Interview Questions:

1.What happens when a spark job is submitted

2.What are the things to be taken care of while running a spark-submit

3.what are transformations and actions

4.Diff between Map and flatMap

5. l=[1,2,3,4]

apply both map and flatmap and print output

6.orderBy Vs SortBy

7.where and having

8.Wordcount Program in Pyspark

9.repartition vs colaesce

10.Different file formats: Avro vs orc vs Parquet

11.Cluster config:

100 gb  
1000 nodes  
16 cores  
32 gb m

No of executors --?

No of executor memory-?

12. df=readcsv inferschema -->  
  
df1=df.filter()--  
df2=df1.groupby()--  
df.write

How many jobs, stages and tasks would be created for the above LOC

13. df1 : order df2: order dtls  
id,dpt,location  
100 mn  
90% loc null

how to handle the data during join as most of the records are here null.

14. sort palindrome words in sentence  
# # Input - “please refer to the madam to know the level”  
# # Output - “please level to the madam to know the refer”

write prog in Python

15. Write a SQL query to delete all duplicate email entries in a table named Person, keeping only unique emails based on its smallest Id.

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

| Id | Email |

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

| 1 | john@example.com |

| 2 | bob@example.com |

| 3 | john@example.com |

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

Id is the primary key column for this table.

For example, after running your query, the above Person table should have the following rows:

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

| Id | Email

| 1 | john@example.com |

| 2 | bob@example.com |

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

16. Name Age Education Year

Ram 28 BE 2012

Rakesh 53 MBA 1985

Madhu 22 B.Com 2018

Rakesh 53 MBA 1985

Find duplicates using SQL query and pyspark.

17.

x = [‘ab’, ‘cd’]

print(len(list(map(list, x))))

Write Output:

18. which one is valid set syntax:

set([[1,2],[3,4],[4,5]])

set([1,2,2,3,4,5])

{1,2,3,4}

set((1,2,3,4))

19. keyword = 'aeioubcdfg'

print (keyword [:3] + keyword [3:])

Write output.

20. find its first non-repeating character in string.

Geeksforgeeks

output=f

21. What is the main advantage of column oriented format ? What does it mean ?

22. Single CSV file [ 100 KB ] , stored in s3 location “s3://test/sample.csv”, containing 1 lac rows.

Paths

s3://test/1.mov

s3://test/2.mov …

WAP : Single csv file ,

UDF:def getSizeInPath(path): Long // immutable // bytes // 10 seconds to execute

O/P: Paths, size

Optimize it as well.

23.

T1

Id, name, value, day

1, “one”, “val1” , 13-dec-2021

2, “two”, “val2” , 13-dec-2021

T2

1, “one”, “val4” , 14-dec-2021

3, “three”, “val3” , 14-dec-2021

WAQ [ get latest records ]

1, “one”, “val4” , 14-dec-2021

2, “two”, “val2” , 13-dec-2021

3 “three”, “val3” , 14-dec-2021

24.

cust: customer table(Parquet file)

lookup file: c1,c2,c3,c4,.....(Having columns only in single row)

Write a program in Pyspark to add column missing in cust table from lookup.

25. top 5 city of each state based on city name

T: city,state

26.

loan:

loan\_no,name,add,city

loan\_hist:

loan\_no,name,add,city,loan\_dt

Write a query to print all records of loan which are not in null and all records of loan\_hist which are not in loan in a single query.

27.

input: "i have a book"  
# S2: "book"  
# output:

Write a program to fing if substring 2 is in substring one or not.

28.

student:  
id,name,subject,marks  
1, A, Maths, 65  
1, A, Phy, 63  
2, B, Maths, 75  
2, B, Hindi, 48  
3, C, Humanity,88  
  
O/P:  
total marks (subject)  
  
A 128 Maths  
B 123 Maths  
C 88 Hum

Write a query to get above output.

29.

Problem2 ->  
Size n  
n=4 - Size of an array  
K = 6  
  
List = [1,5,7,1]  
1+5 = 6  
A[0]+a[1]  
  
5+1=6  
a[1]+a[3]  
  
o/p -> 2

30.

Table 1 -> Product\_id,Product\_Price (Catalog)  
Table 2 -> User\_id,Product\_id,Purchase Date (User interaction)  
Table 3 -> User\_id,User\_name,User Address (User Meta)  
  
Also Table2 data is present at s3://user\_spend/ day=<yyy-mm-dd> (User Interaction) ( day is the partition column )  
 Table1 s3://product-catalogue/ (Static Dump) [ file path ] -> parquet format  
 Table 3 s3://user-metada/ (Static Dump) -> parquet format  
  
Top x spenders per month in 2021  
Purchase\_date - yyyy-mm-dd  
 2021-08-13  
  
Top x spenders in a particular month of a particular year  
  
function(x : 3,month = october ,year = 2021)  
{

}  
  
  
O/p -> Shubham , $560  
 Shrishti ,$340,  
 Sagar ,$120

31.

s='PythonforallPythonMustforall'  
  
nonrep(s)

WAP to find first non repeating character.

32. Write a query to fetch values of user\_id column in table table\_a that are and not in table\_b without using the NOT keyword.

33.Table with 100k uid  
  
if id in (3,4,5,6,7) print true else False.

34. CustomerID  
CusotmerName  
TotalSpend\_LC  
TotalDiscount\_LC  
TotalBalance\_LC  
  
TotalSpend\_USD  
TotalDiscount\_USD  
TotalBalance\_USD  
  
exchange\_rate  
0.9  
  
  
  
df.select((TotalSpend\_USD)\*exchange\_rate ).alias(TotalSpend\_USD),(TotalDiscount\_USD)\*exchange\_rate.alias('TotalDiscount\_USD')  
,(TotalBalance\_USD)\*exchange\_rate ).alias(TotalBalance\_USD)).show()  
  
  
Pyspark: Find country with top international voice usage  
subscriber\_id, country, international\_voice\_usage  
1, Nepal, 100  
1, Nepal, 150  
2, India, 50  
3, Peru, 25

35. capture scenario where rename\_col dict contains 100 elements.

36. dob = "1985-04-01"  
I have DOB. Extract age in years