CSE427 – Homework 7

M. Neumann

Due THU 03/31/2016 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 `hw7`.

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!).

Good News

In this homework you will help your new employer – *Dualcore Inc.* – to **save money** and **attract new customers** by writing PIG scripts that analyze data from two online ad networks to optimize advertising. If you haven't heard about your new job, find the instructions for Lab 5 on the course webpage and catch up!

Preparation (not for credit - but for your convenience)

If you haven't done so yet, complete **Lab 5: PIG for ETL** as you will need the data for this homework.

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

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

Indicating Group Work

Use the file `partners.txt` to indicate group work. Follow these instructions exactly, to ensure to get credit!

- `partners.txt` needs to include up to 2 wustlkeys in the first two lines (one line per wustlkey)
- **first line/wustlkey is the repository, where the solution is located**. We will **only** consider the submission in this repository!

- Every student in a group needs to have **the same** `partners.txt` in the hwx folder in their repository (indicating that the partnership are **mutually accepted**)!
- If you do not have a partner, try to find one. If you want to submit on your own, indicate your wusltkey in the first line of `partners.txt` and leave the second line blank.

Problem 1: Find Low Cost Sites (50%)

Both ad networks charge a fee only when a user clicks on *Dualcore's* ad. This is ideal for *Dualcore* since their goal is to bring new customers to their site. However, some sites and keywords are more effective than others at attracting people interested in the new tablet being advertised by *Dualcore*. With this in mind, you will begin by identifying which sites have the lowest total cost.

The directory for this homework is:

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

(a) Obtain a local subset of the ad data stored in the `dualcore` folder in HDFS. This test data should comprise the first 100 lines of all parts of `ad_data1`. Store this data in `test_ad_data.txt`. Provide your command in the **`hw7.txt`** file in your SVN repository. (HINT: this can be achieved with UNIX commands; PIG is not needed.)

(b) It is way faster to test PIG scripts by using a local subset of the input data. Describe why this is the case.
Note: Creating local data subsets will not be listed as parts of the actual problems in upcoming homeworks. However, doing so will help you do the problems more quickly!

(c) Edit the PIG script `low_cost_sites.pig` to perform the following operations:

- Modify the LOAD statement to read the sample data generated in the previous part.
- Add a line that creates a new relation to include only records where `was_clicked` has a value of 1.
- Group this filtered relation by the `display_site` field.
- Create a new relation that includes two fields: the `display_site` and the total cost of all clicks on that site.
- Sort that new relation by cost (in ascending order).
- Display just the first four records to the screen.

Test your script locally against the sample data `test_ad_data.txt`. What gets displayed on the screen?

(d) Run your script against the full data in HDFS. To achieve this comment out the LOAD statement and add a new LOAD statement using the path with a file glob loading both ad data sets (`ad_data1` and `ad_data2`) simultaneously. Which four sites have the lowest overall cost?

Submit your answer to (a), (b), and parts of (c) and (d) by editing the **`hw7.txt`** file in your SVN repository. Submit your answers to (c) and (d) by adding the `low_cost_sites.pig` to the `hw7` folder in your SVN repository.

To add the .pig file to your SVN repo before committing run:

```
$ svn add low_cost_sites.pig
$ svn commit -m ’hw7 submission’ .
```

Problem 2: Find High Cost Keywords (20%)

The terms users type when doing searches may prompt the site to display a *Dualcore* advertisement. Since online advertisers compete for the same set of keywords, some of them cost more than others. You will now write some PIG Latin to determine which keywords have been the most expensive for *Dualcore* overall.

(a) Write a PIG script called `high_cost_keywords.pig` that groups the ad data by `keyword` and sorts it in descending order of cost.

(b) Which three keywords have the highest overall cost? Modify `high_cost_keywords.pig` to display these top-3 results on the screen.

Submit your answers to (a) and (b) by adding the `high_cost_keywords.pig` to the `hw7` folder in your SVN repository. Submit your answer to (b) by editing the **`hw7.txt`** file in your SVN repository.
To add the .pig file to your SVN repo before committing run:

```
$ svn add high_cost_keywords.pig
$ svn commit -m ’hw7 submission’ .
```

Problem 3: Count Ad Clicks (10%)

One important statistic we haven’t yet calculated is the total number of clicks the ads have received. Doing so will help the marketing director plan the next ad campaign budget.

(a) Edit the PIG script `total_click_count.pig` in the `bonus_01` folder to perform the following operations:

- Group the records (filtered by `was_clicked == 1`) so that you can call the aggregate function in the next step.
- Invoke the `COUNT` function to calculate the total of clicked ads (Hint: Because we should not have any null records, you can use the `COUNT` function instead of `COUNT_STAR`, and the choice of field you supply to the function is arbitrary).
- Display the result to the screen

(b) How many clicks did we receive?

Submit your answer for (a) adding the `total_click_count.pig` to the `hw7` folder in your SVN repository and for (b) by editing the **`hw7.txt`**.
To add the .pig file to your SVN repo before committing run:

```
$ svn add total_click_count.pig
$ svn commit -m ’hw7 submission’ .
```

Problem 4: Estimate the Maximum Cost of the Next Ad Campaign (20%)

When you reported the total number of clicks, the Marketing Director said that the goal is to get about three times that amount during the next campaign. Unfortunately, because the cost is based on the site and keyword, it is not clear how much to budget for that campaign. You can help by estimating the worst case (most expensive) cost based on 50,000 clicks. You will do this by finding the most expensive ad and then multiplying it by the number of clicks desired in the next campaign.

(a) Create a PIG script called `project_next_campaign_cost.pig` to get the maximum value in `cpc` and multiply this by the total number of clicks we expect to have in the next campaign. Display the result on the screen.

(b) What is the maximum cost you expect for this campaign?

Submit your answer to (a) by adding the `project_next_campaign_cost.pig` to the `hw7` folder in your SVN repository. Submit your answer to (b) by editing the **`hw7.txt`** file in your SVN repository. **To add the .pig file to your SVN repo before committing run:**

```
$ svn add project_next_campaign_cost.pig
$ svn commit -m ’hw7 submission’ .
```

Bonus Problem (5% up to a max. of 100%) - no group work!

Write a review for this homework and store it in the file `hw7_review.txt` provided in your SVN repository (and commit your changes). This file should only include the review, **no other information** such as name, wustlkey, etc. Remember that you are not graded for the content of your review, solely it’s completion.

You can only earn bonus points if you write **at least 50 words**. Bonus points are given to the **owner of the repository only** (no group work!).