**Topics:**

1. What is Apache Spark?
2. Difference between Hadoop-Spark
3. Benefits of Spark over Map-reduce
4. RDD, Data Frame, Data Set – Features, RDD Lineage, Schema and Pair RDD
5. Spark Architecture – Driver, Cluster manager, Executor, Executor Memory
6. Partition
7. Operations Transformations( Narrow and Wide Transformation) and Actions
8. Modes of Spark – Spark Core, SQL, MLIB, Graph
9. File System
10. Demerits of Spark
11. How to minimize data Transfer with Spark
12. Persistence level
13. Checkpoints

**List of Apache Spark Interview Questions and Answers**

1. **What is Apache Spark?**
2. **What are the features and characteristics of Apache Spark?**
3. **What are the languages in which Apache Spark create API?**
4. **Compare Apache Hadoop and Apache Spark.**
5. **Can we run Apache Spark without Hadoop?**
6. **What are the benefits of Spark over MapReduce?**
7. **What are the drawbacks of Apache Spark?**
8. **Explain various Apache Spark ecosystem components. In which scenarios can we use these components?**
9. **How do we represent data in Spark?**
10. **How many types of Transformation are there?**
11. **What is lineage graph?**
12. **How can data transfer be minimized when working with Apache Spark?**
13. **How does Apache Spark handles accumulated Metadata?**
14. **How is RDD in Apache Spark different from Distributed Storage Management?**
15. **How does pipe operation writes the result to standard output in Apache Spark?**
16. **How many partitions are created by default in Apache Spark RDD?**
17. **How to split single HDFS block into partitions RDD?**
18. **Define Partition and Partitioner in Apache Spark.**
19. **How to split single HDFS block into partitions RDD?**
20. **Define paired RDD in Apache Spark?**
21. **What are the differences between Caching and Persistence method in Apache Spark?**
22. **What is the use of Spark driver, where it gets executed on the cluster?**
23. **Define Spark Session in Apache Spark? Why is it needed?**
24. **What is catalyst query optimizer in Apache Spark?**
25. **What is a DataSet? What are its advantages over DataFrame and RDD?**
26. **What are the ways to run Spark over Hadoop?**
27. **What is a DStream?**
28. **Explain write ahead log(journaling) in Spark?**
29. **Define the level of parallelism and its need in Spark Streaming.**
30. **Define the common faults of the developer while using Apache Spark?**
31. **What are the various types of shared variable in Apache Spark?**