Name - Shubham Khot | Roll no - 3025
class - TE(IT) T2 | sub - DSBDA
Date - 21/4/25

B 20 Mid

Date : 21/4/25

* Theory Assignment 2 *

1. What is need of big data analysis? Explain different types of analysis techniques.

=>

(1) Big Data analysis is essential for organizations to uncover valuable insights, pattern & trends from large & complex datasets.

(2) Enable them to make informed decisions. This process involves handling structured, semi-structured & un-structured data, which traditional data processing method cannot efficiently manage.

(3) Types - of Data analysis techniques -

1) **Descriptive Analysis** -

- Focus on summarizing historical data to understand what has happen in past
- It uses aggregation, data mining & visualization techniques to understood trend pattern & key performance.
- It helps you understand your current situation & make informed decisions

2) **Predictive Analysis** -

- It involve using historical / current data to find patterns & make predictions about future.

Date :

- Accuracy of prediction depends on input variable & type of model used.
- It is useful for forecasting trends, outcomes based on historical data.

3) **Realtime Analysis** -

- This refers to processing, analyzing of data as it's generated.
- Realtime analytics is useful in setting data which are generated quickly.

4) **Diagnostic analysis** -

- It typically comes after analysis, taking initial finding & inverting with certain patterns in data happen.
- It seeks to answer question "why this happen?" by taking more in depth of data to uncover subtle pattern.

Q.2. Explain data analysis life cycle in big data.

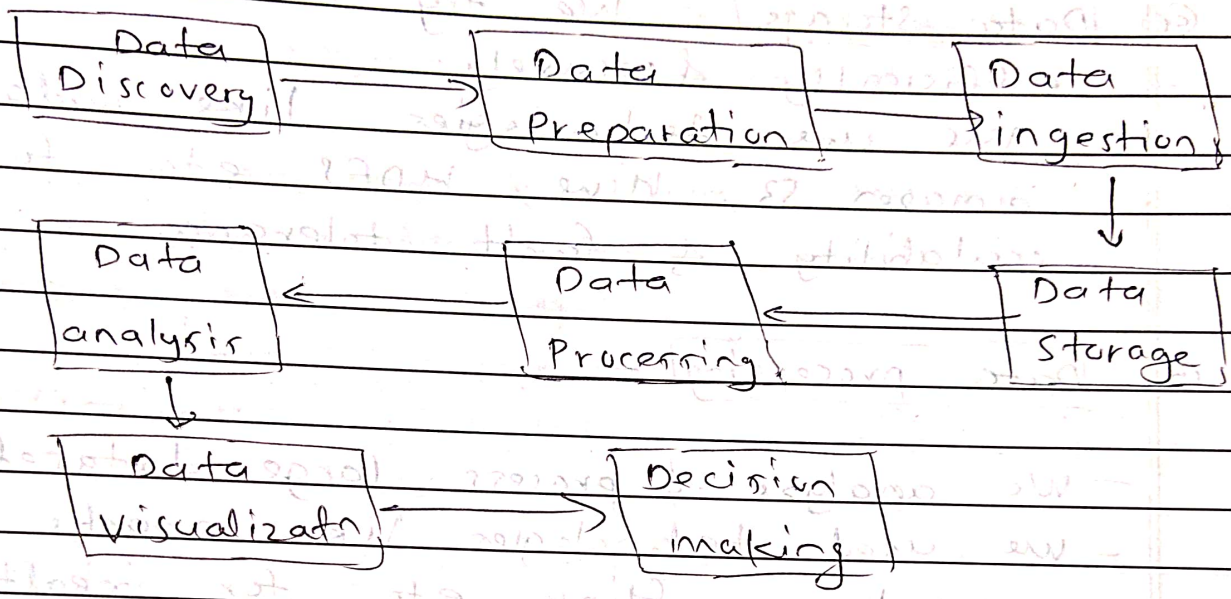
=>

1) **Data Discovery** -

- We understand data & identify relevant data sources
- Basically we define our business

Date : _____

* objectives.
- Then we try to explore available data which can be structured or unstructured.
- Assess data volume, velocity, variety, veracity & value (5 V's of big data)



* Data Analysis life cycle *

(2) Data preparation -

- In which, we try to make data usable for analysis.
- In which we do data data cleaning like handling missing values, null values.
- Data integration from multiple sources are combined, then we perform data transformat & normalization on above data.

(3) Data Ingestion - In which, we move data into storage system for processing.

Date :

- We perform batch / real-time data inge
& use tools like apache kafka, flume
- Data is stored in storages like data
lakes, HDFS, NoSQL dB etc.

④ **Data Storage** - We try to store data efficiently & securely.
- We use technologies like hadoop, amazon S3, Hive, HDFS etc. to ensu
scalability & fault tolerance

⑤ **Data processing** -
- We analyze & process large datasets.
- We used technologies like apache spark, mapreduce, flink etc. for realtime & batch processing capabilities.

⑥ **Data Analysis** - Extract meaningful insights
- And patterns from processing data.
- Here, we perform different analysis on data like descriptive, predictive.
- In which we use different machine learning models & find out different statical outcome.

⑦ **Data Visualization** - & Decision making
- It represents insight

Date : _____

- It uses tools like tableau, powerbi etc. to generate dashboards & interactive reports
- To drive buriness strategy based on insights.

3. Explain min-max scaling for following data.
$X = [24, 28, 53, 30, 40, 18, 15, 21]$

=>

- Min-max scaling is a normalization technique used to scale data within fixed range, usually 0 to 1
- Formula - $$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$$
- $X = [24, 28, 53, 30, 40, 18, 15, 21]$
- $X_{min} = 15$, $X_{max} = 53$

X	calculation formula	Scaled X
24	9/38	0.23
28	13/38	0.34
53	38/38	1
30	15/38	0.39
40	25/38	0.65
18	3/38	0.07
15	0/38	0
21	6/38	0.15

∴ Scaled value = X_{scaled} = [0.23, 0.34, 1, 0.39, 0.65, 0.07, 0, 0.15]

Date :

Q.4. Data Wrangling : Explain its need & methods.

=>

- Data Wrangling is the process of cleaning, transforming & raw data into format that is more useful for analysis.

- **Need** for Data Wrangling

1. Raw Data is messy (Real world data)
2. Different formats (CSV, JSON, XML)
3. Improve Data quality (Better insights)
4. Data Integration (Wrangling helps in combining multisource to single data)
5. Efficiency (clean data speeds up analysis).

- **Methods** of Data Wrangling.

1. **Data cleaning** -
   - Fixing or removing incorrect, corrupted or missing data.
   - Removing duplicates, correcting typos.
   - Fix spelling mistakes in customer name
2. **Data Transformation**
   - Changing format / structure of data
   - Ex- converting text dates to datetime

Date : _____

format, normalizing numeric datasets.

(3) Merging

- combining data from multiple datasets into one unified dataset
- Ex- Joining customer & order data based on CustomerID.
- Inner, outer, left/right join are used.

(4) Data Reduction

- Removing irrelevant, redundant features
- It helps in improving processing speed, focusing on only imp. features
- techniques - Feature selection, Dimensionality Reduction (PCA, LDA) Filtering out liers.