CS 6350- Big Data Analytics and Management

Summer 2015
Due July 12,
11:59pm.

Homework/Assignment #2

Teaching Assistant
GbadeboAyoade,email:gga110020@utdallas.edu
Office hours: Tuesday /Thursday 1-3 pm

Supplementary Materials
In this homework you will learn how to use Pig Latin, Hive and Cassandra. There are slides on eLearning to help with every of these tools. First, take a look at ConnectToPigHiveServer.pdf so that you know how to connect to UTD Hadoop servers.
To connect to server outside from college network first connect to cs1.utdallas.edu server and then
ssh the cs6360.utdallas.edu using **ssh <Netid>@cs6360** command. e.g.. ssh gga110020@cs6360, where gga110020 is my netid.

HDFS commands
Here are commands to work with Hadoop filesystem (HDFS).

List files:
```
hadoop fs -ls /
```
or
```
hadoop fs -ls /some/other/path
```

Create a directory in Hadoop filesystem:
```
hadoop fs -mkdir /xyz100200
```

Copy from local home directory to the hdfs:
```
hadoop fs -copyFromLocal my.dat /xyz100200/my.dat
```

And so on http://hadoop.apache.org/docs/r0.19.0/hdfs_shell.html

Instructions to work:
1) Create and use only one directory with net id per user on the cluster. (e.g /xyz100200)
2) Do not modify/add files in the **/SUMMER2015_HW2_PIG/** directory or other users directories (all logs are tracked)
3) Please, pay attention that the tasks depend on your NetID. Namely, first letter is denotes as <**L**>, first digit is denoted as <**X**>, and last digit is denoted as <Y>. For example, for TA NetID the values are: <NetId> = vmk130030 then <L> = v, <X>=1, <Y>=0

Dataset

We will use the datasets located under **/SUMMER2015_HW2_PIG/** in the HDFS in the Programming/Master
Node CS6360.utdallas.edu. Please use this folder and don't copy/modify any other folder on the server.

All dataset files are caret (^) separated.

The files are business2.csv, user2.csv, review2.csv.

The csv files has 24 columns, namely
Column id : Name of Column
Column 0 :review_id
Column 1: text
Column 2: business_id
Column 3: full_address
Column 4: schools
Column 5: longitude
Column 6: average_stars//this is for the business entity type only
Column 7: date
Column 8: user_id
Column 9: open
Column10: categories
Column11: photo_urlColumn12: city
Column13: review_count
Column14: name
Column15: neighborhoods
Column 16: url
Column 17: votes.cool
Column 18: votes.funny
Column 19: state
Column 20: stars:: //this is for review entity type only
Column 21: latitude
Column 22: type
Column 23: votes.useful

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** using the average ratings. This will require you to use review2.csv and business2.csv files.

Please answer the question by **calculating the average ratings** given to each business using the review2.csv file. Do not use the already calculated ratings (average_stars) contained in the business entity rows.

Q2:
Using Pig Latin script, Implement co-group command on business_id for the datasets **review** and **business.** Print first 5+<**X**> rows.

Q3:
Repeat Question 2 (implement join) with co-group commands. Print first 5+<**X**> rows.

Q4:
Write a UDF(User Define Function) FORMAT_CAT in Pig which basically formats the categories in business in the following:

Before formatting: ['Photographers', 'Event Planning & Services']
After formatting: 1) **['Photographers'|'Event Planning & Services']**
Before formatting: ['Print Media', 'Mass Media']
After formatting: 1) **['Print Media'| 'Mass Media']**

Using Pig Latin script, use the FORMAT_CAT function on business dataset and print the business_id, full_address and the categories **Limit your result to 10 rows.**

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 1
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: Hive

Dataset

The datasets are located under **/tmp/SUMMER_2015_HW3_HIVE/** in the Local UNIX file System. Please use this folder and don't copy to any other folder on the server. All datasets are caret (^) separated.
All dataset files are caret (^) separated.

The files are business2.csv, user2.csv, review2.csv.

The csv files has 24 columns, namely
Column id : Name of Column
Column 0 :review_id
Column 1: text
Column 2: business_id
Column 3: full_address
Column 4: schools
Column 5: longitude
Column 6: average_stars:: //this is for the business entity type only
Column 7: date
Column 8: user_id
Column 9: open
Column10: categories
Column11: photo_urlColumn12: city

Column13: review_count
Column14: name
Column15: neighborhoods
Column 16: url
Column 17: votes.cool
Column 18: votes.funny
Column 19: state
Column 20: stars:: //this is for review entity type only
Column 21: latitude
Column 22: type
Column 23: votes.useful

Q5:
List the business_id , full address of the **Top 10 businesses** using the average ratings. This will require you to use review2.csv and business2.csv files. (Show the create table command, load from local, and the Hive query).

Please answer the question by **calculating the average ratings** given to each business using the review data. Do not use the already calculated ratings (average_stars) contained in the business entity rows.

Q6:
Using Hive script, List the 'business_id' and 'categories' of businesses located in **'Stanford'**. (Show the create table command, load from local, and the Hive query).

Q7:
Dataset:
We will use the yelp datasets here. The datasets are located under **/tmp/HW_3_Summer_Data/partition/ (the file names are business2013.csv, business2014.csv and business2015.csv)** in hadoop file System. Please use these files to write your query. **The path contains three files for the partitioned year 2013, 2014 and 2015.** The datasets are **caret (^)** separated and each line has the following **24 columns as previously described.**

Requirement:
Using Hive script, create one table **partitioned** by year. (Show the create table with **one** command, load from local with **three** commands, and **one** Hive query that selects all columns from the table for the virtual column month of year 2013).

Q8:
Requirement:
Create three tables that have 3 columns each (namely business_id,full_address and longitude).

Each table will represent a year. The three years are 2013, 2014 and 2015.

Using Hive multi-table insert, insert values from columns (business_id,full_address and longitude) from **the table you created in Q7** to these three tables
(each table should have names in form of business_year e.g. business_2013 etc. for the specified year).

Q9:
Write a UDF(User Define Function) FORMAT_CAT in Pig which basically formats the categories in business in the following:

Before formatting: ['Photographers', 'Event Planning & Services']
After formatting: 1) **['Photographers'|'Event Planning & Services']**
Before formatting: ['Print Media', 'Mass Media']
After formatting: 1) **['Print Media'|'Mass Media']**
Using Hive script, use the FORMAT_CAT function on business dataset and print the business_id, full_address and the categories. Limit your result to 10 rows.

Submission: Please upload the following to eLearning:
- Script file for each Question as follows: Qx.pig or Qx.hive where x is the Question number.
- Text file with results of the script for each Question: Qx.res.
- Give a readme file for how to run the program.
- You will need to show your demo to TA.

Part 3: 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.

Q10: Cassandra CQL3
We will use the yelp user dataset . The dataset is located under **/tmp/SUMMER_2015_HW3_HIVE/** in the **local** file System. Please use **business2.csv** file under this folder. The dataset is caret(^) separated and each line has the following 24 columns: as shown previously.

{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 this 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 'axPZazfSZFnynOV52mbe2Q'
4- Delete all rows in the table.
5- Drop the table.

Q11: 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 Q10.
- One file with all commands for Q11.