**CSC 555: Mining Big Data**

Project, Phase 2 (due Sunday March 24th)

In this part of the project, you will execute queries using Hive, Pig and Hadoop streaming and develop a custom version of KMeans clustering. The schema is available below, but don’t forget to apply the correct delimiter:

<http://rasinsrv07.cstcis.cti.depaul.edu/CSC555/SSBM1/SSBM_schema_hive.sql>

The data is available at (this is Scale1, the smallest denomination of this benchmark)

<http://rasinsrv07.cstcis.cti.depaul.edu/CSC555/SSBM1/>

In your submission, please note what cluster you are using. Please be sure to submit all code (pig, python and Hive). You should also submit the command lines you use and a screenshot of a completed run (just the last page, do not worry about capturing the whole output). An answer without code will not receive credit.

**I highly recommend creating a small sample input** (e.g., by running head lineorder.tbl > lineorder.tbl.sample) and testing your code with it. You can run **head -n 500 lineorder.tbl** to get a specific number of lines.

NOTE: the total number of points adds up to 70 because Phase I is worth 30 of the project.

# Part 1: Data Transformation (15 pts)

Transform part.tbl table into a \*-separated (‘\*’) file: Use Hive, MapReduce with HadoopStreaming and Pig (i.e. 3 different solutions).

In all solutions you must switch odd and even columns (i.e., switch the positions of columns 1 and 2, columns 3 and 4, etc.). You do not need to transform the columns in any way, just a new data file.

Using my multi-node cluster

**Hive**

CREATE TABLE part (

p\_partkey INT,

p\_name VARCHAR(22),

p\_mfgr VARCHAR(6),

p\_category VARCHAR(7),

p\_brand1 VARCHAR(9),

p\_color VARCHAR(11),

p\_type VARCHAR(25),

p\_size INT,

p\_container VARCHAR(10)

)ROW FORMAT DELIMITED FIELDS

TERMINATED BY '|' STORED AS TEXTFILE;

LOAD DATA LOCAL INPATH '/home/ec2-user/part.tbl' OVERWRITE INTO TABLE part;

**Python code (colSwitcher.py):**

#!/usr/bin/python

import sys

for line in sys.stdin:

line = line.strip().split('\t')

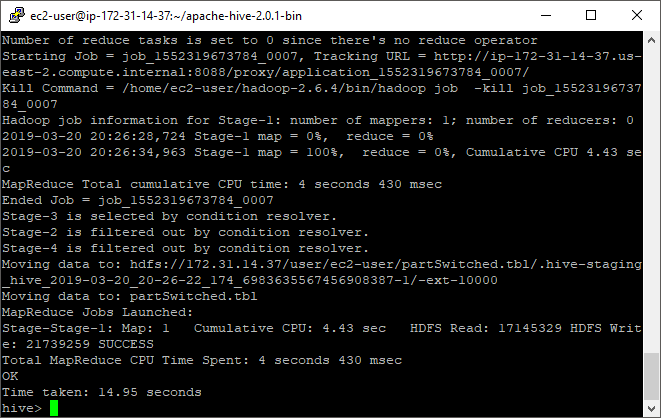
print '\*'.join([line[1], line[0], line[3], line[2], line[5], line[4], line[7], line[6], line[8]])

**Commands:**

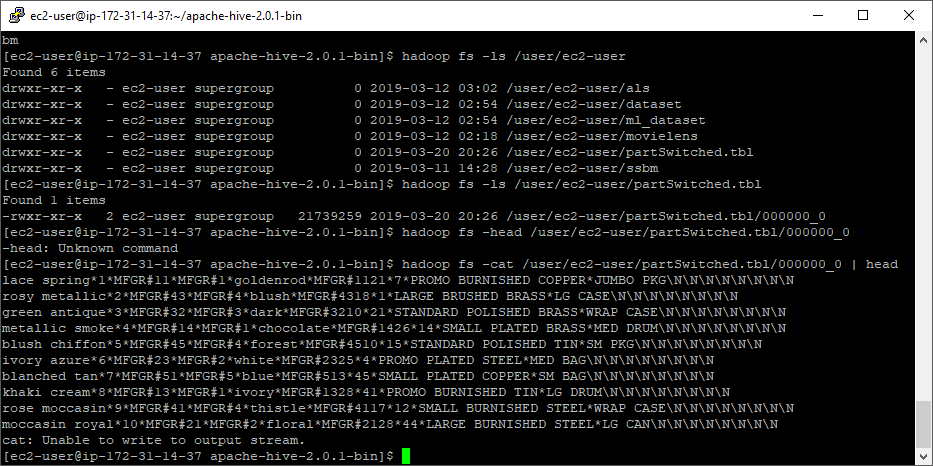
ADD FILE /home/ec2-user/colSwitcher.py;

INSERT OVERWRITE DIRECTORY 'partSwitched.tbl' SELECT TRANSFORM (p\_partkey, p\_name, p\_mfgr, p\_category, p\_brand1, p\_color, p\_type, p\_size, p\_container) USING 'colSwitcher.py' AS (p\_name, p\_partkey, p\_category, p\_mfgr, p\_color, p\_brand1, p\_size, p\_type, p\_container) FROM part;

# Completed Run:



**Output, first ten rows:**



**Hadoop Streaming**

There is no need for a custom mapper for this exercise, so I used the linux cat function as the mapper. The reducer code is:

**colSwitcherReducer.py**

#!/usr/bin/python

import sys

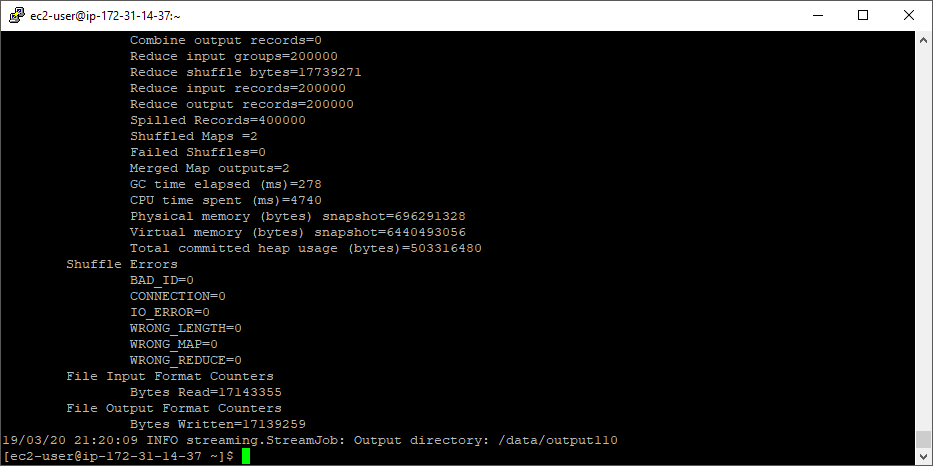
for line in sys.stdin:

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

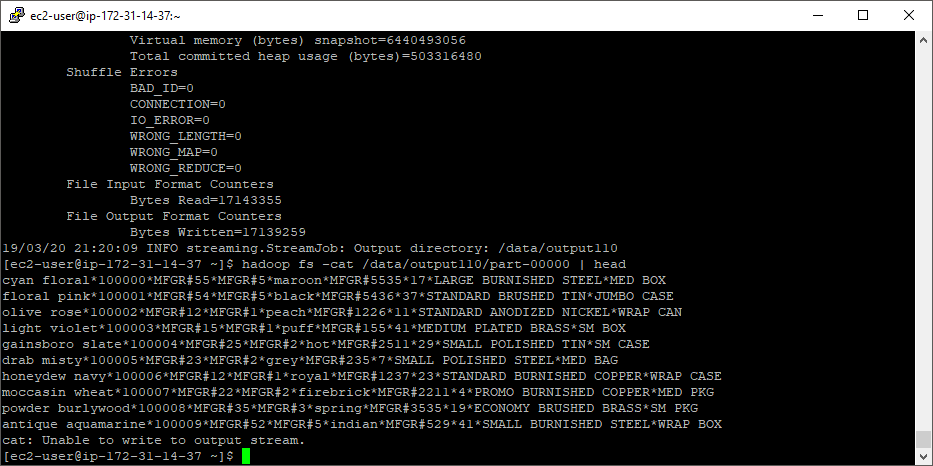
print "%s\*%s\*%s\*%s\*%s\*%s\*%s\*%s\*%s" % (line[1],line[0],line[3],line[2],line[5],line[4],line[7],line[6],line[8])

**Command:**

hadoop jar hadoop-streaming-2.6.4.jar -input /user/ec2-user/ssbm/part.tbl -output /data/output110 -mapper /bin/cat -reducer colSwitcherReducer.py -file colSwitcherReducer.py



**Output, first ten rows:**



**Pig**

**Load the Data:**

PartData = LOAD '/user/ec2-user/ssbm/part.tbl' USING PigStorage('|') AS (p\_partkey:int, p\_name:chararray, p\_mfgr:chararray, p\_category:chararray, p\_brand1:chararray, p\_color:chararray, p\_type:chararray, p\_size:int, p\_container:chararray);

**Verify Data Loaded:**

PartG = GROUP PartData ALL;

Count = FOREACH PartG GENERATE COUNT(PartData);

DUMP Count;

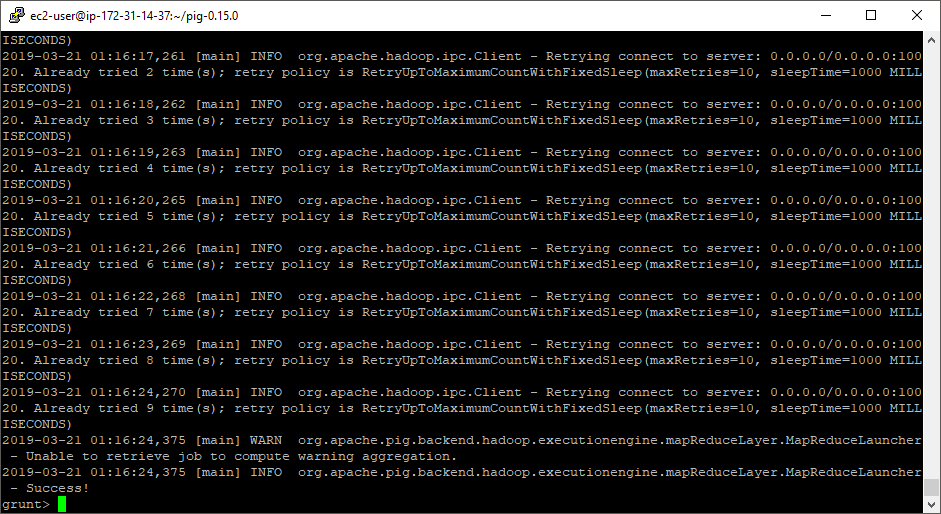
# 

**Switch Columns:**

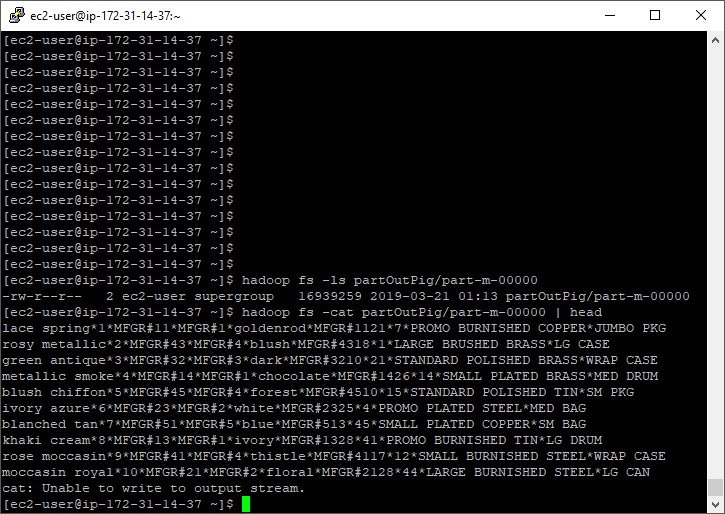
PartSwitchedPig = FOREACH PartData GENERATE p\_name, p\_partkey, p\_category, p\_mfgr, p\_color, p\_brand1, p\_size, p\_type, p\_container;

**Write to file:**

STORE PartSwitchedPig INTO 'partOutPig' USING PigStorage('\*');



**Output, first ten rows:**



# Part 2: Querying (25 pts)

Implement the following query:

select lo\_quantity, c\_nation, sum(lo\_revenue)

from customer, lineorder

where lo\_custkey = c\_custkey

and c\_region = 'AMERICA'

and lo\_discount BETWEEN 3 and 5

group by lo\_quantity, c\_nation;

using Hive, MapReduce with HadoopStreaming and Pig (i.e. 3 different solutions). I Hive, this merely requires pasting the query into the Hive prompt and timing it. In Hadoop streaming, this will require a total of 2 passes (one for join and another one for GROUP BY).

**Hive:**

**Create and load tables:**

CREATE TABLE lineorder (

lo\_orderkey INT,

lo\_linenumber INT,

lo\_custkey INT,

lo\_partkey INT,

lo\_suppkey INT,

lo\_orderdate INT,

lo\_orderpriority VARCHAR(15),

lo\_shippriority VARCHAR(1),

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 VARCHAR(10)

)

ROW FORMAT DELIMITED FIELDS

TERMINATED BY '|' STORED AS TEXTFILE;

LOAD DATA LOCAL INPATH '/home/ec2-user/lineorder.tbl' OVERWRITE INTO TABLE lineorder;

# CREATE TABLE customer (

# c\_custkey INT,

# c\_name VARCHAR(25),

# c\_address VARCHAR (25),

# c\_city VARCHAR (10),

# c\_nation VARCHAR (15),

# c\_region VARCHAR (12),

# c\_phone VARCHAR (15),

# c\_mktsegment VARCHAR (10)

# )

# ROW FORMAT DELIMITED FIELDS

# TERMINATED BY '|' STORED AS TEXTFILE;

LOAD DATA LOCAL INPATH '/home/ec2-user/customer.tbl' OVERWRITE INTO TABLE customer;

**Execute sql statement:**

select lo\_quantity, c\_nation, sum(lo\_revenue)

from customer, lineorder

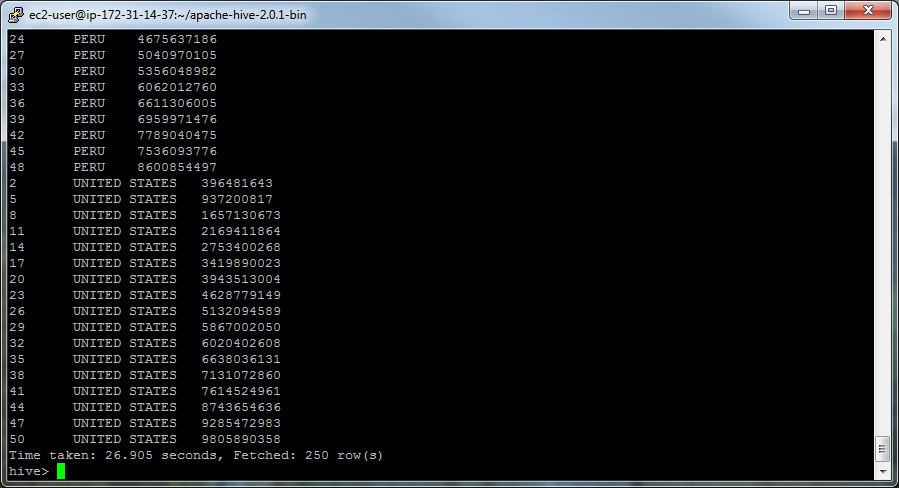
where lo\_custkey = c\_custkey

and c\_region = 'AMERICA'

and lo\_discount BETWEEN 3 and 5

group by lo\_quantity, c\_nation;

**End of output with time taken:**



**Hadoop Streaming:**

**Join**

**lineCustMapJoin.py**

#!/usr/bin/python

import sys

# input comes from STDIN (standard input)

for line in sys.stdin:

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

if line[1].startswith('Customer#'):

if line[5] == 'AMERICA': # Return on matching records

print line[0], '\t', line[4], '\t', 'customer'

# lineorder

else:

if 3 <= int(line[11]) <= 5: # Return on matching records

print line[2], '\t', line[8], '\t', line[12], '\t', 'lineorder'

**lineCustReduceJoin.py**

#!/usr/bin/python

import sys

currentKey = None

quantity = []

revenue = []

nation = ''

# input comes from STDIN

for line in sys.stdin:

split = line.strip().split('\t')

key = split[0] # key is customer id

value = '\t'.join(split[1:])

if currentKey == key: # Same key

if value.endswith('lineorder'):

quantity.extend([split[1]])

revenue.extend([split[2]])

if value.endswith('customer'):

nation = split[1]

else:

# Do not print anything until all records

# for a key have been seen, this is signaled

# by currentKey != key

# Check for values and then iterate results

lenQuantity = len(quantity)

lenNation = len(nation)

if (lenQuantity\*lenNation > 0):

i = 0

while i < lenQuantity:

print quantity[i], '\t', nation, '\t', revenue[i]

i += 1

# reset values

quantity = []

revenue = []

nation = ''

if value.endswith('lineorder'):

quantity.extend([split[1]])

revenue.extend([split[2]])

if value.endswith('customer'):

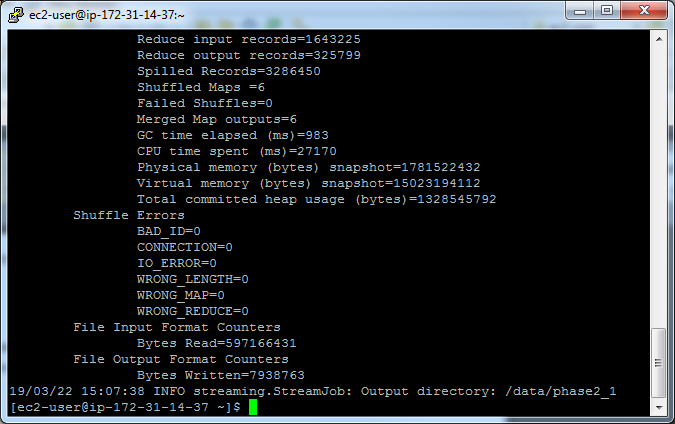
nation = split[1]

# set the current key at the end of each iteration

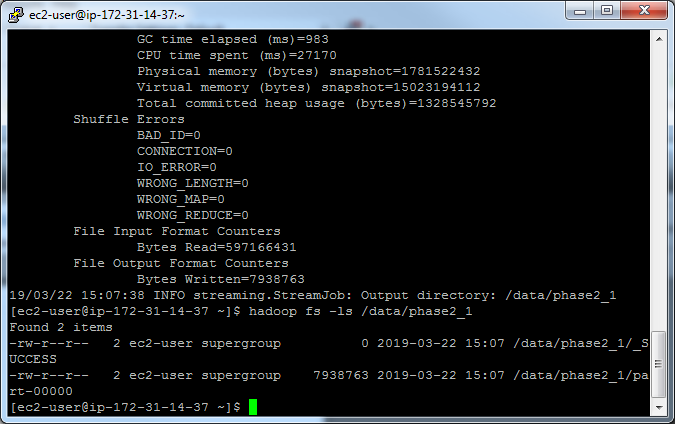
currentKey = key

**Commands**:

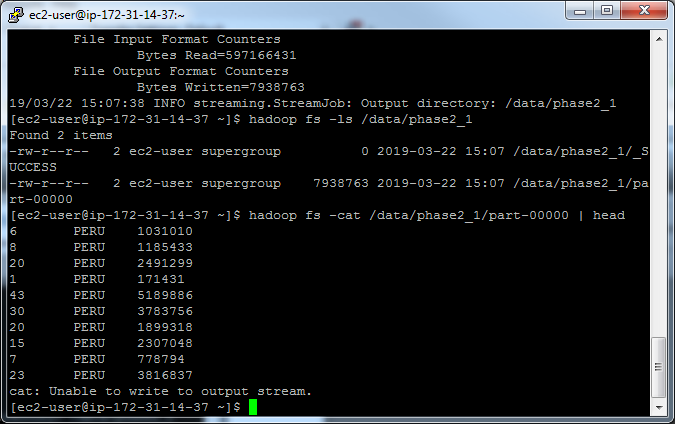
hadoop jar hadoop-streaming-2.6.4.jar -input /user/ec2-user/phase2 -output /data/phase2\_1 -mapper lineCustMapJoin.py -reducer lineCustReduceJoin.py -file lineCustMapJoin.py -file lineCustReduceJoin.py



hadoop fs -ls /data/phase2\_1



hadoop fs -cat /data/phase2\_1/part-00000 | head



**Group:**

**lineCustReduceGroup.py**

#!/usr/bin/python

import sys

curr\_id = None

curr\_tot = 0

id = None

# The input comes from standard input (line by line)

for line in sys.stdin:

# parse the line and split it by '\t'

line = line.strip().split('\t')

# grab the key

# values include some whitespace, removing here

id = line[0].strip() + '\t' + line[1].strip()

# grab the value (int)

val = int(line[2])

if curr\_id == id:

curr\_tot += val

else:

if curr\_id: # output the sum, single key completed

print '%s\t%d' % (curr\_id, curr\_tot)

curr\_tot = val

# set curr\_id to id at end of each iteration

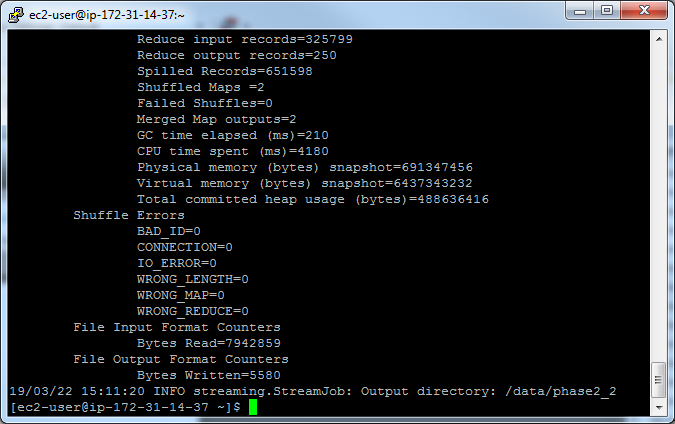
curr\_id = id

# output the last key

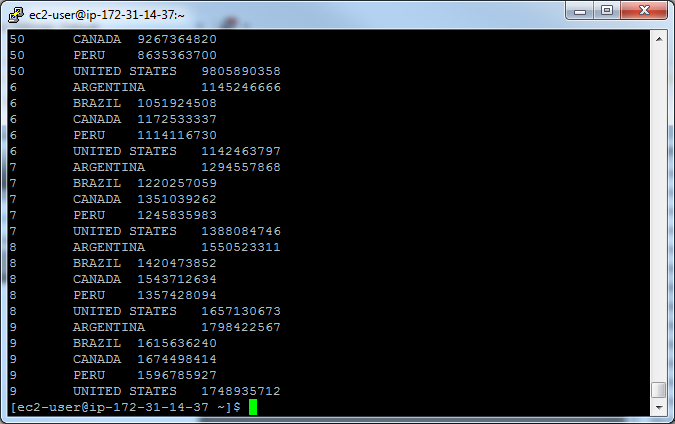
if curr\_id == id:

print '%s\t%d' % (curr\_id, curr\_tot)

hadoop jar hadoop-streaming-2.6.4.jar -D stream.num.map.output.key.fields=2 -input /data/phase2\_01/part-00000 -output /data/phase2\_06 -mapper /bin/cat -reducer lineCustReduceGroup.py -file lineCustReduceGroup.py



hadoop fs -cat /data/phase2\_2/part-00000



**Pig:**

**Load Tables:**

lineorder = LOAD '/user/ec2-user/ssbm/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

);

customer = LOAD '/user/ec2-user/ssbm/customer.tbl' USING PigStorage('|')

AS (c\_custkey:int,

c\_name:chararray,

c\_address:chararray,

c\_city:chararray,

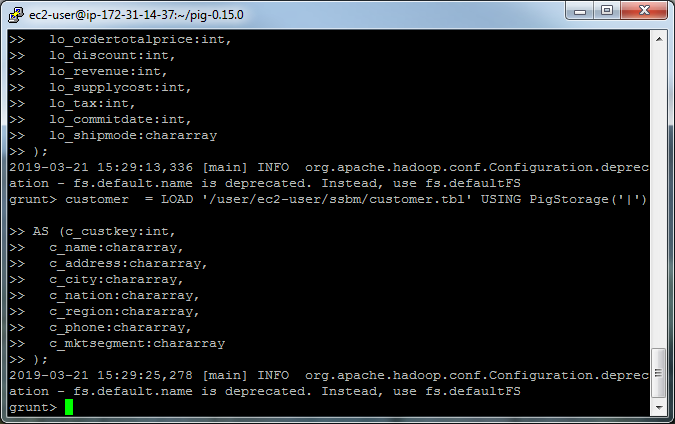
c\_nation:chararray,

c\_region:chararray,

c\_phone:chararray,

c\_mktsegment:chararray

);



**Execution steps:**

FilteredLineorder = FILTER lineorder BY lo\_discount >= 3 AND lo\_discount <= 5;

FilteredCustomer = FILTER customer BY c\_region == 'AMERICA';

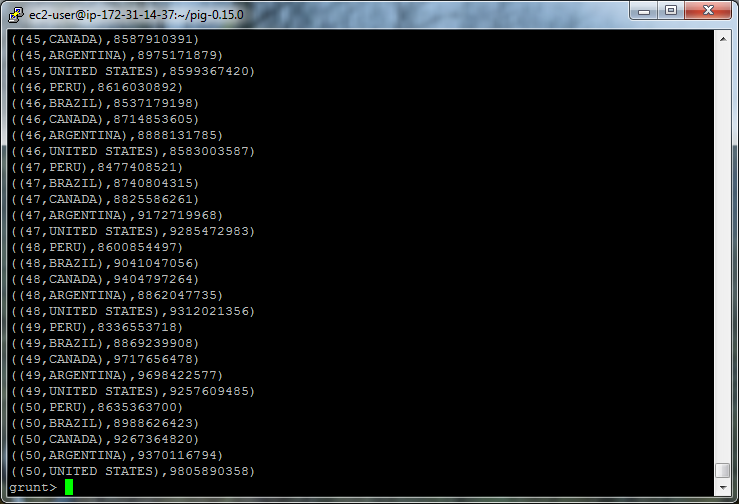
JoinedData = JOIN FilteredLineorder BY (lo\_custkey), FilteredCustomer BY (c\_custkey);

GroupedData = GROUP JoinedData BY (lo\_quantity, c\_nation);

Result = FOREACH GroupedData GENERATE group, SUM(JoinedData.lo\_revenue) as rev;

DUMP Result;

I had difficulty with displaying the non-summed columns so I left them out of the command. What’s displayed still includes the grouped columns. Moving on in the interest of time.



# Part 3: Clustering (30 pts)

Create a new numeric file with 25,000 rows and 3 columns, separated by space – you can generate numeric data as you prefer, but submit whatever code that you have used.

1. (5 pts) Using Mahout synthetic clustering as you have in a previous assignment on sample data. This entails running the **same** clustering command, but substituting your own input data instead of the sample.

For this exercise, I used the wine data set from UCI: <https://archive.ics.uci.edu/ml/datasets/wine>

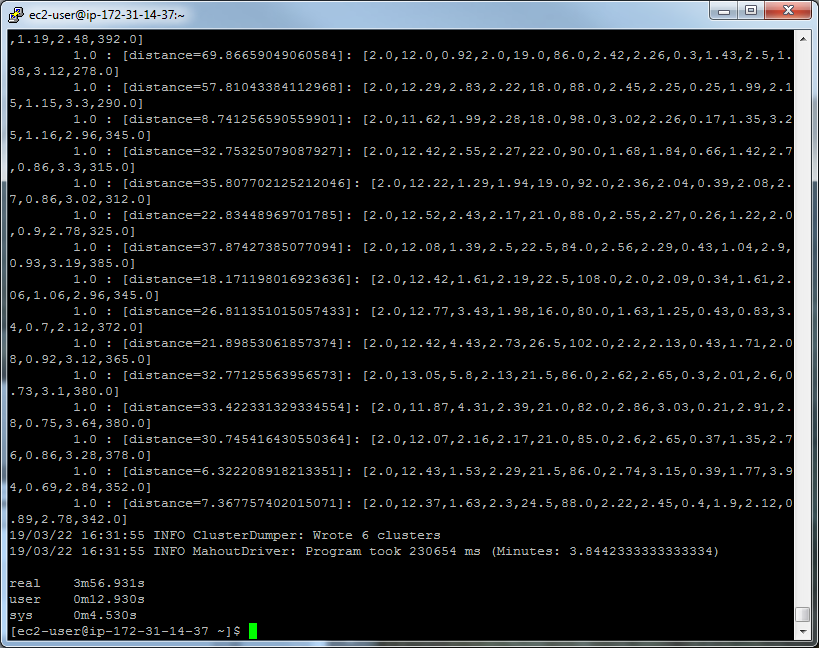
**Commands:**

wget <https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data>

**Note**: mahout did not like commas to separate the variables, so I used nano to replace commas with a space.

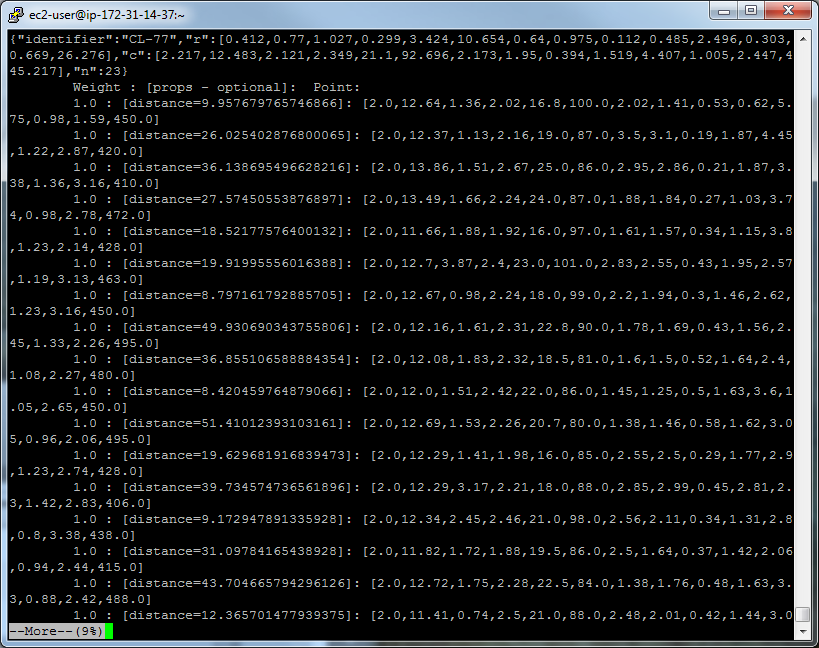
hadoop fs -put wine.data testdata/

time mahout org.apache.mahout.clustering.syntheticcontrol.kmeans.Job



mahout clusterdump --input output/clusters-10-final --pointsDir output/clusteredPoints --output clusteranalyze.txt

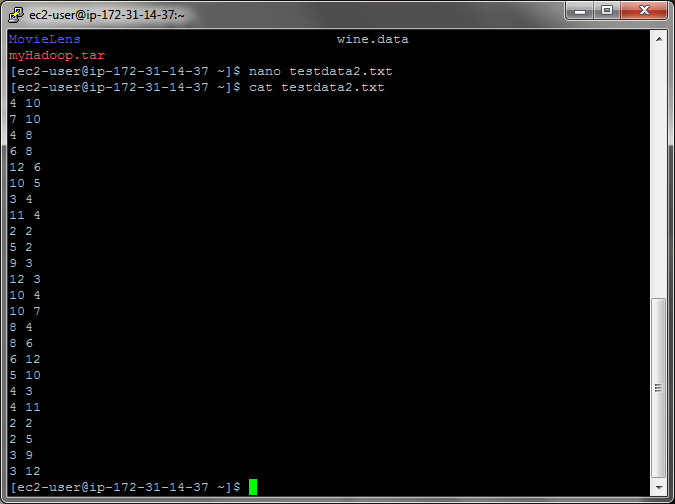
more clusteranalyze.txt



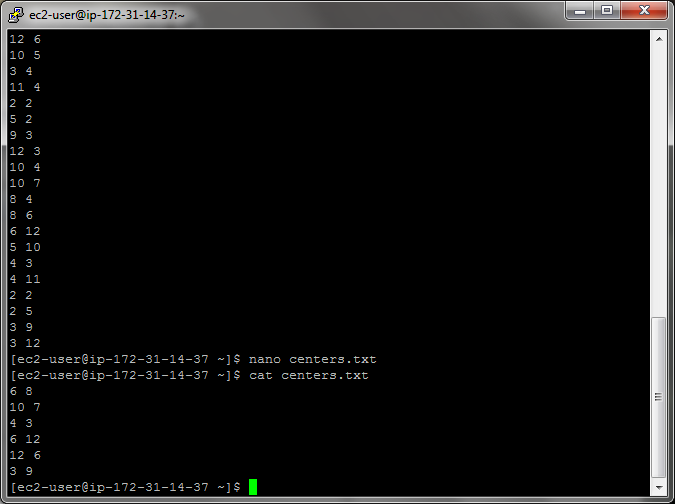
1. (25 pts) Using Hadoop streaming perform four iterations manually **using 6 centers** (initially with randomly chosen centers). This would require passing a text file with cluster centers using -file option, opening the centers.txt in the mapper with open(‘centers.txt’, ‘r’) and assigning a key to each point based on which center is the closest to each particular point. Your reducer would then compute the new centers, and at that point the iteration is done and the output of the reducer with new centers can be given to the next pass of the same code.

The only difference between first and subsequent iteration is that in first iteration you have to pick the initial centers. Starting from 2nd iteration, the centers will be given to you by a previous pass of KMeans.

I used the sample data from chapter 7 in the book, but to add more data points, I swapped the x and y values to double the number of entries:



Starting Centers:



**Extra credit (7 pts)**: Create the equivalent of KMeans driver from Mahout. That is, write a python script that will automatically execute the hadoop streaming command, then get the new centers from HDFS and repeat the command. This will be easiest to do if you write your reducer to output just the centers (without the key) to HDFS. This way, all you have to do is to execute the get command to get the new centers (you can hard-code the locations of output in HDFS into your script).

Submit a single document containing your written answers. Be sure that this document contains your name and “CSC 555 Project Phase 2” at the top.