MIDS-W261-2015-HWK-Week03-Anderson_Wen_Sakhamuri

February 4, 2016

```
In [167]: ## HW3 DATSCI W261
```

Names Safyre Anderson, Howard Wen , Vamsi Sakhamuri
Emails safyre@berkelye.edu, howard.wen1@gmail.com, vamsi@ischool.berkeley.edu
Time of Initial Submission: February 4nd, 2016 8am PST
Section W261-3, Spring 2016
Week 3 Homework

0.0.1 HW 3.0.

What is a merge sort? Where is it used in Hadoop?
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.
What is the Hadoop shuffle?

0.0.2 HW 3.0.1

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

- Merge-sort is a sorting algorithm which runs in O(nlog(n)) time. In hadoop, it is used in between the map outputs and the reduce inputs. Output from each map output is sorted by the key and then merged with the other sorted outputs from other map outputs.

0.0.3 HW 3.0.2

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.

- Combiner function is used after the map output is generated. The values from the same keys are then aggregated to reduce network communication in their transfer to the reduce nodes. Network communication can be a bottleneck and a combiner helps alleviate this problem. Whether a combiner function is run or not is entirely dependent on the execution framework. To be totally sure that intermediate map outputs are combined, in-mapper combining function can be employed. This ensures that aggregation will happen 100% before being shuffled across the network to the various reduce nodes.

- An example where combiners (either the optional one provided by the execution framework or the in-mapper combiner) is helpful would be the word count application. Without any sort of combiners, we would be transfer the same key multiple times within the same map task.

  For example, if the map input is "hello hello hi hi hello hello",

  Without combiners, the following pairs will be emitted across the network and into the reduce nodes:

  hello,1

  hello,1

  hi,1

hi,1

hello,1

hello,1

With combiners, the following pairs will be emitted across the network and into the reduce nodes:

hello,4

hi,2

This is an incredible reduction in network traffic.

0.0.4 HW 3.0.3

What is the Hadoop shuffle?

- Hadoop Shuffle: Map-reduce makes the guarantee that the input to every reducer is sorted by the key. And that the pairs with the same key are routed to the same reducer. The process by which the system performs this sort and merge (across various mappers) and transfers the map outputs to the reducers is known as the hadoop shuffle.

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

Counters are lightweight objects in Hadoop that allow you to keep track of system progress in both the map and reduce stages of processing. By default, Hadoop defines a number of standard counters in "groups"; these show up in the jobtracker webapp, giving you information such as "Map input records", "Map output records", etc. While processing information/data using MapReduce job, it is a challenge to monitor the progress of parallel threads running across nodes of distributed clusters. Moreover, it is also complicated to distinguish between the data that has been processed and the data which is yet to be processed. The MapReduce Framework offers a provision of user-defined Counters, which can be effectively utilized to monitor the progress of data across nodes of distributed clusters. Use the Consumer Complaints Dataset provide here to complete this question:

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

The consumer complaints dataset consists of diverse consumer complaints, which have been reported across the United States regarding various types of loans. The dataset consists of records of the form:
Complaint ID,Product,Sub-product,Issue,Sub-issue,State,ZIP code,Submitted via,Date received,Date sent to company,Company,Company response,Timely response?,Consumer disputed?
Here's is the first few lines of the of the Consumer Complaints Dataset:

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

```
In [168]: #Start hdfs

          !/Users/Vamsi/Downloads/hadoop-2.7.1/sbin/start-dfs.sh
```

```
16/02/04 08:14:48 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
Starting namenodes on [localhost]
localhost: namenode running as process 1498. Stop it first.
localhost: datanode running as process 1042. Stop it first.
Starting secondary namenodes [0.0.0.0]
0.0.0.0: secondarynamenode running as process 1162. Stop it first.
16/02/04 08:14:53 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
```

```
In [169]: #start the jobtracker

          !/Users/Vamsi/Downloads/hadoop-2.7.1/sbin/mr-jobhistory-daemon.sh --config /Users/Vamsi/Downlo
```

```
historyserver running as process 1275. Stop it first.
```

```
In [170]: #Service check
          !jps
```

```
1042 DataNode
1162 SecondaryNameNode
1498 NameNode
1275 JobHistoryServer
4719 Jps
```

```
In [171]: #Start afresh
          !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -rm -r -f  /user/hw3
```

```
16/02/04 08:14:56 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
16/02/04 08:14:57 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minute
Deleted /user/hw3
```

```
In [172]: #Create a directory for hw3
          !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -mkdir -p /user/hw3
```

```
16/02/04 08:14:59 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
```

```
In [173]: #Load the dataset for 3.1
          !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -put Consumer_Complaints.csv /user/hw3
```

```
16/02/04 08:15:02 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
```

```
In [174]: %%writefile mapper.py
          #!/usr/bin/python
          import sys
          import re
          import csv

          csv_iter = csv.reader(sys.stdin)

          #Grabbing the header names
          header = next(csv_iter)

          for line in csv_iter:
              if(len(line)==len(header)):   #continue only if all the rows have all the entries
                  if(line[1].lower()=="debt collection"):
                      sys.stderr.write("reporter:counter:Product,Debt,1\n")
                  elif(line[1].lower() == "mortgage"):
                      sys.stderr.write("reporter:counter:Product,Mortgage,1\n")
                  else:
                      sys.stderr.write("reporter:counter:Product,Other,1\n")
```

```
Overwriting mapper.py
```

```
In [175]: !chmod a+x mapper.py
```

```
In [176]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hadoop jar /Users/Vamsi/Downloads/hadoop-2.7.1/bin/ha
          -D mapred.reduce.tasks=1 \
          -input /user/hw3/Consumer_Complaints.csv \
          -output /user/hw3/output_3_1 \
          -mapper mapper.py \
          -reducer org.apache.hadoop.mapred.lib.IdentityReducer
```

```
16/02/04 08:15:07 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
16/02/04 08:15:08 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.ses
16/02/04 08:15:08 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/02/04 08:15:08 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessio
16/02/04 08:15:09 INFO mapred.FileInputFormat: Total input paths to process : 1
16/02/04 08:15:09 INFO mapreduce.JobSubmitter: number of splits:1
16/02/04 08:15:09 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapred
16/02/04 08:15:09 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local12979890_0001
16/02/04 08:15:09 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/02/04 08:15:09 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/02/04 08:15:09 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputComm
16/02/04 08:15:09 INFO mapreduce.Job: Running job: job_local12979890_0001
16/02/04 08:15:09 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:15:10 INFO mapred.LocalJobRunner: Waiting for map tasks
16/02/04 08:15:10 INFO mapred.LocalJobRunner: Starting task: attempt_local12979890_0001_m_000000_0
16/02/04 08:15:10 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:15:10 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only o
16/02/04 08:15:10 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:15:10 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/hw3/Consumer_Complai
16/02/04 08:15:10 INFO mapred.MapTask: numReduceTasks: 1
16/02/04 08:15:10 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/02/04 08:15:10 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/02/04 08:15:10 INFO mapred.MapTask: soft limit at 83886080
16/02/04 08:15:10 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/02/04 08:15:10 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/02/04 08:15:10 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$Map
16/02/04 08:15:10 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./mapper.p
16/02/04 08:15:10 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.ta
16/02/04 08:15:10 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce
16/02/04 08:15:10 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.m
16/02/04 08:15:10 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.j
16/02/04 08:15:10 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce
16/02/04 08:15:10 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use map
16/02/04 08:15:10 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce
16/02/04 08:15:10 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.jo
16/02/04 08:15:10 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.us
16/02/04 08:15:10 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapredu
16/02/04 08:15:10 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.t
16/02/04 08:15:10 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapr
16/02/04 08:15:10 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:15:10 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:15:10 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:15:10 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:15:10 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
```

```
16/02/04 08:15:10 INFO mapreduce.Job: Job job_local12979890_0001 running in uber mode : false
16/02/04 08:15:10 INFO mapreduce.Job:  map 0% reduce 0%
16/02/04 08:15:12 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:100000=100000/1 [rec/s] out:0=0/1 [rec
16/02/04 08:15:13 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:100000=200000/2 [rec/s] out:0=0/2 [rec
16/02/04 08:15:14 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:100000=300000/3 [rec/s] out:0=0/3 [rec
16/02/04 08:15:14 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:15:14 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:15:14 INFO mapred.LocalJobRunner:
16/02/04 08:15:14 INFO mapred.MapTask: Starting flush of map output
16/02/04 08:15:14 INFO mapred.Task: Task:attempt_local12979890_0001_m_000000_0 is done. And is in the pro
16/02/04 08:15:14 INFO mapred.LocalJobRunner: hdfs://localhost:9000/user/hw3/Consumer_Complaints.csv:0+5
16/02/04 08:15:14 INFO mapred.Task: Task 'attempt_local12979890_0001_m_000000_0' done.
16/02/04 08:15:14 INFO mapred.LocalJobRunner: Finishing task: attempt_local12979890_0001_m_000000_0
16/02/04 08:15:14 INFO mapred.LocalJobRunner: map task executor complete.
16/02/04 08:15:14 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/02/04 08:15:14 INFO mapred.LocalJobRunner: Starting task: attempt_local12979890_0001_r_000000_0
16/02/04 08:15:14 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:15:14 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
16/02/04 08:15:14 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:15:14 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task
16/02/04 08:15:14 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleL
16/02/04 08:15:14 INFO reduce.EventFetcher: attempt_local12979890_0001_r_000000_0 Thread started: EventFe
16/02/04 08:15:14 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1
16/02/04 08:15:14 INFO reduce.InMemoryMapOutput: Read 2 bytes from map-output for attempt_local12979890_
16/02/04 08:15:14 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 2, inMemoryMap
16/02/04 08:15:14 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/02/04 08:15:14 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/04 08:15:14 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-
16/02/04 08:15:14 INFO mapred.Merger: Merging 1 sorted segments
16/02/04 08:15:14 INFO mapred.Merger: Down to the last merge-pass, with 0 segments left of total size:
16/02/04 08:15:14 INFO reduce.MergeManagerImpl: Merged 1 segments, 2 bytes to disk to satisfy reduce mem
16/02/04 08:15:14 INFO reduce.MergeManagerImpl: Merging 1 files, 6 bytes from disk
16/02/04 08:15:14 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/02/04 08:15:14 INFO mapred.Merger: Merging 1 sorted segments
16/02/04 08:15:14 INFO mapred.Merger: Down to the last merge-pass, with 0 segments left of total size:
16/02/04 08:15:14 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/04 08:15:14 INFO mapred.Task: Task:attempt_local12979890_0001_r_000000_0 is done. And is in the pro
16/02/04 08:15:14 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/04 08:15:14 INFO mapred.Task: Task attempt_local12979890_0001_r_000000_0 is allowed to commit now
16/02/04 08:15:14 INFO output.FileOutputCommitter: Saved output of task 'attempt_local12979890_0001_r_00
16/02/04 08:15:14 INFO mapred.LocalJobRunner: reduce > reduce
16/02/04 08:15:14 INFO mapred.Task: Task 'attempt_local12979890_0001_r_000000_0' done.
16/02/04 08:15:14 INFO mapred.LocalJobRunner: Finishing task: attempt_local12979890_0001_r_000000_0
16/02/04 08:15:14 INFO mapred.LocalJobRunner: reduce task executor complete.
16/02/04 08:15:14 INFO mapreduce.Job:  map 100% reduce 100%
16/02/04 08:15:15 INFO mapreduce.Job: Job job_local12979890_0001 completed successfully
16/02/04 08:15:15 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=211842
                FILE: Number of bytes written=764874
                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=0
                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=0
                Map output bytes=0
                Map output materialized bytes=6
                Input split bytes=106
                Combine input records=0
                Combine output records=0
                Reduce input groups=0
                Reduce shuffle bytes=6
                Reduce input records=0
                Reduce output records=0
                Spilled Records=0
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=20
                Total committed heap usage (bytes)=559415296
        Product
                Debt=44372
                Mortgage=125752
                Other=142788
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=50906486
        File Output Format Counters
                Bytes Written=0
16/02/04 08:15:15 INFO streaming.StreamJob: Output directory: /user/hw3/output_3_1
```

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

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

0.0.7 3.2.1

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

```
In [177]: #Create a file
          !echo "foo foo quux labs foo bar quux" >input_3_2
```

```
In [178]: #Load the file into hdfs
          !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -put input_3_2 /user/hw3
```

```
16/02/04 08:15:16 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
```

```
In [179]: %%writefile mapper.py
          #!/usr/bin/python
          import sys
          import re

          sys.stderr.write("reporter:counter:MAPPER,mapper_calls,1\n")   #trigger counter

          for line in sys.stdin:
              line_s = re.split(r'[\s]',line)   #split on whitespace
              for l in line_s:
                  print "%s\t1" %l.strip()    #Emit word,1
```

```
Overwriting mapper.py
```

```
In [180]: !chmod a+x mapper.py
```

```
In [181]: %%writefile reducer.py
          #!/usr/bin/python
          import sys
          import re

          count = 0

          prev_string = None

          sys.stderr.write("reporter:counter:REDUCER,reducer_calls,1\n")  #trigger counter

          for line in sys.stdin:
              line_s = re.split(r'[\t]',line)   #split on tab

              if((prev_string!=None) and (prev_string != line_s[0])): # Check whether the a new key is e
                  print prev_string,count  # If a new key is emitted then commit the previous key and c
                  count = 0   #reset the counter

              count += 1    #increment the counter
              prev_string = line_s[0]
          print prev_string,count
```

```
Overwriting reducer.py
```

```
In [182]: !chmod a+x reducer.py
```

```
In [183]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hadoop jar /Users/Vamsi/Downloads/hadoop-2.7.1/bin/ha
          -D mapred.reduce.tasks=4 \
          -input /user/hw3/input_3_2 \
          -output /user/hw3/output_3_2_1 \
          -mapper mapper.py \
          -reducer reducer.py
```

```
16/02/04 08:15:19 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
16/02/04 08:15:20 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.se
```

```
16/02/04 08:15:20 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/02/04 08:15:20 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessi
16/02/04 08:15:20 INFO mapred.FileInputFormat: Total input paths to process : 1
16/02/04 08:15:20 INFO mapreduce.JobSubmitter: number of splits:1
16/02/04 08:15:20 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapre
16/02/04 08:15:20 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local167321748_0001
16/02/04 08:15:21 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/02/04 08:15:21 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/02/04 08:15:21 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCom
16/02/04 08:15:21 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:15:21 INFO mapreduce.Job: Running job: job_local167321748_0001
16/02/04 08:15:21 INFO mapred.LocalJobRunner: Waiting for map tasks
16/02/04 08:15:21 INFO mapred.LocalJobRunner: Starting task: attempt_local167321748_0001_m_000000_0
16/02/04 08:15:21 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:15:21 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
16/02/04 08:15:21 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:15:21 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/hw3/input_3_2:0+31
16/02/04 08:15:21 INFO mapred.MapTask: numReduceTasks: 4
16/02/04 08:15:21 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/02/04 08:15:21 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/02/04 08:15:21 INFO mapred.MapTask: soft limit at 83886080
16/02/04 08:15:21 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/02/04 08:15:21 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/02/04 08:15:21 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$Ma
16/02/04 08:15:21 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./mapper.
16/02/04 08:15:21 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.t
16/02/04 08:15:21 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduc
16/02/04 08:15:21 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.
16/02/04 08:15:21 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.
16/02/04 08:15:21 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduc
16/02/04 08:15:21 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use ma
16/02/04 08:15:21 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce
16/02/04 08:15:21 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.j
16/02/04 08:15:21 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.u
16/02/04 08:15:21 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapred
16/02/04 08:15:21 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.
16/02/04 08:15:21 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use map
16/02/04 08:15:21 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:15:21 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:15:21 INFO streaming.PipeMapRed: Records R/W=1/1
16/02/04 08:15:21 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:15:21 INFO mapred.LocalJobRunner:
16/02/04 08:15:21 INFO mapred.MapTask: Starting flush of map output
16/02/04 08:15:21 INFO mapred.MapTask: Spilling map output
16/02/04 08:15:21 INFO mapred.MapTask: bufstart = 0; bufend = 48; bufvoid = 104857600
16/02/04 08:15:21 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214368(104857472); leng
16/02/04 08:15:21 INFO mapred.MapTask: Finished spill 0
16/02/04 08:15:21 INFO mapred.Task: Task:attempt_local167321748_0001_m_000000_0 is done. And is in the p
16/02/04 08:15:21 INFO mapred.LocalJobRunner: Records R/W=1/1
16/02/04 08:15:21 INFO mapred.Task: Task 'attempt_local167321748_0001_m_000000_0' done.
16/02/04 08:15:21 INFO mapred.LocalJobRunner: Finishing task: attempt_local167321748_0001_m_000000_0
16/02/04 08:15:21 INFO mapred.LocalJobRunner: map task executor complete.
16/02/04 08:15:21 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/02/04 08:15:21 INFO mapred.LocalJobRunner: Starting task: attempt_local167321748_0001_r_000000_0
```

```
16/02/04 08:15:21 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:15:21 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only o
16/02/04 08:15:21 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:15:21 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task
16/02/04 08:15:21 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLi
16/02/04 08:15:21 INFO reduce.EventFetcher: attempt_local167321748_0001_r_000000_0 Thread started: EventF
16/02/04 08:15:21 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1
16/02/04 08:15:21 INFO reduce.InMemoryMapOutput: Read 20 bytes from map-output for attempt_local1673217
16/02/04 08:15:21 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 20, inMemoryMa
16/02/04 08:15:21 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/02/04 08:15:21 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/04 08:15:21 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on
16/02/04 08:15:21 INFO mapred.Merger: Merging 1 sorted segments
16/02/04 08:15:21 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size:
16/02/04 08:15:21 INFO reduce.MergeManagerImpl: Merged 1 segments, 20 bytes to disk to satisfy reduce me
16/02/04 08:15:21 INFO reduce.MergeManagerImpl: Merging 1 files, 24 bytes from disk
16/02/04 08:15:21 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/02/04 08:15:21 INFO mapred.Merger: Merging 1 sorted segments
16/02/04 08:15:21 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size:
16/02/04 08:15:21 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/04 08:15:21 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./reducer
16/02/04 08:15:21 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapredu
16/02/04 08:15:21 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce
16/02/04 08:15:21 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:15:21 INFO streaming.PipeMapRed: Records R/W=2/1
16/02/04 08:15:21 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:15:21 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:15:21 INFO mapred.Task: Task:attempt_local167321748_0001_r_000000_0 is done. And is in the pr
16/02/04 08:15:21 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/04 08:15:21 INFO mapred.Task: Task attempt_local167321748_0001_r_000000_0 is allowed to commit now
16/02/04 08:15:21 INFO output.FileOutputCommitter: Saved output of task 'attempt_local167321748_0001_r_00
16/02/04 08:15:21 INFO mapred.LocalJobRunner: Records R/W=2/1 > reduce
16/02/04 08:15:21 INFO mapred.Task: Task 'attempt_local167321748_0001_r_000000_0' done.
16/02/04 08:15:21 INFO mapred.LocalJobRunner: Finishing task: attempt_local167321748_0001_r_000000_0
16/02/04 08:15:21 INFO mapred.LocalJobRunner: Starting task: attempt_local167321748_0001_r_000001_0
16/02/04 08:15:21 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:15:21 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only o
16/02/04 08:15:21 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:15:21 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task
16/02/04 08:15:21 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLi
16/02/04 08:15:21 INFO reduce.EventFetcher: attempt_local167321748_0001_r_000001_0 Thread started: EventF
16/02/04 08:15:21 INFO reduce.LocalFetcher: localfetcher#2 about to shuffle output of map attempt_local1
16/02/04 08:15:21 INFO reduce.InMemoryMapOutput: Read 31 bytes from map-output for attempt_local1673217
16/02/04 08:15:21 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 31, inMemoryMa
16/02/04 08:15:21 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/02/04 08:15:21 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/04 08:15:21 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on
16/02/04 08:15:21 INFO mapred.Merger: Merging 1 sorted segments
16/02/04 08:15:21 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size:
16/02/04 08:15:21 INFO reduce.MergeManagerImpl: Merged 1 segments, 31 bytes to disk to satisfy reduce me
16/02/04 08:15:21 INFO reduce.MergeManagerImpl: Merging 1 files, 35 bytes from disk
16/02/04 08:15:21 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/02/04 08:15:21 INFO mapred.Merger: Merging 1 sorted segments
16/02/04 08:15:21 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size:
```

```
16/02/04 08:15:21 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/04 08:15:21 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./reducer
16/02/04 08:15:21 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:15:22 INFO streaming.PipeMapRed: Records R/W=4/1
16/02/04 08:15:22 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:15:22 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:15:22 INFO mapreduce.Job: Job job_local167321748_0001 running in uber mode : false
16/02/04 08:15:22 INFO mapreduce.Job:  map 100% reduce 25%
16/02/04 08:15:22 INFO mapred.Task: Task:attempt_local167321748_0001_r_000001_0 is done. And is in the pr
16/02/04 08:15:22 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/04 08:15:22 INFO mapred.Task: Task attempt_local167321748_0001_r_000001_0 is allowed to commit now
16/02/04 08:15:22 INFO output.FileOutputCommitter: Saved output of task 'attempt_local167321748_0001_r_00
16/02/04 08:15:22 INFO mapred.LocalJobRunner: Records R/W=4/1 > reduce
16/02/04 08:15:22 INFO mapred.Task: Task 'attempt_local167321748_0001_r_000001_0' done.
16/02/04 08:15:22 INFO mapred.LocalJobRunner: Finishing task: attempt_local167321748_0001_r_000001_0
16/02/04 08:15:22 INFO mapred.LocalJobRunner: Starting task: attempt_local167321748_0001_r_000002_0
16/02/04 08:15:22 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:15:22 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only o
16/02/04 08:15:22 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:15:22 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task
16/02/04 08:15:22 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLi
16/02/04 08:15:22 INFO reduce.EventFetcher: attempt_local167321748_0001_r_000002_0 Thread started: EventF
16/02/04 08:15:22 INFO reduce.LocalFetcher: localfetcher#3 about to shuffle output of map attempt_local1
16/02/04 08:15:22 INFO reduce.InMemoryMapOutput: Read 10 bytes from map-output for attempt_local16732174
16/02/04 08:15:22 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 10, inMemoryMap
16/02/04 08:15:22 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/02/04 08:15:22 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/04 08:15:22 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-
16/02/04 08:15:22 INFO mapred.Merger: Merging 1 sorted segments
16/02/04 08:15:22 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 4
16/02/04 08:15:22 INFO reduce.MergeManagerImpl: Merged 1 segments, 10 bytes to disk to satisfy reduce me
16/02/04 08:15:22 INFO reduce.MergeManagerImpl: Merging 1 files, 14 bytes from disk
16/02/04 08:15:22 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/02/04 08:15:22 INFO mapred.Merger: Merging 1 sorted segments
16/02/04 08:15:22 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 4
16/02/04 08:15:22 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/04 08:15:22 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./reducer
16/02/04 08:15:22 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:15:22 INFO streaming.PipeMapRed: Records R/W=1/1
16/02/04 08:15:22 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:15:22 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:15:22 INFO mapred.Task: Task:attempt_local167321748_0001_r_000002_0 is done. And is in the pr
16/02/04 08:15:22 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/04 08:15:22 INFO mapred.Task: Task attempt_local167321748_0001_r_000002_0 is allowed to commit now
16/02/04 08:15:22 INFO output.FileOutputCommitter: Saved output of task 'attempt_local167321748_0001_r_00
16/02/04 08:15:22 INFO mapred.LocalJobRunner: Records R/W=1/1 > reduce
16/02/04 08:15:22 INFO mapred.Task: Task 'attempt_local167321748_0001_r_000002_0' done.
16/02/04 08:15:22 INFO mapred.LocalJobRunner: Finishing task: attempt_local167321748_0001_r_000002_0
16/02/04 08:15:22 INFO mapred.LocalJobRunner: Starting task: attempt_local167321748_0001_r_000003_0
16/02/04 08:15:22 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:15:22 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only o
16/02/04 08:15:22 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:15:22 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task
16/02/04 08:15:22 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLi
```

```
16/02/04 08:15:22 INFO reduce.EventFetcher: attempt_local167321748_0001_r_000003_0 Thread started: EventF
16/02/04 08:15:22 INFO reduce.LocalFetcher: localfetcher#4 about to shuffle output of map attempt_local1
16/02/04 08:15:22 INFO reduce.InMemoryMapOutput: Read 11 bytes from map-output for attempt_local16732174
16/02/04 08:15:22 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 11, inMemoryMap
16/02/04 08:15:22 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/02/04 08:15:22 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/04 08:15:22 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-
16/02/04 08:15:22 INFO mapred.Merger: Merging 1 sorted segments
16/02/04 08:15:22 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 4
16/02/04 08:15:22 INFO reduce.MergeManagerImpl: Merged 1 segments, 11 bytes to disk to satisfy reduce me
16/02/04 08:15:22 INFO reduce.MergeManagerImpl: Merging 1 files, 15 bytes from disk
16/02/04 08:15:22 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/02/04 08:15:22 INFO mapred.Merger: Merging 1 sorted segments
16/02/04 08:15:22 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 4
16/02/04 08:15:22 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/04 08:15:22 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./reducer
16/02/04 08:15:22 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:15:22 INFO streaming.PipeMapRed: Records R/W=1/1
16/02/04 08:15:22 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:15:22 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:15:22 INFO mapred.Task: Task:attempt_local167321748_0001_r_000003_0 is done. And is in the pr
16/02/04 08:15:22 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/04 08:15:22 INFO mapred.Task: Task attempt_local167321748_0001_r_000003_0 is allowed to commit now
16/02/04 08:15:22 INFO output.FileOutputCommitter: Saved output of task 'attempt_local167321748_0001_r_00
16/02/04 08:15:22 INFO mapred.LocalJobRunner: Records R/W=1/1 > reduce
16/02/04 08:15:22 INFO mapred.Task: Task 'attempt_local167321748_0001_r_000003_0' done.
16/02/04 08:15:22 INFO mapred.LocalJobRunner: Finishing task: attempt_local167321748_0001_r_000003_0
16/02/04 08:15:22 INFO mapred.LocalJobRunner: reduce task executor complete.
16/02/04 08:15:23 INFO mapreduce.Job:  map 100% reduce 100%
16/02/04 08:15:23 INFO mapreduce.Job: Job job_local167321748_0001 completed successfully
16/02/04 08:15:23 INFO mapreduce.Job: Counters: 37
        File System Counters
                FILE: Number of bytes read=531306
                FILE: Number of bytes written=1926094
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=155
                HDFS: Number of bytes written=87
                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=8
                Map output bytes=48
                Map output materialized bytes=88
                Input split bytes=92
                Combine input records=0
                Combine output records=0
                Reduce input groups=5
                Reduce shuffle bytes=88
                Reduce input records=8
                Reduce output records=5
                Spilled Records=16
                Shuffled Maps =4
                Failed Shuffles=0
                Merged Map outputs=4
                GC time elapsed (ms)=8
                Total committed heap usage (bytes)=1418723328
        MAPPER
                mapper_calls=1
        REDUCER
                reducer_calls=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=31
        File Output Format Counters
                Bytes Written=34
16/02/04 08:15:23 INFO streaming.StreamJob: Output directory: /user/hw3/output_3_2_1
```

The number of mapper calls is 1 because the file is small enough to be processed by just 1 mapper. The number of reducer calls is 4 because there are 4 unique words and 4 reducers are used.

** Following 4 cells would output each of the reducer outputs **

```
In [184]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -cat /user/hw3/output_3_2_1/part-00000
```

```
16/02/04 08:15:24 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
quux 2
```

```
In [185]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -cat /user/hw3/output_3_2_1/part-00001
```

```
16/02/04 08:15:26 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
 1
foo 3
```

```
In [186]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -cat /user/hw3/output_3_2_1/part-00002
```

```
16/02/04 08:15:29 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
bar 1
```

```
In [187]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -cat /user/hw3/output_3_2_1/part-00003
```

```
16/02/04 08:15:31 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
labs 1
```

0.0.8 3.2.2

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

```
In [188]: #Removing the header from the Consumer_Complaints.csv file
          !tail -n +2 Consumer_Complaints.csv > Consumer_Complaints_no_header.csv
```

```
In [189]: #Split the file into 4 equally sized files
          !split -l 78228 Consumer_Complaints_no_header.csv
```

```
In [190]: #Load the splitted files into hdfs
          !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -put xa* /user/hw3
```

```
16/02/04 08:15:41 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
```

```
In [191]: %%writefile mapper.py
          #!/usr/bin/python
          import sys
          import re
          import csv

          csv_iter = csv.reader(sys.stdin)

          for line in csv_iter:
              sys.stderr.write("reporter:counter:MAPPER,mapper_calls,1\n")
              if(len(line)==14):   #continue only if all the rows have all the entries
                  issues = re.split(r'[\s,./]+',line[3].strip())      #splitting the issues column
                  for w in issues:
                      print "%s\t1" %w #Emit word,1 to the reducer
```

```
Overwriting mapper.py
```

```
In [192]: !chmod a+x mapper.py
```

```
In [193]: %%writefile reducer.py
          #!/usr/bin/python
          import sys
          import re
          import csv

          count = 0

          prev_string = None

          for line in sys.stdin:
              sys.stderr.write("reporter:counter:REDUCER,reducer_calls,1\n")
              line_s = re.split(r'[\t]',line.strip())

              if((prev_string!=None) and (prev_string != line_s[0])):
                  print prev_string,count
                  count = 0

              count += 1
              prev_string = line_s[0]
          print prev_string,count
```

```
Overwriting reducer.py
```

```
In [194]: !chmod a+x reducer.py
```

```
In [195]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -rm -r -f /user/hw3/output_3_2_2
```

```
16/02/04 08:15:45 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
```

```
In [196]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hadoop jar /Users/Vamsi/Downloads/hadoop-2.7.1/bin/ha
          -D mapred.reduce.tasks=4 \
          -input /user/hw3/xa* \
          -output /user/hw3/output_3_2_2 \
          -mapper mapper.py \
          -reducer reducer.py
```

```
16/02/04 08:15:48 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
16/02/04 08:15:49 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.ses
16/02/04 08:15:49 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/02/04 08:15:49 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessio
16/02/04 08:15:49 INFO mapred.FileInputFormat: Total input paths to process : 4
16/02/04 08:15:49 INFO mapreduce.JobSubmitter: number of splits:4
16/02/04 08:15:49 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapred
16/02/04 08:15:49 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local983135479_0001
16/02/04 08:15:50 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/02/04 08:15:50 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/02/04 08:15:50 INFO mapreduce.Job: Running job: job_local983135479_0001
16/02/04 08:15:50 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputComm
16/02/04 08:15:50 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:15:50 INFO mapred.LocalJobRunner: Waiting for map tasks
16/02/04 08:15:50 INFO mapred.LocalJobRunner: Starting task: attempt_local983135479_0001_m_000000_0
16/02/04 08:15:50 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:15:50 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only o
16/02/04 08:15:50 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:15:50 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/hw3/xab:0+13160945
16/02/04 08:15:50 INFO mapred.MapTask: numReduceTasks: 4
16/02/04 08:15:50 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/02/04 08:15:50 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/02/04 08:15:50 INFO mapred.MapTask: soft limit at 83886080
16/02/04 08:15:50 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/02/04 08:15:50 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/02/04 08:15:50 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$Map
16/02/04 08:15:50 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./mapper.
16/02/04 08:15:50 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.ta
16/02/04 08:15:50 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce
16/02/04 08:15:50 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.m
16/02/04 08:15:50 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.j
16/02/04 08:15:50 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce
16/02/04 08:15:50 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use map
16/02/04 08:15:50 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce
16/02/04 08:15:50 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.jo
16/02/04 08:15:50 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.us
16/02/04 08:15:50 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapredu
16/02/04 08:15:50 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.t
16/02/04 08:15:50 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapr
16/02/04 08:15:50 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:15:50 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:15:50 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:15:50 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:15:50 INFO streaming.PipeMapRed: Records R/W=1552/1
16/02/04 08:15:50 INFO streaming.PipeMapRed: R/W/S=10000/38814/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:15:51 INFO mapreduce.Job: Job job_local983135479_0001 running in uber mode : false
```

```
16/02/04 08:15:51 INFO mapreduce.Job:  map 0% reduce 0%
16/02/04 08:15:52 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:15:52 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:15:52 INFO mapred.LocalJobRunner:
16/02/04 08:15:52 INFO mapred.MapTask: Starting flush of map output
16/02/04 08:15:52 INFO mapred.MapTask: Spilling map output
16/02/04 08:15:52 INFO mapred.MapTask: bufstart = 0; bufend = 3385621; bufvoid = 104857600
16/02/04 08:15:52 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 24820088(99280352); lengt
16/02/04 08:15:53 INFO mapred.MapTask: Finished spill 0
16/02/04 08:15:53 INFO mapred.Task: Task:attempt_local983135479_0001_m_000000_0 is done. And is in the pr
16/02/04 08:15:53 INFO mapred.LocalJobRunner: Records R/W=1552/1
16/02/04 08:15:53 INFO mapred.Task: Task 'attempt_local983135479_0001_m_000000_0' done.
16/02/04 08:15:53 INFO mapred.LocalJobRunner: Finishing task: attempt_local983135479_0001_m_000000_0
16/02/04 08:15:53 INFO mapred.LocalJobRunner: Starting task: attempt_local983135479_0001_m_000001_0
16/02/04 08:15:53 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:15:53 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only o
16/02/04 08:15:53 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:15:53 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/hw3/xaa:0+13123345
16/02/04 08:15:53 INFO mapred.MapTask: numReduceTasks: 4
16/02/04 08:15:53 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/02/04 08:15:53 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/02/04 08:15:53 INFO mapred.MapTask: soft limit at 83886080
16/02/04 08:15:53 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/02/04 08:15:53 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/02/04 08:15:53 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$Ma
16/02/04 08:15:53 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./mapper.
16/02/04 08:15:53 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:15:53 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:15:53 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:15:53 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:15:53 INFO streaming.PipeMapRed: Records R/W=1571/1
16/02/04 08:15:53 INFO streaming.PipeMapRed: R/W/S=10000/39416/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:15:54 INFO mapreduce.Job:  map 100% reduce 0%
16/02/04 08:15:55 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:15:55 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:15:55 INFO mapred.LocalJobRunner:
16/02/04 08:15:55 INFO mapred.MapTask: Starting flush of map output
16/02/04 08:15:55 INFO mapred.MapTask: Spilling map output
16/02/04 08:15:55 INFO mapred.MapTask: bufstart = 0; bufend = 3419326; bufvoid = 104857600
16/02/04 08:15:55 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 24795408(99181632); lengt
16/02/04 08:15:56 INFO mapreduce.Job:  map 25% reduce 0%
16/02/04 08:15:56 INFO mapred.MapTask: Finished spill 0
16/02/04 08:15:56 INFO mapred.Task: Task:attempt_local983135479_0001_m_000001_0 is done. And is in the pr
16/02/04 08:15:56 INFO mapred.LocalJobRunner: Records R/W=1571/1
16/02/04 08:15:56 INFO mapred.Task: Task 'attempt_local983135479_0001_m_000001_0' done.
16/02/04 08:15:56 INFO mapred.LocalJobRunner: Finishing task: attempt_local983135479_0001_m_000001_0
16/02/04 08:15:56 INFO mapred.LocalJobRunner: Starting task: attempt_local983135479_0001_m_000002_0
16/02/04 08:15:56 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:15:56 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only o
16/02/04 08:15:56 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:15:56 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/hw3/xac:0+12615536
16/02/04 08:15:56 INFO mapred.MapTask: numReduceTasks: 4
16/02/04 08:15:56 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/02/04 08:15:56 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
```

```
16/02/04 08:15:56 INFO mapred.MapTask: soft limit at 83886080
16/02/04 08:15:56 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/02/04 08:15:56 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/02/04 08:15:56 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$Ma
16/02/04 08:15:56 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./mapper.
16/02/04 08:15:56 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:15:56 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:15:56 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:15:56 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:15:56 INFO streaming.PipeMapRed: Records R/W=1593/1
16/02/04 08:15:57 INFO mapreduce.Job:  map 100% reduce 0%
16/02/04 08:15:57 INFO streaming.PipeMapRed: R/W/S=10000/36510/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:15:59 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:15:59 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:15:59 INFO mapred.LocalJobRunner:
16/02/04 08:15:59 INFO mapred.MapTask: Starting flush of map output
16/02/04 08:15:59 INFO mapred.MapTask: Spilling map output
16/02/04 08:15:59 INFO mapred.MapTask: bufstart = 0; bufend = 3408082; bufvoid = 104857600
16/02/04 08:15:59 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 24883332(99533328); lengt
16/02/04 08:16:00 INFO mapreduce.Job:  map 50% reduce 0%
16/02/04 08:16:00 INFO mapred.MapTask: Finished spill 0
16/02/04 08:16:00 INFO mapred.Task: Task:attempt_local983135479_0001_m_000002_0 is done. And is in the pr
16/02/04 08:16:00 INFO mapred.LocalJobRunner: Records R/W=1593/1
16/02/04 08:16:00 INFO mapred.Task: Task 'attempt_local983135479_0001_m_000002_0' done.
16/02/04 08:16:00 INFO mapred.LocalJobRunner: Finishing task: attempt_local983135479_0001_m_000002_0
16/02/04 08:16:00 INFO mapred.LocalJobRunner: Starting task: attempt_local983135479_0001_m_000003_0
16/02/04 08:16:00 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:16:00 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
16/02/04 08:16:00 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:16:00 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/hw3/xad:0+12006486
16/02/04 08:16:00 INFO mapred.MapTask: numReduceTasks: 4
16/02/04 08:16:00 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/02/04 08:16:00 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/02/04 08:16:00 INFO mapred.MapTask: soft limit at 83886080
16/02/04 08:16:00 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/02/04 08:16:00 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/02/04 08:16:00 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$Ma
16/02/04 08:16:00 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./mapper.
16/02/04 08:16:00 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:00 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:00 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:00 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:00 INFO streaming.PipeMapRed: Records R/W=1639/1
16/02/04 08:16:00 INFO streaming.PipeMapRed: R/W/S=10000/36380/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:01 INFO mapreduce.Job:  map 100% reduce 0%
16/02/04 08:16:02 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:16:02 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:16:02 INFO mapred.LocalJobRunner:
16/02/04 08:16:02 INFO mapred.MapTask: Starting flush of map output
16/02/04 08:16:02 INFO mapred.MapTask: Spilling map output
16/02/04 08:16:02 INFO mapred.MapTask: bufstart = 0; bufend = 3211690; bufvoid = 104857600
16/02/04 08:16:02 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 24965524(99862096); lengt
16/02/04 08:16:03 INFO mapred.MapTask: Finished spill 0
16/02/04 08:16:03 INFO mapred.Task: Task:attempt_local983135479_0001_m_000003_0 is done. And is in the pr
```

```
16/02/04 08:16:03 INFO mapred.LocalJobRunner: Records R/W=1639/1
16/02/04 08:16:03 INFO mapred.Task: Task 'attempt_local983135479_0001_m_000003_0' done.
16/02/04 08:16:03 INFO mapred.LocalJobRunner: Finishing task: attempt_local983135479_0001_m_000003_0
16/02/04 08:16:03 INFO mapred.LocalJobRunner: map task executor complete.
16/02/04 08:16:03 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/02/04 08:16:03 INFO mapred.LocalJobRunner: Starting task: attempt_local983135479_0001_r_000000_0
16/02/04 08:16:03 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:16:03 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only o
16/02/04 08:16:03 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:16:03 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task
16/02/04 08:16:03 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=368836608, maxSingleShuffleLi
16/02/04 08:16:03 INFO reduce.EventFetcher: attempt_local983135479_0001_r_000000_0 Thread started: EventF
16/02/04 08:16:03 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local9
16/02/04 08:16:03 INFO reduce.InMemoryMapOutput: Read 1068325 bytes from map-output for attempt_local983
16/02/04 08:16:03 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 1068325, inMemo
16/02/04 08:16:03 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local9
16/02/04 08:16:03 INFO reduce.InMemoryMapOutput: Read 644986 bytes from map-output for attempt_local9831
16/02/04 08:16:03 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 644986, inMemor
16/02/04 08:16:03 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local9
16/02/04 08:16:03 INFO reduce.InMemoryMapOutput: Read 1104543 bytes from map-output for attempt_local983
16/02/04 08:16:03 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 1104543, inMemo
16/02/04 08:16:03 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local9
16/02/04 08:16:03 INFO reduce.InMemoryMapOutput: Read 772090 bytes from map-output for attempt_local9831
16/02/04 08:16:03 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 772090, inMemor
16/02/04 08:16:03 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/02/04 08:16:03 INFO mapred.LocalJobRunner: 4 / 4 copied.
16/02/04 08:16:03 INFO reduce.MergeManagerImpl: finalMerge called with 4 in-memory map-outputs and 0 on-
16/02/04 08:16:03 INFO mapred.Merger: Merging 4 sorted segments
16/02/04 08:16:03 INFO mapred.Merger: Down to the last merge-pass, with 4 segments left of total size: 3
16/02/04 08:16:03 INFO reduce.MergeManagerImpl: Merged 4 segments, 3589944 bytes to disk to satisfy redu
16/02/04 08:16:03 INFO reduce.MergeManagerImpl: Merging 1 files, 3589942 bytes from disk
16/02/04 08:16:03 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/02/04 08:16:03 INFO mapred.Merger: Merging 1 sorted segments
16/02/04 08:16:03 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 3
16/02/04 08:16:03 INFO mapred.LocalJobRunner: 4 / 4 copied.
16/02/04 08:16:03 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./reducer
16/02/04 08:16:03 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapredu
16/02/04 08:16:03 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce
16/02/04 08:16:03 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:03 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:03 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:03 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:03 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:04 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:05 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:200000=200000/1 [rec/s] out:0=0/1 [rec/
16/02/04 08:16:05 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:300000=300000/1 [rec/s] out:0=0/1 [rec/
16/02/04 08:16:06 INFO streaming.PipeMapRed: Records R/W=346976/1
16/02/04 08:16:06 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:16:06 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:16:06 INFO mapred.Task: Task:attempt_local983135479_0001_r_000000_0 is done. And is in the pr
16/02/04 08:16:06 INFO mapred.LocalJobRunner: 4 / 4 copied.
16/02/04 08:16:06 INFO mapred.Task: Task attempt_local983135479_0001_r_000000_0 is allowed to commit now
16/02/04 08:16:06 INFO output.FileOutputCommitter: Saved output of task 'attempt_local983135479_0001_r_00
16/02/04 08:16:06 INFO mapred.LocalJobRunner: Records R/W=346976/1 > reduce
```

```
16/02/04 08:16:06 INFO mapred.Task: Task 'attempt_local983135479_0001_r_000000_0' done.
16/02/04 08:16:06 INFO mapred.LocalJobRunner: Finishing task: attempt_local983135479_0001_r_000000_0
16/02/04 08:16:06 INFO mapred.LocalJobRunner: Starting task: attempt_local983135479_0001_r_000001_0
16/02/04 08:16:06 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:16:06 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
16/02/04 08:16:06 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:16:06 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task
16/02/04 08:16:06 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=369203616, maxSingleShuffleL
16/02/04 08:16:06 INFO reduce.EventFetcher: attempt_local983135479_0001_r_000001_0 Thread started: EventF
16/02/04 08:16:06 INFO reduce.LocalFetcher: localfetcher#2 about to shuffle output of map attempt_local9
16/02/04 08:16:06 INFO reduce.InMemoryMapOutput: Read 1241234 bytes from map-output for attempt_local983
16/02/04 08:16:06 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 1241234, inMem
16/02/04 08:16:06 INFO reduce.LocalFetcher: localfetcher#2 about to shuffle output of map attempt_local9
16/02/04 08:16:06 INFO reduce.InMemoryMapOutput: Read 1332981 bytes from map-output for attempt_local983
16/02/04 08:16:06 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 1332981, inMem
16/02/04 08:16:06 INFO reduce.LocalFetcher: localfetcher#2 about to shuffle output of map attempt_local9
16/02/04 08:16:06 INFO reduce.InMemoryMapOutput: Read 1243189 bytes from map-output for attempt_local983
16/02/04 08:16:06 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 1243189, inMem
16/02/04 08:16:06 INFO reduce.LocalFetcher: localfetcher#2 about to shuffle output of map attempt_local9
16/02/04 08:16:06 INFO reduce.InMemoryMapOutput: Read 1411496 bytes from map-output for attempt_local983
16/02/04 08:16:06 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 1411496, inMem
16/02/04 08:16:06 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/02/04 08:16:06 INFO mapred.LocalJobRunner: 4 / 4 copied.
16/02/04 08:16:06 INFO reduce.MergeManagerImpl: finalMerge called with 4 in-memory map-outputs and 0 on
16/02/04 08:16:06 INFO mapred.Merger: Merging 4 sorted segments
16/02/04 08:16:06 INFO mapred.Merger: Down to the last merge-pass, with 4 segments left of total size: 5
16/02/04 08:16:06 INFO reduce.MergeManagerImpl: Merged 4 segments, 5228900 bytes to disk to satisfy redu
16/02/04 08:16:06 INFO reduce.MergeManagerImpl: Merging 1 files, 5228898 bytes from disk
16/02/04 08:16:06 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/02/04 08:16:06 INFO mapred.Merger: Merging 1 sorted segments
16/02/04 08:16:06 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 5
16/02/04 08:16:06 INFO mapred.LocalJobRunner: 4 / 4 copied.
16/02/04 08:16:06 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./reducer
16/02/04 08:16:06 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:06 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:06 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:06 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:06 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:07 INFO mapreduce.Job:  map 100% reduce 25%
16/02/04 08:16:07 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:08 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:200000=200000/1 [rec/s] out:0=0/1 [rec
16/02/04 08:16:08 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:150000=300000/2 [rec/s] out:0=0/2 [rec
16/02/04 08:16:09 INFO streaming.PipeMapRed: R/W/S=400000/0/0 in:200000=400000/2 [rec/s] out:0=0/2 [rec
16/02/04 08:16:09 INFO streaming.PipeMapRed: Records R/W=428414/1
16/02/04 08:16:09 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:16:09 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:16:10 INFO mapred.Task: Task:attempt_local983135479_0001_r_000001_0 is done. And is in the pr
16/02/04 08:16:10 INFO mapred.LocalJobRunner: 4 / 4 copied.
16/02/04 08:16:10 INFO mapred.Task: Task attempt_local983135479_0001_r_000001_0 is allowed to commit now
16/02/04 08:16:10 INFO output.FileOutputCommitter: Saved output of task 'attempt_local983135479_0001_r_00
16/02/04 08:16:10 INFO mapred.LocalJobRunner: Records R/W=428414/1 > reduce
16/02/04 08:16:10 INFO mapred.Task: Task 'attempt_local983135479_0001_r_000001_0' done.
16/02/04 08:16:10 INFO mapred.LocalJobRunner: Finishing task: attempt_local983135479_0001_r_000001_0
16/02/04 08:16:10 INFO mapred.LocalJobRunner: Starting task: attempt_local983135479_0001_r_000002_0
```

```
16/02/04 08:16:10 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:16:10 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
16/02/04 08:16:10 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:16:10 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task
16/02/04 08:16:10 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=369570592, maxSingleShuffleL
16/02/04 08:16:10 INFO reduce.EventFetcher: attempt_local983135479_0001_r_000002_0 Thread started: Event
16/02/04 08:16:10 INFO reduce.LocalFetcher: localfetcher#3 about to shuffle output of map attempt_local9
16/02/04 08:16:10 INFO reduce.InMemoryMapOutput: Read 1032138 bytes from map-output for attempt_local983
16/02/04 08:16:10 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 1032138, inMem
16/02/04 08:16:10 INFO reduce.LocalFetcher: localfetcher#3 about to shuffle output of map attempt_local9
16/02/04 08:16:10 INFO reduce.InMemoryMapOutput: Read 1068906 bytes from map-output for attempt_local983
16/02/04 08:16:10 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 1068906, inMem
16/02/04 08:16:10 INFO reduce.LocalFetcher: localfetcher#3 about to shuffle output of map attempt_local9
16/02/04 08:16:10 INFO reduce.InMemoryMapOutput: Read 1011330 bytes from map-output for attempt_local983
16/02/04 08:16:10 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 1011330, inMem
16/02/04 08:16:10 INFO reduce.LocalFetcher: localfetcher#3 about to shuffle output of map attempt_local9
16/02/04 08:16:10 INFO reduce.InMemoryMapOutput: Read 1096104 bytes from map-output for attempt_local983
16/02/04 08:16:10 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 1096104, inMem
16/02/04 08:16:10 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/02/04 08:16:10 INFO mapred.LocalJobRunner: 4 / 4 copied.
16/02/04 08:16:10 INFO reduce.MergeManagerImpl: finalMerge called with 4 in-memory map-outputs and 0 on
16/02/04 08:16:10 INFO mapred.Merger: Merging 4 sorted segments
16/02/04 08:16:10 INFO mapred.Merger: Down to the last merge-pass, with 4 segments left of total size:
16/02/04 08:16:10 INFO reduce.MergeManagerImpl: Merged 4 segments, 4208478 bytes to disk to satisfy redu
16/02/04 08:16:10 INFO reduce.MergeManagerImpl: Merging 1 files, 4208476 bytes from disk
16/02/04 08:16:10 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/02/04 08:16:10 INFO mapred.Merger: Merging 1 sorted segments
16/02/04 08:16:10 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size:
16/02/04 08:16:10 INFO mapred.LocalJobRunner: 4 / 4 copied.
16/02/04 08:16:10 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./reducer
16/02/04 08:16:10 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:10 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:10 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:10 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:10 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:11 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:11 INFO mapreduce.Job:  map 100% reduce 50%
16/02/04 08:16:11 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:200000=200000/1 [rec/s] out:0=0/1 [rec/
16/02/04 08:16:12 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:150000=300000/2 [rec/s] out:0=0/2 [rec/
16/02/04 08:16:13 INFO streaming.PipeMapRed: Records R/W=340472/1
16/02/04 08:16:13 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:16:13 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:16:13 INFO mapred.Task: Task:attempt_local983135479_0001_r_000002_0 is done. And is in the pr
16/02/04 08:16:13 INFO mapred.LocalJobRunner: 4 / 4 copied.
16/02/04 08:16:13 INFO mapred.Task: Task attempt_local983135479_0001_r_000002_0 is allowed to commit now
16/02/04 08:16:13 INFO output.FileOutputCommitter: Saved output of task 'attempt_local983135479_0001_r_00
16/02/04 08:16:13 INFO mapred.LocalJobRunner: Records R/W=340472/1 > reduce
16/02/04 08:16:13 INFO mapred.Task: Task 'attempt_local983135479_0001_r_000002_0' done.
16/02/04 08:16:13 INFO mapred.LocalJobRunner: Finishing task: attempt_local983135479_0001_r_000002_0
16/02/04 08:16:13 INFO mapred.LocalJobRunner: Starting task: attempt_local983135479_0001_r_000003_0
16/02/04 08:16:13 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:16:13 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
16/02/04 08:16:13 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:16:13 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task
```

```
16/02/04 08:16:13 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=369203616, maxSingleShuffleLi
16/02/04 08:16:13 INFO reduce.EventFetcher: attempt_local983135479_0001_r_000003_0 Thread started: EventF
16/02/04 08:16:13 INFO reduce.LocalFetcher: localfetcher#4 about to shuffle output of map attempt_local9
16/02/04 08:16:13 INFO reduce.InMemoryMapOutput: Read 741088 bytes from map-output for attempt_local9831
16/02/04 08:16:13 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 741088, inMemor
16/02/04 08:16:13 INFO reduce.LocalFetcher: localfetcher#4 about to shuffle output of map attempt_local9
16/02/04 08:16:13 INFO reduce.InMemoryMapOutput: Read 789263 bytes from map-output for attempt_local9831
16/02/04 08:16:13 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 789263, inMemor
16/02/04 08:16:13 INFO reduce.LocalFetcher: localfetcher#4 about to shuffle output of map attempt_local9
16/02/04 08:16:13 INFO reduce.InMemoryMapOutput: Read 769768 bytes from map-output for attempt_local9831
16/02/04 08:16:13 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 769768, inMemor
16/02/04 08:16:13 INFO reduce.LocalFetcher: localfetcher#4 about to shuffle output of map attempt_local9
16/02/04 08:16:13 INFO reduce.InMemoryMapOutput: Read 793934 bytes from map-output for attempt_local9831
16/02/04 08:16:13 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 793934, inMemor
16/02/04 08:16:13 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/02/04 08:16:13 INFO mapred.LocalJobRunner: 4 / 4 copied.
16/02/04 08:16:13 INFO reduce.MergeManagerImpl: finalMerge called with 4 in-memory map-outputs and 0 on-
16/02/04 08:16:13 INFO mapred.Merger: Merging 4 sorted segments
16/02/04 08:16:13 INFO mapred.Merger: Down to the last merge-pass, with 4 segments left of total size: 3
16/02/04 08:16:13 INFO mapreduce.Job:  map 100% reduce 75%
16/02/04 08:16:13 INFO reduce.MergeManagerImpl: Merged 4 segments, 3094053 bytes to disk to satisfy redu
16/02/04 08:16:13 INFO reduce.MergeManagerImpl: Merging 1 files, 3094051 bytes from disk
16/02/04 08:16:13 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/02/04 08:16:13 INFO mapred.Merger: Merging 1 sorted segments
16/02/04 08:16:13 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 3
16/02/04 08:16:13 INFO mapred.LocalJobRunner: 4 / 4 copied.
16/02/04 08:16:13 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./reducer
16/02/04 08:16:13 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:13 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:13 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:13 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:13 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:13 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:14 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:200000=200000/1 [rec/s] out:0=0/1 [rec/
16/02/04 08:16:14 INFO streaming.PipeMapRed: Records R/W=232450/1
16/02/04 08:16:14 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:16:14 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:16:14 INFO mapred.Task: Task:attempt_local983135479_0001_r_000003_0 is done. And is in the pr
16/02/04 08:16:14 INFO mapred.LocalJobRunner: 4 / 4 copied.
16/02/04 08:16:14 INFO mapred.Task: Task attempt_local983135479_0001_r_000003_0 is allowed to commit now
16/02/04 08:16:14 INFO output.FileOutputCommitter: Saved output of task 'attempt_local983135479_0001_r_00
16/02/04 08:16:14 INFO mapred.LocalJobRunner: Records R/W=232450/1 > reduce
16/02/04 08:16:14 INFO mapred.Task: Task 'attempt_local983135479_0001_r_000003_0' done.
16/02/04 08:16:14 INFO mapred.LocalJobRunner: Finishing task: attempt_local983135479_0001_r_000003_0
16/02/04 08:16:14 INFO mapred.LocalJobRunner: reduce task executor complete.
16/02/04 08:16:15 INFO mapreduce.Job:  map 100% reduce 100%
16/02/04 08:16:15 INFO mapreduce.Job: Job job_local983135479_0001 completed successfully
16/02/04 08:16:15 INFO mapreduce.Job: Counters: 37
        File System Counters
                FILE: Number of bytes read=84122588
                FILE: Number of bytes written=149829219
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=332876621
                HDFS: Number of bytes written=6302
                HDFS: Number of read operations=130
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=28
        Map-Reduce Framework
                Map input records=312912
                Map output records=1348312
                Map output bytes=13424719
                Map output materialized bytes=16121439
                Input split bytes=344
                Combine input records=0
                Combine output records=0
                Reduce input groups=188
                Reduce shuffle bytes=16121439
                Reduce input records=1348312
                Reduce output records=188
                Spilled Records=2696624
                Shuffled Maps =16
                Failed Shuffles=0
                Merged Map outputs=16
                GC time elapsed (ms)=773
                Total committed heap usage (bytes)=3833069568
        MAPPER
                mapper_calls=312912
        REDUCER
                reducer_calls=1348312
        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=50906312
        File Output Format Counters
                Bytes Written=2526
16/02/04 08:16:15 INFO streaming.StreamJob: Output directory: /user/hw3/output_3_2_2
```

** The next four cells output the first 10 rows of each of the 4 reducer outputs **

```
In [197]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -cat /user/hw3/output_3_2_2/part-00000 | hea
```

```
16/02/04 08:16:17 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
Account 16555
Applied 139
Can't 1999
Cash 240
Closing 2795
Cont'd 17972
Debt 1343
Delinquent 1061
I 925
Incorrect 29133
```

```
In [198]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -cat /user/hw3/output_3_2_2/part-00001 | hea
```

```
16/02/04 08:16:20 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
1 4
ATM 2422
Cancelling 2795
Communication 8671
Dealing 1944
Improper 4966
Loan 107254
Payment 92
Problems 9484
Shopping 672
```

```
In [199]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -cat /user/hw3/output_3_2_2/part-00002 | hea
```

```
16/02/04 08:16:23 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
APR 3431
Arbitration 168
Bankruptcy 222
Billing 8158
Convenience 75
Credit 14768
Deposits 10555
Disclosure 7655
False 3621
Overlimit 127
```

```
In [200]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -cat /user/hw3/output_3_2_2/part-00003 | hea
```

```
16/02/04 08:16:27 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
Advertising 1193
Application 8868
Balance 597
Charged 878
Collection 1907
Customer 2734
Embezzlement 3276
Forbearance 350
Fraud 3842
Getting 291
```

0.0.9 3.2.3

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.

```
In [201]: %%writefile mapper.py
          #!/usr/bin/python
          import sys
          import re
          import csv

          csv_iter = csv.reader(sys.stdin)
          for line in csv_iter:
              sys.stderr.write("reporter:counter:MAPPER,mapper_calls,1\n")
              if(len(line)==14):   #continue only if all the rows have all the entries
                  issues = re.split(r'[\s,./]+',line[3])
                  for w in issues:
                      print "%s\t1" %w
```

```
Overwriting mapper.py
```

```
In [202]: !chmod a+x mapper.py
```

```
In [203]: %%writefile combiner.py
          #!/usr/bin/python
          import sys
          import re

          count = 0

          prev_string = None

          for line in sys.stdin:
              sys.stderr.write("reporter:counter:COMBINER,combiner_calls,1\n")
              line_s = re.split(r'[\t]',line.strip())

              if((prev_string!=None) and (prev_string != line_s[0])):
                  print "%s\t%s" %(prev_string,count)
                  count = 0
              count += 1
              prev_string = line_s[0]
          print "%s\t%s" %(prev_string,count)
```

```
Overwriting combiner.py
```

```
In [204]: !chmod a+x combiner.py
```

```
In [205]: %%writefile reducer.py
          #!/usr/bin/python
          import sys
          import re
          import csv

          count = 0

          prev_string = None

          for line in sys.stdin:
              sys.stderr.write("reporter:counter:REDUCER,reducer_calls,1\n")

              line_s = re.split(r'[\t]',line.strip())

              if((prev_string!=None) and (prev_string != line_s[0])):
                  print prev_string,count
                  count = 0
              count += int(line_s[1])
              prev_string = line_s[0]
          print prev_string,count
```

```
Overwriting reducer.py
```

```
In [206]: !chmod a+x reducer.py
```

```
In [207]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -rm -r -f /user/hw3/output_3_2_3
```

```
16/02/04 08:16:30 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
```

```
In [208]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hadoop jar /Users/Vamsi/Downloads/hadoop-2.7.1/bin/ha
          -D mapred.reduce.tasks=4 \
          -input /user/hw3/xa* \
          -output /user/hw3/output_3_2_3 \
          -mapper mapper.py \
          -combiner combiner.py \
          -reducer reducer.py
```

```
16/02/04 08:16:33 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
16/02/04 08:16:33 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.ses
16/02/04 08:16:33 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/02/04 08:16:33 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessio
16/02/04 08:16:34 INFO mapred.FileInputFormat: Total input paths to process : 4
16/02/04 08:16:34 INFO mapreduce.JobSubmitter: number of splits:4
16/02/04 08:16:34 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapred
16/02/04 08:16:34 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local2063302883_0001
16/02/04 08:16:34 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/02/04 08:16:34 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/02/04 08:16:34 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputComm
16/02/04 08:16:34 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:16:34 INFO mapreduce.Job: Running job: job_local2063302883_0001
16/02/04 08:16:34 INFO mapred.LocalJobRunner: Waiting for map tasks
16/02/04 08:16:34 INFO mapred.LocalJobRunner: Starting task: attempt_local2063302883_0001_m_000000_0
16/02/04 08:16:34 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:16:34 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only o
16/02/04 08:16:34 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:16:34 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/hw3/xab:0+13160945
16/02/04 08:16:34 INFO mapred.MapTask: numReduceTasks: 4
16/02/04 08:16:35 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/02/04 08:16:35 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/02/04 08:16:35 INFO mapred.MapTask: soft limit at 83886080
16/02/04 08:16:35 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/02/04 08:16:35 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/02/04 08:16:35 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$Map
16/02/04 08:16:35 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./mapper.p
16/02/04 08:16:35 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.ta
16/02/04 08:16:35 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce
16/02/04 08:16:35 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.m
16/02/04 08:16:35 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.j
16/02/04 08:16:35 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce
16/02/04 08:16:35 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use map
16/02/04 08:16:35 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce
16/02/04 08:16:35 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.jo
16/02/04 08:16:35 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.us
16/02/04 08:16:35 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapredu
16/02/04 08:16:35 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.t
16/02/04 08:16:35 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use map
```

```
16/02/04 08:16:35 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:35 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:35 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:35 INFO streaming.PipeMapRed: Records R/W=773/1
16/02/04 08:16:35 INFO streaming.PipeMapRed: R/W/S=1000/938/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:35 INFO streaming.PipeMapRed: R/W/S=10000/38814/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:35 INFO mapreduce.Job: Job job_local2063302883_0001 running in uber mode : false
16/02/04 08:16:35 INFO mapreduce.Job:  map 0% reduce 0%
16/02/04 08:16:37 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:16:37 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:16:37 INFO mapred.LocalJobRunner:
16/02/04 08:16:37 INFO mapred.MapTask: Starting flush of map output
16/02/04 08:16:37 INFO mapred.MapTask: Spilling map output
16/02/04 08:16:37 INFO mapred.MapTask: bufstart = 0; bufend = 3385621; bufvoid = 104857600
16/02/04 08:16:37 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 24820088(99280352); lengt
16/02/04 08:16:37 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./combine
16/02/04 08:16:37 INFO Configuration.deprecation: mapred.skip.map.auto.incr.proc.count is deprecated. I
16/02/04 08:16:37 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:37 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:37 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:37 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:37 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:38 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:38 INFO streaming.PipeMapRed: Records R/W=104547/1
16/02/04 08:16:38 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:16:38 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:16:38 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./combine
16/02/04 08:16:38 INFO Configuration.deprecation: mapred.skip.reduce.auto.incr.proc.count is deprecated
16/02/04 08:16:38 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:38 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:38 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:38 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:38 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:38 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:39 INFO streaming.PipeMapRed: Records R/W=101901/1
16/02/04 08:16:39 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:16:39 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:16:39 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./combine
16/02/04 08:16:39 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:39 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:39 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:39 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:39 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:39 INFO streaming.PipeMapRed: Records R/W=85831/1
16/02/04 08:16:39 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:16:39 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:16:39 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./combine
16/02/04 08:16:39 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:39 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:39 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:39 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:39 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:40 INFO streaming.PipeMapRed: Records R/W=56299/1
16/02/04 08:16:40 INFO streaming.PipeMapRed: MRErrorThread done
```

```
16/02/04 08:16:40 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:16:40 INFO mapred.MapTask: Finished spill 0
16/02/04 08:16:40 INFO mapred.Task: Task:attempt_local2063302883_0001_m_000000_0 is done. And is in the p
16/02/04 08:16:40 INFO mapred.LocalJobRunner: Records R/W=56299/1
16/02/04 08:16:40 INFO mapred.Task: Task 'attempt_local2063302883_0001_m_000000_0' done.
16/02/04 08:16:40 INFO mapred.LocalJobRunner: Finishing task: attempt_local2063302883_0001_m_000000_0
16/02/04 08:16:40 INFO mapred.LocalJobRunner: Starting task: attempt_local2063302883_0001_m_000001_0
16/02/04 08:16:40 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:16:40 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
16/02/04 08:16:40 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:16:40 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/hw3/xaa:0+13123345
16/02/04 08:16:40 INFO mapred.MapTask: numReduceTasks: 4
16/02/04 08:16:40 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/02/04 08:16:40 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/02/04 08:16:40 INFO mapred.MapTask: soft limit at 83886080
16/02/04 08:16:40 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/02/04 08:16:40 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/02/04 08:16:40 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$Ma
16/02/04 08:16:40 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./mapper.
16/02/04 08:16:40 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:40 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:40 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:40 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:40 INFO streaming.PipeMapRed: Records R/W=1571/1
16/02/04 08:16:40 INFO streaming.PipeMapRed: R/W/S=10000/39416/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:40 INFO mapreduce.Job:  map 100% reduce 0%
16/02/04 08:16:42 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:16:42 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:16:42 INFO mapred.LocalJobRunner:
16/02/04 08:16:42 INFO mapred.MapTask: Starting flush of map output
16/02/04 08:16:42 INFO mapred.MapTask: Spilling map output
16/02/04 08:16:42 INFO mapred.MapTask: bufstart = 0; bufend = 3419326; bufvoid = 104857600
16/02/04 08:16:42 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 24795408(99181632); lengt
16/02/04 08:16:42 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./combine
16/02/04 08:16:42 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:42 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:42 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:42 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:42 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:42 INFO mapreduce.Job:  map 25% reduce 0%
16/02/04 08:16:43 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:43 INFO streaming.PipeMapRed: Records R/W=108893/1
16/02/04 08:16:43 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:16:43 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:16:43 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./combine
16/02/04 08:16:43 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:43 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:43 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:43 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:43 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:44 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:44 INFO streaming.PipeMapRed: Records R/W=103027/1
16/02/04 08:16:44 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:16:44 INFO streaming.PipeMapRed: mapRedFinished
```

```
16/02/04 08:16:44 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./combine
16/02/04 08:16:44 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:44 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:44 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:44 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:44 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:44 INFO streaming.PipeMapRed: Records R/W=84622/1
16/02/04 08:16:44 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:16:44 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:16:44 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./combine
16/02/04 08:16:44 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:44 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:44 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:44 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:44 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:45 INFO streaming.PipeMapRed: Records R/W=58206/1
16/02/04 08:16:45 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:16:45 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:16:45 INFO mapred.MapTask: Finished spill 0
16/02/04 08:16:45 INFO mapred.Task: Task:attempt_local2063302883_0001_m_000001_0 is done. And is in the p
16/02/04 08:16:45 INFO mapred.LocalJobRunner: Records R/W=58206/1
16/02/04 08:16:45 INFO mapred.Task: Task 'attempt_local2063302883_0001_m_000001_0' done.
16/02/04 08:16:45 INFO mapred.LocalJobRunner: Finishing task: attempt_local2063302883_0001_m_000001_0
16/02/04 08:16:45 INFO mapred.LocalJobRunner: Starting task: attempt_local2063302883_0001_m_000002_0
16/02/04 08:16:45 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:16:45 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only 
16/02/04 08:16:45 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:16:45 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/hw3/xac:0+12615536
16/02/04 08:16:45 INFO mapred.MapTask: numReduceTasks: 4
16/02/04 08:16:45 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/02/04 08:16:45 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/02/04 08:16:45 INFO mapred.MapTask: soft limit at 83886080
16/02/04 08:16:45 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/02/04 08:16:45 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/02/04 08:16:45 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$Ma
16/02/04 08:16:45 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./mapper.
16/02/04 08:16:45 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:45 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:45 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:45 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:45 INFO streaming.PipeMapRed: Records R/W=1593/1
16/02/04 08:16:45 INFO streaming.PipeMapRed: R/W/S=10000/36510/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:45 INFO mapreduce.Job:  map 100% reduce 0%
16/02/04 08:16:47 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:16:47 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:16:47 INFO mapred.LocalJobRunner:
16/02/04 08:16:47 INFO mapred.MapTask: Starting flush of map output
16/02/04 08:16:47 INFO mapred.MapTask: Spilling map output
16/02/04 08:16:47 INFO mapred.MapTask: bufstart = 0; bufend = 3408082; bufvoid = 104857600
16/02/04 08:16:47 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 24883332(99533328); lengt
16/02/04 08:16:47 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./combine
16/02/04 08:16:47 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:47 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:47 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
```

```
16/02/04 08:16:47 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:47 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:47 INFO mapreduce.Job:  map 50% reduce 0%
16/02/04 08:16:48 INFO streaming.PipeMapRed: Records R/W=73207/1
16/02/04 08:16:48 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:16:48 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:16:48 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./combine
16/02/04 08:16:48 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:48 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:48 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:48 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:48 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:48 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:49 INFO streaming.PipeMapRed: Records R/W=114751/1
16/02/04 08:16:49 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:16:49 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:16:49 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./combine
16/02/04 08:16:49 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:49 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:49 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:49 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:49 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:49 INFO streaming.PipeMapRed: Records R/W=86427/1
16/02/04 08:16:49 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:16:49 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:16:49 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./combine
16/02/04 08:16:49 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:49 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:49 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:49 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:49 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:50 INFO streaming.PipeMapRed: Records R/W=58382/1
16/02/04 08:16:50 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:16:50 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:16:50 INFO mapred.MapTask: Finished spill 0
16/02/04 08:16:50 INFO mapred.Task: Task:attempt_local2063302883_0001_m_000002_0 is done. And is in the p
16/02/04 08:16:50 INFO mapred.LocalJobRunner: Records R/W=58382/1
16/02/04 08:16:50 INFO mapred.Task: Task 'attempt_local2063302883_0001_m_000002_0' done.
16/02/04 08:16:50 INFO mapred.LocalJobRunner: Finishing task: attempt_local2063302883_0001_m_000002_0
16/02/04 08:16:50 INFO mapred.LocalJobRunner: Starting task: attempt_local2063302883_0001_m_000003_0
16/02/04 08:16:50 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:16:50 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
16/02/04 08:16:50 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:16:50 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/hw3/xad:0+12006486
16/02/04 08:16:50 INFO mapred.MapTask: numReduceTasks: 4
16/02/04 08:16:50 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/02/04 08:16:50 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/02/04 08:16:50 INFO mapred.MapTask: soft limit at 83886080
16/02/04 08:16:50 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/02/04 08:16:50 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/02/04 08:16:50 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$Ma
16/02/04 08:16:50 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./mapper.
16/02/04 08:16:50 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:50 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
```

```
16/02/04 08:16:50 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:50 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:50 INFO streaming.PipeMapRed: Records R/W=1639/1
16/02/04 08:16:50 INFO streaming.PipeMapRed: R/W/S=10000/36380/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:50 INFO mapreduce.Job:  map 100% reduce 0%
16/02/04 08:16:52 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:16:52 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:16:52 INFO mapred.LocalJobRunner:
16/02/04 08:16:52 INFO mapred.MapTask: Starting flush of map output
16/02/04 08:16:52 INFO mapred.MapTask: Spilling map output
16/02/04 08:16:52 INFO mapred.MapTask: bufstart = 0; bufend = 3211690; bufvoid = 104857600
16/02/04 08:16:52 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 24965524(99862096); lengt
16/02/04 08:16:52 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./combine
16/02/04 08:16:52 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:52 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:52 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:52 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:52 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:52 INFO streaming.PipeMapRed: Records R/W=60329/1
16/02/04 08:16:52 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:16:52 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:16:52 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./combine
16/02/04 08:16:52 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:52 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:52 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:52 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:52 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:52 INFO mapreduce.Job:  map 75% reduce 0%
16/02/04 08:16:53 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:53 INFO streaming.PipeMapRed: Records R/W=108735/1
16/02/04 08:16:53 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:16:53 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:16:53 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./combine
16/02/04 08:16:53 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:53 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:53 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:53 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:53 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:54 INFO streaming.PipeMapRed: Records R/W=83592/1
16/02/04 08:16:54 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:16:54 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:16:54 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./combine
16/02/04 08:16:54 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:54 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:54 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:54 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:54 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:55 INFO streaming.PipeMapRed: Records R/W=59563/1
16/02/04 08:16:55 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:16:55 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:16:55 INFO mapred.MapTask: Finished spill 0
16/02/04 08:16:55 INFO mapred.Task: Task:attempt_local2063302883_0001_m_000003_0 is done. And is in the p
16/02/04 08:16:55 INFO mapred.LocalJobRunner: Records R/W=59563/1
16/02/04 08:16:55 INFO mapred.Task: Task 'attempt_local2063302883_0001_m_000003_0' done.
```

```
16/02/04 08:16:55 INFO mapred.LocalJobRunner: Finishing task: attempt_local2063302883_0001_m_000003_0
16/02/04 08:16:55 INFO mapred.LocalJobRunner: map task executor complete.
16/02/04 08:16:55 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/02/04 08:16:55 INFO mapred.LocalJobRunner: Starting task: attempt_local2063302883_0001_r_000000_0
16/02/04 08:16:55 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:16:55 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only o
16/02/04 08:16:55 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:16:55 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task
16/02/04 08:16:55 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLi
16/02/04 08:16:55 INFO reduce.EventFetcher: attempt_local2063302883_0001_r_000000_0 Thread started: Event
16/02/04 08:16:55 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local2
16/02/04 08:16:55 INFO reduce.InMemoryMapOutput: Read 447 bytes from map-output for attempt_local2063302
16/02/04 08:16:55 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 447, inMemoryMa
16/02/04 08:16:55 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local2
16/02/04 08:16:55 INFO reduce.InMemoryMapOutput: Read 591 bytes from map-output for attempt_local2063302
16/02/04 08:16:55 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 591, inMemoryMa
16/02/04 08:16:55 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local2
16/02/04 08:16:55 INFO reduce.InMemoryMapOutput: Read 596 bytes from map-output for attempt_local2063302
16/02/04 08:16:55 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 596, inMemoryMa
16/02/04 08:16:55 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local2
16/02/04 08:16:55 INFO reduce.InMemoryMapOutput: Read 526 bytes from map-output for attempt_local2063302
16/02/04 08:16:55 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 526, inMemoryMa
16/02/04 08:16:55 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/02/04 08:16:55 INFO mapred.LocalJobRunner: 4 / 4 copied.
16/02/04 08:16:55 INFO reduce.MergeManagerImpl: finalMerge called with 4 in-memory map-outputs and 0 on-
16/02/04 08:16:55 INFO mapred.Merger: Merging 4 sorted segments
16/02/04 08:16:55 INFO mapred.Merger: Down to the last merge-pass, with 4 segments left of total size: 2
16/02/04 08:16:55 INFO reduce.MergeManagerImpl: Merged 4 segments, 2160 bytes to disk to satisfy reduce
16/02/04 08:16:55 INFO reduce.MergeManagerImpl: Merging 1 files, 2158 bytes from disk
16/02/04 08:16:55 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/02/04 08:16:55 INFO mapred.Merger: Merging 1 sorted segments
16/02/04 08:16:55 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 2
16/02/04 08:16:55 INFO mapred.LocalJobRunner: 4 / 4 copied.
16/02/04 08:16:55 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./reducer
16/02/04 08:16:55 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapredu
16/02/04 08:16:55 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce
16/02/04 08:16:55 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:55 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:55 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:55 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:16:55 INFO streaming.PipeMapRed: Records R/W=166/1
16/02/04 08:16:55 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:16:55 INFO mapred.Task: Task:attempt_local2063302883_0001_r_000000_0 is done. And is in the p
16/02/04 08:16:55 INFO mapred.LocalJobRunner: 4 / 4 copied.
16/02/04 08:16:55 INFO mapred.Task: Task attempt_local2063302883_0001_r_000000_0 is allowed to commit now
16/02/04 08:16:55 INFO output.FileOutputCommitter: Saved output of task 'attempt_local2063302883_0001_r_0
16/02/04 08:16:55 INFO mapred.LocalJobRunner: Records R/W=166/1 > reduce
16/02/04 08:16:55 INFO mapred.Task: Task 'attempt_local2063302883_0001_r_000000_0' done.
16/02/04 08:16:55 INFO mapred.LocalJobRunner: Finishing task: attempt_local2063302883_0001_r_000000_0
16/02/04 08:16:55 INFO mapred.LocalJobRunner: Starting task: attempt_local2063302883_0001_r_000001_0
16/02/04 08:16:55 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:16:55 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only o
16/02/04 08:16:55 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:16:55 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task
```

```
16/02/04 08:16:55 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLi
16/02/04 08:16:55 INFO reduce.EventFetcher: attempt_local2063302883_0001_r_000001_0 Thread started: Event
16/02/04 08:16:55 INFO reduce.LocalFetcher: localfetcher#2 about to shuffle output of map attempt_local2
16/02/04 08:16:55 INFO reduce.InMemoryMapOutput: Read 508 bytes from map-output for attempt_local2063302
16/02/04 08:16:55 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 508, inMemoryMa
16/02/04 08:16:55 INFO reduce.LocalFetcher: localfetcher#2 about to shuffle output of map attempt_local2
16/02/04 08:16:55 INFO reduce.InMemoryMapOutput: Read 734 bytes from map-output for attempt_local2063302
16/02/04 08:16:55 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 734, inMemoryMa
16/02/04 08:16:55 INFO reduce.LocalFetcher: localfetcher#2 about to shuffle output of map attempt_local2
16/02/04 08:16:55 INFO reduce.InMemoryMapOutput: Read 744 bytes from map-output for attempt_local2063302
16/02/04 08:16:55 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 744, inMemoryMa
16/02/04 08:16:55 INFO reduce.LocalFetcher: localfetcher#2 about to shuffle output of map attempt_local2
16/02/04 08:16:55 INFO reduce.InMemoryMapOutput: Read 642 bytes from map-output for attempt_local2063302
16/02/04 08:16:55 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 642, inMemoryMa
16/02/04 08:16:55 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/02/04 08:16:55 INFO mapred.LocalJobRunner: 4 / 4 copied.
16/02/04 08:16:55 INFO reduce.MergeManagerImpl: finalMerge called with 4 in-memory map-outputs and 0 on-
16/02/04 08:16:55 INFO mapred.Merger: Merging 4 sorted segments
16/02/04 08:16:55 INFO mapred.Merger: Down to the last merge-pass, with 4 segments left of total size: 2
16/02/04 08:16:55 INFO reduce.MergeManagerImpl: Merged 4 segments, 2628 bytes to disk to satisfy reduce
16/02/04 08:16:55 INFO reduce.MergeManagerImpl: Merging 1 files, 2626 bytes from disk
16/02/04 08:16:55 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/02/04 08:16:55 INFO mapred.Merger: Merging 1 sorted segments
16/02/04 08:16:55 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 2
16/02/04 08:16:55 INFO mapred.LocalJobRunner: 4 / 4 copied.
16/02/04 08:16:55 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./reducer
16/02/04 08:16:55 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:55 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:55 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:55 INFO streaming.PipeMapRed: Records R/W=182/1
16/02/04 08:16:55 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:16:55 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:16:55 INFO mapred.Task: Task:attempt_local2063302883_0001_r_000001_0 is done. And is in the p
16/02/04 08:16:55 INFO mapred.LocalJobRunner: 4 / 4 copied.
16/02/04 08:16:55 INFO mapred.Task: Task attempt_local2063302883_0001_r_000001_0 is allowed to commit now
16/02/04 08:16:55 INFO output.FileOutputCommitter: Saved output of task ’attempt_local2063302883_0001_r_0
16/02/04 08:16:55 INFO mapred.LocalJobRunner: Records R/W=182/1 > reduce
16/02/04 08:16:55 INFO mapred.Task: Task ’attempt_local2063302883_0001_r_000001_0’ done.
16/02/04 08:16:55 INFO mapred.LocalJobRunner: Finishing task: attempt_local2063302883_0001_r_000001_0
16/02/04 08:16:55 INFO mapred.LocalJobRunner: Starting task: attempt_local2063302883_0001_r_000002_0
16/02/04 08:16:55 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:16:55 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only o
16/02/04 08:16:55 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:16:55 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task
16/02/04 08:16:55 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLi
16/02/04 08:16:55 INFO reduce.EventFetcher: attempt_local2063302883_0001_r_000002_0 Thread started: Event
16/02/04 08:16:55 INFO reduce.LocalFetcher: localfetcher#3 about to shuffle output of map attempt_local2
16/02/04 08:16:55 INFO reduce.InMemoryMapOutput: Read 435 bytes from map-output for attempt_local2063302
16/02/04 08:16:55 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 435, inMemoryMa
16/02/04 08:16:55 INFO reduce.LocalFetcher: localfetcher#3 about to shuffle output of map attempt_local2
16/02/04 08:16:55 INFO reduce.InMemoryMapOutput: Read 650 bytes from map-output for attempt_local2063302
16/02/04 08:16:55 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 650, inMemoryMa
16/02/04 08:16:55 INFO reduce.LocalFetcher: localfetcher#3 about to shuffle output of map attempt_local2
16/02/04 08:16:55 INFO reduce.InMemoryMapOutput: Read 649 bytes from map-output for attempt_local2063302
```

```
16/02/04 08:16:55 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 649, inMemoryMa
16/02/04 08:16:55 INFO reduce.LocalFetcher: localfetcher#3 about to shuffle output of map attempt_local2
16/02/04 08:16:55 INFO reduce.InMemoryMapOutput: Read 562 bytes from map-output for attempt_local2063302
16/02/04 08:16:55 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 562, inMemoryMa
16/02/04 08:16:55 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/02/04 08:16:55 INFO mapred.LocalJobRunner: 4 / 4 copied.
16/02/04 08:16:55 INFO reduce.MergeManagerImpl: finalMerge called with 4 in-memory map-outputs and 0 on-
16/02/04 08:16:55 INFO mapred.Merger: Merging 4 sorted segments
16/02/04 08:16:55 INFO mapred.Merger: Down to the last merge-pass, with 4 segments left of total size: 2
16/02/04 08:16:55 INFO reduce.MergeManagerImpl: Merged 4 segments, 2296 bytes to disk to satisfy reduce
16/02/04 08:16:55 INFO reduce.MergeManagerImpl: Merging 1 files, 2294 bytes from disk
16/02/04 08:16:55 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/02/04 08:16:55 INFO mapred.Merger: Merging 1 sorted segments
16/02/04 08:16:55 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 2
16/02/04 08:16:55 INFO mapred.LocalJobRunner: 4 / 4 copied.
16/02/04 08:16:55 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./reducer
16/02/04 08:16:55 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:55 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:55 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:55 INFO streaming.PipeMapRed: Records R/W=157/1
16/02/04 08:16:55 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:16:55 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:16:55 INFO mapred.Task: Task:attempt_local2063302883_0001_r_000002_0 is done. And is in the p
16/02/04 08:16:55 INFO mapred.LocalJobRunner: 4 / 4 copied.
16/02/04 08:16:55 INFO mapred.Task: Task attempt_local2063302883_0001_r_000002_0 is allowed to commit now
16/02/04 08:16:55 INFO output.FileOutputCommitter: Saved output of task 'attempt_local2063302883_0001_r_0
16/02/04 08:16:55 INFO mapred.LocalJobRunner: Records R/W=157/1 > reduce
16/02/04 08:16:55 INFO mapred.Task: Task 'attempt_local2063302883_0001_r_000002_0' done.
16/02/04 08:16:55 INFO mapred.LocalJobRunner: Finishing task: attempt_local2063302883_0001_r_000002_0
16/02/04 08:16:55 INFO mapred.LocalJobRunner: Starting task: attempt_local2063302883_0001_r_000003_0
16/02/04 08:16:55 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:16:55 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only o
16/02/04 08:16:55 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:16:55 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task
16/02/04 08:16:55 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLi
16/02/04 08:16:55 INFO reduce.EventFetcher: attempt_local2063302883_0001_r_000003_0 Thread started: Event
16/02/04 08:16:55 INFO reduce.LocalFetcher: localfetcher#4 about to shuffle output of map attempt_local2
16/02/04 08:16:55 INFO reduce.InMemoryMapOutput: Read 476 bytes from map-output for attempt_local2063302
16/02/04 08:16:55 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 476, inMemoryMa
16/02/04 08:16:55 INFO reduce.LocalFetcher: localfetcher#4 about to shuffle output of map attempt_local2
16/02/04 08:16:55 INFO reduce.InMemoryMapOutput: Read 579 bytes from map-output for attempt_local2063302
16/02/04 08:16:55 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 579, inMemoryMa
16/02/04 08:16:55 INFO reduce.LocalFetcher: localfetcher#4 about to shuffle output of map attempt_local2
16/02/04 08:16:55 INFO reduce.InMemoryMapOutput: Read 583 bytes from map-output for attempt_local2063302
16/02/04 08:16:55 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 583, inMemoryMa
16/02/04 08:16:55 INFO reduce.LocalFetcher: localfetcher#4 about to shuffle output of map attempt_local2
16/02/04 08:16:55 INFO reduce.InMemoryMapOutput: Read 573 bytes from map-output for attempt_local2063302
16/02/04 08:16:55 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 573, inMemoryMa
16/02/04 08:16:55 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/02/04 08:16:55 INFO mapred.LocalJobRunner: 4 / 4 copied.
16/02/04 08:16:55 INFO reduce.MergeManagerImpl: finalMerge called with 4 in-memory map-outputs and 0 on-
16/02/04 08:16:55 INFO mapred.Merger: Merging 4 sorted segments
16/02/04 08:16:55 INFO mapred.Merger: Down to the last merge-pass, with 4 segments left of total size: 2
16/02/04 08:16:55 INFO reduce.MergeManagerImpl: Merged 4 segments, 2211 bytes to disk to satisfy reduce
```

```
16/02/04 08:16:55 INFO reduce.MergeManagerImpl: Merging 1 files, 2209 bytes from disk
16/02/04 08:16:55 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/02/04 08:16:55 INFO mapred.Merger: Merging 1 sorted segments
16/02/04 08:16:55 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 
16/02/04 08:16:55 INFO mapred.LocalJobRunner: 4 / 4 copied.
16/02/04 08:16:55 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./reducer
16/02/04 08:16:55 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:55 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:55 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:16:55 INFO streaming.PipeMapRed: Records R/W=156/1
16/02/04 08:16:55 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:16:55 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:16:55 INFO mapred.Task: Task:attempt_local2063302883_0001_r_000003_0 is done. And is in the p
16/02/04 08:16:55 INFO mapred.LocalJobRunner: 4 / 4 copied.
16/02/04 08:16:55 INFO mapred.Task: Task attempt_local2063302883_0001_r_000003_0 is allowed to commit now
16/02/04 08:16:55 INFO output.FileOutputCommitter: Saved output of task 'attempt_local2063302883_0001_r_0
16/02/04 08:16:55 INFO mapred.LocalJobRunner: Records R/W=156/1 > reduce
16/02/04 08:16:55 INFO mapred.Task: Task 'attempt_local2063302883_0001_r_000003_0' done.
16/02/04 08:16:55 INFO mapred.LocalJobRunner: Finishing task: attempt_local2063302883_0001_r_000003_0
16/02/04 08:16:55 INFO mapred.LocalJobRunner: reduce task executor complete.
16/02/04 08:16:55 INFO mapreduce.Job:  map 100% reduce 100%
16/02/04 08:16:56 INFO mapreduce.Job: Job job_local2063302883_0001 completed successfully
16/02/04 08:16:56 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=953763
                FILE: Number of bytes written=3188275
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=332876621
                HDFS: Number of bytes written=6302
                HDFS: Number of read operations=130
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=28
        Map-Reduce Framework
                Map input records=312912
                Map output records=1348312
                Map output bytes=13424719
                Map output materialized bytes=9359
                Input split bytes=344
                Combine input records=1348312
                Combine output records=661
                Reduce input groups=188
                Reduce shuffle bytes=9359
                Reduce input records=661
                Reduce output records=188
                Spilled Records=1322
                Shuffled Maps =16
                Failed Shuffles=0
                Merged Map outputs=16
                GC time elapsed (ms)=245
                Total committed heap usage (bytes)=3197632512
        COMBINER
                combiner_calls=1348312
        MAPPER
                mapper_calls=312912
        REDUCER
                reducer_calls=661
        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=50906312
        File Output Format Counters
                Bytes Written=2526
16/02/04 08:16:56 INFO streaming.StreamJob: Output directory: /user/hw3/output_3_2_3
```

The power of combiners is evident above. The reducer_calls reduces to 661!

0.0.10 3.2.4

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. If there are ties please sort the tokens in alphanumeric/string order. Present bottom 10 tokens (least frequent items).

```
In [209]: %%writefile mapper.py
          #!/usr/bin/python
          import sys
          import re
          import csv

          csv_iter = csv.reader(sys.stdin)

          for line in csv_iter:
              if(len(line)==14):   #continue only if all the rows have all the entries
                  issues = re.split(r'[\s,./]+',line[3])
                  for w in issues:
                      print "%s,1" %w
                      print "*,1"
```

```
Overwriting mapper.py
```

```
In [210]: !chmod a+x mapper.py
```

```
In [211]: %%writefile combiner.py
          #!/usr/bin/python
          import sys
          import re

          count = 0

          prev_string = None

          for line in sys.stdin:
              line_s = re.split(r'[,]',line.strip())
```

```
                  if((prev_string!=None) and (prev_string != line_s[0])):
                      print "%s,%s" %(prev_string,count)
                      count = 0
                  count += 1
                  prev_string = line_s[0]
          print "%s,%s" %(prev_string,count)
```

```
Overwriting combiner.py
```

```
In [212]: !chmod a+x combiner.py
```

```
In [213]: %%writefile reducer.py
          #!/usr/bin/python
          import sys
          import re
          import csv

          count = 0

          prev_string = None

          for line in sys.stdin:
              line_s = re.split(r'[,]',line.strip())

              if((prev_string!=None) and (prev_string != line_s[0])):
                  print prev_string,count
                  count = 0
              count += int(line_s[1])
              prev_string = line_s[0]
          print prev_string,count
```

```
Overwriting reducer.py
```

```
In [214]: !chmod a+x reducer.py
```

```
In [215]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -rm -r -f /user/hw3/output_3_2_4
```

```
16/02/04 08:16:59 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
```

```
In [216]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hadoop jar /Users/Vamsi/Downloads/hadoop-2.7.1/bin/ha
          -D mapred.reduce.tasks=1 \
          -input /user/hw3/xa* \
          -output /user/hw3/output_3_2_4 \
          -mapper mapper.py \
          -combiner combiner.py \
          -reducer reducer.py
```

```
16/02/04 08:17:01 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
16/02/04 08:17:02 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.ses
16/02/04 08:17:02 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/02/04 08:17:02 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessio
16/02/04 08:17:02 INFO mapred.FileInputFormat: Total input paths to process : 4
16/02/04 08:17:03 INFO mapreduce.JobSubmitter: number of splits:4
16/02/04 08:17:03 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapred
16/02/04 08:17:03 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1641761722_0001
16/02/04 08:17:03 INFO mapred.LocalJobRunner: OutputCommitter set in config null
```

```
16/02/04 08:17:03 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/02/04 08:17:03 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputComm
16/02/04 08:17:03 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:17:03 INFO mapreduce.Job: Running job: job_local1641761722_0001
16/02/04 08:17:03 INFO mapred.LocalJobRunner: Waiting for map tasks
16/02/04 08:17:03 INFO mapred.LocalJobRunner: Starting task: attempt_local1641761722_0001_m_000000_0
16/02/04 08:17:03 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:17:03 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
16/02/04 08:17:03 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:17:03 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/hw3/xab:0+13160945
16/02/04 08:17:03 INFO mapred.MapTask: numReduceTasks: 1
16/02/04 08:17:03 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/02/04 08:17:03 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/02/04 08:17:03 INFO mapred.MapTask: soft limit at 83886080
16/02/04 08:17:03 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/02/04 08:17:03 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/02/04 08:17:03 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$Ma
16/02/04 08:17:03 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./mapper.
16/02/04 08:17:03 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.t
16/02/04 08:17:03 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduc
16/02/04 08:17:03 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.
16/02/04 08:17:03 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.
16/02/04 08:17:03 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduc
16/02/04 08:17:03 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use ma
16/02/04 08:17:03 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce
16/02/04 08:17:03 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.j
16/02/04 08:17:03 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.u
16/02/04 08:17:03 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapred
16/02/04 08:17:03 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.
16/02/04 08:17:03 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use map
16/02/04 08:17:03 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:03 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:03 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:03 INFO streaming.PipeMapRed: Records R/W=773/1
16/02/04 08:17:03 INFO streaming.PipeMapRed: R/W/S=1000/960/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:04 INFO streaming.PipeMapRed: R/W/S=10000/76488/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:04 INFO mapreduce.Job: Job job_local1641761722_0001 running in uber mode : false
16/02/04 08:17:04 INFO mapreduce.Job:  map 0% reduce 0%
16/02/04 08:17:05 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:17:05 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:17:05 INFO mapred.LocalJobRunner:
16/02/04 08:17:05 INFO mapred.MapTask: Starting flush of map output
16/02/04 08:17:05 INFO mapred.MapTask: Spilling map output
16/02/04 08:17:05 INFO mapred.MapTask: bufstart = 0; bufend = 5477089; bufvoid = 104857600
16/02/04 08:17:05 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 23425776(93703104); lengt
16/02/04 08:17:06 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./combine
16/02/04 08:17:06 INFO Configuration.deprecation: mapred.skip.map.auto.incr.proc.count is deprecated. I
16/02/04 08:17:06 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:06 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:06 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:06 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:06 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:06 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:06 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:NA [rec/s] out:NA [rec/s]
```

```
16/02/04 08:17:06 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:07 INFO streaming.PipeMapRed: R/W/S=400000/0/0 in:400000=400000/1 [rec/s] out:0=0/1 [rec
16/02/04 08:17:07 INFO streaming.PipeMapRed: R/W/S=500000/0/0 in:500000=500000/1 [rec/s] out:0=0/1 [rec
16/02/04 08:17:07 INFO streaming.PipeMapRed: R/W/S=600000/0/0 in:600000=600000/1 [rec/s] out:0=0/1 [rec
16/02/04 08:17:08 INFO streaming.PipeMapRed: Records R/W=697156/1
16/02/04 08:17:08 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:17:08 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:17:08 INFO mapred.MapTask: Finished spill 0
16/02/04 08:17:08 INFO mapred.Task: Task:attempt_local1641761722_0001_m_000000_0 is done. And is in the p
16/02/04 08:17:08 INFO mapred.LocalJobRunner: Records R/W=697156/1
16/02/04 08:17:08 INFO mapred.Task: Task 'attempt_local1641761722_0001_m_000000_0' done.
16/02/04 08:17:08 INFO mapred.LocalJobRunner: Finishing task: attempt_local1641761722_0001_m_000000_0
16/02/04 08:17:08 INFO mapred.LocalJobRunner: Starting task: attempt_local1641761722_0001_m_000001_0
16/02/04 08:17:08 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:17:08 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
16/02/04 08:17:08 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:17:08 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/hw3/xaa:0+13123345
16/02/04 08:17:08 INFO mapred.MapTask: numReduceTasks: 1
16/02/04 08:17:08 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/02/04 08:17:08 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/02/04 08:17:08 INFO mapred.MapTask: soft limit at 83886080
16/02/04 08:17:08 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/02/04 08:17:08 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/02/04 08:17:08 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$Ma
16/02/04 08:17:08 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./mapper.
16/02/04 08:17:08 INFO mapreduce.Job:  map 100% reduce 0%
16/02/04 08:17:08 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:08 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:08 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:08 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:08 INFO streaming.PipeMapRed: Records R/W=1571/1
16/02/04 08:17:08 INFO streaming.PipeMapRed: R/W/S=10000/79729/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:10 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:17:10 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:17:10 INFO mapred.LocalJobRunner:
16/02/04 08:17:10 INFO mapred.MapTask: Starting flush of map output
16/02/04 08:17:10 INFO mapred.MapTask: Spilling map output
16/02/04 08:17:10 INFO mapred.MapTask: bufstart = 0; bufend = 5547814; bufvoid = 104857600
16/02/04 08:17:10 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 23376416(93505664); lengt
16/02/04 08:17:10 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./combine
16/02/04 08:17:10 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:10 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:10 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:10 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:10 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:10 INFO mapreduce.Job:  map 25% reduce 0%
16/02/04 08:17:10 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:11 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:11 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:11 INFO streaming.PipeMapRed: R/W/S=400000/0/0 in:400000=400000/1 [rec/s] out:0=0/1 [rec
16/02/04 08:17:12 INFO streaming.PipeMapRed: R/W/S=500000/0/0 in:500000=500000/1 [rec/s] out:0=0/1 [rec
16/02/04 08:17:12 INFO streaming.PipeMapRed: R/W/S=600000/0/0 in:600000=600000/1 [rec/s] out:0=0/1 [rec
16/02/04 08:17:12 INFO streaming.PipeMapRed: R/W/S=700000/0/0 in:350000=700000/2 [rec/s] out:0=0/2 [rec
16/02/04 08:17:12 INFO streaming.PipeMapRed: Records R/W=709496/1
```

```
16/02/04 08:17:12 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:17:12 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:17:12 INFO mapred.MapTask: Finished spill 0
16/02/04 08:17:12 INFO mapred.Task: Task:attempt_local1641761722_0001_m_000001_0 is done. And is in the p
16/02/04 08:17:12 INFO mapred.LocalJobRunner: Records R/W=709496/1
16/02/04 08:17:12 INFO mapred.Task: Task 'attempt_local1641761722_0001_m_000001_0' done.
16/02/04 08:17:12 INFO mapred.LocalJobRunner: Finishing task: attempt_local1641761722_0001_m_000001_0
16/02/04 08:17:12 INFO mapred.LocalJobRunner: Starting task: attempt_local1641761722_0001_m_000002_0
16/02/04 08:17:12 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:17:12 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only o
16/02/04 08:17:12 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:17:12 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/hw3/xac:0+12615536
16/02/04 08:17:12 INFO mapred.MapTask: numReduceTasks: 1
16/02/04 08:17:12 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/02/04 08:17:12 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/02/04 08:17:12 INFO mapred.MapTask: soft limit at 83886080
16/02/04 08:17:12 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/02/04 08:17:12 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/02/04 08:17:12 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$Map
16/02/04 08:17:12 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./mapper.
16/02/04 08:17:12 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:12 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:12 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:12 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:12 INFO streaming.PipeMapRed: Records R/W=1593/1
16/02/04 08:17:13 INFO streaming.PipeMapRed: R/W/S=10000/75602/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:13 INFO mapreduce.Job:  map 100% reduce 0%
16/02/04 08:17:14 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:17:14 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:17:14 INFO mapred.LocalJobRunner:
16/02/04 08:17:14 INFO mapred.MapTask: Starting flush of map output
16/02/04 08:17:14 INFO mapred.MapTask: Spilling map output
16/02/04 08:17:14 INFO mapred.MapTask: bufstart = 0; bufend = 5404684; bufvoid = 104857600
16/02/04 08:17:14 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 23552264(94209056); lengt
16/02/04 08:17:14 INFO mapreduce.Job:  map 50% reduce 0%
16/02/04 08:17:14 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./combine
16/02/04 08:17:14 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:14 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:14 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:14 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:14 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:15 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:15 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:15 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:300000=300000/1 [rec/s] out:0=0/1 [rec/
16/02/04 08:17:16 INFO streaming.PipeMapRed: R/W/S=400000/0/0 in:400000=400000/1 [rec/s] out:0=0/1 [rec/
16/02/04 08:17:16 INFO streaming.PipeMapRed: R/W/S=500000/0/0 in:500000=500000/1 [rec/s] out:0=0/1 [rec/
16/02/04 08:17:16 INFO streaming.PipeMapRed: R/W/S=600000/0/0 in:300000=600000/2 [rec/s] out:0=0/2 [rec/
16/02/04 08:17:17 INFO streaming.PipeMapRed: Records R/W=665534/1
16/02/04 08:17:17 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:17:17 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:17:17 INFO mapred.MapTask: Finished spill 0
16/02/04 08:17:17 INFO mapred.Task: Task:attempt_local1641761722_0001_m_000002_0 is done. And is in the p
16/02/04 08:17:17 INFO mapred.LocalJobRunner: Records R/W=665534/1
16/02/04 08:17:17 INFO mapred.Task: Task 'attempt_local1641761722_0001_m_000002_0' done.
```

```
16/02/04 08:17:17 INFO mapred.LocalJobRunner: Finishing task: attempt_local1641761722_0001_m_000002_0
16/02/04 08:17:17 INFO mapred.LocalJobRunner: Starting task: attempt_local1641761722_0001_m_000003_0
16/02/04 08:17:17 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:17:17 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only o
16/02/04 08:17:17 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:17:17 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/hw3/xad:0+12006486
16/02/04 08:17:17 INFO mapred.MapTask: numReduceTasks: 1
16/02/04 08:17:17 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/02/04 08:17:17 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/02/04 08:17:17 INFO mapred.MapTask: soft limit at 83886080
16/02/04 08:17:17 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/02/04 08:17:17 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/02/04 08:17:17 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$Ma
16/02/04 08:17:17 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./mapper.
16/02/04 08:17:17 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:17 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:17 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:17 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:17 INFO streaming.PipeMapRed: Records R/W=1639/1
16/02/04 08:17:17 INFO mapreduce.Job:  map 100% reduce 0%
16/02/04 08:17:17 INFO streaming.PipeMapRed: R/W/S=10000/75347/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:19 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:17:19 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:17:19 INFO mapred.LocalJobRunner:
16/02/04 08:17:19 INFO mapred.MapTask: Starting flush of map output
16/02/04 08:17:19 INFO mapred.MapTask: Spilling map output
16/02/04 08:17:19 INFO mapred.MapTask: bufstart = 0; bufend = 5085004; bufvoid = 104857600
16/02/04 08:17:19 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 23716648(94866592); lengt
16/02/04 08:17:19 INFO mapreduce.Job:  map 75% reduce 0%
16/02/04 08:17:19 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./combine
16/02/04 08:17:19 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:19 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:19 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:19 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:19 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:20 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:20 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:20 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:300000=300000/1 [rec/s] out:0=0/1 [rec
16/02/04 08:17:21 INFO streaming.PipeMapRed: R/W/S=400000/0/0 in:400000=400000/1 [rec/s] out:0=0/1 [rec
16/02/04 08:17:21 INFO streaming.PipeMapRed: R/W/S=500000/0/0 in:500000=500000/1 [rec/s] out:0=0/1 [rec
16/02/04 08:17:21 INFO streaming.PipeMapRed: R/W/S=600000/0/0 in:300000=600000/2 [rec/s] out:0=0/2 [rec
16/02/04 08:17:22 INFO streaming.PipeMapRed: Records R/W=624438/1
16/02/04 08:17:22 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:17:22 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:17:22 INFO mapred.MapTask: Finished spill 0
16/02/04 08:17:22 INFO mapred.Task: Task:attempt_local1641761722_0001_m_000003_0 is done. And is in the p
16/02/04 08:17:22 INFO mapred.LocalJobRunner: Records R/W=624438/1
16/02/04 08:17:22 INFO mapred.Task: Task 'attempt_local1641761722_0001_m_000003_0' done.
16/02/04 08:17:22 INFO mapred.LocalJobRunner: Finishing task: attempt_local1641761722_0001_m_000003_0
16/02/04 08:17:22 INFO mapred.LocalJobRunner: map task executor complete.
16/02/04 08:17:22 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/02/04 08:17:22 INFO mapred.LocalJobRunner: Starting task: attempt_local1641761722_0001_r_000000_0
16/02/04 08:17:22 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:17:22 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only o
```

```
16/02/04 08:17:22 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:17:22 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task
16/02/04 08:17:22 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLi
16/02/04 08:17:22 INFO reduce.EventFetcher: attempt_local1641761722_0001_r_000000_0 Thread started: Event
16/02/04 08:17:22 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1
16/02/04 08:17:22 INFO reduce.InMemoryMapOutput: Read 2762 bytes from map-output for attempt_local164176
16/02/04 08:17:22 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 2762, inMemory
16/02/04 08:17:22 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1
16/02/04 08:17:22 INFO reduce.InMemoryMapOutput: Read 2471 bytes from map-output for attempt_local164176
16/02/04 08:17:22 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 2471, inMemory
16/02/04 08:17:22 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1
16/02/04 08:17:22 INFO reduce.InMemoryMapOutput: Read 2002 bytes from map-output for attempt_local164176
16/02/04 08:17:22 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 2002, inMemory
16/02/04 08:17:22 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1
16/02/04 08:17:22 INFO reduce.InMemoryMapOutput: Read 2744 bytes from map-output for attempt_local164176
16/02/04 08:17:22 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 2744, inMemory
16/02/04 08:17:22 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/02/04 08:17:22 INFO mapred.LocalJobRunner: 4 / 4 copied.
16/02/04 08:17:22 INFO reduce.MergeManagerImpl: finalMerge called with 4 in-memory map-outputs and 0 on-
16/02/04 08:17:22 INFO mapred.Merger: Merging 4 sorted segments
16/02/04 08:17:22 INFO mapred.Merger: Down to the last merge-pass, with 4 segments left of total size: 9
16/02/04 08:17:22 INFO reduce.MergeManagerImpl: Merged 4 segments, 9979 bytes to disk to satisfy reduce
16/02/04 08:17:22 INFO reduce.MergeManagerImpl: Merging 1 files, 9977 bytes from disk
16/02/04 08:17:22 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/02/04 08:17:22 INFO mapred.Merger: Merging 1 sorted segments
16/02/04 08:17:22 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 9
16/02/04 08:17:22 INFO mapred.LocalJobRunner: 4 / 4 copied.
16/02/04 08:17:22 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./reducer
16/02/04 08:17:22 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapredu
16/02/04 08:17:22 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce
16/02/04 08:17:22 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:22 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:22 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:22 INFO streaming.PipeMapRed: Records R/W=665/1
16/02/04 08:17:22 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:17:22 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:17:22 INFO mapred.Task: Task:attempt_local1641761722_0001_r_000000_0 is done. And is in the p
16/02/04 08:17:22 INFO mapred.LocalJobRunner: 4 / 4 copied.
16/02/04 08:17:22 INFO mapred.Task: Task attempt_local1641761722_0001_r_000000_0 is allowed to commit now
16/02/04 08:17:22 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1641761722_0001_r_0
16/02/04 08:17:22 INFO mapred.LocalJobRunner: Records R/W=665/1 > reduce
16/02/04 08:17:22 INFO mapred.Task: Task 'attempt_local1641761722_0001_r_000000_0' done.
16/02/04 08:17:22 INFO mapred.LocalJobRunner: Finishing task: attempt_local1641761722_0001_r_000000_0
16/02/04 08:17:22 INFO mapred.LocalJobRunner: reduce task executor complete.
16/02/04 08:17:22 INFO mapreduce.Job:  map 100% reduce 100%
16/02/04 08:17:22 INFO mapreduce.Job: Job job_local1641761722_0001 completed successfully
16/02/04 08:17:22 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=554287
                FILE: Number of bytes written=1984331
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=180157685
                HDFS: Number of bytes written=2536
                HDFS: Number of read operations=61
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=7
        Map-Reduce Framework
                Map input records=312912
                Map output records=2696624
                Map output bytes=21514591
                Map output materialized bytes=9995
                Input split bytes=344
                Combine input records=2696624
                Combine output records=665
                Reduce input groups=656
                Reduce shuffle bytes=9995
                Reduce input records=665
                Reduce output records=189
                Spilled Records=1330
                Shuffled Maps =4
                Failed Shuffles=0
                Merged Map outputs=4
                GC time elapsed (ms)=462
                Total committed heap usage (bytes)=1824522240
        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=50906312
        File Output Format Counters
                Bytes Written=2536
16/02/04 08:17:22 INFO streaming.StreamJob: Output directory: /user/hw3/output_3_2_4
```

```
In [217]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -cat /user/hw3/output_3_2_4/part-00000 | hea
```

```
16/02/04 08:17:24 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
* 1348312
 4
APR 3431
ATM 2422
Account 16555
Advertising 1193
Application 8868
Applied 139
Arbitration 168
Balance 597
```

```
In [218]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -cat /user/hw3/output_3_2_4/part-00000 > re
```

```
16/02/04 08:17:26 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
```

** Running a second map-reduce job to sort the word counts **

```
In [219]: %%writefile mapper.py
          #!/usr/bin/python
          import sys
          import re
          import csv

          for line in sys.stdin:
              line = re.split(r'[\s]+',line)
              print "%s,%s" %(line[0],line[1])
```

```
Overwriting mapper.py
```

```
In [220]: %%writefile reducer.py
          #!/usr/bin/python
          import sys
          import re
          i=0
          for line in sys.stdin:
              line = re.split(r'[,]',line.strip())

              if(i==0):
                  total_count = int(line[1])
                  total_count_read = 1
                  i = i+1
              if(total_count_read==1):
                  rel_freq = float(line[1])/float(total_count)
                  print "%s,%s,%s,%s" %(line[0],line[1],rel_freq,total_count)
```

```
Overwriting reducer.py
```

```
In [221]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -put reduce_3_2_4 /user/hw3
```

```
16/02/04 08:17:29 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
```

```
In [222]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -rm -r -f /user/hw3/output_3_2_4f
```

```
16/02/04 08:17:32 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
```

```
In [223]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hadoop jar /Users/Vamsi/Downloads/hadoop-2.7.1/bin/ha
          -D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator \
          -D stream.map.output.field.separator=, \
          -D stream.num.map.output.key.fields=2 \
          -D map.output.key.field.separator=, \
          -D mapred.text.key.comparator.options=-k2,2nr \
          -D mapred.reduce.tasks=1 \
          -input /user/hw3/reduce_3_2_4 \
          -output /user/hw3/output_3_2_4f \
          -mapper mapper.py \
          -reducer reducer.py
```

```
16/02/04 08:17:34 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
16/02/04 08:17:35 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.ses
16/02/04 08:17:35 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/02/04 08:17:35 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessio
16/02/04 08:17:35 INFO mapred.FileInputFormat: Total input paths to process : 1
16/02/04 08:17:35 INFO mapreduce.JobSubmitter: number of splits:1
```

```
16/02/04 08:17:36 INFO Configuration.deprecation: map.output.key.field.separator is deprecated. Instead
16/02/04 08:17:36 INFO Configuration.deprecation: mapred.text.key.comparator.options is deprecated. Inst
16/02/04 08:17:36 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapred
16/02/04 08:17:36 INFO Configuration.deprecation: mapred.output.key.comparator.class is deprecated. Inst
16/02/04 08:17:36 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local2004021949_0001
16/02/04 08:17:36 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/02/04 08:17:36 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/02/04 08:17:36 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputComm
16/02/04 08:17:36 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:17:36 INFO mapreduce.Job: Running job: job_local2004021949_0001
16/02/04 08:17:36 INFO mapred.LocalJobRunner: Waiting for map tasks
16/02/04 08:17:36 INFO mapred.LocalJobRunner: Starting task: attempt_local2004021949_0001_m_000000_0
16/02/04 08:17:36 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:17:36 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
16/02/04 08:17:36 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:17:36 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/hw3/reduce_3_2_4:0+2
16/02/04 08:17:36 INFO mapred.MapTask: numReduceTasks: 1
16/02/04 08:17:36 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/02/04 08:17:36 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/02/04 08:17:36 INFO mapred.MapTask: soft limit at 83886080
16/02/04 08:17:36 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/02/04 08:17:36 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/02/04 08:17:36 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$Map
16/02/04 08:17:36 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./mapper.
16/02/04 08:17:36 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use ma
16/02/04 08:17:36 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce
16/02/04 08:17:36 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapredu
16/02/04 08:17:36 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.
16/02/04 08:17:36 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.ta
16/02/04 08:17:36 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce
16/02/04 08:17:36 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.
16/02/04 08:17:36 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.
16/02/04 08:17:36 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce
16/02/04 08:17:36 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.jo
16/02/04 08:17:36 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.us
16/02/04 08:17:36 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use map
16/02/04 08:17:36 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:36 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:36 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:36 INFO streaming.PipeMapRed: Records R/W=189/1
16/02/04 08:17:36 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:17:36 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:17:36 INFO mapred.LocalJobRunner:
16/02/04 08:17:36 INFO mapred.MapTask: Starting flush of map output
16/02/04 08:17:36 INFO mapred.MapTask: Spilling map output
16/02/04 08:17:36 INFO mapred.MapTask: bufstart = 0; bufend = 2536; bufvoid = 104857600
16/02/04 08:17:36 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26213644(104854576); leng
16/02/04 08:17:36 INFO mapred.MapTask: Finished spill 0
16/02/04 08:17:36 INFO mapred.Task: Task:attempt_local2004021949_0001_m_000000_0 is done. And is in the p
16/02/04 08:17:36 INFO mapred.LocalJobRunner: Records R/W=189/1
16/02/04 08:17:36 INFO mapred.Task: Task 'attempt_local2004021949_0001_m_000000_0' done.
16/02/04 08:17:36 INFO mapred.LocalJobRunner: Finishing task: attempt_local2004021949_0001_m_000000_0
16/02/04 08:17:36 INFO mapred.LocalJobRunner: map task executor complete.
16/02/04 08:17:36 INFO mapred.LocalJobRunner: Waiting for reduce tasks
```

```
16/02/04 08:17:36 INFO mapred.LocalJobRunner: Starting task: attempt_local2004021949_0001_r_000000_0
16/02/04 08:17:36 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:17:36 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
16/02/04 08:17:36 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:17:37 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task
16/02/04 08:17:37 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleL
16/02/04 08:17:37 INFO reduce.EventFetcher: attempt_local2004021949_0001_r_000000_0 Thread started: Event
16/02/04 08:17:37 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local2
16/02/04 08:17:37 INFO reduce.InMemoryMapOutput: Read 2916 bytes from map-output for attempt_local200402
16/02/04 08:17:37 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 2916, inMemory
16/02/04 08:17:37 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/02/04 08:17:37 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/04 08:17:37 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on
16/02/04 08:17:37 INFO mapred.Merger: Merging 1 sorted segments
16/02/04 08:17:37 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size:
16/02/04 08:17:37 INFO reduce.MergeManagerImpl: Merged 1 segments, 2916 bytes to disk to satisfy reduce
16/02/04 08:17:37 INFO reduce.MergeManagerImpl: Merging 1 files, 2920 bytes from disk
16/02/04 08:17:37 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/02/04 08:17:37 INFO mapred.Merger: Merging 1 sorted segments
16/02/04 08:17:37 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size:
16/02/04 08:17:37 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/04 08:17:37 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./reducer
16/02/04 08:17:37 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapredu
16/02/04 08:17:37 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduc
16/02/04 08:17:37 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:37 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:37 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:37 INFO streaming.PipeMapRed: Records R/W=189/1
16/02/04 08:17:37 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:17:37 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:17:37 INFO mapred.Task: Task:attempt_local2004021949_0001_r_000000_0 is done. And is in the
16/02/04 08:17:37 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/04 08:17:37 INFO mapred.Task: Task attempt_local2004021949_0001_r_000000_0 is allowed to commit no
16/02/04 08:17:37 INFO output.FileOutputCommitter: Saved output of task ’attempt_local2004021949_0001_r_
16/02/04 08:17:37 INFO mapred.LocalJobRunner: Records R/W=189/1 > reduce
16/02/04 08:17:37 INFO mapred.Task: Task ’attempt_local2004021949_0001_r_000000_0’ done.
16/02/04 08:17:37 INFO mapred.LocalJobRunner: Finishing task: attempt_local2004021949_0001_r_000000_0
16/02/04 08:17:37 INFO mapred.LocalJobRunner: reduce task executor complete.
16/02/04 08:17:37 INFO mapreduce.Job: Job job_local2004021949_0001 running in uber mode : false
16/02/04 08:17:37 INFO mapreduce.Job:  map 100% reduce 100%
16/02/04 08:17:37 INFO mapreduce.Job: Job job_local2004021949_0001 completed successfully
16/02/04 08:17:37 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=217644
                FILE: Number of bytes written=786182
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=5072
                HDFS: Number of bytes written=7276
                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=189
                Map output records=189
                Map output bytes=2536
                Map output materialized bytes=2920
                Input split bytes=95
                Combine input records=0
                Combine output records=0
                Reduce input groups=189
                Reduce shuffle bytes=2920
                Reduce input records=189
                Reduce output records=189
                Spilled Records=378
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=5
                Total committed heap usage (bytes)=495976448
        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=2536
        File Output Format Counters
                Bytes Written=7276
16/02/04 08:17:37 INFO streaming.StreamJob: Output directory: /user/hw3/output_3_2_4f
```

** Printing out the 50 most frequent words alongwith their frequencies and relative frequencies below. **

The format of the output is:
Word, Frequency , Relative Frequency, Total Count of words

```
In [224]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -cat /user/hw3/output_3_2_4f/part-00000 | h
```

```
16/02/04 08:17:39 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
*,1348312,1.0,1348312
Loan,107254,0.0795468704573,1348312
foreclosure,70487,0.0522779594041,1348312
collection,70487,0.0522779594041,1348312
modification,70487,0.0522779594041,1348312
account,40893,0.0303290336361,1348312
or,40508,0.0300434914174,1348312
credit,40483,0.0300249497149,1348312
payments,39993,0.0296615323456,1348312
escrow,36767,0.0272689110532,1348312
servicing,36767,0.0272689110532,1348312
report,34903,0.0258864417138,1348312
Incorrect,29133,0.0216070167736,1348312
on,29069,0.0215595500151,1348312
information,29069,0.0215595500151,1348312
debt,26531,0.0196771963759,1348312
not,18477,0.013703801494,1348312
```

```
owed,17972,0.0133292591032,1348312
collect,17972,0.0133292591032,1348312
attempts,17972,0.0133292591032,1348312
Cont'd,17972,0.0133292591032,1348312
Account,16555,0.0122783154047,1348312
and,16448,0.012198956918,1348312
opening,16205,0.0120187315695,1348312
closing,16205,0.0120187315695,1348312
management,16205,0.0120187315695,1348312
Credit,14768,0.010952954509,1348312
of,13983,0.0103707450501,1348312
loan,12376,0.00917888441251,1348312
my,10731,0.00795884038709,1348312
Deposits,10555,0.00782830680139,1348312
withdrawals,10555,0.00782830680139,1348312
Problems,9484,0.0070339802657,1348312
Application,8868,0.00657711271575,1348312
Communication,8671,0.00643100409994,1348312
tactics,8671,0.00643100409994,1348312
broker,8625,0.00639688736732,1348312
originator,8625,0.00639688736732,1348312
mortgage,8625,0.00639688736732,1348312
to,8401,0.00623075371279,1348312
Billing,8158,0.00605052836435,1348312
Other,7886,0.005848794641,1348312
Disclosure,7655,0.00567746930977,1348312
verification,7655,0.00567746930977,1348312
disputes,6938,0.00514569328167,1348312
reporting,6559,0.00486460107156,1348312
lease,6337,0.00469995075324,1348312
the,6248,0.00463394229229,1348312
caused,5663,0.00420006645346,1348312
by,5663,0.00420006645346,1348312
being,5663,0.00420006645346,1348312
```

```
In [225]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -cat /user/hw3/output_3_2_4f/part-00000 | ta
```

```
16/02/04 08:17:41 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
credited,92,6.82334652514e-05,1348312
Payment,92,6.82334652514e-05,1348312
Convenience,75,5.56251075419e-05,1348312
checks,75,5.56251075419e-05,1348312
wrong,71,5.26584351396e-05,1348312
day,71,5.26584351396e-05,1348312
amt,71,5.26584351396e-05,1348312
missing,64,4.74667584357e-05,1348312
disclosures,64,4.74667584357e-05,1348312
,4,2.96667240223e-06,1348312
```

0.0.11 HW3.3. Shopping Cart Analysis

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

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

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

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

Here are the first few lines of the ProductPurchaseData

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

Do some exploratory data analysis of this dataset.

How many unique items are available from this supplier?

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

In [226]:

```
#Loading the data-set
!/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -put ProductPurchaseData.txt /user/hw3
```

```
16/02/04 08:17:44 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
```

In [227]:

```
%%writefile mapper.py
#!/usr/bin/python
import sys
import re

for line in sys.stdin:
    line = re.split(r'[\s]',line.strip())
    for l in line:
        print "%s\t1" %l
        print "*\t1"
```

```
Overwriting mapper.py
```

In [228]:

```
%%writefile reducer.py
#!/usr/bin/python
import sys
import re

count = 0
prev_id = None

for line in sys.stdin:
    line = re.split(r'[\t]',line.strip())
    if((prev_id !=None) and (line[0] !=prev_id)):
        print "%s\t%s" %(prev_id,count)
        count = 0
    count +=1
    prev_id = line[0]
print "%s\t%s" %(prev_id,count)
```

```
Overwriting reducer.py
```

```
In [229]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hadoop jar /Users/Vamsi/Downloads/hadoop-2.7.1/bin/ha
          -D mapred.reduce.tasks=1 \
          -input /user/hw3/ProductPurchaseData.txt \
          -output /user/hw3/output_3_3 \
          -mapper mapper.py \
          -reducer reducer.py
```

```
16/02/04 08:17:46 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
16/02/04 08:17:47 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.se
16/02/04 08:17:47 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/02/04 08:17:47 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessio
16/02/04 08:17:48 INFO mapred.FileInputFormat: Total input paths to process : 1
16/02/04 08:17:48 INFO mapreduce.JobSubmitter: number of splits:1
16/02/04 08:17:48 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapred
16/02/04 08:17:48 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local38036051_0001
16/02/04 08:17:48 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/02/04 08:17:48 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/02/04 08:17:48 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputComm
16/02/04 08:17:48 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:17:48 INFO mapreduce.Job: Running job: job_local38036051_0001
16/02/04 08:17:48 INFO mapred.LocalJobRunner: Waiting for map tasks
16/02/04 08:17:48 INFO mapred.LocalJobRunner: Starting task: attempt_local38036051_0001_m_000000_0
16/02/04 08:17:49 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:17:49 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only o
16/02/04 08:17:49 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:17:49 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/hw3/ProductPurchaseD
16/02/04 08:17:49 INFO mapred.MapTask: numReduceTasks: 1
16/02/04 08:17:49 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/02/04 08:17:49 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/02/04 08:17:49 INFO mapred.MapTask: soft limit at 83886080
16/02/04 08:17:49 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/02/04 08:17:49 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/02/04 08:17:49 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$Map
16/02/04 08:17:49 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./mapper.p
16/02/04 08:17:49 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.ta
16/02/04 08:17:49 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce
16/02/04 08:17:49 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.m
16/02/04 08:17:49 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.j
16/02/04 08:17:49 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce
16/02/04 08:17:49 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use map
16/02/04 08:17:49 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce
16/02/04 08:17:49 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.jo
16/02/04 08:17:49 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.us
16/02/04 08:17:49 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapredu
16/02/04 08:17:49 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.t
16/02/04 08:17:49 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapr
16/02/04 08:17:49 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:49 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:49 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:49 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:49 INFO streaming.PipeMapRed: Records R/W=1216/1
16/02/04 08:17:49 INFO mapreduce.Job: Job job_local38036051_0001 running in uber mode : false
16/02/04 08:17:49 INFO mapreduce.Job:  map 0% reduce 0%
16/02/04 08:17:50 INFO streaming.PipeMapRed: R/W/S=10000/238936/0 in:10000=10000/1 [rec/s] out:238955=23
```

```
16/02/04 08:17:51 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:17:51 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:17:51 INFO mapred.LocalJobRunner:
16/02/04 08:17:51 INFO mapred.MapTask: Starting flush of map output
16/02/04 08:17:51 INFO mapred.MapTask: Spilling map output
16/02/04 08:17:51 INFO mapred.MapTask: bufstart = 0; bufend = 5712360; bufvoid = 104857600
16/02/04 08:17:51 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 23167808(92671232); lengt
16/02/04 08:17:52 INFO mapred.MapTask: Finished spill 0
16/02/04 08:17:52 INFO mapred.Task: Task:attempt_local38036051_0001_m_000000_0 is done. And is in the pro
16/02/04 08:17:52 INFO mapred.LocalJobRunner: Records R/W=1216/1
16/02/04 08:17:52 INFO mapred.Task: Task 'attempt_local38036051_0001_m_000000_0' done.
16/02/04 08:17:52 INFO mapred.LocalJobRunner: Finishing task: attempt_local38036051_0001_m_000000_0
16/02/04 08:17:52 INFO mapred.LocalJobRunner: map task executor complete.
16/02/04 08:17:52 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/02/04 08:17:52 INFO mapred.LocalJobRunner: Starting task: attempt_local38036051_0001_r_000000_0
16/02/04 08:17:52 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:17:52 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only 
16/02/04 08:17:52 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:17:52 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task
16/02/04 08:17:52 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLi
16/02/04 08:17:52 INFO reduce.EventFetcher: attempt_local38036051_0001_r_000000_0 Thread started: EventFe
16/02/04 08:17:52 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local3
16/02/04 08:17:52 INFO reduce.InMemoryMapOutput: Read 7235658 bytes from map-output for attempt_local38
16/02/04 08:17:52 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 7235658, inMem
16/02/04 08:17:52 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/02/04 08:17:52 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/04 08:17:52 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on
16/02/04 08:17:52 INFO mapred.Merger: Merging 1 sorted segments
16/02/04 08:17:52 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 
16/02/04 08:17:52 INFO reduce.MergeManagerImpl: Merged 1 segments, 7235658 bytes to disk to satisfy red
16/02/04 08:17:52 INFO reduce.MergeManagerImpl: Merging 1 files, 7235662 bytes from disk
16/02/04 08:17:52 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/02/04 08:17:52 INFO mapred.Merger: Merging 1 sorted segments
16/02/04 08:17:52 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 
16/02/04 08:17:52 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/04 08:17:52 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./reducer
16/02/04 08:17:52 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapred
16/02/04 08:17:52 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduc
16/02/04 08:17:52 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:52 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:52 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:52 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:52 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:52 INFO mapreduce.Job:  map 100% reduce 0%
16/02/04 08:17:53 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:53 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:17:53 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:300000=300000/1 [rec/s] out:0=0/1 [rec
16/02/04 08:17:54 INFO streaming.PipeMapRed: R/W/S=400000/0/0 in:400000=400000/1 [rec/s] out:0=0/1 [rec
16/02/04 08:17:54 INFO streaming.PipeMapRed: Records R/W=432991/1
16/02/04 08:17:54 INFO streaming.PipeMapRed: R/W/S=500000/2889/0 in:250000=500000/2 [rec/s] out:1444=28
16/02/04 08:17:55 INFO streaming.PipeMapRed: R/W/S=600000/5780/0 in:300000=600000/2 [rec/s] out:2890=57
16/02/04 08:17:55 INFO streaming.PipeMapRed: R/W/S=700000/8675/0 in:350000=700000/2 [rec/s] out:4337=86
16/02/04 08:17:55 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:17:55 INFO streaming.PipeMapRed: mapRedFinished
```

```
16/02/04 08:17:55 INFO mapred.Task: Task:attempt_local38036051_0001_r_000000_0 is done. And is in the pro
16/02/04 08:17:55 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/04 08:17:55 INFO mapred.Task: Task attempt_local38036051_0001_r_000000_0 is allowed to commit now
16/02/04 08:17:55 INFO output.FileOutputCommitter: Saved output of task 'attempt_local38036051_0001_r_000
16/02/04 08:17:55 INFO mapred.LocalJobRunner: Records R/W=432991/1 > reduce
16/02/04 08:17:55 INFO mapred.Task: Task 'attempt_local38036051_0001_r_000000_0' done.
16/02/04 08:17:55 INFO mapred.LocalJobRunner: Finishing task: attempt_local38036051_0001_r_000000_0
16/02/04 08:17:55 INFO mapred.LocalJobRunner: reduce task executor complete.
16/02/04 08:17:55 INFO mapreduce.Job:  map 100% reduce 100%
16/02/04 08:17:55 INFO mapreduce.Job: Job job_local38036051_0001 completed successfully
16/02/04 08:17:55 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=14683152
                FILE: Number of bytes written=22474080
                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=142667
                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=761648
                Map output bytes=5712360
                Map output materialized bytes=7235662
                Input split bytes=106
                Combine input records=0
                Combine output records=0
                Reduce input groups=12593
                Reduce shuffle bytes=7235662
                Reduce input records=761648
                Reduce output records=12593
                Spilled Records=1523296
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=5
                Total committed heap usage (bytes)=494927872
        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=142667
16/02/04 08:17:55 INFO streaming.StreamJob: Output directory: /user/hw3/output_3_3
```

```
In [230]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -cat /user/hw3/output_3_3/part-00000 > coun
```

```
16/02/04 08:17:57 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
```

```
In [231]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -put counts_3_3 /user/hw3
```

```
16/02/04 08:18:00 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
```

** Using a second map-reduce job to perform the sort **

```
In [232]: %%writefile mapper.py
          #!/usr/bin/python
          import sys
          import re
          import csv

          i = 0
          for line in sys.stdin:
              line = re.split(r'[\t]',line.strip())
              print "%s,%s" %(line[0],line[1])
```

```
Overwriting mapper.py
```

```
In [233]: %%writefile reducer.py
          #!/usr/bin/python
          import sys
          import re
          i=0
          for line in sys.stdin:
              line = re.split(r'[,]',line.strip())

              if(i==0):
                  total_count = int(line[1])
                  i = i+1
              elif(i==1):
                  rel_freq = float(line[1])/float(total_count)
                  print "%s,%s,%s,%s" %(line[0],line[1],rel_freq,total_count)
```

```
Overwriting reducer.py
```

```
In [234]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hadoop jar /Users/Vamsi/Downloads/hadoop-2.7.1/bin/ha
          -D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator \
          -D stream.map.output.field.separator=, \
          -D stream.num.map.output.key.fields=2 \
          -D map.output.key.field.separator=, \
          -D mapred.text.key.comparator.options=-k2,2nr \
          -D mapred.reduce.tasks=1 \
          -input /user/hw3/counts_3_3 \
          -output /user/hw3/output_3_3_out \
          -mapper mapper.py \
          -reducer reducer.py
```

```
16/02/04 08:18:02 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
16/02/04 08:18:03 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.se
16/02/04 08:18:03 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/02/04 08:18:03 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessi
16/02/04 08:18:03 INFO mapred.FileInputFormat: Total input paths to process : 1
16/02/04 08:18:03 INFO mapreduce.JobSubmitter: number of splits:1
```

```
16/02/04 08:18:04 INFO Configuration.deprecation: map.output.key.field.separator is deprecated. Instead
16/02/04 08:18:04 INFO Configuration.deprecation: mapred.text.key.comparator.options is deprecated. Ins
16/02/04 08:18:04 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapre
16/02/04 08:18:04 INFO Configuration.deprecation: mapred.output.key.comparator.class is deprecated. Ins
16/02/04 08:18:04 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1594728140_0001
16/02/04 08:18:04 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/02/04 08:18:04 INFO mapreduce.Job: Running job: job_local1594728140_0001
16/02/04 08:18:04 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/02/04 08:18:04 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCom
16/02/04 08:18:04 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:18:04 INFO mapred.LocalJobRunner: Waiting for map tasks
16/02/04 08:18:04 INFO mapred.LocalJobRunner: Starting task: attempt_local1594728140_0001_m_000000_0
16/02/04 08:18:04 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:18:04 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
16/02/04 08:18:04 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:18:04 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/hw3/counts_3_3:0+142
16/02/04 08:18:04 INFO mapred.MapTask: numReduceTasks: 1
16/02/04 08:18:04 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/02/04 08:18:04 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/02/04 08:18:04 INFO mapred.MapTask: soft limit at 83886080
16/02/04 08:18:04 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/02/04 08:18:04 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/02/04 08:18:04 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$Ma
16/02/04 08:18:04 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./mapper.
16/02/04 08:18:04 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use ma
16/02/04 08:18:04 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce
16/02/04 08:18:04 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapred
16/02/04 08:18:04 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.
16/02/04 08:18:04 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.t
16/02/04 08:18:04 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduc
16/02/04 08:18:04 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.
16/02/04 08:18:04 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.
16/02/04 08:18:04 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduc
16/02/04 08:18:04 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.j
16/02/04 08:18:04 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.u
16/02/04 08:18:04 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use map
16/02/04 08:18:04 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:18:04 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:18:04 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:18:04 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:18:04 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:18:04 INFO streaming.PipeMapRed: Records R/W=11568/1
16/02/04 08:18:05 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:18:05 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:18:05 INFO mapred.LocalJobRunner:
16/02/04 08:18:05 INFO mapred.MapTask: Starting flush of map output
16/02/04 08:18:05 INFO mapred.MapTask: Spilling map output
16/02/04 08:18:05 INFO mapred.MapTask: bufstart = 0; bufend = 155260; bufvoid = 104857600
16/02/04 08:18:05 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26164028(104656112); leng
16/02/04 08:18:05 INFO mapred.MapTask: Finished spill 0
16/02/04 08:18:05 INFO mapred.Task: Task:attempt_local1594728140_0001_m_000000_0 is done. And is in the
16/02/04 08:18:05 INFO mapred.LocalJobRunner: Records R/W=11568/1
16/02/04 08:18:05 INFO mapred.Task: Task 'attempt_local1594728140_0001_m_000000_0' done.
16/02/04 08:18:05 INFO mapred.LocalJobRunner: Finishing task: attempt_local1594728140_0001_m_000000_0
```

```
16/02/04 08:18:05 INFO mapred.LocalJobRunner: map task executor complete.
16/02/04 08:18:05 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/02/04 08:18:05 INFO mapred.LocalJobRunner: Starting task: attempt_local1594728140_0001_r_000000_0
16/02/04 08:18:05 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:18:05 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only o
16/02/04 08:18:05 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:18:05 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task
16/02/04 08:18:05 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLi
16/02/04 08:18:05 INFO reduce.EventFetcher: attempt_local1594728140_0001_r_000000_0 Thread started: Event
16/02/04 08:18:05 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1
16/02/04 08:18:05 INFO reduce.InMemoryMapOutput: Read 180448 bytes from map-output for attempt_local1594
16/02/04 08:18:05 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 180448, inMemor
16/02/04 08:18:05 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/02/04 08:18:05 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/04 08:18:05 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-
16/02/04 08:18:05 INFO mapred.Merger: Merging 1 sorted segments
16/02/04 08:18:05 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 
16/02/04 08:18:05 INFO reduce.MergeManagerImpl: Merged 1 segments, 180448 bytes to disk to satisfy reduc
16/02/04 08:18:05 INFO reduce.MergeManagerImpl: Merging 1 files, 180452 bytes from disk
16/02/04 08:18:05 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/02/04 08:18:05 INFO mapred.Merger: Merging 1 sorted segments
16/02/04 08:18:05 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 
16/02/04 08:18:05 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/04 08:18:05 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./reducer
16/02/04 08:18:05 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapredu
16/02/04 08:18:05 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce
16/02/04 08:18:05 INFO mapreduce.Job: Job job_local1594728140_0001 running in uber mode : false
16/02/04 08:18:05 INFO mapreduce.Job:  map 100% reduce 0%
16/02/04 08:18:05 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:18:05 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:18:05 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:18:05 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:18:05 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:18:05 INFO streaming.PipeMapRed: Records R/W=10578/1
16/02/04 08:18:05 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:18:05 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:18:05 INFO mapred.Task: Task:attempt_local1594728140_0001_r_000000_0 is done. And is in the p
16/02/04 08:18:05 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/04 08:18:05 INFO mapred.Task: Task attempt_local1594728140_0001_r_000000_0 is allowed to commit now
16/02/04 08:18:05 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1594728140_0001_r_0
16/02/04 08:18:05 INFO mapred.LocalJobRunner: Records R/W=10578/1 > reduce
16/02/04 08:18:05 INFO mapred.Task: Task 'attempt_local1594728140_0001_r_000000_0' done.
16/02/04 08:18:05 INFO mapred.LocalJobRunner: Finishing task: attempt_local1594728140_0001_r_000000_0
16/02/04 08:18:05 INFO mapred.LocalJobRunner: reduce task executor complete.
16/02/04 08:18:06 INFO mapreduce.Job:  map 100% reduce 100%
16/02/04 08:18:06 INFO mapreduce.Job: Job job_local1594728140_0001 completed successfully
16/02/04 08:18:06 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=572706
                FILE: Number of bytes written=1318772
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=285334
                HDFS: Number of bytes written=469371
                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=12593
                Map output records=12593
                Map output bytes=155260
                Map output materialized bytes=180452
                Input split bytes=93
                Combine input records=0
                Combine output records=0
                Reduce input groups=12593
                Reduce shuffle bytes=180452
                Reduce input records=12593
                Reduce output records=12592
                Spilled Records=25186
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=5
                Total committed heap usage (bytes)=492830720
        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=142667
        File Output Format Counters
                Bytes Written=469371
16/02/04 08:18:06 INFO streaming.StreamJob: Output directory: /user/hw3/output_3_3_out
```

```
In [235]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -cat /user/hw3/output_3_3_out/part-00000 | w
```

```
16/02/04 08:18:07 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
   12592
```

** The number of unique items is 12592 **

** The output below is the top 50 most encountered product ID's alongwith their frequencies and relative frequencies **

** The format is Product ID, Frequency, Relative Frequency, Total Number of ID's **

```
In [236]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -cat /user/hw3/output_3_3_out/part-00000 | h
```

```
16/02/04 08:18:10 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
DAI62779,6667,0.0175067747831,380824
FRO40251,3881,0.010191059387,380824
ELE17451,3875,0.0101753040775,380824
GRO73461,3602,0.00945843749344,380824
SNA80324,3044,0.00799319370628,380824
ELE32164,2851,0.0074863979161,380824
DAI75645,2736,0.00718442114993,380824
```

```
SNA45677,2455,0.0064465474865,380824
FRO31317,2330,0.0061183118711,380824
DAI85309,2293,0.00602115412894,380824
ELE26917,2292,0.00601852824402,380824
FRO80039,2233,0.00586360103355,380824
GRO21487,2115,0.00555374661261,380824
SNA99873,2083,0.00546971829507,380824
GRO59710,2004,0.00526227338613,380824
GRO71621,1920,0.00504169905258,380824
FRO85978,1918,0.00503644728273,380824
GRO30386,1840,0.00483162825872,380824
ELE74009,1816,0.00476860702057,380824
GRO56726,1784,0.00468457870302,380824
DAI63921,1773,0.00465569396887,380824
GRO46854,1756,0.00461105392517,380824
ELE66600,1713,0.00449814087347,380824
DAI83733,1712,0.00449551498855,380824
FRO32293,1702,0.00446925613932,380824
ELE66810,1697,0.0044561267147,380824
SNA55762,1646,0.00432220658362,380824
DAI22177,1627,0.00427231477008,380824
FRO78087,1531,0.00402022981745,380824
ELE99737,1516,0.0039808415436,380824
ELE34057,1489,0.00390994265067,380824
GRO94758,1489,0.00390994265067,380824
FRO35904,1436,0.00377077074974,380824
FRO53271,1420,0.00372875659097,380824
SNA93860,1407,0.00369462008697,380824
SNA90094,1390,0.00364998004327,380824
GRO38814,1352,0.00355019641619,380824
ELE56788,1345,0.00353181522173,380824
GRO61133,1321,0.00346879398357,380824
DAI88807,1316,0.00345566455896,380824
ELE74482,1316,0.00345566455896,380824
ELE59935,1311,0.00344253513434,380824
SNA96271,1295,0.00340052097557,380824
DAI43223,1290,0.00338739155095,380824
ELE91337,1289,0.00338476566603,380824
GRO15017,1275,0.0033480032771,380824
DAI31081,1261,0.00331124088818,380824
GRO81087,1220,0.00320357960633,380824
DAI22896,1219,0.0032009537214,380824
GRO85051,1214,0.00318782429679,380824
cat: Unable to write to output stream.
```

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

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

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

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

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

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

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

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

```
In [237]: %%writefile mapper.py
          #!/usr/bin/python
          import sys
          import re

          sys.stderr.write("reporter:counter:group,Num mapper calls,1\n")

          for line in sys.stdin:
              line = re.split(r'[\s]',line.strip())
              for l in line:
                  for p in line:
                      if(p > l):
                          print "%s,%s\t1" %(l,p)
                      elif (l > p):
                           print "%s,%s\t1" %(p,l)
                          #print "%s,*\t1" %l   Use this only if you want total number of tuples in the
              print "BASKET,*\t1" #Use this if you want basket in the denominator for support calculati
```

```
Overwriting mapper.py
```

```
In [238]: %%writefile reducer.py
          #!/usr/bin/python
          import sys
          import re

          count = 0
          prev_id = None

          sys.stderr.write("reporter:counter:group,Num reducer calls,1\n")

          for line in sys.stdin:
              line = re.split(r'[\t]',line.strip())
              if((prev_id !=None) and (line[0] !=prev_id)):
                  print "%s\t%s" %(prev_id,count)
                  count = 0
              count +=1
              prev_id = line[0]
          print "%s\t%s" %(prev_id,count)
```

```
Overwriting reducer.py
```

```
In [239]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -rm -r -f /user/hw3/output_3_4i
```

```
16/02/04 08:18:12 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
```

```
In [240]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hadoop jar /Users/Vamsi/Downloads/hadoop-2.7.1/bin/ha
          -D mapred.reduce.tasks=1 \
          -input /user/hw3/ProductPurchaseData.txt \
          -output /user/hw3/output_3_4i \
          -mapper mapper.py \
          -reducer reducer.py
```

```
16/02/04 08:18:14 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
16/02/04 08:18:15 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.ses
16/02/04 08:18:15 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/02/04 08:18:15 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessio
16/02/04 08:18:16 INFO mapred.FileInputFormat: Total input paths to process : 1
16/02/04 08:18:16 INFO mapreduce.JobSubmitter: number of splits:1
16/02/04 08:18:16 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapred
16/02/04 08:18:16 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1172650280_0001
16/02/04 08:18:16 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/02/04 08:18:16 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/02/04 08:18:16 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputComm
16/02/04 08:18:16 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:18:16 INFO mapreduce.Job: Running job: job_local1172650280_0001
16/02/04 08:18:16 INFO mapred.LocalJobRunner: Waiting for map tasks
16/02/04 08:18:16 INFO mapred.LocalJobRunner: Starting task: attempt_local1172650280_0001_m_000000_0
16/02/04 08:18:16 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:18:16 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only o
16/02/04 08:18:16 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:18:16 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/hw3/ProductPurchaseD
16/02/04 08:18:16 INFO mapred.MapTask: numReduceTasks: 1
16/02/04 08:18:16 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/02/04 08:18:16 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/02/04 08:18:16 INFO mapred.MapTask: soft limit at 83886080
16/02/04 08:18:16 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/02/04 08:18:16 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/02/04 08:18:16 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$Map
16/02/04 08:18:16 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./mapper.p
16/02/04 08:18:16 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.ta
16/02/04 08:18:16 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce
16/02/04 08:18:16 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.m
16/02/04 08:18:16 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.j
16/02/04 08:18:16 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce
16/02/04 08:18:16 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use map
16/02/04 08:18:16 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce
16/02/04 08:18:16 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.jo
16/02/04 08:18:16 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.us
16/02/04 08:18:16 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapredu
16/02/04 08:18:16 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.t
16/02/04 08:18:16 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapr
16/02/04 08:18:16 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:18:16 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:18:16 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:18:16 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:18:17 INFO streaming.PipeMapRed: Records R/W=2349/1
16/02/04 08:18:17 INFO mapreduce.Job: Job job_local1172650280_0001 running in uber mode : false
16/02/04 08:18:17 INFO mapreduce.Job:  map 0% reduce 0%
```

```
16/02/04 08:18:19 INFO streaming.PipeMapRed: R/W/S=10000/1631948/0 in:5000=10000/2 [rec/s] out:815981=1
16/02/04 08:18:20 INFO mapred.MapTask: Spilling map output
16/02/04 08:18:20 INFO mapred.MapTask: bufstart = 0; bufend = 46554754; bufvoid = 104857600
16/02/04 08:18:20 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 16881572(67526288); lengt
16/02/04 08:18:20 INFO mapred.MapTask: (EQUATOR) 56141730 kvi 14035428(56141712)
16/02/04 08:18:22 INFO mapred.LocalJobRunner: Records R/W=2349/1 > map
16/02/04 08:18:23 INFO mapreduce.Job:  map 43% reduce 0%
16/02/04 08:18:25 INFO mapred.MapTask: Finished spill 0
16/02/04 08:18:25 INFO mapred.MapTask: (RESET) equator 56141730 kv 14035428(56141712) kvi 11748300(4699
16/02/04 08:18:25 INFO mapred.LocalJobRunner: Records R/W=2349/1 > map
16/02/04 08:18:27 INFO streaming.PipeMapRed: Records R/W=25559/4091103
16/02/04 08:18:27 INFO mapred.MapTask: Spilling map output
16/02/04 08:18:27 INFO mapred.MapTask: bufstart = 56141730; bufend = 102681917; bufvoid = 104857600
16/02/04 08:18:27 INFO mapred.MapTask: kvstart = 14035428(56141712); kvend = 4698960(18795840); length
16/02/04 08:18:27 INFO mapred.MapTask: (EQUATOR) 7411277 kvi 1852812(7411248)
16/02/04 08:18:28 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:18:28 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:18:28 INFO mapred.LocalJobRunner: Records R/W=2349/1 > map
16/02/04 08:18:28 INFO mapred.MapTask: Starting flush of map output
16/02/04 08:18:28 INFO mapred.LocalJobRunner: Records R/W=25559/4091103 > sort
16/02/04 08:18:29 INFO mapreduce.Job:  map 67% reduce 0%
16/02/04 08:18:31 INFO mapred.LocalJobRunner: Records R/W=25559/4091103 > sort
16/02/04 08:18:32 INFO mapred.MapTask: Finished spill 1
16/02/04 08:18:32 INFO mapred.MapTask: (RESET) equator 7411277 kv 1852812(7411248) kvi 125276(501104)
16/02/04 08:18:32 INFO mapred.MapTask: Spilling map output
16/02/04 08:18:32 INFO mapred.MapTask: bufstart = 7411277; bufend = 16020607; bufvoid = 104857600
16/02/04 08:18:32 INFO mapred.MapTask: kvstart = 1852812(7411248); kvend = 125280(501120); length = 172
16/02/04 08:18:33 INFO mapred.MapTask: Finished spill 2
16/02/04 08:18:33 INFO mapred.Merger: Merging 3 sorted segments
16/02/04 08:18:33 INFO mapred.Merger: Down to the last merge-pass, with 3 segments left of total size:
16/02/04 08:18:34 INFO mapred.LocalJobRunner: Records R/W=25559/4091103 > sort >
16/02/04 08:18:35 INFO mapreduce.Job:  map 77% reduce 0%
16/02/04 08:18:37 INFO mapred.LocalJobRunner: Records R/W=25559/4091103 > sort >
16/02/04 08:18:37 INFO mapred.Task: Task:attempt_local1172650280_0001_m_000000_0 is done. And is in the p
16/02/04 08:18:37 INFO mapred.LocalJobRunner: Records R/W=25559/4091103 > sort
16/02/04 08:18:37 INFO mapred.Task: Task ’attempt_local1172650280_0001_m_000000_0’ done.
16/02/04 08:18:37 INFO mapred.LocalJobRunner: Finishing task: attempt_local1172650280_0001_m_000000_0
16/02/04 08:18:37 INFO mapred.LocalJobRunner: map task executor complete.
16/02/04 08:18:37 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/02/04 08:18:37 INFO mapred.LocalJobRunner: Starting task: attempt_local1172650280_0001_r_000000_0
16/02/04 08:18:37 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:18:37 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
16/02/04 08:18:37 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:18:37 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task
16/02/04 08:18:37 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleL
16/02/04 08:18:37 INFO reduce.EventFetcher: attempt_local1172650280_0001_r_000000_0 Thread started: Even
16/02/04 08:18:37 INFO reduce.MergeManagerImpl: attempt_local1172650280_0001_m_000000_0: Shuffling to di
16/02/04 08:18:37 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local
16/02/04 08:18:38 INFO mapreduce.Job:  map 100% reduce 0%
16/02/04 08:18:38 INFO reduce.OnDiskMapOutput: Read 111902695 bytes from map-output for attempt_local11
16/02/04 08:18:38 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/02/04 08:18:38 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/04 08:18:38 INFO reduce.MergeManagerImpl: finalMerge called with 0 in-memory map-outputs and 1 on
16/02/04 08:18:38 INFO reduce.MergeManagerImpl: Merging 1 files, 111902695 bytes from disk
```

```
16/02/04 08:18:38 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/02/04 08:18:38 INFO mapred.Merger: Merging 1 sorted segments
16/02/04 08:18:38 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size:
16/02/04 08:18:38 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/04 08:18:38 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./reducer
16/02/04 08:18:38 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapredu
16/02/04 08:18:38 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce
16/02/04 08:18:38 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:18:38 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:18:38 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:18:38 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:18:38 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:18:39 INFO streaming.PipeMapRed: Records R/W=46762/1
16/02/04 08:18:39 INFO streaming.PipeMapRed: R/W/S=100000/14685/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:18:39 INFO streaming.PipeMapRed: R/W/S=200000/35899/0 in:200000=200000/1 [rec/s] out:35899=
16/02/04 08:18:40 INFO streaming.PipeMapRed: R/W/S=300000/48905/0 in:300000=300000/1 [rec/s] out:48905=
16/02/04 08:18:40 INFO streaming.PipeMapRed: R/W/S=400000/70115/0 in:400000=400000/1 [rec/s] out:70115=
16/02/04 08:18:41 INFO streaming.PipeMapRed: R/W/S=500000/88855/0 in:250000=500000/2 [rec/s] out:44427=
16/02/04 08:18:41 INFO streaming.PipeMapRed: R/W/S=600000/110063/0 in:300000=600000/2 [rec/s] out:55031
16/02/04 08:18:41 INFO streaming.PipeMapRed: R/W/S=700000/123884/0 in:350000=700000/2 [rec/s] out:61942
16/02/04 08:18:42 INFO streaming.PipeMapRed: R/W/S=800000/143451/0 in:266666=800000/3 [rec/s] out:47817
16/02/04 08:18:42 INFO streaming.PipeMapRed: R/W/S=900000/163002/0 in:300000=900000/3 [rec/s] out:54334
16/02/04 08:18:42 INFO streaming.PipeMapRed: R/W/S=1000000/180931/0 in:333333=1000000/3 [rec/s] out:603
16/02/04 08:18:43 INFO streaming.PipeMapRed: R/W/S=1100000/192320/0 in:275000=1100000/4 [rec/s] out:480
16/02/04 08:18:43 INFO streaming.PipeMapRed: R/W/S=1200000/195530/0 in:300000=1200000/4 [rec/s] out:488
16/02/04 08:18:43 INFO streaming.PipeMapRed: R/W/S=1300000/211821/0 in:325000=1300000/4 [rec/s] out:529
16/02/04 08:18:43 INFO mapred.LocalJobRunner: Records R/W=46762/1 > reduce
16/02/04 08:18:44 INFO streaming.PipeMapRed: R/W/S=1400000/232202/0 in:280000=1400000/5 [rec/s] out:464
16/02/04 08:18:44 INFO streaming.PipeMapRed: R/W/S=1500000/244411/0 in:300000=1500000/5 [rec/s] out:488
16/02/04 08:18:44 INFO mapreduce.Job:  map 100% reduce 75%
16/02/04 08:18:44 INFO streaming.PipeMapRed: R/W/S=1600000/260696/0 in:320000=1600000/5 [rec/s] out:521
16/02/04 08:18:45 INFO streaming.PipeMapRed: R/W/S=1700000/272082/0 in:283333=1700000/6 [rec/s] out:453
16/02/04 08:18:45 INFO streaming.PipeMapRed: R/W/S=1800000/287541/0 in:300000=1800000/6 [rec/s] out:479
16/02/04 08:18:45 INFO streaming.PipeMapRed: R/W/S=1900000/305471/0 in:316666=1900000/6 [rec/s] out:509
16/02/04 08:18:46 INFO streaming.PipeMapRed: R/W/S=2000000/325841/0 in:285714=2000000/7 [rec/s] out:465
16/02/04 08:18:46 INFO streaming.PipeMapRed: R/W/S=2100000/342952/0 in:300000=2100000/7 [rec/s] out:489
16/02/04 08:18:46 INFO mapred.LocalJobRunner: Records R/W=46762/1 > reduce
16/02/04 08:18:46 INFO streaming.PipeMapRed: R/W/S=2200000/351879/0 in:275000=2200000/8 [rec/s] out:439
16/02/04 08:18:47 INFO streaming.PipeMapRed: R/W/S=2300000/369799/0 in:287500=2300000/8 [rec/s] out:462
16/02/04 08:18:47 INFO streaming.PipeMapRed: R/W/S=2400000/385985/0 in:300000=2400000/8 [rec/s] out:482
16/02/04 08:18:47 INFO mapreduce.Job:  map 100% reduce 81%
16/02/04 08:18:47 INFO streaming.PipeMapRed: R/W/S=2500000/402381/0 in:277777=2500000/9 [rec/s] out:447
16/02/04 08:18:48 INFO streaming.PipeMapRed: R/W/S=2600000/417845/0 in:288888=2600000/9 [rec/s] out:464
16/02/04 08:18:48 INFO streaming.PipeMapRed: R/W/S=2700000/436592/0 in:300000=2700000/9 [rec/s] out:485
16/02/04 08:18:49 INFO streaming.PipeMapRed: R/W/S=2800000/459441/0 in:280000=2800000/10 [rec/s] out:45
16/02/04 08:18:49 INFO streaming.PipeMapRed: Records R/W=2864552/470840
16/02/04 08:18:49 INFO streaming.PipeMapRed: R/W/S=2900000/477363/0 in:290000=2900000/10 [rec/s] out:47
16/02/04 08:18:49 INFO streaming.PipeMapRed: R/W/S=3000000/496102/0 in:300000=3000000/10 [rec/s] out:49
16/02/04 08:18:49 INFO mapred.LocalJobRunner: Records R/W=2864552/470840 > reduce
16/02/04 08:18:50 INFO streaming.PipeMapRed: R/W/S=3100000/509111/0 in:281818=3100000/11 [rec/s] out:46
16/02/04 08:18:50 INFO streaming.PipeMapRed: R/W/S=3200000/526214/0 in:290909=3200000/11 [rec/s] out:47
16/02/04 08:18:50 INFO mapreduce.Job:  map 100% reduce 87%
16/02/04 08:18:50 INFO streaming.PipeMapRed: R/W/S=3300000/549060/0 in:300000=3300000/11 [rec/s] out:49
16/02/04 08:18:51 INFO streaming.PipeMapRed: R/W/S=3400000/565357/0 in:283333=3400000/12 [rec/s] out:47
```

```
16/02/04 08:18:51 INFO streaming.PipeMapRed: R/W/S=3500000/587379/0 in:291666=3500000/12 [rec/s] out:48
16/02/04 08:18:52 INFO streaming.PipeMapRed: R/W/S=3600000/608575/0 in:276923=3600000/13 [rec/s] out:46
16/02/04 08:18:52 INFO streaming.PipeMapRed: R/W/S=3700000/624050/0 in:284615=3700000/13 [rec/s] out:48
16/02/04 08:18:52 INFO mapred.LocalJobRunner: Records R/W=2864552/470840 > reduce
16/02/04 08:18:53 INFO streaming.PipeMapRed: R/W/S=3800000/639533/0 in:271428=3800000/14 [rec/s] out:45
16/02/04 08:18:53 INFO streaming.PipeMapRed: R/W/S=3900000/658276/0 in:278571=3900000/14 [rec/s] out:47
16/02/04 08:18:53 INFO mapreduce.Job:  map 100% reduce 91%
16/02/04 08:18:53 INFO streaming.PipeMapRed: R/W/S=4000000/680307/0 in:285714=4000000/14 [rec/s] out:48
16/02/04 08:18:54 INFO streaming.PipeMapRed: R/W/S=4100000/700693/0 in:273333=4100000/15 [rec/s] out:46
16/02/04 08:18:54 INFO streaming.PipeMapRed: R/W/S=4200000/716164/0 in:280000=4200000/15 [rec/s] out:47
16/02/04 08:18:54 INFO streaming.PipeMapRed: R/W/S=4300000/734909/0 in:268750=4300000/16 [rec/s] out:45
16/02/04 08:18:55 INFO streaming.PipeMapRed: R/W/S=4400000/752830/0 in:275000=4400000/16 [rec/s] out:47
16/02/04 08:18:55 INFO streaming.PipeMapRed: R/W/S=4500000/769128/0 in:281250=4500000/16 [rec/s] out:48
16/02/04 08:18:55 INFO mapred.LocalJobRunner: Records R/W=2864552/470840 > reduce
16/02/04 08:18:55 INFO streaming.PipeMapRed: R/W/S=4600000/787054/0 in:270588=4600000/17 [rec/s] out:46
16/02/04 08:18:56 INFO streaming.PipeMapRed: R/W/S=4700000/804154/0 in:276470=4700000/17 [rec/s] out:47
16/02/04 08:18:56 INFO streaming.PipeMapRed: R/W/S=4800000/817996/0 in:282352=4800000/17 [rec/s] out:48
16/02/04 08:18:56 INFO mapreduce.Job:  map 100% reduce 97%
16/02/04 08:18:57 INFO streaming.PipeMapRed: R/W/S=4900000/836739/0 in:272222=4900000/18 [rec/s] out:46
16/02/04 08:18:57 INFO streaming.PipeMapRed: R/W/S=5000000/857947/0 in:277777=5000000/18 [rec/s] out:47
16/02/04 08:18:57 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:18:57 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:18:57 INFO mapred.Task: Task:attempt_local1172650280_0001_r_000000_0 is done. And is in the p
16/02/04 08:18:57 INFO mapred.LocalJobRunner: Records R/W=2864552/470840 > reduce
16/02/04 08:18:57 INFO mapred.Task: Task attempt_local1172650280_0001_r_000000_0 is allowed to commit now
16/02/04 08:18:57 INFO output.FileOutputCommitter: Saved output of task ’attempt_local1172650280_0001_r_
16/02/04 08:18:57 INFO mapred.LocalJobRunner: Records R/W=2864552/470840 > reduce
16/02/04 08:18:57 INFO mapred.Task: Task ’attempt_local1172650280_0001_r_000000_0’ done.
16/02/04 08:18:57 INFO mapred.LocalJobRunner: Finishing task: attempt_local1172650280_0001_r_000000_0
16/02/04 08:18:57 INFO mapred.LocalJobRunner: reduce task executor complete.
16/02/04 08:18:58 INFO mapreduce.Job:  map 100% reduce 100%
16/02/04 08:18:58 INFO mapreduce.Job: Job job_local1172650280_0001 completed successfully
16/02/04 08:18:58 INFO mapreduce.Job: Counters: 37
        File System Counters
                FILE: Number of bytes read=447822632
                FILE: Number of bytes written=560286613
                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=17637495
                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=5099209
                Map output bytes=101704271
                Map output materialized bytes=111902695
                Input split bytes=106
                Combine input records=0
                Combine output records=0
                Reduce input groups=877096
                Reduce shuffle bytes=111902695
                Reduce input records=5099209
                Reduce output records=877096
                Spilled Records=15297627
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=14
                Total committed heap usage (bytes)=605552640
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        group
                Num mapper calls=1
                Num reducer calls=1
        File Input Format Counters
                Bytes Read=3458517
        File Output Format Counters
                Bytes Written=17637495
16/02/04 08:18:58 INFO streaming.StreamJob: Output directory: /user/hw3/output_3_4i
```

```
In [241]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -cat /user/hw3/output_3_4i/part-00000 > pai
```

```
16/02/04 08:19:00 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
```

```
In [242]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -put pairs_unsorted /user/hw3
```

```
16/02/04 08:19:02 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
```

This second Map-reduce job is used to sort the pairs

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

          sys.stderr.write("reporter:counter:group,Num mapper calls,1\n")

          for line in sys.stdin:
              line = re.split(r'[\t]',line.strip())
              print "%s,%s" %(line[0],line[1])
```

```
Overwriting mapper.py
```

```
In [244]: %%writefile reducer.py
          #!/usr/bin/python
          import sys
          import re

          sys.stderr.write("reporter:counter:group,Num reducer calls,1\n")

          for line in sys.stdin:
              line = re.split(r'[,]',line.strip())
              if(line[1]=="*"):
                  total = int(line[2])
              else:
                  rel_freq = float(line[2])/float(total)
                  print "%s,%s,%s,%s,%s" %(line[0],line[1],line[2],rel_freq,total)
```

```
Overwriting reducer.py
```

In [245]:
```
!/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hadoop jar /Users/Vamsi/Downloads/hadoop-2.7.1/bin/ha
-D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator \
-D stream.map.output.field.separator=, \
-D stream.num.map.output.key.fields=3 \
-D map.output.key.field.separator=, \
-D mapred.text.key.comparator.options=-k3,3nr \
-D mapred.reduce.tasks=1 \
-input /user/hw3/pairs_unsorted \
-output /user/hw3/output_3_4_out \
-mapper mapper.py \
-reducer reducer.py
```

```
16/02/04 08:19:05 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
16/02/04 08:19:06 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.ses
16/02/04 08:19:06 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/02/04 08:19:06 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessio
16/02/04 08:19:06 INFO mapred.FileInputFormat: Total input paths to process : 1
16/02/04 08:19:06 INFO mapreduce.JobSubmitter: number of splits:1
16/02/04 08:19:06 INFO Configuration.deprecation: map.output.key.field.separator is deprecated. Instead
16/02/04 08:19:06 INFO Configuration.deprecation: mapred.text.key.comparator.options is deprecated. Inst
16/02/04 08:19:06 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapred
16/02/04 08:19:06 INFO Configuration.deprecation: mapred.output.key.comparator.class is deprecated. Inst
16/02/04 08:19:07 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local774545073_0001
16/02/04 08:19:07 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/02/04 08:19:07 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/02/04 08:19:07 INFO mapreduce.Job: Running job: job_local774545073_0001
16/02/04 08:19:07 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputComm
16/02/04 08:19:07 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:19:07 INFO mapred.LocalJobRunner: Waiting for map tasks
16/02/04 08:19:07 INFO mapred.LocalJobRunner: Starting task: attempt_local774545073_0001_m_000000_0
16/02/04 08:19:07 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:19:07 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
16/02/04 08:19:07 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:19:07 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/hw3/pairs_unsorted:
16/02/04 08:19:07 INFO mapred.MapTask: numReduceTasks: 1
16/02/04 08:19:07 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/02/04 08:19:07 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/02/04 08:19:07 INFO mapred.MapTask: soft limit at 83886080
16/02/04 08:19:07 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/02/04 08:19:07 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/02/04 08:19:07 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$Ma
16/02/04 08:19:07 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./mapper.
16/02/04 08:19:07 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use ma
16/02/04 08:19:07 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce
16/02/04 08:19:07 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapred
16/02/04 08:19:07 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.
```

```
16/02/04 08:19:07 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.ta
16/02/04 08:19:07 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce
16/02/04 08:19:07 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.
16/02/04 08:19:07 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.
16/02/04 08:19:07 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce
16/02/04 08:19:07 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.jo
16/02/04 08:19:07 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.us
16/02/04 08:19:07 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use map
16/02/04 08:19:07 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:19:07 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:19:07 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:19:07 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:19:07 INFO streaming.PipeMapRed: Records R/W=6532/1
16/02/04 08:19:07 INFO streaming.PipeMapRed: R/W/S=10000/5717/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:19:08 INFO mapreduce.Job: Job job_local774545073_0001 running in uber mode : false
16/02/04 08:19:08 INFO mapreduce.Job:  map 0% reduce 0%
16/02/04 08:19:08 INFO streaming.PipeMapRed: R/W/S=100000/92927/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:19:08 INFO streaming.PipeMapRed: R/W/S=200000/191519/0 in:200000=200000/1 [rec/s] out:19151
16/02/04 08:19:09 INFO streaming.PipeMapRed: R/W/S=300000/294871/0 in:300000=300000/1 [rec/s] out:29487
16/02/04 08:19:09 INFO streaming.PipeMapRed: R/W/S=400000/392616/0 in:200000=400000/2 [rec/s] out:19630
16/02/04 08:19:10 INFO streaming.PipeMapRed: R/W/S=500000/491207/0 in:250000=500000/2 [rec/s] out:24560
16/02/04 08:19:10 INFO streaming.PipeMapRed: R/W/S=600000/594719/0 in:300000=600000/2 [rec/s] out:29735
16/02/04 08:19:11 INFO streaming.PipeMapRed: R/W/S=700000/693360/0 in:233333=700000/3 [rec/s] out:23112
16/02/04 08:19:11 INFO streaming.PipeMapRed: R/W/S=800000/791124/0 in:266666=800000/3 [rec/s] out:26370
16/02/04 08:19:11 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:19:11 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:19:11 INFO mapred.LocalJobRunner:
16/02/04 08:19:11 INFO mapred.MapTask: Starting flush of map output
16/02/04 08:19:11 INFO mapred.MapTask: Spilling map output
16/02/04 08:19:11 INFO mapred.MapTask: bufstart = 0; bufend = 18514591; bufvoid = 104857600
16/02/04 08:19:11 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 22706016(90824064); lengt
16/02/04 08:19:12 INFO mapred.MapTask: Finished spill 0
16/02/04 08:19:12 INFO mapred.Task: Task:attempt_local774545073_0001_m_000000_0 is done. And is in the pr
16/02/04 08:19:12 INFO mapred.LocalJobRunner: Records R/W=6532/1
16/02/04 08:19:12 INFO mapred.Task: Task ’attempt_local774545073_0001_m_000000_0’ done.
16/02/04 08:19:12 INFO mapred.LocalJobRunner: Finishing task: attempt_local774545073_0001_m_000000_0
16/02/04 08:19:12 INFO mapred.LocalJobRunner: map task executor complete.
16/02/04 08:19:12 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/02/04 08:19:12 INFO mapred.LocalJobRunner: Starting task: attempt_local774545073_0001_r_000000_0
16/02/04 08:19:13 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:19:13 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
16/02/04 08:19:13 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:19:13 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task
16/02/04 08:19:13 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLi
16/02/04 08:19:13 INFO reduce.EventFetcher: attempt_local774545073_0001_r_000000_0 Thread started: EventF
16/02/04 08:19:13 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local7
16/02/04 08:19:13 INFO reduce.InMemoryMapOutput: Read 20268785 bytes from map-output for attempt_local77
16/02/04 08:19:13 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 20268785, inMem
16/02/04 08:19:13 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/02/04 08:19:13 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/04 08:19:13 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-
16/02/04 08:19:13 INFO mapred.Merger: Merging 1 sorted segments
16/02/04 08:19:13 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size:
16/02/04 08:19:13 INFO mapreduce.Job:  map 100% reduce 0%
```

```
16/02/04 08:19:13 INFO reduce.MergeManagerImpl: Merged 1 segments, 20268785 bytes to disk to satisfy re
16/02/04 08:19:13 INFO reduce.MergeManagerImpl: Merging 1 files, 20268789 bytes from disk
16/02/04 08:19:13 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/02/04 08:19:13 INFO mapred.Merger: Merging 1 sorted segments
16/02/04 08:19:13 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 
16/02/04 08:19:13 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/04 08:19:13 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./reducer
16/02/04 08:19:13 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapredu
16/02/04 08:19:13 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduc
16/02/04 08:19:13 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:19:13 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:19:13 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:19:13 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:19:13 INFO streaming.PipeMapRed: Records R/W=9316/1
16/02/04 08:19:13 INFO streaming.PipeMapRed: R/W/S=10000/56/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:19:14 INFO streaming.PipeMapRed: R/W/S=100000/89070/0 in:100000=100000/1 [rec/s] out:89086=
16/02/04 08:19:15 INFO streaming.PipeMapRed: R/W/S=200000/188965/0 in:200000=200000/1 [rec/s] out:18896
16/02/04 08:19:16 INFO streaming.PipeMapRed: R/W/S=300000/288793/0 in:150000=300000/2 [rec/s] out:14439
16/02/04 08:19:17 INFO streaming.PipeMapRed: R/W/S=400000/388315/0 in:133333=400000/3 [rec/s] out:12943
16/02/04 08:19:18 INFO streaming.PipeMapRed: R/W/S=500000/488481/0 in:125000=500000/4 [rec/s] out:12212
16/02/04 08:19:18 INFO streaming.PipeMapRed: R/W/S=600000/588274/0 in:120000=600000/5 [rec/s] out:11765
16/02/04 08:19:19 INFO mapred.LocalJobRunner: Records R/W=9316/1 > reduce
16/02/04 08:19:19 INFO mapreduce.Job:  map 100% reduce 90%
16/02/04 08:19:19 INFO streaming.PipeMapRed: R/W/S=700000/688068/0 in:116666=700000/6 [rec/s] out:11467
16/02/04 08:19:20 INFO streaming.PipeMapRed: R/W/S=800000/787861/0 in:133333=800000/6 [rec/s] out:13131
16/02/04 08:19:21 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:19:21 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:19:21 INFO mapred.Task: Task:attempt_local774545073_0001_r_000000_0 is done. And is in the pr
16/02/04 08:19:21 INFO mapred.LocalJobRunner: Records R/W=9316/1 > reduce
16/02/04 08:19:21 INFO mapred.Task: Task attempt_local774545073_0001_r_000000_0 is allowed to commit now
16/02/04 08:19:21 INFO output.FileOutputCommitter: Saved output of task 'attempt_local774545073_0001_r_0
16/02/04 08:19:21 INFO mapred.LocalJobRunner: Records R/W=9316/1 > reduce
16/02/04 08:19:21 INFO mapred.Task: Task 'attempt_local774545073_0001_r_000000_0' done.
16/02/04 08:19:21 INFO mapred.LocalJobRunner: Finishing task: attempt_local774545073_0001_r_000000_0
16/02/04 08:19:21 INFO mapred.LocalJobRunner: reduce task executor complete.
16/02/04 08:19:22 INFO mapreduce.Job:  map 100% reduce 100%
16/02/04 08:19:22 INFO mapreduce.Job: Job job_local774545073_0001 completed successfully
16/02/04 08:19:22 INFO mapreduce.Job: Counters: 37
        File System Counters
                FILE: Number of bytes read=40749390
                FILE: Number of bytes written=61580781
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=35274990
                HDFS: Number of bytes written=39406367
                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=877096
                Map output records=877096
                Map output bytes=18514591
                Map output materialized bytes=20268789
                Input split bytes=97
                Combine input records=0
                Combine output records=0
                Reduce input groups=877096
                Reduce shuffle bytes=20268789
                Reduce input records=877096
                Reduce output records=877095
                Spilled Records=1754192
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=33
                Total committed heap usage (bytes)=567279616
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        group
                Num mapper calls=1
                Num reducer calls=1
        File Input Format Counters
                Bytes Read=17637495
        File Output Format Counters
                Bytes Written=39406367
16/02/04 08:19:22 INFO streaming.StreamJob: Output directory: /user/hw3/output_3_4_out
```

** Computational Setup : **

```
MacBook Air
Processor : 1.8 Ghz Intel Core i5 , 2 Cores
Memory : 4 GB 1600 Mhz DDR3
Disk : SSD (Flash Storage) 128 GB
```

The output is shown below
Format = Pairs, Support Count, Support, Total number of baskets

```
In [246]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -cat /user/hw3/output_3_4_out/part-00000 | I
```

```
16/02/04 08:19:23 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
DAI62779,ELE17451,3184,0.102376129385,31101
FRO40251,SNA80324,2824,0.0908009388766,31101
DAI75645,FRO40251,2508,0.0806404938748,31101
FRO40251,GRO85051,2426,0.0780039227035,31101
DAI62779,GRO73461,2278,0.0732452332722,31101
DAI75645,SNA80324,2260,0.0726664737468,31101
DAI62779,FRO40251,2140,0.0688080769107,31101
DAI62779,SNA80324,1846,0.0593550046622,31101
DAI62779,DAI85309,1836,0.0590334715926,31101
ELE32164,GRO59710,1822,0.058583325295,31101
FRO40251,GRO73461,1764,0.0567184334909,31101
DAI62779,DAI75645,1764,0.0567184334909,31101
DAI62779,ELE92920,1754,0.0563969004212,31101
```

```
FRO40251,FRO92469,1670,0.0536960226359,31101
DAI62779,ELE32164,1664,0.0535031027941,31101
DAI75645,GRO73461,1424,0.0457863091219,31101
DAI43223,ELE32164,1422,0.045722002508,31101
DAI62779,GRO30386,1418,0.0455933892801,31101
ELE17451,FRO40251,1394,0.0448217099129,31101
DAI85309,ELE99737,1318,0.0423780585833,31101
DAI62779,ELE26917,1300,0.0417992990579,31101
GRO21487,GRO73461,1262,0.0405774733931,31101
DAI62779,SNA45677,1208,0.0388411948169,31101
ELE17451,SNA80324,1194,0.0383910485193,31101
DAI62779,GRO71621,1190,0.0382624352915,31101
DAI62779,SNA55762,1186,0.0381338220636,31101
DAI62779,DAI83733,1172,0.0376836757661,31101
ELE17451,GRO73461,1160,0.0372978360824,31101
GRO73461,SNA80324,1124,0.0361403170316,31101
DAI62779,GRO59710,1122,0.0360760104177,31101
DAI62779,FRO80039,1100,0.0353686376644,31101
DAI75645,ELE17451,1094,0.0351757178226,31101
DAI62779,SNA93860,1074,0.0345326516832,31101
DAI55148,DAI62779,1052,0.0338252789299,31101
DAI43223,GRO59710,1024,0.0329249863348,31101
ELE17451,ELE32164,1022,0.0328606797209,31101
DAI62779,SNA18336,1012,0.0325391466512,31101
ELE32164,GRO73461,972,0.0312530143725,31101
DAI62779,FRO78087,964,0.0309957879168,31101
DAI85309,ELE17451,964,0.0309957879168,31101
DAI62779,GRO94758,958,0.030802868075,31101
DAI62779,GRO21487,942,0.0302884151635,31101
GRO85051,SNA80324,942,0.0302884151635,31101
ELE17451,GRO30386,936,0.0300954953217,31101
FRO85978,SNA95666,926,0.029773962252,31101
DAI62779,FRO19221,924,0.0297096556381,31101
DAI62779,GRO46854,922,0.0296453490241,31101
DAI43223,DAI62779,918,0.0295167357963,31101
ELE92920,SNA18336,910,0.0292595093405,31101
DAI88079,FRO40251,892,0.0286807498151,31101
cat: Unable to write to output stream.
```

0.0.13 HW3.5: Stripes

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

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

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

```
In [247]: %%writefile mapper.py
          #!/usr/bin/python
          import sys
          import re

          sys.stderr.write("reporter:counter:group,Num mapper calls,1\n")

          for line in sys.stdin:
              line = re.split(r'[\s]',line.strip())

              for l in line:
                  co_stripe = {}
                  for p in line:
                      if(l != p):
                          if p not in co_stripe.keys():
                              co_stripe[p] = 1
                          else:
                              co_stripe[p] += 1
                  print "%s\t%s" %(l,co_stripe)
                          #print "%s,*\t1" %l   Use this only if you want total number of tuples in the
              print "BASKET\t{'*':1}" #Use this if you want basket in the denominator for support calcul
```

```
Overwriting mapper.py
```

```
In [248]: %%writefile reducer.py
          #!/usr/bin/python
          import sys
          import re
          import ast

          count = {}
          prev_id = None

          sys.stderr.write("reporter:counter:group,Num mapper calls,1\n")

          for line in sys.stdin:
              line = re.split(r'[\t]',line.strip())
              co_strip = ast.literal_eval(line[1])
              if((prev_id !=None) and (line[0] !=prev_id)):
                  for w in count.keys():
                      if(prev_id < w):
                          print "%s,%s\t%s" %(prev_id,w,count[w])
                      elif(prev_id >w):
                          print "%s,%s\t%s" %(w,prev_id,count[w])
                  count = {}
              for w in co_strip:
                  if w in count.keys():
                      count[w] +=1
                  else:
                      count[w] = 1
              prev_id = line[0]

          for w in count.keys():
              if(prev_id <w):
                  print "%s,%s\t%s" %(prev_id,w,count[w])
              elif(prev_id >w):
                  print "%s,%s\t%s" %(w,prev_id,count[w])
```

```
Overwriting reducer.py
```

```
In [249]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hadoop jar /Users/Vamsi/Downloads/hadoop-2.7.1/bin/ha
          -D mapred.reduce.tasks=1 \
          -input /user/hw3/ProductPurchaseData.txt \
          -output /user/hw3/output_3_5i \
          -mapper mapper.py \
          -reducer reducer.py

16/02/04 08:19:26 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
16/02/04 08:19:27 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.ses
16/02/04 08:19:27 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/02/04 08:19:27 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessio
16/02/04 08:19:27 INFO mapred.FileInputFormat: Total input paths to process : 1
16/02/04 08:19:28 INFO mapreduce.JobSubmitter: number of splits:1
16/02/04 08:19:28 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapred
16/02/04 08:19:28 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local112720032_0001
16/02/04 08:19:28 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/02/04 08:19:28 INFO mapreduce.Job: Running job: job_local112720032_0001
16/02/04 08:19:28 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/02/04 08:19:28 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputComm
16/02/04 08:19:28 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:19:28 INFO mapred.LocalJobRunner: Starting task: attempt_local112720032_0001_m_000000_0
16/02/04 08:19:28 INFO mapred.LocalJobRunner: Waiting for map tasks
16/02/04 08:19:28 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:19:28 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only o
16/02/04 08:19:28 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:19:28 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/hw3/ProductPurchaseD
16/02/04 08:19:28 INFO mapred.MapTask: numReduceTasks: 1
16/02/04 08:19:28 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/02/04 08:19:28 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/02/04 08:19:28 INFO mapred.MapTask: soft limit at 83886080
16/02/04 08:19:28 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/02/04 08:19:28 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/02/04 08:19:28 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$Map
16/02/04 08:19:28 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./mapper.p
16/02/04 08:19:28 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.ta
16/02/04 08:19:28 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce
16/02/04 08:19:28 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.m
16/02/04 08:19:28 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.j
16/02/04 08:19:28 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce
16/02/04 08:19:28 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use map
16/02/04 08:19:28 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce
16/02/04 08:19:28 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.jo
16/02/04 08:19:28 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.us
16/02/04 08:19:28 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapredu
16/02/04 08:19:28 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.t
16/02/04 08:19:28 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapr
16/02/04 08:19:28 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:19:28 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:19:28 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:19:28 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:19:28 INFO streaming.PipeMapRed: Records R/W=1216/1
16/02/04 08:19:29 INFO mapreduce.Job: Job job_local112720032_0001 running in uber mode : false
16/02/04 08:19:29 INFO mapreduce.Job:  map 0% reduce 0%
16/02/04 08:19:31 INFO streaming.PipeMapRed: R/W/S=10000/129308/0 in:5000=10000/2 [rec/s] out:64654=1293
16/02/04 08:19:34 INFO mapred.LocalJobRunner: Records R/W=1216/1 > map
16/02/04 08:19:35 INFO mapreduce.Job:  map 53% reduce 0%
16/02/04 08:19:35 INFO mapred.MapTask: Spilling map output
```

```
16/02/04 08:19:35 INFO mapred.MapTask: bufstart = 0; bufend = 77559378; bufvoid = 104857600
16/02/04 08:19:35 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 24632720(98530880); length
16/02/04 08:19:35 INFO mapred.MapTask: (EQUATOR) 79142130 kvi 19785528(79142112)
16/02/04 08:19:36 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:19:36 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:19:36 INFO mapred.LocalJobRunner: Records R/W=1216/1 > map
16/02/04 08:19:36 INFO mapred.MapTask: Starting flush of map output
16/02/04 08:19:37 INFO mapred.LocalJobRunner: Records R/W=1216/1 > sort
16/02/04 08:19:37 INFO mapred.MapTask: Finished spill 0
16/02/04 08:19:37 INFO mapred.MapTask: (RESET) equator 79142130 kv 19785528(79142112) kvi 19719508(7887
16/02/04 08:19:37 INFO mapred.MapTask: Spilling map output
16/02/04 08:19:37 INFO mapred.MapTask: bufstart = 79142130; bufend = 82263698; bufvoid = 104857600
16/02/04 08:19:37 INFO mapred.MapTask: kvstart = 19785528(79142112); kvend = 19719512(78878048); length
16/02/04 08:19:37 INFO mapred.MapTask: Finished spill 1
16/02/04 08:19:37 INFO mapred.Merger: Merging 2 sorted segments
16/02/04 08:19:37 INFO mapred.Merger: Down to the last merge-pass, with 2 segments left of total size: 
16/02/04 08:19:38 INFO mapreduce.Job:  map 67% reduce 0%
16/02/04 08:19:39 INFO mapred.Task: Task:attempt_local112720032_0001_m_000000_0 is done. And is in the pr
16/02/04 08:19:39 INFO mapred.LocalJobRunner: Records R/W=1216/1 > sort
16/02/04 08:19:39 INFO mapred.Task: Task 'attempt_local112720032_0001_m_000000_0' done.
16/02/04 08:19:39 INFO mapred.LocalJobRunner: Finishing task: attempt_local112720032_0001_m_000000_0
16/02/04 08:19:39 INFO mapred.LocalJobRunner: map task executor complete.
16/02/04 08:19:39 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/02/04 08:19:39 INFO mapred.LocalJobRunner: Starting task: attempt_local112720032_0001_r_000000_0
16/02/04 08:19:39 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:19:39 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only 
16/02/04 08:19:39 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:19:39 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task
16/02/04 08:19:39 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleL
16/02/04 08:19:39 INFO reduce.EventFetcher: attempt_local112720032_0001_r_000000_0 Thread started: Event
16/02/04 08:19:39 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local
16/02/04 08:19:39 INFO reduce.InMemoryMapOutput: Read 81909481 bytes from map-output for attempt_local1
16/02/04 08:19:39 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 81909481, inMe
16/02/04 08:19:39 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/02/04 08:19:39 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/04 08:19:39 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on
16/02/04 08:19:39 INFO mapred.Merger: Merging 1 sorted segments
16/02/04 08:19:39 INFO mapreduce.Job:  map 100% reduce 0%
16/02/04 08:19:39 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 
16/02/04 08:19:40 INFO reduce.MergeManagerImpl: Merged 1 segments, 81909481 bytes to disk to satisfy re
16/02/04 08:19:40 INFO reduce.MergeManagerImpl: Merging 1 files, 81909485 bytes from disk
16/02/04 08:19:40 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/02/04 08:19:40 INFO mapred.Merger: Merging 1 sorted segments
16/02/04 08:19:40 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 
16/02/04 08:19:40 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/04 08:19:40 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./reducer
16/02/04 08:19:40 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapred
16/02/04 08:19:40 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduc
16/02/04 08:19:40 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:19:40 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:19:40 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:19:40 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:19:40 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:19:41 INFO streaming.PipeMapRed: Records R/W=32019/1
```

```
16/02/04 08:19:45 INFO mapred.LocalJobRunner: Records R/W=32019/1 > reduce
16/02/04 08:19:45 INFO mapreduce.Job:  map 100% reduce 68%
16/02/04 08:19:48 INFO mapred.LocalJobRunner: Records R/W=32019/1 > reduce
16/02/04 08:19:51 INFO mapred.LocalJobRunner: Records R/W=32019/1 > reduce
16/02/04 08:19:51 INFO mapreduce.Job:  map 100% reduce 69%
16/02/04 08:19:51 INFO streaming.PipeMapRed: Records R/W=57237/121015
16/02/04 08:19:54 INFO mapred.LocalJobRunner: Records R/W=57237/121015 > reduce
16/02/04 08:19:54 INFO mapreduce.Job:  map 100% reduce 70%
16/02/04 08:19:57 INFO mapred.LocalJobRunner: Records R/W=57237/121015 > reduce
16/02/04 08:20:00 INFO mapred.LocalJobRunner: Records R/W=57237/121015 > reduce
16/02/04 08:20:03 INFO mapred.LocalJobRunner: Records R/W=57237/121015 > reduce
16/02/04 08:20:06 INFO streaming.PipeMapRed: Records R/W=79275/202758
16/02/04 08:20:06 INFO mapred.LocalJobRunner: Records R/W=79275/202758 > reduce
16/02/04 08:20:06 INFO mapreduce.Job:  map 100% reduce 71%
16/02/04 08:20:09 INFO mapred.LocalJobRunner: Records R/W=79275/202758 > reduce
16/02/04 08:20:09 INFO mapreduce.Job:  map 100% reduce 72%
16/02/04 08:20:12 INFO mapred.LocalJobRunner: Records R/W=79275/202758 > reduce
16/02/04 08:20:15 INFO mapred.LocalJobRunner: Records R/W=79275/202758 > reduce
16/02/04 08:20:18 INFO mapred.LocalJobRunner: Records R/W=79275/202758 > reduce
16/02/04 08:20:18 INFO streaming.PipeMapRed: Records R/W=99022/286063
16/02/04 08:20:18 INFO streaming.PipeMapRed: R/W/S=100000/295034/0 in:2702=100000/37 [rec/s] out:7973=29
16/02/04 08:20:18 INFO mapreduce.Job:  map 100% reduce 73%
16/02/04 08:20:21 INFO mapred.LocalJobRunner: Records R/W=99022/286063 > reduce
16/02/04 08:20:24 INFO mapred.LocalJobRunner: Records R/W=99022/286063 > reduce
16/02/04 08:20:24 INFO mapreduce.Job:  map 100% reduce 74%
16/02/04 08:20:27 INFO mapred.LocalJobRunner: Records R/W=99022/286063 > reduce
16/02/04 08:20:27 INFO mapreduce.Job:  map 100% reduce 75%
16/02/04 08:20:30 INFO mapred.LocalJobRunner: Records R/W=99022/286063 > reduce
16/02/04 08:20:32 INFO streaming.PipeMapRed: Records R/W=124029/388200
16/02/04 08:20:33 INFO mapred.LocalJobRunner: Records R/W=124029/388200 > reduce
16/02/04 08:20:36 INFO mapred.LocalJobRunner: Records R/W=124029/388200 > reduce
16/02/04 08:20:36 INFO mapreduce.Job:  map 100% reduce 76%
16/02/04 08:20:39 INFO mapred.LocalJobRunner: Records R/W=124029/388200 > reduce
16/02/04 08:20:42 INFO mapred.LocalJobRunner: Records R/W=124029/388200 > reduce
16/02/04 08:20:42 INFO streaming.PipeMapRed: Records R/W=141329/468241
16/02/04 08:20:45 INFO mapred.LocalJobRunner: Records R/W=141329/468241 > reduce
16/02/04 08:20:45 INFO mapreduce.Job:  map 100% reduce 77%
16/02/04 08:20:48 INFO mapred.LocalJobRunner: Records R/W=141329/468241 > reduce
16/02/04 08:20:48 INFO mapreduce.Job:  map 100% reduce 78%
16/02/04 08:20:51 INFO mapred.LocalJobRunner: Records R/W=141329/468241 > reduce
16/02/04 08:20:52 INFO streaming.PipeMapRed: Records R/W=165275/582690
16/02/04 08:20:54 INFO mapred.LocalJobRunner: Records R/W=165275/582690 > reduce
16/02/04 08:20:54 INFO mapreduce.Job:  map 100% reduce 79%
16/02/04 08:20:57 INFO mapred.LocalJobRunner: Records R/W=165275/582690 > reduce
16/02/04 08:21:00 INFO mapred.LocalJobRunner: Records R/W=165275/582690 > reduce
16/02/04 08:21:00 INFO mapreduce.Job:  map 100% reduce 80%
16/02/04 08:21:03 INFO streaming.PipeMapRed: Records R/W=185676/681585
16/02/04 08:21:03 INFO mapred.LocalJobRunner: Records R/W=185676/681585 > reduce
16/02/04 08:21:06 INFO mapred.LocalJobRunner: Records R/W=185676/681585 > reduce
16/02/04 08:21:06 INFO mapreduce.Job:  map 100% reduce 81%
16/02/04 08:21:08 INFO streaming.PipeMapRed: R/W/S=200000/755990/0 in:2272=200000/88 [rec/s] out:8590=75
16/02/04 08:21:09 INFO mapred.LocalJobRunner: Records R/W=185676/681585 > reduce
16/02/04 08:21:09 INFO mapreduce.Job:  map 100% reduce 82%
16/02/04 08:21:12 INFO mapred.LocalJobRunner: Records R/W=185676/681585 > reduce
```

```
16/02/04 08:21:13 INFO streaming.PipeMapRed: Records R/W=210419/800158
16/02/04 08:21:15 INFO mapred.LocalJobRunner: Records R/W=210419/800158 > reduce
16/02/04 08:21:15 INFO mapreduce.Job:  map 100% reduce 83%
16/02/04 08:21:18 INFO mapred.LocalJobRunner: Records R/W=210419/800158 > reduce
16/02/04 08:21:21 INFO mapred.LocalJobRunner: Records R/W=210419/800158 > reduce
16/02/04 08:21:23 INFO streaming.PipeMapRed: Records R/W=221447/844241
16/02/04 08:21:24 INFO mapred.LocalJobRunner: Records R/W=221447/844241 > reduce
16/02/04 08:21:24 INFO mapreduce.Job:  map 100% reduce 84%
16/02/04 08:21:27 INFO mapred.LocalJobRunner: Records R/W=221447/844241 > reduce
16/02/04 08:21:30 INFO mapred.LocalJobRunner: Records R/W=221447/844241 > reduce
16/02/04 08:21:30 INFO mapreduce.Job:  map 100% reduce 85%
16/02/04 08:21:33 INFO mapred.LocalJobRunner: Records R/W=221447/844241 > reduce
16/02/04 08:21:33 INFO mapreduce.Job:  map 100% reduce 86%
16/02/04 08:21:35 INFO streaming.PipeMapRed: Records R/W=248945/984896
16/02/04 08:21:36 INFO mapred.LocalJobRunner: Records R/W=248945/984896 > reduce
16/02/04 08:21:39 INFO mapred.LocalJobRunner: Records R/W=248945/984896 > reduce
16/02/04 08:21:39 INFO mapreduce.Job:  map 100% reduce 87%
16/02/04 08:21:42 INFO mapred.LocalJobRunner: Records R/W=248945/984896 > reduce
16/02/04 08:21:45 INFO mapred.LocalJobRunner: Records R/W=248945/984896 > reduce
16/02/04 08:21:46 INFO streaming.PipeMapRed: Records R/W=269514/1087881
16/02/04 08:21:48 INFO mapred.LocalJobRunner: Records R/W=269514/1087881 > reduce
16/02/04 08:21:48 INFO mapreduce.Job:  map 100% reduce 88%
16/02/04 08:21:51 INFO mapred.LocalJobRunner: Records R/W=269514/1087881 > reduce
16/02/04 08:21:54 INFO mapred.LocalJobRunner: Records R/W=269514/1087881 > reduce
16/02/04 08:21:54 INFO mapreduce.Job:  map 100% reduce 89%
16/02/04 08:21:56 INFO streaming.PipeMapRed: Records R/W=286340/1164718
16/02/04 08:21:57 INFO mapred.LocalJobRunner: Records R/W=286340/1164718 > reduce
16/02/04 08:22:00 INFO mapred.LocalJobRunner: Records R/W=286340/1164718 > reduce
16/02/04 08:22:00 INFO mapreduce.Job:  map 100% reduce 90%
16/02/04 08:22:03 INFO mapred.LocalJobRunner: Records R/W=286340/1164718 > reduce
16/02/04 08:22:03 INFO streaming.PipeMapRed: R/W/S=300000/1226025/0 in:2097=300000/143 [rec/s] out:8573
16/02/04 08:22:06 INFO mapred.LocalJobRunner: Records R/W=286340/1164718 > reduce
16/02/04 08:22:06 INFO mapreduce.Job:  map 100% reduce 91%
16/02/04 08:22:07 INFO streaming.PipeMapRed: Records R/W=305643/1248068
16/02/04 08:22:09 INFO mapred.LocalJobRunner: Records R/W=305643/1248068 > reduce
16/02/04 08:22:12 INFO mapred.LocalJobRunner: Records R/W=305643/1248068 > reduce
16/02/04 08:22:12 INFO mapreduce.Job:  map 100% reduce 92%
16/02/04 08:22:15 INFO mapred.LocalJobRunner: Records R/W=305643/1248068 > reduce
16/02/04 08:22:18 INFO mapred.LocalJobRunner: Records R/W=305643/1248068 > reduce
16/02/04 08:22:18 INFO streaming.PipeMapRed: Records R/W=323637/1321600
16/02/04 08:22:21 INFO mapred.LocalJobRunner: Records R/W=323637/1321600 > reduce
16/02/04 08:22:21 INFO mapreduce.Job:  map 100% reduce 93%
16/02/04 08:22:24 INFO mapred.LocalJobRunner: Records R/W=323637/1321600 > reduce
16/02/04 08:22:24 INFO mapreduce.Job:  map 100% reduce 94%
16/02/04 08:22:27 INFO mapred.LocalJobRunner: Records R/W=323637/1321600 > reduce
16/02/04 08:22:27 INFO mapreduce.Job:  map 100% reduce 95%
16/02/04 08:22:28 INFO streaming.PipeMapRed: Records R/W=353056/1478580
16/02/04 08:22:30 INFO mapred.LocalJobRunner: Records R/W=353056/1478580 > reduce
16/02/04 08:22:33 INFO mapred.LocalJobRunner: Records R/W=353056/1478580 > reduce
16/02/04 08:22:33 INFO mapreduce.Job:  map 100% reduce 96%
16/02/04 08:22:36 INFO mapred.LocalJobRunner: Records R/W=353056/1478580 > reduce
16/02/04 08:22:38 INFO streaming.PipeMapRed: Records R/W=375824/1581571
16/02/04 08:22:39 INFO mapred.LocalJobRunner: Records R/W=375824/1581571 > reduce
16/02/04 08:22:39 INFO mapreduce.Job:  map 100% reduce 97%
```

```
16/02/04 08:22:42 INFO mapred.LocalJobRunner: Records R/W=375824/1581571 > reduce
16/02/04 08:22:43 INFO mapreduce.Job:  map 100% reduce 98%
16/02/04 08:22:45 INFO mapred.LocalJobRunner: Records R/W=375824/1581571 > reduce
16/02/04 08:22:48 INFO mapred.LocalJobRunner: Records R/W=375824/1581571 > reduce
16/02/04 08:22:48 INFO streaming.PipeMapRed: Records R/W=396352/1682944
16/02/04 08:22:49 INFO mapreduce.Job:  map 100% reduce 99%
16/02/04 08:22:49 INFO streaming.PipeMapRed: R/W/S=400000/1704208/0 in:2116=400000/189 [rec/s] out:9016=
16/02/04 08:22:51 INFO mapred.LocalJobRunner: Records R/W=396352/1682944 > reduce
16/02/04 08:22:54 INFO mapred.LocalJobRunner: Records R/W=396352/1682944 > reduce
16/02/04 08:22:55 INFO mapreduce.Job:  map 100% reduce 100%
16/02/04 08:22:57 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:22:57 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:22:57 INFO mapred.Task: Task:attempt_local112720032_0001_r_000000_0 is done. And is in the pr
16/02/04 08:22:57 INFO mapred.LocalJobRunner: Records R/W=396352/1682944 > reduce
16/02/04 08:22:57 INFO mapred.Task: Task attempt_local112720032_0001_r_000000_0 is allowed to commit now
16/02/04 08:22:57 INFO output.FileOutputCommitter: Saved output of task 'attempt_local112720032_0001_r_00
16/02/04 08:22:57 INFO mapred.LocalJobRunner: Records R/W=396352/1682944 > reduce
16/02/04 08:22:57 INFO mapred.Task: Task 'attempt_local112720032_0001_r_000000_0' done.
16/02/04 08:22:57 INFO mapred.LocalJobRunner: Finishing task: attempt_local112720032_0001_r_000000_0
16/02/04 08:22:57 INFO mapred.LocalJobRunner: reduce task executor complete.
16/02/04 08:22:58 INFO mapreduce.Job: Job job_local112720032_0001 completed successfully
16/02/04 08:22:58 INFO mapreduce.Job: Counters: 36
        File System Counters
                FILE: Number of bytes read=327849780
                FILE: Number of bytes written=410317543
                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=35163286
                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=411925
                Map output bytes=80680946
                Map output materialized bytes=81909485
                Input split bytes=106
                Combine input records=0
                Combine output records=0
                Reduce input groups=12593
                Reduce shuffle bytes=81909485
                Reduce input records=411925
                Reduce output records=1754191
                Spilled Records=1235775
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=30
                Total committed heap usage (bytes)=602931200
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        group
                Num mapper calls=2
        File Input Format Counters
                Bytes Read=3458517
        File Output Format Counters
                Bytes Written=35163286
16/02/04 08:22:58 INFO streaming.StreamJob: Output directory: /user/hw3/output_3_5i
```

```
In [250]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -cat /user/hw3/output_3_5i/part-00000 > str
```

```
16/02/04 08:22:59 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
```

```
In [251]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -put stripes_unsorted /user/hw3
```

```
16/02/04 08:23:02 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
```

** Running a second map-reduce job to merge same pairs **
For example if there are two keys
A:{'B':2,'C':1}
B:{'A':1,'C':1}
The first map reduce job would output the following
A,B 2
A,B 1
etc...
The second map reduce job would merge these two pairs to give
A,B 3
And the third map reduce job is to sort the counts!

```
In [252]: %%writefile mapper.py
          #!/usr/bin/python
          import sys
          import re

          sys.stderr.write("reporter:counter:group,Num mapper calls,1\n")

          for line in sys.stdin:
              line = re.split(r'[\t]',line.strip())
              print "%s\t%s" %(line[0],line[1])
```

```
Overwriting mapper.py
```

```
In [253]: %%writefile reducer.py
          #!/usr/bin/python
          import sys
          import re

          sys.stderr.write("reporter:counter:group,Num mapper calls,1\n")

          prev_id = None
          count = 0
          for line in sys.stdin:
              line = re.split(r'[\t]',line.strip())
```

```
            if((prev_id !=None) and (line[0] !=prev_id)):
                print "%s\t%s" %(prev_id,count)
                count = 0
            count += int(line[1])
            prev_id = line[0]
        print "%s\t%s" %(prev_id,count)
```

```
Overwriting reducer.py
```

```
In [254]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hadoop jar /Users/Vamsi/Downloads/hadoop-2.7.1/bin/ha
          -D mapred.reduce.tasks=1 \
          -input /user/hw3/stripes_unsorted \
          -output /user/hw3/stripes_unsorted_int \
          -mapper mapper.py \
          -reducer reducer.py
```

```
16/02/04 08:23:06 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
16/02/04 08:23:07 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.ses
16/02/04 08:23:07 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/02/04 08:23:07 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessio
16/02/04 08:23:08 INFO mapred.FileInputFormat: Total input paths to process : 1
16/02/04 08:23:08 INFO mapreduce.JobSubmitter: number of splits:1
16/02/04 08:23:08 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapred
16/02/04 08:23:08 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1940809252_0001
16/02/04 08:23:08 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/02/04 08:23:08 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/02/04 08:23:08 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputComm
16/02/04 08:23:08 INFO mapreduce.Job: Running job: job_local1940809252_0001
16/02/04 08:23:08 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:23:08 INFO mapred.LocalJobRunner: Waiting for map tasks
16/02/04 08:23:08 INFO mapred.LocalJobRunner: Starting task: attempt_local1940809252_0001_m_000000_0
16/02/04 08:23:08 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:23:08 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only o
16/02/04 08:23:08 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:23:08 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/hw3/stripes_unsorted
16/02/04 08:23:08 INFO mapred.MapTask: numReduceTasks: 1
16/02/04 08:23:08 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/02/04 08:23:08 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/02/04 08:23:08 INFO mapred.MapTask: soft limit at 83886080
16/02/04 08:23:08 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/02/04 08:23:08 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/02/04 08:23:08 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$Map
16/02/04 08:23:08 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./mapper.p
16/02/04 08:23:08 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.ta
16/02/04 08:23:08 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce
16/02/04 08:23:08 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.m
16/02/04 08:23:08 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.j
16/02/04 08:23:08 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce
16/02/04 08:23:08 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use map
16/02/04 08:23:08 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce
16/02/04 08:23:08 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.jo
16/02/04 08:23:08 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.us
16/02/04 08:23:08 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapredu
16/02/04 08:23:08 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.t
16/02/04 08:23:08 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapr
```

```
16/02/04 08:23:09 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:23:09 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:23:09 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:23:09 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:23:09 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:23:09 INFO streaming.PipeMapRed: Records R/W=13090/1
16/02/04 08:23:09 INFO mapreduce.Job: Job job_local1940809252_0001 running in uber mode : false
16/02/04 08:23:09 INFO mapreduce.Job:  map 0% reduce 0%
16/02/04 08:23:09 INFO streaming.PipeMapRed: R/W/S=100000/90762/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:23:10 INFO streaming.PipeMapRed: R/W/S=200000/189675/0 in:200000=200000/1 [rec/s] out:189675
16/02/04 08:23:10 INFO streaming.PipeMapRed: R/W/S=300000/287689/0 in:300000=300000/1 [rec/s] out:287689
16/02/04 08:23:11 INFO streaming.PipeMapRed: R/W/S=400000/391449/0 in:200000=400000/2 [rec/s] out:195724
16/02/04 08:23:11 INFO streaming.PipeMapRed: R/W/S=500000/489556/0 in:250000=500000/2 [rec/s] out:24480
16/02/04 08:23:12 INFO streaming.PipeMapRed: R/W/S=600000/588409/0 in:200000=600000/3 [rec/s] out:19613
16/02/04 08:23:12 INFO streaming.PipeMapRed: R/W/S=700000/692204/0 in:233333=700000/3 [rec/s] out:230734
16/02/04 08:23:13 INFO streaming.PipeMapRed: R/W/S=800000/790332/0 in:200000=800000/4 [rec/s] out:19758
16/02/04 08:23:13 INFO streaming.PipeMapRed: R/W/S=900000/889223/0 in:180000=900000/5 [rec/s] out:177844
16/02/04 08:23:14 INFO streaming.PipeMapRed: R/W/S=1000000/986522/0 in:200000=1000000/5 [rec/s] out:1973
16/02/04 08:23:14 INFO mapred.LocalJobRunner: Records R/W=13090/1 > map
16/02/04 08:23:14 INFO streaming.PipeMapRed: R/W/S=1100000/1091577/0 in:183333=1100000/6 [rec/s] out:18
16/02/04 08:23:15 INFO streaming.PipeMapRed: R/W/S=1200000/1190049/0 in:200000=1200000/6 [rec/s] out:19
16/02/04 08:23:15 INFO mapreduce.Job:  map 40% reduce 0%
16/02/04 08:23:16 INFO streaming.PipeMapRed: R/W/S=1300000/1288112/0 in:185714=1300000/7 [rec/s] out:18
16/02/04 08:23:16 INFO streaming.PipeMapRed: R/W/S=1400000/1392571/0 in:200000=1400000/7 [rec/s] out:19
16/02/04 08:23:17 INFO streaming.PipeMapRed: R/W/S=1500000/1490858/0 in:187500=1500000/8 [rec/s] out:18
16/02/04 08:23:17 INFO streaming.PipeMapRed: R/W/S=1600000/1588905/0 in:200000=1600000/8 [rec/s] out:19
16/02/04 08:23:17 INFO mapred.LocalJobRunner: Records R/W=13090/1 > map
16/02/04 08:23:18 INFO streaming.PipeMapRed: R/W/S=1700000/1686989/0 in:188888=1700000/9 [rec/s] out:18
16/02/04 08:23:18 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:23:18 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:23:18 INFO mapred.LocalJobRunner: Records R/W=13090/1 > map
16/02/04 08:23:18 INFO mapred.MapTask: Starting flush of map output
16/02/04 08:23:18 INFO mapred.MapTask: Spilling map output
16/02/04 08:23:18 INFO mapred.MapTask: bufstart = 0; bufend = 35163286; bufvoid = 104857600
16/02/04 08:23:18 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 19197636(76790544); lengt
16/02/04 08:23:18 INFO mapreduce.Job:  map 62% reduce 0%
16/02/04 08:23:20 INFO mapred.LocalJobRunner: Records R/W=13090/1 > sort
16/02/04 08:23:21 INFO mapreduce.Job:  map 67% reduce 0%
16/02/04 08:23:22 INFO mapred.MapTask: Finished spill 0
16/02/04 08:23:23 INFO mapred.Task: Task:attempt_local1940809252_0001_m_000000_0 is done. And is in the p
16/02/04 08:23:23 INFO mapred.LocalJobRunner: Records R/W=13090/1
16/02/04 08:23:23 INFO mapred.Task: Task 'attempt_local1940809252_0001_m_000000_0' done.
16/02/04 08:23:23 INFO mapred.LocalJobRunner: Finishing task: attempt_local1940809252_0001_m_000000_0
16/02/04 08:23:23 INFO mapred.LocalJobRunner: map task executor complete.
16/02/04 08:23:23 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/02/04 08:23:23 INFO mapred.LocalJobRunner: Starting task: attempt_local1940809252_0001_r_000000_0
16/02/04 08:23:23 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:23:23 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
16/02/04 08:23:23 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:23:23 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task
16/02/04 08:23:23 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleL
16/02/04 08:23:23 INFO reduce.EventFetcher: attempt_local1940809252_0001_r_000000_0 Thread started: Event
16/02/04 08:23:23 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1
16/02/04 08:23:23 INFO reduce.InMemoryMapOutput: Read 38671670 bytes from map-output for attempt_local1
```

```
16/02/04 08:23:23 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 38671670, inMe
16/02/04 08:23:23 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/02/04 08:23:23 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/04 08:23:23 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on
16/02/04 08:23:23 INFO mapred.Merger: Merging 1 sorted segments
16/02/04 08:23:23 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 
16/02/04 08:23:23 INFO mapreduce.Job:  map 100% reduce 0%
16/02/04 08:23:24 INFO reduce.MergeManagerImpl: Merged 1 segments, 38671670 bytes to disk to satisfy re
16/02/04 08:23:24 INFO reduce.MergeManagerImpl: Merging 1 files, 38671674 bytes from disk
16/02/04 08:23:24 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/02/04 08:23:24 INFO mapred.Merger: Merging 1 sorted segments
16/02/04 08:23:24 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 
16/02/04 08:23:24 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/04 08:23:24 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./reducer
16/02/04 08:23:24 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapredu
16/02/04 08:23:24 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce
16/02/04 08:23:24 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:23:24 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:23:24 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:23:24 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:23:24 INFO streaming.PipeMapRed: Records R/W=6546/1
16/02/04 08:23:24 INFO streaming.PipeMapRed: R/W/S=10000/2448/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:23:25 INFO streaming.PipeMapRed: R/W/S=100000/46455/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:23:26 INFO streaming.PipeMapRed: R/W/S=200000/95380/0 in:200000=200000/1 [rec/s] out:95380=
16/02/04 08:23:26 INFO streaming.PipeMapRed: R/W/S=300000/145073/0 in:150000=300000/2 [rec/s] out:72536
16/02/04 08:23:27 INFO streaming.PipeMapRed: R/W/S=400000/196547/0 in:200000=400000/2 [rec/s] out:98276
16/02/04 08:23:27 INFO streaming.PipeMapRed: R/W/S=500000/246046/0 in:166666=500000/3 [rec/s] out:82015
16/02/04 08:23:28 INFO streaming.PipeMapRed: R/W/S=600000/294871/0 in:200000=600000/3 [rec/s] out:98290
16/02/04 08:23:28 INFO streaming.PipeMapRed: R/W/S=700000/347012/0 in:175000=700000/4 [rec/s] out:86753
16/02/04 08:23:29 INFO mapred.LocalJobRunner: Records R/W=6546/1 > reduce
16/02/04 08:23:29 INFO streaming.PipeMapRed: R/W/S=800000/396688/0 in:200000=800000/4 [rec/s] out:99172
16/02/04 08:23:29 INFO mapreduce.Job:  map 100% reduce 81%
16/02/04 08:23:30 INFO streaming.PipeMapRed: R/W/S=900000/445584/0 in:180000=900000/5 [rec/s] out:89116
16/02/04 08:23:30 INFO streaming.PipeMapRed: R/W/S=1000000/494472/0 in:166666=1000000/6 [rec/s] out:824
16/02/04 08:23:31 INFO streaming.PipeMapRed: R/W/S=1100000/546619/0 in:183333=1100000/6 [rec/s] out:911
16/02/04 08:23:31 INFO streaming.PipeMapRed: R/W/S=1200000/595534/0 in:171428=1200000/7 [rec/s] out:850
16/02/04 08:23:32 INFO mapred.LocalJobRunner: Records R/W=6546/1 > reduce
16/02/04 08:23:32 INFO streaming.PipeMapRed: R/W/S=1300000/645224/0 in:162500=1300000/8 [rec/s] out:806
16/02/04 08:23:32 INFO mapreduce.Job:  map 100% reduce 90%
16/02/04 08:23:33 INFO streaming.PipeMapRed: R/W/S=1400000/697433/0 in:175000=1400000/8 [rec/s] out:871
16/02/04 08:23:33 INFO streaming.PipeMapRed: R/W/S=1500000/746305/0 in:166666=1500000/9 [rec/s] out:829
16/02/04 08:23:34 INFO streaming.PipeMapRed: Records R/W=1588834/789491
16/02/04 08:23:34 INFO streaming.PipeMapRed: R/W/S=1600000/795203/0 in:160000=1600000/10 [rec/s] out:79
16/02/04 08:23:35 INFO mapred.LocalJobRunner: Records R/W=1588834/789491 > reduce
16/02/04 08:23:35 INFO streaming.PipeMapRed: R/W/S=1700000/847339/0 in:170000=1700000/10 [rec/s] out:84
16/02/04 08:23:35 INFO mapreduce.Job:  map 100% reduce 98%
16/02/04 08:23:35 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:23:35 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:23:35 INFO mapred.Task: Task:attempt_local1940809252_0001_r_000000_0 is done. And is in the p
16/02/04 08:23:35 INFO mapred.LocalJobRunner: Records R/W=1588834/789491 > reduce
16/02/04 08:23:35 INFO mapred.Task: Task attempt_local1940809252_0001_r_000000_0 is allowed to commit now
16/02/04 08:23:35 INFO output.FileOutputCommitter: Saved output of task ’attempt_local1940809252_0001_r_
16/02/04 08:23:35 INFO mapred.LocalJobRunner: Records R/W=1588834/789491 > reduce
16/02/04 08:23:35 INFO mapred.Task: Task ’attempt_local1940809252_0001_r_000000_0’ done.
```

```
16/02/04 08:23:35 INFO mapred.LocalJobRunner: Finishing task: attempt_local1940809252_0001_r_000000_0
16/02/04 08:23:35 INFO mapred.LocalJobRunner: reduce task executor complete.
16/02/04 08:23:36 INFO mapreduce.Job:  map 100% reduce 100%
16/02/04 08:23:36 INFO mapreduce.Job: Job job_local1940809252_0001 completed successfully
16/02/04 08:23:36 INFO mapreduce.Job: Counters: 36
        File System Counters
                FILE: Number of bytes read=77555164
                FILE: Number of bytes written=116788132
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=70326572
                HDFS: Number of bytes written=17637494
                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=1754191
                Map output records=1754191
                Map output bytes=35163286
                Map output materialized bytes=38671674
                Input split bytes=99
                Combine input records=0
                Combine output records=0
                Reduce input groups=877096
                Reduce shuffle bytes=38671674
                Reduce input records=1754191
                Reduce output records=877096
                Spilled Records=3508382
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=14
                Total committed heap usage (bytes)=526909440
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        group
                Num mapper calls=2
        File Input Format Counters
                Bytes Read=35163286
        File Output Format Counters
                Bytes Written=17637494
16/02/04 08:23:36 INFO streaming.StreamJob: Output directory: /user/hw3/stripes_unsorted_int
```

```
In [255]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -cat /user/hw3/stripes_unsorted_int/part-00
```

```
16/02/04 08:23:38 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
```

```
In [256]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -rm -r -f /user/hw3/stripes_unsorted_int
```

```
16/02/04 08:23:40 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
16/02/04 08:23:41 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minut
Deleted /user/hw3/stripes_unsorted_int
```

```
In [257]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -put stripes_unsorted_int /user/hw3
```

```
16/02/04 08:23:43 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
```

Third Map reduce job to perform the sort

```
In [258]: %%writefile mapper.py
          #!/usr/bin/python
          import sys
          import re

          sys.stderr.write("reporter:counter:group,Num mapper calls,1\n")

          for line in sys.stdin:
              line = re.split(r'[\t]',line.strip())
              print "%s,%s" %(line[0],line[1])
```

```
Overwriting mapper.py
```

```
In [259]: %%writefile reducer.py
          #!/usr/bin/python
          import sys
          import re

          sys.stderr.write("reporter:counter:group,Num mapper calls,1\n")

          for line in sys.stdin:
              line = re.split(r'[,]',line.strip())
              if(line[0]=="*"):
                  total = int(line[2])
              else:
                  rel_freq = float(line[2])/float(total)
                  print "%s,%s,%s,%s,%s" %(line[0],line[1],line[2],rel_freq,total)
```

```
Overwriting reducer.py
```

```
In [260]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hadoop jar /Users/Vamsi/Downloads/hadoop-2.7.1/bin/ha
          -D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator \
          -D stream.map.output.field.separator=, \
          -D stream.num.map.output.key.fields=3 \
          -D map.output.key.field.separator=, \
          -D mapred.text.key.comparator.options=-k3,3nr \
          -D mapred.reduce.tasks=1 \
          -input /user/hw3/stripes_unsorted_int \
          -output /user/hw3/output_3_5_o \
          -mapper mapper.py \
          -reducer reducer.py
```

```
16/02/04 08:23:46 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
16/02/04 08:23:46 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.se
16/02/04 08:23:46 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/02/04 08:23:46 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessi
```

```
16/02/04 08:23:47 INFO mapred.FileInputFormat: Total input paths to process : 1
16/02/04 08:23:47 INFO mapreduce.JobSubmitter: number of splits:1
16/02/04 08:23:47 INFO Configuration.deprecation: map.output.key.field.separator is deprecated. Instead
16/02/04 08:23:47 INFO Configuration.deprecation: mapred.text.key.comparator.options is deprecated. Inst
16/02/04 08:23:47 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapred
16/02/04 08:23:47 INFO Configuration.deprecation: mapred.output.key.comparator.class is deprecated. Inst
16/02/04 08:23:47 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1642583775_0001
16/02/04 08:23:47 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/02/04 08:23:47 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/02/04 08:23:47 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputComm
16/02/04 08:23:47 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:23:47 INFO mapreduce.Job: Running job: job_local1642583775_0001
16/02/04 08:23:47 INFO mapred.LocalJobRunner: Waiting for map tasks
16/02/04 08:23:47 INFO mapred.LocalJobRunner: Starting task: attempt_local1642583775_0001_m_000000_0
16/02/04 08:23:47 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:23:47 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
16/02/04 08:23:47 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:23:47 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/hw3/stripes_unsorte
16/02/04 08:23:47 INFO mapred.MapTask: numReduceTasks: 1
16/02/04 08:23:48 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/02/04 08:23:48 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/02/04 08:23:48 INFO mapred.MapTask: soft limit at 83886080
16/02/04 08:23:48 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/02/04 08:23:48 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/02/04 08:23:48 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$Ma
16/02/04 08:23:48 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./mapper.
16/02/04 08:23:48 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use map
16/02/04 08:23:48 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce
16/02/04 08:23:48 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapredu
16/02/04 08:23:48 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.
16/02/04 08:23:48 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.ta
16/02/04 08:23:48 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce
16/02/04 08:23:48 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.
16/02/04 08:23:48 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.
16/02/04 08:23:48 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce
16/02/04 08:23:48 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.jo
16/02/04 08:23:48 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.us
16/02/04 08:23:48 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use map
16/02/04 08:23:48 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:23:48 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:23:48 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:23:48 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:23:48 INFO streaming.PipeMapRed: Records R/W=6532/1
16/02/04 08:23:48 INFO streaming.PipeMapRed: R/W/S=10000/5717/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:23:48 INFO streaming.PipeMapRed: R/W/S=100000/92927/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:23:48 INFO mapreduce.Job: Job job_local1642583775_0001 running in uber mode : false
16/02/04 08:23:48 INFO mapreduce.Job:  map 0% reduce 0%
16/02/04 08:23:49 INFO streaming.PipeMapRed: R/W/S=200000/191519/0 in:200000=200000/1 [rec/s] out:19151
16/02/04 08:23:50 INFO streaming.PipeMapRed: R/W/S=300000/294871/0 in:300000=300000/1 [rec/s] out:29487
16/02/04 08:23:50 INFO streaming.PipeMapRed: R/W/S=400000/392616/0 in:200000=400000/2 [rec/s] out:19630
16/02/04 08:23:50 INFO streaming.PipeMapRed: R/W/S=500000/490389/0 in:250000=500000/2 [rec/s] out:24519
16/02/04 08:23:51 INFO streaming.PipeMapRed: R/W/S=600000/594719/0 in:200000=600000/3 [rec/s] out:19823
16/02/04 08:23:51 INFO streaming.PipeMapRed: R/W/S=700000/692918/0 in:233333=700000/3 [rec/s] out:23097
16/02/04 08:23:52 INFO streaming.PipeMapRed: R/W/S=800000/790307/0 in:200000=800000/4 [rec/s] out:19757
```

```
16/02/04 08:23:52 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:23:52 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:23:52 INFO mapred.LocalJobRunner:
16/02/04 08:23:52 INFO mapred.MapTask: Starting flush of map output
16/02/04 08:23:52 INFO mapred.MapTask: Spilling map output
16/02/04 08:23:52 INFO mapred.MapTask: bufstart = 0; bufend = 18514590; bufvoid = 104857600
16/02/04 08:23:52 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 22706016(90824064); lengt
16/02/04 08:23:53 INFO mapred.MapTask: Finished spill 0
16/02/04 08:23:53 INFO mapred.LocalJobRunner: Records R/W=6532/1 > sort
16/02/04 08:23:53 INFO mapred.Task: Task:attempt_local1642583775_0001_m_000000_0 is done. And is in the p
16/02/04 08:23:53 INFO mapred.LocalJobRunner: Records R/W=6532/1
16/02/04 08:23:53 INFO mapred.Task: Task 'attempt_local1642583775_0001_m_000000_0' done.
16/02/04 08:23:53 INFO mapred.LocalJobRunner: Finishing task: attempt_local1642583775_0001_m_000000_0
16/02/04 08:23:53 INFO mapred.LocalJobRunner: map task executor complete.
16/02/04 08:23:53 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/02/04 08:23:53 INFO mapred.LocalJobRunner: Starting task: attempt_local1642583775_0001_r_000000_0
16/02/04 08:23:53 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/04 08:23:53 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
16/02/04 08:23:53 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/02/04 08:23:53 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task
16/02/04 08:23:54 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleL
16/02/04 08:23:54 INFO reduce.EventFetcher: attempt_local1642583775_0001_r_000000_0 Thread started: Event
16/02/04 08:23:54 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1
16/02/04 08:23:54 INFO reduce.InMemoryMapOutput: Read 20268784 bytes from map-output for attempt_local16
16/02/04 08:23:54 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 20268784, inMe
16/02/04 08:23:54 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/02/04 08:23:54 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/04 08:23:54 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on
16/02/04 08:23:54 INFO mapred.Merger: Merging 1 sorted segments
16/02/04 08:23:54 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size:
16/02/04 08:23:54 INFO reduce.MergeManagerImpl: Merged 1 segments, 20268784 bytes to disk to satisfy re
16/02/04 08:23:54 INFO reduce.MergeManagerImpl: Merging 1 files, 20268788 bytes from disk
16/02/04 08:23:54 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/02/04 08:23:54 INFO mapred.Merger: Merging 1 sorted segments
16/02/04 08:23:54 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size:
16/02/04 08:23:54 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/02/04 08:23:54 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw3/./reducer
16/02/04 08:23:54 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapredu
16/02/04 08:23:54 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce
16/02/04 08:23:54 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:23:54 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:23:54 INFO mapreduce.Job:  map 100% reduce 0%
16/02/04 08:23:54 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:23:54 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:23:55 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/02/04 08:23:55 INFO streaming.PipeMapRed: Records R/W=10234/1
16/02/04 08:23:56 INFO streaming.PipeMapRed: R/W/S=100000/89304/0 in:100000=100000/1 [rec/s] out:89304
16/02/04 08:23:57 INFO streaming.PipeMapRed: R/W/S=200000/188808/0 in:100000=200000/2 [rec/s] out:94406
16/02/04 08:23:58 INFO streaming.PipeMapRed: R/W/S=300000/288412/0 in:100000=300000/3 [rec/s] out:96137
16/02/04 08:23:58 INFO streaming.PipeMapRed: R/W/S=400000/388315/0 in:100000=400000/4 [rec/s] out:97078
16/02/04 08:23:59 INFO streaming.PipeMapRed: R/W/S=500000/488480/0 in:100000=500000/5 [rec/s] out:97696
16/02/04 08:23:59 INFO mapred.LocalJobRunner: Records R/W=10234/1 > reduce
16/02/04 08:24:00 INFO streaming.PipeMapRed: R/W/S=600000/588274/0 in:120000=600000/5 [rec/s] out:11765
16/02/04 08:24:00 INFO mapreduce.Job:  map 100% reduce 86%
```

```
16/02/04 08:24:01 INFO streaming.PipeMapRed: R/W/S=700000/688067/0 in:100000=700000/7 [rec/s] out:98295
16/02/04 08:24:02 INFO mapred.LocalJobRunner: Records R/W=10234/1 > reduce
16/02/04 08:24:03 INFO streaming.PipeMapRed: R/W/S=800000/787861/0 in:100000=800000/8 [rec/s] out:98482
16/02/04 08:24:03 INFO mapreduce.Job:  map 100% reduce 96%
16/02/04 08:24:04 INFO streaming.PipeMapRed: MRErrorThread done
16/02/04 08:24:04 INFO streaming.PipeMapRed: mapRedFinished
16/02/04 08:24:04 INFO mapred.Task: Task:attempt_local1642583775_0001_r_000000_0 is done. And is in the p
16/02/04 08:24:04 INFO mapred.LocalJobRunner: Records R/W=10234/1 > reduce
16/02/04 08:24:04 INFO mapred.Task: Task attempt_local1642583775_0001_r_000000_0 is allowed to commit now
16/02/04 08:24:04 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1642583775_0001_r_0
16/02/04 08:24:04 INFO mapred.LocalJobRunner: Records R/W=10234/1 > reduce
16/02/04 08:24:04 INFO mapred.Task: Task 'attempt_local1642583775_0001_r_000000_0' done.
16/02/04 08:24:04 INFO mapred.LocalJobRunner: Finishing task: attempt_local1642583775_0001_r_000000_0
16/02/04 08:24:04 INFO mapred.LocalJobRunner: reduce task executor complete.
16/02/04 08:24:04 INFO mapreduce.Job:  map 100% reduce 100%
16/02/04 08:24:04 INFO mapreduce.Job: Job job_local1642583775_0001 completed successfully
16/02/04 08:24:04 INFO mapreduce.Job: Counters: 36
        File System Counters
                FILE: Number of bytes read=40749400
                FILE: Number of bytes written=61583834
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=35274988
                HDFS: Number of bytes written=39406387
                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=877096
                Map output records=877096
                Map output bytes=18514590
                Map output materialized bytes=20268788
                Input split bytes=103
                Combine input records=0
                Combine output records=0
                Reduce input groups=877096
                Reduce shuffle bytes=20268788
                Reduce input records=877096
                Reduce output records=877095
                Spilled Records=1754192
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=20
                Total committed heap usage (bytes)=574619648
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        group
```

```
                Num mapper calls=2
        File Input Format Counters
                Bytes Read=17637494
        File Output Format Counters
                Bytes Written=39406387
16/02/04 08:24:04 INFO streaming.StreamJob: Output directory: /user/hw3/output_3_5_o
```

** Computational Setup : **

```
MacBook Air
Processor : 1.8 Ghz Intel Core i5 , 2 Cores
Memory : 4 GB 1600 Mhz DDR3
Disk : SSD (Flash Storage) 128 GB
```

The output is shown below
Format = Pairs, Support Count, Support, Total number of baskets

```
In [261]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -cat /user/hw3/output_3_5_o/part-00000 | hea
```

```
16/02/04 08:24:06 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
DAI62779,ELE17451,3184,0.102376129385,31101
FRO40251,SNA80324,2824,0.0908009388766,31101
DAI75645,FRO40251,2508,0.0806404938748,31101
FRO40251,GRO85051,2426,0.0780039227035,31101
DAI62779,GRO73461,2278,0.0732452332722,31101
DAI75645,SNA80324,2260,0.0726664737468,31101
DAI62779,FRO40251,2140,0.0688080769107,31101
DAI62779,SNA80324,1846,0.0593550046622,31101
DAI62779,DAI85309,1836,0.0590334715926,31101
ELE32164,GRO59710,1822,0.058583325295,31101
DAI62779,DAI75645,1764,0.0567184334909,31101
FRO40251,GRO73461,1764,0.0567184334909,31101
DAI62779,ELE92920,1754,0.0563969004212,31101
FRO40251,FRO92469,1670,0.0536960226359,31101
DAI62779,ELE32164,1664,0.0535031027941,31101
DAI75645,GRO73461,1424,0.0457863091219,31101
DAI43223,ELE32164,1422,0.045722002508,31101
DAI62779,GRO30386,1418,0.0455933892801,31101
ELE17451,FRO40251,1394,0.0448217099129,31101
DAI85309,ELE99737,1318,0.0423780585833,31101
DAI62779,ELE26917,1300,0.0417992990579,31101
GRO21487,GRO73461,1262,0.0405774733931,31101
DAI62779,SNA45677,1208,0.0388411948169,31101
ELE17451,SNA80324,1194,0.0383910485193,31101
DAI62779,GRO71621,1190,0.0382624352915,31101
DAI62779,SNA55762,1186,0.0381338220636,31101
DAI62779,DAI83733,1172,0.0376836757661,31101
ELE17451,GRO73461,1160,0.0372978360824,31101
GRO73461,SNA80324,1124,0.0361403170316,31101
DAI62779,GRO59710,1122,0.0360760104177,31101
DAI62779,FRO80039,1100,0.0353686376644,31101
DAI75645,ELE17451,1094,0.0351757178226,31101
DAI62779,SNA93860,1074,0.0345326516832,31101
DAI55148,DAI62779,1052,0.0338252789299,31101
DAI43223,GRO59710,1024,0.0329249863348,31101
```

```
ELE17451,ELE32164,1022,0.0328606797209,31101
DAI62779,SNA18336,1012,0.0325391466512,31101
ELE32164,GRO73461,972,0.0312530143725,31101
DAI85309,ELE17451,964,0.0309957879168,31101
DAI62779,FRO78087,964,0.0309957879168,31101
DAI62779,GRO94758,958,0.030802868075,31101
GRO85051,SNA80324,942,0.0302884151635,31101
DAI62779,GRO21487,942,0.0302884151635,31101
ELE17451,GRO30386,936,0.0300954953217,31101
FRO85978,SNA95666,926,0.029773962252,31101
DAI62779,FRO19221,924,0.0297096556381,31101
DAI62779,GRO46854,922,0.0296453490241,31101
DAI43223,DAI62779,918,0.0295167357963,31101
ELE92920,SNA18336,910,0.0292595093405,31101
DAI88079,FRO40251,892,0.0286807498151,31101
cat: Unable to write to output stream.
```

```
In [262]: !/Users/Vamsi/Downloads/hadoop-2.7.1/sbin/stop-dfs.sh
```

```
16/02/04 08:24:09 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
Stopping namenodes on [localhost]
localhost: stopping namenode
localhost: stopping datanode
Stopping secondary namenodes [0.0.0.0]
0.0.0.0: stopping secondarynamenode
16/02/04 08:24:30 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
```