Cài đặt Hadoop trên Linux (Ubuntu)

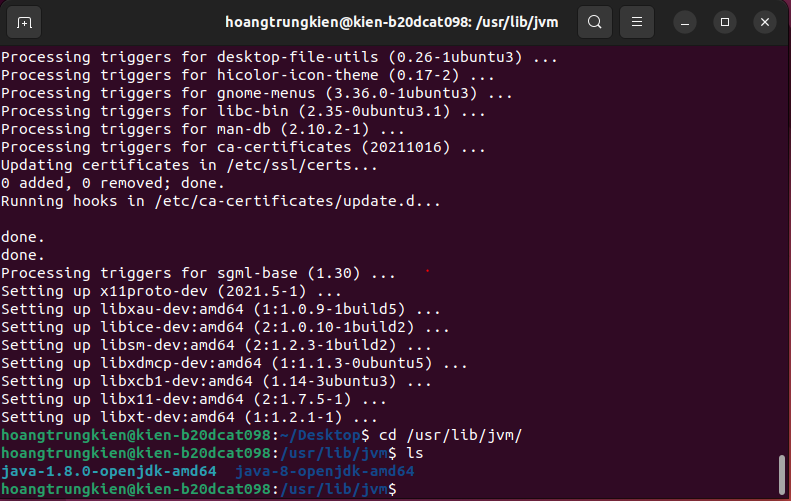
Hadoop mạnh mẽ và hữu dụng chỉ khi được cài đặt và khai thác nó trên nhiều node, tuy nhiên với người bắt đầu thì Hadoop Single node là sự khởi đầu tuyệt vời để làm quen với hadoop. Bài viết này mình sẽ hướng dẫn các bạn triển khai Hadoop trên 1 node ( Hadoop Single node).

Đầu tiên cài đặt Openjdk (bản 8, 11 hay 15 đều được) theo câu lệnh sau:



Cài đặt thành công

(/usr/lib/jvm/ thư mục của Openjdk11)

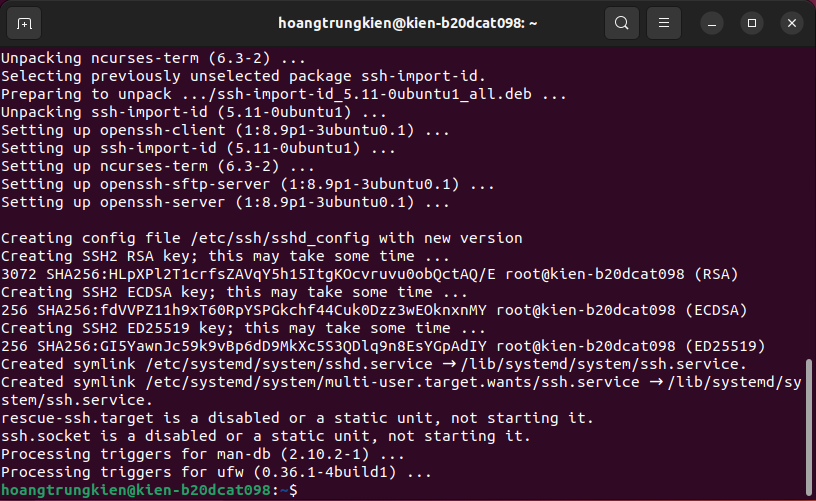


Cài đặt SSH client và SSH server theo câu lệnh:



File /etc/ssh/sshd\_config

Cài đặt thành công



Thiết lập User cho Hadoop:

Tạo cặp khóa SSH và xác định vị trí sẽ được lưu trữ:



Hệ thống sẽ tiến hành tạo và lưu cặp khóa SSH :



Sử dụng lệnh ***cat***để lưu ***public key*** vào ***authorized\_keys*** trong thư mục của SSH:

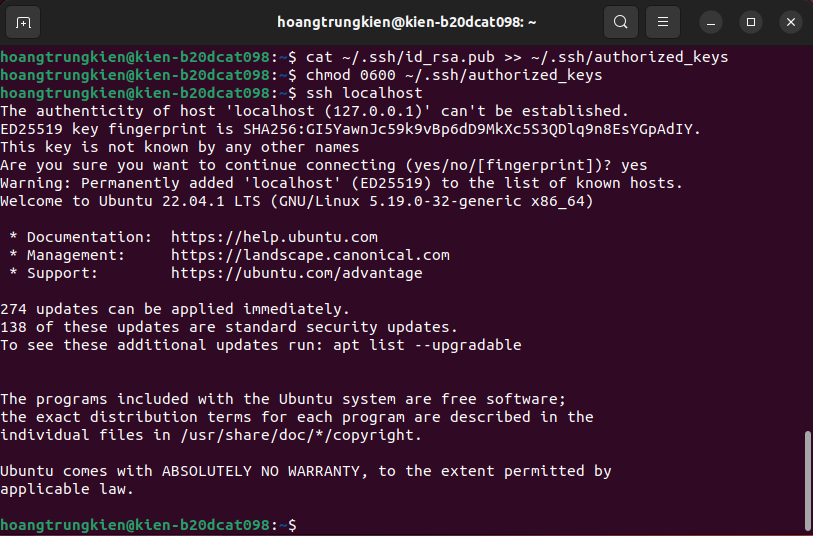


Phân quyền cho người dùng với lệnh ***chmod****:*



Xác minh mọi thứ được thiết lập chính xác bằng cách ssh đến localhost:

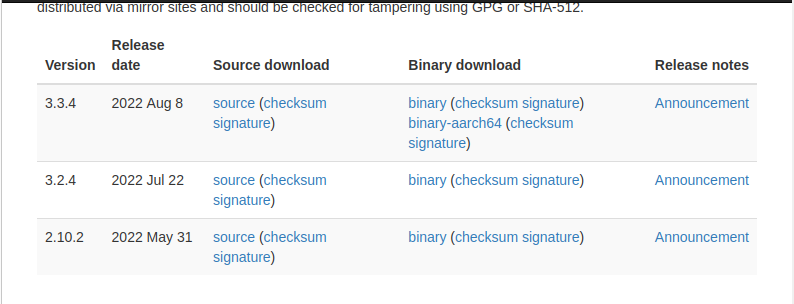




Download và cài đặt Hadoop trên Ubuntu

Tải về một phiên bản Hadoop trên trang phân phối chính thức của Hadoop tại: <https://hadoop.apache.org/releases.html>

Ấn vào phần ***binary*** trong ***Binary download***



Giải nén file vừa tải



Giải nén thành công



# **Cấu hình và triển khai Hadoop Single Node (Pseudo-Distributed Mode)**

Để cấu hình Hadoop cho chế độ phân phối giả chúng ta sẽ chỉnh sửa các tập file cấu hình của Hadoop trong đường dẫn ***etc/hadoop*** và trong file cấu hình môi trường gồm các file sau :

.bashrc

hadoop-env.sh

core-site.xml

hdfs-site.xml

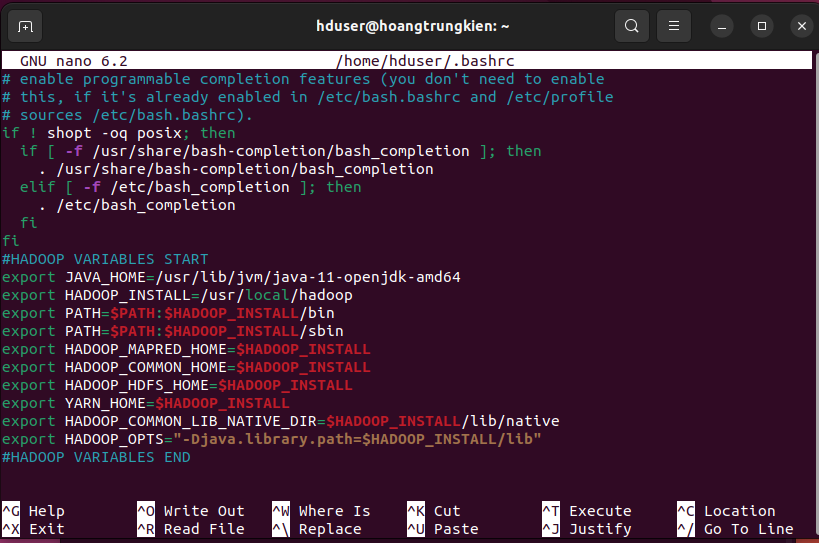
mapred-site.xml

**Cấu hình biến môi trường Hadoop ( file .bashrc)**

Mở file của .bashrc của bạn bằng trình soạn thảo nano :



Xác định biến mỗi trường ***Hadoop*** bằng cách thêm các biến sau vào cuối file ( nhớ chỉnh sửa đường dẫn ***$HADOOP\_HOME*** cho đúng với đường dẫn mà bạn đã đặt hadoop)



Áp dụng các thay đổi trên bằng cách thực hiện lệnh sau :



### **Chỉnh sửa file hadoop-env.sh**

Mở file hadoop-env.sh bằng trình soạn thảo nano :

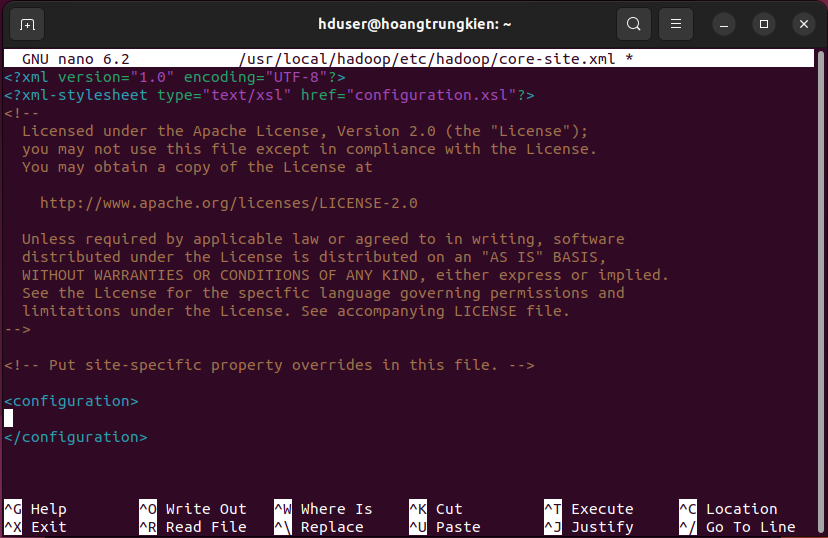


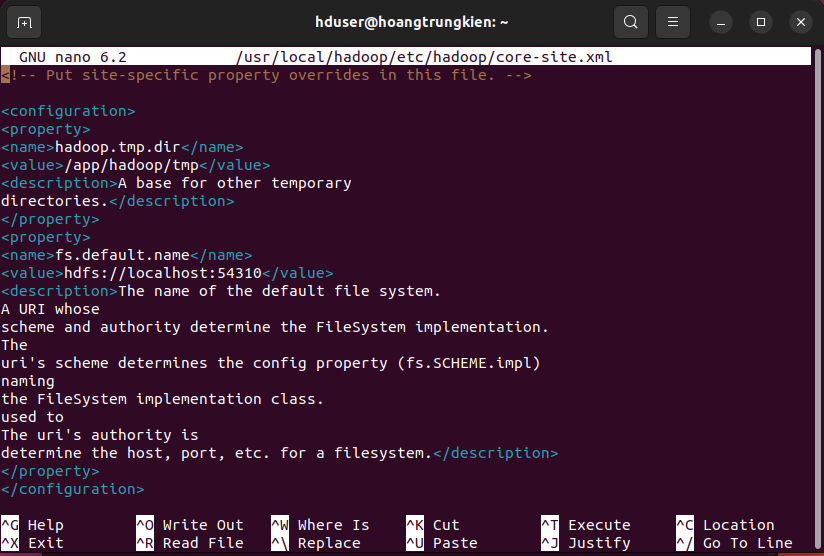


### **Chỉnh sửa file core-site.xml**

Mở file core-site.xml bằng trình soạn thảo nano :

sudo nano /usr/local/hadoop/etc/hadoop/core-site.xml

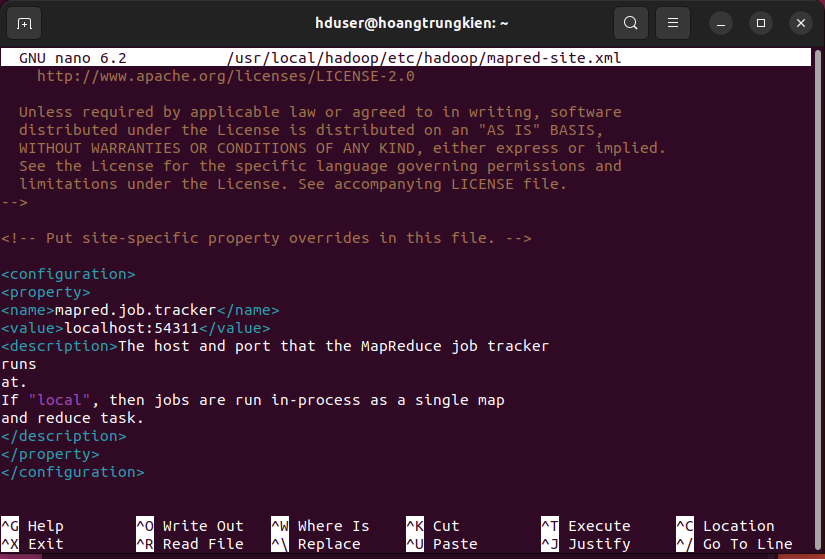




### **Chỉnh sửa file mapred-site.xml**

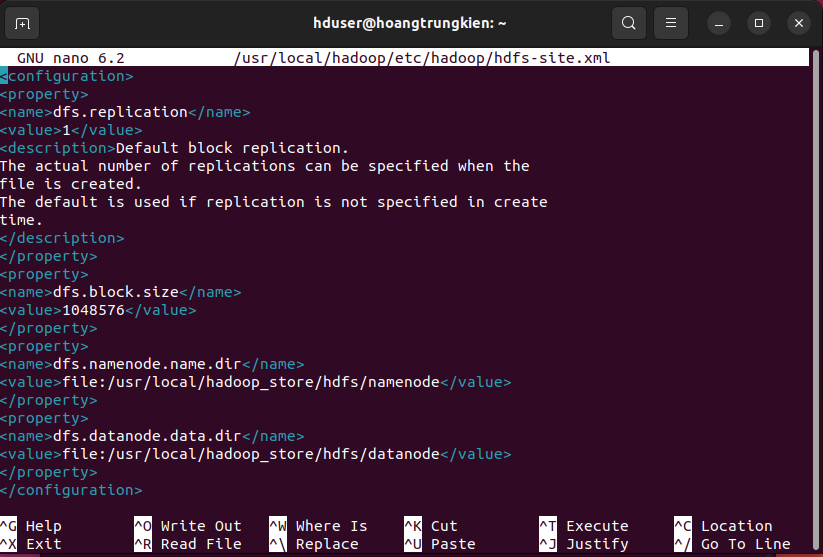
Mở file core-site.xml bằng trình soạn thảo nano :

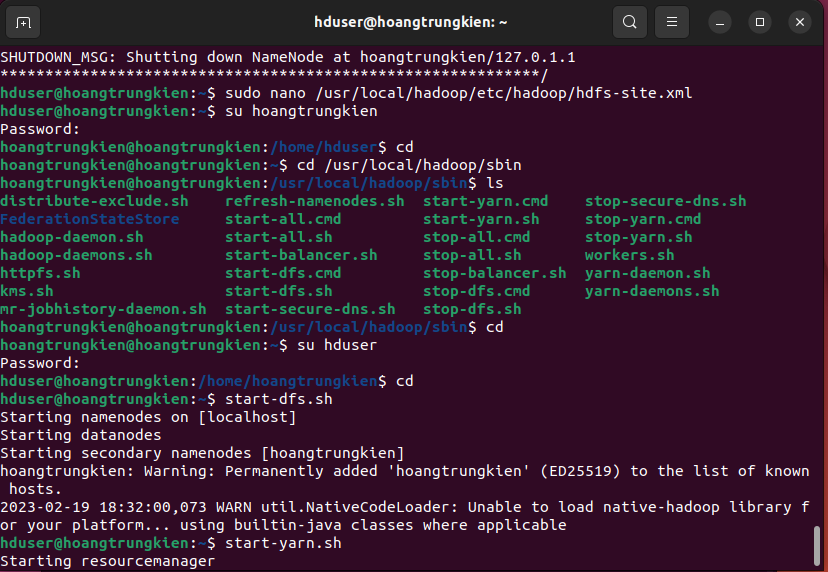
sudo nano /usr/local/hadoop/etc/hadoop/mapred-site.xml

**chỉnh sửa file hdfs-site.xml**

Mở file core-site.xml bằng trình soạn thảo nano :

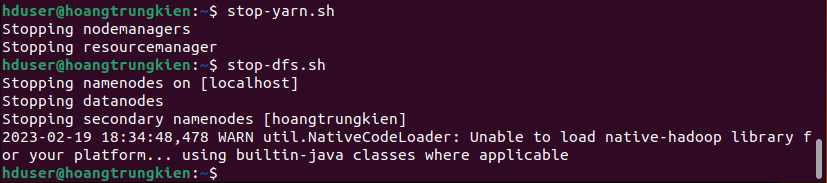
sudo nano /usr/local/hadoop/etc/hadoop/hdfs-site.xml





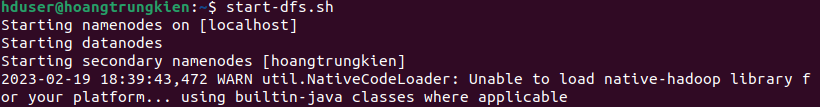


**YARN** (Yet-Another-**Resource**-Negotiator) **là** một framework hỗ trợ phát triển ứng dụng phân tán **YARN** cung cấp daemons và APIs cần thiết cho việc phát triển ứng dụng phân tán, đồng thời xử lý và lập lịch sử dụng tài nguyên tính toán (CPU hay memory) cũng như giám sát quá trình thực thi các ứng dụng đó.



DFS **là tên viết tắt của hệ thống tệp phân tán, là hệ thống tệp lưu trữ dữ liệu trên máy chủ**. Truy cập và xử lý dữ liệu như được lưu trữ trên máy khách cục bộ. Thông qua DFS, bạn có thể dễ dàng chia sẻ thông tin và tệp giữa những người dùng trên mạng một cách có kiểm soát và được ủy quyền.

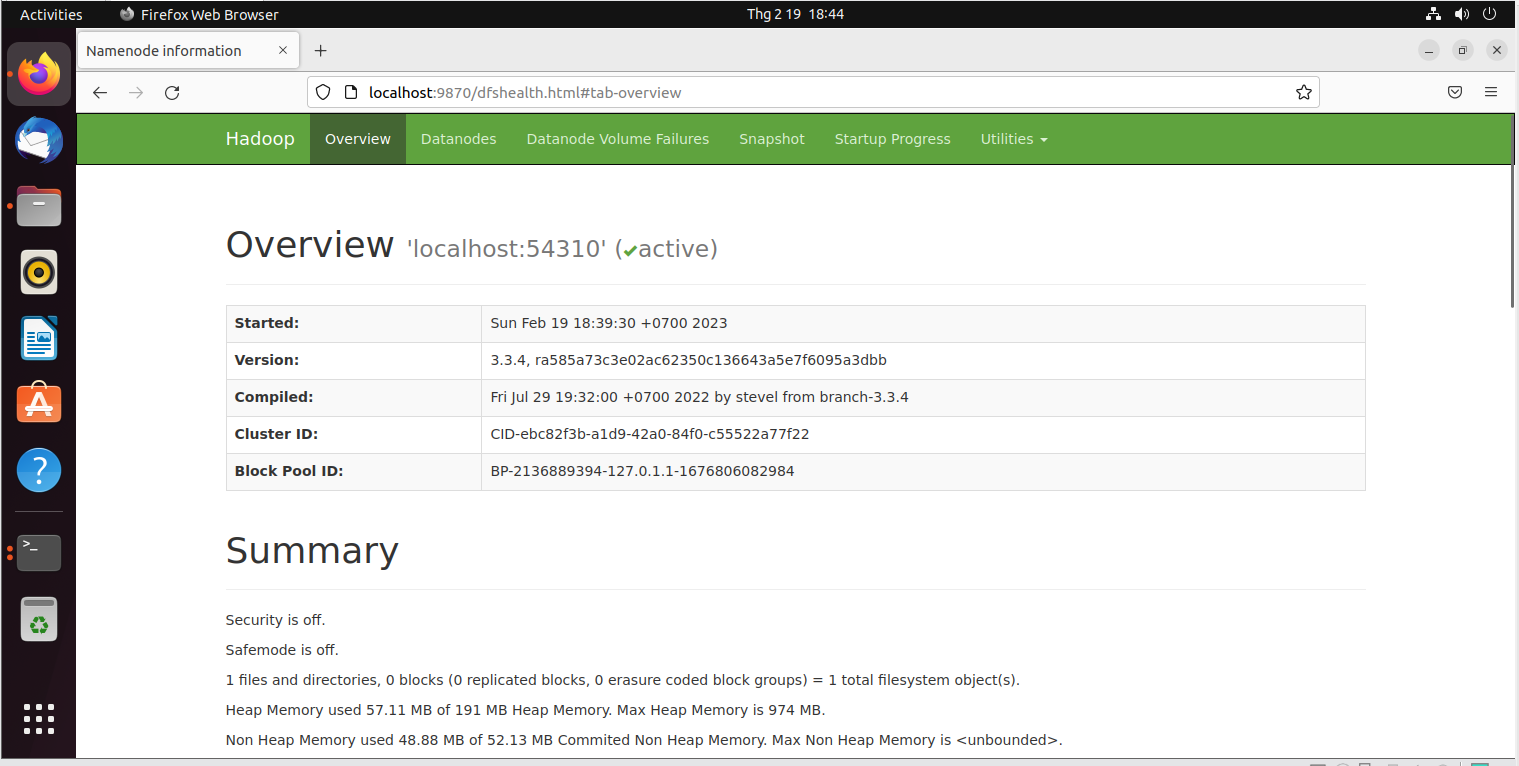
Khởi động dfs.sh



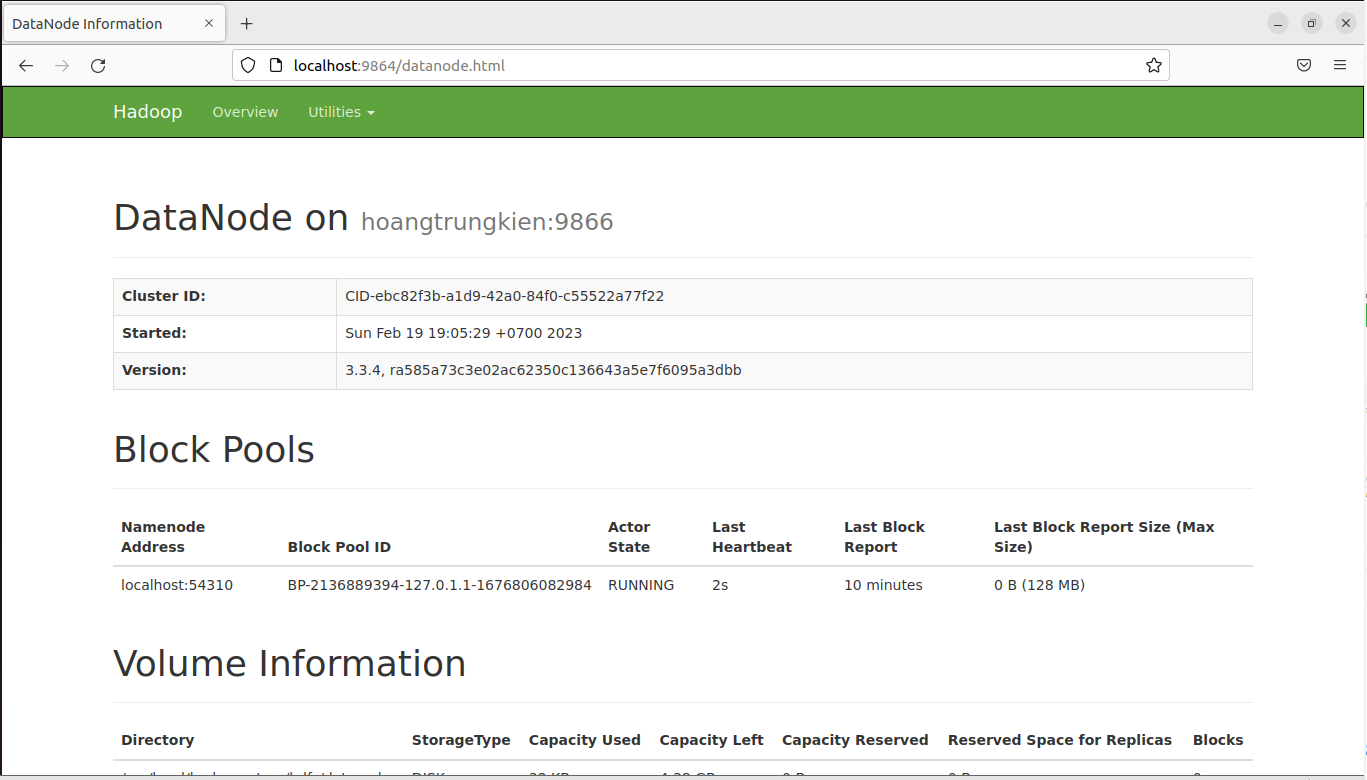
### **Truy cập Hadoop UI từ trình duyệt**

Truy cập <http://localhost:9870/>

Hiển thị là đã cài đặt thành công

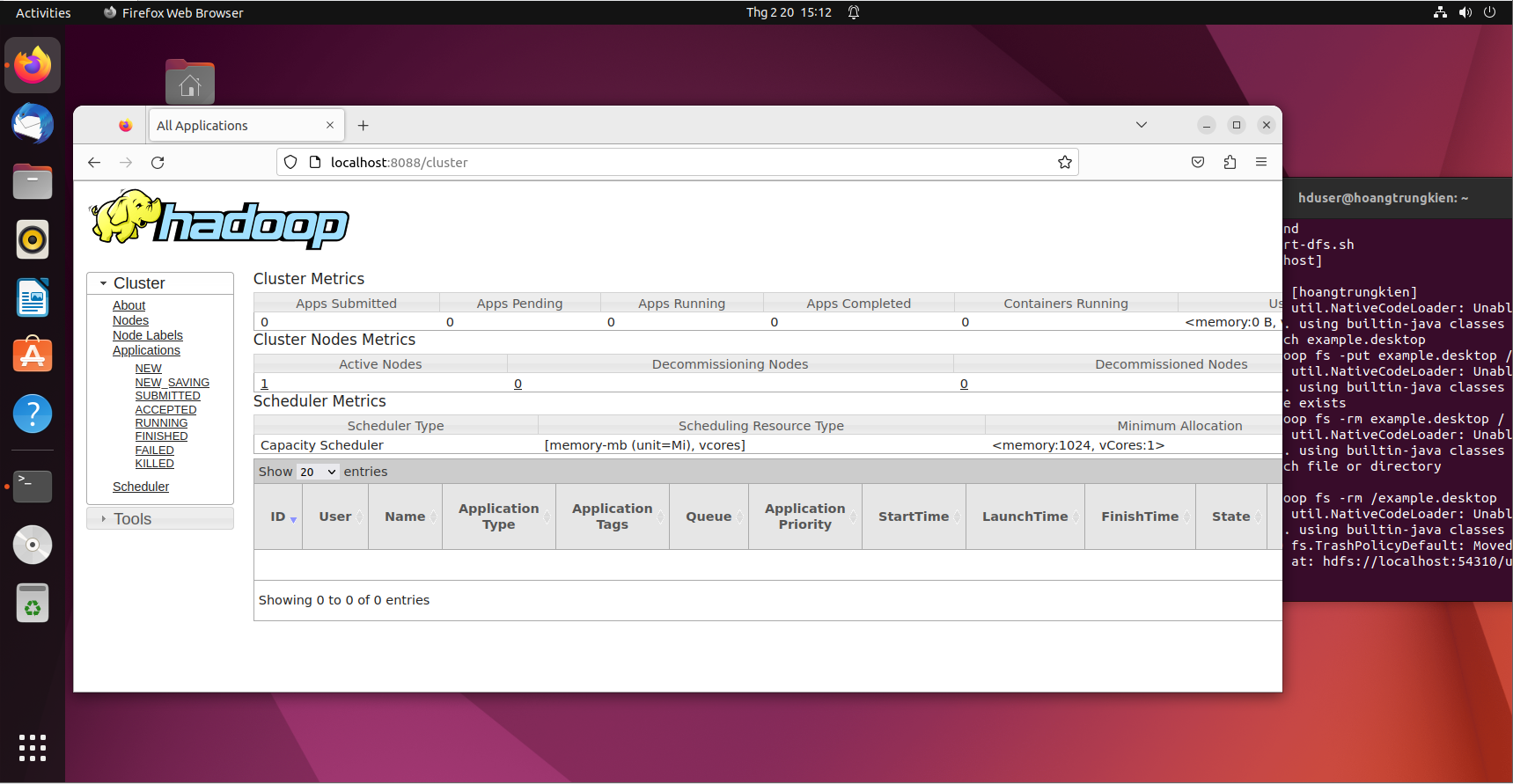


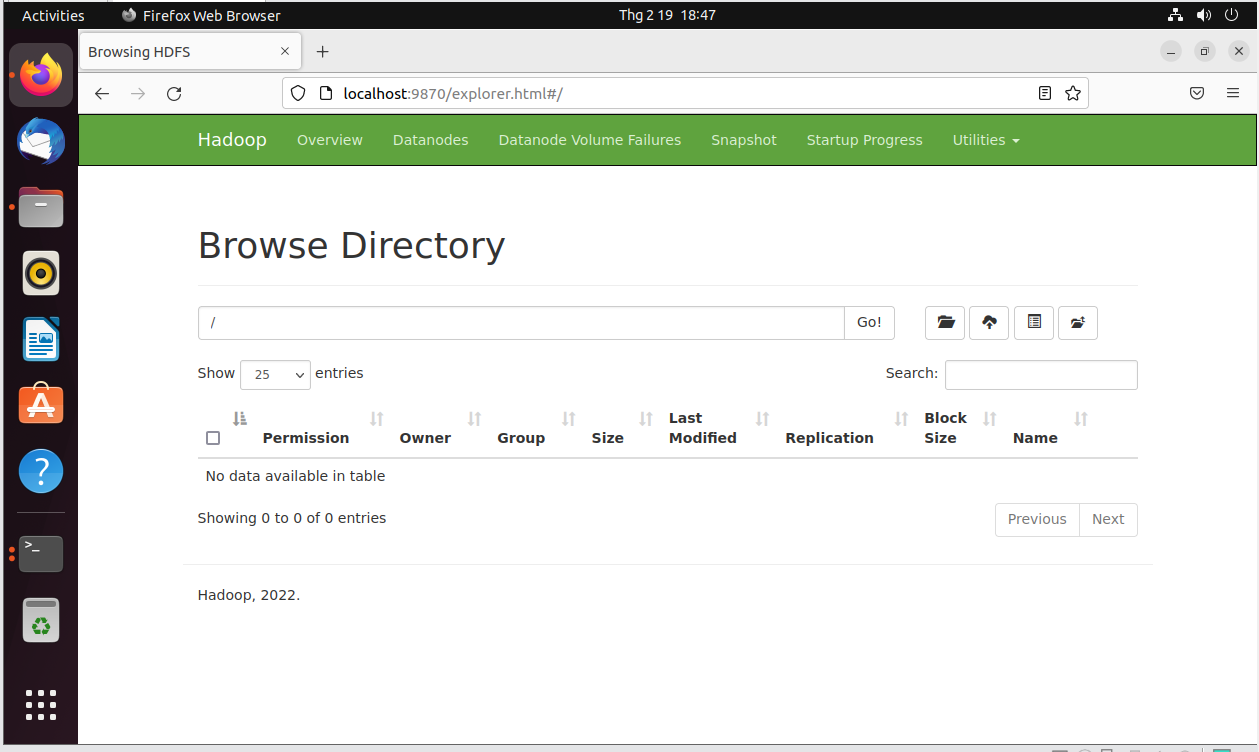
Kiểm tra datanode tại cổng mặc định 9864 :

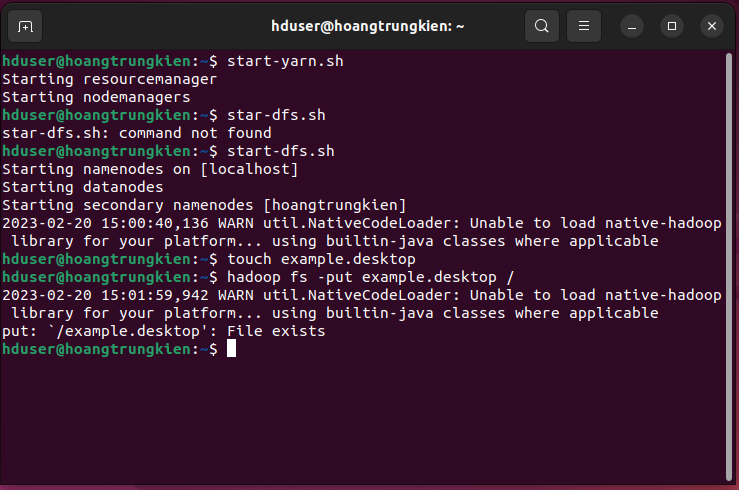


Kiểm tra Yarn resource manager tại cổng 8088 :

locahost:8088









Thêm vào giữa 2 thẻ configuration của **file core-site.xml**

<property>

<name>fs.trash.interval</name>

<value>3</value>

</property>

<property>

<name>fs.trash.checkpoint.interval</name>

<value>1</value>

</property>

<property>

<name>hadoop.http.staticuser.user</name>

<value>hduser</value>

</property>

Thêm vào giữa 2 thẻ configuration của **file hdfs-site.xml**

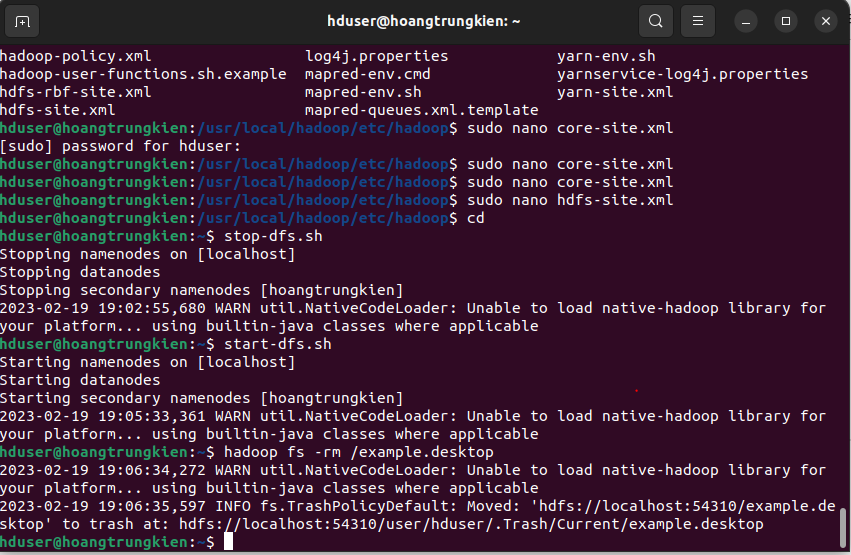
<property>

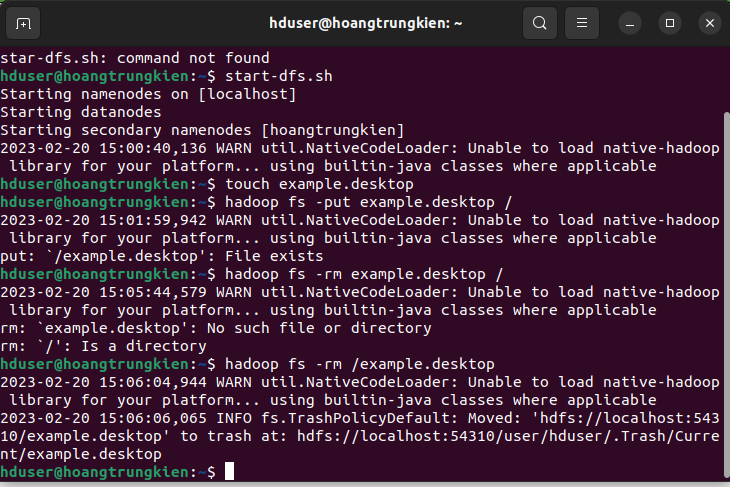
<name>dfs.permissions.enabled</name>

<value>true</value>

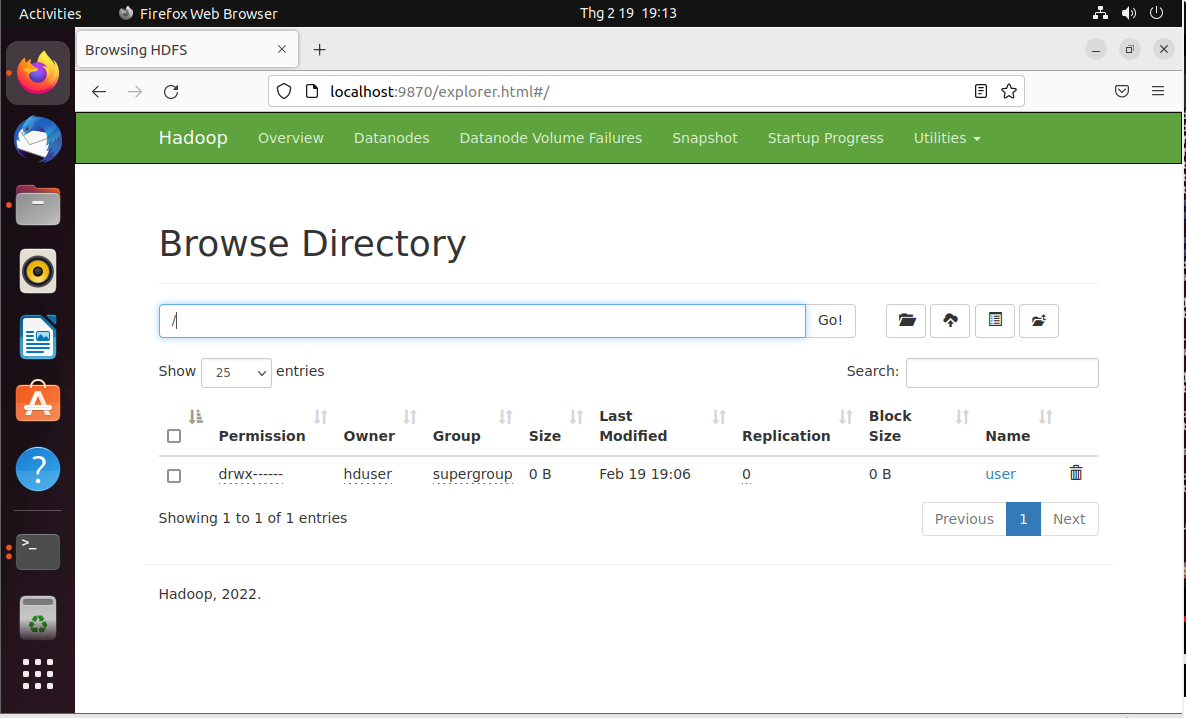
</property>

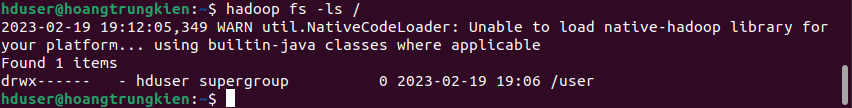
Tạm dừng và khởi động lại:

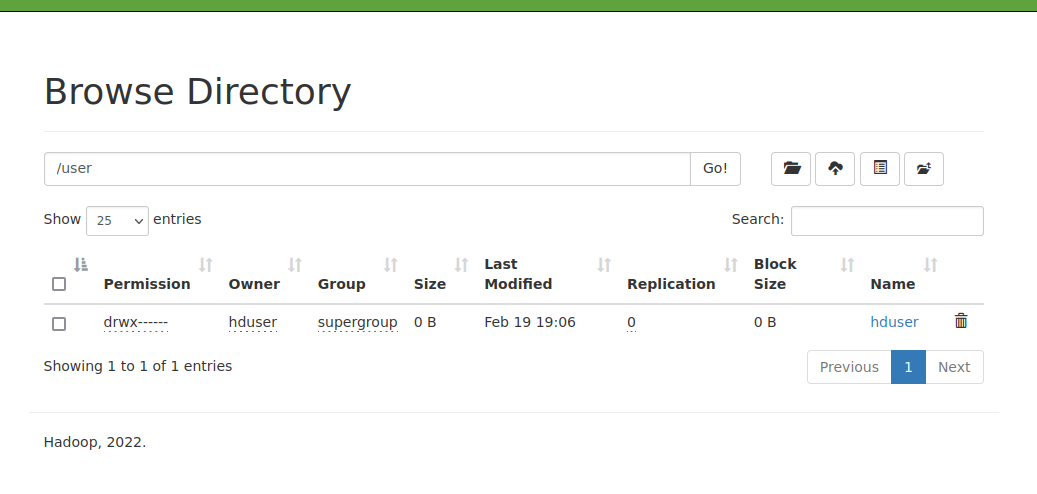


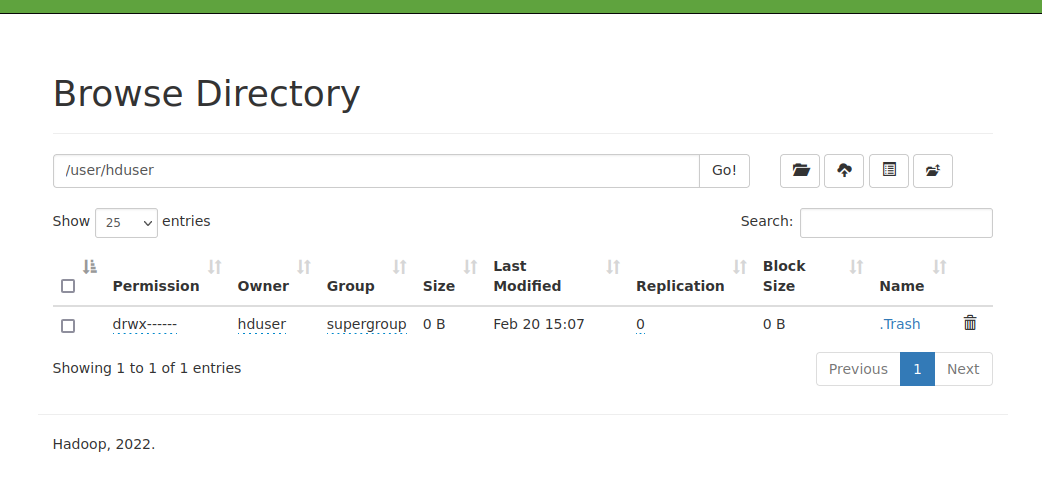


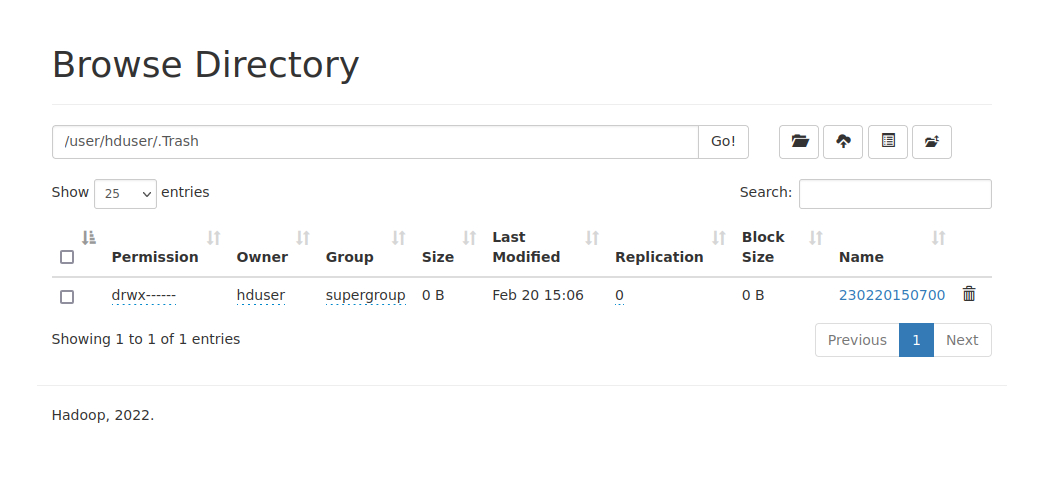
Đã chuyển: 'hdfs://localhost:54310/example.desktop' vào thùng rác tại: hdfs://localhost:54310/user/hduser/.Trash/Current /example.desktop













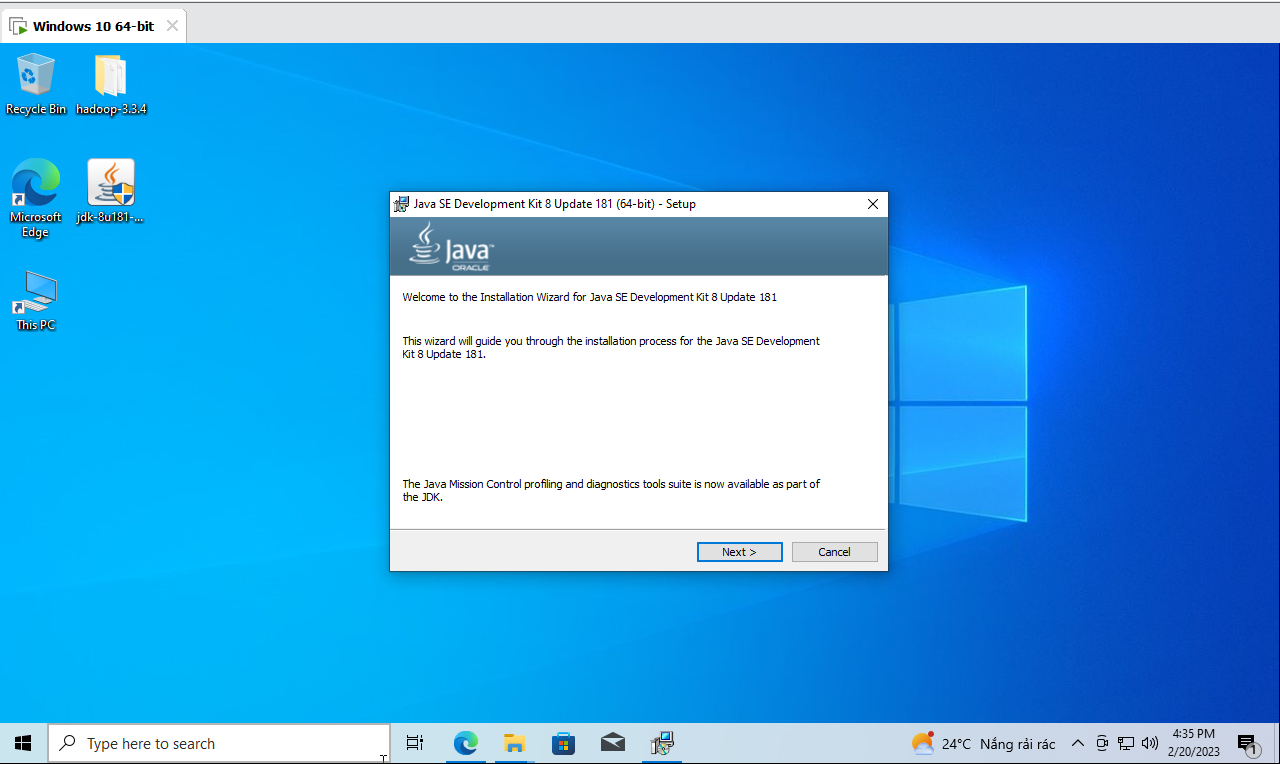
Đầu tiên ta tải Hadoop và jdk 1.8

[**https://dlcdn.apache.org/hadoop/common/hadoop-3.3.4/hadoop-3.3.4.tar.gz**](https://dlcdn.apache.org/hadoop/common/hadoop-3.3.4/hadoop-3.3.4.tar.gz)

<https://www.oracle.com/java/technologies/javase/javase8-archive-downloads.html>

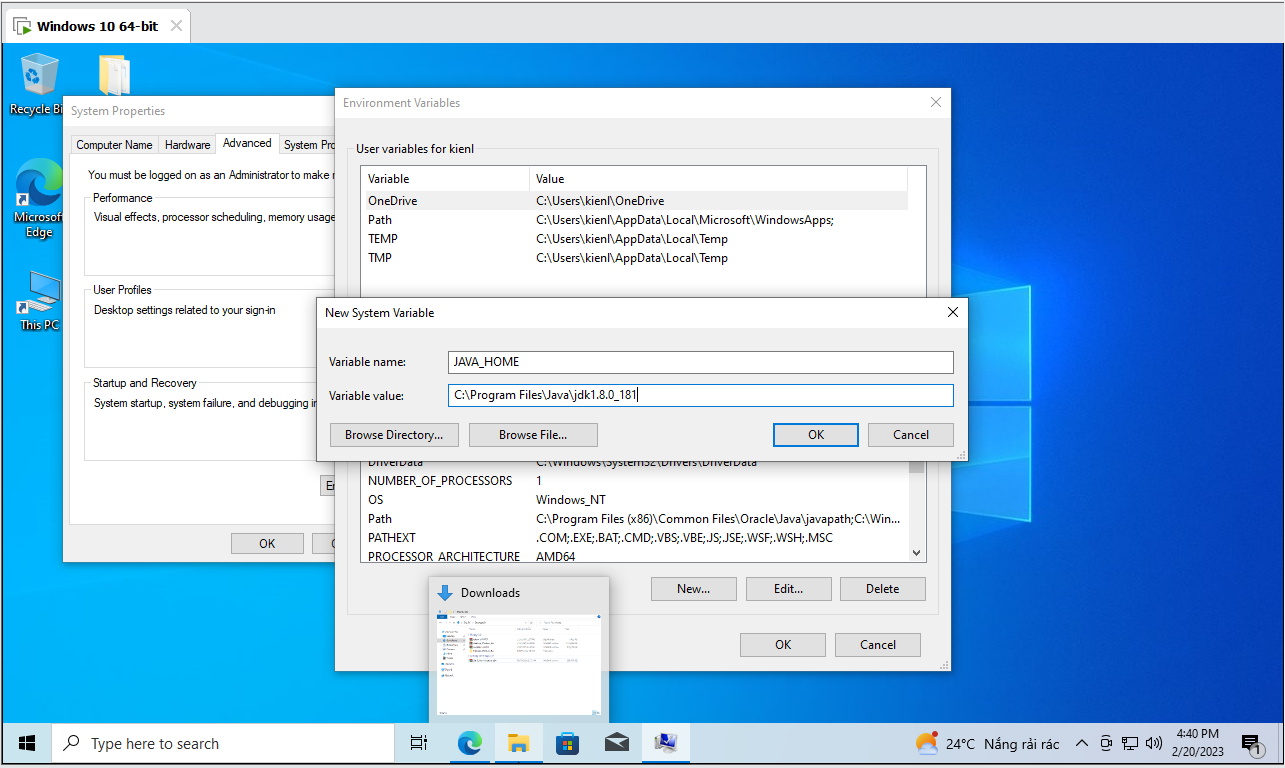


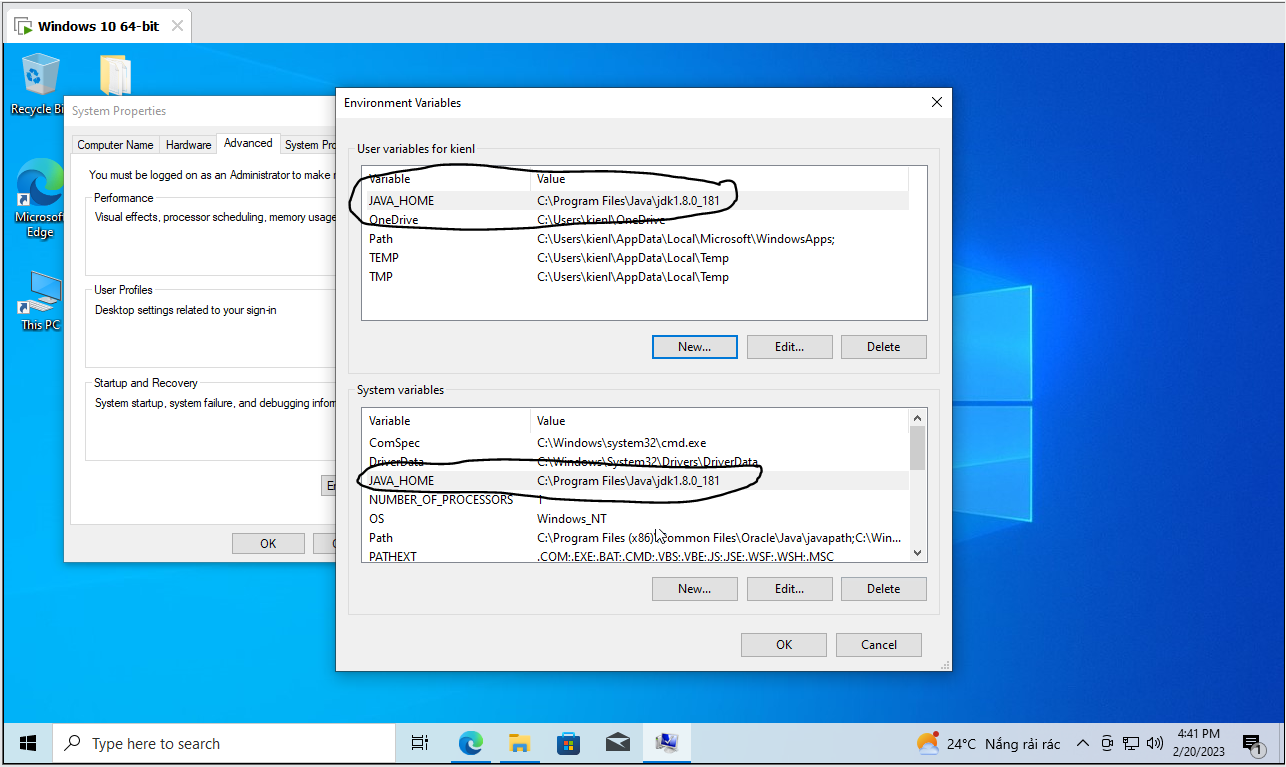
Sau khi tải xong, chúng ta giải nén và cài đặt jdk



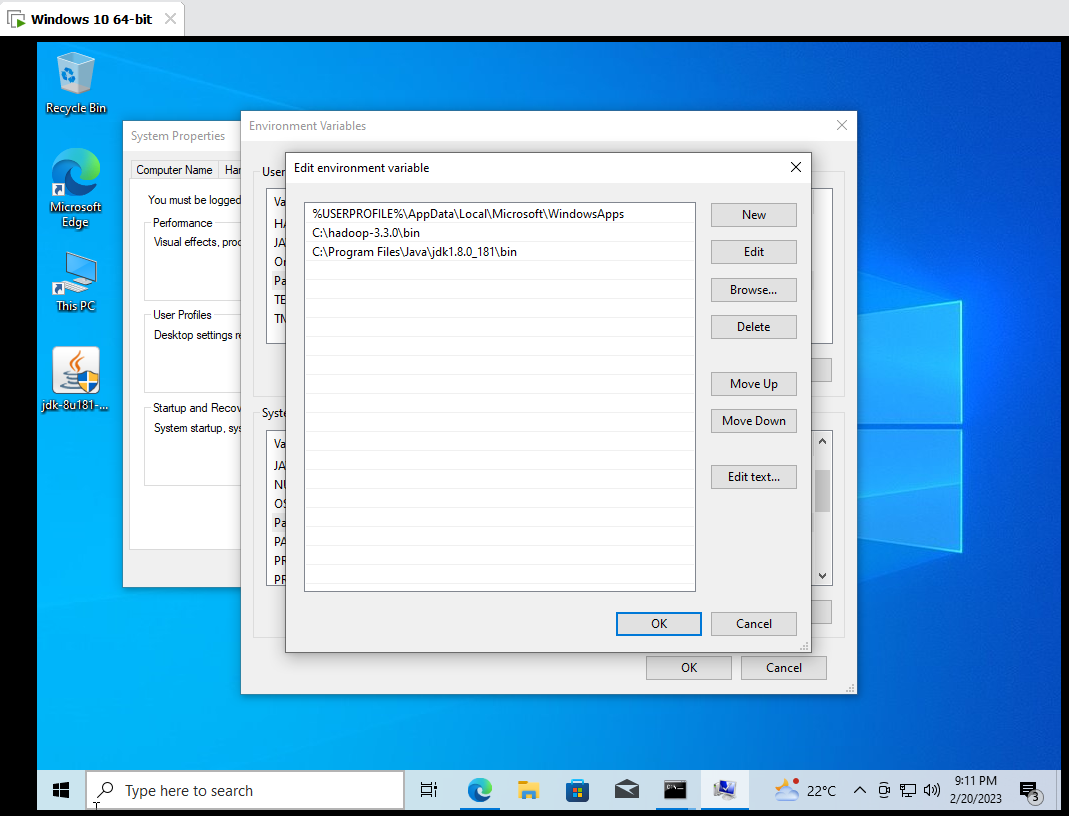
Sau khi cài đặt xong, ta Thiết lập biến môi trường cho Java JDK

Cần cấu hình biến môi trường JAVA\_HOME cho Java JDK





Tiếp theo cấu hình Path (cho cả user và system variable). Tìm tới biến Path, nhấn Edit:



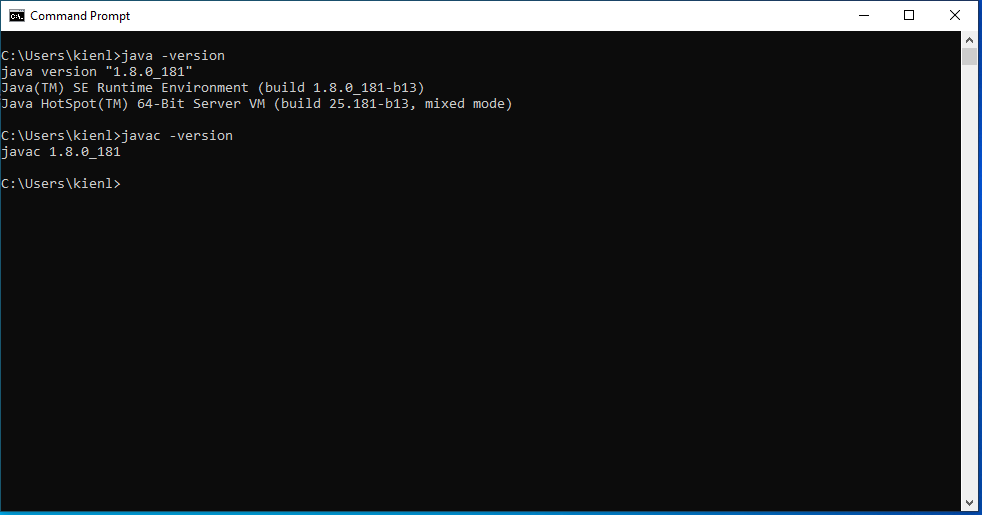
nhấn OK

Kiểm tra lại quá trình cấu hình bằng cách Gõ phím Windows +R–>gõ cmd:

Trong màn hình command line lên gõ các lệnh trên để thấy kết quả:

**java -version**

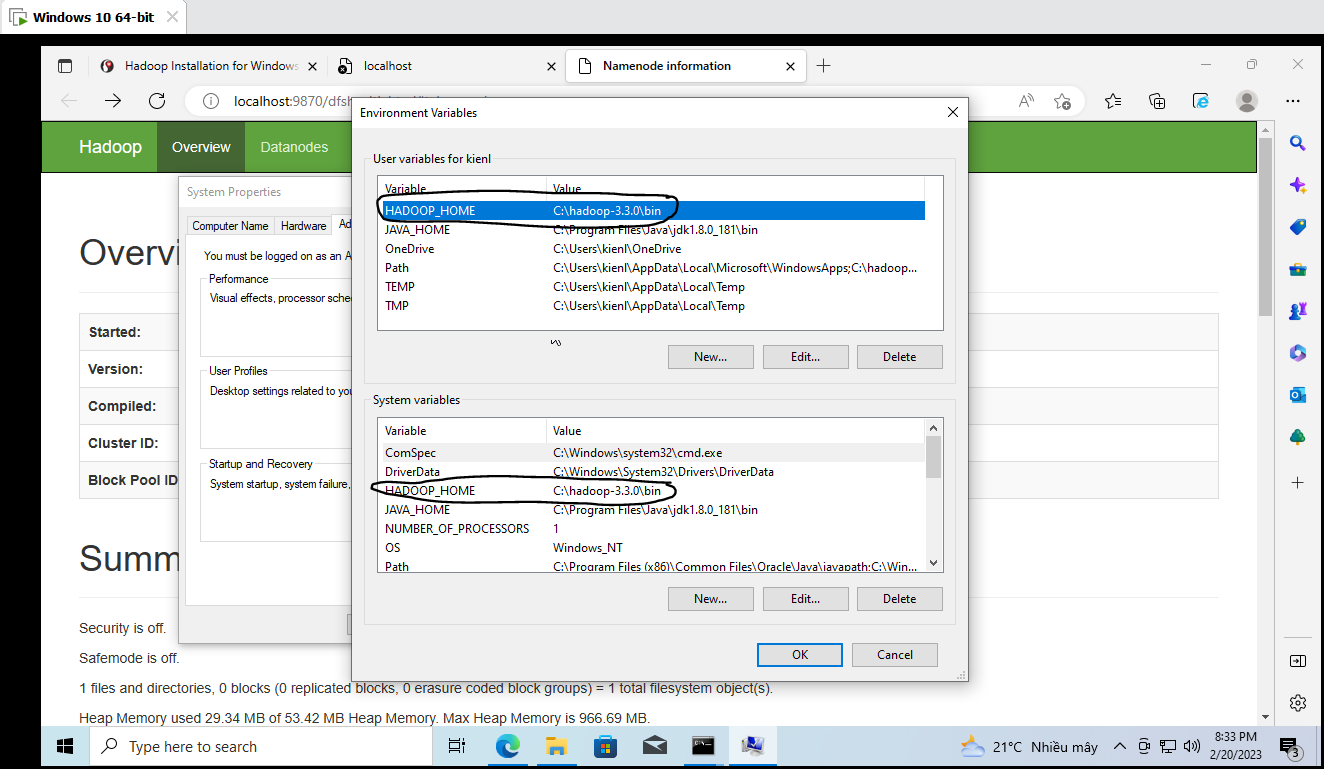
**javac -version**



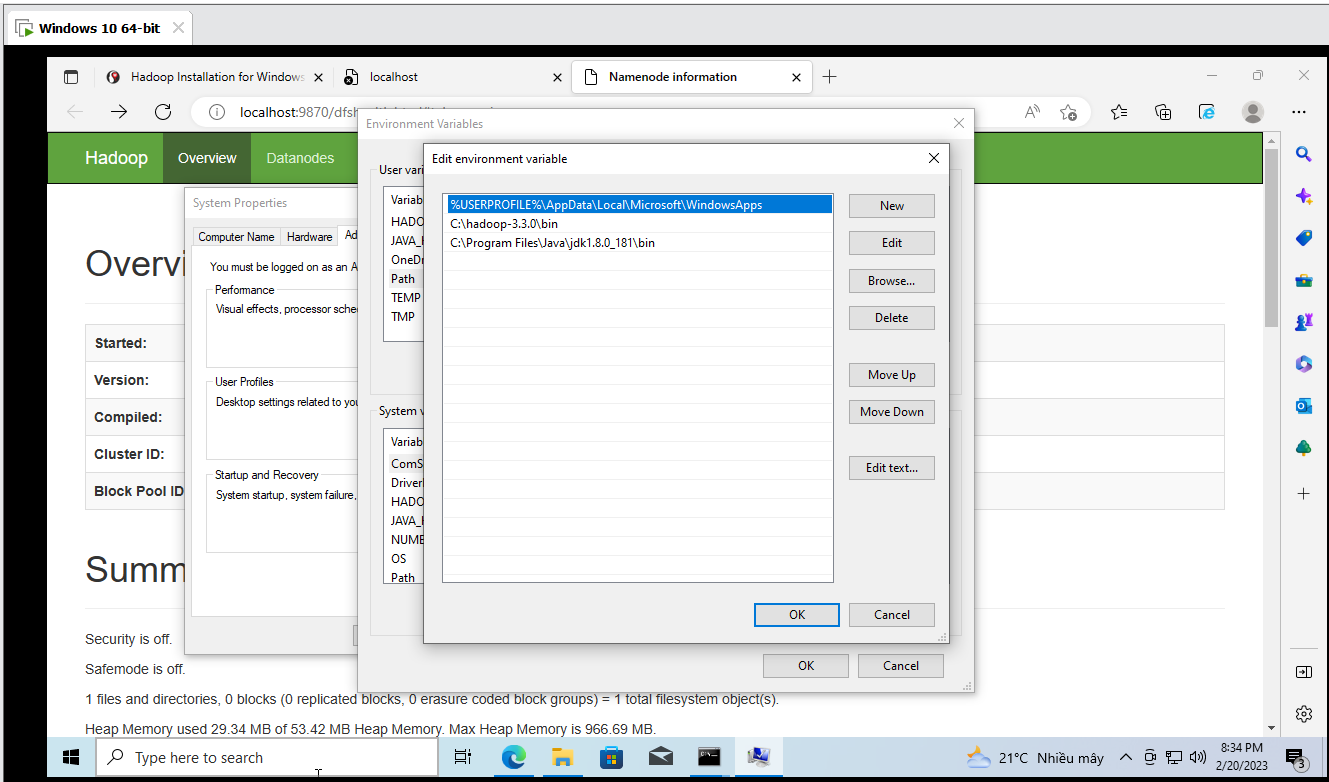
Tiếp theo là cài đặt Hadoop

Tiếp thiết lập biến môi trường cho Hadoop

Tương tự như JAVA JDK, ta cần cấu hình biến môi trường cho Hadoop (**HADOOP\_HOME)**

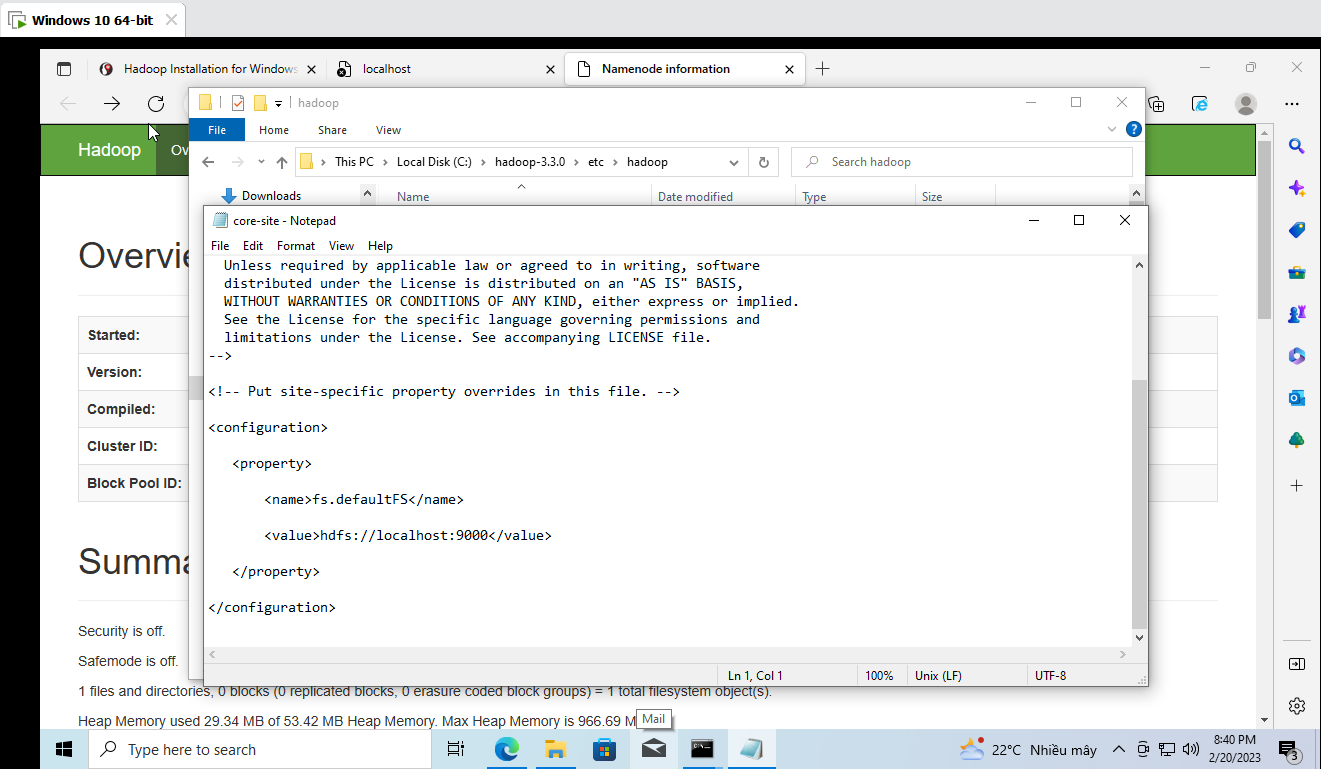


Lần lượt trong user và system variable thêm biến HADOOP\_HOME có giá trị là C:\hadoop-3.3.0\bin mà ta giải nén ở trên.

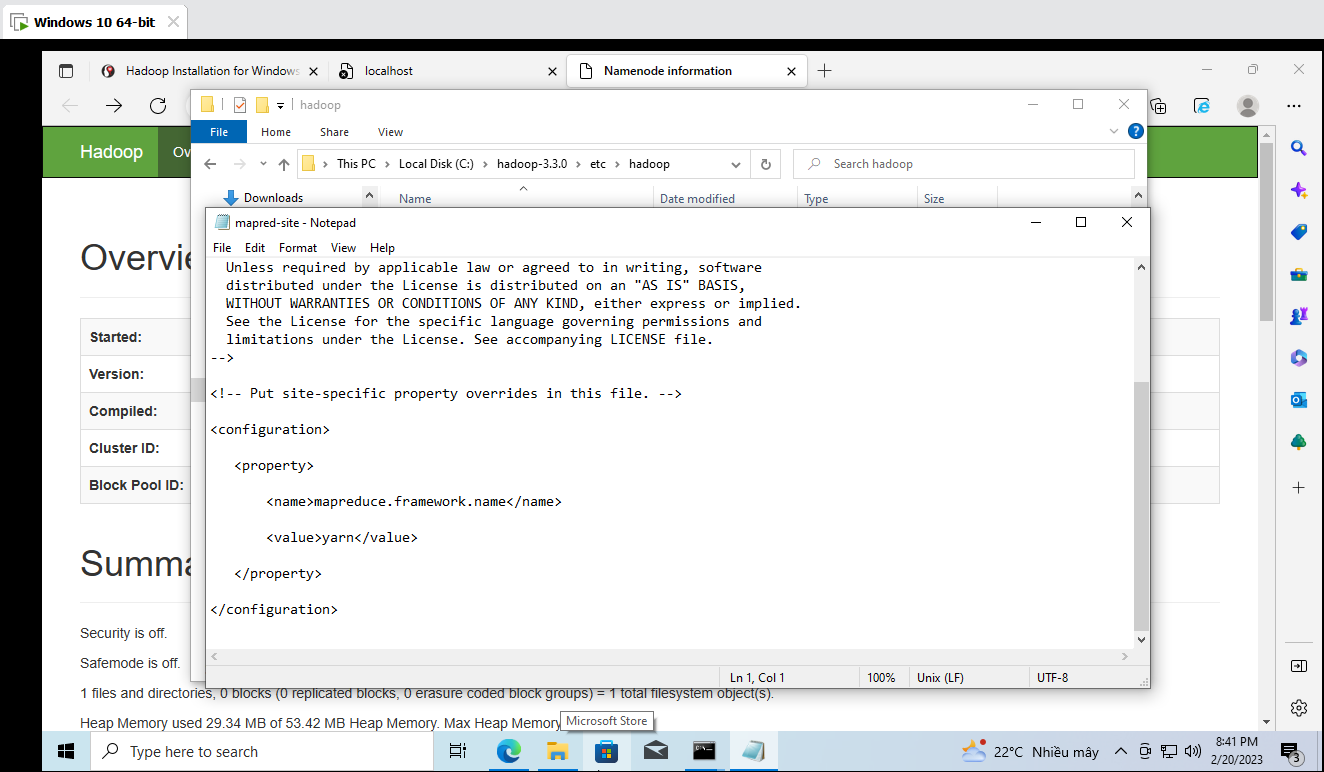


Lần lượt trong user và system variable thêm biến HADOOP\_HOME có giá trị là C:\hadoop-3.3.0\bin mà ta giải nén ở trên.

**Sửa file C:/Hadoop-3.3.0/etc/hadoop/core-site.xml**



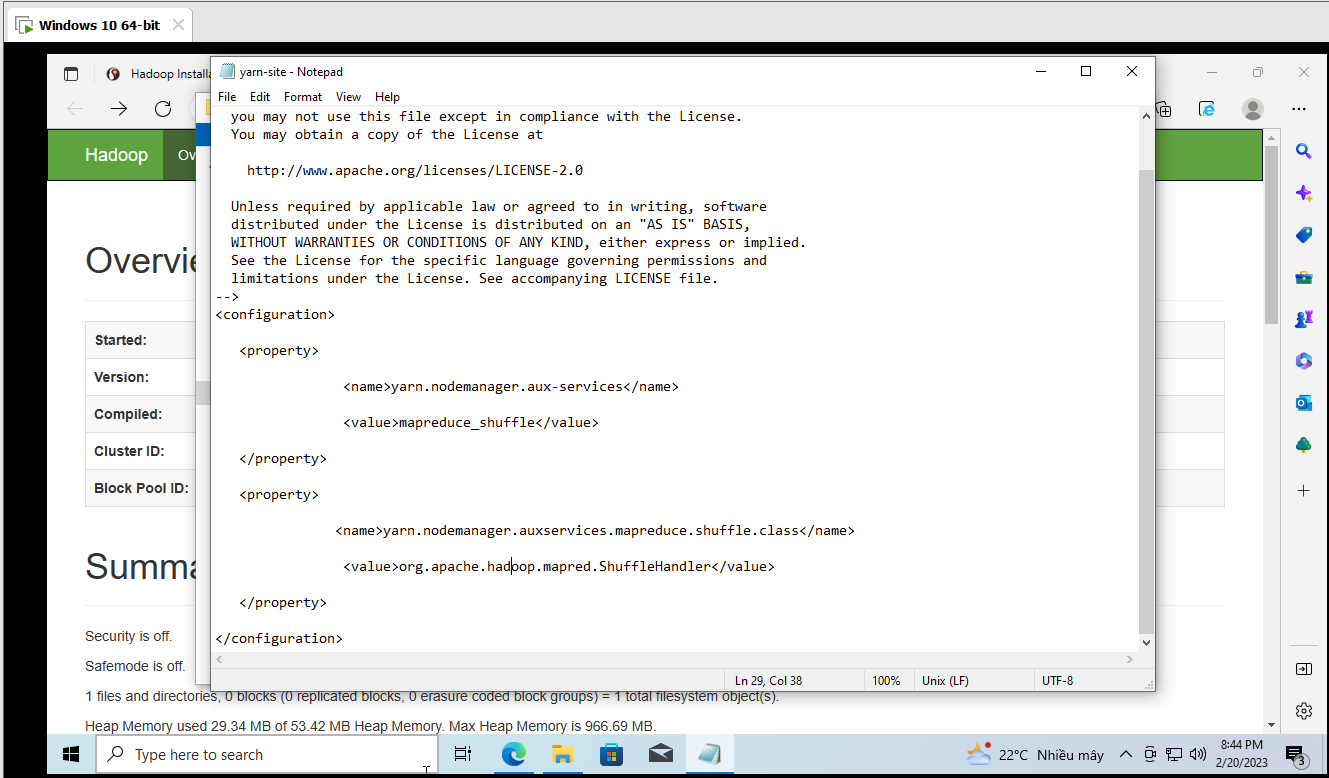
**Sửa file C:/Hadoop-3.3.0/etc/hadoop/mapred-site.xml**



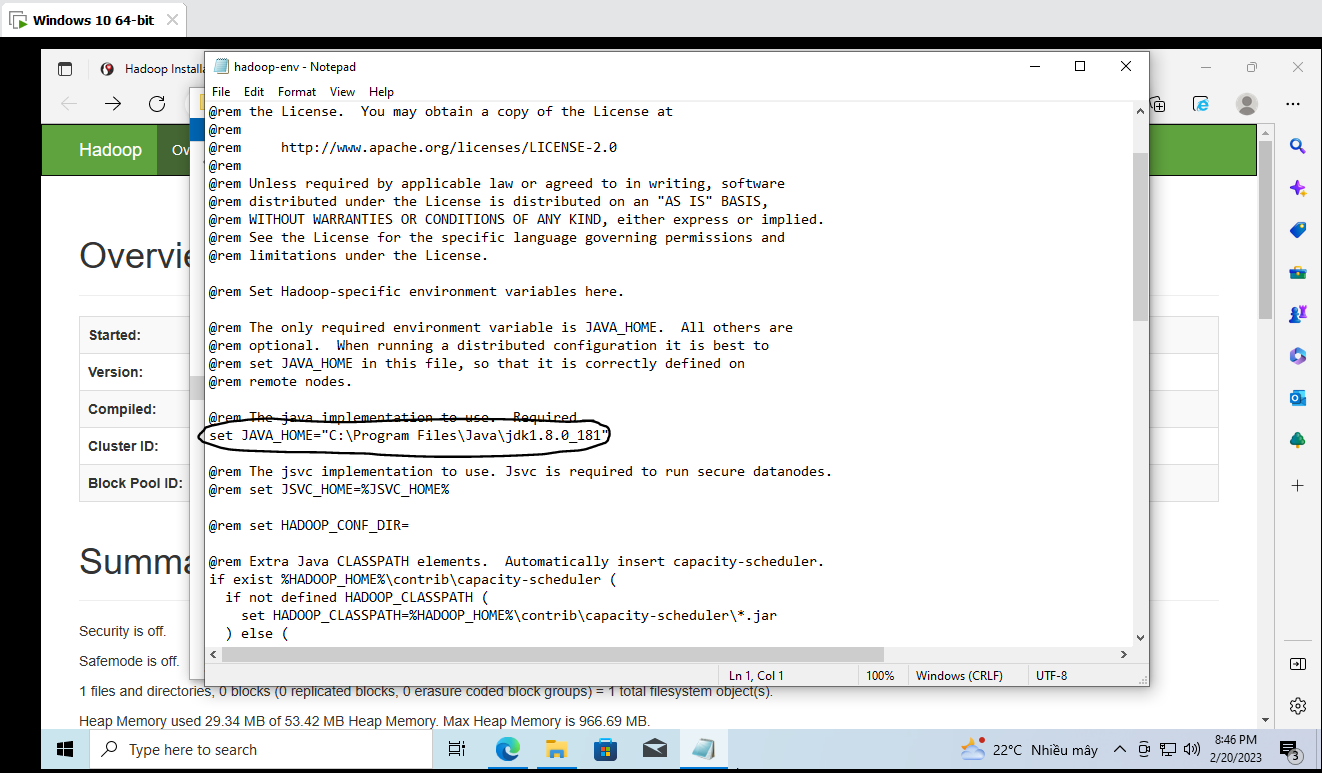
**Sửa file C:\Hadoop-3.3.0/etc/hadoop/hdfs-site.xml**



**Sửa file C:/Hadoop-3.3.0/etc/hadoop/yarn-site.xml**



**Sửa file C:/Hadoop-3.3.0/etc/hadoop/hadoop-env.cmd**

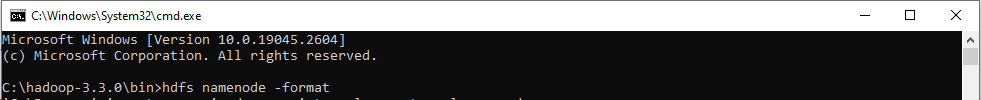


**Create folder “data” under “C:\Hadoop-3.3.0”**

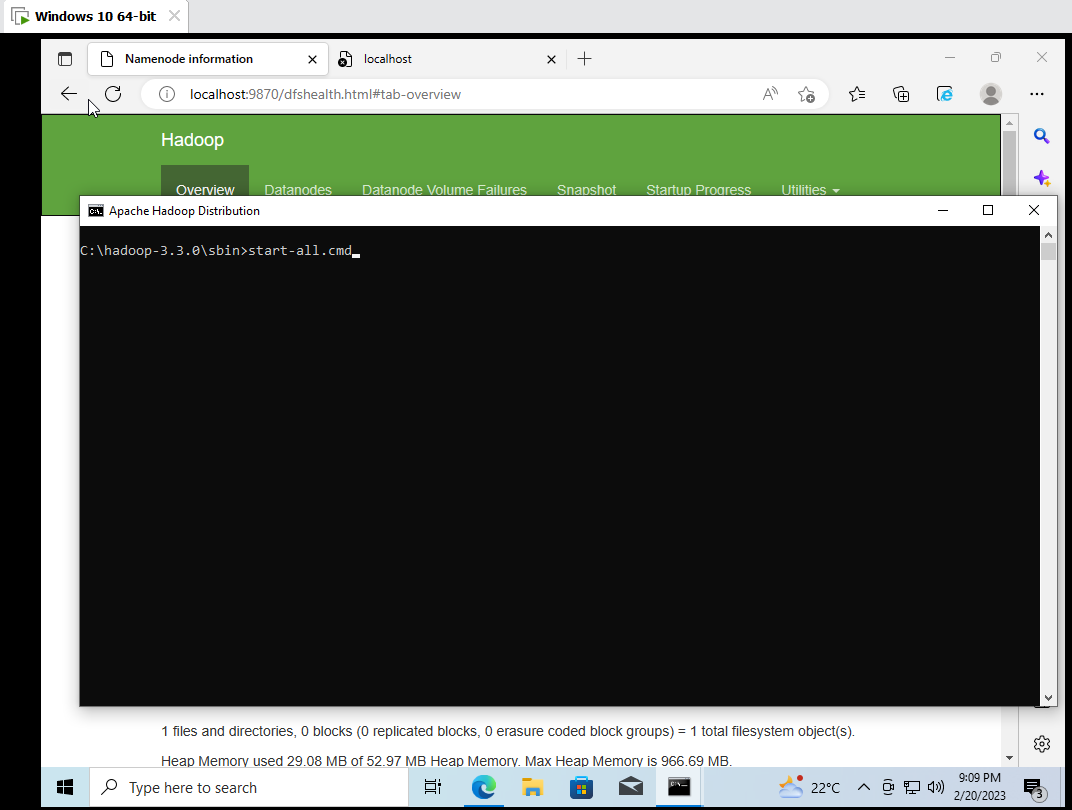
**Create folder “datanode” under “C:\Hadoop-3.3.0\data”**

**Create folder “namenode” under “C:\Hadoop-3.3.0\data”**

Sau đó format lại namenode và datanode là mở File C:\hadoop-3.3.0:\bin bằng cmd và chạy lệnh hdfs namenode –format



**Cuối cùng là mở File C:\hadoop-3.3.0:\sbin bằng cmd và chạy lệnh start-all.cmd**



Sau khi gõ lệnh trên, hệ thống sẽ chạy Hadoop

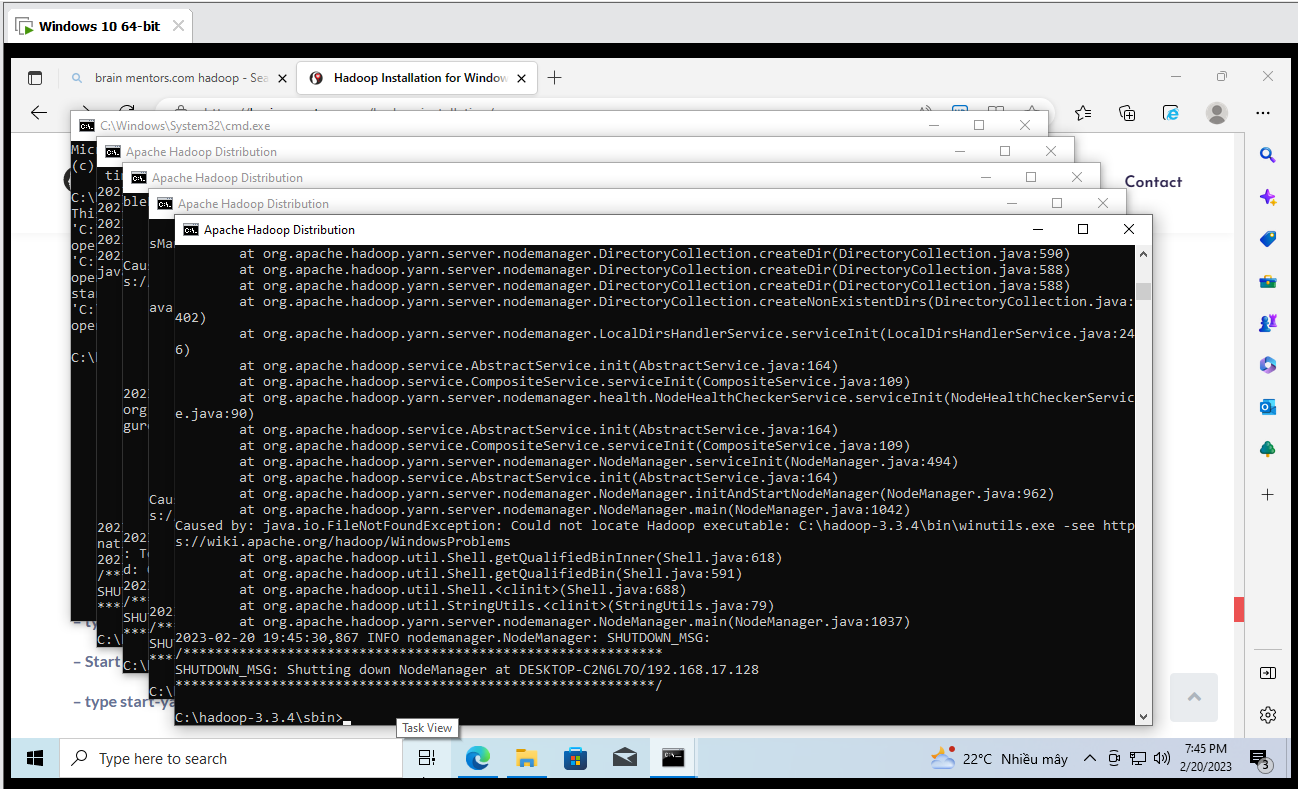
Phải đảm bảo các ứng dụng sau được chạy:

– Hadoop Namenode

– Hadoop datanode

– YARN Resource Manager

– YARN Node Manager



Truy cập <http://localhost:8088>

**Hiển thị là Hadoop đã được cài đặt thành công**

