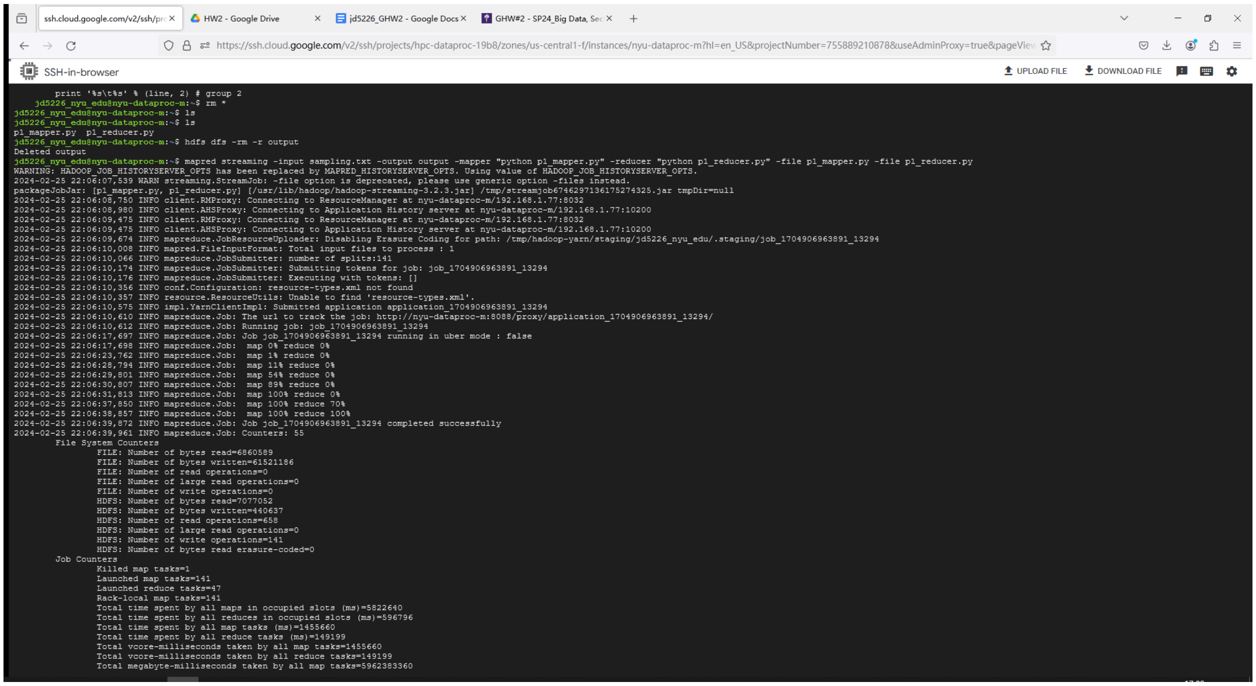
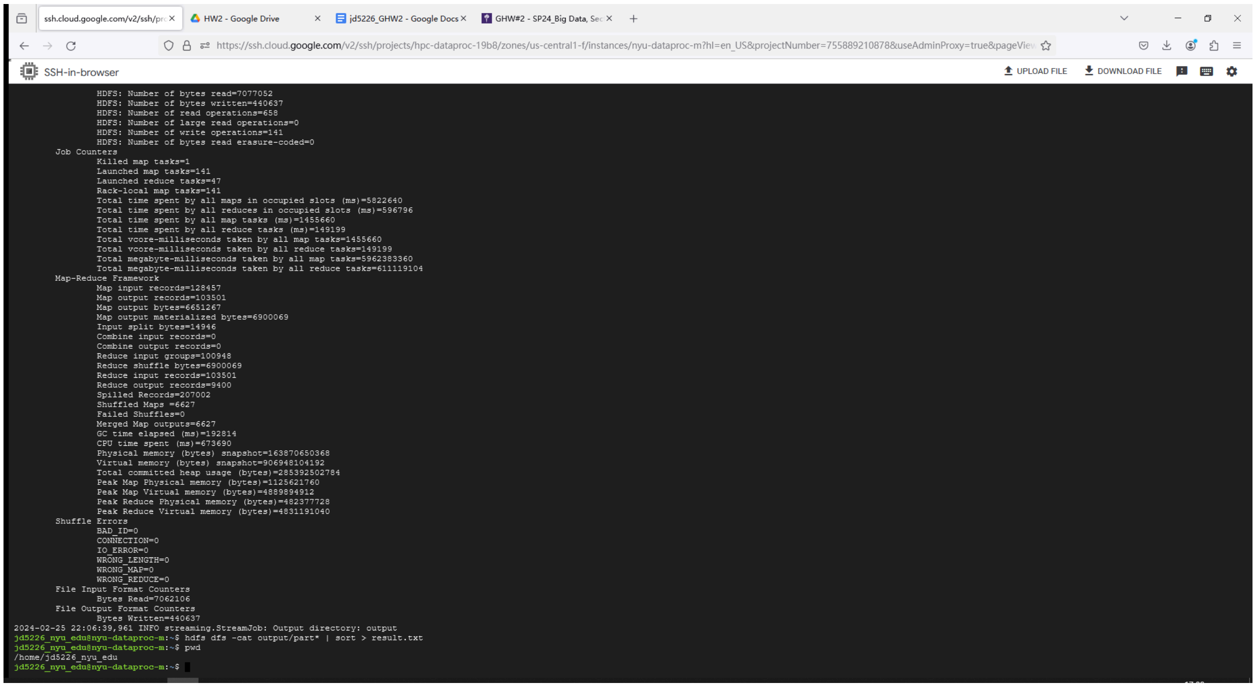
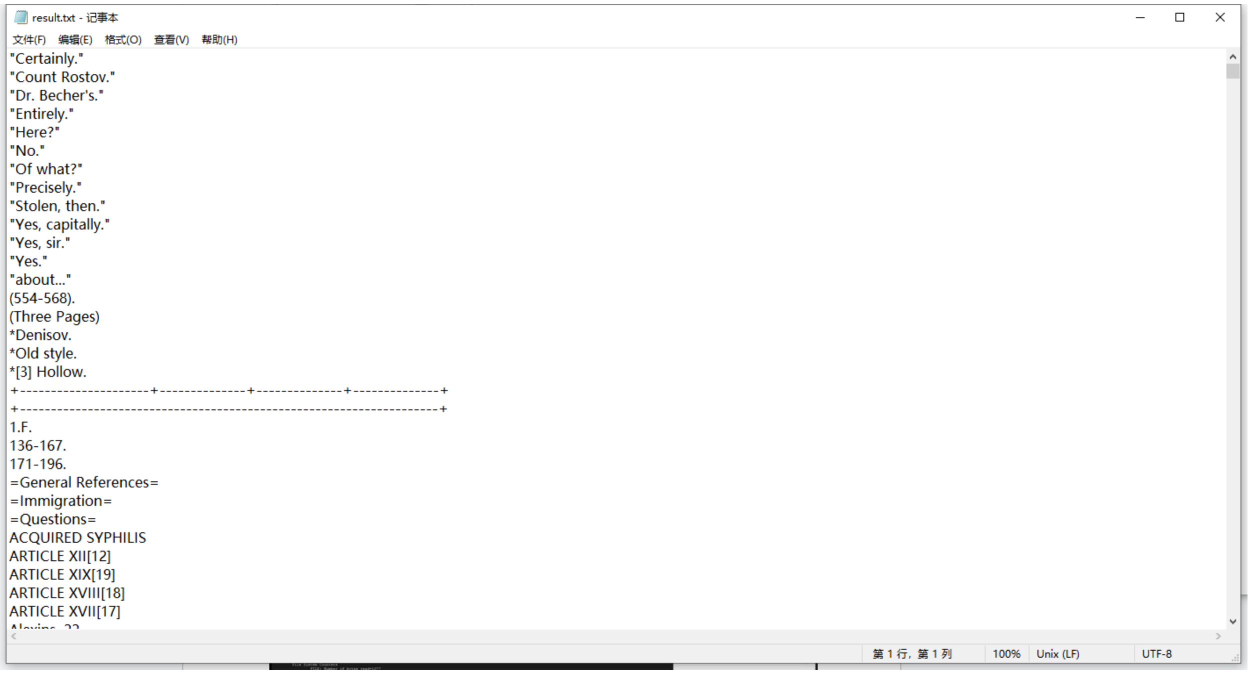

ssh.cloud.google.com/v2/ssh/pr
HW2 - Google Drive
jd5226_GHW2 - Google Docs
GHW#2 - SP24_Big Data, Sec
https://ssh.cloud.google.com/v2/ssh/projects/hpc-dataproc-19b8/zones/us-central1-f/instances/nyu-dataproc-m?hl=en_US&projectNumber=755889210878&useAdminProxy=true&pageView
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE

```
        print '%s\t%s' % (line, 2) # group 2
    jd5226_nyu_edu@nyu-dataproc-m:~$ rm *
jd5226_nyu_edu@nyu-dataproc-m:~$ ls
jd5226_nyu_edu@nyu-dataproc-m:~$ ls
p1_mapper.py  p1_reducer.py
jd5226_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -rm -r output
Deleted output
jd5226_nyu_edu@nyu-dataproc-m:~$ mapred streaming -input sampling.txt -output output -mapper "python p1_mapper.py" -reducer "python p1_reducer.py" -file p1_mapper.py -file p1_reducer.py
WARNING: HADOOP_JOB_HISTORYSERVER_OPTS has been replaced by MAPRED_HISTORYSERVER_OPTS. Using value of HADOOP_JOB_HISTORYSERVER_OPTS.
2024-02-25 22:06:07,539 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [p1_mapper.py, p1_reducer.py] [/usr/lib/hadoop/hadoop-streaming-3.2.3.jar] /tmp/streamjob6746297136175274325.jar tmpDir=null
2024-02-25 22:06:08,750 INFO client.RMProxy: Connecting to ResourceManager at nyu-dataproc-m/192.168.1.77:8032
2024-02-25 22:06:08,980 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m/192.168.1.77:10200
2024-02-25 22:06:09,475 INFO client.RMProxy: Connecting to ResourceManager at nyu-dataproc-m/192.168.1.77:8032
2024-02-25 22:06:09,475 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m/192.168.1.77:10200
2024-02-25 22:06:09,674 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/jd5226_nyu_edu/.staging/job_1704906963891_13294
2024-02-25 22:06:10,008 INFO mapred.FileInputFormat: Total input files to process : 1
2024-02-25 22:06:10,066 INFO mapreduce.JobSubmitter: number of splits:141
2024-02-25 22:06:10,174 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1704906963891_13294
2024-02-25 22:06:10,176 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-02-25 22:06:10,356 INFO conf.Configuration: resource-types.xml not found
2024-02-25 22:06:10,357 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-02-25 22:06:10,575 INFO impl.YarnClientImpl: Submitted application application_1704906963891_13294
2024-02-25 22:06:10,610 INFO mapreduce.Job: The url to track the job: http://nyu-dataproc-m:8088/proxy/application_1704906963891_13294/
2024-02-25 22:06:10,612 INFO mapreduce.Job: Running job: job_1704906963891_13294
2024-02-25 22:06:17,697 INFO mapreduce.Job: Job job_1704906963891_13294 running in uber mode : false
2024-02-25 22:06:17,698 INFO mapreduce.Job:  map 0% reduce 0%
2024-02-25 22:06:23,762 INFO mapreduce.Job:  map 1% reduce 0%
2024-02-25 22:06:28,794 INFO mapreduce.Job:  map 11% reduce 0%
2024-02-25 22:06:29,801 INFO mapreduce.Job:  map 54% reduce 0%
2024-02-25 22:06:30,807 INFO mapreduce.Job:  map 89% reduce 0%
2024-02-25 22:06:31,813 INFO mapreduce.Job:  map 100% reduce 0%
2024-02-25 22:06:37,850 INFO mapreduce.Job:  map 100% reduce 70%
2024-02-25 22:06:38,857 INFO mapreduce.Job:  map 100% reduce 100%
2024-02-25 22:06:39,872 INFO mapreduce.Job: Job job_1704906963891_13294 completed successfully
2024-02-25 22:06:39,961 INFO mapreduce.Job: Counters: 55
        File System Counters
                FILE: Number of bytes read=6860589
                FILE: Number of bytes written=61521186
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=7077052
                HDFS: Number of bytes written=440637
                HDFS: Number of read operations=658
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=141
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=1
                Launched map tasks=141
                Launched reduce tasks=47
                Rack-local map tasks=141
                Total time spent by all maps in occupied slots (ms)=5822640
                Total time spent by all reduces in occupied slots (ms)=596796
                Total time spent by all map tasks (ms)=1455660
                Total time spent by all reduce tasks (ms)=149199
                Total vcore-milliseconds taken by all map tasks=1455660
                Total vcore-milliseconds taken by all reduce tasks=149199
                Total megabyte-milliseconds taken by all map tasks=5962383360
```


ssh.cloud.google.com/v2/ssh/pr
HW2 - Google Drive
jd5226_GHW2 - Google Docs
GHW#2 - SP24_Big Data, Sec
https://ssh.cloud.google.com/v2/ssh/projects/hpc-dataproc-19b8/zones/us-central1-f/instances/nyu-dataproc-m?hl=en_US&projectNumber=755889210878&useAdminProxy=true&pageView
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE

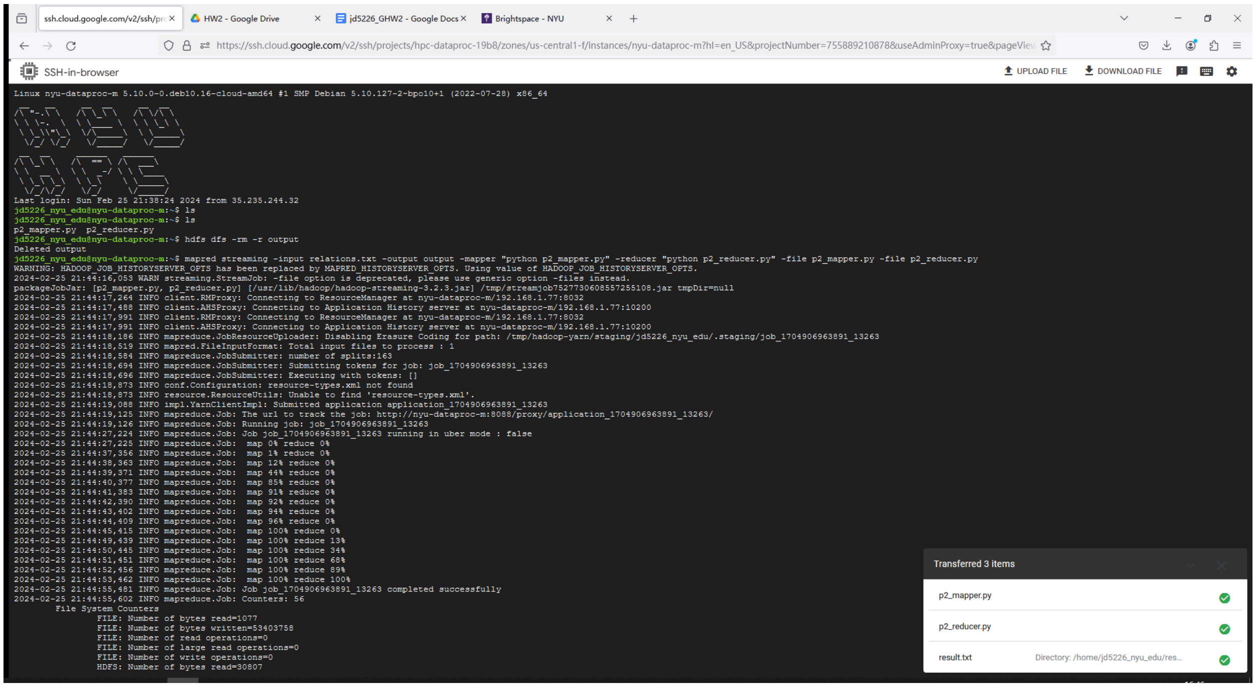
```
                HDFS: Number of bytes read=7077052
                HDFS: Number of bytes written=440637
                HDFS: Number of read operations=658
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=141
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=1
                Launched map tasks=141
                Launched reduce tasks=47
                Rack-local map tasks=141
                Total time spent by all maps in occupied slots (ms)=5822640
                Total time spent by all reduces in occupied slots (ms)=596796
                Total time spent by all map tasks (ms)=1455660
                Total time spent by all reduce tasks (ms)=149199
                Total vcore-milliseconds taken by all map tasks=1455660
                Total vcore-milliseconds taken by all reduce tasks=149199
                Total megabyte-milliseconds taken by all map tasks=5962383360
                Total megabyte-milliseconds taken by all reduce tasks=611119104
        Map-Reduce Framework
                Map input records=128457
                Map output records=103501
                Map output bytes=6651267
                Map output materialized bytes=6900069
                Input split bytes=14946
                Combine input records=0
                Combine output records=0
                Reduce input groups=100948
                Reduce shuffle bytes=6900069
                Reduce input records=103501
                Reduce output records=9400
                Spilled Records=207002
                Shuffled Maps =6627
                Failed Shuffles=0
                Merged Map outputs=6627
                GC time elapsed (ms)=192814
                CPU time spent (ms)=673690
                Physical memory (bytes) snapshot=163870650368
                Virtual memory (bytes) snapshot=906948104192
                Total committed heap usage (bytes)=285392502784
                Peak Map Physical memory (bytes)=1125621760
                Peak Map Virtual memory (bytes)=4889894912
                Peak Reduce Physical memory (bytes)=482377728
                Peak Reduce Virtual memory (bytes)=4831191040
        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=7062106
        File Output Format Counters
                Bytes Written=440637
2024-02-25 22:06:39,961 INFO streaming.StreamJob: Output directory: output
jd5226_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -cat output/part* | sort > result.txt
jd5226_nyu_edu@nyu-dataproc-m:~$ pwd
/home/jd5226_nyu_edu
jd5226_nyu_edu@nyu-dataproc-m:~$ 
```

P1 running on Dataproc

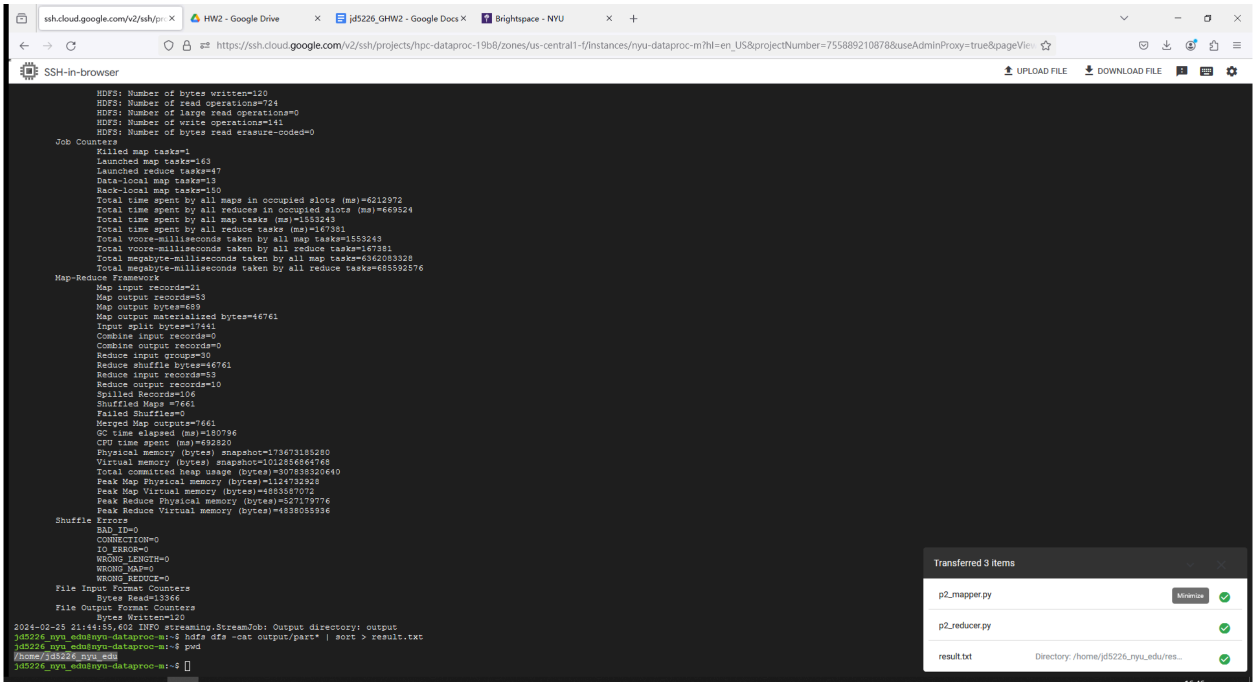


```
"Certainly."
"Count Rostov."
"Dr. Becher's."
"Entirely."
"Here?"
"No."
"Of what?"
"Precisely."
"Stolen, then."
"Yes, capitally."
"Yes, sir."
"Yes."
"about..."
(554-568).
(Three Pages)
*Denisov.
*Old style.
*[3] Hollow.
+--------------------+-------------+-------------+-------------+
+--------------------------------------------------------------+
1.F.
136-167.
171-196.
=General References=
=Immigration=
=Questions=
ACQUIRED SYPHILIS
ARTICLE XII[12]
ARTICLE XIX[19]
ARTICLE XVIII[18]
ARTICLE XVII[17]
```

P1 result

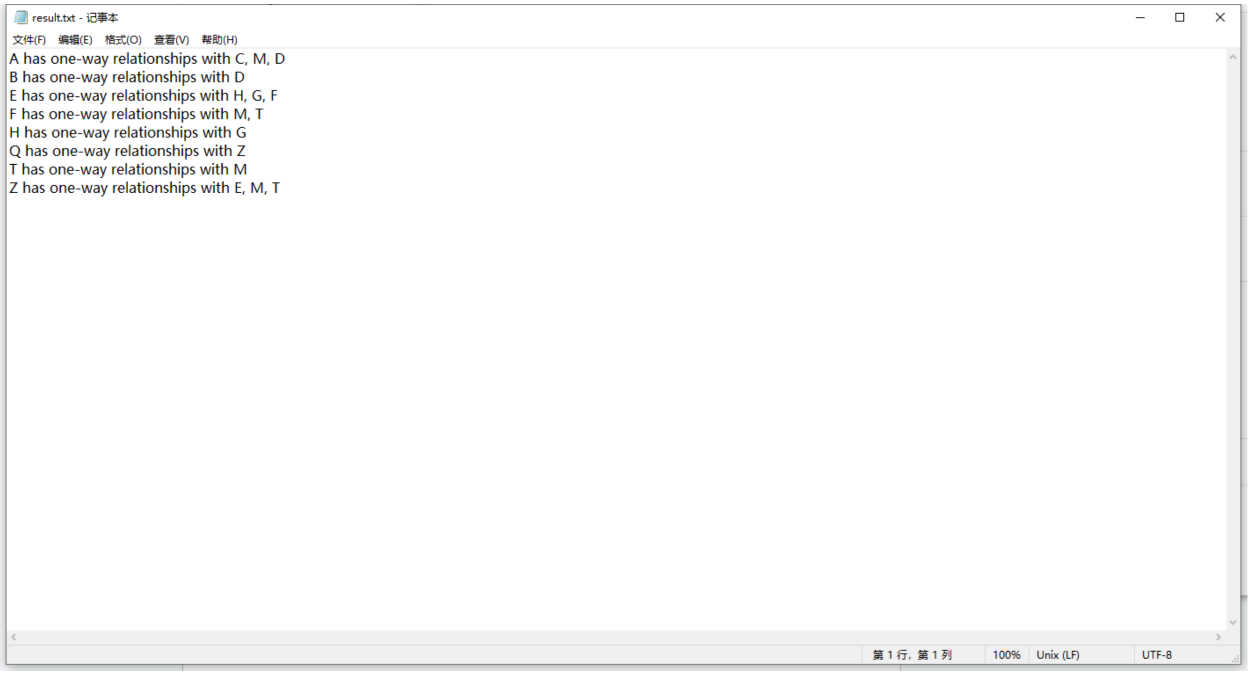

ssh.cloud.google.com/v2/ssh/prc
HW2 - Google Drive
jd5226_GHW2 - Google Docs
Brightspace - NYU
https://ssh.cloud.google.com/v2/ssh/projects/hpc-dataproc-19b8/zones/us-central1-f/instances/nyu-dataproc-m?hl=en_US&projectNumber=755889210878&useAdminProxy=true&pageView
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE
Transferred 3 items
p2_mapper.py
p2_reducer.py
result.txt
Directory: /home/jd5226_nyu_edu/res...

```
Linux nyu-dataproc-m 5.10.0-0.deb10.16-cloud-amd64 #1 SMP Debian 5.10.127-2~bpo10+1 (2022-07-28) x86_64
 __   __     __  __     __  __
/\ "-.\ \   /\ \_\ \   /\ \/\ \
\ \ \-.  \  \ \____ \  \ \ \_\ \
 \ \_\\"\_\  \/\_____\  \ \_____\
  \/_/ \/_/   \/_____/   \/_____/
 __  __     ______   ______
/\ \_\ \   /\  == \ /\  ___\
\ \  __ \  \ \  _-/ \ \ \____
 \ \_\ \_\  \ \_\    \ \_____\
  \/_/\/_/   \/_/     \/_____/
Last login: Sun Feb 25 21:38:24 2024 from 35.235.244.32
jd5226_nyu_edu@nyu-dataproc-m:~$ ls
jd5226_nyu_edu@nyu-dataproc-m:~$ ls
p2_mapper.py  p2_reducer.py
jd5226_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -rm -r output
Deleted output
jd5226_nyu_edu@nyu-dataproc-m:~$ mapred streaming -input relations.txt -output output -mapper "python p2_mapper.py" -reducer "python p2_reducer.py" -file p2_mapper.py -file p2_reducer.py
WARNING: HADOOP_JOB_HISTORYSERVER_OPTS has been replaced by MAPRED_HISTORYSERVER_OPTS. Using value of HADOOP_JOB_HISTORYSERVER_OPTS.
2024-02-25 21:44:16,053 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [p2_mapper.py, p2_reducer.py] [/usr/lib/hadoop/hadoop-streaming-3.2.3.jar] /tmp/streamjob7527730608557255108.jar tmpDir=null
2024-02-25 21:44:17,264 INFO client.RMProxy: Connecting to ResourceManager at nyu-dataproc-m/192.168.1.77:8032
2024-02-25 21:44:17,488 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m/192.168.1.77:10200
2024-02-25 21:44:17,991 INFO client.RMProxy: Connecting to ResourceManager at nyu-dataproc-m/192.168.1.77:8032
2024-02-25 21:44:17,991 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m/192.168.1.77:10200
2024-02-25 21:44:18,186 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/jd5226_nyu_edu/.staging/job_1704906963891_13263
2024-02-25 21:44:18,519 INFO mapred.FileInputFormat: Total input files to process : 1
2024-02-25 21:44:18,584 INFO mapreduce.JobSubmitter: number of splits:163
2024-02-25 21:44:18,694 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1704906963891_13263
2024-02-25 21:44:18,696 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-02-25 21:44:18,873 INFO conf.Configuration: resource-types.xml not found
2024-02-25 21:44:18,873 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-02-25 21:44:19,088 INFO impl.YarnClientImpl: Submitted application application_1704906963891_13263
2024-02-25 21:44:19,125 INFO mapreduce.Job: The url to track the job: http://nyu-dataproc-m:8088/proxy/application_1704906963891_13263/
2024-02-25 21:44:19,126 INFO mapreduce.Job: Running job: job_1704906963891_13263
2024-02-25 21:44:27,224 INFO mapreduce.Job: Job job_1704906963891_13263 running in uber mode : false
2024-02-25 21:44:27,225 INFO mapreduce.Job:  map 0% reduce 0%
2024-02-25 21:44:37,356 INFO mapreduce.Job:  map 1% reduce 0%
2024-02-25 21:44:38,363 INFO mapreduce.Job:  map 12% reduce 0%
2024-02-25 21:44:39,371 INFO mapreduce.Job:  map 44% reduce 0%
2024-02-25 21:44:40,377 INFO mapreduce.Job:  map 85% reduce 0%
2024-02-25 21:44:41,383 INFO mapreduce.Job:  map 91% reduce 0%
2024-02-25 21:44:42,390 INFO mapreduce.Job:  map 92% reduce 0%
2024-02-25 21:44:43,402 INFO mapreduce.Job:  map 94% reduce 0%
2024-02-25 21:44:44,409 INFO mapreduce.Job:  map 96% reduce 0%
2024-02-25 21:44:45,415 INFO mapreduce.Job:  map 100% reduce 0%
2024-02-25 21:44:49,439 INFO mapreduce.Job:  map 100% reduce 13%
2024-02-25 21:44:50,445 INFO mapreduce.Job:  map 100% reduce 34%
2024-02-25 21:44:51,451 INFO mapreduce.Job:  map 100% reduce 68%
2024-02-25 21:44:52,456 INFO mapreduce.Job:  map 100% reduce 89%
2024-02-25 21:44:53,462 INFO mapreduce.Job:  map 100% reduce 100%
2024-02-25 21:44:55,481 INFO mapreduce.Job: Job job_1704906963891_13263 completed successfully
2024-02-25 21:44:55,602 INFO mapreduce.Job: Counters: 56
        File System Counters
                FILE: Number of bytes read=1077
                FILE: Number of bytes written=53403758
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=30807
```


ssh.cloud.google.com/v2/ssh/pr
HW2 - Google Drive
jd5226_GHW2 - Google Docs
Brightspace - NYU
https://ssh.cloud.google.com/v2/ssh/projects/hpc-dataproc-19b8/zones/us-central1-f/instances/nyu-dataproc-m?hl=en_US&projectNumber=755889210878&useAdminProxy=true&pageView
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE
Transferred 3 items
p2_mapper.py
Minimize
p2_reducer.py
result.txt
Directory: /home/jd5226_nyu_edu/res...

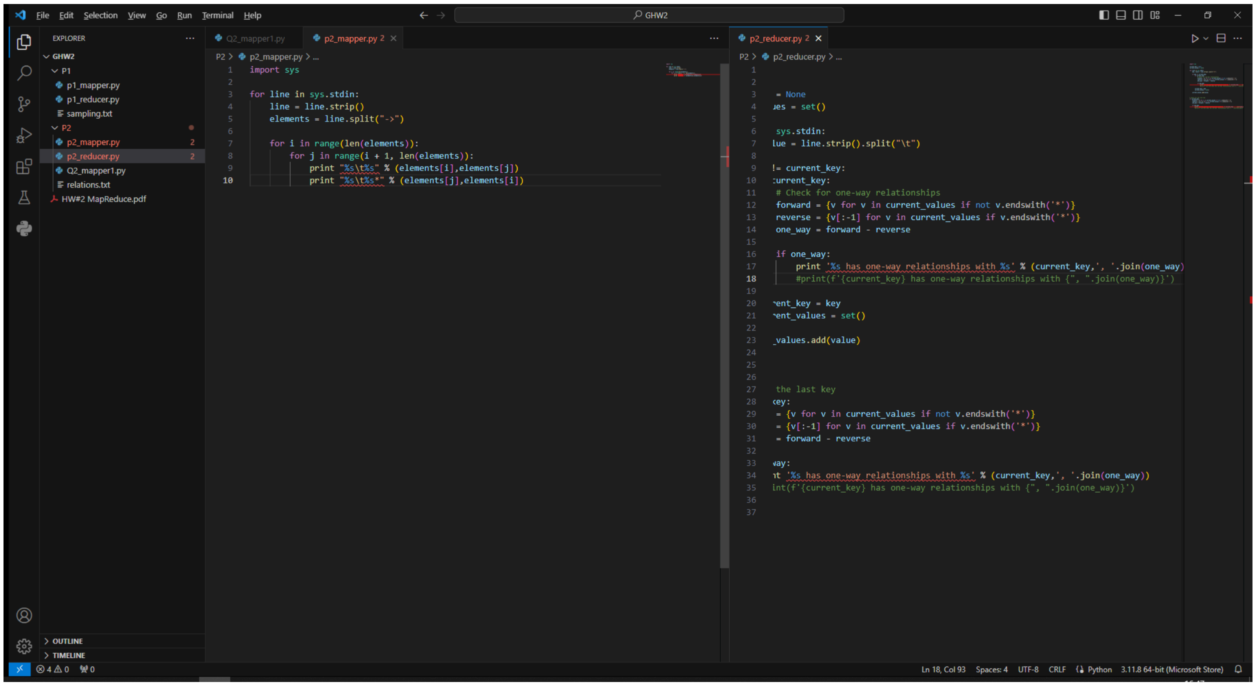
```
                HDFS: Number of bytes written=120
                HDFS: Number of read operations=724
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=141
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=1
                Launched map tasks=163
                Launched reduce tasks=47
                Data-local map tasks=13
                Rack-local map tasks=150
                Total time spent by all maps in occupied slots (ms)=6212972
                Total time spent by all reduces in occupied slots (ms)=669524
                Total time spent by all map tasks (ms)=1553243
                Total time spent by all reduce tasks (ms)=167381
                Total vcore-milliseconds taken by all map tasks=1553243
                Total vcore-milliseconds taken by all reduce tasks=167381
                Total megabyte-milliseconds taken by all map tasks=6362083328
                Total megabyte-milliseconds taken by all reduce tasks=685592576
        Map-Reduce Framework
                Map input records=21
                Map output records=53
                Map output bytes=689
                Map output materialized bytes=46761
                Input split bytes=17441
                Combine input records=0
                Combine output records=0
                Reduce input groups=30
                Reduce shuffle bytes=46761
                Reduce input records=53
                Reduce output records=10
                Spilled Records=106
                Shuffled Maps =7661
                Failed Shuffles=0
                Merged Map outputs=7661
                GC time elapsed (ms)=180796
                CPU time spent (ms)=692820
                Physical memory (bytes) snapshot=173673185280
                Virtual memory (bytes) snapshot=1012856864768
                Total committed heap usage (bytes)=307838320640
                Peak Map Physical memory (bytes)=1124732928
                Peak Map Virtual memory (bytes)=4883587072
                Peak Reduce Physical memory (bytes)=527179776
                Peak Reduce Virtual memory (bytes)=4838055936
        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=13366
        File Output Format Counters
                Bytes Written=120
2024-02-25 21:44:55,602 INFO streaming.StreamJob: Output directory: output
jd5226_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -cat output/part* | sort > result.txt
jd5226_nyu_edu@nyu-dataproc-m:~$ pwd
/home/jd5226_nyu_edu
jd5226_nyu_edu@nyu-dataproc-m:~$ 
```

P2 Running on dataproc



```
A has one-way relationships with C, M, D
B has one-way relationships with D
E has one-way relationships with H, G, F
F has one-way relationships with M, T
H has one-way relationships with G
Q has one-way relationships with Z
T has one-way relationships with M
Z has one-way relationships with E, M, T
```

Result file snippet



```
import sys

for line in sys.stdin:
    line = line.strip()
    elements = line.split("->")

    for i in range(len(elements)):
        for j in range(i + 1, len(elements)):
            print "%s\t%s" % (elements[i],elements[j])
            print "%s\t%s*" % (elements[j],elements[i])
```

```


 = None
ues = set()

 sys.stdin:
lue = line.strip().split("\t")

!= current_key:
current_key:
 # Check for one-way relationships
 forward = {v for v in current_values if not v.endswith('*')}
 reverse = {v[:-1] for v in current_values if v.endswith('*')}
 one_way = forward - reverse

 if one_way:
     print '%s has one-way relationships with %s' % (current_key,', '.join(one_way)
     #print(f'{current_key} has one-way relationships with {", ".join(one_way)}')

rent_key = key
rent_values = set()

_values.add(value)



 the last key
key:
 = {v for v in current_values if not v.endswith('*')}
 = {v[:-1] for v in current_values if v.endswith('*')}
 = forward - reverse

way:
nt '%s has one-way relationships with %s' % (current_key,', '.join(one_way))
int(f'{current_key} has one-way relationships with {", ".join(one_way)}')


```

P2 code snippet