1. What is Big Data?

Big Data refers to **datasets that are too large, fast, or complex** for traditional databases to handle.
It requires **distributed storage, parallel processing, and new technologies** to manage, analyze, and extract value.

---

2. Characteristics (The V's of Big Data)

- **Volume** → Size of data (TB, PB, ZB).
- **Velocity** → Speed of data generation & processing (real-time, batch).
- **Variety** → Multiple data types: structured (tables), semi-structured (JSON/XML), unstructured (images, videos, text).
- **Veracity** → Trustworthiness & quality of data.
- **Value** → Extracting business insights.

---

3. Challenges in Big Data

- **Storage** → Handling massive data efficiently.
- **Processing Speed** → Need for parallel, distributed frameworks.
- **Data Integration** → Combining data from multiple sources.
- **Data Quality** → Inconsistent, incomplete, or noisy data.
- **Security & Privacy** → Protecting sensitive information.
- **Scalability** → Supporting growth without performance issues.

---

4. Big Data Computing Modes

1. **Batch Processing** – Process stored data in large chunks.
   Tech: MapReduce, Spark.
2. **Stream Processing** – Real-time event/data handling.
   Tech: Spark Streaming, Flink, Storm.

3. **Graph Processing** – Analyzing relationships in large networks.
   Tech: GraphX, Neo4j.
4. **Query & Analysis** – Interactive querying & reporting.
   Tech: Hive, Impala, Presto.

---

5. Big Data Ecosystem Components

🔷 **Storage**

- **HDFS (Hadoop Distributed File System)**
  - Distributed, fault-tolerant storage.
  - Splits files into blocks, replicates for reliability.

🔷 **Processing**

- **MapReduce**
  - Batch-oriented, divides into map & reduce tasks.
  - Reliable but slower.
- **Apache Spark**
  - In-memory processing → 100x faster than MapReduce.
  - Unified platform for batch, streaming, ML, graphs.

🔷 **Data Management**

- **Hive** → SQL-like queries on Hadoop.
- **HBase** → NoSQL DB for random access, real-time read/write.

🔷 **Coordination**

- **Zookeeper** → Manages cluster coordination, leader election, configs.

---

6. Benefits of Big Data

- Faster & better decision making.
- Real-time insights into operations/customers.

- Fraud detection, predictive maintenance, recommendations.
- Optimized business processes & personalization.

---

7. Applications / Use Cases

- **Finance** → Fraud detection, algorithmic trading.
- **Healthcare** → Predictive analytics, patient monitoring.
- **Retail & E-commerce** → Recommendation engines, customer analysis.
- **Telecom** → Network optimization, churn prediction.
- **Social Media** → Sentiment analysis, targeted ads.

---

8. Interview Tip – "Ecosystem in One Line"

👉 *HDFS stores data → MapReduce/Spark process it → Hive analyzes it → HBase stores real-time data → Zookeeper manages coordination*.