М. М. Гурьев, В. А. Иванов, А. А. Оленников
Тюменский государственный университет, г. Тюмень
УДК 004.054

АНАЛИЗ ЛОГ-ФАЙЛОВ С ПРИМЕНЕНИЕМ МАШИННОГО ОБУЧЕНИЯ

Аннотация. В этой статье рассматривается рынок решений систем управления событиями безопасности. Далее, на основе этого рассмотрения, с учетом плюсов и минусов этих решений будут сформированы критерии и методы реализации для собственного решения на основе машинного обучения для анализа лог-файлов.

Ключевые слова: машинное обучение, лог-файлы, управление событиями безопасности, анализ текстовых данных.

Введение. В нынешних реалиях все больше и больше устройств задействуется для создания сложных сетей внутри государственных и медицинских учреждений, промышленных предприятий и т. д. Параллельно этому с каждым годом увеличивается количество кибератак на эти сети [1, 2]. Исходя из усложнения инфраструктур, сетевым оборудованием генерируется все больше и больше лог-файлов, из-за количества которых сотрудник информационной безопасности не в состоянии оперативно реагировать на возможные инциденты. Как следствие этого увеличивается количество успешных инцидентов, негативно влияющих на работу систем. В данной ситуации искусственный интеллект (ИИ) мог бы существенно оптимизировать работу сотрудников ИБ. На данный момент существуют разные решения по обучению ИИ для анализа лог-файлов, например, DeepLog [3], LogAnomaly, Logsy и др. [4, 5]. Для оценки лог-файлов искусственным интеллектом необходима предварительная их обработка. Выбор решений программы-парсера на рынке довольно широк. Основываясь на сравнениях из исследований J. Zhu и др. [6] и P. He и др. [7], можно подобрать наиболее подходящее решение для обработки лог-файлов.

Проблема исследования. Снижается общий уровень информационной безопасности инфраструктуры. Поэтому наличие программного решения внутри инфраструктуры для анализа лог-файлов,

который будет способствовать оперативному реагированию на инциденты либо выявлению аномального поведения, является особо важным.

Материалы и методы. Для начала обратимся к сравнительной таблице 1, основой которой является другая сравнительная таблица из статьи González-Granadillo G. и др. [8].

Таблица 1

Сравнение SIEM решений

	Обработка online	*Визуализация*	*Криминалистика*	*Сложность развертывания*	*Масштабируемость*	*Анализ рисков*	*Возможности реагирования и отчетности*	*Машинное обучение*
ArcSight	+	-	-	-	+	-	-	-
QRadar	+	+/-	+	+/-	+	+/-	-	-
McAfee	+	+/-	+	+/-	+	+/-	+	-
LogRhythm	+	+/-	+/-	+	+	+/-	+	+
USM-OSSIM	+	+/-	+	+/-	-	-	-	-
RSA	+	+/-	+	+/-	+	+/-	+/-	
Splunk	+	+	+/-	+	+	-	+/-	-
SolarWinds	+	+/-	+/-	+/-	+	+/-	+/-	-

В ней выделены несколько критериев и их степень реализации в соответствующих SIEM решениях («+» — высокий уровень реализации; «-» — низкий уровень реализации или ее отсутствие; «+/-» — средний уровень реализации). Основываясь на этой сравнительной таблице, можно сделать вывод, что на рынке каждое решение имеет свои недостатки. Разработка нашего решения будет учитывать, как положительные моменты в тех или иных коммерческих решениях, так и отрицательные. Главной отличительной чертой будет являться наличие машинного обучения. Это даст преимущество нашего решения над другими.

В ходе работы была спроектирована модель будущего клиент-серверного приложения для предотвращения инцидентов информационной безопасности, которая представлена на рис. 1. Модель разделяется на две составляющие: мониторинг трафика в реальном времени и анализ трафика в реальном времени. Работа модели начинается в части мониторинга трафика, где выделенные серверы генерируют в себе лог-файлы. Клиент, установленный на данном сервере, который будет проводить обработку этих лог-файлов. Далее, лог-файлы с обработчика передаются на процесс машинного обучения, где на выходе получается уже полностью обработанный поток лог-файлов. Вся эта информация передается на сервер с разделом анализа трафика в реальном времени, где входе этого анализа создается визуализация поступающего потока и постоянные попытки обнаружения аномалий, при обнаружении которых происходит оповещение и предупреждение.

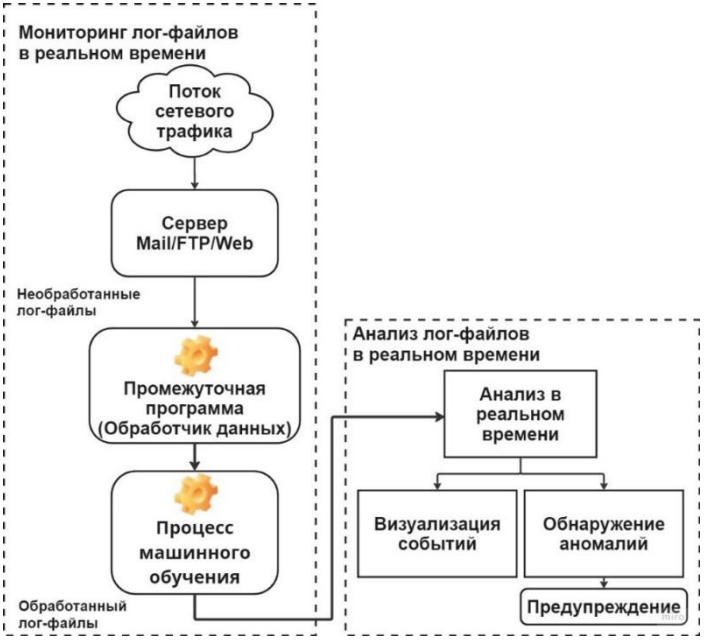

Рис. 1. Модель работы приложения

В качестве промежуточного сервиса по обработке лог-файлов планируется использовать готовые решения с открытым кодом. При выборе мы будем опираться на исследование, посвященное сравнению инструментов, предназначенных для парсинга лог-файлов [6, 7]. Для парсинга будут использованы логи-файлы с серверов FTP, WEB, Mail и DNS. Особое внимание будет уделяться таким полям, как: хост, время подключения, методы подключения и операции, которые происходят во время подключения.

Для машинного обучения приято решение использовать методы, предложенные в статье Chen Z. и др. [3-5].

Результаты. В результате разработки концепта предлагаемого программного решения были сформированы следующие задачи:

1. Изучить архитектуру существующих открытых систем.
2. Учесть минусы и плюсы существующих решений.
3. Собрать базу лог-файлов для обучения искусственного интеллекта.
4. Разработать парсер для разных типов лог-файлов.
5. Внедрить визуализацию результатов работы искусственного интеллекта.
6. Протестировать систему на пробной сети.
7. Реализовать возможность проведения аудита системы.

Заключение. В итоге работы было положено начало разработки программного решения для анализа лог-файлов с использованием машинного обучения. Этот сервис позволит полноценно проводить тотальный мониторинг и анализ трафика в существующих локально-вычислительных сетях, а также будет полезен при проведении компьютерно-криминалистических экспертиз. В последующих публикациях будут подробно описаны архитектура программного решения и результаты анализа лог-фалов выгруженных с сетевых узлов различных инфраструктур.

СПИСОК ЛИТЕРАТУРЫ

1. Отчет об атаках на онлайн-ресурсы российских компаний за 2022 год. — Текст : электронный // Ростелеком-Солар официальный сайт компании. Ростелеком-Солар-Гарантия Кибербезопасности:

[сайт]. — URL: https://rt-solar.ru/analytics/reports/3289/ (дата обращения: 19.05.2023).
2. Статистика НКЦКИ по информационной безопасности. — Текст: электронный // Интернет-портал по информационной безопасности в сети | Для пользователей и специалистов: [сайт]. — URL: https://safe-surf.ru/users-of/media/nkcki-statistics/ (дата обращения: 19.05.2023).
3. DeepLog: Anomaly Detection and Diagnosis from System Logs through Deep Learning / M. Du, F. Li, G. Zheng, V. Srikumar. — Текст: непосредственный // CCS '17: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. — New York: Association for Computing Machinery, 2017. — С. 1285-1298.
4. LogAI: A Library for Log Analytics and Intelligence / Q. Cheng. — Текст : электронный // arXiv.org e-Print archive: [сайт]. — URL: https://arxiv.org/abs/2301.13415 (дата обращения: 19.05.2023).
5. Robust Log-Based Anomaly Detection with Hierarchical Contrastive Learning / Y. Zhao, R. Yang, N. Yang [и др.]. — Текст: непосредственный // ICASSP 2023 — 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). — Rhodes Island: IEEE, 2023. — С. 1-5.
6. Tools and Benchmarks for Automated Log Parsing / J. Zhu, S. He, J. Liu [и др.]. — Текст: непосредственный // 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSESEIP). — Montreal: IEEE, 2019. — С. 121-130.
7. An Evaluation Study on Log Parsing and Its Use in Log Mining / P. He, J. Zhu, S. He [и др.]. — Текст: непосредственный // 2016 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks. — Toulouse: IEEE, 2016. — С. 654-661.
8. González-Granadillo G. Security Information and Event Management (SIEM): Analysis, Trends, and Usage in Critical Infrastructures / G. González-Granadillo, S. González-Zarzosa, R. Diaz. — Текст: непосредственный // Sensors. — 2021. — № 14. — С. 3-8.