# Fifth Homework Assignment (15% of grade)

Due on Friday, June 9 at 5 pm.

**Submission Guidelines:**

Please follow carefully the instructions posted on the course's github page when submitting your solutions (https://github.com/MSIA/bigdatacourse/blob/master/README.md). Failure to follow the instructions will result in lost points.

**Deliverables:**

1. Your hbase and hive source code file: name the file lastname_x.(hbs,sql) (For hbase make sure that the script also includes lines for database model creation.)
2. Output files: name them lastname_x.txt

Here 'x' is the number of the task.

## Enron Emails

Folder /home/public/course/enron contains individual emails of former Enron employees. There is one file per email and the name of the employee is listed as the subfolder name. These emails were used in the actual trail but the judge decided to release them for public consumption. (I have all of them but I am sharing with you only a few of them so that potentially you do not need to write a script to load the data. You are encouraged to write a script, but because you have only a few emails you can insert them manually.)

Load in an appropriate hbase model the following attributes of each email: name of the employee (obtained from the name of the folder), email address of the sender, date of the email, send to as string, and email body as string.

Recently I read in the newspaper FakeInnovations that entrepreneur Bogus John Enron wants to fund a company with the same employees (those jailed will work from the jail). Bogus John needs an hbase database that will track all the emails. The management of the company wants to be able to quickly query emails for a user, all emails during a time period, and all emails for a given user during a period of time.

You have to perform the following tasks:

1. Create an hbase database model.
2. Import all emails in hbase.

3. Return the bodies of all emails for a user of your choice (as a single text file).
4. Return the bodies of all emails written during a particular month of your choice (as a single text file).
5. Return the bodies of all emails of a given user during a particular month both of your choice (as a single text file).

Here are 2 options that you can choose from.

1. Write an hbase script for all of the tasks. This is the route of least effort but also least rewarding.
2. The hbase on wolf offers restful access. You can either use python or java to perform the task.

   Python: use module starbase: http://tinyurl.com/nk92a9r or HappyBase https://happybase.readthedocs.io/en/latest/

   The restful interface on wolf for hbase runs on port 20550 and thus you have initialize the connection as c = Connection(port=20550)

To fetch a single record, use 'get' but to fetch a range of records, use 'scan.'

## Recommendation Engine

File /home/public/course/recommendationEngine contains recommendations of users. The data comes from GroupLens Research Project at the University of Minnesota.

They are movie ratings of users (ratings are 1-5 with 943 users and 1682 movies). The data has format: user id \t movie id \t rating \t timestamp

Given a user recommend related movies.

You can use the following simple algorithm, but you are clearly encouraged to be more creative: Find all pairs of movies rated by the same person with ranking higher than 3. You need to design a strategy which movie to actually recommend based on these counts.

You have to do this in Hive (the command line interface for hive is "beeline"). You have to submit your hive code and a sample output.

REMINDER: Access all necessary files directly from /home/public (and do not copy them into your local folder).