

Week 14: Optional Homework 1: Project: PySpark: DataFrames / SparkSQL + GraphFrames / GraphX

https://hc.labnet.sfbu.edu/~henry/sfbu/course/pyspark_sql_recipes/graphframes/slide/exercise_graphframes.html

Q2 ==> PySpark: DataFrames / SparkSQL + GraphFrames / GraphX

Project: DataFrames / SparkSQL + GraphFrames / GraphX

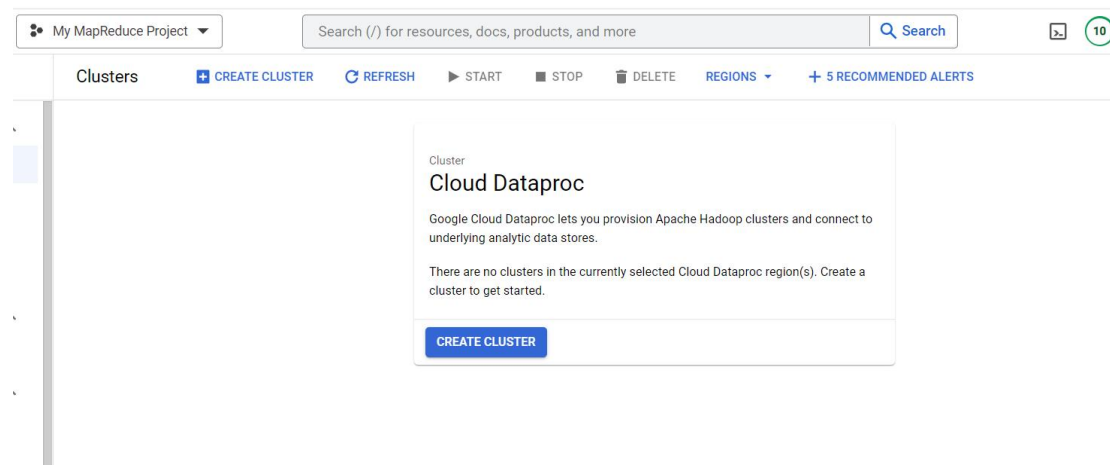
- Step 1: [PySpark: DataFrames / SparkSQL + GraphFrames / GraphX](#)
- Step 2: [Update your portfolio about this project](#)
- Step 3: Submit a PDF file document showing the procedure as part of the homework answers.
- Step 4: Submit the URL of your GitHub webpage as part of the homework answers.

- GitHub directory structure
-

Cloud Computing

Apache Spark + GraphFrame + GraphX

1. Create cluster on GCP



Create Dataproc cluster

Select the infrastructure service that you want to use.

Cluster on Compute Engine

Create the cluster on Compute Engine.

CREATE

Cluster on GKE

Create the cluster on Google Kubernetes Engine (GKE).

CREATE

CANCEL

gle Cloud

My MapReduce Project

Search (/) for resources, docs, products, and more

Search

10

proc

Create a Dataproc cluster on Compute Engine

Set up cluster
Begin by providing basic information.

Configure nodes (optional)
Change node compute and storage capabilities.

Customize cluster (optional)
Add cluster properties, features, and actions.

Manage security (optional)
Change access, encryption, and security settings.

CREATE

CANCEL

EQUIVALENT COMMAND LINE

Name

Cluster Name *
cluster-84ef

Location

Region *
us-central1

Zone *
us-central1-f

Cluster type

☒ Standard (1 master, N workers)

☐ Single Node (1 master, 0 workers)
Provides one node that acts as both master and worker. Good for proof-of-concept or small-scale processing

☐ High Availability (3 masters, N workers)
Hadoop High Availability mode provides uninterrupted YARN and HDFS operations despite single-node failures or reboots

Autoscaling

Automates cluster resource management based on an autoscaling policy.

Policy
None

My MapReduce Project

Search (/) for resources, docs, products, and more

Search

10

?

Clusters

CREATE CLUSTER

REFRESH

START

STOP

DELETE

REGIONS

+ 5 RECOMMENDED ALERTS

SHOW INFO PANE

Filter

Search clusters, press Enter

	Name	Status	Region	Zone	Total worker nodes	Scheduled deletion	Cloud Storage staging bucket	Created
<input type="checkbox"/>	cluster-84ef	Running	us-central1	us-central1-f	2	Off	dataproc-staging-us-central1-166467073201-mbd34cmp	Dec 17, 2022, 1:57:53 PM

My MapReduce Project Search (/) for resources, docs, products, and more Search 10 ? ⋮ R

[Cluster details](#) [SUBMIT JOB](#) [REFRESH](#) [START](#) [STOP](#) [DELETE](#) [VIEW LOGS](#)

Consider using Auto Zone rather than selecting a zone manually. See <https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/auto-zone> [MORE](#)

Name	cluster-84ef
Cluster UUID	66d7707e-7d4a-47e7-ae6e-a8d947e23045
Type	Dataproc Cluster
Status	Running

MONITORING JOBS **VM INSTANCES** CONFIGURATION WEB INTERFACES

Filter Filter instances ? ⋮

	Name	Role	SSH
✓	cluster-84ef-m	Master	SSH
✓	cluster-84ef-w-0	Worker	
✓	cluster-84ef-w-1	Worker	

[EQUIVALENT REST](#)

- Open in browser window
- Open in browser window on custom port
- Open in browser window using provided private SSH key
- View gcloud command
- Use another SSH client

ssh.cloud.google.com/v2/ssh/projects/my-mapreduce-project-367604/zones/us-central1-f/instances/cluster-84ef-m?authuser=0&hl=en...

SSH-in-browser [UPLOAD FILE](#) [DOWNLOAD FILE](#) ! ⌨ ⚙

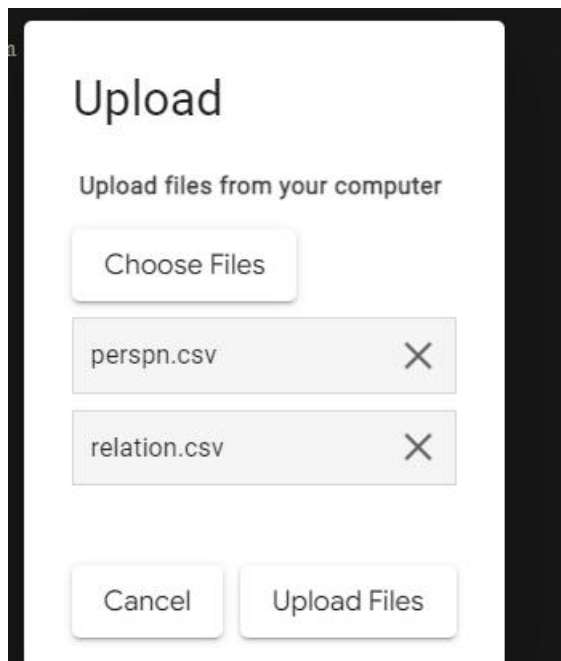
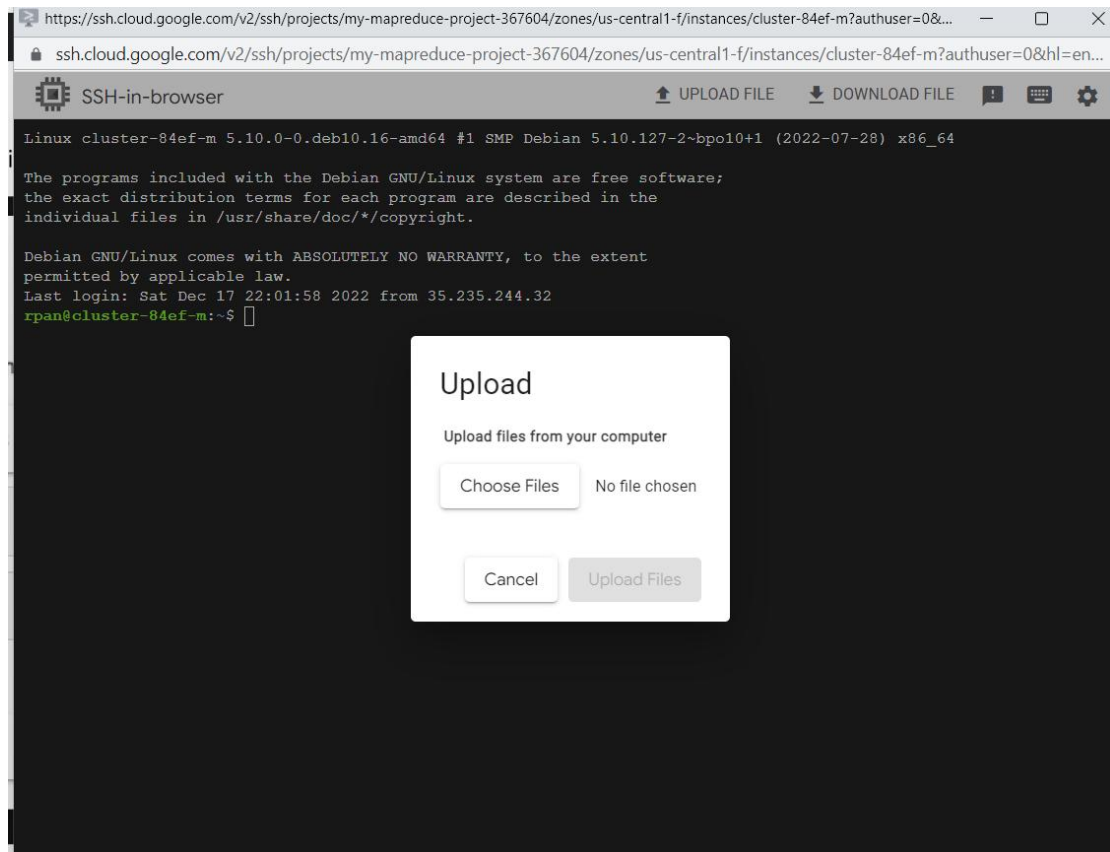
```
Linux cluster-84ef-m 5.10.0-0.deb10.16-amd64 #1 SMP Debian 5.10.127-2~bpo10+1 (2022-07-28) x86_64

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Sat Dec 17 22:01:58 2022 from 35.235.244.32
rpan@cluster-84ef-m:~$
```

2. Prepare Data

Upload csv data from local to cluster





Check uploaded file:

```
rpan@cluster-84ef-m:~$ ls
perspn.csv relation.csv
rpan@cluster-84ef-m:~$
```

3. Create HDFS file system and copy the data to hdfs:

\$ hdfs dfs -mkdir hdfs:///mydata

```
rpan@cluster-84ef-m:~$ hdfs dfs -mkdir dhfs:///mydata
mkdir: No FileSystem for scheme "dhfs"
rpan@cluster-84ef-m:~$ hdfs dfs -mkdir hdfs:///mydata
rpan@cluster-84ef-m:~$ hdfs dfs -put ./csv hdfs:///mydata/
rpan@cluster-84ef-m:~$ hdfs dfs -ls hdfs:///mydata
Found 2 items
-rw-r--r--  2 rpan  hadoop      7168 2022-12-17 22:18 hdfs:///mydata/perspn.csv
-rw-r--r--  2 rpan  hadoop      207 2022-12-17 22:18 hdfs:///mydata/relation.csv
rpan@cluster-84ef-m:~$
```

rpan@cluster-84ef-m:~\$ hdfs dfs -mkdir hdfs:///mydata

rpan@cluster-84ef-m:~\$ hdfs dfs -put ./csv hdfs:///mydata/

rpan@cluster-84ef-m:~\$ hdfs dfs -ls hdfs:///mydata

4. Create the graphdemo.py file and change the path for data files:


```
rpan@cluster-84ef-m:~$ pyspark --version
Welcome to

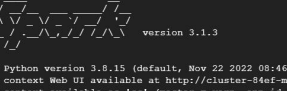
 version 3.1.3

Using Scala version 2.12.14, OpenJDK 64-Bit Server VM, 1.8.0_352
Branch HEAD
Compiled by user on 2022-11-01T22:00:39Z
Revision b28f046c307a8374984c0231d76debe3a33be97
Url https://bigdataoss-internal.googleusercontent.com/third_party/apache/spark
Type --help for more information.
rpan@cluster-84ef-m:~$
```

Transferred 2 items

rpan@cluster-84ef-m:~\$ pyspark --packages graphframes:graphframes:0.8.2-spark3.1-s_2.12

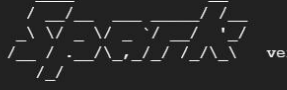
```
Python 3.8.15 | packaged by conda-forge | (default, Nov 22 2022, 08:46:39)
[GCC 10.4.0] on linux
Type "help", "copyright", "credits" or "license()" for more information.
:: loading settings: url = jar:file:/usr/lib/spark/jars/ivy-2.4.0.jar!/org/apache/ivy/core/settings/ivysettings.xml
Ivy Default Cache set to: /home/rpan/.ivy2/cache
The jars for the packages stored in: /home/rpan/.ivy2/jars
graphframes:graphframes added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-469ad0e0-babf-475d-a652-9df60606d73b;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
  downloading https://repos.spark-packages.org/graphframes/graphframes/0.8.2-spark3.1-s_2.12/graphframes-0.8.2-spark3.1-s_2.12.jar ...
    [SUCCESSFUL ] graphframes:graphframes:0.8.2-spark3.1-s_2.12!graphframes.jar (58ms)
  downloading https://repo1.maven.org/maven2/org/slf4j/slf4j-api/1.7.16/slf4j-api-1.7.16.jar ...
    [SUCCESSFUL ] org.slf4j:slf4j-api:1.7.16!slf4j-api.jar (22ms)
  :: resolution report :: resolve 1844ms :: artifacts dl 84ms
    :: 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]
  |-----|
  |  conf  | | number | search | modules | | artifacts | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
  | default | | 2      | 2      | 2      | | 0        | | 2        | | 2        |
  |-----|
  :: retrieving :: org.apache.spark#spark-submit-parent-469ad0e0-babf-475d-a652-9df60606d73b
    confs: [default]
    2 artifacts copied, 0 already retrieved (281kB/6ms)
Setting default log level to "WARN"
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
22/12/17 22:54:23 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
22/12/17 22:54:23 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
22/12/17 22:54:23 INFO org.apache.spark.SparkEnv: Registering BlockManagerMasterHeartbeat
22/12/17 22:54:23 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
22/12/17 22:54:25 WARN org.apache.spark.deploy.yarn.Client: Same path resource file:///home/rpan/.ivy2/jars/graphframes_graphframes-0.8.2-spark3.1-s_2.12.jar added multiple times to distributed
d cache.
22/12/17 22:54:25 WARN org.apache.spark.deploy.yarn.Client: Same path resource file:///home/rpan/.ivy2/jars/org.slf4j_slf4j-api-1.7.16.jar added multiple times to distributed cache.
Welcome to

 version 3.1.3

Using Python version 3.8.15 (default, Nov 22 2022 08:46:39)
Spark context Web UI available at http://cluster-84ef-m.us-central1-f.c.my-mapreduce-project-367604.internal:39027
Spark context available as 'sc' (master = yarn, app id = application_1671314333219_0002).
SparkSession available as 'spark'.
>>>
```

```
https://ssh.cloud.google.com/v2/ssh/projects/my-mapreduce-project-367604/zones/us-central1-a/instances/cluster-d3b4-m?authuser=0...
ssh.cloud.google.com/v2/ssh/projects/my-mapreduce-project-367604/zones/us-central1-a/instances/cluster-d3b4-m?authuser=0&hl=e...
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE

phframes_graphframes-0.8.2-spark2.4-s_2.11.jar added multiple times to distributed cache.
22/12/19 17:59:57 WARN org.apache.spark.deploy.yarn.Client: Same path resource file:///home/rpan/.ivy2/jars/org
.slf4j_slf4j-api-1.7.16.jar added multiple times to distributed cache.
Welcome to

 version 3.1.3

Using Python version 3.8.15 (default, Nov 22 2022 08:46:39)
Spark context Web UI available at http://cluster-d3b4-m.us-central1-a.c.my-mapreduce-project-367604.internal:41
787
Spark context available as 'sc' (master = yarn, app id = application_1671469931148_0003).
SparkSession available as 'spark'.
>>> graph=GraphFrame(personDF,relationDF)
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
NameError: name 'GraphFrame' is not defined
>>> from graphframes import *
>>> personDF=spark.read.csv('hdfs:///mydata/person.csv',header=True,inferSchema=True)
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
  File "/usr/lib/spark/python/pyspark/sql/readwriter.py", line 737, in csv
    return self._df(self._reader.csv(self._spark._sc._jvm.PythonUtils.toSeq(path)))
  File "/usr/lib/spark/python/lib/py4j-0.10.9-src.zip/py4j/java_gateway.py", line 1304, in __call__
  File "/usr/lib/spark/python/pyspark/sql/utils.py", line 117, in deco
    raise converted from None
pyspark.sql.utils.AnalysisException: Path does not exist: hdfs://cluster-d3b4-m/mydata/person.csv
>>> ss
```

Pyspark runs successfully, run remaining code:


```

Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
AttributeError: 'SparkContext' object has no attribute 'setLogLevel'
>>> sc.setLogLevel("ERROR")
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
AttributeError: 'SparkContext' object has no attribute 'setLogLevel'
>>> sc.setLogLevel("ERROR")
>>> sqlContext = SQLContext(sc)
>>> personsDf = spark.read.csv('hdfs://mydata/person.csv',header=True, inferSchema=True)
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
  File "/usr/lib/spark/python/pyspark/sql/readwriter.py", line 737, in csv
    return self._df(self._jreader.csv(self._spark._sc._jvm.PythonUtils.toSeq(path)))
  File "/usr/lib/spark/python/lib/py4j-0.10.9-src.zip/py4j/java_gateway.py", line 1304, in __call__
  File "/usr/lib/spark/python/pyspark/sql/utils.py", line 117, in deco
    raise converted from None
pyspark.sql.utils.IllegalArgumentException: java.net.UnknownHostException: mydata
>>> personsDf = spark.read.csv('hostname/rpan/mydata/person.csv',header=True, inferSchema=True)
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
  File "/usr/lib/spark/python/pyspark/sql/readwriter.py", line 737, in csv
    return self._df(self._jreader.csv(self._spark._sc._jvm.PythonUtils.toSeq(path)))
  File "/usr/lib/spark/python/lib/py4j-0.10.9-src.zip/py4j/java_gateway.py", line 1304, in __call__
  File "/usr/lib/spark/python/pyspark/sql/utils.py", line 117, in deco
    raise converted from None
pyspark.sql.utils.IllegalArgumentException: java.net.UnknownHostException: hostname
>>> personsDf = spark.read.csv('hdfs://35.235.241.19/rpan/mydata/person.csv',header=True, inferSchema=True)

```

```

Using Python version 3.8.15 (default, Nov 22 2022 08:46:39)
Spark context Web UI available at http://cluster-d3b4-m.us-central1-a.c.my-mapreduce-project-367604.internal:37647
Spark context available as 'sc' (master = yarn, app id = application_1671495488392_0006).
SparkSession available as 'spark'.
>>> from graphframes import *
>>> personsDf = spark.read.csv('hdfs:///mydata/person.csv',header=True,inferSchema=True)
>>> personsDf = spark.read.csv('hdfs:///mydata/person.csv',header=True,inferSchema=True)
>>> personsDf.createOrReplaceTempView("persons")
>>> spark.sql("select * from persons").show()
+----+-----+----+
| id | Name | Age |
+----+-----+----+
| 1 | Andrew | 45 |
| 2 | Sierra | 43 |
| 3 | Bob | 12 |
| 4 | Emily | 10 |
| 5 | William | 35 |
| 6 | Rachel | 32 |
+----+-----+----+
>>>

```

```

>>> graph=GraphFrame(personsDf,relationshipDf)
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
NameError: name 'relationshipDf' is not defined
>>> relationshipDf=spark.read.csv("hdfs:///mydata/relation.csv",header=True,inferSchema=True)
>>> relationshipDf.createOrReplaceTempView("relationship")
>>> spark.sql.filter("id=1").show()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
AttributeError: 'function' object has no attribute 'filter'
>>> spark.sql("select * from relationship").show()
+----+---+-----+
|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 |
+----+---+-----+
>>>

```

```
https://ssh.cloud.google.com/v2/ssh/projects/my-mapreduce-project-367604/zones/us-central1-a/instances/cluster-d3b4-m?authuser=0&hl=en_US&projectNumber=166467073201&useAdminProxy=true&troubleshoot255Enabled=tru...
ssh.cloud.google.com/v2/ssh/projects/my-mapreduce-project-367604/zones/us-central1-a/instances/cluster-d3b4-m?authuser=0&hl=en_US&projectNumber=166467073201&useAdminProxy=true&troubleshoot255Ena...
SSH-in-browser
[...]
```

```
>>> relationshipDF=spark.read.csv("hdfs:///mydata/relation.csv",header=True,inferSchema=True)
>>> relationshipDF.createOrReplaceTempView("relationship")
>>> spark.sql("select * from relationship").show()
+---+-----+
|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|
+---+-----+
```

Deleted the incorrect file perspn.csv and re-uploaded person.csv:

```
rpan@cluster-d3b4-m:~$ ls
graphdemo.py perspn.csv relation.csv
rpan@cluster-d3b4-m:~$ hdfs dfs -rm hdfs:///mydata/perspn.csv
Deleted hdfs:///mydata/perspn.csv
rpan@cluster-d3b4-m:~$ ls
graphdemo.py perspn.csv relation.csv
rpan@cluster-d3b4-m:~$ hdfs dfs -ls hdfs:///mydata/
Found 1 items
-rw-r--r--  2 rpan hadoop    207 2022-12-19 17:21 hdfs:///mydata/relation.csv
rpan@cluster-d3b4-m:~$ hdfs dfs -put ./person.csv hdfs:///mydata/
rpan@cluster-d3b4-m:~$ hdfs dfs -ls hdfs:///mydata
Found 2 items
-rw-r--r--  2 rpan hadoop    88 2022-12-20 01:22 hdfs:///mydata/person.csv
-rw-r--r--  2 rpan hadoop    207 2022-12-19 17:21 hdfs:///mydata/relation.csv
```

```
rpan@cluster-d3b4-m:~$ ls
graphdemo.py perspn.csv relation.csv
rpan@cluster-d3b4-m:~$ hdfs dfs -rm hdfs:///mydata/perspn.csv
Deleted hdfs:///mydata/perspn.csv
rpan@cluster-d3b4-m:~$ ls
graphdemo.py perspn.csv relation.csv
rpan@cluster-d3b4-m:~$ hdfs dfs -ls hdfs:///mydata/
Found 1 items
-rw-r--r--  2 rpan hadoop    207 2022-12-19 17:21 hdfs:///mydata/relation.csv
rpan@cluster-d3b4-m:~$ hdfs dfs -put ./person.csv hdfs:///mydata/
rpan@cluster-d3b4-m:~$ hdfs dfs -ls hdfs:///mydata
Found 2 items
-rw-r--r--  2 rpan hadoop    88 2022-12-20 01:22 hdfs:///mydata/person.csv
-rw-r--r--  2 rpan hadoop    207 2022-12-19 17:21 hdfs:///mydata/relation.csv
```

```
>>> graph=GraphFrame(personsDF,relationshipDF)
>>> graph.degrees.filter("id=1").show()
+---+-----+
| id|degree|
+---+-----+
|  1|    10|
+---+-----+
```

```
>>> graph.inDegrees.filter("id=1").show()
+---+-----+
| id|inDegree|
+---+-----+
|  1|     5|
+---+-----+
```

```
>>> graph.outDegrees.filter("id=1").show()
+---+-----+
| id|outDegree|
+---+-----+
|  1|         5|
+---+-----+
```

```
>>> personsTriangleCountDf=graph.triangleCount()
>>> personTriangleCountDf.show()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
NameError: name 'personTriangleCountDf' is not defined
>>> personsTriangleCountDf.show()
+-----+---+-----+---+
|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|
+-----+---+-----+---+
```

```
>>> personsTriangleCountDf.createOrReplaceTempView("personsTriangleCount")
>>> maxCountDf=spark.sql("select max(count) as max_count from personsTriangleCount")
File "<stdin>", line 1
    maxCountDf=spark.sql("select max(count) as max_count from personsTriangleCount")
                                     ^
SyntaxError: EOL while scanning string literal
>>> 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()
+-----+---+-----+---+
|count| id|   Name|Age|max_count|
+-----+---+-----+---+
|    3|  1| Andrew| 45|         3|
|    3|  2| Sierra| 43|         3|
+-----+---+-----+---+
>>> []
```

```
>>> pageRank = graph.pageRank(resetProbability=0.20,maxIter=10)
>>> pageRank.vertices.printSchema()
root
 |-- id: integer (nullable = true)
 |-- Name: string (nullable = true)
 |-- Age: integer (nullable = true)
 |-- pagerank: double (nullable = true)
```

```
>>> pageRank.vertices.orderBy("pagerank",ascending=False).show()
+---+-----+---+-----+---+
| 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|
+---+-----+---+-----+---+
```

```
>>> pageRank.edges.orderBy("weight",ascending=False).show()
+---+-----+
|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|
+---+-----+
```

```
>>> graph.bfs(fromExpr="Name='Bob'",toExpr="Name='William'").show()
+-----+-----+-----+-----+-----+
|      from|      e0|      v1|      e1|      to|
+-----+-----+-----+-----+-----+
|{3, Bob, 12}|{3, 1, Son}|{1, Andrew, 45}|{1, 5, Friend}|{5, William, 35}|
+-----+-----+-----+-----+-----+
```

```
>>> graph.bfs(fromExpr="age<20",toExpr="name='Rachel'").show()
+-----+-----+-----+-----+-----+
|      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}|
+-----+-----+-----+-----+-----+
```

```
>>> graph.bfs(fromExpr="age<20",toExpr="name='Rachel'",edgeFilter="relation !='son'").show()
+-----+-----+-----+-----+-----+
|      from|      e0|      v1|      e1|      to|
+-----+-----+-----+-----+-----+
| {3, Bob, 12}| {3, 1, Son}|{1, Andrew, 45}|{1, 6, Friend}|{6, Rachel, 32}|
| {3, Bob, 12}| {3, 2, Son}|{2, Sierra, 43}|{2, 6, Friend}|{6, Rachel, 32}|
|{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}|
+-----+-----+-----+-----+-----+
```

6. Modify graphdemo.py

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

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

from graphframes import *

#####
# Recipe 9-1. Create GraphFrames
#####
```

7. Run the code(graphdemo.py) with spark-submit

\$ spark-submit --packages graphframes:graphframes:0.8.2-spark3.1-s_2.12 graphdemo.py

```
rpan@cluster-d3b4-m:~$ spark-submit --packages graphframes:graphframes:0.8.2-spark3.1-s_2.12 graphdemo.py
:: loading settings :: url = jar:file:/usr/lib/spark/jars/ivy-2.4.0.jar!/org/apache/ivy/core/settings/ivysettings.xml
Ivy Default Cache set to: /home/rpan/.ivy2/cache
The jars for the packages stored in: /home/rpan/.ivy2/jars
graphframes#graphframes added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-e9372f06-af47-4dd7-adff-e01fb99fd7e6;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 181ms :: artifacts dl 5ms
  :: 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-e9372f06-af47-4dd7-adff-e01fb99fd7e6
  confs: [default]
  0 artifacts copied, 2 already retrieved (0KB/5ms)
22/12/20 02:50:43 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
22/12/20 02:50:43 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
22/12/20 02:50:43 INFO org.apache.spark.SparkEnv: Registering BlockManagerMasterHeartbeat
```

```
:: retrieving :: org.apache.spark#spark-submit-parent-e9372f06-af47-4dd7-adff-e01fb99fd7e6
  confs: [default]
  0 artifacts copied, 2 already retrieved (0KB/5ms)
22/12/20 02:50:43 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
22/12/20 02:50:43 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
22/12/20 02:50:43 INFO org.apache.spark.SparkEnv: Registering BlockManagerMasterHeartbeat
22/12/20 02:50:43 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
22/12/20 02:50:43 INFO org.sparkproject.jetty.util.log: Logging initialized @4021ms to org.sparkproject.jetty.util.log.Slf4jLog
22/12/20 02:50:43 INFO org.sparkproject.jetty.server.Server: jetty-9.4.40.v20210413; built: 2021-04-13T20:42:42.668Z; git: b881a572662e1943a14ae12e7e1207989f218b74; jvm 1.8
2-b08
22/12/20 02:50:43 INFO org.sparkproject.jetty.server.Server: Started @4156ms
22/12/20 02:50:43 INFO org.sparkproject.jetty.server.AbstractConnector: Started ServerConnector@7481d1e[HTTP/1.1, (http/1.1)](0.0.0.0:46549)
22/12/20 02:50:45 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageImpl: Ignoring exception of type GoogleJsonResponseException;
ied object already exists with desired state.
+---+---+---+
| 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|
```

```
+---+---+---+
|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|
+---+-----+

+---+---+-----+---+
|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|
+---+---+-----+---+

```

```

+---+---+-----+---+-----+
|    3|  1|Andrew| 45|        3|
|    3|  2|Sierra| 43|        3|
+---+---+-----+---+-----+

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|

```

```

+---+---+-----+-----+
|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|
+---+---+-----+-----+

```

22/12/20 02:51:02 INFO org.graphframes.lib.BFS\$: GraphFrame.bfs found path of length 2.

```

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

```

22/12/20 02:51:04 INFO org.graphframes.lib.BFS\$: GraphFrame.bfs found path of length 2.

```

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

```

22/12/20 02:51:05 INFO org.graphframes.lib.BFS\$: GraphFrame.bfs found path of length 2.

```

+---+---+-----+-----+
| 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}|
+---+---+-----+-----+

```

22/12/20 02:51:02 INFO org.graphframes.lib.BFS\$: GraphFrame.bfs found path of length 2.

```

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

```

22/12/20 02:51:04 INFO org.graphframes.lib.BFS\$: GraphFrame.bfs found path of length 2.

```

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

```

22/12/20 02:51:05 INFO org.graphframes.lib.BFS\$: GraphFrame.bfs found path of length 2.

```

+---+---+-----+-----+
| 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}|
+---+---+-----+-----+

```

22/12/20 02:51:06 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@7481d41e(HTTP/1.1, (http/1.1)){0.0.0.0:0}
rpan@cluster-d3b4-m:~\$