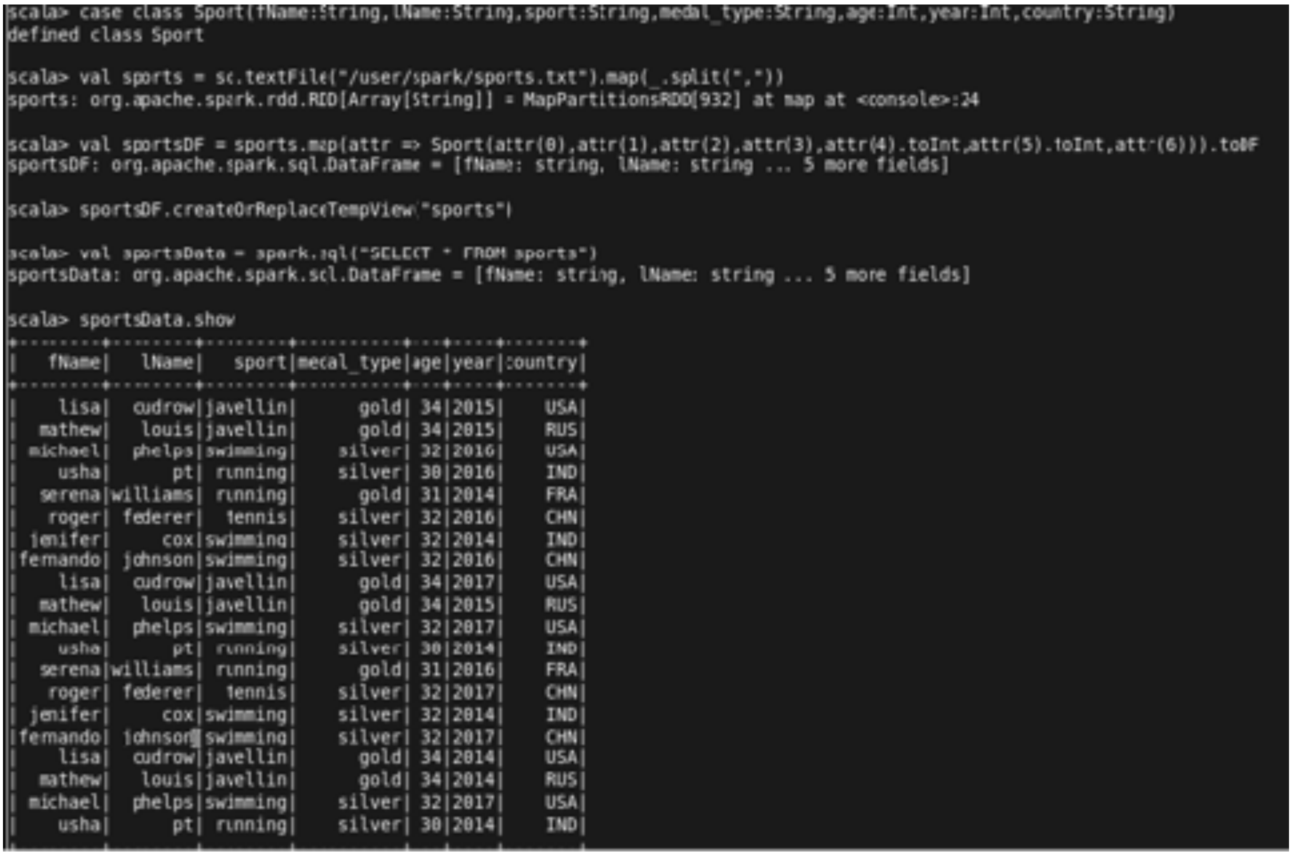
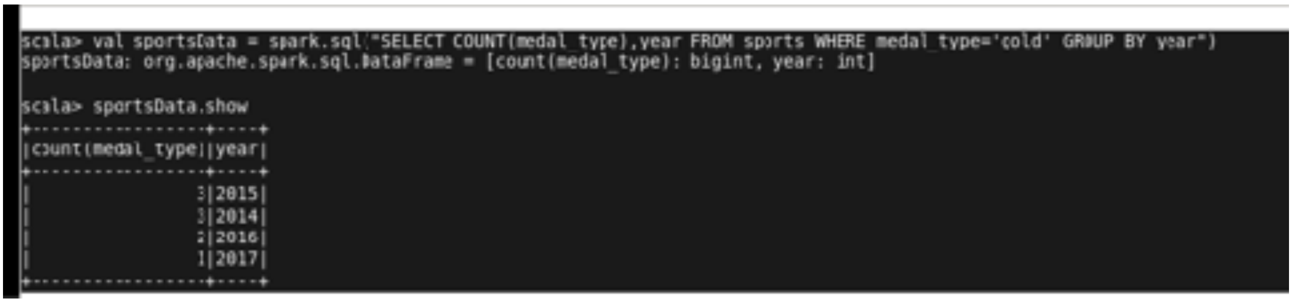
Task 1

Using spark-sql, Find:

1. What are the total number of gold medal winners every year

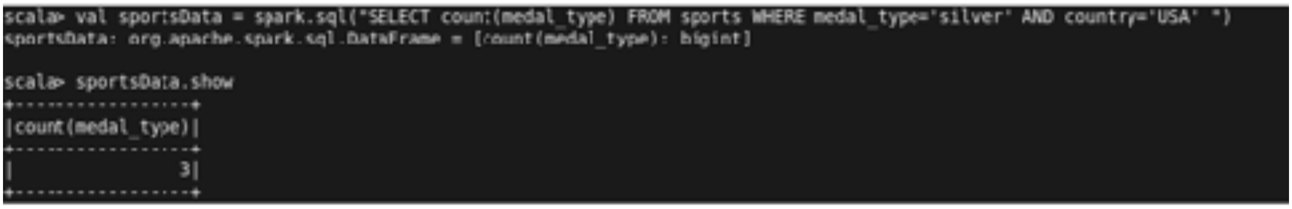


```
scala> case class Sport(fName:String,lName:String,sport:String,medal_type:String,age:Int,year:Int,country:String)
defined class Sport

scala> val sports = sc.textFile("/user/spark/sports.txt").map(_.split(","))
sports: org.apache.spark.rdd.RDD[Array[String]] = MapPartitionsRDD[932] at map at <console>:24

scala> val sportsDF = sports.map(attr => Sport(attr(0),attr(1),attr(2),attr(3),attr(4).toInt,attr(5).toInt,attr(6))).toDF
sportsDF: org.apache.spark.sql.DataFrame = [fName: string, lName: string ... 5 more fields]

scala> sportsDF.createOrReplaceTempView("sports")

scala> val sportsData = spark.sql("SELECT * FROM sports")
sportsData: org.apache.spark.sql.DataFrame = [fName: string, lName: string ... 5 more fields]

scala> sportsData.show
+--------+--------+--------+----------+---+----+-------+
|   fName|   lName|   sport|medal_type|age|year|country|
+--------+--------+--------+----------+---+----+-------+
|    lisa|  cudrow|javellin|      gold| 34|2015|    USA|
|  mathew|   louis|javellin|      gold| 34|2015|    RUS|
| michael|  phelps|swimming|    silver| 32|2016|    USA|
|    usha|      pt| running|    silver| 30|2016|    IND|
|  serena|williams| running|      gold| 31|2014|    FRA|
|   roger| federer|  tennis|    silver| 32|2016|    CHN|
|  jenifer|     cox|swimming|    silver| 32|2014|    IND|
|fernando| johnson|swimming|    silver| 32|2016|    CHN|
|    lisa|  cudrow|javellin|      gold| 34|2017|    USA|
|  mathew|   louis|javellin|      gold| 34|2015|    RUS|
| michael|  phelps|swimming|    silver| 32|2017|    USA|
|    usha|      pt| running|    silver| 30|2014|    IND|
|  serena|williams| running|      gold| 31|2016|    FRA|
|   roger| federer|  tennis|    silver| 32|2017|    CHN|
| jenifer|     cox|swimming|    silver| 32|2014|    IND|
|fernando| johnson|swimming|    silver| 32|2017|    CHN|
|    lisa|  cudrow|javellin|      gold| 34|2014|    USA|
|  mathew|   louis|javellin|      gold| 34|2014|    RUS|
| michael|  phelps|swimming|    silver| 32|2017|    USA|
|    usha|      pt| running|    silver| 30|2014|    IND|
```



```
scala> val sportsData = spark.sql("SELECT COUNT(medal_type),year FROM sports WHERE medal_type='gold' GROUP BY year")
sportsData: org.apache.spark.sql.DataFrame = [count(medal_type): bigint, year: int]

scala> sportsData.show
+-----------------+----+
|count(medal_type)|year|
+-----------------+----+
|                3|2015|
|                3|2014|
|                2|2016|
|                1|2017|
+-----------------+----+
```

2. How many silver medals have been won by USA in each sport

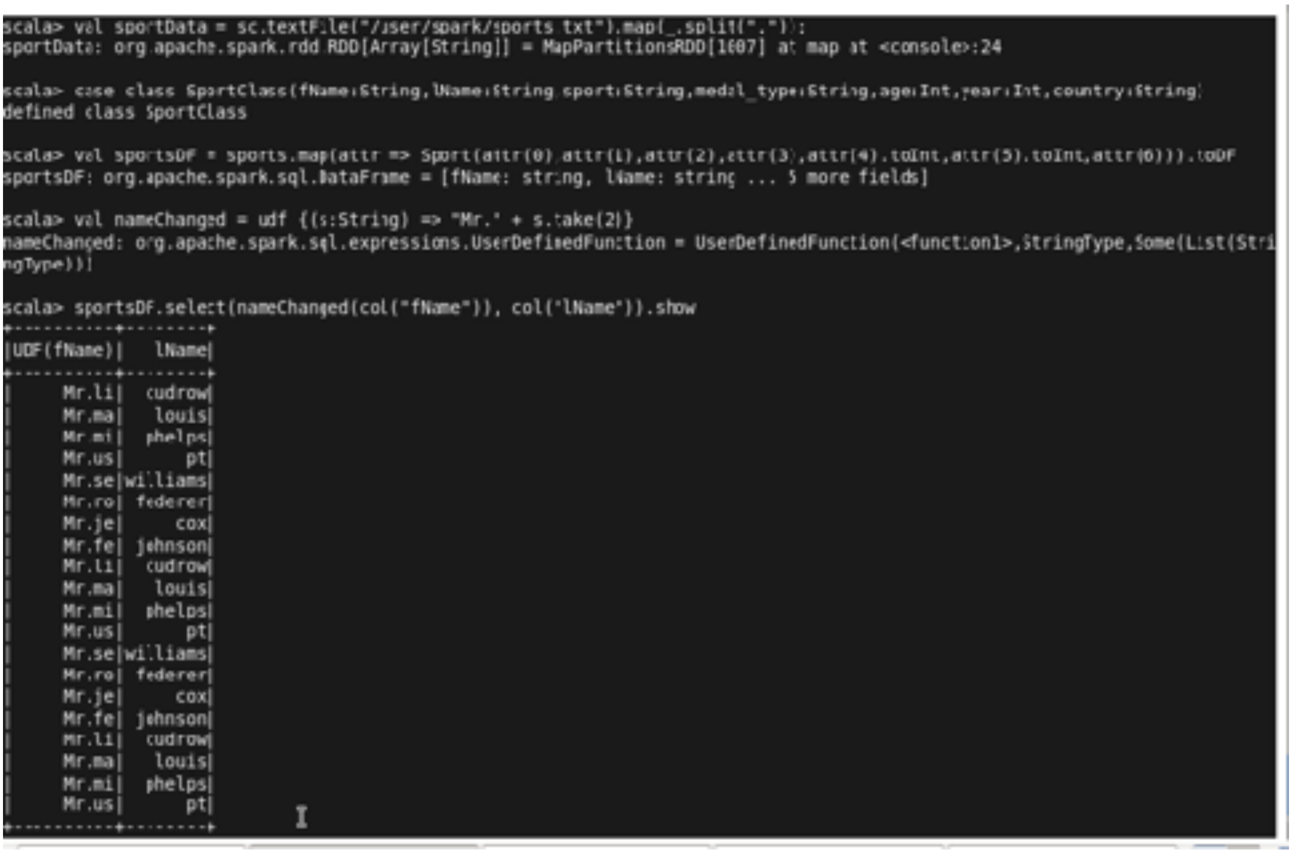


```
scala> val sportsData = spark.sql("SELECT count(medal_type) FROM sports WHERE medal_type='silver' AND country='USA' ")
sportsData: org.apache.spark.sql.DataFrame = [count(medal_type): bigint]

scala> sportsData.show
+-----------------+
|count(medal_type)|
+-----------------+
|                3|
+-----------------+
```

Task 2

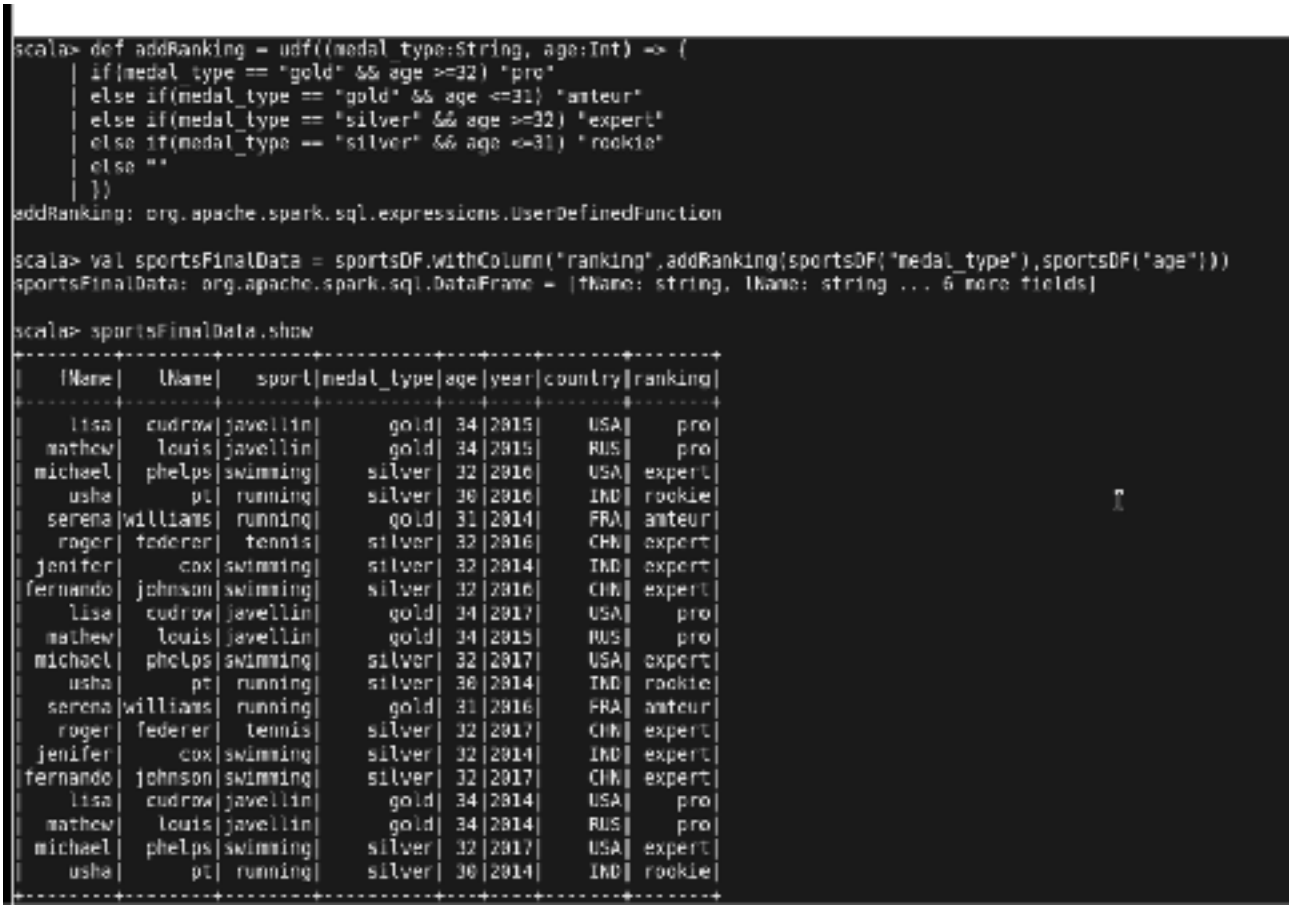
Using udfs on dataframe

1. Change firstname, lastname columns into
Mr.first_two_letters_of_firstname<space>lastname
for example - michael, phelps becomes Mr.mi phelps



```
scala> val sportData = sc.textFile("/user/spark/sports.txt").map(_.split(","));
sportData: org.apache.spark.rdd.RDD[Array[String]] = MapPartitionsRDD[1607] at map at <console>:24

scala> case class SportClass(fName:String,lName:String,sport:String,medal_type:String,age:Int,year:Int,country:String)
defined class SportClass

scala> val sportsDF = sports.map(attr => Sport(attr(0),attr(1),attr(2),attr(3),attr(4).toInt,attr(5).toInt,attr(6))).toDF
sportsDF: org.apache.spark.sql.DataFrame = [fName: string, lName: string ... 5 more fields]

scala> val nameChanged = udf {(s:String) => "Mr." + s.take(2)}
nameChanged: org.apache.spark.sql.expressions.UserDefinedFunction = UserDefinedFunction(<function1>,StringType,Some(List(StringType)))

scala> sportsDF.select(nameChanged(col("fName")), col("lName")).show
+----------+--------+
|UDF(fName)|   lName|
+----------+--------+
|     Mr.li|  cudrow|
|     Mr.ma|   louis|
|     Mr.mi|  phelps|
|     Mr.us|      pt|
|     Mr.se|williams|
|     Mr.ro| federer|
|     Mr.je|     cox|
|     Mr.fe| johnson|
|     Mr.li|  cudrow|
|     Mr.ma|   louis|
|     Mr.mi|  phelps|
|     Mr.us|      pt|
|     Mr.se|williams|
|     Mr.ro| federer|
|     Mr.je|     cox|
|     Mr.fe| johnson|
|     Mr.li|  cudrow|
|     Mr.ma|   louis|
|     Mr.mi|  phelps|
|     Mr.us|      pt|
+----------+--------+
```

2. Add a new column called ranking using udfs on dataframe, where : gold medalist, with age >= 32 are ranked as pro
gold medalists, with age <= 31 are ranked amateur
silver medalist, with age >= 32 are ranked as expert
silver medalists, with age <= 31 are ranked rookie



```
scala> def addRanking = udf((medal_type:String, age:Int) => {
     | if(medal_type == "gold" && age >=32) "pro"
     | else if(medal_type == "gold" && age <=31) "amteur"
     | else if(medal_type == "silver" && age >=32) "expert"
     | else if(medal_type == "silver" && age <=31) "rookie"
     | else ""
     | })
addRanking: org.apache.spark.sql.expressions.UserDefinedFunction

scala> val sportsFinalData = sportsDF.withColumn("ranking",addRanking(sportsDF("medal_type"),sportsDF("age")))
sportsFinalData: org.apache.spark.sql.DataFrame = [fName: string, lName: string ... 6 more fields]

scala> sportsFinalData.show
+--------+--------+--------+----------+---+----+-------+-------+
|   fName|   lName|   sport|medal_type|age|year|country|ranking|
+--------+--------+--------+----------+---+----+-------+-------+
|    lisa|  cudrow|javellin|      gold| 34|2015|    USA|    pro|
|  mathew|   louis|javellin|      gold| 34|2015|    RUS|    pro|
| michael|  phelps|swimming|    silver| 32|2016|    USA| expert|
|    usha|      pt| running|    silver| 30|2016|    IND| rookie|
|  serena|williams| running|      gold| 31|2014|    FRA| amteur|
|   roger| federer|  tennis|    silver| 32|2016|    CHN| expert|
| jenifer|     cox|swimming|    silver| 32|2014|    IND| expert|
|fernando| johnson|swimming|    silver| 32|2016|    CHN| expert|
|    lisa|  cudrow|javellin|      gold| 34|2017|    USA|    pro|
|  mathew|   louis|javellin|      gold| 34|2015|    RUS|    pro|
| michael|  phelps|swimming|    silver| 32|2017|    USA| expert|
|    usha|      pt| running|    silver| 30|2014|    IND| rookie|
|  serena|williams| running|      gold| 31|2016|    FRA| amteur|
|   roger| federer|  tennis|    silver| 32|2017|    CHN| expert|
| jenifer|     cox|swimming|    silver| 32|2014|    IND| expert|
|fernando| johnson|swimming|    silver| 32|2017|    CHN| expert|
|    lisa|  cudrow|javellin|      gold| 34|2014|    USA|    pro|
|  mathew|   louis|javellin|      gold| 34|2014|    RUS|    pro|
| michael|  phelps|swimming|    silver| 32|2017|    USA| expert|
|    usha|      pt| running|    silver| 30|2014|    IND| rookie|
+--------+--------+--------+----------+---+----+-------+-------+
```