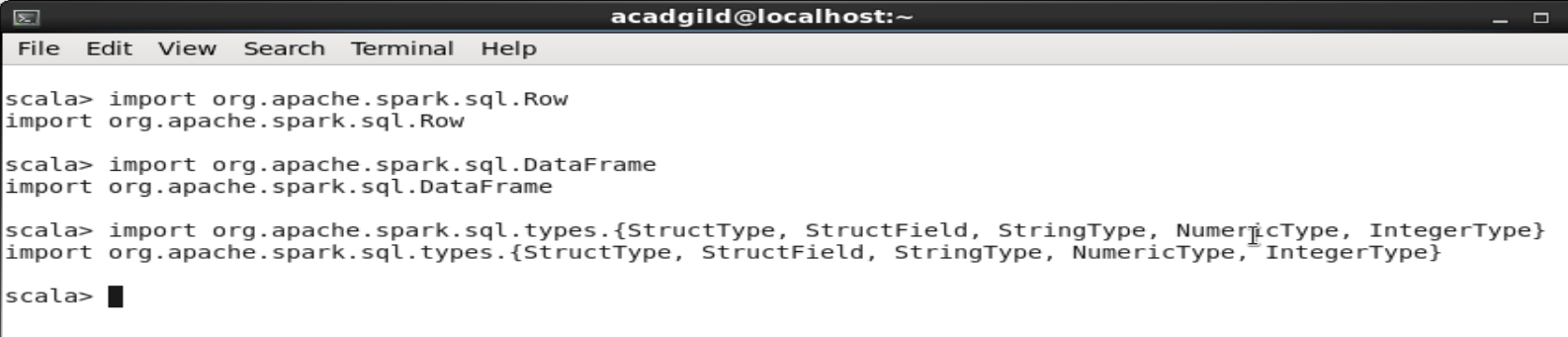
**Session 19 Assignment 1**

**// Importing d**

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**

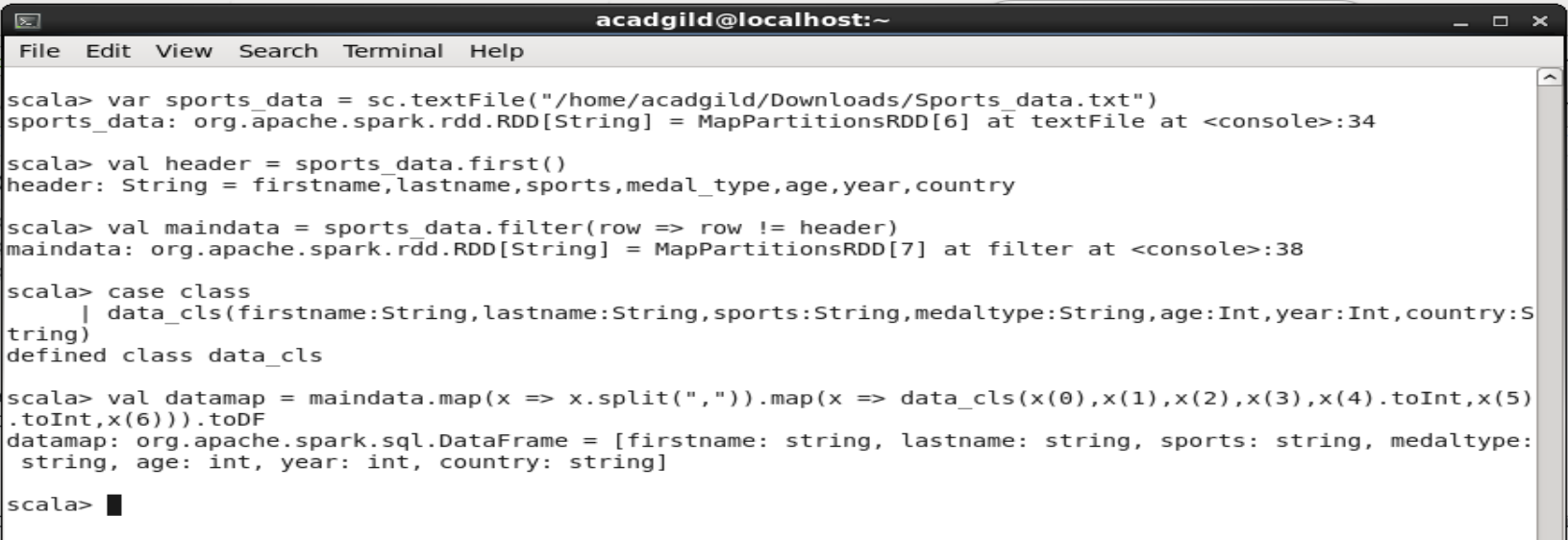
val header = sports\_data.first()

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)

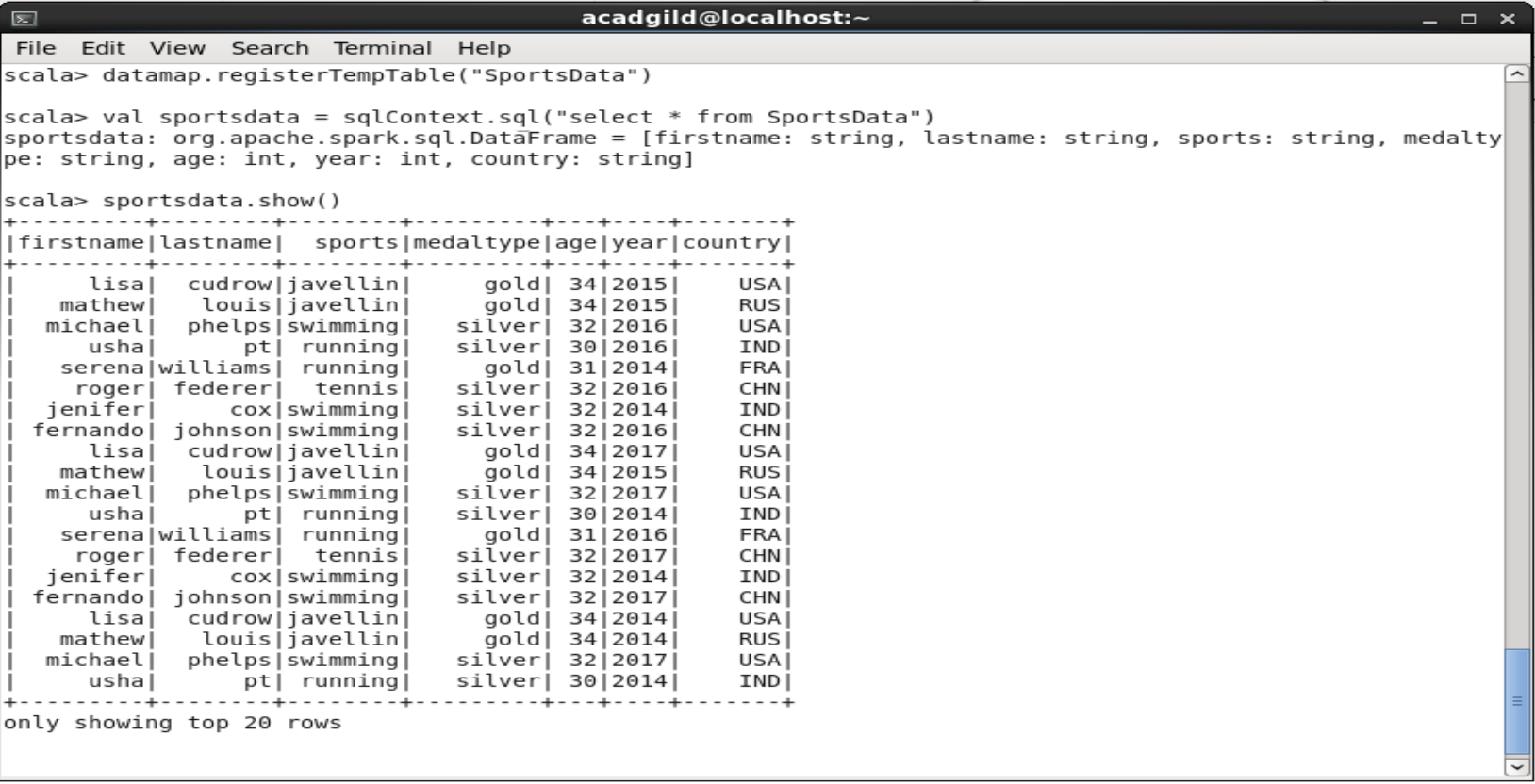
val datamap = maindata.map(x => x.split(",")).map(x => data\_cls(x(0),x(1),x(2),x(3),x(4).toInt,x(5).toInt,x(6))).toDF



**// creating a temp table and assigning select query result to variable**

datamap.registerTempTable("SportsData")

val sportsdata = sqlContext.sql("select \* from SportsData")



* **Task – 1: Total number of gold medal winners every year**

scala> val totGoldPerYr = sqlContext.sql("select year,medaltype,count(medaltype) as Total\_Gold\_Medals from SportsData where medaltype = 'gold' group by year,medaltype")

totGoldPerYr: org.apache.spark.sql.DataFrame = [year: int, medaltype: string, Total\_Gold\_Medals: bigint]

scala> totGoldPerYr.show()

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

|year|medaltype|Total\_Gold\_Medals|

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

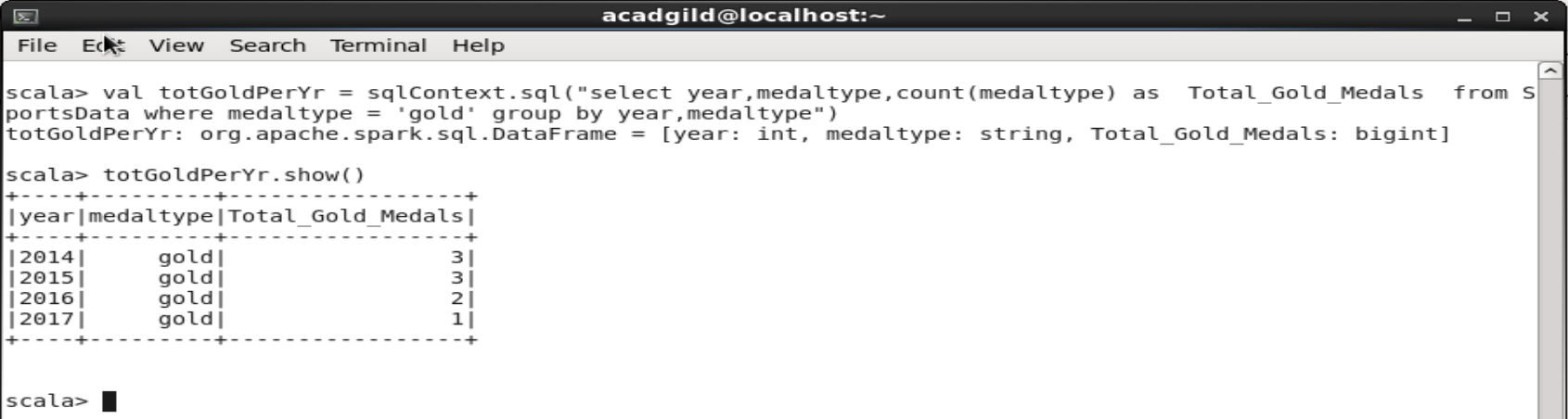
|2014| gold| 3|

|2015| gold| 3|

|2016| gold| 2|

|2017| gold| 1|

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



* **Task – 2: Total Silver medals won by USA in each sport**

scala> val totSilverUSAPerSport = sqlContext.sql("select country,sports,medaltype,count(medaltype) as Total\_Silver\_Medals\_USA\_Sport from SportsData where country = 'USA' and medaltype = 'silver' group by country,sports,medaltype")

totSilverUSAPerSport: org.apache.spark.sql.DataFrame = [country: string, sports: string, medaltype: string, Total\_Silver\_Medals\_USA\_Sport: bigint]

scala> totSilverUSAPerSport.show()

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

|country| sports|medaltype|Total\_Silver\_Medals\_USA\_Sport|

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

| USA|swimming| silver| 3|

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

