

Assignment 19.1:

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

Initial Steps:

```
val sports_data_with_header = sc.textFile("/home/acadgild/spark/Sports_data.txt")

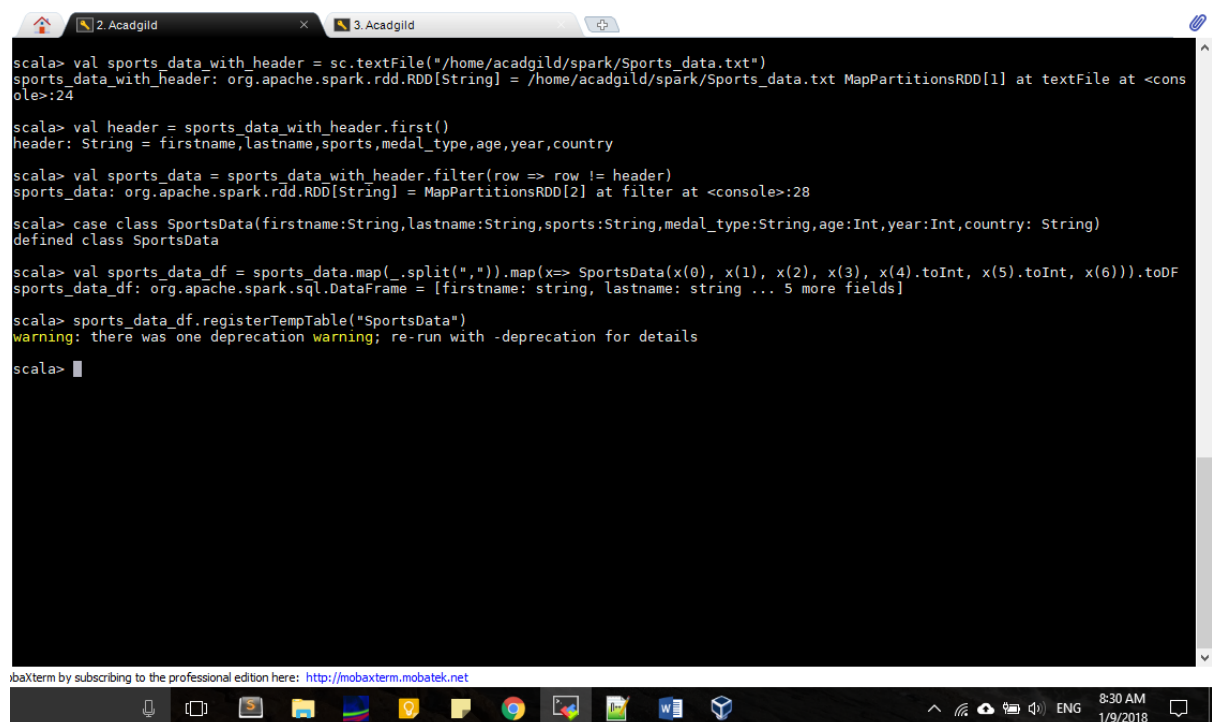
val header = sports_data_with_header.first()

val sports_data = sports_data_with_header.filter(row => row != header)

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

val sports_data_df = sports_data.map(_.split(",")).map(x=> SportsData(x(0), x(1), x(2), x(3),
x(4).toInt, x(5).toInt, x(6))).toDF

sports_data_df.registerTempTable("SportsData")
```



```
scala> val sports_data_with_header = sc.textFile("/home/acadgild/spark/Sports_data.txt")
sports_data_with_header: org.apache.spark.rdd.RDD[String] = /home/acadgild/spark/Sports_data.txt MapPartitionsRDD[1] at textFile at <console>:24

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

scala> val sports_data = sports_data_with_header.filter(row => row != header)
sports_data: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[2] at filter at <console>:28

scala> case class SportsData(firstname:String,lastname:String,sports:String,medal_type:String,age:Int,year:Int,country: String)
defined class SportsData

scala> val sports_data_df = sports_data.map(_.split(",")).map(x=> SportsData(x(0), x(1), x(2), x(3), x(4).toInt, x(5).toInt, x(6))).toDF
sports_data_df: org.apache.spark.sql.DataFrame = [firstname: string, lastname: string ... 5 more fields]

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

scala>
```

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

Using count function on SportsData table, select total number of gold medal by having query criteria medal_type as gold and group by year

Query:

spark.sqlContext.sql("SELECT year, count(*) AS no_of_gold_medals FROM SportsData WHERE medal_type='gold' GROUP BY year ORDER BY year").show()

```
scala> spark.sqlContext.sql("SELECT year, count(*) AS no_of_gold_medals FROM SportsData WHERE medal_type='gold' GROUP BY year ORDER BY year").show()
+-----+
|year|no_of_gold_medals|
+-----+
|2014|3|
|2015|3|
|2016|2|
|2017|1|
+-----+

scala> █
```

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

Using count function on SportsData table, select total number of silver medal by having query criteria medal_type as gold, country as USA and group by sports.

Query:

spark.sqlContext.sql("SELECT sports, count(*) AS no_of_silver_medals FROM SportsData WHERE country='USA' and medal_type='silver' GROUP BY sports ORDER BY sports").show()

```
scala> spark.sqlContext.sql("SELECT sports, count(*) AS no_of_silver_medals FROM SportsData WHERE country='USA' and medal_type='silver' GROUP BY sports ORDER BY sports").show()
+-----+
|sports|no_of_silver_medals|
+-----+
|swimming|3|
+-----+

scala> █
```