CSC 555: HW2

Name: Nachiketh Reddy
ID: 2117731

QUESTION 1.

Consider a database schema consisting of two tables, Employee (ID, Name, Address), Project (PID, Name, Deadline), Assign(EID, PID, Date). Assign.EID is a foreign key referencing employee's ID and Assign.PID is a foreign key referencing the Project.PID Write SQL queries for:

Employee Table:

SQL | All Rows Fetched: 5 in 0.016 seconds

	ID	NAME	ADDRESS
1	4	Tom Cruise	123 Birch St, Schaumburg, IL
2	5	Michael Jordan	456 Maple St, Naperville, IL
3	6	Emma Watson	789 Cali St, Naperville, IL
4	7	Chris Evans	321 Rose St, Schaumburg, IL
5	8	Tom Brady	654 Monte Carlo St, Naperville, IL

Project Table:

SQL | All Rows Fetched: 4 in 0.028 seconds

	PID	NAME	DEADLINE
1	104	Project AI	15-JUN-24
2	105	Manhattan Project	30-JUL-24
3	106	Project Mine	20-AUG-24
4	107	Project X	25-SEP-24

Assignment Table:

SQL | All Rows Fetched

	EID	PID	ASSIGN_DATE
1	4	104	05-JAN-24
2	5	105	10-JAN-24
3	6	104	15-JAN-24
4	7	105	20-JAN-24
5	7	106	05-FEB-24
6	4	105	05-JAN-24
7	5	104	05-FEB-24

(a) Find projects that are not assigned to any employees (Name and Deadline of the project).

```
-- Query 1
SELECT Name, Deadline FROM Project
WHERE PID NOT IN (SELECT PID FROM Assignment);
```

SQL | All Rows Fetched:

	NAME	DEADLINE
1	Project X	25-SEP-24

(b) For each date, find how many assignments were made that day.

```
-- Query 2
SELECT Assign_Date, COUNT(*) AS Emp_Assigned FROM Assignment
GROUP BY Assign_Date;
```

SQL | All Rows Fetched: 5 in

	ASSIGN_DATE	EMP_ASSIGNED
1	10-JAN-24	1
2	20-JAN-24	1
3	05-JAN-24	2
4	15-JAN-24	1
5	05-FEB-24	2

(c) Find all projects that have fewer than 2 employees assigned to them (note that the answer should include 0 or 1 employees to be correct).

```
-- Query 3
SELECT Z.Name, (SELECT COUNT(Y.EID) FROM Assignment Y
      WHERE Y.PID = Z.PID) AS Num_Employees
FROM Project Z WHERE (SELECT COUNT(Y.EID) FROM Assignment Y
     WHERE Y.PID = Z.PID) < 2;
```

SQL | All Rows Fetched: 2 in 0.015 s

	NAME	NUM_EMPLOYEES
1	Project Mine	1
2	Project X	0

QUESTION 2.

MMDS Book. Exercise 2.2.1 all three parts.

Exercise 2.2.1: Suppose we execute the word-count MapReduce program described in this section on a large repository such as a copy of the Web. We shall use 100 Map tasks and some number of Reduce tasks.

(a) Suppose we do not use a combiner at the Map tasks. Do you expect there to be significant skew in the times taken by the various reducers to process their value list? Why or why not?

Ans: There is definitely a skew in the time it takes. The multiple reducers processing speeds could differ greatly if there was no combiner at the Map tasks. Reducers get unaggregated data, which is caused by the various lengths of value lists associated with distinct keys. If there isn't a combiner, reducers have to handle every single value associated with every key. Processing times may skew as a result of certain reducers having significantly more data to process than others due to this unequal workload distribution.

(b) If we combine the reducers into a small number of Reduce tasks, say 10 tasks, at random, do you expect the skew to be significant? What if we instead combine the reducers into 10,000 Reduce tasks?

Ans: Even if we randomize the reducers into smaller reduce tasks, such as ten tasks, the skew will persist, especially if the data distribution from the input file varies significantly. Consequently, certain reduce tasks may bear a heavier load of key-value pairs than others. However, with 10,000 reduce tasks, the data is more evenly distributed among them, thereby reducing the workloads on individual tasks. This increased distribution enhances the likelihood of balancing out the total processing time across tasks.

(c) Suppose we do use a combiner at the 100 Map tasks. Do you expect skew to be significant? Why or why not?

Ans: If we use a combiner at 100 Map tasks, the skew is significantly reduced, as a certain amount of intermediate key-value pairs produced by Map Tasks are aggregated before they move on to the reducers. With the combination of a combiner and a reducer, the combiner helps distribute the workload to reducers

more evenly, balancing the workload. Thus, it provides a streamlined input to reducers and reduces the total processing time.

MMDS Book. Exercise 2.3.1 and 2.3.5.

Exercise 2.3.1: Design MapReduce algorithms to take a very large file of integers and produce as output:

(a) Find the largest integer:

Map function:

```
# emit integer-key pairs such that parse all integers
For each integer in chunk:
        Emit(integer,key);
```

Reduce function:

```
#Find the largest integer among all the input integer-key pairs
Max = 0;
For each integer in values:
        If integer[b] > integer [a]
                Max = integer[b]
Emit(Max);
```

(b) Calculate the average of all integers:

Map function:

```
# emit integer-key pairs
For each integer in chunk:
        Emit(integer,key);
```

Reduce Function:

```
# calculate sum
Sum =0
# to find total number of integers
Count = 0
For each integer in values:
        Sum += integer;
```

```
        Count += 1
Avg = sum/count;
Emit(Avg)
```

(c) Remove duplicates from the set of integers:

Map Function:

 # This function emits key-value pairs but here the key is integer and value is 1.

```
For each ineteger in chunk:
        Emit(integer, 1)
```

Reduce Function:

 # For each key the values are combined [(integer1, [1,1,1,1…]),(integer2, [1,1,1,1…])]

 # The unique keys are emitted

```
For each integer in values:
        Emit(intger,none)
```

(d) Count the number of distinct integers:

Map Function:

 # This function emits key-value pairs but here the key is integer and value is 1.

```
For each ineteger in chunk:
        Emit(integer, 1)
```

Reduce Function:

 # we are counting the number of occurrences of each unique key

```
Count = 0
For each value in values:
        Count += 1
Emit(Count)
```

Exercise 2.3.5: The relational-algebra operation R (A, B) ⊲⊳ B<C S(C, D) produces all tuples (a, b, c, d) such that tuple (a, b) is in relation R, tuple (c, d) is. in S, and b < c. Give a MapReduce implementation of this operation, assuming. R and S are sets.

Mapper Function:

```
# For input file R
For each tuple (a,b) in relation with R:
 # b is the key "R" signifies its from relation R and a is the value
        emit(B,("R",A))
# For input file S
For each tuple (c,d) in relation with S:
 # c is the key "S" signifies its from relation R and d is the value
        emit(C,("S",D))
```

Reducer Function:

```
# input from mapper key(B/C)
# the function identifies values based on key B/C, value list
[(source_relation,value)…] [(R,A),(R,B)…]
list R_Values = []
list S_Values = []
For each (Source_Relation,value) in value list:
        If Source_Relation == "R":
                Add_value_to(R_Values)
        Else:
                Add_value_to(S_Values)

# nested loop to verify B<C and emit results
For each i in R_Values:
        For each k in S_Values:
                If i.b < k.c:
                        Emit((r.A,r.B,s.C,s.D))
```

QUESTION 3:

Consider a Hadoop job that processes an input data file of size equal to 165 disk blocks (165 different blocks, you can assume that HDFS replication factor is set to 1). The mapper in this job requires 1 minute to read and process a single block of data. For the purposes of this assignment, you can assume that the Reduce part of this job takes zero time.

(a) Approximately how long will it take to process the file if you had

(i) 30 nodes and all nodes participate in map tasks;

Number of nodes = 30
Number of blocks processed by each node = (165/30)
Time taken to process 1 block = 1 min
Therefore, Time taken to process = Number of nodes x time taken to process one block x number of blocks processed by each node
= 30 x 1 x (165/30) = 165 mins.

(ii) 100 nodes and all nodes participate in map tasks.

Number of nodes = 100
Number of blocks processed by each node = (165/100)
Time taken to process 1 block = 1 min
Therefore, Time taken to process = Number of nodes x time taken to process one block x number of blocks processed by each node
= 100 x 1 x (165/100) = 165 mins.

(b) Now suppose you were told that the replication factor has been changed to 3. That is, each block is stored in triplicate, but file size is still 165 blocks. Which of the answers (if any) in the part above will have to change? You can ignore the network transfer costs and other potential overheads as well as the possibility of node failure. State any assumptions you make.

Replication factors of three indicate that fault tolerance and redundancy are provided by replicating each block three times. By spreading out over the Hadoop cluster, these copies improve data availability and resilience against node failures. The number of blocks to be processed, however, stays unchanged at 165 in our context of processing time analysis and is not impacted by changes in the replication factor. As a result, the processing time of the file, as determined in part (a) for scenarios (i) and (ii), will continue to exist.

QUESTION 4:

Implement the following SQL queries by writing the corresponding mapper and reducer code to achieve the equivalent result using Hadoop Streaming. The output should have the column names at the top.

QUESTION 4 a:

SELECT lo_quantity, lo_linenumber FROM lineorder
WHERE lo discount < 10 AND lo tax > 2

```
aws  Services  Search [Alt+S]  N. Virginia ▾  voclabs/user3013569=nparamah@depaul.edu @ 2854-6811-230
$ nano reducer.py
[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-03 23:58:53
$ hadoop jar hadoop-streaming-2.6.4.jar -input /data/lineorder.tbl -output /data/output02 -file mapper.py -file reducer.py -mapper mapp
er.py -reducer reducer.py
24/02/03 23:59:50 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [mapper.py, reducer.py] [] /tmp/streamjob1821579747807034762.jar tmpDir=null
24/02/03 23:59:51 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
24/02/03 23:59:51 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
24/02/03 23:59:51 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
24/02/03 23:59:51 INFO mapred.FileInputFormat: Total input paths to process : 1
24/02/03 23:59:51 INFO mapreduce.JobSubmitter: number of splits:5
24/02/03 23:59:51 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local2028928857_0001
24/02/03 23:59:51 INFO mapred.LocalDistributedCacheManager: Localized file:/home/ec2-user/CSC_555_HW2/mapper.py as file:/tmp/hadoop-ec2
-user/mapred/local/1707004791488/mapper.py
24/02/03 23:59:51 INFO mapred.LocalDistributedCacheManager: Localized file:/home/ec2-user/CSC_555_HW2/reducer.py as file:/tmp/hadoop-ec
2-user/mapred/local/1707004791489/reducer.py
24/02/03 23:59:51 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
24/02/03 23:59:51 INFO mapreduce.Job: Running job: job_local2028928857_0001
24/02/03 23:59:51 INFO mapred.LocalJobRunner: OutputCommitter set in config null
24/02/03 23:59:51 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
24/02/03 23:59:51 INFO mapred.LocalJobRunner: Waiting for map tasks
24/02/03 23:59:51 INFO mapred.LocalJobRunner: Starting task: attempt_local2028928857_0001_m_000000_0
24/02/03 23:59:51 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
24/02/03 23:59:51 INFO mapred.MapTask: Processing split: hdfs://localhost/data/lineorder.tbl:0+134217728
24/02/03 23:59:51 INFO mapred.MapTask: numReduceTasks: 1
24/02/03 23:59:51 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
24/02/03 23:59:51 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
24/02/03 23:59:51 INFO mapred.MapTask: soft limit at 83886080
24/02/03 23:59:51 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
24/02/03 23:59:51 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
24/02/03 23:59:51 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
24/02/03 23:59:51 INFO streaming.PipeMapRed: PipeMapRed exec [/home/ec2-user/CSC_555_HW2/./mapper.py]
24/02/03 23:59:51 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
24/02/03 23:59:51 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir
24/02/03 23:59:51 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
24/02/03 23:59:51 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
24/02/03 23:59:51 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.length
24/02/03 23:59:51 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.output.dir
24/02/03 23:59:51 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
24/02/03 23:59:51 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
24/02/03 23:59:51 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
24/02/03 23:59:51 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
24/02/03 23:59:51 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
24/02/03 23:59:51 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
24/02/03 23:59:51 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/03 23:59:51 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/03 23:59:51 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/03 23:59:51 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/03 23:59:51 INFO streaming.PipeMapRed: Records R/W=4098/1
24/02/03 23:59:51 INFO streaming.PipeMapRed: R/W/S=10000/5097/0 in:NA [rec/s] out:NA [rec/s]
24/02/03 23:59:52 INFO streaming.PipeMapRed: R/W/S=100000/59468/0 in:NA [rec/s] out:NA [rec/s]
24/02/03 23:59:52 INFO streaming.PipeMapRed: R/W/S=200000/120111/0 in:NA [rec/s] out:NA [rec/s]
24/02/03 23:59:52 INFO mapreduce.Job: Job job_local2028928857_0001 running in uber mode : false
24/02/03 23:59:52 INFO mapreduce.Job:  map 0% reduce 0%
24/02/03 23:59:52 INFO streaming.PipeMapRed: R/W/S=300000/180108/0 in:NA [rec/s] out:NA [rec/s]
24/02/03 23:59:53 INFO streaming.PipeMapRed: R/W/S=400000/241259/0 in:400000=400000/1 [rec/s] out:241259=241259/1 [rec/s]
24/02/03 23:59:53 INFO streaming.PipeMapRed: R/W/S=500000/302462/0 in:500000=500000/1 [rec/s] out:302462=302462/1 [rec/s]
24/02/03 23:59:53 INFO streaming.PipeMapRed: R/W/S=600000/361923/0 in:600000=600000/1 [rec/s] out:361923=361923/1 [rec/s]
24/02/03 23:59:53 INFO streaming.PipeMapRed: R/W/S=700000/423050/0 in:700000=700000/1 [rec/s] out:423050=423050/1 [rec/s]
24/02/03 23:59:54 INFO streaming.PipeMapRed: R/W/S=800000/484218/0 in:400000=800000/2 [rec/s] out:242109=484218/2 [rec/s]
24/02/03 23:59:54 INFO streaming.PipeMapRed: R/W/S=900000/545351/0 in:450000=900000/2 [rec/s] out:272675=545351/2 [rec/s]
i-062d65a3dd2e0d019 (nachiketh_server)
PublicIPs: 54.160.108.70  PrivateIPs: 172.31.62.135
```

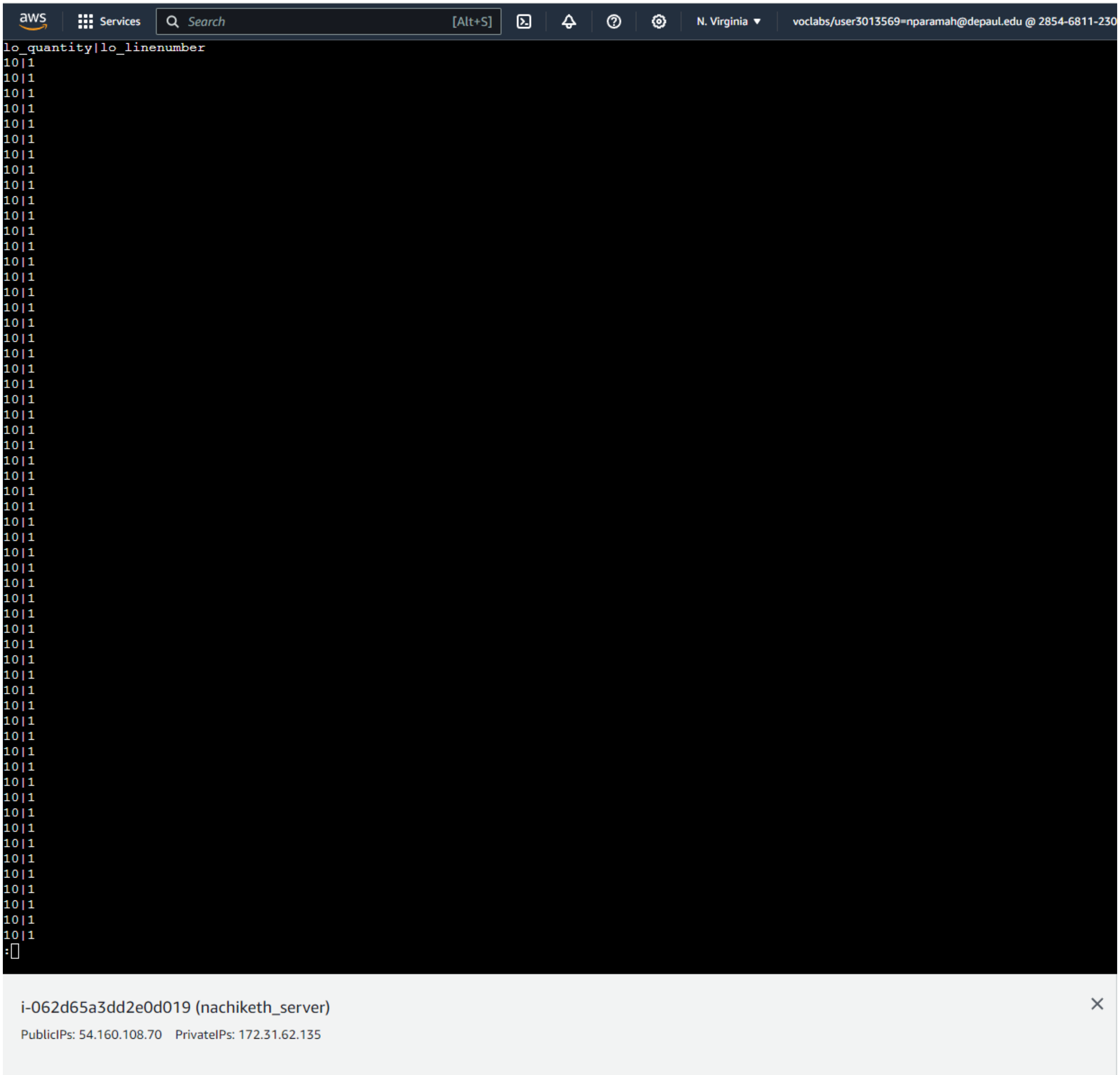
aws Services Search [Alt+S] N. Virginia voclabs/user3013569=nparamah@depaul.edu @ 2854-6811-230

```
24/02/04 00:00:14 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=56917604
                FILE: Number of bytes written=151299653
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=2530877010
                HDFS: Number of bytes written=21174324
                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=3638030
                Map output bytes=21174297
                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=7276060
                Shuffled Maps =5
                Failed Shuffles=0
                Merged Map outputs=5
                GC time elapsed (ms)=100
                CPU time spent (ms)=0
                Physical memory (bytes) snapshot=0
                Virtual memory (bytes) snapshot=0
                Total committed heap usage (bytes)=2587885568
        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=21174324
24/02/04 00:00:14 INFO streaming.StreamJob: Output directory: /data/output02
[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-04 00:00:15
$ hadoop fs -ls /data
Found 7 items
-rw-r--r--   3 ec2-user supergroup  594313001 2024-02-03 23:29 /data/lineorder.tbl
drwxr-xr-x   - ec2-user supergroup          0 2024-02-03 23:46 /data/output01
drwxr-xr-x   - ec2-user supergroup          0 2024-02-04 00:00 /data/output02
drwxr-xr-x   - ec2-user supergroup          0 2024-02-03 23:30 /data/output1
drwxr-xr-x   - ec2-user supergroup          0 2024-02-03 23:32 /data/output2
drwxr-xr-x   - ec2-user supergroup          0 2024-02-03 23:34 /data/output3
drwxr-xr-x   - ec2-user supergroup          0 2024-02-03 23:39 /data/output5
[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-04 00:00:34
$ hadoop fs -cat /data/output02/part-00000 | less

[4]+  Stopped                 hadoop fs -cat /data/output02/part-00000 | less
[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-04 00:02:20
$
```

i-062d65a3dd2e0d019 (nachiketh_server)

PublicIPs: 54.160.108.70 PrivateIPs: 172.31.62.135



```
lo_quantity|lo_linenumber
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
10|1
:
```

i-062d65a3dd2e0d019 (nachiketh_server)

PublicIPs: 54.160.108.70 PrivateIPs: 172.31.62.135

Mapper code:

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

for line in sys.stdin:
  columnName = line.strip().split('|')
  lo_linenumber = columnName[1]
  lo_quantity = columnName[8]
  lo_discount = columnName[11]
  lo_tax = columnName[14]

  if int(lo_discount) < 10 and int(lo_tax) > 2:
    print(f"{lo_quantity}|{lo_linenumber}")
```

Reducer Code:

```
#!/usr/bin/python3
import sys
# printing header
print('lo_quantity|lo_linenumber')
# Process input from mapper line by line
for line in sys.stdin:
   print(line.strip())
```

QUESTION 4b:

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

```
$ hadoop jar hadoop-streaming-2.6.4.jar -input /data/part.tbl -output /data/output21 -file mapper.py -file reducer.py -mapper mapper.py
 -reducer reducer.py
24/02/04 01:12:35 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [mapper.py, reducer.py] [] /tmp/streamjob4638767896681317916.jar tmpDir=null
24/02/04 01:12:36 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
24/02/04 01:12:36 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
24/02/04 01:12:36 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
24/02/04 01:12:36 INFO mapred.FileInputFormat: Total input paths to process : 1
24/02/04 01:12:36 INFO mapreduce.JobSubmitter: number of splits:1
24/02/04 01:12:36 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1938148423_0001
24/02/04 01:12:37 INFO mapred.LocalDistributedCacheManager: Localized file:/home/ec2-user/CSC_555_HW2/mapper.py as file:/tmp/hadoop-ec2
-user/mapred/local/1707009157008/mapper.py
24/02/04 01:12:37 INFO mapred.LocalDistributedCacheManager: Localized file:/home/ec2-user/CSC_555_HW2/reducer.py as file:/tmp/hadoop-ec
2-user/mapred/local/1707009157009/reducer.py
24/02/04 01:12:37 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
24/02/04 01:12:37 INFO mapreduce.Job: Running job: job_local1938148423_0001
24/02/04 01:12:37 INFO mapred.LocalJobRunner: OutputCommitter set in config null
24/02/04 01:12:37 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
24/02/04 01:12:37 INFO mapred.LocalJobRunner: Waiting for map tasks
24/02/04 01:12:37 INFO mapred.LocalJobRunner: Starting task: attempt_local1938148423_0001_m_000000_0
24/02/04 01:12:37 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
24/02/04 01:12:37 INFO mapred.MapTask: Processing split: hdfs://localhost/data/part.tbl:0+17139259
24/02/04 01:12:37 INFO mapred.MapTask: numReduceTasks: 1
24/02/04 01:12:37 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
24/02/04 01:12:37 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
24/02/04 01:12:37 INFO mapred.MapTask: soft limit at 83886080
24/02/04 01:12:37 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
24/02/04 01:12:37 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
24/02/04 01:12:37 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
24/02/04 01:12:37 INFO streaming.PipeMapRed: PipeMapRed exec [/home/ec2-user/CSC_555_HW2/./mapper.py]
24/02/04 01:12:37 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
24/02/04 01:12:37 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir
24/02/04 01:12:37 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
24/02/04 01:12:37 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
24/02/04 01:12:37 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.length
24/02/04 01:12:37 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.output.dir
24/02/04 01:12:37 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
24/02/04 01:12:37 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
24/02/04 01:12:37 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
24/02/04 01:12:37 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
24/02/04 01:12:37 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
24/02/04 01:12:37 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
24/02/04 01:12:37 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 01:12:37 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 01:12:37 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 01:12:37 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 01:12:37 INFO streaming.PipeMapRed: Records R/W=1568/1
24/02/04 01:12:37 INFO streaming.PipeMapRed: R/W/S=10000/5200/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 01:12:37 INFO streaming.PipeMapRed: R/W/S=100000/97297/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 01:12:38 INFO streaming.PipeMapRed: R/W/S=200000/198663/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 01:12:38 INFO streaming.PipeMapRed: MRErrorThread done
24/02/04 01:12:38 INFO streaming.PipeMapRed: mapRedFinished
24/02/04 01:12:38 INFO mapred.LocalJobRunner:
24/02/04 01:12:38 INFO mapred.MapTask: Starting flush of map output
24/02/04 01:12:38 INFO mapred.MapTask: Spilling map output
24/02/04 01:12:38 INFO mapred.MapTask: bufstart = 0; bufend = 6119946; bufvoid = 104857600
24/02/04 01:12:38 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 25414400(101657600); length = 799997/6553600
24/02/04 01:12:38 INFO mapreduce.Job: Job job_local1938148423_0001 running in uber mode : false
```

i-062d65a3dd2e0d019 (nachiketh_server)

PublicIPs: 54.160.108.70 PrivateIPs: 172.31.62.135

aws Services Search [Alt+S] N. Virginia voclabs/user3013569=nparamah@depaul.edu @ 2854-6811-230

```
                Reduce input records=200000
                Reduce output records=26
                Spilled Records=400000
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=6
                CPU time spent (ms)=0
                Physical memory (bytes) snapshot=0
                Virtual memory (bytes) snapshot=0
                Total committed heap usage (bytes)=599785472
        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=17139259
        File Output Format Counters
                Bytes Written=376
24/02/04 01:18:22 INFO streaming.StreamJob: Output directory: /data/output22
[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-04 01:18:23
$ hadoop fs -cat /data/output22/part-00000 | less

[6]+  Stopped                 hadoop fs -cat /data/output22/part-00000 | less
[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-04 01:20:17
$ hadoop fs -cat /data/output22/part-00000 | less

[7]+  Stopped                 hadoop fs -cat /data/output22/part-00000 | less
[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-04 01:22:53
$ hadoop fs -cat /data/output22/part-00000
p_category|p_type(count)
MFGR#11|8158
MFGR#12|7883
MFGR#13|8086
MFGR#14|8029
MFGR#15|7947
MFGR#21|7901
MFGR#22|7982
MFGR#23|8139
MFGR#24|8000
MFGR#25|7920
MFGR#31|7966
MFGR#32|8039
MFGR#33|7986
MFGR#34|7964
MFGR#35|8148
MFGR#41|8092
MFGR#42|8021
MFGR#43|7962
MFGR#44|7814
MFGR#45|7964
MFGR#51|8041
MFGR#52|7934
MFGR#53|7959
MFGR#54|8152
MFGR#55|7913
[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-04 01:23:01
$
```

i-062d65a3dd2e0d019 (nachiketh_server)

PublicIPs: 54.160.108.70 PrivateIPs: 172.31.62.135

```
p_category|p_type(count)
MFGR#11|8158
MFGR#12|7883
MFGR#13|8086
MFGR#14|8029
MFGR#15|7947
MFGR#21|7901
MFGR#22|7982
MFGR#23|8139
MFGR#24|8000
MFGR#25|7920
MFGR#31|7966
MFGR#32|8039
MFGR#33|7986
MFGR#34|7964
MFGR#35|8148
MFGR#41|8092
MFGR#42|8021
MFGR#43|7962
:
```

i-062d65a3dd2e0d019 (nachiketh_server)

PublicIPs: 54.160.108.70 PrivateIPs: 172.31.62.135

Mapper Code:

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

for line in sys.stdin:
    columnName = line.strip().split('|')
    p_category = columnName[3]
    p_type = columnName[6]
    print(f"{p_category}|{p_type}")
```

Reducer Code:

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

# initialize a counter dictionary that will store count of p_type for each p_category
count = {}
print("p_category|p_type(count)")
# Process input from mapper line by line
for line in sys.stdin:
    p_category, p_type = line.strip().split('|')
    if p_category in count:
        count[p_category] += 1
    else:
        count[p_category] = 1
for p_category, cat_count in count.items():
    print(f"{p_category}|{cat_count}")
```

QUESTION 4c:

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

```
$ ls
hadoop-streaming-2.6.4.jar  lineorder.tbl  mapper.py  part.tbl  reducer.py  supplier.tbl
[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-04 04:30:13
$ nano reducer.py
[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-04 04:34:06
$ hadoop fs -put lineorder.tbl /data/
put: `/data/': No such file or directory
[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-04 04:34:41
$ hadoop fs -mkdir /data
[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-04 04:35:07
$ hadoop fs -put lineorder.tbl /data/
[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-04 04:35:32
$ 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 red
ucer.py
24/02/04 04:36:59 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [mapper.py, reducer.py] [] /tmp/streamjob9086976362438202848.jar tmpDir=null
24/02/04 04:36:59 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
24/02/04 04:36:59 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
24/02/04 04:36:59 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
24/02/04 04:36:59 INFO mapred.FileInputFormat: Total input paths to process : 1
24/02/04 04:37:00 INFO mapreduce.JobSubmitter: number of splits:5
24/02/04 04:37:00 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1410278700_0001
24/02/04 04:37:00 INFO mapred.LocalDistributedCacheManager: Localized file:/home/ec2-user/CSC_555_HW2/mapper.py as file:/tmp/hadoop-ec2-user/mapred/local
/1707021420358/mapper.py
24/02/04 04:37:00 INFO mapred.LocalDistributedCacheManager: Localized file:/home/ec2-user/CSC_555_HW2/reducer.py as file:/tmp/hadoop-ec2-user/mapred/loca
l/1707021420359/reducer.py
24/02/04 04:37:00 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
24/02/04 04:37:00 INFO mapreduce.Job: Running job: job_local1410278700_0001
24/02/04 04:37:00 INFO mapred.LocalJobRunner: OutputCommitter set in config null
24/02/04 04:37:00 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
24/02/04 04:37:00 INFO mapred.LocalJobRunner: Waiting for map tasks
24/02/04 04:37:00 INFO mapred.LocalJobRunner: Starting task: attempt_local1410278700_0001_m_000000_0
24/02/04 04:37:00 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
24/02/04 04:37:00 INFO mapred.MapTask: Processing split: hdfs://localhost/data/lineorder.tbl:0+134217728
24/02/04 04:37:00 INFO mapred.MapTask: numReduceTasks: 1
24/02/04 04:37:00 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
24/02/04 04:37:00 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
24/02/04 04:37:00 INFO mapred.MapTask: soft limit at 83886080
24/02/04 04:37:00 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
24/02/04 04:37:00 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
24/02/04 04:37:00 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
24/02/04 04:37:00 INFO streaming.PipeMapRed: PipeMapRed exec [/home/ec2-user/CSC_555_HW2/./mapper.py]
24/02/04 04:37:00 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
24/02/04 04:37:00 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir
24/02/04 04:37:00 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
24/02/04 04:37:00 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
24/02/04 04:37:00 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.length
24/02/04 04:37:00 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.output.dir
24/02/04 04:37:00 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
24/02/04 04:37:00 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
24/02/04 04:37:00 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
24/02/04 04:37:00 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
24/02/04 04:37:00 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
24/02/04 04:37:00 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
24/02/04 04:37:00 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 04:37:00 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 04:37:00 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 04:37:00 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 04:37:00 INFO streaming.PipeMapRed: Records R/W=1376/1
24/02/04 04:37:00 INFO streaming.PipeMapRed: R/W/S=10000/1609/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 04:37:01 INFO streaming.PipeMapRed: R/W/S=100000/98627/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 04:37:01 INFO streaming.PipeMapRed: R/W/S=200000/198612/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 04:37:01 INFO mapreduce.Job: Job job_local1410278700_0001 running in uber mode : false
24/02/04 04:37:01 INFO mapreduce.Job:  map 0% reduce 0%
24/02/04 04:37:01 INFO streaming.PipeMapRed: R/W/S=300000/298435/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 04:37:01 INFO streaming.PipeMapRed: R/W/S=400000/397818/0 in:400000=400000/1 [rec/s] out:397818=397818/1 [rec/s]
```

i-062d65a3dd2e0d019 (nachiketh_server)

PublicIPs: 18.233.65.225 PrivateIPs: 172.31.62.135

aws Services Search [Alt+S] N. Virginia voclabs/user3013569=nparamah@depaul.edu @ 2854-6811-23

```
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=2530877010
                HDFS: Number of bytes written=273
                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=6001215
                Map output bytes=65786684
                Map output materialized bytes=77789144
                Input split bytes=435
                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 Records=12002430
                Shuffled Maps =5
                Failed Shuffles=0
                Merged Map outputs=5
                GC time elapsed (ms)=215
                CPU time spent (ms)=0
                Physical memory (bytes) snapshot=0
                Virtual memory (bytes) snapshot=0
                Total committed heap usage (bytes)=2476212224
        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=273
24/02/04 04:37:31 INFO streaming.StreamJob: Output directory: /data/output31
[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-04 04:37:32
$ hadoop fs -ls /data
Found 2 items
-rw-r--r--   3 ec2-user supergroup  594313001 2024-02-04 04:35 /data/lineorder.tbl
drwxr-xr-x   - ec2-user supergroup          0 2024-02-04 04:37 /data/output31
[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-04 04:38:41
$ hadoop fs -cat /data/output31/part-00000 | less

[1]+  Stopped                 hadoop fs -cat /data/output31/part-00000 | less
[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-04 04:40:55
$ hadoop fs -cat /data/output31/part-00000
lo_discount|lo_extendedprice(avg)
0|3829093.534080523
10|3820012.2906442657
1|3825221.6960687684
2|3825348.6166251353
3|3830409.842713917
4|3823516.7737106928
5|3827676.635869655
6|3826467.937980072
7|3828488.6385758123
8|3821327.8374953885
9|3823085.546772564
[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-04 04:41:13
$
```

i-062d65a3dd2e0d019 (nachiketh_server)

PublicIPs: 18.233.65.225 PrivateIPs: 172.31.62.135

Mapper code:

```
#!/usr/bin/python3
import sys
for line in sys.stdin:
    columnName = line.strip().split('|')
    lo_discount = columnName[11]
    lo_extendedprice = columnName[9]
    print(f"{lo_discount}|{lo_extendedprice}")
```

Reducer Code:

```
#!/usr/bin/python3
import sys
# Initialize a counter dictionary to store sum and count of lo_extendedprice for each lo_discount
count = {}
#printing header
print("lo_discount|lo_extendedprice(avg)")
# Process input from mapper line by line
for line in sys.stdin:
    lo_discount, lo_extendedprice = line.strip().split('|')
    lo_extendedprice = float(lo_extendedprice)

    if lo_discount in count:
        count[lo_discount][0] += lo_extendedprice
        count[lo_discount][1] += 1
    else:
        count[lo_discount] = [lo_extendedprice, 1]

for lo_discount, (total, count_disc) in count.items():
    avg = total / count_disc
    print(f"{lo_discount}|{avg}")
```

QUESTION 4d:

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

```
aws  Services  Search [Alt+S]  N. Virginia ▾  voclabs/user3013569=nparamah@depaul.edu @ 2854-6811-23
hadoop-streaming-2.6.4.jar  lineorder.tbl  mapper.py  part.tbl  reducer.py  supplier.tbl
[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-04 05:27:34
$ hadoop jar hadoop-streaming-2.6.4.jar -input /data/lineorder.tbl -output /data/output42 -file mapper.py -file reducer.py -mapper mapper.py -reducer red
ucer.py
24/02/04 05:27:56 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [mapper.py, reducer.py] [] /tmp/streamjob7353542986826524830.jar tmpDir=null
24/02/04 05:27:57 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
24/02/04 05:27:57 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
24/02/04 05:27:57 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
24/02/04 05:27:57 INFO mapred.FileInputFormat: Total input paths to process : 1
24/02/04 05:27:57 INFO mapreduce.JobSubmitter: number of splits:5
24/02/04 05:27:57 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1066191393_0001
24/02/04 05:27:58 INFO mapred.LocalDistributedCacheManager: Localized file:/home/ec2-user/CSC_555_HW2/mapper.py as file:/tmp/hadoop-ec2-user/mapred/local
/1707024477991/mapper.py
24/02/04 05:27:58 INFO mapred.LocalDistributedCacheManager: Localized file:/home/ec2-user/CSC_555_HW2/reducer.py as file:/tmp/hadoop-ec2-user/mapred/loca
l/1707024477992/reducer.py
24/02/04 05:27:58 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
24/02/04 05:27:58 INFO mapreduce.Job: Running job: job_local1066191393_0001
24/02/04 05:27:58 INFO mapred.LocalJobRunner: OutputCommitter set in config null
24/02/04 05:27:58 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
24/02/04 05:27:58 INFO mapred.LocalJobRunner: Waiting for map tasks
24/02/04 05:27:58 INFO mapred.LocalJobRunner: Starting task: attempt_local1066191393_0001_m_000000_0
24/02/04 05:27:58 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
24/02/04 05:27:58 INFO mapred.MapTask: Processing split: hdfs://localhost/data/lineorder.tbl:0+134217728
24/02/04 05:27:58 INFO mapred.MapTask: numReduceTasks: 1
24/02/04 05:27:58 INFO mapred.MapTask:  (EQUATOR) 0 kvi 26214396(104857584)
24/02/04 05:27:58 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
24/02/04 05:27:58 INFO mapred.MapTask: soft limit at 83886080
24/02/04 05:27:58 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
24/02/04 05:27:58 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
24/02/04 05:27:58 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
24/02/04 05:27:58 INFO streaming.PipeMapRed: PipeMapRed exec [/home/ec2-user/CSC_555_HW2/./mapper.py]
24/02/04 05:27:58 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
24/02/04 05:27:58 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir
24/02/04 05:27:58 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
24/02/04 05:27:58 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
24/02/04 05:27:58 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.length
24/02/04 05:27:58 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.output.dir
24/02/04 05:27:58 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
24/02/04 05:27:58 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
24/02/04 05:27:58 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
24/02/04 05:27:58 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
24/02/04 05:27:58 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
24/02/04 05:27:58 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
24/02/04 05:27:58 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 05:27:58 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 05:27:58 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 05:27:58 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 05:27:58 INFO streaming.PipeMapRed: Records R/W=3060/1
24/02/04 05:27:58 INFO streaming.PipeMapRed: R/W/S=10000/1884/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 05:27:58 INFO streaming.PipeMapRed: R/W/S=100000/21351/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 05:27:58 INFO streaming.PipeMapRed: R/W/S=200000/43337/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 05:27:59 INFO streaming.PipeMapRed: R/W/S=300000/65310/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 05:27:59 INFO mapreduce.Job: Job job_local1066191393_0001 running in uber mode : false
24/02/04 05:27:59 INFO mapreduce.Job:  map 0% reduce 0%
24/02/04 05:27:59 INFO streaming.PipeMapRed: R/W/S=400000/86654/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 05:27:59 INFO streaming.PipeMapRed: R/W/S=500000/108632/0 in:500000=500000/1 [rec/s] out:108632=108632/1 [rec/s]
24/02/04 05:27:59 INFO streaming.PipeMapRed: R/W/S=600000/130592/0 in:600000=600000/1 [rec/s] out:130592=130592/1 [rec/s]
24/02/04 05:27:59 INFO streaming.PipeMapRed: R/W/S=700000/152543/0 in:700000=700000/1 [rec/s] out:152543=152543/1 [rec/s]
24/02/04 05:28:00 INFO streaming.PipeMapRed: R/W/S=800000/174510/0 in:800000=800000/1 [rec/s] out:174510=174510/1 [rec/s]
24/02/04 05:28:00 INFO streaming.PipeMapRed: R/W/S=900000/195867/0 in:900000=900000/1 [rec/s] out:195867=195867/1 [rec/s]
24/02/04 05:28:00 INFO streaming.PipeMapRed: R/W/S=1000000/217841/0 in:1000000=1000000/1 [rec/s] out:217841=217841/1 [rec/s]
24/02/04 05:28:00 INFO streaming.PipeMapRed: R/W/S=1100000/239803/0 in:550000=1100000/2 [rec/s] out:119901=239803/2 [rec/s]
24/02/04 05:28:00 INFO streaming.PipeMapRed: R/W/S=1200000/261786/0 in:600000=1200000/2 [rec/s] out:130893=261786/2 [rec/s]
24/02/04 05:28:00 INFO streaming.PipeMapRed: R/W/S=1300000/283751/0 in:650000=1300000/2 [rec/s] out:141875=283751/2 [rec/s]
24/02/04 05:28:00 INFO streaming.PipeMapRed: MRErrorThread done

i-062d65a3dd2e0d019 (nachiketh_server)
PublicIPs: 18.233.65.225  PrivateIPs: 172.31.62.135
```

aws Services Search [Alt+S] N. Virginia voclabs/user3013569=nparamah@depaul.edu @ 2854-6811-230

```
24/02/04 05:28:13 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1066191393_0001_r_000000_0' to hdfs://localhost/data/output42/_temp
orary/0/task_local1066191393_0001_r_000000
24/02/04 05:28:13 INFO mapred.LocalJobRunner: Records R/W=1318492/1 > reduce
24/02/04 05:28:13 INFO mapred.Task: Task 'attempt_local1066191393_0001_r_000000_0' done.
24/02/04 05:28:13 INFO mapred.LocalJobRunner: Finishing task: attempt_local1066191393_0001_r_000000_0
24/02/04 05:28:13 INFO mapred.LocalJobRunner: reduce task executor complete.
24/02/04 05:28:13 INFO mapreduce.Job:  map 100% reduce 100%
24/02/04 05:28:13 INFO mapreduce.Job: Job job_local1066191393_0001 completed successfully
24/02/04 05:28:13 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=42319578
                FILE: Number of bytes written=112867113
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=2530877010
                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=18512824
                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
                GC time elapsed (ms)=93
                CPU time spent (ms)=0
                Physical memory (bytes) snapshot=0
                Virtual memory (bytes) snapshot=0
                Total committed heap usage (bytes)=2583166976
        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=313156
24/02/04 05:28:13 INFO streaming.StreamJob: Output directory: /data/output42
[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-04 05:28:13
$ hadoop fs -ls /data
Found 4 items
-rw-r--r--   3 ec2-user supergroup  594313001 2024-02-04 04:35 /data/lineorder.tbl
drwxr-xr-x   - ec2-user supergroup          0 2024-02-04 04:37 /data/output31
drwxr-xr-x   - ec2-user supergroup          0 2024-02-04 05:23 /data/output41
drwxr-xr-x   - ec2-user supergroup          0 2024-02-04 05:28 /data/output42
[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-04 05:29:33
$ hadoop fs -cat /data/output42/part-00000 | less

[3]+  Stopped                 hadoop fs -cat /data/output42/part-00000 | less
[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-04 05:30:30
$
```

i-062d65a3dd2e0d019 (nachiketh_server)

PublicIPs: 18.233.65.225 PrivateIPs: 172.31.62.135

aws Services Search [Alt+S] N. Virginia voclabs/user3013569=nparamah@depaul.edu @ 2854-6811-230

```
lo_custkey|revenue
10000|98534642
10001|54061584
10003|72660665
10004|25007708
10006|86838777
10007|53385453
10009|78503954
1000|85197774
10010|44115429
10012|119204668
10013|45215883
10015|82313185
10016|45338853
10018|77046005
10019|31436012
1001|46912006
10021|57579583
10022|56618633
10024|72271302
10025|41958201
10027|86559157
10028|41262754
10030|76012976
10031|52861420
10033|80167356
10034|41342906
10036|50821597
10037|40383469
10039|63595590
1003|85207956
10040|36517322
10042|88735247
10043|52726368
10045|79831947
10046|62266294
10048|68032990
10049|38911700
1004|37810226
10051|79222997
10052|25829745
10054|103435744
10055|55394758
10057|76262488
10058|43165445
10060|94693497
10061|44829744
10063|97458249
10064|32627879
10066|109396705
10067|39612671
10069|91145200
1006|66493030
10070|38779489
10072|54751838
10073|34453092
10075|91605863
10076|45423551
10078|80433882
10079|46842287
1007|25822264
10081|97878592
10082|47241617
10084|63934992
10085|36168304
:
```

i-062d65a3dd2e0d019 (nachiketh_server)

PublicIPs: 18.233.65.225 PrivateIPs: 172.31.62.135

Mapper code:

```
#!/usr/bin/python3
import sys
for line in sys.stdin:
    columnName = line.strip().split('|')
    lo_custkey = columnName[2]
    lo_quantity = columnName[8]
    lo_extendedprice = columnName[9]

    if int(lo_quantity) <12:
        print(f"{lo_custkey}|{lo_extendedprice}")
```

Reducer code:

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

# Initialize a counter dictionary to store sum and count of lo_extendedprice for each lo_discount
sum_price = {}
#printing header
print("lo_custkey|revenue")
# Process input from mapper line by line
for line in sys.stdin:
    lo_custkey, lo_extendedprice = line.strip().split('|')
    lo_extendedprice = int(lo_extendedprice)

    if lo_custkey in sum_price:
        sum_price[lo_custkey] += lo_extendedprice
    else:
        sum_price[lo_custkey] = lo_extendedprice

for lo_custkey, sum_total in sum_price.items():
    print(f"{lo_custkey}|{sum_total}")
```

QUESTION 4e:

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

```
aws  Services  Search [Alt+S]  N. Virginia ▾  voclabs/user3013569=nparamah@depaul.edu @ 2854-6811-230
[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-04 18:02:20
$ nano reducer.py
[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-04 18:03:01
$ hadoop jar hadoop-streaming-2.6.4.jar -input /data/supplier.tbl,/data/lineorder.tbl -output /data/output59 -file mapper.py -file reducer.py -mapper map
per.py -reducer reducer.py
24/02/04 18:03:17 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [mapper.py, reducer.py] [] /tmp/streamjob8591845795689805296.jar tmpDir=null
24/02/04 18:03:18 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
24/02/04 18:03:18 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
24/02/04 18:03:18 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
24/02/04 18:03:18 INFO mapred.FileInputFormat: Total input paths to process : 2
24/02/04 18:03:18 INFO mapreduce.JobSubmitter: number of splits:6
24/02/04 18:03:18 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1094549684_0001
24/02/04 18:03:18 INFO mapred.LocalDistributedCacheManager: Localized file:/home/ec2-user/CSC_555_HW2/mapper.py as file:/tmp/hadoop-ec2-user/mapred/local
/1707069798551/mapper.py
24/02/04 18:03:18 INFO mapred.LocalDistributedCacheManager: Localized file:/home/ec2-user/CSC_555_HW2/reducer.py as file:/tmp/hadoop-ec2-user/mapred/loca
l/1707069798552/reducer.py
24/02/04 18:03:18 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
24/02/04 18:03:18 INFO mapreduce.Job: Running job: job_local1094549684_0001
24/02/04 18:03:18 INFO mapred.LocalJobRunner: OutputCommitter set in config null
24/02/04 18:03:18 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
24/02/04 18:03:18 INFO mapred.LocalJobRunner: Waiting for map tasks
24/02/04 18:03:18 INFO mapred.LocalJobRunner: Starting task: attempt_local1094549684_0001_m_000000_0
24/02/04 18:03:18 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
24/02/04 18:03:18 INFO mapred.MapTask: Processing split: hdfs://localhost/data/lineorder.tbl:0+134217728
24/02/04 18:03:18 INFO mapred.MapTask: numReduceTasks: 1
24/02/04 18:03:18 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
24/02/04 18:03:18 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
24/02/04 18:03:18 INFO mapred.MapTask: soft limit at 83886080
24/02/04 18:03:18 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
24/02/04 18:03:18 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
24/02/04 18:03:18 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
24/02/04 18:03:18 INFO streaming.PipeMapRed: PipeMapRed exec [/home/ec2-user/CSC_555_HW2/./mapper.py]
24/02/04 18:03:18 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
24/02/04 18:03:18 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir
24/02/04 18:03:18 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
24/02/04 18:03:18 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
24/02/04 18:03:18 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.length
24/02/04 18:03:18 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.output.dir
24/02/04 18:03:18 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
24/02/04 18:03:18 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
24/02/04 18:03:18 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
24/02/04 18:03:18 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
24/02/04 18:03:18 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
24/02/04 18:03:18 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
24/02/04 18:03:18 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 18:03:18 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 18:03:18 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 18:03:18 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 18:03:19 INFO streaming.PipeMapRed: Records R/W=2735/1
24/02/04 18:03:19 INFO streaming.PipeMapRed: R/W/S=10000/7698/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 18:03:19 INFO streaming.PipeMapRed: R/W/S=100000/89107/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 18:03:19 INFO mapreduce.Job: Job job_local1094549684_0001 running in uber mode : false
24/02/04 18:03:19 INFO mapreduce.Job:  map 0% reduce 0%
24/02/04 18:03:19 INFO streaming.PipeMapRed: R/W/S=200000/180363/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 18:03:20 INFO streaming.PipeMapRed: R/W/S=300000/271665/0 in:300000=300000/1 [rec/s] out:271665=271665/1 [rec/s]
24/02/04 18:03:20 INFO streaming.PipeMapRed: R/W/S=400000/360738/0 in:400000=400000/1 [rec/s] out:360738=360738/1 [rec/s]
24/02/04 18:03:20 INFO streaming.PipeMapRed: R/W/S=500000/451999/0 in:500000=500000/1 [rec/s] out:451999=451999/1 [rec/s]
24/02/04 18:03:20 INFO streaming.PipeMapRed: R/W/S=600000/543289/0 in:600000=600000/1 [rec/s] out:543289=543289/1 [rec/s]
24/02/04 18:03:21 INFO streaming.PipeMapRed: R/W/S=700000/633996/0 in:350000=700000/2 [rec/s] out:316998=633996/2 [rec/s]
24/02/04 18:03:21 INFO streaming.PipeMapRed: R/W/S=800000/724781/0 in:400000=800000/2 [rec/s] out:362390=724781/2 [rec/s]
24/02/04 18:03:21 INFO streaming.PipeMapRed: R/W/S=900000/816059/0 in:450000=900000/2 [rec/s] out:408029=816059/2 [rec/s]
24/02/04 18:03:21 INFO streaming.PipeMapRed: R/W/S=1000000/906429/0 in:500000=1000000/2 [rec/s] out:453214=906429/2 [rec/s]
24/02/04 18:03:22 INFO streaming.PipeMapRed: R/W/S=1100000/997517/0 in:366666=1100000/3 [rec/s] out:332505=997517/3 [rec/s]
24/02/04 18:03:22 INFO streaming.PipeMapRed: R/W/S=1200000/1088787/0 in:400000=1200000/3 [rec/s] out:362929=1088787/3 [rec/s]
24/02/04 18:03:22 INFO streaming.PipeMapRed: R/W/S=1300000/1178977/0 in:433333=1300000/3 [rec/s] out:392992=1178977/3 [rec/s]
i-062d65a3dd2e0d019 (nachiketh_server)
PublicIPs: 100.25.191.181  PrivateIPs: 172.31.62.135
```

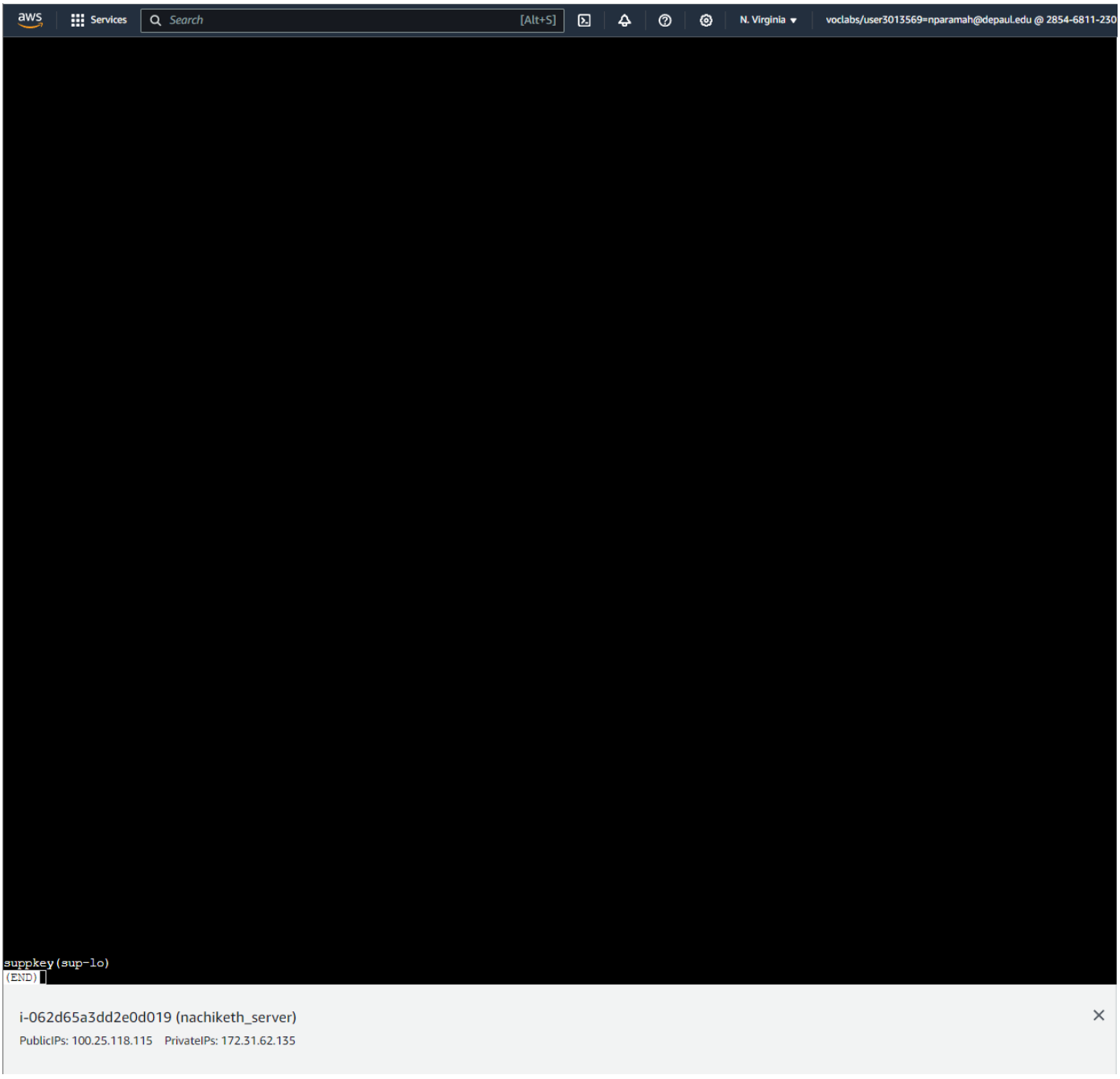
aws Services Search [Alt+S] N. Virginia voclabs/user3013569=nparamah@depaul.edu @ 2854-6811-230

```
24/02/04 18:03:45 INFO streaming.PipeMapRed: R/W/S=5400000/0/0 in:1800000=5400000/3 [rec/s] out:0=0/3 [rec/s]
24/02/04 18:03:45 INFO streaming.PipeMapRed: Records R/W=5457400/1
24/02/04 18:03:45 INFO streaming.PipeMapRed: MRErrorThread done
24/02/04 18:03:45 INFO streaming.PipeMapRed: mapRedFinished
24/02/04 18:03:45 INFO mapred.Task: Task:attempt_local1094549684_0001_r_000000_0 is done. And is in the process of committing
24/02/04 18:03:45 INFO mapred.LocalJobRunner: 6 / 6 copied.
24/02/04 18:03:45 INFO mapred.Task: Task attempt_local1094549684_0001_r_000000_0 is allowed to commit now
24/02/04 18:03:45 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1094549684_0001_r_000000_0' to hdfs://localhost/data/output59/_temp
orary/0/task_local1094549684_0001_r_000000
24/02/04 18:03:45 INFO mapred.LocalJobRunner: Records R/W=5457400/1 > reduce
24/02/04 18:03:45 INFO mapred.Task: Task 'attempt_local1094549684_0001_r_000000_0' done.
24/02/04 18:03:45 INFO mapred.LocalJobRunner: Finishing task: attempt_local1094549684_0001_r_000000_0
24/02/04 18:03:45 INFO mapred.LocalJobRunner: reduce task executor complete.
24/02/04 18:03:45 INFO mapreduce.Job:  map 100% reduce 100%
24/02/04 18:03:45 INFO mapreduce.Job: Job job_local1094549684_0001 completed successfully
24/02/04 18:03:45 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=103132443
                FILE: Number of bytes written=324614813
                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=92
                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=40637510
                Map output materialized bytes=51552346
                Input split bytes=521
                Combine input records=0
                Combine output records=0
                Reduce input groups=2000
                Reduce shuffle bytes=51552346
                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)=103
                CPU time spent (ms)=0
                Physical memory (bytes) snapshot=0
                Virtual memory (bytes) snapshot=0
                Total committed heap usage (bytes)=3196059648
        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=594496061
        File Output Format Counters
                Bytes Written=17
24/02/04 18:03:45 INFO streaming.StreamJob: Output directory: /data/output59
[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-04 18:03:46
$ hadoop fs -cat /data/output59/part-00000 | less

[5]+  Stopped                 hadoop fs -cat /data/output59/part-00000 | less
[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-04 18:04:41
$
```

i-062d65a3dd2e0d019 (nachiketh_server)

PublicIPs: 100.25.191.181 PrivateIPs: 172.31.62.135



Note: The output is empty because all elements in s_suppkey are also found in lo_suppkey.

Mapper code:

```
#!/usr/bin/python3
import sys
import os
input_file = os.environ['map_input_file']

if input_file.endswith("supplier.tbl"):
  for line in sys.stdin:
      s_suppkey = line.strip().split('|')[0]
      print(f"{s_suppkey}\tsup")

else:
   for line in sys.stdin:
      lo_suppkey = line.strip().split('|')[4]
      lo_discount = line.strip().split('|')[11]
      if int(lo_discount) < 10:
         print(f"{lo_suppkey}\tlo")
```

Reducer code:

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

lo_suppkeys = set()
s_suppkeys = set()
# header
print("suppkey(sup-lo)")
# Process input from mapper line by line
for line in sys.stdin:
   suppkey, flag = line.strip().split('\t')
   if flag == "lo":
      lo_suppkeys.add(suppkey)
   elif flag == "sup":
      s_suppkeys.add(suppkey)

# performing set difference operation
result = s_suppkeys - lo_suppkeys

# printing the final result
for lo_suppkey in result:
   print(lo_suppkey)
```

QUESTION 5d:

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

```
aws  Services  Search [Alt+S]  N. Virginia  voclabs/user3013569=nparamah@depaul.edu @ 2854-6811-230
Last login: Sun Feb  4 18:36:34 2024 from 18.206.107.28
[ec2-user@ip-172-31-62-135 ~] 2024-02-04 19:34:44
$ pwd
/home/ec2-user
[ec2-user@ip-172-31-62-135 ~] 2024-02-04 19:34:48
$ cd CSC_555_HW2
[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-04 19:34:54
$ ls
combiner.py  hadoop-streaming-2.6.4.jar  lineorder.tbl  mapper.py  part.tbl  reducer.py  supplier.tbl
[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-04 19:34:56
$ nano reducer.py
[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-04 19:37:28
$ hadoop jar hadoop-streaming-2.6.4.jar -input /data/lineorder.tbl -output /data/output62 -file mapper.py -file combiner.py -file reducer.py -mapper mapp
er.py -combiner combiner.py -reducer reducer.py
24/02/04 19:38:03 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [mapper.py, combiner.py, reducer.py] [] /tmp/streamjob5163254229467769956.jar tmpDir=null
24/02/04 19:38:04 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
24/02/04 19:38:04 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
24/02/04 19:38:04 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
24/02/04 19:38:04 INFO mapred.FileInputFormat: Total input paths to process : 1
24/02/04 19:38:04 INFO mapreduce.JobSubmitter: number of splits:5
24/02/04 19:38:04 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local426872266_0001
24/02/04 19:38:05 INFO mapred.LocalDistributedCacheManager: Localized file:/home/ec2-user/CSC_555_HW2/mapper.py as file:/tmp/hadoop-ec2-user/mapred/local
/1707075484946/mapper.py
24/02/04 19:38:05 INFO mapred.LocalDistributedCacheManager: Localized file:/home/ec2-user/CSC_555_HW2/combiner.py as file:/tmp/hadoop-ec2-user/mapred/loc
al/1707075484947/combiner.py
24/02/04 19:38:05 INFO mapred.LocalDistributedCacheManager: Localized file:/home/ec2-user/CSC_555_HW2/reducer.py as file:/tmp/hadoop-ec2-user/mapred/loca
l/1707075484948/reducer.py
24/02/04 19:38:05 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
24/02/04 19:38:05 INFO mapreduce.Job: Running job: job_local426872266_0001
24/02/04 19:38:05 INFO mapred.LocalJobRunner: OutputCommitter set in config null
24/02/04 19:38:05 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
24/02/04 19:38:05 INFO mapred.LocalJobRunner: Waiting for map tasks
24/02/04 19:38:05 INFO mapred.LocalJobRunner: Starting task: attempt_local426872266_0001_m_000000_0
24/02/04 19:38:05 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
24/02/04 19:38:05 INFO mapred.MapTask: Processing split: hdfs://localhost/data/lineorder.tbl:0+134217728
24/02/04 19:38:05 INFO mapred.MapTask: numReduceTasks: 1
24/02/04 19:38:05 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
24/02/04 19:38:05 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
24/02/04 19:38:05 INFO mapred.MapTask: soft limit at 83886080
24/02/04 19:38:05 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
24/02/04 19:38:05 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
24/02/04 19:38:05 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
24/02/04 19:38:05 INFO streaming.PipeMapRed: PipeMapRed exec [/home/ec2-user/CSC_555_HW2/./mapper.py]
24/02/04 19:38:05 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
24/02/04 19:38:05 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir
24/02/04 19:38:05 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
24/02/04 19:38:05 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
24/02/04 19:38:05 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.length
24/02/04 19:38:05 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.output.dir
24/02/04 19:38:05 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
24/02/04 19:38:05 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
24/02/04 19:38:05 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
24/02/04 19:38:05 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
24/02/04 19:38:05 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
24/02/04 19:38:05 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
24/02/04 19:38:05 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 19:38:05 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 19:38:05 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 19:38:05 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 19:38:05 INFO streaming.PipeMapRed: Records R/W=2927/1
24/02/04 19:38:05 INFO streaming.PipeMapRed: R/W/S=10000/1884/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 19:38:05 INFO streaming.PipeMapRed: R/W/S=100000/21351/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 19:38:05 INFO streaming.PipeMapRed: R/W/S=200000/43337/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 19:38:06 INFO streaming.PipeMapRed: R/W/S=300000/65310/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 19:38:06 INFO mapreduce.Job: Job job_local426872266_0001 running in uber mode : false

i-062d65a3dd2e0d019 (nachiketh_server)
PublicIPs: 100.25.191.181  PrivateIPs: 172.31.62.135
```

aws Services Search [Alt+S] N. Virginia voclabs/user3013569=nparamah@depaul.edu @ 2854-6811-230

```
24/02/04 19:38:19 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 19:38:19 INFO streaming.PipeMapRed: Records R/W=99305/1
24/02/04 19:38:19 INFO streaming.PipeMapRed: MRErrorThread done
24/02/04 19:38:19 INFO streaming.PipeMapRed: mapRedFinished
24/02/04 19:38:19 INFO mapred.Task: Task:attempt_local426872266_0001_r_000000_0 is done. And is in the process of committing
24/02/04 19:38:19 INFO mapred.LocalJobRunner: 5 / 5 copied.
24/02/04 19:38:19 INFO mapred.Task: Task attempt_local426872266_0001_r_000000_0 is allowed to commit now
24/02/04 19:38:19 INFO output.FileOutputCommitter: Saved output of task 'attempt_local426872266_0001_r_000000_0' to hdfs://localhost/data/output62/_tempo
rary/0/task_local426872266_0001_r_000000
24/02/04 19:38:19 INFO mapred.LocalJobRunner: Records R/W=99305/1 > reduce
24/02/04 19:38:19 INFO mapred.Task: Task 'attempt_local426872266_0001_r_000000_0' done.
24/02/04 19:38:19 INFO mapred.LocalJobRunner: Finishing task: attempt_local426872266_0001_r_000000_0
24/02/04 19:38:19 INFO mapred.LocalJobRunner: reduce task executor complete.
24/02/04 19:38:20 INFO mapreduce.Job:  map 100% reduce 100%
24/02/04 19:38:20 INFO mapreduce.Job: Job job_local426872266_0001 completed successfully
24/02/04 19:38:20 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=3432830
                FILE: Number of bytes written=10174011
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=2530877010
                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=18512824
                Map output materialized bytes=1704373
                Input split bytes=435
                Combine input records=1318492
                Combine output records=99305
                Reduce input groups=99305
                Reduce shuffle bytes=1704373
                Reduce input records=99305
                Reduce output records=20001
                Spilled Records=198610
                Shuffled Maps =5
                Failed Shuffles=0
                Merged Map outputs=5
                GC time elapsed (ms)=87
                CPU time spent (ms)=0
                Physical memory (bytes) snapshot=0
                Virtual memory (bytes) snapshot=0
                Total committed heap usage (bytes)=2604662784
        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=313156
24/02/04 19:38:20 INFO streaming.StreamJob: Output directory: /data/output62
[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-04 19:38:20
$ hadoop fs -cat /data/output62/part-00000 | less

[1]+  Stopped                 hadoop fs -cat /data/output62/part-00000 | less
[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-04 19:39:27
$
```

i-062d65a3dd2e0d019 (nachiketh_server)

PublicIPs: 100.25.191.181 PrivateIPs: 172.31.62.135

aws Services Search [Alt+S] N. Virginia voclabs/user3013569=nparamah@depaul.edu @ 2854-6811-230

```
lo_custkey|revenue
10000|98534642
10001|54061584
10003|72660665
10004|25007708
10006|86838777
10007|53385453
10009|78503954
1000|85197774
10010|44115429
10012|119204668
10013|45215883
10015|82313185
10016|45338853
10018|77046005
10019|31436012
1001|46912006
10021|57579583
10022|56618633
10024|72271302
10025|41958201
10027|86559157
10028|41262754
10030|76012976
10031|52861420
10033|80167356
10034|41342906
10036|50821597
10037|40383469
10039|63595590
1003|85207956
10040|36517322
10042|88735247
10043|52726368
10045|79831947
10046|62266294
10048|68032990
10049|38911700
1004|37810226
10051|79222997
10052|25829745
10054|103435744
10055|55394758
10057|76262488
10058|43165445
10060|94693497
10061|44829744
10063|97458249
10064|32627879
10066|109396705
10067|39612671
10069|91145200
1006|66493030
10070|38779489
10072|54751838
10073|34453092
10075|91605863
10076|45423551
10078|80433882
10079|46842287
1007|25822264
10081|97878592
10082|47241617
10084|63934992
10085|36168304
:
```

i-062d65a3dd2e0d019 (nachiketh_server)

PublicIPs: 100.25.191.181 PrivateIPs: 172.31.62.135

Mapper code:

```
#!/usr/bin/python3
import sys
for line in sys.stdin:
    columns = line.strip().split('|')
    lo_custkey = columns[2]
    lo_quantity = columns[8]
    lo_extendedprice = columns[9]

    if int(lo_quantity) < 12:
        print(f"{lo_custkey}|{lo_extendedprice}")
```

Combiner Code:

```
#!/usr/bin/python3
import sys
sum_price = {}
for line in sys.stdin:
    lo_custkey, lo_extendedprice = line.strip().split('|')
    lo_extendedprice = int(lo_extendedprice)

    if lo_custkey in sum_price:
        sum_price[lo_custkey] += lo_extendedprice
    else:
        sum_price[lo_custkey] = lo_extendedprice
for lo_custkey, sum_total in sum_price.items():
    print(f"{lo_custkey}|{sum_total}")
```

Reducer code:

```
#!/usr/bin/python3
import sys
print("lo_custkey|revenue")
# Initialize a counter dictionary to store sum of lo_extendedprice for each lo_custkey
sum_price = {}

# Pprocessing input from mapper & combiner
for line in sys.stdin:
    lo_custkey, lo_extendedprice = line.strip().split('|')
    lo_extendedprice = int(lo_extendedprice)
    if lo_custkey in sum_price:
        sum_price[lo_custkey] += lo_extendedprice
    else:
        sum_price[lo_custkey] = lo_extendedprice
for lo_custkey, sum_total in sum_price.items():
    print(f"{lo_custkey}|{sum_total}")
```

QUESTION 5e:

```
[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-04 21:26:13
$ ls
combiner.py  hadoop-streaming-2.6.4.jar  lineorder.tbl  mapper.py  part.tbl  reducer.py  supplier.tbl
[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-04 21:26:17
$ hadoop fs -ls /data
Found 3 items
-rw-r--r--   3 ec2-user supergroup  594313001 2024-02-04 21:07 /data/lineorder.tbl
drwxr-xr-x   - ec2-user supergroup          0 2024-02-04 21:18 /data/output65
-rw-r--r--   3 ec2-user supergroup     166676 2024-02-04 21:08 /data/supplier.tbl
[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-04 21:26:35
$ hadoop jar hadoop-streaming-2.6.4.jar -input /data/supplier.tbl,/data/lineorder.tbl -output /data/output66 -file mapper.py -file combiner.py -file redu
cer.py -mapper mapper.py -combiner combiner.py -reducer reducer.py
24/02/04 21:27:04 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [mapper.py, combiner.py, reducer.py] [] /tmp/streamjob3934338183148672808.jar tmpDir=null
24/02/04 21:27:05 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
24/02/04 21:27:05 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
24/02/04 21:27:05 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
24/02/04 21:27:05 INFO mapred.FileInputFormat: Total input paths to process : 2
24/02/04 21:27:05 INFO mapreduce.JobSubmitter: number of splits:6
24/02/04 21:27:05 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1827315187_0001
24/02/04 21:27:05 INFO mapred.LocalDistributedCacheManager: Localized file:/home/ec2-user/CSC_555_HW2/mapper.py as file:/tmp/hadoop-ec2-user/mapred/local
/1707082025658/mapper.py
24/02/04 21:27:05 INFO mapred.LocalDistributedCacheManager: Localized file:/home/ec2-user/CSC_555_HW2/combiner.py as file:/tmp/hadoop-ec2-user/mapred/loc
al/1707082025659/combiner.py
24/02/04 21:27:05 INFO mapred.LocalDistributedCacheManager: Localized file:/home/ec2-user/CSC_555_HW2/reducer.py as file:/tmp/hadoop-ec2-user/mapred/loca
l/1707082025660/reducer.py
24/02/04 21:27:05 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
24/02/04 21:27:05 INFO mapreduce.Job: Running job: job_local1827315187_0001
24/02/04 21:27:05 INFO mapred.LocalJobRunner: OutputCommitter set in config null
24/02/04 21:27:05 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
24/02/04 21:27:05 INFO mapred.LocalJobRunner: Waiting for map tasks
24/02/04 21:27:05 INFO mapred.LocalJobRunner: Starting task: attempt_local1827315187_0001_m_000000_0
24/02/04 21:27:05 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
24/02/04 21:27:05 INFO mapred.MapTask: Processing split: hdfs://localhost/data/lineorder.tbl:0+134217728
24/02/04 21:27:05 INFO mapred.MapTask: numReduceTasks: 1
24/02/04 21:27:06 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
24/02/04 21:27:06 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
24/02/04 21:27:06 INFO mapred.MapTask: soft limit at 83886080
24/02/04 21:27:06 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
24/02/04 21:27:06 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
24/02/04 21:27:06 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
24/02/04 21:27:06 INFO streaming.PipeMapRed: PipeMapRed exec [/home/ec2-user/CSC_555_HW2/./mapper.py]
24/02/04 21:27:06 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
24/02/04 21:27:06 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir
24/02/04 21:27:06 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
24/02/04 21:27:06 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
24/02/04 21:27:06 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.length
24/02/04 21:27:06 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.output.dir
24/02/04 21:27:06 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
24/02/04 21:27:06 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
24/02/04 21:27:06 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
24/02/04 21:27:06 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
24/02/04 21:27:06 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
24/02/04 21:27:06 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
24/02/04 21:27:06 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 21:27:06 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 21:27:06 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 21:27:06 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 21:27:06 INFO streaming.PipeMapRed: Records R/W=1376/1
24/02/04 21:27:06 INFO streaming.PipeMapRed: R/W/S=10000/7698/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 21:27:06 INFO streaming.PipeMapRed: R/W/S=100000/89107/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 21:27:06 INFO mapreduce.Job: Job job_local1827315187_0001 running in uber mode : false
24/02/04 21:27:06 INFO mapreduce.Job:  map 0% reduce 0%
24/02/04 21:27:06 INFO streaming.PipeMapRed: R/W/S=200000/180363/0 in:NA [rec/s] out:NA [rec/s]
24/02/04 21:27:07 INFO streaming.PipeMapRed: R/W/S=300000/271665/0 in:300000=300000/1 [rec/s] out:271665=271665/1 [rec/s]
24/02/04 21:27:07 INFO streaming.PipeMapRed: R/W/S=400000/360738/0 in:400000=400000/1 [rec/s] out:360738=360738/1 [rec/s]
```

i-062d65a3dd2e0d019 (nachiketh_server)

PublicIPs: 100.25.118.115 PrivateIPs: 172.31.62.135

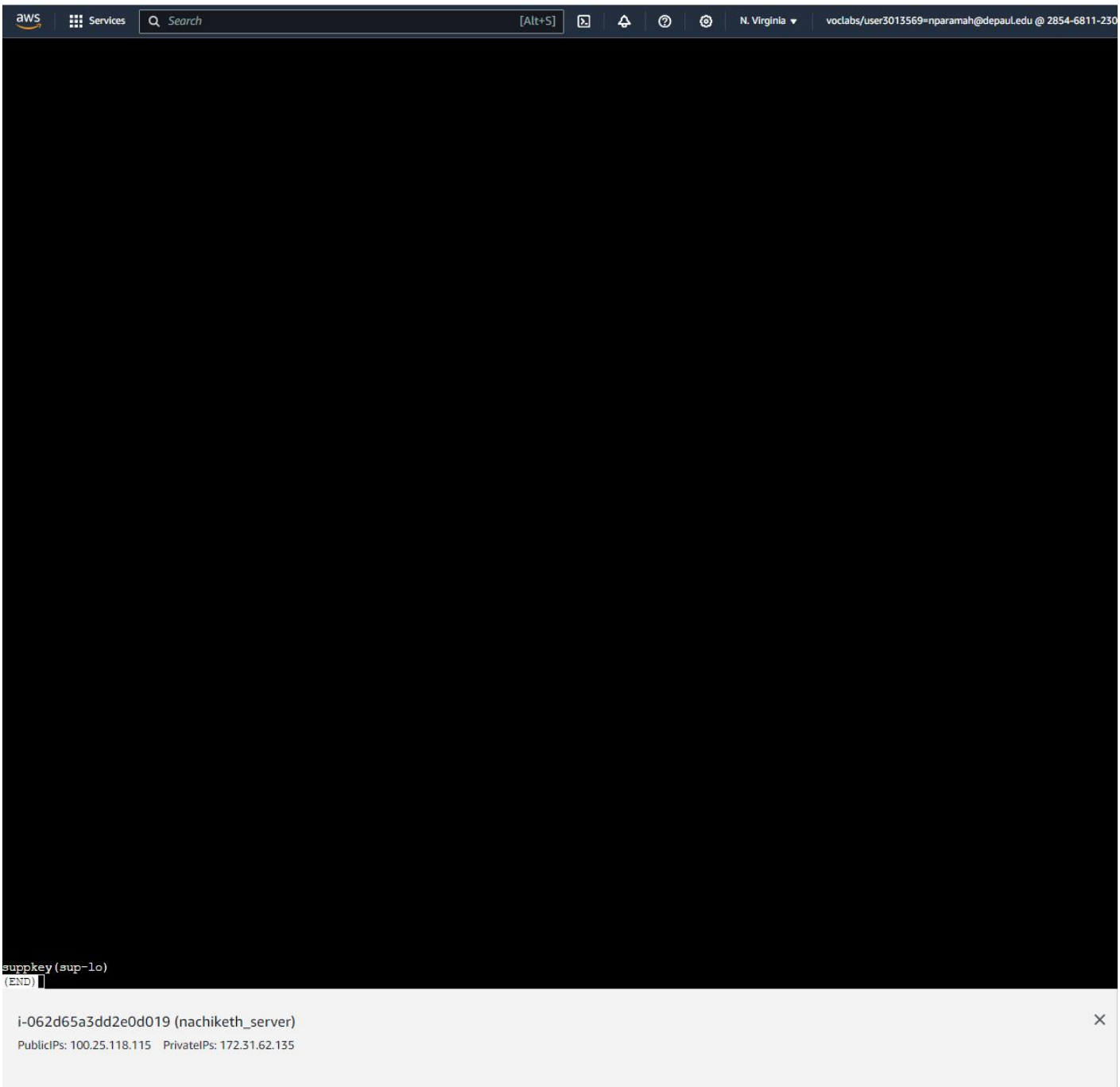
aws Services Search [Alt+S] N. Virginia voclabs/user3013569=nparamah@depaul.edu @ 2854-6811-230

```
24/02/04 21:27:30 INFO streaming.PipeMapRed: mapRedFinished
24/02/04 21:27:30 INFO mapred.Task: Task:attempt_local1827315187_0001_r_000000_0 is done. And is in the process of committing
24/02/04 21:27:30 INFO mapred.LocalJobRunner: 6 / 6 copied.
24/02/04 21:27:30 INFO mapred.Task: Task attempt_local1827315187_0001_r_000000_0 is allowed to commit now
24/02/04 21:27:30 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1827315187_0001_r_000000_0' to hdfs://localhost/data/output66/_temp
orary/0/task_local1827315187_0001_r_000000
24/02/04 21:27:30 INFO mapred.LocalJobRunner: Records R/W=12000/1 > reduce
24/02/04 21:27:30 INFO mapred.Task: Task 'attempt_local1827315187_0001_r_000000_0' done.
24/02/04 21:27:30 INFO mapred.LocalJobRunner: Finishing task: attempt_local1827315187_0001_r_000000_0
24/02/04 21:27:30 INFO mapred.LocalJobRunner: reduce task executor complete.
24/02/04 21:27:30 INFO mapreduce.Job:  map 100% reduce 100%
24/02/04 21:27:30 INFO mapreduce.Job: Job job_local1827315187_0001 completed successfully
24/02/04 21:27:30 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=263208
                FILE: Number of bytes written=2518191
                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=92
                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=40641510
                Map output materialized bytes=115394
                Input split bytes=521
                Combine input records=5457400
                Combine output records=12000
                Reduce input groups=12000
                Reduce shuffle bytes=115394
                Reduce input records=12000
                Reduce output records=1
                Spilled Records=24000
                Shuffled Maps =6
                Failed Shuffles=0
                Merged Map outputs=6
                GC time elapsed (ms)=92
                CPU time spent (ms)=0
                Physical memory (bytes) snapshot=0
                Virtual memory (bytes) snapshot=0
                Total committed heap usage (bytes)=3165650944
        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=594496061
        File Output Format Counters
                Bytes Written=17
24/02/04 21:27:30 INFO streaming.StreamJob: Output directory: /data/output66
[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-04 21:27:31
$ hadoop fs -cat /data/output66/part-00000 | less

[2]+  Stopped                 hadoop fs -cat /data/output66/part-00000 | less
[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-04 21:29:24
$ hadoop fs -cat /data/output66/part-00000
suppkey(sup-lo)
[ec2-user@ip-172-31-62-135 CSC_555_HW2] 2024-02-04 21:30:05
$
```

i-062d65a3dd2e0d019 (nachiketh_server)

PublicIPs: 100.25.118.115 PrivateIPs: 172.31.62.135



Note: The output is empty because all elements in s_suppkey are also found in lo_suppkey.

Mapper code:

```
#!/usr/bin/python3

import sys
import os
for line in sys.stdin:
    input_file = os.environ['map_input_file']
```

```
    if input_file.endswith("supplier.tbl"):
        s_suppkey = line.strip().split('|')[0]
        print(f"{s_suppkey}\tsup")
    else:
        lo_suppkey = line.strip().split('|')[4]
        lo_discount = line.strip().split('|')[11]
        if int(lo_discount) < 10:
            print(f"{lo_suppkey}\tlo")
```

Combiner code:

```
#!/usr/bin/python3
import sys
lo_suppkeys = set()
s_suppkeys = set()
for line in sys.stdin:
    suppkey, flag = line.strip().split('\t')
    if flag == "lo":
        lo_suppkeys.add(suppkey)
    elif flag == "sup":
        s_suppkeys.add(suppkey)

for lo_suppkey in lo_suppkeys:
    print(f"{lo_suppkey}\tlo")
for s_suppkey in s_suppkeys:
    print(f"{s_suppkey}\tsup")
```

Reducer code:

```
#!/usr/bin/python3
import sys
lo_suppkeys = set()
s_suppkeys = set()
for line in sys.stdin:
    suppkey, flag = line.strip().split('\t')
    if flag == "lo":
        lo_suppkeys.add(suppkey)
    elif flag == "sup":
        s_suppkeys.add(suppkey)

result = s_suppkeys - lo_suppkeys
print("suppkey(sup-lo)")
for lo_suppkey in result:
    print(lo_suppkey)
```