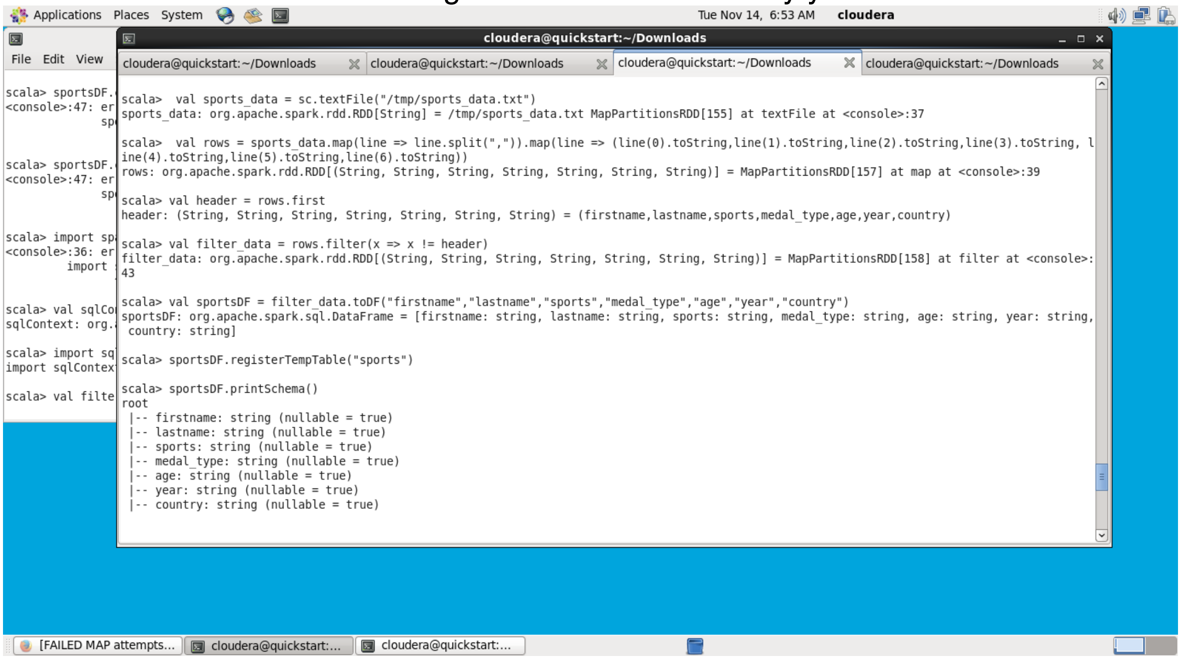
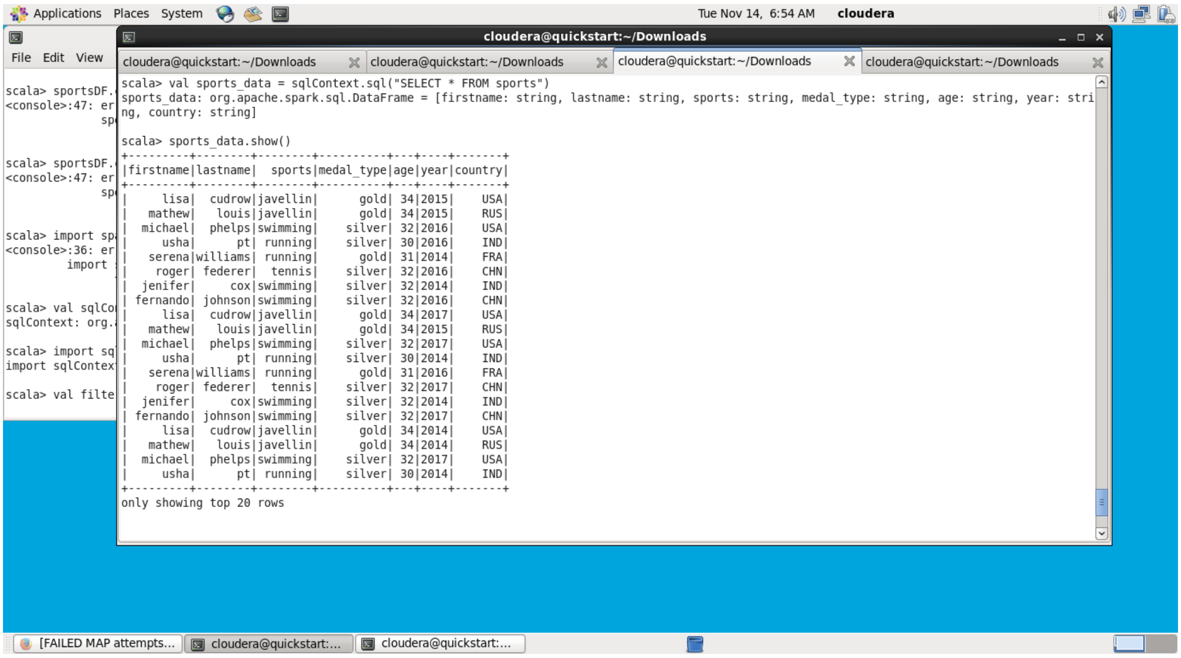
Using spark-sql, Find:

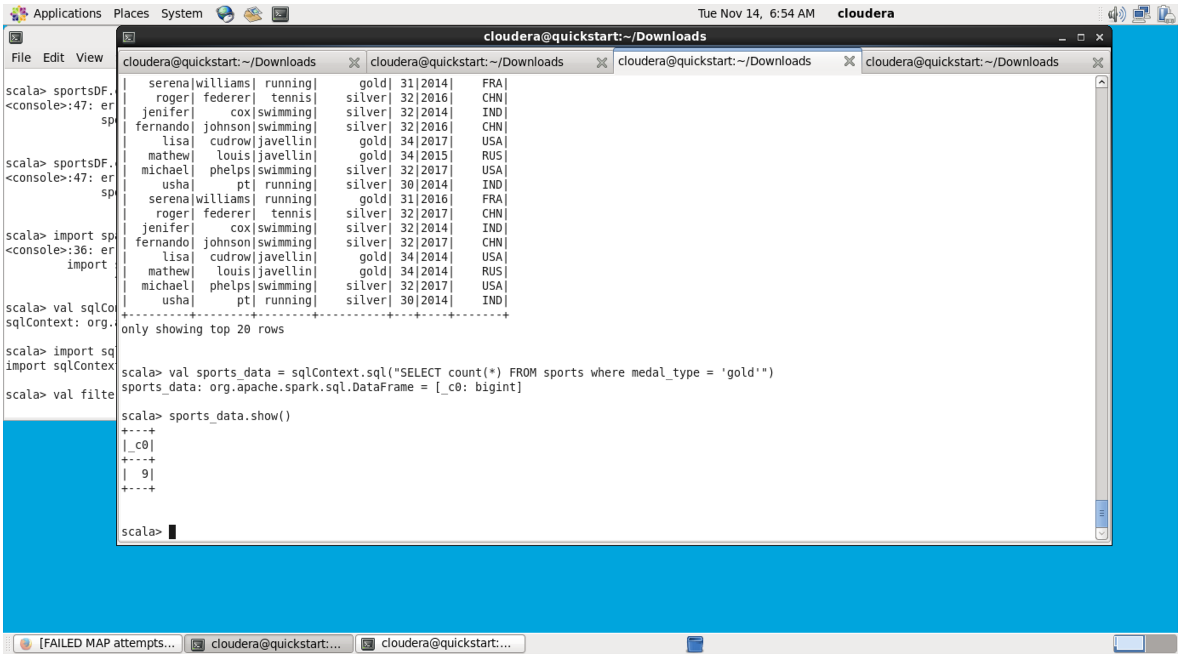
1. What are the total number of gold medal winners every year


Applications Places System
Tue Nov 14, 6:53 AM cloudera
cloudera@quickstart:~/Downloads
cloudera@quickstart:~/Downloads
cloudera@quickstart:~/Downloads
cloudera@quickstart:~/Downloads
cloudera@quickstart:~/Downloads
File Edit View
[FAILED MAP attempts...
cloudera@quickstart:...
cloudera@quickstart:...

```
scala>  val sports_data = sc.textFile("/tmp/sports_data.txt")
sports_data: org.apache.spark.rdd.RDD[String] = /tmp/sports_data.txt MapPartitionsRDD[155] at textFile at <console>:37

scala>  val rows = sports_data.map(line => line.split(",")).map(line => (line(0).toString,line(1).toString,line(2).toString,line(3).toString, l
ine(4).toString,line(5).toString,line(6).toString))
rows: org.apache.spark.rdd.RDD[(String, String, String, String, String, String, String)] = MapPartitionsRDD[157] at map at <console>:39

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

scala> val filter_data = rows.filter(x => x != header)
filter_data: org.apache.spark.rdd.RDD[(String, String, String, String, String, String, String)] = MapPartitionsRDD[158] at filter at <console>:
43

scala> val sportsDF = filter_data.toDF("firstname","lastname","sports","medal_type","age","year","country")
sportsDF: org.apache.spark.sql.DataFrame = [firstname: string, lastname: string, sports: string, medal_type: string, age: string, year: string,
 country: string]

scala> sportsDF.registerTempTable("sports")

scala> sportsDF.printSchema()
root
 |-- firstname: string (nullable = true)
 |-- lastname: string (nullable = true)
 |-- sports: string (nullable = true)
 |-- medal_type: string (nullable = true)
 |-- age: string (nullable = true)
 |-- year: string (nullable = true)
 |-- country: string (nullable = true)
```

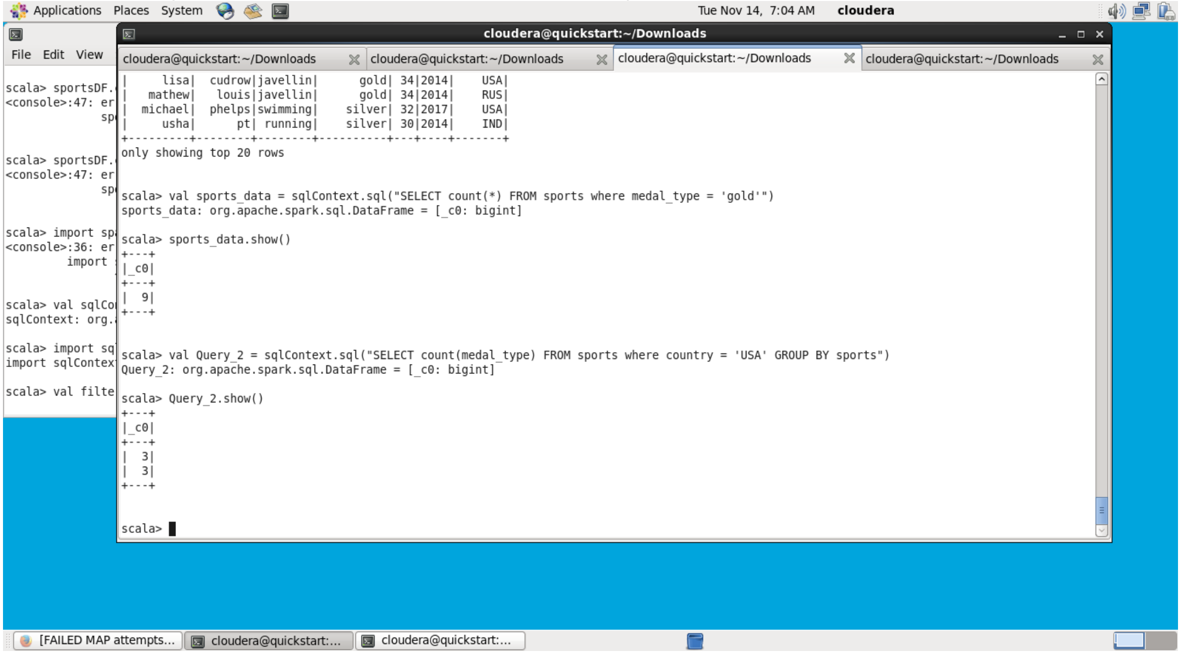

Applications Places System
Tue Nov 14, 6:54 AM cloudera
cloudera@quickstart:~/Downloads
cloudera@quickstart:~/Downloads
cloudera@quickstart:~/Downloads
cloudera@quickstart:~/Downloads
cloudera@quickstart:~/Downloads
File Edit View
[FAILED MAP attempts...
cloudera@quickstart:...
cloudera@quickstart:...

```
scala> val sports_data = sqlContext.sql("SELECT * FROM sports")
sports_data: org.apache.spark.sql.DataFrame = [firstname: string, lastname: string, sports: string, medal_type: string, age: string, year: stri
ng, country: string]

scala> sports_data.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
```


Applications Places System
Tue Nov 14, 6:54 AM cloudera
cloudera@quickstart:~/Downloads
File Edit View
[FAILED MAP attempts...
cloudera@quickstart:...
cloudera@quickstart:...

```
|    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> val sports_data = sqlContext.sql("SELECT count(*) FROM sports where medal_type = 'gold'")
sports_data: org.apache.spark.sql.DataFrame = [_c0: bigint]

scala> sports_data.show()
+---+
|_c0|
+---+
|  9|
+---+


scala> 
```

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


Applications Places System
Tue Nov 14, 7:04 AM cloudera
cloudera@quickstart:~/Downloads
File Edit View
[FAILED MAP attempts...
cloudera@quickstart:...
cloudera@quickstart:...

```
|      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> val sports_data = sqlContext.sql("SELECT count(*) FROM sports where medal_type = 'gold'")
sports_data: org.apache.spark.sql.DataFrame = [_c0: bigint]

scala> sports_data.show()
+---+
|_c0|
+---+
|  9|
+---+


scala> val Query_2 = sqlContext.sql("SELECT count(medal_type) FROM sports where country = 'USA' GROUP BY sports")
Query_2: org.apache.spark.sql.DataFrame = [_c0: bigint]

scala> Query_2.show()
+---+
|_c0|
+---+
|  3|
|  3|
+---+


scala> 
```