**User Travel Data analysis**

* Extract the text file in rdd.

**val** travelData=sc.textFile("C:\\Users\\Parab\\Desktop\\acadglid\\new\\s18\_rdd cont\\S18\_Dataset.txt")

* map each column in RDD to case object.

**val** travels=travelData.map(\_.split(",")).map(r=> **TravelData**(r(0).toInt,r(1),r(2),r(3),r(4).toInt,r(5).toInt))

* convert Rdd to Dataframe

**val** travelsDF=travels.toDF();

**val** userData=sc.textFile("C:\\Users\\Parab\\Desktop\\acadglid\\new\\s18\_rdd cont\\S18\_Dataset\_User\_details.txt")

**val** users=userData.map(\_.split(",")).map(r=> **UserData**(r(0).toInt,r(1),r(2).toInt))

**val** userDF=users.toDF();

userDF.~~registerTempTable~~("userDF")

* Case classes for both travel and user data.

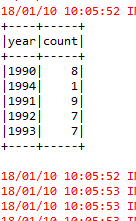
**case** **class** **UserData**(userId:Int, userName:*String*, userAge:Int)

**case** **class** **TravelData**(user: Int, source:*String*, destination:*String*, mode:*String*, distance:Int, year:Int);

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

travelsDF.filter(($"mode").like("airplane%")).groupBy("year").count().show()

O/P:



1. What is the total air distance covered by each user per year.

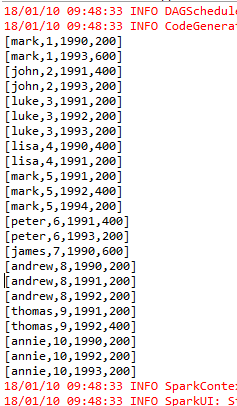
**val** userPerYearDistance= travelsDF.groupBy("user","year").sum("distance")

userPerYearDistance.~~registerTempTable~~("userPerYearDistance")

**val** finalResult=sql.sql("select u.userName,t.\* from userPerYearDistance t join userDF u on t.user=u.userId order by t.user,t.year").collect()

finalResult.foreach(println)

O/P:



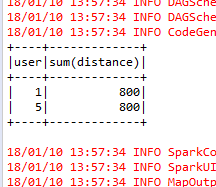
1. Which user has travelled the largest distance till date.

**val** userDistance=travelsDF.groupBy("user").sum("distance").alias("dist")

**val** maxDist=(userDistance.agg(Map("sum(distance)" -> "max")).collect())(0).getLong(0)

**val** userMostDistance=userDistance.filter($"sum(distance)">=maxDist)

userMostDistance.show()



1. What is the most preferred destination for all users.

**val** prefDest=travelsDF.groupBy("destination").count().sort(($"count").desc);

prefDest.show(1);

