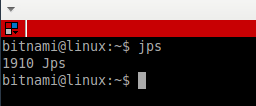
|  |  |
| --- | --- |
| Lab 4 – Apache Pig | Name: LOW, Zhi HaoID: 54924670 |

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.



The jps(Java Virtual Machine Process Status Tool) command lists the instrumented Java HotSpot VMs on the target system. The command is limited to reporting information on JVMs for which it has the access permissions.

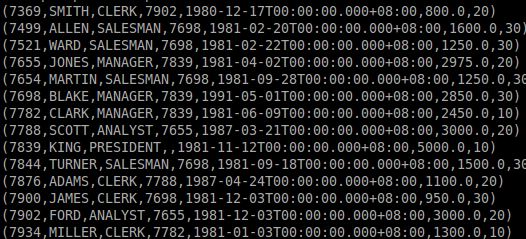
1. Check if you have the necessary environment for running Hadoop.
2. Fill in the following table by navigating the related information on the web.

|  |  |
| --- | --- |
| **Shell Command** | **Meaning** |
| >start-dfs.sh | Starts the Hadoop DFS daemons, the namenode and datanodes. |
| >start-yarn.sh | Starts Yet Another Resource Negotiator(YARN) daemons and NodeManager daemons. |
| >mr-jobhistory-daemon.sh start historyserver | Start the MapReduce JobHistory Server |

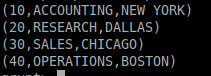
1. You may use the above commands for helping you setup the Hadoop environment.
2. In the terminal, type “wget [www.cs.toronto.edu/~wkc/emp\_dept.tar.gz](http://www.cs.toronto.edu/~wkc/emp_dept.tar.gz)” to get the data file. Alternatively, you can transfer the data file from CANVAS to your Hadoop environment.
3. Decompress the data file by issuing the command “tar xzf emp\_dept.tar.gz”.
4. Put the data into the HDFS for Hadoop by issuing the command “hdfs dfs -put emp\_dept”.
5. Once you have put the data, you can go into the Apache Pig environment by typing “pig”.
6. 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); | Loading employee data from 'ex\_data/emp\_dept/emp.csv' and assigning the type to the column |
| dept = LOAD 'ex\_data/emp\_dept/dept.csv' AS (deptno:INT, dname:CHARARRAY, loc: CHARARRAY); | Loading department data 'ex\_data/emp\_dept/dept.csv' and assigning the type to the column |
| salgrade = LOAD 'ex\_data/emp\_dept/salgrade.csv' AS (grade:INT, losal:INT, hisal:INT); | Loading the sales data  'ex\_data/emp\_dept/salgrade.csv' and assigning the type to the column |

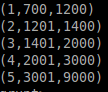
1. **\*\*\*Please feel free to ignore the version depreciation messages. It will not affect your results**.\*\*\*
2. Once you have loaded the data, you can type the dump commands for testing (e.g. dump emp;)
3. If you get 14 rows from emp, 4 rows from dept, and 5 rows from salgrade then you are doing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



DUMP emp;



DUMP dept;



DUMP salgrade;

1. Write down the Apache Pig statement(s) to get Smith’s employment date.
2. Write down the Apache Pig statement(s) to get Ford’s job title.
3. Write down the Apache Pig statement(s) to get the first employee (by the hiredate).

grunt> result1 = FOREACH emp GENERATE hiredate;

grunt> result2 = ORDER result1 BY hiredate ASC;

grunt> result3 = LIMIT result2 1;

1. Write down the Apache Pig statement(s) to get the number of employees in each department.
2. Write down the Apache Pig statement(s) to get the number of employees in each city.
3. (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.
4. The average salary in each city.
5. The highest paid employee in each department
6. The managers whose subordinates have at least one subordinate
7. The number of employees for each hiring year
8. The pay grade of each employee
9. This is the end; please also upload this sheet with your answers to the submission system.