General Big Data Questions:

1. **What is Big Data?**
   Big Data refers to large and complex datasets that traditional data processing systems cannot handle effectively.
2. **What are the 3 Vs of Big Data?**
   Volume, Velocity, and Variety.
3. **What is the difference between structured, unstructured, and semi-structured data?**
   Structured data is organized (e.g., databases), unstructured lacks structure (e.g., emails), and semi-structured contains elements of both (e.g., XML).
4. **What are the challenges of Big Data?**
   Challenges include data growth, integration from multiple sources, tool selection, and securing data.

Hadoop Questions:

5. **What is Hadoop?**
   Hadoop is an open-source framework for distributed storage and processing of Big Data using commodity hardware.
6. **What are the main components of Hadoop?**
   HDFS (storage), YARN (resource management), MapReduce (data processing), and Hadoop Common (utilities).
7. **What is the purpose of HDFS?**
   To store data reliably across multiple nodes with fault tolerance.
8. **Explain the roles of NameNode and DataNode in HDFS.**
   NameNode manages metadata, while DataNodes store the actual data blocks.
9. **What is rack awareness in Hadoop?**
   A concept in HDFS ensuring data replication across racks for fault tolerance.

MapReduce Questions:

10. **What is MapReduce?**
    A programming model in Hadoop for parallel processing of large datasets.
11. **Explain the roles of Mapper and Reducer.**
    The Mapper processes input data into key-value pairs, while the Reducer consolidates intermediate outputs to generate results.
12. **What is the output of a Mapper?**
    A set of intermediate key-value pairs.

13. **What is the purpose of a combiner in MapReduce?**
    To reduce the amount of data transferred to the Reducer by combining intermediate data locally.

NoSQL Databases:

14. **What is NoSQL?**
    NoSQL databases are non-relational databases designed for distributed data storage and horizontal scalability.
15. **What are the types of NoSQL databases?**
    Key-value stores, document stores, column-family stores, and graph databases.
16. **What is HBase?**
    An open-source, distributed, column-oriented NoSQL database built on top of HDFS.

Apache Spark Questions:

17. **What is Apache Spark?**
    A fast, in-memory data processing framework for large-scale data analytics.
18. **What are RDDs in Spark?**
    Resilient Distributed Datasets, the fundamental data structure in Spark for distributed data processing.
19. **What is lazy evaluation in Spark?**
    Spark defers computation until an action (e.g., collect) is executed to optimize execution plans.
20. **Explain Spark's transformations and actions.**
    Transformations create new RDDs (e.g., map, filter), while actions compute results (e.g., collect, save).

Visualization Questions:

21. **What is the role of visualization in Big Data?**
    Visualization helps interpret large datasets and extract insights using tools like Tableau.
22. **What are the types of charts you can create in Tableau?**
    Bar charts, line charts, pie charts, scatter plots, and dashboards.
23. **What is a Tableau dashboard?**
    A combination of multiple visualizations to present data insights interactively

Based on Practical

Experiment 1: Hadoop Installation and HDFS Commands

1. **What is the purpose of HDFS?**
   HDFS stores large datasets across distributed systems, ensuring scalability and fault tolerance.
2. **How do you format the NameNode in HDFS?**
   Use the command hdfs namenode -format. It is done only once during installation.
3. **List some commonly used HDFS commands.**
   - hdfs dfs -mkdir: Create directories.
   - hdfs dfs -put: Upload files.
   - hdfs dfs -get: Retrieve files.
   - hdfs dfs -ls: List files.
   - hdfs dfs -rm: Remove files.
4. **What is the significance of replication in HDFS?**
   Replication ensures data reliability and availability by duplicating blocks across DataNodes.

Experiment 2: MapReduce - Word Count

5. **What is the input and output format of the Word Count program?**
   Input: Text files.
   Output: Key-value pairs where the key is a word, and the value is its count.
6. **How does a Mapper work in Word Count?**
   It splits lines into words and emits each word as a key with a value of 1.
7. **What is the role of the Reducer in Word Count?**
   It aggregates the counts of each word emitted by the Mapper.

Experiment 3: MapReduce - Union and Intersection

8. **What is the difference between union and intersection in Spark?**
   - Union combines elements from two datasets.
   - Intersection extracts common elements from two datasets.
9. **How do you perform a union operation in Spark?**
   Use RDD.union(otherRDD) to merge two RDDs.

10. **What is the significance of set operations in Big Data?**
    Set operations help process and analyze relationships between large datasets.

Experiment 4: MapReduce - Matrix Multiplication

11. **How does MapReduce handle matrix multiplication?**
    The Mapper emits intermediate key-value pairs for each matrix element, while the Reducer combines these to compute results.
12. **Why is matrix multiplication important in analytics?**
    It is used in machine learning, recommendation systems, and graph algorithms.
13. **What are the configurations required for matrix multiplication in MapReduce?**
    Set the number of rows and columns for both matrices in the configuration.

Experiment 5: MongoDB - Database Creation

14. **What is MongoDB?**
    MongoDB is a NoSQL database that stores data in JSON-like documents.
15. **How do you create a collection in MongoDB?**
    Use db.createCollection("collection_name").
16. **List the CRUD operations in MongoDB.**
    - **Create**: db.collection.insertOne().
    - **Read**: db.collection.find().
    - **Update**: db.collection.updateOne().
    - **Delete**: db.collection.deleteOne().

Experiment 6: Hive - Database and Table Creation

17. **What is Hive?**
    Hive is a data warehouse tool built on Hadoop for querying and managing large datasets using HiveQL.
18. **How do you create a table in Hive?**
    Use CREATE TABLE with schema definition. Example:

sql

Copy code

CREATE TABLE students (id INT, name STRING, marks FLOAT);

19. **What is partitioning in Hive?**
    Partitioning divides a table into parts based on column values, improving query performance.

Experiment 7: Apache Spark - Word Count

20. **What is the key difference between Spark and MapReduce?**
    Spark processes data in-memory, making it faster, while MapReduce writes intermediate results to disk.
21. **What is an action in Spark?**
    An action triggers execution, such as collect() or count().
22. **What is lazy evaluation in Spark?**
    Spark delays computation until an action is invoked to optimize the execution plan.

Experiment 8: Tableau - Visualization

23. **What are the steps to create a chart in Tableau?**
    - Import data.
    - Drag fields to rows and columns.
    - Select the chart type.
24. **What is the difference between a Tableau dashboard and a story?**
    - A **dashboard** combines multiple visualizations into one view.
    - A **story** sequences dashboards and visualizations to narrate a data-driven story.
25. **How does Tableau handle big datasets?**
    Tableau connects to big data sources using live connections or extracts for faster performance.

Additional Generic Questions

26. **What is YARN in Hadoop?**
    YARN manages cluster resources and job scheduling in Hadoop.
27. **What is the significance of Spark RDDs?**
    RDDs allow fault-tolerant, distributed data processing in Spark.
28. **What is Pig in the Hadoop ecosystem?**
    Pig is a high-level scripting language for processing data in Hadoop.

29. **What is a combiner in MapReduce?**
    It is an optional component that performs local aggregation of Mapper output.
30. **How is real-time data streaming handled in Big Data?**
    Tools like Apache Kafka and Spark Streaming process real-time data streams efficiently.