



CSC17106 – XỬ LÝ PHÂN TÍCH DỮ LIỆU TRỰC TUYẾN

HƯỚNG DẪN THỰC HÀNH

HADOOP CƠ BẢN

I. Thông tin chung

Mã số:	HD03
Thời lượng dự kiến:	3 tiếng
Deadline nộp bài:	-
Hình thức:	-
Hình thức nộp bài:	-
GV phụ trách:	Phạm Minh Tú
Thông tin liên lạc với GV:	pmtu@fit.hcmus.edu.vn

II. Chuẩn đầu ra cần đạt

Bài hướng dẫn này nhằm mục tiêu đạt giúp sinh viên được các mục tiêu sau:

1. Cài đặt Hadoop trên môi trường window
2. Cấu hình Hadoop
3. Chạy Hadoop và thực hiện các lệnh cơ bản

III. Mô tả

Hadoop là một framework mã nguồn mở được thiết kế để xử lý và lưu trữ dữ liệu lớn trên các cụm máy tính phân tán. Nó ra đời từ dự án Apache Hadoop và đã trở thành một trong những công cụ quan trọng nhất trong lĩnh vực xử lý dữ liệu lớn (big data) và tích hợp dữ liệu.

Các bước cần chuẩn bị trước khi cài đặt hadoop trên Window:

Cài đặt Hadoop trên Windows có thể khá phức tạp do Hadoop ban đầu được phát triển cho môi trường Linux. Tuy nhiên, có một số cách để bạn có thể cài đặt Hadoop trên Windows.

Hadoop chạy trên nền tảng Java, vì vậy bạn cần cài đặt JDK. Hadoop thường tương thích với JDK 8. Hãy đảm bảo bạn đã cài đặt JDK và đã thiết lập biến môi trường **JAVA_HOME** để trỏ đến thư mục JDK

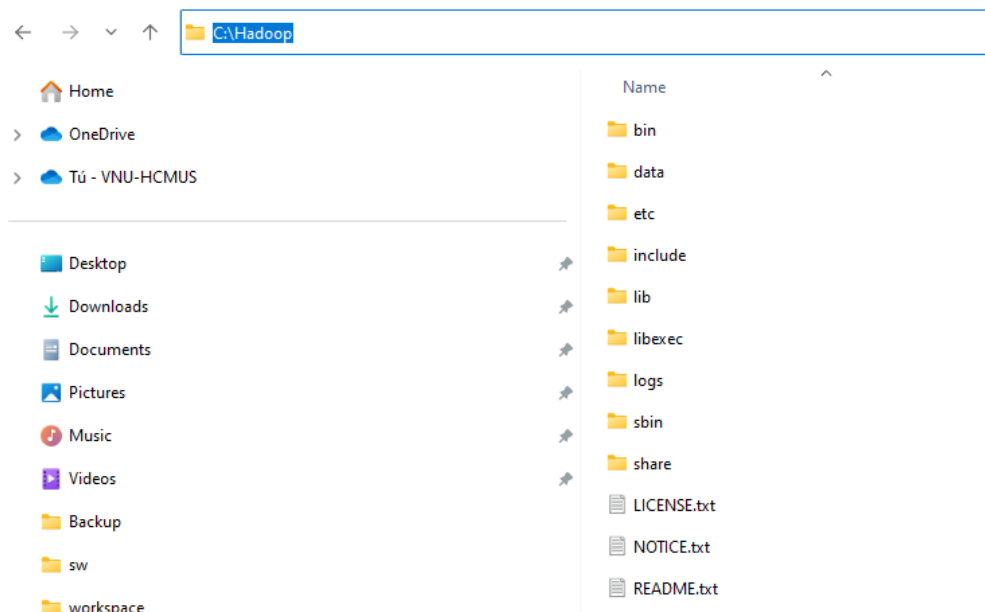
Tải xuống phiên bản Hadoop phù hợp với nhu cầu của bạn từ trang web chính thức của Apache Hadoop.

<https://www.apache.org/dyn/closer.cgi/hadoop/common/hadoop-3.2.4/hadoop-3.2.4.tar.gz>

Link trên tải hadoop phiên bản 3.2.4 tương thích Java 8.

Sau khi tải về, giải nén và lưu vào đường dẫn: C:\Hadoop

Cụ thể như hình bên dưới.



Bước tiếp theo cần cấu hình thông tin, bước này rất quan trọng vì cần cấu hình đúng thì hadoop mới chạy đúng.

Trước tiên vào đường dẫn sau: C:\Hadoop\etc\hadoop

File: core-site.xml, cấu hình như bên dưới.

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

File: mapred-site.xml

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

File: yarn-site.xml

```
<configuration>

<!-- Site specific YARN configuration properties -->
  <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>
```



```
</property>
```

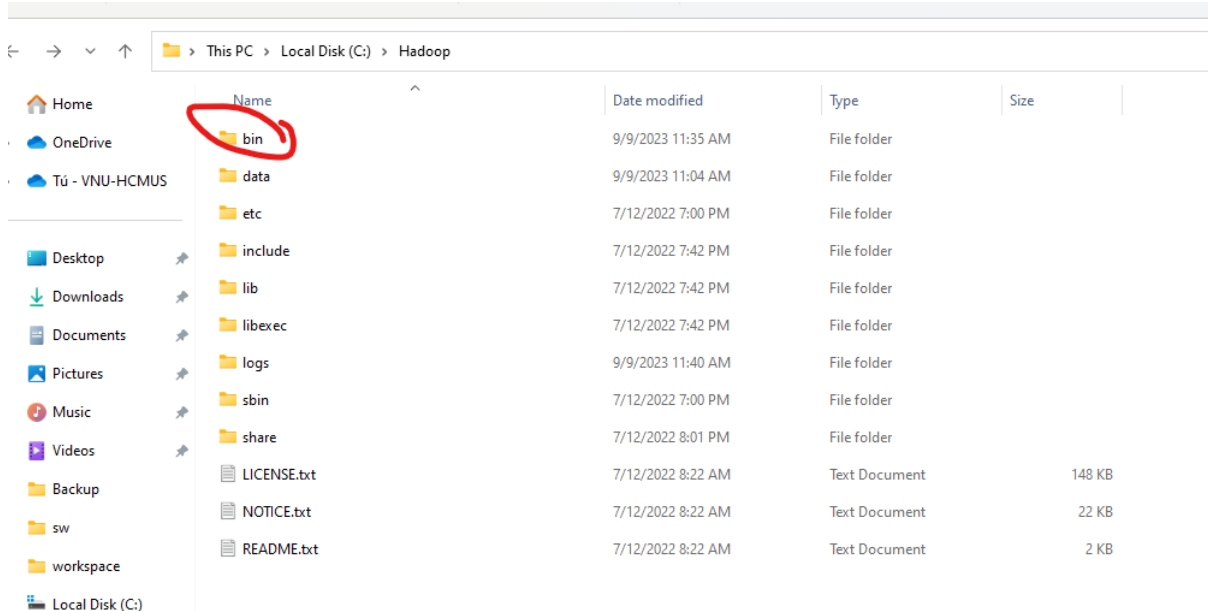
```
</configuration>
```

File: httpfs-site.xml

```
<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>C:\Hadoop\data\namenode</value>
  </property>
  <property>
    <name>dfs.datanode.data.dir</name>
    <value>C:\Hadoop\data\datanode</value>
  </property>
</configuration>
```

Lưu ý: Tự tạo các folder theo đường dẫn trên. VD: C:\Hadoop\data\namenode

Bước quan trọng cuối cùng, hãy để ý folder “bin” của Hadoop

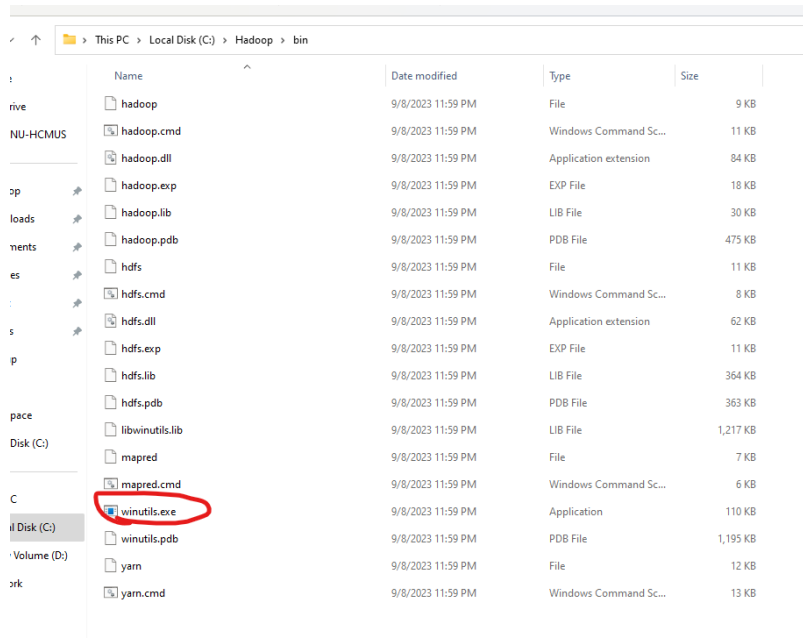


Nó là folder cài đặt dành cho Linux, chúng ta thay thế “bin” có hỗ trợ window. Hãy truy cập đường dẫn sau để tải về và copy override vào folder bin ban đầu.

<https://github.com/cdarlint/winutils/tree/master>

Tải bin [hadoop-3.3.5/bin](#)

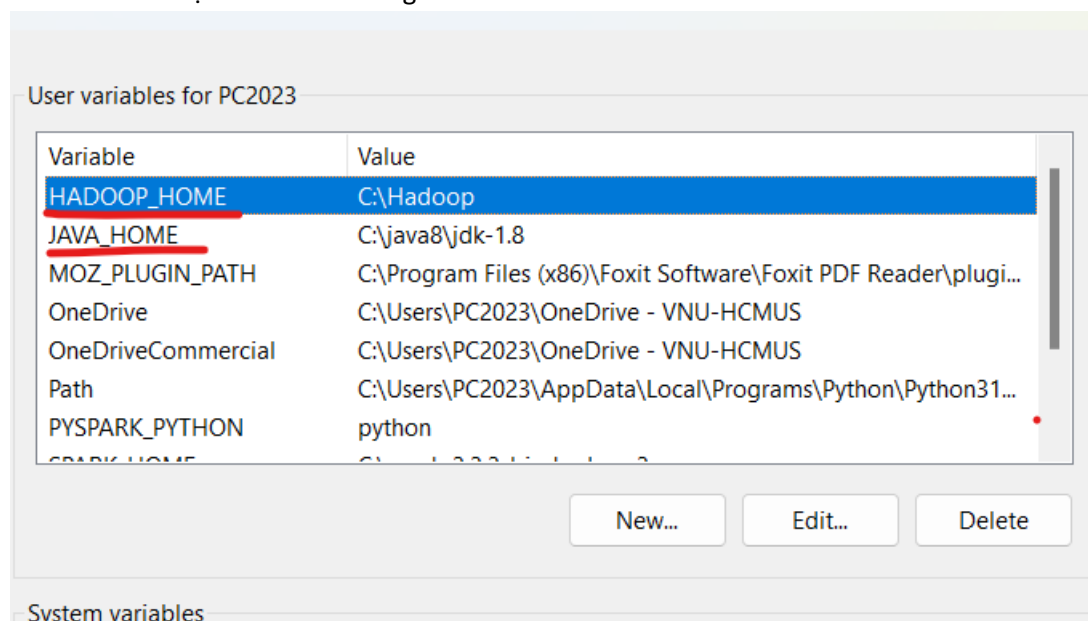
Sau khi override ta thấy có file winutils.exe (chính là file hỗ trợ chạy môi trường window)

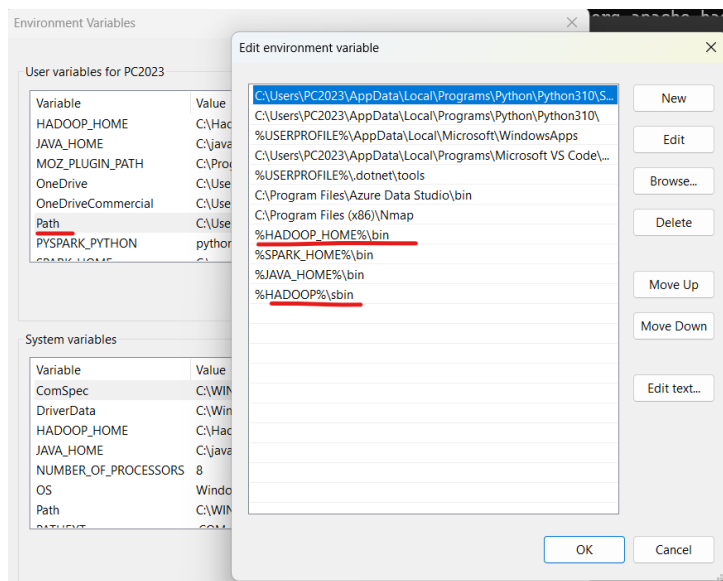


Name	Date modified	Type	Size
hadoop	9/8/2023 11:59 PM	File	9 KB
hadoop.cmd	9/8/2023 11:59 PM	Windows Command Sc...	11 KB
hadoop.dll	9/8/2023 11:59 PM	Application extension	84 KB
hadoop.exp	9/8/2023 11:59 PM	EXP File	18 KB
hadoop.lib	9/8/2023 11:59 PM	LIB File	30 KB
hadoop.pdb	9/8/2023 11:59 PM	PDB File	475 KB
hdfs	9/8/2023 11:59 PM	File	11 KB
hdfs.cmd	9/8/2023 11:59 PM	Windows Command Sc...	8 KB
hdfs.dll	9/8/2023 11:59 PM	Application extension	62 KB
hdfs.exp	9/8/2023 11:59 PM	EXP File	11 KB
hdfs.lib	9/8/2023 11:59 PM	LIB File	364 KB
hdfs.pdb	9/8/2023 11:59 PM	PDB File	363 KB
libwinutils.lib	9/8/2023 11:59 PM	LIB File	1,217 KB
mapred	9/8/2023 11:59 PM	File	7 KB
mapred.cmd	9/8/2023 11:59 PM	Windows Command Sc...	6 KB
winutils.exe	9/8/2023 11:59 PM	Application	110 KB
winutils.pdb	9/8/2023 11:59 PM	PDB File	1,195 KB
yarn	9/8/2023 11:59 PM	File	12 KB
yarn.cmd	9/8/2023 11:59 PM	Windows Command Sc...	13 KB

Còn 1 bước nữa: copy hadoop.dll vào thư mục Window/System32

Tiến hành cài đặt biến môi trường như sau:





Chúng ta đã hoàn tất bước cấu hình hadoop, tiếp theo khởi động hadoop bằng những câu lệnh cơ bản như sau:

Mở CMD và gõ các lệnh sau:

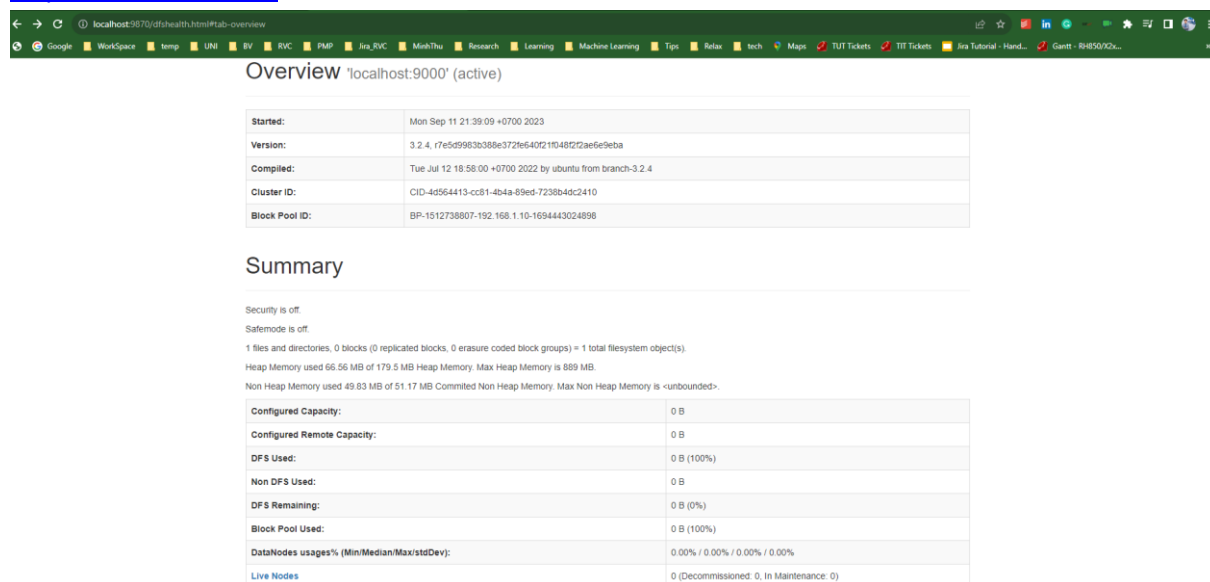
Lệnh 1: `hdfs namenode -format`

Lệnh 2: `start-dfs.cmd`

Lệnh 3: `start-yarn.cmd`

Mở trình duyệt để xem thông tin HDFS

<http://localhost:9870/>



Overview 'localhost:9000' (active)

Started:	Mon Sep 11 21:39:09 +0700 2023
Version:	3.2.4, r7e5d9983b388e372fe64021f048d2ca6e5e9ba
Compiled:	Tue Jul 12 18:58:00 +0700 2022 by ubuntu from branch-3.2.4
Cluster ID:	CID-4d564413-cc81-4b4a-89ed-7238b4dc2410
Block Pool ID:	BP-1512738807-192.168.1.10-1694443024898

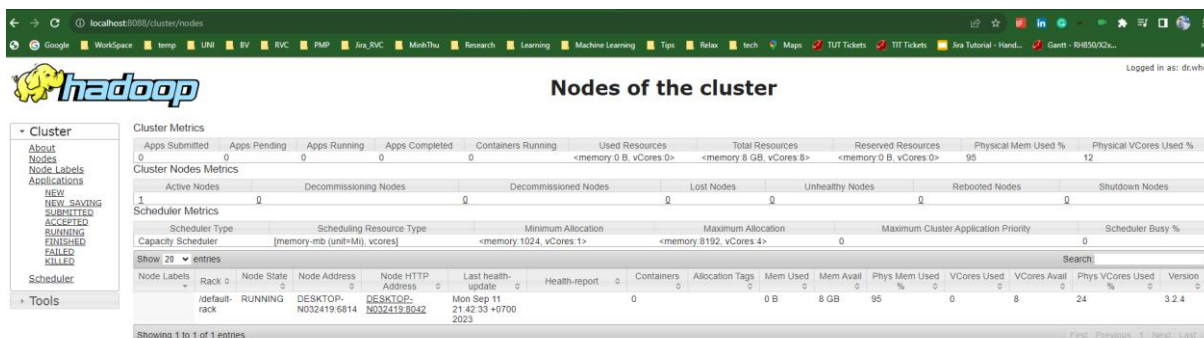
Summary

Security is off.
 Safemode is off.
 1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).
 Heap Memory used 66.56 MB of 179.5 MB Heap Memory. Max Heap Memory is 889 MB.
 Non Heap Memory used 49.83 MB of 51.17 MB Committed Non Heap Memory. Max Non Heap Memory is 'unbounded'.

Configured Capacity:	0 B
Configured Remote Capacity:	0 B
DFS Used:	0 B (100%)
Non DFS Used:	0 B
DFS Remaining:	0 B (0%)
Block Pool Used:	0 B (100%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	0 (Decommissioned: 0, In Maintenance: 0)

Xem thông tin các node

<http://localhost:8088/cluster/nodes>



The screenshot shows the Hadoop cluster metrics page. It includes a sidebar with navigation links like 'Cluster', 'Nodes', 'Node Labels', 'Applications', and 'Tools'. The main content area displays various metrics including 'Cluster Metrics', 'Cluster Nodes Metrics', and 'Scheduler Metrics'. A table at the bottom lists individual nodes with columns for Node Labels, Rack, Node State, Node Address, Node HTTP Address, Last health-update, Health-report, Containers, Allocation Tags, Mem Used, Mem Avail, Phys Mem Used, V-Cores Used, V-Cores Avail, Phys V-Cores Used, and Version.

Các lệnh cơ bản

Giới thiệu HDFS

HDFS (Hadoop Distributed File System) là một hệ thống lưu trữ phân tán được phát triển bởi dự án Apache Hadoop. Nó là một phần quan trọng của hệ thống Hadoop và được thiết kế để lưu trữ và quản lý dữ liệu lớn trên một cụm máy tính phân tán.

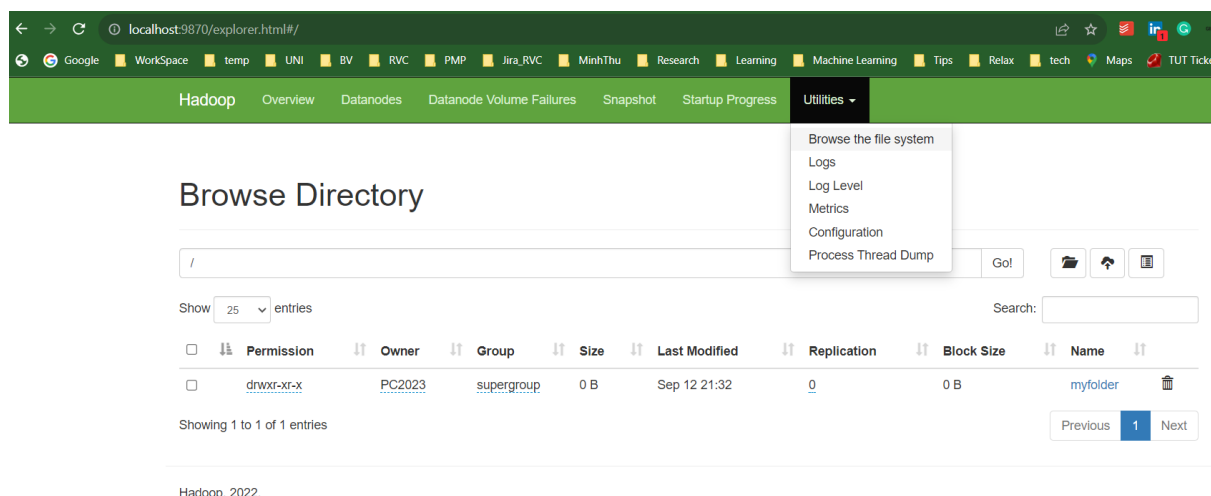
HDFS có một kiến trúc master-slave với hai thành phần chính:

- **NameNode:** Là máy chủ quản lý siêu dữ liệu (metadata) của hệ thống tệp và tên của tất cả các tệp và thư mục. NameNode duy trì danh sách các block của tất cả các tệp và quản lý vị trí của chúng trên các máy chủ DataNode.
- **DataNode:** Là các máy chủ lưu trữ thực tế cho các block dữ liệu. Chúng duy trì và quản lý các block dữ liệu và báo cáo trạng thái của chúng cho NameNode.

Tạo một file từ local system đến HDFS. Ví dụ ta có 1 file test.txt lưu tại local D:\test.txt, ta copy file này vào HDFS dùng lệnh sau:

Trước tiên tạo thư mục myfolder trên HDFS, mở CMD và thực hiện lệnh sau:

```
C:\Users\PC2023>hadoop fs -mkdir /myfolder
C:\Users\PC2023>
```



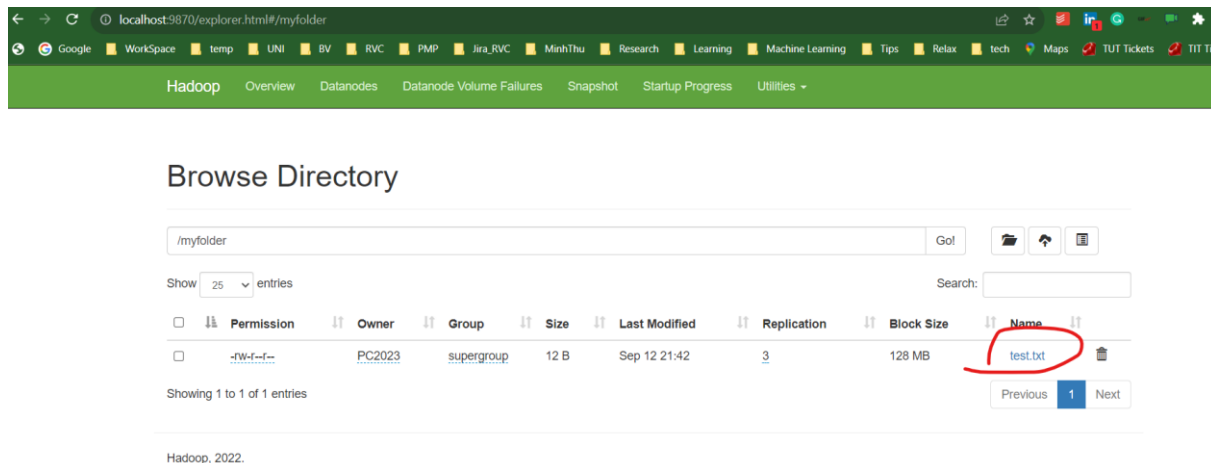
The screenshot shows the Hadoop web interface for browsing the file system. It includes a navigation bar with links like 'Hadoop', 'Overview', 'Datanodes', 'Datanode Volume Failures', 'Snapshot', 'Startup Progress', and 'Utilities'. The 'Browse Directory' section shows a table of files and directories. The table has columns for Permission, Owner, Group, Size, Last Modified, Replication, Block Size, and Name. A directory named 'myfolder' is listed with permissions 'drwxr-xr-x', owner 'PC2023', group 'supergroup', size '0 B', and last modified 'Sep 12 21:32'.

Sau khi tạo thư mục, chúng ta có thể copy file test.txt từ local đến hdfs dùng lệnh như sau:



```
C:\Users\PC2023>hadoop fs -copyFromLocal D:\test.txt /myfolder/test.txt  
C:\Users\PC2023>
```

Có thể xem lại thông tin trên trình duyệt Web



Xóa một file hoặc một folder trong hdfs

```
C:\Users\PC2023>hadoop fs -rm /myfolder/test.txt  
Deleted /myfolder/test.txt  
C:\Users\PC2023>
```

IV. Tài liệu tham khảo

1. <https://hadoop.apache.org/>
2. <https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/CommandsManual.html>
3. <https://community.cloudera.com/t5/Community-Articles/tkb-p/community-articles>
4. Youtube and google

V. Bài tập

Yêu cầu cơ bản:

Sinh viên hãy thực hiện lại bài hướng dẫn thực hành để cài đặt Hadoop và quản lý tài nguyên trên HDFS như tạo folder, tạo file, copy file, xóa file,..... có thể tìm hiểu thêm một số lệnh trong tài liệu tham khảo.

Yêu cầu nâng cao:

Bài toán: Giả sử bạn là quản trị hệ thống cho một công ty thương mại điện tử, cần lưu trữ và quản lý log truy cập website trên HDFS để phân tích sau này.

Yêu cầu:



1. Tạo folder /logs/2025/03 trên HDFS.
2. Upload các file log giả lập vào đó.
3. Kiểm tra dung lượng đã sử dụng trên HDFS.
4. Kiểm tra quyền truy cập của các file (xem ai có quyền đọc, ghi, thực thi).
5. Sử dụng lệnh du và dfsadmin -report để kiểm tra dung lượng từng thư mục và tình trạng của cụm Hadoop.
6. Giả lập tình huống có một file rất lớn (trên 1GB) và yêu cầu:
 - Chia file thành các phần nhỏ (split).
 - Tải từng phần lên HDFS và hợp nhất lại.
 - Kiểm tra sự phân phối các block của file trên các DataNode bằng lệnh hdfs fsck.
7. Giả sử một DataNode bị lỗi, hãy tìm hiểu cách Hadoop đảm bảo dữ liệu không bị mất?