# CS 6350- Big Data Analytics and Management

## Fall 2015 Homework #2

## Topic: Spark Systems

## Due date: Oct 16, 2015

In this homework you will learn how to solve problems using Apache Spark. Please apply ***Apache Spark*** interactive shell (for **scala** or **python**) / run from command line (**scala**/**java**/**python**) to derive some statistics from **Yelp dataset**. The dataset files are located in hdfs in the following path,

**/yelpdatafall/business/business.csv.**

**/yelpdatafall/review/review.csv.**

**/yelpdatafall/user/user.csv.**

A copy of the dataset will also be uploaded to elearning.

**All dataset files are (^) separated.**

### Dataset Description.

The dataset comprises of **three** csv files, namely user.csv, business.csv and review.csv.

**Business.csv** file contain basic information about local businesses.

**Business.csv** file contains the following columns
"business_id","full_address","categories"

'business_id': (a unique identifier for the business)

'full_address': (localized address),

'categories': [(localized category names)]

**review.csv** file contains the star rating given by a user to a business. Use user_id to associate this review with others by the same user. Use business_id to associate this review with others of the same business.

**review.csv** file contains the following columns "review_id","user_id","business_id","stars"

'review_id': (a unique identifier for the review)

'user_id': (the identifier of the reviewed business),

'business_id': (the identifier of the authoring user),

'stars': (star rating, integer 1-5),the rating given by the user to a business

**user.csv file** contains aggregate information about a single user across all of Yelp

**user.csv file** contains the following columns "user_id","name","url"

user_id': (unique user identifier),

'name': (first name, last initial, like 'Matt J.'), this column has been made anonymous to preserve privacy

**'url': url of the user on yelp**

***Q1.*** Given input ***address (any part of the address e.g., city or state)***, find all the ***business ids located at the address***. You must take the input ***address*** in the command line. [For example, if the input ***address*** is ***Stanford*** then you need to find all businesses with stanford ***in the address column***] [You only need ***business.csv*** file to get the answer.]

***Q2***.

**a. Start spark-shell in local mode using all the processor cores on your system or the cluster. (Very important)**

**List the business_id  of the Top 10 businesses** using the average ratings. This will require you to use review.csv. Please answer the question by calculating the average ratings given to each business using the review.csv file. Next, sort the output based on the business_id before taking the top 10  businesses using the average ratings.

**b. Rerun Q2a using Yarn mode. Please solve using our cs6360 cluster. This questions shows how spark can be used on multiple systems in a cluster.**

**Load all the dataset to hadoop cluster as you did in homework1.**

**Use the address of the file on the cluster as input to your scala script.**

Start spark-shell in **YARN mode** using Cs6360 spark cluster.

This spark cluster consist  **6 hadoop machine nodes**. Using the following parameters  Rerun your scala script from question 2a.

Set executor memory =2G

executor cores = 6.

num-executors = 6

For example, the command is as follows.

*spark-shell --master yarn-client --executor-memory 4G --executor-cores 7 --num-executors 6*

**Please measure the execution time of your program using the local mode (Q2a )and yarn mode (Q2b). Please compare the measured time.**

Note: Spark supports only scala or java in YARN mode.

## Q3a:

## List the  business_id , full address and categories of the Top 10 businesses using the average ratings.

## Use the files business.csv and review.csv.

Please sort the output based on the business_id before taking the top 10  business using the average ratings.

**Q3b:**

**Using broadcast variable in spark to store the business data, re implement your solution to question 3a.**

**Please measure the execution time of your program when using spark RDD (Q3a )and when using Broadcast variable (Q3b). Please compare the measured time.**

***Submission***:

You have to upload your submission via e-learning before due date. Please upload the following to eLearning:

1. Three scripting file like, Q1.txt, and Q2.txt and Q3.txt separately if you use scala / python/java interactive shell. Each file contains the scala /python/java code. If you use java, then submit all the java files.
2. Also, submit the commands to start the spark shell for Question 2a. and 2b.