

Practical on MapReduce

Name: Atharva Paliwal

Roll No: B40

Problem statement:

1. Installation of Apache Hadoop.
2. Creation of two virtual machines and connection using NAT.
3. Implementation of mapreduce method by performing word count program on the client machine using files stored on a server machine.

Abstract:

Hadoop is an Apache open-source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models. The Hadoop framework application works in an environment that provides distributed *storage* and *computation* across clusters of computers. Hadoop is designed to scale up from a single server to thousands of machines, each offering local computation and storage. In this system, Hadoop is used to demonstrate the word count example using two virtual machines which are connected using Network Address Translator (NAT) and SSH server.

What is mapreduce in Hadoop?

Mapreduce is a software framework and programming model used for processing huge amounts of data. Mapreduce program work in two phases, namely, Map and Reduce. Map tasks deal with splitting and mapping of data while Reduce tasks shuffle and reduce the data.

Hadoop is capable of running mapreduce programs written in various languages: Java, Ruby, Python, and C++. The programs of Map Reduce in cloud computing are parallel in nature, thus are very useful for performing large-scale data analysis using multiple machines in the cluster.

The input to each phase is key-value pairs. In addition, every programmer needs to specify two functions: map function and reduce function.

Output:

```
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following additional packages will be installed:
  ca-certificates-java java-common openjdk-11-jre-headless
Suggested packages:
  default-jre openjdk-11-demo openjdk-11-source fonts-dejavu-extra
  fonts-ipafont-gothic fonts-ipafont-mincho fonts-wqy-microhei
  | fonts-wqy-zenhei
The following NEW packages will be installed:
  ca-certificates-java java-common openjdk-11-jdk-headless
  openjdk-11-jre-headless
0 upgraded, 4 newly installed, 0 to remove and 501 not upgraded.
Need to get 258 MB of archives.
After this operation, 400 MB of additional disk space will be used.
Do you want to continue? [Y/n] y
Get:1 http://in.archive.ubuntu.com/ubuntu bionic-updates/main amd64 java-common
  all 0.68ubuntu1~18.04.1 [14.5 kB]
Get:2 http://in.archive.ubuntu.com/ubuntu bionic-updates/main amd64 openjdk-11-
  jre-headless amd64 11.0.11+9-0ubuntu2~18.04 [37.2 MB]
6% [2 openjdk-11-jre-headless 4,446 kB/37.2 MB 12%] 97.6 kB/s 43min 15s
```

```
Adding user `hadoopuser' ...
Adding new user `hadoopuser' (1010) with group `hadoop' ...
Creating home directory `/home/hadoopuser' ...
Copying files from `/etc/skel' ...
Enter new UNIX password:
Retype new UNIX password:
passwd: password updated successfully
Changing the user information for hadoopuser
Enter the new value, or press ENTER for the default
  Full Name []: Ruqaiya
  Room Number []:
  Work Phone []:
  Home Phone []:
  Other []:
Is the information correct? [Y/n] y
```

```
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following additional packages will be installed:
  ncurses-term openssh-sftp-server ssh-import-id
Suggested packages:
  molly-guard monkeysphere rssh ssh-askpass
The following NEW packages will be installed:
  ncurses-term openssh-server openssh-sftp-server ssh-import-id
0 upgraded, 4 newly installed, 0 to remove and 501 not upgraded.
Need to get 637 kB of archives.
After this operation, 5,316 kB of additional disk space will be used.
Do you want to continue? [Y/n] y
Get:1 http://in.archive.ubuntu.com/ubuntu bionic-updates/main amd64 ncurses-ter
  m all 6.1-1ubuntu1.18.04 [248 kB]
Ign:2 http://in.archive.ubuntu.com/ubuntu bionic-updates/main amd64 openssh-sft
  p-server amd64 1:7.6p1-4ubuntu0.3
Ign:3 http://in.archive.ubuntu.com/ubuntu bionic-updates/main amd64 openssh-ser
  ver amd64 1:7.6p1-4ubuntu0.3
Get:4 http://in.archive.ubuntu.com/ubuntu bionic-updates/main amd64 ssh-import-
  id all 5.7-0ubuntu1.1 [10.9 kB]
Err:2 http://security.ubuntu.com/ubuntu bionic-updates/main amd64 openssh-sftp-
```

```
Generating public/private rsa key pair.  
Enter file in which to save the key (/home/hadoopuser/.ssh/id_rsa):  
Created directory '/home/hadoopuser/.ssh'.  
Your identification has been saved in /home/hadoopuser/.ssh/id_rsa.  
Your public key has been saved in /home/hadoopuser/.ssh/id_rsa.pub.  
The key fingerprint is:  
SHA256:dn2kbohhuEX6zMfU5wYFjcgM5Pe/iYXn6ey0YUFd3y4 hadoopuser@ruqaiya-VirtualBo
```

```
x  
The key's randomart image is:
```

```
+---[RSA 2048]---+  
|             .o. .o+o|  
|            =.++=..*|  
|           + BoB +.=|  
|          . X * ..= |  
|         S = o E.+|  
+---+-----+---
```

```
The authenticity of host 'localhost (127.0.0.1)' can't be established.  
ECDSA key fingerprint is SHA256:KKNRT9jk/SRRd01zC8NX26gBYjJkqcdVbjGC4CSlaDc.  
Are you sure you want to continue connecting (yes/no)? y  
Please type 'yes' or 'no': yes  
Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.  
hadoopuser@localhost's password:  
Welcome to Ubuntu 18.04.3 LTS (GNU/Linux 5.0.0-23-generic x86_64)
```

```
* Documentation:  https://help.ubuntu.com  
* Management:    https://landscape.canonical.com  
* Support:        https://ubuntu.com/advantage
```

```
* Canonical Livepatch is available for installation.  
- Reduce system reboots and improve kernel security. Activate at:  
  https://ubuntu.com/livepatch
```

```
515 packages can be updated.  
419 updates are security updates.
```

```
Your Hardware Enablement Stack (HWE) is supported until April 2023.
```

```
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.
```



We suggest the following mirror site for your download:

<https://dlcdn.apache.org/hadoop/common/hadoop-2.10.1/hadoop-2.10.1.tar.gz>

Other mirror sites are suggested below.

It is essential that you verify the integrity of the downloaded file using the PGP signature (`.asc` file) or a hash (`.md5` or `.sha*` file).

Please only use the backup mirrors to download KEYS, PGP signatures and hashes (SHA* etc) -- or if no other mirrors are working.

HTTP

File Edit View Search Terminal Help

8.0M).

Deleted archived journal /var/log/journal/ba22f501e3d843e29ee9c0ed17976f6b/user-1000@b1106fc0ec53489a9d15c81cebe84514-0000000000000426-0005acbb26b4a88d.journal (8.0M).

Vacuuming done, freed 392.1M of archived journals from /var/log/journal/ba22f501e3d843e29ee9c0ed17976f6b.


```

Open ▾ .bashrc Save
# ~/.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 -oq 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-11-openjdk-amd64
export HADOOP_HOME=/usr/local/hadoop
export PATH=$PATH:$HADOOP_HOME/bin
export PATH=$PATH:$HADOOP_HOME/sbin
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib";

```

```

Open ▾ hadoop-env.sh Save
/usr/local/hadoop/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.

sh ▾ Tab Width: 8 ▾ Ln 1, Col 1 ▾ INS

```





```
Open ▾  hadoop-env.sh /usr/local/hadoop/etc/hadoop Save   
defined on
# remote nodes.

# The java implementation to use.
export JAVA_HOME=/usr/lib/jvm/java-11-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"}

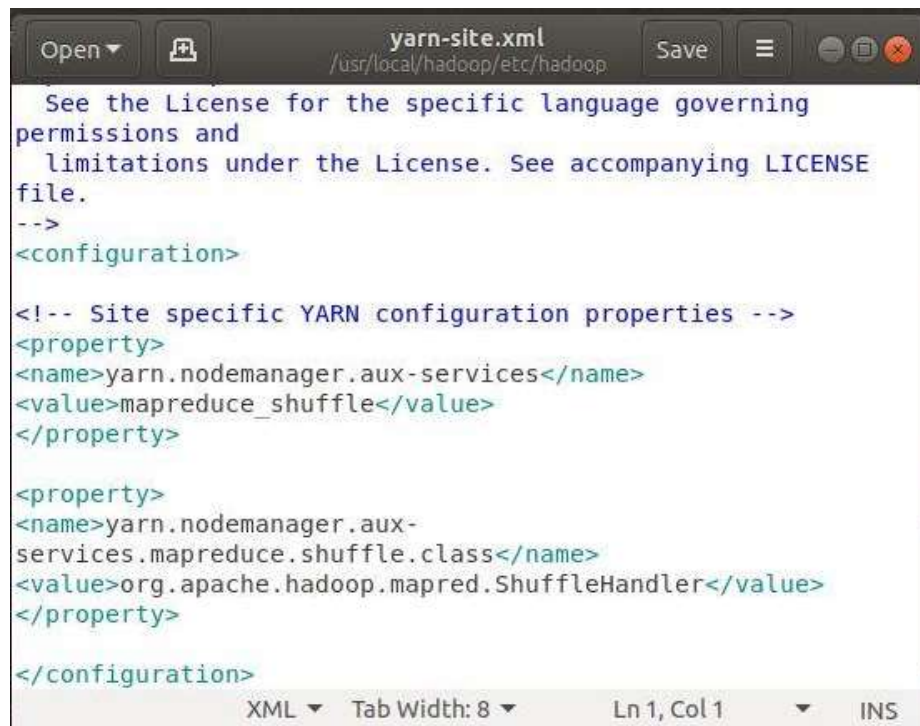
# Extra Java CLASSPATH elements. Automatically insert
capacity-scheduler.
for f in $HADOOP_HOME/contrib/capacity-scheduler/*.jar; do
  if [ "$HADOOP_CLASSPATH" != "" ] then
sh ▾ Tab Width: 8 ▾ Ln 25, Col 52 ▾ INS
```





```
Open ▾  core-site.xml /usr/local/hadoop/etc/hadoop Save   

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.defaultFS</name>
<value>hdfs://localhost:9000</value>
</property>
</configuration>
XML ▾ Tab Width: 8 ▾ Ln 1, Col 1 ▾ INS
```



```
Open ▾  yarn-site.xml /usr/local/hadoop/etc/hadoop Save     
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>  
  
</configuration>  
XML ▾ Tab Width: 8 ▾ Ln 1, Col 1 ▾ INS
```

```
doop/etc/hadoop/mapred-site.xml  
  
** (gedit:4111): WARNING **: 14:54:34.336: Set document metadata failed: Settin  
g attribute metadata::gedit-spell-language not supported  
  
** (gedit:4111): WARNING **: 14:54:34.338: Set document metadata failed: Settin  
g attribute metadata::gedit-encoding not supported  
  
** (gedit:4111): WARNING **: 14:54:38.790: Set document metadata failed: Settin  
g attribute metadata::gedit-position not supported
```



```

at
21/10/17 14:59:27 INFO namenode.NameNode: STARTUP_MSG:
/*****
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = ruqaiya-VirtualBox/127.0.1.1
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 2.10.1
STARTUP_MSG:   classpath = /usr/local/hadoop/etc/hadoop:/usr/local/hadoop/share
/hadoop/common/lib/jsr305-3.0.2.jar:/usr/local/hadoop/share/hadoop/common/lib/j
ackson-mapper-asl-1.9.13.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-
codec-1.4.jar:/usr/local/hadoop/share/hadoop/common/lib/jersey-server-1.9.jar:/
usr/local/hadoop/share/hadoop/common/lib/nimbus-jose-jwt-7.9.jar:/usr/local/had
oop/share/hadoop/common/lib/jaxb-api-2.2.2.jar:/usr/local/hadoop/share/hadoop/c
ommon/lib/jackson-core-asl-1.9.13.jar:/usr/local/hadoop/share/hadoop/common/lib
/jersey-core-1.9.jar:/usr/local/hadoop/share/hadoop/common/lib/woodstox-core-5.
0.3.jar:/usr/local/hadoop/share/hadoop/common/lib/curator-recipes-2.13.0.jar:/u
sr/local/hadoop/share/hadoop/common/lib/spotbugs-annotations-3.1.9.jar:/usr/loc
al/hadoop/share/hadoop/common/lib/xmlenc-0.52.jar:/usr/local/hadoop/share/hadoo
p/common/lib/jsp-api-2.1.jar:/usr/local/hadoop/share/hadoop/common/lib/activati
on-1.1.jar:/usr/local/hadoop/share/hadoop/common/lib/mockito-all-1.8.5.jar:/usr
/local/hadoop/share/hadoop/common/lib/java-xmlbuilder-0.4.jar:/usr/local/hadoop
/share/hadoop/common/lib/api-util-1.0.0-M20.jar:/usr/local/hadoop/share/hadoop/
common/lib/commons-compress-1.19.jar:/usr/local/hadoop/share/hadoop/common/lib/
htrace-core4-4.1.0-incubating.jar:/usr/local/hadoop/share/hadoop/common/lib/jet
s3t-0.9.0.jar:/usr/local/hadoop/share/hadoop/common/lib/jsch-0.1.55.jar:/usr/lo
cal/hadoop/share/hadoop/common/lib/snappy-java-1.0.5.jar:/usr/local/hadoop/shar

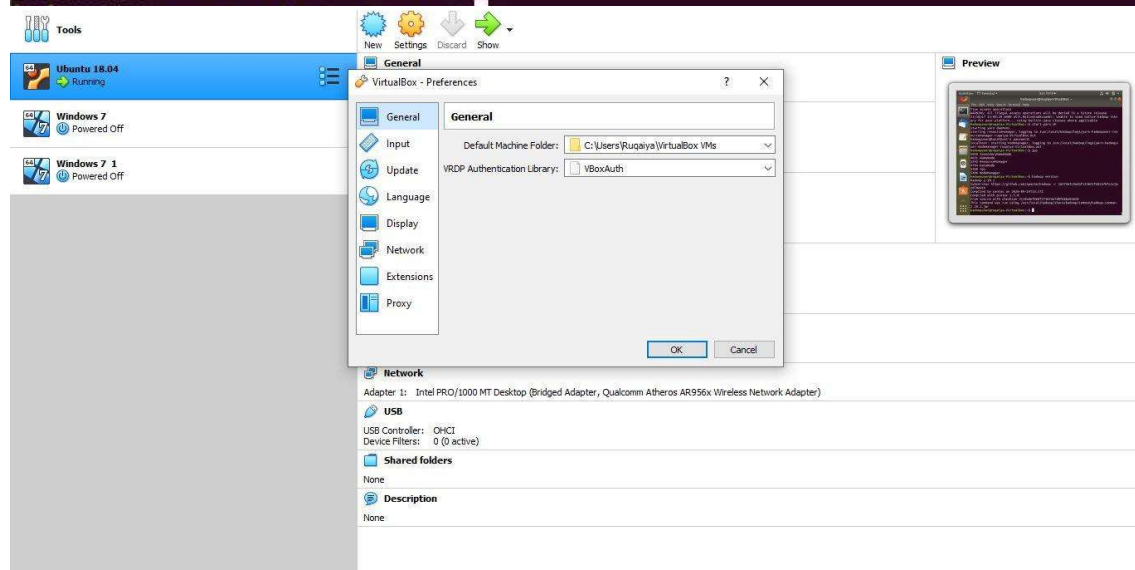
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication
.util.KerberosUtil (file:/usr/local/hadoop/share/hadoop/common/lib/hadoop-auth-
2.10.1.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop
.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflec
tive access operations
WARNING: All illegal access operations will be denied in a future release
21/10/17 15:08:19 WARN util.NativeCodeLoader: Unable to load native-hadoop libr
ary for your platform... using builtin-java classes where applicable
Starting namenodes on [localhost]
hadoopuser@localhost's password:
localhost: starting namenode, logging to /usr/local/hadoop/logs/hadoop-hadoopus
er-namenode-ruqaiya-VirtualBox.out
hadoopuser@localhost's password:
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-hadoopus
er-datanode-ruqaiya-VirtualBox.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 SHA256:KKNRT9jk/SRRd01zC8NX26gBYjJkqcdVbjGC4CSlaDc.
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 host
s.
hadoopuser@0.0.0.0's password:
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-h
adoopuser-secondarynamenode-ruqaiya-VirtualBox.out
WARNING: An illegal reflective access operation has occurred

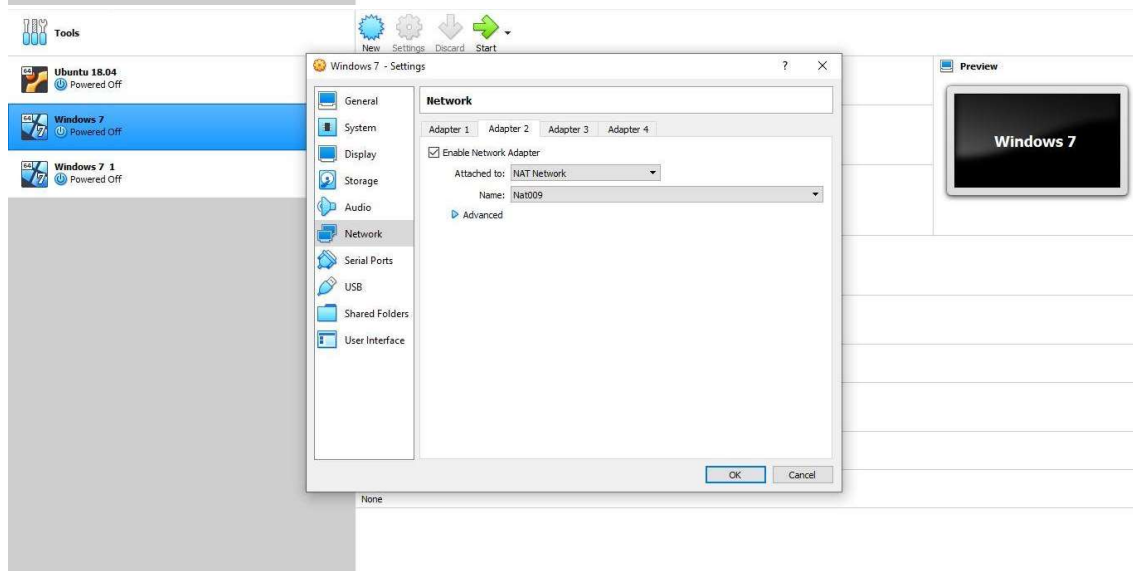
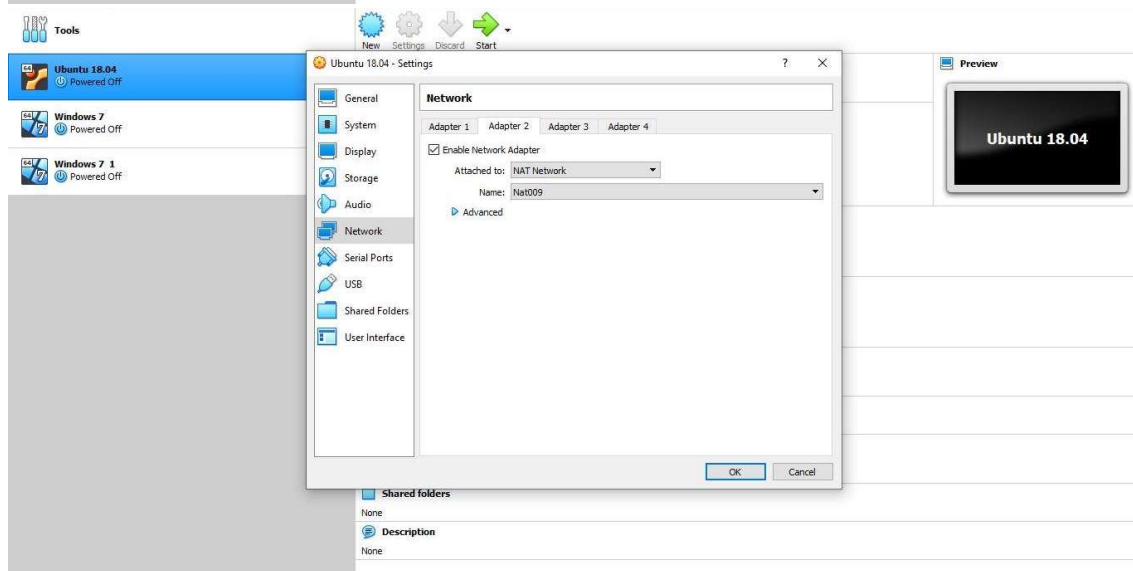
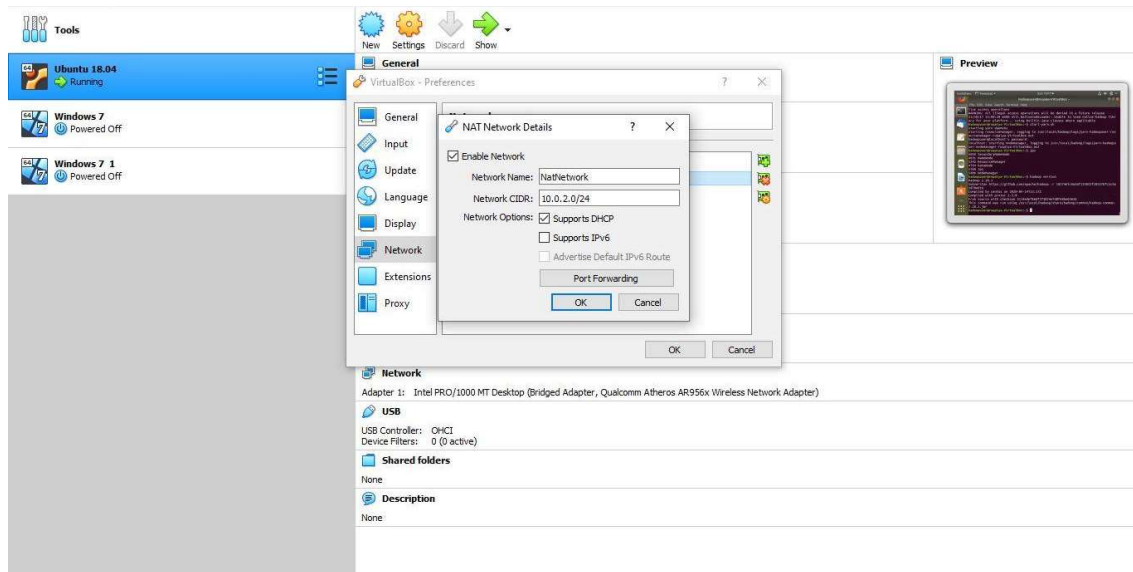
```



```
File Edit View Search Terminal Help
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 SHA256:KKNRT9jk/SRRd01zC8NX26gBYjKqcdVbjGC4CSlaDc.
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 host
s.
hadoopuser@0.0.0.0's password:
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-h
adoopuser-secondarynamenode-ruqaiya-VirtualBox.out
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication
.util.KerberosUtil (file:/usr/local/hadoop/share/hadoop/common/lib/hadoop-auth-
2.10.1.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop
.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflec
tive access operations
WARNING: All illegal access operations will be denied in a future release
21/10/17 15:09:20 WARN util.NativeCodeLoader: Unable to load native-hadoop libr
ary for your platform... using builtin-java classes where applicable

Hadoop 2.10.1
Subversion https://github.com/apache/hadoop -r 1827467c9a56f133025f28557bfc2c56
2d78e816
Compiled by centos on 2020-09-14T13:17Z
Compiled with protoc 2.5.0
From source with checksum 3114edef868f1f3824e7d0f68be03650
This command was run using /usr/local/hadoop/share/hadoop/common/hadoop-common-
2.10.1.jar
```





```
The authenticity of host '10.0.2.6 (10.0.2.6)' can't be est
ablished.
ECDSA key fingerprint is SHA256:KKNRT9jk/SRRd01zC8NX26gBYjJ
kqcdVbjGC4CSlaDc.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added '10.0.2.6' (ECDSA) to the list o
f known hosts.
hadoopuser@10.0.2.6's password:
Welcome to Ubuntu 18.04.3 LTS (GNU/Linux 5.0.0-23-generic x
86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:    https://landscape.canonical.com
 * Support:       https://ubuntu.com/advantage

 * Canonical Livepatch is available for installation.
   - Reduce system reboots and improve kernel security. Act
Ubuntu Dummy [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
64 bytes from 10.0.2.4: icmp_seq=91 ttl=64 time=0.473 ms
64 bytes from 10.0.2.4: icmp_seq=92 ttl=64 time=0.474 ms
64 bytes from 10.0.2.4: icmp_seq=93 ttl=64 time=0.539 ms
64 bytes from 10.0.2.4: icmp_seq=94 ttl=64 time=0.470 ms
64 bytes from 10.0.2.4: icmp_seq=95 ttl=64 time=0.560 ms
64 bytes from 10.0.2.4: icmp_seq=96 ttl=64 time=0.482 ms
64 bytes from 10.0.2.4: icmp_seq=97 ttl=64 time=0.433 ms
64 bytes from 10.0.2.4: icmp_seq=98 ttl=64 time=0.451 ms
64 bytes from 10.0.2.4: icmp_seq=99 ttl=64 time=0.473 ms
64 bytes from 10.0.2.4: icmp_seq=100 ttl=64 time=0.370 ms
64 bytes from 10.0.2.4: icmp_seq=101 ttl=64 time=1.13 ms
64 bytes from 10.0.2.4: icmp_seq=102 ttl=64 time=0.653 ms
64 bytes from 10.0.2.4: icmp_seq=103 ttl=64 time=0.471 ms
64 bytes from 10.0.2.4: icmp_seq=104 ttl=64 time=0.661 ms
64 bytes from 10.0.2.4: icmp_seq=105 ttl=64 time=5.40 ms
64 bytes from 10.0.2.4: icmp_seq=106 ttl=64 time=0.471 ms
^C
--- 10.0.2.4 ping statistics ---
106 packets transmitted, 106 received, 0% packet loss, time 109878ms
rtt min/avg/max/mdev = 0.370/0.597/5.403/0.498 ms
```