**Hadoop and Big Data**

1. **Explain “Big Data” and what are five V’s of Big Data?**
2. **What is Hadoop and its components?**
3. **List the difference between Hadoop 1 and Hadoop 2?**
4. **What is a checkpoint?**
5. **What is the difference between an “HDFS Block” and an “Input Split”?**
6. **What is Big Data?**
7. **What is the difference between Hadoop and Traditional RDBMS?**
8. **Differentiate between Structured and Unstructured data?**
9. **On what concept the Hadoop framework works?**
10. **What are the main components of a Hadoop Application?**
11. **What is Hadoop streaming?**
12. **What are the most commonly defined input formats in Hadoop?**
13. **What are the steps involved in a big data pipeline?**
14. **What are CSV Files, when would you use csv file?**
15. **What are JSON Files, when would you use JSON file?**
16. **What are Avro Files, when would you use Avro file?**
17. **What are Parquet Files, when would you use Parquet file?**
18. **Hadoop architecture?**
19. **Difference between Hadoop and RDBMS?**

**Hadoop HDFS**

1. **What is a block and block scanner in HDFS?**
2. **HDFS architecture?**
3. **Explain the difference between NameNode, Backup Node and Checkpoint NameNode.**
4. **What is the port number for NameNode, Task Tracker and Job Tracker?**
5. **How can you overwrite the replication factors in HDFS?**
6. **Explain what happens if during the PUT operation, HDFS block is assigned a replication factor 1 instead of the default value 3?**
7. **What is the process to change the files at arbitrary locations in HDFS?**
8. **Explain about the indexing process in HDFS**
9. **What is a rack awareness and on what basis is data stored in a rack?**
10. **What happens to a NameNode that has no data?**
11. **What happens when a user submits a Hadoop job when the NameNode is down?**
12. **What happens when a user submits a Hadoop job when the Job Tracker is down?**
13. **Whenever a client submits a hadoop job, who receives it?**
14. **What is an edge nodes in Hadoop?**

**Hadoop MapReduce**

1. **Explain the usage of Context Object?**
2. **What are the core methods of a Reducer?**
3. **Explain about the partitioning, shuffle and sort?**
4. **How to write a custom partitioner for a Hadoop MapReduce job?**
5. **What are side data distribution techniques in Hadoop?**
6. **Explain 3 Stages of MapReduce?**

**Flume Interview**

1. **Explain about the core components of Flume?**
2. **Explain about the different channel types in Flume. Which channel type is faster?**
3. **Which is the reliable channel in Flume to ensure that there is no data loss and why?**
4. **Flume architecture?**
5. **Explain about the replication and multiplexing selectors in Flume?**
6. **How multi-hop agent can be setup in Flume?**
7. **Differentiate between FileSink and FileRollSink?**