BIG DATA

SECTION D

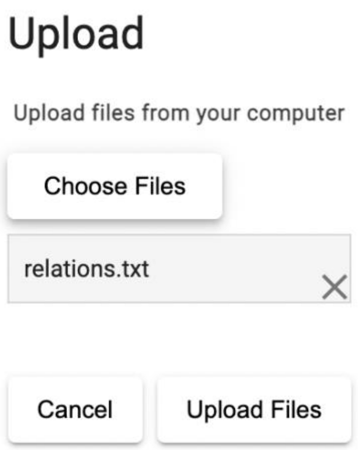
Fall’2023

Yogya Sharma

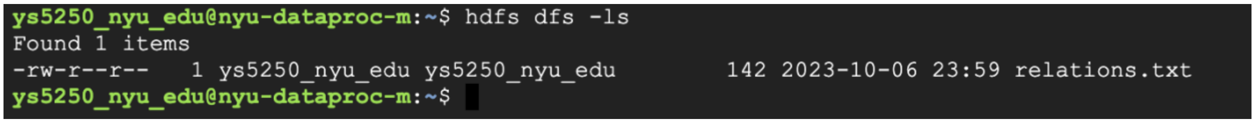
ys5250

GHW#2

Uploaded the text file relations.txt:



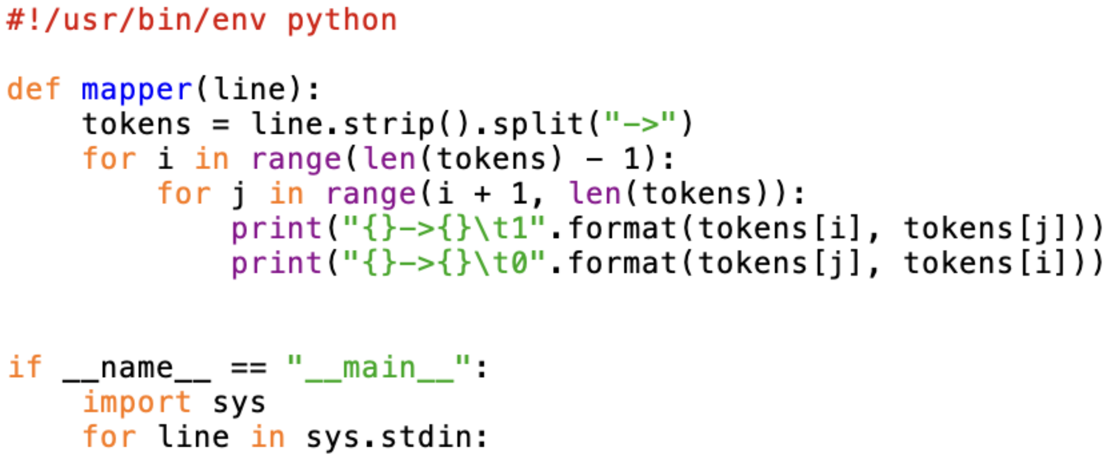
Put the file on HDFS:



```
ys5250_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -ls
Found 1 items
-rw-r--r--   1 ys5250_nyu_edu ys5250_nyu_edu        142 2023-10-06 23:59 relations.txt
ys5250_nyu_edu@nyu-dataproc-m:~$ 
```

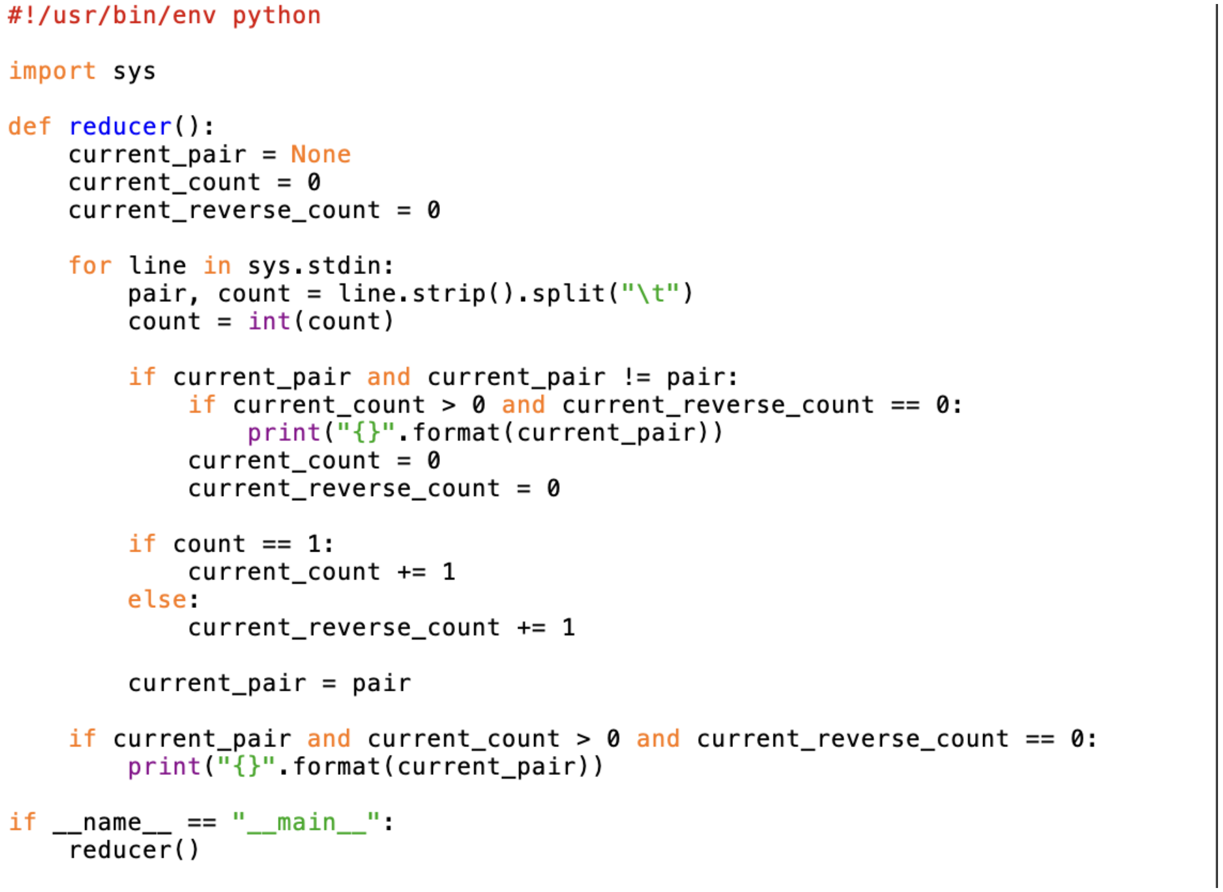
Create mapper and reducer files:

mapper.py



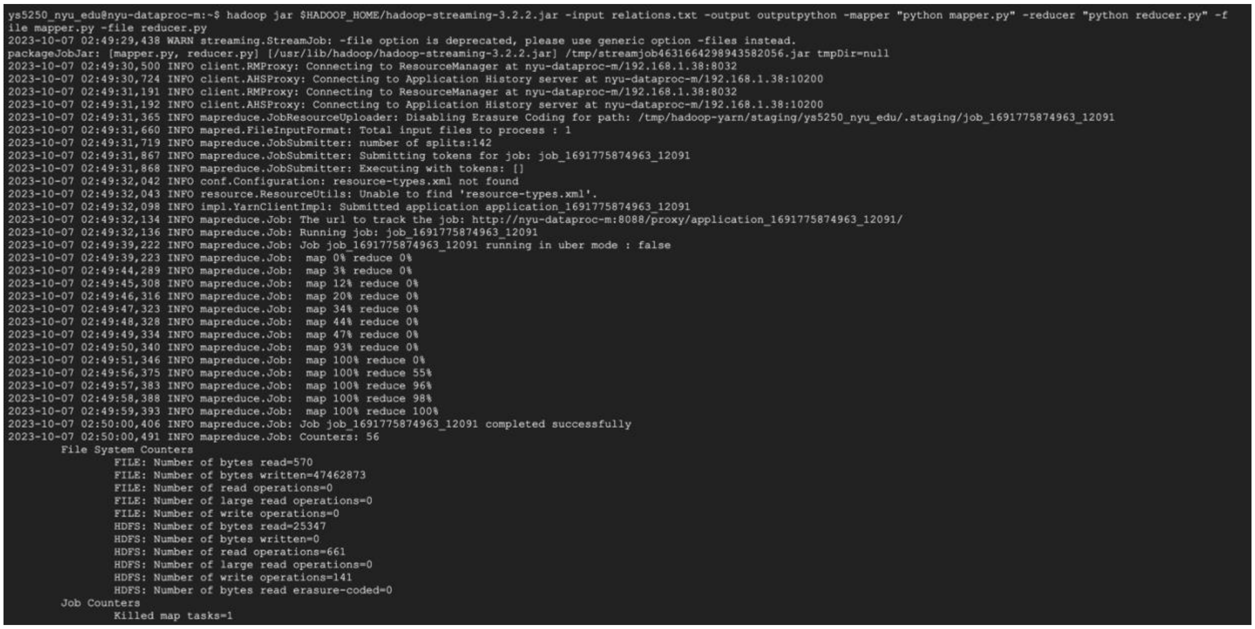
```
#!/usr/bin/env python

def mapper(line):
    tokens = line.strip().split("->")
    for i in range(len(tokens) - 1):
        for j in range(i + 1, len(tokens)):
            print("{}->{}\t1".format(tokens[i], tokens[j]))
            print("{}->{}\t0".format(tokens[j], tokens[i]))


if __name__ == "__main__":
    import sys
    for line in sys.stdin:
```

reducer.py

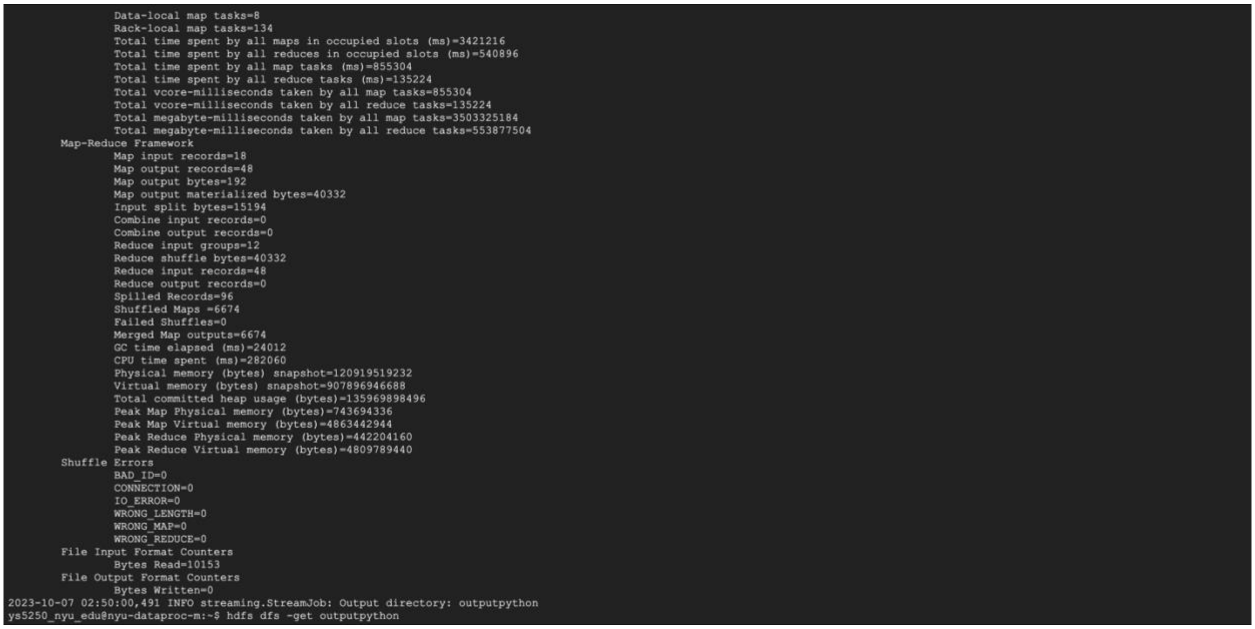


```python
#!/usr/bin/env python

import sys

def reducer():
    current_pair = None
    current_count = 0
    current_reverse_count = 0

    for line in sys.stdin:
        pair, count = line.strip().split("\t")
        count = int(count)

        if current_pair and current_pair != pair:
            if current_count > 0 and current_reverse_count == 0:
                print("{}".format(current_pair))
            current_count = 0
            current_reverse_count = 0

        if count == 1:
            current_count += 1
        else:
            current_reverse_count += 1

        current_pair = pair

    if current_pair and current_count > 0 and current_reverse_count == 0:
        print("{}".format(current_pair))

if __name__ == "__main__":
    reducer()
```

Run the map and reduce job:

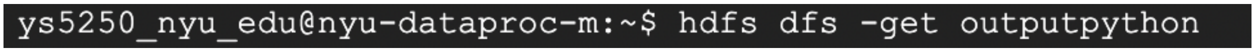


```
ys5250_nyu_edu@nyu-dataproc-m:~$ hadoop jar $HADOOP_HOME/hadoop-streaming-3.2.2.jar -input relations.txt -output outputpython -mapper "python mapper.py" -reducer "python reducer.py" -f
ile mapper.py -file reducer.py
2023-10-07 02:49:29,438 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [mapper.py, reducer.py] [/usr/lib/hadoop/hadoop-streaming-3.2.2.jar] /tmp/streamjob4631664298943582056.jar tmpDir=null
2023-10-07 02:49:30,500 INFO client.RMProxy: Connecting to ResourceManager at nyu-dataproc-m/192.168.1.38:8032
2023-10-07 02:49:30,724 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m/192.168.1.38:10200
2023-10-07 02:49:31,191 INFO client.RMProxy: Connecting to ResourceManager at nyu-dataproc-m/192.168.1.38:8032
2023-10-07 02:49:31,192 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m/192.168.1.38:10200
2023-10-07 02:49:31,365 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/ys5250_nyu_edu/.staging/job_1691775874963_12091
2023-10-07 02:49:31,660 INFO mapred.FileInputFormat: Total input files to process : 1
2023-10-07 02:49:31,719 INFO mapreduce.JobSubmitter: number of splits:142
2023-10-07 02:49:31,867 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1691775874963_12091
2023-10-07 02:49:31,868 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-10-07 02:49:32,042 INFO conf.Configuration: resource-types.xml not found
2023-10-07 02:49:32,043 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-10-07 02:49:32,098 INFO impl.YarnClientImpl: Submitted application application_1691775874963_12091
2023-10-07 02:49:32,134 INFO mapreduce.Job: The url to track the job: http://nyu-dataproc-m:8088/proxy/application_1691775874963_12091/
2023-10-07 02:49:32,136 INFO mapreduce.Job: Running job: job_1691775874963_12091
2023-10-07 02:49:39,222 INFO mapreduce.Job: Job job_1691775874963_12091 running in uber mode : false
2023-10-07 02:49:39,223 INFO mapreduce.Job:  map 0% reduce 0%
2023-10-07 02:49:44,289 INFO mapreduce.Job:  map 3% reduce 0%
2023-10-07 02:49:45,308 INFO mapreduce.Job:  map 12% reduce 0%
2023-10-07 02:49:46,316 INFO mapreduce.Job:  map 20% reduce 0%
2023-10-07 02:49:47,323 INFO mapreduce.Job:  map 34% reduce 0%
2023-10-07 02:49:48,328 INFO mapreduce.Job:  map 44% reduce 0%
2023-10-07 02:49:49,334 INFO mapreduce.Job:  map 47% reduce 0%
2023-10-07 02:49:50,340 INFO mapreduce.Job:  map 93% reduce 0%
2023-10-07 02:49:51,346 INFO mapreduce.Job:  map 100% reduce 0%
2023-10-07 02:49:56,375 INFO mapreduce.Job:  map 100% reduce 55%
2023-10-07 02:49:57,383 INFO mapreduce.Job:  map 100% reduce 96%
2023-10-07 02:49:58,388 INFO mapreduce.Job:  map 100% reduce 98%
2023-10-07 02:49:59,393 INFO mapreduce.Job:  map 100% reduce 100%
2023-10-07 02:50:00,406 INFO mapreduce.Job: Job job_1691775874963_12091 completed successfully
2023-10-07 02:50:00,491 INFO mapreduce.Job: Counters: 56
        File System Counters
                FILE: Number of bytes read=570
                FILE: Number of bytes written=47462873
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=25347
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=661
                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
```



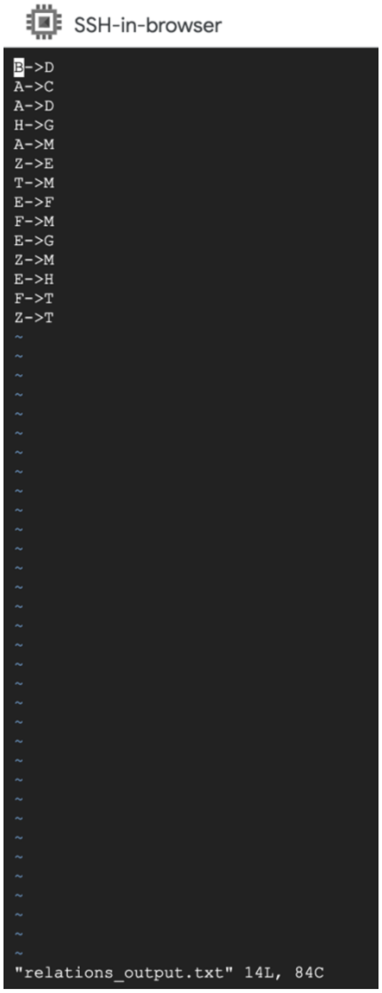
```
                Data-local map tasks=8
                Rack-local map tasks=134
                Total time spent by all maps in occupied slots (ms)=3421216
                Total time spent by all reduces in occupied slots (ms)=540896
                Total time spent by all map tasks (ms)=855304
                Total time spent by all reduce tasks (ms)=135224
                Total vcore-milliseconds taken by all map tasks=855304
                Total vcore-milliseconds taken by all reduce tasks=135224
                Total megabyte-milliseconds taken by all map tasks=3503325184
                Total megabyte-milliseconds taken by all reduce tasks=553877504
        Map-Reduce Framework
                Map input records=18
                Map output records=48
                Map output bytes=192
                Map output materialized bytes=40332
                Input split bytes=15194
                Combine input records=0
                Combine output records=0
                Reduce input groups=12
                Reduce shuffle bytes=40332
                Reduce input records=48
                Reduce output records=0
                Spilled Records=96
                Shuffled Maps =6674
                Failed Shuffles=0
                Merged Map outputs=6674
                GC time elapsed (ms)=24012
                CPU time spent (ms)=282060
                Physical memory (bytes) snapshot=120919519232
                Virtual memory (bytes) snapshot=907896946688
                Total committed heap usage (bytes)=135969898496
                Peak Map Physical memory (bytes)=743694336
                Peak Map Virtual memory (bytes)=4863442944
                Peak Reduce Physical memory (bytes)=442204160
                Peak Reduce Virtual memory (bytes)=4809789440
        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=10153
        File Output Format Counters
                Bytes Written=0
2023-10-07 02:50:00,491 INFO streaming.StreamJob: Output directory: outputpython
ys5250_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -get outputpython
```

Get the output file:



```
ys5250_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -get outputpython
```

Output:



```
B->D
A->C
A->D
H->G
A->M
Z->E
T->M
E->F
F->M
E->G
Z->M
E->H
F->T
Z->T
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
"relations_output.txt" 14L, 84C
```