PES University
UE18CS322:Big Data
Page Rank Algorithm implementation with Map Reduce

SRN: PES2201800681

NAME: AJITESH NAIR

SECTION: B

PageRank (PR) is an algorithm used by Google Search to rank websites in their search engine results. PageRank was named after Larry Page, one of the founders of Google. PageRank is a way of measuring the importance of website pages. According to Google:

PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.

It is not the only algorithm used by Google to order search engine results, but it is the first algorithm that was used by the company, and it is the best-known.
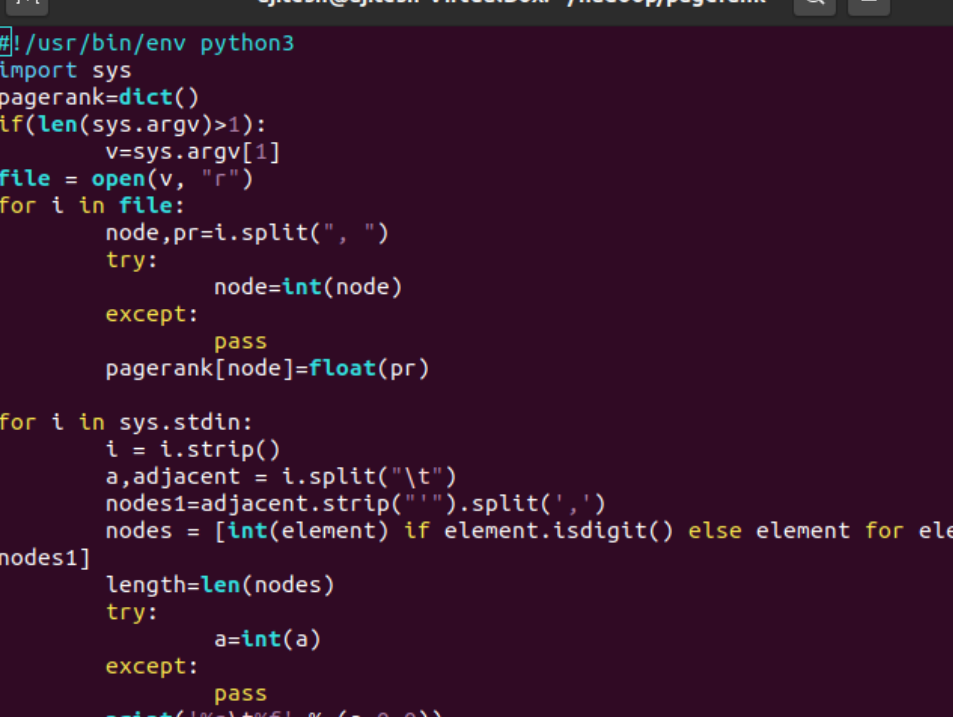
Algorithm

The PageRank algorithm outputs a probability distribution used to represent the likelihood that a person randomly clicking on links will arrive at any particular page. PageRank can be calculated for collections of documents of any size. It is assumed in several research papers that the distribution is evenly divided among all documents in the collection at the beginning of the computational process. The PageRank computations require several passes, called "iterations", through the collection to adjust approximate PageRank values to more closely reflect the theoretical true value.

All codes have been attached in the zip file.

Mapper Code:

```python
#!/usr/bin/env python3
import sys
for i in sys.stdin:
        if i.startswith('#'):
                continue
        i = i.strip()
        try:
                a,b = i.split('\t')
        except:
                continue
        print ('%s\t%s' % (a,b))
```

```
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
```

"mapper_t1.py" 11L, 177C   8,2-16   All
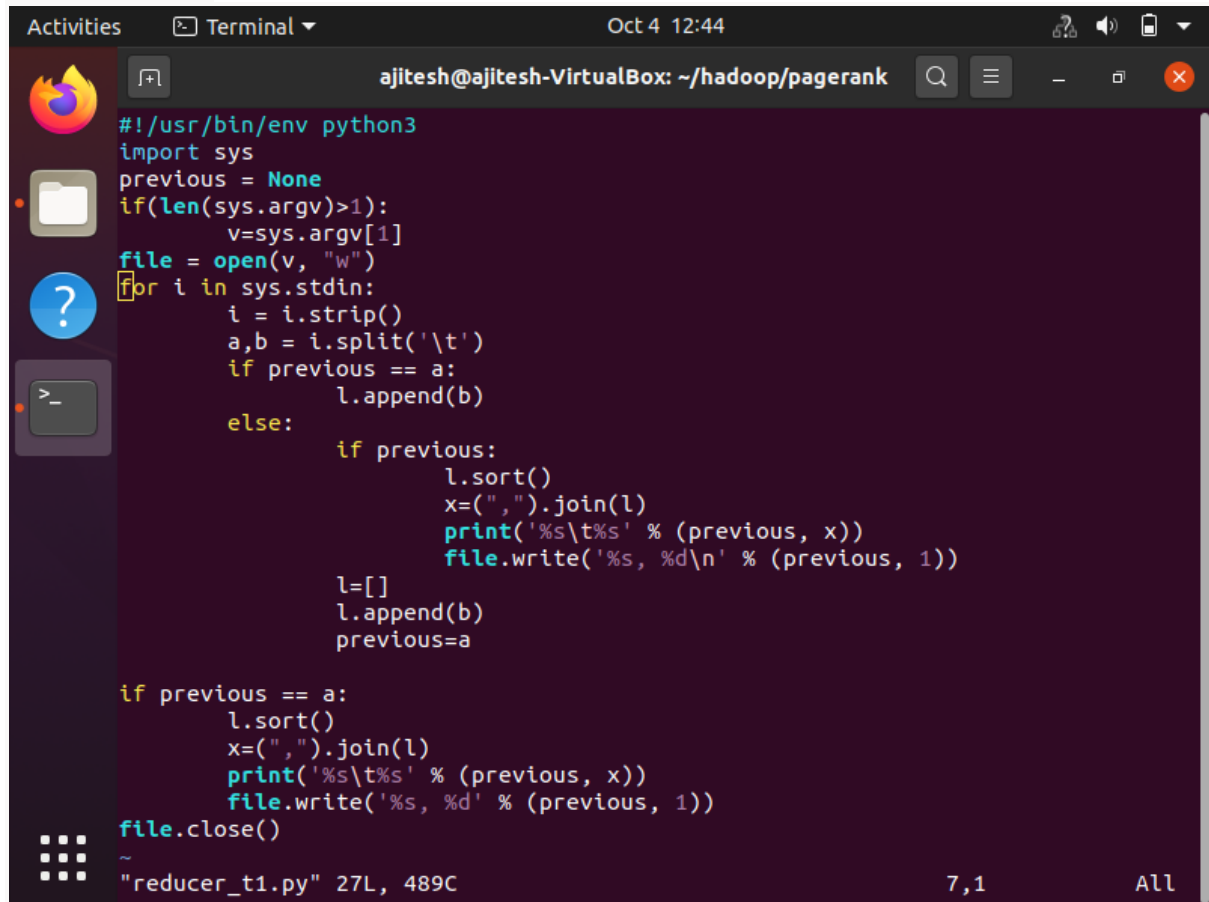
```python
#!/usr/bin/env python3
import sys
pagerank=dict()
if(len(sys.argv)>1):
        v=sys.argv[1]
file = open(v, "r")
for i in file:
        node,pr=i.split(", ")
        try:
                node=int(node)
        except:
                pass
        pagerank[node]=float(pr)

for i in sys.stdin:
        i = i.strip()
        a,adjacent = i.split("\t")
        nodes1=adjacent.strip("'").split(',')
        nodes = [int(element) if element.isdigit() else element for element in
nodes1]
        length=len(nodes)
        try:
                a=int(a)
        except:
                pass
        print('%s\t%f' % (a,0.0))
        for word in nodes:
                try:
```

1,1   Top

Reducer Code:

```python
#!/usr/bin/env python3
import sys
previous = None
if(len(sys.argv)>1):
        v=sys.argv[1]
file = open(v, "w")
for i in sys.stdin:
        i = i.strip()
        a,b = i.split('\t')
        if previous == a:
                l.append(b)
        else:
                if previous:
                        l.sort()
                        x=(",").join(l)
                        print('%s\t%s' % (previous, x))
                        file.write('%s, %d\n' % (previous, 1))
                l=[]
                l.append(b)
                previous=a

if previous == a:
        l.sort()
        x=(",").join(l)
        print('%s\t%s' % (previous, x))
        file.write('%s, %d' % (previous, 1))
file.close()
~
"reducer_t1.py" 27L, 489C                                    7,1            All
```

```python
#!/usr/bin/env python3
import sys
node = None
current_node = None
for i in sys.stdin:
        i = i.strip()
        node, rank = i.split('\t')
        try:
                rank = float(rank)
        except ValueError:
                continue
        if current_node==node:
                cumulative += rank
        else:
                if current_node:
                        updated_page_rank=0.15+0.85*cumulative
                        round(updated_page_rank,5)
                        print('%s, %f' % (current_node, updated_page_rank))
                cumulative = rank
                current_node= node
if current_node==node:
        updated_page_rank=0.15+0.85*cumulative
        round(updated_page_rank,5)
        print('%s, %f' % (current_node, updated_page_rank))
~
~
~
~
"reducer_t2.py" 24L, 570C                                      1,1            All
```

Driver Code:

```sh
#!/bin/sh
CONVERGE=1
rm v* log*
I=1
#$HADOOP_HOME/sbin/start-all.sh
$HADOOP_HOME/bin/hadoop dfsadmin -safemode leave
hdfs dfs -rm -r /output*

$HADOOP_HOME/bin/hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-*streami
ng*.jar \
-mapper "python3 /home/ajitesh/hadoop/pagerank/mapper_t1.py" \
-reducer "python3 /home/ajitesh/hadoop/pagerank/reducer_t1.py '/home/ajitesh/ha
doop/pagerank/v'"  \
-input input/my-web.txt \
-output /output1 #has adjacency list


while [ "$CONVERGE" -ne 0 ]
do
        echo $I
        $HADOOP_HOME/bin/hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-
*streaming*.jar \
        -mapper "python3 /home/ajitesh/hadoop/pagerank/mapper_t2.py '/home/ajit
esh/hadoop/pagerank/v' " \
        -reducer "python3 /home/ajitesh/hadoop/pagerank/reducer_t2.py" \
        -input /output1 \
        -output /output2
        touch v1
"iterate-hadoop.sh" 32L, 1002C                              19,1-8          Top
```

Convergence Code:


```python
import shutil
import os
count=0
n=0
conv =0.5 #this value will vary for different test cases in the backend
def rewrite_pagerank():
        os.remove("/home/ajitesh/hadoop/pagerank/v")

        source = "/home/ajitesh/hadoop/pagerank/v1"
        destination = "/home/ajitesh/hadoop/pagerank/v"
        dest = shutil.copyfile(source, destination)


with open("/home/ajitesh/hadoop/pagerank/v") as file1, open("/home/ajitesh/hado
op/pagerank/v1") as file2:
        for line1, line2 in zip(file1, file2):
                count+=1
                old_pagerank=float(line1.split(",")[1])
                new_pagerank=float(line2.split(",")[1])

                if(abs(old_pagerank-new_pagerank) < conv):
                        n+=1

        if(n==count):
                print(0)
        else:
                rewrite_pagerank()
"check_conv.py" 28L, 683C                              4,1              Top
```
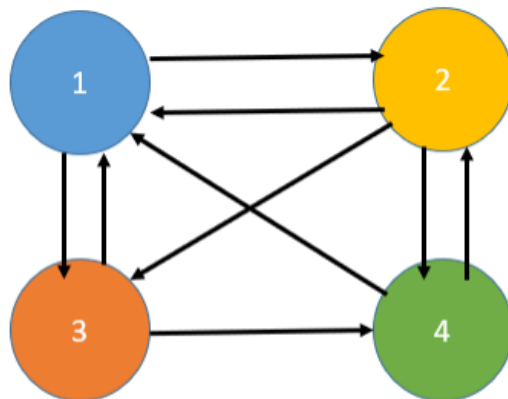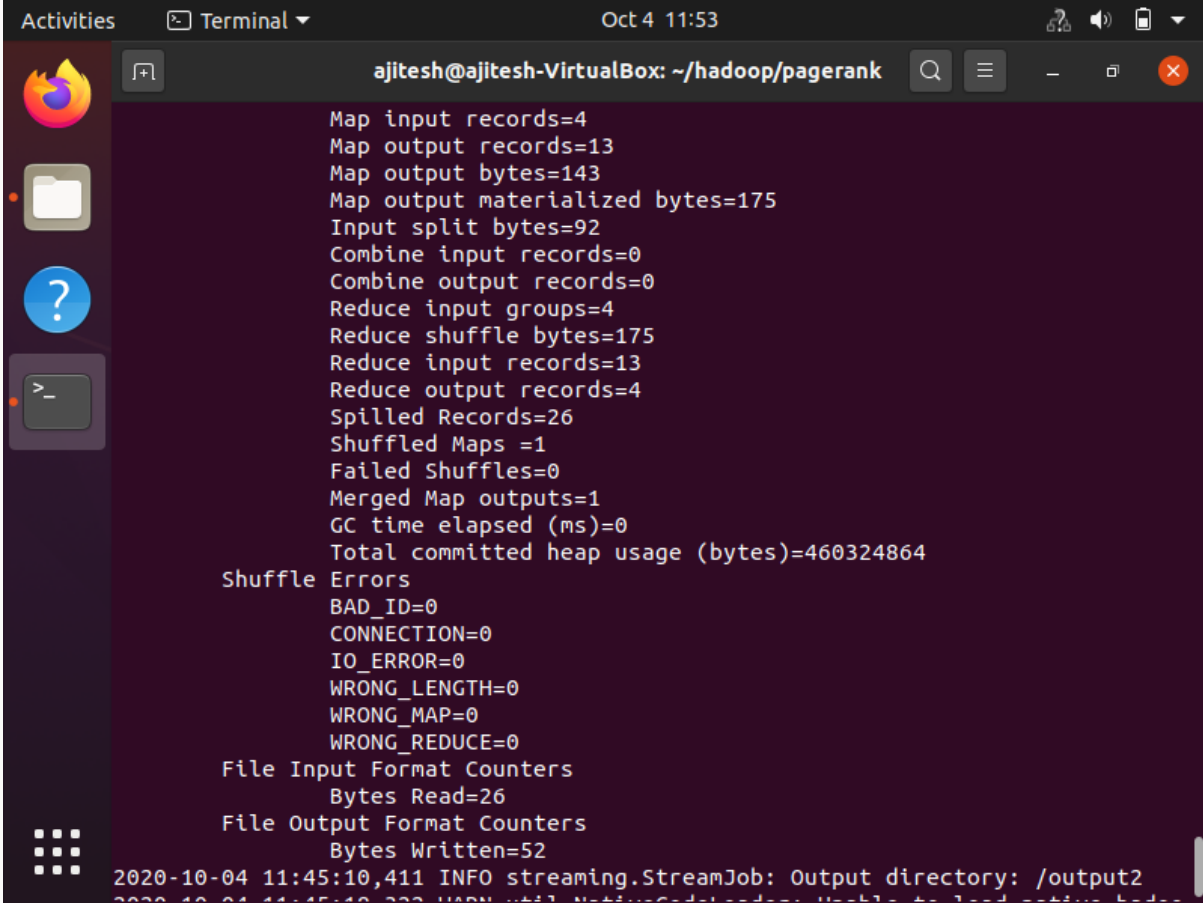
4 node web graph implementation:

Input:

ajitesh@ajitesh-VirtualBox: ~/hadoop/pagerank

```
1       2
1       3
2       1
2       3
2       4
3       1
3       4
4       1
4       2
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
"my-web.txt" 9L, 36C                          9,3-9        All
```

Execution


```
                Map input records=4
                Map output records=13
                Map output bytes=143
                Map output materialized bytes=175
                Input split bytes=92
                Combine input records=0
                Combine output records=0
                Reduce input groups=4
                Reduce shuffle bytes=175
                Reduce input records=13
                Reduce output records=4
                Spilled Records=26
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=460324864
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=26
        File Output Format Counters
                Bytes Written=52
2020-10-04 11:45:10,411 INFO streaming.StreamJob: Output directory: /output2
```

⊞   **ajitesh@ajitesh-VirtualBox: ~/hadoop/pagerank**   Q ☰   _ ⬜ ✕

```
2020-10-04 11:45:09,585 INFO mapred.LocalJobRunner: Finishing task: attempt_loc
al1588668936_0001_r_000000_0
2020-10-04 11:45:09,606 INFO mapred.LocalJobRunner: reduce task executor comple
te.
2020-10-04 11:45:10,179 INFO mapreduce.Job:  map 100% reduce 100%
2020-10-04 11:45:10,182 INFO mapreduce.Job: Job job_local1588668936_0001 comple
ted successfully
2020-10-04 11:45:10,357 INFO mapreduce.Job: Counters: 36
        File System Counters
                FILE: Number of bytes read=353884
                FILE: Number of bytes written=1417411
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=52
                HDFS: Number of bytes written=52
                HDFS: Number of read operations=15
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=4
                Map output records=13
                Map output bytes=143
                Map output materialized bytes=175
                Input split bytes=92
                Combine input records=0
                Combine output records=0
                Reduce input groups=4
```

Output

```
Activities        Terminal ▾                         Oct 4 11:56

                    ajitesh@ajitesh-VirtualBox: ~/hadoop/pagerank          Q  ≡   _  □  ×

ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$ hdfs dfs -ls /output1
2020-10-04 11:53:39,300 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 ajitesh supergroup          0 2020-10-04 11:44 /output1/_SUCCESS
-rw-r--r--   1 ajitesh supergroup         26 2020-10-04 11:44 /output1/part-000
00
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$ hdfs dfs -cat /output1/part-00000
2020-10-04 11:54:33,165 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
2020-10-04 11:54:40,657 INFO sasl.SaslDataTransferClient: SASL encryption trust
 check: localHostTrusted = false, remoteHostTrusted = false
1       2,3
2       1,3,4
3       1,4
4       1,2
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$ ls
check_conv.py      mapper_t1.py  my-web.txt      reducer_t2.py  v1
iterate-hadoop.sh  mapper_t2.py  reducer_t1.py   v              web-Google.txt
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$ cat v
1, 1
2, 1
3, 1
4, 1ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$ cat v1
1, 1.283333
2, 1.000000
3, 0.858333
4, 0.858333
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$ ▯
```
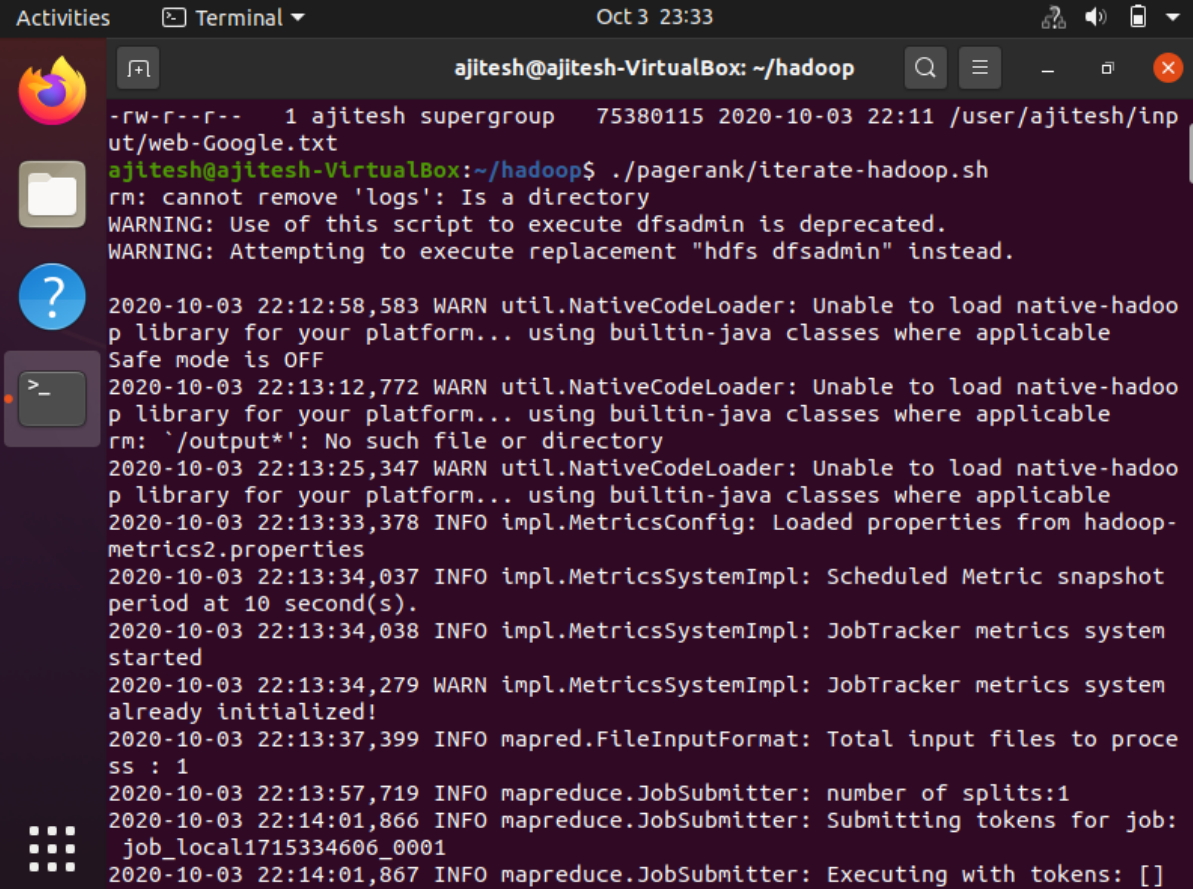
It can be seen that page 1 has the highest rank.

SNAP dataset page rank implementation:

Execution


```
Activities        Terminal                    Oct 3 23:33

                        ajitesh@ajitesh-VirtualBox: ~/hadoop

-rw-r--r--   1 ajitesh supergroup    75380115 2020-10-03 22:11 /user/ajitesh/inp
ut/web-Google.txt
ajitesh@ajitesh-VirtualBox:~/hadoop$ ./pagerank/iterate-hadoop.sh
rm: cannot remove 'logs': Is a directory
WARNING: Use of this script to execute dfsadmin is deprecated.
WARNING: Attempting to execute replacement "hdfs dfsadmin" instead.

2020-10-03 22:12:58,583 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
Safe mode is OFF
2020-10-03 22:13:12,772 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
rm: `/output*': No such file or directory
2020-10-03 22:13:25,347 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
2020-10-03 22:13:33,378 INFO impl.MetricsConfig: Loaded properties from hadoop-
metrics2.properties
2020-10-03 22:13:34,037 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot
period at 10 second(s).
2020-10-03 22:13:34,038 INFO impl.MetricsSystemImpl: JobTracker metrics system
started
2020-10-03 22:13:34,279 WARN impl.MetricsSystemImpl: JobTracker metrics system
already initialized!
2020-10-03 22:13:37,399 INFO mapred.FileInputFormat: Total input files to proce
ss : 1
2020-10-03 22:13:57,719 INFO mapreduce.JobSubmitter: number of splits:1
2020-10-03 22:14:01,866 INFO mapreduce.JobSubmitter: Submitting tokens for job:
  job_local1715334606_0001
2020-10-03 22:14:01,867 INFO mapreduce.JobSubmitter: Executing with tokens: []
```
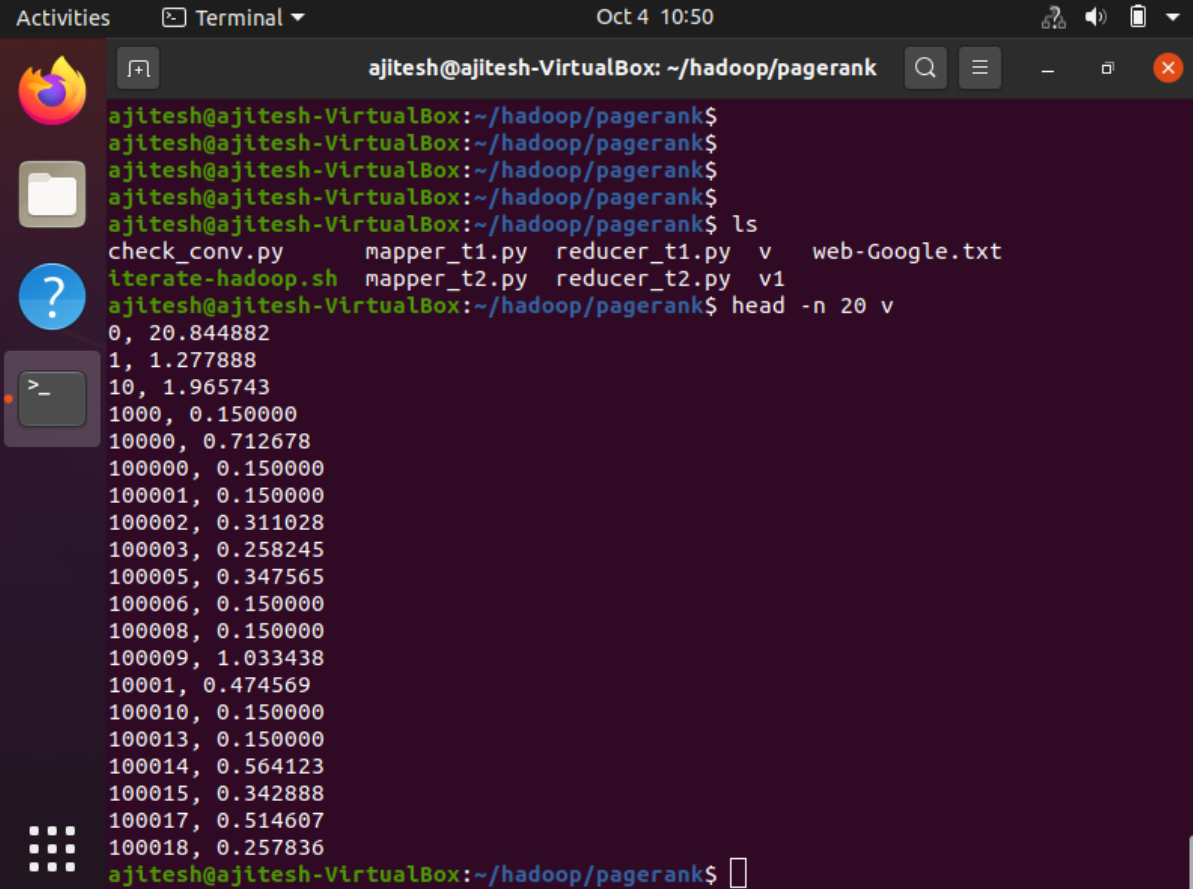
```
2020-10-03 22:25:59,148 INFO mapreduce.Job:  map 100% reduce 100%
2020-10-03 22:26:00,151 INFO mapreduce.Job: Job job_local93301098_0001 complete
d successfully
2020-10-03 22:26:00,309 INFO mapreduce.Job: Counters: 36
        File System Counters
                FILE: Number of bytes read=395175602
                FILE: Number of bytes written=494934215
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=80492266
                HDFS: Number of bytes written=13226340
                HDFS: Number of read operations=15
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=739454
                Map output records=5518768
                Map output bytes=87667967
                Map output materialized bytes=98705509
                Input split bytes=92
                Combine input records=0
                Combine output records=0
                Reduce input groups=739454
                Reduce shuffle bytes=98705509
                Reduce input records=5518768
                Reduce output records=739454
                Spilled Records=16556304
                Shuffled Maps =1
```

```
                Reduce input records=5518768
                Reduce output records=739454
                Spilled Records=16556304
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=132
                Total committed heap usage (bytes)=508559360
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=40246133
        File Output Format Counters
                Bytes Written=13225172
2020-10-04 02:54:05,482 INFO streaming.StreamJob: Output directory: /output2
2020-10-04 02:54:12,847 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
2020-10-04 02:54:19,999 INFO sasl.SaslDataTransferClient: SASL encryption trust
 check: localHostTrusted = false, remoteHostTrusted = false
2020-10-04 02:54:43,658 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
Deleted /output2
0
ajitesh@ajitesh-VirtualBox:~/hadoop$ []
```

Output

```
Activities          Terminal ▾                          Oct 4 10:50

                    ajitesh@ajitesh-VirtualBox: ~/hadoop/pagerank

ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$ ls
check_conv.py       mapper_t1.py  reducer_t1.py  v    web-Google.txt
iterate-hadoop.sh   mapper_t2.py  reducer_t2.py  v1
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$ head -n 20 v1
0, 20.844572
1, 1.277863
10, 1.965705
1000, 0.150000
10000, 0.712673
100000, 0.150000
100001, 0.150000
100002, 0.311027
100003, 0.258245
100005, 0.347565
100006, 0.150000
100008, 0.150000
100009, 1.033438
10001, 0.474293
100010, 0.150000
100013, 0.150000
100014, 0.564115
100015, 0.342886
100017, 0.514607
100018, 0.257871
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$
```

```
Activities          Terminal ▾                          Oct 4 10:56

                    ajitesh@ajitesh-VirtualBox: ~/hadoop/pagerank

ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$ cat v |awk '{print $2" " $1}'|sor
t -r|head
9.999733 207849,
9.999146 567525,
9.997385 8505,
9.995015 811255,
9.995015 650836,
9.995015 3518,
9.992176 176491,
9.991802 706514,
9.988013 877130,
9.984774 522701,
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$
```
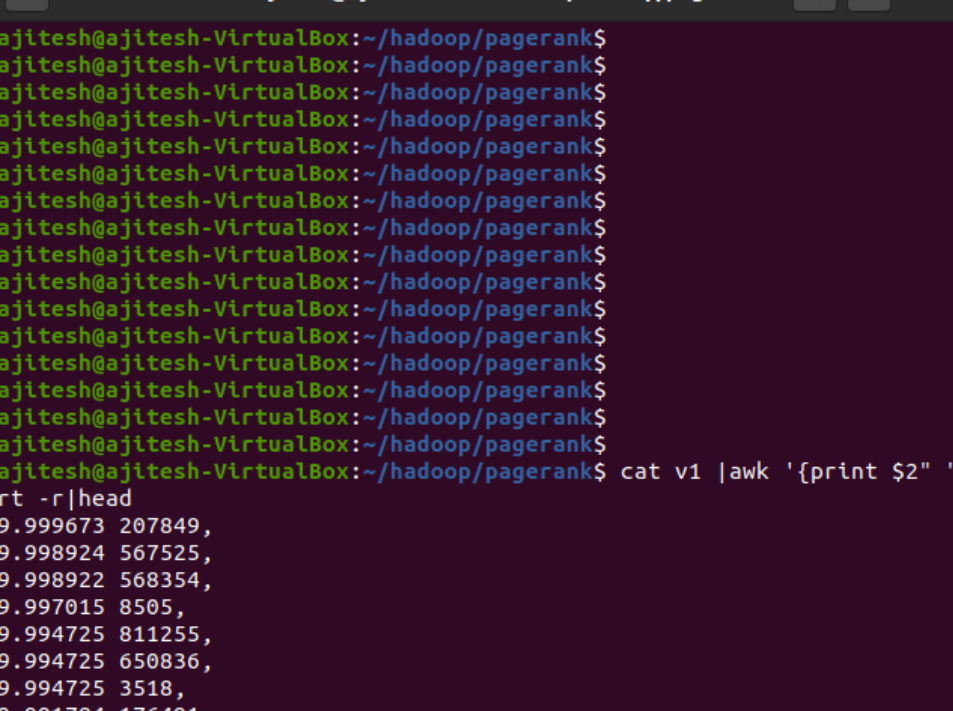
```
ajitesh@ajitesh-VirtualBox: ~/hadoop/pagerank
```

```
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$ cat v1 |awk '{print $2" " $1}'|so
rt -r|head
9.999673 207849,
9.998924 567525,
9.998922 568354,
9.997015 8505,
9.994725 811255,
9.994725 650836,
9.994725 3518,
9.991784 176491,
9.991698 706514,
9.987846 877130,
ajitesh@ajitesh-VirtualBox:~/hadoop/pagerank$ []
```

```
ajitesh@ajitesh-VirtualBox: ~/hadoop
```

```
ajitesh@ajitesh-VirtualBox:~/hadoop$
ajitesh@ajitesh-VirtualBox:~/hadoop$
ajitesh@ajitesh-VirtualBox:~/hadoop$
ajitesh@ajitesh-VirtualBox:~/hadoop$
ajitesh@ajitesh-VirtualBox:~/hadoop$
ajitesh@ajitesh-VirtualBox:~/hadoop$
ajitesh@ajitesh-VirtualBox:~/hadoop$
ajitesh@ajitesh-VirtualBox:~/hadoop$
ajitesh@ajitesh-VirtualBox:~/hadoop$
ajitesh@ajitesh-VirtualBox:~/hadoop$
ajitesh@ajitesh-VirtualBox:~/hadoop$
ajitesh@ajitesh-VirtualBox:~/hadoop$
ajitesh@ajitesh-VirtualBox:~/hadoop$
ajitesh@ajitesh-VirtualBox:~/hadoop$
ajitesh@ajitesh-VirtualBox:~/hadoop$
ajitesh@ajitesh-VirtualBox:~/hadoop$
ajitesh@ajitesh-VirtualBox:~/hadoop$
ajitesh@ajitesh-VirtualBox:~/hadoop$
ajitesh@ajitesh-VirtualBox:~/hadoop$ hdfs dfs -ls /output1
2020-10-04 11:36:23,259 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 ajitesh supergroup          0 2020-10-03 22:19 /output1/_SUCCESS
-rw-r--r--   1 ajitesh supergroup   40246133 2020-10-03 22:19 /output1/part-000
00
ajitesh@ajitesh-VirtualBox:~/hadoop$ []
```

It took 40 iterations for the values to converge.It took several hours for the 40 iterations to complete.Node 207849 is found to have the highest pagerank.V and v1 are the output files which are the pageranks before and after the final iteration.