Nội dung thực hành

[I. Thao tác với HDFS 1](#_Toc9600669)

[Bài 1: Liệt kê thư mục trên HDFS 1](#_Toc9600670)

[Bài 2: Đẩy file vào HDFS 2](#_Toc9600671)

[Bài 3: Xem và xóa file trên HDFS 3](#_Toc9600672)

[Bài 4: Liệt kê thư mục trong HDFS sử dụng hdfs dfs 4](#_Toc9600673)

[Bài 5: Đẩy file lên HDFS sử dụng hdfs dfs 5](#_Toc9600674)

[Bài 6: Xem và xóa file trên HDFS sử dụng hdfs dfs 6](#_Toc9600675)

[Bài 7: Sử dụng HUE để thao tác file trên HDFS 6](#_Toc9600676)

[II. Quản trị HDFS. 6](#_Toc9600677)

# **I. Thao tác với HDFS**

Bài tập này giúp bạn làm quen với một số lệnh HDFS như:

- Liệt kê thư mục trong HDFS.

- Đẩy file vào HDFS.

- Xem và xóa file trên HDFS

Chúng ta có thể thao tác với HDFS trong Hadoop bằng cách sử dụng *hadoop fs* hoặc *hdfs dfs*

*hadoop fs* được thiết kế để làm việc với nhiều loại file hệ thống như HDFS, S3, FTP…

*hdfs dfs* được thiết kế chỉ để làm việc với HDFS.

Trong thực tế có thể dùng cách nào cũng được tùy vào thói quen. Bài tập 1, 2, 3 thực hành dùng *hadoop fs.* Bài tập 4,5,6 thực hành sử dụng *hdfs dfs*

## Bài 1: Liệt kê thư mục trên HDFS

1. Bật máy ảo rồi mở Terminal window.
2. Trong màn hình hình terminal gõ lệnh sau để xem thông tin help của FsShell

hadoop fs

1. Lệnh dưới đây sẽ liệt kê nội dung thư mục root trong HDFS. Một trong những kết quả trả ra là */user*. Mỗi user sẽ có 1 thư mục là thư mục con của */user*. Ví dụ, chúng ta đang dùng username là training thì trên HDFS sẽ có thư mục */user/training*

hadoop fs -ls /

1. Liệt kê thư mục của tất cả HDFS user

hadoop fs -ls /user

1. Liệt kê nội dung của user training. Hiện giờ, thư mục này sẽ chưa chứa data. Trong trường hợp bạn liệt kê thư mục không tồn tại ví dụ /user/test thì error message sẽ hiện ra.

hadoop fs -ls /user/training

## Bài 2: Đẩy file vào HDFS

1. Vẫn trong màn hình Terminal, thay đổi đường dẫn đến thư mục chứa dữ liệu

cd ~/training\_materials/developer/data

1. Unzip *shakespeare.tar.gz* bằng lệnh dưới. Sau khi unzip thư mục tên *shakespeare/* chứa các file dữ liệu như *comedies, poems,…*

tar zxvf shakespeare.tar.gz

1. Đẩy dữ liệu từ thư mục *shakespeare/* vào HDFS:

hadoop fs -put shakespeare /user/training/shakespeare

1. Lúc này, dữ liệu từ thư mục *shakespeare* đã được đẩy vào thư mục */user/training* trên HDFS. Xem lại những dữ liệu vừa được đẩy lên bằng lệnh

hadoop fs -ls /user/training

1. Nếu bạn sử dụng lệnh *hadoop fs -ls* thì vẫn cho kết quả giống như trên. Vì Hadoop mặc định sử dụng home directory. Do ở đây chúng ta dùng user là training nên home directory sẽ là */user/training*

hadoop fs -ls

Từ bước 6 trở đi, chúng ta sẽ thực hành cách tạo thư mục trên HDFS có tên là *weblog* và đẩy dữ liệu vào đây.

1. Tạo thư mực weblog trên HDFS bằng lệnh sau

hadoop fs -mkdir weblog

1. Giải nén file *access\_log.gz* và đẩy lên HDFS trong cùng 1 câu lệnh. Lựa chọn -c của gunzip sẽ cho phép giải nén và đẩy ra standard output, trong khi dấu gạch ngang (-) sau lệnh Hadoop fs -put có nghĩa là Hadoop sẽ lấy tất cả file được gửi từ standard input và đẩy lên HDFS.

gunzip -c access\_log.gz | hadoop fs -put – weblog/access\_log

1. Chạy lệnh liệt kê file access\_log vừa được đẩy lên.

Hadoop fs -ls /user/training/weblog

1. File access\_log khá lớn (500MB), chúng ta có thể upload 1 phần của file này (5000 dòng đầu tiên) lên bằng lệnh sau.

hadoop fs -mkdir testlog

gunzip -c access\_log.gz | head -n 500 | hadoop fs -put – testlog/test\_access\_log

## Bài 3: Xem và xóa file trên HDFS

1. Xem các file có trong thư mục *shakespeare* đã làm ở bài tập 2 bằng lệnh dưới. Hệ thống sẽ liệt kê nội dung trong thư mục */user/training/shakespeare*, gồm các file: *comedies, glossary, histories, poem* và *tragedies*.

hadoop fs -ls shakespeare

1. Chúng ta có thể xóa file *glossary* bằng lệnh:

hadoop fs -rm shakespeare/glossary

1. Chúng ta có thể xem nội dung file trên hdfs bằng lệnh hadoop *fs -cat*. Lệnh dưới sẽ cho chúng ta xem 50 dòng cuối của file *histories*. Trong thực tế, 1 file trên HDFS có thể rất lớn nên chúng ta hay sử dụng lệnh *tail* hoặc lệnh *head* đi kèm

hadoop fs -cat shakespeare/histories | tail -n 50

1. Để download 1 file từ HDFS trở về local ta có thể dùng lệnh *fs -get*. Lệnh dưới đây sẽ download file *poems* trên HDFS và lưu thành file shakepoem.txt ở thư mục gốc trên máy local.

hadoop fs -get shakespeare/poems ~/shakepoems.txt

1. Xem lại nội dung trong file shakepoems.txt bằng lệnh:

less ~/shakepoems.txt

## Bài 4: Liệt kê thư mục trong HDFS sử dụng hdfs dfs

1. Mở Terminal window.
2. Lệnh dưới đây sẽ liệt kê nội dung thư mục root trong HDFS. Một trong những kết quả trả ra là */user*. Mỗi user sẽ có 1 thư mục là thư mục con của */user*. Ví dụ, chúng ta đang dùng username là training thì trên HDFS sẽ có thư mục */user/training*

hdfs dfs -ls /

1. Liệt kê thư mục của tất cả HDFS user

hdfs dfs -ls /user

1. Liệt kê nội dung của user *training*.

hdfs dfs -ls /user/training

## Bài 5: Đẩy file lên HDFS sử dụng hdfs dfs

1. Vẫn trong màn hình Terminal, thay đổi đường dẫn đến thư mục chứa dữ liệu

cd ~/traning\_materials/developer/data

1. Đẩy dữ liệu từ thư mục shakespeare/ vào HDFS (chúng ta dùng shakespeare2 trên HDFS vì bài trên đã có tên là shakespeare)

hdfs dfs -put shakespeare /user/training/shakespeare2

1. Lúc này, dữ liệu từ thư mục shakespeare đã được đẩy vào thư mục /user/training trên HDFS. Xem lại những dữ liệu vừa được đẩy lên bằng lệnh

hdfs dfs -ls /user/training

1. Nếu bạn sử dụng lệnh hdfs dfs -ls thì vẫn cho kết quả giống như trên. Vì Hadoop mặc định sử dụng home directory. Do ở đây chúng ta dùng user là training nên home directory sẽ là /user/training

hdfs dfs -ls

Từ bước 6 trở đi, chúng ta sẽ thực hành cách tạo thư mục trên HDFS có tên là *weblog2* và đẩy dữ liệu vào đây.

1. Tạo thư mực *weblog2* trên HDFS bằng lệnh sau

hdfs dfs -mkdir weblog2

1. Giải nén file *access\_log.gz* và đẩy lên HDFS trong cùng 1 câu lệnh. Lựa chọn -c của gunzip sẽ cho phép giải nén và đẩy ra standard output, trong khi dấu gạch ngang (-) sau lệnh hdfs dfs -put có nghĩa là Hadoop sẽ lấy tất cả file được gửi từ standard input và đẩy lên HDFS.

gunzip -c access\_log.gz | hdfs dfs -put – weblog2/access\_log

1. Chạy lệnh liệt kê file *access\_log2* vừa được đẩy lên.

hdfs dfs -ls /user/training/weblog2/

## Bài 6: Xem và xóa file trên HDFS sử dụng hdfs dfs

1. Xem các file có trong thư mục *shakespeare2* đã làm ở bài tập 5 bằng lệnh dưới. Hệ thống sẽ liệt kê nội dung trong thư mục */user/training/shakespeare*, gồm các file: *comedies, glossary, histories, poem* và *tragedies*.

hdfs dfs -ls shakespeare2

1. Chúng ta có thể xóa file *glossary* bằng lệnh:

hdfs dfs -rm shakespeare2/glossary

1. Chúng ta có thể xem nội dung file trên hdfs bằng lệnh *hdfs dfs -cat.* Lệnh dưới sẽ cho chúng ta xem 50 dòng cuối của file *histories*. Trong thực tế, 1 file trên HDFS có thể rất lớn nên chúng ta hay sử dụng lệnh *tail* hoặc lệnh *head* đi kèm

hdfs dfs -cat shakespeare2/histories | tail -n 50

1. Để download 1 file từ HDFS trở về local ta có thể dùng lệnh *fs -get*. Lệnh dưới đây sẽ download file *poems* trên HDFS và lưu thành file shakepoem.txt ở thư mục gốc trên máy local.

hdfs dfs -get shakespeare2/poems ~/shakepoems.txt

1. Xem lại nội dung trong file shakepoems.txt bằng lệnh:

less ~/shakepoems.txt

## Bài 7: Sử dụng HUE để thao tác file trên HDFS

1. Dùng Firefox và gõ url như dưới để truy nhập HUE. Username và password là admin.

http://localhost:8888

# **II. Quản trị HDFS.**

1. Lấy thông tin namenodes. Lệnh sau sẽ hiển thị tên namenode

hdfs getconf -namenodes

Lệnh hdfs fsck dùng để kiểm tra các file, block trên HDFS

1. Lệnh dưới đây liệt kê các file và trạng thái trên HDFS.

hdfs fsck / -files

d

1. Lệnh dưới liệt kê các file và blocks của các file đó trên HDFS

hdfs fsck / -files -blocks

d

1. Lệnh dưới liệt kê các blocks của file weblog và vị trí (location) của các block này

hdfs fsck / -files -blocks -locations

d

1. Chúng ta có thể xem báo cáo về trạng thái của cluster bằng cách sử dụng lệnh *dfsadmin*. Để dùng lệnh này thì cần phải là super user. Trong HDFS, supert user là hdfs user. Lệnh *sudo -u hdfs* có nghĩa là chúng ta chuyển từ user *training* sang *hdfs* user.

sudo -u hdfs hdfs dfsadmin -report

d