**CS 6350- Big Data Analytics and Management**

**Fall 2015**

**Due Date: Nov 10, 2015**

**Homework/Assignment #3 PART A.**

In this homework you will learn how to use Pig Latin and Cassandra. There are slides on

eLearning to help with every of these tools.

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** csvfiles, 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**

**Part 1: Pig Latin**

Start pig in mapreduce mode by typing pig at command line.

**Q1:**

List the business\_id , full address and categories of the **Top 10 businesses** located in **CA** using the average ratings. This will require you to use review.csv and business.csv files.

Please answer the question by **calculating the average rating**s given to each business using the

review.csv file. Do not use the already calculated ratings (average\_stars) contained in the

business entity rows.

**Q2:**

List the business\_id , full address and categories of the top 10 most reviewed businesses not located in **CA**. Please answer the question by counting the reviews given to each business id not located in CA.

**Q3:**

Using Pig Latin script, Implement co-group command on business\_id for the datasets **review** and **business.** Print first 5rows.

**Q4:**

Repeat Question 2 (implement join) with co-group commands. Print first 5rows.

**NOTE: if dump command does not display result, use the store command to store result into hdfs and then cat the output just like in hw 2**

**e.g**

**>>store E into '/yournetid/casQ1';**

**then exit pig command line and use hdfs command to output your result as shown below.**

**hdfs dfs -cat / yournetid/casQ1/\***

**Part 2: Cassandra**

In this homework you will learn how to use Cassandra. Please use the

“Apache\_Cassandra\_1.2.pdf” for reference and help.

Cassandra 2.05 has been installed and you can access it through cs6360.utdallas.edu. It has four nodes: csac0, csac1, csac2, and csac3. The path is /usr/local/apache-cassandra-2.0.5

**\*\*You are going to create a keyspace with your net ID** (i.e., abc112233) and do all work in this

keyspace. Replication factor should be 1.

**Q5: Cassandra CQL3**

***{cs6360:~} /usr/local/apache-cassandra-2.0.5/bin/cqlsh csac0***

**Requirements**:

Using Cassandra CQL3, write commands to do the following:

1- Create a table for business.csv dataset. Use (business\_id) as the Primary Key.

2- Load all records in the dataset to this table.

3- Select the tuple which has business id 'HPWmjuivv3xJ279qSVfNaQ'

4- Delete all rows in the table.

5- Drop the table.

**Q6:**

Using Cassandra CQL3, write commands to do the following:

1. Create a table for review.csv dataset using the user\_id,business\_id as the primary key and the stars as the sorting key.

2. Create index on column stars.

3. Select any row where the rating is 4.0 limit display result to 10.

4. Delete all rows in the table

5. Drop the table.

**Q7: Cassandra Administration**

1) Run nodetool command and determine how much unbalanced the cluster is.

**Submission:**

Please upload the following to eLearning:

One file with all commands for Q6.

One file with all commands for Q7.