

PRACTICAL EXAM BDAV

Name: zubair Isaque kazi

Roll No: 25

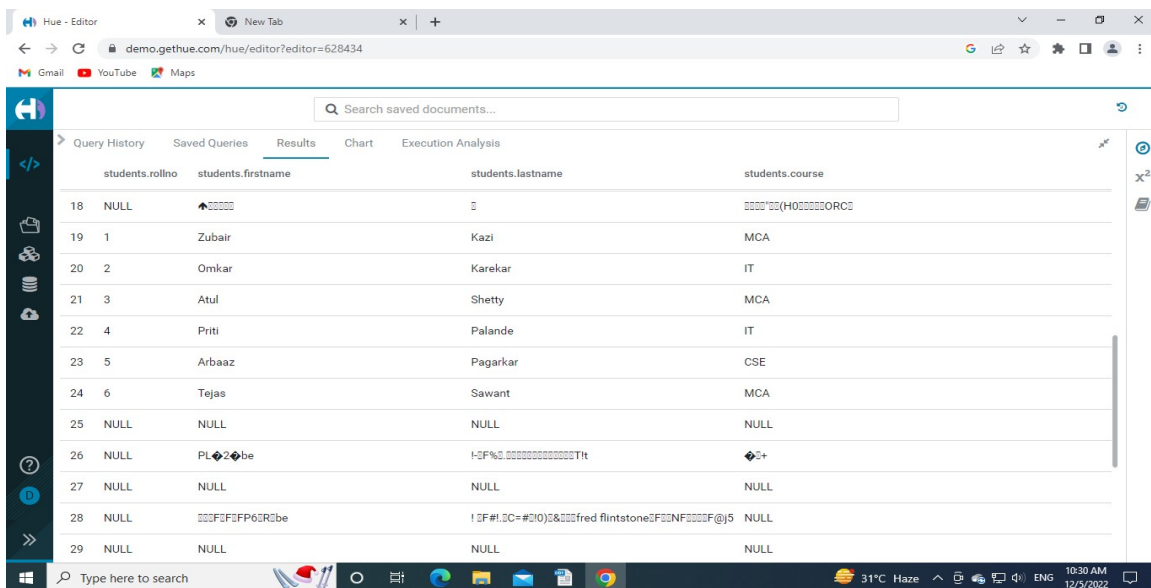
A) Using Student structure do the following task in HIVE

→ Creation of Table Students

1. create table students (Rollno int, Firstname string, Lastname string, course string)

→ Inserting of Data into tables

2. INSERT INTO students VALUES ("01",'Zubair','Kazi','MCA'),
("02",'Omkar','Karekar','IT'),
("03",'Atul','Shetty','MCA'),
("04",'Priti','Palande','IT'),
("05",'Arbaaz','Pagarkar','CSE'),
("06",'Tejas','Sawant','MCA');



students.rollno	students.firstname	students.lastname	students.course
18	NULL		
19	1	Zubair	Kazi
20	2	Omkar	Karekar
21	3	Atul	Shetty
22	4	Priti	Palande
23	5	Arbaaz	Pagarkar
24	6	Tejas	Sawant
25	NULL	NULL	NULL
26	NULL	PL 2 be	+
27	NULL	NULL	NULL
28	NULL	FP6SR3be	! F#1.3C=#210)3&333fred flintstone3F32NF333F@j5
29	NULL	NULL	NULL

→ Displaying of Table Data

3. select * from students;

→ Prepare a the list of all the course and their respective number of student studying for that course

SELECT count(*) as total_record FROM students;

→ Displaying the full name and roll no of those student who is studying in the MCA Course

The screenshot shows the Hue web interface for Hive. The main query editor contains the following SQL query:

```
SELECT * FROM students WHERE course='MCA';
```

The query has been executed, and the results are displayed in a table with 3 rows. The table has 5 columns: students.rolno, students.firstname, students.lastname, students.course, and students.rollno.

	students.rolno	students.firstname	students.lastname	students.course	students.rollno
1	1	Zubair	Kazi	MCA	
2	3	Atul	Shetty	MCA	
3	6	Tejas	Sawant	MCA	

The right sidebar shows the 'Tables' section with a list of tables and their data types:

- Filter...
- default.students
 - rolno: int
 - firstname: string
 - lastname: string
 - course: string

The 'Query Analysis' section shows the query being executed: `Query doing a SELECT *`. It also provides a tip: `Select only a subset of columns instead of all of them` and a link to `Fix me`.

```
- CREATE view students select * FROM students WHERE rollno>"03";
- CREATE VIEW roll_2 AS SELECT rollno FROM course WHERE rollno>"02";
```

```
- CREATE INDEX students ON TABLE students (Rollno, ...) AS 'index.handler.class.name'
[WITH DEFERRED REBUILD] [IDXPROPERTIES (property_name=property_value, ...)] [students]
[PARTITIONED BY (course, ...)]
```

	student.rollno	student.firstname	student.lastname	student.course
1	11	ashwin	Bhavana	MCA

////////////////////////////////////

→ Firstly we need to start Hadoop system

- 1 - start-dfs
- 2 - start-yarn
- 3 - jps

1 - hadoop version

2 - Creating of new Directory

- `hadoop fs -mkdir /newDirectory`

```
C:\windows\system32>hadoop fs -ls /
'C:\Program' is not recognized as an internal or external command,
operable program or batch file.
Found 4 items
drwxr-xr-x   - root      hadoop           0 2022-02-06 20:45 /datasets
drwxrwxrwx   - jinoy    supergroup      0 2022-02-06 20:27 /jinoy
drwxr-xr-x   - student  supergroup      0 2022-12-05 10:15 /newDirectory
drwxr-xr-x   - hp       supergroup      0 2022-11-23 12:52 /pigdata
```

3 - Listing out the Directories

- `hadoop fs -ls /`
- `hadoop fs -ls -R/`

4 - Creating of file

- `hdfs dfs -touchz /newDirectory/myNewFile.txt`

```
C:\windows\system32>hdfs dfs -ls /newDirectory/
'C:\Program' is not recognized as an internal or external command,
operable program or batch file.
Found 1 items
-rw-r--r--   1 student  supergroup      0 2022-12-05 11:29 /newDirectory/myNewFile.txt
```

5 - Checking last modified time of current directory

- `hdfs dfs -ls -u /newDirectory`

```
C:\windows\system32>hdfs dfs -ls -u /
'C:\Program' is not recognized as an internal or external command,
operable program or batch file.
Found 6 items
drwxr-xr-x   - root      hadoop           0 1970-01-01 05:30 /datasets
drwxrwxrwx   - jinoy    supergroup      0 1970-01-01 05:30 /jinoy
drwxr-xr-x   - student  supergroup      0 1970-01-01 05:30 /newDirectory
-rw-r--r--   1 student  supergroup      0 2022-12-05 10:33 /newFie
-rw-r--r--   1 student  supergroup      0 2022-12-05 10:32 /newFie.txt
drwxr-xr-x   - hp       supergroup      0 1970-01-01 05:30 /pigdata
```

6 - HDFS command to get any help for any command on HDFS

- `hadoop fs -ls/`
- OR
- `hdfs dfs -help`

c) Create a dataset of Customers{cust_id, cust_name, cust_add, cust_contact_no) process the same on apache spark

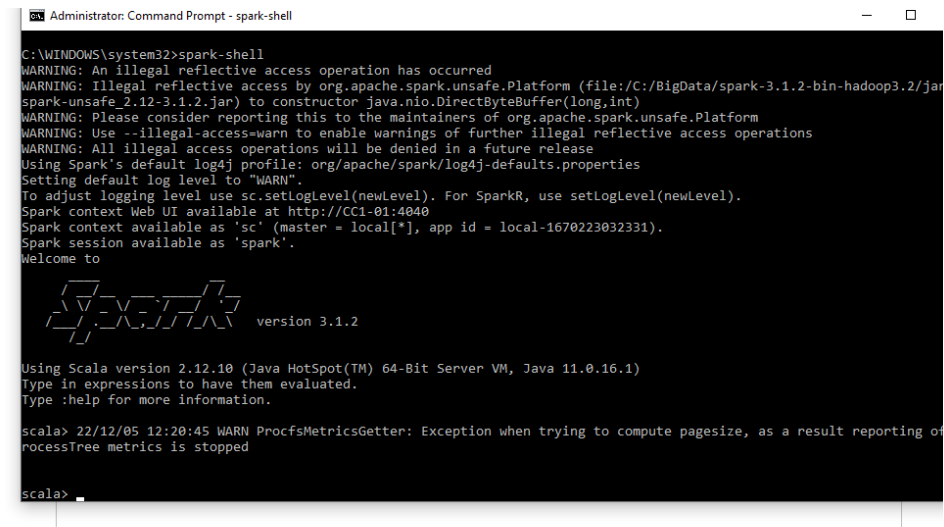
→ Firstly we need to start Hadoop system

Start Hadoop

- 1 - start-dfs
- 2 - start-yarn
- 3 - jps

Run Sparks cmd

- 1) To get into Spark
 - spark-shell



```
C:\WINDOWS\system32>spark-shell
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/C:/BigData/spark-3.1.2-bin-hadoop3.2/jar
spark-unsafe_2.12-3.1.2.jar) to constructor java.nio.DirectByteBuffer(long,int)
WARNING: Please consider reporting this to the maintainers of org.apache.spark.unsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Spark context Web UI available at http://CC1-01:4040
Spark context available as 'sc' (master = local[*], app id = local-1670223032331).
Spark session available as 'spark'.
Welcome to

  ____
 /    \ Spark version 3.1.2
/_    _/

Using Scala version 2.12.10 (Java HotSpot(TM) 64-Bit Server VM, Java 11.0.16.1)
Type in expressions to have them evaluated.
Type :help for more information.

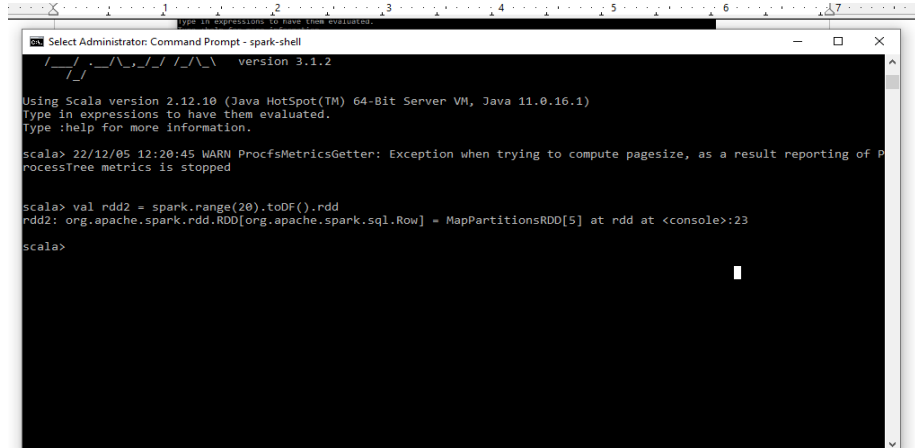
scala> 22/12/05 12:20:45 WARN ProcfsMetricsGetter: Exception when trying to compute pagesize, as a result reporting of
processTree metrics is stopped

scala>
```

- 2) To assign sequence in the rdd
 - val rdd = spark.sparkContext.parallelize(Seq(("cust_id"),("cust_name"),("cust_add"),(cust_contact_no)))
- 3) To print values
 - rdd.foreach(println)
- 4) To read text file as rdd
 - val rdd = spark.sparkContext.textFile("D:/cust.txt")
- 5) To count number of lines in txt
 - rdd.count()
- 6) To show output
 - rdd.foreach(println)
- 7) To read RDD from another RDD
 - val rdd3 = rdd.map(row=>{row._1+100})

8) To show output
- rdd3.foreach(println)

9) To create data frame or dataset in rdd
- val rdd2 = spark.range(20).toDF().rdd



```
Select Administrator: Command Prompt - spark-shell
C:\> spark-shell version 3.1.2

Using Scala version 2.12.10 (Java HotSpot(TM) 64-Bit Server VM, Java 11.0.16.1)
Type in expressions to have them evaluated.
Type :help for more information.

scala> 22/12/05 12:20:45 WARN ProcsMetricsGetter: Exception when trying to compute pagesize, as a result reporting of P
rocessFree metrics is stopped

scala> val rdd2 = spark.range(20).toDF().rdd
rdd2: org.apache.spark.rdd.RDD[org.apache.spark.sql.Row] = MapPartitionsRDD[5] at rdd at <console>:23
scala>
```

END