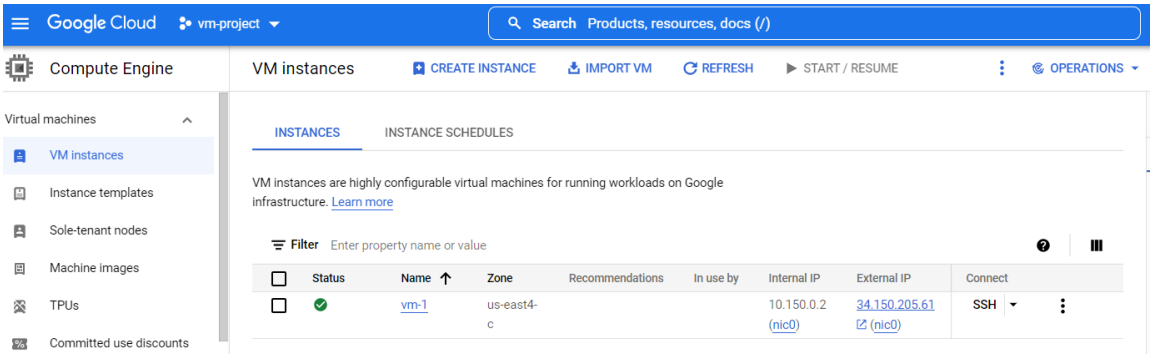
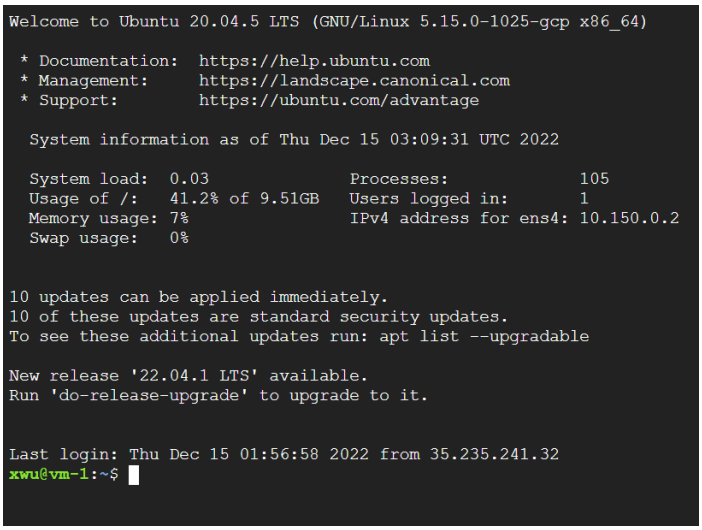
Week14 Homework1: PySpark: DataFrames / SparkSQL + GraphFrames / GraphX

1. Create a GCP project and **compute engine VM instance** “vm-1”, and ssh login to vm-1 terminal

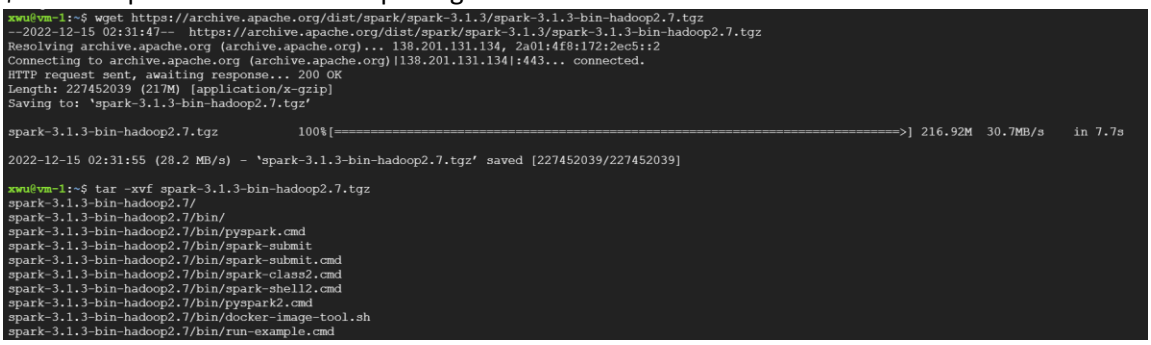




2. Install PySpark, java 11

$ wget https://archive.apache.org/dist/spark/spark-3.1.3/spark-3.1.3-bin-hadoop2.7.tgz

$ tar -xvf spark-3.1.3-bin-hadoop2.7.tgz



Set environmental variables in .bashrc

$ ln -s spark-3.1.3-bin-hadoop2.7 spark

$ vi .bashrc

$ source .bashrc

$ cat .bashrc

```
xwu@vm-1:~$ vi .bashrc
xwu@vm-1:~$ source .bashrc
xwu@vm-1:~$ cat .bashrc
export SPARK_HOME=/home/xwu/spark
export PATH=$SPARK_HOME/bin:$PATH
export PATH=$SPARK_HOME/sbin:$PATH
export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
```

Verify pyspark

```
xwu@vm-1:~$ pyspark
Python 3.8.10 (default, Jun 22 2022, 20:18:18)
[GCC 9.4.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/home/xwu/spark-3.1.3-bin-hadoop2.7/jars/spark-unsafe_2.12-3.1.3.jar) to construct
or java.nio.DirectByteBuffer(long,int)
WARNING: Please consider reporting this to the maintainers of org.apache.spark.unsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
22/12/15 02:32:49 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.1.3
      /_/

Using Python version 3.8.10 (default, Jun 22 2022 20:18:18)
Spark context Web UI available at http://vm-1.us-east4-c.c.vm-project-370900.internal:4040
Spark context available as 'sc' (master = local[*], app id = local-1671071572082).
SparkSession available as 'spark'.
>>>
```

3. Prepare input data files

$ mkdir in

$ cd in

$ vi person.csv

$ vi relationship.csv

$ cat person.csv

$ cat relationship.csv

```
xwu@vm-1:~$ cd in
xwu@vm-1:~/in$ cat person.csv
id,Name,Age
1,Andrew,45
2,Sierra,43
3,Bob,12
4,Emily,10
5,William,35
6,Rachel,32
xwu@vm-1:~/in$ cat relationship.csv
src,dst,relation
1,2,Husband
1,3,Father
1,4,Father
1,5,Friend
1,6,Friend
2,1,Wife
2,3,Mother
2,4,Mother
2,6,Friend
3,1,Son
3,2,Son
4,1,Daughter
4,2,Daughter
5,1,Friend
6,1,Friend
6,2,Friend
xwu@vm-1:~/in$
```

4. Prepare script file - pyspark_graphX.py

```
# Import PySpark
import pyspark
from pyspark.sql import SparkSession

#Create SparkSession
spark =
SparkSession.builder.master("local[1]").appName("pysparkGraphX").getOrCreate()

from graphframes import *

# Recipe 9-1. Create GraphFrames
#     person dataframe : id, Name, age
personsDf = spark.read.csv('in/person.csv',header=True, inferSchema=True)

# Create a "persons" SQL table from personsDF DataFrame
personsDf.createOrReplaceTempView("persons")
spark.sql("select * from persons").show()

# relationship dataframe : src, dst, relation
relationshipDf = spark.read.csv('in/relationship.csv',header=True, inferSchema=True)
relationshipDf.createOrReplaceTempView("relationship")
spark.sql("select * from relationship").show()

# - Create a GraphFrame from both person and relationship dataframes
#     >>> graph
#     GraphFrame(v:[id: int, Name: string ... 1 more field], e:[src:
#     int, dst: int ... 1 more field])
# - A GraphFrame that contains v and e.
#   + The v represents vertices and e represents edges.
graph = GraphFrame(personsDf, relationshipDf)

# - Degrees represent the number of edges that are connected to a vertex.
#   + GraphFrame supports inDegrees and outDegrees.
#     - inDegrees give you the number of incoming links to a vertex.
#     - outDegrees give the number of outgoing edges from a node.
```

```
# - Find all the edges connected to Andrew.
graph.degrees.filter("id = 1").show()

# Find the number of incoming links to Andrew
graph.inDegrees.filter("id = 1").show()

# Find the number of links coming out from Andrew using the outDegrees
graph.outDegrees.filter("id = 1").show()

# Recipe 9-2. Apply Triangle Counting in a GraphFrame
# - Find how many triangle relationships the vertex is participating in
personsTriangleCountDf = graph.triangleCount()
personsTriangleCountDf.show()

# Create a "personsTriangleCount" SQL table from the
# personsTriangleCountDf DataFrame
personsTriangleCountDf.createOrReplaceTempView("personsTriangleCount")

# Create a "personsMaxTriangleCount" SQL table from the
# maxCountDf DataFrame
maxCountDf = spark.sql("select max(count) as max_count from personsTriangleCount")
maxCountDf.createOrReplaceTempView("personsMaxTriangleCount")

spark.sql("select * from personsTriangleCount P JOIN (select * from
personsMaxTriangleCount) M ON (M.max_count = P.count) ").show()

# Recipe 9-3. Apply a PageRank Algorithm
pageRank = graph.pageRank(resetProbability=0.20, maxIter=10)
pageRank.vertices.printSchema()

pageRank.vertices.orderBy("pagerank",ascending=False).show()

pageRank.edges.orderBy("weight",ascending=False).show()

# Recipe 9-4. Apply the Breadth First Algorithm
graph.bfs(fromExpr = "Name='Bob'",toExpr = "Name='William'").show()

graph.bfs(fromExpr = "age < 20", toExpr = "name = 'Rachel'").show()
graph.bfs(fromExpr = "age < 20", toExpr = "name = 'Rachel'", edgeFilter = "relation !=
'Son'").show()
```

5. Run pyspark_graphX.py
   In case that there is no numpy library installed in your virtual machine:
   $ sudo apt install python3-pip

$ pip3 install numpy

Submit the job using the command:
$ spark-submit --packages graphframes:graphframes:0.8.2-spark3.1-s_2.12 pyspark_graphX.py

Note: Please choose the compatible version of graphframes package to match with pyspark version. ([graphframes (spark-packages.org)]

```
xwu@vm-1:~$ spark-submit --packages graphframes:graphframes:0.8.2-spark3.1-s_2.12 pyspark_graphX.py
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/home/xwu/spark-3.1.3-bin-hadoop2.7/jars/spark-unsafe_2.12-3.1.
r java.nio.DirectByteBuffer(long,int)
WARNING: Please consider reporting this to the maintainers of org.apache.spark.unsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
:: loading settings :: url = jar:file:/home/xwu/spark-3.1.3-bin-hadoop2.7/jars/ivy-2.4.0.jar!/org/apache/ivy/core/settings/ivysettings.xml
Ivy Default Cache set to: /home/xwu/.ivy2/cache
The jars for the packages stored in: /home/xwu/.ivy2/jars
graphframes#graphframes added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-9d703660-0d8c-416e-ad5e-bbffc5f34078;1.0
        confs: [default]
        found graphframes#graphframes;0.8.2-spark3.1-s_2.12 in spark-packages
        found org.slf4j#slf4j-api;1.7.16 in central
:: resolution report :: resolve 460ms :: artifacts dl 12ms
        :: modules in use:
        graphframes#graphframes;0.8.2-spark3.1-s_2.12 from spark-packages in [default]
        org.slf4j#slf4j-api;1.7.16 from central in [default]
        ---------------------------------------------------------------------
        |                  |            modules            ||   artifacts   |
        |       conf       | number| search|dwnlded|evicted|| number|dwnlded|
        ---------------------------------------------------------------------
        |      default     |   2   |   0   |   0   |   0   ||   2   |   0   |
        ---------------------------------------------------------------------
:: retrieving :: org.apache.spark#spark-submit-parent-9d703660-0d8c-416e-ad5e-bbffc5f34078
        confs: [default]
        0 artifacts copied, 2 already retrieved (0kB/10ms)
22/12/15 11:20:15 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicab
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
```

6. Result
   GraphFrame:

```
+---+-------+---+
| id|   Name|Age|
+---+-------+---+
|  1| Andrew| 45|
|  2| Sierra| 43|
|  3|    Bob| 12|
|  4|  Emily| 10|
|  5|William| 35|
|  6| Rachel| 32|
+---+-------+---+

+---+---+--------+
|src|dst|relation|
+---+---+--------+
|  1|  2| Husband|
|  1|  3|  Father|
|  1|  4|  Father|
|  1|  5|  Friend|
|  1|  6|  Friend|
|  2|  1|    Wife|
|  2|  3|  Mother|
|  2|  4|  Mother|
|  2|  6|  Friend|
|  3|  1|     Son|
|  3|  2|     Son|
|  4|  1|Daughter|
|  4|  2|Daughter|
|  5|  1|  Friend|
|  6|  1|  Friend|
|  6|  2|  Friend|
+---+---+--------+
```

```
+---+------+
| id|degree|
+---+------+
|  1|    10|
+---+------+

+---+--------+
| id|inDegree|
+---+--------+
|  1|       5|
+---+--------+

+---+---------+
| id|outDegree|
+---+---------+
|  1|        5|
+---+---------+
```

TriangleCount:

```
+-----+---+-------+---+
|count| id|   Name|Age|
+-----+---+-------+---+
|    3|  1| Andrew| 45|
|    1|  6| Rachel| 32|
|    1|  3|    Bob| 12|
|    0|  5|William| 35|
|    1|  4|  Emily| 10|
|    3|  2| Sierra| 43|
+-----+---+-------+---+

+-----+---+------+---+---------+
|count| id|  Name|Age|max_count|
+-----+---+------+---+---------+
|    3|  1|Andrew| 45|        3|
|    3|  2|Sierra| 43|        3|
+-----+---+------+---+---------+
```

PageRank:

```
root
 |-- id: integer (nullable = true)
 |-- Name: string (nullable = true)
 |-- Age: integer (nullable = true)
 |-- pagerank: double (nullable = true)

+---+-------+---+------------------+
| id|   Name|Age|          pagerank|
+---+-------+---+------------------+
|  1| Andrew| 45| 1.787923121897472|
|  2| Sierra| 43| 1.406016795082752|
|  6| Rachel| 32|0.7723665979473922|
|  4|  Emily| 10|0.7723665979473922|
|  3|    Bob| 12|0.7723665979473922|
|  5|William| 35|0.4889602891776001|
+---+-------+---+------------------+
```

```
+---+---+--------+------+
|src|dst|relation|weight|
+---+---+--------+------+
|  5|  1|  Friend|   1.0|
|  3|  1|     Son|   0.5|
|  4|  1|Daughter|   0.5|
|  4|  2|Daughter|   0.5|
|  6|  1|  Friend|   0.5|
|  3|  2|     Son|   0.5|
|  6|  2|  Friend|   0.5|
|  2|  3|  Mother|  0.25|
|  2|  4|  Mother|  0.25|
|  2|  1|    Wife|  0.25|
|  2|  6|  Friend|  0.25|
|  1|  2| Husband|   0.2|
|  1|  6|  Friend|   0.2|
|  1|  3|  Father|   0.2|
|  1|  4|  Father|   0.2|
|  1|  5|  Friend|   0.2|
+---+---+--------+------+
```

BFS:

```
+------------+----------+--------------+-------------+----------------+
|        from|        e0|            v1|           e1|              to|
+------------+----------+--------------+-------------+----------------+
|{3, Bob, 12}|{3, 1, Son}|{1, Andrew, 45}|{1, 5, Friend}|{5, William, 35}|
+------------+----------+--------------+-------------+----------------+

+--------------+----------------+--------------+-------------+---------------+
|          from|              e0|            v1|           e1|             to|
+--------------+----------------+--------------+-------------+---------------+
|{4, Emily, 10}|{4, 1, Daughter}|{1, Andrew, 45}|{1, 6, Friend}|{6, Rachel, 32}|
|  {3, Bob, 12}|     {3, 1, Son}|{1, Andrew, 45}|{1, 6, Friend}|{6, Rachel, 32}|
|{4, Emily, 10}|{4, 2, Daughter}|{2, Sierra, 43}|{2, 6, Friend}|{6, Rachel, 32}|
|  {3, Bob, 12}|     {3, 2, Son}|{2, Sierra, 43}|{2, 6, Friend}|{6, Rachel, 32}|
+--------------+----------------+--------------+-------------+---------------+

+--------------+----------------+--------------+-------------+---------------+
|          from|              e0|            v1|           e1|             to|
+--------------+----------------+--------------+-------------+---------------+
|{4, Emily, 10}|{4, 1, Daughter}|{1, Andrew, 45}|{1, 6, Friend}|{6, Rachel, 32}|
|{4, Emily, 10}|{4, 2, Daughter}|{2, Sierra, 43}|{2, 6, Friend}|{6, Rachel, 32}|
+--------------+----------------+--------------+-------------+---------------+
```

DONE!!!

7.