CS 6301- Big Data Analytics and Management

Fall 2014

Homework/Assignment# 3

Due: Nov 12, 2014 (11:59 p.m.)

Teaching Assistant

Tatiana Erekhinskaya, email: txe110230@utdallas.edu
Office hours: Monday 1:00pm-3:00pm, Wednesday 1:00pm-3:00pm, ECSS 2.103B1

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.

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 /txe110230
```

Copy from local home directory to the hdfs:

```
hadoop fs -copyFromLocal my.dat /txe110230/my.dat
```

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

Task Customization

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> = txe110230
<L> = t

CS 6301- Big Data Analytics and Management

Fall 2014

Homework/Assignment# 3

Due: Nov 12, 2014 (11:59 p.m.)

<X> = 1
<Y> = 0

Part 1: Pig Latin

Dataset

We will use the datasets located under **/Fall2014_HW-3-Pig/** in the HDFS in the Programming/Master Node CS6360.utdallas.edu. Please use this folder and don't copy to any other folder on the server. All datasets are **semi-colon (;)** separated.
There are three modified files in the **/Fall2014_HW-3-Pig/** folder in HDFS with following format.

movies_new: **MovieID#Title#Genres**

ratings_new: **UserID#MovieID#Rating#Timestamp**

users_new: **UserID#Gender#Age#Occupation#Zip-code**

Q1:
Using Pig Latin script, list the unique userid of female users whose age between 20 - 35 and who has rated the highest rated Action AND War movies. (You should consider all movies that has Action **AND** War both in its genre list.) Print only users whose zip starts with **<X>**.

Consider average rating to calculate the highest rated movies. While finding the Action and War movies, you should count all users not only the female users.

Q2:
Using Pig Latin script, Implement cogroup command on UserID for the datasets **ratings_new** and **users_new.** Print first 10+**<X>** rows.

CS 6301- Big Data Analytics and Management

Fall 2014

Homework/Assignment# 3

Due: Nov 12, 2014 (11:59 p.m.)

Q3:

Repeat Question 2 (implement join) with cogroup commands. Print first 10+<**X**> rows.

Q4:

Write a UDF(User Define Function) FORMAT_GENRE in Pig which basically formats the genre in movies_new in the following:

Before formatting:	Children's
After formatting:	Children's <**NetId**>
Before formatting:	Animation|Children's
After formatting:	Children's & Animation <**NetId**>
Before formatting:	Children's|Adventure|Animation
After formatting:	Children's, Adventure & Animation <**NetId**>

Using Pig Latin script, use the FORMAT_GENRE function on movies_new dataset and print the movie name with its genre(s).

Part 2: Hive

Dataset

The datasets are located under **/tmp/Fall2014_HW-3_Hive/** in the **Local** UNIX System. Please use this folder and don't copy to any other folder on the server. All datasets are **semi-colon (;)** separated.

CS 6301- Big Data Analytics and Management

Fall 2014

Homework/Assignment# 3

Due: Nov 12, 2014 (11:59 p.m.)

There are three modified files in the **/tmp/Fall2014_HW-3_Hive/** folder in **Local** UNIX System with following format.

movies_new: **MovieID#Title#Genres**

ratings_new: **UserID#MovieID#Rating#Timestamp**

users_new: **UserID#Gender#Age;Occupation#Zip-code**

Q5:

Using Hive script, find top 10+<**Y**> average rated "Action" movies with descending order of rating. (Show the create table command, load from local, and the Hive query).

Q6:

Using Hive script, List all the movies with its genre where the movie genre is Action or Drama and the average movie rating is in between 4.4 - 4.9 and only the male users rate the movie. (Show the create table command, load from local, and the Hive query).

Q7:

Dataset:

We will use the movie datasets here. The datasets are located under **/tmp/HW_3_Data/partition (the file names are 2009, 2010 and 2011)** in the **Local** UNIX System. Please use this folder and don't copy to any other folder on the server. **The path contains three files for the partitioned**

years 2009, 2010 and 2011. The datasets are **semi-colon (;)** separated and each line has the following 3 columns **MovieID;Title;Genres**

Requirement:
Using Hive script, create one table **partitioned** by year. (Show the create table **one** command, load from local **three** commands, and **one** Hive query that selects all columns from the table for the virtual column year of 2009).

Q8:

Requirement:
Create three tables that have three columns each (MovieID, MovieName, Genre). Each table will represent a year. The three years are 2009, 2010 and 2011.

Using Hive multi-table insert, insert values from **the table you created in Q7** to these three tables (each table should have names of movies e.g. movies_2009 etc. for the specified year).

Q9:

Write a UDF(User Define Function) FORMAT_GENRE in Hive which basically formats the genre in movies_new in the following:

Before formatting:	Children's
After formatting:	Children's

Before formatting:	Animation|Children's
After formatting:	Animation, & Children's - <**NetId**>

CS 6301- Big Data Analytics and Management

Fall 2014

Homework/Assignment# 3

Due: Nov 12, 2014 (11:59 p.m.)

Before formatting: Adventure|Animation|Children's
After formatting: Adventure, Animation, & Children's - <**NetId**>

Using Hive script, use the FORMAT_GENRE function on movies_new dataset and print the movie name with its genre(s).

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 1.1.6 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-1.1.6
****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.

Dataset:

CS 6301- Big Data Analytics and Management

Fall 2014

Homework/Assignment# 3

Due: Nov 12, 2014 (11:59 p.m.)

We will use the IMDB movie dataset given in previous HWs. The dataset is located under under **/tmp/Fall2014_HW-3_Hive/** in the **Local** Unix System. Please use this folder and don't copy to any other folder on the server. The dataset is # separated and each line has the following 3 columns : MovieID,Title,Genres.

Q10: Cassandra CLI

{cs6360:~} /usr/local/apache-cassandra-1.1.6/bin/cassandra-cli --host csac0

Requirements:

Using Cassandra CLI, write commands to do the following:

1- Create a COLUMN FAMILY for this dataset.
2- Insert the following to the column family created in step 1. Use MovieID as the key.
   i. "70#From Dusk Till Dawn (1996)#Action|Comedy|Crime|Horror|Thriller"
   ii. "83#Once Upon a Time When We Were Colored (1995)#Drama"
   iii. "112#Escape from New York (1981)#Action|Adventure|Sci-Fi|Thriller" with time to live (ttl) clause after 300 seconds
3- Show the following:
   i. Get the movie name and genre for the movie id 70 ?
   ii. Retrieve all rows and columns.
   iii. Delete column Genres for the movie id 83.
   iv. Drop the column family.

4. Use describe keyspace command with your netid and show content.

Q11: Cassandra CQL3

{cs6360:~} /usr/local/apache-cassandra-1.1.6/bin/cqlsh –3 csac0

Requirements:

Using Cassandra CQL3, write commands to do the following:

CS 6301- Big Data Analytics and Management

Fall 2014

Homework/Assignment# 3

Due: Nov 12, 2014 (11:59 p.m.)

1- Create a table for this dataset. Use (MovieID) as the Primary Key.
2- Load all records in the dataset to this table.
3- Insert record “1162#New Comedy Movie#Comedy" to the table.
4- Select the tuple which has movie id 1150
5- Delete all rows in the table.
6- Drop the table.

Q12: 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 Q1.
- One file with all commands for Q2.
- One file with all commands for Q3.