


Session 19 Assignment 2

Problem 1 – To read the data from input file and concatenate the first two characters of firstname and lastname using UDF

 bigdata@localhost:~/spark-2.3.0-bin-hadoop2.7/bin

[bigdata@localhost bin]\$./spark-shell -i /home/bigdata/deepak/docs/Acadgild/Session19Assignment2/Session19Assignment2_1.scala 

```
2018-05-31 02:37:42 WARN ObjectStore:568 - Failed to get database global_temp, returning NoSuchObjectException
sportsdata: org.apache.spark.sql.DataFrame = [firstname: string, lastname: string ... 5 more fields]
header: org.apache.spark.sql.Row = [firstname,lastname,sports,medal_type,age,year,country]
sportsdataDF: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [firstname: string, lastname: string ...
concatFirstNameLastName: (fname: String, lname: String)String
output: org.apache.spark.sql.expressions.UserDefinedFunction = UserDefinedFunction(<function2>,StringType,Some(I
result: org.apache.spark.sql.DataFrame = [firstname: string, lastname: string ... 1 more field]
+-----+-----+-----+
|firstname|lastname|UDF(firstname, lastname)|
+-----+-----+-----+
| lisa| cudrow| Mr.li cudrow|
| mathew| louis| Mr.ma louis|
| michael| phelps| Mr.mi phelps|
| usha| pt| Mr.us pt|
| serena|williams| Mr.se williams|
| roger| federer| Mr.ro federer|
| jenifer| cox| Mr.je cox|
| fernando| johnson| Mr.fe johnson|
| lisa| cudrow| Mr.li cudrow|
| mathew| louis| Mr.ma louis|
| michael| phelps| Mr.mi phelps|
| usha| pt| Mr.us pt|
| serena|williams| Mr.se williams|
| roger| federer| Mr.ro federer|
| jenifer| cox| Mr.je cox|
| fernando| johnson| Mr.fe johnson|
| lisa| cudrow| Mr.li cudrow|
| mathew| louis| Mr.ma louis|
| michael| phelps| Mr.mi phelps|
| usha| pt| Mr.us pt|
+-----+-----+-----+
only showing top 20 rows
```

Code Snapshot

```
/* Session19 Assignment 2 - UDFs on Dataframe */

//Problem 1 - Use UDF to concatenate first two characters of firstname and lastname
import spark.implicits._

//Read the input file
val sportsdata=spark.read.csv("/home/bigdata/deepak/docs/Acadgild/Session19Assignment2/Sports_data.txt").toDF(
("firstname","lastname","sports","medal_type","age","year","country")

//Get the header and skip it
val header = sportsdata.first()
val sportsdataDF = sportsdata.filter(line => line != header)

//UDF to perform the concatenation
def concatFirstNameLastName(fname: String, lname: String) = "Mr."+fname.substring(0,2) +" "+ lname

//Register the UDF
val output = udf(concatFirstNameLastName(_:String, _:String))

//Call the UDF to display the result
val result=sportsdataDF.select($"firstname",$"lastname",output(sportsdataDF("firstname"),sportsdataDF(
("lastname"))))
result.show
```