Session 18 : SPARK SQL

Assignment 1

**Problem Statement**

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

Use the dataset given below:



**Solution:-**

Starting Spark shell

Commands –

export SPARK\_HOME=/usr/local/spark/spark-2.2.0-bin-hadoop2.7

spark-shell

(using latest version of spark for running spark SQL)



Loading the data:-

**scala>** val sportsRDD = sc.textFile("Sports\_data.txt");

**scala>** val header = sportsRDD.first() ;

**scala>** val sportsRDDFiltered = sportsRDD.filter(row => row != header);



Loading the data into DataFrame

**Scala>** val sportsDF= sportsRDDFiltered.map(lines=>lines.split(",")).map(arrays => (arrays(0),arrays(1),arrays(2),arrays(3),arrays(4),arrays(5),arrays(6))).toDF("firstname","lastname","sports","medal\_type","age","year","country");

Registering DataFrame as table:

**Scala>** sportsDF.registerTempTable("sports");



Based on table “**sports**” we can find solution to the problem statements

Using spark-sql, Find:

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

**scala>** val problem1DF = spark.sql("SELECT year, count(\*) FROM sports where medal\_type='gold' group by year order by year");

**scala>** problem1DF.show;



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

**scala>** val problem2DF = spark.sql("SELECT country, sports, count(\*) FROM sports where country='USA' and medal\_type='silver' group by country,sports");

**scala>** problem2DF.show;

