

## Assignment19.1

Using spark-sql, Find:

1. What is the total number of gold medal winners every year?
2. How many silver medals have been won by USA in each sport?

```
scala> val rdd=sc.textFile("Sports_data.txt")
rdd: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[1] at textFile at <console>:27
```

```
scala> val head=rdd.first()
head: String = firstname,lastname,sports,medal_type,age,year,country
```

```
scala> val filter_set=rdd.filter(x=>x!=head)
filter_set: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[2] at filter at <console>:31
```

```
scala> val sports=filter_set.map(x=>x.split(",")).map(arrays
=>(arrays(0),arrays(1),arrays(2),arrays(3),arrays(4),arrays(5),arrays(6))).toDF("firstname","lastname","
sports","medal_type","age","year","country")
sports: org.apache.spark.sql.DataFrame = [firstname: string, lastname: string, sports: string,
medal_type: string, age: string, year: string, country: string]
```

```
scala> sports.registerTempTable("sportstable")
```

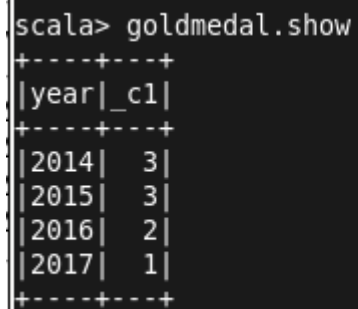
```
scala> sports.show
+-----+-----+-----+-----+-----+-----+
|firstname|lastname| sports|medal_type|age|year|country|
+-----+-----+-----+-----+-----+-----+
|lisa|cudrow|javellin|gold|34|2015|USA|
|mathew|louis|javellin|gold|34|2015|RUS|
|michael|phelps|swimming|silver|32|2016|USA|
|usha|pt|running|silver|30|2016|IND|
|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|
+-----+-----+-----+-----+-----+-----+
```

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

```
scala> val goldmedal= sqlContext.sql("SELECT year, count(*) FROM sportsTable where  
medal_type='gold' group by year order by year")  
goldmedal: org.apache.spark.sql.DataFrame = [year: string, _c1: bigint]
```

```
scala> goldmedal.show
```

```
+----+----+  
|year|_c1|  
+----+----+  
|2014| 3|  
|2015| 3|  
|2016| 2|  
|2017| 1|  
+----+----+
```



```
scala> goldmedal.show  
+----+----+  
|year|_c1|  
+----+----+  
|2014| 3|  
|2015| 3|  
|2016| 2|  
|2017| 1|  
+----+----+
```

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

```
scala> val silver= sqlContext.sql("SELECT country, sports, count(*) FROM sportsTable where
```

```
country='USA' and medal_type='silver' group by country,sports");  
silver: org.apache.spark.sql.DataFrame = [country: string, sports: string, _c2: bigint]
```

```
scala> silver.show  
+-----+-----+----+  
|country| sports|_c2|  
+-----+-----+----+  
|   USA|swimming| 3|  
+-----+-----+----+
```

```
scala> silver.show  
+-----+-----+----+  
|country| sports|_c2|  
+-----+-----+----+  
|   USA|swimming| 3|  
+-----+-----+----+
```