Session 19

Assignment 1

|  |  |
| --- | --- |
| **Prepared For:** | AcadGild |
|  |  |
| **Document Approval:** | **AcadGild** |
|  |  |
|  |  |
|  |  |
|  |  |
| **Project Title:** | Session 19 – Assignment 1 |
|  |  |
| **Prepared By:** | Duncan Burgess |
|  |  |
|  | dburgess@duncb.com |
|  |  |
| **Primary Engineer:** | Duncan Burgess |
|  |  |
| **Document Reference:** | **Session 19 – Assignment 1** |
|  |  |
| **Start Date:** | 18/10/2017 |
|  |  |
|  |  |



# 

# Contents

[Contents 2](#_Toc496297595)

[Change History 3](#_Toc496297596)

[1. Problem Statement 4](#_Toc496297597)

[2. Datasets 4](#_Toc496297598)

[3. Solution 5](#_Toc496297599)

[4. Results 6](#_Toc496297600)

# Change History

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Document Revision** | **Date** | **Authored By** | **Authorised By** | **Sections Affected** | **Reason for Change** |
| Rev 01 | 19/10/2017 | Duncan Burgess |  | All | Initial release. |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

# Problem Statement

* What is the total number of gold medal winners every year
* How many silver medals have been won by USA in each sport Using spark-sql

Using the datasets below:

# Datasets

**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

# Solution

I have deliberately commented and printed out all steps for enhanced learning

**Code written**

**package** com.duncb.spark

**import** org.apache.spark.\_

**import** org.apache.spark.SparkContext.\_

**import** org.apache.spark.sql.\_

**import** org.apache.log4j.\_

**object** SparkSQL {

**case** **class** **Sports**(firstname:*String*, lastname:*String*, sports:*String*,medal\_type:*String*,age:Int,year:Int,country:*String*)

**def** mapper(line:*String*): **Sports** = {

**val** fields = line.split(',')

**val** Osports:**Sports** = **Sports**(fields(0).toString, fields(1).toString, fields(2).toString,fields(3).toString, fields(4).toInt,fields(5).toInt,fields(6).toString)

//firstname,lastname,sports,medal\_type,age,year,country

**return** Osports

}

/\*\* Our main function where all the action happens \*/

**def** main(args: Array[*String*]) {

// Set the log level to only print errors

Logger.getLogger("org").setLevel(Level.ERROR)

// Use new SparkSession interface in Spark 2.0

**val** spark = SparkSession

.builder

.appName("SparkSQL")

.master("local[1]")

.config("spark.sql.warehouse.dir", "file:///C:/Temp") // Necessary to work around a Windows bug in Spark 2.0.0; omit if you're not on Windows.

.getOrCreate()

**val** lines = spark.sparkContext.textFile("file:///N:/Datasets/Sports\_data.txt")

//Filter to remove the header from the textfile

**val** skipable\_first\_row = lines.first()

**val** Osports1 = lines.filter(row => row != skipable\_first\_row)

**val** Osports = Osports1.map(mapper)

// Infer the schema, and register the DataSet as a table.

**import** spark.implicits.\_

**val** schemasports = Osports.toDS

println(" The Schema created")

schemasports.printSchema()

schemasports.createOrReplaceTempView("Osports")

println("--------------------------------------------------------")

// SQL can be run over DataFrames that have been registered as a table

// Now can generate a value using spark.sql("SELECT \* FROM Osports using any SQL query)

**val** res1 = spark.sql("SELECT \* FROM Osports")

**val** res2 =res1.collect()

res2.foreach(println)

println("--------------------------------------------------------")

// query to find gold medals for each year

println(" assignment 19.1 problem 1")

**val** GMPY =spark.sql("SELECT COUNT(\*) AS GMY,year FROM Osports WHERE medal\_type='gold' GROUP BY year ORDER BY year")

println("--------------------------------------------------------")

GMPY.show

println(" Number of Golds per yearr")

println(" assignment 19.1 problem 2")

**val** SMPY =spark.sql("SELECT sports,year,COUNT(\*) AS SMY FROM Osports WHERE country='USA' and medal\_type='silver' GROUP BY sports,year")

println("--------------------------------------------------------")

SMPY.show

//Added per year as it looked more informative

println(" Number of Silvers won by USA per year Total is 3")

spark.stop()

}

}

# Results

Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties

The Schema created

root

|-- firstname: string (nullable = true)

|-- lastname: string (nullable = true)

|-- sports: string (nullable = true)

|-- medal\_type: string (nullable = true)

|-- age: integer (nullable = false)

|-- year: integer (nullable = false)

|-- country: string (nullable = true)

--------------------------------------------------------

[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]

--------------------------------------------------------

assignment 19.1 problem 1

--------------------------------------------------------

[Stage 3:=======> (25 + 1) / 200]

[Stage 3:=================> (64 + 1) / 200]

[Stage 3:=====================================> (137 + 1) / 200]

[Stage 3:==============================================> (169 + 1) / 200]

[Stage 3:======================================================>(197 + 1) / 200]

+---+----+

|GMY|year|

+---+----+

| 3|2014|

| 3|2015|

| 2|2016|

| 1|2017|

+---+----+

Number of Golds per yearr

assignment 19.1 problem 2

--------------------------------------------------------

[Stage 11:==============================================> (85 + 1) / 100]

+--------+----+---+

| sports|year|SMY|

+--------+----+---+

|swimming|2016| 1|

|swimming|2017| 2|

+--------+----+---+

Number of Silvers won by USA per year Total is 3