

BỘ GIÁO DỤC & ĐÀO TẠO
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG-HCM
KHOA CÔNG NGHỆ THÔNG TIN



Hadoop

Môn học: Xử lý phân tích dữ liệu trực tuyến
Bộ môn: Hệ thống thông tin
Mã lớp: CQ2022/1
GVHD: Phạm Minh Tú
Họ và tên: Nguyễn Đăng Trí
MSSV: 22120383

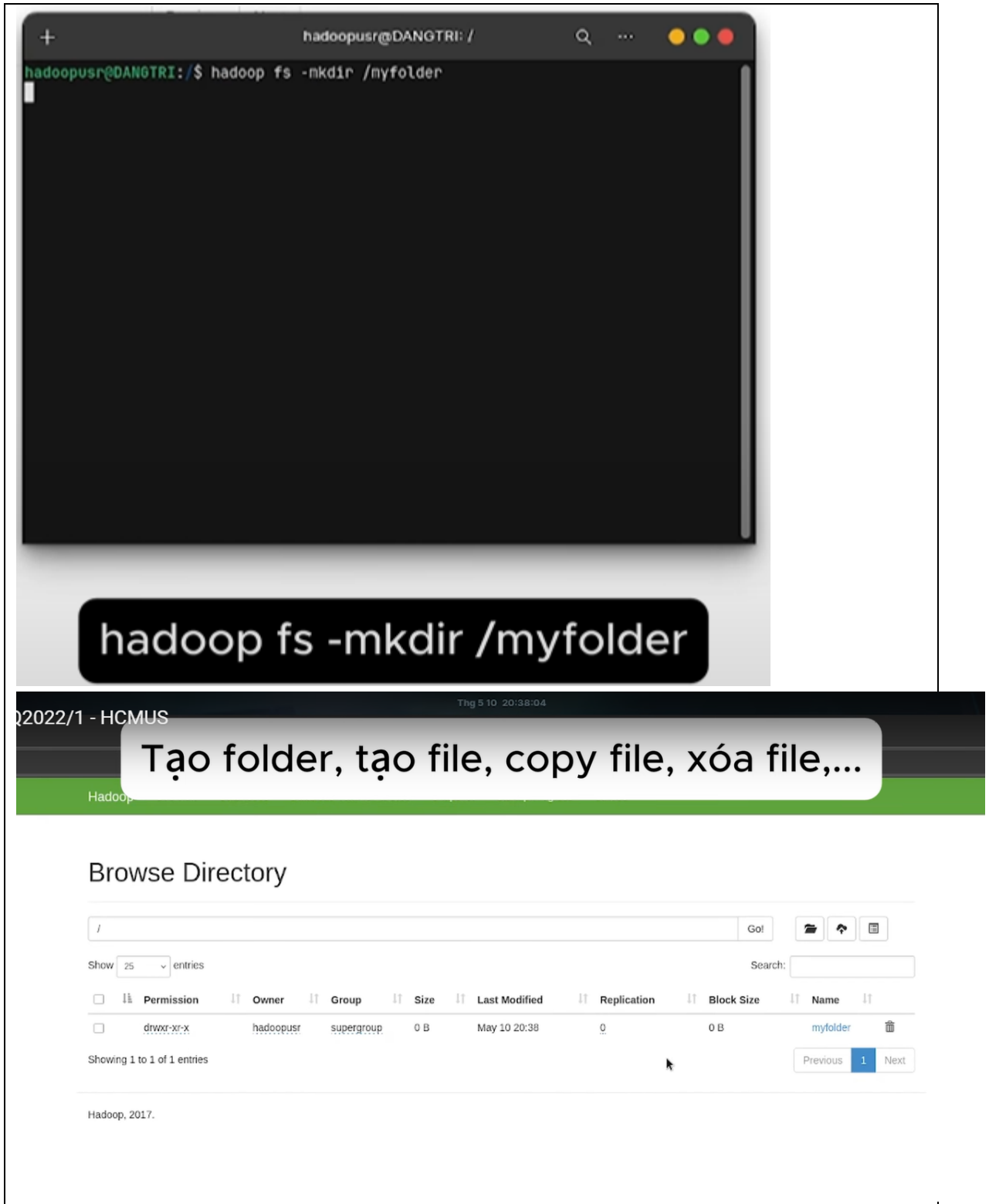
Thành phố Hồ Chí Minh, ngày 10 tháng 5 năm 2025

Mục lục

| | |
|---|---|
| 1. Yêu cầu cơ bản | 1 |
| 1.1. Tạo folder | 1 |
| 1.2. Copy file | 2 |
| 1.3. Xoá file | 3 |
| 2. Yêu cầu nâng cao..... | 4 |
| 2.1. Tạo folder /logs/2025/03 trên HDFS. | 4 |
| 2.2. Upload các file log giả lập vào đó..... | 4 |
| 2.3. Kiểm tra dung lượng đã sử dụng trên HDFS. | 4 |
| 2.4. Kiểm tra quyền truy cập của các file (xem ai có quyền đọc, ghi, thực thi). | 4 |
| 2.5. Sử dụng lệnh du và dfsadmin -report để kiểm tra dung lượng từng thư mục và tình trạng của cụm Hadoop. | 4 |
| 2.6. Giả lập tình huống có một file rất lớn (trên 1GB)..... | 5 |
| 2.6.1. Yêu cầu:..... | 5 |
| 2.6.2. Chia nhỏ file lớn (split): | 5 |
| 2.6.3. Upload từng phần lên HDFS | 5 |
| 2.6.4. Hợp nhất các phần lại trong HDFS | 5 |
| 2.6.5. Kiểm tra phân phối block của file | 5 |
| 2.7. Giả sử một DataNode bị lỗi, hãy tìm hiểu cách Hadoop đảm bảo dữ liệu không bị mất | 5 |
| 3. Video thực hành..... | 6 |

1. Yêu cầu cơ bản

1.1. Tạo folder

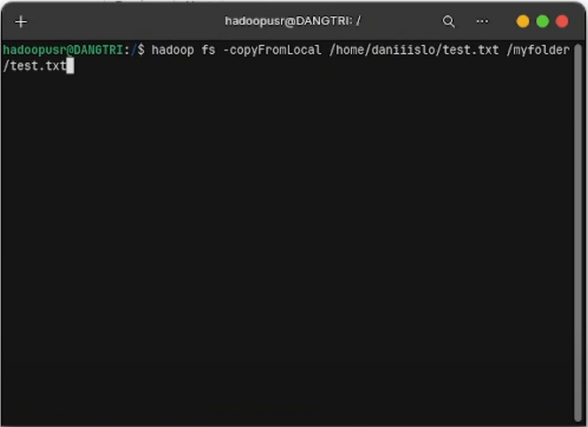


The image illustrates the process of creating a folder in Hadoop. It consists of two main parts: a terminal window and a web interface screenshot.

Terminal Window: The terminal shows the command `hadoop fs -mkdir /myfolder` being executed. Below the terminal, the command is repeated in a large, stylized font: `hadoop fs -mkdir /myfolder`.

Web Interface: The screenshot shows the Hadoop web interface. At the top, it says "2022/1 - HCMUS" and "Thg 5 10 20:38:04". Below this, there is a text box with the text "Tạo folder, tạo file, copy file, xóa file,...". The main section is titled "Browse Directory" and shows a table of files and folders. The table has columns for Permission, Owner, Group, Size, Last Modified, Replication, Block Size, and Name. The table shows one entry: a folder named "myfolder" with permissions "drwxr-xr-x", owner "hadoopusr", group "supergroup", size "0 B", last modified "May 10 20:38", and replication "0". The table also shows "Showing 1 to 1 of 1 entries".

1.2. Copy file



```
hadoopusr@DANGTRI: /  
hadoopusr@DANGTRI: $ hadoop fs -copyFromLocal /home/daniislo/test.txt /myfolder  
/test.txt
```




`hadoop fs -copyFromLocal /home/daniislo/test.txt /myfolder/test.txt`

/1 - HCMUS

Tạo folder, tạo file, copy file, xóa file,...

Hadoop

Browse Directory

/myfolder Go!   

Show 25 entries Search:

| <input type="checkbox"/> | Permission | Owner | Group | Size | Last Modified | Replication | Block Size | Name | <input type="checkbox"/> |
|--------------------------|------------|-----------|------------|------|---------------|-------------|------------|----------|--------------------------|
| <input type="checkbox"/> | -rw-r--r-- | hadoopusr | supergroup | 0 B | May 10 20:40 | 1 | 128 MB | test.txt | <input type="checkbox"/> |

Showing 1 to 1 of 1 entries Previous 1 Next

Hadoop, 2017.

1.3. Xóa file

```
hadoopusr@DANGTRI: /  
hadoopusr@DANGTRI:/$ hadoop fs -copyFromLocal /home/daniislo/test.txt /myfolder/  
test.txt  
WARNING: An illegal reflective access operation has occurred  
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.  
util.KerberosUtil (file:/usr/local/hadoop/share/hadoop/common/lib/hadoop-auth-2.  
9.0.jar) to method sun.security.krb5.Config.getInstance()  
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.  
security.authentication.util.KerberosUtil  
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflect  
ive access operations  
WARNING: All illegal access operations will be denied in a future release  
25/05/10 20:40:40 WARN util.NativeCodeLoader: Unable to load native-hadoop libra  
ry for your platform... using builtin-java classes where applicable  
hadoopusr@DANGTRI:/$ hadoop fs -rm /myfolder/test.txt
```

hadoop fs -m /myfolder/test.txt

Tạo folder, tạo file, copy file, xóa file,...

Browse Directory

Show 25 entries

Search:

| <input type="checkbox"/> | Permission | Owner | Group | Size | Last Modified | Replication | Block Size | Name |
|-----------------------------|------------|-------|-------|------|---------------|-------------|------------|-------------------------|
| No data available in table | | | | | | | | |
| Showing 0 to 0 of 0 entries | | | | | | | | |
| | | | | | | | | <div>PreviousNext</div> |

Hadoop, 2017.

2. Yêu cầu nâng cao

2.1. Tạo folder /logs/2025/03 trên HDFS.

```
hdfs dfs -mkdir -p /logs/2025/03
```

Tham số -p là viết tắt của parents. Khi sử dụng tham số này, HDFS sẽ tự động tạo các thư mục cha cần thiết nếu chúng chưa tồn tại.

2.2. Upload các file log giả lập vào đó.

```
hdfs dfs -put ~/fake-logs/* /logs/2025/03/
```

2.3. Kiểm tra dung lượng đã sử dụng trên HDFS.

```
hdfs dfs -du -h /logs/2025/03
```

Tham số -du là viết tắt của disk usage giúp hiển thị dung lượng sử dụng.

Tham số -h là viết tắt của human-readable, nghĩa là dung lượng theo định dạng dễ đọc hơn (KB, MB, GB...) thay vì số byte thuần túy.

2.4. Kiểm tra quyền truy cập của các file (xem ai có quyền đọc, ghi, thực thi).

```
hdfs dfs -ls /logs/2025/03
```

Lệnh này sẽ hiển thị quyền truy cập (rwx), chủ sở hữu (owner), nhóm (group).

Ví dụ đầu ra:

```
-rw-r--r--  3  hadoop  hadoopusr      1048576  2025-05-10  10:00  
/logs/2025/03/log1.txt
```

2.5. Sử dụng lệnh du và dfsadmin -report để kiểm tra dung lượng từng thư mục và tình trạng của cụm Hadoop.

Dùng lệnh du:

```
hdfs dfs -du -h /logs/2025
```

Dùng lệnh dfsadmin -report:

```
hdfs dfsadmin -report
```

Lệnh này cung cấp thông tin chi tiết về dung lượng còn trống, đã dùng, số lượng DataNode đang online/offline,...

2.6. Giả lập tình huống có một file rất lớn (trên 1GB)

2.6.1. 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`.

2.6.2. Chia nhỏ file lớn (split):

Giả sử file `big_log.txt` lớn hơn 1GB:

```
split -b 256M big_log.txt part_
```

File sẽ được chia thành các phần `part_aa`, `part_ab`,...

2.6.3. Upload từng phần lên HDFS

```
hdfs dfs -mkdir /logs/2025/03/bigfile_parts
hdfs dfs -put part_* /logs/2025/03/bigfile_parts/
```

2.6.4. Hợp nhất các phần lại trong HDFS

```
hdfs dfs -cat /logs/2025/03/bigfile_parts/part_* | hdfs dfs -put
- /logs/2025/03/big_log_combined.txt
```

2.6.5. Kiểm tra phân phối block của file

```
hdfs fsck /logs/2025/03/big_log_combined.txt -files -blocks -
locations
```

Lệnh trên hiển thị các block của file và DataNode tương ứng đang lưu trữ block đó.

2.7. Giả sử một DataNode bị lỗi, hãy tìm hiểu cách Hadoop đảm bảo dữ liệu không bị mất

Nếu một DataNode bị lỗi, Hadoop đảm bảo an toàn dữ liệu như thế nào?

Hadoop sử dụng cơ chế replication: Mỗi block của file sẽ được sao chép lên n bản trên các DataNode khác nhau.

Khi một DataNode bị lỗi:

- NameNode phát hiện lỗi qua việc không nhận heartbeat.
- Hadoop sẽ tự động replicate các block bị mất sang DataNode còn sống để đảm bảo số lượng bản sao luôn đủ.
- Dữ liệu không bị mất miễn là có ít nhất 1 bản sao còn tồn tại.

Kiểm tra replication của file bằng:

```
hdfs fsck /logs/2025/03/big_log_combined.txt -files -blocks -locations
```

Kết quả:

```
Status: HEALTHY
Total size:      1200000000 B
Total dirs:      0
Total files:     1
Total symlinks:   0
Total blocks (validated): 9 (avg. block size 133333333 B)
Minimally replicated blocks: 9 (100.0 %)
Over-replicated blocks: 0 (0.0 %)
Under-replicated blocks: 0 (0.0 %)
Mis-replicated blocks: 0 (0.0 %)
Default replication factor: 1
Average block replication: 1.0
Corrupt blocks: 0
Missing replicas: 0 (0.0 %)
Number of data-nodes: 1
Number of racks: 1
FSCK ended at Sat May 10 21:26:00 ICT 2025 in 1 milliseconds
```

Ở đây, file `big_log_file_combined.txt` chỉ có 1 bản sao (Default replication factor: 1), do đó nếu DataNode duy nhất bị lỗi, dữ liệu sẽ bị mất. Để tăng số replication factor, ta chỉnh cấu hình HDFS (`hdfs-site.xml`)

```
<property>
  <name>dfs.replication</name>
  <value>3</value>
</property>
```

Hoặc có thể thay đổi trên 1 file cụ thể bằng lệnh

```
hdfs dfs -setrep -R 3 <file_path>
```

3. Video thực hành

[Quá trình thực hành phân bài tập](#)