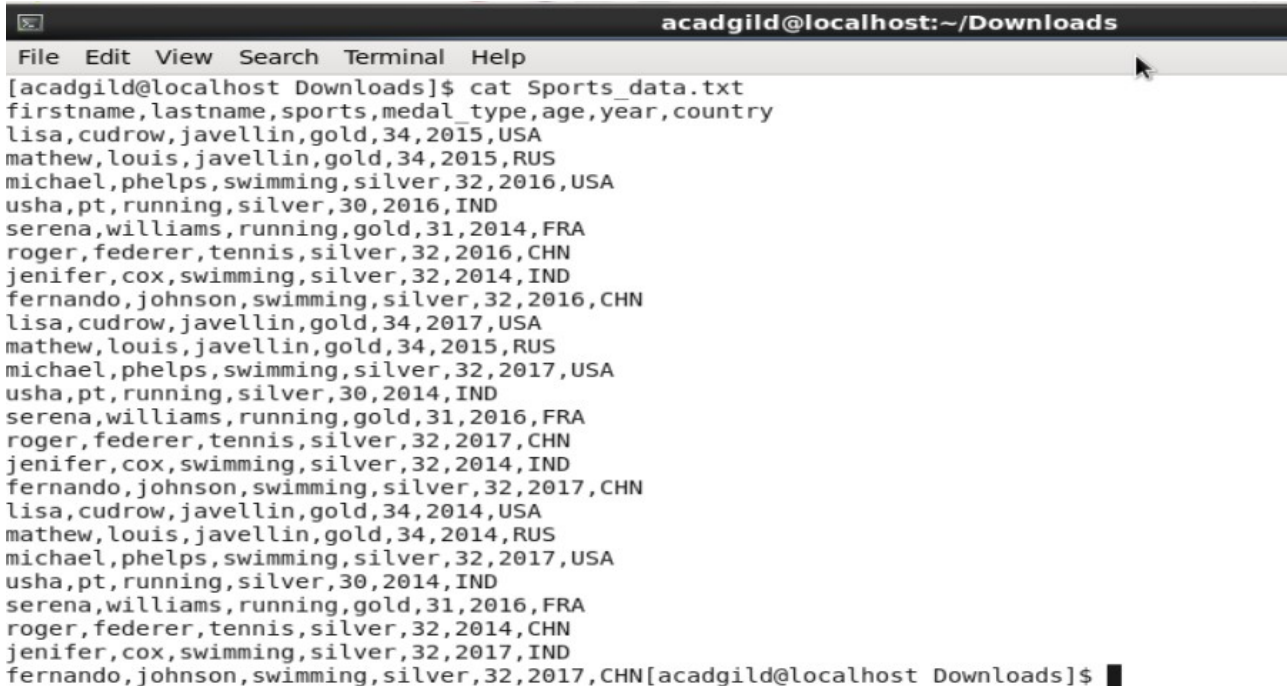


# BigData Assignment 7.5

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

The content of the dataset



```
acadgild@localhost:~/Downloads
File Edit View Search Terminal Help
[acadgild@localhost Downloads]$ cat 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 Downloads]$
```

Solution -

```
//Read the dataset.
```

```
var sports_data =
```

```
sc.textFile("/home/acadgild/Downloads/Sports_data.txt")
```

```
//Stored the header
```

```
val header = sports_data.first()
```

```
//Filtered the data without header
```

```
val actual_data = sports_data.filter(row => row != header)
```

```
//Created the dataframe where the rows are splitted and converted to its appropriate type and namme is assigned to the columns.
```

```
val sportsDF = actual_data.map(x => x.split(",")).map(x =>
```

**x(0),x(1),x(2),x(3),x(4).toInt,x(5).toInt,x(6))).toDF("firstname","lastname","sports","medal\_type","age","year","country")**

//Created a table of the dataframe

**sportsDF.createOrReplaceTempView("SportsData")**

```
acadgild@localhost:~/RITESH/7.1
File Edit View Search Terminal Help

scala> val sports_data = sc.textFile("file:///home/acadgild/Downloads/Sports_data.txt")
sports_data: org.apache.spark.rdd.RDD[String] = file:///home/acadgild/Downloads/Sports_data.txt MapPartitionsRDD[45] at textFile at <console>:24

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

scala> val actual_data = sports_data.filter(row=> row!=header)
actual_data: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[46] at filter at <console>:28

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

scala> sportsDF.createOrReplaceTempView("SportsData")

scala> █
```

1. Selected the rows medal type = gold , then grouped the data to find its count.

**val goldPerYear = spark.sql("select year, medal\_type, count(medal\_type) from SportsData where medal\_type = 'gold' group by year,medal\_type")**

```
scala> val goldPerYear = spark.sql("select year , medal_type , count(medal_type) from SportsData where medal_type = 'gold' group by year , medal_type")
goldPerYear: org.apache.spark.sql.DataFrame = [year: int, medal_type: string ... 1 more field]

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

2. Selected the rows where country = USA and medal type = silver , then grouped the data according to sports to find its count sport wise

**val SilverUsaSport = spark.sql("select country, sports, medaltype, count(medaltype) from SportsData where country = 'USA' and medal\_type = 'silver' group by country,sports,medal\_type")**

```
scala> val silverUsaSport = spark.sql("select country, sports, medal_type , count(medal_type) from SportsData where country = 'USA' and medal_type = 'silver' group by country,sports,medal_type")
silverUsaSport: org.apache.spark.sql.DataFrame = [country: string, sports: string ... 2 more fields]

scala> silverUsaSport.show
+-----+-----+-----+-----+
|country|sports|medal_type|count(medal_type)|
+-----+-----+-----+-----+
|USA|swimming|silver|3|
+-----+-----+-----+-----+
```