*Spark*

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

Session 19: Spark SQL

Assignment 2

**Page 1**

*Spark*

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

*Session 19: Spark SQL*

*Assignment 2*

**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 .................................................... **Error! Bookmark not defined.**

**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**

**Link**

**5. Problem Statement**

Using udfs on dataframe

1. Change firstname, lastname columns into

Mr.first\_two\_letters\_of\_firstname<space>lastname

for example - michael, phelps becomes Mr.mi phelps

2. Add a new column called ranking using udfs on dataframe, where :

gold medalist, with age >= 32 are ranked as pro

gold medalists, with age <= 31 are ranked amateur

silver medalist, with age >= 32 are ranked as expert

silver medalists, with age <= 31 are ranked rookie

**6. Expected output**

Query with screenshots

**Page 3**

**A.load file into RDD**



scala> val colNames = SportRDD.first()

**B.pick the Header**



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

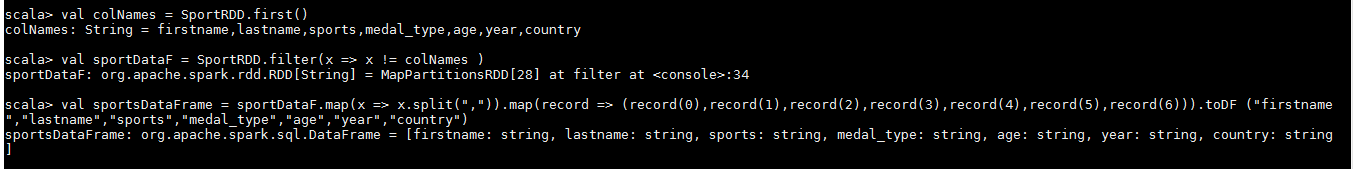
scala> val sportDataF = SportRDD.filter(x => x != colNames )

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")

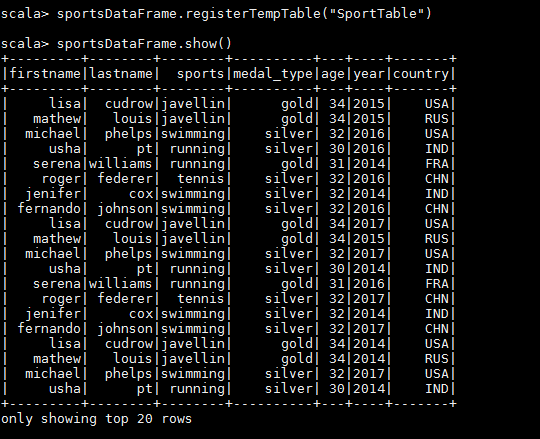
1. Change firstname, lastname columns into

Mr.first\_two\_letters\_of\_firstname<space>lastname

****

**def ChangName = udf((firstname: String, lastname: String) =>{"Mr."+firstname+" " +lastname )})**

**val SportsChangName = sportsDataFrame.withcolumn("name",ChangeName(sportsDataFrame("firstname").substr(1,2),sportsDataFrame("lastname"))).drop(sportsDataFrame("firstname")).drop(sportsDataFrame("lastname"))**



scala> sportsDataFrame.show()



2. Add a new column called ranking using udfs on dataframe, where :

gold medalist, with age >= 32 are ranked as pro

gold medalists, with age <= 31 are ranked amateur

silver medalist, with age >= 32 are ranked as expert

silver medalists, with age <= 31 are ranked rookie

