Assignment 7

The goal of this assignment is to install, configure and use Hive for data science activities.
NOTE: The latest version of Hive is 2.1.0. But, we will be working with Hive 1.2.1. So, please download 1.2.1. The main difference between 1.2.1 and the new 2.x is that the new Hive has added a few additional features such as default MySQL database for Schema, Security, new internal file formats etc. These are additional features to make Hive work with HBase, Spark etc. We don't have to worry about all that for this class. So, to keep installation simple please work with Hive 1.2.1. As you install Hive there is a place where you are asked to create directories in hdfs because the HDFS is where the data is stored for Hive Processing. So, in Setting up Hive module, when you are ready to create the directories add the –p option to create the complete Path:

```
$ hadoop fs -mkdir  -p/tmp
$ hadoop fs -mkdir -p /user/hive/warehouse
```

Ok so you have installed Hive. Now it is time to work on the actual Hive analysis of ml-100k data.

You are already familiar with u.data file located in ~/Downloads/ml-100k folder in your local file system. If you recall the u.data file contains 100000 ratings by 943 users on 1682 items. Each user has rated at least 20 movies. Users and items are numbered consecutively from 1. The data is randomly ordered. This is a tab separated list of user id, item id, rating, timestamp. The time stamps are unix seconds since 1/1/1970 UTC.

There is another file in ~/Downloads/ml-100k folder called u.item. This file contains about the items (movies). This is a pipe separated list of movie id, movie title, release date, video release date, IMDb URL, unknown, Action, Adventure, Animation, Children's, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War, Western. The last 19 fields are the genres, a 1 indicates the movie is of that genre, a 0 indicates it is not; movies can be in several genres at once. The movie ids are the ones used in the u.data data set.

Write Hive schema and statements to print out the complete list of all movies rated with the following information per line -- movie id, movie title, release date, IMDB URL, average rating for the movie. You will have to do the following:

1. Create a database called "movielens"
2. Create table for "rating"
3. Create table for "item"
4. Load data into rating table

5. Load data into item
6. Execute the Select statement with join and group by to extract list of all movies rated with the following information per line -- movie id, movie title, release date, IMDB URL, average rating for the movie.

What to turn in?

1) Sample run with output as a Word document. I have given you a template for assignment 7 solutions. Follow the guidelines, fill the document and submit.