Date: 03/03/2025.

Assignment - UNIT 05 & 06.

Name - Piyusha Supe.
Roll No - 2300315.

30/30

UNIT - 05 -

Q1 What is clustering?

→ • Clustering is a type of unsupervised learning that is used to group data into clusters (groups) based on similarity.

• The objective of clustering is to ensure that objects within the same cluster are more similar to each other than to those in different clusters.

• Importance of Clustering -

1. Helps in identifying patterns in unlabeled data
2. Used for customer segmentation, image segmentation, and anomaly detection.
3. Reduces dimensionality by grouping similar data points together.

• Types of Clustering -

1. Partitioning Clustering -
   - divides data into K clusters where each data point belongs to exactly one cluster.
   - Example - K-means, K-medoids.

2. Heirarchical Clustering -

- Builds a heirarchy of clusters using either Agglomerative (bottom up) or divisive (top down) approaches.

3. Density Based Clustering -

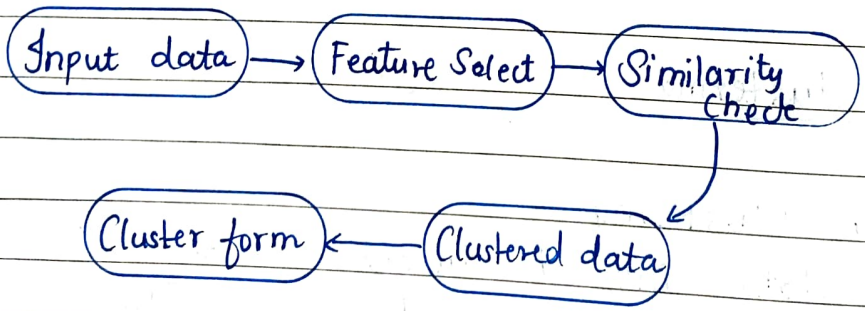
- Forms clusters based on data density and can detect arbitrarily shaped clusters.
- Eg - DBSCAN (Density based Spatial Clustering of applications with noise).

4. Grid Based clustering -

- Divides the space into a grid structure and clusters are formed within grids.
- Example - STING (Statistical Information grid).

5. Model Based Clustering -

- Assumes that data is generated from a mixture of several probabilistic distributions
- Ex - Gaussian Mixture Models.



Q2. Explain K-means clustering algorithms with use cases.

→ ∘ K-means is a partition based clustering algorithm that divides a dataset into K distinct clusters by minimizing variance within each cluster.

• Working principle of K-means -

1. Select K random points as initial centroids.
2. Assign each data point to the nearest centroid.
3. Recalculate the centroid of each cluster.
4. Repeat steps 2 and 3 until convergence (centroids stop moving).

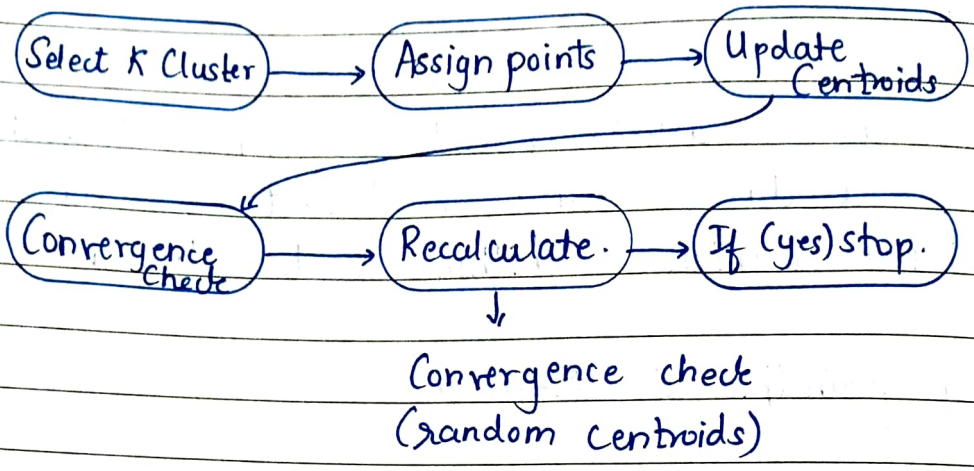
• Algorithm for K-means -

1. Input - Dataset X - containing n data points.
   Output - Number of clusters K., K clusters with their centroids.

2. Steps - (1) Select K initial centroids randomly from the dataset.
   (2) Repeat until convergence -
   - Assign each data point to nearest centroid.
   - Compute new centroid by taking the mean of all points in that cluster.
   - Stop if centroids remain same.

Block diagram:

→



- Use Cases of K-Means -

1. Customer Segmentation - E-commerce companies use K-means to group customers based on purchase behaviour.

2. Image Segmentation - K-means is widely used for object detection and image classification.

3. Anomaly detection - Used in fraud detection by identifying data points that do not belong to any cluster.

4. Document Clustering - Organizing articles or books into relevant categories.

- Hence this is how K-means is used in various cases.

Q3. How K-means algorithm works?

→ Step-by step Execution.

1. Initialize centroids - Choose K random points from the dataset as initial centroids.

2. Assign data points to clusters -
Calculate the distance of each data point from the centroids using Euclidean distance.

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

- Assign each data point to nearest centroid.

3. Update centroids -
Compute new centroid for each cluster by taking the mean of all points in that cluster.

4. Check convergence -
If centroids do not change significantly, stop, otherwise, repeat steps 2 and 3.

* Mathematical representation -

Centroid C_j of Cluster j -

$$C_j = \frac{1}{N_j} \sum_{i=1}^{N_j} X_i$$

where N_j is the number of points in cluster j.

Q4 What are drawbacks of K-means algorithm?

→ * The drawbacks of K-means algorithm is -

(1) Choosing K is difficult - No specific rule to determine the optimal value of K.

(2). Sensitive to outliers - A single outlier can drastically affect the centroid.

(3). Assumes spherical clusters - Cannot handle clusters of irregular shapes.

(4). Computational complexity - large datasets require multiple iterations making it slow.

(5). Initial centroids affect results - poor centroids initialization can lead to bad clustering.

(6). Curse of dimensionality - In high-dimensional spaces, distance metrics become less meaningful.

* Solutions to these drawbacks -

- Using the Elbow method - to determine the optimal K.
- Using K-medoids instead of K-means - to reduce sensitivity to outliers.
- Using advanced methods like DBSCAN for non-spherical clusters.

Q5 Explain UEX 5.2.1.

→ 1. UEX 5.2.1 in Clustering and Machine Learning.

- User experience (UEX) Clustering -
  - Groups users based on behavioral patterns on a website or app.
  - Uses clustering techniques like K-means, DBSCAN or Heirarchical Clustering to group users based on clicks, time spent and navigation flow.
  - Helps in personalizing user interfaces.
- Segmenting users based on Feature Interactions. - UEX clustering could help group users into catagories such as -
  - 1. Frequent useres - Engage with most features.
  - 2. New users - Have limited interactions.
  - 3. Inactive users - Rarely interact with the system.

2. UEX 5.2.1 as a UX standard in Software engineering-

1. User interface clustering - Grouping UI components for better accessibility.
2. User testing methods - Evaluating user experience with machine learning.
3. Adaptive UI clustering - Interfaces that change dynamically based on user clusters.

3. UEX 5.2.1 in database and Big data processing

- Optimized query clustering - using clustering algorithms to group similar queries.
- Log-based clustering - Grouping logs for anomaly detection in cloud applications.

Example -

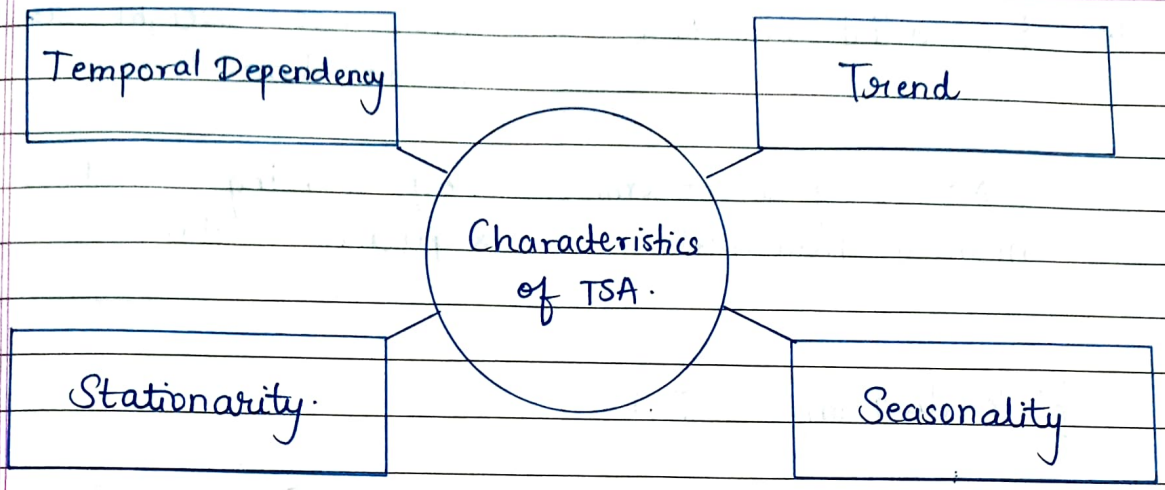
1. Input - Website user logs (clicks, session duration, navigation).
2. Feature Extraction - Click stream data, mouse movement.
3. Clustering Algorithm - Apply K-means or DBSCAN.
4. Output - User personas based on behavioral clusters.

Q6. What is time series analysis?

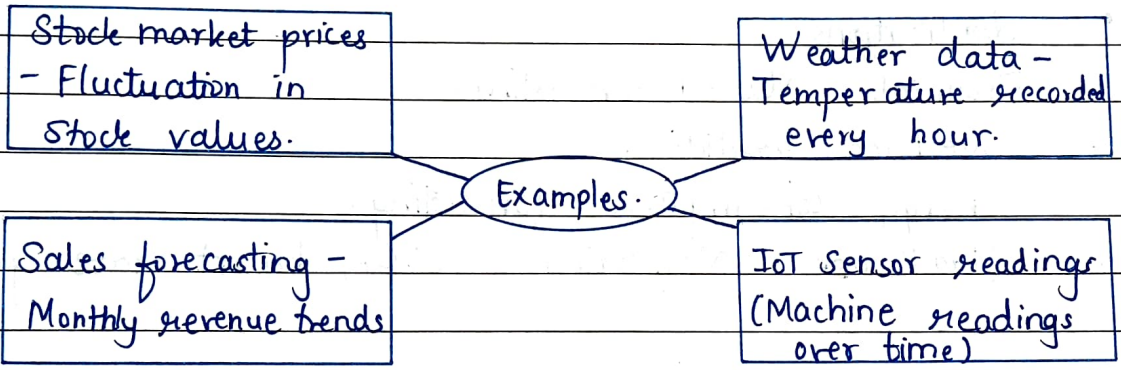
→. Time Series analysis (TSA) is a statistical technique used to analyze data points collected over time at regular intervals. It helps in identifying trends, patterns and seasonal variations of data to make predictions.

- Characteristics of Time series data -

1. Data is collected at different timestamps.
2. Long term increase or decrease in data values
3. Recurring patterns at fixed intervals (eg. daily, monthly)
4. Statistical properties like mean and variance remain constant.



- Examples of Time series data -

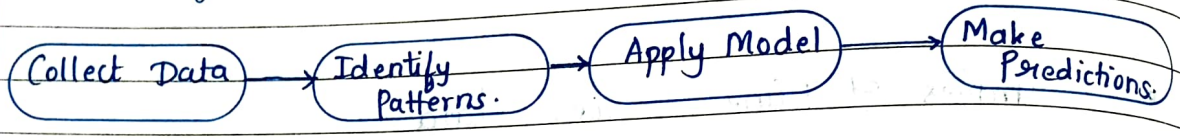


- Time Series Analysis Methods -

Method.	Description.
1. Moving Average (MA)	Smoothens data by averaging past values.
2. Auto regressive (AR)	Predicts future values using past observations.
3. ARIMA (Autoregressive Integrated Moving Average)	Combines AR and MA for better forecasting.

4.	Exponential Smoothing.	Gives more weight to recent data.
5.	LSTMs (Long Short term memory networks)	Deep learning method for sequential data.

- Block diagram - TSA.



- Applications -

1. Financial Market forecasting.
2. Weather prediction.
3. Energy consumption forecasting.
4. Healthcare Monitoring (ECG, heart rate patterns)

Q7. Why social network analysis is important?

→ 1. Social Network analysis (SNA) is a method used to analyze relationships between individuals, groups, or organizations. It helps in understanding interactions and influence in networks.

2. Importance of SNA -

a. Detects influential nodes (Identifies key people in a network)

b. Predicts Trends and Virality - (How information spreads)

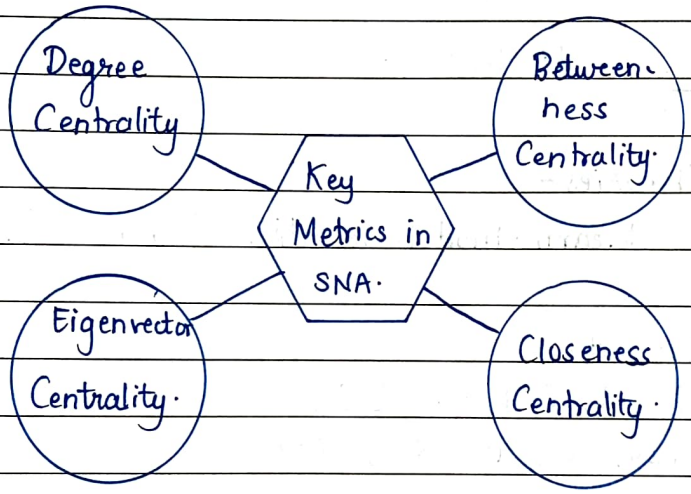
c. Detects fake accounts or Bots - (In Cybersecurity)

d. Improves Marketting strategies (Targetting right audience)

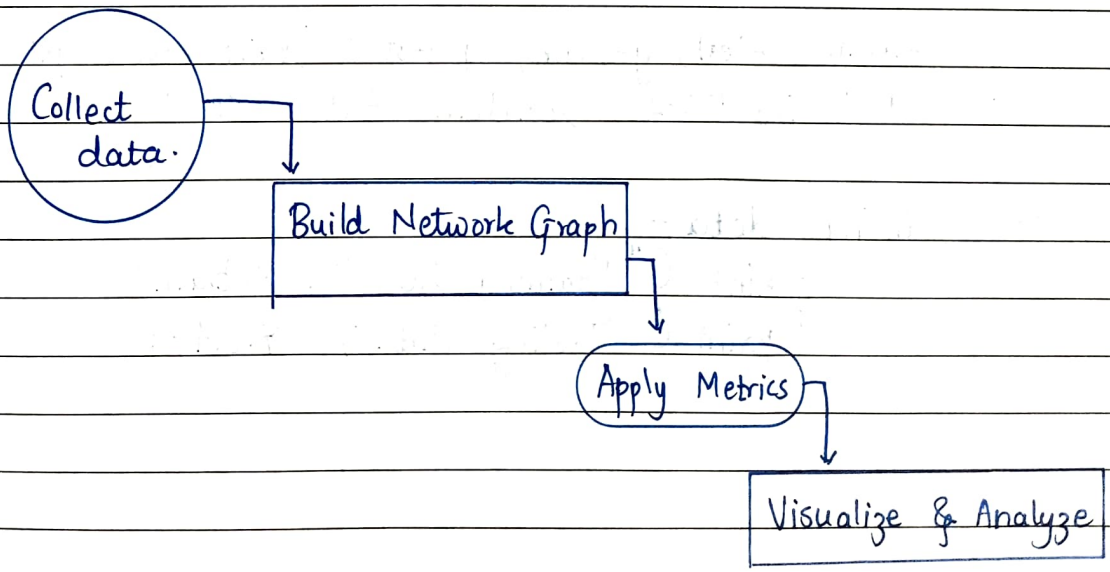
e. Analyzes community structures (Detects groups with similar interests).

---

3. Key metrics in SNA -



4. Block diagram - SNA :



98. What is the holdout method?

→ 1. Introduction

The Holddout method is a technique used in machine learning model evaluation. It involves Splitting dataset into two parts.

(1) Training set (Usually 70% to 80%) – Used to train the model.

(2) Testing set (Usually 20% to 30%) – Used to evaluate performance.

2. Algorithm for Holdout –

(1.) Import libraries –

from sklearn.model_selection import train_test_split.

(2). Sample dataset.

X = [[1], [2], [3], [4], [5], [6]].

Y = [10, 20, 30, 40, 50, 60].

(3) Split data (80% Train ; 20% Test).

X_train, X_test, y_train, y_test = train_test_split (X, y, test_size = 0.2, random_state = 42)

(4). Printing data –

print ("Training data", X_train)

print ("Testing data", X_test)

3. Pros and Cons -

Pros

(1) Simple and easy to implement.
(2). Faster compared to cross-validation.

Cons.

(1) Performance depends on split randomness.
(2). Less effective for small datasets.

Q9. Explain random sub-sampling -

→ 1. It is a technique used for evaluating models or handling large datasets. It involves repeatedly selecting random subsets (samples) from orignal dataset to perform tasks such as validation, training or statistical analysis.

2. Key aspects of random subsampling -

(1). Random selection - A portion of dataset is selected randomly multiple times.

(2). Multiple iterations - The process is repeated several times to reduce bias and increase robustness.

(3). No gurantee of unique data - Some data points may appear in multiple subsets while others might not appear at all.

(4). Flexible sample size - The size of each subset can be adjusted based on the use case.

3. Use cases of in data science and Big data -

(1). Model Evaluation (Cross validation Alternative) -

- Random sub sampling is often used as a simple alternative to K-fold cross validation.
- Instead of dividing data into K-folds., the dataset is randomly split into training and validation sets multiple times.

(2). Handling large Datasets -

- When working with massive datasets, it is computationally expensive to use the entire data.
- Instead, random subsampling allows analysis on a smaller representative portion, reducing processing time while maintaining statistical significance.

(3). Statistical Estimation and Hypothesis testing -

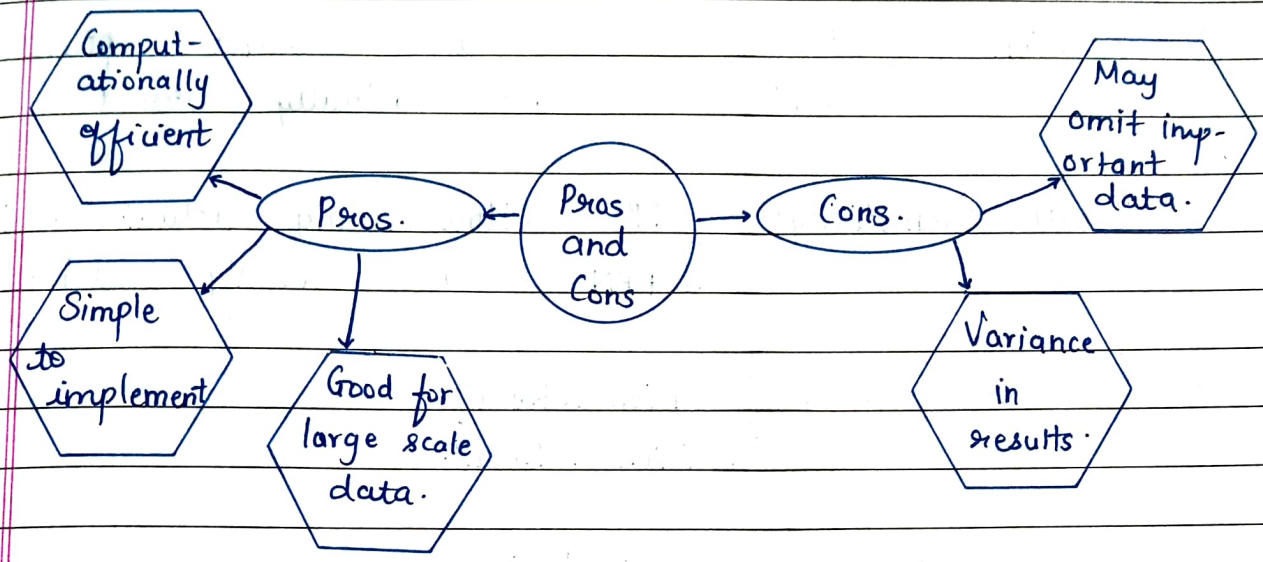
- In inferential statistics, random subsampling is used to estimate population parameters.

4. Advantages.

- Computationally efficient - Reduces dataset size, making computation faster.
- Simple to implement.
- Good for large scale data.

5. Disadvantages -

1. May omit important data points.
2. Variance in results.



Q10. What is confusion matrix ? Explain with suitable example.

→ 1. Confusion matrix is a table used to evaluate the performance of a classification model. It compares actual labels vs. predicted labels.

2. Confusion matrix structure -

Actual / Predicted.	Positive (1).	Negative (0).
Positive (1)	TP (True Positive)	FN (False Negative)
Negative (0)	FP (False positive)	TN (True Negative

3. Key metrics derived from confusion matrix -

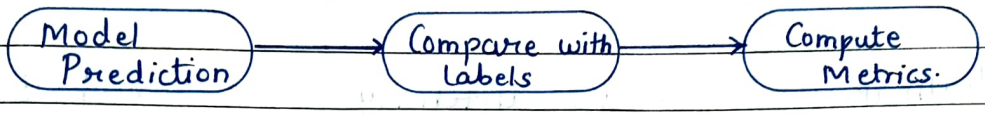
- Accuracy = $\frac{TP + TN}{TP + TN + FP + FN}$
- Precision = $\frac{TP}{TP + FP}$ (How many predicted positives are actually correct)
- Recall (Sensitivity) = $\frac{TP}{TP + FN}$ (How many actual positives are correctly identified?)
- F1 Score = $2 \times \frac{Precision \times Recall}{Precision + Recall}$

4. Example of Confusion Matrix -
Suppose a model classifies spam emails (1 = Spam, 0 = Not Spam).

Actual/Predicted	Spam (1)	Not Spam (0).
Spam (1)	50 (TP)	10 (FN)
Not Spam (0)	5 (FP)	100 (TN).

- Accuracy = $\frac{50 + 100}{50 + 100 + 10 + 5}$ = 90.9 %.
- Precision = $\frac{50}{50 + 5}$ = 90.9 %.
- Recall = $\frac{50}{50 + 10}$ = 88.3 %.

5. Block diagram -



UNIT - VI

Q1 What is data visualization?

→ 1. Data visualization is the graphical representation of information and data. It helps in understanding trends, patterns and insights in a visually engaging manner.

2. Importance of Data Visualization -

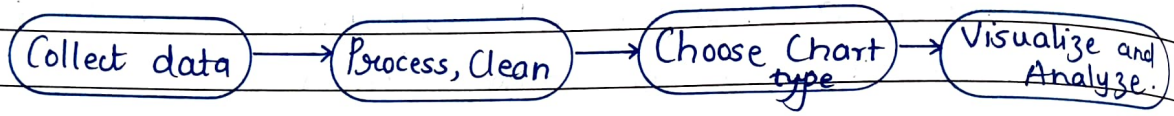
- Converts complex data into intuitive graphics.
- Helps in identifying trends, correlations and outliers.
- Aids in better decision making.
- Improves data storytelling.

3. Types of Data Visualization -

Type	Description	Example.
Bar chart.	Compares categorical data.	Sales per region
Line graph.	Shows trends over time	Stock market trends.

Pie chart.	Represents proportions.	Market share distribution.
Heatmaps.	Shows intensity variations.	Website users engagement.
Scatter plot.	Identifies relationships between variables.	Exam scores vs study time

4. Block diagram -



5. Example in python - line graph -

```
import matplotlib.pyplot as plt.
years = [2018, 2019, 2020, 2021, 2022]
sales = [500, 600, 700, 850, 900]

plt.plot (years, sales, marker='o', linestyle='-',
          color='b')
plt.xlabel ("Year").
plt.ylabel ("Sales ($)")
plt.show().
```

Hence, this is how data visualization is important part of data science.

Q2. What are challenges of data visualization?

→ * The challenges are given below -

1. Choosing the right visualization - Using incorrect chart types leads to misinterpretation.
2. Handling large datasets - Big data requires advanced tools like Hadoop and Spark.
3. Data cleaning issues - Missing or incorrect data affects accuracy.
4. Overloading with information - Too much data in one graph makes it confusing.
5. Bias in data representation - Incorrect sales or misleading graphs cause misinterpretation.

* Example -

Misleading data visualization -

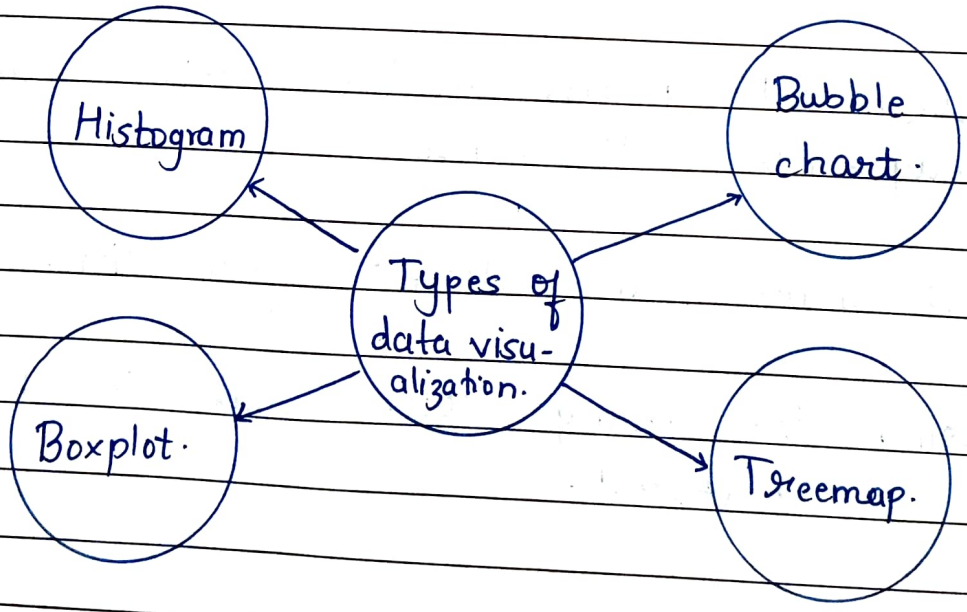
- A truncated y-axis in bar charts can exaggerate differences.
- 3D charts can distort perception.
- Using too many colors makes it hard to interpret.

Q3. Explain data visualization techniques?

→ 1. Data visualization techniques are used to represent data effectively. The choice of technique depends on the nature of data and purpose.

2. Types of data visualization Techniques -

Technique.	Description.	Best for.
Histogram	Shows distribution of data.	Exam scores distribution.
Boxplot.	Displays data spread and outliers.	Comparing salary distribution.
Treemap.	Shows Heirarchical data.	Portfolio diversification.
Bubble chart.	Adds extra dimension with bubble size.	Sales vs profit analysis.



3. Example python for Histogram -

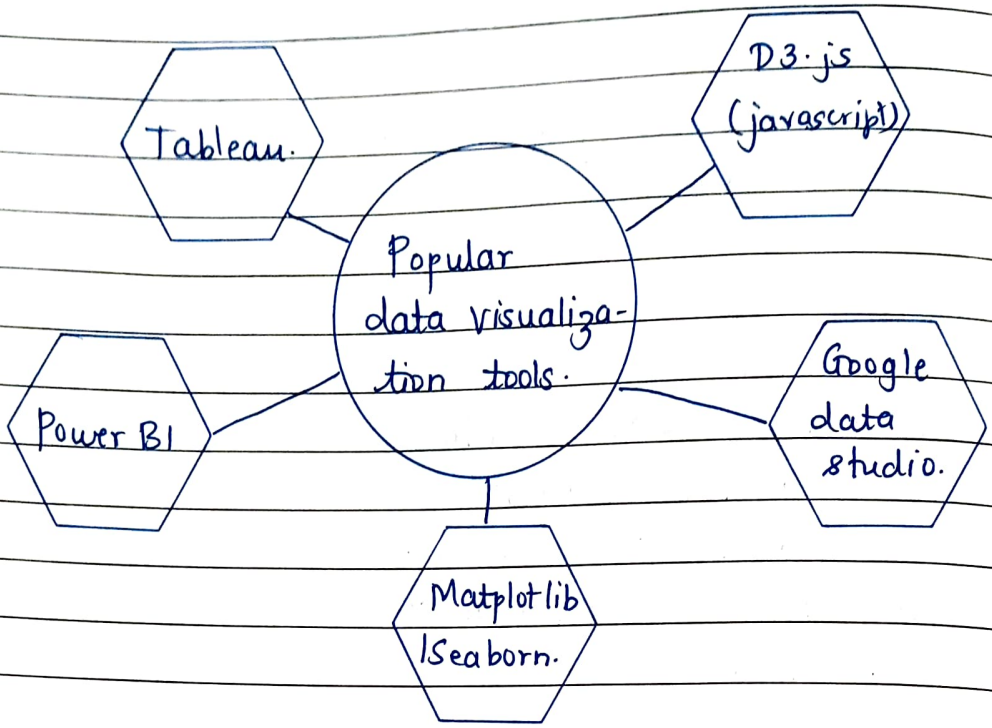
```
import matplotlib.pyplot as plt.
import numpy as np.

data = np.random.randn(1000)
# Histogram.
plt.hist(data, bins=30, color='blue', alpha=0.7)
plt.xlabel("Values")
plt.ylabel("Frequency")
plt.title("Histogram").
plt.show().
```

Q4. What are data visualization tools?

→ 1. Popular data visualization tools -

Tool	Description.	Use Case.
Tableau.	Advanced Analytics and dashboarding.	Business Intelligence.
Power BI.	Microsofts interactive visualization tool.	Enterprise reporting.
Matplotlib/ Seaborn.	Python libraries for plotting.	Data science and AI.
Google data Studios.	Free tool for google data integration.	Marketting Analytics.
D3.js.	Javascript based visualization	Interactive web Dashboards.



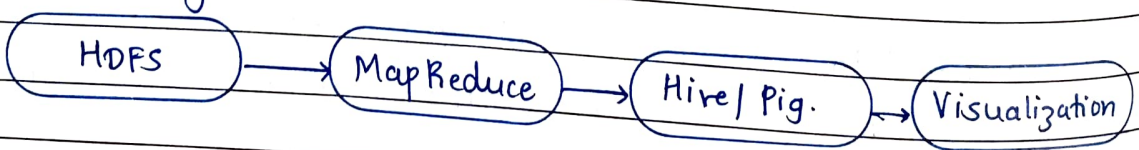
2. Example - Power BI dashboard components -

- Data Sources - (Excel, SQL, API)
- Data Modeling (Transform, clean, filter)
- Visualization (Charts, KPIs)
- Publishing and Sharing (Reports, Dashboards).

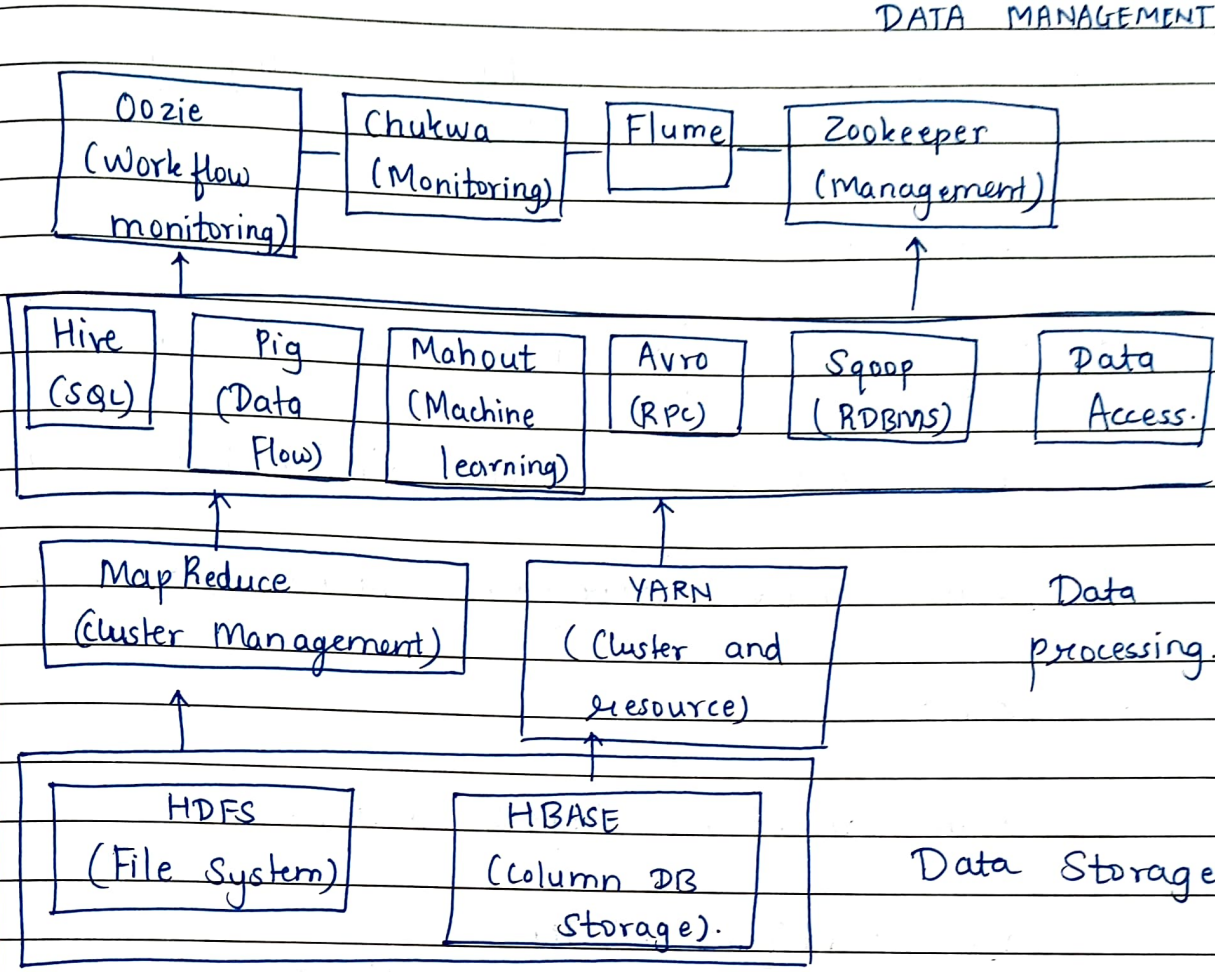
Q5. Explain the Hadoop Ecosystem in detail with (Pig, hive, HBase, Mahout)

→ 1. Hadoop is an open source big data framework that provides distributed storage (HDFS) and parallel processing (MapReduce).

2. Block diagram -



3. HADOOP ECOSYSTEM.



4. Components of Hadoop Ecosystem -

(1). HDFS (Distributed file System) - Stores large datasets across clusters.

(2). Mapreduce - Processes data in parallel.

(3). Pig - High level scripting language for big data.

(4). Hive - SQL like querying on Hadoop.

(5). HBase - NoSQL database for real time data.

(6). Mahout - Machine learning on Hadoop.

Q6. Explain mapreduce paradigm with example –

→ 1. Mapreduce is a parallel processing model used for large scale data processing in Hadoop.

2. Mapreduce Workflow –

- Map phase – Splits data into key value pairs.
- Shuffle and sort – Groups similar keys together.
- Reduce phase – Aggregates values for each key.

3. Example –

```
from mrjob.job import MRjob.

class WordCount (MR Job):
    def mapper (self, –, line):
        for word in line.split():
            yeild word, 1.

    def reducer (self, word, counts):
        yeild word, sum(counts)

if –name– == "– – main – –":
    WordCount.run()
```

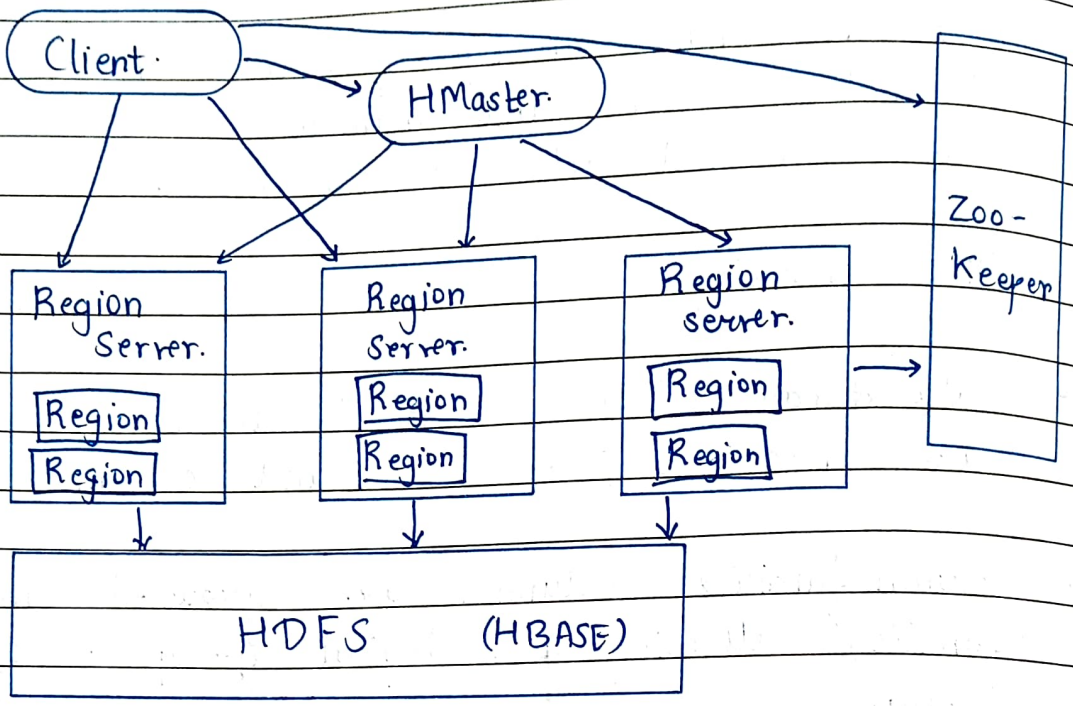
Q7 Explain HBase -

→ 1. Apache HBase is a distributed, scalable, NoSQL data base built on top of Apache Hadoop, designed to store and manage large amounts of data, modeled after google's bigtable.

2. HBase working -

- Data storage - Data is stored in tables, which are organized into rows and columns.
- Column families - Columns are grouped into column families allowing for logical grouping of related data.
- Row Keys: Each row is identified by a unique row key, which is used to locate the data.
- Data distribution - data is distributed across multiple servers in a cluster enabling horizontal scaling.
- Write ahead logging (WAL) - HBase uses a write-ahead log to ensure data durability and consistency.
- Memstore and HFile: Data is initially written to a memory based store (Memstore) and then flushed to disk as HFiles (HBase files).

3. Architecture of HBase -



Q8. What is Analytical technique use. in big data Visualizations?

→ 1. Big data visualization requires advanced analytical techniques for better insights.

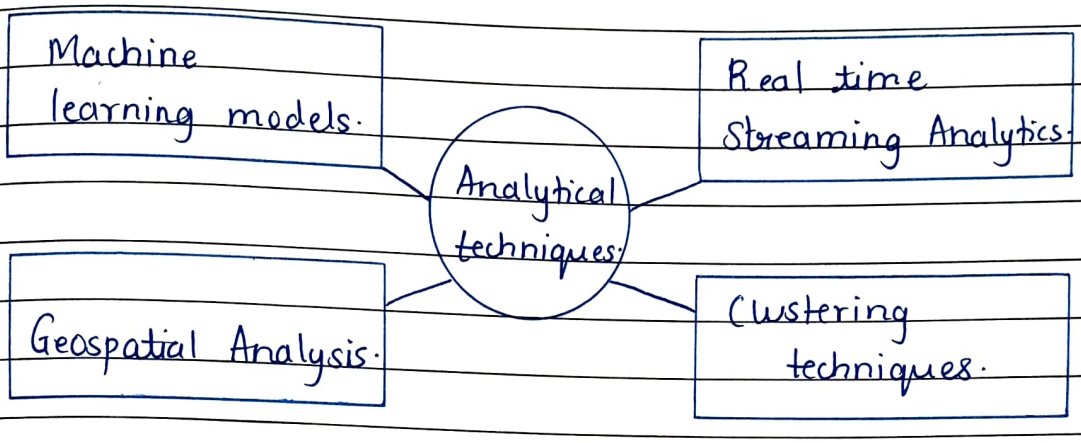
2. Techniques used -

(1). Machine learning models - Predict trends and anomalies.

(2). Real time streaming analytics - Visualize IoT and social media data.

(3) Clustering techniques - Identify patterns in unstructured data.

(4). Geospatial analysis - Visualizing geographic data



3. More techniques include -

(1). Descriptive analytics - focuses on summarizing past data. , historical events.

(2). Diagnostic Analytics - Delves deeper to uncover factors and variables that contributed to specific outcome.

(3). Predictive analytics - This aims to forecast future trends and outcomes by analyzing historical data.

(4). Prescriptive analytics - This goes beyond prediction to recommend actions and strategies to optimize outcomes and acheive desired results.

(5). Time series analysis - Analyses data collected over time to identify trends, seasonality and patterns. enabling forecasting and understanding underlying patterns.

* * * * * * *