|  |  |  |
| --- | --- | --- |
| Lab 4 – Apache Pig | Name: SHUYUE JIAID: 56846018 |  |

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.

~ » jps

2265 Jps

607

Answer:

jps is the Java Virtual Machine Process Status Tool. It is the “host identifier of the host for which the process report should be generated. The hostid may include optional components that indicate the **communications protocol**, **port number**, and other **implementation specific data**.” [1]

Reference:

[1] https://docs.oracle.com/javase/7/docs/technotes/tools/share/jps.html

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 | Start the process of HDFS (for Storage), i.e., start NameNode, DataNode, and secondaryNameNode processes. |
| >start-yarn.sh | Start the process of YARN (for resources scheduling). |
| >mr-jobhistory-daemon.sh start historyserver | Start the process to monitor and check previous (history) work or processes. |

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); | The formatted (AS) emp.csv file is loaded (LOAD) into Apache Pig framework. The format is set with type INT, ename and job CHARARRAY, date DATETIME, etc. Then the data is assigned to emp variable. |
| dept = LOAD 'ex\_data/emp\_dept/dept.csv' AS (deptno:INT, dname:CHARARRAY, loc: CHARARRAY); | The dept.csv file is loaded (LOAD) into the Pig with formatted data type (AS). Then the data is assigned to dept variable. |
| salgrade = LOAD 'ex\_data/emp\_dept/salgrade.csv' AS (grade:INT, losal:INT, hisal:INT); | The salgrade.csv file is loaded (LOAD) into the Pig with formatted data type (AS). Then the data is assigned to salgrade variable. |

1. **\*\*\*Please feel free to ignore the version depreciation messages. It will not affect your results**.\*\*\*

Okay!

1. Once you have loaded the data, you can type the dump commands for testing (e.g. dump emp;)
2. If you get 14 rows from emp, 4 rows from dept, and 5 rows from salgrade then you are doing fine.
3. Write down the Apache Pig statement(s) to get Smith’s employment date.

employee\_smith = FILTER emp BY ename == 'smith';

employee\_date = FOREACH employee\_smith GENERATE ename, hiredate;

DUMP employee\_date;

1. Write down the Apache Pig statement(s) to get Ford’s job title.

employee\_ford = FILTER emp BY ename == 'ford';

employee\_ford\_title = FOREACH employee\_ford GENERATE ename, job;

DUMP employee\_ford\_title;

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

hire\_date\_temp = FOREACH emp GENERATE hiredate;

hire\_date = ORDER hire\_date\_temp BY hiredate ASC;

first\_employee\_date = LIMIT hire\_date 1;

first\_employee = JOIN emp BY hiredate, first\_employee\_date BY hiredate;

name = FOREACH first\_employee GENERATE $0, $1, $2, $4;

DUMP name;

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

dept = GROUP emp BY deptno;

num = FOREACH dept GENERATE group AS deptno, COUNT(emp) AS empCnt;

DUMP num;

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

dept\_name = GROUP emp BY deptno;

deptEmp = FOREACH dept\_name GENERATE group AS deptno, COUNT(emp) AS count;

cityEmp = JOIN dept\_name BY deptno, dept BY deptno;

city = GROUP cityEmp BY loc;

city\_emp\_num = FOREACH city GENERATE group AS city, SUM(cityEmp.count) AS number;

DUMP city\_emp\_num;

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