



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 lên bao gồm 2 file movies.dat và ratings.dat 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
1	Toy Story (1995)	Animation Children's Comedy
2	Jumanji (1995)	Adventure Children's Fantasy
3	Grumpier Old Men (1995)	Comedy Romance
4	Waiting to Exhale (1995)	Comedy Drama
5	Father of the Bride Part II (1995)	Comedy
6	Heat (1995)	Action Crime Thriller
7	Sabrina (1995)	Comedy Romance
8	Tom and Huck (1995)	Adventure Children's
9	Sudden Death (1995)	Action
10	GoldenEye (1995)	Action Adventure Thriller

```
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,timestamp 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 **scoop_workspace**

```
create database scoop_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;
OK
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'
OUTPUTFORMAT
  'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'
LOCATION
  '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)
hive>
```

- 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;
OK
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 bảng 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:

Show entries
Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
<input type="checkbox"/>	-rw-r--r--	root	supergroup	0 B	Nov 30 15:05	2	128 MB	_SUCCESS
<input type="checkbox"/>	-rw-r--r--	root	supergroup	39.17 KB	Nov 30 15:05	2	128 MB	part-m-00000
<input type="checkbox"/>	-rw-r--r--	root	supergroup	38.63 KB	Nov 30 15:05	2	128 MB	part-m-00001
<input type="checkbox"/>	-rw-r--r--	root	supergroup	41.68 KB	Nov 30 15:05	2	128 MB	part-m-00002
<input type="checkbox"/>	-rw-r--r--	root	supergroup	40.28 KB	Nov 30 15:05	2	128 MB	part-m-00003

- 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:

Show entries
Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	-rw-f--f--	root	supergroup	0 B	Nov 30 15:09	2	128 MB	_SUCCESS	
<input type="checkbox"/>	-rw-f--f--	root	supergroup	5.04 MB	Nov 30 15:09	2	128 MB	part-m-00000	
<input type="checkbox"/>	-rw-f--f--	root	supergroup	5 MB	Nov 30 15:09	2	128 MB	part-m-00001	
<input type="checkbox"/>	-rw-f--f--	root	supergroup	5.62 MB	Nov 30 15:09	2	128 MB	part-m-00002	
<input type="checkbox"/>	-rw-f--f--	root	supergroup	4.93 MB	Nov 30 15:09	2	128 MB	part-m-00003	

▼ 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

▼ Một 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 sung vào file conf của hadoop là file **mapred-site.xml**:

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

```
preduce/lib/*,$HADOOP_HOME/share/hadoop/common/*,$HADOOP_HOME/share/hadoop/commo  
n/lib/*,$HADOOP_HOME/share/hadoop/yarn/*,$HADOOP_HOME/share/hadoop/yarn/lib/*,$HA  
DOOP_HOME/share/hadoop/hdfs/*,$HADOOP_HOME/share/hadoop/hdfs/lib/*</value>  
</property>
```

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