Assignment – Spark SQL

Task 1

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.

5)Which route is generating the most revenue per year

6) What is the total amount spent by every user on air-travel per year

Code:

```
package spark.basic.cl18
import org.apache.spark.sql.SQLContext
import org.apache.spark.sql.SparkSession
import org.apache.spark.sql.expressions.Window
import org.apache.spark.sql.types.IntegerType
import org.apache.spark.sql.functions._


  object ssql extends App {

    val sparkSession = SparkSession.builder.master("local")
      .appName("spark").getOrCreate()

    val sparkcontext = sparkSession.sparkContext

    val transport =
sparkSession.read.csv("/home/acadgild/Downloads/S20_Dataset_Transport.txt")
      .toDF("Transport_mode","Cost_per_unit")
    //transport.show()

    val holiday =
sparkSession.read.csv("/home/acadgild/Downloads/S20_Dataset_Holidays.txt")
      .toDF("Id","Source","Destination","Transport_mode","Distance","Year")
    //holiday.show()

    val userdetails =
sparkSession.read.csv("/home/acadgild/Downloads/S20_Dataset_User_details.txt")
      .toDF("Id","Name","Age")
    //userdetails.show()
    import sparkSession.implicits._

    //query 1
    val uh = userdetails.join(holiday,"Id").
    filter($"Transport_mode" === "airplane")
       .groupBy("Year").count().select($"Year", $"count" / 10 * 100 , $"count" ).show()

    //query2

    val q2 = userdetails.join(holiday,"Id")
      .select($"Name",$"Distance".cast(IntegerType),$"Year")
```

```
    .where($"Transport_mode" === "airplane")
    .groupBy("Name","Year")
    .sum("Distance").orderBy("Name","Year").show()

  //Query3


  val q3 = userdetails.join(holiday,"Id")
    .select($"Name",$"Distance".cast(IntegerType),$"Year")
    .groupBy("Name")
    .sum("Distance")
    .sort($"sum(Distance)".desc)
    .limit(1)
    .show()




  //Query 4

  val q4 = userdetails.join(holiday,"Id")
    .select($"Name",$"Destination")
    .groupBy($"Destination")
    .count()
    .orderBy($"count".desc)
    .limit(1).show()

  //query 5


  val q5 = transport.join(holiday,"Transport_mode")
    .select($"Source",$"Destination",$"Year",$"Cost_per_unit".cast(IntegerType))
    .groupBy("Source","Destination","Year")
    .sum("Cost_per_unit")
    .sort($"sum(Cost_per_unit)".desc)
    .limit(1)
    .show()

  //query6

  val q6 = userdetails.join(holiday,"Id")
    .join(transport,"Transport_mode")
    .select($"Name",$"Cost_per_unit".cast(IntegerType),$"Year")
    .where($"Transport_mode" === "airplane")
    .groupBy($"Name",$"year")
    .sum("Cost_per_unit")
    .sort($"sum(Cost_per_unit)".desc)
    .show()


}
```

Output:

1

```
18/08/02 17:56:19 INFO TaskSchedulerImpl: Removed TaskSet 13.
18/08/02 17:56:19 INFO DAGScheduler: ResultStage 13 (show at
18/08/02 17:56:19 INFO DAGScheduler: Job 8 finished: show at
+----+-------------------+-----+
|Year|((count / 10) * 100)|count|
+----+-------------------+-----+
|1992|               70.0|    7|
|1994|               10.0|    1|
|1993|               70.0|    7|
|1990|               80.0|    8|
|1991|               90.0|    9|
+----+-------------------+-----+

18/08/02 17:56:20 INFO FileSourceStrategy: Pruning directorie
18/08/02 17:56:20 INFO FileSourceStrategy: Post-Scan Filters:
```

2

```
18/08/02 17:56:24 INFO DAGScheduler: Job 10 fin
18/08/02 17:56:24 INFO TaskSchedulerImpl: Remov
+------+----+-------------+
|  Name|Year|sum(Distance)|
+------+----+-------------+
|andrew|1990|          200|
|andrew|1991|          200|
|andrew|1992|          200|
| annie|1990|          200|
| annie|1992|          200|
| annie|1993|          200|
| james|1990|          600|
|  john|1991|          400|
|  john|1993|          200|
|  lisa|1990|          400|
|  lisa|1991|          200|
|  luke|1991|          200|
|  luke|1992|          200|
|  luke|1993|          200|
|  mark|1990|          200|
|  mark|1991|          200|
|  mark|1992|          400|
|  mark|1993|          600|
|  mark|1994|          200|
| peter|1991|          400|
+------+----+-------------+
only showing top 20 rows

18/08/02 17:56:24 INFO FileSourceStrategy: Prun
18/08/02 17:56:24 INFO FileSourceStrategy: Post
18/08/02 17:56:24 INFO FileSourceStrategy: Outp
```

3

```
18/08/02 17:56:27 INFO TaskSchedulerImpl: Removed TaskSet
18/08/02 17:56:27 INFO DAGScheduler: ResultStage 19 (show
18/08/02 17:56:27 INFO DAGScheduler: Job 12 finished: show
+----+-------------+
|Name|sum(Distance)|
+----+-------------+
|mark|         1600|
+----+-------------+

18/08/02 17:56:27 INFO FileSourceStrategy: Pruning directo
18/08/02 17:56:27 INFO FileSourceStrategy: Post-Scan Filte
18/08/02 17:56:27 INFO FileSourceStrategy: Output Data Sch
```

4

```
18/08/02 17:56:30 INFO ShuffleBlockFetcherIterator: Star
18/08/02 17:56:30 INFO Executor: Finished task 183.0 in
+-----------+-----+
|Destination|count|
+-----------+-----+
|        IND|    9|
+-----------+-----+

18/08/02 17:56:30 INFO TaskSetManager: Finished task 183
18/08/02 17:56:30 INFO TaskSchedulerImpl: Removed TaskSe
18/08/02 17:56:30 INFO DAGScheduler: ResultStage 22 (sho
18/08/02 17:56:30 INFO DAGScheduler: Job 14 finished: sh
18/08/02 17:56:30 INFO FileSourceStrategy: Pruning direc
18/08/02 17:56:30 INFO FileSourceStrategy: Post-Scan Fil
```

5

```
18/08/02 17:56:33 INFO DAGScheduler: Job 16 finished: show at ss(
+------+-----------+----+------------------+
|Source|Destination|Year|sum(Cost_per_unit)|
+------+-----------+----+------------------+
|   CHN|        RUS|1992|               340|
+------+-----------+----+------------------+

18/08/02 17:56:33 INFO FileSourceStrategy: Pruning directories w:
18/08/02 17:56:33 INFO FileSourceStrategy: Post-Scan Filters: isr
18/08/02 17:56:33 INFO FileSourceStrategy: Output Data Schema: s:
18/08/02 17:56:33 INFO FileSourceScanExec: Pushed Filters: IsNotN
18/08/02 17:56:33 INFO FileSourceStrategy: Pruning directories w:
```

6

```
18/08/02 17:56:36 INFO DAGScheduler: ResultStage 2
18/08/02 17:56:36 INFO DAGScheduler: Job 19 finish
+------+----+------------------+
|  Name|year|sum(Cost_per_unit)|
+------+----+------------------+
|  mark|1993|               510|
| james|1990|               510|
| peter|1991|               340|
|  lisa|1990|               340|
|  john|1991|               340|
|thomas|1992|               340|
|  mark|1992|               340|
|andrew|1990|               170|
|  lisa|1991|               170|
| annie|1993|               170|
|  mark|1990|               170|
|andrew|1992|               170|
| annie|1990|               170|
|  john|1993|               170|
|thomas|1991|               170|
| peter|1993|               170|
|  mark|1991|               170|
| annie|1992|               170|
|  luke|1991|               170|
|  luke|1993|               170|
+------+----+------------------+
only showing top 20 rows

18/08/02 17:56:36 INFO SparkContext: Invoking stop
18/08/02 17:56:37 INFO SparkUI: Stopped Spark web
18/08/02 17:56:37 INFO MapOutputTrackerMasterEndpo
18/08/02 17:56:37 INFO MemoryStore: MemoryStore cl
18/08/02 17:56:37 INFO BlockManager: BlockManager
```