

Lab 4 - Apache Pig

Name:
ID:


1. Read and setup your Hadoop machine environment according to the lab 3 setup guide in CANVAS.
2. Login into your machine with Hadoop and open up a terminal (e.g. ctrl+alt+t)
3. Issue the shell command “jps”. What is it? You could search for its meaning on the web.
4. Check if you have the necessary environment for running Hadoop.
5. Fill in the following table by navigating the related information on the web.

Shell Command	Meaning
>start-dfs. sh	
>start-yarn. sh	
>mr-jobhistory-daemon. sh start historyserver	

6. You may use the above commands for helping you setup the Hadoop environment.
7. In the terminal, type “wget www.cs.toronto.edu/~wkc/emp_dept.tar.gz” to get the data file. Alternatively, you can transfer the data file from CANVAS to your Hadoop environment.
8. Decompress the data file by issuing the command “tar xzf emp_dept.tar.gz”.
9. Put the data into the HDFS for Hadoop by issuing the command “hdfs dfs -put emp_dept”.
10. Once you have put the data, you can go into the Apache Pig environment by typing “pig”.
11. Load your data into the Apache Pig environment; for example, fill in the following table

Apache Pig Statement	Meaning
emp = LOAD 'ex_data/emp_dept/emp.csv' AS (empno:INT, ename:CHARARRAY, job:CHARARRAY, mgr:INT, hiredate:DATETIME, sal:FLOAT, deptno:INT);	
dept = LOAD 'ex_data/emp_dept/dept.csv' AS (deptno:INT, dname:CHARARRAY, loc: CHARARRAY);	
salgrade = LOAD 'ex_data/emp_dept/salgrade.csv' AS (grade:INT, losal:INT, hisal:INT);	

12. *****Please feel free to ignore the version depreciation messages. It will not affect your results.*****
13. Once you have loaded the data, you can type the dump commands for testing (e.g. dump emp;)
14. If you get 14 rows from emp, 4 rows from dept, and 5 rows from salgrade then you are doing fine.

- 
15. Write down the Apache Pig statement(s) to get Smith's employment date.

 16. Write down the Apache Pig statement(s) to get Ford's job title.

 17. Write down the Apache Pig statement(s) to get the first employee (by the hiredate).

 18. Write down the Apache Pig statement(s) to get the number of employees in each department.

 19. Write down the Apache Pig statement(s) to get the number of employees in each city.

 20. (Optional) If you are interested, you could try to see if you can write down the Apache Pig statement(s) for getting the following data outputs. At the end, you can wrap all Apache Pig statements in this lab sheet into a single file "emp_dept.pig" which can be executed by typing "pig -x mapreduce emp_dept.pig" in the shell command of your machine.
 - (1) The average salary in each city.
 - (2) The highest paid employee in each department
 - (3) The managers whose subordinates have at least one subordinate
 - (4) The number of employees for each hiring year
 - (5) The pay grade of each employee
 21. This is the end; please also upload this sheet with your answers to the submission system.
- 