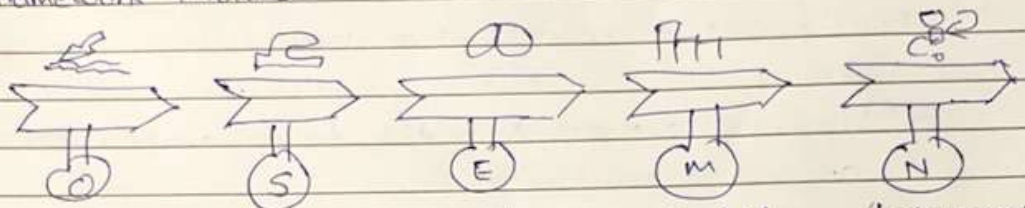
DATA SCIENCE - UNIT 1.

Q1. Explain Data Science Framework.

ANS: A Data Science Framework is a structured path/approach that guides the entire data science process from data collection to deploying and maintaining models. It provides a systematic way to tackle data problems by defining steps and methodologies to ensure consistent and reliable outcomes.

Framework Involves:



Using the above model of framework we can transform the raw data into actionable insights. Understanding it using examples

(i) Obtain:

Collect data from various sources such as databases, API's or online repositories.

Ex: Gathering customer transaction data from a company's database.

(ii) Scrub:

Clean and preprocess the data to remove inaccuracies, inconsistencies and missing values.

Ex: Removing duplicate enteries and filling in missing values in transaction data.

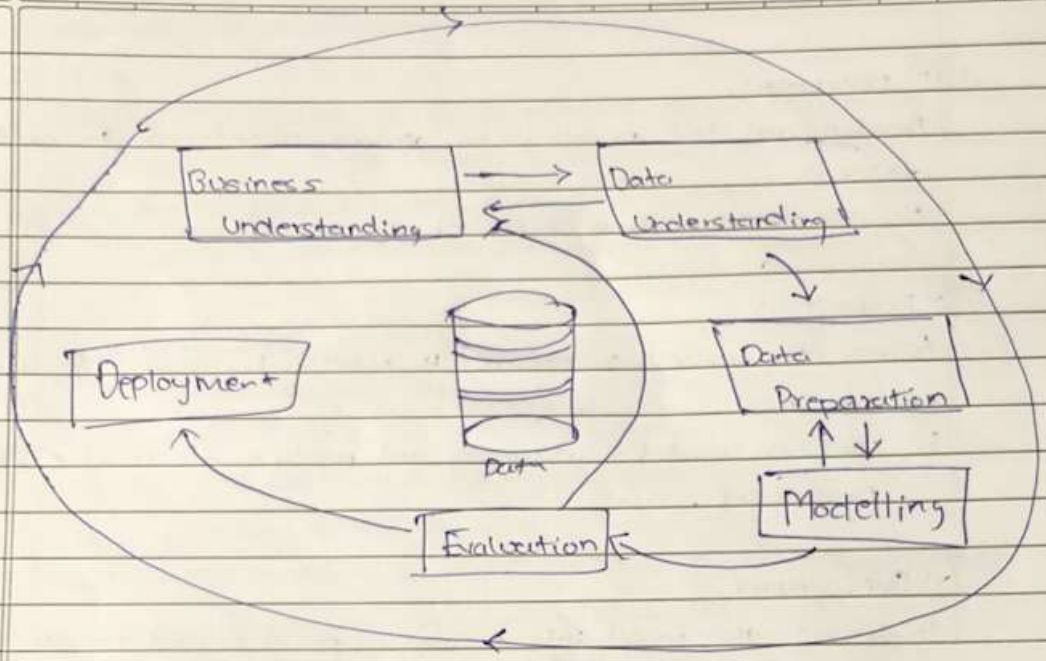
(iv) Explore
Perform explanatory data analysis [EDA] to understand patterns, trends and anomalies in the data.
Ex: Creating visualization to see spending patterns and distribution of transaction amount.

(iv) Model
Build predective or descriptive models using learning machine algorithm.
Ex: Developing a model to predict future customer spending based on past transaction data.

(v) Interpret
Analyze and interpret the results to derive meaningfull insights and actionable recommendations.
Ex: ~~High~~ Identifying high-value customers and ~~tailor~~ tailoring marketing strategies to them based on model predictions.

These steps ensure a structured approach to data science projects, maximizing value of extracted from data.

Q. Discuss the Cross-Industry Standard Process for Data Science [CRISP-DM]
Ans: The CRISP-DM is an widely used methodology for data scienc, mining and analytic projects. It provides a structured approach to planning and executing projects, ensuring consistency and reliability.
Phases of CRISP-DM.



① BUSINESS UNDERSTANDING:

Define project objective and allign them with busines goals, understand what the business ~~eterm~~ aims to achieve and translate this into data mining objective.

Ex: A company, A retail company wants to reduce customer churn

② DATA UNDERSTANDING

Collect and explore new data or initial Data, Analyze its characteristics spot patterns, and verify its quality.

Ex: Gather customer transaction data and demographics.

③ Data PREP

Clean preprocess and prepare data for modelling, Handle missing values, remove duplicates, and create new features.

Ex: Create feature like purchase freq & avg spending.

(iv) MODELING:
Develop and test models using various algorithms to predict outcomes.
Ex: Use decision tree to predict customer churn.

(v) Evaluation
Assess the model to ensure it meets business objectives and perform well.
Ex: Evaluate model's accuracy and position for reliable churn predictions.

(vi) DEPLOYMENT
Implement the model into business process and monitor its performances.
Ex: Integrate the model into CRM system to identify at risk customers.

Q. Organize SuperStep.

ANS Organize SuperStep typically refers to a phase with parallel processing frameworks, like Apache. Where distributed computations are co-ordinated and synchronized.
In parallel processing, a Superstep is an co-ordinated phase where each processing unit (often called a node or vertex) perform computations and communicates with other units. During the organize Superstep, the system co-ordinates these tasks to ensure data consistency and efficient computation

Ex: Imagine you're using a parallel graph processing framework to analyze social network data, such as finding the shortest path between friends in a large network. Each vertex represents a user, and edges represents friendship.

STEPS IN BSS.

(i) Computation

Each vertex performs computation based on incoming messages from previous superstep.

Ex: A vertex might calculate a shortest path to its neighbour

(ii) Communication

Vertices send messages to their connected neighbours.

Ex: A vertex updates its neighbour with new shortest path info.

(iii) Synchronization

All vertices wait until ~~they~~ every vertex has completed its computation and communication. This ensures that no vertex proceeds to next superstep until all are synchronized.

Q. Dicuss Rule of European Union General Data Protection Regulation GDPR - EU.

Ans The Eu - GDPR is an comprehensive data protection law in EU. ensuring individual's privacy and controll over their personal data.

Here Are Some Key Rules :-

(i) Lawfullness, Fairness & Transparency.
Data processing must be lawfull, fair and transparent to the data subject

(ii) Purpose Limitation
Data collected for specified explicit and legitimate purposes must not be further processed in manner incompatible with these purposes.

(iii) Data Minimization
Only that data is necessary for purpose of processing, should be collected an processed

(iv) Accuracy
Personal Data must be accurate and kept upto date.

(v) Storage Limitation
Personal data should be kept in a form that permits identification of data subjects for no longer necessary

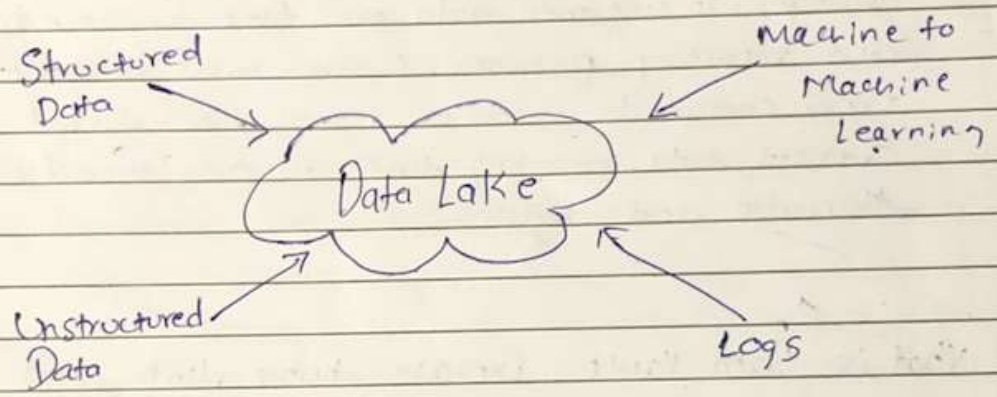
(vi) Integrity & Confidentiality
Data must be processed in a manner that ensures security, including protection against unauthorized or unlawfull processing

Accountability : Data Controller is responsible for demonstrating compliance with all GDPR principle

Q. Explain Data lake & Data Vault.

Ans. Data lake :

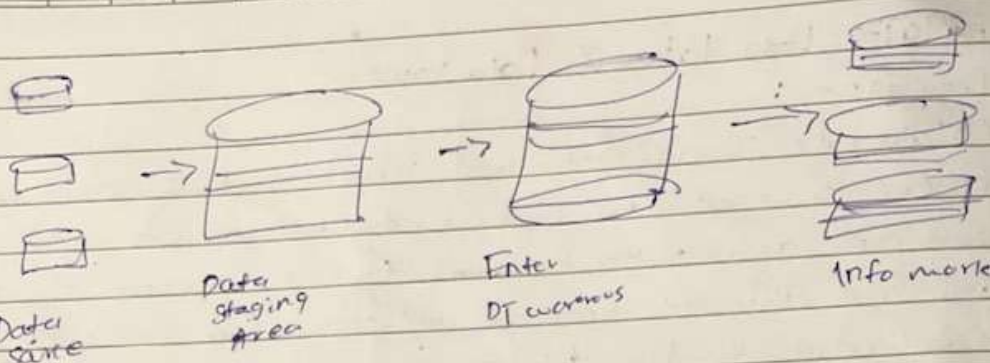
A Data lake is an centralized repository that allows you to store all your structured and unstructured data at any scale. You can store data as-is without having to first structre the data, and run different types of analytics, from dashboards and visualization to big data processing, real time analytics, and machine learning.



Ex: Imagine a company collects massive amounts of data from various sources like social media, IOT devices, and transaction records. All this data, in its raw form, it is stored in an data lake. Analysts and data scientist can then access this data to perform their analysis and extract insights without worrying about predefined structures.

Data Vault :

Data vault is DB modelling method that is designed to provide long-term historical storage of data coming in from multiple OS. It focuses on capturing all changes over time and providing a scalable, flexible, and consistent data repository.



Ex: Consider a financial institution that needs to track changes in customer data over time for compliance and reporting purposes. A data vault is model captures every change to customer records, including historical & current data, ensuring that any data modification and traceable and audits.

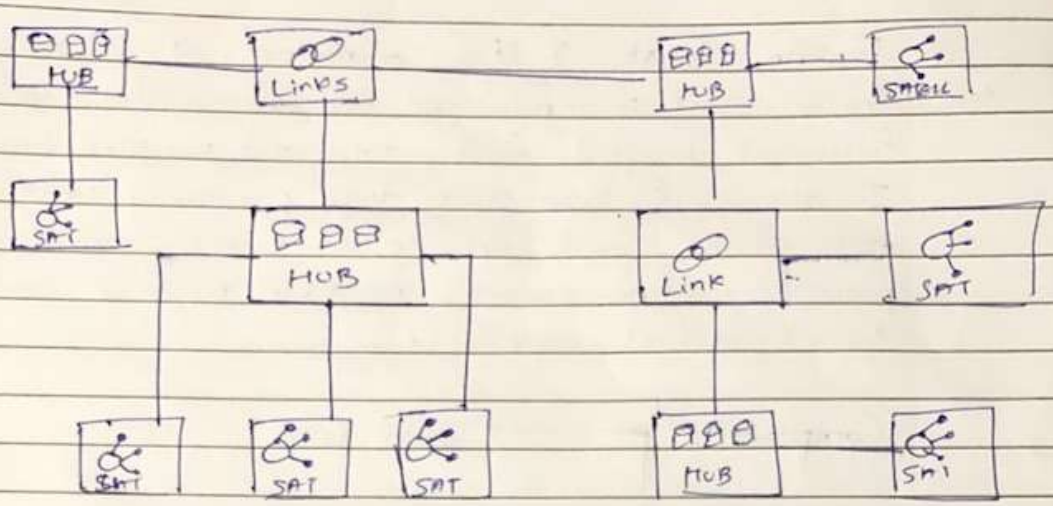
Q. What is Data Vault. Explain hubs, links and satellites with respect to data vault

Ans A Data Vault is data modelling approach and methodology used in enterprise data warehousing to handle complex and varying data structures.

~~This~~ It helps and provides a flexible, agile and scalable solution for integrating and managing large volume of data from diverse sources to support enterprise-scale analytics.

Data Vault Modelling Consist of 3 entities:

- HUBS
- LINKS
- SATTELITES.



① HUBS.

Hubs Represents core business concepts or entities and store their unique business keys.

Ex: In Retail System, a hub might represent entities like Cust, Prod, or Store. Each hub stores business key (Cust ID) and its metadata.

② LINKS

Links capture the relationships between hubs. They connect hubs together to represent business process and association.

Ex: A link might repr relat btwn customer and order, where each order is linked to a customer. The like table may store Cust & order ID.

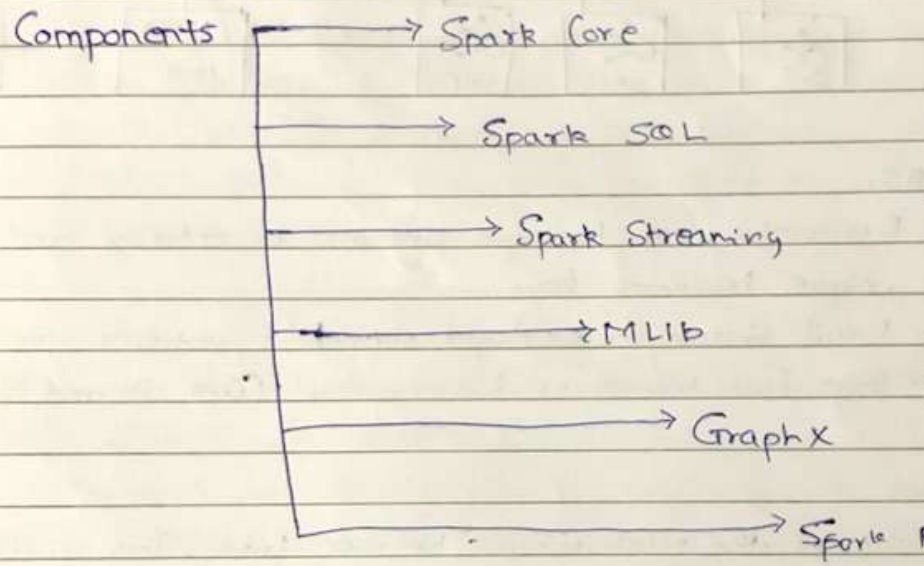
③ SATELLITE

Satellites contain descreptive that related to hub or link and store historical context.

Ex: A sat is link to cust hub might store Cust attri like name, add and contact detail, capturing changes over time to make a historical records.

Q. Explain Spark & it's Component as OS process tools.

ANS: Apache Spark is a powerfull open-source data processing framework designed for big data and machine learning applica. It allows for large scale data processing with high speed ease of use. Sparks operate on clusters, which means it processes data across multiple machine and is built on top of several component.



(1) SPARK CORE

The foundation of spark, responsible for I/O responsibilit-s, task Scheduling and memory management. It manages job execution Provides API's for creating & managing RDD's and handle fault recov

(11) SPARK SQL

A Module for structured data processing. It allows quering data Via SQL and supports data frame and datasets. It integrates

with various data sources like Hive, Parquet & JSON.
Ex: Running SQL queries on large datasets stored in HDFS or other storage system.

(iii) SPARK STREAMING
A module for processing real-time data streams. Enable real-time analytics and integrates with kafka, and TCP sockets.
Ex: Real time log analysis, monitoring application logs for insights and alerts.

(iv) MLlib
A scalable machine learning library. It Provides tools for ML algo's and utilities, including classification, regression, clustering and collaborative filtering.
Ex: Building and deploying ML models for predicting cust churn or product recommendations.

(v) GRAPH - X
A library for graph processing. It supports the creation, manipulation and analysis of graph: parallel computation.
Ex: Analyzing social networks to find influential nodes.

(vi) SPARK - R
An R package that provides a frontend to use SPARK with R. Role Allow data scientists to leverage the Power of SPARK within the prog lang.

Ex:

Dat Coll — DC: Collecting sales, customer, and product data from various sources.

Dat Proce — DP: Using Spark Core and Spark SQL to clean and transform the data into a structured format.

Mac lear — ML: Leveraging MLlib to build models for customer segmentation and targeted marketing.

Grap An — GA: Applying GraphX to understand social connection and influence among customer for improved marketing strategies