DSBDA UNIT – 6 PYQ's

➢ *MAY / JUN 2022*

Q7)

a) With a suitable example explain Histogram and explain its usages. [8]

A **Histogram** is a type of bar chart that represents the distribution of numerical data by grouping values into ranges (called bins). It shows the frequency of data within each bin.

Characteristics of a Histogram :

- Data is grouped into continuous intervals.
- No gaps between bars (unlike bar charts).
- Useful for understanding the **distribution**, **skewness**, **spread**, and **outliers** in the data.

Steps to Draw a Histogram

Histogram is the basic tool of representing data, and we can easily draw a histogram by following the steps added below:

Step 1*:* Collect the data you wish to display in the histogram. This might range from test results to population distribution.

- For example: Assume you get the following test scores: 14, 20, 12, 26, 8, 7, 2, 28, 30, 16, 18, 23.
- First arrange it in ascending order.
  Exam results: 2, 7, 8, 12, 14, 16, 18, 19, 23, 26, 28 and 30.

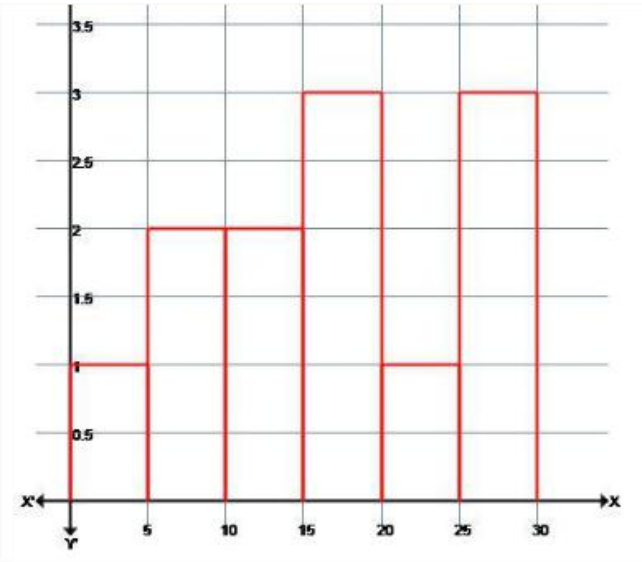
Step 2: Determine the number of intervals, or "bins," you wish to split your data into. This is determined by the scope and distribution of your data, as well as the amount of information you choose to display. Assume we wish to divide the scores into 5 bins.

Step 3: Determine the limits of each bin. These bounds should encompass the complete range of your data and be regularly spaced.
[0-5 - 10 - 15 - 20 - 25 - 30].

Step 4: Count the number of data points that belong in each bin.

Class Interval	***Frequency***
0-5	*1*
5-10	*2*
10-15	*2*
15-20	*3*
20-25	*1*
25-30	*3*

Step 5: On a graph, show the bin borders on the x-axis and the frequency of data points in each bin on the y-axis.
Create bars for each bin, with the height of each bar representing the frequency of data points in that bin.



In this histogram, the x-axis depicts the bins, while the y-axis indicates the frequency of data points falling within each bin. The bars represent the sample data's distribution across the given bins.

Use Case	Description
1. Data Distribution	Understand how data points are spread over an interval.
2. Detect Skewness	Helps identify whether the data is symmetric or skewed.
3. Outlier Detection	Bins with very low frequency or isolated bars may indicate outliers.
4. Compare Features	Compare the distributions of different features in a dataset.
5. Preprocessing Insight	Guides decisions like normalization or binning before ML modeling.

b) Describe the Data Visualization Tool "Tableau". Explain its applications in brief. [9]

Tableau is a powerful and widely used data visualization tool that helps users convert raw data into interactive and shareable dashboards, charts, and graphs. It is widely adopted for Business Intelligence (BI) and Data Analytics tasks.

✓ **Key Features of Tableau :**

- **Drag-and-Drop Interface**: Easy to use with minimal coding required.
- **Data Connectivity**: Connects to various data sources (Excel, SQL, cloud databases, etc.).
- **Real-Time Data Analysis**: Supports live and in-memory data analysis.
- **Interactive Dashboards**: Allows filters, actions, and storytelling.
- **Mobile Support**: Dashboards can be optimized for mobile devices.
- **Integration**: Supports integration with R, Python, and other tools.

✓ **Components of Tableau :**

Component	Description
Tableau Desktop	Used to create dashboards and visualizations.
Tableau Public	Free version for sharing dashboards publicly online.
Tableau Server	Allows sharing and collaboration within an organization.
Tableau Online	Cloud-based version of Tableau Server.
Tableau Prep	Used for data cleaning and preparation before visualization.

✓ **Applications of Tableau :**

Area	Application Example
Business Intelligence	Visualizing KPIs, financial metrics, and sales trends.
Healthcare	Tracking patient data, hospital performance, and disease trends.
Education	Analyzing student performance, attendance, and faculty efficiency.
Government	Displaying demographic data, election analysis, and public policies.
Marketing	Analyzing campaign performance, customer segmentation, and ROI.
E-commerce	Tracking customer behavior, order trends, and inventory levels.

✓ **Advantages of Using Tableau**

- Easy to learn and user-friendly
- Powerful visual analytics and storytelling
- Suitable for both technical and non-technical users
- Strong community and enterprise support

Q8)

a) With a suitable example explain and draw a Box plot and explain its usages. [8]

A Box Plot (or Box-and-Whisker Plot) is a graphical representation of data that shows its central tendency, spread, and potential outliers using five summary statistics:

- Minimum
- First Quartile (Q1)
- Median (Q2)
- Third Quartile (Q3)
- Maximum

✓ **Characteristics of a Box Plot**

- The box represents the **interquartile range (IQR)** = Q3 − Q1.
- A vertical line inside the box shows the **median** (Q2).
- **Whiskers** extend to the minimum and maximum values that are not outliers.
- **Outliers** (if any) are plotted as individual points beyond whiskers.

How to create a box plots?

Let us take a sample data to understand how to create a box plot.

Here are the runs scored by a cricket team in a league of 12 matches - ***100, 120, 110, 150, 110, 140, 130, 170, 120, 220, 140, 110.***

To draw a box plot for the given data first we need to arrange the data in ascending order and then find the minimum, first quartile, median, third quartile and the maximum.

```
Ascending Order
100, 110, 110, 110, 120, 120, 130, 140, 140, 150, 170, 220

Median (Q2) = (120+130)/2 = 125; Since there were even values
```

To find the First Quartile we take the first six values and find their median.

```
Q1 = (110+110)/2 = 110
```

For the Third Quartile, we take the next six and find their median.

```
Q3 = (140+150)/2 = 145
```

Note: If the total number of values is odd then we exclude the Median while calculating Q1 and Q3. Here since there were two central values we included them. Now, we need to calculate the Inter Quartile Range.

```
IQR = Q3-Q1 = 145-110 = 35
```

We can now calculate the Upper and Lower Limits to find the minimum and maximum values and also the outliers if any

```
Lower Limit = Q1-1.5*IQR = 110-1.5*35 = 57.5
Upper Limit = Q3+1.5*IQR = 145+1.5*35 = 197.5
```

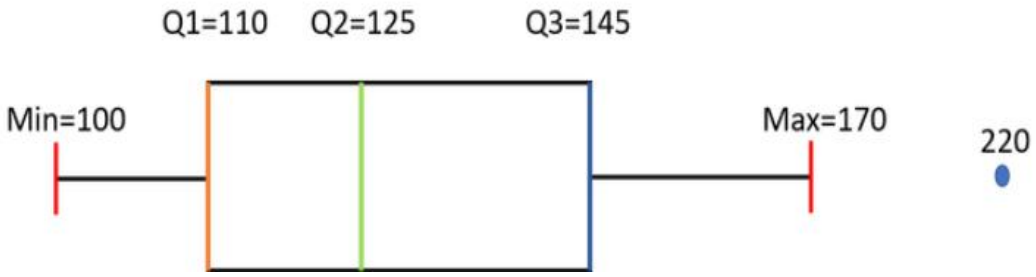
So, the minimum and maximum between the range [57.5,197.5] for our given data are –

```
Minimum = 100
Maximum = 170
```

The outliers which are outside this range are –

```
Outliers = 220
```

we can draw the box plot which is as below-



Usages of Box Plot

- Compare distributions between different groups
- skewness of the data (if median is not centered)
- Detect outliers easily
- Summarize large data sets in compact visual form
- Commonly used in exploratory data analysis (EDA)

Q8

b) Describe the challenges of data visualization. Draw box plot and explain its usages. [9]

Challenges in Big Data Visualization

1. **Volume:** Visualizing immense datasets is difficult due to sheer size and complexity.
2. **Variety:** Combining structured, semi-structured, and unstructured data sources into a coherent visual is challenging.
3. **Velocity:** Real-time data streams require fast processing and updating of visuals.
4. **Visual Noise:** High density of data points makes it hard to distinguish individual objects.
5. **Information Loss:** Data reduction techniques may simplify visuals but risk losing important details.
6. **Large Image Perception:** Physical and device constraints limit how much data can be effectively displayed.
7. **High Rate of Change:** Rapid data updates can overwhelm users' ability to track changes.
8. **Performance Requirements:** High computational power is needed to generate and update visuals quickly.

----------**DRAW BOX PLOT WALA IS COVERED IN PREVIOUS QUESTION Q8 a) !!** ----------

---

➢ NOV / DEC 2022

Q7

a) List the data visualization tools and discuss any four applications of data visualization along with the use of the suitable plot. [9]

Data Visualization Tools:

1. **Tableau –** Drag-and-drop interface for interactive dashboards.
2. **Power BI** – Best for business reporting and Microsoft integration**.**
3. **Microsoft Excel –** Simple and widely used for charts and pivot tables**.**
4. **Google Data Studio –** Great for free, web-based dashboarding.
5. **Python (Matplotlib, Seaborn) –** Used for custom, code-based plotting.
6. **R (ggplot2) –** Statistical data visualization in research.

Applications of Data Visualization (with suitable plots):

1. **Business Sales Analysis :**
    - **Application:** Track sales performance by region or product.
    - **Plot Used: Bar Chart** to compare sales across regions or time periods.
    - **Use:** Helps decision-makers identify high-performing areas and underperformers.
2. **Website Traffic Monitoring :**
    - **Application:** Analyze web traffic trends over time.
    - **Plot Used: Line Chart** to show user visits over days/weeks.
    - **Use:** Reveals traffic spikes, seasonal trends, or impact of promotions.
3. **Customer Feedback and Ratings:**
    - **Application:** Analyze sentiment or satisfaction levels.
    - **Plot Used: Pie Chart** or **Donut Chart** to show rating distribution.
    - **Use:** Helps understand customer satisfaction and areas needing improvement.
4. **Outlier Detection in Finance or Healthcare :**

- **Application:** Detect abnormal spending or unusual health metrics.
- **Plot Used: Box Plot** to identify outliers in numerical data.
- **Use:** Useful for fraud detection or identifying patients needing attention.

Thus, data visualization helps simplify decision-making by presenting complex data in an easy-to-understand graphical format.

b) List the challenges of data visualization explain the types of visualization with example.[9]

Challenges in Big Data Visualization

1. **Volume:** Visualizing immense datasets is difficult due to sheer size and complexity.
2. **Variety:** Combining structured, semi-structured, and unstructured data sources into a coherent visual is challenging.
3. **Velocity:** Real-time data streams require fast processing and updating of visuals.
4. **Visual Noise:** High density of data points makes it hard to distinguish individual objects.
5. **Information Loss:** Data reduction techniques may simplify visuals but risk losing important details.
6. **Large Image Perception:** Physical and device constraints limit how much data can be effectively displayed.
7. **High Rate of Change:** Rapid data updates can overwhelm users' ability to track changes.

Types of Data Visualization with Examples :

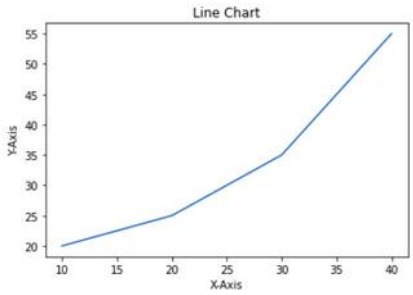
1. Line Chart

Line chart is one of the basic plots and can be created using the plot() function. It is used to represent a relationship between two data X and Y on a different axis.

Example:

```
import matplotlib.pyplot as plt

x = [10, 20, 30, 40]
y = [20, 25, 35, 55]

plt.plot(x, y)

plt.title("Line Chart")

plt.ylabel('Y-Axis')

plt.xlabel('X-Axis')
plt.show()
```

Output:



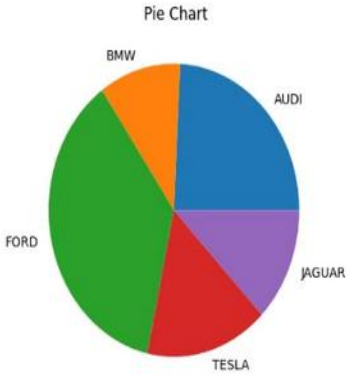
2.Pie Chart

Pie chart is a circular chart used to display only one series of data. The area of slices of the pie represents the percentage of the parts of the data. The slices of pie are called **wedges**. It can be created using the **pie()** method.

Example:

```
import matplotlib.pyplot as plt
import pandas as pd

data = pd.read_csv('/content/tip.csv')

cars = ['AUDI', 'BMW', 'FORD',
        'TESLA', 'JAGUAR',]
data = [23, 10, 35, 15, 12]

plt.pie(data, labels=cars)

plt.title(" Pie Chart")

plt.show()
```

Output:



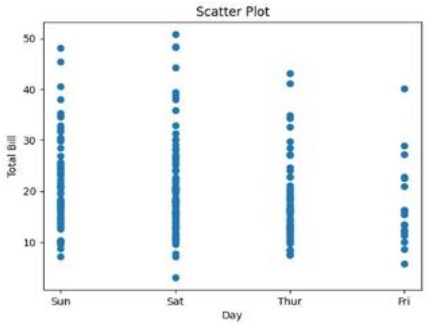
3. Scatter Plot

Scatter plots are used to observe relationships between variables. The **scatter()** method in the matplotlib library is used to draw a scatter plot.

Example:

```
import matplotlib.pyplot as plt
import pandas as pd

data = pd.read_csv('/content/tip.csv')

x = data['day']
y = data['total_bill']

plt.scatter(x, y)

plt.title("Scatter Plot")

plt.ylabel('Total Bill')

plt.xlabel('Day')

plt.show()
```

Output:



ALTERNATIVE (FOR SADA – SIMPLE ANSWER) :

Types of Data Visualization with Examples

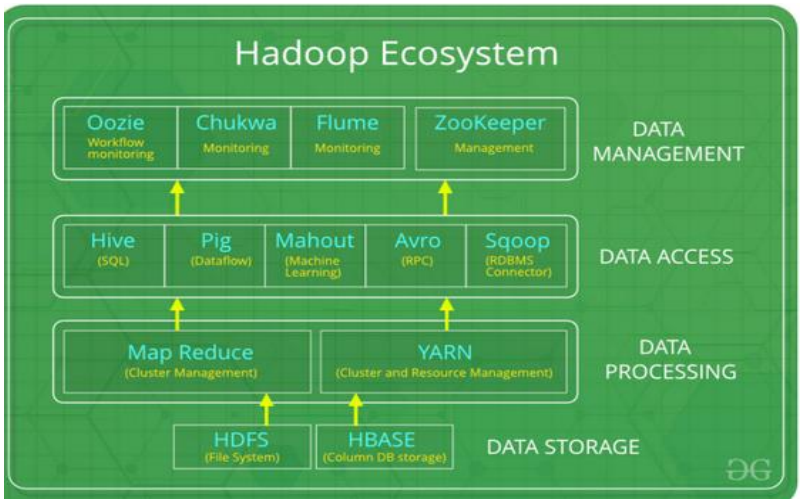
1. **Bar Chart**
   - **Purpose:** Compare quantities across categories.
   - **Example:** Sales revenue by product category.
2. **Line Chart**
   - **Purpose:** Show trends over time.
   - **Example:** Monthly website visitors over a year.
3. **Pie Chart**
   - **Purpose:** Show percentage or proportional data.
   - **Example:** Market share distribution among companies.
4. **Scatter Plot**
   - **Purpose:** Show relationship between two variables.
   - **Example:** Plotting advertising spend vs. sales.
5. **Histogram**
   - **Purpose:** Show distribution of numerical data.
   - **Example:** Frequency of customer ages.

Q8)

a) Explain in detail the Hadoop Ecosystem with suitable diagram [9]

Hadoop Ecosystem

The Hadoop Ecosystem is a suite of components built around the Hadoop platform to store, manage, process, and analyze large-scale data efficiently. It includes several open-source tools that work together to solve big data problems.



Main Layers & Components :

1. Data Storage

- **HDFS**: Hadoop Distributed File System – stores large data across multiple machines.
- **HBase**: Column-oriented NoSQL database on top of HDFS.

2. Data Processing

- **MapReduce**: Programming model for batch processing.
- **YARN**: Resource manager that schedules and manages clusters.

3. Data Access

- **Hive**: SQL-like query language for data warehousing.
- **Pig**: High-level platform for creating MapReduce programs.
- **Mahout**: Library for scalable machine learning.
- **Avro**: Framework for data serialization.
- **Sqoop**: Transfers data between RDBMS and Hadoop.

4. Data Management & Coordination

- **Oozie**: Workflow scheduler.

- **Chukwa**: Data collection system for monitoring.
- **Flume**: Ingests large volumes of log data.
- **ZooKeeper**: Maintains configuration and synchronization.

The Hadoop Ecosystem enables scalable and fault-tolerant data storage and processing. Each tool has a specific role, and together they form a powerful framework for managing big data.

b) Write a short note on the following [9]

i) MapReduce
ii) Pig
iii) Hive

i) MapReduce :

MapReduce is a programming model in the Hadoop ecosystem used for processing large-scale data in parallel across distributed clusters.

- It works in two phases: **Map phase** (filters and sorts data) and **Reduce phase** (aggregates and summarizes results).
- It ensures fault tolerance, scalability, and can handle structured and unstructured data efficiently.
  Example: Counting word frequency in a large text file using mappers and reducers.

ii) Pig

Pig is a high-level scripting platform that simplifies writing MapReduce programs.

- Uses a scripting language called **Pig Latin** for data transformation, aggregation, and analysis.
- Suitable for semi-structured data and integrates with HDFS.
- Automatically converts Pig Latin scripts into MapReduce jobs.
  Example: Loading data, filtering rows, and grouping data by fields in a simple script.

iii) Hive

Hive is a data warehousing tool built on top of Hadoop for querying and managing large datasets using **HiveQL (SQL-like language)**.

- Ideal for users familiar with SQL but not Java or MapReduce.
- Converts SQL queries into MapReduce or Tez/Spark jobs internally.

- Supports partitioning, indexing, and schema evolution.
  Example: Running SQL-like queries on logs stored in HDFS.

➢ MAY / JUN 2023

Q7)

a) With a suitable example, draw a Histogram, boxplot and explain its usages. [9]

→ REPEATED !

b) Describe the data visualization tool Tableau. List of data visualization tools. [9]
➔ REPEATED !

Q8)

a) What is Data Visualization? Describe the challenges of data visualization. [9]

Data Visualization is the graphical representation of information and data using visual elements like charts, graphs, and maps. It helps users understand complex data sets quickly by identifying patterns, trends, and outliers.

Key Points:

- Makes large or complex data easier to understand.
- Common tools include bar charts, pie charts, line graphs, histograms, heatmaps, etc.
- Used in business intelligence, data analysis, and decision-making.
- Enhances communication of data-driven insights.
  Example: A line chart showing monthly sales growth helps managers identify trends and make decisions accordingly.

Disadvantages of Data Visualization

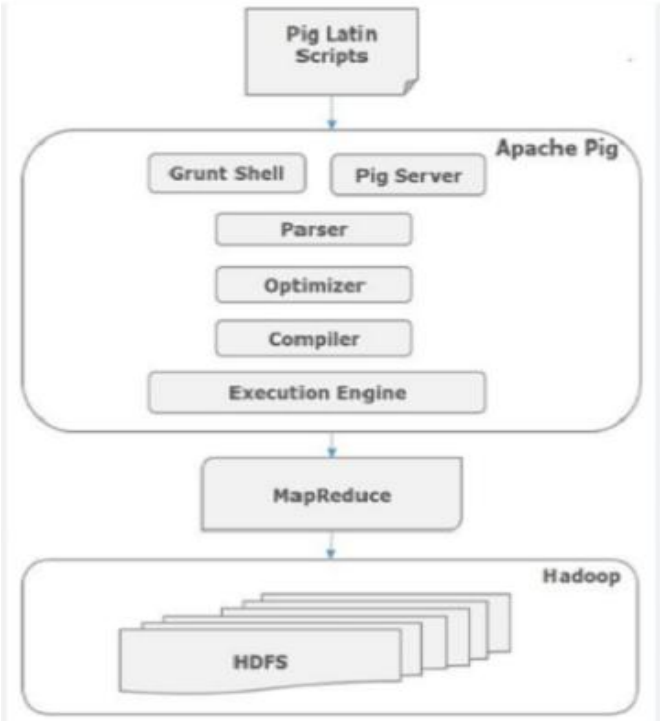
- May lead to misinterpretation if the wrong chart type is used.
- Complex visuals can overwhelm users.
- Visualization tools may have limitations with large or real-time data.

Challenges in Big Data Visualization

1. **Volume:** Visualizing immense datasets is difficult due to sheer size and complexity.
2. **Variety:** Combining structured, semi-structured, and unstructured data sources into a coherent visual is challenging.
3. **Velocity:** Real-time data streams require fast processing and updating of visuals.
4. **Visual Noise:** High density of data points makes it hard to distinguish individual objects.

5. **Information Loss:** Data reduction techniques may simplify visuals but risk losing important details.
6. **Large Image Perception:** Physical and device constraints limit how much data can be effectively displayed.
7. **High Rate of Change:** Rapid data updates can overwhelm users' ability to track changes.

b) Explain architecture of Apache-Pig. [9]



Apache Pig Architecture

Apache Pig is a high-level platform for creating MapReduce programs used with Hadoop. It uses a scripting language called Pig Latin that simplifies the coding required for processing large datasets.

Architecture Components of Apache Pig:

1. **Pig Latin Scripts:**
   - Users write data analysis logic in Pig Latin, a data flow language.
   - Example: LOAD, FILTER, GROUP, JOIN.
2. **Grunt Shell**:
   This is the interactive command-line interface where users write and execute Pig Latin scripts.

3. **Pig Server**:
   It acts as the interface between the user and the Pig runtime environment, managing the execution of Pig scripts.
4. **Parser**:
   The parser takes the Pig Latin script and converts it into a logical plan, which is an internal representation of the data flow.

5. **Optimizer**:
   This component improves the logical plan by applying optimization rules to make the data processing more efficient.

6. **Compiler**:
   The compiler translates the optimized logical plan into a series of MapReduce jobs.

7. **Execution Engine**:
   This engine runs the compiled MapReduce jobs on the Hadoop cluster, processing the data stored in **HDFS** (Hadoop Distributed File System).

8. **MapReduce** executes the data processing tasks generated by Apache Pig
9. **HDFS (Hadoop Distributed File System)** stores the input and output data.

Key Points:

- Pig acts as a bridge between the user and Hadoop.
- It reduces the complexity of writing raw MapReduce code.
- Supports both batch and interactive modes.

➢ NOV / DEC 2023

Q7)

a) List the few data visualization tools and discuss any four applications of data visualization along with the use of the various plots with Python/R or suitable tool

b) List the challenges of Data Visualization. Explain the types of visualization with example.

Q8)

a) Explain in detail the Hadoop Ecosystem with suitable diagram along with the various components. [9]

b) Write a short note on the following. [9]

a) Map Reduce

b) Pig

"I THINK ALL THE QUESTIONS ARE REPEATED. IF YOU FIND ANY CHANGES IN THE QUESTIONS, PLEASE STUDY THEM " !!

➢ MAY / JUN 2024

Q7)

a) What is a histogram? How is it used to visualize the distribution of data? How is it different from a density plot? [9]

A histogram is a graphical representation of the distribution of numerical data. It divides the entire range of data into intervals called bins and counts how many data points fall into each bin. The counts are shown as vertical bars, where the height of each bar represents the frequency of data points within that bin.

Use of Histogram to Visualize Distribution:
Histograms help in understanding the overall shape and spread of the data. They show how data points are distributed across different value ranges, which can reveal key characteristics such as:

- **Central tendency:** Where most data points lie.
- **Spread:** The range over which the data varies.
- **Skewness:** Whether the data is symmetric or leans more to one side.
- **Modality:** Whether the data has one or more peaks (modes).
- **Outliers:** Values that are significantly different from the rest of the data.

By examining a histogram, one can quickly identify patterns like normal distribution, uniform distribution, or skewed data.

Difference from Density Plot:

- A **density plot** is a smooth, continuous curve that estimates the probability density function of the data. It shows the relative likelihood of data points at different values, providing a smoothed version of the histogram.
- **Histogram** shows discrete frequency counts within bins, while **density plots** show a continuous probability distribution.

- Density plots are better for comparing multiple distributions and understanding the overall data shape without the binning effect.

Feature	Histogram	Density Plot
Representation	Vertical bars for frequency counts	Smooth, continuous curve
Data Grouping	Data divided into bins	No bins; smooth estimation
Visualization Type	Discrete intervals	Continuous distribution estimate
Sensitivity	Depends on number and width of bins	Smoother, less affected by bins

b) What is the Hadoop ecosystem, and what are its primary components? What is MapReduce, and how does it fit into the Hadoop ecosystem? [9]

HADOOP ECOSYSTEM AND ITS COMPONENTS

ALREADY DONE

What is MapReduce, and how does it fit into the Hadoop ecosystem :-

MapReduce is a programming model and processing technique used to process and generate large datasets in a distributed environment. It breaks down data processing into two main tasks:

- **Map phase:** The input data is divided into smaller chunks, and the **Map function** processes these chunks in parallel to produce intermediate key-value pairs.
- **Reduce phase:** The **Reduce function** collects and merges all intermediate values associated with the same key to produce the final output.

MapReduce enables scalable and efficient data processing by distributing the work across multiple nodes in a cluster, allowing parallel computation.

How MapReduce fits into the Hadoop ecosystem:

1. **Data Processing Engine:**
   MapReduce is the core processing framework of Hadoop that processes large datasets stored in HDFS.
2. **Works with HDFS:**
   It reads input data directly from HDFS and writes the output back to HDFS, ensuring efficient data storage and retrieval.

3. **Distributed Processing:**
   It divides tasks into smaller sub-tasks (Map and Reduce), which are executed in parallel across multiple nodes in the Hadoop cluster.
4. **Data Locality Optimization:**
   Hadoop runs Map tasks on the nodes where data blocks are stored to minimize network congestion and improve processing speed.
5. **Fault Tolerance:**
   If any task fails during execution, MapReduce automatically re-executes the failed task on another node, ensuring reliable processing.
6. **Job Scheduling and Resource Management:**
   MapReduce manages the scheduling of tasks and allocates cluster resources to optimize overall job execution.
7. **Scalability:**
   It allows processing of petabytes of data by scaling horizontally across many commodity hardware nodes.

ACCORDING TO MARKS SORT THE ANSWER !! IF ITS TOO BIG !!

Q8)

a) What is a box plot? Explain the different components of a box plot? How do you interpret the median, quartiles, and whiskers in a box plot? What does the interquartile range (IQR) represent in a box plot? [9]

A Box Plot (or Box-and-Whisker Plot) is a graphical representation of data that shows its central tendency, spread, and potential outliers using five summary statistics:

- Minimum
- First Quartile (Q1)
- Median (Q2)
- Third Quartile (Q3)
- Maximum

Components of a Box Plot:

1. **Median (Q2):**
   The line inside the box represents the median, which is the middle value when the data is sorted. It divides the dataset into two equal halves.
2. **Quartiles:**
   - **Lower Quartile (Q1):** The 25th percentile, below which 25% of the data falls.

- **Upper Quartile (Q3):** The 75th percentile, below which 75% of the data falls.

3. **Interquartile Range (IQR):**
   The length of the box, calculated as IQR=Q3−Q1IQR = Q3 - Q1IQR=Q3−Q1. It measures the spread of the middle 50% of the data and indicates variability.
4. **Whiskers:**
   Lines extending from the box to the smallest and largest values within 1.5 * IQR from the quartiles. Whiskers show the range of most data points.
5. **Outliers:**
   Data points outside the whiskers (beyond 1.5 * IQR from Q1 or Q3) are considered outliers and are plotted individually.

Interpretation:

- **Median:** Shows the central value of the dataset. If the median is closer to Q1 or Q3, it indicates skewness.
- **Quartiles:** Indicate how the data is spread in the lower and upper halves.
- **Whiskers:** Represent the range of typical data values, excluding outliers.
- **IQR:** Represents the range of the central 50% of data, highlighting data concentration and variability. A larger IQR means more spread.

b) Explain the role of Apache Pig in data processing workflows on Hadoop? What is Apache Spark, and how does it complement Hadoop for big data processing? [9]

Role of Apache Pig in Hadoop Data Processing:
Apache Pig is a high-level scripting platform that simplifies writing MapReduce programs for processing large datasets stored in Hadoop's HDFS.

It uses a language called **Pig Latin**, which is easier to write and understand compared to Java-based MapReduce code. Pig allows data analysts and developers to express data transformations such as filtering, grouping, joining, and sorting in a concise way without worrying about the low-level details of MapReduce.

- Pig scripts are converted by the **Pig engine** into MapReduce jobs automatically.
- It improves developer productivity by abstracting complex MapReduce logic.
- Pig is especially useful for iterative data processing and rapid prototyping.

Apache Spark
Apache Spark is an open-source, fast, and general-purpose distributed computing system designed for big data processing.

Unlike MapReduce, Spark processes data in-memory, which significantly speeds up data processing tasks.

Spark supports multiple programming languages (Scala, Python, Java, R) and offers built-in libraries for SQL, machine learning, graph processing, and streaming.

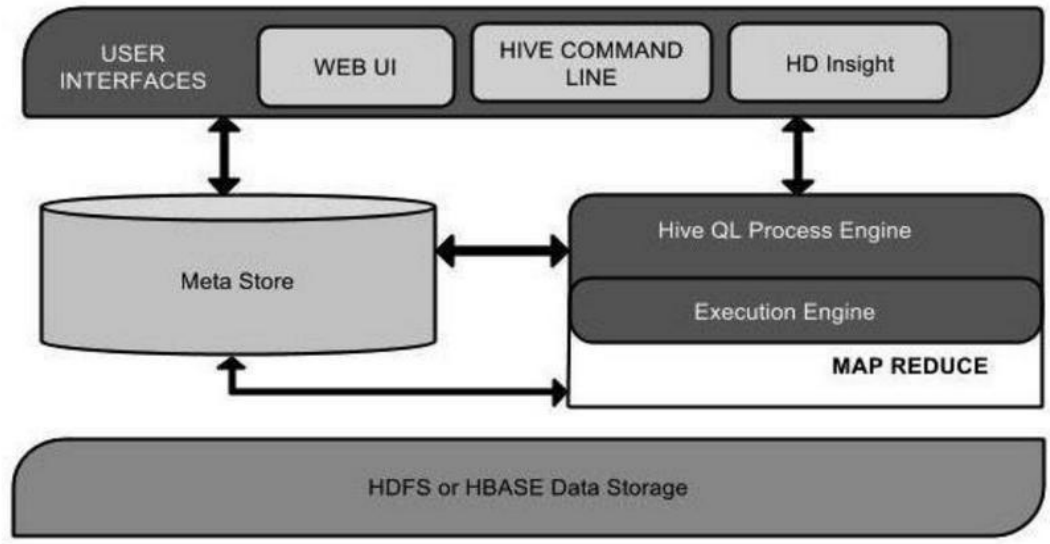
How Apache Spark Complements Hadoop:

- **Speed:** Spark can run workloads up to 100 times faster than traditional MapReduce due to in-memory computation.
- **Compatibility:** Spark can run on Hadoop clusters and use HDFS as its storage layer, making it easy to integrate.
- **Ease of Use:** Spark provides high-level APIs that simplify complex data processing and support interactive data analysis.
- **Advanced Analytics:** Spark supports real-time streaming and machine learning, which MapReduce does not natively support.
- **Flexible Workloads:** Spark handles batch processing, iterative algorithms, and stream processing in one unified engine.

➢ NOV / DEC 2024

Q7)

a) Explain Hive architecture with suitable diagram. Describe characteristics and features of hive. [9]



1. USER INTERFACE

Hive provides different user interfaces to interact with the system:

- **Command Line Interface (CLI)** – Used for writing and running HiveQL queries directly.
- **Web UI** – A graphical interface to interact with Hive.
- **HDInsight** – Used in Windows servers for executing Hive queries.

2. META STORE

- Stores **metadata** about Hive tables (schemas, data types, table locations in HDFS).
- Managed using a traditional **RDBMS** like MySQL.
- Acts as a **lookup system** during query execution and optimization.

3. HIVEQL PROCESS ENGINE

- Accepts and processes **HiveQL (SQL-like)** queries.
- Parses, compiles, and optimizes the queries.
- Converts HiveQL queries into execution plans (DAGs of MapReduce jobs).

4. EXECUTION ENGINE

- Works with the process engine and MapReduce.
- **Executes the query plans** by launching the corresponding jobs.
- Handles job monitoring and collects the final result.

5. MAPREDUCE

- It is the **default execution framework** for Hive.
- HiveQL queries are internally converted into **MapReduce programs**.
- Enables **parallel processing** of large data sets across the Hadoop cluster.

6. HDFS / HBASE (DATA STORAGE)

- **HDFS (Hadoop Distributed File System):** Default storage system for Hive data.
- **HBase:** Used when real-time read/write access is needed instead of batch processing.
- Stores data in **tables** in a distributed format for faster access and scalability.

✓ **CHARACTERISTICS OF HIVE:**

- **SQL-Like Language (HiveQL):** Easy for users familiar with SQL.
- **Schema-on-Read:** Schema is applied when reading data.
- **Batch Processing:** Suitable for large-scale offline analysis.
- **Metadata Management:** Metadata is stored in an RDBMS-based Metastore.

✓ **FEATURES OF HIVE:**

1. **Scalability:** Handles massive datasets on Hadoop clusters.
2. **Compatibility with Hadoop:** Integrates well with HDFS and MapReduce.
3. **Data Warehousing Capabilities:** Supports summarization, querying, and reporting.
4. **Partitioning & Bucketing:** Improves performance of data access.

b) Describe the data visualization tool Tableau. [9]

➔ **Repeated !!**

Q8)

a) What is data visualization and objectives of data visualization? Why it is difficult visualize Big Data? [9]

Data visualization is the graphical representation of information and data using charts, graphs, maps, and dashboards. It helps in understanding patterns, trends, and insights from large datasets more easily and effectively than raw data.

Objectives of Data Visualization:

1. **Simplify complex data**: To make large or complicated datasets easier to interpret.
2. **Identify patterns and trends**: To quickly detect relationships, correlations, and changes over time.
3. **Support decision-making**: Helps stakeholders make data-driven decisions based on visual insights.
4. **Enhance communication**: Makes it easier to share data findings with non-technical users.
5. **Detect outliers and anomalies**: Useful in spotting data errors or unusual behaviors.

Difficult to Visualize Big Data

1. **Volume**: Big Data consists of massive datasets that are difficult to load, process, and visualize using traditional tools.
2. **Variety**: It comes from diverse sources (text, images, videos, logs), making it hard to standardize for visualization.
3. **Velocity**: Big Data is generated at high speed (real-time), which demands dynamic and fast visualization updates.
4. **Complexity**: Data with many dimensions and variables is harder to represent clearly without oversimplification.
5. **Tool limitations**: Many traditional visualization tools struggle with scalability and performance for large datasets.
6. **Data preprocessing needs**: Big Data often requires extensive cleaning and transformation before visualization.

b) Write a note on Microsoft Power BI and Qlik [9]

1. Microsoft Power BI

Definition:
Power BI is a powerful Business Intelligence (BI) and data visualization tool developed by Microsoft. It enables users to connect to multiple data sources, transform raw data, and create interactive dashboards and reports.

Key Features:

- **Data Connectivity:** Connects to Excel, SQL Server, cloud services, APIs, and more.
- **Data Modeling:** Offers DAX (Data Analysis Expressions) for data manipulation and modeling.
- **Interactive Dashboards:** Users can build dynamic and shareable visual reports.

- **Real-time Analytics:** Supports real-time data streaming and monitoring.
- **AI Integration:** Includes AI-driven insights with natural language queries.
- **Power BI Service:** A cloud-based platform to publish and share dashboards online.

Use Cases:

- Business performance tracking
- Sales analysis
- Financial reporting
- Operational monitoring

2. Qlik (QlikView & Qlik Sense)

Definition:
Qlik is a data analytics and business intelligence platform known for its associative data model, which allows users to explore data without being limited to predefined queries.

Key Features:

- **Associative Data Engine:** Enables free-form data exploration and discovery.
- **In-memory Processing:** Fast analytics using in-memory technology.
- **Data Integration:** Connects with numerous data sources for centralized analysis.
- **Self-Service BI:** Users can create their own reports and dashboards.
- **QlikView vs Qlik Sense:**
    - **QlikView** is script-based and developer-driven.
    - **Qlik Sense** is more modern, user-friendly, and focused on self-service analytics.

Use Cases:

- Market and customer analysis
- Healthcare analytics
- Supply chain optimization
- Risk management