Pig Basis

1. Load Customer records:

a. cust = LOAD '/input/custs' using PigStorage(',') AS (custid:chararray, firstname:chararray, lastname:chararray, age:long, profession:chararray);

2. Select no. of records

a. amt = LIMIT cust 100;
b. dump amt;

3. Group customer records by profession

a. groupbyprofession = GROUP cust BY profession;
b. describe groupbyprofession;
c. groupbyboth = GROUP cust BY (profession,age);
d. describe groupbyboth;

4. Count no of customers by profession

a. Countbyprofession = FOREACH groupbyprofession GENERATE group, COUNT(cust);
b. Dump countbyprofession;
c. Note: group is the name of column in the BY section [C] used earlier.

d. Countbyage = FOREACH groupbyboth GENERATE group.age, COUNT(cust);

5. How to work with the sample log file:

1. log = LOAD '/input/sample.log' using TextLoader();

2. LEVELS = foreach log generate REGEX_EXTRACT($0,'(TRACE|DEBUG|INFO|WARN|ERROR|FATAL)',1) as LOGLEVEL;

3. FILTEREDLEVELS = FILTER LEVELS by LOGLEVEL is not NULL;

   GROUPEDLEVELS = GROUP FILTEREDLEVELS by LOGLEVEL;
4. FREQ = foreach GROUPEDLEVELS generate group as LOGLEVEL, COUNT(FILTEREDLEVELS.LOGLEVEL) as COUNT;

5. RESULT = order FREQ by COUNT desc;

6. STORE RESULT into '/output/pig_log1';