HƯỚNG DẪN CÀI ĐẶT HADOOP 2.9.1 TRÊN WINDOWS

1. DOWNLOAD HADOOP 2.9.1 VÀ JDK

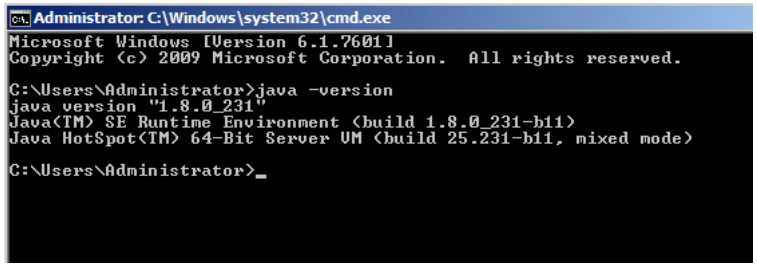
- Truy cập vào đường dẫn sau và download Hadoop 2.9.1 :

```
https://archive.apache.org/dist/hadoop/core/hadoop-2.9.1/hadoop-2.9.1.tar.gz
```

- Kiểm tra Java đã được cài đặt chưa?
  - *Mở cửa sổ **cmd** (command prompt của windows), chạy lệnh sau:*

```
java -version
```

  - *Nếu xuất hiện thông tin sau có nghĩa là máy đã có Java:*



- Nếu máy chưa cài đặt Java, download với từ khóa "***jdk***" từ internet như hình sau (chọn phiên bản 32 bit hoặc 64 bit tùy thuộc vào hệ điều hành đang dùng):

Java SE Development Kit 8u231

You must accept the Oracle Technology Network License Agreement for Oracle Java SE to download this software.

○ Accept License Agreement ◉ Decline License Agreement

Product / File Description	File Size	Download
Linux ARM 32 Hard Float ABI	72.9 MB	jdk-8u231-linux-arm32-vfp-hflt.tar.gz
Linux ARM 64 Hard Float ABI	69.8 MB	jdk-8u231-linux-arm64-vfp-hflt.tar.gz
Linux x86	170.93 MB	jdk-8u231-linux-i586.rpm
Linux x86	185.75 MB	jdk-8u231-linux-i586.tar.gz
Linux x64	170.32 MB	jdk-8u231-linux-x64.rpm
Linux x64	185.16 MB	jdk-8u231-linux-x64.tar.gz
Mac OS X x64	253.4 MB	jdk-8u231-macosx-x64.dmg
Solaris SPARC 64-bit (SVR4 package)	132.98 MB	jdk-8u231-solaris-sparcv9.tar.Z
Solaris SPARC 64-bit	94.16 MB	jdk-8u231-solaris-sparcv9.tar.gz
Solaris x64 (SVR4 package)	133.73 MB	jdk-8u231-solaris-x64.tar.Z
Solaris x64	91.96 MB	jdk-8u231-solaris-x64.tar.gz
Windows x86	200.22 MB	jdk-8u231-windows-i586.exe
Windows x64	210.18 MB	jdk-8u231-windows-x64.exe

Tiến hành cài đặt, đường dẫn sau khi cài đặt mặc định như sau:

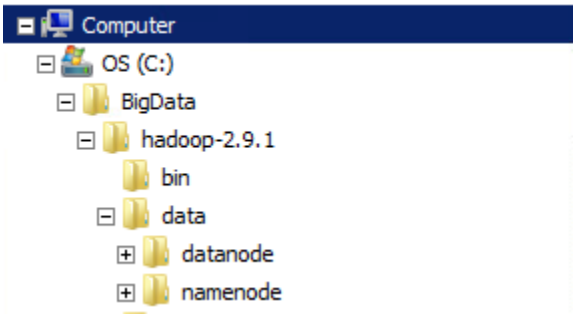
```
C:\Program Files\Java\jdk1.8.0_231
```

2. SETUP ĐƯỜNG DẪN VÀ CÀI ĐẶT BIẾN MÔI TRƯỜNG

- Giải nén file ***hadoop-2.9.1.tar.gz*** đã download tại đường dẫn:

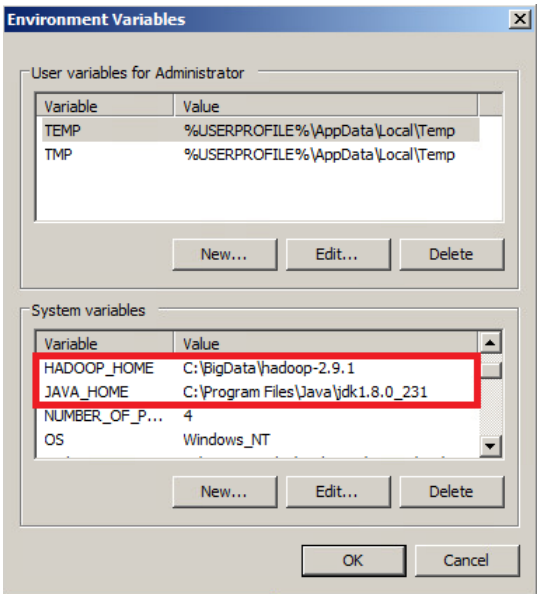
```
C:\BigData\hadoop-2.9.1
```

- Truy cập vào đường dẫn trên, tạo mới thư mục **data**, trong thư mục data, tạo tiếp 2 thư mục **datanode** và **namenode**.



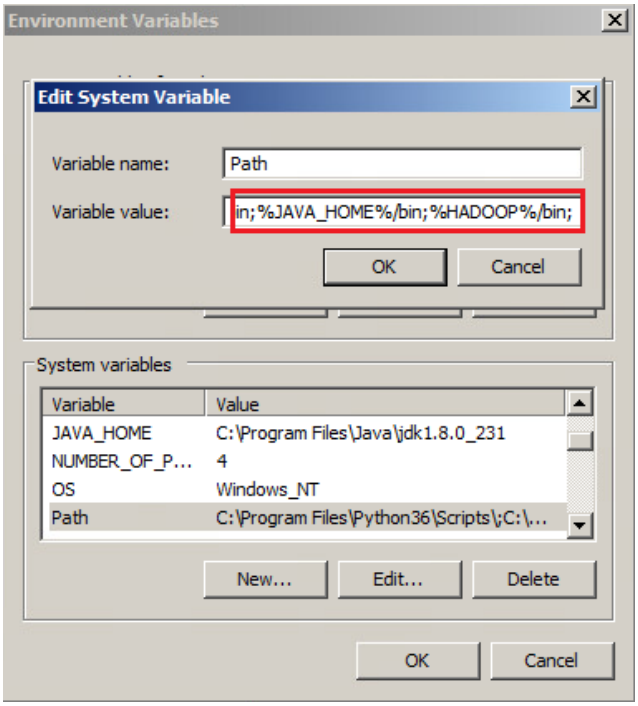
- Cài đặt biến môi trường:

```
HADOOP_HOME="C:\BigData\hadoop-2.9.1"
JAVA_HOME="C:\Program Files\Java\jdk1.8.0_231"
```



- Cập nhật thêm 2 thông tin vào biến môi trường PATH

```
%JAVA_HOME%/bin;%HADOOP_HOME%/bin;
```



3. CÀI ĐẶT HADOOP

Trong thư mục đã giải nén, cần cấu hình 4 file cần thiết sau:

- ✓ *hadoop-env.cmd*
- ✓ *core-site.xml*
- ✓ *hdfs-site.xml*
- ✓ *mapred-site.xml*

Tất cả 4 file này đều đặt trong đường dẫn:

```
C:\BigData\hadoop-2.9.1\etc\hadoop
```

- Cấu hình ***hadoop-env.cmd***: Khai báo set JAVA_HOME theo đường dẫn đã cài đặt ở trên

```
@echo off
@rem Licensed to the Apache Software Foundation (ASF) under one or more
@rem contributor license agreements.  See the NOTICE file distributed with
@rem this work for additional information regarding copyright ownership.
@rem The ASF licenses this file to You under the Apache License, Version 2.0
@rem (the "License"); you may not use this file except in compliance with
@rem the License.  You may obtain a copy of the License at
@rem
@rem     http://www.apache.org/licenses/LICENSE-2.0
@rem
@rem Unless required by applicable law or agreed to in writing, software
@rem distributed under the License is distributed on an "AS IS" BASIS,
@rem WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
@rem See the License for the specific language governing permissions and
@rem limitations under the License.

@rem Set Hadoop-specific environment variables here.

@rem The only required environment variable is JAVA_HOME.  All others are
@rem optional.  When running a distributed configuration it is best to
@rem set JAVA_HOME in this file, so that it is correctly defined on
@rem remote nodes.

@rem The java implementation to use.  Required.
set JAVA_HOME=C:\Progra~1\Java\jdk1.8.0_231

@rem The jsvc implementation to use. Jsvc is required to run secure datanodes.
@rem set JSVC_HOME=%JSVC_HOME%

@rem set HADOOP_CONF_DIR=
```

- Cấu hình ***core-site.xml***

```
<configuration>
   <property>
     <name>fs.default.name</name>
     <value>hdfs://<IP>:<PORT></value>
   </property>
</configuration>
```

IP: IP của máy đang dùng làm Hadoop Server

Port: Lựa chọn Port chưa sử dụng trên máy Server đang cấu hình để tránh tình trạng xung đột port.

- Cấu hình ***hdfs-site.xml***

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

```
        <name>dfs.namenode.name.dir</name>
        <value>C:\BigData\hadoop-2.9.1\data\namenode</value>
    </property>
    <property>
        <name>dfs.datanode.data.dir</name>
        <value>C:\BigData\hadoop-2.9.1\data\datanode</value>
    </property>
</configuration>
```

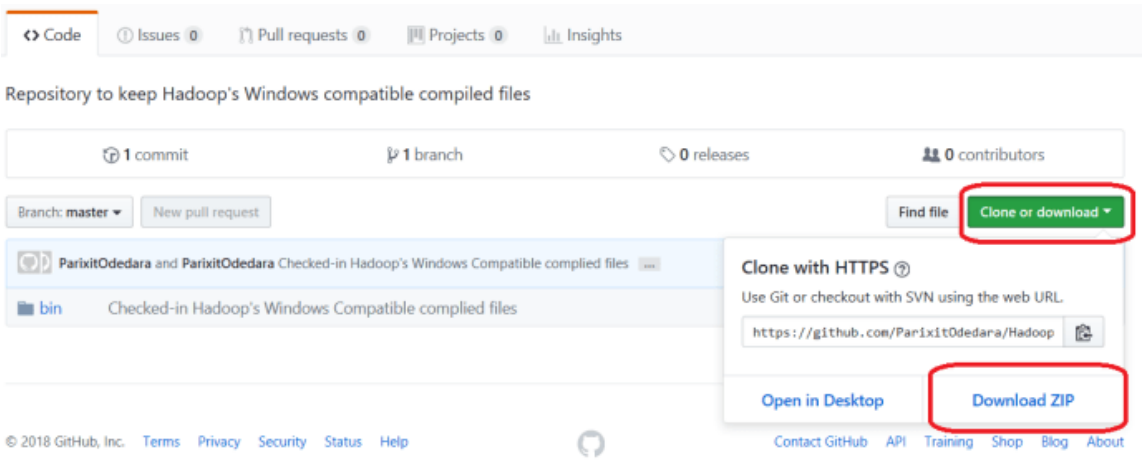
- Cấu hình ***mapred-site.xml***: Nếu tìm không thấy mapred-site.xml thì tìm file mapred-site.xml.template, sau đó đổi tên lại thành mapred-site.xml

```
<configuration>
    <property>
        <name>mapreduce.job.user.name</name>
        <value>%USERNAME%</value>
    </property>
    <property>
        <name>mapreduce.framework.name</name>
        <value>yarn</value>
    </property>
    <property>
        <name>yarn.apps.stagingDir</name>
        <value>/user/%USERNAME%/staging</value>
    </property>
    <property>
        <name>mapreduce.jobtracker.address</name>
        <value>local</value>
    </property>
</configuration>
```

4. FORMAT NAMENODE

- Truy cập vào link sau và download gói cấu hình:

```
https://github.com/ParixitOdedara/Hadoop
```

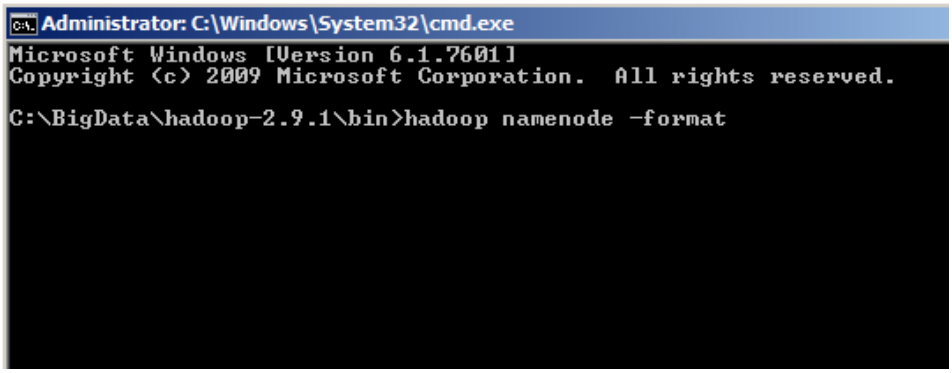


- Giải nén file vừa download về và copy toàn bộ nội dung bên trong thư mục bin, dán vào đường dẫn:

```
C:\BigData\hadoop-2.9.1\bin
```

- Mở cửa sổ cmd, truy cập đến đường dẫn C:\BigData\hadoop-2.9.1\bin, gõ lệnh:

```
hadoop namenode -format
```

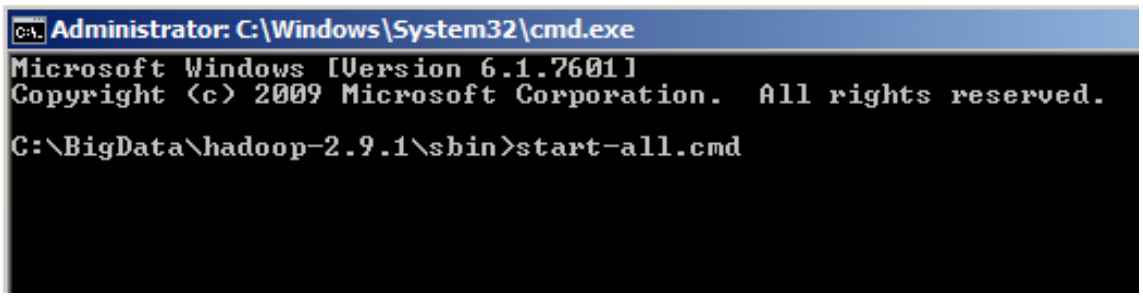


- Chờ quá trình format hoàn thành.

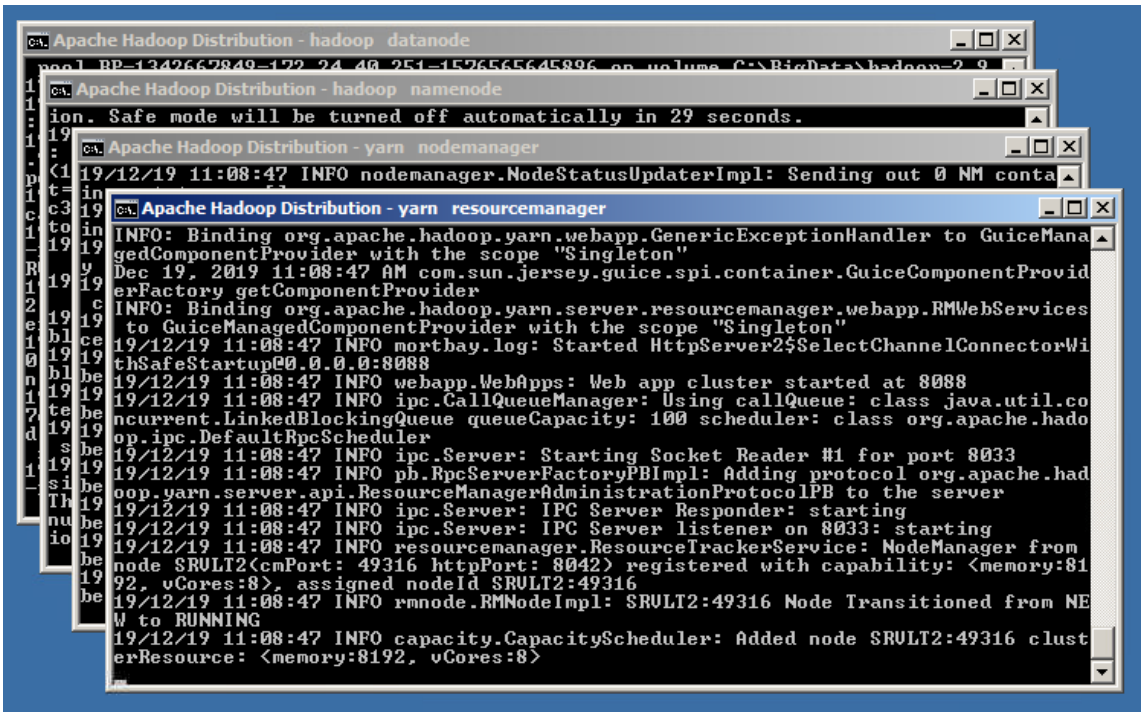
5. KHỞI ĐỘNG HADOOP

- Truy cập đến đường dẫn ***C:\BigData\hadoop-2.9.1\sbin*** *và gõ lệnh*:

```
start-all.cmd
```



- Xuất hiện 4 hộp thoại command sau là dịch vụ đã chạy thành công:



6. TRUY CẬP HADOOP SERVER BẰNG GIAO DIỆN GUI

- Từ máy client, chúng ta có thể truy cập đến máy Hadoop Server bằng trình duyệt theo cú pháp sau:

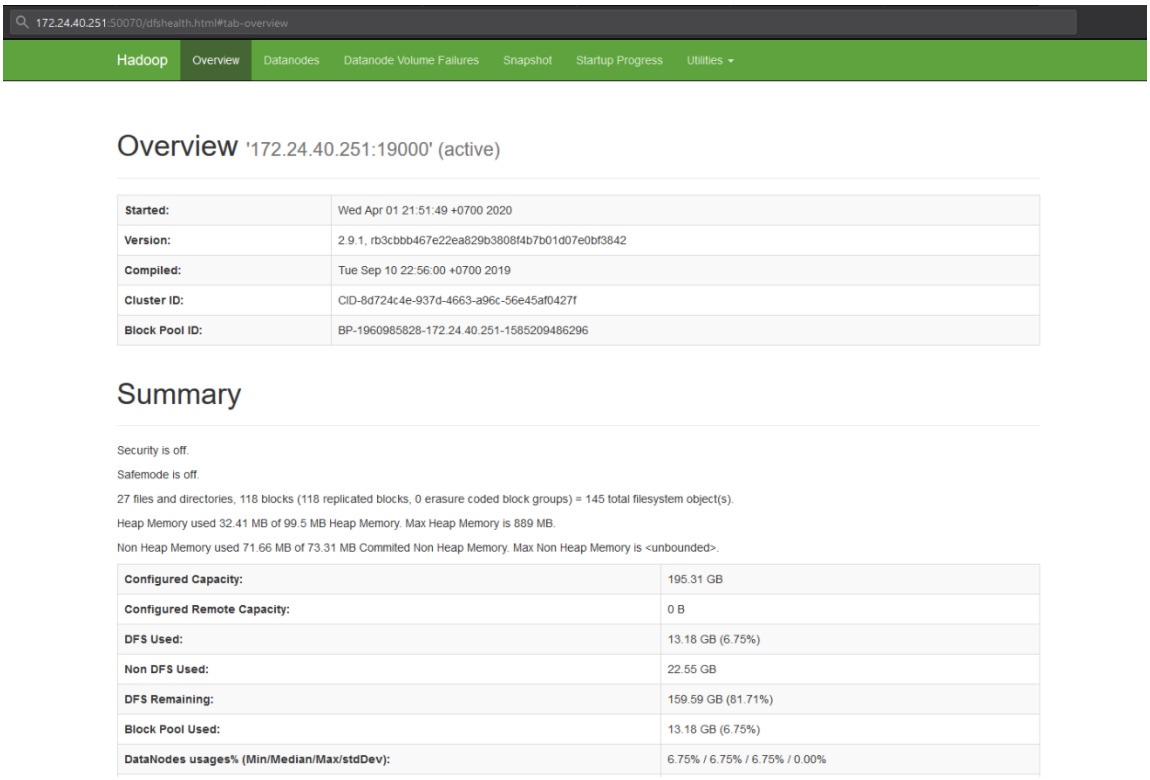
```
http://IP:50070
```

IP: IP của máy Hadoop Server.

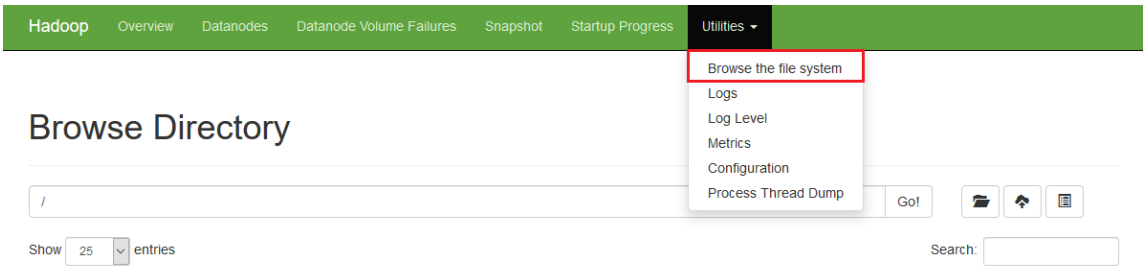
Port 50070: Là port mặc định của để truy cập đến giao diện GUI của Hadoop Server.

Tham khảo các port khác tại link: *https://hadoop.apache.org/docs/r2.9.1/hadoop-project-dist/hadoop-hdfs/hdfs-default.xml*

- Giao diện truy cập thành công sẽ có giao diện như sau:



- Truy cập đến trang upload file:

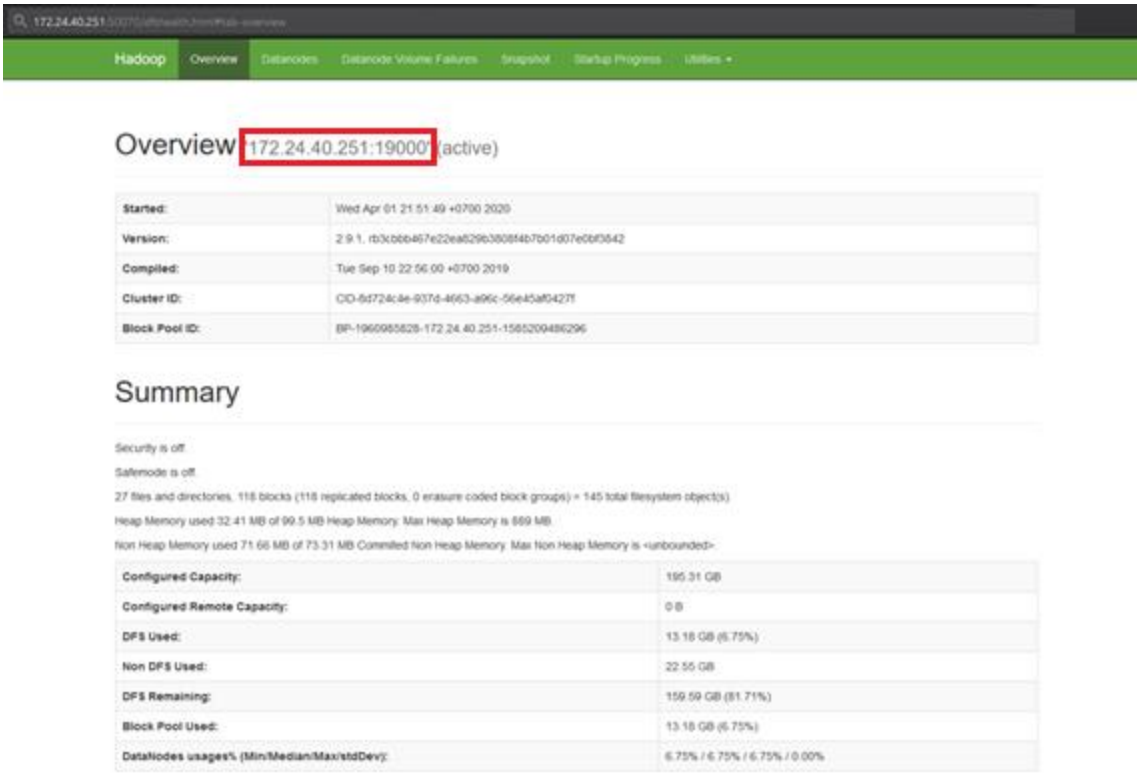


- Truy cập đến file từ tại Hadoop Server theo cú pháp:

```
hdfs://IP:PORT
```

IP: IP của máy Hadoop Server.

PORT: Port đã cấu hình trên file *core-site.xml*. Cũng có thể xem tại giao diện GUI tại trang chủ



--- Hết ---