УДК 004.042

# ВИЯВЛЕННЯ ЗАГРОЗ НА ОСНОВІ АНАЛІЗУ ЛОГ-ФАЙЛІВ APACHE

А. А. Пронченко[1], М. В. Коломицев[1]

[1] *Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Фізико-технічний інститут*

## Анотація

Проведено аналіз та запропонована методика виявлення загроз на основі лог-файлів Apache з використанням інструментів обробки великих даних таких, як Elastic Stack.

*Ключові слова*: аналіз, лог-файл, обробка даних, Elastic Stack

## Вступ

Після того, як відбувається інцидент порушення безпеки на сервері, необхідно провести аналіз лог-файлів сервера для виявлення вразливостей, оцінки збитку та збору доказів. Збільшення обсягу інтернет-трафіку, за рахунок зростаючого числа веб-серверів в центрах обробки даних, призводить до великої кількості журналів серверів, які важко аналізувати за допомогою традиційних послідовних методів обробки.

## 1. Проблеми при виявленні вторгнень та інцидентів

Важко розкрити нові моделі атак, які ми не можемо передбачити і які можуть обійти існуючі засоби захисту. Програми, які використовуються для запобігання вторгнень в режимі реального часу, мають важливе значення, але вони мають істотний недолік. Як правило, вони виявляють лише відомі шаблони атак або відомі вектори атак. Хоча розгортання засобів захисту і запобігання від вторгнень в режимі реального часу вкрай важливо, але цього недостатньо. Аналітики повинні докласти творчих зусиль, щоб розкрити нові атаки, які успішно обходять існуючі засоби захисту. Процес включає в себе аналіз даних, зібраних з систем, таких як лог-файли серверів і мережевих пристроїв, а також інформацію з персональних накопичувачів.

У цій роботі ми зосередилися на аналізі даних, а не на їх зборі. Оскільки більшість систем вже збирають багато даних про мережевий і серверний трафік, наша задача зрозуміти, що робити з цими даними для отримання потрібної інформації.

## 2. Формат лог-файлів Apache

Хоча не існує єдиного стандарту для форматів журналів сервера, тим не менш, існує кілька форматів, які часто використовують. Серверне програмне забезпечення Apache, зазвичай, створює файли журналів в одному з декількох форматів: загальний або комбінований. Ці формати можуть бути змінені користувачем, але здебільшого вони використовуються без змін. Комбінований формат відрізняється від загального тим, що до нього додається два поля referrer і user agent. Referrer вказує сайт, з якого клієнт перейшов. User agent показує ідентифікаційну інформацію про браузер клієнта. Лог-файли, що використовуються у роботі, представлені в популярному комбінованому форматі Apache [1].

## Поля комбінованого лог-файлу:

1) Ім'я віддаленого вузла або IP-адреса користувача
2) Логін віддаленого користувача
3) Ідентифікатор користувача, що запитує документ
4) Дата і час, коли був зроблений запит
5) Рядок запиту від клієнта
6) Код стану http, який сервер повернув клієнту
7) Розмір в байтах документа, який був переданий з сервера клієнту
8) URL-адресу сайту, з якого був направлений клієнт
9) Ідентифікаційна інформація в клієнтському браузері або user agent

## 3. Методи аналізу, що використовувалися:

До найбільш популярних методів аналізу належать [2]:

- Нечіткий пошук, за ключовими словами, які пов'язані з відомими векторами атак. Ці вектори можуть включати в себе ін'єкційні атаки, Directory Traversal, отруєння кешу, включення або виконання шкідливого файлу, а також DoS-атаки.
- Підрахунок кількості і відстеження станів невдалих запитів.
- Сортування, фільтрація і об'єднання даних для виявлення потенційних проблемних місць.

## 4. Інструменти, що були використані для аналізу

Використовували для аналізу програмний стек Elastic Stack, що складається з таких програмних

продуктів, як Elasticsearch, Kibana і Filebeat [3]. З його допомогою ми можемо одночасно аналізувати дуже великі набори журналів серверів.

Elasticsearch – це відмовостійка пошукова система на базі бібліотеки Apache Lucene, яка дозволяє організувати ефективну роботу з великим об'ємом даних [4].

Kibana – це сервіс візуалізації і навігації по даним, які проіндексовані в пошуковій системі Elasticsearch. Цей сервіс дозволяє краще аналізувати дані в зручному для користувача інтерфейсі. Kibana надає інформацію у вигляді діаграм різних видів та може візуалізувати дані на географічній карті [5].

## 5. Підрахунок та відстеження станів невдалих запитів

Спроби атак на сервер можуть привести до кодів стану, що вказують як на успіх, так і на помилку. Індикаторами стану в журналах сервера є тризначні коди статусу, які сервер відправляє назад клієнту. Ця інформація дуже цінна, тому що вона показує, чи призвів запит до успішної відповіді (коди, що починаються з 2), перенаправлення (коди, що починаються з 3), помилки, що викликані клієнтом (коди, що починаються з 4), або помилки на сервері ( коди, що починаються з 5) [1]. Часто зловмисник під час проведення атаки може отримати ряд помилок перед тим, як знайти правильну комбінацію, що призведе до успішної атаки. Таким чином, ми можемо сортувати IP-адреси, щоб визначити які з них викликають найбільшу кількість збоїв. Хоча велика кількість збоїв не є остаточною ознакою того, що була проведена атака, але вона може послужити відправною точкою для подальшого розслідування.

Крім проблем безпеки, це також може допомогти визначити, де користувачі відчувають труднощі при користуванні системою, а також ті IP-адреси, які можуть навантажувати систему найбільше.

### 5.1. Аналіз IP-адрес з найбільшою кількістю невдалих запитів

Ще одним із методів аналізу є пошук IP-адрес з найбільшою кількістю невдалих запитів. Під час аналізу наших лог-файлів, ми можемо отримати відсоткове значення відношення невдалих запитів до їх загальної кількості. Найбільший відсоток невдалих запитів буде тоді, коли клієнти зробили тільки декілька запитів або навіть один. У більшості випадків ці записи навряд чи будуть проблемою. Але, якщо б ми побачили дуже велику кількість таких запитів за короткий час, це могло б вказувати на розподілену DoS-атаку, коли бот запрограмований відправляти запити доступу з великої кількості різних IP-адрес.

### 5.2. Аналіз IP-адрес з найбільшою кількістю невдалих запитів за день або за місяць

При аналізі ми можемо виділити IP-адреси з найбільшою кількістю невдалих запитів за день або за місяць. Якщо в який-небудь день надходила дуже велика кількість запитів з деякого IP-адресу вузла, то це могло б вказувати на спробу атаки, наприклад на DoS-атаку. В цьому випадку можна побачити, що IP-адреси з найбільшою кількістю невдалих запитів в день є тими ж, що генерують велику кількість запитів загалом.

## 6. Пошук за ключовими словами

Найбільш безпосереднім способом виявлення спроб атаки в журналах сервера є пошук в полі «request». В цьому полі відображається інформація про запит користувача до серверу, через це багато атак залишають за собою сліди в цьому полі, які нам потрібно знайти. Був сформований ряд запитів для пошуку по відомим векторам атак за допомогою регулярних виразів "regexp"та Query DSL (Domain Specific Language) на основі JSON.

## Висновки

Використовуючи можливості Elastic Stack вдалося реалізувати гнучку методику аналізу логів для пошуку атак у великій кількості лог-файлів серверів. Через те,що вектори атак досить різноманітні, ця методика є дуже корисним та гнучким інструментом, що дозволяє виконувати деталізацію і спеціальний аналіз найшвидше. Проте, також корисно мати набір методів для аналізу загальних векторів атак в якості відправної точки.

## Перелік використаних джерел

1. Apache. Combined Log Format — `https://httpd.apache.org/docs/2.4/logs.html`.
2. M. Talabis, Robert McPherson, I. Miyamoto, J. Martin, D. Kaye. Information Security Analytics. — 2015.
3. Filebeat — `https://www.elastic.co/beats/filebeat`.
4. Elasticsearch — `https://www.elastic.co/elasticsearch/`.
5. Kibana — `https://www.elastic.co/kibana`.