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

Code Of Question 1 And 2:

```
import org.apache.spark.sql.Row;

import
org.apache.spark.sql.types.{StructType,StructField,StringType,NumericType,IntegerType};

val Sports_data = sc.textFile("/home/acadgild/Assignment-20/Sports_data.txt")

val schemaString =
"firstname:string,lastname:string,sports:string,medal_type:string,age:integer,year:integer,coun
try:string"

val schema = StructType(schemaString.split(",").map(fieldInfo =>
StructField(fieldInfo.split(":")(0), if (fieldInfo.split(":")(1).equals("string")) StringType else
IntegerType, true)))

val rowRDD = Sports_data.map(_.split(",")).map(r => Row(r(0), r(1), r(2), r(3), r(4).toInt,
r(5).toInt, r(6)))

val SportsDF = spark.createDataFrame(rowRDD, schema)

SportsDF.createOrReplaceTempView("SportsData")

val resultDF = spark.sql("SELECT * FROM SportsData")

resultDF.show()
```

Question 1:

```
val no_of_gold = spark.sql("SELECT year, count(*) FROM SportsData WHERE medal_type ==
'gold' GROUP BY year").show()
```

Question 2:

```
val silver =spark.sql("SELECT sports, count(*) FROM SportsData WHERE country ='USA' AND
medal_type = 'silver' GROUP BY sports").show()
```

Screen-Shot

```
scala> import org.apache.spark.sql.Row;
import org.apache.spark.sql.Row

scala> import org.apache.spark.sql.types.{StructType,StructField,StringType,NumericType,IntegerType};
import org.apache.spark.sql.types.{StructType, StructField, StringType, NumericType, IntegerType}

scala>

scala> val Sports_data = sc.textFile("/home/acadgild/Assignment-20/Sports_data.txt")
Sports_data: org.apache.spark.rdd.RDD[String] = /home/acadgild/Assignment-20/Sports_data.txt MapPartitionsRDD[1] at textFile at <console>
:26

scala>

scala> val schemaString = "firstname:string,lastname:string,sports:string,medal_type:string,age:integer,year:integer,country:string"
schemaString: String = firstname:string,lastname:string,sports:string,medal_type:string,age:integer,year:integer,country:string

scala>

scala> val schema = StructType(schemaString.split(",").map(fieldInfo => StructField(fieldInfo.split(":")(0), if (fieldInfo.split(":")(1).
equals("string")) StringType else IntegerType, true)))
schema: org.apache.spark.sql.types.StructType = StructType(StructField(firstname,StringType,true), StructField(lastname,StringType,true),
 StructField(sports,StringType,true), StructField(medal_type,StringType,true), StructField(age,IntegerType,true), StructField(year,Intege
rType,true), StructField(country,StringType,true))

scala>

scala> val rowRDD = Sports_data.map(_.split(",")).map(r => Row(r(0), r(1), r(2), r(3), r(4).toInt, r(5).toInt, r(6)))
rowRDD: org.apache.spark.rdd.RDD[org.apache.spark.sql.Row] = MapPartitionsRDD[3] at map at <console>:28

scala>

scala> val SportsDF = spark.createDataFrame(rowRDD, schema)
SportsDF: org.apache.spark.sql.DataFrame = [firstname: string, lastname: string ... 5 more fields]
```

```
scala> SportsDF.createOrReplaceTempView("SportsData")

scala>

scala> val resultDF = spark.sql("SELECT * FROM SportsData")
resultDF: org.apache.spark.sql.DataFrame = [firstname: string, lastname: string ... 5 more fields]

scala>

scala> resultDF.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
```



```
scala> //Question 1

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

no_of_gold: Unit = ()

scala> //Question 2

scala> val silver =spark.sql("SELECT sports, count(*) FROM SportsData WHERE country ='USA' AND medal_type = 'silver' GROUP BY sports").sh
ow()
+--------+--------+
|  sports|count(1)|
+--------+--------+
|swimming|       3|
+--------+--------+

silver: Unit = ()

scala>
```