УДК 004.7 doi: 10.15827/0236-235X.148.495-503 2024. Т. 37. № 4. С. 495–503

Высокопроизводительный сервис сбора и анализа файлов журналов сетевого и серверного оборудования в национальной исследовательской компьютерной сети

А.Г. Абрамов [1, 2]✉

[1] Межведомственный суперкомпьютерный центр РАН, Санкт-Петербургское отделение, г. Санкт-Петербург, 199034, Россия
[2] Национальный исследовательский центр «Курчатовский институт», г. Москва, 123182, Россия

Ссылка для цитирования
Абрамов А.Г. Высокопроизводительный сервис сбора и анализа файлов журналов сетевого и серверного оборудования в национальной исследовательской компьютерной сети // Программные продукты и системы. 2024. Т. 37. № 4. С. 495–503. doi: 10.15827/0236-235X.148.495-503
Информация о статье

Группа специальностей ВАК: 2.3.5
Поступила в редакцию: 30.07.2024 После доработки: 10.09.2024 Принята к публикации: 16.09.2024

Аннотация. Построение бесперебойного и производительного решения для сбора, интеллектуальной обработки и анализа данных системных и сервисных журналов представляет собой содержательную и многоаспектную исследовательскую и прикладную задачу. Ее решение позволит обеспечить надежное функционирование научных телекоммуникационных сетей и сервисов на их основе. В настоящей статье приведен обзор разработанных и эксплуатируемых методик, технологий и инструментов работы с журналами с акцентом на программное обеспечение с открытым исходным кодом. Рассмотрены некоторые аспекты работы служб журналирования в операционных системах семейства Unix, основанных на протоколе syslog. Обозначены особенности построения и примеры типовых современных программных конвейеров обработки журналов и выполняемые ключевые функции, в том числе при задействовании методов и технологий машинного обучения. Приведено схематическое и детальное текстовое описание разработанного и внедренного в национальной исследовательской компьютерной сети России специализированного сервиса. Представленный сервис основан на наборе открытого программного обеспечения в составе широко применяемого в практике системного администрирования пакета сбора и обработки данных журналов Rsyslog, на высокопроизводительной колоночной СУБД ClickHouse и системе визуализации, статистики и аналитики данных Grafana.

Ключевые слова: сетевой мониторинг, системные и сервисные журналы, централизованный сбор и анализ журналов, национальная исследовательская компьютерная сеть, НИКС, свободно распространяемое программное обеспечение, Rsyslog, ClickHouse, Grafana

Благодарности. Публикация подготовлена в рамках госзадания СПбО МСЦ РАН и НИЦ «Курчатовский институт» по теме № FNEF-2024-0014 с использованием ресурсов Центра коллективного пользования научным оборудованием НИКС

Введение. В условиях массового распространения технологий и устройств Интернета вещей (*Internet of Things*, IoT), технологий виртуализации и облачных вычислений, искусственного интеллекта, киберфизических систем и иных интеллектуальных технологий концепции «Индустрия 4.0», мобильных сетей новых поколений информационно-коммуникационные системы стали включать в себя огромное количество распределенных компонентов, предназначенных для непрерывного предоставления потребителям разнообразных цифровых услуг и сервисов.

Постоянно повышается сложность, возрастает степень критичности и ответственности при мониторинге и управлении крупными цифровыми инфраструктурами. На решение этих задач нацелены наукоемкие методы, технологии и реализующие их коммерческие или базирующиеся на разработках с открытым исходным кодом программные инструменты [1, 2].

Развитые платформы могут реализовывать различные стратегии мониторинга и управления, которые способны органично дополнять друг друга и повышать общую эффективность и результативность процессов – методы на основе сетевых протоколов SNMP и ICMP, анализа системных и сервисных журналов, сетевой телеметрии (NetFlow, IPFIX и др.), а также на основе специальных подходов, предполагающих дублирование трафика, захват сетевых пакетов и выполнение глубокой инспекции трафика [3, 4].

В настоящей работе акцент сделан на методиках и программных решениях для централизованного сбора, обработки и анализа системных и сервисных журналов (логов), поступающих в платформы мониторинга от гетерогенного мультивендорного оборудования в составе наблюдаемых инфраструктур. Специалистам хорошо известно, что журналы обычно представляют собой текстовые файлы, которые со-

держат множество однострочных (в некоторых случаях – многострочных) записей, фиксирующих происходящие на оборудовании события аппаратного и программного уровней [5, 6]. Помимо файлов журналов, системы мониторинга могут агрегировать и представлять в своих пользовательских интерфейсах служебную информацию в виде SNMP-ловушек (трапов), MQTT-сообщений и т.п.

Крупные цифровые инфраструктуры способны генерировать десятки тысяч журнальных событий в минуту, причем далеко не все из них представляют ценность для целей мониторинга. Постоянно растущие объемы, слабая структурированность (или ее отсутствие), низкая плотность представления информации, широкий спектр форматов файлов журналов приводят к существенному повышению стоимости ресурсов долгосрочного хранения и к объективным трудностям при обработке и анализе данных.

Решению соответствующих задач посвящено достаточно большое число исследований и разработок, вместе с тем построение высокопроизводительного и надежного аппаратно-программного конвейера для автоматической работы с журналами в режиме реального времени по-прежнему является актуальной и важной для практики задачей.

Разработке методик, технологий и программных инструментов работы с большими данными, генерируемыми в научных телекоммуникационных сетях, включая информативные для целей аналитики данные о сетевых потоках и в большей степени служебное содержимое файлов журналов, уделяется существенное внимание при реализации проекта развития Национальной исследовательской компьютерной сети России (НИКС) [7, 8].

Обзор методов и инструментов сбора и анализа журналов

Основные методики сбора, обработки и анализа данных журналов. Следует отметить, что какой-либо единый стандарт для формирования полей, составляющих запись о событиях, отсутствует, хотя попытки унификации формата журналов периодически предпринимаются, в том числе в отношении конкретного оборудования и сервисов. Несогласованность форматов записей создает сложности на пути выполнения типичных операций мониторинга на основе журнальных данных, например, классификация некоторого события как критичного для инфраструктуры инцидента с оповещением ответственных служб, с генерацией проблемных билетов и при возможности с автоматическим принятием необходимых мер [9, 10].

Большинство доступных алгоритмов обработки и аналитики данных предполагают взаимодействие со структурированной информацией, кроме того, они неприменимы без определенных доработок и адаптаций к особенностям журналов. Наряду с большими объемами информации это обусловливает необходимость разработки специализированных алгоритмов, методик и программных пакетов, в том числе базирующихся на параллельной многопоточной обработке данных и машинном обучении.

К числу технических задач мониторинга, при решении которых могут оказаться полезными данные журналов, отнесем в частности

- выявление инцидентов на основе журнальных событий;
- обнаружение поведенческих аномалий, проблем с аппаратным и программным обеспечением и их производительностью;
- инициирование отправки оповещений о критических инцидентах и открытия проблемных билетов;
- анализ первопричин;
- диагностика и проактивное прогнозирование комплексных сбоев с целью предупреждения перерывов в предоставлении сервисов;
- анализ статистики использования сервисов [11–13].

Важным направлением является использование информации из журналов для обеспечения информационной безопасности, включая детектирование и предупреждение несанкционированных вторжений и действий пользователей, сетевых атак разных типов и целей и т.п. [14–16].

Поэтапный процесс обработки журнальных данных в рамках автоматизированного подхода обычно включает в себя централизованный сбор и хранение, предварительную обработку, анализ и представление (отчеты, визуализация) [17–19].

Сбор «сырых» журналов осуществляется на специально настроенных серверах-коллекторах, получающих данные от разнородных конечных устройств (сетевые маршрутизаторы и коммутаторы, серверы, системы IP-телефонии и видеонаблюдения и др.), на которых формируются события системного и прикладного уровней (операционные системы, среды виртуализации и контейнеризации, БД, сетевые сервисы и др.), а также из внешних систем сбора

журналов. Доставка данных в коллекторы может производиться в агентном и безагентном режимах с последующим сохранением в архивированном виде (в том числе для целей исполнения законодательства и технического учета), а также с обработкой и анализом в режиме реального времени.

Обработка преимущественно неструктурированной по своей природе первичной журнальной информации, ее автоматическое преобразование в структурированный унифицированный формат с компактным представлением и генерацией стандартных событий производятся в системах мониторинга с помощью сформированных правил и методов, которые выполняют операции по разбору (парсингу) журналов.

Основная идея парсинга заключается в классификации входных записей журнала на основе определенных процедур и синтаксиса известных событий, на их преобразовании в некий шаблон, что позволяет произвести семантическую интерпретацию содержимого и применить к данным доступные методы обнаружения. В условиях постоянного роста объемов журналов и применения большого числа разных форматов разработка парсеров вручную неэффективна и представляет собой задачу трудновыполнимую и требующую глубоких знаний в предметной области. В связи с этим передовые методы и алгоритмы генерации парсеров анализируют образцы данных журнала и автоматически создают шаблоны событий с выделением статических и динамических (переменных) частей. Наряду с рутинными регулярными выражениями и подготовленными вручную правилами при обработке и анализе журналов в последние годы стали широко применяться вычислительно эффективные алгоритмы, модели и библиотеки машинного обучения [20–22].

В процессе разбора и анализа журналов может производиться целый набор действий, таких как нормализация, фильтрация, сжатие, дедупликация, классификация, извлечение признаков и проч.

К полезным функциям работы с журналами также можно отнести автоматическую отправку оповещений о детектированных инцидентах и эскалации, шаблоны для быстрой настройки интеграций с внешними системами, генерацию отчетов на основании задаваемых критериев, импорт/экспорт данных журналов, архивирование данных для длительного хранения и возможности восстановления, управление глубиной хранения.

Доступные протоколы и инструменты для работы с журналами. Особенности структуры записей в журналах обусловливают использование протокола syslog как стандарта регистрации и отправки сообщений о происходящих событиях, используемого в Unix-подобных операционных системах и в большинстве сетевых устройств [23]. В соответствии со стандартом источники формируют текстовые сообщения о событиях и передают их на обработку локальному серверу syslog по протоколам UDP/TCP. Сообщения также могут передаваться на внешние syslog-серверы или в системы мониторинга, например, для целей централизованного сбора и обработки.

Типичная запись журнала может содержать отделяемые друг от друга символами-разделителями наборы полей в составе:

- временная метка, указывающая на момент наступления события;
- идентификатор и/или имя (IP-адрес) источника сообщения;
- идентификатор и/или имя сервиса/приложения (sshd, nginx, postfix, named, mysql и др.);
- уровень важности/критичности сообщения (alert, critical, error, warning, notice, info, debug и др.);
- идентификатор и/или имя категории источника записи (auth, cron, daemon, kern, mail, ntp, security, user и др.)
- идентификаторы и/или имена пользователей (при наличии);
- полезное содержание сообщения.

Место локального хранения журналов для разных типов устройств и платформ может отличаться, а в Unix-подобных системах для этих целей используется каталог файловой системы /var/log. Форматы хранения (текстовые или бинарные файлы, хранение в СУБД) и конкретные каталоги могут назначаться в каждом случае индивидуально в конфигурационных файлах службы журналирования. Специальная системная утилита logrotate ответственна за осуществление ротации и сжатия файлов журналов в соответствии с настроенными правилами.

Созданы десятки систем для работы с журнальной информацией, которые имеют разное качество и функциональные возможности, количество их растет, в первую очередь, в контексте решения задач информационной безопасности. Программные инструменты можно классифицировать по способу лицензирования, варианту доставки и по функциональным возможностям [24, 25]. Ключевыми критери-

ями качества систем являются аккуратность обработки журналов и производительность по количеству обрабатываемых событий в секунду.

Среди набора свободно распространяемых решений с открытым исходным кодом отметим пакеты Logstash, Fluentd, Graylog, Syslog-ng, NXlog. На машинном обучении базируются открытые методы и инструменты LogPAI, Logsy, LogAnomaly, DeepLog [26], DeepSyslog [27], DLLog [28].

Примерами коммерческих систем, традиционно отличающихся широким функционалом и развитыми пользовательскими интерфейсами, являются SolarWinds Loggly, ManageEngine EventLog Analyzer, Splunk Log Observer, Nagios Log Server; DataDog Log Management, Logic Monitor, Logz.io, Sematext Logs, Sumo Logic, Fluent Bit, FileBit.

Следует отметить открытые наборы датасетов с информацией из журналов, собранной из разных систем и программных продуктов, такие как Loghub, CFDR, OpenStack Fault Injection Dataset, SecRepo. Они могут служить полезным источником для машинного обучения моделей и их внедрения в практику мониторинга.

Описание разработанного сервиса и его внедрение в НИКС

Общие сведения о разработанном решении. До недавнего времени в сети НИКС файлы журналов собирались на оборудовании и хранились разрозненно, форматы далеко не всегда были согласованы, что создавало серьезные трудности для администраторов сервисов и сетевых инженеров. В целях преодоления этой проблемы был проработан проект создания специализированного сервиса для централизованного сбора и анализа файлов журналов и реализован освещаемый в настоящей статье первый этап проекта.

На этом этапе разработаны методики и первая версия инструментария на основе обоснованно выбранного и протестированного стека ПО с открытым исходным кодом. В качестве целей создания сервиса были определены:

- реализация возможностей расширенного многоаспектного статистического и аналитического учета функционирования компонентов инфраструктуры НИКС и предоставляемых на их основе сервисов;
- централизованная агрегация и представление соответствующей оперативной и исторической информации с учетом категорий пользователей;
- предоставление возможностей качественной визуализации информации;
- повышение эффективности и результативности функционирования службы технической поддержки.

В сервисе реализуются следующие основные функции:

– авторизация и управление доступом пользователей к функциональным возможностям в соответствии с ролевым профилем;

– получение от сетевого и серверного оборудования информации о функционировании и использовании компонентов инфраструктуры и сервисов НИКС и ее централизованная статистическая обработка;

– отображение информации в различных видах с широкими возможностями настройки, формирования наглядных информационных панелей;

– интеграционное взаимодействие и обмен данными с внешними системами;

Сервис предоставил возможности для содержательного анализа статистики использования инфраструктуры и сервисов НИКС на базе накапливаемой информации журналов, оперативного мониторинга функционирования сетевого оборудования и совершаемых операций, а также для комплексной оценки востребованности отдельных организаций-пользователей и уровня вовлеченности в использование сервисных решений.

Аппаратно-программный комплекс сервиса развернут на нескольких выделенных виртуальных вычислительных серверах в собственной высокопроизводительной инфраструктуре МСЦ РАН (г. Москва). Инфраструктура обеспечения облачной платформы реализована на основе популярного решения виртуализации Proxmox VE, в качестве системного ПО используется операционная система семейства GNU/Linux.

Прикладное ПО сервисов (рис. 1) базируется на взаимодействующих свободно распространяемых решениях с открытым исходным кодом – системе сбора и обработки данных журналов rsyslog (https://www.rsyslog.com) и платформе визуализации, статистики и аналитики данных Grafana (https://grafana.com). Кроме того, задействуются аппаратно-программные ресурсы высокопроизводительной колоночной СУБД ClickHouse (https://clickhouse.com), которая используется в работе других высоконагруженных сервисов НИКС, в частности, сервиса статистики и аналитики данных о сетевых потоках [29, 30].

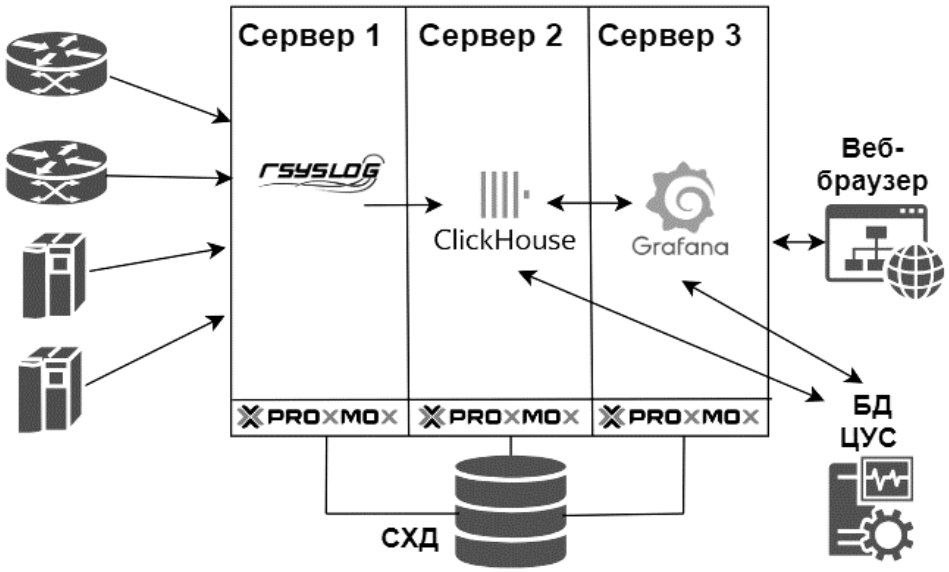

Рис. 1. Архитектурная схема решения с указанием внешних систем и сервисов

Fig. 1. Architectural scheme of the solution indicating external systems and services

Журналы пересылаются с сетевого и серверного оборудования по протоколу UDP на два виртуальных сервера с развернутым и настроенным решением rsyslog. Этот пакет имеет модульную архитектуру, расширение функционала и интеграция с внешними системами осуществляются путем установки дополнительных модулей. В данном случае на серверах используются модули препроцессинга rsyslog-omclickhouse (передача журналов в СУБД ClickHouse) и rsyslog-mmnormalize (парсинг логов для целей статистики). Типичная интенсивность поступления журналов составляет несколько сотен записей в секунду.

На сервере с платформой Grafana установлен и настроен специальный плагин источника данных, обеспечивающий поддержку СУБД ClickHouse. Сервисы обеспечивают управление содержимым с помощью встроенных элементов веб-интерфейса Grafana и позволяют выполнять в нем SQL-запросы к СУБД ClickHouse для отбора требующих отображения данных и уточнения сформированных ранее и сохраненных запросов.

К преимуществам развернутого решения и его компонентов, помимо общей высокой производительности конвейера обработки, можно отнести доступный для использования совместно с rsyslog хорошо подходящий для обработки событий и настройки процесса обработки скриптовый язык RainerScript (https://www.rsyslog.com/doc/rainerscript/). Кроме того, достоинствами являются инвертированные индексы СУБД ClickHouse, позволяющие эффективно производить полнотекстовый поиск по журналам, настраиваемые визуализации в пакете Grafana с разными типами графиков и диаграмм, а также потенциальная возможность формирования и отправки оповещений о выявленных в процессе анализа журналов и классифицированных инцидентах в наблюдаемой инфраструктуре.

Визуализация журналов на информационных табло

Пример настроенного информационного табло (дашборда) с представлением журнальной информации, статистических данных и с элементами управления приведен на рисунке 2.

Два однотипных табло содержат сведения о событиях, сгруппированных по их источникам как серверные и сетевые. На табло размещено табличное представление записей с их сортировкой по временной метке. Каждую запись можно раскрыть для просмотра статических и динамических полей. В результате предварительной обработки и парсинга среди статических полей выделяются

- host (имя источника записи);
- facility (имя категории источника записи);
- priority (уровень важности, приоритет);
- service (имя сервиса/приложения, записанное в файл журнала).

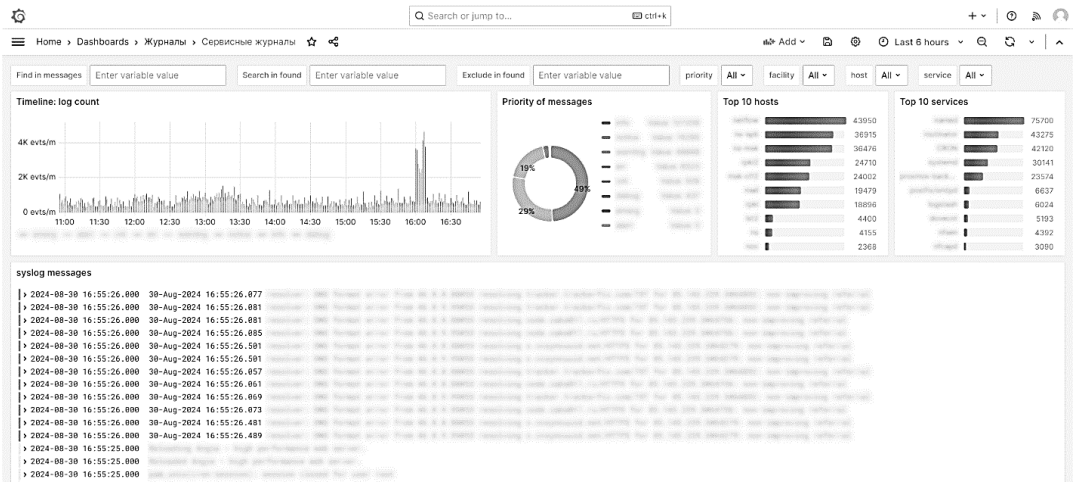

Рис. 2. Общий вид информационной панели с отображением и визуализацией информации из журналов

Fig. 2. Dashboard with log data display

Помимо таблицы, на табло доступны для просмотра и аналитики статистические данные в форме временного ряда с агрегацией событий по уровню важности за назначенный период времени, кольцевая диаграмма с соответствующими интегрированными данными, а также топ-списки с обобщением в разрезе источников (хостов) и сервисов.

Элементы управления позволяют выполнять полнотекстовый поиск информации с возможностью уточнений вида включить/исключить из результатов, а также независимый множественный выбор из выпадающих списков по перечисленным выше статическим полям. Доступны возможности назначения диапазона времени для отображения информации, уточнения настроенных SQL-запросов, экспорта данных в формате CSV и ряд других, стандартно предоставляемых в интерфейсе Grafana.

Дополнительно сформировано специальное табло для работы с событиями информационной безопасности, в которое выводятся отобранные из общих журналов записи, рассматриваемые в качестве потенциальных кандидатов на инциденты безопасности, а также реализована интеграция с системой класса SIEM (*Security Information and Event Management*) в части экспорта журнальных данных.

В отношении некоторых сервисов НИКС организован альтернативный сбор и обработка данных журналов и созданы отдельные информационные панели, отражающие специфику их функционирования и использования [31], в том числе обслуживающие разные сервисы веб-серверы, службы доменных имен (DNS) и точного времени (NTP), сервис тестирования пропускной способности сети iPerf, сервис роуминга в Wi-Fi-сетях для научно-образовательного сообщества Eduroam, сервис вебинаров BigBlueButton.

Заключение

Результатом первого этапа выполнения представленного научно-технического проекта стали разработка и внедрение в НИКС высокопроизводительного сервиса, успешно решающего задачи полного цикла работы с журнальной информацией, которая накапливается на эксплуатируемом сетевом и серверном оборудовании. Сервис, построенный на компонентах свободно распространяемого ПО с открытым исходным кодом, реализует централизованный сбор, многопараметрическую обработку перенаправляемых от устройств журналов, а также элементы аналитики, визуализации и отчетности.

Реализация проекта предоставила сетевым инженерам и администраторам сервисов современные возможности по работе с журналами в едином интерфейсе с удобным для мониторинга и контроля отображением информации о функционировании и возникающих проблемах с оборудованием, системным и прикладным ПО. Доступная статистика обращений к сервисам является ценным источником сведений об их использовании, дает возможность обоснованно принимать решения о развитии сервисного направления НИКС и его отдельных компонентов.

На следующих этапах проекта предполагается интегрировать сервис с сайтом центра управления сетью НИКС (*Network Operations Center*, NOC) для отображения выявленных инцидентов на панели службы технической поддержки, а также реализовать автоматические оповещения об инцидентах средствами электронной почты, иных доступных механизмов и расширить возможности аналитики и визуализации.

В качестве одного из перспективных направлений развития рассматривается привлечение для автоматического анализа журналов методов машинного обучения. Накапливаемые наборы данных позволяют адаптировать и использовать их для обучения моделей машинного обучения с помощью разных подходов и методик в целях интеллектуализации анализа и проактивного предсказания аномалий в сетевой инфраструктуре.

Список литературы

1. Julian M. Practical Monitoring: Effective Strategies for the Real World. California, Sebastopol, O'Reilly Media Publ., 2017, 229 p.

2. Blokdyk G. Network Performance Monitoring and Diagnostics Tools. Toronto, 5STARCooks Publ., 2022, 324 p.

3. D'Alconzo A., Drago I., Morichetta A., Melli M., Casas P. A survey on big data for network traffic monitoring and analysis. IEEE TNSM, 2019, vol. 16, no. 3, pp. 800–813. doi: 10.1109/TNSM.2019.2933358.

4. Wilkins P. Logging in Action: with Fluentd, Kubernetes and more. Manning Publ., 2022, 392 p.

5. Абрамов А.Г., Гончар А.А., Евсеев А.В., Шабанов Б.М. Национальная исследовательская компьютерная сеть нового поколения: текущее состояние и концепция развития // Информационные технологии. 2021. Т. 27. № 3. С. 115–124. doi: 10.17587/it.27.115-124.

6. Абрамов А.Г., Гончар А.А., Евсеев А.В., Шабанов Б.М. Основные результаты первых этапов проекта развития национальной исследовательской компьютерной сети // Информационные технологии и вычислительные системы. 2024. № 1. С. 3–10. doi: 10.14357/20718632240101.

7. Kubacki M., Sosnowski J. Holistic processing and exploring event logs. In: LNPSE. Proc. SERENE, 2017, vol. 10479, pp. 184–200. doi: 10.1007/978-3-319-65948-0_12.

8. He S., He P., Chen Z. et al. A survey on automated log analysis for reliability engineering. ACM CSUR, 2021, vol. 54, no. 6, art. 130. doi: 10.1145/3460345.

9. Lin H., Yan Z., Chen Y., Zhang L. A survey on network security-related data collection technologies. IEEE Access, 2018, vol. 6, pp. 18345–18365. doi: 10.1109/ACCESS.2018.2817921.

10. Драчев Г.А. Разработка алгоритма выделения и кодирования данных из журнальных сообщений вычислительной системы для систем обнаружения аномалий // Информационные технологии. 2023. Т. 29. № 7. С. 351–359. doi: 10.17587/it.29.351-359.

11. Zhang T., Qiu H., Castellano G. et al. System log parsing: a survey. IEEE Transactions on Knowledge & Data Engineering, 2023, vol. 35, no. 8, pp. 8596–8614. doi: 10.1109/TKDE.2022.3222417.

12. Ma J., Liu Y., Wan H., Sun G. Automatic parsing and utilization of system log features in log analysis: a survey. Appl. Sci., 2023, vol. 13, no. 8, art. 4930. doi: 10.3390/app13084930.

13. Skopik F., Landauer M., Wurzenberger M. Online log data analysis with efficient machine learning: a review. IEEE Security & Privacy, 2022, vol. 20, no. 3, pp. 80–90. doi: 10.1109/MSEC.2021.3113275.

14. Худяков Д.А. Разработка системы выявления аномалий на основе распределенной трассировки логов // Вестн. НГУ. Сер.: Информационные технологии. 2023. Т. 21. № 1. С. 62–72. doi: 10.25205/1818-7900-2023-21-1-62-72.

15. Chen B., Jiang Z.M. A survey of software log instrumentation. ACM CSUR, 2021, vol. 54, no. 4, art. 90. doi: 10.1145/3448976.

16. Du M., Li F., Zheng G., Srikumar V. DeepLog: anomaly detection and diagnosis from system logs through deep learning. Proc. ACM SIGSAC Conf. CCS, 2017, pp. 1285–1298. doi: 10.1145/3133956.3134015.

17. Zhou J., Qian Y., Zou Q., Liu P., Xiang J. DeepSyslog: deep anomaly detection on syslog using sentence embedding and metadata. IEEE Transactions on Information Forensics and Security, 2022, vol. 17, pp. 3051–3061. doi: 10.1109/TIFS.2022.3201379.

18. Cheng H., Ying Sh., Duan X., Yuan W. DLLog: an online log parsing approach for large-scale system. Int. J. of Intelligent Systems, 2024, vol. 2024, art. 5961993. doi: 10.1155/2024/5961993.

19. Abramov A.G. Collection, analysis and interactive visualization of NetFlow data: Experience with big data on the base of the National Research Computer Network of Russia. Lobachevskii J. Math., 2020, vol. 41, pp. 2525–2534. doi: 10.1134/S1995080220120021.

20. Abramov A.G., Porkhachev V.A., Yastrebov Yu.V. Methods and high-performance tools for collecting, analysis and visualization of data exchange with a focus on research and education networks. Lobachevskii J. Math., 2023, vol. 44, pp. 4930–4938. doi: 10.1134/S1995080223110021.

21. Abramov A.G. Service portfolios of leading National Research and Education Networks and implementation on the basis of the National Research Computer Network of Russia. Lobachevskii J. Math., 2021, vol. 42, pp. 2481–2492. doi: 10.1134/S1995080221110032.

Software & Systems doi: 10.15827/0236-235X.148.495-503 2024, 37(4), pp. 495–503

High-performance service for collecting and analyzing network and server hardware log files on a National Research Computer Network

Aleksey G. Abramov [1, 2]✉

[1] Joint Supercomputer Center of RAS – St. Petersburg Branch, St. Petersburg, 199034, Russian Federation
[2] National Research Centre "Kurchatov Institute", Moscow, 123182, Russian Federation

For citation
Abramov, A.G. (2024) 'High-performance service for collecting and analyzing network and server hardware log files on a National Research Computer Network', *Software & Systems*, 37(4), pp. 495–503 (in Russ.). doi: 10.15827/0236-235X.148.495-503

Article info
Received: 30.07.2024 After revision: 10.09.2024 Accepted: 16.09.2024

Abstract. Building a seamless and productive solution for collecting, intelligent processing and analyzing system and service log data is a meaningful and multidimensional research and application problem. This solution will ensure reliable functioning of scientific telecommunication networks and services based on them. This paper provides an overview of developed and operated journaling techniques, technologies and tools with an emphasis on open source software. It also considers some aspects of logging services in Unix operating systems based on the Rsyslog protocol. There are also outlined construction features and examples of typical modern software log processing pipelines and their key functions, including those using machine learning methods and technologies. The author gives a schematic and detailed textual description of a special-purpose service developed and implemented in the national research computer network of Russia. This service is based on a set of open source software comprising the widely used in system administration practice of Rsyslog log data collection and processing package, high-performance columnar ClickHouse DBMS and Grafana data visualization, statistics and analytics system.

Keywords: network monitoring, system and service log data, centralized log collection and analysis, National Research Computer Network, NIKS, free software, Rsyslog, ClickHouse, Grafana

Acknowledgements. The publication has been prepared within the framework of the state assignment of the Joint Supercomputer Center of RAS and the National Research Center "Kurchatov Institute" on topic no. FNEF-2024-0014 using the resources of the Center for collective use of scientific equipment "National Research Computer Network"

References

1. Julian, M. (2017) *Practical Monitoring: Effective Strategies for the Real World*. California, Sebastopol: O'Reilly Media Publ., 229 p.
2. Blokdyk, G. (2022) *Network Performance Monitoring and Diagnostics Tools*. Toronto: 5STARCooks Publ., 324 p.
3. D'Alconzo, A., Drago, I., Morichetta, A., Melli, M., Casas, P. (2019) 'A survey on big data for network traffic monitoring and analysis', *IEEE TNSM*, 16(3), pp. 800–813. doi: 10.1109/TNSM.2019.2933358.
4. Wilkins, P. (2022) *Logging in Action: with Fluentd, Kubernetes and more*. Manning Publ., 392 p.
5. Abramov, A.G., Gonchar, A.A., Evseev, A.V., Shabanov, B.M. (2021) 'The new generation national research computer network: Current status and concept for the development', *Inform. Tech.*, 27(3), pp. 115–124 (in Russ.). doi: 10.17587/it.27.115-124.
6. Abramov, A.G., Gonchar, A.A., Evseev, A.V., Shabanov, B.M. (2024) 'Main results of the first stages of the project for the development of national research computer network', *J. of Inform. Tech. and Computing Systems*, (1), pp. 3–10 (in Russ.). doi: 10.14357/20718632240101.
7. Kubacki, M., Sosnowski, J. (2017) 'Holistic processing and exploring event logs', in *LNPSE. Proc. SERENE*, 10479, pp. 184–200. doi: 10.1007/978-3-319-65948-0_12.
8. He, S., He, P., Chen, Z. et al. (2021) 'A survey on automated log analysis for reliability engineering', *ACM CSUR*, 54(6), art. 130. doi: 10.1145/3460345.
9. Lin, H., Yan, Z., Chen, Y., Zhang, L. (2018) 'A survey on network security-related data collection technologies', *IEEE Access*, 6, pp. 18345–18365. doi: 10.1109/ACCESS.2018.2817921.
10. Drachev, G.A. (2023) 'Development of an algorithm for extracting and encoding data from log messages of a computing system for anomaly detection systems', *Inform. Tech.*, 29(7), pp. 351–359 (in Russ.). doi: 10.17587/it.29.351-359.
11. Zhang, T., Qiu, H., Castellano, G. et al. (2023) 'System log parsing: a survey', *IEEE Transactions on Knowledge & Data Engineering*, 35(8), pp. 8596–8614. doi: 10.1109/TKDE.2022.3222417.
12. Ma, J., Liu, Y., Wan, H., Sun, G. (2023) 'Automatic parsing and utilization of system log features in log analysis: a survey', *Appl. Sci.*, 13(8), art. 4930. doi: 10.3390/app13084930.
13. Skopik, F., Landauer, M., Wurzenberger, M. (2022) 'Online log data analysis with efficient machine learning: a review', *IEEE Security & Privacy*, 20(3), pp. 80–90. doi: 10.1109/MSEC.2021.3113275.
14. Khudyakov, D.A. (2023) 'Development of anomaly detection system based on distributed log tracing', *Vestn. NSU. Ser.: Inform. Tech.*, 21(1), pp. 62–72 (in Russ.). doi: 10.25205/1818-7900-2023-21-1-62-72.
15. Chen, B., Jiang, Z.M. (2021) 'A survey of software log instrumentation', *ACM CSUR*, 54(4), art. 90. doi: 10.1145/3448976.

16. Du, M., Li, F., Zheng, G., Srikumar, V. (2017) 'DeepLog: anomaly detection and diagnosis from system logs through deep learning', *Proc. ACM SIGSAC Conf. CCS*, pp. 1285–1298. doi: 10.1145/3133956.3134015.

17. Zhou, J., Qian, Y., Zou, Q., Liu, P., Xiang, J. (2022) 'DeepSyslog: deep anomaly detection on syslog using sentence embedding and metadata', *IEEE Transactions on Information Forensics and Security*, 17, pp. 3051–3061. doi: 10.1109/TIFS.2022.3201379.

18. Cheng, H., Ying, Sh., Duan, X., Yuan, W. (2024) 'DLLog: an online log parsing approach for large-scale system', *Int. J. of Intelligent Systems*, 2024, art. 5961993. doi: 10.1155/2024/59619933.

19. Abramov, A.G. (2020) 'Collection, analysis and interactive visualization of NetFlow data: Experience with big data on the base of the National Research Computer Network of Russia', *Lobachevskii J. Math*., 41, pp. 2525–2534. doi: 10.1134/S1995080220120021.

20. Abramov, A.G., Porkhachev, V.A., Yastrebov, Yu.V. (2023) 'Methods and high-performance tools for collecting, analysis and visualization of data exchange with a focus on research and education networks', *Lobachevskii J. Math*., 44, pp. 4930–4938. doi: 10.1134/S1995080223110021.

21. Abramov, A.G. (2021) 'Service portfolios of leading National Research and Education Networks and implementation on the basis of the National Research Computer Network of Russia', *Lobachevskii J. Math.*, 42, pp. 2481–2492. doi: 10.1134/S1995080221110032.

Авторы
Абрамов Алексей Геннадьевич [1, 2],
к.ф.-м.н., доцент,
ведущий научный сотрудник,
abramov@niks.su

[1] Межведомственный суперкомпьютерный центр РАН, Санкт-Петербургское отделение, г. Санкт-Петербург, 199034, Россия
[2] Национальный исследовательский центр «Курчатовский институт», г. Москва, 123182, Россия

Authors
Aleksey G. Abramov [1, 2],
Cand. of Sci. (Physics and Mathematics),
Associate Professor, Leading Researcher,
abramov@niks.su

[1] Joint Supercomputer Center of RAS – St. Petersburg Branch, St. Petersburg, 199034, Russian Federation
[2] National Research Centre "Kurchatov Institute", Moscow, 123182, Russian Federation