

ASSIGNMENT #2 REPORT

Team Members and Contribution:

- Sankalp Heranjal (A20332999) – Hadoop
- Akshay Ganji (A20325252) – Swift
- Sai Abhilash Putrevu (A20324224) – Java

HADOOP DOCUMENTATION AND EVALUATION

Author: Sankalp Heranjal

Operating Systems Used: Windows 7 – 64Bit and Ubuntu 14.04 LTS – 64Bit (VM)

Software Used: Hadoop v-1.2.1, Java v-1.7.0_65, Eclipse 3.8, PuTTY, FileZilla

ANT Version: 1.2.1

Steps Taken (With Screenshots):

Launching Amazon EC2 Instances:

We started with creating an Amazon AWS Account. After creation of the account, we started with the Cluster Setup. We started with the 16node setup first. We launched an Ubuntu 14.04 c3.large instance with 32GB EBS Storage and created a key file keypairaws.pem which we uploaded in the instance using FileZilla.

All the following execution is done in PuTTY.

Java and Hadoop Installation:

We installed Java v-1.7.0_65 and Hadoop v-1.2.1. The following screenshots show the Java and Hadoop Versions.

CS553 Programming Assignment #2

```
ubuntu@ip-172-31-47-197: ~$
Setting up libgnomevfs2-0:amd64 (1:2.24.4-1ubuntu6) ...
Setting up libgnome2-common (2.32.1-4ubuntu1) ...
Setting up libgnome2-bin (2.32.1-4ubuntu1) ...
Setting up libgnome2-0:amd64 (2.32.1-4ubuntu1) ...
Setting up libatk-wrapper-java-gnome:amd64 (0.30.4-4) ...
Setting up openjdk-7-jre:amd64 (7u71-2.5.3-0ubuntu0.14.04.1) ...
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/jre/bin/policytool to provide /usr/bin/policytool (policytool) in auto mode
Setting up openjdk-7-jdk:amd64 (7u71-2.5.3-0ubuntu0.14.04.1) ...
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/appletviewer to provide /usr/bin/appletviewer (appletviewer) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/extractcheck to provide /usr/bin/extractcheck (extractcheck) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/idlj to provide /usr/bin/idlj (idlj) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/jar to provide /usr/bin/jar (jar) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/jarsigner to provide /usr/bin/jarsigner (jarsigner) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/javac to provide /usr/bin/javac (javac) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/javadoc to provide /usr/bin/javadoc (javadoc) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/javah to provide /usr/bin/javah (javah) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/javap to provide /usr/bin/javap (javap) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/jcmd to provide /usr/bin/jcmd (jcmd) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/jconsole to provide /usr/bin/jconsole (jconsole) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/jdb to provide /usr/bin/jdb (jdb) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/jhat to provide /usr/bin/jhat (jhat) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/jinfo to provide /usr/bin/jinfo (jinfo) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/jmap to provide /usr/bin/jmap (jmap) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/jps to provide /usr/bin/jps (jps) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/jrunscript to provide /usr/bin/jrunscript (jrunscript) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/jstack to provide /usr/bin/jstack (jstack) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/jstat to provide /usr/bin/jstat (jstat) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/jstatd to provide /usr/bin/jstatd (jstatd) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/native2ascii to provide /usr/bin/native2ascii (native2ascii) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/rmic to provide /usr/bin/rmic (rmic) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/schemagen to provide /usr/bin/schemagen (schemagen) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/serialver to provide /usr/bin/serialver (serialver) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/wsgen to provide /usr/bin/wsgen (wsgen) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/wsimport to provide /usr/bin/wsimport (wsimport) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/xjc to provide /usr/bin/xjc (xjc) in auto mode
Processing triggers for libc-bin (2.19-0ubuntu6.3) ...
ubuntu@ip-172-31-47-197:~$ jps
10330 Jps
ubuntu@ip-172-31-47-197:~$ java -version
java version "1.7.0_65"
OpenJDK Runtime Environment (IcedTea 2.5.3) (7u71-2.5.3-0ubuntu0.14.04.1)
OpenJDK 64-Bit Server VM (build 24.65-b04, mixed mode)
ubuntu@ip-172-31-47-197:~$
```

```
ubuntu@ip-172-31-47-197: ~$
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/jarsigner to provide /usr/bin/jarsigner (jarsigner) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/javac to provide /usr/bin/javac (javac) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/javadoc to provide /usr/bin/javadoc (javadoc) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/javah to provide /usr/bin/javah (javah) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/javap to provide /usr/bin/javap (javap) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/jcmd to provide /usr/bin/jcmd (jcmd) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/jconsole to provide /usr/bin/jconsole (jconsole) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/jdb to provide /usr/bin/jdb (jdb) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/jhat to provide /usr/bin/jhat (jhat) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/jinfo to provide /usr/bin/jinfo (jinfo) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/jmap to provide /usr/bin/jmap (jmap) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/jps to provide /usr/bin/jps (jps) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/jrunscript to provide /usr/bin/jrunscript (jrunscript) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/jstack to provide /usr/bin/jstack (jstack) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/jstat to provide /usr/bin/jstat (jstat) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/jstatd to provide /usr/bin/jstatd (jstatd) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/native2ascii to provide /usr/bin/native2ascii (native2ascii) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/rmic to provide /usr/bin/rmic (rmic) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/schemagen to provide /usr/bin/schemagen (schemagen) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/serialver to provide /usr/bin/serialver (serialver) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/wsgen to provide /usr/bin/wsgen (wsgen) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/wsimport to provide /usr/bin/wsimport (wsimport) in auto mode
update-alternatives: using /usr/lib/jvm/java-7-openjdk-amd64/bin/xjc to provide /usr/bin/xjc (xjc) in auto mode
Processing triggers for libc-bin (2.19-0ubuntu6.3) ...
ubuntu@ip-172-31-47-197:~$ jps
10330 Jps
ubuntu@ip-172-31-47-197:~$ java -version
java version "1.7.0_65"
OpenJDK Runtime Environment (IcedTea 2.5.3) (7u71-2.5.3-0ubuntu0.14.04.1)
OpenJDK 64-Bit Server VM (build 24.65-b04, mixed mode)
ubuntu@ip-172-31-47-197:~$ wget http://apache.mirror.gtccomm.net/hadoop/common/hadoop-1.2.1/hadoop-1.2.1.tar.gz
--2014-10-23 21:09:06-- http://apache.mirror.gtccomm.net/hadoop/common/hadoop-1.2.1/hadoop-1.2.1.tar.gz
Resolving apache.mirror.gtccomm.net (apache.mirror.gtccomm.net)... 67.215.8.196
Connecting to apache.mirror.gtccomm.net (apache.mirror.gtccomm.net):67.215.8.196:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 63851630 (61M) [application/x-gzip]
Saving to: 'hadoop-1.2.1.tar.gz'

100%[=====] 63,851,630 4.88MB/s in 13s

2014-10-23 21:09:19 (4.75 MB/s) - 'hadoop-1.2.1.tar.gz' saved [63851630/63851630]

ubuntu@ip-172-31-47-197:~$
```

```

ubuntu@ip-172-31-47-197: ~
hadoop-1.2.1/src/contrib/ec2/bin/hadoop-ec2-env.sh
hadoop-1.2.1/src/contrib/ec2/bin/hadoop-ec2-init-remote.sh
hadoop-1.2.1/src/contrib/ec2/bin/image/create-hadoop-image-remote
hadoop-1.2.1/src/contrib/ec2/bin/image/ec2-run-user-data
hadoop-1.2.1/src/contrib/ec2/bin/launch-hadoop-cluster
hadoop-1.2.1/src/contrib/ec2/bin/launch-hadoop-master
hadoop-1.2.1/src/contrib/ec2/bin/launch-hadoop-slaves
hadoop-1.2.1/src/contrib/ec2/bin/list-hadoop-clusters
hadoop-1.2.1/src/contrib/ec2/bin/terminate-hadoop-cluster
ubuntu@ip-172-31-47-197:~$ mv hadoop-1.2.1 hadoop
ubuntu@ip-172-31-47-197:~$ ls
hadoop  hadoop-1.2.1.tar.gz
ubuntu@ip-172-31-47-197:~$ hadoop/bin/hadoop
Usage: hadoop [--config confdir] COMMAND
where COMMAND is one of:
    namenode -format      format the DFS filesystem
    secondarynamenode    run the DFS secondary namenode
    namenode              run the DFS namenode
    datanode              run a DFS datanode
    dfsadmin             run a DFS admin client
    mradmin              run a MapReduce admin client
    fsck                 run a DFS filesystem checking utility
    fs                   run a generic filesystem user client
    balancer              run a cluster balancing utility
    oiv                  apply the offline fsimage viewer to an fsimage
    fetchdt              fetch a delegation token from the NameNode
    jobtracker           run the MapReduce job Tracker node
    pipes                run a Pipes job
    tasktracker           run a MapReduce task Tracker node
    historyserver        run job history servers as a standalone daemon
    job                  manipulate MapReduce jobs
    queue                get information regarding JobQueues
    version              print the version
    jar <jar>            run a jar file
    distcp <srcurl> <desturl> copy file or directories recursively
    distcp2 <srcurl> <desturl> DistCp version 2
    archive -archiveName NAME -p <parent path> <src>* <dest> create a hadoop archive
    classpath             prints the class path needed to get the
                        Hadoop jar and the required libraries
    daemonlog             get/set the log level for each daemon
    or
    CLASSNAME            run the class named CLASSNAME
Most commands print help when invoked w/o parameters.
ubuntu@ip-172-31-47-197:~$

```

Setting up of Password-less SSH

We need to add the AWS EC2 Key File `keypairaws.pem` to SSH profile. In order to do that we will need to use following ssh utilities:

‘ssh-agent’ : Used as a background program that handles passwords for SSH private keys.

‘ssh-add’ : prompts the user for a private key password and adds it to the list maintained by ssh-agent.

Ssh Session is lost everytime upon shell exit and we have to repeat ssh-agent and ssh-add commands.

```

or
CLASSNAME run the class named CLASSNAME
Most commands print help when invoked w/o parameters.
ubuntu@ip-172-31-47-197:~$ cd
ubuntu@ip-172-31-47-197:~$ sudo bi .bashrc
sudo: bi: command not found
ubuntu@ip-172-31-47-197:~$ sudo vi .bashrc
ubuntu@ip-172-31-47-197:~$ source ~/.bashrc
ubuntu@ip-172-31-47-197:~$ echo $HADOOP_PREFIX
/home/ubuntu/hadoop
ubuntu@ip-172-31-47-197:~$ echo $HADOOP_CONF
/home/ubuntu/hadoop/conf
ubuntu@ip-172-31-47-197:~$ eval `ssh-agent`
Agent pid 10457
ubuntu@ip-172-31-47-197:~$ ssh-add keypairaws.pem
keypairaws.pem: No such file or directory
ubuntu@ip-172-31-47-197:~$ ssh-add keypairaws.pem
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
@           WARNING: UNPROTECTED PRIVATE KEY FILE!          @
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
Permissions 0664 for 'keypairaws.pem' are too open.
It is required that your private key files are NOT accessible by others.
This private key will be ignored.
ubuntu@ip-172-31-47-197:~$ chmod 644 authorized_keys
chmod: cannot access 'authorized_keys': No such file or directory
ubuntu@ip-172-31-47-197:~$ chmod 644 .ssh/authorized_keys
ubuntu@ip-172-31-47-197:~$ chmod 400 keypairaws.pem
ubuntu@ip-172-31-47-197:~$ eval `ssh-agent`
Agent pid 10519
ubuntu@ip-172-31-47-197:~$ ssh-add keypairaws.pem
Identity added: keypairaws.pem (keypairaws.pem)
ubuntu@ip-172-31-47-197:~$ vi $HADOOP_CONF/hadoop-env.sh
ubuntu@ip-172-31-47-197:~$ vi $HADOOP_CONF/core-site.xml
ubuntu@ip-172-31-47-197:~$ mkdir hdfsmp
ubuntu@ip-172-31-47-197:~$ ls
hadoop  hadoop-1.2.1.tar.gz  hdfsmp  keypairaws.pem
ubuntu@ip-172-31-47-197:~$ vi $HADOOP_CONF/core-site.xml
ubuntu@ip-172-31-47-197:~$ vi $HADOOP_CONF/hdfs-site.xml
ubuntu@ip-172-31-47-197:~$ vi $HADOOP_CONF/mapred-site.xml
ubuntu@ip-172-31-47-197:~$ vi $HADOOP_CONF/mapred-site.xml
ubuntu@ip-172-31-47-197:~$ vi $HADOOP_CONF/core-site.xml
ubuntu@ip-172-31-47-197:~$ vi $HADOOP_CONF/hdfs-site.xml
ubuntu@ip-172-31-47-197:~$ vi $HADOOP_CONF/mapred-site.xml
ubuntu@ip-172-31-47-197:~$

```

Setup for Running Word Count on 16 Nodes:

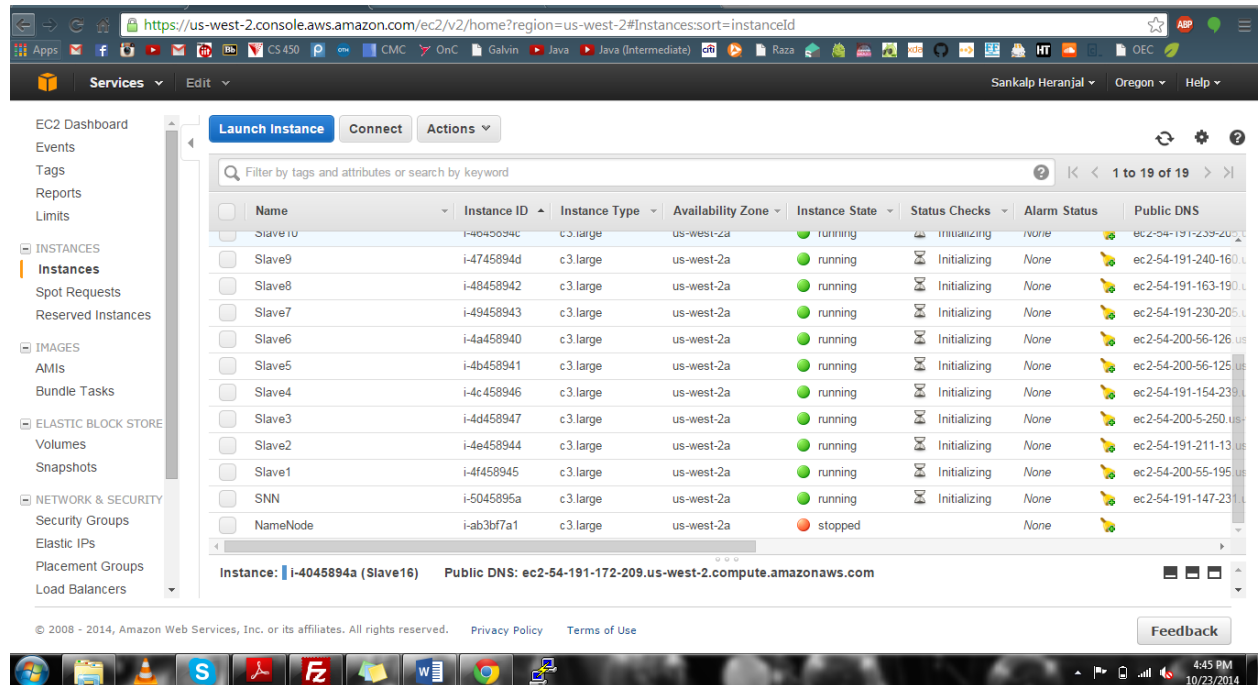
1. Hadoop Cluster Setup

We went to the `hadoop/conf` location and edited the following Configuration Files:

- Hadoop-env.sh**
 In this file we gave the path of the Java Home Directory. This file contains the environment variable settings. We use this to change the aspect of Hadoop daemon behavior, such as where log files are stored, the maximum amount of heap used etc.
- Core-site.xml**
 This file is used for NameNode Configuration. We are providing the Private IP of the instance we are creating and the Port Number i.e. 9000.
- Hdfs-site.xml**
 This file is used for the configuration of HDFS daemons, the NameNode, SecondaryNameNode and Data nodes. We have provided 2 properties i.e. `dfs.permissions.enabled` and `dfs.replication`.
 The first one is set to false which means the user can do anything they want to HDFS and the latter one is set to 16 since we are using 16 slaves.
- Mapred-site.xml**
 This file contains the configuration settings for MapReduce daemons, the job tracker and the task-trackers. We are providing the Private IP of the instance we are creating and the Port Number i.e. 9001 on which the Job Tracker runs.

2. Image Creation and Launching the 16 Slaves

Once we were done with the configuration of the instance, we kept this as the Master or the NameNode. Next, we created an image of the instance and launched 17 more instances with that image. These 17 instances comprised of a SecondaryNameNode and 16 Slaves.

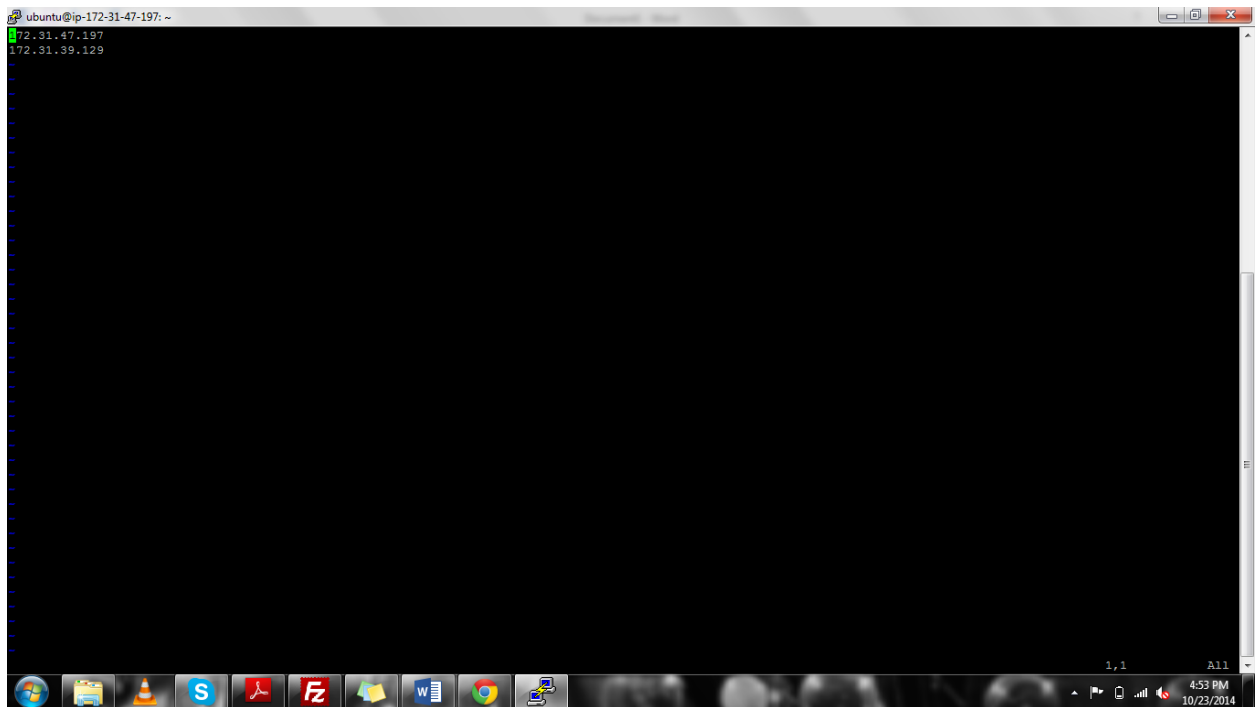


3. Configuration of Masters and Slaves Files

After launching all the instances, we configured the conf/Masters and conf/Slaves files on all the instances.

- Master/NameNode:**
 We took the master file and gave the private IPs of both NameNode and the SecondaryNameNode one after the other. In the Slaves file, we gave the private IPs of all the 16 Slaves.
- SecondaryNameNode:**
 In SNN, we gave the same configuration as that of the Master.
- Slaves/DataNode:**
 We provided the private IPs of each slave in the conf/Slaves file of that slave and left the conf/Masters file blank.

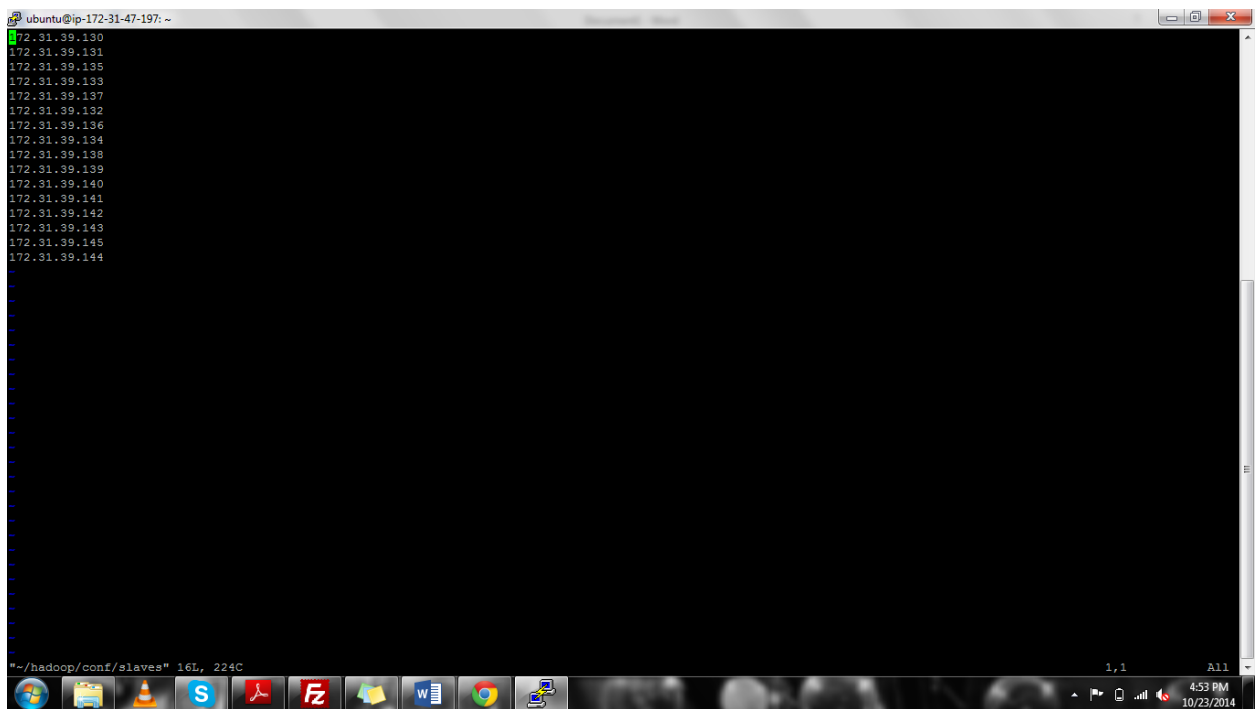
The NameNode Master and Slave File:



A terminal window titled 'ubuntu@ip-172-31-47-197: ~' showing the following output:

```
172.31.47.197
172.31.39.129
```

The terminal has a dark background with blue text. The Ubuntu logo is visible in the top left corner of the window. The bottom status bar shows '1,1 All' and the system clock '4:53 PM 10/23/2014'.



A terminal window titled 'ubuntu@ip-172-31-47-197: ~' showing a list of IP addresses:

```
172.31.39.130
172.31.39.131
172.31.39.135
172.31.39.133
172.31.39.137
172.31.39.132
172.31.39.136
172.31.39.134
172.31.39.138
172.31.39.139
172.31.39.140
172.31.39.141
172.31.39.142
172.31.39.143
172.31.39.145
172.31.39.144
```

The terminal has a dark background with blue text. The Ubuntu logo is visible in the top left corner of the window. The bottom status bar shows '1,1 All' and the system clock '4:53 PM 10/23/2014'.

4. Hadoop Daemon Startup

After configuring all the conf files, we went to the NameNode and Started the Hadoop File System. We did this by formatting the NameNode. After Formatting the NameNode, we started the Hadoop daemons by using bin/start.sh

CS553 Programming Assignment #2

```
ubuntu@ip-172-31-47-197: ~/hadoop/conf
ubuntu@ip-172-31-47-197:~$ vi $HADOOP_CONF/masters
ubuntu@ip-172-31-47-197:~$ vi $HADOOP_CONF/slaves
ubuntu@ip-172-31-47-197:~$ vi $HADOOP_CONF/masters
ubuntu@ip-172-31-47-197:~$ vi $HADOOP_CONF/slaves
ubuntu@ip-172-31-47-197:~$ cd /hadoop
ubuntu@ip-172-31-47-197:~/hadoop$ cd /conf
ubuntu@ip-172-31-47-197:~/hadoop/conf$ scp masters slaves ubuntu@ec2-54-191-147-231.us-west-2.compute.amazonaws.com:/home/ubuntu/hadoop/conf
The authenticity of host 'ec2-54-191-147-231.us-west-2.compute.amazonaws.com (172.31.39.129)' can't be established.
ECDSA key fingerprint is 93:73:15:43:ab:bd:53:3a:cb:41:32:1c:7b:c5:d1:02.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'ec2-54-191-147-231.us-west-2.compute.amazonaws.com,172.31.39.129' (ECDSA) to the list of known hosts.
masters      100% 28   0.0KB/s   00:00
slaves       100% 224  0.2KB/s   00:00
ubuntu@ip-172-31-47-197:~/hadoop/conf$ hadoop namenode -format
14/10/23 22:25:36 INFO namenode.NameNode: STARTUP MSG:
/*****
STARTUP_MSG: Starting NameNode
STARTUP_MSG: host = ip-172-31-47-197/172.31.47.197
STARTUP_MSG: args = [-format]
STARTUP_MSG: version = 1.2.1
STARTUP_MSG: build = https://svn.apache.org/repos/asf/hadoop/common/branches/branch-1.2 -r 1503152; compiled by 'mattf' on Mon Jul 22 15:23:09 PDT 2013
STARTUP_MSG: java = 1.7.0_65
*****/
14/10/23 22:25:37 INFO util.GSet: Computing capacity for map BlocksMap
14/10/23 22:25:37 INFO util.GSet: VM type = 64-bit
14/10/23 22:25:37 INFO util.GSet: 2.0% max memory = 932184064
14/10/23 22:25:37 INFO util.GSet: capacity = 2^21 = 2097152 entries
14/10/23 22:25:37 INFO util.GSet: recommended=2097152, actual=2097152
14/10/23 22:25:37 INFO namenode.FSNamesystem: fsOwner=ubuntu
14/10/23 22:25:37 INFO namenode.FSNamesystem: supergroup=supergroup
14/10/23 22:25:37 INFO namenode.FSNamesystem: isPermissionEnabled=false
14/10/23 22:25:37 INFO namenode.FSNamesystem: dfs.block.invalidate.limit=100
14/10/23 22:25:37 INFO namenode.FSNamesystem: sAccessTokenEnabled=false accessKeyUpdateInterval=0 min(s), accessTokenLifetime=0 min(s)
14/10/23 22:25:37 INFO namenode.FSEditLog: dfs.namenode.edits.tolerance.length = 0
14/10/23 22:25:37 INFO namenode.NameNode: Caching file names occurring more than 10 times
14/10/23 22:25:37 INFO common.Storage: Image file /home/ubuntu/hdfstmp/dfs/name/current/fsimage of size 112 bytes saved in 0 seconds.
14/10/23 22:25:37 INFO namenode.FSEditLog: closing edit log: position=4, editLog=/home/ubuntu/hdfstmp/dfs/name/current/edits
14/10/23 22:25:37 INFO namenode.FSEditLog: close success: truncate to 4, editLog=/home/ubuntu/hdfstmp/dfs/name/current/edits
14/10/23 22:25:37 INFO common.Storage: Storage directory /home/ubuntu/hdfstmp/dfs/name has been successfully formatted.
14/10/23 22:25:37 INFO namenode.NameNode: SHUTDOWN MSG:
/*****
SHUTDOWN_MSG: Shutting down NameNode at ip-172-31-47-197/172.31.47.197
*****/
ubuntu@ip-172-31-47-197:~/hadoop/conf$
```

We then ran jps on every instance to check whether the connection is established by the master with the slaves.

Master:

```
ubuntu@ip-172-31-47-197: ~/hadoop/conf
*****
ubuntu@ip-172-31-47-197:~/hadoop/conf$ start-all.sh
starting namenode, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-namenode-ip-172-31-47-197.out
172.31.39.136: starting datanode, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-datanode-ip-172-31-39-136.out
172.31.39.144: starting datanode, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-datanode-ip-172-31-39-144.out
172.31.39.138: starting datanode, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-datanode-ip-172-31-39-138.out
172.31.39.134: starting datanode, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-datanode-ip-172-31-39-134.out
172.31.39.141: starting datanode, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-datanode-ip-172-31-39-141.out
172.31.39.132: starting datanode, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-datanode-ip-172-31-39-132.out
172.31.39.140: starting datanode, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-datanode-ip-172-31-39-140.out
172.31.39.139: starting datanode, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-datanode-ip-172-31-39-139.out
172.31.39.142: starting datanode, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-datanode-ip-172-31-39-142.out
172.31.39.145: starting datanode, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-datanode-ip-172-31-39-145.out
172.31.39.137: starting datanode, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-datanode-ip-172-31-39-137.out
172.31.39.143: starting datanode, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-datanode-ip-172-31-39-143.out
172.31.39.130: starting datanode, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-datanode-ip-172-31-39-130.out
172.31.39.131: starting datanode, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-datanode-ip-172-31-39-131.out
172.31.39.133: starting datanode, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-datanode-ip-172-31-39-133.out
172.31.39.135: starting datanode, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-datanode-ip-172-31-39-135.out
172.31.47.197: starting secondarynamenode, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-secondarynamenode-ip-172-31-47-197.out
172.31.39.129: starting secondarynamenode, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-secondarynamenode-ip-172-31-39-129.out
starting jobtracker, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-jobtracker-ip-172-31-47-197.out
172.31.39.139: starting tasktracker, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-tasktracker-ip-172-31-39-139.out
172.31.39.138: starting tasktracker, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-tasktracker-ip-172-31-39-138.out
172.31.39.140: starting tasktracker, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-tasktracker-ip-172-31-39-140.out
172.31.39.143: starting tasktracker, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-tasktracker-ip-172-31-39-143.out
172.31.39.137: starting tasktracker, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-tasktracker-ip-172-31-39-137.out
172.31.39.134: starting tasktracker, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-tasktracker-ip-172-31-39-134.out
172.31.39.133: starting tasktracker, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-tasktracker-ip-172-31-39-133.out
172.31.39.130: starting tasktracker, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-tasktracker-ip-172-31-39-130.out
172.31.39.141: starting tasktracker, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-tasktracker-ip-172-31-39-141.out
172.31.39.136: starting tasktracker, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-tasktracker-ip-172-31-39-136.out
172.31.39.142: starting tasktracker, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-tasktracker-ip-172-31-39-142.out
172.31.39.135: starting tasktracker, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-tasktracker-ip-172-31-39-135.out
172.31.39.144: starting tasktracker, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-tasktracker-ip-172-31-39-144.out
172.31.39.132: starting tasktracker, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-tasktracker-ip-172-31-39-132.out
172.31.39.145: starting tasktracker, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-tasktracker-ip-172-31-39-145.out
172.31.39.131: starting tasktracker, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-tasktracker-ip-172-31-39-131.out
ubuntu@ip-172-31-47-197:~/hadoop/conf$ jps
2611 SecondaryNameNode
2634 Jps
2369 NameNode
2689 JobTracker
ubuntu@ip-172-31-47-197:~/hadoop/conf$
```


SecondaryNameNode:

```

ubuntu@ip-172-31-39-129: ~
login as: ubuntu
Authenticating with public key "imported-openssh-key"
Welcome to Ubuntu 14.04.1 LTS (GNU/Linux 3.13.0-36-generic x86_64)

 * Documentation:  https://help.ubuntu.com/

System information as of Thu Oct 23 21:56:42 UTC 2014

System load:  0.0           Processes:    106
Usage of /:   4.5% of 31.36GB Users logged in:   0
Memory usage: 2%           IP address for eth0: 172.31.39.129
Swap usage:   0%

Graph this data and manage this system at:
https://landscape.canonical.com/

Get cloud support with Ubuntu Advantage Cloud Guest:
http://www.ubuntu.com/business/services/cloud

40 packages can be updated.
20 updates are security updates.

Last login: Thu Oct 23 20:53:58 2014 from dhcpl17.nwvn2.lit.edu
ubuntu@ip-172-31-39-129:~$ eval `ssh-agent`
Agent pid 1481
ubuntu@ip-172-31-39-129:~$ ssh-add keypairaws.pem
Identity added: keypairaws.pem (keypairaws.pem)
ubuntu@ip-172-31-39-129:~$ jps
1483 Jps
ubuntu@ip-172-31-39-129:~$ vi $HADOOP_CONF/slaves
ubuntu@ip-172-31-39-129:~$ vi $HADOOP_CONF/master
ubuntu@ip-172-31-39-129:~$ vi $HADOOP_CONF/masters
ubuntu@ip-172-31-39-129:~$ jps
1656 SecondaryNameNode
1705 Jps
ubuntu@ip-172-31-39-129:~$

```

16 Slaves:

```

ubuntu@ip-172-31-39-130: ~
enable verbose output
-version          print product version and exit
-version:<value>  require the specified version to run
-showversion      print product version and continue
-jre-restrict-search | -no-jre-restrict-search
                  include/exclude user private JREs in the version search
-? -help          print this help message
-X               print help on non-standard options
-ea[:<packagename>...[:<classname>]]
-enableassertions[:<packagename>...[:<classname>]]
                  enable assertions with specified granularity
-da[:<packagename>...[:<classname>]]
-disableassertions[:<packagename>...[:<classname>]]
                  disable assertions with specified granularity
-esa | -enableassertions
                  enable system assertions
-dsa | -disableassertions
                  disable system assertions
-agentlib:<libname>[=<options>]
                  load native agent library <libname>, e.g. -agentlib:hprof
                  see also, -agentlib:jdwp=help and -agentlib:hprof=help
-agentpath:<pathname>[=<options>]
                  load native agent library by full pathname
-javaagent:<jarpath>[=<options>]
                  load Java programming language agent, see java.lang.instrument
-splash:<imagepath>
                  show splash screen with specified image
See http://www.oracle.com/technetwork/java/javase/documentation/index.html for more details.
ubuntu@ip-172-31-39-130:~$ vi $HADOOP_CONF/slaves
ubuntu@ip-172-31-39-130:~$ eval `ssh-agent`
Agent pid 1412
ubuntu@ip-172-31-39-130:~$ ssh-add keypairaws.pem
Identity added: keypairaws.pem (keypairaws.pem)
ubuntu@ip-172-31-39-130:~$ vi $HADOOP_CONF/slaves
ubuntu@ip-172-31-39-130:~$ vi $HADOOP_CONF/masters
ubuntu@ip-172-31-39-130:~$ jps
1432 Jps
ubuntu@ip-172-31-39-130:~$ jps
1820 TaskTracker
1681 DataNode
1893 Jps
ubuntu@ip-172-31-39-130:~$

```

Problem and Solution: At this point of time, we had an issue with the program to run. But we thought it was a problem in the Master Node Configuration. So we had to terminate the old NameNode (172.31.39.130) because of some configuration problems and created a new one (172.31.42.33) and

changed the configuration and masters/slaves files for the NameNode, SecondaryNameNode and the DataNodes according to that. It worked fine after that.

The following are the screenshots of the instances running after we created another master:

The screenshot shows the AWS Management Console interface. On the left, the navigation menu includes EC2 Dashboard, Events, Tags, Reports, Limits, INSTANCES, Spot Requests, Reserved Instances, IMAGES, AMIs, Bundle Tasks, ELASTIC BLOCK STORE, Volumes, Snapshots, NETWORK & SECURITY, Security Groups, Elastic IPs, Placement Groups, and Load Balancers. The main content area displays a list of EC2 instances. The 'NameNode' instance is selected, and its details are shown in the right pane. The details include:

Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Status	Public DNS
i-8ec40b84	c3.large	us-west-2a	running	2/2 checks...	None	ec2-54-191-185-126.us-west-2.compute.amazonaws.com

Below the table, the 'Description' tab is active, showing the instance's configuration, including the Private DNS, Private IP, and Security groups.

Master: NameNode, JobTracker and SecondaryNameNode are running

The screenshot shows a terminal window with the following output:

```

ubuntu@ip-172-31-42-33: ~/hadoop/bin
172.31.42.33: starting secondarynamenode, logging to /home/ubuntu/hadoop/libexec
./logs/hadoop-ubuntu-secondarynamenode-ip-172-31-42-33.out
172.31.39.129: starting secondarynamenode, logging to /home/ubuntu/hadoop/libexec
c:/logs/hadoop-ubuntu-secondarynamenode-ip-172-31-39-129.out
starting jobtracker, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu
u-jobtracker-ip-172-31-42-33.out
172.31.39.134: starting tasktracker, logging to /home/ubuntu/hadoop/libexec/./l
ogs/hadoop-ubuntu-tasktracker-ip-172-31-39-134.out
172.31.39.139: starting tasktracker, logging to /home/ubuntu/hadoop/libexec/./l
ogs/hadoop-ubuntu-tasktracker-ip-172-31-39-139.out
172.31.39.141: starting tasktracker, logging to /home/ubuntu/hadoop/libexec/./l
ogs/hadoop-ubuntu-tasktracker-ip-172-31-39-141.out
172.31.39.135: starting tasktracker, logging to /home/ubuntu/hadoop/libexec/./l
ogs/hadoop-ubuntu-tasktracker-ip-172-31-39-135.out
172.31.39.131: starting tasktracker, logging to /home/ubuntu/hadoop/libexec/./l
ogs/hadoop-ubuntu-tasktracker-ip-172-31-39-131.out
172.31.39.138: starting tasktracker, logging to /home/ubuntu/hadoop/libexec/./l
ogs/hadoop-ubuntu-tasktracker-ip-172-31-39-138.out
172.31.39.137: starting tasktracker, logging to /home/ubuntu/hadoop/libexec/./l
ogs/hadoop-ubuntu-tasktracker-ip-172-31-39-137.out
172.31.39.132: starting tasktracker, logging to /home/ubuntu/hadoop/libexec/./l
ogs/hadoop-ubuntu-tasktracker-ip-172-31-39-132.out
172.31.39.130: starting tasktracker, logging to /home/ubuntu/hadoop/libexec/./l
ogs/hadoop-ubuntu-tasktracker-ip-172-31-39-130.out
172.31.39.143: starting tasktracker, logging to /home/ubuntu/hadoop/libexec/./l
ogs/hadoop-ubuntu-tasktracker-ip-172-31-39-143.out
172.31.39.142: starting tasktracker, logging to /home/ubuntu/hadoop/libexec/./l
ogs/hadoop-ubuntu-tasktracker-ip-172-31-39-142.out
ubuntu@ip-172-31-42-33:~/hadoop/conf$ jps
1370 NameNode
1353 Jps
1707 JobTracker
1620 SecondaryNameNode
ubuntu@ip-172-31-42-33:~/hadoop/conf$ cd

```

16 Slaves: DataNode and TaskTracker are running

```

ubuntu@ip-172-31-39-134: ~
login as: ubuntu
Authenticating with public key "imported-openssh-key"
Welcome to Ubuntu 14.04.1 LTS (GNU/Linux 3.13.0-36-generic x86_64)

 * Documentation:  https://help.ubuntu.com/

System information as of Fri Oct 24 15:44:45 UTC 2014

System load:  0.09          Processes:      110
Usage of /:   35.9% of 31.36GB    Users logged in:  0
Memory usage: 4%              IP address for eth0: 172.31.39.134
Swap usage:   0%

Graph this data and manage this system at:
  https://landscape.canonical.com/

Get cloud support with Ubuntu Advantage Cloud Guest:
  http://www.ubuntu.com/business/services/cloud

40 packages can be updated.
20 updates are security updates.

Last login: Fri Oct 24 04:55:44 2014 from 208-59-145-165.c3-0.mcm-ubr1.chi-mcm.i
l.cable.rcn.com
ubuntu@ip-172-31-39-134:~$ vi $HADOOP_CONF/core-site.xml
ubuntu@ip-172-31-39-134:~$ ssh-add keypairaws.pem
Could not open a connection to your authentication agent.
ubuntu@ip-172-31-39-134:~$ eval `ssh-agent`
Agent pid 1572
ubuntu@ip-172-31-39-134:~$ ssh-add keypairaws.pem
Identity added: keypairaws.pem (keypairaws.pem)
ubuntu@ip-172-31-39-134:~$ jps
1426 TaskTracker
1574 Jps
1278 DataNode
ubuntu@ip-172-31-39-134:~$

```

5. Running the Word Count Program:

We created a Hadoop MapReduce word count program on eclipse and created a JAR file of the program > countword.jar

We copied the 10GB i.e. wiki10gb file into the Hadoop Filesystem in a folder Input.

Then we transferred the JAR file countword.jar in the home/ubuntu on our master instance.

After copying the required files, we ran the Jar file:

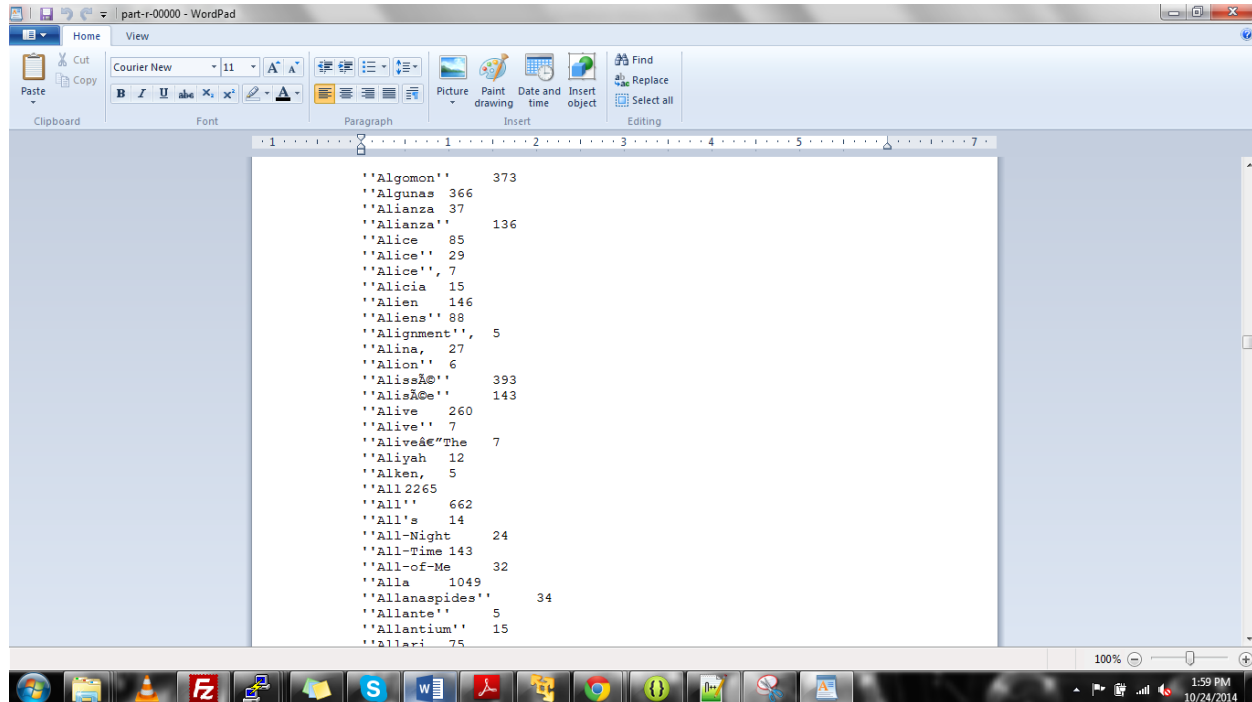
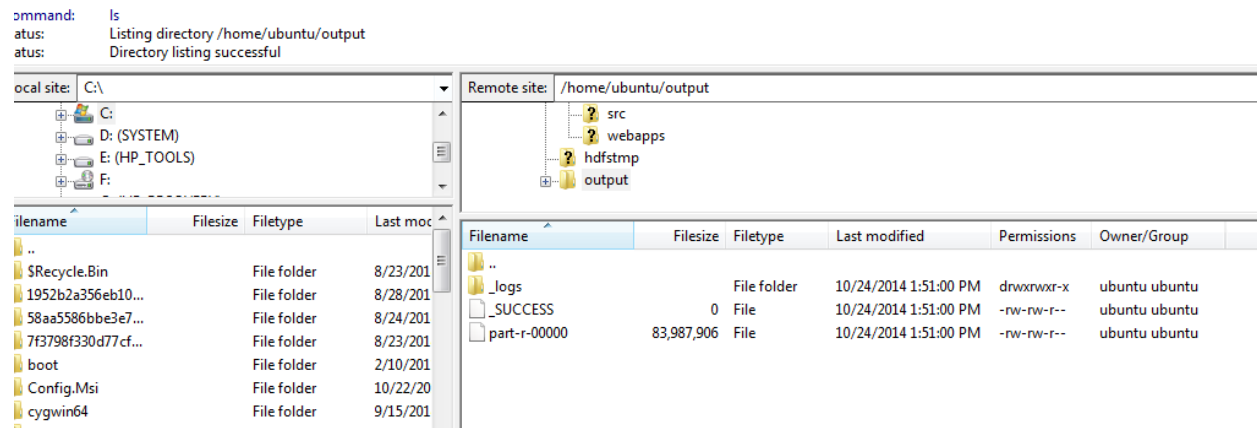
CS553 Programming Assignment #2

```
ubuntu@ip-172-31-42-33: ~  
ubuntu@ip-172-31-42-33:~/hadoop$ bin/hadoop jar countword.jar wordcount input output3  
14/10/24 18:43:23 INFO input.FileInputFormat: Total input paths to process : 1  
14/10/24 18:43:23 INFO util.NativeCodeLoader: Loaded the native-hadoop library  
14/10/24 18:43:23 WARN snappy.LoadSnappy: Snappy native library not loaded  
14/10/24 18:43:23 WARN split.JobSplitWriter: Max block location exceeded for split: hdfs://172.31.42.33:9000/user/ubuntu/input:0+67108864 splitsize: 16 maxsize: 10  
14/10/24 18:43:23 WARN split.JobSplitWriter: Max block location exceeded for split: hdfs://172.31.42.33:9000/user/ubuntu/input:67108864+67108864 splitsize: 16 maxsize: 10  
14/10/24 18:43:23 WARN split.JobSplitWriter: Max block location exceeded for split: hdfs://172.31.42.33:9000/user/ubuntu/input:134217728+67108864 splitsize: 16 maxsize: 10  
14/10/24 18:43:23 WARN split.JobSplitWriter: Max block location exceeded for split: hdfs://172.31.42.33:9000/user/ubuntu/input:201326592+67108864 splitsize: 16 maxsize: 10  
14/10/24 18:43:23 WARN split.JobSplitWriter: Max block location exceeded for split: hdfs://172.31.42.33:9000/user/ubuntu/input:268435456+67108864 splitsize: 16 maxsize: 10  
14/10/24 18:43:23 WARN split.JobSplitWriter: Max block location exceeded for split: hdfs://172.31.42.33:9000/user/ubuntu/input:335544320+67108864 splitsize: 16 maxsize: 10  
14/10/24 18:43:23 WARN split.JobSplitWriter: Max block location exceeded for split: hdfs://172.31.42.33:9000/user/ubuntu/input:402653184+67108864 splitsize: 16 maxsize: 10  
14/10/24 18:43:23 WARN split.JobSplitWriter: Max block location exceeded for split: hdfs://172.31.42.33:9000/user/ubuntu/input:469762048+67108864 splitsize: 16 maxsize: 10  
14/10/24 18:43:23 WARN split.JobSplitWriter: Max block location exceeded for split: hdfs://172.31.42.33:9000/user/ubuntu/input:536870912+67108864 splitsize: 16 maxsize: 10  
14/10/24 18:43:23 WARN split.JobSplitWriter: Max block location exceeded for split: hdfs://172.31.42.33:9000/user/ubuntu/input:603979776+67108864 splitsize: 16 maxsize: 10  
14/10/24 18:43:23 WARN split.JobSplitWriter: Max block location exceeded for split: hdfs://172.31.42.33:9000/user/ubuntu/input:671088640+67108864 splitsize: 16 maxsize: 10  
14/10/24 18:43:23 WARN split.JobSplitWriter: Max block location exceeded for split: hdfs://172.31.42.33:9000/user/ubuntu/input:738197504+67108864 splitsize: 16 maxsize: 10  
14/10/24 18:43:23 WARN split.JobSplitWriter: Max block location exceeded for split: hdfs://172.31.42.33:9000/user/ubuntu/input:805306368+67108864 splitsize: 16 maxsize: 10  
14/10/24 18:43:23 WARN split.JobSplitWriter: Max block location exceeded for split: hdfs://172.31.42.33:9000/user/ubuntu/input:872415232+67108864 splitsize: 16 maxsize: 10  
14/10/24 18:43:23 WARN split.JobSplitWriter: Max block location exceeded for split: hdfs://172.31.42.33:9000/user/ubuntu/input:939524096+67108864 splitsize: 16 maxsize: 10  
14/10/24 18:43:23 WARN split.JobSplitWriter: Max block location exceeded for split: hdfs://172.31.42.33:9000/user/ubuntu/input:1006632960+67108864 splitsize: 16 maxsize: 10  
14/10/24 18:43:23 WARN split.JobSplitWriter: Max block location exceeded for split: hdfs://172.31.42.33:9000/user/ubuntu/input:1073741824+67108864 splitsize: 16 maxsize: 10  
14/10/24 18:43:23 WARN split.JobSplitWriter: Max block location exceeded for split: hdfs://172.31.42.33:9000/user/ubuntu/input:1140850688+67108864 splitsize: 16 maxsize: 10  
14/10/24 18:43:23 WARN split.JobSplitWriter: Max block location exceeded for split: hdfs://172.31.42.33:9000/user/ubuntu/input:1207959552+67108864 splitsize: 16 maxsize: 10  
14/10/24 18:43:23 WARN split.JobSplitWriter: Max block location exceeded for split: hdfs://172.31.42.33:9000/user/ubuntu/input:1275068416+67108864 splitsize: 16 maxsize: 10  
14/10/24 18:43:23 WARN split.JobSplitWriter: Max block location exceeded for split: hdfs://172.31.42.33:9000/user/ubuntu/input:1342177280+67108864 splitsize: 16 maxsize: 10
```

```
ubuntu@ip-172-31-42-33: ~/hadoop  
14/10/24 18:45:57 INFO mapred.JobClient: Physical memory (bytes) snapshot=33567150080  
14/10/24 18:45:57 INFO mapred.JobClient: Reduce output records=3757036  
14/10/24 18:45:57 INFO mapred.JobClient: Virtual memory (bytes) snapshot=122504351744  
14/10/24 18:45:57 INFO mapred.JobClient: Map output records=1515550661  
Word Count Hadoop time: 155773 ms  
14/10/24 18:45:57 INFO mapred.JobClient: Running job: job_201410241544_0004  
14/10/24 18:45:57 INFO mapred.JobClient: Job complete: job_201410241544_0004  
14/10/24 18:45:57 INFO mapred.JobClient: Counters: 30  
14/10/24 18:45:57 INFO mapred.JobClient: Job Counters  
14/10/24 18:45:57 INFO mapred.JobClient: Launched reduce tasks=1  
14/10/24 18:45:57 INFO mapred.JobClient: SLOTS_MILLIS_MAPS=4176986  
14/10/24 18:45:57 INFO mapred.JobClient: Total time spent by all reduces waiting after reserving slots (ms)=0  
14/10/24 18:45:57 INFO mapred.JobClient: Total time spent by all maps waiting after reserving slots (ms)=0  
14/10/24 18:45:57 INFO mapred.JobClient: Rack-local map tasks=2  
14/10/24 18:45:57 INFO mapred.JobClient: Launched map tasks=173  
14/10/24 18:45:57 INFO mapred.JobClient: Data-local map tasks=171  
14/10/24 18:45:57 INFO mapred.JobClient: SLOTS_MILLIS_REDUCE=122553  
14/10/24 18:45:57 INFO mapred.JobClient: File Output Format Counters  
14/10/24 18:45:57 INFO mapred.JobClient: Bytes Written=83987906  
14/10/24 18:45:57 INFO mapred.JobClient: FileSystemCounters  
14/10/24 18:45:57 INFO mapred.JobClient: FILE_BYTES_READ=830064381  
14/10/24 18:45:57 INFO mapred.JobClient: HDFS_BYTES_READ=10400651465  
14/10/24 18:45:57 INFO mapred.JobClient: FILE_BYTES_WRITTEN=994687387  
14/10/24 18:45:57 INFO mapred.JobClient: HDFS_BYTES_WRITTEN=83987906  
14/10/24 18:45:57 INFO mapred.JobClient: File Input Format Counters  
14/10/24 18:45:57 INFO mapred.JobClient: Bytes Read=10400634880  
14/10/24 18:45:57 INFO mapred.JobClient: Map-Reduce Framework  
14/10/24 18:45:57 INFO mapred.JobClient: Map output materialized bytes=155929195  
14/10/24 18:45:57 INFO mapred.JobClient: Map input records=123015884  
14/10/24 18:45:57 INFO mapred.JobClient: Reduce shuffle bytes=155929195  
14/10/24 18:45:57 INFO mapred.JobClient: Spilled Records=55304477  
14/10/24 18:45:57 INFO mapred.JobClient: Map output bytes=15858375070  
14/10/24 18:45:57 INFO mapred.JobClient: Total committed heap usage (bytes)=27374649344  
14/10/24 18:45:57 INFO mapred.JobClient: CPU time spent (ms)=3651320  
14/10/24 18:45:57 INFO mapred.JobClient: Combine input records=1543959260  
14/10/24 18:45:57 INFO mapred.JobClient: SPLIT_RAW_BYTES=16585  
14/10/24 18:45:57 INFO mapred.JobClient: Reduce input records=5484981  
14/10/24 18:45:57 INFO mapred.JobClient: Reduce input groups=3757036  
14/10/24 18:45:57 INFO mapred.JobClient: Combine output records=33893580  
14/10/24 18:45:57 INFO mapred.JobClient: Physical memory (bytes) snapshot=33567150080  
14/10/24 18:45:57 INFO mapred.JobClient: Reduce output records=3757036  
14/10/24 18:45:57 INFO mapred.JobClient: Virtual memory (bytes) snapshot=122504351744  
14/10/24 18:45:57 INFO mapred.JobClient: Map output records=1515550661
```

The Program executed successfully on the 16 Nodes and we got an execution Time of **155773 milliseconds** which you can see in the screenshot above.

The Output Folder and the Output are shown below:



6. Stopping all the Daemons:

After we got the Output, we stopped all the daemons using bin/stop.sh

```

ubuntu@ip-172-31-42-33: ~/hadoop/bin
ubuntu@ip-172-31-42-33:~$ ls output
logs part-r-00000 SUCCESS
ubuntu@ip-172-31-42-33:~$ scp output/part-r-00000 wordcount-hadoop.txt
ubuntu@ip-172-31-42-33:~$ ls
hadoop hadoop-1.2.1.tar.gz hdfsmap keypairaws.pem output wiki10gb wordcount-hadoop.txt
ubuntu@ip-172-31-42-33:~$ cd hadoop/bin
ubuntu@ip-172-31-42-33:~/hadoop/bin$ stop-all.sh
stopping jobtracker
172.31.39.130: stopping tasktracker
172.31.39.144: stopping tasktracker
172.31.39.141: stopping tasktracker
172.31.39.140: stopping tasktracker
172.31.39.132: stopping tasktracker
172.31.39.143: stopping tasktracker
172.31.39.133: stopping tasktracker
172.31.39.137: stopping tasktracker
172.31.39.138: stopping tasktracker
172.31.39.134: stopping tasktracker
172.31.39.131: stopping tasktracker
172.31.39.142: stopping tasktracker
172.31.39.139: stopping tasktracker
172.31.39.136: stopping tasktracker
172.31.39.135: stopping tasktracker
172.31.39.145: stopping tasktracker
stopping namenode
172.31.39.139: stopping datanode
172.31.39.130: stopping datanode
172.31.39.141: stopping datanode
172.31.39.136: stopping datanode
172.31.39.132: stopping datanode
172.31.39.140: stopping datanode
172.31.39.143: stopping datanode
172.31.39.137: stopping datanode
172.31.39.133: stopping datanode
172.31.39.145: stopping datanode
172.31.39.142: stopping datanode
172.31.39.134: stopping datanode
172.31.39.144: stopping datanode
172.31.39.135: stopping datanode
172.31.39.138: stopping datanode
172.31.39.131: stopping datanode
172.31.42.33: stopping secondarynamenode
172.31.39.129: stopping secondarynamenode
ubuntu@ip-172-31-42-33:~/hadoop/bin$

```

Setup for Running Word Count on a Single Node:

1. Launch an Instance

We launched another Ubuntu 14.04 c3.large instance using the image with 32GB EBS Storage and used the same key file keypairaws.pem and uploaded it in the instance using FileZilla.

The screenshot shows the AWS Management Console interface. The left sidebar contains navigation links for EC2 Dashboard, Events, Tags, Reports, Limits, INSTANCES, Spot Requests, Reserved Instances, IMAGES, AMIs, Bundle Tasks, ELASTIC BLOCK STORE, Volumes, Snapshots, NETWORK & SECURITY, Security Groups, Elastic IPs, Placement Groups, and Load Balancers. The main content area displays a table of EC2 instances. The instance 'HadoopEC2SingleNodeCluster' is selected, and its details are shown on the right.

Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Status	Public DNS
Slave4	i-4c458946	c3.large	us-west-2a	stopped	2/2 checks...	None	ec2-54-69-197-214.us-west-2.compute.amazonaws.com
Slave3	i-4d458947	c3.large	us-west-2a	stopped	2/2 checks...	None	ec2-54-69-197-214.us-west-2.compute.amazonaws.com
Slave2	i-4e458944	c3.large	us-west-2a	stopped	2/2 checks...	None	ec2-54-69-197-214.us-west-2.compute.amazonaws.com
Slave1	i-4f458945	c3.large	us-west-2a	stopped	2/2 checks...	None	ec2-54-69-197-214.us-west-2.compute.amazonaws.com
SNN	i-5045895a	c3.large	us-west-2a	stopped	2/2 checks...	None	ec2-54-69-197-214.us-west-2.compute.amazonaws.com
NameNode	i-8ec40b84	c3.large	us-west-2a	stopped	2/2 checks...	None	ec2-54-69-197-214.us-west-2.compute.amazonaws.com
HadoopEC2SingleNodeCluster	i-e306c8e9	c3.large	us-west-2a	running	2/2 checks...	None	ec2-54-69-197-214.us-west-2.compute.amazonaws.com

Details for HadoopEC2SingleNodeCluster:

- Instance state: running
- Instance type: c3.large
- Private DNS: ip-172-31-47-146.us-west-2.compute.internal
- Private IPs: 172.31.47.146
- Secondary private IPs: -
- VPC ID: vpc-dfcb01ba
- Subnet ID: subnet-bbd472de
- Public IP: 54.69.197.214
- Elastic IP: -
- Availability zone: us-west-2a
- Security groups: launch-wizard-14, view rules
- Scheduled events: No scheduled events
- AMI ID: Slaves (ami-77783647)
- Platform: -

Then we started puTTY to start with our Hadoop word count. We setup a passphraseless ssh to the localhost using:

```
ssh-keygen -t dsa -P "" -f ~/.ssh/id_dsa
```

```
cat ~/.ssh/id_dsa.pub >> ~/.ssh/authorized_keys
```

2. Hadoop Configuration Files

We went to the `hadoop/conf` location and edited the following Configuration Files again for single node (since we created an instance of the image we took last time):

- **Core-site.xml**
We have kept it same as that in 16Nodes
- **Hdfs-site.xml**
We have provided the 2 properties i.e. `dfs.permissions.enabled` and `dfs.replication`.
The first one is set to false and the latter one is set to 1 since we are using 1 Node.
- **Mapred-site.xml**
Even this file is kept the same as 16 Nodes.
- **Masters**
We removed IPs given earlier and gave the localhost
- **Slaves**
Even in Slaves we removed the previous slave IPs and gave the localhost.

3. Hadoop Daemon Startup

After configuring all the conf files, we started the Hadoop File System. We did this by formatting the NameNode. After Formatting the NameNode, we started the Hadoop daemons by using `bin/start.sh`

CS553 Programming Assignment #2

```

ubuntu@ip-172-31-47-146: ~$ hadoop
Graph this data and manage this system at:
  https://landscape.canonical.com/

Get cloud support with Ubuntu Advantage Cloud Guest:
  http://www.ubuntu.com/business/services/cloud

40 packages can be updated.
20 updates are security updates.

Last login: Fri Oct 24 20:23:10 2014 from 208-59-146-165.c3-0.mcm-ubr1.chi-mcm.i
1.cable.rcn.com
ubuntu@ip-172-31-47-146:~$ cd hadoop
ubuntu@ip-172-31-47-146:~/hadoop$ bin/hadoop namenode -format
14/10/24 20:31:04 INFO namenode.NameNode: STARTUP_MSG:
/*****
STARTUP_MSG: Starting NameNode
STARTUP_MSG: host = ip-172-31-47-146/172.31.47.146
STARTUP_MSG: args = [-format]
STARTUP_MSG: version = 1.2.1
STARTUP_MSG: build = https://svn.apache.org/repos/asf/hadoop/common/branches/branch-1.2 -r 1503152; compiled by 'mattf' on Mon Jul 22 15:23:09 PDT 2013
STARTUP_MSG: java = 1.7.0_65
*****/
14/10/24 20:31:04 INFO util.GSet: Computing capacity for map BlocksMap
14/10/24 20:31:04 INFO util.GSet: VM type = 64-bit
14/10/24 20:31:04 INFO util.GSet: 2.0% max memory = 932184064
14/10/24 20:31:04 INFO util.GSet: capacity = 2^21 = 2097152 entries
14/10/24 20:31:04 INFO util.GSet: recommended=2097152, actual=2097152
14/10/24 20:31:04 INFO namenode.FSNamesystem: fsOwner=ubuntu
14/10/24 20:31:04 INFO namenode.FSNamesystem: supergroup=supergroup
14/10/24 20:31:04 INFO namenode.FSNamesystem: isPermissionEnabled=true
14/10/24 20:31:04 INFO namenode.FSNamesystem: dfs.block.invalidate.limit=100
14/10/24 20:31:04 INFO namenode.FSNamesystem: isAccessTokenEnabled=false accessKeyUpdateInterval=0 min(s), accessTokenLifetime=0 min(s)
14/10/24 20:31:04 INFO namenode.FSEditLog: dfs.namenode.edits.tolerant.length = 0
14/10/24 20:31:04 INFO namenode.NameNode: Caching file names occurring more than 10 times
14/10/24 20:31:05 INFO common.Storage: Tag: file /tmp/hadoop-ubuntu/dfs/name/current/fsimage of size 112 bytes saved in 0 seconds.
14/10/24 20:31:05 INFO namenode.FSEditLog: closing edit log: position=4, editlog=/tmp/hadoop-ubuntu/dfs/name/current/edits
14/10/24 20:31:05 INFO namenode.FSEditLog: close success: truncate to 4, editlog=/tmp/hadoop-ubuntu/dfs/name/current/edits
14/10/24 20:31:05 INFO common.Storage: Storage directory /tmp/hadoop-ubuntu/dfs/name has been successfully formatted.
14/10/24 20:31:05 INFO namenode.NameNode: SHUTDOWN_MSG:
*****/
SHUTDOWN_MSG: Shutting down NameNode at ip-172-31-47-146/172.31.47.146
*****/
ubuntu@ip-172-31-47-146:~/hadoop$ bin/start-all.sh

```

```

ubuntu@ip-172-31-47-146: ~/hadoo
. . OE = .
. B S = .
+ + o o o
. . .
-----
ubuntu@ip-172-31-47-146:~/hadoop$ cat ~/.ssh/id_dsa.pub >> ~/.ssh/authorized_keys
ubuntu@ip-172-31-47-146:~/hadoop$ bin/stop-all.sh
stopping jobtracker
localhost: no tasktracker to stop
stopping namenode
localhost: no datanode to stop
localhost: no secondarynamenode to stop
ubuntu@ip-172-31-47-146:~/hadoop$ bin/hadoop namenode -format
14/10/24 20:37:06 INFO namenode.NameNode: STARTUP_MSG:
/*****
STARTUP_MSG: Starting NameNode
STARTUP_MSG: host = ip-172-31-47-146/172.31.47.146
STARTUP_MSG: args = [-format]
STARTUP_MSG: version = 1.2.1
STARTUP_MSG: build = https://svn.apache.org/repos/asf/hadoop/common/branches/branch-1.2 -r 1503152; compiled by 'mattf' on Mon Jul 22 15:23:09 PDT 2013
STARTUP_MSG: java = 1.7.0_65
*****/
Re-format filesystem in /tmp/hadoop-ubuntu/dfs/name ? (Y or N) y
format aborted in /tmp/hadoop-ubuntu/dfs/name
14/10/24 20:37:09 INFO namenode.NameNode: SHUTDOWN_MSG:
*****/
SHUTDOWN_MSG: Shutting down NameNode at ip-172-31-47-146/172.31.47.146
*****/
ubuntu@ip-172-31-47-146:~/hadoop$ bin/start-all.sh
starting namenode, logging to /home/ubuntu/hadoop/libexec/../logs/hadoop-ubuntu-namenode-ip-172-31-47-146.out
localhost: starting datanode, logging to /home/ubuntu/hadoop/libexec/../logs/hadoop-ubuntu-datanode-ip-172-31-47-146.out
localhost: starting secondarynamenode, logging to /home/ubuntu/hadoop/libexec/../logs/hadoop-ubuntu-secondarynamenode-ip-172-31-47-146.out
starting jobtracker, logging to /home/ubuntu/hadoop/libexec/../logs/hadoop-ubuntu-jobtracker-ip-172-31-47-146.out
localhost: starting tasktracker, logging to /home/ubuntu/hadoop/libexec/../logs/hadoop-ubuntu-tasktracker-ip-172-31-47-146.out
ubuntu@ip-172-31-47-146:~/hadoop$ jps
2690 JobTracker
2457 NameNode
3158 Jps
3045 TaskTracker
2618 DataNode
2789 SecondaryNameNode
ubuntu@ip-172-31-47-146:~/hadoop$

```

We then ran `jps` and we can see that the `NameNode`, `SecondaryNameNode`, `JobTracker`, `TaskTracker` and the `DataNode` is working in the above screen.

4. Running the Word Count Program:

We took the previously used Hadoop MapReduce word count program > countword.jar for the single node.

We again copied the 10GB i.e. wiki10gb file into the Hadoop Filesystem in a folder Input.

Then we transferred the JAR file countword.jar in the home/ubuntu on our instance.

After copying the required files, we ran the Jar file:

```
ubuntu@ip-172-31-47-146:~$
localhost: starting secondarynamenode, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-secondarynamenode-ip-172-31-47-146.out
starting jobtracker, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-jobtracker-ip-172-31-47-146.out
localhost: starting tasktracker, logging to /home/ubuntu/hadoop/libexec/./logs/hadoop-ubuntu-tasktracker-ip-172-31-47-146.out
ubuntu@ip-172-31-47-146:~/hadoop$ jps
2880 JobTracker
2457 NameNode
3158 Jps
3045 TaskTracker
2618 DataNode
2789 SecondaryNameNode
ubuntu@ip-172-31-47-146:~/hadoop$ cd
ubuntu@ip-172-31-47-146:~$ wget https://s3.amazonaws.com/cs-553/wiki10gb.xz
--2014-10-24 20:41:52-- https://s3.amazonaws.com/cs-553/wiki10gb.xz
Resolving s3.amazonaws.com (s3.amazonaws.com)... 54.231.244.8
Connecting to s3.amazonaws.com (s3.amazonaws.com)[54.231.244.8]:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 86933472 (83M) [application/octet-stream]
Saving to: 'wiki10gb.xz'

100%[=====] 86,933,472 3.25MB/s in 26s

2014-10-24 20:42:19 (3.21 MB/s) - 'wiki10gb.xz' saved [86933472/86933472]

ubuntu@ip-172-31-47-146:~$ xz -d wiki10gb.xz
ubuntu@ip-172-31-47-146:~$ ls
hadoop  hadoop-1.2.1.tar.gz  hdfsmap  keypairaws.pem  wiki10gb
ubuntu@ip-172-31-47-146:~$ cd hadoop
ubuntu@ip-172-31-47-146:~/hadoop$ bin/hadoop fs -put conf ^Cput
ubuntu@ip-172-31-47-146:~/hadoop$ cd
ubuntu@ip-172-31-47-146:~$ hadoop/bin/hadoop fs -put wiki10gb input
ubuntu@ip-172-31-47-146:~$ hadoop/bin/hadoop fs -ls /
Found 2 items
drwxr-xr-x - ubuntu supergroup 0 2014-10-24 20:35 /tmp
drwxr-xr-x - ubuntu supergroup 0 2014-10-24 20:46 /user
ubuntu@ip-172-31-47-146:~$ hadoop/bin/hadoop fs -ls /user/ubuntu
Found 1 items
-rw-r--r-- 1 ubuntu supergroup 10400000000 2014-10-24 20:46 /user/ubuntu/input
ubuntu@ip-172-31-47-146:~$ hadoop/bin/hadoop jar hadoop/countword.jar Wordcount input output
14/10/24 20:58:44 INFO input.FileInputFormat: Total input paths to process : 1
14/10/24 20:58:44 INFO util.NativeCodeLoader: Loaded the native-hadoop library
14/10/24 20:58:44 WARN snappy.LoadSnappy: Snappy native library not loaded
14/10/24 20:58:44 INFO mapred.JobClient: Running job: job_201410242037_0001
14/10/24 20:58:45 INFO mapred.JobClient: map 0% reduce 0%
```

CS553 Programming Assignment #2

```
ubuntu@ip-172-31-47-146: ~  
14/10/24 21:25:45 INFO mapred.JobClient: map 77% reduce 25%  
14/10/24 21:26:05 INFO mapred.JobClient: map 78% reduce 25%  
14/10/24 21:26:17 INFO mapred.JobClient: map 78% reduce 26%  
14/10/24 21:26:25 INFO mapred.JobClient: map 79% reduce 26%  
14/10/24 21:26:46 INFO mapred.JobClient: map 80% reduce 26%  
14/10/24 21:27:04 INFO mapred.JobClient: map 81% reduce 26%  
14/10/24 21:27:20 INFO mapred.JobClient: map 81% reduce 27%  
14/10/24 21:27:29 INFO mapred.JobClient: map 82% reduce 27%  
14/10/24 21:27:53 INFO mapred.JobClient: map 83% reduce 27%  
14/10/24 21:28:13 INFO mapred.JobClient: map 84% reduce 27%  
14/10/24 21:28:29 INFO mapred.JobClient: map 85% reduce 27%  
14/10/24 21:28:35 INFO mapred.JobClient: map 85% reduce 28%  
14/10/24 21:28:53 INFO mapred.JobClient: map 86% reduce 28%  
14/10/24 21:29:13 INFO mapred.JobClient: map 87% reduce 28%  
14/10/24 21:29:27 INFO mapred.JobClient: map 87% reduce 29%  
14/10/24 21:29:37 INFO mapred.JobClient: map 88% reduce 29%  
14/10/24 21:29:56 INFO mapred.JobClient: map 89% reduce 29%  
14/10/24 21:30:15 INFO mapred.JobClient: map 90% reduce 29%  
14/10/24 21:30:27 INFO mapred.JobClient: map 90% reduce 30%  
14/10/24 21:30:38 INFO mapred.JobClient: map 91% reduce 30%  
14/10/24 21:31:00 INFO mapred.JobClient: map 92% reduce 30%  
14/10/24 21:31:20 INFO mapred.JobClient: map 93% reduce 30%  
14/10/24 21:31:42 INFO mapred.JobClient: map 94% reduce 31%  
14/10/24 21:32:03 INFO mapred.JobClient: map 95% reduce 31%  
14/10/24 21:32:25 INFO mapred.JobClient: map 96% reduce 31%  
14/10/24 21:32:36 INFO mapred.JobClient: map 96% reduce 32%  
14/10/24 21:32:46 INFO mapred.JobClient: map 97% reduce 32%  
14/10/24 21:33:06 INFO mapred.JobClient: map 98% reduce 32%  
14/10/24 21:33:23 INFO mapred.JobClient: map 99% reduce 32%  
14/10/24 21:33:38 INFO mapred.JobClient: map 99% reduce 33%  
14/10/24 21:33:44 INFO mapred.JobClient: map 100% reduce 33%  
14/10/24 21:33:52 INFO mapred.JobClient: map 100% reduce 34%  
14/10/24 21:33:54 INFO mapred.JobClient: map 100% reduce 100%  
14/10/24 21:33:55 INFO mapred.JobClient: Job complete: job_201410242037_0001  
14/10/24 21:33:55 INFO mapred.JobClient: Counters: 29  
14/10/24 21:33:55 INFO mapred.JobClient: Job Counters  
14/10/24 21:33:55 INFO mapred.JobClient: Launched reduce tasks=1  
14/10/24 21:33:55 INFO mapred.JobClient: SLOTS_MILLIS_MAPS=4157938  
14/10/24 21:33:55 INFO mapred.JobClient: Total time spent by all reduces waiting after reserving slots (ms)=0  
14/10/24 21:33:55 INFO mapred.JobClient: Total time spent by all maps waiting after reserving slots (ms)=0  
14/10/24 21:33:55 INFO mapred.JobClient: Launched map tasks=155  
14/10/24 21:33:55 INFO mapred.JobClient: Data-local map tasks=155  
14/10/24 21:33:55 INFO mapred.JobClient: SLOTS_MILLIS_REDUCE=1988790  
14/10/24 21:33:55 INFO mapred.JobClient: File Output Format Counters
```

```
ubuntu@ip-172-31-47-146: ~  
14/10/24 21:33:55 INFO mapred.JobClient: Combine output records=33899938  
14/10/24 21:33:55 INFO mapred.JobClient: Physical memory (bytes) snapshot=33987137536  
14/10/24 21:33:55 INFO mapred.JobClient: Reduce output records=3757036  
14/10/24 21:33:55 INFO mapred.JobClient: Virtual memory (bytes) snapshot=122518085632  
14/10/24 21:33:55 INFO mapred.JobClient: Map output records=1515550661  
Word Count Hadoop time: 2111712 ms  
14/10/24 21:33:55 INFO mapred.JobClient: Running job: job_201410242037_0001  
14/10/24 21:33:55 INFO mapred.JobClient: Job complete: job_201410242037_0001  
14/10/24 21:33:55 INFO mapred.JobClient: Counters: 29  
14/10/24 21:33:55 INFO mapred.JobClient: Job Counters  
14/10/24 21:33:55 INFO mapred.JobClient: Launched reduce tasks=1  
14/10/24 21:33:55 INFO mapred.JobClient: SLOTS_MILLIS_MAPS=4157938  
14/10/24 21:33:55 INFO mapred.JobClient: Total time spent by all reduces waiting after reserving slots (ms)=0  
14/10/24 21:33:55 INFO mapred.JobClient: Total time spent by all maps waiting after reserving slots (ms)=0  
14/10/24 21:33:55 INFO mapred.JobClient: Launched map tasks=155  
14/10/24 21:33:55 INFO mapred.JobClient: Data-local map tasks=155  
14/10/24 21:33:55 INFO mapred.JobClient: SLOTS_MILLIS_REDUCE=1988790  
14/10/24 21:33:55 INFO mapred.JobClient: File Output Format Counters  
14/10/24 21:33:55 INFO mapred.JobClient: Bytes Written=83987906  
14/10/24 21:33:55 INFO mapred.JobClient: FileSystemCounters  
14/10/24 21:33:55 INFO mapred.JobClient: FILE_BYTES_READ=829718296  
14/10/24 21:33:55 INFO mapred.JobClient: HDFS_BYTES_READ=10400651000  
14/10/24 21:33:55 INFO mapred.JobClient: FILE_BYTES_WRITTEN=994337714  
14/10/24 21:33:55 INFO mapred.JobClient: HDFS_BYTES_WRITTEN=83987906  
14/10/24 21:33:55 INFO mapred.JobClient: File Input Format Counters  
14/10/24 21:33:55 INFO mapred.JobClient: Bytes Read=10400634880  
14/10/24 21:33:55 INFO mapred.JobClient: Map-Reduce Framework  
14/10/24 21:33:55 INFO mapred.JobClient: Map output materialized bytes=155929195  
14/10/24 21:33:55 INFO mapred.JobClient: Map input records=123015884  
14/10/24 21:33:55 INFO mapred.JobClient: Reduce shuffle bytes=155929195  
14/10/24 21:33:55 INFO mapred.JobClient: Spilled Records=55282978  
14/10/24 21:33:55 INFO mapred.JobClient: Map output bytes=15858375070  
14/10/24 21:33:55 INFO mapred.JobClient: Total committed heap usage (bytes)=27292336128  
14/10/24 21:33:55 INFO mapred.JobClient: CPU time spent (ms)=3663720  
14/10/24 21:33:55 INFO mapred.JobClient: Combine input records=1543987117  
14/10/24 21:33:55 INFO mapred.JobClient: SPLIT_RAW_BYTES=16120  
14/10/24 21:33:55 INFO mapred.JobClient: Reduce input records=5463482  
14/10/24 21:33:55 INFO mapred.JobClient: Reduce input groups=3757036  
14/10/24 21:33:55 INFO mapred.JobClient: Combine output records=33899938  
14/10/24 21:33:55 INFO mapred.JobClient: Physical memory (bytes) snapshot=33987137536  
14/10/24 21:33:55 INFO mapred.JobClient: Reduce output records=3757036  
14/10/24 21:33:55 INFO mapred.JobClient: Virtual memory (bytes) snapshot=122518085632  
14/10/24 21:33:55 INFO mapred.JobClient: Map output records=1515550661  
ubuntu@ip-172-31-47-146:~$
```

The Program executed successfully on the Single Node and we got an execution Time of **2111712 milliseconds** which you can see in the screenshot above.

The Output Folder and the Output are shown below:

```
ubuntu@ip-172-31-47-146:~$ hadoop/bin/hadoop fs -ls /user/ubuntu
Found 2 items
-rw-r--r-- 1 ubuntu supergroup 1040000000 2014-10-24 20:46 /user/ubuntu/input
drwxr-xr-x - ubuntu supergroup 0 2014-10-24 21:33 /user/ubuntu/output
ubuntu@ip-172-31-47-146:~$ hadoop/bin/hadoop fs -ls /user/ubuntu/output
Found 3 items
-rw-r--r-- 1 ubuntu supergroup 0 2014-10-24 21:33 /user/ubuntu/output/_SUCCESS
drwxr-xr-x - ubuntu supergroup 0 2014-10-24 20:58 /user/ubuntu/output/_logs
-rw-r--r-- 1 ubuntu supergroup 83987906 2014-10-24 21:33 /user/ubuntu/output/part-r-00000
ubuntu@ip-172-31-47-146:~$ hadoop/bin/hadoop fs -put output output
put: File output does not exist.
ubuntu@ip-172-31-47-146:~$ hadoop/bin/hadoop fs -get output output
ubuntu@ip-172-31-47-146:~$ ls
hadoop hadoop-1.2.1.tar.gz hdfsmp keypairaws.pem output wiki10gb
ubuntu@ip-172-31-47-146:~$ ls output
logs part-r-00000 SUCCESS
ubuntu@ip-172-31-47-146:~$
```

New directory is: "/home/ubuntu/output"
ls
Listing directory /home/ubuntu/output
Directory listing successful

Documents
puter

Remote site: /home/ubuntu/output

ssh

hadoop

hdfsmp

output

Filename	Filesize	Filetype	Last modified	Permissions	Owner/Group
..					
.._logs		File folder	10/24/2014 4:38:00 PM	drwxrwxr-x	ubuntu ubuntu
.._SUCCESS	0	File	10/24/2014 4:38:00 PM	-rw-rw-r--	ubuntu ubuntu
..part-r-00000	83,987,906	File	10/24/2014 4:38:00 PM	-rw-rw-r--	ubuntu ubuntu

wordcount-hadoop - WordPad

Home View

Courier New 11

Cut Copy

Font Paragraph Insert Editing

1 2 3 4 5 6 7

'''Categorize''' 25

'''Categorize''' 13

'''Category''' 172

'''Category''' 110

'''Catel-Manzke''' 22

'''Catel&Manzke''' 3

'''Caterpillar''' 477

'''Catharpin''' 21

'''Cathedral''' 4

'''Catherine''' 558

'''Catholic''' 27

'''Cathrin''' 247

'''Catlyin''' 5

'''Catsuit''' 1

'''Caulder''' 6

'''Cavenger''' 162

'''Caves''' 32

'''Cav&jos''' 135

'''Caxias''' 44

'''Cayman''' 211

'''Cayuga''' 14

'''Cazin''' 26

'''Ceoil''' 47

'''Cecilia''' 102

'''Celebrities''' 33

'''Celebrity''' 115

'''Celestia''' 54

'''Celica''' 76

'''Cellphone''' 201

'''Census''' 15

'''Centara''' 109

100%

4:50 PM 10/24/2014

5. Stopping all the Daemons:

After we got the Output, we stopped all the daemons using bin/stop.sh

```
ubuntu@ip-172-31-47-146:~$
14/10/24 21:33:55 INFO mapred.JobClient: Bytes Read=10400634880
14/10/24 21:33:55 INFO mapred.JobClient: Map-Reduce Framework
14/10/24 21:33:55 INFO mapred.JobClient: Map output materialized bytes=155929195
14/10/24 21:33:55 INFO mapred.JobClient: Map input records=123015884
14/10/24 21:33:55 INFO mapred.JobClient: Reduce shuffle bytes=155929195
14/10/24 21:33:55 INFO mapred.JobClient: Spilled Records=55282978
14/10/24 21:33:55 INFO mapred.JobClient: Map output bytes=155929195
14/10/24 21:33:55 INFO mapred.JobClient: Total committed heap usage (bytes)=27292336128
14/10/24 21:33:55 INFO mapred.JobClient: CPU time spent (ms)=3663720
14/10/24 21:33:55 INFO mapred.JobClient: Combine input records=1543987117
14/10/24 21:33:55 INFO mapred.JobClient: SPLIT_RAW_BYTES=16120
14/10/24 21:33:55 INFO mapred.JobClient: Reduce input records=5463482
14/10/24 21:33:55 INFO mapred.JobClient: Reduce input groups=3757036
14/10/24 21:33:55 INFO mapred.JobClient: Combine output records=33899938
14/10/24 21:33:55 INFO mapred.JobClient: Physical memory (bytes) snapshot=33987137536
14/10/24 21:33:55 INFO mapred.JobClient: Reduce output records=3757036
14/10/24 21:33:55 INFO mapred.JobClient: Virtual memory (bytes) snapshot=122518085632
14/10/24 21:33:55 INFO mapred.JobClient: Map output records=1515550661
ubuntu@ip-172-31-47-146:~$ hadoop/bin/hadoop fs -ls /user/ubuntu
Found 2 items
-rw-r--r-- 1 ubuntu supergroup 10400000000 2014-10-24 20:46 /user/ubuntu/input
drwxr-xr-x - ubuntu supergroup 0 2014-10-24 21:33 /user/ubuntu/output
ubuntu@ip-172-31-47-146:~$ hadoop/bin/hadoop fs -ls /user/ubuntu/output
Found 5 items
-rw-r--r-- 1 ubuntu supergroup 0 2014-10-24 21:33 /user/ubuntu/output/_SUCCESS
drwxr-xr-x - ubuntu supergroup 0 2014-10-24 20:58 /user/ubuntu/output/_logs
-rw-r--r-- 1 ubuntu supergroup 83987906 2014-10-24 21:33 /user/ubuntu/output/part-r-00000
ubuntu@ip-172-31-47-146:~$ hadoop/bin/hadoop fs -put output output
put: File output does not exist.
ubuntu@ip-172-31-47-146:~$ hadoop/bin/hadoop fs -get output output
ubuntu@ip-172-31-47-146:~$ ls
hadoop-1.2.1.tar.gz hdfsmap keypairaws.pem output wiki10gb
ubuntu@ip-172-31-47-146:~$ ls output
logs part-r-00000 _SUCCESS
ubuntu@ip-172-31-47-146:~$ scp output/part-r-00000 wordcount-hadoop.txt
ubuntu@ip-172-31-47-146:~$ ls
hadoop-1.2.1.tar.gz hdfsmap keypairaws.pem output wiki10gb wordcount-hadoop.txt
ubuntu@ip-172-31-47-146:~$ hadoop/bin/stop-all.sh
stopping jobtracker
localhost: stopping tasktracker
stopping namenode
localhost: stopping datanode
localhost: stopping secondarynamenode
ubuntu@ip-172-31-47-146:~$
```

After the Execution of the Program, we stopped all the running instances:

The screenshot shows the AWS Management Console interface. The left sidebar contains navigation links for EC2 Dashboard, Events, Tags, Reports, Limits, INSTANCES, Spot Requests, Reserved Instances, IMAGES, AMIs, Bundle Tasks, ELASTIC BLOCK STORE, Volumes, Snapshots, NETWORK & SECURITY, Security Groups, Elastic IPs, Placement Groups, and Load Balancers. The main content area displays a table of EC2 instances. The table has columns for Name, Instance ID, Instance Type, Availability Zone, Instance State, Status Checks, Alarm Status, and Public DNS. The instances listed are Slave9, Slave8, Slave7, Slave6, Slave5, Slave4, Slave3, Slave2, Slave1, SNN, NameNode, and HadoopEC2SingleNodeCluster. The HadoopEC2SingleNodeCluster instance is selected, and its details are shown at the bottom: Instance: i-e306c8e9 (HadoopEC2SingleNodeCluster) Private IP: 172.31.47.146.

Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Status	Public DNS
Slave9	i-4745894d	c3.large	us-west-2a	stopped	None		
Slave8	i-48458942	c3.large	us-west-2a	stopped	None		
Slave7	i-49458943	c3.large	us-west-2a	stopped	None		
Slave6	i-4a458940	c3.large	us-west-2a	stopped	None		
Slave5	i-4b458941	c3.large	us-west-2a	stopped	None		
Slave4	i-4c458946	c3.large	us-west-2a	stopped	None		
Slave3	i-4d458947	c3.large	us-west-2a	stopped	None		
Slave2	i-4e458944	c3.large	us-west-2a	stopped	None		
Slave1	i-4f458945	c3.large	us-west-2a	stopped	None		
SNN	i-5045895a	c3.large	us-west-2a	stopped	None		
NameNode	i-8ec40b84	c3.large	us-west-2a	stopped	None		
HadoopEC2SingleNodeCluster	i-e306c8e9	c3.large	us-west-2a	stopped	None		

Instance: i-e306c8e9 (HadoopEC2SingleNodeCluster) Private IP: 172.31.47.146

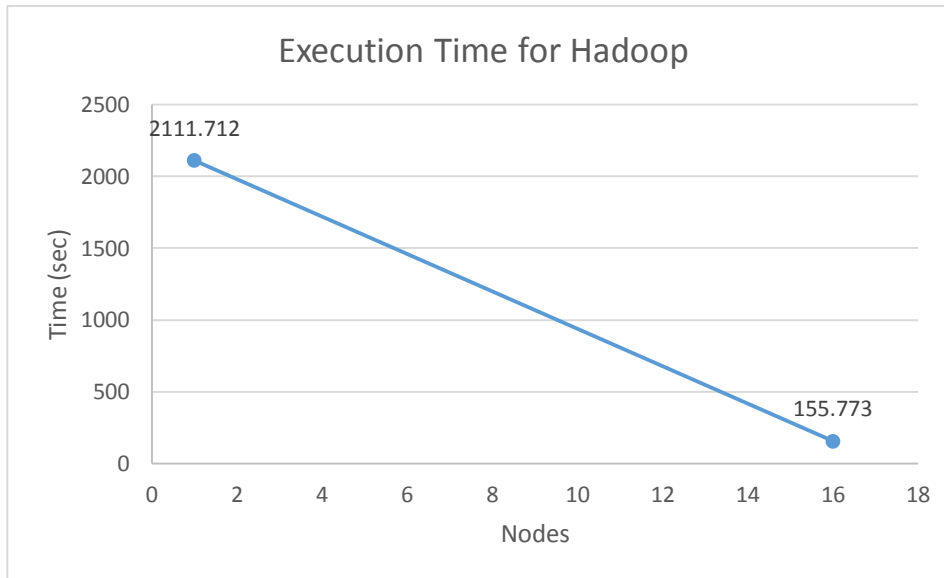
Performance Evaluation:

The Execution Time for running the Word Count Program on a Single node = **2111712 milliseconds**

= 2111.712 seconds

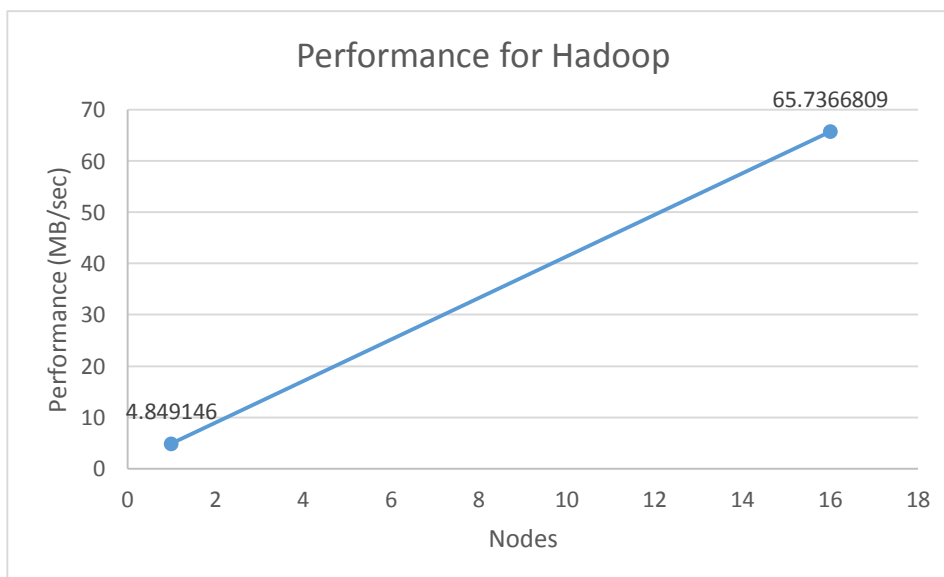
The Execution Time for running the Word Count Program 16 nodes = **155773 milliseconds**

= 155.773 seconds

Graph of Execution Time for Hadoop vs Number of Nodes:

Performance in a Single Node = Total Size of File / Execution Time = **4.849146MB/sec**

Performance in 16 Nodes = Total Size of File / Execution Time = **65.7366809MB/sec**

Graph of Performance for Hadoop vs Number of Nodes:

Output Table:

Number of Nodes	File Size	Execution Time	Performance
1	10240MB	2111.712sec	4.849146MB/sec
16	10240MB	155.773sec	65.7366809MB/sec

Observation:

We observe that we get a better performance on a 16 node system than a single node system. Since there are 16 slaves working on the same problem, the work is equally divided between all the slaves and at a time each slave is doing the computation. This reduces the overall execution time when 16 Nodes are used. In a single node, that node is doing the whole computation by itself, so it takes a lot of time to execute.

What is a Master node? What is a Slaves node?

Master Node: The Master Node is the one which stores all the data in the HDFS (Hadoop Distributed File System) and runs Map Reduce on all the data. Here the NameNode is used to store the data and the JobTracker does the parallel computation on that data i.e. Map Reduce.

Slave Node: The Slave Nodes are the ones which do all the computation work. The Slaves consists of both Data Node and Task Tracker. They receive the instructions from the master and do the required computations.

How can we change the number of mappers and reducers from the configuration file?

The number of reducers is controlled by `mapred.reduce.tasks` and the number of maps is controlled by `mapred.map.tasks` in the `mapred` conf file in the `mapred-site.xml` file.

Why do we need to set unique available ports to those configuration files on a shared environment?

The port number is used by Hadoop to find the path on the HDFS whose NameNode is running at `<HDFSip>:<port>`

Conclusion: We haven't provided the output file for the word count in Hadoop as it was 80MB each and according to the TA's mail, we have to keep the assignment folder small. So we have kept a Screenshot of the output in the Documentation. We have the copy of the output files in our system if you need it in the future.