

Install Hadoop Single Node

Prepared By:

Sirojul Munir S.SI, M.KOM | rojulman@nurulfikri.ac.id

Persiapan

Pada tutorial ini akan dilakukan instalasi Hadoop modus Single Node pada platform sistem operasi linux. Berikut lingkungan server serta software pendukung yang digunakan.

1. Linux Ubuntu 18.04
2. Java JDK 1.8
3. Hadoop 3.2.0

Setingan Komputer Host Node Master

Seting penamaan komputer node master anda dengan mengedit file `/etc/hosts`. Misalkan IP Komputer anda adalah 192.0.2.1 dan nama hostnya adalah node-master.

```
$sudo gedit /etc/hosts
```

File `/etc/hosts` tambahkan baris berikut

```
192.0.2.1 node-master
```

User Sistem

Diperlukan account user dedicated hadoop yang akan melakukan proses administrasi sistem hadoop.

1. Buat account group user hadoop

```
$ sudo addgroup hadoop
```

2. Buat user hduser dan masukan dalam group hadoop

```
$ sudo adduser --ingroup hadoop hduser
```

```
Adding user `hduser' ...  
Adding new user `hduser' (1002) with group `hadoop' ...  
Creating home directory `/home/hduser' ...  
Copying files from `/etc/skel' ...  
Enter new UNIX password:  
Retype new UNIX password:  
passwd: password updated successfully
```

```
Changing the user information for hduser
Enter the new value, or press ENTER for the default
    Full Name []: hduser
    Room Number []:
    Work Phone []:
    Home Phone []:
    Other []:
Is the information correct? [Y/n] y
```

Konfigurasi SSH

Sistem Hadoop menghendaki akses melalui SSH untuk pengelolaan setiap node-nya. Artinya untuk setiap user yang akses melalui komputer lokal (maupun jaringan) haruslah terlebih dahulu melakukan konfigurasi SSH untuk dapat di akses dari komputer lokal untuk user **hduser**.

1. Instalasi SSH

```
$ sudo apt update
$ sudo apt install openssh-server
```

2. Aktifkan service SSH dan cek apakah SSH sudah running

```
$ sudo service ssh start
$ sudo systemctl status ssh
```

SSH berjalan pada port 22

3. Generate Key user hduser

```
$ su - hduser
$ ssh-keygen -t rsa -P ""
Generating public/private rsa key pair.
Enter file in which to save the key (/home/hduser/.ssh/id_rsa):
Created directory '/home/hduser/.ssh'.
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:
SHA256:0szKN+mecZQNRioQWce2KqrX030sWSZK6xqZTDwf78U
hduser@endjanuary
The key's randomart image is:
+---[RSA 2048]-----+
|  o*..  .             |
| .. =   o             |
|  . . . o             |
| . . . = . +          |
| .+.. . S o .         |
|++o o..o o            |
|=. .= ooE= .          |
|.o..o +oo=            |
|o++o +. =+           |
+---[SHA256]-----+
```

4. Beri hak akses otorisasi SSH untuk user hduser ke komputer lokal

```
$cat /home/hduser/.ssh/id_rsa.pub >> /home/hduser/.ssh/authorized_keys
```

5. Test koneksi SHH ke komputer lokal dengan user hduser

```
$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be
established.
ECDSA key fingerprint is
SHA256:YKmt9RyduY64Q3lCShWW5Qs5v8tyaI5WHiQd2rFSXKk.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'localhost' (ECDSA) to the list of
known hosts.
Welcome to Ubuntu 18.04.2 LTS (GNU/Linux 4.15.0-46-generic
x86_64)
```

Non Aktifkan IPV6

Salah satu permasalahan yang dihadapi pada saat install hadoop pada sistem ubuntu adalah IPV6, di ubuntu menggunakan IP 0.0.0.0. Karenanya penggunaan IPV6 perlu di non aktifkan. Untuk menonaktifkan IPV6 bisa melakukan konfigurasi pada file `/etc/systcl.conf`, dengan editor buka file dan lakukan setingan seperti berikut ini:

1. Buka file `/etc/systcl.conf` dengan editor text.

```
$sudo gedit /etc/systcl.conf
```

2. Tambahkan baris berikut pada file untuk men-disable IPV6

```
# disable ipv6
net.ipv6.conf.all.disable_ipv6 = 1
net.ipv6.conf.default.disable_ipv6 = 1
net.ipv6.conf.lo.disable_ipv6 = 1
```

3. Reboot komputer anda
4. Cek apakah IPV6 sudah tidak aktif

```
$ cat /proc/sys/net/ipv6/conf/all/disable_ipv6
```

Jika angka 1 yang dikembalikan maka IPV6 artinya disabled, dan 0 artinya IPV6 enabled.

Instalasi Java

Hadoop dalam proses bekerjanya membutuhkan Java terinstall dalam komputer anda. Pada tutorial ini menggunakan Java 1.8 (JDK 1.8).

1. Download paket binary Java JDK 1.8 (jdk-8u201-linux-x64.tar.gz) di website oracle (<https://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html>)

2. Urai paket tarball dalam direktori /opt

```
$ sudo cp lokasi_download/jdk-8u201-linux-x64.tar.gz /opt
$ sudo su -
# cd /opt
# tar -xvzf jdk-8u201-linux-x64.tar.gz
```

3. Tambahkan PATH Java pada sistem komputer anda (untuk user superuser root & hduser)

Tambahkan baris berikut pada akhir file **/root/.bashrc** dan **/home/hduser/.bashrc**

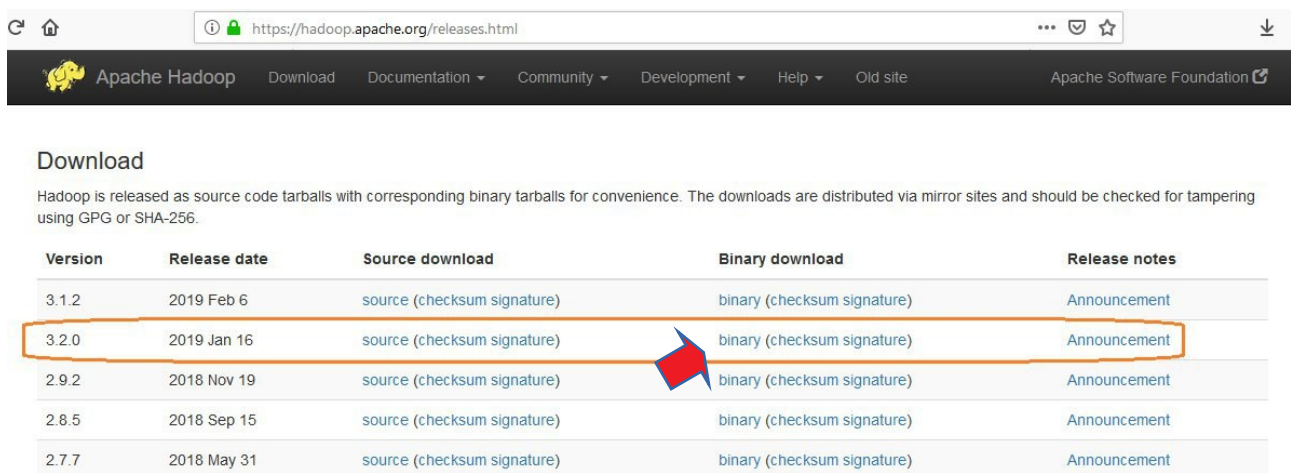
```
JAVA_HOME=/opt/jdk1.8.0_201
export JAVA_HOME
export PATH=$PATH:$JAVA_HOME/bin
```

4. Cek apakah environment user root dan hduser sudah mendukung Java

```
# java -version
java version "1.8.0_201"
Java(TM) SE Runtime Environment (build 1.8.0_201-b09)
Java HotSpot(TM) 64-Bit Server VM (build 25.201-b09, mixed mode)
```

Instalasi Hadoop

1. Download **binary hadoop 3.2.0** pada site: <http://hadoop.apache.org>



Download

Hadoop is released as source code tarballs with corresponding binary tarballs for convenience. The downloads are distributed via mirror sites and should be checked for tampering using GPG or SHA-256.

Version	Release date	Source download	Binary download	Release notes
3.1.2	2019 Feb 6	source (checksum signature)	binary (checksum signature)	Announcement
3.2.0	2019 Jan 16	source (checksum signature)	binary (checksum signature)	Announcement
2.9.2	2018 Nov 19	source (checksum signature)	binary (checksum signature)	Announcement
2.8.5	2018 Sep 15	source (checksum signature)	binary (checksum signature)	Announcement
2.7.7	2018 May 31	source (checksum signature)	binary (checksum signature)	Announcement

2. Ekstrak paket hadoop yang telah didownload dalam direktori instalasi yang dipilih (/usr/local/hadoop)

```
$ sudo tar -xvzf hadoop-3.2.0.tar.gz /usr/local
```

3. Ubah kepemilikan direktori dan file ke user **hduser** dan ke group **hadoop**

```
$ sudo cd /usr/local  
$ sudo mv hadoop-3.2.0 hadoop  
$ sudo chown -R hduser:hadoop hadoop
```

4. Tambahkan PATH hadoop untuk direktori eksekusi **/usr/local/hadoop/bin** dan **/usr/local/hadoop/sbin** pada sistem komputer anda dengan mengedit file **/root/.bashrc** dan **/home/hduser/.bashrc**. Tambahkan setingan berikut ini:

```
HADOOP_HOME=/usr/local/hadoop  
export HADOOP_HOME  
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin
```

5. Tes PATH Java dan Hadoop dalam lingkungan user **hduser**

```
$ sudo su - hduser  
$ echo $JAVA_HOME  
/opt/jdk1.8.0_201 --> output  
$ echo $HADOOP_HOME  
/usr/local/hadoop --> output
```

Seting Konfigurasi Master Node

Seting file **/etc/hadoop/hadoop-env.sh**

Buka file **/usr/local/hadoop/etc/hadoop/hadoop-env.sh** dengan editor favorit anda (misal: vim, atom, gedit) , dan tambahkan PATH lokasi untuk Java dan Hadoop.

```
$sudo gedit /usr/local/hadoop/etc/hadoop/hadoop-env.sh
```

```
# variable is REQUIRED on ALL platforms except OS X!  
export JAVA_HOME=/opt/jdk1.8.0_201
```

```
# this location based upon its execution path  
export HADOOP_HOME=/usr/local/hadoop
```

Seting Lokasi NameNode

Buka file **/usr/local/hadoop/etc/hadoop/core-site.xml** dan set lokasi NameNode ke **node-master** pad port 9000

```
$sudo gedit /usr/local/hadoop/etc/core-site.xml
```

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
  <configuration>
    <property>
      <name>fs.default.name</name>
      <value>hdfs://node-master:9000</value>
    </property>
  </configuration>
```

Seting PATH HDFS

Buka file `/usr/local/hadoop/etc/hadoop/hdfs-site.xml` dan set PATH dari HDFS

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

```
<configuration>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>/home/hduser/data/nameNode</value>
  </property>

  <property>
    <name>dfs.datanode.data.dir</name>
    <value>/home/hduser/data/dataNode</value>
  </property>

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

Arti dari nilai 1 dari setingan `dfs.replication` adalah mengindikasikan berapa kali replikasi dilakukan dalam cluster. Angka setingan disesuaikan dengan jumlah slave nodes.

Seting Yarn - Job Scheduler

1. Edit file `/usr/local/hadoop/etc/hadoop/mapred-site.xml` dan seting default dari framework operasi mapreduce.

```
$sudo gedit /usr/local/hadoop/etc/hadoop/mapred-site.xml
```

```
<configuration>
  <property>
    <name>mapreduce.framework.name</name>
    <value>yarn</value>
  </property>
  <property>
    <name>yarn.app.mapreduce.am.env</name>
    <value>HADOOP_MAPRED_HOME=${HADOOP_HOME}</value>
  </property>
  <property>
    <name>mapreduce.map.env</name>
    <value>HADOOP_MAPRED_HOME=${HADOOP_HOME}</value>
  </property>
```

```
<property>
  <name>mapreduce.reduce.env</name>
  <value>HADOOP_MAPRED_HOME=${HADOOP_HOME}</value>
</property>
</configuration>
```

2. Edit file /usr/local/hadoop/etc/hadoop/yarn-site.xml

```
$sudo gedit /usr/local/hadoop/etc/hadoop/yarn-site.xml
```

```
<configuration>
  <property>
    <name>yarn.acl.enable</name>
    <value>0</value>
  </property>

  <property>
    <name>yarn.resourcemanager.hostname</name>
    <value>node-master</value>
  </property>

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

Format HDFS

Sistem HDFS perlu di lakukan format seperti format file sistem lain. Jalankan perintah berikut untuk melakukan format HDFS.

```
$ sudo su - hduser
$ hdfs namenode -format
```

Menjalankan dan memonitor HDFS

Menjalankan HDFS

1. Jalankan HDFS dengan eksekusi skrip **start-dfs.sh**

```
$sudo su - hduser
$start-dfs.sh

Starting namenodes on [node-master]
Starting datanodes
Starting secondary namenodes [nama komputer]
```

2. Cek node yang dijalankan dengan perintah jps pada lingkungan user hduser

```
$ jps
```

```
19043 Jps
18394 NameNode
18605 DataNode
18863 SecondaryNameNode
```

3. Mematikan node HDFS dengan menjalankan skrip stop-dfs.sh

```
$stop-dfs.sh
```

Memonitoring HDFS

1. Cek service Hadoop dengan menjalankan skrip berikut pada user sistem linux

Buka konsole baru , dan jalankan berikut ini

```
$netstat -tanp | grep java
tcp        0      0 0.0.0.0:9864          0.0.0.0:*             LISTEN      18605/java
tcp        0      0 192.0.2.1:9000        0.0.0.0:*             LISTEN      18394/java
tcp        0      0 0.0.0.0:9866          0.0.0.0:*             LISTEN      18605/java
tcp        0      0 0.0.0.0:9867          0.0.0.0:*             LISTEN      18605/java
tcp        0      0 0.0.0.0:9868          0.0.0.0:*             LISTEN      18863/java
tcp        0      0 0.0.0.0:9870          0.0.0.0:*             LISTEN      18394/java
tcp        0      0 127.0.0.1:34303       0.0.0.0:*             LISTEN      18605/java
tcp        0      0 192.0.2.1:9000        192.0.2.1:39150       ESTABLISHED 18394/java
tcp        0      0 192.0.2.1:39150       192.0.2.1:9000        ESTABLISHED 18605/java
```

Pada contoh ini monitoring dapat dilihat melalui browser pada lokal komputer 127.0.0.1:34303

DataNode on endjanuary:9866

Cluster ID:	CID-36219b23-4f9e-4b0a-85de-2d051a86446c
Version:	3.2.0, re97acb3bd8f3befd27418996fa5d4b50bf2e17bf

Block Pools

Namenode Address	Block Pool ID	Actor State	Last Heartbeat	Last Block Report	Last Block Report Size (Max Size)
node-master:9000	BP-1876099162-127.0.1.1-1554117376888	RUNNING	1s	13 minutes	0 B (64 MB)

Volume Information

Directory	StorageType	Capacity Used	Capacity Left	Capacity Reserved	Reserved Space for Replicas	Blocks
/home/hduser/data/dataNode	DISK	32 KB	158.65 GB	0 B	0 B	0

Hadoop, 2019.

2. Informasi cluster HDFS juga dapat dilihat dengan menjalankan skrip **hdfs dfsadmin** pada konsole hduser

```
$ hdfs dfsadmin -report
```

outputnya sebagai berikut:

```
Configured Capacity: 180814237696 (168.40 GB)
Present Capacity: 170353012736 (158.65 GB)
DFS Remaining: 170352979968 (158.65 GB)
DFS Used: 32768 (32 KB)
DFS Used%: 0.00%
Replicated Blocks:
  Under replicated blocks: 0
  Blocks with corrupt replicas: 0
  Missing blocks: 0
  Missing blocks (with replication factor 1): 0
  Low redundancy blocks with highest priority to recover: 0
  Pending deletion blocks: 0
Erasure Coded Block Groups:
  Low redundancy block groups: 0
  Block groups with corrupt internal blocks: 0
  Missing block groups: 0
  Low redundancy blocks with highest priority to recover: 0
  Pending deletion blocks: 0
```

Live datanodes (1):

```
Name: 192.0.2.1:9866 (node-master)
Hostname: endjanuary
Decommission Status : Normal
Configured Capacity: 180814237696 (168.40 GB)
DFS Used: 32768 (32 KB)
Non DFS Used: 1205026816 (1.12 GB)
DFS Remaining: 170352979968 (158.65 GB)
DFS Used%: 0.00%
DFS Remaining%: 94.21%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Mon Apr 01 19:49:19 WIB 2019
Last Block Report: Mon Apr 01 19:29:28 WIB 2019
Num of Blocks: 0
```

Menjalankan Yarn

1. Untuk menjalankan Yarn jalankan perintah berikut pada konsol hduser

```
$start-yarn.sh
```

dan stop yarn dengan perintah berikut:

```
$stop-yarn.sh
```

2. Cek lewat perintah service yang dijalankan oleh java untuk mendapatkan port aplikasi yarn

```
$netstat -tanp | grep java
```

output perintah: yarn berjalan di komputer IP 192.0.2.1 port 8088

```
tcp        0      0 0.0.0.0:8040          0.0.0.0:*             LISTEN      20584/java
tcp        0      0 0.0.0.0:9864          0.0.0.0:*             LISTEN      18605/java
tcp        0      0 192.0.2.1:9000        0.0.0.0:*             LISTEN      18394/java
tcp        0      0 0.0.0.0:8042          0.0.0.0:*             LISTEN      20584/java
tcp        0      0 0.0.0.0:9866          0.0.0.0:*             LISTEN      18605/java
tcp        0      0 0.0.0.0:9867          0.0.0.0:*             LISTEN      18605/java
tcp        0      0 0.0.0.0:9868          0.0.0.0:*             LISTEN      18863/java
tcp        0      0 0.0.0.0:9870          0.0.0.0:*             LISTEN      18394/java
tcp        0      0 0.0.0.0:42327         0.0.0.0:*             LISTEN      20584/java
tcp        0      0 192.0.2.1:8088        0.0.0.0:*             LISTEN      20217/java
tcp        0      0 0.0.0.0:13562         0.0.0.0:*             LISTEN      20584/java
tcp        0      0 192.0.2.1:8030        0.0.0.0:*             LISTEN      20217/java
tcp        0      0 192.0.2.1:8031        0.0.0.0:*             LISTEN      20217/java
tcp        0      0 127.0.0.1:34303       0.0.0.0:*             LISTEN      18605/java
tcp        0      0 192.0.2.1:8032        0.0.0.0:*             LISTEN      20217/java
tcp        0      0 192.0.2.1:8033        0.0.0.0:*             LISTEN      20217/java
tcp        0      0 192.0.2.1:9000        192.0.2.1:39150       ESTABLISHED 18394/java
tcp        0      0 192.0.2.1:39150       192.0.2.1:9000        ESTABLISHED 18605/java
tcp        0      0 192.0.2.1:8031        192.0.2.1:52722       ESTABLISHED 20217/java
tcp        0      0 192.0.2.1:52722       192.0.2.1:8031        ESTABLISHED 20584/java
```

3. Buka browser anda dan ketik url : 192.0.2.1:8088

The screenshot shows the Hadoop cluster management web interface. The top navigation bar includes the Hadoop logo and the title 'Nodes of the cluster'. The main content area displays 'Cluster Metrics' and 'Cluster Nodes Metrics'. The 'Cluster Nodes Metrics' section shows a table with columns for Node Labels, Rack, Node State, Node Address, Node HTTP Address, Last health-update, Health-report, Containers, Allocation Tags, Mem Used, Mem Avail, VCores Used, VCores Avail, and Version. The table shows one node in a 'RUNNING' state with the address 'endjanuary:42327' and 'endjanuary:8042'. The bottom of the page indicates 'Showing 1 to 1 of 1 entries'.

Referensi:

1. <https://www.linode.com/docs/databases/hadoop/how-to-install-and-set-up-hadoop-cluster/>
2. <https://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-single-node-cluster/>