AVD GROUP
EDUCATION . EXCELLENCE . EMPLOYMENT

Interview Questions

Project Questions:

1. Project flow common question
2. Your role and responsibility in projects
3. What are the considerations has to take while data migration.
4. Data migration & processing pipeline -- which tools, technologies will use and explain why?

==========

Spark:

1. Spark architecture
2. write a spark code to create a new column by applying operations on existing columns of DF
3. How to write DF? different modes to write it?
4. How to rename the columns od DF?
5. difference between coalesce & repartition ? Have you used in project and how?
6. What is skewness ? how to resolve it?
7. Difference between map & flatmap?
8. Spark optimization techniques?
9. What is OOM issue and how to resolve it?
10. What is Cache & persist? difference between them?
11. Difference between DF & Dataser?
12. What is brodcast variables?
13. What is different types of joins in spark? classical joins
14. What is braodcast join, sort merge join in spark?
15. How to remove duplicates from DF?

16. If we create a new column and give a same name for it which is already exists in DF, then what will happen ?

17. UDF functionality in spark? Have you used in project if yes then explain ?

18. What is the advantage of lazy evaluation in spark?

19. What are the memory optimization technique's?

20. There are 2 DFs emp, dept and write a code to join them

21. What is spark session? how it is initialize?

22. What are the issues u have faced in you project n how you resolved those?

23. XML & Json file reading n writing in spark?

24. Write a code to read n write a RDBMS data in spark?

25. Write a code to achieve below o/p

26, Explode function

input dataset:

|MOBILE_NO |CUST_NAME |ITEMS |

+----------+------------------+----------------------+

|7718802201|Riyaz |Chocolates |

|7718802201|Riyaz |Cream Biscuits & Wafers |

Output: -

{ "MobileNumber":"7718802201","Data":[{ "Key":"CUST_NAME","Value":"Riyaz"} ,{ "Key":"ITEMS","Value":"Chocolates,Cream Biscuits & Wafers"}]}

26. Encryption method while writing DF?

27. Which encryption is best and why ?

28. What is the issue with small file processing in spark and how to tackle it?

29. Difference between spark & MapReduce?

30. How to process the CSV which has multiple delimiters.

31. Write word count program.

32. Ways to create DF and create a DF with some data- write a code

==========

Sqoop:

1. How many default mappers ?

2. What is boundary query ?

3. What will happen if we made no. of mapper to 1?

4. If there is no any primary key in RDMBS table then how to import data through sqoop?

5. IF we increase the no. of mappers what will happen?

6. Optimizations in Sqoop?

7. What is split by?

8. How custom query we can mentioned in Sqoop import ? write it down

==========

SQL:

1. Write a SQL query to find a 2nd highest salary

2. Write a SQL query to find Dept wise highest salary

3. Find out highest salary and make null for rest salaries apart from highest salary, write a SQL query

4. lets say there is recs like

name, log in time, log out time

xyz, 12.20. 4.30

abx, 3.00, 10.00

xyz, 6.00, 8,00

.

.

write a query to find out the total time of emp spent in the office.

==========

Hadoop, HDFS, Hive:

1. Difference between cp & distcp?

2. If already a file is present in HDFS then how to write same file on HDFS?

3. What is NM and edge node?

4. How to find a specific word of the file in HDFS?

5. What is data locality

6. Internal & external tables in Hive? difference between them

7. Hive optimization techniques

8. If we delete the table than what will happen with internal & external table?