

Tài liệu kỹ thuật

CÁCH CÀI ĐẶT HADOOP-3.3.0 TRÊN WINDOWS

Big Data đang trở thành một phần thể mạnh và là tài sản rất lớn của mỗi công ty, tổ chức, và Hadoop là công nghệ cốt lõi cho việc lưu trữ và truy cập dữ liệu lớn.

Tài liệu này sẽ trình bày sơ lược lý thuyết về Hadoop và cách cài đặt Hadoop trên hệ điều hành Windows.

Nội dung tài liệu gồm các phần dưới đây:

1. Giới thiệu Hadoop
2. Cài đặt Java JDK 1.8
3. Thiết lập biến môi trường cho Java JDK
4. Tải Hadoop và giải nén vào ổ C
5. Thiết lập biến môi trường cho Hadoop
6. Cấu hình các tập tin cho Hadoop
7. Cập nhật các Hadoop Configurations
8. Hoàn thành cài đặt Hadoop và test thử nghiệm với start-all.cmd

Lưu ý: Mọi thư mục cài đặt: Không có dấu + không có khoảng trắng

Mục Lục

1. Giới thiệu Hadoop	3
1.1 Hadoop là gì?.....	3
1.2 Chức năng nhiệm vụ của Hadoop.....	3
1.3 Kiến trúc Hadoop.....	3
1.4 Cơ chế hoạt động của Hadoop.....	5
2. Cài đặt JDK bản 1.8 (bắt buộc)	6
3) Thiết lập biến môi trường cho Java JDK.....	11
4) Tải Hadoop và giải nén vào ổ C.....	15
5) Thiết lập biến môi trường cho Hadoop	17
6) Cấu hình các tập tin cho Hadoop	19
7) Cập nhật các Hadoop Configurations.....	21
8) Hoàn thành cài đặt Hadoop và test thử nghiệm với start-all.cmd.....	22

1. Giới thiệu Hadoop

Lý thuyết phần Hadoop được tham khảo từ internet, được tổng hợp lại. Không phải do tác giả tự nghĩ ra. Phần lý thuyết của mục này mọi người có thể tìm thấy trên internet.

1.1 Hadoop là gì?

Hadoop là một Apache framework mã nguồn mở cho phép phát triển các ứng dụng phân tán để lưu trữ và quản lý các tập dữ liệu lớn. Hadoop hiện thực mô hình MapReduce, mô hình mà ứng dụng sẽ được chia nhỏ ra thành nhiều phân đoạn khác nhau được chạy song song trên nhiều node khác nhau.

Hadoop được viết bằng Java tuy nhiên vẫn hỗ trợ C++, Python, Perl bằng cơ chế streaming.

Hadoop có các điểm lợi sau:

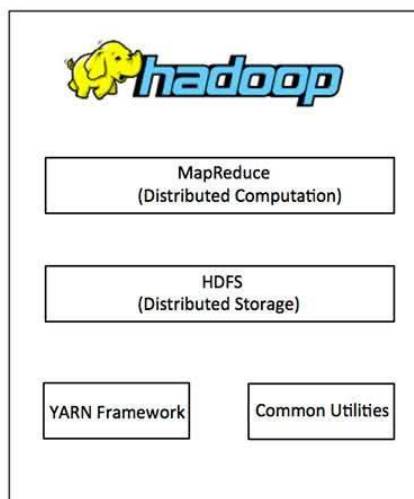
- Robust and Scalable – Có thể thêm node mới và thay đổi chúng khi cần.
- Affordable and Cost Effective – Không cần phần cứng đặc biệt để chạy Hadoop.
- Adaptive and Flexible – Hadoop được xây dựng với tiêu chí xử lý dữ liệu có cấu trúc và không cấu trúc.
- Highly Available and Fault Tolerant – Khi 1 node lỗi, nền tảng Hadoop tự động chuyển sang node khác.

1.2 Chức năng nhiệm vụ của Hadoop

- Xử lý và làm việc khối lượng dữ liệu khổng lồ tính bằng Petabyte.
- Xử lý trong môi trường phân tán, dữ liệu lưu trữ ở nhiều phần cứng khác nhau, yêu cầu xử lý đồng bộ
- Các lỗi xuất hiện thường xuyên.
- Bảng thông giữa các phần cứng vật lý chứa dữ liệu phân tán có giới hạn.

1.3 Kiến trúc Hadoop

Một cụm Hadoop nhỏ gồm 1 master node và nhiều worker/slave node. Toàn bộ cụm chứa 2 lớp, một lớp MapReduce Layer và lớp kia là HDFS Layer. Mỗi lớp có các thành phần liên quan riêng. Master node gồm JobTracker, TaskTracker, NameNode, và DataNode. Slave/worker node gồm DataNode, và TaskTracker. Cũng có thể slave/worker node chỉ là dữ liệu hoặc node để tính toán.



Hadoop framework gồm 4 module:

Module 1: Hadoop Distributed File System (HDFS)

Đây là hệ thống file phân tán cung cấp truy cập thông lượng cao cho ứng dụng khai thác dữ liệu. **Hadoop Distributed File System (HDFS)** là hệ thống tập tin ảo. Khi chúng ta di chuyển 1 tập tin trên HDFS, nó tự động chia thành nhiều mảnh nhỏ. Các đoạn nhỏ của tập tin sẽ được nhân rộng và lưu trữ trên nhiều máy chủ khác để tăng sức chịu lỗi và tính sẵn sàng cao.

HDFS sử dụng kiến trúc master/slave, trong đó master gồm một NameNode để quản lý hệ thống file metadata và một hay nhiều slave DataNodes để lưu trữ dữ liệu thực tại.

Một tập tin với định dạng HDFS được chia thành nhiều khối và những khối này được lưu trữ trong một tập các DataNodes. NameNode định nghĩa ánh xạ từ các khối đến các DataNode. Các DataNode điều hành các tác vụ đọc và ghi dữ liệu lên hệ thống file. Chúng cũng quản lý việc tạo, hủy, và nhân rộng các khối thông qua các chỉ thị từ NameNode.

Module 2: Hadoop MapReduce

Đây là hệ thống dựa trên YARN dùng để xử lý song song các tập dữ liệu lớn. Là cách chia một vấn đề dữ liệu lớn hơn thành các đoạn nhỏ hơn và phân tán nó trên nhiều máy chủ. Mỗi máy chủ có 1 tập tài nguyên riêng và máy chủ xử lý dữ liệu trên cục bộ. Khi máy chủ xử lý xong dữ liệu, chúng sẽ gửi trở về máy chủ chính.

MapReduce gồm một single master (máy chủ) JobTracker và các slave (máy trạm) TaskTracker trên mỗi cluster-node. Master có nhiệm vụ quản lý tài nguyên, theo dõi quá trình tiêu thụ tài nguyên và lập lịch quản lý các tác vụ trên các máy trạm, theo dõi chúng

và thực thi lại các tác vụ bị lỗi. Những máy slave TaskTracker thực thi các tác vụ được master chỉ định và cung cấp thông tin trạng thái tác vụ (task-status) để master theo dõi.

JobTracker là một điểm yếu của Hadoop Mapreduce. Nếu JobTracker bị lỗi thì mọi công việc liên quan sẽ bị ngắt quãng.

Module 3: Hadoop Common

Đây là các thư viện và tiện ích cần thiết của Java để các module khác sử dụng. Những thư viện này cung cấp hệ thống file và lớp OS trừu tượng, đồng thời chứa các mã lệnh Java để khởi động Hadoop.

Module 4: Hadoop YARN

Quản lý tài nguyên của các hệ thống lưu trữ dữ liệu và chạy phân tích.

1.4 Cơ chế hoạt động của Hadoop

Giai đoạn 1:

Một user hay một ứng dụng có thể submit một job lên Hadoop (hadoop job client) với yêu cầu xử lý cùng các thông tin cơ bản:

Nơi lưu (location) dữ liệu input, output trên hệ thống dữ liệu phân tán.

Các java class ở định dạng jar chứa các dòng lệnh thực thi các hàm map và reduce.

Các thiết lập cụ thể liên quan đến job thông qua các thông số truyền vào.

Giai đoạn 2:

Hadoop job client submit job (file jar, file thực thi) và các thiết lập cho JobTracker. Sau đó, master sẽ phân phối tác vụ đến các máy slave để theo dõi và quản lý tiến trình các máy này, đồng thời cung cấp thông tin về tình trạng và chẩn đoán liên quan đến job-client.

Giai đoạn 3:

TaskTrackers trên các node khác nhau thực thi tác vụ MapReduce và trả về kết quả output được lưu trong hệ thống file.

Khi “chạy Hadoop” có nghĩa là chạy một tập các trình nền – daemon, hoặc các chương trình thường trú, trên các máy chủ khác nhau trên mạng của bạn. Những trình nền có vai trò cụ thể, một số chỉ tồn tại trên một máy chủ, một số có thể tồn tại trên nhiều máy chủ.

Các daemon bao gồm:

NameNode

DataNode

SecondaryNameNode














JobTracker

TaskTracker

2. Cài đặt JDK bản 1.8 (bắt buộc)

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

Ví dụ cài bản Java SE Development Kit 8u201:

Java SE Development Kit 8u201		
This software is licensed under the Oracle Binary Code License Agreement for Java SE Platform Products		
Product / File Description	File Size	Download
Linux ARM v6/v7 Soft Float ABI	72.98 MB	 jdk-8u201-linux-arm32-vfp-hflt.tar.gz
Linux ARM v6/v7 Soft Float ABI	69.92 MB	 jdk-8u201-linux-arm64-vfp-hflt.tar.gz
Linux x86	170.98 MB	 jdk-8u201-linux-i586.rpm
Linux x86	185.77 MB	 jdk-8u201-linux-i586.tar.gz
Linux x64	168.05 MB	 jdk-8u201-linux-x64.rpm
Linux x64	182.93 MB	 jdk-8u201-linux-x64.tar.gz
Mac OS X x64	245.92 MB	 jdk-8u201-macosx-x64.dmg
Solaris SPARC 64-bit (SVR4 package)	125.33 MB	 jdk-8u201-solaris-sparcv9.tar.Z
Solaris SPARC 64-bit	88.31 MB	 jdk-8u201-solaris-sparcv9.tar.gz
Solaris x64 (SVR4 package)	133.99 MB	 jdk-8u201-solaris-x64.tar.Z
Solaris x64	92.16 MB	 jdk-8u201-solaris-x64.tar.gz
Windows x86	197.66 MB	 jdk-8u201-windows-i586.exe
Windows x64	207.46 MB	 jdk-8u201-windows-x64.exe

Bấm vào link để tải, chương trình xuất hiện như bên dưới:

Mac OS X x64	245.92 MB	jdk-8u201-macosx-x64.dmg
Solaris SPARC 64-bit (SVR4 package)	125.33 MB	jdk-8u201-solaris-sparcv9.tar.Z
Solaris SPARC 64-bit	88.31 MB	jdk-8u201-solaris-sparcv9.tar.gz
Solaris x64 (SVR4 package)		jdk-8u201-solaris-sparcv9.tar.Z
Solaris x64		jdk-8u201-solaris-sparcv9.tar.gz
Windows x86		jdk-8u201-windows-x86.exe
Windows x64		jdk-8u201-windows-x64.exe

You must accept the [Oracle Binary Code License Agreement for the Java SE Platform Products](#) to download this software.

☒ I reviewed and accept the Oracle Binary Code License Agreement for the Java SE Platform Products

You will be redirected to the login screen in order to download the file.

[Download jdk-8u201-windows-x64.exe](#)

Java SE Runtime Environment 8u201

This software is licensed under the [Oracle Binary Code License Agreement for Java SE Platform Products](#)

Product / File Description	File Size	Download
https://www.oracle.com/webapps/redirect/signon?nexturl=https://download...	68.1 MB	jre-8u201-linux-i586.rpm

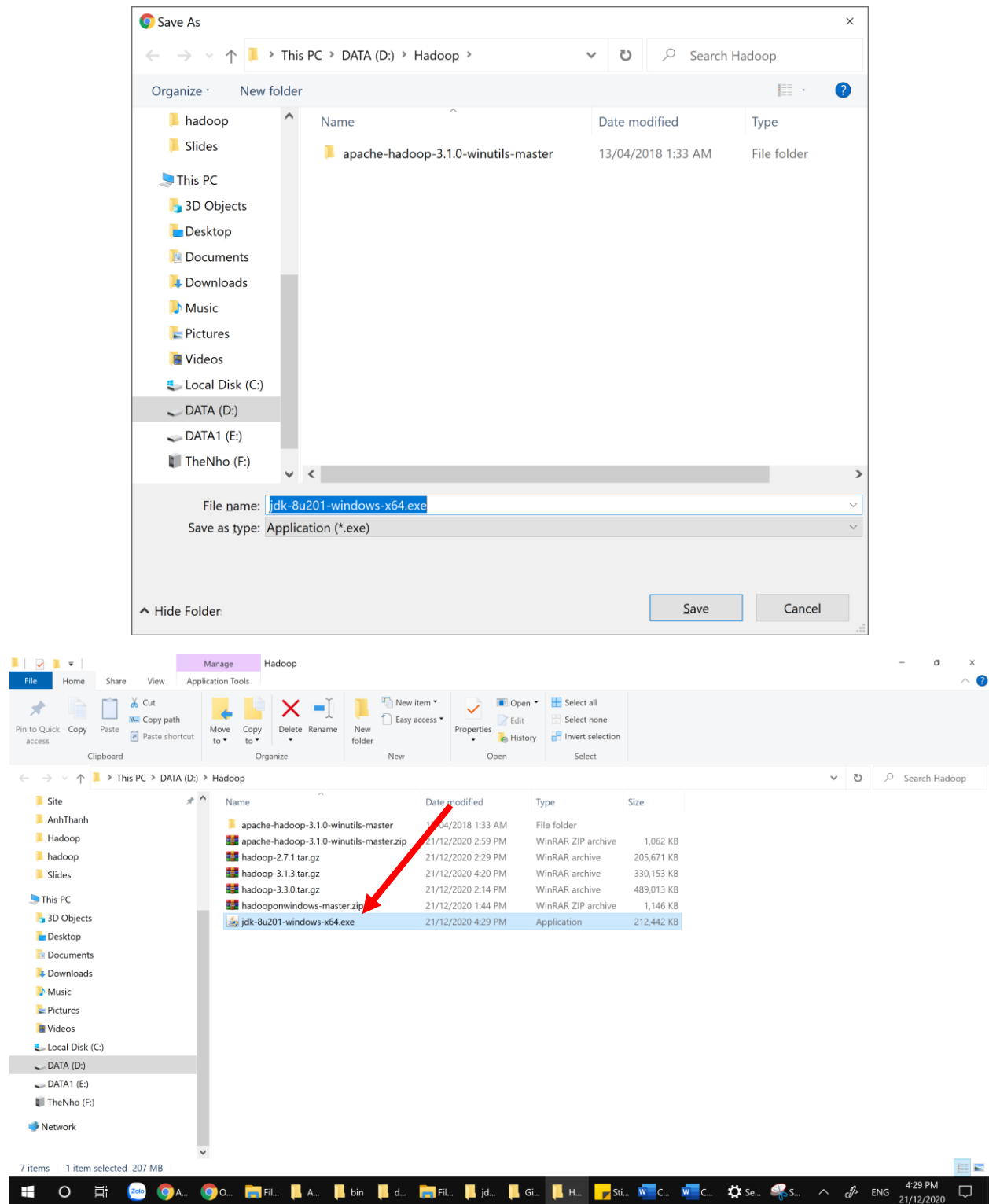
Tick vào “I reviewed and accept the Oracle...”

Rồi bấm download

Chương trình yêu cầu đăng nhập:

Nếu chưa có tài khoản thì cứ đăng ký

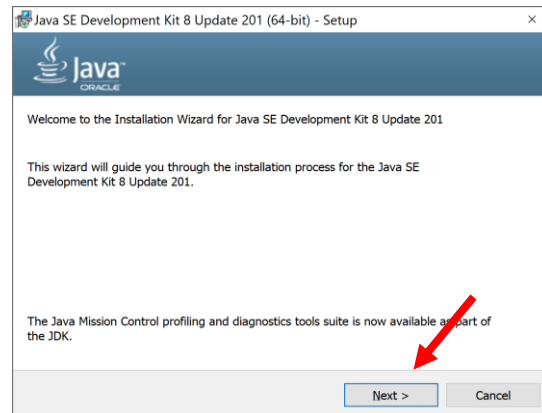
Khi đăng nhập thành công, Oracle



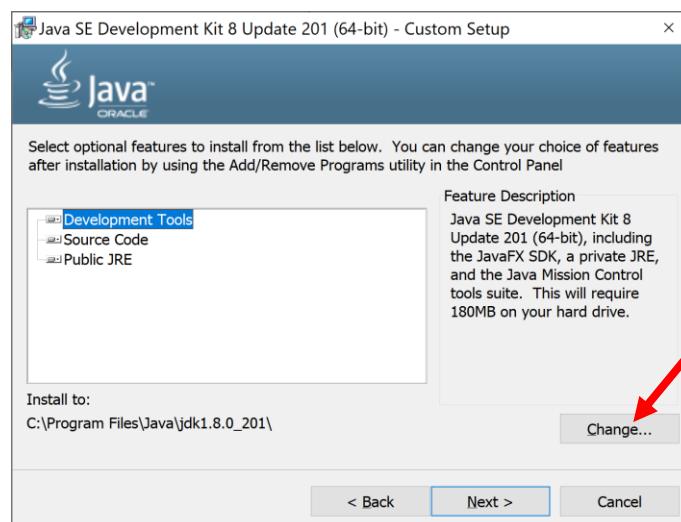
Tiến hành cài đặt:

Double click vào file vừa tải về:

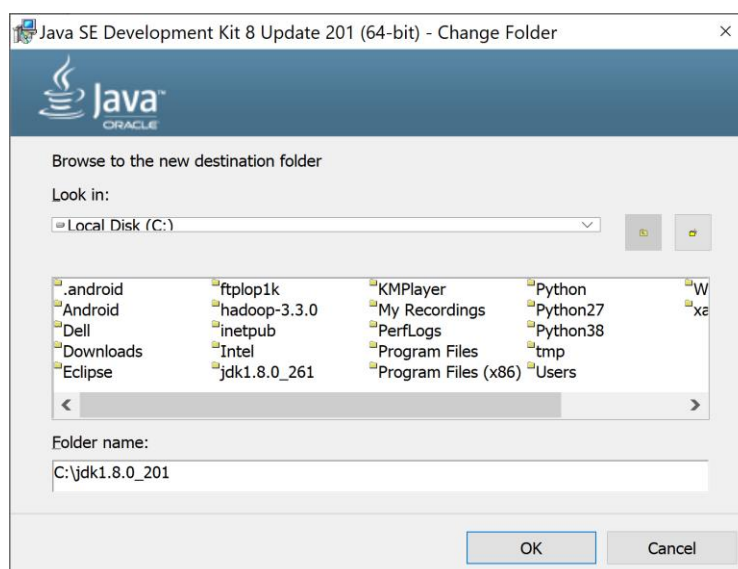
<https://duythanhcse.wordpress.com/big-data/>

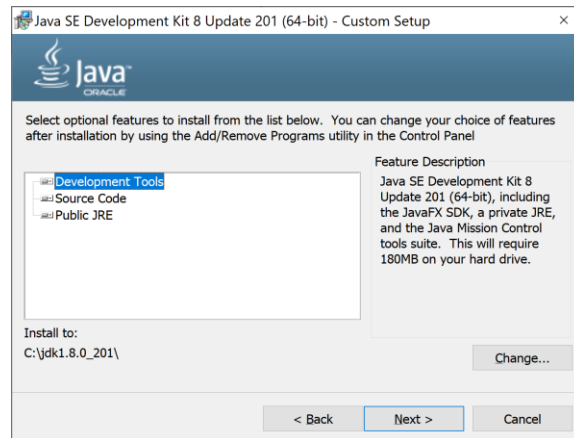


Nhấn Next để cài đặt

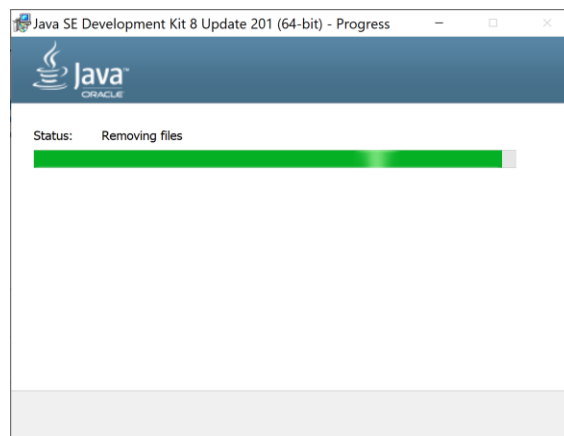


Tới chỗ này nhớ chỉnh vào Ổ C, không có dấu và không khoảng trắng

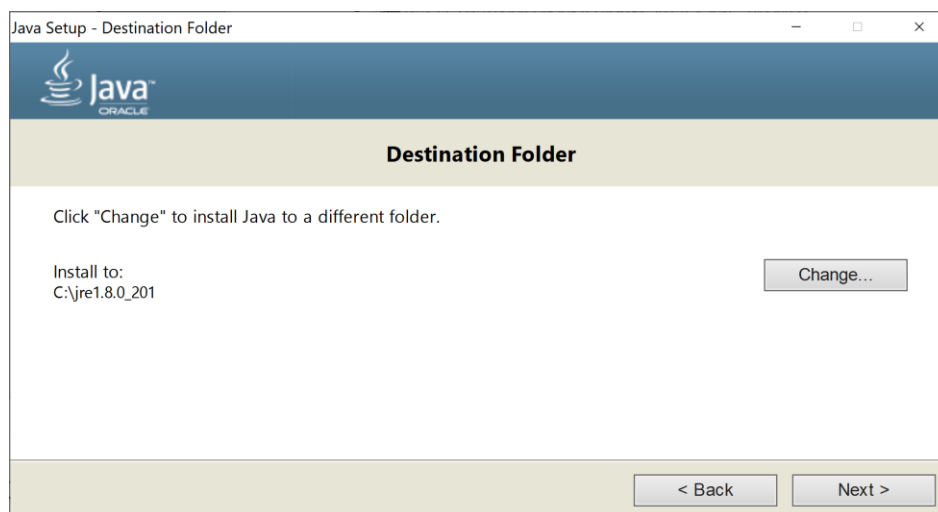




Bấm Next để tiếp tục



Chờ chương trình cài đặt hoàn tất



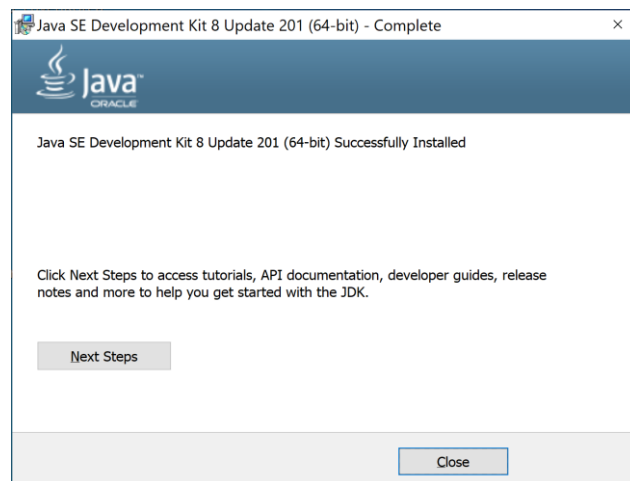
Nếu jre yêu cầu cài thì cũng chỉnh vào ổ C như trên

Bấm Next

Tiếp tục chờ



Khi xuất hiện màn hình dưới đây tức là đã hoàn tất quá trình cài đặt JDK

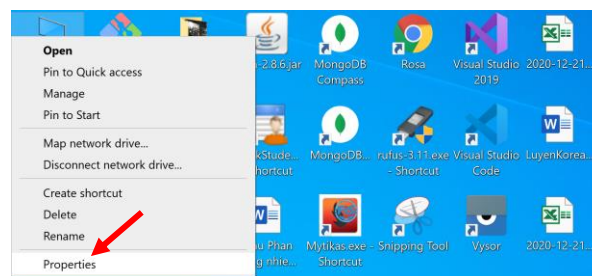


Bấm Close để hoàn tất

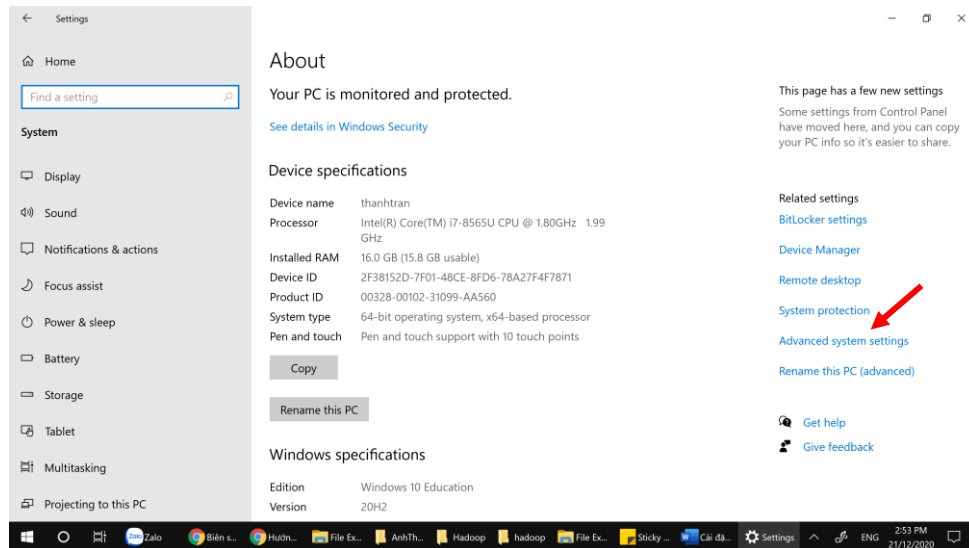
Như vậy đã cài đặt xong

3) Thiết lập biến môi trường cho Java JDK

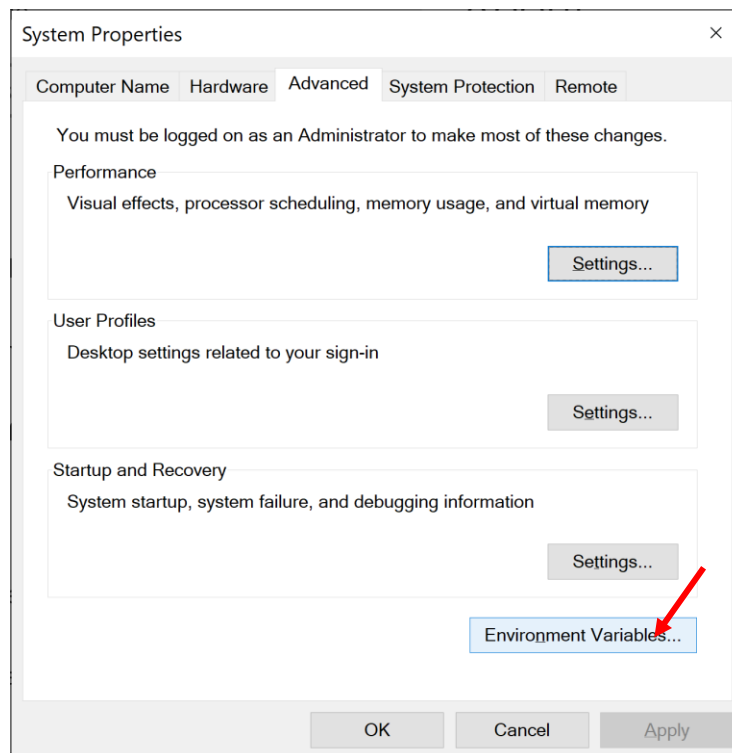
Cần cấu hình biến môi trường JAVA_HOME cho Java JDK



Bấm chuột phải vào Computer / chọn Properties

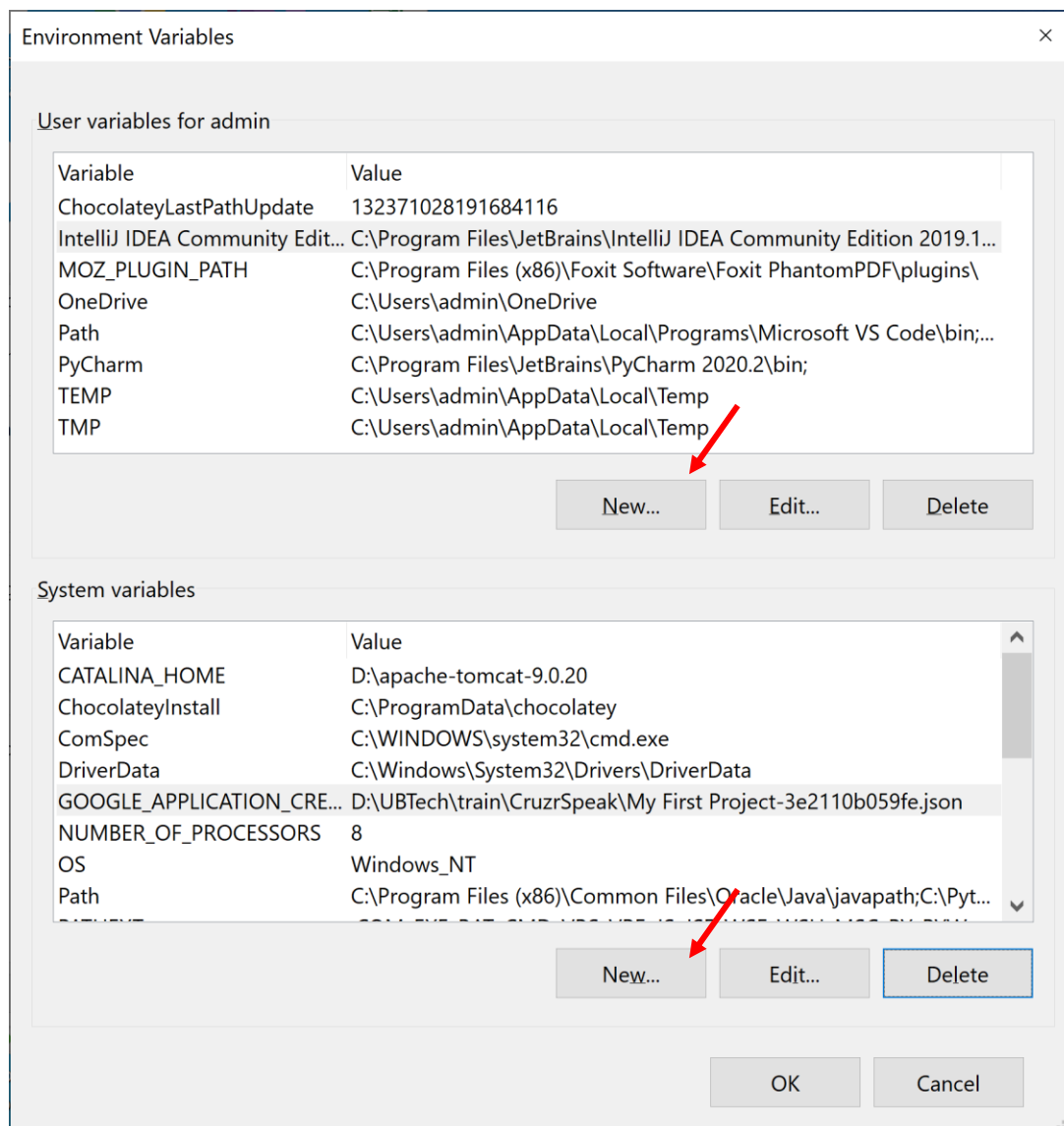


Chọn Advanced System Settings

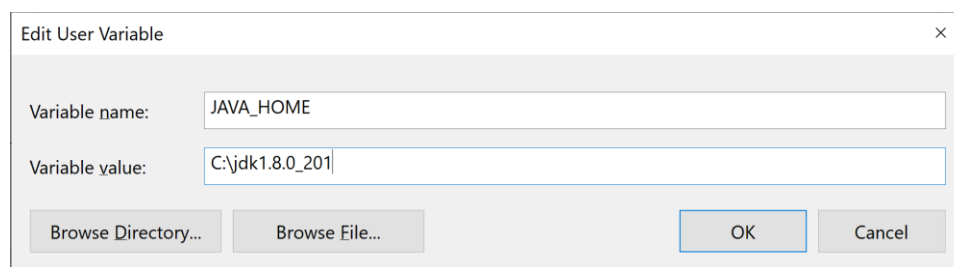


Chọn Environment Variables...

Màn hình Environment Variables sẽ xuất hiện như dưới đây:



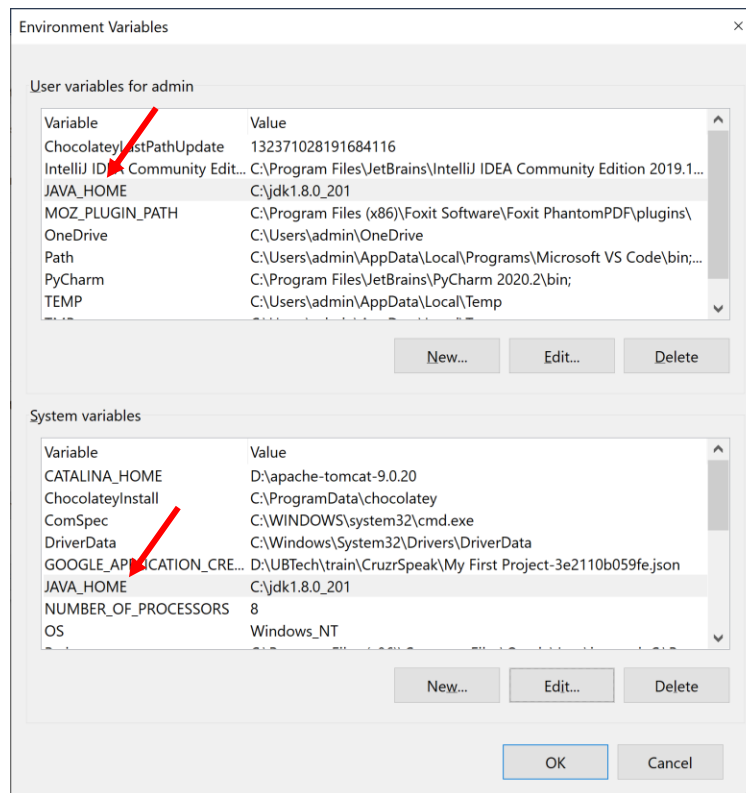
Trong mục user và system variables ta cấu hình JAVA_HOME trở tới nơi cài đặt JDK (bằng cách nhấn vào nút New...)



Variable name: JAVA_HOME

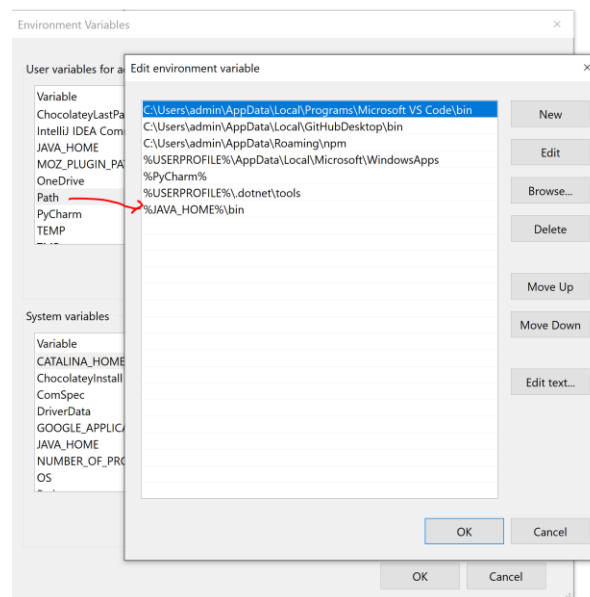
Variable value: C:\jdk1.8.0_201

Kết quả:



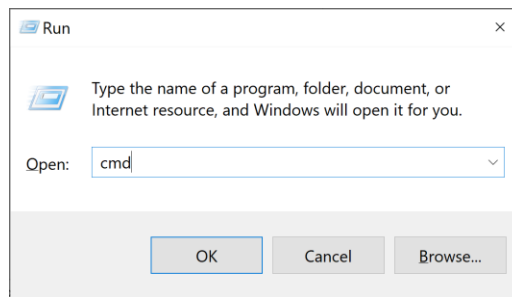
Sau đó bấm OK liên tục để đóng các cửa sổ cũng như xác nhận sự thay đổi

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



Thêm lệnh: %JAVA_HOME%\bin

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

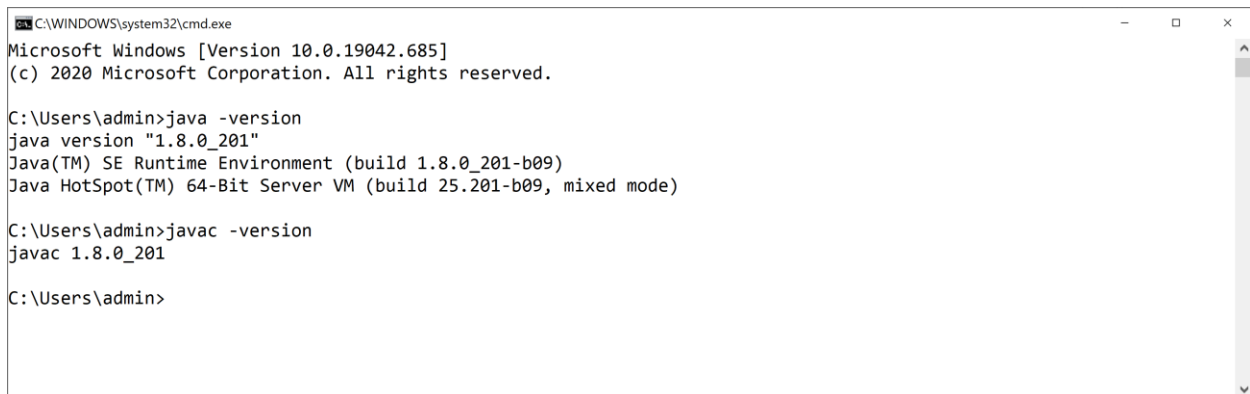


Trong màn hình command line lên gõ các lệnh trên để thấy kết quả:

```
java -version
```

```
javac -version
```

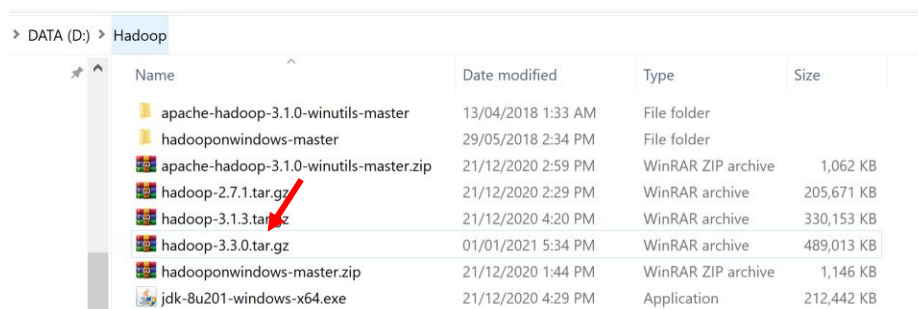
Kết quả:



4) Tải Hadoop và giải nén vào ổ C

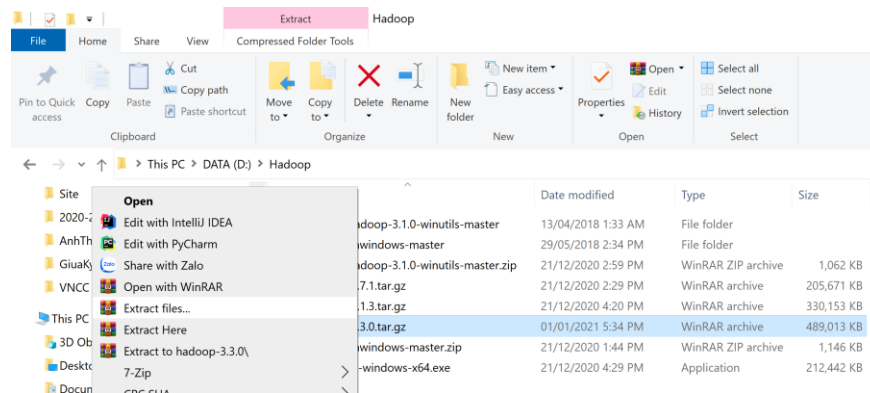
Vào link sau tải hadoop 3.3.0 về:

<https://mirror.downloadvn.com/apache/hadoop/common/hadoop-3.3.0/hadoop-3.3.0.tar.gz>

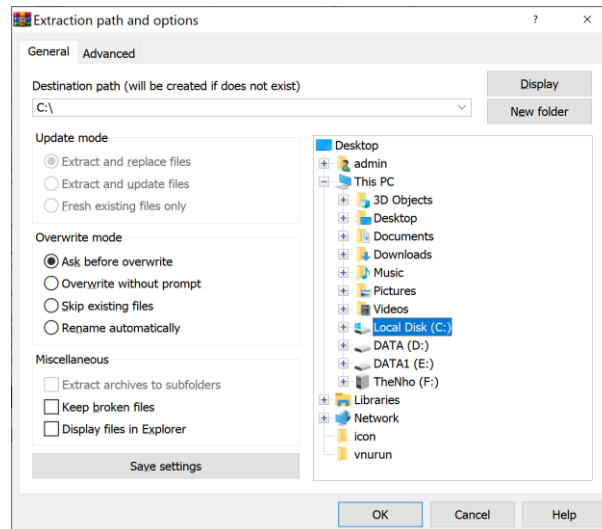


Giải nén vào ổ C

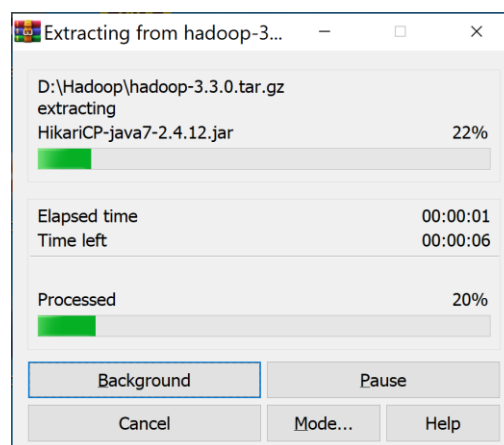
Bấm chuột phải vào “hadoop-3.3.0.tar.gz”



Chọn Extract files...



Chỉnh qua Ổ C rồi bấm OK:



Mở ổ C lên → thấy thư mục hadoop-3.3.0

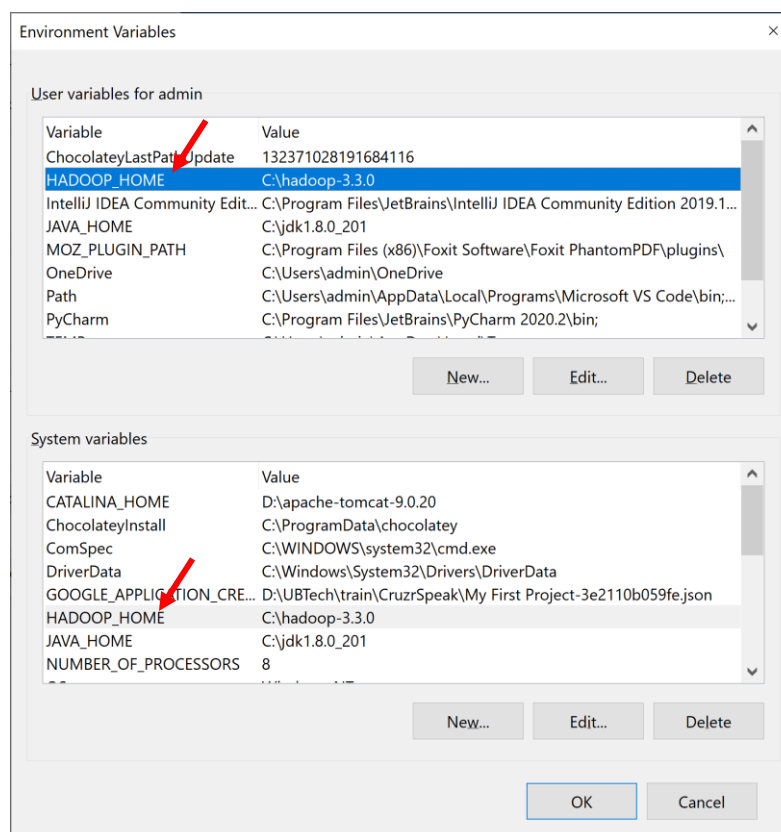
HK1

This PC > Local Disk (C:) > hadoop-3.3.0

Name	Date modified	Type	Size
bin	01/01/2021 5:43 PM	File folder	
etc	01/01/2021 5:44 PM	File folder	
include	01/01/2021 5:43 PM	File folder	
lib	01/01/2021 5:44 PM	File folder	
libexec	01/01/2021 5:43 PM	File folder	
licenses-binary	01/01/2021 5:43 PM	File folder	
sbin	01/01/2021 5:43 PM	File folder	
share	01/01/2021 5:43 PM	File folder	
LICENSE.txt	25/03/2020 12:23 AM	Text Document	16 KB
LICENSE-binary	05/07/2020 12:29 AM	File	23 KB
NOTICE.txt	25/03/2020 12:23 AM	Text Document	2 KB
NOTICE-binary	25/03/2020 12:23 AM	File	27 KB
README.txt	25/03/2020 12:23 AM	Text Document	1 KB

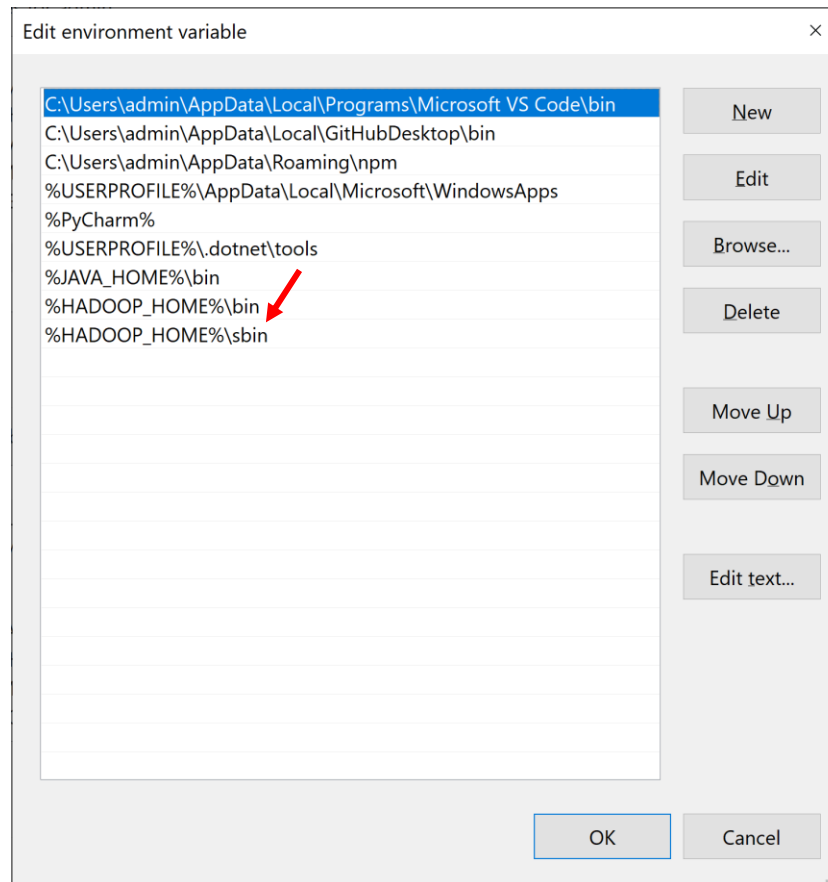
5) Thiết lập biến môi trường cho Hadoop

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 mà ta giải nén ở trên.

Sau đó chỉnh sửa biến path cho cả user và system variable. Bổ sung thêm:



```
%HADOOP_HOME%\bin
%HADOOP_HOME%\sbin
```

Mở CMD để test lại:

```
C:\WINDOWS\system32\cmd.exe

C:\Users\admin>hadoop version
Hadoop 3.3.0
Source code repository https://gitbox.apache.org/repos/asf/hadoop.git -r aa96f1871bfd858f9bac59cf2a81ec470da649af
Compiled by brahma on 2020-07-06T18:44Z
Compiled with protoc 3.7.1
From source with checksum 5dc29b802d6ccd77b262ef9d04d19c4
This command was run using /C:/hadoop-3.3.0/share/hadoop/common/hadoop-common-3.3.0.jar

C:\Users\admin>
```

Gõ lệnh: `hadoop version`

Ta thấy kết quả là hadoop có version 3.3.0, như vậy cấu hình biến môi trường đã xong.

6) Cấu hình các tập tin cho Hadoop

Trong thư mục C:/Hadoop-3.3.0/etc/hadoop lần lượt chỉnh sửa các file:

- core-site.xml
- mapred-site.xml
- hdfs-site.xml
- yarn-site.xml
- hadoop-env.cmd

Cấu hình core-site.xml như dưới đây:

```
<configuration>
  <property>
    <name>fs.defaultFS</name>
    <value>hdfs://localhost:9000</value>
  </property>
</configuration>
```

Cấu hình mapred-site.xml như dưới đây:

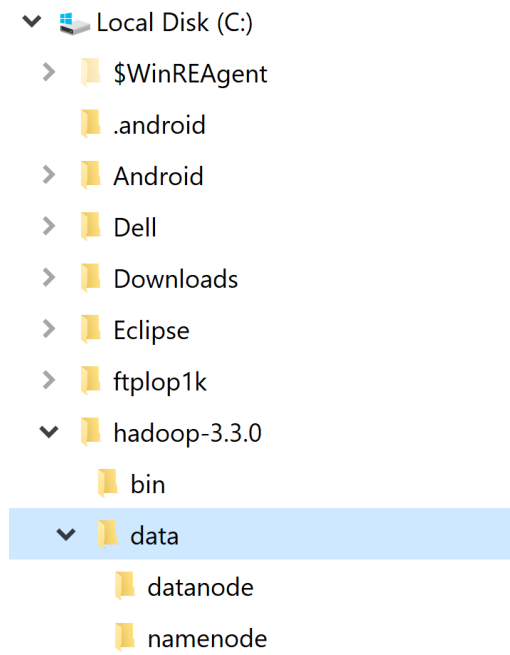
```
<configuration>
  <property>
    <name>mapreduce.framework.name</name>
    <value>yarn</value>
  </property>
</configuration>
```

Cấu hình hdfs-site.xml như dưới đây:

Tạo thư mục “data” trong “C:\Hadoop-3.3.0”

Tạo thư mục con “datanode” trong “C:\Hadoop-3.3.0\data”

Tạo thư mục con “namenode” trong “C:\Hadoop-3.3.0\data”



Sau đó cấu hình hdfs-site.xml như sau:

```
<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>/hadoop-3.3.0/data/namenode</value>
  </property>
  <property>
    <name>dfs.datanode.data.dir</name>
    <value>/hadoop-3.3.0/data/datanode</value>
  </property>
</configuration>
```

Cấu hình yarn-site.xml như dưới đây:

```
<configuration>
  <property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
  </property>
  <property>
    <name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</name>
    <value>org.apache.hadoop.mapred.ShuffleHandler</value>
  </property>
</configuration>
```

Cấu hình **hadoop-env.cmd**:

Mở file này lên và tìm tới lệnh:

```
set JAVA_HOME=%JAVA_HOME%
```

sửa thành:

```
set JAVA_HOME=C:\jdk1.8.0_201
```

7) Cập nhật các Hadoop Configurations

Tải <https://github.com/s911415/apache-hadoop-3.1.0-winutils>

Tải về giải nén ra thấy thư mục **bin** ở bên trong

Chép đè thư mục bin này trong thư mục bin của C:\hadoop-3.3.0\bin

Sau đó format lại **namenode** và **datanode**: mở command line lên, gõ lệnh sau:

```
hdfs namenode -format
```

```
hdfs datanode -format
```

```
C:\WINDOWS\system32\cmd.exe
Microsoft Windows [Version 10.0.19042.685]
(c) 2020 Microsoft Corporation. All rights reserved.

C:\Users\admin>hdfs namenode -format
2021-01-01 18:44:31,000 INFO namenode.NameNode: STARTUP_MSG:
/*****
STARTUP_MSG: Starting NameNode
STARTUP_MSG: host = thanhtran/172.20.224.1
STARTUP_MSG: args = [-format]
STARTUP_MSG: version = 3.3.0
STARTUP_MSG: classpath = C:\hadoop-3.3.0\etc\hadoop;C:\hadoop-3.3.0\share\hadoop\common;C:\hadoop-3.3.0\share\hadoo
p\common\lib\accessors-smart-1.2.jar;C:\hadoop-3.3.0\share\hadoop\common\lib\animal-sniffer-annotations-1.17.jar;C:\h
adoop-3.3.0\share\hadoop\common\lib\asm-5.0.4.jar;C:\hadoop-3.3.0\share\hadoop\common\lib\audience-annotations-0.5.0.
jar;C:\hadoop-3.3.0\share\hadoop\common\lib\avro-1.7.7.jar;C:\hadoop-3.3.0\share\hadoop\common\lib\checker-qual-2.5.2
.jar;C:\hadoop-3.3.0\share\hadoop\common\lib\commons-beanutils-1.9.4.jar;C:\hadoop-3.3.0\share\hadoop\common\lib\comm
ons-cli-1.2.jar;C:\hadoop-3.3.0\share\hadoop\common\lib\commons-codec-1.11.jar;C:\hadoop-3.3.0\share\hadoop\common\li
```

Bước format này chỉ cần làm 1 lần.

*Tiếp theo sao chép file:

“C:\hadoop-3.3.0\share\hadoop\yarn\timelineservice\hadoop-yarn-server-timelineservice-3.3.0.jar”

vào “C:\hadoop-3.3.0\share\hadoop\yarn\hadoop-yarn-server-timelineservice-3.3.0.jar”

8) Hoàn thành cài đặt Hadoop và test thử nghiệm với start-all.cmd

Để test Hadoop, ta mở command line và di chuyển tới thư mục

C:\hadoop-3.3.0\sbin

Sau đó gõ lệnh:

Start-all.cmd

Chi tiết xem hình các lệnh dưới đây:

```
C:\WINDOWS\system32\cmd.exe
Microsoft Windows [Version 10.0.19042.685]
(c) 2020 Microsoft Corporation. All rights reserved.

C:\Users\admin>cd\

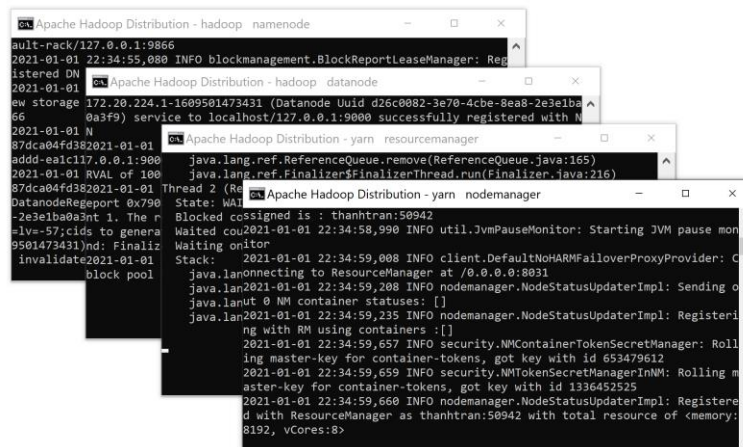
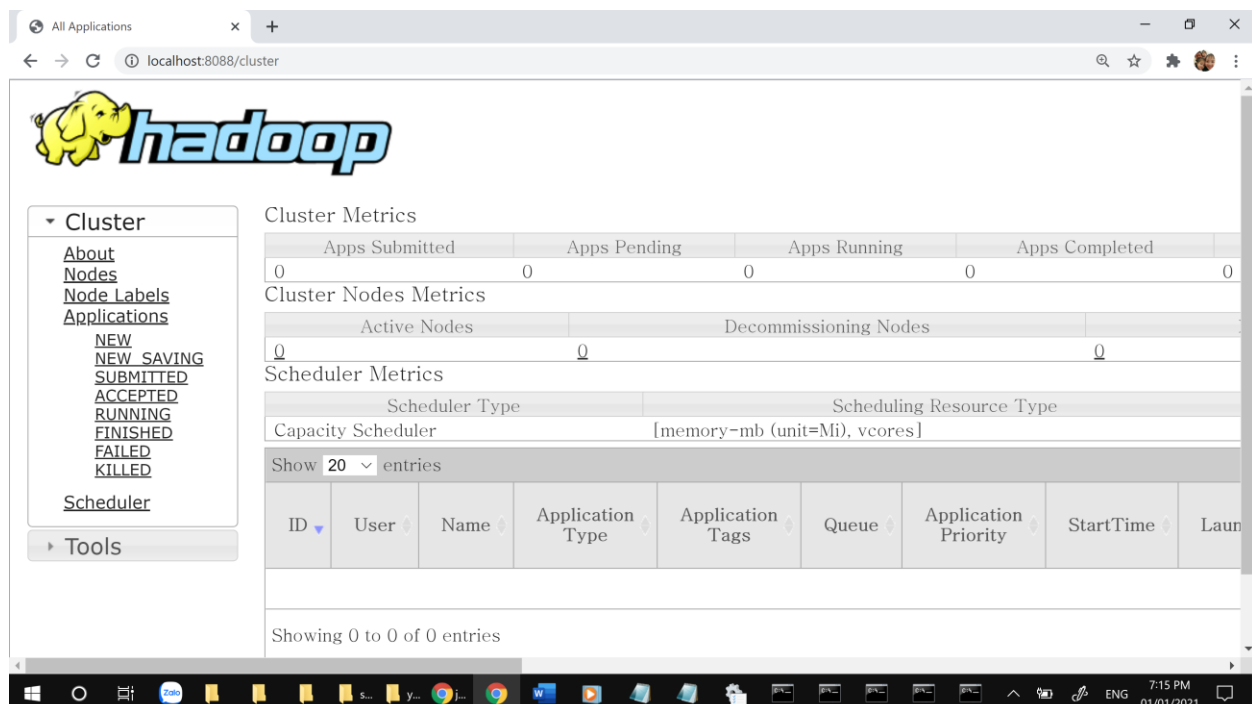
C:\>cd hadoop-3.3.0

C:\hadoop-3.3.0>cd sbin

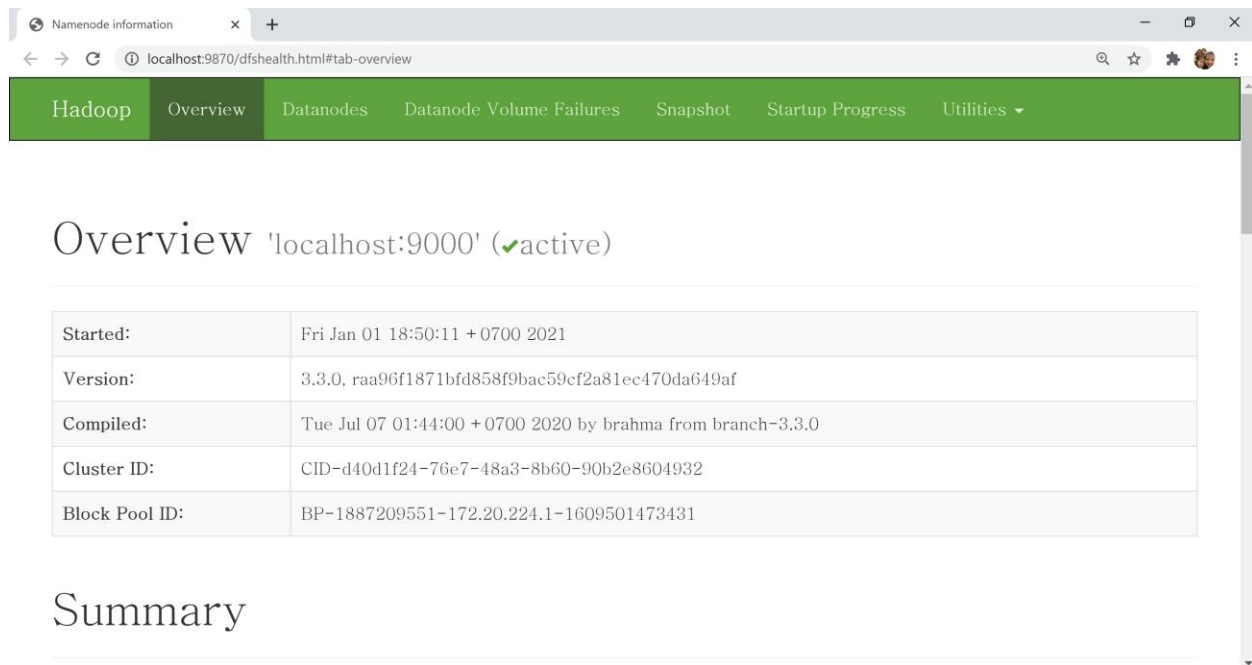
C:\hadoop-3.3.0\sbin>start-all.cmd
```

Sau khi gõ lệnh trên, hệ thống sẽ chạy Hadoop

- Hadoop Namenode
- Hadoop datanode
- YARN Resource Manager
- YARN Node Manager

<http://localhost:8088>

<http://localhost:9870>



The screenshot shows the Hadoop NameNode Overview page. The browser address bar indicates the URL is <http://localhost:9870/dfshealth.html#tab-overview>. The page has a green navigation bar with tabs: Hadoop, Overview (selected), Datanodes, Datanode Volume Failures, Snapshot, Startup Progress, and Utilities. The main content area is titled 'Overview 'localhost:9000' (✓active)'. Below the title is a table with the following information:

Started:	Fri Jan 01 18:50:11 +0700 2021
Version:	3.3.0, raa96f1871bfd858f9bac59cf2a81ec470da649af
Compiled:	Tue Jul 07 01:44:00 +0700 2020 by brahma from branch-3.3.0
Cluster ID:	CID-d40d1f24-76e7-48a3-8b60-90b2e8604932
Block Pool ID:	BP-1887209551-172.20.224.1-1609501473431

Below the table is a section titled 'Summary'.

Ngoài ra ta có thể tách chạy 2 lệnh:

– Khởi động namenode và datanode :

`start-dfs.cmd`

– Khởi động yarn bằng lệnh:

`start-yarn.cmd`