

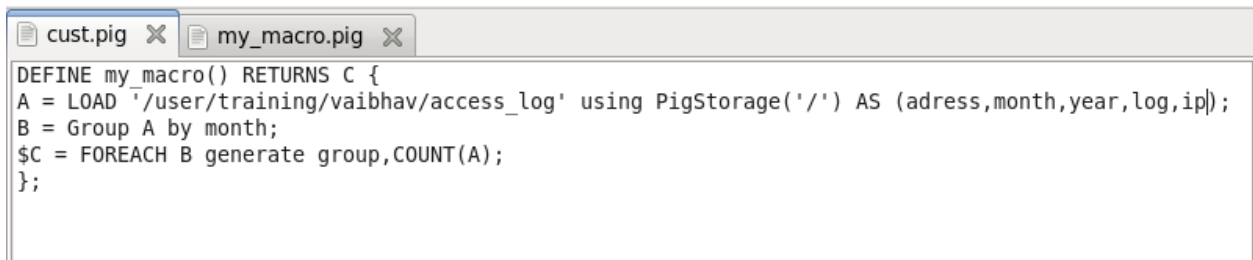
Assignment

Web log + Cust Analysis using Pig

Date:07-09-2022

Find Out the hit count per month

```
DEFINE my_macro() RETURNS C {  
A = LOAD '/user/training/vaibhav/access_log' using PigStorage('/') AS  
(adress,month,year,log,ip);  
B = Group A by month;  
$C = FOREACH B generate group,COUNT(A);  
};
```

A screenshot of a Pig script editor window. The window has two tabs: 'cust.pig' and 'my_macro.pig'. The 'my_macro.pig' tab is active, showing the following Pig script:

```
DEFINE my_macro() RETURNS C {  
A = LOAD '/user/training/vaibhav/access_log' using PigStorage('/') AS (adress,month,year,log,ip);  
B = Group A by month;  
$C = FOREACH B generate group,COUNT(A);  
};
```

```
import '/desktop/cust.pig';  
Count = my_macro();  
dump count;
```

```
grunt> import '/home/training/Desktop/cust.pig';  
2022-09-07 10:01:15,370 [main] WARN org.apache.hadoop.conf.Configuration - dfs.  
df.interval is deprecated. Instead, use fs.df.interval  
2022-09-07 10:01:15,370 [main] WARN org.apache.hadoop.conf.Configuration - dfs.  
max.objects is deprecated. Instead, use dfs.namenode.max.objects  
2022-09-07 10:01:15,370 [main] WARN org.apache.hadoop.conf.Configuration - had  
op.native.lib is deprecated. Instead, use io.native.lib.available  
2022-09-07 10:01:15,370 [main] WARN org.apache.hadoop.conf.Configuration - dfs.  
data.dir is deprecated. Instead, use dfs.datanode.data.dir  
2022-09-07 10:01:15,370 [main] WARN org.apache.hadoop.conf.Configuration - dfs.  
name.dir is deprecated. Instead, use dfs.namenode.name.dir  
2022-09-07 10:01:15,370 [main] WARN org.apache.hadoop.conf.Configuration - fs.d  
efault.name is deprecated. Instead, use fs.defaultFS  
2022-09-07 10:01:15,370 [main] WARN org.apache.hadoop.conf.Configuration - fs.c  
heckpoint.dir is deprecated. Instead, use dfs.namenode.checkpoint.dir  
2022-09-07 10:01:15,370 [main] WARN org.apache.hadoop.conf.Configuration - dfs.  
block.size is deprecated. Instead, use dfs.blocksize  
2022-09-07 10:01:15,370 [main] WARN org.apache.hadoop.conf.Configuration - dfs.  
access.time.precision is deprecated. Instead, use dfs.namenode.accesstime.precis  
ion  
2022-09-07 10:01:15,370 [main] WARN org.apache.hadoop.conf.Configuration - dfs.  
replication.min is deprecated. Instead, use dfs.namenode.replication.min  
2022-09-07 10:01:15,370 [main] WARN org.apache.hadoop.conf.Configuration - dfs.  
name.edits.dir is deprecated. Instead, use dfs.namenode.edits.dir
```

```
|grunt> count = my_macro();  
|grunt> dump count;  
2022-09-07 10:02:53,950 [main] INFO org.apache.pig.tools.pigst  
| Pig features used in the script: GROUP_BY  
2022-09-07 10:02:54,062 [main] INFO org.apache.pig.backend.had  
| ne.mapReduceLayer.MRCompiler - File concatenation threshold: 10  
| se  
2022-09-07 10:02:54,072 [main] INFO org.apache.pig.backend.had  
| ne.mapReduceLayer.CombinerOptimizer - Choosing to move algebr  
| iner  
2022-09-07 10:02:54,126 [main] INFO org.apache.pig.backend.had  
| ne.mapReduceLayer.MultiQueryOptimizer - MR plan size before opt
```

Output:

```
2022-09-07 10:03:57,600 [main] INFO org.apache.pig.backe
.MapReduceLauncher - Success!
2022-09-07 10:03:57,612 [main] INFO org.apache.hadoop.ma
l input paths to process : 1
2022-09-07 10:03:57,612 [main] INFO org.apache.pig.backe
l - Total input paths to process : 1
(Apr,301451)
(Aug,470275)
(Dec,186973)
(Feb,350885)
(Jan,273096)
(Jul,445653)
(Jun,386088)
(Mar,460257)
(May,339551)
(Nov,421626)
(Oct,411456)
(Sep,430529)
( [26,1)
(* [28,1)
(2|pq{jvk@-1.-@lvo))1.(.1--'1()@+)* [29,1)
grunt> █
```

Find Out the hit count per status messages (2XX, 3XX, 4XX, 5XX)

Log= LOAD '/user/training/vaibhav/access_log' using PigStorage(' ') as(ip:chararray, hp1:chararray, hp2:chararray, date:chararray, timezone:chararray, request1:chararray, request2:chararray, request3:chararray, code:int, bt:chararray);

B= GROUP Log by code;
 CC = FOREACH B generate group, COUNT(Log);
 Rcc = FILTER CC BY (\$0>400 AND \$0<500);
 Pdm = GROUP Rcc ALL;
 final = FOREACH Pdm GENERATE SUM(Rcc.\$1);

```

grunt> Log= LOAD '/user/training/vaibhav/access_log' using PigStorage(' ') as(ip:chararray, hp1:chararray, hp2:chararray, date:chararray, timezone:chararray, request1:chararray, request2:chararray, request3:chararray, code:int, bt:chararray);
2022-09-07 10:52:17,252 [main] WARN org.apache.pig.PigServer - Encountered Warning IMPLICIT_CAST_TO_INT 2 time(s).
grunt> describe Log;
2022-09-07 10:52:23,809 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1005: No plan for log to describe
Details at logfile: /home/training/pig_1662524156742.log
grunt> describe Log;
2022-09-07 10:52:30,529 [main] WARN org.apache.pig.PigServer - Encountered Warning IMPLICIT_CAST_TO_INT 2 time(s).
Log: {ip: chararray, hp1: chararray, hp2: chararray, date: chararray, timezone: chararray, request1: chararray, request2: chararray, request3: chararray, code: int, bt: chararray}
grunt> x= GROUP Log by code;
2022-09-07 10:53:42,120 [main] WARN org.apache.pig.PigServer - Encountered Warning IMPLICIT_CAST_TO_INT 2 time(s).
grunt> summ = FOREACH x generate group, COUNT( Log );
2022-09-07 10:53:56,194 [main] WARN org.apache.pig.PigServer - Encountered Warning IMPLICIT_CAST_TO_INT 2 time(s).
grunt> dump summ;
2022-09-07 10:54:04,474 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP BY
2022-09-07 10:54:04,474 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce.LaunchedMapReduceJobExecution - Pig features used in the script: GROUP BY
grunt> B= GROUP Log by code;
grunt> CC = FOREACH B generate group, COUNT( Log );
grunt> Rcc = FILTER CC BY ($0>400 AND $0<500);
grunt> Pdm = GROUP Rcc ALL;
grunt> █

-----
grunt> describe Pdm;
Pdm: {group: chararray, Rcc: {(group: int, long)}}
grunt> describe Rcc;
Rcc: {group: int, long}
grunt> █

-----
grunt> final = FOREACH Pdm GENERATE SUM(Rcc.$1);
grunt> dump final;
2022-09-07 12:05:38,741 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP BY
2022-09-07 12:05:38,893 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce.LaunchedMapReduceJobExecution - Pig features used in the script: GROUP BY
2022-09-07 12:05:38,902 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce.LaunchedMapReduceJobExecution - Pig features used in the script: GROUP BY
2022-09-07 12:05:38,908 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce.LaunchedMapReduceJobExecution - Pig features used in the script: GROUP BY
2022-09-07 12:05:38,928 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce.LaunchedMapReduceJobExecution - Pig features used in the script: GROUP BY
2022-09-07 12:05:38,928 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce.LaunchedMapReduceJobExecution - Pig features used in the script: GROUP BY
2022-09-07 12:05:39,004 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP BY
2022-09-07 12:05:39,010 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce.LaunchedMapReduceJobExecution - Pig features used in the script: GROUP BY
2022-09-07 12:05:39,014 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce.LaunchedMapReduceJobExecution - Pig features used in the script: GROUP BY

```

Output:

```
2022-09-07 12:06:55,046 [main] WARN  org.apache.pig.backend.hadoop
2022-09-07 12:06:55,046 [main] INFO  org.apache.pig.backend.hadoop
2022-09-07 12:06:55,050 [main] INFO  org.apache.hadoop.mapreduce
2022-09-07 12:06:55,050 [main] INFO  org.apache.pig.backend.hadoop
(84970)
grunt> █
```

Customer dataset:**Find out age group-wise customer count**

```
L = LOAD '/user/training/vaibhav/custs' using PigStorage(',') AS (id,fname,lname,age,profession);
X = FILTER L BY (age>30 AND age<=40);
G = GROUP X BY age;
C = FOREACH G generate group,COUNT(X);
Mapp = GROUP C ALL;
describe Mapp;
dump Mapp;
N = FOREACH Mapp GENERATE SUM(C.$1);
dump N;
```

```
grunt> L = LOAD '/user/training/vaibhav/custs' using PigStorage(',') AS (id,fname,lname,age,profession);
2022-09-07 11:33:54,114 [main] WARN org.apache.hadoop.conf.Configuration - io.bytes.per.checksum is deprecated. Inst
2022-09-07 11:33:54,114 [main] WARN org.apache.hadoop.conf.Configuration - dfs.permissions.supergroup is deprecated.
2022-09-07 11:33:54,114 [main] WARN org.apache.hadoop.conf.Configuration - dfs.max.objects is deprecated. Instead, u
2022-09-07 11:33:54,114 [main] WARN org.apache.hadoop.conf.Configuration - dfs.replication.interval is deprecated. I
2022-09-07 11:33:54,114 [main] WARN org.apache.hadoop.conf.Configuration - dfs.data.dir is deprecated. Instead, use
2022-09-07 11:33:54,114 [main] WARN org.apache.hadoop.conf.Configuration - dfs.access.time.precision is deprecated.
2022-09-07 11:33:54,114 [main] WARN org.apache.hadoop.conf.Configuration - dfs.replication.min is deprecated. Inst
```

```
grunt> X = FILTER L BY (age>30 AND age<=40);
2022-09-07 11:51:59,931 [main] WARN org.apache.
grunt> █
```

```
grunt> G = GROUP X BY age;
2022-09-07 11:52:25,426 [main] WARN
grunt> C = FOREACH G generate group,(
2022-09-07 11:52:30,733 [main] WARN
grunt> Mapp = GROUP C ALL;
```

```

2022-09-07 11:39:11,571 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Ma
2022-09-07 11:39:11,575 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total i
2022-09-07 11:39:11,575 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil -
(all, {(21,114), (22,100), (23,108), (24,110), (25,176), (26,210), (27,181), (28,190), (29,203), (30,211)})
grunt> █

```

```

grunt> N = FOREACH Mapp GENERATE SUM(C.$1);
2022-09-07 11:37:36,756 [main] WARN org.apache.pig.
grunt> █

```

```

grunt> dump N;
2022-09-07 11:37:58,423 [main] WARN org.apache.pig.PigServer - Enc
2022-09-07 11:37:58,424 [main] INFO org.apache.pig.tools.pigstats.
2022-09-07 11:37:58,551 [main] INFO org.apache.pig.backend.hadoop.
2022-09-07 11:37:58,562 [main] INFO org.apache.pig.backend.hadoop.
2022-09-07 11:37:58,564 [main] INFO org.apache.pig.backend.hadoop.
2022-09-07 11:37:58,580 [main] INFO org.apache.pig.backend.hadoop.
2022-09-07 11:37:58,580 [main] INFO org.apache.pig.backend.hadoop.
2022-09-07 11:37:58,681 [main] INFO org.apache.pig.tools.pigstats.
2022-09-07 11:37:58,685 [main] INFO org.apache.pig.backend.hadoop.
2022-09-07 11:37:58,686 [main] INFO org.apache.pig.backend.hadoop.
2022-09-07 11:38:00,946 [main] INFO org.apache.pig.backend.hadoop.
2022-09-07 11:38:00,957 [main] INFO org.apache.pig.backend.hadoop.
2022-09-07 11:38:00,977 [main] INFO org.apache.pig.backend.hadoop.
2022-09-07 11:38:00,977 [main] INFO org.apache.pig.backend.hadoop.
f reducers to 1

```

Output:

```

2022-09-07 11:53:44,856 [main] INFO |
2022-09-07 11:53:44,858 [main] INFO |
2022-09-07 11:53:44,858 [main] INFO |
(2061)
grunt> █

```