**REPORT: DE Training (28/10/2024)**

**DAY 1**

**C.Sai Anand**

**Data Engineer**: Gathers data from various sources, processes it, and loads it into databases. This process is known as **ETL** (Extract, Transform, Load) and is crucial for organizing raw data and making it ready for analysis.

**Data Analyst**: Focuses on analyzing data to find insights. They also perform an **ELT** process, where data is loaded and then transformed inside the database to make it usable for reports and analysis.

**Data Scientist**: Combines data engineering with ML models. They particularly utilise historical data to build models and make predictions, such as forecasting future trends or customer behaviors.  
  
**Structured Data**: Organized in tables with the help of rows and columns, where each row is a record and columns are fields. For example, a customer table has rows for each customer and columns for details like name and address. This data is stored in **relational databases** using primary and foreign keys.

**Semi-structured Data**: Not organized in tables but still has a clear structure, like **JSON** and **XML** files.

**Unstructured Data**: Does not follow a particular structure; includes images, audio, video, and text files. Often stored in **non-relational databases**.

**Database**: An organized collection of data where we can perform CRUD operations (Create, Read, Update, Delete) on the data we are having.

**SQL (Structured Query Language)**: A language to manage and retrieve structured data from relational databases.

**Data Warehouse**: A large database used to store **historical** data, which allows companies to keep a record of past data for long-term analysis. Like if database is fulled we can transfer the record date to ware house which directly leaves the space to database and store more data.

**Data Mart**: **A subset of a data warehouse** designed to serve specific departments or data needs.

**OLTP (Online Transactional Processing)**: Handles daily, real-time data transactions from users. It's quick and ideal for ongoing operations (e.g., online purchases or bank transactions).

**OLAP (Online Analytical Processing)**: Used for analyzing historical data, supporting analysts in generating reports and insights. It's slower but designed to handle large volumes of data.

### **Examples of HDFC Bank**

* **OLTP Example**: Daily transaction data (e.g., current bank statements).
* **OLAP Example**: Historical data, such as comparing monthly statements from last year to this year.

**Operational Data Store :** ODS : USED FOR OPERATIONAL REPORTING AND SUPPORTS CURRENT OR NEAL REALTIME REPORTING REQUIREMENTS

Just go through for better understanding

ARCH : AFTER EVERY SIX MONTHS, NEED TO SEND DATA FROM OLTP TO OLAP

JAN FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC

1 2 3 4 5 6 7 8 9 10 11 12

OLTP

OLTP : AUG SEPT OCT NOV DEC JAN (2024)

OLAP : JAN TO JULY 2023

CEO : COMPARE JAN (2023) WITH JAN (2024)

ODS : ANOTHER SAGING DATABASE WHERE YOU CAN SEND THE DATA OF JAN 2024 MONTH TO ODS AND FROM ODS DATA AND FROM DWH (data warehouse) DATA WE CAN CREATE A REPORT.

OLTP : ERD : ENTITY RELATIONSHIP DIAGRAM : MASTER AND TRANSACTIONAL TABLES

OLAP : DIMENSION AND FACT TABLES

### 

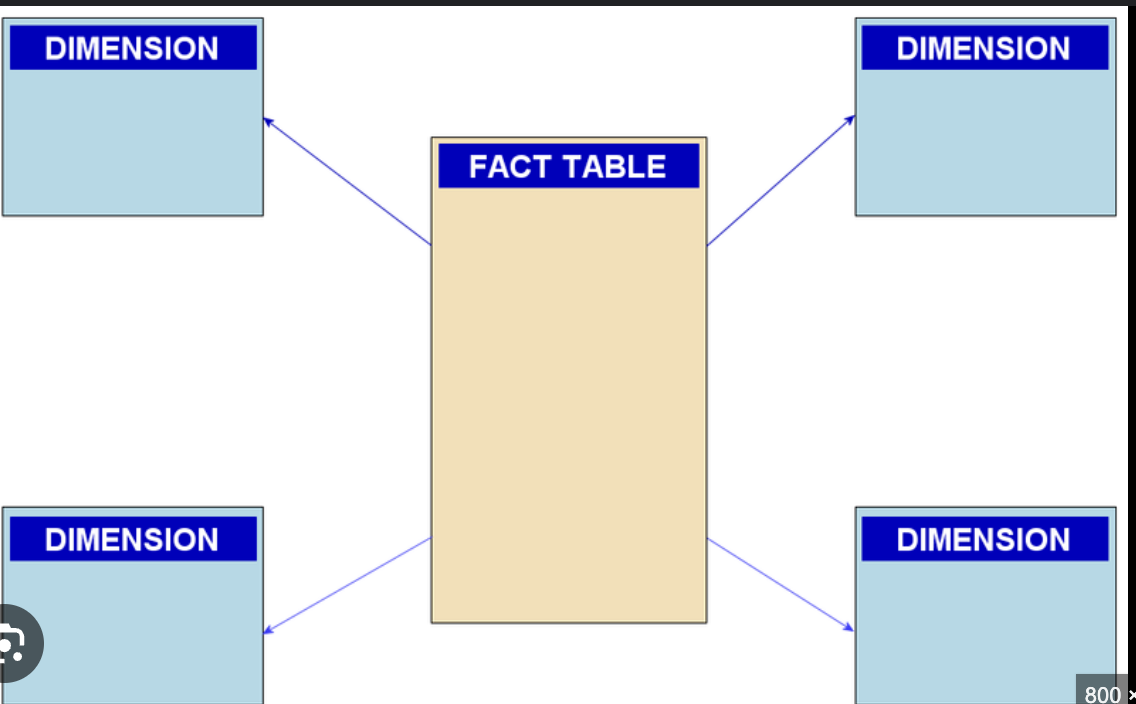
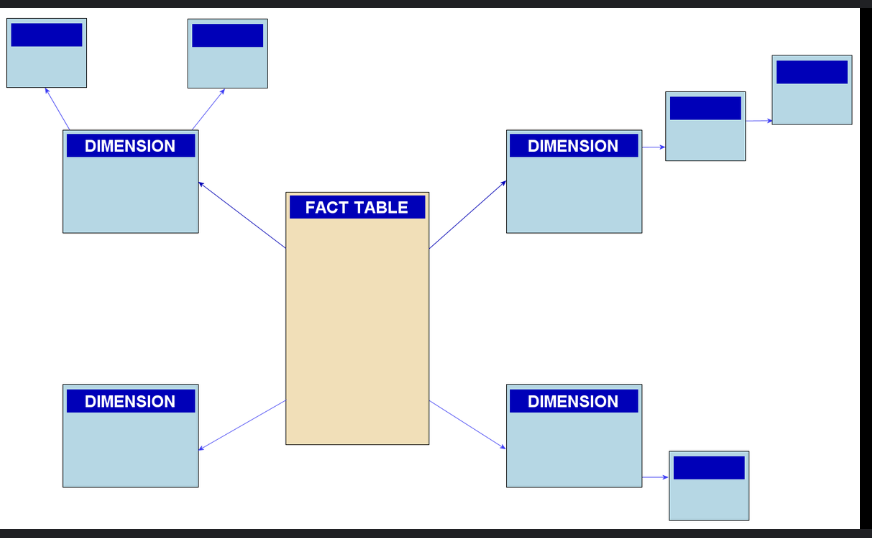
### 

### 

### **Database Schemas and Diagrams**

* **ERD (Entity Relationship Diagram)**: Used in OLTP databases to show relationships between tables. Includes:
  + **Master Tables**: Basic information, like customers or products, with less data.
  + **Transactional tables**: store actual transactions, like sales.
* **Dimensional and Fact Tables**: In OLAP:
  + **Dimension Tables**: Provide descriptive details (e.g., customer information).
  + **Fact Tables**: Store measurable data, like sales numbers, linked to dimension tables.

### **Schema Types**

* **Star Schema**: **Dimension** tables connect **directly** to the **central fact** table.****
* **Snowflake Schema**: A variation where **dimension tables** are further **broken down** into **additional tables**
* .
* **Galaxy Schema**: A complex schema used in large-scale databases with **multiple fact tables.**
* 