Session 21:

SPARK SQL 2

Assignment 1

Task 1

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

Solution: // Create a case class globally.

```
  //Inferring the Schema Using Reflection.Automatically converting an RDD
containing case classes to a DataFrame.
  // The case class defines the schema of the table. The names of the arguments
to the case class are read using reflection
  // and become the names of the columns.
```

SOLUTION:

```
package SQL

//import org.apache.spark.sql.SparkSession

import org.apache.spark.sql.SparkSession

object Sports_Winner {

  //Create a case class globally to be used inside the main method
  //Inferring the Schema Using Reflection.Automatically converting an RDD containing case classes
to a DataFrame.
  // The case class defines the schema of the table. The names of the arguments to the case class
are read using reflection
  // and become the names of the columns.
  // Main method - The execution entry point for the program
```

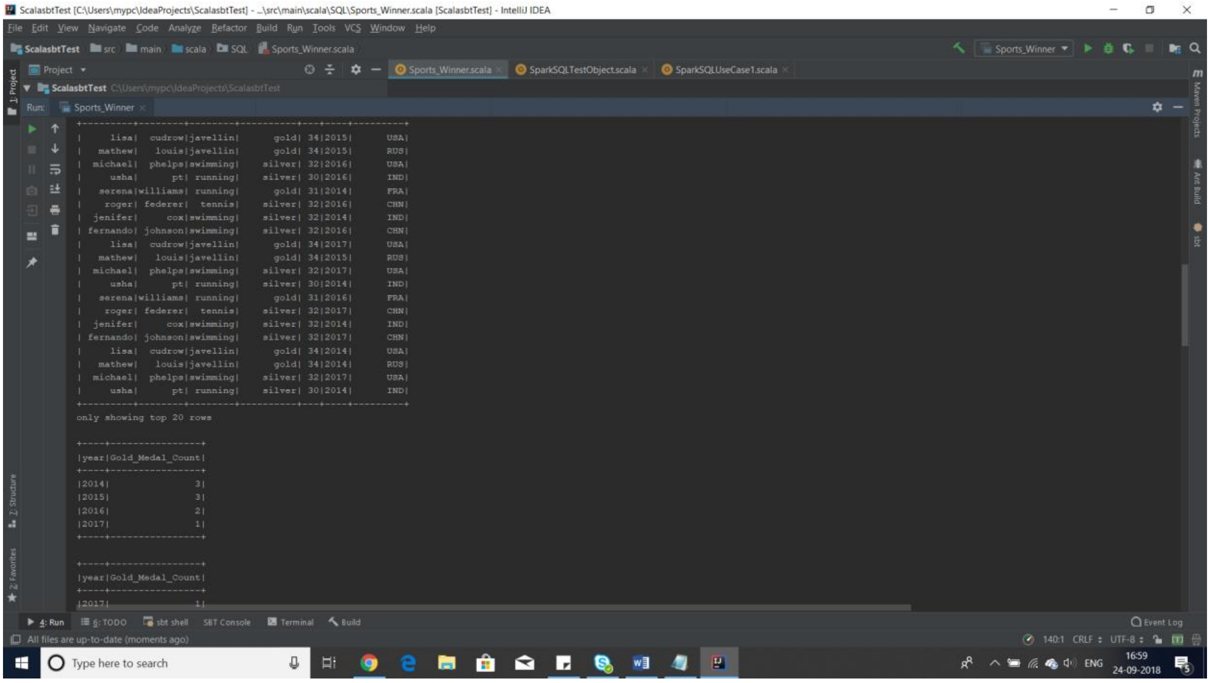
```
 case class
Sports(firstname:String,lastname:String,sports:String,medal_type:String,age:Int,year:Long,country
id:String)

 def main(args: Array[String]): Unit = {

  val spark = SparkSession

    .builder()

    .master("local")

    .appName("Sports_data")

    .config("spark.some.config.option", "some-value")

    .getOrCreate()

  println("spark session object created")

  spark.sparkContext.setLogLevel("WARN")


  //println("spark session object created")

  import spark.implicits._

  val data = spark.sparkContext.textFile("C:/Users/mypc/Desktop/Sports_data.txt")

  val header = data.first()

  val header1 = data.filter(row => row != header)

  val sports_data = header1.map(x => x.split(",")).map(x => Sports(x(0), x(1), x(2), x(3), x(4).toInt,
x(5).toLong, x(6))).toDF()

  sports_data.show()

  //Converting the above created schema into an SQL view named sport

  sports_data.createOrReplaceTempView("sport")
```

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

```
  // Selecting year & couting the occurance of each year by filtering medal_type condition as gold.

  // grouping by year & ordering the result based upon the year.

  val a1=spark.sql("Select year, count(year)as Gold_Medal_Count from sport  where medal_type
='gold' group by year order by year ASC")

  val a2=spark.sql("Select year, count(year)as Gold_Medal_Count from sport  where medal_type
='gold' group by year order by Gold_Medal_Count ASC")

  a1.show()
```

a2.show()

Screenshot:

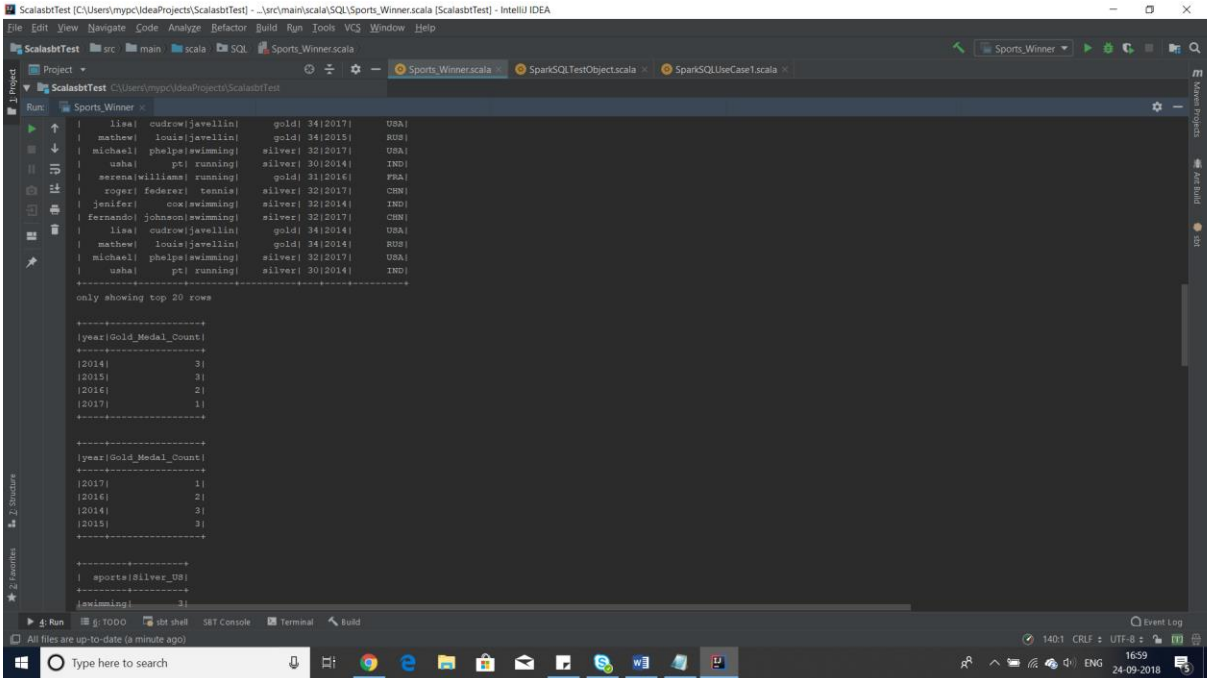


ScalasbtTest [C:\Users\mypc\IdeaProjects\ScalasbtTest] - ...\src\main\scala\SQL\Sports_Winner.scala [ScalasbtTest] - IntelliJ IDEA

```
+---------+--------+--------+---------+---+----+---------+
|     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

+----+----------------+
|year|Gold_Medal_Count|
+----+----------------+
|2014|               3|
|2015|               3|
|2016|               2|
|2017|               1|
+----+----------------+


+----+----------------+
|year|Gold_Medal_Count|
+----+----------------+
|2017|               1|
```

All files are up-to-date (moments ago)

140:1 CRLF UTF-8 ENG 16:59 24-09-2018

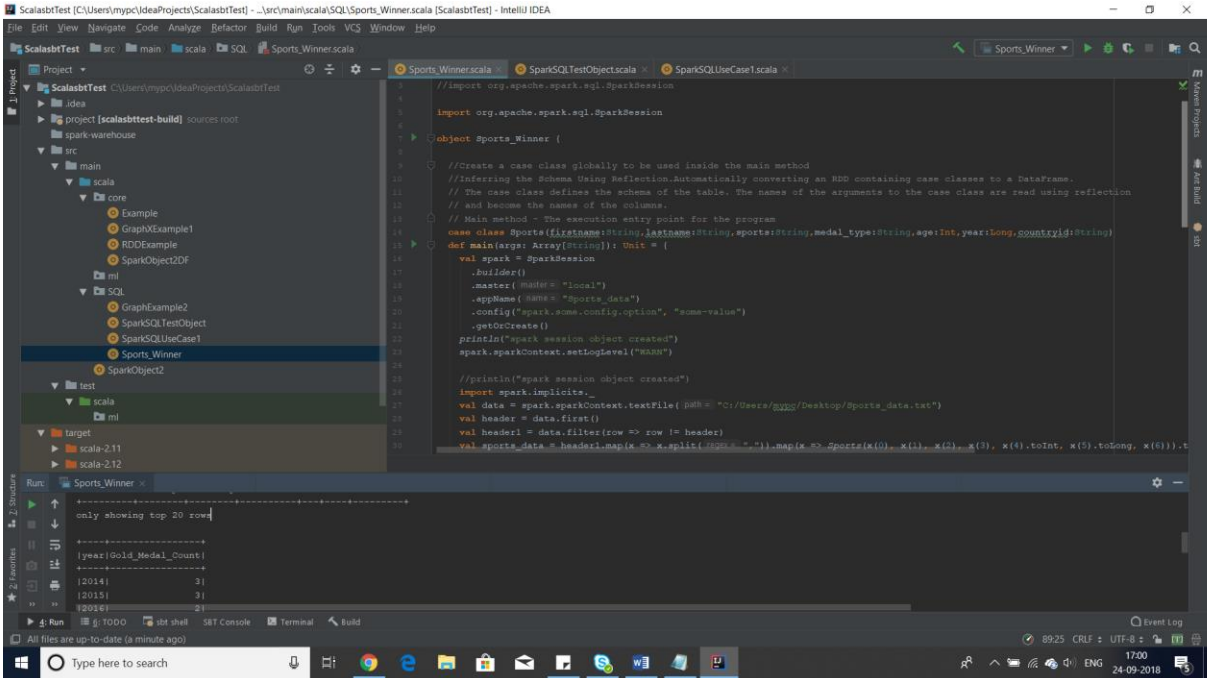


ScalasbtTest [C:\Users\mypc\IdeaProjects\ScalasbtTest] - ...\src\main\scala\SQL\Sports_Winner.scala [ScalasbtTest] - IntelliJ IDEA

```
|     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

+----+----------------+
|year|Gold_Medal_Count|
+----+----------------+
|2014|               3|
|2015|               3|
|2016|               2|
|2017|               1|
+----+----------------+


+----+----------------+
|year|Gold_Medal_Count|
+----+----------------+
|2017|               1|
|2016|               2|
|2014|               3|
|2015|               3|
+----+----------------+


+--------+---------+
|  sports|Silver_US|
+--------+---------+
|swimming|        3|
```

All files are up-to-date (a minute ago)

140:1 CRLF UTF-8 ENG 16:59 24-09-2018



ScalasbtTest [C:\Users\mypc\IdeaProjects\ScalasbtTest] - ...\src\main\scala\SQL\Sports_Winner.scala [ScalasbtTest] - IntelliJ IDEA

```
//import org.apache.spark.sql.SparkSession

import org.apache.spark.sql.SparkSession

object Sports_Winner {

  //Create a case class globally to be used inside the main method
  //Inferring the Schema Using Reflection.Automatically converting an RDD containing case classes to a DataFrame.
  // The case class defines the schema of the table. The names of the arguments to the case class are read using reflection
  // and become the names of the columns.
  // Main method - The execution entry point for the program
  case class Sports(firstname:String,lastname:String,sports:String,medal_type:String,age:Int,year:Long,countryid:String)
  def main(args: Array[String]): Unit = {
    val spark = SparkSession
      .builder()
      .master("local")
      .appName("Sports_data")
      .config("spark.some.config.option", "some-value")
      .getOrCreate()
    println("spark session object created")
    spark.sparkContext.setLogLevel("WARN")

    //println("spark session object created")
    import spark.implicits._
    val data = spark.sparkContext.textFile("C:/Users/mypc/Desktop/Sports_data.txt")
    val header = data.first()
    val header1 = data.filter(row => row != header)
    val sports_data = header1.map(x => x.split(",")).map(x => Sports(x(0), x(1), x(2), x(3), x(4).toInt, x(5).toLong, x(6))).t
```

```
+----------+--------+--------+----------+---+----+----------+
only showing top 20 rows

+----+----------------+
|year|Gold_Medal_Count|
+----+----------------+
|2014|               3|
|2015|               3|
|2016|               2|
```

All files are up-to-date (a minute ago)

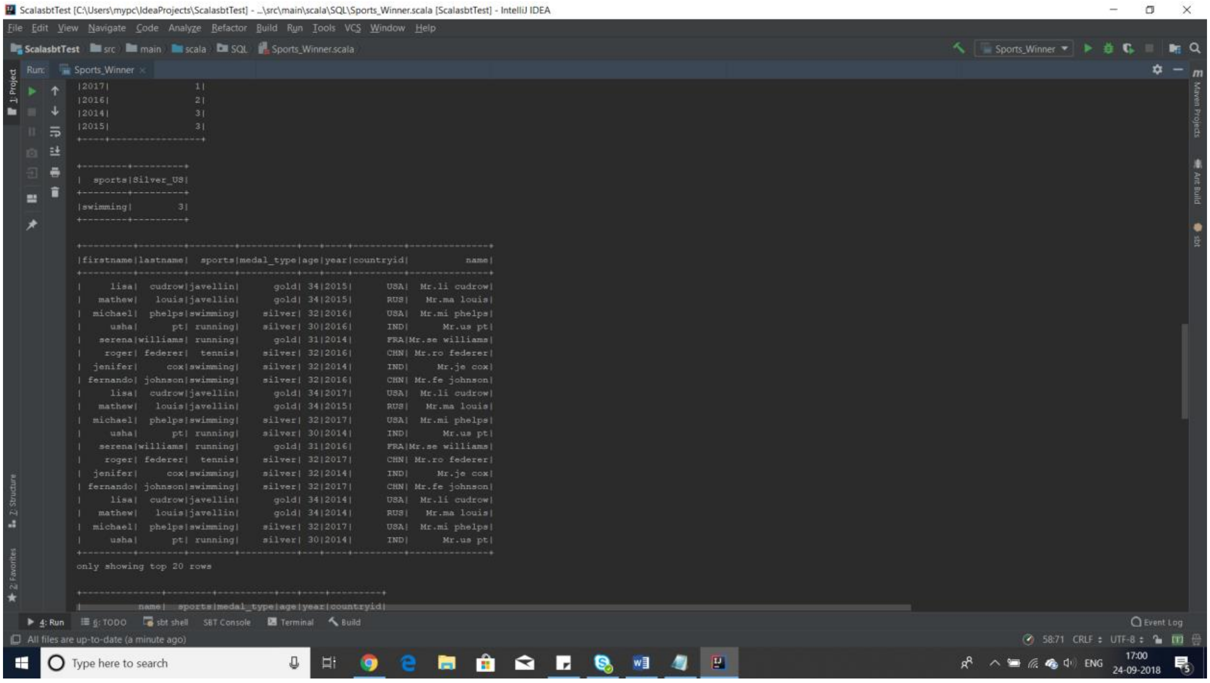
//2. How many silver medals have been won by USA in each sport?

// Selecting sports & count of sports as Silver_US from sports view. Provided the filter as medal_type = 'silver' &

// coutry as USA, grouping by count & ordering the result based upon count of medals won.

```
val a3 = spark.sql("select sports, count(sports) as Silver_US from sport where (medal_type = 'silver' AND countryid = 'USA') group by sports order by Silver_US ASC")

a3.show()

}}
```

Note: //Silver_US column in screenshot shows number of silver medal won by USA.


ScalasbtTest [C:\Users\mypc\IdeaProjects\ScalasbtTest] - ...\src\main\scala\SQL\Sports_Winner.scala [ScalasbtTest] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
ScalasbtTest src main scala SQL Sports_Winner.scala
Run: Sports_Winner
4: Run 6: TODO sbt shell SBT Console Terminal Build
Event Log
All files are up-to-date (a minute ago)
58:71 CRLF UTF-8
Type here to search
ENG 17:00 24-09-2018

```
|2017|               1|
|2016|               2|
|2014|               3|
|2015|               3|
+----+----------------+


+--------+---------+
|  sports|Silver_US|
+--------+---------+
|swimming|        3|
+--------+---------+


+---------+--------+--------+----------+---+----+---------+--------------+
|firstname|lastname|  sports|medal_type|age|year|countryid|          name|
+---------+--------+--------+----------+---+----+---------+--------------+
|     lisa|  cudrow|javellin|      gold| 34|2015|      USA|  Mr.li cudrow|
|   mathew|   louis|javellin|      gold| 34|2015|      RUS|   Mr.ma louis|
|  michael|  phelps|swimming|    silver| 32|2016|      USA|  Mr.mi phelps|
|     usha|      pt| running|    silver| 30|2016|      IND|      Mr.us pt|
|   serena|williams| running|      gold| 31|2014|      FRA|Mr.se williams|
|    roger| federer|  tennis|    silver| 32|2016|      CHN| Mr.ro federer|
|  jenifer|     cox|swimming|    silver| 32|2014|      IND|     Mr.je cox|
| fernando| johnson|swimming|    silver| 32|2016|      CHN| Mr.fe johnson|
|     lisa|  cudrow|javellin|      gold| 34|2017|      USA|  Mr.li cudrow|
|   mathew|   louis|javellin|      gold| 34|2015|      RUS|   Mr.ma louis|
|  michael|  phelps|swimming|    silver| 32|2017|      USA|  Mr.mi phelps|
|     usha|      pt| running|    silver| 30|2014|      IND|      Mr.us pt|
|   serena|williams| running|      gold| 31|2016|      FRA|Mr.se williams|
|    roger| federer|  tennis|    silver| 32|2017|      CHN| Mr.ro federer|
|  jenifer|     cox|swimming|    silver| 32|2014|      IND|     Mr.je cox|
| fernando| johnson|swimming|    silver| 32|2017|      CHN| Mr.fe johnson|
|     lisa|  cudrow|javellin|      gold| 34|2014|      USA|  Mr.li cudrow|
|   mathew|   louis|javellin|      gold| 34|2014|      RUS|   Mr.ma louis|
|  michael|  phelps|swimming|    silver| 32|2017|      USA|  Mr.mi phelps|
|     usha|      pt| running|    silver| 30|2014|      IND|      Mr.us pt|
+---------+--------+--------+----------+---+----+---------+--------------+
only showing top 20 rows

+--------------+--------+----------+---+----+---------+
|          name|  sports|medal_type|age|year|countryid|
```

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

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

Solution:

/*Using udfs on dataframe

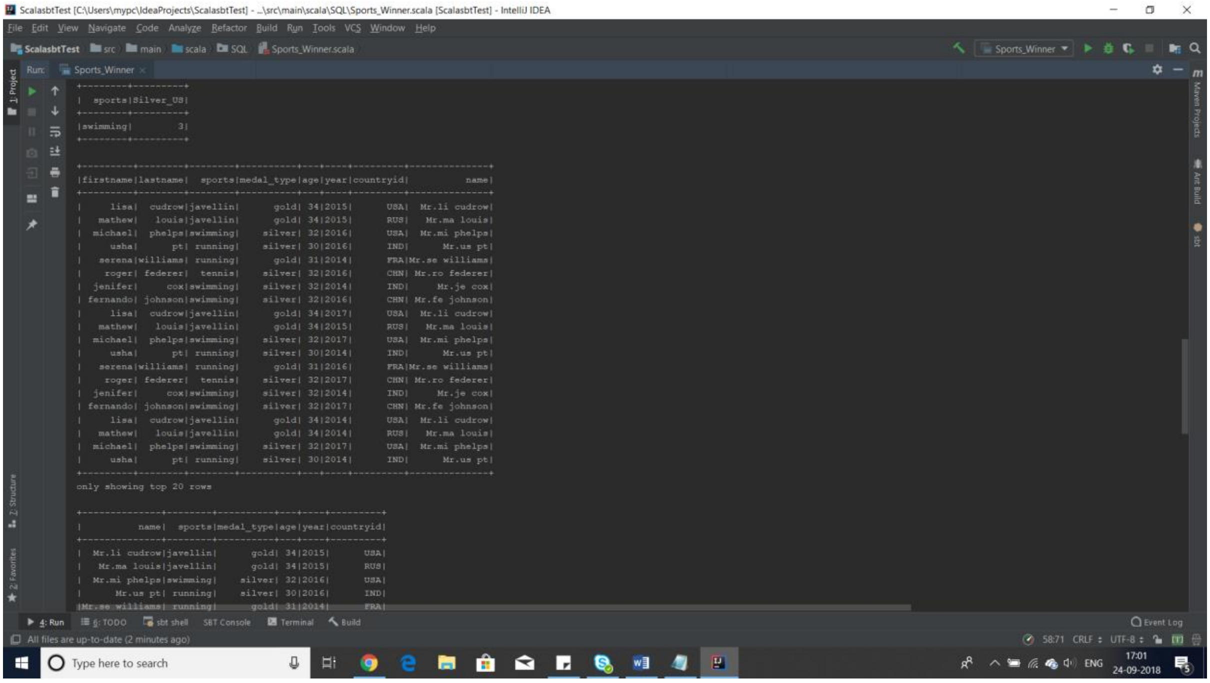
1. Change firstname, lastname columns into

Mr.first_two_letters_of_firstname<space>lastname

for example - michael, phelps becomes Mr.mi phelps*/

```
  import org.apache.spark.sql.functions.udf
   def udf_change_columns = udf((firstname:String,lastname:String)=>{
     val twochars_Firstname=firstname.substring(0,2)
     val name ="Mr."+twochars_Firstname+" "+lastname
      name})
   val df
=sports_data.withColumn("name",udf_change_columns($"firstname",$"last
name"))
   df.show()
   val df1=df.select("name","sports","medal_type","age","year","countryid")
  // val df2 = df.drop("firstname","lastname")
   df1.show()


   val udf_add_columns = udf((medal_type:String,age:Int)=>{
    val ranking= if (medal_type.equals("gold") && age >= 32 ) "pro"
    else if(medal_type.equals("gold") && age <= 31 ) "amateur"
    else if(medal_type.equals("silver") && age >= 32 ) "expert"
    else if(medal_type.equals("silver") && age <= 31 ) "rookie"
    else "NA"
     ranking
   })
val added_column
=df1.withColumn("ranking",udf_add_columns($"medal_type", $"age"))
   added_column.show()
   val added_column1=
sports_data.withColumn("ranking",udf_add_columns($"medal_type",$"age"
))
   added_column1.show()
  }}
```

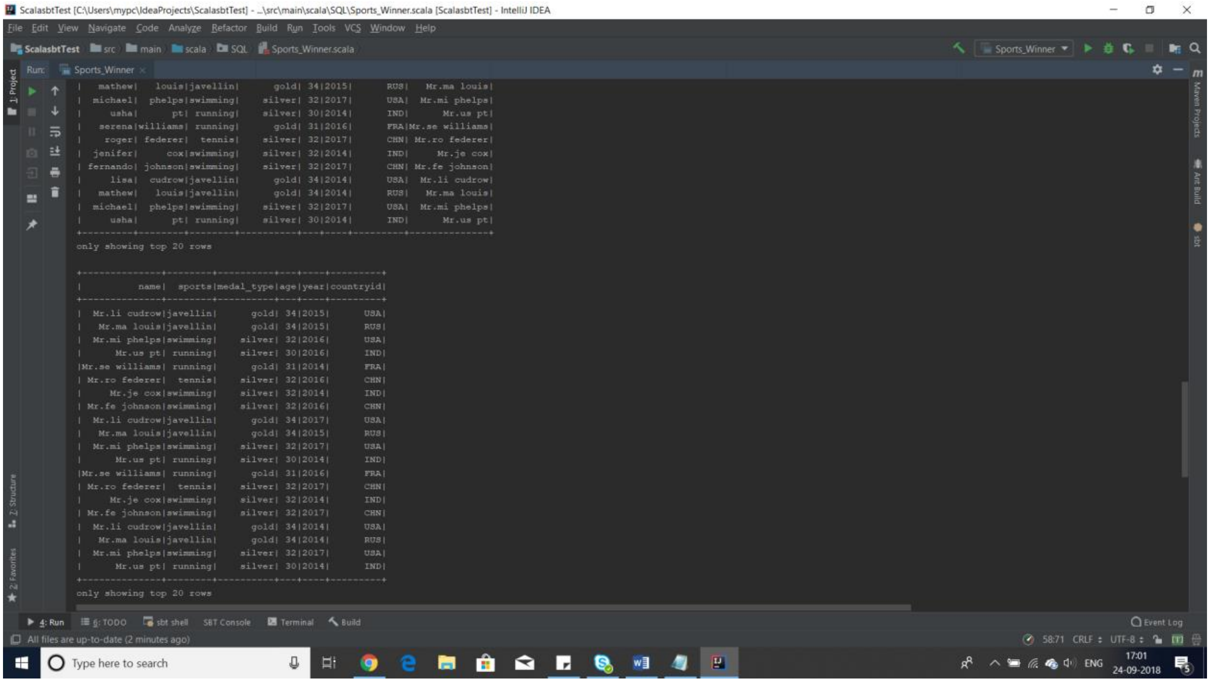
Screenshot:



ScalasbtTest [C:\Users\mypc\IdeaProjects\ScalasbtTest] - ...\src\main\scala\SQL\Sports_Winner.scala [ScalasbtTest] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
ScalasbtTest src main scala SQL Sports_Winner.scala
Sports_Winner
Run: Sports_Winner
4: Run
6: TODO
sbt shell
SBT Console
Terminal
Build
Event Log
58:71 CRLF UTF-8
Type here to search
ENG
17:01
24-09-2018

```
+--------+---------+
|  sports|Silver_US|
+--------+---------+
|swimming|        3|
+--------+---------+

+---------+--------+--------+----------+---+----+---------+--------------+
|firstname|lastname|  sports|medal_type|age|year|countryid|          name|
+---------+--------+--------+----------+---+----+---------+--------------+
|     lisa|  cudrow|javellin|      gold| 34|2015|      USA|  Mr.li cudrow|
|   mathew|   louis|javellin|      gold| 34|2015|      RUS|   Mr.ma louis|
|  michael|  phelps|swimming|    silver| 32|2016|      USA|  Mr.mi phelps|
|     usha|      pt| running|    silver| 30|2016|      IND|      Mr.us pt|
|   serena|williams| running|      gold| 31|2014|      FRA|Mr.se williams|
|    roger| federer|  tennis|    silver| 32|2016|      CHN| Mr.ro federer|
|  jenifer|     cox|swimming|    silver| 32|2014|      IND|     Mr.je cox|
| fernando| johnson|swimming|    silver| 32|2016|      CHN| Mr.fe johnson|
|     lisa|  cudrow|javellin|      gold| 34|2017|      USA|  Mr.li cudrow|
|   mathew|   louis|javellin|      gold| 34|2015|      RUS|   Mr.ma louis|
|  michael|  phelps|swimming|    silver| 32|2017|      USA|  Mr.mi phelps|
|     usha|      pt| running|    silver| 30|2014|      IND|      Mr.us pt|
|   serena|williams| running|      gold| 31|2016|      FRA|Mr.se williams|
|    roger| federer|  tennis|    silver| 32|2017|      CHN| Mr.ro federer|
|  jenifer|     cox|swimming|    silver| 32|2014|      IND|     Mr.je cox|
| fernando| johnson|swimming|    silver| 32|2017|      CHN| Mr.fe johnson|
|     lisa|  cudrow|javellin|      gold| 34|2014|      USA|  Mr.li cudrow|
|   mathew|   louis|javellin|      gold| 34|2014|      RUS|   Mr.ma louis|
|  michael|  phelps|swimming|    silver| 32|2017|      USA|  Mr.mi phelps|
|     usha|      pt| running|    silver| 30|2014|      IND|      Mr.us pt|
+---------+--------+--------+----------+---+----+---------+--------------+
only showing top 20 rows

+--------------+--------+----------+---+----+---------+
|          name|  sports|medal_type|age|year|countryid|
+--------------+--------+----------+---+----+---------+
|  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|
```

All files are up-to-date (2 minutes ago)

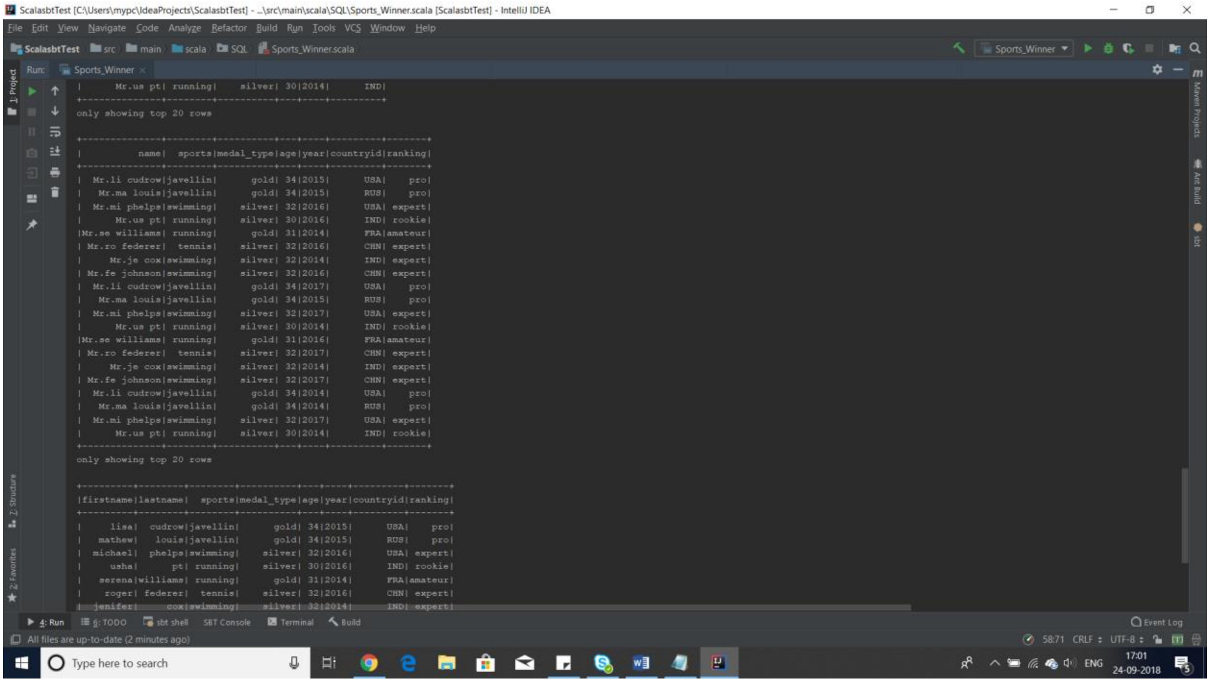


ScalasbtTest [C:\Users\mypc\IdeaProjects\ScalasbtTest] - ...\src\main\scala\SQL\Sports_Winner.scala [ScalasbtTest] - IntelliJ IDEA

File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
ScalasbtTest src main scala SQL Sports_Winner.scala
Sports_Winner
Run: Sports_Winner
4: Run
6: TODO
sbt shell
SBT Console
Terminal
Build
Event Log
Type here to search
ENG
17:01
24-09-2018

```
|   mathew|   louis|javellin|      gold| 34|2015|      RUS|   Mr.ma louis|
|  michael|  phelps|swimming|    silver| 32|2017|      USA|  Mr.mi phelps|
|     usha|      pt| running|    silver| 30|2014|      IND|      Mr.us pt|
|   serena|williams| running|      gold| 31|2016|      FRA|Mr.se williams|
|    roger| federer|  tennis|    silver| 32|2017|      CHN| Mr.ro federer|
|  jenifer|     cox|swimming|    silver| 32|2014|      IND|     Mr.je cox|
| fernando| johnson|swimming|    silver| 32|2017|      CHN| Mr.fe johnson|
|     lisa|  cudrow|javellin|      gold| 34|2014|      USA|  Mr.li cudrow|
|   mathew|   louis|javellin|      gold| 34|2014|      RUS|   Mr.ma louis|
|  michael|  phelps|swimming|    silver| 32|2017|      USA|  Mr.mi phelps|
|     usha|      pt| running|    silver| 30|2014|      IND|      Mr.us pt|
+---------+--------+--------+----------+---+----+---------+--------------+
only showing top 20 rows

+--------------+--------+----------+---+----+---------+
|          name|  sports|medal_type|age|year|countryid|
+--------------+--------+----------+---+----+---------+
|  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
```

All files are up-to-date (2 minutes ago)

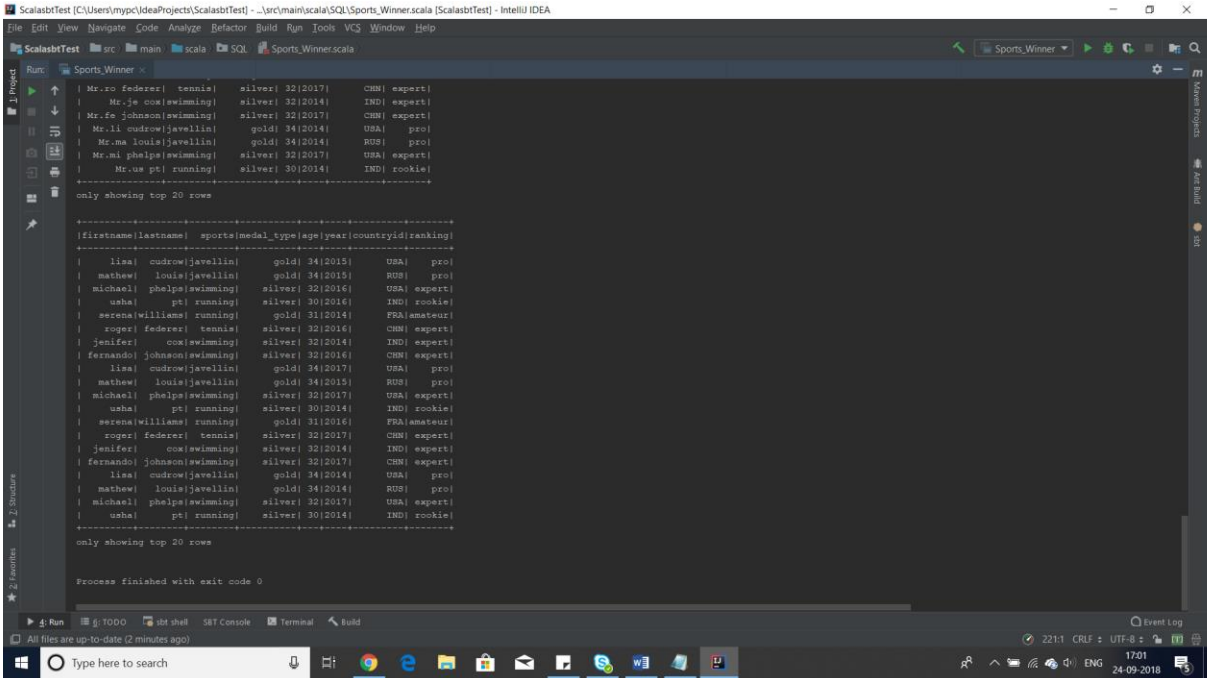


ScalasbtTest [C:\Users\mypc\IdeaProjects\ScalasbtTest] - ...\src\main\scala\SQL\Sports_Winner.scala [ScalasbtTest] - IntelliJ IDEA

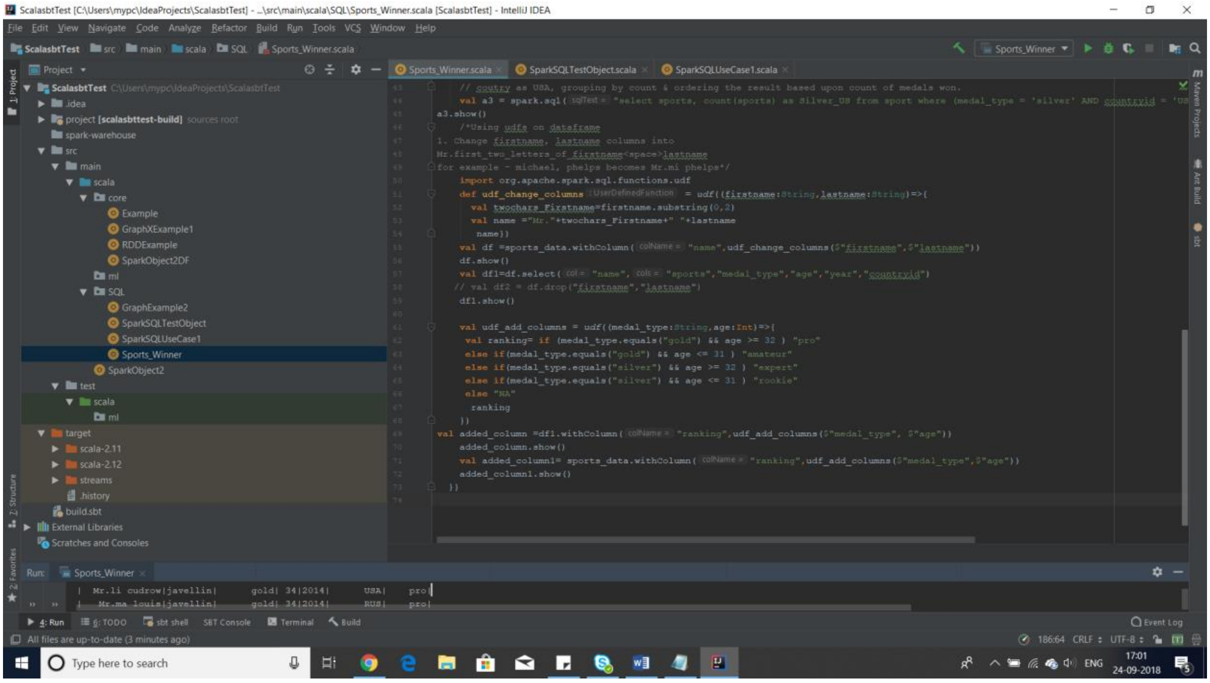
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
ScalasbtTest src main scala SQL Sports_Winner.scala
Sports_Winner
Run: Sports_Winner
4: Run
6: TODO
sbt shell
SBT Console
Terminal
Build
Event Log
Type here to search
ENG
17:01
24-09-2018

```
|      Mr.us pt| running|    silver| 30|2014|      IND|
+--------------+--------+----------+---+----+---------+
only showing top 20 rows

+--------------+--------+----------+---+----+---------+-------+
|          name|  sports|medal_type|age|year|countryid|ranking|
+--------------+--------+----------+---+----+---------+-------+
|  Mr.li cudrow|javellin|      gold| 34|2015|      USA|    pro|
|   Mr.ma louis|javellin|      gold| 34|2015|      RUS|    pro|
|  Mr.mi phelps|swimming|    silver| 32|2016|      USA| expert|
|      Mr.us pt| running|    silver| 30|2016|      IND| rookie|
|Mr.se williams| running|      gold| 31|2014|      FRA|amateur|
| Mr.ro federer|  tennis|    silver| 32|2016|      CHN| expert|
|     Mr.je cox|swimming|    silver| 32|2014|      IND| expert|
| Mr.fe johnson|swimming|    silver| 32|2016|      CHN| expert|
|  Mr.li cudrow|javellin|      gold| 34|2017|      USA|    pro|
|   Mr.ma louis|javellin|      gold| 34|2015|      RUS|    pro|
|  Mr.mi phelps|swimming|    silver| 32|2017|      USA| expert|
|      Mr.us pt| running|    silver| 30|2014|      IND| rookie|
|Mr.se williams| running|      gold| 31|2016|      FRA|amateur|
| Mr.ro federer|  tennis|    silver| 32|2017|      CHN| expert|
|     Mr.je cox|swimming|    silver| 32|2014|      IND| expert|
| Mr.fe johnson|swimming|    silver| 32|2017|      CHN| expert|
|  Mr.li cudrow|javellin|      gold| 34|2014|      USA|    pro|
|   Mr.ma louis|javellin|      gold| 34|2014|      RUS|    pro|
|  Mr.mi phelps|swimming|    silver| 32|2017|      USA| expert|
|      Mr.us pt| running|    silver| 30|2014|      IND| rookie|
+--------------+--------+----------+---+----+---------+-------+
only showing top 20 rows

+---------+--------+--------+----------+---+----+---------+-------+
|firstname|lastname|  sports|medal_type|age|year|countryid|ranking|
+---------+--------+--------+----------+---+----+---------+-------+
|     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|
```

All files are up-to-date (2 minutes ago)

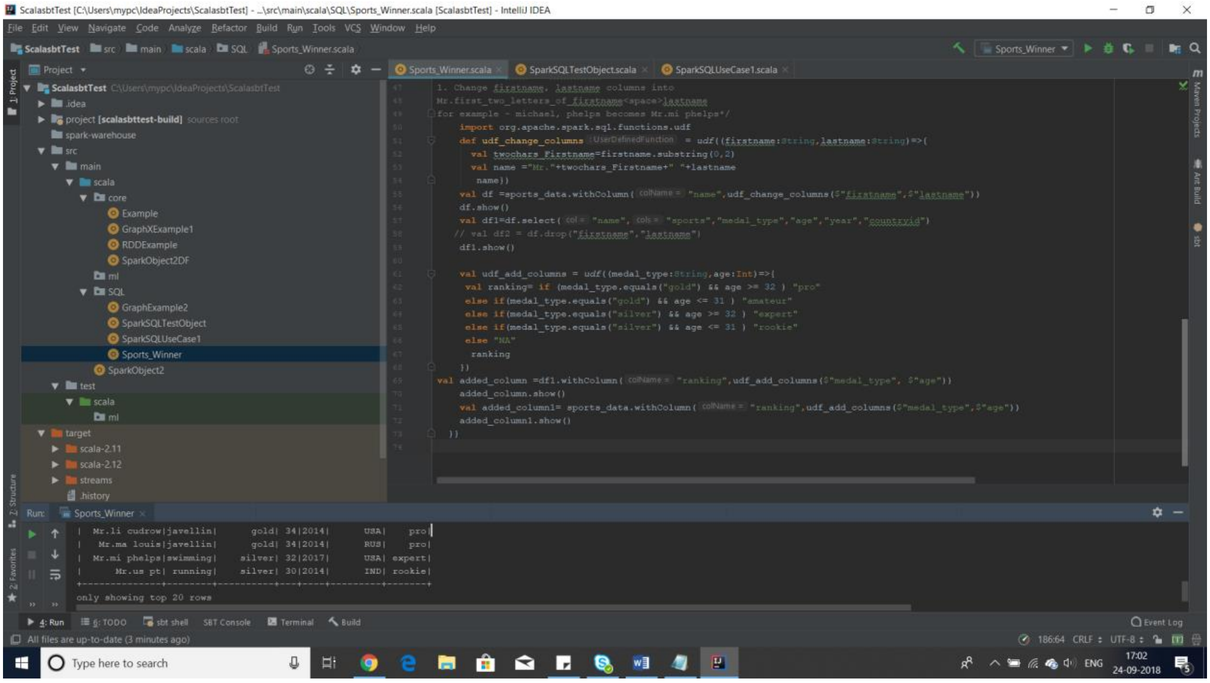

ScalasbtTest [C:\Users\mypc\IdeaProjects\ScalasbtTest] - ...\src\main\scala\SQL\Sports_Winner.scala [ScalasbtTest] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
ScalasbtTest src main scala SQL Sports_Winner.scala
Sports_Winner
Run: Sports_Winner
4: Run 6: TODO sbt shell SBT Console Terminal Build
Event Log
All files are up-to-date (2 minutes ago)
221:1 CRLF UTF-8
Type here to search
ENG 17:01 24-09-2018

```
|  Mr.ro federer|  tennis|    silver| 32|2017|       CHN| expert|
|     Mr.je cox|swimming|    silver| 32|2014|       IND| expert|
| Mr.fe johnson|swimming|    silver| 32|2017|       CHN| expert|
|  Mr.li cudrow|javellin|      gold| 34|2014|       USA|    pro|
|   Mr.ma louis|javellin|      gold| 34|2014|       RUS|    pro|
|  Mr.mi phelps|swimming|    silver| 32|2017|       USA| expert|
|      Mr.us pt| running|    silver| 30|2014|       IND| rookie|
+--------------+--------+----------+---+----+----------+-------+
only showing top 20 rows

+---------+--------+--------+----------+---+----+---------+-------+
|firstname|lastname|  sports|medal_type|age|year|countryid|ranking|
+---------+--------+--------+----------+---+----+---------+-------+
|     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


Process finished with exit code 0
```


ScalasbtTest [C:\Users\mypc\IdeaProjects\ScalasbtTest] - ...\src\main\scala\SQL\Sports_Winner.scala [ScalasbtTest] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
ScalasbtTest src main scala SQL Sports_Winner.scala
Sports_Winner
Project
ScalasbtTest C:\Users\mypc\IdeaProjects\ScalasbtTest
.idea
project [scalasbttest-build] sources root
spark-warehouse
src
main
scala
core
Example
GraphXExample1
RDDExample
SparkObject2DF
ml
SQL
GraphExample2
SparkSQLTestObject
SparkSQLUseCase1
Sports_Winner
SparkObject2
test
scala
ml
target
scala-2.11
scala-2.12
streams
.history
build.sbt
External Libraries
Scratches and Consoles
Sports_Winner.scala
SparkSQLTestObject.scala
SparkSQLUseCase1.scala
Run: Sports_Winner
4: Run 6: TODO sbt shell SBT Console Terminal Build
Event Log
All files are up-to-date (3 minutes ago)
186:64 CRLF UTF-8
Type here to search
ENG 17:01 24-09-2018

```
    // coutry as USA, grouping by count & ordering the result based upon count of medals won.
    val a3 = spark.sql( "select sports, count(sports) as Silver_US from sport where (medal_type = 'silver' AND countryid = 'US
a3.show()
    /*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*/
    import org.apache.spark.sql.functions.udf
    def udf_change_columns  = udf((firstname:String,lastname:String)=>{
      val twochars_Firstname=firstname.substring(0,2)
      val name ="Mr."+twochars_Firstname+" "+lastname
       name})
    val df =sports_data.withColumn( "name",udf_change_columns($"firstname",$"lastname"))
    df.show()
    val df1=df.select( "name",  "sports","medal_type","age","year","countryid")
   // val df2 = df.drop("firstname","lastname")
    df1.show()

    val udf_add_columns = udf((medal_type:String,age:Int)=>{
     val ranking= if (medal_type.equals("gold") && age >= 32 ) "pro"
     else if(medal_type.equals("gold") && age <= 31 ) "amateur"
     else if(medal_type.equals("silver") && age >= 32 ) "expert"
     else if(medal_type.equals("silver") && age <= 31 ) "rookie"
     else "NA"
      ranking
    })
val added_column =df1.withColumn( "ranking",udf_add_columns($"medal_type", $"age"))
    added_column.show()
    val added_column1= sports_data.withColumn( "ranking",udf_add_columns($"medal_type",$"age"))
    added_column1.show()
  }}
```

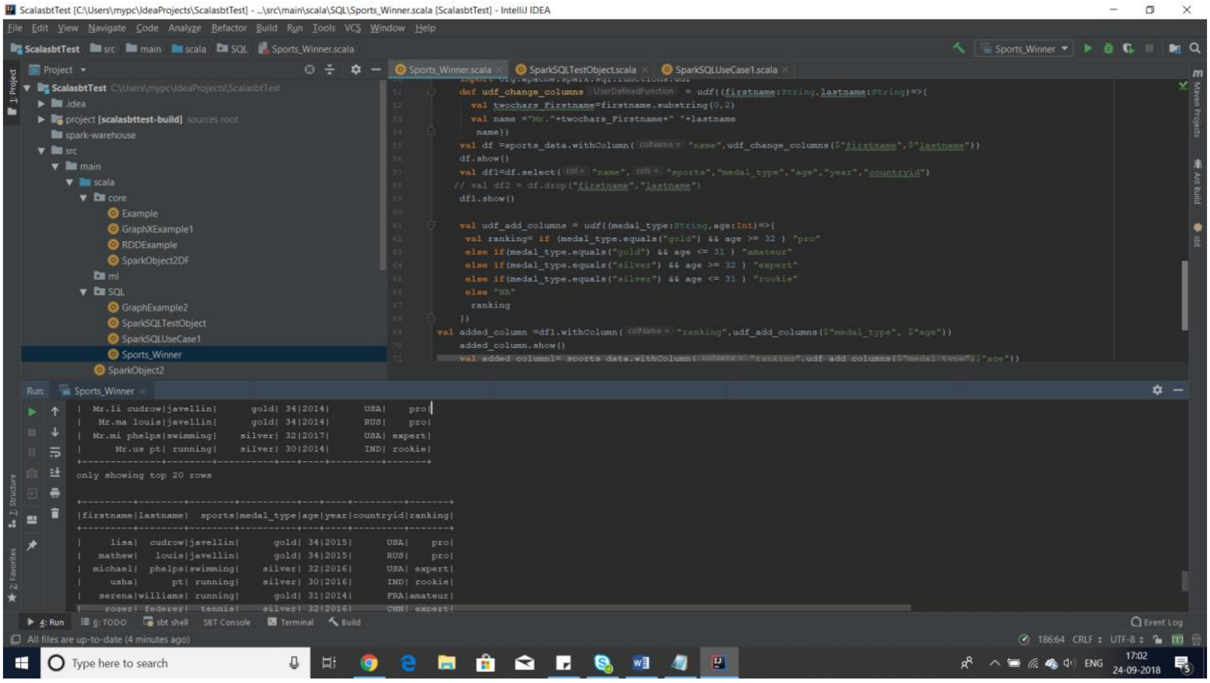
```
|  Mr.li cudrow|javellin|      gold| 34|2014|       USA|    pro|
|   Mr.ma louis|javellin|      gold| 34|2014|       RUS|    pro|
```



```
1. Change firstname, lastname columns into
Mr.first_two_letters_of_firstname<space>lastname
for example - michael, phelps becomes Mr.mi phelps*/
    import org.apache.spark.sql.functions.udf
    def udf_change_columns UserDefinedFunction  = udf((firstname:String,lastname:String)=>{
      val twochars_Firstname=firstname.substring(0,2)
      val name ="Mr."+twochars_Firstname+" "+lastname
       name})
    val df =sports_data.withColumn( colName = "name",udf_change_columns($"firstname",$"lastname"))
    df.show()
    val df1=df.select( col = "name", cols = "sports","medal_type","age","year","countryid")
   // val df2 = df.drop("firstname","lastname")
    df1.show()

    val udf_add_columns = udf((medal_type:String,age:Int)=>{
     val ranking= if (medal_type.equals("gold") && age >= 32 ) "pro"
     else if(medal_type.equals("gold") && age <= 31 ) "amateur"
     else if(medal_type.equals("silver") && age >= 32 ) "expert"
     else if(medal_type.equals("silver") && age <= 31 ) "rookie"
     else "NA"
      ranking
    })
val added_column =df1.withColumn( colName = "ranking",udf_add_columns($"medal_type", $"age"))
    added_column.show()
    val added_column1= sports_data.withColumn( colName = "ranking",udf_add_columns($"medal_type",$"age"))
    added_column1.show()
  })
```

```
|  Mr.li cudrow|javellin|      gold| 34|2014|       USA|    pro|
|   Mr.ma louis|javellin|      gold| 34|2014|       RUS|    pro|
|  Mr.mi phelps|swimming|    silver| 32|2017|       USA| expert|
|      Mr.us pt| running|    silver| 30|2014|       IND| rookie|
+--------------+--------+----------+---+----+----------+-------+
only showing top 20 rows
```

All files are up-to-date (3 minutes ago)



```
    def udf_change_columns UserDefinedFunction  = udf((firstname:String,lastname:String)=>{
      val twochars_Firstname=firstname.substring(0,2)
      val name ="Mr."+twochars_Firstname+" "+lastname
       name})
    val df =sports_data.withColumn( colName = "name",udf_change_columns($"firstname",$"lastname"))
    df.show()
    val df1=df.select( col = "name", cols = "sports","medal_type","age","year","countryid")
   // val df2 = df.drop("firstname","lastname")
    df1.show()

    val udf_add_columns = udf((medal_type:String,age:Int)=>{
     val ranking= if (medal_type.equals("gold") && age >= 32 ) "pro"
     else if(medal_type.equals("gold") && age <= 31 ) "amateur"
     else if(medal_type.equals("silver") && age >= 32 ) "expert"
     else if(medal_type.equals("silver") && age <= 31 ) "rookie"
     else "NA"
      ranking
    })
val added_column =df1.withColumn( colName = "ranking",udf_add_columns($"medal_type", $"age"))
    added_column.show()
    val added_column1= sports_data.withColumn( colName = "ranking",udf_add_columns($"medal_type",$"age"))
```

```
|  Mr.li cudrow|javellin|      gold| 34|2014|       USA|    pro|
|   Mr.ma louis|javellin|      gold| 34|2014|       RUS|    pro|
|  Mr.mi phelps|swimming|    silver| 32|2017|       USA| expert|
|      Mr.us pt| running|    silver| 30|2014|       IND| rookie|
+--------------+--------+----------+---+----+----------+-------+
only showing top 20 rows
```

```
+---------+--------+--------+----------+---+----+---------+-------+
|firstname|lastname|  sports|medal_type|age|year|countryid|ranking|
+---------+--------+--------+----------+---+----+---------+-------+
|     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|
```

All files are up-to-date (4 minutes ago)