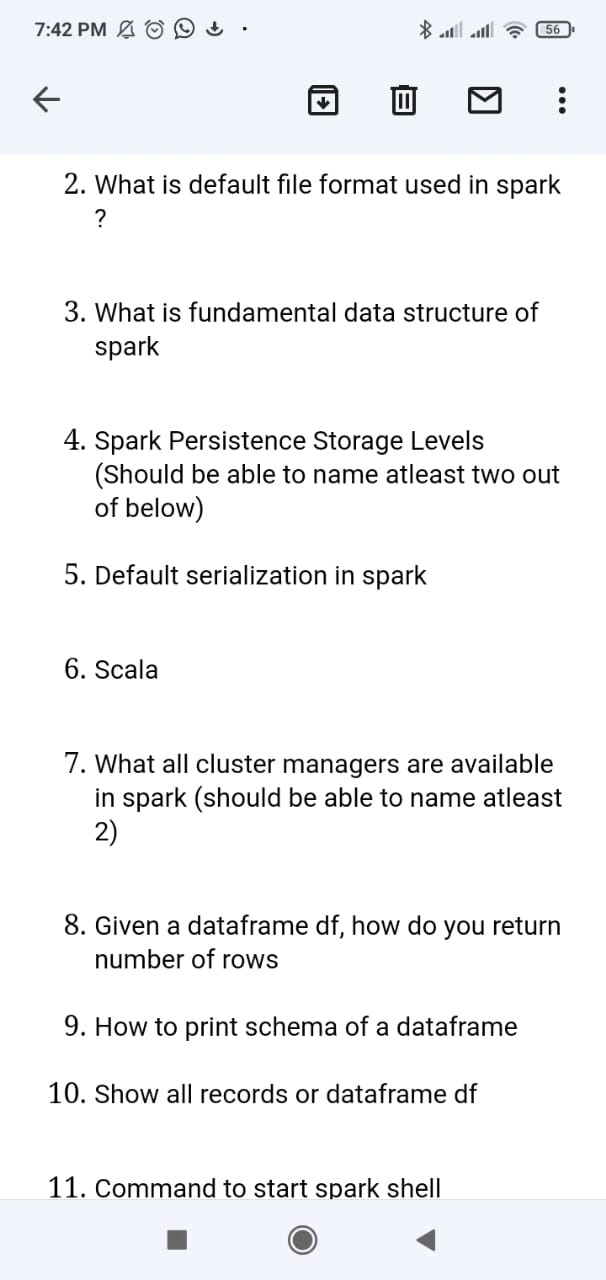
# Which kind of keys does hive support?

# In spark streaming, which of the following is not an output operation on a dstream?

# How many MapReduce will create when hive joins 3 tables

# What is cache vs persist?

# 



No

What is default file format used in spark ?

· Parquet

What is fundamental data structure of spark

· RDD

Spark Persistence Storage Levels (Should be able to name atleast two out of below)

· MEMORY\_ONLY

· DISK\_ONLY

· MEMORY\_AND\_DISK

· MEMORY\_ONLY\_SER

· MEMORY\_AND\_DISK\_SER

· OFF\_HEAP

Default serialization in spark

· Java Serializer

Scala

· Difference between var and val

Val : value

Var : Variable

What all cluster managers are available in spark (should be able to name atleast 2)

· YARN

· Kubernates

· MESOS

· Standalone

Given a dataframe df, how do you return number of rows

df.count()

How to print schema of a dataframe

df.printschema()

Show all records or dataframe df

df.show()

Command to start spark shell

Command : `spark-shell`

Command to clone a git repository

Command : `Git clone <repository name>`

Q2. In spark streaming, which of the following is not an output operation on a dstream?

Answer:- ReduceByKeyAndWindow

Firstly it depends on if the files can be split by Hadoop (splittable) or not. Most files can be split. Example of files that cannot be split are some zipped or compressed files.

Splittable files

1) Calculate the total size of the input files by adding the size of all the files

2) No of Mappers = Total size calculated / Input split size defined in Hadoop configuration

(e.g)

Total size calculated = 1GB (1024MB)

Input split size = 128MB

No of Mappers = 8 (1024 / 128)

(NOTE 1)

FileInputFormat is the bb vb ase class for all implementations of InputFormat that use files as their data source.

1.how to create a folder recursively in Hive ? (ex /test/test1/test2) ( Its one my interview question..i know the answer just try guys)

2.when to choose avro ? and when to choose parquet ?

3.how you will read property files in scala ?

4.when you use small file in Hadoop can you give example ?

Global Logic Interview Q.

1. What are the deployment modes?
2. Partition vs Bucketing
3. About your project.
4. Write down the spark submit command.
5. Lazy evaluation.
6. Flat map and map.
7. Managed table and external tables
8. How do we do optimization in spark?

Scala Programs for

Word Count

Character Count

String Reverse

Prime Numbers

Fibonacci Numbers

Factorial Number

(Select distinct t1.persons\_data t1

From person\_data t1   
Inner join person\_data t2

On t1.persons=t2.persons

Where t1.fruit= ‘Banana’ and t2.fruit=’Apple’) tb1

How to establish mapper

Firstly it depends on if the files can be split by Hadoop (splittable) or not. Most files can be split. Example of files that cannot be split are some zipped or compressed files.

Splittable files

1) Calculate the total size of the input files by adding the size of all the files

2) No of Mappers = Total size calculated / Input split size defined in Hadoop configuration

(e.g)

Total size calculated = 1GB (1024MB)

Input split size = 128MB

No of Mappers = 8 (1024 / 128)

(NOTE 1)

FileInputFormat is the base class for all implementations of InputFormat that use files as their data source.

# What is the output of the program below, given the inputs sequence of strings are

val input : Seq[String] = Seq("metadata\_column", "in\_this\_column-name", "\_index\_column\_name")

Program:

input.map{ columnName =>

val underscorePrefix : Boolean = columnName.take(1) == "\_"

columnName.split("\_").foldLeft( "" ){ (a,b) =>

if ( a.isEmpty ){

if ( underscorePrefix ) "\_" + b else b

}

else {

if ( a.takeRight(1) == "\_" ){

a + b

}

else {

a + b.capitalize

}

}

}

}

# how to create folder recursively in Hive ? (ex /test/test1/test2) ( Its one my interview question..i know the answer just try guys)

# .when to choose avro ? and when to choose parquet ?

# how you will read property files in scala ?

# when you will its a small file in Hadoop can you give example ?

# 

# Global logic :

What are the deployment modes?

Spark is written in which language

Partitioning vs bucketing

About yout project

Write down spark submit command

Lazy evolution

Flatmap and map

Managed table external table

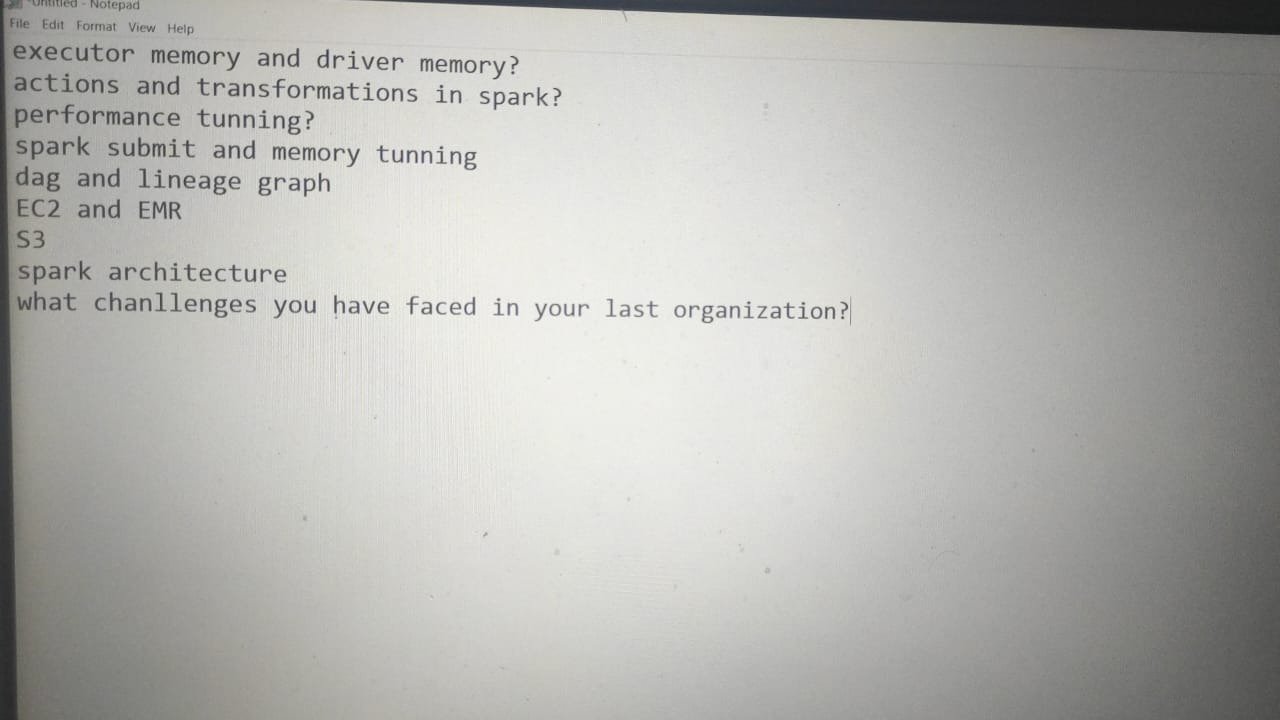
How do we do optimization in spark?

# Given two sorted array: A = [1, 5, 9, 11, 13] B = [3,6,7,10,12] Write a function which accepts two array and return a single merged sorted array without using Arrays.sort or Collection.sort.

# Can anyone tell the difference between Distinct() and dropDuplicate() function in Spark?

# 

# Coforge interview questions..



Distinct applied on then whole dataframe

Drop duplicates can applied On the column level.

And it keeps the first occurance of a value in the column and Drop other rows

# 

# Infosys

Explain your project architecture

2.What re services running on the edge node

3.To filter out the Error logs from log table using pyspark.

4.Few python code as I am comfortable with python.

5.Sqoop related questions.

6.Hive questions

7.File formats

8. And few architecture based questions.

# Capegemini --1). Read csv file in scala. File is without schema, so have to read it with schema.

Scala currying, tuple and write program used in your organization taking any table

2) Concept of Rack awareness

3) Hdfs architecture

4) How the file is distributed in hdfs

5) Yarn Architecture

6) Speculative execution

7) Sqoop work with syntax and file formats you can use with it.

SELECT region, SUM(revenue)

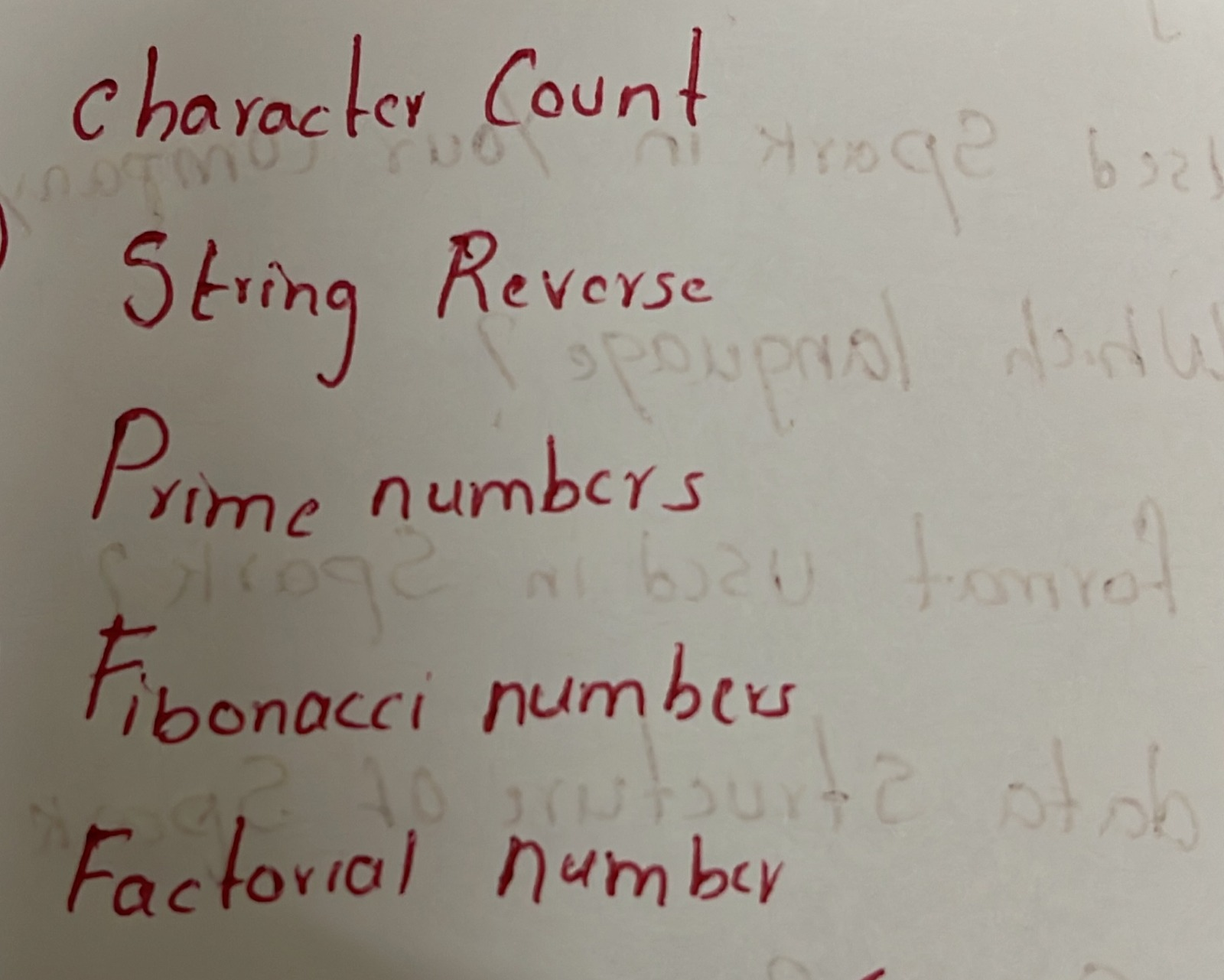
FROM region

WHERE date BETWEEN NOW() - INTERVAL 28 DAY AND NOW()

GROUP BY region

ORDER BY SUM(revenue) DESC

LIMIT 4, 1;



# 

# HCL

1.Hadoop fs -mkdir -p /test1/test2

2.Avro is row format it’s faster for write

3.Source.fromFile("app\_prop.txt").getLines()

4. When it’s smaller than mapper size

In Apache Spark, setMaster is a configuration option that is used to specify the connection to the cluster. There are several different values that can be used with setMaster to specify different types of cluster connections. Here are some common options:

# local: Run Spark in local mode, with a single worker thread. This is useful for testing and debugging, but it is not suitable for distributed processing.

local[\*]: Run Spark in local mode, with as many worker threads as there are cores on the machine. This is useful for testing and debugging, but it is not suitable for distributed processing.

yarn: Connect to a YARN cluster and request resources to execute tasks. YARN will allocate the necessary resources and schedule the tasks on the cluster. This allows Spark to scale out and take advantage of the resources available on the cluster.

mesos://host:port: Connect to a Mesos cluster and request resources to execute tasks. Mesos will allocate the necessary resources and schedule the tasks on the cluster. This allows Spark to scale out and take advantage of the resources available on the cluster.

spark://host:port: Connect to a standalone Spark cluster and request resources to execute tasks. The standalone Spark cluster will allocate the necessary resources and schedule the tasks on the cluster. This allows Spark to scale out and take advantage of the resources available on the cluster.

# Scala Questions:-

Higher Order Functions

HOF take function as an arguments and return funtion as a result

Currying

Currying is technique of transforming a function that takes multiple arguments into a function that takes a single argument

PAF - Partially Applied Functions

Few arguments are passed initially & account to the requirement remaining arguments are passed later

Clousers

A clouser is a function which uses one/more variables which are declared outside the function

Var Vs Val

Var Mutable(Value can be changed later) & Val immutable

Recursion

Function calling itself

Tuple is a collection of elements. Tuples are heterogeneous data structures, i.e., is they can store elements of different data types. A tuple is immutable, unlike an array in scala which is mutable.

Ex:- val name = (15, "Chandan", true)

Map is a collection of key-value pairs.

Ex:- variable = Map(key\_1 -> value\_1, key\_2 -> value\_2,

key\_3 -> value\_3, ....)

Tech Mahindra:

1) Project architecture

2) Difference between internal and external table

3) Functional flow

4) Spark and its importance

5) Difference between repartition and coalesce

6) Suppose for increasing the partition size if you use coalesce what will happen

7) How many default partitions will be created for rdd

8) What is meant by input split

9) How many number of mappers are required for transferring 10GB of data from RDBMS to HDFS

10) How you will find the 7th highest salary from a table

11) How you will delete the duplicates using ROW number

12) Your cluster size

13) Configuration parameters in your project

14) What is the build tool used in your project

15) What schedulers you will use in your project

16) Oozie workflow means what ?

17) How you will schedule the hive script in oozie ? tell me the steps

18) What is compactions in hive ?how you will overcome that ? why delta files are getting created in compactions

# Big Data Project

[Big Data Project Explained -1](https://youtu.be/qJ8Qvbm8wwY)

# 

# Spark Scenario

1,2020-12-30T09:00:00,IN

1,2020-12-30T13:00:00,OUT

1,2020-12-30T14:00:00,IN

1,2020-12-30T17:00:00,OUT

1,2020-12-30T17:30:00,IN

1,2020-12-30T20:00:00,OUT

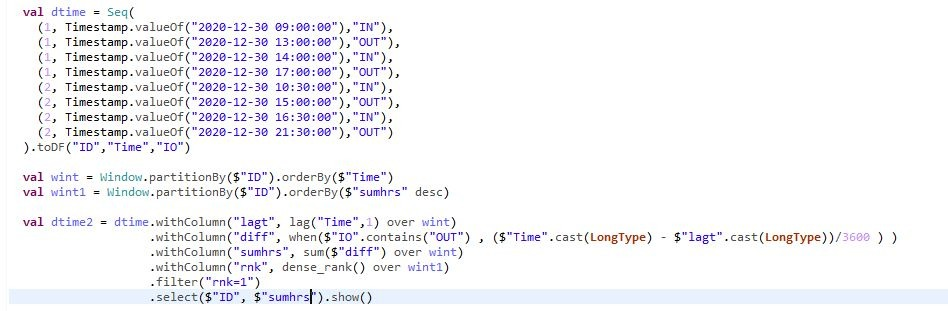
2,2020-12-30T10:30:00,IN

2,2020-12-30T15:00:00,OUT

2,2020-12-30T16:30:00,IN

2,2020-12-30T21:30:00,IN

emplyees login at 9 am in the morning and logout at 1 pm and login back at 2 pm so break timing is 1 hour........Find out how many hours each employee worked



sample file :

Eid, TS, Event

1,2020-12-30T09:00:00, IN

1,2020-12-30T13:00:00, OUT

1,2020-12-30T14:00:00, IN

1,2020-12-30T17:00:00, OUT

1,2020-12-30T17:30:00, IN

1,2020-12-30T20:00:00, OUT

2,2020-12-30T10:30:00, IN

2,2020-12-30T15:00:00, OUT

Output:

1,2 hrs, Out

1, 5 hrs, In

1. Find the difference in start and end time for "In" filter per employee wise.

2. Then sum the time for that employee

Eid, TS, Event

1,2020-12-30 09:00:00, IN

1,2020-12-30 13:00:00, OUT

1,2020-12-30 14:00:00, IN

1,2020-12-30 17:00:00, OUT

1,2020-12-30 17:30:00, IN

1,2020-12-30 20:00:00, OUT

2,2020-12-30 10:30:00, IN

2,2020-12-30 15:00:00, OUT

create table log\_data

(

eid int not null ,

tdate\_ts timestamp,

in\_out\_ind varchar(3)

);

insert into log\_data values (1,'2020-12-30 09:00:00','IN');

insert into log\_data values (1,'2020-12-30 13:00:00','OUT');

insert into log\_data values (1,'2020-12-30 14:00:00','IN');

insert into log\_data values (1,'2020-12-30 17:00:00','OUT');

insert into log\_data values (1,'2020-12-30 17:30:00','IN');

insert into log\_data values (1,'2020-12-30 20:00:00','OUT');

insert into log\_data values (2,'2020-12-30 10:30:00','IN');

insert into log\_data values (2,'2020-12-30 15:00:00','OUT');

select eid,in\_out\_ind,SEC\_TO\_TIME( sum(time\_to\_sec(lhrs))) as login\_hrs

from

(select eid,in\_out\_ind,timediff(out\_time,tdate\_ts) as lhrs

from

(

select \*,

lead(tdate\_ts) OVER (partition by eid order by eid,tdate\_ts ) as out\_time from log\_data

order by eid,tdate\_ts,in\_out\_ind

)aa

)bb

group by eid,in\_out\_ind;

sensorid timestamp values

11111 2021-01-15 10

11111 2021-01-16 15

11111 2021-01-17 30

11112 2021-01-15 10

11112 2021-01-16 20

11112 2021-01-17 30

Output :

sensorid timestamp values

11111 2021-01-15 05

11111 2021-01-16 15

11112 2021-01-15 10

11112 2021-01-16 10

what will be coding for this output

ase class senschema(sensorid: String, timestamp:String, values: String)

def main(args:Array[String]):Unit = {

val conf = new SparkConf().setAppName("first").setMaster("local[\*]")

val sc = new SparkContext(conf)

sc.setLogLevel("ERROR")

val spark = SparkSession.builder().getOrCreate()

import spark.implicits.\_

val rawdata = sc.textFile("file:///c:/sample/interview/sensor.txt")

val sdf = rawdata.map(x=>x.split(","))

val rowdata = sdf.map(x => senschema(x(0),x(1),x(2)))

val df = rowdata.toDF()

df.show()

df.createOrReplaceTempView("sensor")

val df\_result = spark.sql("select sensorid,timestamp,ceil(((lead(values) OVER (partition by sensorid order by values)) - values)) as final\_values from sensor")

df\_result.createTempView("result")

val res\_df = spark.sql("select sensorid,timestamp,final\_values as values from result where final\_values is not null")

res\_df.show()

import org.apache.spark.sql.functions.\_

import org.apache.spark.sql.SparkSession

object obj{

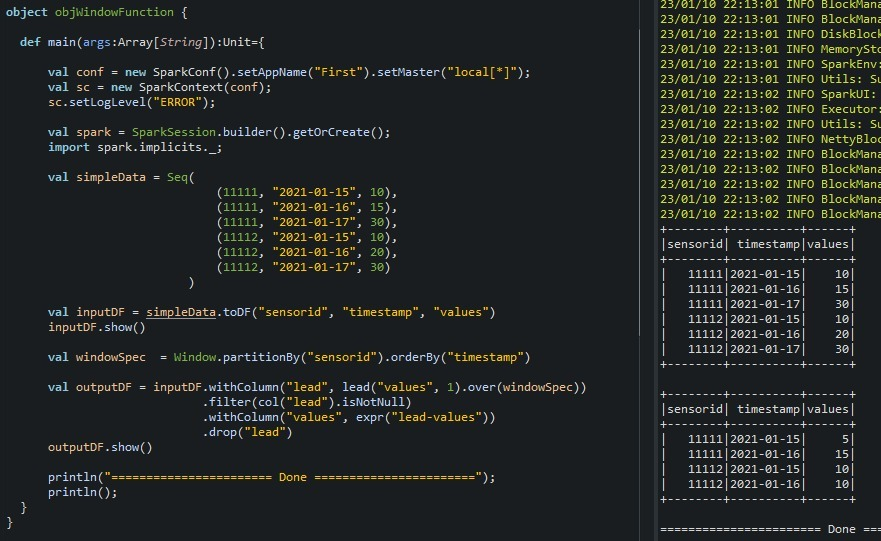
val df = spark.read.format("csv").option("header", "true").load("sensors.csv")

val avgDf = df.groupBy("sensorid", "timestamp").agg(avg("values").as("avg\_values"))

avgDf.show()

}

TCS



# How to calculate executor memory

Based on the recommendations mentioned above, Let’s assign 5 core per executors => --executor-cores = 5 (for good HDFS throughput)

Leave 1 core per node for Hadoop/Yarn daemons => Num cores available per node = 16-1 = 15

So, Total available of cores in cluster = 15 x 10 = 150

Number of available executors = (total cores/num-cores-per-executor) = 150/5 = 30

Leaving 1 executor for ApplicationManager => --num-executors = 29

Number of executors per node = 30/10 = 3

Memory per executor = 64GB/3 = 21GB

Counting off heap overhead = 7% of 21GB = 3GB. So, actual --executor-memory = 21 - 3 = 18GB

# \*Difficulties faced in the project\*

You can say :

Data skewness

Executor mis configuration

Join/shuffles etc

And also let them know how you solved these problems

# Today Interviw in Altrimetrix:

Types od Data loads in Hive table

Sortby vs Orderby

Types of Table

Partitioning and Types

Bucketing

How we will create a new Table in Hive

Save Modes in Spark

# 

# TCS

1. Hdfs layer

2. How the data process in map reduce

3.partition and bucketing

4.external and manage table

5.wht is difference between static and dynamic

6. Lazy evaluation

7.sql view

8. Different types of database

9.tell about yourself and project explanations

Cache and persist

Partition and bucketing when use these

Where and having- difference

Optimization in hive

Please Share Sql Query for this input

Input:

name

a

a

a

b

b

c

c

d

e

OutPut

name,total

a,a3

a,a3

a,a3

b,b2

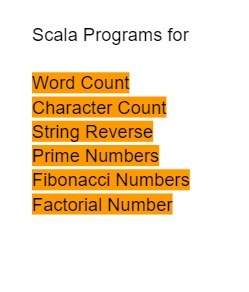
b,b2

C,c2

c,c2

d,d1

e,e1

Steps of Tables in hive

1. Start your Hadoop system or domain

2. Launch hive

(hive)

3. Then create database

(CREATE DATABASE dbname)

4. SHOW DATABASES to check

5. Then ( USE DATABSE)

6. CREATE TABLE tablename and so on

# Interview Question by CTS:

Components of Spark

Components of Hadoop

Spark partition

Spark bucketing

# 

# Accenture

hive external and internal table

Where we use give real time example

Spark session start practical

Sql join output questions

Scoop inport statement

Hive partition syntax

Bucketting syntax

Oops

Polymorphism example

Convert complex sql query to dataframe

Different type of transformations you have used in spark

What is tupple, higher order fuction

How to convert internal table to external

# Write a spark program for below requirements,

File1.csv ~100MB

id,name,age,email

1,abc,31,abc@gmail.com

2,def,23,defyahoo.com

3,xyz,26,xyz@gmail.com

4,qwe,34,qwegmail.com

5,iop,24,iop@gmail.com

File2.csv ~500MB

id,name,age,email,salary

11,jkl,22,abc@gmail.com,1000

12,vbn,33,vbn@yahoo.com,3000

13,wer,27,wer,2000

14,zxc,30,zxc.com,2000

15,lkj,29,lkj@outlook.com,2000

1.Read data from above file into dataframes(df1 and df2).

2.Display number of partitions in df1.

3.Create a new dataframe df3 from df1, along with a new column salary, and keep it constant 1000

4.append df2 and df3, and form df4

5.Remove records which have invalid email from df4, emails with @ are considered to be valid.

6.Write df4 to a target location, by partitioning on salary.

Solution

spark = SparkSession.builder.appName("example\_program").getOrCreate()

//# Read data from file1.csv into a dataframe called df1

df1 = spark.read.csv("file1.csv", header=True)

# Read data from file2.csv into a dataframe called df2

df2 = spark.read.csv("file2.csv", header=True)

# Display the number of partitions in df1

print("Number of partitions in df1:", df1.rdd.getNumPartitions())

# Create a new dataframe df3 from df1, along with a new column salary, and keep it constant 1000

df3 = df1.withColumn("salary", 1000)

# Append df2 and df3, and form df4

df4 = df2.union(df3)

# Remove records which have invalid email from df4, emails with @ are considered to be valid.

df4 = df4.filter(df4.email.contains("@"))

# Write df4 to a target location, by partitioning on salary

df4.write.partitionBy("salary").csv("target\_location"

dataframe uses in spark

rank option in spark

a table which has 2nd highest salary in sql, spark

rank dense rank and row number

partition and bucketing

maximum bucketing in hive

vectorization in spark

optimization in spark

joins out put

coalesce and repartition

X

1

1

1

1

Table 1

y

1

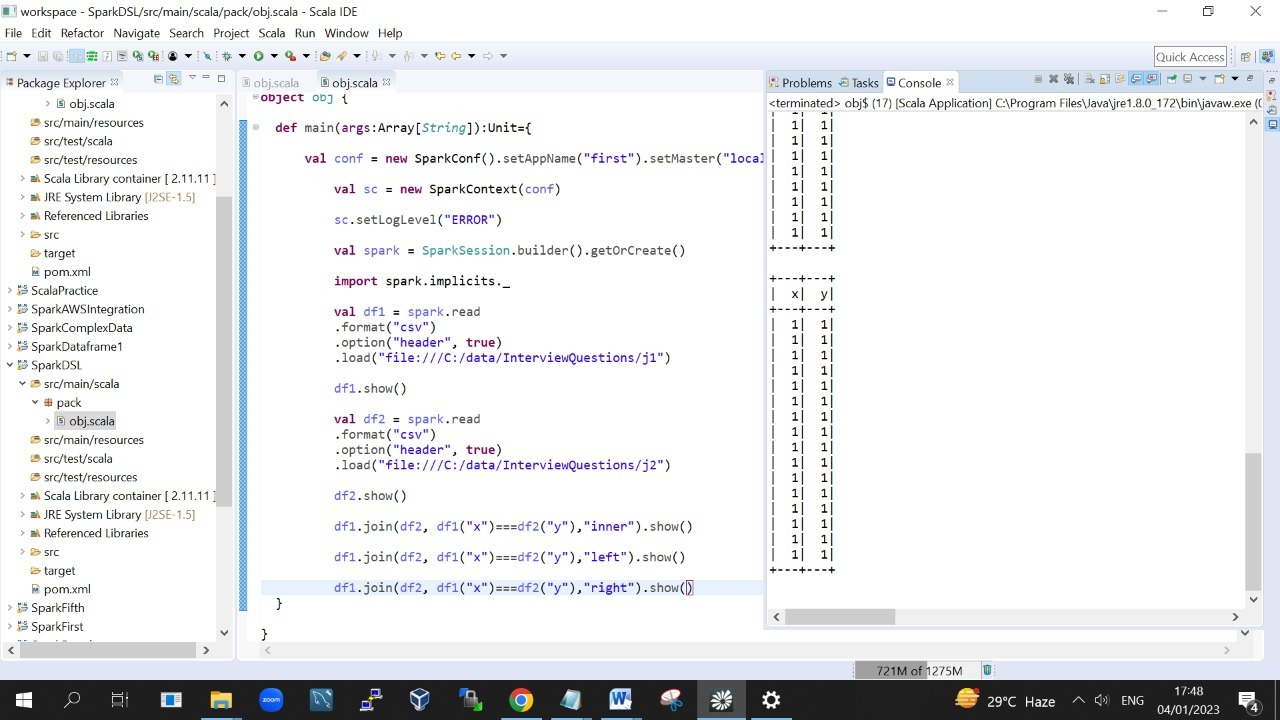
1

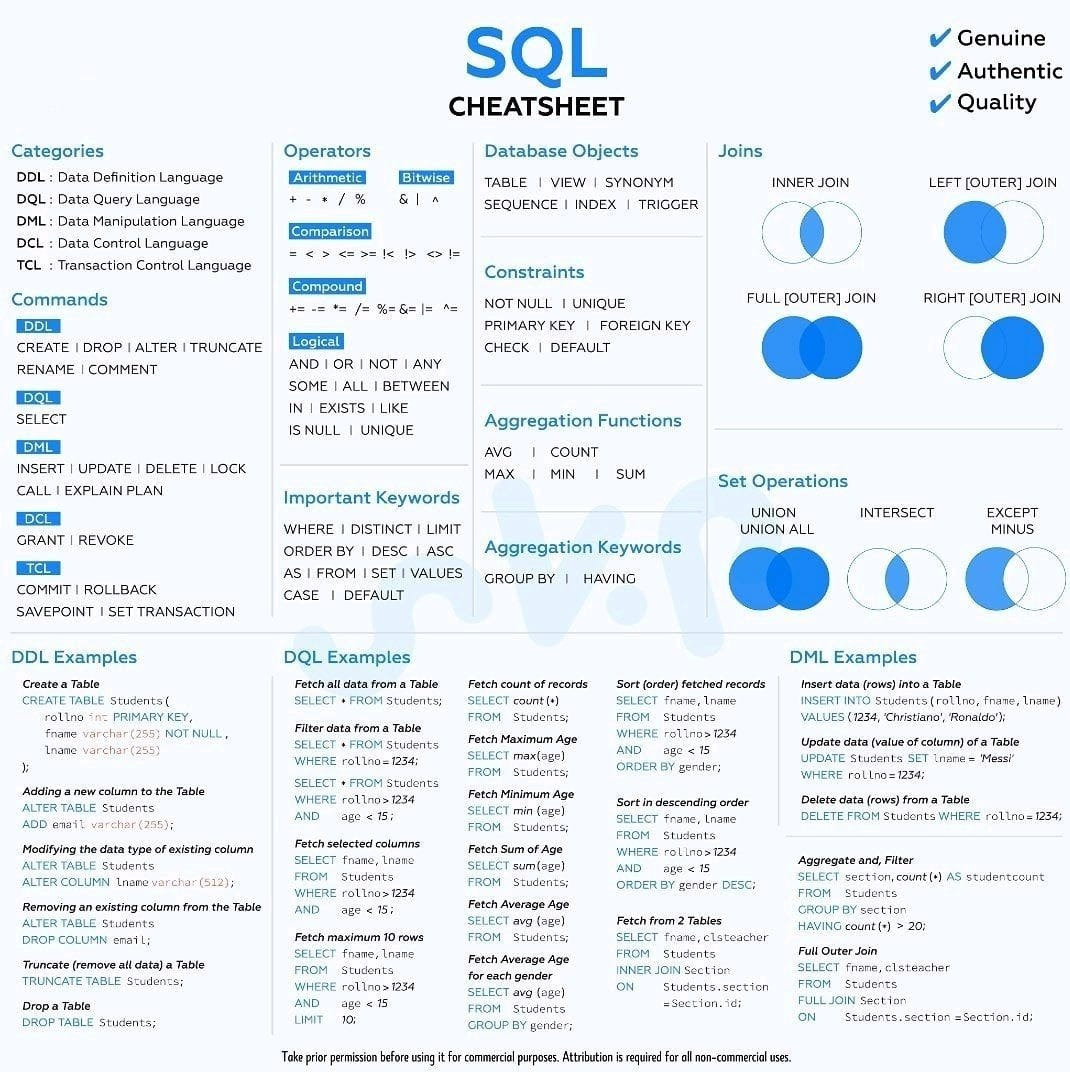
1

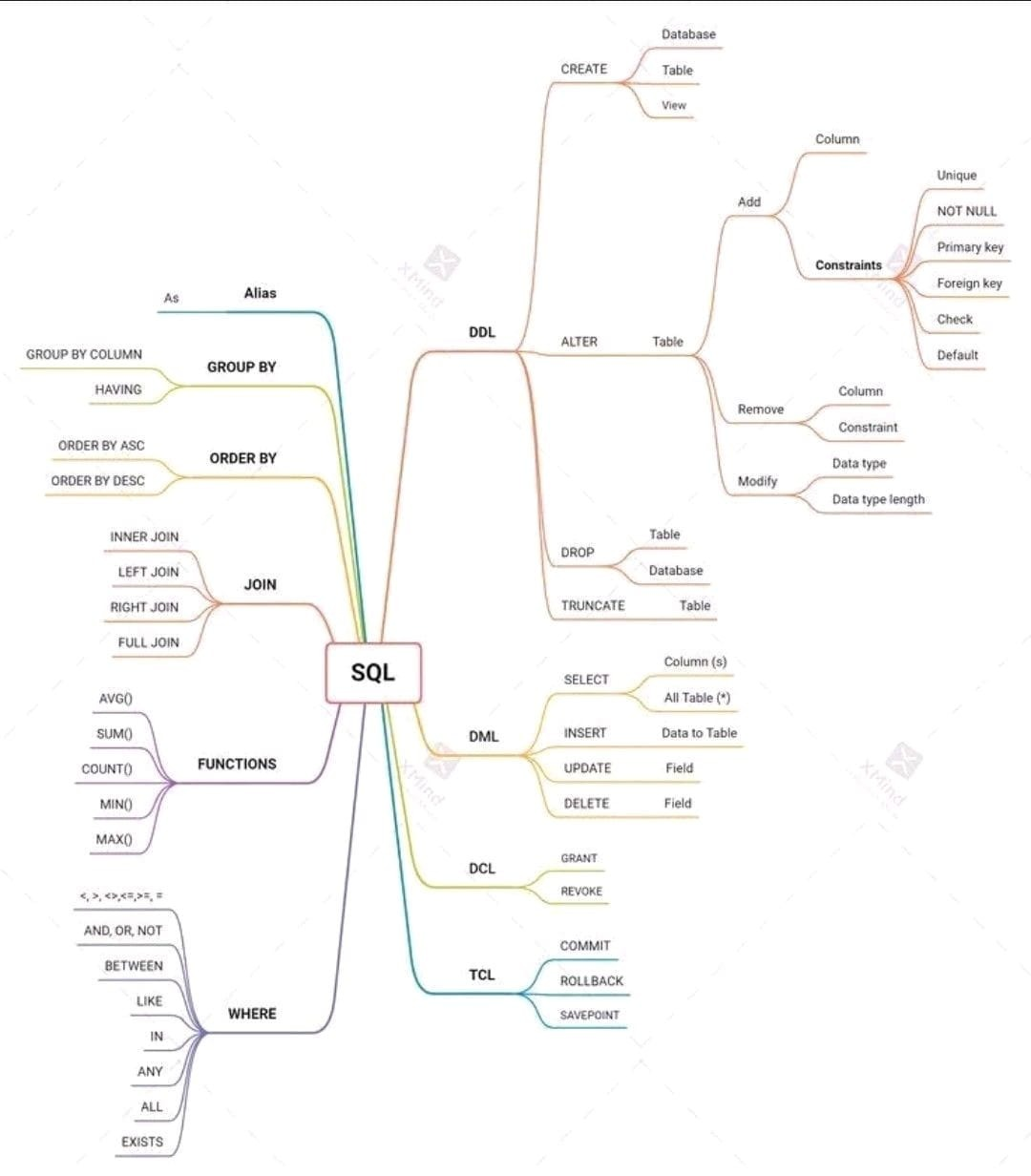
1

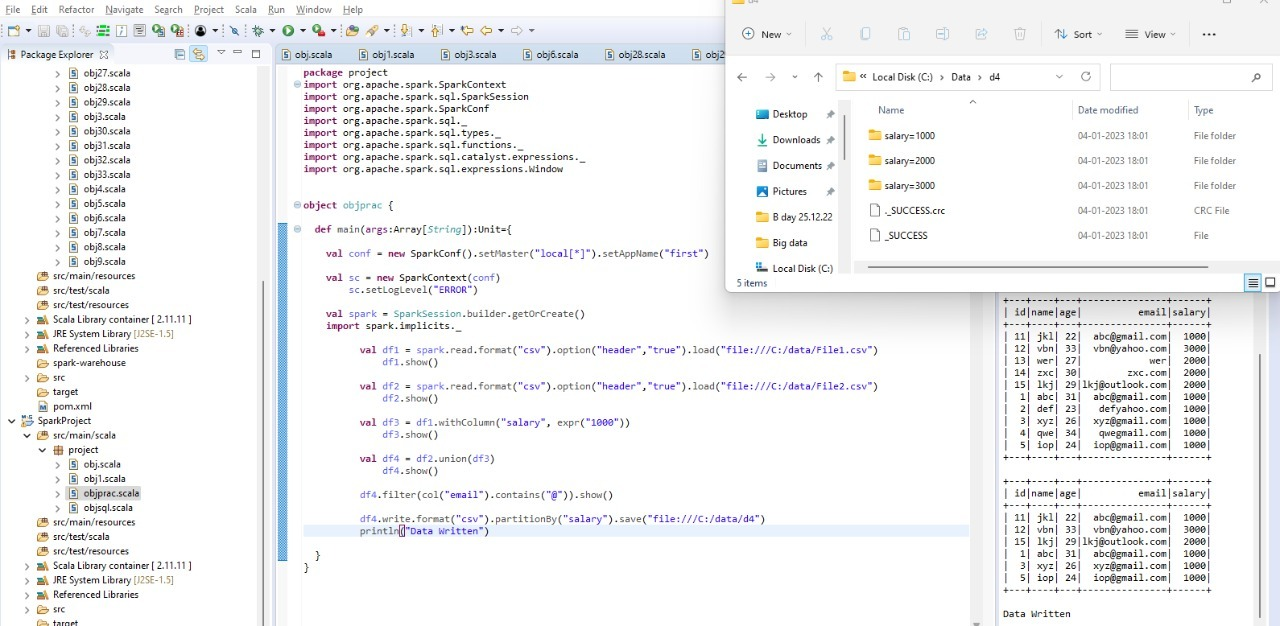
Table 2

what is the output if we use left join,right join, inner join









[23:20, 04/01/2023] +91 89784 10889: Msck repair

[23:20, 04/01/2023] +91 89784 10889: Performance tuning

[23:20, 04/01/2023] +91 89784 10889: File formats

[23:20, 04/01/2023] +91 89784 10889: What's the main difference between avro and other file formats

[23:21, 04/01/2023] +91 89784 10889: And he asked one scenario relates to SQL

[06:51, 05/01/2023] +91 73973 00987: Spark architecture, difference between data frame and data set, parquet and across file format difference, broadcast variable. Partition and buckting usage. Types of tables.

[06:52, 05/01/2023] +91 73973 00987: In sql scenarios find the second Highest salary based in the department.

Select max(salary) as salary from employees where salary not in (select max (salary) as salary from employees) ;

[06:53, 05/01/2023] +91 73973 00987: How can we combine the datas from two tables if one table is having extra column?

[06:55, 05/01/2023] +91 73973 00987: Write the word count program in Scala, and they provide an array of integer and wants to rotate the array.

WITH T AS

(

SELECT \*

DENSE\_RANK() OVER (ORDER BY Salary Desc) AS Rnk

FROM Employees

)

SELECT Name

FROM T

WHERE Rnk=2;

What is the output of the program below, given the inputs sequence of strings are

val input : Seq[String] = Seq("metadata\_column", "in\_this\_column-name", "\_index\_column\_name")

Program:

input.map{ columnName =>

val underscorePrefix : Boolean = columnName.take(1) == "\_"

columnName.split("\_").foldLeft( "" ){ (a,b) =>

if ( a.isEmpty ){

if ( underscorePrefix ) "\_" + b else b

}

else {

if ( a.takeRight(1) == "\_" ){

a + b

}

else {

a + b.capitalize

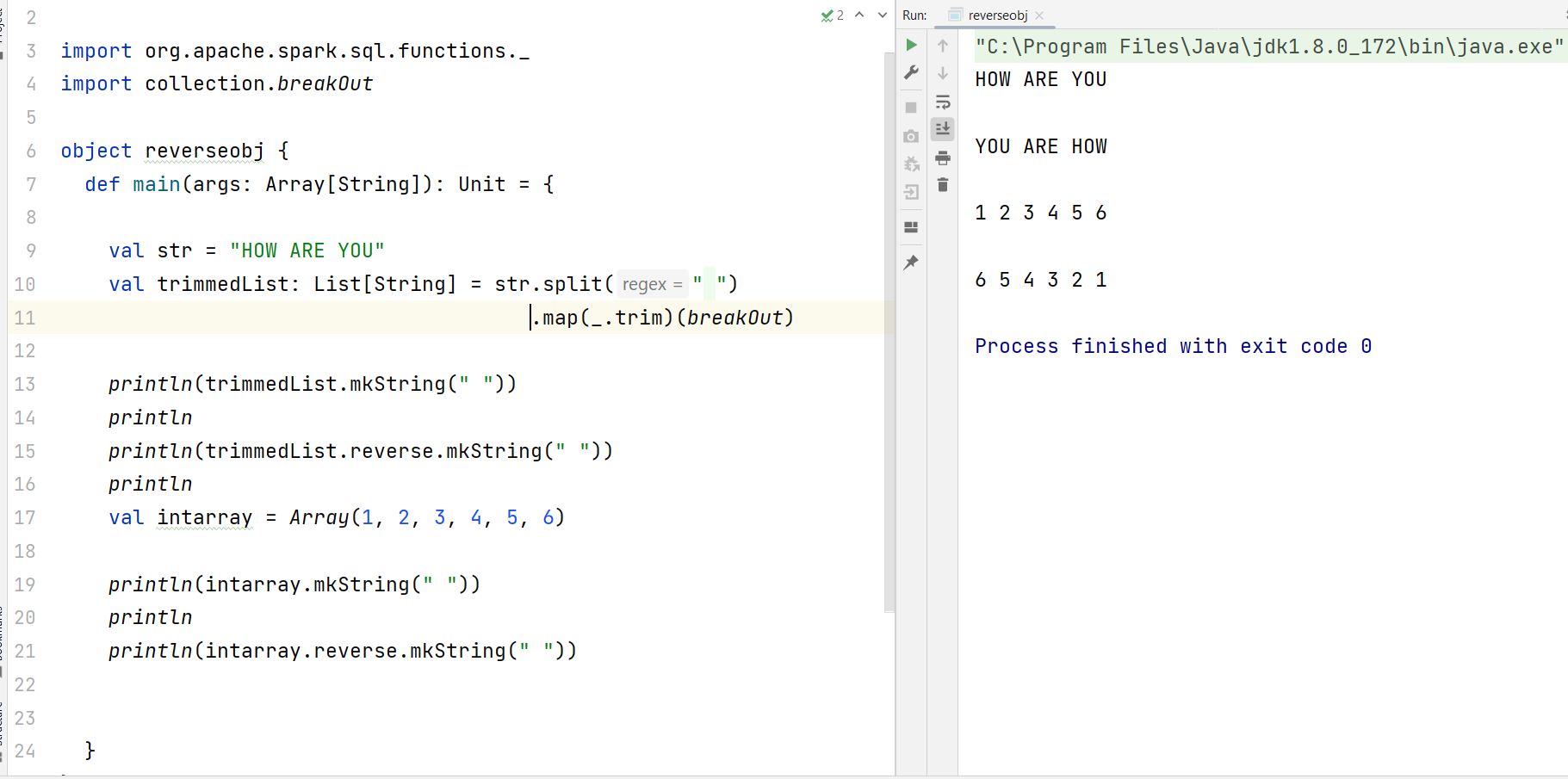
}

}

}

}

Hello, Can anyone provide the scala program to reverse string like " How are you" to " you are how"



how to insert insert data in textfile in cloudera

Having three letter code.writw a Scala code to concat one list to another list also there is a list of 3 numbers 1,2,3

Three letter list a,b,c

Expected output a1,b1,c1,a2,b2,c2,a3,b3,c3



Siemens questions

1. Explain previous project.

2. How spark submit can done for AWS or in GCP.

3. Is dataframe transformation can be possible?

4. Write a spark submit command with object name as part1 for 5 node cluster in emr.

5. How to add jars using spark context.

6. Is there any difference between Cloud storage or S3 and hdfs, which one is more efficient and cost effective.

7. Difference by rank() and dense rank()

8. What DAG and its uses, where we can find it in EMR or Dataproc.

9. Spark lineage

10. Broadcast in spark

11. Map and map transformation.

12. Why ORC is not used, as much as parquet. And explain about avro and parquet.

13. Who creates clusters in real-time is we creates or other teams,.

14. How to connect compute engine without downloading or adding ssh key?

15. What is the difference between big query and hadoop, which one is good and why?

16. To process complex data did we require any jars.

17. In own premises u have a 10 node cluster and 64 core which use intel xeon processor, if u move to the cloud will u use the same confin or there any difference.

Tcs interview 12/01/2023

Questions are

1. Project passage

2. Coalse vs repartion

3. Performance tuning in spark

4. Cache vs persistent

5. What is RDD

6. Lazy evaluation

7. Spark default format, and can we process other formats.

8. Shuffling in spark

9. Managed and external table in hive.

10. Mention one production issue faced in any of ur project.

Cache and persist

Partition and bucketing when use these

Where and having- difference

Optimization in hive

Mutable and immutable

When ur using external and managed table

In hive where it is save

Scenario - 2 nd highest salary

1.Is spark is lazy and why

2.Difference between joins and union

3. Groupby and reduceby keys

4.How to print schema of a dataframe

5.Difference between internal and external table .

6.What is meant by input split.

7.How you will delete the duplicates using ROW number = https://stackoverflow.com/questions/39564777/need-to-delete-duplicate-records-from-the-table-using-row-number

1) if spark memory in 4gb how do you process 10gb of file

2) Multiple spark session objects can be created in a single program

3) diff betwen clustering & bucketing

4) what is No hub

5) how to unzip a file using putty

6) advantages of pyspark RDD

7) What is pyspark spark session

8) how do u handled small files?

9) what is percentile in Hive

10) different type of connections in spark

11) what are implicit classes n how do they work

12) Hive window functions.

13) traits

1) Project Passage

2) Cluster size

3) Difference between RDD and Dataframe

4) how to convert RDD to dataframe

5) what is schema RDD, Row RDD, how to convert them to dataframe

6) types of Transformations, their examples

7) which actions you used in project

8) what collect does

9) broadcast, what will happen if we broadcast larger table, how to take care of it

10) how will you check whether dataframe is empty?

11) what will you prefer SQL/ DSL?

12) scenario- how you will take care of driver out of memory

13) scenario- suppose you have 100 nodes, 1 node goes down, how you will take care of it?

Today I had interview in tcs that is my first interview

1. Why scala is high functional language

2. How you will drop duplicates records in scala spark

3.how to send data from edge node to hdfs when data some data is already exit (means how u will do write)

4. What is sql analytics function

5.dense rank ,rank

6.joins

7.how to find how may nodes you in your hadoop cluster he is asking through which command

8. Coalesce and repartition

9. If we will deleted our external table how we will get same Data from hdfs to hive

10.you faced any any challenge in real time = I said about memory buffer issue

11. What is the difference between Prdd and rdd = https://data-flair.training/blogs/spark-paired-rdd/amp/

how many tables do you handle in your project? = Approximately 6-7 tables.

how to get the schema of the managed Hive table, if that data accidently gets dropped? = The data is available in trash folder for 7 days. We can get it

on 11/01/2023 TCS asked

1. Hdfs layer

2. How the data process in map reduce

3.partition and bucketing

4.external and manage table

5.wht is difference between static and dynamic

6. Lazy evaluation

7.sql view

8. Different types of database

9.tell about yourself and project explanation

11/01/2023, What is the sdlc process in u r project asked in tcs interview ,

It's software development life cycles, u should explain from scrum discussion-building code - deployment of code..everything along with testing

The 7 Stages of the Software Development Life Cycle (SDLC)

Requirements & Analysis.

Project Planning.

Design.

Coding & Implementation.

Testing.

Deployment.

Maintenance.

1.Built-in functions in hive - https://techvidvan.com/tutorials/apache-hive-built-in-functions/ = Round (), floor(), ceil()

2.Udf and udaf function in hive

UDF operates on a single row and produces a single row it's output

UDAF - works on multiple input rows and creates a single output row

3.In employee table if we apply udf in salary how many rows it will return - Only one row

4.How does data stored in spark - spark uses HDFS file system for storing. Spark doesn't have any storage because its a in memory process so that it uses HDFS to store the data.

5.Is rdd is mutable r immutable - RDD is immutable

6.is spark is lazy and why - spark is lazy beacuse transformation like narrow ,wide filter is lazy

7.cache and persist with eh

8. Syntax for cache an persisit in RDD

1.Wat is Null NIL NONE NOTHING UNIT in scala

2.Wat is the use of options in scala

3.managed table vs external table

which eninge we are using in spark for data processing?

Spark engine powered with catalyst optimizer.

1)How will RDD partitions happen basically and where are they store ?

2)Suppose if i give colesce(1) within the data frame which means I require one partition without shuffling what if there is huge data or huge size ? Can we use repartition?

val animals = List("dog","cat","deer","tiger")

val animal = "deer"

println(animals.contains(animal))

if(animals.contains(animal)){

println(animal+" is present in the list")

}else{

println(animal+" is not present in the list")

}

cognizant interview questions

1. Project Passage

2. Types od joins and a join scenario

3. what is S3, EC2 and EMR

4. how do you create a SparkSession

5. difference between SparkSession and SparkContext

6. Scenario

7. How do u deploy the code

you have a dataframe with values in one of the columns as 1 to 100 and you have a list which has value of 1,2,3,4,5.

output - in the dataframe filter the records which are present in list

logic-

using sc.parallelize - convert list of string into rdd

using toDF() - convert rdd to dataframe

once we have the dataframe ready. Use right join to filter the values

Infosys Question

Explain your project architecture

2.What re services running on the edge node -hive server, hue server

3.To filter out the Error logs from log table using pyspark.

4.Few python code as I am comfortable with python.

5.Sqoop related questions.

6.Hive questions

7.File formats

8. And few architecture based questions.

read df

student

id name marks

Write Code for

id name marks grade

>75 Grade A

>65 Grade B

<65 Grade C

CTS Interview Ques

1. SPARK API's - RDD, DF, DS

2. SPARK REPL - Read Evaluate Print loop or Spark shell

3. Do inner, left, right join for x 1,1,1 y 1,1 - output 6 records

4. HIVE ACID tables

5. Partition

6. Spark driver program

7. Repartition

8. What happens when multiple clients try to write in same HDFS file at same time?

HDFS actually works on ‘write once read many’ model. It means that only one client can write a line at a time. Multiple clients can’t write into HDFS file at the similar time.

9. Write spark code to find highest col value

a. Val df = spark.read.format.load

b. Val high = df.groupby("col")

c. .agg(max("col1")).as max\_col

d. high.show()

Yash Technology Interview questions (All Realtime only)

1) Project Passage

2) How you will fetch incremental data using spark

-> RDBS fetch every day and save in shadow table

3) how you do scheduling

4) how much time spark job takes in your project

5) How much data you process in that time

6) how many spark jobs scheduled in a day in your project

7) who defined data pipeline in your project

8) Day to day activities

9) what optimization you use in your project

10) what steps you do during EMR step execution - explain 3 steps

we have done memory tuning calculated and scheduled job

If input data is fluctuating like 10gb one other day 50 go like wise and how do we manage spark job to perfom better?

I answered we give one more parameter as dynamic resource allocation ….

Is this it or do we have something else ?

If the spark job is running in AWS EMR we can configure Auto scalling group(ASG) for executer node. So that the cluster will automatically scale up/down number of executer nodes based on data load

Which one is better ,and why

10 gb - 4 partitions each 2.5 gb

10 gb - 3 gb + 3 gb + 2 gb + 1 gb + 1 gb

[20:12, 17/01/2023] Sai Aditya Course: 4 partitions each

[20:12, 17/01/2023] Sai Aditya Course: It equally distribute to cores

[20:12, 17/01/2023] Sai Aditya Course: Or else it would lead to skewnes

Hi Team,

Few days back someone posted question In doubts session … Do cache/ persist is transformation or Action ?

Actually that day I said it as Action after some discussion.. but recently from Google I came to know that it is Transformation not Action.. kindly note this

<https://m.youtube.com/watch?v=Uia6t-J6a1g>

1)How will RDD partitions happen basically and where are they store ?

2)Suppose if i give colesce(1) within the data frame which means I require one partition without shuffling what if there is huge data or huge size ? Can we use repartition?

1. Performance tuning in hive and spark

2. What is Kafka

3. What is spark

4. Real time challenges

5. Managed vs external table

6. All kinds of joins with a table

7. dag

8. Orc vs parque, which is faster

9. Command to move from one cluster to another cluster

10. Deploy modes

Spark variable?

Spark vs mprd diffrence?

Partitioning the Data

Caching and Unpersist

Increasing Executor Memory

Spilling Data to Disk

External Data Source

HCL Interview Question

1. Explain Project

2. How to read a Web Complex data

3. Which library is using in that is it a spark or scala

4. In AWS, What is big data hadoop offering and how u done spark submit

5. What are things need to read from AWS S3.

6. what are challenges in Production u faced

7. Hive or Spark is more efficient? why

8. Can you change the block size in Spark

9. Performance tuning in spark

10. What is spark Linage

11. What is RDD and DF

12. in which language Hive is written

13. Find the second highest salary

14. Write function a scala to print n fibonacci series using Recursion.

Cognizant interview

Ur self working in bigdata

Coalese and repartion

Spark submit command

Cross join

What is map join?

Broadcast variable

What is Narrow and wide transformation your working?

Ways to process spark your working?

Types of partions and

Difference between static and dynamic?

Types of table in Hive?

Sqoop command Optimization?

1)Project explanation

2)What are the transformations you have used in spark ?

3)What are the optimization techniques in spark?

4)RDD vs Vs datasets vs dataframes ?

5)Git merge conflict. ?

6)Hive (Array, Collections)?

7)Spark arthitecture ?

8)3rd highest salary using spark sql?

9)What are the deployment steps ?

10)What node are you using?

11) Joins used in spark

Project Passage

Map Vs Flat Map

Bucketing vs Partition ``

coalsec vs partition

RDD Vs Data Frame and Datasets

Dag vs Lineage

Hive connection with spark

UDF and Wirte a example

Highest salary record from a table based on department wise

What is ORC file and why we use it

# TCS interview

# 1:- if we have to process 10 files with same code to HDFS. How we can do through spark.

2:- if we have data in text file for eg (select hey who select) how we can count streing value.. Like select is repating with two time.. Then it should

select 2

hey 1

3:- left anti join

4:- sql scenrio based

Optimization in spark?

2. What is lazy evaluation?

3. What are the types of tables in hive? what is the difference between managed and external table?

4. What is difference between group by key() and reduced by key()?

5. What is map side join in hive?

6. What is client mode and cluster mode?

7. What are Accumulators?

8. What is the use of lambda?

9. Cache and Persisting?

10. Difference between Partitioning and Bucketing?

11. How to copy files in git to the local?

12. What is gitpull?

Scenarios:-

Q1: Write a query to find the customers whose first\_name or last\_name begin with ‘R’ and ends with ‘t’ and sort the results by first\_name in descending order Customers

Q2: Delete duplicate records from table id,name,insert\_date 1,A,2023-01-10 2,B,2023-01-12 3,C,2023-01-13 2,D,2023-01-13

Q3: Find 2nd highest salary in table employee emp\_name,salary A,1000 B,2300 C,3500 D,5000 E,5000

Q4: customer\_id, first\_name, last\_name, age, country, monthly\_salary

1, John, Doe, 31, USA, 1000

2, Robert, Luna, 22, USA, 2000

3, David, Robinson, 22, UK, 2500

4, John, Reinhardt, 25, UK, 1500

5, Betty, Doe, 28, UAE, 1200



Is there any difference between container and executor

CTS

1)In scala , create an abstract class and take two interfaces and overwrite it .

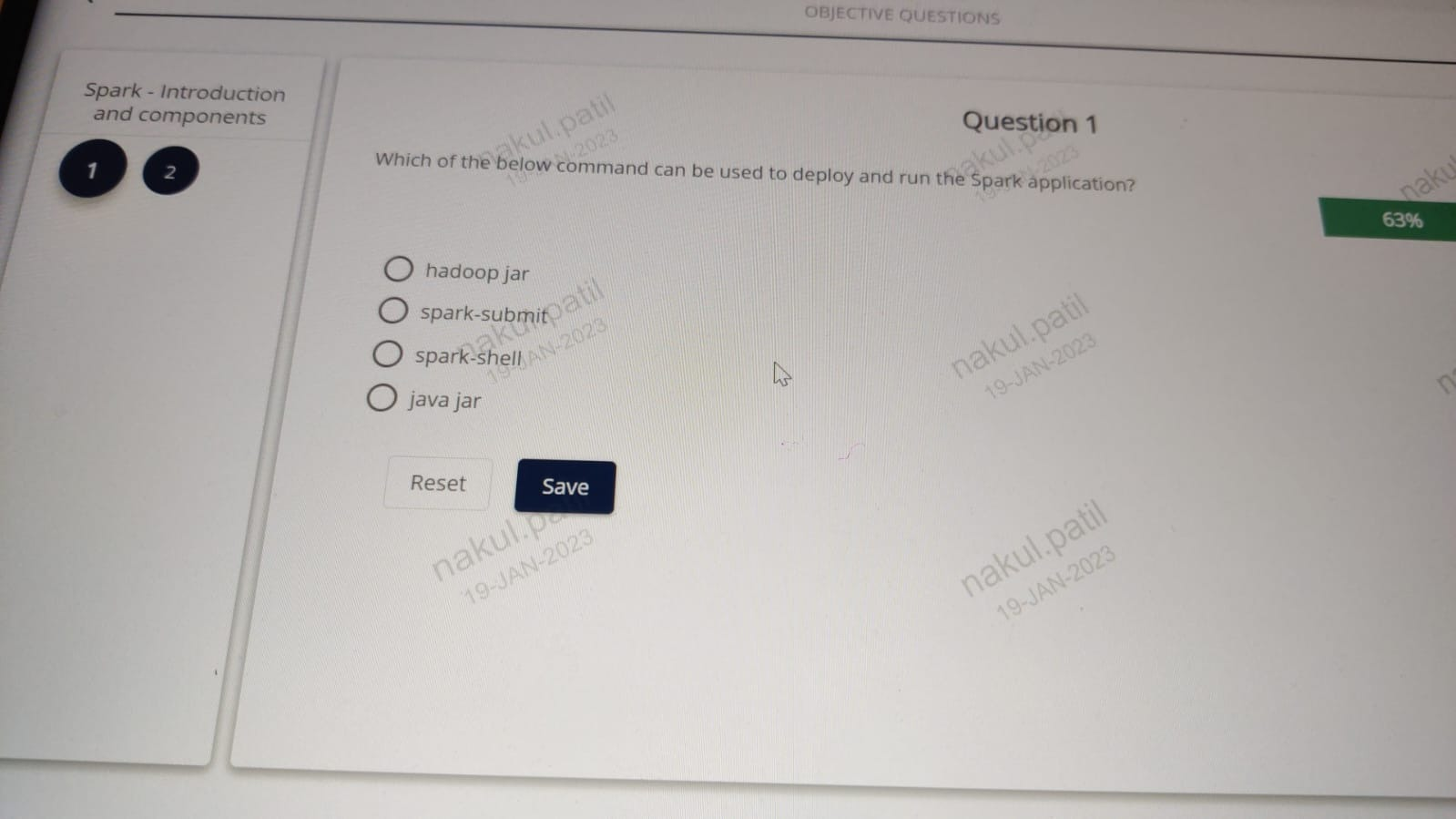
2)

List (filpkart, google, 1234, amazon , 6644, dell)

Remove integers and print strings with first letter starts with Capital

Output should be Flipkart, Google, Amazon, Dell

3)In a CSV file it contains only id and name columns and I'd column doesn't contain any values and name column is having some values Now generate random values in I'd column and print it.



IBM Questions:

1. project Passage

2. Managed ad external tables in HIVE

3. Default table location of managed table in HIVE

4. Partition in the hive and how we can select partition column is the specific one

5. Bucketing in the hive

6. why can't use partitions in RDBMS

7. difference b/w Hive and RDBMS

8. CREATE table in hive command and write partition for that

9. Spark Architecture

val df1 = spark.sql("select \* from table1")

val df2 = spark.sql("select \* from table2 ").filter("filter condition")

val df3 = df2.groupby("city").agg("") // some agg functions

val df4 = df3.join(df1, Seq("id") )

val df5 = df4.union(df3)

df4.saveastable("")

df5.count()

project Passage

Driver Memory vs Excetuor memory?

how do you set extra options to spark-submit in AWS?

How do you check if any Spark Job is failed in AWS?

Where can we see yarn log files in AWS?

what will you do if any spark job fails in AWS?

AVRO vs Parquet?

What is the AVRC file in Avro file format and where we can find it in HIVE ??

Find a list of prime numbers using scala from the below list

x = list ( 15, 17,19,20,25)

1. Explain you’re current project and you’re roles and responsibilities in the project?

2.What are the considerations has to take while data migration.

3. Explain data migration and processing pipeline in you’re project?

4.Which ETL tools, technologies will use and explain why?

5.What is Spark and explain spark architecture? and convert into you’re project?

6.Write a Spark code to create a new column by applying operations on existing columns in Data Frame

7.How to write Data Frame? explain different modes to write it?

8.How to rename the columns on Data Frame?

9.Please explain difference between coalesce and repartition? have you used in project and how?

10.What is Skewness? how to resolve it?

11.Explain Spark optimization techniques?

12.What is broadcast variable? explain with example?

13.What is different types of joins in Apache Spark? explain classical joins?

14.What is broadcast join? explain sort merge join in Spark?

15.How to remove duplicates from Data Frame?

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

17.Explain User Defined Functions (UDF) 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 techniques in Spark?

20.Scenari based question: There are 2 Data Frames emp, department and write a code to join them simply?

21. What is Spark session, how it is initialize?

22. What are the issues you have faced in you’re project and how you resolved those

What is serialization? ...

How to remove the duplicate records from a hive table? ...

How to find the number of delimiter from a file? ...

Replace a certain word from a file using Unix? ...

How to import a table without a primary key? ...

How to write a UDF in Hive?

Could u please tell ans for

- Considerations has to take while data migration

- which ETL tools or technologies will use and xplain why

- broadcast variable and explan with example

- if created a new column and gave same name for it which already exists then what will happen.

- there are 2 dataframes (emp , department) write code to join them simply ?

- what is serialisation

- replace a certain word from file using Unix ?

How to import table without primary key

How to remove duplicates from hive table

How can we write a script for migrating data from hdfs to ozone?

Write a spark session and read csv file and have to merge two tables, one df have only 2 column and another df is three columns

b) Max salary by deptment wise in SQL.

c) Write a hadoop command write data from local to clusterpath.

d) how to find a string in hadoop or particular file.

e) What is scheduling and why using?

f) Hive Partition and Bucketing

g) optimization in sparks

e) Asked about ETL scenario questions as well

Coforge interview questions 1st half basic question included

cache, persist

Tuple

What is Serde

Types of table in hive

Partiotioning vs Bucketing

Higher order function

Why external table in HIVE

Explain Lineage graph

What is lazy evolution

Difference between spark and mapreduce

What is action in spark etc

Narrow transformation and wide transformation

2nd half

Program question included

Fetch re ord from employee table having columns manager id, employee id and need to fetch how many people reporting to the particular manger and total employees count

Spark text file read and find one record from 10000's of record

a) Write a spark session and read csv file and have to merge two tables, one df have only 2 column and another df is three columns

b) Max salary by deptment wise in SQL.

c) Write a hadoop command write data from local to clusterpath.

d) how to find a string in hadoop or particular file.

e) What is scheduling and why using?

f) Hive Partition and Bucketing

g) optimization in sparks

e) Asked about ETL scenario questions as well

Coforge

1st half basic question included

cache, persist

Tuple

What is Serde

Types of table in hive

Partiotioning vs Bucketing

Higher order function

Why external table in HIVE

Explain Lineage graph

What is lazy evolution

Difference between spark and mapreduce

What is action in spark etc

Narrow transformation and wide transformation

2nd half

Program question included

Fetch re ord from employee table having columns manager id, employee id and need to fetch how many people reporting to the particular manger and total employees count

Spark text file read and find one record from 10000's of record

nterview Question by CTS:

Components of Spark

Components of Hadoop

Spark partition

Spark bucketing

Cognizant 1st round interview questions on 12/01/203

1. Difference between RDD and dataframe

2. Given two tables and asked about joins

3.Spark variable

4. How to read CSV file

5. How to write CSV to give table

cognizant interview questions DATE 17/01/2023 BY ZAID

1. Project Passage

2. Types od joins and a join scenario

3. what is S3, EC2 and EMR

4. how do you create a SparkSession

5. difference between SparkSession and SparkContext

6. Scenario

7. How do u deploy the code

you have a dataframe with values in one of the columns as 1 to 100 and you have a list which has value of 1,2,3,4,5.

output - in the dataframe filter the records which are present in list

logic-

using sc.parallelize - convert list of string into rdd

using toDF() - convert rdd to dataframe

once we have the dataframe ready. Use right join to filter the values

1. SPARK API's - RDD, DF, DS

2. SPARK REPL - Read Evaluate Print loop or Spark shell

3. Do inner, left, right join for x 1,1,1 y 1,1 - output 6 records

4. HIVE ACID tables

5. Partition

6. Spark driver program

7. Repartition

8. What happens when multiple clients try to write in same HDFS file at same time?

HDFS actually works on ‘write once read many’ model. It means that only one client can write a line at a time. Multiple clients can’t write into HDFS file at the similar time.

9. Write spark code to find highest col value

a. Val df = spark.read.format.load

b. Val high = df.groupby("col")

c. .agg(max("col1")).as max\_col

d. high.show()

DATE 18/01/2023

Cognizant interview

Ur self working in bigdata

Coalese and repartion

Spark submit command

Cross join

What is map join?

Broadcast variable

What is Narrow and wide transformation your working?

Ways to process spark your working?

Types of partions and

Difference between static and dynamic?

Types of table in Hive?

Sqoop command Optimization?

DATE 19/01/2023

1)In scala , create an abstract class and take two interfaces and overwrite it .

2)

List (filpkart, google, 1234, amazon , 6644, dell)

Remove integers and print strings with first letter starts with Capital

Output should be Flipkart, Google, Amazon, Dell

3)In a CSV file it contains only id and name columns and I'd column doesn't contain any values and name column is having some values Now generate random values in I'd column and print it.

. What issues you get in sqoop import and how you debug?

2. How password is managed in production environment and is password file good option for production.

3. How can we find number of stages in query

4. In your project you have sqoop,hive,pyspark - how do you integrate all these components

5. Dataframe - prod, city, sold

Output

Kolkata, 50

Pune, 20

Total, 70

Coforge

1st half basic question included

cache, persist

Tuple

What is Serde

Types of table in hive

Partiotioning vs Bucketing

Higher order function

Why external table in HIVE

Explain Lineage graph

What is lazy evolution

Difference between spark and mapreduce

What is action in spark etc

Narrow transformation and wide transformation

2nd half

Program question included

Fetch re ord from employee table having columns manager id, employee id and need to fetch how many people reporting to the particular manger and total employees count

Spark text file read and find one record from 10000's of record

Conclusion

Enable ACID Transaction Manager (DbTxnManager) on hive session.

Enable Concurrency.

Create Table by enabling transactional (TBLPROPERTIES ('transactional'='true'))

Create Table with ORC storage type.

Insert data into a table using INSERT INTO.

Finally, Run UPDATE and DELETE HiveQL queries on the table.

1)Project passage?

2)What is Rdd ?

3)What are the Issues you faced in data ingestion (rdbms to hive using sqoop ) ??

4)What are the production issues you have faced ?

5)What is your source ? (MySQL database)

6)What is the Parquet file format?

7)In Scala what are the libraries you have used ??

8)How to find the top 2 ranks in the table.?

9)What are AWS: DEPLOYMENT STEPS ??

10)What are the LAUNCH MODES IN AWS CLUSTER (persist and transient)??

11)What are classes in S3 storage ??

12)How do you estimate User stories ??

13)What are the things you take into consideration while doing development.. ??

14)Day to day activities??

1)How will RDD partitions happen basically and where are they store ?

2)Suppose if i give colesce(1) within the data frame which means I require one partition without shuffling what if there is huge data or huge size ? Can we use repartition?

What is advantage of deploy mode - cluster

Emp\_id, Commission\_Amt, Month\_last\_Date

1 , 300,3 1-Jan-2021

1, 400, 28-Feb-2021

1, 200,31- Mar-2021

2, 1000, 31-Oct-2021

2, 900 ,31-Dec-2021

-Output

1, 200,31-Mar-2021

2, 900,31-Dec-2021

select Emp\_id,min(Commission\_Amt) as amount, Month\_last\_date from table groupBy(Emp\_id)

Out of memory

When you submit in client mode and any performed collect action may lead lead to OOM in of client node if that we can’t process any other transaction

If the same happens in cluster mode submit then that specific slave node will be affected

Can we write two different files formats with different data in the same directory ? Can we use append mode here?

. We have employee table with

Name. Dept. Salary

Write 2 highest salary in spark Scala for each dept

Fractal interview questions:

1. How you handle insert/upsert/ delete of data using joins

2. Fibonacci series logic using scala programming

3. Rank, Dense rank, Row number differentiation with sample data

4. Table:

ID, Name, Salary, Manager ID

1, Joe , 10000, 3

2, Mark, 20000, 4

3, Kim, 40000, NULL

4, Mike, 25000, NULL

OUTPUT: Name, Manager Name

Joe, Kim

Mark, Mike

5. Spark memory optimiser

6. Which scheduling tool you are using in your project

7.

Output= 1,2,3,3,5,5,6,7,8,9,10,11,12

1. What issues you get in sqoop import and how you debug?

2. How password is managed in production environment and is password file good option for production.

3. How can we find number of stages in query

4. In your project you have sqoop,hive,pyspark - how do you integrate all these components

5. Dataframe - prod, city, sold

Output

Kolkata, 50

Pune, 20

Infogain :

What should you do if you get a symantic error while inserting data into a partition table?

Can someone send code with col operation 2 highest salary for dept wise

Select \* from table\_name ORDER BY Commission\_Amt DESC LIMIT 2 OFFSET 3;

A friend of mine gave data engineering interview for a banking company. He was asked to explain how to construct data pipe line in Scala for ETL. Anyone can give answer to this question.

spark.sql("select e.Name,m.Name as ManagerName from df e join df m on e.ManagerID=m.ID").show() +----+-----------+

|Name|ManagerName|

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

| Joe| Kim|

|Mark| Mike|

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

HI Guys,Incedo Interview Questions:

1. Optimization in spark?

2. What is lazy evaluation?

3. What are the types of tables in hive? what is the difference between managed and external table?

4. What is difference between group by key() and reduced by key()?

5. What is map side join in hive?

6. What is client mode and cluster mode?

7. What are Accumulators?

8. What is the use of lambda?

9. Cache and Persisting?

10. Difference between Partitioning and Bucketing?

11. How to copy files in git to the local?

12. What is gitpull?

Scenarios:-

Q1: Write a query to find the customers whose first\_name or last\_name begin with ‘R’ and ends with ‘t’ and sort the results by first\_name in descending order Customers

Q2: Delete duplicate records from table id,name,insert\_date 1,A,2023-01-10 2,B,2023-01-12 3,C,2023-01-13 2,D,2023-01-13

Q3: Find 2nd highest salary in table employee emp\_name,salary A,1000 B,2300 C,3500 D,5000 E,5000

Q4: customer\_id, first\_name, last\_name, age, country, monthly\_salary

1, John, Doe, 31, USA, 1000

2, Robert, Luna, 22, USA, 2000

3, David, Robinson, 22, UK, 2500

4, John, Reinhardt, 25, UK, 1500

5, Betty, Doe, 28, UAE, 1200

Extract transfer and load first we extract data with the help of Sqoop or Kafka then we process with the help of spark and load it in hdfs or hbase

[19:56, 25/01/2023] +91 95894 94080: CSV File

C1,C2

11,D|E

223,D

355,A|B|E

664,C|B|E

Output:

D1,D2

A,1

B,2

C,1

D,2

E,3

[19:57, 25/01/2023] +91 95894 94080: we find output using c2 column with spark

In Hive, performance wise which table is good whether Internal Or External... - I think external more powerful than internal bcoz you know drop column possibility banefits,also connected by location prefer.

In hive, Once u drop some data in table and that table not working properly.. Why and how u will resolve? He actually, answered like need to correct the statistics...

Does parquet perform better than ORC in Hive?

1st was selected 2nd round tec they ask simple id,name,value = i used inner join but that interview person told we can use groupby also

100,a,25

101,b,20

100,a,55

Out put:

id,name,value

100,a,80

101,b,20

What is polymorphism?

val y = b.withColumn("rank",dense\_rank() over(Window.partitionBy(col("salary")).orderBy(col("salary").desc))).where(col("rank") === 2)

Spark - Questions

Which of the following Spark properties is used to configure the broadcasting of a df without the use of

the broadcas() Operation

1.Spark.sql.autoBroadcastjoinThreshold

2.spark.sql.broadcastTimeout

3.spark.broadcast.blocksize

4.spark.broadcast.compress

Which of the following code will not always return the excat number of distint values in column Divison?

1. sdf.agg(approx\_count\_distinct(col("Divison")).alias("DivisionDistint"))

2. sdf.agg(countDistinct(col("Divison")).alias("DivisionDistint"))

3. sdf.select("divison").dropDuplicate().count()

4. sdf.select("divison").distinct().count()

In spark Jobs, what is the need of Checkpointing

1. To recover Driver From Failure

2. To recover Exceutors from failure

3. To recover both driver and excutors

4. None of the above

What is the maximum partistion size used for shuffle block?

1. 1GB

2. 2GB

3. 256 MB

4. 3GB

In Spark SQL, what would be the output of below:

SELECT true <=> NULL;

1. True

2. False

3. Synatax Error

4. None of the above

What happens when we use cache in spark jobs?

1. 2 seperate jobs created. one for regular and other for caching

2. Caching and normal execution will happen in same job but different stages will be created

3. Two seperated stages will be created

4. Two different tasks will be created

What happens when we register Datafram as temporary table or view

1. Materialized view is created in Memory

2. Pointer to the data is created in Memory

3. It starts caching the data on diske and register meta data in memory

4. nothing happens

While reading data from file placed on localfile system, Which property will help in deciding number of partition

1.spark.default.paralleism

2.spark.sql.files.maxPartition Bytes

3.Both 1& 2

4. None of the above

What will happen in following scenario when complete path is scanned everyday.

Base Path: /Path/to/parquet

Under a folder data is partitioned on date basis

What will happen in following scenario when complete path is scanned everyday: date=2022-01-02: Schema- name, city, ten time,com\_amount

Pick One Option

val readDF = spark.read.option("mergeSchema", "true")("/path/to/parquet")

1.Program will break as it is unable to read dasa due to conflict schema

2.It will read succes fully but it will eliminate in time as it is extra column

3.It will read succesfully and schema will be in order: name,city,txn\_amount,txn\_time..

4.It will read successfully and schema will be in order : name,city,txn\_time,txn\_amount

AWS

What is the largest size of object that can be uploaded to AWS in single upload?

1.5TB

2.5GB

3.10GB

4.10TB

Which configuration is most optimized while allocating instances of EMR's Core and Task nodes? CORE-SPOT, TASK SPOT

1.CORE - On Demand, TASK - On Demand

2.CORE SPOT, TASK SPOT

3.CORE- On Demand, TASK SPOT

4.CORE SPOT, TASK On Demand

Is AWS S3 offers regional buckets?

1. Yes

2. No

Out of the following which protocol is fastest and secure while interacting with S3

1.S3://

2.S3a://

3.S3n://

4.S3s://

Hive

Truncate table command will drop both data and schema for a managed table?

1.Yes

2.No

There's an employee table with 10 columns in total including "employee\_dept" and "employee\_designation" as partition columns. In the mentioned optio in hdfs location?

1.Under table name folder, files containing rows with data for all 10 columns

2.Under table name folder, nested folders of partitioned columns and then files containing rows with data for all 10 columns

3.Under table name folder, nested folders of partitioned columns and then files containing rows with data for only 8 columns which are not partitioned

4.There's no table folder for partitioned tables and directly partitioned folders are stored in DB folder with data for all 10 columns in each partitioned folder

Hadoop / HDFS

Suppose you want set mapreduce.map.memory.mb=4096mandindo Beduce.reduce.memory.mb-6096mb,Which of the following files should contain these settings?

1.core-site.xml

2.hdfs-site.xml

3.mapred-site.xml

4.Yarn-site.xml

Resource Manager and History server runs on which port?

1.RM on 8088, History server on 18080

2.RM on 8098, History server on 18008

3.RM on 8008, History server on 18088

4.RM on 8068, History server on 18088

SQL (NOT SPARK)

Election Exit Poll Distribution Report needs a list of US states with candidates and their vote counts.As part of HackerPoll's election exit poll analytics, a team needs a list of US states with The result should be in the following format state, votes.

candidates and their vote counts.

-> votes is a comma-separated list of candidate full names (candidates first name, The result should be in the following format: state, votesdate in a particular state.

sort the candidates descending by the number of votes, and ascending by name.

Be careful to format the data exactly as shown. There is a single space between the first and Results last name and on both sides of the 'x'

SCALA

Scala has singleton objects, which are essentially classes with only one instance.

1.True

2.False

What is the maximum number of arguments a case class can have?

1.22

2.100

3.33

4.No limit

PROGRAMME

Select one question from 23 to 29, Choose your language of intersets

An array of n positive integers, arr[n] can be partitioned into k number of subsegments that all have an equal sum and subsegments completely covers the array a[n].

Find all the possible substrings

For example : Consider an array

Array arr = {7,3,5,12,2,1,5,3,8,4,6,4}

number of segemnet k = 3

then

S1 = {1,2,3,4,4,6}

S2 = {5,7,8}

S3 = {3,5,12}

Selected Language of interset : C++

Function Description

Complete the Function getArraySubsegments in the editor below. The Function must state what must be

returened or printed .

getArraySubsegments has the following parameter(s):

number[numbers[0],....,numbers[n-1]]: an array of integers

numberofSegmanet : integer

Constraints

1. i<=k<=n

2. o<=arr[i]<=10000

3. Elements should be in sorted form

<https://leetcode.com/problems/employees-earning-more-than-their-managers/description/>

<https://leetcode.com/problems/delete-duplicate-emails/solutions/?envType=study-plan&id=sql-i>

sqoop import --connect jdbc://teradata\_host:port/db\_name --username xyz --passwosrd-file path --hive-import --hive-database --hive-table tableName

Sample input: x =['Null', 'Null', 1, 2, 'Null','Null', 'Null', 3, 4, 'Null', 5, 'Null', 'Null']

Sample output: x =['Null', 'Null', 1, 2, 2, 2, 2, 3, 4, 4, 5, 5, 5]

sample\_output=[]

for i in range(len(sample\_input)):

if sample\_input[i] != 'Null':

sample\_output.append(sample\_input[i])

elif sample\_input[i] == 'Null':

if

sample\_output.append(sample\_input[i])

else:

sample\_out.append(sample\_input[i-

city1 city2 distance

bangalore kolkata 200

Kolkata Bangalore 200

Chennai Bangalore 100

Bangalore Chennai 100

Bangalore Pune 100

select \* from t1 where city1 < city2 union all

select \* from t1 where city1 > city2 and not exists(select 1 from tableName t2 where t2.city1 = t1.city2 and t2.city2 = t1.city1)

Tiger Analytics Questions:

1.DE

a. What do you look in the Spark UI to identify bottlenecks?

Sol: The performance bottlenecks are identified using Stages view in Spark UI

b. Can we perform update/delete on Hive table? How to update values in Managed and External table?

2.Scala

1. Given a String "Now a days big data analytics is used a lot"

Find the Largest string in the given statement

2. Write a UDF to convert first letter to Capital case from given input

3. Spark

1. Use lamba function and sort the list of tuples with Hieght > specific value

list =(("john",150),.....)

2. Given a table change the headers/column names

from

Emp ID | Emp Name | Dept

to

emp\_id | emp\_name | dept

Replace space to underscore and lower case the column name

4. SQL queries based on aggregates mostly

(Cant recall)

Coforge Interview questions

1.Tell me about your current project

2 . What's the client requirements for this project

3 . When u r using on premises where you downloaded hive?

4.How u process the data in spark.

Coalesce vs reparation

Broadcast variable

Cache vs persist

Parquet file format

Orc file format

What is enabled in hive?

Coforge

1: broadcast variable

2: difference between cloudera and aws

3:what is shuffling

4:do you have install any softwares while using cloudera

5: without scal spark works or not

6: memory tunning

7: project passage