PingER Data Warehouse using Big Data with Cloudera on Nebula

This documentation is no longer valid. Please refer to this one.

Logging on

There are 4 VMs.

1. `dhcp-nebula-129-209.slac.stanford.edu`
2. `dhcp-nebula-129-210.slac.stanford.edu #application (e.g. Apache) runs on this one`
3. `dhcp-nebula-129-213.slac.stanford.edu`
4. `dhcp-nebula-129-214.slac.stanford.edu`

Note these machines are not visible (e.g. for the web) from outside SLAC. They are accessible via ssh from outside SLAC. Les can ssh with the usual priv /public ssh key.

Command mode

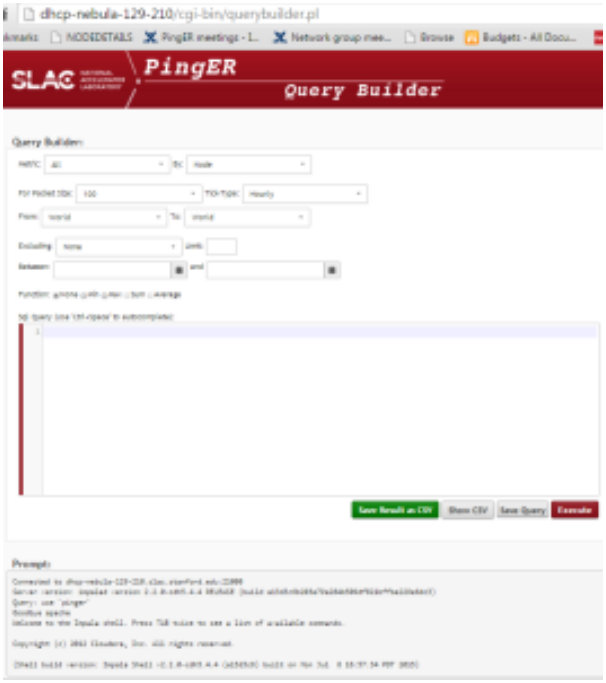
```
177cottrell@pinger:~$ssh dhcp-nebula-129-210.slac.stanford.edu
Warning: Permanently added 'dhcp-nebula-129-210.slac.stanford.edu' (RSA) to the list of known hosts.
Last login: Thu Aug  6 17:24:51 2015 from pinger.slac.stanford.edu
================================================================================
                              NOTICE TO USERS
This is a Federal computer system and is the property of the United States
...
conditions of use.  LOG OFF IMMEDIATELY if you do not agree to the conditions
stated in this warning.
================================================================================
RHEL Server 6.7 (Santiago) 2.6.32-573.3.1.el6.x86_64 (4x2099MHz OpenStack Nova)
================================================================================
WINSTART: Undefined variable.
169cottrell@dhcp-nebula-129-210:~$
```

Accessing Web Application

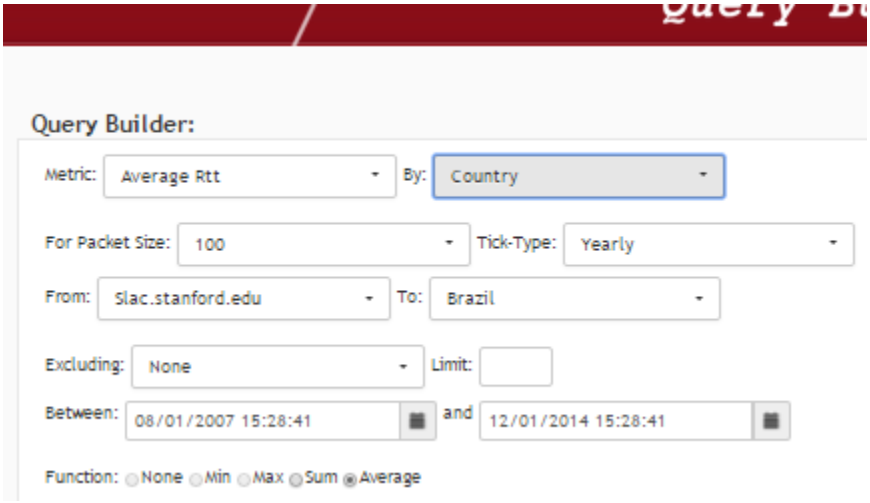
This is at

http://dhcp-nebula-129-210/cgi-bin/querybuilder.pl. We hope to move this to http://www-iepm.slac.stanford.edu/cgi-wrap/querybuilder.pl

Page looks like



I.e. very similar to pingtable.pl. Make your choices, for example



The above query responds with:

Prompt:

```
Connected to dhcp-nebula-129-210.slac.stanford.edu:21000
Server version: impalad version 2.2.0-cdh5.4.4 RELEASE (build a13d3c6b203e79a284b509df821bffbe229e6dc3)
Query: use `pinger`
Query: select src_continent.name, src_country.name, dst_continent.name, dst_country.name, t.year, avg(meas.average_rtt)
FROM pinger_measurement meas, time t, source_node src, destination_node dst, country src_country, country dst_country, continent src_continent, con
tinent dst_continent
WHERE meas.time_id = t.id
AND meas.source_country_id = src_country.id
AND meas.destination_country_id = dst_country.id
AND src_country.continent_code = src_continent.continent_code
AND dst_country.continent_code = dst_continent.continent_code
AND meas.source_node_id = src.id
AND meas.destination_node_id = dst.id
AND t.time_stamp BETWEEN '2007-01-08 15:28:41' AND '2014-12-01 15:28:41'
AND src.site_name = 'slac.stanford.edu'
AND dst_country.name = 'Brazil'
GROUP BY src_continent.name, src_country.name, dst_continent.name, dst_country.name, t.year
ORDER BY src_country.name, dst_country.name, t.year
Fetched 8 row(s) in 13.56s
```

Query Result Charts

Name	Name	Name	Name	Year	Avg(meas.average_rtt)
North America	United States	South America	Brazil	2007	225.210709083972
North America	United States	South America	Brazil	2008	231.5826930714022
North America	United States	South America	Brazil	2009	217.9829438172272
North America	United States	South America	Brazil	2010	225.8728057497396
North America	United States	South America	Brazil	2011	223.3030898666812
North America	United States	South America	Brazil	2012	219.8530903068268
North America	United States	South America	Brazil	2013	216.1611796843303
North America	United States	South America	Brazil	2014	218.8203957794926

There is also an attempt to separate the web server from the VMs. This version is at

/afs/slac.stanford.edu/g/www/cgi-wrap-bin/net/offsite_mon/newtraceping.pl. It places the images, Javascript, css in /afs/slac/g/www/www-iepm/pinger /querybuilder. However, we were unable to execute the impala-shell command remote on the cluster from www-wanmon. This included trying with ssh but ran into known-hot problems,

Accessing Hue application

http://dhcp-nebula-129-210:8888/

User name pinger, password in iepm escrow

This enables one to look at the tables created (lh column), to enter queries, and to see recent queries.

For example choose Query Editor and then Impala. Can choose metrics to chart.

May be useful to use web application to create the SQL then copy to Hue to get chart.

Accessing Cloudera Manager

http://dhcp-nebula-129-210:7180/cmf/home

User name pinger, password in iepm escrow

To look at activity of IO, network, CPU performance etc. Can also control the servers.

If you see a red in the circle for service on the left, try going to manager (at the top) , click on the down arrow for the service and Restart.

If have problems with querybuilder, it is running under Apache so will need to restart Apache. To do this ssh dhcp-nebula-129-210

sudo /sbin/service httpd [stop | start | restart]

Problems

The java installed with Cloudera at /usr/java/jdk1.7.0_67-cloudera/bin/java appears to have a vulnerability, see INC0055183. The version at /usr/local/bin/java installed as part of Taylor appears to be OK. I therefore made a copy of the the cloudera java, removed it and planned to make a logical link from /usr/java/jdk1.7.0_67-cloudera/bin/java to /usr/local/bin/java

/usr/local/bin/java

```
192cottrell@dhcp-nebula-129-210:~$cp /usr/java/jdk1.7.0_67-cloudera/bin/java bin/
192cottrell@dhcp-nebula-129-210:~$sudo rm /usr/java/jdk1.7.0_67-cloudera/bin/java
194cottrell@dhcp-nebula-129-210:~$sudo ln -s /usr/local/bin/java /usr/java/jdk1.7.0_67-cloudera/bin/java
Sorry, user cottrell is not allowed to execute '/bin/ln -s /usr/local/bin/java /usr/java/jdk1.7.0_67-cloudera
/bin/java' as root on dhcp-nebula-129-210.
```

Karl Amrhein gave Les Cottrell sudo privs

```
$ls /usr/java/jdk1.7.0_67-cloudera/bin/
ControlPanel@  javafxpackager*  jmap*        keytool*       servertool*
appletviewer*  javah*           jmc*         native2ascii*  tnameserv*
apt*           javap*           jmc.ini*     orbd*          unpack200*
extcheck*      javaws*          jps*         pack200*       wsgen*
idlj*          jcmd*            jrunscript*  policytool*    wsimport*
jar*           jconsole*        jsadebugd*   rmic*          xjc*
jarsigner*     jcontrol*        jstack*      rmid*
java-rmi.cgi*  jdb*             jstat*       rmiregistry*
javac*         jhat*            jstatd*      schemagen*
javadoc*       jinfo*           jvisualvm*   serialver*
#i.e. java was no longer there
190cottrell@dhcp-nebula-129-210:~$/usr/java/jdk1.7.0_67-cloudera/bin/
190cottrell@dhcp-nebula-129-210:~$sudo ln -s /usr/local/bin/java /usr/java/jdk1.7.0_67-cloudera/bin/java
[sudo] password for cottrell:
191cottrell@dhcp-nebula-129-210:~$ls /usr/java/jdk1.7.0_67-cloudera/bin/        ControlPanel@  javadoc*
jinfo*       jvisualvm*     serialver*
appletviewer*  javafxpackager*  jmap*        keytool*       servertool*
apt*           javah*           jmc*         native2ascii*  tnameserv*
extcheck*      javap*           jmc.ini*     orbd*          unpack200*
idlj*          javaws*          jps*         pack200*       wsgen*
jar*           jcmd*            jrunscript*  policytool*    wsimport*
jarsigner*     jconsole*        jsadebugd*   rmic*          xjc*
java@          jcontrol*        jstack*      rmid*
java-rmi.cgi*  jdb*             jstat*       rmiregistry*
javac*         jhat*            jstatd*      schemagen*
192cottrell@dhcp-nebula-129-210:~$ls -l  /usr/java/jdk1.7.0_67-cloudera/bin/java
lrwxrwxrwx 1 root bin 19 Oct  4 19:50 /usr/java/jdk1.7.0_67-cloudera/bin/java -> /usr/local/bin/java*
193cottrell@dhcp-nebula-129-210:~$
#Now the new version is in place
```

This was repeated on: dhcp-nebula-129-209, dhcp-nebula-129-213, dhcp-nebula-129-214

See ServiceNow INC0055183 for m ore details.

Database

The database only contains by-node and 100Bytes. The by-site is done by Impala. It is missing most of the data for 2002.

Creation

The data was transformed from PingER hourly flat files to the database format using a scicumulus workflow (a parallelization tool) to Extract the flat files from the FTP space, Transform them following a schema (see the paper) and Loaded into the HDFS file space. Renan has the scripts. We need a cronjob to update.

Backup

This is at

193cottrell@pinger:~tbarbosa$ls -l /afs/slac/public/users/cottrell/pinger-datawhs/
total 5373504
-rw-r--r-- 1 cottrell sf 896878 Aug 29 14:43 dimensions.rar
-rw-r--r-- 1 cottrell sf 5501570738 Aug 29 15:00 years.rar

It contains data for 1998-2014.