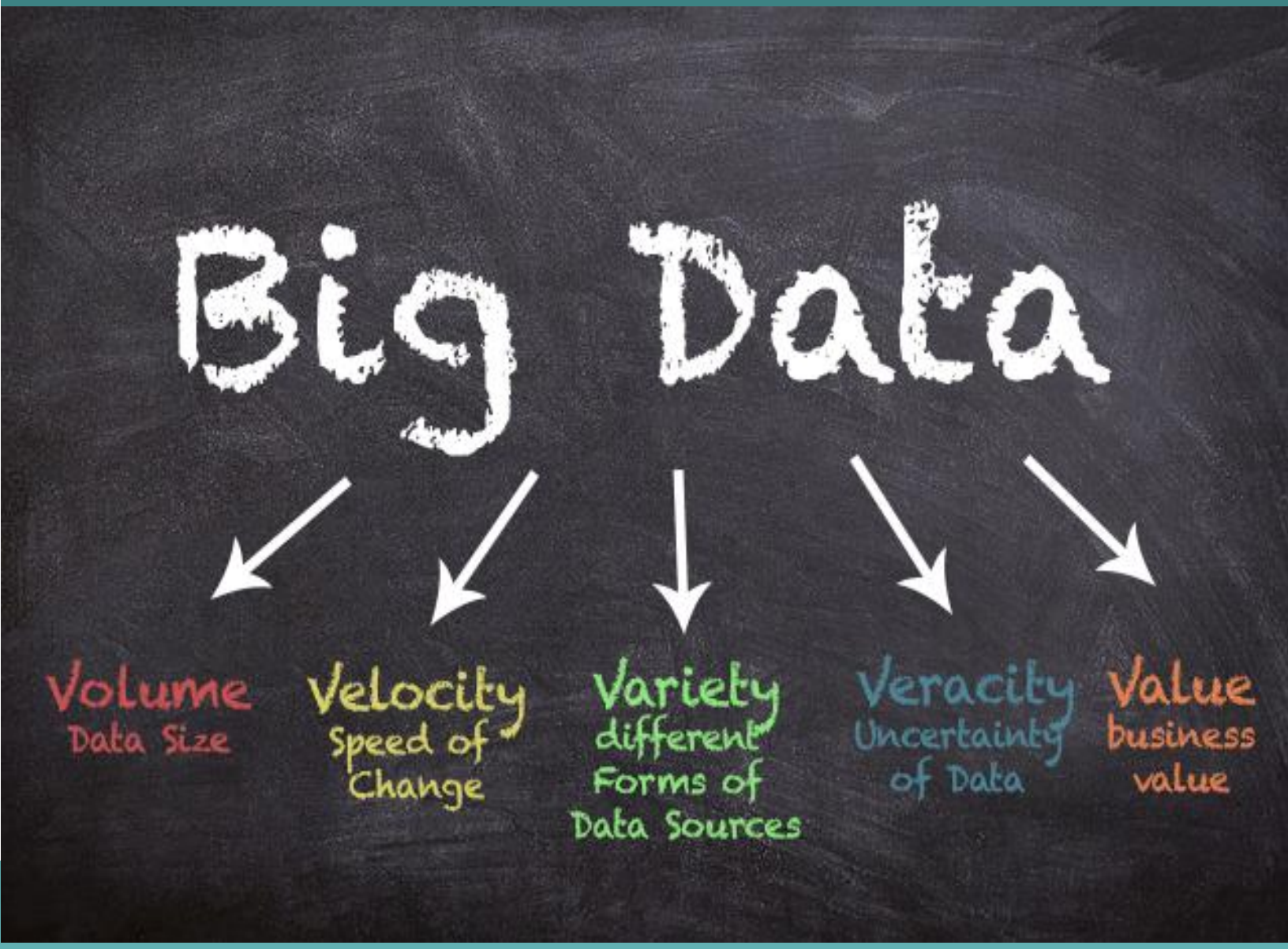
09/10/2018 | By: Ajit K Prasad



Big Data Engineering with Hadoop & Spark

Assignment on Spark SQL

Acadgild

Session 21: Assignment 21.1

This assignment is aimed at consolidating the concepts that was learnt during the Apache Spark SQL session of the course.

Associated Data Files

Datasets can be downloaded from this ***link***.

- Jump to the ***Source Code***.
- Jump to the ***output*** of the dataset.

Task 1:

Problem Statement:

Using Spark-SQL, find:

1. What are the total number of gold medal winners every year?
2. How many silver medals have been won by USA in each sport?

Task 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

Source Code

```
package SQL

import org.apache.spark.sql.SparkSession
import org.apache.spark.sql.functions.col
import org.apache.spark.sql.functions.udf

object SportsSQL {

  case class SportsDetails(FName: String, LName: String, Sports:
String, MedalType: String, sAge:Long, Year: Long, Nationality: String)

  def Ranking(Age: Int, Medal: String): String = {
    if(Medal == "gold" && Age >= 32) "Pro"
    else if(Medal == "gold" && Age <= 31) "Amateur"
    else if(Medal == "silver" && Age >= 32)"Expert"
    else if(Medal == "silver" && Age <= 31)"Rookie"
    else ""
  }

  def main(args: Array[String]): Unit = {
    val spark = SparkSession
      .builder()
      .master("local")
      .appName("Spark SQL basic example")
      .config("spark.some.config.option", "some-value")
      .getOrCreate()

    /***Set the log level as warning***/
    spark.sparkContext.setLogLevel("WARN")

    println("Spark Session Object Created for Sports Data Analysis")

    val sportsData = spark.sparkContext

.textFile("D:\\AcadGild\\ScalaCaseStudies\\Datasets\\Sports\\s21\\Spor
tsDataset.txt")

    val header = sportsData.first()
    val newData = sportsData.filter(row => row != header)

    println("\nHeader removed from the data!")

    /***For implicit conversions like converting RDDs and sequences to
DataFrames***/
    import spark.implicits._

    /***Loading Sports Dataset***/
    val sportsDF = newData
      .map(x => x.split(","))
      .map(x => SportsDetails(x(0), x(1), x(2), x(3), x(4).trim.toInt,
x(5).trim.toInt, x(6)))
```

```
      .toDF()
    print("\nSports dataframe displayed\n")
    sportsDF.show()

    /***What are the total number of gold medal winners every
year?***/
    println("\nList of Gold medal winners every year")
    sportsDF.filter(col("MedalType") === "gold")
      .show()

    val medal = sportsDF.filter(col("MedalType") === "gold")
    println("Total number of Gold medal winners every year: " +
medal.count())

    /***How many silver medals have been won by USA in each
sports?***/
    println("\n\nList of Silver medal winners from USA")
    sportsDF.filter((col("MedalType") === "silver") &&
col("Nationality") === "USA")
      .show()

    val silverMedals = sportsDF
      .filter((col("MedalType") === "silver") && col("Nationality")
=== "USA")
    println("Total number Silver medals won by USA in each sport: " +
silverMedals.count())


    /***Change firstname, lastname columns into
      Mr.first_two_letters_of_firstname<space>lastname
      for example - michael, phelps becomes Mr.mi phelps***/

    println("\n\nUsing udfs on dataframe")
    val Name = udf((FName: String, LName: String) => "Mr. "
      .concat(FName.substring(0, 2).concat(" ")
        .concat(LName)))

    println("\nDisplaying list by concatenating names")
    sportsDF.registerTempTable("SportsData")
    sportsDF.withColumn("FNameConcatLName", Name($"FName", $"LName"))
      .select("FNameConcatLName", "Sports", "MedalType", "sAge",
"Year", "Nationality")
      .show()

    /***Add a new column called Ranking using UDFs on dataframe
defined under Ranking function***/
    val Rank = udf(Ranking(_: Int, _: String))
    println("\nRanking the players accordingly")
    sportsDF.withColumn("Ranks", Rank($"sAge", $"MedalType"))
      .select("Ranks", "FName", "LName", "sAge", "MedalType")
      .show()
  }
}
```

Output of Dataset Used

SportsDataset.txt

SportsSQL

```
Header removed from the data!

Sports dataframe displayed
+--------+--------+--------+---------+----+----+-----------+
|   FName|   LName|  Sports|MedalType|sAge|Year|Nationality|
+--------+--------+--------+---------+----+----+-----------+
|    lisa|  cudrow|javellin|     gold|  34|2015|        USA|
|  mathew|   louis|javellin|     gold|  34|2015|        RUS|
| michael|  phelps|swimming|   silver|  32|2016|        USA|
|    usha|      pt| running|   silver|  30|2016|        IND|
|  serena|williams| running|     gold|  31|2014|        FRA|
|   roger| federer|  tennis|   silver|  32|2016|        CHN|
| jenifer|     cox|swimming|   silver|  32|2014|        IND|
|fernando| johnson|swimming|   silver|  32|2016|        CHN|
|    lisa|  cudrow|javellin|     gold|  34|2017|        USA|
|  mathew|   louis|javellin|     gold|  34|2015|        RUS|
| michael|  phelps|swimming|   silver|  32|2017|        USA|
|    usha|      pt| running|   silver|  30|2014|        IND|
|  serena|williams| running|     gold|  31|2016|        FRA|
|   roger| federer|  tennis|   silver|  32|2017|        CHN|
| jenifer|     cox|swimming|   silver|  32|2014|        IND|
|fernando| johnson|swimming|   silver|  32|2017|        CHN|
|    lisa|  cudrow|javellin|     gold|  34|2014|        USA|
|  mathew|   louis|javellin|     gold|  34|2014|        RUS|
| michael|  phelps|swimming|   silver|  32|2017|        USA|
|    usha|      pt| running|   silver|  30|2014|        IND|
+--------+--------+--------+---------+----+----+-----------+
only showing top 20 rows
```

Output of Task 1.1

SportsSQL

```
List of Gold medal winners every year
+------+--------+--------+---------+----+----+-----------+
| FName|   LName|  Sports|MedalType|sAge|Year|Nationality|
+------+--------+--------+---------+----+----+-----------+
|  lisa|  cudrow|javellin|     gold|  34|2015|        USA|
|mathew|   louis|javellin|     gold|  34|2015|        RUS|
|serena|williams| running|     gold|  31|2014|        FRA|
|  lisa|  cudrow|javellin|     gold|  34|2017|        USA|
|mathew|   louis|javellin|     gold|  34|2015|        RUS|
|serena|williams| running|     gold|  31|2016|        FRA|
|  lisa|  cudrow|javellin|     gold|  34|2014|        USA|
|mathew|   louis|javellin|     gold|  34|2014|        RUS|
|serena|williams| running|     gold|  31|2016|        FRA|
+------+--------+--------+---------+----+----+-----------+

Total number of Gold medal winners every year: 9
```

Output of Task 1.2

SportsSQL

```
List of Silver medal winners from USA
+-------+------+--------+---------+----+----+-----------+
|  FName| LName|  Sports|MedalType|sAge|Year|Nationality|
+-------+------+--------+---------+----+----+-----------+
|michael|phelps|swimming|   silver|  32|2016|        USA|
|michael|phelps|swimming|   silver|  32|2017|        USA|
|michael|phelps|swimming|   silver|  32|2017|        USA|
+-------+------+--------+---------+----+----+-----------+

Total number Silver medals won by USA in each sport: 3
```

Output of Task 2.1

SportsSQL

```
Using udfs on dataframe

Displaying list by concatenating names
+----------------+--------+---------+----+----+-----------+
|FNameConcatLName|  Sports|MedalType|sAge|Year|Nationality|
+----------------+--------+---------+----+----+-----------+
|   Mr. li cudrow|javellin|     gold|  34|2015|        USA|
|    Mr. ma louis|javellin|     gold|  34|2015|        RUS|
|   Mr. mi phelps|swimming|   silver|  32|2016|        USA|
|       Mr. us pt| running|   silver|  30|2016|        IND|
| Mr. se williams| running|     gold|  31|2014|        FRA|
|  Mr. ro federer|  tennis|   silver|  32|2016|        CHN|
|      Mr. je cox|swimming|   silver|  32|2014|        IND|
|  Mr. fe johnson|swimming|   silver|  32|2016|        CHN|
|   Mr. li cudrow|javellin|     gold|  34|2017|        USA|
|    Mr. ma louis|javellin|     gold|  34|2015|        RUS|
|   Mr. mi phelps|swimming|   silver|  32|2017|        USA|
|       Mr. us pt| running|   silver|  30|2014|        IND|
| Mr. se williams| running|     gold|  31|2016|        FRA|
|  Mr. ro federer|  tennis|   silver|  32|2017|        CHN|
|      Mr. je cox|swimming|   silver|  32|2014|        IND|
|  Mr. fe johnson|swimming|   silver|  32|2017|        CHN|
|   Mr. li cudrow|javellin|     gold|  34|2014|        USA|
|    Mr. ma louis|javellin|     gold|  34|2014|        RUS|
|   Mr. mi phelps|swimming|   silver|  32|2017|        USA|
|       Mr. us pt| running|   silver|  30|2014|        IND|
+----------------+--------+---------+----+----+-----------+
only showing top 20 rows
```

Output of Task 2.2

SportsSQL

```
Ranking the players accordingly
+-------+--------+--------+----+---------+
|  Ranks|   FName|   LName|sAge|MedalType|
+-------+--------+--------+----+---------+
|    Pro|    lisa|  cudrow|  34|     gold|
|    Pro|  mathew|   louis|  34|     gold|
| Expert| michael|  phelps|  32|   silver|
| Rookie|    usha|      pt|  30|   silver|
|Amateur|  serena|williams|  31|     gold|
| Expert|   roger| federer|  32|   silver|
| Expert| jenifer|     cox|  32|   silver|
| Expert|fernando| johnson|  32|   silver|
|    Pro|    lisa|  cudrow|  34|     gold|
|    Pro|  mathew|   louis|  34|     gold|
| Expert| michael|  phelps|  32|   silver|
| Rookie|    usha|      pt|  30|   silver|
|Amateur|  serena|williams|  31|     gold|
| Expert|   roger| federer|  32|   silver|
| Expert| jenifer|     cox|  32|   silver|
| Expert|fernando| johnson|  32|   silver|
|    Pro|    lisa|  cudrow|  34|     gold|
|    Pro|  mathew|   louis|  34|     gold|
| Expert| michael|  phelps|  32|   silver|
| Rookie|    usha|      pt|  30|   silver|
+-------+--------+--------+----+---------+
only showing top 20 rows


Process finished with exit code 0
```