# ELK
# 镜像站日志可视化分析

主讲人：潘富

南阳理工学院镜像站

2024-08-17 南京大学 开源软件论坛


# 目录

- 背景介绍
- 技术选型
- 日志收集与分析
- 日志可视化展示
- 后续工作

# 背景介绍

- 日志在镜像站中的重要性
- 在成千上万条数据中检索过于繁琐
- 被恶意刷流后不方便查看涉事IP的行为
- 镜像站总体的数据分析

# 技术选型

- ElasticSearch（基于Lucene的搜索服务器）
- Kibana（ELK栈中的可视化工具）
- Filebeat（日志数据收集器）
- Grafana（数据展示）

# 日志收集与分析

- 安装并配置filtbeat
- 启用 nginx 模块 `filebeat modules enable nginx`
- 日志整体流向为：`Filebeat -> ElasticSearch -> Grafana`
- 使用 `Kibana` 对来自 `Filebeat` 的数据流进行处理

# 日志收集与分析

- NYIST 所采用的日志格式
- ‘$remote_addr - $remote_user [$time_local] “$request” $status $body_bytes_sent “$sent_http_content_type” “$http_referer” “$http_user_agent” - $scheme “$host” - “$request_time” “$upstream_response_time” “$http_x_forwarded_for”’

- 114.5.1.4 - - [17/Aug/2024:14:30:57 +0800] "GET /CTAN/fonts/greek/cbfonts/fonts/tfm/cbgreek/glto0800.tfm HTTP/1.1" 200 2032 "application/octet-stream" "-" "Mozilla/5.0 (Macintosh; Intel Mac OS X 11_3_0) AppleWebKit/537.36 (KHTML, like Gecko) Brave Chrome/89.0.4389.114 Safari/537.36" - https "mirror.nyist.edu.cn" - "0.053" "0.053" "-"

# 日志收集与分析

- Kibana对日志进行GORK
- (%{NGINX_HOST} )?"?(?:%{NGINX_ADDRESS_LIST:nginx.access.remote_ip_list}|%{NOTSPACE:source.address}) - (-|%{DATA:user.name}) \[%{HTTPDATE:nginx.access.time}\] "%{DATA:nginx.access.info}" %{NUMBER:http.response.status_code:long} %{NUMBER:http.response.body.bytes:long} "(-|%{DATA:http.content.type})" "(-|%{DATA:http.request.referrer})" "(-|%{DATA:user_agent.original})" - (-|%{DATA:http.scheme}) "(-|%{DATA:http.url.domain})" - "(-|%{NUMBER:http.request.time:float})" "(-|%{NUMBER:http.upstream_response.time:float})" "(-|%{NOTSPACE:http.xff})

- {
- "NGINX_HOST": "(?:%{IP:destination.ip})(:%{NUMBER:destination.port})?",
- "NGINX_NOTSEPARATOR": """[^ ,:]+""",
- "NGINX_ADDRESS_LIST": """(?:%{IP}|%{WORD})("?,?\s*(?:%{IP}|%{WORD}))*"""
- }

# 日志收集与分析

- (%{NGINX_HOST} )?"?
- 可选项，匹配 NGINX_HOST 或空格
- (?: ... )
- 是非捕获分组，用于匹配 IP 地址或者 NGINX_NOTSEPARATOR 定义的域名
- %{IP:destination.ip}
- 匹配 IP 地址，并将其存储到 destination.ip 字段
- (:%{NUMBER:destination.port})?
- 匹配可选的端口号，并将其存储到 destination.port 字段

# 日志收集与分析

- (%{NGINX_HOST} )?"?
- NGINX_NOTSEPARATOR
- 用于定义 NGINX 主机中不能是分隔符的部分
- NGINX_ADDRESS_LIST
- 用于匹配 NGINX 主机地址列表
- (?:%{IP}|%{WORD}) 匹配 IP 地址或者单词（WORD）
- ("?,?\s*(?:%{IP}|%{WORD}))* 是一个重复模式，用于匹配多个地址，包括逗号、空格等分隔符

# 日志收集与分析

- \[%{HTTPDATE:nginx.access.time}\]
- 匹配日期时间，格式为 HTTPDATE，提取到 nginx.access.time 字段
- "%{DATA:nginx.access.info}“
- 匹配任意数据作为 nginx.access.info 字段的值
- %{NUMBER:http.response.status_code:long}
- 匹配 HTTP 响应状态码，存储为 long 类型的 http.response.status_code

# 日志收集与分析

- 对nginx.access.info再次进行GORK
- "GET /epel/9/Everything/x86_64/Packages/g/ghc-js-dgtable-devel-0.5.2-23.el9.x86_64.rpm HTTP/2.0"
- %{WORD:http.request.method} %{DATA:url.path} HTTP/%{NUMBER:http.version}
- url.original里再次细分
- /%{DATA:http.request.repo}/
- 其余GORK同理

# 日志收集与分析

- (%{NGINX_HOST} )?"?(?:%{NGINX_ADDRESS_LIST:nginx.access.remote_ip_list}|%{NOTSPACE:source.address}) - (-|%{DATA:user.name}) \[%{HTTPDATE:nginx.access.time}\] "%{DATA:nginx.access.info}" %{NUMBER:http.response.status_code:long} %{NUMBER:http.response.body.bytes:long} "(-|%{DATA:http.content.type})" "(-|%{DATA:http.request.referrer})" "(-|%{DATA:user_agent.original})" - (-|%{DATA:http.scheme}) "(-|%{DATA:http.url.domain})" - "(-|%{NUMBER:http.request.time:float})" "(-|%{NUMBER:http.upstream_response.time:float})" "(-|%{NOTSPACE:http.xff})

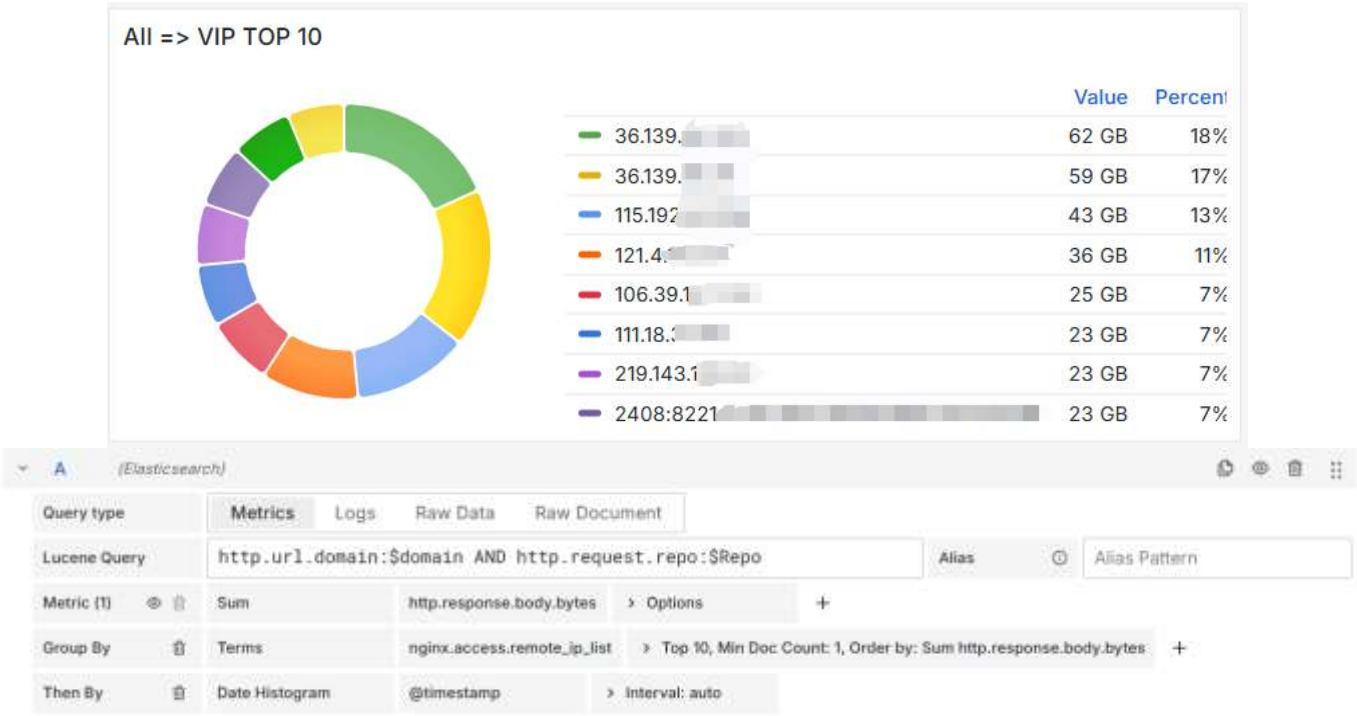
- '$remote_addr - $remote_user [$time_local] "$request" $status $body_bytes_sent "$sent_http_content_type" "$http_referer" "$http_user_agent" - $scheme "$host" - "$request_time" "$upstream_response_time" "$http_x_forwarded_for"'

# 日志可视化展示

- 在Grafana中新增数据源，索引名为filebeat-*
- 新建dashboard（mirrorslog）进行总体数据展示
- 新建变量
- Domain : {"find":"terms","field":"http.url.domain"}
- 创建Ad hoc filters类型变量

# 日志可视化展示

- VIP是需要重点关注的对象！！



# 日志可视化展示

访问 PV（相对）

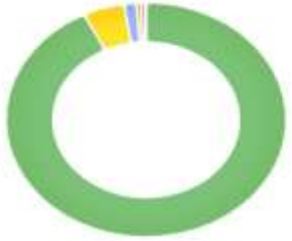

访问 UV（相对）


All => 平均响应时间(s)


基本概况（相对）

| 域名 | pv | uv | 处理时间(s) | 响应时间(s) | 出站流量之和 |
|---|---|---|---|---|---|
| mirror.nyist.edu.cn | 4,104,594 | 209,905 | 4.01 | 2.456 | 7.09 TB |
| mirror4.nyist.edu.cn | 4,116 | 1,363 | 7.223 | 0.738 | 12.1 GB |
| mirrors.nyist.edu.cn | 10,109 | 29 | 3.219 | 0.465 | 1.84 GB |
| mirror6.nyist.edu.cn | 408 | 56 | 0.248 | 0.304 | 831 MB |

All => 国家/地区访问占比



| | Value | Percent |
|---|---|---|
| Asia | 3,816,604 | 93% |
| North America | 191,626 | 5% |
| Europe | 56,173 | 1% |
| South America | 16,526 | 0% |
| Africa | 14,723 | 0% |
| Oceania | 4,400 | 0% |

All => 访问前十的省份

| | Value | Percent |
|---|---|---|
| Guangdong | 539 GB | 23% |
| Beijing | 388 GB | 16% |
| Shanghai | 255 GB | 11% |
| Henan | 255 GB | 11% |
| Jiangsu | 241 GB | 10% |
| Zhejiang | 200 GB | 8% |
| Shandong | 164 GB | 7% |
| Fujian | 156 GB | 7% |

All => 访问前十的城市

| | Value | Percent |
|---|---|---|
| Beijing | 385 GB | 22% |
| Shanghai | 249 GB | 14% |
| Shenzhen | 222 GB | 12% |
| Guangzhou | 198 GB | 11% |
| Zhengzhou | 194 GB | 11% |
| Hangzhou | 153 GB | 9% |
| Nanjing | 106 GB | 6% |
| Chengdu | 98 GB | 6% |

All => VIP TOP 10

| | Value | Percen |
|---|---|---|
| 36.139 | 62 GB | 18% |
| 36.139. | 59 GB | 17% |
| 115.192. | 43 GB | 13% |
| 121.4. | 36 GB | 11% |
| 106.39. | 25 GB | 7% |
| 111.18. | 23 GB | 7% |
| 219.143. | 23 GB | 7% |
| 2408:8221: | 23 GB | 7% |

All 系统 TOP 10

| | Value |
|---|---|
| Windows | 3 TB |
| Linux | 1 TB |
| CentOS | 508 GB |
| Fedora | 358 GB |
| Mac OS X | 218 GB |
| Android | 140 GB |
| Red Hat | 54 GB |
| Ubuntu | 50 GB |

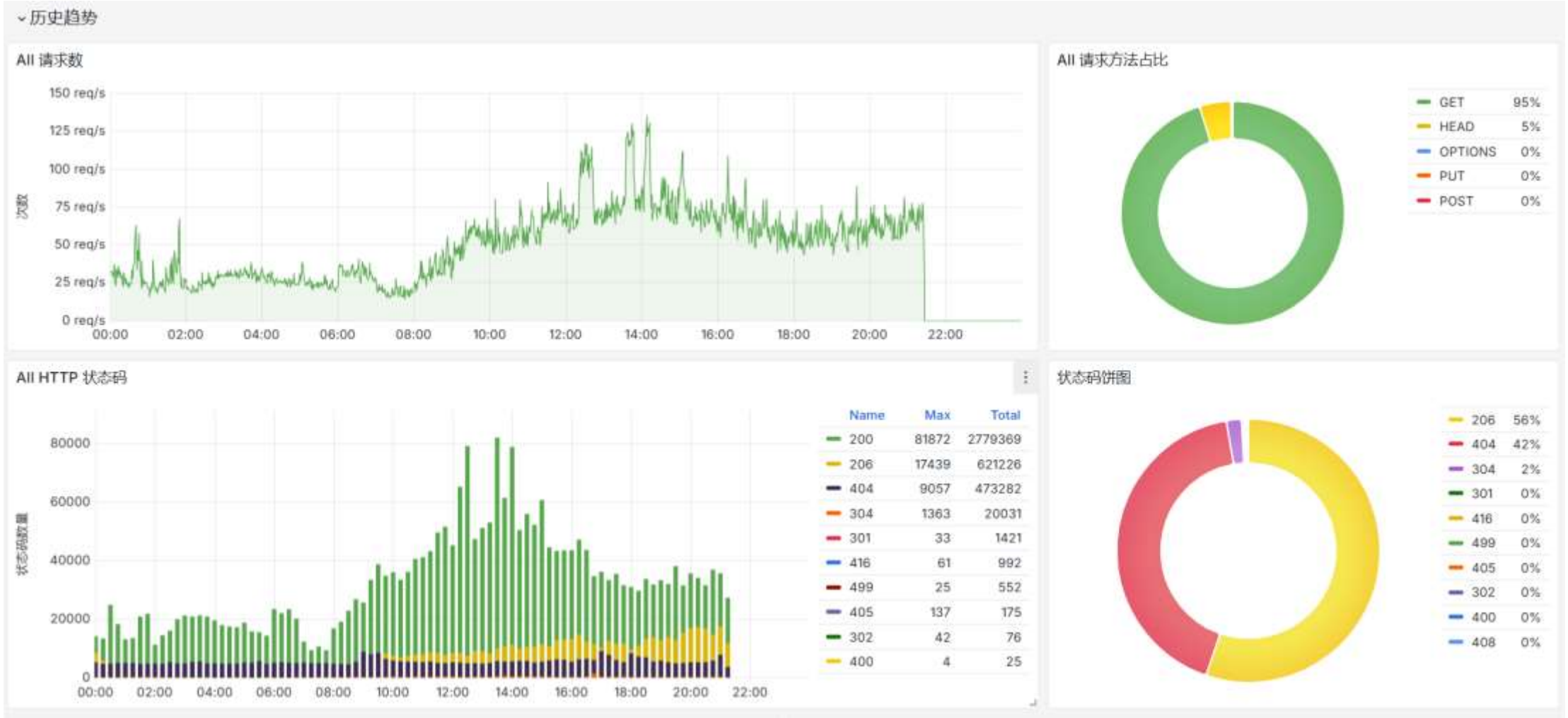
All 移动端 TOP 10

| | Value |
|---|---|
| Other | 7 TB |
| Mac | 218 GB |
| Generic Smartphone | 105 GB |
| Spider | 20 GB |
| 23049RAD8C | 11 GB |
| K | 10 GB |
| XiaoMi Redmi Note 8 Pro | 2 GB |
| Generic Feature Phone | 2 GB |

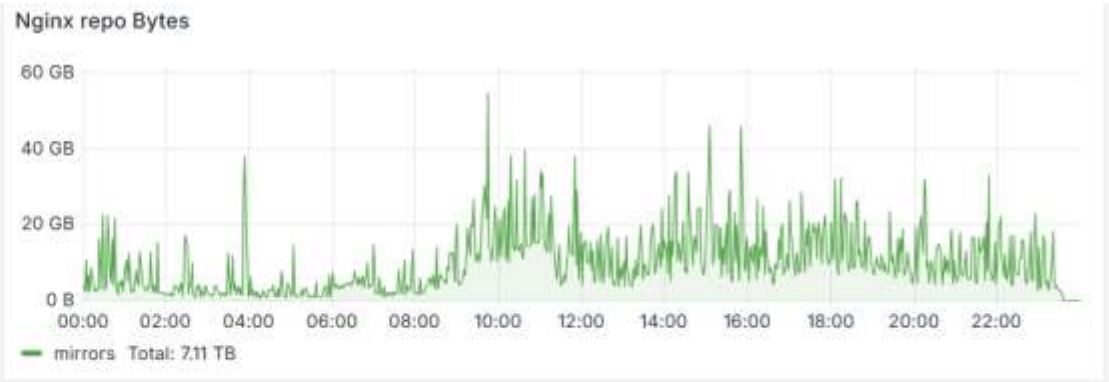
# 日志可视化展示


PV 与 UV
All 今日访问 PV（每10分钟）
60 K
40 K
20 K
0
00:00
03:00
06:00
09:00
12:00
15:00
18:00
21:00
Count
All 今日访问 UV（每10分钟）
6 K
4 K
2 K
0
00:00
02:00
04:00
06:00
08:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
Unique Count source.address
All 7天访问 PV（每10分钟）
Last 7 days
80 K
60 K
40 K
20 K
0
06/27 00:00
06/27 12:00
06/28 00:00
06/28 12:00
06/29 00:00
06/29 12:00
06/30 00:00
06/30 12:00
07/01 00:00
07/01 12:00
07/02 00:00
07/02 12:00
07/03 00:00
07/03 12:00
Count

# 日志可视化展示


历史趋势
All 请求数
150 req/s
125 req/s
100 req/s
75 req/s
50 req/s
25 req/s
0 req/s
次数
00:00
02:00
04:00
06:00
08:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
All 请求方法占比
GET 95%
HEAD 5%
OPTIONS 0%
PUT 0%
POST 0%
All HTTP 状态码
80000
60000
40000
20000
0
状态码数量
Name Max Total
200 81872 2779369
206 17439 621226
404 9057 473282
304 1363 20031
301 33 1421
416 61 992
499 25 552
405 137 175
302 42 76
400 4 25
状态码饼图
206 56%
404 42%
304 2%
301 0%
416 0%
499 0%
405 0%
302 0%
400 0%
408 0%

# 日志可视化展示


数据报表分析

All => 客户端数据详请

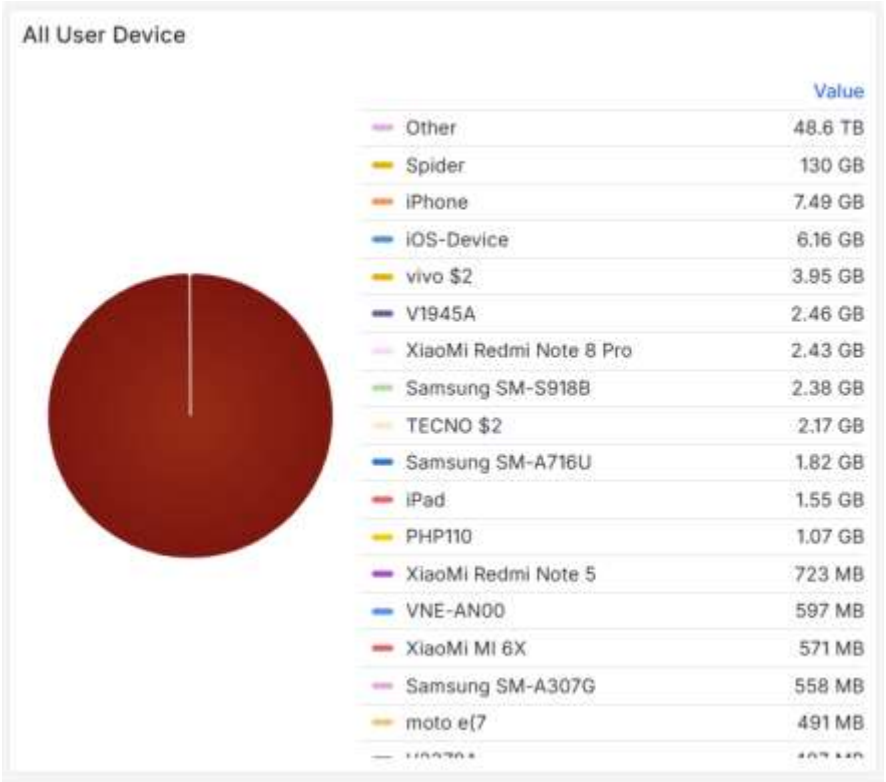
| 客户端IP | 请求次数 | url总数(count) | 返回客户端总流量 |
|---|---|---|---|
| 36.139. | 72 | 4 | 61.74 GB |
| 36.139. | 37 | 4 | 58.93 GB |
| 115.192. | 779 | 757 | 43.38 GB |
| 121.4. | 13,740 | 8 | 36.31 GB |
| 106.39. | 69 | 2 | 25.17 GB |
| 111.18. | 68 | 18 | 22.86 GB |
| 219.143 | 17 | 13 | 22.73 GB |
| 2408:8221. | 4 | 1 | 22.71 GB |
| 119.136. | 2 | 1 | 22.70 GB |

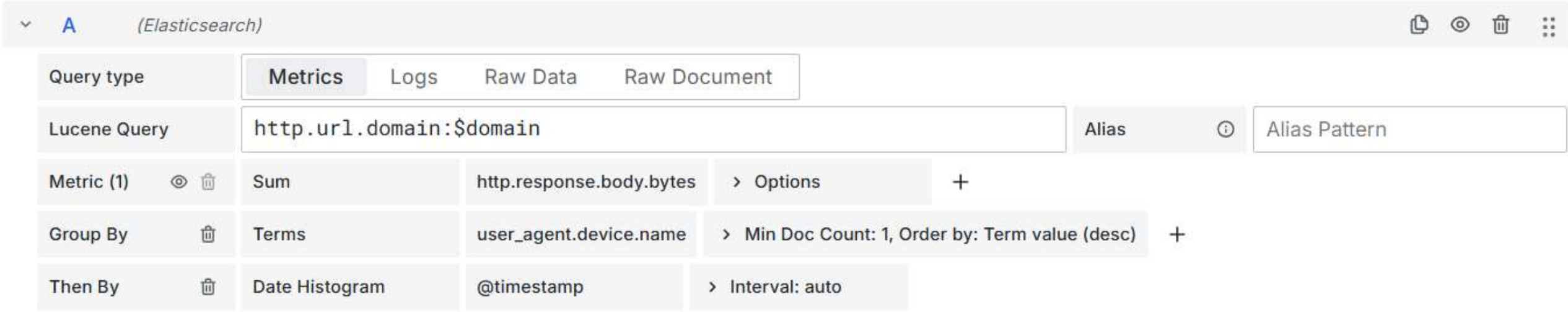
Request Url 数据详请

| 请求URL名称 | 请求总数(pv) | ip总数(uv) | 返回客户端流量 ↓ | 平均响应时间(s) |
|---|---|---|---|---|
| /centos-stream/9-stream/BaseOS/x86_64/iso/CentOS- | 14,538 | 588 | 2.04 TB | 1.94 mins |
| /ubuntu-releases/24.04/ubuntu-24.04-desktop-amd64 | 1,039 | 170 | 617.46 GB | 5.11 mins |
| /fedora/releases/39/Workstation/x86_64/iso/Fedora-W | 4,160 | 590 | 384.54 GB | 52.9 s |
| /epel/8/Everything/x86_64/repodata/cb3c021daed1f23 | 19,320 | 10,963 | 205.99 GB | 14.6 s |
| /fedora/releases/40/Workstation/x86_64/iso/Fedora-W | 1,249 | 151 | 163.79 GB | 4.00 mins |
| /rocky/8.10/AppStream/x86_64/os/repodata/293d7a3b | 19,219 | 6,862 | 159.93 GB | 3.49 s |
| /centos-stream/9-stream/BaseOS/aarch64/iso/CentOS | 1,258 | 66 | 151.41 GB | 1.81 mins |
| /epel/8/Everything/x86_64/repodata/3e09cb5febaec7 | 14,118 | 7,798 | 150.83 GB | 16.2 s |
| /epel/9/Everything/x86_64/repodata/5d56cd107f83d2 | 6,692 | 4,732 | 110.45 GB | 15.5 s |

# 日志可视化展示

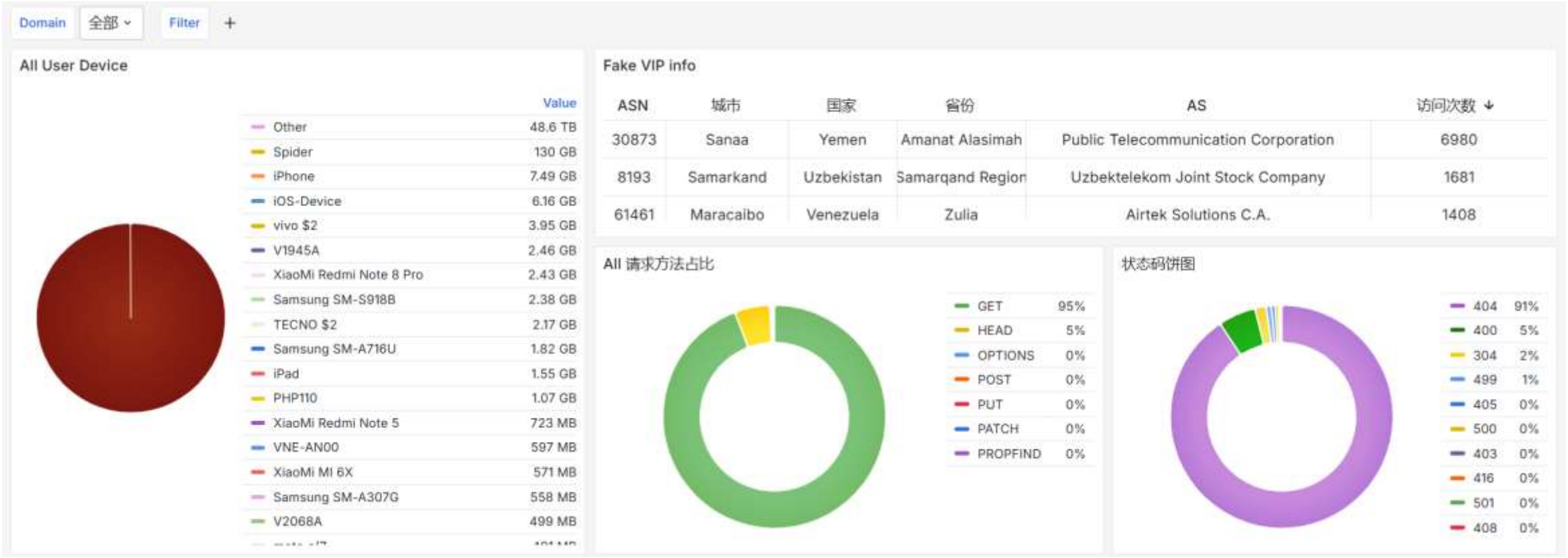
- 对单个 IP/CIDR 的行为进行分析
- UA
- Repo
- Referer
- 以 UA_Device 为例



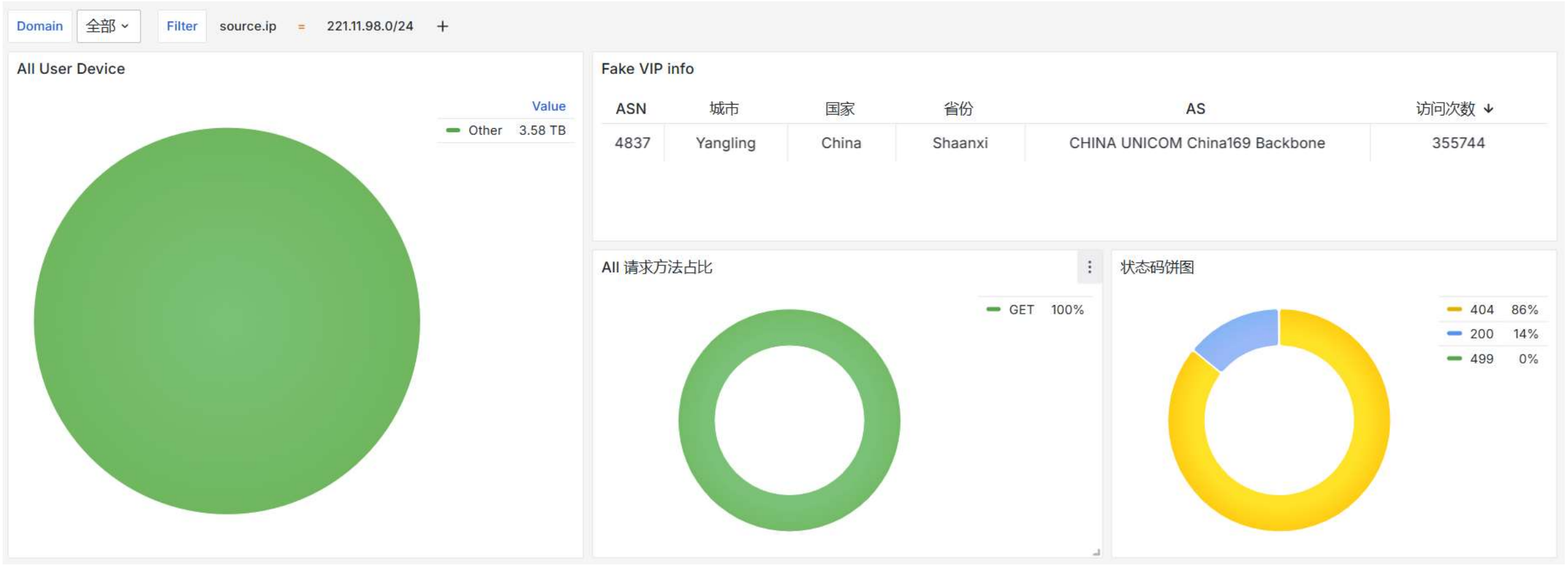
# 日志可视化展示

- 通过filter来对各个panel进行细分查询



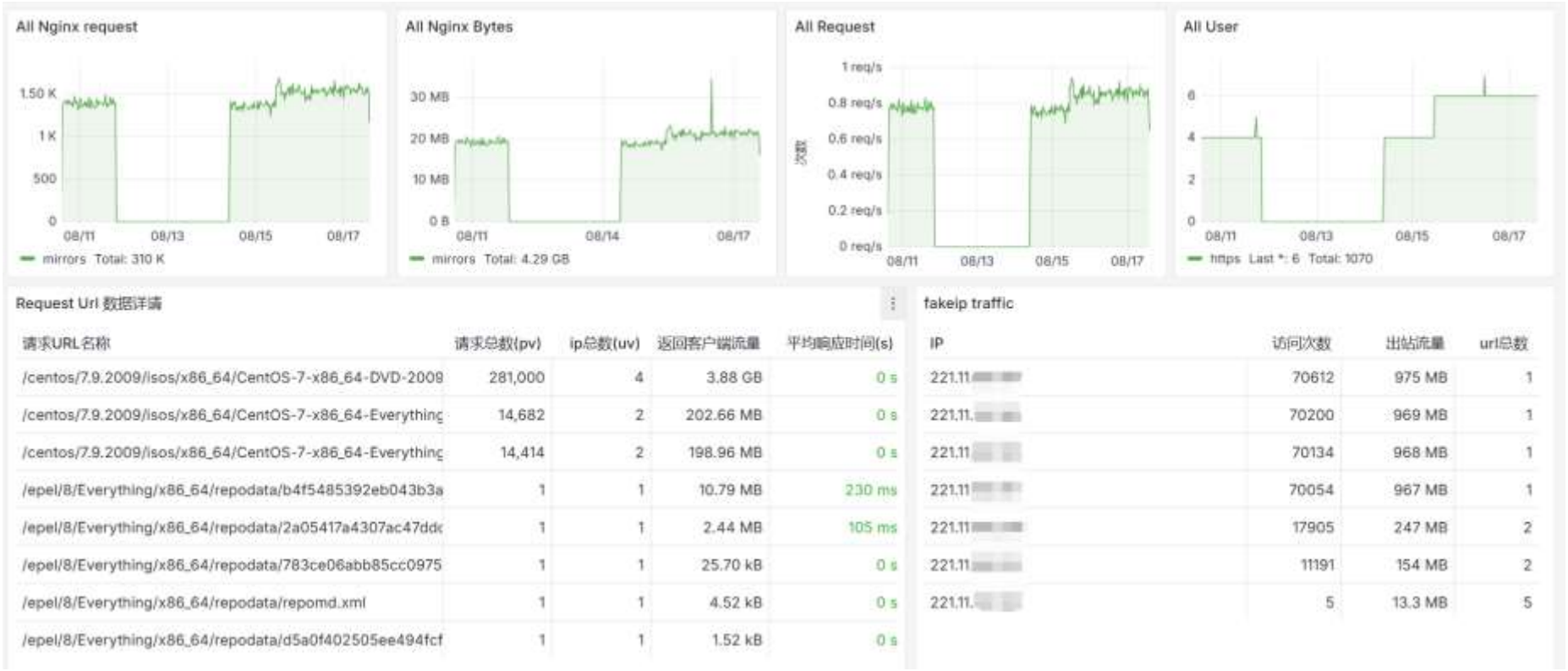
# 日志可视化展示

- 使用CIDR地址段



# 日志可视化展示

- 使用CIDR地址段



# 后续工作

- 恶意流量对策联合工作小组：依靠大脸萌
- 能否将各高校镜像站的日志统计在一起发现潜在问题
- 性能调优、故障转移
- 镜像站双机热备列入实验项目

# 讨论/提问