MODUL 9
Instalasi Hadoop

A. Tujuan :

1. Memahami tentang konsep Hadoop
2. Mampu melakukan instalasi Hadoop

B. Dasar Teori

Hadoop adalah library software (pustaka software) yang merupakan framework open source dari bahasa pemrograman Java di bawah lisensi Apache yang digunakan untuk melakukan pemrosesan big data menggunakan model pemrograman sederhana.
Hadoop dirancang untuk melakukan komputasi data dari satu server sampai ribuan server yang terhubung antara satu sama lainnya. Hal ini dapat memberikan kemudahan dari sisi penyimpanan data untuk melakukan analisis data. Selain itu Hadoop dapat memberikan informasi real time untuk mendeteksi kesalahan terkait kegagalan akses dan ketersediaan data pada masing-masing server.
Hadoop merupakan sebuah framework yang terus dikembangkan untuk melakukan pemrosesan big data. Berikut produk utama yang dikembangkan dalam Hadoop.

1. Hadoop Common

Hadoop Common adalah library-library umum yang mendukung library lainnya untuk dapat digunakan. Ini terkait perintah-perintah dasar yang ada pada Hadoop.

2. Hadoop Distributed File System (HDFS™)

Berbeda dengan system file data pada umumnya yaitu FAT32 dan NTFS yang dapat menyimpan 1 file data berkisaran antara 4 GB hingga 16 TB. HDFS adalah format sistem file yang dapat menampung 1 file data yang sangat besar dengan mengecilkan cluster sekelompok host data storage.

3. Hadoop YARN

Hadoop YARN adalah framework yang digunakan untuk mengatur pekerjaan secara terjadwal (schedule) dan manajemen cluster data.

4. Hadoop MapReduce

Hadoop MapReduce adalah paradigma pemrosesan data yang mengambil spesifikasi big data untuk menentukan bagaimana data tersebut dijadikan input dan output untuk diterapkan. MapReduce terintegrasi erat dengan HDFS untuk menyimpan data yang diperlukan.

Berikut beberapa produk yang dapat disandingkan dengan Hadoop:
1. Ambari™
Produk ini digunakan pada sistem yang berbasis web untuk penyediaan, pengelolaan, dan pemantauan cluster Apache Hadoop yang mencakup dukungan untuk HDFS Hadoop, Hadoop MapReduce, Hive, HCatalog, HBase, Zookeeper, Oozie, Pig, dan Sqoop. Ambari juga menyediakan dashboard untuk melihat kondisi klaster seperti

heatmap dan kemampuan untuk melihat kondisi aplikasi MapReduce, Babi dan Hive secara visual. Ambari juga dilengkapi fitur untuk mendiagnosis karakteristik kinerja Hadoop dengan antarmuka yang ramah.

2. Avro™
Avro™ adalah sistem serialisasi data.

3. Cassandra™
Cassandra™ adalah database multi-master yang dapat diukur untuk mengelola data yang berkapasitas besar.

4. Chukwa™
Chukwa™ adalah sistem pengumpulan data untuk mengelola sistem terdistribusi yang besar.

5. HBase™
HBase™ adalah database yang dapat diukur untuk mendukung penyimpanan data terstruktur dengan tabel yang besar.

6. Hive ™
Hive™ adalah Infrastruktur data warehouse yang menyediakan data summarization dan ad hoc querying.

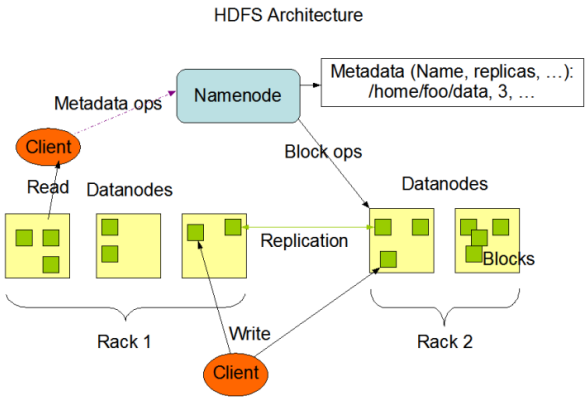
7. Mahout™
Mahout™ adalah library machine learning dan data mining.

8. Pig™
Pig™ adalah bahasa pemrograman tinggi aliran data (data-flow) yang digunakan melakukan eksekusi framework untuk melakukan komputasi data secara paralel.

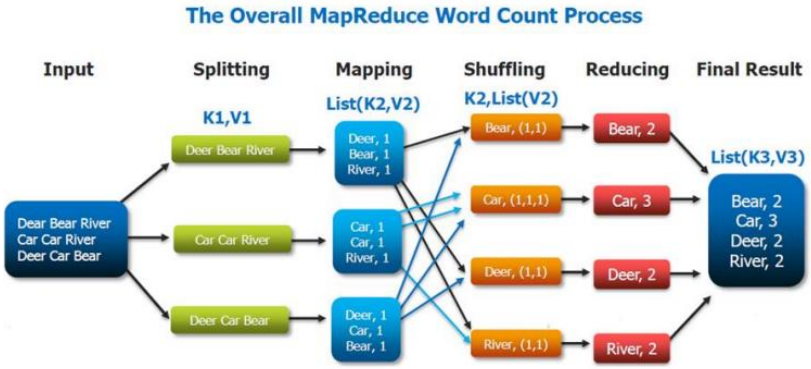
9. Spark™
Spark™ adalah model pemrograman yang digunakan untuk menghitung data dengan cepat. Spark menyediakan model pemrograman yang sederhana dan ekspresif yang mendukung berbagai aplikasi, termasuk ETL, machine learning, stream processing, dan graph computation.

10. Tez™
Tez™ adalah framework bahasa pemrograman untuk membangun data-flow.

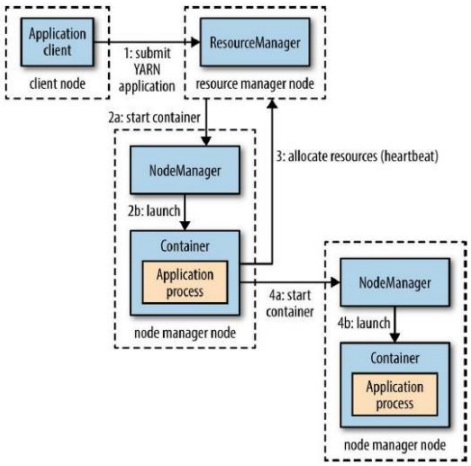
11. ZooKeeper™
ZooKeeper™ adalah layanan koordinasi untuk pendistribusian aplikasi dengan performa tinggi.



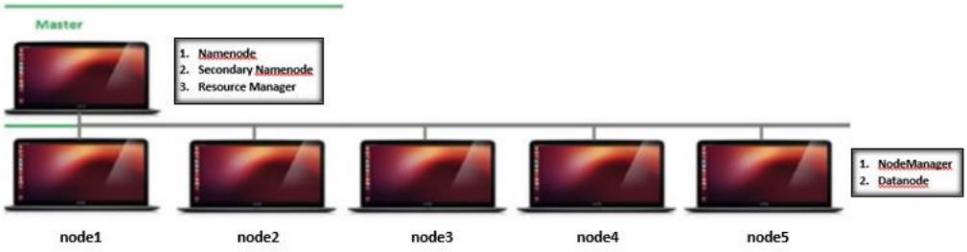
Gambar 1. Arsitektur HDFS



Gambar 2. Cara kerja pemrograman pada MapReduce



Gambar 3. Cara kerja YARN pada sebuah aplikasi



Gambar 4. Teknologi Hadoop Multinode

C. Tugas Pendahuluan

Pelajari konsep Hadoop dengan baik.

D. Percobaan

Instalasi Hadoop

1. Lakukan instalasi java:
   $ sudo apt install openjdk-8-jdk

```
zenhadi@zenhadi-virtual-machine:~$ sudo apt install openjdk-8-jdk
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following package was automatically installed and is no longer required:
  systemd-hwe-hwdb
```

2. Cek versi java:
   $ java -version

```
zenhadi@zenhadi-virtual-machine:~$ java -version
openjdk version "1.8.0_352"
OpenJDK Runtime Environment (build 1.8.0_352-8u352-ga-1~22.04-b08)
OpenJDK 64-Bit Server VM (build 25.352-b08, mixed mode)
```

3. Instalasi Hadoop

a. Download Hadoop

$wget https://dlcdn.apache.org/hadoop/common/hadoop-3.2.3/hadoop-3.2.3.tar.gz

```
hduser@zenhadi-virtual-machine:~$ wget https://dlcdn.apache.org/hadoop/common/h
adoop-3.2.3/hadoop-3.2.3.tar.gz
--2023-02-12 16:23:30--  https://dlcdn.apache.org/hadoop/common/hadoop-3.2.3/ha
doop-3.2.3.tar.gz
Resolving dlcdn.apache.org (dlcdn.apache.org)... 151.101.2.132, 2a04:4e42::644
Connecting to dlcdn.apache.org (dlcdn.apache.org)|151.101.2.132|:443... connect
ed.
HTTP request sent, awaiting response... 200 OK
Length: 492241961 (469M) [application/x-gzip]
Saving to: ‘hadoop-3.2.3.tar.gz’

hadoop-3.2.3.tar.gz     30%[======>                  ] 143,33M   293KB/s    eta
hadoop-3hadoop-3.2.3.tar.gz    67%[================>          ] 316,78M   251KB/s
hadoop-3.2.3.tar.gz 100%[===================>] 469,44M  76,7KB/s    in 27m 47s

2023-02-12 16:51:19 (288 KB/s) - ‘hadoop-3.2.3.tar.gz’ saved [492241961/4922419
61]
```

b. Extract dan install Hadoop
   $ tar -xvzf hadoop-3.2.3.tar.gz

```
zenhadi@zenhadi-virtual-machine:~$ tar -zxvf hadoop-3.2.3.tar.gz
```

Rename nama folder menjadi hadoop agar lebih sederhana:
$ sudo mv hadoop-3.2.3 hadoop

```
zenhadi@zenhadi-virtual-machine:~$ sudo mv hadoop-3.2.3 hadoop
```

Pastikan folder hadoop seperti ini:

```
zenhadi@zenhadi-virtual-machine:~$ ls /home/zenhadi/hadoop
bin  include  libexec      logs         README.txt  share
etc  lib      LICENSE.txt  NOTICE.txt   sbin
```

Berikut ini adalah file-file yang digunakan untuk konfigurasi single-node hadoop cluster:

> .bashrc
> core-site.xml => NAMENODE
> mapred-site.xml => RESOURCE MANAGER
> hdfs-site.xml
> yarn-site.xml

file-file tersebut bisa ditemukan pada direktori berikut:

$ ls /home/zenhadi/hadoop/etc/hadoop

```
zenhadi@zenhadi-virtual-machine:~$ ls /home/zenhadi/hadoop/etc/hadoop/
capacity-scheduler.xml            kms-log4j.properties
configuration.xsl                 kms-site.xml
container-executor.cfg            log4j.properties
core-site.xml                     mapred-env.cmd
hadoop-env.cmd                    mapred-env.sh
hadoop-env.sh                     mapred-queues.xml.template
hadoop-metrics2.properties        mapred-site.xml
hadoop-policy.xml                 shellprofile.d
hadoop-user-functions.sh.example  ssl-client.xml.example
hdfs-site.xml                     ssl-server.xml.example
httpfs-env.sh                     user_ec_policies.xml.template
httpfs-log4j.properties           workers
httpfs-signature.secret           yarn-env.cmd
httpfs-site.xml                   yarn-env.sh
kms-acls.xml                      yarnservice-log4j.properties
kms-env.sh                        yarn-site.xml
```

4. Update file bashrc
   $ nano .bashrc

```
#HADOOP VARIABLES START
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export HADOOP_INSTALL=/usr/local/hadoop
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/native”

export HADOOP_CONF_DIR=$HADOOP_INSTALL/etc/hadoop
export HADOOP_LOG_DIR=$HADOOP_INSTALL/logs
export PDSH_RCMD_TYPE=ssh
#HADOOP VARIABLES END
```

Pastikan versi java sudah sesuai di /usr/lib/jvm/

```
#HADOOP VARIABLES START
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export HADOOP_INSTALL=/home/zenhadi/hadoop
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/native"

export HADOOP_CONF_DIR=$HADOOP_INSTALL/etc/hadoop
export HADOOP_LOG_DIR=$HADOOP_INSTALL/logs
export PDSH_RCMD_TYPE=ssh
#HADOOP VARIABLES END
```

5. Selanjutnya masukkan perintah ini

```
$ cd hadoop/etc/hadoop
```

```
zenhadi@zenhadi-virtual-machine:~$ cd hadoop/etc/hadoop
```

```
$ nano hadoop-env.sh
```

a. Copy kode berikut lalu simpan

```
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
```

```
# For example, to limit who can execute the namenode command,
# export HDFS_NAMENODE_USER=hdfs
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
```

b. Edit file core-site.xml

$ nano core-site.xml

Copy kode berikut lalu simpan

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

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

c. Edit file mapred-site.xml dan masukkan perintah berikut

```
$ nano mapred-site.xml
```

Copy kode berikut lalu simpan

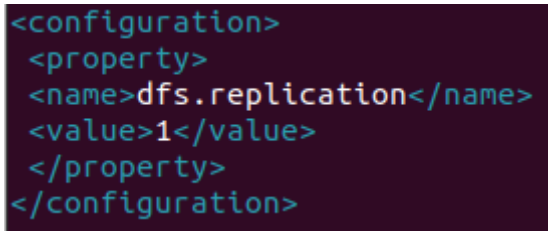
```
<configuration>
 <property>
 <name>mapreduce.framework.name</name>  <value>yarn</value>
 </property>
 <property>
 <name>mapreduce.application.classpath</name>

<value>$HADOOP_MAPRED_HOME/share/hadoop/mapreduce/*:$HADOOP_MAPRED_HOME/sh
are/hadoop/mapreduce/lib/*</value>
 </property>
</configuration>
```

d. Edit file hdfs-site.xml dan masukkan perintah berikut

```
$ nano hdfs-site.xml
```

Copy kode berikut lalu simpan

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



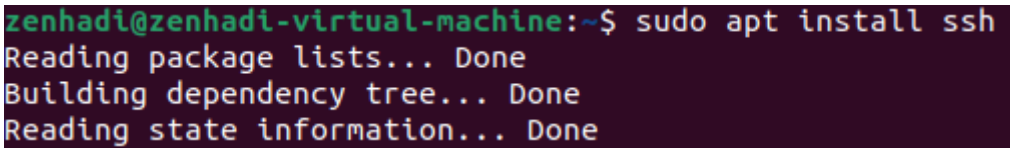
e. Edit file yarn-site.xml dan masukkan perintah berikut

$ nano yarn-site.xml

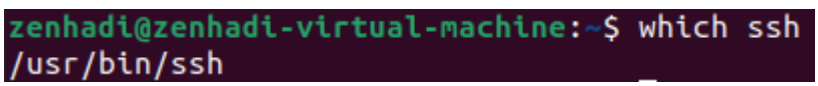
```
<configuration>
 <property>
 <name>yarn.nodemanager.aux-services</name>
 <value>mapreduce_shuffle</value>
 </property>
 <property>
 <name>yarn.nodemanager.env-whitelist</name>

<value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_CONF_DIR,CL
ASSPATH_PREP END_DISTCACHE,HADOOP_YARN_HOME,HADOOP_MAPRED_HOME
</value>
 </property>
</configuration>
```

6. Install ssh

$ sudo apt install ssh



Memastikan ssh sudah terinstal
$ which ssh



Generate ssh key untuk user Hadoop

$ ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
$cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
$chmod 0600 ~/.ssh/authorized_keys

```
hduser@zenhadi-virtual-machine:~$ ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
Generating public/private rsa key pair.
/home/hduser/.ssh/id_rsa already exists.
Overwrite (y/n)? y
Your identification has been saved in /home/hduser/.ssh/id_rsa
Your public key has been saved in /home/hduser/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:24qM9Kzm2FHAbZdQM5ubhP7dv5OZKx2gk7/QusEsntg hduser@zenhadi-virtual-machin
e
The key's randomart image is:
+---[RSA 3072]----+
|      ..+        |
|   . . o *       |
|    o + *        |
|     + o o  .    |
|      o S  o .   |
|     . . *+o  .  |
|    o   + Boo. = |
|   +.* = + +o.*  |
|  .o=.* E o..o++ |
+----[SHA256]-----+
hduser@zenhadi-virtual-machine:~$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
hduser@zenhadi-virtual-machine:~$ chmod 0600 ~/.ssh/authorized_keys
hduser@zenhadi-virtual-machine:/home/zenhadi$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ED25519 key fingerprint is SHA256:PVdyhNbV2VjYFjAIEOCMfe3ENms+1zXjDHWBc90zk3o.
This key is not known by any other names
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'localhost' (ED25519) to the list of known hosts.
Welcome to Ubuntu 22.04.1 LTS (GNU/Linux 5.15.0-60-generic x86_64)
```

Ketik yes lalu enter.

7. Restart vm

Setelah konfigurasi selesai restart OS anda
$ sudo reboot

8. Jalankan hadoop

$ /home/zenhadi/hadoop/bin/hdfs namenode -format

```
zenhadi@zenhadi-virtual-machine:~$ /home/zenhadi/hadoop/bin/hdfs namenode -form
at
WARNING: /home/zenhadi/hadoop/logs does not exist. Creating.
2023-02-20 09:07:46,567 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = zenhadi-virtual-machine/127.0.1.1
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 3.2.3
```

```
2023-02-20 09:07:48,427 INFO util.GSet: capacity      = 2^14 = 16384 entries
2023-02-20 09:07:48,492 INFO namenode.FSImage: Allocated new BlockPoolId: BP-20
29802879-127.0.1.1-1676858868471
2023-02-20 09:07:48,523 INFO common.Storage: Storage directory /tmp/hadoop-zenh
adi/dfs/name has been successfully formatted.
2023-02-20 09:07:48,608 INFO namenode.FSImageFormatProtobuf: Saving image file
/tmp/hadoop-zenhadi/dfs/name/current/fsimage.ckpt_0000000000000000000 using no
compression
2023-02-20 09:07:48,743 INFO namenode.FSImageFormatProtobuf: Image file /tmp/ha
doop-zenhadi/dfs/name/current/fsimage.ckpt_0000000000000000000 of size 402 byte
s saved in 0 seconds .
2023-02-20 09:07:48,758 INFO namenode.NNStorageRetentionManager: Going to retai
n 1 images with txid >= 0
2023-02-20 09:07:48,813 INFO namenode.FSNamesystem: Stopping services started f
or active state
2023-02-20 09:07:48,814 INFO namenode.FSNamesystem: Stopping services started f
or standby state
2023-02-20 09:07:48,824 INFO namenode.FSImage: FSImageSaver clean checkpoint: t
xid=0 when meet shutdown.
2023-02-20 09:07:48,828 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at zenhadi-virtual-machine/127.0.1.1
************************************************************/
```

Simpan file di /home/hduser/hadoop_dir/namenode-dir/

11. Mulai Hadoop

$ start-all.sh

```
zenhadi@zenhadi-virtual-machine:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as zenhadi in 10 seconds
.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [zenhadi-virtual-machine]
Starting resourcemanager
Starting nodemanagers
```

Untuk memverifikasi bahwa daemon namenode dan datanode berjalan, jalankan perintah diatas di terminal. Ini menampilkan proses Java yang sedang berjalan pada sistem.

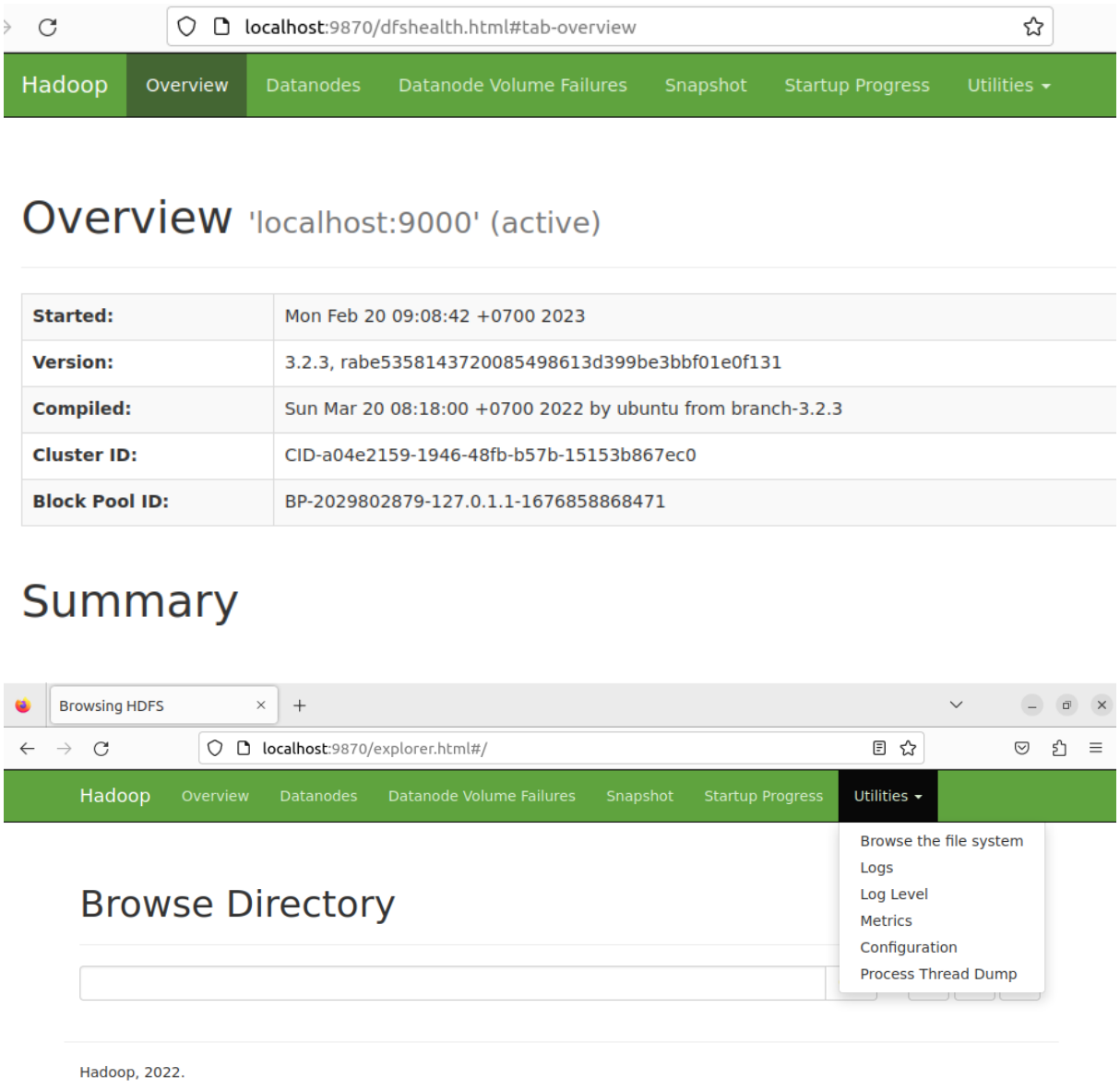
```
zenhadi@zenhadi-virtual-machine:~$ netstat -plten | grep java
(Not all processes could be identified, non-owned process info
 will not be shown, you would have to be root to see it all.)
tcp        0      0 0.0.0.0:13562           0.0.0.0:*               LISTEN
 1000       66214      3646/java
tcp        0      0 127.0.0.1:39687         0.0.0.0:*               LISTEN
 1000       58475      3156/java
tcp        0      0 127.0.0.1:9000          0.0.0.0:*               LISTEN
 1000       57413      3039/java
tcp        0      0 0.0.0.0:9870            0.0.0.0:*               LISTEN
 1000       56457      3039/java
tcp        0      0 0.0.0.0:9868            0.0.0.0:*               LISTEN
 1000       59893      3335/java
tcp        0      0 0.0.0.0:9866            0.0.0.0:*               LISTEN
 1000       58446      3156/java
tcp        0      0 0.0.0.0:9867            0.0.0.0:*               LISTEN
 1000       57933      3156/java
tcp        0      0 0.0.0.0:9864            0.0.0.0:*               LISTEN
 1000       58690      3156/java
tcp        0      0 0.0.0.0:37783           0.0.0.0:*               LISTEN
 1000       64622      3646/java
tcp        0      0 0.0.0.0:8042            0.0.0.0:*               LISTEN
 1000       64665      3646/java
```

JPS: Java Virtual Machine Process Status

```
zenhadi@zenhadi-virtual-machine:~$ jps
3538 ResourceManager
3156 DataNode
3335 SecondaryNameNode
5512 Jps
3646 NodeManager
3039 NameNode
```

Terlihat bahwa datanode dan namenode terletak di server yang sama saat diaplikasikan pada single node Hadoop. Saat berjalan di cluster, namenode tidak mengandung datanode. Jika namenode atau datanode belum berjalan, lihat file log selama start-dfs.sh berjalan.

12. Jalankan Hadoop pada browser



Untuk menghentikan Hadoop:
$ /usr/local/hadoop/sbin/stop-dfs.sh

```
hduser@zenhadi-virtual-machine:/home/zenhadi$ /usr/local/hadoop/sbin/stop-all.sh
WARNING: Stopping all Apache Hadoop daemons as hduser in 10 seconds.
WARNING: Use CTRL-C to abort.
Stopping namenodes on [localhost]
Stopping datanodes
Stopping secondary namenodes [zenhadi-virtual-machine]
Stopping nodemanagers
localhost: WARNING: nodemanager did not stop gracefully after 5 seconds: Trying to kill with kill -9
Stopping resourcemanager
```

stop-all.sh
sudo /usr/local/hadoop/bin/hadoop namenode -format
start-all.sh

E. Laporan Resmi :

1. Analisalah semua langkah-langkah instalasi diatas dan buat kesimpulan.