Lab #3 – Hadoop

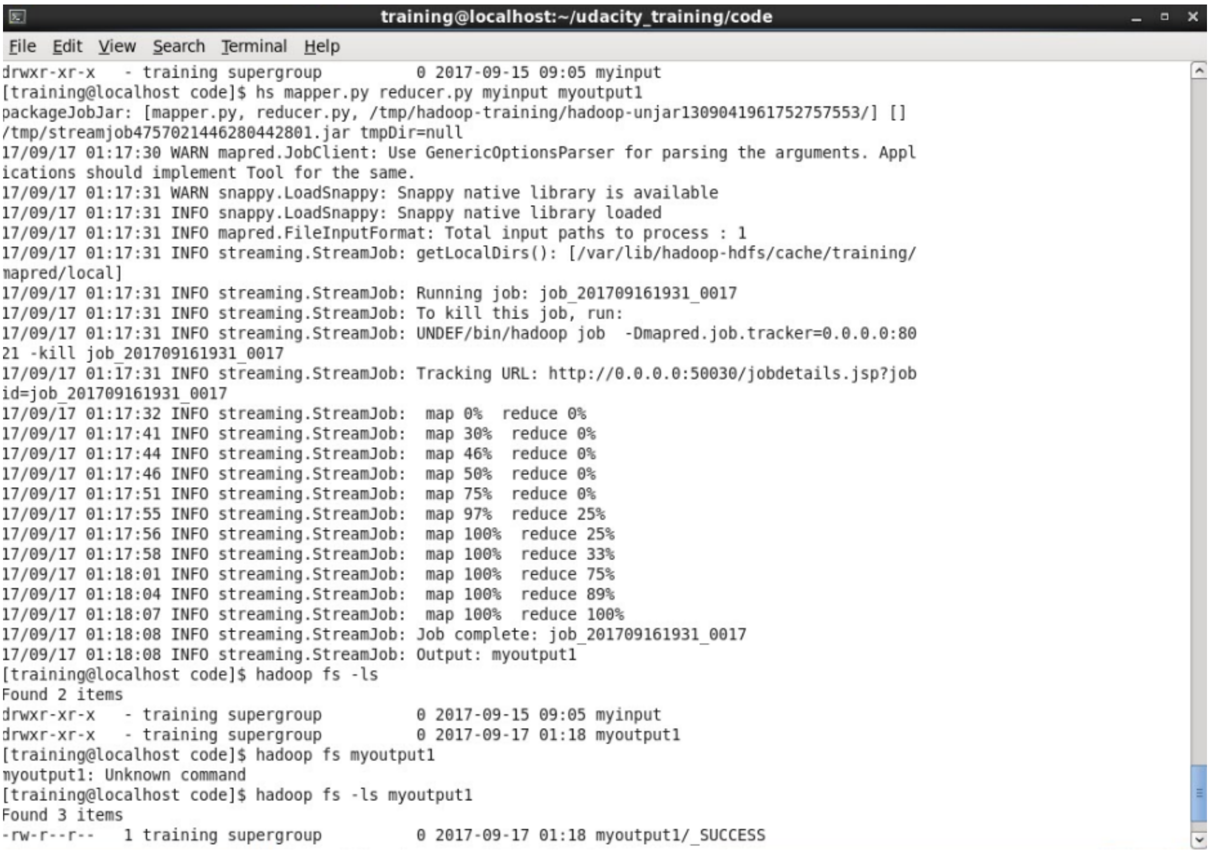
1. Write Python program for the following three questions of project part 1.

   1) Find a sales breakdown by product category across all of our stores.

   2) Find the monetary value for the highest individual sale for each separate store.

   3) Find the total sales value across all the stores, and the total number of sales. (Assume there is only one reducer.

2. Take a screen shot after running MapReduce code for question 1. Copy and paste the mapper and reducer code for question 1. Copy and paste the result for question 1.


training@localhost:~/udacity_training/code
File Edit View Search Terminal Help

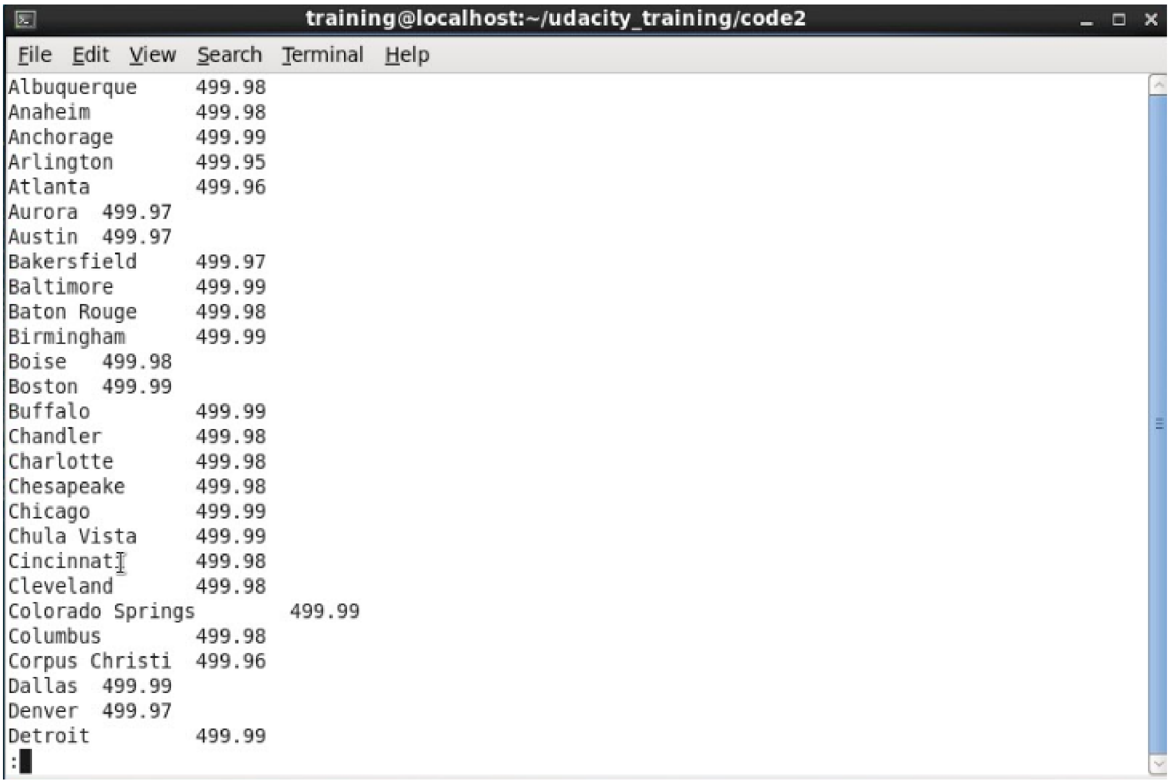
```
drwxr-xr-x   - training supergroup          0 2017-09-15 09:05 myinput
[training@localhost code]$ hs mapper.py reducer.py myinput myoutput1
packageJobJar: [mapper.py, reducer.py, /tmp/hadoop-training/hadoop-unjar1309041961752757553/] []
/tmp/streamjob4757021446280442801.jar tmpDir=null
17/09/17 01:17:30 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. Appl
ications should implement Tool for the same.
17/09/17 01:17:31 WARN snappy.LoadSnappy: Snappy native library is available
17/09/17 01:17:31 INFO snappy.LoadSnappy: Snappy native library loaded
17/09/17 01:17:31 INFO mapred.FileInputFormat: Total input paths to process : 1
17/09/17 01:17:31 INFO streaming.StreamJob: getLocalDirs(): [/var/lib/hadoop-hdfs/cache/training/
mapred/local]
17/09/17 01:17:31 INFO streaming.StreamJob: Running job: job_201709161931_0017
17/09/17 01:17:31 INFO streaming.StreamJob: To kill this job, run:
17/09/17 01:17:31 INFO streaming.StreamJob: UNDEF/bin/hadoop job  -Dmapred.job.tracker=0.0.0.0:80
21 -kill job_201709161931_0017
17/09/17 01:17:31 INFO streaming.StreamJob: Tracking URL: http://0.0.0.0:50030/jobdetails.jsp?job
id=job_201709161931_0017
17/09/17 01:17:32 INFO streaming.StreamJob:  map 0%  reduce 0%
17/09/17 01:17:41 INFO streaming.StreamJob:  map 30%  reduce 0%
17/09/17 01:17:44 INFO streaming.StreamJob:  map 46%  reduce 0%
17/09/17 01:17:46 INFO streaming.StreamJob:  map 50%  reduce 0%
17/09/17 01:17:51 INFO streaming.StreamJob:  map 75%  reduce 0%
17/09/17 01:17:55 INFO streaming.StreamJob:  map 97%  reduce 25%
17/09/17 01:17:56 INFO streaming.StreamJob:  map 100%  reduce 25%
17/09/17 01:17:58 INFO streaming.StreamJob:  map 100%  reduce 33%
17/09/17 01:18:01 INFO streaming.StreamJob:  map 100%  reduce 75%
17/09/17 01:18:04 INFO streaming.StreamJob:  map 100%  reduce 89%
17/09/17 01:18:07 INFO streaming.StreamJob:  map 100%  reduce 100%
17/09/17 01:18:08 INFO streaming.StreamJob: Job complete: job_201709161931_0017
17/09/17 01:18:08 INFO streaming.StreamJob: Output: myoutput1
[training@localhost code]$ hadoop fs -ls
Found 2 items
drwxr-xr-x   - training supergroup          0 2017-09-15 09:05 myinput
drwxr-xr-x   - training supergroup          0 2017-09-17 01:18 myoutput1
[training@localhost code]$ hadoop fs myoutput1
myoutput1: Unknown command
[training@localhost code]$ hadoop fs -ls myoutput1
Found 3 items
-rw-r--r--   1 training supergroup          0 2017-09-17 01:18 myoutput1/_SUCCESS
```

Mapper.py:

```
#!/usr/bin/python


import sys

for line in sys.stdin:
    data = line.strip().split("\t")
    if len(data) == 6:
        date, time, store, item, cost, payment = data
        print "{0}\t{1}".format(item, cost)
```

Reducer.py:

```
#!/usr/bin/python

import sys

salesTotal = 0
oldKey = None


for line in sys.stdin:
    data_mapped = line.strip().split("\t")
    if len(data_mapped) != 2:
        # Something has gone wrong. Skip this line.
        continue

    thisKey, thisSale = data_mapped

    if oldKey and oldKey != thisKey:
        print oldKey, "\t", salesTotal
        oldKey = thisKey;
        salesTotal = 0

    oldKey = thisKey
    salesTotal += float(thisSale)

if oldKey != None:
    print oldKey, "\t", salesTotal
```

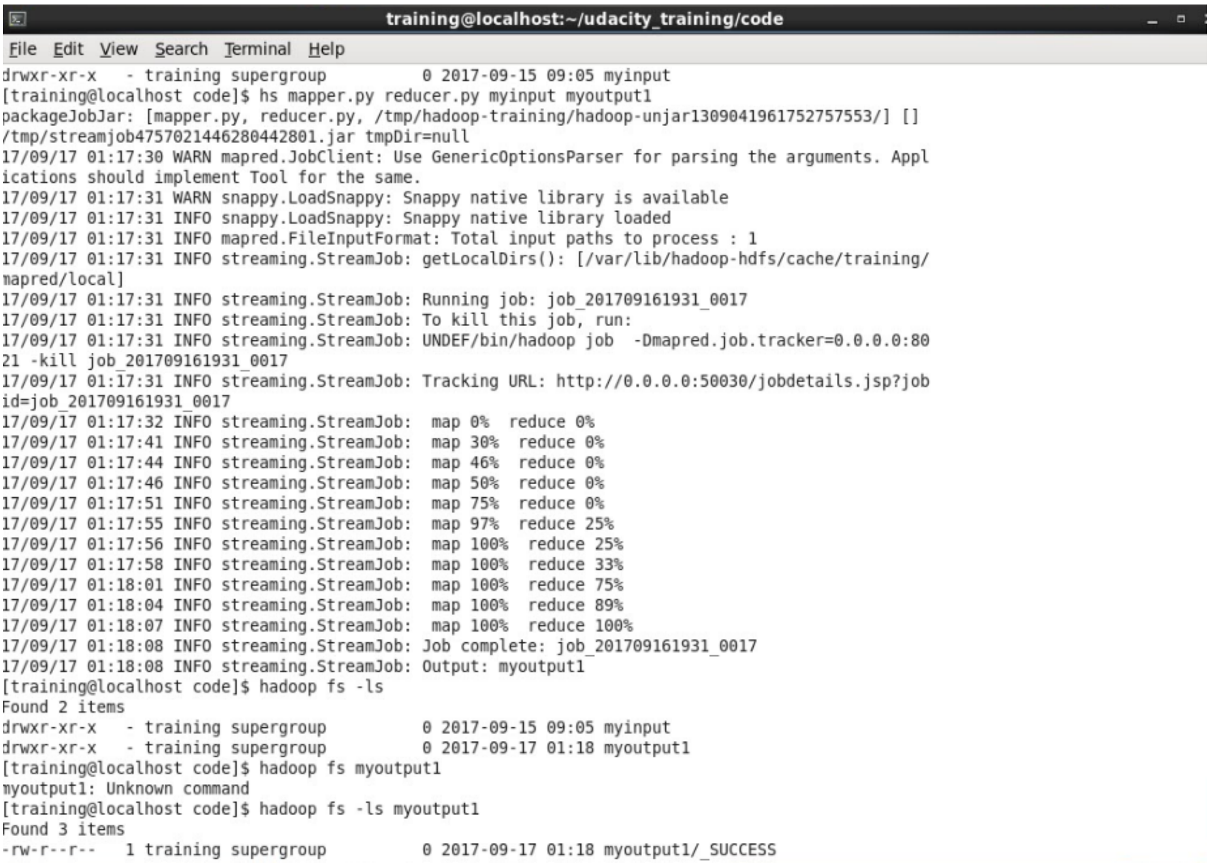

training@localhost:~/udacity_training/code2
File Edit View Search Terminal Help

```
Albuquerque     499.98
Anaheim         499.98
Anchorage       499.99
Arlington       499.95
Atlanta         499.96
Aurora  499.97
Austin  499.97
Bakersfield     499.97
Baltimore       499.99
Baton Rouge     499.98
Birmingham      499.99
Boise   499.98
Boston  499.99
Buffalo         499.99
Chandler        499.98
Charlotte       499.98
Chesapeake      499.98
Chicago         499.99
Chula Vista     499.99
Cincinnati      499.98
Cleveland       499.98
Colorado Springs        499.99
Columbus        499.98
Corpus Christi  499.96
Dallas  499.99
Denver  499.97
Detroit         499.99
:
```

3. Take a screen shot after running MapReduce code for question 2. Copy and paste the mapper and reducer code for question 2. What are the values for the following store:

Anchorage **499.99**

Bakersfield	**499.97**
Colorado Springs	**499.99**



training@localhost:~/udacity_training/code
File Edit View Search Terminal Help

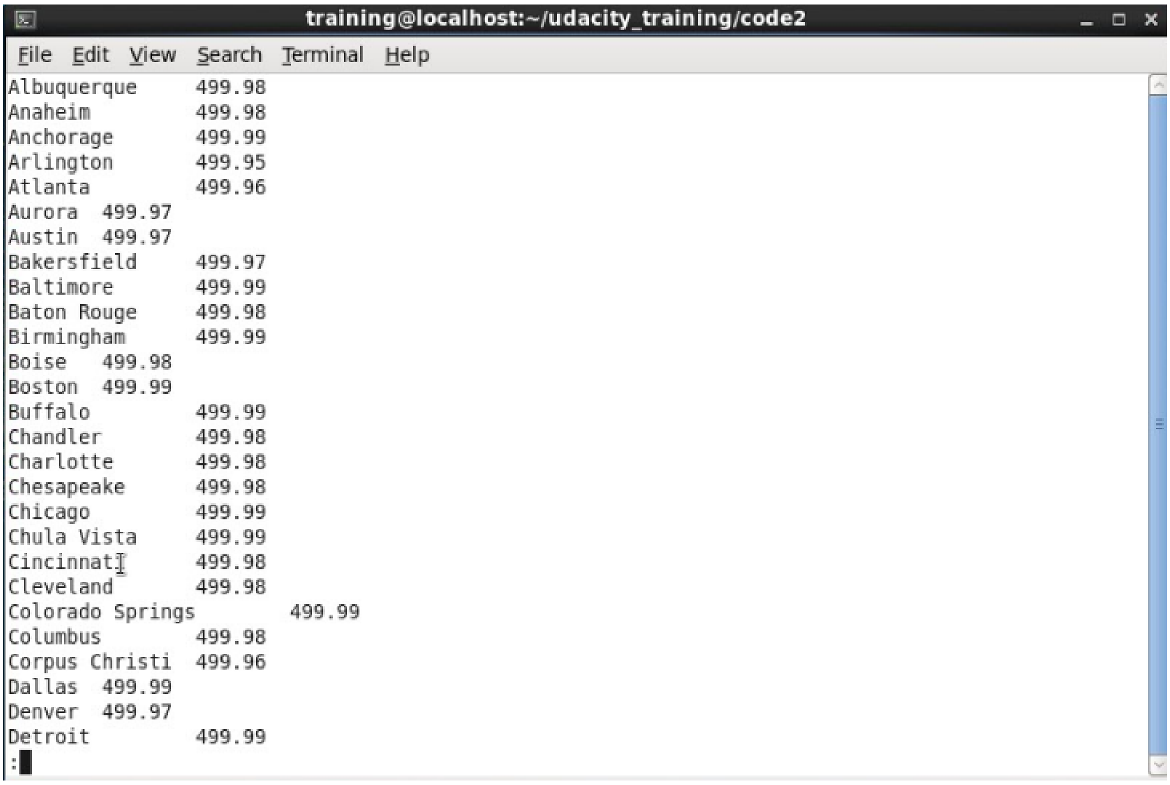
```
drwxr-xr-x   - training supergroup          0 2017-09-15 09:05 myinput
[training@localhost code]$ hs mapper.py reducer.py myinput myoutput1
packageJobJar: [mapper.py, reducer.py, /tmp/hadoop-training/hadoop-unjar1309041961752757553/] []
/tmp/streamjob4757021446280442801.jar tmpDir=null
17/09/17 01:17:30 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. Appl
ications should implement Tool for the same.
17/09/17 01:17:31 WARN snappy.LoadSnappy: Snappy native library is available
17/09/17 01:17:31 INFO snappy.LoadSnappy: Snappy native library loaded
17/09/17 01:17:31 INFO mapred.FileInputFormat: Total input paths to process : 1
17/09/17 01:17:31 INFO streaming.StreamJob: getLocalDirs(): [/var/lib/hadoop-hdfs/cache/training/
mapred/local]
17/09/17 01:17:31 INFO streaming.StreamJob: Running job: job_201709161931_0017
17/09/17 01:17:31 INFO streaming.StreamJob: To kill this job, run:
17/09/17 01:17:31 INFO streaming.StreamJob: UNDEF/bin/hadoop job  -Dmapred.job.tracker=0.0.0.0:80
21 -kill job_201709161931_0017
17/09/17 01:17:31 INFO streaming.StreamJob: Tracking URL: http://0.0.0.0:50030/jobdetails.jsp?job
id=job_201709161931_0017
17/09/17 01:17:32 INFO streaming.StreamJob:  map 0%  reduce 0%
17/09/17 01:17:41 INFO streaming.StreamJob:  map 30%  reduce 0%
17/09/17 01:17:44 INFO streaming.StreamJob:  map 46%  reduce 0%
17/09/17 01:17:46 INFO streaming.StreamJob:  map 50%  reduce 0%
17/09/17 01:17:51 INFO streaming.StreamJob:  map 75%  reduce 0%
17/09/17 01:17:55 INFO streaming.StreamJob:  map 97%  reduce 25%
17/09/17 01:17:56 INFO streaming.StreamJob:  map 100%  reduce 25%
17/09/17 01:17:58 INFO streaming.StreamJob:  map 100%  reduce 33%
17/09/17 01:18:01 INFO streaming.StreamJob:  map 100%  reduce 75%
17/09/17 01:18:04 INFO streaming.StreamJob:  map 100%  reduce 89%
17/09/17 01:18:07 INFO streaming.StreamJob:  map 100%  reduce 100%
17/09/17 01:18:08 INFO streaming.StreamJob: Job complete: job_201709161931_0017
17/09/17 01:18:08 INFO streaming.StreamJob: Output: myoutput1
[training@localhost code]$ hadoop fs -ls
Found 2 items
drwxr-xr-x   - training supergroup          0 2017-09-15 09:05 myinput
drwxr-xr-x   - training supergroup          0 2017-09-17 01:18 myoutput1
[training@localhost code]$ hadoop fs myoutput1
myoutput1: Unknown command
[training@localhost code]$ hadoop fs -ls myoutput1
Found 3 items
-rw-r--r--   1 training supergroup          0 2017-09-17 01:18 myoutput1/_SUCCESS
```

Mapper.py:

```
#!/usr/bin/python


import sys

for line in sys.stdin:
    data = line.strip().split("\t")
    if len(data) == 6:
        date, time, store, item, cost, payment = data
        print "{0}\t{1}".format(store, cost)
```

Reducer.py:

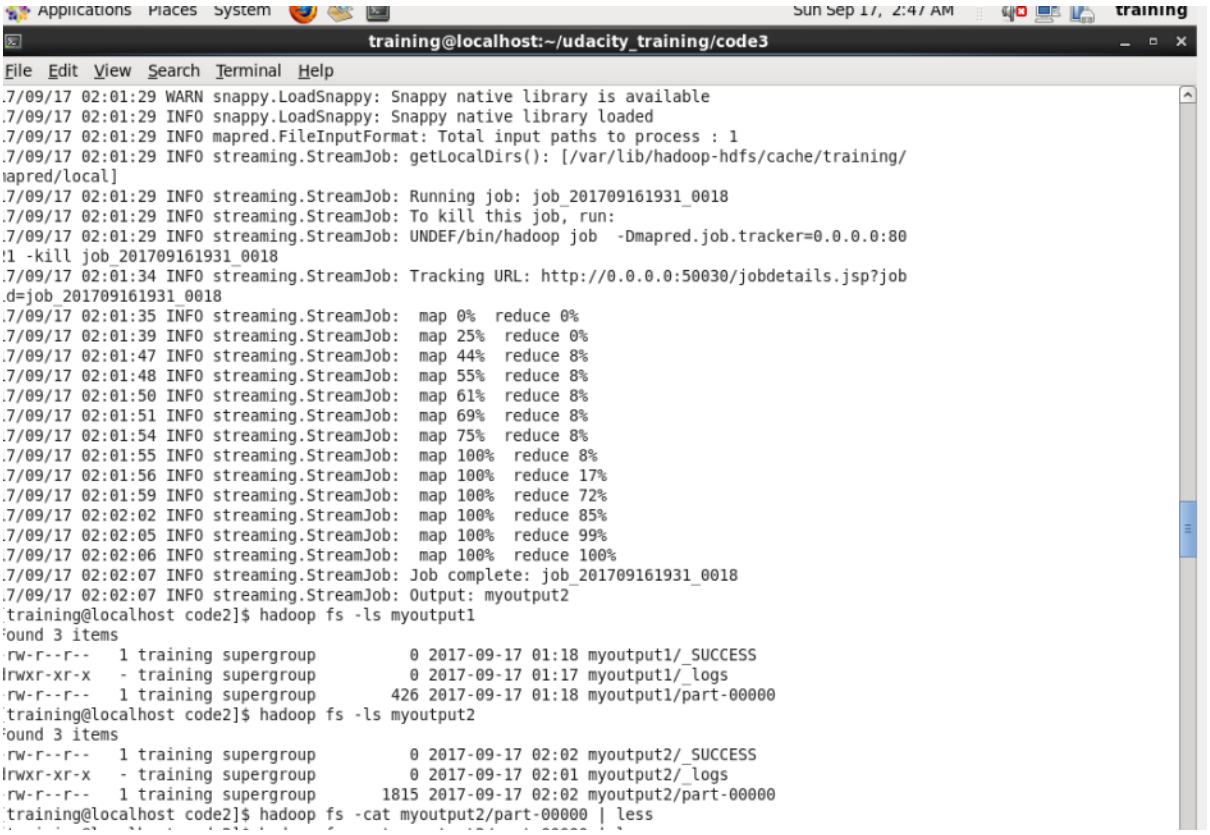
```
#!/usr/bin/python

import sys

sales = 0
oldKey = None


for line in sys.stdin:
    data_mapped = line.strip().split("\t")
    if len(data_mapped) != 2:
        continue

    thisKeyStore, thisSale = data_mapped

    if oldKey and oldKey != thisKeyStore:
        print oldKey, "\t", sales
        oldKey = thisKeyStore;
        sales = 0

    oldKey = thisKeyStore
    #sales += float(thisSale)
    if sales < float(thisSale):
    sales = float(thisSale)

if oldKey != None:
    print oldKey, "\t", sales
```



training@localhost:~/udacity_training/code2
File Edit View Search Terminal Help

```
Albuquerque     499.98
Anaheim         499.98
Anchorage       499.99
Arlington       499.95
Atlanta         499.96
Aurora  499.97
Austin  499.97
Bakersfield     499.97
Baltimore       499.99
Baton Rouge     499.98
Birmingham      499.99
Boise   499.98
Boston  499.99
Buffalo         499.99
Chandler        499.98
Charlotte       499.98
Chesapeake      499.98
Chicago         499.99
Chula Vista     499.99
Cincinnati      499.98
Cleveland       499.98
Colorado Springs        499.99
Columbus        499.98
Corpus Christi  499.96
Dallas  499.99
Denver  499.97
Detroit         499.99
:
```

4. Take a screen shot after running MapReduce code for question 3. Copy and paste the mapper and reducer code for question 3. What is the total number of sales and the total sales value from all the stores?


Applications Places System
Sun Sep 17, 2:47 AM
training
training@localhost:~/udacity_training/code3
File Edit View Search Terminal Help

```
.7/09/17 02:01:29 WARN snappy.LoadSnappy: Snappy native library is available
.7/09/17 02:01:29 INFO snappy.LoadSnappy: Snappy native library loaded
.7/09/17 02:01:29 INFO mapred.FileInputFormat: Total input paths to process : 1
.7/09/17 02:01:29 INFO streaming.StreamJob: getLocalDirs(): [/var/lib/hadoop-hdfs/cache/training/
apred/local]
.7/09/17 02:01:29 INFO streaming.StreamJob: Running job: job_201709161931_0018
.7/09/17 02:01:29 INFO streaming.StreamJob: To kill this job, run:
.7/09/17 02:01:29 INFO streaming.StreamJob: UNDEF/bin/hadoop job  -Dmapred.job.tracker=0.0.0.0:80
!1 -kill job_201709161931_0018
.7/09/17 02:01:34 INFO streaming.StreamJob: Tracking URL: http://0.0.0.0:50030/jobdetails.jsp?job
.d=job_201709161931_0018
.7/09/17 02:01:35 INFO streaming.StreamJob:  map 0%  reduce 0%
.7/09/17 02:01:39 INFO streaming.StreamJob:  map 25%  reduce 0%
.7/09/17 02:01:47 INFO streaming.StreamJob:  map 44%  reduce 8%
.7/09/17 02:01:48 INFO streaming.StreamJob:  map 55%  reduce 8%
.7/09/17 02:01:50 INFO streaming.StreamJob:  map 61%  reduce 8%
.7/09/17 02:01:51 INFO streaming.StreamJob:  map 69%  reduce 8%
.7/09/17 02:01:54 INFO streaming.StreamJob:  map 75%  reduce 8%
.7/09/17 02:01:55 INFO streaming.StreamJob:  map 100%  reduce 8%
.7/09/17 02:01:56 INFO streaming.StreamJob:  map 100%  reduce 17%
.7/09/17 02:01:59 INFO streaming.StreamJob:  map 100%  reduce 72%
.7/09/17 02:02:02 INFO streaming.StreamJob:  map 100%  reduce 85%
.7/09/17 02:02:05 INFO streaming.StreamJob:  map 100%  reduce 99%
.7/09/17 02:02:06 INFO streaming.StreamJob:  map 100%  reduce 100%
.7/09/17 02:02:07 INFO streaming.StreamJob: Job complete: job_201709161931_0018
.7/09/17 02:02:07 INFO streaming.StreamJob: Output: myoutput2
training@localhost code2]$ hadoop fs -ls myoutput1
ound 3 items
rw-r--r--   1 training supergroup          0 2017-09-17 01:18 myoutput1/_SUCCESS
lrwxr-xr-x   - training supergroup          0 2017-09-17 01:17 myoutput1/_logs
rw-r--r--   1 training supergroup        426 2017-09-17 01:18 myoutput1/part-00000
training@localhost code2]$ hadoop fs -ls myoutput2
ound 3 items
rw-r--r--   1 training supergroup          0 2017-09-17 02:02 myoutput2/_SUCCESS
lrwxr-xr-x   - training supergroup          0 2017-09-17 02:01 myoutput2/_logs
rw-r--r--   1 training supergroup       1815 2017-09-17 02:02 myoutput2/part-00000
training@localhost code2]$ hadoop fs -cat myoutput2/part-00000 | less
```

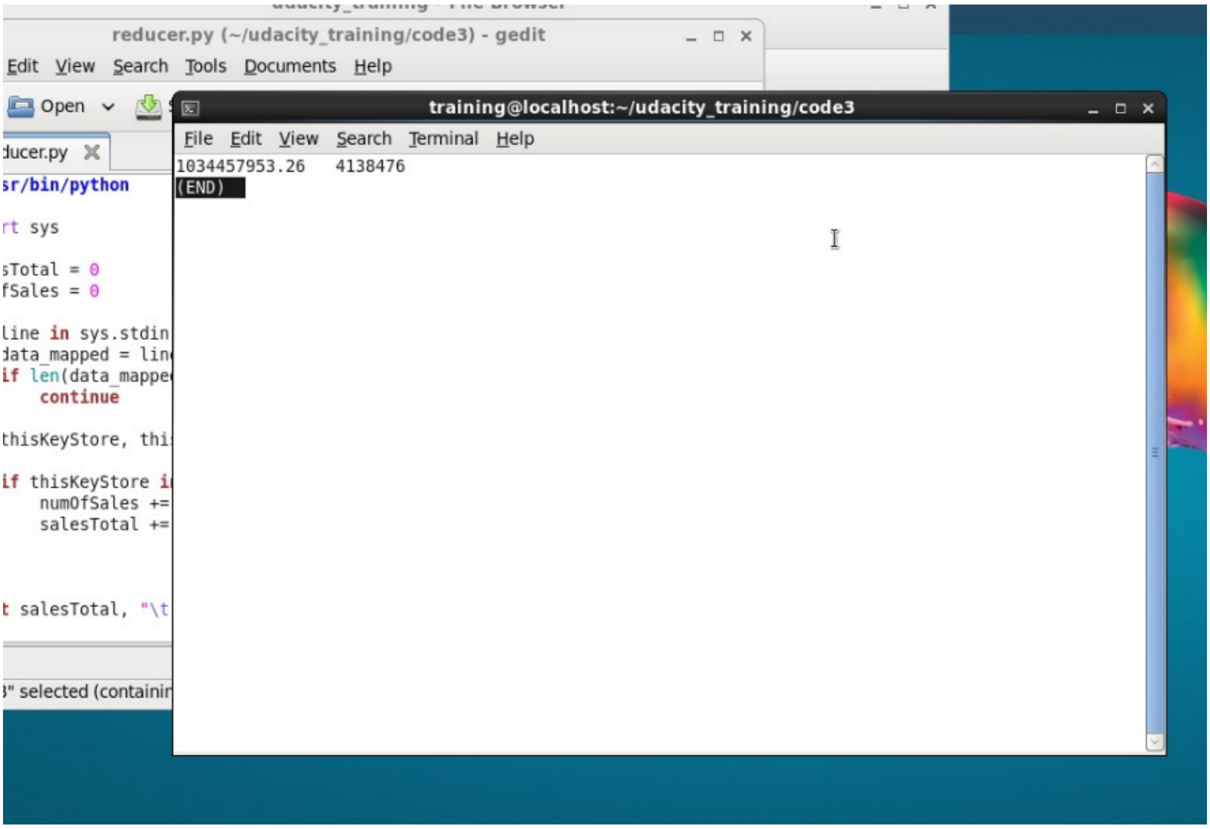
Mapper.py:

```
#!/usr/bin/python


import sys

for line in sys.stdin:
    data = line.strip().split("\t")
    if len(data) == 6:
        date, time, store, item, cost, payment = data
        print "{0}\t{1}".format(store, cost)
```

Reducer.py:

```
#!/usr/bin/python

import sys

salesTotal = 0
numOfSales = 0

for line in sys.stdin:
    data_mapped = line.strip().split("\t")
    if len(data_mapped) != 2:
        continue

    thisKeyStore, thisCost = data_mapped

    if thisKeyStore in line:
    numOfSales += 1
```

```
        salesTotal += float(thisCost)



print salesTotal, "\t", numOfSales
```


reducer.py (~/udacity_training/code3) - gedit
Edit View Search Tools Documents Help
Open
ducer.py
training@localhost:~/udacity_training/code3
File Edit View Search Terminal Help
3" selected (containir

```
sr/bin/python

rt sys

sTotal = 0
fSales = 0

line in sys.stdin
data_mapped = lin
if len(data_mappe
    continue

thisKeyStore, thi

if thisKeyStore i
    numOfSales +=
    salesTotal +=


t salesTotal, "\t
```

```
1034457953.26    4138476
(END)
```