CS 6301- Big Data Analytics and Management

Spring 2014

Homework/Assignment# 3

Due: March 24, 2014 (11:59 p.m.)

In this homework you will learn how to use Pig Latin and Hive.

Part 1: Pig Latin

Dataset

We will use the datasets located under **/Spring2014_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 **/Spring2014_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 - 30 and who has rated the highest rated Action AND War movies. (You should consider all movies that has at Action **AND** War both in its genre list.)

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.**

CS 6301- Big Data Analytics and Management

Spring 2014

Homework/Assignment# 3

Due: March 24, 2014 (11:59 p.m.)

Q3:

Repeat Question 2 (implement join) with cogroup commands.

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
Before formatting:	Animation|Children's
After formatting:	Children's and Animation
Before formatting:	Adventure|Animation|Children's
After formatting:	Children's, Adventure and Animation

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

CS 6301- Big Data Analytics and Management

Spring 2014

Homework/Assignment# 3

Due: March 24, 2014 (11:59 p.m.)

Part 2: Hive

Dataset

The datasets are located under **/tmp/Spring2014_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.

There are three modified files in the **/tmp/Spring2014_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 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 - 4.7 and only the male users rate the movie. (Show the create table command, load from local, and the Hive query).

CS 6301- Big Data Analytics and Management

Spring 2014

Homework/Assignment# 3

Due: March 24, 2014 (11:59 p.m.)

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).

CS 6301- Big Data Analytics and Management

Spring 2014

Homework/Assignment# 3

Due: March 24, 2014 (11:59 p.m.)

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, and Children's
Before formatting:	Adventure|Animation|Children's
After formatting:	Adventure, Animation, and Children's

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.
- Give a readme file for how to run the program.
- You will need to show your demo to TA.