Q1] Explain Data Science & Application of Data Science

→ • Data is a collection of facts and figures which relay something specific. It can be number, word, measurement or description of something. Data is a raw material.

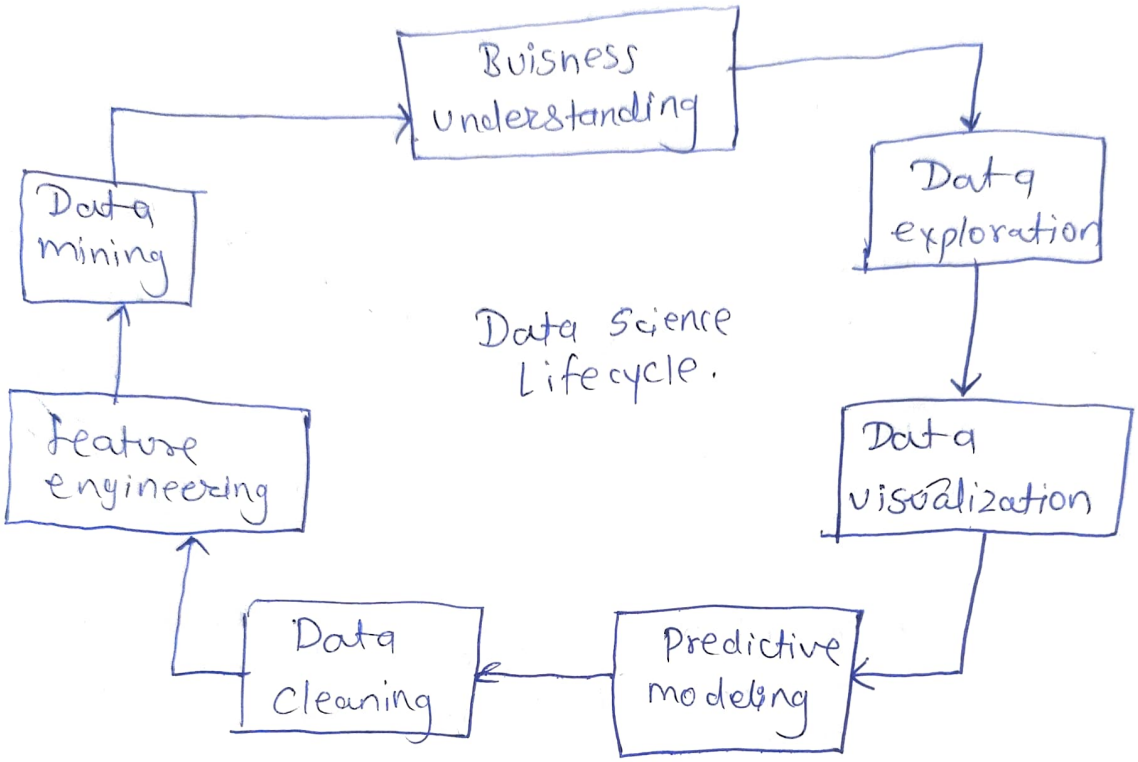
- Types of data are record data, data matrix document data, ordered data, graph data etc..
- Data Science is the field to extract the ~~data~~ information from various form of data.
- Data science aims to discover Knowledge from the data which can be used in buisness decision & prediction.
- Data science can help in buisness Analysis. From historical data, instead of knowing how many product sold it forecast the future sale of product.

Applications of Data science. →

1] Heathcare
2] Gaming
3] Image Recognition
4] Logistics
5] Predict future market trends
6] Reccomendation System

Q] Explain Data Science Life cycle.

→ Datascience lifecycle is as follow



Data Science Lifecycle.

a] Business Understanding → Understand problem to solve

b] Data exploration → Understand pattern & bias your data.

c] Data Visualization → Create and study of the visual representation of data.

d] Predictive modeling → It is the step where the machine learning finally comes into your data

e] Data Cleaning → Detect and correct corrupt or inaccurate data.

f] Feature engineering → Process of cutting down the feature.

9] Data mining → Gathering your data from different sources.

8] Data explosion plays major Role in big data. justify the Statement with proper explaination along with examples. & Factor responsible for data explosion

or

8 Explain 7 V's of Big Data.

→ • Data explosion means rapid growth of the data from different resources and stored in computer system is called data explosion.
• Data is generated automatically through mobile devices, computer System. etc...
• The phenomena of exponetial multiplication of data that get stored is termed as data explosion.
• Sending email, making phone calls, collecting info each day we create massive ammount of data.

* Role of Data Explosion in Big Data *

1] Volume : Volume refers to the size of the dataset.
• It could consists of billions of rows & millions of columns.
• Usually this dataset is stored in multitiered storage.
• for example imagine a Sale on flipkart, a person views 10 mobile a day of 5 different companies each it would make 50 data points now let's suppose flipkart have 1 million active users on Sale day it would make 50 million data points on Single day.

2] Velocity → Velocity refers to the rate of speed at which data is generated & processed. It can be based on real time or based on historical in nature.

- Examples → In stock market millions of transaction are made in real time.

3] Variety → It refers to the data accumulated from multiple data sources. It can be structured, semi-structured or unstructured.
eg → Ecommerse collect structured (sale record), sem-struct (review) unstructured (photos & videos).

4] Veracity → It refers to the measure of data quality & usefulness of the data.
- The data should be true or relevant we could not perform useful analysis if incomming data is false or has error.
eg → fake news.

5] Value → Value measures the usefulness of data which helps in improving business decisions or enhancing AI models.
eg → Netflix uses data to recommend shows to user.

6] Variability → The changing nature of data, including seasonal trends & inconsitencies
eg → Google news.

7] Visualization → The ability to represent complex Big data in visualizing or understandable way using charts, graphs & dashboard.
eg → COVID-19 uses heed map to show spread of virus

*Factors responsible for data Explosion *

1] Growth of internet
2] Social media & Digital Content
3] Smart Devices
4] Ecommerce & Online Transaction
5] AI & ML

Q] List and explain data processing infrastructure challenges in Big data with suitable example.

→ • Big Data requires a strong infrastructure to process large volumes of data efficiently.
• Challenges in big data infrastructure are →

1) Data storage → The increase in volume of data, increases the need for storing data. It requires the medium with higher I/O speed to store the data. Tradition database struggle to handle such large datasets.

eg → Facebook stores petabyte of data which requires hadoop HDFS.

2] Processing Speed → Processing high-velocity data in real time is difficult, especially in applications that requires instant responses.

eg → Stock market platform have millions of transactions per second, so they uses Apache spark

3] Data Integration → Big data comes with various sources like databases, sensors, api's etc. Integrating different format & structure is complex
eg → E-commerce platform integrates data from review, rating, purchase in real time update.

4] Scalability → As data growth exponentially. traditional IT infrastructure fails to scale effectively.
eg → Netflix uses cloud based Aws for streaming.

5] Data privacy & Security → Protecting data from cyber threats, unauthorized access is crucial. Data must be encrypted & should not shared.

6] Higher Cost → Growth in data increased the demand of data storage which is directly proportional to Higher cost.

7] Data Quality & Cleaning → Big Data comes with errors, missing values, inconsistency. which affects decision making.

Q] Difference between Big Data & Small data. with processing architecture.

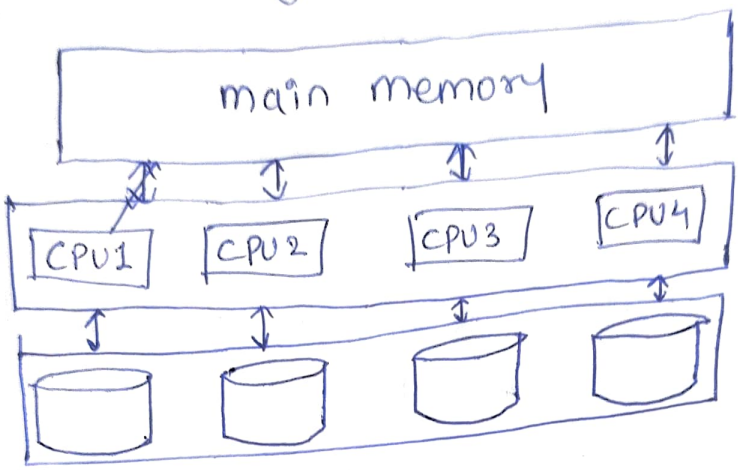
→ feature	Big Data	Small Data.
1] Size.	Very large.	Small.
2] Speed.	Processed in real time.	Slowly or manually
3] Sources	Social media, transaction etc..	Excel files, small databases
4] Processing method	Uses distributed computing (Hadoop, Spark)	Traditional processing (Excel, SQL)
5] Storage	Stored in Cloud, HDFS or NoSQL databases	Stored in local files, Spreadsheet, SQL DB
6] Complexity	Structured, semi-structured or Unstructured	Mostly Structured
7] Scalability	Highly Scalable	Limited scalability
8] Example	Netflix recommendation, fraud detection in banking	Student attendance, sale tracking

Q] Explain Big Data processing Architecture

or

Explain shared Nothing, shared everything. architecture.

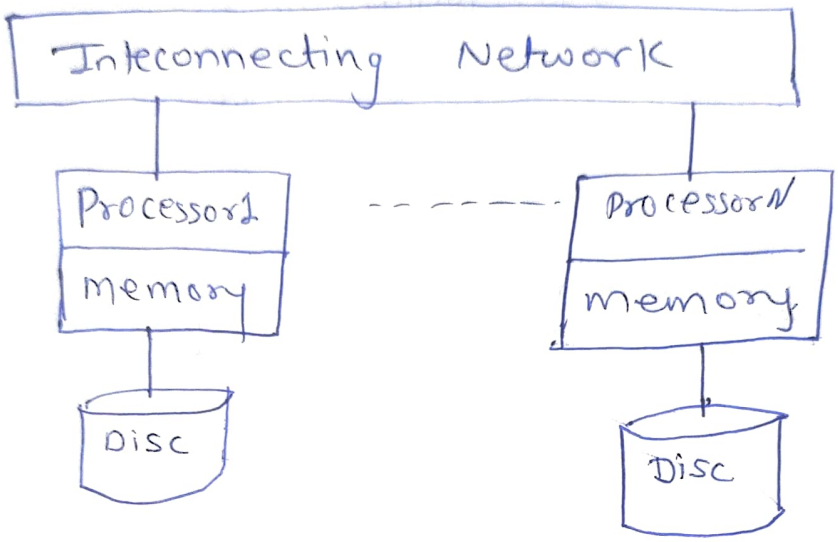
→ Shared everything Architecture



- ~~This~~ In shared every thing architecture, all server are ~~int~~ connected or all server shares the same memory. storage & has access to the store.
- Main idea behind this system is maximom resource utilization. Disadvantage is peeformance.
- Scalability is main problem
- Symentric multiprocessing & Distributed shared memory are two types.
- In Symentric multiprocessing (SMP) architecture, CPU shares a single pool of memory for read-write operations. sometimes also called as uniform Memory Access (UMA) architecture.

- Distributed shared memory (DSM) addresses the scalability problem by providing multiple pools of memory for processors to use. It is also called as Non-Uniform memory access (NUMA) Architecture

* Shared Nothing Architecture *



- In shared nothing architecture it consists of multiple nodes having ~~all~~ own OS.
- Each node is connected with other using interconnecting network
- Each node contains its own memory (M), processor (CPU) & ~~shared~~ storage device.
- Each node is in control of its own OS.
- Data is partitioned horizontally across nodes.

NIKS

Q] Difference Between Data Warehouse & Data mining

~~Definition~~ function	Data Warehouse.	Data mining.
1) Definition	A storage system that collects and organise data.	A process of analysing data to find patterns & trends.
2) Purpose	Storage large amount of data.	Extract useful insights from stored data.
3) process	Data is collected, cleaned & stored from diff resources	Data is analysed using algorithms to find hidden patterns.
4] Technique used	OLAP.	ML, Clustering, Classification
5] Tools.	Snowflake, Google Big Query	Python, R.
6] who use it?	Data engineer & analyst.	Data scientist & Ai/ML experts.
7] Example.	A bank store transaction details in datawarehouse	A bank used Data mining for fraud detection.

Q] Define the relation between Artificial Intelligence, statistical Learning & machine learning

→ AI, statistics & ML are interconnected fields, but they differ in scope & application.

AI → It is the field that focuses on Creating machine that can perform tasks that typically requires human intelligence such as reasoning, problem Solving decision making

ML → ML is a subset of AI that enables computers to learn from data without being explicitly program

Statistical Learning → It is a model that helps in predictions & relationships using statistical models

Relation

1] AI → Intelligence which include ML model to fullfill requirement

2] ML → Uses Statistical Learning technique to train model from data.

• Statistical Learning → provides model to ML Algo's

eg → fraud detection system in bank.

1] Statistical Learning → Logistic Regression analyses past fraudelent transaction

2] ML → Random forest improves accuracy

3] AI → make decision based on ML model.