Unit 1: Data Warehousing

1. Introduction to Data Warehouse

1.1 Definition

A **Data Warehouse (DW)** is a centralized repository that stores integrated, historical, and structured data from multiple sources to support business intelligence, reporting, and decision-making processes. It is optimized for **analytical processing** rather than transaction processing.

1.2 Features of a Data Warehouse

- **Subject-Oriented**: Organized around specific business subjects (e.g., sales, inventory).
- **Integrated**: Data from multiple sources is combined and standardized.
- **Time-Variant**: Stores historical data over a long period.
- **Non-Volatile**: Data is read-only and not frequently updated or deleted.
- **Optimized for Analysis**: Supports decision-making rather than daily operations.

1.3 Importance of a Data Warehouse

- Enhances **decision-making** by providing reliable insights.
- Consolidates **heterogeneous data sources** into a single, structured format.
- Enables **trend analysis** by storing historical data.
- Improves **query performance** for large datasets.

---

2. Data Warehouse Characteristics

2.1 Key Characteristics

1. **Data Consolidation**: Integrates data from various sources (databases, spreadsheets, etc.).
2. **Query Performance**: Optimized for complex analytical queries.
3. **Scalability**: Designed to handle massive amounts of data.
4. **Security and Access Control**: Ensures restricted access to sensitive data.
5. **Data Summarization**: Stores aggregated and detailed data for faster reporting.

2.2 Differences Between a Database and a Data Warehouse

Feature	Database	Data Warehouse
Purpose	Transaction processing (OLTP)	Analytical processing (OLAP)
Data Type	Real-time, current data	Historical, time-variant data
Normalization	Highly normalized (3NF)	Denormalized for fast querying
Query Performance	Fast for small transactions	Optimized for large analytical queries

3. Scope of Data Warehousing

3.1 Scope and Applications

- **Business Intelligence (BI)**: Helps organizations make data-driven decisions.
- **Customer Relationship Management (CRM)**: Stores and analyzes customer behavior.
- **Healthcare Analytics**: Manages patient data and historical health records.
- **Retail & E-commerce**: Tracks sales trends and inventory management.
- **Financial & Banking**: Detects fraud and analyzes financial data.

3.2 Benefits

- **Improved Decision-Making**: Provides a single source of truth.
- **Data Consistency**: Reduces discrepancies from multiple data sources.
- **Performance Optimization**: Faster query execution for analytics.
- **Historical Data Storage**: Supports long-term trend analysis.

4. Data Cube Technology

4.1 Definition

A **Data Cube** is a multidimensional representation of data used in **Online Analytical Processing (OLAP)**. It enables complex data analysis and visualization across multiple dimensions.

4.2 Features

- **Multidimensional View**: Allows viewing data from multiple perspectives.
- **Data Aggregation**: Summarizes data at different levels.
- **Fast Query Processing**: Optimized for analytical queries.

4.3 Operations on Data Cubes

1. **Roll-up**: Aggregates data to a higher level (e.g., weekly to monthly).
2. **Drill-down**: Goes deeper into data granularity (e.g., yearly to monthly).
3. **Slice**: Extracts a subset of the data cube based on a single dimension.
4. **Dice**: Extracts a subset using multiple dimensions.
5. **Pivot (Rotate)**: Changes the orientation of the data cube for different views.

---

5. Planning of Data Warehouse

5.1 Steps in Data Warehouse Planning

1. **Business Requirement Analysis**: Identify business goals and data needs.
2. **Data Source Identification**: Determine sources such as databases, ERP systems.
3. **Data Modeling**: Define schema and relationships (Star, Snowflake).
4. **ETL Process Design**: Extract, Transform, Load data from sources.
5. **Data Storage and Management**: Choose hardware and database solutions.
6. **Security and Access Control**: Implement role-based access.
7. **Testing and Deployment**: Validate and optimize the system.

---

6. Data Warehouse Designing Approaches

6.1 Top-Down Approach

- **Developed by Bill Inmon** (Father of Data Warehousing).
- Starts with an **enterprise-wide data warehouse**.
- Data Marts are created **after** the warehouse.
- **Advantages**:
  - Provides a comprehensive view of data.
  - Highly scalable.
- **Disadvantages**:

- High initial cost.
- Long implementation time.

6.2 Bottom-Up Approach

- **Proposed by Ralph Kimball**.
- Starts with **individual Data Marts**.
- Later, Data Marts are combined into a Data Warehouse.
- **Advantages**:
  - Faster implementation.
  - Cost-effective for small businesses.
- **Disadvantages**:
  - Integration challenges when scaling.

---

7. Data Warehouse Delivery Methods

7.1 Types of Data Warehouse Delivery

1. **Enterprise Data Warehouse (EDW)**:
   - A **centralized** warehouse for the entire organization.
   - Supports complex analytics across departments.
2. **Data Mart**:
   - A **subset** of a data warehouse focusing on a specific domain (e.g., sales).
   - Can be independent or dependent on a central data warehouse.
3. **Virtual Data Warehouse**:
   - Provides a **real-time** view of data.
   - No physical storage, retrieves data on demand.

7.2 Factors Affecting Delivery Methods

- **Business Requirements**: Determines if a company needs an EDW or Data Mart.
- **Budget and Resources**: Larger warehouses require significant investment.
- **Data Complexity**: Affects integration and retrieval speed.
