

CSEE5590/CS490: Big Data Programming

LAB ASSIGNMENT #2

Team 4:

Aparna Manda(11)
Lohitha Yenugu(19)

Task 1

- Implement MapReduce algorithm for finding Facebook common friends problem and run the MapReduce job on Apache Spark.

Input

The terminal window displays the following data:

| Line Number | Content |
|-------------|---------|
| 1 | 0 1 |
| 2 | 0 2 |
| 3 | 0 3 |
| 4 | 0 4 |
| 5 | 0 5 |
| 6 | 0 6 |
| 7 | 0 7 |
| 8 | 0 8 |
| 9 | 0 9 |
| 10 | 0 10 |
| 11 | 0 11 |
| 12 | 0 12 |
| 13 | 0 13 |
| 14 | 0 14 |
| 15 | 0 15 |
| 16 | 0 16 |
| 17 | 0 17 |
| 18 | 0 18 |
| 19 | 0 19 |
| 20 | 0 20 |
| 21 | 0 21 |
| 22 | 0 22 |
| 23 | 0 23 |
| 24 | 0 24 |
| 25 | 0 25 |
| 26 | 0 26 |
| 27 | 0 27 |
| 28 | 0 28 |
| 29 | 0 29 |

Code:

```
facebook_combined.txt x  facebook_friends_mutual.py x  part-00000 x
1  from builtins import print, list, len, set
2  from pyspark import SparkContext
3
4
5  def map_friends(line):
6      people = line.split(" ")
7      person = people[0]
8      friend = line[1]
9      return person, friend
10
11
12 def map_togroup(tuple):
13     person = tuple[0]
14     final = []
15     for friend in tuple[1]:
16         if person < friend:
17             key = person + "," + friend
18         else:
19             key = friend + "," + person
20         value = key, list(tuple[1])
21         final.append(value)
22
23
return final
```

```

24
25     def reduce(key, value):
26         return list(set(key) & set(value))
27
28
29     def run(input, output):
30         sc = SparkContext.getOrCreate()
31         lines = sc.textFile(input, 1)
32         mapped_friends = lines.map(map_friends).groupByKey()
33         grouped_friends = mapped_friends.flatMap(map_to_group)
34         mutual_friends = grouped_friends.reduceByKey(reduce).filter(lambda x: len(x[1]) > 0)
35         print(mutual_friends.collect())
36         mutual_friends.coalesce(1).saveAsTextFile(output)
37
38
39 ▶   if __name__ == "__main__":
40       run("facebook_combined.txt", "facebook_mutual_friends")
41

```

Output:

The screenshot shows the PyCharm IDE interface. On the left, the Project tool window displays a file structure under 'Task1' with files like 'facebook_mutual_friends', 'facebook_combined.txt', and 'facebook_friends_mutual.py'. In the center, there's a terminal window titled 'part-00000' showing the output of the script. On the right, there are tabs for 'facebook_combined.txt' and 'facebook_friends_mutual.py'.

```

part-00000
10  (' ,9', [' '])
11  ('0,10', ['0'])
12  ('13,3', ['3'])
13  ('14,4', ['4'])
14  ('16,6', ['6'])
15  ('17,7', ['7'])
16  ('19,9', ['9'])
17  ('0,20', ['0'])
18  ('1,21', ['1'])
19  ('2,22', ['2'])
20  ('23,3', ['3'])

```

Task 2

- Create a Spark DataFrame
- Perform 10 intuitive queries in Dataset
- Using 5 queries compare between Spark RDD's and Spark DataFrames.

10 Queries

1. The country which was winner for highest number of times.
2. The year in which there was maximum craze and attendance for the games.
3. Home Team Goals and Away Team Goals for each stage
4. Referees with highest number of matches
5. Home Team Goals for Winners
6. Countries that hosted highest world cup matches
7. No of matches in year 1934 and in Estadio Centenario stadium
8. Captains of Teams of Final Matches for all Years
9. Number of matches of referee Macias Jose
10. Display all Goal-Keeper Names

5 Queries Comparing RDD's and DF's

1. Winner of 1930
2. Years Uruguay was winner
3. Display Year, Country when Attendance was greater than 500.00
4. Display Year Country Winner where Hosted Country was the Winner
5. Display Year, Winner, Runners-Up, Third, Fourth where USA was in Winner, Runner-Up, Third, Fourth

Input

WorldCups.csv as WC View

IntelliJ IDEA Project View showing the Task2 and Task4 modules. The Task2 module contains a Task2.scala file and a WorldCupMatches.csv file. The Task4 module contains a Task4.scala file and a WorldCupPlayers.csv file. The WorldCupMatches.csv file contains the following data:

```

1 Year,Country,Winner,Runners-Up,Third,Fourth,GoalsScored,QualifiedTeams,MatchesPlayed,Attendance
2 1930,Uruguay,Uruguay,Argentina,USA,Yugoslavia,70,13,18,590.549
3 1934,Italy,Italy,Czechoslovakia,Germany,Austria,70,16,17,363.000
4 1938,France,Italy,Hungary,Brazil,Sweden,84,15,18,375.700
5 1950,Brazil,Uruguay,Brazil,Sweden,Spain,88,13,22,1.045.246
6 1954,Switzerland,Germany FR,Hungary,Austria,Uruguay,140,16,26,768.607
7 1958,Sweden,Brazil,Sweden,France,Germany FR,126,16,35,819.810
8 1962,Chile,Brazil,Czechoslovakia,Chile,Yugoslavia,89,16,32,893.172
9 1966,England,England,Germany FR,Portugal,Soviet Union,89,16,32,1.563.135
10 1970,Mexico,Brazil,Italy,Germany FR,Uruguay,95,16,32,1.683.975
11 1974,Germany,Germany FR,Netherlands,Poland,Brazil,97,16,38,1.865.753
12 1978,Argentina,Argentina,Netherlands,Brazil,Italy,102,16,38,1.545.791
13 1982,Spain,Italy,Germany FR,Poland,France,146,24,52,2.109.723
14 1986,Mexico,Argentina,Germany FR,France,Belgium,132,24,52,2.394.031
15 1990,Italy,Germany FR,Argentina,Italy,England,115,24,52,2.516.215
16 1994,USA,Brazil,Italy,Sweden,Bulgaria,141,24,52,3.587.538
17 1998,France,France,Brazil,Croatia,Netherlands,171,32,64,2.785.100
18 2002,Korea/Japan,Brazil,Germany,Turkey,Korea Republic,161,32,64,2.705.197
19 2006,Germany,Italy,France,Germany,Portugal,147,32,64,3.359.439
20 2010,South Africa,Spain,Netherlands,Germany,Uruguay,145,32,64,3.178.856
21 2014,Brazil,Germany,Argentina,Netherlands,Brazil,171,32,64,3.386.810
22

```

WorldCupMatches.csv as Matches View

IntelliJ IDEA Project View showing the Task2 and Task4 modules. The Task2 module contains a Task2.scala file and a WorldCupMatches.csv file. The Task4 module contains a Task4.scala file and a meta-groups.csv file. The WorldCupMatches.csv file contains the following data:

```

1 Year,Datetime,Stage,Stadium,City,Home Team Name,Home Team Goals,Away Team Name,Away Team Goals,Win conditions,Attendance,Half-time Home Goals,Half-time Away Goals
2 1930,"13 Jul 1930 - 15:00 ",Group 1,Pocitos,"Montevideo ",France,4,1,Mexico," ",4444,3,0,LOMBARDI Domingo (URU),CRISTOPHE Henry (BEL),REGO Gilberto (BRA),SAUCEDO Ulises (BOL),RADULESCU Ionut (ROM),VALLARINO Ricardo (URU),BALWAY Thomas (FRA),CRISTOPHE He Validation
3 1930,"13 Jul 1930 - 15:00 ",Group 4,Parque Central,"Montevideo ",USA,3,0,Belgium," ",18346,2,0,MACIAS Jose (ARG),MATEUCCI Francisco (URU),WARNKEN Alberto (CHI),VALLARINO Ricardo (URU),BALWAY Thomas (FRA),CRISTOPHE He Validation
4 1930,"14 Jul 1930 - 12:45 ",Group 2,Parque Central,"Montevideo ",Yugoslavia,2,1,Brazil," ",24059,2,0,TEJADA Anibal (URU),VALLARINO Ricardo (URU),BALWAY Thomas (FRA),CRISTOPHE He Validation
5 1930,"14 Jul 1930 - 14:50 ",Group 3,Pocitos,"Montevideo ",Romania,3,1,Peru," ",25491,1,0,WARNKEN Alberto (CHI),LANGENUS Jean (BEL),MATEUCCI Francisco (URU),VALLARINO Ricardo (URU),BALWAY Thomas (FRA),CRISTOPHE He Validation
6 1930,"15 Jul 1930 - 16:00 ",Group 1,Parque Central,"Montevideo ",Argentina,1,0,France," ",23409,0,0,REGO Gilberto (BRA),SAUCEDO Ulises (BOL),RADULESCU Ionut (ROM),VALLARINO Ricardo (URU),BALWAY Thomas (FRA),CRISTOPHE He Validation
7 1930,"16 Jul 1930 - 14:45 ",Group 1,Parque Central,"Montevideo ",Chile,3,0,Mexico," ",9249,1,0,CRISTOPHE Henry (BEL),APHESTEGUY Martin (URU),LANGENUS Jean (BEL),VALLARINO Ricardo (URU),BALWAY Thomas (FRA),CRISTOPHE He Validation
8 1930,"17 Jul 1930 - 12:45 ",Group 2,Parque Central,"Montevideo ",Yugoslavia,4,0,Bolivia," ",18306,0,0,MATEUCCI Francisco (URU),LOMBARDI Domingo (URU),WAHLER Max (GER),VALLARINO Ricardo (URU),TEJADA Anibal (URU),BALWAY Thomas (FRA),CRISTOPHE He Validation
9 1930,"17 Jul 1930 - 14:45 ",Group 4,Parque Central,"Montevideo ",USA,3,0,Paraguay," ",18306,2,0,MACIAS Jose (ARG),APHESTEGUY Martin (URU),TEJADA Anibal (URU),VALLARINO Ricardo (URU),BALWAY Thomas (FRA),CRISTOPHE He Validation
10 1930,"18 Jul 1930 - 14:30 ",Group 3,Estadio Centenario,"Montevideo ",Uruguay,1,0,Peru," ",57735,0,0,LANGENUS Jean (BEL),BALWAY Thomas (FRA),CRISTOPHE He Validation
11 1930,"19 Jul 1930 - 12:50 ",Group 1,Estadio Centenario,"Montevideo ",Chile,1,0,France," ",2000,0,0,TEJADA Anibal (URU),LOMBARDI Domingo (URU),REGO Gilberto (BRA),VALLARINO Ricardo (URU),BALWAY Thomas (FRA),CRISTOPHE He Validation
12 1930,"19 Jul 1930 - 15:00 ",Group 1,Estadio Centenario,"Montevideo ",Argentina,6,3,Mexico," ",42180,3,1,SAUCEDO Ulises (BOL),ALONSO Gualberto (URU),RADULESCU Ionut (ROM),VALLARINO Ricardo (URU),BALWAY Thomas (FRA),CRISTOPHE He Validation
13 1930,"20 Jul 1930 - 13:00 ",Group 2,Estadio Centenario,"Montevideo ",Brazil,4,0,Bolivia," ",25466,1,0,BALWAY Thomas (FRA),MATEUCCI Francisco (URU),VALLEJO Gaspar (MEX),VALLARINO Ricardo (URU),BALWAY Thomas (FRA),CRISTOPHE He Validation
14 1930,"20 Jul 1930 - 15:00 ",Group 4,Estadio Centenario,"Montevideo ",Paraguay,1,0,Belgium," ",12000,1,0,VALLARINO Ricardo (URU),MACIAS Jose (ARG),LOMBARDI Domingo (URU),WAHLER Max (GER),VALLARINO Ricardo (URU),TEJADA Anibal (URU),BALWAY Thomas (FRA),CRISTOPHE He Validation
15 1930,"21 Jul 1930 - 14:50 ",Group 3,Estadio Centenario,"Montevideo ",Uruguay,4,0,Romania," ",70022,4,0,REGO Gilberto (BRA),WARNKEN Alberto (CHI),SAUCEDO Ulises (BOL),ALONSO Gualberto (URU),RADULESCU Ionut (ROM),VALLARINO Ricardo (URU),TEJADA Anibal (URU),BALWAY Thomas (FRA),CRISTOPHE He Validation
16 1930,"22 Jul 1930 - 14:45 ",Group 1,Estadio Centenario,"Montevideo ",Argentina,3,1,Chile," ",41459,2,1,LANGENUS Jean (BEL),CRISTOPHE Henry (BEL),SAUCEDO Ulises (BOL),ALONSO Gualberto (URU),RADULESCU Ionut (ROM),VALLARINO Ricardo (URU),TEJADA Anibal (URU),BALWAY Thomas (FRA),CRISTOPHE He Validation
17 1930,"26 Jul 1930 - 14:45 ",Semi-Finals,Estadio Centenario,"Montevideo ",Argentina,6,1,USA," ",72886,1,0,LANGENUS Jean (BEL),VALLEJO Gaspar (MEX),WARNKEN Alberto (CHI),TEJADA Anibal (URU),BALWAY Thomas (FRA),CRISTOPHE He Validation
18 1930,"27 Jul 1930 - 14:45 ",Semi-Finals,Estadio Centenario,"Montevideo ",Uruguay,6,1,Yugoslavia," ",79867,3,Type: In word 'LANGENUS'
19 1930,"30 Jul 1930 - 14:15 ",Final,Estadio Centenario,"Montevideo ",Uruguay,6,1,Yugoslavia," ",68346,1,2,LANGENUS Jean (BEL),VALLEJO Gaspar (MEX),WARNKEN Alberto (CHI),TEJADA Anibal (URU),BALWAY Thomas (FRA),CRISTOPHE He Validation
20 1934,"27 May 1934 - 16:30 ",Preliminary round,Stadio Benito Mussolini,Turin," ",Austria,3,2,France,Austria," ",68346,1,2,LANGENUS Jean (BEL),VALLEJO Gaspar (MEX),WARNKEN Alberto (CHI),TEJADA Anibal (URU),BALWAY Thomas (FRA),CRISTOPHE He Validation
21 1934,"27 May 1934 - 16:30 ",Preliminary round,Giorgio Ascarelli,Naples," ",Hungary,4,2,Egypt," ",9800,2,2,BARLASSINA Rinaldo (ITA),DATTILO Generoso (ITA),VALLARINO Ricardo (URU),BALWAY Thomas (FRA),CRISTOPHE He Validation
22 1934,"27 May 1934 - 16:30 ",Preliminary round,San Siro,Milan," ",Switzerland,3,2,Netherlands," ",33008,2,1,EKLIND Ivan (SWE),BERANEK Alois (AUT),BONIVENTURE Georges (FRA),VALLARINO Ricardo (URU),TEJADA Anibal (URU),BALWAY Thomas (FRA),CRISTOPHE He Validation
23 1934,"27 May 1934 - 16:30 ",Preliminary round,Littoriale,Bologna," ",Sweden,3,2,Argentina," ",14000,1,2,BRAUN Eugen (AUT),CARRARO Albino (ITA),TURBIANI Giacomo (ITA),VALLARINO Ricardo (URU),TEJADA Anibal (URU),BALWAY Thomas (FRA),CRISTOPHE He Validation
24 1934,"27 May 1934 - 16:30 ",Preliminary round,Giovanni Berta,Florence," ",Germany,5,2,Belgium," ",8000,1,2,MATTEA Francesco (ITA),MELANDRI Ermengildo (ITA),VALLARINO Ricardo (URU),TEJADA Anibal (URU),BALWAY Thomas (FRA),CRISTOPHE He Validation
25 1934,"27 May 1934 - 16:30 ",Preliminary round,Luigi Ferraris,Genoa," ",Spain,3,1,Brazil," ",21000,3,0,BIRLEM Alfred (GER),CARMINATI Ettore (ITA),IVANCSIC János (HUN),VALLARINO Ricardo (URU),TEJADA Anibal (URU),BALWAY Thomas (FRA),CRISTOPHE He Validation
26 1934,"27 May 1934 - 16:30 ",Preliminary round,Nazionale PNF,Rome," ",Italy,7,1,USA," ",25000,3,0,MERCET Rene (SUI),ESCATIN Pedro (ESP),ZENISEK Bohumil (CZE),VALLARINO Ricardo (URU),TEJADA Anibal (URU),BALWAY Thomas (FRA),CRISTOPHE He Validation

```

WorldCupPlayers.csv as Players View

The screenshot shows the IntelliJ IDEA interface with the following details:

- File Menu:** File, Edit, View, Navigate, Code, Analyze, Refactor, Build, Run, Tools, VCS, Window, Help.
- Project Bar:** Task2 and Task4 [Task2], fifa-world-cup, WorldCupPlayers.csv.
- Toolbars:** Standard toolbar with icons for Open, Save, Run, etc.
- Code Editor:** The main window displays the content of `WorldCupPlayers.csv`. The first few lines of the CSV data are:

```
1 RoundID,MatchID,Team Initials,Coach Name,Line up,Shirt Number,Player Name,Position,Event
2 201,1096,FRA,CAUDRON Raoul (FRA),S,0,Alex THEPOT,GK,
3 201,1096,MEX,LUQUE Juan (MEX),S,0,Oscar BONFIGLIO,GK,
4 201,1096,FRA,CAUDRON Raoul (FRA),S,0,Marcel LANGILLER,,G40'
5 201,1096,MEX,LUQUE Juan (MEX),S,0,Juan CARRENO,,G70'
6 201,1096,FRA,CAUDRON Raoul (FRA),S,0,Ernest LIBERATI,,
```
- Sidebar:** Shows the project structure with `fifa-world-cup`, `nashville-meetup`, and `project [Task2-build]`.
- Bottom Navigation:** Text, Data, Git, TODO, Run, Terminal, sbt shell, Build.
- Status Bar:** All files are up-to-date (10 minutes ago), 7 chars, 2:21 LF, UTF-8, 4 spaces, 5:59 PM, 5/9/2020, ENG.

Code

```
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help Task2 - Task2.scala [Task2] - IntelliJ IDEA

Task2 and Task4 src main scala Task2.scala

Project
  Task2 and Task4 [Task2]
    .idea
    fifa-world-cup
      WorldCupMatches.csv
      WorldCupPlayers.csv
      WorldCups.csv
    nashville-meetup
      group-edges.csv
      member-edges.csv
      member-to-group-edges.csv
      meta-events.csv
      meta-groups.csv
      meta-members.csv
      rsvps.csv
  project [Task2-build] sources root
    src
      main
        scala
          Task2
          Task4
    test
    target
    build.sbt
  External Libraries
  Scratches and Consoles

Task2.scala
  1 import ...
  2
  3 object Task2 {
  4
  5   def main(args: Array[String]): Unit = {
  6
  7     //Setting up the Spark Session and Spark Context
  8     val conf = new SparkConf().setMaster("local[2]").setAppName("Task2")
  9     val sc = new SparkContext(conf)
 10     val spark = SparkSession
 11       .builder()
 12       .appName( name = "Task2" )
 13       .config(conf = conf)
 14       .getOrCreate()
 15
 16     Logger.getLogger( name = "org" ).setLevel(Level.ERROR)
 17     Logger.getLogger( name = "akka" ).setLevel(Level.ERROR)
 18
 19     // We are using all 3 Fifa dataset given on Kaggle Repository
 20     //a.Import the dataset and create df and print Schema
 21
 22     val df1 = spark.read
 23       .format( source= "csv" )
 24       .option("header", "true") //reading the headers
 25       .option("mode", "DROPMALFORMED")
 26       .load( path = "fifa-world-cup/WorldCups.csv" )
 27
 28     val df2 = spark.read
 29
```

Task2 and Task4 src main scala Task2.scala

```

19 // We are using all 3 Fifa dataset given on Kaggle Repository
20 // a.Import the dataset and create df and print Schema
21
22 val df1 = spark.read
23   .format( source= "csv")
24   .option("header", "true") //reading the headers
25   .option("mode", "DROPMALFORMED")
26   .load( path= "fifa-world-cup/WorldCups.csv")
27
28 val df2 = spark.read
29   .format( source= "csv")
30   .option("header", "true") //reading the headers
31   .option("mode", "DROPMALFORMED")
32   .load( path= "fifa-world-cup/WorldCupPlayers.csv")
33
34 val df3 = spark.read
35   .format( source= "csv")
36   .option("header", "true") //reading the headers
37   .option("mode", "DROPMALFORMED")
38   .load( path= "fifa-world-cup/WorldCupMatches.csv")
39
40 // Printing the Schema
41 df1.printSchema()
42 df2.printSchema()
43 df3.printSchema()
44
45

```

Task2 > main(args: Array[String])

All files are up-to-date (11 minutes ago)

Event Log

47:43 LF UTF-8 2 spaces ENG 6:01 PM 5/9/2020

Task2 and Task4 src main scala Task2.scala

```

47 //b.Perform 10 intuitive queries in Dataset
48 //For this problem we have used the Spark SQL on DataFrames
49
50 //First of all create three Temp View
51 df1.createOrReplaceTempView( viewName= "WC")
52 df2.createOrReplaceTempView( viewName= "Players")
53 df3.createOrReplaceTempView( viewName= "Matches")
54
55
56 // Find the country which was winner for highest number of times
57 val Q = spark.sql( sqlText= "select Winner, Count(*) as HighestCount from WC group by Winner Order By HighestCount")
58 Q.show()
59
60 //Find the year in which there was maximum craze and attendance for the games
61 val Q1 = spark.sql( sqlText= "select Year, Attendance from WC Order By Attendance desc")
62 Q1.show()
63
64 //Sum of Home Team Goals and Away Team Goals for each stage
65 val Q2 = spark.sql( sqlText= "select Stage,Sum('Home Team Goals'),Sum('Away Team Goals') from Matches Group By Stage")
66 Q2.show()
67
68 // Referees with highest number of matches
69 val Q3 = spark.sql( sqlText= "select Referee, Count(Referee) from Matches group by referee order by 2 desc limit 1")
70 Q3.show()
71
72 //Home Team Goals for Winners
73 val Q4 =spark.sql( sqlText= "select WC.Year, WC.Winner, Sum('Home Team Goals') from Matches join WC on Matches.Year=WC.Year group by WC.Winner, WC.Year")
74
75

```

Task2 > main(args: Array[String])

All files are up-to-date (12 minutes ago)

Event Log

47:43 LF UTF-8 2 spaces ENG 6:01 PM 5/9/2020

The screenshot shows the IntelliJ IDEA interface with the following details:

- Project Structure:** The project is named "Task2 and Task4 [Task2]". It contains several sub-directories like ".idea", "fifa-world-cup", "nashville-meetup", and "src/main/scala".
- Code Editor:** The main editor window displays "Task2.scala" with the following Scala code:

```
//Home Team Goals for Winners
val Q4 = spark.sql( sqlText= "select WC.Year, WC.Winner, Sum('Home Team Goals') from Matches join WC on Matches.'H
Q4.show()

//Countries that hosted highest world cup matches
val Q5 = spark.sql( sqlText= "select Country, Count(Country) from WC Group By Country Order by Count(Country) des
Q5.show()

//No of matches in year 1934 and in Estadio Centenario stadium
val Q6 = spark.sql( sqlText= "select count(*) from Matches where year=1934 AND Stadium = 'Estadio Centenario' ")
Q6.show()

//Captains of Teams of Final Matches for all Years
val Q7 = spark.sql( sqlText= "select Matches.Year , Players.'Player Name' from Matches join Players on Matches.St
Q7.show()

//Count of matches of referee Macias Jose
val Q8 = spark.sql( sqlText= "select Count(*) from Matches where referee like '%MACIAS Jose%'")
Q8.show()

//Stadium with highest number of matches
val Q9 = spark.sql( sqlText= "select Stadium from Matches Group By Stadium order by Count(*) desc limit 1")
Q9.show()

//Display all Goal-Keepers
val Q10 = spark.sql( sqlText= "select 'Player Name' from Players where Position = 'GK' ")
Q10.show()
```

- Toolbars and Status Bar:** The bottom toolbar includes icons for Git, TODO, Run, Terminal, sbt shell, and Build. The status bar shows "Event Log", "47:43 LF", "UTF-8", "2 spaces", "6:01 PM", and "5/9/2020".

The screenshot shows the IntelliJ IDEA interface with the following details:

- Project Structure:** The project is named "Task2 and Task4 [Task2]". It contains several sub-directories like ".idea", "fifa-world-cup", "nashville-meetup", and "src/main/scala".
- Code Editor:** The main editor window displays "Task2.scala" with the following Scala code, highlighting a portion of the code between lines 108 and 127:

```
//Perform any 5 queries in Spark RDD's and Spark Data Frames.
// To Solve this Problem we first create the rdd as we already have Dataframe df1 created above code
// RDD creation

val csv = sc.textFile( path= "fifa-world-cup/WorldCups.csv")

val h1 = csv.first()

val data = csv.filter(line => line != h1)

data.foreach(println)

val rdd = data.map(line=>line.split( regex = "\r\n")).collect()

//rdd.foreach(println)
//Winner of 1930
val rdd1 = data.filter(line => line.split( regex = "\r\n")(0) == "1930").map(line => (line.split( regex = "\r\n")(0),
| (line.split( regex = "\r\n")(1), (line.split( regex = "\r\n")(2)) )
rdd1.foreach(println)

// Dataframe
df1.select( col= "Year", cols= "Country", "Winner").filter( conditionExpr= "Year =1930").show()

// Dataframe SQL
val df01 = spark.sql( sqlText= "select Year, Country, Winner FROM WC WHERE Year = 1930 ").show()
```

- Toolbars and Status Bar:** The bottom toolbar includes icons for Git, TODO, Run, Terminal, sbt shell, and Build. The status bar shows "Event Log", "10:25 LF", "UTF-8", "2 spaces", "6:02 PM", and "5/9/2020".

```

127 // When Uruguay won
128 // Using RDD
129 val rdd2 = data.filter(line => (line.split( regex = "\n")(2) == "Uruguay" ))
130   .map(line=> (line.split( regex = "\n")(0),line.split( regex = "\n")(2),line.split( regex = "\n")(3),line.split( regex = "\n")(4)))
131   .collect()
132 rdd2.foreach(println)

133 // Using Dataframe
134 df1.select( col= "Year", cols= "Winner", "Runners-Up", "Third", "Fourth").filter( conditionExpr= "Winner == 'Uruguay' ")

135 // usig Spark SQL
136 val DFQ2 = spark.sql( sqlText= "select Year, Winner, Runners-Up, Third, Fourth from WC where Winner = 'Uruguay' ")

137 // Attendance > 500.00
138 // RDD
139 val rdd3 = data.filter(line => (line.split( regex = "\n")(9)>"500.000" ))
140   .map(line=> (line.split( regex = "\n")(0),line.split( regex = "\n")(1))).collect()
141 rdd3.foreach(println)

142 //DataFrame
143 df1.select( col= "Year", cols= "Country").filter( conditionExpr= "Attendance > 500.000").show( numRows = 10)

144 //DF - SQL
145 val DFQ3 = spark.sql( sqlText= "SELECT Year, Country from WC where Attendance > 500.000 ").show( numRows = 10)

```

The screenshot shows the IntelliJ IDEA interface with the Task2.scala file open. The code filters for Uruguay wins using RDD, Dataframe, and Spark SQL. It also filters for attendance greater than 500.000 using RDD and DataFrame.

```

153 //Country==Winner
154 //Rdd
155 val rdd4 = data.filter(line => line.split( regex = "\n")(1)==line.split( regex = "\n")(2))
156   .map(line => (line.split( regex = "\n")(0),line.split( regex = "\n")(1), line.split( regex = "\n")(2)))
157   .collect()
158 rdd4.foreach(println)

159 // Using Dataframe
160 df1.select( col= "Year", cols= "Country", "Winner").filter( conditionExpr= "Country==Winner").show( numRows = 10)

161 // usig Spark SQL
162 val DFQ4 = spark.sql( sqlText= "select Year,Country,Winner from WC where Country = Winner order by Year").show()

163 //Matches won by USA
164 //RDD
165 val rdd5 = data.filter(line=>line.split( regex = "\n")(2) == "USA" || line.split( regex = "\n")(3) == "USA" || line.s
166   .map(line=> (line.split( regex = "\n")(0),line.split( regex = "\n")(2),line.split( regex = "\n")(3),line.split( regex = "\n")
167   .collect()
168 rdd5.foreach(println)

169 // DataFrame
170 df1.filter( conditionExpr= "Winner == 'USA' OR 'Runners-Up' == 'USA' OR Third == 'USA' OR Fourth == 'USA' ").show()

171 // Spark SQL
172 val DFQ5 = spark.sql( sqlText= " Select Year, Country, Winner, 'Runners-Up', Third, Fourth from WC where Winner = 'USA' ")
173
174
175
176
177
178
179
180

```

The screenshot shows the IntelliJ IDEA interface with the Task2.scala file open. The code filters for USA wins using RDD, Dataframe, and Spark SQL.

Output

```
20/05/09 18:05:32 INFO BlockManagerMasterEndpoint: Registering block manager DESKTOP-KJ4BB42:53868 with 895.2 MB RAM, BlockManagerId(driver, DESKTOP-KJ4BB42, 53868)
20/05/09 18:05:32 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, DESKTOP-KJ4BB42, 53868, None)
20/05/09 18:05:32 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, DESKTOP-KJ4BB42, 53868, None)
20/05/09 18:05:32 WARN SparkContext: Using an existing SparkContext; some configuration may not take effect.

root
|-- Year: string (nullable = true)
|-- Country: string (nullable = true)
|-- Winner: string (nullable = true)
|-- Runners-Up: string (nullable = true)
|-- Third: string (nullable = true)
|-- Fourth: string (nullable = true)
|-- GoalsScored: string (nullable = true)
|-- QualifiedTeams: string (nullable = true)
|-- MatchesPlayed: string (nullable = true)
|-- Attendance: string (nullable = true)

root
|-- RoundID: string (nullable = true)
|-- MatchID: string (nullable = true)
|-- Team Initials: string (nullable = true)
|-- Coach Name: string (nullable = true)
|-- Line-up: string (nullable = true)
|-- Shirt Number: string (nullable = true)
|-- Player Name: string (nullable = true)
|-- Position: string (nullable = true)
|-- Event: string (nullable = true)

Build completed successfully in 13 s 237 ms (12 minutes ago)
```

```
-- Player Name: string (nullable = true)
|-- Position: string (nullable = true)
|-- Event: string (nullable = true)

root
|-- Year: string (nullable = true)
|-- Datetime: string (nullable = true)
|-- Stage: string (nullable = true)
|-- Stadium: string (nullable = true)
|-- City: string (nullable = true)
|-- Home Team Name: string (nullable = true)
|-- Home Team Goals: string (nullable = true)
|-- Away Team Goals: string (nullable = true)
|-- Away Team Name: string (nullable = true)
|-- Win conditions: string (nullable = true)
|-- Attendance: string (nullable = true)
|-- Half-time Home Goals: string (nullable = true)
|-- Half-time Away Goals: string (nullable = true)
|-- Referee: string (nullable = true)
|-- Assistant 1: string (nullable = true)
|-- Assistant 2: string (nullable = true)
|-- RoundID: string (nullable = true)
|-- MatchID: string (nullable = true)
|-- Home Team Initials: string (nullable = true)
|-- Away Team Initials: string (nullable = true)

Build completed successfully in 13 s 237 ms (12 minutes ago)
```

File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help Task2 - Task2.scala [Task2] - IntelliJ IDEA

Task2 and Task4 src main scala Task2.scala

Run: Task2

```
+-----+  
| Winner|HighestCount|  
+-----+  
| Brazil|      5|  
| Italy|       4|  
| Germany FR| 3|  
| Uruguay|     2|  
| Argentina|   2|  
| Germany|    1|  
| France|     1|  
| Spain|      1|  
| England|    1|  
+-----+  
  
+-----+  
|Year|Attendance|  
+-----+  
|1962| 893.172|  
|1958| 819.810|  
|1954| 768.607|  
|1930| 590.549|  
|1938| 375.700|  
|1934| 363.000|  
|1994| 3.587.538|  
|2014| 3.366.810|  
|2006| 3.359.439|  
|2010| 3.190.051|  
+-----+
```

Git & TODO Run Terminal sbt shell Build Event Log

Build completed successfully in 13 s 237 ms (13 minutes ago)

Type here to search

File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help Task2 - Task2.scala [Task2] - IntelliJ IDEA

Task2 and Task4 src main scala Task2.scala

Run: Task2

```
+-----+  
| Stage|sum(CAST(Home Team Goals AS DOUBLE))|sum(CAST(Away Team Goals AS DOUBLE))|  
+-----+  
| Group 1|          129.0|           45.0|  
| Final|          43.0|           26.0|  
| Group H|          39.0|           33.0|  
| Group 5|          5.0|            4.0|  
| null|          null|           null|  
| Group 6|          43.0|           14.0|  
| Group A|          77.0|           83.0|  
| Third place|         5.0|            4.0|  
| Group 3|          116.0|           38.0|  
| First round|         30.0|           14.0|  
| Group E|          65.0|           56.0|  
| Group 4|          127.0|           49.0|  
| Round of 16|         114.0|           68.0|  
| Preliminary round|        30.0|           13.0|  
| Play-off for thir...|        0.0|            6.0|  
| Semi-finals|         76.0|           55.0|  
| Group D|          73.0|           50.0|  
| Match for third p...|         38.0|           21.0|  
| Group 2|          142.0|           45.0|  
| Quarter-finals|        118.0|           62.0|  
+-----+  
only showing top 20 rows
```

Git & TODO Run Terminal sbt shell Build Event Log

Build completed successfully in 13 s 237 ms (13 minutes ago)

Type here to search

```
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help Task2 - Task2.scala [Task2] - IntelliJ IDEA
Task2 and Task4 src main scala Task2.scala
Project Run: Task2
+-----+
| Referee|count(Referee)|
+-----+
|Ravshan IRMATOV (...) | 10|
|LARRIONDA Jorge (...) | 8|
|ARCHUNDIA Benito (...) | 8|
| QUINIOU Joel (FRA)| 8|
|RODRIGUEZ Marco (...) | 8|
+-----+
+-----+
|Year| Winner|sum(CAST(Home Team Goals AS DOUBLE))|
+-----+
|1930| Uruguay| 62.0|
|1934| Italy| 99.0|
|1938| Italy| 99.0|
|1950| Uruguay| 62.0|
|1954| Germany FR| 99.0|
|1958| Brazil| 188.0|
|1962| Brazil| 188.0|
|1966| England| 54.0|
|1970| Brazil| 188.0|
|1974| Germany FR| 99.0|
|1978| Argentina| 111.0|
|1982| Italy| 99.0|
|1986| Argentina| 111.0|
|1990| Germany FR| 99.0|
+-----+
Event Log
Build completed successfully in 13 s 237 ms (13 minutes ago)
Type here to search
376:1 LF UTF-8 2 spaces ENG 6:18 PM 5/9/2020
```

```
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help Task2 - Task2.scala [Task2] - IntelliJ IDEA
Task2 and Task4 src main scala Task2.scala
Project Run: Task2
+-----+
| Country|count(Country)|
+-----+
|Germany| 2|
|Mexico| 2|
|France| 2|
|Brazil| 2|
|Italy| 2|
|Sweden| 1|
|Argentina| 1|
|Chile| 1|
|Spain| 1|
|Korea/Japan| 1|
|Switzerland| 1|
|South Africa| 1|
|USA| 1|
|Uruguay| 1|
|England| 1|
+-----+
+-----+
|count(1)|
+-----+
| 0|
+-----+
+-----+
Event Log
Build completed successfully in 13 s 237 ms (13 minutes ago)
Type here to search
376:1 LF UTF-8 2 spaces ENG 6:18 PM 5/9/2020
```

File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help Task2 - Task2.scala [Task2] - IntelliJ IDEA

Task2 and Task4 src main scala Task2.scala

Run: Task2

```
+---+-----+
|Year|    Player Name|
+---+-----+
|1930|    Manuel FERREIRA|
|1930|    Jose NASAZZI|
|1938|    Gyorgy SAROSI|
|1938|    Giuseppe MEAZZA|
|1954|    Ferenc PUSKAS|
|1954|    Fritz WALTER|
|1958|    BELINI|
|1958|    Nils LIEDHOLM|
|1962|    MAURO RAMOS|
|1962|    Ladislav NOVAK|
|1966|    Bobby MOORE|
|1966|    Uwe SEELER|
|1970|    Giacinto FACCHETTI|
|1970|    CARLOS ALBERTO|
|1974|    Franz BECKENBAUER|
|1974|    Johan CRUYFF|
|1978|    Ruud KROL|
|1978|    Daniel PASSARELLA|
|1982|    Karl-Heinz RUMMENIGGE|
|1986|    Diego MARADONA|
+---+-----+
only showing top 20 rows
```

Git TODO Run Terminal sbt shell Build Event Log

Build completed successfully in 13 s 237 ms (13 minutes ago)

Type here to search

File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help Task2 - Task2.scala [Task2] - IntelliJ IDEA

Task2 and Task4 src main scala Task2.scala

Run: Task2

```
+----+
|count(1)|
+----+
| 2|
+----+
```

```
+----+
|Stadium|
+----+
| null|
+----+
```

```
+----+
|    Player Name|
+----+
| Alex THEPOT|
| Oscar BONFIGLIO|
| Jimmy DOUGLAS|
| Arnold BADJOU|
| Milovan JAKSIC|
| JOEL|
| Ion LAPUSNEANU|
| Juan VALDIVIESO|
| Angel BOSSIO|
| Alex THEPOT|
| Roberto CORTES|
| Isidoro SOTA|
```

Git TODO Run Terminal sbt shell Build Event Log

Build completed successfully in 13 s 237 ms (14 minutes ago)

Type here to search

File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help Task2 - Task2.scala [Task2] - IntelliJ IDEA

Task2 and Task4 src main scala Task2.scala

Run: Task2

```
+-----+
| Player Name|
+-----+
| Alex THEPOT|
| Oscar BONFIGLIO|
| Jimmy DOUGLAS|
| Arnold BADJOU|
| Milovan JAKSIC|
| JOEL|
| Ion LAPUSNEANU|
| Juan VALDIVIESO|
| Angel BOSSIO|
| Alex THEPOT|
| Roberto CORTES|
| Isidoro SOTA|
| Milovan JAKSIC|
| Jesus BERNUDEZ|
| Jimmy DOUGLAS|
| Modesto DENIS|
| Enrique BALLESTREIRO|
| Jorge PARDON|
| Roberto CORTES|
| Alex THEPOT|
+-----+
only showing top 20 rows
```

1978,Argentina,Argentina,Netherlands,Brazil,Italy,102,16,38,1.545.791

Git & TODO Run Terminal sbt shell Build Event Log

Build completed successfully in 13 s 237 ms (14 minutes ago) 256:22 LF UTF-8 2 spaces ENG 6:19 PM 5/9/2020

Type here to search

File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help Task2 - Task2.scala [Task2] - IntelliJ IDEA

Task2 and Task4 src main scala Task2.scala

Run: Task2

```
+-----+
| 1998,France,France,Brazil,Croatia,Netherlands,171,32,64,2.785.108|
| 2002,Korea/Japan,Brazil,Germany,Turkey,Korea Republic,161,32,64,2.705.197|
| 2006,Germany,Italy,France,Germany,Portugal,147,32,64,3.359.439|
| 2010,South Africa,Spain,Netherlands,Germany,Uruguay,145,32,64,3.178.856|
| 2014,Brazil,Germany,Argentina,Netherlands,Brazil,171,32,64,3.386.810|
| 1958,Uruguay,Uruguay,Argentina,USA,Yugoslavia,70,13,18,590.549|
| 1934,Italy,Italy,Czechoslovakia,Germany,Austria,70,16,17,363.000|
| 1938,France,Italy,Hungary,Brazil,Sweden,84,15,18,375.700|
| 1950,Brazil,Uruguay,Brazil,Sweden,Spain,88,13,22,1.045.246|
| 1954,Switzerland,Germany,FR,Hungary,Austria,Uruguay,140,16,26,768.607|
| 1958,Sweden,Brazil,Sweden,France,Germany,FR,126,16,35,819.810|
| 1962,Chile,Brazil,Czechoslovakia,Chile,Yugoslavia,89,16,32,893.172|
| 1966,England,England,Germany,FR,Portugal,Soviet Union,89,16,32,1.563.135|
| 1970,Mexico,Brazil,Italy,Germany,FR,Uruguay,95,16,32,1.603.975|
| 1974,Germany,Germany,FR,Netherlands,Poland,Brazil,97,16,38,1.865.753|
|(1930,Uruguay,Uruguay)|
+-----+
|Year|Country| Winner|
+-----+
|1930|Uruguay|Uruguay|
+-----+
+-----+
|Year|Country| Winner|
+-----+
|1930|Uruguay|Uruguay|
+-----+
```

Git & TODO Run Terminal sbt shell Build Event Log

Build completed successfully in 13 s 237 ms (15 minutes ago) 256:22 LF UTF-8 2 spaces ENG 6:20 PM 5/9/2020

Type here to search

File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help Task2 - Task2.scala [Task2] - IntelliJ IDEA

Task2 and Task4 > src > main > scala > Task2.scala

Run: Task2

Project Z Structure Favorites

```
(1930, Uruguay, Argentina, USA, Yugoslavia, Yugoslavia)
(1950, Uruguay, Brazil, Sweden, Spain, Spain)
+---+---+---+---+---+
|Year| Winner|Runners-Up| Third| Fourth|
+---+---+---+---+---+
|1930| Uruguay| Argentina| USA|Yugoslavia|
|1950| Uruguay| Brazil|Sweden| Spain|
+---+---+---+---+---+
+---+---+---+---+---+
|Year| Winner|Runners-Up| Third| Fourth|
+---+---+---+---+---+
|1930| Uruguay| Argentina| USA|Yugoslavia|
|1950| Uruguay| Brazil|Sweden| Spain|
+---+---+---+---+---+
(1930, Uruguay)
(1954, Switzerland)
(1958, Sweden)
(1962, Chile)
+---+---+
|Year| Country|
+---+---+
|1930| Uruguay|
|1954| Switzerland|
|1958| Sweden|
|1962| Chile|
+---+---+
```

Git TODO Run Terminal sbt shell Build Event Log

Build completed successfully in 13 s 237 ms (15 minutes ago)

Type here to search

File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help Task2 - Task2.scala [Task2] - IntelliJ IDEA

Task2 and Task4 > src > main > scala > Task2.scala

Run: Task2

Project Z Structure Favorites

```
(1930, Uruguay) 0.0211(Sweden) Spain
+---+---+---+---+
(1930, Uruguay)
(1954, Switzerland)
(1958, Sweden)
(1962, Chile)
+---+---+
|Year| Country|
+---+---+
|1930| Uruguay|
|1954| Switzerland|
|1958| Sweden|
|1962| Chile|
+---+---+
+---+---+
|Year| Country|
+---+---+
|1930| Uruguay|
|1954| Switzerland|
|1958| Sweden|
|1962| Chile|
+---+---+
(1930, Uruguay, Uruguay)
(1934, Italy, Italy)
(1966, Fndland, Fndland)
```

Git TODO Run Terminal sbt shell Build Event Log

Build completed successfully in 13 s 237 ms (16 minutes ago)

Type here to search

```
(1930,Uruguay,Uruguay)
(1934,Italy,Italy)
(1966,England,England)
(1978,Argentina,Argentina)
(1998,France,France)

+---+---+---+
|Year| Country| Winner|
+---+---+---+
|1930| Uruguay| Uruguay|
|1934| Italy| Italy|
|1966| England| England|
|1978| Argentina| Argentina|
|1998| France| France|
+---+---+---+


+---+---+---+
|Year| Country| Winner|
+---+---+---+
|1930| Uruguay| Uruguay|
|1934| Italy| Italy|
|1966| England| England|
|1978| Argentina| Argentina|
|1998| France| France|
+---+---+---+


(1930,Uruguay,Argentina,USA,Yugoslavia)
```

```
+---+---+---+
|Year| Country| Winner|
+---+---+---+
|1930| Uruguay| Uruguay|
|1934| Italy| Italy|
|1966| England| England|
|1978| Argentina| Argentina|
|1998| France| France|
+---+---+---+


(1930,Uruguay,Argentina,USA,Yugoslavia)
+---+---+---+---+---+---+---+
|Year|Country| Winner|Runners-Up|Third| Fourth|GoalsScored|QualifiedTeams|MatchesPlayed|Attendance|
+---+---+---+---+---+---+---+---+---+
|1930|Uruguay|Uruguay| Argentina| USA|Yugoslavia| 70| 13| 18| 590.549|
+---+---+---+---+---+---+---+---+---+


+---+---+---+---+---+
|Year|Country| Winner|Runners-Up|Third| Fourth|
+---+---+---+---+---+---+
|1930|Uruguay|Uruguay| Argentina| USA|Yugoslavia|
+---+---+---+---+---+---+


Process finished with exit code 0
```

Task 3

a. Perform Word-Count on Twitter Streaming Data using Spark.

Code:

```
Twitter.py × TwitterStreaming.py ×
1  from builtins import BaseException, print, str
2
3  from tweepy import OAuthHandler
4  from tweepy import Stream
5  from tweepy.streaming import StreamListener
6  import socket
7  import json
8  import time
9
10 # authorization tokens
11 access_token = "1045756888510541824-1zbS5Pi4hL6pYUwHu041pwNcK4CNvm"
12 access_token_secret = "yer1Vmp2te61Wat52olEpmpkjYqLumSZ3ry9XpyKTJ7oH"
13 consumer_key = "WAlQxKnCMdIFRsx1NyS7bdAEV"
14 consumer_secret = "cJNYNLYHhb0McLJaQ3sN9uWiAR6DmFDhzmN6CXNUCCIPOezM18"
15
16 auth = OAuthHandler(consumer_key, consumer_secret)
17 auth.set_access_token(access_token, access_token_secret)
18
19 o class StreamListener(StreamListener):
20
21 o     def __init__(self, socket):
22         self.socket = socket
23
```

```
24  def on_data(self, data):
25      try:
26          msg = json.loads(data)
27          print(msg['text'].encode('utf-8'))
28          self.socket.send(msg['text'].encode('utf-8'))
29          return True
30      except BaseException as e:
31          print("Encountered error in on_data: %s" % str(e))
32          return True
33
34  def on_error(self, status_code):
35      print("Encountered streaming error (", status_code, ")")
36      return True
37
38  def readData(socket):
39      auth = OAuthHandler(consumer_key, consumer_secret)
40      auth.set_access_token(access_token, access_token_secret)
41      stream = Stream(auth, StreamListener(socket))
42      tags = ["layoffs", "covid", "corona", "wfh"]
43      stream.filter(track=tags)
44
45  if __name__ == "__main__":
46      s = socket.socket()
47      host = "localhost"
48      port = 6000
49      s.bind((host, port))
50      print("Listening on the port: %s" % str(port))
51      s.listen(5)
52      socket, addr = s.accept()
53      print("Received the request from: " + str(addr))
54      time.sleep(5)
55      readData(socket)
```

```

Twitter.py × TwitterStreaming.py ×
1 import os
2
3 from pyspark import SparkContext
4 from pyspark.streaming import StreamingContext
5
6 from collections import namedtuple
7
8 os.environ["SPARK_HOME"] = "C:\Spark"
9
10 def main():
11     sc = SparkContext(appName="WordCount")
12     ssc = StreamingContext(sc, 5)
13     lines = ssc.socketTextStream("localhost", 6000)
14     fields = ("word", "count")
15     Tweet = namedtuple('Text', fields)
16
17     counts = lines.flatMap(lambda text: text.split(" "))\
18         .map(lambda word: (word, 1))\
19         .reduceByKey(lambda x, y: x + y).map(lambda rec: Tweet(rec[0], rec[1]))
20     counts.pprint()
21     ssc.start()
22     ssc.awaitTermination()
23
24 if __name__ == "__main__":
25     main()
26

```

Output:

```

Run: Twitter × TwitterStreaming ×
C:\Users\lohit\AppData\Local\Programs\Python\Python37\python.exe C:/Users/Lohit/Documents/GitHub/BigDataProgramming/Lab2/Task3/Twitter.py
Listening on the port: 6000
Received the request from: ('127.0.0.1', 51519)
b'RT @kroordarshan: All News18 debates by Amish Devgan (Aar Paar) between 17th March (1st debate on Corona) to 8th May '20\n\xe0\x9a4\xaf\xe0\x9a5\x87 \xe0\x9a4\x9a8\xe0\x9a5\x
b'RT @mjtreza1: Feliz Noche!\nAHORA\nReuni\xc3\xb3n de balance de la lucha contra el COVID-19 y evaluaci\xc3\xb3n de distintos temas de inter\xc3\x9a9s nacionales.\n#V\xe2\x80\x99
b'RT @cafecomferri: N\xc3\xba3o foi decis\xc3\xba3o do STF o Brasil n\xc3\xba3o ter pol\xc3\xadtica \x9c3\xbanica pro COVID19 @claudiodantas ?'
b'RT @RealLucyLawless: Hilarious and TERRIFYING: EVERYTHING I NEED TO KNOW TO SURVIVE COVID-19 I LEARNED BY WATCHING SCIFI... https://t.co/QFB\xe2\x80\x99
b'RT @trtworld: Turkish companies have successfully achieved mass production of ventilators, which are critical in treating Covid-19. Obaida\xe2\x80\x99
b'@lucasrohan Quando pega corona e corre risco de morte: mas se o pt n tivesse roubado agenti teriamos ospitais'
b'Este hombre diciendo que no se endeudan pero reparte cr\xc3\xadditos que van a endeudar por 3 al\xc3\xbdlos a sus beneficiarios...\xe2\x80\x99 https://t.co/7ghv1TEdu5
b'RT @LunaMate10: Por COVID-19, Invea suspende comercios no esenciales en 5 alcald\xc3\xadas de CdMx\nINVEA @inveacdmx\n y ELEKTRA cu\xc3\x9a1ndo?\n#ElektraCu\xe2\x80\x99
b'RT @MinSaludCol: #ReporteCOVID19 \xf0\x9f\x9a6\x9a0 Para este 9 de mayo, confirmamos en Colombia:\n\n145 recuperados,\n444 nuevos casos y\n17 fallecidos.\n\nAs\xc3\x9a1d,\n\x
b'RT @bernamaradio: Rakyat Eropah yang berkongsikan pengalaman ketika menjalani rawatan COVID-19 di Malaysia menyatakan mereka menerima layanan\xe2\x80\x99
b'RT @GP_Ansor_Malang: JeLang PSBB Malang Raya, @NU_MalangKota mendukung dengan mendirikan Media Centre Satgas NU Peduli Covid-19 Malang Raya\xe2\x80\x99
b'RT @ithembwa2926319: Renowned German Forensic doctor destroys media Covid_19 "Killer virus lies"\n#Checkpoint #Covid19SA #Covid_19SA #Covid\xe2\x80\x99
b'RT @f_passerini94: @GianricoCarof guardi, zia sta guidando una regione che \x9c3\x9a8 best practice nel mondo forse, in italia sicuro, nella lotta\xe2\x80\x99
b'\xe2\x98\x9a\x98\x9a\xef\x9b\x8f DIA DAS M\xc3\x83ES DE LUTO \xf0\x9f\x8f\x9b\x94\n\nAs fam\xc3\x9a3adias dos mais de 10 mil pacientes graves que morreram por Covid-19 \n\x9c3\x9a3o ter\x99

```

```

Run: Twitter x TwitterStreaming x
05/09 17:55:37 WARN BlockManager: Block input-0-1589064937200 replicated to only 0 peer(s) instead of 1 peers
-----
e: 2020-05-09 17:55:10
-----
t(word='March', count=1)
t(word='नूज़', count=1)
t(word='अ...RT', count=1)
t(word='', count=15)
t(word='el', count=4)
t(word='y', count=5)
t(word='Não', count=1)
t(word='foi', count=1)
t(word='Brasil', count=1)
t(word='ter', count=1)

```

Task 4

a. Perform Page Rank

b. State importance of using graphx on the chosen dataset.

Input

Meta-groups.csv as vertices View

Vertices of graph are 100 records of *Meta-groups.csv*

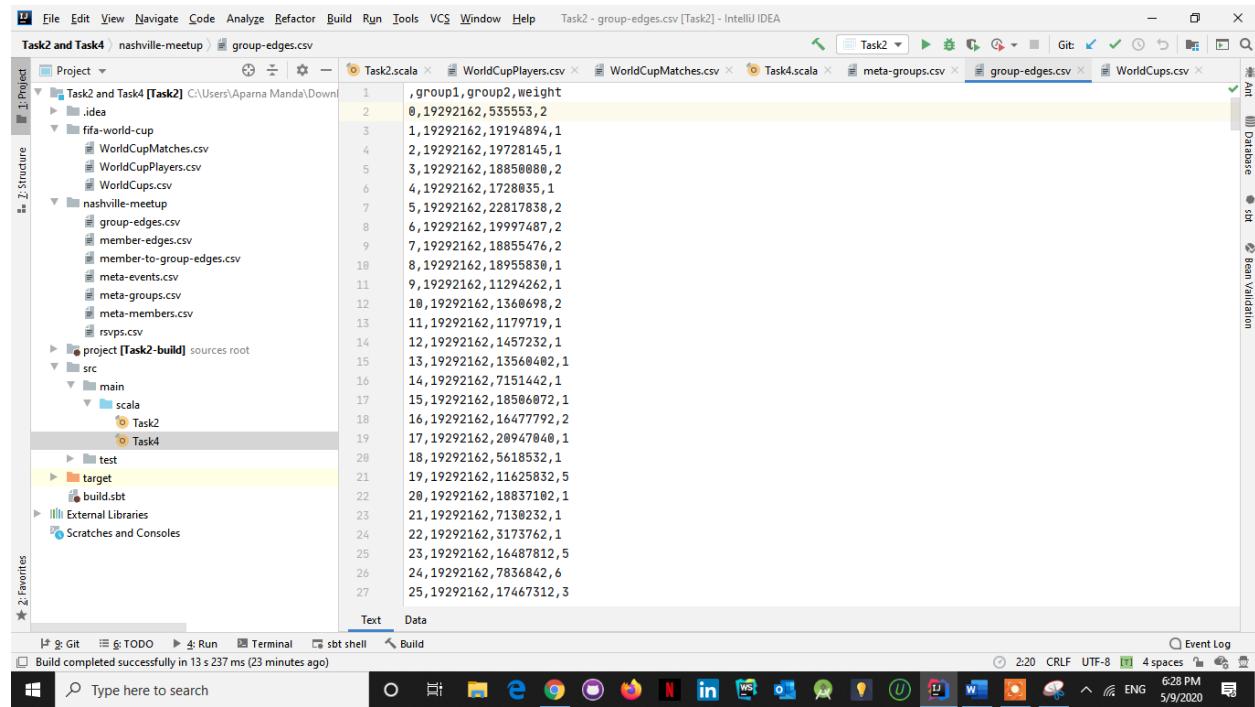
```

group_id,group_name,num_members,category_id,category_name,organizer_id,group_urlname
33981, Nashville Hiking Meetup, 15838, 23, Outdoors & Adventure, 4353803, nashville-hiking
19728145, Stepping Out Social Dance Meetup, 1778, 5, Dancing, 118484462, steppingoutsocialdance
6335372, Nashville soccer, 2869, 32, Sports & Recreation, 108448382, Nashville-soccer
10016242, NashJS, 1975, 34, Tech, 8111102, nashjs
21174496, 20's & 30's Women looking for girlfriends, 2782, 31, Socializing, 184580248, new-friends-in-Nashville
11077852, Sunday Assembly Nashville, 918, 28, Religion & Beliefs, 4765912, Sunday-Assembly-Nashville
22197221, Team Green Adventures, 1812, 23, Outdoors & Adventure, 199336381, TeamGreenAdventures
1585196, Tennessee Hiking Group, 4828, 23, Outdoors & Adventure, 13537265, TennesseeHikingGroup
526316, ¡Diablos Que Bajan! (Salsa Nashville), 3472, 5, Dancing, 12229328, diablos-que-bajan
1763198, Nashville Tennis Meetup, 1563, 32, Sports & Recreation, 9890725, Nashville-Tennis-Meetup
18243826, Middle TN 40+ singles, 2583, 30, Singles, 198309808, MTN-40
11625832, PyNash, 1442, 34, Tech, 215201845, PyNash
168014, The Nashville Writers Meetup, 3286, 36, Writing, 1281684, nashvillewriters
19218850, Greater Nashville Healthcare Analytics, 764, 34, Tech, 12825115, Greater-Nashville-Healthcare-Analytics
1526075, Nashville Area Gamer Association - NAGA, 2730, 11, Games, 10764011, NAGACentral
1102353, Nashville Backpacker Meetup, 3861, 23, Outdoors & Adventure, 7528310, NashvilleBackpacker
18955830, Eat Love Nash, 5008, 31, Socializing, 13814459, EatLoveNash
18495240, Middle Tennessee Hiking Meetup, 1576, 23, Outdoors & Adventure, 183268581, Middle-Tennessee-Hiking-Outdoor-Meetup
18562307, Nashville Young Professionals Meetup, 3210, 2, Career & Business, 8736052, Nashville-Young-Professionals-Meetup
18589616, Agile Nashville User Group, 862, 34, Tech, 126249582, Agile-Nashville-User-Group
18297014, Nashville Christian Singles, 1683, 23, Outdoors & Adventure, 183268581, Nashville-Active-Christian-Singles
47094, The Greater Nashville RPG and Board Gamers Group, 2375, 11, Games, 185299351, dnd-49
20181560, Tails of the Trail, 1241, 26, Pets & Animals, 365419, Tails-of-the-Trail
19277993, Nashville DevOps Meetup, 502, 34, Tech, 183378188, NashDevOps
7836842, Nashville UX, 1318, 34, Tech, 21627131, nashville-ux
18616278, Franklin Developer Lunch & Learn, 629, 34, Tech, 170855672, franklin-developer-lunch

```

Group-edges.csv as edges View

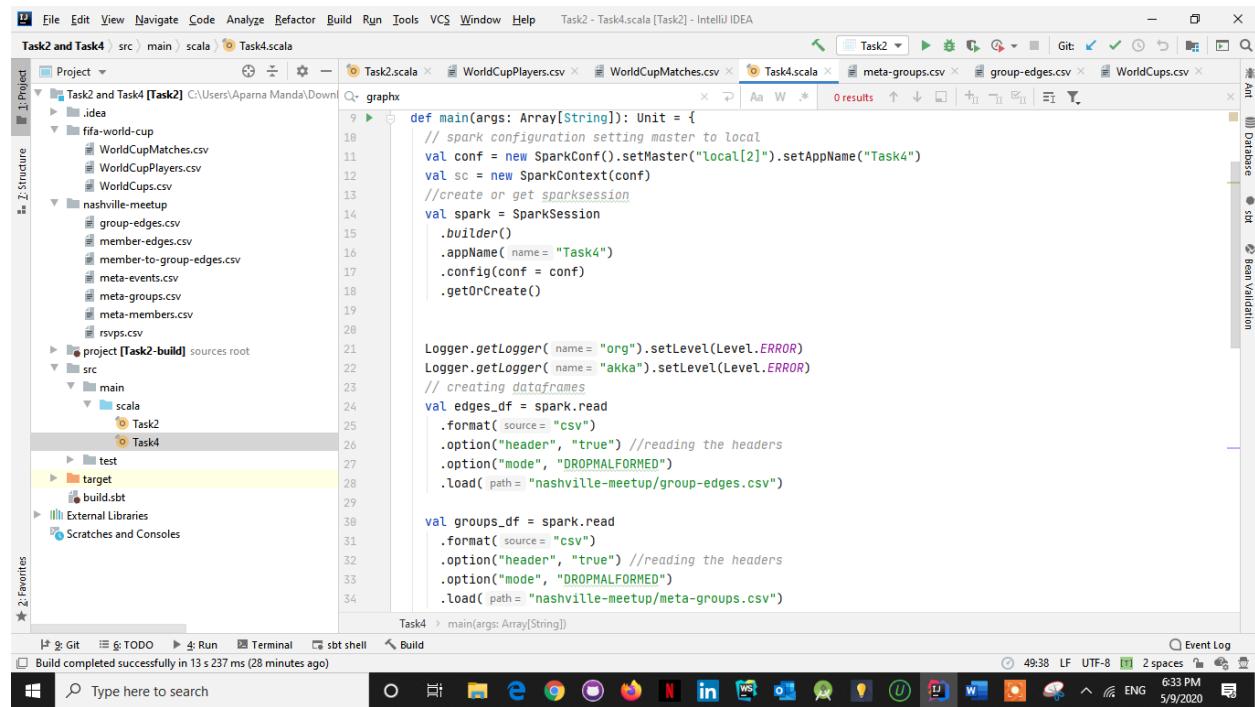
Edges of graph are 500 records of Group-edges.csv



The screenshot shows the IntelliJ IDEA interface with the 'group-edges.csv' file open in the center editor pane. The file contains 500 records of edge data, each consisting of three fields: source, target, and weight. The records are numbered from 1 to 500. The project structure on the left shows various CSV files and Scala code files like Task2.scala and Task4.scala. The bottom status bar indicates a successful build.

| Record | Source | Target | Weight |
|--------|--------------|-----------|---------------|
| 1 | , | group1 | group2,weight |
| 2 | 0,19292162, | 535553, | 2 |
| 3 | 1,19292162, | 19194894, | 1 |
| 4 | 2,19292162, | 19728145, | 1 |
| 5 | 3,19292162, | 18856080, | 2 |
| 6 | 4,19292162, | 17288035, | 1 |
| 7 | 5,19292162, | 22817838, | 2 |
| 8 | 6,19292162, | 19997487, | 2 |
| 9 | 7,19292162, | 18855476, | 2 |
| 10 | 8,19292162, | 18955838, | 1 |
| 11 | 9,19292162, | 11294262, | 1 |
| 12 | 10,19292162, | 1368698, | 2 |
| 13 | 11,19292162, | 1179719, | 1 |
| 14 | 12,19292162, | 1457232, | 1 |
| 15 | 13,19292162, | 13560402, | 1 |
| 16 | 14,19292162, | 7151442, | 1 |
| 17 | 15,19292162, | 18566872, | 1 |
| 18 | 16,19292162, | 16477792, | 2 |
| 19 | 17,19292162, | 28947840, | 1 |
| 20 | 18,19292162, | 5618532, | 1 |
| 21 | 19,19292162, | 11625632, | 5 |
| 22 | 20,19292162, | 18837182, | 1 |
| 23 | 21,19292162, | 7130232, | 1 |
| 24 | 22,19292162, | 3173762, | 1 |
| 25 | 23,19292162, | 16487812, | 5 |
| 26 | 24,19292162, | 7836842, | 6 |
| 27 | 25,19292162, | 17467312, | 3 |

Code



The screenshot shows the IntelliJ IDEA interface with the 'Task4.scala' file open in the center editor pane. The code defines a main function that sets up a SparkContext, reads two CSV files into DataFrames, and performs some basic operations on them. The project structure on the left shows the 'Task4' directory containing the Scala file. The bottom status bar indicates a successful build.

```
def main(args: Array[String]): Unit = {
    // spark configuration setting master to local
    val conf = new SparkConf().setMaster("local[2]").setAppName("Task4")
    val sc = new SparkContext(conf)
    //create or get sparksession
    val spark = SparkSession
        .builder()
        .appName( name = "Task4" )
        .config(conf = conf)
        .getOrCreate()

    Logger.getLogger( name = "org" ).setLevel(Level.ERROR)
    Logger.getLogger( name = "akka" ).setLevel(Level.ERROR)
    // creating dataframes
    val edges_df = spark.read
        .format( source = "csv" )
        .option("header", "true") //reading the headers
        .option("mode", "DROPMALFORMED")
        .load( path = "nashville-meetup/group-edges.csv" )

    val groups_df = spark.read
        .format( source = "csv" )
        .option("header", "true") //reading the headers
        .option("mode", "DROPMALFORMED")
        .load( path = "nashville-meetup/meta-groups.csv" )
}
```

The screenshot shows the IntelliJ IDEA interface with the following details:

- Project Structure:** The project is named "Task2 and Task4 [Task2]". It contains a "fifa-world-cup" directory with CSV files like "WorldCupMatches.csv", "WorldCupPlayers.csv", and "WorldCups.csv". It also has a "nashville-meetup" directory with CSV files like "group-edges.csv", "member-edges.csv", "member-to-group-edges.csv", "meta-events.csv", "meta-groups.csv", "meta-members.csv", and "rsvps.csv".
- Code Editor:** The file "Task4.scala" is open. The code reads CSV files into DataFrames, creates temporary views for edges and vertices, and then creates a GraphFrame. It prints the total number of vertices and edges.

```
// Printing the Schema
edges_df.printSchema()
groups_df.printSchema()
edges_df.createOrReplaceTempView( viewName = "edges" )
groups_df.createOrReplaceTempView( viewName = "vertices" )
val edges_val = spark.sql( sqlText = "select * from edges" )
val vertices_val = spark.sql( sqlText = "select * from vertices" )
//replacing column names
val vertices = vertices_val
.withColumnRenamed( existingName = "group_id", newName = "id" ).limit(100)
.distinct()

val edges = edges_val
.withColumnRenamed( existingName = "group1", newName = "src" ).limit(500).distinct()
.withColumnRenamed( existingName = "group2", newName = "dst" ).limit(500).distinct()

val graph = GraphFrame(vertices, edges)

edges.cache()
vertices.cache()
graph.vertices.show()
graph.edges.show()

println("Total Number of vertices count is : " + graph.vertices.count)
println("Total Number of edges count is: " + graph.edges.count)
```

- Toolbars and Status Bar:** The toolbar includes File, Edit, View, Navigate, Code, Analyze, Refactor, Build, Run, Tools, VCS, Window, Help, and Git. The status bar at the bottom shows "Build completed successfully in 13 s 237 ms (29 minutes ago)" and the current time as 6:34 PM on 5/9/2020.

This screenshot shows the same IntelliJ IDEA environment as the first one, but with additional code added to the "Task4.scala" file:

```
.distinct()

val edges = edges_val
.withColumnRenamed( existingName = "group1", newName = "src" ).limit(500).distinct()
.withColumnRenamed( existingName = "group2", newName = "dst" ).limit(500).distinct()

val graph = GraphFrame(vertices, edges)

edges.cache()
vertices.cache()
graph.vertices.show()
graph.edges.show()

println("Total Number of vertices count is : " + graph.vertices.count)
println("Total Number of edges count is: " + graph.edges.count)

val stationPageRank = graph.pageRank.resetProbability( value = 0.15 ).tol( value = 0.01 ).run()
stationPageRank.vertices.show()
stationPageRank.edges.show()
```

Output

```
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help Task2 - Task4.scala [Task2] - IntelliJ IDEA
Task2 and Task4 src main scala Task4.scala Task2.scala WorldCupPlayers.csv WorldCupMatches.csv Task4.scala meta-groups.csv group-edges.csv WorldCups.csv
Project Run Task4
1 Structure 2 Favorites
root
  |-- _c0: string (nullable = true)
  |-- group1: string (nullable = true)
  |-- group2: string (nullable = true)
  |-- weight: string (nullable = true)

root
  |-- group_id: string (nullable = true)
  |-- group_name: string (nullable = true)
  |-- num_members: string (nullable = true)
  |-- category_id: string (nullable = true)
  |-- category_name: string (nullable = true)
  |-- organizer_id: string (nullable = true)
  |-- group_urlname: string (nullable = true)

+-----+-----+-----+-----+
| id | group_name|num_members|category_id| category_name|organizer_id| group_urlname|
+-----+-----+-----+-----+
| 339811|Nashville Hiking ...| 15838| 23|Outdoors & Adventure| 4353803| nashville-hiking|
|19728145|Stepping Out Soci...| 1778| 5| Dancing| 118484462|steppingoutsocial...|
| 6335372| Nashville soccer| 2869| 32| Sports & Recreation| 1084448302| Nashville-soccer|
|18016242| NashJS| 1975| 34| Tech| 8111102| nashjs|
|21174496|20's & 30's Women...| 2782| 31| Socializing| 184580248|new-friends-in-Na...|
|11077852|Sunday Assembly N...| 918| 28| Religion & Beliefs| 4765912|Sunday-Assembly-N...|
|22197221|Team Green Advent...| 1812| 23|Outdoors & Adventure| 199336381| TeamGreenAdventures|
| 1585196|Tennessee Hiking ...| 4828| 23|Outdoors & Adventure| 13537265|TennesseeHikingGroup|
| 526316|Diablos Que Baila...| 3472| 5| Dancing| 12229328| diablos-que-bailan|
+-----+-----+-----+-----+
1 2 Git & TODO 4 Run Terminal sbt shell Build Event Log
Build completed successfully in 33 s 271 ms (2 minutes ago) 49:38 LF UTF-8 2 spaces 6:37 PM 5/9/2020
Type here to search
```

```
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help Task2 - Task4.scala [Task2] - IntelliJ IDEA
Task2 and Task4 src main scala Task4.scala Task2.scala WorldCupPlayers.csv WorldCupMatches.csv Task4.scala meta-groups.csv group-edges.csv WorldCups.csv
Project Run Task4
1 Structure 2 Favorites
+-----+-----+-----+-----+
| id | group_name|num_members|category_id| category_name|organizer_id| group_urlname|
+-----+-----+-----+-----+
| 339811|Nashville Hiking ...| 15838| 23|Outdoors & Adventure| 4353803| nashville-hiking|
|19728145|Stepping Out Soci...| 1778| 5| Dancing| 118484462|steppingoutsocial...|
| 6335372| Nashville soccer| 2869| 32| Sports & Recreation| 1084448302| Nashville-soccer|
|18016242| NashJS| 1975| 34| Tech| 8111102| nashjs|
|21174496|20's & 30's Women...| 2782| 31| Socializing| 184580248|new-friends-in-Na...|
|11077852|Sunday Assembly N...| 918| 28| Religion & Beliefs| 4765912|Sunday-Assembly-N...|
|22197221|Team Green Advent...| 1812| 23|Outdoors & Adventure| 199336381| TeamGreenAdventures|
| 1585196|Tennessee Hiking ...| 4828| 23|Outdoors & Adventure| 13537265|TennesseeHikingGroup|
| 526316|Diablos Que Baila...| 3472| 5| Dancing| 12229328| diablos-que-bailan|
| 1763190|Nashville Tennis ...| 1563| 32| Sports & Recreation| 9890725|Nashville-Tennis...|
|18243826|Middle TN 40+ sin...| 2583| 30| Singles| 198309808| MTN-40|
|11625832| PyNash| 1442| 34| Tech| 215201845| PyNash|
| 168014|The Nashville Wri...| 3286| 36| Writing| 1281684| nashvillewriters|
|19218850|Greater Nashville...| 764| 34| Tech| 12825115|Greater-Nashville...|
| 1526075|Nashville Area Ga...| 2738| 11| Games| 10764011| NAGACentral|
| 1102353|Nashville Backpacker| 3861| 23|Outdoors & Adventure| 7528310| NashvilleBackpacker|
|18955830| Eat Love Nash| 5008| 31| Socializing| 13814459| EatLoveNash|
|18495240|Middle Tennessee ...| 1576| 23|Outdoors & Adventure| 183268581|Middle-Tennessee...|
|18562307|Nashville Young P...| 3210| 21| Career & Business| 8736052|Nashville-Young-P...|
|18589616|Agile Nashville U...| 862| 34| Tech| 126249582|Agile-Nashville-U...|
+-----+-----+-----+-----+
only showing top 20 rows
1 2 Git & TODO 4 Run Terminal sbt shell Build Event Log
Build completed successfully in 33 s 271 ms (3 minutes ago) 49:38 LF UTF-8 2 spaces 6:37 PM 5/9/2020
Type here to search
```

File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help Task2 - Task4.scala [Task2] - IntelliJ IDEA

Task2 and Task4 > src > main > scala > Task4.scala

Run: Task4

```
+---+---+---+
| _0 | src | dst | weight |
+---+---+---+
| 0|19292162| 535553 | 2|
| 1|19292162|19194894| 1|
| 2|19292162|19728145| 1|
| 3|19292162|18850080| 2|
| 4|19292162| 1728835| 1|
| 5|19292162|22817838| 2|
| 6|19292162|19997487| 2|
| 7|19292162|18855476| 2|
| 8|19292162|18955830| 1|
| 9|19292162|11294262| 1|
| 10|19292162| 1360698| 2|
| 11|19292162| 1179719| 1|
| 12|19292162| 1457232| 1|
| 13|19292162|13560402| 1|
| 14|19292162| 7151442| 1|
| 15|19292162|18506872| 1|
| 16|19292162|16477792| 2|
| 17|19292162|20947840| 1|
| 18|19292162| 5618532| 1|
| 19|19292162|11625832| 5|
+---+---+---+
only showing top 20 rows
```

Total Number of vertices count is : 100

Build completed successfully in 33 s 271 ms (3 minutes ago)

Event Log

49:38 LF UTF-8 2 spaces ENG 6:38 PM 5/9/2020

File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help Task2 - Task4.scala [Task2] - IntelliJ IDEA

Task2 and Task4 > src > main > scala > Task4.scala

Run: Task4

```
+-----+
| id | group_name|num_members|category_id| category_name|organizer_id| group_urlname| pagerank|
+-----+
| 18616279|Franklin Develop...| 629| 34| Tech| 170855672|franklin-develope...|1.0155433646812946|
| 5882552|Nashville and Ten...| 1422| 6| Education & Learning| 59643462|Tennessee-Histori...|0.967117983945831|
| 22736876|Business Connecti...| 126| 2| Career & Business| 191532521|Brentwood-Rowdy-R...|0.9812912692589866|
| 19266390|Nashville Network...| 1444| 2| Career & Business| 11759637|Nashville-Network...|0.9812912692589866|
| 18529135|Franklin AM - Net...| 368| 2| Career & Business| 34583172|Franklin-AM-Netwo...|0.9812912692589866|
| 11625832| PvNash| 1442| 34| Tech| 215201845| PvNash|1.0397213143911592|
| 535553| Nashrb| 881| 34| Tech| 14344641| nashrb| 1.025548833526756|
| 18314164| NashBI| 784| 34| Tech| 183427754| NashBI|0.9812912692589866|
| 6707902|Data Science Nash...| 1046| 34| Tech| 14589429|Data-Science-Nash...|1.0155433646812946|
| 541319|The Nashville Son...| 2644| 21| Music| 2984170| vocalists-164|1.0245570740029548|
| 4126912|Nashville Online| 1532| 2| Career & Business| 44942272| nashville-online|1.0245570740029548|
| 22197221|Team Green Advent...| 1812| 23|Outdoors & Adventure| 199336381| TeamGreenAdventures|0.9812912692589866|
| 18955830| Eat Love Nash| 5008| 31| Socializing| 13814459| EatLoveNash|1.0829871191351275|
| 227528|Nashville Spanish...| 1417| 16|Language & Ethnic...| 2434993| spanish-570| 1.034561742848416|
| 18361585|Make Nashville Me...| 1684| 34| Tech| 5908662|Make-Nashville-Me...|0.967117983945831|
| 18850080| NashReact| 438| 34| Tech| 100083866| NashReact-Meetup|1.0013700838168913|
| 17228835| WordPress Nashville| 1643| 34| Tech| 72560962| NashvilleWordpress| 1.06881383270724|
| 19218850|Greater Nashville...| 764| 34| Tech| 12825115|Greater-Nashville...|0.9812912692589866|
| 3173762|Nashville Java Us...| 1041| 34| Tech| 40336202| nashvillejug|1.0013700838168913|
| 19556549|Inner Engineering...| 1103| 14| Health & Wellbeing| 184777575|Isha-Yoga-Inner-E...|0.9912959381044477|
+-----+
only showing top 20 rows
```

Build completed successfully in 33 s 271 ms (3 minutes ago)

Event Log

49:38 LF UTF-8 2 spaces ENG 6:38 PM 5/9/2020

```
+---+---+---+
|_c0|src|dst|weight|
+---+---+---+
|371|1179719|23674779|3|0.029411764785882353|
|6|19292162|19997487|2|0.041666666666666664|
|495|168014|4126912|1|0.05263157894736842|
|383|1585196|526316|1|0.017241379310344827|
|256|1585196|18562307|3|0.017241379310344827|
|28|19292162|67079702|2|0.041666666666666664|
|2|19292162|19728145|1|0.041666666666666664|
|248|1585196|21174496|3|0.017241379310344827|
|253|1585196|15297782|1|0.017241379310344827|
|272|1585196|11625832|3|0.017241379310344827|
|23|19292162|16487812|5|0.041666666666666664|
|377|1179719|18476981|1|0.029411764785882353|
|357|1179719|1526075|1|0.029411764785882353|
|273|1585196|18529135|1|0.017241379310344827|
|473|168014|929402|1|0.05263157894736842|
|259|1585196|19728145|18|0.017241379310344827|
|22|19292162|3173762|1|0.041666666666666664|
|0|19292162|535553|2|0.041666666666666664|
|378|1179719|9376702|1|0.029411764785882353|
|221|1585196|929402|8|0.017241379310344827|
+---+---+---+
only showing top 20 rows
```

b. Importance of GraphX

The importance of a group can be measured by the weights of the groups it is connected to and thus GraphX would be the best-fit for the dataset and Page Rank algorithm would be the perfect algorithm to determine the importance of the groups.