Cloud Computing Labs: Pig Programming for Relational Data Analysis

The primary purpose of this assignment is to get familiar with Pig programming for analyzing relational data. This project consists of two parts: the first part is to work with the pig tutorial; the second part will apply Pig Latin to answer a few queries on the given data. Note: The pig executable has been installed at /usr/local/pig/ on nimbus17.cs.wright.edu. You can check /home/hadoop/.profile for the setting of environment variables.

Part 1: Getting Familiar with Pig/Pig Latin

To begin with, you need to check the following items.

- First, make sure the environment variables are setup. In the file .profile under your home directory, you should have the following items:

```
export PIG_HOME=/usr/local/pig/
export PIG_CLASSPATH=/usr/local/hadoop/conf
export HADOOP_CONF_DIR=/usr/local/hadoop/conf
```

- Next, work through the "Getting Start" tutorial and make sure you are familiar with the basic environment.
- Read the details of the Pig Latin Language and the lecture slides. Make sure you fully understand the usage of LOAD, STORE, DUMP, FILTER, FOREACH-GENERATE, GROUP, JOIN, DISTINCT, ORDER, LIMIT, which might be used in solving the problems in the project.

Now, answer the following questions:

Question 1.1 The GROUP-BY operator groups together tuples that have the same group key (key field). What is the default name for the group key in the grouping result?

Question 1.2 Assume the input file is a text document. Please explain what the following piece of code does, line by line. What is the output of the script?

```
A = load './input.txt';
B = foreach A generate flatten(TOKENIZE((chararray)$0)) as token;
C = group B by token;
D = foreach C generate group, COUNT(B);
store D into './output';
```

Question 1.3 How much time did you spend on the task 1.1 and 1.2?

Question 1.4 How useful is this task to your understanding of the Pig Latin language?

Part 2: Analyzing Book Purchasing Records with Pig

In this task, you will use Pig to solve the analytic problems for a set of linked tables. Now we have three data files with the following schema: customer(cid, name, age, city, sex), book(isbn, name), and purchase(year, cid, isbn, seller, price), where purchase.cid is the foreign key to customer.cid and purchase.isbn is the foreign key to book.isbn. These datasets are also in nimbus17: /home/hadoop/project2/. The fields in the dataset are separated by "\t". Please try to use pig scripts to answer the following queries. The pig local mode is recommended for better performance:

```
pig -x local
```

Question 2.1 How much did each seller earn? Write down the code and the result here.

Question 2.2 How much did each family spend on books? Assume all users with the same last name are from the same family. Implement your own UDF that extracts the last name from the full name. You will need to learn how to implement UDF with Java. Write down the code and the result here.

Question 2.3 Find the names of the books that Amazon gives the lowest price among all sellers (You should exclude the case that Amazon is the only seller; it counts that Amazon and other sellers give the same price). Write down the code and the result here.

Question 2.4 Who also bought ALL the books that Harry bought? Write down the code and the result here.

Question 2.5 How much time did you spend on each of the tasks 2.1-2.4, respectively?

Question 2.6 How useful is this task to your understanding of Pig programming?

Final Survey Questions

Question 3.1 Your level of interest in this lab exercise (high, average, low);

Question 3.2 How challenging is this lab exercise? (high, average, low);

Question 3.3 How valuable is this lab as a part of the course (high, average, low).

Question 3.4 Are the supporting materials and lectures helpful for you to finish the project? (very helpful, somewhat helpful, not helpful);

Question 3.5 How much time in total did you spend in completing the lab exercise;

Quertion 3.6 Do you feel confident on applying the skills learned in the lab to solve other problems?

Deliverables

Turn in the PDF report that answers all the questions. The code should be embedded for questions 2.1-2.4.

This page, first created: 2 Oct, 2014; last updated: 2 Oct, 2014