Session 18 : RDD'S CONTD. & INTRODUCTION TO DATAFRAMES

Assignment 1

**Problem Statement**

1) What is the distribution of the total number of air-travelers per year

2) What is the total air distance covered by each user per year

3) Which user has travelled the largest distance till date

4) What is the most preferred destination for all users.

Use below link to download the dataset:

https://drive.google.com/drive/folders/0B\_P3pWagdIrrVThBaUdVSUtzbms

**Solution:-**

Starting Spark shell

Command - spark-shell



Loading data to dataframe:-

scala> val holidaysRDD = sc.textFile("S18\_Dataset\_Holidays.txt");

scala> val holidaysDF= holidaysRDD.map(lines=>lines.split(",")).map(arrays => (arrays(0),arrays(1),arrays(2),arrays(3),arrays(4),arrays(5))).toDF("Person\_ID","Source","Destination","Mode","Fare","Year");



scala> val transportRDD = sc.textFile("S18\_Dataset\_Transport.txt");

scala> val transportDF= transportRDD.map(lines=>lines.split(",")).map(arrays=>(arrays(0),arrays(1))).toDF("Transport\_Name","Transport\_ID");



scala> val userRDD = sc.textFile("S18\_Dataset\_User\_details.txt");

scala> val userDF= userRDD.map(lines=>lines.split(",")).map(arrays=>(arrays(0),arrays(1),arrays(2))).toDF("Person\_ID","Name","Age");



Creating temporary view out of all data frames.

scala> holidaysDF.registerTempTable("holidays");

scala> transportDF.registerTempTable("transport");

scala> userDF.registerTempTable("user");



All the datasets have been loaded in to temporary tables.

These temporary tables and dataframes would be used to find solution to problem statements

Listing out temporary tables and column names:-

holidays -> "Person\_ID","Source","Destination","Mode","Fare","Year"

transport -> "Transport\_Name","Transport\_ID"

user -> "Person\_ID","Name","Age"

Starting with problem statements now:-

**1) What is the distribution of the total number of air-travelers per year**

**scala>** holidaysDF.groupBy("year").count().show;



**2) What is the total air distance covered by each user per year**

By mistake I have mentioned the column name as “Fare”. Its actually the distance travelled.

So using “Fare” column as distance.

**scala>** val sqlContext = new org.apache.spark.sql.SQLContext(sc)

**scala>** import sqlContext.implicits.\_

**scala>** val joinDF = holidaysDF.as("d1").join(userDF.as("d2"), $"d1.Person\_ID" === $"d2.Person\_ID").select($"d2.Name", $"d1.Year", $"d1.Fare");

**scala>** val problem2DF= joinDF.groupBy("Name","Year").agg(sum("Fare"));

**scala>** problem2DF.collect.foreach(println);







Rest of output --->>>>>



**3) Which user has travelled the largest distance till date**

**scala>** val problem3DF= joinDF.groupBy("Name").agg(sum("Fare")).orderBy($"sum(Fare)".desc).show(1);



***Mark has travelled the most with 1600 Km.***

**4) What is the most preferred destination for all users.**

**scala>** val problem4DF = holidaysDF.groupBy("Destination").count().orderBy($"count".desc).show(1);



***IND (India) with visit count of 9 is the most preferred destination for all users***.