UNIT 2 - BASICS OF DATA WAREHOUSING

(1) Definition and examples of data

- Data is raw, unprocessed facts, figures, symbols, or observations like numbers, words, images, sounds
- Data become meaningful information when organized and analyzed, forming the basis for decisions
- Example: student test scores (raw data) becoming an average grade (information).

(2) How data becomes information

- Data becomes information when raw facts are processed, organized, and structured to transform them into meaningful insights for decision-making.
- Example: list of temperature readings (data) becoming a trend of rising global warming (information) after analysis

(3) What is knowledge

- Knowledge is patterns, rules, or insights discovered from large datasets
- Example: by analyzing patterns and knowing which products customers often buy together

(4) Real-life examples of data, Information and knowledge

- Data : The smartwatch collects raw data such as the number of steps taken, heart rate, and sleep duration
- Information : The smartwatch app organizes and structures the data, displaying it in a comprehensible format, such as daily step count, average heart rate, and hours of sleep per night
- Knowledge : Analyzing and interpreting the information may reveal patterns, such as increased step count leading to improved sleep quality or a correlation between heart rate and workout intensity

(5) Data warehouse

- Meaning : A data warehouse (DW) is a centralized, enterprise-wide repository that consolidates large volumes of historical and current data from diverse sources for reporting, analysis, and business intelligence (BI).

- Purpose :
  - Business Intelligence & Analytics: Reports and ad-hoc analysis for deeper insights
  - Decision-Making: Provides reliable data for strategic choices, improving business outcomes.
  - Data Quality & Consistency: Cleans and standardizes data, creating a single, trusted version
  - Historical Trend Analysis: Allows tracking performance over time, essential for forecasting and planning
  - Support for AI/ML: Supplies clean and structured data for training models
  - Operational Efficiency: Separates analytical workloads from live transactional databases, improving performance

- Key features :
  - Subject-Oriented: Data is organized around major subjects. e.g. customers, products, sales
  - Integrated: Data from disparate sources (like CRM, ERP systems) is cleaned, transformed, and consolidated into a single, consistent format
  - Time-Variant: Store data over long periods (years, quarters, days) to track changes and analyze trends over time
  - Non-Volatile: Once data is loaded, it's permanent; it's not updated or deleted, only added to, ensuring historical records remain intact for consistent analysis
  - Centralized Repository: A single source for an organization's data

- Optimized for Analysis: Designed for fast querying and reporting (OLAP) rather than online transaction processing (OLTP)

(6) Difference between database and data warehouse

	Database	Data Warehouse
Purpose	Operational processing (OLTP)	Analytical processing (OLAP)
Data Type	Current, real-time operational data	Historical and aggregated data
Data Source	Single source or application	Multiple sources
Schema	Highly normalized to reduce redundancy	Denormalized for faster queries
Query Type	Simple, transactional queries	Complex, analytical queries
Users	Operational staff, application developers	Business analysts, data scientists, decision-makers

(7) Importance of data warehouse in decision-making

- Unified View & Data Integration:
  - Combines data from disparate systems (CRM, ERP, etc.) into one place
- Improved Data Quality:
  - Cleanses and standardizes data, removing errors and inconsistencies, ensuring trustworthy information for reliable decisions
- Faster, Easier Access:
  - Centralized, optimized structure allows quick querying and reporting, empowering users to get insights faster
- Historical Analysis:
  - Stores vast amounts of historical data, crucial for trend analysis, performance tracking, and forecasting future outcomes
- Predictive & Strategic Insights:
  - Enables advanced analytics, modeling, and forecasting to support proactive planning and strategic choices
- Enhanced Collaboration:
  - Ensures all departments work from the same consistent data, fostering better alignment and cross-functional decision-making

- Reduced Costs:
  - Automates data consolidation, saving time and resources, and allows skilled staff to focus on analysis rather than data wrangling

(8) OLTP and OLAP

- Meaning :
  - OLTP (Online Transaction Processing) manages real-time, day-to-day operations with high volumes of small transactions (inserts, updates, deletes) for fast, consistent data entry, like an e-commerce checkout.
  - OLAP (Online Analytical Processing) analyzes large historical datasets for complex queries, reporting, and business intelligence, focusing on insights such as sales trend analysis.
  - OLTP used for daily business and OLAP used for strategic decisions

- Differences :

	OLTP	**OLAP**
Full Form	Online Transaction Processing	Online Analytical Processing
Meaning	OLTP is about recording and managing current activities (e.g., "What's the current stock level for product X?").	OLAP is about analyzing past activities to understand trends (e.g., "How have sales of product X trended over the last five years across different regions?")
Purpose	Day-to-day operations, data capture, and real-time processing	Business intelligence, data analysis, trend identification, and decision support
Data	Current, detailed, often normalized data	Large volumes of historical, aggregated, often denormalized data (from data warehouses).
Operations	Insert, Update, Delete (CRUD)	Complex SELECT queries, aggregations (SUM, AVG, COUNT).
Focus	Speed, availability, data integrity for transactions	Query performance, multidimensional analysis,

		providing insights.
Examples	Banking transactions (deposits, withdrawals), Airline reservation systems, E-commerce order processing, Point-of-sale (POS) systems	Sales forecasting and marketing analysis, Financial reporting and budgeting, Recommendation engines (Netflix, Amazon), Customer behavior analysis

(9) Data Marts

- Meaning of a data mart :
  - A data mart is a focused, subject-oriented database containing a subset of an organization's data, designed for a specific department (like Sales, Marketing, or Finance) or business unit to enable faster, more efficient analysis and business intelligence.
  - Instead of digging through massive, complex data warehouses, a data mart provides a simplified view, like having a specific display case for your marketing team's customer data. This curated data allows analysts and business users to quickly find insights, identify trends, and make informed decisions relevant to their specific area.
  - Key Characteristics:
    - Subject-Oriented: Concentrates on a single business process or subject area (e.g., customer data, product sales).
    - Subset of Data Warehouse: Often drawn from a central data warehouse but contains only what a specific team needs.
    - Departmental Focus: Supports the analytical needs of a particular business unit.
    - Faster Performance: Smaller, targeted datasets allow for quicker querying and reporting.
    - Simpler & Cheaper: Easier and less costly to build and maintain than an entire data warehouse.

- Difference between data warehouse and data mart :

	Data Warehouse	Data Mart

Meaning	A Data Warehouse is a large, centralized, enterprise-wide repository for integrated data, supporting strategic decisions across the entire organization.	Data Mart is a smaller, subject-focused subset of a data warehouse (or independent) designed for specific departments (like sales or finance) to provide quicker, targeted insights for tactical needs.
Scope	Enterprise-wide, holistic view	Departmental or subject-specific (e.g., marketing, finance)
Data Sources	Integrates data from numerous, diverse operational systems	Subset from the data warehouse or fewer operational systems.
Purpose	Strategic, long-term analysis, single source of truth	Tactical, departmental reporting, quick insights.
Size	Large (terabytes)	Small (under 100 GB)
Complexity	High	Lower
Build Time	Long (months to years)	Short (weeks to months).
Cost	Costly	Cheaper

➢ Types of data marts :

- Data marts are primarily categorized into three types based on their data source and relationship with a central data warehouse: Dependent Data Marts, fed from a central DW (top-down); Independent Data Marts, built directly from operational systems (bottom-up); and Hybrid Data Marts, which combine data from both the data warehouse and other operational sources for flexibility.

(1) Dependent Data Mart :

- Source: Data is extracted from an existing, centralized data warehouse.
- Approach: Follows a top-down approach, ensuring data consistency and quality from a single source of truth.
- Best For: Large organizations needing consistent, integrated data for specific departments, leveraging the DW's centralized governance.

(2) Independent Data Mart :

- Source: Built directly from operational systems or external sources, bypassing a central data warehouse.
- Approach: Follows a bottom-up approach, built for specific departmental needs quickly.
- Best For: Smaller teams or specific projects needing rapid deployment; can lead to data redundancy.

(3) Hybrid Data Mart :

- Source: Pulls data from both the central data warehouse and other operational or external sources.
- Approach: Combines the best of both worlds, offering flexibility while maintaining some central control.
- Best For: Situations requiring integration of new data or ad-hoc analysis that needs both historical (DW) and real-time (operational) data.

➢ Uses of data marts in organizations :

- Data marts help organizations by providing focused, fast, and efficient access to data for specific departments like sales or marketing, enabling quicker, smarter tactical decisions, improving operational efficiency.

- Key Uses and Benefits:
  - Faster, Focused Analytics: Provide smaller, relevant data subsets for quicker query performance and faster insights
  - Improved Decision-Making: Enable data-driven decisions by delivering critical information directly relevant to a department's needs (e.g., marketing campaign performance, sales trends).
  - Enhanced Efficiency: Reduce data preparation time and IT workload by empowering business users with self-service tools.

- Data Accuracy & Governance: Create a single, reliable source of truth for a specific domain, improving data integrity and ensuring consistency.
- Cost-Effectiveness: Cheaper and faster to implement and maintain than a full data warehouse.
- Data Security & Compliance: Allow for segmented, controlled access to sensitive data.
- Departmental Autonomy: Give business units control over their own data environment, tailored to their specific workflows.
- Examples of Use Cases:
  - Sales Team: A sales data mart provides focused data on regional performance, product sales, and customer buying patterns.
  - Marketing Team: A marketing data mart offers insights into campaign effectiveness, customer segmentation, and ROI.
  - Finance Department: A finance data mart focuses on budget, revenue, and expense reporting.