**Session 19: SPARK SQL**

Assignment 19.2

**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.miphelps

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***

**Dataset**

****

Task – 1 - Change firstname, lastname columns

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

***For example - michael, phelps becomes Mr.miphelps***

Please see the codes used below,

***1. valSportsData = sc.textFile("/home/acadgild/hadoop/Sports\_data.txt")***

***2. valschemaString = "firstname:string,lastname:string,sports:string,medal\_type:string,age:string,year:string,country:string"***

***3. val schema = StructType(schemaString.split(",").map(x => StructField(x.split(":")(0),if(x.split(":")(1).equals("string"))StringType else IntegerType, true)))***

***4. valrowRDD = SportsData.map(\_.split(",")).map(r => Row(r(0), r(1), r(2), r(3), r(4), r(5), r(6)))***

***5. valSportsDataDF = spark.createDataFrame(rowRDD, schema)***

***6. SportsDataDF.createOrReplaceTempView("Sports\_Data")***

***7. val Name = udf((firstname:String, lastname:String)=>"Mr. ".concat(firstname.substring(0,2)).concat(" ")concat(lastname))***

***8. spark.udf.register("Full\_Name", Name)***

***9. valfname = spark.sql("SELECT Full\_Name(firstname, lastname) FROM SportsData").show()***

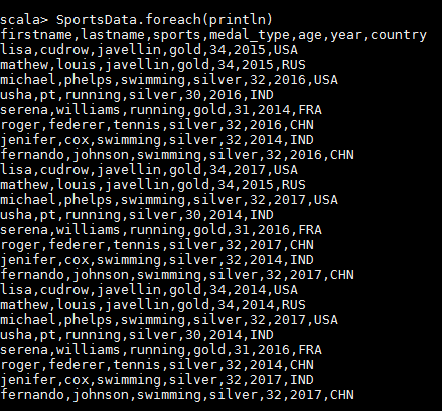
We will proceed with the tasks,

In order to proceed we need to import some dependencies as shown below,

* **import org.apache.spark.sql.Row;**
* **import org.apache.spark.sql.types.{StructType,StructField,StringType,NumericType,IntegerType};**
* **import org.apache.spark.sql.functions.udf**

****

**Step -1** – we are creating a RDD from Input DataSet,



**Step -2** – we are defining a schema since it is a text file and splitting the input file using the delimiters and extracting the rows from it.





We have created the **dataframe**by passing the RDD which reads the file and schema to spark session object-

The schema of the created **Dataframe**can be seen below.



**Step – 3** - Here we are defining the UDF which will take 2 strings (columns) as input and will concatenate them with Mr. appended in it and now we need to register the UDF. Here we doing the same and giving it an alias as **Full\_Name.**

Finally we can apply this UDF on the columns to give the required result-

Output

