

PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

A20360111

DESCRIPTION:

The purpose of this document is to describe steps to install, configure and capture the results before and after running Sorting implementation on files greater than memory size in Hadoop, Spark, Java. Also, by comparing the performance of Shared Memory Sort Java, Hadoop and Spark and determine which one is better.

Operating System: Linux Operating System.

Java version: 1.7.0

Hadoop version: 2.7.2

Spark version 1.6.1

Instance type c3.large

Linux distribution kernel: 3.19.0-25-generic

HADOOP:

Hadoop single node is implemented using Java language. It consists of mapper and reducer phase. In the mapper phase the text input format is used and the key value pair for one record is extracted and push it to output collector. In the reducer phase, iterate through values with respect to the key and push to output collector in sorted order.

Configuration File Description:

1. hadoop-env.sh- This file requires environment variables settings such as JAVA_HOME which affects the Hadoop daemon behavior while storing files and increasing heap size.
2. Slaves- This file consists the public DNS of Master and Slave where the Hadoop slave daemons will run. This configuration required for multimode setup. In case of single node it consists of localhost.
3. Core-site.xml- This file consists of I/O configuration setting for Hadoop Core common to HDFS and MapReduce.
4. Hdfs-site.xml- This file contains the configuration setting for HDFS daemon, the namenode, the secondary namenode and the datanode nodes.
5. Mapred-site.xml- This file contains the configuration setting for Hadoop daemon job tracker and task tracker
6. Yarn-site.xml- This file contains the configuration setting for YARN daemon the ResourceManager, NodeManager, and WebAppProxy.

The modifications that needed when configuring files on multi node are as follows:

For each Slave Node

1. Update the Hosts file with the private IP and public DNS of master and slave.
2. Update the Slaves file with the public DNS of master and slave.

For the Master Node

PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

A20360111

1. Update the Hosts file with the private IP and public DNS of master and all the slave nodes.
2. Update the Slaves file with the public DNS of master and all the slave nodes.

Questions:

1. What is a Master node? What is a Slaves node?

Answer: Master node stores lots of data (HDFS) and runs parallel computation on all that data. Master node assigns work to Slaves. Slave node is the place where data is stored and where data processing takes place.

2. Why do we need to set unique available ports to those configuration files on a shared environment? What errors or side-effects will show if we use same port number for each user?

In the configuration files, MapReduce job tracker is set to 9001, ResourceManager's resource tracker is set to 9025 and ResourceManager's scheduler is set to 9030. All are using different ports to avoid port collision. The bandwidth of the port will be shared and will raise security concerns.

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

The mappers and reducers can be changed by updating the

mapred.tasktracker.map.tasks.maximum and

mapred.tasktracker.reduce.tasks.maximum in the mapred-site.xml.

1. Installation of Hadoop 1-node:

Step1: sudo apt-get update

sudo apt-get install default-jdk

The screenshot shows the AWS EC2 Management Console interface. On the left, there is a sidebar with various navigation links: EC2 Dashboard, Events, Tags, Reports, Limits, Instances (which is selected), Spot Requests, Reserved Instances, Scheduled Instances, Commands, Dedicated Hosts, IMAGES, AMIs, Bundle Tasks, ELASTIC BLOCK STORE, Volumes, Snapshots, and NETWORK & SECURITY, Security Groups. The main content area displays a table of instances. The table has columns: Name, Instance ID, Instance Type, Availability Zone, Instance State, Status Checks, Alarm Status, and Public DNS. There are three rows:

Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Status	Public DNS
Spark	i-0e8de98d	t2.micro	us-east-1c	running	2/2 checks ...	None	ec2-54-209-83-238.com
	i-3def267	c3.large	us-east-1d	running	2/2 checks ...	None	ec2-52-23-225-126.com
	i-e1d3160	c3.large	us-east-1d	terminated		None	

Below the table, for the selected instance i-3def267, there is a detailed view with tabs: Description, Status Checks, Monitoring, and Tags. The Description tab shows the following details:

Attribute	Value
Instance ID	i-3def267
Instance state	running
Instance type	c3.large
Private DNS	ip-172-31-5-30.ec2.internal
Private IPs	172.31.5.30
Secondary private IPs	
VPC ID	vpc-00b3f564
Subnet ID	subnet-9f43f6e9

On the right side of the instance details, there are several status indicators and links:

- Public DNS: ec2-52-23-225-126.compute-1.amazonaws.com
- Public IP: 52.23.225.126
- Elastic IP: -
- Availability zone: us-east-1d
- Security groups: launch-wizard-50, view rules
- Scheduled events: No scheduled events
- AMI ID: ubuntu-trusty-14.04-amd64-server-20160114.5 (ami-fce3c696)
- Platform: -

At the bottom of the page, there are links for Feedback, English, Mozilla Firefox seems slow... to start, Learn How to Speed It Up, Don't Tell Me Again, Privacy Policy, Terms of Use, and a copyright notice: © 2008 - 2016, Amazon Web Services, Inc. or its affiliates. All rights reserved.

PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

A20360111

```
ubuntu@ip-172-31-5-30: ~
Get:12 http://us-east-1.ec2.archive.ubuntu.com trusty-updates/multiverse Translation-en [7,227 B]
Get:13 http://us-east-1.ec2.archive.ubuntu.com trusty-updates/restricted Translation-en [3,699 B]
Get:14 http://us-east-1.ec2.archive.ubuntu.com trusty-updates/universe Translation-en [186 kB]
Get:15 http://us-east-1.ec2.archive.ubuntu.com trusty-backports/main Sources [8,696 B]
Get:16 http://security.ubuntu.com trusty-security InRelease [65.9 kB]
Get:17 http://us-east-1.ec2.archive.ubuntu.com trusty-backports/restricted Sources [28 B]
Get:18 http://us-east-1.ec2.archive.ubuntu.com trusty-backports/universe Sources [34.5 kB]
Get:19 http://us-east-1.ec2.archive.ubuntu.com trusty-backports/multiverse Sources [1,898 B]
Get:20 http://us-east-1.ec2.archive.ubuntu.com trusty-backports/main amd64 Packages [9,782 B]
Get:21 http://us-east-1.ec2.archive.ubuntu.com trusty-backports/restricted amd64 Packages [28 B]
Get:22 http://us-east-1.ec2.archive.ubuntu.com trusty-backports/universe amd64 Packages [41.3 kB]
Get:23 http://us-east-1.ec2.archive.ubuntu.com trusty-backports/multiverse amd64 Packages [1,571 B]
Get:24 http://us-east-1.ec2.archive.ubuntu.com trusty-backports/main Translation-en [5,843 B]
Get:25 http://us-east-1.ec2.archive.ubuntu.com trusty-backports/multiverse Translation-en [1,215 B]
Get:26 http://us-east-1.ec2.archive.ubuntu.com trusty-backports/restricted Translation-en [28 B]
Get:27 http://us-east-1.ec2.archive.ubuntu.com trusty-backports/universe Translation-en [35.9 kB]
Get:28 http://us-east-1.ec2.archive.ubuntu.com trusty/main Sources [1,064 kB]
Get:29 http://us-east-1.ec2.archive.ubuntu.com trusty/restricted Sources [5,433 B]
Get:30 http://us-east-1.ec2.archive.ubuntu.com trusty/universe Sources [6,399 kB]
Get:31 http://security.ubuntu.com trusty-security/main Sources [10 kB]
Get:32 http://us-east-1.ec2.archive.ubuntu.com trusty/multiverse Sources [174 kB]
Hit http://us-east-1.ec2.archive.ubuntu.com trusty/main amd64 Packages
Hit http://us-east-1.ec2.archive.ubuntu.com trusty/restricted amd64 Packages
Hit http://us-east-1.ec2.archive.ubuntu.com trusty/universe amd64 Packages
Hit http://us-east-1.ec2.archive.ubuntu.com trusty/multiverse amd64 Packages
Hit http://us-east-1.ec2.archive.ubuntu.com trusty/main Translation-en
Hit http://us-east-1.ec2.archive.ubuntu.com trusty/multiverse Translation-en
Hit http://us-east-1.ec2.archive.ubuntu.com trusty/restricted Translation-en
Hit http://us-east-1.ec2.archive.ubuntu.com trusty/universe Translation-en
Get:33 http://security.ubuntu.com trusty-security/universe Sources [34.1 kB]
Ign http://us-east-1.ec2.archive.ubuntu.com trusty/main Translation-en_US
Ign http://us-east-1.ec2.archive.ubuntu.com trusty/multiverse Translation-en_US
Ign http://us-east-1.ec2.archive.ubuntu.com trusty/restricted Translation-en_US
Ign http://us-east-1.ec2.archive.ubuntu.com trusty/universe Translation-en_US
Get:34 http://security.ubuntu.com trusty-security/main amd64 Packages [448 kB]
Get:35 http://security.ubuntu.com trusty-security/universe amd64 Packages [125 kB]
Get:36 http://security.ubuntu.com trusty-security/main Translation-en [244 kB]
Get:37 http://security.ubuntu.com trusty-security/universe Translation-en [74.0 kB]
Fetched 11.1 MB in 3s (3,411 kB/s)
Reading package lists... Done
ubuntu@ip-172-31-5-30:~$
```

```
ubuntu@ip-172-31-5-30: ~
Adding debian:Camerfirma_Chambers_of_Commerce_Root.pem
Adding debian:AddTrust_Low_Value_Services_Root.pem
Adding debian:Thawte_Server_CA.pem
Adding debian:S-TRUST_Authentication_and_Encryption_Root_CA_2005_PN.pem
Adding debian:ApplicationCA_-_Japanese_Government.pem
Adding debian:Deutsche_Telekom_Root_CA_2.pem
Adding debian:AffirmTrust_Commercial.pem
Adding debian:VeriSign_Universal_Root_Certification_Authority.pem
Adding debian:Starfield_Services_Root_Certificate_Authority_-_G2.pem
Adding debian:D-TRUST_Root_Class_3_CA_2_2009.pem
Adding debian:GlobalSign_Root_CA_-_R3.pem
Adding debian:QuoVadis_Root_CA_3_G3.pem
Adding debian:TUBITAK_UKEKAE_Kok_Sertifika_Hizmet_Saglayicisi_-_50r0m_3.pem
Adding debian:Entrust_Root_Certification_Authority.pem
Adding debian:Autoridad_de_Certificacion_Firmaprofesional_CIF_A62634068.pem
Adding debian:UTN_DATAcorp_SGC_Root_CA.pem
Adding debian:D-TRUST_Root_Class_3_CA_2_EV_2009.pem
Adding debian:CA_Disig_Root_R1.pem
Adding debian:NetLock_Notary_-_Class_A_=Root.pem
Adding debian:Tatwan_GRCAs.pem
Adding debian:SG_TRUST_SERVICES_RACINE.pem
Adding debian:AffirmTrust_Premium_ECC.pem
Adding debian:Verisign_Class_3_Public_Primary_Certification_Authority_-_G2.pem
Adding debian:Chambers_of_Commerce_Root_-_2008.pem
Adding debian:A-Trust-nQual-03.pem
Adding debian:E-Guvet_Kok_Elektronik_Sertifika_Hizmet_Saglayicisi.pem
Adding debian:America_Online_Root_Certification_Authority_2.pem
Adding debian:NetLock_Qualified_=Class_QA_=Root.pem
Adding debian:Baltimore_CyberTrust_Root.pem
Adding debian:GlobalSign_Root_CA_-_R2.pem
Adding debian:CNNIC_ROOT.pem
done.
Setting up libatk-wrapper-jav (0.30.4-4) ...
Setting up libatk-wrapper-jav-jni:amd64 (0.30.4-4) ...
Processing triggers for libc-bin (2.19-0ubuntu6.6) ...
Processing triggers for ca-certificates (20140109ubuntu0.14.04.1) ...
Updating certificates in /etc/ssl/certs... 0 added, 0 removed; done.
Running hooks in /etc/ca-certificates/update.d.....
done.
done.
done.
ubuntu@ip-172-31-5-30:~$ sudo apt-get install default-jdk
```

PROGRAMMING ASSIGNMENT 2

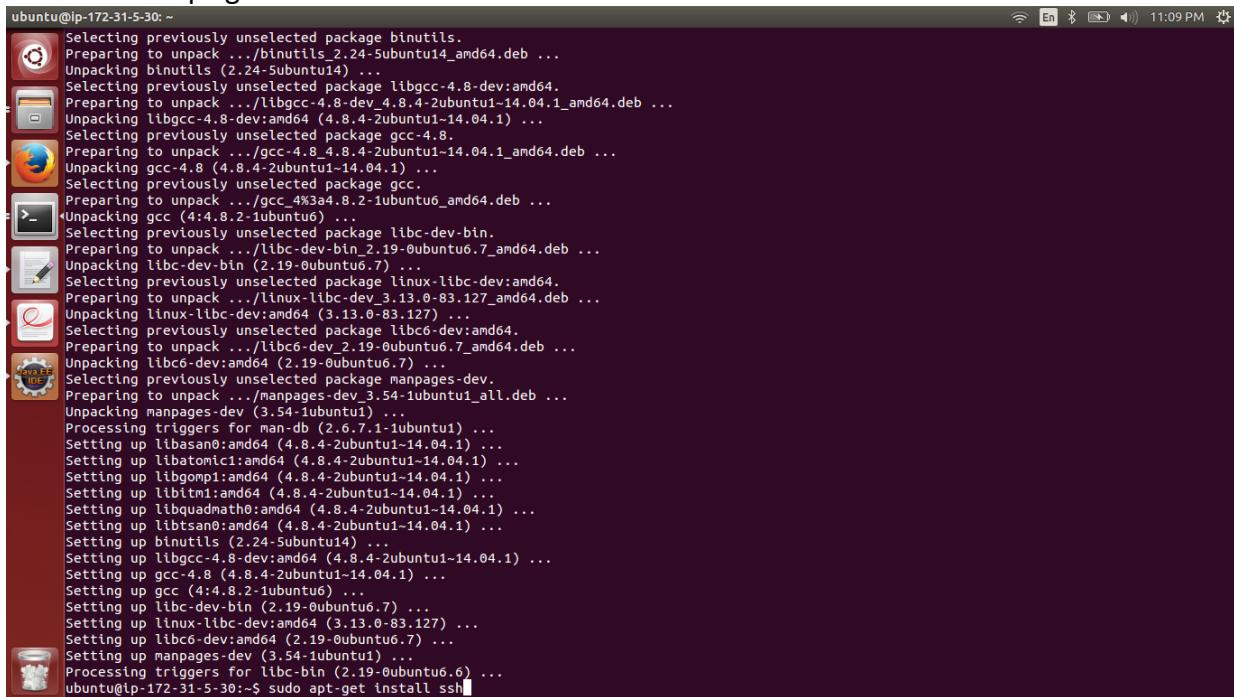
Sort on HADOOP/SPARK

Snehal Sonawane

A20360111

Step2: Install ssh

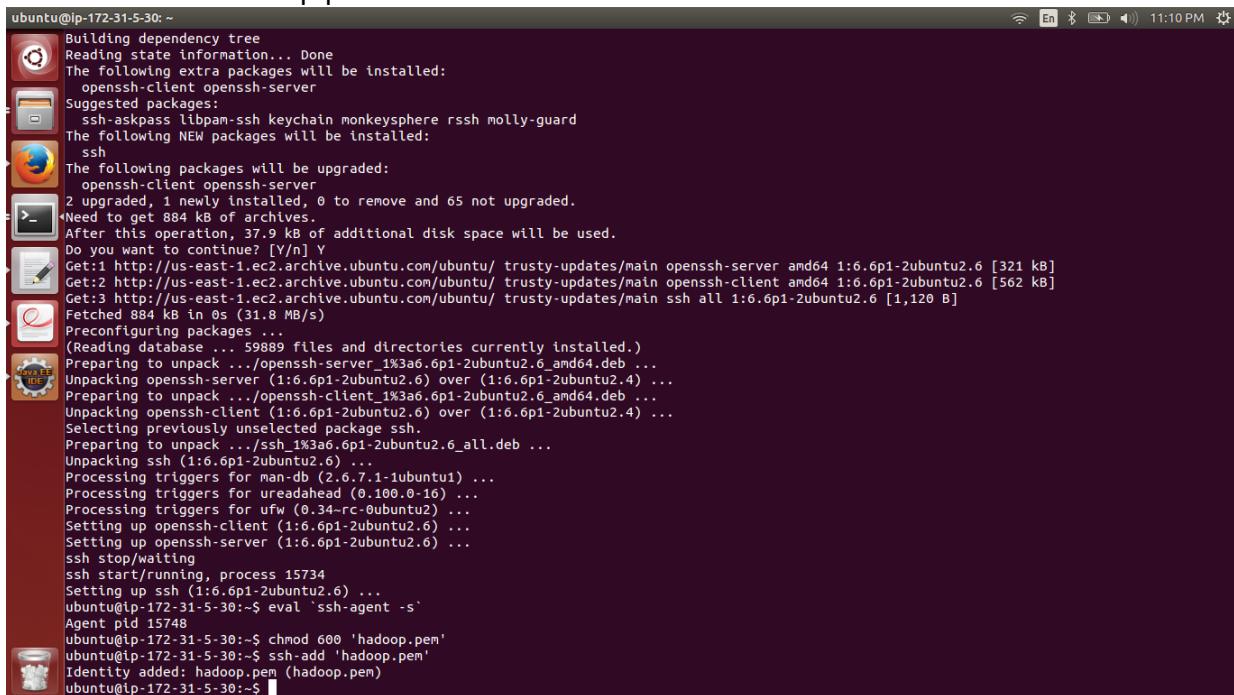
```
sudo apt-get install ssh
```



```
ubuntu@ip-172-31-5-30: ~
Selecting previously unselected package binutils.
Preparing to unpack .../binutils_2.24-5ubuntu14_amd64.deb ...
Unpacking binutils (2.24-5ubuntu14) ...
Selecting previously unselected package libgcc-4.8-dev:amd64.
Preparing to unpack .../libgcc-4.8-dev_4.8.4-2ubuntu1-14.04.1_amd64.deb ...
Unpacking libgcc-4.8-dev:amd64 (4.8.4-2ubuntu1-14.04.1) ...
Selecting previously unselected package gcc-4.8.
Preparing to unpack .../gcc-4.8_4.8.4-2ubuntu1-14.04.1_amd64.deb ...
Unpacking gcc-4.8 (4.8.4-2ubuntu1-14.04.1) ...
Selecting previously unselected package gcc.
Preparing to unpack .../gcc_4%3a4.8.2-1ubuntu6_amd64.deb ...
Unpacking gcc (4:4.8.2-1ubuntu6) ...
Selecting previously unselected package libc-dev-bin.
Preparing to unpack .../libc-dev-bin_2.19-0ubuntu6.7_amd64.deb ...
Unpacking libc-dev-bin (2.19-0ubuntu6.7) ...
Selecting previously unselected package linux-libc-dev:amd64.
Preparing to unpack .../linux-libc-dev_3.13.0-83.127_amd64.deb ...
Unpacking linux-libc-dev:amd64 (3.13.0-83.127) ...
Selecting previously unselected package libc6-dev:amd64.
Preparing to unpack .../libc6-dev_2.19-0ubuntu6.7_amd64.deb ...
Unpacking libc6-dev:amd64 (2.19-0ubuntu6.7) ...
Selecting previously unselected package manpages-dev.
Preparing to unpack .../manpages-dev_3.54-1ubuntu1_all.deb ...
Unpacking manpages-dev (3.54-1ubuntu1) ...
Processing triggers for man-db (2.6.7.1-1ubuntu1) ...
Setting up libasan0:amd64 (4.8.4-2ubuntu1-14.04.1) ...
Setting up libatomic1:amd64 (4.8.4-2ubuntu1-14.04.1) ...
Setting up libomp1:amd64 (4.8.4-2ubuntu1-14.04.1) ...
Setting up libitm1:amd64 (4.8.4-2ubuntu1-14.04.1) ...
Setting up libquadmath0:amd64 (4.8.4-2ubuntu1-14.04.1) ...
Setting up libtsan0:amd64 (4.8.4-2ubuntu1-14.04.1) ...
Setting up binutils (2.24-5ubuntu14) ...
Setting up libgcc-4.8-dev:amd64 (4.8.4-2ubuntu1-14.04.1) ...
Setting up gcc-4.8 (4.8.4-2ubuntu1-14.04.1) ...
Setting up gcc (4:4.8.2-1ubuntu6) ...
Setting up libc-dev-bin (2.19-0ubuntu6.7) ...
Setting up linux-libc-dev:amd64 (3.13.0-83.127) ...
Setting up libc6-dev:amd64 (2.19-0ubuntu6.7) ...
Setting up manpages-dev (3.54-1ubuntu1) ...
Processing triggers for libc-bin (2.19-0ubuntu6.6) ...
ubuntu@ip-172-31-5-30:~$ sudo apt-get install ssh
```

Step 3: add pem file to ssh

```
eval `ssh-agent -s`
chmod 600 'hadoop.pem'
ssh-add 'hadoop.pem'
```



```
ubuntu@ip-172-31-5-30: ~
Building dependency tree
Reading state information... Done
The following extra packages will be installed:
  openssh-client openssh-server
Suggested packages:
  ssh-askpass libpam-ssh keychain monkeysphere rssh molly-guard
The following NEW packages will be installed:
  ssh
The following packages will be upgraded:
  openssh-client openssh-server
2 upgraded, 1 newly installed, 0 to remove and 65 not upgraded.
Need to get 884 kB of archives.
After this operation, 37.9 kB of additional disk space will be used.
Do you want to continue? [Y/n] Y
Get:1 http://us-east-1.ec2.archive.ubuntu.com/ubuntu/ trusty-updates/main openssh-server amd64 1:6.6p1-2ubuntu2.6 [321 kB]
Get:2 http://us-east-1.ec2.archive.ubuntu.com/ubuntu/ trusty-updates/main openssh-client amd64 1:6.6p1-2ubuntu2.6 [562 kB]
Get:3 http://us-east-1.ec2.archive.ubuntu.com/ubuntu/ trusty-updates/main ssh all 1:6.6p1-2ubuntu2.6 [1,120 kB]
Fetched 884 kB in 0s (31.8 MB/s)
Preconfiguring packages ...
(Reading database ... 59889 files and directories currently installed.)
Preparing to unpack .../openssh-server_1%3a6.6p1-2ubuntu2.6_amd64.deb ...
Unpacking openssh-server (1:6.6p1-2ubuntu2.6) over (1:6.6p1-2ubuntu2.4) ...
Preparing to unpack .../openssh-client_1%3a6.6p1-2ubuntu2.6_amd64.deb ...
Unpacking openssh-client (1:6.6p1-2ubuntu2.6) over (1:6.6p1-2ubuntu2.4) ...
Selecting previously unselected package ssh.
Preparing to unpack .../ssh_1%3a6.6p1-2ubuntu2.6_all.deb ...
Unpacking ssh (1:6.6p1-2ubuntu2.6) ...
Processing triggers for man-db (2.6.7.1-1ubuntu1) ...
Processing triggers for ureadahead (0.100.0-16) ...
Processing triggers for ufw (0.34-rc-0ubuntu2) ...
Setting up openssh-client (1:6.6p1-2ubuntu2.6) ...
Setting up openssh-server (1:6.6p1-2ubuntu2.6) ...
ssh stop/waiting
ssh start/running, process 15734
Setting up ssh (1:6.6p1-2ubuntu2.6) ...
ubuntu@ip-172-31-5-30:~$ eval `ssh-agent -s`
Agent pid 15748
ubuntu@ip-172-31-5-30:~$ chmod 600 'hadoop.pem'
ubuntu@ip-172-31-5-30:~$ ssh-add 'hadoop.pem'
Identity added: hadoop.pem (hadoop.pem)
ubuntu@ip-172-31-5-30:~$
```

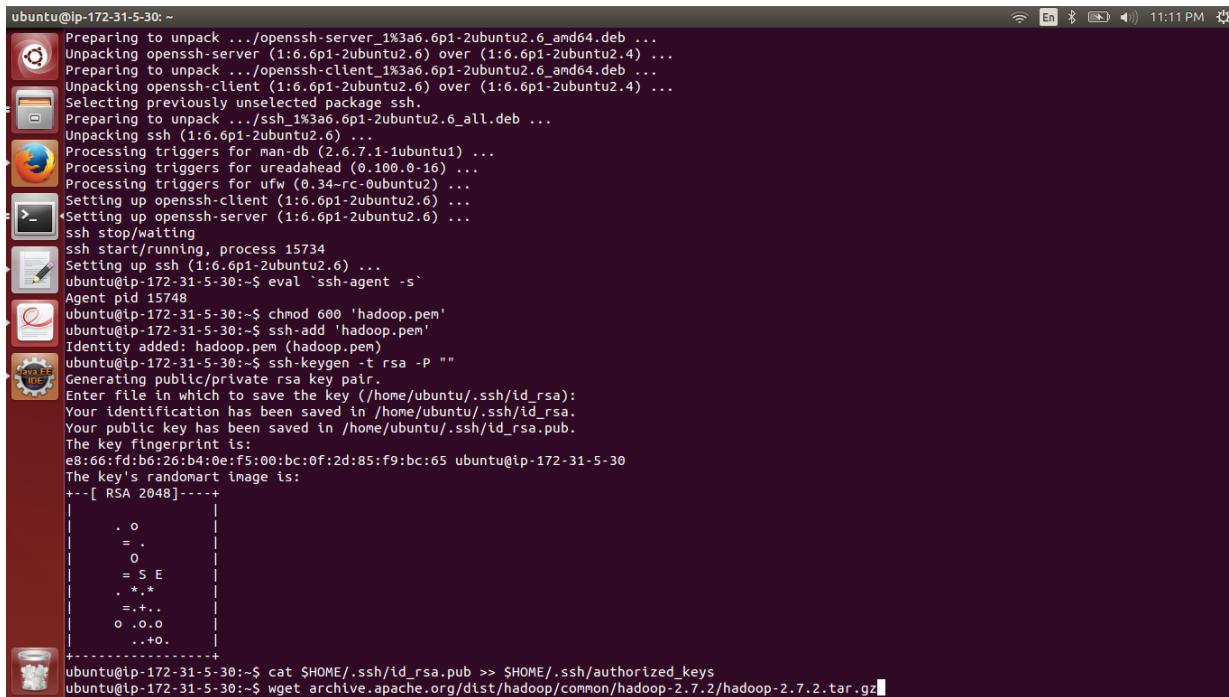
PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

A20360111

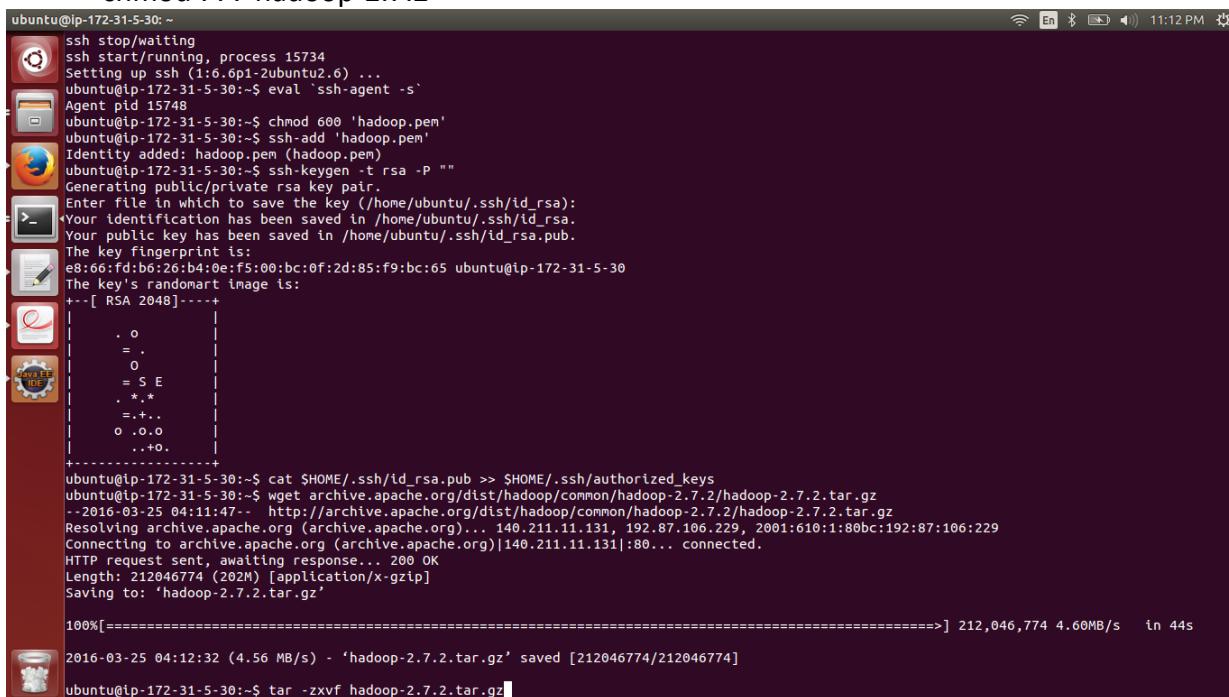
Step 4: Set up SSH certificates to avoid entering password each time:



```
ubuntu@ip-172-31-5-30: ~
Preparing to unpack .../openssh-server_1%3a6.6p1-2ubuntu2.6_and64.deb ...
Unpacking openssh-server (1:6.6p1-2ubuntu2.6) over (1:6.6p1-2ubuntu2.4) ...
Preparing to unpack .../openssh-client_1%3a6.6p1-2ubuntu2.6_and64.deb ...
Unpacking openssh-client (1:6.6p1-2ubuntu2.6) over (1:6.6p1-2ubuntu2.4) ...
Selecting previously unselected package ssh.
Preparing to unpack .../ssh_1%3a6.6p1-2ubuntu2.6_all.deb ...
Unpacking ssh (1:6.6p1-2ubuntu2.6) ...
Processing triggers for man-db (2.6.7.1-1ubuntu1) ...
Processing triggers for ureadahead (0.100.0-16) ...
Setting up openssh-client (1:6.6p1-2ubuntu2.6) ...
Setting up openssh-server (1:6.6p1-2ubuntu2.6) ...
ssh stop/waiting
ssh start/running, process 15734
Setting up ssh (1:6.6p1-2ubuntu2.6) ...
ubuntu@ip-172-31-5-30:~ eval `ssh-agent -s`
Agent pid 15748
ubuntu@ip-172-31-5-30:~ chmod 600 'hadoop.pem'
ubuntu@ip-172-31-5-30:~ ssh-add 'hadoop.pem'
Identity added: hadoop.pem (hadoop.pem)
ubuntu@ip-172-31-5-30:~ ssh-keygen -t rsa -P ""
Generating public/private rsa key pair.
Enter file in which to save the key (/home/ubuntu/.ssh/id_rsa):
Your identification has been saved in /home/ubuntu/.ssh/id_rsa.
Your public key has been saved in /home/ubuntu/.ssh/id_rsa.pub.
The key fingerprint is:
e8:66:fd:b6:26:b4:0e:f5:00:bc:0f:2d:85:f9:bc:65 ubuntu@ip-172-31-5-30
The key's randomart image is:
+--[ RSA 2048]----+
| . o |
| = . |
| O |
| = S E |
| . *.* |
| =+.+ |
| o .o.o |
| ..+o. |
+-----+
ubuntu@ip-172-31-5-30:~ cat $HOME/.ssh/id_rsa.pub >> $HOME/.ssh/authorized_keys
ubuntu@ip-172-31-5-30:~ wget archive.apache.org/dist/hadoop/common/hadoop-2.7.2/hadoop-2.7.2.tar.gz
```

Step 5: Download and install Hadoop

```
wget archive.apache.org/dist/hadoop/common/hadoop-2.7.2/hadoop-2.7.2.tar.gz
tar -zvxf hadoop-2.7.2.tar.gz
chmod 777 hadoop-2.7.2
```



```
ubuntu@ip-172-31-5-30: ~
ssh stop/waiting
ssh start/running, process 15734
Setting up ssh (1:6.6p1-2ubuntu2.6) ...
ubuntu@ip-172-31-5-30:~ eval `ssh-agent -s`
Agent pid 15748
ubuntu@ip-172-31-5-30:~ chmod 600 'hadoop.pem'
ubuntu@ip-172-31-5-30:~ ssh-add 'hadoop.pem'
Identity added: hadoop.pem (hadoop.pem)
ubuntu@ip-172-31-5-30:~ ssh-keygen -t rsa -P ""
Generating public/private rsa key pair.
Enter file in which to save the key (/home/ubuntu/.ssh/id_rsa):
Your identification has been saved in /home/ubuntu/.ssh/id_rsa.
Your public key has been saved in /home/ubuntu/.ssh/id_rsa.pub.
The key fingerprint is:
e8:66:fd:b6:26:b4:0e:f5:00:bc:0f:2d:85:f9:bc:65 ubuntu@ip-172-31-5-30
The key's randomart image is:
+--[ RSA 2048]----+
| . o |
| = . |
| O |
| = S E |
| . *.* |
| =+.+ |
| o .o.o |
| ..+o. |
+-----+
ubuntu@ip-172-31-5-30:~ cat $HOME/.ssh/id_rsa.pub >> $HOME/.ssh/authorized_keys
ubuntu@ip-172-31-5-30:~ wget archive.apache.org/dist/hadoop/common/hadoop-2.7.2/hadoop-2.7.2.tar.gz
--2016-03-25 04:11:47- http://archive.apache.org/dist/hadoop/common/hadoop-2.7.2/hadoop-2.7.2.tar.gz
Resolving archive.apache.org (archive.apache.org)... 140.211.11.131, 192.87.106.229, 2001:610:1:80bc:192:87:106:229
Connecting to archive.apache.org (archive.apache.org)|140.211.11.131|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 212046774 (202M) [application/x-gzip]
Saving to: 'hadoop-2.7.2.tar.gz'

100%[=====] 212,046,774 4.60MB/s  in 44s
2016-03-25 04:12:32 (4.56 MB/s) - 'hadoop-2.7.2.tar.gz' saved [212046774/212046774]
ubuntu@ip-172-31-5-30:~ tar -zvxf hadoop-2.7.2.tar.gz
```

PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

A20360111

Step 6: Setup Configuration Files:

6.1. Edit environmental variables in bashrc file :

sudo

```
export JAVA_HOME=/usr/lib/jvm/java-7-openjdk-amd64
export HADOOP_INSTALL=/home/ubuntu/hadoop-2.7.2
export PATH=$PATH:$HADOOP_INSTALL/bin
export PATH=$PATH:$HADOOP_INSTALL/sbin
export HADOOP_MAPRED_HOME=$HADOOP_INSTALL
export HADOOP_COMMON_HOME=$HADOOP_INSTALL
export HADOOP_HDFS_HOME=$HADOOP_INSTALL
export YARN_HOME=$HADOOP_INSTALL
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_INSTALL/lib/native
export HADOOP_OPTS='-Djava.library.path=$HADOOP_INSTALL/lib'
```

6.2. Edit core-site.xml

```
ubuntu@ip-172-31-5-30: ~/hadoop-2.7.2/etc/hadoop
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
 Licensed under the Apache License, Version 2.0 (the "License");
 you may not use this file except in compliance with the License.
 You may obtain a copy of the License at

 http://www.apache.org/licenses/LICENSE-2.0

 Unless required by applicable law or agreed to in writing, software
 distributed under the License is distributed on an "AS IS" BASIS,
 WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
 See the License for the specific language governing permissions and
 limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->
<configuration>
<property>
  <name>fs.default.name</name>
  <value>hdfs://ec2-52-23-225-126.compute-1.amazonaws.com:9000</value>
</property>
<property>
  <name>hadoop.tmp.dir</name>
  <value>/data</value>
  <description>base location for other hdfs directories.</description>
</property>
</configuration>
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
:wq
```

PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

A20360111

6.3 Edit hadoop-env.sh

```
ubuntu@ip-172-31-5-30: ~/hadoop-2.7.2/etc/hadoop
# Licensed to the Apache Software Foundation (ASF) under one
# or more contributor license agreements. See the NOTICE file
# distributed with this work for additional information
# regarding copyright ownership. The ASF licenses this file
# to you under the Apache License, Version 2.0 (the
# "License"); you may not use this file except in compliance
# with the License. You may obtain a copy of the License at
#
#     http://www.apache.org/licenses/LICENSE-2.0
#
# Unless required by applicable law or agreed to in writing, software
# distributed under the License is distributed on an "AS IS" BASIS,
# WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
# See the License for the specific language governing permissions and
# limitations under the License.

# Set Hadoop-specific environment variables here.

# The only required environment variable is JAVA_HOME. All others are
# optional. When running a distributed configuration it is best to
# set JAVA_HOME in this file, so that it is correctly defined on
# remote nodes.

# The java implementation to use.
#export JAVA_HOME=${JAVA_HOME}
export JAVA_HOME=/usr/lib/jvm/java-7-openjdk-amd64
# The jsvc implementation to use. Jsvc is required to run secure datanodes
# that bind to privileged ports to provide authentication of data transfer
# protocol. Jsvc is not required if SASL is configured for authentication of
# data transfer protocol using non-privileged ports.
#export JSVC_HOME=${JSVC_HOME}

export HADOOP_CONF_DIR=$HADOOP_CONF_DIR:-"/etc/hadoop"

# Extra Java CLASSPATH elements. Automatically insert capacity-scheduler.
for f in $HADOOP_HOME/contrib/capacity-scheduler/*.jar; do
  if [ "$HADOOP_CLASSPATH" ]; then
    export HADOOP_CLASSPATH=$HADOOP_CLASSPATH:$f
  else
    export HADOOP_CLASSPATH=$f
  fi
done

26,50      Top
```

6.4 Edit hdfs-site.xml

```
ubuntu@ip-172-31-5-30: ~/hadoop-2.7.2/etc/hadoop
<?xml version= "1.0" encoding="UTF-8"?>
<xmlelement type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
<property>
  <name>dfs.replication</name>
  <value>1</value>
</property>
<property>
  <name>dfs.permissions</name>
  <value>false</value>
</property>
</configuration>
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~>
~
```

27,12-19 All

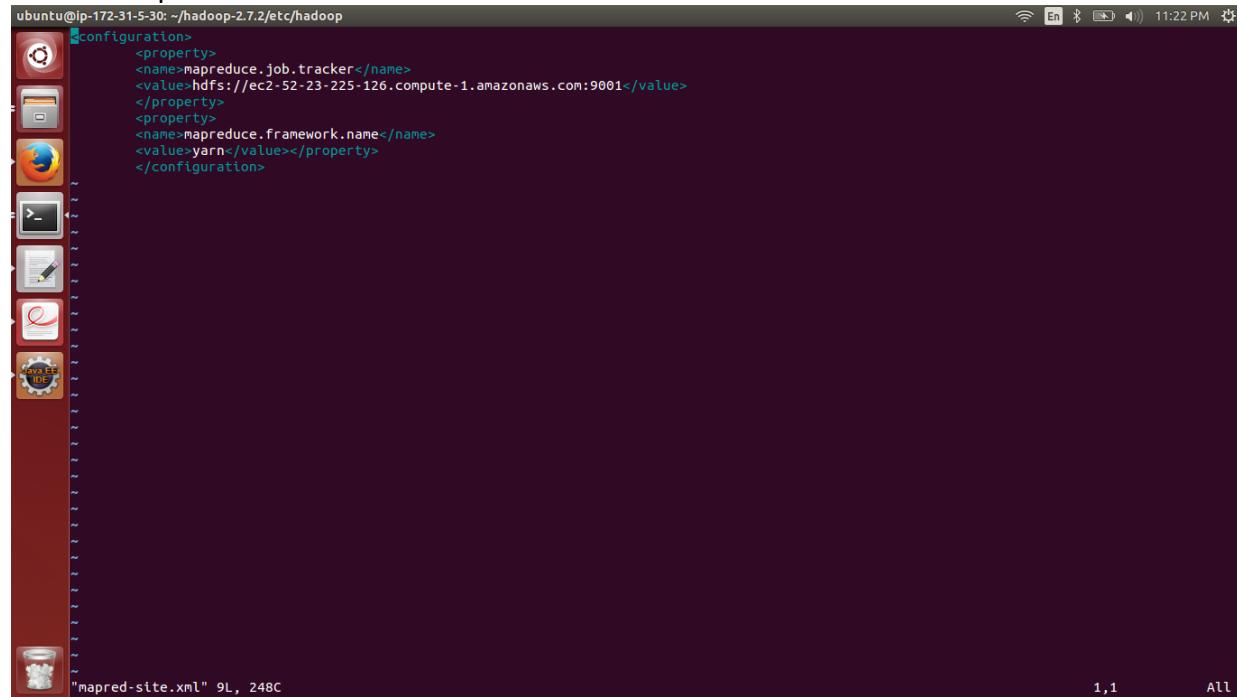
PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

A20360111

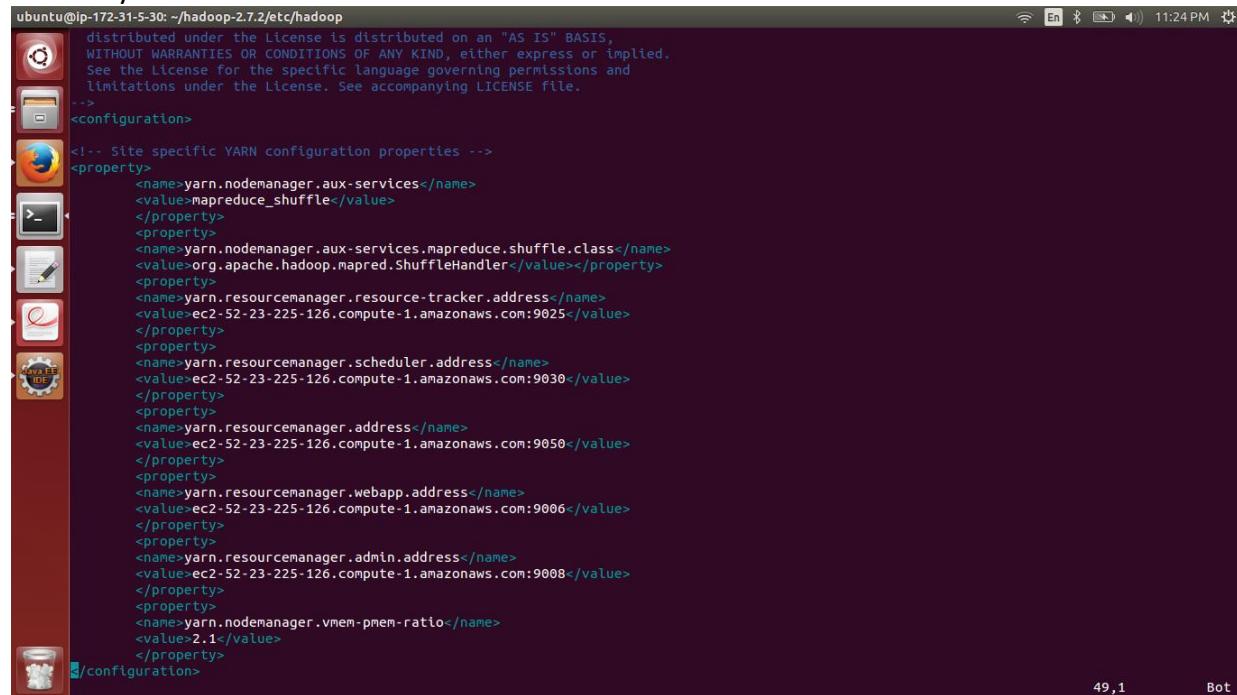
6.5 Edit mapred-site.xml



```
ubuntu@ip-172-31-5-30: ~/hadoop-2.7.2/etc/hadoop
<configuration>
  <property>
    <name>mapreduce.job.tracker</name>
    <value>hdfs://ec2-52-23-225-126.compute-1.amazonaws.com:9001</value>
  </property>
  <property>
    <name>mapreduce.framework.name</name>
    <value>yarn</value>
  </property>
</configuration>
~
```

"mapred-site.xml" 9L, 248C 1,1 All

6.6 Edit yarn-site.xml



```
ubuntu@ip-172-31-5-30: ~/hadoop-2.7.2/etc/hadoop
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.

-->
<configuration>
  <!-- Site specific YARN configuration properties -->
  <property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
  </property>
  <property>
    <name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
    <value>org.apache.hadoop.mapred.ShuffleHandler</value>
  </property>
  <property>
    <name>yarn.resourcemanager.resource-tracker.address</name>
    <value>ec2-52-23-225-126.compute-1.amazonaws.com:9025</value>
  </property>
  <property>
    <name>yarn.resourcemanager.scheduler.address</name>
    <value>ec2-52-23-225-126.compute-1.amazonaws.com:9030</value>
  </property>
  <property>
    <name>yarn.resourcemanager.address</name>
    <value>ec2-52-23-225-126.compute-1.amazonaws.com:9050</value>
  </property>
  <property>
    <name>yarn.resourcemanager.webapp.address</name>
    <value>ec2-52-23-225-126.compute-1.amazonaws.com:9006</value>
  </property>
  <property>
    <name>yarn.resourcemanager.admin.address</name>
    <value>ec2-52-23-225-126.compute-1.amazonaws.com:9008</value>
  </property>
  <property>
    <name>yarn.nodemanager.vmem-pmem-ratio</name>
    <value>2.1</value>
  </property>
</configuration>
```

49,1 Bot

PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

A20360111

6.7 Format the Namenode

./hdfs namenode -format

```
ubuntu@ip-172-31-5-30:~/hadoop-2.7.2/bin
16/03/25 04:26:33 INFO util.GSet: Computing capacity for map cachedBlocks
16/03/25 04:26:33 INFO util.GSet: VM type      = 64-bit
16/03/25 04:26:33 INFO util.GSet: 0.25% max memory 889 MB = 2.2 MB
16/03/25 04:26:33 INFO util.GSet: capacity     = 2^18 = 262144 entries
16/03/25 04:26:33 INFO namenode.FSNamesystem: dfs.namenode.safemode.threshold-pct = 0.9990000128746033
16/03/25 04:26:33 INFO namenode.FSNamesystem: dfs.namenode.min.datanodes = 0
16/03/25 04:26:33 INFO namenode.FSNamesystem: dfs.namenode.safemode.extension = 30000
16/03/25 04:26:33 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.window.num.buckets = 10
16/03/25 04:26:33 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.num.users = 10
16/03/25 04:26:33 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.windows.minutes = 1,5,25
16/03/25 04:26:33 INFO namenode.FSNamesystem: Retry cache on namenode is enabled
16/03/25 04:26:33 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total heap and retry cache entry expiry time is 600000 millis
16/03/25 04:26:33 INFO util.GSet: Computing capacity for map NameNodeRetryCache
16/03/25 04:26:33 INFO util.GSet: VM type      = 64-bit
16/03/25 04:26:33 INFO util.GSet: 0.029999999329447746% max memory 889 MB = 273.1 KB
16/03/25 04:26:33 INFO util.GSet: capacity     = 2^15 = 32768 entries
16/03/25 04:26:33 INFO namenode.FSImage: Allocated new BlockPoolId: BP-276328556-172.31.5.30-1458879993639
16/03/25 04:26:33 WARN namenode.NameNode: Encountered exception during format:
java.io.IOException: Cannot create directory /data/dfs/name/current
        at org.apache.hadoop.hdfs.server.common.Storage$StorageDirectory.clearDirectory(Storage.java:337)
        at org.apache.hadoop.hdfs.server.namenode.NNStorage.format(NNStorage.java:548)
        at org.apache.hadoop.hdfs.server.namenode.NNStorage.format(NNStorage.java:569)
        at org.apache.hadoop.hdfs.server.namenode.FSImage.format(FSImage.java:161)
        at org.apache.hadoop.hdfs.server.namenode.NameNode.format(NameNode.java:991)
        at org.apache.hadoop.hdfs.server.namenode.NameNode.createNameNode(NameNode.java:1429)
        at org.apache.hadoop.hdfs.server.namenode.NameNode.main(NameNode.java:1554)
16/03/25 04:26:33 ERROR namenode.NameNode: Failed to start namenode.
java.io.IOException: Cannot create directory /data/dfs/name/current
        at org.apache.hadoop.hdfs.server.common.Storage$StorageDirectory.clearDirectory(Storage.java:337)
        at org.apache.hadoop.hdfs.server.namenode.NNStorage.format(NNStorage.java:548)
        at org.apache.hadoop.hdfs.server.namenode.NNStorage.format(NNStorage.java:569)
        at org.apache.hadoop.hdfs.server.namenode.FSImage.format(FSImage.java:161)
        at org.apache.hadoop.hdfs.server.namenode.NameNode.format(NameNode.java:991)
        at org.apache.hadoop.hdfs.server.namenode.NameNode.createNameNode(NameNode.java:1429)
        at org.apache.hadoop.hdfs.server.namenode.NameNode.main(NameNode.java:1554)
16/03/25 04:26:33 INFO util.ExitUtil: Exiting with status 1
16/03/25 04:26:33 INFO namenode.NameNode: SHUTDOWN_MSG:
*****SHUTDOWN_MSG: Shutting down NameNode at ip-172-31-5-30.ec2.internal/172.31.5.30
*****
```

6.8 Start dfs and yarn and generate data

./start-dfs.sh

./start-yarn.sh

./gensort -a 100000000 /data/input1

```
ubuntu@ip-172-31-5-30:-
16/03/25 04:32:42 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.windows.minutes = 1,5,25
16/03/25 04:32:42 INFO namenode.FSNamesystem: Retry cache on namenode is enabled
16/03/25 04:32:42 INFO util.GSet: Computing capacity for map NameNodeRetryCache
16/03/25 04:32:42 INFO util.GSet: VM type      = 64-bit
16/03/25 04:32:42 INFO util.GSet: 0.029999999329447746% max memory 889 MB = 273.1 KB
16/03/25 04:32:42 INFO util.GSet: capacity     = 2^15 = 32768 entries
16/03/25 04:32:42 INFO namenode.FSImage: Allocated new BlockPoolId: BP-703763242-172.31.5.30-1458880362465
16/03/25 04:32:42 INFO common.Storage: Storage directory /data/dfs/name has been successfully formatted.
16/03/25 04:32:42 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
16/03/25 04:32:42 INFO util.ExitUtil: Exiting with status 0
16/03/25 04:32:42 INFO namenode.NameNode: SHUTDOWN_MSG:
*****SHUTDOWN_MSG: Shutting down NameNode at ip-172-31-5-30.ec2.internal/172.31.5.30
*****
```

ubuntu@ip-172-31-5-30:~/hadoop-2.7.2/sbin\$ cd /home/ubuntu/hadoop-2.7.2/sbin/
ubuntu@ip-172-31-5-30:~/hadoop-2.7.2/sbin\$./start-dfs.sh
Starting namenodes on [ec2-52-23-225-126.compute-1.amazonaws.com]
ec2-52-23-225-126.compute-1.amazonaws.com: starting namenode, logging to /home/ubuntu/hadoop-2.7.2/logs/hadoop-ubuntu-namenode-ip-172-31-5-30.out
localhost: starting datanode, logging to /home/ubuntu/hadoop-2.7.2/logs/hadoop-ubuntu-datanode-ip-172-31-5-30.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /home/ubuntu/hadoop-2.7.2/logs/hadoop-ubuntu-secondarynamenode-ip-172-31-5-30.out
ubuntu@ip-172-31-5-30:~/hadoop-2.7.2/sbin\$./start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /home/ubuntu/hadoop-2.7.2/logs/yarn-ubuntu-resourcemanager-ip-172-31-5-30.out
localhost: starting nodemanager, logging to /home/ubuntu/hadoop-2.7.2/logs/yarn-ubuntu-nodemanager-ip-172-31-5-30.out
ubuntu@ip-172-31-5-30:~/hadoop-2.7.2/sbin\$ jps
18979 ResourceManager
19417 Jps
18823 SecondaryNameNode
18615 DataNode
19126 NodeManager
18443 NameNode
ubuntu@ip-172-31-5-30:~/hadoop-2.7.2/sbin\$ cd /home/ubuntu/hadoop-2.7.2/
ubuntu@ip-172-31-5-30:~/hadoop-2.7.2\$ bin/hadoop fs -mkdir -p /user/ubuntu/gutenberg
ubuntu@ip-172-31-5-30:~/hadoop-2.7.2\$./gensort -a 100000000 /data/input1
-bash: ./gensort: No such file or directory
ubuntu@ip-172-31-5-30:~/hadoop-2.7.2\$ cd ..
ubuntu@ip-172-31-5-30:~/ ./.gensort -a 100000000 /data/input1

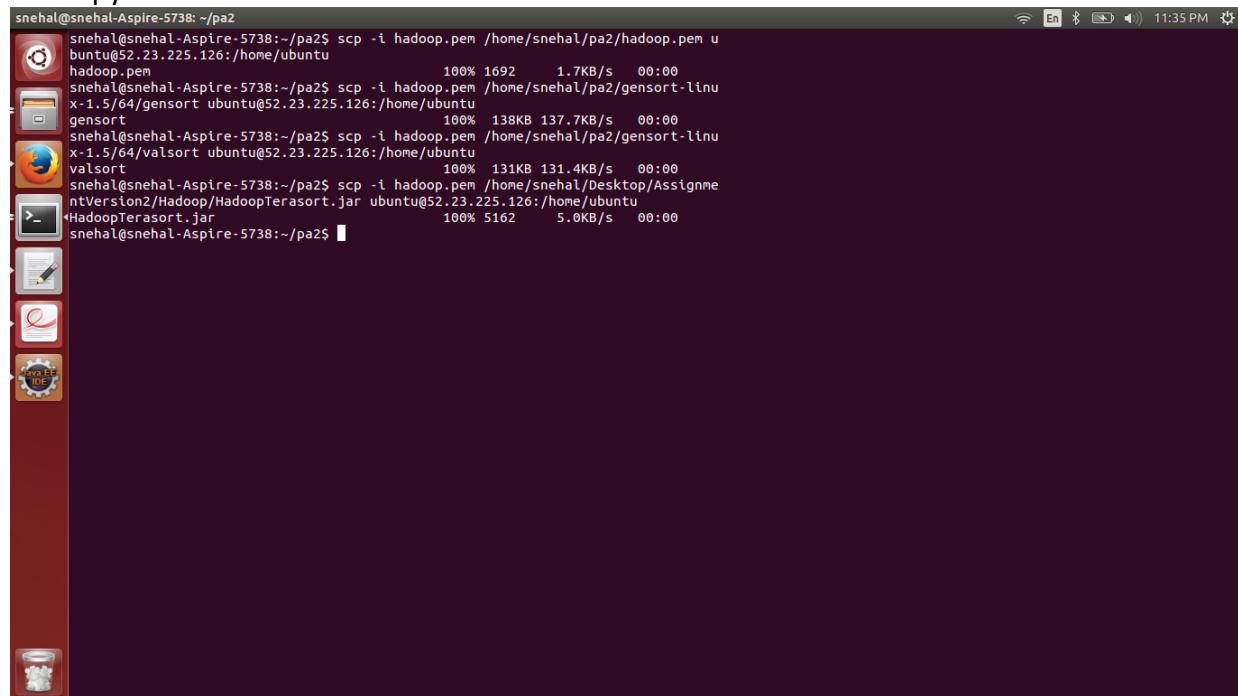
PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

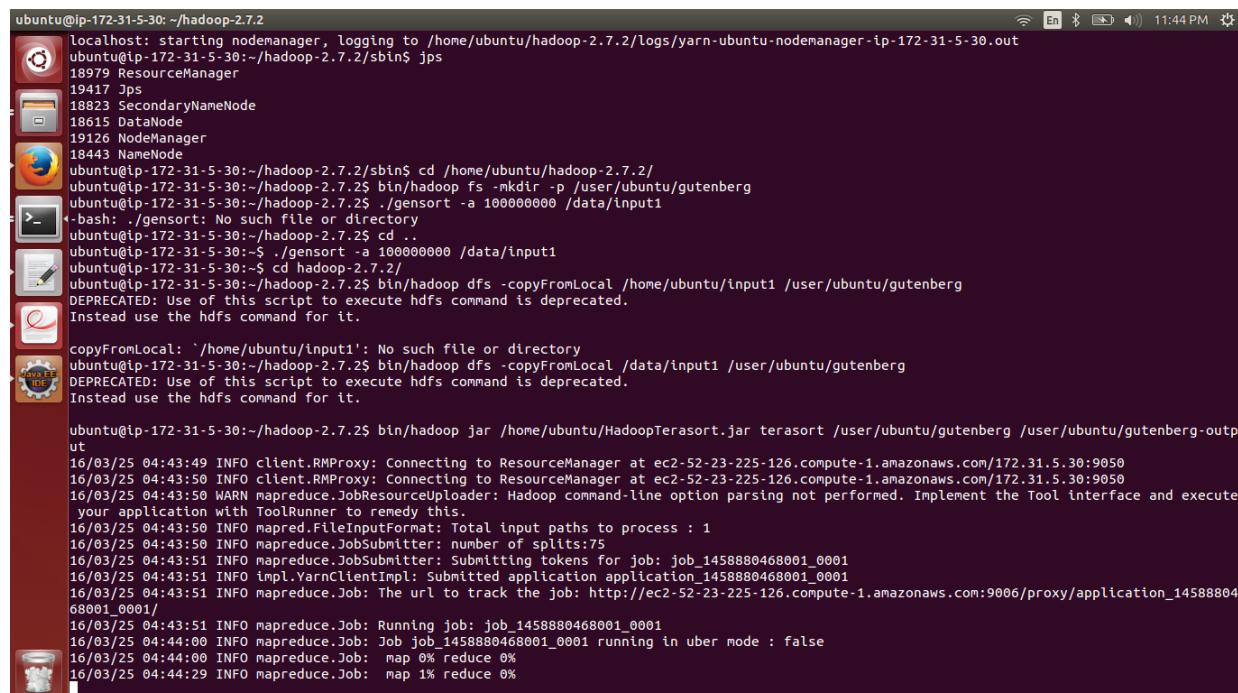
A20360111

6.9 Copy files from local to instance



```
snehal@snehal-Aspire-5738:~/pa2$ scp -i hadoop.pem /home/snehal/pa2/hadoop.pem ubuntu@52.23.225.126:/home/ubuntu
hadoop.pem                                100% 1692      1.7KB/s  00:00
snehal@snehal-Aspire-5738:~/pa2$ scp -i hadoop.pem /home/snehal/pa2/gensort-linu
x-1.5/64/gensort ubuntu@52.23.225.126:/home/ubuntu
gensort                                    100%   138KB 137.7KB/s  00:00
snehal@snehal-Aspire-5738:~/pa2$ scp -i hadoop.pem /home/snehal/pa2/gensort-linu
x-1.5/64/valsort ubuntu@52.23.225.126:/home/ubuntu
valsort                                     100%   131KB 131.4KB/s  00:00
snehal@snehal-Aspire-5738:~/pa2$ scp -i hadoop.pem /home/snehal/Desktop/Assignment2/HadoopTerasort.jar ubuntu@52.23.225.126:/home/ubuntu
HadoopTerasort.jar                         100% 5162      5.0KB/s  00:00
snehal@snehal-Aspire-5738:~/pa2$
```

Output



```
ubuntu@ip-172-31-5-30:~/hadoop-2.7.2
localhost: starting nodemanager, logging to /home/ubuntu/hadoop-2.7.2/logs/yarn-ubuntu-nodemanager-ip-172-31-5-30.out
18979 ResourceManager
19417 Jps
18823 SecondaryNameNode
18615 DataNode
19126 NodeManager
18443 NameNode
ubuntu@ip-172-31-5-30:~/hadoop-2.7.2/sbin$ cd /home/ubuntu/hadoop-2.7.2/
ubuntu@ip-172-31-5-30:~/hadoop-2.7.2$ bin/hadoop fs -mkdirr -p /user/ubuntu/gutenberg
ubuntu@ip-172-31-5-30:~/hadoop-2.7.2$ ./gensort -a 100000000 /data/input1
-bash: ./gensort: No such file or directory
ubuntu@ip-172-31-5-30:~/hadoop-2.7.2$ cd ..
ubuntu@ip-172-31-5-30:~/hadoop-2.7.2$ ./gensort -a 100000000 /data/input1
ubuntu@ip-172-31-5-30:~/hadoop-2.7.2$ cd hadoop-2.7.2
ubuntu@ip-172-31-5-30:~/hadoop-2.7.2$ bin/hadoop dfs -copyFromLocal /home/ubuntu/input1 /user/ubuntu/gutenberg
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

copyFromLocal: '/home/ubuntu/input1': No such file or directory
ubuntu@ip-172-31-5-30:~/hadoop-2.7.2$ bin/hadoop dfs -copyFromLocal /data/input1 /user/ubuntu/gutenberg
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

ubuntu@ip-172-31-5-30:~/hadoop-2.7.2$ bin/hadoop jar /home/ubuntu/HadoopTerasort.jar terasort /user/ubuntu/gutenberg /user/ubuntu/gutenberg-out
ut
16/03/25 04:43:49 INFO client.RMProxy: Connecting to ResourceManager at ec2-52-23-225-126.compute-1.amazonaws.com/172.31.5.30:9050
16/03/25 04:43:50 INFO client.RMProxy: Connecting to ResourceManager at ec2-52-23-225-126.compute-1.amazonaws.com/172.31.5.30:9050
16/03/25 04:43:50 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute
your application with ToolRunner to remedy this.
16/03/25 04:43:50 INFO mapred.FileInputFormat: Total input paths to process : 1
16/03/25 04:43:50 INFO mapreduce.JobSubmitter: number of splits:75
16/03/25 04:43:51 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1458880468001_0001
16/03/25 04:43:51 INFO impl.YarnClientImpl: Submitted application application_1458880468001_0001
16/03/25 04:43:51 INFO mapreduce.Job: The url to track the job: http://ec2-52-23-225-126.compute-1.amazonaws.com:9006/proxy/application_14588804
68001_0001/
16/03/25 04:43:51 INFO mapreduce.Job: Running job: job_1458880468001_0001
16/03/25 04:44:00 INFO mapreduce.Job: Job job_1458880468001_0001 running in uber mode : false
16/03/25 04:44:00 INFO mapreduce.Job: map 0% reduce 0%
16/03/25 04:44:29 INFO mapreduce.Job: map 1% reduce 0%
```

PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

A20360111

Namenode Information - Mozilla Firefox

Error: Browser Prob... CS 553 (131 unread) EC2 Management C... All Applications Namenode information 11:52 PM

Safemode is off.

23 files and directories, 81 blocks = 104 total filesystem object(s).

Heap Memory used 72.9 MB of 165.5 MB Heap Memory. Max Heap Memory is 889 MB.

Non Heap Memory used 31.27 MB of 32.44 MB Committed Non Heap Memory. Max Non Heap Memory is 214 MB.

Configured Capacity:	73.7 GB
DFS Used:	9.39 GB (12.74%)
Non DFS Used:	18.07 GB
DFS Remaining:	46.24 GB (62.74%)
Block Pool Used:	9.39 GB (12.74%)
DataNodes usages% (Min/Median/Max/stdDev):	12.74% / 12.74% / 12.74% / 0.00%
Live Nodes	1 (Decommissioned: 0)
Dead Nodes	0 (Decommissioned: 0)
Decommissioning Nodes	0
Total Datanode Volume Failures	0 (0 B)
Number of Under-Replicated Blocks	2
Number of Blocks Pending Deletion	0
Block Deletion Start Time	3/24/2016, 11:34:06 PM

Mozilla Firefox seems slow... to... start. Learn How to Speed It Up Don't Tell Me Again

```
ubuntu@ip-172-31-5-30:~/hadoop-2.7.2
16/03/25 04:54:49 INFO mapreduce.Job: map 66% reduce 21%
16/03/25 04:54:52 INFO mapreduce.Job: map 67% reduce 21%
16/03/25 04:55:05 INFO mapreduce.Job: map 68% reduce 21%
16/03/25 04:55:17 INFO mapreduce.Job: map 69% reduce 21%
16/03/25 04:55:20 INFO mapreduce.Job: map 70% reduce 21%
16/03/25 04:55:26 INFO mapreduce.Job: map 71% reduce 21%
16/03/25 04:55:41 INFO mapreduce.Job: map 71% reduce 22%
16/03/25 04:55:47 INFO mapreduce.Job: map 71% reduce 23%
16/03/25 04:55:54 INFO mapreduce.Job: map 71% reduce 24%
16/03/25 04:56:02 INFO mapreduce.Job: map 72% reduce 24%
16/03/25 04:56:03 INFO mapreduce.Job: map 73% reduce 24%
16/03/25 04:56:06 INFO mapreduce.Job: map 74% reduce 24%
16/03/25 04:56:18 INFO mapreduce.Job: map 75% reduce 24%
16/03/25 04:56:27 INFO mapreduce.Job: map 76% reduce 24%
16/03/25 04:56:30 INFO mapreduce.Job: map 77% reduce 24%
16/03/25 04:56:48 INFO mapreduce.Job: map 77% reduce 25%
16/03/25 04:56:54 INFO mapreduce.Job: map 77% reduce 26%
16/03/25 04:57:09 INFO mapreduce.Job: map 78% reduce 26%
16/03/25 04:57:12 INFO mapreduce.Job: map 79% reduce 26%
16/03/25 04:57:14 INFO mapreduce.Job: map 80% reduce 26%
16/03/25 04:57:28 INFO mapreduce.Job: map 81% reduce 26%
16/03/25 04:57:31 INFO mapreduce.Job: map 82% reduce 26%
16/03/25 04:57:44 INFO mapreduce.Job: map 83% reduce 26%
16/03/25 04:57:49 INFO mapreduce.Job: map 84% reduce 26%
16/03/25 04:57:56 INFO mapreduce.Job: map 84% reduce 27%
16/03/25 04:58:02 INFO mapreduce.Job: map 84% reduce 28%
16/03/25 04:58:19 INFO mapreduce.Job: map 85% reduce 28%
16/03/25 04:58:22 INFO mapreduce.Job: map 86% reduce 28%
16/03/25 04:58:23 INFO mapreduce.Job: map 87% reduce 28%
16/03/25 04:58:35 INFO mapreduce.Job: map 88% reduce 28%
16/03/25 04:58:44 INFO mapreduce.Job: map 89% reduce 28%
16/03/25 04:58:48 INFO mapreduce.Job: map 90% reduce 28%
16/03/25 04:58:53 INFO mapreduce.Job: map 91% reduce 28%
16/03/25 04:59:00 INFO mapreduce.Job: map 91% reduce 29%
16/03/25 04:59:03 INFO mapreduce.Job: map 91% reduce 30%
16/03/25 04:59:24 INFO mapreduce.Job: map 93% reduce 30%
16/03/25 04:59:30 INFO mapreduce.Job: map 94% reduce 30%
16/03/25 04:59:42 INFO mapreduce.Job: map 95% reduce 30%
16/03/25 04:59:55 INFO mapreduce.Job: map 96% reduce 30%
16/03/25 04:59:58 INFO mapreduce.Job: map 97% reduce 30%
```

PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

A20360111

```
ubuntu@ip-172-31-5-30: ~/hadoop-2.7.2
16/03/25 05:04:16 INFO mapreduce.Job: Job job_1458880468001_0001 completed successfully
16/03/25 05:04:17 INFO mapreduce.Job: Counters: 50
File System Counters
    FILE: Number of bytes read=30256951212
    FILE: Number of bytes written=40465923320
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=10000313154
    HDFS: Number of bytes written=10000000000
    HDFS: Number of read operations=228
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
Job Counters
    Killed map tasks=1
    Launched map tasks=76
    Launched reduce tasks=1
    Data-local map tasks=76
    Total time spent by all maps in occupied slots (ms)=5003456
    Total time spent by all reduces in occupied slots (ms)=1008307
    Total time spent by all map tasks (ms)=5003456
    Total time spent by all reduce tasks (ms)=1008307
    Total vcore-milliseconds taken by all map tasks=5003456
    Total vcore-milliseconds taken by all reduce tasks=1008307
    Total megabyte-milliseconds taken by all map tasks=5123538944
    Total megabyte-milliseconds taken by all reduce tasks=1032506368
Map-Reduce Framework
    Map input records=100000000
    Map output records=100000000
    Map output bytes=10000000000
    Map output materialized bytes=10200000450
    Input split bytes=10050
    Combine input records=100000000
    Combine output records=100000000
    Reduce input groups=100000000
    Reduce shuffle bytes=10200000450
    Reduce input records=100000000
    Reduce output records=100000000
    Spilled Records=396636763
    Shuffled Maps =75
    Failed Shuffles=0
    Merged Map outputs=75
    GC time elapsed (ms)=66650
    CPU time spent (ms)=1496740
    Physical memory (bytes) snapshot=19365974016
    Virtual memory (bytes) snapshot=62888722432
    Total committed heap usage (bytes)=15564537856
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=10000303104
File Output Format Counters
    Bytes Written=10000000000
ubuntu@ip-172-31-5-30:~/hadoop-2.7.2$
```

```
ubuntu@ip-172-31-5-30: ~/hadoop-2.7.2
Total time spent by all maps in occupied slots (ms)=5003456
Total time spent by all reduces in occupied slots (ms)=1008307
Total time spent by all map tasks (ms)=5003456
Total time spent by all reduce tasks (ms)=1008307
Total vcore-milliseconds taken by all map tasks=5003456
Total vcore-milliseconds taken by all reduce tasks=1008307
Total megabyte-milliseconds taken by all map tasks=5123538944
Total megabyte-milliseconds taken by all reduce tasks=1032506368
Map-Reduce Framework
    Map input records=100000000
    Map output records=100000000
    Map output bytes=10000000000
    Map output materialized bytes=10200000450
    Input split bytes=10050
    Combine input records=100000000
    Combine output records=100000000
    Reduce input groups=100000000
    Reduce shuffle bytes=10200000450
    Reduce input records=100000000
    Reduce output records=100000000
    Spilled Records=396636763
    Shuffled Maps =75
    Failed Shuffles=0
    Merged Map outputs=75
    GC time elapsed (ms)=66650
    CPU time spent (ms)=1496740
    Physical memory (bytes) snapshot=19365974016
    Virtual memory (bytes) snapshot=62888722432
    Total committed heap usage (bytes)=15564537856
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=10000303104
File Output Format Counters
    Bytes Written=10000000000
ubuntu@ip-172-31-5-30:~/hadoop-2.7.2$
```

PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

A20360111

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	Vcores Used	Vcores Total	Vcores Reserved	Active Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Reboot Node
1	0	1	0	1	2 GB	8 GB	0 B	1	8	0	1	0	0	0	0

Scheduler Type		Scheduling Resource Type		Minimum Allocation				Maximum Allocation					
Capacity Scheduler		[MEMORY]		<memory:1024, vCores:1>				<memory:8192, vCores:8>					
Show 20 entries													
Search:													
ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Progress	Tracking UI	Blacklisted Nodes	History	0
application_145880468001_0001	ubuntu	Terasort	MAPREDUCE	default	Thu Mar 24 23:43:51 -0500 2016	Fri Mar 25 00:04:15 -0500 2016	FINISHED	SUCCEEDED					

Showing 1 to 1 of 1 entries

First Previous 1 Next Last

Mozilla Firefox seems slow... to... start.

Valsort

```

ubuntu@ip-172-31-5-30: ~
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=100000303104
File Output Format Counters
Bytes Written=10000000000
ubuntu@ip-172-31-5-30:~/hadoop-2.7.2$ cd ..
ubuntu@ip-172-31-5-30:~$ cd /data
ubuntu@ip-172-31-5-30:/data$ cd ..
ubuntu@ip-172-31-5-30:/$ cd /home/ubuntu
ubuntu@ip-172-31-5-30:~$ cd hadoop-2.7.2/
ubuntu@ip-172-31-5-30:~/hadoop-2.7.2$ bin/hadoop dfs -ls /user/ubuntu/gutenberg-output
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Found 2 items
-rw-r--r-- 1 ubuntu supergroup 0 2016-03-25 05:04 /user/ubuntu/gutenberg-output/_SUCCESS
-rw-r--r-- 1 ubuntu supergroup 10000000000 2016-03-25 05:04 /user/ubuntu/gutenberg-output/part-00000
ubuntu@ip-172-31-5-30:~/hadoop-2.7.2$ cd ..
ubuntu@ip-172-31-5-30:~$ cd hadoop-2.7.2/
ubuntu@ip-172-31-5-30:~/hadoop-2.7.2$ bin/hadoop dfs -getmerge /user/ubuntu/gutenberg-output /data/result
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

ubuntu@ip-172-31-5-30:~/hadoop-2.7.2$ cd ..
ubuntu@ip-172-31-5-30:~$ cd /data
ubuntu@ip-172-31-5-30:~/data$ ls
dfs lost+found nn-local-dir result
ubuntu@ip-172-31-5-30:~/data$ ls -l
total 9765656
drwx----- 2 root root 16384 Mar 25 04:31 lost+found
drwxrwxr-x 5 ubuntu ubuntu 4096 Mar 25 04:34 dfs
-rw-r--r-- 1 ubuntu ubuntu 10000000000 Mar 25 05:13 result
drwxr-xr-x 5 ubuntu ubuntu 4096 Mar 25 05:14 nn-local-dir
ubuntu@ip-172-31-5-30:~/data$ cd ..
ubuntu@ip-172-31-5-30:~/cd /home/ubuntu
ubuntu@ip-172-31-5-30:~$ ./valsort /data/result
Records: 100000000
Checksum: 2fb0574596d67c8
Duplicate keys: 0
SUCCESS - all records are in order
ubuntu@ip-172-31-5-30:~$ 
```

PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

A20360111

2. Installation of Hadoop 16-nodes

Step 1: sudo apt-get update

```
sudo apt-get install default-jdk  
sudo apt-get install ssh
```

Step 2: Set up SSH certificates to avoid entering password each time.

```
ubuntu@ip-172-31-31-1: ~  
Processing triggers for ufw (0.34-rc2ubuntu2) ...  
Setting up openssh-client (1:6.6p1-2ubuntu2.6) ...  
Setting up openssh-server (1:6.6p1-2ubuntu2.6) ...  
ssh stop/waiting  
ssh start/running, process 10989  
Setting up ssh (1:6.6p1-2ubuntu2.6) ...  
ubuntu@ip-172-31-31-1:~$ sudo apt-get install vim  
Reading package lists... Done  
Building dependency tree  
Reading state information... Done  
vim is already the newest version.  
0 upgraded, 0 newly installed, 0 to remove and 67 not upgraded.  
ubuntu@ip-172-31-31-1:~$ ssh-keygen -t rsa -P ""  
Generating public/private rsa key pair.  
Enter file in which to save the key (/home/ubuntu/.ssh/id_rsa):  
Your identification has been saved in /home/ubuntu/.ssh/id_rsa.  
Your public key has been saved in /home/ubuntu/.ssh/id_rsa.pub.  
The key fingerprint is:  
11:de:d9:06:70:75:4b:c6:e9:61:1f:41:fb:2e:f5:0c ubuntu@ip-172-31-31-1  
The key's randomart image is:  
++ [ RSA 2048] -----+  
o . o . + . |  
. + + o . o . |  
o o o . . . |  
. . = .. |  
S o E o + |  
. = o |  
. + |  
. |  
+-----+  
ubuntu@ip-172-31-31-1:~$ cat $HOME/.ssh/id_rsa.pub >> $HOME/.ssh/authorized_keys  
ubuntu@ip-172-31-31-1:~$ eval `ssh-agent -s`  
Agent pid 11010  
ubuntu@ip-172-31-31-1:~$ chmod 400 hadoop.pem  
chmod: cannot access 'hadoop.pem': No such file or directory  
ubuntu@ip-172-31-31-1:~$ chmod 400 hadoop.pem  
ubuntu@ip-172-31-31-1:~$ ssh-add hadoop.pem  
Identity added: hadoop.pem (hadoop.pem)  
ubuntu@ip-172-31-31-1:~$ cd /home/ubuntu  
ubuntu@ip-172-31-31-1:~$ wget archive.apache.org/dist/hadoop/common/hadoop-2.7.2/hadoop-2.7.2.tar.gz
```

Step 3: Install Hadoop, unzip and give permissions

```
ubuntu@ip-172-31-31-1: ~  
hadoop-2.7.2/share/hadoop/common/lib/commons-lang-2.6.jar  
hadoop-2.7.2/share/hadoop/common/lib/xz-1.0.jar  
hadoop-2.7.2/share/hadoop/common/lib/jackson-xc-1.9.13.jar  
hadoop-2.7.2/share/hadoop/common/lib/hadoop-annotations-2.7.2.jar  
hadoop-2.7.2/share/hadoop/common/lib/jaxb-api-2.2.2.jar  
hadoop-2.7.2/share/hadoop/common/lib/jersey-json-1.9.jar  
hadoop-2.7.2/share/hadoop/common/lib/protobuf-java-2.5.0.jar  
hadoop-2.7.2/share/hadoop/common/lib/httpcore-4.2.5.jar  
hadoop-2.7.2/share/hadoop/common/lib/avro-1.7.4.jar  
hadoop-2.7.2/share/hadoop/common/lib/commons-beanutils-core-1.8.0.jar  
hadoop-2.7.2/share/hadoop/common/lib/servlet-api-2.5.jar  
hadoop-2.7.2/share/hadoop/common/lib/api-asn1-api-1.0.0-M20.jar  
hadoop-2.7.2/share/hadoop/common/lib/gson-2.2.4.jar  
hadoop-2.7.2/share/hadoop/common/lib/commons-cli-1.2.jar  
hadoop-2.7.2/share/hadoop/common/lib/junit-4.11.jar  
hadoop-2.7.2/share/hadoop/common/lib/jettison-1.1.jar  
hadoop-2.7.2/share/hadoop/common/lib/jsr305-3.0.0.jar  
hadoop-2.7.2/share/hadoop/common/lib/commons-logging-1.1.3.jar  
hadoop-2.7.2/share/hadoop/common/lib/stf4j-log4j12-1.7.10.jar  
hadoop-2.7.2/share/hadoop/common/lib/hancrest-core-1.3.jar  
hadoop-2.7.2/share/hadoop/common/lib/stf4j-api-1.7.10.jar  
hadoop-2.7.2/share/hadoop/common/lib/commons-httpclient-3.1.jar  
hadoop-2.7.2/share/hadoop/common/lib/commons-beanutils-1.7.0.jar  
hadoop-2.7.2/share/hadoop/common/lib/paranamer-2.3.jar  
hadoop-2.7.2/share/hadoop/common/hadoop-common-2.7.2-tests.jar  
hadoop-2.7.2/share/hadoop/common/sources/  
hadoop-2.7.2/share/hadoop/common/sources/hadoop-common-2.7.2-sources.jar  
hadoop-2.7.2/share/hadoop/common/sources/hadoop-common-2.7.2-test-sources.jar  
hadoop-2.7.2/lib/  
hadoop-2.7.2/lib/native/  
hadoop-2.7.2/lib/native/libhdfs.so  
hadoop-2.7.2/lib/native/libhadooputils.a  
hadoop-2.7.2/lib/native/libhdfs.so.0.0.0  
hadoop-2.7.2/lib/native/libhadoop.so.1.0.0  
hadoop-2.7.2/lib/native/libhadoop.a  
hadoop-2.7.2/lib/native/libhdfs.a  
hadoop-2.7.2/lib/native/libhadoop.so  
hadoop-2.7.2/lib/native/libhadooppipes.a  
hadoop-2.7.2/LICENSE.txt  
ubuntu@ip-172-31-31-1:~$ chmod 777 hadoop-2.7.2  
ubuntu@ip-172-31-31-1:~$
```

PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

A20360111

Step 4: Mount EBS volume

```
ubuntu@ip-172-31-31-1: ~
7864320 blocks (5.00%) reserved for the super user
First data block=0
Maximum filesystem blocks=4294967296
4800 block groups
32768 blocks per group, 32768 fragments per group
8192 inodes per group
Superblock backups stored on blocks:
    32768, 98304, 163840, 229376, 294912, 819200, 884736, 1605632, 2654208,
    4096000, 7962624, 11239424, 20480000, 23887872, 71663616, 78675968,
    102400000

#Allocating group tables: done
Writing inode tables: done
Creating journal (32768 blocks): done
Writing superblocks and filesystem accounting information: done

ubuntu@ip-172-31-31-1:~$ sudo mke2fs -F -t ext4 /dev/xvdd
mke2fs 1.42.9 (4-Feb-2014)
Filesystem label=
OS type: Linux
Block size=4096 (log=2)
Fragment size=4096 (log=2)
Stride=0 blocks, Stripe width=0 blocks
39321600 inodes, 157286400 blocks
7864320 blocks (5.00%) reserved for the super user
First data block=0
Maximum filesystem blocks=4294967296
4800 block groups
32768 blocks per group, 32768 fragments per group
8192 inodes per group
Superblock backups stored on blocks:
    32768, 98304, 163840, 229376, 294912, 819200, 884736, 1605632, 2654208,
    4096000, 7962624, 11239424, 20480000, 23887872, 71663616, 78675968,
    102400000

Allocating group tables: done
Writing inode tables: done
Creating journal (32768 blocks): done
Writing superblocks and filesystem accounting information: done

ubuntu@ip-172-31-31-1:~$
```

Step 5: Edit environmental variables in `~/.bashrc` file :

```
ubuntu@ip-172-31-31-1: ~
alias l='ls -CF'

# Add an "alert" alias for long running commands.  Use like so:
# sleep 10; alert
alias alert="notify-send --urgency=low -i "$( [ $? = 0 ] && echo terminal || echo error)" "${history|tail -n1|sed -e '\'s/^/\s*[0-9]\+\s*//;s/[;8]*/\s*alert'/'\""

# Alias definitions.
# You may want to put all your additions into a separate file like
# ~/.bash_aliases, instead of adding them here directly.
# See /usr/share/doc/bash-doc/examples in the bash-doc package.

if [ -f ~/.bash_aliases ]; then
. ~/.bash_aliases
fi

# enable programmable completion features (you don't need to enable
# this, if it's already enabled in /etc/bash.bashrc and /etc/profile
# sources /etc/bash.bashrc).
if ! shopt -q posix; then
if [ -f /usr/share/bash-completion/bash_completion ]; then
. /usr/share/bash-completion/bash_completion
elif [ -f /etc/bash_completion ]; then
. /etc/bash_completion
fi
fi


export JAVA_HOME=/usr/lib/jvm/java-7-openjdk-amd64
export HADOOP_INSTALL=/home/ubuntu/hadoop-2.7.2
export PATH=$PATH:$HADOOP_INSTALL/bin
export PATH=$PATH:$HADOOP_INSTALL/sbin
export HADOOP_MAPRED_HOME=$HADOOP_INSTALL
export HADOOP_COMMON_HOME=$HADOOP_INSTALL
export HADOOP_HDFS_HOME=$HADOOP_INSTALL
export YARN_HOME=$HADOOP_INSTALL
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_INSTALL/lib/native
export HADOOP_OPTS='-Djava.library.path=$HADOOP_INSTALL/lib'

-- INSERT --
```

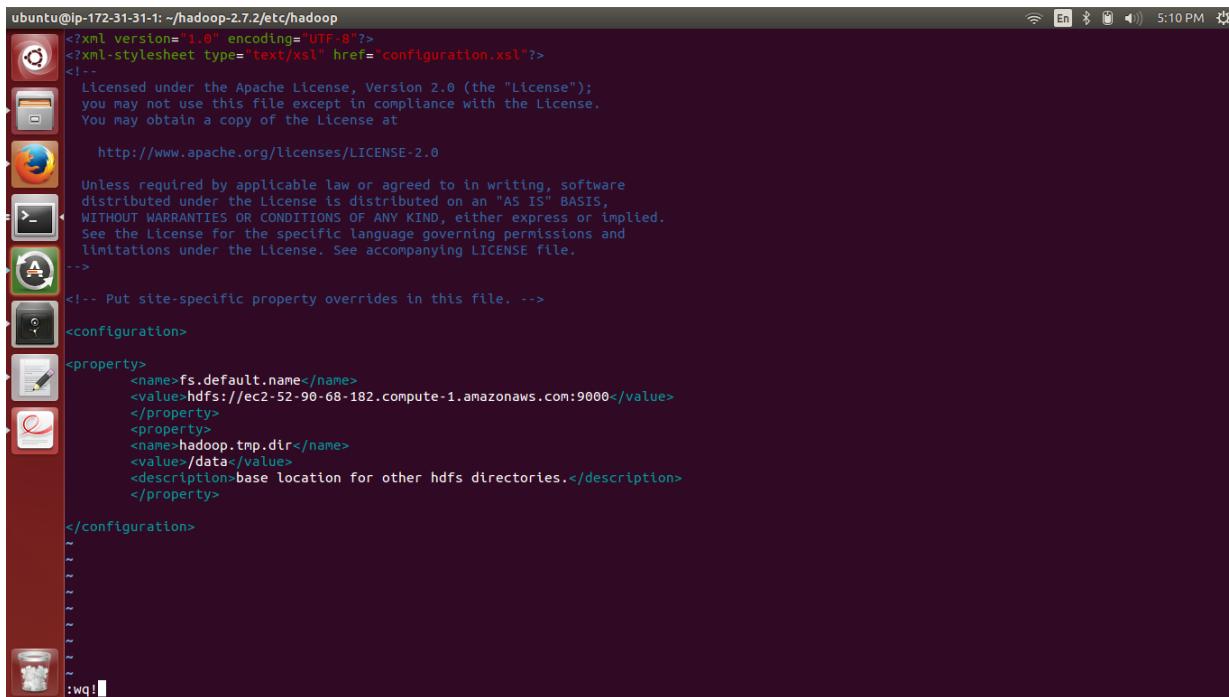
PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

A20360111

Step 6: Edit core-site.xml

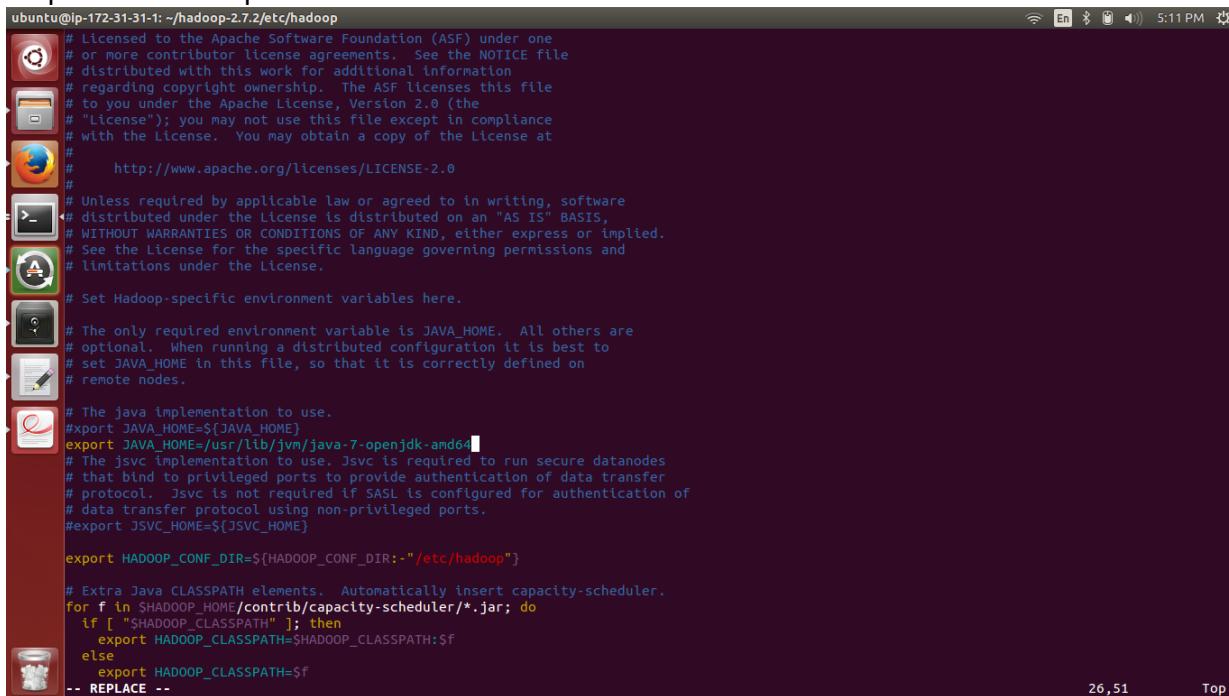


```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
 Licensed under the Apache License, Version 2.0 (the "License");
 you may not use this file except in compliance with the License.
 You may obtain a copy of the License at

 http://www.apache.org/licenses/LICENSE-2.0

 Unless required by applicable law or agreed to in writing, software
 distributed under the License is distributed on an "AS IS" BASIS,
 WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
 See the License for the specific language governing permissions and
 limitations under the License. See accompanying LICENSE file.
-->
<!-- Put site-specific property overrides in this file. -->
<configuration>
<property>
<name>fs.default.name</name>
<value>hdfs://ec2-52-90-68-182.compute-1.amazonaws.com:9000</value>
</property>
<property>
<name>hadoop.tmp.dir</name>
<value>/data</value>
<description>base location for other hdfs directories.</description>
</property>
</configuration>
~
```

Step 7: Edit hadoop-env.sh



```
# Licensed to the Apache Software Foundation (ASF) under one
# or more contributor license agreements. See the NOTICE file
# distributed with this work for additional information
# regarding copyright ownership. The ASF licenses this file
# to you under the Apache License, Version 2.0 (the
# "License"); you may not use this file except in compliance
# with the License. You may obtain a copy of the License at
#
#     http://www.apache.org/licenses/LICENSE-2.0
#
# Unless required by applicable law or agreed to in writing, software
# distributed under the License is distributed on an "AS IS" BASIS,
# WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
# See the License for the specific language governing permissions and
# limitations under the License.

# Set Hadoop-specific environment variables here.

# The only required environment variable is JAVA_HOME. All others are
# optional. When running a distributed configuration it is best to
# set JAVA_HOME in this file, so that it is correctly defined on
# remote nodes.

# The java implementation to use.
#export JAVA_HOME=${JAVA_HOME}
#export JAVA_HOME=/usr/lib/jvm/java-7-openjdk-amd64
# The jsvc implementation to use. Jsvc is required to run secure datanodes
# that bind to privileged ports to provide authentication of data transfer
# protocol. Jsvc is not required if SASL is configured for authentication of
# data transfer protocol using non-privileged ports.
#export JSVC_HOME=${JSVC_HOME}

export HADOOP_CONF_DIR=${HADOOP_CONF_DIR:-"/etc/hadoop"}

# Extra Java CLASSPATH elements. Automatically insert capacity-scheduler.
for f in $HADOOP_HOME/contrib/capacity-scheduler/*.jar; do
  if [ "$HADOOP_CLASSPATH" ]; then
    export HADOOP_CLASSPATH=$HADOOP_CLASSPATH:$f
  else
    export HADOOP_CLASSPATH=$f
  fi
done
-- REPLACE --
```

Step 8: Edit hdfs-site.xml

PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

A20360111

```
ubuntu@lp-172-31-31-1: ~/hadoop-2.7.2/etc/hadoop
  xmlns="http://www.w3.org/2001/XMLSchema">
  <?xml version="1.0" encoding="UTF-8"?>
  <xsl:stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
  <property>
    <name>dfs.permissions</name>
    <value>false</value>
  </property>
</configuration>
~
```

Step 9: Edit mapred-site.xml

```
ubuntu@ip-172-31-31-1: ~/hadoop-2.7.2/etc/hadoop
<configuration>
  <property>
    <name>mapreduce.job.tracker</name>
    <value>hdfs://ec2-52-90-68-182.compute-1.amazonaws.com:9001</value>
  </property>
  <property>
    <name>mapreduce.framework.name</name>
    <value>yarn</value></property>
</configuration>
```

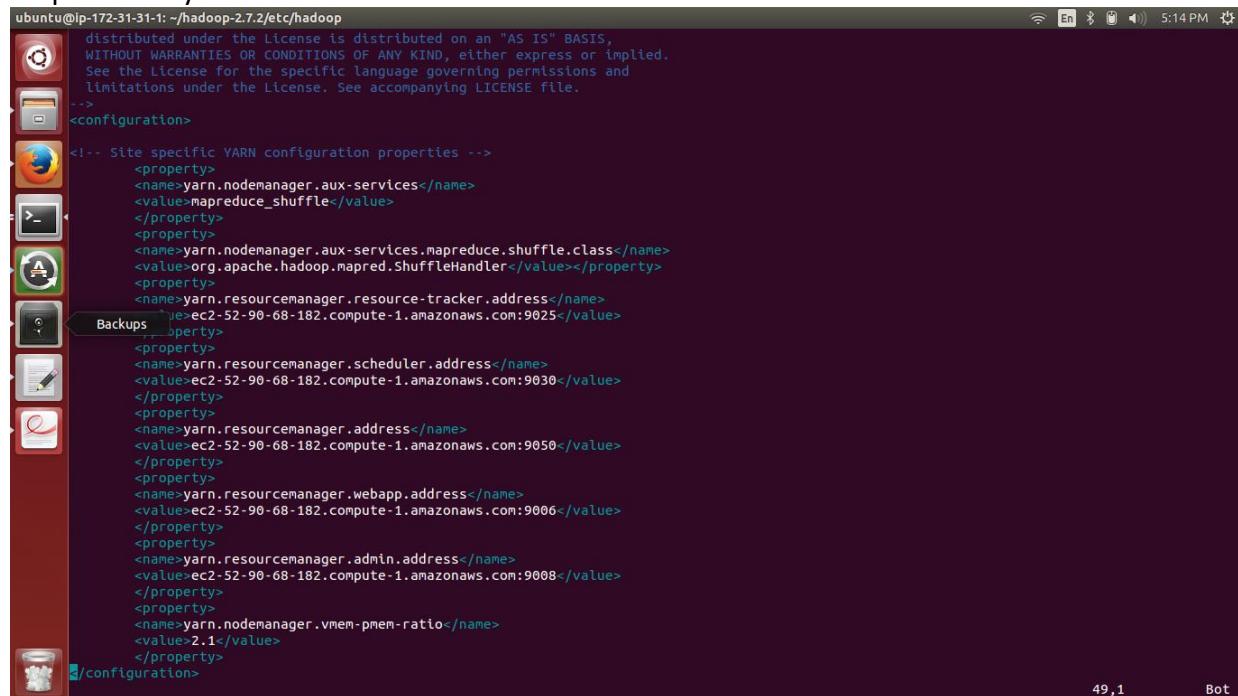
PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

A20360111

Step 10: Edit yarn-site.xml



```
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.

-->
<configuration>

<!-- Site specific YARN configuration properties -->
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
</property>
<property>
<name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
<value>org.apache.hadoop.mapred.ShuffleHandler</value></property>
<property>
<name>yarn.resourcemanager.resource-tracker.address</name>
<value>ec2-52-90-68-182.compute-1.amazonaws.com:9025</value>
</property>
<!-- Backups -->
<property>
<name>yarn.resourcemanager.scheduler.address</name>
<value>ec2-52-90-68-182.compute-1.amazonaws.com:9030</value>
</property>
<property>
<name>yarn.resourcemanager.address</name>
<value>ec2-52-90-68-182.compute-1.amazonaws.com:9050</value>
</property>
<property>
<name>yarn.resourcemanager.webapp.address</name>
<value>ec2-52-90-68-182.compute-1.amazonaws.com:9006</value>
</property>
<property>
<name>yarn.resourcemanager.admin.address</name>
<value>ec2-52-90-68-182.compute-1.amazonaws.com:9008</value>
</property>
<property>
<name>yarn.nodemanager.vmem-pmem-ratio</name>
<value>2.1</value>
</property>
</configuration>
```

Step 11: Create image of the Master Node

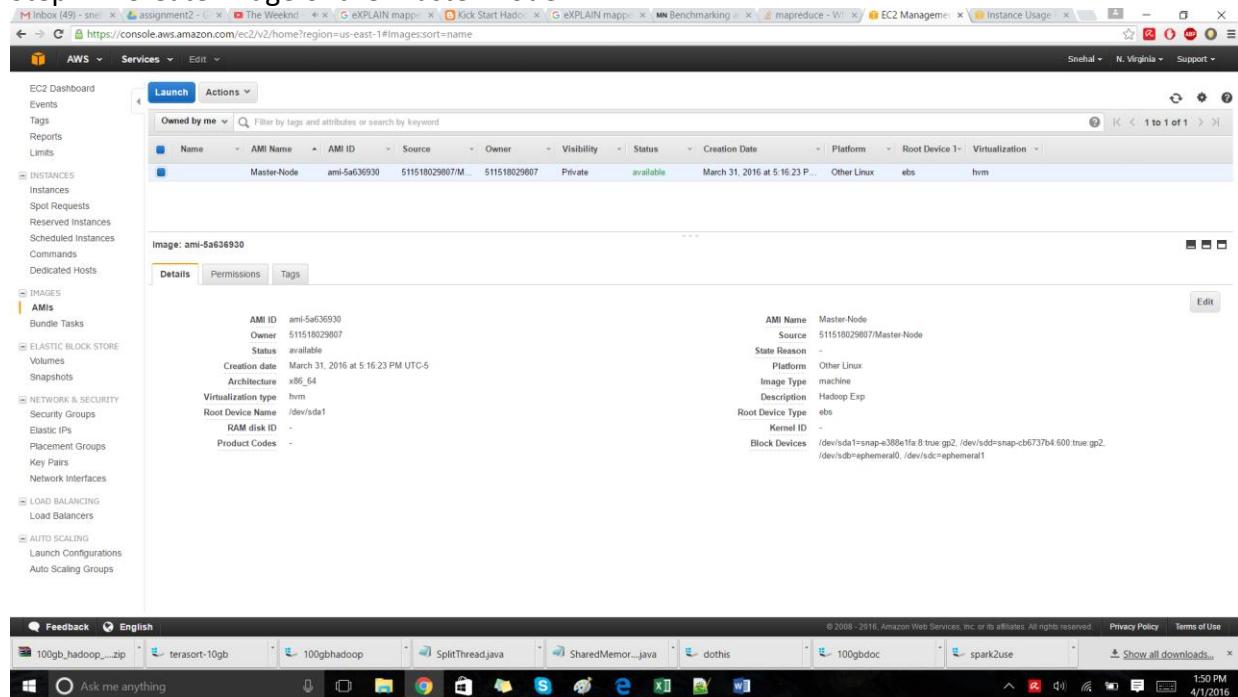


Image: ami-5a636930

AMI ID: ami-5a636930
Owner: 511518029807
Status: available
Creation date: March 31, 2016 at 5:16:23 PM UTC-5
Architecture: x86_64
Virtualization type: hvm
Root Device Name: /dev/sda1
RAM disk ID: -
Product Codes: -

AMI Name: Master-Node
Source: 511518029807/Master-Node
Start Reason: -
Platform: Other Linux
Image Type: machine
Description: Hadoop Exp
Root Device Type: ebs
Kernel ID: -
Block Devices: /dev/sda1=snap-e398e1fa:8 true gp2, /dev/sdd=snap-cb6737b4:600 true gp2, /dev/sdb=ephemeral0, /dev/sdc=ephemeral1

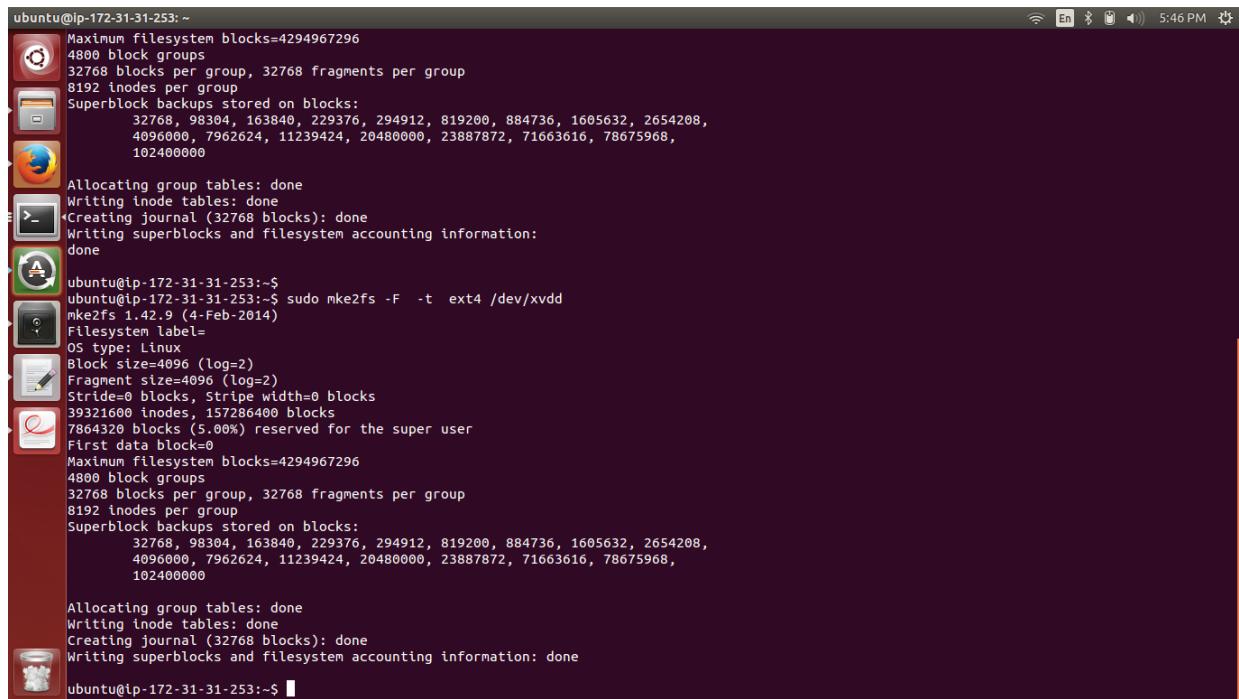
PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

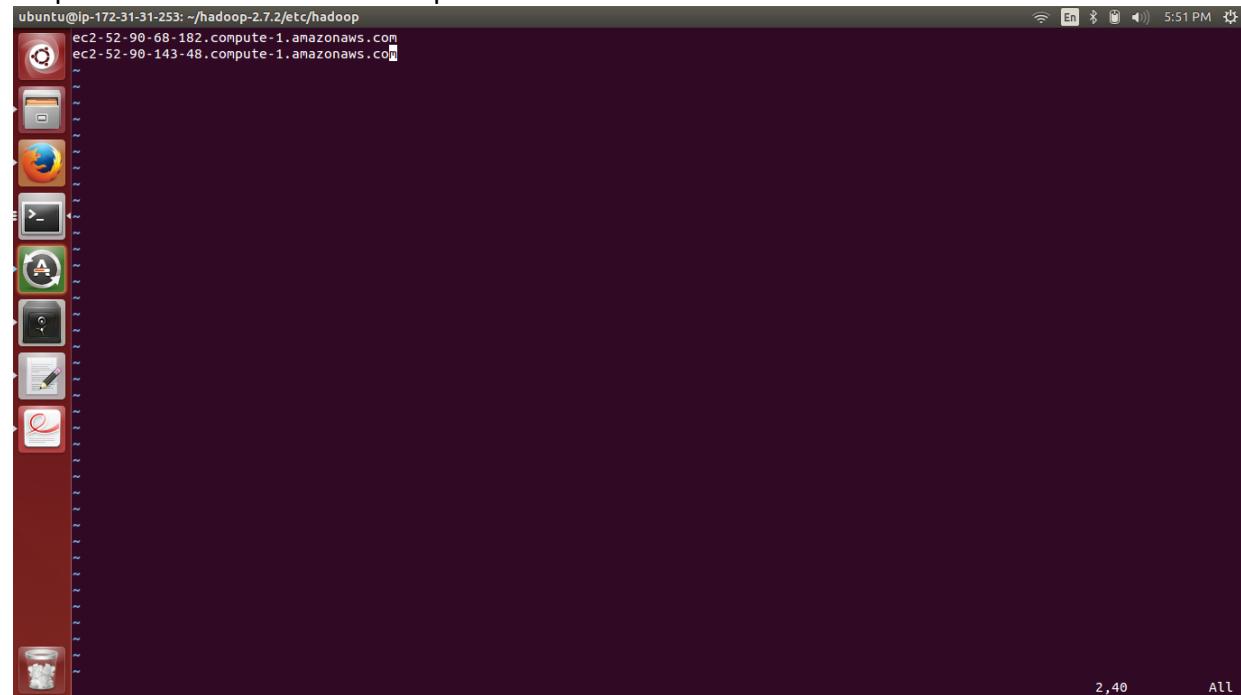
A20360111

Step 12: Connect to each slave and mount the EBS volume.



```
ubuntu@ip-172-31-31-253:~$ Maximum filesystem blocks=4294967296
4800 block groups
32768 blocks per group, 32768 fragments per group
8192 inodes per group
Superblock backups stored on blocks:
      32768, 98304, 163840, 229376, 294912, 819200, 884736, 1605632, 2654208,
      4096000, 7962624, 11239424, 20480000, 23887872, 71663616, 78675968,
      102400000
Allocating group tables: done
Writing inode tables: done
Creating journal (32768 blocks): done
Writing superblocks and filesystem accounting information: done
ubuntu@ip-172-31-31-253:~$ sudo mke2fs -F -t ext4 /dev/xvdd
mke2fs 1.42.9 (4-Feb-2014)
Filesystem label=
OS type: Linux
Block size=4096 (log=2)
Fragment size=4096 (log=2)
Stride=0 blocks, Stripe width=0 blocks
39321600 inodes, 157286400 blocks
7864320 blocks (5.00%) reserved for the super user
First data block=0
Maximum filesystem blocks=4294967296
4800 block groups
32768 blocks per group, 32768 fragments per group
8192 inodes per group
Superblock backups stored on blocks:
      32768, 98304, 163840, 229376, 294912, 819200, 884736, 1605632, 2654208,
      4096000, 7962624, 11239424, 20480000, 23887872, 71663616, 78675968,
      102400000
Allocating group tables: done
Writing inode tables: done
Creating journal (32768 blocks): done
Writing superblocks and filesystem accounting information: done
ubuntu@ip-172-31-31-253:~$
```

Step 13: Edit the slaves file with public DNS of master and slave



```
ubuntu@ip-172-31-31-253:~/hadoop-2.7.2/etc/hadoop$ ec2-52-90-68-182.compute-1.amazonaws.com
ec2-52-90-143-48.compute-1.amazonaws.com
```

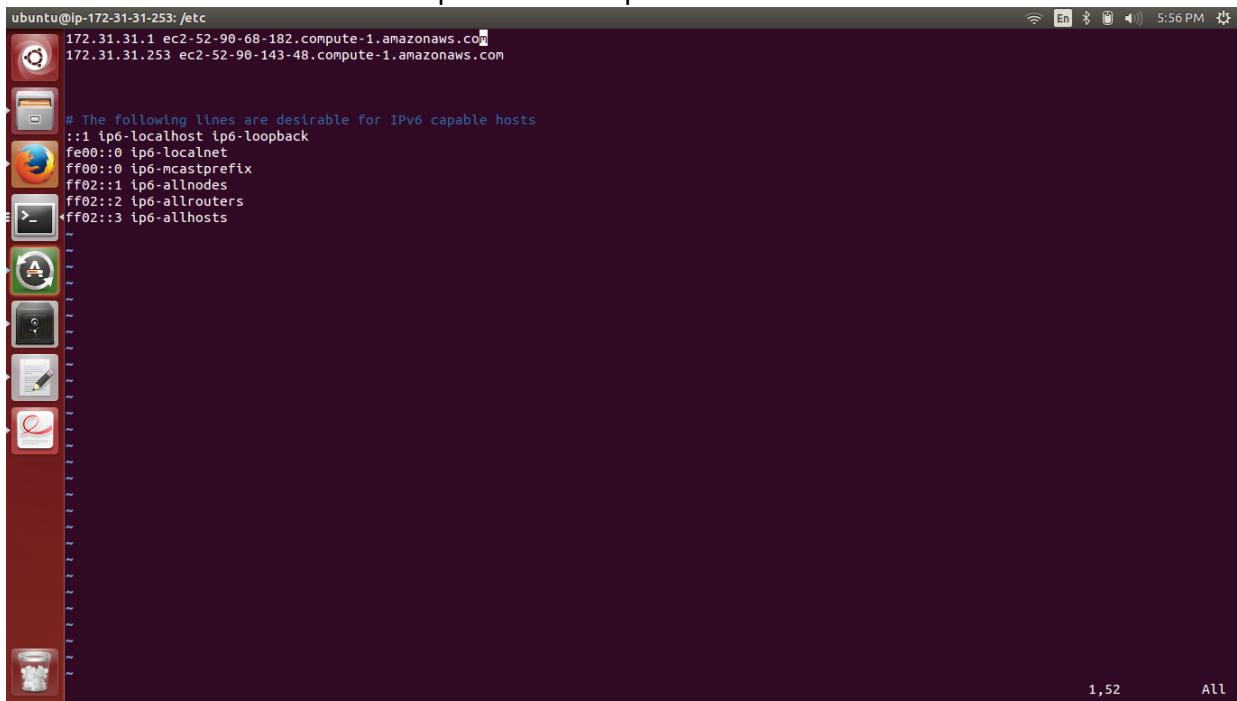
PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

A20360111

Step 14: Edit the hosts file with the with private IP and public DNS

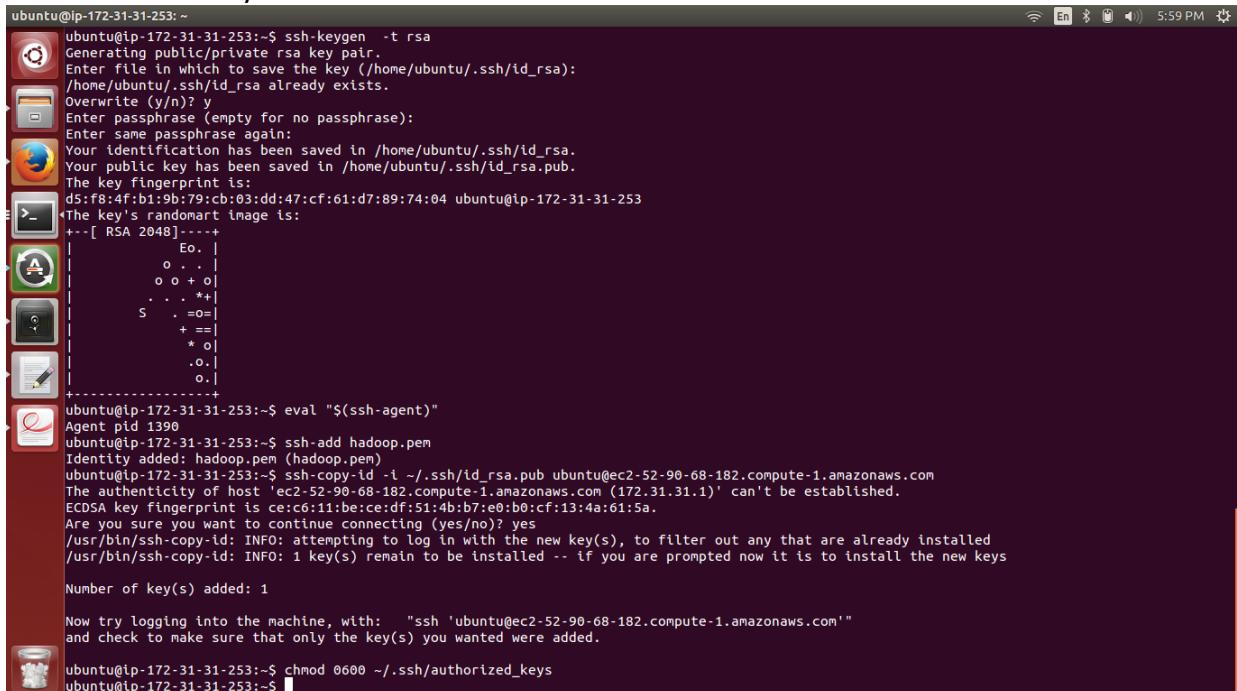


ubuntu@ip-172-31-31-253: /etc

```
172.31.31.1 ec2-52-90-68-182.compute-1.amazonaws.com
172.31.31.253 ec2-52-90-143-48.compute-1.amazonaws.com

# The following lines are desirable for IPv6 capable hosts
::1 ip6-localhost ip6-loopback
fe00::0 ip6-localnet
ff00::1 ip6-allnodes
ff02::1 ip6-allrouters
ff02::2 ip6-allrouters
ff02::3 ip6-allhosts
```

Step 15: Generate RSA key on each slave



```
ubuntu@ip-172-31-31-253: ~
ubuntu@ip-172-31-31-253:~$ ssh-keygen -t rsa
Generating public/private rsa key pair.
Enter file in which to save the key (/home/ubuntu/.ssh/id_rsa):
/home/ubuntu/.ssh/id_rsa already exists.
Overwrite (y/n)? y
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/ubuntu/.ssh/id_rsa.
Your public key has been saved in /home/ubuntu/.ssh/id_rsa.pub.
The key fingerprint is:
d5:f8:4f:b1:9b:79:cb:03:dd:47:cf:61:d7:89:74:04 ubuntu@ip-172-31-31-253
The key's randomart image is:
++[ RSA 2048]---+
Eo. |
o . . |
o o + o|
. . . *+|
S . =o=|
+ ==|
* o|
. o|
o.|
```

```
ubuntu@ip-172-31-31-253:~$ eval "$(ssh-agent)"
Agent pid 1390
ubuntu@ip-172-31-31-253:~$ ssh-add hadoop.pem
Identity added: hadoop.pem (hadoop.pem)
ubuntu@ip-172-31-31-253:~$ ssh-copy-id -l ~/ssh/id_rsa.pub ubuntu@ec2-52-90-68-182.compute-1.amazonaws.com
The authenticity of host 'ec2-52-90-68-182.compute-1.amazonaws.com (172.31.31.1)' can't be established.
ECDSA key fingerprint is ce:c6:11:be:ce:df:51:4b:b7:e0:b6:cf:13:4a:61:5a.
Are you sure you want to continue connecting (yes/no)? yes
/usr/bin/ssh-copy-id: INFO: attempting to log in with the new key(s), to filter out any that are already installed
/usr/bin/ssh-copy-id: INFO: 1 key(s) remain to be installed -- if you are prompted now it is to install the new keys

Number of key(s) added: 1

Now try logging into the machine, with: "ssh 'ubuntu@ec2-52-90-68-182.compute-1.amazonaws.com'"
and check to make sure that only the key(s) you wanted were added.
```

```
ubuntu@ip-172-31-31-253:~$ chmod 0600 ~/.ssh/authorized_keys
ubuntu@ip-172-31-31-253:~$
```

PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

A20360111

Step 16: Generate rsa on Master and copy to all slaves

```
ubuntu@ip-172-31-31-1:~/hadoop-2.7.2/bin
and check to make sure that only the key(s) you wanted were added.
ubuntu@ip-172-31-31-1:~$ ssh-copy-id -i ~/ssh/id_rsa.pub ubuntu@ec2-52-90-141-123.compute-1.amazonaws.com
The authenticity of host 'ec2-52-90-141-123.compute-1.amazonaws.com (172.31.31.242)' can't be established.
ECDSA key fingerprint is 26:07:63:ca:78:8d:6f:dd:bd:42:id:if:59:7e:id:c1.
Are you sure you want to continue connecting (yes/no)? yes
/usr/bin/ssh-copy-id: INFO: attempting to log in with the new key(s), to filter out any that are already installed
/usr/bin/ssh-copy-id: INFO: 1 key(s) remain to be installed -- if you are prompted now it is to install the new keys

Number of key(s) added: 1

Now try logging into the machine, with: "ssh 'ubuntu@ec2-52-90-141-123.compute-1.amazonaws.com'"
and check to make sure that only the key(s) you wanted were added.

ubuntu@ip-172-31-31-1:~$ ssh-copy-id -i ~/ssh/id_rsa.pub ubuntu@ec2-54-85-72-182.compute-1.amazonaws.com
The authenticity of host 'ec2-54-85-72-182.compute-1.amazonaws.com (172.31.31.243)' can't be established.
ECDSA key fingerprint is aa:3d:8b:7b:38:fc:bf:aab7:02:ed:af:22:05:26:bd.
Are you sure you want to continue connecting (yes/no)? yes
/usr/bin/ssh-copy-id: INFO: attempting to log in with the new key(s), to filter out any that are already installed
/usr/bin/ssh-copy-id: INFO: 1 key(s) remain to be installed -- if you are prompted now it is to install the new keys

Number of key(s) added: 1

Now try logging into the machine, with: "ssh 'ubuntu@ec2-54-85-72-182.compute-1.amazonaws.com'"
and check to make sure that only the key(s) you wanted were added.

ubuntu@ip-172-31-31-1:~$ ssh-copy-id -i ~/ssh/id_rsa.pub ubuntu@ec2-54-173-128-111.compute-1.amazonaws.com
The authenticity of host 'ec2-54-173-128-111.compute-1.amazonaws.com (172.31.31.244)' can't be established.
ECDSA key fingerprint is 62:f6:54:af:82:3b:98:56:73:31:7e:8d:e2:b5:56:45.
Are you sure you want to continue connecting (yes/no)? yes
/usr/bin/ssh-copy-id: INFO: attempting to log in with the new key(s), to filter out any that are already installed
/usr/bin/ssh-copy-id: INFO: 1 key(s) remain to be installed -- if you are prompted now it is to install the new keys

Number of key(s) added: 1

Now try logging into the machine, with: "ssh 'ubuntu@ec2-54-173-128-111.compute-1.amazonaws.com'"
and check to make sure that only the key(s) you wanted were added.

ubuntu@ip-172-31-31-1:~$ chmod 600 ~/.ssh/authorized_keys
ubuntu@ip-172-31-31-1:~$ cd /home/ubuntu/hadoop-2.7.2/bin
ubuntu@ip-172-31-31-1:~/hadoop-2.7.2/bin$ ./hdfs namenode -format
```

Step 17: Start dfs and yarn

```
ubuntu@ip-172-31-31-1:~/hadoop-2.7.2/sbin
16/04/01 00:53:07 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.window.num.buckets = 10
16/04/01 00:53:07 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.num.users = 10
16/04/01 00:53:07 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.windows.minutes = 1,5,25
16/04/01 00:53:07 INFO namenode.FSNamesystem: Retry cache on namenode is enabled
16/04/01 00:53:07 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total heap and retry cache entry expiry time is 600000 millis
16/04/01 00:53:07 INFO util.GSet: Computing capacity for map NameNodeRetryCache
16/04/01 00:53:07 INFO util.GSet: VM type          = 64-bit
16/04/01 00:53:07 INFO util.GSet: capacity        = 2'15 = 32768 entries
16/04/01 00:53:07 INFO util.GSet: capacity        = 2'15 = 32768 entries
16/04/01 00:53:07 INFO namenode.FSImage: Allocated new BlockPoolId: BP-516609664-172.31.31.1-1459471987542
16/04/01 00:53:07 INFO common.Storage: Storage directory /data/dfs/name has been successfully formatted.
16/04/01 00:53:07 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
16/04/01 00:53:07 INFO util.ExitUtil: Exiting with status 0
16/04/01 00:53:07 INFO namenode.NameNode: SHUTDOWN_MSG:
*****SHUTDOWN_MSG: Shutting down NameNode at ip-172-31-31-1.ec2.internal/172.31.31.1*****
*****SHUTDOWN_MSG: Shutting down NameNode at ip-172-31-31-1.ec2.internal/172.31.31.1*****
ubuntu@ip-172-31-31-1:~/hadoop-2.7.2/bin$ cd /home/ubuntu/hadoop-2.7.2/sbin/
ubuntu@ip-172-31-31-1:~/hadoop-2.7.2/sbin$ ./start-dfs.sh
16/04/01 00:53:51 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Starting namenodes on [ec2-52-90-68-182.compute-1.amazonaws.com]
ec2-52-90-68-182.compute-1.amazonaws.com: starting namenode, logging to /home/ubuntu/hadoop-2.7.2/logs/hadoop-ubuntu-namenode-ip-172-31-31-1.out
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ECDSA key fingerprint is ce:6:11:be:ce:df:51:4b:b7:e0:b0:cf:13:4a:61:5a.
Are you sure you want to continue connecting (yes/no)? yes
localhost: Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
localhost: starting datanode, logging to /home/ubuntu/hadoop-2.7.2/logs/hadoop-ubuntu-datanode-ip-172-31-31-1.out
Starting secondary namenodes [0.0.0.0]
The authenticity of host '0.0.0.0 (0.0.0.0)' can't be established.
ECDSA key fingerprint is ce:6:11:be:ce:df:51:4b:b7:e0:b0:cf:13:4a:61:5a.
Are you sure you want to continue connecting (yes/no)? yes
0.0.0.0: Warning: Permanently added '0.0.0.0' (ECDSA) to the list of known hosts.
0.0.0.0: starting secondarynamenode, logging to /home/ubuntu/hadoop-2.7.2/logs/hadoop-ubuntu-secondarynamenode-ip-172-31-31-1.out
16/04/01 00:54:14 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
ubuntu@ip-172-31-31-1:~/hadoop-2.7.2/sbin$ ./start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /home/ubuntu/hadoop-2.7.2/logs/yarn-ubuntu-resourcemanager-ip-172-31-31-1.out
localhost: starting nodemanager, logging to /home/ubuntu/hadoop-2.7.2/logs/yarn-ubuntu-nodemanager-ip-172-31-31-1.out
ubuntu@ip-172-31-31-1:~/hadoop-2.7.2/sbin$
```

PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

A20360111

Step 18: Check if the mount was successful

The screenshot shows a Mozilla Firefox window with the title "Namenode information - Mozilla Firefox". The address bar indicates the URL is "ec2-52-90-68-182.compute-1.amazonaws.com:50070/dfshealth.html#tab-overview". The main content area has two sections: "Overview" and "Summary".

Overview section (for 'ec2-52-90-68-182.compute-1.amazonaws.com:9000' (active)):

Started:	Fri Apr 01 01:34:59 UTC 2016
Version:	2.7.2, rb165c4fe8a74265c792ce23f546c64604acf0e41
Compiled:	2016-01-26T00:08Z by jenkins from (detached from b165c4f)
Cluster ID:	CID-1b62435a-3c98-4fd3-a188-d97d8c6973ad
Block Pool ID:	BP-324177185-172.31.1-1459474349799

Summary section:

Security is off.
Safemode is off.
1 files and directories, 0 blocks = 1 total filesystem object(s).
Heap Memory used 67.19 MB of 159.5 MB Heap Memory. Max Heap Memory is 889 MB.
Non Heap Memory used 31.09 MB of 32.44 MB Committed Non Heap Memory. Max Non Heap Memory is 214 MB.

Configured Capacity:	9.8 TB
DFS Used:	408 KB (0%)

Step 19: Generate data using gensort

The screenshot shows a terminal window on an Ubuntu system with the prompt "ubuntu@ip-172-31-31-1:~>".

```
45.out
ec2-54-173-95-125.compute-1.amazonaws.com: starting nodemanager, logging to /home/ubuntu/hadoop-2.7.2/logs/yarn-ubuntu-nodemanager-ip-172-31-31-2
239.out
ec2-54-165-65-45.compute-1.amazonaws.com: starting nodemanager, logging to /home/ubuntu/hadoop-2.7.2/logs/yarn-ubuntu-nodemanager-ip-172-31-31-2
49.out
ec2-54-165-7-232.compute-1.amazonaws.com: starting nodemanager, logging to /home/ubuntu/hadoop-2.7.2/logs/yarn-ubuntu-nodemanager-ip-172-31-31-2
54.out
ec2-54-173-128-111.compute-1.amazonaws.com: starting nodemanager, logging to /home/ubuntu/hadoop-2.7.2/logs/yarn-ubuntu-nodemanager-ip-172-31-31-2
-244.out
ec2-54-89-122-163.compute-1.amazonaws.com: starting nodemanager, logging to /home/ubuntu/hadoop-2.7.2/logs/yarn-ubuntu-nodemanager-ip-172-31-31-2
248.out
ec2-52-90-143-48.compute-1.amazonaws.com: starting nodemanager, logging to /home/ubuntu/hadoop-2.7.2/logs/yarn-ubuntu-nodemanager-ip-172-31-31-2
53.out
ec2-54-85-72-182.compute-1.amazonaws.com: starting nodemanager, logging to /home/ubuntu/hadoop-2.7.2/logs/yarn-ubuntu-nodemanager-ip-172-31-31-2
43.out
ec2-52-90-141-123.compute-1.amazonaws.com: starting nodemanager, logging to /home/ubuntu/hadoop-2.7.2/logs/yarn-ubuntu-nodemanager-ip-172-31-31-2
242.out
ec2-54-173-132-3.compute-1.amazonaws.com: starting nodemanager, logging to /home/ubuntu/hadoop-2.7.2/logs/yarn-ubuntu-nodemanager-ip-172-31-31-2
52.out
ec2-54-173-231-189.compute-1.amazonaws.com: starting nodemanager, logging to /home/ubuntu/hadoop-2.7.2/logs/yarn-ubuntu-nodemanager-ip-172-31-31-2
-251.out
ec2-54-89-112-24.compute-1.amazonaws.com: starting nodemanager, logging to /home/ubuntu/hadoop-2.7.2/logs/yarn-ubuntu-nodemanager-ip-172-31-31-2
40.out
ec2-54-85-106-15.compute-1.amazonaws.com: starting nodemanager, logging to /home/ubuntu/hadoop-2.7.2/logs/yarn-ubuntu-nodemanager-ip-172-31-31-2
50.out
ec2-54-173-68-102.compute-1.amazonaws.com: starting nodemanager, logging to /home/ubuntu/hadoop-2.7.2/logs/yarn-ubuntu-nodemanager-ip-172-31-31-2
241.out
ec2-52-90-68-182.compute-1.amazonaws.com: starting nodemanager, logging to /home/ubuntu/hadoop-2.7.2/logs/yarn-ubuntu-nodemanager-ip-172-31-31-1
.out
ubuntu@ip-172-31-31-1:~/hadoop-2.7.2/sbin$ jps
15902 DataNode
16746 Jps
16445 NodeManager
16266 ResourceManager
16110 SecondaryNameNode
15698 NameNode
ubuntu@ip-172-31-31-1:~/hadoop-2.7.2/sbin$ cd ..
ubuntu@ip-172-31-31-1:~/hadoop-2.7.2$ cd ..
ubuntu@ip-172-31-31-1:~ ls
gensort hadoop-2.7.2.tar.gz hadoop.pem valsort
ubuntu@ip-172-31-31-1:~ ./gensort -a 1000000000 /data/dataset
```

PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

A20360111

Step 20: Run the HadoopTerasort jar

```
ubuntu@ip-172-31-31-1:~/hadoop-2.7.2
drwxrwxr-x 5 ubuntu ubuntu 4096 Apr  1 01:35 dfs
-rwxrwxr-x 1 ubuntu ubuntu 100000000000 Apr  1 02:06 dataset
drwxr-xr-x 5 ubuntu ubuntu 4096 Apr  1 02:10 nn-local-dir
ubuntu@ip-172-31-31-1:/data$ cd /home/ubuntu/
ubuntu@ip-172-31-31-1:$ cd hadoop-2.7.2/
ubuntu@ip-172-31-31-1:~/hadoop-2.7.2$ bin/hadoop fs -mkdir -p /user/ubuntu/gutenberg
16/04/01 02:14:50 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
ubuntu@ip-172-31-31-1:~/hadoop-2.7.2$ bin/hadoop dfs -copyFromLocal /data/dataset /user/ubuntu/gutenberg
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
16/04/01 02:15:07 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
ubuntu@ip-172-31-31-1:~/hadoop-2.7.2$ bin/hadoop dfs -ls /user/ubuntu/gutenberg
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
16/04/01 02:47:00 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 1 items
-rw-r--r-- 1 ubuntu supergroup 100000000000 2016-04-01 02:42 /user/ubuntu/gutenberg/dataset
ubuntu@ip-172-31-31-1:~/hadoop-2.7.2$ bin/hadoop jar /home/ubuntu/HadoopTerasort.jar
16/04/01 02:55:35 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/04/01 02:55:36 INFO client.RMProxy: Connecting to ResourceManager at ec2-52-90-68-182.compute-1.amazonaws.com/172.31.31.1:9050
16/04/01 02:55:36 INFO client.RMProxy: Connecting to ResourceManager at ec2-52-90-68-182.compute-1.amazonaws.com/172.31.31.1:9050
16/04/01 02:55:37 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
16/04/01 02:55:37 INFO mapred.FileInputFormat: Total input paths to process : 1
16/04/01 02:55:37 INFO net.NetworkTopology: Adding a new node: /default-rack/172.31.31.1:50010
16/04/01 02:55:38 INFO mapreduce.JobSubmitter: number of splits:745
16/04/01 02:55:38 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1459474565538_0001
16/04/01 02:55:39 INFO impl.YarnClientImpl: Submitted application application_1459474565538_0001
16/04/01 02:55:39 INFO mapreduce.Job: The url to track the job: http://ec2-52-90-68-182.compute-1.amazonaws.com:9006/proxy/application_1459474565538_0001/
16/04/01 02:55:39 INFO mapreduce.Job: Running job: job_1459474565538_0001
16/04/01 02:55:47 INFO mapreduce.Job: Job job_1459474565538_0001 running in uber mode : false
16/04/01 02:55:47 INFO mapreduce.Job: map 0% reduce 0%
```

Output

All Applications - Mozilla Firefox

ec2-52-90-68-182.compute-1.amazonaws.com:9006/cluster

Logged in as: 

All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	Vcores Used	Vcores Total	Vcores Reserved	Active Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Reboot Nodes
1	0	1	0	39	40 GB	136 GB	0 B	39	136	0	17	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[MEMORY]	<memory:1024, vCores:1>	<memory:8192, vCores:8>

Show 20 entries

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Progress	Tracking UI	Blacklist Nodes
application_1459474565538_0001	ubuntu	Terasort	MAPREDUCE	default	Thu Mar 31 21:55:39 -0500 2016	N/A	RUNNING	UNDEFINED		ApplicationMaster	0

Showing 1 to 1 of 1 entries

First Previous 1 Next Last

PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

A20360111

The screenshot shows a Linux desktop environment with a dark theme. On the left is a vertical dock containing icons for various applications: GIMP, Nautilus, Firefox, Java IDE, Eclipse, and others. The main window is a terminal window titled "ubuntu@ip-172-31-31-1: ~/hadoop-2.7.2". It displays a log of Hadoop MapReduce job progress from March 1, 2016, at 03:18:23 to 03:31:05. The log entries show the status of map and reduce tasks, with percentages indicating completion. Below the terminal is a Mozilla Firefox browser window titled "All Applications - Mozilla Firefox". The address bar shows the URL "ec2-52-90-68-182.compute-1.amazonaws.com:9006/cluster". The page content is titled "All Applications" and displays two tables: "Cluster Metrics" and "Scheduler Metrics". The "Cluster Metrics" table provides a summary of cluster resources. The "Scheduler Metrics" table lists application details, including ID, User, Name, Application Type, Queue, Start Time, Finish Time, State, Final Status, Progress, Tracking UI, and Blacklist Nodes. One entry is shown: "application_1459474565538_0001" by user "ubuntu" named "Terasort" with type "MAPREDUCE" in the default queue, finished successfully on Thursday, March 31, 2016, at 23:56:29.

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Progress	Tracking UI	Blacklist Nodes
application_1459474565538_0001	ubuntu	Terasort	MAPREDUCE	default	Thu Mar 31 21:55:39 -0500 2016	Thu Mar 31 23:56:29 -0500 2016	FINISHED	SUCCEEDED		History	N/A

PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

A20360111

```
ubuntu@ip-172-31-31-1: ~/hadoop-2.7.2
Total time spent by all maps in occupied slots (ms)=198300071
Total time spent by all reduces in occupied slots (ms)=6947146
Total time spent by all map tasks (ms)=198300071
Total time spent by all reduce tasks (ms)=6947146
Total vcore-milliseconds taken by all map tasks=198300071
Total vcore-milliseconds taken by all reduce tasks=6947146
Total megabyte-milliseconds taken by all map tasks=20305927704
Total megabyte-milliseconds taken by all reduce tasks=7113877504
Map-Reduce Framework
  Map Input records=1000000000
  Map output records=1000000000
  Map output bytes=100000000000
  Map output materialized bytes=102000004470
  Input split bytes=99830
  Combine input records=1000000000
  Combine output records=1000000000
  Reduce input groups=1000000000
  Reduce shuffle bytes=102000004470
  Reduce input records=1000000000
  Reduce output records=1000000000
  Spilled Records=5026843532
  Shuffled Maps =745
  Failed Shuffles=0
  Merged Map outputs=745
  GC time elapsed (ms)=327372
  CPU time spent (ms)=14326750
  Physical memory (bytes) snapshot=193963106304
  Virtual memory (bytes) snapshot=618489397248
  Total committed heap usage (bytes)=135624916992
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=100003047424
File Output Format Counters
  Bytes Written=100000000000
ubuntu@ip-172-31-31-1:~/hadoop-2.7.2$
```

Valsort

```
ubuntu@ip-172-31-31-1: ~
ubuntu@ip-172-31-31-1:~$ ./valsrt /data/result
Records: 1000000000
Checksum: 1cd601caeae2f059
Duplicate keys: 0
SUCCESS - all records are in order
ubuntu@ip-172-31-31-1:~$ ubuntu@ip-172-31-31-1:~$ ubuntu@ip-172-31-31-1:~$ *ubuntu@ip-172-31-31-1:~$ cd /data
ubuntu@ip-172-31-31-1:~/data$ head -10 /data/result
!4:ABv          0000000000000000000000000000000000177E829  EEEE33334444111122288833334444666633332222DDDEEE
"0!uve         00000000000000000000000000000000001228D4  7777888800002224440DDDDDEEE00000000CC7777DDDD
%!$sU(        00000000000000000000000000000000002E6C821C  22223333777444455511119999CCC4444EEEEFF11115555
&5rx|X        000000000000000000000000000000000399BC288  5555CCCCBBB99999999DD11100001111EEE7777DDDD9999
'ic%So         00000000000000000000000000000000031F06B7D  EEEEBBBAAA88880DD0DD0777722244411116664444AAA
*0G1Io         00000000000000000000000000385E85A1  1111AAA9999CCCBBB111199991111333399991111AAAA6666
,(GHT          00000000000000000000000000000000020172DC  1111CCC1111DDDCCCCEE9999CCC8888CCCF5555555
0*Vvm3         0000000000000000000000000000026061578  DDD7777AAAEEEEE6666AAA2222CCC55555522229999
2C>Bd          00000000000000000000000000000026C79E66  444400001111CCC6666BBB5557776666CCC2222AAAABB
PMd32=         000000000000000000000000000003440C1  FFFFFEEE6666CCCCBBB99993335555DDDDDD0777788886666
ubuntu@ip-172-31-31-1:~/data$ tail -10f /data/result
----#iay1X      000000000000000000000000025035EDF  6666AAA5555999977700002223338888FFFF999922220000
----+@()@       0000000000000000000000000085426F4  77773333551111110000CCCC55559999AAA7777DDDDDD
----,R^?n      0000000000000000000000000001034E347  1111111999900011118888AAA5555444EEE99993338888
----,Ey^-)     000000000000000000000000001f0E66B  CCC6666DDDD22220DD11118889999EEEeeeeEEBBB4444
----4!KA7x     00000000000000000000000001F1A1E26  EEE777711117777BBB1111EEE88884444DDDDDDDEEEBBB
----8I!l@       0000000000000000000000000001F05932F  11119999BBB444477700011114444CCCCAA6666DDDD0000
----<I'5>F    000000000000000000000000000008C82293  888833388811116666999988885555888888882228888CCCC
----G-)m^)     00000000000000000000000000013397F73  DDDDFFFFBBBBCCCCFFFF44446666AAA111133333333AAAAACCC
----c+1&P     0000000000000000000000000000074BD64   88880000555500000DD22227777AAA000033332222AAAADD
----hb&5X*     00000000000000000000000032C0E06B  7777BBBBBBB9999EEEAAAAAAA0000CCCCDDDD4444BBBB4444
[0] 0:tall*
```

PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

SHARED MEMORY SORT

A20360111

This Shared Memory Sort is implemented in java. Multithreading is used for its implementation. ExecutorsService class for is used for multithreading.

ExecutorService exec = Executors.newFixedThreadPool(noOfThreads);

The Shared Memory Sort is tested for 1, 2, 4, and 8 threads. The implementation is designed for single node with 10GB dataset. I ran the code simultaneously on 5 instance .

The screenshot shows the AWS EC2 Instances page. On the left, there's a sidebar with navigation links for EC2 Dashboard, Events, Tags, Reports, Limits, Instances, Images, Elastic Block Store, Network & Security, and Load Balancing. The Instances section is currently selected. The main pane displays a table of instances with columns: Name, Instance ID, Instance Type, Availability Zone, Instance State, Status Checks, Alarm Status, Public DNS, and Public IP. There are five instances listed:

Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Status	Public DNS	Public IP
Shared 6	i-026b6f86	c3.large	us-east-1d	running	2/2 checks ...	None	ec2-54-152-158-115.co...	54.15
Shared 8	i-58080cdc	c3.large	us-east-1d	running	2/2 checks ...	None	ec2-52-90-1-247.comp...	52.90
Shared 2	i-650206e1	c3.large	us-east-1d	running	2/2 checks ...	None	ec2-54-173-232-8.comp...	54.17
	i-04171300	c3.large	us-east-1d	running	2/2 checks ...	None	ec2-52-90-163-45.comp...	52.90
	i-91797d15	c3.large	us-east-1d	running	2/2 checks ...	None	ec2-54-86-6-201.comp...	54.86

Below the table, the details for the first instance, Shared 6, are expanded. The 'Description' tab is selected, showing the following information:

Attribute	Value
Instance ID	i-026b6f86
Instance state	running
Instance type	c3.large
Private DNS	ip-172-31-10-44.ec2.internal
Private IPs	172.31.10.44
Secondary private IPs	
VPC ID	vpc-00b3f564
Subnet ID	subnet-9f43f6e9
Network Interfaces	eth0
Source/dest. check	True

On the right side of the expanded view, there are additional details:

Attribute	Value
Public DNS	ec2-54-152-158-115.compute-1.amazonaws.com
Public IP	54.152.158.115
Elastic IP	-
Availability zone	us-east-1d
Security groups	launch-wizard-80, view rules
Scheduled events	No scheduled events
AMI ID	ubuntu-trusty-14.04-amd64-server-20160114.5 (ami-fce3c696)
Platform	-
IAM role	-
Key pair name	hadoop
Owner	511518029807

Compiling the code and running the code with 6 threads.

PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

A20360111

```
ubuntu@ip-172-31-10-44: ~
ubuntu@ip-172-31-10-44:~$ sudo mke2fs -F -t ext4 /dev/xvdd
mke2fs 1.42.9 (4-Feb-2014)
Filesystem label=
OS type: Linux
Block size=4096 (log=2)
Fragment size=4096 (log=2)
Stride=0 blocks, Stripe width=0 blocks
13107200 inodes, 52428800 blocks
2621440 blocks (5.00%) reserved for the super user
First data block=0
Maximum filesystem blocks=4294967296
1600 block groups
32768 blocks per group, 32768 fragments per group
8192 inodes per group
Superblock backups stored on blocks:
      32768, 98304, 163840, 229376, 294912, 819200, 884736, 1605632, 2654208,
      4096000, 7962624, 11239424, 20480000, 23887872

Allocating group tables: done
Writing inode tables: done
Creating journal (32768 blocks): done
Writing superblocks and filesystem accounting information: done

ubuntu@ip-172-31-10-44:~$ sudo mkdir /data
ubuntu@ip-172-31-10-44:~$ sudo mount /dev/xvdd /data
ubuntu@ip-172-31-10-44:~$ sudo chmod 777 /data
ubuntu@ip-172-31-10-44:~$ cd /data
Snap
ubuntu@ip-172-31-10-44:/data$ sudo mkdir output
ubuntu@ip-172-31-10-44:/data$ cd ..
ubuntu@ip-172-31-10-44:$ cd /home/ubuntu
ubuntu@ip-172-31-10-44:~$ javac SharedMemory.java
```

Valsort

```
ubuntu@ip-172-31-10-44: ~
Records: 10000000
Checksum: 2faf0ab746e89a8
Duplicate keys: 0
SUCCESS - all records are in order
ubuntu@ip-172-31-10-44:~$ clear

ubuntu@ip-172-31-10-44:~$ ./valsrt /data/output/output
Records: 10000000
Checksum: 2faf0ab746e89a8
Duplicate keys: 0
SUCCESS - all records are in order
ubuntu@ip-172-31-10-44:~$
```

PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

A20360111

```
ubuntu@ip-172-31-10-44: ~
  http://www.ubuntu.com/business/services/cloud
0 packages can be updated.
0 updates are security updates.

Last login: Wed Mar 30 09:06:32 2016 from 208-59-149-170.c3-0.mcm-ubr1.chi-mcm.i
l.cable.rcn.com
ubuntu@ip-172-31-10-44:~$ head -10 /data/output/output
'Olive 0000000000000000000000000000122804 77778880000222244440DDDDDDDEEE00
000000CCCC7777DDDD
  PMD32= 000000000000000000000000000003440CC1 FFFFFEEE666CCCCBBB999933335555DD
DDDDDD77778886666
  ^3C0], 00000000000000000000000000000158C5C5 5555AAA9999EEEE888822229999CCCCD
DD66655554442222
  I&S3[]] 000000000000000000000000000002145D78 8888BBBBDDDD1111CCC5556666BBBB11
11EEEEDDDD22229999
  !=#U,.9 000000000000000000000000000001907E3 33332222FFFBBBBB0000FFFFAAAA666655
553333DDDD3333CCCC
  !Of[ITo 000000000000000000000000000003CAA4B 9999FFFF55553337777CCCC4444BBBB77
77EEEEBBBBDDDD4444
  !f6Suy2 000000000000000000000000000003ABFD84 EEEESSSSSS5556666AAA5555BBBBDDDD00
0011116666000000DD
  #SNIq. 000000000000000000000000000003B36FB9 1111000033334444111166666666AAAAAA
AA00001111CCCCEEE
  #`^cL` 000000000000000000000000000002EDC5C8 8888AAA11114444FFFF77773333EEEE44
440000FFFF9999999
  $`-'Q) 000000000000000000000000000005F1265D CCC6666EEE222200000DDDAAAA888866
66BBBBB00006666AAA
ubuntu@ip-172-31-10-44:~$ tail -10 /data/output/output
~~u2k#=U 000000000000000000000000000002C06745 99991110DD2221110000FFFEFFFF33337777CCCC2222
~~v/0&nn 000000000000000000000000000004709701 CCCCB888333FFF00000000009999111FFFF7774446666
~~yKoL;qE 000000000000000000000000000002048B4F CCCC11114444888822226666BBBB888855557777EEEBBBBB0000
Place~~yK^H_il 00000000000000000000000000000463D004 44440000FFF333399944447777DDDFFFAAA11118880DD00
Key~~yL;C'XE 000000000000000000000000000005B0D211 2222EEE333300002221111CCCCFFF55557774444BBBB6666
~~zBa_Tt 000000000000000000000000000000097F9F4F BBBBCCCC666655559999FFF8888AAA11116666AAAABB0000
Netw~~ze0^fEq 0000000000000000000000000000000106130 4444CCCCBBBB999922288885558888CCCCFFF00001111111
~~)GxJWH 00000000000000000000000000000000CA1345 777711118888AAAAAA222111BBBB0000222BBBBCCC2222
~~)P;jg0 000000000000000000000000000000040DA3E4 4444FFF444466663333EEE88888888D0DDEEEE44442222DD00
~~)KU!K<p 000000000000000000000000000005E4A0AA 0000666655551111BBBB88889999AAA5555000033335557777
```

Result

```
Key ubuntu@ip-172-31-10-44:~$ ./gensort -a 1000000000 /data/input
ubuntu@ip-172-31-10-44:~$ cd /data
Networks ubuntu@ip-172-31-10-44:~/data$ sudo chown -R ubuntu /data/output
ubuntu@ip-172-31-10-44:~/data$ cd /home/ubuntu
ubuntu@ip-172-31-10-44:~$ java SharedMemory 6
Time taken to sort 10GB : 1961841for threads6
ubuntu@ip-172-31-10-44:~$
```

```
ubuntu@ip-172-31-5-202: ~
ubuntu@ip-172-31-5-202:~$ java SharedMemory 4
Time taken to sort 10GB : 1966064for threads4
ubuntu@ip-172-31-5-202:~$
```

```
ubuntu@ip-172-31-4-150: ~
ubuntu@ip-172-31-4-150:~$ java SharedMemory 8
Time taken to sort 10GB : 2021289for threads8
ubuntu@ip-172-31-4-150:~$ ./valsrt /data/output/output
Records: 100000000
Checksum: 2faf0ab746e89a8
```

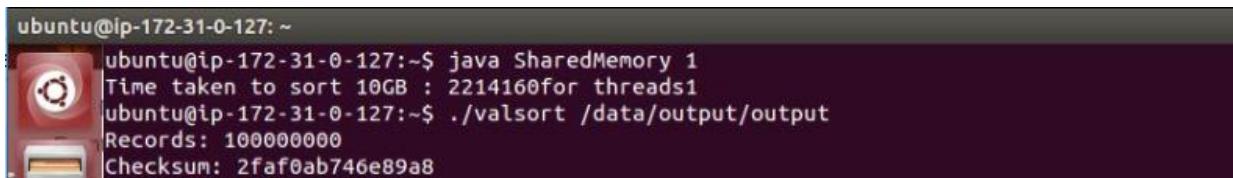
```
ubuntu@ip-172-31-8-133: ~
ubuntu@ip-172-31-8-133:~$ java SharedMemory 2
Time taken to sort 10GB : 1969362for threads2
ubuntu@ip-172-31-8-133:~$
```

PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

A20360111



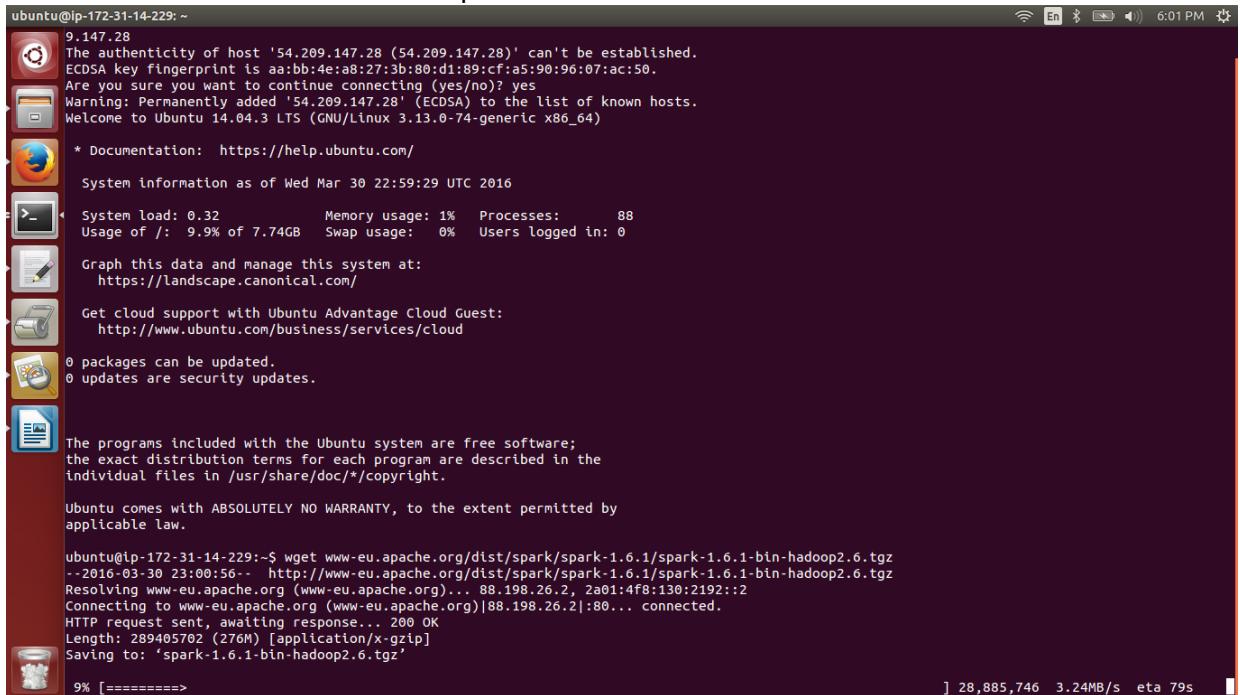
```
ubuntu@ip-172-31-0-127: ~
ubuntu@ip-172-31-0-127:~$ java SharedMemory 1
Time taken to sort 10GB : 2214160for threads1
ubuntu@ip-172-31-0-127:~$ ./valsrt /data/output/output
Records: 100000000
Checksum: 2faf0ab746e89a8
```

SPARK

This Spark code is written in Scala. The inbuilt function `sortByKey` is used to sort the input file generated using `gensort`.

Installation Single Node

1. Create an instance and download spark



```
9.147.28
The authenticity of host '54.209.147.28 (54.209.147.28)' can't be established.
ECDSA key fingerprint is aa:bb:4e:a8:27:3b:80:d1:89:cf:a5:90:96:07:ac:50.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added '54.209.147.28' (ECDSA) to the list of known hosts.
Welcome to Ubuntu 14.04.3 LTS (GNU/Linux 3.13.0-74-generic x86_64)

 * Documentation:  https://help.ubuntu.com/
System information as of Wed Mar 30 22:59:29 UTC 2016
System load: 0.32      Memory usage: 1%   Processes:     88
Usage of /:  9.9% of 7.74GB  Swap usage:  0%   Users logged in: 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
0 packages can be updated,
0 updates are security updates.

The programs included with the Ubuntu system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*copyright.

Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by
applicable law.

ubuntu@ip-172-31-14-229:~$ wget www-eu.apache.org/dist/spark/spark-1.6.1/spark-1.6.1-bin-hadoop2.6.tgz
--2016-03-30 23:00:56--  http://www-eu.apache.org/dist/spark/spark-1.6.1/spark-1.6.1-bin-hadoop2.6.tgz
Resolving www-eu.apache.org (www-eu.apache.org)... 88.198.26.2, 2a01:4f8:130:2192::2
Connecting to www-eu.apache.org (www-eu.apache.org)|88.198.26.2|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 289405702 (276M) [application/x-gzip]
Saving to: 'spark-1.6.1-bin-hadoop2.6.tgz'

9% [=====]> 28,885,746  3.24MB/s  eta 79s
```

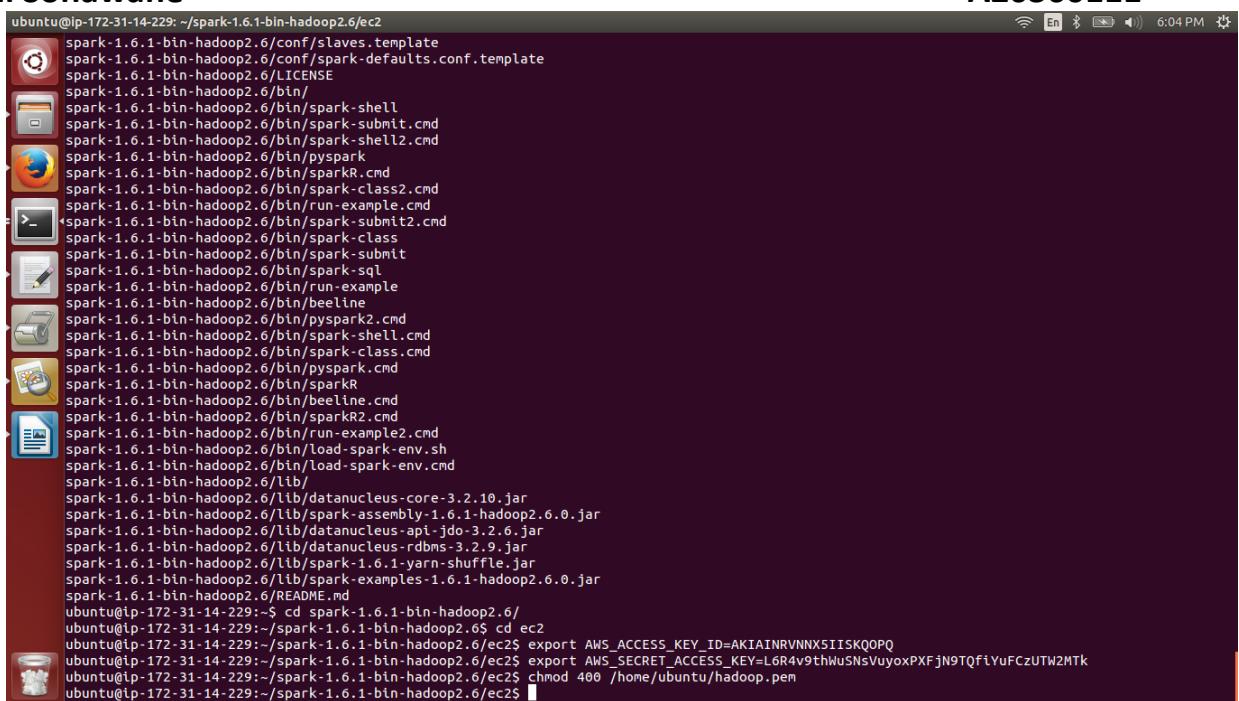
2. Unzip the downloaded file and export ACCESS KEY and SECRET ACCESS KEY

PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

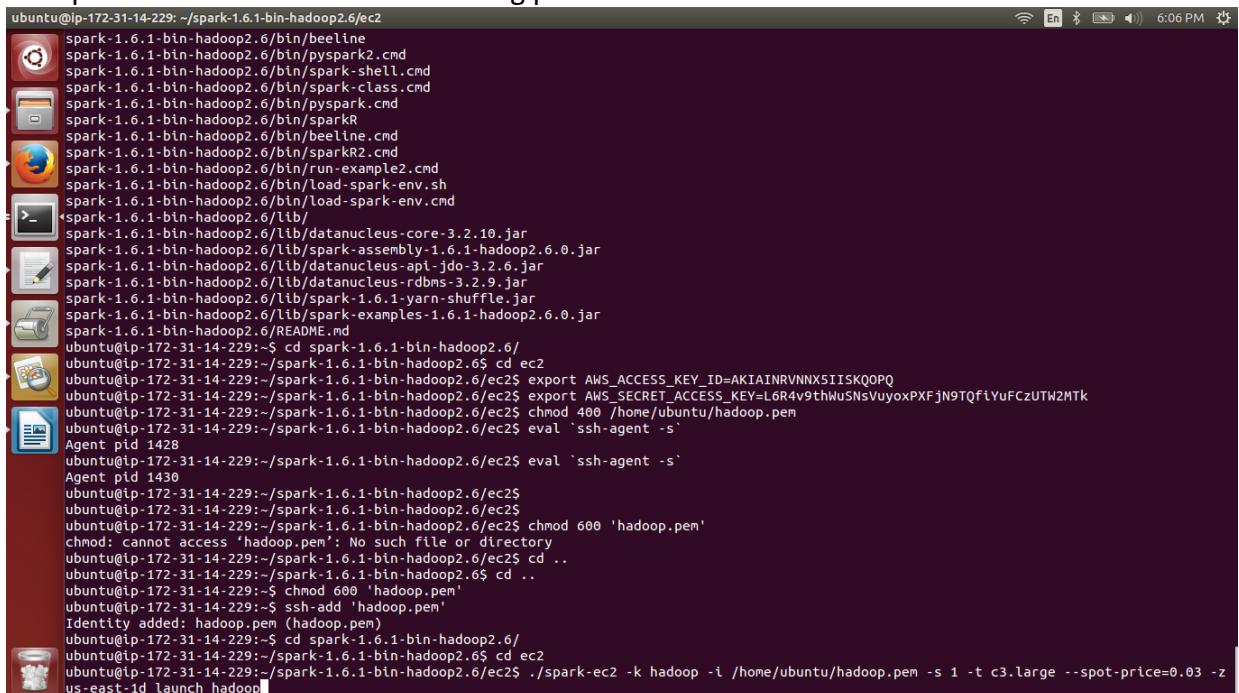
Snehal Sonawane

A20360111



```
ubuntu@ip-172-31-14-229: ~/spark-1.6.1-bin-hadoop2.6/ec2
spark-1.6.1-bin-hadoop2.6/conf/slaves.template
spark-1.6.1-bin-hadoop2.6/conf/spark-defaults.conf.template
spark-1.6.1-bin-hadoop2.6/LICENSE
spark-1.6.1-bin-hadoop2.6/bin/
spark-1.6.1-bin-hadoop2.6/bin/spark-shell
spark-1.6.1-bin-hadoop2.6/bin/spark-submit.cmd
spark-1.6.1-bin-hadoop2.6/bin/spark-shell2.cmd
spark-1.6.1-bin-hadoop2.6/bin/pyspark
spark-1.6.1-bin-hadoop2.6/bin/sparkR.cmd
spark-1.6.1-bin-hadoop2.6/bin/spark-class2.cmd
spark-1.6.1-bin-hadoop2.6/bin/run-example.cmd
spark-1.6.1-bin-hadoop2.6/bin/spark-submit2.cmd
spark-1.6.1-bin-hadoop2.6/bin/spark-class
spark-1.6.1-bin-hadoop2.6/bin/spark-submit
spark-1.6.1-bin-hadoop2.6/bin/spark-sql
spark-1.6.1-bin-hadoop2.6/bin/run-example
spark-1.6.1-bin-hadoop2.6/bin/beeline
spark-1.6.1-bin-hadoop2.6/bin/pyspark2.cmd
spark-1.6.1-bin-hadoop2.6/bin/spark-shell.cmd
spark-1.6.1-bin-hadoop2.6/bin/spark-class.cmd
spark-1.6.1-bin-hadoop2.6/bin/pyspark.cmd
spark-1.6.1-bin-hadoop2.6/bin/sparkR
spark-1.6.1-bin-hadoop2.6/bin/beeline.cmd
spark-1.6.1-bin-hadoop2.6/bin/sparkR2.cmd
spark-1.6.1-bin-hadoop2.6/bin/run-example2.cmd
spark-1.6.1-bin-hadoop2.6/bin/load-spark-env.sh
spark-1.6.1-bin-hadoop2.6/bin/load-spark-env.cmd
spark-1.6.1-bin-hadoop2.6/lib/
spark-1.6.1-bin-hadoop2.6/lib/datanucleus-core-3.2.10.jar
spark-1.6.1-bin-hadoop2.6/lib/spark-assembly-1.6.1-hadoop2.6.0.jar
spark-1.6.1-bin-hadoop2.6/lib/datanucleus-api-jdo-3.2.6.jar
spark-1.6.1-bin-hadoop2.6/lib/datanucleus-rdbms-3.2.9.jar
spark-1.6.1-bin-hadoop2.6/lib/spark-1.6.1-yarn-shuffle.jar
spark-1.6.1-bin-hadoop2.6/lib/spark-examples-1.6.1-hadoop2.6.0.jar
spark-1.6.1-bin-hadoop2.6/README.md
ubuntu@ip-172-31-14-229:~$ cd spark-1.6.1-bin-hadoop2.6/
ubuntu@ip-172-31-14-229:~/spark-1.6.1-bin-hadoop2.6/ec2$ export AWS_ACCESS_KEY_ID=AKIAINRVNNX5IISKQ0PQ
ubuntu@ip-172-31-14-229:~/spark-1.6.1-bin-hadoop2.6/ec2$ export AWS_SECRET_ACCESS_KEY=L6R4v9thWuSNsVuyoxPXFjN9TQfIYuFCzUTW2MTk
ubuntu@ip-172-31-14-229:~/spark-1.6.1-bin-hadoop2.6/ec2$ chmod 400 /home/ubuntu/hadoop.pem
ubuntu@ip-172-31-14-229:~/spark-1.6.1-bin-hadoop2.6/ec2$
```

3. Set up SSH certificates to avoid entering password each time:



```
ubuntu@ip-172-31-14-229:~/spark-1.6.1-bin-hadoop2.6/ec2
spark-1.6.1-bin-hadoop2.6/bin/beeline
spark-1.6.1-bin-hadoop2.6/bin/pyspark2.cmd
spark-1.6.1-bin-hadoop2.6/bin/spark-shell.cmd
spark-1.6.1-bin-hadoop2.6/bin/spark-class.cmd
spark-1.6.1-bin-hadoop2.6/bin/pyspark.cmd
spark-1.6.1-bin-hadoop2.6/bin/sparkR
spark-1.6.1-bin-hadoop2.6/bin/beeline.cmd
spark-1.6.1-bin-hadoop2.6/bin/sparkR2.cmd
spark-1.6.1-bin-hadoop2.6/bin/run-example2.cmd
spark-1.6.1-bin-hadoop2.6/bin/load-spark-env.sh
spark-1.6.1-bin-hadoop2.6/bin/load-spark-env.cmd
spark-1.6.1-bin-hadoop2.6/lib/datanucleus-core-3.2.10.jar
spark-1.6.1-bin-hadoop2.6/lib/spark-assembly-1.6.1-hadoop2.6.0.jar
spark-1.6.1-bin-hadoop2.6/lib/datanucleus-api-jdo-3.2.6.jar
spark-1.6.1-bin-hadoop2.6/lib/datanucleus-rdbms-3.2.9.jar
spark-1.6.1-bin-hadoop2.6/lib/spark-1.6.1-yarn-shuffle.jar
spark-1.6.1-bin-hadoop2.6/lib/spark-examples-1.6.1-hadoop2.6.0.jar
spark-1.6.1-bin-hadoop2.6/README.md
ubuntu@ip-172-31-14-229:~$ cd spark-1.6.1-bin-hadoop2.6/
ubuntu@ip-172-31-14-229:~/spark-1.6.1-bin-hadoop2.6/ec2$ cd ec2
ubuntu@ip-172-31-14-229:~/spark-1.6.1-bin-hadoop2.6/ec2$ export AWS_ACCESS_KEY_ID=AKIAINRVNNX5IISKQ0PQ
ubuntu@ip-172-31-14-229:~/spark-1.6.1-bin-hadoop2.6/ec2$ export AWS_SECRET_ACCESS_KEY=L6R4v9thWuSNsVuyoxPXFjN9TQfIYuFCzUTW2MTk
ubuntu@ip-172-31-14-229:~/spark-1.6.1-bin-hadoop2.6/ec2$ chmod 400 /home/ubuntu/hadoop.pem
ubuntu@ip-172-31-14-229:~/spark-1.6.1-bin-hadoop2.6/ec2$ eval `ssh-agent -s`
Agent pid 1428
ubuntu@ip-172-31-14-229:~/spark-1.6.1-bin-hadoop2.6/ec2$ Agent pid 1430
ubuntu@ip-172-31-14-229:~/spark-1.6.1-bin-hadoop2.6/ec2$ 
ubuntu@ip-172-31-14-229:~/spark-1.6.1-bin-hadoop2.6/ec2$ 
ubuntu@ip-172-31-14-229:~/spark-1.6.1-bin-hadoop2.6/ec2$ chmod 600 'hadoop.pem'
chmod: cannot access 'hadoop.pem': No such file or directory
ubuntu@ip-172-31-14-229:~/spark-1.6.1-bin-hadoop2.6/ec2$ cd ..
ubuntu@ip-172-31-14-229:~/spark-1.6.1-bin-hadoop2.6$ cd ..
ubuntu@ip-172-31-14-229:~/spark-1.6.1-bin-hadoop2.6$ chmod 600 'hadoop.pem'
ubuntu@ip-172-31-14-229:~/spark-1.6.1-bin-hadoop2.6$ ssh-add 'hadoop.pem'
Identity added: hadoop.pem (hadoop.pem)
ubuntu@ip-172-31-14-229:~/spark-1.6.1-bin-hadoop2.6$ cd spark-1.6.1-bin-hadoop2.6
ubuntu@ip-172-31-14-229:~/spark-1.6.1-bin-hadoop2.6$ cd ec2
ubuntu@ip-172-31-14-229:~/spark-1.6.1-bin-hadoop2.6$ ./spark-ec2 -k hadoop -i /home/ubuntu/hadoop.pem -s 1 -t c3.large --spot-price=0.03 -z us-east-1d launch hadoop
```

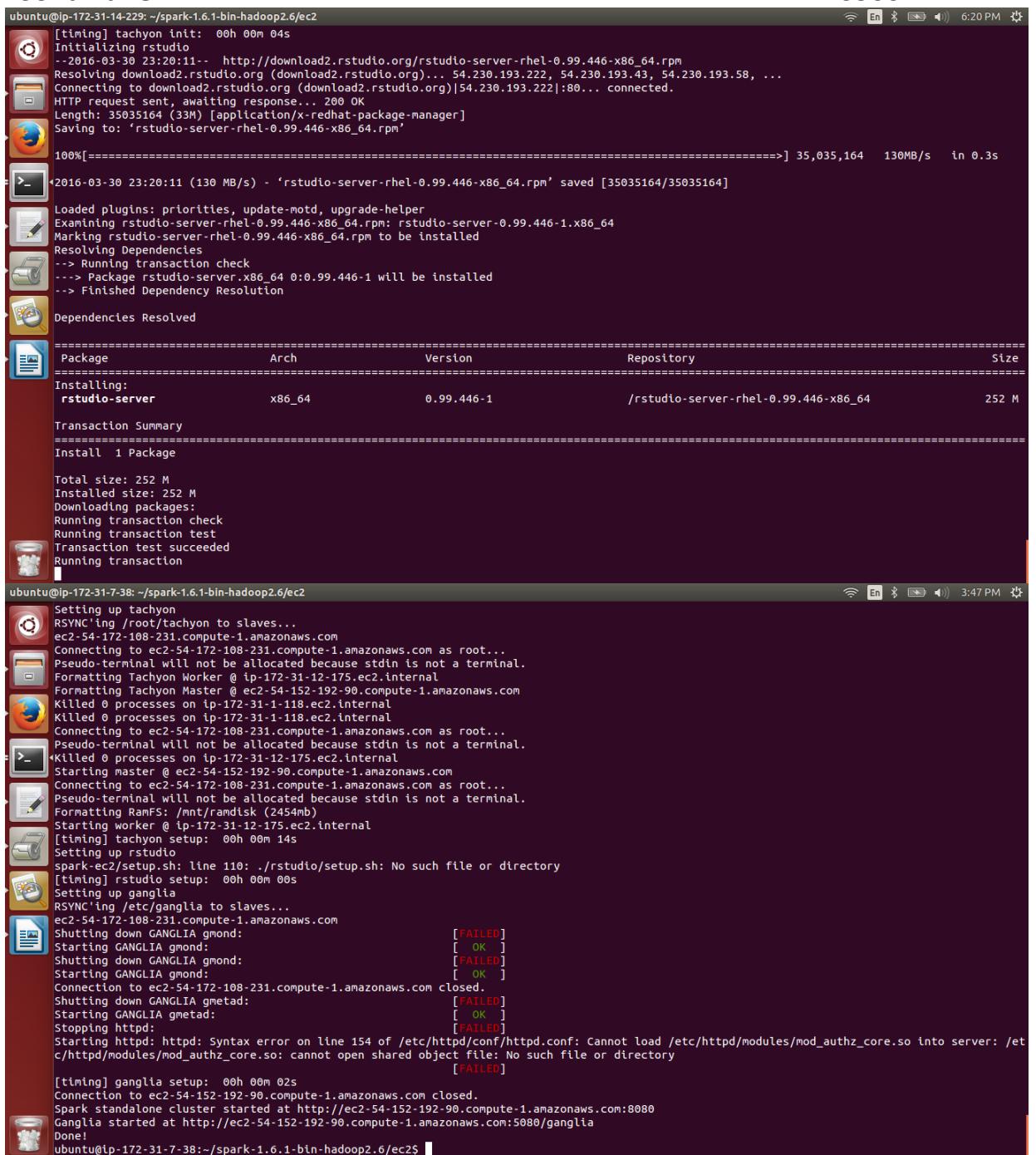
4. Launch Spark

PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

A20360111



```

ubuntu@ip-172-31-14-229: ~/spark-1.6.1-bin-hadoop2.6/ec2
[timing] tachyon init: 00h 00m 04s
Initialzing rstdio
--2016-03-30 23:20:11- http://download2.rstudio.org/rstudio-server-rhel-0.99.446-x86_64.rpm
Resolving download2.rstudio.org (download2.rstudio.org)... 54.230.193.222, 54.230.193.43, 54.230.193.58, ...
Connecting to download2.rstudio.org (download2.rstudio.org)|54.230.193.222|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 35035164 (33M) [application/x-redhat-package-manager]
Saving to: 'rstudio-server-rhel-0.99.446-x86_64.rpm'

100%[=====] 35,035,164 130MB/s in 0.3s
[2016-03-30 23:20:11 (130 MB/s) - 'rstudio-server-rhel-0.99.446-x86_64.rpm' saved [35035164/35035164]

Loaded plugins: priorities, update-motd, upgrade-helper
Examining rstudion-server-rhel-0.99.446-x86_64.rpm: rstudion-server-0.99.446-1.x86_64
Marking rstudion-server-rhel-0.99.446-x86_64.rpm to be installed
Resolving Dependencies
--> Running transaction check
--> Package rstudion-server.x86_64 0:0.99.446-1 will be installed
--> Finished Dependency Resolution
Dependencies Resolved

=====
Package          Arch      Version           Repository      Size
=====
Installing:
rstudio-server   x86_64   0.99.446-1       /rstudio-server-rhel-0.99.446-x86_64  252 M

Transaction Summary
=====
Install 1 Package

Total size: 252 M
Installed size: 252 M
Downloading packages:
Running transaction check
Running transaction test
Transaction test succeeded
Running transaction
[2016-03-30 23:20:11] Transaction complete

ubuntu@ip-172-31-7-38: ~/spark-1.6.1-bin-hadoop2.6/ec2
Setting up tachyon
RSYNC'ing /root/tachyon to slaves...
ec2-54-172-108-231.compute-1.amazonaws.com
Connecting to ec2-54-172-108-231.compute-1.amazonaws.com as root...
Pseudo-terminal will not be allocated because stdin is not a terminal.
Formatting Tachyon Worker @ ip-172-31-12-175.ec2.internal
Formatting Tachyon Master @ ec2-54-152-192-90.compute-1.amazonaws.com
Killed 0 processes on ip-172-31-1-118.ec2.internal
Killed 0 processes on ip-172-31-1-118.ec2.internal
Connecting to ec2-54-172-108-231.compute-1.amazonaws.com as root...
Pseudo-terminal will not be allocated because stdin is not a terminal.
*killed 0 processes on ip-172-31-12-175.ec2.internal
Starting master @ ec2-54-152-192-90.compute-1.amazonaws.com
Connecting to ec2-54-172-108-231.compute-1.amazonaws.com as root...
Pseudo-terminal will not be allocated because stdin is not a terminal.
Formatting RamFS: /mnt/randisk (2454mb)
Starting worker @ ip-172-31-12-175.ec2.internal
[timing] tachyon setup: 00h 00m 14s
Setting up rstdio
spark-ec2/setup.sh: line 110: ./rstudio/setup.sh: No such file or directory
[timing] rstudio setup: 00h 00m 00s
Setting up ganglia
RSYNC'ing /etc/ganglia to slaves...
ec2-54-172-108-231.compute-1.amazonaws.com
Shutting down GANGLIA gmond: [FAILED]
Starting GANGLIA gmond: [OK]
Shutting down GANGLIA gmond: [FAILED]
Starting GANGLIA gmond: [OK]
Connection to ec2-54-172-108-231.compute-1.amazonaws.com closed.
Shutting down GANGLIA gmetad: [FAILED]
Starting GANGLIA gmetad: [OK]
Stopping httpd: [FAILED]
Starting httpd: httpd: syntax error on line 154 of /etc/httpd/conf/httpd.conf: Cannot load /etc/httpd/modules/mod_authz_core.so into server: /etc/httpd/modules/mod_authz_core.so: cannot open shared object file: No such file or directory [FAILED]
[timing] ganglia setup: 00h 00m 02s
Connection to ec2-54-152-192-90.compute-1.amazonaws.com closed.
Spark standalone cluster started at http://ec2-54-152-192-90.compute-1.amazonaws.com:8080
Ganglia started at http://ec2-54-152-192-90.compute-1.amazonaws.com:5080/ganglia
Done!
ubuntu@ip-172-31-7-38:~/spark-1.6.1-bin-hadoop2.6/ec2$ 
```

5. Spark has created a master and slave

PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

A20360111

Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Status	Public DNS
hadoop-mast...	i-11e2e395	c3.large	us-east-1d	running	Initializing	None	ec2-54-89-147-12.com
hadoop-mast...	i-1cb3b598	c3.large	us-east-1d	terminated		None	
hadoop-mast...	i-454445c1	c3.large	us-east-1d	terminated		None	
hadoop-slave...	i-554544d1	c3.large	us-east-1d	terminated		None	
hadoop-slave...	i-91d9d815	c3.large	us-east-1d	running	2/2 checks ...	None	ec2-54-209-147-28.com
hadoop-slave...	i-5e5e471	c3.large	us-east-1d	running	Initializing	None	ec2-54-172-202-240.co...

- Mount the EBS volume on Master Node and generate data using gensort

```
ubuntu@ip-172-31-14-229: ~/spark-1.6.1-bin-hadoop2.6/ec2
Writing inode tables: done
Creating journal (32768 blocks): done
Writing superblocks and filesystem accounting information: done

root@ip-172-31-8-83:~$ sudo mke2fs -F -t ext4 /dev/xvdp
mke2fs 1.42.3 (14-May-2012)
Filesystem label=
OS type: Linux
Block size=4096 (log=2)
Fragment size=4096 (log=2)
Stride=0 blocks, Stripe width=0 blocks
16553600 inodes, 26214400 blocks
1310720 blocks (5.00%) reserved for the super user
First data block=0
Maximum filesystem blocks=4294967296
800 block groups
32768 blocks per group, 32768 fragments per group
8192 inodes per group
Superblock backups stored on blocks:
    32768, 98304, 163840, 229376, 294912, 819200, 884736, 1605632, 2654208,
    4096000, 7962624, 11239424, 20480000, 23887872

Allocating group tables: done
Writing inode tables: done
Creating journal (32768 blocks): done
Writing superblocks and filesystem accounting information: done

root@ip-172-31-8-83:~$ sudo mkdir /data
root@ip-172-31-8-83:~$ sudo mount /dev/xvdp /data
root@ip-172-31-8-83:~$ sudo chmod 777 /data
root@ip-172-31-8-83:~$ ls
ephemeral-hdfs  gensort  hadoop-native  mapreduce  persistent-hdfs  scala  spark  SparkCode.scala  spark-ec2  tachyon  valsort
root@ip-172-31-8-83:~$ ./gensort -a 100000000 /data/dataset
root@ip-172-31-8-83:~$ cd ephemeral-hdfs/
root@ip-172-31-8-83:~/ephemeral-hdfs$ bin/hadoop fs -mkdir -p /user/ubuntu/gutenberg
Warning: $HADOOP_HOME is deprecated.

root@ip-172-31-8-83:~/ephemeral-hdfs$ bin/hadoop fs -Ddfs.replication=1 -put /data/dataset /user/ubuntu/gutenberg
Warning: $HADOOP_HOME is deprecated.
```

- Run the Scala code

PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

A20360111

```
ubuntu@ip-172-31-14-229: ~/spark-1.6.1-bin-hadoop2.6/ec2
root@ip-172-31-8-83 ~]$ cd ephemeral-hdfs/
root@ip-172-31-8-83 ephemeral-hdfs]$ bin/hadoop fs -mkdir -p /user/ubuntu/gutenberg
Warning: $HADOOP_HOME is deprecated.

root@ip-172-31-8-83 ephemeral-hdfs]$ bin/hadoop fs -Ddfs.replication=1 -put /data/dataset /user/ubuntu/gutenberg
Warning: $HADOOP_HOME is deprecated.

root@ip-172-31-8-83 ephemeral-hdfs]$ cd ..
root@ip-172-31-8-83 ~]$ cd spark
root@ip-172-31-8-83 spark]$ cd bin
root@ip-172-31-8-83 btnj]$ cd ..
root@ip-172-31-8-83 spark]$ cd ..
root@ip-172-31-8-83 ~]$ sudo vim SparkCode.scala
root@ip-172-31-8-83 ~]$ cd spark
root@ip-172-31-8-83 spark]$ cd bin
root@ip-172-31-8-83 bin]$ ./spark-shell -i /root/SparkCode.scala
16/03/30 23:37:14 INFO spark.SecurityManager: Changing view acls to: root
16/03/30 23:37:14 INFO spark.SecurityManager: Changing modify acls to: root
16/03/30 23:37:14 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(root)
; users with modify permissions: Set(root)
16/03/30 23:37:14 INFO spark.HttpServer: Starting HTTP Server
16/03/30 23:37:14 INFO server.Server: jetty-8.y.z-SNAPSHOT
16/03/30 23:37:14 INFO server.AbstractConnector: Started SocketConnector@0.0.0.0:49493
16/03/30 23:37:14 INFO util.Utils: Successfully started service 'HTTP class server' on port 49493.
Welcome to

           _/\_ 
          / \ \_ 
         /   \ \_ 
        /     \ \_ 
       /       \ \_ 
      /         \ \_ 
     /           \ \_ 
    /             \ \_ 
   /               \ \_ 
  /                 \ \_ 
 /                   \ \_ 
version 1.6.1
/ / 
/ / 

Using Scala version 2.10.5 (OpenJDK 64-Bit Server VM, Java 1.7.0_99)
Type in expressions to have them evaluated.
Type :help for more information.
16/03/30 23:37:20 INFO spark.SparkContext: Running Spark version 1.6.1
16/03/30 23:37:20 WARN spark.SparkConf:
SPARK_WORKER_INSTANCES was detected (set to '1').
This is deprecated in Spark 1.0+.

Please instead use:
  ./spark-submit with --num-executors to specify the number of executors
```

Output

```
ubuntu@ip-172-31-14-229: ~/spark-1.6.1-bin-hadoop2.6/ec2
1894 bytes)
16/03/30 23:49:12 INFO scheduler.TaskSetManager: Finished task 63.0 in stage 2.0 (TID 213) in 14569 ms on ip-172-31-0-66.ec2.internal (64/75)
16/03/30 23:49:20 INFO scheduler.TaskSetManager: Starting task 66.0 in stage 2.0 (TID 216, ip-172-31-0-66.ec2.internal, partition 66,NODE_LOCAL,
1894 bytes)
16/03/30 23:49:20 INFO scheduler.TaskSetManager: Finished task 64.0 in stage 2.0 (TID 214) in 12858 ms on ip-172-31-0-66.ec2.internal (65/75)
16/03/30 23:49:24 INFO scheduler.TaskSetManager: Starting task 67.0 in stage 2.0 (TID 217, ip-172-31-0-66.ec2.internal, partition 67,NODE_LOCAL,
1894 bytes)
16/03/30 23:49:24 INFO scheduler.TaskSetManager: Finished task 65.0 in stage 2.0 (TID 215) in 11607 ms on ip-172-31-0-66.ec2.internal (66/75)
16/03/30 23:49:32 INFO scheduler.TaskSetManager: Starting task 68.0 in stage 2.0 (TID 218, ip-172-31-0-66.ec2.internal, partition 68,NODE_LOCAL,
1894 bytes)
16/03/30 23:49:32 INFO scheduler.TaskSetManager: Finished task 66.0 in stage 2.0 (TID 216) in 11622 ms on ip-172-31-0-66.ec2.internal (67/75)
16/03/30 23:49:37 INFO scheduler.TaskSetManager: Starting task 69.0 in stage 2.0 (TID 219, ip-172-31-0-66.ec2.internal, partition 69,NODE_LOCAL,
1894 bytes)
16/03/30 23:49:37 INFO scheduler.TaskSetManager: Finished task 67.0 in stage 2.0 (TID 217) in 13346 ms on ip-172-31-0-66.ec2.internal (68/75)
16/03/30 23:49:45 INFO scheduler.TaskSetManager: Starting task 70.0 in stage 2.0 (TID 220, ip-172-31-0-66.ec2.internal, partition 70,NODE_LOCAL,
1894 bytes)
16/03/30 23:49:45 INFO scheduler.TaskSetManager: Finished task 68.0 in stage 2.0 (TID 218) in 13066 ms on ip-172-31-0-66.ec2.internal (69/75)
16/03/30 23:49:50 INFO scheduler.TaskSetManager: Starting task 71.0 in stage 2.0 (TID 221, ip-172-31-0-66.ec2.internal, partition 71,NODE_LOCAL,
1894 bytes)
16/03/30 23:49:50 INFO scheduler.TaskSetManager: Finished task 69.0 in stage 2.0 (TID 219) in 13020 ms on ip-172-31-0-66.ec2.internal (70/75)
16/03/30 23:50:00 INFO scheduler.TaskSetManager: Starting task 72.0 in stage 2.0 (TID 222, ip-172-31-0-66.ec2.internal, partition 72,NODE_LOCAL,
1894 bytes)
16/03/30 23:50:00 INFO scheduler.TaskSetManager: Finished task 70.0 in stage 2.0 (TID 220) in 14964 ms on ip-172-31-0-66.ec2.internal (71/75)
16/03/30 23:50:04 INFO scheduler.TaskSetManager: Starting task 73.0 in stage 2.0 (TID 223, ip-172-31-0-66.ec2.internal, partition 73,NODE_LOCAL,
1894 bytes)
16/03/30 23:50:04 INFO scheduler.TaskSetManager: Finished task 71.0 in stage 2.0 (TID 221) in 13766 ms on ip-172-31-0-66.ec2.internal (72/75)
16/03/30 23:50:15 INFO scheduler.TaskSetManager: Starting task 74.0 in stage 2.0 (TID 224, ip-172-31-0-66.ec2.internal, partition 74,NODE_LOCAL,
1894 bytes)
16/03/30 23:50:15 INFO scheduler.TaskSetManager: Finished task 72.0 in stage 2.0 (TID 222) in 15310 ms on ip-172-31-0-66.ec2.internal (73/75)
16/03/30 23:50:15 INFO scheduler.TaskSetManager: Finished task 73.0 in stage 2.0 (TID 223) in 11262 ms on ip-172-31-0-66.ec2.internal (74/75)
16/03/30 23:50:21 INFO scheduler.TaskSetManager: Finished task 74.0 in stage 2.0 (TID 224) in 5578 ms on ip-172-31-0-66.ec2.internal (75/75)
16/03/30 23:50:21 INFO scheduler.DAGScheduler: ResultStage 2 (saveAsTextFile at <console>:30) finished in 454.424 s
16/03/30 23:50:21 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 2.0, whose tasks have all completed, from pool
16/03/30 23:50:21 INFO scheduler.DAGScheduler: Job 1 finished: saveAsTextFile at <console>:30, took 706.576095 s
end_time: Long = 1459381821521
Time taken to Sort : 767174ms

scala>
scala>
```

PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

A20360111

Valsort

Mulit Node Spark

1. Create an instance and download spark

```
ubuntu@ip-172-31-1-85: ~
spark-1.6.1-bin-hadoop2.6/conf/fairscheduler.xml.template
spark-1.6.1-bin-hadoop2.6/conf/metrics.properties.template
spark-1.6.1-bin-hadoop2.6/conf/spark-env.sh.template
spark-1.6.1-bin-hadoop2.6/conf/log4j.properties.template
spark-1.6.1-bin-hadoop2.6/conf/docker.properties.template
spark-1.6.1-bin-hadoop2.6/conf/slaves.template
spark-1.6.1-bin-hadoop2.6/conf/spark-defaults.conf.template
spark-1.6.1-bin-hadoop2.6/LICENSE
spark-1.6.1-bin-hadoop2.6/bin/
spark-1.6.1-bin-hadoop2.6/bin/spark-shell
spark-1.6.1-bin-hadoop2.6/bin/spark-submit.cmd
*spark-1.6.1-bin-hadoop2.6/bin/spark-shell2.cmd
spark-1.6.1-bin-hadoop2.6/bin/pyspark
spark-1.6.1-bin-hadoop2.6/bin/sparkR.cmd
spark-1.6.1-bin-hadoop2.6/bin/spark-class.cmd
spark-1.6.1-bin-hadoop2.6/bin/run-example.cmd
spark-1.6.1-bin-hadoop2.6/bin/spark-submit2.cmd
spark-1.6.1-bin-hadoop2.6/bin/spark-class
spark-1.6.1-bin-hadoop2.6/bin/spark-submit
spark-1.6.1-bin-hadoop2.6/bin/spark-sql
spark-1.6.1-bin-hadoop2.6/bin/run-example
spark-1.6.1-bin-hadoop2.6/bin/beeline
spark-1.6.1-bin-hadoop2.6/bin/pyspark2.cmd
spark-1.6.1-bin-hadoop2.6/bin/spark-shell.cmd
spark-1.6.1-bin-hadoop2.6/bin/spark-class.cmd
spark-1.6.1-bin-hadoop2.6/bin/pyspark.cmd
spark-1.6.1-bin-hadoop2.6/bin/sparkR
spark-1.6.1-bin-hadoop2.6/bin/beeline.cmd
spark-1.6.1-bin-hadoop2.6/bin/sparkR2.cmd
spark-1.6.1-bin-hadoop2.6/bin/run-example2.cmd
spark-1.6.1-bin-hadoop2.6/bin/load-spark-env.sh
spark-1.6.1-bin-hadoop2.6/bin/load-spark-env.cmd
spark-1.6.1-bin-hadoop2.6/lib/
spark-1.6.1-bin-hadoop2.6/lib/datanucleus-core-3.2.10.jar
spark-1.6.1-bin-hadoop2.6/lib/spark-assembly-1.6.1-hadoop2.6.0.jar
spark-1.6.1-bin-hadoop2.6/lib/datanucleus-api-jdo-3.2.6.jar
spark-1.6.1-bin-hadoop2.6/lib/datanucleus-rdbms-3.9.9.jar
spark-1.6.1-bin-hadoop2.6/lib/spark-1.6.1-yarn-shuffle.jar
spark-1.6.1-bin-hadoop2.6/lib/spark-examples-1.6.1-hadoop2.6.0.jar
spark-1.6.1-bin-hadoop2.6/README.md
ubuntu@ip-172-31-1-85: ~
```

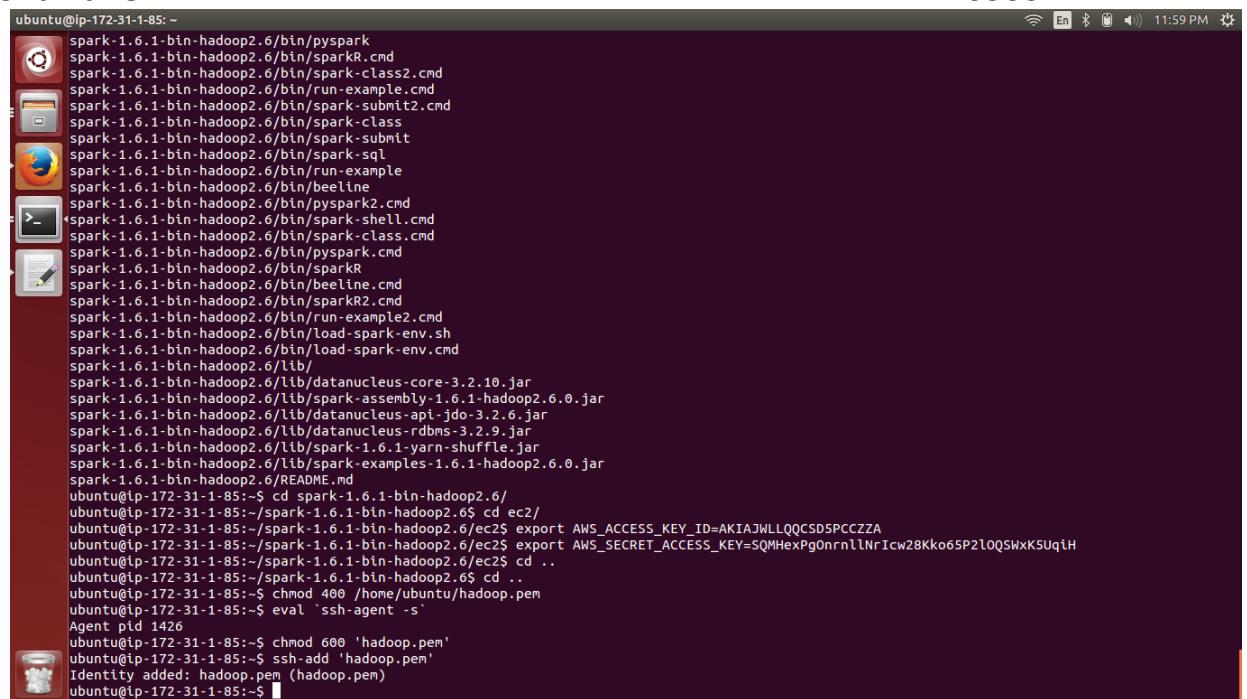
2. Unzip the downloaded file and export ACCESS KEY and SECRET ACCESS KEY

PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

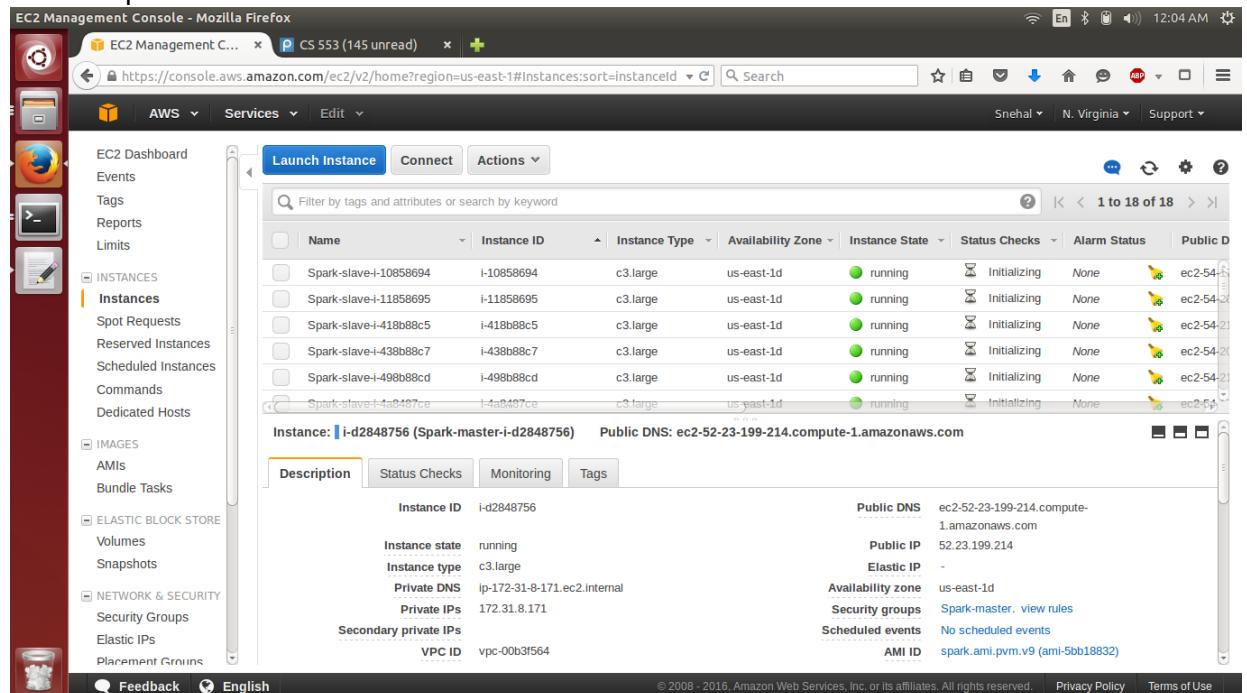
A20360111



```

ubuntu@ip-172-31-1-85: ~
spark-1.6.1-bin-hadoop2.6/bin/pyspark
spark-1.6.1-bin-hadoop2.6/bin/sparkR.cmd
spark-1.6.1-bin-hadoop2.6/bin/spark-class2.cmd
spark-1.6.1-bin-hadoop2.6/bin/run-example.cmd
spark-1.6.1-bin-hadoop2.6/bin/spark-submit2.cmd
spark-1.6.1-bin-hadoop2.6/bin/spark-class
spark-1.6.1-bin-hadoop2.6/bin/spark-submit
spark-1.6.1-bin-hadoop2.6/bin/spark-sql
spark-1.6.1-bin-hadoop2.6/bin/run-example
spark-1.6.1-bin-hadoop2.6/bin/beeline
spark-1.6.1-bin-hadoop2.6/bin/pyspark2.cmd
spark-1.6.1-bin-hadoop2.6/bin/spark-shell.cmd
spark-1.6.1-bin-hadoop2.6/bin/spark-class.cmd
spark-1.6.1-bin-hadoop2.6/bin/pyspark.cmd
spark-1.6.1-bin-hadoop2.6/bin/sparkR
spark-1.6.1-bin-hadoop2.6/bin/beeline.cmd
spark-1.6.1-bin-hadoop2.6/bin/sparkR2.cmd
spark-1.6.1-bin-hadoop2.6/bin/run-example2.cmd
spark-1.6.1-bin-hadoop2.6/bin/load-spark-env.sh
spark-1.6.1-bin-hadoop2.6/bin/load-spark-env.cmd
spark-1.6.1-bin-hadoop2.6/lib/
spark-1.6.1-bin-hadoop2.6/lib/datanucleus-core-3.2.10.jar
spark-1.6.1-bin-hadoop2.6/lib/spark-assembly-1.6.1-hadoop2.6.0.jar
spark-1.6.1-bin-hadoop2.6/lib/datanucleus-api-jdo-3.2.6.jar
spark-1.6.1-bin-hadoop2.6/lib/datanucleus-rdbms-3.2.9.jar
spark-1.6.1-bin-hadoop2.6/lib/spark-1.6.1-yarn-shuffle.jar
spark-1.6.1-bin-hadoop2.6/lib/spark-examples-1.6.1-hadoop2.6.0.jar
spark-1.6.1-bin-hadoop2.6/README.md
ubuntu@ip-172-31-1-85:~$ cd spark-1.6.1-bin-hadoop2.6/
ubuntu@ip-172-31-1-85:~/spark-1.6.1-bin-hadoop2.6$ cd ec2
ubuntu@ip-172-31-1-85:~/spark-1.6.1-bin-hadoop2.6/ec2$ export AWS_ACCESS_KEY_ID=AKIAJWLLQQCSD5PCCZZA
ubuntu@ip-172-31-1-85:~/spark-1.6.1-bin-hadoop2.6/ec2$ export AWS_SECRET_ACCESS_KEY=SQMHexpG0nrnllNrIcw2KKo65P2l0QSwK5UqtlH
ubuntu@ip-172-31-1-85:~/spark-1.6.1-bin-hadoop2.6/ec2$ cd ..
ubuntu@ip-172-31-1-85:~/spark-1.6.1-bin-hadoop2.6$ chmod 400 /home/ubuntu/hadoop.pem
ubuntu@ip-172-31-1-85:~/spark-1.6.1-bin-hadoop2.6$ eval `ssh-agent -s`
Agent pid 1426
ubuntu@ip-172-31-1-85:~/spark-1.6.1-bin-hadoop2.6$ chmod 600 'hadoop.pem'
ubuntu@ip-172-31-1-85:~/spark-1.6.1-bin-hadoop2.6$ ssh-add 'hadoop.pem'
Identity added: hadoop.pem (hadoop.pem)
ubuntu@ip-172-31-1-85:~/spark-1.6.1-bin-hadoop2.6$
```

3. Launch Spark to result in 16 Nodes



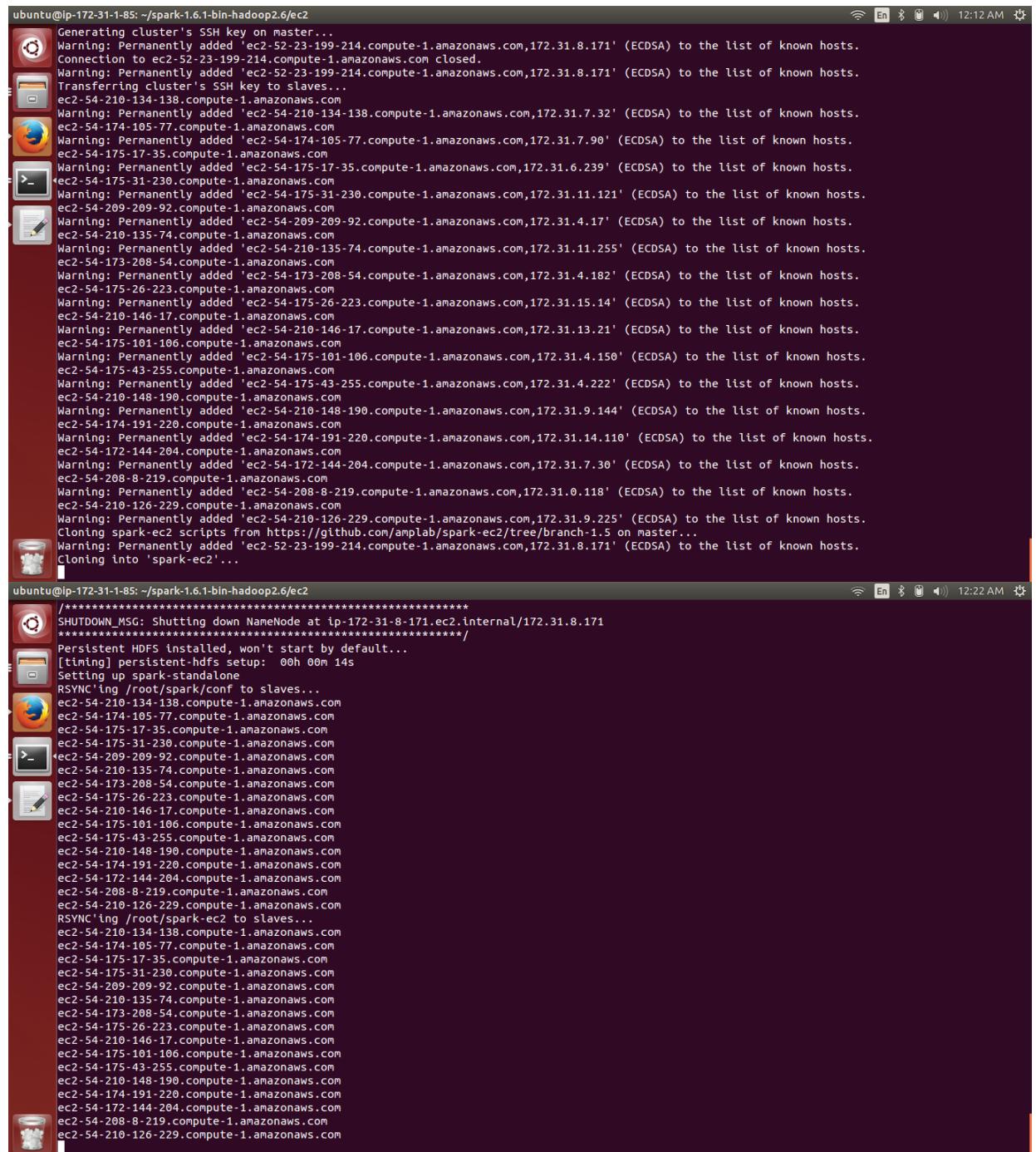
The screenshot shows the AWS EC2 Management Console interface. On the left, there's a sidebar with navigation links like EC2 Dashboard, Events, Tags, Reports, Limits, INSTANCES, IMAGES, AMIs, and ELASTIC BLOCK STORE. The main area is titled "Launch Instance" and shows a table of existing instances. The table includes columns for Name, Instance ID, Instance Type, Availability Zone, Instance State, Status Checks, Alarm Status, and Public DNS. All listed instances are in the "running" state. Below the table, there's a detailed view for instance i-d2848756, which is a "Spark-master" type. The details include Public DNS (ec2-52-23-199-214.compute-1.amazonaws.com), Public IP (52.23.199.214), and various network and security settings.

PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

A20360111



The image shows a dual-terminal setup on an Ubuntu desktop. The top terminal window is titled "ubuntu@ip-172-31-1-85: ~/spark-1.6.1-bin-hadoop2.6/ec2" and displays a log of SSH key additions to the master host. The bottom terminal window is titled "ubuntu@ip-172-31-1-85: ~/spark-1.6.1-bin-hadoop2.6/ec2" and shows the execution of a spark-ec2 script to start a NameNode and a persistent HDFS instance.

```
ubuntu@ip-172-31-1-85: ~/spark-1.6.1-bin-hadoop2.6/ec2
Generating cluster's SSH key on master...
Warning: Permanently added 'ec2-52-23-199-214.compute-1.amazonaws.com,172.31.8.171' (ECDSA) to the list of known hosts.
Connection to ec2-52-23-199-214.compute-1.amazonaws.com closed.
Warning: Permanently added 'ec2-52-23-199-214.compute-1.amazonaws.com,172.31.8.171' (ECDSA) to the list of known hosts.
Transferring cluster's SSH key to slaves...
ec2-54-210-134-138.compute-1.amazonaws.com
Warning: Permanently added 'ec2-54-210-134-138.compute-1.amazonaws.com,172.31.7.32' (ECDSA) to the list of known hosts.
ec2-54-174-105-77.compute-1.amazonaws.com
Warning: Permanently added 'ec2-54-174-105-77.compute-1.amazonaws.com,172.31.7.90' (ECDSA) to the list of known hosts.
ec2-54-175-17-35.compute-1.amazonaws.com
Warning: Permanently added 'ec2-54-175-17-35.compute-1.amazonaws.com,172.31.6.239' (ECDSA) to the list of known hosts.
ec2-54-175-31-230.compute-1.amazonaws.com
Warning: Permanently added 'ec2-54-175-31-230.compute-1.amazonaws.com,172.31.11.121' (ECDSA) to the list of known hosts.
ec2-54-209-209-92.compute-1.amazonaws.com
Warning: Permanently added 'ec2-54-209-209-92.compute-1.amazonaws.com,172.31.4.17' (ECDSA) to the list of known hosts.
ec2-54-210-135-74.compute-1.amazonaws.com
Warning: Permanently added 'ec2-54-210-135-74.compute-1.amazonaws.com,172.31.11.255' (ECDSA) to the list of known hosts.
ec2-54-173-208-54.compute-1.amazonaws.com
Warning: Permanently added 'ec2-54-173-208-54.compute-1.amazonaws.com,172.31.4.182' (ECDSA) to the list of known hosts.
ec2-54-175-26-223.compute-1.amazonaws.com
Warning: Permanently added 'ec2-54-175-26-223.compute-1.amazonaws.com,172.31.15.14' (ECDSA) to the list of known hosts.
ec2-54-210-146-17.compute-1.amazonaws.com
Warning: Permanently added 'ec2-54-210-146-17.compute-1.amazonaws.com,172.31.13.21' (ECDSA) to the list of known hosts.
ec2-54-175-101-106.compute-1.amazonaws.com
Warning: Permanently added 'ec2-54-175-101-106.compute-1.amazonaws.com,172.31.4.150' (ECDSA) to the list of known hosts.
ec2-54-175-43-255.compute-1.amazonaws.com
Warning: Permanently added 'ec2-54-175-43-255.compute-1.amazonaws.com,172.31.4.222' (ECDSA) to the list of known hosts.
ec2-54-210-148-190.compute-1.amazonaws.com
Warning: Permanently added 'ec2-54-210-148-190.compute-1.amazonaws.com,172.31.9.144' (ECDSA) to the list of known hosts.
ec2-54-174-191-220.compute-1.amazonaws.com
Warning: Permanently added 'ec2-54-174-191-220.compute-1.amazonaws.com,172.31.14.110' (ECDSA) to the list of known hosts.
ec2-54-172-144-204.compute-1.amazonaws.com
Warning: Permanently added 'ec2-54-172-144-204.compute-1.amazonaws.com,172.31.7.30' (ECDSA) to the list of known hosts.
ec2-54-208-8-219.compute-1.amazonaws.com
Warning: Permanently added 'ec2-54-208-8-219.compute-1.amazonaws.com,172.31.0.118' (ECDSA) to the list of known hosts.
ec2-54-210-126-229.compute-1.amazonaws.com
Warning: Permanently added 'ec2-54-210-126-229.compute-1.amazonaws.com,172.31.9.225' (ECDSA) to the list of known hosts.
Cloning spark-ec2 scripts from https://github.com/amplab/spark-ec2/tree/branch-1.5 on master...
Warning: Permanently added 'ec2-52-23-199-214.compute-1.amazonaws.com,172.31.8.171' (ECDSA) to the list of known hosts.
Cloning into 'spark-ec2'...

```



```
ubuntu@ip-172-31-1-85: ~/spark-1.6.1-bin-hadoop2.6/ec2
*****
SHUTDOWN MSG: Shutting down NameNode at ip-172-31-8-171.ec2.internal/172.31.8.171
*****
Persistent HDFS Installd, won't start by default...
[timing] persistent-hdfs setup: 00h 00m 14s
Setting up spark-standalone
RSYNC'ing /root/spark/conf to slaves...
ec2-54-210-134-138.compute-1.amazonaws.com
ec2-54-174-105-77.compute-1.amazonaws.com
ec2-54-175-17-35.compute-1.amazonaws.com
ec2-54-175-31-230.compute-1.amazonaws.com
ec2-54-209-209-92.compute-1.amazonaws.com
ec2-54-210-135-74.compute-1.amazonaws.com
ec2-54-173-208-54.compute-1.amazonaws.com
ec2-54-175-26-223.compute-1.amazonaws.com
ec2-54-210-146-17.compute-1.amazonaws.com
ec2-54-175-101-106.compute-1.amazonaws.com
ec2-54-175-43-255.compute-1.amazonaws.com
ec2-54-210-148-190.compute-1.amazonaws.com
ec2-54-174-191-220.compute-1.amazonaws.com
ec2-54-172-144-204.compute-1.amazonaws.com
ec2-54-208-8-219.compute-1.amazonaws.com
ec2-54-210-126-229.compute-1.amazonaws.com
RSYNC'ing /root/spark-ec2 to slaves...
ec2-54-210-134-138.compute-1.amazonaws.com
ec2-54-174-105-77.compute-1.amazonaws.com
ec2-54-175-17-35.compute-1.amazonaws.com
ec2-54-175-31-230.compute-1.amazonaws.com
ec2-54-209-209-92.compute-1.amazonaws.com
ec2-54-210-135-74.compute-1.amazonaws.com
ec2-54-173-208-54.compute-1.amazonaws.com
ec2-54-175-26-223.compute-1.amazonaws.com
ec2-54-210-146-17.compute-1.amazonaws.com
ec2-54-175-101-106.compute-1.amazonaws.com
ec2-54-175-43-255.compute-1.amazonaws.com
ec2-54-210-148-190.compute-1.amazonaws.com
ec2-54-174-191-220.compute-1.amazonaws.com
ec2-54-172-144-204.compute-1.amazonaws.com
ec2-54-208-8-219.compute-1.amazonaws.com
ec2-54-210-126-229.compute-1.amazonaws.com

```

PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

A20360111

```
ubuntu@ip-172-31-1-85:~/spark-1.6.1-bin-hadoop2.6/ec2
Connection to ec2-54-174-191-220.compute-1.amazonaws.com closed.
Shutting down GANGLIA gmond: [FAILED]
Starting GANGLIA gmond: [OK]
Connection to ec2-54-172-144-204.compute-1.amazonaws.com closed.
Shutting down GANGLIA gmond: [FAILED]
Starting GANGLIA gmond: [OK]
Connection to ec2-54-208-8-219.compute-1.amazonaws.com closed.
Shutting down GANGLIA gmond: [FAILED]
Starting GANGLIA gmond: [OK]
Connection to ec2-54-210-126-229.compute-1.amazonaws.com closed.
Shutting down GANGLIA gmetad: [FAILED]
Starting GANGLIA gmetad: [OK]
Stopping httpd: [FAILED]
Starting httpd: httpd: Syntax error on line 154 of /etc/httpd/conf/httpd.conf: Cannot load /etc/httpd/modules/mod_authz_core.so into server: /etc/httpd/modules/mod_authz_core.so: cannot open shared object file: No such file or directory [FAILED]

[timing] ganglia setup: 00h 00m 12s
Connection to ec2-52-23-199-214.compute-1.amazonaws.com closed.
Spark standalone cluster started at http://ec2-52-23-199-214.compute-1.amazonaws.com:8080
Ganglia started at http://ec2-52-23-199-214.compute-1.amazonaws.com:5080/ganglia
Done!
ubuntu@ip-172-31-1-85:~/spark-1.6.1-bin-hadoop2.6/ec2$ ./spark-ec2 -k hadoop -i /home/ubuntu/hadoop.pem login hadoop
Searching for existing cluster hadoop in region us-east-1...
ERROR: Could not find a master for cluster hadoop in region us-east-1.
ubuntu@ip-172-31-1-85:~/spark-1.6.1-bin-hadoop2.6/ec2$ ./spark-ec2 -k hadoop -i /home/ubuntu/hadoop.pem login hadoop
Searching for existing cluster hadoop in region us-east-1...
ERROR: Could not find a master for cluster hadoop in region us-east-1.
ubuntu@ip-172-31-1-85:~/spark-1.6.1-bin-hadoop2.6/ec2$ ./spark-ec2 -k hadoop -i /home/ubuntu/hadoop.pem login Spark
Searching for existing cluster Spark in region us-east-1...
Found 1 master, 16 slaves.
Logging into master ec2-52-23-199-214.compute-1.amazonaws.com...
Warning: Permanently added 'ec2-52-23-199-214.compute-1.amazonaws.com,172.31.8.171' (ECDSA) to the list of known hosts.
Last login: Thu Mar 31 05:13:02 2016 from ip-172-31-8-171.ec2.internal

 _I _|_ )
 _I ( _|_ / Amazon Linux AMI
 __|__|_|_|

https://aws.amazon.com/amazon-linux-ami/2013.03-release-notes/
Amazon Linux version 2016.03 ts available.
root@ip-172-31-8-171 ~]$
```

4. Mount the EBS volume on Master Node and generate data using gensor

```
ubuntu@ip-172-31-1-85:~/spark-1.6.1-bin-hadoop2.6/ec2
4000 block groups
32768 blocks per group, 32768 fragments per group
8192 inodes per group
Superblock backups stored on blocks:
    32768, 98304, 163840, 229376, 294912, 819200, 884736, 1605632, 2654208,
    4096000, 7962624, 11239424, 20480000, 23887872, 71663616, 78675968,
    102400000

Allocating group tables: done
Writing inode tables: done
Creating journal (32768 blocks): done
Writing superblocks and filesystem accounting information: done

root@ip-172-31-8-171 ~] $ sudo mke2fs -F -t ext4 /dev/xvdp
mke2fs 1.42.3 (14-May-2012)
Filesystem label=
OS type: Linux
Block size=4096 (log=2)
Fragment size=4096 (log=2)
Stride=0 blocks, Stripe width=0 blocks
32768000 inodes, 131072000 blocks
6553600 blocks (5.00%) reserved for the super user
First data block=0
Maximum filesystem blocks=4294967296
4000 block groups
32768 blocks per group, 32768 fragments per group
8192 inodes per group
Superblock backups stored on blocks:
    32768, 98304, 163840, 229376, 294912, 819200, 884736, 1605632, 2654208,
    4096000, 7962624, 11239424, 20480000, 23887872, 71663616, 78675968,
    102400000

Allocating group tables: done
Writing inode tables: done
Creating journal (32768 blocks): done
Writing superblocks and filesystem accounting information: done

root@ip-172-31-8-171 ~] $ sudo mkdir /data
root@ip-172-31-8-171 ~] $ sudo mount /dev/xvdp /data
root@ip-172-31-8-171 ~] $ sudo chmod 777 /data
root@ip-172-31-8-171 ~] $
```

PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

A20360111

5. Output

```
ubuntu@ip-172-31-1-85:~/spark-1.6.1-bin-hadoop2.6/ec2
45)
16/03/31 07:00:29 INFO scheduler.TaskSetManager: Finished task 731.0 in stage 2.0 (TID 2221) in 18502 ms on ip-172-31-7-32.ec2.internal (733/745
)
16/03/31 07:00:29 INFO scheduler.TaskSetManager: Finished task 729.0 in stage 2.0 (TID 2219) in 19311 ms on ip-172-31-11-121.ec2.internal (734/7
45)
16/03/31 07:00:29 INFO scheduler.TaskSetManager: Finished task 744.0 in stage 2.0 (TID 2234) in 10871 ms on ip-172-31-7-32.ec2.internal (735/745
)
16/03/31 07:00:29 INFO scheduler.TaskSetManager: Finished task 735.0 in stage 2.0 (TID 2225) in 17439 ms on ip-172-31-4-222.ec2.internal (736/74
5)
16/03/31 07:00:30 INFO scheduler.TaskSetManager: Finished task 741.0 in stage 2.0 (TID 2231) in 13585 ms on ip-172-31-7-30.ec2.internal (737/745
)
16/03/31 07:00:30 INFO scheduler.TaskSetManager: Finished task 727.0 in stage 2.0 (TID 2217) in 22089 ms on ip-172-31-7-30.ec2.internal (738/745
)
16/03/31 07:00:30 INFO scheduler.TaskSetManager: Finished task 736.0 in stage 2.0 (TID 2226) in 17318 ms on ip-172-31-7-90.ec2.internal (739/745
)
16/03/31 07:00:31 INFO scheduler.TaskSetManager: Finished task 718.0 in stage 2.0 (TID 2208) in 26757 ms on ip-172-31-4-182.ec2.internal (740/74
5)
16/03/31 07:00:32 INFO scheduler.TaskSetManager: Finished task 740.0 in stage 2.0 (TID 2230) in 16284 ms on ip-172-31-9-225.ec2.internal (741/74
5)
16/03/31 07:00:33 INFO scheduler.TaskSetManager: Finished task 726.0 in stage 2.0 (TID 2216) in 25867 ms on ip-172-31-7-90.ec2.internal (742/785
)
16/03/31 07:00:33 INFO scheduler.TaskSetManager: Finished task 743.0 in stage 2.0 (TID 2233) in 16480 ms on ip-172-31-13-21.ec2.internal (743/74
5)
16/03/31 07:00:35 INFO scheduler.TaskSetManager: Finished task 739.0 in stage 2.0 (TID 2229) in 19957 ms on ip-172-31-11-121.ec2.internal (744/7
45)
16/03/31 07:00:36 INFO scheduler.TaskSetManager: Finished task 742.0 in stage 2.0 (TID 2232) in 18759 ms on ip-172-31-9-144.ec2.internal (745/74
5)
16/03/31 07:00:36 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 2.0, whose tasks have all completed, from pool
16/03/31 07:00:36 INFO scheduler.DAGScheduler: ResultStage 2 (saveAsTextFile at <console>:30) finished in 392.854 s
16/03/31 07:00:36 INFO scheduler.DAGScheduler: Job 1 finished: saveAsTextFile at <console>:30, took 664.389320 s
end_time: Long = 1459407636399
Time taken to Sort :712597ms
```

Valsort

PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

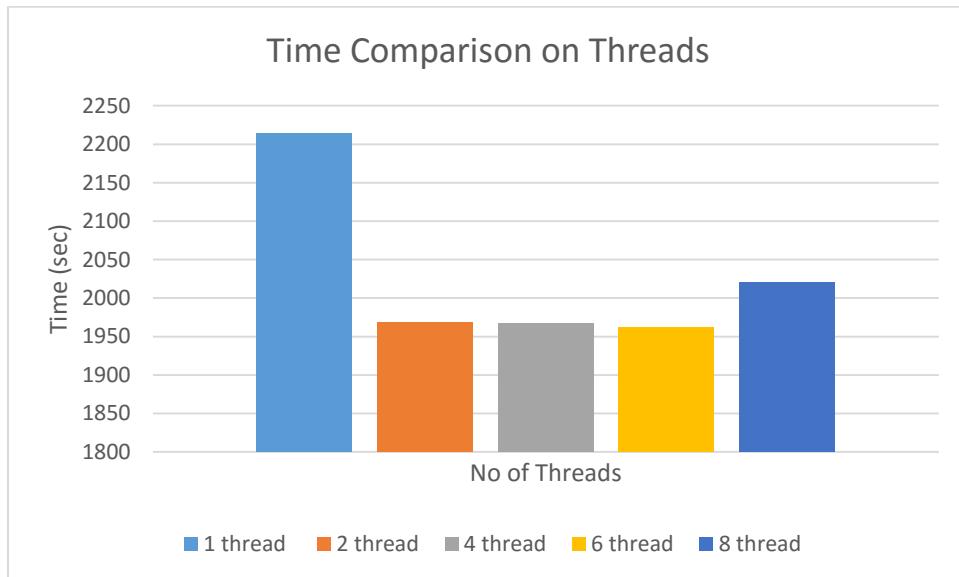
A20360111

PERFORMANCE

1. Shared Memory

The time required to sort 10 GB dataset on 1 –node using multiple threads is as follows:

Number of Threads	Time Required (sec)
1 Thread	2214
2 Thread	1969.2
4 Thread	1966.8
6 Thread	1961.4
8 Thread	2020.8



Observation:

It is observed that the Shared Memory Sort runs the fastest for 6 threads. Thus, concluding 6 threads as the best performance.

PROGRAMMING ASSIGNMENT 2

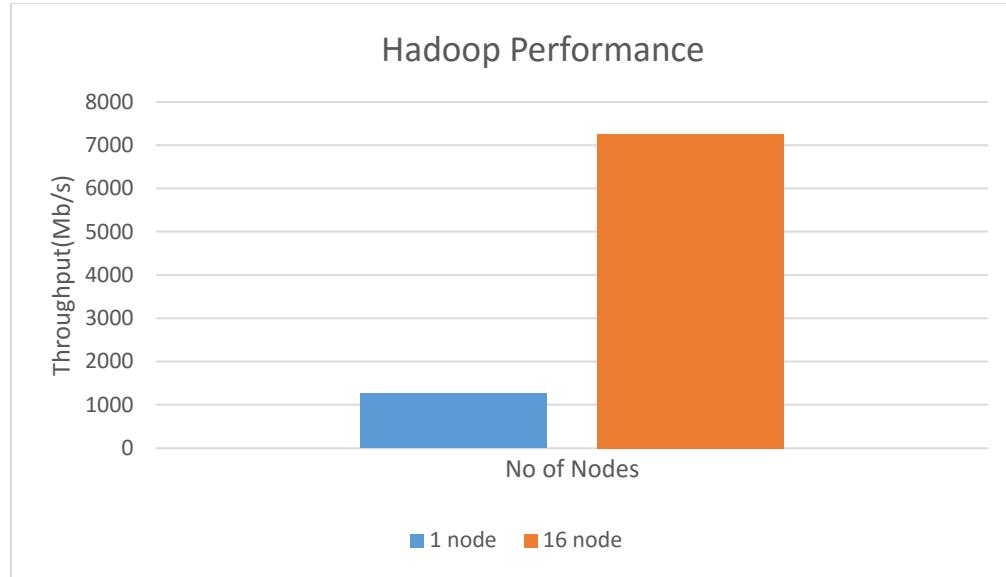
Sort on HADOOP/SPARK

Snehal Sonawane

A20360111

2. Hadoop

Performance on Single Node and Multi Node

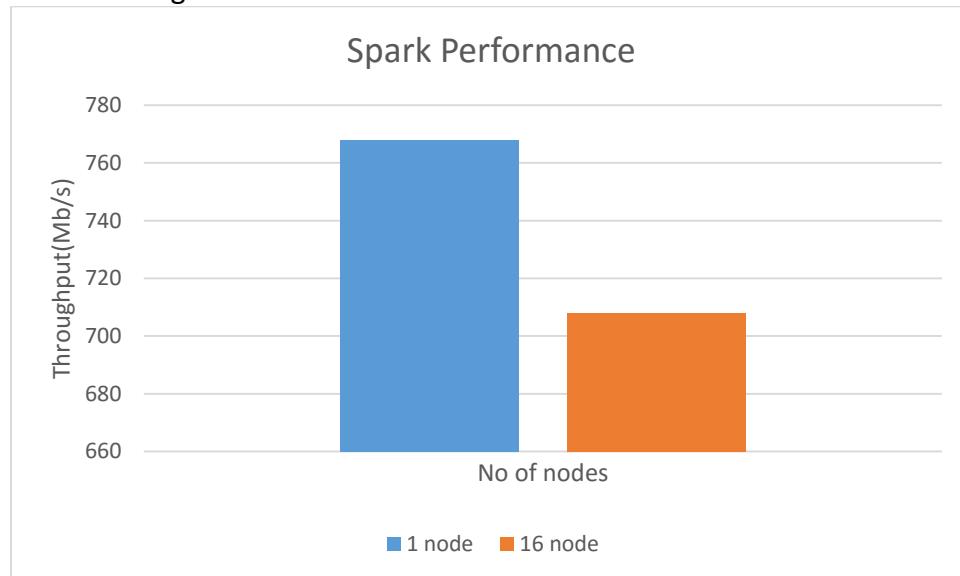


Observation:

1. It can be observed that throughput is higher for 16 nodes as we scale up from 1-node.

3. Spark

Performance on Single Node and Multi Node



Observation:

1. It can be observed that the throughput is higher on 1- node. But here is no

PROGRAMMING ASSIGNMENT 2

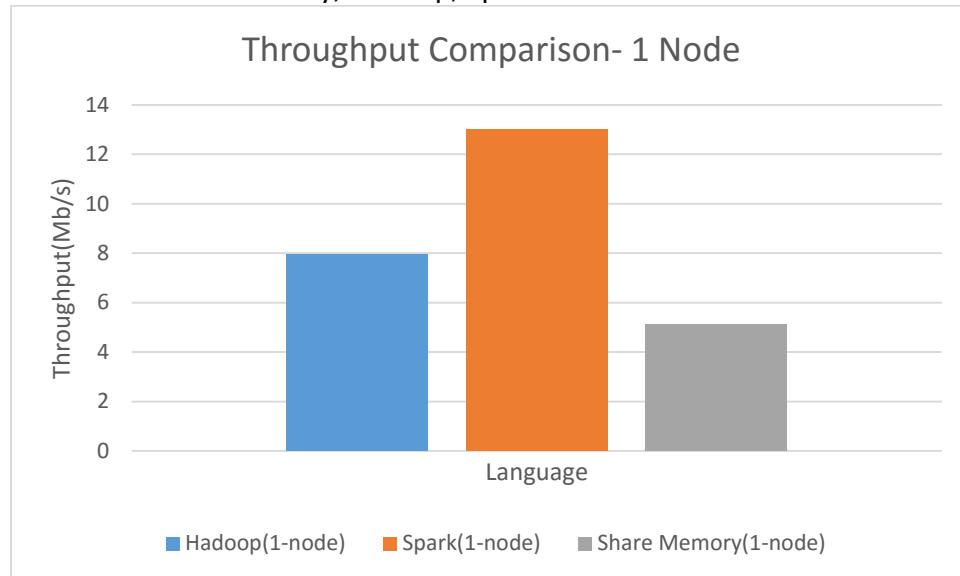
Sort on HADOOP/SPARK

Snehal Sonawane

A20360111

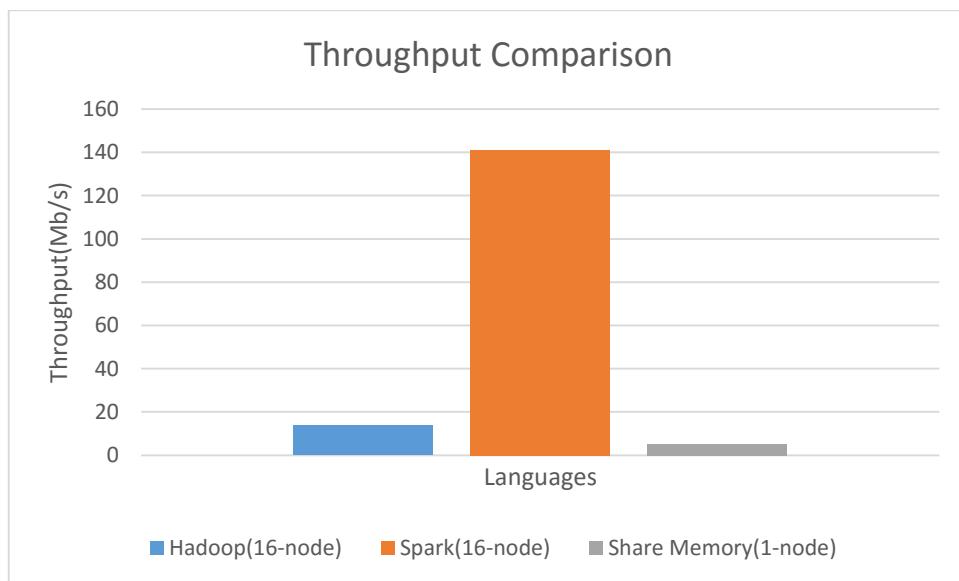
drastic change in the values.

4. Performance of Shared Memory, Hadoop, Spark



Observation:

1. Throughput is highest for Spark in comparison to Hadoop and Shared Memory on a 1 node scale.



Observation:

1. Throughput is highest for 16 node Spark in comparison to 16 Node Hadoop and 1 node Shared Memory
2. Throughput of Spark is high in spite of the scalability from 1-node to 16-node.

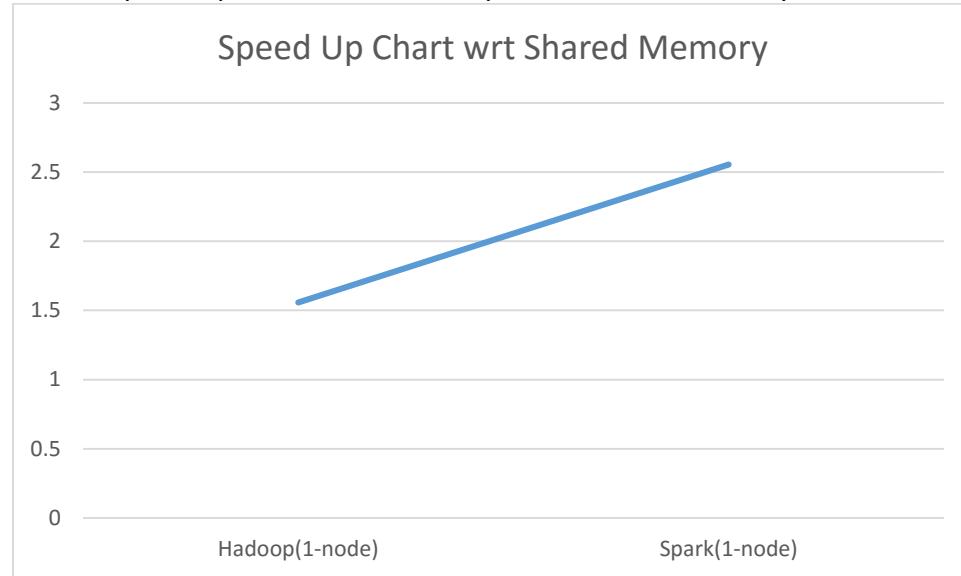
PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

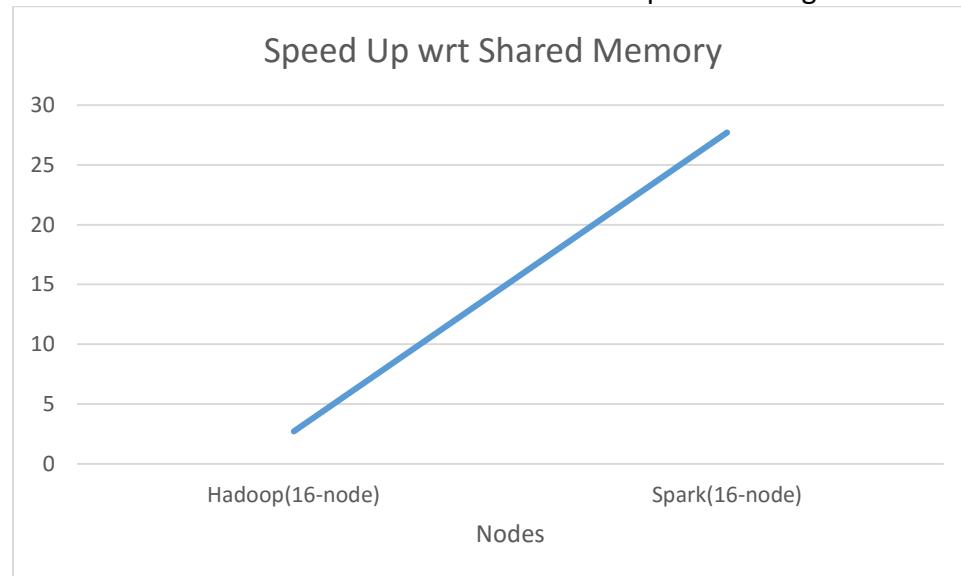
A20360111

5. Speed Up Line Chart with respect to Shared Memory on 1- node



Observation:

1. It is observed that the speed up is greater than 1 for both Hadoop and Spark with respect to Shared Memory.
2. It can be concluded that the code is not optimized to give the best solution.



Observation:

1. It is observed that the speed up is greater than 1 for both Hadoop and Spark with respect to Shared Memory.
2. For Spark it is immensely high.

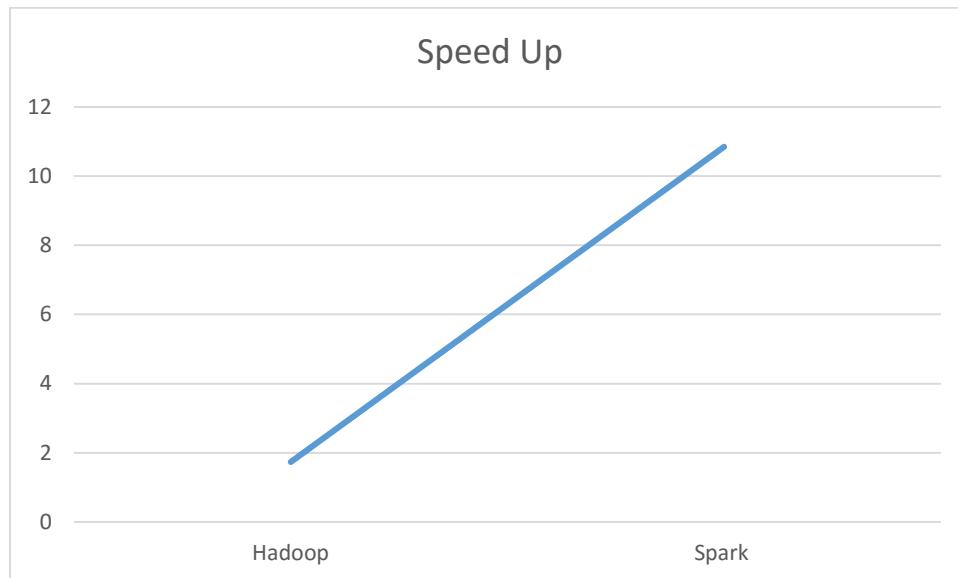
PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

A20360111

3. It can be concluded that the code is not optimized to give the best solution.



Observation:

1. It is observed that the speed up is greater than 1 for both Hadoop 16-node and Spark 16 node with respect to Hadoop 1-node and Spark 1-node.
2. It follows the expected trend.

Questions:

1. What conclusions can you draw? Which seems to be best at 1 node scale? How about 16 nodes? Can you predict which would be best at 100 node scale? How about 1000 node scales?

Answer: It can be concluded that Spark is the best for 1-node and 16-node scale and 100 node scale based on the statistics. Spark is relatively new so the performance of Spark cannot be determined to be the best on a 1000 node scale.

2. Compare your results with those from the Sort Benchmark [9], specifically the winners in 2013 and 2014 who used Hadoop and Spark.

Answer:

	Spark	Hadoop
2014 & 2013	100 TB in 1,406 seconds 207 Amazon EC2 i2.8xlarge nodes	102.5 TB in 4,328 seconds 2100 nodes
My Results	10 GB in 768 seconds 1 node	10 GB in 1260 seconds on 1 node

100 TB in how many seconds using my code?

PROGRAMMING ASSIGNMENT 2

Sort on HADOOP/SPARK

Snehal Sonawane

A20360111

10 gb → 768 sec

100000 gb → 7680000 sec on 1 –node

Thus, with the increase in the number of nodes the throughput increases for both Hadoop and Spark. As my code won't be able to give the best results in comparison with the sort benchmark for both spark and Hadoop.

3. Also, what can you learn from the CloudSort benchmark.

Cloudsort benchmark sorts records at a cheap cost on public clouds. Since the clouds are not well supported to work for IO intensive workloads and also since we face total –cost of ownership. Cloudsort benchmark is innovative and supports the total cost of ownership