

# Hadoop Installation:

Name: อวยศัย ภิรมย์รัตน์

Tel. : 086-813-5354

e-mail : p.Auoychai@gmail.com

Big Data

# Wrap up



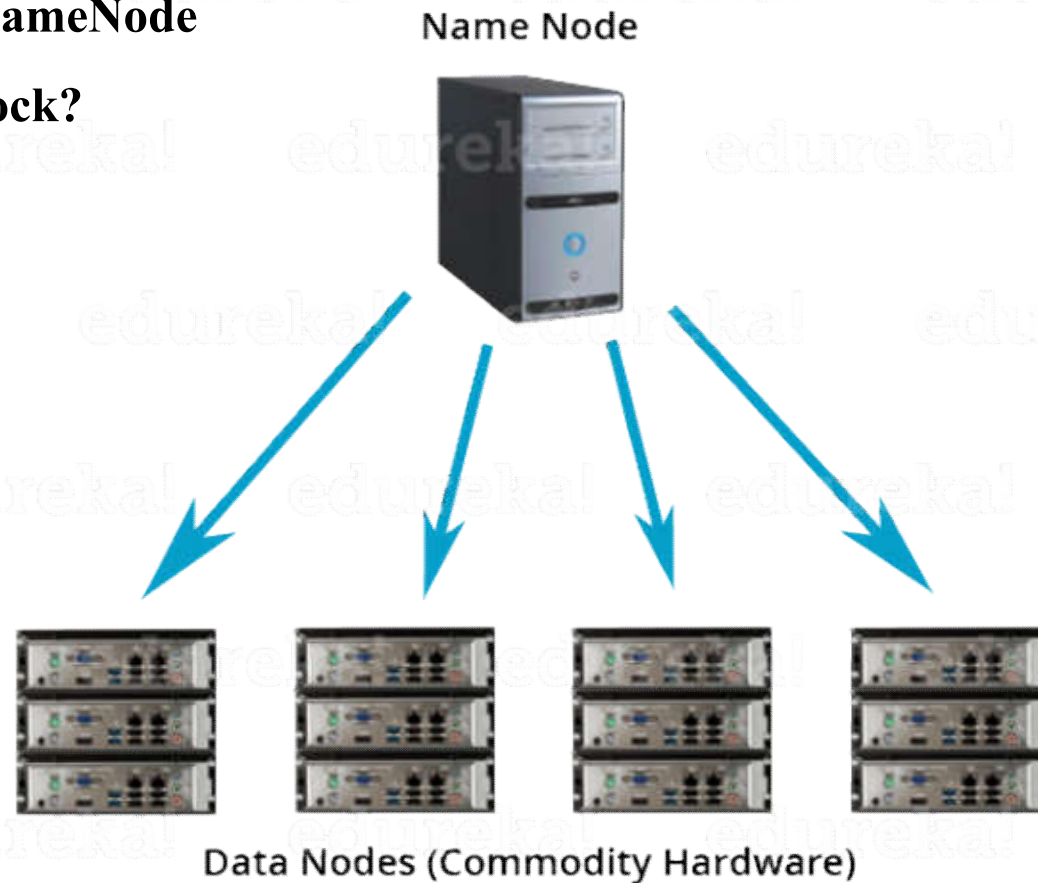
## # System Topology :

---

### HDFS Master/Slave Architecture

**NameNode, DataNode and  
Secondary NameNode**

**What is a block?**



# **Hadoop Installation :**

---

## **# Installation Mode :**

---

Hadoop Installation Mode:

Local standalone mode:

Pseudo-distributed mode:

Fully distributed mode:

# Hadoop Installation :

---

## # Hadoop Installation:

---

# Define Hadoop System Topology

- Master Node: IP=
- Slave Node#1: IP=
- Slave Node#2: IP=
- Slave Node#3: IP=

## # Hadoop Installation:

- Main Package Directory : /usr/local/hadoop
- Data Directory : /var/hadoop\_data/namenode  
: /var/hadoop\_data/datanode
- Log file Directory : /var/log/hadoop

## # Hadoop Installation:

---

### #Hadoop Configuration Review:

Core-site.xml

```
<configuration>
  <property>
    <name>fs.defaultFS</name>
    <value>hdfs://localhost:9000</value>
  </property>
  <property>
    <name>hadoop.proxyuser.root.hosts</name>
    <value>*</value>
  </property>
  <property>
    <name>hadoop.proxyuser.root.groups</name>
    <value>*</value>
  </property>
</configuration>
```

## # Hadoop Installation:

---

### #Hadoop Configuration Review:

hdfs-site.xml

```
<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>file:/var/hadoop_data/namenode</value>
  </property>
  <property>
    <name>dfs.datanode.data.dir</name>
    <value>file:/var/hadoop_data/datanode</value>
  </property>
  <property>
    <name>dfs.namenode.acls.enabled</name>
    <value>true</value>
  </property>
</configuration>
```



# Hadoop Installation :

## # Hadoop Installation:

### #Hadoop Configuration Review:

yarn-site.xml

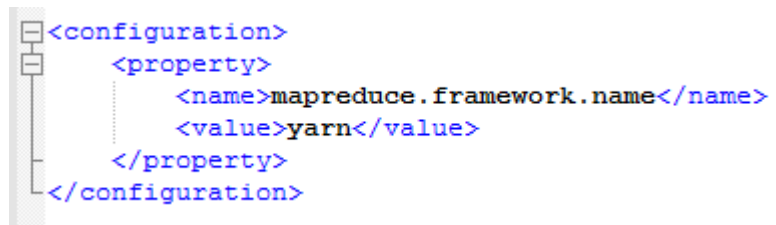
```
<configuration>
  <!-- Site specific YARN configuration properties -->
  <property>
    <name>yarn.resourcemanager.hostname</name>
    <value>localhost</value>
  </property>
  <property>
    <name>yarn.resourcemanager.scheduler.address</name>
    <value>localhost:8030</value>
  </property>
  <property>
    <name>yarn.resourcemanager.resource-tracker.address</name>
    <value>localhost:8031</value>
  </property>
  <property>
    <name>yarn.resourcemanager.address</name>
    <value>localhost:8032</value>
  </property>
  <property>
    <name>yarn.resourcemanager.admin.address</name>
    <value>localhost:8033</value>
  </property>
  <property>
    <name>yarn.resourcemanager.webapp.address</name>
    <value>localhost:8088</value>
  </property>
  <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>
</configuration>
```

### # Hadoop Installation:

---

#### #Hadoop Configuration Review:

mapred-site.xml

The image shows a snippet of XML code from a file named mapred-site.xml. On the left side, there is a vertical toolbar with icons for editing XML, including a tree view icon, a copy icon, a paste icon, and a search icon. The XML code is as follows:

```
<configuration>
  <property>
    <name>mapreduce.framework.name</name>
    <value>yarn</value>
  </property>
</configuration>
```

### # Hadoop Installation:

---

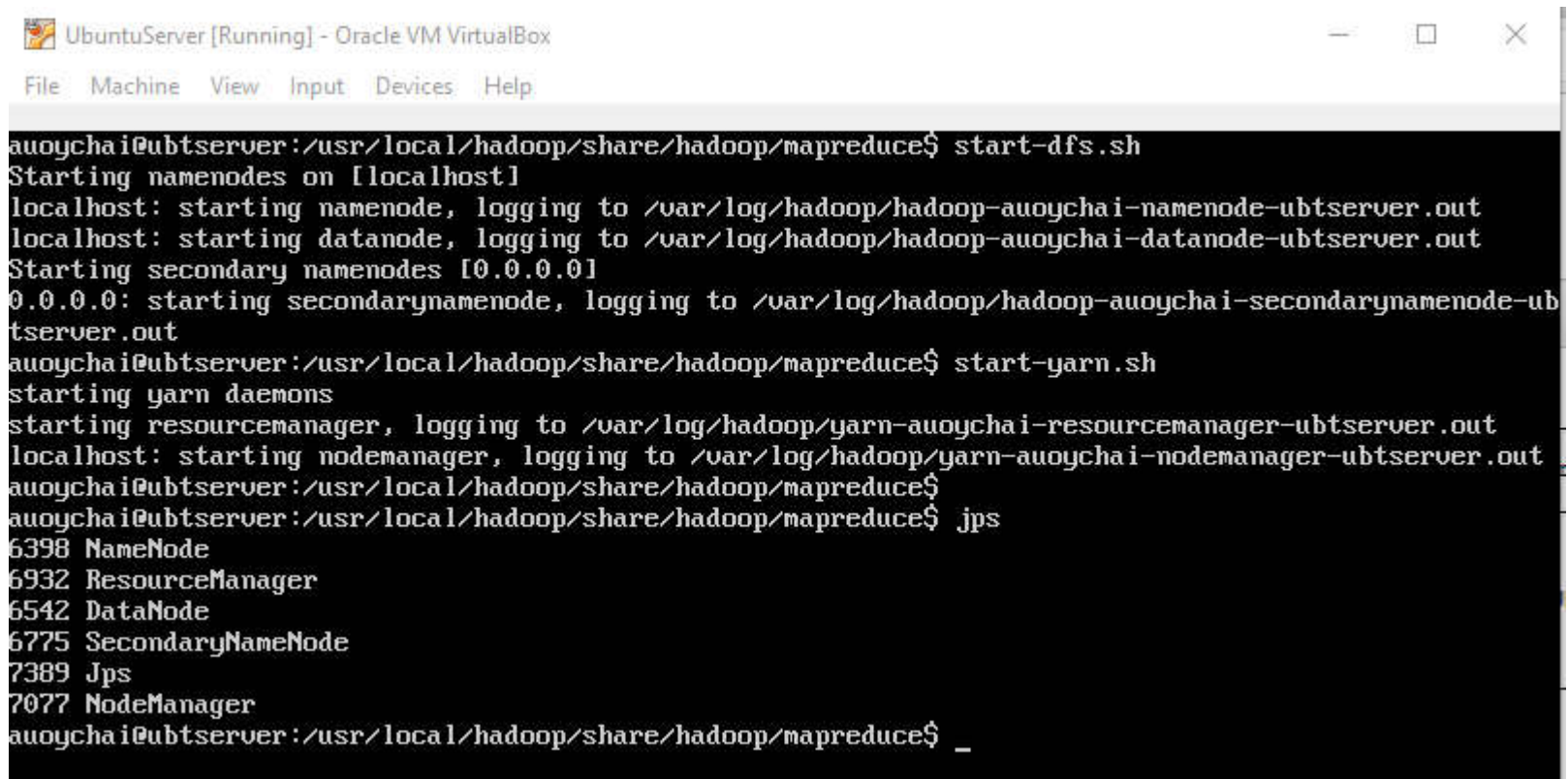
# Initial HDFS : HDFS Formatting

\$hadoop namenode -format

### # Post Test Installation:

How to Start&Stop Hadoop :

- start-dfs.sh , start-yarn.sh
- stop-yarn.sh , stop-dfs.sh

A screenshot of a terminal window titled "UbuntuServer [Running] - Oracle VM VirtualBox". The terminal shows the execution of Hadoop start scripts. The user runs "start-dfs.sh", which starts the NameNode and DataNodes on localhost. Then, the user runs "start-yarn.sh", which starts the Yarn daemons (ResourceManager and NodeManager) on localhost. Finally, the user runs "jps", which lists the running processes: NameNode (6398), ResourceManager (6932), DataNode (6542), SecondaryNameNode (6775), Jps (7389), and NodeManager (7077).

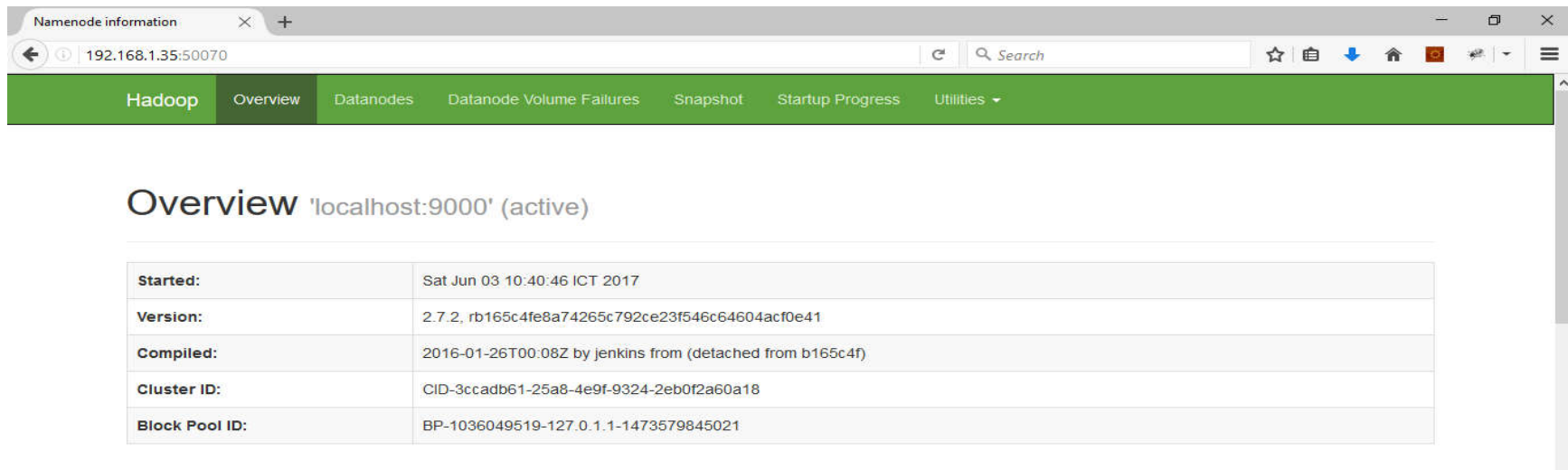
```
auoychai@ubtserver:/usr/local/hadoop/share/hadoop/mapreduce$ start-dfs.sh
Starting namenodes on [localhost]
localhost: starting namenode, logging to /var/log/hadoop/hadoop-auoychai-namenode-ubtserver.out
localhost: starting datanode, logging to /var/log/hadoop/hadoop-auoychai-datanode-ubtserver.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /var/log/hadoop/hadoop-auoychai-secondarynamenode-ubtserver.out
auoychai@ubtserver:/usr/local/hadoop/share/hadoop/mapreduce$ start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /var/log/hadoop/yarn-auoychai-resourcemanager-ubtserver.out
localhost: starting nodemanager, logging to /var/log/hadoop/yarn-auoychai-nodemanager-ubtserver.out
auoychai@ubtserver:/usr/local/hadoop/share/hadoop/mapreduce$
auoychai@ubtserver:/usr/local/hadoop/share/hadoop/mapreduce$ jps
6398 NameNode
6932 ResourceManager
6542 DataNode
6775 SecondaryNameNode
7389 Jps
7077 NodeManager
auoychai@ubtserver:/usr/local/hadoop/share/hadoop/mapreduce$ _
```

Web UI: <IP>:50070

# Hadoop Installation :

## # Post Test Installation:

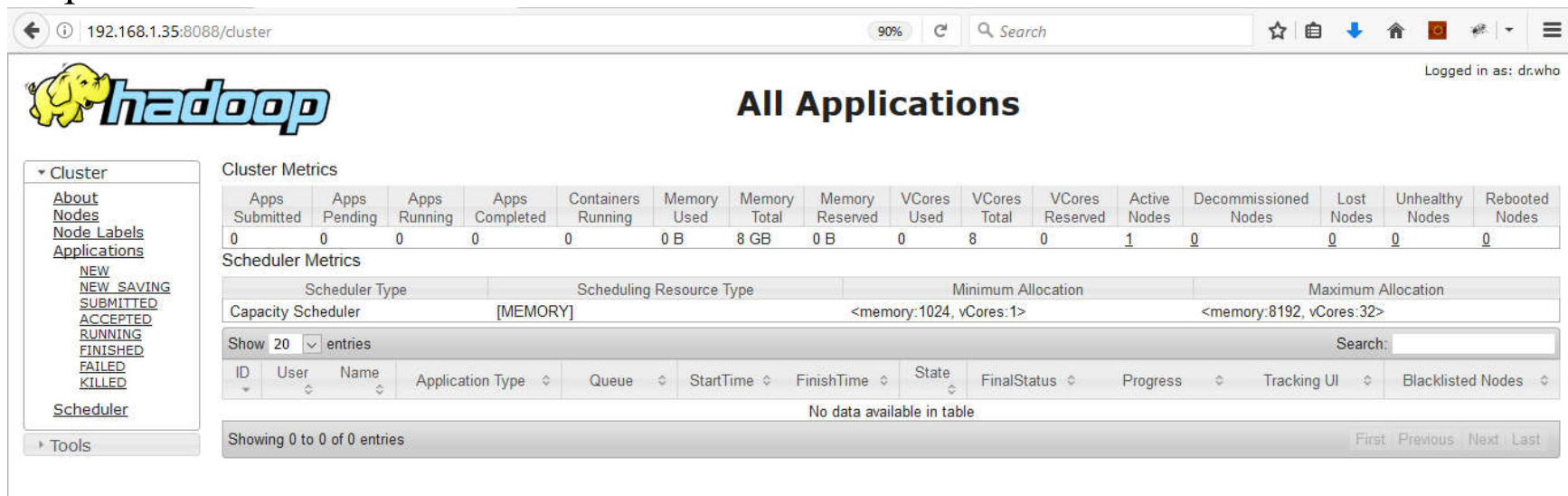
http://< IP >:50070/



The screenshot shows the 'NameNode information' page in a web browser. The address bar displays '192.168.1.35:50070'. The page has a green navigation bar with tabs: 'Hadoop', 'Overview', 'Datanodes', 'Datanode Volume Failures', 'Snapshot', 'Startup Progress', and 'Utilities'. The 'Overview' tab is selected, showing 'localhost:9000' (active). Below the navigation bar, there is a table with the following information:

Started:	Sat Jun 03 10:40:46 ICT 2017
Version:	2.7.2, rb165c4fe8a74265c792ce23f546c64604acf0e41
Compiled:	2016-01-26T00:08Z by jenkins from (detached from b165c4f)
Cluster ID:	CID-3ccadb61-25a8-4e9f-9324-2eb0f2a60a18
Block Pool ID:	BP-1036049519-127.0.1.1-1473579845021

http://< IP >:8088



The screenshot shows the 'All Applications' page in a web browser. The address bar displays '192.168.1.35:8088/cluster'. The page features the Hadoop logo and the text 'All Applications'. Below the logo, there is a sidebar with a 'Cluster' section containing links: 'About', 'Nodes', 'Node Labels', 'Applications', 'NEW', 'NEW SAVING', 'SUBMITTED', 'ACCEPTED', 'RUNNING', 'FINISHED', 'FAILED', 'KILLED', and 'Scheduler'. The main content area displays 'Cluster Metrics' and 'Scheduler Metrics'. The 'Cluster Metrics' table shows various metrics for the cluster, including Apps Submitted, Apps Pending, Apps Running, Apps Completed, Containers Running, Memory Used, Memory Total, Memory Reserved, VCoers Used, VCoers Total, VCoers Reserved, Active Nodes, Decommissioned Nodes, Lost Nodes, Unhealthy Nodes, and Rebooted Nodes. The 'Scheduler Metrics' section shows the Scheduler Type (Capacity Scheduler), Scheduling Resource Type (MEMORY), Minimum Allocation, and Maximum Allocation. Below these metrics, there is a table with columns: ID, User, Name, Application Type, Queue, StartTime, FinishTime, State, FinalStatus, Progress, Tracking UI, and Blacklisted Nodes. The table is currently empty, showing 'No data available in table'.

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCoers Used	VCoers Total	VCoers Reserved	Active Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes
0	0	0	0	0	0 B	8 GB	0 B	0	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:32>

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Progress	Tracking UI	Blacklisted Nodes
No data available in table											

### # Hand-On Plan:

---

A). Pseudo-distributed mode:

B). Fully distributed mode:

## Hadoop Installation :

### # Download & Install Hadoop : Pseudo-distributed mode

---

- Update Package
- Setup SSH
- Setup Java
- Setup Hadoop

### # Download & Install Hadoop :

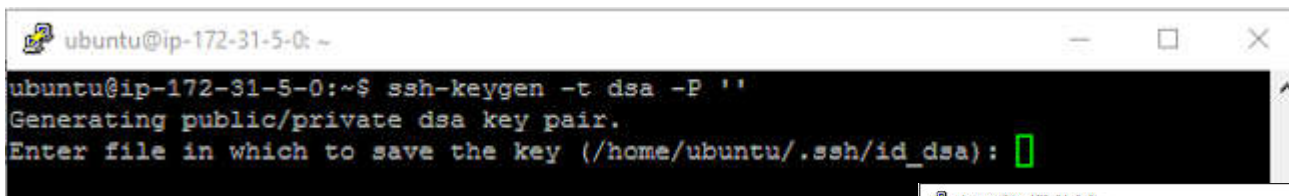
- Update Package

`sudo apt-get update`

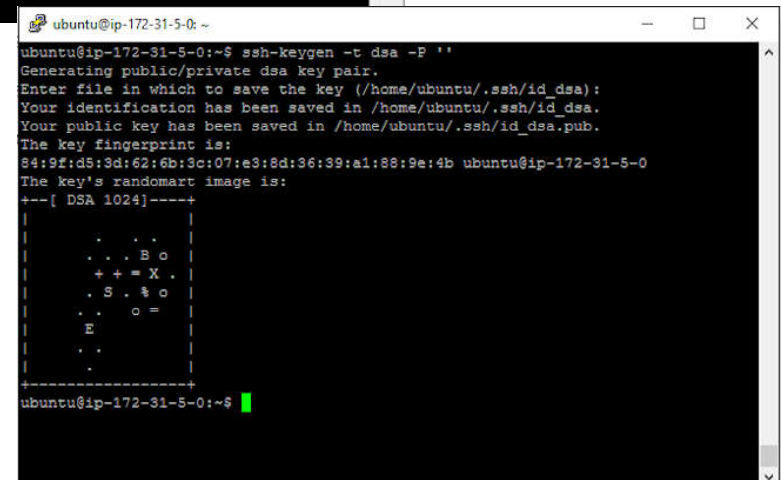
- Setup SSH

`sudo apt-get install -y openssh-server`

`ssh-keygen -t dsa -P ''`



```
ubuntu@ip-172-31-5-0: ~  
ubuntu@ip-172-31-5-0:~$ ssh-keygen -t dsa -P ''  
Generating public/private dsa key pair.  
Enter file in which to save the key (/home/ubuntu/.ssh/id_dsa):
```



```
ubuntu@ip-172-31-5-0: ~  
ubuntu@ip-172-31-5-0:~$ ssh-keygen -t dsa -P ''  
Generating public/private dsa key pair.  
Enter file in which to save the key (/home/ubuntu/.ssh/id_dsa):  
Your identification has been saved in /home/ubuntu/.ssh/id_dsa.  
Your public key has been saved in /home/ubuntu/.ssh/id_dsa.pub.  
The key fingerprint is:  
84:9f:d5:3d:62:6b:3c:07:e3:8d:36:39:a1:88:9e:4b ubuntu@ip-172-31-5-0  
The key's randomart image is:  
+--[ DSA 1024]-----+  
| . . . |  
| . . B o |  
| + + = X . |  
| . S . % o |  
| . . o = |  
| E |  
| . . |  
| . |  
+-----+  
ubuntu@ip-172-31-5-0:~$
```



## # Download & Install Hadoop :

### ■ Setup SSH

-- โหลด Public Key เข้า Key-Store

```
$cat .ssh/id_data.pub >> .ssh/authorized_key
```

```
$ssh localhost
```

```
ubuntu@ip-172-31-4-254: ~
* Documentation: https://help.ubuntu.com/

System information as of Tue Aug 16 03:20:56 UTC 2016

System load:  0.0           Processes:           97
Usage of /:   2.7% of 29.39GB Users logged in:      0
Memory usage: 5%           IP address for eth0: 172.31.4.254
Swap usage:   0%

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

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

0 packages can be updated.
0 updates are security updates.

New release '16.04.1 LTS' available.
Run 'do-release-upgrade' to upgrade to it.

Last login: Tue Aug 16 03:20:57 2016 from 49.229.230.93
ubuntu@ip-172-31-4-254:~$
```

```
ubuntu@ip-172-31-26-239: ~
Usage of /:   6.1% of 29.39GB  Users logged in:      0
Memory usage: 5%             IP address for eth0: 172.31.26.239
Swap usage:   0%

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

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

25 packages can be updated.
14 updates are security updates.

New release '16.04.1 LTS' available.
Run 'do-release-upgrade' to upgrade to it.

Last login: Sun Aug 21 10:39:13 2016 from node-d7x.pool-125-24.dynamic.totbb.net
ubuntu@ip-172-31-26-239:~$ eixt
eixt: command not found
ubuntu@ip-172-31-26-239:~$ exit
logout
Connection to localhost closed.
ubuntu@ip-172-31-26-239:~$
```

### # Download & Install Hadoop :

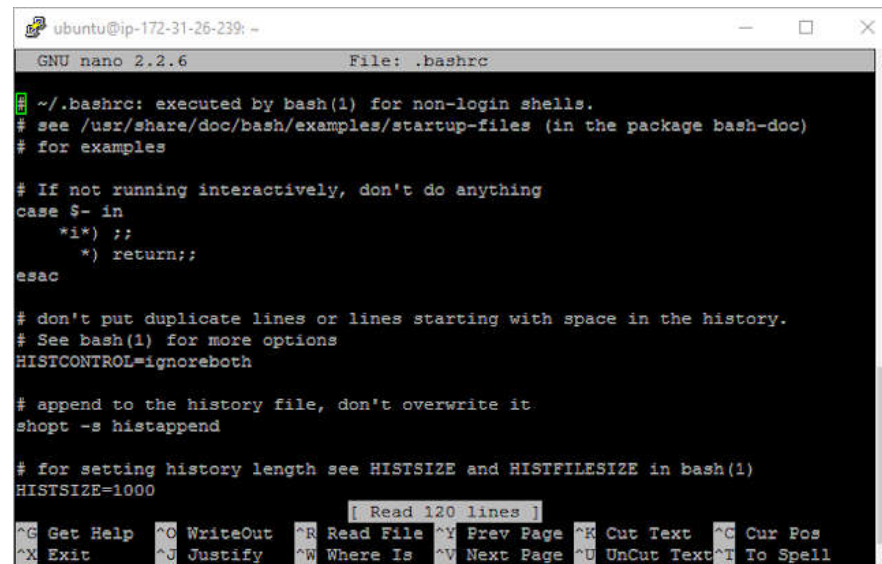
#### ■ Setup Java

`sudo apt-get install -y openjdk-7-jdk`

-- Add environment variable ที่ไฟล์ .bashrc

`export JAVA_HOME = usr/lib/jvm/java-7-openjdk-amd64`

`export PATH=$PATH:$JAVA_HOME/bin`



```
ubuntu@ip-172-31-26-239: ~
GNU nano 2.2.6 File: .bashrc

# ~/.bashrc: executed by bash(1) for non-login shells.
# see /usr/share/doc/bash/examples/startup-files (in the package bash-doc)
# for examples

# If not running interactively, don't do anything
case $- in
  *) ;;
  *) return;;
esac

# don't put duplicate lines or lines starting with space in the history.
# See bash(1) for more options
HISTCONTROL=ignoreboth

# append to the history file, don't overwrite it
shopt -s histappend

# for setting history length see HISTSIZE and HISTFILESIZE in bash(1)
HISTSIZE=1000

[ Read 120 lines ]
^G Get Help  ^O WriteOut  ^R Read File ^Y Prev Page ^K Cut Text   ^C Cur Pos
^X Exit      ^J Justify   ^W Where Is ^V Next Page ^U UnCut Text^T To Spell
```

### # Download & Install Hadoop :

---

#### ■ Setup Hadoop

##### ii-1). Download Hadoop

wget

<http://mirror.cc.columbia.edu/pub/software/apache/hadoop/common/hadoop-2.7.2/hadoop-2.7.2.tar.gz>

##### ii-2). สร้าง Folder สำหรับให้ Hadoop พิมพ์คำสั่งบรรทัดต่อไปนี้ที่หน้าจอ Console

```
sudo mkdir -p /usr/local/Hadoop // โฟลเดอร์สำหรับ Hadoop
```

```
sudo chown ubuntu:ubuntu -R /usr/local/hadoop
```

```
sudo mkdir /var/log/Hadoop // โฟลเดอร์สำหรับ Log file ของ Hadoop
```

```
sudo chown -R ubuntu:ubuntu /var/log/hadoop
```

```
sudo mkdir -p /var/hadoop_data // โฟลเดอร์สำหรับ Data file ที่ HDFS จัดการ
```

```
sudo mkdir -p /var/hadoop_data/namenode
```

```
sudo mkdir -p /var/hadoop_data/datanode
```

```
sudo chown ubuntu:ubuntu -R /var/hadoop_data
```

### # Download & Install Hadoop :

---

#### ■ Setup Hadoop

##### ii-3). Install Hadoop

```
tar -xvf hadoop-2.7.2.tar.gz
```

```
sudo mv ./hadoop-2.7.2/* /usr/local/hadoop
```

```
sudo chown ubuntu:ubuntu -R /usr/local/hadoop
```

##### ii-4). Setup Hadoop Variable Environment

เพิ่ม Environment Variable ให้กับ Hadoop ลงใน .bashrc file ตามนี้ ด้วยการพิมพ์คำสั่ง nano .bashrc เพิ่ม 2 บรรทัดด้านล่างลงไปตามนี้

```
export HADOOP_HOME=/usr/local/hadoop
```

```
export PATH=$PATH:$HADOOP_HOME/bin
```

```
export PATH=$PATH:$HADOOP_HOME/sbin
```

ii-5). แก้ไข script สำหรับกำหนด Environment Variable ให้กับ Hadoop ที่ไฟล์ hadoop-env.sh ที่อยู่ใน /usr/local/hadoop/etc/hadoop โดยพิมพ์ nano hadoop-env.sh และ เพิ่มคำสั่ง 2 บรรทัดด้านล่างนี้ แสดงตามรูปที่ ii5-1

```
export JAVA_HOME=/usr/lib/jvm/java-7-openjdk-amd64
```

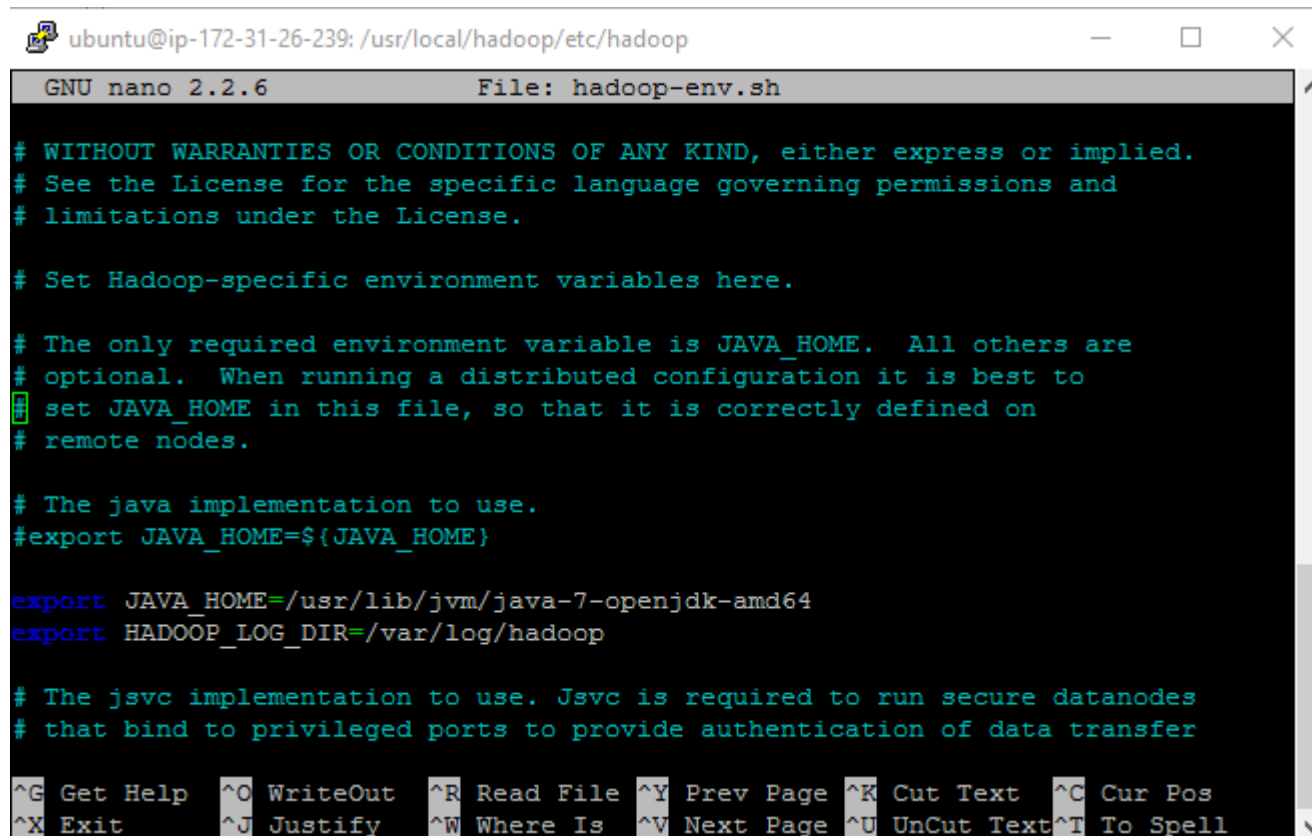
```
export HADOOP_LOG_DIR=/var/log/hadoop
```

### # Download & Install Hadoop :

ii-5). แก้ไข script สำหรับกำหนด Environment Variable ให้กับ Hadoop ที่ไฟล์ `hadoop-env.sh` ที่อยู่ใน `/usr/local/hadoop/etc/hadoop` โดยพิมพ์ `nano hadoop-env.sh`

```
export JAVA_HOME=/usr/lib/jvm/java-7-openjdk-amd64
```

```
export HADOOP_LOG_DIR=/var/log/hadoop
```



```
ubuntu@ip-172-31-26-239: /usr/local/hadoop/etc/hadoop
GNU nano 2.2.6      File: hadoop-env.sh

# 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
export HADOOP_LOG_DIR=/var/log/hadoop

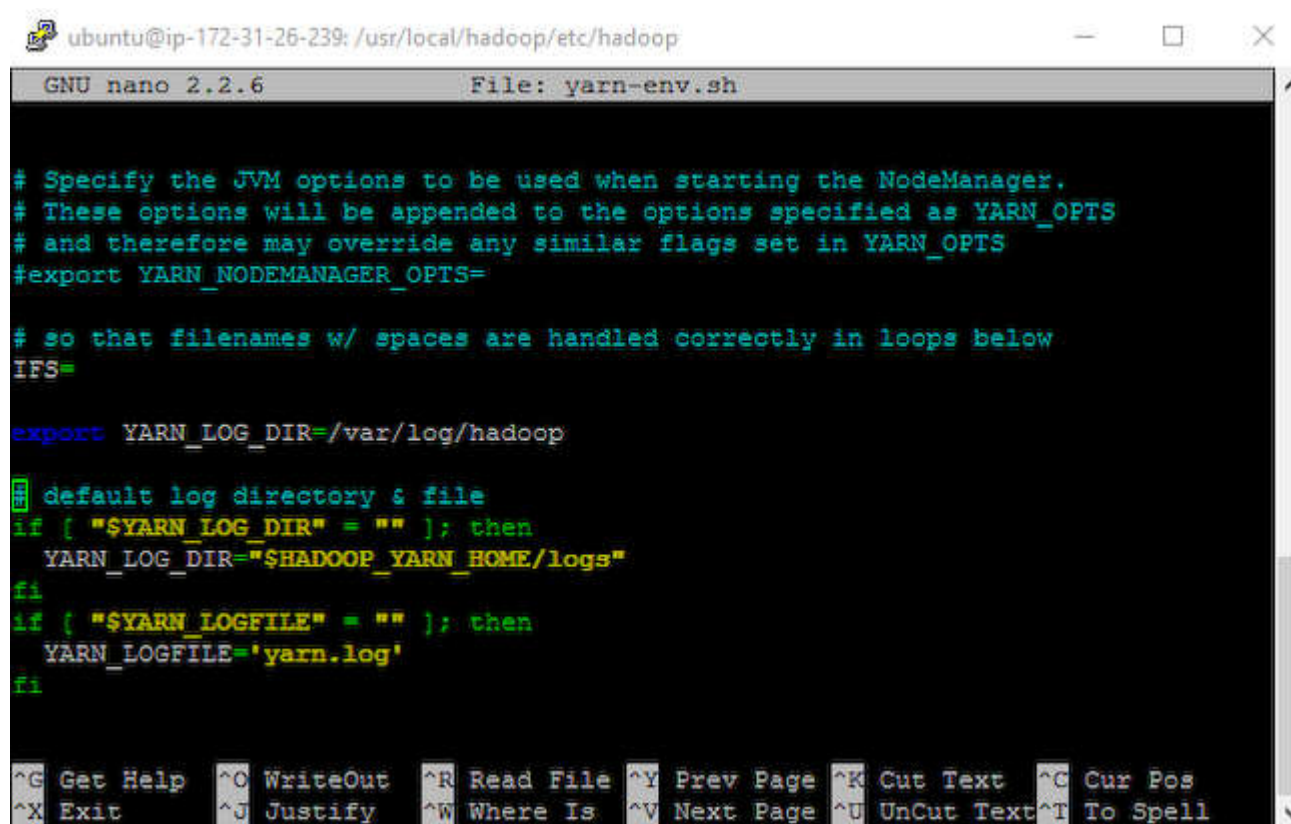
# The jsvc implementation to use. Jsvc is required to run secure datanodes
# that bind to privileged ports to provide authentication of data transfer

^G Get Help  ^O WriteOut  ^R Read File  ^Y Prev Page  ^K Cut Text   ^C Cur Pos
^X Exit      ^J Justify   ^W Where Is   ^V Next Page  ^U UnCut Text ^T To Spell
```

### # Download & Install Hadoop :

ii-6). แก้ไข shell script สำหรับกำหนด Environment Variable ให้กับ Yarn ที่ไฟล์ yarn-env.sh ที่อยู่ใน /usr/local/hadoop/etc/hadoop โดย nano yarn-env.sh

```
export YARN_LOG_DIR = /var/log/hadoop
```



```
ubuntu@ip-172-31-26-239: /usr/local/hadoop/etc/hadoop
GNU nano 2.2.6 File: yarn-env.sh

# Specify the JVM options to be used when starting the NodeManager.
# These options will be appended to the options specified as YARN_OPTS
# and therefore may override any similar flags set in YARN_OPTS
#export YARN_NODEMANAGER_OPTS=

# so that filenames w/ spaces are handled correctly in loops below
IFS=

export YARN_LOG_DIR=/var/log/hadoop

# default log directory & file
if [ "$YARN_LOG_DIR" = "" ]; then
    YARN_LOG_DIR="$HADOOP_YARN_HOME/logs"
fi
if [ "$YARN_LOGFILE" = "" ]; then
    YARN_LOGFILE='yarn.log'
fi

^G Get Help ^O WriteOut ^R Read File ^Y Prev Page ^K Cut Text ^C Cur Pos
^X Exit ^J Justify ^W Where Is ^V Next Page ^U UnCut Text ^T To Spell
```

### # Initial HFDS :

---

ii-8). Format ระบบไฟล์ ( ภาพเหมือนกับเรา Format Harddisk )

```
Last login: Sun Aug 21 11:52:06 2016 from node-d7x.pool-125-24.dynamic.totbb.net
ubuntu@ip-172-31-26-239:~$ cd /usr/local/hadoop/etc/hadoop/
ubuntu@ip-172-31-26-239:/usr/local/hadoop/etc/hadoop$ hdfs namenode -format
```

```
16/08/27 02:53:36 INFO util.GSet: capacity      = 2^15 = 32768 entries
Re-format filesystem in Storage Directory /var/hadoop_data/namenode ? (Y or N) Y
16/08/27 02:53:41 INFO namenode.FSImage: Allocated new BlockPoolId: BP-903428081-172.31.26.239-1472266421834
16/08/27 02:53:41 INFO common.Storage: Storage directory /var/hadoop_data/namenode has been successfully formatted.
16/08/27 02:53:42 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
16/08/27 02:53:42 INFO util.ExitUtil: Exiting with status 0
16/08/27 02:53:42 INFO namenode.NameNode: SHUTDOWN_MSG:
/*****
SHUTDOWN_MSG: Shutting down NameNode at ip-172-31-26-239.us-west-2.compute.internal/172.31.26.239
*****/
ubuntu@ip-172-31-26-239:/usr/local/hadoop/etc/hadoop$
```

### # Start&Stop Hadoop :

---

ii-9). Start Hadoop ( ที่ /usr/local/hadoop/etc/hadoop )

ii9-1. ทำการ Start HDFS ด้วยการ Run Script นี้ start-dfs.sh และ สังเกตดูว่า Process NameNode และ DataNode ทำงานอยู่ด้วย คำสั่ง jps

ii9-2. ทำการ Start YARN ด้วยการ Run Script นี้ start-yarn.sh และ สังเกตดูว่า Process ResourceManager และ NodeManager ทำงานอยู่ด้วยคำสั่ง jps

ii-11). Stop Hadoop

1). stop-yarn.sh และ 2). stop-dfs.sh ที่ Command Console และ เมื่อขอดู Hadoop Process ด้วยคำสั่ง jps



### # Hello MapReduce :

---

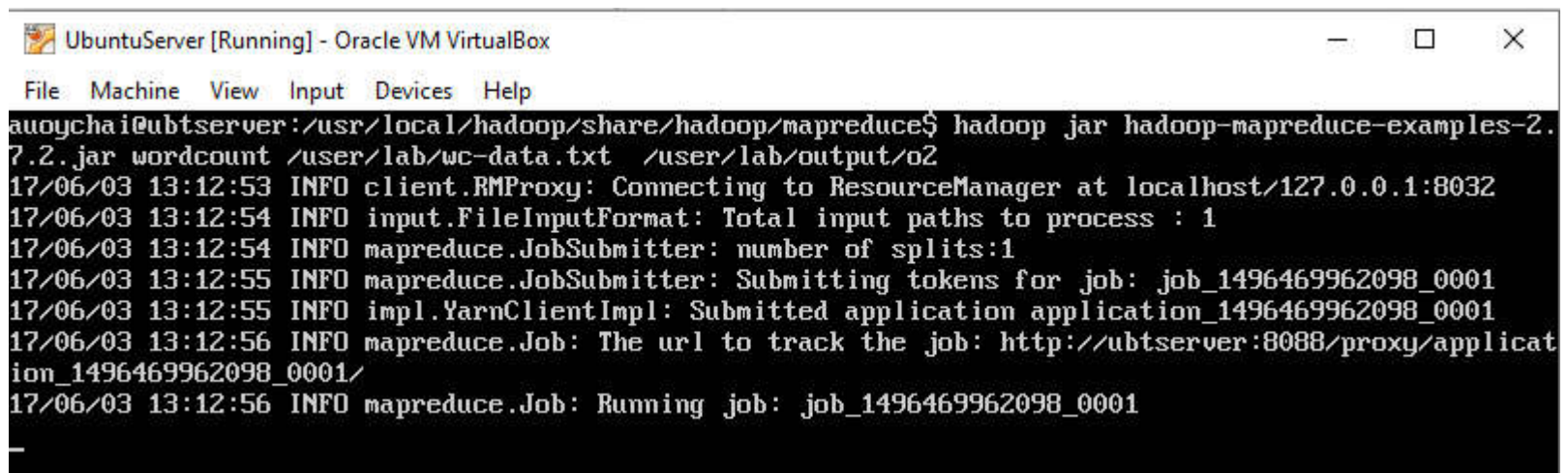
*Package Directory:* \$HADOOP\_HOME/share/hadoop/mapreduce

*Execute Command :*

```
$HADOOP_HOME/share/hadoop/mapreduce$ hadoop jar hadoop-mapreduce-examples-2.7.2.jar  
pi 10 1000
```

\*\*\* Observe terminal screen \*\*\*

```
$HADOOP_HOME/share/hadoop/mapreduce$ hadoop jar hadoop-mapreduce-examples-2.7.2.jar  
wordcount /user/lab/wc-data.txt /user/lab/output/o2
```



The screenshot shows a terminal window titled "UbuntuServer [Running] - Oracle VM VirtualBox". The terminal output displays the execution of the Hadoop MapReduce job "wordcount". The command executed is `hadoop jar hadoop-mapreduce-examples-2.7.2.jar wordcount /user/lab/wc-data.txt /user/lab/output/o2`. The output shows the job's progress, including connecting to the Resource Manager, submitting the application, and running the job. The job ID is `job_1496469962098_0001`.

```
auoychai@ubtserver: /usr/local/hadoop/share/hadoop/mapreduce$ hadoop jar hadoop-mapreduce-examples-2.7.2.jar wordcount /user/lab/wc-data.txt /user/lab/output/o2
17/06/03 13:12:53 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
17/06/03 13:12:54 INFO input.FileInputFormat: Total input paths to process : 1
17/06/03 13:12:54 INFO mapreduce.JobSubmitter: number of splits:1
17/06/03 13:12:55 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1496469962098_0001
17/06/03 13:12:55 INFO impl.YarnClientImpl: Submitted application application_1496469962098_0001
17/06/03 13:12:56 INFO mapreduce.Job: The url to track the job: http://ubtserver:8088/proxy/application_1496469962098_0001/
17/06/03 13:12:56 INFO mapreduce.Job: Running job: job_1496469962098_0001
```

## Hadoop Installation :

# Download & Install Hadoop : Fully distributed mode

---

- X

# The End

Big  
data

Shift