链滴

配置 NGINX 拒绝恶意访问 / 爬取网站

作者：washmore
原文链接：https://ld246.com/article/1502257483568
来源网站：链滴

最近有点忙,一段时间没管博客了,今天上来看了一下access.log,多了一些牛鬼蛇神,之前因为博客访问少,没怎么弄,看来是时候带一波节奏了

之前的做法

以前就已经陆续发现一些恶意用户访问了,比如:

- 认为后端是java对tomcat的/manager进行访问的,
- 认为后端是php做一些eval或者爆破操作的,
- 一些独狼/个人蜘蛛用户不分时段对网站进行大规模爬取的

由于都是一些零散的访问,针对这些行为 在nginx.conf同目录下创建了一个denyIpList.conf配置文件,容形式如下:

```
# 针对单个ip的形式
deny 171.94.171.205;
deny 115.29.166.101;
deny 182.247.251.48;
deny 61.147.89.17;
# 针对网段的形式
deny 66.249.227.0/24;
```

然后在nginx.conf合适的位置引入此配置文件:

```
http {
    include       mime.types;
    include       denyIpList.conf;

    default_type  application/octet-stream;
...以下省略
```

重启nginx后生效,这样,当这些ip/网段发起访问后,直接返回403;

现在的做法

现象

今早上来看了一下访问记录后,发现了几组丧心病狂的内容:

1. user-agent为 Baidu-YunGuanCe-SLABot(ce.baidu.com) 的访问;
2. 来自美国66.249.*.*网段的访问;
3. user-agent为 Mozilla/5.0 (compatible; MJ12bot/v1.4.7; ... 的访问
4. user-agent为空(正常浏览器访问不会为空的)

分析

其中第一个我刚开始以为是我配的百度云观测网站定期健康检查的访问记录,但是简单统计了一下,数也太大了,而且不分时段都有,初步怀疑是有人闲着无聊借用百度云观测提供的工具对本站进行了友e情y压测...即便不是,我也不需要云观测提供的特殊服务,准备直接ban掉;

第二个,应该是谷歌的爬虫(ua判断),以前也看到过访问记录,频率比较低,直接deny访问地址的,但是最访问的ip也太多了,根本ban不过来;

第三个,MJ12bot比较常用的爬虫工具,访问ip也是来自世界各地;

解决方案

根据以上分析,发现大部分恶意请求可以通过user-agent来判断,因此,考虑通过nginx提供的一些内置量进行配置:

- 首先我们还是新建一个文件denyUaList.conf在nginx.conf同目录下;
- 编写denyUaList.conf规则内容:

```
#禁止常用工具的抓取
if ($http_user_agent ~* (Scrapy|Curl|HttpClient|Java)) {
  return 403;
}
#禁止指定UA及UA为空的访问
if ($http_user_agent ~* "Baidu-YunGuanCe|FeedDemon|JikeSpider|Indy Library|Alexa Toolbar|
skTbFXTV|AhrefsBot|CrawlDaddy|CoolpadWebkit|Feedly|UniversalFeedParser|ApacheBench|M
crosoft URL Control|Swiftbot|ZmEu|oBot|jaunty|Python-urllib|lightDeckReports Bot|YYSpider|D
gExt|YisouSpider|MJ12bot|heritrix|EasouSpider|LinkpadBot|Ezooms|^$" )
{
  return 403;
}
```

tips:注意书写格式if和(之间有空格,|Ezooms|最后面有个|破折号,用于禁止空ua访问

- 在nginx.conf中合适的位置引入denyUaList.conf

```
...以上省略
    server {
        listen          443 ssl;
        server_name     example.com;#你的域名

        ...ssl配置省略...

        include         denyUaList.conf;
...以下省略
```

tips:注意因为denyUaList.conf包含if等控制语句,因此不能和denyIpList.conf一样放在根节点,需要行根据需要放在server节点中

以上就是目前的配置方案,如果后续有优化升级,会在本帖更新,如果有朋友有更合适的方案,请在留言中复!

测试

在Baidu-YunGuanCe前增加chrome|,然后使用谷歌浏览器访问博客地址,返回403forbidden,切换为i
访问,正常进入(不过d大@88250竟然给了一个超low的提示(/ □)),测
通过,去掉chrome|,重启nginx,收工!