

LAPORAN PRAKTIKUM 2 ANALISIS BIG DATA

MEMASUKAN DATA KE HDFS

Dosen Pengampu: Sevi Nurafni ST., M.Si., M.Sc.



IKOPIN University

Disusun Oleh Ryan Fadhilah Faizal Hakim

Program Studi Sains Data

Fakultas Sains dan Teknologi

©2024 Ryan Hakim All Rights Reserved

Pendahuluan

Pada praktikum 1 analisis big data yaitu instalasi Hadoop ternyata masih ada beberapa kesalahan dalam instalasi sehingga masih terdapat *error* ketika dilakukan proses lebih lanjut. Untuk itu, pada praktikum ke-2 ini akan dilakukan perbaikan terhadap *error* tersebut dimulai dari pengecekan ulang *environment variable* sampai dengan penginstalan bagian yang terkena *bug* atau *error*.

Setelah dilakukan *setup environment variable* yang baru, akan dilanjutkan dengan praktikum cara memasukan data ke HDFS. Untuk mengingat kembali bahwa Hadoop Distributed File System (HDFS) merupakan proyek open source yang dikembangkan oleh Apache Software Foundation yang juga merupakan subproyek dari Apache Hadoop. HDFS dirancang untuk menyimpan data dalam skala besar di seluruh kluster komputer. Sistem ini membagi data menjadi blok-blok yang kemudian didistribusikan dan disimpan di berbagai node dalam kluster.

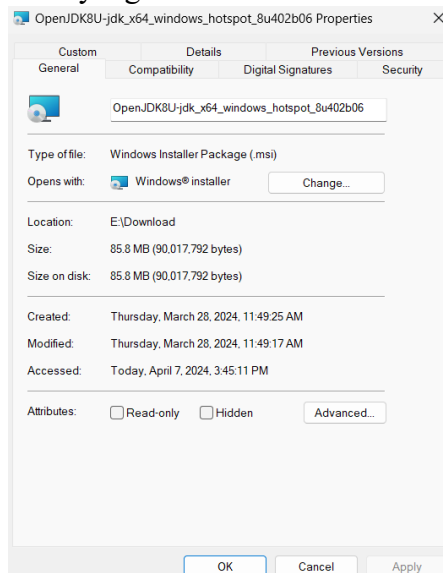
HDFS memiliki dua komponen utama, yaitu **namenode** dan **datanode**. Namenode bertindak sebagai master dan bertanggung jawab untuk menyimpan semua metadata, sedangkan datanode bertugas untuk menyimpan blok data. Jika salah satu datanode rusak, sistem tetap dapat berjalan karena setiap blok data direplikasi di beberapa mesin lainnya.

Pada praktikum kali ini akan disampaikan cara memasukan data ke HDFS.

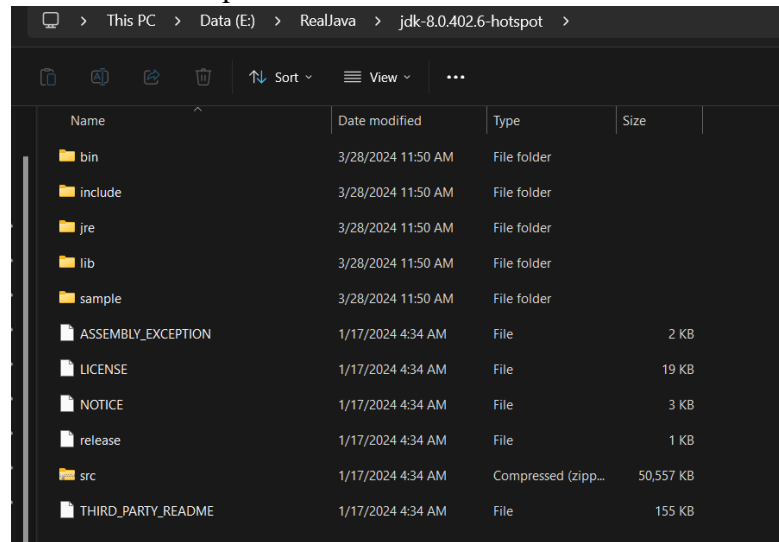
Metode

Pada praktikum sebelumnya masih terdapat kesalahan yang saya lakukan ketika melakukan instalasi Hadoop salah satunya adalah *environment variable* yaitu JAVA_HOME masih memiliki masalah sehingga Hadoop tidak dapat membaca lokasi JAVA. Berikut adalah langkah yang saya lakukan untuk memperbaikinya:

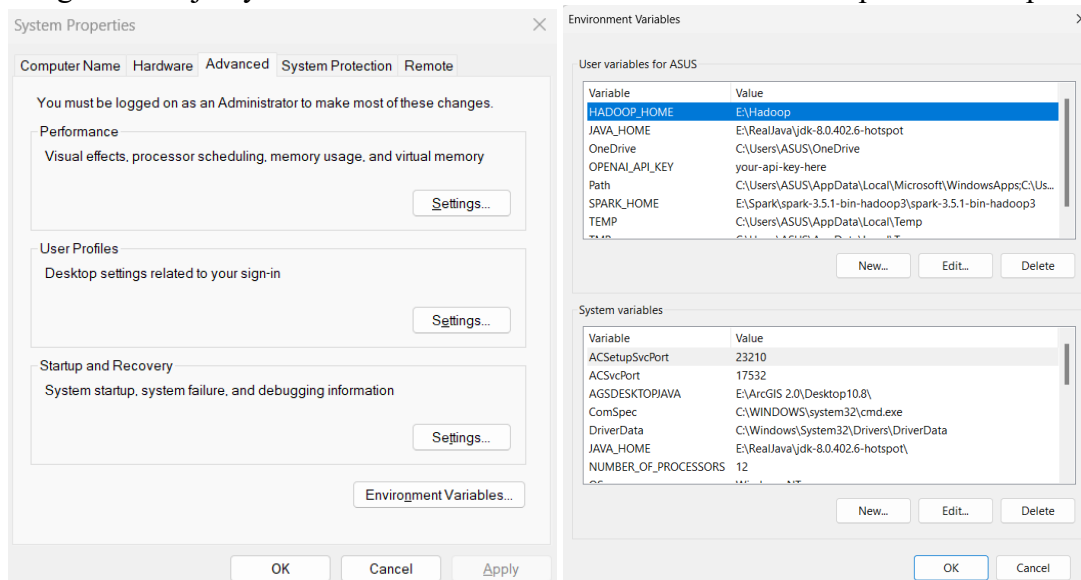
1. Menginstal kembali JDK versi 8 yang disediakan oleh laboran.



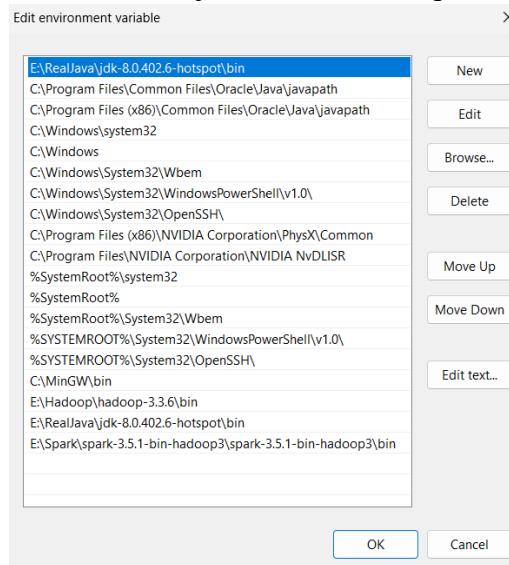
2. Setelah dilakukan penginstalan ulang, JDK versi 8 yang saya instal tersimpan di direktori E:\RealJava\jdk-8.0.402.6-hotspot



3. Langkah selanjutnya adalah melakukan edit *environment variable* pada control panel



4. Pada bagian *User variable for ...*, bagian JAVA_HOME dilakukan edit *value* dengan direktori E:\RealJava\jdk-8.0.402.6-hotspot serta melakukan penggantian path dari *System Variables* dengan direktori E:\RealJava\jdk-8.0.402.6-hotspot\bin.



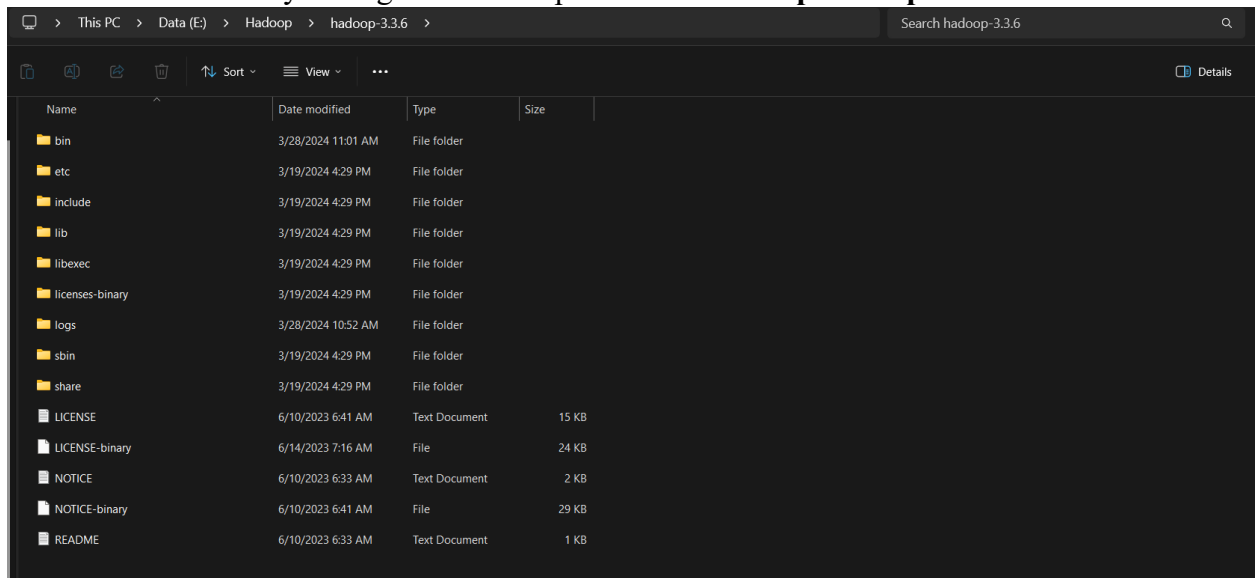
5. Setelah itu lakukan pengecekan dengan masuk ke command prompt dan mengetikan “java -version”.

```
Microsoft Windows [Version 10.0.22631.3374]
(c) Microsoft Corporation. All rights reserved.

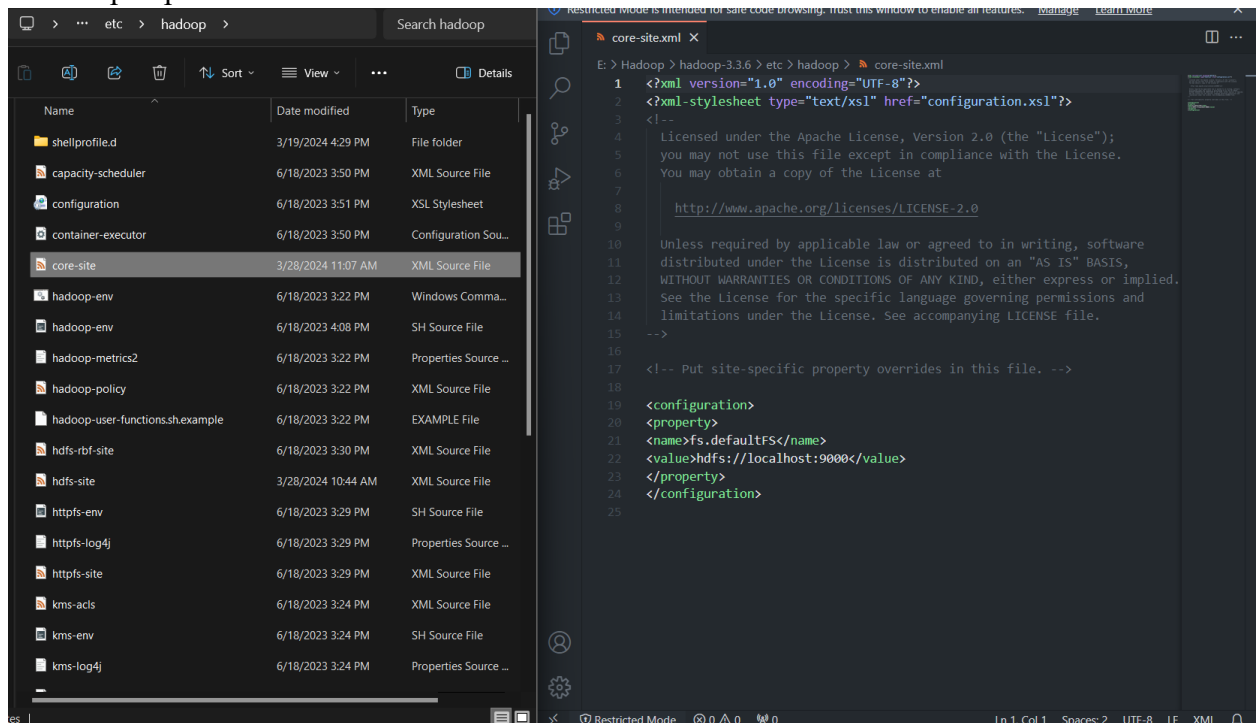
C:\Users\ASUS>java -version
openjdk version "1.8.0_402"
OpenJDK Runtime Environment (Temurin)(build 1.8.0_402-b06)
OpenJDK 64-Bit Server VM (Temurin)(build 25.402-b06, mixed mode)
```

6. Langkah selanjutnya yaitu pergi ke folder dimana kita menginstal Hadoop.

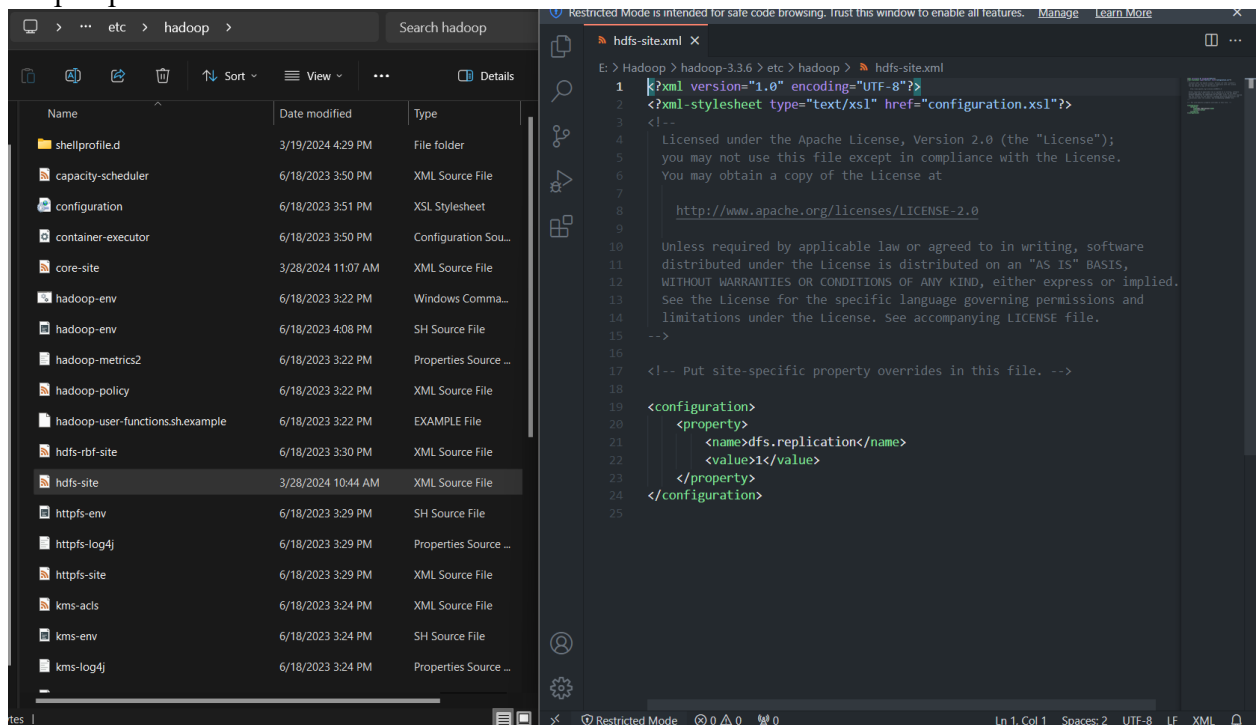
Note: Disini lokasi saya menginstal Hadoop ada di **E:\Hadoop\hadoop-3.3.6**



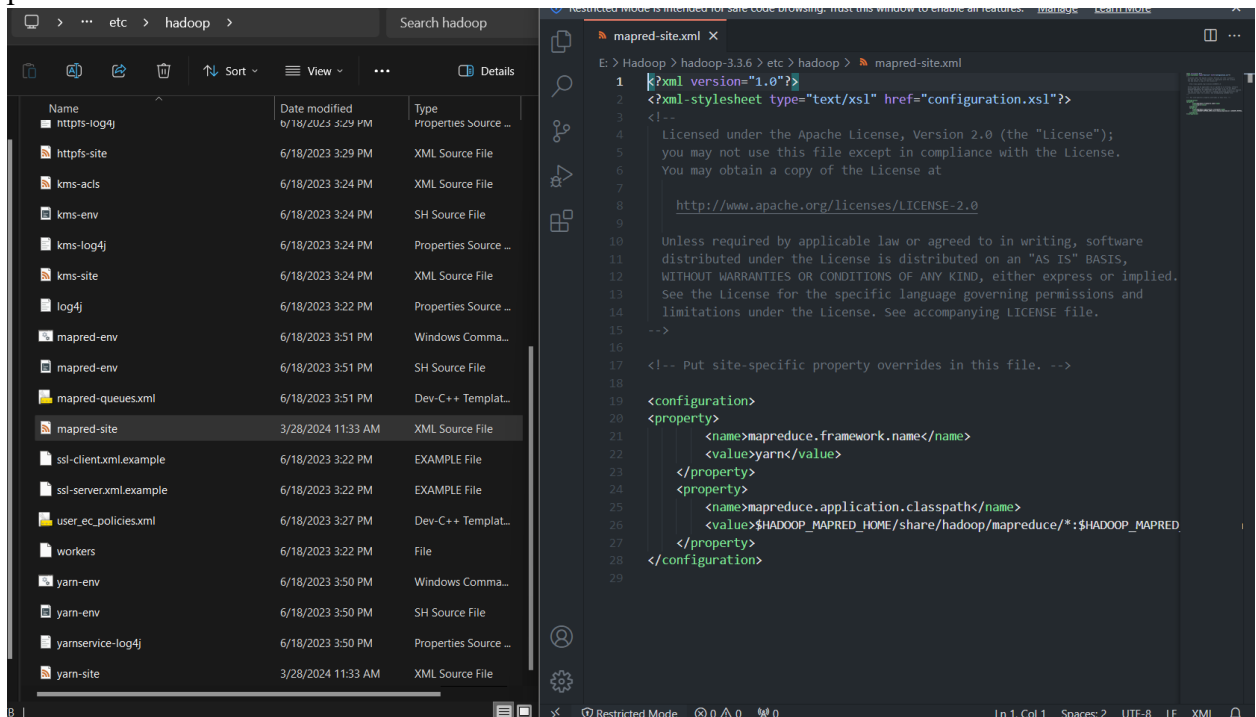
7. Pada bagian folder Hadoop tersebut, buka folder **etc** lalu buka file bernama **core-site.xml** dengan code editor dan tambahkan bagian **configuration** dengan code seperti di bawah lalu simpan perubahan.



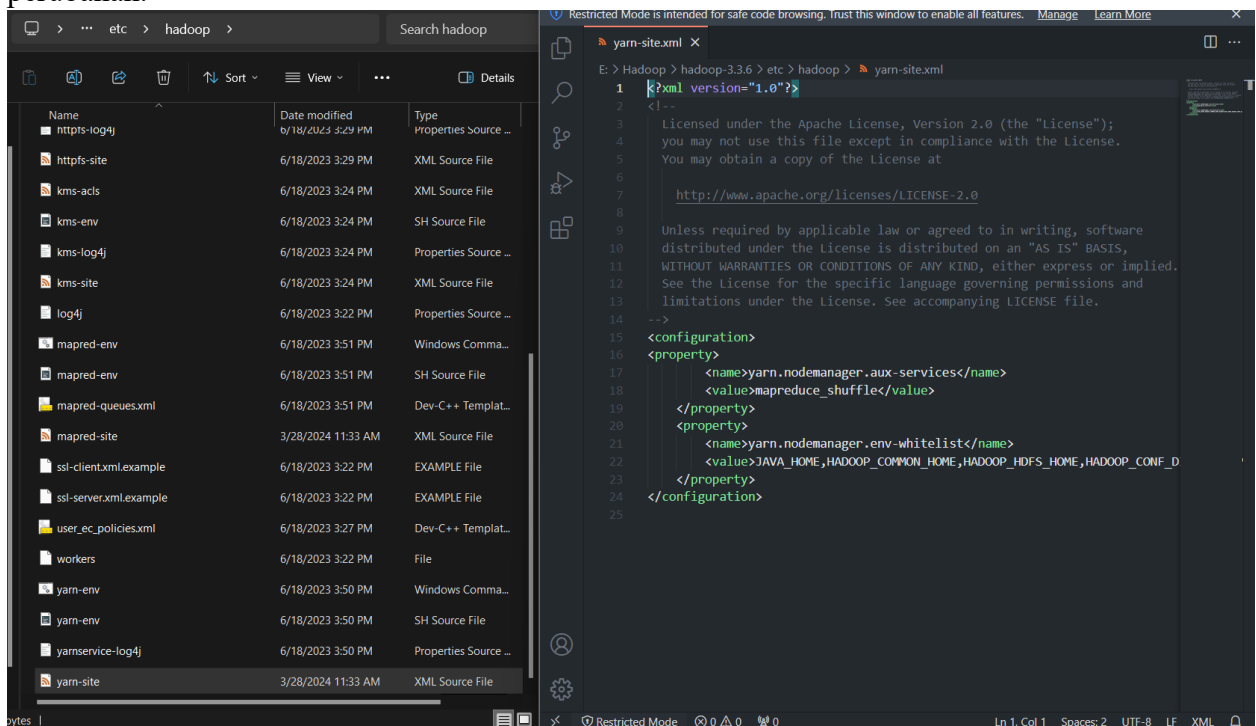
8. Selanjutnya buka kembali folder **etc** lalu buka file bernama **hdfs-site.xml** dengan code editor dan tambahkan pada bagian **configuration** dengan code seperti di bawah lalu simpan perubahan.



9. Setelah itu kembali lagi ke folder **etc** dan buka file bernama **mapred-site.xml** dengan code editor dan tambahkan pada bagian **configuration** dengan code seperti di bawah lalu simpan perubahan.



10. Setelah itu kembali lagi ke folder **etc** dan buka file bernama **yarn-site.xml** dengan code editor dan tambahkan pada bagian **configuration** dengan code seperti di bawah lalu simpan perubahan.



11. Setelah berhasil, langkah selanjutnya yaitu membuka “Terminal” atau Window Powershell

```
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\ASUS> |
```

12. Dikarenakan saya menginstal hadoop di direktori E:, maka perlu mengganti direktori dengan cara mengetik **cd E:/** pada powershell

```
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\ASUS> cd E:/
PS E:\> |
```

13. Setelah itu masuk ke file direktori Hadoop yang diinstal dengan mengetik **cd .\Hadoop\hadoop-3.3.6** lalu tekan enter

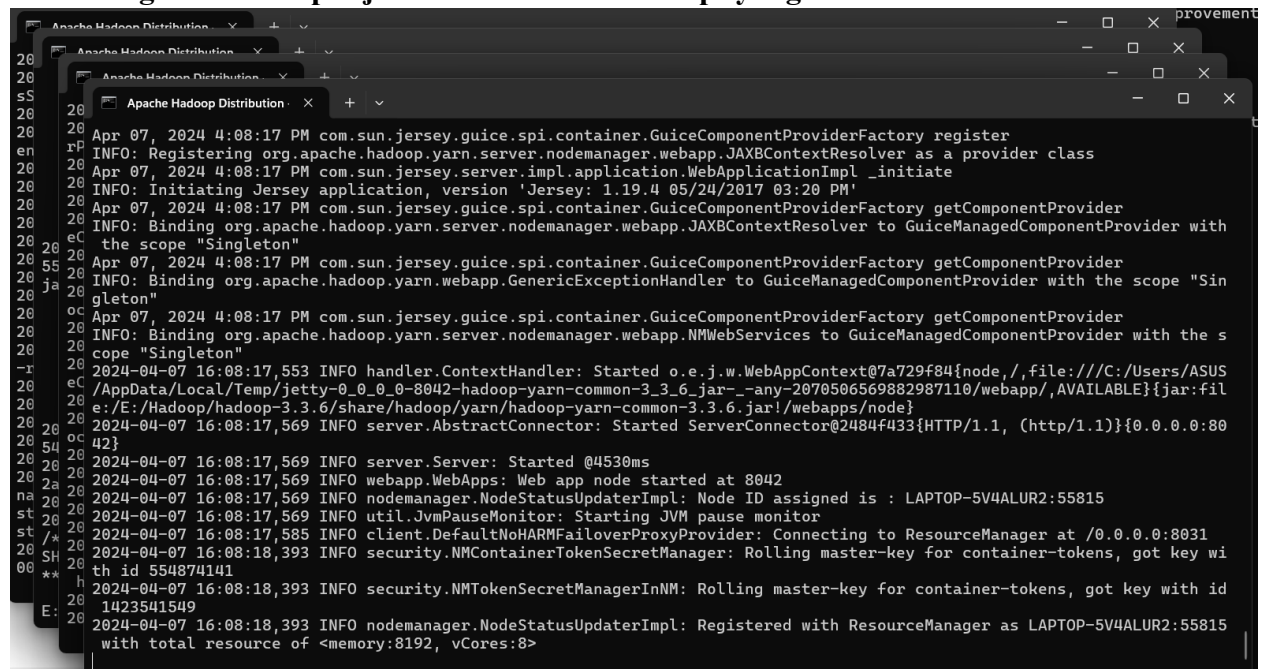
```
PS C:\Users\ASUS> cd E:/
PS E:\> cd .\Hadoop\hadoop-3.3.6\
PS E:\Hadoop\hadoop-3.3.6> |
```

14. Setelah masuk direktori hadoop, ketik **hadoop namenode -format** untuk melakukan format namenode pada sistem Hadoop sehingga muncul tampilan seperti di bawah ini.

```
PS E:\Hadoop\hadoop-3.3.6> hadoop namenode -format
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
2024-04-07 16:04:34,112 INFO namenode.NameNode: STARTUP_MSG:
/*****
STARTUP_MSG: Starting NameNode
STARTUP_MSG:  host = LAPTOP-5V4ALUR2/192.168.43.24
STARTUP_MSG:  args = [-format]
STARTUP_MSG:  version = 3.3.6
STARTUP_MSG:  classpath = E:\Hadoop\hadoop-3.3.6\etc\hadoop;E:\Hadoop\hadoop-3.3.6\share\hadoop\common;E:\Hadoop\hadoop-3.3.6\share\hadoop\common\lib\animal-sniffer-annotations-1.17.jar;E:\Hadoop\hadoop-3.3.6\share\hadoop\common\lib\audience-annotations-0.5.0.jar;E:\Hadoop\hadoop-3.3.6\share\hadoop\common\lib\avro-1.7.7.jar;E:\Hadoop\hadoop-3.3.6\share\hadoop\common\lib\checker-qual-2.5.2.jar;E:\Hadoop\hadoop-3.3.6\share\hadoop\common\lib\commons-beanutils-1.9.4.jar;E:\Hadoop\hadoop-3.3.6\share\hadoop\common\lib\commons-cli-1
```

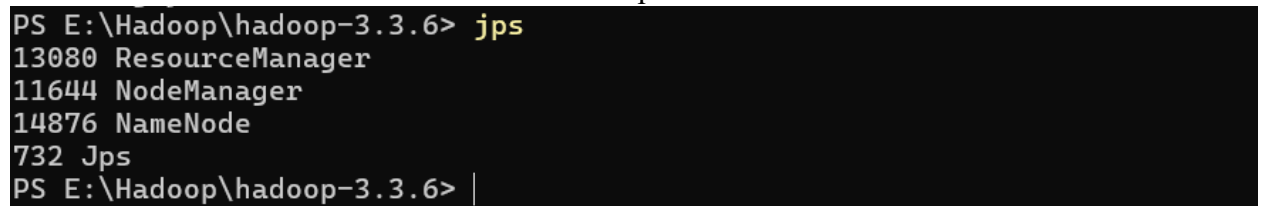
15. Setelah melakukan format namenode pada Hadoop, masukan kata kunci **sbin/start-all** sehingga muncul empat jendela command prompt baru seperti gambar berikut.

Note: **Jangan close empat jendela Command Prompt yang muncul**

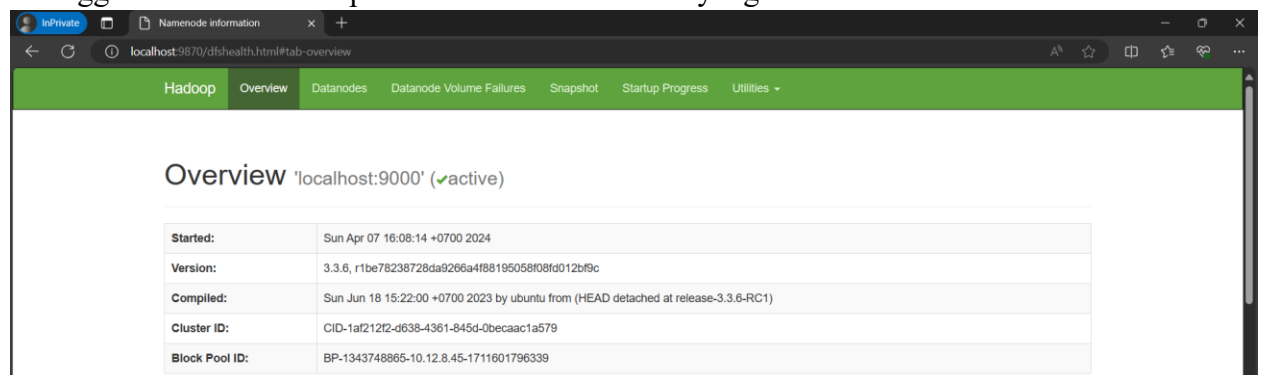


16. Setelah muncul, kembali ke jendela powershell tadi lalu ketik **jps** untuk memastikan bahwa HDFS sudah bisa dipakai.

Note: Jika sudah berhasil maka akan muncul tampilan berikut

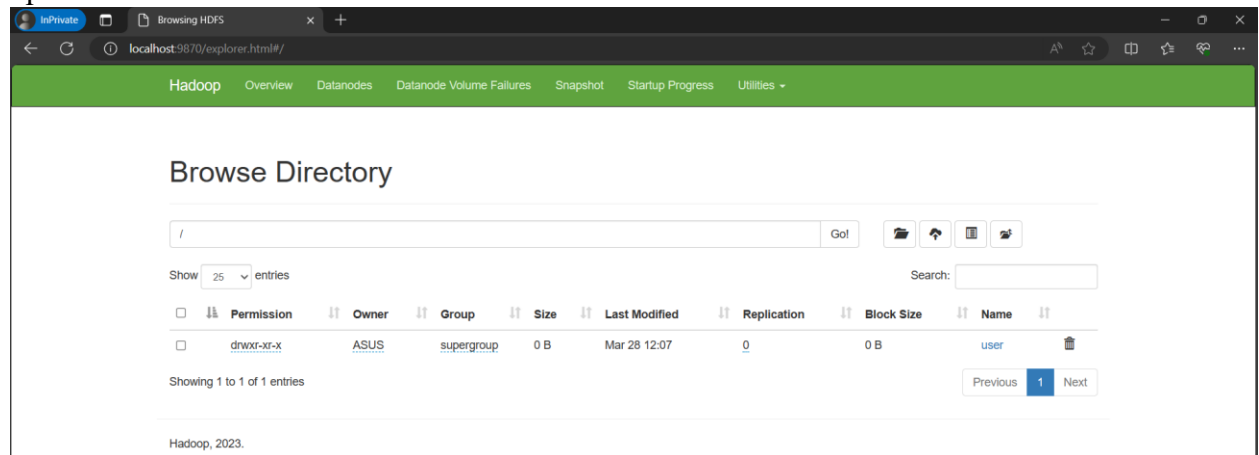


17. Setelah itu buka browser lalu melakukan search dengan kata kunci **localhost:9870** sehingga akan muncul tampilan local host dari HDFS yang dibuat dan akan berstatus aktif.



18. Pada local host tersebut, pergi ke bagian navigasi **Utilities** dan pilih **Browse the file system** sehingga muncul tampilan berikut

Note: disana telah terdapat file yang telah diupload karena saya telah melakukan proses upload tersebut. Jadi diasumsikan bahwa file tersebut belum dibuat.



19. Langkah selanjutnya yaitu pergi ke Powershell kembali lalu buat direktori file HDFS dengan mengetikkan **hadoop fs -mkdir /user**

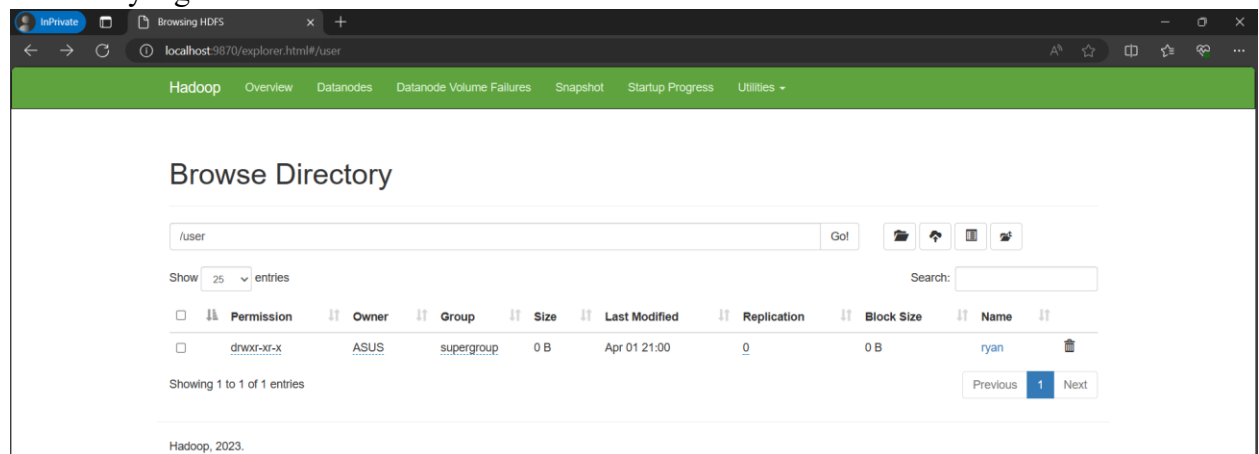
```
PS E:\Hadoop\hadoop-3.3.6> hadoop fs -mkdir /user
```

20. Setelah itu ketikkan **hadoop fs -mkdir /user/nama**

Note: Pada bagian “nama”, itu dapat diisi sesuai pengguna (User define)

```
PS E:\Hadoop\hadoop-3.3.6> hadoop fs -mkdir /user/ryan
```

21. Setelah melakukan langkah tersebut, kembali ke browse dimana kita mengakses **Browse the file system** dan akan muncul direktori baru dalam browse tersebut dengan nama direktori yang telah ditentukan user



22. Setelah semua langkah selesai dilakukan, kembali ke powershell dan ketikkan **sbin/stop-all** untuk mengakhiri semua aktivitas hadoop.

```
PS E:\Hadoop\hadoop-3.3.6> sbin/stop-all
This script is Deprecated. Instead use stop-dfs.cmd and stop-yarn.cmd

INFO: No tasks running with the specified criteria.

INFO: No tasks running with the specified criteria.
stopping yarn daemons

INFO: No tasks running with the specified criteria.

INFO: No tasks running with the specified criteria.

INFO: No tasks running with the specified criteria.
```

Hasil

Berdasarkan praktikum yang telah saya coba lakukan seperti yang telah saya sebutkan langkah-langkahnya, bahwa proses untuk menyimpan data ke dalam HDFS berhasil berjalan dengan baik sehingga HDFS siap digunakan untuk menyimpan data.

Diskusi

Poin-poin yang perlu didiskusikan pada praktikum ini diantaranya:

- Dengan memiliki akses terhadap lingkungan Hadoop, apakah masih terdapat batasan untuk menyimpan data?
- Apabila Hadoop ini dapat digunakan secara open source, mengapa masih banyak user menyimpan datanya secara local dengan pertimbangan bahwa data tersebut berukuran besar?
- Bagaimana cara menghubungkan file yang berada di HDFS tersebut dengan software lainnya?
- Apakah harus menjaga server agar HDFS dapat berjalan terus menerus?

Kesimpulan

Dalam praktikum ini, Saya melakukan perbaikan dan penyesuaian pada instalasi Hadoop yang sebelumnya mengalami beberapa kesalahan. Saya melakukan pengecekan ulang environment variable dan melakukan penginstalan ulang pada bagian yang mengalami bug atau error. Setelah melakukan setup environment variable yang baru, Saya melanjutkan praktikum dengan memasukkan data ke Hadoop Distributed File System (HDFS), sebuah sistem penyimpanan data skala besar yang membagi data menjadi blok-blok dan mendistribusikannya di berbagai node dalam kluster.

Saya berhasil melakukan format namenode pada Hadoop dan memastikan bahwa HDFS sudah bisa dipakai. Saya juga berhasil membuat direktori baru dalam HDFS dan menyimpan data ke dalamnya. Dengan demikian, HDFS siap digunakan untuk menyimpan data dalam skala besar.

Namun, masih ada beberapa pertanyaan yang perlu didiskusikan lebih lanjut, seperti batasan penyimpanan data pada Hadoop, alasan pengguna menyimpan data secara lokal meskipun Hadoop tersedia sebagai open source, cara menghubungkan file yang berada di HDFS dengan software lain, dan apakah perlu menjaga server agar HDFS dapat berjalan terus menerus. Kesimpulan ini memberikan gambaran umum tentang proses dan hasil praktikum, serta pertanyaan yang muncul selama proses tersebut.

Referensi

Databricks.com. (2024). Hadoop Distributed File System (HDFS). Diakses dari Situs web Databricks: <https://www.databricks.com/glossary/hadoop-distributed-file-system-hdfs>.