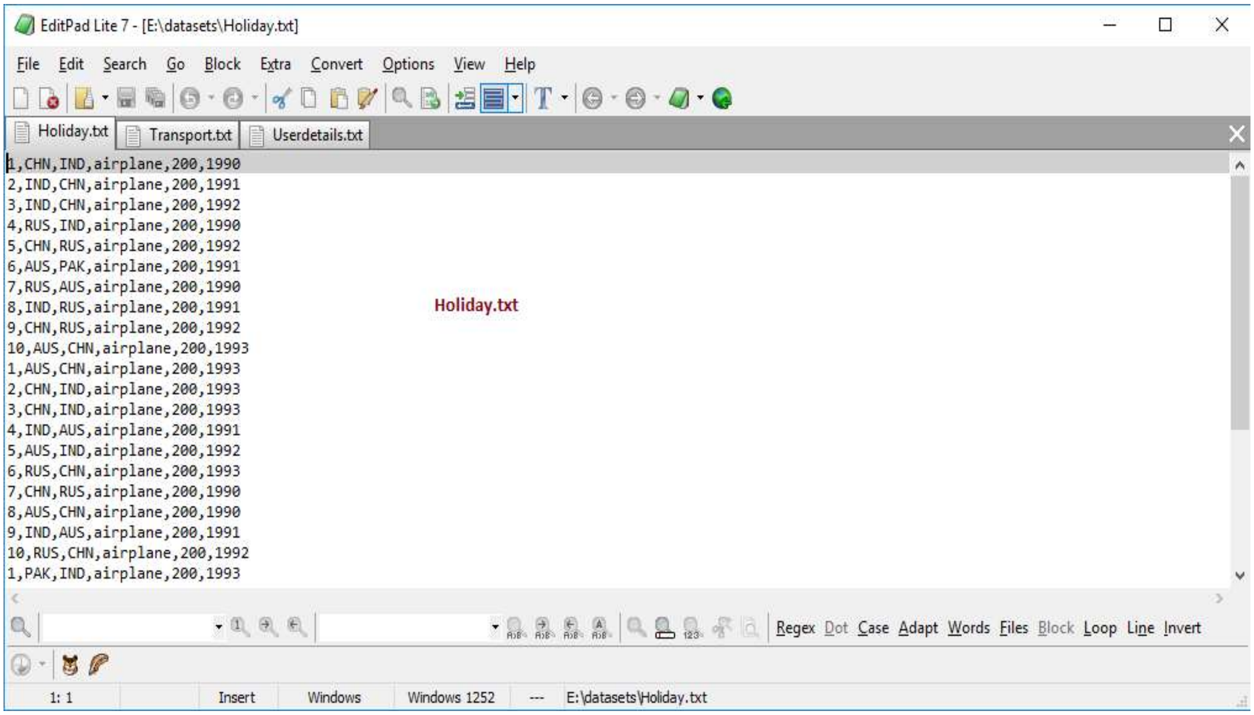
BIG DATA HADOOP & SPARK TRAINING

ASSIGNMENT ON SPARK SQL I

Rashmi Krishna

Task 1

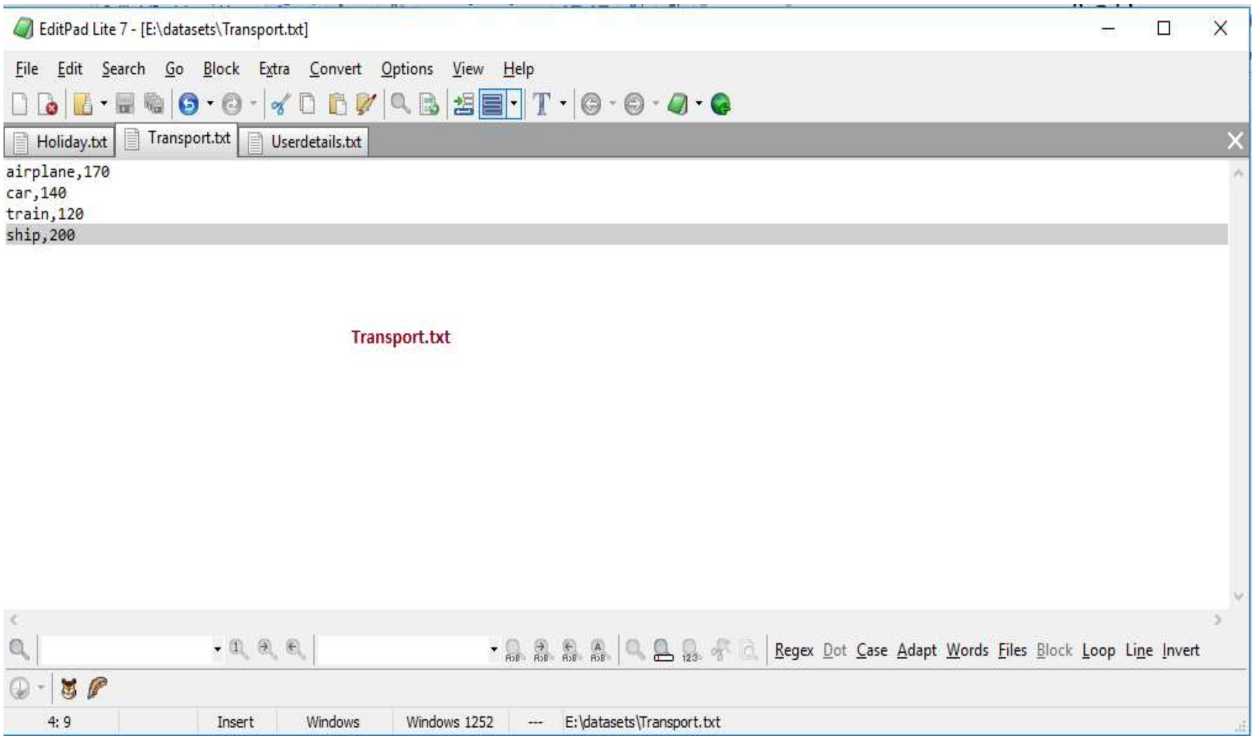
Input files required to perform the below operations:


EditPad Lite 7 - [E:\datasets\Holiday.txt]
File Edit Search Go Block Extra Convert Options View Help
Holiday.txt
Transport.txt
Userdetails.txt
Regex Dot Case Adapt Words Files Block Loop Line Invert
1: 1
Insert
Windows
Windows 1252

E:\datasets\Holiday.txt

```
1,CHN,IND,airplane,200,1990
2,IND,CHN,airplane,200,1991
3,IND,CHN,airplane,200,1992
4,RUS,IND,airplane,200,1990
5,CHN,RUS,airplane,200,1992
6,AUS,PAK,airplane,200,1991
7,RUS,AUS,airplane,200,1990
8,IND,RUS,airplane,200,1991
9,CHN,RUS,airplane,200,1992
10,AUS,CHN,airplane,200,1993
1,AUS,CHN,airplane,200,1993
2,CHN,IND,airplane,200,1993
3,CHN,IND,airplane,200,1993
4,IND,AUS,airplane,200,1991
5,AUS,IND,airplane,200,1992
6,RUS,CHN,airplane,200,1993
7,CHN,RUS,airplane,200,1990
8,AUS,CHN,airplane,200,1990
9,IND,AUS,airplane,200,1991
10,RUS,CHN,airplane,200,1992
1,PAK,IND,airplane,200,1993
```

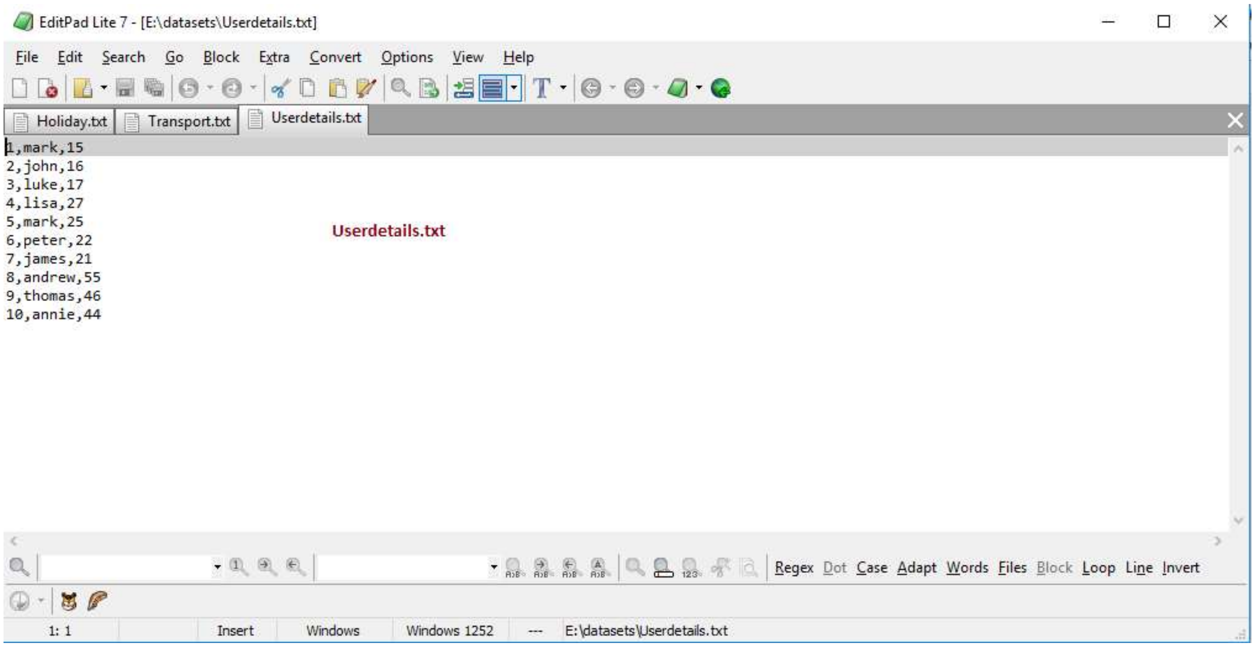
Holiday.txt


EditPad Lite 7 - [E:\datasets\Transport.txt]
File Edit Search Go Block Extra Convert Options View Help
Holiday.txt
Transport.txt
Userdetails.txt
Regex Dot Case Adapt Words Files Block Loop Line Invert
4: 9
Insert
Windows
Windows 1252

E:\datasets\Transport.txt

```
airplane,170
car,140
train,120
ship,200
```

Transport.txt


EditPad Lite 7 - [E:\datasets\Userdetails.txt]
File Edit Search Go Block Extra Convert Options View Help
Holiday.txt
Transport.txt
Userdetails.txt
Regex Dot Case Adapt Words Files Block Loop Line Invert
1: 1
Insert
Windows
Windows 1252

E:\datasets\Userdetails.txt

```
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
```

Userdetails.txt

Below tasks are performed in **IntelliJ IDEA.**

Created a Spark-Scala Application and performed following steps to read the above mentioned text data to spark.

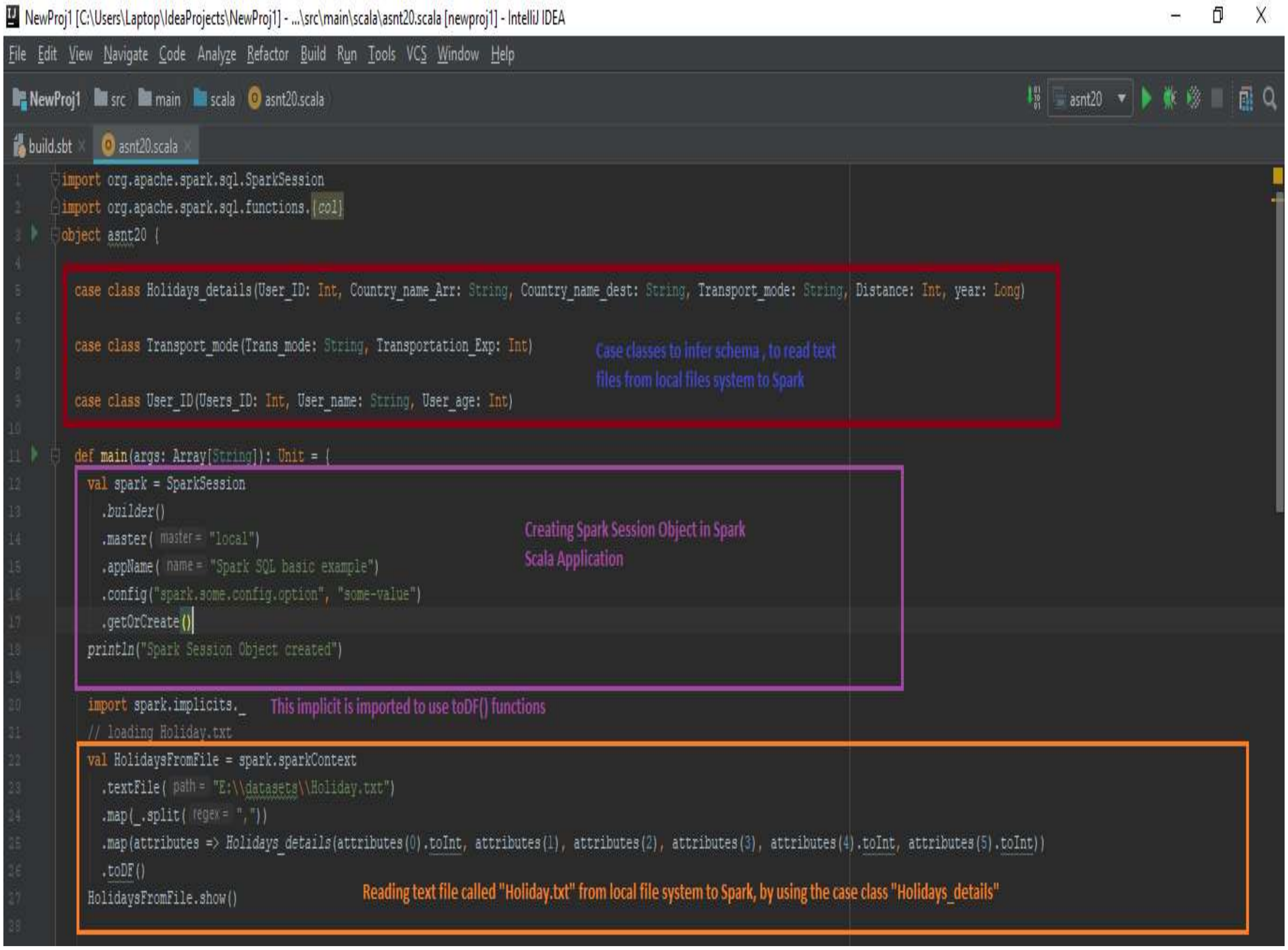
- ➢ Created case class for each text files which represents the schema for the respective text files.
- ○ **case class** Holidays_details(User_ID: Int, Country_name_Arr: String, Country_name_dest: String, Transport_mode: String, Distance: Int, year: Long)
- ○ **case class** Transport_mode(Trans_mode: String, Transportation_Exp: Int)
- ○ **case class** User_ID(Users_ID: Int, User_name: String, User_age: Int)
- ➢ created a spark session object for spark application
  - ○ val spark = SparkSession
  - ○ .builder()
  - ○ .master("local")
  - ○ .appName("Spark SQL basic example")
  - ○ .config("spark.some.config.option", "some-value")
  - ○ .getOrCreate()
  - ○ println("Spark Session Object created")

- ➢ Imported spark implicits to convert RDD's to DataFrames implicitly
  - ○ import spark.implicits._

- ➢ created a spark context object to read contents of text file to Spark
  - ○ val HolidaysFromFile = spark.sparkContext // create spark context object

- .textFile("E:\\datasets\\Holiday.txt")// path of the text file in the local file system
- .map(_.split(","))//splitting the input file base on ‘,’ seperator
- .map(attributes => Holidays_details(attributes(0).toInt, attributes(1), attributes(2), attributes(3), attributes(4).toInt, attributes(5).toInt)) //assisgning the inputs to attributes of case class “Holidays_details
- .toDF()// converting RDD to Data Frame
- HolidaysFromFile.show()
  // displaying contents of the variable in which contents of text file is stored

- Similarly different DataFrames are created for other two text files as well, as shown in the below screen shot.


NewProj1 [C:\Users\Laptop\IdeaProjects\NewProj1] - ...\src\main\scala\asnt20.scala [newproj1] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
NewProj1 src main scala asnt20.scala
asnt20
build.sbt
asnt20.scala

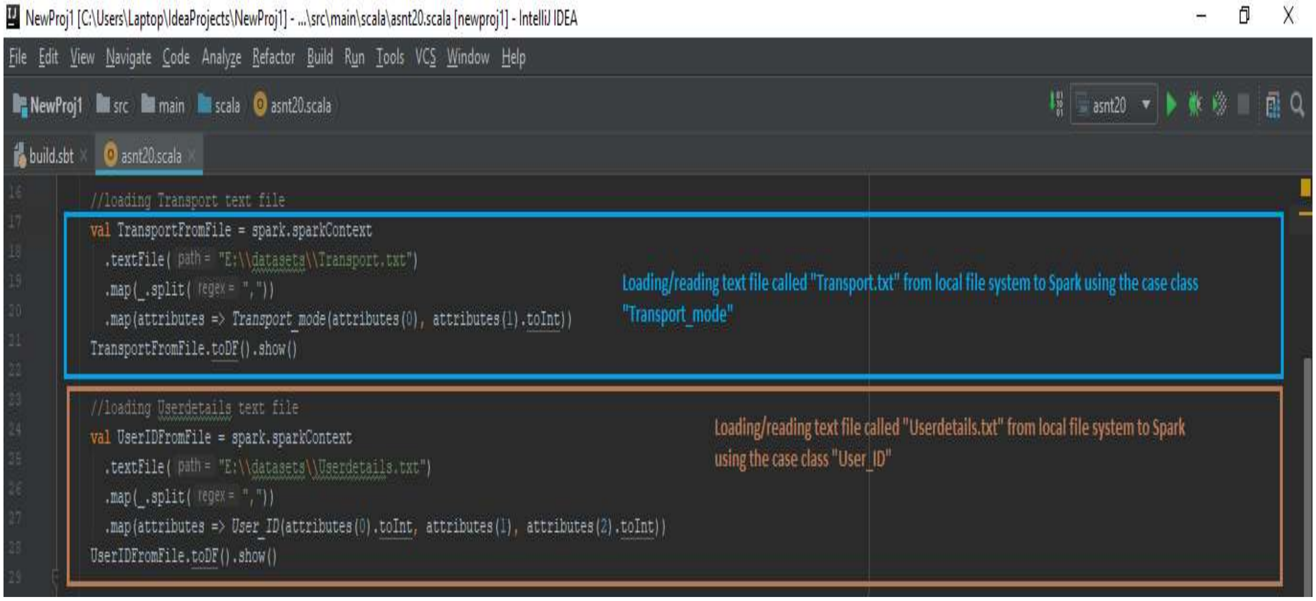
```scala
import org.apache.spark.sql.SparkSession
import org.apache.spark.sql.functions.{col}
object asnt20 {

  case class Holidays_details(User_ID: Int, Country_name_Arr: String, Country_name_dest: String, Transport_mode: String, Distance: Int, year: Long)

  case class Transport_mode(Trans_mode: String, Transportation_Exp: Int)

  case class User_ID(Users_ID: Int, User_name: String, User_age: Int)

  def main(args: Array[String]): Unit = {
    val spark = SparkSession
      .builder()
      .master("local")
      .appName("Spark SQL basic example")
      .config("spark.some.config.option", "some-value")
      .getOrCreate()
    println("Spark Session Object created")

    import spark.implicits._
    // loading Holiday.txt
    val HolidaysFromFile = spark.sparkContext
      .textFile("E:\\datasets\\Holiday.txt")
      .map(_.split(","))
      .map(attributes => Holidays_details(attributes(0).toInt, attributes(1), attributes(2), attributes(3), attributes(4).toInt, attributes(5).toInt))
      .toDF()
    HolidaysFromFile.show()
```

Case classes to infer schema , to read text files from local files system to Spark

Creating Spark Session Object in Spark Scala Application

This implicit is imported to use toDF() functions

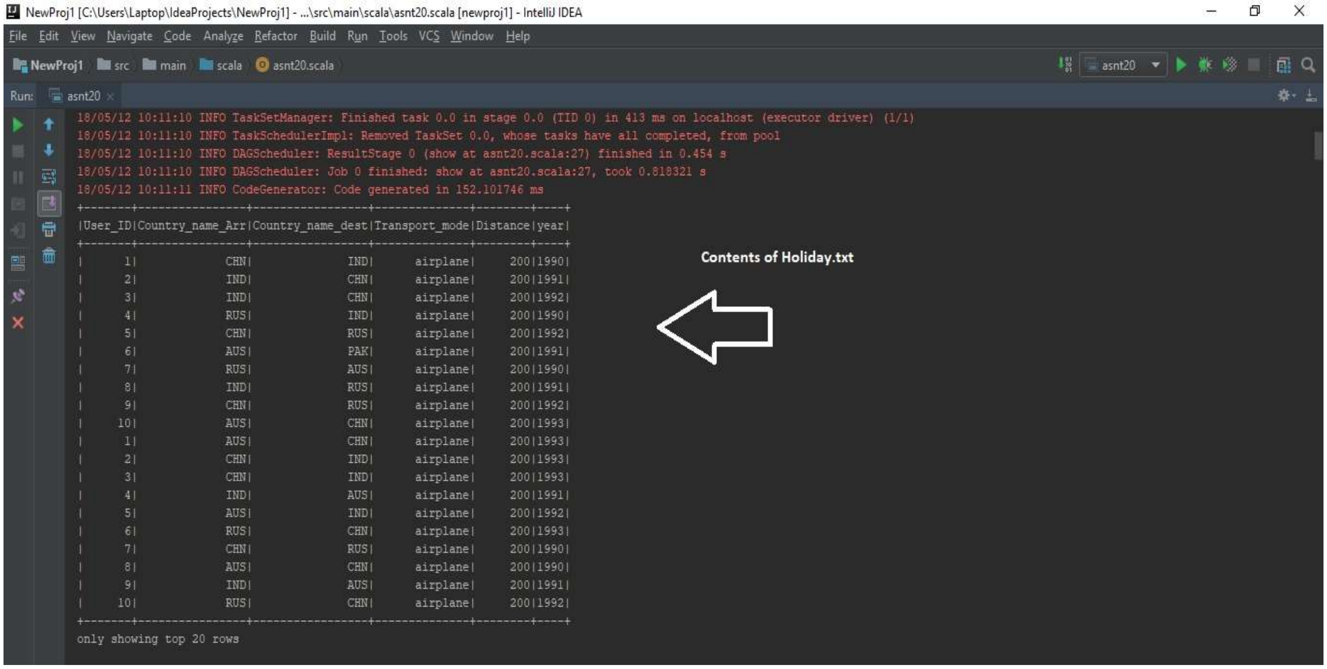
Reading text file called "Holiday.txt" from local file system to Spark, by using the case class "Holidays_details"


NewProj1 [C:\Users\Laptop\IdeaProjects\NewProj1] - ...\src\main\scala\asnt20.scala [newproj1] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
NewProj1 src main scala asnt20.scala
asnt20
build.sbt
asnt20.scala

```scala
    //loading Transport text file
    val TransportFromFile = spark.sparkContext
      .textFile( path = "E:\\datasets\\Transport.txt")
      .map(_.split( regex = ","))
      .map(attributes => Transport_mode(attributes(0), attributes(1).toInt))
    TransportFromFile.toDF().show()

    //loading Userdetails text file
    val UserIDFromFile = spark.sparkContext
      .textFile( path = "E:\\datasets\\Userdetails.txt")
      .map(_.split( regex = ","))
      .map(attributes => User_ID(attributes(0).toInt, attributes(1), attributes(2).toInt))
    UserIDFromFile.toDF().show()
```

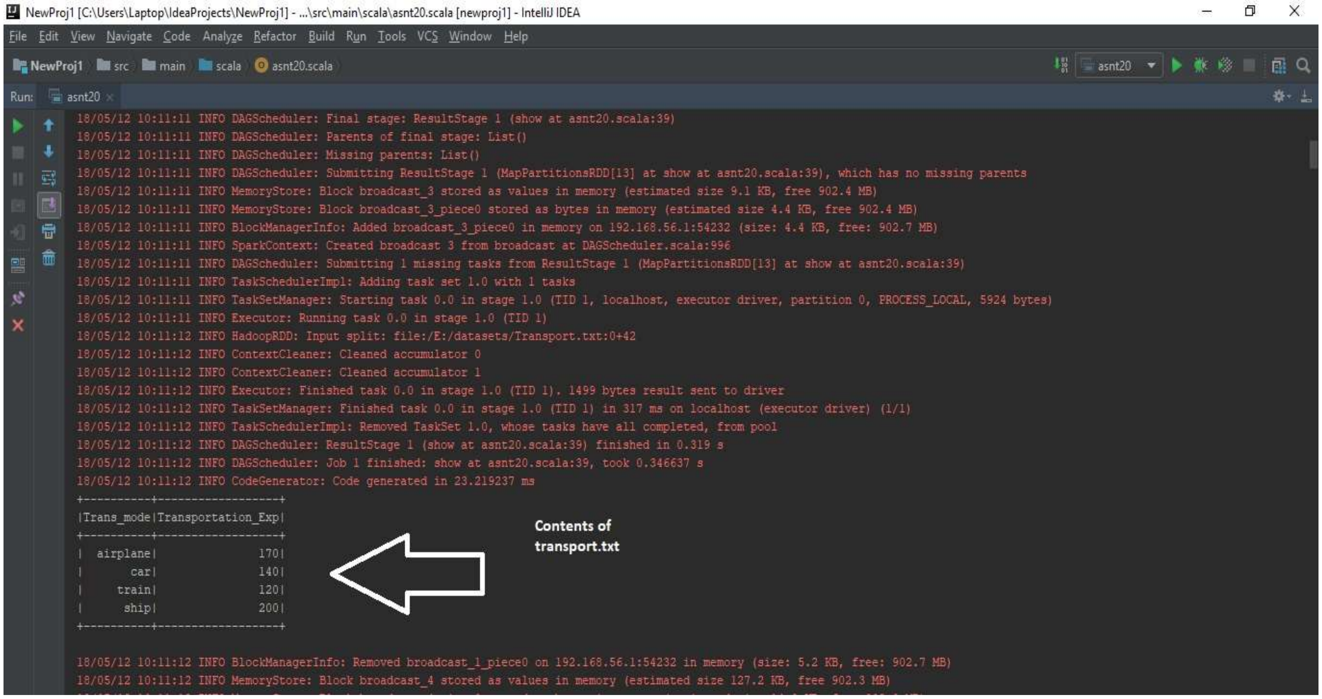
Loading/reading text file called "Transport.txt" from local file system to Spark using the case class "Transport_mode"

Loading/reading text file called "Userdetails.txt" from local file system to Spark using the case class "User_ID"


NewProj1 [C:\Users\Laptop\IdeaProjects\NewProj1] - ...\src\main\scala\asnt20.scala [newproj1] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
NewProj1 src main scala asnt20.scala
asnt20
Run: asnt20

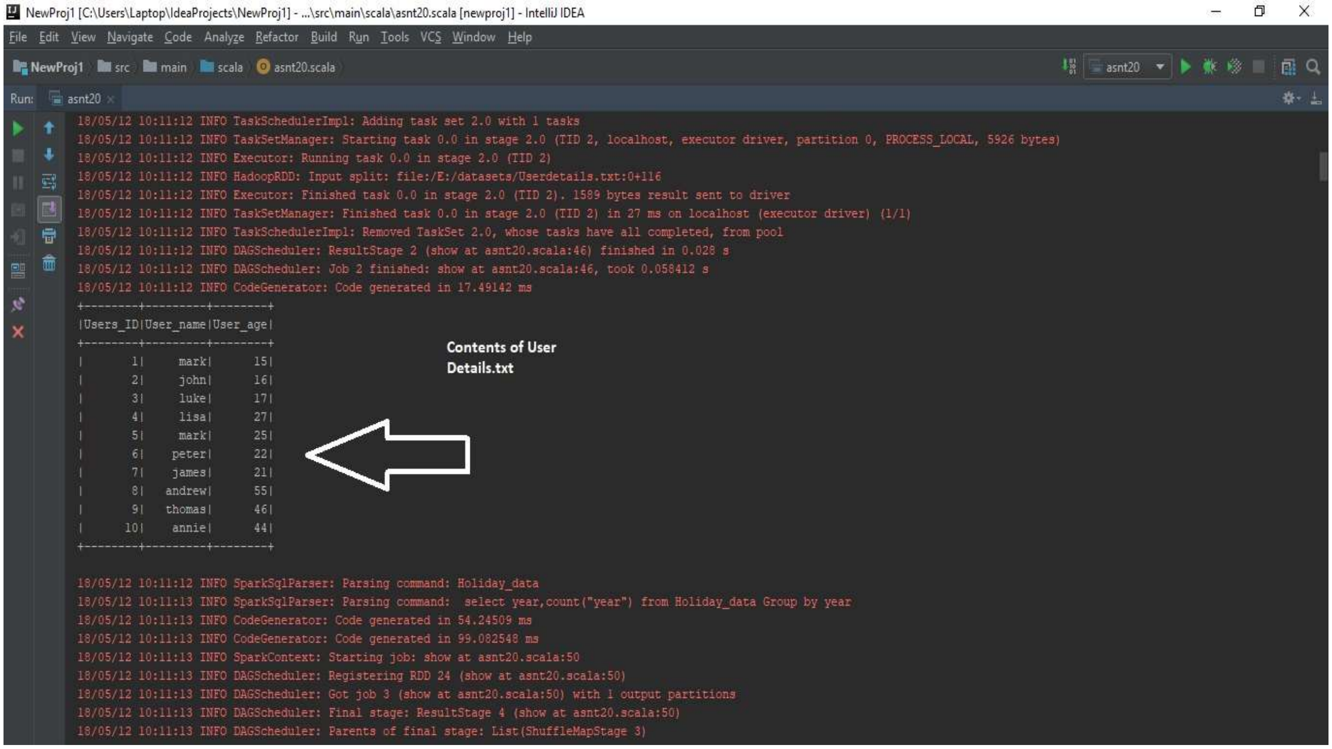
```
18/05/12 10:11:10 INFO TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 413 ms on localhost (executor driver) (1/1)
18/05/12 10:11:10 INFO TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool
18/05/12 10:11:10 INFO DAGScheduler: ResultStage 0 (show at asnt20.scala:27) finished in 0.454 s
18/05/12 10:11:10 INFO DAGScheduler: Job 0 finished: show at asnt20.scala:27, took 0.818321 s
18/05/12 10:11:11 INFO CodeGenerator: Code generated in 152.101746 ms
+-------+----------------+-----------------+--------------+--------+----+
|User_ID|Country_name_Arr|Country_name_dest|Transport_mode|Distance|year|
+-------+----------------+-----------------+--------------+--------+----+
|      1|             CHN|              IND|      airplane|     200|1990|
|      2|             IND|              CHN|      airplane|     200|1991|
|      3|             IND|              CHN|      airplane|     200|1992|
|      4|             RUS|              IND|      airplane|     200|1990|
|      5|             CHN|              RUS|      airplane|     200|1992|
|      6|             AUS|              PAK|      airplane|     200|1991|
|      7|             RUS|              AUS|      airplane|     200|1990|
|      8|             IND|              RUS|      airplane|     200|1991|
|      9|             CHN|              RUS|      airplane|     200|1992|
|     10|             AUS|              CHN|      airplane|     200|1993|
|      1|             AUS|              CHN|      airplane|     200|1993|
|      2|             CHN|              IND|      airplane|     200|1993|
|      3|             CHN|              IND|      airplane|     200|1993|
|      4|             IND|              AUS|      airplane|     200|1991|
|      5|             AUS|              IND|      airplane|     200|1992|
|      6|             RUS|              CHN|      airplane|     200|1993|
|      7|             CHN|              RUS|      airplane|     200|1990|
|      8|             AUS|              CHN|      airplane|     200|1990|
|      9|             IND|              AUS|      airplane|     200|1991|
|     10|             RUS|              CHN|      airplane|     200|1992|
+-------+----------------+-----------------+--------------+--------+----+
only showing top 20 rows
```

Contents of Holiday.txt


NewProj1 [C:\Users\Laptop\IdeaProjects\NewProj1] - ...\src\main\scala\asnt20.scala [newproj1] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
NewProj1 src main scala asnt20.scala
asnt20
Run: asnt20

```
18/05/12 10:11:11 INFO DAGScheduler: Final stage: ResultStage 1 (show at asnt20.scala:39)
18/05/12 10:11:11 INFO DAGScheduler: Parents of final stage: List()
18/05/12 10:11:11 INFO DAGScheduler: Missing parents: List()
18/05/12 10:11:11 INFO DAGScheduler: Submitting ResultStage 1 (MapPartitionsRDD[13] at show at asnt20.scala:39), which has no missing parents
18/05/12 10:11:11 INFO MemoryStore: Block broadcast_3 stored as values in memory (estimated size 9.1 KB, free 902.4 MB)
18/05/12 10:11:11 INFO MemoryStore: Block broadcast_3_piece0 stored as bytes in memory (estimated size 4.4 KB, free 902.4 MB)
18/05/12 10:11:11 INFO BlockManagerInfo: Added broadcast_3_piece0 in memory on 192.168.56.1:54232 (size: 4.4 KB, free: 902.7 MB)
18/05/12 10:11:11 INFO SparkContext: Created broadcast 3 from broadcast at DAGScheduler.scala:996
18/05/12 10:11:11 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 1 (MapPartitionsRDD[13] at show at asnt20.scala:39)
18/05/12 10:11:11 INFO TaskSchedulerImpl: Adding task set 1.0 with 1 tasks
18/05/12 10:11:11 INFO TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1, localhost, executor driver, partition 0, PROCESS_LOCAL, 5924 bytes)
18/05/12 10:11:11 INFO Executor: Running task 0.0 in stage 1.0 (TID 1)
18/05/12 10:11:12 INFO HadoopRDD: Input split: file:/E:/datasets/Transport.txt:0+42
18/05/12 10:11:12 INFO ContextCleaner: Cleaned accumulator 0
18/05/12 10:11:12 INFO ContextCleaner: Cleaned accumulator 1
18/05/12 10:11:12 INFO Executor: Finished task 0.0 in stage 1.0 (TID 1). 1499 bytes result sent to driver
18/05/12 10:11:12 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 317 ms on localhost (executor driver) (1/1)
18/05/12 10:11:12 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
18/05/12 10:11:12 INFO DAGScheduler: ResultStage 1 (show at asnt20.scala:39) finished in 0.319 s
18/05/12 10:11:12 INFO DAGScheduler: Job 1 finished: show at asnt20.scala:39, took 0.346637 s
18/05/12 10:11:12 INFO CodeGenerator: Code generated in 23.219237 ms
+----------+------------------+
|Trans_mode|Transportation_Exp|
+----------+------------------+
|  airplane|               170|
|       car|               140|
|     train|               120|
|      ship|               200|
+----------+------------------+

18/05/12 10:11:12 INFO BlockManagerInfo: Removed broadcast_1_piece0 on 192.168.56.1:54232 in memory (size: 5.2 KB, free: 902.7 MB)
18/05/12 10:11:12 INFO MemoryStore: Block broadcast_4 stored as values in memory (estimated size 127.2 KB, free 902.3 MB)
```

Contents of transport.txt


NewProj1 [C:\Users\Laptop\IdeaProjects\NewProj1] - ...\src\main\scala\asnt20.scala [newproj1] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
NewProj1 src main scala asnt20.scala
asnt20
Run: asnt20

```
18/05/12 10:11:12 INFO TaskSchedulerImpl: Adding task set 2.0 with 1 tasks
18/05/12 10:11:12 INFO TaskSetManager: Starting task 0.0 in stage 2.0 (TID 2, localhost, executor driver, partition 0, PROCESS_LOCAL, 5926 bytes)
18/05/12 10:11:12 INFO Executor: Running task 0.0 in stage 2.0 (TID 2)
18/05/12 10:11:12 INFO HadoopRDD: Input split: file:/E:/datasets/Userdetails.txt:0+116
18/05/12 10:11:12 INFO Executor: Finished task 0.0 in stage 2.0 (TID 2). 1589 bytes result sent to driver
18/05/12 10:11:12 INFO TaskSetManager: Finished task 0.0 in stage 2.0 (TID 2) in 27 ms on localhost (executor driver) (1/1)
18/05/12 10:11:12 INFO TaskSchedulerImpl: Removed TaskSet 2.0, whose tasks have all completed, from pool
18/05/12 10:11:12 INFO DAGScheduler: ResultStage 2 (show at asnt20.scala:46) finished in 0.028 s
18/05/12 10:11:12 INFO DAGScheduler: Job 2 finished: show at asnt20.scala:46, took 0.058412 s
18/05/12 10:11:12 INFO CodeGenerator: Code generated in 17.49142 ms
+--------+---------+--------+
|Users_ID|User_name|User_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|
+--------+---------+--------+

18/05/12 10:11:12 INFO SparkSqlParser: Parsing command: Holiday_data
18/05/12 10:11:13 INFO SparkSqlParser: Parsing command:  select year,count("year") from Holiday_data Group by year
18/05/12 10:11:13 INFO CodeGenerator: Code generated in 54.24509 ms
18/05/12 10:11:13 INFO CodeGenerator: Code generated in 99.082548 ms
18/05/12 10:11:13 INFO SparkContext: Starting job: show at asnt20.scala:50
18/05/12 10:11:13 INFO DAGScheduler: Registering RDD 24 (show at asnt20.scala:50)
18/05/12 10:11:13 INFO DAGScheduler: Got job 3 (show at asnt20.scala:50) with 1 output partitions
18/05/12 10:11:13 INFO DAGScheduler: Final stage: ResultStage 4 (show at asnt20.scala:50)
18/05/12 10:11:13 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 3)
```

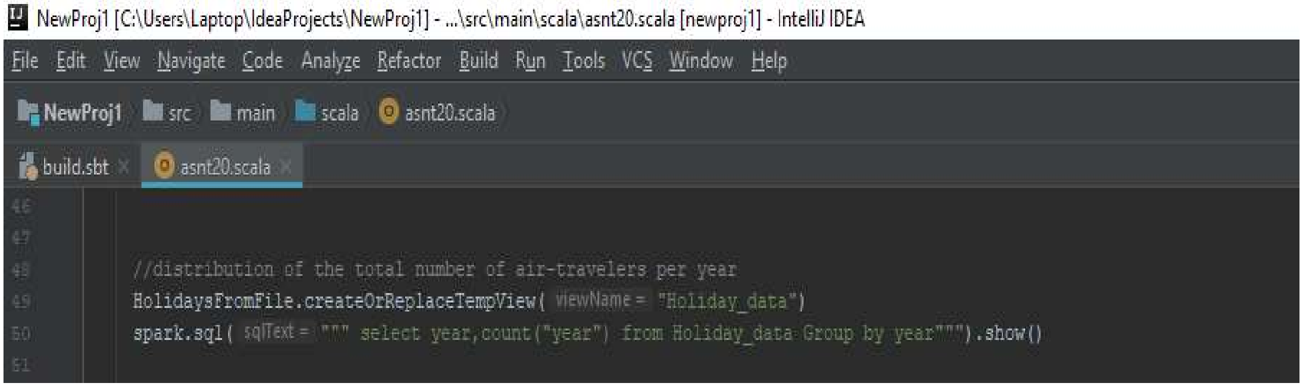
Contents of User Details.txt

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

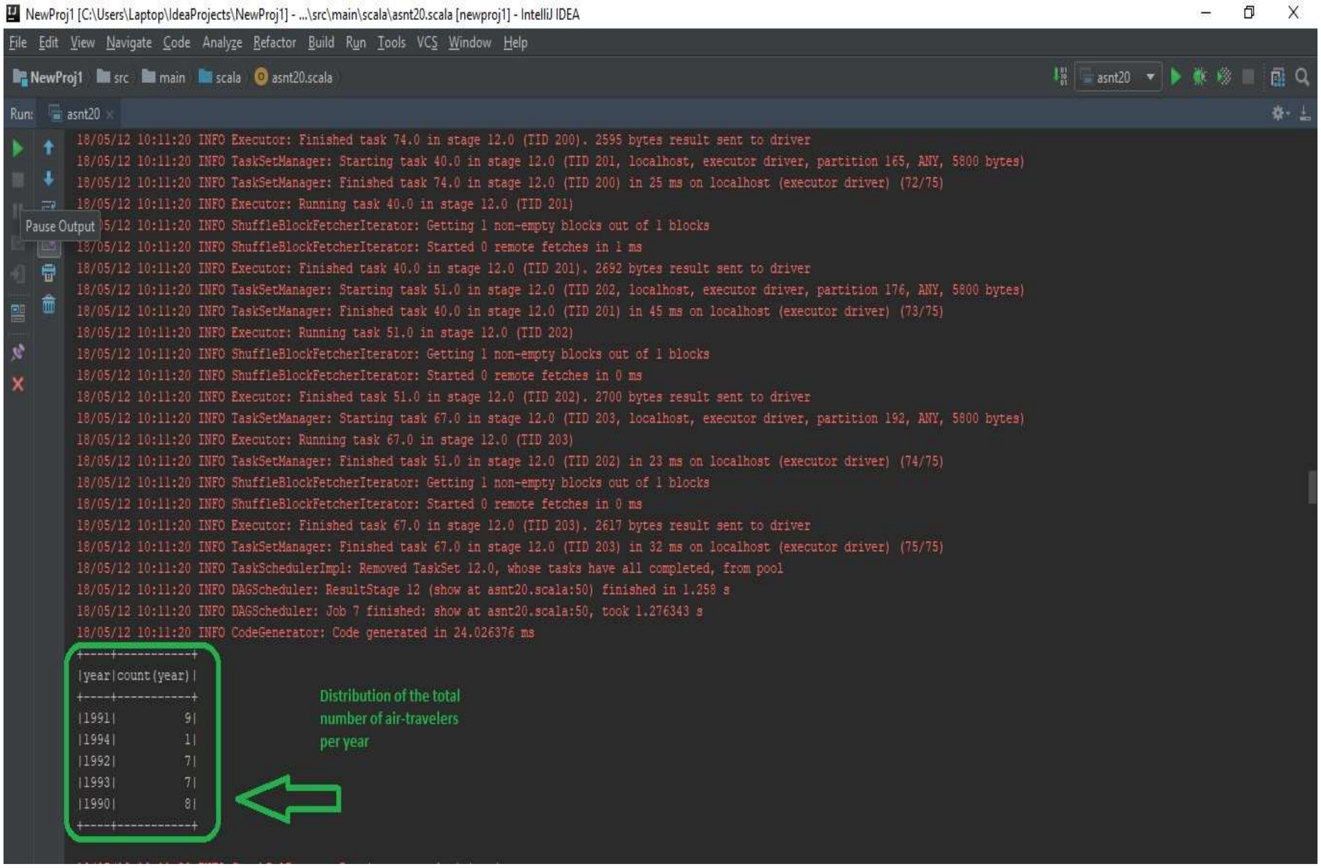
Below code will generate the solution for this task:

- ➢ HolidaysFromFile.createOrReplaceTempView("Holiday_data")
  //create a view on the Data Frames called "Holiday_data"
- ➢ spark.sql(""" select year,count("year") from Holiday_data Group by year""").show()

  Write a query to select year and calculate count of year from the view created


NewProj1 [C:\Users\Laptop\IdeaProjects\NewProj1] - ...\src\main\scala\asnt20.scala [newproj1] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
NewProj1 src main scala asnt20.scala
build.sbt asnt20.scala

```
//distribution of the total number of air-travelers per year
HolidaysFromFile.createOrReplaceTempView( viewName = "Holiday_data")
spark.sql( sqlText = """ select year,count("year") from Holiday_data Group by year""").show()
```


NewProj1 [C:\Users\Laptop\IdeaProjects\NewProj1] - ...\src\main\scala\asnt20.scala [newproj1] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
NewProj1 src main scala asnt20.scala
asnt20
Run: asnt20
Pause Output

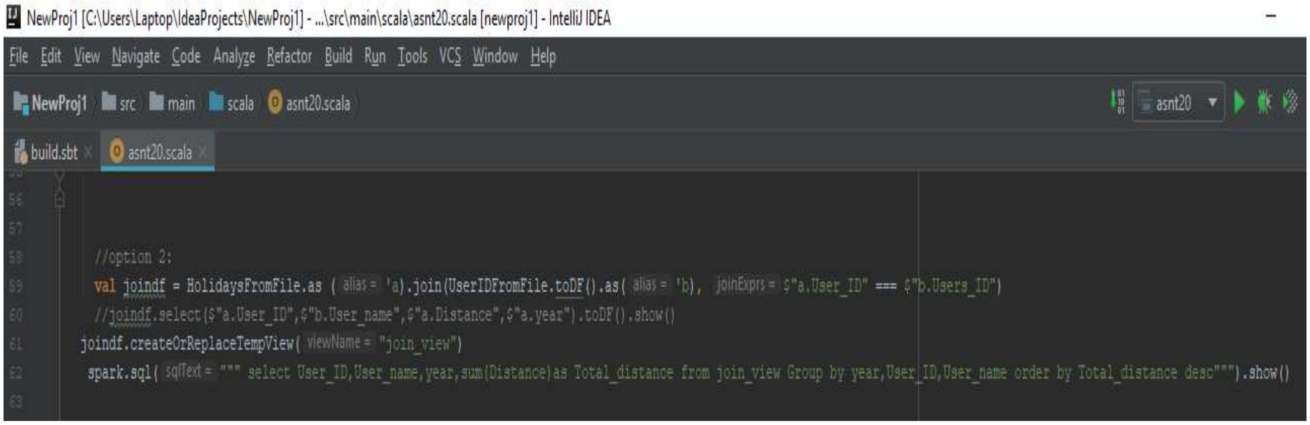
```
18/05/12 10:11:20 INFO Executor: Finished task 74.0 in stage 12.0 (TID 200). 2595 bytes result sent to driver
18/05/12 10:11:20 INFO TaskSetManager: Starting task 40.0 in stage 12.0 (TID 201, localhost, executor driver, partition 165, ANY, 5800 bytes)
18/05/12 10:11:20 INFO TaskSetManager: Finished task 74.0 in stage 12.0 (TID 200) in 25 ms on localhost (executor driver) (72/75)
18/05/12 10:11:20 INFO Executor: Running task 40.0 in stage 12.0 (TID 201)
18/05/12 10:11:20 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/05/12 10:11:20 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
18/05/12 10:11:20 INFO Executor: Finished task 40.0 in stage 12.0 (TID 201). 2692 bytes result sent to driver
18/05/12 10:11:20 INFO TaskSetManager: Starting task 51.0 in stage 12.0 (TID 202, localhost, executor driver, partition 176, ANY, 5800 bytes)
18/05/12 10:11:20 INFO TaskSetManager: Finished task 40.0 in stage 12.0 (TID 201) in 45 ms on localhost (executor driver) (73/75)
18/05/12 10:11:20 INFO Executor: Running task 51.0 in stage 12.0 (TID 202)
18/05/12 10:11:20 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/05/12 10:11:20 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/05/12 10:11:20 INFO Executor: Finished task 51.0 in stage 12.0 (TID 202). 2700 bytes result sent to driver
18/05/12 10:11:20 INFO TaskSetManager: Starting task 67.0 in stage 12.0 (TID 203, localhost, executor driver, partition 192, ANY, 5800 bytes)
18/05/12 10:11:20 INFO Executor: Running task 67.0 in stage 12.0 (TID 203)
18/05/12 10:11:20 INFO TaskSetManager: Finished task 51.0 in stage 12.0 (TID 202) in 23 ms on localhost (executor driver) (74/75)
18/05/12 10:11:20 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/05/12 10:11:20 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/05/12 10:11:20 INFO Executor: Finished task 67.0 in stage 12.0 (TID 203). 2617 bytes result sent to driver
18/05/12 10:11:20 INFO TaskSetManager: Finished task 67.0 in stage 12.0 (TID 203) in 32 ms on localhost (executor driver) (75/75)
18/05/12 10:11:20 INFO TaskSchedulerImpl: Removed TaskSet 12.0, whose tasks have all completed, from pool
18/05/12 10:11:20 INFO DAGScheduler: ResultStage 12 (show at asnt20.scala:50) finished in 1.258 s
18/05/12 10:11:20 INFO DAGScheduler: Job 7 finished: show at asnt20.scala:50, took 1.276343 s
18/05/12 10:11:20 INFO CodeGenerator: Code generated in 24.026376 ms
+----+-----------+
|year|count(year)|
+----+-----------+
|1991|          9|
|1994|          1|
|1992|          7|
|1993|          7|
|1990|          8|
+----+-----------+
```

Distribution of the total number of air-travelers per year

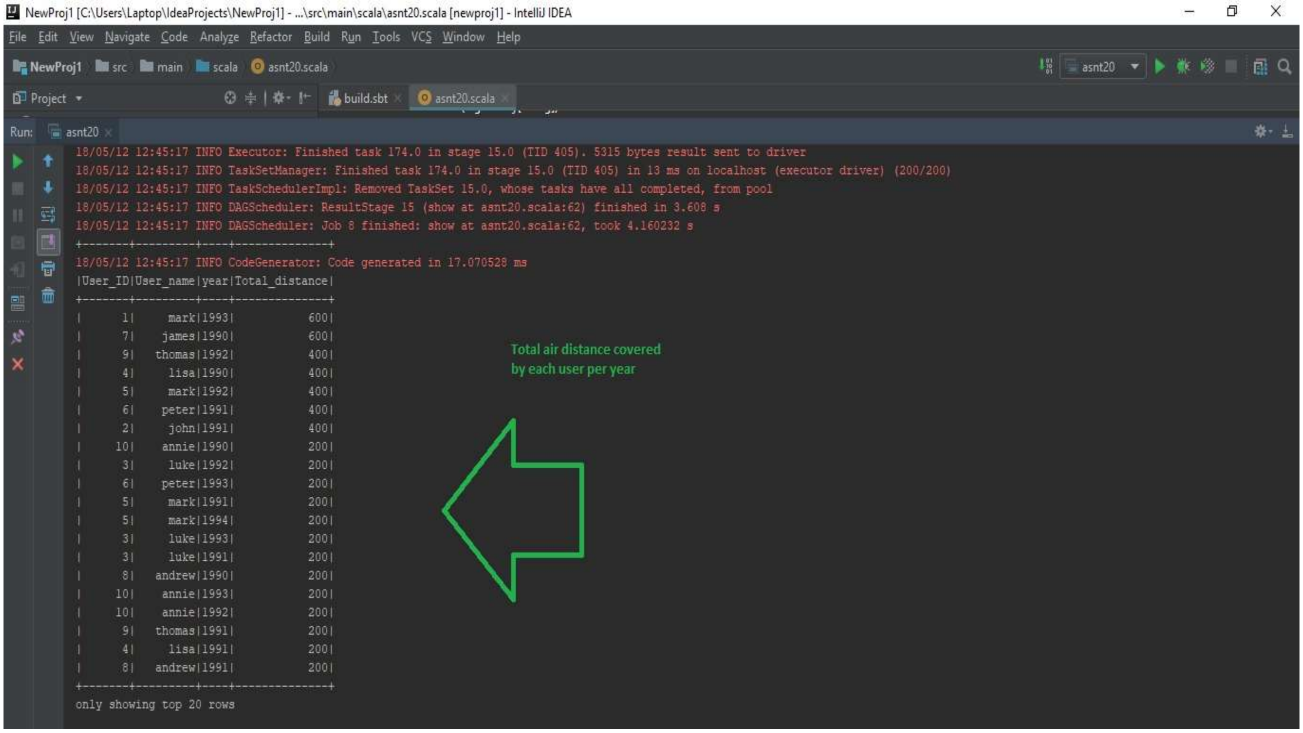
2) What is the total air distance covered by each user per year

Below code will generate the solution for this task:

- val joindf = HolidaysFromFile.as ('a).join(UserIDFromFile.toDF().as('b), $"a.User_ID" === $"b.Users_ID")//join two data frames based on User_ID
- joindf.createOrReplaceTempView("join_view")//create a view on the joindf variable
- spark.sql(""" select User_ID,User_name,year,sum(Distance)as Total_distance from join_view Group by year,User_ID,User_name order by Total_distance desc""").show()   //Write a SQL query to find the total air distance covered by each user per year


NewProj1 [C:\Users\Laptop\IdeaProjects\NewProj1] - ...\src\main\scala\asnt20.scala [newproj1] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
NewProj1 src main scala asnt20.scala
asnt20
build.sbt
asnt20.scala

```
    //option 2:
    val joindf = HolidaysFromFile.as ( alias = 'a).join(UserIDFromFile.toDF().as( alias = 'b),  joinExprs = $"a.User_ID" === $"b.Users_ID")
    //joindf.select($"a.User_ID",$"b.User_name",$"a.Distance",$"a.year").toDF().show()
  joindf.createOrReplaceTempView( viewName = "join_view")
   spark.sql( sqlText = """ select User_ID,User_name,year,sum(Distance)as Total_distance from join_view Group by year,User_ID,User_name order by Total_distance desc""").show()
```


NewProj1 [C:\Users\Laptop\IdeaProjects\NewProj1] - ...\src\main\scala\asnt20.scala [newproj1] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
NewProj1 src main scala asnt20.scala
asnt20
Project
build.sbt
asnt20.scala
Run:
asnt20

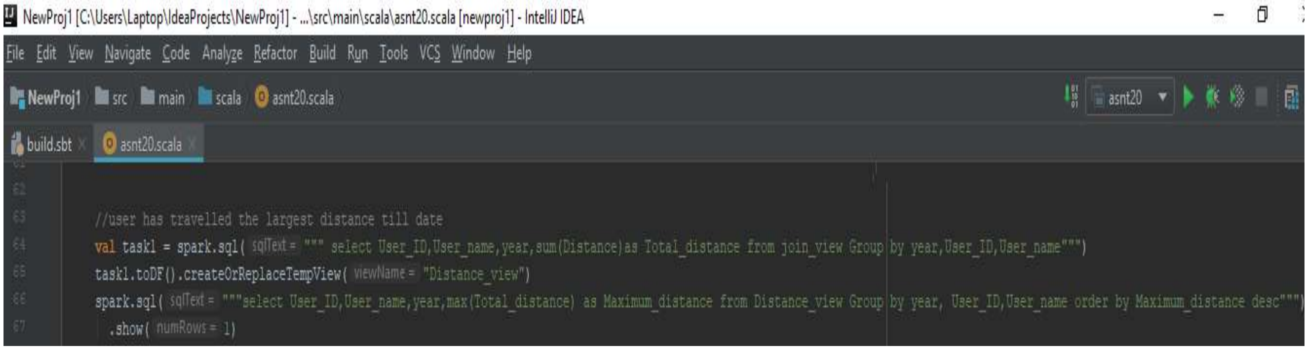
```
18/05/12 12:45:17 INFO Executor: Finished task 174.0 in stage 15.0 (TID 405). 5315 bytes result sent to driver
18/05/12 12:45:17 INFO TaskSetManager: Finished task 174.0 in stage 15.0 (TID 405) in 13 ms on localhost (executor driver) (200/200)
18/05/12 12:45:17 INFO TaskSchedulerImpl: Removed TaskSet 15.0, whose tasks have all completed, from pool
18/05/12 12:45:17 INFO DAGScheduler: ResultStage 15 (show at asnt20.scala:62) finished in 3.608 s
18/05/12 12:45:17 INFO DAGScheduler: Job 8 finished: show at asnt20.scala:62, took 4.160232 s
+-------+---------+----+--------------+
18/05/12 12:45:17 INFO CodeGenerator: Code generated in 17.070528 ms
|User_ID|User_name|year|Total_distance|
+-------+---------+----+--------------+
|      1|     mark|1993|           600|
|      7|    james|1990|           600|
|      9|   thomas|1992|           400|
|      4|     lisa|1990|           400|
|      5|     mark|1992|           400|
|      6|    peter|1991|           400|
|      2|     john|1991|           400|
|     10|    annie|1990|           200|
|      3|     luke|1992|           200|
|      6|    peter|1993|           200|
|      5|     mark|1991|           200|
|      5|     mark|1994|           200|
|      3|     luke|1993|           200|
|      3|     luke|1991|           200|
|      8|   andrew|1990|           200|
|     10|    annie|1993|           200|
|     10|    annie|1992|           200|
|      9|   thomas|1991|           200|
|      4|     lisa|1991|           200|
|      8|   andrew|1991|           200|
+-------+---------+----+--------------+
only showing top 20 rows
```

Total air distance covered
by each user per year

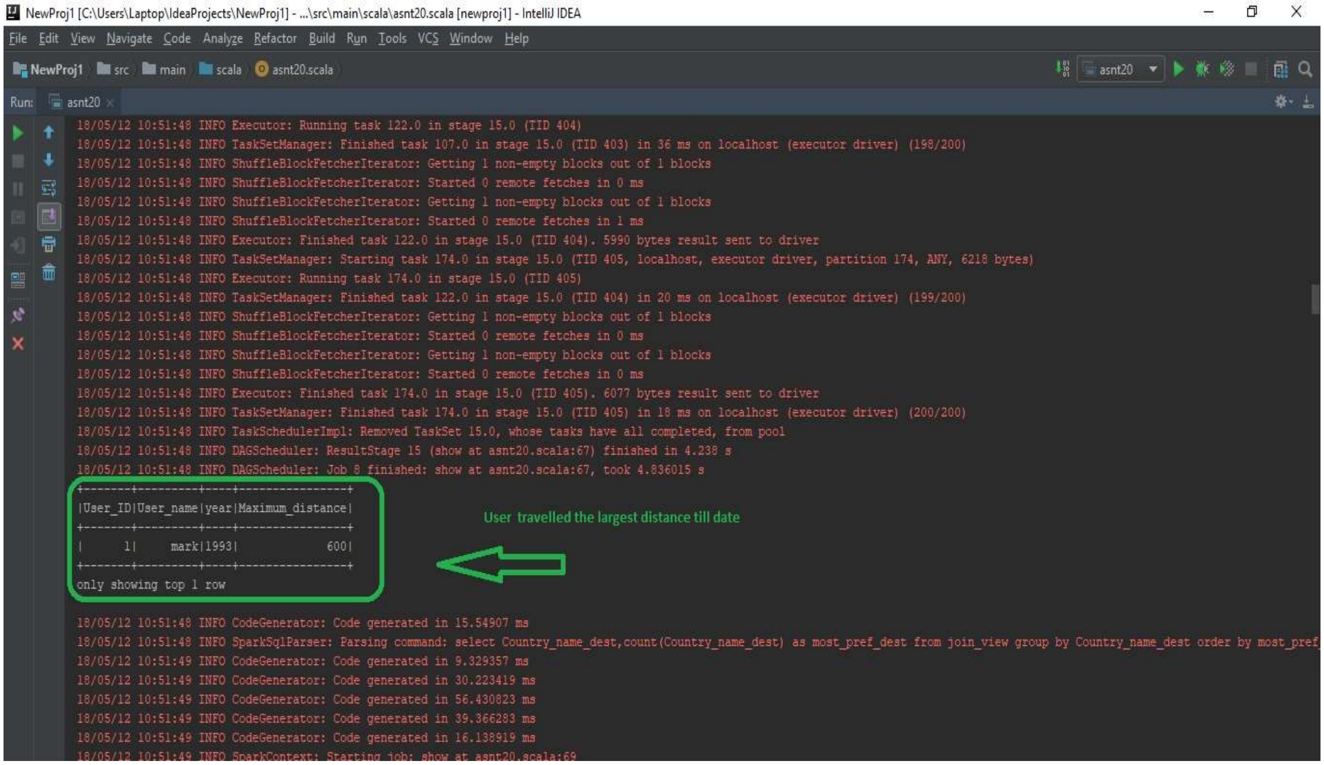
3) Which user has travelled the largest distance till date

Below code will generate the solution for this task:

- val task1 = spark.sql(""" select User_ID,User_name,year,sum(Distance)as Total_distance from join_view Group by year,User_ID,User_name""")//saving the SQL query in a variable "task1"
- task1.toDF().createOrReplaceTempView("Distance_view")// creating view on task1 variable
- spark.sql("""select User_ID,User_name,year,max(Total_distance) as Maximum_distance from Distance_view Group by year, User_ID,User_name order by Maximum_distance desc""")
- .show(1)// creating a SQL query to find which user has travelled the largest distance till date


NewProj1 [C:\Users\Laptop\IdeaProjects\NewProj1] - ...\src\main\scala\asnt20.scala [newproj1] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
NewProj1 src main scala asnt20.scala
asnt20
build.sbt
asnt20.scala

```
//user has travelled the largest distance till date
val task1 = spark.sql( sqlText = """ select User_ID,User_name,year,sum(Distance)as Total_distance from join_view Group by year,User_ID,User_name""")
task1.toDF().createOrReplaceTempView( viewName = "Distance_view")
spark.sql( sqlText = """select User_ID,User_name,year,max(Total_distance) as Maximum_distance from Distance_view Group by year, User_ID,User_name order by Maximum_distance desc""")
  .show( numRows = 1)
```


NewProj1 [C:\Users\Laptop\IdeaProjects\NewProj1] - ...\src\main\scala\asnt20.scala [newproj1] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
NewProj1 src main scala asnt20.scala
asnt20
Run: asnt20

```
18/05/12 10:51:48 INFO Executor: Running task 122.0 in stage 15.0 (TID 404)
18/05/12 10:51:48 INFO TaskSetManager: Finished task 107.0 in stage 15.0 (TID 403) in 36 ms on localhost (executor driver) (198/200)
18/05/12 10:51:48 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/05/12 10:51:48 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/05/12 10:51:48 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/05/12 10:51:48 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
18/05/12 10:51:48 INFO Executor: Finished task 122.0 in stage 15.0 (TID 404). 5990 bytes result sent to driver
18/05/12 10:51:48 INFO TaskSetManager: Starting task 174.0 in stage 15.0 (TID 405, localhost, executor driver, partition 174, ANY, 6218 bytes)
18/05/12 10:51:48 INFO Executor: Running task 174.0 in stage 15.0 (TID 405)
18/05/12 10:51:48 INFO TaskSetManager: Finished task 122.0 in stage 15.0 (TID 404) in 20 ms on localhost (executor driver) (199/200)
18/05/12 10:51:48 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/05/12 10:51:48 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/05/12 10:51:48 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/05/12 10:51:48 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/05/12 10:51:48 INFO Executor: Finished task 174.0 in stage 15.0 (TID 405). 6077 bytes result sent to driver
18/05/12 10:51:48 INFO TaskSetManager: Finished task 174.0 in stage 15.0 (TID 405) in 18 ms on localhost (executor driver) (200/200)
18/05/12 10:51:48 INFO TaskSchedulerImpl: Removed TaskSet 15.0, whose tasks have all completed, from pool
18/05/12 10:51:48 INFO DAGScheduler: ResultStage 15 (show at asnt20.scala:67) finished in 4.238 s
18/05/12 10:51:48 INFO DAGScheduler: Job 8 finished: show at asnt20.scala:67, took 4.836015 s
+-------+---------+----+----------------+
|User_ID|User_name|year|Maximum_distance|
+-------+---------+----+----------------+
|      1|     mark|1993|             600|
+-------+---------+----+----------------+
only showing top 1 row

18/05/12 10:51:48 INFO CodeGenerator: Code generated in 15.54907 ms
18/05/12 10:51:48 INFO SparkSqlParser: Parsing command: select Country_name_dest,count(Country_name_dest) as most_pref_dest from join_view group by Country_name_dest order by most_pref
18/05/12 10:51:49 INFO CodeGenerator: Code generated in 9.329357 ms
18/05/12 10:51:49 INFO CodeGenerator: Code generated in 30.223419 ms
18/05/12 10:51:49 INFO CodeGenerator: Code generated in 56.430823 ms
18/05/12 10:51:49 INFO CodeGenerator: Code generated in 39.366283 ms
18/05/12 10:51:49 INFO CodeGenerator: Code generated in 16.138919 ms
18/05/12 10:51:49 INFO SparkContext: Starting job: show at asnt20.scala:69
```

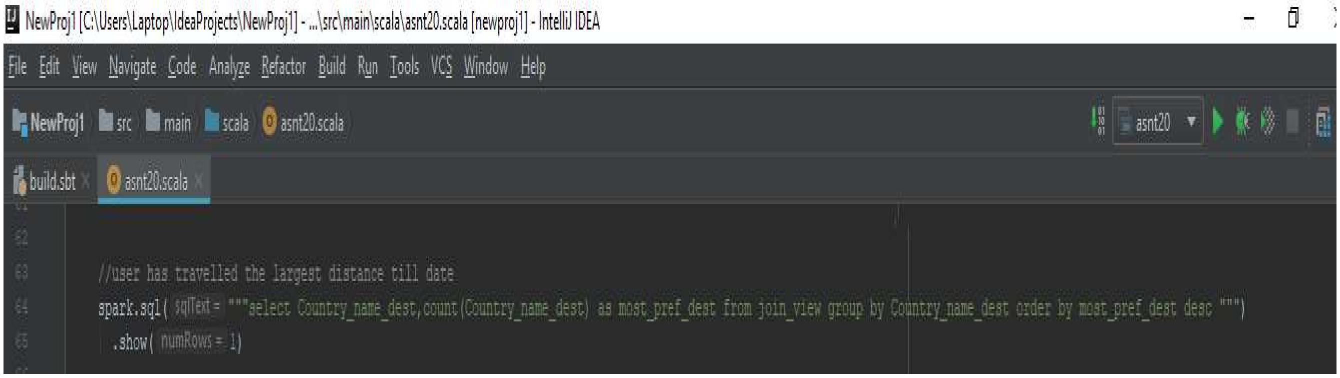
User travelled the largest distance till date

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

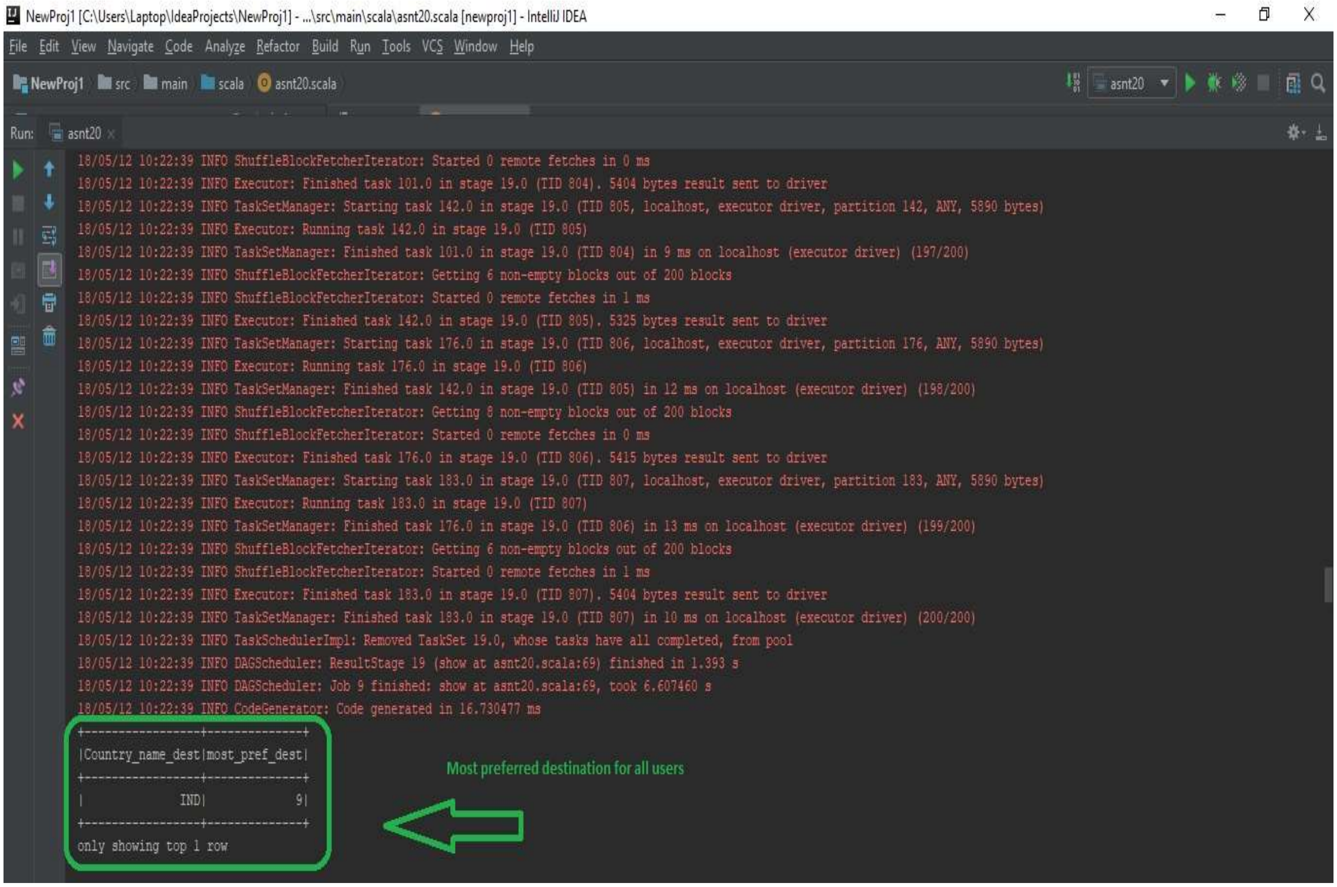
Below code will generate the solution for this task:

spark.sql("""select Country_name_dest,count(Country_name_dest) as most_pref_dest from join_view group by Country_name_dest order by most_pref_dest desc """) .show(1)

// write an sql query to find most preferred destination for all users from the view previously created //called "join_view"


NewProj1 [C:\Users\Laptop\IdeaProjects\NewProj1] - ...\src\main\scala\asnt20.scala [newproj1] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
NewProj1 src main scala asnt20.scala
asnt20
build.sbt
asnt20.scala

```
//user has travelled the largest distance till date
spark.sql( sqlText = """select Country_name_dest,count(Country_name_dest) as most_pref_dest from join_view group by Country_name_dest order by most_pref_dest desc """)
 .show( numRows = 1)
```


NewProj1 [C:\Users\Laptop\IdeaProjects\NewProj1] - ...\src\main\scala\asnt20.scala [newproj1] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
NewProj1 src main scala asnt20.scala
asnt20
Run: asnt20

```
18/05/12 10:22:39 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/05/12 10:22:39 INFO Executor: Finished task 101.0 in stage 19.0 (TID 804). 5404 bytes result sent to driver
18/05/12 10:22:39 INFO TaskSetManager: Starting task 142.0 in stage 19.0 (TID 805, localhost, executor driver, partition 142, ANY, 5890 bytes)
18/05/12 10:22:39 INFO Executor: Running task 142.0 in stage 19.0 (TID 805)
18/05/12 10:22:39 INFO TaskSetManager: Finished task 101.0 in stage 19.0 (TID 804) in 9 ms on localhost (executor driver) (197/200)
18/05/12 10:22:39 INFO ShuffleBlockFetcherIterator: Getting 6 non-empty blocks out of 200 blocks
18/05/12 10:22:39 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
18/05/12 10:22:39 INFO Executor: Finished task 142.0 in stage 19.0 (TID 805). 5325 bytes result sent to driver
18/05/12 10:22:39 INFO TaskSetManager: Starting task 176.0 in stage 19.0 (TID 806, localhost, executor driver, partition 176, ANY, 5890 bytes)
18/05/12 10:22:39 INFO Executor: Running task 176.0 in stage 19.0 (TID 806)
18/05/12 10:22:39 INFO TaskSetManager: Finished task 142.0 in stage 19.0 (TID 805) in 12 ms on localhost (executor driver) (198/200)
18/05/12 10:22:39 INFO ShuffleBlockFetcherIterator: Getting 8 non-empty blocks out of 200 blocks
18/05/12 10:22:39 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/05/12 10:22:39 INFO Executor: Finished task 176.0 in stage 19.0 (TID 806). 5415 bytes result sent to driver
18/05/12 10:22:39 INFO TaskSetManager: Starting task 183.0 in stage 19.0 (TID 807, localhost, executor driver, partition 183, ANY, 5890 bytes)
18/05/12 10:22:39 INFO Executor: Running task 183.0 in stage 19.0 (TID 807)
18/05/12 10:22:39 INFO TaskSetManager: Finished task 176.0 in stage 19.0 (TID 806) in 13 ms on localhost (executor driver) (199/200)
18/05/12 10:22:39 INFO ShuffleBlockFetcherIterator: Getting 6 non-empty blocks out of 200 blocks
18/05/12 10:22:39 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
18/05/12 10:22:39 INFO Executor: Finished task 183.0 in stage 19.0 (TID 807). 5404 bytes result sent to driver
18/05/12 10:22:39 INFO TaskSetManager: Finished task 183.0 in stage 19.0 (TID 807) in 10 ms on localhost (executor driver) (200/200)
18/05/12 10:22:39 INFO TaskSchedulerImpl: Removed TaskSet 19.0, whose tasks have all completed, from pool
18/05/12 10:22:39 INFO DAGScheduler: ResultStage 19 (show at asnt20.scala:69) finished in 1.393 s
18/05/12 10:22:39 INFO DAGScheduler: Job 9 finished: show at asnt20.scala:69, took 6.607460 s
18/05/12 10:22:39 INFO CodeGenerator: Code generated in 16.730477 ms
+-----------------+--------------+
|Country_name_dest|most_pref_dest|
+-----------------+--------------+
|              IND|             9|
+-----------------+--------------+
only showing top 1 row
```

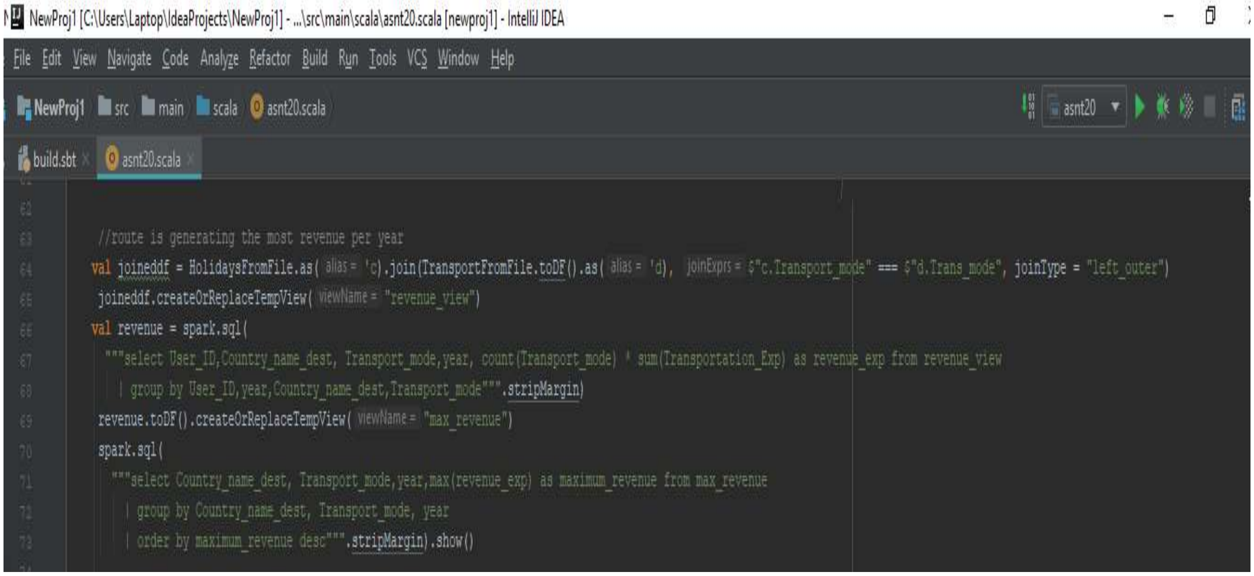
Most preferred destination for all users

5)Which route is generating the most revenue per year

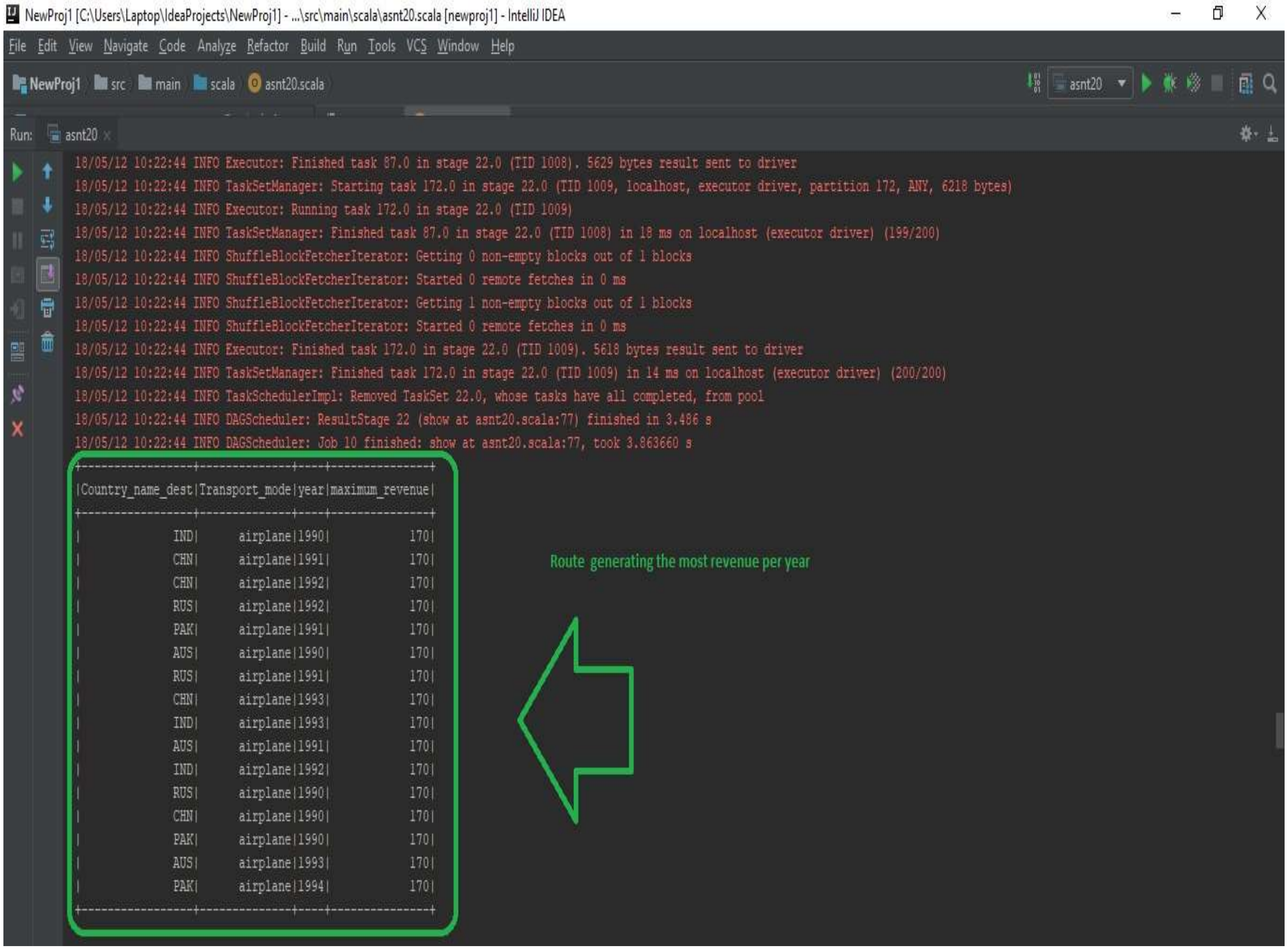
Below code will generate the solution for this task:

- val joineddf = HolidaysFromFile.as('c).join(TransportFromFile.toDF().as('d), $"c.Transport_mode" === $"d.Trans_mode", joinType = "left_outer")
  // to join two Dataframes based on Transportation mode
- joineddf.createOrReplaceTempView("revenue_view")
  //creating a view on the join previously created
- val revenue = spark.sql(// creating an sql query to calculate the revenue
- """select User_ID,Country_name_dest, Transport_mode,year, count(Transport_mode) * sum(Transportation_Exp) as revenue_exp from revenue_view
- | group by User_ID,year,Country_name_dest,Transport_mode""".stripMargin)
- revenue.toDF().createOrReplaceTempView("max_revenue")// converting the previously evaluated variable to DataFrame and create a view on the same called"max_revenue"
- spark.sql(
- """select Country_name_dest, Transport_mode,year,max(revenue_exp) as maximum_revenue from max_revenue
- | group by Country_name_dest, Transport_mode, year
- | order by maximum_revenue desc""".stripMargin).show()

// write an sql query to find which route is generating the most revenue per year


NewProj1 [C:\Users\Laptop\IdeaProjects\NewProj1] - ...\src\main\scala\asnt20.scala [newproj1] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
NewProj1 src main scala asnt20.scala
asnt20
build.sbt
asnt20.scala

```
//route is generating the most revenue per year
val joineddf = HolidaysFromFile.as( alias = 'c).join(TransportFromFile.toDF().as( alias = 'd),  joinExprs = $"c.Transport_mode" === $"d.Trans_mode", joinType = "left_outer")
 joineddf.createOrReplaceTempView( viewName = "revenue_view")
val revenue = spark.sql(
  """select User_ID,Country_name_dest, Transport_mode,year, count(Transport_mode) * sum(Transportation_Exp) as revenue_exp from revenue_view
   | group by User_ID,year,Country_name_dest,Transport_mode""".stripMargin)
 revenue.toDF().createOrReplaceTempView( viewName = "max_revenue")
 spark.sql(
   """select Country_name_dest, Transport_mode,year,max(revenue_exp) as maximum_revenue from max_revenue
    | group by Country_name_dest, Transport_mode, year
    | order by maximum_revenue desc""".stripMargin).show()
```


NewProj1 [C:\Users\Laptop\IdeaProjects\NewProj1] - ...\src\main\scala\asnt20.scala [newproj1] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
NewProj1 src main scala asnt20.scala
asnt20
Run: asnt20

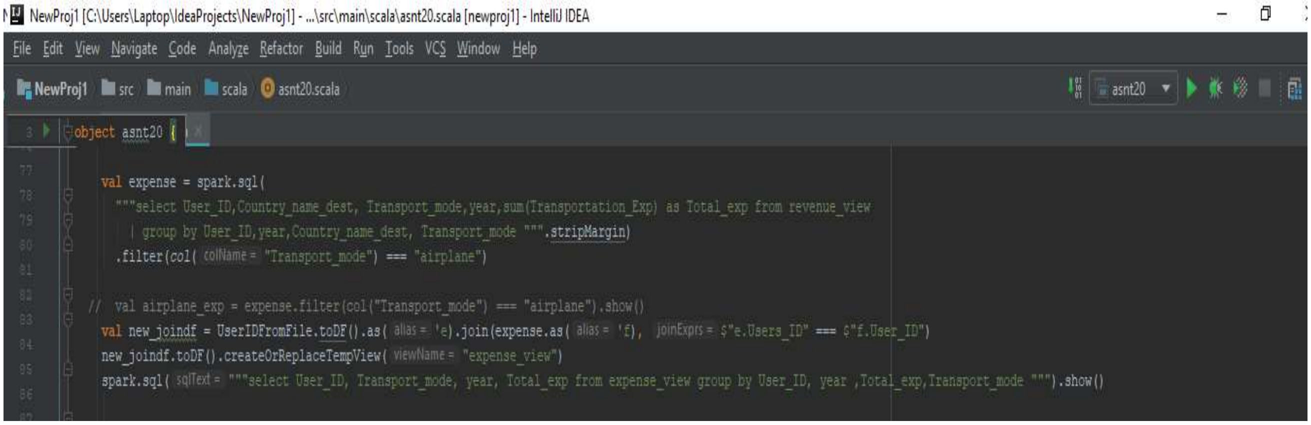
```
18/05/12 10:22:44 INFO Executor: Finished task 87.0 in stage 22.0 (TID 1008). 5629 bytes result sent to driver
18/05/12 10:22:44 INFO TaskSetManager: Starting task 172.0 in stage 22.0 (TID 1009, localhost, executor driver, partition 172, ANY, 6218 bytes)
18/05/12 10:22:44 INFO Executor: Running task 172.0 in stage 22.0 (TID 1009)
18/05/12 10:22:44 INFO TaskSetManager: Finished task 87.0 in stage 22.0 (TID 1008) in 18 ms on localhost (executor driver) (199/200)
18/05/12 10:22:44 INFO ShuffleBlockFetcherIterator: Getting 0 non-empty blocks out of 1 blocks
18/05/12 10:22:44 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/05/12 10:22:44 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/05/12 10:22:44 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/05/12 10:22:44 INFO Executor: Finished task 172.0 in stage 22.0 (TID 1009). 5618 bytes result sent to driver
18/05/12 10:22:44 INFO TaskSetManager: Finished task 172.0 in stage 22.0 (TID 1009) in 14 ms on localhost (executor driver) (200/200)
18/05/12 10:22:44 INFO TaskSchedulerImpl: Removed TaskSet 22.0, whose tasks have all completed, from pool
18/05/12 10:22:44 INFO DAGScheduler: ResultStage 22 (show at asnt20.scala:77) finished in 3.486 s
18/05/12 10:22:44 INFO DAGScheduler: Job 10 finished: show at asnt20.scala:77, took 3.863660 s
+----------------+--------------+----+---------------+
|Country_name_dest|Transport_mode|year|maximum_revenue|
+----------------+--------------+----+---------------+
|             IND|      airplane|1990|            170|
|             CHN|      airplane|1991|            170|
|             CHN|      airplane|1992|            170|
|             RUS|      airplane|1992|            170|
|             PAK|      airplane|1991|            170|
|             AUS|      airplane|1990|            170|
|             RUS|      airplane|1991|            170|
|             CHN|      airplane|1993|            170|
|             IND|      airplane|1993|            170|
|             AUS|      airplane|1991|            170|
|             IND|      airplane|1992|            170|
|             RUS|      airplane|1990|            170|
|             CHN|      airplane|1990|            170|
|             PAK|      airplane|1990|            170|
|             AUS|      airplane|1993|            170|
|             PAK|      airplane|1994|            170|
+----------------+--------------+----+---------------+
```

Route generating the most revenue per year

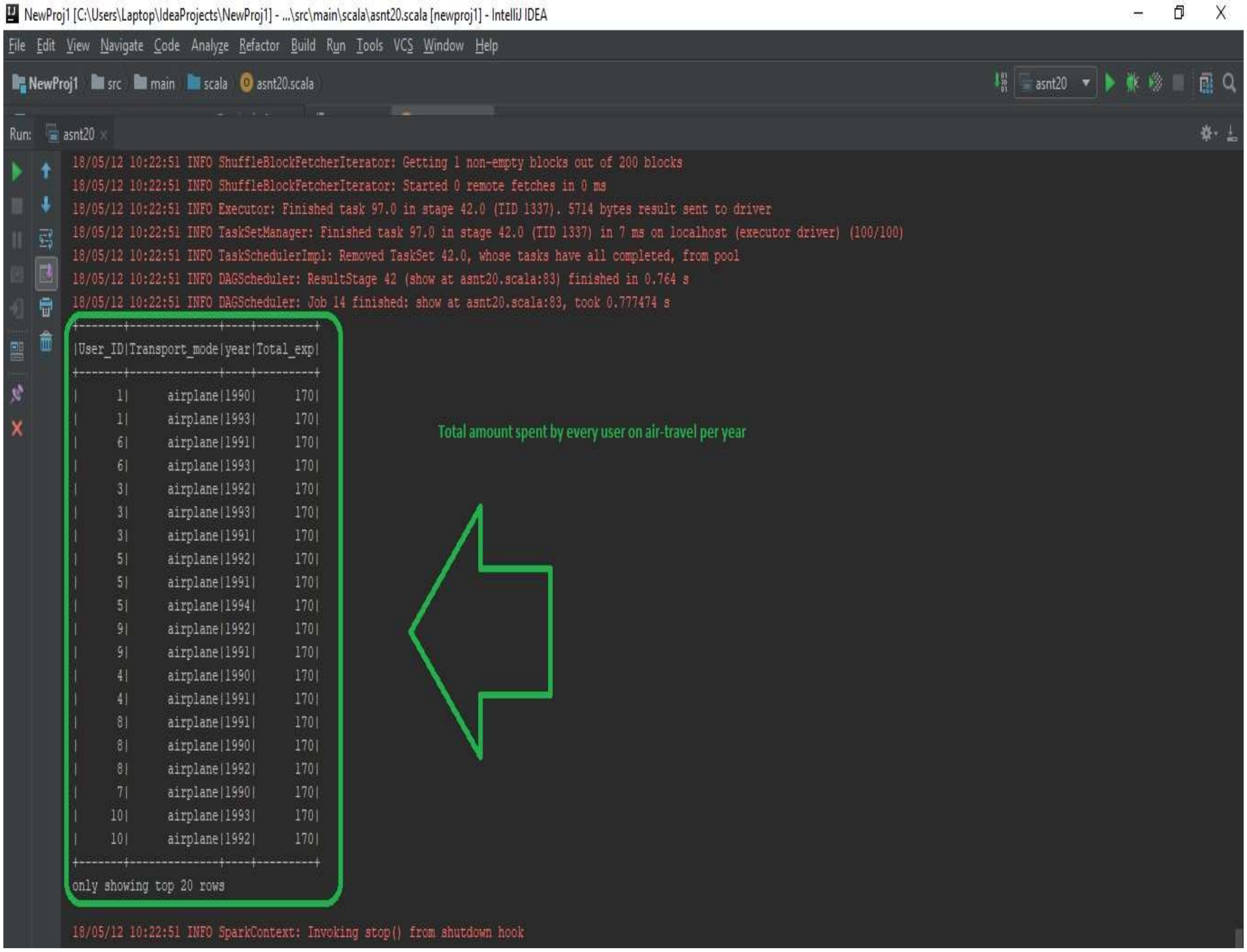
5) What is the total amount spent by every user on air-travel per year

Below code will generate the solution for this task:

- val expense = spark.sql(// create a sql query to calculate the transportation expenses from //previously created revenue_view
- """select User_ID,Country_name_dest, Transport_mode,year,sum(Transportation_Exp) as Total_exp from revenue_view
- | group by User_ID,year,Country_name_dest, Transport_mode """.stripMargin)
- .filter(col("Transport_mode") === "airplane")// filter transport mode as airplane
- val new_joindf = UserIDFromFile.toDF().as('e).join(expense.as('f), $"e.Users_ID" === $"f.User_ID")// join two data frames based on UserID
- new_joindf.toDF().createOrReplaceTempView("expense_view") //create a view on variable created
- spark.sql("""select User_ID, Transport_mode, year, Total_exp from expense_view group by User_ID, year ,Total_exp,Transport_mode """).show() //an sql query to find total expenses spent by user for air-travel


NewProj1 [C:\Users\Laptop\IdeaProjects\NewProj1] - ...\src\main\scala\asnt20.scala [newproj1] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
NewProj1 src main scala asnt20.scala
asnt20

```scala
object asnt20 {

    val expense = spark.sql(
      """select User_ID,Country_name_dest, Transport_mode,year,sum(Transportation_Exp) as Total_exp from revenue_view
        | group by User_ID,year,Country_name_dest, Transport_mode """.stripMargin)
      .filter(col("Transport_mode") === "airplane")

  //  val airplane_exp = expense.filter(col("Transport_mode") === "airplane").show()
    val new_joindf = UserIDFromFile.toDF().as('e).join(expense.as('f), $"e.Users_ID" === $"f.User_ID")
    new_joindf.toDF().createOrReplaceTempView("expense_view")
    spark.sql("""select User_ID, Transport_mode, year, Total_exp from expense_view group by User_ID, year ,Total_exp,Transport_mode """).show()
```


NewProj1 [C:\Users\Laptop\IdeaProjects\NewProj1] - ...\src\main\scala\asnt20.scala [newproj1] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
NewProj1 src main scala asnt20.scala
asnt20
Run: asnt20

```
18/05/12 10:22:51 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 200 blocks
18/05/12 10:22:51 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/05/12 10:22:51 INFO Executor: Finished task 97.0 in stage 42.0 (TID 1337). 5714 bytes result sent to driver
18/05/12 10:22:51 INFO TaskSetManager: Finished task 97.0 in stage 42.0 (TID 1337) in 7 ms on localhost (executor driver) (100/100)
18/05/12 10:22:51 INFO TaskSchedulerImpl: Removed TaskSet 42.0, whose tasks have all completed, from pool
18/05/12 10:22:51 INFO DAGScheduler: ResultStage 42 (show at asnt20.scala:83) finished in 0.764 s
18/05/12 10:22:51 INFO DAGScheduler: Job 14 finished: show at asnt20.scala:83, took 0.777474 s
+-------+--------------+----+---------+
|User_ID|Transport_mode|year|Total_exp|
+-------+--------------+----+---------+
|      1|      airplane|1990|      170|
|      1|      airplane|1993|      170|
|      6|      airplane|1991|      170|
|      6|      airplane|1993|      170|
|      3|      airplane|1992|      170|
|      3|      airplane|1993|      170|
|      3|      airplane|1991|      170|
|      5|      airplane|1992|      170|
|      5|      airplane|1991|      170|
|      5|      airplane|1994|      170|
|      9|      airplane|1992|      170|
|      9|      airplane|1991|      170|
|      4|      airplane|1990|      170|
|      4|      airplane|1991|      170|
|      8|      airplane|1991|      170|
|      8|      airplane|1990|      170|
|      8|      airplane|1992|      170|
|      7|      airplane|1990|      170|
|     10|      airplane|1993|      170|
|     10|      airplane|1992|      170|
+-------+--------------+----+---------+
only showing top 20 rows

18/05/12 10:22:51 INFO SparkContext: Invoking stop() from shutdown hook
```

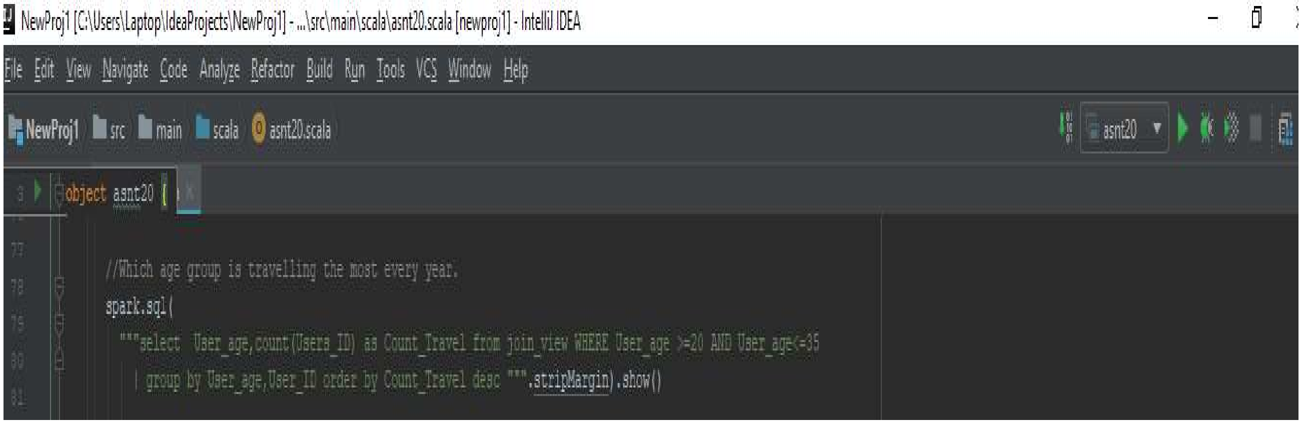
Total amount spent by every user on air-travel per year

7) Considering age groups of < 20 , 20-35, 35 > ,Which age group is travelling the most

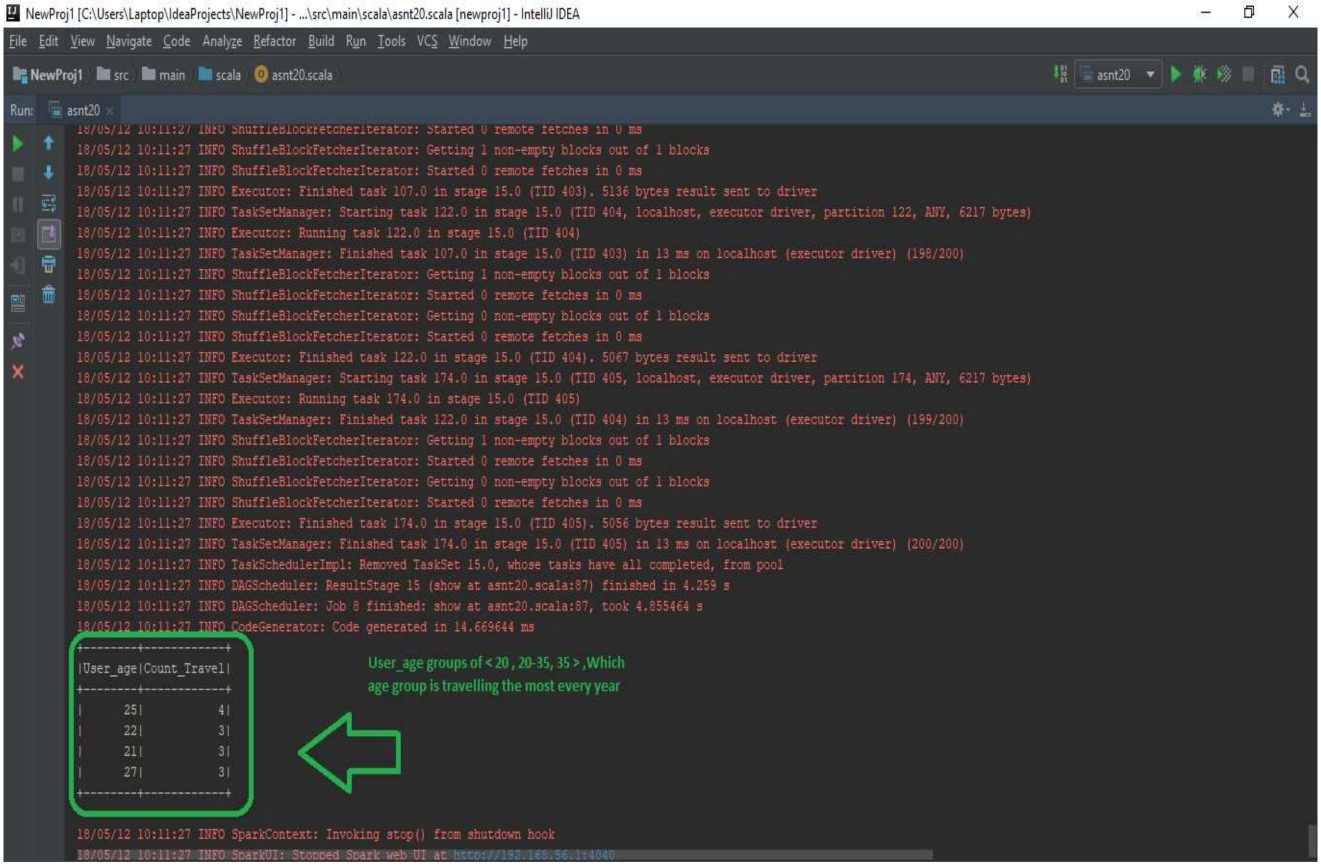
every year.

spark.sql("""select User_age,count(Users_ID) as Count_Travel from join_view WHERE User_age >=20 AND User_age<=35 group by User_age,User_ID order by Count_Travel desc """).show()

// sql query to take count of number of users from previously created join_view and filter the records //based on Users_age between 20 to 35, to find which age group is travelling the most.


NewProj1 [C:\Users\Laptop\IdeaProjects\NewProj1] - ...\src\main\scala\asnt20.scala [newproj1] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
NewProj1 src main scala asnt20.scala
asnt20

```
object asnt20 {

    //Which age group is travelling the most every year.
    spark.sql(
      """select  User_age,count(Users_ID) as Count_Travel from join_view WHERE User_age >=20 AND User_age<=35
        | group by User_age,User_ID order by Count_Travel desc """.stripMargin).show()
```


NewProj1 [C:\Users\Laptop\IdeaProjects\NewProj1] - ...\src\main\scala\asnt20.scala [newproj1] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
NewProj1 src main scala asnt20.scala
asnt20
Run: asnt20

```
18/05/12 10:11:27 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/05/12 10:11:27 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/05/12 10:11:27 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/05/12 10:11:27 INFO Executor: Finished task 107.0 in stage 15.0 (TID 403). 5136 bytes result sent to driver
18/05/12 10:11:27 INFO TaskSetManager: Starting task 122.0 in stage 15.0 (TID 404, localhost, executor driver, partition 122, ANY, 6217 bytes)
18/05/12 10:11:27 INFO Executor: Running task 122.0 in stage 15.0 (TID 404)
18/05/12 10:11:27 INFO TaskSetManager: Finished task 107.0 in stage 15.0 (TID 403) in 13 ms on localhost (executor driver) (198/200)
18/05/12 10:11:27 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/05/12 10:11:27 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/05/12 10:11:27 INFO ShuffleBlockFetcherIterator: Getting 0 non-empty blocks out of 1 blocks
18/05/12 10:11:27 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/05/12 10:11:27 INFO Executor: Finished task 122.0 in stage 15.0 (TID 404). 5067 bytes result sent to driver
18/05/12 10:11:27 INFO TaskSetManager: Starting task 174.0 in stage 15.0 (TID 405, localhost, executor driver, partition 174, ANY, 6217 bytes)
18/05/12 10:11:27 INFO Executor: Running task 174.0 in stage 15.0 (TID 405)
18/05/12 10:11:27 INFO TaskSetManager: Finished task 122.0 in stage 15.0 (TID 404) in 13 ms on localhost (executor driver) (199/200)
18/05/12 10:11:27 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/05/12 10:11:27 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/05/12 10:11:27 INFO ShuffleBlockFetcherIterator: Getting 0 non-empty blocks out of 1 blocks
18/05/12 10:11:27 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/05/12 10:11:27 INFO Executor: Finished task 174.0 in stage 15.0 (TID 405). 5056 bytes result sent to driver
18/05/12 10:11:27 INFO TaskSetManager: Finished task 174.0 in stage 15.0 (TID 405) in 13 ms on localhost (executor driver) (200/200)
18/05/12 10:11:27 INFO TaskSchedulerImpl: Removed TaskSet 15.0, whose tasks have all completed, from pool
18/05/12 10:11:27 INFO DAGScheduler: ResultStage 15 (show at asnt20.scala:87) finished in 4.259 s
18/05/12 10:11:27 INFO DAGScheduler: Job 8 finished: show at asnt20.scala:87, took 4.855464 s
18/05/12 10:11:27 INFO CodeGenerator: Code generated in 14.669644 ms
+--------+------------+
|User_age|Count_Travel|
+--------+------------+
|      25|           4|
|      22|           3|
|      21|           3|
|      27|           3|
+--------+------------+

18/05/12 10:11:27 INFO SparkContext: Invoking stop() from shutdown hook
18/05/12 10:11:27 INFO SparkUI: Stopped Spark web UI at http://192.168.56.1:4040
```

User_age groups of < 20 , 20-35, 35 > ,Which age group is travelling the most every year