CS 410/510 Cloud and Cluster Data Management

Spring 2018 Quarter

Assignment 5 Pig Latin and HiveQL

Due: Thursday, 31 May 2018 at the beginning of class

You should do this assignment individually or in pairs of two students. (Pairs can be from different sections.) You should only talk to the instructors or your partner about this assignment. You may also post questions and comments to the course discussion list on Piazza.

Please turn in your completed assignments on paper. Put your last name, first name, the assignment number in that order in the first line of your assignment.

This assignment involves writing queries in Pig Latin or HiveQL (your choice).
Information on Pig Latin:
http://pig.apache.org/docs/r0.17.0/

Information on HiveQL:
https://cwiki.apache.org/confluence/display/Hive/LanguageManual

The questions below concern a dataset `SearchClicks` with the following structure:

```
{(searchTerm: String, topLinks: Map<url: string → clicks: int>)}
```

That is, it is a set of records, each containing a search string plus a map (associative array) that takes a URL to the number of clicks on that URL for that search.

Question 1 (50 points): Give a Pig Latin or HiveQL program that figures out for which searches did `pdx.edu` have more than 5% of the total clicks, and what was the percentage. (You don't need to actually run the program.)

Provide comments on what each statement in your program does. If you use a User-Defined Function, explain what that function does and give

pseudocode.

Question 2 (50 points): Show what your program does on the following example data. Include the intermediate result of each statement.

```
{
  (searchTerm: 'Portland',
   topClicks: <

    portlandonline.com → 131

    en.wikipedia.org → 96

    travelportland.com → 96

    flypdx.com → 81

    portlandmarathon.com → 9 >),

  (searchTerm: 'Portland University',
   topClicks: <

    up.edu → 76

    pdx.edu → 53

    portlandpilots.com → 45

    en.wikipedia.org → 12

    pdx.uoregon.edu → 9 >),
  (searchTerm: 'Portland Computer',
   topClicks: <

    freegeek.org → 21

    enunic.com → 19

    pacficsolutions.com → 19

    1201.com → 12

    boltportland.com → 9 >),
  (searchTerm: 'Portland Vikings',
   topClicks: <
```

```
 goviks.com → 36

 en.wikipedia.org → 31

 go.espn.com → 22

 pdx.edu → 3

 oregonlive.com → 3 >),
(searchTerm: ‘Portland Degrees’,
 topClicks: <
 twodegrees.com → 52

 pcc.edu → 30

 pdx.edu → 11

 up.edu → 6

 artinstitutes.com → 2 >)}
```