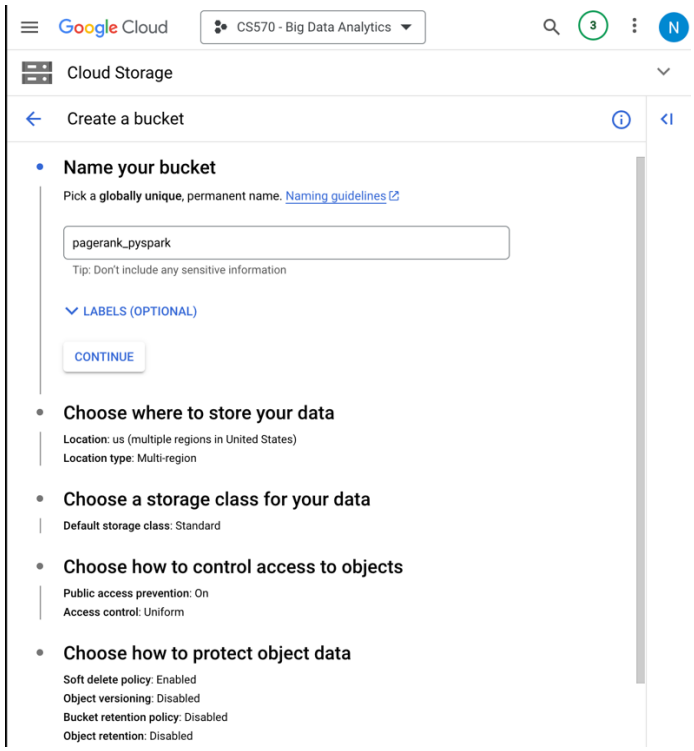
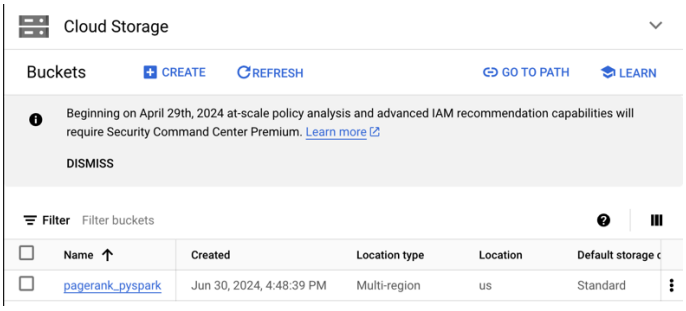
PART ONE – RUN USING PYSPARK

1. Create a bucket.





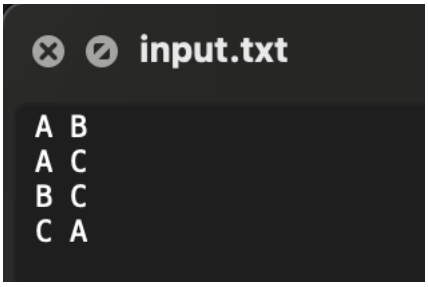
2. Create a cluster.

```
gcloud dataproc clusters create pagerank-cluster \
    --region=us-central1 \
    --zone=us-central1-a \
    --single-node \
    --master-machine-type=n1-standard-4 \
    --master-boot-disk-size=50GB \
    --image-version=1.5-debian10
```

```
nhaile96456@cloudshell:~ (cs570-big-data-analytics)$ gcloud dataproc clusters create pager
ank-cluster \
    --region=us-central1 \
    --zone=us-central1-a \
    --single-node \
    --master-machine-type=n1-standard-4 \
    --master-boot-disk-size=50GB \
    --image-version=1.5-debian10
Waiting on operation [projects/cs570-big-data-analytics/regions/us-central1/operations/0b0
e8137-3f64-3271-b601-11f71a914619].
Waiting for cluster creation operation...
WARNING: Consider using Auto Zone rather than selecting a zone manually. See https://cloud
.google.com/dataproc/docs/concepts/configuring-clusters/auto-zone
WARNING: Failed to validate permissions required for default service account: '48943335059
7-compute@developer.gserviceaccount.com'. Cluster creation could still be successful if re
quired permissions have been granted to the respective service accounts as mentioned in th
e document https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/service-ac
counts#dataproc_service_accounts_2. This could be due to Cloud Resource Manager API hasn't
 been enabled in your project '489433350597' before or it is disabled. Enable it by visiti
ng 'https://console.developers.google.com/apis/api/cloudresourcemanager.googleapis.com/ove
rview?project=489433350597'.
WARNING: For PD-Standard without local SSDs, we strongly recommend provisioning 1TB or lar
ger to ensure consistently high I/O performance. See https://cloud.google.com/compute/docs
/disks/performance for information on disk I/O performance.
WARNING: The firewall rules for specified network or subnetwork would allow ingress traffi
c from 0.0.0.0/0, which could be a security risk.
WARNING: Unable to validate the staging bucket lifecycle configuration of the bucket 'data
proc-staging-us-central1-489433350597-rvbwz4vc' due to an internal error, Please make sure
 that the provided bucket doesn't have any delete rules set.
Waiting for cluster creation operation...working...
```

```
Waiting for cluster creation operation...done.
Created [https://dataproc.googleapis.com/v1/projects/cs570-big-data-analytics/regions/us-c
entral1/clusters/pagerank-cluster] Cluster placed in zone [us-central1-a].
nhaile96456@cloudshell:~ (cs570-big-data-analytics)$ gcloud dataproc clusters list --regio
n=us-central1
NAME: pagerank-cluster
PLATFORM: GCE
PRIMARY_WORKER_COUNT:
SECONDARY_WORKER_COUNT:
STATUS: RUNNING
ZONE: us-central1-a
SCHEDULED_DELETE:
nhaile96456@cloudshell:~ (cs570-big-data-analytics)$
```

3. Upload the following files to your bucket:
   a. input.txt



   b. pagerank.py

```python
import re
import sys
from operator import add
from pyspark.sql import SparkSession

def computeContribs(urls, rank):
    num_urls = len(urls)
    for url in urls:
        yield (url, rank / num_urls)

def parseNeighbors(urls):
    parts = re.split(r'\s+', urls)
    return parts[0], parts[1]

if __name__ == "__main__":
    if len(sys.argv) != 3:
        print("Usage: pagerank <file> <iterations>", file=sys.stderr)
        sys.exit(-1)

    spark = SparkSession.builder.appName("PythonPageRank").getOrCreate()

    lines = spark.read.text(sys.argv[1]).rdd.map(lambda r: r[0])
    links = lines.map(lambda urls: parseNeighbors(urls)).distinct().groupByKey().cache()
    ranks = links.map(lambda url_neighbors: (url_neighbors[0], 1.0))

    for iteration in range(int(sys.argv[2])):
        contribs = links.join(ranks).flatMap(
            lambda url_urls_rank: computeContribs(url_urls_rank[1][0], url_urls_rank[1][1]
))
        ranks = contribs.reduceByKey(add).mapValues(lambda rank: rank * 0.85 + 0.15)

    for (link, rank) in ranks.collect():
        print("%s has rank: %s." % (link, rank))

    spark.stop()
```

This script implements the PageRank algorithm using PySpark. Here's an explanation of the key components and functionality:

a) Imports and Function Definitions:
   - import re: Regular expression module.
   - import sys: System-specific parameters and functions.
   - from operator import add: Importing the addition operator for reduceByKey.
   - from pyspark.sql import SparkSession: Importing SparkSession to create a Spark session.

b) Function computeContribs(urls, rank):
   - Computes contributions of each URL to the rank of other URLs.
   - urls: List of URLs.
   - rank: The rank of the current URL.
   - num_urls: Number of URLs linked from the current URL.
   - Yields tuples of (url, rank / num_urls) for each URL in the list.

c) Function parseNeighbors(urls):
   - Parses input URLs to extract links.
   - Splits the input string on whitespace to separate the URL and its neighbor.
   - Returns a tuple (parts[0], parts[1]).

d) Main Execution Block:

- Checks command-line arguments for the input file and number of iterations.
- Initializes a Spark session named "PythonPageRank".
- Reads the input file and creates an RDD of lines.
- Parses lines to create links and caches the result.
- Initializes ranks with a rank of 1.0 for each URL.

e) PageRank Iterations:

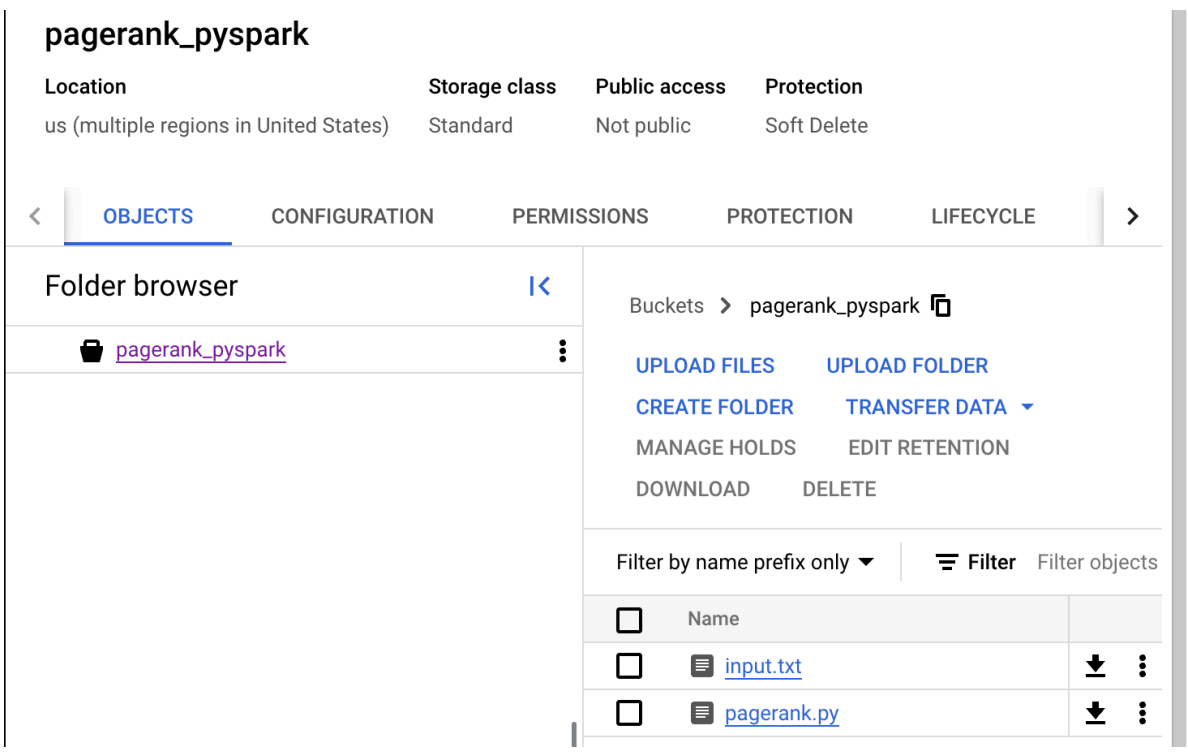
- Iterates for the specified number of iterations.
- Computes contributions of URLs to their neighbors.
- Updates ranks based on the contributions, applying the PageRank formula.

f) Output and Cleanup:

- Collects and prints the final ranks of each URL.
- Stops the Spark session.

Uploading the files:

```
nhaile96456@cloudshell:~ (cs570-big-data-analytics)$ gsutil cp pagerank.py gs://pagerank_p
yspark/
Copying file://pagerank.py [Content-Type=text/x-python]...
/ [1 files][  1.1 KiB/  1.1 KiB]
Operation completed over 1 objects/1.1 KiB.
nhaile96456@cloudshell:~ (cs570-big-data-analytics)$
```



4. Submit the PySpark job.

*gcloud dataproc jobs submit pyspark gs://pagerank_pyspark/pagerank.py *

*--cluster=pagerank-cluster *

*--region=us-central1 *

-- gs://pagerank_pyspark/input.txt 10

```
nhaile96456@cloudshell:~ (cs570-big-data-analytics)$ gcloud dataproc jobs submit pyspark g
s://pagerank_pyspark/pagerank.py \
    --cluster=pagerank-cluster \
    --region=us-central1 \
    -- gs://pagerank_pyspark/input.txt 10
Job [3d24e898dedf48f1a5d2bcf98b8e51c5] submitted.
Waiting for job output...
24/07/01 00:12:06 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
24/07/01 00:12:06 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
24/07/01 00:12:06 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
24/07/01 00:12:06 INFO org.spark_project.jetty.util.log: Logging initialized @4338ms to or
g.spark_project.jetty.util.log.Slf4jLog
24/07/01 00:12:06 INFO org.spark_project.jetty.server.Server: jetty-9.4.z-SNAPSHOT; built:
 unknown; git: unknown; jvm 1.8.0_382-b05
24/07/01 00:12:06 INFO org.spark_project.jetty.server.Server: Started @4450ms
24/07/01 00:12:06 INFO org.spark_project.jetty.server.AbstractConnector: Started ServerCon
nector@711be86e{HTTP/1.1, (http/1.1)}{0.0.0.0:38817}
24/07/01 00:12:08 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManage
r at pagerank-cluster-m/10.128.0.6:8032
24/07/01 00:12:08 INFO org.apache.hadoop.yarn.client.AHSProxy: Connecting to Application H
istory server at pagerank-cluster-m/10.128.0.6:10200
24/07/01 00:12:08 INFO org.apache.hadoop.conf.Configuration: resource-types.xml not found
24/07/01 00:12:08 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Unable to find
'resource-types.xml'.
24/07/01 00:12:08 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Adding resource
 type - name = memory-mb, units = Mi, type = COUNTABLE
24/07/01 00:12:08 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Adding resource
 type - name = vcores, units = , type = COUNTABLE
24/07/01 00:12:11 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted ap
plication application_1719791996059_0001
A has rank: 1.1667391764027368.
B has rank: 0.6432494117885129.
C has rank: 1.1900114118087488.
24/07/01 00:12:40 INFO org.spark_project.jetty.server.AbstractConnector: Stopped Spark@711
be86e{HTTP/1.1, (http/1.1)}{0.0.0.0:0}
Job [3d24e898dedf48f1a5d2bcf98b8e51c5] finished successfully.
done: true
driverControlFilesUri: gs://dataproc-staging-us-central1-489433350597-rvbwz4vc/google-clou
d-dataproc-metainfo/b50d301f-7ab8-4451-9b66-3fbb00ebd4d6/jobs/3d24e898dedf48f1a5d2bcf98b8e
51c5/
driverOutputResourceUri: gs://dataproc-staging-us-central1-489433350597-rvbwz4vc/google-cl
oud-dataproc-metainfo/b50d301f-7ab8-4451-9b66-3fbb00ebd4d6/jobs/3d24e898dedf48f1a5d2bcf98b
8e51c5/driveroutput
jobUuid: 58b963bd-4481-3a63-9e6f-1a2b1345f58c
placement:
```

```
placement:
  clusterName: pagerank-cluster
  clusterUuid: b50d301f-7ab8-4451-9b66-3fbb00ebd4d6
pysparkJob:
  args:
  - gs://pagerank_pyspark/input.txt
  - '10'
  mainPythonFileUri: gs://pagerank_pyspark/pagerank.py
reference:
  jobId: 3d24e898dedf48f1a5d2bcf98b8e51c5
  projectId: cs570-big-data-analytics
status:
  state: DONE
  stateStartTime: '2024-07-01T00:12:45.533264Z'
statusHistory:
- state: PENDING
  stateStartTime: '2024-07-01T00:11:59.975966Z'
- state: SETUP_DONE
  stateStartTime: '2024-07-01T00:12:00.016518Z'
- details: Agent reported job success
  state: RUNNING
  stateStartTime: '2024-07-01T00:12:00.522710Z'
yarnApplications:
- name: PythonPageRank
  progress: 1.0
  state: FINISHED
  trackingUrl: http://pagerank-cluster-m:8088/proxy/application_1719791996059_0001/
nhaile96456@cloudshell:~ (cs570-big-data-analytics)$
```

5. Let's check and confirm the output files.

```
nhaile96456@cloudshell:~ (cs570-big-data-analytics)$ gsutil ls gs://dataproc-staging-us-ce
ntral1-489433350597-rvbwz4vc/google-cloud-dataproc-metainfo/b50d301f-7ab8-4451-9b66-3fbb00
ebd4d6/jobs/3d24e898dedf48f1a5d2bcf98b8e51c5/
gs://dataproc-staging-us-central1-489433350597-rvbwz4vc/google-cloud-dataproc-metainfo/b50
d301f-7ab8-4451-9b66-3fbb00ebd4d6/jobs/3d24e898dedf48f1a5d2bcf98b8e51c5/driveroutput.00000
0000
gs://dataproc-staging-us-central1-489433350597-rvbwz4vc/google-cloud-dataproc-metainfo/b50
d301f-7ab8-4451-9b66-3fbb00ebd4d6/jobs/3d24e898dedf48f1a5d2bcf98b8e51c5/driveroutput.00000
0001
```

```
24/07/01 00:12:11 INFO org.apache.hadoop.yarn.client
plication application_1719791996059_0001
A has rank: 1.1667391764027368.
B has rank: 0.6432494117885129.
C has rank: 1.1900114118087488.
```

6. Experiment with different number of iterations.
   - 1 iteration

*gcloud dataproc jobs submit pyspark gs://pagerank_pyspark/pagerank.py *

*--cluster=pagerank-cluster *

*--region=us-central1 *

-- gs://pagerank_pyspark/input.txt 1

```
24/07/01 00:33:25 INFO org.apache.hadoop.yarn.client
plication application_1719791996059_0002
C has rank: 1.4249999999999998.
A has rank: 1.0.
B has rank: 0.575.
```

   - 50 iterations

*gcloud dataproc jobs submit pyspark gs://pagerank_pyspark/pagerank.py *

*--cluster=pagerank-cluster *

*--region=us-central1 *

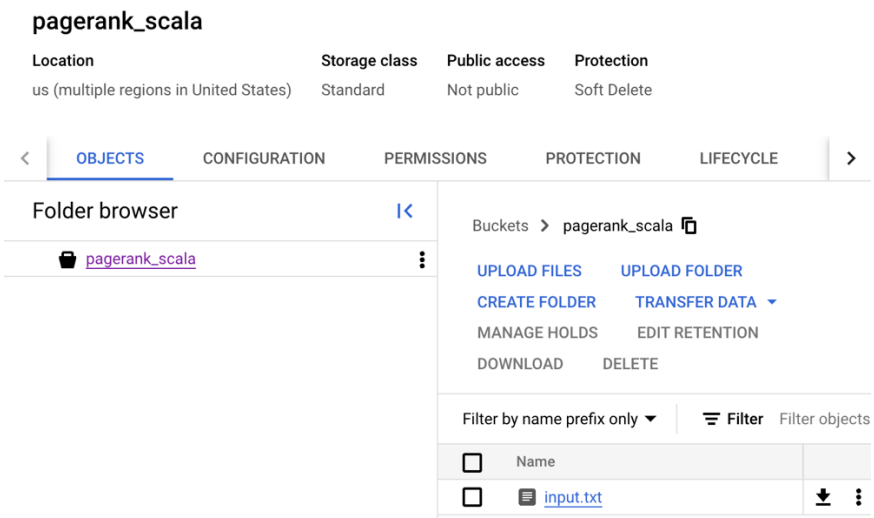
-- gs://pagerank_pyspark/input.txt 50

```
24/07/01 00:36:17 INFO org.apache.hadoop.yarn.client
plication application_1719791996059_0003
B has rank: 0.6444318824177515.
C has rank: 1.1921989824728403.
A has rank: 1.1633691351094062.
```

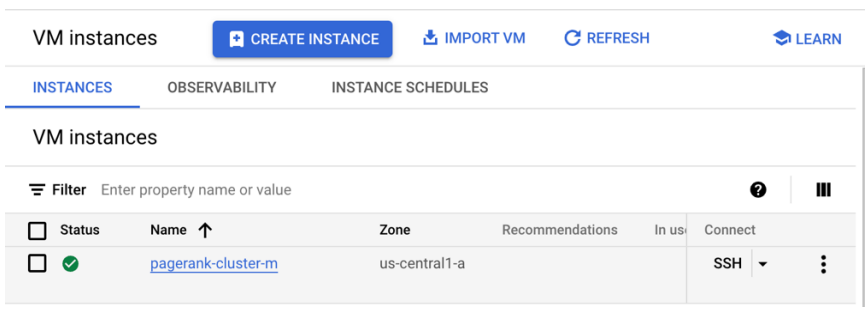
PART TWO – RUN USING SCALA

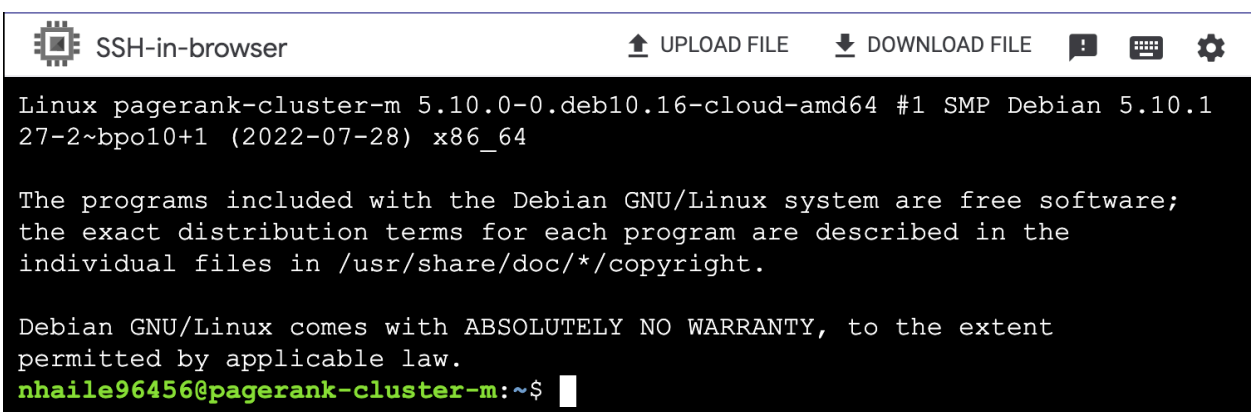
7. Create a new bucket and upload the input.txt file to it.

```
nhaile96456@cloudshell:~ (cs570-big-data-analytics)$ gsutil mb gs://pagerank_scala/
Creating gs://pagerank_scala/...
nhaile96456@cloudshell:~ (cs570-big-data-analytics)$
```



8. SSH into the cluster.





9. Install scala.

- *sudo apt-get update*
- *sudo apt-get install scala*

```
nhaile96456@pagerank-cluster-m:~$ sudo apt-get install scala
Reading package lists... Done
Building dependency tree
Reading state information... Done
scala is already the newest version (2.12.10-400).
0 upgraded, 0 newly installed, 0 to remove and 120 not upgraded.
nhaile96456@pagerank-cluster-m:~$
```

10. Add the Scala SBT (Simple Build Tool) repository to your system's package sources list.

- *echo "deb https://repo.scala-sbt.org/scalasbt/debian all main" | sudo tee /etc/apt/sources.list.d/sbt.list*

To ensure you can install sbt successfully, you should also import the public key used by the package management system:

- *curl -sL "https://keyserver.ubuntu.com/pks/lookup?op=get&search=0x99E82A75642AC823" | sudo apt-key add*

```
nhaile96456@pagerank-cluster-m:~$ echo "deb https://repo.scala-sbt.org/scal
asbt/debian all main" | sudo tee /etc/apt/sources.list.d/sbt.list
deb https://repo.scala-sbt.org/scalasbt/debian all main
nhaile96456@pagerank-cluster-m:~$ curl -sL "https://keyserver.ubuntu.com/pk
s/lookup?op=get&search=0x99E82A75642AC823" | sudo apt-key add -
OK
nhaile96456@pagerank-cluster-m:~$
```

Then, update the package list and install sbt:

- *sudo apt-get update*
- *sudo apt-get install sbt*

```
nhaile96456@pagerank-cluster-m:~$ sudo apt-get install sbt
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following NEW packages will be installed:
  sbt
0 upgraded, 1 newly installed, 0 to remove and 120 not upgraded.
Need to get 20.0 kB of archives.
After this operation, 50.2 kB of additional disk space will be used.
Get:1 https://scala.jfrog.io/artifactory/debian all/main amd64 sbt all 1.10
.0 [20.0 kB]
Fetched 20.0 kB in 1s (30.5 kB/s)
Selecting previously unselected package sbt.
(Reading database ... 167133 files and directories currently installed.)
Preparing to unpack .../archives/sbt_1.10.0_all.deb ...
Unpacking sbt (1.10.0) ...
Setting up sbt (1.10.0) ...
Creating system group: sbt
Creating system user: sbt in sbt with sbt daemon-user and shell /bin/false
Processing triggers for man-db (2.8.5-2) ...
nhaile96456@pagerank-cluster-m:~$
```

11. To setup the project structure and compile the code, create the project directories.

```
nhaile96456@pagerank-cluster-m:~$ mkdir pagerank
nhaile96456@pagerank-cluster-m:~$ cd pagerank/
nhaile96456@pagerank-cluster-m:~/pagerank$ mkdir -p src/main/scala
nhaile96456@pagerank-cluster-m:~/pagerank$
```

12. Create build.sbt.

```
nhaile96456@pagerank-cluster-m:~/pagerank$ vi build.sbt
nhaile96456@pagerank-cluster-m:~/pagerank$ cat build.sbt
name := "SparkPageRank"

version := "1.0"

scalaVersion := "2.12.10"

libraryDependencies ++= Seq(
  "org.apache.spark" %% "spark-core" % "3.1.2",
  "org.apache.spark" %% "spark-sql" % "3.1.2"
)
```

13. Create ScalaPageRank.scala.

```
nhaile96456@pagerank-cluster-m:~/pagerank$ vi src/main/scala/SparkPageRank.
scala
nhaile96456@pagerank-cluster-m:~/pagerank$ cat src/main/scala/SparkPageRank
.scala
package org.apache.spark.examples

import org.apache.spark.SparkContext._
import org.apache.spark.{SparkConf, SparkContext}

object SparkPageRank {

  def showWarning() {
    System.err.println(
      """WARN: This is a naive implementation of PageRank and is given as a
n example!
        |Please use the PageRank implementation found in org.apache.spark.g
raphx.lib.PageRank
        |for more conventional use.
      """.stripMargin)
  }

  def main(args: Array[String]) {
    if (args.length < 1) {
      System.err.println("Usage: SparkPageRank <file> <iter>")
      System.exit(1)
    }

    showWarning()

    val sparkConf = new SparkConf().setAppName("PageRank")
    val iters = if (args.length > 1) args(1).toInt else 10
    val ctx = new SparkContext(sparkConf)
    val lines = ctx.textFile(args(0), 1)

    val links = lines.map { s =>
      val parts = s.split("\\s+")
      (parts(0), parts(1))
    }.distinct().groupByKey().cache()
```

```
    var ranks = links.mapValues(v => 1.0)

    for (i <- 1 to iters) {
      val contribs = links.join(ranks).values.flatMap { case (urls, rank) =
>
        val size = urls.size
        urls.map(url => (url, rank / size))
      }
      ranks = contribs.reduceByKey(_ + _).mapValues(0.15 + 0.85 * _)
    }

    val output = ranks.collect()
    output.foreach(tup => println(tup._1 + " has rank: " + tup._2 + "."))

    ctx.stop()
  }
}
```

14. Compile the project.

```
nhaile96456@pagerank-cluster-m:~/pagerank$ sbt package
downloading sbt launcher 1.10.0
[info] [launcher] getting org.scala-sbt sbt 1.10.0  (this may take some tim
e)...
[info] [launcher] getting Scala 2.12.19 (for sbt)...
[info] Updated file /home/nhaile96456/pagerank/project/build.properties: se
t sbt.version to 1.10.0
[info] welcome to sbt 1.10.0 (Temurin Java 1.8.0_382)
[info] loading project definition from /home/nhaile96456/pagerank/project
[info] Updating pagerank-build
https://repo1.maven.org/maven2/jline/jline/2.14.6/jline-2.14.6.pom
  100.0% [##########] 19.4 KiB (188.6 KiB / s)
[info] Resolved pagerank-build dependencies
[info] Fetching artifacts of pagerank-build
[info] Fetched artifacts of pagerank-build
[info] loading settings for project pagerank from build.sbt ...
[info] set current project to SparkPageRank (in build file:/home/nhaile9645
6/pagerank/)
```

```
https://repo1.maven.org/maven2/org/apache/hadoop/hadoop-yarn-client/3.2.0/…
  100.0% [##########] 310.5 KiB (6.1 MiB / s)
https://repo1.maven.org/maven2/org/apache/httpcomponents/httpclient/4.5.2/…
  100.0% [##########] 719.4 KiB (18.0 MiB / s)
https://repo1.maven.org/maven2/org/apache/spark/spark-tags_2.12/3.1.2/spar…
  100.0% [##########] 14.8 KiB (672.7 KiB / s)
https://repo1.maven.org/maven2/org/apache/yetus/audience-annotations/0.5.0…
  100.0% [##########] 20.0 KiB (486.8 KiB / s)
https://repo1.maven.org/maven2/org/scala-lang/modules/scala-xml_2.12/1.2.0…
  100.0% [##########] 543.5 KiB (12.1 MiB / s)
https://repo1.maven.org/maven2/org/apache/commons/commons-crypto/1.1.0/com…
  100.0% [##########] 162.3 KiB (4.4 MiB / s)
https://repo1.maven.org/maven2/org/json4s/json4s-scalap_2.12/3.7.0-M5/json…
  100.0% [##########] 340.9 KiB (9.8 MiB / s)
[info] Fetched artifacts of sparkpagerank_2.12
[info] compiling 1 Scala source to /home/nhaile96456/pagerank/target/scala-
2.12/classes ...
[info] Non-compiled module 'compiler-bridge_2.12' for Scala 2.12.10. Compil
ing...
[info]   Compilation completed in 13.557s.
[success] Total time: 25 s, completed Jul 1, 2024 2:12:19 AM
nhaile96456@pagerank-cluster-m:~/pagerank$
```

15. To upload a compile JAR file to Google Cloud Storage, copy the compiled JAR file to a Google Cloud Storage bucket.

```
nhaile96456@pagerank-cluster-m:~/pagerank$ gsutil cp target/scala-2.12/spar
kpagerank_2.12-1.0.jar gs://pagerank_scala/
WARNING:  Python 3.5-3.7 will be deprecated on August 8th, 2023. Please use
 Python version 3.8 and up.

If you have a compatible Python interpreter installed, you can use it by se
tting
the CLOUDSDK_PYTHON environment variable to point to it.

Copying file://target/scala-2.12/sparkpagerank_2.12-1.0.jar [Content-Type=a
pplication/java-archive]...
/ [0 files][    0.0 B/  5.4 KiB]
/ [1 files][  5.4 KiB/  5.4 KiB]

Operation completed over 1 objects/5.4 KiB.

nhaile96456@pagerank-cluster-m:~/pagerank$
```

16. Submit spark job on Dataproc.

- *gcloud dataproc jobs submit spark --cluster=pagerank-cluster --region=us-central1 --jars=gs://pagerank_scala/sparkpagerank_2.12-1.0.jar --class=org.apache.spark.examples.SparkPageRank -- gs://pagerank_scala/input.txt 10*

```
nhaile96456@cloudshell:~ (cs570-big-data-analytics)$ gcloud dataproc jobs submit spark --clust
er=pagerank-cluster --region=us-central1 --jars=gs://pagerank_scala/sparkpagerank_2.12-1.0.jar
 --class=org.apache.spark.examples.SparkPageRank -- gs://pagerank_scala/input.txt 10
Job [954ffdb0b8fc48289f01353eaa6e3f90] submitted.
Waiting for job output...
WARN: This is a naive implementation of PageRank and is given as an example!
Please use the PageRank implementation found in org.apache.spark.graphx.lib.PageRank
for more conventional use.

24/07/01 02:18:26 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
24/07/01 02:18:26 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
24/07/01 02:18:26 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
24/07/01 02:18:27 INFO org.spark_project.jetty.util.log: Logging initialized @3259ms to org.sp
ark_project.jetty.util.log.Slf4jLog
24/07/01 02:18:27 INFO org.spark_project.jetty.server.Server: jetty-9.4.z-SNAPSHOT; built: unk
nown; git: unknown; jvm 1.8.0_382-b05
24/07/01 02:18:27 INFO org.spark_project.jetty.server.Server: Started @3496ms
24/07/01 02:18:27 INFO org.spark_project.jetty.server.AbstractConnector: Started ServerConnect
or@39109136{HTTP/1.1, (http/1.1)}{0.0.0.0:42611}
24/07/01 02:18:28 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at
 pagerank-cluster-m/10.128.0.6:8032
24/07/01 02:18:28 INFO org.apache.hadoop.yarn.client.AHSProxy: Connecting to Application Histo
ry server at pagerank-cluster-m/10.128.0.6:10200
24/07/01 02:18:28 INFO org.apache.hadoop.conf.Configuration: resource-types.xml not found
24/07/01 02:18:28 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Unable to find 'res
ource-types.xml'.
24/07/01 02:18:28 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Adding resource typ
e - name = memory-mb, units = Mi, type = COUNTABLE
24/07/01 02:18:28 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Adding resource typ
e - name = vcores, units = , type = COUNTABLE
24/07/01 02:18:30 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted applic
ation application_1719791996059_0004
24/07/01 02:18:38 INFO org.apache.hadoop.mapred.FileInputFormat: Total input files to process
: 1
B has rank: 0.6432494117885129.
A has rank: 1.1667391764027368.
C has rank: 1.1900114118087488.
24/07/01 02:18:45 INFO org.spark_project.jetty.server.AbstractConnector: Stopped Spark@3910913
6{HTTP/1.1, (http/1.1)}{0.0.0.0:0}
Job [954ffdb0b8fc48289f01353eaa6e3f90] finished successfully.
done: true
```

```
driverControlFilesUri: gs://dataproc-staging-us-central1-489433350597-rvbwz4vc/google-cloud-da
taproc-metainfo/b50d301f-7ab8-4451-9b66-3fbb00ebd4d6/jobs/954ffdb0b8fc48289f01353eaa6e3f90/
driverOutputResourceUri: gs://dataproc-staging-us-central1-489433350597-rvbwz4vc/google-cloud-
dataproc-metainfo/b50d301f-7ab8-4451-9b66-3fbb00ebd4d6/jobs/954ffdb0b8fc48289f01353eaa6e3f90/d
riveroutput
jobUuid: 8a528566-0da1-300b-8e3d-91b886020cda
placement:
  clusterName: pagerank-cluster
  clusterUuid: b50d301f-7ab8-4451-9b66-3fbb00ebd4d6
reference:
  jobId: 954ffdb0b8fc48289f01353eaa6e3f90
  projectId: cs570-big-data-analytics
sparkJob:
  args:
  - gs://pagerank_scala/input.txt
  - '10'
  jarFileUris:
  - gs://pagerank_scala/sparkpagerank_2.12-1.0.jar
  mainClass: org.apache.spark.examples.SparkPageRank
status:
  state: DONE
  stateStartTime: '2024-07-01T02:18:48.909406Z'
statusHistory:
- state: PENDING
  stateStartTime: '2024-07-01T02:18:23.093755Z'
- state: SETUP_DONE
  stateStartTime: '2024-07-01T02:18:23.133932Z'
- details: Agent reported job success
  state: RUNNING
  stateStartTime: '2024-07-01T02:18:23.403562Z'
yarnApplications:
- name: PageRank
  progress: 1.0
  state: FINISHED
  trackingUrl: http://pagerank-cluster-m:8088/proxy/application_1719791996059_0004/
nhaile96456@cloudshell:~ (cs570-big-data-analytics)$
```

```
24/07/01 02:18:38 INFO org.apache.hadoop
: 1
B has rank: 0.6432494117885129.
A has rank: 1.1667391764027368.
C has rank: 1.1900114118087488.
```