Assignment 5
CSC 555
Sarah Cummings

1) Consider a Hadoop job that processes a job that will result in 50 blocks of output to HDFS—writing takes 2 mins per block and the replication factor is set to two.

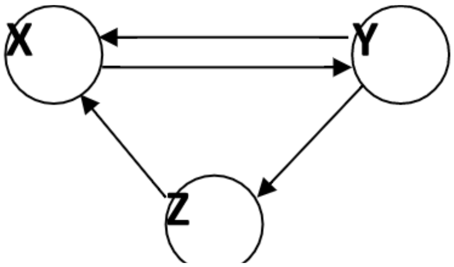
a. How long will it take for the reducer to write the job output on a 1-node Hadoop cluster?
Assuming no failure, 100 minutes. Two minutes for each of the 50 blocks.

b. How long will it take for the reducer to write the job to 5 Hadoop worker nodes cluster with replication set to one?
20 minutes

c. How long will it take for reducer(s) to write the job output to 5 Hadoop worker nodes with replication set to two?
20 minutes. We only use the replicated data if there is failure.

2) a. Page rank for the following:

$$M = \begin{array}{ccc} X & Y & Z \\ 0 & 1/2 & 0 \\ 1 & 0 & 0 \\ 0 & 1/2 & 0 \end{array} \qquad V = \begin{array}{c} 1/3 \\ 1/3 \\ 1/3 \end{array}$$

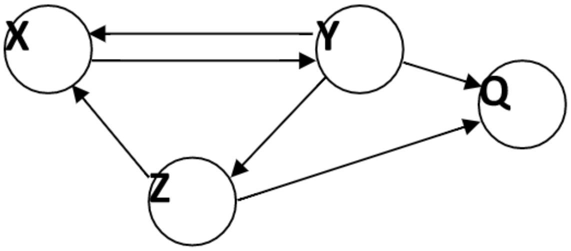


Using my graphing calculator, I get M^50 * v =
a 3*1 vector that is a column of nearly zeros. It gives 9.93 * 10 ^ -9 in each row.

The matrix will converge to zero because Z is a dead end node.

b. Page rank for the following, with Q as a dead end node:

With a dead end node, the matrix converges to zero and hence the page rank is 0.

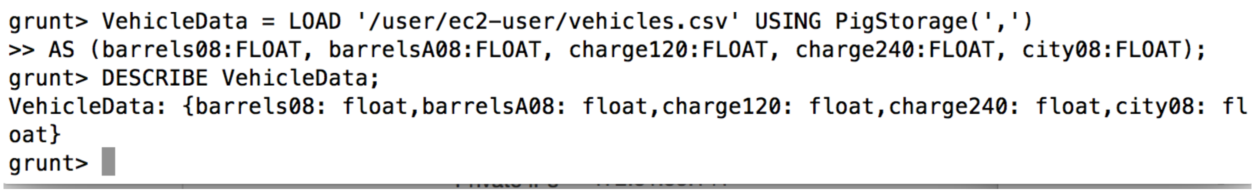


3) Implement and run MapReduce to compute The count of number of distinct integers in the input file

I don’t know how to do this. Might submit another file late with this part of the assignment.

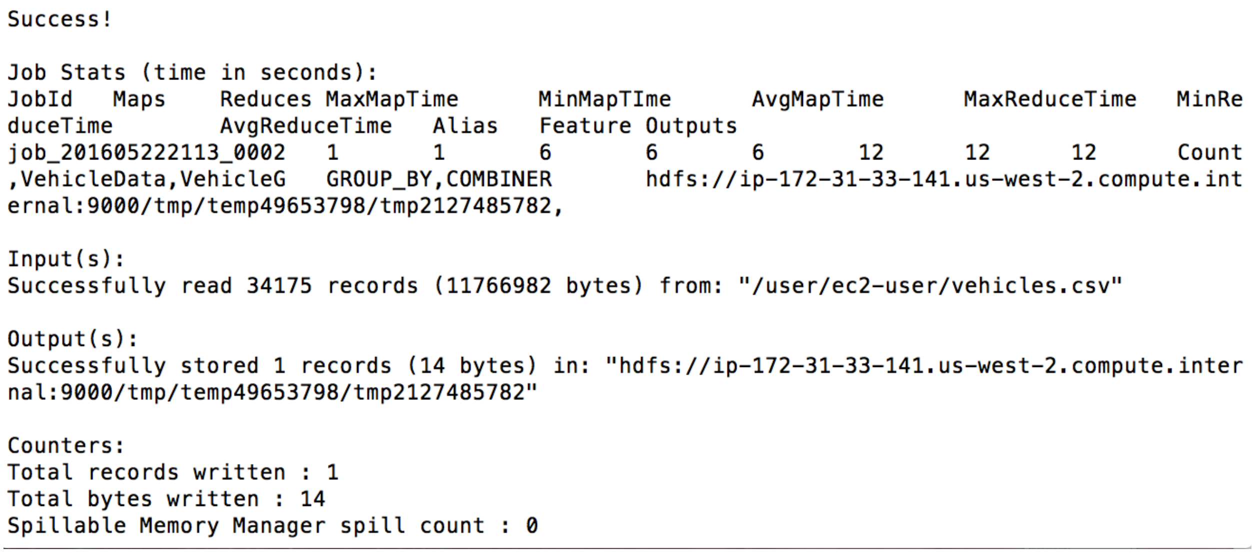
4) Pig set up and vehicle table problem

Describing Vehicle data:



```
grunt> VehicleData = LOAD '/user/ec2-user/vehicles.csv' USING PigStorage(',')
>> AS (barrels08:FLOAT, barrelsA08:FLOAT, charge120:FLOAT, charge240:FLOAT, city08:FLOAT);
grunt> DESCRIBE VehicleData;
VehicleData: {barrels08: float,barrelsA08: float,charge120: float,charge240: float,city08: fl
oat}
grunt> 
```

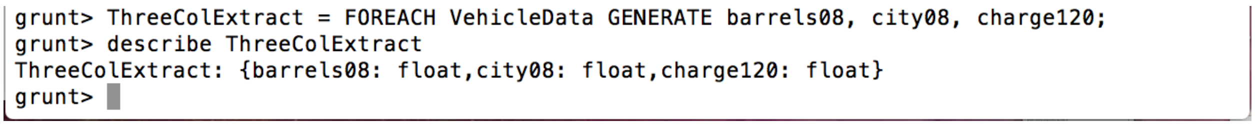
Verifying the data has loaded:



```
Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTIme      AvgMapTime      MaxReduceTime   MinRe
duceTime        AvgReduceTime   Alias   Feature Outputs
job_201605222113_0002   1       1       6       6       6       12      12      12      Count
,VehicleData,VehicleG   GROUP_BY,COMBINER       hdfs://ip-172-31-33-141.us-west-2.compute.int
ernal:9000/tmp/temp49653798/tmp2127485782,

Input(s):
Successfully read 34175 records (11766982 bytes) from: "/user/ec2-user/vehicles.csv"

Output(s):
Successfully stored 1 records (14 bytes) in: "hdfs://ip-172-31-33-141.us-west-2.compute.inter
nal:9000/tmp/temp49653798/tmp2127485782"

Counters:
Total records written : 1
Total bytes written : 14
Spillable Memory Manager spill count : 0
```

There are 34175 records.

ThreeColumn Table:

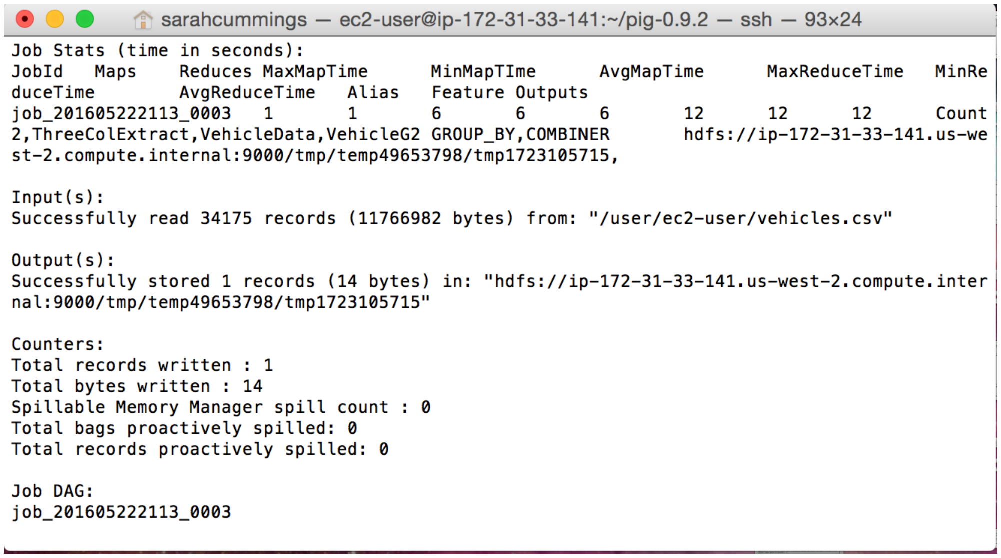


```
grunt> ThreeColExtract = FOREACH VehicleData GENERATE barrels08, city08, charge120;
grunt> describe ThreeColExtract
ThreeColExtract: {barrels08: float,city08: float,charge120: float}
grunt> 
```

I don’t know how to “verify its been created”
so i did something similar to what we did above. I ran

```
VehicleG2 = GROUP ThreeColExtract ALL;
Count2= FOREACH VehicleG2 GENERATE COUNT(ThreeColExtract);
Dump Count2;
```

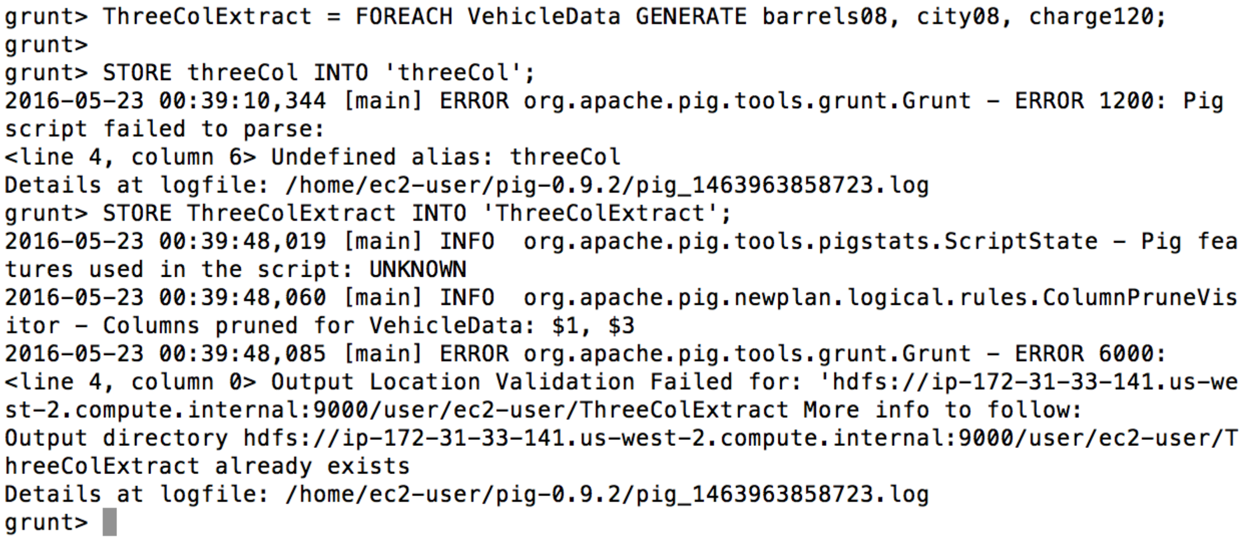
And I got a similar response as before, confirming again we have all of our rows of our data.



sarahcummings — ec2-user@ip-172-31-33-141:~/pig-0.9.2 — ssh — 93×24

```
Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTIme      AvgMapTime      MaxReduceTime   MinRe
duceTime        AvgReduceTime   Alias   Feature Outputs
job_20160522113_0003    1       1       6       6       6       12      12      12      Count
2,ThreeColExtract,VehicleData,VehicleG2 GROUP_BY,COMBINER       hdfs://ip-172-31-33-141.us-we
st-2.compute.internal:9000/tmp/temp49653798/tmp1723105715,

Input(s):
Successfully read 34175 records (11766982 bytes) from: "/user/ec2-user/vehicles.csv"

Output(s):
Successfully stored 1 records (14 bytes) in: "hdfs://ip-172-31-33-141.us-west-2.compute.inter
nal:9000/tmp/temp49653798/tmp1723105715"

Counters:
Total records written : 1
Total bytes written : 14
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_20160522113_0003
```

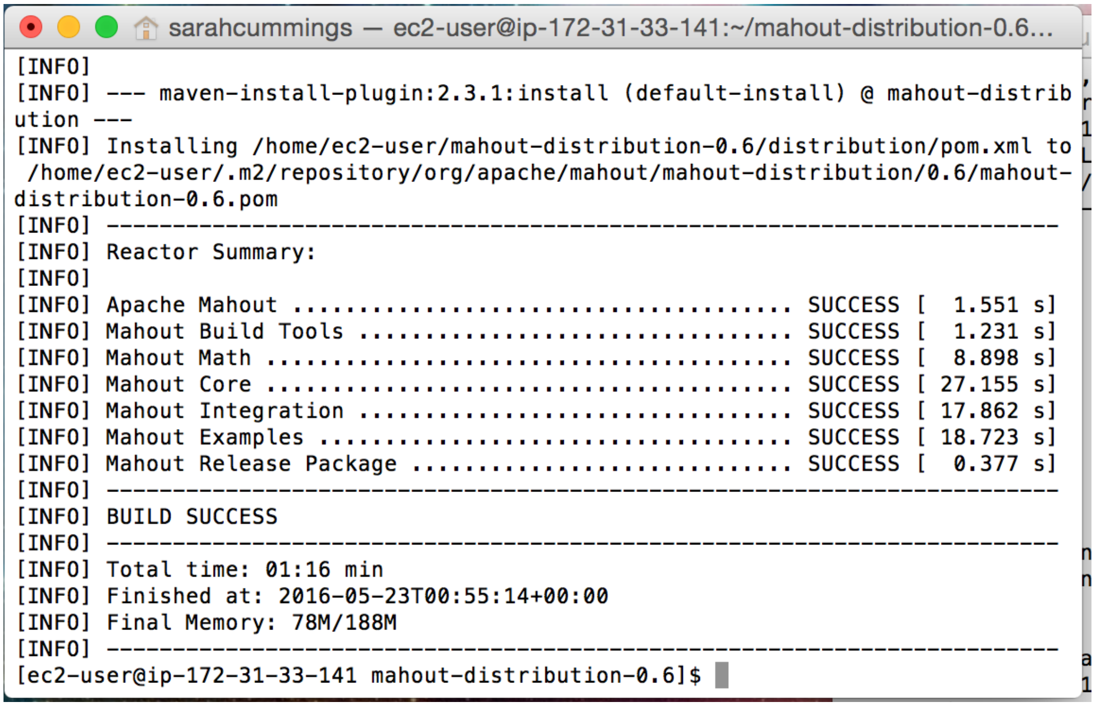
But also using **store** I get:



```
grunt> ThreeColExtract = FOREACH VehicleData GENERATE barrels08, city08, charge120;
grunt>
grunt> STORE threeCol INTO 'threeCol';
2016-05-23 00:39:10,344 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1200: Pig
script failed to parse:
<line 4, column 6> Undefined alias: threeCol
Details at logfile: /home/ec2-user/pig-0.9.2/pig_1463963858723.log
grunt> STORE ThreeColExtract INTO 'ThreeColExtract';
2016-05-23 00:39:48,019 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig fea
tures used in the script: UNKNOWN
2016-05-23 00:39:48,060 [main] INFO  org.apache.pig.newplan.logical.rules.ColumnPruneVis
itor - Columns pruned for VehicleData: $1, $3
2016-05-23 00:39:48,085 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 6000:
<line 4, column 0> Output Location Validation Failed for: 'hdfs://ip-172-31-33-141.us-we
st-2.compute.internal:9000/user/ec2-user/ThreeColExtract More info to follow:
Output directory hdfs://ip-172-31-33-141.us-west-2.compute.internal:9000/user/ec2-user/T
hreeColExtract already exists
Details at logfile: /home/ec2-user/pig-0.9.2/pig_1463963858723.log
grunt> 
```

5) Mahout set up

a. Setting up mahout:

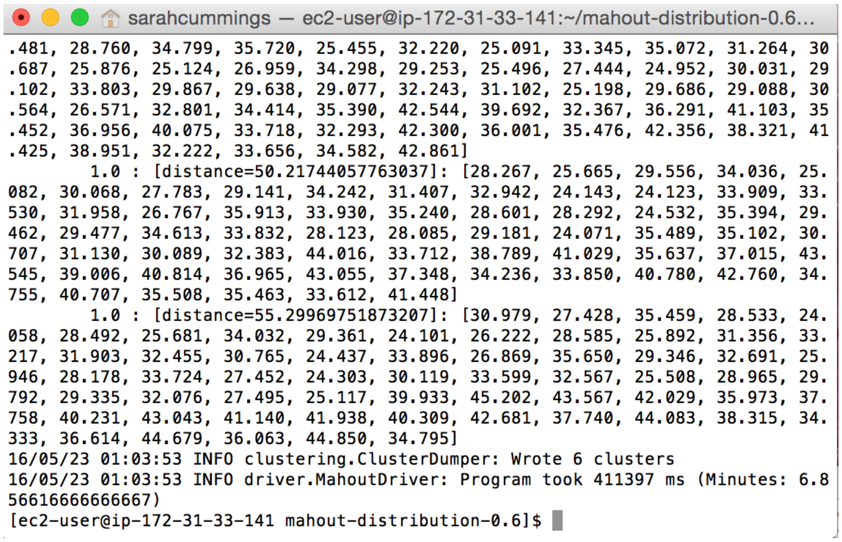

sarahcummings — ec2-user@ip-172-31-33-141:~/mahout-distribution-0.6...

```
[INFO]
[INFO] --- maven-install-plugin:2.3.1:install (default-install) @ mahout-distrib
ution ---
[INFO] Installing /home/ec2-user/mahout-distribution-0.6/distribution/pom.xml to
 /home/ec2-user/.m2/repository/org/apache/mahout/mahout-distribution/0.6/mahout-
distribution-0.6.pom
[INFO] ------------------------------------------------------------------------
[INFO] Reactor Summary:
[INFO]
[INFO] Apache Mahout ..................................... SUCCESS [  1.551 s]
[INFO] Mahout Build Tools ................................ SUCCESS [  1.231 s]
[INFO] Mahout Math ....................................... SUCCESS [  8.898 s]
[INFO] Mahout Core ....................................... SUCCESS [ 27.155 s]
[INFO] Mahout Integration ................................ SUCCESS [ 17.862 s]
[INFO] Mahout Examples ................................... SUCCESS [ 18.723 s]
[INFO] Mahout Release Package ............................ SUCCESS [  0.377 s]
[INFO] ------------------------------------------------------------------------
[INFO] BUILD SUCCESS
[INFO] ------------------------------------------------------------------------
[INFO] Total time: 01:16 min
[INFO] Finished at: 2016-05-23T00:55:14+00:00
[INFO] Final Memory: 78M/188M
[INFO] ------------------------------------------------------------------------
[ec2-user@ip-172-31-33-141 mahout-distribution-0.6]$
```

b. Doing the clustering example:

Running $MAHOUT_HOME/bin/mahouto
rg.apache.mahout.clustering.syntheticcontrol.kmeans.Job

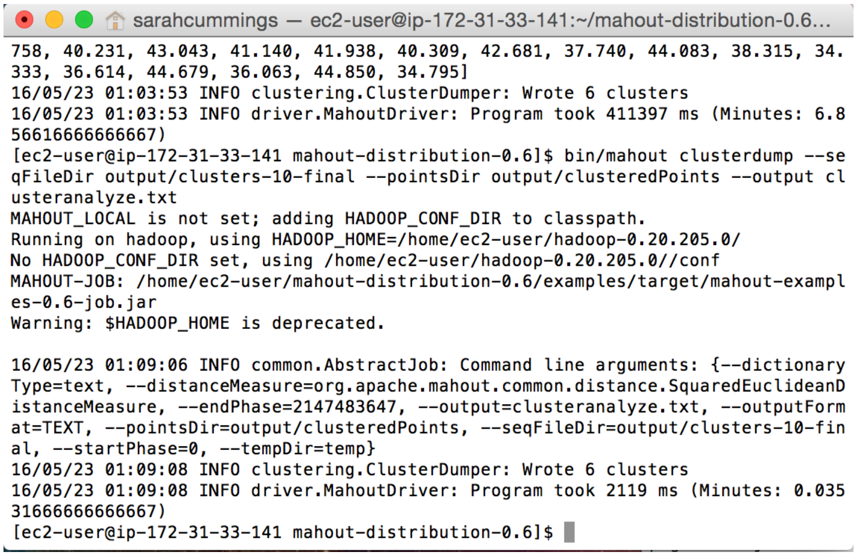
Took 6.8 minutes


sarahcummings — ec2-user@ip-172-31-33-141:~/mahout-distribution-0.6...

```
.481, 28.760, 34.799, 35.720, 25.455, 32.220, 25.091, 33.345, 35.072, 31.264, 30
.687, 25.876, 25.124, 26.959, 34.298, 29.253, 25.496, 27.444, 24.952, 30.031, 29
.102, 33.803, 29.867, 29.638, 29.077, 32.243, 31.102, 25.198, 29.686, 29.088, 30
.564, 26.571, 32.801, 34.414, 35.390, 42.544, 39.692, 32.367, 36.291, 41.103, 35
.452, 36.956, 40.075, 33.718, 32.293, 42.300, 36.001, 35.476, 42.356, 38.321, 41
.425, 38.951, 32.222, 33.656, 34.582, 42.861]
        1.0 : [distance=50.21744057763037]: [28.267, 25.665, 29.556, 34.036, 25.
082, 30.068, 27.783, 29.141, 34.242, 31.407, 32.942, 24.143, 24.123, 33.909, 33.
530, 31.958, 26.767, 35.913, 33.930, 35.240, 28.601, 28.292, 24.532, 35.394, 29.
462, 29.477, 34.613, 33.832, 28.123, 28.085, 29.181, 24.071, 35.489, 35.102, 30.
707, 31.130, 30.089, 32.383, 44.016, 33.712, 38.789, 41.029, 35.637, 37.015, 43.
545, 39.006, 40.814, 36.965, 43.055, 37.348, 34.236, 33.850, 40.780, 42.760, 34.
755, 40.707, 35.508, 35.463, 33.612, 41.448]
        1.0 : [distance=55.29969751873207]: [30.979, 27.428, 35.459, 28.533, 24.
058, 28.492, 25.681, 34.032, 29.361, 24.101, 26.222, 28.585, 25.892, 31.356, 33.
217, 31.903, 32.455, 30.765, 24.437, 33.896, 26.869, 35.650, 29.346, 32.691, 25.
946, 28.178, 33.724, 27.452, 24.303, 30.119, 33.599, 32.567, 25.508, 28.965, 29.
792, 29.335, 32.076, 27.495, 25.117, 39.933, 45.202, 43.567, 42.029, 35.973, 37.
758, 40.231, 43.043, 41.140, 41.938, 40.309, 42.681, 37.740, 44.083, 38.315, 34.
333, 36.614, 44.679, 36.063, 44.850, 34.795]
16/05/23 01:03:53 INFO clustering.ClusterDumper: Wrote 6 clusters
16/05/23 01:03:53 INFO driver.MahoutDriver: Program took 411397 ms (Minutes: 6.8
56616666666667)
[ec2-user@ip-172-31-33-141 mahout-distribution-0.6]$
```

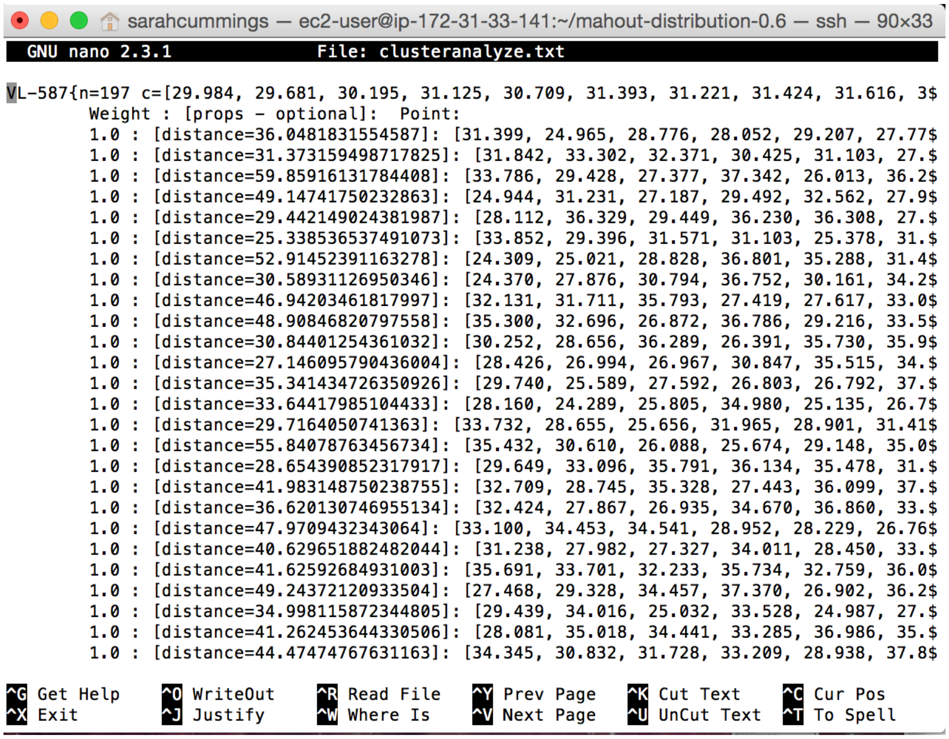
Running

$MAHOUT_HOME/bin/mahout clusterdump --seqFileDir output/clusters-10-final --pointsDir output/clusteredPoints --output clusteranalyze.txt


sarahcummings — ec2-user@ip-172-31-33-141:~/mahout-distribution-0.6...

```
758, 40.231, 43.043, 41.140, 41.938, 40.309, 42.681, 37.740, 44.083, 38.315, 34.
333, 36.614, 44.679, 36.063, 44.850, 34.795]
16/05/23 01:03:53 INFO clustering.ClusterDumper: Wrote 6 clusters
16/05/23 01:03:53 INFO driver.MahoutDriver: Program took 411397 ms (Minutes: 6.8
5661666666667)
[ec2-user@ip-172-31-33-141 mahout-distribution-0.6]$ bin/mahout clusterdump --se
qFileDir output/clusters-10-final --pointsDir output/clusteredPoints --output cl
usteranalyze.txt
MAHOUT_LOCAL is not set; adding HADOOP_CONF_DIR to classpath.
Running on hadoop, using HADOOP_HOME=/home/ec2-user/hadoop-0.20.205.0/
No HADOOP_CONF_DIR set, using /home/ec2-user/hadoop-0.20.205.0//conf
MAHOUT-JOB: /home/ec2-user/mahout-distribution-0.6/examples/target/mahout-exampl
es-0.6-job.jar
Warning: $HADOOP_HOME is deprecated.

16/05/23 01:09:06 INFO common.AbstractJob: Command line arguments: {--dictionary
Type=text, --distanceMeasure=org.apache.mahout.common.distance.SquaredEuclideanD
istanceMeasure, --endPhase=2147483647, --output=clusteranalyze.txt, --outputForm
at=TEXT, --pointsDir=output/clusteredPoints, --seqFileDir=output/clusters-10-fin
al, --startPhase=0, --tempDir=temp}
16/05/23 01:09:08 INFO clustering.ClusterDumper: Wrote 6 clusters
16/05/23 01:09:08 INFO driver.MahoutDriver: Program took 2119 ms (Minutes: 0.035
31666666666667)
[ec2-user@ip-172-31-33-141 mahout-distribution-0.6]$
```

First page of the ClusterAnalyze.txt


sarahcummings — ec2-user@ip-172-31-33-141:~/mahout-distribution-0.6 — ssh — 90×33

```
  GNU nano 2.3.1             File: clusteranalyze.txt

VL-587{n=197 c=[29.984, 29.681, 30.195, 31.125, 30.709, 31.393, 31.221, 31.424, 31.616, 3$
        Weight : [props - optional]:  Point:
        1.0 : [distance=36.0481831554587]: [31.399, 24.965, 28.776, 28.052, 29.207, 27.77$
        1.0 : [distance=31.373159498717825]: [31.842, 33.302, 32.371, 30.425, 31.103, 27.$
        1.0 : [distance=59.85916131784408]: [33.786, 29.428, 27.377, 37.342, 26.013, 36.2$
        1.0 : [distance=49.14741750232863]: [24.944, 31.231, 27.187, 29.492, 32.562, 27.9$
        1.0 : [distance=29.442149024381987]: [28.112, 36.329, 29.449, 36.230, 36.308, 27.$
        1.0 : [distance=25.338536537491073]: [33.852, 29.396, 31.571, 31.103, 25.378, 31.$
        1.0 : [distance=52.91452391163278]: [24.309, 25.021, 28.828, 36.801, 35.288, 31.4$
        1.0 : [distance=30.58931126950346]: [24.370, 27.876, 30.794, 36.752, 30.161, 34.2$
        1.0 : [distance=46.94203461817997]: [32.131, 31.711, 35.793, 27.419, 27.617, 33.0$
        1.0 : [distance=48.90846820797558]: [35.300, 32.696, 26.872, 36.786, 29.216, 33.5$
        1.0 : [distance=30.84401254361032]: [30.252, 28.656, 36.289, 26.391, 35.730, 35.9$
        1.0 : [distance=27.146095790436004]: [28.426, 26.994, 26.967, 30.847, 35.515, 34.$
        1.0 : [distance=35.341434726350926]: [29.740, 25.589, 27.592, 26.803, 26.792, 37.$
        1.0 : [distance=33.64417985104433]: [28.160, 24.289, 25.805, 34.980, 25.135, 26.7$
        1.0 : [distance=29.7164050741363]: [33.732, 28.655, 25.656, 31.965, 28.901, 31.41$
        1.0 : [distance=55.84078763456734]: [35.432, 30.610, 26.088, 25.674, 29.148, 35.0$
        1.0 : [distance=28.654390852317917]: [29.649, 33.096, 35.791, 36.134, 35.478, 31.$
        1.0 : [distance=41.983148750238755]: [32.709, 28.745, 35.328, 27.443, 36.099, 37.$
        1.0 : [distance=36.62013074695134]: [32.424, 27.867, 26.935, 34.670, 36.860, 33.$
        1.0 : [distance=47.9709432343064]: [33.100, 34.453, 34.541, 28.952, 28.229, 26.76$
        1.0 : [distance=40.629651882482044]: [31.238, 27.982, 27.327, 34.011, 28.450, 33.$
        1.0 : [distance=41.62592684931003]: [35.691, 33.701, 32.233, 35.734, 32.759, 36.0$
        1.0 : [distance=49.24372120933504]: [27.468, 29.328, 34.457, 37.370, 26.902, 36.2$
        1.0 : [distance=34.998115872344805]: [29.439, 34.016, 25.032, 33.528, 24.987, 27.$
        1.0 : [distance=41.262453644330506]: [28.081, 35.018, 34.441, 33.285, 36.986, 35.$
        1.0 : [distance=44.47474767631163]: [34.345, 30.832, 31.728, 33.209, 28.938, 37.8$
```