Assignment 19.1

Initial Steps:

* Load the Dataset Sport\_data.txt into RDD sports\_data\_with\_header
* Get the header from the sports\_data\_with\_header
* Remove the header from sports\_data\_with\_header by filtering and create RDD sports\_data
* Create class SportData with schema firstname, lastname,sports,medal\_type,age,year,country
* Splitting the fields of sports\_data records base on field separator comma (,) Populate SportsData class and convert to Data Frame sports\_data\_df
* Create temporary table SportsData based on sports\_data\_df

Code is as below:

val sports\_data\_with\_header = sc.textFile("/home/acadgild/assignment\_19.1/Sports\_data.txt")

val header = sports\_data\_with\_header.first()

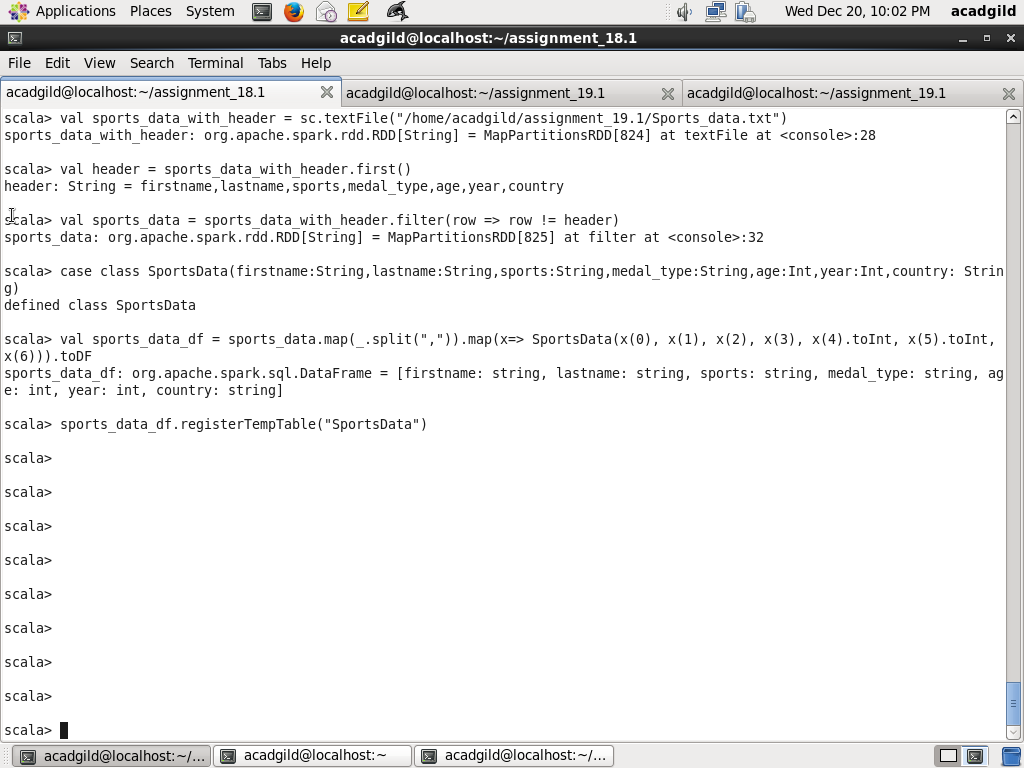
val sports\_data = sports\_data\_with\_header.filter(row => row != header)

case class SportsData(firstname:String,lastname:String,sports:String,medal\_type:String,age:Int,year:Int,country: String)

val sports\_data\_df = sports\_data.map(\_.split(",")).map(x=> SportsData(x(0), x(1), x(2), x(3), x(4).toInt, x(5).toInt, x(6))).toDF

sports\_data\_df.registerTempTable("SportsData")

Screenshot is as below:



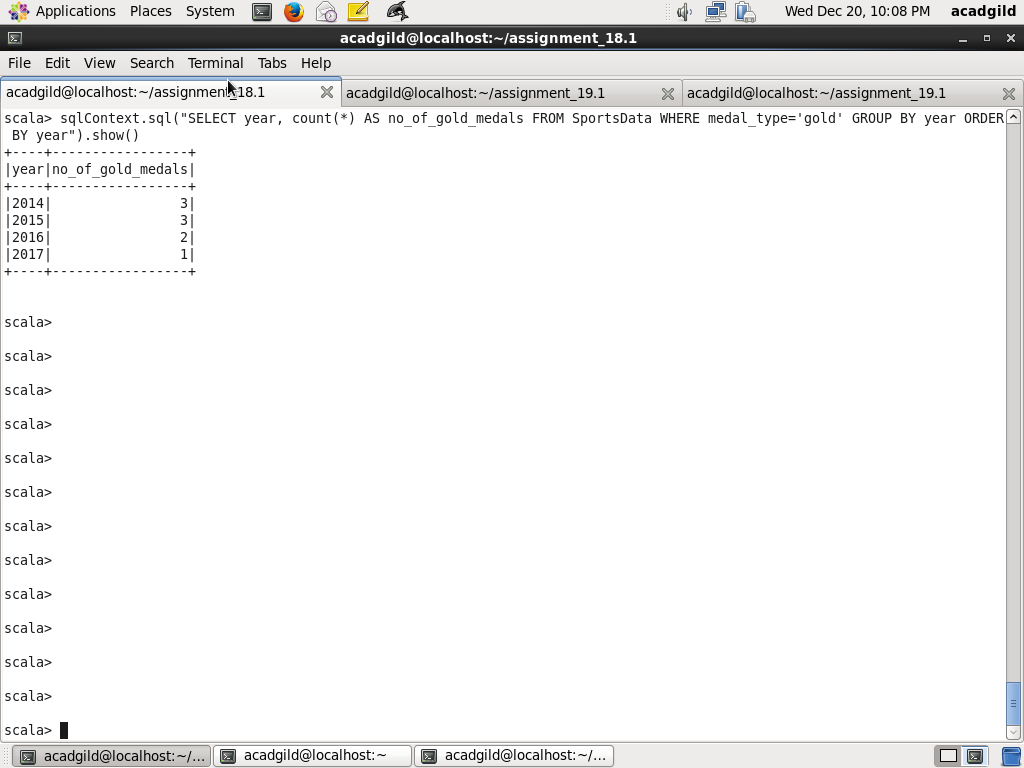
Task1: What are the total number of gold medal winners every year

Using count function on SportsData table, select total number of gold medal by having query criteria medal\_type as gold and group by year

Code is as below:

sqlContext.sql("SELECT year, count(\*) AS no\_of\_gold\_medals FROM SportsData WHERE medal\_type='gold' GROUP BY year ORDER BY year").show()

Screenshot is as below:



Task2: How many silver medal won by US in each sports

Using count function on SportsData table, select total number of silver medal by having query criteria medal\_type as gold, country as USA and group by sports

Code is as below:

sqlContext.sql("SELECT sports, count(\*) AS no\_of\_silver\_medals FROM SportsData WHERE country='USA' and medal\_type='silver' GROUP BY sports ORDER BY sports").show()

Screenshot is as below:

