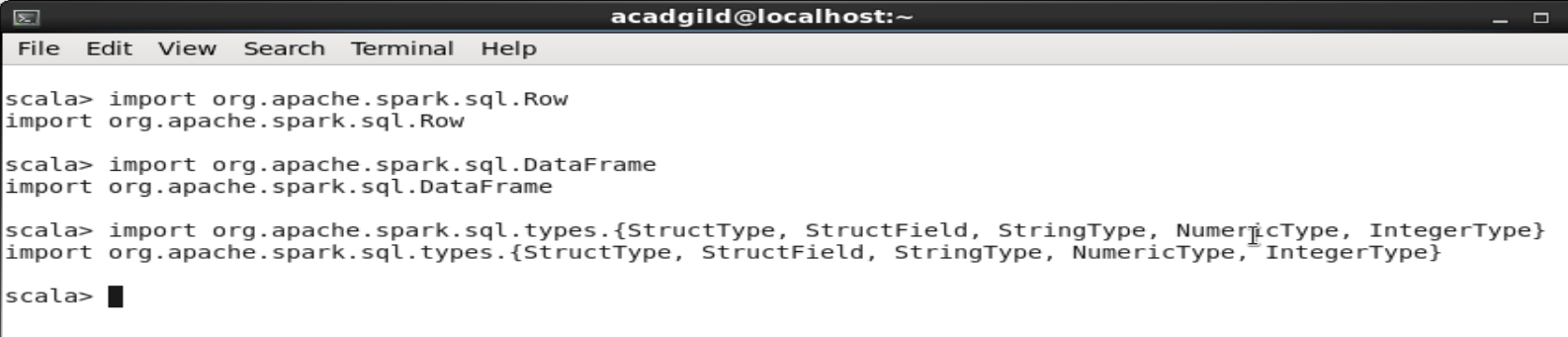
**Session 19 Assignment 2**

**// Importing API**

scala> import org.apache.spark.sql.Row  
import org.apache.spark.sql.Row  
  
scala> import org.apache.spark.sql.DataFrame  
import org.apache.spark.sql.DataFrame  
  
scala> import org.apache.spark.sql.types.{StructType, StructField, StringType, NumericType, IntegerType}  
import org.apache.spark.sql.types.{StructType, StructField, StringType, NumericType, IntegerType}



**// Importing data**

var sports\_data = sc.textFile("/home/acadgild/Downloads/Sports\_data.txt")

**// Considering main data excluding header row**

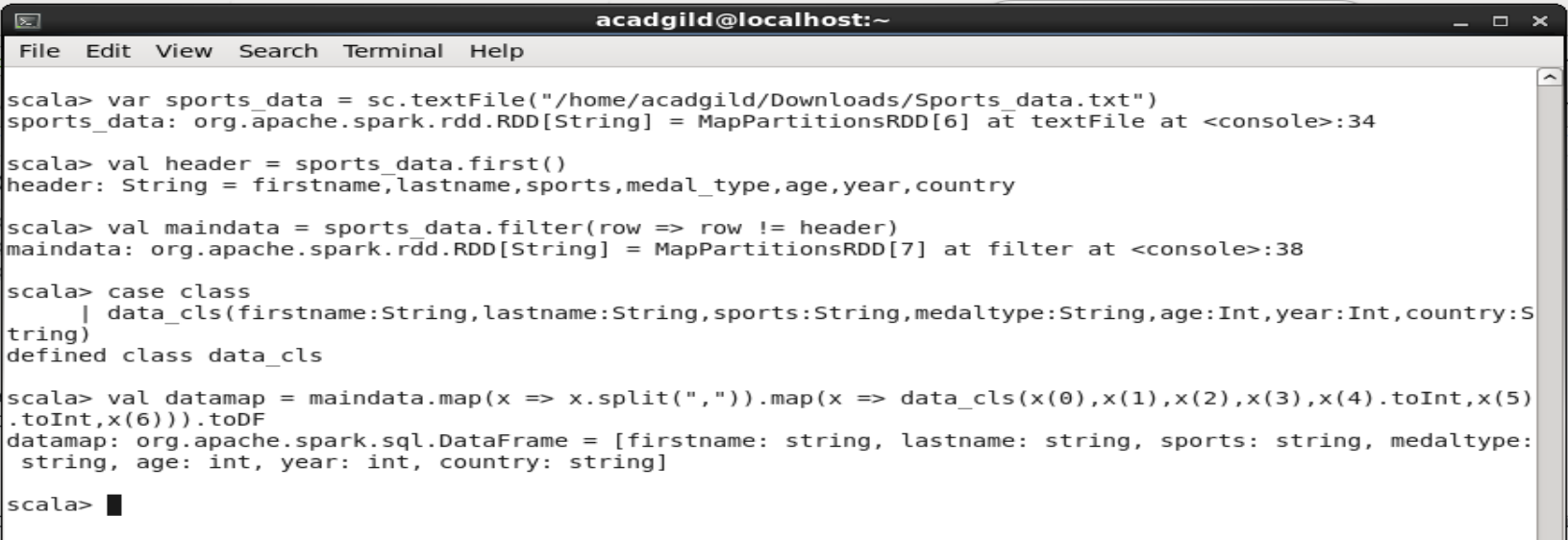
val header = sports\_data.first()

val maindata = sports\_data.filter(row => row != header)

**// Creating Case class and Data frame**

case class data\_cls(firstname:String,lastname:String,sports:String,medaltype:String,age:Int,year:Int,country:String)

val datamap = maindata.map(x => x.split(",")).map(x => data\_cls(x(0),x(1),x(2),x(3),x(4).toInt,x(5).toInt,x(6))).toDF



* **Task – 1: New Name format**

scala> def newName = udf((firstname:String, lastname:String) => {"Mr." + firstname + " " + lastname})

newName: org.apache.spark.sql.UserDefinedFunction

scala> val nameChange = datamap.withColumn("name", newName(datamap("firstname").substr(1,2),datamap("lastname"))).drop(datamap("firstname")).drop(datamap("lastname"))

nameChange: org.apache.spark.sql.DataFrame = [sports: string, medaltype: string, age: int, year: int, country: string, name: string]

scala> nameChange.show()

+--------+---------+---+----+-------+--------------+

| sports|medaltype|age|year|country| name|

+--------+---------+---+----+-------+--------------+

|javellin| gold| 34|2015| USA| Mr.li cudrow|

|javellin| gold| 34|2015| RUS| Mr.ma louis|

|swimming| silver| 32|2016| USA| Mr.mi phelps|

| running| silver| 30|2016| IND| Mr.us pt|

| running| gold| 31|2014| FRA|Mr.se williams|

| tennis| silver| 32|2016| CHN| Mr.ro federer|

|swimming| silver| 32|2014| IND| Mr.je cox|

|swimming| silver| 32|2016| CHN| Mr.fe johnson|

|javellin| gold| 34|2017| USA| Mr.li cudrow|

|javellin| gold| 34|2015| RUS| Mr.ma louis|

|swimming| silver| 32|2017| USA| Mr.mi phelps|

| running| silver| 30|2014| IND| Mr.us pt|

| running| gold| 31|2016| FRA|Mr.se williams|

| tennis| silver| 32|2017| CHN| Mr.ro federer|

|swimming| silver| 32|2014| IND| Mr.je cox|

|swimming| silver| 32|2017| CHN| Mr.fe johnson|

|javellin| gold| 34|2014| USA| Mr.li cudrow|

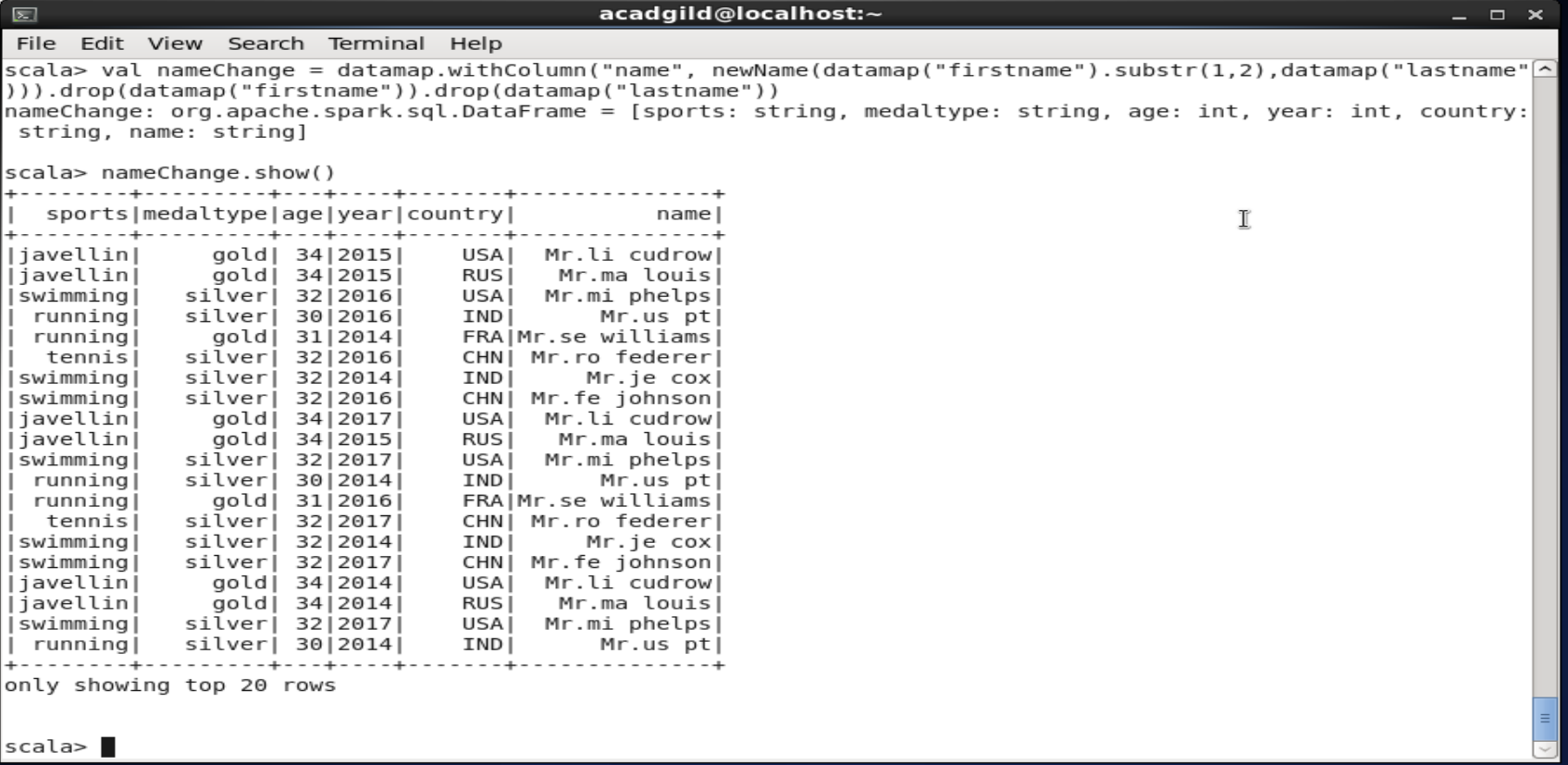
|javellin| gold| 34|2014| RUS| Mr.ma louis|

|swimming| silver| 32|2017| USA| Mr.mi phelps|

| running| silver| 30|2014| IND| Mr.us pt|

+--------+---------+---+----+-------+--------------+

only showing top 20 rows



* **Task – 2: New ranking column**

scala> def rankUDF = udf((medaltype:String,age:Int) => {

| if (medaltype == "gold" && age >=32) "pro"

| else if (medaltype == "gold" && age <=31) "amateur"

| else if (medaltype == "silver" && age >=32) "expert"

| else if (medaltype == "silver" && age <=31) "rookie"

| else ""

| })

rankUDF: org.apache.spark.sql.UserDefinedFunction

scala> val newRanking = datamap.withColumn("ranking", rankUDF(datamap("medaltype"),datamap("age")))

newRanking: org.apache.spark.sql.DataFrame = [firstname: string, lastname: string, sports: string, medaltype: string, age: int, year: int, country: string, ranking: string]

scala> newRanking.show()

+---------+--------+--------+---------+---+----+-------+-------+

|firstname|lastname| sports|medaltype|age|year|country|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

