**1. What is Data Visualization?**

**Answer:**  
Data Visualization is the graphical representation of information and data using visual elements like charts, graphs, and maps. It enables users to understand complex data patterns, identify trends, spot outliers, and make data-driven decisions quickly.

**Applications of Data Visualization with Tools and Plots:**

1. **Business Intelligence (BI):**
   * **Use:** Track KPIs, revenue growth, and customer behavior.
   * **Tool/Plot:** Tableau/Power BI – Bar charts, line charts.
2. **Healthcare:**
   * **Use:** Monitor patient data, disease spread, and hospital performance.
   * **Tool/Plot:** Seaborn/ggplot2 – Heatmaps, line plots.
3. **Finance:**
   * **Use:** Analyze stock market trends and portfolio risks.
   * **Tool/Plot:** Plotly/Matplotlib – Line charts, candlestick charts.
4. **Social Media Analytics:**
   * **Use:** Understand user engagement and sentiment.
   * **Tool/Plot:** Python, Tableau – Word clouds, bar charts.

**Types of Data Visualization with Examples:**

* **Bar Chart:** Used to compare categories (e.g., sales in different regions).
* **Line Chart:** Shows trends over time (e.g., monthly website traffic).
* **Pie Chart:** Displays part-to-whole relationships (e.g., browser usage share).
* **Heatmap:** Displays data intensity using color (e.g., correlation matrix).

**2. Describe the Challenges of Data Visualization**

**Answer:**

1. **Choosing the Right Visualization:** Misuse of charts may mislead viewers.
2. **Handling Big Data:** Visual clutter and performance issues with large datasets.
3. **Data Quality Issues:** Inaccurate or incomplete data impacts insight quality.
4. **Overloading Information:** Too much detail can confuse users.
5. **Understanding Audience:** Complex visuals might not suit all audiences.
6. **Tool Limitations:** Not all tools support advanced visualizations.
7. **Bias and Misinterpretation:** Visualization can unintentionally lead to biased conclusions.

**3. With a suitable example explain Histogram and its usages. How is it used to visualize the distribution of data? How is it different from a density plot?**

**Answer:**  
A **Histogram** is a plot that shows the **frequency distribution** of numerical data by grouping values into bins.

**Example in Python:**

import matplotlib.pyplot as plt marks = [55, 67, 73, 45, 82, 89, 76, 61, 70] plt.hist(marks, bins=5) plt.xlabel("Marks") plt.ylabel("Frequency") plt.title("Marks Distribution") plt.show()

**Usages:**

* Understand data distribution.
* Detect skewness and outliers.
* Find mode(s) in data.

**Difference from Density Plot:**

* **Histogram:** Uses bars, shows actual frequency.
* **Density Plot:** Uses a smooth curve, shows probability density, better for continuous data visualization.

**4. With a suitable example explain and draw a Box plot and its usages. Explain the different components of a box plot. How do you interpret the median, quartiles, and whiskers? What does the interquartile range (IQR) represent?**

**Answer:**  
A **Box Plot** shows the spread and skewness of data based on a five-number summary: **min, Q1, median, Q3, max**.

**Python Example:**

import matplotlib.pyplot as plt data = [55, 67, 73, 45, 82, 76, 61, 70, 55] plt.boxplot(data) plt.title("Box Plot of Scores") plt.ylabel("Scores") plt.show()

**Usages:**

* Compare distributions across groups.
* Detect outliers.
* Show data symmetry and skewness.

**Components:**

* **Median (Q2):** Line inside the box (middle value).
* **Q1 & Q3 (1st & 3rd Quartile):** Edges of the box (25th and 75th percentile).
* **Whiskers:** Extend to the smallest and largest non-outlier values.
* **Outliers:** Data points outside 1.5×IQR range.

**IQR (Interquartile Range):**  
IQR = Q3 - Q1, representing the middle 50% of the data. It is used to detect variability and outliers.

**5. Describe the Data Visualization Tool “Tableau”. Explain its applications in brief. List the data visualization tools and discuss any four applications of data visualization along with the use of the suitable plot.**

**Answer:**  
**Tableau** is a powerful data visualization tool used for analyzing, visualizing, and sharing data through interactive dashboards and reports.

**Applications of Tableau:**

* **Business Analytics:** Sales and performance dashboards.
* **Healthcare:** Patient data monitoring.
* **Marketing:** Campaign performance tracking.
* **Education:** Student progress analysis.

**Popular Data Visualization Tools:**

* Tableau
* Power BI
* Python (Matplotlib, Seaborn)
* R (ggplot2)
* Google Data Studio
* Excel

**Four Applications with Plots:**

1. **Sales Analysis** – Bar Chart (Power BI/Tableau)
2. **Customer Segmentation** – Pie Chart (Python/Excel)
3. **Website Traffic** – Line Chart (Google Data Studio)
4. **Correlation Analysis** – Heatmap (Seaborn/ggplot2)

**6. Explain in detail the Hadoop Ecosystem with suitable diagram along with the various components. Write a short note on: i) MapReduce, ii) Pig, iii) Hive. Explain architecture of Apache Pig. What is MapReduce, and how does it fit into the Hadoop ecosystem? 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?**

**Answer:**

**Hadoop Ecosystem Diagram (Text Format)**

Data Sources

↓

Data Ingestion (Sqoop, Flume)

↓

Storage (HDFS, HBase)

↓

Processing (MapReduce, Apache Spark, Tez)

↓

Query Tools (Hive, Pig, Impala)

↓

Visualization (Tableau, Power BI)

**Core Hadoop Components:**

1. **HDFS:** Hadoop Distributed File System for storing massive data.
2. **YARN:** Resource manager for job scheduling.
3. **MapReduce:** Core processing model for batch data.
4. **Pig:** High-level scripting for data transformation (Pig Latin).
5. **Hive:** SQL-like query engine for HDFS data.
6. **HBase:** NoSQL DB for real-time access.
7. **Sqoop & Flume:** Tools for importing/exporting structured & unstructured data.
8. **Spark:** Fast, in-memory data processing engine.

**Short Notes:**

**i) MapReduce:**

A programming model in Hadoop to process large-scale data in parallel.

* **Map:** Converts data into key-value pairs.
* **Reduce:** Aggregates the output.

**ii) Pig:**

A high-level platform using **Pig Latin** scripts for ETL and data manipulation over Hadoop.

**iii) Hive:**

Data warehouse software that uses **HiveQL** (similar to SQL) to manage and query structured data in HDFS.

**Apache Pig Architecture:**

1. **Pig Latin Scripts** →
2. **Parser/Compiler** →
3. **Execution Engine (MapReduce or Tez)** →
4. **HDFS Output**

**MapReduce in Hadoop:**

* Works on top of HDFS.
* Handles batch processing by splitting jobs into map and reduce tasks.
* Automatically parallelized.

**Role of Apache Pig:**

* Converts Pig Latin scripts to MapReduce jobs.
* Ideal for cleaning, transforming, and analyzing raw data.
* Simplifies data processing compared to writing Java MapReduce code.

**Apache Spark:**

* An open-source, fast, in-memory data processing engine.
* Supports batch, streaming, ML, and graph analytics.
* Integrates with HDFS and Hive, complementing Hadoop by providing **real-time, iterative processing** much faster than MapReduce.

**FOR EXTRA MARKS ADD BELOW DIAG/FLOWCHARTS IN ANS.**

**1.Hadoop Ecosystem Flowchart**

Data Sources

↓

Data Ingestion (Sqoop, Flume)

↓

Storage (HDFS, HBase)

↓

Processing (MapReduce, Apache Spark, Tez)

↓

Query Tools (Hive, Pig, Impala)

↓

Visualization (Tableau, Power BI)

**2. Box Plot Components Diagram**

|---------|--------------------|--------------|--------------------|---------|

Min Q1 (25th percentile) Median (Q2) Q3 (75th percentile) Max

- Box edges at Q1 and Q3 (Interquartile Range, IQR)

- Line inside the box is Median

- Whiskers extend from min to max (excluding outliers)

- Outliers shown as dots beyond whiskers

**3. Histogram vs Density Plot**

| **Histogram** | **Density Plot** |
| --- | --- |
| Uses bars | Uses a smooth curve |
| Shows frequency (counts) | Shows probability density |
| Good for discrete binning | Good for continuous data smoothness |
| Dependent on bin width | No bins, continuous estimation |

**4. Types of Data Visualization (Examples)**

| **Type** | **Use Case** | **Example Plot** |
| --- | --- | --- |
| Bar Chart | Compare categories | Sales by region |
| Line Chart | Show trends over time | Website traffic |
| Pie Chart | Show part-to-whole relationships | Market share |
| Heatmap | Show data intensity by color | Correlation matrix |

**5. MapReduce Workflow**

Input Data (Split into chunks)

↓

Map Phase

(Processes each chunk, outputs key-value pairs)

↓

Shuffle and Sort

(Groups key-value pairs by key)

↓

Reduce Phase

(Aggregates grouped data)

↓

Output Results

**6. Apache Pig Architecture**

Pig Latin Script

↓

Parser & Compiler

↓

Execution Engine (MapReduce/YARN)

↓

HDFS (Input/Output Storage)

**7. Apache Spark Features**

| **Feature** | **Description** |
| --- | --- |
| In-Memory Processing | Much faster than disk-based Hadoop MapReduce |
| Supports Batch and Stream Processing | Can handle real-time and batch data |
| Compatible with Hadoop Ecosystem | Uses HDFS, Yarn, Hive |
| Advanced Analytics | Machine Learning, Graph Processing |

**8. Box Plot Interpretation Example**

Data Set: [2, 4, 5, 7, 9, 12, 15, 18, 22, 30]

Median (Q2): 10.5

Q1: 5

Q3: 18

IQR: 18 - 5 = 13

Whiskers:

- Lower whisker: 2 (lowest non-outlier)

- Upper whisker: 30 (highest non-outlier)

Outliers: None if data points fall within 1.5\*IQR range beyond Q1 and Q3

**9. Tableau Data Flow**

Data Connection (Excel, SQL, Cloud)

↓

Data Preparation & Cleaning

↓

Data Visualization (Drag and Drop interface)

↓

Interactive Dashboards & Reports

↓

Sharing & Collaboration (Web, Mobile)

**10. Data Visualization Process**

Raw Data

↓

Data Cleaning

↓

Selecting Visualization Type

↓

Creating Charts/Graphs

↓

Interpreting Results

↓

Decision Making