BigDataTraining.IN

Pig Log Analysis Session – Notes

```
Saravanans-MacBook-Pro:~ saravanans$ ssh hadoop@ec2-54-205-176-184.compute-1.amazonaws.com
The authenticity of host 'ec2-54-205-176-184.compute-1.amazonaws.com (54.205.176.184)' can't be established.
RSA key fingerprint is c1:01:ae:f5:53:1a:62:74:8f:be:e2:ac:ac:42:58:35.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'ec2-54-205-176-184.compute-1.amazonaws.com,54.205.176.184' (RSA) to the list of known hosts.
hadoop@ec2-54-205-176-184.compute-1.amazonaws.com's password:
Last login: Sun May 26 13:04:29 2013 from 182.65.186.58

       __|  __|_  )
       _|  (     /   Amazon Linux AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-ami/2012.03-release-notes/
There are 39 security update(s) out of 241 total update(s) available
Run "sudo yum update" to apply all updates.
Amazon Linux version 2013.09 is available.
-bash: warning: setlocale: LC_CTYPE: cannot change locale (UTF-8)
[hadoop@ip-10-235-53-189 ~]$ ls
[hadoop@ip-10-235-53-189 ~]$ cd /data/
```

1) website that is most used

```
users = LOAD '/users.txt' as (uid:chararray, name:chararray, age:int, ipaddr:chararray);

 dump users;
(100,govind,23,20.10.225.1)
(200,saketh,52,20.10.225.2)
(300,program,65,20.10.225.3)
```

(400,varun,24,20.10.225.4)
(500,rajiv,22,20.10.225.5)
(600,sharkar,22,20.10.225.6)
(700,kumar,22,20.10.225.7)
(800,ganesh,24,20.10.225.8)
(900,raja,23,20.10.225.9)
(101,sukumar,51,20.10.225.10)

loading logs
logs = LOAD '/logs.txt' as (date:chararray, ipaddr:chararray, uid:chararray, url:chararray, trancode:chararray, desc:chararray);
dump logs;

(10152013,20.10.225.1,100,google.com,tran101,describegoogle)
(10102013,20.10.225.2,200,yahoo.com,tran102,describeyahoo)
(10102013,20.10.225.2,200,gmail.com,tran103,describegmail)
(10102013,20.10.225.3,300,rediff.com,tran104,descriverediff)
(10102013,20.10.225.5,500,ubuntu.com,tran104,descriverediff)
(10102013,20.10.225.5,500,google.com,tran104,descriverediff)
(10102013,20.10.225.6,600,gmail.com,tran104,descriverediff)
(10102013,20.10.225.7,700,linux.com,tran104,descriverediff)

A = GROUP logs by (url);
(gmail.com,{(10102013,20.10.225.2,200,gmail.com,tran103,describegmail),(10102013,20.10.225.6,600,gmail.com,tran104,descriverediff)})
(linux.com,{(10102013,20.10.225.7,700,linux.com,tran104,descriverediff)})
(yahoo.com,{(10102013,20.10.225.2,200,yahoo.com,tran102,describeyahoo)})
(google.com,{(10152013,20.10.225.1,100,google.com,tran101,describegoogle),(10102013,20.10.225.5,500,google.com,tran104,descriverediff)})
(rediff.com,{(10102013,20.10.225.3,300,rediff.com,tran104,descriverediff)})
(ubuntu.com,{(10102013,20.10.225.5,500,ubuntu.com,tran104,descriverediff)})

B = FOREACH A GENERATE group as url, COUNT($1) as urlcount;
(gmail.com,2)
(linux.com,1)
(yahoo.com,1)
(google.com,2)
(rediff.com,1)
(ubuntu.com,1)

C = ORDER B by urlcount ASC;
(linux.com,1)
(yahoo.com,1)
(rediff.com,1)
(ubuntu.com,1)
(gmail.com,2)
(google.com,2)

1) For Users Most Active

A = GROUP logs BY uid;
(100,{(10152013,20.10.225.1,100,google.com,tran101,describegoogle)})
(200,{(10102013,20.10.225.2,200,yahoo.com,tran102,describeyahoo),(10102013,20.10.225.2,200,gmail.com,tran103,describegmail)})
(300,{(10102013,20.10.225.3,300,rediff.com,tran104,describerediff)})
(500,{(10102013,20.10.225.5,500,ubuntu.com,tran104,describerediff),(10102013,20.10.225.5,500,google.com,tran104,describerediff)})
(600,{(10102013,20.10.225.6,600,gmail.com,tran104,describerediff)})
(700,{(10102013,20.10.225.7,700,linux.com,tran104,describerediff)})
B = FOREACH A generate group as loguid, COUNT($1) as countid;
(100,1)
(200,2)
(300,1)
(500,2)
(600,1)
(700,1)
C = JOIN B by loguid,users by uid;
(100,1,100,govind,23,20.10.225.1)

(200,2,200,saketh,52,20.10.225.2)
(300,1,300,program,65,20.10.225.3)
(500,2,500,rajiv,22,20.10.225.5)
(600,1,600,sharkar,22,20.10.225.6)
(700,1,700,kumar,22,20.10.225.7)
D = ORDER C BY countid DESC;
(200,2,200,saketh,52,20.10.225.2)
(500,2,500,rajiv,22,20.10.225.5)
(100,1,100,govind,23,20.10.225.1)
(300,1,300,program,65,20.10.225.3)
(600,1,600,sharkar,22,20.10.225.6)
(700,1,700,kumar,22,20.10.225.7)

1) for which age group of users most active

E = FOREACH D GENERATE age, countid;
(52,2)
(22,2)
(23,1)
(65,1)
(22,1)
(22,1)