1. What is MapReduce?

Hadoop MapReduce is a framework used to process large data sets (big data) across a Hadoop cluster.

2. Mention three benefits/advantages of MapReduce.

The three significant benefits of MapReduce are:

●Highly scalable: Stores and distributes enormous data sets across thousands of servers.

●Cost-effective: Allows data storage and processing at affordable prices.

●Secure: It allows only approved users to operate on the data and incorporates HDFS and HBase security.

3. What are the main components of MapReduce?

The three main components of MapReduce are:

●Main Driver Class: The Main Driver Class provides the job configuration parameters.

●Mapper Class: This class is used for mapping purposes.

●Reducer Class: Reducer class divides the data into splits.

4. What are the configuration parameters required to be specified in MapReduce?

The required configuration parameters that need to be specified are:

●The job's input and output location in HDFS

●The input and output format

●The classes containing the map and reduce functions

●The .JAR file for driver, mapper, and reducer classes.

5. Define shuffling in MapReduce.

Shuffling is the process of transferring data from Mapper to Reducer. It is part of the first phase of the framework.

6. What is meant by HDFS?

HDFS stands for Hadoop Distributed File System. It is one of the most critical components in Hadoop architecture and is responsible for data storage.

7. What do you mean by a heartbeat in HDFS?

Heartbeat is the signal sent by the datanode to the namenode to indicate that it's alive. It is used to detect failures and ensure that the link between the two nodes is intact.

8. Can you tell us about the distributed cache in MapReduce?

A distributed cache is a service offered by the MapReduce framework to cache files such as text, jars, etc., needed by applications.

9. What do you mean by a combiner?

Combiner is an optional class that accepts input from the Map class and passes the output key-value pairs to the Reducer class. It is used to increase the efficiency of the MapReduce program. However, the execution of the combiner is not guaranteed.

10. Is the renaming of the output file possible?

Yes, the implementation of multiple format output class makes it possible to rename the output file.

11. What is meant by JobTracker?

JobTracker is a service that is used for processing MapReduce jobs in a cluster. The JobTracker performs the following functions:

- Accept jobs submitted by client applications
- Communicate with NameNode to know the data location
- Locate TaskTracker nodes that are near the data or are available
- Submit the work to the chosen nodes

●If a TaskTracker node notifies failure, JobTracker decides the steps be taken next.

●It updates the status of the job after completion.

If the JobTracker fails, all running jobs are stopped.

12. Can you tell us about MapReduce Partitioner and its role?

The phase that controls the partitioning of intermediate map-reduce output keys is known as a partitioner. The process also helps to provide the input data to the reducer. The default partitioner in Hadoop is the 'Hash' partitioner.

13. Can Reducers communicate with each other?

No, Reducers can't communicate with each other as they work in isolation.

14. What do you mean by InputFormat? What are the types of InputFormat in MapReduce?

InputFormat is a feature in MapReduce that defines the input specifications for a job. The eight different types of InputFormat in MapReduce are:

●FileInputFormat

●TextInputFormat

●SequenceFileInputFormat

●SequenceFileAsTextInputFormat

●SequenceFileAsBinaryInputFormat

●DBInputFormat

●NLineInputFormat

●KeyValueTextInputFormat

15. How does MapReduce work?

MapReduce works in two phases — the map phase and the reduce phase. In the map phase, MapReduce counts the words in each document. In the reduce phase, it reduces the data and segregates them.