GraphFrames in Jupyter: a practical guide (https://towardsdatascience.com/graphframes-in-jupyter-a-practical-guide-9b3b346cebc5)

In [1]:
```
from pyspark.sql import SparkSession
import pyspark.sql.functions as F
from pyspark.sql.types import *

spark = SparkSession\
    .builder\
    .appName("chapter-30-graph")\
    .getOrCreate()

import os
SPARK_BOOK_DATA_PATH = os.environ['SPARK_BOOK_DATA_PATH']
```

In [2]:
```
bikeStations = spark.read\
  .option("header","true")\
  .csv(SPARK_BOOK_DATA_PATH + "/data/bike-data/201508_station_data.csv")
```

In [3]:
```
bikeStations.show(3, False)
```

```
+----------+---------------------------------+---------+-----------+---------+--------+------------
+
|station_id|name                             |lat      |long       |dockcount|landmark|installation
|
+----------+---------------------------------+---------+-----------+---------+--------+------------
+
|2         |San Jose Diridon Caltrain Station|37.329732|-121.901782|27       |San Jose|8/6/2013
|
|3         |San Jose Civic Center            |37.330698|-121.888979|15       |San Jose|8/5/2013
|
|4         |Santa Clara at Almaden           |37.333988|-121.894902|11       |San Jose|8/6/2013
|
+----------+---------------------------------+---------+-----------+---------+--------+------------
+
only showing top 3 rows
```

In [4]:
```
tripData = spark.read\
  .option("header","true")\
  .csv(SPARK_BOOK_DATA_PATH + "/data/bike-data/201508_trip_data.csv")
```

In [5]:
```
tripData.show(3, False)
```

```
+-------+--------+--------------+-----------------------------------+--------------+-------------
--+----------------------------------------+-----------+------+---------------+--------+
|Trip ID|Duration|Start Date     |Start Station                      |Start Terminal|End Date
|End Station                             |End Terminal|Bike #|Subscriber Type|Zip Code|
+-------+--------+--------------+-----------------------------------+--------------+-------------
--+----------------------------------------+-----------+------+---------------+--------+
|913460 |765     |8/31/2015 23:26|Harry Bridges Plaza (Ferry Building)|50            |8/31/2015 23:
39|San Francisco Caltrain (Townsend at 4th)|70          |288   |Subscriber     |2139    |
|913459 |1036    |8/31/2015 23:11|San Antonio Shopping Center         |31            |8/31/2015 23:
28|Mountain View City Hall                 |27          |35    |Subscriber     |95032   |
|913455 |307     |8/31/2015 23:13|Post at Kearny                      |47            |8/31/2015 23:
18|2nd at South Park                       |64          |468   |Subscriber     |94107   |
+-------+--------+--------------+-----------------------------------+--------------+-------------
--+----------------------------------------+-----------+------+---------------+--------+
only showing top 3 rows
```

In [6]:
```
# COMMAND ----------

stationVertices = bikeStations.withColumnRenamed("name", "id").distinct()
```

In [7]:
```
stationVertices.show(3)
```

```
+----------+--------------------+---------+-----------+---------+-------------+------------+
|station_id|                  id|      lat|       long|dockcount|     landmark|installation|
+----------+--------------------+---------+-----------+---------+-------------+------------+
|        51|Embarcadero at Fo...|37.791464|-122.391034|       19|San Francisco|   8/20/2013|
|        58|San Francisco Cit...| 37.77865|-122.418235|       19|San Francisco|   8/21/2013|
|        60|Embarcadero at Sa...| 37.80477|-122.403234|       15|San Francisco|   8/21/2013|
+----------+--------------------+---------+-----------+---------+-------------+------------+
only showing top 3 rows
```

In [8]:
```
tripEdges = tripData\
  .withColumnRenamed("Start Station", "src")\
  .withColumnRenamed("End Station", "dst")
```

In [9]:
```
tripEdges.show(3)
```

```
+-------+--------+--------------+--------------------+--------------+--------------+-------------
-------+------------+------+--------------+--------+
|Trip ID|Duration|     Start Date|                 src|Start Terminal|       End Date|
dst|End Terminal|Bike #|Subscriber Type|Zip Code|
+-------+--------+--------------+--------------------+--------------+--------------+-------------
-------+------------+------+--------------+--------+
| 913460|     765|8/31/2015 23:26|Harry Bridges Pla...|           50|8/31/2015 23:39|San Francisco
Cal...|          70|   288|     Subscriber|    2139|
| 913459|    1036|8/31/2015 23:11|San Antonio Shopp...|           31|8/31/2015 23:28|Mountain View
Cit...|          27|    35|     Subscriber|   95032|
| 913455|     307|8/31/2015 23:13|      Post at Kearny|           47|8/31/2015 23:18|   2nd at Sou
th Park|          64|   468|     Subscriber|   94107|
+-------+--------+--------------+--------------------+--------------+--------------+-------------
-------+------------+------+--------------+--------+
only showing top 3 rows
```

install **graphframes** pkg from https://spark-packages.org/package/graphframes/graphframes (https://spark-packages.org/package/graphframes/graphframes)

download graphframes-0.7.0-spark2.4-s_2.11.jar and copy it to `~/spark/jars` folder

Re-start jupyter notebook

```
$ PYSPARK_DRIVER_PYTHON="jupyter" PYSPARK_DRIVER_PYTHON_OPTS="notebook" pyspark --packages g
raphframes:graphframes:0.7.0-spark2.4-s_2.11
```

In [10]:
```
# COMMAND ----------

from graphframes import GraphFrame
```

In [11]:
```
stationGraph = GraphFrame(stationVertices, tripEdges)
stationGraph.cache()
```

Out[11]:
```
GraphFrame(v:[id: string, station_id: string ... 5 more fields], e:[src: string, dst: string ... 9 more fields])
```

In [12]:
```
# COMMAND ----------

print ("Total Number of Trips in Original Data: " + str(tripData.count()))
print ("Total Number of Stations: " + str(stationGraph.vertices.count()))
print ("Total Number of Trips in Graph: " + str(stationGraph.edges.count()))
```

```
Total Number of Trips in Original Data: 354152
Total Number of Stations: 70
Total Number of Trips in Graph: 354152
```

In [13]:
```
# COMMAND ----------

from pyspark.sql.functions import desc
stationGraph.edges.groupBy("src", "dst").count().orderBy(desc("count")).show(10,False)
```

```
+----------------------------------------------+----------------------------------------+-----+
|src                                           |dst                                     |count|
+----------------------------------------------+----------------------------------------+-----+
|San Francisco Caltrain 2 (330 Townsend)       |Townsend at 7th                         |3748 |
|Harry Bridges Plaza (Ferry Building)          |Embarcadero at Sansome                  |3145 |
|2nd at Townsend                               |Harry Bridges Plaza (Ferry Building)    |2973 |
|Townsend at 7th                               |San Francisco Caltrain 2 (330 Townsend) |2734 |
|Harry Bridges Plaza (Ferry Building)          |2nd at Townsend                         |2640 |
|Embarcadero at Folsom                         |San Francisco Caltrain (Townsend at 4th)|2439 |
|Steuart at Market                             |2nd at Townsend                         |2356 |
|Embarcadero at Sansome                        |Steuart at Market                       |2330 |
|Townsend at 7th                               |San Francisco Caltrain (Townsend at 4th)|2192 |
|Temporary Transbay Terminal (Howard at Beale)|San Francisco Caltrain (Townsend at 4th)|2184 |
+----------------------------------------------+----------------------------------------+-----+
only showing top 10 rows
```

In [14]:
```
# COMMAND ----------

stationGraph.edges\
  .where("src = 'Townsend at 7th' OR dst = 'Townsend at 7th'")\
  .groupBy("src", "dst").count()\
  .orderBy(desc("count"))\
  .show(10,False)
```

```
+---------------------------------------------+---------------------------------------------+-----+
|src                                          |dst                                          |count|
+---------------------------------------------+---------------------------------------------+-----+
|San Francisco Caltrain 2 (330 Townsend)      |Townsend at 7th                              |3748 |
|Townsend at 7th                              |San Francisco Caltrain 2 (330 Townsend)      |2734 |
|Townsend at 7th                              |San Francisco Caltrain (Townsend at 4th)     |2192 |
|Townsend at 7th                              |Civic Center BART (7th at Market)            |1844 |
|Civic Center BART (7th at Market)            |Townsend at 7th                              |1765 |
|San Francisco Caltrain (Townsend at 4th)     |Townsend at 7th                              |1198 |
|Temporary Transbay Terminal (Howard at Beale)|Townsend at 7th                              |834  |
|Townsend at 7th                              |Harry Bridges Plaza (Ferry Building)         |827  |
|Steuart at Market                            |Townsend at 7th                              |746  |
|Townsend at 7th                              |Temporary Transbay Terminal (Howard at Beale)|740  |
+---------------------------------------------+---------------------------------------------+-----+
only showing top 10 rows
```

In [15]:
```
# COMMAND ----------

townAnd7thEdges = stationGraph.edges\
  .where("src = 'Townsend at 7th' OR dst = 'Townsend at 7th'")
subgraph = GraphFrame(stationGraph.vertices, townAnd7thEdges)
```

In [16]:
```
# COMMAND ----------

motifs = stationGraph.find("(a)-[ab]->(b); (b)-[bc]->(c); (c)-[ca]->(a)")
```

In [17]:
```
# COMMAND ----------

from pyspark.sql.functions import expr
motifs.selectExpr("*",
    "to_timestamp(ab.`Start Date`, 'MM/dd/yyyy HH:mm') as abStart",
    "to_timestamp(bc.`Start Date`, 'MM/dd/yyyy HH:mm') as bcStart",
    "to_timestamp(ca.`Start Date`, 'MM/dd/yyyy HH:mm') as caStart")\
  .where("ca.`Bike #` = bc.`Bike #`").where("ab.`Bike #` = bc.`Bike #`")\
  .where("a.id != b.id").where("b.id != c.id")\
  .where("abStart < bcStart").where("bcStart < caStart")\
  .orderBy(expr("cast(caStart as long) - cast(abStart as long)"))\
  .selectExpr("a.id", "b.id", "c.id", "ab.`Start Date`", "ca.`End Date`")\
  .limit(1).show(1, False)
```

```
+----------------------------------------+--------------+----------------------------------------+-
--------------+--------------+
|id                                      |id            |id                                      |S
tart Date     |End Date      |
+----------------------------------------+--------------+----------------------------------------+-
--------------+--------------+
|San Francisco Caltrain 2 (330 Townsend)|Townsend at 7th|San Francisco Caltrain (Townsend at 4th)|
5/19/2015 16:09|5/19/2015 16:33|
+----------------------------------------+--------------+----------------------------------------+-
--------------+--------------+
```

In [18]:
```
# COMMAND ----------

from pyspark.sql.functions import desc
ranks = stationGraph.pageRank(resetProbability=0.15, maxIter=10)
ranks.vertices.orderBy(desc("pagerank")).select("id", "pagerank").show(10)
```

```
+--------------------+------------------+
|                  id|          pagerank|
+--------------------+------------------+
|San Jose Diridon ...| 4.051504835990019|
|San Francisco Cal...|3.3511832964286965|
|Mountain View Cal...|2.5143907710155435|
|Redwood City Calt...| 2.326308771371171|
|San Francisco Cal...| 2.231144291369883|
|Harry Bridges Pla...|1.8251120118882473|
|     2nd at Townsend|1.5821217785038688|
|Santa Clara at Al...|1.5730074084907584|
|     Townsend at 7th|1.5684565805340545|
|Embarcadero at Sa...|1.5414242087748589|
+--------------------+------------------+
only showing top 10 rows
```

In [19]:
```
# COMMAND ----------

inDeg = stationGraph.inDegrees
inDeg.orderBy(desc("inDegree")).show(5, False)
```

```
+----------------------------------------+--------+
|id                                      |inDegree|
+----------------------------------------+--------+
|San Francisco Caltrain (Townsend at 4th)|34810   |
|San Francisco Caltrain 2 (330 Townsend) |22523   |
|Harry Bridges Plaza (Ferry Building)    |17810   |
|2nd at Townsend                         |15463   |
|Townsend at 7th                         |15422   |
+----------------------------------------+--------+
only showing top 5 rows
```

In [20]:
```
# COMMAND ----------

outDeg = stationGraph.outDegrees
outDeg.orderBy(desc("outDegree")).show(5, False)
```

```
+----------------------------------------------+---------+
|id                                            |outDegree|
+----------------------------------------------+---------+
|San Francisco Caltrain (Townsend at 4th)      |26304    |
|San Francisco Caltrain 2 (330 Townsend)       |21758    |
|Harry Bridges Plaza (Ferry Building)          |17255    |
|Temporary Transbay Terminal (Howard at Beale)|14436    |
|Embarcadero at Sansome                        |14158    |
+----------------------------------------------+---------+
only showing top 5 rows
```

In [21]:
```
# COMMAND ----------

degreeRatio = inDeg.join(outDeg, "id")\
  .selectExpr("id", "double(inDegree)/double(outDegree) as degreeRatio")
degreeRatio.orderBy(desc("degreeRatio")).show(10, False)
```

```
+---------------------------------------+------------------+
|id                                     |degreeRatio       |
+---------------------------------------+------------------+
|Redwood City Medical Center            |1.5333333333333334|
|San Mateo County Center                |1.4724409448818898|
|SJSU 4th at San Carlos                 |1.3621052631578947|
|San Francisco Caltrain (Townsend at 4th)|1.3233728710462287|
|Washington at Kearny                   |1.3086466165413533|
|Paseo de San Antonio                   |1.2535046728971964|
|California Ave Caltrain Station        |1.24              |
|Franklin at Maple                      |1.2345679012345678|
|Embarcadero at Vallejo                 |1.2201707365495336|
|Market at Sansome                      |1.2173913043478262|
+---------------------------------------+------------------+
only showing top 10 rows
```

In [22]:
```
degreeRatio.orderBy("degreeRatio").show(10, False)
```

```
+------------------------------+------------------+
|id                            |degreeRatio       |
+------------------------------+------------------+
|Grant Avenue at Columbus Avenue|0.5180520570948782|
|2nd at Folsom                 |0.5909488686085761|
|Powell at Post (Union Square) |0.6434241245136186|
|Mezes Park                    |0.6839622641509434|
|Evelyn Park and Ride          |0.7413087934560327|
|Beale at Market               |0.75726761574351  |
|Golden Gate at Polk           |0.7822270981897971|
|Ryland Park                   |0.7857142857142857|
|San Francisco City Hall       |0.7928849902534113|
|Palo Alto Caltrain Station    |0.8064516129032258|
+------------------------------+------------------+
only showing top 10 rows
```

In [23]:
```
# COMMAND ----------

stationGraph.bfs(fromExpr="id = 'Townsend at 7th'",
  toExpr="id = 'Spear at Folsom'", maxPathLength=2).show(10)
```

```
+--------------------+--------------------+--------------------+
|                from|                  e0|                  to|
+--------------------+--------------------+--------------------+
|[65, Townsend at ...|[913371, 663, 8/3...|[49, Spear at Fol...|
|[65, Townsend at ...|[913265, 658, 8/3...|[49, Spear at Fol...|
|[65, Townsend at ...|[911919, 722, 8/3...|[49, Spear at Fol...|
|[65, Townsend at ...|[910777, 704, 8/2...|[49, Spear at Fol...|
|[65, Townsend at ...|[908994, 1115, 8/...|[49, Spear at Fol...|
|[65, Townsend at ...|[906912, 892, 8/2...|[49, Spear at Fol...|
|[65, Townsend at ...|[905201, 980, 8/2...|[49, Spear at Fol...|
|[65, Townsend at ...|[904010, 969, 8/2...|[49, Spear at Fol...|
|[65, Townsend at ...|[903375, 850, 8/2...|[49, Spear at Fol...|
|[65, Townsend at ...|[899944, 910, 8/2...|[49, Spear at Fol...|
+--------------------+--------------------+--------------------+
only showing top 10 rows
```

In [24]:
```
# COMMAND ----------

spark.sparkContext.setCheckpointDir("/tmp/checkpoints")
```

In [25]:
```
# COMMAND ----------

minGraph = GraphFrame(stationVertices, tripEdges.sample(False, 0.1))
cc = minGraph.connectedComponents()
```

In [26]:
```
# COMMAND ----------

cc.where("component != 0").show(5)
```

```
+----------+--------------------+---------+----------+---------+------------+------------+-------
-----+
|station_id|                  id|      lat|      long|dockcount|    landmark|installation|   comp
onent|
+----------+--------------------+---------+----------+---------+------------+------------+-------
-----+
|        47|     Post at Kearney|37.788975|-122.403452|       19|San Francisco|   8/19/2013|3178275
79904|
|        46|Washington at Kea...|37.795425|-122.404767|       15|San Francisco|   8/19/2013| 171798
69184|
+----------+--------------------+---------+----------+---------+------------+------------+-------
-----+
```

In [27]:
```
# COMMAND ----------

scc = minGraph.stronglyConnectedComponents(maxIter=3)
```

In [28]:
```
# COMMAND ----------
```

In [30]:
```
type(scc)
```

Out[30]:
```
pyspark.sql.dataframe.DataFrame
```

In [31]:
```
scc.count()
```

Out[31]:
```
70
```

In [33]:
```
scc.show(10,False)
```

```
+----------+-----------------------------------+--------+----------+---------+------------+-----
-------+------------+
|station_id|id                                 |lat     |long      |dockcount|landmark    |insta
llation|component   |
+----------+-----------------------------------+--------+----------+---------+------------+-----
-------+------------+
|11        |MLK Library                        |37.335885|-121.88566 |19       |San Jose    |8/6/2
013    |128849018880|
|80        |Santa Clara County Civic Center    |37.352601|-121.905733|15       |San Jose    |12/3
1/2013  |128849018880|
|64        |2nd at South Park                  |37.782259|-122.392738|15       |San Francisco|8/22/
2013   |0           |
|36        |California Ave Caltrain Station    |37.429082|-122.142805|15       |Palo Alto   |8/14/
2013   |0           |
|62        |2nd at Folsom                      |37.785299|-122.396236|19       |San Francisco|8/22/
2013   |0           |
|5         |Adobe on Almaden                   |37.331415|-121.8932  |19       |San Jose    |8/5/2
013    |128849018880|
|75        |Mechanics Plaza (Market at Battery)|37.7913  |-122.399051|19       |San Francisco|8/25/
2013   |0           |
|63        |Howard at 2nd                      |37.786978|-122.398108|19       |San Francisco|8/22/
2013   |0           |
|83        |Mezes Park                         |37.491269|-122.236234|15       |Redwood City |2/20/
2014   |0           |
|42        |Davis at Jackson                   |37.79728 |-122.398436|15       |San Francisco|8/19/
2013   |0           |
+----------+-----------------------------------+--------+----------+---------+------------+-----
-------+------------+
only showing top 10 rows
```

In [ ]:
```

```