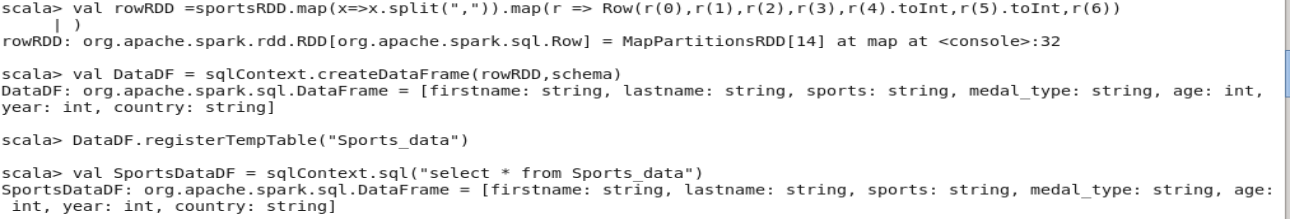
**Assignmwnt 03**

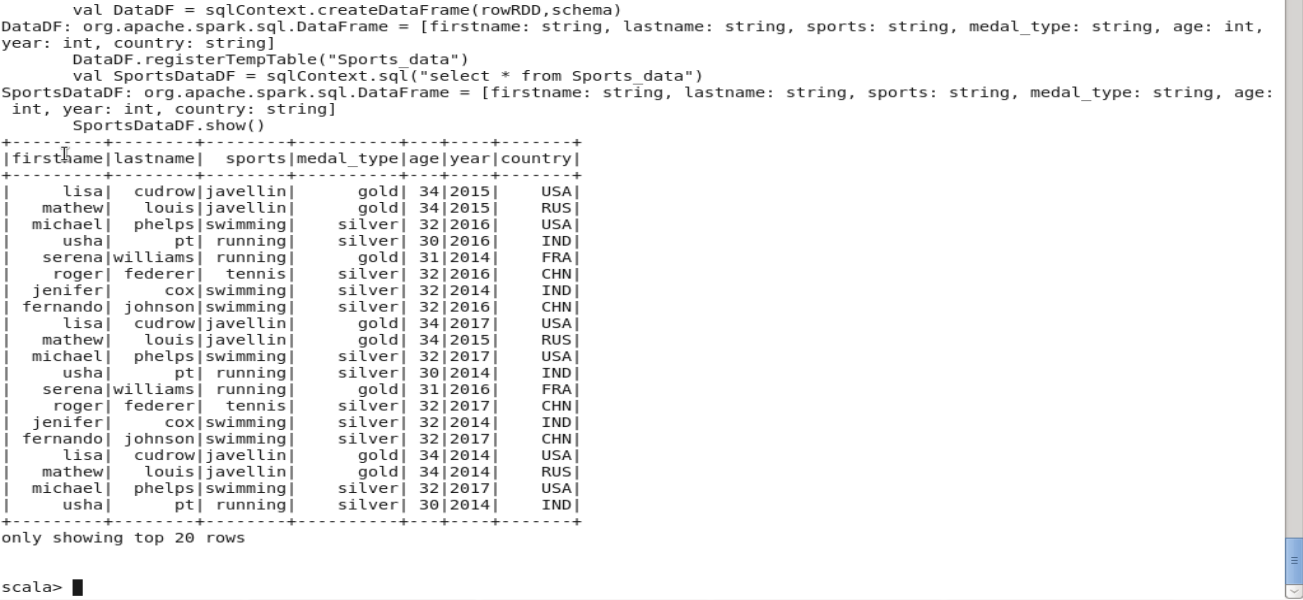
**Data Frame Creation:**

1. Import the spark sql Row, DataFrame and types objects.
2. Create an RDD that will hold the contents of the Sports Dataset(text file format) sportsRDD
3. Create the schema (column name and type) for the data. schemaString
4. Create the schema structure based on the schemaString. schema
5. Create an RDD that will split the data in sportsRDD by ‘,’ and map columns in rows as in schema. rowRDD
6. Create the DataFrame from the rowRDD and schema created before using the “sqlContext” object. SportsDataDF
7. Create a temporary table (Sports\_data) from the DataFrame SportsDataDF





To view the contents of the temp table Sports\_Data, use a sql SELECT Statement with the sqlContext



**Create a dataframe with 1 to 100 and save as parquet file.**

Create an RDD to hold numbers from 1 to 100 by using the spark context object **sc** and the **parallelize** method. pdataRDD

Create the DataFrame pDataDF from the RDD pdataRDD by using the function **toDF()**.

To save the data as a **parquet file**, use the **write** method with data type: parquet and filename as parameter: *[filename]*.parquet

To read the data that was saved as a parquet file, use the **read** method with data type: parquet and filename (given before) as parameter: *[filename]*.parquet. readFile

Then use the .show() method with readFile to display the contents of the file.

You can use .show(100, false) to view the entire data in the file.



