Assignment 1

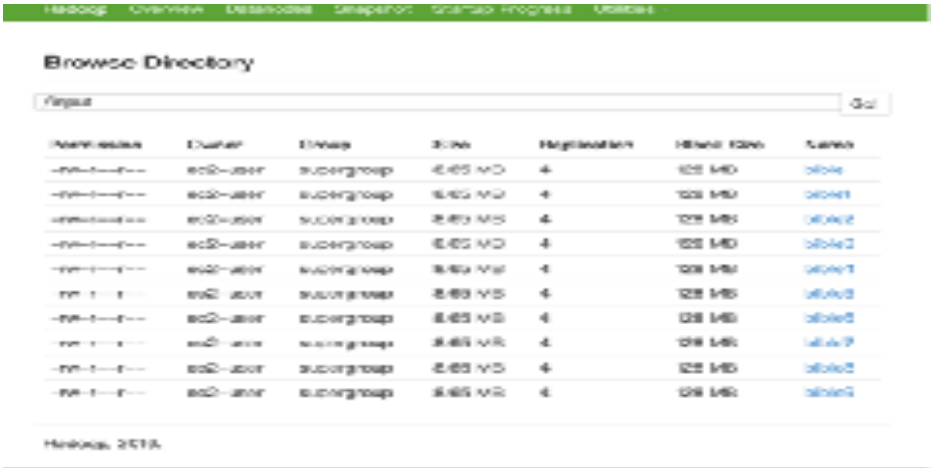
Jingyi Ding 69364001

Codes and outputs in https://github.com/dingjy94/EEL6761_Cloud_Computing

1. Some aws steps:

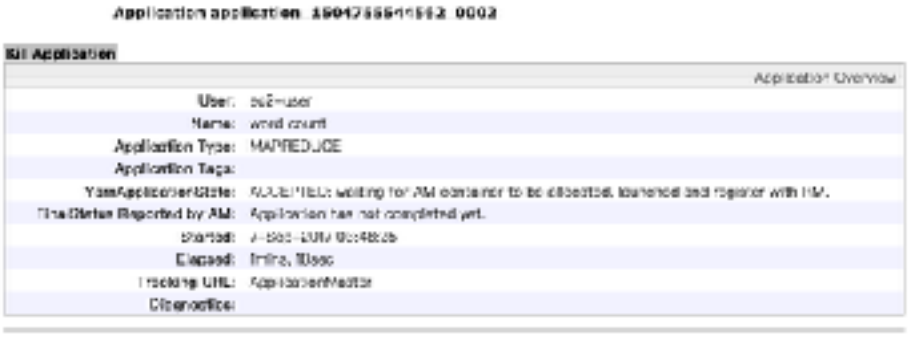
3 aws sc2 instances, one for namenode, two for datanodes;

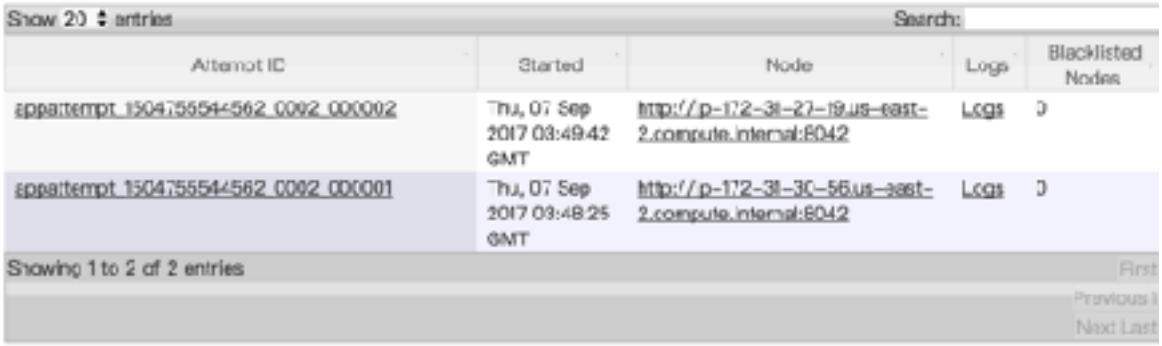
2 datanode in operation when hdfs started;



10 bibles (input files).

2. part1

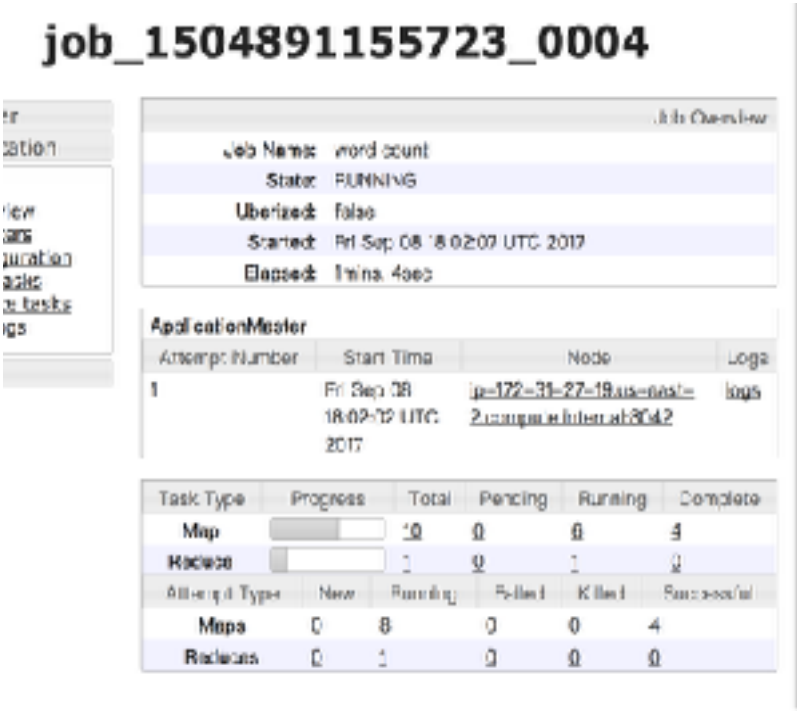


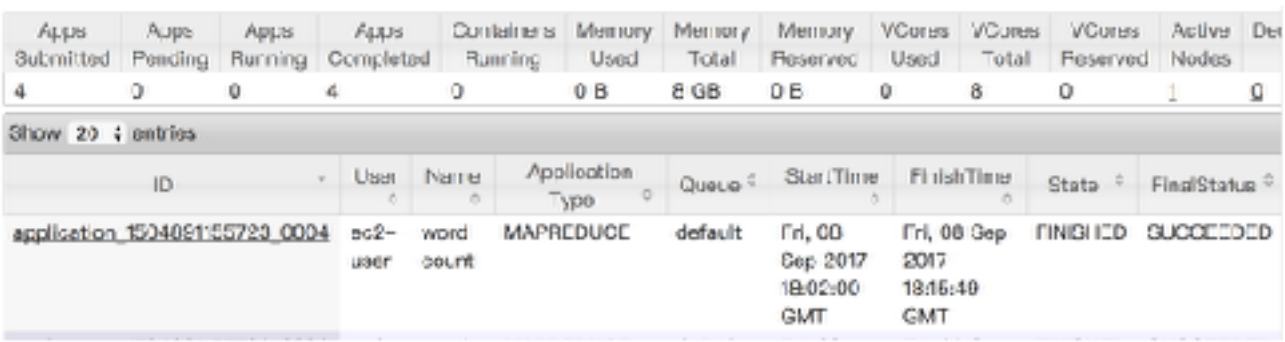


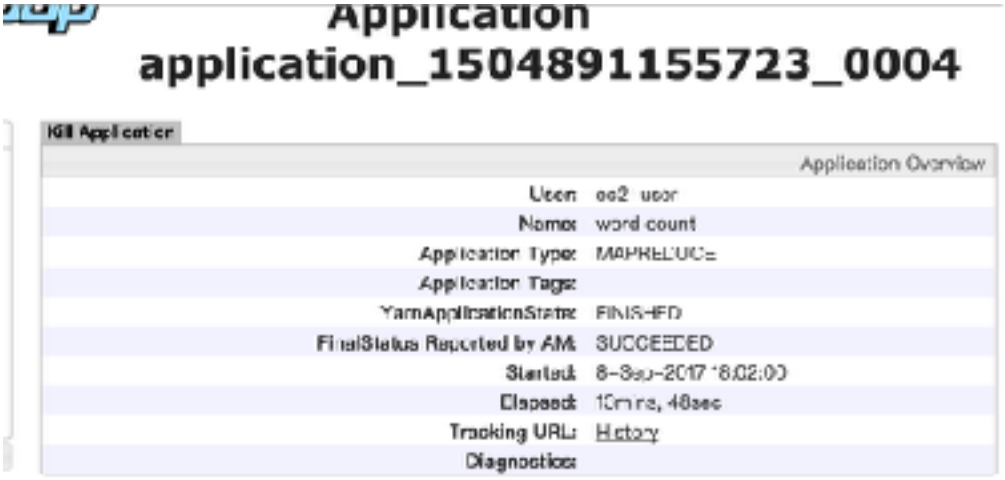
The code is just similar to Hadoop's sample WordCount.

3. part2

My solution is sample, just use two StringTokenizer to read the input file, one StringTokenizer is one word faster than another. Thus, we can easily get double words key by combine these two StringTokenizer.







4. part3

The program get three arguments, input, output and word-pattern file's paths instead of only two arguments. Then, add word-pattern into local cache. In mapper class, when setup the map, read the local file and turns it to HashSet. Then, map method is still similier to WordCount, but only write the word contained in HashSet