DATASCI W261: Machine Learning at Scale

Group: 10
Names:
Prabhakar Gundugola,
Yi Jin,
Jaime Villalpando

Emails:
prabhakar@berkeley.edu,
yjin@ischool.berkeley.edu,
jaimegvl@ischool.berkeley.edu

Time of Initial Submission: Feb 2, 2016
Week 3: Homework 3
Date: February 4, 2016
Time of Submission: 00:10 AM PT

HW3.0

What is a merge sort?

Merge sort is an efficient, general-purpose, comparison based sorting algorithm for rearranging lists into a specified order.

Mergesort algorithm

Mergesort works as follows:

- Divide the unsorted list into n sublists, each containing only 1 element.
- Merge sublists repeatedly into sorted sublists until there is only 1 sublist remaining.

Where is it used in Hadoop? Mergesort is used in sort and shuffle phase of hadoop between Map and Reduce phases.

How is a combiner function in the context of Hadoop?

A combiner, also known as a semi-reducer, accepts the inputs from the Map procedure and thereafter passes the output of key,value pairs to the Reduce procedure.

It is used in between Map and Reduce procedures to reduce the volume of data transfer between Map and Reduce when the output of Map phase is very large.

Combiner - Multiple reducers

Give an example where it can be used and justify why it should be used in the context of this problem.

An example where a combiner is required is word count in large number of documents. A map emits a (key, value) pair with (word, 1) for each and every word in the document. The output of Map phase is very large and to reduce the volume of data transfer to reduce phase, we need a combiner that aggregates the values by key.

What is the Hadoop shuffle?

Hadoop shuffle is the process of transferring data from mappers to reducers based on a partitioning function. It sorts and combines all the data based on a partitioning key and ensures that all the (key, value) pairs of the same key are sent to the same reducer.

HW 3.1. Use Counters to do EDA (exploratory data analysis and to monitor progress)

Counters are lightweight objects in Hadoop that allow you to keep track of system progress in both the map and reduce stages of processing. By default, Hadoop defines a number of standard counters in "groups"; these show up in the jobtracker webapp, giving you information such as "Map input records", "Map output records", etc.

While processing information/data using MapReduce job, it is a challenge to monitor the progress of parallel threads running across nodes of distributed clusters. Moreover, it is also complicated to distinguish between the data that has been processed and the data which is yet to be processed. The MapReduce Framework offers a provision of user-defined Counters, which can be effectively utilized to monitor the progress of data across nodes of distributed clusters.

Use the Consumer Complaints Dataset provide here to complete this question:

```
https://www.dropbox.com/s/vbalm3yva2rr86m/Consumer_Complaints.csv?dl=0
```

The consumer complaints dataset consists of diverse consumer complaints, which have been reported across the United States regarding various types of loans. The dataset consists of records of the form:

Complaint ID,Product,Sub-product,Issue,Sub-issue,State,ZIP code,Submitted via,Date received,Date sent to company,Company,Company response,Timely response?,Consumer disputed?

Here’s is the first few lines of the of the Consumer Complaints Dataset:

Complaint ID,Product,Sub-product,Issue,Sub-issue,State,ZIP code,Submitted via,Date received,Date sent to company,Company,Company response,Timely response?,Consumer disputed? 1114245,Debt collection,Medical,Disclosure verification of debt,Not given enough info to verify debt,FL,32219,Web,11/13/2014,11/13/2014,"Choice Recovery, Inc.",Closed with explanation,Yes, 1114488,Debt collection,Medical,Disclosure verification of debt,Right to dispute notice not received,TX,75006,Web,11/13/2014,11/13/2014,"Expert Global Solutions, Inc.",In progress,Yes, 1114255,Bank account or service,Checking account,Deposits and withdrawals,,NY,11102,Web,11/13/2014,11/13/2014,"FNIS (Fidelity National Information Services, Inc.)",In progress,Yes, 1115106,Debt collection,"Other (phone, health club, etc.)",Communication tactics,Frequent or repeated calls,GA,31721,Web,11/13/2014,11/13/2014,"Expert Global Solutions, Inc.",In progress,Yes,

User-defined Counters

Now, let’s use Hadoop Counters to identify the number of complaints pertaining to debt collection, mortgage and other categories (all other categories get lumped into this one) in the consumer complaints dataset. Basically produce the distribution of the Product column in this dataset using counters (limited to 3 counters here).

Hadoop offers Job Tracker, an UI tool to determine the status and statistics of all jobs. Using the job tracker UI, developers can view the Counters that have been created. Screenshot your job tracker UI as your job completes and include it here. Make sure that your user defined counters are visible.

In [71]:

```
%%writefile mapper31.py
#!/usr/bin/python
## mapper31.py
## Author: Prabhakar Gundugola
## Description: mapper code for HW3.1

import sys

for line in sys.stdin:
    tokens = line.strip().split(",")

    # Skip the Header
    if tokens[0] == 'Complaint ID':
        continue

    product = 'none'
    if 'Debt' in tokens[1]:
        product = 'debt'
    elif 'Mortgage' in tokens[1]:
        product = 'mortgage'
    else:
        product = 'others'

    sys.stderr.write("reporter:counter:MapperTokens," + product +
',1\n')
    print product + '\t' + str(1)
```

```
Overwriting mapper31.py
```

In [72]:

```
%%writefile reducer31.py
#!/usr/bin/python
## reducer31.py
## Author: Prabhakar Gundugola
## Description: reducer code for HW3.1

import sys

prev_word = None
counts = 0
for line in sys.stdin:
    word, value = line.strip().split('\t')

    if prev_word != word:
        if prev_word is not None:
            print prev_word + '\t' + str(counts)
            sys.stderr.write('reporter:counter:ReducerTokens,'
                             + prev_word + ',' + str(counts) + '\n')

        prev_word = word
        counts = 0
    counts += 1

print prev_word + '\t' + str(counts)
sys.stderr.write('reporter:counter:ReducerTokens,' + prev_word + ',' + s
tr(counts) + '\n')
```

```
Overwriting reducer31.py
```

In [73]:

```
!chmod a+x mapper31.py
!chmod a+x reducer31.py
```

In [4]:

```
# Ensure hw31 folder doesn't exist
!hdfs dfs -rm -r /user/root/wk3/hw31

# Create HDFS input and src folder
!hdfs dfs -mkdir -p /user/root/wk3/hw31/input

# Copy the input file, mapper.py, reducer.py
!hdfs dfs -put Consumer_Complaints.csv /user/root/wk3/hw31/input
```

```
16/01/31 23:17:41 INFO fs.TrashPolicyDefault: Namenode trash configurat
ion: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/root/wk3/hw31
```

Run Hadoop Streaming job

In [5]:

```
# Ensure output folder doesn't exist
!hdfs dfs -rm -r /user/root/wk3/hw31/output

# Run Hadoop Streaming job
!hadoop jar hadoop-streaming-2.7.1.jar \
-mapper /root/hw3/mapper31.py \
-reducer /root/hw3/reducer31.py \
-input /user/root/wk3/hw31/input \
-output /user/root/wk3/hw31/output
```

```
rm: `/user/root/wk3/hw31/output': No such file or directory
packageJobJar: [/tmp/hadoop-unjar5817919191444807337/] [] /tmp/streamjo
b5060557907352090611.jar tmpDir=null
16/01/31 23:17:53 INFO client.RMProxy: Connecting to ResourceManager at
/0.0.0.0:8032
16/01/31 23:17:54 INFO client.RMProxy: Connecting to ResourceManager at
/0.0.0.0:8032
16/01/31 23:17:54 INFO mapred.FileInputFormat: Total input paths to pro
cess : 1
16/01/31 23:17:54 INFO mapreduce.JobSubmitter: number of splits:2
16/01/31 23:17:54 INFO mapreduce.JobSubmitter: Submitting tokens for jo
b: job_1454301000890_0001
16/01/31 23:17:55 INFO impl.YarnClientImpl: Submitted application appli
cation_1454301000890_0001
16/01/31 23:17:55 INFO mapreduce.Job: The url to track the job: htt
p://prabhakar:8088/proxy/application_1454301000890_0001/
16/01/31 23:17:55 INFO mapreduce.Job: Running job: job_1454301000890_00
01
16/01/31 23:18:03 INFO mapreduce.Job: Job job_1454301000890_0001 runnin
g in uber mode : false
16/01/31 23:18:03 INFO mapreduce.Job:  map 0% reduce 0%
16/01/31 23:18:11 INFO mapreduce.Job:  map 100% reduce 0%
16/01/31 23:18:18 INFO mapreduce.Job:  map 100% reduce 100%
16/01/31 23:18:18 INFO mapreduce.Job: Job job_1454301000890_0001 comple
ted successfully
16/01/31 23:18:18 INFO mapreduce.Job: Counters: 55
        File System Counters
                FILE: Number of bytes read=3604798
                FILE: Number of bytes written=7562077
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=50910129
                HDFS: Number of bytes written=41
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=118
00
                Total time spent by all reduces in occupied slots (m
s)=4541
                Total time spent by all map tasks (ms)=11800
                Total time spent by all reduce tasks (ms)=4541
                Total vcore-seconds taken by all map tasks=11800
                Total vcore-seconds taken by all reduce tasks=4541
                Total megabyte-seconds taken by all map tasks=12083200
                Total megabyte-seconds taken by all reduce tasks=464998
4
        Map-Reduce Framework
```

```
                Map input records=312913
                Map output records=312912
                Map output bytes=2978968
                Map output materialized bytes=3604804
                Input split bytes=246
                Combine input records=0
                Combine output records=0
                Reduce input groups=3
                Reduce shuffle bytes=3604804
                Reduce input records=312912
                Reduce output records=3
                Spilled Records=625824
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=248
                CPU time spent (ms)=9170
                Physical memory (bytes) snapshot=690049024
                Virtual memory (bytes) snapshot=2528423936
                Total committed heap usage (bytes)=598212608
        MapperTokens
                debt=44372
                mortgage=125752
                others=142788
        ReducerTokens
                debt=44372
                mortgage=125752
                others=142788
        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=50909883
        File Output Format Counters
                Bytes Written=41
16/01/31 23:18:18 INFO streaming.StreamJob: Output directory: /user/roo
t/wk3/hw31/output
```

As shown in the output:

- debt=44372
- mortgage=125752
- others=142788

HW 3.2. Analyze the performance of your Mappers, Combiners and Reducers using Counters

a) For this brief study the Input file will be one record (the next line only): foo foo quux labs foo bar quux

Perform a word count analysis of this single record dataset using a Mapper and Reducer based WordCount (i.e., no combiners are used here) using user defined Counters to count up how many time the mapper and reducer are called. What is the value of your user defined Mapper Counter, and Reducer Counter after completing this word count job. The answer should be 1 and 4 respectively. Please explain.

In [6]:
```
!echo "foo foo quux labs foo bar quux" > input_data.txt
```

In [6]:
```
%%writefile mapper32a.py
#!/usr/bin/python
## mapper32a.py
## Author: Prabhakar Gundugola
## Description: mapper code for HW3.2
import sys

sys.stderr.write('reporter:counter:mapper,Mapper,1\n')

for line in sys.stdin:
    words = line.strip().split()

    for word in words:
        print word + '\t' + str(1)
```

```
Overwriting mapper32a.py
```

In [7]:
```
%%writefile reducer32a.py
#!/usr/bin/python
## reducer32a.py
## Author: Prabhakar Gundugola
## Description: reducer code for HW3.2
import sys

for line in sys.stdin:
    print line

sys.stderr.write('reporter:counter:mapper,Reducer,1\n')
```

```
Overwriting reducer32a.py
```

In [8]:

```
!chmod a+x mapper32a.py
!chmod a+x reducer32a.py
```

In [9]:

```
# Ensure the input folder doesn't exist
!hdfs dfs -rm -r /user/root/wk3/hw32a

# Create HDFS directory for input folder
!hdfs dfs -mkdir -p /user/root/wk3/hw32a/input

# Copy input data
!hdfs dfs -put input_data.txt /user/root/wk3/hw32a/input
```

```
16/01/31 14:27:35 INFO fs.TrashPolicyDefault: Namenode trash configurat
ion: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/root/wk3/hw32a
```

In [10]:

```
# Ensure the output folder doesn't exist
!hdfs dfs -rm -r /user/root/wk3/hw32a/output

# Run Hadoop Streaming job
!hadoop jar hadoop-streaming-2.7.1.jar \
-D mapred.map.tasks=1 \
-D mapred.reduce.tasks=4 \
-mapper /root/hw3/mapper32a.py \
-reducer /root/hw3/reducer32a.py \
-input /user/root/wk3/hw32a/input \
-output /user/root/wk3/hw32a/output
```

```
rm: `/user/root/wk3/hw32a/output': No such file or directory
packageJobJar: [/tmp/hadoop-unjar2188181261195739559/] [] /tmp/streamjo
b9135154725253275562.jar tmpDir=null
16/01/31 14:27:47 INFO client.RMProxy: Connecting to ResourceManager at
/0.0.0.0:8032
16/01/31 14:27:47 INFO client.RMProxy: Connecting to ResourceManager at
/0.0.0.0:8032
16/01/31 14:27:48 INFO mapred.FileInputFormat: Total input paths to pro
cess : 1
16/01/31 14:27:48 INFO mapreduce.JobSubmitter: number of splits:1
16/01/31 14:27:48 INFO Configuration.deprecation: mapred.reduce.tasks i
s deprecated. Instead, use mapreduce.job.reduces
16/01/31 14:27:48 INFO Configuration.deprecation: mapred.map.tasks is d
eprecated. Instead, use mapreduce.job.maps
16/01/31 14:27:48 INFO mapreduce.JobSubmitter: Submitting tokens for jo
b: job_1454270249092_0011
16/01/31 14:27:48 INFO impl.YarnClientImpl: Submitted application appli
cation_1454270249092_0011
16/01/31 14:27:48 INFO mapreduce.Job: The url to track the job: htt
p://prabhakar:8088/proxy/application_1454270249092_0011/
16/01/31 14:27:48 INFO mapreduce.Job: Running job: job_1454270249092_00
11
16/01/31 14:27:54 INFO mapreduce.Job: Job job_1454270249092_0011 runnin
g in uber mode : false
16/01/31 14:27:54 INFO mapreduce.Job:  map 0% reduce 0%
16/01/31 14:28:00 INFO mapreduce.Job:  map 100% reduce 0%
16/01/31 14:28:07 INFO mapreduce.Job:  map 100% reduce 25%
16/01/31 14:28:09 INFO mapreduce.Job:  map 100% reduce 50%
16/01/31 14:28:10 INFO mapreduce.Job:  map 100% reduce 100%
16/01/31 14:28:10 INFO mapreduce.Job: Job job_1454270249092_0011 comple
ted successfully
16/01/31 14:28:10 INFO mapreduce.Job: Counters: 51
        File System Counters
                FILE: Number of bytes read=83
                FILE: Number of bytes written=587583
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=146
                HDFS: Number of bytes written=59
                HDFS: Number of read operations=15
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=8
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=4
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=329
6
                Total time spent by all reduces in occupied slots (m
s)=20191
                Total time spent by all map tasks (ms)=3296
                Total time spent by all reduce tasks (ms)=20191
                Total vcore-seconds taken by all map tasks=3296
                Total vcore-seconds taken by all reduce tasks=20191
                Total megabyte-seconds taken by all map tasks=3375104
                Total megabyte-seconds taken by all reduce tasks=206755
84
        Map-Reduce Framework
                Map input records=1
                Map output records=7
                Map output bytes=45
                Map output materialized bytes=83
                Input split bytes=115
                Combine input records=0
                Combine output records=0
                Reduce input groups=4
                Reduce shuffle bytes=83
                Reduce input records=7
                Reduce output records=14
                Spilled Records=14
                Shuffled Maps =4
                Failed Shuffles=0
                Merged Map outputs=4
                GC time elapsed (ms)=402
                CPU time spent (ms)=5090
                Physical memory (bytes) snapshot=938647552
                Virtual memory (bytes) snapshot=4204924928
                Total committed heap usage (bytes)=1006632960
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        mapper
                Mapper=1
                Reducer=4
        File Input Format Counters
                Bytes Read=31
        File Output Format Counters
                Bytes Written=59
16/01/31 14:28:10 INFO streaming.StreamJob: Output directory: /user/roo
t/wk3/hw32a/output
```

Mapper Counter: 1
Reducer Counter: 4

The default output counters for mapper and reducer are 2 and 1 when I didn't pass the properties for mapper and reducer in Hadoop Streaming command. I had to explicitly set the counters for mapper and reducer as 1 and 4 to produce the output counters 1 and 4.

b) Please use mulitple mappers and reducers for these jobs (at least 2 mappers and 2 reducers).
Perform a word count analysis of the Issue column of the Consumer Complaints Dataset using a Mapper and Reducer based WordCount (i.e., no combiners used anywhere) using user defined Counters to count up how many time the mapper and reducer are called. What is the value of your user defined Mapper Counter, and Reducer Counter after completing your word count job.

In [36]:

```
%%writefile mapper32b.py
#!/usr/bin/python
## mapper32b.py
## Author: Prabhakar Gundugola
## Description: mapper code for HW3.2b
import sys
import string

sys.stderr.write('reporter:counter:mapper32b,Mapper,1\n')
total_words = 0
for line in sys.stdin:
    tokens = line.strip().split(",")
    if 'Complaint' in tokens[0]:
        continue

    word_string = tokens[3].replace(',', ' ').replace('/', ' ').replac
e('"', '')
    for word in word_string.lower().split():
        total_words += 1
        print word + '\t' + str(1)
print '0000TOTALWORDS' + '\t' + str(total_words)
```

```
Overwriting mapper32b.py
```

In [17]:

```
%%writefile reducer32b.py
#!/usr/bin/python
## reducer32b.py
## Author: Prabhakar Gundugola
## Description: reducer code for HW3.2b
import sys

sys.stderr.write('reporter:counter:reducer32b,Reducer,1\n')
prev_word = None
counts = 0

for line in sys.stdin:
    word, value = line.strip().split('\t')

    if prev_word != word:
        if prev_word is not None:
            print prev_word + '\t' + str(counts)
        prev_word = word
        counts = 0
    counts += eval(value)
print prev_word + '\t' + str(counts)
```

```
Overwriting reducer32b.py
```

In [13]:

```
!chmod a+x mapper32b.py
!chmod a+x reducer32b.py
```

In [14]:

```
# Ensure the input folder doesn't exist
!hdfs dfs -rm -r /user/root/wk3/hw32b

# Create Input folder
!hdfs dfs -mkdir -p /user/root/wk3/hw32b/input

# Copy the input file to input folder
!hdfs dfs -put Consumer_Complaints.csv /user/root/wk3/hw32b/input
```

```
16/01/31 14:29:01 INFO fs.TrashPolicyDefault: Namenode trash configurat
ion: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/root/wk3/hw32b
```

In [37]:

```
# Ensure the output folder doesn't exist
!hdfs dfs -rm -r /user/root/wk3/hw32b/output


# Run Hadoop Streaming job
!hadoop jar hadoop-streaming-2.7.1.jar \
-D mapred.reduce.tasks=4 \
-mapper /root/hw3/mapper32b.py \
-reducer /root/hw3/reducer32b.py \
-input /user/root/wk3/hw32b/input \
-output /user/root/wk3/hw32b/output
```

```
16/01/31 15:07:28 INFO fs.TrashPolicyDefault: Namenode trash configurat
ion: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/root/wk3/hw32b/output
packageJobJar: [/tmp/hadoop-unjar3847755633682334463/] [] /tmp/streamjo
b5197053891155922287.jar tmpDir=null
16/01/31 15:07:31 INFO client.RMProxy: Connecting to ResourceManager at
/0.0.0.0:8032
16/01/31 15:07:31 INFO client.RMProxy: Connecting to ResourceManager at
/0.0.0.0:8032
16/01/31 15:07:32 INFO mapred.FileInputFormat: Total input paths to pro
cess : 1
16/01/31 15:07:32 INFO mapreduce.JobSubmitter: number of splits:2
16/01/31 15:07:32 INFO Configuration.deprecation: mapred.reduce.tasks i
s deprecated. Instead, use mapreduce.job.reduces
16/01/31 15:07:32 INFO mapreduce.JobSubmitter: Submitting tokens for jo
b: job_1454270249092_0022
16/01/31 15:07:32 INFO impl.YarnClientImpl: Submitted application appli
cation_1454270249092_0022
16/01/31 15:07:32 INFO mapreduce.Job: The url to track the job: htt
p://prabhakar:8088/proxy/application_1454270249092_0022/
16/01/31 15:07:32 INFO mapreduce.Job: Running job: job_1454270249092_00
22
16/01/31 15:07:38 INFO mapreduce.Job: Job job_1454270249092_0022 runnin
g in uber mode : false
16/01/31 15:07:38 INFO mapreduce.Job:  map 0% reduce 0%
16/01/31 15:07:46 INFO mapreduce.Job:  map 100% reduce 0%
16/01/31 15:07:56 INFO mapreduce.Job:  map 100% reduce 25%
16/01/31 15:07:59 INFO mapreduce.Job:  map 100% reduce 100%
16/01/31 15:07:59 INFO mapreduce.Job: Job job_1454270249092_0022 comple
ted successfully
16/01/31 15:07:59 INFO mapreduce.Job: Counters: 51
        File System Counters
                FILE: Number of bytes read=11233537
                FILE: Number of bytes written=23172100
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=50910131
                HDFS: Number of bytes written=2113
                HDFS: Number of read operations=18
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=8
        Job Counters 
                Launched map tasks=2
                Launched reduce tasks=4
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=113
78
                Total time spent by all reduces in occupied slots (m
s)=32035
                Total time spent by all map tasks (ms)=11378
                Total time spent by all reduce tasks (ms)=32035
                Total vcore-seconds taken by all map tasks=11378
                Total vcore-seconds taken by all reduce tasks=32035
                Total megabyte-seconds taken by all map tasks=11651072
                Total megabyte-seconds taken by all reduce tasks=328038
40
        Map-Reduce Framework
                Map input records=312913
                Map output records=980484
                Map output bytes=9272545
                Map output materialized bytes=11233561
                Input split bytes=248
                Combine input records=0
                Combine output records=0
                Reduce input groups=170
                Reduce shuffle bytes=11233561
                Reduce input records=980484
                Reduce output records=170
                Spilled Records=1960968
                Shuffled Maps =8
                Failed Shuffles=0
                Merged Map outputs=8
                GC time elapsed (ms)=527
                CPU time spent (ms)=14760
                Physical memory (bytes) snapshot=1190985728
                Virtual memory (bytes) snapshot=5073149952
                Total committed heap usage (bytes)=1204813824
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        mapper32b
                Mapper=2
        File Input Format Counters
                Bytes Read=50909883
        File Output Format Counters
                Bytes Written=2113
        reducer32b
                Reducer=4
16/01/31 15:07:59 INFO streaming.StreamJob: Output directory: /user/roo
t/wk3/hw32b/output
```

Mapper Counter: 2
Reducer Counter: 4

The default output counters for mapper and reducer are 2 and 1. I had to explicitly set the counter for reducer to 4 to produce the output counters 2 and 4 for mapper and reducer.

c) Perform a word count analysis of the Issue column of the Consumer Complaints Dataset using a Mapper, Reducer, and standalone combiner (i.e., not an in-memory combiner) based WordCount using user defined Counters to count up how many time the mapper, combiner, reducer are called.
What is the value of your user defined Mapper Counter, and Reducer Counter after completing your word count job.

In [19]:

```
%%writefile combiner32c.py
#!/usr/bin/python
## combiner32c.py
## Author: Prabhakar Gundugola
## Description: combiner code for HW3.2c
import sys

sys.stderr.write('reporter:counter:combiner32c,Combiner,1\n')
prev_word = None
counts = 0

for line in sys.stdin:
    word, value = line.strip().split('\t')

    if prev_word != word:
        if prev_word is not None:
            print prev_word + '\t' + str(counts)
        prev_word = word
        counts = 0
    counts += eval(value)
print prev_word + '\t' + str(counts)
```

```
Overwriting combiner32c.py
```

In [44]:

```
!chmod a+x combiner32c.py
```

In [24]:

```
# Ensure the input folder doesn't exist
!hdfs dfs -rm -r /user/root/wk3/hw32c

# Create Input folder
!hdfs dfs -mkdir -p /user/root/wk3/hw32c/input

# Copy the input file to input folder
!hdfs dfs -put Consumer_Complaints.csv /user/root/wk3/hw32c/input
```

```
16/01/31 14:10:43 INFO fs.TrashPolicyDefault: Namenode trash configurat
ion: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/root/wk3/hw32c
```

In [20]:

```
# Ensure the output folder doesn't exist
!hdfs dfs -rm -r /user/root/wk3/hw32c/output


# Run Hadoop Streaming job.
!hadoop jar hadoop-streaming-2.7.1.jar \
-D mapred.reduce.tasks=4 \
-mapper /root/hw3/mapper32b.py \
-combiner /root/hw3/combiner32c.py \
-reducer /root/hw3/reducer32b.py \
-input /user/root/wk3/hw32c/input \
-output /user/root/wk3/hw32c/output
```

```
16/01/31 14:31:51 INFO fs.TrashPolicyDefault: Namenode trash configurat
ion: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/root/wk3/hw32c/output
packageJobJar: [/tmp/hadoop-unjar7939414701922345542/] [] /tmp/streamjo
b8041904735065720818.jar tmpDir=null
16/01/31 14:31:53 INFO client.RMProxy: Connecting to ResourceManager at
/0.0.0.0:8032
16/01/31 14:31:54 INFO client.RMProxy: Connecting to ResourceManager at
/0.0.0.0:8032
16/01/31 14:31:54 INFO mapred.FileInputFormat: Total input paths to pro
cess : 1
16/01/31 14:31:54 INFO mapreduce.JobSubmitter: number of splits:2
16/01/31 14:31:54 INFO Configuration.deprecation: mapred.reduce.tasks i
s deprecated. Instead, use mapreduce.job.reduces
16/01/31 14:31:54 INFO mapreduce.JobSubmitter: Submitting tokens for jo
b: job_1454270249092_0014
16/01/31 14:31:54 INFO impl.YarnClientImpl: Submitted application appli
cation_1454270249092_0014
16/01/31 14:31:55 INFO mapreduce.Job: The url to track the job: htt
p://prabhakar:8088/proxy/application_1454270249092_0014/
16/01/31 14:31:55 INFO mapreduce.Job: Running job: job_1454270249092_00
14
16/01/31 14:32:01 INFO mapreduce.Job: Job job_1454270249092_0014 runnin
g in uber mode : false
16/01/31 14:32:01 INFO mapreduce.Job:  map 0% reduce 0%
16/01/31 14:32:12 INFO mapreduce.Job:  map 100% reduce 0%
16/01/31 14:32:18 INFO mapreduce.Job:  map 100% reduce 25%
16/01/31 14:32:20 INFO mapreduce.Job:  map 100% reduce 50%
16/01/31 14:32:21 INFO mapreduce.Job:  map 100% reduce 100%
16/01/31 14:32:22 INFO mapreduce.Job: Job job_1454270249092_0014 comple
ted successfully
16/01/31 14:32:22 INFO mapreduce.Job: Counters: 52
        File System Counters
                FILE: Number of bytes read=4525
                FILE: Number of bytes written=716098
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=50910131
                HDFS: Number of bytes written=2128
                HDFS: Number of read operations=18
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=8
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=4
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=176
64
                Total time spent by all reduces in occupied slots (m
s)=19239
                Total time spent by all map tasks (ms)=17664
                Total time spent by all reduce tasks (ms)=19239
                Total vcore-seconds taken by all map tasks=17664
                Total vcore-seconds taken by all reduce tasks=19239
                Total megabyte-seconds taken by all map tasks=18087936
                Total megabyte-seconds taken by all reduce tasks=197007
36
        Map-Reduce Framework
                Map input records=312913
                Map output records=966249
                Map output bytes=9210210
                Map output materialized bytes=4549
                Input split bytes=248
                Combine input records=966249
                Combine output records=311
                Reduce input groups=168
                Reduce shuffle bytes=4549
                Reduce input records=311
                Reduce output records=168
                Spilled Records=622
                Shuffled Maps =8
                Failed Shuffles=0
                Merged Map outputs=8
                GC time elapsed (ms)=448
                CPU time spent (ms)=10780
                Physical memory (bytes) snapshot=1188081664
                Virtual memory (bytes) snapshot=5059936256
                Total committed heap usage (bytes)=1207959552
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        combiner32c
                Combiner=8
        mapper32b
                Mapper=2
        File Input Format Counters
                Bytes Read=50909883
        File Output Format Counters
                Bytes Written=2128
        reducer32b
                Reducer=4
16/01/31 14:32:22 INFO streaming.StreamJob: Output directory: /user/roo
t/wk3/hw32c/output
```

The counters produced by Hadoop Mapreduce job are:

- Mapper - 2
- Combiner - 8
- Reducer - 4

Using a single reducer: What are the top 50 most frequent terms in your word count analysis? Present the top 50 terms and their frequency and their relative frequency. Present the top 50 terms and their frequency and their relative frequency. If there are ties please sort the tokens in alphanumeric/string order. Present bottom 10 tokens (least frequent items).

In [96]:

```
%%writefile mapper32d.py
#!/usr/bin/python
## mapper32d.py
## Author: Prabhakar Gundugola
## Description: mapper code for HW3.2d that takes the output of HW3.2c a
s input
import sys

sys.stderr.write('reporter:counter:mapper,Mapper32d,1\n')

for line in sys.stdin:
    word, value = line.strip().split('\t')
    print value + '\t' + word
```

```
Overwriting mapper32d.py
```

In [97]:

```
%%writefile reducer32d.py
#!/usr/bin/python
## reducer32d.py
## Author: Prabhakar Gundugola
## Description: reducer code for HW3.2d
import sys

sys.stderr.write('reporter:counter:reducer,Reducer32d,1\n')

total = 0
for line in sys.stdin:
    value, word = line.strip().split('\t')
    # First word should be 0000TOTALWORDS
    if word == '0000TOTALWORDS':
        total = int(value)
    else:
        term_freq = 100.0 * int(value)/total
        print word.ljust(20) + '\t' + value + '\t' + str(round(term_fre
q,4)) + '%'
```

```
Overwriting reducer32d.py
```

In [59]:

```
!chmod a+x mapper32d.py
!chmod a+x reducer32d.py
```

In [29]:

```
# Ensure the input folder doesn't exist
!hdfs dfs -rm -r /user/root/wk3/hw32d

# Create Input folder
!hdfs dfs -mkdir -p /user/root/wk3/hw32d/input

# Copy the input file to input folder
!hdfs dfs -put Consumer_Complaints.csv /user/root/wk3/hw32d/input
```

```
16/01/31 14:59:47 INFO fs.TrashPolicyDefault: Namenode trash configurat
ion: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/root/wk3/hw32d
```

In [98]:

```
# Ensure the output folder doesn't exist
!hdfs dfs -rm -r /user/root/wk3/hw32d/output


# Run Hadoop Streaming job.
!hadoop jar hadoop-streaming-2.7.1.jar \
-D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator \
-D mapred.text.key.partitioner.options=-k1,1 \
-D stream.num.map.output.key.fields=2 \
-D mapred.text.key.comparator.options='-k1,1nr -k2,2n' \
-mapper /root/hw3/mapper32d.py \
-reducer /root/hw3/reducer32d.py \
-input /user/root/wk3/hw32b/output/part* \
-output /user/root/wk3/hw32d/output
```

```
16/01/31 16:52:54 INFO fs.TrashPolicyDefault: Namenode trash configurat
ion: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/root/wk3/hw32d/output
packageJobJar: [/tmp/hadoop-unjar6767512836284184782/] [] /tmp/streamjo
b1590576393051470913.jar tmpDir=null
16/01/31 16:52:57 INFO client.RMProxy: Connecting to ResourceManager at
/0.0.0.0:8032
16/01/31 16:52:57 INFO client.RMProxy: Connecting to ResourceManager at
/0.0.0.0:8032
16/01/31 16:52:58 INFO mapred.FileInputFormat: Total input paths to pro
cess : 4
16/01/31 16:52:58 INFO mapreduce.JobSubmitter: number of splits:4
16/01/31 16:52:58 INFO Configuration.deprecation: mapred.output.key.com
parator.class is deprecated. Instead, use mapreduce.job.output.key.comp
arator.class
16/01/31 16:52:58 INFO Configuration.deprecation: mapred.text.key.compa
rator.options is deprecated. Instead, use mapreduce.partition.keycompar
ator.options
16/01/31 16:52:58 INFO Configuration.deprecation: mapred.text.key.parti
tioner.options is deprecated. Instead, use mapreduce.partition.keyparti
tioner.options
16/01/31 16:52:58 INFO mapreduce.JobSubmitter: Submitting tokens for jo
b: job_1454270249092_0045
16/01/31 16:52:58 INFO impl.YarnClientImpl: Submitted application appli
cation_1454270249092_0045
16/01/31 16:52:58 INFO mapreduce.Job: The url to track the job: htt
p://prabhakar:8088/proxy/application_1454270249092_0045/
16/01/31 16:52:58 INFO mapreduce.Job: Running job: job_1454270249092_00
45
16/01/31 16:53:04 INFO mapreduce.Job: Job job_1454270249092_0045 runnin
g in uber mode : false
16/01/31 16:53:04 INFO mapreduce.Job:  map 0% reduce 0%
16/01/31 16:53:12 INFO mapreduce.Job:  map 100% reduce 0%
16/01/31 16:53:19 INFO mapreduce.Job:  map 100% reduce 100%
16/01/31 16:53:19 INFO mapreduce.Job: Job job_1454270249092_0045 comple
ted successfully
16/01/31 16:53:20 INFO mapreduce.Job: Counters: 51
        File System Counters
                FILE: Number of bytes read=2629
                FILE: Number of bytes written=595976
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=2561
                HDFS: Number of bytes written=5694
                HDFS: Number of read operations=15
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=4
                Launched reduce tasks=1
                Data-local map tasks=4
                Total time spent by all maps in occupied slots (ms)=240
```

```
43
                Total time spent by all reduces in occupied slots (m
s)=3360
                Total time spent by all map tasks (ms)=24043
                Total time spent by all reduce tasks (ms)=3360
                Total vcore-seconds taken by all map tasks=24043
                Total vcore-seconds taken by all reduce tasks=3360
                Total megabyte-seconds taken by all map tasks=24620032
                Total megabyte-seconds taken by all reduce tasks=344064
0
        Map-Reduce Framework
                Map input records=170
                Map output records=170
                Map output bytes=2283
                Map output materialized bytes=2647
                Input split bytes=448
                Combine input records=0
                Combine output records=0
                Reduce input groups=170
                Reduce shuffle bytes=2647
                Reduce input records=170
                Reduce output records=169
                Spilled Records=340
                Shuffled Maps =4
                Failed Shuffles=0
                Merged Map outputs=4
                GC time elapsed (ms)=314
                CPU time spent (ms)=3940
                Physical memory (bytes) snapshot=1191534592
                Virtual memory (bytes) snapshot=4193255424
                Total committed heap usage (bytes)=1006632960
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        mapper
                Mapper32d=4
        File Input Format Counters
                Bytes Read=2113
        File Output Format Counters
                Bytes Written=5694
        reducer
                Reducer32d=1
16/01/31 16:53:20 INFO streaming.StreamJob: Output directory: /user/roo
t/wk3/hw32d/output
```

In [99]:

```
!hdfs dfs -tail /user/root/wk3/hw32d/output/part-00000 |tail -10
```

```
apply                   118     0.012%
amount                  98      0.01%
credited                92      0.0094%
payment                 92      0.0094%
convenience             75      0.0076%
checks                  75      0.0076%
amt                     71      0.0072%
day                     71      0.0072%
disclosures             64      0.0065%
missing                 64      0.0065%
```

In [100]:

```
!hdfs dfs -cat /user/root/wk3/hw32d/output/part-00000 |head -50
```

```
loan                    119630  12.2011%
modification            70487   7.189%
credit                  55251   5.6351%
servicing               36767   3.7499%
report                  34903   3.5598%
incorrect               29133   2.9713%
information             29069   2.9648%
on                      29069   2.9648%
or                      22533   2.2982%
account                 20681   2.1093%
debt                    19309   1.9693%
and                     16448   1.6775%
opening                 16205   1.6528%
club                    12545   1.2795%
health                  12545   1.2795%
not                     12353   1.2599%
attempts                11848   1.2084%
collect                 11848   1.2084%
cont'd                  11848   1.2084%
owed                    11848   1.2084%
of                      10885   1.1102%
my                      10731   1.0945%
deposits                10555   1.0765%
withdrawals             10555   1.0765%
problems                9484    0.9673%
application             8868    0.9045%
to                      8401    0.8568%
unable                  8178    0.8341%
billing                 8158    0.832%
other                   7886    0.8043%
disputes                6938    0.7076%
communication           6920    0.7058%
tactics                 6920    0.7058%
reporting               6559    0.669%
lease                   6337    0.6463%
the                     6248    0.6372%
by                      5663    0.5776%
being                   5663    0.5776%
caused                  5663    0.5776%
funds                   5663    0.5776%
low                     5663    0.5776%
process                 5505    0.5615%
disclosure              5214    0.5318%
verification            5214    0.5318%
managing                5006    0.5106%
company's               4858    0.4955%
investigation           4858    0.4955%
identity                4729    0.4823%
card                    4405    0.4493%
get                     4357    0.4444%
```

In [74]:

```
!hdfs dfs -cat /user/root/wk3/hw32d/output/part-00000|wc
```

```
    169     507    5694
```

HW3.3. Shopping Cart Analysis

Product Recommendations: The action or practice of selling additional products or services to existing customers is called cross-selling. Giving product recommendation is one of the examples of cross-selling that are frequently used by online retailers. One simple method to give product recommendations is to recommend products that are frequently browsed together by the customers.

For this homework use the online browsing behavior dataset located at:

```
https://www.dropbox.com/s/zlfyiwa70poqg74/ProductPurchaseData.txt?dl=0
```

Each line in this dataset represents a browsing session of a customer. On each line, each string of 8 characters represents the id of an item browsed during that session. The items are separated by spaces.

Here are the first few lines of the ProductPurchaseData FRO11987 ELE17451 ELE89019 SNA90258 GRO99222 GRO99222 GRO12298 FRO12685 ELE91550 SNA11465 ELE26917 ELE52966 FRO90334 SNA30755 ELE17451 FRO84225 SNA80192 ELE17451 GRO73461 DAI22896 SNA99873 FRO86643 ELE17451 ELE37798 FRO86643 GRO56989 ELE23393 SNA11465 ELE17451 SNA69641 FRO86643 FRO78087 SNA11465 GRO39357 ELE28573 ELE11375 DAI54444

Do some exploratory data analysis of this dataset.

How many unique items are available from this supplier?

Using a single reducer: Report your findings such as number of unique products; largest basket; report the top 50 most frequently purchased items, their frequency, and their relative frequency (break ties by sorting the products alphabetical order) etc. using Hadoop Map-Reduce.

In [118]:

```
%%writefile mapper33a.py
#!/usr/bin/python
## mapper33a.py
## Author: Prabhakar Gundugola
## Description: mapper code for HW3.3
import sys

sys.stderr.write('reporter:counter:mapper,mapper331,1\n')
total_products = 0
basket = 0
largest_basket = 0
for line in sys.stdin:
    products = line.strip().split()
    for product in products:
        total_products += 1
        basket += 1
        print product + '\t' + str(1)
    if basket > largest_basket:
        largest_basket = basket
    basket = 0
print '0000TOTALPRODUCTS' + '\t' + str(total_products)
print '0000LARGESTBASKET' + '\t' + str(largest_basket)
```

```
Overwriting mapper33a.py
```

In [119]:

```
%%writefile reducer33a.py
#!/usr/bin/python
## reducer33.py
## Author: Prabhakar Gundugola
## Description: reducer code for HW3.3
import sys

sys.stderr.write('reporter:counter:reducer,reducer33,1\n')
prev_product = None
counts = 0
total = 0
unique_count = 0
largest_basket = 0
for line in sys.stdin:
    product, value = line.strip().split('\t')
    if prev_product != product:
        if prev_product is not None:
            if prev_product != '0000LARGESTBASKET':
                print prev_product + '\t' + str(counts)
                if prev_product != '0000TOTALWORDS':
                    unique_count += 1
            else:
                print prev_product + '\t' + str(largest_basket)

        prev_product = product
        counts = 0
    if product == '0000LARGESTBASKET':
        if int(value) > largest_basket:
            largest_basket = int(value)
    else:
        counts += int(value)
unique_count += 1
print prev_product + '\t' + str(counts)
print '0000UNIQUECOUNT' + '\t' + str(unique_count)
```

```
Overwriting reducer33a.py
```

In [107]:

```
!chmod a+x mapper33a.py
!chmod a+x reducer33a.py
```

In [105]:

```
# Ensure the input folder doesn't exist
!hdfs dfs -rm -r /user/root/wk3/hw33a

# Create Input folder
!hdfs dfs -mkdir -p /user/root/wk3/hw33a/input

# Copy the input file to input folder
!hdfs dfs -put ProductPurchaseData.txt /user/root/wk3/hw33a/input
```

```
rm: `/user/root/wk3/hw33a': No such file or directory
```

In [120]:

```
# Ensure output folder doesn't exist
!hdfs dfs -rm -r /user/root/wk3/hw33a/output

# Run Hadoop Streaming job
!hadoop jar hadoop-streaming-2.7.1.jar \
-mapper /root/hw3/mapper33a.py \
-reducer /root/hw3/reducer33a.py \
-input /user/root/wk3/hw33a/input \
-output /user/root/wk3/hw33a/output

#-D mapred.text.key.comparator.options='-k1,1n -k2,2nr' \
16/01/31 17:16:35 INFO fs.TrashPolicyDefault: Namenode trash configurat
ion: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/root/wk3/hw33a/output
packageJobJar: [/tmp/hadoop-unjar5488039784815785888/] [] /tmp/streamjo
b1187851644835129363.jar tmpDir=null
16/01/31 17:16:38 INFO client.RMProxy: Connecting to ResourceManager at
/0.0.0.0:8032
16/01/31 17:16:38 INFO client.RMProxy: Connecting to ResourceManager at
/0.0.0.0:8032
16/01/31 17:16:39 INFO mapred.FileInputFormat: Total input paths to pro
cess : 1
16/01/31 17:16:39 INFO mapreduce.JobSubmitter: number of splits:2
16/01/31 17:16:39 INFO mapreduce.JobSubmitter: Submitting tokens for jo
b: job_1454270249092_0051
16/01/31 17:16:39 INFO impl.YarnClientImpl: Submitted application appli
cation_1454270249092_0051
16/01/31 17:16:39 INFO mapreduce.Job: The url to track the job: htt
p://prabhakar:8088/proxy/application_1454270249092_0051/
16/01/31 17:16:39 INFO mapreduce.Job: Running job: job_1454270249092_00
51
16/01/31 17:16:45 INFO mapreduce.Job: Job job_1454270249092_0051 runnin
g in uber mode : false
16/01/31 17:16:45 INFO mapreduce.Job:  map 0% reduce 0%
16/01/31 17:16:52 INFO mapreduce.Job:  map 100% reduce 0%
16/01/31 17:16:59 INFO mapreduce.Job:  map 100% reduce 100%
16/01/31 17:16:59 INFO mapreduce.Job: Job job_1454270249092_0051 comple
ted successfully
16/01/31 17:17:00 INFO mapreduce.Job: Counters: 51
        File System Counters
                FILE: Number of bytes read=4950818
                FILE: Number of bytes written=10254129
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=3462115
                HDFS: Number of bytes written=142726
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=928
8
                Total time spent by all reduces in occupied slots (m
s)=4539
                Total time spent by all map tasks (ms)=9288
                Total time spent by all reduce tasks (ms)=4539
                Total vcore-seconds taken by all map tasks=9288
                Total vcore-seconds taken by all reduce tasks=4539
                Total megabyte-seconds taken by all map tasks=9510912
                Total megabyte-seconds taken by all reduce tasks=464793
```

```
6
        Map-Reduce Framework
                Map input records=31101
                Map output records=380828
                Map output bytes=4189156
                Map output materialized bytes=4950824
                Input split bytes=248
                Combine input records=0
                Combine output records=0
                Reduce input groups=12594
                Reduce shuffle bytes=4950824
                Reduce input records=380828
                Reduce output records=12595
                Spilled Records=761656
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=106
                CPU time spent (ms)=6730
                Physical memory (bytes) snapshot=699924480
                Virtual memory (bytes) snapshot=2515263488
                Total committed heap usage (bytes)=601882624
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        mapper
                mapper331=2
        File Input Format Counters
                Bytes Read=3461867
        File Output Format Counters
                Bytes Written=142726
        reducer
                reducer33=1
16/01/31 17:17:00 INFO streaming.StreamJob: Output directory: /user/roo
t/wk3/hw33a/output
```

In [17]:

```
%%writefile reducer33b.py
#!/usr/bin/python
## reducer33b.py
## Author: Prabhakar Gundugola
## Description: reducer code for HW3.3
import sys

sys.stderr.write('reporter:counter:reducer,Reducer32d,1\n')

total = 0
for line in sys.stdin:
    value, word = line.strip().split('\t')
    # First word should be 0000TOTALWORDS
    if word == '0000TOTALPRODUCTS':
        total = int(value)
    elif word == '0000UNIQUECOUNT':
        print word.ljust(20) + '\t' + value
    elif word == '0000LARGESTBASKET':
        print word.ljust(20) + '\t' + value
    else:
        term_freq = round(100.0 * int(value)/total, 3)
        print word.ljust(20) + '\t' + value + '\t' + str(term_freq) +
'%'
```

```
Overwriting reducer33b.py
```

In [135]:

```
!chmod a+x reducer33b.py
```

In [144]:

```
# Ensure the output folder doesn't exist
!hdfs dfs -rm -r /user/root/wk3/hw33b/output



# Run Hadoop Streaming job.
!hadoop jar hadoop-streaming-2.7.1.jar \
-D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFi
eldBasedComparator \
-D mapred.text.key.partitioner.options=-k1,1 \
-D stream.num.map.output.key.fields=2 \
-D mapred.text.key.comparator.options='-k1,1nr -k2,2n' \
-mapper /root/hw3/mapper32d.py \
-reducer /root/hw3/reducer33b.py \
-input /user/root/wk3/hw33a/output/part* \
-output /user/root/wk3/hw33b/output
```

```
16/01/31 17:36:54 INFO fs.TrashPolicyDefault: Namenode trash configurat
ion: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/root/wk3/hw33b/output
packageJobJar: [/tmp/hadoop-unjar2487017460346512839/] [] /tmp/streamjo
b7490952225334260520.jar tmpDir=null
16/01/31 17:36:57 INFO client.RMProxy: Connecting to ResourceManager at
/0.0.0.0:8032
16/01/31 17:36:58 INFO client.RMProxy: Connecting to ResourceManager at
/0.0.0.0:8032
16/01/31 17:36:58 INFO mapred.FileInputFormat: Total input paths to pro
cess : 1
16/01/31 17:36:58 INFO mapreduce.JobSubmitter: number of splits:2
16/01/31 17:36:58 INFO Configuration.deprecation: mapred.output.key.com
parator.class is deprecated. Instead, use mapreduce.job.output.key.comp
arator.class
16/01/31 17:36:58 INFO Configuration.deprecation: mapred.text.key.compa
rator.options is deprecated. Instead, use mapreduce.partition.keycompar
ator.options
16/01/31 17:36:58 INFO Configuration.deprecation: mapred.text.key.parti
tioner.options is deprecated. Instead, use mapreduce.partition.keyparti
tioner.options
16/01/31 17:36:58 INFO mapreduce.JobSubmitter: Submitting tokens for jo
b: job_1454270249092_0055
16/01/31 17:36:58 INFO impl.YarnClientImpl: Submitted application appli
cation_1454270249092_0055
16/01/31 17:36:58 INFO mapreduce.Job: The url to track the job: htt
p://prabhakar:8088/proxy/application_1454270249092_0055/
16/01/31 17:36:58 INFO mapreduce.Job: Running job: job_1454270249092_00
55
16/01/31 17:37:05 INFO mapreduce.Job: Job job_1454270249092_0055 runnin
g in uber mode : false
16/01/31 17:37:05 INFO mapreduce.Job:  map 0% reduce 0%
16/01/31 17:37:11 INFO mapreduce.Job:  map 100% reduce 0%
16/01/31 17:37:18 INFO mapreduce.Job:  map 100% reduce 100%
16/01/31 17:37:18 INFO mapreduce.Job: Job job_1454270249092_0055 comple
ted successfully
16/01/31 17:37:18 INFO mapreduce.Job: Counters: 51
        File System Counters
                FILE: Number of bytes read=180517
                FILE: Number of bytes written=715438
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=145315
                HDFS: Number of bytes written=373700
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=810
```

```
6
                Total time spent by all reduces in occupied slots (m
s)=3782
                Total time spent by all map tasks (ms)=8106
                Total time spent by all reduce tasks (ms)=3782
                Total vcore-seconds taken by all map tasks=8106
                Total vcore-seconds taken by all reduce tasks=3782
                Total megabyte-seconds taken by all map tasks=8300544
                Total megabyte-seconds taken by all reduce tasks=387276
8
        Map-Reduce Framework
                Map input records=12595
                Map output records=12595
                Map output bytes=155321
                Map output materialized bytes=180523
                Input split bytes=224
                Combine input records=0
                Combine output records=0
                Reduce input groups=12595
                Reduce shuffle bytes=180523
                Reduce input records=12595
                Reduce output records=12594
                Spilled Records=25190
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=104
                CPU time spent (ms)=4750
                Physical memory (bytes) snapshot=692658176
                Virtual memory (bytes) snapshot=2528735232
                Total committed heap usage (bytes)=603979776
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        mapper
                Mapper32d=2
        File Input Format Counters
                Bytes Read=145091
        File Output Format Counters
                Bytes Written=373700
        reducer
                Reducer32d=1
16/01/31 17:37:18 INFO streaming.StreamJob: Output directory: /user/roo
t/wk3/hw33b/output
```

In [150]:

```
!rm output33a.txt
!rm output33b.txt
!hdfs dfs -copyToLocal /user/root/wk3/hw33a/output/part-00000 output33a.txt
!hdfs dfs -copyToLocal /user/root/wk3/hw33b/output/part-00000 output33b.txt
```

```
rm: cannot remove 'output33a.txt': No such file or directory
16/01/31 17:40:14 WARN hdfs.DFSClient: DFSInputStream has been closed already
16/01/31 17:40:17 WARN hdfs.DFSClient: DFSInputStream has been closed already
```

In [151]:
```
!head -51 output33b.txt|tail -50
```

```
DAI62779                6667    1.751%
FRO40251                3881    1.019%
ELE17451                3875    1.018%
GRO73461                3602    0.946%
SNA80324                3044    0.799%
ELE32164                2851    0.749%
DAI75645                2736    0.718%
SNA45677                2455    0.645%
FRO31317                2330    0.612%
DAI85309                2293    0.602%
ELE26917                2292    0.602%
FRO80039                2233    0.586%
GRO21487                2115    0.555%
SNA99873                2083    0.547%
GRO59710                2004    0.526%
GRO71621                1920    0.504%
FRO85978                1918    0.504%
GRO30386                1840    0.483%
ELE74009                1816    0.477%
GRO56726                1784    0.468%
DAI63921                1773    0.466%
GRO46854                1756    0.461%
ELE66600                1713    0.45%
DAI83733                1712    0.45%
FRO32293                1702    0.447%
ELE66810                1697    0.446%
SNA55762                1646    0.432%
DAI22177                1627    0.427%
FRO78087                1531    0.402%
ELE99737                1516    0.398%
ELE34057                1489    0.391%
GRO94758                1489    0.391%
FRO35904                1436    0.377%
FRO53271                1420    0.373%
SNA93860                1407    0.369%
SNA90094                1390    0.365%
GRO38814                1352    0.355%
ELE56788                1345    0.353%
GRO61133                1321    0.347%
DAI88807                1316    0.346%
ELE74482                1316    0.346%
ELE59935                1311    0.344%
SNA96271                1295    0.34%
DAI43223                1290    0.339%
ELE91337                1289    0.338%
GRO15017                1275    0.335%
DAI31081                1261    0.331%
GRO81087                1220    0.32%
DAI22896                1219    0.32%
GRO85051                1214    0.319%
```

In [166]:

```
!echo "Unique Product count: " `head -1 output33b.txt|cut -f 2`
!echo "Largest Basket: " `head -1 output33a.txt|cut -f 2`
```

```
Unique Product count:  12593
Largest Basket:  37
```

HW 3.4. (Computationally prohibitive but then again Hadoop can handle this) Pairs

Suppose we want to recommend new products to the customer based on the products they have already browsed on the online website. Write a map-reduce program to find products which are frequently browsed together. Fix the support count (cooccurence count) to s = 100 (i.e. product pairs need to occur together at least 100 times to be considered frequent) and find pairs of items (sometimes referred to itemsets of size 2 in association rule mining) that have a support count of 100 or more.

List the top 50 product pairs with corresponding support count (aka frequency), and relative frequency or support (number of records where they coccur, the number of records where they coccur/the number of baskets in the dataset) in decreasing order of support for frequent (100>count) itemsets of size 2.

Use the Pairs pattern (lecture 3) to extract these frequent itemsets of size 2. Free free to use combiners if they bring value. Instrument your code with counters for count the number of times your mapper, combiner and reducers are called.

Please output records of the following form for the top 50 pairs (itemsets of size 2):

```
item1, item2, support count, support
```

Fix the ordering of the pairs lexicographically (left to right), and break ties in support (between pairs, if any exist) by taking the first ones in lexicographically increasing order.

Report the compute time for the Pairs job. Describe the computational setup used (E.g., single computer; dual core; linux, number of mappers, number of reducers) Instrument your mapper, combiner, and reducer to count how many times each is called using Counters and report these counts.

In [70]:

```
%%writefile mapper34a.py
#!/usr/bin/python
## mapper34a.py
## Author: Prabhakar Gundugola
## Description: mapper code for HW3.4
import sys
import itertools

sys.stderr.write('reporter:counter:mapper,mapper34a,1\n')

record_count = 0

for line in sys.stdin:
    products = line.strip().split()

    product_pairs = list(itertools.combinations(set(products),2))

    for product_pair in product_pairs:
        pair = sorted(product_pair)
        print pair[0] + ', ' + pair[1] + '\t' + str(1)

    record_count += 1

print '0000RECORDCOUNT' +'\t' + str(record_count)
```

```
Overwriting mapper34a.py
```

In [9]:

```
%%writefile reducer34a.py
#!/usr/bin/python
## reducer34a.py
## Author: Prabhakar Gundugola
## Description: reducer code for HW3.4
import sys

sys.stderr.write('reporter:counter:reducer,reducer34a,1\n')

prev_pair = None
counts = 0
for line in sys.stdin:
    pair, value = line.strip().split('\t')

    if prev_pair != pair:
        if prev_pair is not None:
            print prev_pair + '\t' + str(counts)
        counts = 0
        prev_pair = pair
    counts += eval(value)
print prev_pair + '\t' + str(counts)
```

```
Overwriting reducer34a.py
```

In [10]:

```
!chmod a+x mapper34a.py
!chmod a+x reducer34a.py
```

In [11]:

```
# Ensure the input folder doesn't exist
!hdfs dfs -rm -r /user/root/wk3/hw34a

# Create the input folder
!hdfs dfs -mkdir -p /user/root/wk3/hw34a/input

# Copy the input data file to HDFS input folder
!hdfs dfs -put ProductPurchaseData.txt /user/root/wk3/hw34a/input
```

```
16/01/31 20:00:09 INFO fs.TrashPolicyDefault: Namenode trash configurat
ion: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/root/wk3/hw34a
```

In [75]:

```
# Ensure output folder doesn't exist
!hdfs dfs -rm -r /user/root/wk3/hw34a/output

# Run Hadoop Streaming job
!hadoop jar hadoop-streaming-2.7.1.jar \
-mapper /root/hw3/mapper34a.py \
-reducer /root/hw3/reducer34a.py \
-input /user/root/wk3/hw34a/input \
-output /user/root/wk3/hw34a/output
```

```
16/02/04 01:38:43 INFO fs.TrashPolicyDefault: Namenode trash configurat
ion: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/root/wk3/hw34a/output
packageJobJar: [/tmp/hadoop-unjar5547206564294289197/] [] /tmp/streamjo
b5742543864462449860.jar tmpDir=null
16/02/04 01:38:46 INFO client.RMProxy: Connecting to ResourceManager at
/0.0.0.0:8032
16/02/04 01:38:46 INFO client.RMProxy: Connecting to ResourceManager at
/0.0.0.0:8032
16/02/04 01:38:46 INFO mapred.FileInputFormat: Total input paths to pro
cess : 1
16/02/04 01:38:46 INFO mapreduce.JobSubmitter: number of splits:2
16/02/04 01:38:47 INFO mapreduce.JobSubmitter: Submitting tokens for jo
b: job_1454525374165_0020
16/02/04 01:38:47 INFO impl.YarnClientImpl: Submitted application appli
cation_1454525374165_0020
16/02/04 01:38:47 INFO mapreduce.Job: The url to track the job: htt
p://prabhakar:8088/proxy/application_1454525374165_0020/
16/02/04 01:38:47 INFO mapreduce.Job: Running job: job_1454525374165_00
20
16/02/04 01:38:53 INFO mapreduce.Job: Job job_1454525374165_0020 runnin
g in uber mode : false
16/02/04 01:38:53 INFO mapreduce.Job:  map 0% reduce 0%
16/02/04 01:39:03 INFO mapreduce.Job:  map 50% reduce 0%
16/02/04 01:39:04 INFO mapreduce.Job:  map 100% reduce 0%
16/02/04 01:39:14 INFO mapreduce.Job:  map 100% reduce 72%
16/02/04 01:39:18 INFO mapreduce.Job:  map 100% reduce 77%
16/02/04 01:39:21 INFO mapreduce.Job:  map 100% reduce 83%
16/02/04 01:39:24 INFO mapreduce.Job:  map 100% reduce 88%
16/02/04 01:39:27 INFO mapreduce.Job:  map 100% reduce 93%
16/02/04 01:39:30 INFO mapreduce.Job:  map 100% reduce 100%
16/02/04 01:39:31 INFO mapreduce.Job: Job job_1454525374165_0020 comple
ted successfully
16/02/04 01:39:31 INFO mapreduce.Job: Counters: 51
        File System Counters
                FILE: Number of bytes read=58282376
                FILE: Number of bytes written=116917245
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=3462115
                HDFS: Number of bytes written=18458752
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=177
79
                Total time spent by all reduces in occupied slots (m
s)=24959
```

```
                Total time spent by all map tasks (ms)=17779
                Total time spent by all reduce tasks (ms)=24959
                Total vcore-seconds taken by all map tasks=17779
                Total vcore-seconds taken by all reduce tasks=24959
                Total megabyte-seconds taken by all map tasks=18205696
                Total megabyte-seconds taken by all reduce tasks=255580
16
        Map-Reduce Framework
                Map input records=31101
                Map output records=2534016
                Map output bytes=53214338
                Map output materialized bytes=58282382
                Input split bytes=248
                Combine input records=0
                Combine output records=0
                Reduce input groups=877096
                Reduce shuffle bytes=58282382
                Reduce input records=2534016
                Reduce output records=877096
                Spilled Records=5068032
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=181
                CPU time spent (ms)=38050
                Physical memory (bytes) snapshot=728227840
                Virtual memory (bytes) snapshot=2531340288
                Total committed heap usage (bytes)=591921152
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        mapper
                mapper34a=2
        File Input Format Counters
                Bytes Read=3461867
        File Output Format Counters
                Bytes Written=18458752
        reducer
                reducer34a=1
16/02/04 01:39:31 INFO streaming.StreamJob: Output directory: /user/roo
t/wk3/hw34a/output
```

In [18]:

```
%%writefile reducer34b.py
#!/usr/bin/python
## reducer34b.py
## Author: Prabhakar Gundugola
## Description: reducer code for HW3.4
import sys

sys.stderr.write('reporter:counter:reducer,Reducer34b,1\n')

total = 0
for line in sys.stdin:
    value, pair = line.strip().split('\t')
    # First word should be 0000TOTALWORDS
    if pair == '0000RECORDCOUNT':
        total = int(value)
    else:
        term_freq = round(100.0 * int(value)/total, 3)
        print pair.ljust(20) + '\t' + value + '\t' + str(term_freq) +
'%'
```

```
Writing reducer34b.py
```

In [19]:

```
!chmod a+x reducer34b.py
```

In [20]:

```
# Ensure the output folder doesn't exist
!hdfs dfs -rm -r /user/root/wk3/hw34b/output

# Run Hadoop Streaming job.
!hadoop jar hadoop-streaming-2.7.1.jar \
-D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator \
-D mapred.text.key.partitioner.options=-k1,1 \
-D stream.num.map.output.key.fields=2 \
-D mapred.text.key.comparator.options='-k1,1nr -k2,2n' \
-mapper /root/hw3/mapper32d.py \
-reducer /root/hw3/reducer34b.py \
-input /user/root/wk3/hw34a/output/part* \
-output /user/root/wk3/hw34b/output
```

```
rm: `/user/root/wk3/hw34b/output': No such file or directory
packageJobJar: [/tmp/hadoop-unjar3536631726763484975/] [] /tmp/streamjo
b8537272369429628869.jar tmpDir=null
16/01/31 20:15:27 INFO client.RMProxy: Connecting to ResourceManager at
/0.0.0.0:8032
16/01/31 20:15:27 INFO client.RMProxy: Connecting to ResourceManager at
/0.0.0.0:8032
16/01/31 20:15:27 INFO mapred.FileInputFormat: Total input paths to pro
cess : 1
16/01/31 20:15:27 INFO mapreduce.JobSubmitter: number of splits:2
16/01/31 20:15:27 INFO Configuration.deprecation: mapred.output.key.com
parator.class is deprecated. Instead, use mapreduce.job.output.key.comp
arator.class
16/01/31 20:15:27 INFO Configuration.deprecation: mapred.text.key.compa
rator.options is deprecated. Instead, use mapreduce.partition.keycompar
ator.options
16/01/31 20:15:27 INFO Configuration.deprecation: mapred.text.key.parti
tioner.options is deprecated. Instead, use mapreduce.partition.keyparti
tioner.options
16/01/31 20:15:28 INFO mapreduce.JobSubmitter: Submitting tokens for jo
b: job_1454270249092_0061
16/01/31 20:15:28 INFO impl.YarnClientImpl: Submitted application appli
cation_1454270249092_0061
16/01/31 20:15:28 INFO mapreduce.Job: The url to track the job: htt
p://prabhakar:8088/proxy/application_1454270249092_0061/
16/01/31 20:15:28 INFO mapreduce.Job: Running job: job_1454270249092_00
61
16/01/31 20:15:35 INFO mapreduce.Job: Job job_1454270249092_0061 runnin
g in uber mode : false
16/01/31 20:15:35 INFO mapreduce.Job:  map 0% reduce 0%
16/01/31 20:15:42 INFO mapreduce.Job:  map 100% reduce 0%
16/01/31 20:15:53 INFO mapreduce.Job:  map 100% reduce 100%
16/01/31 20:15:53 INFO mapreduce.Job: Job job_1454270249092_0061 comple
ted successfully
16/01/31 20:15:53 INFO mapreduce.Job: Counters: 51
        File System Counters
                FILE: Number of bytes read=20212951
                FILE: Number of bytes written=40780306
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=17585165
                HDFS: Number of bytes written=26295178
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=112
43
                Total time spent by all reduces in occupied slots (m
s)=8731
                Total time spent by all map tasks (ms)=11243
                Total time spent by all reduce tasks (ms)=8731
                Total vcore-seconds taken by all map tasks=11243
                Total vcore-seconds taken by all reduce tasks=8731
                Total megabyte-seconds taken by all map tasks=11512832
                Total megabyte-seconds taken by all reduce tasks=894054
4
        Map-Reduce Framework
                Map input records=877096
                Map output records=877096
                Map output bytes=18458753
                Map output materialized bytes=20212957
                Input split bytes=224
                Combine input records=0
                Combine output records=0
                Reduce input groups=877096
                Reduce shuffle bytes=20212957
                Reduce input records=877096
                Reduce output records=877095
                Spilled Records=1754192
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=260
                CPU time spent (ms)=12900
                Physical memory (bytes) snapshot=723660800
                Virtual memory (bytes) snapshot=2530881536
                Total committed heap usage (bytes)=568852480
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        mapper
                Mapper32d=2
        File Input Format Counters
                Bytes Read=17584941
        File Output Format Counters
                Bytes Written=26295178
        reducer
                Reducer34b=1
16/01/31 20:15:53 INFO streaming.StreamJob: Output directory: /user/roo
t/wk3/hw34b/output
```

**

In [21]:

```
!rm output34b.txt
!hdfs dfs -copyToLocal /user/root/wk3/hw34b/output/part-00000 output34b.txt
```

```
rm: cannot remove 'output34b.txt': No such file or directory
16/01/31 20:17:04 WARN hdfs.DFSClient: DFSInputStream has been closed already
```

In [23]:
```
!head -50 output34b.txt
```

```
DAI62779,ELE17451       1592    5.119%
FRO40251,SNA80324       1412    4.54%
DAI75645,FRO40251       1254    4.032%
FRO40251,GRO85051       1213    3.9%
DAI62779,GRO73461       1139    3.662%
DAI75645,SNA80324       1130    3.633%
DAI62779,FRO40251       1070    3.44%
DAI62779,SNA80324       923     2.968%
DAI62779,DAI85309       918     2.952%
ELE32164,GRO59710       911     2.929%
DAI62779,DAI75645       882     2.836%
FRO40251,GRO73461       882     2.836%
DAI62779,ELE92920       877     2.82%
FRO40251,FRO92469       835     2.685%
DAI62779,ELE32164       832     2.675%
DAI75645,GRO73461       712     2.289%
DAI43223,ELE32164       711     2.286%
DAI62779,GRO30386       709     2.28%
ELE17451,FRO40251       697     2.241%
DAI85309,ELE99737       659     2.119%
DAI62779,ELE26917       650     2.09%
GRO21487,GRO73461       631     2.029%
DAI62779,SNA45677       604     1.942%
ELE17451,SNA80324       597     1.92%
DAI62779,GRO71621       595     1.913%
DAI62779,SNA55762       593     1.907%
DAI62779,DAI83733       586     1.884%
ELE17451,GRO73461       580     1.865%
GRO73461,SNA80324       562     1.807%
DAI62779,GRO59710       561     1.804%
DAI62779,FRO80039       550     1.768%
DAI75645,ELE17451       547     1.759%
DAI62779,SNA93860       537     1.727%
DAI55148,DAI62779       526     1.691%
DAI43223,GRO59710       512     1.646%
ELE17451,ELE32164       511     1.643%
DAI62779,SNA18336       506     1.627%
ELE32164,GRO73461       486     1.563%
DAI85309,ELE17451       482     1.55%
DAI62779,FRO78087       482     1.55%
DAI62779,GRO94758       479     1.54%
DAI62779,GRO21487       471     1.514%
GRO85051,SNA80324       471     1.514%
ELE17451,GRO30386       468     1.505%
FRO85978,SNA95666       463     1.489%
DAI62779,FRO19221       462     1.485%
DAI62779,GRO46854       461     1.482%
DAI43223,DAI62779       459     1.476%
ELE92920,SNA18336       455     1.463%
DAI88079,FRO40251       446     1.434%
```

Report the compute time - Pairs job

1st Map Reduce program:
All map tasks: 17779 ms
All reduce tasks: 24959 ms

2nd Map Reduce program:
All map tasks: 11243 ms
All reduce tasks: 8731 ms

Computational Setup

SoftLayer VM, 4 Core, 32 GB RAM, 2 mappers, 1 reducer

HW3.5. Stripes

Repeat 3.4 using the stripes design pattern for finding cooccuring pairs.

Report the compute times for stripes job versus the Pairs job. Describe the computational setup used (E.g., single computer; dual core; linux, number of mappers, number of reducers)

Instrument your mapper, combiner, and reducer to count how many times each is called using Counters and report these counts. Discuss the differences in these counts between the Pairs and Stripes jobs

In [18]:

```
%%writefile mapper35a.py
#!/usr/bin/python
#HW 3.5

import sys
sys.stderr.write("reporter:counter:Calls,mapper_calls,1\n")
linecount = 0
# input comes from STDIN (standard input)
for line in sys.stdin:
    line = line.strip()
    products = line.split(" ")
    products = sorted(products)
    linecount += 1
    # emit the product
    for item in products:
        for item2 in products[products.index(item)+1:]:
            print "%s,%s\t1" % (item, item2)

print "linecount\t"+str(linecount)
```

```
Overwriting mapper35a.py
```

In [48]:

```
%%writefile reducer35a.py
#!/usr/bin/python
#HW 3.5

import sys

sys.stderr.write("reporter:counter:Calls,reducer_calls,1\n")
stripes = {}
current_key = None
current_count = 0
key = None
linecount = 0

# input comes from STDIN (standard input)
for line in sys.stdin:
    line = line.strip()
    key, count = line.split("\t", 1)
    count = int(count)

    if current_key == key:
        current_count += int(count)
    else:
        if current_key:
            items = current_key.split(",", 1)
            if len(items) == 2:
                stripes.setdefault(items[0], {})
                stripes[items[0]][items[1]]=current_count
            elif items[0] == "linecount":
                linecount = current_count
        current_count = count
        current_key = key

# output the last word
if current_key == key:
    items = current_key.split(",", 1)
    if len(items) == 2:
        stripes.setdefault(items[0], {})
        stripes[items[0]][items[1]]=current_count
    elif items[0] == "linecount":
        linecount = current_count


for key, stripe in stripes.items():
    marg_count = sum(stripe.values())
    for key2, count in stripe.items():
        if count >= 100:
            line_freq = round(100.0*count/marg_count, 4)
            #print "%s\t%s\t%s\t%.4f\t%.4f" % \
            #(key, key2, str(count), count*1.0/linecount, count*1.0/mar
g_count)
            print key + ', ' + key2 + '\t' + str(count) +'\t' + str(lin
e_freq) + '%'
```

```
Overwriting reducer35a.py
```

In [40]:

```
!chmod a+x mapper35a.py
!chmod a+x reducer35a.py
```

In [41]:

```
# Ensure the input folder doesn't exist
!hdfs dfs -rm -r /user/root/wk3/hw35a

# Create the input folder
!hdfs dfs -mkdir -p /user/root/wk3/hw35a/input

# Copy the input data file to HDFS input folder
!hdfs dfs -put ProductPurchaseData.txt /user/root/wk3/hw35a/input
```

```
16/02/04 00:42:02 INFO fs.TrashPolicyDefault: Namenode trash configurat
ion: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/root/wk3/hw35a
```

In [49]:

```
!hdfs dfs -rm -r /user/root/wk3/hw35a/output

!hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7.1.jar \
-D stream.map.output.field.separator="\t" \
-mapper /root/hw3/mapper35a.py \
-reducer /root/hw3/reducer35a.py \
-input /user/root/wk3/hw35a/input/ProductPurchaseData.txt \
-output /user/root/wk3/hw35a/output
```

```
16/02/04 00:45:27 INFO fs.TrashPolicyDefault: Namenode trash configurat
ion: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/root/wk3/hw35a/output
packageJobJar: [/tmp/hadoop-unjar4941323668525768229/] [] /tmp/streamjo
b947624389084947282.jar tmpDir=null
16/02/04 00:45:30 INFO client.RMProxy: Connecting to ResourceManager at
/0.0.0.0:8032
16/02/04 00:45:30 INFO client.RMProxy: Connecting to ResourceManager at
/0.0.0.0:8032
16/02/04 00:45:31 INFO mapred.FileInputFormat: Total input paths to pro
cess : 1
16/02/04 00:45:31 INFO mapreduce.JobSubmitter: number of splits:2
16/02/04 00:45:31 INFO mapreduce.JobSubmitter: Submitting tokens for jo
b: job_1454525374165_0015
16/02/04 00:45:31 INFO impl.YarnClientImpl: Submitted application appli
cation_1454525374165_0015
16/02/04 00:45:31 INFO mapreduce.Job: The url to track the job: htt
p://prabhakar:8088/proxy/application_1454525374165_0015/
16/02/04 00:45:31 INFO mapreduce.Job: Running job: job_1454525374165_00
15
16/02/04 00:45:37 INFO mapreduce.Job: Job job_1454525374165_0015 runnin
g in uber mode : false
16/02/04 00:45:37 INFO mapreduce.Job:  map 0% reduce 0%
16/02/04 00:45:48 INFO mapreduce.Job:  map 100% reduce 0%
16/02/04 00:46:00 INFO mapreduce.Job:  map 100% reduce 88%
16/02/04 00:46:02 INFO mapreduce.Job:  map 100% reduce 100%
16/02/04 00:46:03 INFO mapreduce.Job: Job job_1454525374165_0015 comple
ted successfully
16/02/04 00:46:03 INFO mapreduce.Job: Counters: 51
        File System Counters
                FILE: Number of bytes read=58283424
                FILE: Number of bytes written=116919839
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=3462115
                HDFS: Number of bytes written=41230
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=169
82
                Total time spent by all reduces in occupied slots (m
s)=11127
                Total time spent by all map tasks (ms)=16982
                Total time spent by all reduce tasks (ms)=11127
                Total vcore-seconds taken by all map tasks=16982
                Total vcore-seconds taken by all reduce tasks=11127
                Total megabyte-seconds taken by all map tasks=17389568
                Total megabyte-seconds taken by all reduce tasks=113940
48
        Map-Reduce Framework
                Map input records=31101
                Map output records=2534062
                Map output bytes=53215294
                Map output materialized bytes=58283430
                Input split bytes=248
                Combine input records=0
                Combine output records=0
                Reduce input groups=877100
                Reduce shuffle bytes=58283430
                Reduce input records=2534062
                Reduce output records=1334
                Spilled Records=5068124
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=261
                CPU time spent (ms)=20000
                Physical memory (bytes) snapshot=713805824
                Virtual memory (bytes) snapshot=2528686080
                Total committed heap usage (bytes)=590348288
        Calls
                mapper_calls=2
                reducer_calls=1
        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=3461867
        File Output Format Counters
                Bytes Written=41230
16/02/04 00:46:03 INFO streaming.StreamJob: Output directory: /user/roo
t/wk3/hw35a/output
```

In [50]:

```
!hdfs dfs -cat /user/root/wk3/hw35a/output/part-00000|head -20
```

```
ELE20847, ELE26917      110     1.1777%
ELE20847, GRO73461      187     2.0021%
ELE20847, FRO92469      122     1.3062%
ELE20847, GRO85051      139     1.4882%
ELE20847, SNA80324      410     4.3897%
ELE20847, FRO75586      118     1.2634%
ELE20847, SNA96271      184     1.97%
ELE20847, FRO40251      434     4.6467%
DAI22896, GRO21487      114     0.6891%
DAI22896, GRO38814      223     1.3479%
DAI22896, ELE74009      165     0.9973%
DAI22896, DAI62779      297     1.7952%
DAI22896, GRO73461      304     1.8375%
DAI22896, DAI75645      215     1.2996%
DAI22896, GRO30386      102     0.6165%
DAI22896, SNA80324      195     1.1787%
DAI22896, ELE32164      107     0.6468%
DAI22896, GRO46854      114     0.6891%
DAI22896, FRO53271      123     0.7435%
DAI22896, SNA72163      227     1.3721%
cat: Unable to write to output stream.
```

In [64]:

```
%%writefile mapper35b.py
#!/usr/bin/python
#HW 3.5

import sys

# input comes from STDIN (standard input)
for line in sys.stdin:
    line = line.strip()
    print line
```

```
Overwriting mapper35b.py
```

In [57]:

```
!chmod a+x mapper35b.py
!chmod a+x reducer35b.py
```

In [59]:

```
# Ensure the input folder doesn't exist
!hdfs dfs -rm -r /user/root/wk3/hw35b
# Create the input folder
!hdfs dfs -mkdir -p /user/root/wk3/hw35b/input

# Copy the input data file to HDFS input folder
!hdfs dfs -put ProductPurchaseData.txt /user/root/wk3/hw35b/input
```

```
rm: `/user/root/wk3/hw35b': No such file or directory
```

In [67]:

```
!hdfs dfs -rm -r /user/root/wk3/hw35b/output

!hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-
2.7.1.jar \
-D stream.map.output.field.separator="\t" \
-D mapreduce.job.output.key.comparator.class=\
org.apache.hadoop.mapred.lib.KeyFieldBasedComparator \
-D mapreduce.partition.keycomparator.options="-k2,2nr -k1,1 -k2,2" \
-mapper /root/hw3/mapper35b.py \
-reducer /root/hw3/mapper35b.py \
-input /user/root/wk3/hw35a/output/part-00000 \
-output /user/root/wk3/hw35b/output
```

```
16/02/04 01:15:27 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/root/wk3/hw35b/output
packageJobJar: [/tmp/hadoop-unjar6372001845285142446/] [] /tmp/streamjob7974217546918567144.jar tmpDir=null
16/02/04 01:15:30 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
16/02/04 01:15:30 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
16/02/04 01:15:31 INFO mapred.FileInputFormat: Total input paths to process : 1
16/02/04 01:15:31 INFO mapreduce.JobSubmitter: number of splits:2
16/02/04 01:15:31 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1454525374165_0019
16/02/04 01:15:31 INFO impl.YarnClientImpl: Submitted application application_1454525374165_0019
16/02/04 01:15:31 INFO mapreduce.Job: The url to track the job: http://prabhakar:8088/proxy/application_1454525374165_0019/
16/02/04 01:15:31 INFO mapreduce.Job: Running job: job_1454525374165_0019
16/02/04 01:15:37 INFO mapreduce.Job: Job job_1454525374165_0019 running in uber mode : false
16/02/04 01:15:37 INFO mapreduce.Job:  map 0% reduce 0%
16/02/04 01:15:43 INFO mapreduce.Job:  map 100% reduce 0%
16/02/04 01:15:49 INFO mapreduce.Job:  map 100% reduce 100%
16/02/04 01:15:50 INFO mapreduce.Job: Job job_1454525374165_0019 completed successfully
16/02/04 01:15:51 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=45238
                FILE: Number of bytes written=444550
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=45415
                HDFS: Number of bytes written=41230
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=7822
                Total time spent by all reduces in occupied slots (ms)=3483
                Total time spent by all map tasks (ms)=7822
                Total time spent by all reduce tasks (ms)=3483
                Total vcore-seconds taken by all map tasks=7822
                Total vcore-seconds taken by all reduce tasks=3483
                Total megabyte-seconds taken by all map tasks=8009728
                Total megabyte-seconds taken by all reduce tasks=356659
```

```
2
        Map-Reduce Framework
                Map input records=1334
                Map output records=1334
                Map output bytes=42564
                Map output materialized bytes=45244
                Input split bytes=224
                Combine input records=0
                Combine output records=0
                Reduce input groups=1334
                Reduce shuffle bytes=45244
                Reduce input records=1334
                Reduce output records=1334
                Spilled Records=2668
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=98
                CPU time spent (ms)=2830
                Physical memory (bytes) snapshot=706846720
                Virtual memory (bytes) snapshot=2527629312
                Total committed heap usage (bytes)=603979776
        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=45191
        File Output Format Counters
                Bytes Written=41230
16/02/04 01:15:51 INFO streaming.StreamJob: Output directory: /user/roo
t/wk3/hw35b/output
```

In [69]:

```
!hdfs dfs -cat /user/root/wk3/hw35b/output/part-00000 |head -50
```

```
DAI62779, ELE17451      1592    2.045%
FRO40251, SNA80324      1412    4.888%
DAI75645, FRO40251      1254    3.6456%
FRO40251, GRO85051      1213    4.1991%
DAI62779, GRO73461      1139    1.4631%
DAI75645, SNA80324      1130    3.2851%
DAI62779, FRO40251      1070    1.3744%
DAI62779, SNA80324      923     1.1856%
DAI62779, DAI85309      918     1.1792%
ELE32164, GRO59710      911     3.4435%
DAI62779, DAI75645      882     1.1329%
FRO40251, GRO73461      882     3.0533%
DAI62779, ELE92920      877     1.1265%
FRO40251, FRO92469      835     2.8906%
DAI62779, ELE32164      832     1.0687%
DAI75645, GRO73461      712     2.0699%
DAI43223, ELE32164      711     4.2296%
DAI62779, GRO30386      709     0.9107%
ELE17451, FRO40251      697     1.8024%
DAI85309, ELE99737      659     2.4564%
DAI62779, ELE26917      650     0.8349%
GRO21487, GRO73461      631     5.6908%
DAI62779, SNA45677      604     0.7759%
ELE17451, SNA80324      597     1.5438%
DAI62779, GRO71621      595     0.7643%
DAI62779, SNA55762      593     0.7617%
DAI62779, DAI83733      586     0.7527%
ELE17451, GRO73461      580     1.4998%
GRO73461, SNA80324      562     4.7014%
DAI62779, GRO59710      561     0.7206%
DAI62779, FRO80039      550     0.7065%
DAI75645, ELE17451      547     1.5902%
DAI62779, SNA93860      537     0.6898%
DAI55148, DAI62779      526     4.5166%
DAI43223, GRO59710      512     3.0458%
ELE17451, ELE32164      511     1.3214%
DAI62779, SNA18336      506     0.65%
ELE32164, GRO73461      486     1.837%
DAI62779, FRO78087      482     0.6191%
DAI85309, ELE17451      482     1.7966%
DAI62779, GRO94758      479     0.6153%
DAI62779, GRO21487      471     0.605%
GRO85051, SNA80324      471     12.9645%
ELE17451, GRO30386      468     1.2102%
FRO85978, SNA95666      463     4.1299%
DAI62779, FRO19221      462     0.5934%
DAI62779, GRO46854      461     0.5922%
DAI43223, DAI62779      459     2.7305%
ELE92920, SNA18336      455     4.8194%
DAI88079, FRO40251      446     8.3898%
```

Report the compute time - Stripes job

1st Map Reduce program:

All map tasks: 16982 ms
All reduce tasks: 11127 ms

2nd Map Reduce program:

All map tasks: 7822 ms
All reduce tasks: 3483 ms

Computational Setup

SoftLayer VM, 4 Core, 32 GB RAM, 2 mappers, 1 reducer

```
In [ ]:
```