# Анализ подходов и средств обработки сервисных журналов

А.Н. Шепелев, А.А. Букатов, А.В. Пыхалов, А.Н. Березовский

## 1. Введение

При обслуживании современных информационных систем возрастает роль автоматического анализа информации, полученной из журналов различных служб (сервисных журналов). Как правило, информация в журналах представлена набором записей, имеющих схожую структуру. При рассмотрении конкретного сервисного журнала можно разделить все содержащиеся в нем записи на группы, которые характеризуются различиями в структуре записей. Информация хранится в текстовом виде, поэтому отдельные записи могут быть легко прочитаны и сгруппированы человеком. Однако объемы сервисных журналов велики, их размер может изменяться от десятков до миллионов записей. В результате этого анализ данной информации становится весьма трудоемкой задачей, требующей автоматизации.

### *Области применения средств обработки сервисных журналов*

Основной областью применения рассматриваемых средств является администрирование компьютерных систем и сетей.

В качестве формальных задач можно выделить:

- получение статистики работы сервисов;
- обнаружение ошибок в работе сервисов;
- выявление причин и источников нестандартного поведения системы;

Для достижения данных целей как правило необходимо решение следующих технических задач:

- распознавание данных в текстовом представлении;
- анализ полученной в ходе распознавания информации;

- перевод информации в структурированный вид;
- поиск информации, удовлетворяющей определенным условиям;
- агрегация информации по одному или нескольким параметрам;
- обработка взаимосвязанной информации из нескольких журналов;
- использование логических переходов при обработке информации;
- представление обработанной и проанализированной информации;

Следует отметить, что приведенные технические задачи встают не только при обработке сервисных журналов. Указанные задачи возникают во многих сферах, требующих автоматического анализа и обработки информации, представимой в текстовом виде, например, при обработке данных о сетевом трафике. Данное представление может быть получено на основе генерируемых утилитой tcpdump журналов [1]. Другими источниками информации, требующей обработки, могут служить различные системы сбора информации (сетевые устройства, датчики, счетчики).

Укажем несколько реальных задач, требующих автоматической обработки сервисных журналов.

- Сбор данных о работе конкретной службы за большой период времени. Полученные данные могут быть использованы для оценки и оптимизации работы службы.
- Мониторинг работы сервиса в режиме реального времени. Регулярное обновление данных о работе системы может позволить быстро отреагировать на внезапные отклонения в режиме работы.
- Мониторинг и анализ сетевого трафика. Проанализировав информацию сетевых пакетов за определенный период времени, можно получить данные о текущих отклонениях в работе сети [2]. Помимо этого, именно статистические данные о сетевом трафике помогают выявлять различные виды компьютерных атак. Различные системы обнаружения и предотвращения вторжения анализируют сетевой трафик для обеспечения

безопасности компьютеров и сети. Среди подобных систем следует отметить SNORT [3], REAL Secure [4], Bro [5]. Следует подчеркнуть, что хотя указанные системы и обеспечивают обнаружение атак и вторжений, для выявления источников и последствий атак может потребоваться дополнительный аудит, включающий, в частности, и анализ системных журналов по дополнительным критериям и сценариям. Среди наиболее распространенных атак следует отметить Distributed Denial of Service (DDoS) атаки [6], сниффинг пакетов, IP-спуффинг, а также различные виды сетевой разведки.

- Отслеживание конкретных ошибок в системе для мгновенного оповещения и принятия мер по автоматизированному устранению проблем.

Таким образом, круг задач, решаемых анализаторами сервисных журналов достаточно широк, не ограничен узким перечнем и может расширяться, что делает актуальным разработку и применение универсальных и конфигурируемых средств анализа журналов.

Целью данной работы является анализ существующих средств и подходов к обработке сервисных журналов для выработки подходов к разработке и реализации универсальных и конфигурируемых средств анализа журналов. В ходе анализа требуется оценить возможности и области применения рассмотренных средств, а также определить наиболее удачные технические решения для возможного использования в ходе реализации системы анализа сервисных журналов.

## 2. Требования к средствам автоматической обработки сервисных журналов

Обычно сервисный журнал состоит из набора текстовых записей. Рассмотрим пример записи журнала access.log веб-сервера nginx.

85.26.184.8 - - [08/Nov/2013:02:13:42 +0400] sfedu.ru "GET /www/docs/F13635/50IarhiMG_1003_md.jpg HTTP/1.1" 200 41604

"http://sfedu.ru/" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/30.0.1599.101 Safari/537.36"

В приведенном примере записи журнала содержится представленная в виде текстовых полей информация, указанная ниже при рассмотрении способа разбора записи средствами шаблона. Множество всех записей конкретного журнала можно разделить на подмножества записей, имеющих схожую структуру. Для определения принадлежности записей к различным подмножествам используются шаблоны. Здесь и далее шаблоном будем называть групповое объединение по типу записи, позволяющее идентифицировать и различать записи сервисного журнала и "вычленять" параметры этих записей. Указанные действия выполняются путем сопоставления записи шаблону (т.е. разбора структуры записи в соответствии со структурой шаблона). Как правило, шаблон имеет текстовое представление, например в виде регулярных выражений. Рассмотрим пример шаблона, способного распознать и вычленить параметры приведенной выше записи.

^([\d.]+) - (\S+) \[(.*)\]  (\S+) \"(\S+) (\S+) (\S+)\" (\d{3}) (\d+) "\S+\" "(.*)\"

Под параметром шаблона будем понимать последовательность символов шаблона, определяющую положение и  вид искомого значения в анализируемой записи журнала. В приведенном примере, параметры шаблона заключены в круглые скобки. Разберем искомые значения, определяемые данным шаблоном.

1. IP-адрес клиента, пославшего запрос веб-серверу.
2. Идентификатор удаленного пользователя.
3. Дата обращения к веб-серверу.
4. Виртуальный сервер, к которому ведется обращение.
5. Тип запроса.
6. Запрашиваемый ресурс.
7. Протокол запроса.
8. Код ошибки веб-сервера при ответе на запрос.

9. Объем переданных данных в байтах.
10. Информация о клиенте.

Под параметрами записи будем понимать реальные значения, сопоставленные соответствующим параметром шаблона. Назовем правилами фильтрации условия отбора требуемых пользователю записей. В соответствии со сказанным выше, анализ сервисных журналов и требуемый для его выполнения разбора записей таких журналов выполняется специализированным автоматическим анализатором. Под логикой разбора, переходов и обработки будем понимать набор специализированных команд, задающих поведение автоматического анализатора при работе с сервисным журналом. Считаем, что данные команды заданы посредством встроенного языка программирования, ориентированного на создание инструкций разбора, переходов и обработки сервисных журналов.

Основным требованием к автоматическому анализатору является возможность задавать логику разбора записей сервисных журналов. Под разобранной записью здесь и далее понимается структура данных, являющаяся результатам процесса распознавания записи (сопоставление ее с шаблоном), содержащая в себе параметры записи, определяющие её содержание. Для обеспечения использования анализатора как универсального средства разбора сервисных журналов, требуемое решение должно иметь поддержку встроенного языка программирования, на котором задается логика разбора. Данный язык должен позволять:

- описывать шаблоны записей;
- задавать правила фильтрации записей;
- выполнять вызов обработчика записи;
- выполнять логический разбор записи;

Под логическим разбором записи будем понимать:

- Анализ нескольких взаимосвязанных записей журнала с определением логики переходов между записями на основе параметров записи.

- Анализ информации нескольких взаимосвязанных журналов с определением логики переходов между журналами на основе параметров записи.

Под обработкой записи будем понимать:

- выполнение внешних скриптов с передачей в качестве аргументов параметров записи;
- выполнение вывода сообщений на консоль;
- оповещение администратора сервиса (например, посредством отправки e-mail или SMS сообщений).

## 3. Обзор известных средств автоматизации анализа сервисных журналов

Рассмотрим следующую схему сценария разбора сервисного журнала. При стандартном обходе анализатор по очереди обрабатывает все записи в журнале или ожидает появления новых записей. На рис. 1 показан цикл разбора очередной записи сервисного журнала.

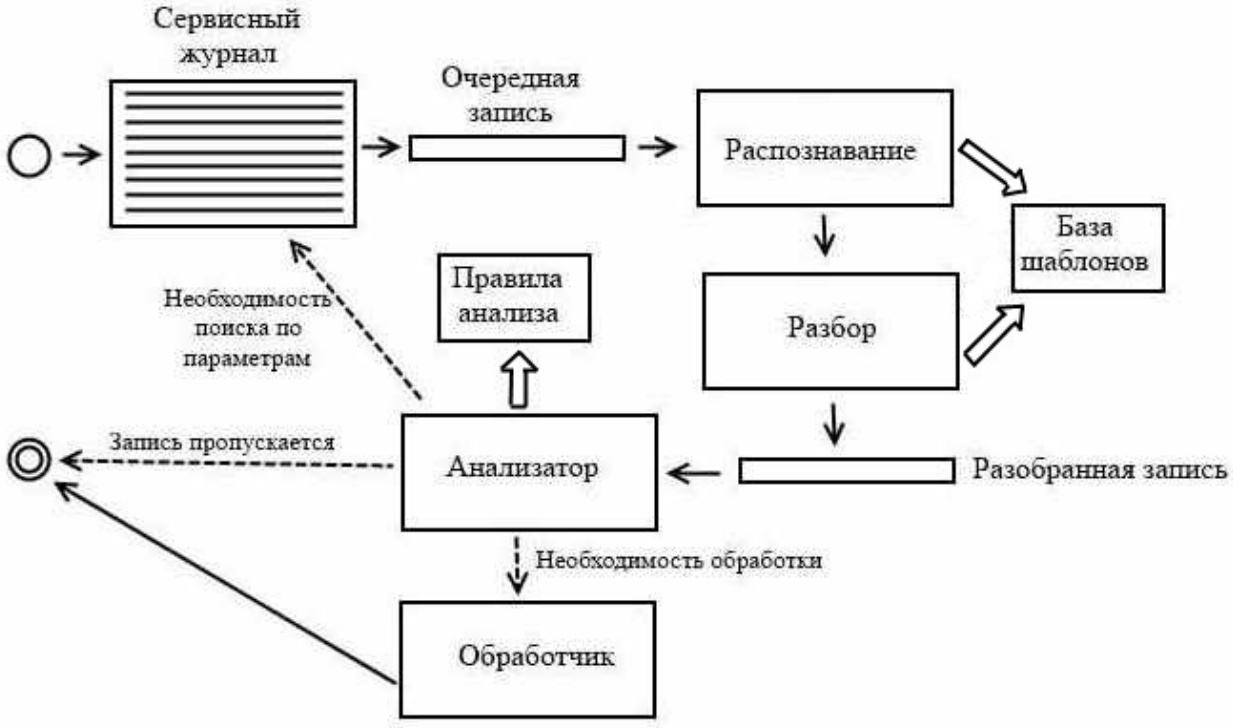


Рис. 1. – Схема цикла обработки записи

В процессе разбора очередной записи сервисного журнала выполняются два основных действия.

- Разбор записи. Сюда следует отнести определение шаблона данной записи, а также уточнение параметров записи. Данный механизм разбора получает на вход запись сервисного журнала в исходном виде. На выходе получаются данные в структурированном виде, подходящим для дальнейшей обработки.
- На следующем этапе работы выполняется анализ разобранной информации. На данном этапе анализатор получает данные в структурированном виде и в зависимости от правил выполняет соответствующие действия. Таким образом, анализатор имеет информацию о типе записи, а также о параметрах записи. На основании этих данных определяется последовательность действий.

При выполнении узкоспециализированных задач, направленных на обработку редких сервисных событий, обнаружение требуемой записи, как правило является простой задачей. Однако в остальных случаях, требующих тщательного разбора записей сервисных журналов, задача разбора оказывается нетривиальной. Для решения простых задач, связанных с анализом сервисных журналов, администраторы могут использовать существующие программные решения, реализующий поиск и анализ информации. В более сложных случаях (например, нестандартные форматы журналов, необходимость выборки информации из нескольких связанных источников) отсутствие готовых решений вынуждает создавать новые утилиты программного разбора. Таким образом, ставится задача о создании более или менее универсального средства, позволяющего задавать простые правила разбора записей. Следует отметить, что у большинства решений механизм разбора основывается на использовании регулярных выражений, которые позволяют практически однозначно задавать шаблоны разбора. К тому же, многие системные администраторы и программисты знакомы с основами формирования регулярных выражений и не испытывают больших

затруднений в понимании формируемого шаблона. На данный момент существует большое число решений, предназначенных для разбора, анализа и обработки записей в сервисных журналах.

Проанализированные средства анализа сервисных журналов были разделены на 3 группы: узкоспециализированные решения, решения по управлению безопасностью и событиями, решения по управлению безопасностью и событиями, универсальные решения с настраиваемой логикой разбора. Приведем характеристику программных средств каждой из этих групп.

### *3.1. Узкоспециализированные решения*

К данной категории следует отнести программные решения имеющие мощные средства для разбора журналов конкретных служб. Данные средства позволяют строить статистику на основе полученных данных и выводить её в различном виде (PDF-файлы, HTML-страницы). К данной категории следует отнести: lire [7], request-log-analyzer [8], а также множество анализаторов журналов веб-серверов (AWStats [9], Open Web Analytics [10], и т.д.). Для понимания работы программных средств данной группы приведем более подробное описание утилиты lire. Данный программный продукт поддерживает наиболее востребованные типы сервисов: Web-сервера, FTP-сервера, почтовые сервера и другие сервисы. Для начала работы данной утилиты необходимо указать путь к сервисному журналу, а также его тип. В процессе работы программы формируется отчет, который содержит наиболее важные статистические данные. Виды генерируемых отчетов являются стандартными и слабо поддаются модификации.

Средства анализа из данной группы обладают готовой базой шаблонов, для обработки журналов поддерживаемых служб. При этом эти шаблоны достаточно четко определяют типы записей, в результате чего для каждого типа служб получается подробная статистика по ключевым полям. Однако данные решения являются неприменимыми в ряде случаев. Во-первых,

сгенерированная статистика не позволяет выполнять поиск по интересующим параметрам. Во-вторых, возможности программы ограничены поддерживаемыми службами, что не позволяет использовать её, например, для не столь распространенных служб или для анализа журналов, представленных в нестандартных форматах. В-третьих, пользователь не имеет возможности изменять существующие шаблоны или указывать действия по обработке различных типов записей. Очевидно, что данные решения не являются универсальными и не подходят для поставленной задачи.

### *3.2. Решения по управлению безопасностью и событиями (Security Information and event management (SIEM))*

К данной категории следует отнести средства, поддерживающие различные интерфейсы получения информации: журналы, серверы, сетевые пакеты, сетевые устройства, различные устройства контроля и сбора данных. К таким продуктам следует отнести: HP Arcsight Logger, IBM QRadar Logger, LogRhythm и т.д. [11]. Данные решения удовлетворяют пользовательским требованиям по анализу безопасности и информации о событиях в режиме реального времени. Они помогают собирать, хранить, анализировать и сообщать информацию в сервисных журналах для быстрого реагирования и прогнозирования. Данные решения обладают рядом преимуществ. Во-первых, приведенные продукты для представления статистики используют веб-интерфейс, что позволяет получать подробные отчеты из любого места, обеспеченного выходом в Интернет. Во-вторых, пользователь может самостоятельно настраивать отображение обработанной информации и регулировать систему согласно своим требованиям. В-третьих, данные продукты обладают расширенными инструментами поиска, что позволяет получать детальную информацию по интересующим параметрам. Тем не менее, указанные программные средства не предоставляют функциональный возможности для модификации логики анализа и обработки, поэтому не

являются универсальными средствами разбора сервисных журналов. Также следует отметить, что указанные решения являются коммерческими и бесплатные версии, как правило, имеют ряд ограничений.

### *3.3. Универсальные решения с настраиваемой логикой разбора*

Анализаторы из этой группы обладают средствами, которые позволяют пользователю задавать логику процесса распознавания и обработки информации путем использования специальных языков или веб-конструкторов. К данным решениям следует отнести: logstash [12], swatch [13], xlogmaster [14], splunk [15]. Далее рассматриваются детали работы этих решений.

Программный продукт logstash предоставляет мощные инструменты для фильтрации записей журналов. Для работы с информацией logstash использует различные средства фильтрации.

- Фильтр выделения записей по контрольным суммам. Данный фильтр позволяет создавать уникальные идентификаторы для записей и использовать их в дальнейшем для выявления аналогичных записей.
- Разбор по параметрам CSV-файла, JSON-файла, XML-файла. Данный фильтр позволяет разбирать структурированную информацию в наиболее популярных формата передачи данных.
- Распознавание времени и даты.
- Разрешения IP-адресов и доменных имен.
- Определение географического положения по IP-адресу.
- Распознавание строк с содержимым вида <key>=<value>.
- Работа с метрическими значениями. Данный фильтр позволяет сравнивать числовые значения, присутствующие в записи для более глубоко анализа сообщений.
- Запуск Ruby-приложения для обработки записи.
- Механизм фильтрации grok. Данный фильтр позволяет создавать шаблоны для нахождения записей журнала. Данный фильтр является более

удобным аналогом регулярных выражений и основан на разборе записи по типам данных.

- Механизм фильрации split. Позволяет разбивать строку на подстроки, используя в качестве разделителя введенную последовательность символов.
- Механизм фильтрации grep. Позволяет использовать регулярные выражения для фильтрации записей.

В качестве входного и выходного потоков данных logstash поддерживает различные структуры: файлы, письма серверов сообщений, TCP или UDP пакеты, а также другие службы, предоставляющие информацию по различным каналам. Данное решение часто используется в связке с такими программными средствами как elasticsearch, позволяющим агрегировать структурированную информацию и анализировать её [16], а также kibana, предоставляющий удобный веб-интерфейс для собранной информации [17]. Программное решение logstash позволяет определять команды для фильтрации записей журналов, а также выполнять некоторую обработку при помощи встроенных модулей. Использование logstash вместе с elasticsearch и kibana позволяет получить удобный интерфейс к отфильтрованным аналитическим и статистическим данным. Данное решение предназначено для фильтрации записей журнала по различным условиям отбора. Отфильтрованные записи могут быть использованы в дальнейшем для анализа и обработки. Описанный программный продукт может быть использован как универсальное средство обработки сервисных журналов. Среди его недостатков стоит отметить отсутствие поддержки логического разбора записей.

Программные решения swatch и xlogmaster используются для обработки сервисных журналов в режиме реального времени. Данные утилиты принимают на вход любой сервисный журнал. Далее, с использованием файла правил выполняется обработка журнала, в двух

режимах: полной обработки одного сервисного журнала с остановкой или постоянным наблюдением за сервисным журналом и немедленной обработкой появляющихся записей. Для указания логики разбора используются стандартные регулярные выражения. Для обработки подходящих под шаблон записей используется простой встроенный язык, позволяющий указывать действия, которые необходимо провести с информацией, в том числе: отправить e-mail с данной записью, вывести необходимую информацию в поток вывода, запустить отдельную программу, передав ей в качестве параметра разобранную запись. Данные решения позволяют решать стандартные задачи разбора и обработки сервисных журналов: распознавание записей и обработка записи. Однако, данные решения не предоставляют встроенных средств для выполнения логического разбора записей и не могут быть использованы для задач сбора статистической информации.

Программное решение splunk имеет широкие возможности и позволяет решать большую часть задач, связанных с анализом сервисных журналов. Во-первых, splunk обладает поддержкой большого числа сервисов и имеет средства для разбора соответствующих журналов. Во-вторых, пользователь имеет возможность модифицировать вид данной статистики, а также выполнять поиск по содержимому структурированного журнала. В-третьих, данное решение поддерживает возможность описывать логику разбора для журналов. Данное решение может применено для различных задач разбора, анализа и обработки информации, однако оно также лишено средств создания сложной логики разбора записей, а также является коммерческим и не предоставляет доступа к исходному коду для модификации его в случае изменения сформулированных выше требований к универсальному логическому анализатору.

## 4. Заключение

Как видно из обзора продуктов логического анализа журналов, на данный момент существует большое количество решений, имеющих разнообразный функционал, ориентированных на различные задачи и обладающие встроенными возможностями для модификации логики разбора и обработки. Проанализировав существующие решения можно сделать следующие выводы:

Для стандартных задач сбора аналитической информации на основе сервисных журналов распространенных программ подходят узкоспециализированные решения, однако они не имеют средств для самостоятельного разбора сервисных журналов, а также не предоставляют возможности задавать пользовательские действия, выполняемые при обнаружении записей определенного типа.

Для задач сбора аналитической информации и выполнения поиска среди разобранных записей сервисов подходят конфигурируемые SIEM решения. Однако они не подходят для выполнения сложного логического разбора, задач обработки или создания собственных указаний для разбора записи.

Универсальные, модифицируемые решения в наибольшей степени удовлетворяют поставленным требованиям, однако ни одно из рассмотренных решений не предоставляет встроенных возможностей задания выполнения логического разбора, в том числе и в нескольких журналах.

Наиболее перспективными инструментами для решения задач анализа и обработки сервисных журналов являются универсальные и конфигурируемые решения. Совокупные возможности данных программных средств удовлетворяют большинству поставленных требований.

## Литература

1. Сайт проекта tcpdump. URL: http://www.tcpdump.org/ (дата обращения 01.11.13)

2. И.М.Ажмухамедов, А.Н. Марьенков Поиск и оценка аномалий сетевого трафика на основе циклического анализа. [Электронный ресурс] // Инженерный вестник Дона, 2012, №2. – Режим доступа: http://www.ivdon.ru/magazine/archive/n2y2012/742 (доступ свободный) – Загл. с экрана. – Яз. Рус.

3. Сайт проекта SNORT. URL: http://www.snort.org/ (дата обращения 14.11.2013)

4. Сайт проекта REAL Secure. URL: http://www.ibm.com/ru/services/iss/realsecure_network.html (дата обращения 14.11.13)

5. Сайт проекта bro. URL: http://www.bro.org/ (дата обращения 14.11.13)

6. И.В. Георгица, С.А. Гончаров, В.А. Мохов Мультиагентное моделирование сетевой атаки типа DDoS. [Электронный ресурс] // «Инженерный вестник Дона», 2012, №2. – Режим доступа: http://www.ivdon.ru/magazine/archive/n3y2013/1852 (доступ свободный) – Загл. с экрана. – Яз. рус.

7. Страница проекта lire. URL: http://www.logreport.org/ (дата обращения 20.09.2013)

8. Страница проекта request-log-analyzer. URL: https://github.com/wvanbergen/request-log-analyzer/ (дата обращения 10.10.2013)

9. Страница проекта awstats. URL: http://awstats.sourceforge.net/ (дата обращения 08.11.2013)

10. Страница проекта Open Web Analyatics. URL: http://www.openwebanalytics.com/ (дата обращения 08.11.2013)

11. Mark Nicolett, Kelly M. Kavanagh Magic Quadrant for Security Information and Event Management [Текст] // Gartner RAS Core Research Note, 2013. – стр. 1-2, 7-9

12. Страница проекта logstash. URL: http://logstash.net/ (дата обращения 20.09.2013)

13. Страница проекта swatch. URL: http://sourceforge.net/projects/swatch/ (дата обращения 20.09.2013)

14. Страница проекта xlogmaster. URL: http://www.gnu.org/software/xlogmaster (дата обращения 10.11.2013)

15. Страница проекта Splunk. URL: http://docs.splunk.com/Documentation/Splunk/ (дата обращения 22.09.2013)

16. Страница проекта Elasticsearch. URL: http://www.elasticsearch.org/ (дата обращения 10.10.2013)

17. Страница проекта Kibana. URL: https://github.com/rashidkpc/Kibana/ (дата обращения 10.10.2013)