** Spark Final Preparation Interview Questions**

1) Explain Spark architecture in details?

2) What is difference between Spark and Map Reduce?

3) What is RDD in Spark and how it is differ from other datasets? How to create RDD?

Different features of RDD?

4) What is Spark session?

5) Differentiate RDD, Data Frame, and Datasets in Spark?

6) What are the types of transformation in Spark? Explain types of Transformation?

7) How you will convert RDD into Data frame and Datasets?

8) What are the types of performance tuning performed by Spark? How it will achieve?

9) What are the actions in spark?

10) How spark understand schemas?

11) What is map and Flatmap in Spark?

12) What is groupby key and reduceby key in Spark?

13) How to find line number for particular word in spark?

14) What is caching in Spark? Difference between Cache and Persistence?

15) What is combined by key in Spark?

16) What is difference between combined by key and aggregated by key?

17) What is Spark pivot?

18) What is shared variables in Spark? Explain Accumulator and Broadcast variables in Spark?

19) Write a syntax for declaring accumulator and broadcast variables?

20) Default size of Accumulator and broadcast variables in Spark?

21) When and how to use Accumulator and Broadcast variables used in Spark?

22) Where the Accumulator and broadcast variables stored in spark?

23) What is lazy evolution in Spark?

24) What is partitions in Spark?

25) What is Spark Context?

26) What is RDD leanage?

27) On which factor Spark decides memory for executer?

28) How to submit Sparks’ job in real time production?

29) How large file process through spark?

30) Difference between Coalease and Repartitions?

31) Which Spark optimization techniques you used in your project?

32) What is the memory level in Spark?

33) Which memory level you used in your project and why?

34) Can we process parquet file format with RDD directly and how to do it?

35) Can parquet support schema evaluation in Spark?

36) Can we process Avro file format with RDD directly?

37) How stages created in spark?

38) I have 8 GB of memory and Data is 10 GB how you will stored in spark?

39) Catalyst optimization in spark (internally work)?

40) What will happened in backend when you submit spark query (BACKEND OPERATION)?

41) Lambda function in Spark?

42) How Spark job run in Background? Like if I submit job on spark then what happened in backend?

43) What are the different spark components?

44) What is Cache memory and RAM memory is same or different? If not similar then what is the difference between them?

45) How to submit Spark job on Cluster mode and also for local file system?

46) Which one is better in Spark, Run query on Cluster mode or on Local mode or edge node?

47) What happens when job submitted on Cluster mode, local mode, and Edge node?

48) Difference between RDD VS LINEAGE VS DAG?

49) How Spark deal with fault tolerance?

50) In Spark I want to use output of one job output to be input for the next job?

51) How to schedule Spark job for the last day of month?

52) Why RDD made as read only?

53) How to create persistence and cache RDD?

54) How to configure executor or set executor in spark?

55) SparkSession vs SparkContext in Apache Spark?

56) By Default, how many partitions are created in RDD in Apache Spark?

57) What is the need for Spark DAG?

58) How can data transfer be minimized when working with Apache Spark?

59) What is the role of Driver program in Spark Application?

60) By Default, how many partitions are created in RDD in Apache Spark?

**Spark Scenario based Questions**

1) Write the Spark program to get bellow output.

Input table

Company Qtr. Sales

A q1 5

A q2 3

A q3 7

A q4 9

Output table should be like

Company Q1 Q2 Q3 Q4

A 5 3 7 9

2) Write the spark program to get below output.

Input-

Col1

A|B|C

Output should be like

Col1 col2 col3

A B C

3) I have variable which following list

Scala> Val a= List (1, 2, 3, 4, 5, 6, 7, 8, 9)

Then how to store odd no. in another RDD?