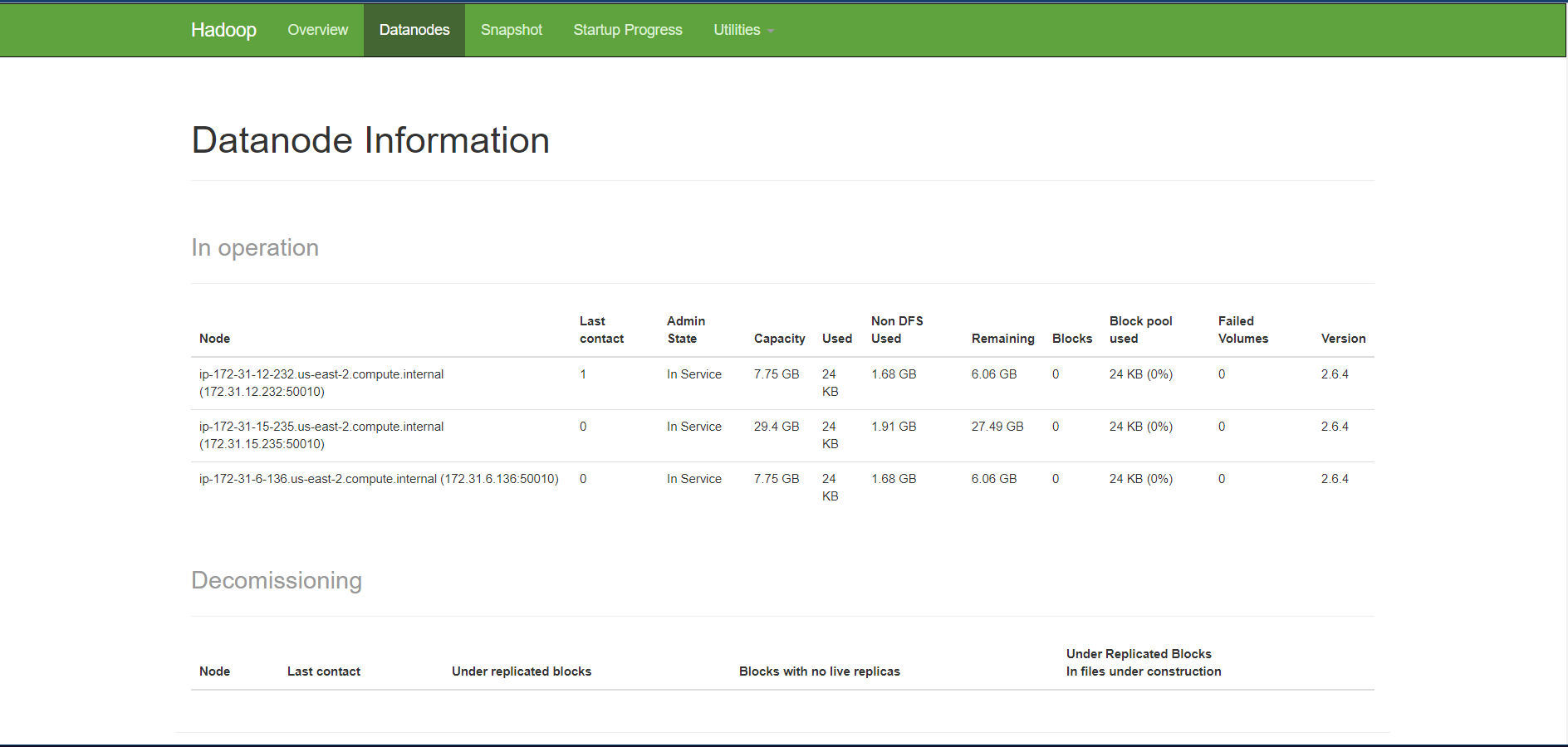
**CSC 555 Phase 1**

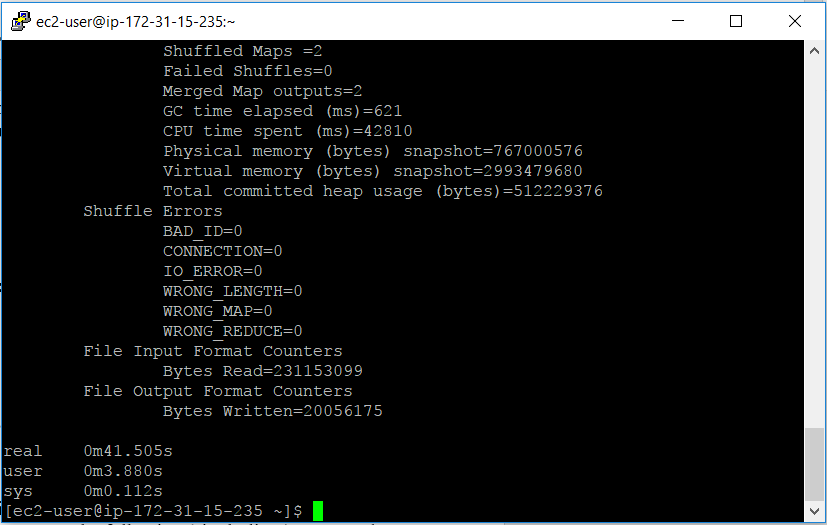
**Yiyang Yang**

**Part 1**

Nodes status screenshot.

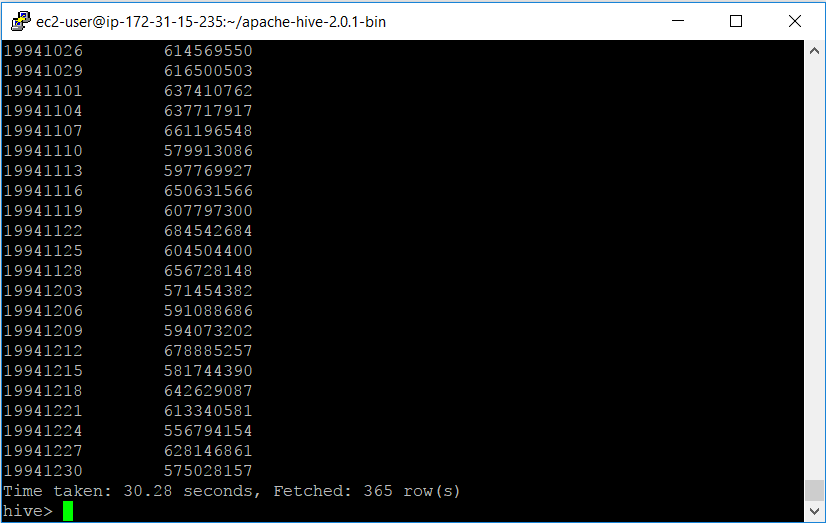


Word count time.

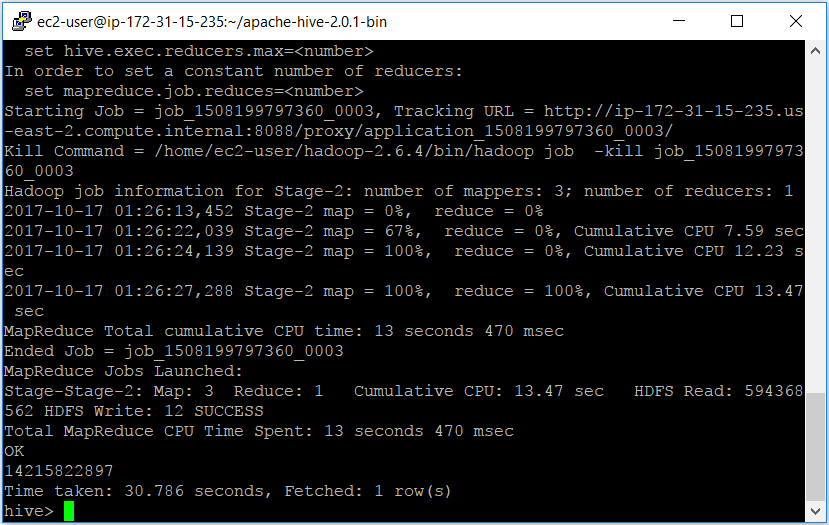


**Part 2**

**1.1**

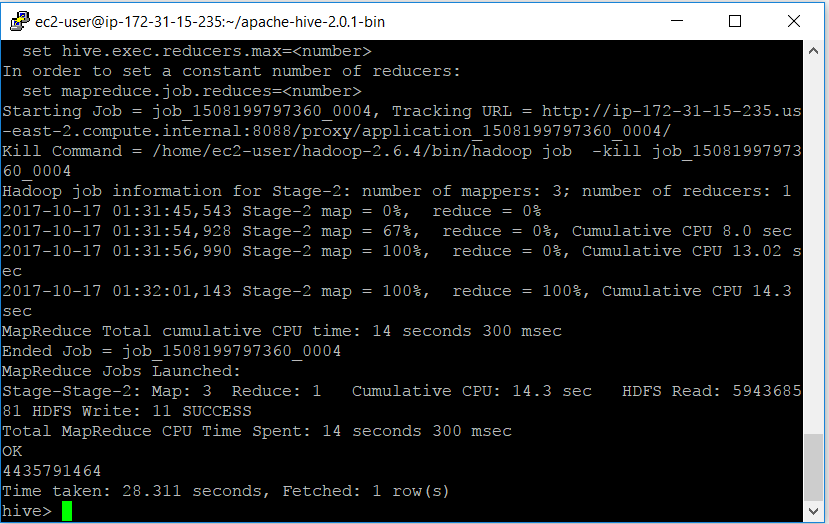
The runtime is 30.28 seconds

**1.2**



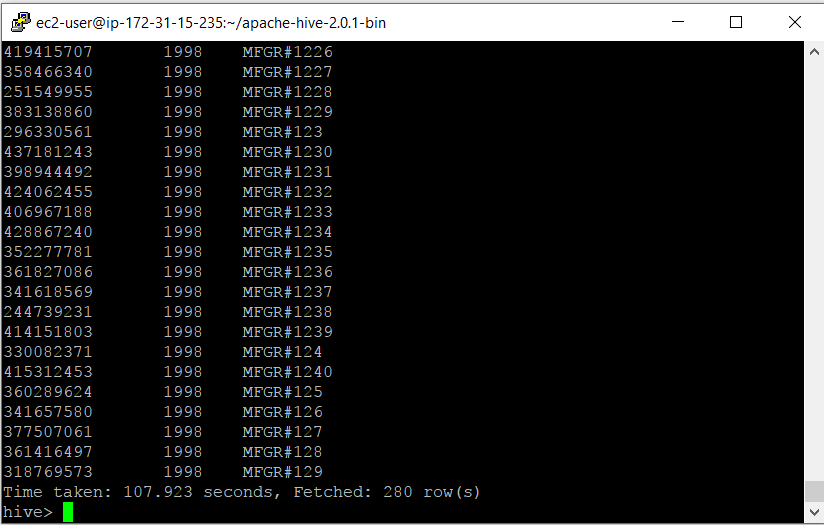
The run time is 30.786 seconds.

**1.3**



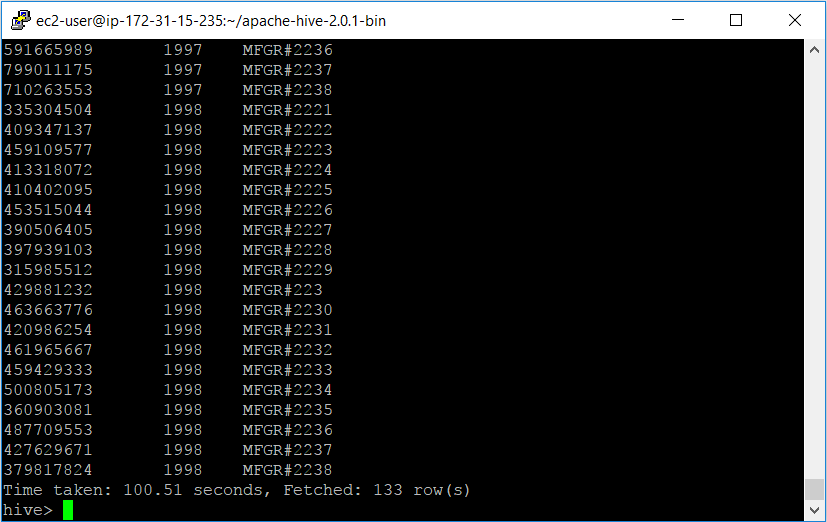
The run time is 28.311 seconds.

**2.1**



The run time is 107.923 seconds.

**2.2**



The run time is 100.51 seconds.

**Part 3**

**0.1**

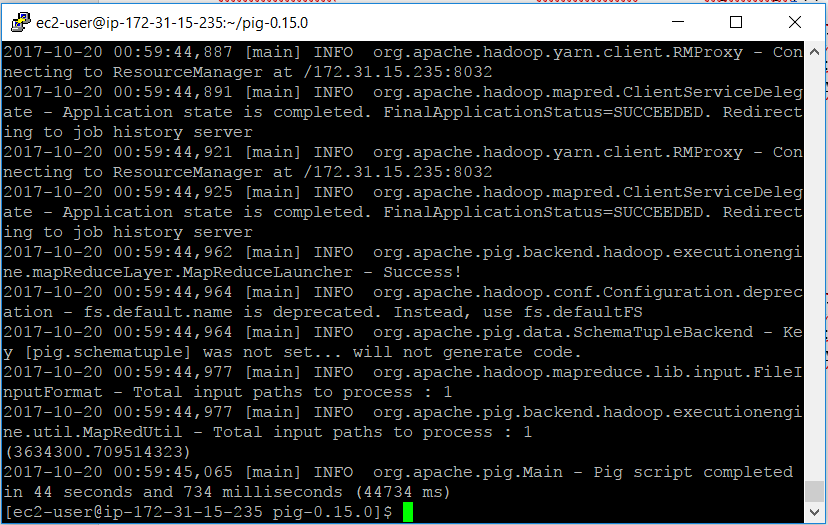
LineorderData = LOAD '/user/ec2-user/lineorder.tbl' USING PigStorage('|')

AS(lo\_orderkey:INT, lo\_linenumber:INT, lo\_custkey:INT, lo\_partkey:INT, lo\_suppkey:INT, lo\_orderdate:INT, lo\_orderpriority:CHARARRAY, lo\_shippriority:CHARARRAY, lo\_quantity:INT, lo\_extendedprice:INT, lo\_ordertotalprice:INT, lo\_discount:INT, lo\_revenue:INT, lo\_supplycost:INT, lo\_tax:INT, lo\_commitdate:INT, lo\_shipmode:CHARARRAY);

LineorderG = GROUP LineorderData ALL;

AvgR = FOREACH LineorderG GENERATE AVG(LineorderData.lo\_revenue);

DUMP AvgR;



The run time is 44 seconds and 734 milliseconds.

**0.2**

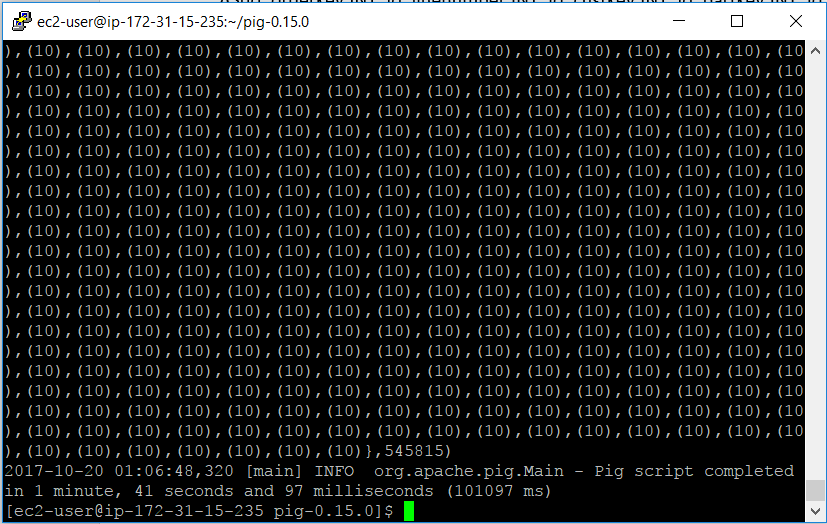
LineorderData = LOAD '/user/ec2-user/lineorder.tbl' USING PigStorage('|')

AS(lo\_orderkey:INT, lo\_linenumber:INT, lo\_custkey:INT, lo\_partkey:INT, lo\_suppkey:INT, lo\_orderdate:INT, lo\_orderpriority:CHARARRAY, lo\_shippriority:CHARARRAY, lo\_quantity:INT, lo\_extendedprice:INT, lo\_ordertotalprice:INT, lo\_discount:INT, lo\_revenue:INT, lo\_supplycost:INT, lo\_tax:INT, lo\_commitdate:INT, lo\_shipmode:CHARARRAY);

LineorderG = GROUP LineorderData BY lo\_discount;

SetD = FOREACH LineorderG GENERATE LineorderData.lo\_discount, COUNT(LineorderData.lo\_extendedprice);

DUMP SetD;



The run time is 1 minute, 41 seconds and 97 milliseconds.

**0.3**

LineorderData = LOAD '/user/ec2-user/lineorder.tbl' USING PigStorage('|')

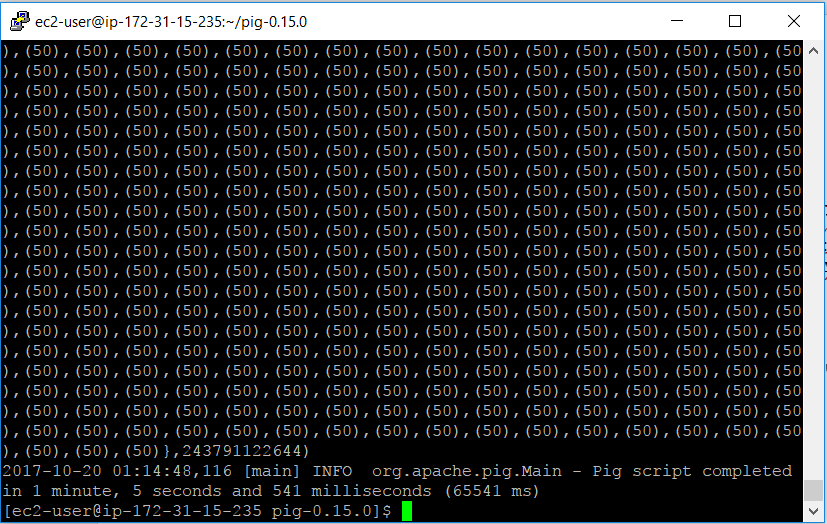
AS(lo\_orderkey:INT, lo\_linenumber:INT, lo\_custkey:INT, lo\_partkey:INT, lo\_suppkey:INT, lo\_orderdate:INT, lo\_orderpriority:CHARARRAY, lo\_shippriority:CHARARRAY, lo\_quantity:INT, lo\_extendedprice:INT, lo\_ordertotalprice:INT, lo\_discount:INT, lo\_revenue:INT, lo\_supplycost:INT, lo\_tax:INT, lo\_commitdate:INT, lo\_shipmode:CHARARRAY);

Discount3 = FILTER LineorderData BY lo\_discount < 3;

DiscountG = GROUP Discount3 BY lo\_quantity;

SetD3 = FOREACH DiscountG GENERATE Discount3.lo\_quantity, SUM(Discount3.lo\_revenue);

DUMP SetD3;



The run time is 1 minute, 5 seconds and 541 milliseconds.

**Part 4**

Mapper.py

#!/usr/bin/python

import sys

for line in sys.stdin:

line = line.strip()

val = line.split("|")

lo\_quantity = val[8]

lo\_revenue = val[12]

lo\_discount = val[11]

print '%s\t%s\t%s' % (lo\_quantity, lo\_revenue, lo\_discount)

Reducer.py

#!/usr/bin/python

import sys

lo\_dic = {}

lo\_list = []

sum\_rev = 0;

for line in sys.stdin:

line = line.strip()

lo\_quantity, lo\_revenue, lo\_discount = line.split("\t")

if int(lo\_discount) < 3:

if lo\_quantity in lo\_list:

lo\_dic[lo\_quantity].append(int(lo\_revenue))

else

lo\_list.append(lo\_quantity)

lo\_dic[lo\_quantity] = []

lo\_dic[lo\_quantity].append(int(lo\_revenue))

for keys, values in lo\_dic.iteritems():

if keys in lo\_list:

sum\_rev = sum\_rev + sum(values)

print '%s' % (sum\_rev)