**Accision Labs:**

1. What is singleton Object
2. How do you compare 2 objects in Scala
3. What is meant by data skew
4. How does skew join resolves the data skew issue
5. How to create the empty table in hive to get schema
6. What are the exceptions you faced in spark and how you have overcome that
7. Suppose you have 2 data frames and 2 actions, how many times data will be processed?
8. How do you list the files in a directory
9. How do you print all the files in a directory in Unix
10. Suppose if you have 3 words in a file, how do you print the second word?
11. What are the joins used in spark
12. What is meant by broadcast join?
13. How you will move the data from RDBMS to spark using scala instead of sqoop
14. Suppose if you update the table in hive, will map reduce program will run or not ?
15. How you will update the table in hive and what are the properties to be set in that?
16. What is the use cache and persist
17. How you have overcome the java memory heap issues in spark

**Standard Chartered Interview Questions:**

1. Explain the project architecture and explain what components you will be using in which part of your project
2. Explain the functional flow of project
3. Hive tuning techniques and how you have used in your project
4. Suppose if you have hive query running in one cluster and another query running in different cluster how you will club both
5. Why block size is 128 MB?
6. What is the serialization technique used in your project
7. How you achieve the schema data evolution using AVRO project
8. What is meant by serde and how it plays important role in your project
9. Explain why parquet is more efficient than AVRO and ORC
10. What and all you know about Kafka?
11. What is topics , offset, broker , controller broker
12. How availability in hdfs is maintained?
13. How you will see the spark logs in your project?
14. Have you worked on Nifi ?
15. What are ACID properties and explain its importance
16. What is delta load and how you will handle it in your project
17. What is compactions in Hive and how you will resolve it
18. How you will update the data in Hive and will mapper and reducer runs while updating the query in hive?
19. How you will initialize the spark session?
20. What are the different partitions available in spark?
21. What is speculative execution in spark and how it plays the important role in your deployment
22. How you will register the kyro serialization in your project

**IBM Interview :**

1. Project architecture
2. Functional flow of project
3. Partitions and bucketing
4. Syntax for creating table for both partitioning and bucketing on same table
5. How you will decide the partitioned column on what basis
6. What is map side join and what its importance
7. What is meant Sort merge Join
8. What is use of resource Manager in Hadoop
9. what are transformations and actions , give examples
10. why you go for vectorization and what are the properties you will set
11. Explain the techniques for doing the hive tuning ?
12. what is indexing in hive
13. what is ACID properties and explain its importance
14. how you will update a table with size 10 TB ?
15. how you will create the empty table in hive
16. how you will transfer the data from RDBMS to HDFS through spark ? can you do that ? if so how you will do that ? tell me the syntax
17. explain the modes of deployment in spark ? in your project which mode you have used and why so you have used that ?
18. what is minimum number of parallel process running in executor
19. suppose if you have null values in your data frame , how you will handle that
20. how you will change the data type of a column in df
21. how to add the new column in a data frame
22. how many types of partitions are there in Hive ? which one you will use and why it is so important ?
23. how you will read the table in spark ?
24. explain spark architecture ?
25. what is the difference between narrow and wide transformations and what role it plays in your spark architecture
26. what is the difference between spark context and spark session
27. how you will initialize the spark session ?
28. what is DAG ?
29. what is cache and persist
30. which storage level is best suited for your project and how you will explain that

**Tech Mahindra:**

1. Project architecture
2. Difference between internal and external table
3. Functional flow
4. Spark and its importance
5. Difference between repartition and coalesce
6. Suppose for increasing the partition size if you use coalesce what will happen
7. How many default partitions will be created for rdd
8. What is meant by input split
9. How many number of mappers are required for transferring 10GB of data from RDBMS to HDFS
10. How you will find the 7th highest salary from a table
11. How you will delete the duplicates using ROW number
12. Your cluster size
13. Configuration parameters in your project
14. What is the build tool used in your project
15. What schedulers you will use in your project
16. Oozie workflow means what ?
17. How you will schedule the hive script in oozie ? tell me the steps
18. What is compactions in hive ?how you will overcome that ? why delta files are getting created in compactions

**HCL :**

1. Project explanation
2. Sqoop Internal process
3. How you will handle the incremental data load in sqoop
4. How will you handle the updated records in sqoop
5. How will you automate the Jobs in sqoop
6. Hive thrift server
7. Difference between RDBMS and HIVE
8. Internal table Vs External Table
9. How will you covert a external table to Internal table
10. How to handle a big table and small table in hive ? which join is useful and how?
11. If you have 200 tables in a database , how will you import except 3 ?
12. Difference between Partitions and Bucketing?
13. Performance wise which is good static partition or dynamic partition?
14. How to add new partition to the existing table ?
15. How you will execute HQL script in command line ?
16. Questions on unix commands , head, tail
17. I have a file , in that some word is repeating many time , how you will find that ? Use grep command
18. Check awk and sed commands
19. Suppose if you have 2 tables, one table is 4 rows and all the values are 1 , another table having 3 rows , and it having all values as 1 , so what will be the output if you do right outer join and left outer join ? answer will be 4 \* 3 = 12 rows of 1’s will be the output
20. Difference between accumulator and broadcast ?
21. Coalesce and repartition and why you coalesce and in which particular situation you will do that?
22. Explain the memory levels in spark ? explain MEMORY\_ONLY\_SER in detail?
23. How you will do serialization in spark ?
24. Spark session initialization?
25. groupBy vs ReduceBy

**HCL (Second Round):**

1. explain the project interview in detail
2. how you will handle the incremental data load in Sqoop ? and how you will automate that ? and how you will trigger that in oozie ? Based on which parameters you will choose that triggering?
3. How you will identify the bad records in hive through shell scripting ? explain the logic in unuix shell scripting ?
4. How you will check whether a record present I HDFS directory or not in spark with scala and unix shell scripting ? write the logic and show us ?
5. Suppose if frequently getting delta load in hive ? How you will do that ? Answer is we can do that using views
6. Have you worked on SCB type1 scenerios?
7. Where you will see the logs in hive?
8. How you will find out the number of partitions in a hive table and what is the commands used for ?
9. How you will add partition to a new table ?
10. Can you convert a parquet table to ORC table?
11. How much of data you handled in your project ?
12. How you have did the partition in your project , explain in detail with data size and how it has reduced your processing time, have you did one partition or 2 partitions , if so what are other parameters you need to take that into account?
13. How you will hardcode your password in hive?
14. How will you handle the small files in unix and how you will automate that job?
15. Explain the spark architecture? difference between spark core architecture and Spark SQL architecture?
16. What is the best optimization technique used in spark?
17. I have a condition, if there is file in hadoop , it need to create the df as per schema , but if there is no file in hadoop directory then it need to create the dataframe with empty schema ? write the code and show me how you will do that ? and you need to use struct type not case class ? write the code and show me or explain the logic
18. What are the difficulties you faced in the project ? how you have overcome those difficulties ?
19. Explain a real time use case you have worked in Hbase ?
20. What is meant by data standardization and why companies use that ?
21. In landing layer , why are you using storage layers?
22. What is meant by data lake?
23. Suppose if I have a semi structured data with nested data , how you will process in spark , explain e the logic?
24. How you will handle data from multiple sources? How you will create the architecture for that ?how you will decide which system will act faster ?
25. From source to presentation layer, write the work flow for one job and explain with code

**MindTree:**

1. Explain the project architrcture
2. Spark architecture explanation
3. Difference between narrow and wide transformations
4. How you will apply map for a code ?
5. How you will choose the number of partitions in spark
6. Have a 10GB file ? how you will choose the number of executors ,number of cores for this ?This is very important question in client rounds
7. How you do the spark submit ? tell me the commands ? what is meant by spark dynamic memor allocation
8. What is unit of caching?
9. How caching takes place and explain the logic with ComputeMethod()?
10. What is lazy transformations?
11. How you will check the spark logs ?
12. What are different types of partitions available in spark?
13. What is hash key practitioner?
14. What is different between merge and fork in oozie?
15. Spark optimization techniques?
16. How you will maintain high availability of partition data in spark ?
17. Command for checking the number of partitions
18. How you will choose the number of buckets in a table?

**Note** : This and all the questions which I have attended recently , I have placed only questions , if you need answers , you can check with aditya or else let me know and mail to [gadirajumidhun@gmail.com](mailto:gadirajumidhun@gmail.com)

I will help you out .

All The bEst … Sai Midhun