

Тестовое задание NGENIX на позицию: Аналитик данных сервисов ИБ

I. Визуализация

Преобразуем файл в формат CSV для удобства работы

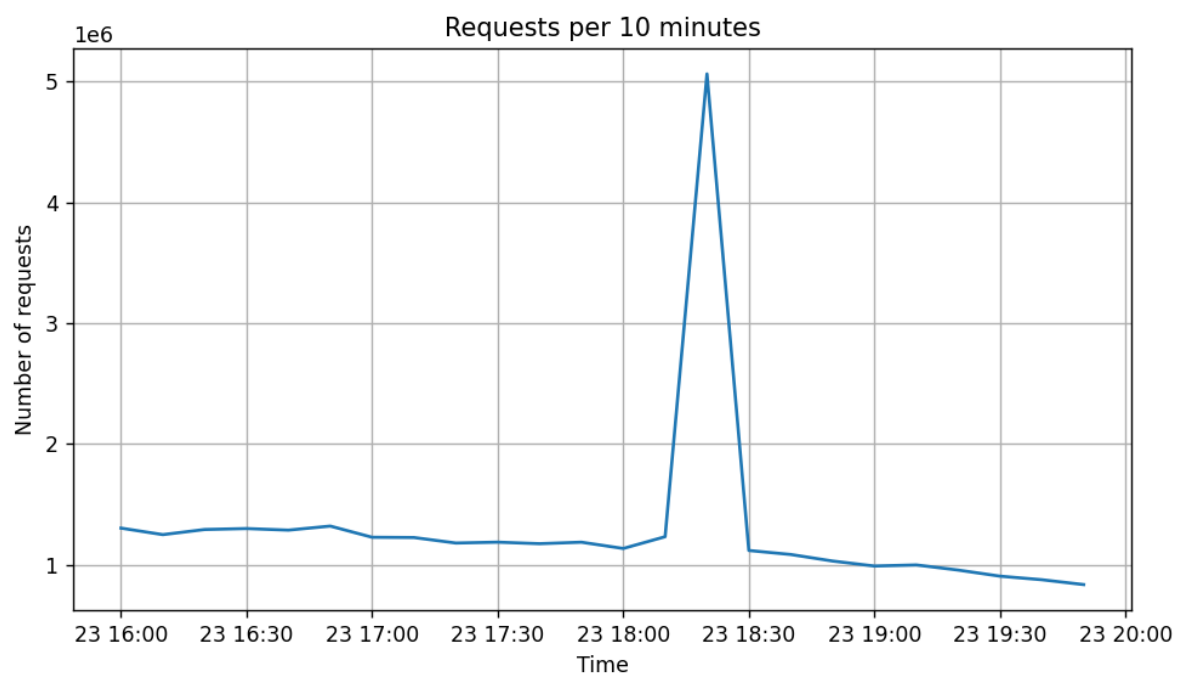
Быстрее всего с точки зрения исполнения запроса это можно сделать через Git bash:

```
MINGW64:/d/Тестовое задание/access.log
Berge@LAPTOP-7ESREHG8 MINGW64 /d/Тестовое задание/access.log
$ cat access.log > access_filtered.csv
```

1.1 Построение общего графика запросов: ось X - время, ось Y - количество запросов

[Код по ссылке](#)

Результат:

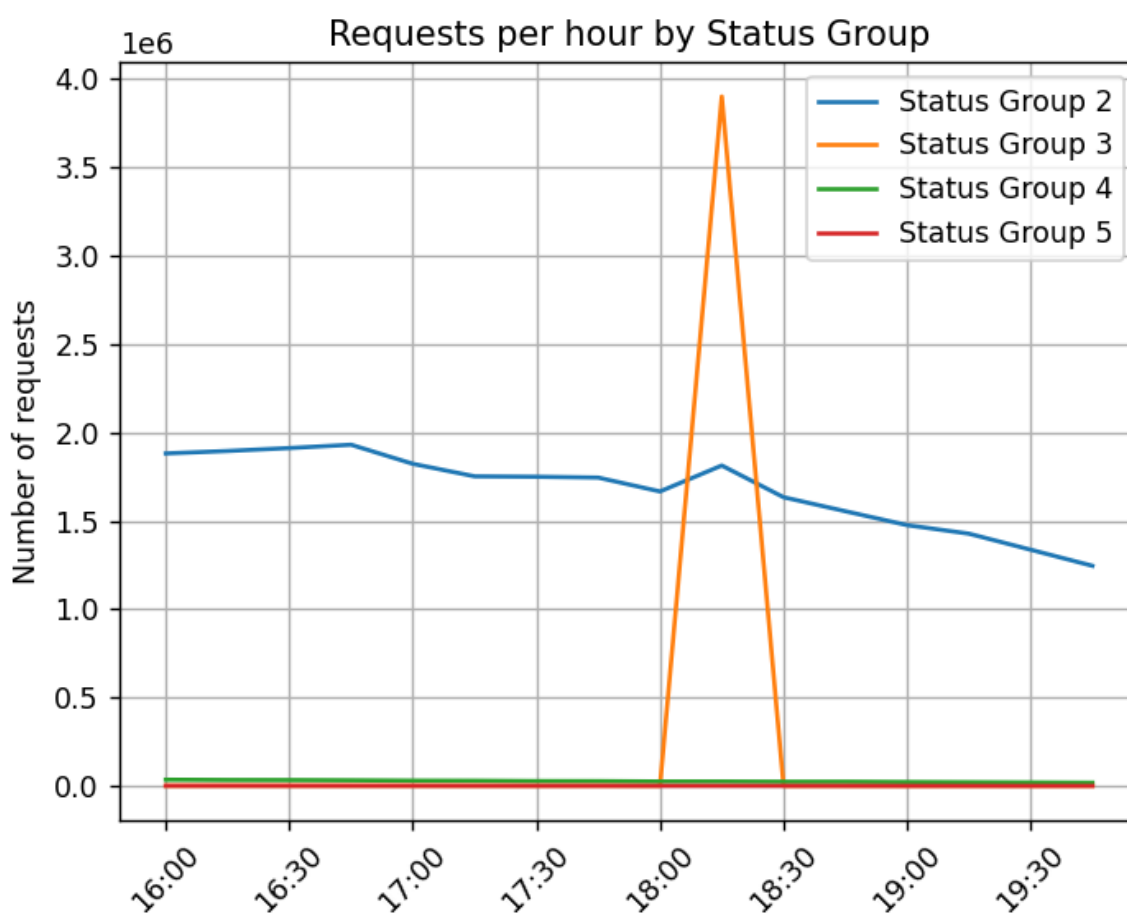


**Ось у отображается логарифмическая, это вызвано наличием больших значений в данных*

1.2 Построение графика количества запросов в зависимости от их группы. Группа запросов определяется первой цифрой: 2xx, 3xx, 4xx, 5xx. Можно на одном графике. Ось X - время, ось Y - количество запросов определенной группы.

[Код по ссылке](#)

Результат:

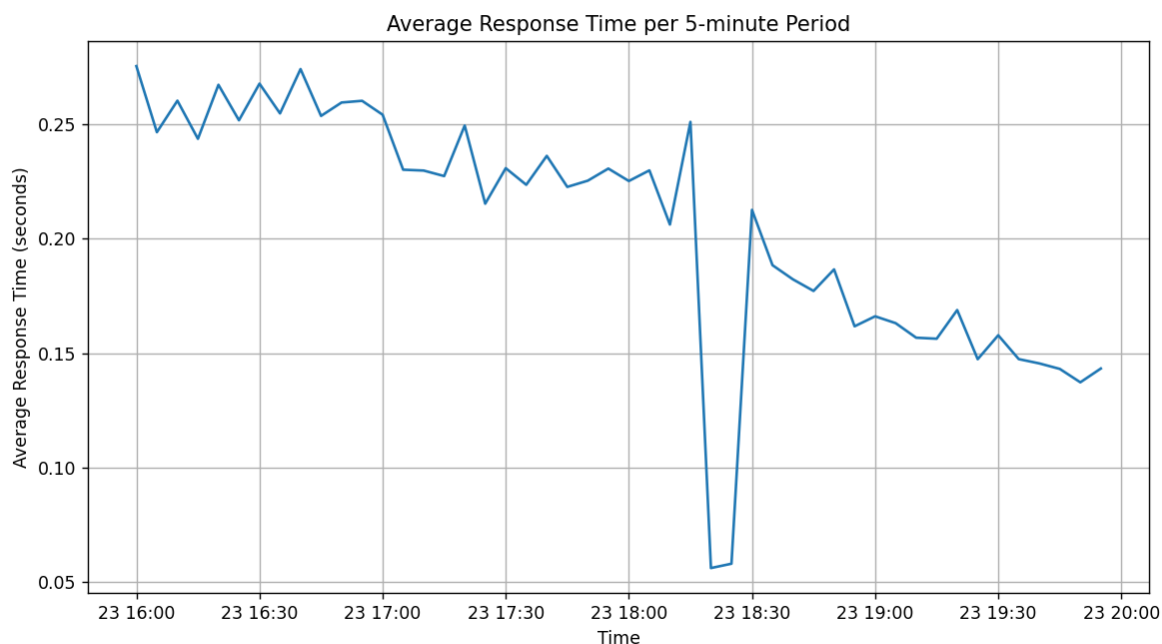


** При построении визуализации статусы 2xx, 3xx, 4xx, 5xx разделены на соответствующие Status Group.*

1.3 Построение графика среднего времени ответа. Значение взяты из поля request time, агрегированы запросы за 5-ти минутный период. Ось X - время, ось Y - среднее время ответа в секундах.

[Код по ссылке](#)

Результат:



II. Анализ

2.1 Определение, были ли аномалии в количестве запросов и временной промежутки, когда они наблюдались.

Начиная с 18:15 и заканчивая примерно в 18:30, согласно 1.1.- общему графику количества запросов, наблюдается резкое повышение количества запросов и резкое снижение.

Такой резкий рост и спад количества запросов со статусом с маской 3xx является аномалией согласно 1.2 График запросов в зависимости от группы, на нем мы можем наблюдать регрессию количества запросов со статусом, имеющих маску: 2xx и аномальный всплеск и резкое понижение количества запросов со статусом, имеющих маску: 3xx.

2.2 Определение паттернов аномалий: характерные признаки запросов, которые относят их к аномалии И описание хода рассуждений по определению аномалий и их признаков.

Аномалия, во временных рамках описанных выше, связана с динамикой количества запросов со status = 3xx (Redirection): Запрос перенаправлен на другой ресурс или местоположение.

2.2.1 Группировка по 'request_method' (HTTP-метод)

Выполним группировку по HTTP-методу и сохраним полученный результат в новый .csv файл:

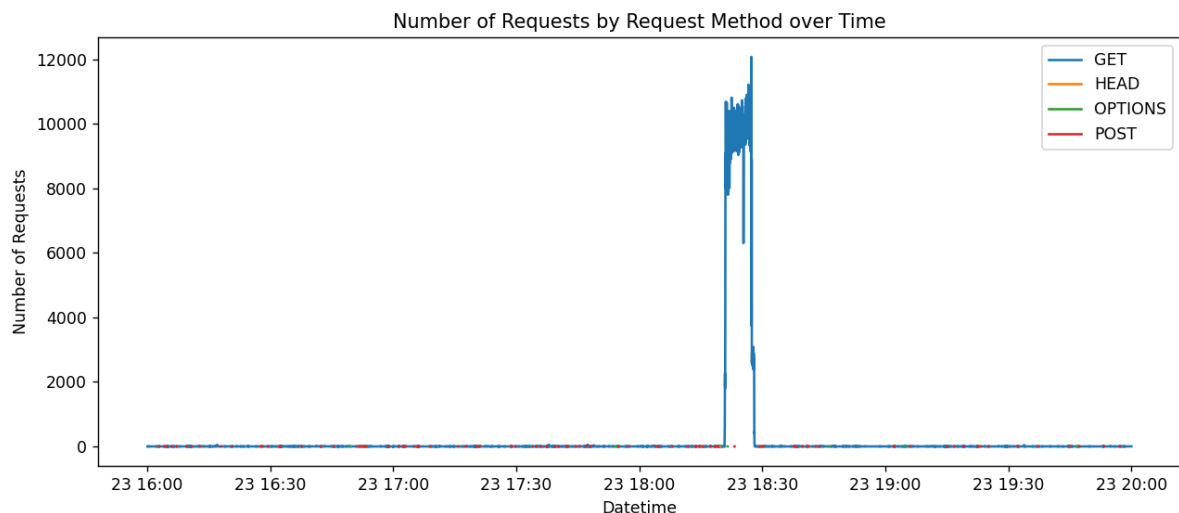
[Код по ссылке](#)

2.2.2 График количества запросов в разрезе групп HTTP-методов

Выполним построим график для визуальной детализации процесса во времени с дальнейшей целью определить конкретный HTTP-метод, являющийся возможным признаком аномалии:

[Код по ссылке](#)

Результат:



Делаем вывод о том, что аномалия наблюдается для единственного Request_method (HTTP-метода) - GET

2.2.3 Выбор запросов, где Request_method - GET за период наблюдения за аномалией

Выделим только запросы, где Request_method = GET за временной период, где наблюдается аномалия и сохраним полученный результат в новый .csv файл (*сужение рассматриваемой выборки запросов или уменьшение количества строк также позволить ускорить скорость выполнения будущих запросов*):

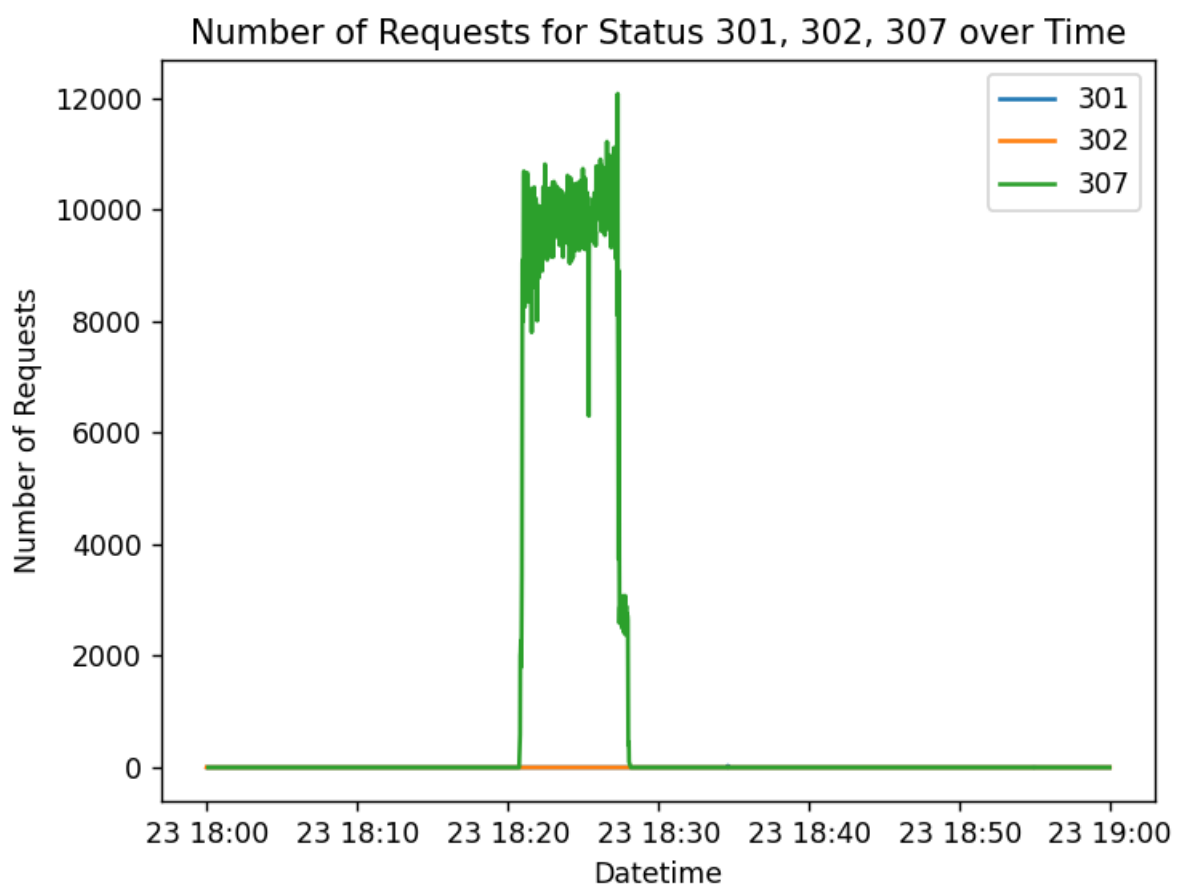
[Код по ссылке](#)

2.2.4 График количества запросов в разрезе групп статусов

Выполним построение графика зависимости от status (301, 302, 307):

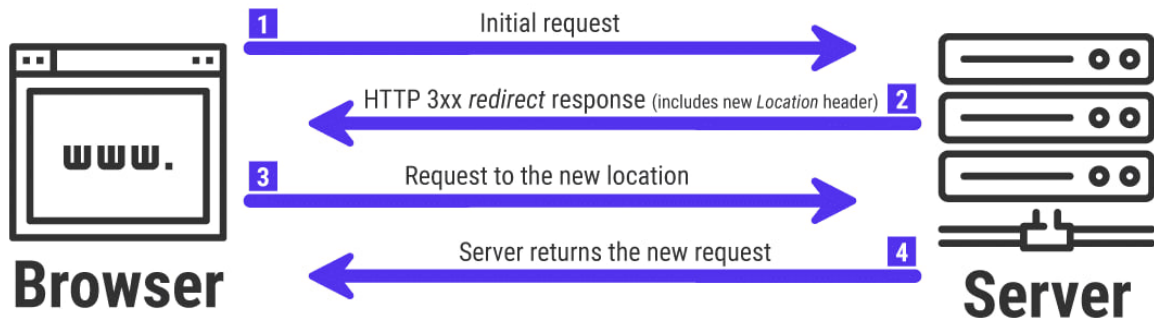
[Код по ссылке](#)

Результат:



Делаем вывод о том, что аномалия наблюдается для единственного status (Статуса) - 307

307 статус (*HTTP 307 Temporary Redirect redirect status response code indicates that the resource requested has been temporarily moved to the URL given by the Location headers*) подразумевает, что будет выполнена переадресация на другой сервер, который отображается в колонке http_referer.



2.2.5 Выбор запросов, где Status - 307

Сузим выборку ([2.2.3](#)) выделив только запросы, где status = 307 и сохраним полученный результат в новый .csv файл (*сужение рассматриваемой выборки запросов или уменьшение количества строк также позволит ускорить скорость выполнения будущих запросов*):

[Код по ссылке](#)

Для визуального удобства подгружаем полученный файл на локальный сервер (dbeaver - Postgre) для дальнейшей аналитики.

2.2.6 Выдвижение гипотезы: запросы со значением `remote_addr` “77.51.117.176” являются причиной аномалии

Было замечено, что для ряда запросов значение в `remote_addr` (IP-адрес источника запроса) равно “77.51.117.176”, см скриншот ниже:

	datetime	host	remote_addr	geo_country_code_variable	connection_requests	connection	ssl_protocol	request_time
1	2022-11-23 18:27:15.000	karelia.somesite.ru	77.51.117.176	RU	153	3064753994	TLSv1.3	0
2	2022-11-23 18:27:15.000	rostov.somesite.ru	77.51.117.176	RU	224	3064753553	TLSv1.3	0
3	2022-11-23 18:27:15.000	karelia.somesite.ru	77.51.117.176	RU	154	3064753994	TLSv1.3	0
4	2022-11-23 18:27:15.000	rostov.somesite.ru	77.51.117.176	RU	225	3064753553	TLSv1.3	0
5	2022-11-23 18:27:15.000	rostov.somesite.ru	77.51.117.176	RU	226	3064753553	TLSv1.3	0
6	2022-11-23 18:27:15.000	rostov.somesite.ru	77.51.117.176	RU	227	3064753553	TLSv1.3	0
7	2022-11-23 18:27:15.000	rostov.somesite.ru	77.51.117.176	RU	228	3064753553	TLSv1.3	0
8	2022-11-23 18:27:15.000	rostov.somesite.ru	77.51.117.176	RU	229	3064753553	TLSv1.3	0
9	2022-11-23 18:27:15.000	rostov.somesite.ru	77.51.117.176	RU	230	3064753553	TLSv1.3	0
10	2022-11-23 18:27:15.000	rostov.somesite.ru	77.51.117.176	RU	231	3064753553	TLSv1.3	0
11	2022-11-23 18:27:15.000	rostov.somesite.ru	77.51.117.176	RU	232	3064753553	TLSv1.3	0
12	2022-11-23 18:27:15.000	karelia.somesite.ru	77.51.117.176	RU	155	3064753994	TLSv1.3	0
13	2022-11-23 18:27:15.000	karelia.somesite.ru	77.51.117.176	RU	156	3064753994	TLSv1.3	0
14	2022-11-23 18:27:15.000	karelia.somesite.ru	77.51.117.176	RU	157	3064753994	TLSv1.3	0
15	2022-11-23 18:27:15.000	karelia.somesite.ru	77.51.117.176	RU	158	3064753994	TLSv1.3	0
16	2022-11-23 18:27:15.000	karelia.somesite.ru	77.51.117.176	RU	159	3064753994	TLSv1.3	0
17	2022-11-23 18:27:15.000	karelia.somesite.ru	77.51.117.176	RU	160	3064753994	TLSv1.3	0
18	2022-11-23 18:27:15.000	kostroma.somesite.ru	77.51.117.176	RU	985	3064628202	TLSv1.3	0
19	2022-11-23 18:27:15.000	karelia.somesite.ru	77.51.117.176	RU	161	3064753994	TLSv1.3	0
20	2022-11-23 18:27:15.000	karelia.somesite.ru	77.51.117.176	RU	162	3064753994	TLSv1.3	0
21	2022-11-23 18:27:15.000	rostov.somesite.ru	77.51.117.176	RU	233	3064753553	TLSv1.3	0
22	2022-11-23 18:27:15.000	karelia.somesite.ru	77.51.117.176	RU	163	3064753994	TLSv1.3	0
23	2022-11-23 18:27:15.000	karelia.somesite.ru	77.51.117.176	RU	164	3064753994	TLSv1.3	0
24	2022-11-23 18:27:15.000	karelia.somesite.ru	77.51.117.176	RU	165	3064753994	TLSv1.3	0
25	2022-11-23 18:27:15.000	karelia.somesite.ru	77.51.117.176	RU	166	3064753994	TLSv1.3	0
26	2022-11-23 18:27:15.000	karelia.somesite.ru	77.51.117.176	RU	167	3064753994	TLSv1.3	0
27	2022-11-23 18:27:15.000	karelia.somesite.ru	77.51.117.176	RU	168	3064753994	TLSv1.3	0
28	2022-11-23 18:27:15.000	karelia.somesite.ru	77.51.117.176	RU	169	3064753994	TLSv1.3	0
29	2022-11-23 18:27:15.000	karelia.somesite.ru	77.51.117.176	RU	170	3064753994	TLSv1.3	0
30	2022-11-23 18:27:15.000	karelia.somesite.ru	77.51.117.176	RU	171	3064753994	TLSv1.3	0

**данный факт был замечен без предварительной фильтрации таблицы по одному из столбцов*

Выдвигаем гипотезу: запросы со значением `remote_addr` (IP-адрес источника запроса) - “77.51.117.176” являются причиной аномалии.

2.2.6.1 Проверка принадлежности запросов со значением `remote_addr` - “77.51.117.176” только для текущей выборки

Выполняем подсчет количества запросов со значением `remote_addr` (IP-адрес источника запроса) - “77.51.117.176” для исходного файла и для текущей выборки, чтобы убедиться, что такой паттерн присущ только для выбранного временного диапазона (18:00-19:00) :

2.2.6.1.1 Подсчет количества запросов со значением `remote_addr` - “77.51.117.176” для исходного файла

Выполняем подсчет количество запросов со значением `remote_addr` (IP-адрес источника запроса) - “77.51.117.176” для исходного файла с использованием Python:

[Код по ссылке](#)

Результат:

```
Количество строк с remote_addr равным '77.51.117.17': 390239  
Process finished with exit code 0
```

2.2.6.1.2 Подсчет количества со значением remote_addr - “77.51.117.176” для выборки

Выполняем подсчет количество запросов со значением remote_addr (IP-адрес источника запроса) - “77.51.117.176” для выборки напрямую в базе данных (для повышения скорости обработки запроса):

[Код по ссылке](#)

Результат: 390233

2.2.6.1.3 Итог проверки

Получаем примерно равное количество в 2.2.7.1 и в 2.2.7.2, что дает основание утверждать, что гипотеза верна и необходимо продолжать определять другие характерные для аномалии признаки в рамках ранее определенной выборки списка транзакции.

2.2.7 Проверка уникальности значения remote_addr “77.51.117.176” для аномалии

2.2.7.1 Группировка по remote_addr и подсчет количества запросов

Определяем является ли значение remote_addr (IP-адрес источника запроса) - “77.51.117.176” единственным для данной аномалии:

[Код по ссылке](#)

Результат:

access_grouped_new3 1		access_group	
select remote_addr, count(remote_addr) from			
Grid		remote_addr	count
1		77.51.117.176	3,900,754
2		85.26.232.251	12
3		31.162.189.85	9
4		93.120.143.196	7
5		176.59.141.251	7
6		84.18.119.28	7
7		176.59.100.42	6
8		128.72.8.22	5
9		5.59.135.174	5

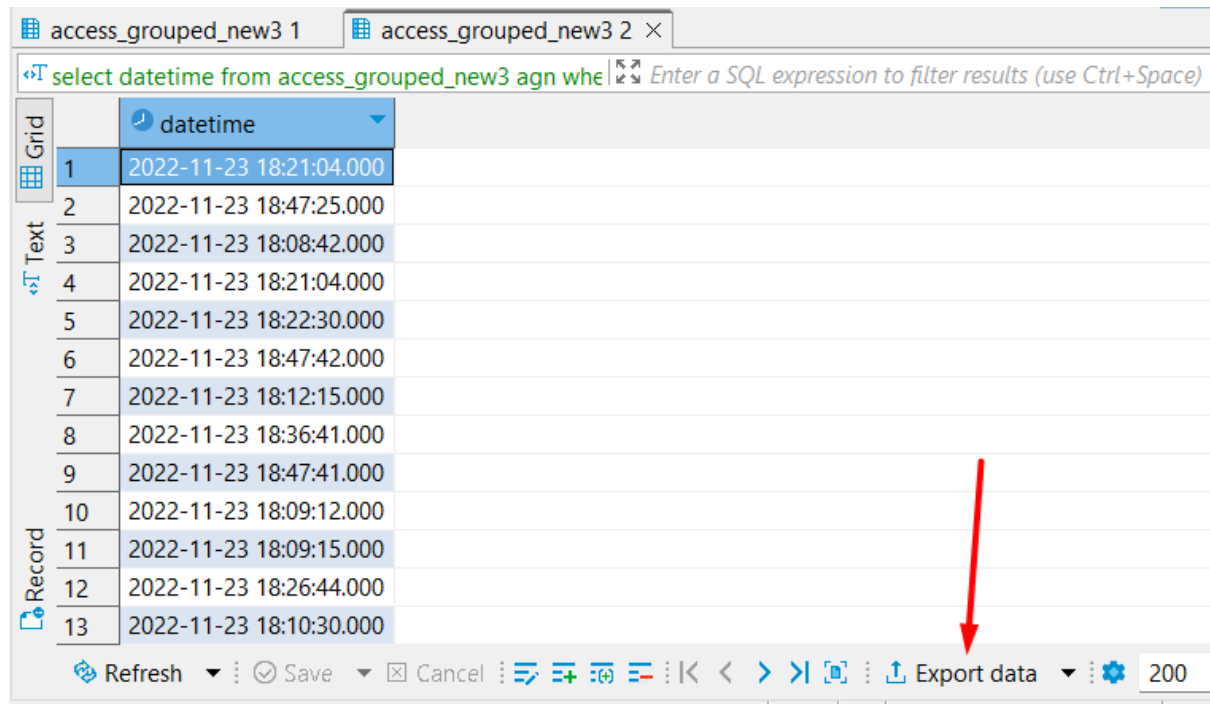
*убеждаемся в уникальности данного признака

2.2.7.2 Выгрузка данных из выборки, исключая значение remote_addr "77.51.117.176"

Напишем и запустим скрипт для выгрузки файла в формате .csv из выборки, исключая значение remote_addr равное "77.51.117.176"

[Код по ссылке](#)

Результат:



	datetime
1	2022-11-23 18:21:04.000
2	2022-11-23 18:47:25.000
3	2022-11-23 18:08:42.000
4	2022-11-23 18:21:04.000
5	2022-11-23 18:22:30.000
6	2022-11-23 18:47:42.000
7	2022-11-23 18:12:15.000
8	2022-11-23 18:36:41.000
9	2022-11-23 18:47:41.000
10	2022-11-23 18:09:12.000
11	2022-11-23 18:09:15.000
12	2022-11-23 18:26:44.000
13	2022-11-23 18:10:30.000

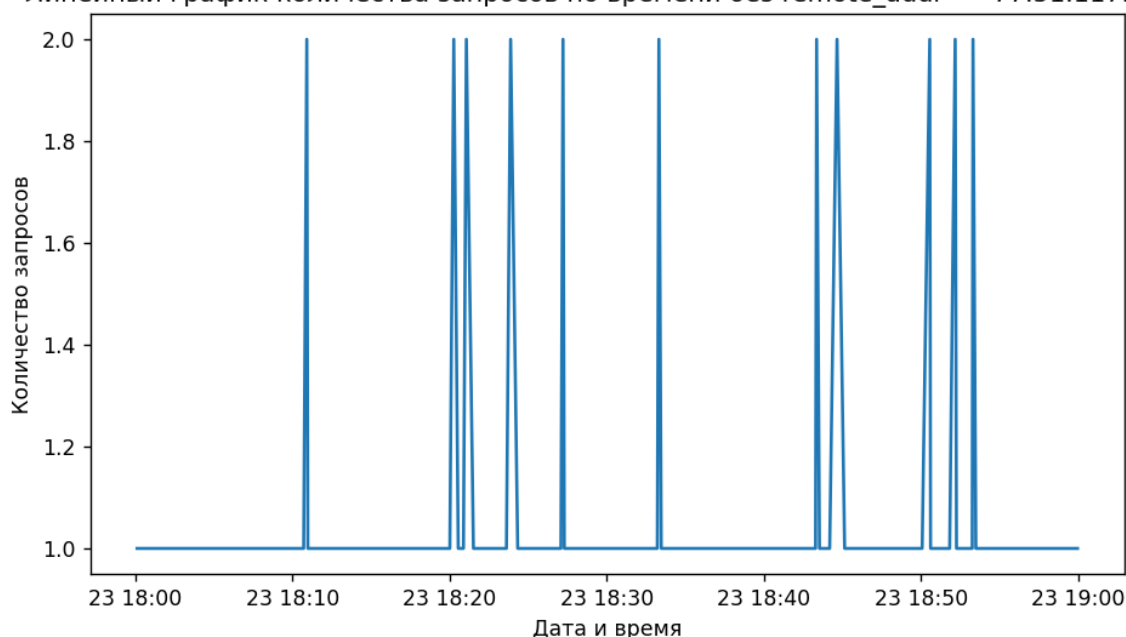
2.2.7.3 График запросов из выборки, исключая значение remote_addr "77.51.117.176"

Выполним построение графика запросов из выборки, исключая значение remote_addr "77.51.117.176" для того, чтобы убедиться, что без данного признака аномалия будет отсутствовать.

[Код по ссылке](#)

Результат:

Линейный график количества запросов по времени без remote_addr = "77.51.117.176"



Таким образом, с помощью данного графика мы убедились в том, что мы локализовали все признаки аномалии

2.2.8 Полный список признаков аномалии

Запросы со следующими признаками являются полными признаками аномалии:

Status - 307

request_method - GET

Datetime - период с 2022-11-23 18:20:49 до 2022-11-23 18:28:03

remote_addr (IP-адрес источника запроса) - "77.51.117.176"

Информацию об IP "77.51.117.176" может быть полезно проверить в открытых источниках:

<https://ru.ipshu.com/ipv4/77.51.117.176#question2>

2.3 Описание алгоритма, который позволит в будущем исключать аномальные запросы такого характера без ручного анализа лог-файла.

Из всех возможных причин, кажется явной использование злоумышленником ботов и сканеров, которые производят автоматические запросы к серверу, обращаясь по различным ссылкам http_referer.

Properties Data ER Diagram postgres Databases			
access_grouped_new3 Enter a SQL expression to filter results (use Ctrl+Space)			
	server_port	http_referer	http_use
10	ost=161&offset=0&limit=20	443 https://kuzbass.somesite.ru/api/exchange/lo	Mozilla/5.0
11	ost=444&offset=0&limit=20	443 https://kuzbass.somesite.ru/api/exchange/lo	Mozilla/5.0
12	ost=641&offset=0&limit=20	443 https://kuzbass.somesite.ru/api/exchange/lo	Mozilla/5.0
13	ost=581&offset=0&limit=20	443 https://kuzbass.somesite.ru/api/exchange/lo	Mozilla/5.0
14	ost=323&offset=0&limit=20	443 https://kuzbass.somesite.ru/api/exchange/lo	Mozilla/5.0
15	ost=180&offset=0&limit=20	443 https://kuzbass.somesite.ru/api/exchange/lo	Mozilla/5.0
16	ost=378&offset=0&limit=20	443 https://kuzbass.somesite.ru/api/exchange/lo	Mozilla/5.0
17	ost=300&offset=0&limit=20	443 https://kuzbass.somesite.ru/api/exchange/lo	Mozilla/5.0
18	st=78&offset=0&limit=20	443 https://kuzbass.somesite.ru/api/exchange/lo	Mozilla/5.0
19	ost=787&offset=0&limit=20	443 https://khakasia.somesite.ru/api/exchange/l	Mozilla/5.0
20	ost=456&offset=0&limit=20	443 https://khakasia.somesite.ru/api/exchange/l	Mozilla/5.0
21	ost=585&offset=0&limit=20	443 https://khakasia.somesite.ru/api/exchange/l	Mozilla/5.0
22	ost=234&offset=0&limit=20	443 -	Mozilla/5.0 (X11; Linux x
23	ost=534&offset=0&limit=20	443 https://khakasia.somesite.ru/api/exchange/l	Mozilla/5.0
24	ost=561&offset=0&limit=20	443 https://volgograd.somesite.ru/api/exchange	Mozilla/5.0
25	ost=624&offset=0&limit=20	443 https://volgograd.somesite.ru/api/exchange	Mozilla/5.0
26	ost=290&offset=0&limit=20	443 https://volgograd.somesite.ru/api/exchange	Mozilla/5.0
27	ost=487&offset=0&limit=20	443 https://volgograd.somesite.ru/api/exchange	Mozilla/5.0
28	ost=189&offset=0&limit=20	443 https://volgograd.somesite.ru/api/exchange	Mozilla/5.0
29	ost=103&offset=0&limit=20	443 https://volgograd.somesite.ru/api/exchange	Mozilla/5.0
30	ost=254&offset=0&limit=20	443 https://volgograd.somesite.ru/api/exchange	Mozilla/5.0
31	ost=524&offset=0&limit=20	443 https://volgograd.somesite.ru/api/exchange	Mozilla/5.0
32	ost=444&offset=0&limit=20	443 -	Mozilla/5.0
33	ost=186&offset=0&limit=20	443 https://volgograd.somesite.ru/api/exchange	Mozilla/5.0
34	ost=624&offset=0&limit=20	443 https://volgograd.somesite.ru/api/exchange	Mozilla/5.0
35	ost=290&offset=0&limit=20	443 https://volgograd.somesite.ru/api/exchange	Mozilla/5.0
36	ost=487&offset=0&limit=20	443 https://volgograd.somesite.ru/api/exchange	Mozilla/5.0
37	ost=189&offset=0&limit=20	443 https://volgograd.somesite.ru/api/exchange	Mozilla/5.0
38	ost=103&offset=0&limit=20	443 https://volgograd.somesite.ru/api/exchange	Mozilla/5.0
39	ost=99&offset=0&limit=20	443 -	Mozilla/5.0

Предлагается создать на web application firewall правило мониторинга, провести его пилотирование и проверить корректность работы.

Как вариант: установить правило, чтобы количество запросов не было более 20 за 1 секунду от одного уникального remote_addr (IP).

Стоит отметить, что данная аномалия является частым случаем, такие сервисы как cloudflare или аналоги уже имеют наборы предустановленных фильтров или правил обработки.