CSEE5590 Big Data Programming

In Class Programming –10 Report
(Jongkook Son)

Project Overview:

Lesson Title: *Data Frame and SQL*
Lesson Description: *Distributed Collection of Data*

Requirements/Task(s):

Part – 1

1. Import the dataset and create data frames directly on import.

2. Save data to file.

3. Check for Duplicate records in the dataset.

4. Apply Union operation on the dataset and order the output by Country

Name alphabetically.

5. Use Group by Query based on treatment.

Part – 2

1. Apply the basic queries related to Joins and aggregate functions (at least 2)

2. Write a query to fetch 13th Row in the dataset.

What I learned in ICP:

I learned How to utilize sparksql using scala. In this Icp I performed some actions and transformations and basic commands and sql for the data frames. Also By Using Spark SQL, I could have loaded the csv file easily and perform some queries

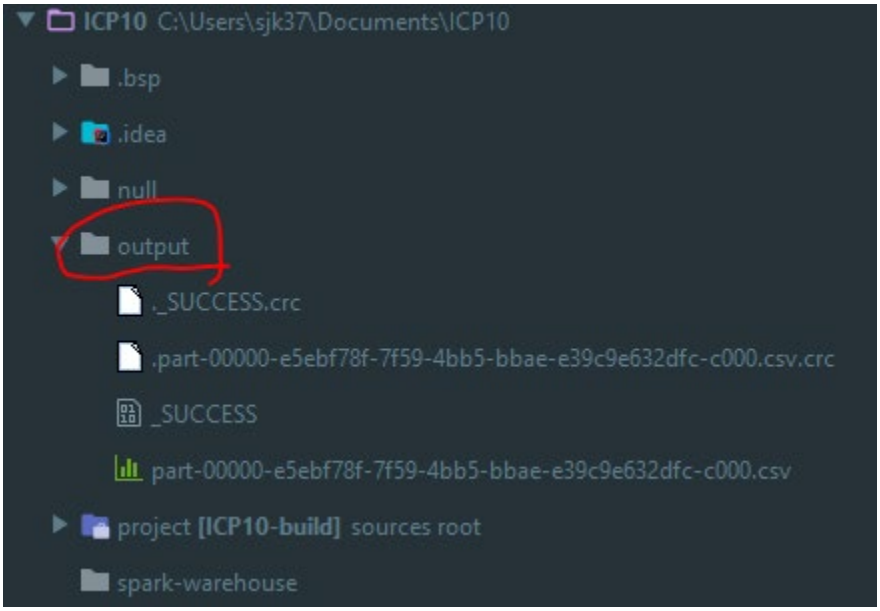
Part – 1

1. Import the dataset and create data frames directly on import.

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

  val spark: SparkSession = SparkSession.builder()
    .master( master = "local[*]")
    .appName( name = "groupProject")
    .getOrCreate()

  val SQLContext = spark.sqlContext

  val filepath = "survey.csv"

  //  1.  Import the dataset and create data frames directly on import.
  val df = SQLContext.read.option("header", true).csv(filepath)
  df.show( numRows = 20)
  df.createOrReplaceTempView( viewName = "survey")
```

2.Save data to file

```
// 2.Save data to file
df.write.mode( saveMode = "overwrite").option("header","true").csv( path = "output")
```



3. Check for Duplicate records in the dataset.

```
// 3. Check for Duplicate records in the dataset.
val distinctDF = df.distinct()
println(distinctDF.count()+"some"+ df.count())
```

```
21/04/06 23:20:00 INFO DAGScheduler: ResultStage 7 (count at SparkSql.scala:27) finished in 0.006 s
21/04/06 23:20:00 INFO DAGScheduler: Job 4 finished: count at SparkSql.scala:27, took 0.025472 s
Distinct count: 1259 Overall Count: 1259
21/04/06 23:20:00 INFO FileSourceStrategy: Pruning directories with:
21/04/06 23:20:00 INFO FileSourceStrategy: Post-Scan Filters: isnotnull(state#14),(state#14 = TX)
21/04/06 23:20:00 INFO FileSourceStrategy: Output Data Schema: struct<Timestamp: string, Age: string, Gender: string, Country: string, st
```

4. Apply Union operation on the dataset and order the output by Country Name alphabetically.

```
21/04/06 23:29:12 INFO ContextCleaner: Cleaned accumulator 333
+-------------------+---+--------------------+--------------+-----+-------------+--------------+---------+--------------+--------------+-----------+---------+----------+---------------+----------+---------+----------+
|          Timestamp|Age|              Gender|       Country|state|self_employed|family_history|treatment|work_interfere|  no_employees|remote_work|tech_company|  benefits|care_options|wellness_program| seek_help| anonymity|
+-------------------+---+--------------------+--------------+-----+-------------+--------------+---------+--------------+--------------+-----------+---------+----------+---------------+----------+---------+----------+
|2014-08-29 09:10:58|  8|A little about you| Bahamas, The|   IL|          Yes|           Yes|      Yes|         Often|           1-5|        Yes|      Yes|       Yes|            Yes|       Yes|      Yes|
|2014-08-27 11:44:55| 22|                   M|United States|   TX|           No|           Yes|      Yes|         Often|         6-25|         No|      Yes|        No|            Yes|        No|       No|      Yes|
|2014-08-28 17:07:11| 23|                   M|United States|   IL|           No|            No|       No|         Never|       100-500|         No|       No|Don't know|            No|  Don't know|Don't know|Don't know|
|2014-08-27 12:51:36| 33|                Male|United States|   TX|           No|            No|       No|            NA|         6-25|        Yes|Yes|Don't know|            No|        No|       No|Don't know|
|2014-08-27 15:25:03| 29|                Male|United States|   TX|           No|            No|       No|         Never|More than 1000|         No|       No|       Yes|     Not sure|        No|Don't know|Don't know|
|2014-08-27 11:43:36| 38|              Female|United States|   TX|           No|           Yes|      Yes|     Sometimes|         26-100|         No|      Yes|       Yes|            Yes|        No|      Yes|       Yes|
|2014-08-27 21:15:09| 34|                Male|United States|   IL|           No|           Yes|      Yes|     Sometimes|         26-100|        Yes|       No|       Yes|             No|        No|       No|Don't know|
|2014-08-27 15:35:21| 24|                Male|United States|   TX|           No|           Yes|      Yes|     Sometimes|       100-500|         No|      Yes|        No|            Yes|        No|       No|Don't know|Some
|2014-08-27 14:10:15| 18|                male|United States|   TX|           No|            No|      Yes|     Sometimes|         6-25|         No|Yes|Don't know|             No|        No|Don't know|Don't know|
|2015-02-21 04:41:28| 32|                Male|United States|   TX|           No|            No|      Yes|         Often|         26-100|        Yes|      Yes|       Yes|     Not sure|        No|Don't know|Don't know|
|2014-08-27 11:29:31| 37|              Female|United States|   IL|           NA|            No|      Yes|         Often|         6-25|         No|      Yes|       Yes|     Not sure|        No|       Yes|       Yes|
|2014-08-27 11:44:43| 30|                male|United States|   IL|           No|           Yes|      Yes|        Rarely|         26-100|         No|      Yes|       Yes|             No|        No|Don't know|Don't know|
|2014-08-28 17:56:02| 25|                Male|United States|   TX|           No|           Yes|      Yes|        Rarely|         26-100|         No|       No|        No|             No|        No|       No|Don't know|
|2014-08-27 12:41:20| 36|                Male|United States|   IL|           No|            No|       No|         Never|         6-25|         No|      Yes|Don't know|     Not sure|  Don't know|Don't know|Don't know|
|2014-08-27 12:42:24| 25|                Male|United States|   IL|           No|           Yes|      Yes|     Sometimes|         26-100|         No|       No|Don't know|             No|        No|Don't know|Don't know|Some
|2014-08-28 16:57:46| 40|                   M|United States|   IL|           No|            No|       No|        Rarely|More than 1000|         No|      Yes|       Yes|            Yes|       Yes|      Yes|       Yes|
|2014-08-28 09:56:21| 39|                   F|United States|   TX|           No|            No|       No|            NA|More than 1000|        Yes|       No|        No|            Yes|        No|       No|Don't know|
|2014-09-09 13:49:50| 29|                Male|United States|   IL|           No|            No|       No|            NA|More than 1000|         No|      Yes|Don't know|     Not sure|        No|       No|Don't know|
|2014-08-27 11:32:39| 42|              Female|United States|   IL|           NA|           Yes|      Yes|     Sometimes|       100-500|         No|      Yes|       Yes|            Yes|        No|       No|        No|
|2014-08-28 17:50:32| 24|                Male|United States|   TX|           No|            No|       No|     Sometimes|         26-100|        Yes|      Yes|Don't know|             No|        No|Don't know|Don't know|
+-------------------+---+--------------------+--------------+-----+-------------+--------------+---------+--------------+--------------+-----------+---------+----------+---------------+----------+---------+----------+
only showing top 20 rows
```

5. Use Groupby Query based on treatment.

```
//5. Use Groupby Query based on treatment.
SQLContext.sql( sqlText = "SELECT treatment, count(*) FROM survey GROUP BY treatment").show( numRows = 20)
```

```
+---------+--------+
|treatment|count(1)|
+---------+--------+
|       No|     622|
|      Yes|     637|
+---------+--------+

21/04/06 23:29:12 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
```

Part – 2

1. Apply the basic queries related to Joins and aggregate functions (at least 2)

```
// Part - 2

// 1. Apply the basic queries related to Joins and aggregate functions (at least 2)

val femaleDF = df.filter( conditionExpr = "Gender LIKE 'f%' OR Gender LIKE 'F%' ")
val maleDF   = df.filter( conditionExpr = "Gender LIKE 'm%' OR Gender LIKE 'M%' ")

val df1 = femaleDF.select( col = "Age" , cols = "Country","Gender","state","family_history")
val df2 = maleDF.select( col = "Age" , cols = "Country","Gender","state", "benefits")
val jointdf = df1.join(df2, df1("Country") === df2("Country"), joinType = "inner")
jointdf.show( truncate = false)

val udf = df2.union(df1)
val uniondf =udf.withColumn( colName = "Age", udf.col( colName = "Age").cast(DataTypes.IntegerType))
uniondf.orderBy( sortCol = "Country").show( numRows = 10)

uniondf.groupBy( col1 = "Country").count().show()
uniondf.groupBy( col1 = "Country").mean( colNames = "Age").show( numRows = 10)
```

```
21/04/06 23:39:46 INFO CodeGenerator: Code generated in 4.535 ms
+---+---------+------+-----+----------+
|Age|  Country|Gender|state|  benefits|
+---+---------+------+-----+----------+
| 27|Australia|  Male|   NA|        No|
| 27|Australia|  Male|   NA|        No|
| 25|Australia|  Male|   NA|        No|
| 48|Australia|  male|   NA|        No|
| 33|Australia|  Male|   NA|Don't know|
| 27|Australia|Female|   NA|       Yes|
| 23|Australia|Female|   NA|       Yes|
| 26|Australia|     F|   NA|       Yes|
| 29|Australia|Female|   NA|       Yes|
| 34|Australia|Female|   NA|       Yes|
+---+---------+------+-----+----------+
only showing top 10 rows
```

```
21/04/06 23:39:47 INFO FileSourceScanExec: Pushed Filters:
+-------------+------------------+
|      Country|          avg(Age)|
+-------------+------------------+
|       Russia|              27.0|
|       Sweden|26.857142857142858|
|  Philippines|              31.0|
|    Singapore|             34.25|
|      Germany|30.522727272727273|
|       France| 31.53846153846154|
|       Greece|              36.5|
|      Belgium|              29.5|
|      Finland|29.333333333333332|
|United States| 33.42798913043478|
+-------------+------------------+
only showing top 10 rows
```

2. Write a query to fetch 13th Row in the dataset.

```
// 2. Write a query to fetch 13th Row in the dataset.
println(df.take(13).last)
```

```
21/04/06 23:39:47 INFO SparkUI: Stopped Spark web UI at http://DESKTOP-IHI3GNU.mshome.net:4040
[2014-08-27 11:33:23,42,female,United States,CA,NA,Yes,Yes,Sometimes,26-100,No,No,Yes,Yes,No,No,Don't know,Somewhat difficult,Yes,Yes,Yes,Yes,Maybe,Maybe,No,Yes,NA]
21/04/06 23:39:47 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
```