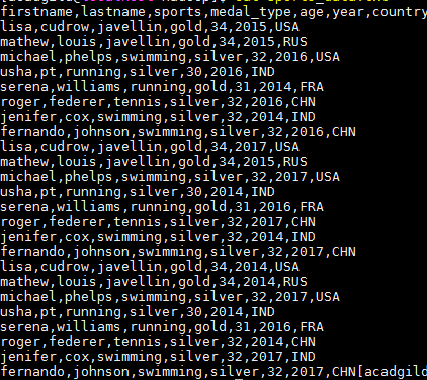
**Session 19: SPARK SQL**

Assignment 19.1

Introduction

Dataset



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

**xSpark**SQL is a Spark module for structured data processing. A **DataFrame**is a **Dataset** organized into named columns. It is conceptually equivalent to a table in a relational database or a data frame in

R/Python.**DataFrames**can be constructed from a wide array of sources such as: structured data files,

tables in Hive, external databases, or existing RDDs.

A row in **DataFrame**is represented by Row object. Row can be used to create a row object by using

named arguments, the fields will be sorted by names.

The fields in it can be accessed like attributes.

In 2.0, SparkSession is the entry point for creation of DataFrames.

Task – 1 - What are the total number of gold medal winners every year

In order to proceed we need to import some dependencies as shown below,

***importorg.apache.spark.sql.Row;***

***importorg.apache.spark.sql.types.{StructType,StructField,StringType,NumericType,IntegerType};***



Please see the codes used below

,

1. valSportsData = sc.textFile("/home/acadgild/hadoop/Sports\_data.txt")-

2. valschemaString =

"firstname:string,lastname:string,sports:string,medal\_type:string,age:string,year:string,

country:string"

3. val schema = StructType(schemaString.split(",").map(x =>

StructField(x.split(":")(0),if(x.split(":")(1).equals("string"))StringType else IntegerType, true)))

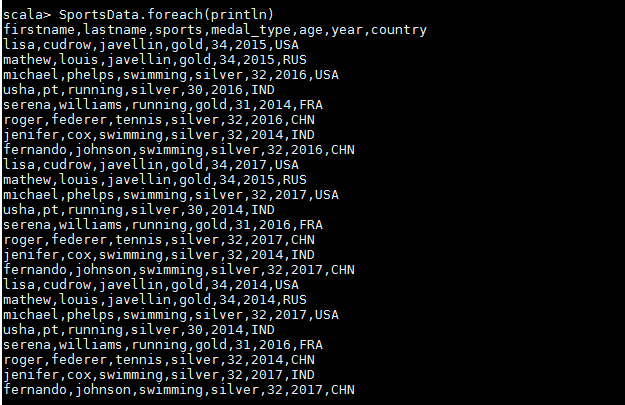
4. valrowRDD = SportsData.map(\_.split(",")).map(r => Row(r(0), r(1), r(2), r(3), r(4), r(5), r(6)))

5. valSportsDataDF = spark.createDataFrame(rowRDD, schema)

6. SportsDataDF.createOrReplaceTempView("SportsData")

7. valresultDF = spark.sql("SELECT year,COUNT (\*) FROM SportsData WHERE medal\_type = 'gold'GROUP BY year")

8. resultDF.show()

******

**Step -2** – we are defining a schema since it is a text file and splitting the input file using the delimiters and extracting the rows from it.





We have created the **dataframe**by passing the RDD which reads the file and schema to spark session

object-

The schema of the created **Dataframe**can be seen below.



Result

Now, we are using the simple SQL query so that we can execute our query by applying it on the temporary table created



Task – 2 - How many silver medals have been won by USA in each sport

Here we use the same **DataFrame**to get the desired result,

We are using the simple SQL query so that we can execute our query by applying it on the temporary table created,

The code used is,

**1. val result2DF = spark.sql("SELECT sports, COUNT (\*) FROM SportsData WHERE medal\_type ='silver' and country ='USA' GROUP BY sports")**

**2. result2DF.show()**

