

Chapter 2 – Loading Data into Hive

Exercise

Part 1

1. This exercise uses the MovieLens data set. This is a free data set provided by the University of Minnesota's GroupLens Research Group, and contains over one million movie ratings. For more information, please see the license information at the end of this exercise.

Extract the data from the MovieLens data set:

```
$ tar xvzf ml-data.tar__0.gz
```

You should now have a subdirectory called ml-data with several files in it.

2. Examine the files u.data and u.item files.
3. Create a table called UserRatings with these column names: userid, movieid, rating, unixtime (The 4th column is a timestamp, represented by the number of seconds since 01/01/1970). Use the appropriate field delimiter for the u.data file.
4. Move the u.data file into HDFS with this command:

```
$ hadoop fs -put ml-data/u.data /user/training/u.data
```
5. Use `LOAD DATA INPATH` to import the u.data file into the UserRatings table.
6. Verify that data was loaded:

```
hive> SELECT * FROM UserRatings LIMIT 10;
```

196	242	3	881250949
186	302	3	891717742
22	377	1	878887116
244	51	2	880606923
166	346	1	886397596
298	474	4	884182806
115	265	2	881171488
253	465	5	891628467
305	451	3	886324817
6	86	3	883603013

Part 2

1. The Movie data (u.info) has already been loaded into MySQL. Log into MySQL and look at the table:

```
1. $ mysql
```

```
2. mysql> use training
```

```
3. mysql> DESCRIBE Movies;
```

2. *Make sure you exit the hive shell before doing this step since we are using the local Derby database as a metastore which does not support multiple concurrent connections.*

Use Sqoop to import the table into Hive from MySQL. The resulting Hive table should use '\001' (a ctrl-A character) between fields and newline (\n) between lines. The username for MySQL is "training", the password is empty.

3. Verify the table was imported into Hive.

4. Use the Hadoop shell to examine the data in HDFS:

```
$ hadoop fs -ls /user/hive/warehouse
$ hadoop fs -ls /user/hive/warehouse/userratings
$ hadoop fs -ls /user/hive/warehouse/movies
```

This exercise uses the MovieLens data set, or subsets thereof. This data is freely available for academic purposes, and used and distributed by Cloudera with the express permission of the UMN GroupLens Research Group. If you would like to use this data for your own research purposes, you are free to do so, so long as you cite the GroupLens Research Group in any resulting publications. If you would like to use this data for commercial purposes, you must obtain explicit permission. You may find the full dataset, as well as detailed license terms at <http://www.grouplens.org/node/73>