Name: Sihang Wang

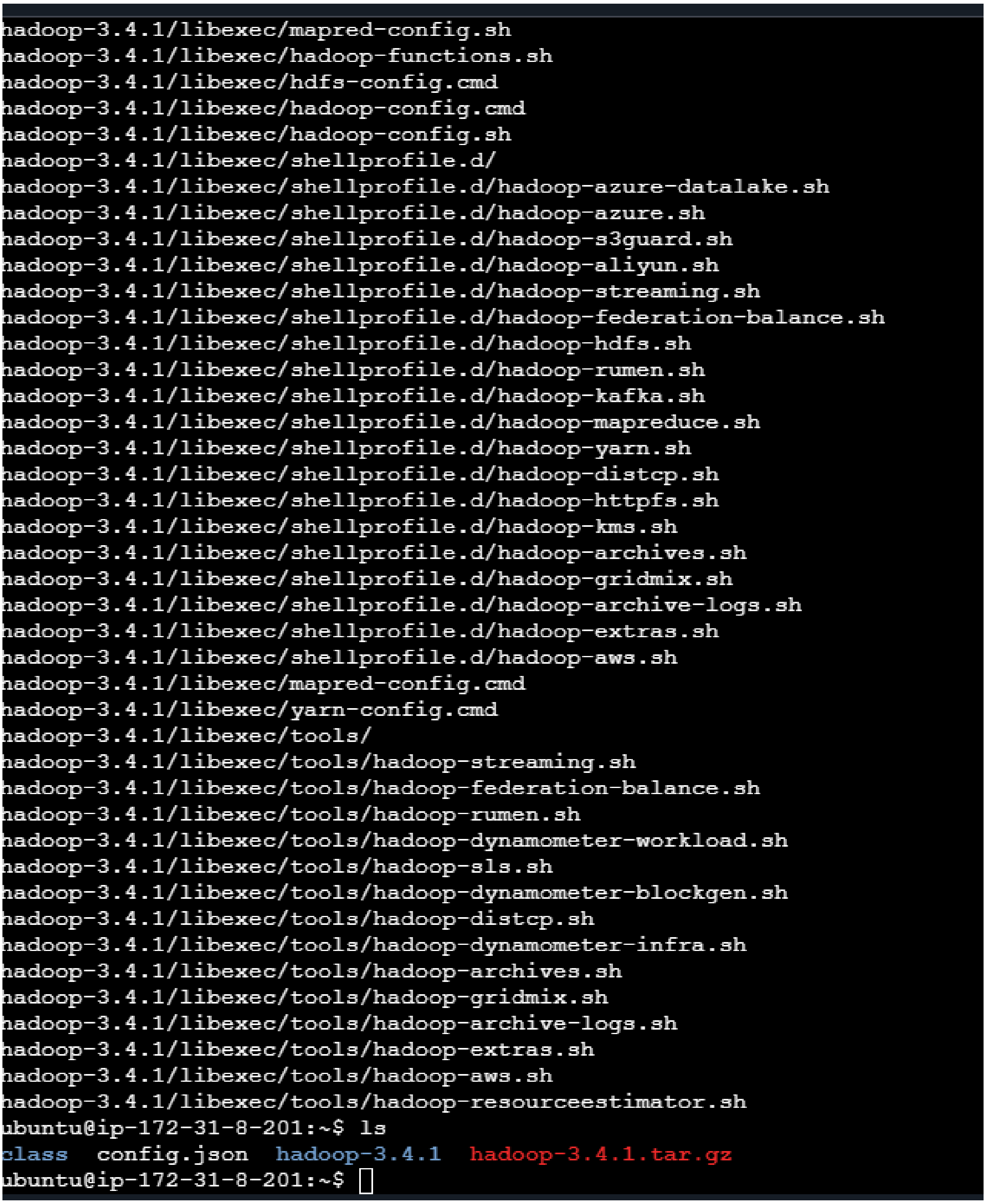
Course: DSCI 551

Assignment: Week 4 Summary

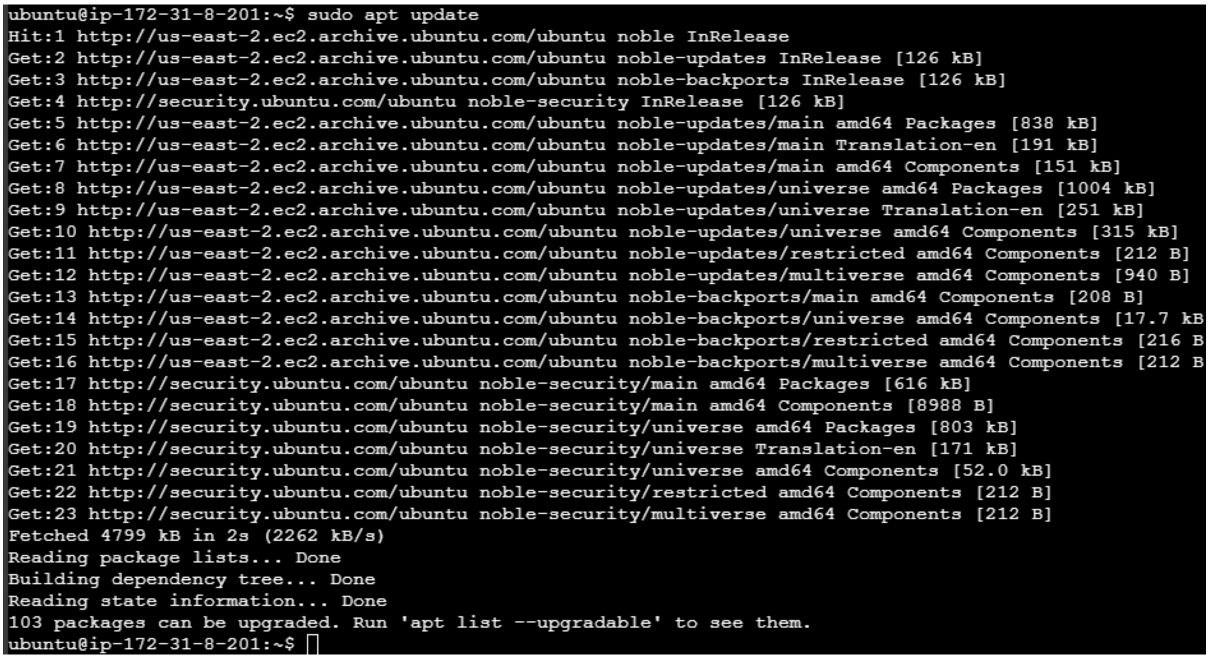
Date: February 9, 2025

Week 4 Summary

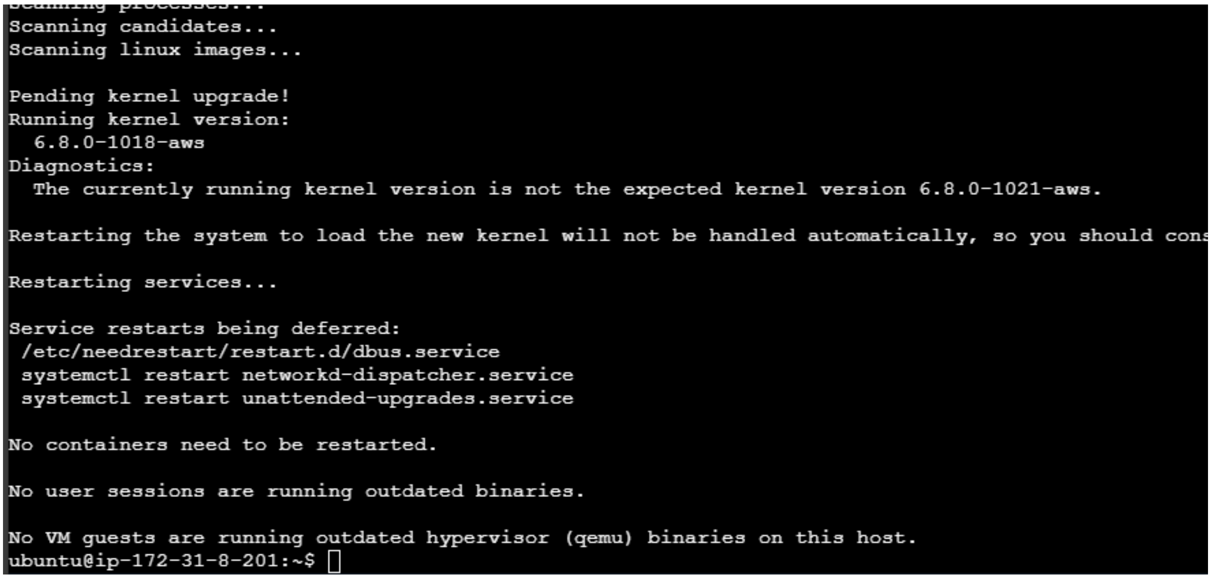
This week's topic is mainly focusing on the Hadoop and HDFS architecture. For the Hadoop part, the lecture mainly explained and operates in the terminal, the EC2 to proceed. The following screen shoots are about technical learning outcomes about Hadoop. HDFS (Hadoop Distributed File System) consists of a single NameNode that stores metadata, including the directory hierarchy, file attributes, and mappings of files to data blocks. It also has multiple DataNodes that store the actual file contents, allowing data to be distributed across nodes for scalability. To ensure fault tolerance, data is replicated with a typical replication factor of 2 or 3, enabling redundancy in case of node failures. Additionally, requests can be directed to any available replica, reducing bottlenecks and improving access efficiency compared to traditional single-server storage systems.



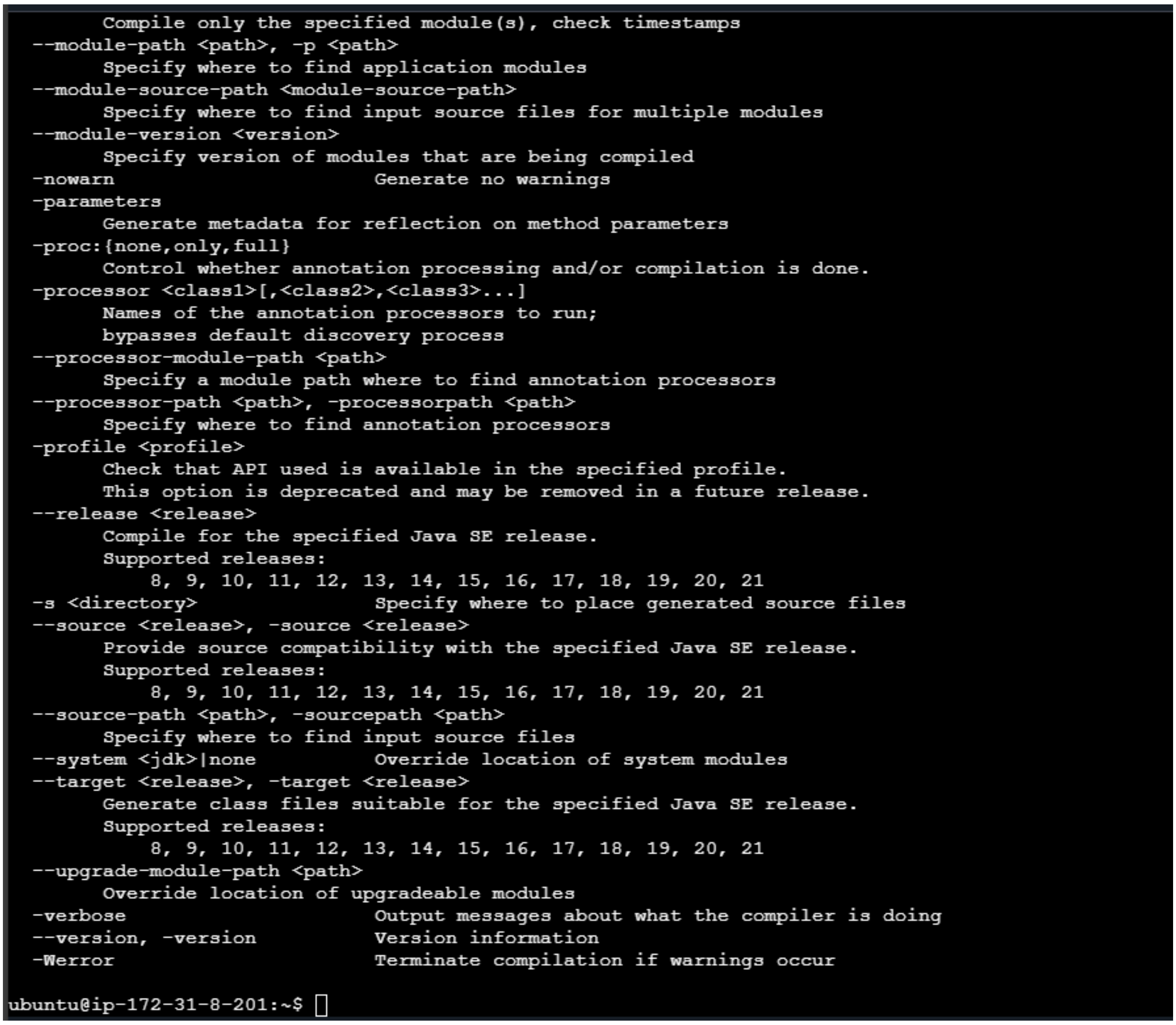
```
hadoop-3.4.1/libexec/mapred-config.sh
hadoop-3.4.1/libexec/hadoop-functions.sh
hadoop-3.4.1/libexec/hdfs-config.cmd
hadoop-3.4.1/libexec/hadoop-config.cmd
hadoop-3.4.1/libexec/hadoop-config.sh
hadoop-3.4.1/libexec/shellprofile.d/
hadoop-3.4.1/libexec/shellprofile.d/hadoop-azure-datalake.sh
hadoop-3.4.1/libexec/shellprofile.d/hadoop-azure.sh
hadoop-3.4.1/libexec/shellprofile.d/hadoop-s3guard.sh
hadoop-3.4.1/libexec/shellprofile.d/hadoop-aliyun.sh
hadoop-3.4.1/libexec/shellprofile.d/hadoop-streaming.sh
hadoop-3.4.1/libexec/shellprofile.d/hadoop-federation-balance.sh
hadoop-3.4.1/libexec/shellprofile.d/hadoop-hdfs.sh
hadoop-3.4.1/libexec/shellprofile.d/hadoop-rumen.sh
hadoop-3.4.1/libexec/shellprofile.d/hadoop-kafka.sh
hadoop-3.4.1/libexec/shellprofile.d/hadoop-mapreduce.sh
hadoop-3.4.1/libexec/shellprofile.d/hadoop-yarn.sh
hadoop-3.4.1/libexec/shellprofile.d/hadoop-distcp.sh
hadoop-3.4.1/libexec/shellprofile.d/hadoop-httpfs.sh
hadoop-3.4.1/libexec/shellprofile.d/hadoop-kms.sh
hadoop-3.4.1/libexec/shellprofile.d/hadoop-archives.sh
hadoop-3.4.1/libexec/shellprofile.d/hadoop-gridmix.sh
hadoop-3.4.1/libexec/shellprofile.d/hadoop-archive-logs.sh
hadoop-3.4.1/libexec/shellprofile.d/hadoop-extras.sh
hadoop-3.4.1/libexec/shellprofile.d/hadoop-aws.sh
hadoop-3.4.1/libexec/mapred-config.cmd
hadoop-3.4.1/libexec/yarn-config.cmd
hadoop-3.4.1/libexec/tools/
hadoop-3.4.1/libexec/tools/hadoop-streaming.sh
hadoop-3.4.1/libexec/tools/hadoop-federation-balance.sh
hadoop-3.4.1/libexec/tools/hadoop-rumen.sh
hadoop-3.4.1/libexec/tools/hadoop-dynamometer-workload.sh
hadoop-3.4.1/libexec/tools/hadoop-sls.sh
hadoop-3.4.1/libexec/tools/hadoop-dynamometer-blockgen.sh
hadoop-3.4.1/libexec/tools/hadoop-distcp.sh
hadoop-3.4.1/libexec/tools/hadoop-dynamometer-infra.sh
hadoop-3.4.1/libexec/tools/hadoop-archives.sh
hadoop-3.4.1/libexec/tools/hadoop-gridmix.sh
hadoop-3.4.1/libexec/tools/hadoop-archive-logs.sh
hadoop-3.4.1/libexec/tools/hadoop-extras.sh
hadoop-3.4.1/libexec/tools/hadoop-aws.sh
hadoop-3.4.1/libexec/tools/hadoop-resourceestimator.sh
ubuntu@ip-172-31-8-201:~$ ls
class  config.json  hadoop-3.4.1  hadoop-3.4.1.tar.gz
ubuntu@ip-172-31-8-201:~$ 
```



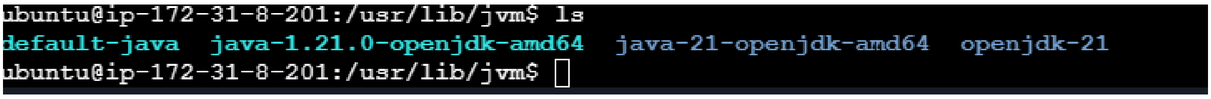
```
ubuntu@ip-172-31-8-201:~$ sudo apt update
Hit:1 http://us-east-2.ec2.archive.ubuntu.com/ubuntu noble InRelease
Get:2 http://us-east-2.ec2.archive.ubuntu.com/ubuntu noble-updates InRelease [126 kB]
Get:3 http://us-east-2.ec2.archive.ubuntu.com/ubuntu noble-backports InRelease [126 kB]
Get:4 http://security.ubuntu.com/ubuntu noble-security InRelease [126 kB]
Get:5 http://us-east-2.ec2.archive.ubuntu.com/ubuntu noble-updates/main amd64 Packages [838 kB]
Get:6 http://us-east-2.ec2.archive.ubuntu.com/ubuntu noble-updates/main Translation-en [191 kB]
Get:7 http://us-east-2.ec2.archive.ubuntu.com/ubuntu noble-updates/main amd64 Components [151 kB]
Get:8 http://us-east-2.ec2.archive.ubuntu.com/ubuntu noble-updates/universe amd64 Packages [1004 kB]
Get:9 http://us-east-2.ec2.archive.ubuntu.com/ubuntu noble-updates/universe Translation-en [251 kB]
Get:10 http://us-east-2.ec2.archive.ubuntu.com/ubuntu noble-updates/universe amd64 Components [315 kB]
Get:11 http://us-east-2.ec2.archive.ubuntu.com/ubuntu noble-updates/restricted amd64 Components [212 B]
Get:12 http://us-east-2.ec2.archive.ubuntu.com/ubuntu noble-updates/multiverse amd64 Components [940 B]
Get:13 http://us-east-2.ec2.archive.ubuntu.com/ubuntu noble-backports/main amd64 Components [208 B]
Get:14 http://us-east-2.ec2.archive.ubuntu.com/ubuntu noble-backports/universe amd64 Components [17.7 kB
Get:15 http://us-east-2.ec2.archive.ubuntu.com/ubuntu noble-backports/restricted amd64 Components [216 B
Get:16 http://us-east-2.ec2.archive.ubuntu.com/ubuntu noble-backports/multiverse amd64 Components [212 B
Get:17 http://security.ubuntu.com/ubuntu noble-security/main amd64 Packages [616 kB]
Get:18 http://security.ubuntu.com/ubuntu noble-security/main amd64 Components [8988 B]
Get:19 http://security.ubuntu.com/ubuntu noble-security/universe amd64 Packages [803 kB]
Get:20 http://security.ubuntu.com/ubuntu noble-security/universe Translation-en [171 kB]
Get:21 http://security.ubuntu.com/ubuntu noble-security/universe amd64 Components [52.0 kB]
Get:22 http://security.ubuntu.com/ubuntu noble-security/restricted amd64 Components [212 B]
Get:23 http://security.ubuntu.com/ubuntu noble-security/multiverse amd64 Components [212 B]
Fetched 4799 kB in 2s (2262 kB/s)
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
103 packages can be upgraded. Run 'apt list --upgradable' to see them.
ubuntu@ip-172-31-8-201:~$ 
```



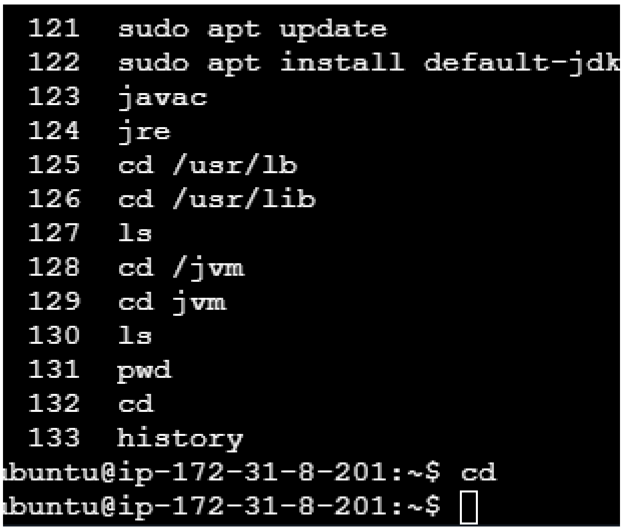
```
Scanning processes...
Scanning candidates...
Scanning linux images...

Pending kernel upgrade!
Running kernel version:
  6.8.0-1018-aws
Diagnostics:
  The currently running kernel version is not the expected kernel version 6.8.0-1021-aws.

Restarting the system to load the new kernel will not be handled automatically, so you should cons

Restarting services...

Service restarts being deferred:
 /etc/needrestart/restart.d/dbus.service
 systemctl restart networkd-dispatcher.service
 systemctl restart unattended-upgrades.service

No containers need to be restarted.

No user sessions are running outdated binaries.

No VM guests are running outdated hypervisor (qemu) binaries on this host.
ubuntu@ip-172-31-8-201:~$ 
```



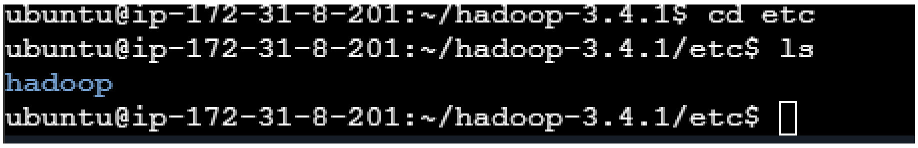
```
        Compile only the specified module(s), check timestamps
  --module-path <path>, -p <path>
        Specify where to find application modules
  --module-source-path <module-source-path>
        Specify where to find input source files for multiple modules
  --module-version <version>
        Specify version of modules that are being compiled
  -nowarn                      Generate no warnings
  -parameters
        Generate metadata for reflection on method parameters
  -proc:{none,only,full}
        Control whether annotation processing and/or compilation is done.
  -processor <class1>[,<class2>,<class3>...]
        Names of the annotation processors to run;
        bypasses default discovery process
  --processor-module-path <path>
        Specify a module path where to find annotation processors
  --processor-path <path>, -processorpath <path>
        Specify where to find annotation processors
  -profile <profile>
        Check that API used is available in the specified profile.
        This option is deprecated and may be removed in a future release.
  --release <release>
        Compile for the specified Java SE release.
        Supported releases:
            8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21
  -s <directory>               Specify where to place generated source files
  --source <release>, -source <release>
        Provide source compatibility with the specified Java SE release.
        Supported releases:
            8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21
  --source-path <path>, -sourcepath <path>
        Specify where to find input source files
  --system <jdk>|none          Override location of system modules
  --target <release>, -target <release>
        Generate class files suitable for the specified Java SE release.
        Supported releases:
            8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21
  --upgrade-module-path <path>
        Override location of upgradeable modules
  -verbose                     Output messages about what the compiler is doing
  --version, -version          Version information
  -Werror                      Terminate compilation if warnings occur

ubuntu@ip-172-31-8-201:~$
```



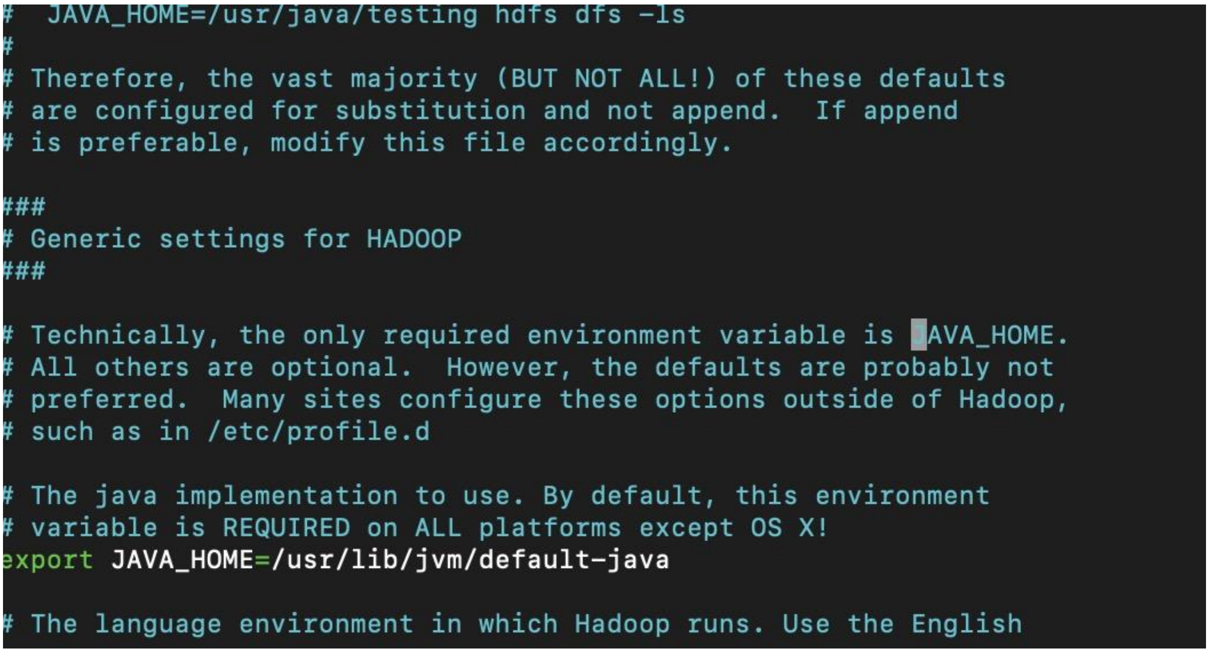
```
ubuntu@ip-172-31-8-201:/usr/lib/jvm$ ls
default-java  java-1.21.0-openjdk-amd64  java-21-openjdk-amd64  openjdk-21
ubuntu@ip-172-31-8-201:/usr/lib/jvm$
```



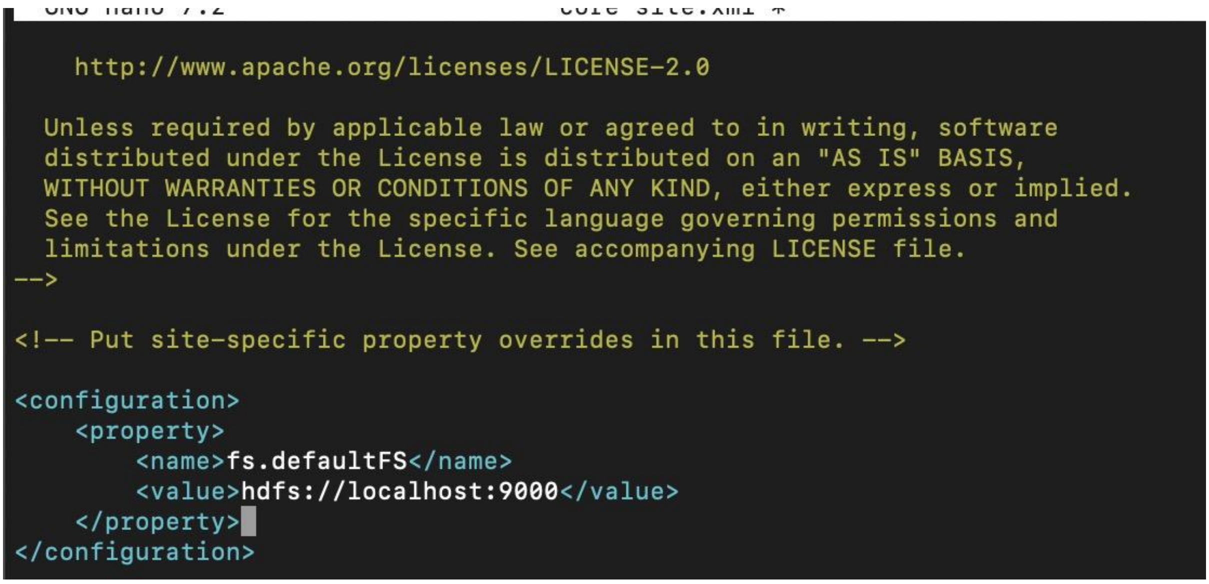
```
  121  sudo apt update
  122  sudo apt install default-jdk
  123  javac
  124  jre
  125  cd /usr/lb
  126  cd /usr/lib
  127  ls
  128  cd /jvm
  129  cd jvm
  130  ls
  131  pwd
  132  cd
  133  history
buntu@ip-172-31-8-201:~$ cd
buntu@ip-172-31-8-201:~$
```



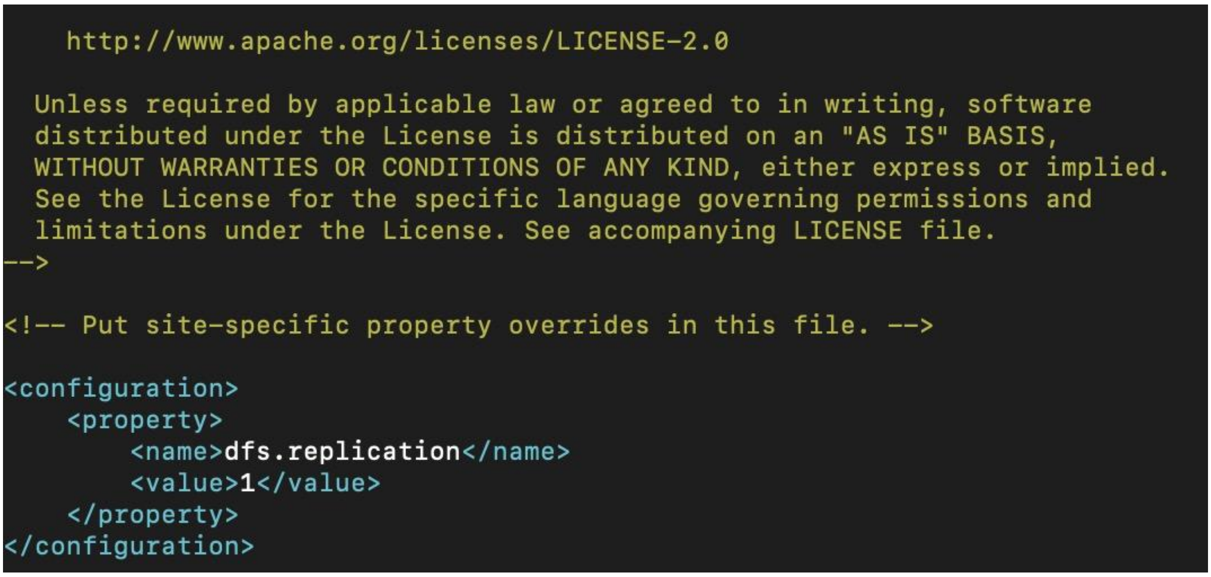
```
ubuntu@ip-172-31-8-201:~/hadoop-3.4.1$ cd etc
ubuntu@ip-172-31-8-201:~/hadoop-3.4.1/etc$ ls
hadoop
ubuntu@ip-172-31-8-201:~/hadoop-3.4.1/etc$ 
```



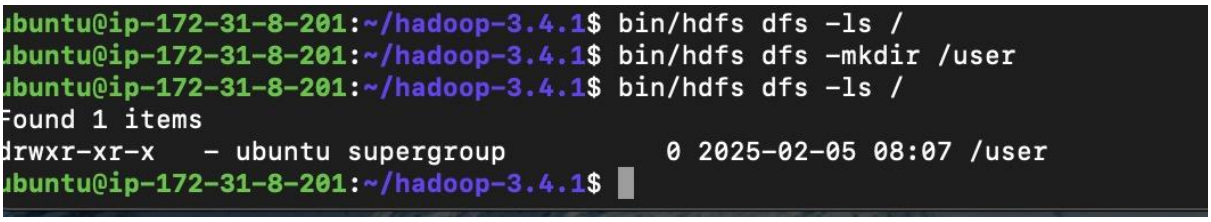
```
#   JAVA_HOME=/usr/java/testing hdfs dfs -ls
#
# Therefore, the vast majority (BUT NOT ALL!) of these defaults
# are configured for substitution and not append.  If append
# is preferable, modify this file accordingly.

###
# Generic settings for HADOOP
###

# Technically, the only required environment variable is JAVA_HOME.
# All others are optional.  However, the defaults are probably not
# preferred.  Many sites configure these options outside of Hadoop,
# such as in /etc/profile.d

# The java implementation to use. By default, this environment
# variable is REQUIRED on ALL platforms except OS X!
export JAVA_HOME=/usr/lib/jvm/default-java

# The language environment in which Hadoop runs. Use the English
```



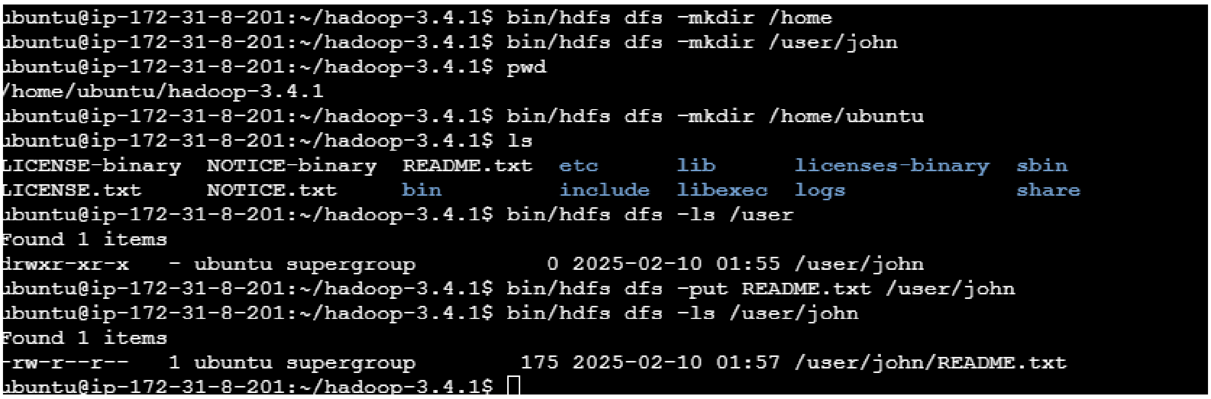
```
  GNU nano 7.2                          core-site.xml *

    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.defaultFS</name>
        <value>hdfs://localhost:9000</value>
    </property>
</configuration>
```



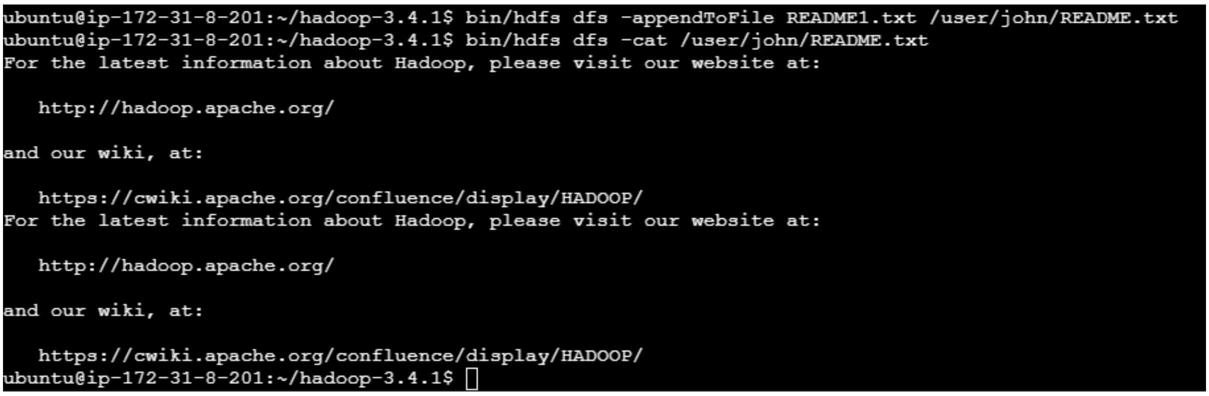
```
    http://www.apache.org/licenses/LICENSE-2.0

  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>
</configuration>
```



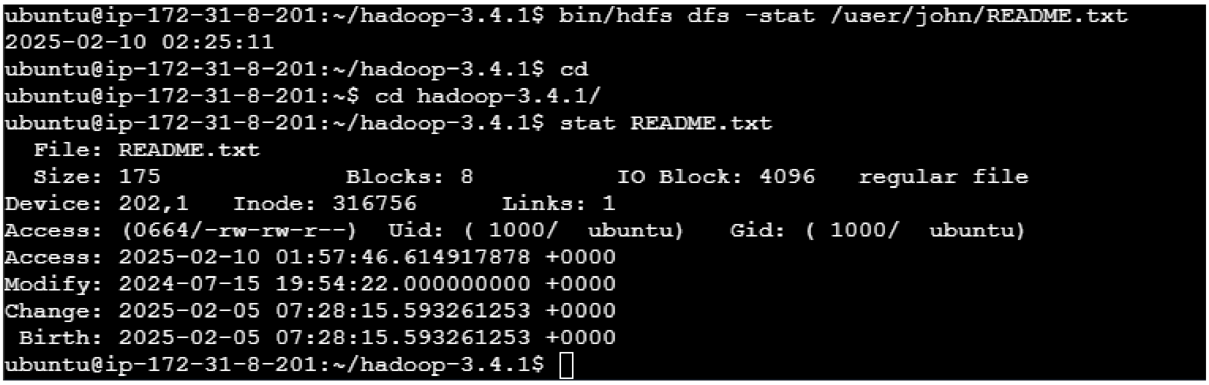
```
ubuntu@ip-172-31-8-201:~/hadoop-3.4.1$ bin/hdfs dfs -ls /
ubuntu@ip-172-31-8-201:~/hadoop-3.4.1$ bin/hdfs dfs -mkdir /user
ubuntu@ip-172-31-8-201:~/hadoop-3.4.1$ bin/hdfs dfs -ls /
Found 1 items
drwxr-xr-x   - ubuntu supergroup          0 2025-02-05 08:07 /user
ubuntu@ip-172-31-8-201:~/hadoop-3.4.1$
```



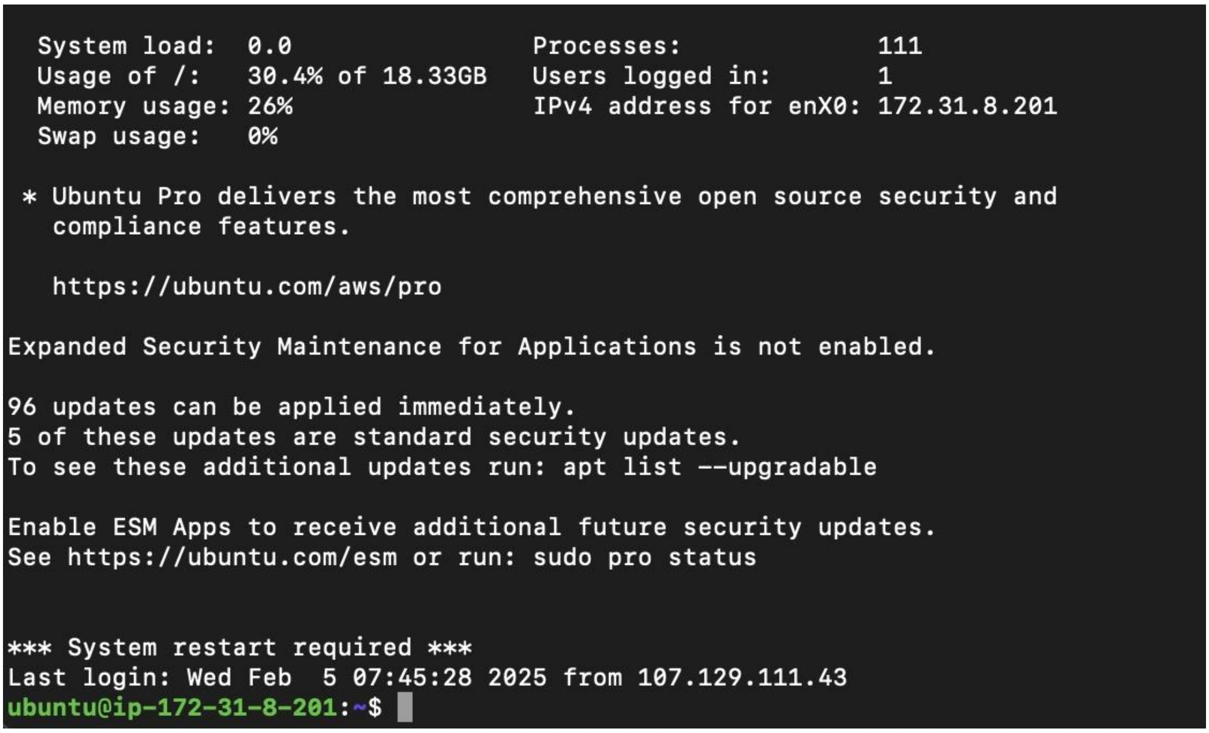
```
ubuntu@ip-172-31-8-201:~/hadoop-3.4.1$ bin/hdfs dfs -mkdir /home
ubuntu@ip-172-31-8-201:~/hadoop-3.4.1$ bin/hdfs dfs -mkdir /user/john
ubuntu@ip-172-31-8-201:~/hadoop-3.4.1$ pwd
/home/ubuntu/hadoop-3.4.1
ubuntu@ip-172-31-8-201:~/hadoop-3.4.1$ bin/hdfs dfs -mkdir /home/ubuntu
ubuntu@ip-172-31-8-201:~/hadoop-3.4.1$ ls
LICENSE-binary  NOTICE-binary  README.txt  etc      lib      licenses-binary  sbin
LICENSE.txt     NOTICE.txt     bin         include  libexec  logs             share
ubuntu@ip-172-31-8-201:~/hadoop-3.4.1$ bin/hdfs dfs -ls /user
Found 1 items
drwxr-xr-x   - ubuntu supergroup          0 2025-02-10 01:55 /user/john
ubuntu@ip-172-31-8-201:~/hadoop-3.4.1$ bin/hdfs dfs -put README.txt /user/john
ubuntu@ip-172-31-8-201:~/hadoop-3.4.1$ bin/hdfs dfs -ls /user/john
Found 1 items
-rw-r--r--   1 ubuntu supergroup        175 2025-02-10 01:57 /user/john/README.txt
ubuntu@ip-172-31-8-201:~/hadoop-3.4.1$
```



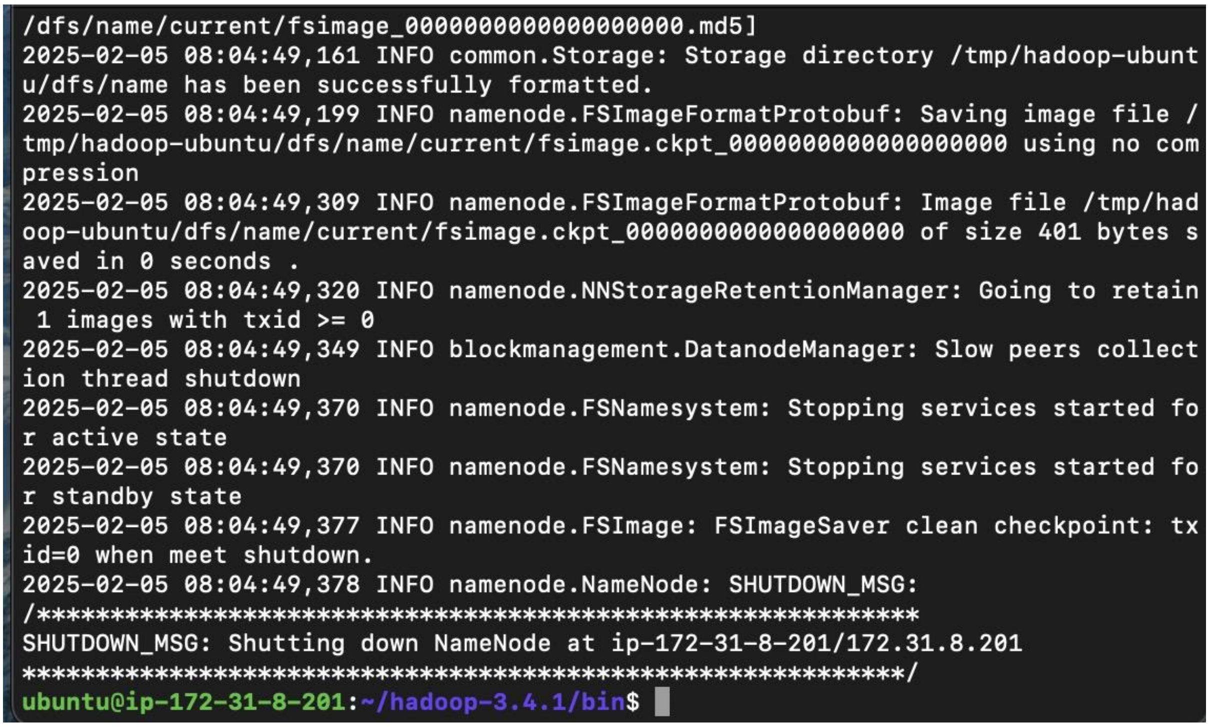
```
ubuntu@ip-172-31-8-201:~/hadoop-3.4.1$ bin/hdfs dfs -appendToFile README1.txt /user/john/README.txt
ubuntu@ip-172-31-8-201:~/hadoop-3.4.1$ bin/hdfs dfs -cat /user/john/README.txt
For the latest information about Hadoop, please visit our website at:

   http://hadoop.apache.org/

and our wiki, at:

   https://cwiki.apache.org/confluence/display/HADOOP/
For the latest information about Hadoop, please visit our website at:

   http://hadoop.apache.org/

and our wiki, at:

   https://cwiki.apache.org/confluence/display/HADOOP/
ubuntu@ip-172-31-8-201:~/hadoop-3.4.1$
```



```
ubuntu@ip-172-31-8-201:~/hadoop-3.4.1$ bin/hdfs dfs -stat /user/john/README.txt
2025-02-10 02:25:11
ubuntu@ip-172-31-8-201:~/hadoop-3.4.1$ cd
ubuntu@ip-172-31-8-201:~$ cd hadoop-3.4.1/
ubuntu@ip-172-31-8-201:~/hadoop-3.4.1$ stat README.txt
  File: README.txt
  Size: 175             Blocks: 8          IO Block: 4096   regular file
Device: 202,1   Inode: 316756      Links: 1
Access: (0664/-rw-rw-r--)  Uid: ( 1000/  ubuntu)   Gid: ( 1000/  ubuntu)
Access: 2025-02-10 01:57:46.614917878 +0000
Modify: 2024-07-15 19:54:22.000000000 +0000
Change: 2025-02-05 07:28:15.593261253 +0000
 Birth: 2025-02-05 07:28:15.593261253 +0000
ubuntu@ip-172-31-8-201:~/hadoop-3.4.1$ 
```



```
  System load:  0.0                Processes:             111
  Usage of /:   30.4% of 18.33GB   Users logged in:       1
  Memory usage: 26%                IPv4 address for enX0: 172.31.8.201
  Swap usage:   0%

 * Ubuntu Pro delivers the most comprehensive open source security and
   compliance features.

   https://ubuntu.com/aws/pro

Expanded Security Maintenance for Applications is not enabled.

96 updates can be applied immediately.
5 of these updates are standard security updates.
To see these additional updates run: apt list --upgradable

Enable ESM Apps to receive additional future security updates.
See https://ubuntu.com/esm or run: sudo pro status


*** System restart required ***
Last login: Wed Feb  5 07:45:28 2025 from 107.129.111.43
ubuntu@ip-172-31-8-201:~$ 
```



```
/dfs/name/current/fsimage_0000000000000000000.md5]
2025-02-05 08:04:49,161 INFO common.Storage: Storage directory /tmp/hadoop-ubunt
u/dfs/name has been successfully formatted.
2025-02-05 08:04:49,199 INFO namenode.FSImageFormatProtobuf: Saving image file /
tmp/hadoop-ubuntu/dfs/name/current/fsimage.ckpt_0000000000000000000 using no com
pression
2025-02-05 08:04:49,309 INFO namenode.FSImageFormatProtobuf: Image file /tmp/had
oop-ubuntu/dfs/name/current/fsimage.ckpt_0000000000000000000 of size 401 bytes s
aved in 0 seconds .
2025-02-05 08:04:49,320 INFO namenode.NNStorageRetentionManager: Going to retain
 1 images with txid >= 0
2025-02-05 08:04:49,349 INFO blockmanagement.DatanodeManager: Slow peers collect
ion thread shutdown
2025-02-05 08:04:49,370 INFO namenode.FSNamesystem: Stopping services started fo
r active state
2025-02-05 08:04:49,370 INFO namenode.FSNamesystem: Stopping services started fo
r standby state
2025-02-05 08:04:49,377 INFO namenode.FSImage: FSImageSaver clean checkpoint: tx
id=0 when meet shutdown.
2025-02-05 08:04:49,378 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at ip-172-31-8-201/172.31.8.201
************************************************************/
ubuntu@ip-172-31-8-201:~/hadoop-3.4.1/bin$
```