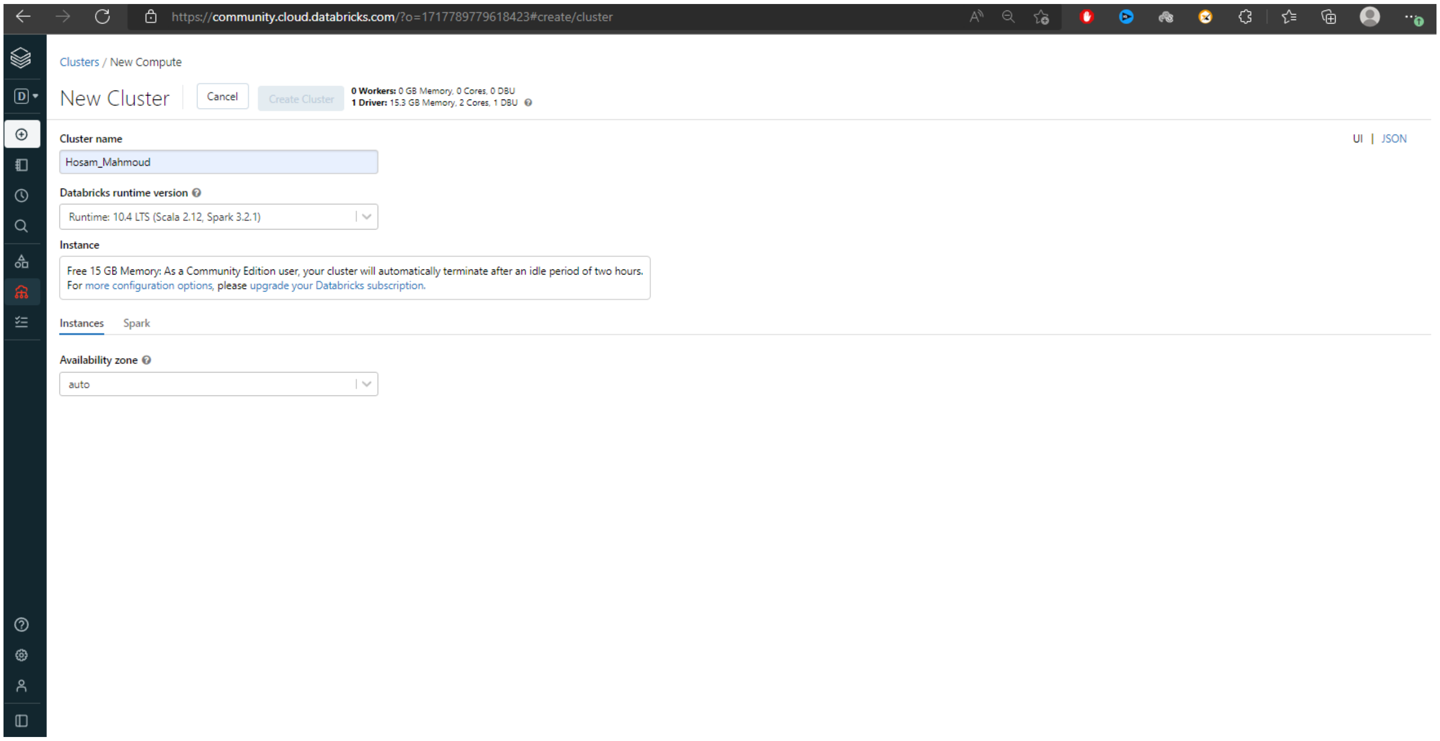
Name/ Hosam Mahmoud Ibrahim Mahmoud.

Please use the provided ***HTML*** Files to see the code and the output, these images are just to show the cluster name for each notebook.

****Note please don't forget to Scroll to see the full output.***

1- Cluster creation.



https://community.cloud.databricks.com/?o=1717789779618423#create/cluster

Clusters / New Compute

New Cluster

0 Workers: 0 GB Memory, 0 Cores, 0 DBU
1 Driver: 15.3 GB Memory, 2 Cores, 1 DBU

Cluster name: Hosam_Mahmoud

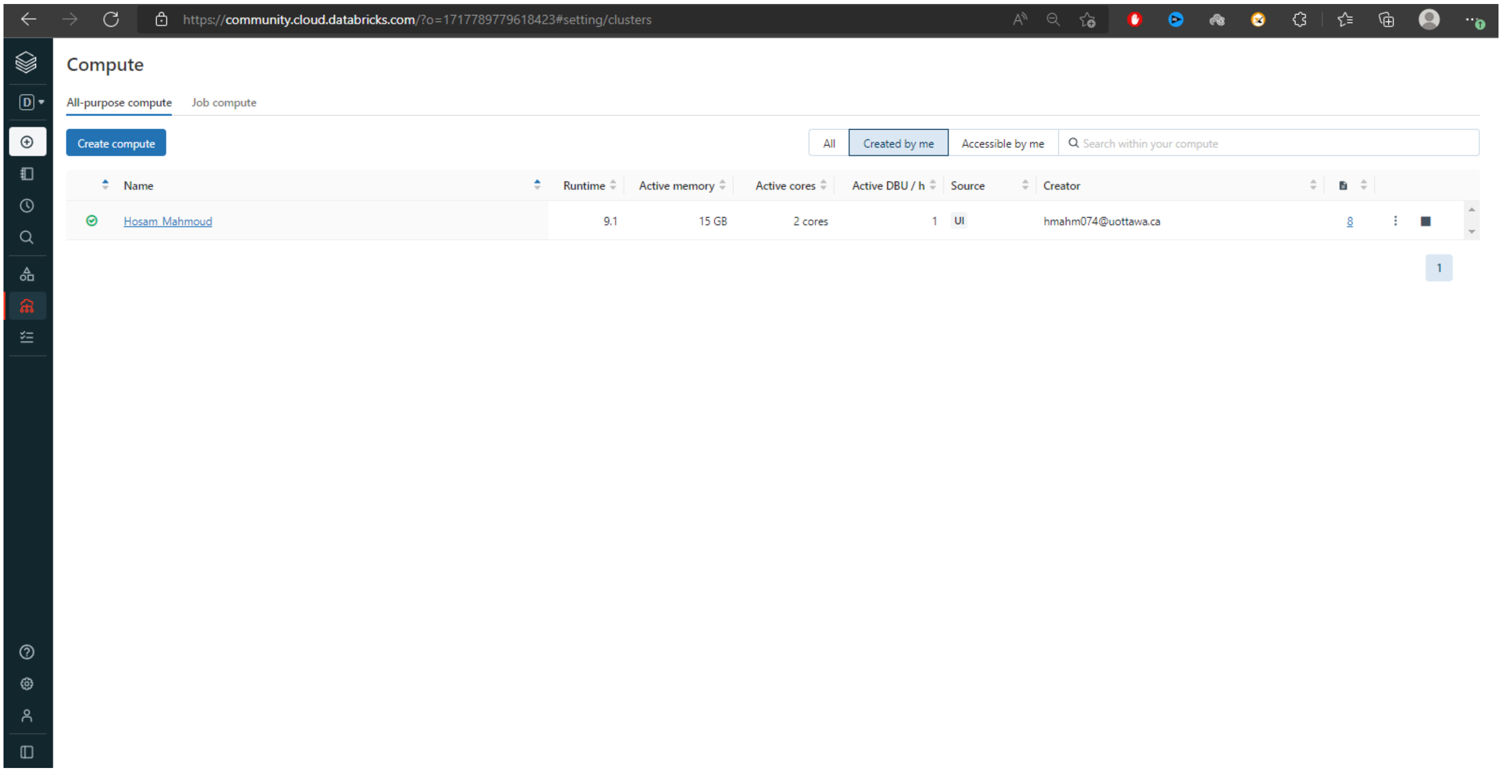
Databricks runtime version: Runtime: 10.4 LTS (Scala 2.12, Spark 3.2.1)

Instance

Free 15 GB Memory: As a Community Edition user, your cluster will automatically terminate after an idle period of two hours.
For more configuration options, please upgrade your Databricks subscription.

Availability zone: auto

2- Cluster Created successfully

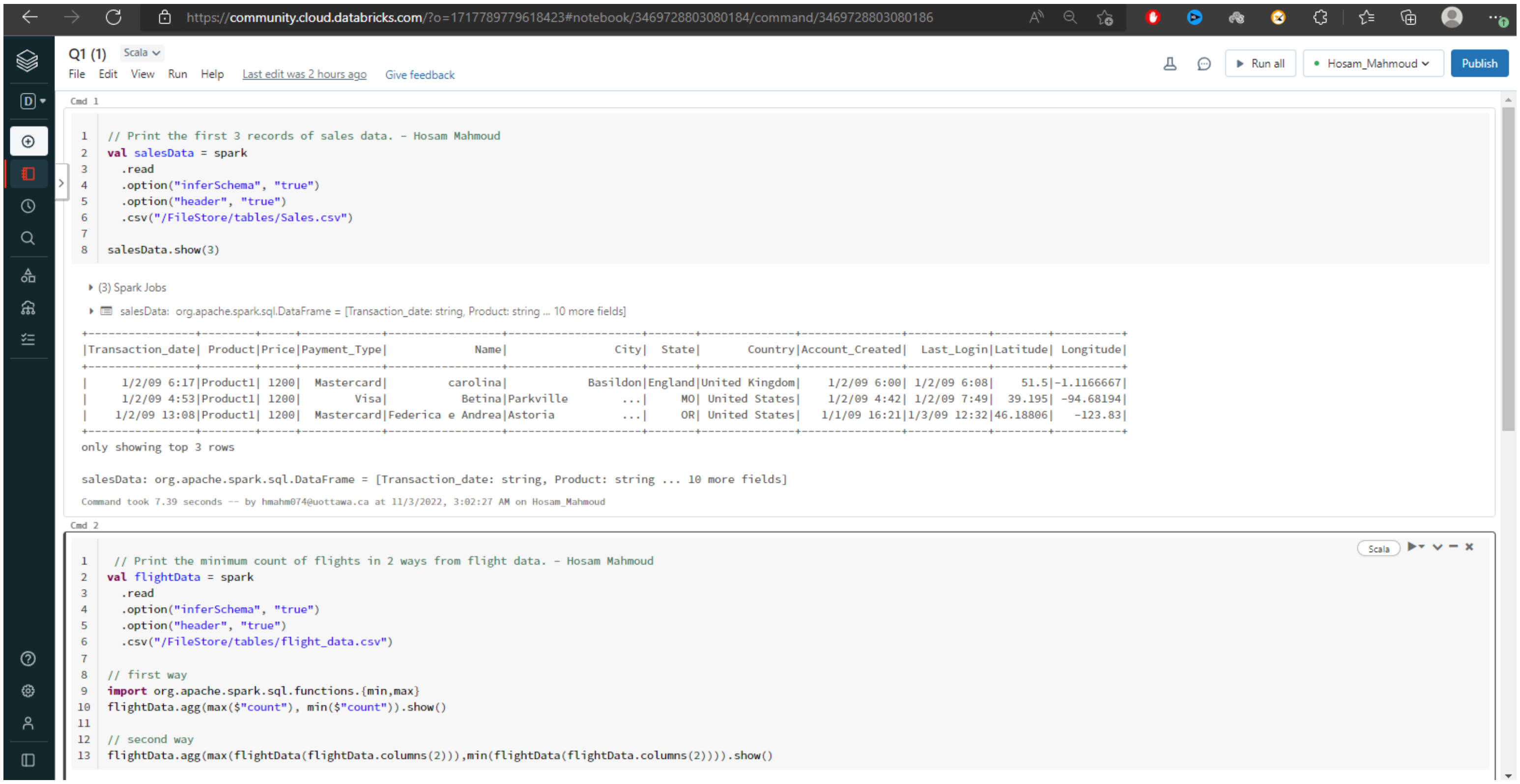


https://community.cloud.databricks.com/?o=1717789779618423#setting/clusters

Compute

	Name	Runtime	Active memory	Active cores	Active DBU / h	Source	Creator	
✓	Hosam Mahmoud	9.1	15 GB	2 cores	1	UI	hmahm074@uottawa.ca	8

3- Q1



https://community.cloud.databricks.com/?o=1717789779618423#notebook/3469728803080184/command/3469728803080186
Q1 (1)
Scala
File
Edit
View
Run
Help
Last edit was 2 hours ago
Give feedback
Run all
Hosam_Mahmoud
Publish
Cmd 1

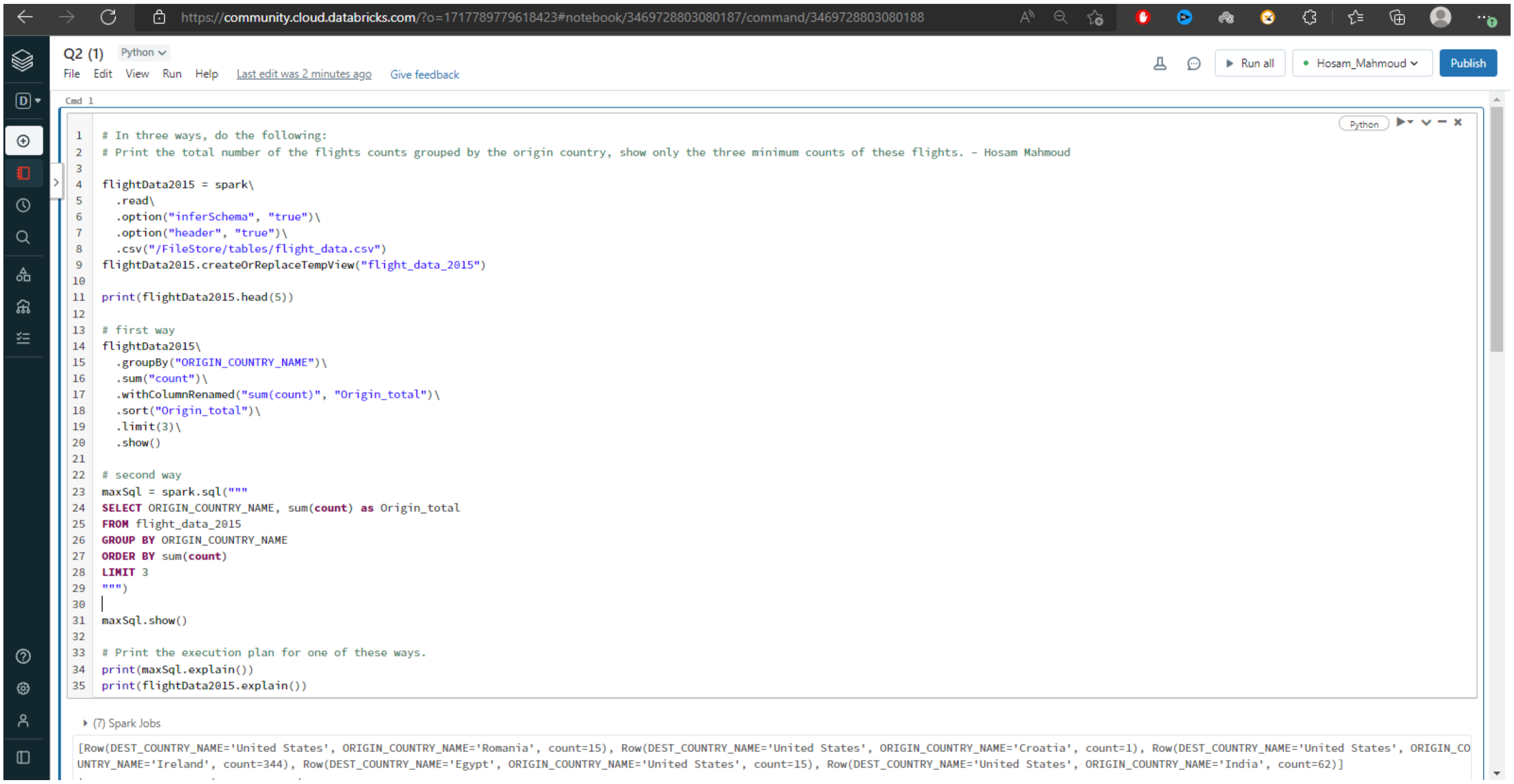
```scala
// Print the first 3 records of sales data. - Hosam Mahmoud
val salesData = spark
  .read
  .option("inferSchema", "true")
  .option("header", "true")
  .csv("/FileStore/tables/Sales.csv")

salesData.show(3)
```

(3) Spark Jobs
salesData: org.apache.spark.sql.DataFrame = [Transaction_date: string, Product: string ... 10 more fields]

```
+----------------+--------+-----+------------+-----------------+--------------------+-------+--------------+---------------+------------+--------+----------+
|Transaction_date| Product|Price|Payment_Type|             Name|                City|  State|       Country|Account_Created|  Last_Login|Latitude| Longitude|
+----------------+--------+-----+------------+-----------------+--------------------+-------+--------------+---------------+------------+--------+----------+
|     1/2/09 6:17|Product1| 1200|  Mastercard|         carolina|            Basildon|England|United Kingdom|    1/2/09 6:00| 1/2/09 6:08|    51.5|-1.1166667|
|     1/2/09 4:53|Product1| 1200|        Visa|           Betina|Parkville        ...|     MO| United States|    1/2/09 4:42| 1/2/09 7:49|  39.195| -94.68194|
|    1/2/09 13:08|Product1| 1200|  Mastercard|Federica e Andrea|Astoria          ...|     OR| United States|   1/1/09 16:21|1/3/09 12:32|46.18806|   -123.83|
+----------------+--------+-----+------------+-----------------+--------------------+-------+--------------+---------------+------------+--------+----------+
only showing top 3 rows

salesData: org.apache.spark.sql.DataFrame = [Transaction_date: string, Product: string ... 10 more fields]
Command took 7.39 seconds -- by hmahm074@uottawa.ca at 11/3/2022, 3:02:27 AM on Hosam_Mahmoud
```

```scala
 // Print the minimum count of flights in 2 ways from flight data. - Hosam Mahmoud
val flightData = spark
  .read
  .option("inferSchema", "true")
  .option("header", "true")
  .csv("/FileStore/tables/flight_data.csv")

// first way
import org.apache.spark.sql.functions.{min,max}
flightData.agg(max($"count"), min($"count")).show()

// second way
flightData.agg(max(flightData(flightData.columns(2))),min(flightData(flightData.columns(2)))).show()
```

4- Q2

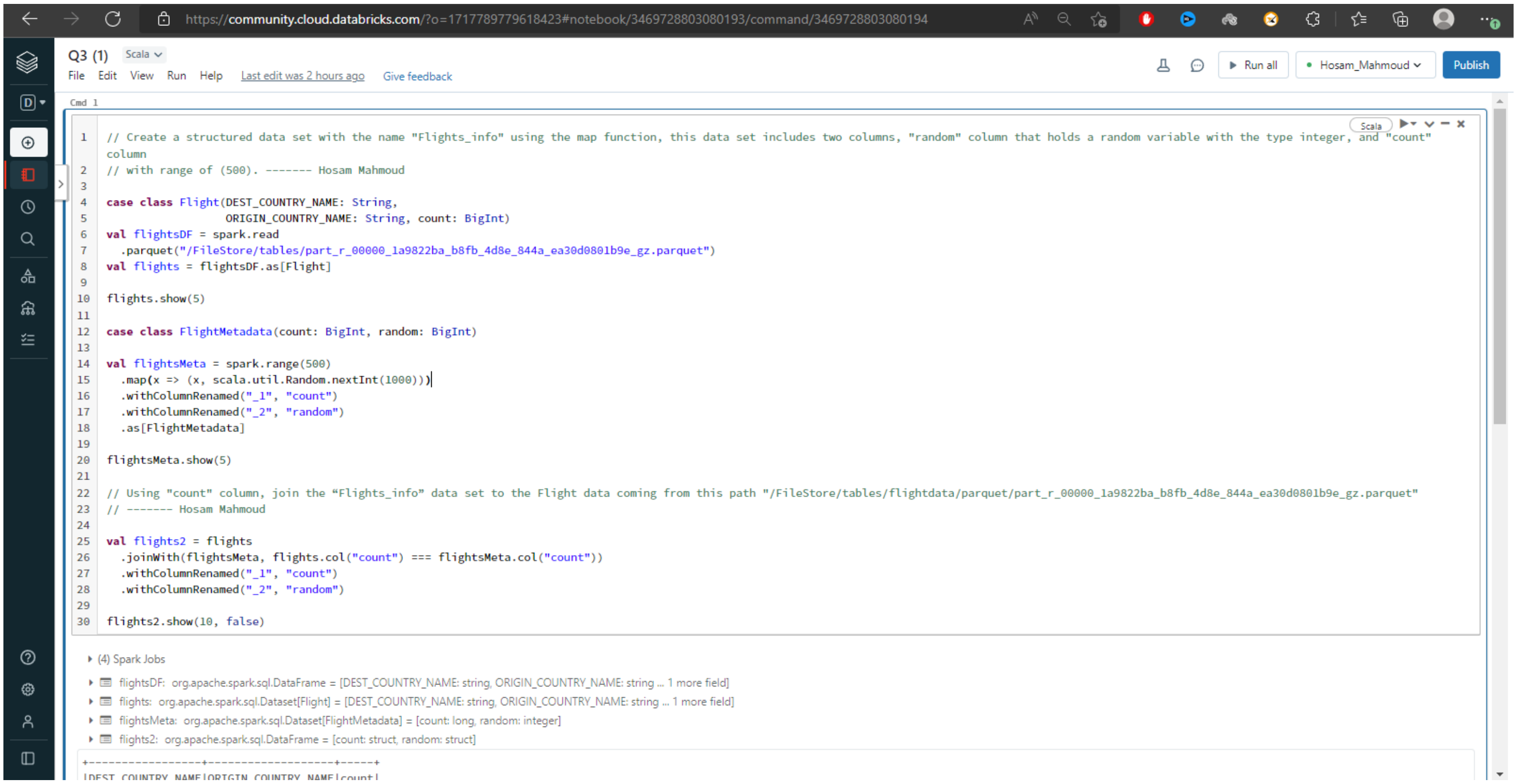


https://community.cloud.databricks.com/?o=1717789779618423#notebook/3469728803080187/command/3469728803080188
Q2 (1)
Python
File
Edit
View
Run
Help
Last edit was 2 minutes ago
Give feedback
Run all
Hosam_Mahmoud
Publish
Cmd 1
Python

```python
# In three ways, do the following:
# Print the total number of the flights counts grouped by the origin country, show only the three minimum counts of these flights. - Hosam Mahmoud

flightData2015 = spark\
  .read\
  .option("inferSchema", "true")\
  .option("header", "true")\
  .csv("/FileStore/tables/flight_data.csv")
flightData2015.createOrReplaceTempView("flight_data_2015")

print(flightData2015.head(5))

# first way
flightData2015\
  .groupBy("ORIGIN_COUNTRY_NAME")\
  .sum("count")\
  .withColumnRenamed("sum(count)", "Origin_total")\
  .sort("Origin_total")\
  .limit(3)\
  .show()

# second way
maxSql = spark.sql("""
SELECT ORIGIN_COUNTRY_NAME, sum(count) as Origin_total
FROM flight_data_2015
GROUP BY ORIGIN_COUNTRY_NAME
ORDER BY sum(count)
LIMIT 3
""")

maxSql.show()

# Print the execution plan for one of these ways.
print(maxSql.explain())
print(flightData2015.explain())
```

```
[Row(DEST_COUNTRY_NAME='United States', ORIGIN_COUNTRY_NAME='Romania', count=15), Row(DEST_COUNTRY_NAME='United States', ORIGIN_COUNTRY_NAME='Croatia', count=1), Row(DEST_COUNTRY_NAME='United States', ORIGIN_CO
UNTRY_NAME='Ireland', count=344), Row(DEST_COUNTRY_NAME='Egypt', ORIGIN_COUNTRY_NAME='United States', count=15), Row(DEST_COUNTRY_NAME='United States', ORIGIN_COUNTRY_NAME='India', count=62)]
```

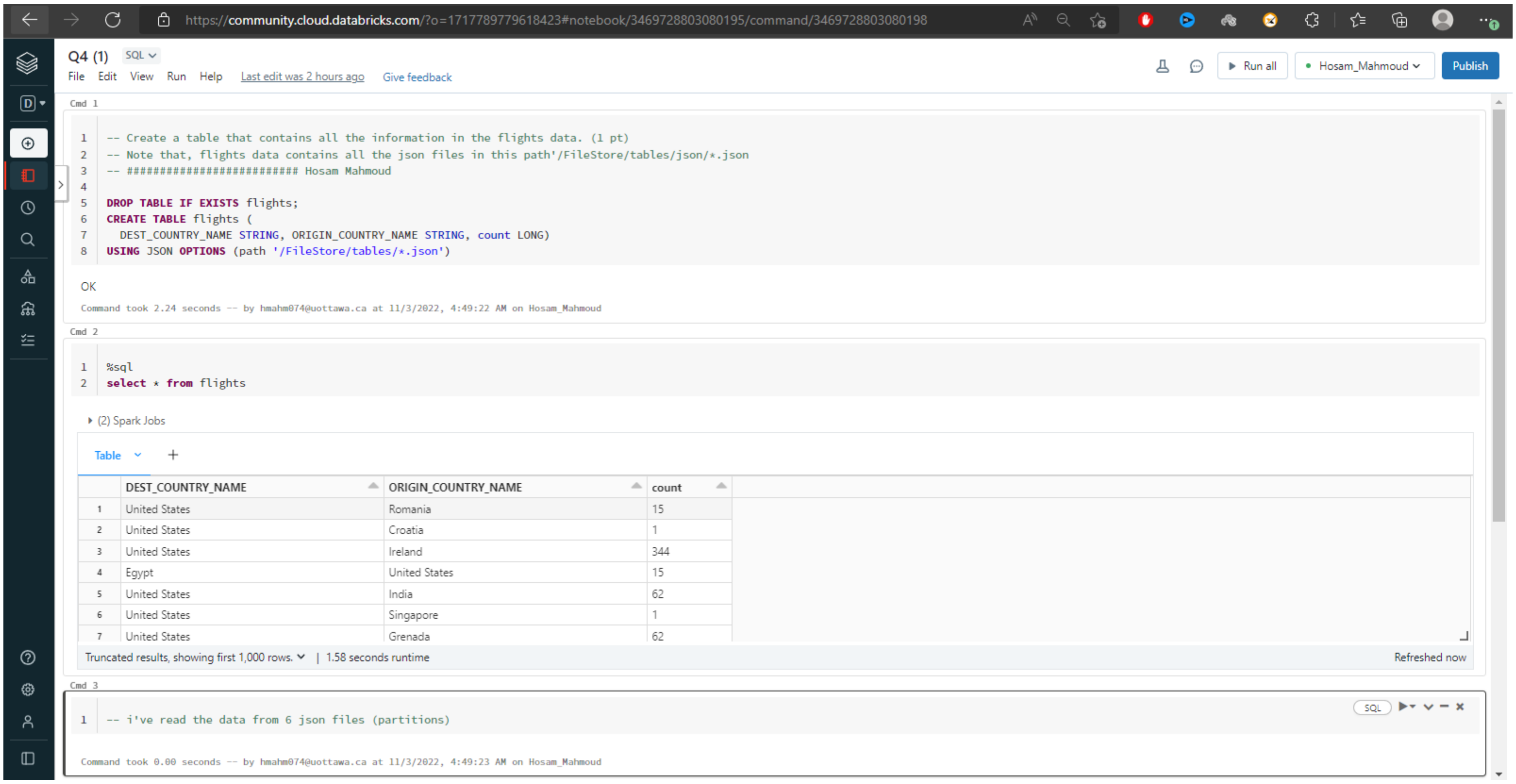
5- Q3


https://community.cloud.databricks.com/?o=1717789779618423#notebook/3469728803080193/command/3469728803080194
Q3 (1)
Scala
File
Edit
View
Run
Help
Last edit was 2 hours ago
Give feedback
Run all
Hosam_Mahmoud
Publish
Cmd 1

```scala
// Create a structured data set with the name "Flights_info" using the map function, this data set includes two columns, "random" column that holds a random variable with the type integer, and "count" column
// with range of (500). ------- Hosam Mahmoud

case class Flight(DEST_COUNTRY_NAME: String,
                  ORIGIN_COUNTRY_NAME: String, count: BigInt)
val flightsDF = spark.read
  .parquet("/FileStore/tables/part_r_00000_1a9822ba_b8fb_4d8e_844a_ea30d0801b9e_gz.parquet")
val flights = flightsDF.as[Flight]

flights.show(5)

case class FlightMetadata(count: BigInt, random: BigInt)

val flightsMeta = spark.range(500)
  .map(x => (x, scala.util.Random.nextInt(1000)))
  .withColumnRenamed("_1", "count")
  .withColumnRenamed("_2", "random")
  .as[FlightMetadata]

flightsMeta.show(5)

// Using "count" column, join the “Flights_info” data set to the Flight data coming from this path "/FileStore/tables/flightdata/parquet/part_r_00000_1a9822ba_b8fb_4d8e_844a_ea30d0801b9e_gz.parquet"
// ------- Hosam Mahmoud

val flights2 = flights
  .joinWith(flightsMeta, flights.col("count") === flightsMeta.col("count"))
  .withColumnRenamed("_1", "count")
  .withColumnRenamed("_2", "random")

flights2.show(10, false)
```

```
(4) Spark Jobs
flightsDF: org.apache.spark.sql.DataFrame = [DEST_COUNTRY_NAME: string, ORIGIN_COUNTRY_NAME: string ... 1 more field]
flights: org.apache.spark.sql.Dataset[Flight] = [DEST_COUNTRY_NAME: string, ORIGIN_COUNTRY_NAME: string ... 1 more field]
flightsMeta: org.apache.spark.sql.Dataset[FlightMetadata] = [count: long, random: integer]
flights2: org.apache.spark.sql.DataFrame = [count: struct, random: struct]
+-----------------+-------------------+-----+
|DEST_COUNTRY_NAME|ORIGIN_COUNTRY_NAME|count|
```

6- Q4


https://community.cloud.databricks.com/?o=1717789779618423#notebook/3469728803080195/command/3469728803080198
Q4 (1)
SQL
File
Edit
View
Run
Help
Last edit was 2 hours ago
Give feedback
Run all
Hosam_Mahmoud
Publish
Cmd 1

```sql
-- Create a table that contains all the information in the flights data. (1 pt)
-- Note that, flights data contains all the json files in this path'/FileStore/tables/json/*.json
-- ######################### Hosam Mahmoud

DROP TABLE IF EXISTS flights;
CREATE TABLE flights (
  DEST_COUNTRY_NAME STRING, ORIGIN_COUNTRY_NAME STRING, count LONG)
USING JSON OPTIONS (path '/FileStore/tables/*.json')
```

```
OK
Command took 2.24 seconds -- by hmahm074@uottawa.ca at 11/3/2022, 4:49:22 AM on Hosam_Mahmoud
```

Cmd 2

```sql
%sql
select * from flights
```

(2) Spark Jobs

	DEST_COUNTRY_NAME	ORIGIN_COUNTRY_NAME	count
1	United States	Romania	15
2	United States	Croatia	1
3	United States	Ireland	344
4	Egypt	United States	15
5	United States	India	62
6	United States	Singapore	1
7	United States	Grenada	62

Truncated results, showing first 1,000 rows. | 1.58 seconds runtime

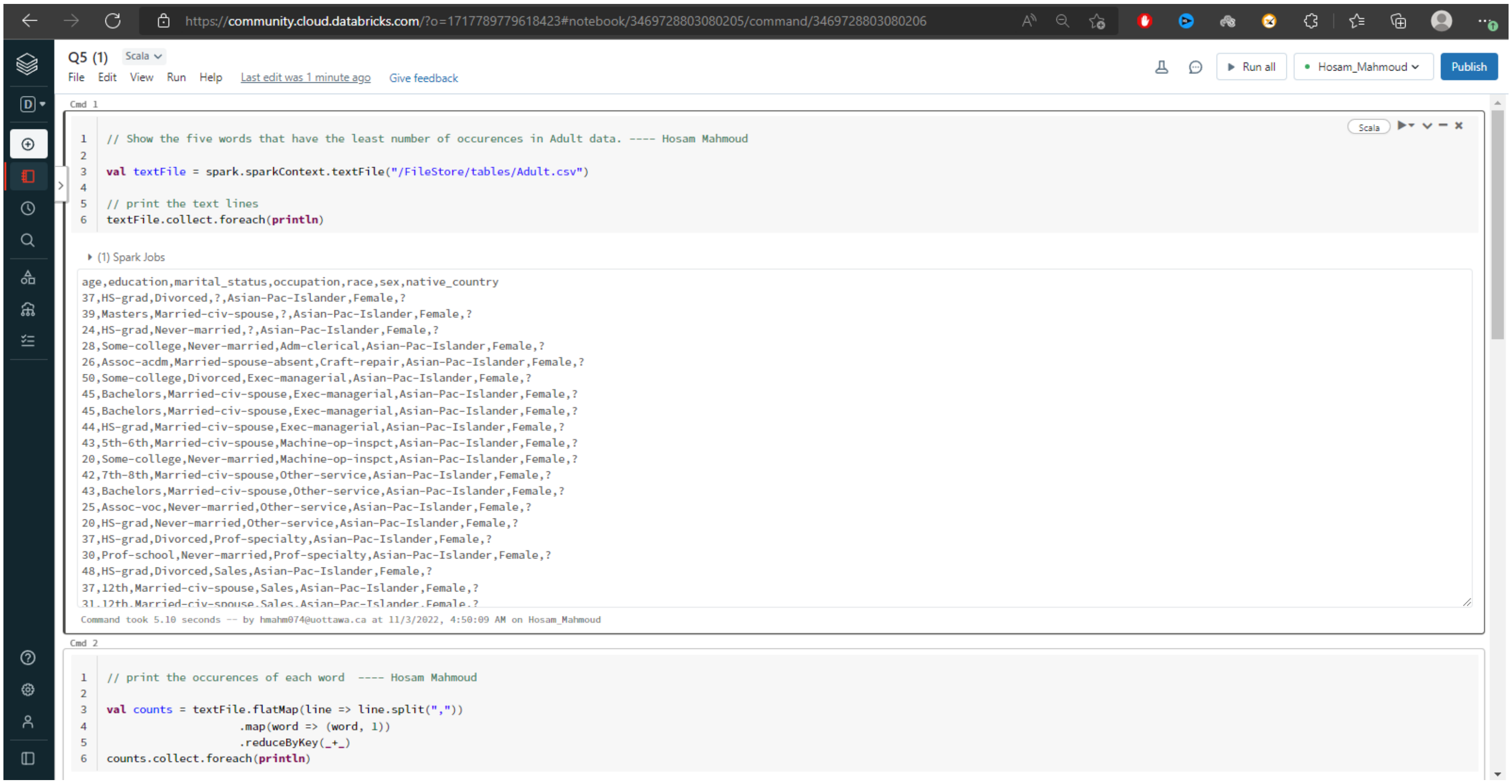
Refreshed now

Cmd 3

```sql
-- i've read the data from 6 json files (partitions)
```

```
Command took 0.00 seconds -- by hmahm074@uottawa.ca at 11/3/2022, 4:49:23 AM on Hosam_Mahmoud
```

7- Q5


https://community.cloud.databricks.com/?o=1717789779618423#notebook/3469728803080205/command/3469728803080206
Q5 (1)
Scala
File
Edit
View
Run
Help
Last edit was 1 minute ago
Give feedback
Run all
Hosam_Mahmoud
Publish
Cmd 1
Scala
(1) Spark Jobs
Cmd 2

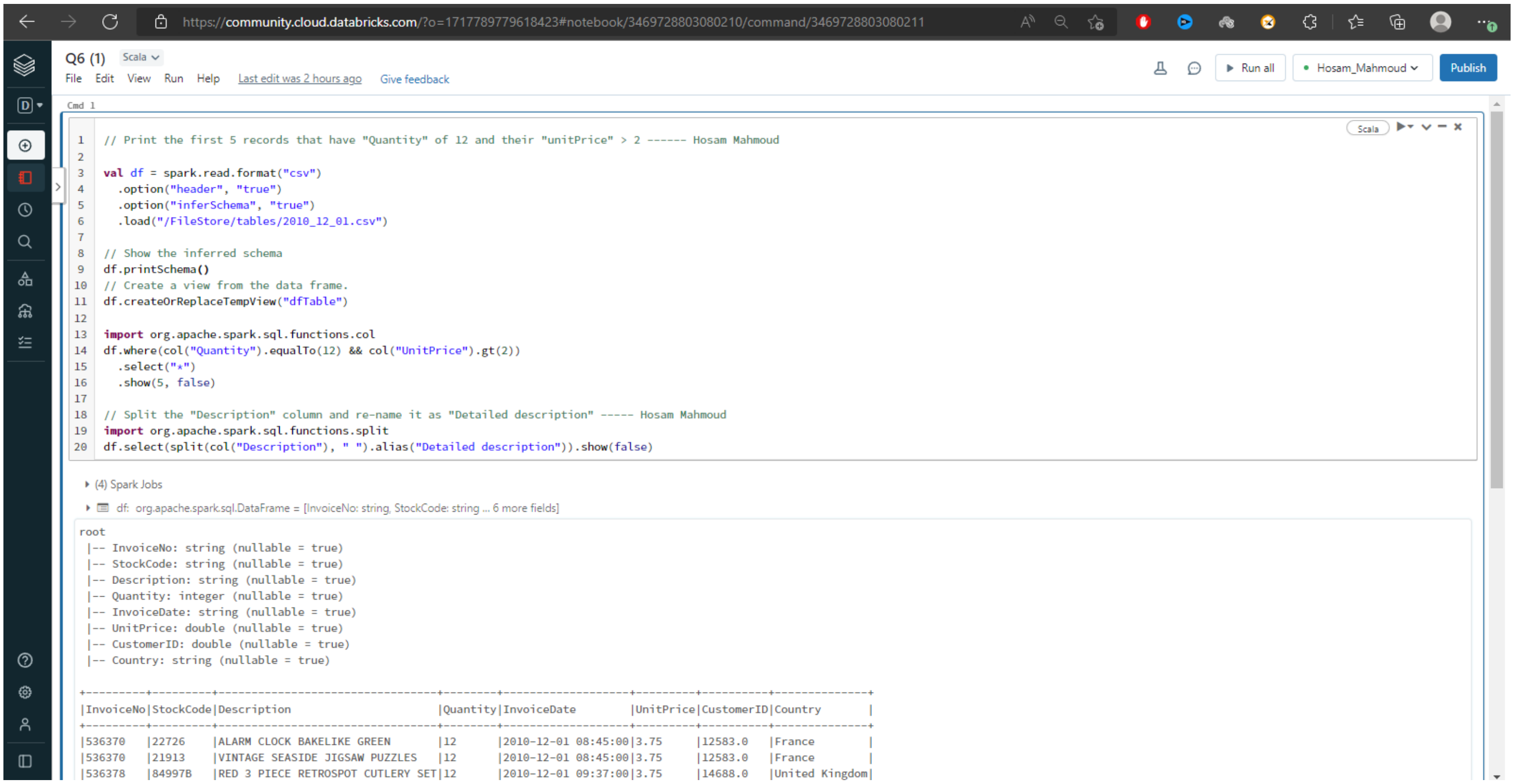
```
// Show the five words that have the least number of occurences in Adult data. ---- Hosam Mahmoud

val textFile = spark.sparkContext.textFile("/FileStore/tables/Adult.csv")

// print the text lines
textFile.collect.foreach(println)
```

```
age,education,marital_status,occupation,race,sex,native_country
37,HS-grad,Divorced,?,Asian-Pac-Islander,Female,?
39,Masters,Married-civ-spouse,?,Asian-Pac-Islander,Female,?
24,HS-grad,Never-married,?,Asian-Pac-Islander,Female,?
28,Some-college,Never-married,Adm-clerical,Asian-Pac-Islander,Female,?
26,Assoc-acdm,Married-spouse-absent,Craft-repair,Asian-Pac-Islander,Female,?
50,Some-college,Divorced,Exec-managerial,Asian-Pac-Islander,Female,?
45,Bachelors,Married-civ-spouse,Exec-managerial,Asian-Pac-Islander,Female,?
45,Bachelors,Married-civ-spouse,Exec-managerial,Asian-Pac-Islander,Female,?
44,HS-grad,Married-civ-spouse,Exec-managerial,Asian-Pac-Islander,Female,?
43,5th-6th,Married-civ-spouse,Machine-op-inspct,Asian-Pac-Islander,Female,?
20,Some-college,Never-married,Machine-op-inspct,Asian-Pac-Islander,Female,?
42,7th-8th,Married-civ-spouse,Other-service,Asian-Pac-Islander,Female,?
43,Bachelors,Married-civ-spouse,Other-service,Asian-Pac-Islander,Female,?
25,Assoc-voc,Never-married,Other-service,Asian-Pac-Islander,Female,?
20,HS-grad,Never-married,Other-service,Asian-Pac-Islander,Female,?
37,HS-grad,Divorced,Prof-specialty,Asian-Pac-Islander,Female,?
30,Prof-school,Never-married,Prof-specialty,Asian-Pac-Islander,Female,?
48,HS-grad,Divorced,Sales,Asian-Pac-Islander,Female,?
37,12th,Married-civ-spouse,Sales,Asian-Pac-Islander,Female,?
31,12th,Married-civ-spouse,Sales,Asian-Pac-Islander,Female,?
```

Command took 5.10 seconds -- by hmahm074@uottawa.ca at 11/3/2022, 4:50:09 AM on Hosam_Mahmoud

```
// print the occurences of each word  ---- Hosam Mahmoud

val counts = textFile.flatMap(line => line.split(","))
                     .map(word => (word, 1))
                     .reduceByKey(_+_)
counts.collect.foreach(println)
```

8- Q6

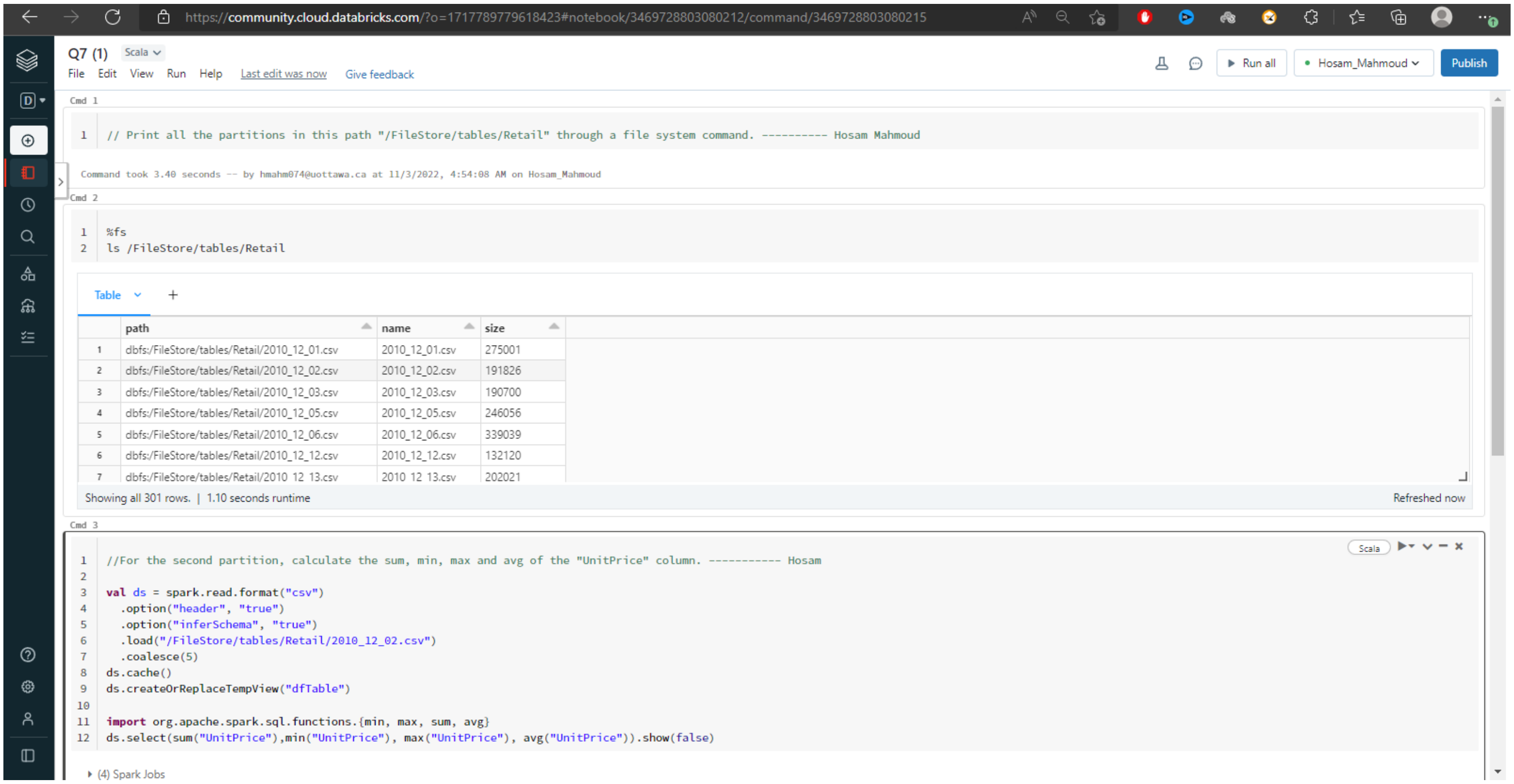

https://community.cloud.databricks.com/?o=1717789779618423#notebook/3469728803080210/command/3469728803080211
Q6 (1)
Scala
File
Edit
View
Run
Help
Last edit was 2 hours ago
Give feedback
Run all
Hosam_Mahmoud
Publish
Cmd 1
Scala
(4) Spark Jobs

```
// Print the first 5 records that have "Quantity" of 12 and their "unitPrice" > 2 ------ Hosam Mahmoud

val df = spark.read.format("csv")
  .option("header", "true")
  .option("inferSchema", "true")
  .load("/FileStore/tables/2010_12_01.csv")

// Show the inferred schema
df.printSchema()
// Create a view from the data frame.
df.createOrReplaceTempView("dfTable")

import org.apache.spark.sql.functions.col
df.where(col("Quantity").equalTo(12) && col("UnitPrice").gt(2))
  .select("*")
  .show(5, false)

// Split the "Description" column and re-name it as "Detailed description" ----- Hosam Mahmoud
import org.apache.spark.sql.functions.split
df.select(split(col("Description"), " ").alias("Detailed description")).show(false)
```

df: org.apache.spark.sql.DataFrame = [InvoiceNo: string, StockCode: string ... 6 more fields]

```
root
 |-- InvoiceNo: string (nullable = true)
 |-- StockCode: string (nullable = true)
 |-- Description: string (nullable = true)
 |-- Quantity: integer (nullable = true)
 |-- InvoiceDate: string (nullable = true)
 |-- UnitPrice: double (nullable = true)
 |-- CustomerID: double (nullable = true)
 |-- Country: string (nullable = true)

+---------+---------+-----------------------------------+--------+-------------------+---------+----------+--------------+
|InvoiceNo|StockCode|Description                        |Quantity|InvoiceDate        |UnitPrice|CustomerID|Country       |
+---------+---------+-----------------------------------+--------+-------------------+---------+----------+--------------+
|536370   |22726    |ALARM CLOCK BAKELIKE GREEN         |12      |2010-12-01 08:45:00|3.75     |12583.0   |France        |
|536370   |21913    |VINTAGE SEASIDE JIGSAW PUZZLES     |12      |2010-12-01 08:45:00|3.75     |12583.0   |France        |
|536378   |84997B   |RED 3 PIECE RETROSPOT CUTLERY SET|12      |2010-12-01 09:37:00|3.75     |14688.0   |United Kingdom|
```

9- Q7


https://community.cloud.databricks.com/?o=1717789779618423#notebook/3469728803080212/command/3469728803080215
Q7 (1)
Scala
File Edit View Run Help
Last edit was now
Give feedback
Run all
Hosam_Mahmoud
Publish

Cmd 1

```
// Print all the partitions in this path "/FileStore/tables/Retail" through a file system command. ---------- Hosam Mahmoud
```

Command took 3.40 seconds -- by hmahm074@uottawa.ca at 11/3/2022, 4:54:08 AM on Hosam_Mahmoud

Cmd 2

```
%fs
ls /FileStore/tables/Retail
```

	path	name	size
1	dbfs:/FileStore/tables/Retail/2010_12_01.csv	2010_12_01.csv	275001
2	dbfs:/FileStore/tables/Retail/2010_12_02.csv	2010_12_02.csv	191826
3	dbfs:/FileStore/tables/Retail/2010_12_03.csv	2010_12_03.csv	190700
4	dbfs:/FileStore/tables/Retail/2010_12_05.csv	2010_12_05.csv	246056
5	dbfs:/FileStore/tables/Retail/2010_12_06.csv	2010_12_06.csv	339039
6	dbfs:/FileStore/tables/Retail/2010_12_12.csv	2010_12_12.csv	132120
7	dbfs:/FileStore/tables/Retail/2010_12_13.csv	2010_12_13.csv	202021

Showing all 301 rows. | 1.10 seconds runtime

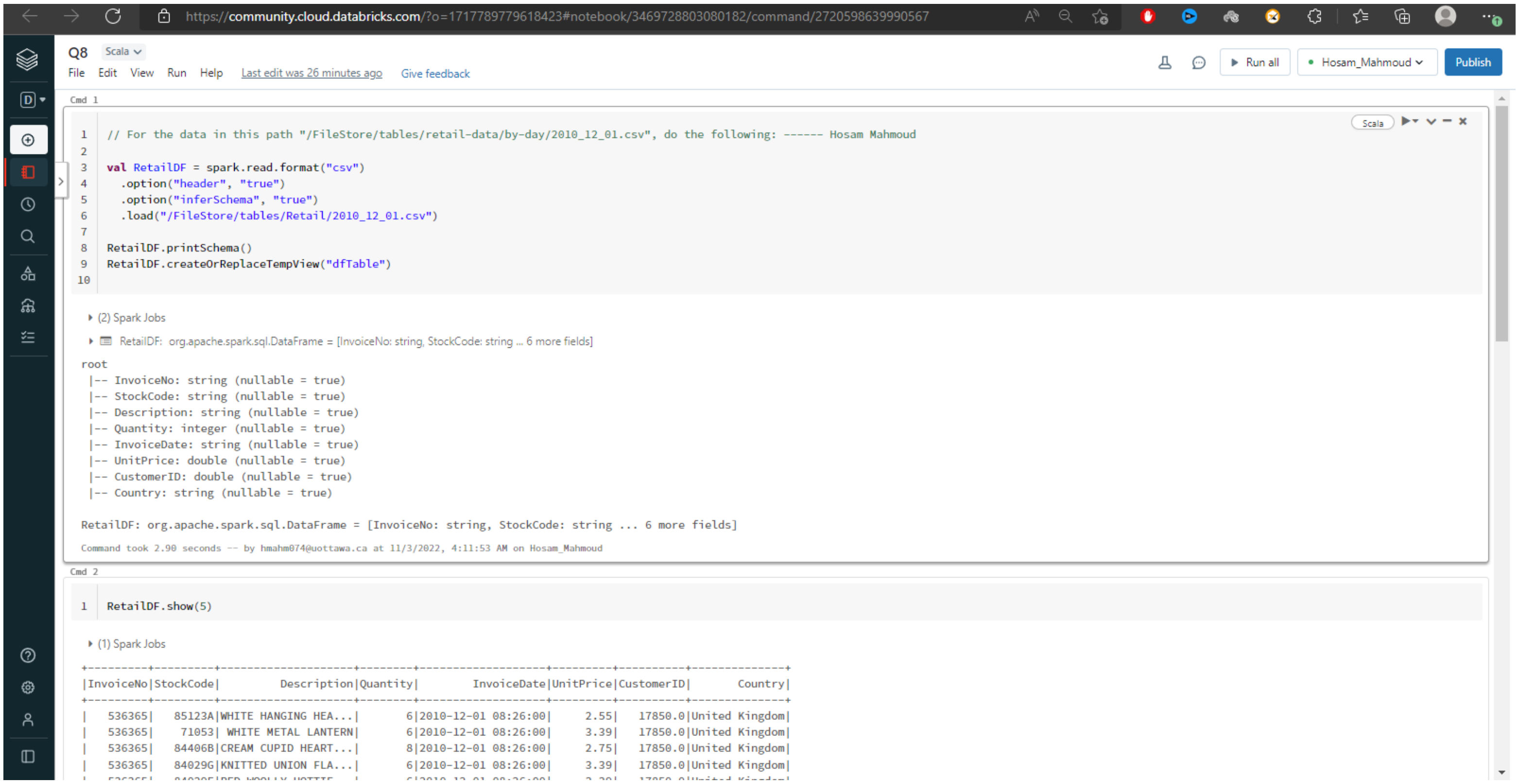
Refreshed now

Cmd 3

```
//For the second partition, calculate the sum, min, max and avg of the "UnitPrice" column. ----------- Hosam

val ds = spark.read.format("csv")
  .option("header", "true")
  .option("inferSchema", "true")
  .load("/FileStore/tables/Retail/2010_12_02.csv")
  .coalesce(5)
ds.cache()
ds.createOrReplaceTempView("dfTable")

import org.apache.spark.sql.functions.{min, max, sum, avg}
ds.select(sum("UnitPrice"),min("UnitPrice"), max("UnitPrice"), avg("UnitPrice")).show(false)
```

▸ (4) Spark Jobs

10- Q8


https://community.cloud.databricks.com/?o=1717789779618423#notebook/3469728803080182/command/2720598639990567
Q8
Scala
File Edit View Run Help
Last edit was 26 minutes ago
Give feedback
Run all
Hosam_Mahmoud
Publish

Cmd 1

```
// For the data in this path "/FileStore/tables/retail-data/by-day/2010_12_01.csv", do the following: ------ Hosam Mahmoud

val RetailDF = spark.read.format("csv")
  .option("header", "true")
  .option("inferSchema", "true")
  .load("/FileStore/tables/Retail/2010_12_01.csv")

RetailDF.printSchema()
RetailDF.createOrReplaceTempView("dfTable")

```

▸ (2) Spark Jobs

▸ RetailDF: org.apache.spark.sql.DataFrame = [InvoiceNo: string, StockCode: string ... 6 more fields]

```
root
 |-- InvoiceNo: string (nullable = true)
 |-- StockCode: string (nullable = true)
 |-- Description: string (nullable = true)
 |-- Quantity: integer (nullable = true)
 |-- InvoiceDate: string (nullable = true)
 |-- UnitPrice: double (nullable = true)
 |-- CustomerID: double (nullable = true)
 |-- Country: string (nullable = true)

RetailDF: org.apache.spark.sql.DataFrame = [InvoiceNo: string, StockCode: string ... 6 more fields]
```

Command took 2.90 seconds -- by hmahm074@uottawa.ca at 11/3/2022, 4:11:53 AM on Hosam_Mahmoud

Cmd 2

```
RetailDF.show(5)
```

▸ (1) Spark Jobs

```
+---------+---------+--------------------+--------+-------------------+---------+----------+--------------+
|InvoiceNo|StockCode|         Description|Quantity|        InvoiceDate|UnitPrice|CustomerID|       Country|
+---------+---------+--------------------+--------+-------------------+---------+----------+--------------+
|   536365|   85123A|WHITE HANGING HEA...|       6|2010-12-01 08:26:00|     2.55|   17850.0|United Kingdom|
|   536365|    71053| WHITE METAL LANTERN|       6|2010-12-01 08:26:00|     3.39|   17850.0|United Kingdom|
|   536365|   84406B|CREAM CUPID HEART...|       8|2010-12-01 08:26:00|     2.75|   17850.0|United Kingdom|
|   536365|   84029G|KNITTED UNION FLA...|       6|2010-12-01 08:26:00|     3.39|   17850.0|United Kingdom|
```