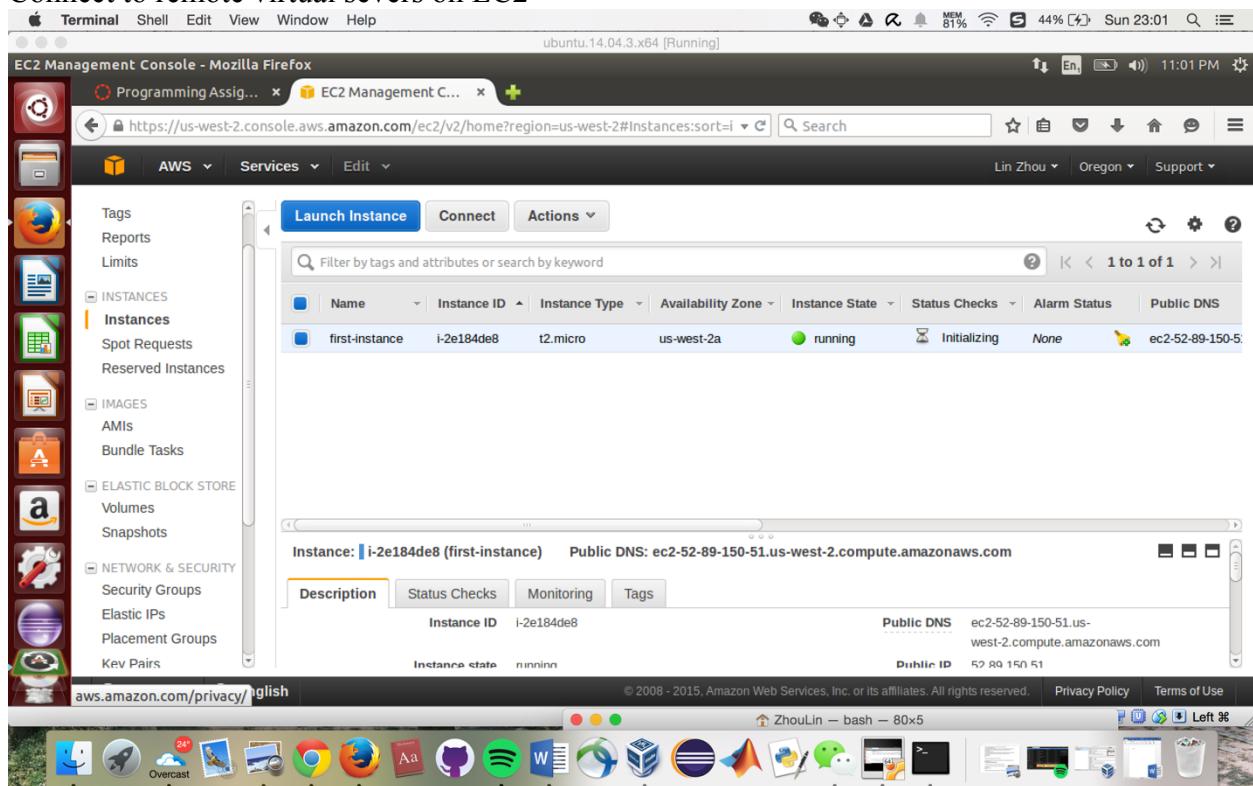
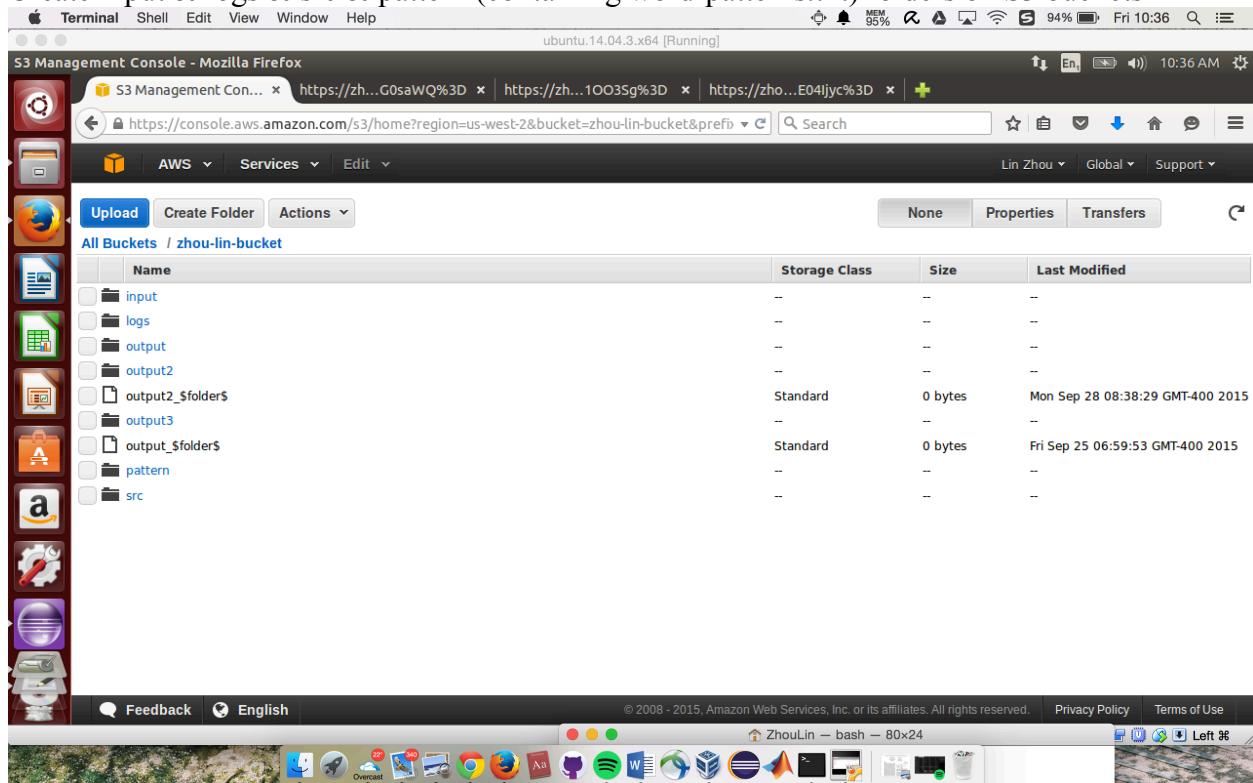


Connect to remote virtual servers on EC2



Create input & logs & src & pattern (containing word-patterns.txt) folders on S3 buckets



Input folder (10 bibles)

The screenshot shows the AWS S3 Management Console in Mozilla Firefox. The URL is <https://console.aws.amazon.com/s3/home?region=us-west-2#bucket=zhou-lin-bucket&prefix=input>. The page displays a list of files in the 'input' folder of the 'zhou-lin-bucket'. There are 10 files, all named 'bible+shakes-[1-10].nopunc' and are 8.6 MB each, last modified on Fri Sep 25 03:59:18 GMT-400 2015.

Name	Storage Class	Size	Last Modified
bible+shakes-1.nopunc	Standard	8.6 MB	Fri Sep 25 03:59:18 GMT-400 2015
bible+shakes-10.nopunc	Standard	8.6 MB	Fri Sep 25 03:59:37 GMT-400 2015
bible+shakes-2.nopunc	Standard	8.6 MB	Fri Sep 25 03:59:22 GMT-400 2015
bible+shakes-3.nopunc	Standard	8.6 MB	Fri Sep 25 03:59:23 GMT-400 2015
bible+shakes-4.nopunc	Standard	8.6 MB	Fri Sep 25 03:59:27 GMT-400 2015
bible+shakes-5.nopunc	Standard	8.6 MB	Fri Sep 25 03:59:30 GMT-400 2015
bible+shakes-6.nopunc	Standard	8.6 MB	Fri Sep 25 03:59:31 GMT-400 2015
bible+shakes-7.nopunc	Standard	8.6 MB	Fri Sep 25 03:59:33 GMT-400 2015
bible+shakes-8.nopunc	Standard	8.6 MB	Fri Sep 25 03:59:34 GMT-400 2015
bible+shakes-9.nopunc	Standard	8.6 MB	Fri Sep 25 03:59:35 GMT-400 2015

Src folder (containing jar files)

The screenshot shows the AWS S3 Management Console in Mozilla Firefox. The URL is <https://console.aws.amazon.com/s3/home?region=us-west-2#bucket=zhou-lin-bucket&prefix=src>. The page displays a list of files in the 'src' folder of the 'zhou-lin-bucket'. There are 7 files: WordCount.java, WordCount2.jar, WordCount2.java, WordCount4.java, WordCount4.java, wordcount.jar, and wordcount2.jar. The sizes range from 2.1 KB to 9.7 KB, and the last modification dates are from Fri Sep 25 03:57:50 GMT-400 2015 to Mon Sep 28 08:35:07 GMT-400 2015.

Name	Storage Class	Size	Last Modified
WordCount.java	Standard	2.1 KB	Fri Sep 25 03:57:50 GMT-400 2015
WordCount2.jar	Standard	9.7 KB	Mon Sep 28 07:26:49 GMT-400 2015
WordCount2.java	Standard	2.3 KB	Fri Sep 25 14:20:31 GMT-400 2015
WordCount4.java	Standard	10.2 KB	Thu Oct 01 23:43:40 GMT-400 2015
WordCount4.java	Standard	3.2 KB	Thu Oct 01 23:43:39 GMT-400 2015
wordcount.jar	Standard	1.4 KB	Fri Sep 25 06:36:44 GMT-400 2015
wordcount2.jar	Standard	9.7 KB	Mon Sep 28 08:35:07 GMT-400 2015

Create a new cluster

Create Cluster - Advanced Options [go to quick options](#) [Configure sample application](#)

Cluster name	<input type="text" value="zhou-lin-cluster"/>	Termination protection	<input checked="" type="radio"/> Yes <input type="radio"/> No	Prevents accidental termination of the cluster: to shut down the cluster, you must turn off termination protection. Learn more
Logging	<input checked="" type="checkbox"/> Enabled	Copy the cluster's log files automatically to S3. Learn more		
Log folder S3 location		<input type="text" value="s3://zhou-lin-bucket/logs/"/> s3://<bucket-name>/<folder>/		
Debugging	<input checked="" type="checkbox"/> Enabled	Index logs to enable console debugging functionality (requires logging). Learn more		

Hardware Configuration

Specify the networking and hardware configuration for your cluster. If you need more than 20 EC2 instances, [complete this form](#). Request Spot Instances (unused EC2 capacity) to save money.

Network	<input type="text" value="vpc-5c2d6d39 (172.31.0.0/16) (default)"/>	Use a Virtual Private Cloud (VPC) to process sensitive data or connect to a private network. Create a VPC
EC2 Subnet	<input type="text" value="No preference (random subnet)"/>	Create a Subnet

Type	Name	EC2 instance type	Count	Request spot	Bid price
Master	Master instance group	m3.xlarge	1	<input type="checkbox"/>	?
Core	Core instance group	m3.xlarge	1	<input type="checkbox"/>	?
Task	Task instance group	m3.xlarge	0	<input type="checkbox"/>	X ?

[Add task instance group](#)

Security and Access

EC2 key pair	<input type="text" value="lecture1"/>	Use an existing EC2 key pair to SSH into the master node of the Amazon EMR cluster. Learn more
IAM user access	<input checked="" type="radio"/> All other IAM users <input type="radio"/> No other IAM users	Control the visibility of this cluster to other IAM users. Learn more

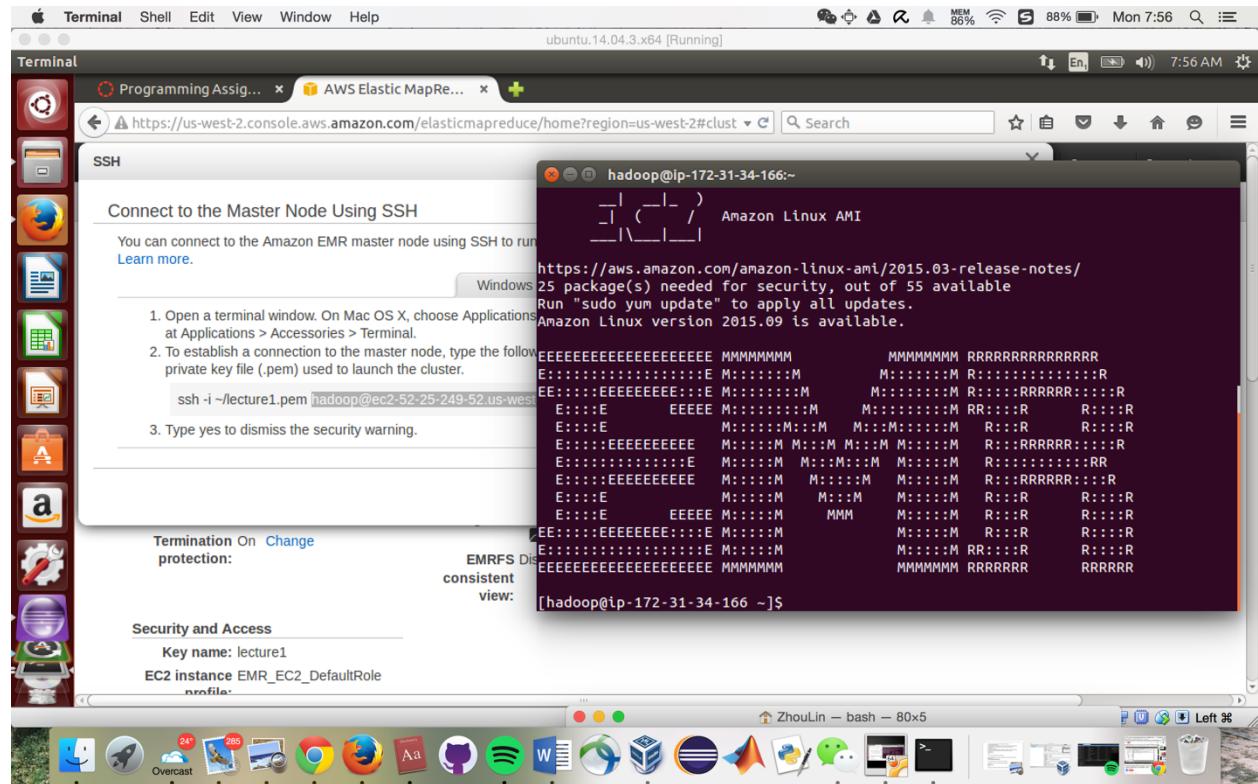
The screenshot shows the AWS Elastic MapReduce Management Console interface. At the top, there's a toolbar with icons for Terminal, Shell, Edit, View, Window, Help, and system status indicators like battery level (77%), signal strength, and time (Mon 9:10). Below the toolbar is the browser address bar showing the URL <https://us-west-2.console.aws.amazon.com/elasticmapreduce/home?region=us-west-2>.

The main content area is titled "Elastic MapReduce" and "Cluster List". It features a "Create cluster" button and a table listing four clusters:

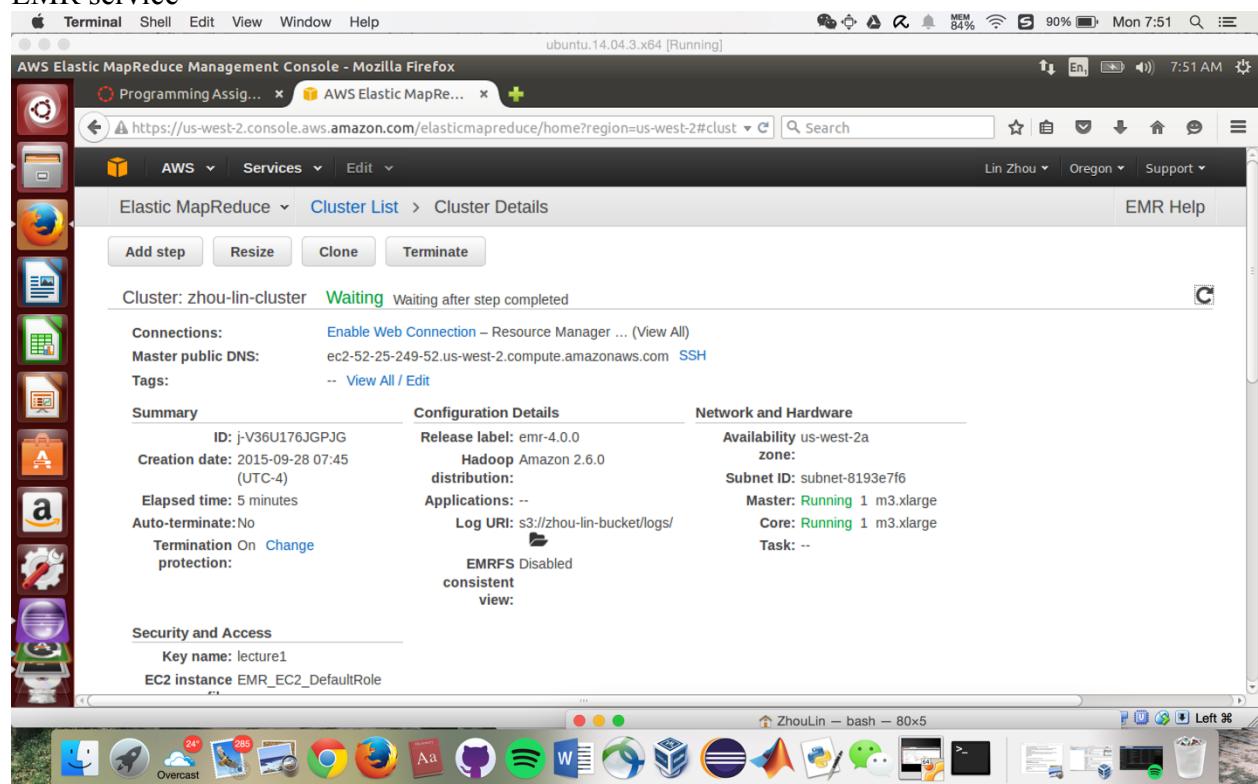
	Name	ID	Status	Creation time (UTC-4)	Elapsed time	Normalized instance hours
<input type="checkbox"/>	zhou-lin-cluster	j-V36U176JGPJG	Waiting	2015-09-28 07:45 (UTC-4)	1 hour, 24 minutes	32
<input type="checkbox"/>	zhou-lin-cluster1	j-3A91HHOEEFE8H	Terminated User request	2015-09-25 04:07 (UTC-4)	13 hours	224
<input type="checkbox"/>	Word count	j-3QWUCVWL9VOU	Terminated All steps completed	2015-09-21 15:54 (UTC-4)	10 minutes	24
<input type="checkbox"/>	Word count	j-2HKN2OLU9UAS	Terminated	2015-09-21 15:06 (UTC-4)	16 minutes	24

At the bottom of the browser window, there are links for Feedback, English, Privacy Policy, and Terms of Use. The Mac OS X dock at the very bottom contains various application icons including Finder, Mail, Safari, and others.

Connect to Master-node



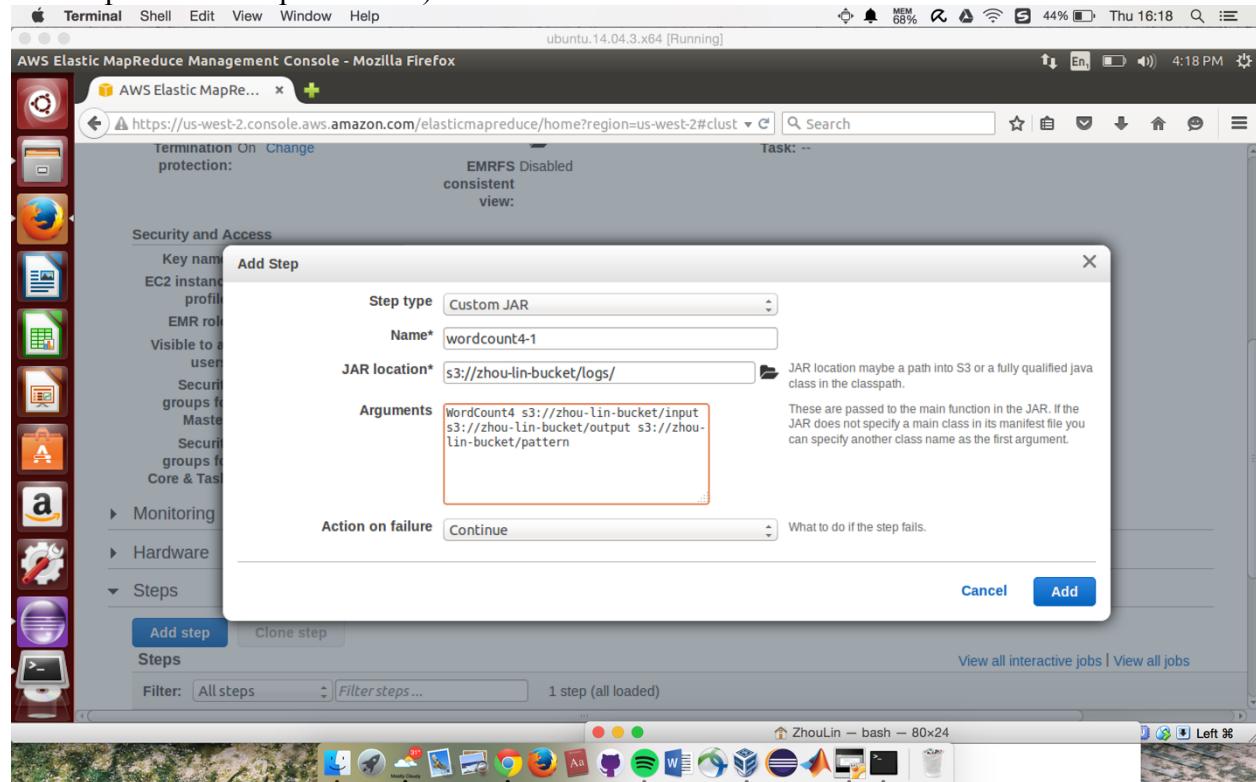
EMR service



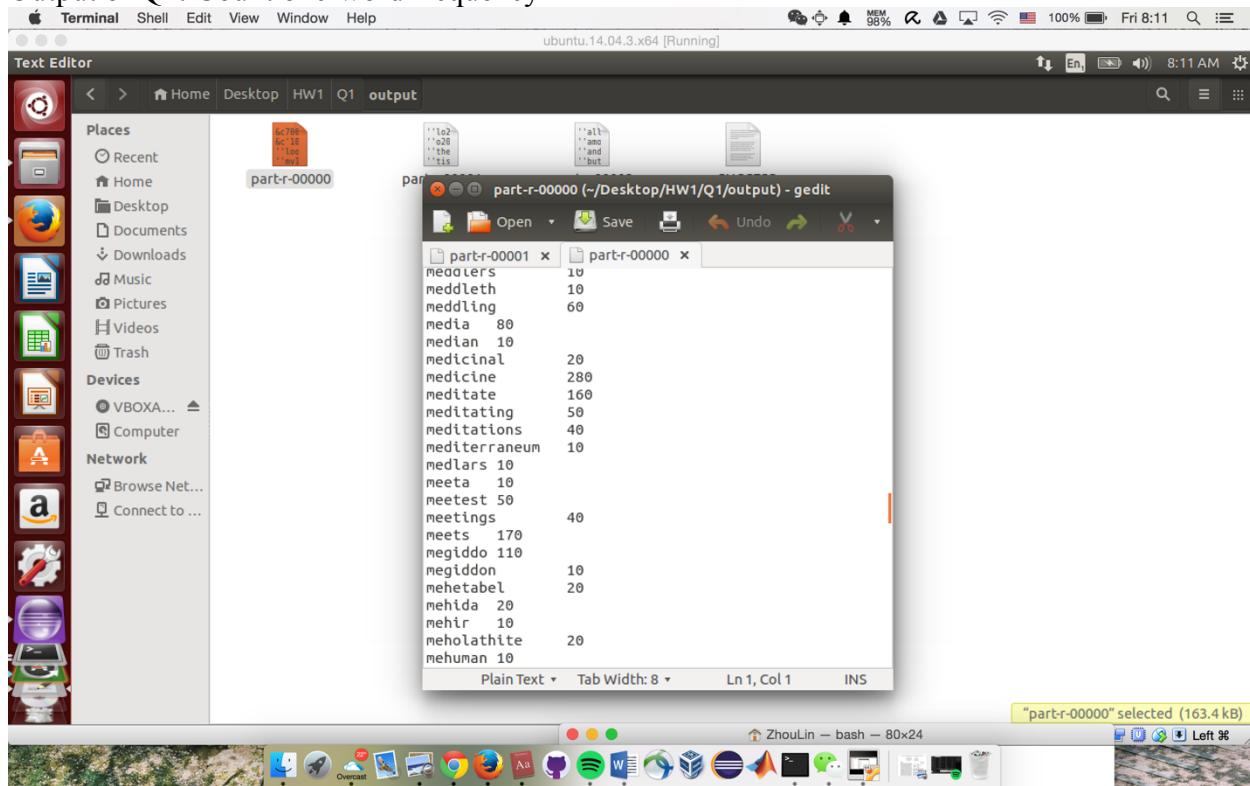
Submit a job to Hadoop cluster (Add step; Configuration of Q3)

(Arguments are change shown as follows:

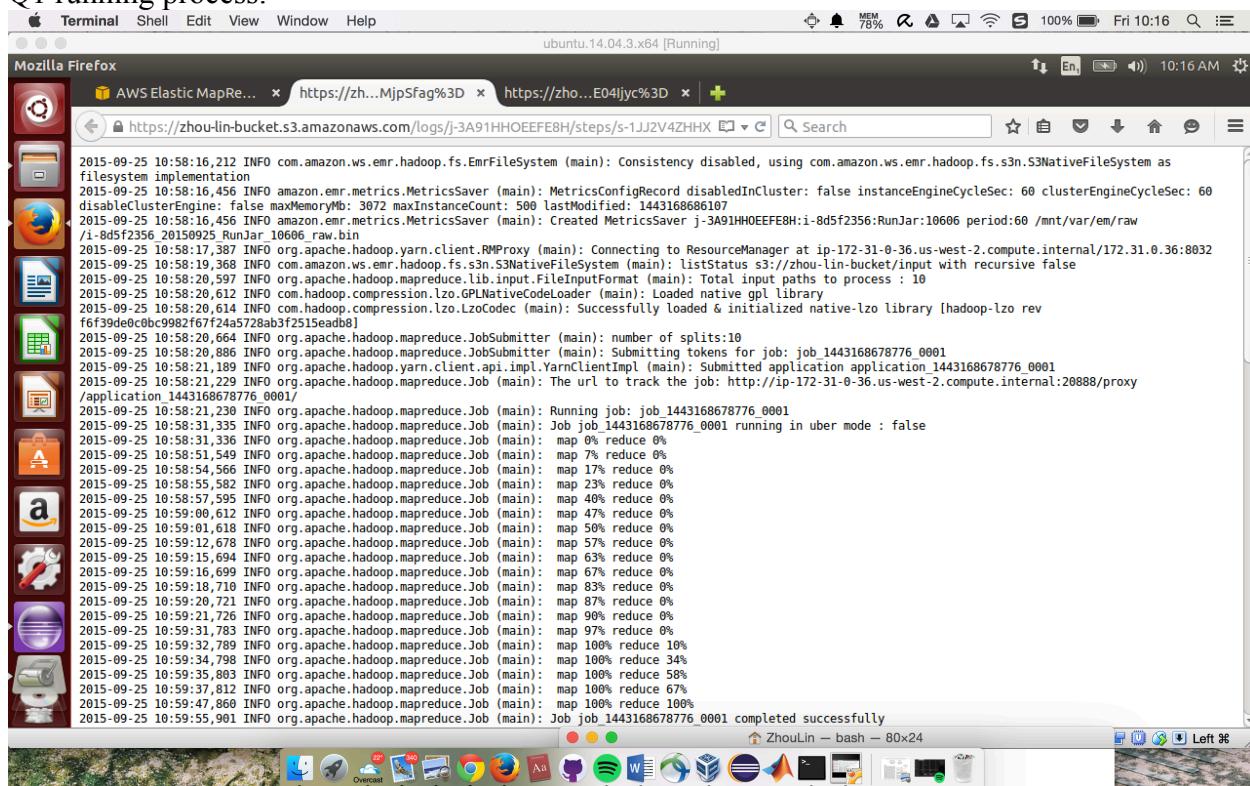
WordCount4 s3://zhou-lin-bucket/input s3://zhou-lin-bucket/output3 s3://zhou-lin-bucket/pattern/word-patterns.txt)



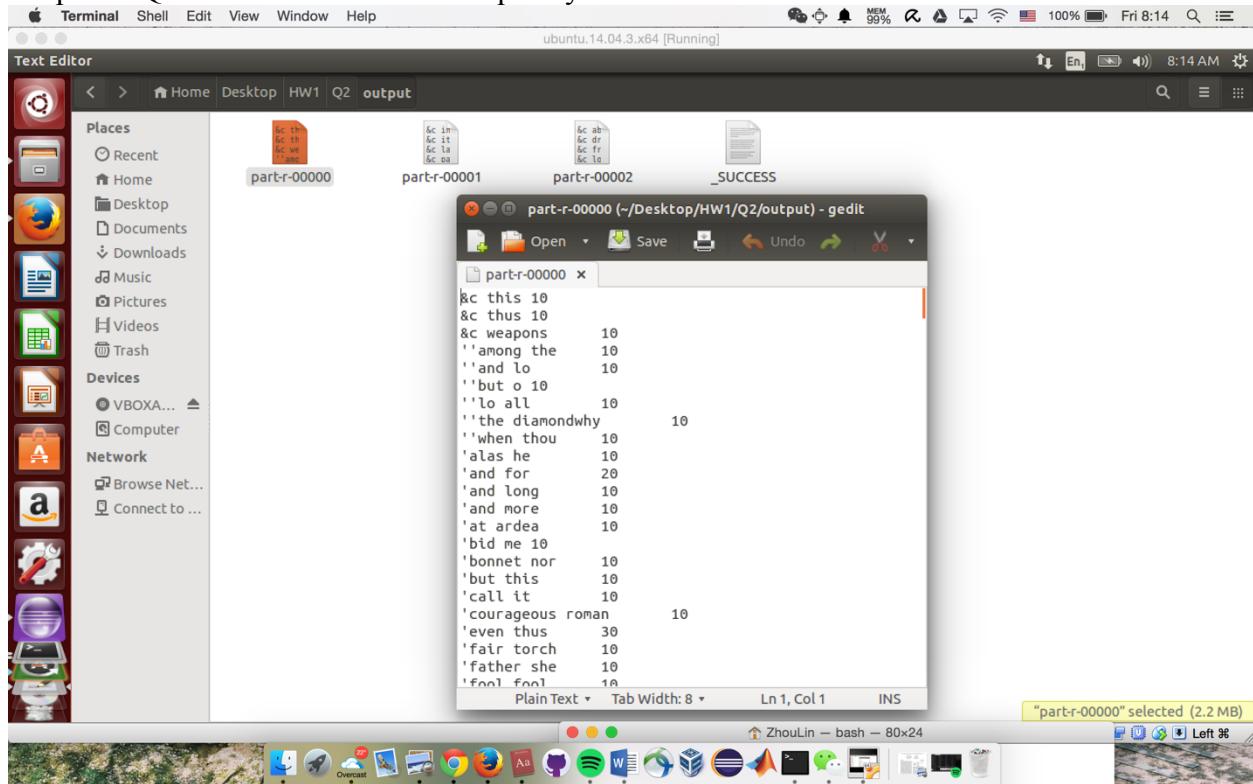
Output of Q1: Count one-word frequency



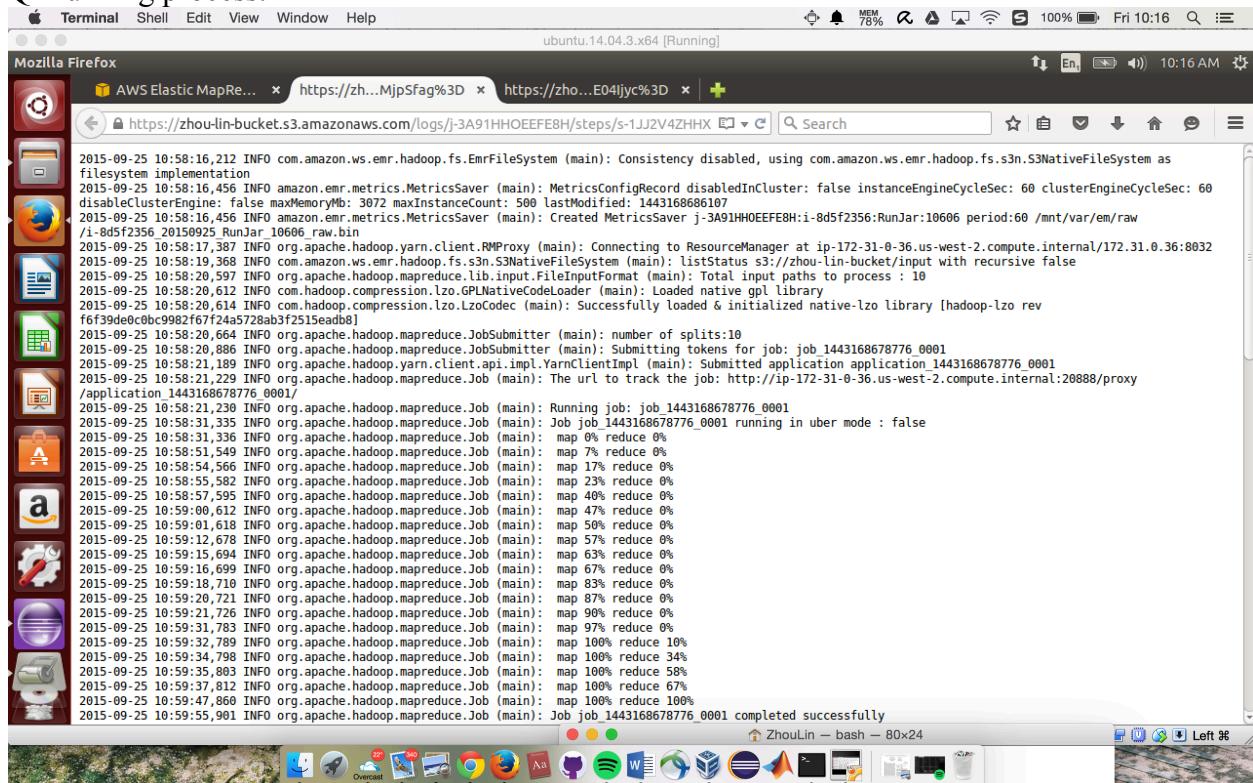
Q1 running process:



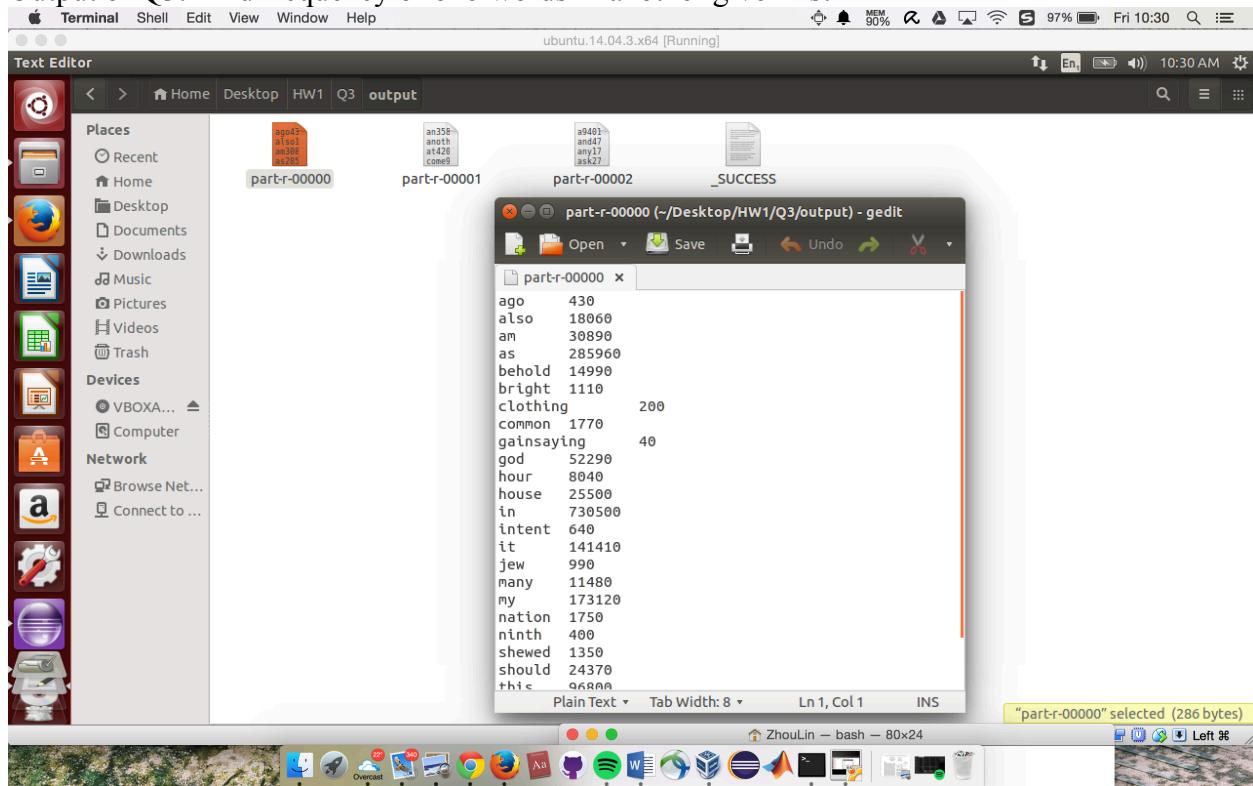
Output of Q2: Count double-word frequency



Q2 running process:



Output of Q3: Find frequency of one-words in another given list



Q3 running process:

