DBDA – Big Data Technologies

1. What is Big data? Explain volume, velocity and variety.
2. Explain Hadoop 2.x architecture and its daemons. How it differ from Hadoop 1.x?
3. What is map-reduce and explain work-flow of map reduce?
4. Explain Hadoop installation modes? Explain installation steps and what are setup configuration files?
5. What is default replication factor in Hadoop? Why?
6. What is the difference between Hadoop and traditional RDBMS?
7. Differentiate between structured, semi-structured and unstructured data?
8. Differentiate between structured, semi-structured and unstructured data?
9. Explain important HDFS commands. How data is ingested into HDFS? If a job is completed with multiple output files (e.g. part001, part002, ..., part00n), how can we extract the output in a single file on local-system?
10. What happens if NameNode is down after submitting hadoop job?
11. How fault tolerance is handled in Hadoop 2.x?
12. How to increase performance in map-reduce? What is spill?
13. How are the number of mappers and splits decided while executing a job in map-reduce?
14. Imagine that you are uploading a file of 500MB into HDFS. 100MB of data is successfully uploaded into HDFS and another client wants to read the uploaded data while the upload is still in progress. What will happen in such a scenario, will the 100 MB of data that is uploaded will it be displayed?
15. How will you decide the size of your hadoop cluster?
16. How to perform incremental import in Sqoop?
17. Name the hadoop ETL Tools. Explain the basic difference between Sqoop and Flume?
18. What is HBase? Differentiate HBase and Hive?
19. What is the difference between Pig and Hive?
20. What are the benefits of Apache Hive/Pig over map reduce?
21. What is UDF and how to write in Pig and Hive?
22. Explain Hive architecture? What is role of meta-store? How Hive data can be accessed using JDBC? What is role of thrift for hive metadata?

23. What is internal & external table in Hive?

24. Difference between Spark and Hadoop?

25. What is Spark and What is RDD? Which programming language is used with Spark?

26. What are the features of Spark over Hadoop?

27. How is streaming implemented in Spark? Explain with examples?

28. Why Spark is faster than Hadoop? Will Apache Spark replace hadoop?

29. Explain Kafka architecture. What is use of zookeeper? What is significance of producer and consumer groups?

30. Explain Spark streaming architecture.