



JURUSAN TEKNOLOGI INFORMASI

Mata Kuliah Big Data

## 03. Infrastruktur Big Data Bagian-1 (HDFS)



# Topik

- Setup HDFS
- Konfigurasi VPN Client
- Mengakses *Cluster* Hadoop
- Sedikit Perintah Dasar



# *Topik-1: Setup HDFS*

---

---

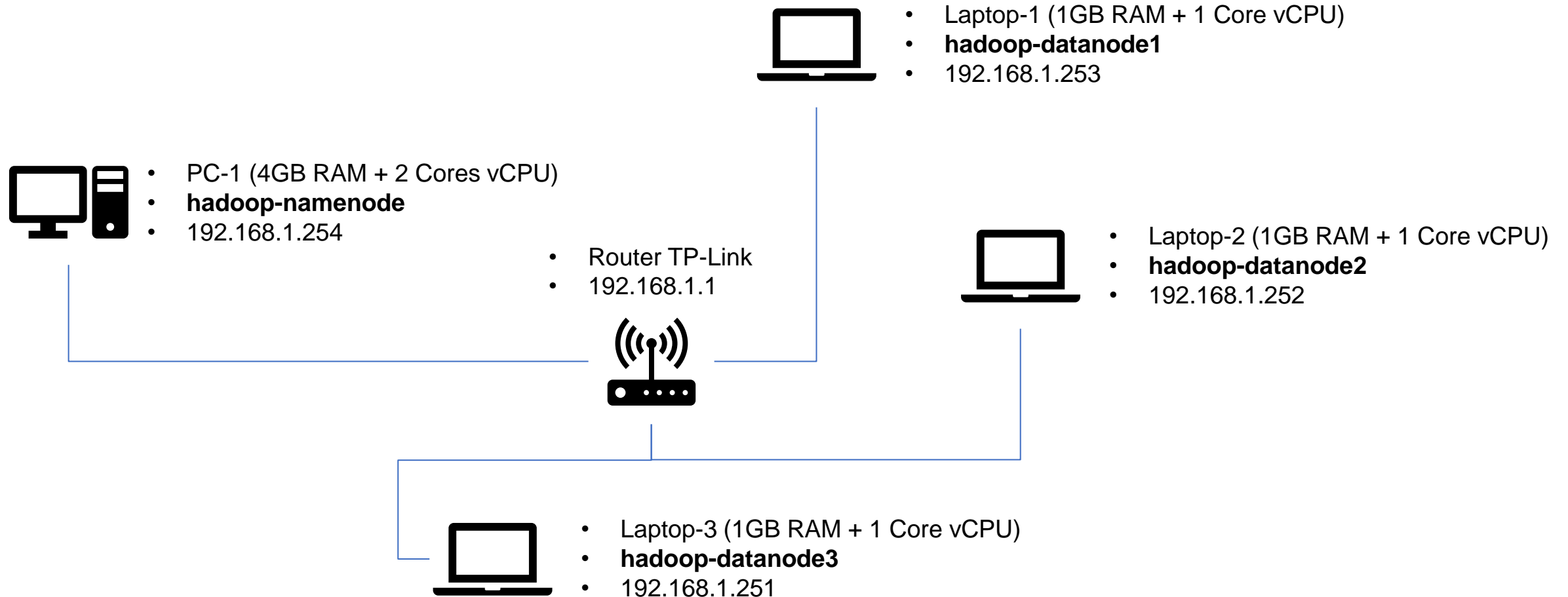
# 1. Setup HDFS

- Seperti sudah dijelaskan pada pertemuan sebelumnya, salah satu komponen sentral dalam sistem Big Data dengan Hadoop adalah **cluster**.
- **Cluster** → Kumpulan mesin/komputer (**nodes**) yang terhubung dalam satu jaringan, yang bekerja bersama-sama untuk menjadi tempat penyimpanan dan pemrosesan data besar (big data).
- Cluster terdiri dari:
  - **Data Node** → Tempat sesungguhnya dimana data disimpan dan diproses.
  - **Name Node** → Yang melakukan supervise terhadap data node, tempat dimana **metadata** disimpan.
- **Metadata** → Informasi tempat dimana **block** diletakkan.
- **Block** → Data yang diupload ke **HDFS** akan dipecah-pecah (split) dan di sebar ke data node.
  - Pecahan tersebut disebut dengan **block**.
  - Setiap block ukuran defaultnya 64 mb.
  - Setiap block secara default akan direplikasi ke minimal 3 data nodes yang berbeda-beda.
- Core Hadoop terdiri dari HDFS & MapReduce.
  - **HDFS** → File system dari Hadoop, yang bertugas mengatur penyimpanan file-file.
  - **MapReduce** → Algoritma, yang bertugas menjalankan pemrosesan terhadap file yang sudah disimpan.
- Untuk bisa berjalan, HDFS memerlukan minimal 1 komputer sebagai name node dan 3 computer sebagai datanode.

# 1. Setup HDFS

## Setup LAN

- Berikut ini adalah contoh setup Hadoop yang minimalis.



# 1. Setup HDFS

## Setup LAN



- Untuk bekerja dengan sebuah cluster Hadoop kita hanya perlu terhubung dengan **Name Node**-nya saja.
- Name Node dapat memerintahkan semua daemon yang terdapat pada data node yang terhubung dengannya untuk aktif maupun berhenti.
- Ketika akan mengupload file besar ke **HDFS**, maka filenya harus terlebih dulu ada pada file system name node.
  - **Contoh:** Jika OS Name Node Anda adalah Windows, maka file Anda harus berada pada drive-drive Windows: C:\mydata.txt, D:\mydata.txt, dlsb.
- Untuk bekerja dengan Name Node, misal: Mengunggah file, mendownload file, maka Anda harus menggunakan terminal/command prompt pada mesin Name Node tersebut.
  - Tentu saja harus login dulu.
- Bagaimana jika ingin mengakses via internet?
  - Hubungkan dengan VPN
  - Gunakan SSH
  - Jika di Windows, install **Putty**.

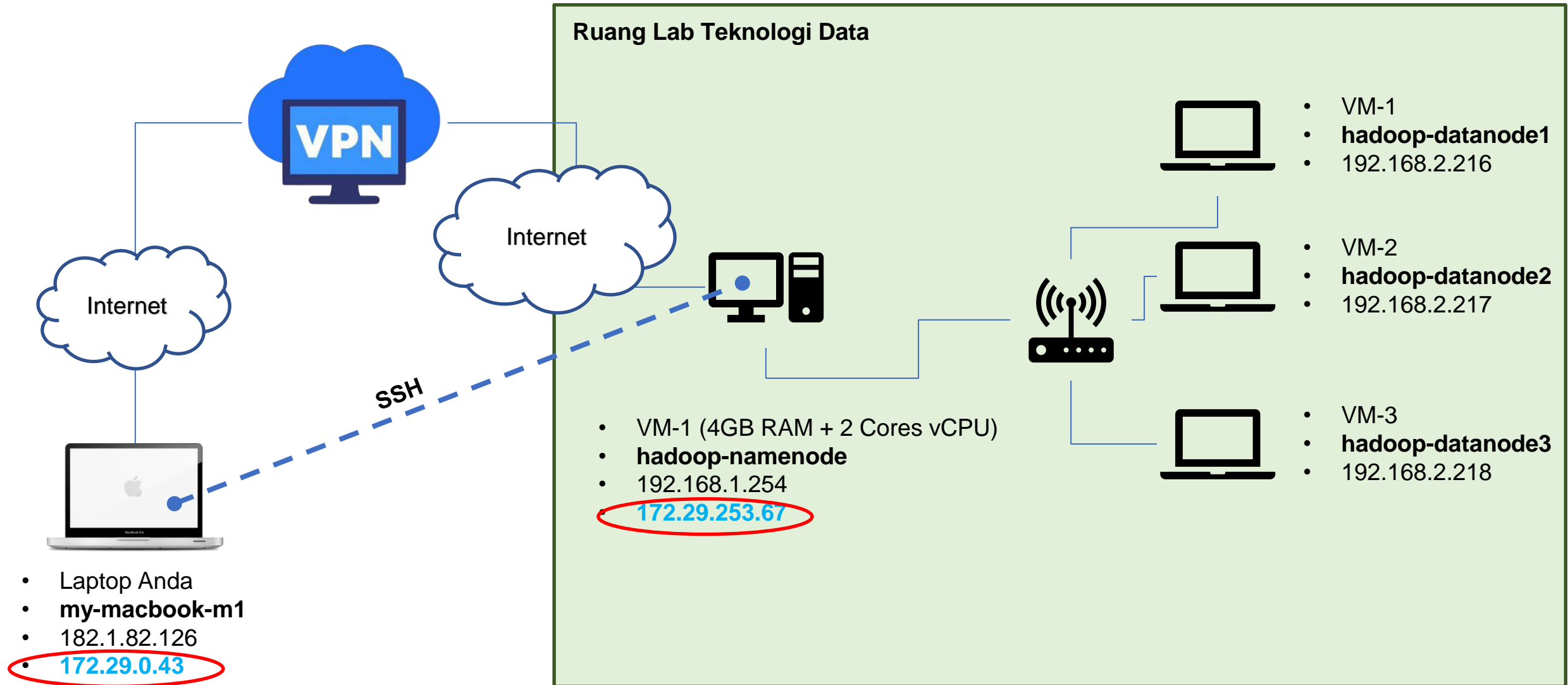
## 1. Setup HDFS

# Mengakses Cluster Hadoop Lab Teknologi Data via Internet

- Tim dosen dari Lab Teknologi Data telah menyiapkan sebuah cluster Hadoop di ruang RT-8 di lantai-8 Gedung ini.
- Cluster Hadoop tersebut terdiri dari:
  - 4 VM yang berjalan di dua computer yang berbeda.
  - Masing-masing VM dapat dianggap sebagai satu **node** yang beroperasi sendiri, independen antara satu dan yang lainnya.
- Sehingga secara keseluruhan terdapat 4 PC yang terdiri dari:
  - **1 Namenode** (2 vCPU, 4 GB RAM, 10 GB Storage)
  - **3 Datanode** (1 vCPU, 1 GB RAM, 10 GB Storage)
- Disebabkan topologi jaringan gedung sipil saat ini kita **tidak bisa mengakses** cluster tersebut dengan IP Local.
  - Oleh karenanya kita memerlukan **VPN** untuk mengakses cluster tersebut **via internet**.
- Untuk terhubung dengan cluster tersebut, dilakukan dengan protokol **SSH**.

# 1. Setup HDFS

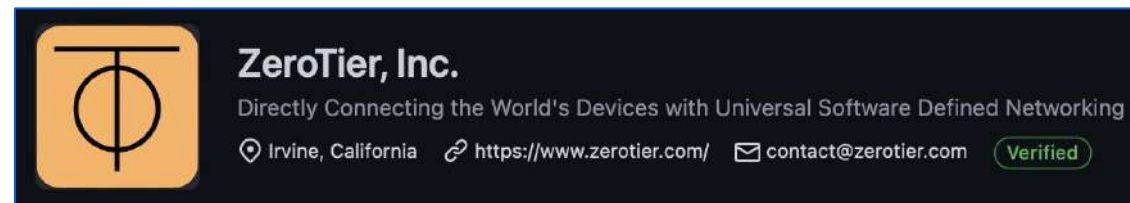
## Mengakses Cluster Hadoop Lab Teknologi Data via Internet





## 1. Setup HDFS VPN

- **VPN** → Virtual Private Network.
- Dengan menggunakan VPN, komputer Anda yang berada di tempat lain akan menjadi seolah-olah dalam satu jaringan LAN.
  - Komputer Anda akan diberi IP lain, yang serupa dengan IP Name Node.
- Dengan demikian, Anda akan dapat menjalankan perintah-perintah Hadoop secara “langsung” pada NameNode melalui SSH.
- Untuk dapat menggunakan VPN Anda harus menginstall VPN client pada computer Anda dan computer Name Node.
- VPN ada yang berbayar dan ada yang gratis. Pada topik ini, kita gunakan VPN gratis bernama **ZeroTier**.



## *Topik-2: Konfigurasi VPN Client*

---

---

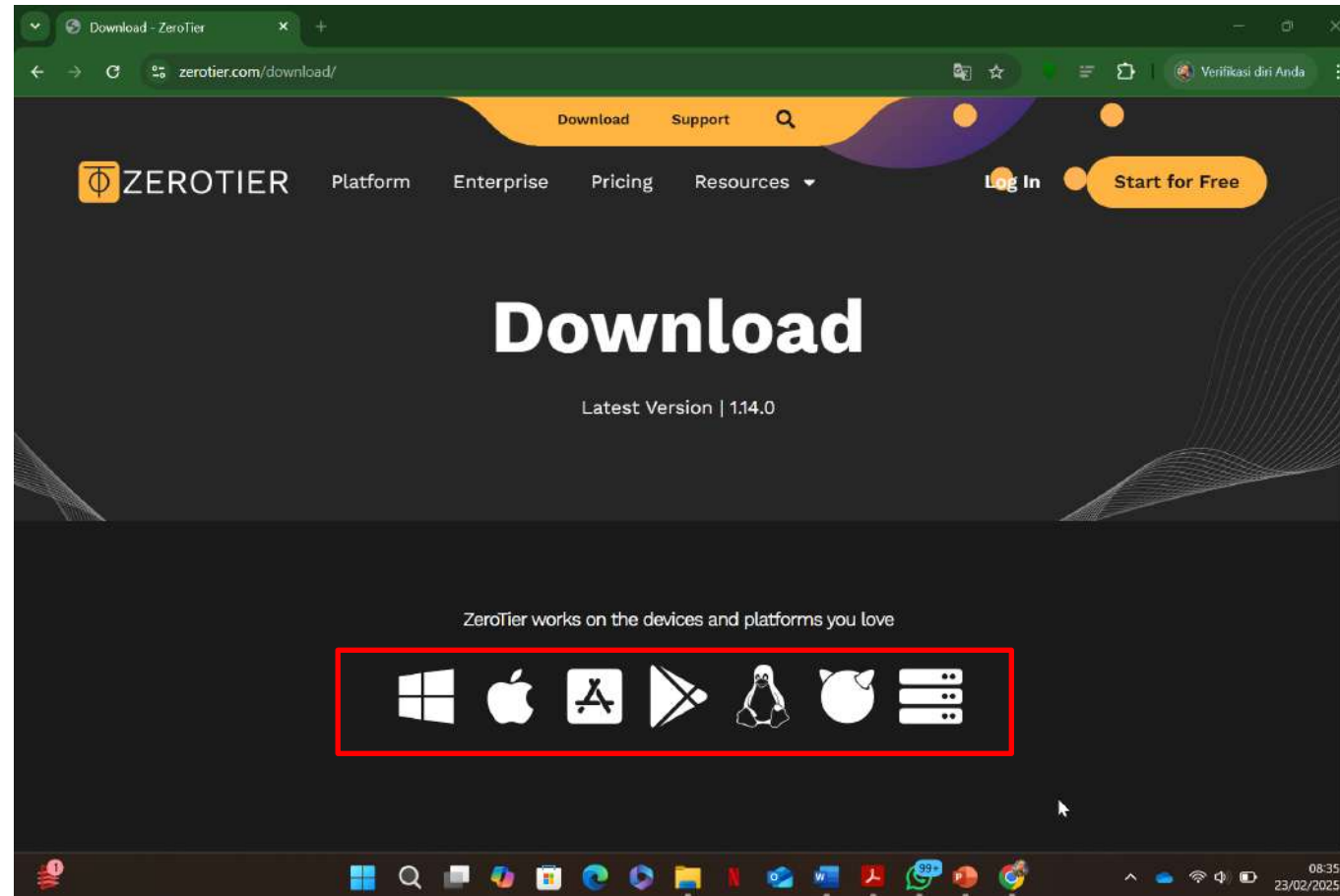
## 2. Konfigurasi VPN Client

- Dalam satu jaringan VPN, terdapat VPN Server dan VPN Client.
- VPN Server berada pada server milik layanan penyedia VPN.
- VPN Client harus diinstall pada komputer Anda dan pada komputer name node pada cluster Hadoop.
  - VPN pada cluster Hadoop sudah dikonfigurasi.
- Anda hanya perlu mengkonfigurasi VPN client di laptop Anda.
- Secara umum Langkah konfigurasi VPN client sangatlah mudah yaitu sebagai berikut:
  1. Unduh VPN Client.
  2. Install VPN Client.
  3. Bergabung ke jaringan VPN.

## 2. Konfigurasi VPN Client

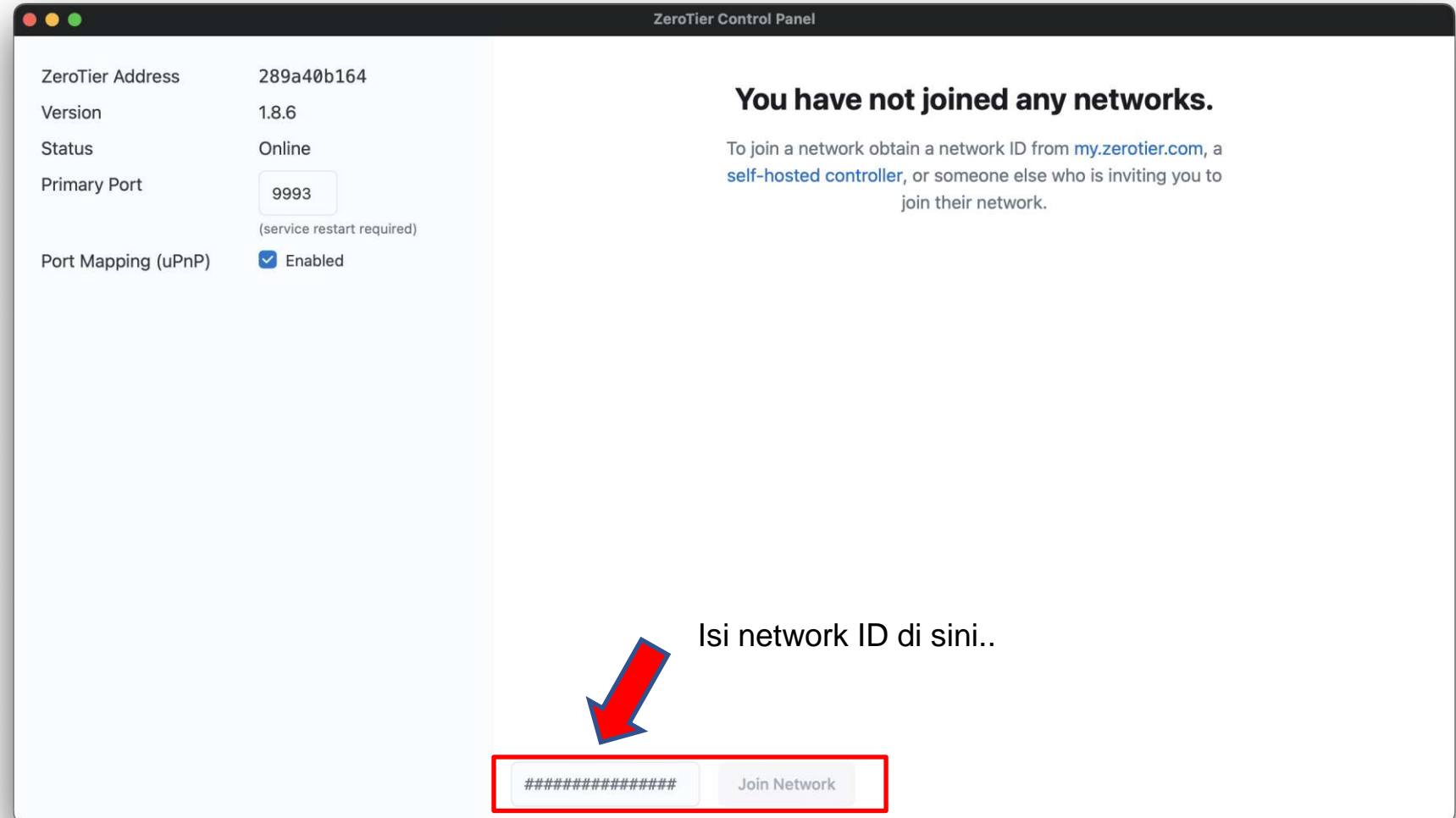
### Unduh VPN Client

- Buka situs ZeroTier: <https://www.zerotier.com/download/>
- Klik pada logo OS yang sesuai dengan OS pada komputer Anda.



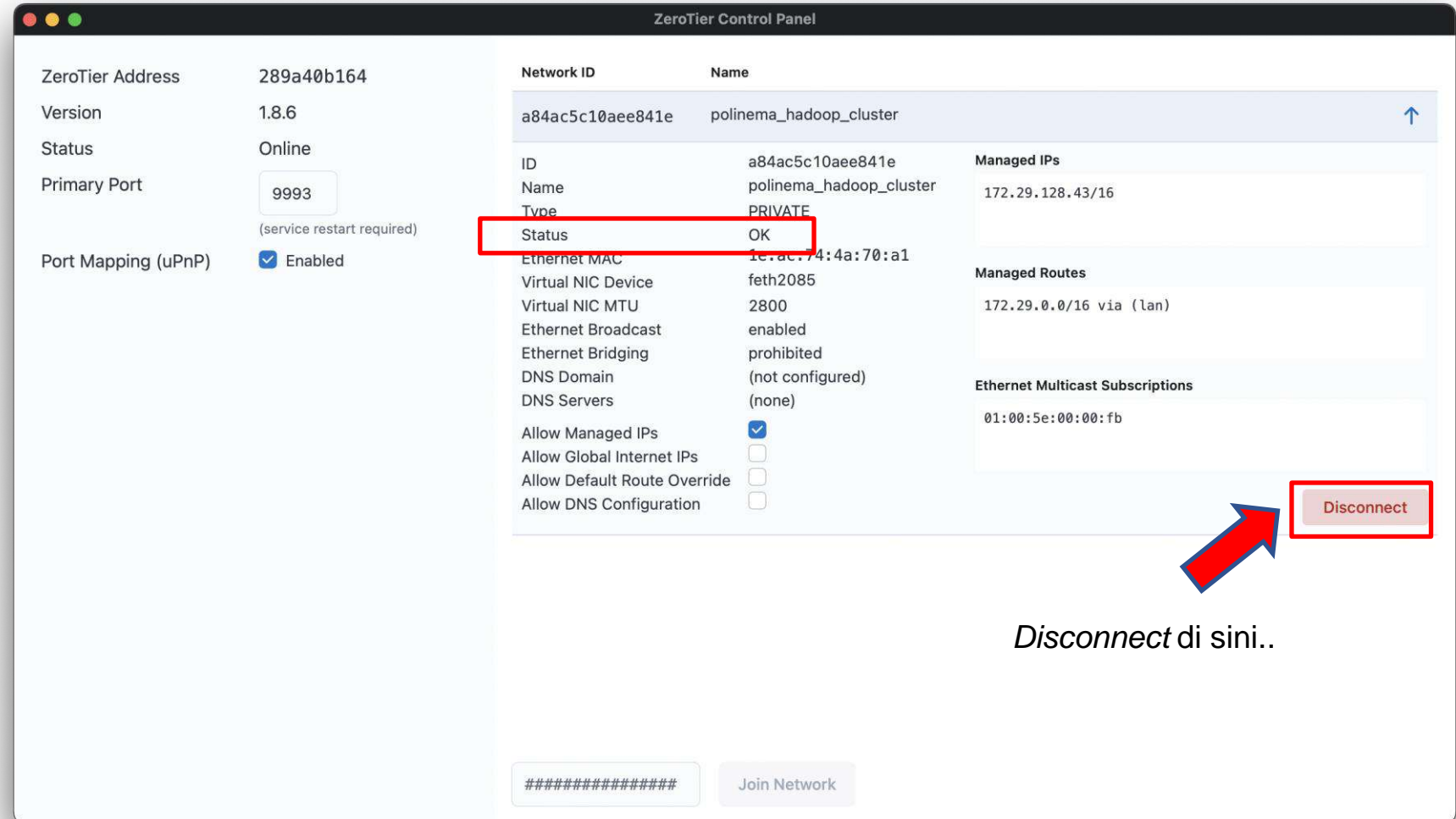
## 2. Konfigurasi VPN Client Install VPN Client

- Install software yang telah diunduh. Prosesnya sangat mudah, cukup klik 2x, next..next..
- Setelah selesai, buka ZeroTier control panel.
- Masukkan Network ID VPN dari saya di kotak bertanda ### di bagian bawah:
  - **a84ac5c10aee841e**
- Klik Join Network.



## 2. Konfigurasi VPN Client Bergabung ke Jaringan VPN

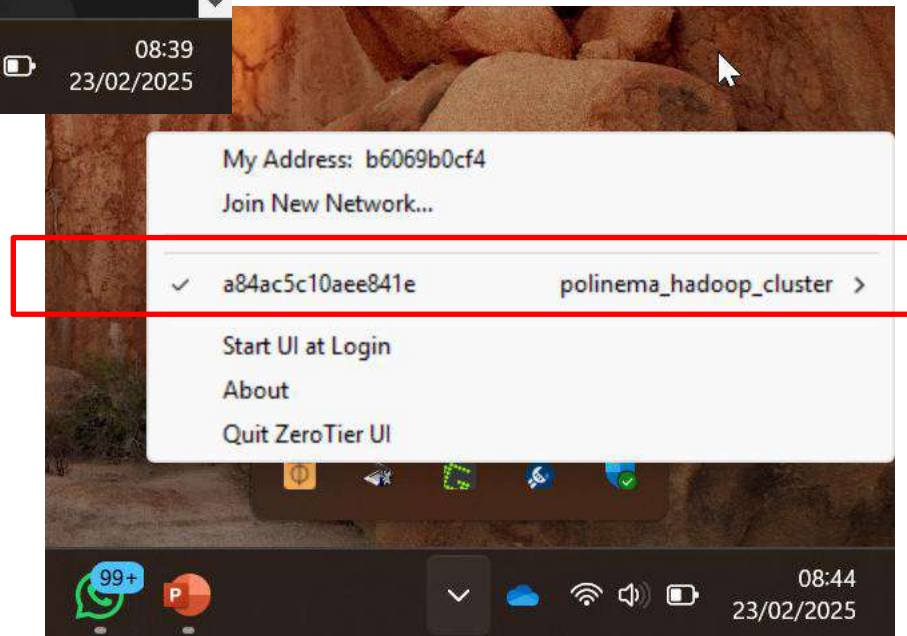
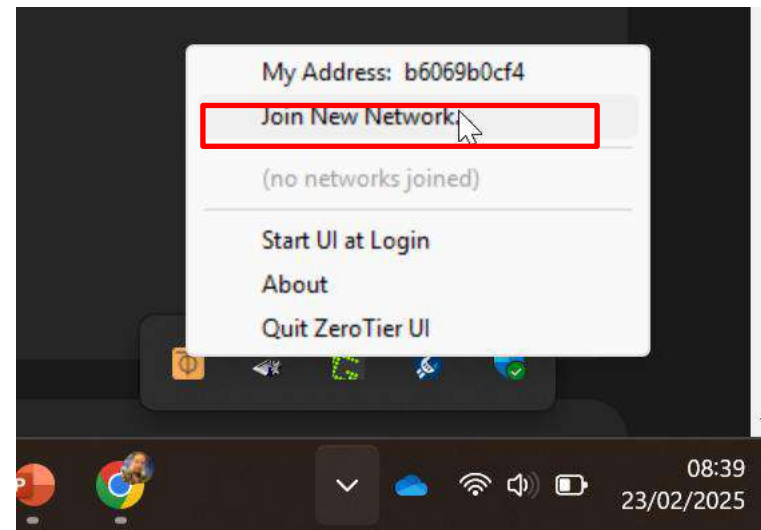
- Jika alamat yang Anda masukkan benar, maka tampilan control panel akan menjadi seperti di samping.
- Status “OK” artinya Anda sudah terhubung.
- Jika “DENIED”, berarti belum diizinkan oleh server.
  - Hubungi saya.
- Jika sudah selesai praktikum, *disconnect*-lah dari jaringan VPN agar browsing Anda tidak menjadi lambat.



Disconnect di sini..

## 2. Konfigurasi VPN *Client* Bergabung ke Jaringan VPN (dari OS Windows)

- Klik kanan ikon ZeroTier di system tray.
- Klik “Join New Network”.
- Masukkan network ID yang sudah dijelaskan sebelumnya.



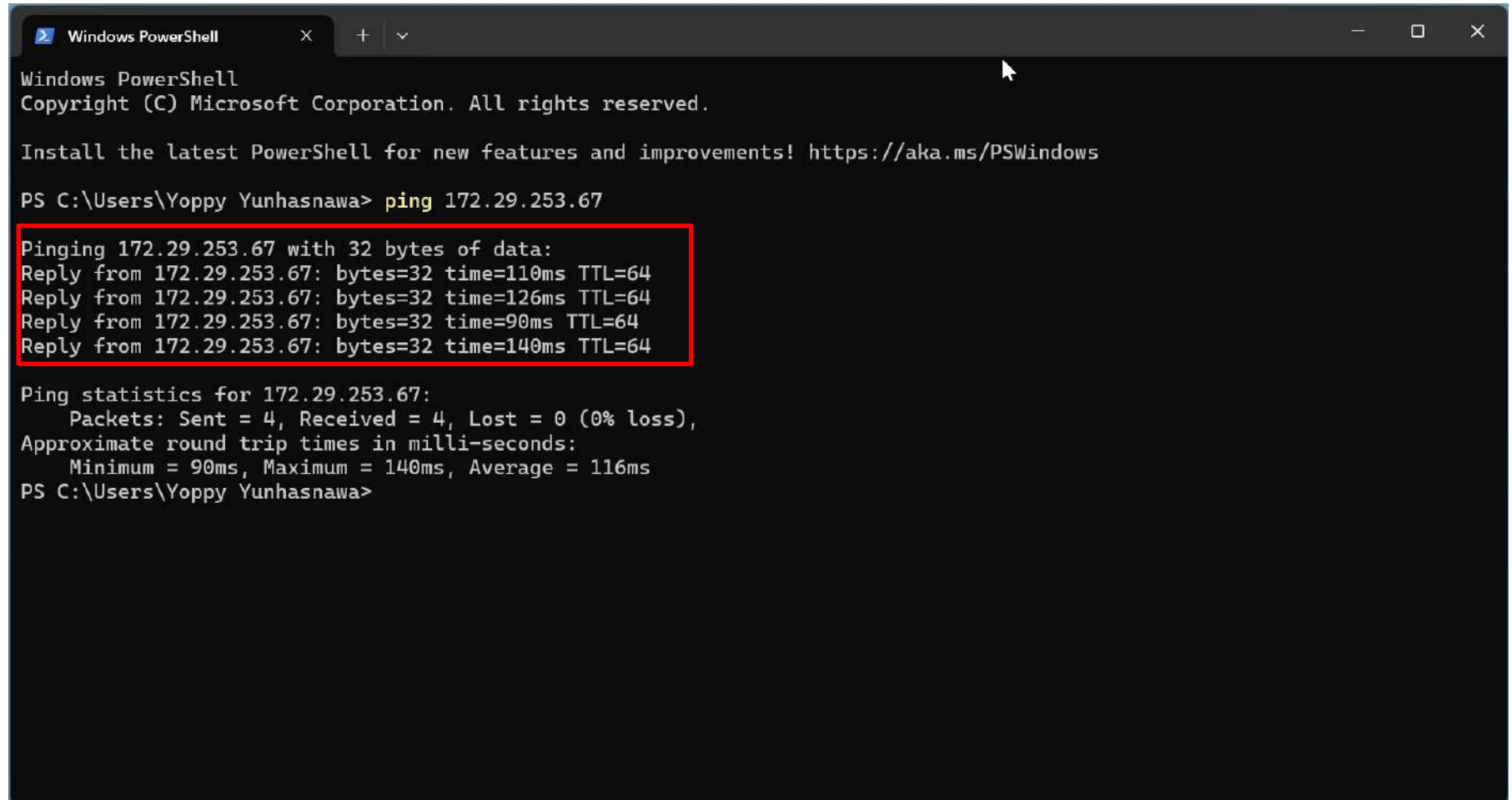
Ini tandanya  
sudah *connect*..



## 2. Konfigurasi VPN Client

### Install VPN Client

- Untuk menguji apakah Anda sudah terhubung dengan jaringan VPN cluster Hadoop ini, bukalah CMD atau terminal di computer Anda. Lalu ping IP berikut:
  - 172.29.253.67



```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Install the latest PowerShell for new features and improvements! https://aka.ms/PSWindows

PS C:\Users\Yoppy Yunhasnawa> ping 172.29.253.67

Pinging 172.29.253.67 with 32 bytes of data:
Reply from 172.29.253.67: bytes=32 time=110ms TTL=64
Reply from 172.29.253.67: bytes=32 time=126ms TTL=64
Reply from 172.29.253.67: bytes=32 time=90ms TTL=64
Reply from 172.29.253.67: bytes=32 time=140ms TTL=64

Ping statistics for 172.29.253.67:
    Packets: Sent = 4, Received = 4, Lost = 0 (0% loss),
    Approximate round trip times in milli-seconds:
        Minimum = 90ms, Maximum = 140ms, Average = 116ms
PS C:\Users\Yoppy Yunhasnawa>
```



# *Topik-3: Mengakses Cluster Hadoop*

---

---

### 3. Mengakses *Cluster* Hadoop

- Untuk mengakses suatu cluster Hadoop, sekali lagi, Anda hanya perlu terhubung dengan **name node**-nya saja.
- Ketika komputer Anda sudah terhubung dengan VPN cluster Hadoop, maka Anda dapat mencoba untuk terhubung dengan name node melalui SSH.
- Jika Anda di macOS/Linux:
  - Buka terminal dan ketikkan perintah: `ssh hadoopuser@172.29.253.67`
  - Masukkan password: **hadoop**
- Jika Anda di Windows, gunakan **PowerShell** atau (*install*) **Putty**.

```
yunhasnawa — hadoopuser@hadoop-namenode: ~ — ssh hadoop...
[yunhasnawa@Yunhasnawa-M1 ~] % ssh hadoopuser@172.29.247.62
[hadoopuser@172.29.247.62's password:
Welcome to Ubuntu 24.04.2 LTS (GNU/Linux 6.8.0-53-generic x86_64)

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

System information as of Sun Feb 23 02:11:36 AM UTC 2025

System load:  0.05          Processes:            111
Usage of /:   61.7% of 9.74GB Users logged in:           2
Memory usage: 20%          IPv4 address for enp0s3: 192.168.2.204
Swap usage:   0%

 * Strictly confined Kubernetes makes edge and IoT secure. Learn how MicroK8s
   just raised the bar for easy, resilient and secure K8s cluster deployment.

https://ubuntu.com/engage/secure-kubernetes-at-the-edge

Expanded Security Maintenance for Applications is not enabled.

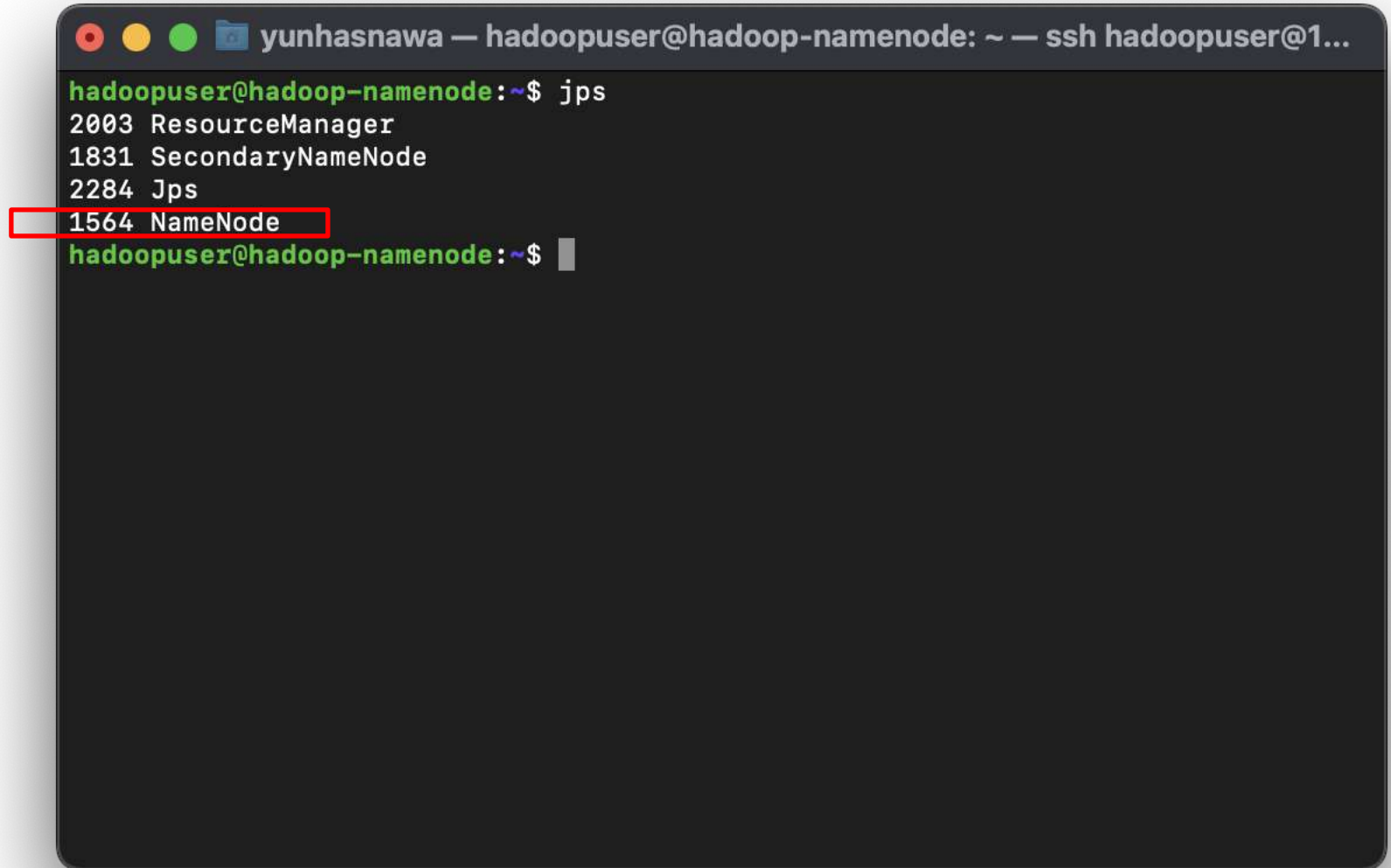
1 update can be applied immediately.
To see these additional updates run: apt list --upgradable

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

Last login: Sun Feb 23 02:11:37 2025 from 172.29.128.43
hadoopuser@hadoop-namenode:~$
```

### 3. Mengakses *Cluster* Hadoop Cek Status Name Node

- Jika Anda berhasil terhubung, maka di terminal anda, *prompt*-nya akan berganti menjadi:
  - **hadoopuser@hadoop-namenode**
- Ketikkan perintah “jps”.
- Akan ditampilkan status mesin saat ini yang bertugas sebagai name node.

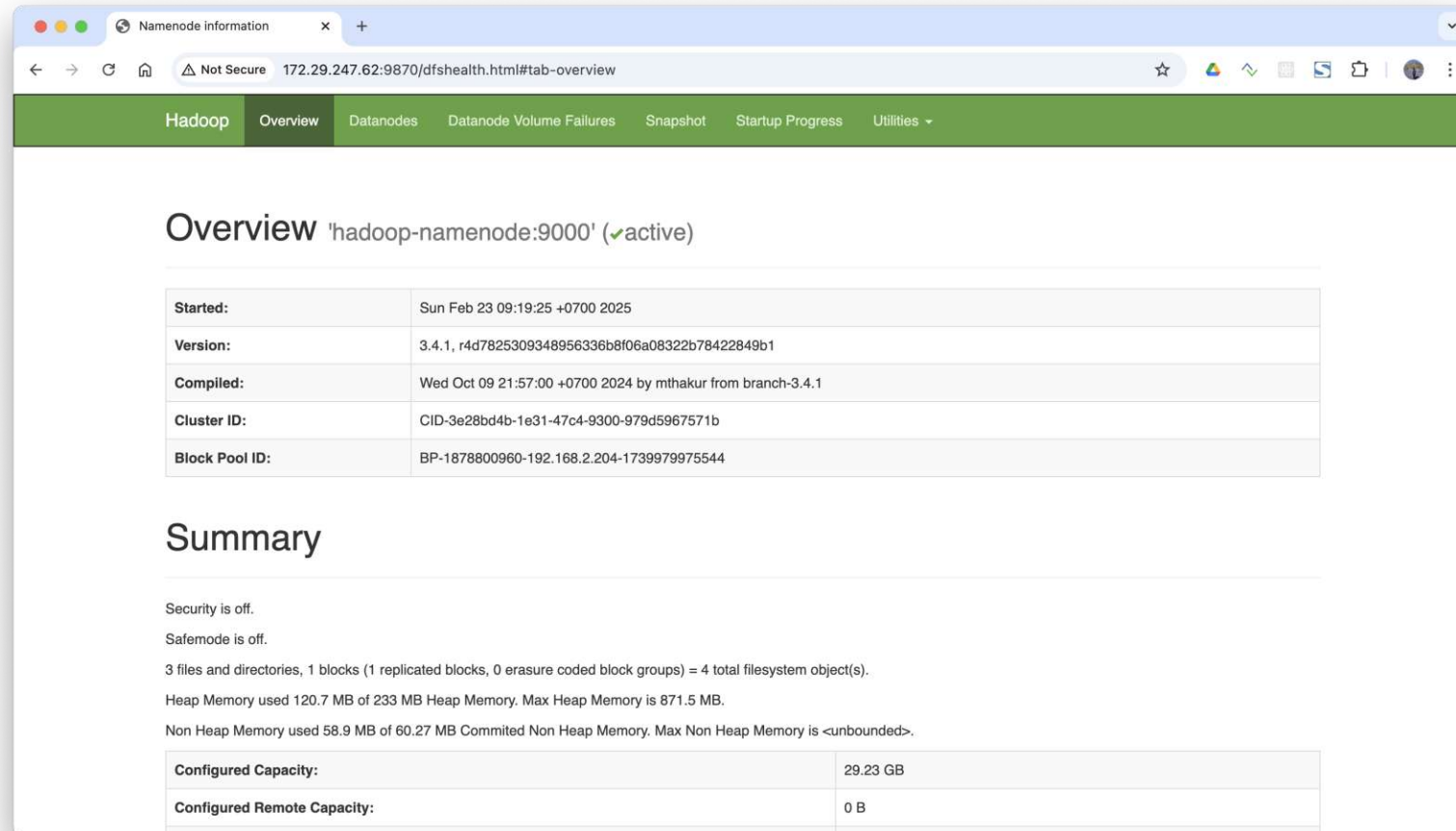


```
yunhasnawa — hadoopuser@hadoop-namenode: ~ — ssh hadoopuser@1...  
hadoopuser@hadoop-namenode:~$ jps  
2003 ResourceManager  
1831 SecondaryNameNode  
2284 Jps  
1564 NameNode  
hadoopuser@hadoop-namenode:~$
```

### 3. Mengakses *Cluster* Hadoop

## Melihat Status Data Nodes

- Anda dapat melihat status data nodes saat ini melalui GUI Web yang dapat diakses melalui alamat:
  - [http:// <IP\\_NameNode>:9870/dfshealth.html](http://<IP_NameNode>:9870/dfshealth.html)



The screenshot shows a web browser window with the title "NameNode Information". The address bar displays "172.29.247.62:9870/dfshealth.html#tab-overview". 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 the title "Overview 'hadoop-namenode:9000' (✓active)".

Started:	Sun Feb 23 09:19:25 +0700 2025
Version:	3.4.1, r4d7825309348956336b8f06a08322b78422849b1
Compiled:	Wed Oct 09 21:57:00 +0700 2024 by mthakur from branch-3.4.1
Cluster ID:	CID-3e28bd4b-1e31-47c4-9300-979d5967571b
Block Pool ID:	BP-1878800960-192.168.2.204-1739979975544

### Summary

Security is off.  
Safemode is off.

3 files and directories, 1 blocks (1 replicated blocks, 0 erasure coded block groups) = 4 total filesystem object(s).

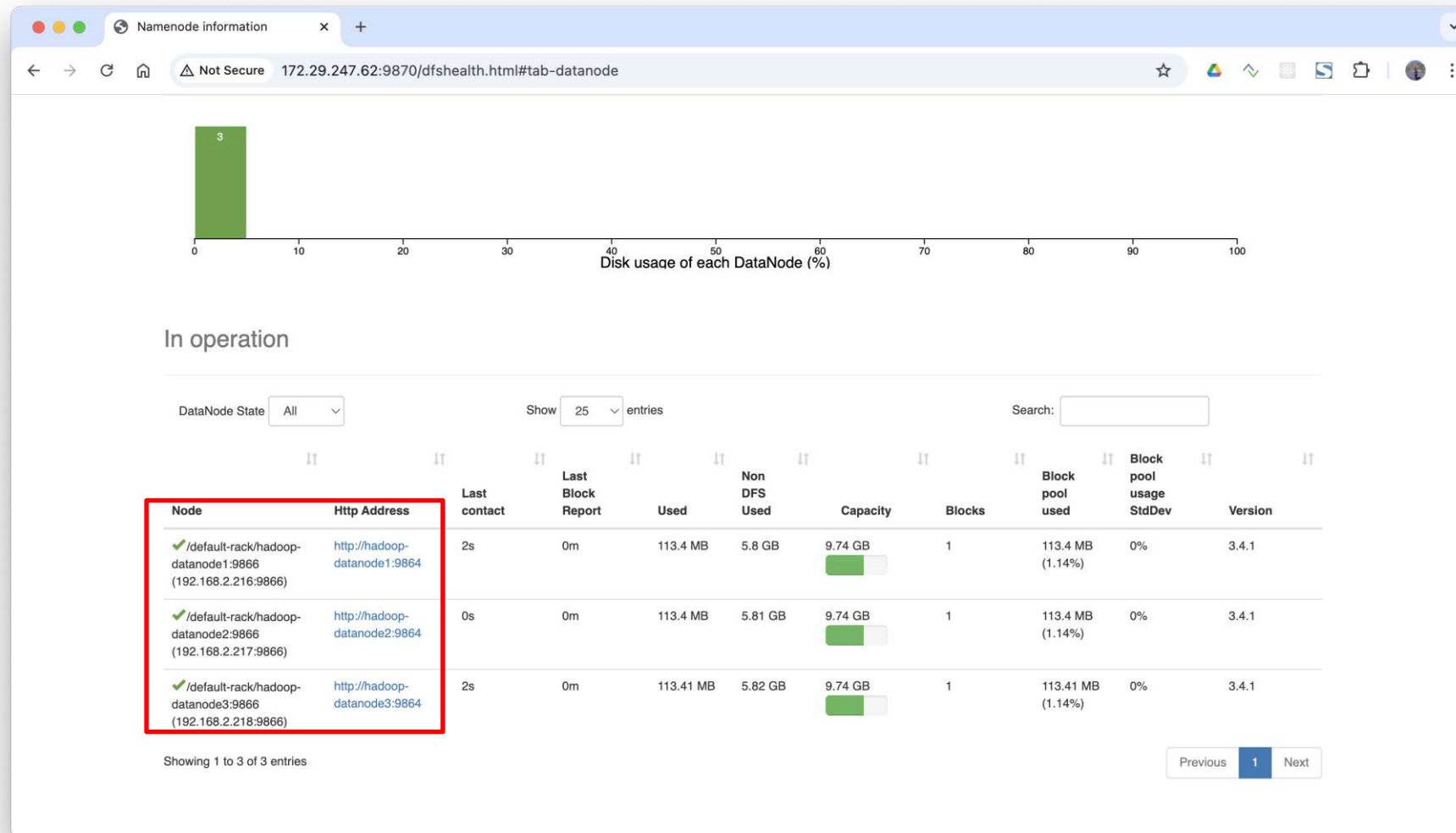
Heap Memory used 120.7 MB of 233 MB Heap Memory. Max Heap Memory is 871.5 MB.

Non Heap Memory used 58.9 MB of 60.27 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	29.23 GB
Configured Remote Capacity:	0 B

### 3. Mengakses *Cluster* Hadoop Melihat Status Data Nodes

- Klik menu **Datanodes** untuk melihat informasi data node yang ada pada cluster.

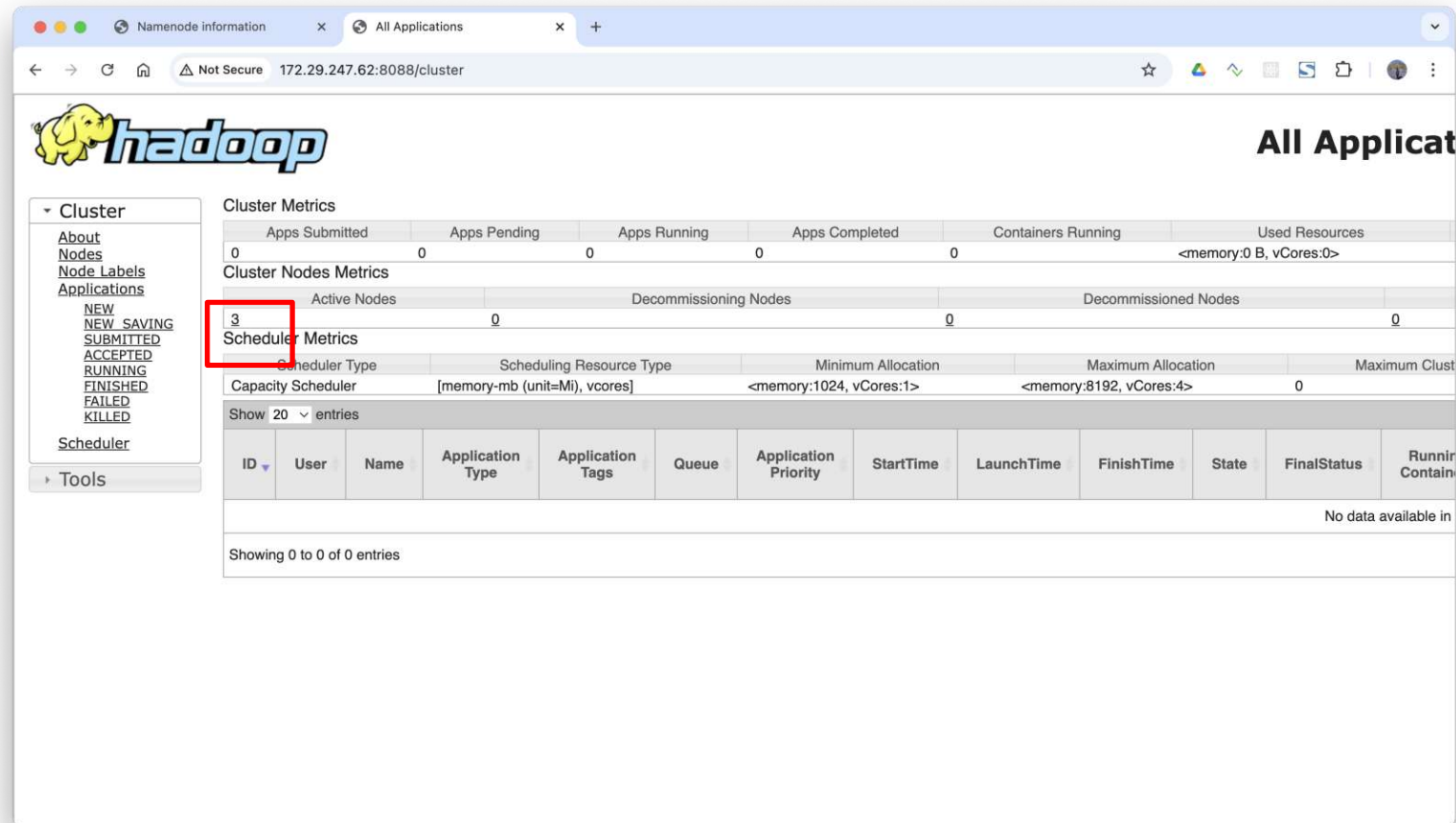




### 3. Mengakses *Cluster* Hadoop

## Melihat Status Pemrosesan MapReduce

- Situs yang sebelumnya terkait dengan penyimpanan data (HDFS) untuk pemrosesan, ada GUI Web yang lain. Anda dapat mengaksesnya melalui browser Anda dengan mengetikkan alamat:
  - [http://<IP\\_NameNode>:8088/cluster](http://<IP_NameNode>:8088/cluster)



The screenshot shows the Hadoop Web UI for 'All Applications'. The page includes a sidebar with navigation links and a main content area with several tables of metrics.

**Cluster Metrics**

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources
0	0	0	0	0	<memory:0 B, vCores:0>

**Cluster Nodes Metrics**

Active Nodes	Decommissioning Nodes	Decommissioned Nodes
3	0	0

**Scheduler Metrics**

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Clust
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>	0

**Applications Table**

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers
No data available in												

Showing 0 to 0 of 0 entries

# *Topik-4: Sedikit Perintah Dasar*

---

---

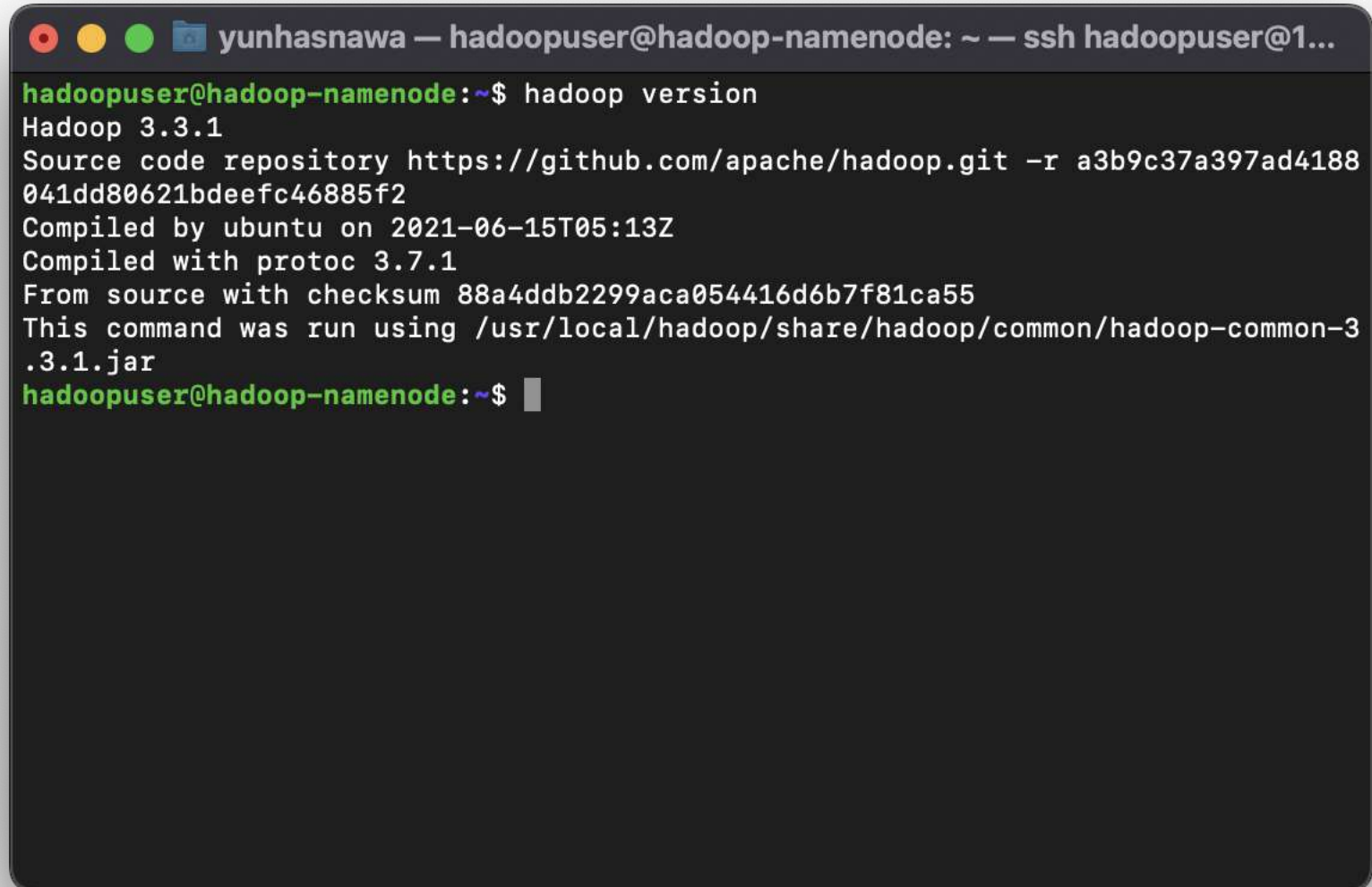
## 4. Sedikit Perintah Dasar

- Setelah Anda berhasil mengakses Hadoop dan melihat-lihat GUI Web HDFS dan MapReduce, berikutnya waktunya mencoba sedikit perintah dasar pada Hadoop.
- Perintah-perintah tersebut adalah:
  - Melihat versi Hadoop
  - Membuat direktori
  - Melihat direktori



### 3. Sedikit Perintah Dasar Melihat Versi Hadoop

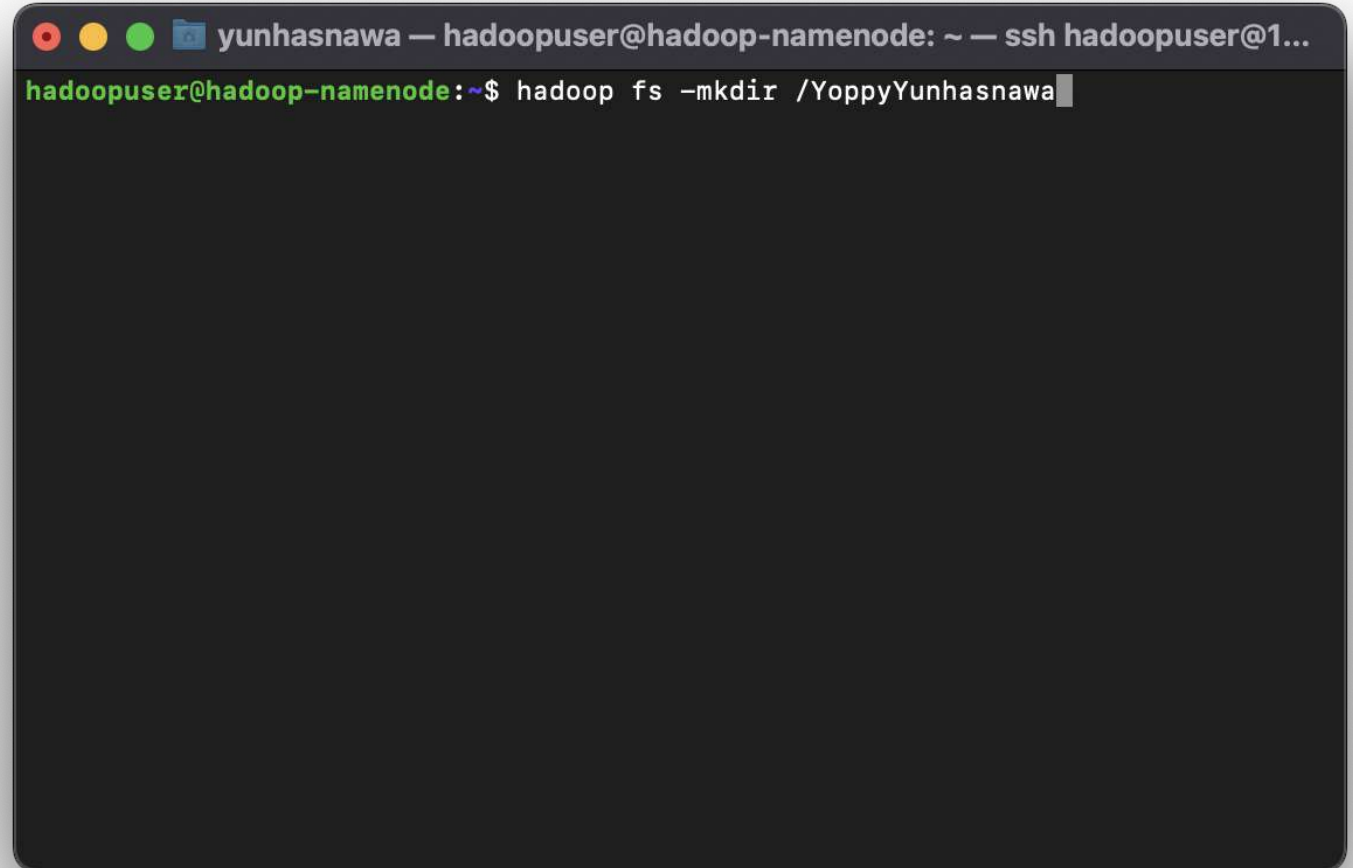
- Ketikkan:
  - `hadoop version`



```
yunhasnawa — hadoopuser@hadoop-namenode: ~ — ssh hadoopuser@1...  
hadoopuser@hadoop-namenode:~$ hadoop version  
Hadoop 3.3.1  
Source code repository https://github.com/apache/hadoop.git -r a3b9c37a397ad4188  
041dd80621bdeefc46885f2  
Compiled by ubuntu on 2021-06-15T05:13Z  
Compiled with protoc 3.7.1  
From source with checksum 88a4ddb2299aca054416d6b7f81ca55  
This command was run using /usr/local/hadoop/share/hadoop/common/hadoop-common-3  
.3.1.jar  
hadoopuser@hadoop-namenode:~$
```

### 3. Sedikit Perintah Dasar Membuat Folder

- Ketikkan:
  - `hadoop fs -mkdir /<NamaFolder>`
- Nama folder jangan ada spasi.
  - Disarankan menggunakan CamelCase.
- “/” adalah direktori teratas alias *root directory*.
  - Semua folder yang dibuat harus **selalu berada** di bawahnya.

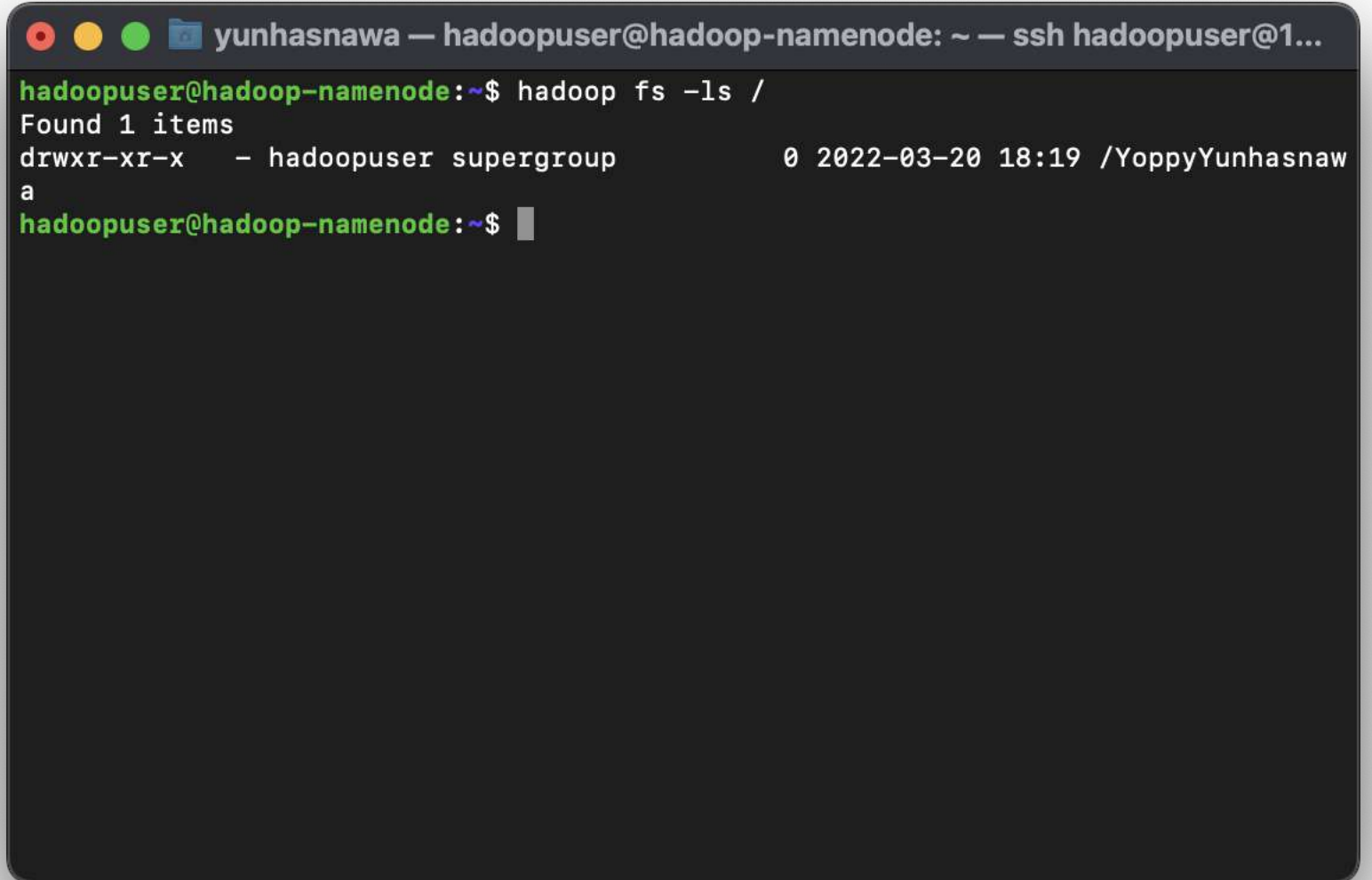


```
yunhasnawa — hadoopuser@hadoop-namenode: ~ — ssh hadoopuser@1...  
hadoopuser@hadoop-namenode:~$ hadoop fs -mkdir /YoppyYunhasnawa
```

### 3. Sedikit Perintah Dasar

## Melihat Folder-folder yang Ada

- Ketikkan:
  - `hadoop fs -ls /`



```
yunhasnawa — hadoopuser@hadoop-namenode: ~ — ssh hadoopuser@1...  
hadoopuser@hadoop-namenode:~$ hadoop fs -ls /  
Found 1 items  
drwxr-xr-x  - hadoopuser supergroup          0 2022-03-20 18:19 /YoppyYunhasnaw  
a  
hadoopuser@hadoop-namenode:~$
```

# Pertanyaan?



*Terima Kasih*

# Latihan

- Cobalah untuk:
  - Terhubung ke cluster Hadoop.
  - Jalankan 3 perintah dasar tadi.
  - Catatan: Untuk folder yang dibuat, beri nama dengan **NoAbsen\_NamaAnda**.
- Buatlah laporan yang berisi:
  - Screenshot dan penjelasan Langkah-Langkah Anda.
- Kumpulkan di Google Classroom.