DATASCI W261, Machine Learning at Scale

Assignement: week #3

Shih Yu Chang

Due: 2016-09-20, 8AM PST

HW3.0.

1. How do you merge two sorted lists/arrays of records of the form [key, value]?
2. Where is this used in Hadoop MapReduce? [Hint within the shuffle]
3. What is a combiner function in the context of Hadoop?
4. Give an example where it can be used and justify why it should be used in the context of this problem.
5. What is the Hadoop shuffle?

HW3.0. Q&A

1.2 Merge sort is a sorting method which combines two sorted lists into a single sorted list of items. Merge sort benefits from distributed computing environment by sorting of the child lists. The merging of child lists into a single sorted list can be finished in linear time. Merge sorting is used in the shuffle stage of Hadoop to rearrange keys prior to sending them to the reducer. Key-value pairs from different mappers are sorted at their mappers, and then distributed across the reducers in a sorted form.

3. Combiners are used for local aggregation during the mapper processes of Hadoop. They are run when the incomplete output from the mapper becomes too large to fit within memory. The combiner is responsible for reducing the data size so that the mapper can run faster by keeping data in memory and so that the network operations overhead in the partitioner is kept as small as possible. Depending on the size and scope of the problem, Hadoop will run combiners any number of times including zero with no input from the user. For this reason, it is important that the combiner is able to receive records in the same format of the mapper's output and emit data in the same format. The combining operation must also be associative and commutative so that the variable number of runs will not affect the final result.

4. Combiners can be used in large word-count operations. A typical mapper output for a word-count problem will, in general, be much greater than the size of the document since it emits each individual word and the counter associated with it. Transferring this data across the network will downgrade performance of this operation, as well as making the subsequent sorting operation take much longer. Adding a combiner can reduce the size of the mapper output.

5. Shuffle happens after all mapper tasks complete, but before reducer tasks start. All key-value pairs are sorted by key, and the same key is guaranteed to be delivered to the same reducer.

HW3.1 consumer complaints dataset: Use Counters to do EDA (exploratory data analysis and to monitor progress)

In [2]:
```
!hdfs dfs -mkdir -p /user/shihyu
!hdfs dfs -put Consumer_Complaints.csv /user/shihyu
```

```
put: `/user/shihyu/Consumer_Complaints.csv': File exists
```

In [9]:
```
%%writefile mappe3r1.py
#!/usr/bin/python

import sys
for line in sys.stdin:
    line=line.strip()
    #Since product at second field, extract from index 1
    product=line.split(',')[1]
    # emit product name as key, no need for value
    print "%s\t%s" %(product, 'na')
```

```
Overwriting mappe3r1.py
```

In [10]:
```
%%writefile reduce3r1.py
#!/usr/bin/python

import sys
for line in sys.stdin:
    product = line.split('\t')[0].strip()
    try:
        #Iterate the counter depending on the product
        if product.lower()=='debt collection':
            sys.stderr.write("reporter:counter:Debt,Total,1\n")
        if product.lower()=='mortgage':
            sys.stderr.write("reporter:counter:Mortgage,Total,1\n")
        else:
            sys.stderr.write("reporter:counter:Others,Total,1\n")
    except:
        # must be a header record so skip it
        pass
```

```
Overwriting reduce3r1.py
```

In [11]:
```
!chmod a+x mappe3r1.py
!chmod a+x reduce3r1.py
!hdfs dfs -rm -r result3s1
!hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
-mapper /home/cloudera/mappe3r1.py \
-reducer /home/cloudera/reduce3r1.py \
-input /user/shihyu/Consumer_Complaints.csv \
-output result3s1
```

```
Deleted result3s1
packageJobJar: [] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.6.0-cdh5.8.0.jar
] /tmp/streamjob5972900019385505293.jar tmpDir=null
16/09/12 20:01:27 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/12 20:01:28 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/12 20:01:28 INFO mapred.FileInputFormat: Total input paths to process : 1
16/09/12 20:01:28 INFO mapreduce.JobSubmitter: number of splits:2
16/09/12 20:01:29 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_14
73444507507_0065
16/09/12 20:01:29 INFO impl.YarnClientImpl: Submitted application application_14
73444507507_0065
16/09/12 20:01:29 INFO mapreduce.Job: The url to track the job: http://quickstar
t.cloudera:8088/proxy/application_1473444507507_0065/
16/09/12 20:01:29 INFO mapreduce.Job: Running job: job_1473444507507_0065
16/09/12 20:01:37 INFO mapreduce.Job: Job job_1473444507507_0065 running in uber
 mode : false
16/09/12 20:01:37 INFO mapreduce.Job:  map 0% reduce 0%
16/09/12 20:01:48 INFO mapreduce.Job:  map 50% reduce 0%
16/09/12 20:01:49 INFO mapreduce.Job:  map 100% reduce 0%
16/09/12 20:01:57 INFO mapreduce.Job:  map 100% reduce 100%
16/09/12 20:01:57 INFO mapreduce.Job: Job job_1473444507507_0065 completed succe
ssfully
16/09/12 20:01:58 INFO mapreduce.Job: Counters: 52
        File System Counters
                FILE: Number of bytes read=5817067
                FILE: Number of bytes written=11989780
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=50910820
                HDFS: Number of bytes written=0
                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)=18726
                Total time spent by all reduces in occupied slots (ms)=6190
                Total time spent by all map tasks (ms)=18726
                Total time spent by all reduce tasks (ms)=6190
                Total vcore-seconds taken by all map tasks=18726
                Total vcore-seconds taken by all reduce tasks=6190
                Total megabyte-seconds taken by all map tasks=19175424
                Total megabyte-seconds taken by all reduce tasks=6338560
        Map-Reduce Framework
                Map input records=312913
                Map output records=312913
                Map output bytes=5191235
                Map output materialized bytes=5817073
                Input split bytes=238
                Combine input records=0
                Combine output records=0
                Reduce input groups=10
                Reduce shuffle bytes=5817073
                Reduce input records=312913
                Reduce output records=0
                Spilled Records=625826
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
```

Above results show that there are 44372 debt records, 125752 mortgage records, and 187161 other records.

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

single record dataset: foo foo quux labs foo bar quux

In [13]:
```
#Create a test file that we can use to test our code
! echo "foo foo quux labs foo bar quux" > testfil3e2a.txt
!hdfs dfs -mkdir -p /user/shihyu
!hdfs dfs -put testfil3e2a.txt /user/shihyu
```

```
put: `/user/shihyu/testfil3e2a.txt': File exists
```

In [22]:
```
%%writefile mappe3r2a.py
#!/usr/bin/python

import sys
#counter
count = 0
#Increment script call counter once when the file runs
sys.stderr.write("reporter:counter:Mapper2a,Script Calls,1\n")
for line in sys.stdin:
    #Increment line call counter when we process a new line
    sys.stderr.write("reporter:counter:Mapper2a,Line Calls,1\n")
    line=line.strip()
    words=line.split()
    for word in words:
        print '%s\t%s' % (word, 1)
```

```
Overwriting mappe3r2a.py
```

In [23]:
```
%%writefile reduce3r2a.py
#!/usr/bin/python

import sys
tmp_word=''
#Counter for the chosen word
count = 0

#Increment script call counter once when the file runs
sys.stderr.write("reporter:counter:Reducer2a,Script Calls,1\n")
for line in sys.stdin:
    #Parse line
    line=line.strip().split('\t')
    word,tmp_count=line
    if tmp_word==word:
        count+=int(tmp_count)
    else:
        if tmp_word:

            #Increment line call counter when we emit a new word
            sys.stderr.write("reporter:counter:Reducer2a,Line Calls,1\n")
            print tmp_word+'\t'+str(count)
        tmp_word=word
        count=int(tmp_count)

#Do not forget  to emit final record
if tmp_word:
    sys.stderr.write("reporter:counter:Reducer2a,Line Calls,1\n")
    print tmp_word+'\t'+str(count)
```

```
Overwriting reduce3r2a.py
```

In [25]:
```
!chmod a+x mappe3r2a.py
!chmod a+x reduce3r2a.py
!hdfs dfs -rm -r result3s2a
!hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
-D mapred.map.tasks=1 \
-D mapred.reduce.tasks=4 \
-mapper /home/cloudera/mappe3r2a.py \
-reducer /home/cloudera/reduce3r2a.py \
-input /user/shihyu/testfil3e2a.txt \
-output result3s2a
```

```
Deleted result3s2a
packageJobJar: [] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.6.0-cdh5.8.0.jar
] /tmp/streamjob6115026488106076061.jar tmpDir=null
16/09/12 22:07:41 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/12 22:07:42 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/12 22:07:42 INFO mapred.FileInputFormat: Total input paths to process : 1
16/09/12 22:07:42 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1281)
        at java.lang.Thread.join(Thread.java:1355)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DF
SOutputStream.java:862)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutpu
tStream.java:600)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStre
am.java:789)
16/09/12 22:07:42 INFO mapreduce.JobSubmitter: number of splits:1
16/09/12 22:07:42 INFO Configuration.deprecation: mapred.reduce.tasks is depreca
ted. Instead, use mapreduce.job.reduces
16/09/12 22:07:42 INFO Configuration.deprecation: mapred.map.tasks is deprecated
. Instead, use mapreduce.job.maps
16/09/12 22:07:43 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_14
73444507507_0069
16/09/12 22:07:43 INFO impl.YarnClientImpl: Submitted application application_14
73444507507_0069
16/09/12 22:07:43 INFO mapreduce.Job: The url to track the job: http://quickstar
t.cloudera:8088/proxy/application_1473444507507_0069/
16/09/12 22:07:43 INFO mapreduce.Job: Running job: job_1473444507507_0069
16/09/12 22:07:51 INFO mapreduce.Job: Job job_1473444507507_0069 running in uber
 mode : false
16/09/12 22:07:51 INFO mapreduce.Job:  map 0% reduce 0%
16/09/12 22:07:57 INFO mapreduce.Job:  map 100% reduce 0%
16/09/12 22:08:11 INFO mapreduce.Job:  map 100% reduce 25%
16/09/12 22:08:13 INFO mapreduce.Job:  map 100% reduce 50%
16/09/12 22:08:15 INFO mapreduce.Job:  map 100% reduce 100%
16/09/12 22:08:15 INFO mapreduce.Job: Job job_1473444507507_0069 completed succe
ssfully
16/09/12 22:08:15 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=83
                FILE: Number of bytes written=592813
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=142
                HDFS: Number of bytes written=26
                HDFS: Number of read operations=15
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=8
        Job Counters
                Killed reduce tasks=1
                Launched map tasks=1
                Launched reduce tasks=4
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=3852
                Total time spent by all reduces in occupied slots (ms)=47084
                Total time spent by all map tasks (ms)=3852
                Total time spent by all reduce tasks (ms)=47084
                Total vcore-seconds taken by all map tasks=3852
                Total vcore-seconds taken by all reduce tasks=47084
                Total megabyte-seconds taken by all map tasks=3944448
```

In [18]:
```
!hdfs dfs -cat result3s2a/part-00000 | head -5
```

```
bar     1
foo     3
labs    1
quux    2
```

Since there are four different words, the value of mapper counter is one and the value of reducer counter is four.

mulitple mappers and reducers

In [26]:
```
%%writefile mappe3r2b.py
#!/usr/bin/python

import sys
import re
from csv import reader
WORD_RE = re.compile(r"[\w']+")

count = 0
sys.stderr.write("reporter:counter:Mapper2b,Script Count,1\n")
for line in reader(sys.stdin):
    sys.stderr.write("reporter:counter:Mapper2b,Line Count,1\n")
    # Considering words in compaints issue
    words = re.findall(WORD_RE, line[3])
    for word in words:
        print '%s\t%s' % (word, 1)
```

```
Writing mappe3r2b.py
```

In [27]:
```
%%writefile reduce3r2b.py
#!/usr/bin/python

import sys
tmp_word=''
#Counter for the chosen word
count = 0

#Increment script call counter once when the file runs
sys.stderr.write("reporter:counter:Reducer2a,Script Calls,1\n")
for line in sys.stdin:
    #Parse line
    line=line.strip().split('\t')
    word,tmp_count=line
    if tmp_word==word:
        count+=int(tmp_count)
    else:
        if tmp_word:

            #Increment line call counter when we emit a new word
            sys.stderr.write("reporter:counter:Reducer2a,Line Calls,1\n")
            print tmp_word+'\t'+str(count)
        tmp_word=word
        count=int(tmp_count)

#Do not forget  to emit final record
if tmp_word:
    sys.stderr.write("reporter:counter:Reducer2a,Line Calls,1\n")
    print tmp_word+'\t'+str(count)
```

```
Writing reduce3r2b.py
```

In [28]:
```
!chmod a+x mappe3r2b.py
!chmod a+x reduce3r2b.py
!hdfs dfs -rm -r result3s2b
!hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
-D mapred.map.tasks=2 \
-D mapred.reduce.tasks=2 \
-mapper /home/cloudera/mappe3r2b.py \
-reducer /home/cloudera/reduce3r2b.py \
-input /user/shihyu/Consumer_Complaints.csv \
-output result3s2b
```

```
rm: `result3s2b': No such file or directory
packageJobJar: [] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.6.0-cdh5.8.0.jar
] /tmp/streamjob8882929842593617907.jar tmpDir=null
16/09/12 22:28:54 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/12 22:28:54 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/12 22:28:55 INFO mapred.FileInputFormat: Total input paths to process : 1
16/09/12 22:28:55 INFO mapreduce.JobSubmitter: number of splits:2
16/09/12 22:28:55 INFO Configuration.deprecation: mapred.reduce.tasks is depreca
ted. Instead, use mapreduce.job.reduces
16/09/12 22:28:55 INFO Configuration.deprecation: mapred.map.tasks is deprecated
. Instead, use mapreduce.job.maps
16/09/12 22:28:55 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_14
73444507507_0070
16/09/12 22:28:56 INFO impl.YarnClientImpl: Submitted application application_14
73444507507_0070
16/09/12 22:28:56 INFO mapreduce.Job: The url to track the job: http://quickstar
t.cloudera:8088/proxy/application_1473444507507_0070/
16/09/12 22:28:56 INFO mapreduce.Job: Running job: job_1473444507507_0070
16/09/12 22:29:05 INFO mapreduce.Job: Job job_1473444507507_0070 running in uber
 mode : false
16/09/12 22:29:05 INFO mapreduce.Job:  map 0% reduce 0%
16/09/12 22:29:20 INFO mapreduce.Job:  map 61% reduce 0%
16/09/12 22:29:22 INFO mapreduce.Job:  map 79% reduce 0%
16/09/12 22:29:23 INFO mapreduce.Job:  map 100% reduce 0%
16/09/12 22:29:34 INFO mapreduce.Job:  map 100% reduce 50%
16/09/12 22:29:35 INFO mapreduce.Job:  map 100% reduce 100%
16/09/12 22:29:35 INFO mapreduce.Job: Job job_1473444507507_0070 completed succe
ssfully
16/09/12 22:29:35 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=16121345
                FILE: Number of bytes written=32716908
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=50910820
                HDFS: Number of bytes written=2342
                HDFS: Number of read operations=12
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Job Counters
                Killed reduce tasks=1
                Launched map tasks=2
                Launched reduce tasks=2
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=28489
                Total time spent by all reduces in occupied slots (ms)=19523
                Total time spent by all map tasks (ms)=28489
                Total time spent by all reduce tasks (ms)=19523
                Total vcore-seconds taken by all map tasks=28489
                Total vcore-seconds taken by all reduce tasks=19523
                Total megabyte-seconds taken by all map tasks=29172736
                Total megabyte-seconds taken by all reduce tasks=19991552
        Map-Reduce Framework
                Map input records=312913
                Map output records=1348309
                Map output bytes=13424715
                Map output materialized bytes=16121357
                Input split bytes=238
                Combine input records=0
                Combine output records=0
                Reduce input groups=188
```

In [29]:
```
!hdfs dfs -cat result3s2b/part-00000 | head -5
```

```
APR     3431
Account 16555
Applied 139
Arbitration     168
Bankruptcy      222
```

Mapper is called twice, and reducer is also called twice. Line count for mapper is 312913, and line count for reducer is 188.

Consumer Complaints Dataset using a Mapper, Reducer, and standalone combiner

In [30]:
```
!chmod a+x mappe3r2b.py
!chmod a+x reduce3r2b.py
!hdfs dfs -rm -r result3s2c
!hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
-D mapred.map.tasks=2 \
-D mapred.reduce.tasks=2 \
-mapper /home/cloudera/mappe3r2b.py \
-combiner /home/cloudera/reduce3r2b.py \
-reducer /home/cloudera/reduce3r2b.py \
-input /user/shihyu/Consumer_Complaints.csv \
-output result3s2c
```

```
rm: `result3s2c': No such file or directory
packageJobJar: [] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.6.0-cdh5.8.0.jar
] /tmp/streamjob2891191680511399467.jar tmpDir=null
16/09/13 07:22:26 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/13 07:22:26 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/13 07:22:27 INFO mapred.FileInputFormat: Total input paths to process : 1
16/09/13 07:22:27 INFO mapreduce.JobSubmitter: number of splits:2
16/09/13 07:22:27 INFO Configuration.deprecation: mapred.reduce.tasks is depreca
ted. Instead, use mapreduce.job.reduces
16/09/13 07:22:27 INFO Configuration.deprecation: mapred.map.tasks is deprecated
. Instead, use mapreduce.job.maps
16/09/13 07:22:27 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_14
73444507507_0071
16/09/13 07:22:27 INFO impl.YarnClientImpl: Submitted application application_14
73444507507_0071
16/09/13 07:22:27 INFO mapreduce.Job: The url to track the job: http://quickstar
t.cloudera:8088/proxy/application_1473444507507_0071/
16/09/13 07:22:27 INFO mapreduce.Job: Running job: job_1473444507507_0071
16/09/13 07:22:36 INFO mapreduce.Job: Job job_1473444507507_0071 running in uber
 mode : false
16/09/13 07:22:36 INFO mapreduce.Job:  map 0% reduce 0%
16/09/13 07:22:50 INFO mapreduce.Job:  map 56% reduce 0%
16/09/13 07:22:53 INFO mapreduce.Job:  map 100% reduce 0%
16/09/13 07:23:03 INFO mapreduce.Job:  map 100% reduce 50%
16/09/13 07:23:04 INFO mapreduce.Job:  map 100% reduce 100%
16/09/13 07:23:04 INFO mapreduce.Job: Job job_1473444507507_0071 completed succe
ssfully
16/09/13 07:23:04 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=4971
                FILE: Number of bytes written=485524
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=50910820
                HDFS: Number of bytes written=2342
                HDFS: Number of read operations=12
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Job Counters
                Killed reduce tasks=1
                Launched map tasks=2
                Launched reduce tasks=2
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=30438
                Total time spent by all reduces in occupied slots (ms)=14295
                Total time spent by all map tasks (ms)=30438
                Total time spent by all reduce tasks (ms)=14295
                Total vcore-seconds taken by all map tasks=30438
                Total vcore-seconds taken by all reduce tasks=14295
                Total megabyte-seconds taken by all map tasks=31168512
                Total megabyte-seconds taken by all reduce tasks=14638080
        Map-Reduce Framework
                Map input records=312913
                Map output records=1348309
                Map output bytes=13424715
                Map output materialized bytes=4983
                Input split bytes=238
                Combine input records=1348309
                Combine output records=347
                Reduce input groups=188
                Reduce shuffle bytes=4983
```

In [31]:
```
!hdfs dfs -cat result3s2c/part-00000 | head -5
```

```
APR     3431
Account 16555
Applied 139
Arbitration     168
Bankruptcy      222
```

This time, when we add the combiner, we see that it runs four times, in addition to the two map and reduce tasks. Since we use the reducer as a combiner, we have seen 2 map tasks and 6 reduce tasks.

Using a single reducer: Top 50 most frequent terms and bottom 10 tokens (least frequent items).

In [45]:
```
%%writefile mappe3r2d.py
#!/usr/bin/python

import sys
import re
from csv import reader
WORD_RE = re.compile(r"[\w']+")
total_words = 0

count = 0
sys.stderr.write("reporter:counter:Mapper2d,Script Count,1\n")
for line in reader(sys.stdin):
    sys.stderr.write("reporter:counter:Mapper2d,Line Count,1\n")
    # Considering words in compaints issue
    words = re.findall(WORD_RE, line[3])
    for word in words:
        print '%s\t%s' % (word, 1)
        total_words = total_words + 1

#Also print out total words

print '%s\t%s' % ('!!Total', str(total_words))
```

```
Overwriting mappe3r2d.py
```

In [46]:
```
%%writefile reduce3r2d.py
#!/usr/bin/python

import sys
tmp_word=''
#Counter for the chosen word
count = 0
total_num_words=0

#Increment script call counter once when the file runs
sys.stderr.write("reporter:counter:Reducer2d,Script Calls,1\n")
for line in sys.stdin:
    #Parse line
    line=line.strip().split('\t')
    word,tmp_count=line
    tmp_count=int(tmp_count)

    if word=='!!Total':
        total_num_words = tmp_count
        continue

    if tmp_word==word:
        count+=int(tmp_count)
    else:
        if tmp_word:

            #Increment line call counter when we emit a new word
            sys.stderr.write("reporter:counter:Reducer2d,Line Calls,1\n")
            print tmp_word+'\t'+str(count)+'\t'+str((count+0.0)/(total_num_words +
0.0))

        tmp_word=word
        count=int(tmp_count)

#Do not forget  to emit final record
if tmp_word:
    sys.stderr.write("reporter:counter:Reducer2d,Line Calls,1\n")
    print tmp_word+'\t'+str(count)+'\t'+str((count+0.0)/(total_num_words + 0.0))
```

```
Overwriting reduce3r2d.py
```

Sorting via the Hadoop Shuffle using break tie mapper/reducers

In [35]:
```
%%writefile break_tie.py
#!/usr/bin/python

import sys
for line in sys.stdin:
    print line.strip()
```

```
Writing break_tie.py
```

In [47]:
```
# Generate Hadoop results without sorting
!chmod a+x mappe3r2d.py
!chmod a+x reduce3r2d.py
!hdfs dfs -rm -r result3s2d
!hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
-D mapred.map.tasks=2 \
-D mapred.reduce.tasks=1 \
-mapper /home/cloudera/mappe3r2d.py \
-reducer /home/cloudera/reduce3r2d.py \
-input /user/shihyu/Consumer_Complaints.csv \
-output result3s2d
```

```
Deleted result3s2d
packageJobJar: [] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.6.0-cdh5.8.0.jar
] /tmp/streamjob4950492321238404515.jar tmpDir=null
16/09/13 10:08:47 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/13 10:08:47 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/13 10:08:48 INFO mapred.FileInputFormat: Total input paths to process : 1
16/09/13 10:08:48 INFO mapreduce.JobSubmitter: number of splits:2
16/09/13 10:08:48 INFO Configuration.deprecation: mapred.reduce.tasks is depreca
ted. Instead, use mapreduce.job.reduces
16/09/13 10:08:48 INFO Configuration.deprecation: mapred.map.tasks is deprecated
. Instead, use mapreduce.job.maps
16/09/13 10:08:48 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_14
73444507507_0076
16/09/13 10:08:49 INFO impl.YarnClientImpl: Submitted application application_14
73444507507_0076
16/09/13 10:08:49 INFO mapreduce.Job: The url to track the job: http://quickstar
t.cloudera:8088/proxy/application_1473444507507_0076/
16/09/13 10:08:49 INFO mapreduce.Job: Running job: job_1473444507507_0076
16/09/13 10:08:58 INFO mapreduce.Job: Job job_1473444507507_0076 running in uber
 mode : false
16/09/13 10:08:58 INFO mapreduce.Job:  map 0% reduce 0%
16/09/13 10:09:14 INFO mapreduce.Job:  map 14% reduce 0%
16/09/13 10:09:17 INFO mapreduce.Job:  map 25% reduce 0%
16/09/13 10:09:18 INFO mapreduce.Job:  map 36% reduce 0%
16/09/13 10:09:20 INFO mapreduce.Job:  map 44% reduce 0%
16/09/13 10:09:22 INFO mapreduce.Job:  map 72% reduce 0%
16/09/13 10:09:25 INFO mapreduce.Job:  map 83% reduce 0%
16/09/13 10:09:29 INFO mapreduce.Job:  map 100% reduce 0%
16/09/13 10:09:39 INFO mapreduce.Job:  map 100% reduce 92%
16/09/13 10:09:40 INFO mapreduce.Job:  map 100% reduce 100%
16/09/13 10:09:41 INFO mapreduce.Job: Job job_1473444507507_0076 completed succe
ssfully
16/09/13 10:09:41 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=16121373
                FILE: Number of bytes written=32598398
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=50910820
                HDFS: Number of bytes written=5508
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=1
                Launched map tasks=3
                Launched reduce tasks=1
                Data-local map tasks=3
                Total time spent by all maps in occupied slots (ms)=53152
                Total time spent by all reduces in occupied slots (ms)=15674
                Total time spent by all map tasks (ms)=53152
                Total time spent by all reduce tasks (ms)=15674
                Total vcore-seconds taken by all map tasks=53152
                Total vcore-seconds taken by all reduce tasks=15674
                Total megabyte-seconds taken by all map tasks=54427648
                Total megabyte-seconds taken by all reduce tasks=16050176
        Map-Reduce Framework
                Map input records=312913
                Map output records=1348311
                Map output bytes=13424745
                Map output materialized bytes=16121379
```

In [48]:
```
!hdfs dfs -cat result3s2d/* | head -5
```

```
APR     3431    0.00514780239401
ATM     2422    0.0036339193816
Account 16555   0.0248387842124
Advertising     1193    0.00178995285807
Application     8868    0.0133053662577
cat: Unable to write to output stream.
```

In [49]:
```
### Begin sorting
!chmod a+x break_tie.py
!hdfs dfs -rm -r result3s2d_sorted
!hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
-D mapred.map.tasks=2 \
-D mapred.reduce.tasks=1 \
-D stream.num.map.output.key.fields=2 \
-D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedCom
parator \
-D  mapred.text.key.comparator.options='-k2,2nr' \
-mapper /home/cloudera/break_tie.py \
-reducer /home/cloudera/break_tie.py \
-input result3s2d \
-output result3s2d_sorted
```

```
Deleted result3s2d_sorted
packageJobJar: [] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.6.0-cdh5.8.0.jar
] /tmp/streamjob4914841511058435259.jar tmpDir=null
16/09/13 10:10:07 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/13 10:10:07 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/13 10:10:08 INFO mapred.FileInputFormat: Total input paths to process : 1
16/09/13 10:10:08 INFO mapreduce.JobSubmitter: number of splits:2
16/09/13 10:10:08 INFO Configuration.deprecation: mapred.reduce.tasks is depreca
ted. Instead, use mapreduce.job.reduces
16/09/13 10:10:08 INFO Configuration.deprecation: mapred.output.key.comparator.c
lass is deprecated. Instead, use mapreduce.job.output.key.comparator.class
16/09/13 10:10:08 INFO Configuration.deprecation: mapred.text.key.comparator.opt
ions is deprecated. Instead, use mapreduce.partition.keycomparator.options
16/09/13 10:10:08 INFO Configuration.deprecation: mapred.map.tasks is deprecated
. Instead, use mapreduce.job.maps
16/09/13 10:10:09 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_14
73444507507_0077
16/09/13 10:10:09 INFO impl.YarnClientImpl: Submitted application application_14
73444507507_0077
16/09/13 10:10:09 INFO mapreduce.Job: The url to track the job: http://quickstar
t.cloudera:8088/proxy/application_1473444507507_0077/
16/09/13 10:10:09 INFO mapreduce.Job: Running job: job_1473444507507_0077
16/09/13 10:10:20 INFO mapreduce.Job: Job job_1473444507507_0077 running in uber
 mode : false
16/09/13 10:10:20 INFO mapreduce.Job:  map 0% reduce 0%
16/09/13 10:10:31 INFO mapreduce.Job:  map 50% reduce 0%
16/09/13 10:10:32 INFO mapreduce.Job:  map 100% reduce 0%
16/09/13 10:10:42 INFO mapreduce.Job:  map 100% reduce 100%
16/09/13 10:10:42 INFO mapreduce.Job: Job job_1473444507507_0077 completed succe
ssfully
16/09/13 10:10:43 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=5890
                FILE: Number of bytes written=368821
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=8500
                HDFS: Number of bytes written=5508
                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)=18116
                Total time spent by all reduces in occupied slots (ms)=7683
                Total time spent by all map tasks (ms)=18116
                Total time spent by all reduce tasks (ms)=7683
                Total vcore-seconds taken by all map tasks=18116
                Total vcore-seconds taken by all reduce tasks=7683
                Total megabyte-seconds taken by all map tasks=18550784
                Total megabyte-seconds taken by all reduce tasks=7867392
        Map-Reduce Framework
                Map input records=188
                Map output records=188
                Map output bytes=5508
                Map output materialized bytes=5896
                Input split bytes=238
                Combine input records=0
                Combine output records=0
```

In [53]:
```
! echo "50 Most Common Words:" # !hdfs dfs -cat result3s2c/part-00000 | head -5
! echo "Word | Frequency | Relative Frequency"
!hdfs dfs -cat result3s2d_sorted/* | head -50
! echo "10 Least Common Words:"
! echo "Word | Frequency | Relative Frequency"
!hdfs dfs -cat result3s2d_sorted/* | tail -10
```

```
50 Most Common Words:
Word | Frequency | Relative Frequency
Loan    107254  0.160921713193
collection      70487   0.105757256586
modification    70487   0.105757256586
foreclosure     70487   0.105757256586
account 40893   0.0613550228208
or      40508   0.0607773766763
credit  40483   0.0607398671864
payments        39993   0.0600046811843
escrow  36767   0.0551644566075
servicing       36767   0.0551644566075
report  34903   0.0523677490405
Incorrect       29133   0.0437105587714
on      29069   0.0436145344772
information     29069   0.0436145344772
debt    26531   0.0398065710625
not     18477   0.027722513796
owed    17972   0.0269648221
Cont'd  17972   0.0269648221
attempts        17972   0.0269648221
collect 17972   0.0269648221
Account 16555   0.0248387842124
and     16448   0.0246782435956
closing 16205   0.0243136513538
management      16205   0.0243136513538
opening 16205   0.0243136513538
Credit  14768   0.0221576058743
of      13983   0.0209798078914
loan    12376   0.0185686978806
my      10731   0.0161005734451
withdrawals     10555   0.0158365066362
Deposits        10555   0.0158365066362
Problems        9484    0.0142296000888
Application     8868    0.0133053662577
Communication   8671    0.0130097914772
tactics 8671    0.0130097914772
originator      8625    0.0129407740158
mortgage        8625    0.0129407740158
broker  8625    0.0129407740158
to      8401    0.0126046889863
Billing 8158    0.0122400967445
Other   7886    0.0118319934944
Disclosure      7655    0.0114854058077
verification    7655    0.0114854058077
disputes        6938    0.0104096336373
reporting       6559    0.00984098977041
lease   6337    0.00950790550009
the     6248    0.00937437171604
funds   5663    0.00849664965236
low     5663    0.00849664965236
by      5663    0.00849664965236
cat: Unable to write to output stream.
10 Least Common Words:
Word | Frequency | Relative Frequency
Payment 92      0.000138034922835
credited        92      0.000138034922835
Convenience     75      0.000112528469703
checks  75      0.000112528469703
amt     71      0.000106526951319
day     71      0.000106526951319
wrong   71      0.000106526951319
disclosures     64      9.60242941464e-05
missing 64      9.60242941464e-05
```

3.2.1

Using **2 reducers**: What are the top **50 most frequent terms** in your word count analysis?

In [132]:
```
%%writefile mappe3r21_t.py
#!/usr/bin/python

import sys
import re
from csv import reader
WORD_RE = re.compile(r"[\w']+")

total_words = 0
sys.stderr.write("reporter:counter:Mapper21,Script Count,1\n")
for line in reader(sys.stdin):
    sys.stderr.write("reporter:counter:Mapper21,Line Count,1\n")
    # Considering words in compaints issue
    words = re.findall(WORD_RE, line[3])
    for word in words:
        print '%s\t%s' % (word, 1)
        if word[0].lower() == 'a':
            print '%s\t%s' % ('*', 1)
        else:
            print '%s\t%s' % ('#', 1)
        total_words = total_words + 1

print '%s\t%s' % ('!!Total', str(total_words))
```

```
Overwriting mappe3r21_t.py
```

In [118]:
```
%%writefile combine3r21.py
#!/usr/bin/env python

from itertools import groupby
from operator import itemgetter
import sys

def read(file, separator='\t'):
    for line in file:
        yield line.rstrip().split(separator, 1)

# input comes from STDIN (standard input)
data = read(sys.stdin, separator='\t')
# groupby groups multiple word-count pairs by word,
for tmp_word, group in groupby(data, itemgetter(0)):
    try:
        total_count = sum(int(count) for tmp_word, count in group)
        sys.stderr.write("reporter:counter:Code Call Counters,combiner pairs,1\n")
        sys.stdout.write("{0}{1}{2}\n".format(tmp_word, '\t', total_count))
        if tmp_word == '*':
            sys.stderr.write("reporter:counter:Code Call Counters,combiner total f
lags,1\n")
    except ValueError:
        sys.stderr.write("reporter:counter:Code Call Counters,combiner skipped pai
rs,1\n")
        # count was not a number, so silently discard this item
        pass
```

```
Overwriting combine3r21.py
```

In [134]:
```
%%writefile reduce3r21_t.py
#!/usr/bin/python

from itertools import groupby
from operator import itemgetter
import sys

def read(file, separator='\t'):
    for line in file:
        yield line.rstrip().split(separator, 1)

total = 1
total_first = True
total_words = 0

# input comes from STDIN (standard input)
data = read(sys.stdin, separator='\t')


#Increment script call counter once when the file runs
sys.stderr.write("reporter:counter:Reducer21,Script Calls,1\n")

for line in sys.stdin:
    #Parse line
    line=line.strip().split('\t')
    word,tmp_count=line
    tmp_count=int(tmp_count)
    if word=='!!Total':
        total_words = tmp_count
        break



for tmp_word, group in groupby(data, itemgetter(0)):
    try:
        total_count = sum(int(count) for tmp_word, count in group)

        if tmp_word == '*':
            total = total_count
            sys.stderr.write("reporter:counter:Reducer total indicators,1\n")
            sys.stderr.write("reporter:counter:Reducer word count,{0}\n".format(to
tal))
            if total_first:
                sys.stderr.write("reporter:counter:Reducer recvd total first,1\n")
        elif tmp_word == '#':
            total =total_count
            sys.stderr.write("reporter:counter:Reducer total indicators,1\n")
            sys.stderr.write("reporter:counter:Reducer word count,{0}\n".format(to
tal))
            if total_first:
                sys.stderr.write("reporter:counter:Reducer recvd total first,1\n")

        else:
            #total = total_1 + total_2
            sys.stderr.write("reporter:counter:Code Call Counters,reducer processe
d,1\n")
            print tmp_word+'\t'+str(total_count)+'\t'+str(float(total_count)/float
(total_words))

            total_first = False

    except ValueError:
        sys.stderr.write("reporter:counter:Code Call Counters,reducer skipped pair
s,1\n")
```

```
Overwriting reduce3r21_t.py
```

In [135]:
```
# Generate Hadoop results without sorting
!chmod a+x mappe3r21_t.py
!chmod a+x combine3r21.py
!chmod a+x reduce3r21_t.py
!hdfs dfs -rm -r result3s21_t
!hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
-D mapred.map.tasks=2 \
-D mapred.reduce.tasks=2 \
-mapper /home/cloudera/mappe3r21_t.py \
-combiner /home/cloudera/combine3r21.py \
-reducer /home/cloudera/reduce3r21_t.py \
-input /user/shihyu/Consumer_Complaints.csv \
-output result3s21_t
```

```
Deleted result3s21_t
packageJobJar: [] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.6.0-cdh5.8.0.jar
] /tmp/streamjob4180188453315537109.jar tmpDir=null
16/09/13 20:24:17 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/13 20:24:17 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/13 20:24:18 INFO mapred.FileInputFormat: Total input paths to process : 1
16/09/13 20:24:18 INFO mapreduce.JobSubmitter: number of splits:2
16/09/13 20:24:18 INFO Configuration.deprecation: mapred.reduce.tasks is depreca
ted. Instead, use mapreduce.job.reduces
16/09/13 20:24:18 INFO Configuration.deprecation: mapred.map.tasks is deprecated
. Instead, use mapreduce.job.maps
16/09/13 20:24:19 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_14
73444507507_0105
16/09/13 20:24:19 INFO impl.YarnClientImpl: Submitted application application_14
73444507507_0105
16/09/13 20:24:19 INFO mapreduce.Job: The url to track the job: http://quickstar
t.cloudera:8088/proxy/application_1473444507507_0105/
16/09/13 20:24:19 INFO mapreduce.Job: Running job: job_1473444507507_0105
16/09/13 20:24:30 INFO mapreduce.Job: Job job_1473444507507_0105 running in uber
 mode : false
16/09/13 20:24:30 INFO mapreduce.Job:  map 0% reduce 0%
16/09/13 20:24:46 INFO mapreduce.Job:  map 9% reduce 0%
16/09/13 20:24:48 INFO mapreduce.Job:  map 16% reduce 0%
16/09/13 20:24:49 INFO mapreduce.Job:  map 22% reduce 0%
16/09/13 20:24:52 INFO mapreduce.Job:  map 34% reduce 0%
16/09/13 20:24:55 INFO mapreduce.Job:  map 46% reduce 0%
16/09/13 20:24:58 INFO mapreduce.Job:  map 58% reduce 0%
16/09/13 20:25:02 INFO mapreduce.Job:  map 65% reduce 0%
16/09/13 20:25:05 INFO mapreduce.Job:  map 67% reduce 0%
16/09/13 20:25:06 INFO mapreduce.Job:  map 83% reduce 0%
16/09/13 20:25:08 INFO mapreduce.Job:  map 100% reduce 0%
16/09/13 20:25:20 INFO mapreduce.Job:  map 100% reduce 100%
16/09/13 20:25:21 INFO mapreduce.Job: Job job_1473444507507_0105 completed succe
ssfully
16/09/13 20:25:21 INFO mapreduce.Job: Counters: 55
        File System Counters
                FILE: Number of bytes read=5047
                FILE: Number of bytes written=485704
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=50910820
                HDFS: Number of bytes written=2868
                HDFS: Number of read operations=12
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=2
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=68922
                Total time spent by all reduces in occupied slots (ms)=20634
                Total time spent by all map tasks (ms)=68922
                Total time spent by all reduce tasks (ms)=20634
                Total vcore-seconds taken by all map tasks=68922
                Total vcore-seconds taken by all reduce tasks=20634
                Total megabyte-seconds taken by all map tasks=70576128
                Total megabyte-seconds taken by all reduce tasks=21129216
        Map-Reduce Framework
                Map input records=312913
                Map output records=2696620
                Map output bytes=18817981
```

In [136]:
```
! echo "50 Most Common Words with 2 reducers:" # !hdfs dfs -cat result3s2c/part-00000 | head -5
! echo "Word | Frequency | Relative Frequency"
!hdfs dfs -cat result3s21_t/* | head -15
```

```
50 Most Common Words with 2 reducers:
Word | Frequency | Relative Frequency
!!Total 681811  1.02297531275
ATM     2422    0.0036339193816
Advertising     1193    0.00178995285807
Application     8868    0.0133053662577
Balance 597     0.000895726618835
Cancelling      2795    0.00419356097093
Charged 878     0.00131733328532
Collection      1907    0.00286122388964
Communication   8671    0.0130097914772
Customer        2734    0.00410203781557
Dealing 1944    0.0029167379347
Embezzlement    3276    0.00491524355662
Forbearance     350     0.000525132858613
Fraud   3842    0.00576445840798
Getting 291     0.000436610462447
```

In [137]:
```
### Begin sorting
!chmod a+x break_tie.py
!hdfs dfs -rm -r result3s21_t_sorted
!hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
-D mapred.map.tasks=2 \
-D mapred.reduce.tasks=1 \
-D stream.num.map.output.key.fields=2 \
-D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedCo
mparator \
-D  mapred.text.key.comparator.options='-k2,2nr' \
-mapper /home/cloudera/break_tie.py \
-reducer /home/cloudera/break_tie.py \
-input result3s21_t \
-output result3s21_t_sorted
```

```
Deleted result3s21_t_sorted
packageJobJar: [] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.6.0-cdh5.8.0.jar
] /tmp/streamjob8926523487061108341.jar tmpDir=null
16/09/13 20:32:37 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/13 20:32:38 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/13 20:32:39 INFO mapred.FileInputFormat: Total input paths to process : 2
16/09/13 20:32:39 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1281)
        at java.lang.Thread.join(Thread.java:1355)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DF
SOutputStream.java:862)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutpu
tStream.java:600)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStre
am.java:789)
16/09/13 20:32:39 INFO mapreduce.JobSubmitter: number of splits:3
16/09/13 20:32:39 INFO Configuration.deprecation: mapred.reduce.tasks is depreca
ted. Instead, use mapreduce.job.reduces
16/09/13 20:32:39 INFO Configuration.deprecation: mapred.output.key.comparator.c
lass is deprecated. Instead, use mapreduce.job.output.key.comparator.class
16/09/13 20:32:39 INFO Configuration.deprecation: mapred.text.key.comparator.opt
ions is deprecated. Instead, use mapreduce.partition.keycomparator.options
16/09/13 20:32:39 INFO Configuration.deprecation: mapred.map.tasks is deprecated
. Instead, use mapreduce.job.maps
16/09/13 20:32:39 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_14
73444507507_0106
16/09/13 20:32:40 INFO impl.YarnClientImpl: Submitted application application_14
73444507507_0106
16/09/13 20:32:40 INFO mapreduce.Job: The url to track the job: http://quickstar
t.cloudera:8088/proxy/application_1473444507507_0106/
16/09/13 20:32:40 INFO mapreduce.Job: Running job: job_1473444507507_0106
16/09/13 20:32:52 INFO mapreduce.Job: Job job_1473444507507_0106 running in uber
 mode : false
16/09/13 20:32:52 INFO mapreduce.Job:  map 0% reduce 0%
16/09/13 20:33:06 INFO mapreduce.Job:  map 33% reduce 0%
16/09/13 20:33:10 INFO mapreduce.Job:  map 67% reduce 0%
16/09/13 20:33:12 INFO mapreduce.Job:  map 100% reduce 0%
16/09/13 20:33:18 INFO mapreduce.Job:  map 100% reduce 100%
16/09/13 20:33:18 INFO mapreduce.Job: Job job_1473444507507_0106 completed succe
ssfully
16/09/13 20:33:18 INFO mapreduce.Job: Counters: 51
        File System Counters
                FILE: Number of bytes read=3068
                FILE: Number of bytes written=482229
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=4665
                HDFS: Number of bytes written=2868
                HDFS: Number of read operations=12
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=1
                Launched map tasks=3
                Launched reduce tasks=1
                Other local map tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=42946
                Total time spent by all reduces in occupied slots (ms)=8876
```

In [139]:
```
! echo "50 Most Common Words with 2 reducers:" # !hdfs dfs -cat result3s2c/part-00
000 | head -5
! echo "Word | Frequency | Relative Frequency"
!hdfs dfs -cat result3s21_t_sorted/* | head -50
```

```
50 Most Common Words with 2 reducers:
Word | Frequency | Relative Frequency
!!Total 681811  1.02297531275
Loan    107254  0.160921713193
collection      70487   0.105757256586
modification    70487   0.105757256586
servicing       36767   0.0551644566075
report  34903   0.0523677490405
information     29069   0.0436145344772
attempts        17972   0.0269648221
collect 17972   0.0269648221
opening 16205   0.0243136513538
loan    12376   0.0185686978806
my      10731   0.0161005734451
withdrawals     10555   0.0158365066362
Problems        9484    0.0142296000888
Application     8868    0.0133053662577
Communication   8671    0.0130097914772
mortgage        8625    0.0129407740158
originator      8625    0.0129407740158
Other   7886    0.0118319934944
reporting       6559    0.00984098977041
lease   6337    0.00950790550009
low     5663    0.00849664965236
funds   5663    0.00849664965236
Managing        5006    0.00751090025777
Improper        4966    0.00745088507392
investigation   4858    0.00728884407755
card    4405    0.00660917212055
score   4357    0.00653715389994
get     4357    0.00653715389994
costs   4350    0.00652665124276
interest        4238    0.00635860872801
Taking  4206    0.00631059658093
when    4095    0.00614405444577
Fraud   3842    0.00576445840798
are     3821    0.00573295043646
pay     3821    0.00573295043646
contact 3710    0.0055664083013
statements      3621    0.00543287451725
info    3553    0.00533084870472
sharing 3489    0.00523482441058
rate    3431    0.00514780239401
money   3365    0.00504877734067
Identity        3276    0.00491524355662
Embezzlement    3276    0.00491524355662
receiving       3226    0.00484022457682
sending 3226    0.00484022457682
fee     3198    0.00479821394813
illegal 2964    0.00444712512266
action  2964    0.00444712512266
Cancelling      2795    0.00419356097093
```

HW3.3. Shopping Cart Analysis

In [88]:
```
# put ProductPurchaseData.txt data
!hdfs dfs -mkdir -p /user/shihyu
!hdfs dfs -put ProductPurchaseData.txt /user/shihyu
```

```
put: `/user/shihyu/ProductPurchaseData.txt': File exists
```

In [166]:
```
%%writefile mappe3r3.py
#!/usr/bin/python

import sys

product_count=0
# cart index from 1
cart_id = 0
for line in sys.stdin:
    line=line.strip()
    products=line.split() #split on whitespace
    cart_id = cart_id + 1
    for product in products:
        product_count+=1
        print product+' '+str(cart_id)+' 1 '+str(len(products))



#Emit total with special key for order inversion
print '**Total '+'0'+' '+str(product_count)+' 0'
```

```
Overwriting mappe3r3.py
```

In [167]:
```
%%writefile reduce3r3.py
#!/usr/bin/python

from __future__ import division
import sys
tmp_product=None
# counter for a paricular product
count = 0
largest_basket_id=0
largest_basket_size=0
unique_products=0
total_product_count=0

for line in sys.stdin:
    #Parse line into fields
    try:
         product,cart_id,product_count,cart_total = line.strip().split(' ')
    except ValueError:
        continue



    cart_total=int(cart_total)

    #Extract total products
    if product=='**Total':
        total_product_count+=int(product_count)
        continue

    #Updated largest cart size and ID
    if cart_total>largest_basket_size:
        largest_basket_size=cart_total
        largest_basket_id=cart_id

    if tmp_product==product:
        count+=int(product_count)
    else:
        if tmp_product and tmp_product!='**Total':
            print tmp_product+'\t'+str(count)+'\t'+ str((count + 0.0)/(total_produ
ct_count + 0.0))
            unique_products+=1
        tmp_product=product
        count=int(product_count)
# Dont forget last one
if tmp_product:
    print tmp_product+'\t'+str(count)+'\t'+ str((count + 0.0)/(total_product_count
 + 0.0))
    unique_products+=1

#Print aggregated stats separately with special key to make them easy to find
print '*Largest Size Cart' + '\t' + str(largest_basket_id) + '\t '+ str(largest_ba
sket_size)
print '*Unique Products'+'\t'+str(unique_products)
```

```
Overwriting reduce3r3.py
```

In [168]:
```
# Generate Hadoop results without sorting
!chmod a+x mappe3r3.py
!chmod a+x reduce3r3.py
!hdfs dfs -rm -r result3s3
!hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
-D mapred.map.tasks=2 \
-D mapred.reduce.tasks=1 \
-mapper /home/cloudera/mappe3r3.py \
-reducer /home/cloudera/reduce3r3.py \
-input /user/shihyu/ProductPurchaseData.txt \
-output result3s3
```

```
Deleted result3s3
packageJobJar: [] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.6.0-cdh5.8.0.jar
] /tmp/streamjob3340586247327220801.jar tmpDir=null
16/09/13 22:25:04 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/13 22:25:04 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/13 22:25:05 INFO mapred.FileInputFormat: Total input paths to process : 1
16/09/13 22:25:06 INFO mapreduce.JobSubmitter: number of splits:2
16/09/13 22:25:06 INFO Configuration.deprecation: mapred.reduce.tasks is depreca
ted. Instead, use mapreduce.job.reduces
16/09/13 22:25:06 INFO Configuration.deprecation: mapred.map.tasks is deprecated
. Instead, use mapreduce.job.maps
16/09/13 22:25:06 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_14
73444507507_0116
16/09/13 22:25:06 INFO impl.YarnClientImpl: Submitted application application_14
73444507507_0116
16/09/13 22:25:07 INFO mapreduce.Job: The url to track the job: http://quickstar
t.cloudera:8088/proxy/application_1473444507507_0116/
16/09/13 22:25:07 INFO mapreduce.Job: Running job: job_1473444507507_0116
16/09/13 22:25:18 INFO mapreduce.Job: Job job_1473444507507_0116 running in uber
 mode : false
16/09/13 22:25:18 INFO mapreduce.Job:  map 0% reduce 0%
16/09/13 22:25:34 INFO mapreduce.Job:  map 50% reduce 0%
16/09/13 22:25:35 INFO mapreduce.Job:  map 100% reduce 0%
16/09/13 22:25:45 INFO mapreduce.Job:  map 100% reduce 100%
16/09/13 22:25:46 INFO mapreduce.Job: Job job_1473444507507_0116 completed succe
ssfully
16/09/13 22:25:46 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=8406183
                FILE: Number of bytes written=17168009
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=3462851
                HDFS: Number of bytes written=368686
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=1
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=26298
                Total time spent by all reduces in occupied slots (ms)=9015
                Total time spent by all map tasks (ms)=26298
                Total time spent by all reduce tasks (ms)=9015
                Total vcore-seconds taken by all map tasks=26298
                Total vcore-seconds taken by all reduce tasks=9015
                Total megabyte-seconds taken by all map tasks=26929152
                Total megabyte-seconds taken by all reduce tasks=9231360
        Map-Reduce Framework
                Map input records=31101
                Map output records=380826
                Map output bytes=7644525
                Map output materialized bytes=8406189
                Input split bytes=238
                Combine input records=0
                Combine output records=0
                Reduce input groups=380577
                Reduce shuffle bytes=8406189
                Reduce input records=380826
```

In [169]:
```
### Begin sorting
!chmod a+x break_tie.py
!hdfs dfs -rm -r result3s3_sorted
!hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
-D mapred.map.tasks=2 \
-D mapred.reduce.tasks=1 \
-D stream.num.map.output.key.fields=2 \
-D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedCo
mparator \
-D  mapred.text.key.comparator.options='-k2,2nr' \
-mapper /home/cloudera/break_tie.py \
-reducer /home/cloudera/break_tie.py \
-input result3s3 \
-output result3s3_sorted
```

```
Deleted result3s3_sorted
packageJobJar: [] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.6.0-cdh5.8.0.jar
] /tmp/streamjob5072414457235328786.jar tmpDir=null
16/09/13 22:26:02 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/13 22:26:03 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/13 22:26:04 INFO mapred.FileInputFormat: Total input paths to process : 1
16/09/13 22:26:04 INFO mapreduce.JobSubmitter: number of splits:2
16/09/13 22:26:04 INFO Configuration.deprecation: mapred.reduce.tasks is depreca
ted. Instead, use mapreduce.job.reduces
16/09/13 22:26:04 INFO Configuration.deprecation: mapred.output.key.comparator.c
lass is deprecated. Instead, use mapreduce.job.output.key.comparator.class
16/09/13 22:26:04 INFO Configuration.deprecation: mapred.text.key.comparator.opt
ions is deprecated. Instead, use mapreduce.partition.keycomparator.options
16/09/13 22:26:04 INFO Configuration.deprecation: mapred.map.tasks is deprecated
. Instead, use mapreduce.job.maps
16/09/13 22:26:05 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_14
73444507507_0117
16/09/13 22:26:05 INFO impl.YarnClientImpl: Submitted application application_14
73444507507_0117
16/09/13 22:26:05 INFO mapreduce.Job: The url to track the job: http://quickstar
t.cloudera:8088/proxy/application_1473444507507_0117/
16/09/13 22:26:05 INFO mapreduce.Job: Running job: job_1473444507507_0117
16/09/13 22:26:15 INFO mapreduce.Job: Job job_1473444507507_0117 running in uber
 mode : false
16/09/13 22:26:15 INFO mapreduce.Job:  map 0% reduce 0%
16/09/13 22:26:32 INFO mapreduce.Job:  map 100% reduce 0%
16/09/13 22:26:45 INFO mapreduce.Job:  map 100% reduce 100%
16/09/13 22:26:46 INFO mapreduce.Job: Job job_1473444507507_0117 completed succe
ssfully
16/09/13 22:26:46 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=393881
                FILE: Number of bytes written=1144797
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=373018
                HDFS: Number of bytes written=368686
                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)=28896
                Total time spent by all reduces in occupied slots (ms)=11081
                Total time spent by all map tasks (ms)=28896
                Total time spent by all reduce tasks (ms)=11081
                Total vcore-seconds taken by all map tasks=28896
                Total vcore-seconds taken by all reduce tasks=11081
                Total megabyte-seconds taken by all map tasks=29589504
                Total megabyte-seconds taken by all reduce tasks=11346944
        Map-Reduce Framework
                Map input records=12594
                Map output records=12594
                Map output bytes=368687
                Map output materialized bytes=393887
                Input split bytes=236
                Combine input records=0
                Combine output records=0
                Reduce input groups=12594
```

In [170]:
```
! echo "50 Most Common Products:"
! echo "Product | Frequency | Relative Frequency"
!hdfs dfs -cat result3s3_sorted/* | head -50
```

```
50 Most Common Products:
Product | Frequency | Relative Frequency
*Unique Products        12592
*Largest Size Cart      6914     37
DAI62779        6667    0.0175067747831
FRO40251        3881    0.010191059387
ELE17451        3875    0.0101753040775
GRO73461        3602    0.00945843749344
SNA80324        3044    0.00799319370628
ELE32164        2851    0.0074863979161
DAI75645        2736    0.00718442114993
SNA45677        2455    0.0064465474865
FRO31317        2330    0.0061183118711
DAI85309        2293    0.00602115412894
ELE26917        2292    0.00601852824402
FRO80039        2233    0.00586360103355
GRO21487        2115    0.00555374661261
SNA99873        2083    0.00546971829507
GRO59710        2004    0.00526227338613
GRO71621        1920    0.00504169905258
FRO85978        1918    0.00503644728273
GRO30386        1840    0.00483162825872
ELE74009        1816    0.00476860702057
GRO56726        1784    0.00468457870302
DAI63921        1773    0.00465569396887
GRO46854        1756    0.00461105392517
ELE66600        1713    0.00449814087347
DAI83733        1712    0.00449551498855
FRO32293        1702    0.00446925613932
ELE66810        1697    0.0044561267147
SNA55762        1646    0.00432220658362
DAI22177        1627    0.00427231477008
FRO78087        1531    0.00402022981745
ELE99737        1516    0.0039808415436
ELE34057        1489    0.00390994265067
GRO94758        1489    0.00390994265067
FRO35904        1436    0.00377077074974
FRO53271        1420    0.00372875659097
SNA93860        1407    0.00369462008697
SNA90094        1390    0.00364998004327
GRO38814        1352    0.00355019641619
ELE56788        1345    0.00353181522173
GRO61133        1321    0.00346879398357
ELE74482        1316    0.00345566455896
DAI88807        1316    0.00345566455896
ELE59935        1311    0.00344253513434
SNA96271        1295    0.00340052097557
DAI43223        1290    0.00338739155095
ELE91337        1289    0.00338476566603
GRO15017        1275    0.0033480032771
DAI31081        1261    0.00331124088818
GRO81087        1220    0.00320357960633
cat: Unable to write to output stream.
```

According to these results, we have 12592 unique products browsed. The largest cart has 37 products, and the most commonly browsed product was DAI62779, which was selected by buyers 6667 times for a relative frequency of 0.0175.

3.3.1 OPTIONAL Using 2 reducers: 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 [ ]:

HW3.4. (Computationally prohibitive but then again Hadoop can handle this) Pairs

In [182]:
```
%%writefile mappe3r4.py
#!/usr/bin/python
import sys

for line in sys.stdin:
    # get all products from the session
    products = line.strip().split(' ')
    cart_size = len(products)
    if cart_size==0:
        continue

    # sort products the pair is lexicographically sound
    products.sort()

    # set pairs of products
    pairs = [[products[i], products[j]] for i in range(cart_size) for j in range(i
+1, cart_size)]

    # dummy record for total products count
    print '%s,%s' %('*', 1)

    # emit product pairs
    for pair in pairs:
        print '%s_%s,%s' %(pair[0], pair[1], 1)
```

```
Overwriting mappe3r4.py
```

In [183]:
```
%%writefile combine3r4.py
#!/usr/bin/python
import sys


tmp_pair = None
tmp_count = 0

for line in sys.stdin:
    # get all products from each line
    pair, count = line.strip().split(',', 1)

    # skip bad value
    try:
        count = int(count)
    except ValueError:
        continue

    # accumulate counts for whatever keys it receives
    if tmp_pair == pair:
        tmp_count += count
    else:
        # previous pair finishes streaming, emit results
        if tmp_pair:
            print '%s,%s' %(tmp_pair, tmp_count)
        # set new pair
        tmp_pair = pair
        tmp_count = count
```

```
Overwriting combine3r4.py
```

In [184]:
```
%%writefile reduce3r4.py
#!/usr/bin/python
import sys


total_products = 0
min_support = 100
tmp_pair = None
tmp_count = 0

for line in sys.stdin:
    # get all products from the session
    pair, count = line.strip().split(',', 1)

    # skip bad count
    try:
        count = int(count)
    except ValueError:
        continue

    # get total sessions/baskets
    if pair == '*':
        total_products += count
        continue

    # get pair count
    if tmp_pair == pair:
        tmp_count += count
    else:
        # previous pair finishes
        if tmp_pair and tmp_count > min_support:
            # emit
            print '%s,%s,%s' %(tmp_pair, tmp_count, str(float(tmp_count)/float(tot
al_products)))
        # reset new pair
        tmp_pair = pair
        tmp_count = count
```

```
Overwriting reduce3r4.py
```

In [185]:
```
# Generate Hadoop results without sorting
!chmod a+x mappe3r4.py
!chmod a+x combine3r4.py
!chmod a+x reduce3r4.py
!hdfs dfs -rm -r result3s4
!hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
-D mapred.map.tasks=2 \
-D mapred.reduce.tasks=1 \
-mapper /home/cloudera/mappe3r4.py \
-combiner /home/cloudera/combine3r4.py \
-reducer /home/cloudera/reduce3r4.py \
-input /user/shihyu/ProductPurchaseData.txt \
-output result3s4
```

```
Deleted result3s4
packageJobJar: [] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.6.0-cdh5.8.0.jar
] /tmp/streamjob7527278688606732722.jar tmpDir=null
16/09/14 19:05:07 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/14 19:05:08 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/14 19:05:09 INFO mapred.FileInputFormat: Total input paths to process : 1
16/09/14 19:05:09 INFO mapreduce.JobSubmitter: number of splits:2
16/09/14 19:05:09 INFO Configuration.deprecation: mapred.reduce.tasks is depreca
ted. Instead, use mapreduce.job.reduces
16/09/14 19:05:09 INFO Configuration.deprecation: mapred.map.tasks is deprecated
. Instead, use mapreduce.job.maps
16/09/14 19:05:10 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_14
73444507507_0122
16/09/14 19:05:10 INFO impl.YarnClientImpl: Submitted application application_14
73444507507_0122
16/09/14 19:05:10 INFO mapreduce.Job: The url to track the job: http://quickstar
t.cloudera:8088/proxy/application_1473444507507_0122/
16/09/14 19:05:10 INFO mapreduce.Job: Running job: job_1473444507507_0122
16/09/14 19:05:23 INFO mapreduce.Job: Job job_1473444507507_0122 running in uber
 mode : false
16/09/14 19:05:23 INFO mapreduce.Job:  map 0% reduce 0%
16/09/14 19:05:42 INFO mapreduce.Job:  map 30% reduce 0%
16/09/14 19:05:44 INFO mapreduce.Job:  map 51% reduce 0%
16/09/14 19:05:45 INFO mapreduce.Job:  map 54% reduce 0%
16/09/14 19:05:47 INFO mapreduce.Job:  map 67% reduce 0%
16/09/14 19:05:55 INFO mapreduce.Job:  map 83% reduce 0%
16/09/14 19:05:59 INFO mapreduce.Job:  map 100% reduce 0%
16/09/14 19:06:09 INFO mapreduce.Job:  map 100% reduce 96%
16/09/14 19:06:10 INFO mapreduce.Job:  map 100% reduce 100%
16/09/14 19:06:11 INFO mapreduce.Job: Job job_1473444507507_0122 completed succe
ssfully
16/09/14 19:06:12 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=23650909
                FILE: Number of bytes written=47658484
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=3462851
                HDFS: Number of bytes written=52179
                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)=64749
                Total time spent by all reduces in occupied slots (ms)=12714
                Total time spent by all map tasks (ms)=64749
                Total time spent by all reduce tasks (ms)=12714
                Total vcore-seconds taken by all map tasks=64749
                Total vcore-seconds taken by all reduce tasks=12714
                Total megabyte-seconds taken by all map tasks=66302976
                Total megabyte-seconds taken by all reduce tasks=13019136
        Map-Reduce Framework
                Map input records=31101
                Map output records=2565158
                Map output bytes=53370702
                Map output materialized bytes=23650915
                Input split bytes=238
                Combine input records=2565158
```

In [188]:
```
%%writefile mappe3r4_s.py
#!/usr/bin/python
import sys

for line in sys.stdin:
    # just emit
    print line.strip()
```

```
Writing mappe3r4_s.py
```

In [189]:
```
%%writefile reduce3r4_s.py
#!/usr/bin/python
import sys

n = 0
top = 50

for line in sys.stdin:
    # parse mappe3r4_s output
    pair, count, relative_freq = line.strip().split(',', 2)
    n += 1
    if n <= top:
        w1, w2 = pair.split('_')
        print '%s\t%s\t%s\t%s' %(w1, w2, count, relative_freq)
```

```
Writing reduce3r4_s.py
```

In [192]:
```
### Begin sorting
!chmod a+x mappe3r4_s.py
!chmod a+x reduce3r4_s.py
!hdfs dfs -rm -r result3s4_sorted
!hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
-D mapred.map.tasks=2 \
-D mapred.reduce.tasks=1 \
-D stream.num.map.output.key.fields=2 \
-D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedCo
mparator \
-D map.output.key.field.separator=',' \
-D map.output.key.value.fields.spec=0-1:2- \
-D mapred.text.key.comparator.options='-k2,2nr -k1,1' \
-mapper /home/cloudera/mappe3r4_s.py \
-reducer /home/cloudera/reduce3r4_s.py \
-input result3s4 \
-output result3s4_sorted
```

```
Deleted result3s4_sorted
packageJobJar: [] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.6.0-cdh5.8.0.jar
] /tmp/streamjob6464004570836655768.jar tmpDir=null
16/09/14 19:52:59 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/14 19:53:00 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/14 19:53:01 INFO mapred.FileInputFormat: Total input paths to process : 1
16/09/14 19:53:01 INFO mapreduce.JobSubmitter: number of splits:2
16/09/14 19:53:01 INFO Configuration.deprecation: mapred.reduce.tasks is depreca
ted. Instead, use mapreduce.job.reduces
16/09/14 19:53:01 INFO Configuration.deprecation: mapred.output.key.comparator.c
lass is deprecated. Instead, use mapreduce.job.output.key.comparator.class
16/09/14 19:53:01 INFO Configuration.deprecation: mapred.text.key.comparator.opt
ions is deprecated. Instead, use mapreduce.partition.keycomparator.options
16/09/14 19:53:01 INFO Configuration.deprecation: map.output.key.field.separator
 is deprecated. Instead, use mapreduce.map.output.key.field.separator
16/09/14 19:53:01 INFO Configuration.deprecation: map.output.key.value.fields.sp
ec is deprecated. Instead, use mapreduce.fieldsel.map.output.key.value.fields.sp
ec
16/09/14 19:53:01 INFO Configuration.deprecation: mapred.map.tasks is deprecated
. Instead, use mapreduce.job.maps
16/09/14 19:53:02 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_14
73444507507_0125
16/09/14 19:53:02 INFO impl.YarnClientImpl: Submitted application application_14
73444507507_0125
16/09/14 19:53:02 INFO mapreduce.Job: The url to track the job: http://quickstar
t.cloudera:8088/proxy/application_1473444507507_0125/
16/09/14 19:53:02 INFO mapreduce.Job: Running job: job_1473444507507_0125
16/09/14 19:53:18 INFO mapreduce.Job: Job job_1473444507507_0125 running in uber
 mode : false
16/09/14 19:53:18 INFO mapreduce.Job:  map 0% reduce 0%
16/09/14 19:53:32 INFO mapreduce.Job:  map 100% reduce 0%
16/09/14 19:53:43 INFO mapreduce.Job:  map 100% reduce 100%
16/09/14 19:53:43 INFO mapreduce.Job: Job job_1473444507507_0125 completed succe
ssfully
16/09/14 19:53:43 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=54807
                FILE: Number of bytes written=467528
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=56511
                HDFS: Number of bytes written=1898
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=1
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=24389
                Total time spent by all reduces in occupied slots (ms)=7963
                Total time spent by all map tasks (ms)=24389
                Total time spent by all reduce tasks (ms)=7963
                Total vcore-seconds taken by all map tasks=24389
                Total vcore-seconds taken by all reduce tasks=7963
                Total megabyte-seconds taken by all map tasks=24974336
                Total megabyte-seconds taken by all reduce tasks=8154112
        Map-Reduce Framework
                Map input records=1311
                Map output records=1311
```

In [193]:
```
! echo "50 Most Common Pairs of Products:"
! echo "Product Pairs| Frequency | Relative Frequency"
!hdfs dfs -cat result3s4_sorted/* | head -50
```

```
50 Most Common Pairs of Products:
Product Pairs| Frequency | Relative Frequency
DAI62779        ELE17451        1592    0.0511880646925
FRO40251        SNA80324        1412    0.0454004694383
DAI75645        FRO40251        1254    0.0403202469374
FRO40251        GRO85051        1213    0.0390019613517
DAI62779        GRO73461        1139    0.0366226166361
DAI75645        SNA80324        1130    0.0363332368734
DAI62779        FRO40251        1070    0.0344040384554
DAI62779        SNA80324        923     0.0296775023311
DAI62779        DAI85309        918     0.0295167357963
ELE32164        GRO59710        911     0.0292916626475
DAI62779        DAI75645        882     0.0283592167454
FRO40251        GRO73461        882     0.0283592167454
DAI62779        ELE92920        877     0.0281984502106
FRO40251        FRO92469        835     0.026848011318
DAI62779        ELE32164        832     0.0267515513971
DAI75645        GRO73461        712     0.0228931545609
DAI43223        ELE32164        711     0.022861001254
DAI62779        GRO30386        709     0.02279669464
ELE17451        FRO40251        697     0.0224108549564
DAI85309        ELE99737        659     0.0211890292917
DAI62779        ELE26917        650     0.020899649529
GRO21487        GRO73461        631     0.0202887366966
DAI62779        SNA45677        604     0.0194205974084
ELE17451        SNA80324        597     0.0191955242597
DAI62779        GRO71621        595     0.0191312176457
DAI62779        SNA55762        593     0.0190669110318
DAI62779        DAI83733        586     0.018841837883
ELE17451        GRO73461        580     0.0186489180412
GRO73461        SNA80324        562     0.0180701585158
DAI62779        GRO59710        561     0.0180380052088
DAI62779        FRO80039        550     0.0176843188322
DAI75645        ELE17451        547     0.0175878589113
DAI62779        SNA93860        537     0.0172663258416
DAI55148        DAI62779        526     0.016912639465
DAI43223        GRO59710        512     0.0164624931674
ELE17451        ELE32164        511     0.0164303398605
DAI62779        SNA18336        506     0.0162695733256
ELE32164        GRO73461        486     0.0156265071863
DAI62779        FRO78087        482     0.0154978939584
DAI85309        ELE17451        482     0.0154978939584
DAI62779        GRO94758        479     0.0154014340375
DAI62779        GRO21487        471     0.0151442075817
GRO85051        SNA80324        471     0.0151442075817
ELE17451        GRO30386        468     0.0150477476608
FRO85978        SNA95666        463     0.014886981126
DAI62779        FRO19221        462     0.014854827819
DAI62779        GRO46854        461     0.0148226745121
DAI43223        DAI62779        459     0.0147583678981
ELE92920        SNA18336        455     0.0146297546703
DAI88079        FRO40251        446     0.0143403749076
```

HW3.5: Stripes

In [208]:
```
%%writefile mappe3r5.py
#!/usr/bin/python
import sys

# mapper counter
sys.stderr.write("reporter:counter:HW3_5, Mapper_counter,1\n")

# Associative array
A = {}

for line in sys.stdin:
    # get all products from the session
    products = line.strip().split(' ')
    product_size = len(products)
    if product_size==0:
        continue

    #  lexicographically sort
    products.sort()

    # get pairs of products
    pairs = [[products[i], products[j]] for i in range(product_size) for j in rang
e(i+1, product_size)]

    # emit dummy record for total count
    print '%s\t%s' %('*', 1)

    # prepare associative arrays
    for w1, w2 in pairs:

        if w1 not in A:
            # if w1 is new, add to associative array
            A[w1] = {}
            A[w1][w2] = 1
        elif w2 not in A[w1]:
            # w1 is not new, but it doesn't have key for w2
            A[w1][w2] = 1
        else:
            # both are there, increase it
            A[w1][w2] += 1

# emit associative arrays
for a in A:
    print '%s\t%s' %(a, str(A[a]))
```

```
Overwriting mappe3r5.py
```

In [209]:
```
%%writefile reduce3r5.py
#!/usr/bin/python

# function to combine associative array
def elementSum(A1, A2):
    # make sure A1 is the long one
    if len(A1)<len(A2):
        A0 = A2
        A2 = A1
        A1 = A0
    # merge shorter one into longer one
    for a in A2:
        if a not in A1:
            A1[a] = A2[a]
        else:
            A1[a] += A2[a]
    # return
    return A1

import sys
import numpy as np

# increase counter for reducer called
sys.stderr.write("reporter:counter:HW3_5,Reducer_counter,1\n")

min_support = 100
tmp_word = None
tmp_Array = None
total_products = 0

for line in sys.stdin:
    # parse keyword and the associative array
    word, Array = line.strip().split('\t', 1)

    # get total basket
    if word == '*':
        total_products += int(Array)
        continue

    # get array into variable
    cmdStr = 'Array = ' + Array
    exec cmdStr

    # merge the associative array
    if tmp_word == word:
        tmp_aArray = elementSum(tmp_Array, Array)
    else:
        # finish one word merge
        if tmp_word:
            # get the top pairs with heap
            for p in tmp_Array:
                if tmp_Array[p] > min_support:

                    print '%s,%s,%s,%s' %(tmp_word, p, tmp_Array[p], str(float(tmp
_Array[p])/float(total_products)))
        # reset for a new word
        tmp_word = word
        tmp_Array = Array
```

```
Overwriting reduce3r5.py
```

In [243]:
```
%%writefile mappe3r5_s.py
#!/usr/bin/python
import sys

sys.stderr.write("reporter:counter:HW3_5,Mapper_s_counter,1\n")

for line in sys.stdin:
    # just emit
    print line.strip()
```

```
Overwriting mappe3r5_s.py
```

In [257]:
```
%%writefile reduce3r5_s.py
#!/usr/bin/python
import sys

sys.stderr.write("reporter:counter:HW3_5,Reducer_s_counter,1\n")

n = 0
top = 50

for line in sys.stdin:
    # parse mapper output
    n += 1
    if n <= top:
        print line.strip().replace(',', '\t')
```

```
Overwriting reduce3r5_s.py
```

In [221]:
```
# Generate Hadoop results without sorting
!chmod a+x mappe3r5.py
!chmod a+x reduce3r5.py
!hdfs dfs -rm -r result3s5
!hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
-D mapred.map.tasks=2 \
-D mapred.reduce.tasks=1 \
-mapper /home/cloudera/mappe3r5.py \
-reducer /home/cloudera/reduce3r5.py \
-input /user/shihyu/ProductPurchaseData.txt \
-output result3s5
```

```
Deleted result3s5
packageJobJar: [] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.6.0-cdh5.8.0.jar
] /tmp/streamjob1235062254260049591.jar tmpDir=null
16/09/14 22:44:52 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/14 22:44:52 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/14 22:44:53 INFO mapred.FileInputFormat: Total input paths to process : 1
16/09/14 22:44:53 INFO mapreduce.JobSubmitter: number of splits:2
16/09/14 22:44:53 INFO Configuration.deprecation: mapred.reduce.tasks is depreca
ted. Instead, use mapreduce.job.reduces
16/09/14 22:44:53 INFO Configuration.deprecation: mapred.map.tasks is deprecated
. Instead, use mapreduce.job.maps
16/09/14 22:44:54 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_14
73444507507_0133
16/09/14 22:44:54 INFO impl.YarnClientImpl: Submitted application application_14
73444507507_0133
16/09/14 22:44:54 INFO mapreduce.Job: The url to track the job: http://quickstar
t.cloudera:8088/proxy/application_1473444507507_0133/
16/09/14 22:44:54 INFO mapreduce.Job: Running job: job_1473444507507_0133
16/09/14 22:45:07 INFO mapreduce.Job: Job job_1473444507507_0133 running in uber
 mode : false
16/09/14 22:45:07 INFO mapreduce.Job:  map 0% reduce 0%
16/09/14 22:45:25 INFO mapreduce.Job:  map 28% reduce 0%
16/09/14 22:45:27 INFO mapreduce.Job:  map 61% reduce 0%
16/09/14 22:45:28 INFO mapreduce.Job:  map 83% reduce 0%
16/09/14 22:45:29 INFO mapreduce.Job:  map 100% reduce 0%
16/09/14 22:45:43 INFO mapreduce.Job:  map 100% reduce 94%
16/09/14 22:45:46 INFO mapreduce.Job:  map 100% reduce 100%
16/09/14 22:45:46 INFO mapreduce.Job: Job job_1473444507507_0133 completed succe
ssfully
16/09/14 22:45:46 INFO mapreduce.Job: Counters: 51
        File System Counters
                FILE: Number of bytes read=15864355
                FILE: Number of bytes written=32084353
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=3462851
                HDFS: Number of bytes written=29770
                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)=39154
                Total time spent by all reduces in occupied slots (ms)=14291
                Total time spent by all map tasks (ms)=39154
                Total time spent by all reduce tasks (ms)=14291
                Total vcore-seconds taken by all map tasks=39154
                Total vcore-seconds taken by all reduce tasks=14291
                Total megabyte-seconds taken by all map tasks=40093696
                Total megabyte-seconds taken by all reduce tasks=14633984
        Map-Reduce Framework
                Map input records=31101
                Map output records=48041
                Map output bytes=15749537
                Map output materialized bytes=15864361
                Input split bytes=238
                Combine input records=0
                Combine output records=0
                Reduce input groups=12012
```

In [255]:
```
!hdfs dfs -cat result3s5/* | head -4
```

```
DAI16732,FRO78087,106,0.00340825053857
DAI22177,DAI62779,129,0.00414777659882
DAI22534,DAI62779,123,0.00395485675702
DAI22896,GRO21487,114,0.00366547699431
cat: Unable to write to output stream.
```

In [258]:
```
### Begin sorting
!chmod a+x mappe3r5_s.py
!chmod a+x reduce3r5_s.py
!hdfs dfs -rm -r result3s5_sorted
!hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
-D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedCo
mparator \
-D map.output.key.field.separator=',' \
-D map.output.key.value.fields.spec=0-2:3- \
-D mapred.text.key.comparator.options='-k3,3nr -k1,1 -k2,2' \
-D mapred.map.tasks=2 \
-D mapred.reduce.tasks=1 \
-files mappe3r5_s.py,reduce3r5_s.py \
-mapper mappe3r5_s.py \
-reducer reduce3r5_s.py \
-input result3s5 \
-output result3s5_sorted
```

```
Deleted result3s5_sorted
16/09/15 07:04:49 INFO Configuration.deprecation: mapred.output.key.comparator.c
lass is deprecated. Instead, use mapreduce.job.output.key.comparator.class
16/09/15 07:04:49 INFO Configuration.deprecation: map.output.key.field.separator
 is deprecated. Instead, use mapreduce.map.output.key.field.separator
16/09/15 07:04:49 INFO Configuration.deprecation: map.output.key.value.fields.sp
ec is deprecated. Instead, use mapreduce.fieldsel.map.output.key.value.fields.sp
ec
16/09/15 07:04:49 INFO Configuration.deprecation: mapred.text.key.comparator.opt
ions is deprecated. Instead, use mapreduce.partition.keycomparator.options
16/09/15 07:04:49 INFO Configuration.deprecation: mapred.map.tasks is deprecated
. Instead, use mapreduce.job.maps
16/09/15 07:04:49 INFO Configuration.deprecation: mapred.reduce.tasks is depreca
ted. Instead, use mapreduce.job.reduces
packageJobJar: [] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.6.0-cdh5.8.0.jar
] /tmp/streamjob5655302219594596207.jar tmpDir=null
16/09/15 07:04:50 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/15 07:04:51 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/15 07:04:52 INFO mapred.FileInputFormat: Total input paths to process : 1
16/09/15 07:04:52 INFO mapreduce.JobSubmitter: number of splits:2
16/09/15 07:04:53 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_14
73444507507_0143
16/09/15 07:04:53 INFO impl.YarnClientImpl: Submitted application application_14
73444507507_0143
16/09/15 07:04:53 INFO mapreduce.Job: The url to track the job: http://quickstar
t.cloudera:8088/proxy/application_1473444507507_0143/
16/09/15 07:04:53 INFO mapreduce.Job: Running job: job_1473444507507_0143
16/09/15 07:05:04 INFO mapreduce.Job: Job job_1473444507507_0143 running in uber
 mode : false
16/09/15 07:05:04 INFO mapreduce.Job:  map 0% reduce 0%
16/09/15 07:05:15 INFO mapreduce.Job:  map 100% reduce 0%
16/09/15 07:05:27 INFO mapreduce.Job:  map 100% reduce 100%
16/09/15 07:05:27 INFO mapreduce.Job: Job job_1473444507507_0143 completed succe
ssfully
16/09/15 07:05:27 INFO mapreduce.Job: Counters: 51
        File System Counters
                FILE: Number of bytes read=31272
                FILE: Number of bytes written=424364
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=34102
                HDFS: Number of bytes written=1897
                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)=19212
                Total time spent by all reduces in occupied slots (ms)=8826
                Total time spent by all map tasks (ms)=19212
                Total time spent by all reduce tasks (ms)=8826
                Total vcore-seconds taken by all map tasks=19212
                Total vcore-seconds taken by all reduce tasks=8826
                Total megabyte-seconds taken by all map tasks=19673088
                Total megabyte-seconds taken by all reduce tasks=9037824
        Map-Reduce Framework
                Map input records=748
                Map output records=748
                Map output bytes=29770
```

In [259]:
```
! echo "50 Most Common Pairs of Products by Strip:"
! echo "Product Pairs| Frequency | Relative Frequency"
!hdfs dfs -cat result3s5_sorted/part-0*
```

```
50 Most Common Pairs of Products by Strip:
Product Pairs| Frequency | Relative Frequency
FRO40251        SNA80324        1412    0.0454004694383
DAI75645        FRO40251        1254    0.0403202469374
FRO40251        GRO85051        1213    0.0390019613517
DAI75645        SNA80324        1130    0.0363332368734
ELE32164        GRO59710        911     0.0292916626475
DAI62779        ELE17451        902     0.0290022828848
FRO40251        GRO73461        882     0.0283592167454
DAI62779        GRO73461        844     0.0271373910807
FRO40251        FRO92469        835     0.026848011318
DAI75645        GRO73461        712     0.0228931545609
DAI43223        ELE32164        711     0.022861001254
DAI62779        FRO40251        658     0.0211568759847
DAI62779        DAI85309        614     0.0197421304781
DAI62779        SNA80324        598     0.0192276775666
DAI62779        DAI75645        560     0.0180058519019
DAI75645        ELE17451        547     0.0175878589113
DAI55148        DAI62779        526     0.016912639465
DAI43223        GRO59710        512     0.0164624931674
ELE32164        GRO73461        486     0.0156265071863
DAI62779        SNA55762        463     0.014886981126
FRO85978        SNA95666        463     0.014886981126
DAI43223        DAI62779        459     0.0147583678981
ELE92920        SNA18336        455     0.0146297546703
ELE17451        FRO40251        453     0.0145654480563
DAI62779        FRO19221        449     0.0144368348285
DAI88079        FRO40251        446     0.0143403749076
FRO73056        GRO44993        438     0.0140831484518
ELE17451        SNA80324        428     0.0137616153821
GRO38814        GRO73461        427     0.0137294620752
ELE17451        GRO73461        409     0.0131507025498
DAI62779        ELE32164        406     0.0130542426289
DAI75645        GRO85051        395     0.0127005562522
FRO31317        GRO73461        395     0.0127005562522
DAI62779        SNA45677        392     0.0126040963313
DAI62779        GRO30386        389     0.0125076364104
GRO46854        GRO73461        389     0.0125076364104
GRO30386        GRO73461        380     0.0122182566477
FRO40251        GRO21487        375     0.0120574901129
DAI62779        ELE26917        371     0.011928876885
DAI62779        DAI83733        363     0.0116716504292
DAI62779        GRO71621        357     0.0114787305874
ELE74482        SNA99873        357     0.0114787305874
FRO92469        SNA80324        352     0.0113179640526
FRO85978        GRO73461        344     0.0110607375969
DAI55148        FRO40251        343     0.0110285842899
DAI55148        SNA80324        339     0.010899971062
DAI62779        FRO80039        327     0.0105141313784
DAI43223        ELE17451        326     0.0104819780714
ELE74482        FRO31317        317     0.0101925983087
DAI62779        ELE99737        315     0.0101282916948
```

HW3.5 Results

- 2 mappers, 1 reducer
- with the same configure, the execution time is reduced to 15 sec. from 25 sec. of pair approach, about 32% improvement

HW3.6 Computing Relative Frequencies on 100K WikiPedia pages (93Meg)

In [301]:
```
!hdfs dfs -mkdir -p /user/shihyu
!hdfs dfs -put wikitext_100k.txt /user/shihyu
# hdfs -cat \usr\cloudera\output\part-r-0000 >\somewhere\results.txt
```

```
put: `/user/shihyu/wikitext_100k.txt': File exists
```

Pairs Method

In [500]:
```
%%writefile mappe3r6_pair.py
#!/usr/bin/python
import sys
import re

# mapper counter
sys.stderr.write("reporter:counter:HW3_6_Pair, Mapper_counter,1\n")

WORD_RE = re.compile(r"[\w']+")

#cleanedHost = re.sub(r"[^a-zA-Z0-9]+", "", host)

for line in sys.stdin:


    try:
        # Get format
        author, bbb, ccc, ddd, eee, body = line.split('\t', -1)
    except ValueError:
        continue

    # get all words from the session (subject + ' ' + body)
    # for word in WORD_RE.findall(subject + ' ' + body):
    #clean_body = re.sub(r"[^a-zA-Z0-9]+", "", body)
    #  re.sub(r"[^A-Za-z]", " ", body.strip())
    #clean_eee = re.sub(r"[^a-zA-Z0-9]+", "", eee)
    # words = re.split(r"[^A-Za-z]", line.strip())

    ### workable
    #line = re.sub(r"[^A-Za-z]", " ", body.strip())
    #words = line.split()

    #clean_author = re.sub(r"[^A-Za-z]", " ", author.strip())
    #clean_bbb = re.sub(r"[^A-Za-z]", " ", bbb.strip())
    #clean_ccc = re.sub(r"[^A-Za-z]", " ", ccc.strip())
    #clean_ddd = re.sub(r"[^A-Za-z]", " ", ddd.strip())
    clean_eee = re.sub(r"[^A-Za-z]", " ", eee.strip())
    clean_body = re.sub(r"[^A-Za-z]", " ", body.strip())

    words = clean_body + ' ' + clean_eee# + ' ' + clean_ddd + ' ' + clean_ccc + '
' + clean_bbb
    words = words.split()

    #line = re.sub(r"[^A-Za-z]", " ", body.strip()) + ' ' + re.sub(r"[^A-Za-z]", "
 ", eee.strip())
    #words = line.split()


    cart_size = len(words)
    if cart_size==0:
        continue

    # sort words the pair is lexicographically sound
    words.sort()

    # set pairs of words
    pairs = [[words[i], words[j]] for i in range(cart_size) for j in range(i+1, ca
rt_size)]

    # dummy record for total products count
    print '%s,%s' %('*', 1)

    # emit product pairs
    for pair in pairs:
```

```
Overwriting mappe3r6_pair.py
```

In [501]:
```
%%writefile combine3r6_pair.py
#!/usr/bin/python
import sys

# combiner counter
sys.stderr.write("reporter:counter:HW3_6_Pair, Combiner_ounter,1\n")

tmp_pair = None
tmp_count = 0

for line in sys.stdin:
    # get all products from each line
    pair, count = line.strip().split(',', 1)

    # skip bad value
    try:
        count = int(count)
    except ValueError:
        continue

    # accumulate counts for whatever keys it receives
    if tmp_pair == pair:
        tmp_count += count
    else:
        # previous pair finishes streaming, emit results
        if tmp_pair:
            print '%s,%s' %(tmp_pair, tmp_count)
        # set new pair
        tmp_pair = pair
        tmp_count = count
```

```
Overwriting combine3r6_pair.py
```

In [502]:
```
%%writefile reduce3r6_pair.py
#!/usr/bin/python
import sys

# reducer counter
sys.stderr.write("reporter:counter:HW3_6_Pair, Reducer_counter,1\n")


total_words = 0
min_support = 100
tmp_pair = None
tmp_count = 0

for line in sys.stdin:
    # get all products from the session
    pair, count = line.strip().split(',', 1)

    # skip bad count
    try:
        count = int(count)
    except ValueError:
        continue

    # get total sessions/baskets
    if pair == '*':
        total_words += count
        continue

    # get pair count
    if tmp_pair == pair:
        tmp_count += count
    else:
        # previous pair finishes
        if tmp_pair and tmp_count > min_support:
            # emit
            print '%s,%s,%s' %(tmp_pair, tmp_count, str(float(tmp_count)/float(tot
al_words)))
        # reset new pair
        tmp_pair = pair
        tmp_count = count
```

```
Overwriting reduce3r6_pair.py
```

In [503]:
```
# Generate Hadoop results without sorting
!chmod a+x mappe3r6_pair.py
!chmod a+x combine3r6_pair.py
!chmod a+x reduce3r6_pair.py
!hdfs dfs -rm -r result3s6_pair
!hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
-D mapred.map.tasks=2 \
-D mapred.reduce.tasks=1 \
-mapper /home/cloudera/mappe3r6_pair.py \
-combiner /home/cloudera/combine3r6_pair.py \
-reducer /home/cloudera/reduce3r6_pair.py \
-input /user/shihyu/wikitext_100k.txt \
-output result3s6_pair

#!hdfs -cat /user/cloudera/result3s6_pair/part-00000 > /user/cloudera/rfpairs.txt
```

```
Deleted result3s6_pair
packageJobJar: [] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.6.0-cdh5.8.0.jar
] /tmp/streamjob6908161091078244484.jar tmpDir=null
16/09/17 14:25:09 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/17 14:25:12 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/17 14:25:15 INFO mapred.FileInputFormat: Total input paths to process : 1
16/09/17 14:25:16 INFO mapreduce.JobSubmitter: number of splits:2
16/09/17 14:25:16 INFO Configuration.deprecation: mapred.reduce.tasks is depreca
ted. Instead, use mapreduce.job.reduces
16/09/17 14:25:16 INFO Configuration.deprecation: mapred.map.tasks is deprecated
. Instead, use mapreduce.job.maps
16/09/17 14:25:17 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_14
73444507507_0211
16/09/17 14:25:18 INFO impl.YarnClientImpl: Submitted application application_14
73444507507_0211
16/09/17 14:25:19 INFO mapreduce.Job: The url to track the job: http://quickstar
t.cloudera:8088/proxy/application_1473444507507_0211/
16/09/17 14:25:19 INFO mapreduce.Job: Running job: job_1473444507507_0211
16/09/17 14:27:42 INFO mapreduce.Job: Job job_1473444507507_0211 running in uber
 mode : false
16/09/17 14:27:42 INFO mapreduce.Job:  map 0% reduce 0%
16/09/17 14:28:09 INFO mapreduce.Job:  map 1% reduce 0%
16/09/17 14:29:18 INFO mapreduce.Job:  map 4% reduce 0%
16/09/17 14:29:24 INFO mapreduce.Job:  map 29% reduce 0%
16/09/17 14:29:46 INFO mapreduce.Job:  map 37% reduce 0%
16/09/17 14:29:59 INFO mapreduce.Job:  map 66% reduce 0%
16/09/17 14:30:02 INFO mapreduce.Job:  map 67% reduce 0%
16/09/17 14:30:23 INFO mapreduce.Job:  map 83% reduce 0%
16/09/17 14:31:05 INFO mapreduce.Job:  map 87% reduce 0%
16/09/17 14:31:09 INFO mapreduce.Job:  map 90% reduce 0%
16/09/17 14:31:12 INFO mapreduce.Job:  map 96% reduce 0%
16/09/17 14:31:14 INFO mapreduce.Job:  map 100% reduce 0%
16/09/17 14:31:40 INFO mapreduce.Job:  map 100% reduce 67%
16/09/17 14:31:44 INFO mapreduce.Job:  map 100% reduce 79%
16/09/17 14:31:47 INFO mapreduce.Job:  map 100% reduce 87%
16/09/17 14:31:50 INFO mapreduce.Job:  map 100% reduce 100%
16/09/17 14:31:52 INFO mapreduce.Job: Job job_1473444507507_0211 completed succe
ssfully
16/09/17 14:31:52 INFO mapreduce.Job: Counters: 53
        File System Counters
                FILE: Number of bytes read=23902990
                FILE: Number of bytes written=37599996
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=92446758
                HDFS: Number of bytes written=237386
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=1
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=295395
                Total time spent by all reduces in occupied slots (ms)=35601
                Total time spent by all map tasks (ms)=295395
                Total time spent by all reduce tasks (ms)=35601
                Total vcore-seconds taken by all map tasks=295395
                Total vcore-seconds taken by all reduce tasks=35601
                Total megabyte-seconds taken by all map tasks=302484480
```

In [504]:
```
!hdfs dfs -cat result3s6_pair/* | head -4
```

```
ABC_the,103,5.72222222222
ALTERNATIVE_the,103,5.72222222222
AL_the,103,5.72222222222
AZ_Gallery,108,6.0
cat: Unable to write to output stream.
```

In [505]:
```
### Begin sorting
!chmod a+x mappe3r4_s.py
!chmod a+x reduce3r4_s.py
!hdfs dfs -rm -r result3s6_pair_sorted
!hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
-D mapred.map.tasks=2 \
-D mapred.reduce.tasks=1 \
-D stream.num.map.output.key.fields=2 \
-D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedCo
mparator \
-D map.output.key.field.separator=',' \
-D map.output.key.value.fields.spec=0-1:2- \
-D mapred.text.key.comparator.options='-k2,2nr -k1,1' \
-mapper /home/cloudera/mappe3r4_s.py \
-reducer /home/cloudera/reduce3r4_s.py \
-input result3s6_pair \
-output result3s6_pair_sorted

!hdfs dfs -cat result3s6_pair_sorted/* | head -4
```

```
Deleted result3s6_pair_sorted
packageJobJar: [] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.6.0-cdh5.8.0.jar
] /tmp/streamjob6382017384913746716.jar tmpDir=null
16/09/17 14:37:28 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/17 14:37:29 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/17 14:37:31 INFO mapred.FileInputFormat: Total input paths to process : 1
16/09/17 14:37:31 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1281)
        at java.lang.Thread.join(Thread.java:1355)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DF
SOutputStream.java:862)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutpu
tStream.java:600)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStre
am.java:789)
16/09/17 14:37:31 INFO mapreduce.JobSubmitter: number of splits:2
16/09/17 14:37:31 INFO Configuration.deprecation: mapred.reduce.tasks is depreca
ted. Instead, use mapreduce.job.reduces
16/09/17 14:37:31 INFO Configuration.deprecation: mapred.output.key.comparator.c
lass is deprecated. Instead, use mapreduce.job.output.key.comparator.class
16/09/17 14:37:31 INFO Configuration.deprecation: mapred.text.key.comparator.opt
ions is deprecated. Instead, use mapreduce.partition.keycomparator.options
16/09/17 14:37:31 INFO Configuration.deprecation: map.output.key.field.separator
 is deprecated. Instead, use mapreduce.map.output.key.field.separator
16/09/17 14:37:31 INFO Configuration.deprecation: map.output.key.value.fields.sp
ec is deprecated. Instead, use mapreduce.fieldsel.map.output.key.value.fields.sp
ec
16/09/17 14:37:31 INFO Configuration.deprecation: mapred.map.tasks is deprecated
. Instead, use mapreduce.job.maps
16/09/17 14:37:31 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_14
73444507507_0212
16/09/17 14:37:32 INFO impl.YarnClientImpl: Submitted application application_14
73444507507_0212
16/09/17 14:37:32 INFO mapreduce.Job: The url to track the job: http://quickstar
t.cloudera:8088/proxy/application_1473444507507_0212/
16/09/17 14:37:32 INFO mapreduce.Job: Running job: job_1473444507507_0212
16/09/17 14:37:46 INFO mapreduce.Job: Job job_1473444507507_0212 running in uber
 mode : false
16/09/17 14:37:46 INFO mapreduce.Job:  map 0% reduce 0%
16/09/17 14:38:03 INFO mapreduce.Job:  map 50% reduce 0%
16/09/17 14:38:04 INFO mapreduce.Job:  map 100% reduce 0%
16/09/17 14:38:14 INFO mapreduce.Job:  map 100% reduce 100%
16/09/17 14:38:15 INFO mapreduce.Job: Job job_1473444507507_0212 completed succe
ssfully
16/09/17 14:38:15 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=254040
                FILE: Number of bytes written=866024
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=241728
                HDFS: Number of bytes written=1588
                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
```

In [412]:
```
!hdfs -cat /user/cloudera/result3s6_pair_sorted/* > /user/cloudera/rfpairs.txt
```

```
/bin/sh: /user/cloudera/rfpairs.txt: No such file or directory
```

Strip Method

In [ ]:
```
%%writefile mappe3r6_Strip.py
#!/usr/bin/python
import sys

# mapper counter
sys.stderr.write("reporter:counter:HW3_6_Strip, Mapper_counter,1\n")

# Associative array
A = {}

for line in sys.stdin:
    # get all products from the session
    products = line.strip().split(' ')
    product_size = len(products)
    if product_size==0:
        continue

    #  lexicographically sort
    products.sort()

    # get pairs of products
    pairs = [[products[i], products[j]] for i in range(product_size) for j in range
(i+1, product_size)]

    # emit dummy record for total count
    print '%s\t%s' %('*', 1)

    # prepare associative arrays
    for w1, w2 in pairs:

        if w1 not in A:
            # if w1 is new, add to associative array
            A[w1] = {}
            A[w1][w2] = 1
        elif w2 not in A[w1]:
            # w1 is not new, but it doesn't have key for w2
            A[w1][w2] = 1
        else:
            # both are there, increase it
            A[w1][w2] += 1

# emit associative arrays
for a in A:
    print '%s\t%s' %(a, str(A[a]))
```

In [ ]:

In [ ]:

In [ ]:

Comparison for Pairs and Strip Method:

HW3.7 Apriori Algorithm

Answer:

- Aprior algorithm is used to find itemsets that appear frequently, each iteration has two scans of data and a filtering in between. They have following steps:
  1. generate a set S_k for itemsets with size k from the output of previous iteration M_{k-1}.
  2. remove all members from the set whose support is less than the user specified threshold t_i
  3. generate the final set M_k for itemset of size k, based on output after filtering.
- For example, to find itemsets of size k from a shopping basket set, the procedure is described as :
  1. count all single product from all baskets, output S_1
  2. remove all words with support below threshold, output M_1
  3. use M_1 to generate set for frequent pair set S_2
  4. remove all pairs with support below threshold to obtain M_2
  5. use M_2 to generate set for frequent triple set S_3
  6. remove all triples with support below threshold, get M_3
  7. Cotinue above to M_k.

HW3.8. Shopping Cart Analysis, Benchmark your results using the pyFIM implementation of the Apriori algorithm

In [260]:
```
%%writefile mappe3r8_1.py
#!/usr/bin/python
import sys


for line in sys.stdin:
    # get products and emit
    for product in line.strip().split(' '):
        print '%s\t%d' %(product, 1)
```

```
Writing mappe3r8_1.py
```

In [261]:
```
%%writefile reduce3r8_1.py
#!/usr/bin/python
import sys


tmp_prod = None
tmp_count = 0
min_support = 100

for line in sys.stdin:
    # get key value pair
    product, count = line.strip().split('\t', 1)

    # skip bad count
    try:
        count = int(count)
    except ValueError:
        continue

    # get count
    if tmp_prod == product:
        tmp_count += count
    else:
        if tmp_prod and tmp_count > min_support:
            # emit product above min support
            print '%s\t%d' %(tmp_prod, tmp_count)
        # reset product and count
        tmp_prod = product
        tmp_count = count
```

```
Writing reduce3r8_1.py
```

In [270]:
```
%%writefile mappe3r8_2.py
#!/usr/bin/python
import sys, subprocess


single = []
cat = subprocess.Popen(["hadoop", "fs", "-cat", "/user/cloudera/result3s8_1/part-0
0000"], stdout=subprocess.PIPE)
for line in cat.stdout:
    single.append(line.strip().split('\t')[0])

# read the input data
for line in sys.stdin:

    line = line.strip()

    # get products for each cart
    product = line.strip().split(' ')

    # keep product from set with single element only
    products = [val for val in product if val in single]
    products.sort()

    # get pairs to emit
    size = len(products)
    pairs = [products[i] + '_' + products[j] for i in range(size) for j in range(i
+1, size)]
    for p in pairs:
        print '%s\t%d' %(p, 1)
```

```
Overwriting mappe3r8_2.py
```

Second stage reducer is reduce3r8_1.py also.

In [273]:
```
%%writefile mappe3r8_3.py
#!/usr/bin/python
import sys, subprocess

# load the frequent frequent Pairs given by Job 2
Freq_Pair = []
cat = subprocess.Popen(["hadoop", "fs", "-cat", "/user/cloudera/result3s8_2/part-0
0000"], stdout=subprocess.PIPE)
for line in cat.stdout:
    Freq_Pair.append(line.strip().split('\t')[0])

# still read frequent freqPairs first, then session data to generate triples
for line in sys.stdin:

    line = line.strip()

    # Get product from each cart
    products = line.split(' ')
    products.sort()
    size = len(products)

    # build Pairs and Triples from the session, in the format of a_b and a_b_c, al
phabetically sorted
    triples = [[products[i],products[j],products[k]] for i in range(size) for j in
 range(i+1,size) for k in range(i+2,size)]
    pairs = [products[i]+'_'+products[j] for i in range(size) for j in range(i+1,s
ize)]

    # processing pairs
    for pair in pairs:
        # if the pair is in frequent Pair, emit a dummy key a_b_*
        if pair in Freq_Pair:
            print '%s_*\t%d' %(pair, 1)

    # processing triples
    for tri in triples:
        # from each triple a_b_c: check if the 3 child-pairs (a_b, b_c, a_c) are i
n the pair set
        # If yes, it is associative rule
        if tri[0]+'_'+tri[1] in Freq_Pair and tri[1]+'_'+tri[2] in Freq_Pair and t
ri[0]+'_'+tri[2] in Freq_Pair:
            # if so, emit the triple a_b_c
            print '%s_%s_%s\t%d' %(tri[0], tri[1], tri[2], 1)
```

```
Overwriting mappe3r8_3.py
```

In [280]:
```
%%writefile reduce3r8_3.py
#!/usr/bin/python
import sys

tmp_prod = None
tmp_dummy = None
tmp_count = 0
min_support = 100
marginal = 0

for line in sys.stdin:

    # get k-v freqPair
    product, count = line.strip().split('\t', 1)

    # skip bad count
    try:
        count = int(count)
    except ValueError:
        continue

    # handle marginal with dummy key
    if '*' == product[-1]:
        if tmp_dummy == product:
            # accumulate marginal
            marginal += count
        else:
            # reset marginal for new dummy key
            tmp_dummy = product
            marginal = count
        continue

    # processing triple and emit rules
    if tmp_prod == product:
        tmp_count += count
    else:
        if tmp_prod and tmp_count > min_support and tmp_count <= marginal: # Remov
ing some mismatch
            # emit triples for the rule
            w1,w2,w3 = tmp_prod.split('_')
            conf = float(tmp_count)/float(marginal)
            print '(%s, %s) => %s, %d, %d, %.2f' %(w1, w2, w3, tmp_count, marginal
, conf)

        # reset for new triple
        tmp_prod = product
        tmp_count = count
```

```
Overwriting reduce3r8_3.py
```

In [265]:
```
# job 1 - get M_1 for frequent singletons
!chmod a+x mappe3r8_1.py
!chmod a+x reduce3r8_1.py
!hdfs dfs -rm -r result3s8_1
!hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
-D mapred.map.tasks=3 \
-D mapred.reduce.tasks=1 \
-files mappe3r8_1.py,reduce3r8_1.py \
-mapper mappe3r8_1.py \
-reducer reduce3r8_1.py \
-combiner reduce3r8_1.py \
-input /user/shihyu/ProductPurchaseData.txt \
-output result3s8_1
```

```
rm: `result3s8_1': No such file or directory
16/09/15 11:37:24 INFO Configuration.deprecation: mapred.map.tasks is deprecated
. Instead, use mapreduce.job.maps
16/09/15 11:37:24 INFO Configuration.deprecation: mapred.reduce.tasks is depreca
ted. Instead, use mapreduce.job.reduces
packageJobJar: [] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.6.0-cdh5.8.0.jar
] /tmp/streamjob7347751964987250476.jar tmpDir=null
16/09/15 11:37:25 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/15 11:37:25 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/15 11:37:26 INFO mapred.FileInputFormat: Total input paths to process : 1
16/09/15 11:37:26 INFO mapreduce.JobSubmitter: number of splits:3
16/09/15 11:37:27 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_14
73444507507_0144
16/09/15 11:37:27 INFO impl.YarnClientImpl: Submitted application application_14
73444507507_0144
16/09/15 11:37:27 INFO mapreduce.Job: The url to track the job: http://quickstar
t.cloudera:8088/proxy/application_1473444507507_0144/
16/09/15 11:37:27 INFO mapreduce.Job: Running job: job_1473444507507_0144
16/09/15 11:37:37 INFO mapreduce.Job: Job job_1473444507507_0144 running in uber
 mode : false
16/09/15 11:37:37 INFO mapreduce.Job:  map 0% reduce 0%
16/09/15 11:37:57 INFO mapreduce.Job:  map 33% reduce 0%
16/09/15 11:37:58 INFO mapreduce.Job:  map 67% reduce 0%
16/09/15 11:38:00 INFO mapreduce.Job:  map 100% reduce 0%
16/09/15 11:38:06 INFO mapreduce.Job:  map 100% reduce 100%
16/09/15 11:38:07 INFO mapreduce.Job: Job job_1473444507507_0144 completed succe
ssfully
16/09/15 11:38:07 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=11017
                FILE: Number of bytes written=502999
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=3467066
                HDFS: Number of bytes written=4791
                HDFS: Number of read operations=12
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=1
                Launched map tasks=3
                Launched reduce tasks=1
                Data-local map tasks=3
                Total time spent by all maps in occupied slots (ms)=56193
                Total time spent by all reduces in occupied slots (ms)=6622
                Total time spent by all map tasks (ms)=56193
                Total time spent by all reduce tasks (ms)=6622
                Total vcore-seconds taken by all map tasks=56193
                Total vcore-seconds taken by all reduce tasks=6622
                Total megabyte-seconds taken by all map tasks=57541632
                Total megabyte-seconds taken by all reduce tasks=6780928
        Map-Reduce Framework
                Map input records=31101
                Map output records=380824
                Map output bytes=4189064
                Map output materialized bytes=11029
                Input split bytes=357
                Combine input records=380824
                Combine output records=733
                Reduce input groups=365
                Reduce shuffle bytes=11029
```

In [271]:
```
# job 2 - get M_2 for frequent pairs
!chmod a+x mappe3r8_2.py
!chmod a+x reduce3r8_1.py
!hdfs dfs -rm -r result3s8_2
!hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
-D mapred.map.tasks=3 \
-D mapred.reduce.tasks=1 \
-files mappe3r8_2.py,reduce3r8_1.py \
-mapper mappe3r8_2.py \
-reducer reduce3r8_1.py \
-input /user/shihyu/ProductPurchaseData.txt \
-output result3s8_2
```

```
Deleted result3s8_2
16/09/15 11:57:06 INFO Configuration.deprecation: mapred.map.tasks is deprecated
. Instead, use mapreduce.job.maps
16/09/15 11:57:06 INFO Configuration.deprecation: mapred.reduce.tasks is depreca
ted. Instead, use mapreduce.job.reduces
packageJobJar: [] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.6.0-cdh5.8.0.jar
] /tmp/streamjob5428113389080488872.jar tmpDir=null
16/09/15 11:57:07 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/15 11:57:08 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/15 11:57:09 INFO mapred.FileInputFormat: Total input paths to process : 1
16/09/15 11:57:09 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1281)
        at java.lang.Thread.join(Thread.java:1355)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DF
SOutputStream.java:862)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutpu
tStream.java:600)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStre
am.java:789)
16/09/15 11:57:09 INFO mapreduce.JobSubmitter: number of splits:3
16/09/15 11:57:09 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_14
73444507507_0146
16/09/15 11:57:09 INFO impl.YarnClientImpl: Submitted application application_14
73444507507_0146
16/09/15 11:57:10 INFO mapreduce.Job: The url to track the job: http://quickstar
t.cloudera:8088/proxy/application_1473444507507_0146/
16/09/15 11:57:10 INFO mapreduce.Job: Running job: job_1473444507507_0146
16/09/15 11:57:21 INFO mapreduce.Job: Job job_1473444507507_0146 running in uber
 mode : false
16/09/15 11:57:21 INFO mapreduce.Job:  map 0% reduce 0%
16/09/15 11:57:43 INFO mapreduce.Job:  map 5% reduce 0%
16/09/15 11:57:45 INFO mapreduce.Job:  map 8% reduce 0%
16/09/15 11:57:51 INFO mapreduce.Job:  map 13% reduce 0%
16/09/15 11:57:52 INFO mapreduce.Job:  map 23% reduce 0%
16/09/15 11:57:55 INFO mapreduce.Job:  map 67% reduce 0%
16/09/15 11:57:56 INFO mapreduce.Job:  map 78% reduce 0%
16/09/15 11:57:58 INFO mapreduce.Job:  map 100% reduce 0%
16/09/15 11:58:09 INFO mapreduce.Job:  map 100% reduce 100%
16/09/15 11:58:10 INFO mapreduce.Job: Job job_1473444507507_0146 completed succe
ssfully
16/09/15 11:58:10 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=21377208
                FILE: Number of bytes written=43234113
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=3467066
                HDFS: Number of bytes written=27705
                HDFS: Number of read operations=12
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=3
                Launched reduce tasks=1
                Data-local map tasks=3
                Total time spent by all maps in occupied slots (ms)=99729
                Total time spent by all reduces in occupied slots (ms)=11827
                Total time spent by all map tasks (ms)=99729
                Total time spent by all reduce tasks (ms)=11827
```

In [281]:
```
# job 3 - get M_3 for frequent triples
!chmod a+x mappe3r8_3.py
!chmod a+x reduce3r8_3.py
!hdfs dfs -rm -r result3s8_3
!hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
-D mapred.map.tasks=3 \
-D mapred.reduce.tasks=1 \
-files mappe3r8_3.py,reduce3r8_3.py \
-mapper mappe3r8_3.py \
-reducer reduce3r8_3.py \
-input /user/shihyu/ProductPurchaseData.txt \
-output result3s8_3
```

```
Deleted result3s8_3
16/09/15 12:45:49 INFO Configuration.deprecation: mapred.map.tasks is deprecated
. Instead, use mapreduce.job.maps
16/09/15 12:45:49 INFO Configuration.deprecation: mapred.reduce.tasks is depreca
ted. Instead, use mapreduce.job.reduces
packageJobJar: [] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.6.0-cdh5.8.0.jar
] /tmp/streamjob1755665626735376634.jar tmpDir=null
16/09/15 12:45:51 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/15 12:45:51 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/15 12:45:52 INFO mapred.FileInputFormat: Total input paths to process : 1
16/09/15 12:45:52 INFO mapreduce.JobSubmitter: number of splits:3
16/09/15 12:45:53 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_14
73444507507_0149
16/09/15 12:45:53 INFO impl.YarnClientImpl: Submitted application application_14
73444507507_0149
16/09/15 12:45:53 INFO mapreduce.Job: The url to track the job: http://quickstar
t.cloudera:8088/proxy/application_1473444507507_0149/
16/09/15 12:45:53 INFO mapreduce.Job: Running job: job_1473444507507_0149
16/09/15 12:46:04 INFO mapreduce.Job: Job job_1473444507507_0149 running in uber
 mode : false
16/09/15 12:46:04 INFO mapreduce.Job:  map 0% reduce 0%
16/09/15 12:46:23 INFO mapreduce.Job:  map 3% reduce 0%
16/09/15 12:46:25 INFO mapreduce.Job:  map 5% reduce 0%
16/09/15 12:46:26 INFO mapreduce.Job:  map 8% reduce 0%
16/09/15 12:47:00 INFO mapreduce.Job:  map 10% reduce 0%
16/09/15 12:47:15 INFO mapreduce.Job:  map 13% reduce 0%
16/09/15 12:47:22 INFO mapreduce.Job:  map 15% reduce 0%
16/09/15 12:48:17 INFO mapreduce.Job:  map 18% reduce 0%
16/09/15 12:48:28 INFO mapreduce.Job:  map 20% reduce 0%
16/09/15 12:49:24 INFO mapreduce.Job:  map 23% reduce 0%
16/09/15 12:49:35 INFO mapreduce.Job:  map 25% reduce 0%
16/09/15 12:49:46 INFO mapreduce.Job:  map 28% reduce 0%
16/09/15 12:50:31 INFO mapreduce.Job:  map 30% reduce 0%
16/09/15 12:50:55 INFO mapreduce.Job:  map 33% reduce 0%
16/09/15 12:51:13 INFO mapreduce.Job:  map 35% reduce 0%
16/09/15 12:51:16 INFO mapreduce.Job:  map 38% reduce 0%
16/09/15 12:51:23 INFO mapreduce.Job:  map 40% reduce 0%
16/09/15 12:51:53 INFO mapreduce.Job:  map 43% reduce 0%
16/09/15 12:52:33 INFO mapreduce.Job:  map 45% reduce 0%
16/09/15 12:52:51 INFO mapreduce.Job:  map 48% reduce 0%
16/09/15 12:53:13 INFO mapreduce.Job:  map 50% reduce 0%
16/09/15 12:53:33 INFO mapreduce.Job:  map 53% reduce 0%
16/09/15 12:54:41 INFO mapreduce.Job:  map 55% reduce 0%
16/09/15 12:54:47 INFO mapreduce.Job:  map 58% reduce 0%
16/09/15 12:54:49 INFO mapreduce.Job:  map 69% reduce 0%
16/09/15 12:55:09 INFO mapreduce.Job:  map 69% reduce 11%
16/09/15 12:55:46 INFO mapreduce.Job:  map 71% reduce 11%
16/09/15 12:55:55 INFO mapreduce.Job:  map 74% reduce 11%
16/09/15 12:56:28 INFO mapreduce.Job:  map 76% reduce 11%
16/09/15 12:56:58 INFO mapreduce.Job:  map 78% reduce 11%
16/09/15 12:57:41 INFO mapreduce.Job:  map 89% reduce 11%
16/09/15 12:57:42 INFO mapreduce.Job:  map 89% reduce 22%
16/09/15 12:58:19 INFO mapreduce.Job:  map 100% reduce 22%
16/09/15 12:58:22 INFO mapreduce.Job:  map 100% reduce 85%
16/09/15 12:58:23 INFO mapreduce.Job:  map 100% reduce 100%
16/09/15 12:58:23 INFO mapreduce.Job: Job job_1473444507507_0149 completed succe
ssfully
16/09/15 12:58:23 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=11884987
                FILE: Number of bytes written=24249679
                FILE: Number of read operations=0
```

In [282]:
```
!hdfs dfs -cat result3s8_3/* | head -4
```

```
(DAI22896, DAI62779) => GRO73461, 101, 297, 0.34
(DAI31081, DAI62779) => ELE17451, 103, 364, 0.28
(DAI31081, DAI75645) => FRO40251, 122, 206, 0.59
(DAI31081, ELE32164) => GRO59710, 112, 312, 0.36
cat: Unable to write to output stream.
```

In [299]:
```
### Begin sorting
!chmod a+x break_tie.py
!hdfs dfs -rm -r result3s8_3_sorted
!hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
-D mapred.map.tasks=2 \
-D mapred.reduce.tasks=1 \
-D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedCo
mparator \
-D map.output.key.field.separator=',' \
-D map.output.key.value.fields.spec=0-2:3- \
-D mapred.text.key.comparator.options='-k5,5r' \
-mapper /home/cloudera/mappe3r5_s.py \
-reducer /home/cloudera/reduce3r5_s.py \
-input result3s8_3 \
-output result3s8_3_sorted

!hdfs dfs -cat result3s8_3_sorted/* | head -10
```

```
Deleted result3s8_3_sorted
packageJobJar: [] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.6.0-cdh5.8.0.jar
] /tmp/streamjob5506711815659657084.jar tmpDir=null
16/09/15 13:50:12 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/15 13:50:13 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/09/15 13:50:15 INFO mapred.FileInputFormat: Total input paths to process : 1
16/09/15 13:50:15 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1281)
        at java.lang.Thread.join(Thread.java:1355)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DF
SOutputStream.java:862)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutpu
tStream.java:600)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStre
am.java:789)
16/09/15 13:50:15 INFO mapreduce.JobSubmitter: number of splits:2
16/09/15 13:50:15 INFO Configuration.deprecation: mapred.reduce.tasks is depreca
ted. Instead, use mapreduce.job.reduces
16/09/15 13:50:15 INFO Configuration.deprecation: mapred.output.key.comparator.c
lass is deprecated. Instead, use mapreduce.job.output.key.comparator.class
16/09/15 13:50:15 INFO Configuration.deprecation: mapred.text.key.comparator.opt
ions is deprecated. Instead, use mapreduce.partition.keycomparator.options
16/09/15 13:50:15 INFO Configuration.deprecation: map.output.key.field.separator
 is deprecated. Instead, use mapreduce.map.output.key.field.separator
16/09/15 13:50:15 INFO Configuration.deprecation: map.output.key.value.fields.sp
ec is deprecated. Instead, use mapreduce.fieldsel.map.output.key.value.fields.sp
ec
16/09/15 13:50:15 INFO Configuration.deprecation: mapred.map.tasks is deprecated
. Instead, use mapreduce.job.maps
16/09/15 13:50:16 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_14
73444507507_0165
16/09/15 13:50:16 INFO impl.YarnClientImpl: Submitted application application_14
73444507507_0165
16/09/15 13:50:17 INFO mapreduce.Job: The url to track the job: http://quickstar
t.cloudera:8088/proxy/application_1473444507507_0165/
16/09/15 13:50:17 INFO mapreduce.Job: Running job: job_1473444507507_0165
16/09/15 13:50:28 INFO mapreduce.Job: Job job_1473444507507_0165 running in uber
 mode : false
16/09/15 13:50:28 INFO mapreduce.Job:  map 0% reduce 0%
16/09/15 13:50:38 INFO mapreduce.Job:  map 50% reduce 0%
16/09/15 13:50:39 INFO mapreduce.Job:  map 100% reduce 0%
16/09/15 13:50:47 INFO mapreduce.Job:  map 100% reduce 100%
16/09/15 13:50:48 INFO mapreduce.Job: Job job_1473444507507_0165 completed succe
ssfully
16/09/15 13:50:49 INFO mapreduce.Job: Counters: 51
        File System Counters
                FILE: Number of bytes read=11549
                FILE: Number of bytes written=380568
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=15437
                HDFS: Number of bytes written=2450
                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
```

In [ ]: