Vijay's Assignment – Spark SQL 2

Task 1

Common areas of the program to declare the SparkConf and SparkSession

```
package vksp1

import org.apache.spark.SparkConf
import org.apache.spark.SparkContext
import org.apache.spark.rdd.RDD.rddToPairRDDFunctions
import org.apache.spark.sql.SparkSession

import org.apache.spark.sql.types._
object vksprksql2{

  case class sports(fname: String, lname: String, Sprts: String, medal_type:
String, age: String, year: String,country: String)

  def main(args: Array[String]) {

    val conf = new SparkConf()
      .setAppName("vjsprksql2")
      .setMaster("local")

    val sc = new SparkContext(conf)

    val spark = SparkSession
      .builder()
      .appName("Spark SQL basic example")
      .config("spark.some.config.option", "some-value")
      .getOrCreate()

    import spark.implicits._

     val sportsDF = spark.sparkContext
      .textFile("C:\\Users\\VIJAYLAKSHMANAN\\spark\\sports.txt")
      .map(_.split(","))
      .map(attributes => sports(attributes(0), attributes(1), attributes(2),
attributes(3), attributes(4), attributes(5), attributes(6)))
      .toDF()

    sportsDF.createOrReplaceTempView("sports")
```

Using spark-sql, Find:

1. What are the total number of gold medal winners every year

```
val query1DF = spark.sql("SELECT year, count(*) as Cnt FROM sports group by year")
    query1DF.show()
```

Output:

```
19/02/20 18:37:39 INFO DAGScheduler: Job 4 finished: show at vjsprksql2.scala:38, took 0.479868 s
+----+---+
|year|Cnt|
+----+---+
|2016|  6|
|2017|  7|
|2014|  8|
|year|  1|
|2015|  3|
+----+---+

19/02/20 18:37:39 INFO SparkContext: Invoking stop() from shutdown hook
```

2. How many silver medals have been won by USA in each sport

```
val query2DF = spark.sql("SELECT country, Sprts,count(*) from sports where country
= 'USA' and medal_type = 'silver' group by country, Sprts")
    query2DF.show()
```

```
19/02/20 18:38:43 INFO DAGScheduler: Job 4 finished: show at vjsprksql2.scala:41, took 0.617728 s
+-------+--------+--------+
|country|   Sprts|count(1)|
+-------+--------+--------+
|    USA|swimming|       3|
+-------+--------+--------+
```

Task 2

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

```
val udf1 = udf((fname1: String, lname1:
String)=>"Mr.".concat(fname1.substring(0,2)).concat(" ")concat(lname1))
    spark.udf.register("udf_vijay", udf1)

    val query3DF = spark.sql("SELECT udf_vijay(fname,lname) from sports where
fname <> 'firstname'")
    query3DF.show()
```

```
19/02/21 11:16:21 INFO DAGScheduler: Job 0 finished: show at vjsprksql2.scala:48, took 0.340272 s
+-----------------+
|UDF(fname, lname)|
+-----------------+
|     Mr.li cudrow|
|      Mr.ma louis|
|     Mr.mi phelps|
|         Mr.us pt|
|   Mr.se williams|
|    Mr.ro federer|
|        Mr.je cox|
|    Mr.fe johnson|
|     Mr.li cudrow|
|      Mr.ma louis|
|     Mr.mi phelps|
|         Mr.us pt|
|   Mr.se williams|
|    Mr.ro federer|
|        Mr.je cox|
|    Mr.fe johnson|
|     Mr.li cudrow|
|      Mr.ma louis|
|     Mr.mi phelps|
|         Mr.us pt|
+-----------------+
only showing top 20 rows
```

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

```
val udf2 = udf((med:String, age1:String)=>(med,age1) match
        { case (med,age1) if med =="gold"&&age1.toInt>=32=>"Pro"
          case (med,age1) if med =="gold"&&age1.toInt<=31=>"Amateur"
          case (med,age1) if med =="silver"&&age1.toInt>=32=>"Expert"
            case (med,age1) if med =="silver"&&age1.toInt<=31=>"Rookie"
         })

    spark.udf.register("udf_vijay1", udf2)

    val query4DF = spark.sql("SELECT fname, lname, Sprts, medal_type, age, year,
country, udf_vijay1(medal_type,age) as Ranks from sports where fname <>
'firstname'")
    query4DF.show()
```

Output:

```
19/02/21 11:40:57 INFO DAGScheduler: Job 0 finished: show at vjsprksql2.scala:60, took 0.283250 s
+--------+--------+--------+----------+---+----+-------+-------+
|   fname|   lname|   Sprts|medal_type|age|year|country|  Ranks|
+--------+--------+--------+----------+---+----+-------+-------+
|    lisa|  cudrow|javellin|      gold| 34|2015|    USA|    Pro|
|  mathew|   louis|javellin|      gold| 34|2015|    RUS|    Pro|
| michael|  phelps|swimming|    silver| 32|2016|    USA| Expert|
|    usha|      pt| running|    silver| 30|2016|    IND| Rookie|
|  serena|williams| running|      gold| 31|2014|    FRA|Amateur|
|   roger| federer|  tennis|    silver| 32|2016|    CHN| Expert|
| jenifer|     cox|swimming|    silver| 32|2014|    IND| Expert|
|fernando| johnson|swimming|    silver| 32|2016|    CHN| Expert|
|    lisa|  cudrow|javellin|      gold| 34|2017|    USA|    Pro|
|  mathew|   louis|javellin|      gold| 34|2015|    RUS|    Pro|
| michael|  phelps|swimming|    silver| 32|2017|    USA| Expert|
|    usha|      pt| running|    silver| 30|2014|    IND| Rookie|
|  serena|williams| running|      gold| 31|2016|    FRA|Amateur|
|   roger| federer|  tennis|    silver| 32|2017|    CHN| Expert|
| jenifer|     cox|swimming|    silver| 32|2014|    IND| Expert|
|fernando| johnson|swimming|    silver| 32|2017|    CHN| Expert|
|    lisa|  cudrow|javellin|      gold| 34|2014|    USA|    Pro|
|  mathew|   louis|javellin|      gold| 34|2014|    RUS|    Pro|
| michael|  phelps|swimming|    silver| 32|2017|    USA| Expert|
|    usha|      pt| running|    silver| 30|2014|    IND| Rookie|
+--------+--------+--------+----------+---+----+-------+-------+
only showing top 20 rows
```