

Homework 1

Created	@November 29, 2022 2:10 PM
∷ Tags	

▼ Khởi tạo hệ thống

- Thông tin dữ liệu:
 - Movies: gồm 2 bảng movies (với 3883 row) và bảng ratings (với 1000209 row)

▼ Import dữ liệu vào MySql

- Copy dữ liệu movies len bao gồm 2 file movies.dat và <u>ratings.dat</u> vào thư mục /var/lib/mysql/ của container MySql
- Vào terminal của container MySql

```
docker exec -it mysql bash
```

• Vào mysql shell. (password của root là admin)

```
mysql --local-infile=1 -u root -p
```

• Set quyền đọc ghi file

```
SET GLOBAL local_infile=1;
```

Vào database db

```
use db;
```

Tạo và ghi dữ liệu vào bảng movies

```
#tao bang movies

create table if not exists db.movies (movieId INT PRIMARY KEY, name TEXT, category TEXT);

#load data vao bang

load data local infile "/var/lib/mysql/movies.dat" into table db.movies fields te rminated by "::";
```

Kết quả bảng:

```
mysql> select * from movies limit 10;
 movieId | name
                                                       category
            Toy Story (1995)
Jumanji (1995)
Grumpier Old Men (1995)
                                                        Animation|Children's|Comedy
        1 |
                                                        Adventure | Children's | Fantasy
                                                       Comedy | Romance
Comedy | Drama
            Waiting to Exhale (1995)
             Father of the Bride Part II (1995)
                                                       Comedy
            Heat (1995)
                                                        Action|Crime|Thriller
            Sabrina (1995)
                                                        Comedy Romance
                                                        Adventure|Children's
        8
            Tom and Huck (1995)
        9
            Sudden Death (1995)
                                                        Action
                                                       Action | Adventure | Thriller
       10 | GoldenEye (1995)
10 rows in set (0.00 sec)
```

Tạo và ghi dữ liệu vào bảng ratings

```
#tao bang ratings

CREATE TABLE IF NOT EXISTS db.ratings (userID INT ,itemID INT ,rating INT,timesta mp INT);

#load data vao bang

load data local infile "/var/lib/mysql/ratings.dat" into table db.ratings fields terminated by "::";
```

Kết quả:

mysql> select * from ratings limit 10;				
userID	itemID	rating	timestamp	
1	1193	5	978300760	
1	661	3	978302109	
1	914	3	978301968	
1	3408	4	978300275	
1	2355	5	978824291	
1	1197	3	978302268	
1	1287	5	978302039	
1	2804	5	978300719	
1	594	4	978302268	
1	919	4	978301368	
++				
10 rows in set (0.01 sec)				

- ▼ Import dữ liệu từ MySql sang HDFS, Hive
 - Vào container master để import:

```
docker exec -it master bash
```

• Gõ hive để vào hive shell , khởi tạo database sqoop_workspace

```
create database sqoop_workspace;
```

- import bảng movies vào Hive và lưu trên HDFS
 - Syntax:

```
sqoop import --connect jdbc:mysql://172.28.1.10:3306/db --username root -P --
split-by movieId --table movies --target-dir /movies --fields-terminated-by
"," --hive-import --create-hive-table --hive-table sqoop_workspace.movies
```

Check bảng đã được tạo trên Hive

```
hive> show create table movies;
CREATE TABLE `movies`(
  `movieid` int,
  `name` string,
  `category` string)
COMMENT 'Imported by sqoop on 2022/11/30 07:08:12'
ROW FORMAT SERDE
  'org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe'
WITH SERDEPROPERTIES (
  'field.delim'=',',
  'line.delim'='\n'
  'serialization.format'=',')
STORED AS INPUTFORMAT
  'org.apache.hadoop.mapred.TextInputFormat'
  'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'
  'hdfs://master:9000/usr/hive/warehouse/sqoop workspace.db/movies'
TBLPROPERTIES (
  'bucketing_version'='2',
  'transient lastDdlTime'='1669792099')
Time taken: 0.389 seconds, Fetched: 20 row(s)
```

- import bảng ratings vào Hive và lưu trên HDFS
 - Syntax:

```
sqoop import --connect jdbc:mysql://172.28.1.10:3306/db --username root -P --split-by userID --table ratings --target-dir /ratings --fields-terminated-by "," --hive-import --create-hive-table --hive-table sqoop_workspace.ratings
```

Check bảng đã tạo:

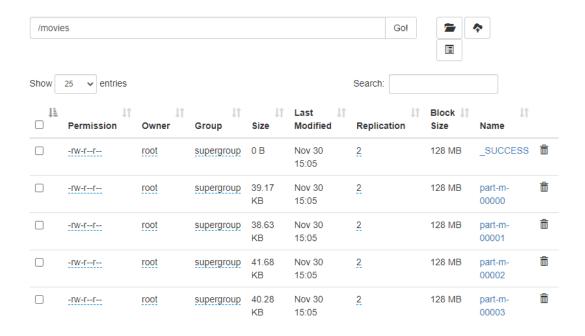
```
hive> show create table ratings;
ОК
CREATE TABLE `ratings`(
  `userid` int,
  `itemid` int,
  rating` int,
  `timestamp` int)
COMMENT 'Imported by sqoop on 2022/11/30 07:23:41'
ROW FORMAT SERDE
  'org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe'
WITH SERDEPROPERTIES (
  'field.delim'=',',
  'line.delim'='\n'
  'serialization.format'=',')
STORED AS INPUTFORMAT
  'org.apache.hadoop.mapred.TextInputFormat'
OUTPUTFORMAT
  'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'
LOCATION
  'hdfs://master:9000/usr/hive/warehouse/sqoop workspace.db/ratings'
TBLPROPERTIES (
  'bucketing_version'='2',
  'transient lastDdlTime'='1669793028')
Time taken: 0.048 seconds, Fetched: 21 row(s)
```

▼ Import thẳng dữ liệu vào HDFS

- Import bång movies vào HDFS
 - Syntax:

```
sqoop import --connect jdbc:mysql://172.28.1.10:3306/db --username root -P --
split-by movieId --table movies --target-dir /movies --fields-terminated-by
","
```

Kết quả:

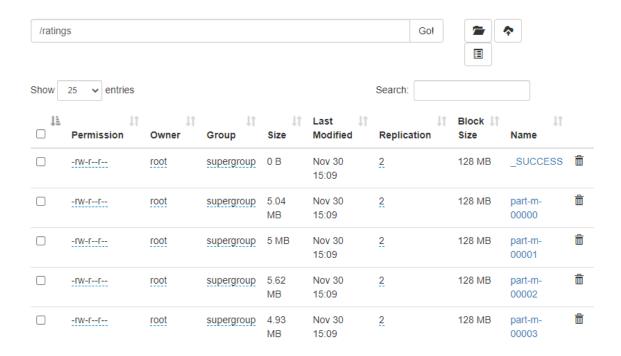


• Import bảng ratings vào HDFS

Syntax:

```
sqoop import --connect jdbc:mysql://172.28.1.10:3306/db --username root -P --
split-by userID --table ratings --target-dir /ratings --fields-terminated-by
","
```

Kết quả:



▼ Thao tác bảng với Hive

Vào container master để vào hive

```
docker exec -it master bash
```

- Vào hive, vào database sqoop_workspace
- ▼ Kiểm thử và so sánh hiệu năng của Hive so với MySql
- ▼ Môt số vấn đề khi cài đặt
 - Khi clone docker file từ github, thì xóa file .git đi vì git sẽ k cho up file .sh lên
 - Nên tải file zip khi clone từ window, hoặc dùng notepad để format lại file .sh về format Unix
 - Khi dùng sqoop import từ mysql vào hive, hdfs, cần thư viện mapreduce, bổ dung vào file conf của hadoop là file mapred-site.xml:

```
<name>mapreduce.application.classpath
<value>$HAD00P_HOME/share/hadoop/mapreduce/*,$HAD00P_HOME/share/hadoop/ma
```

• Trước khi dùng sqoop import thì vào Hive tạo một database, nếu không thì phải dùng database default