CSE427 – Makeup Homework

M. Neumann

Due THU 04/21/2016 10am

Getting Started

Update your SVN repository. Find instructions on how to checkout, update, and commit to your SVN repository here: `http://sites.wustl.edu/neumann/resources/cse427s_resources/`

When needed, you will find additional materials for *homework x* in the folder `hwx`. So, for the current assignment the folder is `hwM`.

Hint: You can **check your submission** to the SVN repository by viewing `https://svn.seas.wustl.edu/repositories/<yourwustlkey>/cse427s_sp16` in a web browser (mind browser caching!).

Run this command in the command line, **WHENEVER you (RE)START your VM**:

```
$ ~/scripts/analyst/toggle_services.sh
```

This makeup homework may replace your lowest score homework (if your score will be higher than your lowest homework score).

Note that Problem 1-3 are very similar to Lab 7.

Problem 1: HIVE Table Import – The Easy Way (20%)

You used SQOOP in an earlier lab to import data from MySQL into HDFS. SQOOP can also create a HIVE table with the same fields as the source table in addition to importing the records, which saves you from having to write a CREATE TABLE statement.

(a) Use SQOOP with the option `--hive-import` (this option has no directory parameter and replaces the `--warehouse-dir` option!) to import the table `suppliers` from MySQL into HIVE (Hint: check **Lab 6, Part II**.) Provide your command.

(b) It is always a good idea to validate the data you have added. Use a HIVE statement to determine the number of suppliers in Missouri. Provide your command and report the result.

(c) Provide a HIVE statement that lists the number of suppliers in each state.

Submit your answers by editing the `hwM.txt` file.

Problem 2: Table Creation – Adding HIVE Support (30%)

You imported data from the `employees` table from MySQL into in HDFS in **Lab 6, Part II**.

(a) Discuss why it is not possible to use this data straightaway in HIVE?

(b) Write a HiveQL statement to create a HIVE table for this data at its current location (the data should not be moved!). Provide your command. Hints:

- Find the field information for this table here: `http://www.cse.wustl.edu/~m.neumann/sp2016/cse427/protected/Labs_Hw_DataModelRef.pdf`.
- You might want to look at the import command you used in Lab 6 to find any other necessary information.

Name a reason/scenario why the data could not be moved?

(c) What are the top four most common job titles at *Dualcore*?

Submit your answers by editing the `hwM.txt` file.

Problem 3: HIVE Table Creation – The Full Process (30%)

Customer ratings and feedback are great sources of information for both customers and retailers like *Dualcore*. One of *Dualcore*'s new projects is to analyze their customers' reviews and product ratings. In this problem you will prepare the data and create the needed HIVE table.

(a) Explain what is meant by the following statement: "In HIVE data location and schema are decoupled". Describe a scenario/use-case where this is advantageous?

(b) Create a table named `ratings` with the following fields (data types in parentheses): `posted` (TIMESTAMP), `cust_id` (INT), `prod_id` (INT), `rating` (TINYINT), `message` (STRING). Provide your command and describe what this command does. Why is it beneficial to use TINYINT instead of INT? What other data types for numeric values does HIVE offer?

(c) Now, you need to put the data into the *right location* in HDFS . Provide the path of this location (Hint: it has to match the command you used in the previous part.). Describe two ways of populating your HIVE table with data from your local file system. Include the respective commands in your answer.

(d) What command would you use to rename the `message` column in the `ratings` table to `review`. How can you verify the change? (Once you verfied that your command works, rename the column back to message.) Provide the both HIVE commands.

(e) What command would you use to delete the `ratings` table? Do **not** execute this command, if you are working final project 2! Provide the command and explain what exactly got deleted from where.

Submit your answers by editing the `hwM.txt` file.

Problem 4: Back to PIG – Calculate Click-Through Rate (20%)

The calculations you did in `hw7` provided a rough idea about the success of the ad campaign at *Dualcore*, but didn't account for the fact that some sites display *Dualcore's* ads more than others. This makes it difficult to determine how effective their ads were by simply counting the number of clicks on one site and comparing it to the number of clicks on another site. One metric that would allow Dualcore to better make such comparisons is the Click-Through Rate (`https://en.wikipedia.org/wiki/Click-through_rate`), commonly abbreviated as CTR. This value is simply the percentage of ads shown that users actually clicked, and can be calculated by dividing the number of clicks by the total number of ads shown.
The directory for this problem is:

```
~/training_materials/analyst/exercises/analyze_ads/bonus_03
```

(a) Edit the `lowest_ctr_by_site.pig` file and implement the following:

- Within the nested FOREACH, filter the records to include only records where the ad was clicked.
- Create a new relation on the line that follows the FILTER statement which counts the number of records within the current group.
- Add another line below that to calculate the CTR in a new field named `ctr`.
- After the nested FOREACH, sort the records in ascending order of CTR and display the first three to the screen.

(b) Once you have made these changes, try running your script against the data in HDFS. Which three sites have the lowest click through rate?

(c) Modify your script to display the three keywords with the highest CTR. Provide your commands. Which keywords are those?

Submit your answers to (a) by adding the `lowest_ctr_by_site.pig` to the `hwM` folder in your SVN repository. Submit your answer to (b) and (c) by editing the **`hwM.txt`** file in your SVN repository.

To add the .pig file to your SVN repo and commit your work run:

```
$ svn add lowest_ctr_by_site.pig
$ svn commit -m 'hwM submission' .
```