CS 6301- Big Data Analytics and Management

Spring 2015

Homework/Assignment# 3

Due: Mar 24, 2015 (11:59 p.m.)

Teaching Assistant
Vishal Karande, email: vmk130030@utdallas.edu
Office hours: Tuesday 2:30 pm-4:00 pm,
Friday 2:30 pm-4:00 pm, Clark Center 1.202C

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 vmk130030@cs6360

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 **/Spring-2015-input/** 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 **/Spring-2015-input/** 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 **single-colon (:)** separated.

movies.dat: **MovieID:Title:Genres**
ratings.dat: **UserID:MovieID:Rating:Timestamp**
users.dat: **UserID:Gender:Age:Occupation:Zip-code**

Part 1: Pig Latin

Q1:

Using Pig Latin script, list the unique userid of male users whose age between 20 - 40 and who has rated the lowest rated Comedy AND Drama movies. (You should consider all movies that has Comedy **AND** Drama both in its genre list.)
Print only users whose zip starts with **<X>**. Consider average rating to calculate the lowest rated movies. While finding the Comedy and Drama movies, you should count all users not only the male users.

Q2:

Using Pig Latin script, Implement co-group command on MovieID for the datasets **ratings** and **movies.** 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_GENRE in Pig which basically formats the genre in movie in the following:

Before formatting: Children's
After formatting: 1) Children's **<NetId>**
Before formatting: Animation|Children's
After formatting: 1) Children's & 2) Animation **<NetId>**
Before formatting: Children's|Adventure|Animation
After formatting: 1) Children's, 2) Adventure & 3) Animation **<NetId>**
Using Pig Latin script, use the FORMAT_GENRE function on movies dataset and print the movie name with its genre(s).

Part 2: Hive

Q5:

Using Hive script, find top 10+<**X**> average rated "Comedy" 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 Comedy or Drama and the average movie rating is in between 4.5 - 4.6 (inclusive) 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 **/hive-partition-hw3/ (the file names are January.dat, February.dat and March.dat)** in hadoop file System. Please use these files to write your query. **The path contains three files for the partitioned months January, February and March.** 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 month. (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 month of March).

Q8:

Requirement:

Create three tables that have three columns each (MovieID, MovieName, Genre). Each table will represent a month. The three months are January, February and March.

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_march etc. for the specified month).

Q9:

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

Before formatting: Children's
After formatting: 1) Children's <**NetId**> :hive
Before formatting: Animation|Children's
After formatting: 1) Children's & 2) Animation <**NetId**> :hive
Before formatting: Children's|Adventure|Animation
After formatting: 1) Children's, 2) Adventure & 3) Animation <**NetId**> :hive

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.

We will use the IMDB user dataset given in previous HWs. The dataset is located under **/cassandra-input/** in the **hadoop** file System. Please use users.dat file under this folder. The dataset is **:** separated and each line has the following 5 columns:
UserID:Gender:Age:Occupation:Zip-code

Q10: Cassandra CLI

{cs6360:~} /usr/local/apache-cassandra-2.0.5/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 UserID as the key.
   i. "13:F:51:1:93334"
   ii. "1471:F:31:17:11116"
   iii. "1496:F:31:17:94118" with time to live (ttl) clause after 300 seconds

3- Show the following:
   i. Get the Gender and Occupation for user with id 13 ?
   ii. Retrieve all rows and columns.
   iii. Delete column Gender for the user id 1471.
   iv. Drop the column family.

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

Q11: Cassandra CQL3

{cs6360:~} /usr/local/apache-cassandra2.0.5/bin/cqlsh –3 csac0

Requirements:
Using Cassandra CQL3, write commands to do the following:

1- Create a table for this dataset. Use (UserID) as the Primary Key.
2- Load all records in the dataset to this table.
3- Insert record "6041:M:32:6:11120" to the table.
4- Select the tuple which has user id 6020
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 Q10.
- One file with all commands for Q11.
- One file with all commands for Q12.