文章编号：2096-1472(2020)-01-35-05
DOI:10.19644/j.cnki.issn2096-1472.2020.01.009

基于ELK的BIND海量日志分析与用户行为态势感知研究

阮晓龙[1]，路景鑫[2]

(1.河南中医药大学信息技术学院，河南 郑州 450046；
2.郑州泰来信息科技有限公司，河南 郑州 450000)

摘　要：在互联网大规模应用的环境下，如何有效分析用户喜欢访问什么网站、在互联网上喜欢做什么、浏览什么是针对用户行为分析的重要内容。DNS作为一种域名解析服务，是互联网业务通信的重要保障，几乎所有互联网业务访问运行均离不开DNS的支持，所以本文通过对DNS海量日志进行收集、清洗、存储全流程处理过程，并结合以ELK为平台、以业务分析模型为基础，从而使DNS日志分析结果可视化清晰呈现，让用户了解互联网业务访问运行趋势，也直观表现出用户实际的访问情况，最终实现用户行为的态势感知。

关键词：ELK；BIND；日志分析；态势感知
中图分类号：TP31　**文献标识码**：A

ELK-based BIND Massive Log Analysis and User Behavior Situation Awareness Research

RUAN Xiaolong[1],LU Jingxin[2]

(1.*School of Information Technology,Henan University of Traditional Chinese Medicine,Zhengzhou* 450046,*China*;
2.*Zhengzhou Tailai Information Technology Co.,Ltd.,Zhengzhou* 450000,*China*)

Abstract: In the environment of large-scale application of Internet,how to effectively analyze what websites users like to visit,what they like to do on the Internet,and what websites users like to brows is an important content of user behavior analysis.DNS,as a domain name resolution service,is an important guarantee for Internet business communication.Almost all Internet business access and operation can not be separated from DNS support.Therefore,through the whole process of collection,cleaning and storage of DNS massive logs,combined with elk platform and business analysis model,this paper makes DNS log analysis results visible and clear.Let users understand the operational trend of Internet business access,and also intuitively show the actual access situation of users,and finally achieve the situation awareness of user behavior.

Keywords:ELK;BIND;log analysis;situational awareness

1 引言(Introduction)

伴随着互联网规模的日益扩大，用户量增加[1]。第43次《中国互联网络发展状况统计报告》中指出，截至2018年12月，网民规模达8.29亿，全年新增网民5653万，互联网普及率为59.6%。针对互联网域名解析系统，报告中也指出，截至2018年12月，我国域名总数为3792.8万个。其中，“.CN”域名总数为2124.3万个，在域名总数中占比56.0%[2]。

域名服务DNS实现域名与IP地址之间的转换，几乎所有互联网应用的正常运行都离不开DNS的支持。DNS是全球互联网服务的基石，是互联网通信的重要保障。域名生命周期的不同阶段中会产生域名操作管理日志、查询服务日志、解析服务日志等数据。这些日志数据记录蕴含用户访问行为特征、域名系统服务质量、互联网访问安全、域名服务发展趋势等大量信息，是对域名管理和域名服务状态的体现，反映域名服务体系的总体特征[3]。

本文基于ELK对BIND海量日志进行数据分析，实现域名服务质量及用户访问行为的态势感知分析与研究。

2 BIND日志(BIND log)

2.1 BIND日志配置

在默认情况下，BIND把日志消息写到/var/log/

messages文件中，而这些日志消息是非常少的，主要包含启动、关闭，以及严重错误的日志消息。如果需要详细记录服务器的运行日志，用户可自行配置日志行为记录内容。

本文基于BIND 9进行DNS服务实现，其日志配置操作较为简单灵活，要详细记录DNS域名解析与服务器的运行状况，要在配置文件中使用“logging”语句来定义所需要的日志记录类型。

(1)常用术语

BIND日志配置中常用术语及其说明，如表1所示。

表1 BIND日志常用术语

Tab.1 Common terms for BIND logs

术语	含义
channel	通道，日志输出的方式，如syslog(系统日志记录)、file(文本文件)、stderr(标准错误输出)或/dev/null(空)
category	类别，日志的消息类别，内置类别分别为default(默认)、general(未明确分类)、client(客户端请求)、config(配置文件分析)、database(内部数据库消息)、dnssec(DNSSEC签名响应)、lame-servers(发现错误授权)、network(网络操作)、notify(异步区变动通知)、queries(查询日志)、resolver(递归查询)、security(认可/非认可的请求)、update(动态更新事件)、xfer-in(服务器的区域接收)、xfer-out(服务器的区域传送)
module	模块，产生消息的来源模块名称
facility	设备，表明Syslog设备名称
severity	严重性，消息的严重性等级，按照严重性依次递减，分别为critical、error、warning、notice、info、debug、dynamic

(2)日志语法格式

BIND的logging日志语法格式如下所示。

```
logging {
channel channel_name {                    //定义通道
file log_file [versions number | unlimited] [size sizespec]; | syslog optional_facility; | null; | stderr; //定义输出方式
severity log_severity;                    //定义消息严重性
[print-time boolean;]                     //是否添加时间前缀
[print-severity boolean;]                 //是否添加消息严重性前缀
[print-category boolean;]                 //是否添加消息类别名前缀
};
category category_name {                  // 定义类别
channel_name;
......
};
};
```

(3)DNS日志配置

本文中主要实现DNS查询日志配置并满足以下要求，日志文件代码如下所示。

①能够获取并存储查询日志，以文本文件方式存储；

②文本文件大小进行限制，不能无限存储，并设置多版本文件；

③存储日志消息记录可添加时间、消息严重性、类别等内容前缀。

```
logging {
channel query{
                                          //查询日志
file "log/query.log" versions 9 size 32m;
//日志存储
severity info;
print-category yes;
print-severity yes;
print-time yes;
};
category default {null;};
category queries {query;};
category network {null;};
category client {null;};
category general {null;};
};
```

2.2 BIND日志解读

查看DNS服务器中“/log/query.log”查询日志记录文件，获取DNS一条查询日志如下所示，根据日志消息记录类型进行拆分每个字段的含义如表2所示。

08-Aug-2019 04:05:10.009 queries: info: client 10.10.3.234#64345 (vip.wps.cn): view openlabs: query: vip.wps.cn IN A + (10.10.3.70)

表2 查询日志消息记录拆分

Tab.2 Query log message record split

消息字段	说明
08-Aug-2019 04:05:10.009	DNS查询日志获取时记录到的时间，格式为日-月-年时:分:秒
queries	日志的消息类别为查询日志日志
info	日志的消息等级为“info(消息)”
client 10.10.3.234	发起请求查询的客户端地址
64345	发起请求查询的客户端端口
view oepnlabs	发起请求查询的客户端来自哪个视图。一个DNS服务器可定义多个view(视图)，每个view中可定义一个或多个zone(区域)。每个view用来匹配一组客户端，多个view内可能需要对同一个区域进行解析，但使用不同的区域解析库文件
query: vip.wps.cn	查询的域名记录
IN A	查询的域名记录类型
+	标识是否设置了递归查询，如果是为+，否则为-
(10.10.3.70)	进行域名查询的DNS服务器的地址

2.3 BIND日志推送

Filebeat是一种轻量型日志采集器，用于转发和汇总日志与文件。本文基于该采集器进行BIND日志数据获取并推送到Logstash服务器中进行日志数据处理，其操作过程如下所示。

(1)获取与安装

获取Filebeat软件可通过官网网址(https://www.elastic.co/cn/downloads/beats/filebeat) 选择相应版本进行下载。下载完成后的软件包，放置到DNS服务器中，并通过以下命令解压安装。

```
# tar-zxvf filebeat-7.3.0-linux-x86_64.tar.gz
```

(2)修改配置文件

进入Filebeat软件解压目录下，对Filebeat配置文件“filebeat.yml”进行编辑修改，设置本地日志文件的路径与输出Logstash服务器地址，具体配置如下所示。

```
filebeat.inputs:
-type: log    #数据类型
  enabled: true    #开启模块
  paths:
    -/var/named/log/query*    #文件路径
  fields:
     document_type: dnslog     #定义日志文件索引值，区分日志文件
output.logstash:
  hosts: ["10.10.2.231:5044"]    #输出Logstash地址
```

(3)执行启动Filebeat

配置执行后台启动Filebeat，其操作命令如下所示。

```
//将所有标准输出及标准错误输出到/dev/null空设备，即没有任何输出
# nohup./filebeat-e-c filebeat.yml>/dev/null 2>&1 &
//或者执行如下命令操作
# nohup ./filebeat-e-c filebeat.yml > filebeat.log &
```

3 日志数据处理(Log data processing)

Logstash作为Elasicsearch常用的实时数据采集引擎，可以采集来自不同数据源的数据，并对数据进行处理后输出到多种输出源。Logstash的数据处理主要是由Inputs(用于从数据源获取数据)、Filters(用户处理数据如格式化、数据派生等)、Outputs(用于数据输出)三部分组成。

在Logstash的“config”目录内创建配置文件“dns-logstash.conf”进行日志数据处理，其操作过程如下所示。

3.1 日志数据接收

在“dns-logstash.conf”配置文件中填写Input内容，用于接收来自DNS服务器的Filebeat数据，配置内容如下所示。

```
input {
  beats {
    port=>5044
  }
}
```

3.2 日志数据格式化

应根据日志字段的含义进行切分格式化处理，可使用Grok或自定义规则，将定义格式化规则内容写入配置文件“filter”字段中，配置内容如下所示。

```
filter {
    if [fields][document_type]== "dnslog" {
            //根据正则规则进行日志字段拆分
            grok {
    match=>{ "message"=>"^(?<dns_time>.*)\s(?<dns_desc>\w*):\s(?<dns_level>\w*):\sclient\s(?<client_ip>\d{1,3}\.\d{1,3}\.\d{1,3}\.\d{1,3})#(?<client_port>\w*)\s\((?<dns_domain>.*)\):\squery:\s(?<dns_domain1>.*?\s)(?<dns_class>.*?\s)(?<dns_type>.*?\s)(?<dns_info>.*?\s)\((?<dns_ip>.*)\)"}
            }
    //DNS日志时间格式化
    date {
            match=>["dns_time" , "dd-MMM-yyyy HH:mm:ss.SSS"]
            target=>"dns_time_local"
            }
    //根据客户端IP获取地理位置
    geoip {
            source=>"client_ip"
            target=>"geoip"
            database=>" /opt/logstash/vendor/bundle/jruby/2.5.0/gems/logstash-filter-geoip-6.0.1-java/vendor/GeoLite2-City.mmdb"
            add_field=>["[geoip][coordinates]", "%{[geoip][longitude]}"]
            add_field=>["[geoip][coordinates]", "%{[geoip][latitude]}"]
```

```
        }
    //对字段信息进行处理，格式化类型、删除字段等
    mutate {
        convert=>["[geoip][coordinates]", "float"]
        remove_field=>["dns_domain1"]
        strip=>["dns_time","dns_desc","dns_level","client_ip","client_port","dns_domain","dns_class","dns_type","dns_info","dns_ip"]
        }
    }
}
```

3.3 日志数据清洗

可对拆分的字段进行深入清洗处理，需在“filter”字段中引用脚本清洗文件“dns.rb”，配置内容如下所示。

```
ruby {
    path=>"/opt/logstash/script/dns.rb"
    remove_field=>["dns_time"]
}
```

脚本清洗处理文件中处理字段部分内容如下所示。

```
//进行多级域名聚合合并
dns_domain=Array.new
dns_domain=event.get("dns_domain").split('.')
event.set('dns_domain_ana_arr',dns_domain)
dns_domain_ana=""
for i in 1..dns_domain.length-1
    dns_domain_ana +=dns_domain[i] +"."
end
event.set('dns_domain_ana',dns_domain_ana[0..dns_domain_ana.length-2])
```

3.4 日志数据输出

如果是在对Logstash配置文件调试阶段，可将处理后的数据输出到控制台中，从而可直观查看处理后数据是否正确，其配置内容如下所示。

```
output {
    if [fields][document_type]== "dnslog"  {
stdout { codec=>rubydebug }
}
}
```

如果需要处理后的数据推送至Elasticsearch服务器中，其配置内容如下所示。

```
output {
    if [fields][document_type]== "dnslog"  {
    elasticsearch{
        hosts=>"10.10.2.231:9200"
        index=>"dnslog"
    }
    }
}
```

将DNS日志文件推送到Elasticsearch服务中，基于Kibana服务进行数据展示，如图1所示。

图1 DNS推送日志文件展示

Fig.1 DNS push log file display

4 分析模型研究(Research on analysis model)

4.1 分析模型

基于采集的DNS日志数据结构并结合DNS作用可创建如表3所示日志数据分析模型。

表3 DNS分析模型

Tab.3 DNS analysis model

序号	分析模型	图标类型
1	日志总量统计分析	指标
2	记录类型占比分析	饼状图
3	记录类型访问次数统计分析	表格
4	独立IP数统计分析	指标
5	域名访问统计排行分析	表格
6	外部域名请求访问统计排行分析	表格
7	区域来访IP统计分析	表格
8	区域访问总量统计分析	表格
9	请求访问来源地区分析	地图
10	请求访问来源城市分析	标签云
11	一分钟日志量统计分析	指标
12	一天中每个小时的访问量统计分析	柱状图
13	一月中每天的访问量统计分析	柱状图
14	记录访问量实时统计分析	折线图
15	记录类型访问量实时统计分析	折线图
16	DNS请求量对比分析	折线图

4.2 数据分析可视化

使用Kibana可快速将获取到日志数据创建直观的分析展

示，并将各种分析图表汇总，形成仪表盘进行展示。以分析“DNS记录访问量统计分析”为例，绘制数据可视化图表，其操作步骤如下所示。

①在“可视化”操作界面中，点击【创建新的可视化】，选择可视化类型为“折线图”；

②选择可视化呈现数据源为“dnslog”；

③调整“存储桶”下“X轴”数据内容，以“Date Historgram”为聚合方式、“dns_time_local”为字段、设置“秒”为最小时间间隔；

④点击左侧“运行”图标可查看数据呈现效果，如图2所示，点击【保存】按钮完成可视化图标配置。

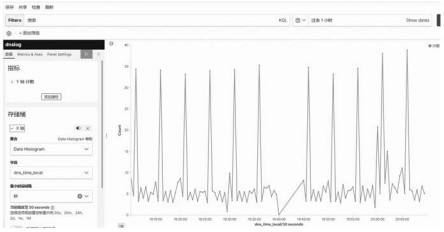

图2 可视化图标呈现

Fig.2 Visual icon rendering

5 态势感知研究(Research on situational awareness)

态势感知过程分为态势采集、态势理解、未来态势预测三个阶段[4]，相对于DNS日志分析来说，分别体现出域名系统“目前是怎样的、为什么这样、将要发生什么”等问题。通过解决这类问题可有效辅助域名系统提高服务质量、降低安全风险、了解用户访问行为等内容，基于BIND海量日志分析的态势感知框架如图3所示。

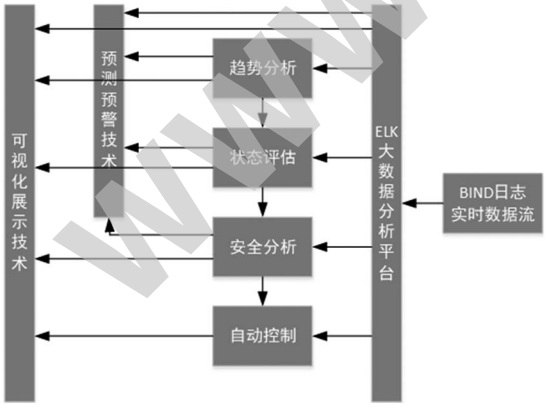

图3 基于ELK的BIND日志分析态势感知框架

Fig.3 ELK-based situation awareness framework for BIND log analysis

5.1 态势采集

态势采集获取环境中的重要元素，是态势感知的基础。高效地采集态势感知大数据并可靠地进行传输、存储、分析、处理，是构建ELK大数据分析平台所不可或缺的，通过该平台可快速实时获取DNS的运行状态、请求查询等日志。当新日志产生时，ELK平台也可快速检测并进行数据的拆分、清洗、格式化等处理并进行存储，为用户行为态势感知提供平台基础。

5.2 态势理解

根据域名解析的实际需求，可以建立一套实用、有效地态势理解指标体系，从而提高判断的准确性与可靠性。态势理解是通过大数据技术获取数据隐含的消息内容，明确DNS域名解析系统所处状态、掌握用户访问行为趋势，并按照指标数据进行态势评估分析。

5.3 态势预测

基于对态势的理解，对DNS域名解析未来的发展趋势进行预测，根据用户请求访问量，能够及时调整链路带宽或发现异常并及时采取措施等内容，此为态势感知的最高层次。

在进行态势预测时，可以将以往的态势数据分为输入、输出两个序列，通过对模型参数的不断调整，使输出序列的值逐步逼近于输入序列，从而得到比较可靠的初步预测模型。在初步预测模型的基础上，采用机器学习方法，可以进一步得到完善的预测模型[5]。

6 结论(Conclusion)

基于ELK大数据分析平台实现对BIND海量日志的分析，并通过对态势感知的了解，探索了海量BIND日志的分析方法，能够及时分析掌握DNS服务的运行性能，了解用户访问互联网的行为趋势，有效地分析用户上网行为。

参考文献(References)

[1] 程琦.基于DNS日志分析调度及优化设计与实现[J].福建电脑,2017,33(11):101-103.

[2] 中国互联网络信息中心 (CNNIC) 在京发布第43次《中国互联网络发展状况统计报告》[R].2019.

[3] 陈文文,吴开超.海量域名日志数据分析与可视化研究及应用[J].计算机应用研究,2016,33(02):335-338.

[4] 周凯,马智远,许中.基于大数据技术的智能电网态势感知分析[J].电器与能效管理技术,2018(21):70-76.

[5] 董超,刘雷.大数据网络安全态势感知中数据融合技术研究[J].网络安全技术与应用,2019(07):60-62.

作者简介：

阮晓龙(1981-)，男，本科，副教授.研究领域：计算机网络，计算机软件，Web技术.

路景鑫(1994-)，男，本科，工程师.研究领域：计算机软件，系统运维.