*Spark*

**A C A D G I L D**

Session 19: Spark SQL

Assignment 1

**Page 1**

*Spark*

**A C A D G I L D**

*Session 19: Spark SQL*

*Assignment 1*

**Table of Contents**

1. Introduction .......................................................................................................................................... 3

2. Objective ............................................................................................................................................... 3

3. Prerequisites ......................................................................................................................................... 3

4. Associated Data Files ............................................................................................................................ 3

5. Problem Statement ............................................................................................................................... 3

6. Approximate Time to Complete Task ................................................................................................... 3

**Page 2**

*Spark*

**A C A D G I L D**

**1. Introduction**

In this assignment, you have to implement the concepts of the session.

**2. Objective**

This assignment will help you to understand the concepts learnt in the session.

**3. Prerequisites**

None

**4. Associated Data Files**

Dataset link

**5. 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

**6. Approximate Time to Complete Task**

**A.load file into RDD**



scala> val colNames = SportRDD.first()

**B.pick the Header**



colNames: String = firstname,lastname,sports,medal\_type,age,year,country

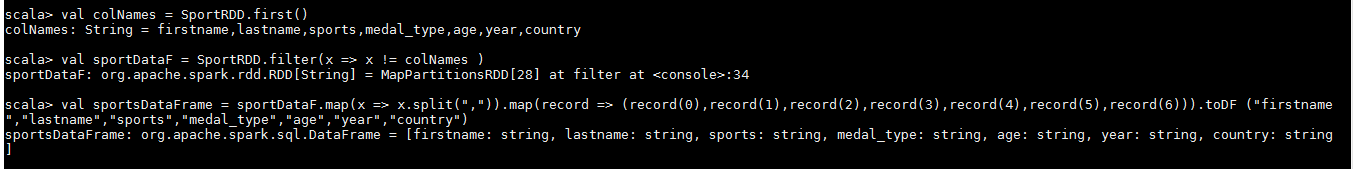
scala> val sportDataF = SportRDD.filter(x => x != colNames )

sportDataF: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[28] at filter at <console>:34

scala> val sportsDataFrame = sportDataF.map(x => x.split(",")).map(record => (record(0),record(1),record(2),record(3),record(4),record(5),record(6))).toDF ("firstname ","lastname","sports","medal\_type","age","year","country")

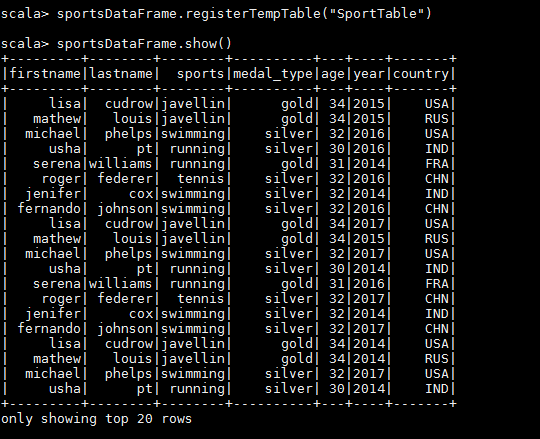
sportsDataFrame: org.apache.spark.sql.DataFrame = [firstname: string, lastname: string, sports: string, medal\_type: string, age: string, year: string, country: string ]

**C:create DATAFRAME**



**D:Register the Table**

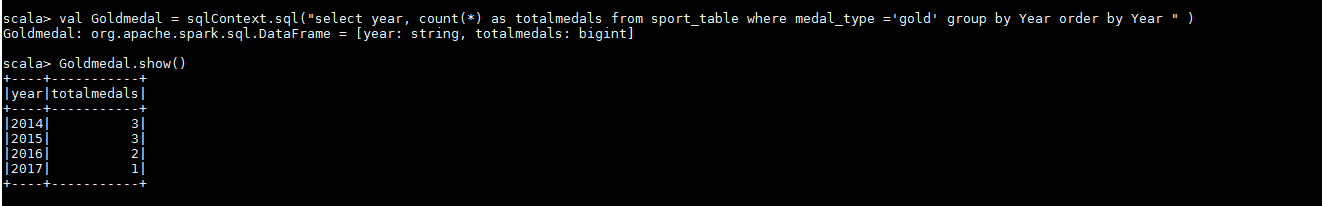
scala> sportsDataFrame.registerTempTable("SportTable")



scala> sportsDataFrame.show()

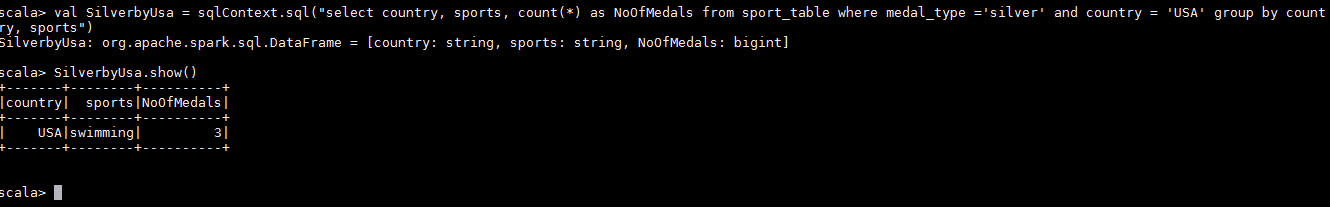
**1)What are the total number of gold medal winners every year()**

**Ans:** **scala> val Goldmedal = sqlContext.sql("select year, count(\*) as totalmedals from sport\_table where medal\_type ='gold' group by Year order by Year " )**



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

**Ans: val SilverbyUsa = sqlContext.sql("select country, sports, count(\*) as NoOfMedals from sport\_table where medal\_type ='silver' and country = 'USA' group by country, sports")**

****

****