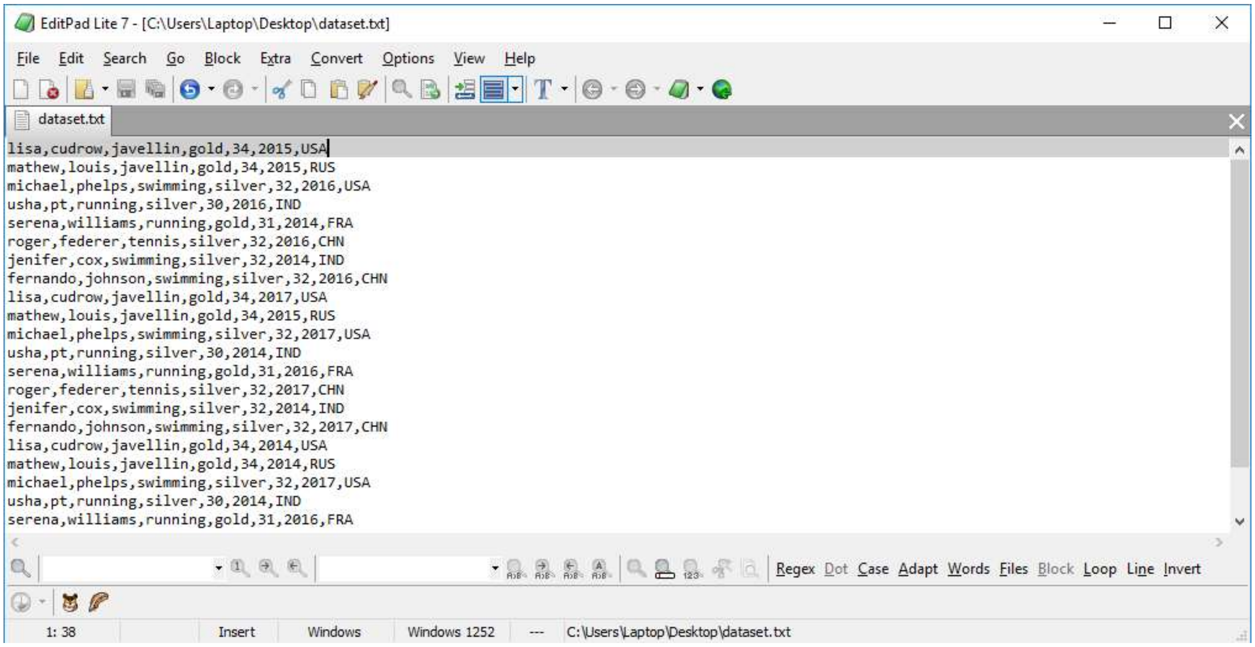
BIGDATA HADOOP & SPARK TRAINING

Assignment on Spark SQL II

Rashmi Krishna

Input file:

This input file is used to perform the below tasks


EditPad Lite 7 - [C:\Users\Laptop\Desktop\dataset.txt]
File Edit Search Go Block Extra Convert Options View Help
dataset.txt
Regex Dot Case Adapt Words Files Block Loop Line Invert
1: 38 Insert Windows Windows 1252 --- C:\Users\Laptop\Desktop\dataset.txt

```
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
serena,williams,running,gold,31,2016,FRA
```

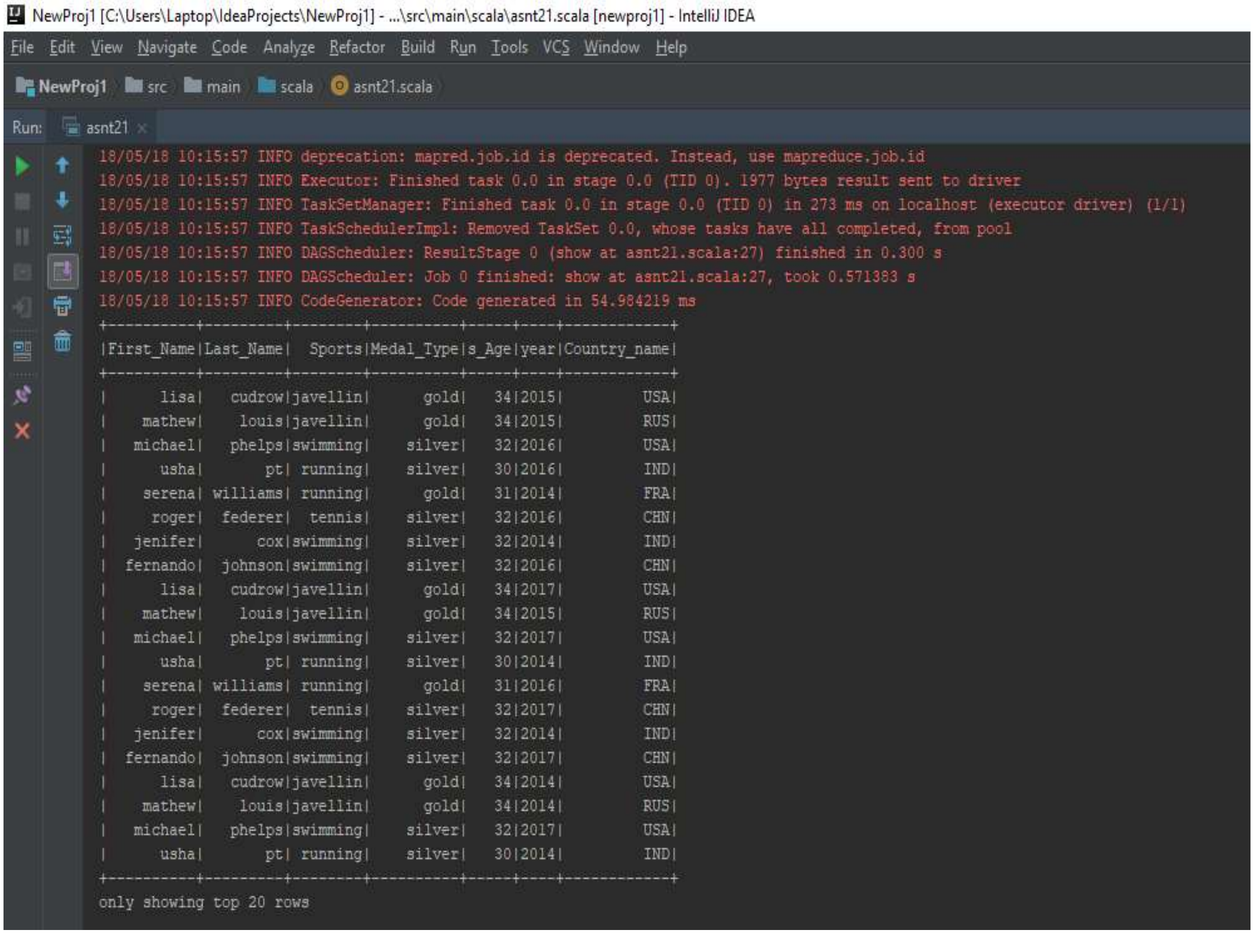
Below tasks are performed in **IntelliJ IDEA.**

Created a Spark-Scala Application and performed following steps to read the above mentioned text data to spark.

- Created case class for each text files which represents the schema for the respective text files.
- Created a Spark Session Object
- Created a spark context to read the files from the local file system to spark by matching the schema from case class

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

object asnt21 {
 case class Sports_details(First_Name: String, Last_Name: String,Sports:String,Medal_Type: String, s_Age:Long, year: Long,Country_name:String)

 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()
  println("Spark Session Object created")

  import spark.implicits._

  val SportsDFFromFile = spark.sparkContext
    .textFile("C:\\Users\\Laptop\\Desktop\\dataset.txt")
    .map(_.split(","))
    .map(attributes => Sports_details(attributes(0), attributes(1), attributes(2), 
attributes(3), attributes(4).trim.toInt, attributes(5).trim.toInt, attributes(6)))
    .toDF()
  SportsDFFromFile.show()
```


NewProj1 [C:\Users\Laptop\IdeaProjects\NewProj1] - ...\src\main\scala\asnt21.scala [newproj1] - IntelliJ IDEA
File
Edit
View
Navigate
Code
Analyze
Refactor
Build
Run
Tools
VCS
Window
Help
NewProj1
src
main
scala
asnt21.scala
Run:
asnt21

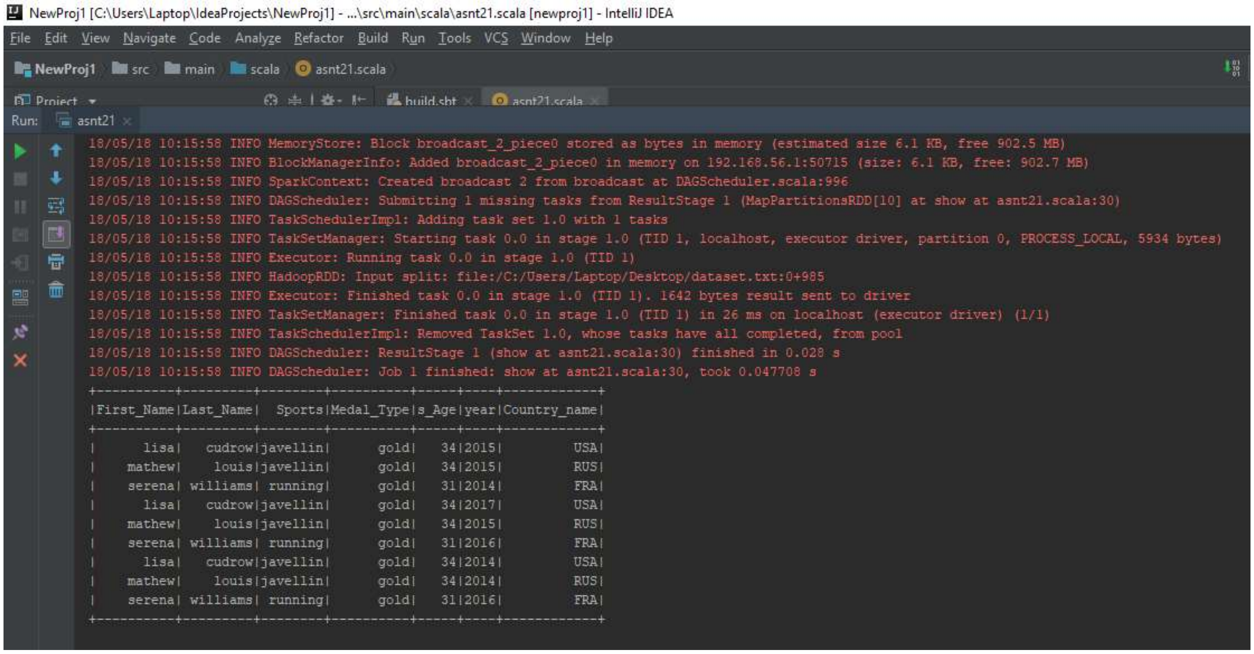
```
18/05/18 10:15:57 INFO deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
18/05/18 10:15:57 INFO Executor: Finished task 0.0 in stage 0.0 (TID 0). 1977 bytes result sent to driver
18/05/18 10:15:57 INFO TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 273 ms on localhost (executor driver) (1/1)
18/05/18 10:15:57 INFO TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool
18/05/18 10:15:57 INFO DAGScheduler: ResultStage 0 (show at asnt21.scala:27) finished in 0.300 s
18/05/18 10:15:57 INFO DAGScheduler: Job 0 finished: show at asnt21.scala:27, took 0.571383 s
18/05/18 10:15:57 INFO CodeGenerator: Code generated in 54.984219 ms
+----------+---------+--------+----------+-----+----+------------+
|First_Name|Last_Name|  Sports|Medal_Type|s_Age|year|Country_name|
+----------+---------+--------+----------+-----+----+------------+
|      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
```

Task 1

Using spark-sql, Find: What are the total number of gold medal winners every year

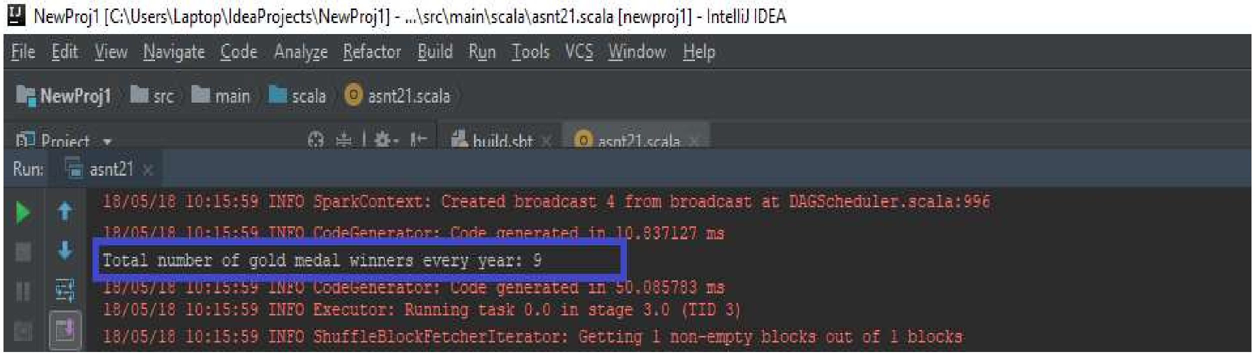
To accomplish this task, we just filter the medal_type column with “gold” and count them

```
SportsDFFromFile.filter(col("Medal_type") === "gold").show()
 val medal = SportsDFFromFile.filter(col("Medal_type") === "gold")
println("Total number of gold medal winners every year: "+medal.count())
```



NewProj1 [C:\Users\Laptop\IdeaProjects\NewProj1] - ...\src\main\scala\asnt21.scala [newproj1] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
NewProj1 src main scala asnt21.scala
Project build.sbt asnt21.scala
Run: asnt21

```
18/05/18 10:15:58 INFO MemoryStore: Block broadcast_2_piece0 stored as bytes in memory (estimated size 6.1 KB, free 902.5 MB)
18/05/18 10:15:58 INFO BlockManagerInfo: Added broadcast_2_piece0 in memory on 192.168.56.1:50715 (size: 6.1 KB, free: 902.7 MB)
18/05/18 10:15:58 INFO SparkContext: Created broadcast 2 from broadcast at DAGScheduler.scala:996
18/05/18 10:15:58 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 1 (MapPartitionsRDD[10] at show at asnt21.scala:30)
18/05/18 10:15:58 INFO TaskSchedulerImpl: Adding task set 1.0 with 1 tasks
18/05/18 10:15:58 INFO TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1, localhost, executor driver, partition 0, PROCESS_LOCAL, 5934 bytes)
18/05/18 10:15:58 INFO Executor: Running task 0.0 in stage 1.0 (TID 1)
18/05/18 10:15:58 INFO HadoopRDD: Input split: file:/C:/Users/Laptop/Desktop/dataset.txt:0+985
18/05/18 10:15:58 INFO Executor: Finished task 0.0 in stage 1.0 (TID 1). 1642 bytes result sent to driver
18/05/18 10:15:58 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 26 ms on localhost (executor driver) (1/1)
18/05/18 10:15:58 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
18/05/18 10:15:58 INFO DAGScheduler: ResultStage 1 (show at asnt21.scala:30) finished in 0.028 s
18/05/18 10:15:58 INFO DAGScheduler: Job 1 finished: show at asnt21.scala:30, took 0.047708 s
+----------+---------+--------+----------+-----+----+------------+
|First_Name|Last_Name|  Sports|Medal_Type|s_Age|year|Country_name|
+----------+---------+--------+----------+-----+----+------------+
|      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|
+----------+---------+--------+----------+-----+----+------------+
```



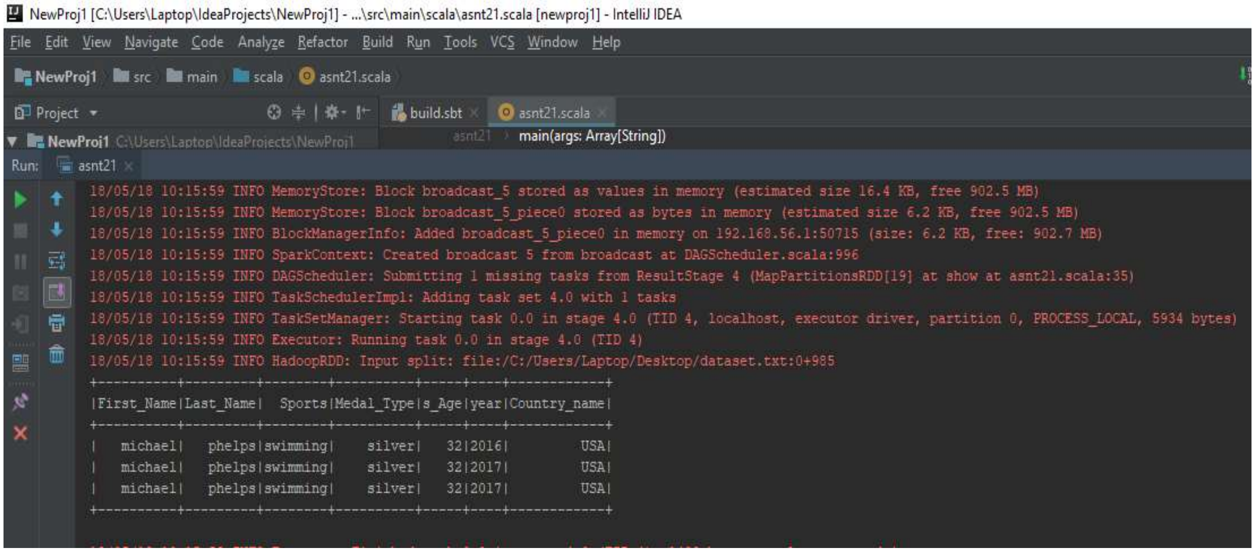
NewProj1 [C:\Users\Laptop\IdeaProjects\NewProj1] - ...\src\main\scala\asnt21.scala [newproj1] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
NewProj1 src main scala asnt21.scala
Project build.sbt asnt21.scala
Run: asnt21

```
18/05/18 10:15:59 INFO SparkContext: Created broadcast 4 from broadcast at DAGScheduler.scala:996
18/05/18 10:15:59 INFO CodeGenerator: Code generated in 10.837127 ms
Total number of gold medal winners every year: 9
18/05/18 10:15:59 INFO CodeGenerator: Code generated in 50.085783 ms
18/05/18 10:15:59 INFO Executor: Running task 0.0 in stage 3.0 (TID 3)
18/05/18 10:15:59 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
```

How many silver medals have been won by USA in each sports?

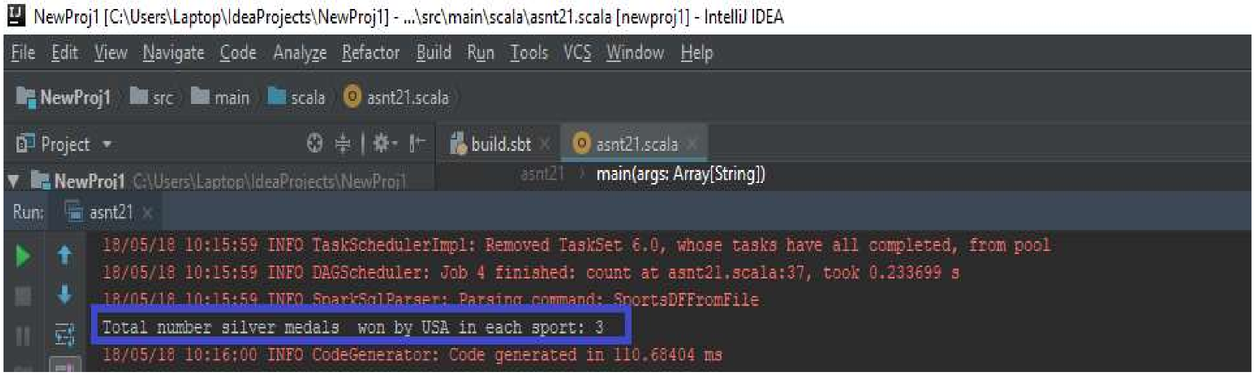
To accomplish this task, we filter Medal_type column with “silver” and Country_name column with “USA” and count them

```
SportsDFFromFile.filter((col("Medal_type") === "silver") && col("Country_name") === "USA").show()
val silver_medals = SportsDFFromFile.filter((col("Medal_type") === "silver") && col("Country_name") === "USA")
println("Total number silver medals  won by USA in each sport: "+silver_medals.count())
```



NewProj1 [C:\Users\Laptop\IdeaProjects\NewProj1] - ...\src\main\scala\asnt21.scala [newproj1] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
NewProj1 src main scala asnt21.scala
Project
build.sbt
asnt21.scala
NewProj1 C:\Users\Laptop\IdeaProjects\NewProj1
asnt21 main(args: Array[String])
Run: asnt21

```
18/05/18 10:15:59 INFO MemoryStore: Block broadcast_5 stored as values in memory (estimated size 16.4 KB, free 902.5 MB)
18/05/18 10:15:59 INFO MemoryStore: Block broadcast_5_piece0 stored as bytes in memory (estimated size 6.2 KB, free 902.5 MB)
18/05/18 10:15:59 INFO BlockManagerInfo: Added broadcast_5_piece0 in memory on 192.168.56.1:50715 (size: 6.2 KB, free: 902.7 MB)
18/05/18 10:15:59 INFO SparkContext: Created broadcast 5 from broadcast at DAGScheduler.scala:996
18/05/18 10:15:59 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 4 (MapPartitionsRDD[19] at show at asnt21.scala:35)
18/05/18 10:15:59 INFO TaskSchedulerImpl: Adding task set 4.0 with 1 tasks
18/05/18 10:15:59 INFO TaskSetManager: Starting task 0.0 in stage 4.0 (TID 4, localhost, executor driver, partition 0, PROCESS_LOCAL, 5934 bytes)
18/05/18 10:15:59 INFO Executor: Running task 0.0 in stage 4.0 (TID 4)
18/05/18 10:15:59 INFO HadoopRDD: Input split: file:/C:/Users/Laptop/Desktop/dataset.txt:0+985
+----------+---------+--------+----------+-----+----+------------+
|First_Name|Last_Name|  Sports|Medal_Type|s_Age|year|Country_name|
+----------+---------+--------+----------+-----+----+------------+
|   michael|   phelps|swimming|    silver|   32|2016|         USA|
|   michael|   phelps|swimming|    silver|   32|2017|         USA|
|   michael|   phelps|swimming|    silver|   32|2017|         USA|
+----------+---------+--------+----------+-----+----+------------+
```



NewProj1 [C:\Users\Laptop\IdeaProjects\NewProj1] - ...\src\main\scala\asnt21.scala [newproj1] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
NewProj1 src main scala asnt21.scala
Project
build.sbt
asnt21.scala
NewProj1 C:\Users\Laptop\IdeaProjects\NewProj1
asnt21 main(args: Array[String])
Run: asnt21

```
18/05/18 10:15:59 INFO TaskSchedulerImpl: Removed TaskSet 6.0, whose tasks have all completed, from pool
18/05/18 10:15:59 INFO DAGScheduler: Job 4 finished: count at asnt21.scala:37, took 0.233699 s
18/05/18 10:15:59 INFO SparkSqlParser: Parsing command: SportsDFFromFile
Total number silver medals  won by USA in each sport: 3
18/05/18 10:16:00 INFO CodeGenerator: Code generated in 110.68404 ms
```

Task 2

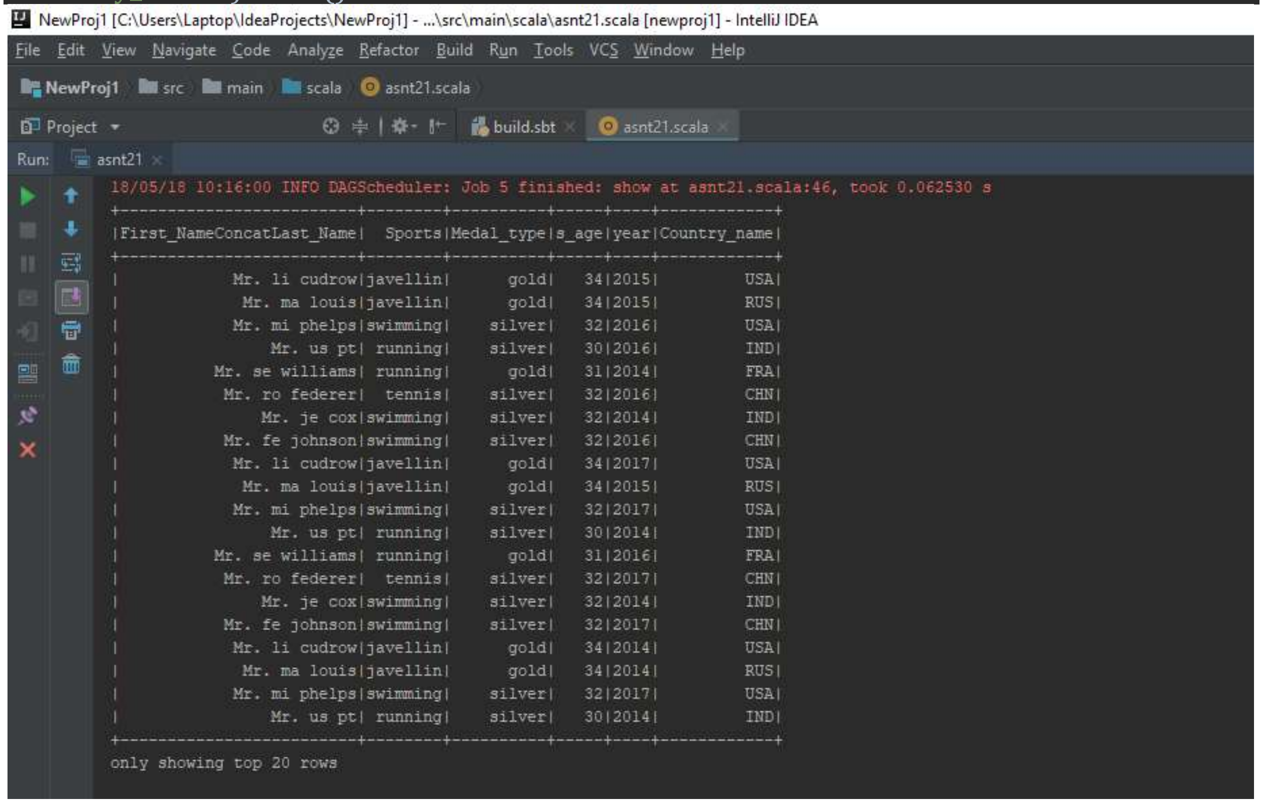
Using udfs on dataframe

1. Change firstname, lastname columns into Mr.first_two_letters_of_firstname<space>lastname

for example - michael, phelps becomes Mr.mi phelps

- To accomplish this task we write UDF which extracts first name and last name of the athlete and add "Mr" and joins the "Mr. firstname lastname".
- Then we register the Data frame to temptable and give the same name as dataframe
- Then we create a column called "first_NameConcatLast_Name" and display it along with other columns in the data frame

```
val Name = udf((First_Name: String,Last_Name: String) => "Mr. "
.concat(First_Name.substring(0,2).concat(" ")
  .concat(Last_Name)))

 SportsDFFromFile.registerTempTable("SportsDFFromFile")

 SportsDFFromFile.withColumn("First_NameConcatLast_Name", Name($"First_Name",
$"Last_Name")).select("First_NameConcatLast_Name","Sports","Medal_type","s_age","year",
"Country_name").show()
```

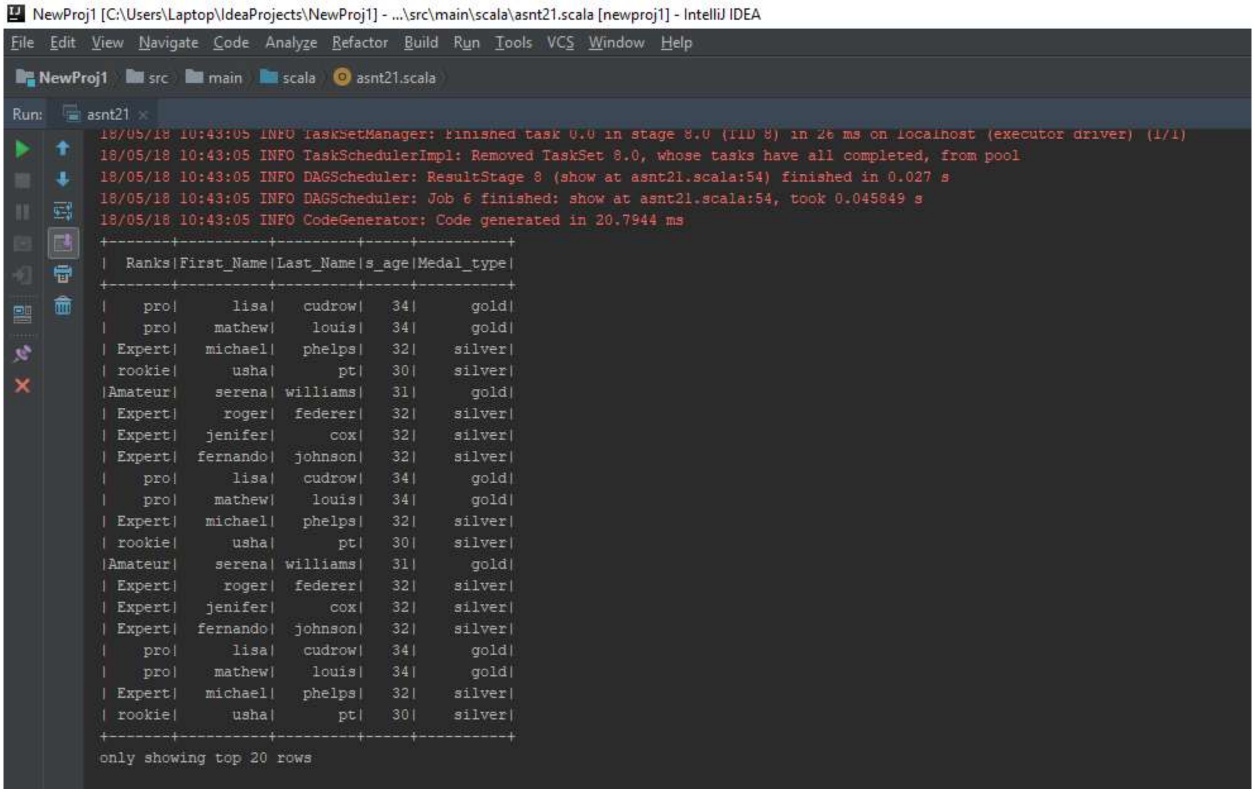

NewProj1 [C:\Users\Laptop\IdeaProjects\NewProj1] - ...\src\main\scala\asnt21.scala [newproj1] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
NewProj1 src main scala asnt21.scala
Project
build.sbt
asnt21.scala
Run: asnt21

```
18/05/18 10:16:00 INFO DAGScheduler: Job 5 finished: show at asnt21.scala:46, took 0.062530 s
+-------------------------+--------+----------+-----+----+------------+
|First_NameConcatLast_Name|  Sports|Medal_type|s_age|year|Country_name|
+-------------------------+--------+----------+-----+----+------------+
|            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
```

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

- To accomplish this task, we write a function called “ranking” which takes two parameters namely Age and Medal, we filter medal by gold, silver and age by <=31 & >=32 as pro, amateur etc..
- We create a UDF on the above function
- We create a column called “Ranks” , call the udf function and display other columns from the temptable called “SportsDFFromFile

```
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 ""}

val Rank = udf(Ranking(_:Int,_:String))
SportsDFFromFile.withColumn("Ranks",
Rank($"s_age",$"Medal_type")).select("Ranks","First_Name","Last_Name","s_age","Medal_type").show()
```


NewProj1 [C:\Users\Laptop\IdeaProjects\NewProj1] - ...\src\main\scala\asnt21.scala [newproj1] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
NewProj1 src main scala asnt21.scala
Run: asnt21

```
18/05/18 10:43:05 INFO TaskSetManager: Finished task 0.0 in stage 8.0 (TID 8) in 26 ms on localhost (executor driver) (1/1)
18/05/18 10:43:05 INFO TaskSchedulerImpl: Removed TaskSet 8.0, whose tasks have all completed, from pool
18/05/18 10:43:05 INFO DAGScheduler: ResultStage 8 (show at asnt21.scala:54) finished in 0.027 s
18/05/18 10:43:05 INFO DAGScheduler: Job 6 finished: show at asnt21.scala:54, took 0.045849 s
18/05/18 10:43:05 INFO CodeGenerator: Code generated in 20.7944 ms
+-------+----------+---------+-----+----------+
|  Ranks|First_Name|Last_Name|s_age|Medal_type|
+-------+----------+---------+-----+----------+
|    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
```