthly totals).
- **Derivation:** compute new fields (profit = revenue − cost, age from DOB).
- **Normalization (scaling)** and **encoding** for numeric attributes where needed.bigdataframework+1

Hindi video – ETL steps with practical example

- Lec-2: ETL (Extract, Transform, Load) – Data Aggregation – Gate Smashers (Hindi) https://www.youtube.com/watch?v=Tq8oCFjP6kQyoutube

5. Data Reduction

Data reduction decreases data volume while keeping the essential information, so queries run faster and storage is saved.mimuw+1

Techniques:

- **Aggregation and cube construction:** store summaries instead of every tiny detail when fine-grain is not needed.
- **Dimensionality reduction / feature selection:** keep only most relevant attributes.
- **Sampling and compression:** store smaller representative subsets or compressed forms.learninglabb+1

Hindi video – Reduction in preprocessing

- Data Preprocessing Techniques in Detail [Hindi] – has a section on data reduction and feature selection.
  https://www.youtube.com/watch?v=3imSHVySLRcyoutube

6. Discretization and Normalization

Discretization converts continuous attributes into categorical ranges, which simplifies some mining algorithms and improves interpretability (e.g., Age → {0–18, 19–35, 36–60, 60+}).egyankosh+1

Normalization (in preprocessing sense) scales numeric values into a standard range, such as 0–1 or z-scores, to avoid attributes with large scales dominating distance-based methods.geeksforgeeks+1

Hindi video – Discretization & normalization

- Data Preprocessing Techniques / Steps in Detail [Hindi] – includes clear explanation and examples of normalization and discretization.
  https://www.youtube.com/watch?v=3imSHVySLRcyoutube

Unit-4

1. OLAP Operations: Roll-up and Drill-down

OLAP lets analysts see data at different levels of detail in a multi-dimensional cube (for example, Sales by Product, Region, Time).geeksforgeeks+1

- **Roll-up (consolidation):** Move from detailed level to a higher summary level in a hierarchy, e.g., from daily sales → monthly → yearly totals or from city → state → country.bcalabs+1
- **Drill-down:** Opposite of roll-up; move from summary to more detailed view, e.g., from yearly sales down to quarter, month or day to investigate trends or problems.aws.amazon+1

Hindi video – "OLAP operations with real life example | Roll-up, Drill-down, Slice, Dice, Pivot"
Link: https://www.youtube.com/watch?v=BLqE2EKiAy4youtube

2. OLAP Operations: Slice, Dice and Pivot

Other core OLAP operations reshape the cube to focus analysis.

- **Slice:** Fix one dimension value to get a 2-D sub-cube, e.g., "sales for 2024 only" across product and region.datacamp+1
- **Dice:** Select a sub-cube by specifying ranges or sets on multiple dimensions, e.g., "sales of Mobile & Laptop in North and West regions during Q1 and Q2".studocu+1
- **Pivot (rotation):** Rotate the cube or report so that different dimensions appear on rows/columns, helping see the same data from a new angle.punjabiuniversity+1

Hindi video – "DM1 CL5 – OLAP OPERATIONS Roll up Drill Down, Slice & Dice, Pivot (with examples)"
Link: https://www.youtube.com/watch?v=HR0CxN0T0pMyoutube

3. Types of OLAP: MOLAP, ROLAP, HOLAP

OLAP servers differ mainly in **how they store and process** data for analysis.wikipedia+1

- **MOLAP (Multidimensional OLAP):**
  - Stores data in specialised multidimensional cubes; pre-computes many aggregates.
  - Very fast query performance but may need more storage and cube-processing time.pwskills+1
- **ROLAP (Relational OLAP):**
  - Keeps data in relational tables; OLAP engine generates SQL queries on the fly.

    - Scales well to huge datasets and reuses existing RDBMS but queries can be slower if aggregates are not optimised.translate.google+1
- **HOLAP (Hybrid OLAP):**
    - Combines both: recent or summary data in MOLAP cubes for speed, detailed or older data in relational form for scalability.snowflake+1

Hindi video – "OLAP Servers | ROLAP, MOLAP & HOLAP explained in Hindi"
Link: https://www.youtube.com/watch?v=HVv2d3blBqsyoutube

Unit-5

1. Definition and Idea of Data Mining

Data mining is the process of automatically discovering useful patterns, trends and relationships from large datasets using algorithms and statistics. It is usually one step inside a bigger Knowledge Discovery in Databases (KDD) process and turns huge raw data into information that can help decision-making (for example, "which customers are likely to churn").ibm+3

Hindi video – "What is Data Mining? ⛏️ | Data Mining Explained in Hindi for Beginners"
Link: https://www.youtube.com/watch?v=KijNedsQBCsyoutube

2. KDD Process (Knowledge Discovery in Databases)

The **KDD process** is the full pipeline from raw data to knowledge. Typical steps:geeksforgeeks+1

1. **Data Selection:** Choose relevant data from databases, warehouses, logs, web, etc.
2. **Preprocessing:** Clean and integrate data (handle missing values, noise, duplicates).
3. **Transformation:** Convert data into suitable forms (normalisation, feature creation).
4. **Data Mining:** Apply algorithms such as classification, clustering, association rules to find patterns.
5. **Pattern Evaluation & Knowledge Representation:** Filter interesting patterns, visualise and interpret them as useful knowledge for business.learninglabb+1

Hindi video – "Introduction to Data Mining 🔥 (Data Warehouse & Mining in Hindi)" – includes KDD overview
Link: https://www.youtube.com/watch?v=z6t_hFDtvewyoutube

3. Types of Data Used in Data Mining

Data mining can work on many kinds of data:wikipedia+1

- **Relational / structured data:** Tables from databases and warehouses (rows and columns).
- **Transactional data:** Market-basket style records (customer, time, list of items).
- **Time-series and sequence data:** Stock prices, sensor readings, click streams.
- **Spatial and spatio-temporal data:** GIS maps, satellite images, location traces.
- **Text, web and multimedia data:** Documents, web pages, images, audio and video, often converted into structured features for mining.uregina+1

Hindi video – "What is Data Mining In Hindi | Data Mining Explained" (talks about data types with real-life examples)
Link: https://www.youtube.com/watch?v=-e607XdHWvsyoutube

4. Types of Patterns Discovered

Common pattern types that data mining looks for:geeksforgeeks+1

- **Association patterns / association rules:** Show items that occur together frequently (e.g., "customers who buy bread and butter often buy jam").
- **Classification patterns:** Learn a model that assigns records to predefined classes (spam vs non-spam, good vs bad credit).
- **Clustering patterns:** Group similar records into clusters without predefined labels (customer segments, document topics).
- **Sequential / time-series patterns:** Capture order and temporal relationships (customers who buy A then B often later buy C).
- **Outlier or anomaly patterns:** Detect unusual records that do not fit general behaviour (fraudulent transactions, sensor faults).learninglabb+1

Hindi video – "Data Mining explained in Hindi in 5 minutes | Great Learning" (covers pattern types briefly)
Link: https://www.youtube.com/watch?v=G6J2Wzl5ukgyoutube

5. Applications of Data Mining

Data mining is used in many domains:geeksforgeeks+1

- **Business & marketing:** Market-basket analysis, customer segmentation, churn prediction, targeted advertising.
- **Finance & banking:** Credit scoring, fraud detection, risk modelling, algorithmic trading.

- **Healthcare:** Disease prediction, patient-risk stratification, discovering side effects from clinical data.
- **Telecom & web:** Network fault detection, recommender systems, click-stream analysis.
- **Manufacturing & science:** Quality control, predictive maintenance, experiment analysis.ibm+1

Hindi video – "L19: Data Mining Introduction | Evolution, Need | Data Warehouse and Data Mining Lectures in Hindi"
Link: https://www.youtube.com/watch?v=G6J2Wzl5ukgyoutube

Unit-6

1. Classification Techniques (Decision Tree, Naive Bayes, KNN)

Classification learns from labelled data and predicts a class for new records (spam/not-spam, loan approved/rejected, etc.).actscidm.math.uconn+1

- **Decision Tree:**
  - Data is split repeatedly on attributes that best separate classes, forming a tree where internal nodes are tests (e.g., age > 30?) and leaves are class labels.
  - Easy to understand and can handle both categorical and numeric inputs.vssut+1
- **Naive Bayes:**
  - Uses Bayes' theorem to compute the probability of each class given the features, assuming features are conditionally independent.
  - Very fast and works well for text and document classification (e.g., spam filters).ijs+1
- **K-Nearest Neighbour (KNN):**
  - Stores all training examples; for a new point, finds the k closest points and assigns the majority class.
  - Simple, non-parametric method but can be slow if dataset is large.geeksforgeeks+1

Hindi video – "Classification: Decision Tree, Naive Bayes, KNN (Data Mining in Hindi)"

- Article + Hindi lecture: https://electuresai.com/classification-decision-trees-naive-bayes-knn/electuresai
  Extra Naive Bayes in Hindi:
- https://www.youtube.com/watch?v=JpT9RqR4P5Myoutube

2. Clustering Techniques (K-Means, Hierarchical)

Clustering is unsupervised: it groups similar records together without predefined labels.geeksforgeeks+1

- **K-Means clustering:**
  - Choose k cluster centres, assign each point to nearest centre, recompute centres; repeat until assignments stabilise.
  - Good for roughly spherical clusters; result depends on choice of k and initial centres.upgrad+1
- **Hierarchical clustering:**
  - **Agglomerative:** start with each point as its own cluster, then repeatedly merge the closest clusters to build a tree (dendrogram).
  - **Divisive:** start with one big cluster and split into smaller ones.
  - Allows you to cut the tree at different levels to get different numbers of clusters.vssut+1

Hindi video – "K-Means and Hierarchical Clustering in Hindi"

- Example playlist that covers both with numericals: https://www.youtube.com/watch?v=jEdZqbqC4Zoyoutube

3. Association Rule Mining (Apriori, FP-Growth)

Association rule mining looks for frequent itemsets and rules like **IF {A, B} THEN {C}** in transaction data such as market-basket records.bcssp10.wordpress+1

Key measures:

- **Support(A→B):** fraction of transactions that contain both A and B.
- **Confidence(A→B):** fraction of transactions with A that also contain B.
- **Lift(A→B):** how many times more likely A and B occur together compared to them being independent.sis.binus+1

Algorithms:

- **Apriori:**
  - Generates candidate itemsets level-by-level, using the rule that all subsets of a frequent itemset must also be frequent.
  - Repeatedly scans the database, so it can be slower on very large data.upgrad+1
- **FP-Growth:**
  - Builds an FP-tree, a compressed structure of the database, then mines frequent itemsets from this tree without generating many candidates.
  - Usually faster than Apriori on big, dense datasets.geeksforgeeks+1

Hindi video – "Association Rule Mining (Apriori Algorithm) in Hindi | Market Basket Analysis"

- Worked support & confidence example: https://www.youtube.com/watch?v=IsSNiWIhHKE youtube

Unit-7

1. Text Mining

Text mining extracts useful information and patterns from large collections of unstructured text such as documents, reviews or social-media posts. Typical steps are tokenisation (splitting into words), removing stop-words, stemming/lemmatisation, then building features (bag-of-words, TF-IDF, embeddings) for tasks like classification, topic modelling or sentiment analysis.geeksforgeeks+2

Hindi video – "Text Mining in Data Mining (Hindi)"
Link: https://www.youtube.com/watch?v=AbfpJ6DJYHU (covers basics of text & web mining together)eduonix

2. Web Mining

Web mining applies data-mining techniques to web data. It has three main types:slideshare+1

- **Web content mining:** analysing page content (text, images, videos) to classify pages, extract information or do sentiment analysis.
- **Web structure mining:** using links between pages to find authorities and communities (e.g., PageRank).
- **Web usage mining:** mining web-server logs and click paths to understand user behaviour and improve sites or recommendations.geeksforgeeks+1

Hindi video – "Web Mining & Text Mining in Data Mining (Hindi)"
Link: https://www.youtube.com/watch?v=AbfpJ6DJYHUeduonix

3. Time-Series & Spatial / Spatio-Temporal Mining

Time-series mining discovers trends, seasonal patterns and anomalies in data collected over time, such as stock prices, sensor readings or website traffic. Techniques include smoothing, autocorrelation analysis, motif discovery and forecasting models.geeksforgeeks+1

Spatial and spatio-temporal mining focus on data with location and sometimes time (GIS maps, satellite images, GPS traces), finding patterns like hot-spots, clusters or movement paths related to geography and time.powerdrill+1

Hindi video – "Time Series Analysis & Data Mining in Hindi"
Link: https://www.youtube.com/watch?v=z6t_hFDtvew (later part discusses time-series style data-mining examples)youtube

4. Big Data Analytics

Big data analytics uses distributed storage and processing frameworks (like Hadoop, Spark) plus advanced algorithms to mine very large, fast or diverse datasets that traditional single-machine tools cannot handle. It supports use-cases such as real-time recommendation, log analytics, IoT sensor analysis and large-scale fraud detection.wikipedia+1

Hindi video – "What is Big Data Analytics? in Hindi"
Link: https://www.youtube.com/watch?v=x1Lede0Xr1c (covers big-data idea and role of analytics/mining)youtube

5. Data Mining Tools

Many **tools** help implement data-mining and machine-learning techniques without coding everything from scratch. Popular ones for academics:vssut+1

- **WEKA:** Java-based open-source GUI tool with many algorithms for classification, clustering, association rules and preprocessing; good for learning and experiments.
- **RapidMiner, Orange, KNIME:** graphical workflow tools for building and evaluating data-mining pipelines; support integration with Python/R and big-data platforms.youtube guvi

Hindi videos – Tools overview

- "WEKA Data Mining Tool in Hindi – Full Tool Explanation in 30 minutes"
  Link: https://www.youtube.com/watch?v=AbfpJ6DJYHU&t=188s (WEKA demo section)youtube
- "WEKA Tutorial in Hindi (Playlist)"
  Link: https://www.youtube.com/playlist?list=PLfX2IHFUV0cHE8lekKsvjE7Xp3mDkd3Tayoutube