**Purpose**

1.Get a list which ip is blocked by aws and built a query website.

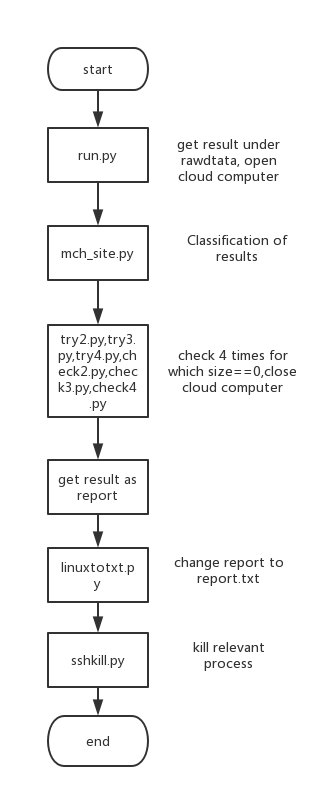
2.Operation regions:

* us-west-2
* us-office
* cn-office

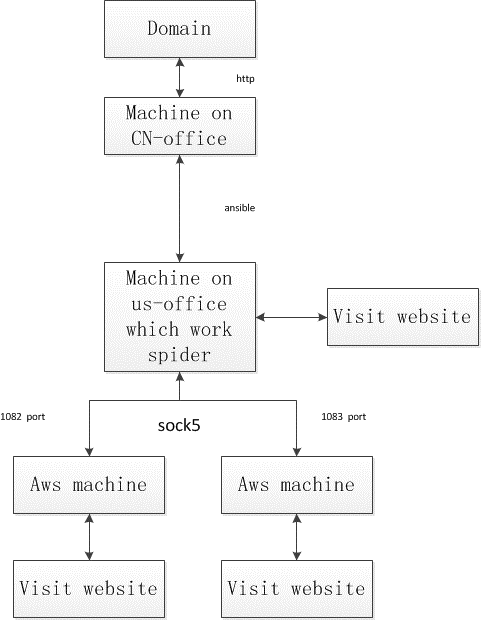
3.You can get all resource from https://wiki.jarvis.trendmicro.com/display/IWSS/List+of+websites+blocking+AWS+ip.

**The flow chart as follows:**

**Spider’s flow chart:**

****

**Website’s flow chart:**

****

**Start or create aws instances**

1. In the first step, you must install python2.7 and pip tools in the current Linux environment. Some Linux has python but no pip. You can install the pip after installing easyinstall first.
2. Use the command pip install awscli --upgrade --user to install the aws client.
3. Add export PATH=~/.local/bin:$PATH to ~/.bashrc . Use the command source ~/.bashrc .
4. After that, configure the AWS client. Use the command aws configure to configure the AWS client. If you have built machines in the aws cloud and set up the relevant security groups (open 22 port). You can configure the AWS client as follows:

AWS\_ACCESS\_KEY\_ID=AKIAJBOALUB26NG3T6MQ

AWS\_SECRET\_ACCESS\_KEY=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

AWS\_DEFAULT\_REGION=us-west-2

DEFAULT\_OUTPUT\_FORMAT=text

1. If you have built machines in the aws cloud, you can also use commands to create one, but you still should set up the relevant security groups and configure fixed IP by yourself.

Use the command aws ec2 start-instances --instance-ids i-0364c04273bae8d75 i-0c9f07975773803e4 to start the aws computer.

Use the command aws ec2 stop-instances --instance-ids i-0364c04273bae8d75 i-0c9f07975773803e4 to stop the aws computer.

Use the command aws ec2 run-instances --image-id ami-bf4193c7 --count 1 --instance-type t2.micro --key-name bin\_test --security-group-ids sg-cbe829b7 --subnet-id subnet-be0aa2d9 to create a new aws computer.( ami-bf4193c7 is a public ami)

**Configure the running environment**

1. Use the command mkdir rawdata 2try 3try 4try 5try webtest to build directory. These directories contain the results of using the system command to access the associated web pages. The directory rawdata contains all the results of the first visit. Directory 2try,3try,4try,5try to store the results of the visit to size=0,5try is the final result. Without these directories, the program will not be executed.
2. If you use Xshell, please use the command yum install lrzsz –y. You can use the command rz, sz to transfer files.
3. use the command rz to transfer files include run.py, black\_list1.txt, white\_list.txt, mch\_site.py, top-1m.csv, try2.py, try3.py, try4.py, check2.py, check3.py, check4.py, work.py, sshkill.py, init.py, linuxtotxt.py should be included under root.
4. Make sure your US computer's 1080,1081(for spider),1082,1083(for website) ports are available. Use the pip install Beautifulsoup to install the Beautifulsoup module, which is used in the program for parsing web text.
5. Use the command rz to transfer bin\_test2.pem, change bin\_test2.pem’s mode chmod 600 bin\_test2.pem.

**Run program**

1. Use the command python work.py to work it.

2. If you want to get a new result, you should delete all files in the rawdata directory, while 2try,3try,4try, and 5try files can be ignored because they are generated from rawdata.

3. Why do I have 2try,3try,4try,5try? The main goal is to avoid the fact that some sites are inaccessible at some point, or that the speed of the site's bandwidth is being preempted by faster access to the site.

4. You can find the result called report.txt under the directory of root.

**Build a query site on CN-Lab.**

If you want to set website on aws, it means the program on aws must call program on US-Lab, because we use two different machine access results to compare. But company does not allow external machine access, the program on aws cannot invoke programs on the company machine. So it will be difficult.

1. Use the command rz to transfer tornadow.py, result.html, form.html on CN-Lab’s machine.

2. Use the command pip install tornado to install tornado. Why did I choose tornado? Because the tornado framework has an asynchronous non-blocking mechanism, the query will take a long time. If a blocking framework, when a user request query website, because of the block, not the second user provide responses, even with the method of multiple threads or processes can't solve the fundamental problem.

3. Use the command pip install ansible to install ansible in order to execute the program on the remote host. Frist of all,we should establish SSH free key communication. Use the command ssh-keygen to produce key . Two times enter, then use the command scp id\_rsa.pub root@ 10.202.241.216:.ssh/id\_rsa.pub to transfer the key(10.202.241.216 is a machine’s ip of US-Lab).

Use the command cat  id\_rsa.pub >> authorized\_keys to change the key. Use the command chmod 600 authorized\_keys. Use the command chmod 700 .ssh. Use the command ssh root@ 10.202.241.216 to test if it is ok.Add content to /ets/ansible/hosts. Vim hosts and add as follows：

[test]

10.202.241.216

Then we use ansible test -m command -a 'ls' to test ansible whether work.

4. Use the supervisor as the process manager. Use the command pip install supervisor to install supervisor. Use echo\_supervisord\_conf > etc/supervisord.conf to generate standard configuration files. Add to etc/supervisord.conf as follows：

[program:app]

command=python /root/tornadow.py

directory=/root/

user=root

autorestart=true

redirect\_stderr=true

stdout\_logfile=/tmp/tornado.log

loglevel=info

5. Use the command supervisord to start supervisor

Use the command supervisorctl start app to start app

Use the command supervisorctl stop app to stop app

Use the command supervisorctl reload to reload app

You can use ps –ef to watch supervisor if work.

1. Use the command wget -c <https://nginx.org/download/nginx-1.10.1.tar.gz> to download nginx.

Use the command tar -zxvf nginx-1.10.1.tar.gz

to get nginx file then cd nginx-1.10.1. Use the command ./configure

to configure. Finally, use the command make && make install to install nginx.

1. Use the command cd /usr/local/nginx/conf to edit nginx config file.

Vim nginx.conf change this file as follows：

#user nobody;

worker\_processes 4;

#error\_log logs/error.log;

#error\_log logs/error.log notice;

#error\_log logs/error.log info;

#pid logs/nginx.pid;

events {

worker\_connections 1024;

}

http {

include mime.types;

default\_type application/octet-stream;

access\_log logs/access.log;

sendfile on;

#tcp\_nopush on;

#keepalive\_timeout 0;

keepalive\_timeout 200;

gzip on;

upstream tornados{

server 127.0.0.1:8000;

}

proxy\_next\_upstream error;

server {

listen 80;

server\_name localhost;

location /{

root html;

index form.html;

proxy\_pass\_header Server;

proxy\_set\_header Host $http\_host;

proxy\_redirect off;

proxy\_set\_header X-Real-IP $remote\_addr;

proxy\_pass http://tornados;

proxy\_connect\_timeout 300s;

proxy\_send\_timeout 300s;

proxy\_read\_timeout 300s;

}

}

}

8. Use the command cd /usr/local/nginx/sbin/ to enter the directory of nginx.

Use the command ./nginx to start nginx.

Use the command ./nginx -s to stop nginx.

1. Use command python init.py on machine of US-Lab.
2. Visit the website.

**Notice.**

1.You may need to manually turn off the cloud server if the reptile is aborted. for instance, Closing the console when the program runs will cause the reptile is aborted because the 'print' function is used to print the message.

2.The domain name of the website does not have regular restrictions, and cannot enter the url of http://www.baidu.com, otherwise it will cause an exception. The reason is that the domain name is divided and only the main domain name is saved. The baidu.com or www.tornadoweb.org/en/stable is considered to be normal. It means you can’t input http:// or https:// before domain.

3.If you need to terminate the program, kill the corresponding child process, and the main process will close the cloud computer and exit automatically.

4.If you change the cloud machine, modify the related configuration in work.py, init.py.

thread\_1 = threading.Thread(target=keep,args=("ssh -D 127.0.0.1:1080 ec2-user@52.10.36.162 -i ./bin\_test2.pem -NCTn ",))

thread\_1.daemon = True

thread\_1.start()

thread\_2 = threading.Thread(target=keep,args=("ssh -D 127.0.0.1:1081 ec2-user@35.161.28.170 -i ./bin\_test2.pem -NCTn ",))

Find this code and change red part as your new aws ip.