DATAWAREHOUSING NOTES BY PRADEEP

1. **Data Warehouses**

A **Data Warehouse** is a large, centralized storage system that collects and stores data from different sources, like transactional systems (for everyday operations) or external systems (e.g., social media data, customer feedback, etc.). The main purpose of a data warehouse is to help in **business intelligence** activities, which means helping a company make smart decisions by analyzing data.

Key Features of a Data Warehouse:

- **Subject-Oriented**: It is organized by specific areas of the business (e.g., sales, finance, marketing).
- **Integrated**: Data comes from many different sources (e.g., databases, files) and is converted into a consistent format.
- **Time-Variant**: Data is stored in a way that helps track changes over time (e.g., sales data from the past year).
- **Non-Volatile**: Once data is entered, it is stable and not frequently changed or deleted.

Advantages of Data Warehouses:

- **Increased Efficiency**: Makes analyzing large amounts of data faster.
- **Enhanced Data Quality**: Ensures data from different sources is consistent and accurate.
- **Improved Decision-Making**: Helps businesses make better decisions using data insights.

---

2. **OLTP vs. OLAP**

more update operations

- **OLTP (Online Transactional Processing)**:

This system is used for managing day-to-day transactions, like bank transfers or order processing in e-commerce. It's fast and handles lots of small, frequent transactions.

- Characteristics:
- Fast queries

- Many updates and inserts
- Highly normalized (data split into many small tables to reduce duplication)

- **OLAP (Online Analytical Processing)**:
This system is used for analyzing large amounts of data for business insights (e.g., sales trends). It is slower compared to OLTP but is used for complex queries.
- Characteristics:
- Large data volume
- Slower queries
- Denormalized (data is stored in fewer, bigger tables for faster retrieval)

Example:
- **OLTP**: Recording a new order in an online store.
- **OLAP**: Analyzing how many orders were placed last year from a specific city.

---

3. **Data Warehouse vs. Data Lake**

- **Data Lake**: A data lake stores data in its raw, unprocessed form. Think of it as a big "dump" of data where you store everything without organizing it first. This is useful for data scientists who might need all the raw data for advanced analysis.
- **Example**: Storing raw logs of user activity from a website.
- Tools: AWS S3, Azure Data Lake (ADLS), Google Cloud Storage.

- **Data Warehouse**: A data warehouse stores data in an organized, structured format that is ready to be used for analysis. It's like having a neat library where all the data is categorized and prepared for reporting.
- **Example**: Summarized sales data for easy reporting.
- Tools: Snowflake, Vertica.

- **Data Lakehouse**: A combination of both, where data can be stored in both raw and

processed forms.

---

4. **ETL (Extract, Transform, Load)**

ETL is the process used to move data from different sources into a data warehouse.
1. **Extract**: Pulling data from various sources (like databases, files).
2. **Transform**: Cleaning and converting the data into a consistent format.
3. **Load**: Loading the transformed data into the data warehouse for analysis.

Example:
- Extract customer data from multiple systems (e.g., CRM, website logs).
- Transform it to match the same format (e.g., standardize date formats).
- Load it into the data warehouse for analysis and reporting.

---

5. **Dimensional Modeling**

Dimensional modeling is a method of organizing data in the warehouse to make it easy and quick to access for reporting and analysis.

Key Components:
- **Fact Table**: Stores numerical data (e.g., sales amount).
- **Dimension Table**: Stores descriptive information (e.g., product name, customer region).

Example:
- **Fact**: Sales amount ($500).
- **Dimension**: Product (Laptop), Region (New York), Time (August 2023).

5.1. Why Use Dimensional Modeling?

The main goal is **faster retrieval** of data for reports, as it is easier to query when data is structured this way.

5.2. Fact Table

- A **Fact Table** contains measurable data like sales amounts or number of items sold. It's large and stores all the key measurements we want to analyze.
- **Example**: The total revenue generated by selling laptops in August 2023.

5.3. Dimension Table

- A **Dimension Table** holds descriptive, often static information, like customer names or product descriptions. This data doesn't change often.
- **Example**: Product details (Laptop specifications) or Customer location (New York).

---

6. **Fact Tables Types**

- **Transaction Fact Table**: Each row represents one event, such as an order or a sale.
- **Example**: A single order in an e-commerce store.
- **Periodic Fact Table**: Data is aggregated over a time period (e.g., week, month).
- **Example**: Monthly sales totals for each region.
- **Accumulation Fact Table**: Tracks the progress of a process over time.
- **Example**: Tracking a customer's journey from placing an order to delivery.

---

7. **Surrogate Key vs. Natural Key**

- **Surrogate Key**: An artificial, unique identifier created in the database (e.g., a random ID assigned to a product or customer).
- **Example**: Instead of using a customer's email as a key, the system might generate "Customer ID: 12345".

- **Natural Key**: A real-world identifier, like a Social Security number or an email address.
- **Example**: Using an email address as a unique identifier for a customer.

---

8. **Date Dimension**

The **Date Dimension** is a special table in the data warehouse that stores all the possible dates (days, months, years) along with useful information like holidays, weekends, or fiscal periods. This helps in reporting and analyzing data based on time periods.

Example:
- **Date**: August 1, 2023
- **Attributes**: IsHoliday = No, DayOfWeek = Tuesday

---

9. **Star Schema vs. Snowflake Schema**

- **Star Schema**: A simple and fast structure where a central fact table connects to multiple dimension tables. It looks like a star.
- **Example**: A fact table for sales connects to dimension tables for time, products, and regions.

- **Snowflake Schema**: Similar to a star schema but with additional levels of normalization, meaning the dimension tables are further split into smaller tables.
- **Example**: Instead of one product dimension, it might have separate tables for product categories and brands.