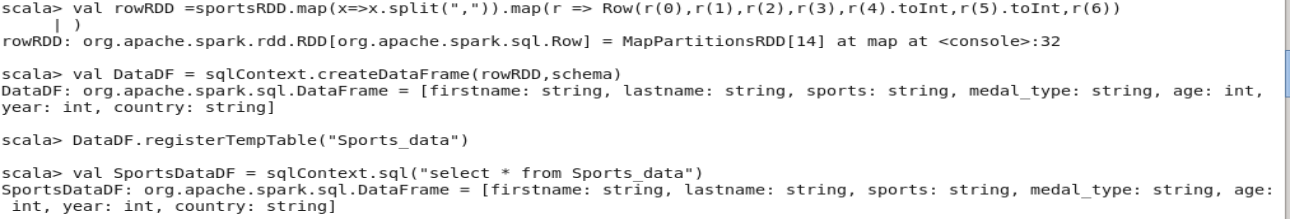
**Assignment 02:**

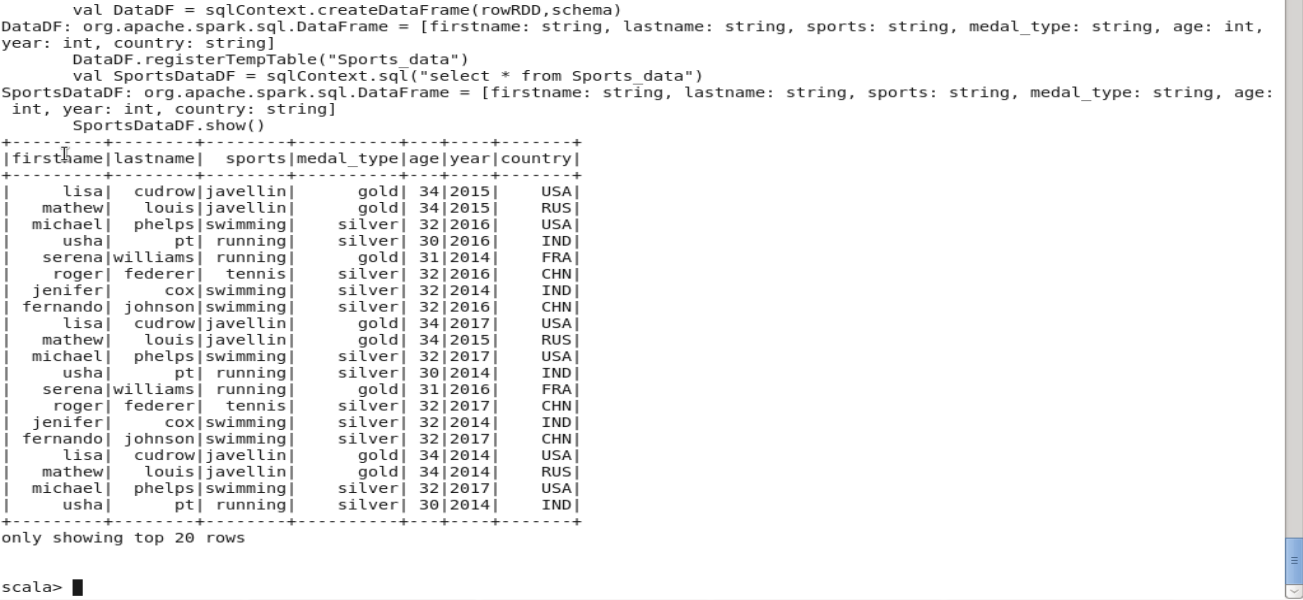
**Data Frame Creation:**

1. Import the spark sql Row, DataFrame and types objects.
2. Create an RDD that will hold the contents of the Sports Dataset(text file format) sportsRDD
3. Create the schema (column name and type) for the data. schemaString
4. Create the schema structure based on the schemaString. schema
5. Create an RDD that will split the data in sportsRDD by ‘,’ and map columns in rows as in schema. rowRDD
6. Create the DataFrame from the rowRDD and schema created before using the “sqlContext” object. SportsDataDF
7. Create a temporary table (Sports\_data) from the DataFrame SportsDataDF





To view the contents of the temp table Sports\_Data, use a sql SELECT Statement with the sqlContext



**Prob01: Change firstname and lastname columns into:**

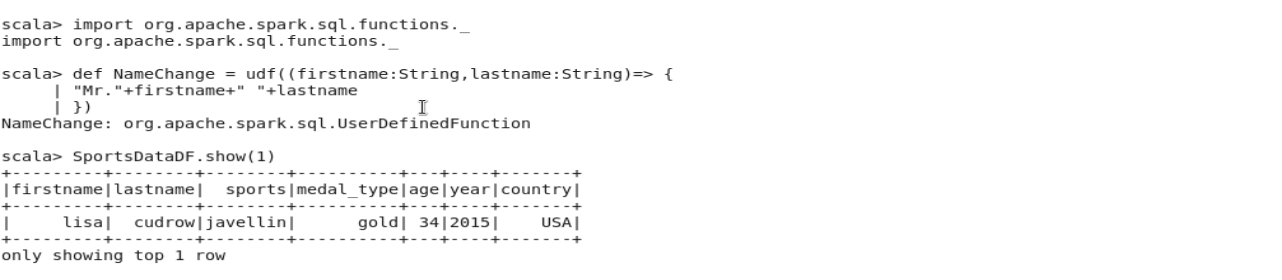
**Mr.first\_two\_letters\_of\_firstname<space>lastname**

Import the spark sql functions object.

Create a Spark SQL UDF that takes the **firstname** and **lastname** as parameters and performs the change as required in the problem statement. ChangeName

Using the DataFrame SportsDataDF created above with the withColumn function, we:

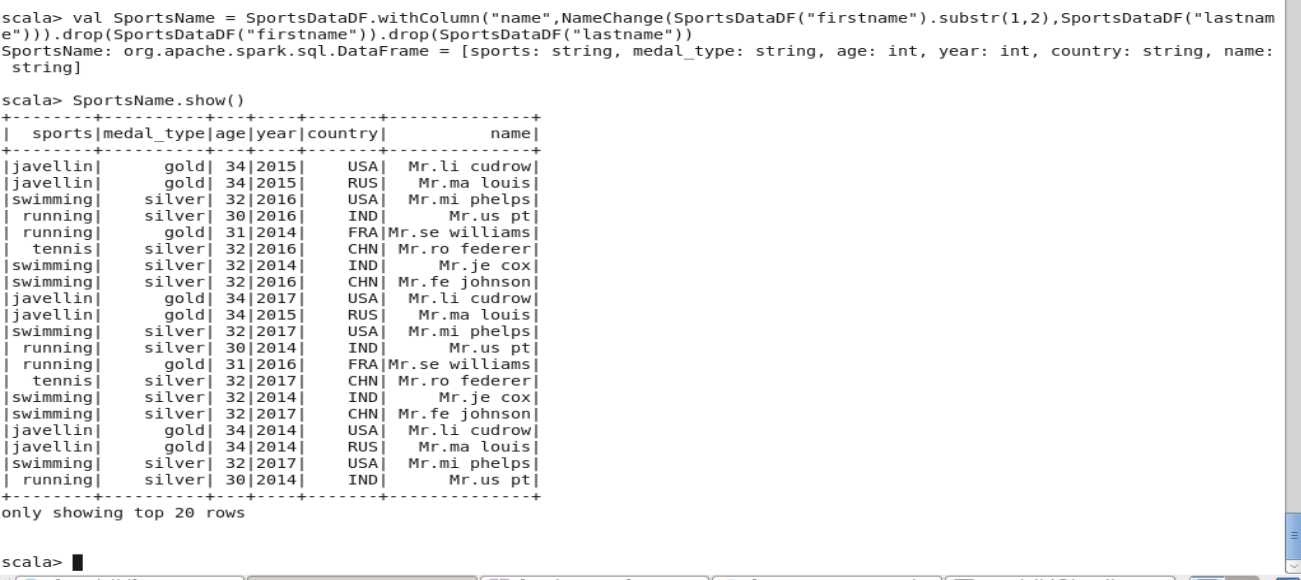
Assign the name for the new column (**name**) *and*



Call the UDF NameChange with the 1st argument as the 1st two letters of firstname and the 2nd argument as the lastname

Drop both existing columns firstname and lastname (as the new column has been created)

SportsChangeName

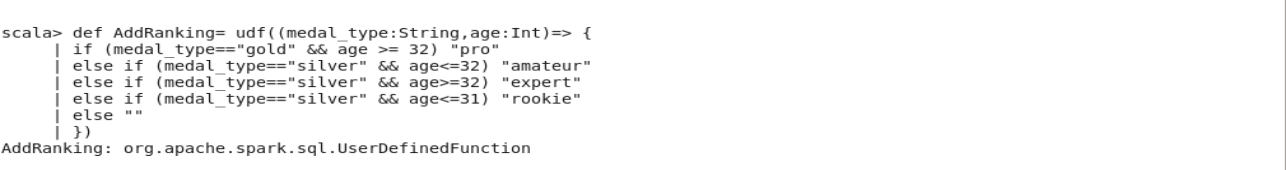


**Prob 02: Add a new column called “ranking” using UDFs on DataFrame, where:** gold medalist, with age >= 32 are ranked as pro, gold medalists, with age <= 31 are ranked amateur, silver medalist, with age >= 32 are ranked as expert and silver medalists, with age <= 31 are ranked rookie

Import the spark sql functions object.

Create a Spark SQL UDF that takes the **medal\_type** and **age** as parameters and performs the change as required in the problem statement. AddRanking

Using the DataFrame SportsDataDF created above with the withColumn function, we:



Assign the name for the new column (**ranking**) *and*

Call the UDF AddRanking with the 1st argument as medal\_type and the 2nd argument as the age

SportsDataFinal

