

## Sqoop

### 1. Install Sqoop

- `# yum install -y sqoop mysql libmysql-java`

### 2. Create symbolic link to MySQL driver JDBC

- `# ls /usr/share/java/mysql-connector-java.jar`
- `# ls /usr/lib/sqoop/lib/mysql-connector-java.jar -l`
- If not exists symbolic link, create it  
`# ln -s /usr/share/java/mysql-connector-java.jar /usr/lib/sqoop/lib/`

### 3. Configure MySQL to start on boot

- `# chkconfig mysqld on`

### 4. Start MySQL

- `# service mysqld start`

### 5. Create password for root

- `# /usr/bin/mysqladmin -u root password 'cloudera'`

### 6. Create database

- `$ mysql -u root -p`
- `mysql> create database musicbrainz;`
- `mysql> grant select,insert,update,delete ON musicbrainz.* TO cloudera@localhost IDENTIFIED BY 'cloudera';`

### 7. Import schema musicbrainz

- `wget ftp://tallergpul/files/musicbrainz.tar.gz`
- `tar zxvf musicbrainz.tar.gz`
- `$ mysql -u root -p musicbrainz < musicbrainz.sql`

### 8. Show data from tables

- `mysql> select t.trackname, t.position, r.album, r.artist from musicbrainz_tracks t JOIN musicbrainz_releases r ON (t.releaseid=r.id) LIMIT 0,10;`

### 9. Export tables to HDFS

- Table musicbrainz\_tracks

```
$ sqoop import --connect jdbc:mysql://localhost/musicbrainz --table musicbrainz_tracks --columns "trackid, releaseid, trackname, position" --fields-terminated-by '\t' --username cloudera --password cloudera
```

- Table musicbrainz\_releases

```
$ sqoop import --connect jdbc:mysql://localhost/musicbrainz --table musicbrainz_releases --columns "id, album, artist" --fields-terminated-by '\t' --username cloudera --password cloudera
```

### 10. Check paths and files in HDFS

- `$ hadoop fs -ls musicbrainz_releases`

- \$ `hadoop fs -tail musicbrainz_releases/part-m-00003`
- \$ `hadoop fs -ls musicbrainz_tracks/`

```

cloudera@elephant:~
File Edit View Search Terminal Tabs Help
cloudera@elephant:~
14/10/08 00:26:33 INFO mapred.JobClient: Counters: 23
14/10/08 00:26:33 INFO mapred.JobClient:   File System Counters
14/10/08 00:26:33 INFO mapred.JobClient:     FILE: Number of bytes read=0
14/10/08 00:26:33 INFO mapred.JobClient:     FILE: Number of bytes written=802448
14/10/08 00:26:33 INFO mapred.JobClient:     FILE: Number of read operations=0
14/10/08 00:26:33 INFO mapred.JobClient:     FILE: Number of large read operations=0
14/10/08 00:26:33 INFO mapred.JobClient:     FILE: Number of write operations=0
14/10/08 00:26:33 INFO mapred.JobClient:     HDFS: Number of bytes read=426
14/10/08 00:26:33 INFO mapred.JobClient:     HDFS: Number of bytes written=11378516
14/10/08 00:26:33 INFO mapred.JobClient:     HDFS: Number of read operations=4
14/10/08 00:26:33 INFO mapred.JobClient:     HDFS: Number of large read operations=0
14/10/08 00:26:33 INFO mapred.JobClient:     HDFS: Number of write operations=4
14/10/08 00:26:33 INFO mapred.JobClient:   Job Counters
14/10/08 00:26:33 INFO mapred.JobClient:     Launched map tasks=10
14/10/08 00:26:33 INFO mapred.JobClient:     Total time spent by all maps in occupied slots (ms)=106789
14/10/08 00:26:33 INFO mapred.JobClient:     Total time spent by all reduces in occupied slots (ms)=0
14/10/08 00:26:33 INFO mapred.JobClient:     Total time spent by all maps waiting after reserving slots (ms)=0
14/10/08 00:26:33 INFO mapred.JobClient:     Total time spent by all reduces waiting after reserving slots (ms)=0
14/10/08 00:26:33 INFO mapred.JobClient:   Map-Reduce Framework
14/10/08 00:26:33 INFO mapred.JobClient:     Map input records=274263
14/10/08 00:26:33 INFO mapred.JobClient:     Map output records=274263
14/10/08 00:26:33 INFO mapred.JobClient:     Input split bytes=426
14/10/08 00:26:33 INFO mapred.JobClient:     Spilled Records=0
14/10/08 00:26:33 INFO mapred.JobClient:     CPU time spent (ms)=12250
14/10/08 00:26:33 INFO mapred.JobClient:     Physical memory (bytes) snapshot=394383360
14/10/08 00:26:33 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=2901127168
14/10/08 00:26:33 INFO mapred.JobClient:     Total committed heap usage (bytes)=95420416
14/10/08 00:26:33 INFO mapreduce.ImportJobBase: Transferred 10.8514 MB in 58.1336 seconds (191.1432 KB/sec)
14/10/08 00:26:33 INFO mapreduce.ImportJobBase: Retrieved 274263 records.
[cloudera@elephant ~]$ hadoop fs -ls
Found 3 items
drwx----- - cloudera supergroup          0 2014-10-08 00:26 .staging
drwxr-xr-x - cloudera supergroup          0 2014-10-08 00:26 musicbrainz_releases
drwxr-xr-x - cloudera supergroup          0 2014-10-07 03:06 wordcount
[cloudera@elephant ~]$
[cloudera@elephant ~]$ hadoop fs -ls musicbrainz_releases
Found 6 items
-rw-r--r--  3 cloudera supergroup          0 2014-10-08 00:26 musicbrainz_releases/ SUCCESS
drwxr-xr-x - cloudera supergroup          0 2014-10-08 00:25 musicbrainz_releases/_logs
-rw-r--r--  3 cloudera supergroup    2699084 2014-10-08 00:26 musicbrainz_releases/part-m-00000
-rw-r--r--  3 cloudera supergroup    3034338 2014-10-08 00:26 musicbrainz_releases/part-m-00001
-rw-r--r--  3 cloudera supergroup    2926631 2014-10-08 00:26 musicbrainz_releases/part-m-00002
-rw-r--r--  3 cloudera supergroup    2718463 2014-10-08 00:26 musicbrainz_releases/part-m-00003
[cloudera@elephant ~]$ hadoop fs -tail musicbrainz_releases/part-m-00003

```