ASSIGNMENT :- 1st

Q1. what is data science? Explain different stages involved in a Data Science Project?

Ans

Data Science

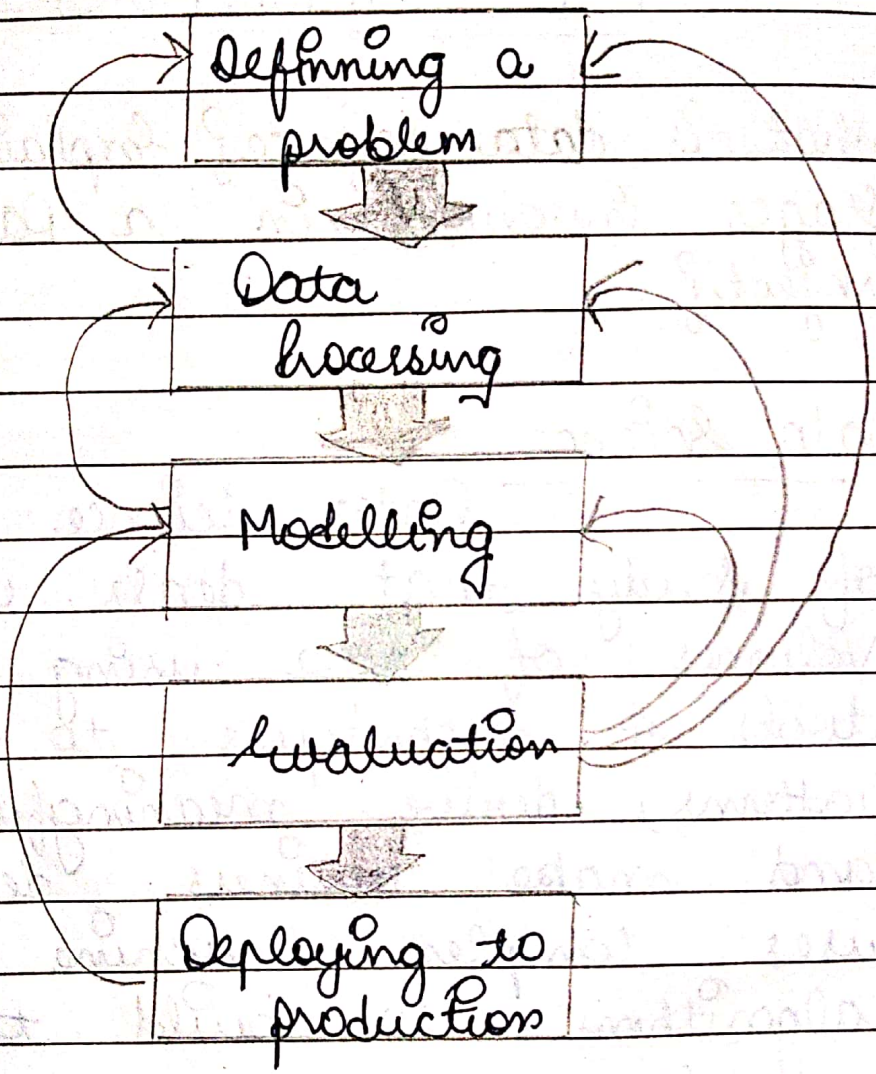
Data science is the domain of study that deals with vast volumes of data using modern tools & techniques to find unseen patterns, derive meaningful information and make buisness decisions. It uses complex machine learning algorithms to build the predictive models.

Eg Driverless Vehicle, Finance, Google maps, Healthcare, Manufacturing etc

Stages in a Data science Project:-

Data science projects can take

on very different challenges & focuses resulting in diff methods are used &



Stages in data science project

(i) Defining a problem:- The first stage of any data science project is to identify and define a problem to be solved. Without a clearly defined problem to solve, it can be difficult to

know how to tackle to the problem. A challenge with this is being able to define a problem small enough that is can be solved / tackled individually. An eg. of this would a manufacturing firm that is not profitable.

(ii) Data Processing :- Once you have your problem, how you are going to measure success, and an idea of the methods you will be using, you can then go about performing all important tasks of data processing. There are a variety of tasks that need to occur at this stage depending on what problem you are going to tackle.

(iii) Modelling :- The next part, and often the used of data science project. The format this will take will depend primarily on what the problem is & how you defined success in first

Step & how processed the data. It is often better to use multiple models at this stage to be able compare & contrast them with comes it evaluation.

(iv) Evaluation :- Once you created & Implemented your models, you need to know how to evaluate it. Depending on how you processed your data and set up your model, you may have a holdout dataset or testing data so that can be used to evaluate your model.

(v) Deployment :- This can mean a variety of things such as whether you use the insights from the model to make changes in your business, whether you use the model to check whether changes that have been successful or whether the model is deployed.

Q2. Explain briefly:-

a) Data Re-Processing

Ans Data Re-Processing, a component of data preparation, describes any type of processing performed on raw data to prepare it for another data processing procedure. Data Re-Processing transforms the data into a format that is more easily and effectively processed in data mining, machine learning & other data science tasks. This techniques are generally used at the earliest stages of machine learning & AI development pipeline to ensure accurate results.

b) Data cleaning

Ans Data cleaning is the process of identifying and fixing incorrect data. It can be incorrect format, duplicates, corrupt, inaccurate, incomplete or irrelevant. Various fixes can be made to the

data values representing correctness in the data. The data cleaning and validation steps undertaken for any data science project are implemented using a data pipeline.

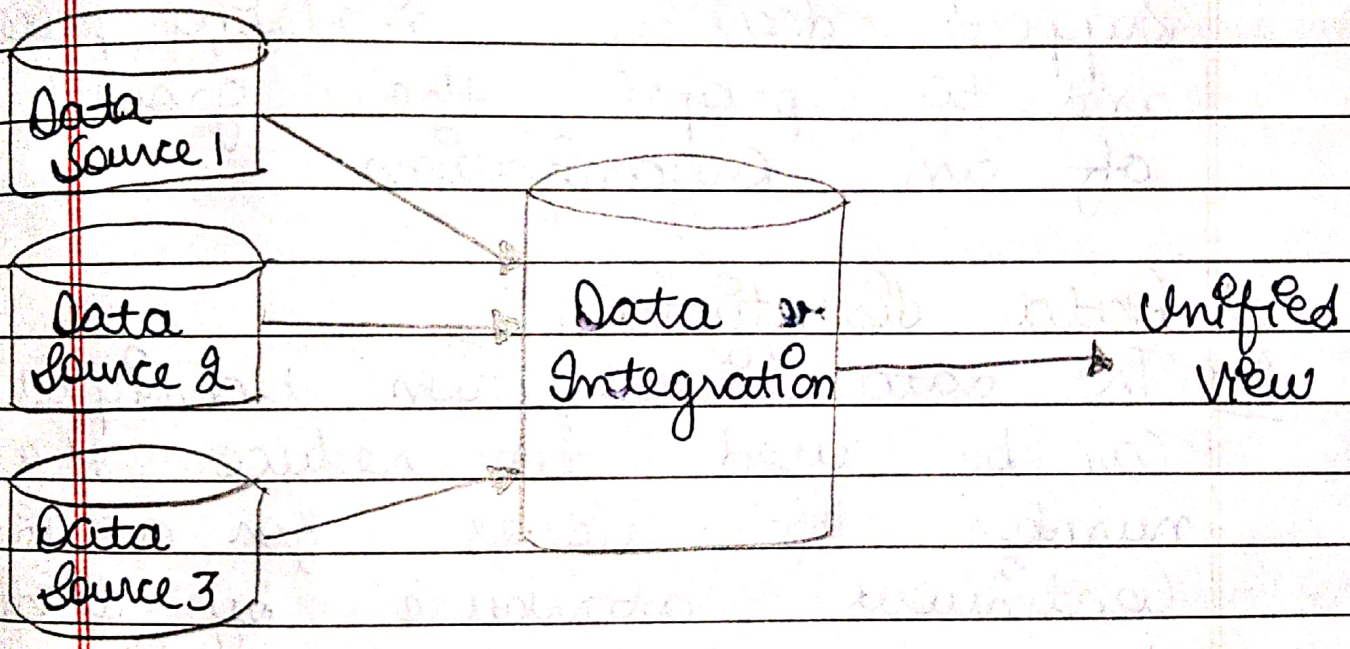
Advantages

(i) your results will be accurate & consistent.
(ii) Maintaining data quality & enabling more precise analytics that support the overall decision making process are made.
(iii) Avoiding unnecessary costs & errors.

Q) Data Integration & transformation

Ans Data Integration is the process of combining data from different sources into a single, unified view. Integration begins with the ingestion process, and includes steps such as cleansing, ETL mapping &

transformation. Data Integration ultimately enables analytics tool to produce effective, actionable business intelligence. There is no universal approach to data integration.



Data transformation is the process of converting data from one format to another, typically from the format of a source system into the required format of a destination system. Data transformation is the component of most data integration

& data management tasks, such as data wrangling & data warehousing.

- It is a process of converting, cleansing, and structuring. data into a usable format that can be analyzed to support decision making process, and to propel the growth of an organization.

d) Data discretization

Ans The data discretization techniques can be used to reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can be used to restore actual data values. It can be restoring multiple values of a continuous attribute with a small no. of interval labels therefore decrease & simplifies the original information.

Q3. What are the applications of data science. Also, write about the data security issues?

Ans

Applications of Data Science

(i) Healthcare :- It uses to build sophisticated instruments to detect & cure diseases.

(ii) Gaming :- Video & computer games are being created.

(iii) Image Recognition :- Detecting the images such as facebook, Instagram & Twitter.

(iv) Recommendation Systems :- Netflix & Amazon give make & product recommendations based on watch, purchase & browse on their platforms.

(v) Logistics :- delivery of products & increase operational efficiency.

(vi) Fraud Detection :- Banking & financial institutions are used.

(vii) Internet Search :- Here are other search engines such as yahoo, Duckduckgo, Bing, AOL, Ask & others.

(viii) Speech recognition :- Have you ever needed a virtual speech assistant like alexa or Siri.

Data security Issues

Big data is crucial for any businesses to succeed on the data driven world. With several advanced infrastructure, organizations have streamlined the flow of data for real time insights delivery & better decision making.

There are 7 main data security Issues are as follows :-

(i) Data Storage :- Businesses are adopting cloud data storage to move their data easily to expedite business operations. Even the slightest mistake in controlling the access of data can allow to get a host of sensitive data.

(ii) Fake Data :- Fake data generation pores a severe threat to businesses as it consumes time that otherwise could be spent to identify or solve other pressing issues.

(iii) Data Privacy :- It is a big challenge in digital world. It aims to safeguard personal or s/w information from cyberattacks, breaches & potential & unintentional data loss. Data privacy is the help of access management services in cloud.

(iv) Data Management :- A Security breach can have crushing consequences on businesses, including the vulnerability of critical business information to completely comprised database. Deploying highly secured databases is vital to ensure data Security at all levels.

(v) Data Access control :- Controlling which data users or can view or edit enables companies to ensure not only data integrity but also preserves its privacy.

(vi) Data Poisioning :- There are several machine learning solutions like chatbots that are trained on colossal amount of data. The results can be catastrophic ranging from logic corruption to data manipulation & data Injection.

(18) Employee Theft :- The risk of an employee leaking sensitive information, intentionally or unintentionally is high. Employee Theft is prevalent not only in big Tech companies but also in startups. To avoid employee theft, companies have to implement legal policies along with securing the networks with virtual private networks.