Interview Questions-Visa

Manager round

1.Explain your project

2.What are the optimizations you have worked on in spark?

3.What is shuffling?explain

4.any scenario where you deployed a code and experienced failure alerts

5.Difference between coalesce and repartition

6.Any lessons learnt while leading the team

7.Broadcast join

8.Explain your project

Technical round

1.Python question

find all unique pairs of numbers in an array,N which sum to a value s=15
n=[1,9,42,6,2,0,14,15]

2.explain hash table and hash function

3.how do you handle long running jobs in spark?

4.how do you handle data skewness?

5. two tables

merchant_volume,merchant_name,volume are columns

merchant_category,category and merchant_name

sql query to select top merchant of each category

a.What happens if different categories have same merchant name

CELEBAL TECHNOLOGIES

1ST TECHNICAL ROUND

1.explain project story

2.Databricks runtime

3.difference between csv and parquet

4.transformations used in project

5.is it possible to union 2 df with different schema?How can we do it?

6.find non matches between 2 df

7.operators used in airflow

8. what happens if a incremental daily file does not come on a day in databricks

9.How is incremental load done in databricks?

10.higher order functions and anonymous functions in scala

11.is pyspark and sparksql same in terms of execution?difference

2nd round:

AWS Design round

1.Different redshift clusters

2.glue crawlers and what happens if schema changes?

3.different ec2 instances

4.why redshift does not allow primary keys?

5.3 data sources are there and client wants one single source of data?how will the data modelling be?

6.what is delta load?

7.difference between incremental load and CDC

8.difference between data lake and delta lake

9.difference between athena and redshift

10.how can we increase execution time of lambda?

11.service used to migrate databases?

12.different s3 storage levels and difference

13.how can glue job be triggered?what if one job depends on another?

SMART CUBE

2nd Round:

1.Explain a challenging situation faced in project

2.What is denormalization?

3.difference between union and union all

4.list comprehension

5.lambda functions

6.data structures used

7.find even elements from a list

8.display only the unmatched records from two list

9.why pandas is preferred over spark?

10.how to explode a nested json into row and column in pyspark?

11.what happens internally when we submit spark job

3rd Round:

1.describe any complex architecture you built

2.aws services used

3.find duplicates in a df

4.top 2 customers per month with highest sales(sql)

5.list=[1,2,3,4,5,6]

Find the sum of the odd indexes with and without built-in functions

COFORGE

ROUND 1

1.DIFFERENCE BETWEEN RANK AND DENSE_Rank

2.DEEP COPY VS SHALLOW COPY

3.DATAFRAME VS SERIES

4.LIST VS TUPLE

5.GROUPBYKEY VS REDUCEBYKEY

6.GLUE RESIDES ON MEMORY?

7.RDD VS DATAFRAME

8.SYNCHRONOUS AND ASYNCHRONOUS FUNCTIONS IN LAMBDA

WALMART

ROUND 1

1.find the maximum length of the subset of array having sum as 0

2.find expiry date by adding remaining days to recharge date in pyspark

3.find the count of top trending hashtags but duplicates would not count in the same line
4.spark architecture

5.spark optimizations

6.partitioning in spark

7.yarn architecture

8.shuffle partitions

9.rank vs dense rank

10.data skewness

11.airflow architecture

PUBLICIS

ROUND 2:

1.SERVICES WORKED ON IN AWS

2.PARTITIONING IN HIVE

3.JOBS,STAGE,TASKS IN SPARK

4.SPARK ARCHITECTURE

5.LST=[a,a,b,b,c,c]

Find count of occurrences in python and pyspark

6.airflow architecture

7.project architecture

EPAM

ROUND 2:

1.AWS GLUE ,3,EMR

2.TRANSIENT AND LONG RUNNING JOB IN EMR

3.STEP EXECUTION IN EMR

4.BACKEND OF LAMBDA

5.SYNCHRONOUS AND ASYNCHRONOUS IN LAMBDA

6.spark optimizations

7.relation between cpu cores and partitions

8.ways to solve data skewness

9.can we do repartition on columns

10.adequate query execution in spark

11.generators and decorators

12.list comprehension

13.scd implementation using pyspark

14.args in python

15.checkpointing in spark

16.limitations of lambda

17.where can we see the logs of emr

18.difference between data lake and delta lake

19.serialisation in spark

20.checkpointing in spark

WALMART

ROUND 2:

1.WHAT IS DATA SPILLING?

2.HOW DO YOU DEFINE THE NUMBER OF SHUFFLE PARTITIONS WITH A FILE OF 500 GB AND 10 GB EXECUTOR MEMORY?

3.Broadcast join and Sort merge join

4.broadcast nested loop join

5.Shuffle partitions concepts

6.spark streaming

7.sql leetcode

8.data spilling

9.how to identify long running jobs

10.how to assign resources to spark jobs

11.case class in scala

12.z-order

13.how to solve out of memory errors in spark?