

```
RDDEspark [C:\Users\Shrutti\IdeaProjects\RDDEspark] - ...src\main\scala\com\sparksql1\assignment\SparkSQL1.scala (RDDEspark) - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
RDDEspark src main scala com sparksql1 assignment SparkSQL1.scala
Project build.sbt SparkSQL1.scala
Project RDDEspark C:\Users\Shrutti\IdeaProjects\RDDEspark
  project [rdde-spark-build] sources root
  spark-warehouse
  src
  main
  scala
  com
    sparkdd.assignment
    RDD_ASSIGNMENT
    sparksql1.assignment
    SparkSQL1
  test
  build.sbt
  External Libraries
  Scratches and Consoles

Search SparkSQL1.scala
1 import org.apache.spark.sql.SparkSession
2 import org.apache.spark.sql.functions._
3 import org.apache.spark.sql.functions.{lit, struct}
4 import org.apache.log4j.{Level, Logger}
5
6 object SparkSQL1 {
7
8   //Case Class for Holidays
9   case class Holidays (UserID:Int, Country_Name_Dept:String, Country_Name_Arrival:String, modeOfTravel:String, Distance:Int, Year:Int)
10
11   //Case Class for Transport Details
12   case class Transport_Details(Transport_Mode:String, Transport_Exp:Int)
13
14   //Case Class for User Details
15   case class User_Details(UserID:Int, User_Name:String, Age:Int)
16
17   def main(args:Array[String]): Unit = {
18
19     //Let us create a spark session object
20     //Create a case class globally to be used inside the main method
21     val spark = SparkSession
22       .builder()
23       .master("local[*]")
24       .appName("Spark SQL - 1")
25       .getOrCreate()
26
27     // Removing all INFO logs in console printing only result sets
28     val rootLogger = Logger.getLogger("org.apache.spark")
29     rootLogger.setLevel(Level.ERROR)
30
31     import spark.implicits._
32     //Read the Holiday Details from Local file
33
34     val data = spark.sparkContext.textFile(s"${System.getProperty("user.dir")}/src/main/resources/V22_Dataset_Holidays.txt")
35     //Create Holiday DF
36
37     SparkSQL1 main(args: Array[String])
```

```
RDDEspark [C:\Users\Shrutti\IdeaProjects\RDDEspark] - ...src\main\scala\com\sparksql1\assignment\SparkSQL1.scala (RDDEspark) - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
RDDEspark src main scala com sparksql1 assignment SparkSQL1.scala
Project build.sbt SparkSQL1.scala
Project RDDEspark C:\Users\Shrutti\IdeaProjects\RDDEspark
  project [rdde-spark-build] sources root
  spark-warehouse
  src
  main
  scala
  com
    sparkdd.assignment
    RDD_ASSIGNMENT
    sparksql1.assignment
    SparkSQL1
  test
  build.sbt
  External Libraries
  Scratches and Consoles

Search SparkSQL1.scala
48 //Create Transport Details DF by loading the Transport Details file
49 val transportDetailsDF = spark.sparkContext.textFile(s"${System.getProperty("user.dir")}/src/main/resources/V22_Dataset_Transport.txt")
50   .map(_.split(",")).map(x=>Transport_Details(Transport_Mode = x(0), Transport_Exp = x(1).toInt)).toDF()
51
52 //Printing data of Transport Mode DF
53 transportDetailsDF.show()
54
55 //Create User Details DF by loading the User file
56 val userDetailsDF = spark.sparkContext.textFile(s"${System.getProperty("user.dir")}/src/main/resources/V22_Dataset_User_Details.txt")
57   .map(_.split(",")).map(x=>User_Details(UserID = x(0).toInt, User_Name = x(1), Age = x(2).toInt)).toDF()
58
59 //Printing data of Transport Mode DF
60 userDetailsDF.show()
61
62 //Task 1
63 //1) What is the distribution of the total number of air-travelers per year
64
65 //This is by using filter and group by operations on DataFrame
66 holidaysDF.filter((colNameDF: "modeOfTravel"=="airplane")).groupBy((colName "Year").count().show()
67
68 //Below approach is by using SQL in spark
69 holidaysDF.createOrReplaceTempView(TableName: "Holiday_Data")
70 println("Using SQL & Temp View")
71 spark.sql(s"SELECT year, count(year) a from Holiday_Data where modeOfTravel='airplane' group By Year ").show()
72
73 //Task 1
74 //2) What is the total air distance covered by each user per year
75
76 //creating Or replacing the view
77 userDetailsDF.createOrReplaceTempView(TableName: "Users_Data")
78
79 //Approach : By joining two DFs
80 println("Below Result is after joining two Data frames")
81 holidaysDF.as(alias: "HD").join(userDetailsDF.as(alias: "UD"), joinExpr: "UD.UserID===HD.UserID")
82   .groupBy((colName "HD.UserID"), (colName "HD.Year"), (colName "UD.User_Name").sum((colName "Distance")).show()
83
84 SparkSQL1 main(args: Array[String])
```



```
RDDEpark [C:\Users\Shrutu\IdeaProjects\RDDEpark] - ...src\main\scala\com\sparksql1\assignment\SparkSQL1.scala (RDDEpark) - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
RDDEpark src main scala com sparksql1 assignment SparkSQL1.scala
Project
  RDDEpark
    src
      main
        scala
          com
            sparksql1
              assignment
                SparkSQL1.scala
          test
          build.sbt
          External Libraries
          Scratches and Consoles
Run
  SparkSQL1 (1)
  Run
  Build
  SBT Shell
  Terminal
  Event Log
  All files are up-to-date (2 minutes ago)
  23:23 CRLF : UTF-8 :
  11:14 AM
  06-08-2018

31 holidaysDF.show()
32
33 //Create Transport Details DF by loading the Transport_Details file
34 val transportDetailsDF = spark.sparkContext.textFile( RMW = "C:\\Users\\Shrutu\\Downloads\\USDataset\\Transport.txt").
35   map(_.split( RMW = ",")).map(x=>Transport_Details(Transport_Mode = x(0),Transport_Exp = x(1).toInt)).toDF()
36
37 //Printing data of Transport Mode DF
38 transportDetailsDF.show()
39
40 //Create User Details DF by loading the User file
41 val userDetailsDF = spark.sparkContext.textFile( RMW = "C:\\Users\\Shrutu\\Downloads\\USDataset\\User_details.txt").
42   map(_.split( RMW = ",")).map(x=>User_Details(UserID = x(0).toInt,User_Name = x(1),Age = x(2).toInt)).toDF()
43
44 //Printing data of Transport Mode DF
45 userDetailsDF.show()
46
47 //Task 1
48 //1) What is the distribution of the total number of air-travelers per year
49
50 //This is by using filter and group by operations on DataFrame
51 holidaysDF.filter( RMW = "modeOfTravel='airplane'").groupBy( RMW = "Year").count().show()
52
53 //Below approach is by using SQL in spark
54 holidaysDF.createOrReplaceTempView( RMW = "Holiday_Data")
55 println("Using SQL & Temp View")
56 spark.sql( RMW = "Select year, count(year) as count from Holiday_Data where modeOfTravel='airplane' group By Year ").show()
57
58 //Task 1
59 //2) What is the total air distance covered by each user per year
60
61 //creating Or replacing the view
62 userDetailsDF.createOrReplaceTempView( RMW = "Users_Data")
63
64
65 //Approach : By Joining two DFs
66 println("Below Result is after joining two Data frames")
67 holidaysDF.as( RMW = "HD").join(userDetailsDF.as( RMW = "UD"), RMW = "UD.UserID===HD.UserID")
68   .groupBy( RMW = "HD.Year", RMW = "UD.User_Name").sum( RMW = "Distance").show()
69
70 SparkSQL1 main(args: Array[String])
```

```
RDDEpark [C:\Users\Shrutu\IdeaProjects\RDDEpark] - ...src\main\scala\com\sparksql1\assignment\SparkSQL1.scala (RDDEpark) - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
RDDEpark src main scala com sparksql1 assignment SparkSQL1.scala
Project
  RDDEpark
    src
      main
        scala
          com
            sparksql1
              assignment
                SparkSQL1.scala
          test
          build.sbt
          External Libraries
          Scratches and Consoles
Run
  SparkSQL1 (1)
  Run
  Build
  SBT Shell
  Terminal
  Event Log
  All files are up-to-date (2 minutes ago)
  1061 CRLF : UTF-8 :
  11:23 AM
  06-08-2018

+-----+
|Year|count|
+-----+
|1990| 8|
|1994| 1|
|1991| 9|
|1992| 7|
|1993| 7|
+-----+

Using SQL & Temp View
+-----+
|year|count|
+-----+
|1990| 8|
|1994| 1|
|1991| 9|
|1992| 7|
|1993| 7|
+-----+

Below Result is after joining two Data frames
+-----+
|UserID|Year|User_Name|sum(Distance)|
+-----+
|1|1990|mark|200|
|1|1993|mark|600|
|6|1991|peter|400|
|6|1993|peter|200|
|3|1992|luke|200|
|3|1993|luke|200|
|3|1991|luke|200|
|5|1992|mark|400|
|5|1991|mark|200|
|5|1994|mark|200|
|9|1992|thomas|400|
|9|1991|thomas|200|
+-----+
```

```
Run: SparkSQL (1) x
[year] a)
-----
1990) 8)
1994) 1)
1991) 9)
1992) 7)
1993) 7)
-----

Below Result is after joining two Data frames
-----
(UserID|Year|User_Name|sum(Distance))
-----
1|1990|mark|200)
1|1993|mark|400)
6|1991|peter|400)
6|1993|peter|200)
3|1992|luke|200)
3|1993|luke|200)
3|1991|luke|200)
5|1992|mark|400)
5|1991|mark|200)
5|1994|mark|200)
5|1992|thomas|400)
5|1991|thomas|200)
4|1990|lisa|400)
4|1991|lisa|200)
8|1991|andrew|200)
8|1990|andrew|200)
8|1992|andrew|200)
7|1990|james|400)
10|1993|annie|200)
10|1992|annie|200)
-----
only showing top 20 rows

[1,1993,mark,600]
-----
(Country_Name_Arrival|count)
-----
```



```
Run: SparkSQL (1) x
[1,1993,mark,600]
-----
|Country_Name_Arrival|count|
-----
|IND|9|
-----
only showing top 1 row

-----
|Route|Total_Exp|
-----
|[CBN, IND]|680|
-----
only showing top 1 row

-----
|UserID|User_Name|Year|sum(Transport_Exp)|
-----
|1|mark|1990|170|
|1|mark|1992|510|
|2|john|1991|340|
|2|john|1993|170|
|3|luke|1991|170|
|3|luke|1992|170|
|3|luke|1993|170|
|4|lisa|1990|340|
|4|lisa|1991|170|
|5|mark|1991|170|
|5|mark|1992|340|
|5|mark|1994|170|
|6|peter|1991|340|
|6|peter|1993|170|
|7|james|1990|510|
|8|andrew|1990|170|
|8|andrew|1991|170|
|8|andrew|1992|170|
|9|thomas|1991|170|
|9|thomas|1992|340|
-----
SQL showing top 20 rows
```

```
RDDEspark [C:\Users\Shruthi\IdeaProjects\RDDspark] - ...src\main\scala\com\sparksql1\assignment\SparkSQL1.scala [RDDspark] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
RDDEspark src main scala com sparksql1 assignment SparkSQL1.scala
Project view: RDDspark, spark-warehouse, src, main, scala, com, sparksql1, assignment, SparkSQL1.scala
// 5) Which route is generating the most revenue per year
// Approach Using Spark SQL Operations
// First create a new DF where two columns Dept Country and Arrival country should be kept in one column to get distinct routes
val routesDF = holidaysDF.withColumn("Route", struct($"Country_Name_Dept", $"Country_Name_Arrival")).toDF()
routesDF.as($"Route").join(transportDetailDF.as($"Route"), $"Route" === $"TD.Transport_Mode" === $"TD.ModeOfTravel")
.groupBy($"Route").sum($"Sum").withColumnRenamed($"Sum", $"Total_Exp").sort($"Total_Exp").show($"numRows")

// Task 1
// 6) What is the total amount spent by every user on air-travel per year
// Approach : Using spark SQL operations
holidaysDF.as($"H").join(userDetailDF.as($"U"), $"UserID" === $"UD.UserID")
.join(transportDetailDF.as($"T"), $"ModeOfTravel" === $"TD.Transport_Mode")
.groupBy($"UserID", $"Year").sum($"Sum").sort($"UserID", $"Year").show()

// Task 1
// 7) Considering age groups of < 20 , 20-35 , 35 + , Which age group is travelling the most
// every year.
// Another Approach of Case Statement
holidaysDF.as($"H").join(userDetailDF.as($"U"), $"UserID" === $"UD.UserID")
.select($"UserID", $"Age", $"Year", $"Transport_Mode")
.when($"Age" > 20 && $"Age" < 35, $"Between20And35")
.when($"Age" > 35, $"Above35").alias($"AgeGroup")
.groupBy($"AgeGroup").count().sort($"AgeGroup").show($"numRows")

SparkSQL1 (1)
main(args: Array[String])
```

```
RDDEspark [C:\Users\Shruthi\IdeaProjects\RDDspark] - ...src\main\scala\com\sparksql1\assignment\SparkSQL1.scala [RDDspark] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
RDDEspark src main scala com sparksql1 assignment SparkSQL1.scala
Run: SparkSQL1 (1)
only showing top 1 row
+-----+
|UserID|User_Name|Year|sum(Transport_Exp)|
+-----+
| 11 | mark | 1990 | 170 |
| 11 | mark | 1991 | 510 |
| 21 | john | 1991 | 340 |
| 21 | john | 1993 | 170 |
| 31 | luke | 1991 | 170 |
| 31 | luke | 1992 | 170 |
| 31 | luke | 1993 | 170 |
| 41 | lisa | 1990 | 340 |
| 41 | lisa | 1991 | 170 |
| 51 | mark | 1991 | 170 |
| 51 | mark | 1992 | 340 |
| 51 | mark | 1994 | 170 |
| 61 | peter | 1991 | 340 |
| 61 | peter | 1993 | 170 |
| 71 | jamaa | 1990 | 510 |
| 81 | andrew | 1990 | 170 |
| 81 | andrew | 1991 | 170 |
| 81 | andrew | 1992 | 170 |
| 91 | thomas | 1991 | 170 |
| 91 | thomas | 1992 | 340 |
+-----+
only showing top 20 rows
+-----+
| AgeGroup|count|
+-----+
|Between20And35| 13 |
+-----+
only showing top 1 row
Process finished with exit code 0
All files are up-to-date (4 minutes ago)
```