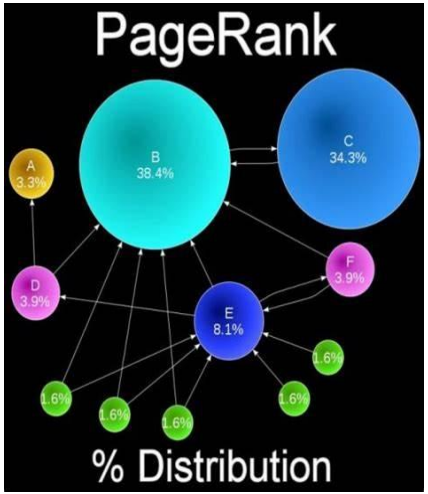
PageRank on GCP



CS570 - Big Data Processing & Analytics

Submitted by: Divya Pandey(19665)
Instructor: Dr. Henry Chang

Table of Content

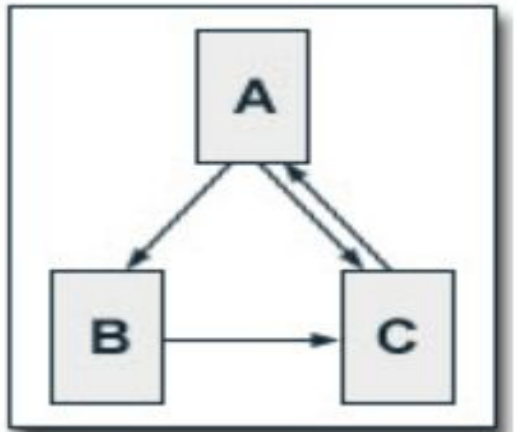
Introduction

- PageRank is essentially an automated formula that determines the search engine rankings of a web page based on the links that connect to it.

- The higher a web page scores on this zero to ten scale, the better the ranked.

Design

- Page A has a link to pages B and C
- Page B has a link to page C
- Page C has links to page A



Design

A's PageRank is

PR(A) =(1-d)+d*(PR(C)/1)

B's PageRank is

PR(B) = (1-d)+d*(PR(A)/2)

C's PageRank is

PR(C) =(1-d)+d*(PR(B)/1+PR(A)/2)

Damping factor is 0.85

Design

1st iteration:

A = 1

B = (1/2) = 0.5

C = 1 + (1/2) = 1.5

PageRank (A) = 1 – 0.85 + 0.85 * 1 = 1

PageRank (B) = 1 – 0.85 + 0.85 * 1 = 0.575

PageRank (C) = 1 – 0.85 + 0.85 * 1.5 = 1.425

Design

2nd iteration:

A = 1

B = (1/2) = 0.5

C = 0.575 + (1/2) = 1.075

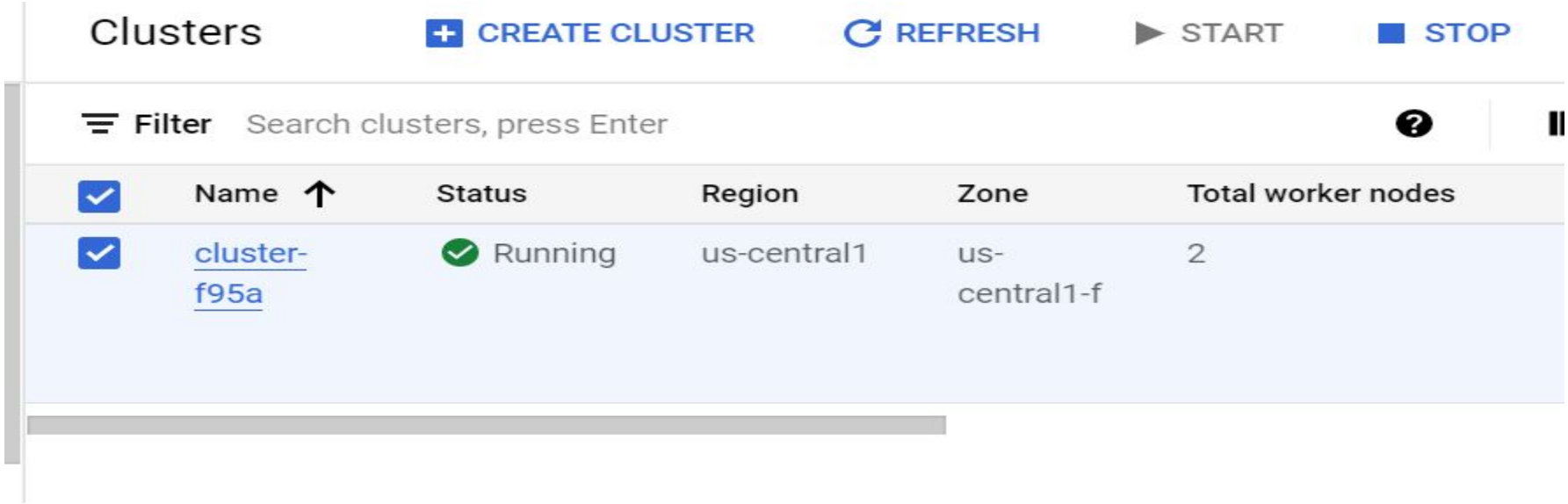
PageRank (A) = 1 – 0.85 + 0.85 * 1.425 = 1.36125

PageRank (B) = 1 – 0.85 + 0.85 * 0.5 = 0.575

PageRank (C) = 1 – 0.85 + 0.85 * 1.075 = 1.06375

Implementation

- **Create Dataproc cluster on GCP**



Implementation(PySpark)

- **Create pagerank.py python program (Pyspark)**
- **Generate pagerank_data.txt input file**
- **Create hdfs directory:**

```
hdfs dfs -mkdir hdfs:///mydata
hdfs dfs -put pagerank_data.txt hdfs:///mydata

hdfs dfs -put pagerank.py hdfs:///mydata
```

- **Submit the Spark Job (10th Iteration)**

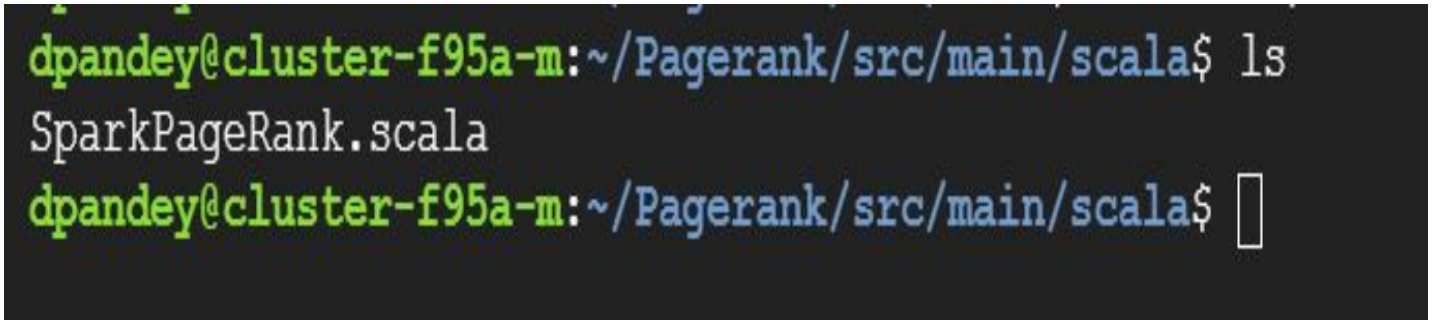
```
spark-submit hdfs:///mydata/pagerank.py
hdfs:///mydata/pagerank_data.txt 10
```

Implementation(PySpark)

Generate the pagerank_data.txt and Save as input file for Hadoop Job

Implementation(Scala)

- Create SparkPageRank.scala Scala program
- Generate pagerank_data.txt input file
- Install sbt
- Create Pagerank directory and add SparkPageRank.scala program into that directory:



Implementation(Scala)

- **Create build.sbt file into Pagerank directory**

  name := "Simple Project"
  version := "1.0"
  scalaVersion := "2.12.14"
  libraryDependencies += "org.apache.spark" %% "spark-core" % "3.1.3"

  Save and exit

- **Create jar file**

  sbt package

```
dpandey@cluster-f95a-m:~/Pagerank$ sbt package
[info] welcome to sbt 1.7.3 (Temurin Java 1.8.0_345)
[info] loading project definition from /home/dpandey/Pagerank/project
[info] loading settings for project pagerank from build.sbt ...
[info] set current project to Simple Project (in build file:/home/dpandey/Pagerank/)
[success] Total time: 2 s, completed Nov 2, 2022 7:43:05 PM
dpandey@cluster-f95a-m:~/Pagerank$
```

Implementation(Scala)

- **Create your bucket**
- **Upload pagerank_data.txt file into your bucket**

  gsutil cp pagerank_data.txt gs://pagerankscala

- **Upload jar file into your bucket**

gsutil cp/home/dpandey/Pagerank/target/scala-2.12/simple-project_2.12-1.0.jar gs://pagerankscala

Test(PySpark)

spark-submit hdfs:///mydata/pagerank.py

hdfs:///mydata/pagerank_data.txt 10 (10th Iteration)

```
22/11/02 19:18:50 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Unable to find 'resource-types.xml'.
22/11/02 19:18:51 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted application application_1667415352394_0007
22/11/02 19:18:52 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at cluster-f95a-m/10.128.0.52:8030
22/11/02 19:18:54 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageImpl: Ignoring exception of type GoogleJsonRespons
eException; verified object already exists with desired state.
A has rank: 1.1667391764027368.
B has rank: 0.6432494117885129.
C has rank: 1.1900114118087488.
22/11/02 19:19:09 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@3bfbalb{HTTP/1.1, (http/1.1)}{0.0.0.0:0}
dpandey@cluster-f95a-m:~$
```

Test(Scala)

gcloud dataproc jobs submit spark --cluster=cluster-f95a
--region=us-central1
--jars=gs://scalapagerank/simple-project_2.12-1.0.jar
--class=org.apache.spark.examples.SparkPageRank --
gs://scalapagerank/pagerank_data.txt 10 (10th Iteration)

```
22/11/02 19:08:40 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at cluster-f95a-m/10.128.0.52:8030
22/11/02 19:08:42 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageImpl: Ignoring exception of type GoogleJsonRespo
eException; verified object already exists with desired state.
22/11/02 19:08:43 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.
22/11/02 19:08:48 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@149f5761{HTTP/1.1, (http/1.1)}{0.0.0.0:0}
Job [6f0eed3fc1a84db1b9eaa58c7939840a] finished successfully.
done: true
driverControlFilesUri: gs://dataproc-staging-us-central1-109026694714-p79rkwvt/google-cloud-dataproc-metainfo/e1140323-5d07-4c2a-a210-0f3107cfa30f/jobs/6f0eed
c1a84db1b9eaa58c7939840a/
driverOutputResourceUri: gs://dataproc-staging-us-central1-109026694714-p79rkwvt/google-cloud-dataproc-metainfo/e1140323-5d07-4c2a-a210-0f3107cfa30f/jobs/6f0e
3fc1a84db1b9eaa58c7939840a/driveroutput
jobUuid: 00dc380f-7a23-31f4-9d8d-90bc99da8d30
```

Enhancement Ideas

PageRank Operation can also performed on the following cloud computing environment:

- AWS
- Azure

Conclusion

- The more input web links better the pagerank of web pages.

References

- *Seo glossary definition page*. Ahrefs. (n.d.). Retrieved November 2, 2022, from https://ahrefs.com/seo/glossary/pagerank
- O'Reilly Media, Inc. (n.d.). *Learning spark*. O'Reilly Online Learning. Retrieved November 2, 2022, from https://www.oreilly.com/library/view/learning-spark/9781449359034/ch04.html