

Hadoop Cluster Setup

Sorayut Glomglome

Content

1. Create Instances from image
2. Configure Hadoop cluster
3. Run MapReduce

Create 2 new instances from Hadoop image

Choose m5.large

The screenshot shows the AWS Management Console interface for creating a new instance. The top navigation bar includes the AWS logo, 'Services', 'Resource Groups', and user information. The main content area is titled 'Step 1: Choose an Amazon Machine Image (AMI)' and includes a search bar and a list of AMIs. The 'BDA-Hadoop' AMI is highlighted, and a red circle with the number '2' is placed over the 'Select' button. The sidebar on the left contains filters for 'Ownership', 'Architecture', and 'Root device type'.

Step 1: Choose an Amazon Machine Image (AMI)

An AMI is a template that contains the software configuration (operating system, application server, and applications) required to launch your instance. You can select an AMI provided by AWS, our user community, or the AWS Marketplace; or you can select one of your own AMIs.

Search for an AMI by entering a search term e.g. "Windows"

Quick Start

My AMIs **1**

BDA-Hadoop - ami-01a0140ac99084f5e

Hadoop Single Node

Root device type: ebs Virtualization type: hvm Owner: 903840611884 ENA Enabled: Yes

2 **Select**

64-bit (x86)

Ownership

- ☒ Owned by me
- ☐ Shared with me

Architecture

- ☐ 32-bit (x86)
- ☐ 64-bit (x86)
- ☐ 64-bit (Arm)

Root device type

- ☐ EBS

Feedback **English (US)**

© 2008 - 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved. [Privacy Policy](#) [Terms of Use](#)

Create 2 new instances from Hadoop image

aws Services Resource Groups

Sorayut Glomglome Oregon Support

1. Choose AMI 2. Choose Instance Type 3. Configure Instance 4. Add Storage 5. Add Tags 6. Configure Security Group 7. Review

Step 3: Configure Instance Details

Configure the instance to suit your requirements. You can launch multiple instances from the same AMI, request Spot instances to take advantage of the lower pricing, assign an access management role to the instance, and more.

Number of instances ⓘ 2 1 [Launch into Auto Scaling Group](#) ⓘ

You may want to consider launching these instances into an Auto Scaling Group to help you maintain application availability and for easy scaling in the future. [Learn how Auto Scaling can help your application stay healthy and cost effective.](#)

Purchasing option ⓘ ☐ Request Spot instances

Network ⓘ vpc-72e14516 (default) [Create new VPC](#)

Subnet ⓘ No preference (default subnet in any Availability Zone) [Create new subnet](#)

Auto-assign Public IP ⓘ Use subnet setting (Enable)

Placement group ⓘ ☐ Add instance to placement group

Capacity Reservation ⓘ Open [Create new Capacity Reservation](#)

IAM role ⓘ None [Create new IAM role](#)

CPU options ⓘ ☐ Specify CPU options

[Cancel](#) [Previous](#) [Review and Launch](#) 2 [Next: Add Storage](#)

Feedback English (US) © 2008 - 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use

Name new instances as Slave 1 & Slave 2

The screenshot shows the AWS Management Console interface. On the left is a navigation sidebar with categories like EC2 Dashboard, INSTANCES, IMAGES, ELASTIC BLOCK STORE, NETWORK & SECURITY, and LOAD BALANCING. The main content area displays a table of EC2 instances. Below the table, the details for the selected instance 'BDA-Hadoop Master' are shown, including its ID, state, type, availability zone, security groups, and network configuration.

Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Status	Public DNS (IF)
BDA-Hadoop Master	i-0dd57077f57bc13c0	m5.large	us-west-2b	running	Initializing	None	ec2-52-40-209-213.us-west-2.compute.amazonaws.com
BDA-Hadoop Slave 1	i-059f5777f55f8d4c6	m5.large	us-west-2a	running	Initializing	None	ec2-52-13-213-4.us-west-2.compute.amazonaws.com
BDA-Hadoop Slave 2	i-0cc4183bd85a608a2	m5.large	us-west-2a	running	Initializing	None	ec2-34-215-181-4.us-west-2.compute.amazonaws.com
	i-0d35056ebb33ba5e2	t2.large	us-west-2a	stopped		None	

Instance: i-0dd57077f57bc13c0 (BDA-Hadoop Master) **Public DNS:** ec2-52-40-209-213.us-west-2.compute.amazonaws.com

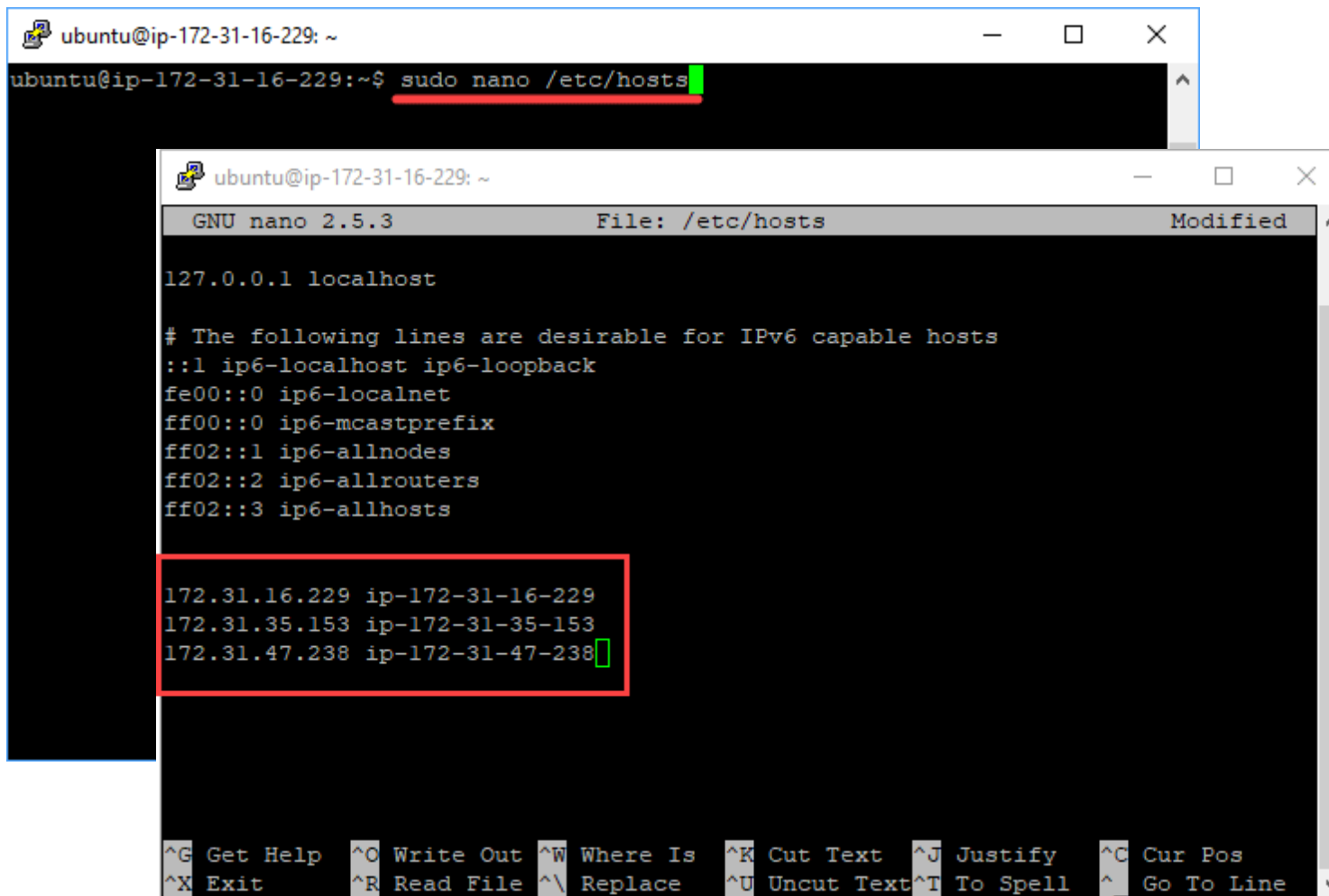
Description | Status Checks | Monitoring | Tags

Instance ID	i-0dd57077f57bc13c0	Public DNS (IPv4)	ec2-52-40-209-213.us-west-2.compute.amazonaws.com
Instance state	running	IPv4 Public IP	52.40.209.213
Instance type	m5.large	IPv6 IPs	-
Elastic IPs		Private DNS	ip-172-31-16-229.us-west-2.compute.internal
Availability zone	us-west-2b	Private IPs	172.31.16.229
Security groups	BDA-SecurityGroup. view inbound rules. view outbound rules	Secondary private IPs	
Scheduled events	No scheduled events	VPC ID	vpc-72e14516
AMI ID	ubuntu/images/hvm-ssd/ubuntu-xenial-16.04-amd64-server-20181114 (ami-076e276d85f524150)	Subnet ID	subnet-256cce41
Platform	-	Network interfaces	eth0
IAM role	-	Source/dest. check	True

Private IPs

Name	Private IP	Private DNS
BDA-Hadoop Master	172.31.16.229	ip-172-31-16-229.us-west-2.compute.internal
BDA-Hadoop Slave 1	172.31.35.153	ip-172-31-35-153.us-west-2.compute.internal
BDA-Hadoop Slave 2	172.31.47.238	ip-172-31-47-238.us-west-2.compute.internal

At master node : Edit host file

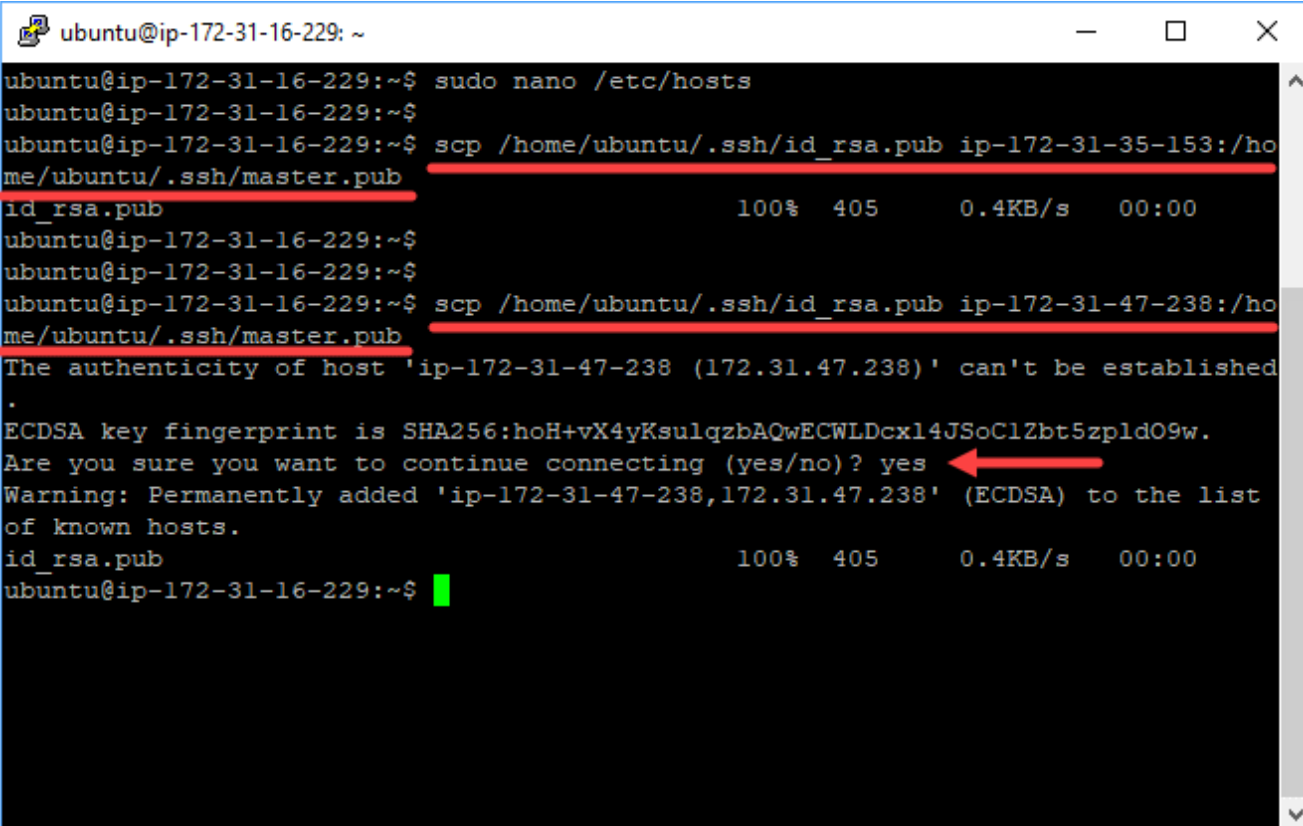


```
ubuntu@ip-172-31-16-229: ~  
ubuntu@ip-172-31-16-229:~$ sudo nano /etc/hosts  
GNU nano 2.5.3 File: /etc/hosts Modified  
127.0.0.1 localhost  
  
# The following lines are desirable for IPv6 capable hosts  
::1 ip6-localhost ip6-loopback  
fe00::0 ip6-localnet  
ff00::0 ip6-mcastprefix  
ff02::1 ip6-allnodes  
ff02::2 ip6-allrouters  
ff02::3 ip6-allhosts  
  
172.31.16.229 ip-172-31-16-229  
172.31.35.153 ip-172-31-35-153  
172.31.47.238 ip-172-31-47-238  
  
^G Get Help ^O Write Out ^W Where Is ^K Cut Text ^J Justify ^C Cur Pos  
^X Exit ^R Read File ^\ Replace ^U Uncut Text ^T To Spell ^_ Go To Line
```

At master node : Copy key file to Slave 1 & Slave 2

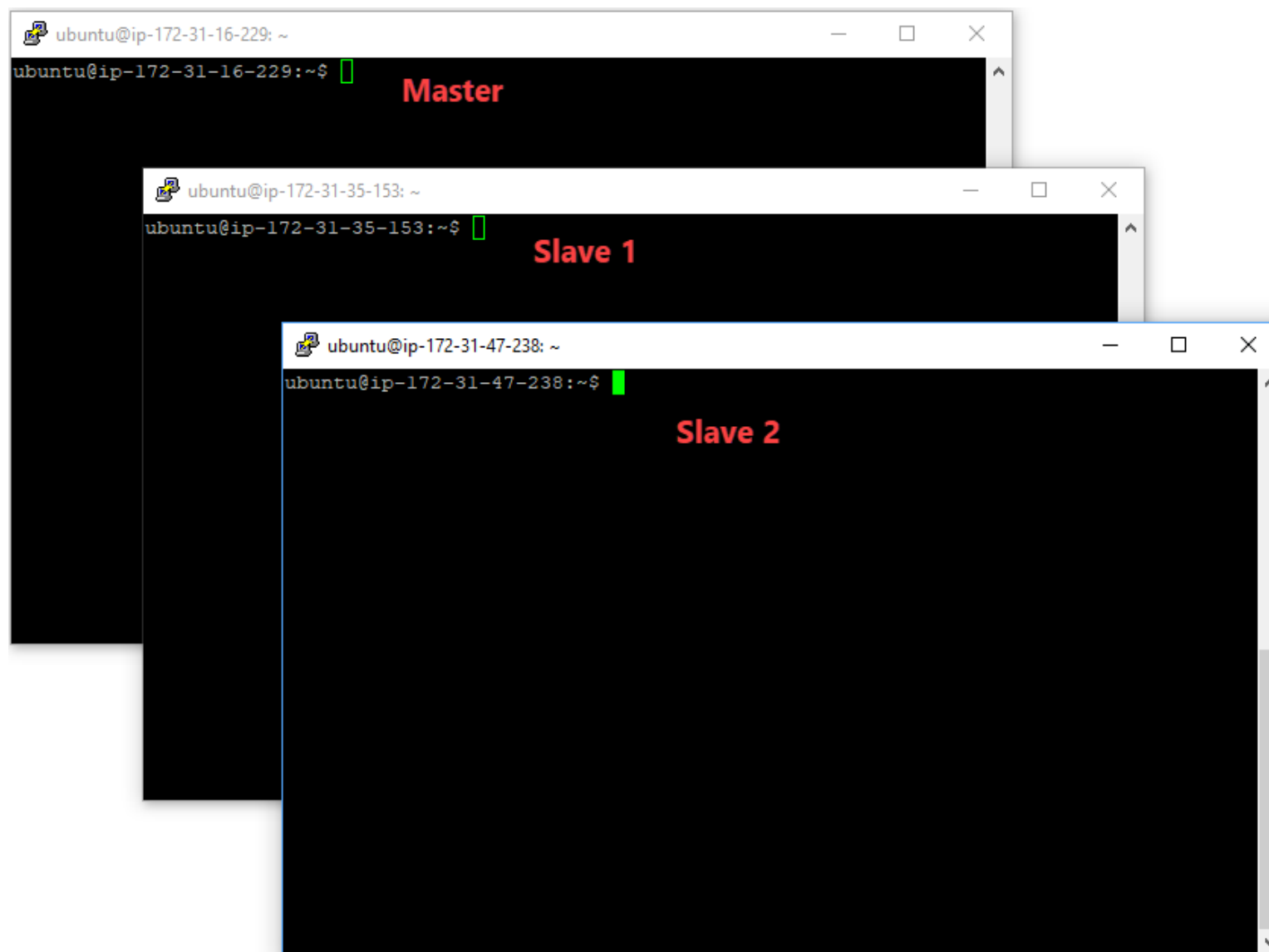
```
$scp /home/ubuntu/.ssh/id_rsa.pub ip-172-31-35-153:/home/ubuntu/.ssh/master.pub
```

```
$scp /home/ubuntu/.ssh/id_rsa.pub ip-172-31-47-238:/home/ubuntu/.ssh/master.pub
```

A terminal window titled 'ubuntu@ip-172-31-16-229: ~' with standard window controls. It shows two successful SCP commands copying 'id_rsa.pub' to slave nodes. The second command is followed by a warning about the host's authenticity, which the user confirms by typing 'yes'.

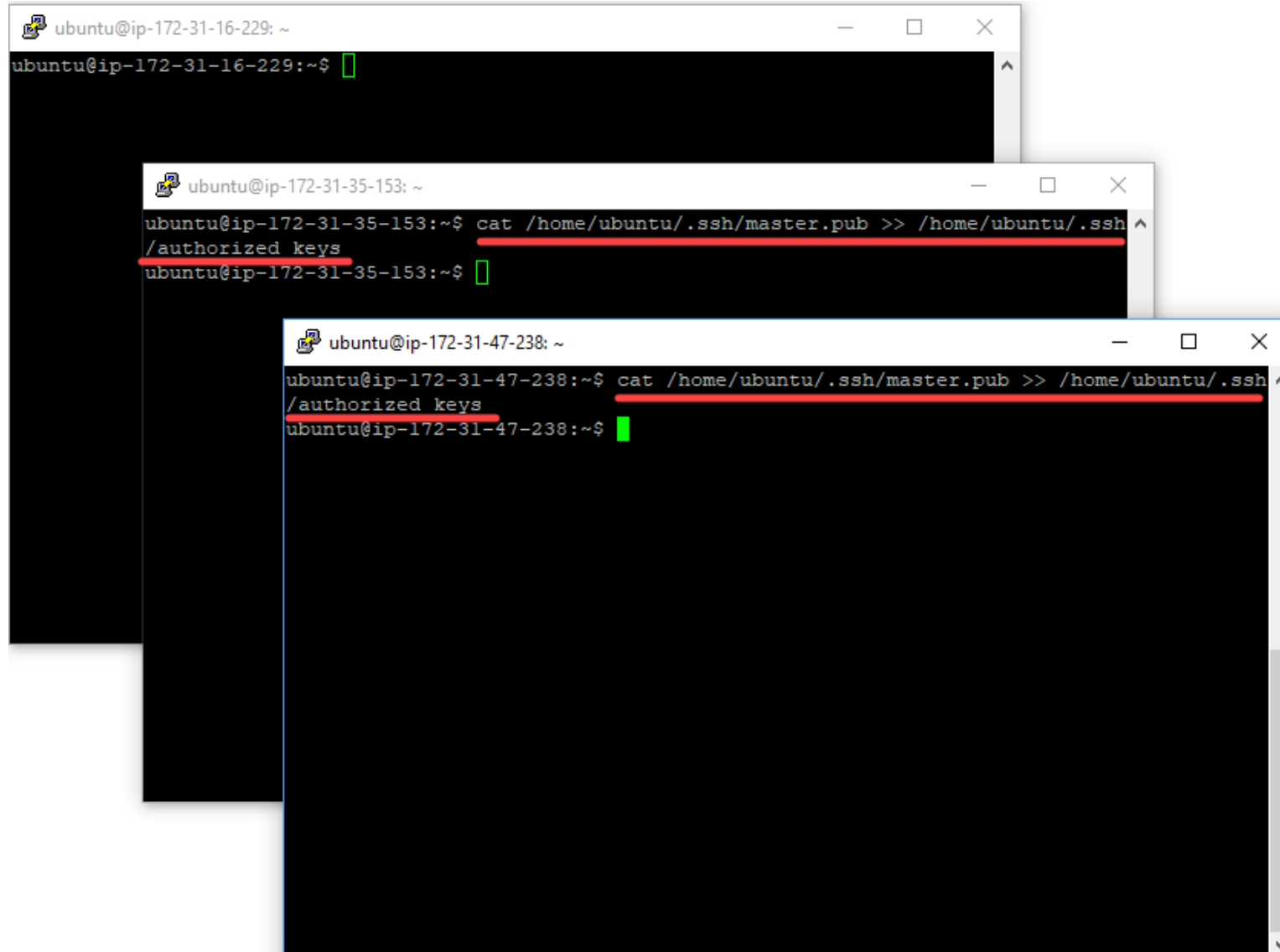
```
ubuntu@ip-172-31-16-229:~$ sudo nano /etc/hosts
ubuntu@ip-172-31-16-229:~$
ubuntu@ip-172-31-16-229:~$ scp /home/ubuntu/.ssh/id_rsa.pub ip-172-31-35-153:/home/ubuntu/.ssh/master.pub
id_rsa.pub                                100% 405      0.4KB/s   00:00
ubuntu@ip-172-31-16-229:~$
ubuntu@ip-172-31-16-229:~$
ubuntu@ip-172-31-16-229:~$ scp /home/ubuntu/.ssh/id_rsa.pub ip-172-31-47-238:/home/ubuntu/.ssh/master.pub
The authenticity of host 'ip-172-31-47-238 (172.31.47.238)' can't be established.
ECDSA key fingerprint is SHA256:hoH+vX4yKsulqzbAQwECWLDcxl4JSocl2bt5zp1dO9w.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'ip-172-31-47-238,172.31.47.238' (ECDSA) to the list of known hosts.
id_rsa.pub                                100% 405      0.4KB/s   00:00
ubuntu@ip-172-31-16-229:~$
```


Arrange SSH terminals



At Slave1 & Slave2 : Append new key to key file

```
$cat /home/ubuntu/.ssh/master.pub >> /home/ubuntu/.ssh/authorized_keys
```



The image displays three overlapping terminal windows, each representing a different slave node in a distributed system. Each window shows the execution of a command to append the contents of the master's public key file to the local authorized_keys file. The command is: `cat /home/ubuntu/.ssh/master.pub >> /home/ubuntu/.ssh/authorized_keys`. The first window (top) is for `ubuntu@ip-172-31-16-229: ~` and shows the command prompt. The second window (middle) is for `ubuntu@ip-172-31-35-153: ~` and shows the command being executed. The third window (bottom) is for `ubuntu@ip-172-31-47-238: ~` and also shows the command being executed. The command is highlighted with a red underline in the middle and bottom windows.

```
ubuntu@ip-172-31-16-229: ~  
ubuntu@ip-172-31-16-229:~$  
  
ubuntu@ip-172-31-35-153: ~  
ubuntu@ip-172-31-35-153:~$ cat /home/ubuntu/.ssh/master.pub >> /home/ubuntu/.ssh/authorized_keys  
ubuntu@ip-172-31-35-153:~$  
  
ubuntu@ip-172-31-47-238: ~  
ubuntu@ip-172-31-47-238:~$ cat /home/ubuntu/.ssh/master.pub >> /home/ubuntu/.ssh/authorized_keys  
ubuntu@ip-172-31-47-238:~$
```

At Master : Test ssh to Slave1 & Slave2

```
$ssh ip-172-31-35-153
```

```
$exit
```

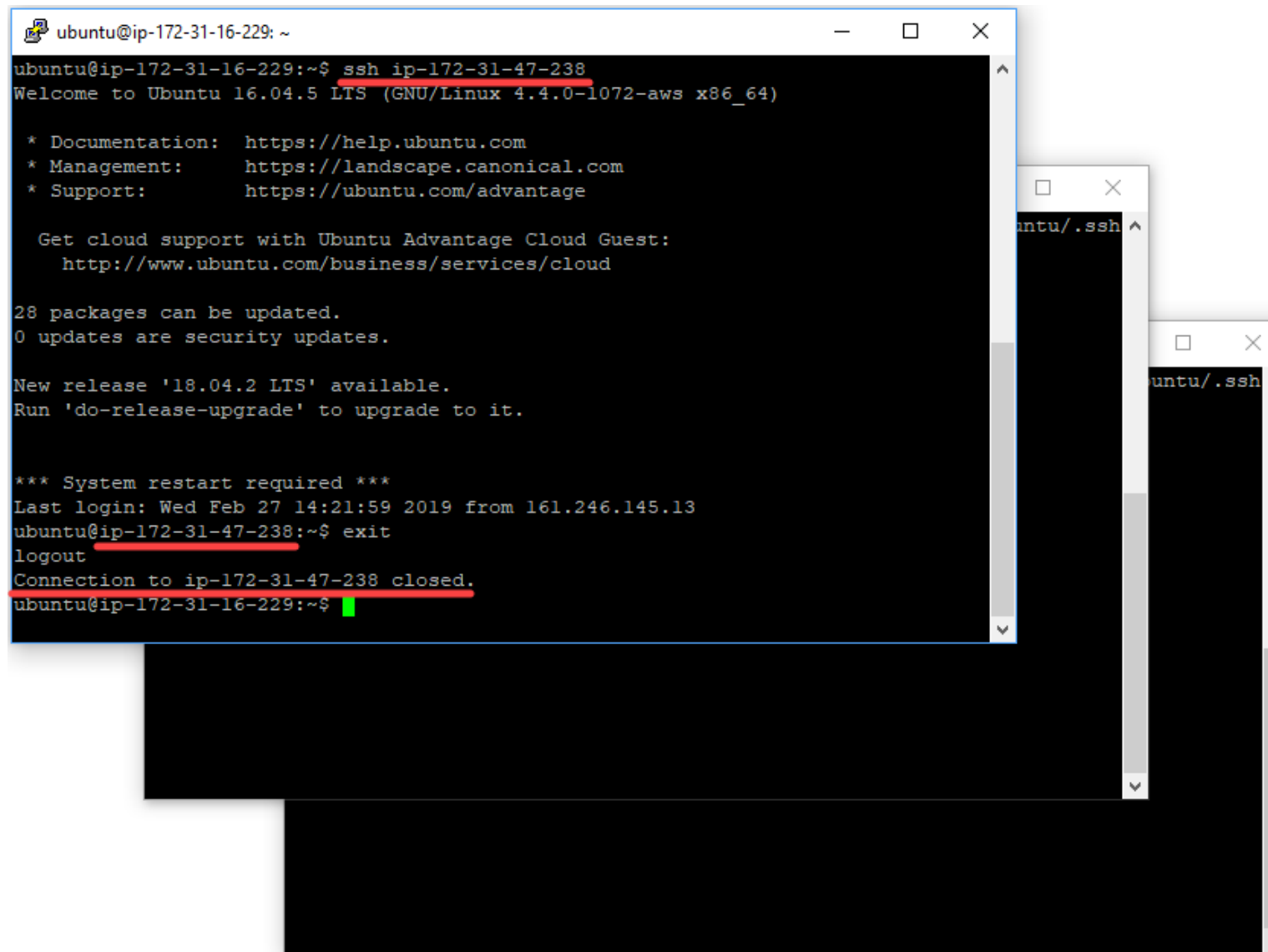
```
$ssh ip-172-31-47-238
```

```
$exit
```

At Master : Test ssh to Slave1

```
ubuntu@ip-172-31-16-229: ~  
ubuntu@ip-172-31-16-229:~$ ssh ip-172-31-35-153  
Welcome to Ubuntu 16.04.5 LTS (GNU/Linux 4.4.0-1072-aws x86_64)  
  
* Documentation:  https://help.ubuntu.com  
* Management:    https://landscape.canonical.com  
* Support:        https://ubuntu.com/advantage  
  
Get cloud support with Ubuntu Advantage Cloud Guest:  
http://www.ubuntu.com/business/services/cloud  
  
28 packages can be updated.  
0 updates are security updates.  
  
New release '18.04.2 LTS' available.  
Run 'do-release-upgrade' to upgrade to it.  
  
*** System restart required ***  
Last login: Wed Feb 27 14:21:18 2019 from 161.246.145.13  
ubuntu@ip-172-31-35-153:~$ exit  
logout  
Connection to ip-172-31-35-153 closed.  
ubuntu@ip-172-31-16-229:~$
```

At Master : Test ssh to Slave2

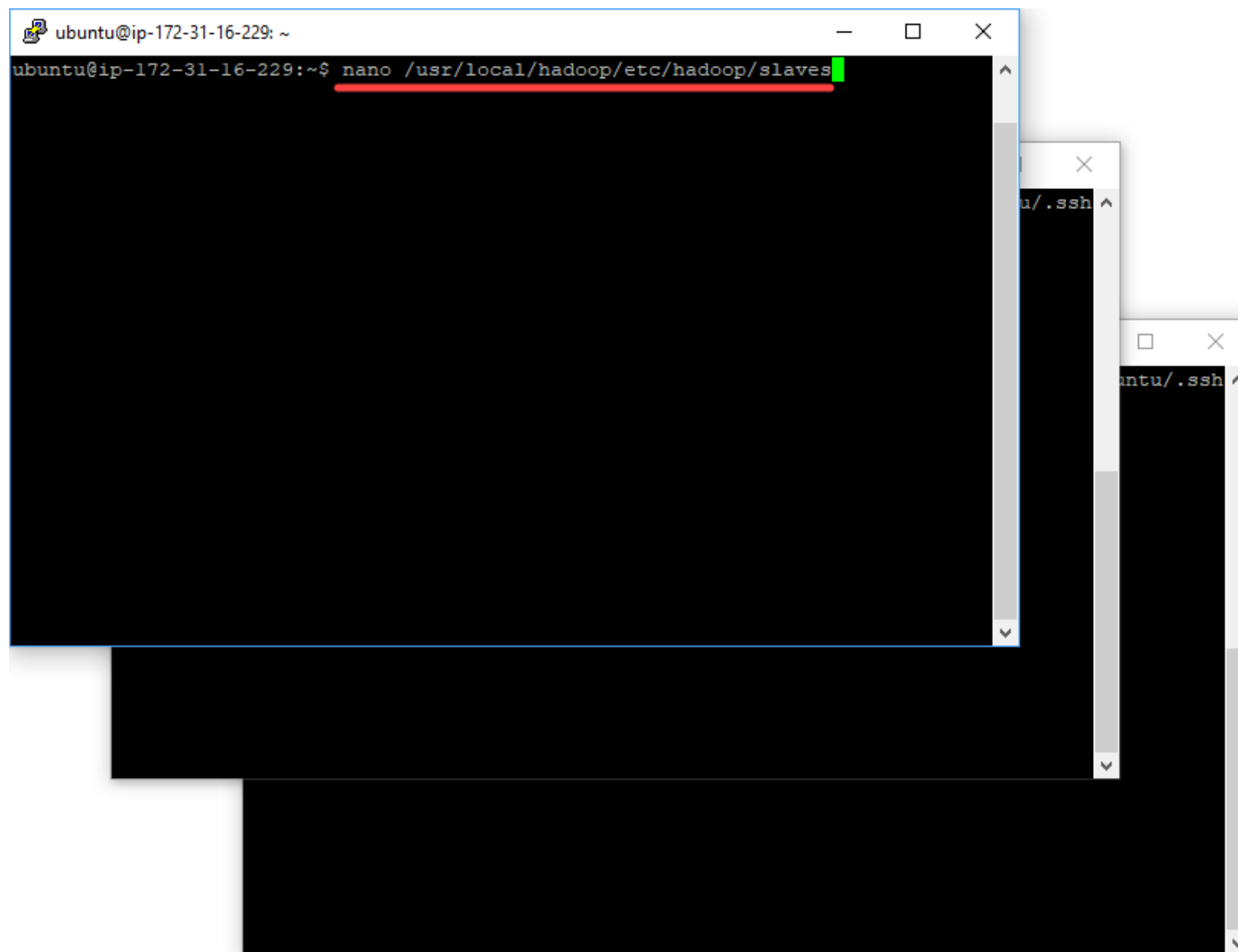


The image shows a terminal window titled 'ubuntu@ip-172-31-16-229: ~'. The user enters the command 'ssh ip-172-31-47-238'. The terminal displays the SSH banner for Ubuntu 16.04.5 LTS, including documentation, management, and support links. It also shows system update information: '28 packages can be updated. 0 updates are security updates. New release '18.04.2 LTS' available. Run 'do-release-upgrade' to upgrade to it.' followed by a system restart requirement. The user then enters 'exit', and the terminal shows 'logout' and 'Connection to ip-172-31-47-238 closed.' before returning to the prompt 'ubuntu@ip-172-31-16-229:~\$' with a green cursor. Two other terminal windows are visible in the background, both showing 'ubuntu/.ssh'.

```
ubuntu@ip-172-31-16-229: ~  
ubuntu@ip-172-31-16-229:~$ ssh ip-172-31-47-238  
Welcome to Ubuntu 16.04.5 LTS (GNU/Linux 4.4.0-1072-aws x86_64)  
  
* Documentation:  https://help.ubuntu.com  
* Management:    https://landscape.canonical.com  
* Support:       https://ubuntu.com/advantage  
  
Get cloud support with Ubuntu Advantage Cloud Guest:  
http://www.ubuntu.com/business/services/cloud  
  
28 packages can be updated.  
0 updates are security updates.  
  
New release '18.04.2 LTS' available.  
Run 'do-release-upgrade' to upgrade to it.  
  
*** System restart required ***  
Last login: Wed Feb 27 14:21:59 2019 from 161.246.145.13  
ubuntu@ip-172-31-47-238:~$ exit  
logout  
Connection to ip-172-31-47-238 closed.  
ubuntu@ip-172-31-16-229:~$
```

At Master : Add Slave1 & Slave2 to Hadoop slave file

```
$nano /usr/local/hadoop/etc/hadoop/slaves
```



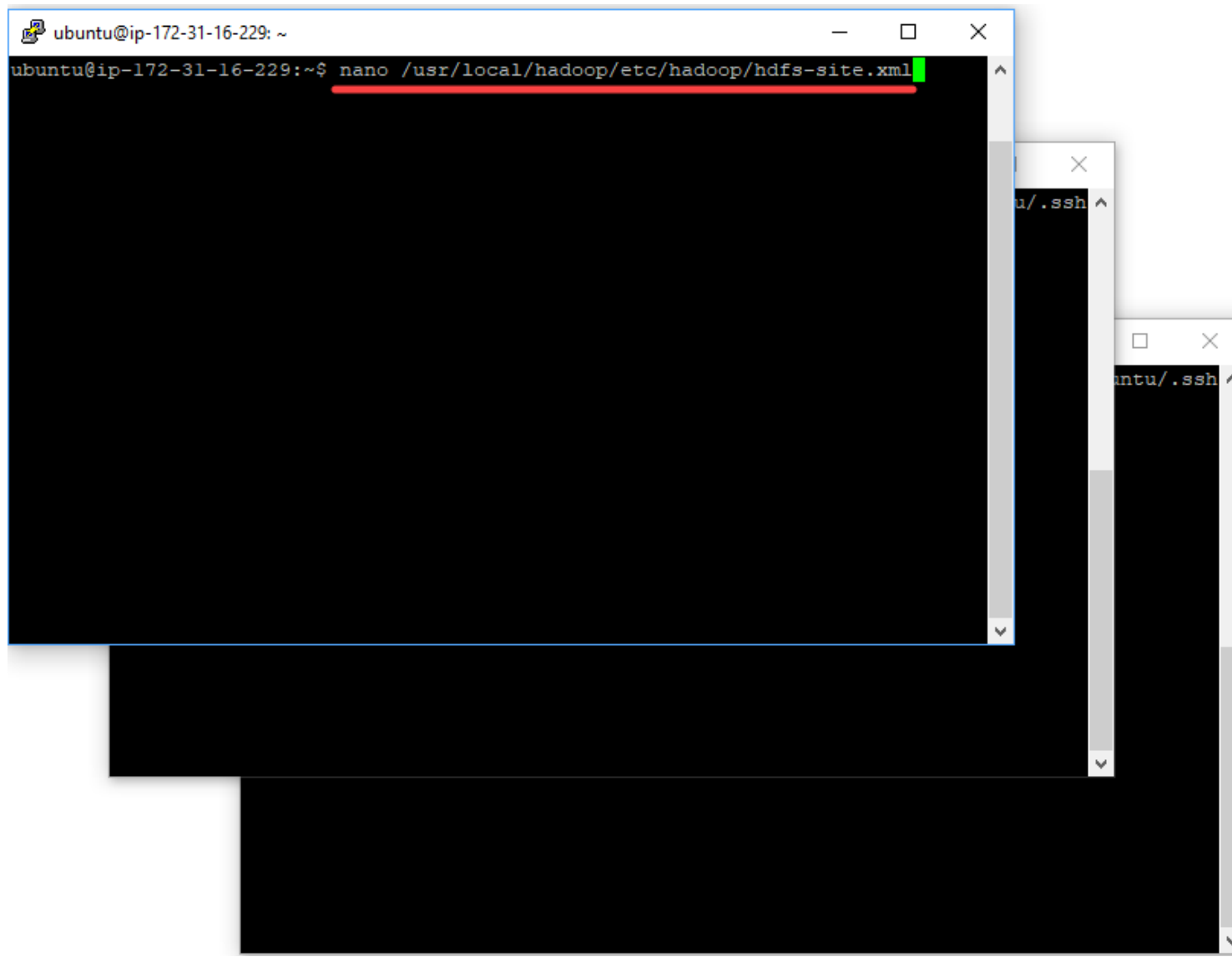
At Master : Add Slave1 & Slave2 to Hadoop slave file

Add slave private DNS

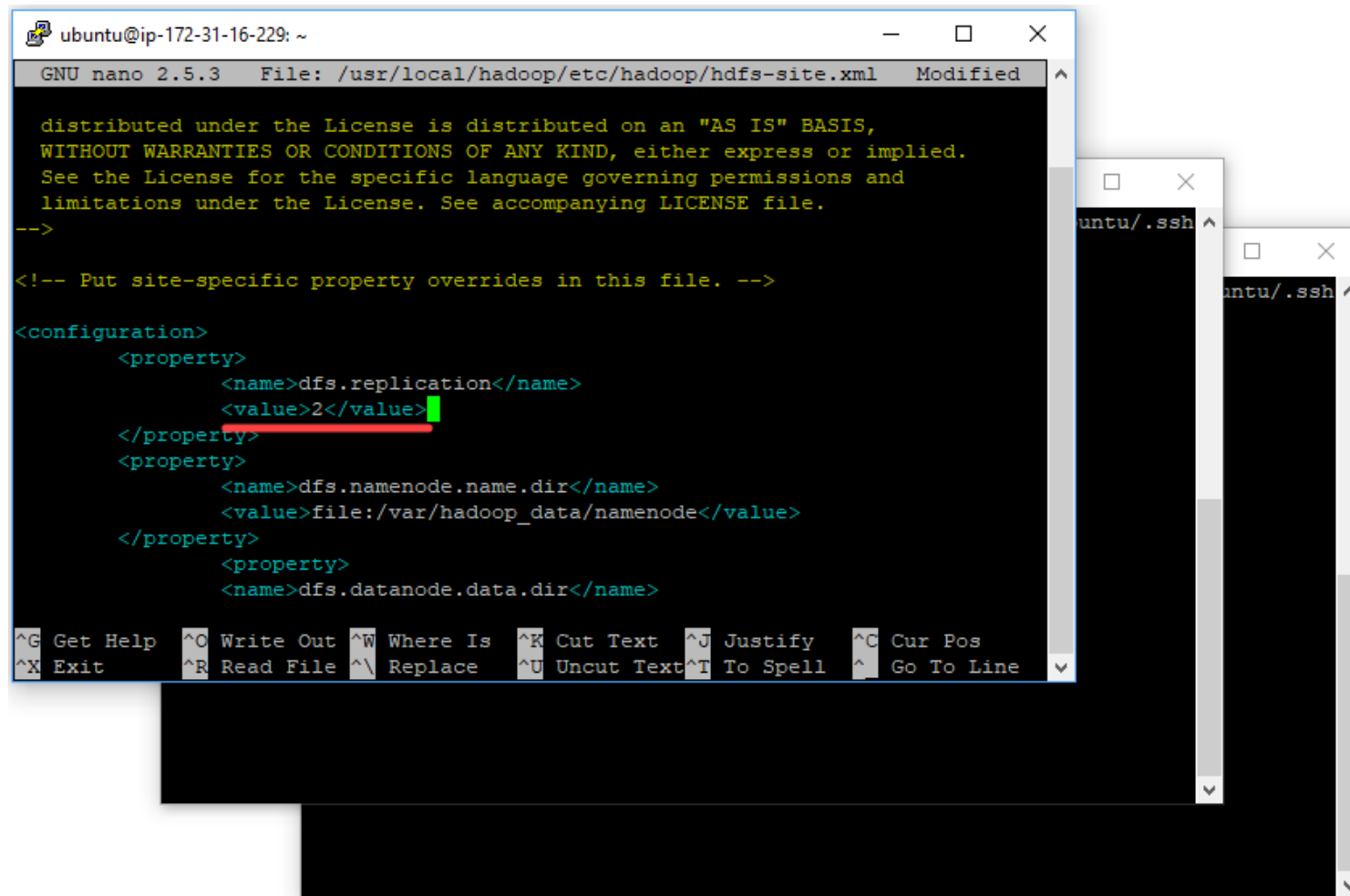
```
ubuntu@ip-172-31-16-229: ~  
GNU nano 2.5.3 File: /usr/local/hadoop/etc/hadoop/slaves Modified  
ip-172-31-35-153  
ip-172-31-47-238  
^G Get Help ^O Write Out ^W Where Is ^K Cut Text ^J Justify ^C Cur Pos  
^X Exit ^R Read File ^\ Replace ^U Uncut Text ^T To Spell ^_ Go To Line
```

At Master : Edit hdfs-site.xml

```
$nano /usr/local/hadoop/etc/hadoop/hdfs-site.xml
```



At Master : Edit replication to 2



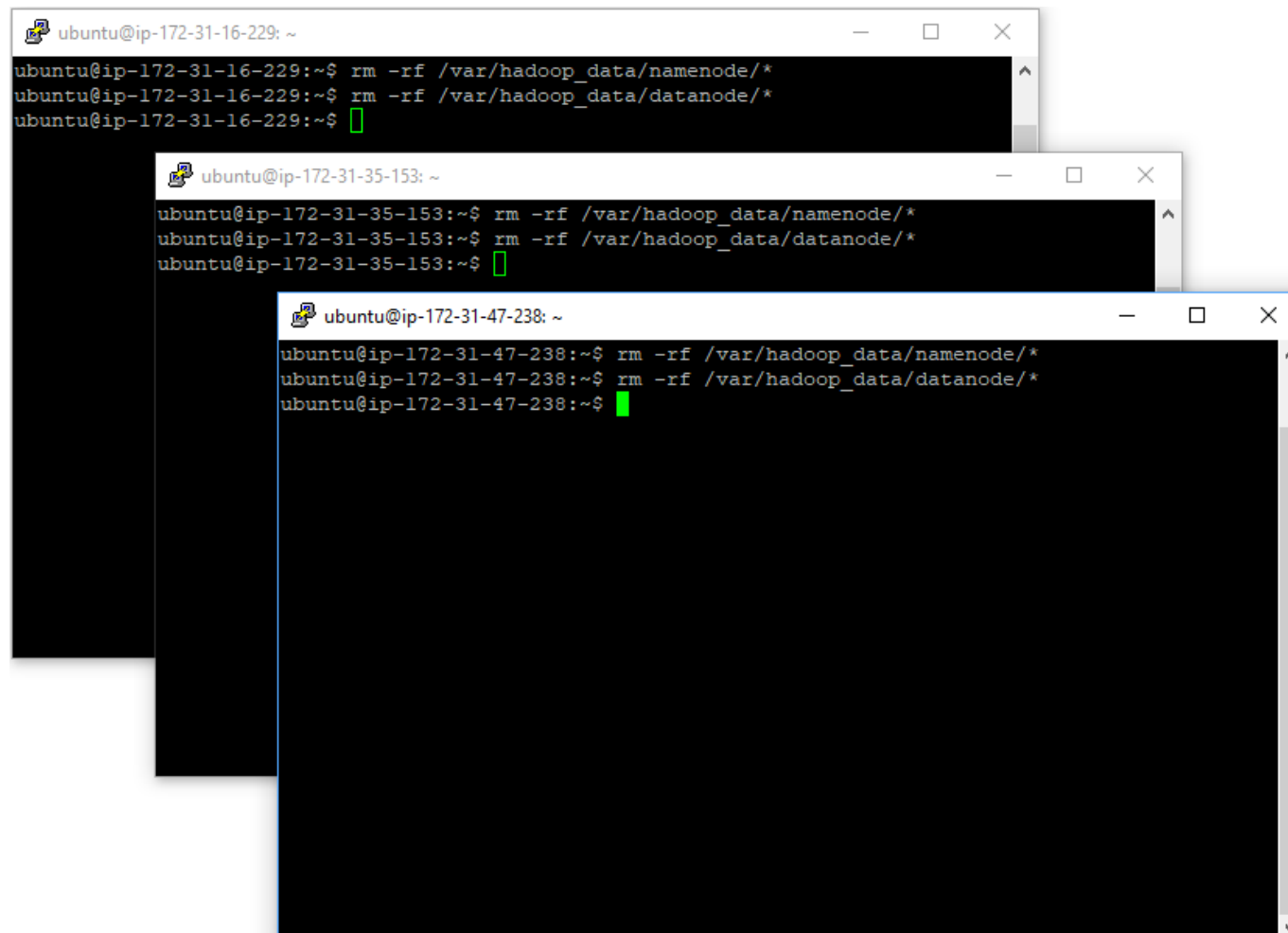
```
ubuntu@ip-172-31-16-229: ~
GNU nano 2.5.3 File: /usr/local/hadoop/etc/hadoop/hdfs-site.xml Modified
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->
<!-- Put site-specific property overrides in this file. -->
<configuration>
  <property>
    <name>dfs.replication</name>
    <value>2</value>
  </property>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>file:/var/hadoop_data/namenode</value>
  </property>
  <property>
    <name>dfs.datanode.data.dir</name>
  </property>

```

At all nodes : Remove directories of namenode and datanode

```
$rm -rf /var/hadoop_data/namenode/*
```

```
$rm -rf /var/hadoop_data/datanode/*
```

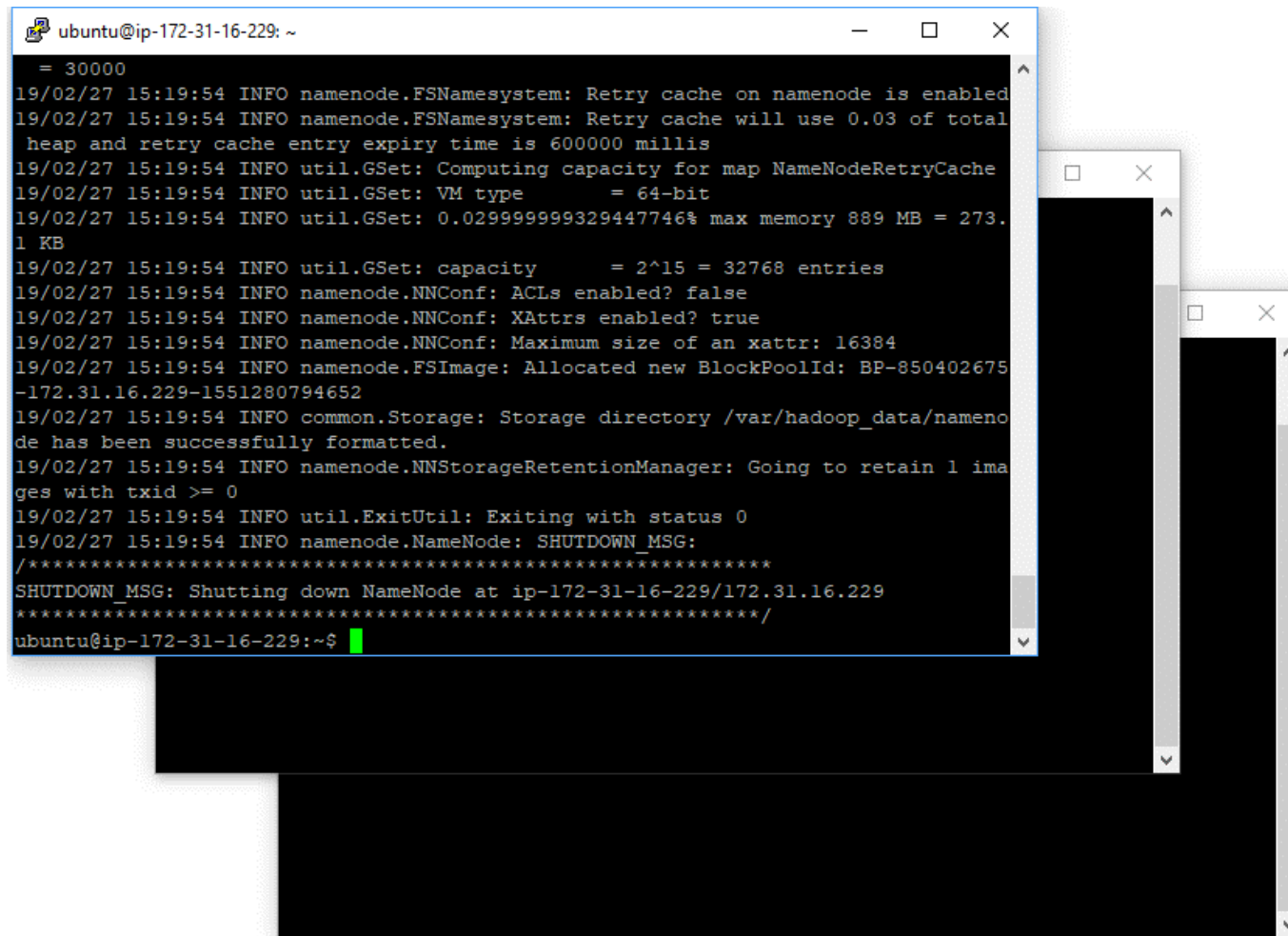


The image displays three overlapping terminal windows, each representing a different node in a Hadoop cluster. Each window shows the execution of two commands to remove Hadoop data directories. The top window is for node ip-172-31-16-229, the middle for ip-172-31-35-153, and the bottom for ip-172-31-47-238. All three nodes show successful execution of the commands, with a green cursor visible on the final line of each terminal.

```
ubuntu@ip-172-31-16-229: ~  
ubuntu@ip-172-31-16-229:~$ rm -rf /var/hadoop_data/namenode/*  
ubuntu@ip-172-31-16-229:~$ rm -rf /var/hadoop_data/datanode/*  
ubuntu@ip-172-31-16-229:~$   
  
ubuntu@ip-172-31-35-153: ~  
ubuntu@ip-172-31-35-153:~$ rm -rf /var/hadoop_data/namenode/*  
ubuntu@ip-172-31-35-153:~$ rm -rf /var/hadoop_data/datanode/*  
ubuntu@ip-172-31-35-153:~$   
  
ubuntu@ip-172-31-47-238: ~  
ubuntu@ip-172-31-47-238:~$ rm -rf /var/hadoop_data/namenode/*  
ubuntu@ip-172-31-47-238:~$ rm -rf /var/hadoop_data/datanode/*  
ubuntu@ip-172-31-47-238:~$
```

At Master: Format namenode

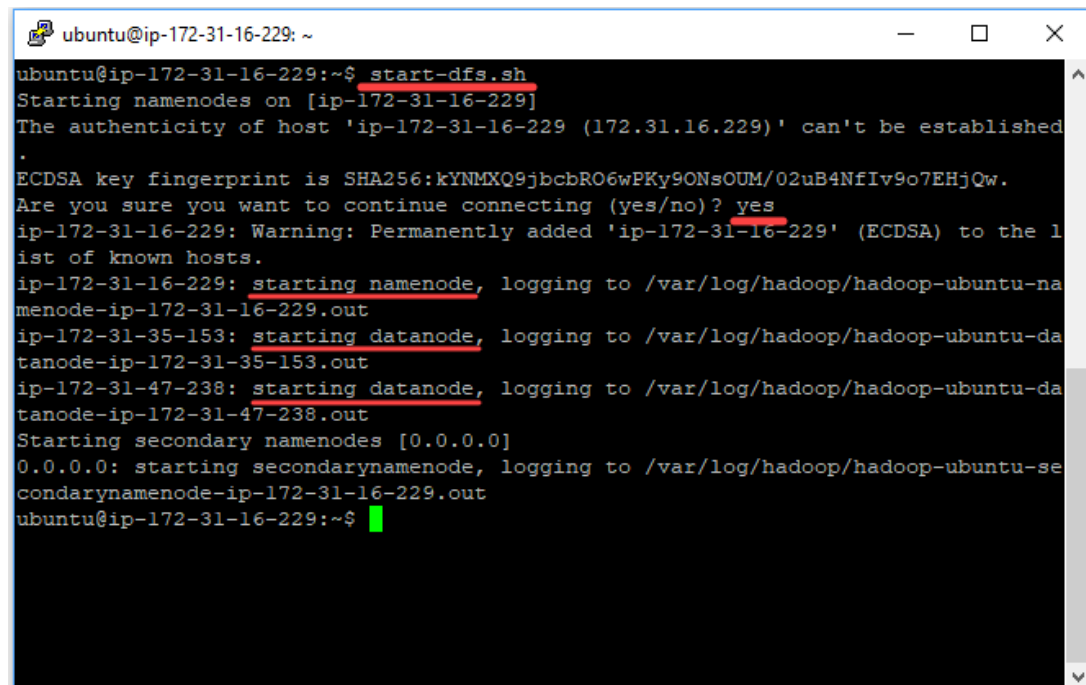
\$hdfs namenode -format



```
ubuntu@ip-172-31-16-229: ~  
= 30000  
19/02/27 15:19:54 INFO namenode.FSNamesystem: Retry cache on namenode is enabled  
19/02/27 15:19:54 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total  
heap and retry cache entry expiry time is 600000 millis  
19/02/27 15:19:54 INFO util.GSet: Computing capacity for map NameNodeRetryCache  
19/02/27 15:19:54 INFO util.GSet: VM type = 64-bit  
19/02/27 15:19:54 INFO util.GSet: 0.0299999999329447746% max memory 889 MB = 273.  
1 KB  
19/02/27 15:19:54 INFO util.GSet: capacity = 2^15 = 32768 entries  
19/02/27 15:19:54 INFO namenode.NNConf: ACLs enabled? false  
19/02/27 15:19:54 INFO namenode.NNConf: XAttrs enabled? true  
19/02/27 15:19:54 INFO namenode.NNConf: Maximum size of an xattr: 16384  
19/02/27 15:19:54 INFO namenode.FSImage: Allocated new BlockPoolId: BP-850402675-  
172.31.16.229-1551280794652  
19/02/27 15:19:54 INFO common.Storage: Storage directory /var/hadoop_data/nameno  
de has been successfully formatted.  
19/02/27 15:19:54 INFO namenode.NNStorageRetentionManager: Going to retain 1 ima  
ges with txid >= 0  
19/02/27 15:19:54 INFO util.ExitUtil: Exiting with status 0  
19/02/27 15:19:54 INFO namenode.NameNode: SHUTDOWN_MSG:  
/*****  
SHUTDOWN_MSG: Shutting down NameNode at ip-172-31-16-229/172.31.16.229  
*****/  
ubuntu@ip-172-31-16-229:~$
```

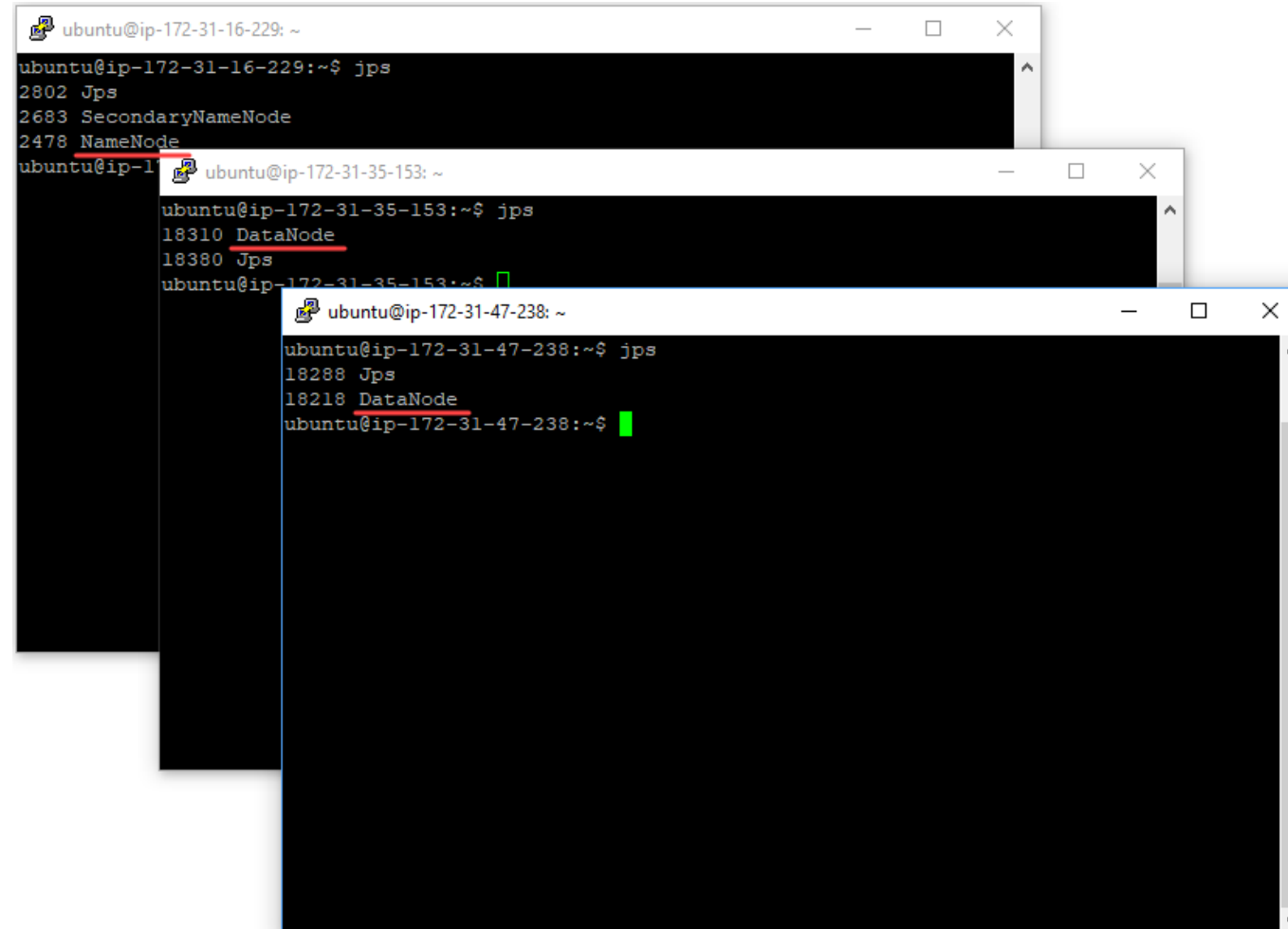
At Master: Execute start-dfs.sh

`$start-dfs.sh`

A terminal window titled 'ubuntu@ip-172-31-16-229: ~' showing the execution of the 'start-dfs.sh' script. The script starts by attempting to start namenodes on the local host, which results in a warning about host authenticity. It then proceeds to start datanodes on three other hosts (ip-172-31-35-153 and ip-172-31-47-238) and secondary namenodes on the local host. The output shows logging for each node and the final state of the cluster.

```
ubuntu@ip-172-31-16-229: ~  
ubuntu@ip-172-31-16-229:~$ start-dfs.sh  
Starting namenodes on [ip-172-31-16-229]  
The authenticity of host 'ip-172-31-16-229 (172.31.16.229)' can't be established  
.  
ECDSA key fingerprint is SHA256:kYNMXQ9jbcbrO6wPKy9ONsOUM/02uB4NfIv9o7EHjQw.  
Are you sure you want to continue connecting (yes/no)? yes  
ip-172-31-16-229: Warning: Permanently added 'ip-172-31-16-229' (ECDSA) to the 1  
ist of known hosts.  
ip-172-31-16-229: starting namenode, logging to /var/log/hadoop/hadoop-ubuntu-na  
menode-ip-172-31-16-229.out  
ip-172-31-35-153: starting datanode, logging to /var/log/hadoop/hadoop-ubuntu-da  
tanode-ip-172-31-35-153.out  
ip-172-31-47-238: starting datanode, logging to /var/log/hadoop/hadoop-ubuntu-da  
tanode-ip-172-31-47-238.out  
Starting secondary namenodes [0.0.0.0]  
0.0.0.0: starting secondarynamenode, logging to /var/log/hadoop/hadoop-ubuntu-se  
condarynamenode-ip-172-31-16-229.out  
ubuntu@ip-172-31-16-229:~$
```

At all nodes: Use `jps` to check result, should see NameNode started on Master and DataNode started on both Slave1&Slave2



The image displays three overlapping terminal windows, each showing the output of the `jps` command on a different node. The windows are titled with their respective IP addresses: `ubuntu@ip-172-31-16-229: ~`, `ubuntu@ip-172-31-35-153: ~`, and `ubuntu@ip-172-31-47-238: ~`.

- The first window (Master node) shows:

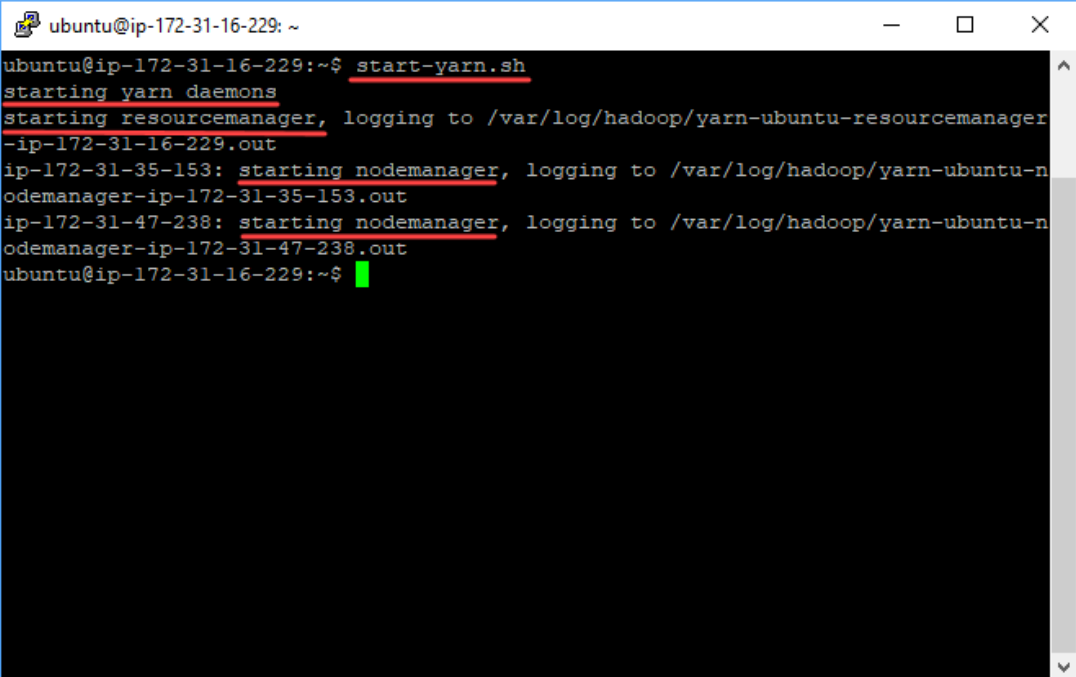
```
ubuntu@ip-172-31-16-229:~$ jps
2802 Jps
2683 SecondaryNameNode
2478 NameNode
```
- The second window (Slave1 node) shows:

```
ubuntu@ip-172-31-35-153:~$ jps
18310 DataNode
18380 Jps
```
- The third window (Slave2 node) shows:

```
ubuntu@ip-172-31-47-238:~$ jps
18288 Jps
18218 DataNode
```

At Master: Execute start-yarn.sh

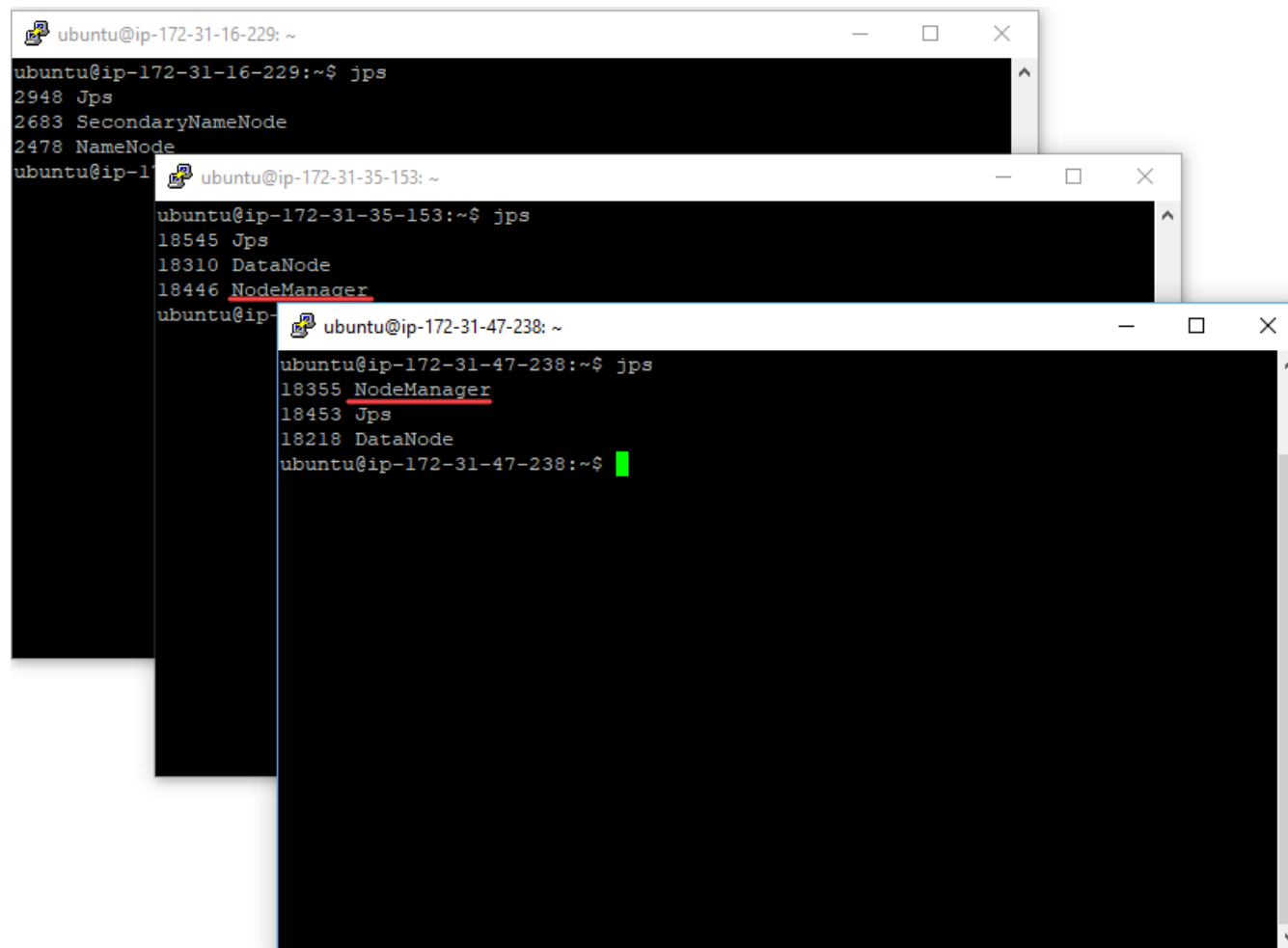
`$start-yarn.sh`



```
ubuntu@ip-172-31-16-229: ~  
ubuntu@ip-172-31-16-229:~$ start-yarn.sh  
starting yarn daemons  
starting resourcemanager, logging to /var/log/hadoop/yarn-ubuntu-resourcemanager-  
-ip-172-31-16-229.out  
ip-172-31-35-153: starting nodemanager, logging to /var/log/hadoop/yarn-ubuntu-n  
odemanager-ip-172-31-35-153.out  
ip-172-31-47-238: starting nodemanager, logging to /var/log/hadoop/yarn-ubuntu-n  
odemanager-ip-172-31-47-238.out  
ubuntu@ip-172-31-16-229:~$
```

The image shows a terminal window on a master node (ip-172-31-16-229) where the command `start-yarn.sh` has been executed. The output shows the initiation of YARN daemons, including the Resource Manager and Node Managers on slave nodes. The terminal window is titled 'ubuntu@ip-172-31-16-229: ~' and has standard window controls. Two other terminal windows are visible in the background, both of which are blacked out.

At all nodes: Use `jps` to check result, should see NameManager started on both Slave1&Slave2



```
ubuntu@ip-172-31-16-229: ~  
ubuntu@ip-172-31-16-229:~$ jps  
2948 Jps  
2683 SecondaryNameNode  
2478 NameNode  
ubuntu@ip-172-31-35-153: ~  
ubuntu@ip-172-31-35-153:~$ jps  
18545 Jps  
18310 DataNode  
18446 NodeManager  
ubuntu@ip-172-31-47-238: ~  
ubuntu@ip-172-31-47-238:~$ jps  
18355 NodeManager  
18453 Jps  
18218 DataNode  
ubuntu@ip-172-31-47-238:~$
```

Namenode Information:

<http://52.26.15.54:50070>

Cluster ID:	CID-3cc3d898-c601-4111-a36e-21754d6cf52d
Block Pool ID:	BP-850402675-172.31.16.229-1551280794652

Summary

Security is off.

Safemode is off.

4 files and directories, 1 blocks = 5 total filesystem object(s).

Heap Memory used 122.49 MB of 207.5 MB Heap Memory. Max Heap Memory is 889 MB.

Non Heap Memory used 39.93 MB of 40.69 MB Committed Non Heap Memory. Max Non Heap Memory is -1 B.

Configured Capacity:	38.65 GB
DFS Used:	48 KB
Non DFS Used:	4.74 GB
DFS Remaining:	33.91 GB
DFS Used%:	0%
DFS Remaining%:	87.74%
Block Pool Used:	48 KB
Block Pool Used%:	0%
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	2 (Decommissioned: 0)
Dead Nodes	0 (Decommissioned: 0)
Decommissioning Nodes	0

Datanode Information

EC2 Management Console | Namenode information | Problem loading page

ec2-52-40-209-213.us-west-2.compute

Hadoop Overview **Datanodes** Snapshot Startup Progress Utilities

Datanode Information

In operation

Node	Last contact	Admin State	Capacity	Used	Non DFS Used	Remaining	Blocks	Block pool used	Failed Volumes	Version
ip-172-31-35-153.us-west-2.compute.internal (172.31.35.153:50010)	2	In Service	19.32 GB	24 KB	2.37 GB	16.96 GB	0	24 KB (0%)	0	2.6.0
ip-172-31-47-238.us-west-2.compute.internal (172.31.47.238:50010)	2	In Service	19.32 GB	24 KB	2.37 GB	16.96 GB	0	24 KB (0%)	0	2.6.0

Decommissioning

Node	Last contact	Under replicated blocks	Blocks with no live replicas	Under Replicated Blocks in files under construction
------	--------------	-------------------------	------------------------------	---

Hadoop, 2014. Legacy UI

Log Files



HadoopOverviewDatanodesSnapshotStartup ProgressUtilities1

Browse the file system2
Logs

Browse Directory

/

Permission	Owner	Group
drwxr-xr-x	ubuntu	supergroup
drwxr-xr-x	ubuntu	supergroup

Hadoop, 2014.

ec2-52-40-209-213.us-west-2.compute.amazonaws.com:50070/logs

Directory: /logs/

SecurityAuth-ubuntu.audit	0 bytes	Feb 13, 2019 9:38:15 PM
hadoop-ubuntu-datanode-ip-172-31-16-229.log	80064 bytes	Feb 20, 2019 6:35:20 PM
hadoop-ubuntu-datanode-ip-172-31-16-229.out	3024 bytes	Feb 20, 2019 1:26:00 PM
hadoop-ubuntu-datanode-ip-172-31-16-229.out.1	718 bytes	Feb 13, 2019 9:38:20 PM
hadoop-ubuntu-namenode-ip-172-31-16-229.log	407331 bytes	Feb 27, 2019 4:11:26 PM
hadoop-ubuntu-namenode-ip-172-31-16-229.out	4908 bytes	Feb 27, 2019 4:11:23 PM
hadoop-ubuntu-namenode-ip-172-31-16-229.out.1	4908 bytes	Feb 20, 2019 1:24:36 PM
hadoop-ubuntu-namenode-ip-172-31-16-229.out.2	4908 bytes	Feb 13, 2019 9:53:23 PM
hadoop-ubuntu-secondarynamenode-ip-172-31-16-229.log	79546 bytes	Feb 27, 2019 3:24:40 PM
hadoop-ubuntu-secondarynamenode-ip-172-31-16-229.out	718 bytes	Feb 27, 2019 3:23:39 PM
hadoop-ubuntu-secondarynamenode-ip-172-31-16-229.out.1	718 bytes	Feb 20, 2019 12:54:18 PM
hadoop-ubuntu-secondarynamenode-ip-172-31-16-229.out.2	718 bytes	Feb 13, 2019 9:38:33 PM
userlogs/	4096 bytes	Feb 20, 2019 1:14:59 PM
yarn-ubuntu-nodemanager-ip-172-31-16-229.log	285816 bytes	Feb 20, 2019 1:15:13 PM
yarn-ubuntu-nodemanager-ip-172-31-16-229.out	702 bytes	Feb 20, 2019 12:54:59 PM
yarn-ubuntu-nodemanager-ip-172-31-16-229.out.1	702 bytes	Feb 13, 2019 10:47:46 PM
yarn-ubuntu-nodemanager-ip-172-31-16-229.out.2	702 bytes	Feb 13, 2019 9:43:43 PM
yarn-ubuntu-resourcemanager-ip-172-31-16-229.log	76232 bytes	Feb 27, 2019 3:30:40 PM
yarn-ubuntu-resourcemanager-ip-172-31-16-229.out	702 bytes	Feb 27, 2019 3:30:40 PM
yarn-ubuntu-resourcemanager-ip-172-31-16-229.out.1	702 bytes	Feb 20, 2019 12:54:58 PM
yarn-ubuntu-resourcemanager-ip-172-31-16-229.out.2	702 bytes	Feb 13, 2019 10:47:45 PM
yarn-ubuntu-resourcemanager-ip-172-31-16-229.out.3	702 bytes	Feb 13, 2019 9:43:42 PM

Next Steps

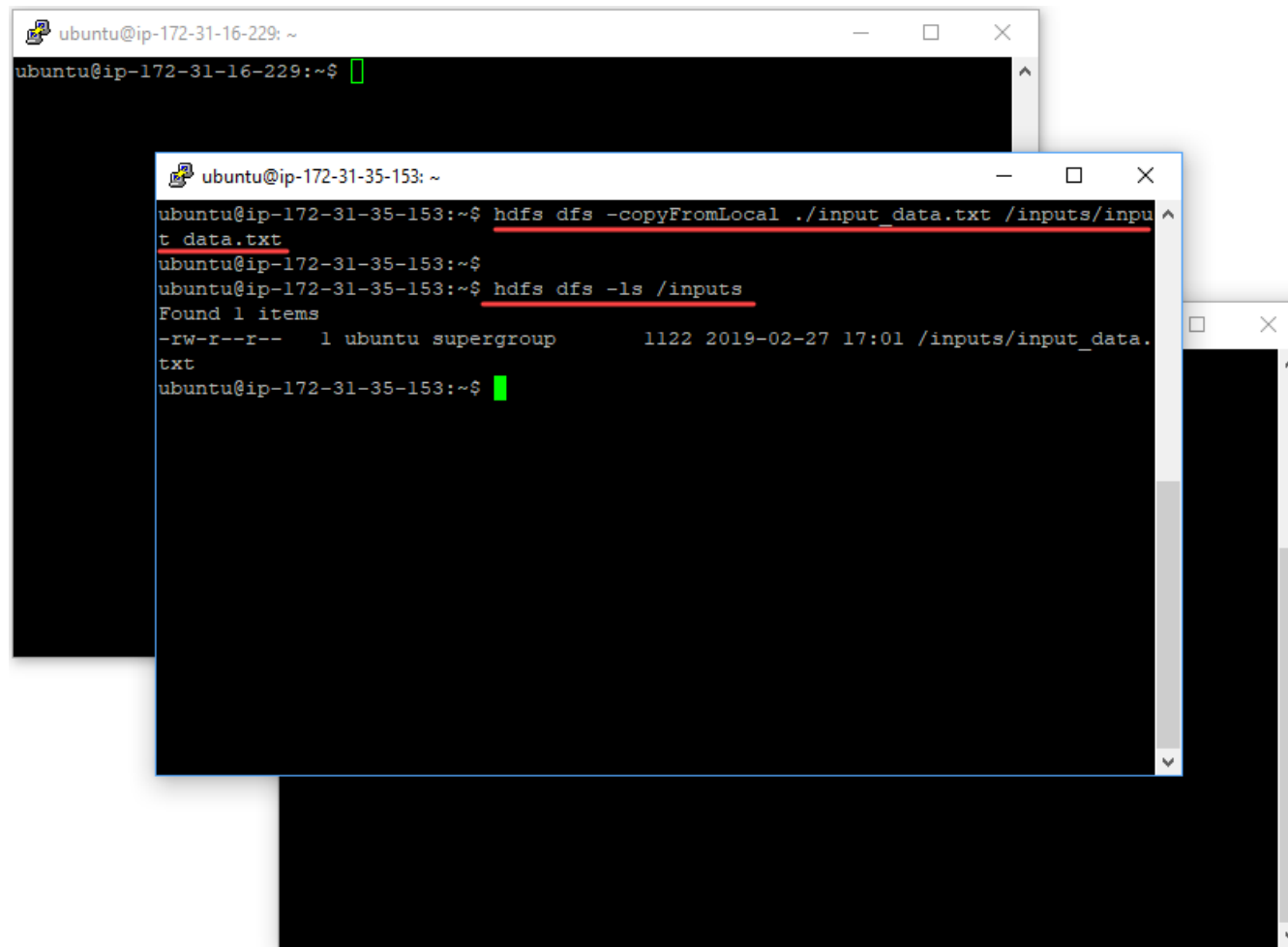
- Import data to Hadoop cluster
- Execute MapReduce
- Compare result
- Stop Yarn & Stop DFS

Use the same commands as single node Hadoop

At Slave1: Import data to Hadoop cluster

```
$hdfs dfs -copyFromLocal ./input_data.txt /inputs/input_data.txt
```

```
$hdfs dfs -ls /inputs
```



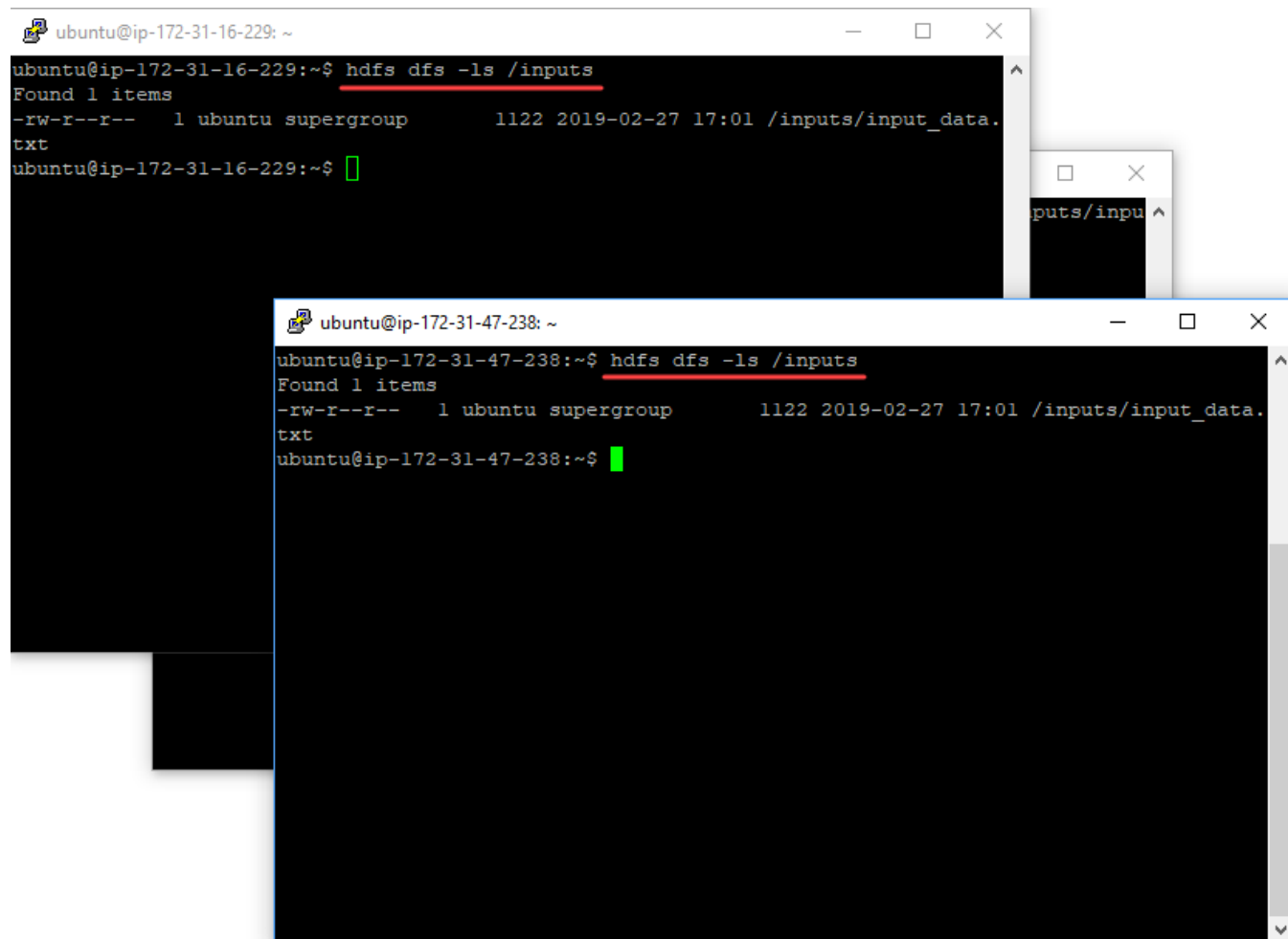
The image shows three overlapping terminal windows. The top window is titled 'ubuntu@ip-172-31-16-229: ~' and shows a prompt. The middle window is titled 'ubuntu@ip-172-31-35-153: ~' and shows the following commands and output:

```
ubuntu@ip-172-31-35-153:~$ hdfs dfs -copyFromLocal ./input_data.txt /inputs/input_data.txt
ubuntu@ip-172-31-35-153:~$ hdfs dfs -ls /inputs
Found 1 items
-rw-r--r--  1 ubuntu supergroup      1122 2019-02-27 17:01 /inputs/input_data.txt
ubuntu@ip-172-31-35-153:~$
```

The bottom window is partially visible and shows a prompt.

At Master & Slave2: Check imported file

```
$hdfs dfs -ls /inputs
```



```
ubuntu@ip-172-31-16-229: ~  
ubuntu@ip-172-31-16-229:~$ hdfs dfs -ls /inputs  
Found 1 items  
-rw-r--r--    1 ubuntu supergroup      1122 2019-02-27 17:01 /inputs/input_data.  
txt  
ubuntu@ip-172-31-16-229:~$  
  
ubuntu@ip-172-31-47-238: ~  
ubuntu@ip-172-31-47-238:~$ hdfs dfs -ls /inputs  
Found 1 items  
-rw-r--r--    1 ubuntu supergroup      1122 2019-02-27 17:01 /inputs/input_data.  
txt  
ubuntu@ip-172-31-47-238:~$
```

At Slave1 & Slave2: Use `jps` to see Application Master and Yarn Child Container

```
ubuntu@ip-172-31-4-165: ~  
Bytes Read=437  
File Output Format Counters  
Bytes Written=513  
ubuntu@ip-172-31-4-165:~$ yarn jar ./wordcount.jar WordCount /inputs/* /outputs/  
wordcount output dir=00  
15/04/17 18: ubuntu@ip-172-31-1-8: ~  
1-4-165/172.  
15/04/17 18: ubuntu@ip-172-31-1-8:~$ jps  
1869 NodeManager  
2240 Jps  
1720 DataNode  
15/04/17 18: ubuntu@ip-172-31-1-8:~$ jps  
2252 Jps  
1869 NodeManager  
29295556906 1720 DataNode  
15/04/17 18: ubuntu@ip-172-31-1-8:~$ jps  
29295556906 2299 YarnChild  
15/04/17 18: 2309 Jps  
-4-165:8088/  
15/04/17 18: 1869 NodeManager  
15/04/17 18: 1720 DataNode  
15/04/17 18: ubuntu@ip-172-31-1-8:~$ jps  
mode : false  
15/04/17 18: 2428 Jps  
15/04/17 18: 2277 MRAppMaster  
ubuntu@ip-172-31-15-16:~$ jps  
1820 NodeManager  
1667 DataNode  
2277 MRAppMaster  
ubuntu@ip-172-31-15-16:~$ jps  
2442 Jps  
1820 NodeManager  
1667 DataNode  
2277 MRAppMaster  
ubuntu@ip-172-31-15-16:~$
```

Don't Forget to
SUSPEND/TERMINATE
Instance !!!!!