part1-wordcount

**Part 1. [5 point] Set up a single node cluster and optionally an eclipse development environment to create and test your programs.**

(a) Get VMWare or VirtualBox (install)

(b) Get Cloudera (install)

(c) Get WordCount (test run)

1. I wrote a WordCount java class, than change it to mywordcount.jar.
2. I wrote some shell script tools :

ls-input.sh, ls-output.sh, rm-output.sh

also a wordcount shart shell : wordcount.sh

This is wordcount.sh

#!/bin/bash

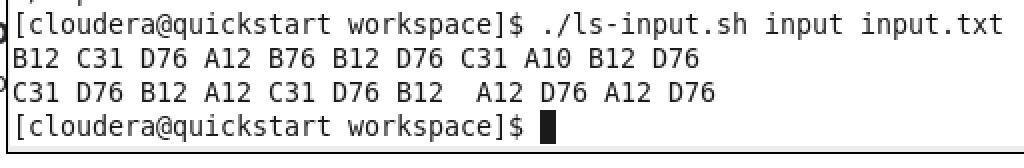
hadoop fs -rm -r /user/cloudera/wordcount/output

hadoop jar mywordcount.jar wordcount.WordCount /user/cloudera/wordcount/input /user/cloudera/wordcount/output

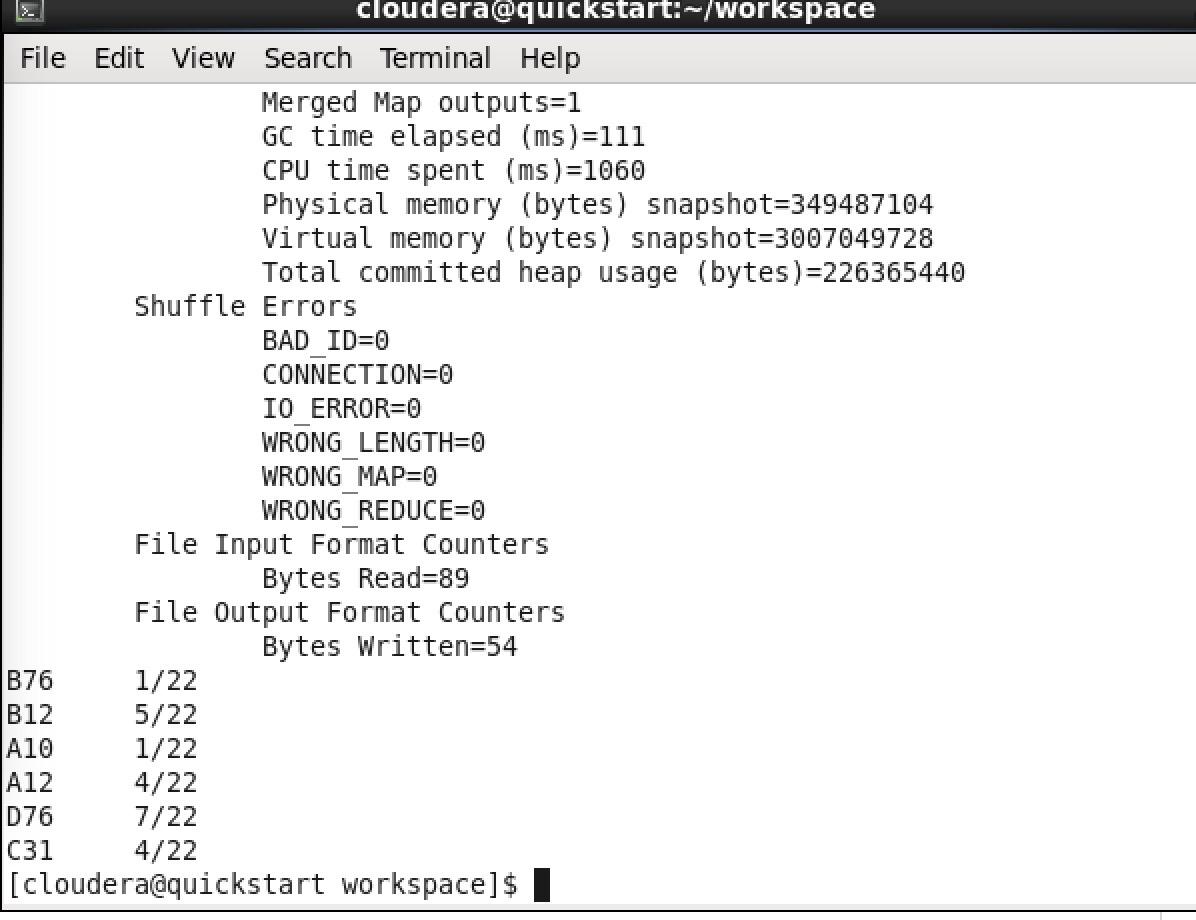
./ls-output.sh part-r-00000

3. I put input.txt to hadoop file system

hadoop fs -put input/input.txt /user/cloudera/input



4.run wordcount.sh to start wordcount program.

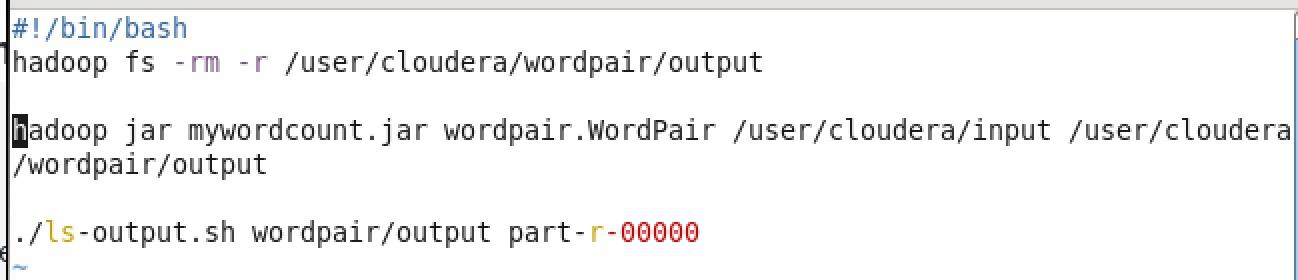


I will add attachment with code and tools.

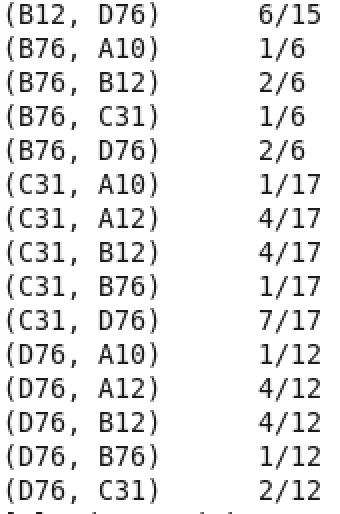
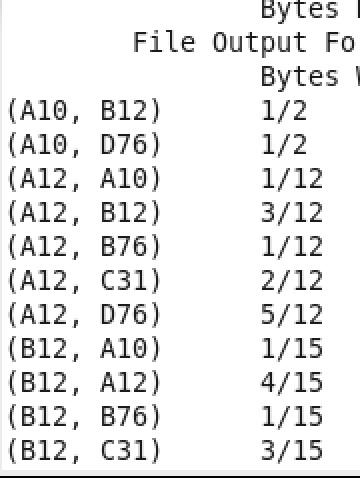
**Part 2. Implement Pairs algorithm to compute relative frequencies.**

1. [2 points] Create Java classes (.java files)
2. [1 points] Show input, output and batch file to execute your program at command line in Hadoop.

This is wordpair.sh. Run wordpair.sh to run word pair.



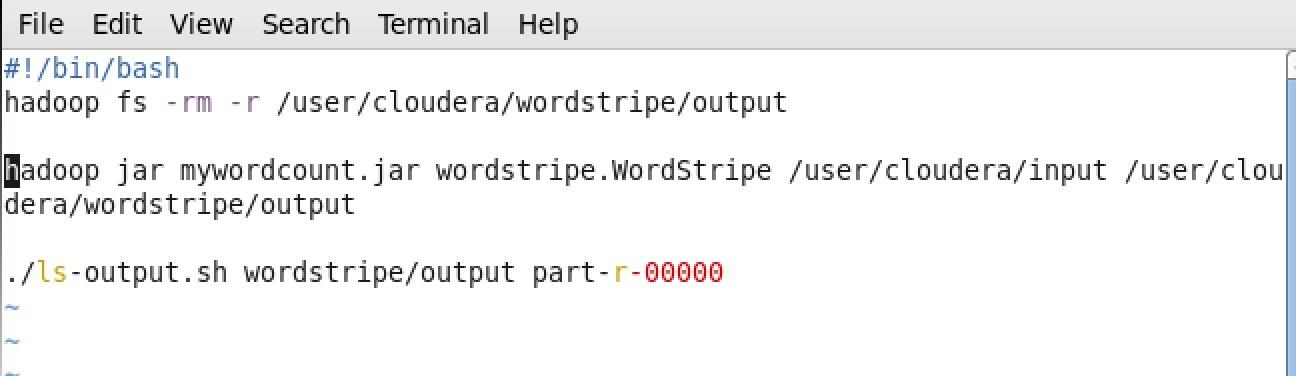
This is word pair output.



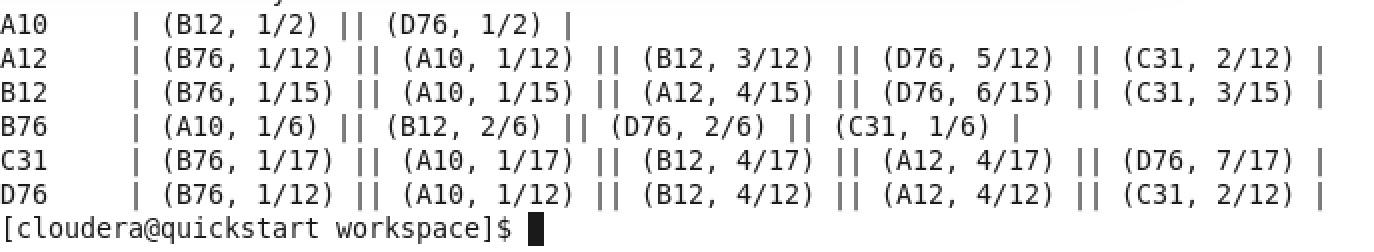
**Part 3. Implement Stripes algorithm to compute relative frequencies.**

1. [2 points] Create Java classes (.java files)
2. [1 points] Show input, output and batch file to execute your program at command line in Hadoop.

Use the command : ./wordstripe.sh to run this shell script.



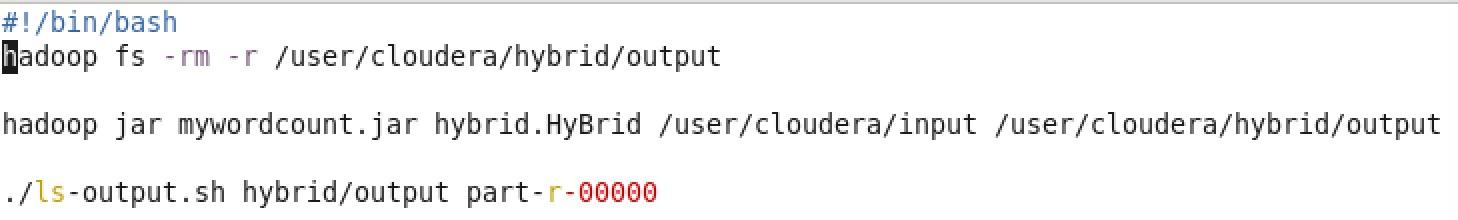
This is output of stripe.



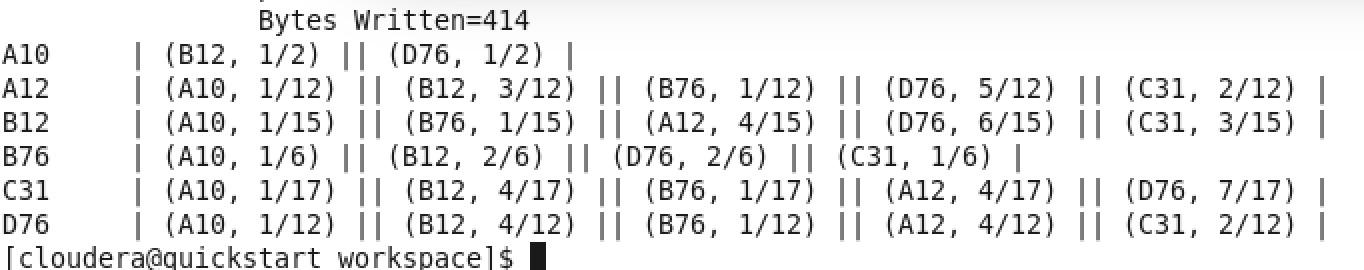
**Part 4. Implement Pairs in Mapper and Stripes in Reducer to compute relative frequencies.**

1. [2 points] Create Java classes (.java files)
2. [1 points] Show input, output and batch file to execute your program at command line in Hadoop.

This is hybrid.sh. Use the command : ./hybrid.sh to run hybrid.



This is hybrid output.



Tools for this project.

ls-output.sh & ls-input.sh

