**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?**
20. **Minimum system requirements for running a Hadoop Cluster with High Availability**
21. **What is active and passive NameNode in Hadoop?**

**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?**
15. **Modifying the block placement strategy of HDFS**
16. **How to Configure Replication Factor and Block Size for HDFS?**
17. **hdfs vs spark?**
18. **What is Checkpoint HDFS?**
19. **how do you balance disk space on HDFS clusters?**
20. **Alter file partition on hdfs?**
21. **Move data from ORACLE to HDFS?**
22. **What are HDFS snapshots?**
23. **What is the role of zookeeper in HDFS?**

**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?**

**Spark**

1. **Your experience with spark streaming?**
2. **How spark execute?**
3. **What is windowing in spark streaming?**
4. **How to join 2 input streams?**
5. **What is a pair RDD?**
6. **What are broadcast variables?**
7. **What are examples of actions in Spark?**
8. **What does lazy mean in spark?**
9. **What are types of transformations?**
10. **What are some challenges you've had with spark?**
11. **What is RDD Lineage?**
12. **Do you have experience with Spark streaming?**
13. **What is an RDD?**
14. **What's the difference between a transformation and an action?**
15. **What is shuffling in Spark?**
16. **What RDD function or method performs shuffling?**
17. **What are window functions is Spark? What are some examples?**
18. **Window aggregation function with scenario**
19. **Difference between sort by and order by**
20. **How to create data frame in spark?**
21. **How to work with JSON file in Spark**
22. **Within spark streaming, how do you have the stream check once on?**
23. **Why use dstream instead of structured streaming?**
24. **Data skewness in spark**
25. **Coalesces vs partition?**
26. **Spark expertise?**
27. **What is RDD vs DF vs DS?**
28. **Specific use case of DS?**
29. **Have you worked with Unstructured data?**
30. **Spark Session vs Spark Context?**
31. **Have used Databricks?**
32. **How to install dependencies in Spark?**
33. **What are spark dynamically allocated sources.**
34. **Explain the Distribution of Executors, Cores and Memory for a Spark Application running in Yarn.**
35. **Cluster spark job executor memory usage throughput.**
36. **How do you code extensions in spark?**
37. **How can you optimized memory issues in spark?**
38. **What is the maximum size for a broadcast object in Spark?**
39. **What are broadcast variables and accumulators in Apache Spark?**
40. **What is shuffling?**
41. **How can you remove shuffling in spark?**
42. **Explain a spark groupBy.**
43. **Why is spark faster than hadoop**
44. **Spark what is an rdd**
45. **What is the difference between map and flatMap and a good use case for each**
46. **Difference of semantics between reduce and reduceByKey**
47. **When are accumulators truly reliable?**
48. **Spark worker nodes and executioners?**
49. **Spark vs MapReduce?**
50. **Spark Checkpoint**

**Flume**

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?**
8. **Sqoop vs flume?**
9. **Kafka vs Flume?**

**Hive**

1. **How to connect HBase with Hive?**
2. **How to optimize hive queries and pig?**
3. **How do you do Join in Hive?**
4. **hive - partitioning and bucketing?**
5. **How do you manage slow changing dimensions in Hive?**
6. **what are the commands to connect and pull the 10 most recent records for hive?**
7. **What it the default delimiter in hive?**
8. **Static vs dynamic partitioning in hive?**
9. **What is the difference between external and internal table in Hive?**
10. **What would happen if we add the data manually to Hive table?**
11. **how to change the meta store in hive?**
12. **Write a hive query to remove duplicate from a table**
13. **difference between SQL in Hive vs Spark?**
14. **what are the different joins in hive?**
15. **difference between hive and Hbase?**
16. **impala vs hive?**
17. **explain vectorization in hive and how it's compiled?**
18. **Differences between Apache hive and Apache Spark?**
19. **Hadoop hive core component?**
20. **hive scripting?**
21. **hive default storage location?**
22. **what is a Dynamic-Partition in hive?**
23. **What is the advantage of partitioning and bucketing Hive Table?**
24. **how to put data into hive?**
25. **how to find the number of rows in hive?**
26. **How to optimize hive queries?**
27. **how many mappers are required for a query in hive?**
28. **how many reducers are required for a query in hive?**
29. **What is the hive bucketing formula to determine volume?**
30. **How does dynamic partition help in hive?**
31. **Why do we need dynamic partition to help in hive?**
32. **Main difference between dynamic partition and static partition in hive.**
33. **How can we create dynamic partition column?**
34. **Difference between orderby and sortby in hive?**
35. **What is a hive map side join?**
36. **Advantages of Hive?**
37. **How to put json into hive?**

**Sqoop**

**Kafka**