CSSE 434 – Lab 5 – Answers

Christopher Keinsley

Question 1. Pig introduced the notion of a macro starting with Pig 0.9.1. What is a macro, explain it with an example?

Macros are a way of replacing small parts of pig latin with a single command. The macro allows users to, essentially, create functions for performing various operations on dynamic data. For instance, say someone wanted to count the number of rows in a student table. In SQL one could just do 'Select COUNT(*) from student;'. In Pig, this takes four lines,

'A = load 'student.txt' as (name, student, gpa);
B = group A all;
C = foreach B generate COUNT(A); **
dump C;'

However, with the addition of macros, this can be condensed to 'DEFINE row_count(X) RETURNS Z { Y = group $X all; $Z = foreach Y generate COUNT($X); };' This is much more nicer to look at, and it can be used for any set of data. Pig macros allow us to create functions that do common tasks for us easily without repeating a bunch of code.

Question 2. Explain the following commands/functions with an example.

a. COGROUP
    a. COGROUP used on a single table is just a regular group by. However, if you use COGROUP on more than one table, it will create a row with bags of matching rows from all tables up to 127 at once. This could then be used to perform a join on the grouped column. Example, imagine we had a data set of animal owners and a data set of animal names. We could cogroup on the type of animal if we wanted to see which people might have which pet names.
    b. C = COGROUP B by B.field, A by A.field;
b. RANK
    a. RANK assigns an ordinal number to every tuple in a relation. This can be done with exact ranks, each element with the same sort value gets the same number, or dense rank, where elements of the same sort value get consecutive numbers. We could sort students by GPA, then rank them to get their class rank. The rank number ends up appended to the front of the tuple when we do this.
    b. B = RANK A BY field;
c. STREAM
    a. STREAM can be used to send the data through an external script or program. We could use pig to clean the data and then send it to some other program for storing or processing via the STREAM command.
    b. A = STREAM stuff THROUGH someOtherThing;