UNIT – 1

1. **Which of the following commands is used to list all files in a directory in HDFS?**  
   a) hadoop fs -ls  
   b) hadoop fs -rm  
   c) hadoop fs -mv  
   d) hadoop fs -cp

Answer: a) hadoop fs -ls

1. **Which command is used to delete a directory and all its contents in HDFS?**  
   a) hadoop fs -rm  
   b) hadoop fs -rmdir  
   c) hadoop fs -rmr  
   d) hadoop fs -rm -R

Answer: d) hadoop fs -rm -R

1. **Which of the following is true about a distributed system?**  
   a) Data is processed by a single machine  
   b) Data and computation are spread across multiple nodes  
   c) It is limited to relational databases  
   d) It cannot handle big data

Answer: b) Data and computation are spread across multiple nodes

1. **Which operation in HDFS is rack-aware?**   
   a) File creation  
   b) File deletion  
   c) File reading  
   d) File copying

Answer: a) File creation

1. **In Hadoop, which of the following is a limitation of MapReduce?**  
   a) Real-time processing  
   b) Scalability  
   c) Fault tolerance  
   d) Data parallelism

Answer: a) Real-time processing

1. **Which operation is performed first when reading a file from HDFS?**   
   a) Data replication  
   b) Block reading from DataNode  
   c) Requesting file metadata from NameNode  
   d) Sorting the file blocks

Answer: c) Requesting file metadata from NameNode

1. **Which of the following tools is primarily used for managing and querying structured data in Hadoop?**   
   a) Hive  
   b) Sqoop  
   c) Flume  
   d) Zookeeper

Answer: a) Hive

1. **What is the purpose of the Secondary NameNode in Hadoop?**   
   a) Backup the DataNode  
   b) Load balancing  
   c) Manage file system namespace  
   d) Merge and checkpoint the filesystem image

Answer: d) Merge and checkpoint the filesystem image

1. **What is the file format used by Hadoop to store large amounts of text data efficiently?**   
   a) XML  
   b) CSV  
   c) JSON  
   d) SequenceFile

Answer: d) SequenceFile

1. **In a distributed system, what does data locality refer to?**   
   a) Data stored near the processing unit  
   b) Data replicated across multiple data centers  
   c) Data available to all nodes in a cluster  
   d) Data processing across different geographical regions

Answer: a) Data stored near the processing unit

UNIT – 2

1. In Hive, what is the result of using the SORT BY clause?

a) It sorts data globally across all reducers  
b) It sorts data within each reducer  
c) It redistributes data among reducers based on a specified column  
d) It randomly shuffles data across reducers

Answer: b) It sorts data within each reducer

2. In Hive, how can you specify that a table should be partitioned by a specific column?

a) PARTITIONED BY  
b) CLUSTERED BY  
c) SORTED BY  
d) ORDERED BY

Answer: a) PARTITIONED BY

3. Which of the following is true about the CLUSTER BY clause in Hive?

a) It orders the data within each partition  
b) It sorts data globally across all partitions  
c) It distributes data among reducers and sorts it  
d) It removes duplicates within a partition

Answer: c) It distributes data among reducers and sorts it

4. What is the role of Hue in the context of Big Data?

a) It is a query language for Hive  
b) It is a web-based interface for interacting with Hadoop  
c) It is a distributed file system like HDFS  
d) It is a programming framework for MapReduce

Answer: b) It is a web-based interface for interacting with Hadoop

5. Which command in Hive allows you to see the structure of an existing table?

a) SHOW STRUCTURE  
b) DESCRIBE TABLE  
c) SHOW TABLE  
d) DESCRIBE FORM

Answer: b) DESCRIBE TABLE

6. In Hadoop, which of the following best describes a Mapper?

a) It sorts the final output data  
b) It processes input data and generates key-value pairs  
c) It aggregates the results of the mapping process  
d) It stores intermediate data between Map and Reduce phases

Answer: b) It processes input data and generates key-value pairs

7. What is the function of the EXTERNAL keyword when creating a table in Hive?

a) It specifies that the table will use data stored outside the Hive warehouse  
b) It indicates that the table should not be used for querying  
c) It allows the table to be accessible by other databases  
d) It ensures the table is deleted along with its data when dropped

Answer: a) It specifies that the table will use data stored outside the Hive warehouse

8. Which of the following commands is used to view all the tables in the current Hive database?

a) SHOW DATABASES  
b) SHOW TABLES  
c) LIST TABLES  
d) DESCRIBE TABLES

Answer: b) SHOW TABLES

9. Which of the following best describes a Reducer in the MapReduce framework?

a) It splits data into smaller tasks for processing  
b) It sorts and shuffles intermediate data  
c) It combines intermediate data to produce final output  
d) It processes the final output to store in HDFS

Answer: c) It combines intermediate data to produce final output

10. In Hive, what is the purpose of the ORDER BY clause?

a) It groups data based on specific columns  
b) It sorts data within each reducer only  
c) It sorts data globally across all reducers  
d) It divides data into partitions for processing

Answer: c) It sorts data globally across all reducers

UNIT – 3

1. How does Hive's "Sort By" operation differ from "Order By"?

a) "Sort By" guarantees global ordering, while "Order By" does not

b) "Order By" guarantees global ordering, while "Sort By" sorts data within partitions

c) "Sort By" only works with numeric data, while "Order By" works with all data types

d) There is no difference; they are interchangeable

Answer: b) "Order By" guarantees global ordering, while "Sort By" sorts data within partitions

2. In Spark, how does the persist method differ from the cache method?

a) persist allows data to be stored in a specified storage level, while cache stores data in memory only

b) cache allows data to be stored in a specified storage level, while persist stores data in memory only

c) Both methods store data in memory but differ in the API

d) persist is used for RDDs, while cache is used for DataFrames

Answer: a) persist allows data to be stored in a specified storage level, while cache stores data in memory only

3. Which Hive function would you use to combine multiple rows of data into a single string?

a) CONCAT()

b) GROUP\_CONCAT()

c) COLLECT\_LIST()

d) STRING\_AGG()

Answer: c) COLLECT\_LIST()

4. How does the groupByKey operation differ from reduceByKey in Spark?

a) groupByKey aggregates data, while reduceByKey groups it

b) reduceByKey requires a combiner function, while groupByKey does not

c) groupByKey groups all the values with the same key, while reduceByKey aggregates them using a function

d) groupByKey is faster than reduceByKey

Answer: c) groupByKey groups all the values with the same key, while reduceByKey aggregates them using a function

5. When integrating Hive with HBase, which storage format is typically used for efficient querying?

a) ORC

b) Parquet

c) AVRO

d) RowKey

Answer: d) RowKey

6. What is the primary advantage of using an external table in Hive for Amazon review data?

a) Hive manages and controls the data

b) Data can be stored outside of Hive's control, preserving the original data location

c) Faster query execution

d) Data is automatically partitioned

Answer: b) Data can be stored outside of Hive's control, preserving the original data location

7. When performing data analysis without partitioning in Hive, what is the most likely outcome compared to using partitions?

a) Increased query execution speed

b) Reduced storage space requirements

c) Slower query execution due to scanning entire datasets

d) Automatic indexing of data

Answer: c) Slower query execution due to scanning entire datasets

8. What is the key difference between Hive and HBase when integrated for data analysis?

a) Hive is used for real-time analytics, while HBase is used for batch processing

b) Hive stores data in a relational format, while HBase stores data in a non-relational format

c) Hive is a NoSQL database, while HBase is an SQL-based database

d) Hive uses row-based storage, while HBase uses column-based storage

Answer: b) Hive stores data in a relational format, while HBase stores data in a non-relational format

9. In Apache Spark, what does lazy evaluation allow you to achieve when working with RDDs?

a) Immediate execution of transformations

b) Improved memory management by delaying execution until necessary

c) Automatically optimize the order of operations

d) Execute operations without requiring any actions

Answer: b) Improved memory management by delaying execution until necessary

10. Which of the following is a key difference between Spark and MapReduce?

a) MapReduce is faster than Spark

b) Spark uses disk-based processing, while MapReduce uses in-memory processing

c) Spark supports iterative algorithms, while MapReduce does not

d) Spark cannot be integrated with Hadoop, while MapReduce can

Answer: c) Spark supports iterative algorithms, while MapReduce does not