

CSC 555: HW3

Name: Nachiketh Reddy

ID: 2117731

QUESTION 1:

Multi-node cluster setup Follow the instructions in the accompanying Word document to set up a 4 node cluster.

1. You should verify that the cluster is running by pointing your browser to the link below. [http://\[insert-the-public-ip-of-master\]:50070/](http://[insert-the-public-ip-of-master]:50070/) Make sure that the cluster is operational. You should see the 3 nodes under Datanodes tab. Submit a screenshot of your cluster status view.

Hadoop

Overview

Datanodes

Snapshot

Startup Progress

Utilities

<

- Repeat the steps for the wordcount example in HW1 and submit the screenshots of running it. Write a short paragraph with a discussion about how the results compare. Are they faster or slower? How much faster or slower?

```
time hadoop jar hadoop-2.6.4/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.6.4.jar wordcount /data/bioproject1.xml /data/wordcount1
```

```
aws Services Search [Alt+S]
wotime hadoop jar hadoop-2.6.4/share/hadoop/mwpreduce/hadoop-mapreduceexamples-2.6.4.jar
wotime hadoop jar hadoop-2.6.4/share/hadoop/mapreduce/hadoop-mapreduceexamples-2.6.4.jar
wohadoop fs -ls /data/^C
[ec2-user@ip-172-31-62-135 ~] 2024-02-23 23:13:22
$ time hadoop jar hadoop-2.6.4/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.6.4.jar wordcount /data/biop
24/02/23 23:13:30 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
24/02/23 23:13:30 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
24/02/23 23:13:30 INFO input.FileInputFormat: Total input paths to process : 1
24/02/23 23:13:30 INFO mapreduce.JobSubmitter: number of splits:2
24/02/23 23:13:31 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1031727439_0001
24/02/23 23:13:31 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
24/02/23 23:13:31 INFO mapreduce.Job: Running job: job_local1031727439_0001
24/02/23 23:13:31 INFO mapred.LocalJobRunner: OutputCommitter set in config null
24/02/23 23:13:31 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOut
24/02/23 23:13:31 INFO mapred.LocalJobRunner: Waiting for map tasks
24/02/23 23:13:31 INFO mapred.LocalJobRunner: Starting task: attempt_local1031727439_0001_m_000000_0
24/02/23 23:13:31 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
24/02/23 23:13:31 INFO mapred.MapTask: Processing split: hdfs://localhost/data/bioproject1.xml:0+134217728
24/02/23 23:13:31 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
24/02/23 23:13:31 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
24/02/23 23:13:31 INFO mapred.MapTask: soft limit at 83886080
24/02/23 23:13:31 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
24/02/23 23:13:31 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
24/02/23 23:13:31 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutput
24/02/23 23:13:32 INFO mapreduce.Job: Job job_local1031727439_0001 running in uber mode : false
24/02/23 23:13:32 INFO mapreduce.Job: map 0% reduce 0%
24/02/23 23:13:32 INFO mapred.MapTask: Spilling map output
24/02/23 23:13:32 INFO mapred.MapTask: bufstart = 0; bufend = 44123799; bufvoid = 104857600

i-062d65a3dd2e0d019 (nachiketh_server)
PublicIPs: 18.207.112.152 PrivateIPs: 172.31.62.135
```

```
aws Services Search [Alt+S]
Reduce input records=1182524
Reduce output records=1040558
Spilled Records=3856070
Shuffled Maps =2
Failed Shuffles=0
Merged Map outputs=2
GC time elapsed (ms)=99
CPU time spent (ms)=0
Physical memory (bytes) snapshot=0
Virtual memory (bytes) snapshot=0
Total committed heap usage (bytes)=1193279488
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=231154875
File Output Format Counters
Bytes Written=20057914
real    0m38.410s
user    0m44.455s
sys     0m0.952s
[ec2-user@ip-172-31-62-135 ~] 2024-02-23 23:14:07
$

i-062d65a3dd2e0d019 (nachiketh_server)
PublicIPs: 18.207.112.152 PrivateIPs: 172.31.62.135
```

aws

Services

Search

[Alt+S]

Saving to: 'bioproject1.xml'

bioproject1.xml

100%[-----]

2024-02-23 23:55:49 (70.9 MB/s) - 'bioproject1.xml' saved [231150779/231150779]

[ec2-user@ip-172-31-52-86 ~]\$

hadoop fs -put bioproject1.xml /data/

[ec2-user@ip-172-31-52-86 ~]\$

hadoop fs -ls /data/

Found 1 items

-rw-r--r-- 3 ec2-user supergroup 231150779 2024-02-23 23:56 /data/bioproject1.xml

[ec2-user@ip-172-31-52-86 ~]\$

time hadoop jar hadoop-2.6.4/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.6.4.jar wordcount /data/bioproject1.xml /data/wordcount1

24/02/23 23:56:57 INFO client.RMProxy: Connecting to ResourceManager at /172.31.52.86:8032

24/02/23 23:56:58 INFO input.FileInputFormat: Total input paths to process : 1

24/02/23 23:56:58 INFO mapreduce.JobSubmitter: number of splits:2

24/02/23 23:56:58 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1708730794501_0001

24/02/23 23:56:59 INFO impl.YarnClientImpl: Submitted application application_1708730794501_0001

24/02/23 23:56:59 INFO mapreduce.Job: The url to track the job: http://ip-172-31-52-86.ec2.internal:8088/proxy/application_1708730794501_0001/

24/02/23 23:56:59 INFO mapreduce.Job: Running job: job_1708730794501_0001

24/02/23 23:57:05 INFO mapreduce.Job: Job job_1708730794501_0001 running in uber mode : false

24/02/23 23:57:05 INFO mapreduce.Job: map 0% reduce 0%

24/02/23 23:57:18 INFO mapreduce.Job: map 26% reduce 0%

24/02/23 23:57:21 INFO mapreduce.Job: map 41% reduce 0%

24/02/23 23:57:24 INFO mapreduce.Job: map 47% reduce 0%

24/02/23 23:57:27 INFO mapreduce.Job: map 60% reduce 0%

24/02/23 23:57:29 INFO mapreduce.Job: map 77% reduce 0%

24/02/23 23:57:30 INFO mapreduce.Job: map 81% reduce 0%

24/02/23 23:57:33 INFO mapreduce.Job: map 83% reduce 0%

24/02/23 23:57:34 INFO mapreduce.Job: map 100% reduce 0%

24/02/23 23:57:37 INFO mapreduce.Job: map 100% reduce 100%

24/02/23 23:57:38 INFO mapreduce.Job: Job job_1708730794501_0001 completed successfully

24/02/23 23:57:38 INFO mapreduce.Job: Counters: 49

File System Counters

FILE: Number of bytes read=59610589

FILE: Number of bytes written=86837036

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=231155085

HDFS: Number of bytes written=20057914

HDFS: Number of read operations=9

HDFS: Number of large read operations=0

HDFS: Number of write operations=2

Job Counters

Launched map tasks=2

Launched reduce tasks=1

Data-local map tasks=2

Total time spent by all maps in occupied slots (ms)=193244

Total time spent by all reduces in occupied slots (ms)=23636

Total time spent by all map tasks (ms)=48311

Total time spent by all reduce tasks (ms)=5909

Total vcore-milliseconds taken by all map tasks=48311

Total vcore-milliseconds taken by all reduce tasks=5909

Total megabyte-milliseconds taken by all map tasks=193244000

Total megabyte-milliseconds taken by all reduce tasks=23636000

Map-Reduce Framework

Map input records=5284641

Map output records=18562590

Map output bytes=279359352

Map output materialized bytes=26905064

Input split bytes=210

Combine input records=20053612

Combine output records=2673546

Reduce input groups=1040558

Reduce shuffle bytes=26905064

Reduce input records=1182524

Reduce output records=1040558

Spilled Records=3856070

Shuffled Maps =2

Failed Shuffles=0

Merged Map outputs=2

GC time elapsed (ms)=407

CPU time spent (ms)=41480

Physical memory (bytes) snapshot=1617174528

Virtual memory (bytes) snapshot=9983971328

Total committed heap usage (bytes)=1462763520

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=231154875

File Output Format Counters

Bytes Written=20057914

real 0m42.693s

user 0m3.947s

sys 0m0.272s

[ec2-user@ip-172-31-52-86 ~]\$

i-01f594425b534e2ef (Master)

PublicIPs: 100.26.241.175 PrivateIPs: 172.31.52.86

SINGLE NODE	MULTI NODE
real 0m38.410s	real 0m42.629s
user 0m44.455s	user 0m3.947s
sys 0m0.952s	sys 0m0.272s

When the wordcount script was executed on the bioproject.xml file using single-node and multi-node setups, the single-node setup finished the operation in 38.410 seconds, with a user time of 44.455 seconds. The multi-node arrangement, on the other hand, finished in 42.629 seconds, although with a much shorter user time (3.947 seconds). The discrepancy implies that even though the multi-node configuration had a minor runtime overhead overall, it effectively used CPU resources in user mode, most likely as a result of parallel processing over several nodes. Regardless of the number of nodes deployed, both setups displayed low system time, showing constant overhead linked to system operations.

3. Run all the queries implemented in HW2. Submit the timings for each query in the form of a table and compare it with the timing using a single node cluster. Here is a sample table:

QUERY	SINGLE NODE	MULTI NODE
SELECT lo_quantity, lo_linenummer FROM lineorder WHERE lo_discount < 10 AND lo_tax > 2	real 0m26.453s user 0m33.889s sys 0m3.267s	real 0m30.124s user 0m3.884s sys 0m0.245s
SELECT p_category, COUNT(p_type) FROM part GROUP BY p_category	real 0m5.508s user 0m7.298s sys 0m0.483s	real 0m20.967s user 0m3.585s sys 0m0.330s
SELECT lo_discount, AVG(lo_extendedprice) FROM lineorder GROUP BY lo_discount	real 0m35.510s user 0m43.745s sys 0m3.915s	real 0m31.081s user 0m3.679s sys 0m0.343s
SELECT lo_custkey, SUM(lo_extendedprice) AS revenue FROM lineorder WHERE lo_quantity < 12 GROUP BY lo custkey	real 0m18.552s user 0m22.576s sys 0m2.064s	real 0m23.018s user 0m3.787s sys 0m0.232s
SELECT s_suppkey FROM supplier MINUS SELECT lo_suppkey FROM lineorder WHERE lo_discount < 10	real 0m30.550s user 0m38.458s sys 0m3.300s	real 0m28.030s user 0m3.865s sys 0m0.300s

Several trends become apparent when comparing the query execution speeds of single-node and multi-node systems. The single-node arrangement performs better than the multi-node setup for queries containing complex filtering conditions such as tax and discount thresholds, however it has greater system and user times. However, even with slightly longer actual times, smaller queries requiring simple aggregations such as averaging or counting benefit from parallel processing in multi-node settings. Because there is less overhead involved in coordinating operations among nodes, the single-node arrangement generally exhibits faster actual timings. On the other hand, multi-node configurations show benefits in terms of user and system times, exhibiting better resource management and parallelism for certain kinds of queries.

SINGLE NODE:

QUERY 1:

```
SELECT lo_quantity, lo_linenumber FROM lineorder WHERE lo_discount < 10 AND lo_tax > 2
```

[illegible]

```

24/02/23 22:08:31 INFO streaming.PipelineBdd: R/W/S=1300000/1253241/0 in:650000=1300000/2 [rec/s] out:626622=1253245/2 [rec/s]
24/02/23 22:08:31 INFO streaming.PipelineBdd: R/W/S=1400000/1438702/0 in:700000=1400000/2 [rec/s] out:681183=143796/2 [rec/s]
24/02/23 22:08:31 INFO streaming.PipelineBdd: R/W/S=1500000/1468233/0 in:750000=1500000/2 [rec/s] out:734119=1468238/2 [rec/s]
24/02/23 22:08:31 INFO streaming.PipelineBdd: R/W/S=1600000/1856554/0 in:800000=1600000/2 [rec/s] out:778280=1856561/2 [rec/s]
24/02/23 22:08:31 INFO streaming.PipelineBdd: R/W/S=1700000/1658927/0 in:850000=1700000/2 [rec/s] out:823166=1658933/2 [rec/s]
24/02/23 22:08:31 INFO streaming.PipelineBdd: R/W/S=1800000/1762752/0 in:900000=1800000/2 [rec/s] out:881378=1762757/2 [rec/s]
24/02/23 22:08:32 INFO streaming.PipelineBdd: R/W/S=1900000/1855253/0 in:950000=1900000/2 [rec/s] out:927632=1855264/2 [rec/s]
24/02/23 22:08:32 INFO streaming.PipelineBdd: R/W/S=2000000/1955604/0 in:1000000=2000000/2 [rec/s] out:979715=1955610/2 [rec/s]
24/02/23 22:08:32 INFO streaming.PipelineBdd: R/W/S=2100000/2068984/0 in:1050000=2100000/2 [rec/s] out:1039453=2068989/2 [rec/s]
24/02/23 22:08:32 INFO streaming.PipelineBdd: R/W/S=2200000/2161646/0 in:1100000=2200000/2 [rec/s] out:1099950=2161652/2 [rec/s]
24/02/23 22:08:32 INFO streaming.PipelineBdd: R/W/S=2300000/2266819/0 in:1150000=2300000/2 [rec/s] out:1158507=2266821/2 [rec/s]
24/02/23 22:08:32 INFO streaming.PipelineBdd: R/W/S=2400000/2361795/0 in:1200000=2400000/2 [rec/s] out:1217267=2361801/2 [rec/s]
24/02/23 22:08:32 INFO streaming.PipelineBdd: R/W/S=2500000/2460632/0 in:1250000=2500000/2 [rec/s] out:1276027=2460638/2 [rec/s]
24/02/23 22:08:33 INFO streaming.PipelineBdd: R/W/S=2600000/2566840/0 in:1300000=2600000/2 [rec/s] out:1335151=2566846/2 [rec/s]
24/02/23 22:08:33 INFO streaming.PipelineBdd: R/W/S=2700000/2659483/0 in:1350000=2700000/2 [rec/s] out:1394496=2659489/2 [rec/s]
24/02/23 22:08:33 INFO streaming.PipelineBdd: R/W/S=2800000/2761493/0 in:1400000=2800000/2 [rec/s] out:1454049=2761499/2 [rec/s]
24/02/23 22:08:33 INFO streaming.PipelineBdd: R/W/S=2900000/2871153/0 in:1450000=2900000/2 [rec/s] out:1513789=2871159/2 [rec/s]
24/02/23 22:08:33 INFO streaming.PipelineBdd: R/W/S=3000000/2964368/0 in:1500000=3000000/2 [rec/s] out:1573433=2964373/2 [rec/s]
24/02/23 22:08:33 INFO streaming.PipelineBdd: R/W/S=3100000/3063958/0 in:1550000=3100000/2 [rec/s] out:1633191=3063964/2 [rec/s]
24/02/23 22:08:33 INFO streaming.PipelineBdd: R/W/S=3200000/3158574/0 in:1600000=3200000/2 [rec/s] out:1693046=3158580/2 [rec/s]
24/02/23 22:08:33 INFO streaming.PipelineBdd: R/W/S=3300000/3253935/0 in:1650000=3300000/2 [rec/s] out:1752944=3253963/2 [rec/s]
24/02/23 22:08:33 INFO streaming.PipelineBdd: R/W/S=3400000/3352582/0 in:1700000=3400000/2 [rec/s] out:1813474=3352588/2 [rec/s]
24/02/23 22:08:34 INFO streaming.PipelineBdd: R/W/S=3500000/3451415/0 in:1750000=3500000/2 [rec/s] out:1874055=3451421/2 [rec/s]
24/02/23 22:08:34 INFO HadoopLocalJobRunner: Records R/W=218461 /> reduce
24/02/23 22:08:34 INFO streaming.PipelineBdd: R/W/S=3600000/3561550/0 in:1800000=3600000/2 [rec/s] out:1890389=3561556/2 [rec/s]
24/02/23 22:08:34 INFO streaming.PipelineBdd: HMRThread done
24/02/23 22:08:34 INFO streaming.PipelineBdd: mapSpdFinished
24/02/23 22:08:34 INFO HadoopTask: Task:attempt_local634139787_0001_r_000000_0 is done. And is in the process of committing
24/02/23 22:08:34 INFO HadoopTask: Task:attempt_local634139787_0001_r_000000_0 is allowed to commit now
24/02/23 22:08:34 INFO output.FileOutputCommitter: Saved output of task 'attempt_local634139787_0001_r_000000_0' to hdfs://localhost/data/output2/_temporary/0/task_local634139787_0001_r_000000
24/02/23 22:08:34 INFO HadoopTask: Task:attempt_local634139787_0001_r_000000_0 done.
24/02/23 22:08:34 INFO HadoopLocalJobRunner: Finishing task: attempt_local634139787_0001_r_000000_0
24/02/23 22:08:34 INFO HadoopLocalJobRunner: reduce task executor complete.
24/02/23 22:08:34 INFO HadoopReduceJob: map 100% reduce 100%
24/02/23 22:08:34 INFO HadoopReduceJob: Job job_local634139787_0001 completed successfully
24/02/23 22:08:34 INFO HadoopReduceJob: Counters: 38
File System Counters
FILE: Number of bytes read=56917628
FILE: Number of bytes written=151291229
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=210877010
HDFS: Number of bytes written=21174324
HDFS: Number of read operations=0
HDFS: Number of large read operations=0
HDFS: Number of write operations=8
Map-Reduce Framework
Map input records=6001215
Map output records=3638030
Map output bytes=2174297
Map output materialized bytes=28450387
Input split bytes=435
Combine input records=0
Combine output records=0
Reduce input groups=350
Reduce shuffle bytes=28450387
Reduce input records=3638030
Reduce output records=3638031
Spilled Records=727600
Shuffled Maps=5
Failed Shuffles=0
Merged Map outputs=0
CPU time elapsed (ms)=127
CPU time spent (ms)=0
Physical memory (bytes) snapshot=0
Virtual memory (bytes) snapshot=0
Total committed heap usage (bytes)=2601517056
Shuffle
Error
RAN ID=0
CONNECTION=0
IO SKIPPED
HRCNG_HRCNTH=0
HRCNG_MAP=0
HRCNG_REDUCE=0
File Input Format Counters
Bytes Read=594323985
File Output Format Counters
Bytes Written=21174324
24/02/23 22:08:34 INFO streaming.StreamJob: Output directory: /data/output2
real 0m26.453s
user 0m33.889s
sys 0m3.267s
[ec2-user@ip-172-31-62-135 CSC_555_HM2] 2024-02-23 22:08:34
$ hadoop fs -cat /data/output2/part-00000 | less
[31] Stopped hadoop fs -cat /data/output2/part-00000 | less
[ec2-user@ip-172-31-62-135 CSC_555_HM2] 2024-02-23 22:13:40
$ nano mapper.py
i-O62d65a3d2e0d019 (nachiketh_server)
PublicPis: 18.207.112.152 PrivatePis: 172.31.62.135

```



```
SELECT p_category, COUNT(p_type) FROM part GROUP BY p_category
```

```

24/02/23 22:19:49 INFO reduce.InMemoryMapOutput: Read 651948 bytes from map-output for attempt_local1930491341_0001_r_000000_0
24/02/23 22:19:49 INFO reduce.MergeManagerImpl: CloseMemoryFile -> map-output of size: 651948, InMemoryMapOutputs.size() -> 1, commitMemory -> 0, useMemory -> 651948
24/02/23 22:19:49 INFO reduce.EventFetcher: EventFetcher is interrupted. Returning
24/02/23 22:19:49 INFO mapred.LocalJobRunner: 1 / 1 copied.
24/02/23 22:19:49 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs
24/02/23 22:19:49 INFO mapred.Merger: Merging 1 sorted segments
24/02/23 22:19:49 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 651991 bytes
24/02/23 22:19:49 INFO reduce.MergeManagerImpl: Merged 1 segments, 651948 bytes to disk to satisfy reduce memory limit
24/02/23 22:19:49 INFO reduce.MergeManagerImpl: Merging 1 files, 651952 bytes from disk
24/02/23 22:19:49 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
24/02/23 22:19:49 INFO mapred.Merger: Merging 1 sorted segments
24/02/23 22:19:49 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 651991 bytes
24/02/23 22:19:49 INFO mapred.LocalJobRunner: 1 / 1 copied.
24/02/23 22:19:49 INFO streaming.PipelineMapRed: PipelineMapRed exec (/home/ec2-user/CSC_555_9M2//rddhcar.py)
24/02/23 22:19:49 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.job.tracker.address
24/02/23 22:19:49 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.map
24/02/23 22:19:49 INFO mapreduce.Job: map 100% reduce 0%
24/02/23 22:19:49 INFO streaming.PipelineMapRed: R/W/S10/O/O in:NA (rec/s) out:NA (rec/s)
24/02/23 22:19:49 INFO streaming.PipelineMapRed: R/W/S10/O/O in:NA (rec/s) out:NA (rec/s)
24/02/23 22:19:49 INFO streaming.PipelineMapRed: R/W/S100/O/O in:NA (rec/s) out:NA (rec/s)
24/02/23 22:19:49 INFO streaming.PipelineMapRed: R/W/S1000/O/O in:NA (rec/s) out:NA (rec/s)
24/02/23 22:19:49 INFO streaming.PipelineMapRed: R/W/S10000/O/O in:NA (rec/s) out:NA (rec/s)
24/02/23 22:19:49 INFO streaming.PipelineMapRed: R/W/S100000/O/O in:NA (rec/s) out:NA (rec/s)
24/02/23 22:19:49 INFO streaming.PipelineMapRed: R/W/S200000/O/O in:NA (rec/s) out:NA (rec/s)
24/02/23 22:19:49 INFO streaming.PipelineMapRed: Records R/W=200000/1
24/02/23 22:19:49 INFO streaming.PipelineMapRed: WRTerrThread done
24/02/23 22:19:49 INFO streaming.PipelineMapRed: mapRedFinished
24/02/23 22:19:49 INFO mapred.Task: Task attempt_local1930491341_0001_r_000000_0 is done. And is in the process of committing
24/02/23 22:19:49 INFO mapred.LocalJobRunner: 1 / 1 copied.
24/02/23 22:19:49 INFO mapred.Task: Task attempt_local1930491341_0001_r_000000_0 is allowed to commit now
24/02/23 22:19:49 INFO FileOutputCommitter: Saved output of task 'attempt_local1930491341_0001_r_000000_0' to hdfs://localhost/data/output3/_temporary/0/task_local1930491341_0001_r_000000
24/02/23 22:19:49 INFO mapred.LocalJobRunner: Records R/W=200000/1 reduce
24/02/23 22:19:49 INFO mapred.Task: Task 'attempt_local1930491341_0001_r_000000_0' done.
24/02/23 22:19:49 INFO mapred.LocalJobRunner: Finishing task: attempt_local1930491341_0001_r_000000_0
24/02/23 22:19:49 INFO mapred.LocalJobRunner: reduce task executor complete.
24/02/23 22:19:50 INFO mapreduce.Job: map 100% reduce 100%
24/02/23 22:19:50 INFO mapreduce.Job: Job job_local1930491341_0001 completed successfully
24/02/23 22:19:50 INFO mapreduce.Job: Counters: 38
File System Counters
  FILE: Number of bytes read=13042828
  FILE: Number of bytes written=2003682
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=34718519
  HDFS: Number of bytes written=376
  HDFS: Number of read operations=19
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=4
Map-Reduce Framework
  Map input records=200000
  Map output records=200000
  Map output bytes=611946
  Map output materialized bytes=651952
  Input split bytes=2
  Combine input records=0
  Combine output records=0
  Reduce input groups=950
  Reduce shuffle bytes=651952
  Reduce input records=200000
  Reduce output records=26
  Spilled Records=400000
  Shuffled Hops=41
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=10
  CPU time spent (ms)=0
  Physical memory (bytes) memspah=0
  Virtual memory (bytes) memspah=0
  Total committed heap usage (bytes)=597689820
Shuffle Error:
  BAD ID=0
  CORRUPTID=0
  IO EXCEPTION=0
  MRCMG_IDR2TH=0
  MRCMG_M4MD=0
  MRCMG_SORTED=0
File Input Format Counters
  Bytes Read=1193250
File Output Format Counters
  Bytes Written=376
24/02/23 22:19:50 INFO streaming.StreamJob: Output directory: /data/output3

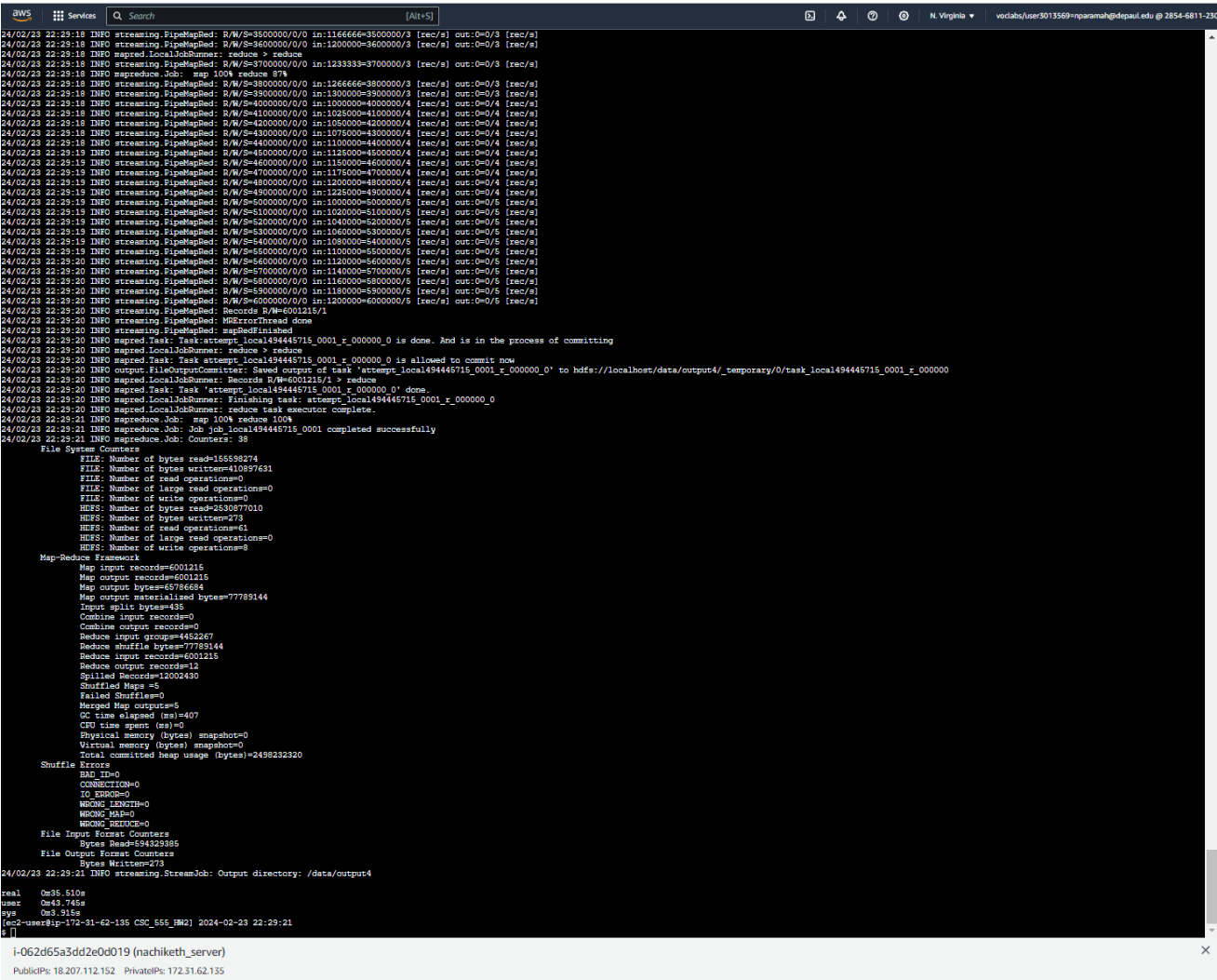
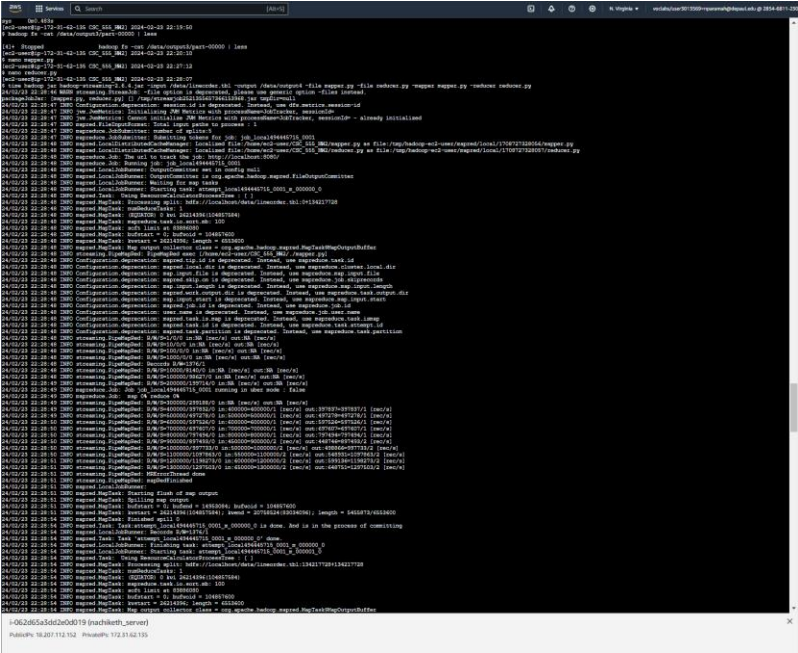
real    0m5.506s
user    0m7.568s
sys     0m0.488s
[ec2-user@ip-172-31-62-135 CSC_555_9M2] # 2024-02-23 22:19:50
$ hadoop fs -cat /data/output3/part-000001 | less

[4] Stopped
hadoop fs -cat /data/output3/part-000001 | less
[ec2-user@ip-172-31-62-135 CSC_555_9M2] # 2024-02-23 22:20:10
$]]

```

QUERY 3:

SELECT lo_discount, AVG(lo_extendedprice) FROM lineorder GROUP BY lo_discount



```
SELECT lo_custkey, SUM(lo_extendedprice) AS revenue FROM lineorder WHERE lo_quantity < 12 GROUP BY lo_custkey
```

```

24/02/23 22:36:12 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 21149794 bytes
24/02/23 22:36:12 INFO mapred.LocalJobRunner: 5 / 5 copied.
24/02/23 22:36:12 INFO streaming.PipelineMaped: PipelineMaped exec: [/home/ec2-user/CSC_555_IBM2/reducer.py]
24/02/23 22:36:12 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
24/02/23 22:36:12 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
24/02/23 22:36:12 INFO streaming.PipelineMaped: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/23 22:36:12 INFO streaming.PipelineMaped: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/23 22:36:12 INFO streaming.PipelineMaped: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/23 22:36:12 INFO streaming.PipelineMaped: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/23 22:36:12 INFO streaming.PipelineMaped: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/23 22:36:12 INFO streaming.PipelineMaped: R/W/S=100000/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/23 22:36:13 INFO mapreduce.Job: map 100A reduce 0A
24/02/23 22:36:13 INFO streaming.PipelineMaped: R/W/S=200000/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/23 22:36:13 INFO streaming.PipelineMaped: R/W/S=300000/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/23 22:36:13 INFO streaming.PipelineMaped: R/W/S=400000/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/23 22:36:13 INFO streaming.PipelineMaped: R/W/S=500000/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/23 22:36:13 INFO streaming.PipelineMaped: R/W/S=600000/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/23 22:36:13 INFO streaming.PipelineMaped: R/W/S=700000/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/23 22:36:13 INFO streaming.PipelineMaped: R/W/S=800000/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/23 22:36:13 INFO streaming.PipelineMaped: R/W/S=900000/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/23 22:36:13 INFO streaming.PipelineMaped: R/W/S=1000000/0/0 in:1000000-1000000/1 [rec/s] out:0=0/1 [rec/s]
24/02/23 22:36:13 INFO streaming.PipelineMaped: R/W/S=1100000/0/0 in:1100000-1100000/1 [rec/s] out:0=0/1 [rec/s]
24/02/23 22:36:13 INFO streaming.PipelineMaped: R/W/S=1200000/0/0 in:1200000-1200000/1 [rec/s] out:0=0/1 [rec/s]
24/02/23 22:36:14 INFO streaming.PipelineMaped: R/W/S=1300000/0/0 in:1300000-1300000/1 [rec/s] out:0=0/1 [rec/s]
24/02/23 22:36:14 INFO streaming.PipelineMaped: Records R/W=1318492/1
24/02/23 22:36:14 INFO streaming.PipelineMaped: MRErrorThread done
24/02/23 22:36:14 INFO streaming.PipelineMaped: mapRedFinished
24/02/23 22:36:14 INFO mapred.Task: Task:attempt_local1126028106_0001_r_000000_0 is done. And is in the process of committing
24/02/23 22:36:14 INFO mapred.LocalJobRunner: 5 / 5 copied.
24/02/23 22:36:14 INFO mapred.Task: Task:attempt_local1126028106_0001_r_000000_0 is allowed to commit now
24/02/23 22:36:14 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1126028106_0001_r_000000_0' to hdfs://localhost/data/output/s/_temporary/O/task_local1126028106_0001_r_000000
24/02/23 22:36:14 INFO mapred.LocalJobRunner: Records R/W=1318492/1 reduce
24/02/23 22:36:14 INFO mapred.Task: Task:attempt_local1126028106_0001_r_000000_0 done.
24/02/23 22:36:14 INFO mapred.LocalJobRunner: Finishing task: attempt_local1126028106_0001_r_000000_0
24/02/23 22:36:14 INFO mapred.LocalJobRunner: reduce task executor complete.
24/02/23 22:36:15 INFO mapreduce.Job: map 100A reduce 100A
24/02/23 22:36:15 INFO mapreduce.Job: Job job_local1126028106_0001 completed successfully
24/02/23 22:36:15 INFO mapreduce.Job: Counters: 38

File System Counters
  FILE: Number of bytes read=42319542
  FILE: Number of bytes written=12867065
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=2530977010
  HDFS: Number of bytes written=313156
  HDFS: Number of read operations=61
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=8

Map-Reduce Framework
  Map input records=6001215
  Map output records=1318492
  Map output bytes=1851284
  Map output materialized bytes=21149838
  Input split bytes=435
  Combine input records=0
  Combine output records=0
  Reduce input groups=1318270
  Reduce shuffle bytes=21149838
  Reduce input records=1318492
  Reduce output records=20001
  Spilled Records=2636984
  Shuffled Maps=5
  Failed Shuffles=0
  Merged Map outputs=5
  CPU time elapsed (ms)=203
  CPU time spent (ms)=0
  Physical memory (bytes) mapped=0
  Virtual memory (bytes) mapped=0
  Total committed heap usage (bytes)=262668280

Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDTYPE=0

File Input Format Counters
  Bytes Read=594329385

File Output Format Counters
  Bytes Written=313156

24/02/23 22:36:15 INFO streaming.StreamJob: Output directory: /data/output5

real    0m18.552s
user    0m22.576s
sys     0m2.046s

[ec2-user@ip-172-31-62-135 CSC_555_IBM2] 2024-02-23 22:36:15
$ hadoop fs -cat /data/output5/part-00000 | less

[6] + Stopped          hadoop fs -cat /data/output5/part-00000 | less
[ec2-user@ip-172-31-62-135 CSC_555_IBM2] 2024-02-23 22:36:32
$ []

```



```
SELECT s_suppkey FROM supplier MINUS SELECT lo_suppkey FROM lineorder WHERE lo_discount < 10
```

```

24/02/23 22:42:49 INFO streaming.PipelineMaped: R/W#=3100000/O/O in:1550000-3100000/2 [rec/s] out:0=0/2 [rec/s]
24/02/23 22:42:49 INFO streaming.PipelineMaped: R/W#=3200000/O/O in:1600000-3200000/2 [rec/s] out:0=0/2 [rec/s]
24/02/23 22:42:49 INFO streaming.PipelineMaped: R/W#=3300000/O/O in:1650000-3300000/2 [rec/s] out:0=0/2 [rec/s]
24/02/23 22:42:49 INFO streaming.PipelineMaped: R/W#=3400000/O/O in:1700000-3400000/2 [rec/s] out:0=0/2 [rec/s]
24/02/23 22:42:49 INFO streaming.PipelineMaped: R/W#=3500000/O/O in:1750000-3500000/2 [rec/s] out:0=0/2 [rec/s]
24/02/23 22:42:49 INFO streaming.PipelineMaped: R/W#=3600000/O/O in:1800000-3600000/2 [rec/s] out:0=0/2 [rec/s]
24/02/23 22:42:49 INFO streaming.PipelineMaped: R/W#=3700000/O/O in:1850000-3700000/2 [rec/s] out:0=0/2 [rec/s]
24/02/23 22:42:49 INFO streaming.PipelineMaped: R/W#=3800000/O/O in:1900000-3800000/2 [rec/s] out:0=0/2 [rec/s]
24/02/23 22:42:49 INFO streaming.PipelineMaped: R/W#=3900000/O/O in:1950000-3900000/2 [rec/s] out:0=0/2 [rec/s]
24/02/23 22:42:49 INFO streaming.PipelineMaped: R/W#=4000000/O/O in:2000000-4000000/2 [rec/s] out:0=0/2 [rec/s]
24/02/23 22:42:49 INFO streaming.PipelineMaped: R/W#=4100000/O/O in:2050000-4100000/2 [rec/s] out:0=0/2 [rec/s]
24/02/23 22:42:49 INFO streaming.PipelineMaped: R/W#=4200000/O/O in:2100000-4200000/2 [rec/s] out:0=0/2 [rec/s]
24/02/23 22:42:50 INFO streaming.PipelineMaped: R/W#=4300000/O/O in:2150000-4300000/2 [rec/s] out:0=0/2 [rec/s]
24/02/23 22:42:50 INFO streaming.PipelineMaped: R/W#=4400000/O/O in:2200000-4400000/2 [rec/s] out:0=0/2 [rec/s]
24/02/23 22:42:50 INFO streaming.PipelineMaped: R/W#=4500000/O/O in:2250000-4500000/2 [rec/s] out:0=0/2 [rec/s]
24/02/23 22:42:50 INFO streaming.PipelineMaped: R/W#=4600000/O/O in:2300000-4600000/2 [rec/s] out:0=0/2 [rec/s]
24/02/23 22:42:50 INFO streaming.PipelineMaped: R/W#=4700000/O/O in:1666666-4700000/3 [rec/s] out:0=0/3 [rec/s]
24/02/23 22:42:50 INFO streaming.PipelineMaped: R/W#=4800000/O/O in:1600000-4800000/3 [rec/s] out:0=0/3 [rec/s]
24/02/23 22:42:50 INFO streaming.PipelineMaped: R/W#=4900000/O/O in:1633333-4900000/3 [rec/s] out:0=0/3 [rec/s]
24/02/23 22:42:50 INFO streaming.PipelineMaped: R/W#=5000000/O/O in:1666666-5000000/3 [rec/s] out:0=0/3 [rec/s]
24/02/23 22:42:50 INFO streaming.PipelineMaped: R/W#=5100000/O/O in:1700000-5100000/3 [rec/s] out:0=0/3 [rec/s]
24/02/23 22:42:50 INFO streaming.PipelineMaped: R/W#=5200000/O/O in:1733333-5200000/3 [rec/s] out:0=0/3 [rec/s]
24/02/23 22:42:50 INFO streaming.PipelineMaped: R/W#=5300000/O/O in:1766666-5300000/3 [rec/s] out:0=0/3 [rec/s]
24/02/23 22:42:50 INFO streaming.PipelineMaped: R/W#=5400000/O/O in:1800000-5400000/3 [rec/s] out:0=0/3 [rec/s]
24/02/23 22:42:50 INFO streaming.PipelineMaped: Records R/W=5457400/1
24/02/23 22:42:50 INFO streaming.PipelineMaped: MRErrorThread done
24/02/23 22:42:50 INFO streaming.PipelineMaped: mapRedFinished
24/02/23 22:42:50 INFO mapred.Task: Task attempt_local12039987978_0001_r_000000_0 is done. And is in the process of committing
24/02/23 22:42:50 INFO mapred.Task: Task attempt_local12039987978_0001_r_000000_0 is allowed to commit now
24/02/23 22:42:50 INFO output.FileOutputCommitter: Saved output of task attempt_local12039987978_0001_r_000000_0 to hdfs://localhost/data/output6/_temporary/0/task_local12039987978_0001_r_000000
24/02/23 22:42:50 INFO mapred.Task: Task attempt_local12039987978_0001_r_000000_0 done.
24/02/23 22:42:50 INFO mapred.LocalJobRunner: Finishing task: attempt_local12039987978_0001_r_000000_0
24/02/23 22:42:50 INFO mapred.LocalJobRunner: reduce task executor complete.
24/02/23 22:42:51 INFO mapreduce.Job: map 100% reduce 100%
24/02/23 22:42:51 INFO mapreduce.Job: Job job_local12039987978_0001 completed successfully
24/02/23 22:42:51 INFO mapreduce.Job: Counters: 38

File System Counters
  FILE: Number of bytes read=103140653
  FILE: Number of bytes written=324626996
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=3125539747
  HDFS: Number of bytes written=17
  HDFS: Number of read operations=29
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=9

Map-Reduce framework
  Map input records=6003215
  Map output records=5457400
  Map output bytes=6441510
  Map output materialized bytes=51556346
  Input split bytes=521
  Combine input records=0
  Combine output records=0
  Reduce input groups=2000
  Reduce shuffle bytes=54556346
  Reduce input records=5457400
  Reduce output records=1
  Spilled Records=10914600
  Shuffled Maps=6
  Failed Shuffles=0
  Merged Map outputs=6
  CPU time elapsed (ms)=120
  CPU time spent (ms)=0
  Physical memory (bytes) memused=0
  Virtual memory (bytes) mapashtot=0
  Total committed heap usage (bytes)=3171942400

Shuffle Error
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0

File Input Format Counters
  Bytes Read=54456061

File Output Format Counters
  Bytes Written=17

24/02/23 22:42:51 INFO streaming.StreamJob: Output directory: /data/output6

real    0m30.550s
user    0m38.458s
sys      0m0.300s

ec2-user@ip-172-31-62-135 CSC_555 [RM2] 2024-02-23 22:42:51
$ hadoop fs -cat /data/output6/part-00000 | less

[7]+  Stopped                  hadoop fs -cat /data/output6/part-00000 | less
[ec2-user@ip-172-31-62-135 CSC_555 [RM2] 2024-02-23 22:43:22
$ !

```

MULTI NODE:

QUERY 1:

SELECT lo_quantity, lo_linenumbers FROM lineorder WHERE lo_discount < 10 AND lo_tax > 2

```
Found 5 items
-rw-r--r-- 3 ec2-user supergroup 231150779 2024-02-23 23:56 /data/bioprospect1.xml
-rw-r--r-- 3 ec2-user supergroup 594313001 2024-02-24 00:15 /data/lineorder.tbl
-rw-r--r-- 3 ec2-user supergroup 171332653 2024-02-24 00:16 /data/part.tbl
-rw-r--r-- 3 ec2-user supergroup 166676 2024-02-24 00:16 /data/supplier.tbl
dmesg -x -w      ec2-user supergroup      0 2024-02-23 23:57 /data/worldcount1

[ec2-user@ip-172-31-52-86 ~]$ nano mapper.py
[ec2-user@ip-172-31-52-86 ~]$ ls
bioproject1.xml  hadoop-2.6.4  hadoop-streaming-2.6.4.jar  lineorder.tbl  myHadoop.tar  part.tbl  supplier.tbl
[ec2-user@ip-172-31-52-86 ~]$ nano mapper.py
[ec2-user@ip-172-31-52-86 ~]$ nano reducer.py
[ec2-user@ip-172-31-52-86 ~]$ time hadoop jar hadoop-streaming-2.6.4.jar -input /data/lineorder.tbl -output /data/output1 -file mapper.py -file reducer.py -mapper mapper.py -reducer reducer.py
24/02/24 00:32:54 INFO streaming.StreamJob: file option is deprecated, please use generic option -files instead,
packageJobJar: [mapper.py, reducer.py, /tmp/hadoop-unjar6669915160572464663/] [] /tmp/streamjob8347225714561902567.jar tmpDir=null
24/02/24 00:32:55 INFO client.HMProxy: Connecting to ResourceManager at /172.31.52.86:8032
24/02/24 00:32:55 INFO client.HMProxy: Connecting to ResourceManager at /172.31.52.86:8032
24/02/24 00:32:56 INFO mapred.FileInputFormat: Total input paths to process : 1
24/02/24 00:32:56 INFO mapreduce.JobSubmitter: number of splits=5
24/02/24 00:32:56 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1708730794501_0002
24/02/24 00:32:56 INFO impl.YarnClientImpl: Submitted application application_1708730794501_0002
24/02/24 00:32:56 INFO mapreduce.Job: The url to track the job: http://ip-172-31-52-86.ec2.internal:8088/proxy/application_1708730794501_0002/
24/02/24 00:32:56 INFO mapreduce.Job: Running job: job_1708730794501_0002
24/02/24 00:33:01 INFO mapreduce.Job: Job job_1708730794501_0002 running in uber mode : false
24/02/24 00:33:01 INFO mapreduce.Job: map 0% reduce 0%
24/02/24 00:33:10 INFO mapreduce.Job: map 20% reduce 0%
24/02/24 00:33:11 INFO mapreduce.Job: map 40% reduce 0%
24/02/24 00:33:12 INFO mapreduce.Job: map 53% reduce 0%
24/02/24 00:33:13 INFO mapreduce.Job: map 80% reduce 0%
24/02/24 00:33:14 INFO mapreduce.Job: map 100% reduce 0%
24/02/24 00:33:21 INFO mapreduce.Job: map 100% reduce 93%
24/02/24 00:33:22 INFO mapreduce.Job: map 100% reduce 100%
24/02/24 00:33:23 INFO mapreduce.Job: Job job_1708730794501_0002 completed successfully
24/02/24 00:33:23 INFO mapreduce.Job: Counters: 50
File System Counters
  FILE: Number of bytes read=28450387
  FILE: Number of bytes written=5762657
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=594323835
  HDFS: Number of bytes written=21174324
  HDFS: Number of read operations=16
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=2
Job Counters
  Killed map tasks=1
  Launched map tasks=5
  Launched reduce tasks=1
  Data-local map tasks=5
  Total time spent by all maps in occupied slots (ms)=183620
  Total time spent by all reduces in occupied slots (ms)=38892
  Total time spent by all map tasks (ms)=5905
  Total time spent by all reduce tasks (ms)=9723
  Total vcore-millisecods taken by all map tasks=45905
  Total vcore-millisecods taken by all reduce tasks=9723
  Total megabyte-millisecods taken by all map tasks=183620000
  Total megabyte-millisecods taken by all reduce tasks=38892000
Map-Reduce Framework
  Map input records=6001215
  Map output records=3638030
  Map output bytes=21174327
  Map output materialized bytes=28450387
  Input split bytes=450
  Combine input records=0
  Combine output records=0
  Reduce input groups=350
  Reduce shuffle bytes=28450387
  Reduce input records=3638030
  Reduce output records=3638031
  Spilled Records=7276060
  Shuffled Maps =5
  Failed Shuffles=0
  Merged Map outputs=5
  GC time elapsed (ms)=688
  CPU time spent (ms)=26850
  Physical memory (bytes) snapshot=1698451456
  Virtual memory (bytes) snapshot=2145311948
  Total committed heap usage (bytes)=1458044928
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=594323835
File Output Format Counters
  Bytes Written=21174324
24/02/24 00:33:23 INFO streaming.StreamJob: Output directory: /data/output1

real    0m30.124s
user    0m3.884s
sys     0m0.246s
[ec2-user@ip-172-31-52-86 ~]$ ^C
[ec2-user@ip-172-31-52-86 ~]$ []
```

```
HEDFS: Number of large read operations=0
HEDFS: Number of write operations=2
Job Counters
  Killed map tasks=1
  Launched map tasks=5
  Launched reduce tasks=1
  Data-local map tasks=5
  Total time spent by all maps in occupied slots (ms)=183620
  Total time spent by all reduces in occupied slots (ms)=38892
  Total time spent by all map tasks (ms)=5905
  Total time spent by all reduce tasks (ms)=9723
  Total vcore-millisecods taken by all map tasks=45905
  Total vcore-millisecods taken by all reduce tasks=9723
  Total megabyte-millisecods taken by all map tasks=183620000
  Total megabyte-millisecods taken by all reduce tasks=38892000
Map-Reduce Framework
  Map input records=6001215
  Map output records=3638030
  Map output bytes=21174327
  Map output materialized bytes=28450387
  Input split bytes=450
  Combine input records=0
  Combine output records=0
  Reduce input groups=350
  Reduce shuffle bytes=28450387
  Reduce input records=3638030
  Reduce output records=3638031
  Spilled Records=7276060
  Shuffled Maps =5
  Failed Shuffles=0
  Merged Map outputs=5
  GC time elapsed (ms)=688
  CPU time spent (ms)=26850
  Physical memory (bytes) snapshot=1698451456
  Virtual memory (bytes) snapshot=2145311948
  Total committed heap usage (bytes)=1458044928
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=594323835
File Output Format Counters
  Bytes Written=21174324
24/02/24 00:33:23 INFO streaming.StreamJob: Output directory: /data/output1

real    0m30.124s
user    0m3.884s
sys     0m0.246s
[ec2-user@ip-172-31-52-86 ~]$ ^C
[ec2-user@ip-172-31-52-86 ~]$ []
```

QUERY 2:

SELECT p_category, COUNT(p_type) FROM part GROUP BY p_category

```
aws Services Q Search [Alt+S]
[ec2-user@ip-172-31-52-86 ~]$ time hadoop jar hadoop-streaming-2.6.4.jar -input /data/part.tbl -output /data/output2 -file mapper.py -file reducer.py -mapper mapper.py -reducer reducer.py
24/02/24 00:52:27 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
hadoopjob.jar: mapper.py, reducer.py, /tmp/hadoop-ubuntu4548506768848514417/1 /tmp/streamjob6394431473145320269.jar tmpDir=null
24/02/24 00:52:28 INFO client.RMProxy: Connecting to ResourceManager at /172.31.52.86:8032
24/02/24 00:52:28 INFO client.RMProxy: Connecting to ResourceManager at /172.31.52.86:8032
24/02/24 00:52:28 INFO mapred.FileInputFormat: Total input paths to process : 1
24/02/24 00:52:28 INFO mapreduce.JobSubmitter: number of splits:2
24/02/24 00:52:28 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1708730794501_0003
24/02/24 00:52:29 INFO impl.YarnClientImpl: Submitted application application_1708730794501_0003
24/02/24 00:52:29 INFO mapreduce.Job: The url to track the job: http://ip-172-31-52-86.ec2.internal:8088/proxy/application_1708730794501_0003/
24/02/24 00:52:29 INFO mapreduce.Job: Running job: job_1708730794501_0003
24/02/24 00:52:34 INFO mapreduce.Job: Job job_1708730794501_0003 running in uber mode : false
24/02/24 00:52:34 INFO mapreduce.Job: map 0% reduce 0%
24/02/24 00:52:41 INFO mapreduce.Job: map 100% reduce 0%
24/02/24 00:52:47 INFO mapreduce.Job: map 100% reduce 100%
24/02/24 00:52:47 INFO mapreduce.Job: Job job_1708730794501_0003 completed successfully
24/02/24 00:52:47 INFO mapreduce.Job: Counters: 49

File System Counters
  FILE: Number of bytes read=6519952
  FILE: Number of bytes written=13370821
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=17143535
  HDFS: Number of bytes written=376
  HDFS: Number of read operations=0
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=2

Job Counters
  Launched map tasks=2
  Launched reduce tasks=1
  Data-local map tasks=2
  Total time spent by all maps in occupied slots (ms)=41636
  Total time spent by all reduces in occupied slots (ms)=12192
  Total time spent by all map tasks (ms)=10409
  Total time spent by all reduce tasks (ms)=3048
  Total vcore-milliseconds taken by all map tasks=10409
  Total vcore-milliseconds taken by all reduce tasks=3048
  Total megabyte-milliseconds taken by all map tasks=41636000
  Total megabyte-milliseconds taken by all reduce tasks=12192000

Map-Reduce framework
  Map input records=200000
  Map output records=200000
  Map output bytes=6119946
  Map output materialized bytes=6519958
  Input split bytes=170
  Combine input records=0
  Combine output records=0
  Reduce input groups=3760
  Reduce shuffle bytes=6519958
  Reduce input records=200000
  Reduce output records=26
  Spilled Records=400000
  Shuffled Maps=2

i-01f594425b534e2ef (Master)
PublicDns: 100.26.241.175 PrivateDns: 172.31.52.86
```

```
aws Services Q Search [Alt+S]
Job Counters
  Launched map tasks=2
  Launched reduce tasks=1
  Data-local map tasks=2
  Total time spent by all maps in occupied slots (ms)=41636
  Total time spent by all reduces in occupied slots (ms)=12192
  Total time spent by all map tasks (ms)=10409
  Total time spent by all reduce tasks (ms)=3048
  Total vcore-milliseconds taken by all map tasks=10409
  Total vcore-milliseconds taken by all reduce tasks=3048
  Total megabyte-milliseconds taken by all map tasks=41636000
  Total megabyte-milliseconds taken by all reduce tasks=12192000

Map-Reduce framework
  Map input records=200000
  Map output records=200000
  Map output bytes=6119946
  Map output materialized bytes=6519958
  Input split bytes=170
  Combine input records=0
  Combine output records=0
  Reduce input groups=3760
  Reduce shuffle bytes=6519958
  Reduce input records=200000
  Reduce output records=26
  Spilled Records=400000
  Shuffled Maps=2
  Failed Shuffles=0
  Merged Map outputs=2
  GC time elapsed (ms)=244
  CPU time spent (ms)=4400
  Physical memory (bytes) snapshot=730370048
  Virtual memory (bytes) snapshot=9574772904
  Total committed heap usage (bytes)=588251136

Shuffle Errors
  BAD_IP=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0

File Input Format Counters
  Bytes Read=17143535
File Output Format Counters
  Bytes Written=376
24/02/24 00:52:47 INFO streaming.StreamJob: Output directory: /data/output2

real    0m20.967s
user    0m3.585s
sys     0m0.330s
[ec2-user@ip-172-31-52-86 ~]$ ^C
[ec2-user@ip-172-31-52-86 ~]$ hadoop fs -cat /data/output2/part-00000 | less
[2]+  Stopped                  hadoop fs -cat /data/output2/part-00000 | less
[ec2-user@ip-172-31-52-86 ~]$

i-01f594425b534e2ef (Master)
PublicDns: 100.26.241.175 PrivateDns: 172.31.52.86
```

QUERY 3:

SELECT lo_discount, AVG(lo_extendedprice) FROM lineorder GROUP BY lo_discount

```
24/02/24 01:15:09 INFO streaming.StreamJob: Output directory: /data/output3
[ec2-user@ip-172-31-52-86 ~]$ time hadoop jar hadoop-streaming-2.6.4.jar -input /data/lineorder.tbl -output /data/output31 -file mapper.py -file reducer.py -mapper mapper.py -reducer reducer.py
24/02/24 01:15:08 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead
packageJobJar: (mapper.py, reducer.py, /tmp/hadoop-unjar2247850230353385676/) [/tmp/streamjob4270551684335060842.jar tmpDir=null]
24/02/24 01:15:59 INFO client.RMProxy: Connecting to ResourceManager at /172.31.52.86:8032
24/02/24 01:15:59 INFO client.RMProxy: Connecting to ResourceManager at /172.31.52.86:8032
24/02/24 01:16:00 INFO mapred.FileInputFormat: Total input paths to process : 1
24/02/24 01:16:00 INFO mapreduce.JobSubmitter: number of splits:5
24/02/24 01:16:00 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1708736861460_0002
24/02/24 01:16:00 INFO mapreduce.Job: Submitted application application_1708736861460_0002
24/02/24 01:16:00 INFO mapreduce.Job: The url to track the job: http://ip-172-31-52-86.ec2.internal:8088/proxy/application_1708736861460_0002/
24/02/24 01:16:00 INFO mapreduce.Job: Running job: job_1708736861460_0002
24/02/24 01:16:05 INFO mapreduce.Job: Job job_1708736861460_0002 running in uber mode : false
24/02/24 01:16:05 INFO mapreduce.Job: map 0% reduce 0%
24/02/24 01:16:12 INFO mapreduce.Job: map 20% reduce 0%
24/02/24 01:16:15 INFO mapreduce.Job: map 60% reduce 0%
24/02/24 01:16:16 INFO mapreduce.Job: map 73% reduce 0%
24/02/24 01:16:17 INFO mapreduce.Job: map 87% reduce 0%
24/02/24 01:16:19 INFO mapreduce.Job: map 100% reduce 0%
24/02/24 01:16:23 INFO mapreduce.Job: map 100% reduce 74%
24/02/24 01:16:26 INFO mapreduce.Job: map 100% reduce 91%
24/02/24 01:16:28 INFO mapreduce.Job: map 100% reduce 100%
24/02/24 01:16:28 INFO mapreduce.Job: Job job_1708736861460_0002 completed successfully
24/02/24 01:16:29 INFO mapreduce.Job: Counters: 50

File System Counters
  FILE: Number of bytes read=77789120
  FILE: Number of bytes written=156240177
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=594323835
  HDFS: Number of bytes written=273
  HDFS: Number of read operations=18
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=2

Job Counters
  Killed map tasks=1
  Launched map tasks=6
  Launched reduce tasks=1
  Data-local map tasks=6
  Total time spent by all maps in occupied slots (ms)=178184
  Total time spent by all reduces in occupied slots (ms)=50652
  Total time spent by all map tasks (ms)=44546
  Total time spent by all reduce tasks (ms)=12663
  Total vcore-milliseconds taken by all map tasks=44546
  Total vcore-milliseconds taken by all reduce tasks=12663
  Total megabyte-milliseconds taken by all map tasks=178184000
  Total megabyte-milliseconds taken by all reduce tasks=50652000

Map-Reduce Framework
  Map input records=6001215
  Map output records=6001215
  Map output bytes=5786684
  Map output materialized bytes=77789144
  Input split bytes=450

i-01f594425b534e2ef (Master)
PublicIPs: 54.165.22.237 PrivateIPs: 172.31.52.86
```

```
Job Counters
  Killed map tasks=1
  Launched map tasks=6
  Launched reduce tasks=1
  Data-local map tasks=6
  Total time spent by all maps in occupied slots (ms)=178184
  Total time spent by all reduces in occupied slots (ms)=50652
  Total time spent by all map tasks (ms)=44546
  Total time spent by all reduce tasks (ms)=12663
  Total vcore-milliseconds taken by all map tasks=44546
  Total vcore-milliseconds taken by all reduce tasks=12663
  Total megabyte-milliseconds taken by all map tasks=178184000
  Total megabyte-milliseconds taken by all reduce tasks=50652000

Map-Reduce Framework
  Map input records=6001215
  Map output records=6001215
  Map output bytes=5786684
  Map output materialized bytes=77789144
  Input split bytes=450
  Combine input records=0
  Combine output records=0
  Reduce input groups=4452267
  Reduce shuffle bytes=77789144
  Reduce input records=6001215
  Reduce output records=12
  Spilled Record=12002430
  Shuffled Maps=6
  Failed Shuffles=0
  Merged Map outputs=5
  GC time elapsed (ms)=616
  CPU time spent (ms)=36250
  Physical memory (bytes) snapshot=1794957312
  Virtual memory (bytes) snapshot=21453033472
  Total committed heap usage (bytes)=1542979584

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=594323835

File Output Format Counters
  Bytes Written=273

24/02/24 01:16:29 INFO streaming.StreamJob: Output directory: /data/output31

real    0m31.081s
user    0m3.679s
sys     0m0.343s
[ec2-user@ip-172-31-52-86 ~]$ hadoop fs -cat /data/output31/part-00000 | less

[1]+  Stopped                  hadoop fs -cat /data/output31/part-00000 | less
[ec2-user@ip-172-31-52-86 ~]$

i-01f594425b534e2ef (Master)
PublicIPs: 54.165.22.237 PrivateIPs: 172.31.52.86
```

QUERY 4:

SELECT lo_custkey, SUM(lo_extendedprice) AS revenue FROM lineorder WHERE lo_quantity < 12 GROUP BY lo_custkey

```
aws Services Search [Alt+S]
[ec2-user@ip-172-31-52-86 ~]$ time hadoop jar hadoop-streaming-2.6.4.jar -input /data/lineorder.tbl -output /data/output43 -file mapper.py -file reducer.py -mapper mapper.py -reducer reducer.py
24/02/24 01:29:24 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
hadoop-streaming-2.6.4.jar: (mapper.py: /tmp/hadoop-mapreduce607321948745029539/) () /tmp/streamjob1981875098436741496.jar tmpDir=null
24/02/24 01:29:25 INFO client.RMProxy: Connecting to ResourceManager at /172.31.52.86:8032
24/02/24 01:29:25 INFO client.RMProxy: Connecting to ResourceManager at /172.31.52.86:8032
24/02/24 01:29:25 INFO mapreduce.JobSubmitter: Total input paths to process : 1
24/02/24 01:29:26 INFO mapreduce.JobSubmitter: number of splits=5
24/02/24 01:29:26 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1708736861460_0005
24/02/24 01:29:26 INFO impl.YarnClientImpl: Submitted application application_1708736861460_0005
24/02/24 01:29:26 INFO mapreduce.Job: The url to track the job: http://ip-172-31-52-86.ec2.internal:8088/proxy/application_1708736861460_0005/
24/02/24 01:29:26 INFO mapreduce.Job: Running job: job_1708736861460_0005
24/02/24 01:29:31 INFO mapreduce.Job: Job job_1708736861460_0005 running in uber mode : false
24/02/24 01:29:31 INFO mapreduce.Job: map 0% reduce 0%
24/02/24 01:29:37 INFO mapreduce.Job: map 20% reduce 0%
24/02/24 01:29:41 INFO mapreduce.Job: map 80% reduce 0%
24/02/24 01:29:42 INFO mapreduce.Job: map 100% reduce 0%
24/02/24 01:29:46 INFO mapreduce.Job: map 100% reduce 100%
24/02/24 01:29:46 INFO mapreduce.Job: Job job_1708736861460_0005 completed successfully
24/02/24 01:29:46 INFO mapreduce.Job: Counters: 51

File System Counters
  FILE: Number of bytes read=21149814
  FILE: Number of bytes written=12961565
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=594329385
  HDFS: Number of bytes written=313156
  HDFS: Number of read operations=18
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=2

Job Counters
  Killed map tasks=1
  Launched map tasks=5
  Launched reduce tasks=1
  Data-local map tasks=4
  Back-local map tasks=1
  Total time spent by all maps in occupied slots (ms)=151460
  Total time spent by all reduces in occupied slots (ms)=27936
  Total time spent by all map tasks (ms)=37865
  Total time spent by all reduce tasks (ms)=6984
  Total vcore-millisecsd taken by all map tasks=37865
  Total vcore-millisecsd taken by all reduce tasks=6984
  Total megabyte-millisecsd taken by all map tasks=151460000
  Total megabyte-millisecsd taken by all reduce tasks=27936000

Map-Reduce Framework
  Map input records=6001215
  Map output records=1318492
  Map output bytes=18512824
  Map output materialized bytes=21149838
  Input split bytes=450
  Combine input records=0
  Combine output records=0
  Reduce input groups=1318270
  Reduce shuffle bytes=21149838
  Reduce input records=1318492
  Reduce output records=20001
  Spilled Records=2636984
  Shuffled Maps =5
  Failed Shuffles=0
  Merged Map outputs=5
  GC time elapsed (ms)=573
  CPU time spent (ms)=16340
  Physical memory (bytes) snapshot=1640034304
  Virtual memory (bytes) snapshot=21464772608
  Total committed heap usage (bytes)=1355130368

Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDOCE=0

File Input Format Counters
  Bytes Read=594329385
File Output Format Counters
  Bytes Written=313156

24/02/24 01:29:46 INFO streaming.StreamJob: Output directory: /data/output43

real    0m23.018s
user    0m3.797s
sys     0m0.232s
[ec2-user@ip-172-31-52-86 ~]$ hadoop fs -cat /data/output43/part-00000 | less
[2]+  Stopped                  hadoop fs -cat /data/output43/part-00000 | less
[ec2-user@ip-172-31-52-86 ~]$
```

i-01f594425b534e2ef (Master)

PublicIPs: 54.165.22.237 PrivateIPs: 172.31.52.86

```
aws Services Search [Alt+S]
Killed map tasks=1
Launched map tasks=5
Launched reduce tasks=1
Data-local map tasks=4
Back-local map tasks=1
Total time spent by all maps in occupied slots (ms)=151460
Total time spent by all reduces in occupied slots (ms)=27936
Total time spent by all map tasks (ms)=37865
Total time spent by all reduce tasks (ms)=6984
Total vcore-millisecsd taken by all map tasks=37865
Total vcore-millisecsd taken by all reduce tasks=6984
Total megabyte-millisecsd taken by all map tasks=151460000
Total megabyte-millisecsd taken by all reduce tasks=27936000

Map-Reduce Framework
  Map input records=6001215
  Map output records=1318492
  Map output bytes=18512824
  Map output materialized bytes=21149838
  Input split bytes=450
  Combine input records=0
  Combine output records=0
  Reduce input groups=1318270
  Reduce shuffle bytes=21149838
  Reduce input records=1318492
  Reduce output records=20001
  Spilled Records=2636984
  Shuffled Maps =5
  Failed Shuffles=0
  Merged Map outputs=5
  GC time elapsed (ms)=573
  CPU time spent (ms)=16340
  Physical memory (bytes) snapshot=1640034304
  Virtual memory (bytes) snapshot=21464772608
  Total committed heap usage (bytes)=1355130368

Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDOCE=0

File Input Format Counters
  Bytes Read=594329385
File Output Format Counters
  Bytes Written=313156

24/02/24 01:29:46 INFO streaming.StreamJob: Output directory: /data/output43

real    0m23.018s
user    0m3.797s
sys     0m0.232s
[ec2-user@ip-172-31-52-86 ~]$ hadoop fs -cat /data/output43/part-00000 | less
[2]+  Stopped                  hadoop fs -cat /data/output43/part-00000 | less
[ec2-user@ip-172-31-52-86 ~]$
```

i-01f594425b534e2ef (Master)

PublicIPs: 54.165.22.237 PrivateIPs: 172.31.52.86

QUERY 5:

SELECT s_suppkey FROM supplier MINUS SELECT lo_suppkey FROM lineorder WHERE lo_discount < 10

```
AWSServicesSearch[Alt+S]N. Virginiavodabs/user3013569-nparamah@depaul.edu @ 2854-6811-230

[ec2-user@ip-172-31-52-86 ~]$ nano reducer.py
[ec2-user@ip-172-31-52-86 ~]$ time hadoop jar hadoop-streaming-2.6.4.jar -input /data/supplier.tbl -output /data/output51 -file mapper.py -file reducer.py -mapper mapper.py -reducer reducer.py
24/02/24 01:33:15 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packagelocaljar: [mapper.py, reducer.py, /tmp/hadoop-unjar609430387114019735/1 /tmp/streamjob728086341595472102.jar tmpDir=null
24/02/24 01:33:15 INFO client.RMProxy: Connecting to ResourceManager at /172.31.52.86:8032
24/02/24 01:33:16 INFO client.RMProxy: Connecting to ResourceManager at /172.31.52.86:8032
24/02/24 01:33:16 INFO mapred.FileInputFormat: Total input paths to process : 2
24/02/24 01:33:16 INFO mapreduce.JobSubmitter: number of splits=6
24/02/24 01:33:16 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1708736861460_0006
24/02/24 01:33:16 INFO impl.YarnClientImpl: Submitted application application_1708736861460_0006
24/02/24 01:33:16 INFO mapreduce.Job: The url to track the job: http://ip-172-31-52-86.ec2.internal:8088/proxy/application_1708736861460_0006/
24/02/24 01:33:16 INFO mapreduce.Job: Running job: job_1708736861460_0006
24/02/24 01:33:21 INFO mapreduce.Job: Job job_1708736861460_0006 running in uber mode : false
24/02/24 01:33:21 INFO mapreduce.Job: map 0% reduce 0%
24/02/24 01:33:28 INFO mapreduce.Job: map 17% reduce 0%
24/02/24 01:33:30 INFO mapreduce.Job: map 33% reduce 0%
24/02/24 01:33:33 INFO mapreduce.Job: map 76% reduce 0%
24/02/24 01:33:38 INFO mapreduce.Job: map 100% reduce 0%
24/02/24 01:33:39 INFO mapreduce.Job: map 100% reduce 73%
24/02/24 01:33:42 INFO mapreduce.Job: map 100% reduce 100%
24/02/24 01:33:42 INFO mapreduce.Job: Job job_1708736861460_0006 completed successfully
24/02/24 01:33:42 INFO mapreduce.Job: Counters: 51

File System Counters
  FILE: Number of bytes read=51556316
  FILE: Number of bytes written=103885162
  FILE: Number of read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=53454600
  HDFS: Number of bytes written=17
  HDFS: Number of read operations=21
  HDFS: Number of write operations=2
  HDFS: Number of large read operations=0
  HDFS: Number of large write operations=0

Job Counters
  Killed map tasks=1
  Launched map tasks=7
  Launched reduce tasks=1
  Data-local map tasks=6
  Rack-local map tasks=1
  Total time spent by all maps in occupied slots (ms)=225304
  Total time spent by all reduces in occupied slots (ms)=46588
  Total time spent by all map tasks (ms)=56326
  Total time spent by all reduce tasks (ms)=11647
  Total vcore-milliseconds taken by all map tasks=56326
  Total vcore-milliseconds taken by all reduce tasks=11647
  Total megabyte-milliseconds taken by all map tasks=225304000
  Total megabyte-milliseconds taken by all reduce tasks=46588000

Map-Reduce Framework
  Map input records=6003215
  Map output records=5457400
  Map output bytes=40641510
  Map output materialized bytes=51556346
  Input split bytes=539
  Combine input records=0

i-01f594425b534e2ef (Master)
PublicIPs: 54.165.22.237 PrivateIPs: 172.31.52.86
```

```
AWSServicesSearch[Alt+S]N. Virginiavodabs/user3013569-nparamah@depaul.edu @ 2854-6811-230

Killed map tasks=1
Launched map tasks=7
Launched reduce tasks=1
Data-local map tasks=6
Rack-local map tasks=1
Total time spent by all maps in occupied slots (ms)=225304
Total time spent by all reduces in occupied slots (ms)=46588
Total time spent by all map tasks (ms)=56326
Total time spent by all reduce tasks (ms)=11647
Total vcore-milliseconds taken by all map tasks=56326
Total vcore-milliseconds taken by all reduce tasks=11647
Total megabyte-milliseconds taken by all map tasks=225304000
Total megabyte-milliseconds taken by all reduce tasks=46588000

Map-Reduce Framework
  Map input records=6003215
  Map output records=5457400
  Map output bytes=40641510
  Map output materialized bytes=51556346
  Input split bytes=539
  Combine input records=0
  Combine output records=0
  Reduce input groups=2000
  Reduce shuffle bytes=51556346
  Reduce input records=5457400
  Reduce output records=1
  Spilled Records=10914800
  Shuffled Maps =6
  Failed Shuffles=0
  Merged Map outputs=6
  GC time elapsed (ms)=733
  CPU time spent (ms)=35030
  Physical memory (bytes) snapshot=1331816960
  Virtual memory (bytes) snapshot=26277317120
  Total committed heap usage (bytes)=1664090112

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=534496061
File Output Format Counters
  Bytes Written=17
24/02/24 01:33:42 INFO streaming.StreamJob: Output directory: /data/output51

real    0m28.090s
user    0m3.866s
sys     0m0.900s
[ec2-user@ip-172-31-52-86 ~]$ hadoop fs -cat /data/output51/part-00000 | less
[3]+  Stopped                  hadoop fs -cat /data/output51/part-00000 | less
[ec2-user@ip-172-31-52-86 ~]$

i-01f594425b534e2ef (Master)
PublicIPs: 54.165.22.237 PrivateIPs: 172.31.52.86
```

QUESTION 2:

Implement, execute, and time the following query using Hadoop streaming with python.

SELECT lo_quantity, MAX(lo_revenue)

FROM (

SELECT lo_revenue, MAX(lo_quantity) as lo_quantity, MAX(lo_discount) as lo_discount

FROM lineorder

WHERE lo_orderpriority LIKE '%URGENT'

GROUP BY lo_revenue)

WHERE lo_discount BETWEEN 5 AND 8

GROUP BY lo_quantity;

SINGLE NODE:

JOB 1:

time hadoop jar hadoop-streaming-2.6.4.jar -input /data/lineorder.tbl -output /data/job_output_1 -mapper mapper1.py -reducer reducer1.py
mapper1.py -reducer reducer1.py -file mapper1.py -file reducer1.py

```
BWS Services Q Search [Alt+S]
[ec2-user@ip-172-31-62-135 CSC 555 IM2] 2024-02-24 21:16:49
$ time hadoop jar hadoop-streaming-2.6.4.jar -input /data/lineorder.tbl -output /data/job_output_1 -mapper mapper1.py -reducer reducer1.py -file mapper1.py -file reducer1.py
24/02/24 21:17:23 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
24/02/24 21:17:24 INFO jvm.JvmMetrics: Cannot initialize Jvm Metrics with processName=JobTracker, sessionId=
24/02/24 21:17:24 INFO jvm.JvmMetrics: Initializing Jvm Metrics with processName=JobTracker, sessionId=
24/02/24 21:17:24 INFO mapred.FileInputFormat: Total input paths to process : 1
24/02/24 21:17:24 INFO mapreduce.JobSubmitter: number of splits:5
24/02/24 21:17:24 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local458332266_0001
24/02/24 21:17:25 INFO mapred.LocalDistributedCacheManager: Localized file:/home/ec2-user/CSC_555_IM2/mapper1.py as file:/tmp/hadoop-ec2-user/mapred/local/1708809444936/mapper1.py
24/02/24 21:17:25 INFO mapred.LocalDistributedCacheManager: Localized file:/home/ec2-user/CSC_555_IM2/reducer1.py as file:/tmp/hadoop-ec2-user/mapred/local/1708809444937/reducer1.py
24/02/24 21:17:25 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
24/02/24 21:17:25 INFO mapred.LocalJobRunner: Waiting for map tasks
24/02/24 21:17:25 INFO mapred.LocalJobRunner: OutputCommiter set in config null
24/02/24 21:17:25 INFO mapred.LocalJobRunner: OutputCommiter is org.apache.hadoop.mapred.FileOutputCommitter
24/02/24 21:17:25 INFO mapred.LocalJobRunner: Starting task: attempt_local458332266_0001_m_000000_0
24/02/24 21:17:25 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
24/02/24 21:17:25 INFO mapred.MapTask: Processing split: hdfs://localhost/data/lineorder.tbl:0:134217728
24/02/24 21:17:25 INFO mapred.MapTask: numReduceTasks: 1
24/02/24 21:17:25 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
24/02/24 21:17:25 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
24/02/24 21:17:25 INFO mapred.MapTask: soft limit at 48986000
24/02/24 21:17:25 INFO mapred.MapTask: hufstart = 0, hufvoid = 104857600
24/02/24 21:17:25 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
24/02/24 21:17:25 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.mapred.MapTaskMapOutputBuffer
24/02/24 21:17:25 INFO streaming.PipeMapRed: PipeMapRed exec [/home/ec2-user/CSC_555_IM2/./mapper1.py]
24/02/24 21:17:25 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
24/02/24 21:17:25 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir
24/02/24 21:17:25 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
24/02/24 21:17:25 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
24/02/24 21:17:25 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.length
24/02/24 21:17:25 INFO Configuration.deprecation: map.work.output.dir is deprecated. Instead, use mapreduce.task.output.dir
24/02/24 21:17:25 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
24/02/24 21:17:25 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
24/02/24 21:17:25 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
24/02/24 21:17:25 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
24/02/24 21:17:25 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
24/02/24 21:17:25 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
24/02/24 21:17:25 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/24 21:17:25 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/24 21:17:25 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/24 21:17:25 INFO streaming.PipeMapRed: Records R/W=4098/1
24/02/24 21:17:25 INFO streaming.PipeMapRed: R/W/S=10000/1719/0 in:NA [rec/s] out:NA [rec/s]
24/02/24 21:17:25 INFO streaming.PipeMapRed: R/W/S=10000/15224/0 in:NA [rec/s] out:NA [rec/s]
24/02/24 21:17:25 INFO streaming.PipeMapRed: R/W/S=30000/33007/0 in:NA [rec/s] out:NA [rec/s]
24/02/24 21:17:25 INFO streaming.PipeMapRed: R/W/S=30000/58946/0 in:NA [rec/s] out:NA [rec/s]
24/02/24 21:17:26 INFO mapreduce.Job: Job job_local458332266_0001 running in uber mode : false
24/02/24 21:17:26 INFO mapreduce.Task: QA reduce 0%
24/02/24 21:17:26 INFO streaming.PipeMapRed: R/W/S=400000/80089/0 in:NA [rec/s] out:NA [rec/s]
24/02/24 21:17:26 INFO streaming.PipeMapRed: R/W/S=500000/99953/0 in:500000=500000/1 [rec/s] out:99953=99953/1 [rec/s]
24/02/24 21:17:26 INFO streaming.PipeMapRed: R/W/S=600000/119823/0 in:600000=600000/1 [rec/s] out:119823=119823/1 [rec/s]
24/02/24 21:17:26 INFO streaming.PipeMapRed: R/W/S=700000/139051/0 in:700000=700000/1 [rec/s] out:139051=139051/1 [rec/s]
24/02/24 21:17:26 INFO streaming.PipeMapRed: R/W/S=800000/159556/0 in:800000=800000/1 [rec/s] out:159556=159556/1 [rec/s]
24/02/24 21:17:27 INFO streaming.PipeMapRed: R/W/S=900000/179405/0 in:900000=900000/1 [rec/s] out:179405=179405/1 [rec/s]
24/02/24 21:17:27 INFO streaming.PipeMapRed: R/W/S=1000000/193916/0 in:1000000=1000000/2 [rec/s] out:193916=193916/2 [rec/s]
24/02/24 21:17:27 INFO streaming.PipeMapRed: R/W/S=1100000/219778/0 in:1100000=1100000/2 [rec/s] out:219778=219778/2 [rec/s]
24/02/24 21:17:27 INFO streaming.PipeMapRed: R/W/S=1200000/239568/0 in:1200000=1200000/2 [rec/s] out:239568=239568/2 [rec/s]
24/02/24 21:17:27 INFO streaming.PipeMapRed: R/W/S=1300000/259519/0 in:1300000=1300000/2 [rec/s] out:259519=259519/2 [rec/s]
24/02/24 21:17:27 INFO streaming.PipeMapRed: HSErrortThread done
24/02/24 21:17:27 INFO streaming.PipeMapRed: mapRedFinished
24/02/24 21:17:27 INFO mapred.Task: mapred.Task done
24/02/24 21:17:27 INFO mapred.MapTask: Starting flush of map output
24/02/24 21:17:27 INFO mapred.MapTask: Spilling map output
24/02/24 21:17:27 INFO mapred.MapTask: hufstart = 0, hufend = 3761045; hufvoid = 104857600
24/02/24 21:17:27 INFO mapred.MapTask: kvstart = 26214396(104857584); kword = 25123264(109501056); length = 1089133/6553600
24/02/24 21:17:28 INFO mapred.MapTask: Finished spill 0
24/02/24 21:17:28 INFO mapred.Task: Task:attempt_local458332266_0001_m_000000_0 is done. And is in the process of committing
24/02/24 21:17:28 INFO mapred.LocalJobRunner: Records R/W=4098/1
24/02/24 21:17:28 INFO mapred.Task: Task 'attempt_local458332266_0001_m_000000_0' done.
24/02/24 21:17:28 INFO mapred.LocalJobRunner: Finishing task: attempt_local458332266_0001_m_000000_0
24/02/24 21:17:28 INFO mapred.LocalJobRunner: Starting task: attempt_local458332266_0001_m_000000_0
24/02/24 21:17:28 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
24/02/24 21:17:28 INFO mapred.MapTask: Processing split: hdfs://localhost/data/lineorder.tbl:1:134217728+134217728
24/02/24 21:17:28 INFO mapred.MapTask: numReduceTasks: 1
24/02/24 21:17:28 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
24/02/24 21:17:28 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
24/02/24 21:17:28 INFO mapred.MapTask: soft limit at 48986000
24/02/24 21:17:28 INFO mapred.MapTask: hufstart = 0, hufvoid = 104857600
24/02/24 21:17:28 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
24/02/24 21:17:28 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.mapred.MapTaskMapOutputBuffer
24/02/24 21:17:28 INFO streaming.PipeMapRed: PipeMapRed exec [/home/ec2-user/CSC_555_IM2/./mapper1.py]
24/02/24 21:17:28 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/24 21:17:28 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/24 21:17:28 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/24 21:17:28 INFO streaming.PipeMapRed: Records R/W=5282/1
24/02/24 21:17:28 INFO streaming.PipeMapRed: R/W/S=10000/1278/0 in:NA [rec/s] out:NA [rec/s]
24/02/24 21:17:28 INFO streaming.PipeMapRed: R/W/S=10000/15224/0 in:NA [rec/s] out:NA [rec/s]
24/02/24 21:17:28 INFO streaming.PipeMapRed: R/W/S=200000/3719/0 in:NA [rec/s] out:NA [rec/s]
I-062d65a3dd2e0d019 (nachiketh_server)
PublicIPs: 54.144.46.55 PrivateIPs: 172.31.62.135
```

```
24/02/24 21:17:38 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
24/02/24 21:17:38 INFO mapred.Merger: Merging 1 sorted segments
24/02/24 21:17:38 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 18953979 bytes
24/02/24 21:17:38 INFO mapred.LocalJobRunner: 5 / 5 copied.
24/02/24 21:17:38 INFO streaming.PipeMapRed: PipeMapRed exec [/home/ec2-user/CSC_555_HM2/./reducer1.py]
24/02/24 21:17:38 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
24/02/24 21:17:38 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
24/02/24 21:17:38 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/24 21:17:38 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/24 21:17:38 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/24 21:17:38 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/24 21:17:38 INFO streaming.PipeMapRed: Records R/W=5946/1
24/02/24 21:17:38 INFO streaming.PipeMapRed: R/W/S=10000/3731/0 in:NA [rec/s] out:NA [rec/s]
24/02/24 21:17:38 INFO streaming.PipeMapRed: R/W/S=100000/79879/0 in:NA [rec/s] out:NA [rec/s]
24/02/24 21:17:38 INFO streaming.PipeMapRed: R/W/S=200000/162059/0 in:NA [rec/s] out:NA [rec/s]
24/02/24 21:17:38 INFO streaming.PipeMapRed: R/W/S=300000/250976/0 in:300000=300000/1 [rec/s] out:250976=250976/1 [rec/s]
24/02/24 21:17:38 INFO streaming.PipeMapRed: R/W/S=400000/339991/0 in:400000=400000/1 [rec/s] out:339991=339991/1 [rec/s]
24/02/24 21:17:40 INFO streaming.PipeMapRed: R/W/S=500000/421109/0 in:500000=500000/1 [rec/s] out:421109=421109/1 [rec/s]
24/02/24 21:17:40 INFO streaming.PipeMapRed: R/W/S=600000/509917/0 in:600000=600000/1 [rec/s] out:509917=509917/1 [rec/s]
24/02/24 21:17:40 INFO streaming.PipeMapRed: R/W/S=700000/590622/0 in:700000=700000/1 [rec/s] out:590622=590622/1 [rec/s]
24/02/24 21:17:40 INFO streaming.PipeMapRed: R/W/S=800000/679722/0 in:800000=800000/1 [rec/s] out:679722=679722/1 [rec/s]
24/02/24 21:17:40 INFO streaming.PipeMapRed: R/W/S=900000/768276/0 in:450000=900000/2 [rec/s] out:384138=768276/2 [rec/s]
24/02/24 21:17:41 INFO streaming.PipeMapRed: R/W/S=1000000/858133/0 in:500000=1000000/2 [rec/s] out:426669=858133/2 [rec/s]
24/02/24 21:17:41 INFO streaming.PipeMapRed: R/W/S=1100000/947072/0 in:550000=1100000/2 [rec/s] out:473536=947072/2 [rec/s]
24/02/24 21:17:41 INFO streaming.PipeMapRed: R/W/S=1200000/1035205/0 in:600000=1200000/2 [rec/s] out:517602=1035205/2 [rec/s]
24/02/24 21:17:41 INFO streaming.PipeMapRed: MRErrorThread done
24/02/24 21:17:41 INFO streaming.PipeMapRed: mapredFinished
24/02/24 21:17:41 INFO mapred.Task: Task:attempt_local458332266_0001_r_000000_0 is done. And is in the process of committing
24/02/24 21:17:41 INFO mapred.LocalJobRunner: 5 / 5 copied.
24/02/24 21:17:41 INFO mapred.Task: Task attempt_local458332266_0001_r_000000_0 is allowed to commit now
24/02/24 21:17:41 INFO output.FileOutputCommitter: Saved output of task 'attempt_local458332266_0001_r_000000_0' to hdfs://localhost/data/job_output_1/temporary/0/task_local458332266_0001_r_000000
24/02/24 21:17:41 INFO mapred.LocalJobRunner: Records R/W=9946/1 > reduce
24/02/24 21:17:41 INFO mapred.Task: Task 'attempt_local458332266_0001_r_000000_0' done.
24/02/24 21:17:41 INFO mapred.LocalJobRunner: Finishing task: attempt_local458332266_0001_r_000000_0
24/02/24 21:17:41 INFO mapred.LocalJobRunner: reduce task executor complete.
24/02/24 21:17:42 INFO mapreduce.Job: map 100% reduce 100%
24/02/24 21:17:42 INFO mapreduce.Job: Job job_local458332266_0001 completed successfully
24/02/24 21:17:42 INFO mapreduce.Job: Counters: 38

File System Counters
  FILE: Number of bytes read=37928342
  FILE: Number of bytes written=101333133
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=263097010
  HDFS: Number of bytes written=14375448
  HDFS: Number of read operations=61
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=8

Map-Reduce Framework
  Map input records=6001215
  Map output records=1201581
  Map output bytes=16550830
  Map output materialized bytes=18954022
  Input split bytes=436
  Combine input records=0
  Combine output records=0
  Reduce input groups=1138804
  Reduce shuffle bytes=18954022
  Reduce input records=1201581
  Reduce output records=1043429
  Spilled Records=2403162
  Shuffled Maps =5
  Failed Shuffles=0
  Merged Map outputs=5
  GC time elapsed (ms)=114
  CPU time spent (ms)=0
  Physical memory (bytes) snapshot=0
  Virtual memory (bytes) snapshot=0
  Total committed heap usage (bytes)=2608857088

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=594329385

File Output Format Counters
  Bytes Written=14375448

24/02/24 21:17:42 INFO streaming.StreamJob: Output directory: /data/job_output_1

real    0m19.680s
user    0m24.478s
sys      0m2.181s
[ec2-user@ip-172-31-62-135 CSC_555_HM2] 2024-02-24 21:17:42
$ hadoop fs -cat /data/job_output_1/part-00000 | less

[1]+  Stopped                  hadoop fs -cat /data/job_output_1/part-00000 | less
[ec2-user@ip-172-31-62-135 CSC_555_HM2] 2024-02-24 21:18:39
$ hadoop fs -cat /data/job_output_1/part-00000 | less]
```

i-062d65a3dd2e0d019 (nachiketh_server)

PublicIPs: 54.144.46.55 PrivateIPs: 172.31.62.135

aws

Services

Search

1000242|6|3

1000248|8|7

100026|1|6

10002713|49|1

1000271|6|3

1000272|8|0

1000276|9|2

1000310|6|4

1000315|9|7

1000320|9|2

1000321|6|4

1000326|7|7

1000329|9|2

1000331|9|8

1000337|6|7

1000340|8|6

1000346|7|7

1000355|8|6

1000362|6|4

1000365|9|7

1000366|6|6

1000382|6|7

1000385|5|1

1000388|10|5

1000395|6|6

1000396|8|10

1000397|9|8

1000405|11|2

1000406|10|4

100040|1|4

1000411|8|10

1000414|6|4

1000418|9|5

1000426|9|5

1000430|8|8

1000434|7|8

1000440|6|6

1000441|8|5

1000445|6|6

1000448|7|2

1000450|5|0

1000452|7|3

1000456|8|5

10004650|50|0

1000469|9|5

1000479|7|8

1000483|8|9

1000486|8|1

1000487|9|10

1000495|6|10

:

i-062d65a3dd2e0d019 (nachiketh_server)

PublicIPs: 54.144.46.55 PrivateIPs: 172.31.62.135

```
time hadoop jar hadoop-streaming-2.6.4.jar -input /data/job_output_1 -output /data/job_output_2 -mapper
mapper2.py -reducer reducer2.py -file mapper2.py -file reducer2.py
```

aws

Services

Search

lo_quantity|max(lo_revenue)

10|1967440

11|2185084

12|2375748

13|2573727

14|2767703

15|2959696

16|3185904

17|3389868

18|3570445

19|3783261

1|197218

20|3951962

21|4181500

22|4347158

23|4568813

24|4737817

25|4982726

26|5167215

27|5355694

28|5503460

29|5752412

2|396148

30|5916571

31|6128515

32|6311009

33|6580333

34|6724827

35|6922616

36|7168285

37|7325224

38|7555693

39|7732297

3|597072

40|7938162

41|8136616

42|8307140

43|8549864

44|8744518

45|8891914

46|9111362

47|9354130

48|9512068

49|9742868

4|795716

50|9965452

5|985620

6|1187874

7|1389178

8|1583832

9|1792926

[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-24 21:26:35

\$

i-062d65a3dd2e0d019 (nachiketh_server)

PublicIPs: 54.144.46.55 PrivateIPs: 172.31.62.135

MULTI NODE RUNS:

JOB 1:

time hadoop jar hadoop-streaming-2.6.4.jar -input /data/lineorder.tbl -output /data/job_output_1 -mapper mapper1.py -reducer reducer1.py -file mapper1.py -file reducer1.py

job 1 output:

aws

Services

Search

[Alt+S]

ec2-user@ip-172-31-52-86 ~]\$ nano reducer2.py

24/02/24 17:34:23 INFO streaming.ShellJob: file option is deprecated, please use generic option -files instead

PackageJobJar: [mapger1.py, reducer1.py, reducer1.py, /tmp/hadoop-userjar1812380378008978708/] [] /tmp/etwaw/job7494271599245041693.jar tmpDir=null

24/02/24 17:34:24 INFO client.RMProxy: Connecting to ResourceManager at /172.31.52.86:8032

24/02/24 17:34:24 INFO client.RMProxy: Connecting to ResourceManager at /172.31.52.86:8032

24/02/24 17:34:25 INFO mapred.FileInputFormat: Total input paths to process : 1

24/02/24 17:34:25 INFO mapreduce.JobSubmitter: number of splits=5

24/02/24 17:34:25 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1708786053669_0006

24/02/24 17:34:25 INFO mapreduce.Job: Submitted application application_1708786053669_0006

24/02/24 17:34:25 INFO mapreduce.Job: The url to track the job: http://ip-172-31-52-86.ec2.internal:8088/proxy/application_1708786053669_0006/

24/02/24 17:34:25 INFO mapreduce.Job: Running job: job_1708786053669_0006

24/02/24 17:34:30 INFO mapreduce.Job: Job job_1708786053669_0006 running in user mode : false

24/02/24 17:34:30 INFO mapreduce.Job: map OK reduce OK

24/02/24 17:34:36 INFO mapreduce.Job: map 32% reduce OK

24/02/24 17:34:40 INFO mapreduce.Job: map 100% reduce OK

24/02/24 17:34:45 INFO mapreduce.Job: map 100% reduce 100%

24/02/24 17:34:45 INFO mapreduce.Job: Job job_1708786053669_0006 completed successfully

24/02/24 17:34:46 INFO mapreduce.Job: Counters: 50

File System Counters

FILE: Number of bytes read=1893398

FILE: Number of bytes written=98370017

FILE: Number of read operations=0

FILE: Number of write operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=19120815

HDFS: Number of bytes written=14375448

HDFS: Number of read operations=19

HDFS: Number of write operations=0

HDFS: Number of large read operations=0

HDFS: Number of write operations=2

Job Counters

Killed map task=1

Launched map task=5

Launched reduce task=1

Data-Local map task=0

Total time spent by all maps in occupied slots (ms)=147628

Total time spent by all reducers in occupied slots (ms)=23292

Total time spent by all map tasks (ms)=4889

Total time spent by all reduce tasks (ms)=8823

Total vcore-milliseconds taken by all map tasks=36892

Total vcore-milliseconds taken by all reduce tasks=6823

Total megabyte-milliseconds taken by all map tasks=147628000

Total megabyte-milliseconds taken by all reduce tasks=23292000

Map-Reduce Framework

Map input records=6001215

Map output records=1201581

Map output bytes=1650890

Map output materialized bytes=18964022

Input split bytes=460

Combine input records=0

Combine output records=0

i-01f594425b534e2ef (Master)

PublicIPs: 3.94.185.130 PrivateIPs: 172.31.52.86

aws

Services

Search

[Alt]

Data-local map tasks=5
Total time spent by all maps in occupied slots (ms)=147528
Total time spent by all reduces in occupied slots (ms)=23292
Total time spent by all map tasks (ms)=36882
Total time spent by all reduce tasks (ms)=5823
Total vcore-milliseconds taken by all map tasks=36882
Total vcore-milliseconds taken by all reduce tasks=5823
Total megabyte-milliseconds taken by all map tasks=147528000
Total megabyte-milliseconds taken by all reduce tasks=23292000

Map-Reduce Framework
Map input records=6001215
Map output records=1201581
Map output bytes=16550830
Map output materialized bytes=18954022
Input split bytes=450
Combine input records=0
Combine output records=0
Reduce input groups=1138804
Reduce shuffle bytes=18954022
Reduce input records=1201581
Reduce output records=1043429
Spilled Records=2403162
Shuffled Maps =5
Failed Shuffles=0
Merged Map outputs=5
GC time elapsed (ms)=579
CPU time spent (ms)=16750
Physical memory (bytes) snapshot=1622245376
Virtual memory (bytes) snapshot=21450235904
Total committed heap usage (bytes)=1417150464

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=594329385

File Output Format Counters
Bytes Written=14375448

24/02/24 17:34:45 INFO streaming.StreamJob: Output directory: /data/job_output_1

real 0m22.996s
user 0m3.864s
sys 0m0.204s

[ec2-user@ip-172-31-52-86 ~]\$ ^C
[ec2-user@ip-172-31-52-86 ~]\$ hadoop fs -cat /data/job_output_1/part-00000 | less

[1]+ Stopped hadoop fs -cat /data/job_output_1/part-00000 | less
[ec2-user@ip-172-31-52-86 ~]\$

i-01f594425b534e2ef (Master)

PublicIPs: 3.94.185.130 PrivateIPs: 172.31.52.86



Services

Search

```
1000242|6|3
1000248|8|7
100026|1|6
10002713|49|1
1000271|6|3
1000272|8|0
1000276|9|2
1000310|6|4
1000315|9|7
1000320|9|2
1000321|6|4
1000326|7|7
1000329|9|2
1000331|9|8
1000337|6|7
1000340|8|6
1000346|7|7
1000355|8|6
1000362|6|4
1000365|9|7
1000366|6|6
1000382|6|7
1000385|5|1
1000388|10|5
1000395|6|6
1000396|8|10
1000397|9|8
1000405|11|2
1000406|10|4
100040|1|4
1000411|8|10
1000414|6|4
1000418|9|5
1000426|9|5
1000430|8|8
1000434|7|8
1000440|6|6
1000441|8|5
1000445|6|6
1000448|7|2
1000450|5|0
1000452|7|3
1000456|8|5
10004650|50|0
1000469|9|5
1000479|7|8
1000483|8|9
1000486|8|1
1000487|9|10
1000495|6|10
=
```

i-01f594425b534e2ef (Master)

PublicIPs: 3.94.185.130 PrivateIPs: 172.31.52.86

JOB 2:

time hadoop jar hadoop-streaming-2.6.4.jar -input /data/job_output_1 -output /data/job2_output -mapper mapper2.py -reducer reducer2.py -file mapper2.py -file reducer2.py

```
[ec2-user@ip-172-31-52-86 ~]$ time hadoop jar hadoop-streaming-2.6.4.jar -input /data/job_output_1 -output /data/job_output2 -mapper mapper2.py -reducer reducer2.py -file mapper2.py -file reducer2.py
24/02/24 17:47:21 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [mapper2.py, reducer2.py, /tmp/hadoop-unjar2015844659589727637/] [] /tmp/streamjob4232451981533053644.jar tmpDir=null
24/02/24 17:47:22 INFO client.RMProxy: Connecting to ResourceManager at /172.31.52.86:8032
24/02/24 17:47:22 INFO client.RMProxy: Connecting to ResourceManager at /172.31.52.86:8032
24/02/24 17:47:22 INFO mapred.FileInputFormat: Total input paths to process : 1
24/02/24 17:47:22 INFO mapreduce.JobSubmitter: number of splits:2
24/02/24 17:47:22 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1708786053669_0009
24/02/24 17:47:23 INFO impl.YarnClientImpl: Submitted application application_1708786053669_0009
24/02/24 17:47:23 INFO mapreduce.Job: The url to track the job: http://ip-172-31-52-86.ec2.internal:8089/proxy/application_1708786053669_0009/
24/02/24 17:47:23 INFO mapreduce.Job: Running job: job_1708786053669_0009
24/02/24 17:47:28 INFO mapreduce.Job: Job job_1708786053669_0009 running in uber mode : false
24/02/24 17:47:28 INFO mapreduce.Job: map 0% reduce 0%
24/02/24 17:47:35 INFO mapreduce.Job: map 50% reduce 0%
24/02/24 17:47:36 INFO mapreduce.Job: map 100% reduce 0%
24/02/24 17:47:41 INFO mapreduce.Job: map 100% reduce 100%
24/02/24 17:47:41 INFO mapreduce.Job: Job job_1708786053669_0009 completed successfully
24/02/24 17:47:41 INFO mapreduce.Job: Counters: 50

File System Counters
  FILE: Number of bytes read=5218287
  FILE: Number of bytes written=10767545
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=14379744
  HDFS: Number of bytes written=615
  HDFS: Number of read operations=9
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=2

Job Counters
  Killed map tasks=1
  Launched map tasks=2
  Launched reduce tasks=1
  Data-local map tasks=2
  Total time spent by all maps in occupied slots (ms)=45372
  Total time spent by all reduces in occupied slots (ms)=13000
  Total time spent by all map tasks (ms)=11343
  Total time spent by all reduce tasks (ms)=3250
  Total vcore-milliseconds taken by all map tasks=11343
  Total vcore-milliseconds taken by all reduce tasks=3250
  Total megabyte-milliseconds taken by all map tasks=45372000
  Total megabyte-milliseconds taken by all reduce tasks=13000000

Map-Reduce Framework
  Map input records=1043429
  Map output records=381226
  Map output bytes=4455829
  Map output materialized bytes=5218293
  Input split bytes=200
  Combine input records=0
  Combine output records=0
  Reduce input groups=381226
  Reduce shuffle bytes=5218293
  Reduce input records=381226
  Reduce output records=51
  Spilled Records=762452
  Shuffled Maps =2
  Failed Shuffles=0
  Merged Map outputs=2
  GC time elapsed (ms)=281
  CPU time spent (ms)=5410
  Physical memory (bytes) snapshot=727285760
  Virtual memory (bytes) snapshot=9971367936
  Total committed heap usage (bytes)=579338240

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=14379544
File Output Format Counters
  Bytes Written=615
24/02/24 17:47:41 INFO streaming.StreamJob: Output directory: /data/job_output2

real    0m20.974s
user    0m3.746s
sys      0m0.276s
[ec2-user@ip-172-31-52-86 ~]$ hadoop fs -cat /data/job_output2/part-00000 | less
[4]+  Stopped                  hadoop fs -cat /data/job_output2/part-00000 | less
[ec2-user@ip-172-31-52-86 ~]$ hadoop fs -cat /data/job_output2/part-00000 | less
[5]+  Stopped                  hadoop fs -cat /data/job_output2/part-00000 | less
[ec2-user@ip-172-31-52-86 ~]$
```

i-01f594425b534e2ef (Master)

PublicIPs: 3.94.185.130 PrivateIPs: 172.31.52.86

```
lo_quantity|max(lo_revenue)
10|1967440
11|2185084
12|2375748
13|2573727
14|2767703
15|2959696
16|3185904
17|3389868
18|3570445
19|3783261
1|197218
20|3951962
21|4181500
22|4347158
23|4568813
24|4737817
25|4982726
26|5167215
27|5355694
28|5503460
29|5752412
2|396148
30|5916571
31|6128515
32|6311009
33|6580333
34|6724827
35|6922616
36|7168285
37|7325224
38|7555693
39|7732297
3|597072
40|7938162
41|8136616
42|8307140
43|8549864
44|8744518
45|8891914
46|9111362
47|9354130
48|9512068
49|9742868
4|795716
50|9965452
5|985620
6|1187874
7|1389178
8|1583832
9|1792926
(END)
```

i-01f594425b534e2ef (Master)

PublicIPs: 3.94.185.130 PrivateIPs: 172.31.52.86

JOB TYPE	SINGLE NODE	MULTI NODE
JOB 1	real 0m19.680s user 0m24.478s sys 0m2.181s	real 0m22.996s user 0m3.864s sys 0m0.204s
JOB 2	real 0m6.361s user 0m9.497s sys 0m0.498s	real 0m20.974s user 0m3.746s sys 0m0.276s

The single-node arrangement outperforms the multi-node system in terms of Job 1 execution times. When using a single node, Job 1's real time is 19.680 seconds; however, when using many nodes, it rises to 22.996 seconds. In a similar vein, the multi-node architecture also has longer system and user times. This suggests that there are more overhead costs involved in spreading the workload across several nodes than there are possible performance benefits, which causes Job 1 in the multi-node system to execute Job 1 more slowly. As an alternative, Job 2 shows different outcomes. Job 2 takes 6.361 real seconds to finish in the single-node setup, and 20.974 real seconds in the multi-node configuration. This significant increase in Job 2's runtime in the multi-node configuration points to inefficiencies or bottlenecks in the map's parallel processing and the distribution of tasks among several nodes.

CODE:**JOB 1:****MAPPER 1:**

```
#!/usr/bin/python3
import sys

for line in sys.stdin:
    fields = line.strip().split('|')

    # Extract relevant fields
    lo_revenue = fields[12]
    lo_quantity = fields[8]
    lo_discount = fields[11]
    lo_orderpriority = fields[6]

    # Check if lo_orderpriority ends with 'URGENT'
    if lo_orderpriority.endswith('URGENT'):
        # Emit key-value pair (revenue, quantity, discount)
        print(f"{lo_revenue}|{lo_quantity}|{lo_discount}")
```

REDUCER 1:

```
#!/usr/bin/python3
import sys

curr_rev = None
max_qty = None
max_disc = None

for line in sys.stdin:
    # Split the input lines into fields
    lo_revenue, lo_quantity, lo_discount = line.strip().split('|')

    # Check if the revenue is the same as the current revenue
    if curr_rev == lo_revenue:
        # Update max quantity and max discount if necessary
        if int(lo_quantity) > max_qty:
            max_qty = int(lo_quantity)
            max_disc = int(lo_discount)
    else:
        # Output the result for the previous revenue if applicable
        if curr_rev:
            print(f"{curr_rev}|{max_qty}|{max_disc}")

        # Update curr_rev and reset max_qty and max_disc
        curr_rev = lo_revenue
        max_qty = int(lo_quantity)
        max_disc = int(lo_discount)
```

```
if curr_rev:
    print(f"{curr_rev}|{max_qty}|{max_disc}")
```

JOB 2:

MAPPER 2:

```
#!/usr/bin/python3
import sys
for line in sys.stdin:
    # Split the input line into fields
    lo_revenue, lo_quantity, lo_discount = line.strip().split('|')

    # Check if discount falls within the range 5 to 8
    if 5 <= int(lo_discount) <= 8:
        # Emit key-value pair (quantity, revenue)
        print(f"{lo_quantity}|{lo_revenue}")
```

REDUCER 2:

```
#!/usr/bin/python3
import sys
curr_qty = None
max_rev = None

# Print header
print("lo_quantity|max(lo_revenue)")

for line in sys.stdin:
    # Split the input line into fields
    lo_quantity, lo_revenue = line.strip().split('|')

    # Check if quantity is the same as the current quantity
    if curr_qty == lo_quantity:
        # Update max revenue if necessary
        max_rev = max(max_rev, int(lo_revenue))
    else:
        # Output the result for the previous quantity if applicable
        if curr_qty:
            print(f"{curr_qty}|{max_rev}")

        # Update curr_qty and reset max_rev
        curr_qty = lo_quantity
        max_rev = int(lo_revenue)

#final rev
if curr_qty:
    print(f"{curr_qty}|{max_rev}")
```

QUESTION 3:

Implement 1-step matrix multiplication using two matrix files matrix1.600 and matrix2.600 (posted in D2L). These matrix files both consist of a 600x600 matrix. Each row in the file matrix1.600 is of the form: i, j, A[i, j]. Similarly, each row in the file matrix2.600 is of the form: j, k, B[j, k] If a specific i, j (or j, k) combination is missing in matrix1.600 and matrix2.600 then its corresponding value is 0. You must implement 1-step matrix multiplication and write the corresponding mapper.py and reducer.py. You must run the matrix multiplication on a single-node and then on your multinode cluster and time its performance. Report the difference observed.

SINGLE NODE:

```
time hadoop jar hadoop-streaming-2.6.4.jar -input /data/matrix6x6_1.txt,/data/matrix6x6_2.txt -output /data/output14 -file mapper3.py -file reducer3.py -mapper mapper3.py -reducer reducer3.py
```

```
aws Services Search [Alt+S] N. Virginia vodlabs/user3013569=nparamah@depau.edu @ 28
[ec2-user@ip-172-31-62-135 ~] 2024-02-27 02:15:04
$ time hadoop jar hadoop-streaming-2.6.4.jar -input /data/matrix6x6_1.txt,/data/matrix6x6_2.txt -output /data/output14 -file mapper3.py -file reducer3.py -mapper mapper3.py -reducer reducer3.py
24/02/27 02:15:17 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [mapper3.py, reducer3.py] [] /tmp/streamjob7730625390512903651.jar tmpDir=null
24/02/27 02:15:18 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
24/02/27 02:15:18 INFO JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
24/02/27 02:15:18 INFO JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
24/02/27 02:15:18 INFO mapred.FileInputFormat: Total input paths to process : 2
24/02/27 02:15:18 INFO mapreduce.JobSubmitter: number of splits:2
24/02/27 02:15:18 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local335694294_0001
24/02/27 02:15:19 INFO mapred.LocalDistributedCacheManager: Localized file:/home/ec2-user/mapper3.py as file:/tmp/hadoop-ec2-user/mapred/local/1709000118998/mapper3.py
24/02/27 02:15:19 INFO mapred.LocalDistributedCacheManager: Localized file:/home/ec2-user/reducer3.py as file:/tmp/hadoop-ec2-user/mapred/local/1709000118999/reducer3.py
24/02/27 02:15:19 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
24/02/27 02:15:19 INFO mapreduce.Job: Running job: job_local335694294_0001
24/02/27 02:15:19 INFO mapred.LocalJobRunner: OutputCommitter set in config null
24/02/27 02:15:19 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
24/02/27 02:15:19 INFO mapred.LocalJobRunner: Waiting for map tasks
24/02/27 02:15:19 INFO mapred.LocalJobRunner: Starting task: attempt_local335694294_0001_m_000000_0
24/02/27 02:15:19 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
24/02/27 02:15:19 INFO mapred.MapTask: Processing split: hdfs://localhost/data/matrix6x6_1.txt:0+176502
24/02/27 02:15:19 INFO mapred.MapTask: numReduceTasks: 1
24/02/27 02:15:19 INFO mapred.MapTask: EQOUTOR 0 kv: 26214396(104857584)
24/02/27 02:15:19 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
24/02/27 02:15:19 INFO mapred.MapTask: soft limit at 83886080
24/02/27 02:15:19 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
24/02/27 02:15:19 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
24/02/27 02:15:19 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
24/02/27 02:15:19 INFO streaming.PipeMapRed: PipeMapRed exec [/home/ec2-user/./mapper3.py]
24/02/27 02:15:19 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
24/02/27 02:15:19 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir
24/02/27 02:15:19 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
24/02/27 02:15:19 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
24/02/27 02:15:19 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.length
24/02/27 02:15:19 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.output.dir
24/02/27 02:15:19 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
24/02/27 02:15:19 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
24/02/27 02:15:19 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
24/02/27 02:15:19 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
24/02/27 02:15:19 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
24/02/27 02:15:19 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
24/02/27 02:15:19 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/27 02:15:19 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/27 02:15:19 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]

i-062d65a3dd2e0d019 (nachiketh_server)
PublicIPs: 18.215.160.65 PrivateIPs: 172.31.62.135
```

```
aws Services Search
Map-Reduce Framework
  Map input records=20000
  Map output records=2000000
  Map output bytes=39140400
  Map output materialized bytes=43140412
  Input split bytes=178
  Combine input records=0
  Combine output records=0
  Reduce input groups=10000
  Reduce shuffle bytes=43140412
  Reduce input records=2000000
  Reduce output records=10000
  Spilled Records=4000000
  Shuffled Maps =2
  Failed Shuffles=0
  Merged Map outputs=2
  GC time elapsed (ms)=12
  CPU time spent (ms)=0
  Physical memory (bytes) snapshot=0
  Virtual memory (bytes) snapshot=0
  Total committed heap usage (bytes)=1031798784
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=353004
File Output Format Counters
  Bytes Written=241677
24/02/27 02:15:29 INFO streaming.StreamJob: Output directory: /data/output14
real    0m12.232s
user    0m15.600s
sys     0m1.076s
[ec2-user@ip-172-31-62-135 ~] 2024-02-27 02:15:29
$ hadoop fs -cat /data/output14/part-00000 | less

[2]+  Stopped                  hadoop fs -cat /data/output14/part-00000 | less
[ec2-user@ip-172-31-62-135 ~] 2024-02-27 02:19:11
$
```

i-062d65a3dd2e0d019 (nachiketh_server)

PublicIPs: 18.215.160.65 PrivateIPs: 172.31.62.135

```
aws Services Search
1,41 32.79625368014268
1,42 34.97789347785779
1,43 33.68042487785594
1,44 37.4353264919306
1,45 34.47521907924633
1,46 33.17310338907201
1,47 38.04089627538999
1,48 38.93006799421027
1,49 37.04835251327836
1,5 34.29721195573056
1,50 36.07394171362464
1,51 30.704899323922387
1,52 35.14953876427255
1,53 36.582406240539264
1,54 35.02608784236
1,55 35.26466537424075
1,56 34.837061575541455
1,57 36.89681249864044
1,58 42.16131936866344
1,59 28.86203340185723
1,6 34.532500665342276
1,60 36.69790135780007
1,61 37.64954289884143
1,62 33.84340987266133
1,63 38.51318741453649
1,64 33.70794239057527
1,65 31.353630827227065
1,66 31.730308075801275
1,67 37.473676737126304
1,68 40.75733020764856
1,69 37.89512030876346
1,7 32.078138156007505
1,70 38.76378847089113
1,71 35.368526103819356
1,72 32.11568069814145
1,73 35.248723457951385
1,74 36.149890032684134
1,75 32.007214976639574
1,76 32.12230221168318
1,77 32.681824880534066
1,78 33.772313791319036
1,79 32.71056395899955
:[]
```

i-062d65a3dd2e0d019 (nachiketh_server)

PublicIPs: 18.215.160.65 PrivateIPs: 172.31.62.135

MULTI NODE:

```
time hadoop jar hadoop-streaming-2.6.4.jar -input /data/matrix6x6_1.txt,/data/matrix6x6_2.txt -output /data/output12 -file mapper3.py -file reducer3.py -mapper mapper3.py -reducer reducer3.py
```

```
aws Services Search [Alt+S] N. Virginia voclabs/user3013569-nparamah@depaul.edu @ 2854-6811
[ec2-user@ip-172-31-52-86 ~]$ nano mapper3.py
[ec2-user@ip-172-31-52-86 ~]$ time hadoop jar hadoop-streaming-2.6.4.jar -input /data/matrix6x6_1.txt,/data/matrix6x6_2.txt -output /data/output12 -file mapper3.py -file reducer3.py -mapper mapper3.py -r
educer reducer3.py
24/02/27 02:07:54 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [mapper3.py, reducer3.py, /tmp/hadoop-unjar6145209747262122501/] [] /tmp/streamjob5726845625831008739.jar tmpDir=null
24/02/27 02:07:55 INFO client.RMProxy: Connecting to ResourceManager at /172.31.52.86:8032
24/02/27 02:07:55 INFO client.RMProxy: Connecting to ResourceManager at /172.31.52.86:8032
24/02/27 02:07:56 INFO mapred.FileInputFormat: Total input paths to process : 2
24/02/27 02:07:56 INFO mapreduce.JobSubmitter: number of splits:2
24/02/27 02:07:56 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1708992538088_0004
24/02/27 02:07:56 INFO impl.YarnClientImpl: Submitted application application_1708992538088_0004
24/02/27 02:07:56 INFO mapreduce.Job: The url to track the job: http://ip-172-31-52-86.ec2.internal:8088/proxy/application_1708992538088_0004/
24/02/27 02:07:56 INFO mapreduce.Job: Running job: job_1708992538088_0004
24/02/27 02:08:01 INFO mapreduce.Job: Job job_1708992538088_0004 running in uber mode : false
24/02/27 02:08:01 INFO mapreduce.Job: map 0% reduce 0%
24/02/27 02:08:08 INFO mapreduce.Job: map 100% reduce 0%
24/02/27 02:08:17 INFO mapreduce.Job: map 100% reduce 100%
24/02/27 02:08:18 INFO mapreduce.Job: Job job_1708992538088_0004 completed successfully
24/02/27 02:08:18 INFO mapreduce.Job: Counters: 49

File System Counters
  FILE: Number of bytes read=43140406
  FILE: Number of bytes written=86611906
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=353188
  HDFS: Number of bytes written=241677
  HDFS: Number of read operations=9
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=2

Job Counters
  Launched map tasks=2
  Launched reduce tasks=1
  Data-local map tasks=2
  Total time spent by all maps in occupied slots (ms)=38052
  Total time spent by all reduces in occupied slots (ms)=30488
  Total time spent by all map tasks (ms)=9513
  Total time spent by all reduce tasks (ms)=7622
  Total vcore-milliseconds taken by all map tasks=9513
  Total vcore-milliseconds taken by all reduce tasks=7622
  Total megabyte-milliseconds taken by all map tasks=38052000
  Total megabyte-milliseconds taken by all reduce tasks=30488000

Map-Reduce Framework
  i-01f594425b534e2ef (Master)
  PublicIPs: 54.173.88.18 PrivateIPs: 172.31.52.86
```

```
aws Services Search
Total megabyte-milliseconds taken by all map tasks=38052000
Total megabyte-milliseconds taken by all reduce tasks=30488000
Map-Reduce Framework
  Map input records=20000
  Map output records=2000000
  Map output bytes=39140400
  Map output materialized bytes=43140412
  Input split bytes=184
  Combine input records=0
  Combine output records=0
  Reduce input groups=10000
  Reduce shuffle bytes=43140412
  Reduce input records=2000000
  Reduce output records=10000
  Spilled Records=4000000
  Shuffled Maps =2
  Failed Shuffles=0
  Merged Map outputs=2
  GC time elapsed (ms)=241
  CPU time spent (ms)=10210
  Physical memory (bytes) snapshot=731951104
  Virtual memory (bytes) snapshot=9969922048
  Total committed heap usage (bytes)=562561024

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=353004
File Output Format Counters
  Bytes Written=241677
24/02/27 02:08:18 INFO streaming.StreamJob: Output directory: /data/output12

real    0m25.075s
user    0m3.709s
sys     0m0.351s
[ec2-user@ip-172-31-52-86 ~]$ hadoop fs -cat /data/output12/part-00000 | less

[4]+  Stopped                  hadoop fs -cat /data/output12/part-00000 | less
[ec2-user@ip-172-31-52-86 ~]$

i-01f594425b534e2ef (Master)
PublicIPs: 54.173.88.18 PrivateIPs: 172.31.52.86
```



```
aws Services Search
1,41 32.79625368014268
1,42 34.97789347785779
1,43 33.68042487785594
1,44 37.4353264919306
1,45 34.47521907924633
1,46 33.17310338907201
1,47 38.04089627538999
1,48 38.93006799421027
1,49 37.04835251327836
1,5 34.29721195573056
1,50 36.07394171362464
1,51 30.704899323922387
1,52 35.14953876427255
1,53 36.582406240539264
1,54 35.02608784236
1,55 35.26466537424075
1,56 34.837061575541455
1,57 36.89681249864044
1,58 42.16131936866344
1,59 28.86203340185723
1,6 34.532500665342276
1,60 36.69790135780007
1,61 37.64954289884143
1,62 33.84340987266133
1,63 38.51318741453649
1,64 33.70794239057527
1,65 31.353630827227065
1,66 31.730308075801275
1,67 37.473676737126304
1,68 40.75733020764856
1,69 37.89512030876346
1,7 32.078138156007505
1,70 38.76378847089113
1,71 35.368526103819356
1,72 32.11568069814145
1,73 35.248723457951385
1,74 36.149890032684134
1,75 32.007214976639574
1,76 32.12230221168318
1,77 32.681824880534066
1,78 33.772313791319036
1,79 32.71056395899955
-[]
i-01f594425b534e2ef (Master)
PublicIPs: 54.173.88.18 PrivateIPs: 172.31.52.86
```

MATRIX SIZE	SINGLE NODE	MULTI NODE
100x100	real 0m12.232s user 0m15.600s sys 0m1.076s	real 0m25.075s user 0m3.709s sys 0m0.351s
600x600 (did not run completely)	real 5m11.354s user 7m30.881s sys 0m43.947s	real 8m12.192s user 0m5.199s sys 0m0.501s

Based on dataset size, the runtime comparisons for matrix multiplication show clear disparities between single-node and multi-node systems. The single-node solution shows shorter total processing times for the smaller 100x100 matrix, suggesting that the expense associated with dividing jobs across several nodes outweighs the benefits of parallelization for smaller datasets. On the other hand, when the size of the matrix reaches 600x600, the computational load on a single node causes a significant increase in runtime; in contrast, multi-node configurations show reduced user and system times but greater actual time, highlighting the efficiency gains from distributed processing. Nevertheless, mapper failures at about 28% completion for both configurations point to possible problems with resource allocation, requiring memory management adjustments.

600x600 single node:

```
aws Services Search [Alt+S]

[ec2-user@ip-172-31-62-135 ~] 2024-02-27 02:32:13
$ time hadoop jar hadoop-streaming-2.6.4.jar -input /data/mat.1.txt,/data/mat.2.txt -output /data/output62 -file mapper3.py -file reducer3.py -mapper mapper3.py -reducer reducer3.py
24/02/27 02:32:23 WARN Streaming.StreamJob: file option is deprecated, please use generic option -files instead.
packageJobJar: [mapper3.py, reducer3.py] (/tmp/streamjob2636449507429413527.jar tmpDir=null)
24/02/27 02:32:24 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
24/02/27 02:32:24 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
24/02/27 02:32:24 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
24/02/27 02:32:24 INFO mapred.FileInputFormat: Total input paths to process : 2
24/02/27 02:32:24 INFO mapreduce.JobSubmitter: number of splits:2
24/02/27 02:32:24 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local246144403_0001
24/02/27 02:32:24 INFO mapred.LocalDistributedCacheManager: Localized file:/home/ec2-user/mapper3.py as file:/tmp/hadoop-ec2-user/mapred/local/1709001144547/mapper3.py
24/02/27 02:32:24 INFO mapred.LocalDistributedCacheManager: Localized file:/home/ec2-user/reducer3.py as file:/tmp/hadoop-ec2-user/mapred/local/1709001144548/reducer3.py
24/02/27 02:32:24 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
24/02/27 02:32:24 INFO mapreduce.Job: Running job: job_local246144403_0001
24/02/27 02:32:24 INFO mapred.LocalJobRunner: OutputCommittee set in Config null
24/02/27 02:32:24 INFO mapred.LocalJobRunner: OutputCommittee is org.apache.hadoop.mapred.FileOutputCommittee
24/02/27 02:32:24 INFO mapred.LocalJobRunner: Waiting for map tasks
24/02/27 02:32:24 INFO mapred.LocalJobRunner: Starting task: attempt_local246144403_0001=0_000000_0
24/02/27 02:32:24 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
24/02/27 02:32:24 INFO mapred.MapTask: Processing split: hdfs://localhost/data/mat.1.txt:0+6939392
24/02/27 02:32:24 INFO mapred.MapTask: numReduceTasks: 1
24/02/27 02:32:24 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
24/02/27 02:32:24 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
24/02/27 02:32:24 INFO mapred.MapTask: soft limit at 83886080
24/02/27 02:32:24 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
24/02/27 02:32:24 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
24/02/27 02:32:24 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
24/02/27 02:32:24 INFO streaming.PipeMapRed: PipeMapRed exec [/home/ec2-user/./mapper3.py]
24/02/27 02:32:24 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
24/02/27 02:32:24 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir
24/02/27 02:32:24 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
24/02/27 02:32:24 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
24/02/27 02:32:24 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.length
24/02/27 02:32:24 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.output.dir
24/02/27 02:32:24 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
24/02/27 02:32:24 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
24/02/27 02:32:24 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
24/02/27 02:32:24 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
24/02/27 02:32:24 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
24/02/27 02:32:24 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
24/02/27 02:32:24 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/27 02:32:24 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/27 02:32:24 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/27 02:32:24 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/27 02:32:24 INFO streaming.PipeMapRed: Records R/W=7804/1
24/02/27 02:32:25 INFO mapreduce.Job: Job job_local246144403_0001 running in uber mode : false
24/02/27 02:32:25 INFO mapreduce.Job: map 0% reduce 0%
24/02/27 02:32:26 INFO streaming.PipeMapRed: R/W/S=10000/2098028/0 in:5000=10000/2 [rec/s] out:1049019=2098039/2 [rec/s]
24/02/27 02:32:27 INFO mapred.MapTask: Spilling map output
```

i-062d65a3dd2e0d019 (nachiketh_server)

PublicIPs: 18.215.160.65 PrivateIPs: 172.31.62.135

```
aws Services Search [Alt+S]

24/02/27 02:37:11 INFO mapred.MapTask: kvstart = 3505540(14022160); kvend = 20909328(83637312); length = 8810613/6553600
24/02/27 02:37:11 INFO mapred.MapTask: (EQUATOR) 71495904 kvi 17873972(71495888)
24/02/27 02:37:12 INFO mapred.LocalJobRunner: Records R/W=280670/162494875 > map
24/02/27 02:37:13 INFO mapreduce.Job: map 26% reduce 0%
24/02/27 02:37:13 INFO mapred.MapTask: Finished spill 73
24/02/27 02:37:13 INFO mapred.MapTask: (RESET) equator 71495904 kv 17873972(71495888) kvi 15698656(62794624)
24/02/27 02:37:15 INFO mapred.MapTask: Spilling map output
24/02/27 02:37:15 INFO mapred.MapTask: bufstart = 71495904; bufend = 15488665; bufvoid = 104857593
24/02/27 02:37:15 INFO mapred.MapTask: kvstart = 17873972(71495888); kvend = 9115040(36460160); length = 8758933/6553600
24/02/27 02:37:15 INFO mapred.MapTask: (EQUATOR) 24318761 kvi 6079684(24318736)
24/02/27 02:37:15 INFO mapred.LocalJobRunner: Records R/W=280670/162494875 > map
24/02/27 02:37:17 INFO mapred.MapTask: Finished spill 74
24/02/27 02:37:17 INFO mapred.MapTask: (RESET) equator 24318761 kv 6079684(24318736) kvi 3929268(15717072)
24/02/27 02:37:18 INFO mapred.LocalJobRunner: Records R/W=280670/162494875 > map
24/02/27 02:37:19 INFO mapred.MapTask: Spilling map output
24/02/27 02:37:19 INFO mapred.MapTask: bufstart = 24318761; bufend = 73175692; bufvoid = 104857600
24/02/27 02:37:19 INFO mapred.MapTask: kvstart = 6079684(24318736); kvend = 23536796(94147184); length = 8757289/6553600
24/02/27 02:37:19 INFO mapred.MapTask: (EQUATOR) 82005772 kvi 20501436(82005744)
24/02/27 02:37:19 INFO mapreduce.Job: map 27% reduce 0%
24/02/27 02:37:21 INFO mapred.MapTask: Finished spill 75
24/02/27 02:37:21 INFO mapred.MapTask: (RESET) equator 82005772 kv 20501436(82005744) kvi 18293928(73175712)
24/02/27 02:37:21 INFO streaming.PipeMapRed: Records R/W=287753/168178167
24/02/27 02:37:21 INFO mapred.LocalJobRunner: Records R/W=287753/168178167 > map
24/02/27 02:37:23 INFO mapred.MapTask: Spilling map output
24/02/27 02:37:23 INFO mapred.MapTask: bufstart = 82005772; bufend = 25807332; bufvoid = 104857599
24/02/27 02:37:23 INFO mapred.MapTask: kvstart = 20501436(82005744); kvend = 11694716(46778864); length = 8806721/6553600
24/02/27 02:37:23 INFO mapred.MapTask: (EQUATOR) 34637444 kvi 8659356(34637424)
24/02/27 02:37:24 INFO mapred.LocalJobRunner: Records R/W=287753/168178167 > map
24/02/27 02:37:25 INFO mapred.MapTask: Finished spill 76
24/02/27 02:37:25 INFO mapred.MapTask: (RESET) equator 34637444 kv 8659356(34637424) kvi 6455032(25820128)
24/02/27 02:37:27 INFO mapred.MapTask: Spilling map output
24/02/27 02:37:27 INFO mapred.MapTask: bufstart = 34637444; bufend = 83423292; bufvoid = 104857600
24/02/27 02:37:27 INFO mapred.MapTask: kvstart = 8659356(34637424); kvend = 26098704(104394816); length = 8775053/6553600
24/02/27 02:37:27 INFO mapred.MapTask: (EQUATOR) 92253388 kvi 23063340(92253360)
24/02/27 02:37:27 INFO mapred.LocalJobRunner: Records R/W=287753/168178167 > map
24/02/27 02:37:29 INFO mapred.MapTask: Finished spill 77
24/02/27 02:37:29 INFO mapred.MapTask: (RESET) equator 92253388 kv 23063340(92253360) kvi 20912592(83650368)
24/02/27 02:37:30 INFO mapred.LocalJobRunner: Records R/W=287753/168178167 > map
24/02/27 02:37:31 INFO mapred.MapTask: Spilling map output
24/02/27 02:37:31 INFO mapred.MapTask: bufstart = 92253388; bufend = 36552807; bufvoid = 104857593
24/02/27 02:37:31 INFO mapred.MapTask: kvstart = 23063340(92253360); kvend = 14381076(57524304); length = 8682265/6553600
24/02/27 02:37:31 INFO mapred.MapTask: (EQUATOR) 45382903 kvi 11345720(45382880)
24/02/27 02:37:31 WARN streaming.PipeMapRed: java.io.IOException: Spill failed
Traceback (most recent call last):
  File "/home/ec2-user/./mapper3.py", line 11, in <module>
    print(f"{i},{k}\t{v},{value}")
BrokenPipeError: [Errno 32] Broken pipe
24/02/27 02:37:31 INFO streaming.PipeMapRed: MRErrorThread done
24/02/27 02:37:31 INFO streaming.PipeMapRed: R/W/S=294791/174192302/0 in:963=294791/306 [rec/s] out:569255=174192302/306 [rec/s]
```

i-062d65a3dd2e0d019 (nachiketh_server)

PublicIPs: 18.215.160.65 PrivateIPs: 172.31.62.135

```
aws Services Search [Alt+S]

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)
at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)
at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)
at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:450)
at org.apache.hadoop.mapred.MapTask.run(MapTask.java:343)
at org.apache.hadoop.mapred.LocalJobRunner$Job$MapTaskRunnable.run(LocalJobRunner.java:243)
at java.util.concurrent.Executors$RunnableAdapter.call(Executors.java:511)
at java.util.concurrent.FutureTask.run(FutureTask.java:266)
at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1149)
at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
at java.lang.Thread.run(Thread.java:750)
24/02/27 02:37:33 INFO mapreduce.Job: Job job_local246144403_0001 failed with state FAILED due to: NA
24/02/27 02:37:33 INFO mapreduce.Job: Counters: 25
File System Counters
  FILE: Number of bytes read=2211
  FILE: Number of bytes written=4135076522
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=5668864
  HDFS: Number of bytes written=0
  HDFS: Number of read operations=6
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=1
Map-Reduce Framework
  Map input records=294791
  Map output records=173913975
  Map output bytes=3833483319
  Map output materialized bytes=0
  Input split bytes=83
  Combine input records=0
  Spilled Records=172021734
  Failed Shuffles=0
  Merged Map outputs=0
  GC time elapsed (ms)=0
  CPU time spent (ms)=0
  Physical memory (bytes) snapshot=0
  Virtual memory (bytes) snapshot=0
  Total committed heap usage (bytes)=298844160
File Input Format Counters
  Bytes Read=5668864
24/02/27 02:37:33 ERROR streaming.StreamJob: Job not successful!
Streaming Command Failed!

real    5m11.354s
user    7m30.881s
sys     0m43.947s
[ec2-user@ip-172-31-62-135 ~] 2024-02-27 02:37:34
$
```

i-062d65a3dd2e0d019 (nachiketh_server)

PublicIPs: 18.215.160.65 PrivateIPs: 172.31.62.135

600x600 multi node:

```
aws Services Search [Alt+S]

[ec2-user@ip-172-31-52-86 ~]$ time hadoop jar hadoop-streaming-2.6.4.jar -input /data/mat-1.txt,/data/mat-2.txt -output /data/output6 -file mapper3.py -file reducer3.py -mapper mapper3.py -reducer reducer3.py
time hadoop jar hadoop-streaming-2.6.4.jar -input /data/mat-1.txt,/data/mat-2.txt -output /data/output6 -file mapper3.py -file reducer3.py -mapper mapper3.py -reducer reducer3.py
24/02/27 02:27:42 WARN streaming.StreamJob: file option is deprecated, please use generic option -files instead.
packageJobJar: [mapper3.py, reducer3.py, /tmp/hadoop-unjar3937863966186627815/1] /tmp/steamjob7945054518371473560.jar tmpDir=null
24/02/27 02:27:42 INFO client.RMProxy: Connecting to ResourceManager at /172.31.52.86:8032
24/02/27 02:27:43 INFO mapred.FileInputFormat: Total input paths to process : 2
24/02/27 02:27:43 INFO mapreduce.JobSubmitter: number of splits:2
24/02/27 02:27:43 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1708992538088_0005
24/02/27 02:27:43 INFO impl.YarnClientImpl: Submitted application application_1708992538088_0005
24/02/27 02:27:43 INFO mapreduce.Job: The url to track the job: http://ip-172-31-52-86.ec2.internal:8088/proxy/application_1708992538088_0005/
24/02/27 02:27:43 INFO mapreduce.Job: Running job: job_1708992538088_0005
24/02/27 02:27:46 INFO mapreduce.Job: Job job_1708992538088_0005 running in uber mode : false
24/02/27 02:27:48 INFO mapreduce.Job: map 0% reduce 0%
24/02/27 02:27:59 INFO mapreduce.Job: map 3% reduce 0%
24/02/27 02:28:08 INFO mapreduce.Job: map 4% reduce 0%
24/02/27 02:28:14 INFO mapreduce.Job: map 5% reduce 0%
24/02/27 02:28:23 INFO mapreduce.Job: map 6% reduce 0%
24/02/27 02:28:29 INFO mapreduce.Job: map 7% reduce 0%
24/02/27 02:28:36 INFO mapreduce.Job: map 9% reduce 0%
24/02/27 02:28:50 INFO mapreduce.Job: map 10% reduce 0%
24/02/27 02:28:53 INFO mapreduce.Job: map 11% reduce 0%
24/02/27 02:29:02 INFO mapreduce.Job: map 12% reduce 0%
24/02/27 02:29:11 INFO mapreduce.Job: map 13% reduce 0%
24/02/27 02:29:17 INFO mapreduce.Job: map 14% reduce 0%
24/02/27 02:29:20 INFO mapreduce.Job: map 15% reduce 0%
24/02/27 02:29:32 INFO mapreduce.Job: map 16% reduce 0%
24/02/27 02:29:36 INFO mapreduce.Job: map 17% reduce 0%
24/02/27 02:29:41 INFO mapreduce.Job: map 18% reduce 0%
24/02/27 02:29:50 INFO mapreduce.Job: map 19% reduce 0%
24/02/27 02:29:56 INFO mapreduce.Job: map 20% reduce 0%
24/02/27 02:30:02 INFO mapreduce.Job: map 21% reduce 0%
24/02/27 02:30:12 INFO mapreduce.Job: map 22% reduce 0%
24/02/27 02:30:15 INFO mapreduce.Job: map 23% reduce 0%
24/02/27 02:30:24 INFO mapreduce.Job: map 24% reduce 0%
24/02/27 02:30:29 INFO mapreduce.Job: Task id : attempt_1708992538088_0005_w_000000_0, Status : FAILED
Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 1
at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:322)
at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:535)
at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)
at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)
at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)
at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:450)
at org.apache.hadoop.mapred.MapTask.run(MapTask.java:343)
at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:163)
at java.security.AccessController.doPrivileged(Native Method)
at java.security.p0 Subject.doAs(Subject.java:422)
at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1656)
at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:158)
```

i-01f594425b534e2ef (Master)

PublicIPs: 54.173.88.18 PrivateIPs: 172.31.52.86

```
aws Services Search [Alt+S]

24/02/27 02:30:30 INFO mapreduce.Job: map 11% reduce 0%
24/02/27 02:30:39 INFO mapreduce.Job: map 13% reduce 0%
24/02/27 02:30:45 INFO mapreduce.Job: map 14% reduce 0%
24/02/27 02:30:51 INFO mapreduce.Job: map 15% reduce 0%
24/02/27 02:31:00 INFO mapreduce.Job: map 16% reduce 0%
24/02/27 02:31:04 INFO mapreduce.Job: map 0% reduce 0%
24/02/27 02:31:15 INFO mapreduce.Job: map 3% reduce 0%
24/02/27 02:31:30 INFO mapreduce.Job: map 4% reduce 0%
24/02/27 02:31:40 INFO mapreduce.Job: map 5% reduce 0%
24/02/27 02:31:52 INFO mapreduce.Job: map 6% reduce 0%
24/02/27 02:32:02 INFO mapreduce.Job: map 7% reduce 0%
24/02/27 02:32:17 INFO mapreduce.Job: map 8% reduce 0%
24/02/27 02:32:20 INFO mapreduce.Job: map 9% reduce 0%
24/02/27 02:32:32 INFO mapreduce.Job: map 10% reduce 0%
24/02/27 02:32:41 INFO mapreduce.Job: map 11% reduce 0%
24/02/27 02:32:53 INFO mapreduce.Job: map 12% reduce 0%
24/02/27 02:33:02 INFO mapreduce.Job: map 13% reduce 0%
24/02/27 02:33:17 INFO mapreduce.Job: map 14% reduce 0%
24/02/27 02:33:20 INFO mapreduce.Job: map 15% reduce 0%
24/02/27 02:33:32 INFO mapreduce.Job: map 16% reduce 0%
24/02/27 02:33:41 INFO mapreduce.Job: map 17% reduce 0%
24/02/27 02:33:53 INFO mapreduce.Job: map 18% reduce 0%
24/02/27 02:34:02 INFO mapreduce.Job: map 19% reduce 0%
24/02/27 02:34:14 INFO mapreduce.Job: map 20% reduce 0%
24/02/27 02:34:17 INFO mapreduce.Job: map 21% reduce 0%
24/02/27 02:34:29 INFO mapreduce.Job: map 22% reduce 0%
24/02/27 02:34:38 INFO mapreduce.Job: map 23% reduce 0%
24/02/27 02:34:50 INFO mapreduce.Job: map 24% reduce 0%
24/02/27 02:34:59 INFO mapreduce.Job: map 25% reduce 0%
24/02/27 02:35:15 INFO mapreduce.Job: map 27% reduce 0%
24/02/27 02:35:30 INFO mapreduce.Job: map 28% reduce 0%
24/02/27 02:35:34 INFO mapreduce.Job: map 13% reduce 0%
24/02/27 02:35:35 INFO mapreduce.Job: Task Id : attempt_1708992538088_0005_m_000001_1000, Status : FAILED
Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 1
    at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:322)
    at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:535)
    at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)
    at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)
    at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)
    at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:450)

i-01f594425b534e2ef (Master)

PublicIPs: 54.173.88.18 PrivateIPs: 172.31.52.86
```

```
aws Services Search [Alt+S]

    at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)
    at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)
    at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:450)
    at org.apache.hadoop.mapred.MapTask.run(MapTask.java:343)
    at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:163)
    at java.security.AccessController.doPrivileged(Native Method)
    at javax.security.auth.Subject.doAs(Subject.java:422)
    at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1656)
    at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:158)

24/02/27 02:35:45 INFO mapreduce.Job: map 1% reduce 0%
24/02/27 02:35:46 INFO mapreduce.Job: Task Id : attempt_1708992538088_0005_m_000000_1002, Status : FAILED
Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 1
    at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:322)
    at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:535)
    at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)
    at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)
    at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)
    at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:450)
    at org.apache.hadoop.mapred.MapTask.run(MapTask.java:343)
    at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:163)
    at java.security.AccessController.doPrivileged(Native Method)
    at javax.security.auth.Subject.doAs(Subject.java:422)
    at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1656)
    at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:158)

24/02/27 02:35:53 INFO mapreduce.Job: map 100% reduce 100%
24/02/27 02:35:53 INFO mapreduce.Job: Job job_1708992538088_0005 failed with state FAILED due to: Task failed task_1708992538088_0005_m_000000
Job failed as tasks failed. failedMaps:1 failedReduces:0

24/02/27 02:35:53 INFO mapreduce.Job: Counters: 14
  Job Counters
    Failed map tasks=5
    Killed map tasks=1
    Killed reduce tasks=1
    Launched map tasks=6
    Other local map tasks=4
    Data-local map tasks=2
    Total time spent by all maps in occupied slots (ms)=2268100
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=567025
    Total time spent by all reduce tasks (ms)=0
    Total vcore-milliseconds taken by all map tasks=567025
    Total vcore-milliseconds taken by all reduce tasks=0
    Total megabyte-milliseconds taken by all map tasks=2268100000
    Total megabyte-milliseconds taken by all reduce tasks=0
24/02/27 02:35:53 ERROR streaming.StreamJob: Job not successful!
Streaming Command Failed!

real    8m12.192s
user    0m5.199s
sys     0m0.501s
[ec2-user@ip-172-31-52-86 ~]$ ^C
[ec2-user@ip-172-31-52-86 ~]$ []

i-01f594425b534e2ef (Master)

PublicIPs: 54.173.88.18 PrivateIPs: 172.31.52.86
```

MAPPER:

```
#!/usr/bin/python3
import sys
import os

input_file = os.environ['map_input_file']
if input_file.endswith("matrix_1.txt"):
    for line in sys.stdin:
        i, j, value = line.strip().split(',')
        for k in range(1, 101): # starting range as 1.
            print(f"{i},{k}\t{j},{value}")
elif input_file.endswith("matrix_2.txt"):
    for line in sys.stdin:
        j, k, value = line.strip().split(',')
        for i in range(0, 100): # Start from 0 to include the first row
            print(f"{i+1},{k}\t{j},{value}") # Increment i by 1 to start from 1
```

REDUCER:

```
#!/usr/bin/python3
import sys

curr_key = None
curr_values = []

# Processing input from mapper
for line in sys.stdin:
    key, value = line.strip().split('\t')
    pair, val = value.split(',')
    if key != curr_key:
        if curr_key is not None:
            result = sum(v1 * v2 for v1, v2 in curr_values)
            if result != 0:
                print(f"{curr_key}\t{result}")
        curr_key = key
        curr_values = []
    curr_values.append((float(val), float(val)))
# for last row values
if curr_key is not None:
    result = sum(v1 * v2 for v1, v2 in curr_values)
    if result != 0:
        print(f"{curr_key}\t{result}")
```


QUESTION 4:

There seems to be memory allocation error I tried different methods but ended up getting blank output (for one step), also cause the termination of master node a couple of times. The Reducer runs upto 85% before it fails.

SINGLE NODE:

```
aws Services Search [Alt+S] N. Virginia voclabs/user3013569=nparamah@depaup.edu @ 2854-6811-2301
24/02/23 02:25:24 INFO reduce.MergeManagerImpl: attempt_local1002961548_0001_m_000000_0: Shuffling to disk since 153997647 is greater than maxSingleShuffleLimit (92209152)
24/02/23 02:25:24 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1002961548_0001_m_000000_0 decomp: 153997647 len: 153997651 to DISK
24/02/23 02:25:25 INFO reduce.OnDiskMapOutput: Read 153997651 bytes from map-output for attempt_local1002961548_0001_m_000000_0
24/02/23 02:25:25 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1002961548_0001_m_000006_0 decomp: 267145 len: 267149 to MEMORY
24/02/23 02:25:25 INFO reduce.InMemoryMapOutput: Read 267145 bytes from map-output for attempt_local1002961548_0001_m_000006_0
24/02/23 02:25:25 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 267145, inMemoryMapOutputs.size() -> 2, commitMemory -> 65828655, usedMemory -> 66095800
24/02/23 02:25:25 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1002961548_0001_m_000005_0 decomp: 20037187 len: 20037191 to MEMORY
24/02/23 02:25:25 INFO reduce.InMemoryMapOutput: Read 20037187 bytes from map-output for attempt_local1002961548_0001_m_000005_0
24/02/23 02:25:25 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 20037187, inMemoryMapOutputs.size() -> 3, commitMemory -> 66095800, usedMemory -> 86132987
24/02/23 02:25:25 INFO reduce.MergeManagerImpl: attempt_local1002961548_0001_m_000002_0: Shuffling to disk since 153840632 is greater than maxSingleShuffleLimit (92209152)
24/02/23 02:25:25 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1002961548_0001_m_000002_0 decomp: 153840632 len: 153840636 to DISK
24/02/23 02:25:25 INFO reduce.OnDiskMapOutput: Read 153840636 bytes from map-output for attempt_local1002961548_0001_m_000002_0
24/02/23 02:25:25 INFO reduce.MergeManagerImpl: attempt_local1002961548_0001_m_000001_0: Shuffling to disk since 153838764 is greater than maxSingleShuffleLimit (92209152)
24/02/23 02:25:25 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1002961548_0001_m_000001_0 decomp: 153838764 len: 153838768 to DISK
24/02/23 02:25:25 INFO reduce.OnDiskMapOutput: Read 153838768 bytes from map-output for attempt_local1002961548_0001_m_000001_0
24/02/23 02:25:25 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
24/02/23 02:25:25 INFO mapred.LocalJobRunner: 7 / 7 copied.
24/02/23 02:25:25 INFO reduce.MergeManagerImpl: finalMerge called with 3 in-memory map-outputs and 4 on-disk map-outputs
24/02/23 02:25:25 INFO mapred.Merger: Merging 3 sorted segments
24/02/23 02:25:25 INFO mapred.Merger: Down to the last merge-pass, with 3 segments left of total size: 86132975 bytes
24/02/23 02:25:25 INFO mapreduce.Job: map 100% reduce 0%
24/02/23 02:25:25 INFO reduce.MergeManagerImpl: Merged 3 segments, 86132987 bytes to disk to satisfy reduce memory limit
24/02/23 02:25:25 INFO reduce.MergeManagerImpl: Merging 5 files, 701660318 bytes from disk
24/02/23 02:25:25 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
24/02/23 02:25:25 INFO mapred.Merger: Merging 5 sorted segments
24/02/23 02:25:25 INFO mapred.Merger: Down to the last merge-pass, with 5 segments left of total size: 701660278 bytes
24/02/23 02:25:25 INFO mapred.LocalJobRunner: 7 / 7 copied.
24/02/23 02:25:25 INFO streaming.PipeMapRed: PipeMapRed exec (/home/ec2-user/CSC_555_HM2/./reducer.py)
24/02/23 02:25:25 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
24/02/23 02:25:25 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
24/02/23 02:25:25 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/23 02:25:25 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/23 02:25:25 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/23 02:25:25 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/23 02:25:25 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/23 02:25:26 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/23 02:25:26 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/23 02:25:26 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:300000=300000/1 [rec/s] out:0=0/1 [rec/s]
24/02/23 02:25:27 INFO streaming.PipeMapRed: R/W/S=400000/0/0 in:400000=400000/1 [rec/s] out:0=0/1 [rec/s]
24/02/23 02:25:27 INFO streaming.PipeMapRed: R/W/S=500000/0/0 in:250000=500000/2 [rec/s] out:0=0/2 [rec/s]
24/02/23 02:25:28 INFO streaming.PipeMapRed: R/W/S=600000/0/0 in:300000=600000/2 [rec/s] out:0=0/2 [rec/s]
24/02/23 02:25:28 INFO streaming.PipeMapRed: R/W/S=700000/0/0 in:233333=700000/3 [rec/s] out:0=0/3 [rec/s]
24/02/23 02:25:28 INFO streaming.PipeMapRed: R/W/S=800000/0/0 in:266666=800000/3 [rec/s] out:0=0/3 [rec/s]
24/02/23 02:25:29 INFO streaming.PipeMapRed: R/W/S=900000/0/0 in:225000=900000/4 [rec/s] out:0=0/4 [rec/s]
24/02/23 02:25:30 INFO mapred.LocalJobRunner: reduce > reduce
24/02/23 02:25:30 INFO streaming.PipeMapRed: R/W/S=1000000/0/0 in:250000=1000000/4 [rec/s] out:0=0/4 [rec/s]
24/02/23 02:25:30 INFO streaming.PipeMapRed: R/W/S=1100000/0/0 in:220000=1100000/5 [rec/s] out:0=0/5 [rec/s]
24/02/23 02:25:31 INFO mapreduce.Job: map 100% reduce 72%
24/02/23 02:25:31 INFO streaming.PipeMapRed: R/W/S=1200000/0/0 in:200000=1200000/6 [rec/s] out:0=0/6 [rec/s]
24/02/23 02:25:32 INFO streaming.PipeMapRed: R/W/S=1300000/0/0 in:216666=1300000/6 [rec/s] out:0=0/6 [rec/s]
24/02/23 02:25:32 INFO streaming.PipeMapRed: R/W/S=1400000/0/0 in:233333=1400000/6 [rec/s] out:0=0/6 [rec/s]
24/02/23 02:25:33 INFO mapred.LocalJobRunner: reduce > reduce
24/02/23 02:25:33 INFO streaming.PipeMapRed: R/W/S=1500000/0/0 in:214285=1500000/7 [rec/s] out:0=0/7 [rec/s]
24/02/23 02:25:33 INFO streaming.PipeMapRed: R/W/S=1600000/0/0 in:200000=1600000/8 [rec/s] out:0=0/8 [rec/s]
24/02/23 02:25:34 INFO streaming.PipeMapRed: R/W/S=1700000/0/0 in:212500=1700000/9 [rec/s] out:0=0/9 [rec/s]
24/02/23 02:25:34 INFO streaming.PipeMapRed: R/W/S=1800000/0/0 in:225000=1800000/8 [rec/s] out:0=0/8 [rec/s]
24/02/23 02:25:34 INFO mapreduce.Job: map 100% reduce 73%
24/02/23 02:25:35 INFO streaming.PipeMapRed: R/W/S=1900000/0/0 in:190000=1900000/10 [rec/s] out:0=0/10 [rec/s]
24/02/23 02:25:36 INFO streaming.PipeMapRed: R/W/S=2000000/0/0 in:200000=2000000/10 [rec/s] out:0=0/10 [rec/s]
24/02/23 02:25:36 INFO streaming.PipeMapRed: R/W/S=2100000/0/0 in:210000=2100000/10 [rec/s] out:0=0/10 [rec/s]
24/02/23 02:25:36 INFO mapred.LocalJobRunner: reduce > reduce
24/02/23 02:25:36 INFO streaming.PipeMapRed: R/W/S=2200000/0/0 in:220000=2200000/10 [rec/s] out:0=0/10 [rec/s]
24/02/23 02:25:37 INFO mapreduce.Job: map 100% reduce 75%
24/02/23 02:25:38 INFO streaming.PipeMapRed: R/W/S=2300000/0/0 in:191666=2300000/12 [rec/s] out:0=0/12 [rec/s]
24/02/23 02:25:38 INFO streaming.PipeMapRed: R/W/S=2400000/0/0 in:200000=2400000/12 [rec/s] out:0=0/12 [rec/s]
24/02/23 02:25:38 INFO streaming.PipeMapRed: R/W/S=2500000/0/0 in:192307=2500000/13 [rec/s] out:0=0/13 [rec/s]
24/02/23 02:25:39 INFO streaming.PipeMapRed: R/W/S=2600000/0/0 in:200000=2600000/13 [rec/s] out:0=0/13 [rec/s]
24/02/23 02:25:39 INFO streaming.PipeMapRed: R/W/S=2700000/0/0 in:207692=2700000/13 [rec/s] out:0=0/13 [rec/s]
24/02/23 02:25:39 INFO mapred.LocalJobRunner: reduce > reduce
24/02/23 02:25:39 INFO streaming.PipeMapRed: R/W/S=2800000/0/0 in:215384=2800000/13 [rec/s] out:0=0/13 [rec/s]
24/02/23 02:25:40 INFO mapreduce.Job: map 100% reduce 80%
24/02/23 02:25:41 INFO streaming.PipeMapRed: R/W/S=2900000/0/0 in:181250=2900000/16 [rec/s] out:0=0/16 [rec/s]
24/02/23 02:25:42 INFO streaming.PipeMapRed: R/W/S=3000000/0/0 in:187500=3000000/16 [rec/s] out:0=0/16 [rec/s]
24/02/23 02:25:42 INFO streaming.PipeMapRed: R/W/S=3100000/0/0 in:193750=3100000/16 [rec/s] out:0=0/16 [rec/s]
24/02/23 02:25:42 INFO streaming.PipeMapRed: R/W/S=3200000/0/0 in:200000=3200000/16 [rec/s] out:0=0/16 [rec/s]
24/02/23 02:25:42 INFO mapred.LocalJobRunner: reduce > reduce
24/02/23 02:25:42 INFO streaming.PipeMapRed: R/W/S=3300000/0/0 in:194117=3300000/17 [rec/s] out:0=0/17 [rec/s]
24/02/23 02:25:43 INFO streaming.PipeMapRed: R/W/S=3400000/0/0 in:200000=3400000/17 [rec/s] out:0=0/17 [rec/s]
24/02/23 02:25:43 INFO streaming.PipeMapRed: R/W/S=3500000/0/0 in:205882=3500000/17 [rec/s] out:0=0/17 [rec/s]
24/02/23 02:25:43 INFO mapreduce.Job: map 100% reduce 83%
24/02/23 02:25:45 INFO mapred.LocalJobRunner: reduce > reduce
24/02/23 02:25:45 INFO streaming.PipeMapRed: R/W/S=3600000/0/0 in:180000=3600000/20 [rec/s] out:0=0/20 [rec/s]
24/02/23 02:25:46 INFO streaming.PipeMapRed: R/W/S=3700000/0/0 in:185000=3700000/20 [rec/s] out:0=0/20 [rec/s]
24/02/23 02:25:46 INFO streaming.PipeMapRed: R/W/S=3800000/0/0 in:190000=3800000/20 [rec/s] out:0=0/20 [rec/s]
24/02/23 02:25:46 INFO mapreduce.Job: map 100% reduce 85%
24/02/23 02:25:46 INFO streaming.PipeMapRed: R/W/S=3900000/0/0 in:195000=3900000/20 [rec/s] out:0=0/20 [rec/s]
24/02/23 02:25:46 INFO streaming.PipeMapRed: R/W/S=4000000/0/0 in:190476=4000000/21 [rec/s] out:0=0/21 [rec/s]
24/02/23 02:26:17 INFO mapred.LocalJobRunner: reduce > reduce
24/02/23 02:26:37 INFO mapreduce.Job: map 100% reduce 87%
24/02/23 02:26:46 INFO mapred.LocalJobRunner: reduce > reduce
```

i-062d65a3dd2e0d019 (nachiketh_server)

PublicIPs: 34.224.83.97 PrivateIPs: 172.31.62.135

MULTI NODE:

```
apper_3.py file reducer.py -mapper mapper_1.py -mapper mapper_2.py -mapper mapper_3.py -reducer reducer.py
24/02/29 02:07:53 WARN streaming.StreamJob: file option is deprecated, please use generic option -files instead.
packageJobJar: [mapper_1.py, mapper_2.py, mapper_3.py, reducer.py, /tmp/hadoop-unjar5898683568139639863/] [] /tmp/streamjob451456333265197780.jar tmpDir=null
24/02/29 02:07:53 INFO client.RMProxy: Connecting to ResourceManager at /172.31.57.221:8032
24/02/29 02:07:54 INFO client.RMProxy: Connecting to ResourceManager at /172.31.57.221:8032
24/02/29 02:07:54 INFO mapred.FileInputFormat: Total input paths to process : 3
24/02/29 02:07:54 INFO mapreduce.JobSubmitter: number of splits=7
24/02/29 02:07:55 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1709172024572_0001
24/02/29 02:07:55 INFO impl.YarnClientImpl: Submitted application application_1709172024572_0001
24/02/29 02:07:55 INFO mapreduce.Job: The url to track the job: http://ip-172-31-57-221.ec2.internal:8088/proxy/application_1709172024572_0001/
24/02/29 02:07:55 INFO mapreduce.Job: Running job: job_1709172024572_0001
24/02/29 02:08:01 INFO mapreduce.Job: Job job_1709172024572_0001 running in uber mode : false
24/02/29 02:08:01 INFO mapreduce.Job: map 0% reduce 0%
24/02/29 02:08:09 INFO mapreduce.Job: map 14% reduce 0%
24/02/29 02:08:10 INFO mapreduce.Job: map 29% reduce 0%
24/02/29 02:08:12 INFO mapreduce.Job: map 43% reduce 0%
24/02/29 02:08:13 INFO mapreduce.Job: map 57% reduce 0%
24/02/29 02:08:14 INFO mapreduce.Job: map 71% reduce 0%
24/02/29 02:08:17 INFO mapreduce.Job: map 90% reduce 0%
24/02/29 02:08:19 INFO mapreduce.Job: map 100% reduce 24%
24/02/29 02:08:22 INFO mapreduce.Job: map 100% reduce 67%
24/02/29 02:08:25 INFO mapreduce.Job: map 100% reduce 70%
24/02/29 02:08:28 INFO mapreduce.Job: map 100% reduce 73%
24/02/29 02:08:31 INFO mapreduce.Job: map 100% reduce 75%
24/02/29 02:08:34 INFO mapreduce.Job: map 100% reduce 78%
24/02/29 02:08:37 INFO mapreduce.Job: map 100% reduce 81%
24/02/29 02:08:40 INFO mapreduce.Job: map 100% reduce 85%
24/02/29 02:08:49 INFO mapreduce.Job: Task Id : attempt_1709172024572_0001_r_000000_0, Status : FAILED
Container [pid=5502,containerID=container_1709172024572_0001_01_000011] is running Beyond physical memory limits. Current usage: 4.4 GB of 4 GB physical memory used; 6.6 GB of 8.4 GB virtual memory used. Killing container.
Dump of the process-tree for container_1709172024572_0001_01_000011 :
|- PID PPID PGPRID SESSID CMD NAME USER MODE TIME(MILLIS) SYSTEM TIME(MILLIS) VMEM USAGE(BYTES) RSSMEM USAGE(PAGES) FULL CMD LINE
Djava.io.tmpdir=/tmp/hadoop-ec2-user/nm-local-dir/usercache/ec2-user/appcache/application_1709172024572_0001/container_1709172024572_0001_01_000011/tmp -Dlog4j.configuration=container-log4j.properties -Dyarn.app.container.log.dir=/home/ec2-user/hadoop-2.6.4/logs/userlogs/application_1709172024572_0001/container_1709172024572_0001_01_000011 -Dyarn.app.container.log.filesize=0 -Dhadoop.root.logg
er=INFO,CLA org.apache.hadoop.mapred.YarnChild 172.31.82.170 45679 attempt_1709172024572_0001_r_000000_0 11
|- 5502 5500 5502 5502 (bash) 0 0 228298752 767 /bin/bash -c /usr/lib/jvm/java-1.8.0-amazon-corretto.x86_64/jre/bin/java -Djava.net.preferIPv4Stack=true -Dhadoop.metrics.log.level=WARN -Xm
x200m -Djava.io.tmpdir=/tmp/hadoop-ec2-user/nm-local-dir/usercache/ec2-user/appcache/application_1709172024572_0001/container_1709172024572_0001_01_000011 -Dlog4j.configuration=container-log4j.p
roperties -Dyarn.app.container.log.dir=/home/ec2-user/hadoop-2.6.4/logs/userlogs/application_1709172024572_0001/container_1709172024572_0001_01_000011 -Dyarn.app.container.log.filesize=0 -Dhadoop.ro
ot.logger=INFO,CLA org.apache.hadoop.mapred.YarnChild 172.31.82.170 45679 attempt_1709172024572_0001_r_000000_0 11 1>/home/ec2-user/hadoop-2.6.4/logs/userlogs/application_1709172024572_0001/containe
r_1709172024572_0001_01_000011/stdout 2>/home/ec2-user/hadoop-2.6.4/logs/userlogs/application_1709172024572_0001/container_1709172024572_0001_01_000011/stderr
|- 5557 5510 5502 5502 (reducer.py) 1833 240 4782956544 1112153 /usr/bin/python3 /tmp/hadoop-ec2-user/nm-local-dir/usercache/ec2-user/appcache/application_1709172024572_0001/container_170917
2024572_0001_01_000011/./reducer.py
Container killed on request. Exit code is 143
Container exited with a non-zero exit code 143

24/02/29 02:08:44 INFO mapreduce.Job: map 100% reduce 0%
24/02/29 02:08:56 INFO mapreduce.Job: map 100% reduce 67%
24/02/29 02:08:59 INFO mapreduce.Job: map 100% reduce 70%
24/02/29 02:09:02 INFO mapreduce.Job: map 100% reduce 73%
24/02/29 02:09:05 INFO mapreduce.Job: map 100% reduce 75%
24/02/29 02:09:08 INFO mapreduce.Job: map 100% reduce 78%
24/02/29 02:09:11 INFO mapreduce.Job: map 100% reduce 81%
24/02/29 02:09:14 INFO mapreduce.Job: map 100% reduce 85%
24/02/29 02:09:16 INFO mapreduce.Job: Task Id : attempt_1709172024572_0001_r_000000_1, Status : FAILED
Container [pid=5581,containerID=container_1709172024572_0001_01_000014] is running Beyond physical memory limits. Current usage: 4.2 GB of 4 GB physical memory used; 6.4 GB of 8.4 GB virtual memory used. Killing container.
Dump of the process-tree for container_1709172024572_0001_01_000014 :
|- PID PPID PGPRID SESSID CMD NAME USER MODE TIME(MILLIS) SYSTEM TIME(MILLIS) VMEM USAGE(BYTES) RSSMEM USAGE(PAGES) FULL CMD LINE
|- 5581 5579 5581 5581 (bash) 0 0 228298752 445 /bin/bash -c /usr/lib/jvm/java-1.8.0-amazon-corretto.x86_64/jre/bin/java -Djava.net.preferIPv4Stack=true -Dhadoop.metrics.log.level=WARN -Xm
x200m -Djava.io.tmpdir=/tmp/hadoop-ec2-user/nm-local-dir/usercache/ec2-user/appcache/application_1709172024572_0001/container_1709172024572_0001_01_000014/tmp -Dlog4j.configuration=container-log4j.p
roperties -Dyarn.app.container.log.dir=/home/ec2-user/hadoop-2.6.4/logs/userlogs/application_1709172024572_0001/container_1709172024572_0001_01_000014 -Dyarn.app.container.log.filesize=0 -Dhadoop.ro
ot.logger=INFO,CLA org.apache.hadoop.mapred.YarnChild 172.31.82.170 45679 attempt_1709172024572_0001_r_000000_1 14 1>/home/ec2-user/hadoop-2.6.4/logs/userlogs/application_1709172024572_0001/containe
r_1709172024572_0001_01_000014/stdout 2>/home/ec2-user/hadoop-2.6.4/logs/userlogs/application_1709172024572_0001/container_1709172024572_0001_01_000014/stderr
|- 5631 5589 5581 5581 (reducer.py) 1853 226 4582158336 1062376 /usr/bin/python3 /tmp/hadoop-ec2-user/nm-local-dir/usercache/ec2-user/appcache/application_1709172024572_0001/container_170917
2024572_0001_01_000014/./reducer.py
|- 5589 5581 5581 5581 (java) 805 197 2081009664 36782 /usr/lib/jvm/java-1.8.0-amazon-corretto.x86_64/jre/bin/java -Djava.net.preferIPv4Stack=true -Dhadoop.metrics.log.level=WARN -Xmx200m -
Djava.io.tmpdir=/tmp/hadoop-ec2-user/nm-local-dir/usercache/ec2-user/appcache/application_1709172024572_0001/container_1709172024572_0001_01_000014/tmp -Dlog4j.configuration=container-log4j.propertie
s -Dyarn.app.container.log.dir=/home/ec2-user/hadoop-2.6.4/logs/userlogs/application_1709172024572_0001/container_1709172024572_0001_01_000014 -Dyarn.app.container.log.filesize=0 -Dhadoop.root.logg
er=INFO,CLA org.apache.hadoop.mapred.YarnChild 172.31.82.170 45679 attempt_1709172024572_0001_r_000000_1 14
Container killed on request. Exit code is 143
Container exited with a non-zero exit code 143

24/02/29 02:09:17 INFO mapreduce.Job: map 100% reduce 0%
24/02/29 02:09:31 INFO mapreduce.Job: map 100% reduce 67%
24/02/29 02:09:34 INFO mapreduce.Job: map 100% reduce 70%
24/02/29 02:09:37 INFO mapreduce.Job: map 100% reduce 73%
24/02/29 02:09:40 INFO mapreduce.Job: map 100% reduce 75%
24/02/29 02:09:43 INFO mapreduce.Job: map 100% reduce 78%
24/02/29 02:10:12 INFO mapreduce.Job: map 100% reduce 81%
24/02/29 02:10:14 INFO mapreduce.Job: map 100% reduce 85%
rc[]
```

i-06fafa7fc816ec90a (Master_1)

PublicIPs: 35.153.66.130 PrivateIPs: 172.31.57.221

MAPPER 1:

```
#!/usr/bin/python3
```

```
import sys
```

```
def hash_function(key, num_buckets):
    return hash(key) % num_buckets
```

b = 4

c = 5

k = b * c

for line in sys.stdin:

```
    fields = line.strip().split('|')
```

```
lo_orderdate = fields[5] # Extracting lo_orderdate
hashed_bucket = hash_function(lo_orderdate, k)
print(f"{hashed_bucket}\tlineorder\t{'|'.join(fields)}")
```

MAPPER 2:

```
#!/usr/bin/python3
```

```
import sys
```

```
def hash_function(key, num_buckets):
    return hash(key) % num_buckets
```

```
b = 4
```

```
c = 5
```

```
k = b * c
```

```
for line in sys.stdin:
```

```
    fields = line.strip().split('|')
```

```
    d_datekey = fields[0] # Extracting d_datekey
```

```
    hashed_bucket = hash_function(d_datekey, k)
```

```
    print(f"{hashed_bucket}\tdwdate\t{'|'.join(fields)}")
```

MAPPER 3:

```
#!/usr/bin/python3
```

```
import sys
```

```
def hash_function(key, num_buckets):
    return hash(key) % num_buckets
```

```
b = 4
```

```
c = 5
```

```
k = b * c
```

```
for line in sys.stdin:
```

```
    fields = line.strip().split('|')
```

```
    p_partkey = fields[0] # Extracting p_partkey
```

```
    hashed_bucket = hash_function(p_partkey, k)
```

```
    print(f"{hashed_bucket}\tpart\t{'|'.join(fields)}")
```

REDUCER:

```
#!/usr/bin/python3
```

```
import sys
```

```
def hash_function(key, num_buckets):
    return hash(key) % num_buckets
```

```
b = 4
```

```
c = 5
```

```
k = b * c
```

```

lineorder_data = {}
dwwdate_data = {}
part_data = {}
for line in sys.stdin:
    hashed_bucket, table_name, record = line.strip().split('\t')
    if table_name == 'lineorder':
        lineorder_data.setdefault(hashed_bucket, []).append(record.split('|'))
    elif table_name == 'dwwdate':
        dwwdate_data.setdefault(hashed_bucket, []).append(record.split('|'))
    elif table_name == 'part':
        part_data.setdefault(hashed_bucket, []).append(record.split('|'))

for bucket in range(k):
    if bucket in lineorder_data and bucket in dwwdate_data and bucket in part_data:
        for lineorder_record in lineorder_data[bucket]:
            for dwwdate_record in dwwdate_data[bucket]:
                for part_record in part_data[bucket]:
                    if dwwdate_record[12] == 'Fall' and part_record[4] == 'MFGR#2123' and lineorder_record[5] ==
dwwdate_record[0] and lineorder_record[3] == part_record[0]:
                        print('\t'.join(lineorder_record + dwwdate_record + part_record))

```