Session 18 : SPARK SQL

Assignment 2

**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

Use the dataset given below:



**Solution:-**

Starting Spark shell

Commands –

export SPARK\_HOME=/usr/local/spark/spark-2.2.0-bin-hadoop2.7

spark-shell

(using latest version of spark for running spark SQL)



Loading the data:-

**scala>** val sportsRDD = sc.textFile("Sports\_data.txt");

**scala>** val header = sportsRDD.first() ;

**scala>** val sportsRDDFiltered = sportsRDD.filter(row => row != header);



Loading the data into DataFrame

**Scala>** val sportsDF= sportsRDDFiltered.map(lines=>lines.split(",")).map(arrays => (arrays(0),arrays(1),arrays(2),arrays(3),arrays(4),arrays(5),arrays(6))).toDF("firstname","lastname","sports","medal\_type","age","year","country");



Based on DataFrame “**sportsDF**” we can find solution to the problem statements

**1. Change firstname, lastname columns into**

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

**for example - michael, phelps becomes Mr.mi phelps**

**Creating UDF :-**

**scala>** val convertName = udf((fname: String, lname: String) => { "Mr."+fname.slice(0,2)+" "+lname+""});

Adding column “**Fullname**” to Data Frame using UDF:-

**scala>** val problem1DF = sportsDF.withColumn("Fullname", convertName($"firstname", $"lastname"));

**Removing columns “firstname” and “lastname” from DataFrame problem1DF**

**scala>** val problem1ResultDF = problem1DF.drop("firstname").drop("lastname");

**scala>**problem1ResultDF.show(50);





**Required output is:-**



**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**

**Creating UDF:-**

**scala>** val playerClass = udf((medal:String, age: String) => { if(age.toInt <= 31 && medal=="silver") { "rookie"; } else { if(age.toInt >= 32 && medal=="silver") { "expert"; } else { if(age.toInt <= 31 && medal=="gold"){"amateur";} else{ if(age.toInt >= 32 && medal=="gold") {"pro";} else {"Could not determine class";}} }}});

Adding new column “**Class**” to DataFrame **problem2DF**

**scala>** val problem2DF = sportsDF.withColumn("Class", playerClass($"medal\_type", $"age"));

**scala>** problem2DF.show(50);



**Required output is:-**

