1. Big Data and Its Characteristics

Big Data means **very large and complex datasets**. It is described by **5 Vs** –

- **Volume** (large size),
- **Velocity** (fast generation),
- **Variety** (different formats),
- **Veracity** (accuracy),
- **Value** (usefulness of data).

---

2. GFS Components (Google File System)

- **Master Node** – manages metadata.
- **Chunk Server** – stores actual data in chunks.
- **Client** – accesses and writes data through master.

---

3. Hadoop Framework

Open-source framework to process big data using:

- **HDFS** (storage),
- **MapReduce** (processing),
- **YARN** (resource management).

---

4. Cluster of Commodity Hardware

A **group of normal/low-cost computers** (not high-end servers) used together to store and process data.

---

5. HDFS Components

- **NameNode** – stores metadata.
- **DataNode** – stores actual data.
- **Secondary NameNode (SNN)** – backs up NameNode's data (not a replacement).

---

6. YARN Components

- **Resource Manager** – allocates resources.
- **Node Manager** – manages tasks on each node.

---

7. Fault Tolerance, Durability, Parallel Processing

- **Fault Tolerance** – system continues even if some nodes fail.
- **Durability** – data is not lost (copies stored).
- **Parallel Processing** – tasks run on multiple nodes at once.

---

8. MapReduce

A programming model:

- **Map** – processes data in parallel.
- **Reduce** – combines outputs.

---

9. HDFS File Storage

Files are split into **blocks** (default 128MB), and stored across **DataNodes** with **3 copies** for safety.

---

10. Hadoop Ecosystem Tools

- **Hive** – SQL-like queries.
- **HBase** – NoSQL database.
- **Pig**, **Sqoop**, **Flume** – data transfer tools.

---

11. Pandas DataFrame

- **Datatype** – int, float, object.
- **Attributes** – .shape, .columns.
- **Methods** – .head(), .info(), .describe().

---

12. Tableau Features & Project Stages

Features – Drag-drop UI, filters, dashboards.
Stages – Data connection → Cleaning → Visualization → Dashboard.

---

13. Tableau Panes

- **Data Pane** – all data fields.
- **Analytics Pane** – trends, averages.

- **View Pane** – where charts are built.
- **Shelves** – Rows, Columns, Filters.

---

14. Measures and Dimensions

- **Dimensions** – categories (e.g., Name, Date).
- **Measures** – numeric data (e.g., Sales, Profit).

---

15. Central Tendency

- **Mean** – average.
- **Median** – middle value.
- **Mode** – most frequent value.

---

16. Why Normalize Data?

To scale values between **0–1** or **standardize** so features are equal.
Methods:

- **Min-Max**,
- **Z-score**,
- **Decimal Scaling**.

---

17. CAP Theorem

A distributed system can have only **2 of 3**:

- **Consistency**,
- **Availability**,
- **Partition Tolerance**.

---

18. NoSQL Databases

Databases for **unstructured or semi-structured** data (not traditional tables).
Examples: **MongoDB**, **Cassandra**.

---

19. Data Visualization (DV)

Process of showing data using **charts/graphs**.
Tools – **Tableau, Power BI, Helical Insight**.
Goal – **make data easy to understand**.

---

20. Types of Charts

- **Bar Chart** – compare categories.
- **Line Chart** – trends over time.
- **Pie Chart** – parts of whole.
- **Histogram** – frequency.
- **Scatter Plot** – relationships.

---

21. What is Dashboard?

A dashboard is a **collection of visual charts** to view data insights.
Used in business for **quick decisions**.
Example: Sales Dashboard.

---

22. Phases of Data Science Project

1. Problem Understanding
2. Data Collection
3. Cleaning
4. Analysis
5. Modeling
6. Evaluation
7. Deployment

---

23. Big Data 4Vs Challenges

- **Volume** – storing huge data.
- **Velocity** – real-time processing.
- **Variety** – different data types.
- **Veracity** – handling inaccurate data.

---

s