SQOOP

1. What is sqoop?
2. What are different sqoop commands?
3. What is the difference between target directory and warehouse directory arguments in sqoop import?
4. How to redirect output and log messages from console to files?
5. What are the different file formats and compression techniques in sqoop import?
6. Explain the arguments --columns, --where, --table in sqoop import?
7. What are free form query imports?
8. Is sqoop import a map-reduce job? Pls explain number of mappers argument in sqoop import.
   1. number of mappers
   2. split by
   3. controlling parallelism
   4. bounding vals query
9. How to format output in sqoop import?
10. How to append to existing data in sqoop import?
11. How to delete existing directory during sqoop import?
12. How to deal with nulls during sqoop import?
13. Explain incremental import in sqoop import.
14. What are the different ways of handling password in sqoop.
15. Explain sqoop import execution flow.
16. What is the concept of staging table in sqoop export and why it is needed?
17. Explain sqoop job.
18. How you schedule sqoop job in your project?

HIVE

1. What is hive? How it places itself in Hadoop ecosystem?
2. Why we use hive in place of MapReduce?
3. Can you explain hive internal architecture in detail?
   1. https://cwiki.apache.org/confluence/display/Hive/Design
4. Difference between hive and rdbms.
5. Difference between hql and sql.
6. What is the concept of a table in hive?
7. What are the different types of tables in hive? And what is the difference between them?
   1. managed
   2. external
   3. temporary
8. What are the 3 different ways to insert data into a hive table?
   1. insert command using values
   2. load command (from local (copy paste), from hdfs (cut paste))
   3. insert command using queries (one table to another table)
9. What are the different ways to connect to hive?
   1. hive shell
   2. hue ui
   3. beeline
10. How you run hive queries in production?
    1. beeline with hive.hql file containing hive queries
11. What are different data types in hive?
    1. https://cwiki.apache.org/confluence/display/Hive/LanguageManual+Types
12. What are hive subqueries?
    1. https://cwiki.apache.org/confluence/display/Hive/LanguageManual+SubQueries
13. What are hive views? Why we use views?
14. What are different set operations in hive?
    1. https://cwiki.apache.org/confluence/display/Hive/LanguageManual+Union
15. What are the different types of functions in hive?
    1. https://cwiki.apache.org/confluence/display/Hive/LanguageManual%20UDF
16. What is lateral view and how we use it?
    1. <https://spark.apache.org/docs/latest/sql-ref-syntax-qry-select-lateral-view.html>
    2. <https://cwiki.apache.org/confluence/display/Hive/LanguageManual+LateralView>
17. How to create temporary as well as permanent custom UDFs in hive?
    1. https://cwiki.apache.org/confluence/display/Hive/HivePlugins
18. What is the difference between normalization and de-normalization and what is preferred in hive?

OPTIMIZATIONS

1. What are the different optimization techniques you have used?
2. What is partitioning?
3. How do you select proper columns for partitioning?
4. What is the difference between static and dynamic partitioning?
5. What is bucketing?
6. How do you select proper columns for bucketing?
7. How to calculate the number of buckets needed?
8. Can we have partitions inside buckets? Can we have buckets inside partitions?
9. Tell some differences between partitioning and bucketing.
10. Given following operations - filter, joins, group by - which of them get optimized with the help of partitioning, which of them get optimized with the help of bucketing?
11. What is broadcast join and when to use it?
12. What is the difference between row and columnar file formats? Why columnar file formats are preferred?

SYSTEM DESIGN

1. Difference between database, data warehouse, datalake.
2. Difference between normalization and de-normalization.
3. What are the different normal forms?

ADF

1. What are the different kinds of triggers in ADF? Explain the working of each one of them.

GENERIC

1. What is bigdata?
2. Difference between transactional vs analytical processing.

PROJECT

1. Explain your project in detail.
2. How much amount of data that you are processing per day?
3. What you mean by incremental data? And how you process incremental data in your project?
4. What are the challenges/bottlenecks you faced while developing your pipeline and how did you resolve them?