Nội dung thực hành

[I. Sử dụng Sqoop để import dữ liệu từ cơ sở dữ liệu quan hệ vào HDFS 1](#_Toc9601104)

[Bài 1: Import 2 bảng của movielens lên HDFS 1](#_Toc9601105)

[Bài 2: Import các bảng từ database dualcore vào HDFS, dùng một số lựa chọn trong Sqoop 3](#_Toc9601106)

# I. Sử dụng Sqoop để import dữ liệu từ cơ sở dữ liệu quan hệ vào HDFS

Bài tập này giúp bạn thực hành việc import dữ liệu từ MySQL vào HDFS sử dụng Apache Sqoop

## Bài 1: Import 2 bảng của movielens lên HDFS

*Lệnh 1 đến 6 giúp bạn nắm được cấu trúc của hai bảng movie và movierating trong movielens*

1. Mở Terminal window và truy nhập vào database *movielens* của MySQL bằng lệnh dưới. Account để logon vào mysql là *training/training*

mysql -utraining -ptraining movielens

1. Lệnh *show tables;* sẽ liệt kê tất cả các bảng của *movielens* (chú ý có dấu ; sau lệnh)

SHOW tables;

1. Lệnh *describe movie;* sẽ hiển thị cấu trúc của bảng *movie*. Gồm 3 trường là *id, name, year*

DESCRIBE movie;

1. Lệnh dưới liệt kê 5 bản ghi của bảng *movie*

SELECT \* FROM movie LIMIT 5;

1. Bạn có thể xem cấu trúc của bảng *movierating* bằng cách lặp lại bước 3, 4 ở trên (thay *movie* bằng *movierating*)
2. Thoát khỏi MySQL bằng lệnh

quit;

*Lệnh 7 đến 9 mô tả việc liệt kê (list) danh sách databases, tables trong MySQL sử dụng Sqoop.*

1. Xem các lệnh trong sqoop như sau (bạn có thể bỏ qua warning về accumulo)

sqoop help

1. Liệt kê tất cả database trên mysql. Lệnh dưới sẽ hiển thị database như *movielens, hue, metastore,…* (chú ý dấu \ có nghĩa là cho phép chúng ta viết lệnh trên nhiều dòng)

sqoop list-database \  
--connect jdbc:mysql://localhost \  
--username training --password training

1. Liệt kê các bảng trong *movielens*. Lệnh dưới sẽ hiển thị 1 số bảng như *movie, movierating*,…

sqoop list-tables \  
--connect jdbc:mysql://localhost/movielens \  
--username training --password training

*Lệnh 10 đến 13 cho thực hành việc đẩy dữ liệu tử bảng movie, movierating lên HDFS*

1. Đẩy dữ liệu từ bảng *movie* vào HDFS. Các trường trong movie được phân chia bởi dấu tab (‘\t’)

sqoop import \  
--connect jdbc:mysql://localhost/movielens \  
--username training --password training \  
--fields-terminated-by ‘\t’ \  
--table movie

1. Lúc này bảng *movie* đã được đẩy lên HDFS trên thư mục */user/training/movie*. Kiểm tra bằng lệnh dưới.

hadoop fs -ls movie

1. Xem 1 số data mẫu bằng lệnh:

hadoop fs -tail movie/part-m-00000

1. Lặp lại bước 10 đến 12 để import bảng *movierating* lên HDFS (thay tên bảng movie thành movierating trong các câu lệnh)

## Bài 2: Import các bảng từ database dualcore vào HDFS, dùng một số lựa chọn trong Sqoop

1. Chạy lệnh sau để đẩy dữ liệu từ bảng *customers* trong database *dualcore* lên HDFS. Mặc định dữ liệu sẽ được đẩy vào thư mục */user/training/customers* trên HDFS. Tuy nhiên chúng ta có thể chuyển sang thư mục khác (trong bài này là */dualcore*) bằng cách sử dụng lựa chọn *--warehouse-dir*

sqoop import \  
--connect jdbc:mysql://localhost/dualcore \  
--username training --password training \  
--fields-terminated-by ‘\t’ \  
--warehouse-dir /dualcore \  
--table customers

1. Làm tương tự với các bảng *products, orders*.
2. Tiếp theo, chúng ta đẩy dữ liệu của bảng *order\_details* lên. Bảng này là bảng trung gian của *orders* và *products*, nó không có khóa chính mà có 2 khóa ngoại từ 2 bảng trên. Do vậy, để đẩy dữ liệu từ bảng này lên chúng ta cần phải chỉ định lựa chọn *--split-by* là *order\_id*

sqoop import \  
--connect jdbc:mysql://localhost/dualcore \  
--username training --password training \  
--fields-terminated-by ‘\t’ \  
--warehouse-dir /dualcore \  
--table order\_details \  
--split-by=order\_id