```
-----------------------------------------------------------------
-----------------------------------------------
-----------------------------------------------------------------
-----------------------------------------------
-- Programming Hive  - Additional Hive Exercise (Optional)

-- 1. In this optional lab exercise, we will work with the
MovieLens dataset
--    The movielens dataset is a collection of movie ratings data
and has been widely used in the industry and
--    academia for experimenting with recommendation algorithms
and we see many publications using this dataset
--    to benchmark the performance of their algorithms
-- 2. For access to full-sized movielens data, go to
http://grouplens.org/datasets/movielens/
-----------------------------------------------------------------
-----------------------------------------------
-----------------------------------------------------------------
-----------------------------------------------

-----------------------------------------------
-- Loading User Ratings Data into Hive - u.data
-----------------------------------------------

-- 1. Upload movielens.tgz file to linux sandbox /home/lab

-- 2. Extract the data from the MovieLens dataset

$ cd /home/lab
$ tar -zxvf movielens.tgz
$ ll

-- 3. Examine the files

$ cd ml-data
$ more u.data
-- You will find two file u.data and u.item
-----------------------------
-- 4. Create a database called ml and table called user_ratings
(tab-delimited)
-- 5. Move file u.data into hadoop
-- 6. Load the u.data into user_ratings hive table
-- 7. Verify the data was loaded into hdfs

---------------------------------
-- loading file u.item into hive
---------------------------------


-- 8. Create a table called movies
-- Read the README file for u.item column description
-- 9. Move the  file u.item into hadoop
```

```
-- 10. Load the u.item into hive table called ml.Movies
-- 11. Verify the data was loaded


--12. Examine both tables on hdfs
------------------------
-- Simple analysis
------------------------

-- 1. how many records in both tables?

-- 2. find the name of all movies released in 1990

-- 3. list the movieid of the 10 most rated films in user_ratings
table

-- 4. use a join to list the titles of the movies you found in
step 3

-- 5. do any movies have no ratings? (hint: outer join and IS
NULL)

-- 6. what is the highest rated sci_fi mvoie

-- 7. what is the highest rated sci_fi movie that has at least 10
user ratings
```