Dataset link

https://drive.google.com/drive/folders/0B\_P3pWagdIrrTEVTVE1VWXI0YVk

5. Problem Statement

Using spark-sql, Find:

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

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

*val rowsRDD = sc.textFile("file:///home/cloudera/sports\_data.txt")*

*val sportsRDD = rowsRDD.map{row => row.split(",")}.map{cols => sports(cols(0), cols(1), cols(2), cols(3), cols(4), cols(5), cols(6) )}*

*val sportsDF = sportsRDD.toDF()*

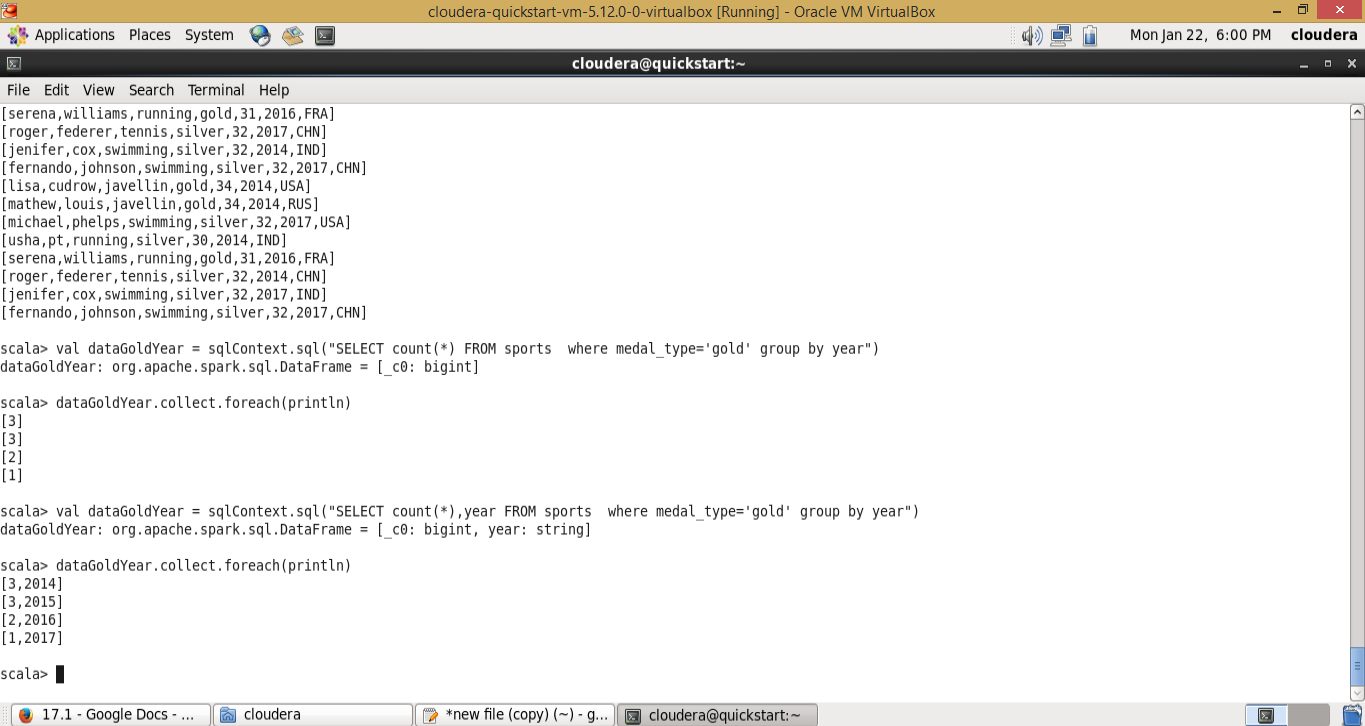
*sportsDF.registerTempTable("sports")*

*val data = sqlContext.sql( "SELECT \* FROM sports")*

*data.collect.foreach(println)*

*val dataGoldYear = sqlContext.sql("SELECT count(\*),year FROM sports where medal\_type='gold' group by year")*

*dataGoldYear.collect.foreach(println)*



1. **How many silver medals have been won by USA in each sport**

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

*val rowsRDD = sc.textFile("file:///home/cloudera/sports\_data.txt")*

*val sportsRDD = rowsRDD.map{row => row.split(",")}.map{cols => sports(cols(0), cols(1), cols(2), cols(3), cols(4), cols(5), cols(6) )}*

*val sportsDF = sportsRDD.toDF()*

*sportsDF.registerTempTable("sports")*

*val data = sqlContext.sql( "SELECT \* FROM sports")*

*data.collect.foreach(println)*

*val dataSilverSport = sqlContext.sql("SELECT count(\*),sports FROM sports where medal\_type='silver' and country='USA' group by sports")*

*dataSilverSport.collect.foreach(println)*

