

# MIDS W261 Spring 2016 Homework Week 3

**Ted Dunmire** glenn.dunmire.iv@gmail.com

**Filip Krunic** fkrunic@ischool.berkeley.edu

**Ron Cordell** ron.cordell@ischool.berkeley.edu

W261-4

January 26, 2016

## HW3.0.

### **What is a merge sort? Where is it used in Hadoop?**

A merge sort merges sorted lists into a single sorted list. The merge sort works by establishing a pointer to the beginning of each sorted list as well as a new "merge" list. The objects or keys in each list referenced by the pointers are compared and the chosen one moved or copied to the location indicated by the pointer of the merge list. The merge list pointer is advanced as is the pointer for the list from which the object was moved. This is repeated until all objects in all list have been moved or copied to the new merge list. The comparator function "chooses" the object from the source lists based on the rules coded into the comparator function such as the largest, the smallest, etc.

Hadoop uses a merge sort during the shuffle process when it takes output from multiple sources such as mappers or combiners and merges them into the sorted streams used by downstream processes.

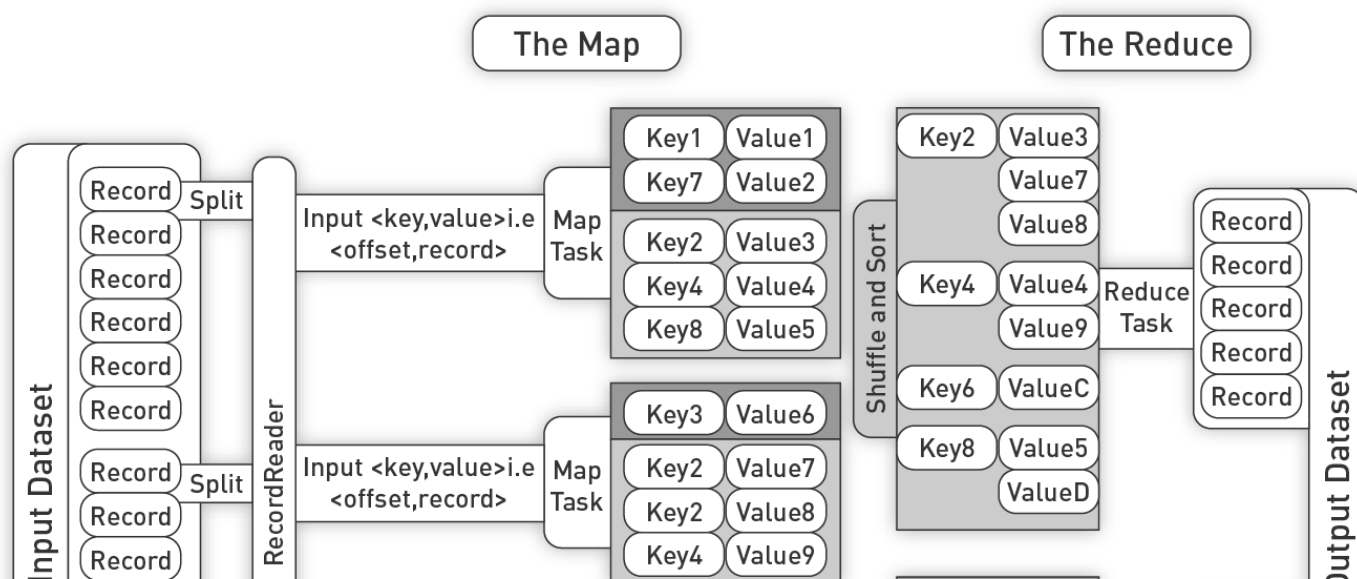
### **How is a combiner function in the context of Hadoop? Give an example where it can be used and justify why it should be used in the context of this problem.**

A combiner function is a function that can be used by Hadoop anywhere between the mappers and producers to help eliminate network and data traffic, especially as part of the shuffle. An example combiner function typically provides a partial aggregation point for data emitted from the mapper to reduce hotspots in the shuffle.

An example where a combiner can be used to good effect is in a word count scenario, where the mapper emits the word as key and a value of 1. A combiner can perform aggregations on the key-value pairs by combining those with the same key and adding their values. This greatly reduces the granularity of the data required to shuffle and sort and provide to the reducers and helps reduce the amount of network and disk traffic of the shuffle, decreasing the overall run time.

### **What is the Hadoop shuffle?**

## MapReduce Workflow



## HW3.1 Use Counters to do EDA (exploratory data analysis and to monitor progress)\*\*

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

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

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

```
Complaint ID,Product,Sub-product,Issue,Sub-issue,State,ZIP code,Submitted via,Date received,Date sent to company,Company,Company response,Time ly response?,Consumer disputed?
1114245,Debt collection,Medical,Disclosure verification of debt,Not giv en enough info to verify debt,FL,32219,Web,11/13/2014,11/13/2014,"Choic e Recovery, Inc.",Closed with explanation,Yes,
1114488,Debt collection,Medical,Disclosure verification of debt,Right t o dispute notice not received,TX,75006,Web,11/13/2014,11/13/2014,"Exper t Global Solutions, Inc.",In progress,Yes,
1114255,Bank account or service,Checking account,Deposits and withdrawa ls,,NY,11102,Web,11/13/2014,11/13/2014,"FNIS (Fidelity National Informa tion Services, Inc.)",In progress,Yes,
1115106,Debt collection,"Other (phone, health club, etc.)",Communicatio n tactics,Frequent or repeated calls,GA,31721,Web,11/13/2014,11/13/201 4,"Expert Global Solutions, Inc.",In progress,Yes,
```

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

## HW3.1 Map Function

```
In [1]: %%writefile mapper.py
#!/usr/bin/python
import sys

line_num = 0
for line in sys.stdin: #read input
    if line_num == 0:
        line_num += 1 #skip the first line, which is a header
        continue
    else:
        line = line.strip() #remove extra chars
        line = line.rstrip()
        line = line.split(',') #split on comma delimiter
        if line[1] == "Debt collection": #line[1] is the issue part
of the complaint
            sys.stderr.write('reporter:counter:Debt-Counter,Total,1
\n') #increment counter based on complaint
            elif line[1] == 'Mortgage':
                sys.stderr.write('reporter:counter:Mortgage-Counter,Tot
al,1\n')
            else:
                sys.stderr.write('reporter:counter:Other-Counter,Total,
1\n') #all other issues are lumped together
                sys.stderr.write("reporter:counter:Tokens,Total,1\n")
                print line[1] + '\t' + '1' #This just prints the issue and
a 1. Will use to check if counters are correct
```

Overwriting mapper.py

## HW3.1 Reduce Function

```
In [2]: %%writefile reducer.py
#!/usr/bin/python
import sys

#keep counters to see how many results we have
debt_counter = 0
mortgage_counter = 0
other_counter = 0

for line in sys.stdin: #read input
    line = line.strip().split('\t')
    if line[0] == "Debt collection": #recall we passed the issue as
line[0] from the mapper, here we parse
        debt_counter +=1
    elif line[0] == 'Mortgage':
        mortgage_counter += 1
    else:
        other_counter += 1
print "Debt collection: " + str(debt_counter) #print results. These
should match the counters from the mapper
print "Mortgage: " + str(mortgage_counter)
print "Other: " + str(other_counter)
```

Overwriting reducer.py

## HW 3.1 Start Hadoop

```
In [3]: #Start hadoop yarn
! /usr/local/Cellar/hadoop/2.7.1/sbin/start-yarn.sh
! /usr/local/Cellar/hadoop/2.7.1/sbin/start-dfs.sh
```

```
starting yarn daemons
starting resourcemanager, logging to /usr/local/Cellar/hadoop/2.7.1/libexec/logs/yarn-dunmireg-resourcemanager-Glenns-Air.home.out
localhost: starting nodemanager, logging to /usr/local/Cellar/hadoop/2.7.1/libexec/logs/yarn-dunmireg-nodemanager-Glenns-Air.home.out
16/01/29 21:04:55 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Starting namenodes on [localhost]
localhost: starting namenode, logging to /usr/local/Cellar/hadoop/2.7.1/libexec/logs/hadoop-dunmireg-namenode-Glenns-Air.home.out
localhost: starting datanode, logging to /usr/local/Cellar/hadoop/2.7.1/libexec/logs/hadoop-dunmireg-datanode-Glenns-Air.home.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/Cellar/hadoop/2.7.1/libexec/logs/hadoop-dunmireg-secondarynamenode-Glenns-Air.home.out
16/01/29 21:05:11 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

```
In [4]: #add input to hdfs
!hdfs dfs -put Consumer_Complaints.csv /user/dunmireg
```

```
16/01/29 21:05:29 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

## HW3.1 Hadoop MapReduce

```
In [5]: !hadoop jar /usr/local/Cellar/hadoop/2.7.1/libexec/share/hadoop/tools/lib/hadoop-streaming-2.7.1.jar \
        -mapper mapper.py \
        -reducer reducer.py \
        -input Consumer_Complaints.csv \
        -output consumer_counters
```





```
16/01/29 21:05:34 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/01/29 21:05:34 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
16/01/29 21:05:34 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/01/29 21:05:34 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
16/01/29 21:05:35 INFO mapred.FileInputFormat: Total input paths to process : 1
16/01/29 21:05:35 INFO mapreduce.JobSubmitter: number of splits:1
16/01/29 21:05:35 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1206054309_0001
16/01/29 21:05:35 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/01/29 21:05:35 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/01/29 21:05:35 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
16/01/29 21:05:35 INFO mapreduce.Job: Running job: job_local1206054309_0001
16/01/29 21:05:35 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/29 21:05:35 INFO mapred.LocalJobRunner: Waiting for map tasks
16/01/29 21:05:35 INFO mapred.LocalJobRunner: Starting task: attempt_local1206054309_0001_m_000000_0
16/01/29 21:05:36 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/29 21:05:36 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
16/01/29 21:05:36 INFO mapred.Task: Using ResourceCalculatorProcessTree : null
16/01/29 21:05:36 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/dunmireg/Consumer_Complaints.csv:0+50906486
16/01/29 21:05:36 INFO mapred.MapTask: numReduceTasks: 1
16/01/29 21:05:36 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/01/29 21:05:36 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/01/29 21:05:36 INFO mapred.MapTask: soft limit at 83886080
16/01/29 21:05:36 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/01/29 21:05:36 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/01/29 21:05:36 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
16/01/29 21:05:36 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/dunmireg/Documents/261HW/HW3/./mapper.py]
16/01/29 21:05:36 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
16/01/29 21:05:36 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir
```

```
16/01/29 21:05:36 INFO Configuration.deprecation: map.input.file i
s deprecated. Instead, use mapreduce.map.input.file
16/01/29 21:05:36 INFO Configuration.deprecation: mapred.skip.on i
s deprecated. Instead, use mapreduce.job.skiprecords
16/01/29 21:05:36 INFO Configuration.deprecation: map.input.length
is deprecated. Instead, use mapreduce.map.input.length
16/01/29 21:05:36 INFO Configuration.deprecation: mapred.work.outp
ut.dir is deprecated. Instead, use mapreduce.task.output.dir
16/01/29 21:05:36 INFO Configuration.deprecation: map.input.start
is deprecated. Instead, use mapreduce.map.input.start
16/01/29 21:05:36 INFO Configuration.deprecation: mapred.job.id is
deprecated. Instead, use mapreduce.job.id
16/01/29 21:05:36 INFO Configuration.deprecation: user.name is dep
recated. Instead, use mapreduce.job.user.name
16/01/29 21:05:36 INFO Configuration.deprecation: mapred.task.is.m
ap is deprecated. Instead, use mapreduce.task.ismap
16/01/29 21:05:36 INFO Configuration.deprecation: mapred.task.id i
s deprecated. Instead, use mapreduce.task.attempt.id
16/01/29 21:05:36 INFO Configuration.deprecation: mapred.task.part
ition is deprecated. Instead, use mapreduce.task.partition
16/01/29 21:05:36 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/01/29 21:05:36 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/29 21:05:36 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/29 21:05:36 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/29 21:05:36 INFO streaming.PipeMapRed: Records R/W=2032/1
16/01/29 21:05:36 INFO streaming.PipeMapRed: R/W/S=10000/7896/0 i
n:NA [rec/s] out:NA [rec/s]
16/01/29 21:05:36 INFO mapreduce.Job: Job job_local1206054309_0001
running in uber mode : false
16/01/29 21:05:36 INFO mapreduce.Job: map 0% reduce 0%
16/01/29 21:05:37 INFO streaming.PipeMapRed: R/W/S=100000/98646/0
in:NA [rec/s] out:NA [rec/s]
16/01/29 21:05:37 INFO streaming.PipeMapRed: R/W/S=200000/198558/0
in:200000=200000/1 [rec/s] out:198558=198558/1 [rec/s]
16/01/29 21:05:38 INFO streaming.PipeMapRed: R/W/S=300000/298306/0
in:150000=300000/2 [rec/s] out:149153=298306/2 [rec/s]
16/01/29 21:05:38 INFO streaming.PipeMapRed: MRErrorThread done
16/01/29 21:05:38 INFO streaming.PipeMapRed: mapRedFinished
16/01/29 21:05:38 INFO mapred.LocalJobRunner:
16/01/29 21:05:38 INFO mapred.MapTask: Starting flush of map output
t
16/01/29 21:05:38 INFO mapred.MapTask: Spilling map output
16/01/29 21:05:38 INFO mapred.MapTask: bufstart = 0; bufend = 4878
322; bufvoid = 104857600
16/01/29 21:05:38 INFO mapred.MapTask: kvstart = 26214396(10485758
4); kvend = 24962748(99850992); length = 1251649/6553600
16/01/29 21:05:38 INFO mapred.MapTask: Finished spill 0
16/01/29 21:05:38 INFO mapred.Task: Task:attempt_local1206054309_0
001_m_000000_0 is done. And is in the process of committing
16/01/29 21:05:38 INFO mapred.LocalJobRunner: Records R/W=2032/1
```

```
16/01/29 21:05:38 INFO mapred.Task: Task 'attempt_local1206054309_0001_m_000000_0' done.
16/01/29 21:05:38 INFO mapred.LocalJobRunner: Finishing task: attempt_local1206054309_0001_m_000000_0
16/01/29 21:05:38 INFO mapred.LocalJobRunner: map task executor complete.
16/01/29 21:05:38 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/01/29 21:05:38 INFO mapred.LocalJobRunner: Starting task: attempt_local1206054309_0001_r_000000_0
16/01/29 21:05:38 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/29 21:05:38 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
16/01/29 21:05:38 INFO mapred.Task: Using ResourceCalculatorProcessTree : null
16/01/29 21:05:38 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@2102fea2
16/01/29 21:05:38 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLimit=83584616, mergeThreshold=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
16/01/29 21:05:38 INFO reduce.EventFetcher: attempt_local1206054309_0001_r_000000_0 Thread started: EventFetcher for fetching Map Completion Events
16/01/29 21:05:38 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1206054309_0001_m_000000_0 decomp: 5504150 len: 5504154 to MEMORY
16/01/29 21:05:38 INFO mapreduce.Job: map 100% reduce 0%
16/01/29 21:05:38 INFO reduce.InMemoryMapOutput: Read 5504150 bytes from map-output for attempt_local1206054309_0001_m_000000_0
16/01/29 21:05:38 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 5504150, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory -> 5504150
16/01/29 21:05:38 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/01/29 21:05:38 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/29 21:05:38 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs
16/01/29 21:05:38 INFO mapred.Merger: Merging 1 sorted segments
16/01/29 21:05:38 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 5504124 bytes
16/01/29 21:05:39 INFO reduce.MergeManagerImpl: Merged 1 segments, 5504150 bytes to disk to satisfy reduce memory limit
16/01/29 21:05:39 INFO reduce.MergeManagerImpl: Merging 1 files, 5504154 bytes from disk
16/01/29 21:05:39 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/01/29 21:05:39 INFO mapred.Merger: Merging 1 sorted segments
16/01/29 21:05:39 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 5504124 bytes
16/01/29 21:05:39 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/29 21:05:39 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/dunmireg/Documents/261HW/HW3/./reducer.py]
16/01/29 21:05:39 INFO Configuration.deprecation: mapred.job.track
```

```

er is deprecated. Instead, use mapreduce.jobtracker.address
16/01/29 21:05:39 INFO Configuration.deprecation: mapred.map.tasks
is deprecated. Instead, use mapreduce.job.maps
16/01/29 21:05:39 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/01/29 21:05:39 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/29 21:05:39 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/29 21:05:39 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/29 21:05:39 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/29 21:05:39 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/01/29 21:05:39 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/01/29 21:05:39 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/01/29 21:05:39 INFO streaming.PipeMapRed: MRErrorThread done
16/01/29 21:05:39 INFO streaming.PipeMapRed: Records R/W=312913/1
16/01/29 21:05:39 INFO streaming.PipeMapRed: mapRedFinished
16/01/29 21:05:39 INFO mapred.Task: Task:attempt_local1206054309_0
001_r_000000_0 is done. And is in the process of committing
16/01/29 21:05:39 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/29 21:05:39 INFO mapred.Task: Task attempt_local1206054309_0
001_r_000000_0 is allowed to commit now
16/01/29 21:05:39 INFO output.FileOutputCommitter: Saved output of
task 'attempt_local1206054309_0001_r_000000_0' to hdfs://localhos
t:9000/user/dunmireg/consumer_counters/_temporary/0/task_local1206
054309_0001_r_000000
16/01/29 21:05:39 INFO mapred.LocalJobRunner: Records R/W=312913/1
> reduce
16/01/29 21:05:39 INFO mapred.Task: Task 'attempt_local1206054309_
0001_r_000000_0' done.
16/01/29 21:05:39 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local1206054309_0001_r_000000_0
16/01/29 21:05:39 INFO mapred.LocalJobRunner: reduce task executor
complete.
16/01/29 21:05:39 INFO mapreduce.Job: map 100% reduce 100%
16/01/29 21:05:39 INFO mapreduce.Job: Job job_local1206054309_0001
completed successfully
16/01/29 21:05:39 INFO mapreduce.Job: Counters: 39

```

#### File System Counters

```

FILE: Number of bytes read=11220438
FILE: Number of bytes written=17314316
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=101812972
HDFS: Number of bytes written=57
HDFS: Number of read operations=13
HDFS: Number of large read operations=0
HDFS: Number of write operations=4

```

## Map-Reduce Framework

```

Map input records=312913
Map output records=312913
Map output bytes=4878322
Map output materialized bytes=5504154
Input split bytes=111
Combine input records=0
Combine output records=0
Reduce input groups=10
Reduce shuffle bytes=5504154
Reduce input records=312913
Reduce output records=3
Spilled Records=625826
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=9
Total committed heap usage (bytes)=671088640

```

## Debt-Counter

```
Total=44372
```

## Mortgage-Counter

```
Total=125752
```

## Other-Counter

```
Total=142789
```

## Shuffle Errors

```

BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0

```

## Tokens

```
Total=312913
```

## File Input Format Counters

```
Bytes Read=50906486
```

## File Output Format Counters

```
Bytes Written=57
```

```
16/01/29 21:05:39 INFO streaming.StreamJob: Output directory: consumer_counters
```

```

In [6]: #show results
!hdfs dfs -cat /user/dunmireg/consumer_counters/part-00000

```

```

16/01/29 21:06:12 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Debt collection: 44372
Mortgage: 125752
Other: 142789

```

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



## Counters for job\_1454175435207\_0002

Logged in as: drwho

Counter Group	Name	Map	Reduce	Total
File System Counters	FILE: Number of bytes read	0	0	0
	FILE: Number of bytes written	236,642	0	236,642
	FILE: Number of large read operations	0	0	0
	FILE: Number of read operations	0	0	0
	FILE: Number of write operations	0	0	0
	HDFS: Number of bytes read	50,910,105	0	50,910,105
	HDFS: Number of bytes written	0	0	0
	HDFS: Number of large read operations	0	0	0
	HDFS: Number of read operations	10	0	10
	HDFS: Number of write operations	4	0	4
Job Counters	Launched map tasks	0	0	2
	Rack-local map tasks	0	0	2
	Total megabyte-seconds taken by all map tasks	0	0	5,151,744
	Total time spent by all map tasks (ms)	0	0	5,031
	Total time spent by all maps in occupied slots (ms)	0	0	5,031
	Total vcore-seconds taken by all map tasks	0	0	5,031
Map-Reduce Framework	CPU time spent (ms)	0	0	0
	Failed Shuffles	0	0	0
	GC time elapsed (ms)	85	0	85
	Input split bytes	222	0	222
	Map input records	312,913	0	312,913
	Map output records	0	0	0
	Merged Map outputs	0	0	0
	Physical memory (bytes) snapshot	0	0	0
	Spilled Records	0	0	0
	Total committed heap usage (bytes)	284,688,384	0	284,688,384
	Virtual memory (bytes) snapshot	0	0	0
Category Counters	debt collection	44,372	0	44,372
	mortgage	125,752	0	125,752
	other	142,788	0	142,788
File Input Format Counters	Bytes Read	50,909,883	0	50,909,883
File Output Format Counters	Bytes Written	0	0	0

Notice that the mapper counters and the results from counting in the reducer match. In total we found:

**Debt collection: 44372**

**Mortgage: 125752**

**Other: 142789**

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

## HW3.2 Part 1

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

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

### Write input string to a file on disk

```
In [1]: #make a basic text file to test
        with open('input_text.txt', 'w') as myfile:
            myfile.write('foo foo quux labs foo bar quux')
```

## HW3.2 Word Count Map Function

```
In [8]: %%writefile mapper.py
        #!/usr/bin/python
        import sys

        sys.stderr.write('reporter:counter:Map-Count,Total,1\n') #increment
        mapper counter
        for line in sys.stdin: #read input
            line = line.strip()
            line = line.rstrip()
            line = line.split() #split on space delimiter
            for word in line:
                print word + '\t' + '1' #emit word and count of 1
```

Overwriting mapper.py

## HW3.2 Word Count Reducer



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

current_word = None #keep track of current word and current count
current_count = None
word = None
sys.stderr.write('reporter:counter:Reduce-Counter,Total,1\n') #increment reducer counter
for line in sys.stdin:
    line = line.strip()
    line = line.rstrip()
    word, count = line.split('\t') #split here on tab

    if current_word == word: #increment the counter
        current_count += int(count)
    else:
        if current_word:
            print '%s\t%s' % (current_word, current_count) #print results when we have found a new word
            current_word = word #change word
            current_count = int(count)

#print final word
if current_word == word:
    print '%s\t%s' % (current_word, current_count)
```

Overwriting reducer.py

```
In [19]: #!/cat input.txt | python mapper.py | sort | python reducer.py
```

## HW3.2 Hadoop MapReduce

```
In [1]: # create input file and add input to hdfs
!echo "foo foo quux labs foo bar quux" >input_text.txt
!hdfs dfs -put input_text.txt /user/dunmireg
```

16/02/01 21:00:38 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

```
In [11]: #run hadoop, manually setting the number of mappers and reducers
!hadoop jar /usr/local/Cellar/hadoop/2.7.1/libexec/share/hadoop/tools/lib/hadoop-streaming-2.7.1.jar \
-D mapred.map.tasks=1 \
-D mapred.reduce.tasks=4 \
-mapper mapper.py \
-reducer reducer.py \
-input input_text.txt \
-output word_count
```



```
16/02/01 22:33:48 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/02/01 22:33:49 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
16/02/01 22:33:49 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/02/01 22:33:49 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
16/02/01 22:33:49 INFO mapred.FileInputFormat: Total input paths to process : 1
16/02/01 22:33:49 INFO mapreduce.JobSubmitter: number of splits:1
16/02/01 22:33:49 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
16/02/01 22:33:49 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
16/02/01 22:33:49 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local927319060_0001
16/02/01 22:33:49 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/02/01 22:33:49 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/02/01 22:33:49 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
16/02/01 22:33:49 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/01 22:33:49 INFO mapreduce.Job: Running job: job_local927319060_0001
16/02/01 22:33:50 INFO mapred.LocalJobRunner: Starting task: attempt_local927319060_0001_m_000000_0
16/02/01 22:33:50 INFO mapred.LocalJobRunner: Waiting for map tasks
16/02/01 22:33:50 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/01 22:33:50 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
16/02/01 22:33:50 INFO mapred.Task: Using ResourceCalculatorProcessTree : null
16/02/01 22:33:50 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/dunmireg/input_text.txt:0+30
16/02/01 22:33:50 INFO mapred.MapTask: numReduceTasks: 4
16/02/01 22:33:50 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/02/01 22:33:50 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/02/01 22:33:50 INFO mapred.MapTask: soft limit at 83886080
16/02/01 22:33:50 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/02/01 22:33:50 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/02/01 22:33:50 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
16/02/01 22:33:50 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/dunmireg/Documents/261HW/HW3/./mapper.py]
```

```
16/02/01 22:33:50 INFO Configuration.deprecation: mapred.tip.id is
deprecated. Instead, use mapreduce.task.id
16/02/01 22:33:50 INFO Configuration.deprecation: mapred.local.dir
is deprecated. Instead, use mapreduce.cluster.local.dir
16/02/01 22:33:50 INFO Configuration.deprecation: map.input.file i
s deprecated. Instead, use mapreduce.map.input.file
16/02/01 22:33:50 INFO Configuration.deprecation: mapred.skip.on i
s deprecated. Instead, use mapreduce.job.skiprecords
16/02/01 22:33:50 INFO Configuration.deprecation: map.input.length
is deprecated. Instead, use mapreduce.map.input.length
16/02/01 22:33:50 INFO Configuration.deprecation: mapred.work.outp
ut.dir is deprecated. Instead, use mapreduce.task.output.dir
16/02/01 22:33:50 INFO Configuration.deprecation: map.input.start
is deprecated. Instead, use mapreduce.map.input.start
16/02/01 22:33:50 INFO Configuration.deprecation: mapred.job.id is
deprecated. Instead, use mapreduce.job.id
16/02/01 22:33:50 INFO Configuration.deprecation: user.name is dep
recated. Instead, use mapreduce.job.user.name
16/02/01 22:33:50 INFO Configuration.deprecation: mapred.task.is.m
ap is deprecated. Instead, use mapreduce.task.ismap
16/02/01 22:33:50 INFO Configuration.deprecation: mapred.task.id i
s deprecated. Instead, use mapreduce.task.attempt.id
16/02/01 22:33:50 INFO Configuration.deprecation: mapred.task.part
ition is deprecated. Instead, use mapreduce.task.partition
16/02/01 22:33:50 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/02/01 22:33:50 INFO streaming.PipeMapRed: MRErrorThread done
16/02/01 22:33:50 INFO streaming.PipeMapRed: Records R/W=1/1
16/02/01 22:33:50 INFO streaming.PipeMapRed: mapRedFinished
16/02/01 22:33:50 INFO mapred.LocalJobRunner:
16/02/01 22:33:50 INFO mapred.MapTask: Starting flush of map output
t
16/02/01 22:33:50 INFO mapred.MapTask: Spilling map output
16/02/01 22:33:50 INFO mapred.MapTask: bufstart = 0; bufend = 45;
bufvoid = 104857600
16/02/01 22:33:50 INFO mapred.MapTask: kvstart = 26214396(10485758
4); kvend = 26214372(104857488); length = 25/6553600
16/02/01 22:33:50 INFO mapred.MapTask: Finished spill 0
16/02/01 22:33:50 INFO mapred.Task: Task:attempt_local927319060_00
01_m_000000_0 is done. And is in the process of committing
16/02/01 22:33:50 INFO mapred.LocalJobRunner: Records R/W=1/1
16/02/01 22:33:50 INFO mapred.Task: Task 'attempt_local927319060_0
001_m_000000_0' done.
16/02/01 22:33:50 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local927319060_0001_m_000000_0
16/02/01 22:33:50 INFO mapred.LocalJobRunner: map task executor co
mplete.
16/02/01 22:33:50 INFO mapred.LocalJobRunner: Waiting for reduce t
asks
16/02/01 22:33:50 INFO mapred.LocalJobRunner: Starting task: attem
pt_local927319060_0001_r_000000_0
16/02/01 22:33:50 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
16/02/01 22:33:50 INFO util.ProcfsBasedProcessTree: ProcfsBasedPro
```

```
cessTree currently is supported only on Linux.
16/02/01 22:33:50 INFO mapred.Task: Using ResourceCalculatorProce
ssTree : null
16/02/01 22:33:50 INFO mapred.ReduceTask: Using ShuffleConsumerPlu
gin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@347fc133
16/02/01 22:33:50 INFO reduce.MergeManagerImpl: MergerManager: mem
oryLimit=334338464, maxSingleShuffleLimit=83584616, mergeThreshold
=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
16/02/01 22:33:50 INFO reduce.EventFetcher: attempt_local927319060
_0001_r_000000_0 Thread started: EventFetcher for fetching Map Com
pletion Events
16/02/01 22:33:50 INFO reduce.LocalFetcher: localfetcher#1 about t
o shuffle output of map attempt_local927319060_0001_m_000000_0 dec
omp: 20 len: 24 to MEMORY
16/02/01 22:33:50 INFO reduce.InMemoryMapOutput: Read 20 bytes fro
m map-output for attempt_local927319060_0001_m_000000_0
16/02/01 22:33:50 INFO reduce.MergeManagerImpl: closeInMemoryFile
-> map-output of size: 20, inMemoryMapOutputs.size() -> 1, commitM
emory -> 0, usedMemory ->20
16/02/01 22:33:50 INFO reduce.EventFetcher: EventFetcher is interr
upted.. Returning
16/02/01 22:33:50 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/01 22:33:50 INFO reduce.MergeManagerImpl: finalMerge called
with 1 in-memory map-outputs and 0 on-disk map-outputs
16/02/01 22:33:50 INFO mapred.Merger: Merging 1 sorted segments
16/02/01 22:33:50 INFO mapred.Merger: Down to the last merge-pass,
with 1 segments left of total size: 13 bytes
16/02/01 22:33:50 INFO reduce.MergeManagerImpl: Merged 1 segments,
20 bytes to disk to satisfy reduce memory limit
16/02/01 22:33:50 INFO reduce.MergeManagerImpl: Merging 1 files, 2
4 bytes from disk
16/02/01 22:33:50 INFO reduce.MergeManagerImpl: Merging 0 segment
s, 0 bytes from memory into reduce
16/02/01 22:33:50 INFO mapred.Merger: Merging 1 sorted segments
16/02/01 22:33:50 INFO mapred.Merger: Down to the last merge-pass,
with 1 segments left of total size: 13 bytes
16/02/01 22:33:50 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/01 22:33:50 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/dunmireg/Documents/261HW/HW3/./reducer.py]
16/02/01 22:33:50 INFO Configuration.deprecation: mapred.job.track
er is deprecated. Instead, use mapreduce.jobtracker.address
16/02/01 22:33:50 INFO Configuration.deprecation: mapred.map.tasks
is deprecated. Instead, use mapreduce.job.maps
16/02/01 22:33:50 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/02/01 22:33:50 INFO streaming.PipeMapRed: Records R/W=2/1
16/02/01 22:33:50 INFO streaming.PipeMapRed: MRErrorThread done
16/02/01 22:33:50 INFO streaming.PipeMapRed: mapRedFinished
16/02/01 22:33:50 INFO mapred.Task: Task:attempt_local927319060_00
01_r_000000_0 is done. And is in the process of committing
16/02/01 22:33:50 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/01 22:33:50 INFO mapred.Task: Task attempt_local927319060_00
01_r_000000_0 is allowed to commit now
16/02/01 22:33:50 INFO output.FileOutputCommitter: Saved output of
```

```
task 'attempt_local927319060_0001_r_000000_0' to hdfs://localhost:
9000/user/dunmireg/word_count/_temporary/0/task_local927319060_000
1_r_000000
16/02/01 22:33:50 INFO mapred.LocalJobRunner: Records R/W=2/1 > re
duce
16/02/01 22:33:50 INFO mapred.Task: Task 'attempt_local927319060_0
001_r_000000_0' done.
16/02/01 22:33:50 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local927319060_0001_r_000000_0
16/02/01 22:33:50 INFO mapred.LocalJobRunner: Starting task: attem
pt_local927319060_0001_r_000001_0
16/02/01 22:33:50 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
16/02/01 22:33:50 INFO util.ProcfsBasedProcessTree: ProcfsBasedPro
cessTree currently is supported only on Linux.
16/02/01 22:33:50 INFO mapred.Task: Using ResourceCalculatorProce
ssTree : null
16/02/01 22:33:50 INFO mapred.ReduceTask: Using ShuffleConsumerPlu
gin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@34751e75
16/02/01 22:33:50 INFO reduce.MergeManagerImpl: MergerManager: mem
oryLimit=334338464, maxSingleShuffleLimit=83584616, mergeThreshold
=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
16/02/01 22:33:50 INFO reduce.EventFetcher: attempt_local927319060
_0001_r_000001_0 Thread started: EventFetcher for fetching Map Com
pletion Events
16/02/01 22:33:50 INFO reduce.LocalFetcher: localfetcher#2 about t
o shuffle output of map attempt_local927319060_0001_m_000000_0 dec
omp: 26 len: 30 to MEMORY
16/02/01 22:33:50 INFO reduce.InMemoryMapOutput: Read 26 bytes fro
m map-output for attempt_local927319060_0001_m_000000_0
16/02/01 22:33:50 INFO reduce.MergeManagerImpl: closeInMemoryFile
-> map-output of size: 26, inMemoryMapOutputs.size() -> 1, commitM
emory -> 0, usedMemory ->26
16/02/01 22:33:50 INFO reduce.EventFetcher: EventFetcher is interr
upted.. Returning
16/02/01 22:33:50 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/01 22:33:50 INFO reduce.MergeManagerImpl: finalMerge called
with 1 in-memory map-outputs and 0 on-disk map-outputs
16/02/01 22:33:50 INFO mapred.Merger: Merging 1 sorted segments
16/02/01 22:33:50 INFO mapred.Merger: Down to the last merge-pass,
with 1 segments left of total size: 20 bytes
16/02/01 22:33:50 INFO reduce.MergeManagerImpl: Merged 1 segments,
26 bytes to disk to satisfy reduce memory limit
16/02/01 22:33:50 INFO reduce.MergeManagerImpl: Merging 1 files, 3
0 bytes from disk
16/02/01 22:33:50 INFO reduce.MergeManagerImpl: Merging 0 segment
s, 0 bytes from memory into reduce
16/02/01 22:33:50 INFO mapred.Merger: Merging 1 sorted segments
16/02/01 22:33:50 INFO mapred.Merger: Down to the last merge-pass,
with 1 segments left of total size: 20 bytes
16/02/01 22:33:50 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/01 22:33:50 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/dunmireg/Documents/261HW/HW3/./reducer.py]
16/02/01 22:33:50 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
```

```
c/s] out:NA [rec/s]
16/02/01 22:33:50 INFO streaming.PipeMapRed: Records R/W=3/1
16/02/01 22:33:50 INFO streaming.PipeMapRed: MRErrorThread done
16/02/01 22:33:50 INFO streaming.PipeMapRed: mapRedFinished
16/02/01 22:33:50 INFO mapred.Task: Task:attempt_local927319060_00
01_r_000001_0 is done. And is in the process of committing
16/02/01 22:33:50 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/01 22:33:50 INFO mapred.Task: Task attempt_local927319060_00
01_r_000001_0 is allowed to commit now
16/02/01 22:33:50 INFO output.FileOutputCommitter: Saved output of
task 'attempt_local927319060_0001_r_000001_0' to hdfs://localhost:
9000/user/dunmireg/word_count/_temporary/0/task_local927319060_000
1_r_000001
16/02/01 22:33:50 INFO mapred.LocalJobRunner: Records R/W=3/1 > re
duce
16/02/01 22:33:50 INFO mapred.Task: Task 'attempt_local927319060_0
001_r_000001_0' done.
16/02/01 22:33:50 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local927319060_0001_r_000001_0
16/02/01 22:33:50 INFO mapred.LocalJobRunner: Starting task: attem
pt_local927319060_0001_r_000002_0
16/02/01 22:33:50 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
16/02/01 22:33:50 INFO util.ProcfsBasedProcessTree: ProcfsBasedPro
cessTree currently is supported only on Linux.
16/02/01 22:33:50 INFO mapred.Task: Using ResourceCalculatorProce
ssTree : null
16/02/01 22:33:50 INFO mapred.ReduceTask: Using ShuffleConsumerPlu
gin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@4e432ccb
16/02/01 22:33:50 INFO reduce.MergeManagerImpl: MergerManager: mem
oryLimit=334338464, maxSingleShuffleLimit=83584616, mergeThreshold
=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
16/02/01 22:33:50 INFO reduce.EventFetcher: attempt_local927319060
_0001_r_000002_0 Thread started: EventFetcher for fetching Map Com
pletion Events
16/02/01 22:33:50 INFO reduce.LocalFetcher: localfetcher#3 about t
o shuffle output of map attempt_local927319060_0001_m_000000_0 dec
omp: 10 len: 14 to MEMORY
16/02/01 22:33:50 INFO reduce.InMemoryMapOutput: Read 10 bytes fro
m map-output for attempt_local927319060_0001_m_000000_0
16/02/01 22:33:50 INFO reduce.MergeManagerImpl: closeInMemoryFile
-> map-output of size: 10, inMemoryMapOutputs.size() -> 1, commitM
emory -> 0, usedMemory ->10
16/02/01 22:33:50 INFO reduce.EventFetcher: EventFetcher is interr
upted.. Returning
16/02/01 22:33:50 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/01 22:33:50 INFO reduce.MergeManagerImpl: finalMerge called
with 1 in-memory map-outputs and 0 on-disk map-outputs
16/02/01 22:33:50 INFO mapred.Merger: Merging 1 sorted segments
16/02/01 22:33:50 INFO mapred.Merger: Down to the last merge-pass,
with 1 segments left of total size: 4 bytes
16/02/01 22:33:50 INFO reduce.MergeManagerImpl: Merged 1 segments,
10 bytes to disk to satisfy reduce memory limit
16/02/01 22:33:50 INFO reduce.MergeManagerImpl: Merging 1 files, 1
```



```
4 bytes from disk
16/02/01 22:33:50 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/02/01 22:33:50 INFO mapred.Merger: Merging 1 sorted segments
16/02/01 22:33:50 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 4 bytes
16/02/01 22:33:50 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/01 22:33:50 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/dunmireg/Documents/261HW/HW3/./reducer.py]
16/02/01 22:33:50 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/01 22:33:50 INFO streaming.PipeMapRed: Records R/W=1/1
16/02/01 22:33:50 INFO streaming.PipeMapRed: MRErrorThread done
16/02/01 22:33:50 INFO streaming.PipeMapRed: mapRedFinished
16/02/01 22:33:50 INFO mapred.Task: Task:attempt_local927319060_0001_r_000002_0 is done. And is in the process of committing
16/02/01 22:33:50 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/01 22:33:50 INFO mapred.Task: Task attempt_local927319060_0001_r_000002_0 is allowed to commit now
16/02/01 22:33:50 INFO output.FileOutputCommitter: Saved output of task 'attempt_local927319060_0001_r_000002_0' to hdfs://localhost:9000/user/dunmireg/word_count/_temporary/0/task_local927319060_0001_r_000002
16/02/01 22:33:50 INFO mapred.LocalJobRunner: Records R/W=1/1 > reduce
16/02/01 22:33:50 INFO mapred.Task: Task 'attempt_local927319060_0001_r_000002_0' done.
16/02/01 22:33:50 INFO mapred.LocalJobRunner: Finishing task: attempt_local927319060_0001_r_000002_0
16/02/01 22:33:50 INFO mapred.LocalJobRunner: Starting task: attempt_local927319060_0001_r_000003_0
16/02/01 22:33:50 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/01 22:33:50 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
16/02/01 22:33:50 INFO mapred.Task: Using ResourceCalculatorProcessTree : null
16/02/01 22:33:50 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@15760f0b
16/02/01 22:33:50 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLimit=83584616, mergeThreshold=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
16/02/01 22:33:50 INFO reduce.EventFetcher: attempt_local927319060_0001_r_000003_0 Thread started: EventFetcher for fetching Map Completion Events
16/02/01 22:33:50 INFO reduce.LocalFetcher: localfetcher#4 about to shuffle output of map attempt_local927319060_0001_m_000000_0 decomp: 11 len: 15 to MEMORY
16/02/01 22:33:50 INFO reduce.InMemoryMapOutput: Read 11 bytes from map-output for attempt_local927319060_0001_m_000000_0
16/02/01 22:33:50 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 11, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory -> 11
16/02/01 22:33:50 INFO reduce.EventFetcher: EventFetcher is interr
```

```

upted.. Returning
16/02/01 22:33:50 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/01 22:33:50 INFO reduce.MergeManagerImpl: finalMerge called
with 1 in-memory map-outputs and 0 on-disk map-outputs
16/02/01 22:33:50 INFO mapred.Merger: Merging 1 sorted segments
16/02/01 22:33:50 INFO mapred.Merger: Down to the last merge-pass,
with 1 segments left of total size: 4 bytes
16/02/01 22:33:50 INFO reduce.MergeManagerImpl: Merged 1 segments,
11 bytes to disk to satisfy reduce memory limit
16/02/01 22:33:50 INFO reduce.MergeManagerImpl: Merging 1 files, 1
5 bytes from disk
16/02/01 22:33:50 INFO reduce.MergeManagerImpl: Merging 0 segment
s, 0 bytes from memory into reduce
16/02/01 22:33:50 INFO mapred.Merger: Merging 1 sorted segments
16/02/01 22:33:50 INFO mapred.Merger: Down to the last merge-pass,
with 1 segments left of total size: 4 bytes
16/02/01 22:33:50 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/01 22:33:50 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/dunmireg/Documents/261HW/HW3/./reducer.py]
16/02/01 22:33:50 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/02/01 22:33:50 INFO streaming.PipeMapRed: Records R/W=1/1
16/02/01 22:33:50 INFO streaming.PipeMapRed: MRErrorThread done
16/02/01 22:33:50 INFO streaming.PipeMapRed: mapRedFinished
16/02/01 22:33:50 INFO mapred.Task: Task:attempt_local927319060_00
01_r_000003_0 is done. And is in the process of committing
16/02/01 22:33:50 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/01 22:33:50 INFO mapred.Task: Task attempt_local927319060_00
01_r_000003_0 is allowed to commit now
16/02/01 22:33:50 INFO output.FileOutputCommitter: Saved output of
task 'attempt_local927319060_0001_r_000003_0' to hdfs://localhost:
9000/user/dunmireg/word_count/_temporary/0/task_local927319060_000
1_r_000003
16/02/01 22:33:50 INFO mapred.LocalJobRunner: Records R/W=1/1 > re
duce
16/02/01 22:33:50 INFO mapred.Task: Task 'attempt_local927319060_0
001_r_000003_0' done.
16/02/01 22:33:50 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local927319060_0001_r_000003_0
16/02/01 22:33:50 INFO mapred.LocalJobRunner: reduce task executor
complete.
16/02/01 22:33:50 INFO mapreduce.Job: Job job_local927319060_0001
running in uber mode : false
16/02/01 22:33:50 INFO mapreduce.Job: map 100% reduce 100%
16/02/01 22:33:50 INFO mapreduce.Job: Job job_local927319060_0001
completed successfully
16/02/01 22:33:51 INFO mapreduce.Job: Counters: 37

```

#### File System Counters

```

FILE: Number of bytes read=532031
FILE: Number of bytes written=1998924
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=150

```

```

HDFS: Number of bytes written=65
HDFS: Number of read operations=55
HDFS: Number of large read operations=0
HDFS: Number of write operations=25

```

#### Map-Reduce Framework

```

Map input records=1
Map output records=7
Map output bytes=45
Map output materialized bytes=83
Input split bytes=102
Combine input records=0
Combine output records=0
Reduce input groups=4
Reduce shuffle bytes=83
Reduce input records=7
Reduce output records=4
Spilled Records=14
Shuffled Maps =4
Failed Shuffles=0
Merged Map outputs=4
GC time elapsed (ms)=7
Total committed heap usage (bytes)=1551892480

```

#### Map-Count

```
Total=1
```

#### Reduce-Counter

```
Total=4
```

#### Shuffle Errors

```

BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0

```

#### File Input Format Counters

```
Bytes Read=30
```

#### File Output Format Counters

```
Bytes Written=26
```

```
16/02/01 22:33:51 INFO streaming.StreamJob: Output directory: word_count
```

**RESULT: Mappers - 1, Reducers - 4**

## HW3.2 Part 2

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

## HW3.2 Part 2 Map Function

```
In [182]: %%writefile mapper32b.py
#!/usr/bin/env python
#
# W261 HW 3.2 MapReduce and Counters for Code Analysis
# Read from a CSV file of consumer complaints with fields as follow
s:
#
# Complaint ID,Product,Sub-product,Issue,Sub-issue,State,ZIP code,S
ubmitted via,Date received,
# Date sent to company,Company,Company response,Timely response?,Co
nsumer disputed?
#
# Use counters to count the number of times the mapper is called
#
# Remember that in Hadoop streaming, to update a counter is to writ
e to STDERR in the format
# reporter:counter:<group>,<counter>,<amount>

import sys
import re

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

# Read data from STDIN and use counters to count the data
def main(separator='\t'):
    for line in sys.stdin:
        fields = line.split(',')
        try:
            # check to see if this is a header by trying to convert
the first field to an integer
            id = int(fields[0])
            # we have a real record, so do some mapping
            for word in WORD_RE.findall(fields[3]):
                sys.stdout.write('{0}{1}{2}\n'.format(word, separat
or, 1))
        except:
            # must be a header record so skip it
            pass

if __name__ == "__main__":
    # increment counter for mapper call, write to STDERR
    sys.stderr.write("reporter:counter:Code Call Counters,mapper,1
\n")
    main()
```

Overwriting mapper32b.py

## HW3.2 Part Reduce Function

```
In [194]: %%writefile reducer32b.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)

def main(separator='\t'):
    # input comes from STDIN (standard input)
    data = read(sys.stdin, separator=separator)
    # groupby groups multiple word-count pairs by word,
    # and creates an iterator that returns consecutive keys and the
    # group:
    #   current_word - string containing a word (the key)
    #   group - iterator yielding all ["<current_word>", "<count>"] items
    for current_word, group in groupby(data, itemgetter(0)):
        try:
            total_count = sum(int(count) for current_word, count in
group)
            sys.stderr.write("reporter:counter:Code Call Counters,red
ucer pairs,1\n")
            sys.stdout.write("{}{}{}\n".format(current_word, sep
arator, total_count))
        except ValueError:
            sys.stderr.write("reporter:counter:Code Call Counters,red
ucer skipped pairs,1\n")
            # count was not a number, so silently discard this item
            pass

if __name__ == "__main__":
    # increment counter for reducer call, write to STDERR
    sys.stderr.write("reporter:counter:Code Call Counters,red
ucer,1\n")
    main()
```

Overwriting reducer32b.py

## HW3.2 Part 2 Hadoop MapReduce

```
In [184]: !chmod a+x mapper32b.py  
          !chmod a+x reducer32b.py
```

```
In [196]: !hdfs dfs -rm -r /user/rcordell/recordsOutput
!yarn jar /usr/local/Cellar/hadoop/2.7.1/libexec/share/hadoop/tools/lib/hadoop-streaming-2.7.1.jar \
-D mapreduce.job.reduces=2 \
-D mapreduce.job.maps=2 \
-files "mapper32b.py, reducer32b.py" \
-mapper mapper32b.py \
-reducer reducer32b.py \
-input Consumer_Complaints.csv \
-output recordsOutput
```





```

16/01/30 19:16:43 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
packageJobJar: [/var/folders/z_/rfp5q2cd6db13d19v6yw0n8w0000gn/T/hadoop-unjar3393828723429341869/] [] /var/folders/z_/rfp5q2cd6db13d19v6yw0n8w0000gn/T/streamjob6639340843874969134.jar tmpDir=null
16/01/30 19:16:44 INFO client.RMPProxy: Connecting to ResourceManager at /0.0.0.0:8032
16/01/30 19:16:44 INFO client.RMPProxy: Connecting to ResourceManager at /0.0.0.0:8032
16/01/30 19:16:44 INFO mapred.FileInputFormat: Total input paths to process : 1
16/01/30 19:16:44 INFO mapreduce.JobSubmitter: number of splits:2
16/01/30 19:16:45 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1454175435207_0018
16/01/30 19:16:45 INFO impl.YarnClientImpl: Submitted application application_1454175435207_0018
16/01/30 19:16:45 INFO mapreduce.Job: The url to track the job: http://Rons-iMac-Retina.local:8088/proxy/application_1454175435207_0018/
16/01/30 19:16:45 INFO mapreduce.Job: Running job: job_1454175435207_0018
16/01/30 19:16:49 INFO mapreduce.Job: Job job_1454175435207_0018 running in uber mode : false
16/01/30 19:16:49 INFO mapreduce.Job:  map 0% reduce 0%
16/01/30 19:16:55 INFO mapreduce.Job:  map 100% reduce 0%
16/01/30 19:17:01 INFO mapreduce.Job:  map 100% reduce 50%
16/01/30 19:17:02 INFO mapreduce.Job:  map 100% reduce 100%
16/01/30 19:17:02 INFO mapreduce.Job: Job job_1454175435207_0018 completed successfully
16/01/30 19:17:02 INFO mapreduce.Job: Counters: 52

```

#### File System Counters

```

FILE: Number of bytes read=11233477
FILE: Number of bytes written=22943108
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=50910105
HDFS: Number of bytes written=2221
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
Rack-local map tasks=2
Total time spent by all maps in occupied slots (ms)=5782
Total time spent by all reduces in occupied slots (ms)=5489
Total time spent by all map tasks (ms)=5782
Total time spent by all reduce tasks (ms)=5489
Total vcore-seconds taken by all map tasks=5782
Total vcore-seconds taken by all reduce tasks=5489

```

```

Total megabyte-seconds taken by all map tasks=5920
768
Total megabyte-seconds taken by all reduce tasks=5
620736
Map-Reduce Framework
  Map input records=312913
  Map output records=980482
  Map output bytes=9272501
  Map output materialized bytes=11233489
  Input split bytes=222
  Combine input records=0
  Combine output records=0
  Reduce input groups=180
  Reduce shuffle bytes=11233489
  Reduce input records=980482
  Reduce output records=180
  Spilled Records=1960964
  Shuffled Maps =4
  Failed Shuffles=0
  Merged Map outputs=4
  GC time elapsed (ms)=142
  CPU time spent (ms)=0
  Physical memory (bytes) snapshot=0
  Virtual memory (bytes) snapshot=0
  Total committed heap usage (bytes)=693108736
Code Call Counters
  mapper=2
  reducer=2
  reducer pairs=180
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=50909883
File Output Format Counters
  Bytes Written=2221
16/01/30 19:17:02 INFO streaming.StreamJob: Output directory: reco
rdsOutput

```

```
In [197]: !hdfs dfs -cat /user/rcordell/recordsOutput/part-00000 | wc -l
```

```

16/01/30 19:17:08 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable

```

86

**RESULTS: Mapper - 2, Reducer - 2, Unique Words - 86**

## HW3.2 Part 3

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

## HW3.2 Part 3 - Combiner Function

The combiner is essentially the same as the aggregation part of the reducer.

```
In [13]: %%writefile combiner.py
#!/usr/bin/python
import sys

#we added a combiner here to perform the exact same function as the
reducer above, the only difference being the counter
current_word = None
current_count = None
word = None
sys.stderr.write('reporter:counter:Combiner-Counter,Total,1\n') #in
crement combiner counter
for line in sys.stdin:
    line = line.split('\t') #split line on tab from standard input
    word = line[0]
    count = int(line[1])

    if current_word == word: #increment word count
        current_count += count
    else:
        if current_word:
            print '%s\t%s' % (current_word, current_count) #print r
esult when found a new word
        current_word = word
        current_count = count

#print last word
if current_word == word:
    print '%s\t%s' % (current_word, current_count)
```

Writing combiner.py

## HW3.2 Part 3 Hadoop MapReduce

```
In [18]: !hadoop jar /usr/local/Cellar/hadoop/2.7.1/libexec/share/hadoop/tools/lib/hadoop-streaming-2.7.1.jar \
-D mapred.reduce.tasks=2 \
-D mapred.map.tasks=2 \
-mapper mapper.py \
-reducer reducer.py \
-combiner combiner.py \
-input Consumer_Complaints.csv \
-output word_count
```



```
16/02/01 22:36:51 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/02/01 22:36:51 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
16/02/01 22:36:51 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/02/01 22:36:51 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
16/02/01 22:36:51 INFO mapred.FileInputFormat: Total input paths to process : 1
16/02/01 22:36:51 INFO mapreduce.JobSubmitter: number of splits:1
16/02/01 22:36:52 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
16/02/01 22:36:52 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
16/02/01 22:36:52 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1588555268_0001
16/02/01 22:36:52 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/02/01 22:36:52 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/02/01 22:36:52 INFO mapreduce.Job: Running job: job_local1588555268_0001
16/02/01 22:36:52 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
16/02/01 22:36:52 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/01 22:36:52 INFO mapred.LocalJobRunner: Starting task: attempt_local1588555268_0001_m_000000_0
16/02/01 22:36:52 INFO mapred.LocalJobRunner: Waiting for map tasks
16/02/01 22:36:52 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/01 22:36:52 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
16/02/01 22:36:52 INFO mapred.Task: Using ResourceCalculatorProcessTree : null
16/02/01 22:36:52 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/dunmireg/Consumer_Complaints.csv:0+50906486
16/02/01 22:36:52 INFO mapred.MapTask: numReduceTasks: 2
16/02/01 22:36:52 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/02/01 22:36:52 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/02/01 22:36:52 INFO mapred.MapTask: soft limit at 83886080
16/02/01 22:36:52 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/02/01 22:36:52 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/02/01 22:36:52 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
16/02/01 22:36:52 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/dunmireg/Documents/261HW/HW3/./mapper.py]
```

```
16/02/01 22:36:52 INFO Configuration.deprecation: mapred.tip.id is
deprecated. Instead, use mapreduce.task.id
16/02/01 22:36:52 INFO Configuration.deprecation: mapred.local.dir
is deprecated. Instead, use mapreduce.cluster.local.dir
16/02/01 22:36:52 INFO Configuration.deprecation: map.input.file i
s deprecated. Instead, use mapreduce.map.input.file
16/02/01 22:36:52 INFO Configuration.deprecation: mapred.skip.on i
s deprecated. Instead, use mapreduce.job.skiprecords
16/02/01 22:36:52 INFO Configuration.deprecation: map.input.length
is deprecated. Instead, use mapreduce.map.input.length
16/02/01 22:36:52 INFO Configuration.deprecation: mapred.work.outp
ut.dir is deprecated. Instead, use mapreduce.task.output.dir
16/02/01 22:36:52 INFO Configuration.deprecation: map.input.start
is deprecated. Instead, use mapreduce.map.input.start
16/02/01 22:36:52 INFO Configuration.deprecation: mapred.job.id is
deprecated. Instead, use mapreduce.job.id
16/02/01 22:36:52 INFO Configuration.deprecation: user.name is dep
recated. Instead, use mapreduce.job.user.name
16/02/01 22:36:52 INFO Configuration.deprecation: mapred.task.is.m
ap is deprecated. Instead, use mapreduce.task.ismap
16/02/01 22:36:52 INFO Configuration.deprecation: mapred.task.id i
s deprecated. Instead, use mapreduce.task.attempt.id
16/02/01 22:36:52 INFO Configuration.deprecation: mapred.task.part
ition is deprecated. Instead, use mapreduce.task.partition
16/02/01 22:36:52 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/02/01 22:36:52 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/02/01 22:36:52 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/02/01 22:36:52 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/02/01 22:36:52 INFO streaming.PipeMapRed: Records R/W=1043/1
16/02/01 22:36:52 INFO streaming.PipeMapRed: R/W/S=10000/39415/0 i
n:NA [rec/s] out:NA [rec/s]
16/02/01 22:36:53 INFO mapreduce.Job: Job job_local1588555268_0001
running in uber mode : false
16/02/01 22:36:53 INFO mapreduce.Job: map 0% reduce 0%
16/02/01 22:36:54 INFO streaming.PipeMapRed: R/W/S=100000/448120/0
in:100000=100000/1 [rec/s] out:448120=448120/1 [rec/s]
16/02/01 22:36:55 INFO streaming.PipeMapRed: R/W/S=200000/884534/0
in:100000=200000/2 [rec/s] out:442267=884534/2 [rec/s]
16/02/01 22:36:56 INFO streaming.PipeMapRed: R/W/S=300000/1296452/
0 in:100000=300000/3 [rec/s] out:432150=1296452/3 [rec/s]
16/02/01 22:36:56 INFO streaming.PipeMapRed: MRErrorThread done
16/02/01 22:36:56 INFO streaming.PipeMapRed: mapRedFinished
16/02/01 22:36:56 INFO mapred.LocalJobRunner:
16/02/01 22:36:56 INFO mapred.MapTask: Starting flush of map outpu
t
16/02/01 22:36:56 INFO mapred.MapTask: Spilling map output
16/02/01 22:36:56 INFO mapred.MapTask: bufstart = 0; bufend = 1342
4739; bufvoid = 104857600
16/02/01 22:36:56 INFO mapred.MapTask: kvstart = 26214396(10485758
4); kvend = 20821152(83284608); length = 5393245/6553600
```

```
16/02/01 22:36:57 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/dunmireg/Documents/261HW/HW3/./combiner.py]
16/02/01 22:36:57 INFO Configuration.deprecation: mapred.skip.map.
auto.incr.proc.count is deprecated. Instead, use mapreduce.map.ski
p.proc-count.auto-incr
16/02/01 22:36:57 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/02/01 22:36:57 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/02/01 22:36:57 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/02/01 22:36:57 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/02/01 22:36:57 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/02/01 22:36:57 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/02/01 22:36:57 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/02/01 22:36:57 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/02/01 22:36:57 INFO streaming.PipeMapRed: R/W/S=400000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/02/01 22:36:57 INFO streaming.PipeMapRed: R/W/S=500000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/02/01 22:36:57 INFO streaming.PipeMapRed: R/W/S=600000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/02/01 22:36:57 INFO streaming.PipeMapRed: Records R/W=687448/1
16/02/01 22:36:57 INFO streaming.PipeMapRed: MRErrorThread done
16/02/01 22:36:57 INFO streaming.PipeMapRed: mapRedFinished
16/02/01 22:36:58 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/dunmireg/Documents/261HW/HW3/./combiner.py]
16/02/01 22:36:58 INFO Configuration.deprecation: mapred.skip.redu
ce.auto.incr.proc.count is deprecated. Instead, use mapreduce.redu
ce.skip.proc-count.auto-incr
16/02/01 22:36:58 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/02/01 22:36:58 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/02/01 22:36:58 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/02/01 22:36:58 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/02/01 22:36:58 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/02/01 22:36:58 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/02/01 22:36:58 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/02/01 22:36:58 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/02/01 22:36:58 INFO mapred.LocalJobRunner: Records R/W=687448/1
> sort
16/02/01 22:36:58 INFO streaming.PipeMapRed: R/W/S=400000/0/0 in:N
```



```
A [rec/s] out:NA [rec/s]
16/02/01 22:36:58 INFO streaming.PipeMapRed: R/W/S=500000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/02/01 22:36:59 INFO streaming.PipeMapRed: R/W/S=600000/0/0 in:6
00000=600000/1 [rec/s] out:0=0/1 [rec/s]
16/02/01 22:36:59 INFO mapreduce.Job: map 67% reduce 0%
16/02/01 22:36:59 INFO streaming.PipeMapRed: Records R/W=660864/1
16/02/01 22:36:59 INFO streaming.PipeMapRed: MRErrorThread done
16/02/01 22:36:59 INFO streaming.PipeMapRed: mapRedFinished
16/02/01 22:36:59 INFO mapred.MapTask: Finished spill 0
16/02/01 22:36:59 INFO mapred.Task: Task:attempt_local1588555268_0
001_m_000000_0 is done. And is in the process of committing
16/02/01 22:36:59 INFO mapred.LocalJobRunner: Records R/W=660864/1
16/02/01 22:36:59 INFO mapred.Task: Task 'attempt_local1588555268_
0001_m_000000_0' done.
16/02/01 22:36:59 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local1588555268_0001_m_000000_0
16/02/01 22:36:59 INFO mapred.LocalJobRunner: map task executor co
mplete.
16/02/01 22:36:59 INFO mapred.LocalJobRunner: Waiting for reduce t
asks
16/02/01 22:36:59 INFO mapred.LocalJobRunner: Starting task: attem
pt_local1588555268_0001_r_000000_0
16/02/01 22:36:59 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
16/02/01 22:36:59 INFO util.ProcfsBasedProcessTree: ProcfsBasedPro
cessTree currently is supported only on Linux.
16/02/01 22:36:59 INFO mapred.Task: Using ResourceCalculatorProce
ssTree : null
16/02/01 22:36:59 INFO mapred.ReduceTask: Using ShuffleConsumerPlu
gin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@448f7373
16/02/01 22:36:59 INFO reduce.MergeManagerImpl: MergerManager: mem
oryLimit=334338464, maxSingleShuffleLimit=83584616, mergeThreshold
=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
16/02/01 22:36:59 INFO reduce.EventFetcher: attempt_local158855526
8_0001_r_000000_0 Thread started: EventFetcher for fetching Map Co
mpletion Events
16/02/01 22:36:59 INFO reduce.LocalFetcher: localfetcher#1 about t
o shuffle output of map attempt_local1588555268_0001_m_000000_0 de
comp: 1201 len: 1205 to MEMORY
16/02/01 22:36:59 INFO reduce.InMemoryMapOutput: Read 1201 bytes f
rom map-output for attempt_local1588555268_0001_m_000000_0
16/02/01 22:36:59 INFO reduce.MergeManagerImpl: closeInMemoryFile
-> map-output of size: 1201, inMemoryMapOutputs.size() -> 1, commi
tMemory -> 0, usedMemory ->1201
16/02/01 22:36:59 INFO reduce.EventFetcher: EventFetcher is interr
upted.. Returning
16/02/01 22:36:59 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/01 22:36:59 INFO reduce.MergeManagerImpl: finalMerge called
with 1 in-memory map-outputs and 0 on-disk map-outputs
16/02/01 22:36:59 INFO mapred.Merger: Merging 1 sorted segments
16/02/01 22:36:59 INFO mapred.Merger: Down to the last merge-pass,
with 1 segments left of total size: 1197 bytes
16/02/01 22:36:59 INFO reduce.MergeManagerImpl: Merged 1 segments,
```

```
1201 bytes to disk to satisfy reduce memory limit
16/02/01 22:36:59 INFO reduce.MergeManagerImpl: Merging 1 files, 1
205 bytes from disk
16/02/01 22:36:59 INFO reduce.MergeManagerImpl: Merging 0 segment
s, 0 bytes from memory into reduce
16/02/01 22:36:59 INFO mapred.Merger: Merging 1 sorted segments
16/02/01 22:36:59 INFO mapred.Merger: Down to the last merge-pass,
with 1 segments left of total size: 1197 bytes
16/02/01 22:36:59 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/01 22:36:59 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/dunmireg/Documents/261HW/HW3/./reducer.py]
16/02/01 22:36:59 INFO Configuration.deprecation: mapred.job.track
er is deprecated. Instead, use mapreduce.jobtracker.address
16/02/01 22:36:59 INFO Configuration.deprecation: mapred.map.tasks
is deprecated. Instead, use mapreduce.job.maps
16/02/01 22:36:59 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/02/01 22:36:59 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/02/01 22:36:59 INFO streaming.PipeMapRed: Records R/W=84/1
16/02/01 22:36:59 INFO streaming.PipeMapRed: MRErrorThread done
16/02/01 22:36:59 INFO streaming.PipeMapRed: mapRedFinished
16/02/01 22:36:59 INFO mapred.Task: Task:attempt_local1588555268_0
001_r_000000_0 is done. And is in the process of committing
16/02/01 22:36:59 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/01 22:36:59 INFO mapred.Task: Task attempt_local1588555268_0
001_r_000000_0 is allowed to commit now
16/02/01 22:36:59 INFO output.FileOutputCommitter: Saved output of
task 'attempt_local1588555268_0001_r_000000_0' to hdfs://localhos
t:9000/user/dunmireg/word_count/_temporary/0/task_local1588555268_
0001_r_000000
16/02/01 22:36:59 INFO mapred.LocalJobRunner: Records R/W=84/1 > r
educe
16/02/01 22:36:59 INFO mapred.Task: Task 'attempt_local1588555268_
0001_r_000000_0' done.
16/02/01 22:36:59 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local1588555268_0001_r_000000_0
16/02/01 22:36:59 INFO mapred.LocalJobRunner: Starting task: attem
pt_local1588555268_0001_r_000001_0
16/02/01 22:36:59 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
16/02/01 22:36:59 INFO util.ProcfsBasedProcessTree: ProcfsBasedPro
cessTree currently is supported only on Linux.
16/02/01 22:36:59 INFO mapred.Task: Using ResourceCalculatorProce
ssTree : null
16/02/01 22:36:59 INFO mapred.ReduceTask: Using ShuffleConsumerPlu
gin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@71770ed3
16/02/01 22:36:59 INFO reduce.MergeManagerImpl: MergerManager: mem
oryLimit=334338464, maxSingleShuffleLimit=83584616, mergeThreshold
=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
16/02/01 22:36:59 INFO reduce.EventFetcher: attempt_local158855526
8_0001_r_000001_0 Thread started: EventFetcher for fetching Map Co
mpletion Events
16/02/01 22:36:59 INFO reduce.LocalFetcher: localfetcher#2 about t
```

```
o shuffle output of map attempt_local1588555268_0001_m_000000_0 de
comp: 1335 len: 1339 to MEMORY
16/02/01 22:36:59 INFO reduce.InMemoryMapOutput: Read 1335 bytes f
rom map-output for attempt_local1588555268_0001_m_000000_0
16/02/01 22:36:59 INFO reduce.MergeManagerImpl: closeInMemoryFile
-> map-output of size: 1335, inMemoryMapOutputs.size() -> 1, commi
tMemory -> 0, usedMemory ->1335
16/02/01 22:36:59 INFO reduce.EventFetcher: EventFetcher is interr
upted.. Returning
16/02/01 22:36:59 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/01 22:36:59 INFO reduce.MergeManagerImpl: finalMerge called
with 1 in-memory map-outputs and 0 on-disk map-outputs
16/02/01 22:36:59 INFO mapred.Merger: Merging 1 sorted segments
16/02/01 22:36:59 INFO mapred.Merger: Down to the last merge-pass,
with 1 segments left of total size: 1326 bytes
16/02/01 22:36:59 INFO reduce.MergeManagerImpl: Merged 1 segments,
1335 bytes to disk to satisfy reduce memory limit
16/02/01 22:36:59 INFO reduce.MergeManagerImpl: Merging 1 files, 1
339 bytes from disk
16/02/01 22:36:59 INFO reduce.MergeManagerImpl: Merging 0 segment
s, 0 bytes from memory into reduce
16/02/01 22:36:59 INFO mapred.Merger: Merging 1 sorted segments
16/02/01 22:36:59 INFO mapred.Merger: Down to the last merge-pass,
with 1 segments left of total size: 1326 bytes
16/02/01 22:36:59 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/01 22:36:59 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/dunmireg/Documents/261HW/HW3/./reducer.py]
16/02/01 22:36:59 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/02/01 22:36:59 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/02/01 22:36:59 INFO streaming.PipeMapRed: Records R/W=91/1
16/02/01 22:36:59 INFO streaming.PipeMapRed: MRErrorThread done
16/02/01 22:36:59 INFO streaming.PipeMapRed: mapRedFinished
16/02/01 22:36:59 INFO mapred.Task: Task:attempt_local1588555268_0
001_r_000001_0 is done. And is in the process of committing
16/02/01 22:36:59 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/01 22:36:59 INFO mapred.Task: Task attempt_local1588555268_0
001_r_000001_0 is allowed to commit now
16/02/01 22:36:59 INFO output.FileOutputCommitter: Saved output of
task 'attempt_local1588555268_0001_r_000001_0' to hdfs://localhos
t:9000/user/dunmireg/word_count/_temporary/0/task_local1588555268_
0001_r_000001
16/02/01 22:36:59 INFO mapred.LocalJobRunner: Records R/W=91/1 > r
educer
16/02/01 22:36:59 INFO mapred.Task: Task 'attempt_local1588555268_
0001_r_000001_0' done.
16/02/01 22:36:59 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local1588555268_0001_r_000001_0
16/02/01 22:36:59 INFO mapred.LocalJobRunner: reduce task executor
complete.
16/02/01 22:37:00 INFO mapreduce.Job: map 100% reduce 100%
16/02/01 22:37:00 INFO mapreduce.Job: Job job_local1588555268_0001
completed successfully
```

16/02/01 22:37:00 INFO mapreduce.Job: Counters: 38

File System Counters

FILE: Number of bytes read=328491  
FILE: Number of bytes written=1217204  
FILE: Number of read operations=0  
FILE: Number of large read operations=0  
FILE: Number of write operations=0  
HDFS: Number of bytes read=152719458  
HDFS: Number of bytes written=3213  
HDFS: Number of read operations=24  
HDFS: Number of large read operations=0  
HDFS: Number of write operations=9

Map-Reduce Framework

Map input records=312913  
Map output records=1348312  
Map output bytes=13424739  
Map output materialized bytes=2544  
Input split bytes=111  
Combine input records=1348312  
Combine output records=175  
Reduce input groups=175  
Reduce shuffle bytes=2544  
Reduce input records=175  
Reduce output records=175  
Spilled Records=350  
Shuffled Maps =2  
Failed Shuffles=0  
Merged Map outputs=2  
GC time elapsed (ms)=5  
Total committed heap usage (bytes)=813170688

Combiner-Counter

Total=2

Map-Count

Total=1

Reduce-Counter

Total=2

Shuffle Errors

BAD\_ID=0  
CONNECTION=0  
IO\_ERROR=0  
WRONG\_LENGTH=0  
WRONG\_MAP=0  
WRONG\_REDUCE=0

File Input Format Counters

Bytes Read=50906486

File Output Format Counters

Bytes Written=2182

16/02/01 22:37:00 INFO streaming.StreamJob: Output directory: word\_count

**RESULTS: Mapper - 1, Combiner - 2, Reducer - 2**

## HW3.2 Part 4

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

### HW3.2 Part 4 Stage 1 Map Function

```
In [2]: %%writefile mapper-3-2-4.py
#!/usr/bin/python
import sys
import re
from csv import reader

#Structure of complaints
#Complaint ID,Product,Sub-product,Issue,Sub-issue,State,ZIP code,Su
bmitted via,Date received,Date sent to company,
#Company,Company response,Timely response?,Consumer disputed?

line_num = 0 #for skipping header
total = 0 #total number of words in issue
WORD_RE = re.compile(r"[\w']+")
for line in reader(sys.stdin): #here we use csv.reader to read the
input of the file
    if line_num == 0: #skip first row, which is a header
        line_num += 1
        continue
    else:
        issue = line[3] #parse the issue of the complaint
        if issue == '': #There are exactly four records where the i
ssue was marked as blank.
            issue = 'Blank' #We felt that setting to blank was appr
opriate
        words = re.findall(WORD_RE, issue)
        for word in words:
            total += 1 #increment total word counter
            print word.lower() + '\t' + str(1) #print the word and
a count of 1
print '*' + '\t' + str(total) #use order inversion to provide total
as first input to reducer
```

Overwriting mapper-3-2-4.py

### HW3.2 Part 4 Stage 1 Reduce Function

In [7]:

```

%%writefile reducer-3-2-4.py
#!/usr/bin/python
import sys
import operator

current_word = None #follows same basic structure as word count we
have worked with before
current_count = None
word = None
total = 0
#wordcount = {} #a dictionary to store counts. This was used in an
earlier version using an in-memory mapper

for line in sys.stdin:
    line = line.split('\t') #split line
    word = line[0] #read word
    count = int(line[1]) #get count
    if word == '*':
        total = count #if the word is * we know this is the total n
umber of words and set this as a field
    else: #otherwise continue as normal
        if current_word == word:
            current_count += count
        else:
            if current_word:
                #wordcount[current_word] = current_count #used in i
n-memory dictionary version
                #structure of result is word + count + relative cou
nt
                print current_word + '\t' + str(current_count) +
'\t' + str(float(current_count)/total) #print result
            current_word = word
            current_count = count

#print last word
if current_word == word:
    #wordcount[current_word] = current_count
    print current_word + '\t' + str(current_count) + '\t' + str(flo
at(current_count)/total)

#Code for in-memory dictionary printing
# largest = 50
# smallest = 10
# sortedWordCount = sorted(wordcount.items(), key = operator.itemge
tter(1))

# print "The Top 50 terms are"
# for i in range(largest):
#     print str(sortedWordCount[-i-1][0]) + '\t' + str(sortedWordCo
unt[-i-1][1]) + '\t' + str(float(sortedWordCount[-i-1][1])/total)

# print '\n'

# print "The bottom 10 terms are"

```

```
# for i in range(smallest):
#     print str(sortedWordCount[i][0]) + '\t' + str(sortedWordCount[i][1]) + '\t' + str(float(sortedWordCount[i][1])/total)
```

Overwriting reducer-3-2-4.py

## HW3.2 Part 4 Stage 2 Map Function

```
In [3]: %%writefile mapper-3-2-4-2.py
#!/usr/bin/python
import sys

#Structure is word + \t + count + \t + relative count, to be read from output of first set of jobs

for line in sys.stdin:
    line = line.strip()
    line = line.split('\t') #split the line
    print line[1] + '\t' + line[0] + '\t' + line[2] #now we are using the number as the key, this will be sorted
```

Overwriting mapper-3-2-4-2.py

## HW3.2 Part 4 Stage 2 Reduce Function

```
In [10]: %%writefile reducer-3-2-4-2.py
#!/usr/bin/python
import sys

print "Issue" + '\t' + "Count" + '\t' + "Relative Count"
for line in sys.stdin: #reads result of second mapper using number as key
    line = line.strip()
    line = line.split('\t')
    #structure currently is count + word + relative count
    print line[1] + '\t' + line[0] + '\t' + line[2] #now reverse back to display word + count + relative count
```

Overwriting reducer-3-2-4-2.py

```
In [11]: #!/cat Consumer_Complaints.csv | python mapper-3-2-4.py | sort | python reducer-3-2-4.py | python mapper-3-2-4-2.py | sort -n | python reducer-3-2-4-2.py > output.txt
```

## HW3.2 Part 4 Hadoop MapReduce Stage 1 - Word Counts and Frequencies



```
In [14]: #run hadoop job
!hadoop jar /usr/local/Cellar/hadoop/2.7.1/libexec/share/hadoop/tools/lib/hadoop-streaming-2.7.1.jar \
-mapper mapper-3-2-4.py \
-reducer reducer-3-2-4.py \
-input Consumer_Complaints.csv \
-output word_count
```



```
16/01/31 23:01:12 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/01/31 23:01:12 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
16/01/31 23:01:12 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/01/31 23:01:12 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
16/01/31 23:01:13 INFO mapred.FileInputFormat: Total input paths to process : 1
16/01/31 23:01:13 INFO mapreduce.JobSubmitter: number of splits:1
16/01/31 23:01:13 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1964160718_0001
16/01/31 23:01:13 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/01/31 23:01:13 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/01/31 23:01:13 INFO mapreduce.Job: Running job: job_local1964160718_0001
16/01/31 23:01:13 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
16/01/31 23:01:13 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/31 23:01:13 INFO mapred.LocalJobRunner: Waiting for map tasks
16/01/31 23:01:13 INFO mapred.LocalJobRunner: Starting task: attempt_local1964160718_0001_m_000000_0
16/01/31 23:01:13 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/31 23:01:13 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
16/01/31 23:01:13 INFO mapred.Task: Using ResourceCalculatorProcessTree : null
16/01/31 23:01:13 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/dunmireg/Consumer_Complaints.csv:0+50906486
16/01/31 23:01:13 INFO mapred.MapTask: numReduceTasks: 1
16/01/31 23:01:14 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/01/31 23:01:14 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/01/31 23:01:14 INFO mapred.MapTask: soft limit at 83886080
16/01/31 23:01:14 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/01/31 23:01:14 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/01/31 23:01:14 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
16/01/31 23:01:14 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/dunmireg/Documents/261HW/HW3/./mapper-3-2-4.py]
16/01/31 23:01:14 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
16/01/31 23:01:14 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir
```

```
16/01/31 23:01:14 INFO Configuration.deprecation: map.input.file i
s deprecated. Instead, use mapreduce.map.input.file
16/01/31 23:01:14 INFO Configuration.deprecation: mapred.skip.on i
s deprecated. Instead, use mapreduce.job.skiprecords
16/01/31 23:01:14 INFO Configuration.deprecation: map.input.length
is deprecated. Instead, use mapreduce.map.input.length
16/01/31 23:01:14 INFO Configuration.deprecation: mapred.work.outp
ut.dir is deprecated. Instead, use mapreduce.task.output.dir
16/01/31 23:01:14 INFO Configuration.deprecation: map.input.start
is deprecated. Instead, use mapreduce.map.input.start
16/01/31 23:01:14 INFO Configuration.deprecation: mapred.job.id is
deprecated. Instead, use mapreduce.job.id
16/01/31 23:01:14 INFO Configuration.deprecation: user.name is dep
recated. Instead, use mapreduce.job.user.name
16/01/31 23:01:14 INFO Configuration.deprecation: mapred.task.is.m
ap is deprecated. Instead, use mapreduce.task.ismap
16/01/31 23:01:14 INFO Configuration.deprecation: mapred.task.id i
s deprecated. Instead, use mapreduce.task.attempt.id
16/01/31 23:01:14 INFO Configuration.deprecation: mapred.task.part
ition is deprecated. Instead, use mapreduce.task.partition
16/01/31 23:01:14 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/01/31 23:01:14 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/31 23:01:14 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/31 23:01:14 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/31 23:01:14 INFO streaming.PipeMapRed: Records R/W=1513/1
16/01/31 23:01:14 INFO streaming.PipeMapRed: R/W/S=10000/41100/0 i
n:NA [rec/s] out:NA [rec/s]
16/01/31 23:01:14 INFO mapreduce.Job: Job job_local1964160718_0001
running in uber mode : false
16/01/31 23:01:14 INFO mapreduce.Job: map 0% reduce 0%
16/01/31 23:01:15 INFO streaming.PipeMapRed: R/W/S=100000/448120/0
in:100000=100000/1 [rec/s] out:448120=448120/1 [rec/s]
16/01/31 23:01:16 INFO streaming.PipeMapRed: R/W/S=200000/884534/0
in:100000=200000/2 [rec/s] out:442267=884534/2 [rec/s]
16/01/31 23:01:17 INFO streaming.PipeMapRed: R/W/S=300000/1296452/
0 in:100000=300000/3 [rec/s] out:432150=1296452/3 [rec/s]
16/01/31 23:01:18 INFO streaming.PipeMapRed: MRErrorThread done
16/01/31 23:01:18 INFO streaming.PipeMapRed: mapRedFinished
16/01/31 23:01:18 INFO mapred.LocalJobRunner:
16/01/31 23:01:18 INFO mapred.MapTask: Starting flush of map output
t
16/01/31 23:01:18 INFO mapred.MapTask: Spilling map output
16/01/31 23:01:18 INFO mapred.MapTask: bufstart = 0; bufend = 1342
4749; bufvoid = 104857600
16/01/31 23:01:18 INFO mapred.MapTask: kvstart = 26214396(10485758
4); kvend = 20821148(83284592); length = 5393249/6553600
16/01/31 23:01:19 INFO mapred.MapTask: Finished spill 0
16/01/31 23:01:19 INFO mapred.Task: Task:attempt_local1964160718_0
001_m_000000_0 is done. And is in the process of committing
16/01/31 23:01:19 INFO mapred.LocalJobRunner: Records R/W=1513/1
```

```
16/01/31 23:01:19 INFO mapred.Task: Task 'attempt_local1964160718_0001_m_000000_0' done.
16/01/31 23:01:19 INFO mapred.LocalJobRunner: Finishing task: attempt_local1964160718_0001_m_000000_0
16/01/31 23:01:19 INFO mapred.LocalJobRunner: map task executor complete.
16/01/31 23:01:19 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/01/31 23:01:19 INFO mapred.LocalJobRunner: Starting task: attempt_local1964160718_0001_r_000000_0
16/01/31 23:01:19 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/31 23:01:19 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
16/01/31 23:01:19 INFO mapred.Task: Using ResourceCalculatorProcessTree : null
16/01/31 23:01:19 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@79e9b608
16/01/31 23:01:19 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLimit=83584616, mergeThreshold=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
16/01/31 23:01:19 INFO reduce.EventFetcher: attempt_local1964160718_0001_r_000000_0 Thread started: EventFetcher for fetching Map Completion Events
16/01/31 23:01:19 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1964160718_0001_m_000000_0 decomp: 16121377 len: 16121381 to MEMORY
16/01/31 23:01:19 INFO reduce.InMemoryMapOutput: Read 16121377 bytes from map-output for attempt_local1964160718_0001_m_000000_0
16/01/31 23:01:19 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 16121377, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory -> 16121377
16/01/31 23:01:19 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/01/31 23:01:19 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/31 23:01:19 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs
16/01/31 23:01:19 INFO mapred.Merger: Merging 1 sorted segments
16/01/31 23:01:19 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 16121373 bytes
16/01/31 23:01:19 INFO reduce.MergeManagerImpl: Merged 1 segments, 16121377 bytes to disk to satisfy reduce memory limit
16/01/31 23:01:19 INFO reduce.MergeManagerImpl: Merging 1 files, 16121381 bytes from disk
16/01/31 23:01:19 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/01/31 23:01:19 INFO mapred.Merger: Merging 1 sorted segments
16/01/31 23:01:19 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 16121373 bytes
16/01/31 23:01:19 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/31 23:01:19 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/dunmireg/Documents/261HW/HW3/./reducer-3-2-4.py]
16/01/31 23:01:19 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
```

```
16/01/31 23:01:19 INFO Configuration.deprecation: mapred.map.tasks
is deprecated. Instead, use mapreduce.job.maps
16/01/31 23:01:19 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/01/31 23:01:19 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/31 23:01:19 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/31 23:01:19 INFO mapreduce.Job: map 100% reduce 0%
16/01/31 23:01:19 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/31 23:01:19 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/31 23:01:20 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/01/31 23:01:20 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/01/31 23:01:20 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/01/31 23:01:20 INFO streaming.PipeMapRed: R/W/S=400000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/01/31 23:01:20 INFO streaming.PipeMapRed: R/W/S=500000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/01/31 23:01:20 INFO streaming.PipeMapRed: R/W/S=600000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/01/31 23:01:20 INFO streaming.PipeMapRed: R/W/S=700000/0/0 in:7
00000=700000/1 [rec/s] out:0=0/1 [rec/s]
16/01/31 23:01:20 INFO streaming.PipeMapRed: R/W/S=800000/0/0 in:8
00000=800000/1 [rec/s] out:0=0/1 [rec/s]
16/01/31 23:01:21 INFO streaming.PipeMapRed: R/W/S=900000/0/0 in:9
00000=900000/1 [rec/s] out:0=0/1 [rec/s]
16/01/31 23:01:21 INFO streaming.PipeMapRed: R/W/S=1000000/0/0 in:
1000000=1000000/1 [rec/s] out:0=0/1 [rec/s]
16/01/31 23:01:21 INFO streaming.PipeMapRed: R/W/S=1100000/0/0 in:
1100000=1100000/1 [rec/s] out:0=0/1 [rec/s]
16/01/31 23:01:21 INFO streaming.PipeMapRed: R/W/S=1200000/0/0 in:
1200000=1200000/1 [rec/s] out:0=0/1 [rec/s]
16/01/31 23:01:21 INFO streaming.PipeMapRed: R/W/S=1300000/0/0 in:
1300000=1300000/1 [rec/s] out:0=0/1 [rec/s]
16/01/31 23:01:21 INFO streaming.PipeMapRed: MRErrorThread done
16/01/31 23:01:21 INFO streaming.PipeMapRed: Records R/W=1348313/1
16/01/31 23:01:21 INFO streaming.PipeMapRed: mapRedFinished
16/01/31 23:01:21 INFO mapred.Task: Task:attempt_local1964160718_0
001_r_000000_0 is done. And is in the process of committing
16/01/31 23:01:21 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/31 23:01:21 INFO mapred.Task: Task attempt_local1964160718_0
001_r_000000_0 is allowed to commit now
16/01/31 23:01:21 INFO output.FileOutputCommitter: Saved output of
task 'attempt_local1964160718_0001_r_000000_0' to hdfs://localhos
t:9000/user/dunmireg/word_count/_temporary/0/task_local1964160718_
0001_r_000000
16/01/31 23:01:21 INFO mapred.LocalJobRunner: Records R/W=1348313/
1 > reduce
16/01/31 23:01:21 INFO mapred.Task: Task 'attempt_local1964160718_
```

```

0001_r_000000_0' done.
16/01/31 23:01:21 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local1964160718_0001_r_000000_0
16/01/31 23:01:21 INFO mapred.LocalJobRunner: reduce task executor
complete.
16/01/31 23:01:21 INFO mapreduce.Job: map 100% reduce 100%
16/01/31 23:01:21 INFO mapreduce.Job: Job job_local1964160718_0001
completed successfully
16/01/31 23:01:21 INFO mapreduce.Job: Counters: 35
    File System Counters
        FILE: Number of bytes read=32454892
        FILE: Number of bytes written=49166017
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=101812972
        HDFS: Number of bytes written=5182
        HDFS: Number of read operations=13
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=4
    Map-Reduce Framework
        Map input records=312913
        Map output records=1348313
        Map output bytes=13424749
        Map output materialized bytes=16121381
        Input split bytes=111
        Combine input records=0
        Combine output records=0
        Reduce input groups=176
        Reduce shuffle bytes=16121381
        Reduce input records=1348313
        Reduce output records=175
        Spilled Records=2696626
        Shuffled Maps =1
        Failed Shuffles=0
        Merged Map outputs=1
        GC time elapsed (ms)=4
        Total committed heap usage (bytes)=541065216
    Shuffle Errors
        BAD_ID=0
        CONNECTION=0
        IO_ERROR=0
        WRONG_LENGTH=0
        WRONG_MAP=0
        WRONG_REDUCE=0
    File Input Format Counters
        Bytes Read=50906486
    File Output Format Counters
        Bytes Written=5182
16/01/31 23:01:21 INFO streaming.StreamJob: Output directory: word
_count

```

## HW3.2 Part 4 Hadoop MapReduce Stage 2 - Final Sort



```
In [15]: #run second job to properly sort by counts
!hadoop jar /usr/local/Cellar/hadoop/2.7.1/libexec/share/hadoop/tools/lib/hadoop-streaming-2.7.1.jar \
-D stream.num.map.output.key.fields=2 \
-D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator \
-D mapred.text.key.comparator.options="-k1nr -k2" \
-mapper mapper-3-2-4-2.py \
-reducer reducer-3-2-4-2.py \
-input /user/dunmireg/word_count/part-00000 \
-output sortedWordCount
```



```
16/01/31 23:01:45 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/01/31 23:01:45 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
16/01/31 23:01:45 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/01/31 23:01:45 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
16/01/31 23:01:45 INFO mapred.FileInputFormat: Total input paths to process : 1
16/01/31 23:01:45 INFO mapreduce.JobSubmitter: number of splits:1
16/01/31 23:01:46 INFO Configuration.deprecation: mapred.text.key.comparator.options is deprecated. Instead, use mapreduce.partition.keycomparator.options
16/01/31 23:01:46 INFO Configuration.deprecation: mapred.output.key.comparator.class is deprecated. Instead, use mapreduce.job.output.key.comparator.class
16/01/31 23:01:46 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1842258814_0001
16/01/31 23:01:46 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/01/31 23:01:46 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
16/01/31 23:01:46 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/01/31 23:01:46 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/31 23:01:46 INFO mapreduce.Job: Running job: job_local1842258814_0001
16/01/31 23:01:46 INFO mapred.LocalJobRunner: Waiting for map tasks
16/01/31 23:01:46 INFO mapred.LocalJobRunner: Starting task: attempt_local1842258814_0001_m_000000_0
16/01/31 23:01:46 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/31 23:01:46 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
16/01/31 23:01:46 INFO mapred.Task: Using ResourceCalculatorProcessTree : null
16/01/31 23:01:46 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/dunmireg/word_count/part-00000:0+5182
16/01/31 23:01:46 INFO mapred.MapTask: numReduceTasks: 1
16/01/31 23:01:46 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/01/31 23:01:46 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/01/31 23:01:46 INFO mapred.MapTask: soft limit at 83886080
16/01/31 23:01:46 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/01/31 23:01:46 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/01/31 23:01:46 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
```

```
16/01/31 23:01:46 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/dunmireg/Documents/261HW/HW3/./mapper-3-2-4-2.py]
16/01/31 23:01:46 INFO Configuration.deprecation: mapred.tip.id is
deprecated. Instead, use mapreduce.task.id
16/01/31 23:01:46 INFO Configuration.deprecation: mapred.local.dir
is deprecated. Instead, use mapreduce.cluster.local.dir
16/01/31 23:01:46 INFO Configuration.deprecation: map.input.file i
s deprecated. Instead, use mapreduce.map.input.file
16/01/31 23:01:46 INFO Configuration.deprecation: mapred.skip.on i
s deprecated. Instead, use mapreduce.job.skiprecords
16/01/31 23:01:46 INFO Configuration.deprecation: map.input.length
is deprecated. Instead, use mapreduce.map.input.length
16/01/31 23:01:46 INFO Configuration.deprecation: mapred.work.outp
ut.dir is deprecated. Instead, use mapreduce.task.output.dir
16/01/31 23:01:46 INFO Configuration.deprecation: map.input.start
is deprecated. Instead, use mapreduce.map.input.start
16/01/31 23:01:46 INFO Configuration.deprecation: mapred.job.id is
deprecated. Instead, use mapreduce.job.id
16/01/31 23:01:46 INFO Configuration.deprecation: user.name is dep
recated. Instead, use mapreduce.job.user.name
16/01/31 23:01:46 INFO Configuration.deprecation: mapred.task.is.m
ap is deprecated. Instead, use mapreduce.task.ismap
16/01/31 23:01:46 INFO Configuration.deprecation: mapred.task.id i
s deprecated. Instead, use mapreduce.task.attempt.id
16/01/31 23:01:46 INFO Configuration.deprecation: mapred.task.part
ition is deprecated. Instead, use mapreduce.task.partition
16/01/31 23:01:46 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/01/31 23:01:46 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/31 23:01:46 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/31 23:01:46 INFO streaming.PipeMapRed: MRErrorThread done
16/01/31 23:01:46 INFO streaming.PipeMapRed: Records R/W=175/1
16/01/31 23:01:46 INFO streaming.PipeMapRed: mapRedFinished
16/01/31 23:01:46 INFO mapred.LocalJobRunner:
16/01/31 23:01:46 INFO mapred.MapTask: Starting flush of map output
t
16/01/31 23:01:46 INFO mapred.MapTask: Spilling map output
16/01/31 23:01:46 INFO mapred.MapTask: bufstart = 0; bufend = 518
2; bufvoid = 104857600
16/01/31 23:01:46 INFO mapred.MapTask: kvstart = 26214396(10485758
4); kvend = 26213700(104854800); length = 697/6553600
16/01/31 23:01:46 INFO mapred.MapTask: Finished spill 0
16/01/31 23:01:46 INFO mapred.Task: Task:attempt_local1842258814_0
001_m_000000_0 is done. And is in the process of committing
16/01/31 23:01:46 INFO mapred.LocalJobRunner: Records R/W=175/1
16/01/31 23:01:46 INFO mapred.Task: Task 'attempt_local1842258814_
0001_m_000000_0' done.
16/01/31 23:01:46 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local1842258814_0001_m_000000_0
16/01/31 23:01:46 INFO mapred.LocalJobRunner: map task executor co
mplete.
16/01/31 23:01:46 INFO mapred.LocalJobRunner: Waiting for reduce t
```

```
asks
16/01/31 23:01:46 INFO mapred.LocalJobRunner: Starting task: attempt_local1842258814_0001_r_000000_0
16/01/31 23:01:46 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/31 23:01:46 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
16/01/31 23:01:46 INFO mapred.Task: Using ResourceCalculatorProcessTree : null
16/01/31 23:01:46 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@562f392a
16/01/31 23:01:46 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLimit=83584616, mergeThreshold=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
16/01/31 23:01:46 INFO reduce.EventFetcher: attempt_local1842258814_0001_r_000000_0 Thread started: EventFetcher for fetching Map Completion Events
16/01/31 23:01:46 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1842258814_0001_m_000000_0 decomp: 5534 len: 5538 to MEMORY
16/01/31 23:01:46 INFO reduce.InMemoryMapOutput: Read 5534 bytes from map-output for attempt_local1842258814_0001_m_000000_0
16/01/31 23:01:46 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 5534, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory -> 5534
16/01/31 23:01:46 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/01/31 23:01:46 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/31 23:01:46 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs
16/01/31 23:01:46 INFO mapred.Merger: Merging 1 sorted segments
16/01/31 23:01:46 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 5530 bytes
16/01/31 23:01:46 INFO reduce.MergeManagerImpl: Merged 1 segments, 5534 bytes to disk to satisfy reduce memory limit
16/01/31 23:01:46 INFO reduce.MergeManagerImpl: Merging 1 files, 5538 bytes from disk
16/01/31 23:01:46 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/01/31 23:01:46 INFO mapred.Merger: Merging 1 sorted segments
16/01/31 23:01:46 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 5530 bytes
16/01/31 23:01:46 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/31 23:01:46 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/dunmireg/Documents/261HW/HW3/./reducer-3-2-4-2.py]
16/01/31 23:01:46 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
16/01/31 23:01:46 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
16/01/31 23:01:46 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/31 23:01:46 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/31 23:01:46 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
```

```
[rec/s] out:NA [rec/s]
16/01/31 23:01:46 INFO streaming.PipeMapRed: Records R/W=175/1
16/01/31 23:01:46 INFO streaming.PipeMapRed: MRErrorThread done
16/01/31 23:01:46 INFO streaming.PipeMapRed: mapRedFinished
16/01/31 23:01:46 INFO mapred.Task: Task:attempt_local1842258814_0
001_r_000000_0 is done. And is in the process of committing
16/01/31 23:01:46 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/31 23:01:46 INFO mapred.Task: Task attempt_local1842258814_0
001_r_000000_0 is allowed to commit now
16/01/31 23:01:46 INFO output.FileOutputCommitter: Saved output of
task 'attempt_local1842258814_0001_r_000000_0' to hdfs://localhos
t:9000/user/dunmireg/sortedWordCount/_temporary/0/task_local184225
8814_0001_r_000000
16/01/31 23:01:46 INFO mapred.LocalJobRunner: Records R/W=175/1 >
reduce
16/01/31 23:01:46 INFO mapred.Task: Task 'attempt_local1842258814_
0001_r_000000_0' done.
16/01/31 23:01:46 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local1842258814_0001_r_000000_0
16/01/31 23:01:46 INFO mapred.LocalJobRunner: reduce task executor
complete.
16/01/31 23:01:47 INFO mapreduce.Job: Job job_local1842258814_0001
running in uber mode : false
16/01/31 23:01:47 INFO mapreduce.Job: map 100% reduce 100%
16/01/31 23:01:47 INFO mapreduce.Job: Job job_local1842258814_0001
completed successfully
16/01/31 23:01:47 INFO mapreduce.Job: Counters: 35
    File System Counters
        FILE: Number of bytes read=223198
        FILE: Number of bytes written=820648
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=10364
        HDFS: Number of bytes written=5209
        HDFS: Number of read operations=13
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=4
    Map-Reduce Framework
        Map input records=175
        Map output records=175
        Map output bytes=5182
        Map output materialized bytes=5538
        Input split bytes=109
        Combine input records=0
        Combine output records=0
        Reduce input groups=116
        Reduce shuffle bytes=5538
        Reduce input records=175
        Reduce output records=176
        Spilled Records=350
        Shuffled Maps =1
        Failed Shuffles=0
        Merged Map outputs=1
```

```
GC time elapsed (ms)=5
Total committed heap usage (bytes)=511705088
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=5182
File Output Format Counters
  Bytes Written=5209
16/01/31 23:01:47 INFO streaming.StreamJob: Output directory: sortedWordCount
```

## HW3.2 Part 4 Results

```
In [ ]: #show results, display top 50 words
!hdfs dfs -cat /user/dunmireg/sortedWordCount/part-00000 | head -50
```

16/02/02 23:03:01 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

loan 119630 0.0887257548698

collection 72394 0.0536923204718

foreclosure 70487 0.0522779594041

modification 70487 0.0522779594041

account 57448 0.0426073490409

credit 55251 0.0409779042239

or 40508 0.0300434914174

payments 39993 0.0296615323456

escrow 36767 0.0272689110532

servicing 36767 0.0272689110532

report 34903 0.0258864417138

incorrect 29133 0.0216070167736

information 29069 0.0215595500151

on 29069 0.0215595500151

debt 27874 0.020673256635

closing 19000 0.0140916939106

not 18477 0.013703801494

attempts 17972 0.0133292591032

collect 17972 0.0133292591032

cont'd 17972 0.0133292591032

owed 17972 0.0133292591032

and 16448 0.012198956918

management 16205 0.0120187315695

opening 16205 0.0120187315695

of 13983 0.0103707450501



my 10731 0.00795884038709

deposits 10555 0.00782830680139

withdrawals 10555 0.00782830680139

problems 9484 0.0070339802657

application 8868 0.00657711271575

communication 8671 0.00643100409994

tactics 8671 0.00643100409994

broker 8625 0.00639688736732

mortgage 8625 0.00639688736732

originator 8625 0.00639688736732

to 8401 0.00623075371279

unable 8178 0.00606536172637

billing 8158 0.00605052836435

other 7886 0.005848794641

disclosure 7655 0.00567746930977

verification 7655 0.00567746930977

disputes 6938 0.00514569328167

reporting 6559 0.00486460107156

lease 6337 0.00469995075324

the 6248 0.00463394229229

being 5663 0.00420006645346

by 5663 0.00420006645346

caused 5663 0.00420006645346

funds 5663 0.00420006645346

low 5663 0.00420006645346

```
In [ ]: #Display bottom 10 counts (least is on top)  
!hdfs dfs -cat /user/dunmireg/sortedWordCount/part-00000 | tail | s  
ort -k2
```

16/02/02 23:06:12 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

blank 4 2.96667240223e-06

disclosures 64 4.74667584357e-05

missing 64 4.74667584357e-05

amt 71 5.26584351396e-05

day 71 5.26584351396e-05

checks 75 5.56251075419e-05

convenience 75 5.56251075419e-05

credited 92 6.82334652514e-05

payment 92 6.82334652514e-05

amount 98 7.26834738547e-05

## HW3.3. Shopping Cart Analysis

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

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

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

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

Here are the first few lines of the ProductPurchaseData

```

FRO11987 ELE17451 ELE89019 SNA90258 GRO99222
GRO99222 GRO12298 FRO12685 ELE91550 SNA11465 ELE26917 ELE52966 FRO90334
SNA30755 ELE17451 FRO84225 SNA80192
ELE17451 GRO73461 DAI22896 SNA99873 FRO86643
ELE17451 ELE37798 FRO86643 GRO56989 ELE23393 SNA11465
ELE17451 SNA69641 FRO86643 FRO78087 SNA11465 GRO39357 ELE28573 ELE11375
DAI54444

```

Do some exploratory data analysis of this dataset.

How many unique items are available from this supplier?

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

## HW3.3 Mapper

Below is the mapper used to accomplish this task. It emits each token along with a basket size to the reducers for additional processing.

```
In [73]: %%writefile mapperQ33.py
#!/usr/bin/env python

import sys
from collections import defaultdict
from itertools import combinations

def readInput(file, separator=None):
    for line in file:
        yield line.split(separator)

if __name__ == "__main__":

    # Read input
    data = readInput(sys.stdin)
    for line in data:
        holdingDict = defaultdict(int)
        basketSize = len(line)

        # Append elements
        for token in line:
            holdingDict[token] += 1

        # Emit results
        for k, v in holdingDict.iteritems():
            basketInfo = str([v, basketSize])
            print '%s%s%s' % (k, '\t', basketInfo)

    # Update counter
    sys.stderr.write("reporter:counter:User-Defined,Number of M
appers,1\n")
```

Overwriting mapperQ33.py

## HW3.3 Reducer

Below is the reducer used to accomplish this task. The reducer uses a `defaultdict` object from `collections` to automatically collate tokens and their counts without having to explicitly instantiate the key. This is convenient when reading from `sys.stdin` as one can yield lines and store the tokens simultaneously.

In [74]:

```
%%writefile reducerQ33.py
#!/usr/bin/env python

from __future__ import division
import sys
from collections import defaultdict
import ast

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

if __name__ == "__main__":

    # Final store
    storingDict = defaultdict(int)
    maxBasket = 0
    totalTerms = 0

    # Read data
    data = readInput(sys.stdin)
    for line in data:

        # Parse value
        token = line[0]
        termCount, basketSize = ast.literal_eval(line[1])
        totalTerms += termCount

        # Store results
        storingDict[token] += termCount
        maxBasket = max(maxBasket, basketSize)

    # Metrics
    numUniqueProducts = len(set(storingDict.keys()))
    largestBasket = maxBasket

    # Compute frequencies
    for k, v in storingDict.iteritems():
        storingDict[k] = v

    # Find most frequent terms
    mostFrequentTerms = [(k, v, round(v / totalTerms, 4)) for
k, v in storingDict.iteritems()]
    mostFrequentTerms = sorted(mostFrequentTerms,
                                key
                                = lambda x: x[1],
                                rev
```

```

erse = True)

    # Get results
    print '\n' + '===== Number of Unique Products ====='
    print 'Answer: ' + str(numUniqueProducts) + '\n'

    print '\n' + '===== Largest Basket =====' + '\n'
    print 'Answer: ' + str(largestBasket) + '\n'

    print '===== Most Frequent Terms =====' + '\n'
    template = "{0:20}|{1:20}|{2:20}"
    print template.format("ITEM", "FREQUENCY", "RELATIVE FREQUE
NCY")

    # Print terms
    for termPair in mostFrequentTerms[:50]:
        print template.format(*termPair)

    # Update counter
    sys.stderr.write("reporter:counter:User-Defined,Number of R
educers,1\n")

```

Overwriting reducerQ33.py

## HW3.3 Wrapper

Below is the bash script that's used to submit the Hadoop Streaming job. It maps user input to variables and specifies the options for the job.

```
In [29]: %%writefile wrapperQ33.sh
#!/bin/bash

# Initialize
RAW_DATA=$1
RAW_MAPPER=$2
RAW_REDUCER=$3

# Hadoop variables
HDFS_DIR="/user/john/notebook"
HDFS_INPUT="$HDFS_DIR/input"
HDFS_OUTPUT="$HDFS_DIR/output"
HDFS_FILES="$HDFS_DIR/files"

# Local variables
PROJECT_DIR="/Users/john/Dropbox/MIDS/W261/Week3/Homework3/notebook"
DATA="$PROJECT_DIR/$RAW_DATA"
MAPPER="$PROJECT_DIR/$RAW_MAPPER"
REDUCER="$PROJECT_DIR/$RAW_REDUCER"


# NAIVE="$PROJECT_DIR/naiveBayes.py"
STREAMING_JAR="$PROJECT_DIR/hadoop-streaming-2.6.0.jar"

# Make directories and put file
hdfs dfs -rm -r $HDFS_DIR
hdfs dfs -mkdir $HDFS_DIR $HDFS_INPUT $HDFS_FILES
hdfs dfs -put $DATA $HDFS_INPUT

# Execute
hadoop jar $STREAMING_JAR \
    -file "$MAPPER" -mapper "$MAPPER" \
    -file "$REDUCER" -reducer "$REDUCER" \
    -input $HDFS_INPUT \
    -output $HDFS_OUTPUT

# Output results
if [ $? -eq 0 ]; then
    hdfs dfs -cat $HDFS_OUTPUT/part-00000
fi
```





```
Overwriting wrapperQ33.sh
```

### HW3.3 Hadoop MapReduce Submit Job

Below we submit the job. The wrapper takes as arguments the data, mapper, and reducer. Later versions of the wrapper also allow combiner arguments.

```
In [40]: !bash wrapperQ33.sh ProductPurchaseData.txt mapperQ33.py reducerQ33.py
```



```
16/01/30 16:05:41 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/01/30 16:05:42 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/john/notebook
16/01/30 16:05:44 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/01/30 16:05:46 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/01/30 16:05:49 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
16/01/30 16:05:49 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
packageJobJar: [/Users/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/mapperQ33.py, /Users/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/reducerQ33.py] [] /var/folders/0w/8hzv7rsj3qgdynsjlqy3gjsc0000gn/T/streamjob3210653833237786513.jar tmpDir=null
16/01/30 16:05:50 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
16/01/30 16:05:50 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/01/30 16:05:50 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
16/01/30 16:05:51 INFO mapred.FileInputFormat: Total input paths to process : 1
16/01/30 16:05:51 INFO mapreduce.JobSubmitter: number of splits:1
16/01/30 16:05:51 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1059638235_0001
16/01/30 16:05:52 INFO mapred.LocalDistributedCacheManager: Localized file:/Users/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/mapperQ33.py as file:/usr/local/Cellar/hadoop/hdfs/tmp/mapred/local/1454187951883/mapperQ33.py
16/01/30 16:05:52 INFO mapred.LocalDistributedCacheManager: Localized file:/Users/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/reducerQ33.py as file:/usr/local/Cellar/hadoop/hdfs/tmp/mapred/local/1454187951884/reducerQ33.py
16/01/30 16:05:52 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/01/30 16:05:52 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
16/01/30 16:05:52 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/30 16:05:52 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/01/30 16:05:52 INFO mapreduce.Job: Running job: job_local1059638235_0001
16/01/30 16:05:52 INFO mapred.LocalJobRunner: Waiting for map tasks
16/01/30 16:05:52 INFO mapred.LocalJobRunner: Starting task: attempt
```

```
pt_local1059638235_0001_m_000000_0
16/01/30 16:05:52 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
16/01/30 16:05:52 INFO util.ProcfsBasedProcessTree: ProcfsBasedPro
cessTree currently is supported only on Linux.
16/01/30 16:05:52 INFO mapred.Task: Using ResourceCalculatorProce
ssTree : null
16/01/30 16:05:52 INFO mapred.MapTask: Processing split: hdfs://lo
calhost:9000/user/john/notebook/input/ProductPurchaseData.txt:0+34
58517
16/01/30 16:05:52 INFO mapred.MapTask: numReduceTasks: 1
16/01/30 16:05:52 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(10
4857584)
16/01/30 16:05:52 INFO mapred.MapTask: mapreduce.task.io.sort.mb:
100
16/01/30 16:05:52 INFO mapred.MapTask: soft limit at 83886080
16/01/30 16:05:52 INFO mapred.MapTask: bufstart = 0; bufvoid = 104
857600
16/01/30 16:05:52 INFO mapred.MapTask: kvstart = 26214396; length
= 6553600
16/01/30 16:05:52 INFO mapred.MapTask: Map output collector class
= org.apache.hadoop.mapred.MapTask$MapOutputBuffer
16/01/30 16:05:53 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/./mapperQ33.py]
16/01/30 16:05:53 INFO Configuration.deprecation: mapred.work.outp
ut.dir is deprecated. Instead, use mapreduce.task.output.dir
16/01/30 16:05:53 INFO Configuration.deprecation: map.input.start
is deprecated. Instead, use mapreduce.map.input.start
16/01/30 16:05:53 INFO Configuration.deprecation: mapred.task.is.m
ap is deprecated. Instead, use mapreduce.task.ismap
16/01/30 16:05:53 INFO Configuration.deprecation: mapred.task.id i
s deprecated. Instead, use mapreduce.task.attempt.id
16/01/30 16:05:53 INFO Configuration.deprecation: mapred.tip.id is
deprecated. Instead, use mapreduce.task.id
16/01/30 16:05:53 INFO Configuration.deprecation: mapred.local.dir
is deprecated. Instead, use mapreduce.cluster.local.dir
16/01/30 16:05:53 INFO Configuration.deprecation: map.input.file i
s deprecated. Instead, use mapreduce.map.input.file
16/01/30 16:05:53 INFO Configuration.deprecation: mapred.skip.on i
s deprecated. Instead, use mapreduce.job.skiprecords
16/01/30 16:05:53 INFO Configuration.deprecation: map.input.length
is deprecated. Instead, use mapreduce.map.input.length
16/01/30 16:05:53 INFO Configuration.deprecation: mapred.job.id is
deprecated. Instead, use mapreduce.job.id
16/01/30 16:05:53 INFO Configuration.deprecation: user.name is dep
recated. Instead, use mapreduce.job.user.name
16/01/30 16:05:53 INFO Configuration.deprecation: mapred.task.part
ition is deprecated. Instead, use mapreduce.task.partition
16/01/30 16:05:53 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/01/30 16:05:53 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/30 16:05:53 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
```

```
16/01/30 16:05:53 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/30 16:05:53 INFO streaming.PipeMapRed: Records R/W=1216/1
16/01/30 16:05:53 INFO mapreduce.Job: Job job_local1059638235_0001
running in uber mode : false
16/01/30 16:05:53 INFO mapreduce.Job: map 0% reduce 0%
16/01/30 16:05:54 INFO streaming.PipeMapRed: R/W/S=10000/119481/0
in:NA [rec/s] out:NA [rec/s]
16/01/30 16:05:55 INFO streaming.PipeMapRed: MRErrorThread done
16/01/30 16:05:55 INFO streaming.PipeMapRed: mapRedFinished
16/01/30 16:05:55 INFO mapred.LocalJobRunner:
16/01/30 16:05:55 INFO mapred.MapTask: Starting flush of map output
16/01/30 16:05:55 INFO mapred.MapTask: Spilling map output
16/01/30 16:05:55 INFO mapred.MapTask: bufstart = 0; bufend = 6397
909; bufvoid = 104857600
16/01/30 16:05:55 INFO mapred.MapTask: kvstart = 26214396(10485758
4); kvend = 24691116(98764464); length = 1523281/6553600
16/01/30 16:05:56 INFO mapred.MapTask: Finished spill 0
16/01/30 16:05:56 INFO mapred.Task: Task:attempt_local1059638235_0
001_m_000000_0 is done. And is in the process of committing
16/01/30 16:05:56 INFO mapred.LocalJobRunner: Records R/W=1216/1
16/01/30 16:05:56 INFO mapred.Task: Task 'attempt_local1059638235_
0001_m_000000_0' done.
16/01/30 16:05:56 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local1059638235_0001_m_000000_0
16/01/30 16:05:56 INFO mapred.LocalJobRunner: map task executor co
mplete.
16/01/30 16:05:56 INFO mapred.LocalJobRunner: Waiting for reduce t
asks
16/01/30 16:05:56 INFO mapred.LocalJobRunner: Starting task: attem
pt_local1059638235_0001_r_000000_0
16/01/30 16:05:56 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
16/01/30 16:05:56 INFO util.ProcfsBasedProcessTree: ProcfsBasedPro
cessTree currently is supported only on Linux.
16/01/30 16:05:56 INFO mapred.Task: Using ResourceCalculatorProce
ssTree : null
16/01/30 16:05:56 INFO mapred.ReduceTask: Using ShuffleConsumerPlu
gin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@70e26d9a
16/01/30 16:05:56 INFO reduce.MergeManagerImpl: MergerManager: mem
oryLimit=334338464, maxSingleShuffleLimit=83584616, mergeThreshold
=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
16/01/30 16:05:56 INFO reduce.EventFetcher: attempt_local105963823
5_0001_r_000000_0 Thread started: EventFetcher for fetching Map Co
mpletion Events
16/01/30 16:05:56 INFO reduce.LocalFetcher: localfetcher#1 about t
o shuffle output of map attempt_local1059638235_0001_m_000000_0 de
comp: 7159553 len: 7159557 to MEMORY
16/01/30 16:05:56 INFO reduce.InMemoryMapOutput: Read 7159553 byte
s from map-output for attempt_local1059638235_0001_m_000000_0
16/01/30 16:05:56 INFO reduce.MergeManagerImpl: closeInMemoryFile
-> map-output of size: 7159553, inMemoryMapOutputs.size() -> 1, co
mmitMemory -> 0, usedMemory -> 7159553
```

```
16/01/30 16:05:56 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/01/30 16:05:56 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/30 16:05:56 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs
16/01/30 16:05:56 INFO mapred.Merger: Merging 1 sorted segments
16/01/30 16:05:56 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 7159542 bytes
16/01/30 16:05:57 INFO mapreduce.Job: map 100% reduce 0%
16/01/30 16:05:57 INFO reduce.MergeManagerImpl: Merged 1 segments, 7159553 bytes to disk to satisfy reduce memory limit
16/01/30 16:05:57 INFO reduce.MergeManagerImpl: Merging 1 files, 7159557 bytes from disk
16/01/30 16:05:57 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/01/30 16:05:57 INFO mapred.Merger: Merging 1 sorted segments
16/01/30 16:05:57 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 7159542 bytes
16/01/30 16:05:57 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/30 16:05:57 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/./reducerQ33.py]
16/01/30 16:05:57 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
16/01/30 16:05:57 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
16/01/30 16:05:57 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 16:05:57 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 16:05:57 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 16:05:57 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 16:05:57 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 16:06:00 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:3333=100000/3 [rec/s] out:0=0/3 [rec/s]
16/01/30 16:06:02 INFO mapred.LocalJobRunner: reduce > reduce
16/01/30 16:06:03 INFO mapreduce.Job: map 100% reduce 82%
16/01/30 16:06:03 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:40000=200000/5 [rec/s] out:0=0/5 [rec/s]
16/01/30 16:06:05 INFO mapred.LocalJobRunner: reduce > reduce
16/01/30 16:06:06 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:37500=300000/8 [rec/s] out:0=0/8 [rec/s]
16/01/30 16:06:06 INFO mapreduce.Job: map 100% reduce 91%
16/01/30 16:06:08 INFO streaming.PipeMapRed: Records R/W=380821/1
16/01/30 16:06:08 INFO streaming.PipeMapRed: MRErrorThread done
16/01/30 16:06:08 INFO streaming.PipeMapRed: mapRedFinished
16/01/30 16:06:08 INFO mapred.LocalJobRunner: Records R/W=380821/1 > reduce
16/01/30 16:06:08 INFO mapred.Task: Task:attempt_local1059638235_001_r_000000_0 is done. And is in the process of committing
16/01/30 16:06:08 INFO mapred.LocalJobRunner: Records R/W=380821/1
```

```
> reduce
16/01/30 16:06:08 INFO mapred.Task: Task attempt_local1059638235_0
001_r_000000_0 is allowed to commit now
16/01/30 16:06:08 INFO output.FileOutputCommitter: Saved output of
task 'attempt_local1059638235_0001_r_000000_0' to hdfs://localhos
t:9000/user/john/notebook/output/_temporary/0/task_local1059638235
_0001_r_000000
16/01/30 16:06:08 INFO mapred.LocalJobRunner: Records R/W=380821/1
> reduce
16/01/30 16:06:08 INFO mapred.Task: Task 'attempt_local1059638235_
0001_r_000000_0' done.
16/01/30 16:06:08 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local1059638235_0001_r_000000_0
16/01/30 16:06:08 INFO mapred.LocalJobRunner: reduce task executor
complete.
16/01/30 16:06:09 INFO mapreduce.Job: map 100% reduce 100%
16/01/30 16:06:09 INFO mapreduce.Job: Job job_local1059638235_0001
completed successfully
16/01/30 16:06:09 INFO mapreduce.Job: Counters: 35
    File System Counters
        FILE: Number of bytes read=14326178
        FILE: Number of bytes written=22073423
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=6917034
        HDFS: Number of bytes written=3434
        HDFS: Number of read operations=13
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=4
    Map-Reduce Framework
        Map input records=31101
        Map output records=380821
        Map output bytes=6397909
        Map output materialized bytes=7159557
        Input split bytes=122
        Combine input records=0
        Combine output records=0
        Reduce input groups=12592
        Reduce shuffle bytes=7159557
        Reduce input records=380821
        Reduce output records=63
        Spilled Records=761642
        Shuffled Maps =1
        Failed Shuffles=0
        Merged Map outputs=1
        GC time elapsed (ms)=0
        Total committed heap usage (bytes)=500170752
    Shuffle Errors
        BAD_ID=0
        CONNECTION=0
        IO_ERROR=0
        WRONG_LENGTH=0
        WRONG_MAP=0
```



WRONG\_REDUCE=0

File Input Format Counters

Bytes Read=3458517

File Output Format Counters

Bytes Written=3434

16/01/30 16:06:09 INFO streaming.StreamJob: Output directory: /usr/john/notebook/output

16/01/30 16:06:11 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

===== Number of Unique Products =====

Answer: 12592

===== Largest Basket =====

Answer: 37

===== Most Frequent Terms =====

ITEM	FREQUENCY	RELATIVE FREQUENCY
DAI62779	6667	0.0175
FRO40251	3881	0.0102
ELE17451	3875	0.0102
GRO73461	3602	0.0095
SNA80324	3044	0.008
ELE32164	2851	0.0075
DAI75645	2736	0.0072
SNA45677	2455	0.0064
FRO31317	2330	0.0061
DAI85309	2293	0.006
ELE26917	2292	0.006
FRO80039	2233	0.0059
GRO21487	2115	0.0056
SNA99873	2083	0.0055
GRO59710	2004	0.0053
GRO71621	1920	0.005
FRO85978	1918	0.005
GRO30386	1840	0.0048
ELE74009	1816	0.0048
GRO56726	1784	0.0047
DAI63921	1773	0.0047
GRO46854	1756	0.0046
ELE66600	1713	0.0045
DAI83733	1712	0.0045
FRO32293	1702	0.0045
ELE66810	1697	0.0045
SNA55762	1646	0.0043
DAI22177	1627	0.0043
FRO78087	1531	0.004
ELE99737	1516	0.004
ELE34057	1489	0.0039

GRO94758	1489	0.0039
FRO35904	1436	0.0038
FRO53271	1420	0.0037
SNA93860	1407	0.0037
SNA90094	1390	0.0036
GRO38814	1352	0.0036
ELE56788	1345	0.0035
GRO61133	1321	0.0035
ELE74482	1316	0.0035
DAI88807	1316	0.0035
ELE59935	1311	0.0034
SNA96271	1295	0.0034
DAI43223	1290	0.0034
ELE91337	1289	0.0034
GRO15017	1275	0.0033
DAI31081	1261	0.0033
GRO81087	1220	0.0032
DAI22896	1219	0.0032
GRO85051	1214	0.0032

## HW3.4

Write a map-reduce program to find products which are frequently browsed together. Fix the support count (cooccurrence count) to  $s = 100$  (i.e. product pairs need to occur together at least 100 times to be considered frequent) and find pairs of items (sometimes referred to itemsets of size 2 in association rule mining) that have a support count of 100 or more.

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

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

## Solution Approach A

### Mapper

Below is the mapper code for the pairs implementation. It generates tuple-combinations from the input lines, sorts them to get unique keys, then emits them to the reducer or optional combiner.

```
In [10]: %%writefile mapperQ34.py
#!/usr/bin/env python

from __future__ import division
import sys
from collections import defaultdict
from itertools import combinations

def readInput(file, separator=None):
    for line in file:
        yield line.split(separator)

if __name__ == "__main__":

    # Read input
    data = readInput(sys.stdin)
    totalBaskets = 0

    for line in data:

        # Increment
        totalBaskets += 1

        # Get unique keys
        pairs = list(combinations(line, 2))

        # Sort keys
        sortedPairs = []
        for pair in pairs:
            pList = list(pair)
            pList.sort()
            sortedPairs.append(tuple(pList))

        # Emit
        for pair in sortedPairs:
            print '%s%s%s' % (pair, '\t', 1)

    # Update counter
    sys.stderr.write("reporter:counter:User-Defined,Number of M
appers,1\n")

    # Emit basket count
    print '%s%s%s' % ('*', '\t', str(totalBaskets))
```

Overwriting mapperQ34.py

## HW3.4 Reducer

Below is the reducer for this process. It uses a simple key-aggregation to collate the results from the mappers, where the keys are unique tuples from the lines read in by the mapper. The uniqueness condition is specified at the line-level when read by the mapper.

In [11]:

```
%%writefile reducerQ34.py
#!/usr/bin/env python

from __future__ import division
import sys
from collections import defaultdict
import ast

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

if __name__ == "__main__":

    # Final store
    storingDict = defaultdict(int)
    support = 100
    totalBaskets = 0

    # Read data
    data = readInput(sys.stdin)
    for line in data:

        # Check for basket
        token = line[0]

        if token == '*':
            totalBaskets += int(line[1])

        else:

            # Parse
            termCount = int(line[1])

            # Store results
            storingDict[token] += termCount

    # Filter
    filterDict = defaultdict(int)
    for k, v in storingDict.iteritems():
        if v >= support:
            filterDict[k] += v

    # Find most frequent terms
    mostFrequentTerms = [(k, v, round(int(v) / totalBaskets,
3)) for k, v in filterDict.iteritems()]
    mostFrequentTerms = sorted(mostFrequentTerms,
                                key
                                = lambda x: x[1],
```

```
erse = True)

    # Get results
    print '\n' + '===== Most Frequent Terms =====' +
'\n'
    template = "{0:30}|{1:20}|{2:20}"
    print template.format("PAIR", "SUPPORT COUNT", "RELATIVE FR
EQUENCY")

    # Print terms
    for termPair in mostFrequentTerms[:50]:
        print template.format(*termPair)

    # Update counter
    sys.stderr.write("reporter:counter:User-Defined,Number of R
educers,1\n")
```

Overwriting reducerQ34.py

### HW3.4 Combiner (Optional)

Below is the code for an optional combiner. It has the same signature as the reducer and does the same essential aggregation.

```
In [12]: %%writefile combinerQ34.py
#!/usr/bin/env python

from __future__ import division
import sys
from collections import defaultdict
import ast

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

if __name__ == "__main__":

    # Final store
    storingDict = defaultdict(int)

    # Read data
    data = readInput(sys.stdin)
    totalBaskets = 0

    for line in data:

        # Check for basket
        token = line[0]

        if token == '*':
            totalBaskets += int(line[1])

        else:

            # Parse
            termCount = int(line[1])


            # Store results
            storingDict[token] += termCount

    # Emit
    for k, v in storingDict.iteritems():
        print '%s%s%s' % (k, '\t', v)

    # Update counter
    sys.stderr.write("reporter:counter:User-Defined,Number of C
ombiners,1\n")

    # Emit basket count
    print '%s%s%s' % ('*', '\t', str(totalBaskets))
```





```
Overwriting combinerQ34.py
```

## HW3.4 Wrapper

Here is the wrapper for this particular submission. The noticeable difference is that it now has been modified to pass a combiner file to Hadoop Streaming.

```
In [52]: %%writefile wrapperQ34.sh
#!/bin/bash

# Initialize
RAW_DATA=$1
RAW_MAPPER=$2
RAW_REDUCER=$3
RAW_COMBINER=$4

# Hadoop variables
HDFS_DIR="/user/john/notebook"
HDFS_INPUT="$HDFS_DIR/input"
HDFS_OUTPUT="$HDFS_DIR/output"
HDFS_FILES="$HDFS_DIR/files"


# Local variables
PROJECT_DIR="/Users/john/Dropbox/MIDS/W261/Week3/Homework3/notebook"
DATA="$PROJECT_DIR/$RAW_DATA"
MAPPER="$PROJECT_DIR/$RAW_MAPPER"
REDUCER="$PROJECT_DIR/$RAW_REDUCER"
COMBINER="$PROJECT_DIR/$RAW_COMBINER"

# NAIVE="$PROJECT_DIR/naiveBayes.py"
STREAMING_JAR="$PROJECT_DIR/hadoop-streaming-2.6.0.jar"

# Make directories and put file
hdfs dfs -rm -r $HDFS_DIR
hdfs dfs -mkdir $HDFS_DIR $HDFS_INPUT $HDFS_FILES
hdfs dfs -put $DATA $HDFS_INPUT

# Execute
hadoop jar $STREAMING_JAR \
    -file "$MAPPER" -mapper "$MAPPER" \
    -file "$REDUCER" -reducer "$REDUCER" \
    -file "$COMBINER" -combiner "$COMBINER" \
    -input $HDFS_INPUT \
    -output $HDFS_OUTPUT \
    -cmdenv mapred.map.max.attempts=1 \
    -cmdenv mapred.reduce.max.attempts=1 \

# Output results
if [ $? -eq 0 ]; then
    hdfs dfs -cat $HDFS_OUTPUT/part-00000
fi
```



```
Overwriting wrapperQ34.sh
```

***Submit Job***

We now submit the job for this process. Notice that an argument has been added for the combiner.

```
In [13]: !bash wrapperQ34.sh ProductPurchaseData.txt mapperQ34.py reducerQ34.py combinerQ34.py
```



```
16/02/01 19:56:52 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/02/01 19:56:53 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/john/notebook
16/02/01 19:56:55 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/02/01 19:56:57 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/02/01 19:57:01 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
16/02/01 19:57:02 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
packageJobJar: [/Users/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/mapperQ34.py, /Users/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/reducerQ34.py, /Users/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/combinerQ34.py] [] /var/folders/0w/8hzv7rsj3qgdynsjlqy3gjsc0000gn/T/streamjob6486516089282375945.jar tmpDir=null
16/02/01 19:57:03 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
16/02/01 19:57:03 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/02/01 19:57:04 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
16/02/01 19:57:05 INFO mapred.FileInputFormat: Total input paths to process : 1
16/02/01 19:57:05 INFO mapreduce.JobSubmitter: number of splits:1
16/02/01 19:57:05 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1887715370_0001
16/02/01 19:57:06 INFO mapred.LocalDistributedCacheManager: Localized file:/Users/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/mapperQ34.py as file:/usr/local/Cellar/hadoop/hdfs/tmp/mapred/local/1454374626230/mapperQ34.py
16/02/01 19:57:06 INFO mapred.LocalDistributedCacheManager: Localized file:/Users/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/reducerQ34.py as file:/usr/local/Cellar/hadoop/hdfs/tmp/mapred/local/1454374626231/reducerQ34.py
16/02/01 19:57:06 INFO mapred.LocalDistributedCacheManager: Localized file:/Users/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/combinerQ34.py as file:/usr/local/Cellar/hadoop/hdfs/tmp/mapred/local/1454374626232/combinerQ34.py
16/02/01 19:57:07 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/02/01 19:57:07 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/02/01 19:57:07 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
16/02/01 19:57:07 INFO mapreduce.Job: Running job: job_local1887715370_0001
```

```
16/02/01 19:57:07 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
16/02/01 19:57:07 INFO mapred.LocalJobRunner: Waiting for map task
s
16/02/01 19:57:07 INFO mapred.LocalJobRunner: Starting task: attem
pt_local1887715370_0001_m_000000_0
16/02/01 19:57:07 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
16/02/01 19:57:07 INFO util.ProcfsBasedProcessTree: ProcfsBasedPro
cessTree currently is supported only on Linux.
16/02/01 19:57:07 INFO mapred.Task: Using ResourceCalculatorProce
ssTree : null
16/02/01 19:57:07 INFO mapred.MapTask: Processing split: hdfs://lo
calhost:9000/user/john/notebook/input/ProductPurchaseData.txt:0+34
58517
16/02/01 19:57:07 INFO mapred.MapTask: numReduceTasks: 1
16/02/01 19:57:07 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(10
4857584)
16/02/01 19:57:07 INFO mapred.MapTask: mapreduce.task.io.sort.mb:
100
16/02/01 19:57:07 INFO mapred.MapTask: soft limit at 83886080
16/02/01 19:57:07 INFO mapred.MapTask: bufstart = 0; bufvoid = 104
857600
16/02/01 19:57:07 INFO mapred.MapTask: kvstart = 26214396; length
= 6553600
16/02/01 19:57:07 INFO mapred.MapTask: Map output collector class
= org.apache.hadoop.mapred.MapTask$MapOutputBuffer
16/02/01 19:57:07 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/./mapperQ34.py]
16/02/01 19:57:07 INFO Configuration.deprecation: mapred.work.outp
ut.dir is deprecated. Instead, use mapreduce.task.output.dir
16/02/01 19:57:07 INFO Configuration.deprecation: map.input.start
is deprecated. Instead, use mapreduce.map.input.start
16/02/01 19:57:07 INFO Configuration.deprecation: mapred.task.is.m
ap is deprecated. Instead, use mapreduce.task.ismap
16/02/01 19:57:07 INFO Configuration.deprecation: mapred.task.id i
s deprecated. Instead, use mapreduce.task.attempt.id
16/02/01 19:57:07 INFO Configuration.deprecation: mapred.tip.id is
deprecated. Instead, use mapreduce.task.id
16/02/01 19:57:07 INFO Configuration.deprecation: mapred.local.dir
is deprecated. Instead, use mapreduce.cluster.local.dir
16/02/01 19:57:07 INFO Configuration.deprecation: map.input.file i
s deprecated. Instead, use mapreduce.map.input.file
16/02/01 19:57:07 INFO Configuration.deprecation: mapred.skip.on i
s deprecated. Instead, use mapreduce.job.skiprecords
16/02/01 19:57:07 INFO Configuration.deprecation: map.input.length
is deprecated. Instead, use mapreduce.map.input.length
16/02/01 19:57:07 INFO Configuration.deprecation: mapred.job.id is
deprecated. Instead, use mapreduce.job.id
16/02/01 19:57:07 INFO Configuration.deprecation: user.name is dep
recated. Instead, use mapreduce.job.user.name
16/02/01 19:57:07 INFO Configuration.deprecation: mapred.task.part
ition is deprecated. Instead, use mapreduce.task.partition
16/02/01 19:57:07 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
```

```
c/s] out:NA [rec/s]
16/02/01 19:57:07 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/02/01 19:57:07 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/02/01 19:57:07 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/02/01 19:57:07 INFO streaming.PipeMapRed: Records R/W=1216/1
16/02/01 19:57:08 INFO mapreduce.Job: Job job_local1887715370_0001
running in uber mode : false
16/02/01 19:57:08 INFO mapreduce.Job: map 0% reduce 0%
16/02/01 19:57:12 INFO streaming.PipeMapRed: R/W/S=10000/854395/0
in:2000=10000/5 [rec/s] out:170879=854395/5 [rec/s]
16/02/01 19:57:13 INFO mapred.LocalJobRunner: Records R/W=1216/1 >
map
16/02/01 19:57:14 INFO mapreduce.Job: map 25% reduce 0%
16/02/01 19:57:16 INFO mapred.LocalJobRunner: Records R/W=1216/1 >
map
16/02/01 19:57:17 INFO mapreduce.Job: map 33% reduce 0%
16/02/01 19:57:17 INFO streaming.PipeMapRed: Records R/W=17860/135
6232
16/02/01 19:57:19 INFO mapred.LocalJobRunner: Records R/W=17860/13
56232 > map
16/02/01 19:57:20 INFO mapreduce.Job: map 43% reduce 0%
16/02/01 19:57:22 INFO mapred.LocalJobRunner: Records R/W=17860/13
56232 > map
16/02/01 19:57:22 INFO mapred.MapTask: Spilling map output
16/02/01 19:57:22 INFO mapred.MapTask: bufstart = 0; bufend = 5267
2653; bufvoid = 104857600
16/02/01 19:57:22 INFO mapred.MapTask: kvstart = 26214396(10485758
4); kvend = 18411044(73644176); length = 7803353/6553600
16/02/01 19:57:22 INFO mapred.MapTask: (EQUATOR) 60475997 kvi 1511
8992(60475968)
16/02/01 19:57:23 INFO mapreduce.Job: map 53% reduce 0%
16/02/01 19:57:25 INFO mapred.LocalJobRunner: Records R/W=17860/13
56232 > map
16/02/01 19:57:26 INFO mapreduce.Job: map 61% reduce 0%
16/02/01 19:57:28 INFO mapred.LocalJobRunner: Records R/W=17860/13
56232 > map
16/02/01 19:57:29 INFO mapreduce.Job: map 67% reduce 0%
16/02/01 19:57:30 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/./combinerQ34.p
y]
16/02/01 19:57:30 INFO Configuration.deprecation: mapred.skip.map.
auto.incr.proc.count is deprecated. Instead, use mapreduce.map.ski
p.proc-count.auto-incr
16/02/01 19:57:30 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/02/01 19:57:30 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/02/01 19:57:30 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/02/01 19:57:30 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA
[rec/s] out:NA [rec/s]
```



```
16/02/01 19:57:31 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/02/01 19:57:31 INFO mapred.LocalJobRunner: Records R/W=17860/13
56232 > map
16/02/01 19:57:31 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:NA
A [rec/s] out:NA [rec/s]
16/02/01 19:57:32 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:2
00000=200000/1 [rec/s] out:0=0/1 [rec/s]
16/02/01 19:57:33 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:1
50000=300000/2 [rec/s] out:0=0/2 [rec/s]
16/02/01 19:57:33 INFO streaming.PipeMapRed: R/W/S=400000/0/0 in:2
00000=400000/2 [rec/s] out:0=0/2 [rec/s]
16/02/01 19:57:34 INFO mapred.LocalJobRunner: Records R/W=17860/13
56232 > map
16/02/01 19:57:34 INFO streaming.PipeMapRed: R/W/S=500000/0/0 in:1
66666=500000/3 [rec/s] out:0=0/3 [rec/s]
16/02/01 19:57:34 INFO streaming.PipeMapRed: R/W/S=600000/0/0 in:2
00000=600000/3 [rec/s] out:0=0/3 [rec/s]
16/02/01 19:57:35 INFO streaming.PipeMapRed: R/W/S=700000/0/0 in:1
75000=700000/4 [rec/s] out:0=0/4 [rec/s]
16/02/01 19:57:35 INFO streaming.PipeMapRed: R/W/S=800000/0/0 in:1
60000=800000/5 [rec/s] out:0=0/5 [rec/s]
16/02/01 19:57:36 INFO streaming.PipeMapRed: R/W/S=900000/0/0 in:1
80000=900000/5 [rec/s] out:0=0/5 [rec/s]
16/02/01 19:57:37 INFO streaming.PipeMapRed: R/W/S=1000000/0/0 in:
166666=1000000/6 [rec/s] out:0=0/6 [rec/s]
16/02/01 19:57:37 INFO mapred.LocalJobRunner: Records R/W=17860/13
56232 > map
16/02/01 19:57:37 INFO streaming.PipeMapRed: R/W/S=1100000/0/0 in:
157142=1100000/7 [rec/s] out:0=0/7 [rec/s]
16/02/01 19:57:38 INFO streaming.PipeMapRed: R/W/S=1200000/0/0 in:
171428=1200000/7 [rec/s] out:0=0/7 [rec/s]
16/02/01 19:57:38 INFO streaming.PipeMapRed: R/W/S=1300000/0/0 in:
185714=1300000/7 [rec/s] out:0=0/7 [rec/s]
16/02/01 19:57:39 INFO streaming.PipeMapRed: R/W/S=1400000/0/0 in:
175000=1400000/8 [rec/s] out:0=0/8 [rec/s]
16/02/01 19:57:40 INFO streaming.PipeMapRed: R/W/S=1500000/0/0 in:
166666=1500000/9 [rec/s] out:0=0/9 [rec/s]
16/02/01 19:57:40 INFO mapred.LocalJobRunner: Records R/W=17860/13
56232 > map
16/02/01 19:57:40 INFO streaming.PipeMapRed: R/W/S=1600000/0/0 in:
160000=1600000/10 [rec/s] out:0=0/10 [rec/s]
16/02/01 19:57:41 INFO streaming.PipeMapRed: R/W/S=1700000/0/0 in:
170000=1700000/10 [rec/s] out:0=0/10 [rec/s]
16/02/01 19:57:42 INFO streaming.PipeMapRed: R/W/S=1800000/0/0 in:
163636=1800000/11 [rec/s] out:0=0/11 [rec/s]
16/02/01 19:57:42 INFO streaming.PipeMapRed: R/W/S=1900000/0/0 in:
172727=1900000/11 [rec/s] out:0=0/11 [rec/s]
16/02/01 19:57:42 INFO streaming.PipeMapRed: Records R/W=1950839/1
16/02/01 19:57:43 INFO mapred.LocalJobRunner: Records R/W=1950839/
1 > map
16/02/01 19:57:45 INFO streaming.PipeMapRed: MRErrorThread done
16/02/01 19:57:45 INFO streaming.PipeMapRed: mapRedFinished
16/02/01 19:57:45 INFO mapred.MapTask: Finished spill 0
```

```
16/02/01 19:57:45 INFO mapred.MapTask: (RESET) equator 60475997 kv
15118992(60475968) kvi 13168168(52672672)
16/02/01 19:57:45 INFO streaming.PipeMapRed: Records R/W=31101/243
8546
16/02/01 19:57:45 INFO streaming.PipeMapRed: MRErrorThread done
16/02/01 19:57:45 INFO streaming.PipeMapRed: mapRedFinished
16/02/01 19:57:45 INFO mapred.LocalJobRunner: Records R/W=1950839/
1 > map
16/02/01 19:57:45 INFO mapred.MapTask: Starting flush of map output
16/02/01 19:57:45 INFO mapred.MapTask: Spilling map output
16/02/01 19:57:45 INFO mapred.MapTask: bufstart = 60475997; bufend
= 76222891; bufvoid = 104857600
16/02/01 19:57:45 INFO mapred.MapTask: kvstart = 15118992(6047596
8); kvend = 12786120(51144480); length = 2332873/6553600
16/02/01 19:57:46 INFO mapred.LocalJobRunner: Records R/W=31101/24
38546 > sort
16/02/01 19:57:47 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/./combinerQ34.p
y]
16/02/01 19:57:47 INFO Configuration.deprecation: mapred.skip.redu
ce.auto.incr.proc.count is deprecated. Instead, use mapreduce.redu
ce.skip.proc-count.auto-incr
16/02/01 19:57:47 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/02/01 19:57:47 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/02/01 19:57:47 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/02/01 19:57:47 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/02/01 19:57:47 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/02/01 19:57:47 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/02/01 19:57:48 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:2
00000=200000/1 [rec/s] out:0=0/1 [rec/s]
16/02/01 19:57:48 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:3
00000=300000/1 [rec/s] out:0=0/1 [rec/s]
16/02/01 19:57:49 INFO streaming.PipeMapRed: R/W/S=400000/0/0 in:4
00000=400000/1 [rec/s] out:0=0/1 [rec/s]
16/02/01 19:57:49 INFO mapred.LocalJobRunner: Records R/W=31101/24
38546 > sort
16/02/01 19:57:49 INFO streaming.PipeMapRed: R/W/S=500000/0/0 in:2
50000=500000/2 [rec/s] out:0=0/2 [rec/s]
16/02/01 19:57:50 INFO streaming.PipeMapRed: Records R/W=583219/1
16/02/01 19:57:51 INFO streaming.PipeMapRed: MRErrorThread done
16/02/01 19:57:51 INFO streaming.PipeMapRed: mapRedFinished
16/02/01 19:57:51 INFO mapred.MapTask: Finished spill 1
16/02/01 19:57:51 INFO mapred.Merger: Merging 2 sorted segments
16/02/01 19:57:51 INFO mapred.Merger: Down to the last merge-pass,
with 2 segments left of total size: 28603089 bytes
16/02/01 19:57:52 INFO mapred.LocalJobRunner: Records R/W=583219/1
> sort >
```

```
16/02/01 19:57:53 INFO mapreduce.Job: map 71% reduce 0%
16/02/01 19:57:55 INFO mapred.LocalJobRunner: Records R/W=583219/1
> sort >
16/02/01 19:57:55 INFO mapred.Task: Task:attempt_local887715370_00
01_m_000000_0 is done. And is in the process of committing
16/02/01 19:57:55 INFO mapred.LocalJobRunner: Records R/W=583219/1
> sort
16/02/01 19:57:55 INFO mapred.Task: Task 'attempt_local887715370_0
001_m_000000_0' done.
16/02/01 19:57:55 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local887715370_0001_m_000000_0
16/02/01 19:57:55 INFO mapred.LocalJobRunner: map task executor co
mplete.
16/02/01 19:57:55 INFO mapred.LocalJobRunner: Waiting for reduce t
asks
16/02/01 19:57:55 INFO mapred.LocalJobRunner: Starting task: attem
pt_local887715370_0001_r_000000_0
16/02/01 19:57:55 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
16/02/01 19:57:55 INFO util.ProcfsBasedProcessTree: ProcfsBasedPro
cessTree currently is supported only on Linux.
16/02/01 19:57:55 INFO mapred.Task: Using ResourceCalculatorProce
ssTree : null
16/02/01 19:57:55 INFO mapred.ReduceTask: Using ShuffleConsumerPlu
gin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@31ba89b4
16/02/01 19:57:55 INFO reduce.MergeManagerImpl: MergerManager: mem
oryLimit=334338464, maxSingleShuffleLimit=83584616, mergeThreshold
=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
16/02/01 19:57:55 INFO reduce.EventFetcher: attempt_local887715370
_0001_r_000000_0 Thread started: EventFetcher for fetching Map Com
pletion Events
16/02/01 19:57:56 INFO reduce.LocalFetcher: localfetcher#1 about t
o shuffle output of map attempt_local887715370_0001_m_000000_0 dec
omp: 28603133 len: 28603137 to MEMORY
16/02/01 19:57:56 INFO mapreduce.Job: map 100% reduce 0%
16/02/01 19:57:56 INFO reduce.InMemoryMapOutput: Read 28603133 byt
es from map-output for attempt_local887715370_0001_m_000000_0
16/02/01 19:57:56 INFO reduce.MergeManagerImpl: closeInMemoryFile
-> map-output of size: 28603133, inMemoryMapOutputs.size() -> 1, c
ommitMemory -> 0, usedMemory ->28603133
16/02/01 19:57:56 INFO reduce.EventFetcher: EventFetcher is interr
upted.. Returning
16/02/01 19:57:56 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/01 19:57:56 INFO reduce.MergeManagerImpl: finalMerge called
with 1 in-memory map-outputs and 0 on-disk map-outputs
16/02/01 19:57:56 INFO mapred.Merger: Merging 1 sorted segments
16/02/01 19:57:56 INFO mapred.Merger: Down to the last merge-pass,
with 1 segments left of total size: 28603106 bytes
16/02/01 19:57:58 INFO reduce.MergeManagerImpl: Merged 1 segments,
28603133 bytes to disk to satisfy reduce memory limit
16/02/01 19:57:58 INFO reduce.MergeManagerImpl: Merging 1 files, 2
8603137 bytes from disk
16/02/01 19:57:58 INFO reduce.MergeManagerImpl: Merging 0 segment
s, 0 bytes from memory into reduce
```

```
16/02/01 19:57:58 INFO mapred.Merger: Merging 1 sorted segments
16/02/01 19:57:58 INFO mapred.Merger: Down to the last merge-pass,
with 1 segments left of total size: 28603106 bytes
16/02/01 19:57:58 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/01 19:57:58 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/./reducerQ34.p
y]
16/02/01 19:57:58 INFO Configuration.deprecation: mapred.job.track
er is deprecated. Instead, use mapreduce.jobtracker.address
16/02/01 19:57:58 INFO Configuration.deprecation: mapred.map.tasks
is deprecated. Instead, use mapreduce.job.maps
16/02/01 19:57:59 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/02/01 19:57:59 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/02/01 19:57:59 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/02/01 19:57:59 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/02/01 19:57:59 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/02/01 19:58:00 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:1
00000=100000/1 [rec/s] out:0=0/1 [rec/s]
16/02/01 19:58:01 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:1
00000=200000/2 [rec/s] out:0=0/2 [rec/s]
16/02/01 19:58:01 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:1
50000=300000/2 [rec/s] out:0=0/2 [rec/s]
16/02/01 19:58:01 INFO mapred.LocalJobRunner: reduce > reduce
16/02/01 19:58:02 INFO mapreduce.Job: map 100% reduce 78%
16/02/01 19:58:02 INFO streaming.PipeMapRed: R/W/S=400000/0/0 in:1
33333=400000/3 [rec/s] out:0=0/3 [rec/s]
16/02/01 19:58:03 INFO streaming.PipeMapRed: R/W/S=500000/0/0 in:1
25000=500000/4 [rec/s] out:0=0/4 [rec/s]
16/02/01 19:58:03 INFO streaming.PipeMapRed: R/W/S=600000/0/0 in:1
50000=600000/4 [rec/s] out:0=0/4 [rec/s]
16/02/01 19:58:04 INFO streaming.PipeMapRed: R/W/S=700000/0/0 in:1
40000=700000/5 [rec/s] out:0=0/5 [rec/s]
16/02/01 19:58:04 INFO mapred.LocalJobRunner: reduce > reduce
16/02/01 19:58:05 INFO streaming.PipeMapRed: R/W/S=800000/0/0 in:1
33333=800000/6 [rec/s] out:0=0/6 [rec/s]
16/02/01 19:58:05 INFO mapreduce.Job: map 100% reduce 93%
16/02/01 19:58:05 INFO streaming.PipeMapRed: R/W/S=900000/0/0 in:1
50000=900000/6 [rec/s] out:0=0/6 [rec/s]
16/02/01 19:58:07 INFO streaming.PipeMapRed: Records R/W=985039/1
16/02/01 19:58:07 INFO streaming.PipeMapRed: MRErrorThread done
16/02/01 19:58:07 INFO streaming.PipeMapRed: mapRedFinished
16/02/01 19:58:07 INFO mapred.Task: Task:attempt_local887715370_00
01_r_000000_0 is done. And is in the process of committing
16/02/01 19:58:07 INFO mapred.LocalJobRunner: reduce > reduce
16/02/01 19:58:07 INFO mapred.Task: Task attempt_local887715370_00
01_r_000000_0 is allowed to commit now
16/02/01 19:58:07 INFO output.FileOutputCommitter: Saved output of
task 'attempt_local887715370_0001_r_000000_0' to hdfs://localhost:
9000/user/john/notebook/output/_temporary/0/task_local887715370_00
```

```
01_r_000000
16/02/01 19:58:07 INFO mapred.LocalJobRunner: Records R/W=985039/1
> reduce
16/02/01 19:58:07 INFO mapred.Task: Task 'attempt_local887715370_0
001_r_000000_0' done.
16/02/01 19:58:07 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local887715370_0001_r_000000_0
16/02/01 19:58:07 INFO mapred.LocalJobRunner: reduce task executor
complete.
16/02/01 19:58:08 INFO mapreduce.Job: map 100% reduce 100%
16/02/01 19:58:08 INFO mapreduce.Job: Job job_local887715370_0001
completed successfully
16/02/01 19:58:08 INFO mapreduce.Job: Counters: 38
```

#### File System Counters

```
FILE: Number of bytes read=114422786
FILE: Number of bytes written=143613365
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=6917034
HDFS: Number of bytes written=3821
HDFS: Number of read operations=13
HDFS: Number of large read operations=0
HDFS: Number of write operations=4
```

#### Map-Reduce Framework

```
Map input records=31101
Map output records=2534058
Map output bytes=68419547
Map output materialized bytes=28603137
Input split bytes=122
Combine input records=2534058
Combine output records=985039
Reduce input groups=985038
Reduce shuffle bytes=28603137
Reduce input records=985039
Reduce output records=54
Spilled Records=2955117
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=20
Total committed heap usage (bytes)=526385152
```

#### Shuffle Errors

```
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
```

#### User-Defined

```
Number of Combiners=2
Number of Mappers=1
Number of Reducers=1
```

#### File Input Format Counters

Bytes Read=3458517

File Output Format Counters

Bytes Written=3821

16/02/01 19:58:08 INFO streaming.StreamJob: Output directory: /user/john/notebook/output

16/02/01 19:58:10 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

===== Most Frequent Terms =====

PAIR ENCY	SUPPORT COUNT	RELATIVE FREQUENCY
('DAI62779', 'ELE17451')	1592	0.051
('FRO40251', 'SNA80324')	1412	0.045
('DAI75645', 'FRO40251')	1254	0.04
('FRO40251', 'GRO85051')	1213	0.039
('DAI62779', 'GRO73461')	1139	0.037
('DAI75645', 'SNA80324')	1130	0.036
('DAI62779', 'FRO40251')	1070	0.034
('DAI62779', 'SNA80324')	923	0.03
('DAI62779', 'DAI85309')	918	0.03
('ELE32164', 'GRO59710')	911	0.029
('FRO40251', 'GRO73461')	882	0.028
('DAI62779', 'DAI75645')	882	0.028
('DAI62779', 'ELE92920')	877	0.028
('FRO40251', 'FRO92469')	835	0.027
('DAI62779', 'ELE32164')	832	0.027
('DAI75645', 'GRO73461')	712	0.023
('DAI43223', 'ELE32164')	711	0.023
('DAI62779', 'GRO30386')	709	0.023
('ELE17451', 'FRO40251')	697	0.022
('DAI85309', 'ELE99737')	659	0.021
('DAI62779', 'ELE26917')	650	

0.021		
('GRO21487', 'GRO73461')		631
0.02		
('DAI62779', 'SNA45677')		604
0.019		
('ELE17451', 'SNA80324')		597
0.019		
('DAI62779', 'GRO71621')		595
0.019		
('DAI62779', 'SNA55762')		593
0.019		
('DAI62779', 'DAI83733')		586
0.019		
('ELE17451', 'GRO73461')		580
0.019		
('GRO73461', 'SNA80324')		562
0.018		
('DAI62779', 'GRO59710')		561
0.018		
('DAI62779', 'FRO80039')		550
0.018		
('DAI75645', 'ELE17451')		547
0.018		
('DAI62779', 'SNA93860')		537
0.017		
('DAI55148', 'DAI62779')		526
0.017		
('DAI43223', 'GRO59710')		512
0.016		
('ELE17451', 'ELE32164')		511
0.016		
('DAI62779', 'SNA18336')		506
0.016		
('ELE32164', 'GRO73461')		486
0.016		
('DAI85309', 'ELE17451')		482
0.015		
('DAI62779', 'FRO78087')		482
0.015		
('DAI62779', 'GRO94758')		479
0.015		
('GRO85051', 'SNA80324')		471
0.015		
('DAI62779', 'GRO21487')		471
0.015		
('ELE17451', 'GRO30386')		468
0.015		
('FRO85978', 'SNA95666')		463
0.015		
('DAI62779', 'FRO19221')		462
0.015		
('DAI62779', 'GRO46854')		461
0.015		
('DAI43223', 'DAI62779')		459

```

0.015
('ELE92920', 'SNA18336')      |          455|
0.015
('DAI88079', 'FRO40251')      |          446|
0.014

```

To view the counters, check the 'User-Defined' category in the job logs appearing just prior to the reducer results. It should appear as follows:

```

User-Defined
  Number of Combiners=2
  Number of Mappers=1
  Number of Reducers=1

```

The job in total took 1 minute and 16 seconds to complete

## HW3.4 Solution Approach B

### Mapper

```

In [22]: %%writefile mapper-3-4.py
          #!/usr/bin/python
          import sys
          from itertools import combinations

          totalBaskets = 0 #field to hold total number of baskets
          sys.stderr.write('reporter:counter:Mapper-Counter,Total,1\n') #incr
          ement mapper counter
          for line in sys.stdin:
              totalBaskets += 1 #increment
              line = line.strip()
              line = line.split()

              pairs = list(combinations(line, 2)) #this give all pair combina
              tions for all items in a basket
              for pair in pairs:
                  pair = sorted(list(pair)) #sort the pairs in lexicographic
              order
                  print pair[0] + ' ' + pair[1] + '\t' + str(1) #print resul
              t: item1 + item2 + count of 1
          print '*' + '\t' + str(totalBaskets) #print total baskets

```

Overwriting mapper-3-4.py



## Reducer

```
In [23]: %%writefile reducer-3-4.py
#!/usr/bin/python
import sys
from collections import defaultdict
import operator

support = 100 #level of support
totalBaskets = 0 #total basket
pairs = defaultdict(int) #in-memory dictionary to hold results. This
wouldn't work as a scalable solution

sys.stderr.write('reporter:counter:Reducer-Counter,Total,1\n') #inc
rement counter
for line in sys.stdin:
    line = line.split('\t')
    if line[0] == '*': #order inversion says this will be the total
        totalBaskets = int(line[1])
    else:
        pairs[line[0]] += int(line[1]) #increment the default dicti
onary for pair by counter.
        #note when using a regular dictionary this takes an extreme
ly long time.

freqDict = {}
for pair, count in pairs.iteritems(): #filter dictionary for only i
tems with support greater than level set
    if count > support:
        freqDict[pair] = count

print "Top 50 item pairs:"
print '\n'
print 'Item Pair' + '\t' + 'Support Count' + '\t' + 'Relative Suppo
rt Count'
print '\n'

sortedFreqDict = sorted(freqDict.items(), key = lambda x: (-x[1], x
[0])) #sort results by number and by lexicographic order
for i in range(50):
    print sortedFreqDict[i][0] + '\t' + str(sortedFreqDict[i][1]) +
'\t' + str(float(sortedFreqDict[i][1])/totalBaskets)
```

Overwriting reducer-3-4.py

```
In [26]: time !hadoop jar /usr/local/Cellar/hadoop/2.7.1/libexec/share/hadoop
tools/lib/hadoop-streaming-2.7.1.jar \
-mapper mapper-3-4.py \
-reducer reducer-3-4.py \
-input ProductPurchaseData.txt \
-output sortedProducts
```



```
16/02/01 17:56:36 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/02/01 17:56:36 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
16/02/01 17:56:36 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/02/01 17:56:36 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
16/02/01 17:56:37 INFO mapred.FileInputFormat: Total input paths to process : 1
16/02/01 17:56:37 INFO mapreduce.JobSubmitter: number of splits:1
16/02/01 17:56:37 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local350127535_0001
16/02/01 17:56:37 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/02/01 17:56:37 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
16/02/01 17:56:37 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/02/01 17:56:37 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/01 17:56:37 INFO mapreduce.Job: Running job: job_local350127535_0001
16/02/01 17:56:37 INFO mapred.LocalJobRunner: Waiting for map tasks
16/02/01 17:56:37 INFO mapred.LocalJobRunner: Starting task: attempt_local350127535_0001_m_000000_0
16/02/01 17:56:37 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/01 17:56:37 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
16/02/01 17:56:37 INFO mapred.Task: Using ResourceCalculatorProcessTree : null
16/02/01 17:56:37 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/dunmireg/ProductPurchaseData.txt:0+3458517
16/02/01 17:56:37 INFO mapred.MapTask: numReduceTasks: 1
16/02/01 17:56:37 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/02/01 17:56:37 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/02/01 17:56:37 INFO mapred.MapTask: soft limit at 83886080
16/02/01 17:56:37 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/02/01 17:56:37 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/02/01 17:56:37 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
16/02/01 17:56:37 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/dunmireg/Documents/261HW/HW3/./mapper-3-4.py]
16/02/01 17:56:37 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
16/02/01 17:56:37 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir
```

```
16/02/01 17:56:37 INFO Configuration.deprecation: map.input.file i
s deprecated. Instead, use mapreduce.map.input.file
16/02/01 17:56:37 INFO Configuration.deprecation: mapred.skip.on i
s deprecated. Instead, use mapreduce.job.skiprecords
16/02/01 17:56:37 INFO Configuration.deprecation: map.input.length
is deprecated. Instead, use mapreduce.map.input.length
16/02/01 17:56:37 INFO Configuration.deprecation: mapred.work.outp
ut.dir is deprecated. Instead, use mapreduce.task.output.dir
16/02/01 17:56:37 INFO Configuration.deprecation: map.input.start
is deprecated. Instead, use mapreduce.map.input.start
16/02/01 17:56:37 INFO Configuration.deprecation: mapred.job.id is
deprecated. Instead, use mapreduce.job.id
16/02/01 17:56:37 INFO Configuration.deprecation: user.name is dep
recated. Instead, use mapreduce.job.user.name
16/02/01 17:56:37 INFO Configuration.deprecation: mapred.task.is.m
ap is deprecated. Instead, use mapreduce.task.ismap
16/02/01 17:56:37 INFO Configuration.deprecation: mapred.task.id i
s deprecated. Instead, use mapreduce.task.attempt.id
16/02/01 17:56:37 INFO Configuration.deprecation: mapred.task.part
ition is deprecated. Instead, use mapreduce.task.partition
16/02/01 17:56:37 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/02/01 17:56:37 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/02/01 17:56:37 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/02/01 17:56:37 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/02/01 17:56:37 INFO streaming.PipeMapRed: Records R/W=1216/1
16/02/01 17:56:38 INFO mapreduce.Job: Job job_local350127535_0001
running in uber mode : false
16/02/01 17:56:38 INFO mapreduce.Job: map 0% reduce 0%
16/02/01 17:56:39 INFO streaming.PipeMapRed: R/W/S=10000/854425/0
in:10000=10000/1 [rec/s] out:854425=854425/1 [rec/s]
16/02/01 17:56:41 INFO mapred.MapTask: Spilling map output
16/02/01 17:56:41 INFO mapred.MapTask: bufstart = 0; bufend = 4660
3380; bufvoid = 104857600
16/02/01 17:56:41 INFO mapred.MapTask: kvstart = 26214396(10485758
4); kvend = 16893724(67574896); length = 9320673/6553600
16/02/01 17:56:41 INFO mapred.MapTask: (EQUATOR) 55924036 kvi 1398
1004(55924016)
16/02/01 17:56:42 INFO streaming.PipeMapRed: MRErrorThread done
16/02/01 17:56:42 INFO streaming.PipeMapRed: mapRedFinished
16/02/01 17:56:42 INFO mapred.LocalJobRunner:
16/02/01 17:56:42 INFO mapred.MapTask: Starting flush of map output
t
16/02/01 17:56:43 INFO mapred.LocalJobRunner: Records R/W=1216/1 >
sort
16/02/01 17:56:44 INFO mapreduce.Job: map 67% reduce 0%
16/02/01 17:56:46 INFO mapred.MapTask: Finished spill 0
16/02/01 17:56:46 INFO mapred.MapTask: (RESET) equator 55924036 kv
13981004(55924016) kvi 13165448(52661792)
16/02/01 17:56:46 INFO mapred.MapTask: Spilling map output
16/02/01 17:56:46 INFO mapred.MapTask: bufstart = 55924036; bufend
```

```
= 60001804; bufvoid = 104857600
16/02/01 17:56:46 INFO mapred.MapTask: kvstart = 13981004(5592401
6); kvend = 13165452(52661808); length = 815553/6553600
16/02/01 17:56:46 INFO mapred.MapTask: Finished spill 1
16/02/01 17:56:46 INFO mapred.Merger: Merging 2 sorted segments
16/02/01 17:56:46 INFO mapred.Merger: Down to the last merge-pass,
with 2 segments left of total size: 55749252 bytes
16/02/01 17:56:46 INFO mapred.LocalJobRunner: Records R/W=1216/1 >
sort >
16/02/01 17:56:47 INFO mapreduce.Job: map 75% reduce 0%
16/02/01 17:56:47 INFO mapred.Task: Task:attempt_local350127535_00
01_m_000000_0 is done. And is in the process of committing
16/02/01 17:56:47 INFO mapred.LocalJobRunner: Records R/W=1216/1 >
sort
16/02/01 17:56:47 INFO mapred.Task: Task 'attempt_local350127535_0
001_m_000000_0' done.
16/02/01 17:56:47 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local350127535_0001_m_000000_0
16/02/01 17:56:47 INFO mapred.LocalJobRunner: map task executor co
mplete.
16/02/01 17:56:47 INFO mapred.LocalJobRunner: Waiting for reduce t
asks
16/02/01 17:56:47 INFO mapred.LocalJobRunner: Starting task: attem
pt_local350127535_0001_r_000000_0
16/02/01 17:56:47 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
16/02/01 17:56:47 INFO util.ProcfsBasedProcessTree: ProcfsBasedPro
cessTree currently is supported only on Linux.
16/02/01 17:56:47 INFO mapred.Task: Using ResourceCalculatorProce
ssTree : null
16/02/01 17:56:47 INFO mapred.ReduceTask: Using ShuffleConsumerPlu
gin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@1a05e6dd
16/02/01 17:56:47 INFO reduce.MergeManagerImpl: MergeManager: mem
oryLimit=334338464, maxSingleShuffleLimit=83584616, mergeThreshold
=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
16/02/01 17:56:47 INFO reduce.EventFetcher: attempt_local350127535
_0001_r_000000_0 Thread started: EventFetcher for fetching Map Com
pletion Events
16/02/01 17:56:47 INFO reduce.LocalFetcher: localfetcher#1 about t
o shuffle output of map attempt_local350127535_0001_m_000000_0 dec
omp: 55749266 len: 55749270 to MEMORY
16/02/01 17:56:47 INFO reduce.InMemoryMapOutput: Read 55749266 byt
es from map-output for attempt_local350127535_0001_m_000000_0
16/02/01 17:56:47 INFO reduce.MergeManagerImpl: closeInMemoryFile
-> map-output of size: 55749266, inMemoryMapOutputs.size() -> 1, c
ommitMemory -> 0, usedMemory ->55749266
16/02/01 17:56:47 INFO reduce.EventFetcher: EventFetcher is interr
upted.. Returning
16/02/01 17:56:47 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/01 17:56:47 INFO reduce.MergeManagerImpl: finalMerge called
with 1 in-memory map-outputs and 0 on-disk map-outputs
16/02/01 17:56:47 INFO mapred.Merger: Merging 1 sorted segments
16/02/01 17:56:47 INFO mapred.Merger: Down to the last merge-pass,
with 1 segments left of total size: 55749262 bytes
```

```
16/02/01 17:56:48 INFO mapreduce.Job: map 100% reduce 0%
16/02/01 17:56:48 INFO reduce.MergeManagerImpl: Merged 1 segments,
55749266 bytes to disk to satisfy reduce memory limit
16/02/01 17:56:48 INFO reduce.MergeManagerImpl: Merging 1 files, 5
5749270 bytes from disk
16/02/01 17:56:48 INFO reduce.MergeManagerImpl: Merging 0 segment
s, 0 bytes from memory into reduce
16/02/01 17:56:48 INFO mapred.Merger: Merging 1 sorted segments
16/02/01 17:56:48 INFO mapred.Merger: Down to the last merge-pass,
with 1 segments left of total size: 55749262 bytes
16/02/01 17:56:48 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/01 17:56:48 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/dunmireg/Documents/261HW/HW3/./reducer-3-4.py]
16/02/01 17:56:48 INFO Configuration.deprecation: mapred.job.track
er is deprecated. Instead, use mapreduce.jobtracker.address
16/02/01 17:56:48 INFO Configuration.deprecation: mapred.map.tasks
is deprecated. Instead, use mapreduce.job.maps
16/02/01 17:56:48 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/02/01 17:56:48 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/02/01 17:56:48 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/02/01 17:56:48 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/02/01 17:56:48 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/02/01 17:56:49 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/02/01 17:56:49 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/02/01 17:56:49 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/02/01 17:56:49 INFO streaming.PipeMapRed: R/W/S=400000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/02/01 17:56:49 INFO streaming.PipeMapRed: R/W/S=500000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/02/01 17:56:49 INFO streaming.PipeMapRed: R/W/S=600000/0/0 in:6
00000=600000/1 [rec/s] out:0=0/1 [rec/s]
16/02/01 17:56:50 INFO streaming.PipeMapRed: R/W/S=700000/0/0 in:7
00000=700000/1 [rec/s] out:0=0/1 [rec/s]
16/02/01 17:56:50 INFO streaming.PipeMapRed: R/W/S=800000/0/0 in:8
00000=800000/1 [rec/s] out:0=0/1 [rec/s]
16/02/01 17:56:50 INFO streaming.PipeMapRed: R/W/S=900000/0/0 in:9
00000=900000/1 [rec/s] out:0=0/1 [rec/s]
16/02/01 17:56:50 INFO streaming.PipeMapRed: R/W/S=1000000/0/0 in:
1000000=1000000/1 [rec/s] out:0=0/1 [rec/s]
16/02/01 17:56:50 INFO streaming.PipeMapRed: R/W/S=1100000/0/0 in:
1100000=1100000/1 [rec/s] out:0=0/1 [rec/s]
16/02/01 17:56:50 INFO streaming.PipeMapRed: R/W/S=1200000/0/0 in:
600000=1200000/2 [rec/s] out:0=0/2 [rec/s]
16/02/01 17:56:51 INFO streaming.PipeMapRed: R/W/S=1300000/0/0 in:
650000=1300000/2 [rec/s] out:0=0/2 [rec/s]
16/02/01 17:56:51 INFO streaming.PipeMapRed: R/W/S=1400000/0/0 in:
```

```
700000=1400000/2 [rec/s] out:0=0/2 [rec/s]
16/02/01 17:56:51 INFO streaming.PipeMapRed: R/W/S=1500000/0/0 in:
750000=1500000/2 [rec/s] out:0=0/2 [rec/s]
16/02/01 17:56:51 INFO streaming.PipeMapRed: R/W/S=1600000/0/0 in:
800000=1600000/2 [rec/s] out:0=0/2 [rec/s]
16/02/01 17:56:51 INFO streaming.PipeMapRed: R/W/S=1700000/0/0 in:
850000=1700000/2 [rec/s] out:0=0/2 [rec/s]
16/02/01 17:56:51 INFO streaming.PipeMapRed: R/W/S=1800000/0/0 in:
600000=1800000/3 [rec/s] out:0=0/3 [rec/s]
16/02/01 17:56:52 INFO streaming.PipeMapRed: R/W/S=1900000/0/0 in:
633333=1900000/3 [rec/s] out:0=0/3 [rec/s]
16/02/01 17:56:52 INFO streaming.PipeMapRed: R/W/S=2000000/0/0 in:
666666=2000000/3 [rec/s] out:0=0/3 [rec/s]
16/02/01 17:56:52 INFO streaming.PipeMapRed: R/W/S=2100000/0/0 in:
700000=2100000/3 [rec/s] out:0=0/3 [rec/s]
16/02/01 17:56:52 INFO streaming.PipeMapRed: R/W/S=2200000/0/0 in:
733333=2200000/3 [rec/s] out:0=0/3 [rec/s]
16/02/01 17:56:52 INFO streaming.PipeMapRed: R/W/S=2300000/0/0 in:
766666=2300000/3 [rec/s] out:0=0/3 [rec/s]
16/02/01 17:56:53 INFO streaming.PipeMapRed: R/W/S=2400000/0/0 in:
600000=2400000/4 [rec/s] out:0=0/4 [rec/s]
16/02/01 17:56:53 INFO streaming.PipeMapRed: R/W/S=2500000/0/0 in:
625000=2500000/4 [rec/s] out:0=0/4 [rec/s]
16/02/01 17:56:53 INFO mapred.LocalJobRunner: reduce > reduce
16/02/01 17:56:54 INFO mapreduce.Job: map 100% reduce 100%
16/02/01 17:56:54 INFO streaming.PipeMapRed: Records R/W=2534058/1
16/02/01 17:56:54 INFO streaming.PipeMapRed: MRErrorThread done
16/02/01 17:56:54 INFO streaming.PipeMapRed: mapRedFinished
16/02/01 17:56:55 INFO mapred.Task: Task:attempt_local350127535_00
01_r_000000_0 is done. And is in the process of committing
16/02/01 17:56:55 INFO mapred.LocalJobRunner: reduce > reduce
16/02/01 17:56:55 INFO mapred.Task: Task attempt_local350127535_00
01_r_000000_0 is allowed to commit now
16/02/01 17:56:55 INFO output.FileOutputCommitter: Saved output of
task 'attempt_local350127535_0001_r_000000_0' to hdfs://localhost:
9000/user/dunmireg/sortedProducts/_temporary/0/task_local350127535
_0001_r_000000
16/02/01 17:56:55 INFO mapred.LocalJobRunner: Records R/W=2534058/
1 > reduce
16/02/01 17:56:55 INFO mapred.Task: Task 'attempt_local350127535_0
001_r_000000_0' done.
16/02/01 17:56:55 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local350127535_0001_r_000000_0
16/02/01 17:56:55 INFO mapred.LocalJobRunner: reduce task executor
complete.
16/02/01 17:56:55 INFO mapreduce.Job: Job job_local350127535_0001
completed successfully
16/02/01 17:56:55 INFO mapreduce.Job: Counters: 35
```

#### File System Counters

```
FILE: Number of bytes read=223209220
FILE: Number of bytes written=279545226
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
```



HDFS: Number of bytes read=6917034  
HDFS: Number of bytes written=1973  
HDFS: Number of read operations=13  
HDFS: Number of large read operations=0  
HDFS: Number of write operations=4

#### Map-Reduce Framework

Map input records=31101  
Map output records=2534058  
Map output bytes=50681148  
Map output materialized bytes=55749270  
Input split bytes=111  
Combine input records=0  
Combine output records=0  
Reduce input groups=877099  
Reduce shuffle bytes=55749270  
Reduce input records=2534058  
Reduce output records=56  
Spilled Records=7602174  
Shuffled Maps =1  
Failed Shuffles=0  
Merged Map outputs=1  
GC time elapsed (ms)=6  
Total committed heap usage (bytes)=620756992

#### Shuffle Errors

BAD\_ID=0  
CONNECTION=0  
IO\_ERROR=0  
WRONG\_LENGTH=0  
WRONG\_MAP=0  
WRONG\_REDUCE=0

#### File Input Format Counters

Bytes Read=3458517

#### File Output Format Counters

Bytes Written=1973

16/02/01 17:56:55 INFO streaming.StreamJob: Output directory: sortedProducts

CPU times: user 107 ms, sys: 28.8 ms, total: 136 ms

Wall time: 20.6 s

```
In [27]: #show results  
!hdfs dfs -cat /user/dunmireg/sortedProducts/part-00000
```



16/02/01 17:57:05 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

Top 50 item pairs:

Item Pair	Support Count	Relative Support Count
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

The total time taken is displayed here:

CPU times: user 112 ms, sys: 28.3 ms, total: 140 ms Wall time: 21.4 s

This code used 1 mapper and 1 reducer. This was run on an 8 GB Macbook Air with a 2.2 GHz Intel Core i7 quad core processor.

---

## HW3.5

*Repeat 3.4 using the stripes design pattern for finding cooccurring pairs.*

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

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

## Solution Approach A:

### Mapper

The mapper for the stripes implementation is noticeably different than the one used for the pairs implementation. In particular, a dictionary stripe is emitted which is then parsed literally by the reducer or optional combiner. From here, the stripes are aggregated per token then divided by two to compensate for the two combinations in which a key can be updated.

```
In [1]: %%writefile mapperQ35.py
#!/usr/bin/env python

import sys
from collections import defaultdict
from itertools import combinations

def readInput(file, separator=None):
    for line in file:
        yield line.split(separator)

if __name__ == "__main__":

    # Read input
    data = readInput(sys.stdin)
    totalBaskets = 0

    for line in data:

        totalBaskets += 1

        for token in line:

            occurrence = defaultdict(int)

            # Remove token from neighbors
            stripe = [x for x in line if x != token]

            # Create co-occurrence array
            for neighbor in stripe:
                occurrence[neighbor] += 1

            # Emit
            cArray = dict(occurrence)

            print '%s%s%s' % (token, '\t', str(cArray))

    # Update counter
    sys.stderr.write("reporter:counter:User-Defined,Number of M
appers,1\n")

    # Emit basket count
    print '%s%s%s' % ('*', '\t', str(totalBaskets))
```

```
Overwriting mapperQ35.py
```

## Reducer

Here, the reducer has been changed to compensate for the adjusted input. The reducer aggregates the stripes by aggregating the incoming dictionaries. It does this efficiently using `defaultdict`, an automatically instantiating key-value dictionary included in `collections`.

In [8]:



```
%%writefile reducerQ35.py
#!/usr/bin/env python

from __future__ import division
import sys
from collections import defaultdict
from itertools import combinations, chain
import ast

# Read input from mapper
def readInput(file, separator='\t'):
    for line in file:
        yield line.split(separator)

if __name__ == "__main__":

    tokenDict = defaultdict(list)
    mergedDict = defaultdict(dict)
    tupleList = defaultdict(float)
    flattenedTerms = defaultdict(int)
    support = 100
    totalBaskets = 0

    # Read data
    data = readInput(sys.stdin)
    for line in data:

        token = line[0]
        if token == '*':

            # Increment basket
            totalBaskets += int(line[1])

        else:

            # Parse normally
            stripe = ast.literal_eval(line[1])

            # Combine dictionaries
            tokenDict[token].append(stripe)

    # Merge stripes
    for k, v in tokenDict.iteritems():
        merged = defaultdict(int)

        # Loop and aggregate
        for stripe in v:
            for stripeKey in stripe:
                merged[stripeKey] += stripe[stripeK
```

```

ey]

mergedDict[k] = merged

# Create key-value pairs
for token, stripe in mergedDict.iteritems():
    for innerToken, count in stripe.iteritems():

        # Get unique keys
        tokenPair = [token, innerToken]
        tokenPair.sort()
        tuplePair = tuple(tokenPair)

        # Overcounting exactly twice per pair
        tupleList[tuplePair] += count / 2

# Find most frequent terms and filter
mostFrequentTerms = [(k, int(v), round(int(v) / totalBasket
s, 3)) for k, v in tupleList.iteritems()
                     if int(v) >= support]

mostFrequentTerms = sorted(mostFrequentTerms,
                           key
                           = lambda x: x[1],
                           rev
                           erse = True)

# Get results
print '\n' + '===== Most Frequent Terms =====' +
'\n'
template = "{0:30}|{1:20}|{2:20}"
print template.format("PAIR", "SUPPORT COUNT", "RELATIVE FR
EQUENCY")

# Print terms
for termPair in mostFrequentTerms[:50]:
    print template.format(*termPair)

# Update counter
sys.stderr.write("reporter:counter:User-Defined,Number of R
educers,1\n")

```

Overwriting reducerQ35.py

## Combiner

Below is the combiner used optionally in the process. It aggregates in a similar way to the reduce and outputs partially aggregated stripes.

In [3]:

```
%%writefile combinerQ35.py
#!/usr/bin/env python

from __future__ import division
import sys
from collections import defaultdict
import ast

# Read input from mapper
def readInput(file, separator='\t'):
    for line in file:
        yield line.split(separator)

if __name__ == "__main__":

    tokenDict = defaultdict(list)
    mergedDict = defaultdict(dict)
    totalBaskets = 0

    # Read data
    data = readInput(sys.stdin)
    for line in data:

        token = line[0]
        if token == '*':

            # Increment basket
            totalBaskets += int(line[1])

        else:

            # Parse normally
            stripe = ast.literal_eval(line[1])

            # Combine dictionaries
            tokenDict[token].append(stripe)

    # # Merge stripes
    for k, v in tokenDict.iteritems():
        merged = defaultdict(int)

        # Loop and aggregate
        for stripe in v:
            for stripeKey in stripe:
                merged[stripeKey] += stripe[stripeKey]

        mergedDict[k] = dict(merged)

    # Emit results
    for k, v in mergedDict.iteritems():
```

```
print '%s%s%s' % (k, '\t', str(v))

# Update counter
sys.stderr.write("reporter:counter:User-Defined,Number of Combiners,1\n")

# Emit baskets
print '%s%s%s' % ('*', '\t', str(totalBaskets))
```

Overwriting combinerQ35.py

### **Submit Job**

Our wrapper remains unchanged from the prior implementation that has an argument input for the combiner. We re-use it here:

```
In [9]: !bash wrapperQ34.sh ProductPurchaseData.txt mapperQ35.py reducerQ35.py combinerQ35.py
```





```
16/02/01 19:50:32 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/02/01 19:50:34 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/john/notebook
16/02/01 19:50:36 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/02/01 19:50:39 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/02/01 19:50:42 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
16/02/01 19:50:42 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
packageJobJar: [/Users/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/mapperQ35.py, /Users/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/reducerQ35.py, /Users/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/combinerQ35.py] [] /var/folders/0w/8hzv7rsj3qgdynsjlqy3gjsc0000gn/T/streamjob6872178788643398508.jar tmpDir=null
16/02/01 19:50:44 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
16/02/01 19:50:44 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/02/01 19:50:44 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
16/02/01 19:50:44 INFO mapred.FileInputFormat: Total input paths to process : 1
16/02/01 19:50:44 INFO mapreduce.JobSubmitter: number of splits:1
16/02/01 19:50:45 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local150026337_0001
16/02/01 19:50:45 INFO mapred.LocalDistributedCacheManager: Localized file:/Users/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/mapperQ35.py as file:/usr/local/Cellar/hadoop/hdfs/tmp/mapred/local/1454374245453/mapperQ35.py
16/02/01 19:50:45 INFO mapred.LocalDistributedCacheManager: Localized file:/Users/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/reducerQ35.py as file:/usr/local/Cellar/hadoop/hdfs/tmp/mapred/local/1454374245454/reducerQ35.py
16/02/01 19:50:45 INFO mapred.LocalDistributedCacheManager: Localized file:/Users/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/combinerQ35.py as file:/usr/local/Cellar/hadoop/hdfs/tmp/mapred/local/1454374245455/combinerQ35.py
16/02/01 19:50:46 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/02/01 19:50:46 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/02/01 19:50:46 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
16/02/01 19:50:46 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
```

```
16/02/01 19:50:46 INFO mapreduce.Job: Running job: job_local150026337_0001
16/02/01 19:50:46 INFO mapred.LocalJobRunner: Waiting for map tasks
16/02/01 19:50:46 INFO mapred.LocalJobRunner: Starting task: attempt_local150026337_0001_m_000000_0
16/02/01 19:50:46 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/01 19:50:46 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
16/02/01 19:50:46 INFO mapred.Task: Using ResourceCalculatorProcessTree : null
16/02/01 19:50:46 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/john/notebook/input/ProductPurchaseData.txt:0+3458517
16/02/01 19:50:46 INFO mapred.MapTask: numReduceTasks: 1
16/02/01 19:50:46 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/02/01 19:50:46 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/02/01 19:50:46 INFO mapred.MapTask: soft limit at 83886080
16/02/01 19:50:46 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/02/01 19:50:46 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/02/01 19:50:46 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
16/02/01 19:50:46 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/./mapperQ35.py]
16/02/01 19:50:46 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.output.dir
16/02/01 19:50:46 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
16/02/01 19:50:46 INFO Configuration.deprecation: mapred.task.ismap is deprecated. Instead, use mapreduce.task.ismap
16/02/01 19:50:46 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
16/02/01 19:50:46 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
16/02/01 19:50:46 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir
16/02/01 19:50:46 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
16/02/01 19:50:46 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
16/02/01 19:50:46 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.length
16/02/01 19:50:46 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
16/02/01 19:50:46 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
16/02/01 19:50:46 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
16/02/01 19:50:46 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
```

```
c/s] out:NA [rec/s]
16/02/01 19:50:46 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/02/01 19:50:46 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/02/01 19:50:46 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/02/01 19:50:46 INFO streaming.PipeMapRed: Records R/W=1216/1
16/02/01 19:50:47 INFO mapreduce.Job: Job job_local150026337_0001
running in uber mode : false
16/02/01 19:50:47 INFO mapreduce.Job: map 0% reduce 0%
16/02/01 19:50:49 INFO streaming.PipeMapRed: R/W/S=10000/120238/0
in:3333=10000/3 [rec/s] out:40079=120238/3 [rec/s]
16/02/01 19:50:52 INFO mapred.LocalJobRunner: Records R/W=1216/1 >
map
16/02/01 19:50:53 INFO mapreduce.Job: map 38% reduce 0%
16/02/01 19:50:55 INFO mapred.LocalJobRunner: Records R/W=1216/1 >
map
16/02/01 19:50:56 INFO mapreduce.Job: map 53% reduce 0%
16/02/01 19:50:56 INFO streaming.PipeMapRed: Records R/W=26917/313
473
16/02/01 19:50:57 INFO mapred.MapTask: Spilling map output
16/02/01 19:50:57 INFO mapred.MapTask: bufstart = 0; bufend = 7796
5406; bufvoid = 104857600
16/02/01 19:50:57 INFO mapred.MapTask: kvstart = 26214396(10485758
4); kvend = 24734212(98936848); length = 1480185/6553600
16/02/01 19:50:57 INFO mapred.MapTask: (EQUATOR) 79450094 kvi 1986
2516(79450064)
16/02/01 19:50:58 INFO mapred.LocalJobRunner: Records R/W=26917/31
3473 > map
16/02/01 19:50:58 INFO streaming.PipeMapRed: MRErrorThread done
16/02/01 19:50:58 INFO streaming.PipeMapRed: mapRedFinished
16/02/01 19:50:58 INFO mapred.LocalJobRunner: Records R/W=26917/31
3473 > map
16/02/01 19:50:58 INFO mapred.MapTask: Starting flush of map outpu
t
16/02/01 19:50:58 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/./combinerQ35.p
y]
16/02/01 19:50:58 INFO Configuration.deprecation: mapred.skip.map.
auto.incr.proc.count is deprecated. Instead, use mapreduce.map.ski
p.proc-count.auto-incr
16/02/01 19:50:58 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/02/01 19:50:58 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/02/01 19:50:58 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/02/01 19:50:59 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/02/01 19:50:59 INFO mapreduce.Job: map 67% reduce 0%
16/02/01 19:51:00 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:10
000=10000/1 [rec/s] out:0=0/1 [rec/s]
16/02/01 19:51:01 INFO mapred.LocalJobRunner: Records R/W=26917/31
```

```
3473 > sort
16/02/01 19:51:04 INFO mapred.LocalJobRunner: Records R/W=26917/31
3473 > sort
16/02/01 19:51:07 INFO mapred.LocalJobRunner: Records R/W=26917/31
3473 > sort
16/02/01 19:51:10 INFO mapred.LocalJobRunner: Records R/W=26917/31
3473 > sort
16/02/01 19:51:11 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:8
333=100000/12 [rec/s] out:0=0/12 [rec/s]
16/02/01 19:51:13 INFO mapred.LocalJobRunner: Records R/W=26917/31
3473 > sort
16/02/01 19:51:16 INFO mapred.LocalJobRunner: Records R/W=26917/31
3473 > sort
16/02/01 19:51:19 INFO mapred.LocalJobRunner: Records R/W=26917/31
3473 > sort
16/02/01 19:51:22 INFO mapred.LocalJobRunner: Records R/W=26917/31
3473 > sort
16/02/01 19:51:25 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:7
692=200000/26 [rec/s] out:0=0/26 [rec/s]
16/02/01 19:51:25 INFO mapred.LocalJobRunner: Records R/W=26917/31
3473 > sort
16/02/01 19:51:28 INFO mapred.LocalJobRunner: Records R/W=26917/31
3473 > sort
16/02/01 19:51:31 INFO mapred.LocalJobRunner: Records R/W=26917/31
3473 > sort
16/02/01 19:51:34 INFO mapred.LocalJobRunner: Records R/W=26917/31
3473 > sort
16/02/01 19:51:37 INFO mapred.LocalJobRunner: Records R/W=26917/31
3473 > sort
16/02/01 19:51:38 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:7
692=300000/39 [rec/s] out:0=0/39 [rec/s]
16/02/01 19:51:40 INFO mapred.LocalJobRunner: Records R/W=26917/31
3473 > sort
16/02/01 19:51:43 INFO mapred.LocalJobRunner: Records R/W=26917/31
3473 > sort
16/02/01 19:51:46 INFO mapred.LocalJobRunner: Records R/W=26917/31
3473 > sort
16/02/01 19:51:49 INFO mapred.LocalJobRunner: Records R/W=26917/31
3473 > sort
16/02/01 19:51:52 INFO streaming.PipeMapRed: Records R/W=370047/1
16/02/01 19:51:52 INFO mapred.LocalJobRunner: Records R/W=370047/1
> sort
16/02/01 19:51:54 INFO streaming.PipeMapRed: MRErrorThread done
16/02/01 19:51:54 INFO streaming.PipeMapRed: mapRedFinished
16/02/01 19:51:54 INFO mapred.MapTask: Finished spill 0
16/02/01 19:51:54 INFO mapred.MapTask: (RESET) equator 79450094 kv
19862516(79450064) kvi 19819404(79277616)
16/02/01 19:51:54 INFO mapred.MapTask: Spilling map output
16/02/01 19:51:54 INFO mapred.MapTask: bufstart = 79450094; bufend
= 81699127; bufvoid = 104857600
16/02/01 19:51:54 INFO mapred.MapTask: kvstart = 19862516(7945006
4); kvend = 19819408(79277632); length = 43109/6553600
16/02/01 19:51:54 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/./combinerQ35.p
```

```
y]
16/02/01 19:51:54 INFO Configuration.deprecation: mapred.skip.reduce.auto.incr.proc.count is deprecated. Instead, use mapreduce.reduce.skip.proc-count.auto-incr
16/02/01 19:51:54 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/01 19:51:54 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/01 19:51:54 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/01 19:51:54 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/01 19:51:55 INFO mapred.LocalJobRunner: Records R/W=370047/1
> sort
16/02/01 19:51:55 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:10000=10000/1 [rec/s] out:0=0/1 [rec/s]
16/02/01 19:51:55 INFO streaming.PipeMapRed: Records R/W=10778/1
16/02/01 19:51:55 INFO streaming.PipeMapRed: MRErrorThread done
16/02/01 19:51:55 INFO streaming.PipeMapRed: mapRedFinished
16/02/01 19:51:55 INFO mapred.MapTask: Finished spill 1
16/02/01 19:51:55 INFO mapred.Merger: Merging 2 sorted segments
16/02/01 19:51:55 INFO mapred.Merger: Down to the last merge-pass, with 2 segments left of total size: 27515029 bytes
16/02/01 19:51:56 INFO mapred.Task: Task:attempt_local150026337_001_m_000000_0 is done. And is in the process of committing
16/02/01 19:51:56 INFO mapred.LocalJobRunner: Records R/W=10778/1
> sort
16/02/01 19:51:56 INFO mapred.Task: Task 'attempt_local150026337_001_m_000000_0' done.
16/02/01 19:51:56 INFO mapred.LocalJobRunner: Finishing task: attempt_local150026337_0001_m_000000_0
16/02/01 19:51:56 INFO mapred.LocalJobRunner: map task executor complete.
16/02/01 19:51:56 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/02/01 19:51:56 INFO mapred.LocalJobRunner: Starting task: attempt_local150026337_0001_r_000000_0
16/02/01 19:51:56 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/01 19:51:56 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
16/02/01 19:51:56 INFO mapred.Task: Using ResourceCalculatorProcessTree : null
16/02/01 19:51:56 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@32592bbb
16/02/01 19:51:56 INFO mapreduce.Job: map 100% reduce 0%
16/02/01 19:51:56 INFO reduce.MergeManagerImpl: MergeManager: memoryLimit=334338464, maxSingleShuffleLimit=83584616, mergeThreshold=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
16/02/01 19:51:56 INFO reduce.EventFetcher: attempt_local150026337_0001_r_000000_0 Thread started: EventFetcher for fetching Map Completion Events
16/02/01 19:51:56 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local150026337_0001_m_000000_0 dec
```

```
omp: 27515044 len: 27515048 to MEMORY
16/02/01 19:51:56 INFO reduce.InMemoryMapOutput: Read 27515044 bytes from map-output for attempt_local150026337_0001_m_000000_0
16/02/01 19:51:56 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 27515044, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory -> 27515044
16/02/01 19:51:56 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/02/01 19:51:56 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/01 19:51:56 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs
16/02/01 19:51:56 INFO mapred.Merger: Merging 1 sorted segments
16/02/01 19:51:56 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 27515031 bytes
16/02/01 19:51:56 INFO reduce.MergeManagerImpl: Merged 1 segments, 27515044 bytes to disk to satisfy reduce memory limit
16/02/01 19:51:56 INFO reduce.MergeManagerImpl: Merging 1 files, 27515048 bytes from disk
16/02/01 19:51:57 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/02/01 19:51:57 INFO mapred.Merger: Merging 1 sorted segments
16/02/01 19:51:57 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 27515031 bytes
16/02/01 19:51:57 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/01 19:51:57 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/./reducerQ35.py]
16/02/01 19:51:57 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
16/02/01 19:51:57 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
16/02/01 19:51:57 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/01 19:51:57 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/01 19:51:57 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/01 19:51:57 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/01 19:52:02 INFO mapred.LocalJobRunner: reduce > reduce
16/02/01 19:52:02 INFO mapreduce.Job: map 100% reduce 78%
16/02/01 19:52:05 INFO mapred.LocalJobRunner: reduce > reduce
16/02/01 19:52:05 INFO mapreduce.Job: map 100% reduce 84%
16/02/01 19:52:07 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:1111=10000/9 [rec/s] out:0=0/9 [rec/s]
16/02/01 19:52:08 INFO mapred.LocalJobRunner: reduce > reduce
16/02/01 19:52:08 INFO mapreduce.Job: map 100% reduce 91%
16/02/01 19:52:11 INFO mapred.LocalJobRunner: reduce > reduce
16/02/01 19:52:11 INFO mapreduce.Job: map 100% reduce 97%
16/02/01 19:52:14 INFO mapred.LocalJobRunner: reduce > reduce
16/02/01 19:52:14 INFO mapreduce.Job: map 100% reduce 100%
16/02/01 19:52:17 INFO mapred.LocalJobRunner: reduce > reduce
16/02/01 19:52:22 INFO streaming.PipeMapRed: Records R/W=14619/1
16/02/01 19:52:22 INFO streaming.PipeMapRed: MRErrorThread done
```

```
16/02/01 19:52:22 INFO streaming.PipeMapRed: mapRedFinished
16/02/01 19:52:22 INFO mapred.Task: Task:attempt_local150026337_00
01_r_000000_0 is done. And is in the process of committing
16/02/01 19:52:22 INFO mapred.LocalJobRunner: reduce > reduce
16/02/01 19:52:22 INFO mapred.Task: Task attempt_local150026337_00
01_r_000000_0 is allowed to commit now
16/02/01 19:52:22 INFO output.FileOutputCommitter: Saved output of
task 'attempt_local150026337_0001_r_000000_0' to hdfs://localhost:
9000/user/john/notebook/output/_temporary/0/task_local150026337_00
01_r_000000
16/02/01 19:52:22 INFO mapred.LocalJobRunner: Records R/W=14619/1
> reduce
16/02/01 19:52:22 INFO mapred.Task: Task 'attempt_local150026337_0
001_r_000000_0' done.
16/02/01 19:52:22 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local150026337_0001_r_000000_0
16/02/01 19:52:22 INFO mapred.LocalJobRunner: reduce task executor
complete.
16/02/01 19:52:23 INFO mapreduce.Job: Job job_local150026337_0001
completed successfully
16/02/01 19:52:23 INFO mapreduce.Job: Counters: 38
```

#### File System Counters

```
FILE: Number of bytes read=110073226
FILE: Number of bytes written=138175744
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=6917034
HDFS: Number of bytes written=3821
HDFS: Number of read operations=13
HDFS: Number of large read operations=0
HDFS: Number of write operations=4
```

#### Map-Reduce Framework

```
Map input records=31101
Map output records=380825
Map output bytes=80214439
Map output materialized bytes=27515048
Input split bytes=122
Combine input records=380825
Combine output records=14619
Reduce input groups=14617
Reduce shuffle bytes=27515048
Reduce input records=14619
Reduce output records=54
Spilled Records=43857
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=26
Total committed heap usage (bytes)=534249472
```

#### Shuffle Errors

```
BAD_ID=0
CONNECTION=0
IO_ERROR=0
```

WRONG\_LENGTH=0

WRONG\_MAP=0

WRONG\_REDUCE=0

User-Defined

Number of Combiners=2

Number of Mappers=1

Number of Reducers=1

File Input Format Counters

Bytes Read=3458517

File Output Format Counters

Bytes Written=3821

16/02/01 19:52:23 INFO streaming.StreamJob: Output directory: /user/john/notebook/output

16/02/01 19:52:25 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

===== Most Frequent Terms =====

PAIR ENCY	SUPPORT COUNT	RELATIVE FREQUENCY
('DAI62779', 'ELE17451')	1592	0.051
('FRO40251', 'SNA80324')	1412	0.045
('DAI75645', 'FRO40251')	1254	0.04
('FRO40251', 'GRO85051')	1213	0.039
('DAI62779', 'GRO73461')	1139	0.037
('DAI75645', 'SNA80324')	1130	0.036
('DAI62779', 'FRO40251')	1070	0.034
('DAI62779', 'SNA80324')	923	0.03
('DAI62779', 'DAI85309')	918	0.03
('ELE32164', 'GRO59710')	911	0.029
('FRO40251', 'GRO73461')	882	0.028
('DAI62779', 'DAI75645')	882	0.028
('DAI62779', 'ELE92920')	877	0.028
('FRO40251', 'FRO92469')	835	0.027
('DAI62779', 'ELE32164')	832	0.027
('DAI75645', 'GRO73461')	712	0.023
('DAI43223', 'ELE32164')	711	



0.023		
('DAI62779', 'GRO30386')		709
0.023		
('ELE17451', 'FRO40251')		697
0.022		
('DAI85309', 'ELE99737')		659
0.021		
('DAI62779', 'ELE26917')		650
0.021		
('GRO21487', 'GRO73461')		631
0.02		
('DAI62779', 'SNA45677')		604
0.019		
('ELE17451', 'SNA80324')		597
0.019		
('DAI62779', 'GRO71621')		595
0.019		
('DAI62779', 'SNA55762')		593
0.019		
('DAI62779', 'DAI83733')		586
0.019		
('ELE17451', 'GRO73461')		580
0.019		
('GRO73461', 'SNA80324')		562
0.018		
('DAI62779', 'GRO59710')		561
0.018		
('DAI62779', 'FRO80039')		550
0.018		
('DAI75645', 'ELE17451')		547
0.018		
('DAI62779', 'SNA93860')		537
0.017		
('DAI55148', 'DAI62779')		526
0.017		
('DAI43223', 'GRO59710')		512
0.016		
('ELE17451', 'ELE32164')		511
0.016		
('DAI62779', 'SNA18336')		506
0.016		
('ELE32164', 'GRO73461')		486
0.016		
('DAI85309', 'ELE17451')		482
0.015		
('DAI62779', 'FRO78087')		482
0.015		
('DAI62779', 'GRO94758')		479
0.015		
('DAI62779', 'GRO21487')		471
0.015		
('GRO85051', 'SNA80324')		471
0.015		
('ELE17451', 'GRO30386')		468

```

0.015
('FRO85978', 'SNA95666') | 463 |
0.015
('DAI62779', 'FRO19221') | 462 |
0.015
('DAI62779', 'GRO46854') | 461 |
0.015
('DAI43223', 'DAI62779') | 459 |
0.015
('ELE92920', 'SNA18336') | 455 |
0.015
('DAI88079', 'FRO40251') | 446 |
0.014

```

The job was run on a Macbook Pro with a 2.66GHz Intel dual-core processor and 4GB of memory.

This job took 1 minute and 51 seconds, so only slightly longer than the previous job.

Based on the results from Question 3.4 and Question 3.5, the two jobs use the same number of combiners, mappers, and reducers to accomplish their respective tasks. These are 2 combiners, 1 mapper, and 1 reducer for both jobs.

---

## HW3.6 (Optional)

## Part A

*What is the Apriori algorithm? Describe an example use in your domain of expertise.*

### Solution:

The Apriori algorithm is an algorithm for associative rule learning over transactional data sets. In particular, it implements a multi-scan approach with different tuning parameters like minimum support and minimum confidence. These are used to improve efficiency and control performance. In particular, the Apriori algorithm will identify frequent terms from the given support, and recursively compute larger baskets and search for additional frequent items. The Apriori algorithm stops when no further baskets can be found that are frequent.

Within the domain of telecommunications, Apriori-like algorithms are used to describe the customer journey from signing-up to porting out. Identifying patterns that lead to churn is very important for Big Telecom, as the market continues to grow more volatile and competitive. Looking at information like clicks, calls to customer support, or billing events and building associative rules is important in finding business insights that improve customer retention.

## Part B

*Define confidence and lift*

### Solution:

Given elements A and B, the confidence of  $A \Rightarrow B$  is equal to the support of A and B divided by the support of A. In words, confidence represents the likelihood of observing a basket containing B given a basket containing A.

Lift is defined as a measure of performance of a model at predicting enhanced responses against a random choice targetting. In other words, Lift describes how well a model performs in identifying association rules that are more confident than selecting association rules at random. Given elements A and B, the lift of  $A \Rightarrow B$  is defined as the confidence of A and B given the average confidence across the entire basket list.

## HW3.7. Shopping Cart Analysis

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

Suppose we want to recommend new products to the customer based on the products they have already browsed on the online website. Write a program using the A-priori algorithm to find products which are frequently browsed together. Fix the support to  $s = 100$  (i.e. product sets need to occur together at least 100 times to be considered frequent) and find itemsets of size 2 and 3.

Then extract association rules from these frequent items.

A rule is of the form:

$(\text{item1}, \text{item5}) \Rightarrow \text{item2}.$

List the top 10 discovered rules in decreasing order of confidence in the following format

$(\text{item1}, \text{item5}) \Rightarrow \text{item2}, \text{supportCount}, \text{support}, \text{confidence}$

## HW3.7 Mapper

We use the pairs implementation here as it requires minimal code changes, albeit being less efficient. To this end, the mapper now outputs both 2 and 3-tuples which are treated the same by the reducer. The reducer changes its printing functionality to this end.

```
In [91]: %%writefile mapperQ37.py
#!/usr/bin/env python

from __future__ import division
import sys
from collections import defaultdict
from itertools import combinations

def readInput(file, separator=None):
    for line in file:
        yield line.split(separator)

if __name__ == "__main__":

    # Read input
    data = readInput(sys.stdin)
    for line in data:

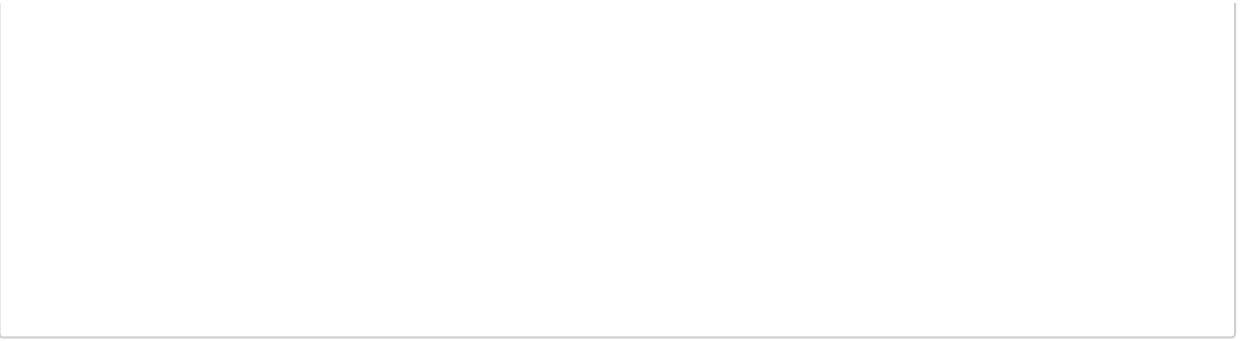
        # Get unique keys
        doublePairs = list(combinations(line, 2))
        triplePairs = list(combinations(line, 3))

        # Sort doubles
        sortedDoubles = []
        for pair in doublePairs:
            pList = list(pair)
            pList.sort()
            sortedDoubles.append(tuple(pList))

        # Emit
        for pair in sortedDoubles:
            print '%s%s%s' % (pair, '\t', 1)

        # Sort triples
        sortedTriples = []
        for pair in triplePairs:
            pList = list(pair)
            pList.sort()
            sortedTriples.append(tuple(pList))

        # Emit
        for pair in sortedTriples:
            print '%s%s%s' % (pair, '\t', 1)
```



Overwriting mapperQ37.py

## HW3.7 Reducer

We modify the reducer so it outputs the top 50 of both the length-2 and length-3 tuples. This is mainly a cosmetic change in the printing functionality.

In [97]:

```

%%writefile reducerQ37.py
#!/usr/bin/env python

from __future__ import division
import sys
from collections import defaultdict
import ast

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

if __name__ == "__main__":

    # Final store
    storingDict = defaultdict(int)
    support = 100

    # Read data
    data = readInput(sys.stdin)
    for line in data:

        # Parse value
        token = line[0]
        termCount = int(line[1])

        # Store results
        storingDict[token] += termCount

    # Filter
    filterDict = defaultdict(int)
    for k, v in storingDict.iteritems():
        if v >= support:
            filterDict[k] += v

    # Find most frequent doubles
    mfgDoubles = [(k, v, support) for k, v in filterDict.iteritems()
                  if k.count(',') == 1]

    mfgDoubles = sorted(mfgDoubles,
                        key=lambda x: x[1],
                        reverse = True)

    # Get results

```



```

print '\n' + '=====  

+ '\n'  

template = "{0:30}|{1:20}|{2:20}"  

print template.format("PAIR", "SUPPORT COUNT", "SUPPORT")  

# Print terms  

for termPair in mfqDoubles[:50]:  

    print template.format(*termPair)  

# Find most frequent triples  

mfqTriples = [(k, v, support) for k, v in filterDict.iteritems()  

                                                        if k.count(',') == 2]  

mfqTriples = sorted(mfqTriples,  

                    key  

                    = lambda x: x[1],  

                    rev  

                    erse = True)  

# Get results  

print '\n' + '=====  

+ '\n'  

template = "{0:50}|{1:20}|{2:20}"  

print template.format("PAIR", "SUPPORT COUNT", "SUPPORT")  

# Print terms  

for termPair in mfqTriples[:50]:  

    print template.format(*termPair)

```

Overwriting reducerQ37.py

## HW3.7 Hadoop MapReduce Submit Job

We now submit the job using the combiner and wrapper defined in Question 3.4.

```
In [98]: !bash wrapperQ34.sh ProductPurchaseData.txt mapperQ37.py reducerQ37.py combinerQ34.py
```



```
16/01/30 20:39:32 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/01/30 20:39:33 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/john/notebook
16/01/30 20:39:35 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/01/30 20:39:39 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/01/30 20:39:42 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
16/01/30 20:39:42 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
packageJobJar: [/Users/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/mapperQ37.py, /Users/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/reducerQ37.py, /Users/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/combinerQ34.py] [] /var/folders/0w/8hzv7rsj3qgdynsjlqy3gjsc0000gn/T/streamjob5966942647828294044.jar tmpDir=null
16/01/30 20:39:44 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
16/01/30 20:39:44 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/01/30 20:39:44 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
16/01/30 20:39:45 INFO mapred.FileInputFormat: Total input paths to process : 1
16/01/30 20:39:45 INFO mapreduce.JobSubmitter: number of splits:1
16/01/30 20:39:45 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local583153488_0001
16/01/30 20:39:45 INFO mapred.LocalDistributedCacheManager: Localized file:/Users/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/mapperQ37.py as file:/usr/local/Cellar/hadoop/hdfs/tmp/mapred/local/1454204385598/mapperQ37.py
16/01/30 20:39:46 INFO mapred.LocalDistributedCacheManager: Localized file:/Users/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/reducerQ37.py as file:/usr/local/Cellar/hadoop/hdfs/tmp/mapred/local/1454204385599/reducerQ37.py
16/01/30 20:39:46 INFO mapred.LocalDistributedCacheManager: Localized file:/Users/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/combinerQ34.py as file:/usr/local/Cellar/hadoop/hdfs/tmp/mapred/local/1454204385600/combinerQ34.py
16/01/30 20:39:46 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/01/30 20:39:46 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/01/30 20:39:46 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
16/01/30 20:39:46 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
```

```
16/01/30 20:39:46 INFO mapred.LocalJobRunner: Waiting for map task
s
16/01/30 20:39:46 INFO mapred.LocalJobRunner: Starting task: attempt_local583153488_0001_m_000000_0
16/01/30 20:39:46 INFO mapreduce.Job: Running job: job_local583153488_0001
16/01/30 20:39:46 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/30 20:39:46 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
16/01/30 20:39:46 INFO mapred.Task: Using ResourceCalculatorProcessTree : null
16/01/30 20:39:46 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/john/notebook/input/ProductPurchaseData.txt:0+3458517
16/01/30 20:39:46 INFO mapred.MapTask: numReduceTasks: 1
16/01/30 20:39:46 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/01/30 20:39:46 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/01/30 20:39:46 INFO mapred.MapTask: soft limit at 83886080
16/01/30 20:39:46 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/01/30 20:39:46 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/01/30 20:39:46 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
16/01/30 20:39:46 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/./mapperQ37.py]
16/01/30 20:39:46 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.output.dir
16/01/30 20:39:46 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
16/01/30 20:39:46 INFO Configuration.deprecation: mapred.task.ismap is deprecated. Instead, use mapreduce.task.ismap
16/01/30 20:39:46 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
16/01/30 20:39:46 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
16/01/30 20:39:46 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir
16/01/30 20:39:46 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
16/01/30 20:39:46 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
16/01/30 20:39:46 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.length
16/01/30 20:39:46 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
16/01/30 20:39:46 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
16/01/30 20:39:46 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
16/01/30 20:39:46 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
```

```
c/s] out:NA [rec/s]
16/01/30 20:39:46 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/30 20:39:46 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/30 20:39:46 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/30 20:39:46 INFO streaming.PipeMapRed: Records R/W=1216/1
16/01/30 20:39:47 INFO mapreduce.Job: Job job_local583153488_0001
running in uber mode : false
16/01/30 20:39:47 INFO mapreduce.Job: map 0% reduce 0%
16/01/30 20:39:52 INFO mapred.LocalJobRunner: Records R/W=1216/1 >
map
16/01/30 20:39:53 INFO mapreduce.Job: map 8% reduce 0%
16/01/30 20:39:54 INFO mapred.MapTask: Spilling map output
16/01/30 20:39:54 INFO mapred.MapTask: bufstart = 0; bufend = 5848
1682; bufvoid = 104857600
16/01/30 20:39:54 INFO mapred.MapTask: kvstart = 26214396(10485758
4); kvend = 19863304(79453216); length = 6351093/6553600
16/01/30 20:39:54 INFO mapred.MapTask: (EQUATOR) 64934450 kvi 1623
3608(64934432)
16/01/30 20:39:55 INFO mapred.LocalJobRunner: Records R/W=1216/1 >
map
16/01/30 20:39:56 INFO mapreduce.Job: map 10% reduce 0%
16/01/30 20:39:56 INFO streaming.PipeMapRed: Records R/W=5743/1928
237
16/01/30 20:39:58 INFO mapred.LocalJobRunner: Records R/W=5743/192
8237 > map
16/01/30 20:39:59 INFO mapreduce.Job: map 13% reduce 0%
16/01/30 20:39:59 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/./combinerQ34.p
y]
16/01/30 20:39:59 INFO Configuration.deprecation: mapred.skip.map.
auto.incr.proc.count is deprecated. Instead, use mapreduce.map.ski
p.proc-count.auto-incr
16/01/30 20:39:59 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/01/30 20:39:59 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/30 20:39:59 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/30 20:39:59 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/30 20:39:59 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/30 20:40:00 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/01/30 20:40:00 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/01/30 20:40:00 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:3
00000=300000/1 [rec/s] out:0=0/1 [rec/s]
16/01/30 20:40:01 INFO streaming.PipeMapRed: R/W/S=400000/0/0 in:4
00000=400000/1 [rec/s] out:0=0/1 [rec/s]
16/01/30 20:40:01 INFO mapred.LocalJobRunner: Records R/W=5743/192
```

```

8237 > map
16/01/30 20:40:01 INFO streaming.PipeMapRed: R/W/S=500000/0/0 in:5
00000=500000/1 [rec/s] out:0=0/1 [rec/s]
16/01/30 20:40:01 INFO streaming.PipeMapRed: R/W/S=600000/0/0 in:3
00000=600000/2 [rec/s] out:0=0/2 [rec/s]
16/01/30 20:40:02 INFO streaming.PipeMapRed: R/W/S=700000/0/0 in:3
50000=700000/2 [rec/s] out:0=0/2 [rec/s]
16/01/30 20:40:02 INFO streaming.PipeMapRed: R/W/S=800000/0/0 in:4
00000=800000/2 [rec/s] out:0=0/2 [rec/s]
16/01/30 20:40:02 INFO streaming.PipeMapRed: R/W/S=900000/0/0 in:3
00000=900000/3 [rec/s] out:0=0/3 [rec/s]
16/01/30 20:40:03 INFO streaming.PipeMapRed: R/W/S=1000000/0/0 in:
333333=1000000/3 [rec/s] out:0=0/3 [rec/s]
16/01/30 20:40:03 INFO streaming.PipeMapRed: R/W/S=1100000/0/0 in:
366666=1100000/3 [rec/s] out:0=0/3 [rec/s]
16/01/30 20:40:03 INFO streaming.PipeMapRed: R/W/S=1200000/0/0 in:
300000=1200000/4 [rec/s] out:0=0/4 [rec/s]
16/01/30 20:40:04 INFO streaming.PipeMapRed: R/W/S=1300000/0/0 in:
325000=1300000/4 [rec/s] out:0=0/4 [rec/s]
16/01/30 20:40:04 INFO mapred.LocalJobRunner: Records R/W=5743/192
8237 > map
16/01/30 20:40:04 INFO streaming.PipeMapRed: R/W/S=1400000/0/0 in:
350000=1400000/4 [rec/s] out:0=0/4 [rec/s]
16/01/30 20:40:04 INFO streaming.PipeMapRed: R/W/S=1500000/0/0 in:
300000=1500000/5 [rec/s] out:0=0/5 [rec/s]
16/01/30 20:40:05 INFO streaming.PipeMapRed: Records R/W=1587774/1
16/01/30 20:40:07 INFO mapred.LocalJobRunner: Records R/W=1587774/
1 > map
16/01/30 20:40:07 INFO streaming.PipeMapRed: MRErrorThread done
16/01/30 20:40:07 INFO streaming.PipeMapRed: mapRedFinished
16/01/30 20:40:07 INFO mapred.MapTask: Finished spill 0
16/01/30 20:40:07 INFO mapred.MapTask: (RESET) equator 64934450 kv
16233608(64934432) kvi 14659840(58639360)
16/01/30 20:40:07 INFO streaming.PipeMapRed: Records R/W=5743/1981
217
16/01/30 20:40:10 INFO mapred.LocalJobRunner: Records R/W=5743/198
1217 > map
16/01/30 20:40:11 INFO mapreduce.Job: map 15% reduce 0%
16/01/30 20:40:13 INFO mapred.MapTask: Spilling map output
16/01/30 20:40:13 INFO mapred.MapTask: bufstart = 64934450; bufend
= 18648555; bufvoid = 104857577
16/01/30 20:40:13 INFO mapred.MapTask: kvstart = 16233608(6493443
2); kvend = 9905012(39620048); length = 6328597/6553600
16/01/30 20:40:13 INFO mapred.MapTask: (EQUATOR) 25101307 kvi 6275
320(25101280)
16/01/30 20:40:13 INFO mapred.LocalJobRunner: Records R/W=5743/198
1217 > map
16/01/30 20:40:13 INFO mapreduce.Job: map 18% reduce 0%
16/01/30 20:40:16 INFO mapred.LocalJobRunner: Records R/W=5743/198
1217 > map
16/01/30 20:40:17 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/./combinerQ34.p
y]
16/01/30 20:40:17 INFO Configuration.deprecation: mapred.skip.redu

```

```
ce.auto.incr.proc.count is deprecated. Instead, use mapreduce.redu
ce.skip.proc-count.auto-incr
16/01/30 20:40:17 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/01/30 20:40:17 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/30 20:40:17 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/30 20:40:17 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/30 20:40:17 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/30 20:40:17 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/01/30 20:40:18 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/01/30 20:40:18 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:3
00000=300000/1 [rec/s] out:0=0/1 [rec/s]
16/01/30 20:40:18 INFO streaming.PipeMapRed: R/W/S=400000/0/0 in:4
00000=400000/1 [rec/s] out:0=0/1 [rec/s]
16/01/30 20:40:19 INFO streaming.PipeMapRed: R/W/S=500000/0/0 in:5
00000=500000/1 [rec/s] out:0=0/1 [rec/s]
16/01/30 20:40:19 INFO mapred.LocalJobRunner: Records R/W=5743/198
1217 > map
16/01/30 20:40:19 INFO streaming.PipeMapRed: R/W/S=600000/0/0 in:3
00000=600000/2 [rec/s] out:0=0/2 [rec/s]
16/01/30 20:40:19 INFO streaming.PipeMapRed: R/W/S=700000/0/0 in:3
50000=700000/2 [rec/s] out:0=0/2 [rec/s]
16/01/30 20:40:20 INFO streaming.PipeMapRed: R/W/S=800000/0/0 in:4
00000=800000/2 [rec/s] out:0=0/2 [rec/s]
16/01/30 20:40:20 INFO streaming.PipeMapRed: R/W/S=900000/0/0 in:3
00000=900000/3 [rec/s] out:0=0/3 [rec/s]
16/01/30 20:40:20 INFO streaming.PipeMapRed: R/W/S=1000000/0/0 in:
333333=1000000/3 [rec/s] out:0=0/3 [rec/s]
16/01/30 20:40:21 INFO streaming.PipeMapRed: R/W/S=1100000/0/0 in:
275000=1100000/4 [rec/s] out:0=0/4 [rec/s]
16/01/30 20:40:21 INFO streaming.PipeMapRed: R/W/S=1200000/0/0 in:
300000=1200000/4 [rec/s] out:0=0/4 [rec/s]
16/01/30 20:40:22 INFO streaming.PipeMapRed: R/W/S=1300000/0/0 in:
325000=1300000/4 [rec/s] out:0=0/4 [rec/s]
16/01/30 20:40:22 INFO mapred.LocalJobRunner: Records R/W=5743/198
1217 > map
16/01/30 20:40:22 INFO streaming.PipeMapRed: R/W/S=1400000/0/0 in:
280000=1400000/5 [rec/s] out:0=0/5 [rec/s]
16/01/30 20:40:23 INFO streaming.PipeMapRed: R/W/S=1500000/0/0 in:
300000=1500000/5 [rec/s] out:0=0/5 [rec/s]
16/01/30 20:40:23 INFO streaming.PipeMapRed: Records R/W=1582150/1
16/01/30 20:40:25 INFO mapred.LocalJobRunner: Records R/W=1582150/
1 > map
16/01/30 20:40:26 INFO streaming.PipeMapRed: MRErrorThread done
16/01/30 20:40:26 INFO streaming.PipeMapRed: mapRedFinished
16/01/30 20:40:26 INFO mapred.MapTask: Finished spill 1
16/01/30 20:40:26 INFO mapred.MapTask: (RESET) equator 25101307 kv
6275320(25101280) kvi 4710532(18842128)
```



```
16/01/30 20:40:26 INFO streaming.PipeMapRed: Records R/W=7778/3561
122
16/01/30 20:40:28 INFO mapred.LocalJobRunner: Records R/W=7778/356
1122 > map
16/01/30 20:40:28 INFO mapreduce.Job: map 20% reduce 0%
16/01/30 20:40:31 INFO mapred.LocalJobRunner: Records R/W=7778/356
1122 > map
16/01/30 20:40:31 INFO mapreduce.Job: map 23% reduce 0%
16/01/30 20:40:32 INFO mapred.MapTask: Spilling map output
16/01/30 20:40:32 INFO mapred.MapTask: bufstart = 25101307; bufend
= 83736112; bufvoid = 104857600
16/01/30 20:40:32 INFO mapred.MapTask: kvstart = 6275320(2510128
0); kvend = 26176912(104707648); length = 6312809/6553600
16/01/30 20:40:32 INFO mapred.MapTask: (EQUATOR) 90067136 kvi 2251
6780(90067120)
16/01/30 20:40:34 INFO mapred.LocalJobRunner: Records R/W=7778/356
1122 > map
16/01/30 20:40:37 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/./combinerQ34.p
y]
16/01/30 20:40:37 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/01/30 20:40:37 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/30 20:40:37 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/30 20:40:37 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/30 20:40:37 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/30 20:40:37 INFO mapred.LocalJobRunner: Records R/W=7778/356
1122 > map
16/01/30 20:40:37 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/01/30 20:40:38 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/01/30 20:40:38 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:3
00000=300000/1 [rec/s] out:0=0/1 [rec/s]
16/01/30 20:40:38 INFO streaming.PipeMapRed: R/W/S=400000/0/0 in:4
00000=400000/1 [rec/s] out:0=0/1 [rec/s]
16/01/30 20:40:39 INFO streaming.PipeMapRed: R/W/S=500000/0/0 in:5
00000=500000/1 [rec/s] out:0=0/1 [rec/s]
16/01/30 20:40:39 INFO streaming.PipeMapRed: R/W/S=600000/0/0 in:3
00000=600000/2 [rec/s] out:0=0/2 [rec/s]
16/01/30 20:40:39 INFO streaming.PipeMapRed: R/W/S=700000/0/0 in:3
50000=700000/2 [rec/s] out:0=0/2 [rec/s]
16/01/30 20:40:40 INFO streaming.PipeMapRed: R/W/S=800000/0/0 in:4
00000=800000/2 [rec/s] out:0=0/2 [rec/s]
16/01/30 20:40:40 INFO mapred.LocalJobRunner: Records R/W=7778/356
1122 > map
16/01/30 20:40:41 INFO streaming.PipeMapRed: R/W/S=900000/0/0 in:3
00000=900000/3 [rec/s] out:0=0/3 [rec/s]
16/01/30 20:40:42 INFO streaming.PipeMapRed: R/W/S=1000000/0/0 in:
200000=1000000/5 [rec/s] out:0=0/5 [rec/s]
```

```
16/01/30 20:40:42 INFO streaming.PipeMapRed: R/W/S=1100000/0/0 in:
220000=1100000/5 [rec/s] out:0=0/5 [rec/s]
16/01/30 20:40:43 INFO mapred.LocalJobRunner: Records R/W=7778/356
1122 > map
16/01/30 20:40:43 INFO streaming.PipeMapRed: R/W/S=1200000/0/0 in:
200000=1200000/6 [rec/s] out:0=0/6 [rec/s]
16/01/30 20:40:43 INFO streaming.PipeMapRed: R/W/S=1300000/0/0 in:
216666=1300000/6 [rec/s] out:0=0/6 [rec/s]
16/01/30 20:40:44 INFO streaming.PipeMapRed: R/W/S=1400000/0/0 in:
200000=1400000/7 [rec/s] out:0=0/7 [rec/s]
16/01/30 20:40:45 INFO streaming.PipeMapRed: R/W/S=1500000/0/0 in:
214285=1500000/7 [rec/s] out:0=0/7 [rec/s]
16/01/30 20:40:45 INFO streaming.PipeMapRed: Records R/W=1578203/1
16/01/30 20:40:46 INFO mapred.LocalJobRunner: Records R/W=1578203/
1 > map
16/01/30 20:40:49 INFO mapred.LocalJobRunner: Records R/W=1578203/
1 > map
16/01/30 20:40:49 INFO streaming.PipeMapRed: MRErrorThread done
16/01/30 20:40:49 INFO streaming.PipeMapRed: mapRedFinished
16/01/30 20:40:49 INFO mapred.MapTask: Finished spill 2
16/01/30 20:40:49 INFO mapred.MapTask: (RESET) equator 90067136 kv
22516780(90067120) kvi 20941520(83766080)
16/01/30 20:40:49 INFO streaming.PipeMapRed: Records R/W=9736/5141
943
16/01/30 20:40:49 INFO streaming.PipeMapRed: R/W/S=10000/5165981/0
in:158=10000/63 [rec/s] out:81999=5165981/63 [rec/s]
16/01/30 20:40:52 INFO mapred.LocalJobRunner: Records R/W=9736/514
1943 > map
16/01/30 20:40:52 INFO mapreduce.Job: map 25% reduce 0%
16/01/30 20:40:55 INFO mapred.LocalJobRunner: Records R/W=9736/514
1943 > map
16/01/30 20:40:55 INFO mapreduce.Job: map 28% reduce 0%
16/01/30 20:40:56 INFO mapred.MapTask: Spilling map output
16/01/30 20:40:56 INFO mapred.MapTask: bufstart = 90067136; bufend
= 43824763; bufvoid = 104857600
16/01/30 20:40:56 INFO mapred.MapTask: kvstart = 22516780(9006712
0); kvend = 16199068(64796272); length = 6317713/6553600
16/01/30 20:40:56 INFO mapred.MapTask: (EQUATOR) 50155771 kvi 1253
8936(50155744)
16/01/30 20:40:58 INFO mapred.LocalJobRunner: Records R/W=9736/514
1943 > map
16/01/30 20:40:58 INFO mapreduce.Job: map 30% reduce 0%
16/01/30 20:41:00 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/./combinerQ34.p
y]
16/01/30 20:41:00 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/01/30 20:41:00 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/30 20:41:00 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/30 20:41:00 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/30 20:41:00 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA
```

```
[rec/s] out:NA [rec/s]
16/01/30 20:41:01 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/01/30 20:41:01 INFO mapred.LocalJobRunner: Records R/W=9736/514
1943 > map
16/01/30 20:41:01 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/01/30 20:41:02 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:3
00000=300000/1 [rec/s] out:0=0/1 [rec/s]
16/01/30 20:41:02 INFO streaming.PipeMapRed: R/W/S=400000/0/0 in:2
00000=400000/2 [rec/s] out:0=0/2 [rec/s]
16/01/30 20:41:03 INFO streaming.PipeMapRed: R/W/S=500000/0/0 in:2
50000=500000/2 [rec/s] out:0=0/2 [rec/s]
16/01/30 20:41:03 INFO streaming.PipeMapRed: R/W/S=600000/0/0 in:2
00000=600000/3 [rec/s] out:0=0/3 [rec/s]
16/01/30 20:41:04 INFO streaming.PipeMapRed: R/W/S=700000/0/0 in:2
33333=700000/3 [rec/s] out:0=0/3 [rec/s]
16/01/30 20:41:04 INFO mapred.LocalJobRunner: Records R/W=9736/514
1943 > map
16/01/30 20:41:04 INFO streaming.PipeMapRed: R/W/S=800000/0/0 in:2
66666=800000/3 [rec/s] out:0=0/3 [rec/s]
16/01/30 20:41:05 INFO streaming.PipeMapRed: R/W/S=900000/0/0 in:2
25000=900000/4 [rec/s] out:0=0/4 [rec/s]
16/01/30 20:41:05 INFO streaming.PipeMapRed: R/W/S=1000000/0/0 in:
250000=1000000/4 [rec/s] out:0=0/4 [rec/s]
16/01/30 20:41:05 INFO streaming.PipeMapRed: R/W/S=1100000/0/0 in:
220000=1100000/5 [rec/s] out:0=0/5 [rec/s]
16/01/30 20:41:06 INFO streaming.PipeMapRed: R/W/S=1200000/0/0 in:
240000=1200000/5 [rec/s] out:0=0/5 [rec/s]
16/01/30 20:41:06 INFO streaming.PipeMapRed: R/W/S=1300000/0/0 in:
260000=1300000/5 [rec/s] out:0=0/5 [rec/s]
16/01/30 20:41:07 INFO streaming.PipeMapRed: R/W/S=1400000/0/0 in:
233333=1400000/6 [rec/s] out:0=0/6 [rec/s]
16/01/30 20:41:07 INFO mapred.LocalJobRunner: Records R/W=9736/514
1943 > map
16/01/30 20:41:07 INFO streaming.PipeMapRed: R/W/S=1500000/0/0 in:
250000=1500000/6 [rec/s] out:0=0/6 [rec/s]
16/01/30 20:41:07 INFO streaming.PipeMapRed: Records R/W=1579429/1
16/01/30 20:41:10 INFO mapred.LocalJobRunner: Records R/W=1579429/
1 > map
16/01/30 20:41:10 INFO streaming.PipeMapRed: MRErrorThread done
16/01/30 20:41:10 INFO streaming.PipeMapRed: mapRedFinished
16/01/30 20:41:10 INFO mapred.MapTask: Finished spill 3
16/01/30 20:41:10 INFO mapred.MapTask: (RESET) equator 50155771 kv
12538936(50155744) kvi 10963592(43854368)
16/01/30 20:41:10 INFO streaming.PipeMapRed: Records R/W=13103/672
1393
16/01/30 20:41:13 INFO mapred.LocalJobRunner: Records R/W=13103/67
21393 > map
16/01/30 20:41:16 INFO mapred.LocalJobRunner: Records R/W=13103/67
21393 > map
16/01/30 20:41:16 INFO mapred.MapTask: Spilling map output
16/01/30 20:41:16 INFO mapred.MapTask: bufstart = 50155771; bufend
= 3874239; bufvoid = 104857588
```

```
16/01/30 20:41:16 INFO mapred.MapTask: kvstart = 12538936(5015574
4); kvend = 6211436(24845744); length = 6327501/6553600
16/01/30 20:41:16 INFO mapred.MapTask: (EQUATOR) 10205247 kvi 2551
304(10205216)
16/01/30 20:41:16 INFO mapreduce.Job: map 35% reduce 0%
16/01/30 20:41:19 INFO mapred.LocalJobRunner: Records R/W=13103/67
21393 > map
16/01/30 20:41:19 INFO mapreduce.Job: map 40% reduce 0%
16/01/30 20:41:20 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/./combinerQ34.p
y]
16/01/30 20:41:20 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/01/30 20:41:20 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/30 20:41:20 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/30 20:41:20 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/30 20:41:20 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/30 20:41:20 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/01/30 20:41:20 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/01/30 20:41:21 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:3
00000=300000/1 [rec/s] out:0=0/1 [rec/s]
16/01/30 20:41:21 INFO streaming.PipeMapRed: R/W/S=400000/0/0 in:4
00000=400000/1 [rec/s] out:0=0/1 [rec/s]
16/01/30 20:41:22 INFO streaming.PipeMapRed: R/W/S=500000/0/0 in:5
00000=500000/1 [rec/s] out:0=0/1 [rec/s]
16/01/30 20:41:22 INFO streaming.PipeMapRed: R/W/S=600000/0/0 in:3
00000=600000/2 [rec/s] out:0=0/2 [rec/s]
16/01/30 20:41:22 INFO mapred.LocalJobRunner: Records R/W=13103/67
21393 > map
16/01/30 20:41:22 INFO streaming.PipeMapRed: R/W/S=700000/0/0 in:3
50000=700000/2 [rec/s] out:0=0/2 [rec/s]
16/01/30 20:41:23 INFO streaming.PipeMapRed: R/W/S=800000/0/0 in:4
00000=800000/2 [rec/s] out:0=0/2 [rec/s]
16/01/30 20:41:23 INFO streaming.PipeMapRed: R/W/S=900000/0/0 in:3
00000=900000/3 [rec/s] out:0=0/3 [rec/s]
16/01/30 20:41:23 INFO streaming.PipeMapRed: R/W/S=1000000/0/0 in:
333333=1000000/3 [rec/s] out:0=0/3 [rec/s]
16/01/30 20:41:24 INFO streaming.PipeMapRed: R/W/S=1100000/0/0 in:
366666=1100000/3 [rec/s] out:0=0/3 [rec/s]
16/01/30 20:41:24 INFO streaming.PipeMapRed: R/W/S=1200000/0/0 in:
300000=1200000/4 [rec/s] out:0=0/4 [rec/s]
16/01/30 20:41:24 INFO streaming.PipeMapRed: R/W/S=1300000/0/0 in:
325000=1300000/4 [rec/s] out:0=0/4 [rec/s]
16/01/30 20:41:25 INFO streaming.PipeMapRed: R/W/S=1400000/0/0 in:
350000=1400000/4 [rec/s] out:0=0/4 [rec/s]
16/01/30 20:41:25 INFO mapred.LocalJobRunner: Records R/W=13103/67
21393 > map
16/01/30 20:41:25 INFO streaming.PipeMapRed: R/W/S=1500000/0/0 in:
```

```
300000=1500000/5 [rec/s] out:0=0/5 [rec/s]
16/01/30 20:41:25 INFO streaming.PipeMapRed: Records R/W=1581876/1
16/01/30 20:41:28 INFO mapred.LocalJobRunner: Records R/W=1581876/
1 > map
16/01/30 20:41:28 INFO streaming.PipeMapRed: MRErrorThread done
16/01/30 20:41:28 INFO streaming.PipeMapRed: mapRedFinished
16/01/30 20:41:28 INFO mapred.MapTask: Finished spill 4
16/01/30 20:41:28 INFO mapred.MapTask: (RESET) equator 10205247 kv
2551304(10205216) kvi 968564(3874256)
16/01/30 20:41:28 INFO streaming.PipeMapRed: Records R/W=18924/830
5118
16/01/30 20:41:31 INFO mapred.LocalJobRunner: Records R/W=18924/83
05118 > map
16/01/30 20:41:31 INFO mapreduce.Job: map 43% reduce 0%
16/01/30 20:41:34 INFO mapred.LocalJobRunner: Records R/W=18924/83
05118 > map
16/01/30 20:41:34 INFO mapreduce.Job: map 45% reduce 0%
16/01/30 20:41:34 INFO mapred.MapTask: Spilling map output
16/01/30 20:41:34 INFO mapred.MapTask: bufstart = 10205247; bufend
= 68611041; bufvoid = 104857600
16/01/30 20:41:34 INFO mapred.MapTask: kvstart = 2551304(1020521
6); kvend = 22395636(89582544); length = 6370069/6553600
16/01/30 20:41:34 INFO mapred.MapTask: (EQUATOR) 75063793 kvi 1876
5944(75063776)
16/01/30 20:41:37 INFO mapred.LocalJobRunner: Records R/W=18924/83
05118 > map
16/01/30 20:41:38 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/./combinerQ34.p
y]
16/01/30 20:41:38 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/01/30 20:41:38 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/30 20:41:38 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/30 20:41:38 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/30 20:41:38 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/30 20:41:38 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/01/30 20:41:39 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/01/30 20:41:39 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:3
00000=300000/1 [rec/s] out:0=0/1 [rec/s]
16/01/30 20:41:39 INFO streaming.PipeMapRed: R/W/S=400000/0/0 in:4
00000=400000/1 [rec/s] out:0=0/1 [rec/s]
16/01/30 20:41:40 INFO streaming.PipeMapRed: R/W/S=500000/0/0 in:5
00000=500000/1 [rec/s] out:0=0/1 [rec/s]
16/01/30 20:41:40 INFO mapred.LocalJobRunner: Records R/W=18924/83
05118 > map
16/01/30 20:41:40 INFO streaming.PipeMapRed: R/W/S=600000/0/0 in:3
00000=600000/2 [rec/s] out:0=0/2 [rec/s]
16/01/30 20:41:41 INFO streaming.PipeMapRed: R/W/S=700000/0/0 in:3
```

```
50000=700000/2 [rec/s] out:0=0/2 [rec/s]
16/01/30 20:41:41 INFO streaming.PipeMapRed: R/W/S=800000/0/0 in:4
00000=800000/2 [rec/s] out:0=0/2 [rec/s]
16/01/30 20:41:41 INFO streaming.PipeMapRed: R/W/S=900000/0/0 in:3
00000=900000/3 [rec/s] out:0=0/3 [rec/s]
16/01/30 20:41:42 INFO streaming.PipeMapRed: R/W/S=1000000/0/0 in:
333333=1000000/3 [rec/s] out:0=0/3 [rec/s]
16/01/30 20:41:42 INFO streaming.PipeMapRed: R/W/S=1100000/0/0 in:
366666=1100000/3 [rec/s] out:0=0/3 [rec/s]
16/01/30 20:41:42 INFO streaming.PipeMapRed: R/W/S=1200000/0/0 in:
300000=1200000/4 [rec/s] out:0=0/4 [rec/s]
16/01/30 20:41:43 INFO streaming.PipeMapRed: R/W/S=1300000/0/0 in:
325000=1300000/4 [rec/s] out:0=0/4 [rec/s]
16/01/30 20:41:43 INFO mapred.LocalJobRunner: Records R/W=18924/83
05118 > map
16/01/30 20:41:43 INFO streaming.PipeMapRed: R/W/S=1400000/0/0 in:
350000=1400000/4 [rec/s] out:0=0/4 [rec/s]
16/01/30 20:41:43 INFO streaming.PipeMapRed: R/W/S=1500000/0/0 in:
300000=1500000/5 [rec/s] out:0=0/5 [rec/s]
16/01/30 20:41:44 INFO streaming.PipeMapRed: Records R/W=1592518/1
16/01/30 20:41:46 INFO mapred.LocalJobRunner: Records R/W=1592518/
1 > map
16/01/30 20:41:46 INFO streaming.PipeMapRed: MRErrorThread done
16/01/30 20:41:46 INFO streaming.PipeMapRed: mapRedFinished
16/01/30 20:41:46 INFO mapred.MapTask: Finished spill 5
16/01/30 20:41:46 INFO mapred.MapTask: (RESET) equator 75063793 kv
18765944(75063776) kvi 17202988(68811952)
16/01/30 20:41:46 INFO streaming.PipeMapRed: Records R/W=20935/989
2690
16/01/30 20:41:49 INFO mapred.LocalJobRunner: Records R/W=20935/98
92690 > map
16/01/30 20:41:49 INFO mapreduce.Job: map 48% reduce 0%
16/01/30 20:41:51 INFO mapred.MapTask: Spilling map output
16/01/30 20:41:51 INFO mapred.MapTask: bufstart = 75063793; bufend
= 28812705; bufvoid = 104857567
16/01/30 20:41:51 INFO mapred.MapTask: kvstart = 18765944(7506377
6); kvend = 12446056(49784224); length = 6319889/6553600
16/01/30 20:41:51 INFO mapred.MapTask: (EQUATOR) 35265473 kvi 8816
364(35265456)
16/01/30 20:41:52 INFO mapred.LocalJobRunner: Records R/W=20935/98
92690 > map
16/01/30 20:41:52 INFO mapreduce.Job: map 51% reduce 0%
16/01/30 20:41:55 INFO mapred.LocalJobRunner: Records R/W=20935/98
92690 > map
16/01/30 20:41:55 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/./combinerQ34.p
y]
16/01/30 20:41:55 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/01/30 20:41:55 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/30 20:41:55 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/30 20:41:55 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA
```

```
[rec/s] out:NA [rec/s]
16/01/30 20:41:55 INFO mapreduce.Job: map 53% reduce 0%
16/01/30 20:41:55 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/30 20:41:56 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:NA
A [rec/s] out:NA [rec/s]
16/01/30 20:41:56 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:NA
A [rec/s] out:NA [rec/s]
16/01/30 20:41:57 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:3
00000=300000/1 [rec/s] out:0=0/1 [rec/s]
16/01/30 20:41:57 INFO streaming.PipeMapRed: R/W/S=400000/0/0 in:2
00000=400000/2 [rec/s] out:0=0/2 [rec/s]
16/01/30 20:41:58 INFO mapred.LocalJobRunner: Records R/W=20935/98
92690 > map
16/01/30 20:41:58 INFO streaming.PipeMapRed: R/W/S=500000/0/0 in:2
50000=500000/2 [rec/s] out:0=0/2 [rec/s]
16/01/30 20:41:59 INFO streaming.PipeMapRed: R/W/S=600000/0/0 in:2
00000=600000/3 [rec/s] out:0=0/3 [rec/s]
16/01/30 20:42:00 INFO streaming.PipeMapRed: R/W/S=700000/0/0 in:1
75000=700000/4 [rec/s] out:0=0/4 [rec/s]
16/01/30 20:42:00 INFO streaming.PipeMapRed: R/W/S=800000/0/0 in:1
60000=800000/5 [rec/s] out:0=0/5 [rec/s]
16/01/30 20:42:01 INFO mapred.LocalJobRunner: Records R/W=20935/98
92690 > map
16/01/30 20:42:01 INFO streaming.PipeMapRed: R/W/S=900000/0/0 in:1
50000=900000/6 [rec/s] out:0=0/6 [rec/s]
16/01/30 20:42:02 INFO streaming.PipeMapRed: R/W/S=1000000/0/0 in:
166666=1000000/6 [rec/s] out:0=0/6 [rec/s]
16/01/30 20:42:02 INFO streaming.PipeMapRed: R/W/S=1100000/0/0 in:
157142=1100000/7 [rec/s] out:0=0/7 [rec/s]
16/01/30 20:42:03 INFO streaming.PipeMapRed: R/W/S=1200000/0/0 in:
171428=1200000/7 [rec/s] out:0=0/7 [rec/s]
16/01/30 20:42:03 INFO streaming.PipeMapRed: R/W/S=1300000/0/0 in:
185714=1300000/7 [rec/s] out:0=0/7 [rec/s]
16/01/30 20:42:03 INFO streaming.PipeMapRed: R/W/S=1400000/0/0 in:
175000=1400000/8 [rec/s] out:0=0/8 [rec/s]
16/01/30 20:42:04 INFO streaming.PipeMapRed: R/W/S=1500000/0/0 in:
187500=1500000/8 [rec/s] out:0=0/8 [rec/s]
16/01/30 20:42:04 INFO mapred.LocalJobRunner: Records R/W=20935/98
92690 > map
16/01/30 20:42:04 INFO streaming.PipeMapRed: Records R/W=1579973/1
16/01/30 20:42:07 INFO streaming.PipeMapRed: MRErrorThread done
16/01/30 20:42:07 INFO streaming.PipeMapRed: mapRedFinished
16/01/30 20:42:07 INFO mapred.MapTask: Finished spill 6
16/01/30 20:42:07 INFO mapred.MapTask: (RESET) equator 35265473 kv
8816364(35265456) kvi 7254956(29019824)
16/01/30 20:42:07 INFO streaming.PipeMapRed: Records R/W=24268/114
72276
16/01/30 20:42:07 INFO mapred.LocalJobRunner: Records R/W=24268/11
472276 > map
16/01/30 20:42:10 INFO mapred.LocalJobRunner: Records R/W=24268/11
472276 > map
16/01/30 20:42:10 INFO mapreduce.Job: map 56% reduce 0%
16/01/30 20:42:11 INFO mapred.MapTask: Spilling map output
```

```
16/01/30 20:42:11 INFO mapred.MapTask: bufstart = 35265473; bufend
= 93773681; bufvoid = 104857600
16/01/30 20:42:11 INFO mapred.MapTask: kvstart = 8816364(3526545
6); kvend = 2471904(9887616); length = 6344461/6553600
16/01/30 20:42:11 INFO mapred.MapTask: (EQUATOR) 100226449 kvi 250
56608(100226432)
16/01/30 20:42:13 INFO mapred.LocalJobRunner: Records R/W=24268/11
472276 > map
16/01/30 20:42:13 INFO mapreduce.Job: map 61% reduce 0%
16/01/30 20:42:15 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/./combinerQ34.p
y]
16/01/30 20:42:15 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/01/30 20:42:15 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/30 20:42:15 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/30 20:42:15 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/30 20:42:15 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/30 20:42:16 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/01/30 20:42:16 INFO mapred.LocalJobRunner: Records R/W=24268/11
472276 > map
16/01/30 20:42:16 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/01/30 20:42:16 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:3
00000=300000/1 [rec/s] out:0=0/1 [rec/s]
16/01/30 20:42:17 INFO streaming.PipeMapRed: R/W/S=400000/0/0 in:4
00000=400000/1 [rec/s] out:0=0/1 [rec/s]
16/01/30 20:42:17 INFO streaming.PipeMapRed: R/W/S=500000/0/0 in:5
00000=500000/1 [rec/s] out:0=0/1 [rec/s]
16/01/30 20:42:17 INFO streaming.PipeMapRed: R/W/S=600000/0/0 in:3
00000=600000/2 [rec/s] out:0=0/2 [rec/s]
16/01/30 20:42:18 INFO streaming.PipeMapRed: R/W/S=700000/0/0 in:3
50000=700000/2 [rec/s] out:0=0/2 [rec/s]
16/01/30 20:42:18 INFO streaming.PipeMapRed: R/W/S=800000/0/0 in:4
00000=800000/2 [rec/s] out:0=0/2 [rec/s]
16/01/30 20:42:19 INFO streaming.PipeMapRed: R/W/S=900000/0/0 in:3
00000=900000/3 [rec/s] out:0=0/3 [rec/s]
16/01/30 20:42:19 INFO streaming.PipeMapRed: R/W/S=1000000/0/0 in:
333333=1000000/3 [rec/s] out:0=0/3 [rec/s]
16/01/30 20:42:19 INFO mapred.LocalJobRunner: Records R/W=24268/11
472276 > map
16/01/30 20:42:19 INFO streaming.PipeMapRed: R/W/S=1100000/0/0 in:
275000=1100000/4 [rec/s] out:0=0/4 [rec/s]
16/01/30 20:42:20 INFO streaming.PipeMapRed: R/W/S=1200000/0/0 in:
300000=1200000/4 [rec/s] out:0=0/4 [rec/s]
16/01/30 20:42:20 INFO streaming.PipeMapRed: R/W/S=1300000/0/0 in:
325000=1300000/4 [rec/s] out:0=0/4 [rec/s]
16/01/30 20:42:20 INFO streaming.PipeMapRed: R/W/S=1400000/0/0 in:
280000=1400000/5 [rec/s] out:0=0/5 [rec/s]
```



```
16/01/30 20:42:21 INFO streaming.PipeMapRed: R/W/S=1500000/0/0 in:
300000=1500000/5 [rec/s] out:0=0/5 [rec/s]
16/01/30 20:42:21 INFO streaming.PipeMapRed: Records R/W=1586116/1
16/01/30 20:42:22 INFO mapred.LocalJobRunner: Records R/W=1586116/
1 > map
16/01/30 20:42:25 INFO streaming.PipeMapRed: MRErrorThread done
16/01/30 20:42:25 INFO streaming.PipeMapRed: mapRedFinished
16/01/30 20:42:25 INFO mapred.MapTask: Finished spill 7
16/01/30 20:42:25 INFO mapred.MapTask: (RESET) equator 100226449 k
v 25056608(100226432) kvi 23468064(93872256)
16/01/30 20:42:25 INFO streaming.PipeMapRed: Records R/W=28078/130
65176
16/01/30 20:42:25 INFO mapred.LocalJobRunner: Records R/W=28078/13
065176 > map
16/01/30 20:42:28 INFO mapred.LocalJobRunner: Records R/W=28078/13
065176 > map
16/01/30 20:42:28 INFO mapreduce.Job: map 63% reduce 0%
16/01/30 20:42:31 INFO mapred.LocalJobRunner: Records R/W=28078/13
065176 > map
16/01/30 20:42:31 INFO mapreduce.Job: map 66% reduce 0%
16/01/30 20:42:32 INFO mapred.MapTask: Spilling map output
16/01/30 20:42:32 INFO mapred.MapTask: bufstart = 100226449; bufen
d = 53764836; bufvoid = 104857582
16/01/30 20:42:32 INFO mapred.MapTask: kvstart = 25056608(10022643
2); kvend = 18684088(74736352); length = 6372521/6553600
16/01/30 20:42:32 INFO mapred.MapTask: (EQUATOR) 60217604 kvi 1505
4396(60217584)
16/01/30 20:42:34 INFO mapred.LocalJobRunner: Records R/W=28078/13
065176 > map
16/01/30 20:42:34 INFO mapreduce.Job: map 67% reduce 0%
16/01/30 20:42:36 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/./combinerQ34.p
y]
16/01/30 20:42:36 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/01/30 20:42:36 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/30 20:42:36 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/30 20:42:36 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/30 20:42:36 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/30 20:42:36 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/01/30 20:42:37 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:N
A [rec/s] out:NA [rec/s]
16/01/30 20:42:37 INFO mapred.LocalJobRunner: Records R/W=28078/13
065176 > map
16/01/30 20:42:37 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:3
00000=300000/1 [rec/s] out:0=0/1 [rec/s]
16/01/30 20:42:37 INFO streaming.PipeMapRed: R/W/S=400000/0/0 in:4
00000=400000/1 [rec/s] out:0=0/1 [rec/s]
16/01/30 20:42:38 INFO streaming.PipeMapRed: R/W/S=500000/0/0 in:5
```

```

00000=500000/1 [rec/s] out:0=0/1 [rec/s]
16/01/30 20:42:38 INFO streaming.PipeMapRed: R/W/S=600000/0/0 in:3
00000=600000/2 [rec/s] out:0=0/2 [rec/s]
16/01/30 20:42:39 INFO streaming.PipeMapRed: R/W/S=700000/0/0 in:3
50000=700000/2 [rec/s] out:0=0/2 [rec/s]
16/01/30 20:42:39 INFO streaming.PipeMapRed: R/W/S=800000/0/0 in:4
00000=800000/2 [rec/s] out:0=0/2 [rec/s]
16/01/30 20:42:39 INFO streaming.PipeMapRed: R/W/S=900000/0/0 in:3
00000=900000/3 [rec/s] out:0=0/3 [rec/s]
16/01/30 20:42:40 INFO streaming.PipeMapRed: R/W/S=1000000/0/0 in:
333333=1000000/3 [rec/s] out:0=0/3 [rec/s]
16/01/30 20:42:40 INFO mapred.LocalJobRunner: Records R/W=28078/13
065176 > map
16/01/30 20:42:40 INFO streaming.PipeMapRed: R/W/S=1100000/0/0 in:
275000=1100000/4 [rec/s] out:0=0/4 [rec/s]
16/01/30 20:42:40 INFO streaming.PipeMapRed: R/W/S=1200000/0/0 in:
300000=1200000/4 [rec/s] out:0=0/4 [rec/s]
16/01/30 20:42:41 INFO streaming.PipeMapRed: R/W/S=1300000/0/0 in:
325000=1300000/4 [rec/s] out:0=0/4 [rec/s]
16/01/30 20:42:41 INFO streaming.PipeMapRed: R/W/S=1400000/0/0 in:
280000=1400000/5 [rec/s] out:0=0/5 [rec/s]
16/01/30 20:42:41 INFO streaming.PipeMapRed: R/W/S=1500000/0/0 in:
300000=1500000/5 [rec/s] out:0=0/5 [rec/s]
16/01/30 20:42:42 INFO streaming.PipeMapRed: Records R/W=1593131/1
16/01/30 20:42:43 INFO mapred.LocalJobRunner: Records R/W=1593131/
1 > map
16/01/30 20:42:44 INFO streaming.PipeMapRed: MRErrorThread done
16/01/30 20:42:44 INFO streaming.PipeMapRed: mapRedFinished
16/01/30 20:42:44 INFO mapred.MapTask: Finished spill 8
16/01/30 20:42:44 INFO mapred.MapTask: (RESET) equator 60217604 kv
15054396(60217584) kvi 13476216(53904864)
16/01/30 20:42:44 INFO streaming.PipeMapRed: Records R/W=31101/146
55716
16/01/30 20:42:45 INFO streaming.PipeMapRed: MRErrorThread done
16/01/30 20:42:45 INFO streaming.PipeMapRed: mapRedFinished
16/01/30 20:42:45 INFO mapred.LocalJobRunner: Records R/W=1593131/
1 > map
16/01/30 20:42:45 INFO mapred.MapTask: Starting flush of map output
16/01/30 20:42:45 INFO mapred.MapTask: Spilling map output
16/01/30 20:42:45 INFO mapred.MapTask: bufstart = 60217604; bufend
= 77432318; bufvoid = 104857600
16/01/30 20:42:45 INFO mapred.MapTask: kvstart = 15054396(6021758
4); kvend = 13184280(52737120); length = 1870117/6553600
16/01/30 20:42:46 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/./combinerQ34.p
y]
16/01/30 20:42:46 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/01/30 20:42:46 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/30 20:42:46 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/30 20:42:46 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA

```

```
[rec/s] out:NA [rec/s]
16/01/30 20:42:46 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/30 20:42:46 INFO mapred.LocalJobRunner: Records R/W=31101/14
655716 > sort
16/01/30 20:42:46 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:NA
A [rec/s] out:NA [rec/s]
16/01/30 20:42:46 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:NA
A [rec/s] out:NA [rec/s]
16/01/30 20:42:47 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:3
00000=300000/1 [rec/s] out:0=0/1 [rec/s]
16/01/30 20:42:47 INFO streaming.PipeMapRed: R/W/S=400000/0/0 in:4
00000=400000/1 [rec/s] out:0=0/1 [rec/s]
16/01/30 20:42:47 INFO streaming.PipeMapRed: Records R/W=467530/1
16/01/30 20:42:48 INFO streaming.PipeMapRed: MRErrorThread done
16/01/30 20:42:48 INFO streaming.PipeMapRed: mapRedFinished
16/01/30 20:42:48 INFO mapred.MapTask: Finished spill 9
16/01/30 20:42:48 INFO mapred.Merger: Merging 10 sorted segments
16/01/30 20:42:48 INFO mapred.Merger: Down to the last merge-pass,
with 10 segments left of total size: 477989196 bytes
16/01/30 20:42:48 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/./combinerQ34.p
Y]
16/01/30 20:42:48 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/01/30 20:42:48 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/30 20:42:48 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/30 20:42:48 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/30 20:42:49 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/30 20:42:49 INFO mapred.LocalJobRunner: Records R/W=467530/1
> sort >
16/01/30 20:42:49 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:NA
A [rec/s] out:NA [rec/s]
16/01/30 20:42:50 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:2
00000=200000/1 [rec/s] out:0=0/1 [rec/s]
16/01/30 20:42:50 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:1
50000=300000/2 [rec/s] out:0=0/2 [rec/s]
16/01/30 20:42:51 INFO streaming.PipeMapRed: R/W/S=400000/0/0 in:2
00000=400000/2 [rec/s] out:0=0/2 [rec/s]
16/01/30 20:42:51 INFO streaming.PipeMapRed: R/W/S=500000/0/0 in:1
66666=500000/3 [rec/s] out:0=0/3 [rec/s]
16/01/30 20:42:52 INFO streaming.PipeMapRed: R/W/S=600000/0/0 in:2
00000=600000/3 [rec/s] out:0=0/3 [rec/s]
16/01/30 20:42:52 INFO mapred.LocalJobRunner: Records R/W=467530/1
> sort >
16/01/30 20:42:52 INFO mapreduce.Job: map 68% reduce 0%
16/01/30 20:42:52 INFO streaming.PipeMapRed: R/W/S=700000/0/0 in:2
33333=700000/3 [rec/s] out:0=0/3 [rec/s]
16/01/30 20:42:54 INFO streaming.PipeMapRed: R/W/S=800000/0/0 in:1
60000=800000/5 [rec/s] out:0=0/5 [rec/s]
```

```
16/01/30 20:42:55 INFO streaming.PipeMapRed: R/W/S=900000/0/0 in:1
50000=900000/6 [rec/s] out:0=0/6 [rec/s]
16/01/30 20:42:55 INFO mapred.LocalJobRunner: Records R/W=467530/1
> sort >
16/01/30 20:42:55 INFO mapreduce.Job: map 69% reduce 0%
16/01/30 20:42:56 INFO streaming.PipeMapRed: R/W/S=1000000/0/0 in:
142857=1000000/7 [rec/s] out:0=0/7 [rec/s]
16/01/30 20:42:56 INFO streaming.PipeMapRed: R/W/S=1100000/0/0 in:
157142=1100000/7 [rec/s] out:0=0/7 [rec/s]
16/01/30 20:42:57 INFO streaming.PipeMapRed: R/W/S=1200000/0/0 in:
150000=1200000/8 [rec/s] out:0=0/8 [rec/s]
16/01/30 20:42:58 INFO streaming.PipeMapRed: R/W/S=1300000/0/0 in:
144444=1300000/9 [rec/s] out:0=0/9 [rec/s]
16/01/30 20:42:58 INFO mapred.LocalJobRunner: Records R/W=467530/1
> sort >
16/01/30 20:42:58 INFO streaming.PipeMapRed: R/W/S=1400000/0/0 in:
155555=1400000/9 [rec/s] out:0=0/9 [rec/s]
16/01/30 20:42:58 INFO mapreduce.Job: map 70% reduce 0%
16/01/30 20:42:59 INFO streaming.PipeMapRed: R/W/S=1500000/0/0 in:
150000=1500000/10 [rec/s] out:0=0/10 [rec/s]
16/01/30 20:43:00 INFO streaming.PipeMapRed: R/W/S=1600000/0/0 in:
145454=1600000/11 [rec/s] out:0=0/11 [rec/s]
16/01/30 20:43:00 INFO streaming.PipeMapRed: R/W/S=1700000/0/0 in:
154545=1700000/11 [rec/s] out:0=0/11 [rec/s]
16/01/30 20:43:01 INFO streaming.PipeMapRed: R/W/S=1800000/0/0 in:
150000=1800000/12 [rec/s] out:0=0/12 [rec/s]
16/01/30 20:43:01 INFO streaming.PipeMapRed: R/W/S=1900000/0/0 in:
158333=1900000/12 [rec/s] out:0=0/12 [rec/s]
16/01/30 20:43:01 INFO mapred.LocalJobRunner: Records R/W=467530/1
> sort >
16/01/30 20:43:01 INFO streaming.PipeMapRed: R/W/S=2000000/0/0 in:
166666=2000000/12 [rec/s] out:0=0/12 [rec/s]
16/01/30 20:43:01 INFO mapreduce.Job: map 72% reduce 0%
16/01/30 20:43:02 INFO streaming.PipeMapRed: R/W/S=2100000/0/0 in:
161538=2100000/13 [rec/s] out:0=0/13 [rec/s]
16/01/30 20:43:02 INFO streaming.PipeMapRed: R/W/S=2200000/0/0 in:
169230=2200000/13 [rec/s] out:0=0/13 [rec/s]
16/01/30 20:43:02 INFO streaming.PipeMapRed: R/W/S=2300000/0/0 in:
164285=2300000/14 [rec/s] out:0=0/14 [rec/s]
16/01/30 20:43:03 INFO streaming.PipeMapRed: R/W/S=2400000/0/0 in:
171428=2400000/14 [rec/s] out:0=0/14 [rec/s]
16/01/30 20:43:03 INFO streaming.PipeMapRed: R/W/S=2500000/0/0 in:
178571=2500000/14 [rec/s] out:0=0/14 [rec/s]
16/01/30 20:43:04 INFO streaming.PipeMapRed: R/W/S=2600000/0/0 in:
173333=2600000/15 [rec/s] out:0=0/15 [rec/s]
16/01/30 20:43:04 INFO mapred.LocalJobRunner: Records R/W=467530/1
> sort >
16/01/30 20:43:04 INFO mapreduce.Job: map 74% reduce 0%
16/01/30 20:43:05 INFO streaming.PipeMapRed: R/W/S=2700000/0/0 in:
168750=2700000/16 [rec/s] out:0=0/16 [rec/s]
16/01/30 20:43:05 INFO streaming.PipeMapRed: R/W/S=2800000/0/0 in:
175000=2800000/16 [rec/s] out:0=0/16 [rec/s]
16/01/30 20:43:05 INFO streaming.PipeMapRed: R/W/S=2900000/0/0 in:
170588=2900000/17 [rec/s] out:0=0/17 [rec/s]
```

```
16/01/30 20:43:06 INFO streaming.PipeMapRed: R/W/S=3000000/0/0 in:
176470=3000000/17 [rec/s] out:0=0/17 [rec/s]
16/01/30 20:43:07 INFO streaming.PipeMapRed: R/W/S=3100000/0/0 in:
172222=3100000/18 [rec/s] out:0=0/18 [rec/s]
16/01/30 20:43:07 INFO mapred.LocalJobRunner: Records R/W=467530/1
> sort >
16/01/30 20:43:07 INFO mapreduce.Job: map 75% reduce 0%
16/01/30 20:43:07 INFO streaming.PipeMapRed: R/W/S=3200000/0/0 in:
168421=3200000/19 [rec/s] out:0=0/19 [rec/s]
16/01/30 20:43:08 INFO streaming.PipeMapRed: R/W/S=3300000/0/0 in:
173684=3300000/19 [rec/s] out:0=0/19 [rec/s]
16/01/30 20:43:09 INFO streaming.PipeMapRed: R/W/S=3400000/0/0 in:
170000=3400000/20 [rec/s] out:0=0/20 [rec/s]
16/01/30 20:43:10 INFO streaming.PipeMapRed: R/W/S=3500000/0/0 in:
166666=3500000/21 [rec/s] out:0=0/21 [rec/s]
16/01/30 20:43:10 INFO mapred.LocalJobRunner: Records R/W=467530/1
> sort >
16/01/30 20:43:10 INFO streaming.PipeMapRed: R/W/S=3600000/0/0 in:
171428=3600000/21 [rec/s] out:0=0/21 [rec/s]
16/01/30 20:43:10 INFO mapreduce.Job: map 77% reduce 0%
16/01/30 20:43:11 INFO streaming.PipeMapRed: R/W/S=3700000/0/0 in:
168181=3700000/22 [rec/s] out:0=0/22 [rec/s]
16/01/30 20:43:11 INFO streaming.PipeMapRed: R/W/S=3800000/0/0 in:
172727=3800000/22 [rec/s] out:0=0/22 [rec/s]
16/01/30 20:43:12 INFO streaming.PipeMapRed: R/W/S=3900000/0/0 in:
169565=3900000/23 [rec/s] out:0=0/23 [rec/s]
16/01/30 20:43:12 INFO streaming.PipeMapRed: R/W/S=4000000/0/0 in:
173913=4000000/23 [rec/s] out:0=0/23 [rec/s]
16/01/30 20:43:12 INFO streaming.PipeMapRed: R/W/S=4100000/0/0 in:
178260=4100000/23 [rec/s] out:0=0/23 [rec/s]
16/01/30 20:43:13 INFO streaming.PipeMapRed: R/W/S=4200000/0/0 in:
175000=4200000/24 [rec/s] out:0=0/24 [rec/s]
16/01/30 20:43:13 INFO mapred.LocalJobRunner: Records R/W=467530/1
> sort >
16/01/30 20:43:13 INFO streaming.PipeMapRed: R/W/S=4300000/0/0 in:
179166=4300000/24 [rec/s] out:0=0/24 [rec/s]
16/01/30 20:43:13 INFO mapreduce.Job: map 78% reduce 0%
16/01/30 20:43:14 INFO streaming.PipeMapRed: R/W/S=4400000/0/0 in:
176000=4400000/25 [rec/s] out:0=0/25 [rec/s]
16/01/30 20:43:14 INFO streaming.PipeMapRed: R/W/S=4500000/0/0 in:
180000=4500000/25 [rec/s] out:0=0/25 [rec/s]
16/01/30 20:43:14 INFO streaming.PipeMapRed: R/W/S=4600000/0/0 in:
176923=4600000/26 [rec/s] out:0=0/26 [rec/s]
16/01/30 20:43:15 INFO streaming.PipeMapRed: R/W/S=4700000/0/0 in:
180769=4700000/26 [rec/s] out:0=0/26 [rec/s]
16/01/30 20:43:15 INFO streaming.PipeMapRed: R/W/S=4800000/0/0 in:
184615=4800000/26 [rec/s] out:0=0/26 [rec/s]
16/01/30 20:43:16 INFO streaming.PipeMapRed: R/W/S=4900000/0/0 in:
181481=4900000/27 [rec/s] out:0=0/27 [rec/s]
16/01/30 20:43:16 INFO streaming.PipeMapRed: R/W/S=5000000/0/0 in:
185185=5000000/27 [rec/s] out:0=0/27 [rec/s]
16/01/30 20:43:16 INFO mapred.LocalJobRunner: Records R/W=467530/1
> sort >
16/01/30 20:43:16 INFO mapreduce.Job: map 81% reduce 0%
```

```
16/01/30 20:43:16 INFO streaming.PipeMapRed: R/W/S=5100000/0/0 in:
182142=5100000/28 [rec/s] out:0=0/28 [rec/s]
16/01/30 20:43:17 INFO streaming.PipeMapRed: R/W/S=5200000/0/0 in:
185714=5200000/28 [rec/s] out:0=0/28 [rec/s]
16/01/30 20:43:18 INFO streaming.PipeMapRed: R/W/S=5300000/0/0 in:
182758=5300000/29 [rec/s] out:0=0/29 [rec/s]
16/01/30 20:43:18 INFO streaming.PipeMapRed: R/W/S=5400000/0/0 in:
186206=5400000/29 [rec/s] out:0=0/29 [rec/s]
16/01/30 20:43:19 INFO streaming.PipeMapRed: R/W/S=5500000/0/0 in:
183333=5500000/30 [rec/s] out:0=0/30 [rec/s]
16/01/30 20:43:19 INFO mapred.LocalJobRunner: Records R/W=467530/1
> sort >
16/01/30 20:43:19 INFO mapreduce.Job: map 82% reduce 0%
16/01/30 20:43:20 INFO streaming.PipeMapRed: R/W/S=5600000/0/0 in:
180645=5600000/31 [rec/s] out:0=0/31 [rec/s]
16/01/30 20:43:20 INFO streaming.PipeMapRed: R/W/S=5700000/0/0 in:
183870=5700000/31 [rec/s] out:0=0/31 [rec/s]
16/01/30 20:43:21 INFO streaming.PipeMapRed: R/W/S=5800000/0/0 in:
181250=5800000/32 [rec/s] out:0=0/32 [rec/s]
16/01/30 20:43:21 INFO streaming.PipeMapRed: R/W/S=5900000/0/0 in:
178787=5900000/33 [rec/s] out:0=0/33 [rec/s]
16/01/30 20:43:22 INFO streaming.PipeMapRed: R/W/S=6000000/0/0 in:
181818=6000000/33 [rec/s] out:0=0/33 [rec/s]
16/01/30 20:43:22 INFO mapred.LocalJobRunner: Records R/W=467530/1
> sort >
16/01/30 20:43:22 INFO mapreduce.Job: map 83% reduce 0%
16/01/30 20:43:23 INFO streaming.PipeMapRed: R/W/S=6100000/0/0 in:
179411=6100000/34 [rec/s] out:0=0/34 [rec/s]
16/01/30 20:43:24 INFO streaming.PipeMapRed: R/W/S=6200000/0/0 in:
177142=6200000/35 [rec/s] out:0=0/35 [rec/s]
16/01/30 20:43:25 INFO mapred.LocalJobRunner: Records R/W=467530/1
> sort >
16/01/30 20:43:25 INFO mapreduce.Job: map 84% reduce 0%
16/01/30 20:43:27 INFO streaming.PipeMapRed: R/W/S=6300000/0/0 in:
165789=6300000/38 [rec/s] out:0=0/38 [rec/s]
16/01/30 20:43:28 INFO streaming.PipeMapRed: R/W/S=6400000/0/0 in:
164102=6400000/39 [rec/s] out:0=0/39 [rec/s]
16/01/30 20:43:28 INFO mapred.LocalJobRunner: Records R/W=467530/1
> sort >
16/01/30 20:43:28 INFO streaming.PipeMapRed: R/W/S=6500000/0/0 in:
166666=6500000/39 [rec/s] out:0=0/39 [rec/s]
16/01/30 20:43:29 INFO streaming.PipeMapRed: R/W/S=6600000/0/0 in:
165000=6600000/40 [rec/s] out:0=0/40 [rec/s]
16/01/30 20:43:29 INFO streaming.PipeMapRed: R/W/S=6700000/0/0 in:
163414=6700000/41 [rec/s] out:0=0/41 [rec/s]
16/01/30 20:43:30 INFO streaming.PipeMapRed: R/W/S=6800000/0/0 in:
165853=6800000/41 [rec/s] out:0=0/41 [rec/s]
16/01/30 20:43:30 INFO streaming.PipeMapRed: R/W/S=6900000/0/0 in:
164285=6900000/42 [rec/s] out:0=0/42 [rec/s]
16/01/30 20:43:31 INFO streaming.PipeMapRed: R/W/S=7000000/0/0 in:
166666=7000000/42 [rec/s] out:0=0/42 [rec/s]
16/01/30 20:43:31 INFO mapred.LocalJobRunner: Records R/W=467530/1
> sort >
16/01/30 20:43:31 INFO mapreduce.Job: map 86% reduce 0%
```

```
16/01/30 20:43:32 INFO streaming.PipeMapRed: R/W/S=7100000/0/0 in:
165116=7100000/43 [rec/s] out:0=0/43 [rec/s]
16/01/30 20:43:32 INFO streaming.PipeMapRed: R/W/S=7200000/0/0 in:
167441=7200000/43 [rec/s] out:0=0/43 [rec/s]
16/01/30 20:43:33 INFO streaming.PipeMapRed: R/W/S=7300000/0/0 in:
165909=7300000/44 [rec/s] out:0=0/44 [rec/s]
16/01/30 20:43:33 INFO streaming.PipeMapRed: R/W/S=7400000/0/0 in:
168181=7400000/44 [rec/s] out:0=0/44 [rec/s]
16/01/30 20:43:34 INFO streaming.PipeMapRed: R/W/S=7500000/0/0 in:
166666=7500000/45 [rec/s] out:0=0/45 [rec/s]
16/01/30 20:43:34 INFO mapred.LocalJobRunner: Records R/W=467530/1
> sort >
16/01/30 20:43:34 INFO streaming.PipeMapRed: R/W/S=7600000/0/0 in:
168888=7600000/45 [rec/s] out:0=0/45 [rec/s]
16/01/30 20:43:34 INFO mapreduce.Job: map 88% reduce 0%
16/01/30 20:43:35 INFO streaming.PipeMapRed: R/W/S=7700000/0/0 in:
167391=7700000/46 [rec/s] out:0=0/46 [rec/s]
16/01/30 20:43:35 INFO streaming.PipeMapRed: R/W/S=7800000/0/0 in:
169565=7800000/46 [rec/s] out:0=0/46 [rec/s]
16/01/30 20:43:36 INFO streaming.PipeMapRed: R/W/S=7900000/0/0 in:
168085=7900000/47 [rec/s] out:0=0/47 [rec/s]
16/01/30 20:43:36 INFO streaming.PipeMapRed: R/W/S=8000000/0/0 in:
170212=8000000/47 [rec/s] out:0=0/47 [rec/s]
16/01/30 20:43:37 INFO streaming.PipeMapRed: R/W/S=8100000/0/0 in:
168750=8100000/48 [rec/s] out:0=0/48 [rec/s]
16/01/30 20:43:37 INFO mapred.LocalJobRunner: Records R/W=467530/1
> sort >
16/01/30 20:43:37 INFO streaming.PipeMapRed: R/W/S=8200000/0/0 in:
170833=8200000/48 [rec/s] out:0=0/48 [rec/s]
16/01/30 20:43:37 INFO mapreduce.Job: map 89% reduce 0%
16/01/30 20:43:38 INFO streaming.PipeMapRed: R/W/S=8300000/0/0 in:
169387=8300000/49 [rec/s] out:0=0/49 [rec/s]
16/01/30 20:43:38 INFO streaming.PipeMapRed: R/W/S=8400000/0/0 in:
171428=8400000/49 [rec/s] out:0=0/49 [rec/s]
16/01/30 20:43:39 INFO streaming.PipeMapRed: R/W/S=8500000/0/0 in:
170000=8500000/50 [rec/s] out:0=0/50 [rec/s]
16/01/30 20:43:39 INFO streaming.PipeMapRed: R/W/S=8600000/0/0 in:
172000=8600000/50 [rec/s] out:0=0/50 [rec/s]
16/01/30 20:43:40 INFO streaming.PipeMapRed: R/W/S=8700000/0/0 in:
170588=8700000/51 [rec/s] out:0=0/51 [rec/s]
16/01/30 20:43:40 INFO mapred.LocalJobRunner: Records R/W=467530/1
> sort >
16/01/30 20:43:40 INFO streaming.PipeMapRed: R/W/S=8800000/0/0 in:
172549=8800000/51 [rec/s] out:0=0/51 [rec/s]
16/01/30 20:43:40 INFO mapreduce.Job: map 91% reduce 0%
16/01/30 20:43:41 INFO streaming.PipeMapRed: R/W/S=8900000/0/0 in:
171153=8900000/52 [rec/s] out:0=0/52 [rec/s]
16/01/30 20:43:41 INFO streaming.PipeMapRed: R/W/S=9000000/0/0 in:
173076=9000000/52 [rec/s] out:0=0/52 [rec/s]
16/01/30 20:43:42 INFO streaming.PipeMapRed: R/W/S=9100000/0/0 in:
171698=9100000/53 [rec/s] out:0=0/53 [rec/s]
16/01/30 20:43:42 INFO streaming.PipeMapRed: R/W/S=9200000/0/0 in:
173584=9200000/53 [rec/s] out:0=0/53 [rec/s]
16/01/30 20:43:43 INFO streaming.PipeMapRed: R/W/S=9300000/0/0 in:
```

```
172222=9300000/54 [rec/s] out:0=0/54 [rec/s]
16/01/30 20:43:43 INFO mapred.LocalJobRunner: Records R/W=467530/1
> sort >
16/01/30 20:43:43 INFO streaming.PipeMapRed: R/W/S=9400000/0/0 in:
174074=9400000/54 [rec/s] out:0=0/54 [rec/s]
16/01/30 20:43:43 INFO mapreduce.Job: map 93% reduce 0%
16/01/30 20:43:44 INFO streaming.PipeMapRed: R/W/S=9500000/0/0 in:
172727=9500000/55 [rec/s] out:0=0/55 [rec/s]
16/01/30 20:43:44 INFO streaming.PipeMapRed: R/W/S=9600000/0/0 in:
174545=9600000/55 [rec/s] out:0=0/55 [rec/s]
16/01/30 20:43:45 INFO streaming.PipeMapRed: R/W/S=9700000/0/0 in:
173214=9700000/56 [rec/s] out:0=0/56 [rec/s]
16/01/30 20:43:45 INFO streaming.PipeMapRed: R/W/S=9800000/0/0 in:
171929=9800000/57 [rec/s] out:0=0/57 [rec/s]
16/01/30 20:43:46 INFO mapred.LocalJobRunner: Records R/W=467530/1
> sort >
16/01/30 20:43:46 INFO streaming.PipeMapRed: R/W/S=9900000/0/0 in:
173684=9900000/57 [rec/s] out:0=0/57 [rec/s]
16/01/30 20:43:46 INFO mapreduce.Job: map 94% reduce 0%
16/01/30 20:43:47 INFO streaming.PipeMapRed: R/W/S=10000000/0/0 i
n:172413=10000000/58 [rec/s] out:0=0/58 [rec/s]
16/01/30 20:43:47 INFO streaming.PipeMapRed: R/W/S=10100000/0/0 i
n:174137=10100000/58 [rec/s] out:0=0/58 [rec/s]
16/01/30 20:43:48 INFO streaming.PipeMapRed: R/W/S=10200000/0/0 i
n:172881=10200000/59 [rec/s] out:0=0/59 [rec/s]
16/01/30 20:43:48 INFO streaming.PipeMapRed: R/W/S=10300000/0/0 i
n:174576=10300000/59 [rec/s] out:0=0/59 [rec/s]
16/01/30 20:43:49 INFO streaming.PipeMapRed: R/W/S=10400000/0/0 i
n:173333=10400000/60 [rec/s] out:0=0/60 [rec/s]
16/01/30 20:43:49 INFO streaming.PipeMapRed: R/W/S=10500000/0/0 i
n:175000=10500000/60 [rec/s] out:0=0/60 [rec/s]
16/01/30 20:43:49 INFO mapred.LocalJobRunner: Records R/W=467530/1
> sort >
16/01/30 20:43:49 INFO mapreduce.Job: map 96% reduce 0%
16/01/30 20:43:49 INFO streaming.PipeMapRed: R/W/S=10600000/0/0 i
n:173770=10600000/61 [rec/s] out:0=0/61 [rec/s]
16/01/30 20:43:50 INFO streaming.PipeMapRed: R/W/S=10700000/0/0 i
n:175409=10700000/61 [rec/s] out:0=0/61 [rec/s]
16/01/30 20:43:50 INFO streaming.PipeMapRed: R/W/S=10800000/0/0 i
n:177049=10800000/61 [rec/s] out:0=0/61 [rec/s]
16/01/30 20:43:51 INFO streaming.PipeMapRed: R/W/S=10900000/0/0 i
n:175806=10900000/62 [rec/s] out:0=0/62 [rec/s]
16/01/30 20:43:51 INFO streaming.PipeMapRed: R/W/S=11000000/0/0 i
n:174603=11000000/63 [rec/s] out:0=0/63 [rec/s]
16/01/30 20:43:52 INFO streaming.PipeMapRed: R/W/S=11100000/0/0 i
n:176190=11100000/63 [rec/s] out:0=0/63 [rec/s]
16/01/30 20:43:52 INFO mapred.LocalJobRunner: Records R/W=467530/1
> sort >
16/01/30 20:43:52 INFO mapreduce.Job: map 97% reduce 0%
16/01/30 20:43:53 INFO streaming.PipeMapRed: R/W/S=11200000/0/0 i
n:175000=11200000/64 [rec/s] out:0=0/64 [rec/s]
16/01/30 20:43:54 INFO streaming.PipeMapRed: R/W/S=11300000/0/0 i
n:173846=11300000/65 [rec/s] out:0=0/65 [rec/s]
16/01/30 20:43:54 INFO streaming.PipeMapRed: R/W/S=11400000/0/0 i
```



```
n:175384=11400000/65 [rec/s] out:0=0/65 [rec/s]
16/01/30 20:43:55 INFO streaming.PipeMapRed: R/W/S=11500000/0/0 i
n:174242=11500000/66 [rec/s] out:0=0/66 [rec/s]
16/01/30 20:43:55 INFO mapred.LocalJobRunner: Records R/W=467530/1
> sort >
16/01/30 20:43:55 INFO mapreduce.Job: map 98% reduce 0%
16/01/30 20:43:55 INFO streaming.PipeMapRed: R/W/S=11600000/0/0 i
n:173134=11600000/67 [rec/s] out:0=0/67 [rec/s]
16/01/30 20:43:56 INFO streaming.PipeMapRed: R/W/S=11700000/0/0 i
n:174626=11700000/67 [rec/s] out:0=0/67 [rec/s]
16/01/30 20:43:56 INFO streaming.PipeMapRed: R/W/S=11800000/0/0 i
n:173529=11800000/68 [rec/s] out:0=0/68 [rec/s]
16/01/30 20:43:57 INFO streaming.PipeMapRed: R/W/S=11900000/0/0 i
n:175000=11900000/68 [rec/s] out:0=0/68 [rec/s]
16/01/30 20:43:57 INFO streaming.PipeMapRed: R/W/S=12000000/0/0 i
n:176470=12000000/68 [rec/s] out:0=0/68 [rec/s]
16/01/30 20:43:58 INFO streaming.PipeMapRed: Records R/W=12070340/
1
16/01/30 20:43:58 INFO mapred.LocalJobRunner: Records R/W=1207034
0/1 > sort >
16/01/30 20:43:58 INFO mapreduce.Job: map 100% reduce 0%
16/01/30 20:44:01 INFO mapred.LocalJobRunner: Records R/W=1207034
0/1 > sort >
16/01/30 20:44:04 INFO mapred.LocalJobRunner: Records R/W=1207034
0/1 > sort >
16/01/30 20:44:07 INFO mapred.LocalJobRunner: Records R/W=1207034
0/1 > sort >
16/01/30 20:44:08 INFO streaming.PipeMapRed: Records R/W=12070340/
5951298
16/01/30 20:44:10 INFO mapred.LocalJobRunner: Records R/W=1207034
0/5951298 > sort >
16/01/30 20:44:13 INFO mapred.LocalJobRunner: Records R/W=1207034
0/5951298 > sort >
16/01/30 20:44:16 INFO mapred.LocalJobRunner: Records R/W=1207034
0/5951298 > sort >
16/01/30 20:44:19 INFO mapred.LocalJobRunner: Records R/W=1207034
0/5951298 > sort >
16/01/30 20:44:19 INFO streaming.PipeMapRed: Records R/W=12070340/
10092674
16/01/30 20:44:19 INFO streaming.PipeMapRed: MRErrorThread done
16/01/30 20:44:19 INFO streaming.PipeMapRed: mapRedFinished
16/01/30 20:44:19 INFO mapred.Task: Task:attempt_local583153488_00
01_m_000000_0 is done. And is in the process of committing
16/01/30 20:44:19 INFO mapred.LocalJobRunner: Records R/W=1207034
0/10092674 > sort
16/01/30 20:44:19 INFO mapred.Task: Task 'attempt_local583153488_0
001_m_000000_0' done.
16/01/30 20:44:19 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local583153488_0001_m_000000_0
16/01/30 20:44:19 INFO mapred.LocalJobRunner: map task executor co
mplete.
16/01/30 20:44:19 INFO mapred.LocalJobRunner: Waiting for reduce t
asks
16/01/30 20:44:19 INFO mapred.LocalJobRunner: Starting task: attem
```

```
pt_local583153488_0001_r_000000_0
16/01/30 20:44:19 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
16/01/30 20:44:19 INFO util.ProcfsBasedProcessTree: ProcfsBasedPro
cessTree currently is supported only on Linux.
16/01/30 20:44:19 INFO mapred.Task: Using ResourceCalculatorProce
ssTree : null
16/01/30 20:44:20 INFO mapred.ReduceTask: Using ShuffleConsumerPlu
gin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@6f95197a
16/01/30 20:44:20 INFO reduce.MergeManagerImpl: MergeManager: mem
oryLimit=334338464, maxSingleShuffleLimit=83584616, mergeThreshold
=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
16/01/30 20:44:20 INFO reduce.EventFetcher: attempt_local583153488
_0001_r_000000_0 Thread started: EventFetcher for fetching Map Com
pletion Events
16/01/30 20:44:20 INFO reduce.MergeManagerImpl: attempt_local58315
3488_0001_m_000000_0: Shuffling to disk since 403369367 is greater
than maxSingleShuffleLimit (83584616)
16/01/30 20:44:20 INFO reduce.LocalFetcher: localfetcher#1 about t
o shuffle output of map attempt_local583153488_0001_m_000000_0 dec
omp: 403369367 len: 403369371 to DISK
16/01/30 20:44:23 INFO reduce.OnDiskMapOutput: Read 403369371 byte
s from map-output for attempt_local583153488_0001_m_000000_0
16/01/30 20:44:23 INFO reduce.EventFetcher: EventFetcher is interr
upted.. Returning
16/01/30 20:44:23 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/30 20:44:23 INFO reduce.MergeManagerImpl: finalMerge called
with 0 in-memory map-outputs and 1 on-disk map-outputs
16/01/30 20:44:23 INFO reduce.MergeManagerImpl: Merging 1 files, 4
03369371 bytes from disk
16/01/30 20:44:23 INFO reduce.MergeManagerImpl: Merging 0 segment
s, 0 bytes from memory into reduce
16/01/30 20:44:23 INFO mapred.Merger: Merging 1 sorted segments
16/01/30 20:44:23 INFO mapred.Merger: Down to the last merge-pass,
with 1 segments left of total size: 403369328 bytes
16/01/30 20:44:23 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/30 20:44:23 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/john/Dropbox/MIDS/W261/Week3/Homework3/notebook/./reducerQ37.p
y]
16/01/30 20:44:23 INFO Configuration.deprecation: mapred.job.track
er is deprecated. Instead, use mapreduce.jobtracker.address
16/01/30 20:44:23 INFO Configuration.deprecation: mapred.map.tasks
is deprecated. Instead, use mapreduce.job.maps
16/01/30 20:44:23 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/01/30 20:44:23 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/30 20:44:23 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/30 20:44:23 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/30 20:44:23 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/30 20:44:24 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:N
```

```
A [rec/s] out:NA [rec/s]
16/01/30 20:44:24 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:2
00000=200000/1 [rec/s] out:0=0/1 [rec/s]
16/01/30 20:44:25 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:3
00000=300000/1 [rec/s] out:0=0/1 [rec/s]
16/01/30 20:44:25 INFO streaming.PipeMapRed: R/W/S=400000/0/0 in:2
00000=400000/2 [rec/s] out:0=0/2 [rec/s]
16/01/30 20:44:26 INFO mapred.LocalJobRunner: reduce > reduce
16/01/30 20:44:26 INFO streaming.PipeMapRed: R/W/S=500000/0/0 in:2
50000=500000/2 [rec/s] out:0=0/2 [rec/s]
16/01/30 20:44:26 INFO streaming.PipeMapRed: R/W/S=600000/0/0 in:3
00000=600000/2 [rec/s] out:0=0/2 [rec/s]
16/01/30 20:44:27 INFO mapreduce.Job: map 100% reduce 68%
16/01/30 20:44:27 INFO streaming.PipeMapRed: R/W/S=700000/0/0 in:2
33333=700000/3 [rec/s] out:0=0/3 [rec/s]
16/01/30 20:44:27 INFO streaming.PipeMapRed: R/W/S=800000/0/0 in:2
66666=800000/3 [rec/s] out:0=0/3 [rec/s]
16/01/30 20:44:27 INFO streaming.PipeMapRed: R/W/S=900000/0/0 in:2
25000=900000/4 [rec/s] out:0=0/4 [rec/s]
16/01/30 20:44:28 INFO streaming.PipeMapRed: R/W/S=1000000/0/0 in:
250000=1000000/4 [rec/s] out:0=0/4 [rec/s]
16/01/30 20:44:28 INFO streaming.PipeMapRed: R/W/S=1100000/0/0 in:
220000=1100000/5 [rec/s] out:0=0/5 [rec/s]
16/01/30 20:44:29 INFO mapred.LocalJobRunner: reduce > reduce
16/01/30 20:44:29 INFO streaming.PipeMapRed: R/W/S=1200000/0/0 in:
240000=1200000/5 [rec/s] out:0=0/5 [rec/s]
16/01/30 20:44:29 INFO streaming.PipeMapRed: R/W/S=1300000/0/0 in:
260000=1300000/5 [rec/s] out:0=0/5 [rec/s]
16/01/30 20:44:29 INFO streaming.PipeMapRed: R/W/S=1400000/0/0 in:
233333=1400000/6 [rec/s] out:0=0/6 [rec/s]
16/01/30 20:44:30 INFO mapreduce.Job: map 100% reduce 71%
16/01/30 20:44:30 INFO streaming.PipeMapRed: R/W/S=1500000/0/0 in:
250000=1500000/6 [rec/s] out:0=0/6 [rec/s]
16/01/30 20:44:30 INFO streaming.PipeMapRed: R/W/S=1600000/0/0 in:
228571=1600000/7 [rec/s] out:0=0/7 [rec/s]
16/01/30 20:44:31 INFO streaming.PipeMapRed: R/W/S=1700000/0/0 in:
212500=1700000/8 [rec/s] out:0=0/8 [rec/s]
16/01/30 20:44:32 INFO mapred.LocalJobRunner: reduce > reduce
16/01/30 20:44:32 INFO mapreduce.Job: map 100% reduce 72%
16/01/30 20:44:32 INFO streaming.PipeMapRed: R/W/S=1800000/0/0 in:
225000=1800000/8 [rec/s] out:0=0/8 [rec/s]
16/01/30 20:44:32 INFO streaming.PipeMapRed: R/W/S=1900000/0/0 in:
211111=1900000/9 [rec/s] out:0=0/9 [rec/s]
16/01/30 20:44:33 INFO streaming.PipeMapRed: R/W/S=2000000/0/0 in:
222222=2000000/9 [rec/s] out:0=0/9 [rec/s]
16/01/30 20:44:33 INFO streaming.PipeMapRed: R/W/S=2100000/0/0 in:
233333=2100000/9 [rec/s] out:0=0/9 [rec/s]
16/01/30 20:44:34 INFO streaming.PipeMapRed: R/W/S=2200000/0/0 in:
220000=2200000/10 [rec/s] out:0=0/10 [rec/s]
16/01/30 20:44:34 INFO streaming.PipeMapRed: R/W/S=2300000/0/0 in:
230000=2300000/10 [rec/s] out:0=0/10 [rec/s]
16/01/30 20:44:34 INFO streaming.PipeMapRed: R/W/S=2400000/0/0 in:
218181=2400000/11 [rec/s] out:0=0/11 [rec/s]
16/01/30 20:44:35 INFO mapred.LocalJobRunner: reduce > reduce
```

```
16/01/30 20:44:35 INFO mapreduce.Job: map 100% reduce 75%
16/01/30 20:44:35 INFO streaming.PipeMapRed: R/W/S=2500000/0/0 in:
227272=2500000/11 [rec/s] out:0=0/11 [rec/s]
16/01/30 20:44:36 INFO streaming.PipeMapRed: R/W/S=2600000/0/0 in:
216666=2600000/12 [rec/s] out:0=0/12 [rec/s]
16/01/30 20:44:36 INFO streaming.PipeMapRed: R/W/S=2700000/0/0 in:
207692=2700000/13 [rec/s] out:0=0/13 [rec/s]
16/01/30 20:44:37 INFO streaming.PipeMapRed: R/W/S=2800000/0/0 in:
215384=2800000/13 [rec/s] out:0=0/13 [rec/s]
16/01/30 20:44:38 INFO streaming.PipeMapRed: R/W/S=2900000/0/0 in:
207142=2900000/14 [rec/s] out:0=0/14 [rec/s]
16/01/30 20:44:38 INFO mapred.LocalJobRunner: reduce > reduce
16/01/30 20:44:38 INFO mapreduce.Job: map 100% reduce 76%
16/01/30 20:44:38 INFO streaming.PipeMapRed: R/W/S=3000000/0/0 in:
214285=3000000/14 [rec/s] out:0=0/14 [rec/s]
16/01/30 20:44:39 INFO streaming.PipeMapRed: R/W/S=3100000/0/0 in:
206666=3100000/15 [rec/s] out:0=0/15 [rec/s]
16/01/30 20:44:39 INFO streaming.PipeMapRed: R/W/S=3200000/0/0 in:
213333=3200000/15 [rec/s] out:0=0/15 [rec/s]
16/01/30 20:44:39 INFO streaming.PipeMapRed: R/W/S=3300000/0/0 in:
206250=3300000/16 [rec/s] out:0=0/16 [rec/s]
16/01/30 20:44:40 INFO streaming.PipeMapRed: R/W/S=3400000/0/0 in:
212500=3400000/16 [rec/s] out:0=0/16 [rec/s]
16/01/30 20:44:40 INFO streaming.PipeMapRed: R/W/S=3500000/0/0 in:
218750=3500000/16 [rec/s] out:0=0/16 [rec/s]
16/01/30 20:44:40 INFO streaming.PipeMapRed: R/W/S=3600000/0/0 in:
211764=3600000/17 [rec/s] out:0=0/17 [rec/s]
16/01/30 20:44:41 INFO mapred.LocalJobRunner: reduce > reduce
16/01/30 20:44:41 INFO mapreduce.Job: map 100% reduce 79%
16/01/30 20:44:41 INFO streaming.PipeMapRed: R/W/S=3700000/0/0 in:
217647=3700000/17 [rec/s] out:0=0/17 [rec/s]
16/01/30 20:44:41 INFO streaming.PipeMapRed: R/W/S=3800000/0/0 in:
223529=3800000/17 [rec/s] out:0=0/17 [rec/s]
16/01/30 20:44:42 INFO streaming.PipeMapRed: R/W/S=3900000/0/0 in:
216666=3900000/18 [rec/s] out:0=0/18 [rec/s]
16/01/30 20:44:42 INFO streaming.PipeMapRed: R/W/S=4000000/0/0 in:
222222=4000000/18 [rec/s] out:0=0/18 [rec/s]
16/01/30 20:44:43 INFO streaming.PipeMapRed: R/W/S=4100000/0/0 in:
215789=4100000/19 [rec/s] out:0=0/19 [rec/s]
16/01/30 20:44:43 INFO streaming.PipeMapRed: R/W/S=4200000/0/0 in:
210000=4200000/20 [rec/s] out:0=0/20 [rec/s]
16/01/30 20:44:44 INFO mapred.LocalJobRunner: reduce > reduce
16/01/30 20:44:44 INFO mapreduce.Job: map 100% reduce 81%
16/01/30 20:44:44 INFO streaming.PipeMapRed: R/W/S=4300000/0/0 in:
215000=4300000/20 [rec/s] out:0=0/20 [rec/s]
16/01/30 20:44:44 INFO streaming.PipeMapRed: R/W/S=4400000/0/0 in:
209523=4400000/21 [rec/s] out:0=0/21 [rec/s]
16/01/30 20:44:45 INFO streaming.PipeMapRed: R/W/S=4500000/0/0 in:
214285=4500000/21 [rec/s] out:0=0/21 [rec/s]
16/01/30 20:44:45 INFO streaming.PipeMapRed: R/W/S=4600000/0/0 in:
219047=4600000/21 [rec/s] out:0=0/21 [rec/s]
16/01/30 20:44:45 INFO streaming.PipeMapRed: R/W/S=4700000/0/0 in:
213636=4700000/22 [rec/s] out:0=0/22 [rec/s]
16/01/30 20:44:46 INFO streaming.PipeMapRed: R/W/S=4800000/0/0 in:
```

```
218181=4800000/22 [rec/s] out:0=0/22 [rec/s]
16/01/30 20:44:46 INFO streaming.PipeMapRed: R/W/S=4900000/0/0 in:
213043=4900000/23 [rec/s] out:0=0/23 [rec/s]
16/01/30 20:44:47 INFO mapred.LocalJobRunner: reduce > reduce
16/01/30 20:44:47 INFO mapreduce.Job: map 100% reduce 83%
16/01/30 20:44:47 INFO streaming.PipeMapRed: R/W/S=5000000/0/0 in:
217391=5000000/23 [rec/s] out:0=0/23 [rec/s]
16/01/30 20:44:47 INFO streaming.PipeMapRed: R/W/S=5100000/0/0 in:
221739=5100000/23 [rec/s] out:0=0/23 [rec/s]
16/01/30 20:44:48 INFO streaming.PipeMapRed: R/W/S=5200000/0/0 in:
216666=5200000/24 [rec/s] out:0=0/24 [rec/s]
16/01/30 20:44:49 INFO streaming.PipeMapRed: R/W/S=5300000/0/0 in:
212000=5300000/25 [rec/s] out:0=0/25 [rec/s]
16/01/30 20:44:50 INFO mapred.LocalJobRunner: reduce > reduce
16/01/30 20:44:50 INFO mapreduce.Job: map 100% reduce 84%
16/01/30 20:44:50 INFO streaming.PipeMapRed: R/W/S=5400000/0/0 in:
207692=5400000/26 [rec/s] out:0=0/26 [rec/s]
16/01/30 20:44:51 INFO streaming.PipeMapRed: R/W/S=5500000/0/0 in:
203703=5500000/27 [rec/s] out:0=0/27 [rec/s]
16/01/30 20:44:53 INFO mapred.LocalJobRunner: reduce > reduce
16/01/30 20:44:53 INFO mapreduce.Job: map 100% reduce 85%
16/01/30 20:44:56 INFO mapred.LocalJobRunner: reduce > reduce
16/01/30 20:44:57 INFO streaming.PipeMapRed: R/W/S=5600000/0/0 in:
169696=5600000/33 [rec/s] out:0=0/33 [rec/s]
16/01/30 20:44:58 INFO streaming.PipeMapRed: R/W/S=5700000/0/0 in:
167647=5700000/34 [rec/s] out:0=0/34 [rec/s]
16/01/30 20:44:59 INFO streaming.PipeMapRed: R/W/S=5800000/0/0 in:
165714=5800000/35 [rec/s] out:0=0/35 [rec/s]
16/01/30 20:44:59 INFO streaming.PipeMapRed: R/W/S=5900000/0/0 in:
163888=5900000/36 [rec/s] out:0=0/36 [rec/s]
16/01/30 20:45:00 INFO streaming.PipeMapRed: R/W/S=6000000/0/0 in:
162162=6000000/37 [rec/s] out:0=0/37 [rec/s]
16/01/30 20:45:01 INFO streaming.PipeMapRed: R/W/S=6100000/0/0 in:
164864=6100000/37 [rec/s] out:0=0/37 [rec/s]
16/01/30 20:45:01 INFO streaming.PipeMapRed: R/W/S=6200000/0/0 in:
163157=6200000/38 [rec/s] out:0=0/38 [rec/s]
16/01/30 20:45:02 INFO mapred.LocalJobRunner: reduce > reduce
16/01/30 20:45:02 INFO mapreduce.Job: map 100% reduce 87%
16/01/30 20:45:02 INFO streaming.PipeMapRed: R/W/S=6300000/0/0 in:
165789=6300000/38 [rec/s] out:0=0/38 [rec/s]
16/01/30 20:45:03 INFO streaming.PipeMapRed: R/W/S=6400000/0/0 in:
164102=6400000/39 [rec/s] out:0=0/39 [rec/s]
16/01/30 20:45:03 INFO streaming.PipeMapRed: R/W/S=6500000/0/0 in:
162500=6500000/40 [rec/s] out:0=0/40 [rec/s]
16/01/30 20:45:04 INFO streaming.PipeMapRed: R/W/S=6600000/0/0 in:
165000=6600000/40 [rec/s] out:0=0/40 [rec/s]
16/01/30 20:45:04 INFO streaming.PipeMapRed: R/W/S=6700000/0/0 in:
167500=6700000/40 [rec/s] out:0=0/40 [rec/s]
16/01/30 20:45:04 INFO streaming.PipeMapRed: R/W/S=6800000/0/0 in:
165853=6800000/41 [rec/s] out:0=0/41 [rec/s]
16/01/30 20:45:05 INFO mapred.LocalJobRunner: reduce > reduce
16/01/30 20:45:05 INFO mapreduce.Job: map 100% reduce 89%
16/01/30 20:45:05 INFO streaming.PipeMapRed: R/W/S=6900000/0/0 in:
168292=6900000/41 [rec/s] out:0=0/41 [rec/s]
```

```
16/01/30 20:45:05 INFO streaming.PipeMapRed: R/W/S=7000000/0/0 in:
170731=7000000/41 [rec/s] out:0=0/41 [rec/s]
16/01/30 20:45:06 INFO streaming.PipeMapRed: R/W/S=7100000/0/0 in:
169047=7100000/42 [rec/s] out:0=0/42 [rec/s]
16/01/30 20:45:06 INFO streaming.PipeMapRed: R/W/S=7200000/0/0 in:
167441=7200000/43 [rec/s] out:0=0/43 [rec/s]
16/01/30 20:45:07 INFO streaming.PipeMapRed: R/W/S=7300000/0/0 in:
169767=7300000/43 [rec/s] out:0=0/43 [rec/s]
16/01/30 20:45:07 INFO streaming.PipeMapRed: R/W/S=7400000/0/0 in:
172093=7400000/43 [rec/s] out:0=0/43 [rec/s]
16/01/30 20:45:08 INFO mapred.LocalJobRunner: reduce > reduce
16/01/30 20:45:08 INFO mapreduce.Job: map 100% reduce 91%
16/01/30 20:45:08 INFO streaming.PipeMapRed: R/W/S=7500000/0/0 in:
170454=7500000/44 [rec/s] out:0=0/44 [rec/s]
16/01/30 20:45:08 INFO streaming.PipeMapRed: R/W/S=7600000/0/0 in:
168888=7600000/45 [rec/s] out:0=0/45 [rec/s]
16/01/30 20:45:09 INFO streaming.PipeMapRed: R/W/S=7700000/0/0 in:
171111=7700000/45 [rec/s] out:0=0/45 [rec/s]
16/01/30 20:45:10 INFO streaming.PipeMapRed: R/W/S=7800000/0/0 in:
169565=7800000/46 [rec/s] out:0=0/46 [rec/s]
16/01/30 20:45:10 INFO streaming.PipeMapRed: R/W/S=7900000/0/0 in:
168085=7900000/47 [rec/s] out:0=0/47 [rec/s]
16/01/30 20:45:11 INFO mapred.LocalJobRunner: reduce > reduce
16/01/30 20:45:11 INFO mapreduce.Job: map 100% reduce 93%
16/01/30 20:45:11 INFO streaming.PipeMapRed: R/W/S=8000000/0/0 in:
170212=8000000/47 [rec/s] out:0=0/47 [rec/s]
16/01/30 20:45:11 INFO streaming.PipeMapRed: R/W/S=8100000/0/0 in:
168750=8100000/48 [rec/s] out:0=0/48 [rec/s]
16/01/30 20:45:12 INFO streaming.PipeMapRed: R/W/S=8200000/0/0 in:
170833=8200000/48 [rec/s] out:0=0/48 [rec/s]
16/01/30 20:45:13 INFO streaming.PipeMapRed: R/W/S=8300000/0/0 in:
169387=8300000/49 [rec/s] out:0=0/49 [rec/s]
16/01/30 20:45:13 INFO streaming.PipeMapRed: R/W/S=8400000/0/0 in:
168000=8400000/50 [rec/s] out:0=0/50 [rec/s]
16/01/30 20:45:14 INFO mapred.LocalJobRunner: reduce > reduce
16/01/30 20:45:14 INFO mapreduce.Job: map 100% reduce 94%
16/01/30 20:45:14 INFO streaming.PipeMapRed: R/W/S=8500000/0/0 in:
166666=8500000/51 [rec/s] out:0=0/51 [rec/s]
16/01/30 20:45:15 INFO streaming.PipeMapRed: R/W/S=8600000/0/0 in:
168627=8600000/51 [rec/s] out:0=0/51 [rec/s]
16/01/30 20:45:15 INFO streaming.PipeMapRed: R/W/S=8700000/0/0 in:
167307=8700000/52 [rec/s] out:0=0/52 [rec/s]
16/01/30 20:45:16 INFO streaming.PipeMapRed: R/W/S=8800000/0/0 in:
169230=8800000/52 [rec/s] out:0=0/52 [rec/s]
16/01/30 20:45:17 INFO streaming.PipeMapRed: R/W/S=8900000/0/0 in:
167924=8900000/53 [rec/s] out:0=0/53 [rec/s]
16/01/30 20:45:17 INFO mapred.LocalJobRunner: reduce > reduce
16/01/30 20:45:17 INFO mapreduce.Job: map 100% reduce 96%
16/01/30 20:45:17 INFO streaming.PipeMapRed: R/W/S=9000000/0/0 in:
166666=9000000/54 [rec/s] out:0=0/54 [rec/s]
16/01/30 20:45:18 INFO streaming.PipeMapRed: R/W/S=9100000/0/0 in:
168518=9100000/54 [rec/s] out:0=0/54 [rec/s]
16/01/30 20:45:19 INFO streaming.PipeMapRed: R/W/S=9200000/0/0 in:
167272=9200000/55 [rec/s] out:0=0/55 [rec/s]
```

```
16/01/30 20:45:19 INFO streaming.PipeMapRed: R/W/S=9300000/0/0 in:
166071=9300000/56 [rec/s] out:0=0/56 [rec/s]
16/01/30 20:45:20 INFO mapred.LocalJobRunner: reduce > reduce
16/01/30 20:45:20 INFO mapreduce.Job: map 100% reduce 98%
16/01/30 20:45:20 INFO streaming.PipeMapRed: R/W/S=9400000/0/0 in:
167857=9400000/56 [rec/s] out:0=0/56 [rec/s]
16/01/30 20:45:21 INFO streaming.PipeMapRed: R/W/S=9500000/0/0 in:
166666=9500000/57 [rec/s] out:0=0/57 [rec/s]
16/01/30 20:45:22 INFO streaming.PipeMapRed: R/W/S=9600000/0/0 in:
165517=9600000/58 [rec/s] out:0=0/58 [rec/s]
16/01/30 20:45:22 INFO streaming.PipeMapRed: R/W/S=9700000/0/0 in:
164406=9700000/59 [rec/s] out:0=0/59 [rec/s]
16/01/30 20:45:23 INFO mapred.LocalJobRunner: reduce > reduce
16/01/30 20:45:23 INFO mapreduce.Job: map 100% reduce 99%
16/01/30 20:45:23 INFO streaming.PipeMapRed: R/W/S=9800000/0/0 in:
166101=9800000/59 [rec/s] out:0=0/59 [rec/s]
16/01/30 20:45:23 INFO streaming.PipeMapRed: R/W/S=9900000/0/0 in:
165000=9900000/60 [rec/s] out:0=0/60 [rec/s]
16/01/30 20:45:24 INFO streaming.PipeMapRed: R/W/S=10000000/0/0 i
n:166666=10000000/60 [rec/s] out:0=0/60 [rec/s]
16/01/30 20:45:26 INFO mapred.LocalJobRunner: reduce > reduce
16/01/30 20:45:26 INFO mapreduce.Job: map 100% reduce 100%
16/01/30 20:45:29 INFO mapred.LocalJobRunner: reduce > reduce
16/01/30 20:45:39 INFO streaming.PipeMapRed: Records R/W=10092961/
1
16/01/30 20:45:39 INFO streaming.PipeMapRed: MRErrorThread done
16/01/30 20:45:39 INFO streaming.PipeMapRed: mapRedFinished
16/01/30 20:45:39 INFO mapred.Task: Task:attempt_local583153488_00
01_r_000000_0 is done. And is in the process of committing
16/01/30 20:45:39 INFO mapred.LocalJobRunner: reduce > reduce
16/01/30 20:45:39 INFO mapred.Task: Task attempt_local583153488_00
01_r_000000_0 is allowed to commit now
16/01/30 20:45:39 INFO output.FileOutputCommitter: Saved output of
task 'attempt_local583153488_0001_r_000000_0' to hdfs://localhost:
9000/user/john/notebook/output/_temporary/0/task_local583153488_00
01_r_000000
16/01/30 20:45:39 INFO mapred.LocalJobRunner: Records R/W=1009296
1/1 > reduce
16/01/30 20:45:39 INFO mapred.Task: Task 'attempt_local583153488_0
001_r_000000_0' done.
16/01/30 20:45:39 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local583153488_0001_r_000000_0
16/01/30 20:45:39 INFO mapred.LocalJobRunner: reduce task executor
complete.
16/01/30 20:45:40 INFO mapreduce.Job: Job job_local583153488_0001
completed successfully
16/01/30 20:45:40 INFO mapreduce.Job: Counters: 36
```

#### File System Counters

```
FILE: Number of bytes read=1762727936
FILE: Number of bytes written=2166684757
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=6917034
```

HDFS: Number of bytes written=8666  
 HDFS: Number of read operations=13  
 HDFS: Number of large read operations=0  
 HDFS: Number of write operations=4

#### Map-Reduce Framework

Map input records=31101  
 Map output records=14728700  
 Map output bytes=544010616  
 Map output materialized bytes=403369371  
 Input split bytes=122  
 Combine input records=26799040  
 Combine output records=22163301  
 Reduce input groups=10092961  
 Reduce shuffle bytes=403369371  
 Reduce input records=10092961  
 Reduce output records=108  
 Spilled Records=32256262  
 Shuffled Maps =1  
 Failed Shuffles=0  
 Merged Map outputs=1  
 GC time elapsed (ms)=52  
 Total committed heap usage (bytes)=526385152

#### Shuffle Errors

BAD\_ID=0  
 CONNECTION=0  
 IO\_ERROR=0  
 WRONG\_LENGTH=0  
 WRONG\_MAP=0  
 WRONG\_REDUCE=0

#### User-Defined

Number of Combiners=11

#### File Input Format Counters

Bytes Read=3458517

#### File Output Format Counters

Bytes Written=8666

16/01/30 20:45:40 INFO streaming.StreamJob: Output directory: /user/john/notebook/output

16/01/30 20:45:42 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

#### ===== Most Frequent Doubles =====

PAIR	SUPPORT COUNT	SUPPORT
('DAI62779', 'ELE17451')	1592	
100		
('FRO40251', 'SNA80324')	1412	
100		
('DAI75645', 'FRO40251')	1254	
100		
('FRO40251', 'GRO85051')	1213	
100		
('DAI62779', 'GRO73461')	1139	
100		



('DAI75645', 'SNA80324')		1130
100		
('DAI62779', 'FRO40251')		1070
100		
('DAI62779', 'SNA80324')		923
100		
('DAI62779', 'DAI85309')		918
100		
('ELE32164', 'GRO59710')		911
100		
('DAI62779', 'DAI75645')		882
100		
('FRO40251', 'GRO73461')		882
100		
('DAI62779', 'ELE92920')		877
100		
('FRO40251', 'FRO92469')		835
100		
('DAI62779', 'ELE32164')		832
100		
('DAI75645', 'GRO73461')		712
100		
('DAI43223', 'ELE32164')		711
100		
('DAI62779', 'GRO30386')		709
100		
('ELE17451', 'FRO40251')		697
100		
('DAI85309', 'ELE99737')		659
100		
('DAI62779', 'ELE26917')		650
100		
('GRO21487', 'GRO73461')		631
100		
('DAI62779', 'SNA45677')		604
100		
('ELE17451', 'SNA80324')		597
100		
('DAI62779', 'GRO71621')		595
100		
('DAI62779', 'SNA55762')		593
100		
('DAI62779', 'DAI83733')		586
100		
('ELE17451', 'GRO73461')		580
100		
('GRO73461', 'SNA80324')		562
100		
('DAI62779', 'GRO59710')		561
100		
('DAI62779', 'FRO80039')		550
100		
('DAI75645', 'ELE17451')		547
100		

('DAI62779', 'SNA93860')		537
100		
('DAI55148', 'DAI62779')		526
100		
('DAI43223', 'GRO59710')		512
100		
('ELE17451', 'ELE32164')		511
100		
('DAI62779', 'SNA18336')		506
100		
('ELE32164', 'GRO73461')		486
100		
('DAI85309', 'ELE17451')		482
100		
('DAI62779', 'FRO78087')		482
100		
('DAI62779', 'GRO94758')		479
100		
('GRO85051', 'SNA80324')		471
100		
('DAI62779', 'GRO21487')		471
100		
('ELE17451', 'GRO30386')		468
100		
('FRO85978', 'SNA95666')		463
100		
('DAI62779', 'FRO19221')		462
100		
('DAI62779', 'GRO46854')		461
100		
('DAI43223', 'DAI62779')		459
100		
('ELE92920', 'SNA18336')		455
100		
('DAI88079', 'FRO40251')		446
100		

===== Most Frequent Triples =====

PAIR		SUPPORT COUNT
SUPPORT		
('DAI75645', 'FRO40251', 'SNA80324')		
550	100	
('DAI62779', 'FRO40251', 'SNA80324')		
476	100	
('FRO40251', 'GRO85051', 'SNA80324')		
471	100	
('DAI62779', 'ELE92920', 'SNA18336')		
432	100	
('DAI62779', 'DAI75645', 'SNA80324')		
421	100	
('DAI62779', 'ELE17451', 'SNA80324')		
417	100	
('DAI62779', 'DAI75645', 'FRO40251')		

412	100	
('DAI62779', 'ELE17451', 'FRO40251')		
406	100	
('DAI75645', 'FRO40251', 'GRO85051')		
395	100	
('DAI62779', 'FRO40251', 'GRO85051')		
381	100	
('ELE17451', 'FRO40251', 'SNA80324')		
353	100	
('DAI62779', 'ELE17451', 'ELE92920')		
345	100	
('FRO40251', 'FRO92469', 'SNA80324')		
343	100	
('DAI62779', 'DAI85309', 'ELE17451')		
339	100	
('DAI62779', 'DAI75645', 'ELE17451')		
328	100	
('DAI62779', 'FRO40251', 'GRO73461')		
315	100	
('DAI62779', 'ELE32164', 'GRO59710')		
301	100	
('DAI75645', 'ELE17451', 'SNA80324')		
300	100	
('DAI75645', 'FRO40251', 'GRO73461')		
293	100	
('DAI75645', 'ELE17451', 'FRO40251')		
292	100	
('DAI43223', 'DAI62779', 'ELE32164')		
287	100	
('DAI43223', 'ELE32164', 'GRO59710')		
287	100	
('DAI62779', 'ELE17451', 'ELE32164')		
277	100	
('DAI62779', 'DAI85309', 'ELE99737')		
272	100	
('DAI62779', 'DAI75645', 'GRO73461')		
261	100	
('DAI75645', 'FRO40251', 'FRO92469')		
251	100	
('DAI62779', 'ELE17451', 'GRO73461')		
245	100	
('DAI62779', 'ELE17451', 'SNA18336')		
244	100	
('DAI62779', 'FRO40251', 'FRO92469')		
238	100	
('ELE20847', 'FRO40251', 'SNA80324')		
232	100	
('FRO40251', 'GRO73461', 'SNA80324')		
232	100	
('DAI75645', 'GRO73461', 'SNA80324')		
230	100	
('ELE17451', 'ELE92920', 'SNA18336')		
228	100	
('DAI43223', 'DAI62779', 'ELE17451')		

```

227| 100
('DAI62779', 'ELE17451', 'GRO30386')
218| 100
('ELE17451', 'FRO40251', 'GRO85051')
217| 100
('DAI62779', 'ELE17451', 'GRO59710')
213| 100
('FRO40251', 'FRO92469', 'GRO73461')
211| 100
('DAI43223', 'ELE17451', 'ELE32164')
206| 100
('DAI62779', 'GRO85051', 'SNA80324')
205| 100
('DAI43223', 'DAI62779', 'GRO59710')
205| 100
('ELE17451', 'ELE32164', 'GRO59710')
202| 100
('DAI62779', 'ELE17451', 'SNA59903')
202| 100
('DAI62779', 'GRO73461', 'SNA80324')
198| 100
('DAI75645', 'GRO85051', 'SNA80324')
192| 100
('DAI62779', 'DAI85309', 'ELE92920')
191| 100
('DAI55148', 'DAI62779', 'FRO40251')
189| 100
('DAI55148', 'DAI62779', 'SNA80324')
188| 100
('DAI62779', 'GRO30386', 'GRO73461')
186| 100
('FRO40251', 'GRO21487', 'GRO73461')
182| 100

```

## Question 3.8 (Optional)

\*\*

### Solution:

Here's some text.