Assignment 18.3:

Problem Statement:

Initial Steps:

Step1: Create a temporary table User

import org.apache.spark.sql.types.{StructType, StringType, IntegerType, StructField}

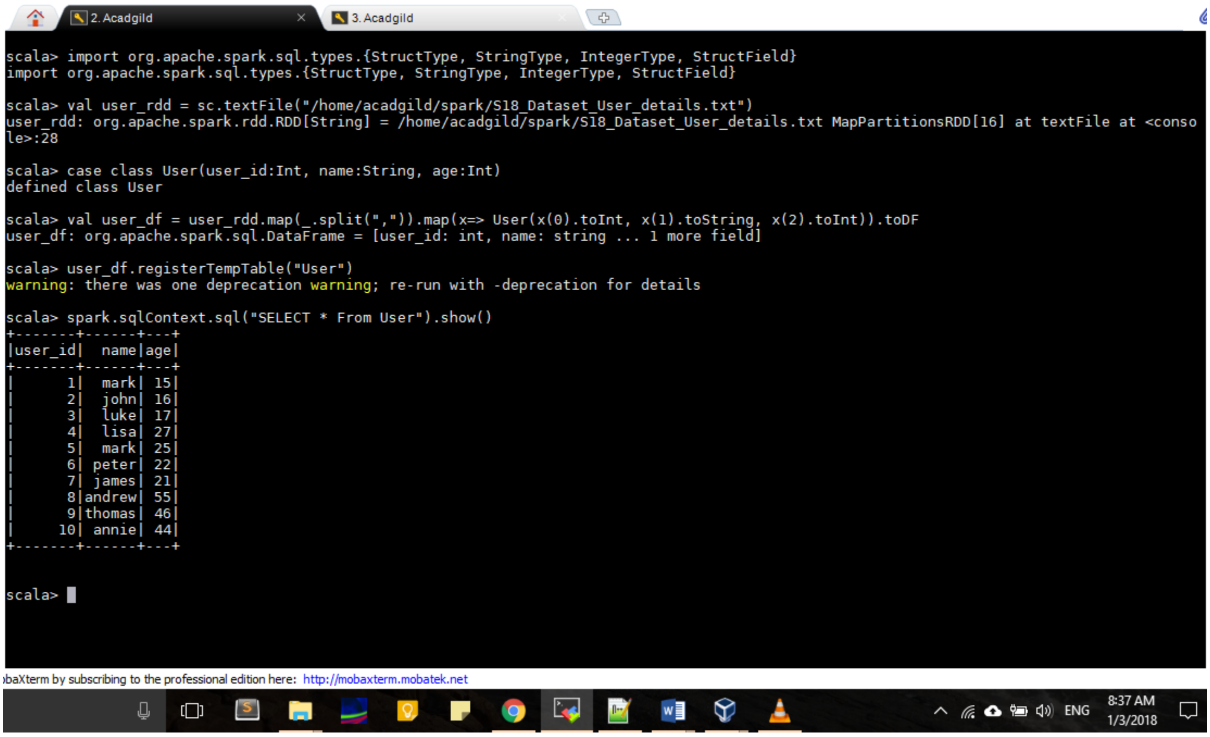
val user_rdd = sc.textFile("/home/acadgild/assignment_18.1/S18_Dataset_User_details.txt")

case class User(user_id:Int, name:String, age:Int)

val user_df = user_rdd.map(_.split(",")).map(x=> User(x(0).toInt, x(1).toString, x(2).toInt)).toDF

user_df.registerTempTable("User")

spark.sqlContext.sql("SELECT * From User").show()



```
scala> import org.apache.spark.sql.types.{StructType, StringType, IntegerType, StructField}
import org.apache.spark.sql.types.{StructType, StringType, IntegerType, StructField}

scala> val user_rdd = sc.textFile("/home/acadgild/spark/S18_Dataset_User_details.txt")
user_rdd: org.apache.spark.rdd.RDD[String] = /home/acadgild/spark/S18_Dataset_User_details.txt MapPartitionsRDD[16] at textFile at <conso
le>:28

scala> case class User(user_id:Int, name:String, age:Int)
defined class User

scala> val user_df = user_rdd.map(_.split(",")).map(x=> User(x(0).toInt, x(1).toString, x(2).toInt)).toDF
user_df: org.apache.spark.sql.DataFrame = [user_id: int, name: string ... 1 more field]

scala> user_df.registerTempTable("User")
warning: there was one deprecation warning; re-run with -deprecation for details

scala> spark.sqlContext.sql("SELECT * From User").show()
+-------+------+---+
|user_id|  name|age|
+-------+------+---+
|      1|  mark| 15|
|      2|  john| 16|
|      3|  luke| 17|
|      4|  lisa| 27|
|      5|  mark| 25|
|      6| peter| 22|
|      7| james| 21|
|      8|andrew| 55|
|      9|thomas| 46|
|     10| annie| 44|
+-------+------+---+


scala> 
```

obaXterm by subscribing to the professional edition here: http://mobaxterm.mobatek.net
ENG
8:37 AM
1/3/2018

Step2: Create a temporary table Travel

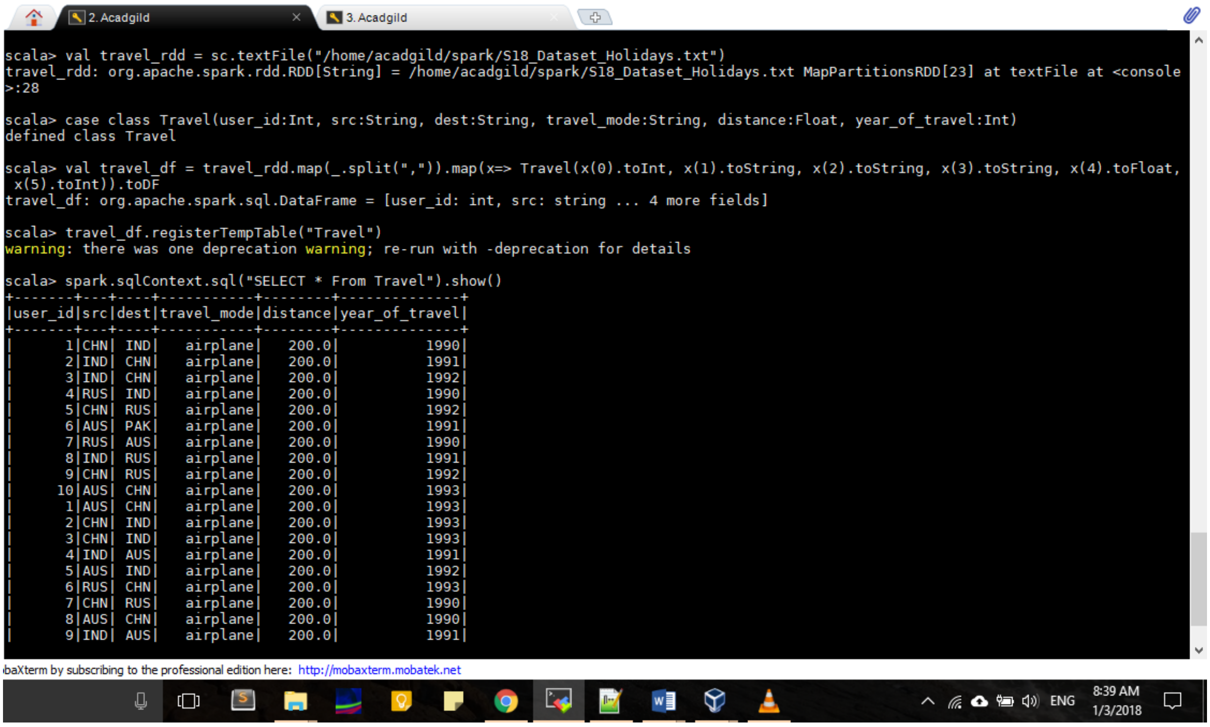
val travel_rdd = sc.textFile("/home/acadgild/spark/S18_Dataset_Holidays.txt")

case class Travel(user_id:Int, src:String, dest:String, travel_mode:String, distance:Float, year_of_travel:Int)

val travel_df = travel_rdd.map(_.split(",")).map(x=> Travel(x(0).toInt, x(1).toString, x(2).toString, x(3).toString, x(4).toFloat, x(5).toInt)).toDF

travel_df.registerTempTable("Travel")

spark.sqlContext.sql("SELECT * From Travel").show()



```
scala> val travel_rdd = sc.textFile("/home/acadgild/spark/S18_Dataset_Holidays.txt")
travel_rdd: org.apache.spark.rdd.RDD[String] = /home/acadgild/spark/S18_Dataset_Holidays.txt MapPartitionsRDD[23] at textFile at <console
>:28

scala> case class Travel(user_id:Int, src:String, dest:String, travel_mode:String, distance:Float, year_of_travel:Int)
defined class Travel

scala> val travel_df = travel_rdd.map(_.split(",")).map(x=> Travel(x(0).toInt, x(1).toString, x(2).toString, x(3).toString, x(4).toFloat,
 x(5).toInt)).toDF
travel_df: org.apache.spark.sql.DataFrame = [user_id: int, src: string ... 4 more fields]

scala> travel_df.registerTempTable("Travel")
warning: there was one deprecation warning; re-run with -deprecation for details

scala> spark.sqlContext.sql("SELECT * From Travel").show()
+-------+---+----+-----------+--------+--------------+
|user_id|src|dest|travel_mode|distance|year_of_travel|
+-------+---+----+-----------+--------+--------------+
|      1|CHN| IND|   airplane|   200.0|          1990|
|      2|IND| CHN|   airplane|   200.0|          1991|
|      3|IND| CHN|   airplane|   200.0|          1992|
|      4|RUS| IND|   airplane|   200.0|          1990|
|      5|CHN| RUS|   airplane|   200.0|          1992|
|      6|AUS| PAK|   airplane|   200.0|          1991|
|      7|RUS| AUS|   airplane|   200.0|          1990|
|      8|IND| RUS|   airplane|   200.0|          1991|
|      9|CHN| RUS|   airplane|   200.0|          1992|
|     10|AUS| CHN|   airplane|   200.0|          1993|
|      1|AUS| CHN|   airplane|   200.0|          1993|
|      2|CHN| IND|   airplane|   200.0|          1993|
|      3|CHN| IND|   airplane|   200.0|          1993|
|      4|IND| AUS|   airplane|   200.0|          1991|
|      5|AUS| IND|   airplane|   200.0|          1992|
|      6|RUS| CHN|   airplane|   200.0|          1993|
|      7|CHN| RUS|   airplane|   200.0|          1990|
|      8|AUS| CHN|   airplane|   200.0|          1990|
|      9|IND| AUS|   airplane|   200.0|          1991|
```

baXterm by subscribing to the professional edition here: http://mobaxterm.mobatek.net

Step3: Create temporary table Transport

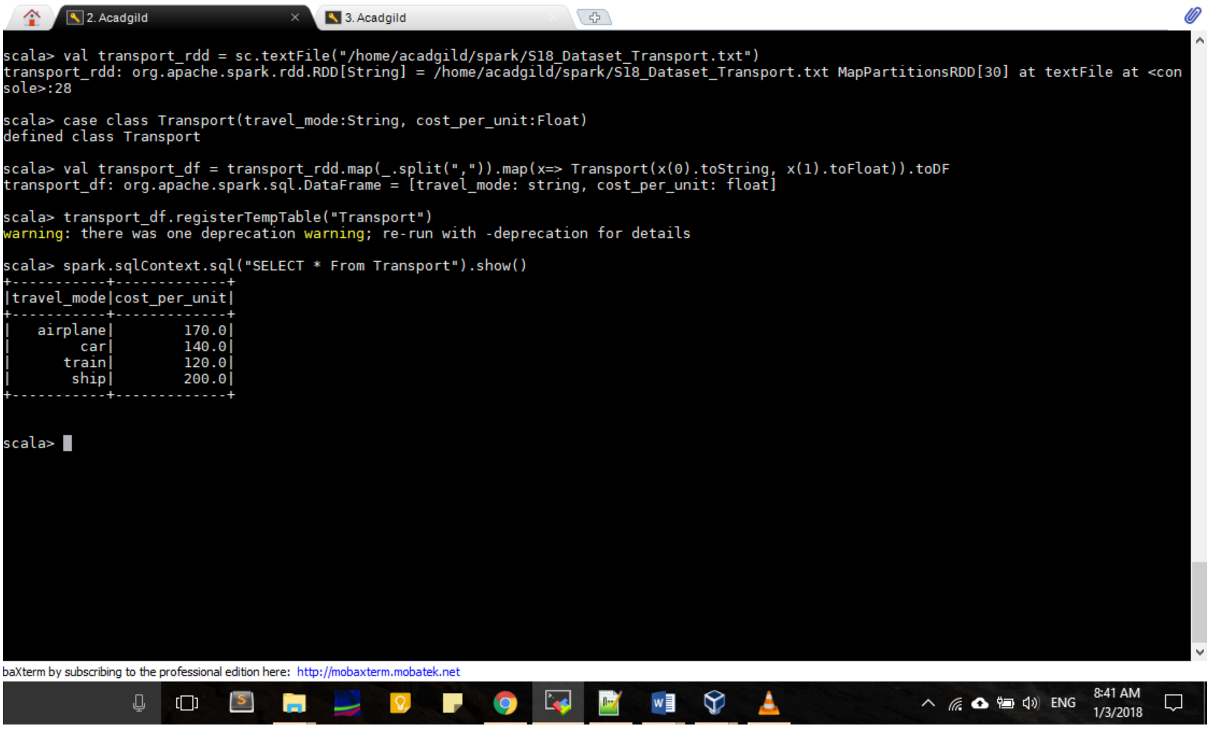
val transport_rdd = sc.textFile("/home/acadgild/spark/S18_Dataset_Transport.txt")

case class Transport(travel_mode:String, cost_per_unit:Float)

val transport_df = transport_rdd.map(_.split(",")).map(x=> Transport(x(0).toString, x(1).toFloat)).toDF

transport_df.registerTempTable("Transport")

spark.sqlContext.sql("SELECT * From Transport").show()

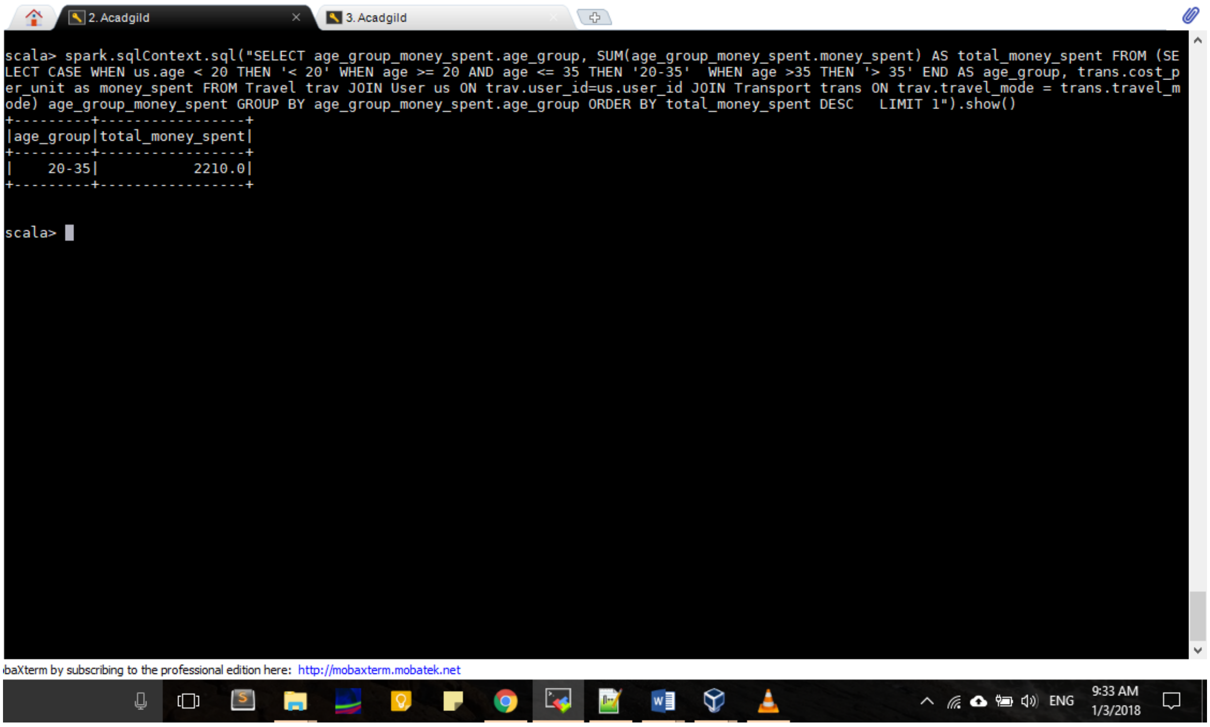


```
scala> val transport_rdd = sc.textFile("/home/acadgild/spark/S18_Dataset_Transport.txt")
transport_rdd: org.apache.spark.rdd.RDD[String] = /home/acadgild/spark/S18_Dataset_Transport.txt MapPartitionsRDD[30] at textFile at <con
sole>:28

scala> case class Transport(travel_mode:String, cost_per_unit:Float)
defined class Transport

scala> val transport_df = transport_rdd.map(_.split(",")).map(x=> Transport(x(0).toString, x(1).toFloat)).toDF
transport_df: org.apache.spark.sql.DataFrame = [travel_mode: string, cost_per_unit: float]

scala> transport_df.registerTempTable("Transport")
warning: there was one deprecation warning; re-run with -deprecation for details

scala> spark.sqlContext.sql("SELECT * From Transport").show()
+-----------+-------------+
|travel_mode|cost_per_unit|
+-----------+-------------+
|   airplane|        170.0|
|        car|        140.0|
|      train|        120.0|
|       ship|        200.0|
+-----------+-------------+


scala> 
```

baXterm by subscribing to the professional edition here: http://mobaxterm.mobatek.net

1) Considering age groups of < 20, 20-35, 35 >, Which age group spends the most amount of money travelling.

```
spark.sqlContext.sql("SELECT age_group_money_spent.age_group,
SUM(age_group_money_spent.money_spent) AS total_money_spent FROM (SELECT CASE
WHEN us.age < 20 THEN '< 20' WHEN age >= 20 AND age <= 35 THEN '20-35'  WHEN age >35
THEN '> 35' END AS age_group, trans.cost_per_unit as money_spent FROM Travel trav JOIN User
us ON trav.user_id=us.user_id JOIN Transport trans ON trav.travel_mode = trans.travel_mode)
age_group_money_spent GROUP BY age_group_money_spent.age_group ORDER BY
total_money_spent DESC   LIMIT 1").show()
```

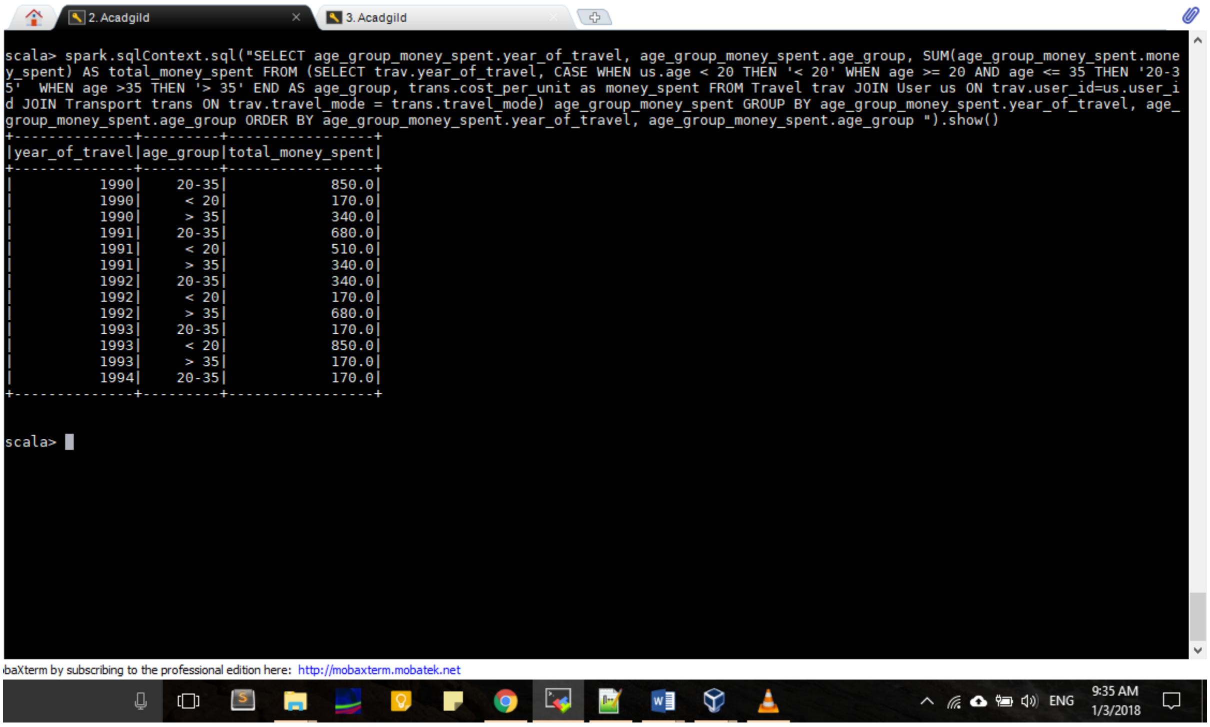


2. Acadgild
3. Acadgild
baXterm by subscribing to the professional edition here: http://mobaxterm.mobatek.net
ENG
9:33 AM
1/3/2018

```
scala> spark.sqlContext.sql("SELECT age_group_money_spent.age_group, SUM(age_group_money_spent.money_spent) AS total_money_spent FROM (SE
LECT CASE WHEN us.age < 20 THEN '< 20' WHEN age >= 20 AND age <= 35 THEN '20-35'  WHEN age >35 THEN '> 35' END AS age_group, trans.cost_p
er_unit as money_spent FROM Travel trav JOIN User us ON trav.user_id=us.user_id JOIN Transport trans ON trav.travel_mode = trans.travel_m
ode) age_group_money_spent GROUP BY age_group_money_spent.age_group ORDER BY total_money_spent DESC   LIMIT 1").show()
+---------+-----------------+
|age_group|total_money_spent|
+---------+-----------------+
|    20-35|           2210.0|
+---------+-----------------+


scala> 
```

2) What is the amount spent by each age-group, every year in travelling?

spark.sqlContext.sql("SELECT age_group_money_spent.year_of_travel, age_group_money_spent.age_group, SUM(age_group_money_spent.money_spent) AS total_money_spent FROM (SELECT trav.year_of_travel, CASE WHEN us.age < 20 THEN '< 20' WHEN age >= 20 AND age <= 35 THEN '20-35' WHEN age >35 THEN '> 35' END AS age_group, trans.cost_per_unit as money_spent FROM Travel trav JOIN User us ON trav.user_id=us.user_id JOIN Transport trans ON trav.travel_mode = trans.travel_mode) age_group_money_spent GROUP BY age_group_money_spent.year_of_travel, age_group_money_spent.age_group ORDER BY age_group_money_spent.year_of_travel, age_group_money_spent.age_group ").show()



```
scala> spark.sqlContext.sql("SELECT age_group_money_spent.year_of_travel, age_group_money_spent.age_group, SUM(age_group_money_spent.mone
y_spent) AS total_money_spent FROM (SELECT trav.year_of_travel, CASE WHEN us.age < 20 THEN '< 20' WHEN age >= 20 AND age <= 35 THEN '20-3
5'  WHEN age >35 THEN '> 35' END AS age_group, trans.cost_per_unit as money_spent FROM Travel trav JOIN User us ON trav.user_id=us.user_i
d JOIN Transport trans ON trav.travel_mode = trans.travel_mode) age_group_money_spent GROUP BY age_group_money_spent.year_of_travel, age_
group_money_spent.age_group ORDER BY age_group_money_spent.year_of_travel, age_group_money_spent.age_group ").show()
+--------------+---------+-----------------+
|year_of_travel|age_group|total_money_spent|
+--------------+---------+-----------------+
|          1990|    20-35|            850.0|
|          1990|     < 20|            170.0|
|          1990|     > 35|            340.0|
|          1991|    20-35|            680.0|
|          1991|     < 20|            510.0|
|          1991|     > 35|            340.0|
|          1992|    20-35|            340.0|
|          1992|     < 20|            170.0|
|          1992|     > 35|            680.0|
|          1993|    20-35|            170.0|
|          1993|     < 20|            850.0|
|          1993|     > 35|            170.0|
|          1994|    20-35|            170.0|
+--------------+---------+-----------------+


scala> 
```

baXterm by subscribing to the professional edition here: http://mobaxterm.mobatek.net
ENG
9:35 AM
1/3/2018