Spark Assignment 19.1

Given the below dataset . Solve the below mentioned problem statement in spark Sql.

19_Sports_Data.txt as a text file
(firstname,lastname,sports,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
serena,williams,running,gold,31,2016,FRA
roger,federer,tennis,silver,32,2014,CHN
jenifer,cox,swimming,silver,32,2017,IND
fernando,johnson,swimming,silver,32,2017,CHN
```

This assignment is done in the spark shell within the acadgildVM.

Steps Followed:

1) Copied the dataset file in the path /home/acadgild/spark/19_Sports_data.txt Then read the text file by using Sparksession(spark.read.textfile) as below and created a dataframe.

```
val df1= spark.read.textFile("/home/acadgild/spark/19_Sports_data.txt")
```

2) Created case class Sport_Data by specifying columns and datatypes, to describe the contents of the rows.

```
case class Sport_Data(firstname: String, lastname: String,sports: String,medal_type: String, age:Int,year:Int, country: String)
```

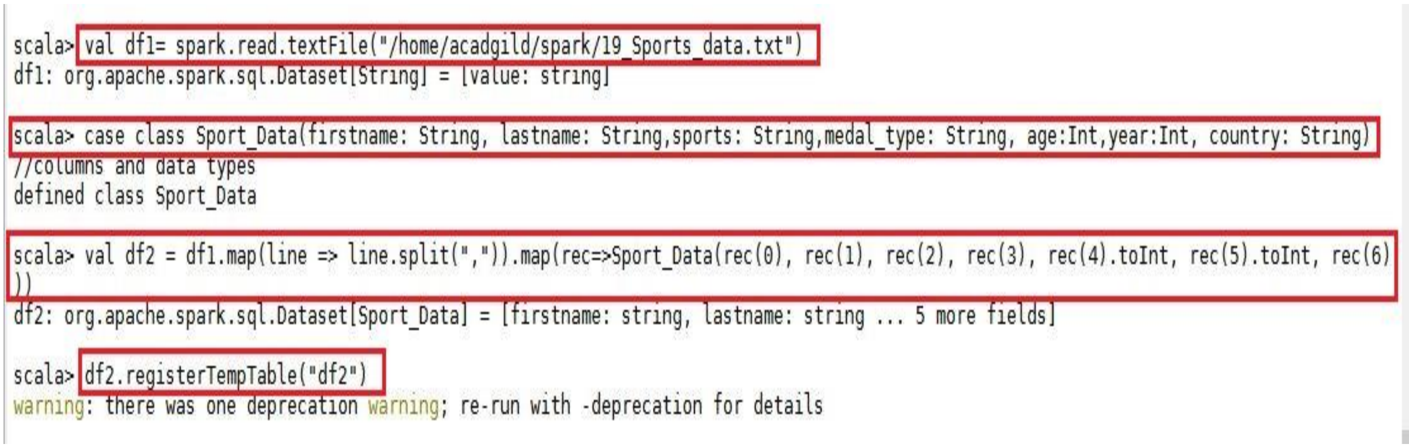
3) Then created drataframe df2, by using map function over the df1.

```
val df2 = df1.map(line => line.split(",")).map(rec=>Sport_Data(rec(0), rec(1), rec(2), rec(3), rec(4).toInt, rec(5).toInt, rec(6)))
```

4) Registered df2 as TempTable as below.

```
df2.registerTempTable("df2")
```

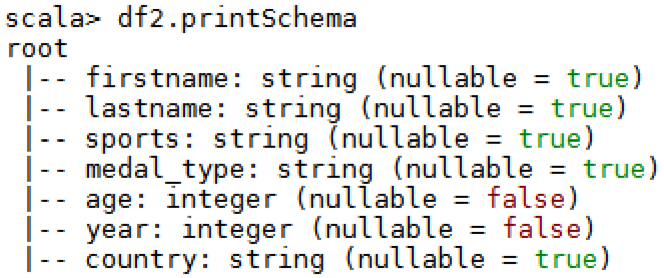
Spark-shell output



```
scala> val df1= spark.read.textFile("/home/acadgild/spark/19_Sports_data.txt")
df1: org.apache.spark.sql.Dataset[String] = [value: string]

scala> case class Sport_Data(firstname: String, lastname: String,sports: String,medal_type: String, age:Int,year:Int, country: String)
//columns and data types
defined class Sport_Data

scala> val df2 = df1.map(line => line.split(",")).map(rec=>Sport_Data(rec(0), rec(1), rec(2), rec(3), rec(4).toInt, rec(5).toInt, rec(6)
))
df2: org.apache.spark.sql.Dataset[Sport_Data] = [firstname: string, lastname: string ... 5 more fields]

scala> df2.registerTempTable("df2")
warning: there was one deprecation warning; re-run with -deprecation for details
```

5) The schema of df2 is as below

Spark-shell output



```
scala> df2.printSchema
root
 |-- firstname: string (nullable = true)
 |-- lastname: string (nullable = true)
 |-- sports: string (nullable = true)
 |-- medal_type: string (nullable = true)
 |-- age: integer (nullable = false)
 |-- year: integer (nullable = false)
 |-- country: string (nullable = true)
```

Problem Statement:

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

To find the total number of gold medal winners every year, we are calling the filter function over df2 and filtering the data with medal type which equals gold and creating a dataframe as df3.

```
val df3 =df2.filter(df2("medal_type")==="gold")
```

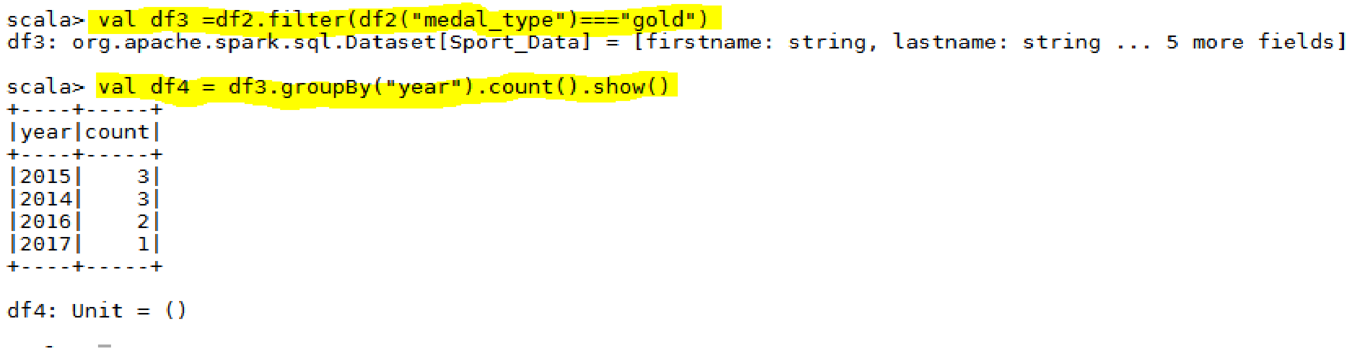
Once filtered, we are calling the groupby using year function over df3 and counting the number of items.

```
val df4 = df3.groupBy("year").count().show()
```

Output

```
|year|count|

|2015|   3|

|2014|   3|

|2016|   2|

|2017|   1|
```

Spark-shell output



```
scala> val df3 =df2.filter(df2("medal_type")==="gold")
df3: org.apache.spark.sql.Dataset[Sport_Data] = [firstname: string, lastname: string ... 5 more fields]

scala> val df4 = df3.groupBy("year").count().show()
+----+-----+
|year|count|
+----+-----+
|2015|    3|
|2014|    3|
|2016|    2|
|2017|    1|
+----+-----+

df4: Unit = ()
```

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

To find the number of silver medal won by USA in each sport, we are calling the filter function over df2 and filtering the data with medal type which equals silver and country which equals USA, then creating a dataframe as df5.

```
val df5 =df2.filter(df2("medal_type")==="silver" && df2("country")==="USA" )
```

Once filtered, we are calling the groupby using sport function over df5 and counting the number of items.

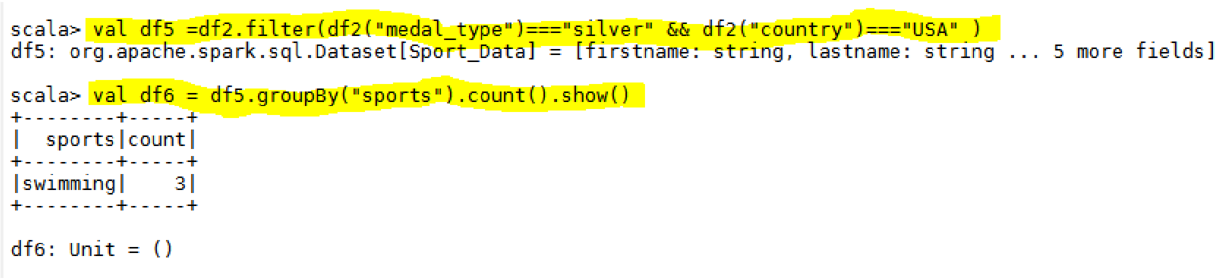
```
val df6 = df5.groupBy("sports").count().show()
```

Output

```
|  sports|count|

+--------+-----+
```

|swimming| 3|

Spark-shell ouput



```
scala> val df5 =df2.filter(df2("medal_type")==="silver" && df2("country")==="USA" )
df5: org.apache.spark.sql.Dataset[Sport_Data] = [firstname: string, lastname: string ... 5 more fields]

scala> val df6 = df5.groupBy("sports").count().show()
+--------+-----+
|  sports|count|
+--------+-----+
|swimming|    3|
+--------+-----+

df6: Unit = ()
```