

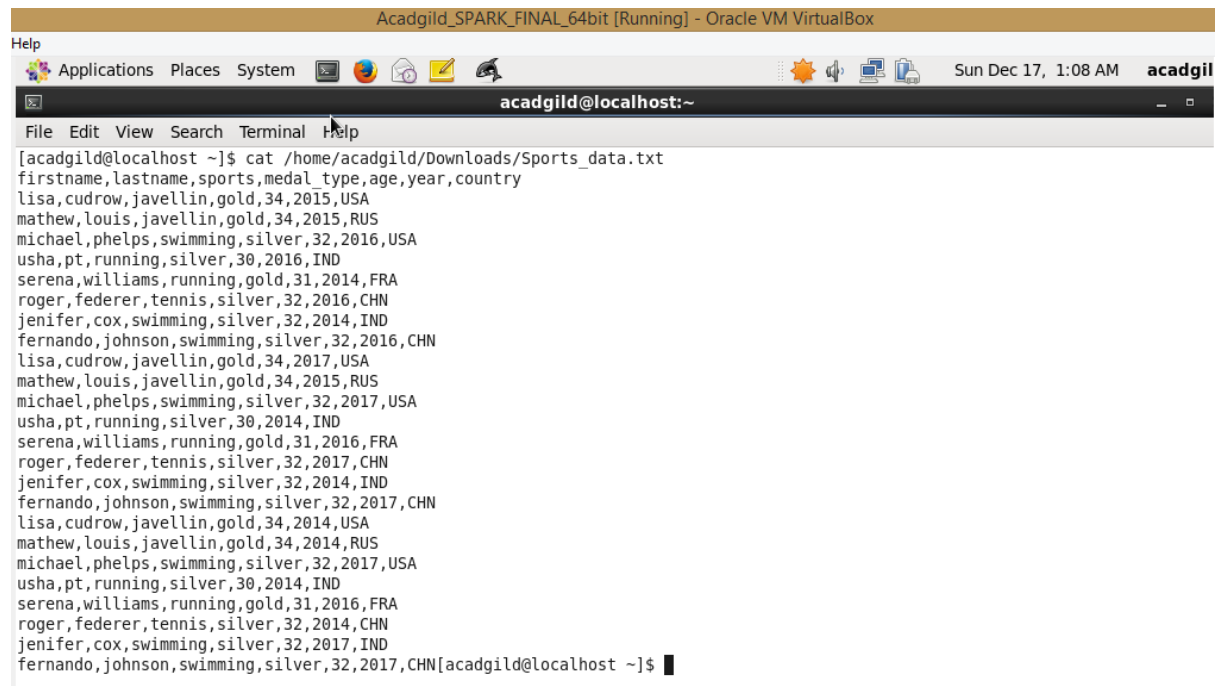
Problem Statement

Using spark-sql, Find:

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

DataSet

The following dataset is a Sports Dataset with columns: **firstname, lastname, sports, medal_type, age, year, country**



```
Acadgild_SPARK_FINAL_64bit [Running] - Oracle VM VirtualBox
Help
Applications Places System [Icons] [Network] [Sound] [USB] [Printer] [Mouse] [Keyboard] [Camera] [Audio] [Video] [Display] [Power] [Sun Dec 17, 1:08 AM] acadgil
acadgild@localhost:~
File Edit View Search Terminal Help
[acadgild@localhost ~]$ cat /home/acadgild/Downloads/Sports_data.txt
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[acadgild@localhost ~]$
```

Importing Spark SQL Packages

```
scala> import org.apache.spark.sql._
import org.apache.spark.sql._

scala>

scala> import sqlContext.implicits._
import sqlContext.implicits._
```

Converting text file into RDD with the help of SPARK CONTEXT object

```
scala> val sportsRDD = sc.textFile("/home/acadgild/Downloads/Sports_data.txt")
sportsRDD: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[163] at textFile at <console>:43
scala>
```

Put the first line of file into RDD which is header and remove the header from RDD for data manipulation

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

scala> val sportsdstarRDD = sportsRDD.filter(record => (record != header))
sportsdstarRDD: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[164] at filter at <console>:47

scala>
```

Printing Rdd which returns Array[String]

```
scala> sportsdstaRDD.foreach(println)
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
```

Class Defination for sports data class and defining its respective schema

```
scala> case class
|   sportsdata(firstname:String,
|   lastname:String,
|   sports:String,
|   medal_type:String,
|   age:Integer,
|   year:Integer,
|   country:String)
defined class sportsdata
scala>
```

Mapping the record into word delimited by comma , defining the columns and converting to dataframe

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

Register temporary table sports for querying data

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

Querying the entire table via sqlContext object and displaying the same

```
scala> sqlContext.sql("SELECT * FROM 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|
+-----+-----+-----+-----+-----+-----+
only showing top 20 rows
```

✓ What are the total number of gold medal winners every year

Code:

```
scala> val total_gold_medals_per_year=sqlContext.sql("SELECT year, COUNT(medal_type) AS total_gold_medals FROM sports where m
edal_type = 'gold' GROUP BY year")
total_gold_medals_per_year: org.apache.spark.sql.DataFrame = [year: int, total_gold_medals: bigint]
```

Output:

```
scala> total_gold_medals_per_year.show()
+-----+-----+
|year|total_gold_medals|
+-----+-----+
|2014| 3|
|2015| 3|
|2016| 2|
|2017| 1|
+-----+-----+

scala> █
```

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

Code:

```
scala> val Total_Silver_Medal_USA_Won = sqlContext.sql("SELECT sports,count(*) as total_silver_medals_won_by_USA FROM sports  
where medal type ='silver' and country = 'USA' GROUP BY sports")  
Total_Silver_Medal_USA_Won: org.apache.spark.sql.DataFrame = [sports: string, total_silver_medals_won_by_USA: bigint]
```

Output:

```
scala> Total_Silver_Medal_USA_Won.show()  
+-----+-----+  
| sports|total_silver_medals_won_by_USA|  
+-----+-----+  
|swimming|3|  
+-----+-----+  
I  
scala> █
```