Generic Questions

1. What are the challenges that you encountered in your project?
2. Did you work on Talend or any other ETL tools before?
3. Source database used in your pervious project.
4. Data flowed in which format from source to destination.
5. Tools used for creating Reports?
6. How we are analyzing the data?
7. What kind of requirements coming from client?

Sqoop

1. How are you able to process the incremental load
2. Scenario: cust table - ID, first name, last name, address
3. sqoop import command if cust changes the address, how will you handle the incremental update - write the command,
4. Difference between append and last update.
5. Ingesting RDBMS data using sqoop... The candidate was asked to explain the ingestion part.
6. Significance of Sqoop Query.
7. How many mappers we can in sqoop maximum?
8. How can you specify mappers in Sqoop?
9. What is incremental load , how it's managed
10. What is a sqoop job
11. Max number of mappers that can be included in sqoop
12. sqoop - USe of split-by
13. How can we import partitioned table in Hive using sqoop?
14. How sqoop will know the latest column values in incremental load?
15. How can we import table without primary key in Hive using sqoop? What if we have only 1 mapper for this import?
16. Sqoop increment [How we can avoid duplicates]
17. How the sqoop import works with hive tables.
18. What are the limitations of sqoop export
19. Why and how to use mappers in sqoop.
20. Can we move the data for partition table in hive from sqoop

Hadoop

1. How HDFS writes files?
2. What do you know about Mapreduce? Explain me about MapReduce architecture.
3. How replications are done by data nodes?
4. If data node failed to write the first copy of data in HDFS, then how it will proceed further?
5. What is the difference between Hadoop 1.x and 2.x?
6. What is single point of failure in Hadoop 1.x?
7. In MapReduce, What are Combiner and Partitioner phases?
8. What is HCatalog?
9. Scenario based, there are 1000 records in a single file, and requirement is to copy every 100 records in a separate file without changing the order.
10. What is appropriate cluster size - how do we derive that?

11. Scenario: Need to process 500 TB of data per day - how many nodes required?
12. What are the various Hadoop container formats and explain its application?
13. Have you triggered Map Reduce jobs?
14. What is Mapper and Reducer?
15. Compression technique used in your project.
16. How to copy data from Local (HDFS command)?
17. What is editlog
18. HDFS : How reading a file in Hadoop works with respect to Namenode and Datanode
19. How replication factor is maintained at writing time to Hadoop.
20. HDFS-CopyFromLocal and Put command, Replication factor
21. Mapreduce- Partitioner , combiner and reducer
22. FSimage
23. Logical Splits
24. What if Name node fails?
25. How to read and write files in HDFS-Flow, There is 10 GB file- How will you store and retrieve from HDFS
26. What the client will do if the Data Node given by Name node fails
27. Rack awareness concept
28. How the Name node knows if particular data fails or not responding to the request – through Heart Beat signal every 3s
29. Mapreduce- Input-file format
30. Mapreduce –Sorting , Shuffling concepts
31. Hadoop Distributed cache
32. What is FS image and edit logs? How is it handled in hadoop 2.x?
33. What is the single point of failure in Hadoop? How can we handle it?

Hive

1. Hive metastore in-depth – how it work and why we using derby why can't other RSBMS for this.
2. Difference between managed and external table in hive? Which scenario you will go for external tables?
3. Where hive store metadata information?
4. Performance tuning in Hive? Partitioning, bucketing and data denormalization.
5. What is Metastore? What metastore we are using to save the table information?
6. Does hive supports non eqi joins?
7. What are the various analytical SQL queries that you worked on?
8. Hive Partitioning, Bucketing & other performance optimization techniques.
9. Types of hive table used in your project.
10. List of Hive functions used by you.
11. Difference between order by and sort by. (FAQ)
12. Different file formats in Hive?
13. How Mapreduce works in Hive?
14. How to load fixed width files / pipe delimited files / XML files in pig and hive?
15. Hive - Implementing Bucketmapjoin.
16. Where does Hive metastore stored?
17. What is the difference between hive and hbase
18. What are the different file formats processed in hive, how the txt data will be loaded into hive table as ORC data?
19. What is the difference between compressed text files and compressed ORC files?

20. In which case we go for partitions and bucketing, if partition is already then why to use bucketing?(where you use partition/bucketing in your project)
21. when you do an update or delete in the existing file, do you perform full scan or partition scan.
22. How to perform update and delete operation.

Spark

1. What is catalyst optimizer?
2. What is Tungsten encoders?
3. What are transformations in spark?
4. How spark applications execute?
5. How DAG creates stages?
6. What are Stages and tasks?
7. Difference between Hadoop and Spark.
8. What is action in Scala?
9. Why Spark is faster than Hadoop?
10. How we will you decide the NO of nodes required for the big data project based on the input volume
    - How many executors you need per node , how will you decide
    - How memory required for per executors
    - Why you increase the executors in per node [he was expecting the calculations]
11. When you have to restart a job? Will you lose data in spark as it works IN MEMORY process?
12. Difference between data frames and RDD?
13. What is difference between dataframe and dataset? Which is better choice? why?
14. Examples on challenges faced in your project and how it was solved?
15. Any UDF used or worked for transformation?
16. What is Mapsidejoin? (FAQ)
17. Where did you use spark in your project?
18. What is RDD?
19. What is ORC file format?
20. Optimization techniques for Spark dataframe joins?
21. How to partition a data frame with 1 million by 10k records per partition with maintaining order?
22. Accumulators and broadcast variables?
23. How to optimize Spark jobs?
24. How to run Spark jobs?
25. RDD vs Dataframe vs Dataset.
26. What is CBO in spark?
27. Once you submit spark application, what will happen?
28. What are transformations and actions? Where are they executed?
29. What is lineage graph?
30. Map vs flatmap
31. Narrow vs wide transformations
32. Coalesce vs repartition
33. Spark shuffling how it works
34. If the spark job is failed in between then how would you handle that scenario?
35. What are Narrow transformations and wide transformations?

36. Map vs flatMap
37. Why to use spark if hive is already there.
38. Use of spark in your project.
39. What is Pair RDD
40. Broadcast variables (Use of Broadcast variable)
41. What is RDD and if we use only one rdd than how it is fault tolerant
42. What are the transformations you have used?
43. Processing Json file format using spark
44. If you receive 100gb data in xml with dynamic schema, how do you handle in spark
45. How do you connect Hbase from spark scala program

Scala

1. What is the necessity to use singleton and companion objects?
2. Scala - Recursive function
3. Scala traits
4. Pattern matching
5. Case class
6. UDF
7. Closures
8. Interpolation
9. Anonymous function

Oozie

1. How do you schedule jobs in your project?
2. Workflows used in oozie?
3. In Oozie, how will run nodes in parallel?
4. How to know if the oozie workflow got complicated?
5. OOzie - How will oozie know if an action is completed or failed?

shell

1. I have a flat file with state, dollar column and other related columns and would need to get the final summarized dollar values for each state using Unix Shell script. How will you do this?
2. UNIX - command to parse a error log file
3. UNIX - How will you extract only few lines from 1000 lines?
4. What is DISTCP? Have you used unix commands and hadoop shell commands?
5. How can we find errors in a file using Unix? If we want to fetch 5 rows before and 5 rows after error messages, how can we find it using unix?
6. UNIX print the all parameters in the single row
7. How to create the subdirectory.(eg: a/b/c/d)

Hbase

1. I have 1m records in Hbase table and I am getting a file which has few records and I would like to get those matched records from hbase, explain

2. Explain the Hbase use case which I worked in the earlier project

pig

1. Difference between join and cogroup
2. How to call pig or hive scripts from Shell?
3. How can we generate sequence number using PIG?
4. What are specialized joins in PIG? What is Map join and Merge join in PIG?

Kafka

1. Where message gets stored in kafka