УДК 004.056
DOI: 10.17586/0021-3454-2018-61-11-1005-1011

КОМБИНИРОВАНИЕ ТЕХНОЛОГИЙ HADOOP И SNORT ДЛЯ ОБНАРУЖЕНИЯ СЕТЕВЫХ АТАК

Н. А. Комашинский

Санкт-Петербургский институт информатики и автоматизации РАН,
199178, Санкт-Петербург, Россия
E-mail: nckkm@ya.ru

Исследуется метод обработки информации с целью обнаружения компьютерных атак на основе технологий больших данных. Обоснована потребность создания специализированных методов проектирования, которые позволят повысить оперативность обработки получаемой информации. Рассматриваются возможности и оценки результативности параллельной обработки данных с целью обнаружения компьютерных воздействий на основе функционального подхода, а также ключевые принципы работы с большими данными. Приведена математическая модель, с помощью которой разработана методика обнаружения вторжений. Описывается принцип реализации задач обработки информации и выявления аномалий на основе интеграции платформ Hadoop, Snort. Изложены основные результаты экспериментальной оценки показателей применяемого метода для обнаружения компьютерных атак.

Ключевые слова: *большие данные, Hadoop, информационная система, информационная безопасность, компьютерная атака, Snort, аномалия, обработка данных*

Введение. В настоящее время одним из феноменов, оказывающих существенное влияние на область технологий обработки данных и выявление компьютерных воздействий, являются большие данные. С появлением вычислительных кластеров параллельные вычисления стали более доступны для массового применения. Для построения кластерных решений, как правило, используются персональные компьютеры, стандартные сетевые технологии, свободно распространяемые библиотеки и протоколы [1]. Следовательно, для решения вычислительно сложных задач в области обнаружения компьютерных атак можно использовать кластерную систему.

Постановка задачи, рассматриваемой в настоящей статье, заключается в следующем. Входными (исходными) данными является поток данных о событиях безопасности, которые относятся к различным типам. На основе анализа данных, характеризующих отклонение от набора элементарных действий, которые свидетельствуют о наличии признаков компьютерной атаки, необходимо синтезировать метод обработки входного потока данных, позволяющий выявлять сетевые атаки и поддерживающий распараллеливание процессов в специализированной среде.

Релевантные работы. Методы обработки данных, направленные на выявление компьютерных атак, можно условно разделить на сигнатурные и эвристические [2]. В этих методах используются различные подходы, основанные на анализе схожести шаблонов, статистическом анализе, интеллектуальном анализе данных и др. [3, 4].

В настоящее время наиболее распространенной является система, работающая на основе технологии Snort [5]. Пакеты, перехваченные Snort, анализируются с помощью системы Hadoop. Для более удобного анализа используется Apache Hive — система управления базами данных на основе платформы Hadoop. Эта система позволяет выполнять запросы, агрегировать и анализировать данные, хранящиеся в Hadoop. Для того чтобы охватить огромное количество

сетевых сообщений, в работе [6] предложена распределенная система обнаружения вторжений (Intrusion Detection System — IDS) с использованием Hadoop, HDFS и нескольких рабочих узлов; в целях повышения эффективности процесса предупреждения атак исследователи усовершенствовали алгоритмы путем внедрения распределенной обработки данных.

Достаточно интересным современным направлением развития методов анализа событий является применение подходов, базирующихся на машинном обучении и интеллектуальном анализе данных, таких как байесовские сети [7—9], иммунные сети [8, 9], искусственные нейронные сети [7, 8—10] и др. Особенность этих подходов заключается в возможности самостоятельной (безусловной) обработки событий с минимизацией ручной настройки. Однако для построения моделей обучения требуется предварительный анализ самих данных, который далеко не всегда можно автоматизировать. Работа [11] посвящена повышению точности и производительности алгоритма „случайного леса“ при его параллельном выполнении на платформе Spark — здесь предлагается подход к уменьшению размерности высокоразмерных данных, в которых экземпляры (или записи) содержат достаточно большое количество признаков. В работе [12] представлена архитектура системы, направленной на обнаружение и предотвращение вторжений в локальных сетях компаний. Анализируются несколько источников: DNS-трафик, HTTP-трафик, записи NetFlow-протокола, трафик с honeypot. Система интегрирует разные хранилища данных в одну систему хранения и обработки данных. Данные используются распределенной системой корреляции для организации широкомасштабной системы мониторинга безопасности.

Как показывает анализ релевантных работ, существует множество методов обработки событий с целью обнаружения компьютерных атак. Разработанные методы обладают различными достоинствами, но имеют и недостатки. Многие подходы основываются на сценариях атак для формирования последовательностей атакующих действий, тем самым относя их к классу сигнатурных методов, и характеризуются значительными временными затратами для настройки и адаптации к целевой инфраструктуре. Важным вектором развития методов обработки информации с целью выявления компьютерных атак является их адаптация к технологиям обработки больших данных и параллельных вычислений.

Модель обнаружения вторжений. Возможность и результативность параллельной обработки данных с целью обнаружения компьютерных воздействий оценивается на основе функционального подхода, при котором события информационной безопасности предполагается рассматривать с точки зрения выполнения элементарных функций, представляющих собой алгоритмы преобразования агрегированного пространства состояний самой системы [13, 14]. Разработка функционального представления информационной системы осуществляется на основе анализа пространства параметров процессов в системе по установленным правилам и выявлении параметров, характеризующих действие компьютерной атаки. Процесс агрегирования, являющийся ключевым понятием функционального подхода, заключается в построении агрегированного пространства состояний информационной системы — $\mathbb{R}e$, которое отличается от настоящего рядом упрощений (укрупнений), но при определенных допущениях может рассматриваться как реальное.

Элементарной функцией $f_i \in F$ будем называть математическое описание соответствующего ей элементарного действия или композицию элементарных действий минимальной длительностью в виде алгоритма преобразования, определенного на всем пространстве агрегированных состояний информационной системы. Областью значений элементарной функции будем называть полное подмножество состояний информационной системы, каждое из которых для данного преобразования есть прообраз. При этом если $S'_f \subset S$ — область определения функции f в пространстве $\mathbb{R}e$, а $S''_f \subset S$ — область значений функции f в пространстве $\mathbb{R}e$, то преобразование в этом пространстве согласно f может быть описано отображением

$$G_f : S'_f \to S''_f .$$

Множества элементарных функций, объединенные в правильные композиции, представляют собой цепочки, отражающие поведение информационной системы во введенном пространстве состояний. При этом изменение характера длины $|l| : l \in F^*$ представляет собой кортеж:

$$l = f_{l_1} \circ f_{l_2} \circ \ldots \circ f_{l_{|l|}} \in F^* ,$$

а l можно обозначить отображением

$$G_l = F_{\text{тр}} \to F ;$$

$$\left(\forall f^n\right)\left(\exists! f_i\right) P\left(G_l\left(f^n\right) = f_i\right).$$

Согласно последнему предикату (P) на каждом шаге функционирования информационной системы существует единственное верное элементарное действие f_i. Несоответствие набору элементарных действий, в свою очередь, свидетельствует о наличии признаков компьютерной атаки [15].

На основе представленной математической модели разработана модель обнаружения вторжений с использованием технологии больших данных. Для реализации используется интеграция технологий Snort и Hadoop. Предполагается, что отправка большого количества подозрительных пакетов источником может рассматриваться как атака [16]. Такая аномалия должна быть идентифицирована, и пакеты из соответствующего источника должны быть заблокированы. Также для блокировки входящих пакетов из определенных источников предполагается генерировать правила для Snort, чтобы иметь возможность предупреждать другие узлы об атаке. Таким образом, система Snort будет основываться как на сигнатурном методе, так и на методе обнаружения аномалий.

Входящие пакеты собираются с помощью запуска Snort в режиме регистрации пакетов. Каждый пакет содержит признаки, такие как временная метка (*t*), протокол, IP-адрес источника (*src*), IP-адрес узла назначения (*dst*), номера портов (*n*), тип пакета (*type*). Когда на локальном диске будет записано большое количество пакетов, все файлы будут либо в формате *tcpdump*, либо в двоичном формате. Поэтому необходимо выполнить некоторую предварительную обработку для преобразования файлов в читаемый формат, что можно реализовать посредством подходящих команд Snort. Для анализа в Hadoop эти файлы должны быть загружены в файловую систему HDFS. В этом случае оператор MapReduce, преобразовывая пакеты и извлекая из них параметры (*t*), (*src*), (*dst*), (*n*), (*type*), определяет количество пакетов *N*, поступивших из конкретного источника и дошедших до адресата определенного типа.

После идентификации IP-адреса источника и IP-адреса узла назначения могут быть сгенерированы новые правила Snort, если на конкретный узел поступает большое количество пакетов. Поэтому когда наступает такой момент, правила Snort уже согласованы, и будут выполняться действия, описанные в новых сгенерированных правилах. Процесс обработки входящих пакетов с помощью HDFS и HIVE представлен на рис. 1. Вначале входящие пакеты хранятся в виде логов. Далее они загружаются в HDFS. На Map-шаге осуществляется предварительная параллельная обработка входной информации, что существенно ускоряет в режиме реального времени процесс реагирования на подозрительную активность при работе с разнородными данными. Для этого один из компьютеров (называемый главным узлом — master node) получает входные данные задачи, разделяет их на части и передает другим компьютерам (рабочим узлам — worker node) для предварительной обработки. На Reduce-шаге происходит свертка предварительно обработанных данных. Главный узел получает ответы от рабочих узлов и на их основе формирует результат, который передается в Apache Hive и HDFS, где и формируются новые правила.

Для обнаружения аномалий технология Snort должна быть дополнена правилами, чтобы при будущей аналогичной атаке можно было предпринять соответствующие меры. Обновление и добавление новых правил должно осуществляться на основе анализа поступающих событий. Таким образом, добавляя соответствующие правила, система Snort будет способна находить новые атаки и сигнализировать пользователю о потенциальной вредоносной активности.

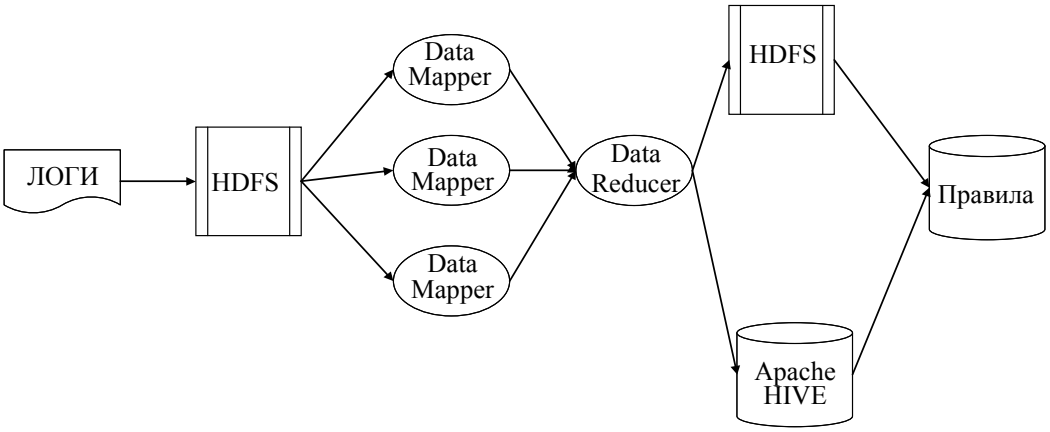

Рис. 1

Результаты экспериментов. Для тестирования системы дополнительно использовались следующие инструменты: Hadoop cdh4.2.0, Snort 2.9.8.3, Hive 1.1.0. Кластер Hadoop функционирует в распределенном режиме. Для генерации пакетов использовался Hping3. Пакеты отправлялись таким образом, чтобы имитировать DoS-атаку. Пакеты регистрировались путем запуска Snort в режиме регистратора пакетов в течение 5 мин. Средний размер каждого из лог-файлов, которые затем преобразовывались в подходящий формат для чтения, составлял 550 Мб. Пакеты имеют разный размер и разный формат; из них были извлечены следующие данные: исходный IP-адрес, IP-адрес назначения, протокол передачи и количество извлеченных пакетов.

Работа MapReduce выполнялась на кластере Hadoop с различными размерами файлов: 454,7 и 888,1 Мб, 1,8 и 2,9 Гб, каждый из которых содержал в среднем от 50 тыс. до двух lakh-пакетов. В табл. 1 показаны время процесса и количество Map-задач.

Таблица 1

Размер входного файла	Количество Map-задач	Время процесса, с
454,7 Мб	7	50
888,1 Мб	14	100
1,8 Гб	22	140
2,9 Гб	38	275

Количество reduce-задач во всех случаях одинаково. На одном узле время выполнения задачи составляет в среднем 15 мин. По мере увеличения количества узлов время обработки значительно сократится, и, следовательно, эффективность системы повысится.

В ходе экспериментов использовались блоки разных размеров: 32, 64 и 128 Мб, размер пакетного файла 420 Мб. В табл. 2 показана производительность системы, при которой количество Map-задач и время процесса соответствуют определенному размеру блока. Как видно, производительность выше при размере блока 128 Мб.

Таблица 2

Размер блока	Количество Map-задач	Время процесса, с
32 Мб	16	64
64 Мб	7	50
128 Мб	4	39

В системе Apache Hive время для извлечения данных очень мало. Результаты экспериментов были загружены в систему. Время выполнения запросов показано в табл. 3, где можно видеть, что для извлечения целых данных потребовалось всего 10 с. Запросы, включающие

объединения и агрегатные функции, выполнялись как задания MapReduce. С помощью второго запроса системой подсчитано 131 091 набор данных за 93 с. Отсюда следует эффективность использования системы Apache Hive при управлении данными определенного формата.

Таблица 3

Номер п/п	Запрос	Время, с
1	Select * from table_name	10
2	Select count(*) from table_name	93
3	Select count(*) from table_name for certain conditions	53

Для узлов с IP-адресами, отправляющих аномально большое количество пакетов, созданы правила Snort. Если количество пакетов из конкретного источника превышает определенное значение, то будут сформированы правила, генерирующие предупреждения об аномальной активности. Это предпринимается во избежание количества ложных срабатываний. Для начального назначения правил Snort было использовано стандартное их количество — 2931, однако ни одно из них не генерировало никаких предупреждений, когда атаки были имитированы системой Hping3.

Затем, после анализа, были созданы новые правила Snort с соответствующими IP-адресами и номерами портов вместе с необходимыми параметрами. На рис. 2 показана вероятность обнаружения различных видов атак, которые активируются путем отправки большого количества пакетов. Новые правила были сформированы с помощью результатов, полученных в ходе анализа, выполненного Hadoop. Как следует из рисунка, созданные новые правила Snort эффективны при обнаружении атак ICMP, Smurf, SYN-flood, атаки на UDP и сканирования портов. Таким образом, генерация правил с помощью анализа является действенным методом.

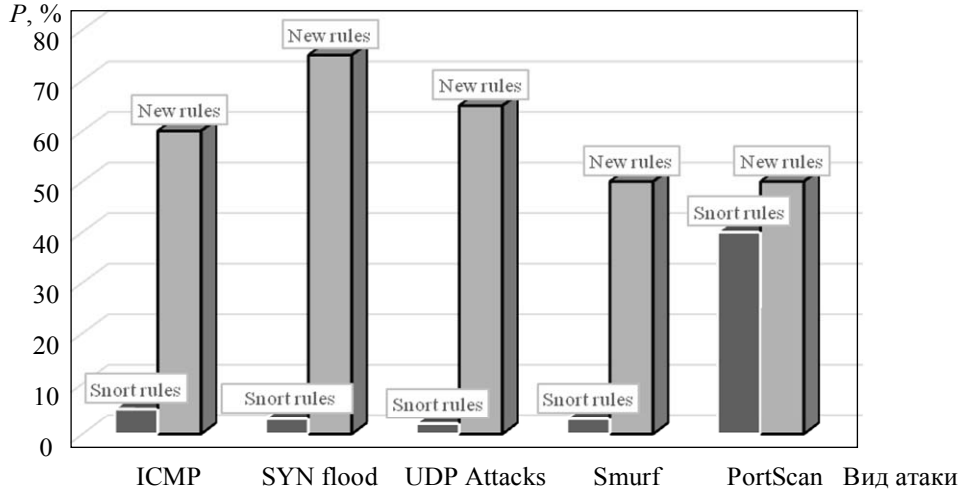

Рис. 2

Заключение. Продемонстрированный в экспериментах и представленный при сравнительном анализе высокий потенциал предложенной системы обнаружения компьютерных атак, при малых затратах вычислительных ресурсов, позволяет получить существенный выигрыш по сравнению с аналогами, что обеспечивает возможность проведения дальнейших исследований при больших объемах поступающей информации и событий безопасности.

Исследование выполнено при частичной финансовой поддержке РФФИ (проекты 16-29-09482 и 18-07-01488) и бюджетной темы № АААА-А16-116033110102-5.

СПИСОК ЛИТЕРАТУРЫ

1. *Котенко И. В., Ушаков И. А.* Технологии больших данных для мониторинга компьютерной безопасности // Защита информации. Инсайд. 2017. № 3. С. 23—33.

2. *Котенко И. В.* Интеллектуальные механизмы управления кибербезопасностью // Тр. Института системного анализа РАН. 2009. Т. 41. С. 74—103.

3. *Kotenko I., Stepashkin M.* Network security evaluation based on simulation of malefactor's behavior // Proc. of the Intern. Conf. on Security and Cryptography (SECRYPT 2006). Setubal, Portugal. 2006. P. 339—344.

4. *Novikova E., Kotenko I.* Analytical visualization techniques for security information and event management // Proc. of the 21st Euromicro Intern. Conf. on Parallel, Distributed, and Network-Based Processing. 2013. P. 519—525.

5. *Veeramachaneni K., Arnaldo I.* et al. AI2: Training a Big Data machine to defend [Электронный ресурс]: <people.csail.mit.edu>, 2016.

6. *Jeong Jin Cheon, D Tae-Young Choe.* Design of a hybrid intrusion detection system using Snort and Hadoop // Intern. Journal of Engineering and Technology. 2013.

7. *Muller A.* Event correlation engine: Master`s Thesis / Swiss Federal Institute of Technology. Zurich, 2009. 165 p.

8. *Guerer D. W., Khan I., Ogler R., Keffer R.* An Artificial Intelligence Approach to Network Fault Management / SRI International, CA USA, 1996. 10 p.

9. *Tiffany M.* A Survey of Event Correlation Techniques and Related Topics [Электронный ресурс]: <http://www.tiffman.com/netman/netman.html>.

10. *Elshoush H. T., Osman I. M.* Alert correlation in collaborative intelligent intrusion detection systems — A survey // Applied Soft Computing. 2011. P. 4349—4365.

11. *Jianguo Chen, Kenli Li* et al. A parallel random forest algorithm for big data in a spark cloud computing environment // IEEE Transact. on Parallel and Distributed Systems. 2016. P. 919—933.

12. *Marchal S., Xiuyan Jiang* et al. A Big Data Architecture for large scale security monitoring // Proc. of the 3rd IEEE Intern. Congress of Big Data. 2014. P. 56—63.

13. *Климов С. М., Сычёв М. П., Астрахов А. В.* Противодействие компьютерным атакам. Методические основы. М.: МГТУ им. Н. Э. Баумана, 2013. 108 с. (Электронное учебное издание.)

14. *Фор А.* Восприятие и распознавание образов / Пер. с франц.; Под ред. *Г. П. Катыса*. М.: Машиностроение, 1989. 272 с.

15. *Мазин А. В., Клочко О. С.* Анализ методов противодействия угрозам и атакам на вычислительные системы // Материалы Всерос. науч.-техн. конф. „Наукоемкие технологии в приборо- и машиностроении и развитие инновационной деятельности в вузе“. 2014. Т. 3. С. 71—75.

16. *Котенко И. В., Полубелова О. В., Саенко И. Б., Чечулин А. А.* Применение онтологий и логического вывода для управления информацией и событиями безопасности // Системы высокой доступности. 2012. Т. 8, № 2. С. 100—108.

Сведения об авторе

Николай Александрович Комашинский — аспирант; СПИИРАН; лаборатория проблем компьютерной безопасности; E-mail: nckkm@ya.ru

Поступила в редакцию
27.08.18 г.

Ссылка для цитирования: *Комашинский Н. А.* Комбинирование технологий Hadoop и Snort для обнаружения сетевых атак // Изв. вузов. Приборостроение. 2018. Т. 61, № 11. С. 1005—1011.

COMBINING HADOOP AND SNORT TECHNOLOGIES FOR DETECTION OF NETWORK ATTACKS

N. A. Komashinsky

St. Petersburg Institute for Informatics and Automation of the RAS,
199178, St. Petersburg, Russia
E-mail: nckkm@ya.ru

A method of information processing on the base of Big Data technologies aimed at computer attacks detection is studied. The need to create specialized approaches and design methods that will improve the efficiency of processing the received information is justified. The possibilities and effectiveness assessments of parallel data processing with the purpose of computer influences detection using a functional approach, as well as the key principles of working with Big Data, are considered. The mathematical model by means of which the technique of intrusion detection is developed is presented. The principle of

implementation of the tasks of information processing and anomaly detection based on integration of Hadoop, Snort platforms is described. Main results of the experimental evaluation of the method used to detect computer attacks are presented.

Keywords: Big Data, Hadoop, information system, information security, computer attack, Snort, anomaly, data processing

REFERENCES

1. Kotenko I.V., Ushakov I.A. *Zašita informacii. Inside*, 2017, no. 3, pp. 23–33. (in Russ.)
2. Kotenko I.V. *Trudy Instituta sistemnogo analiza rossiyskoy akademii nauk* (Proceeding of the Institute for Systems Analysis of the Russian Academy of Science), 2009, no. 41, pp. 74–103. (in Russ.)
3. Kotenko I., Stepashkin M. *Proc. of the Intern. Conf. on Security and Cryptography (SECRYPT 2006),* Setubal, Portuga, 2006, pp. 339–344.
4. Novikova E., Kotenko I. *Proc. of the 21st Euromicro Intern. Conf. on Parallel, Distributed, and Network-Based Processing*, 2013, pp. 519–525.
5. Veeramachaneni K., Arnaldo I. et al. *AI2: Training a big data machine to defend*, 2016, people.csail.mit.edu.
6. Jeong Jin Cheon, D Tae-Young Choe*. Intern. Journal of Engineering and Technology*, 2013.
7. Muller A*. Event correlation engine*, Master`s Thesis, Swiss Federal Institute of Technology, Zurich, 2009, 165 p.
8. Guerer D.W., Khan I., Ogler R., Keffer R. *An Artificial Intelligence Approach to Network Fault Management,* SRI International, CA USA, 1996, 10 p.
9. Tiffany M. *A survey of event correlation techniques and related topics*, http://www.tiffman.com/netman/netman.html.
10. Elshoush H.T., Osman I.M. *Applied Soft Computing*, 2011, pp. 4349–4365.
11. Jianguo Chen, Kenli Li et al*. IEEE Transact. on Parallel and Distributed Systems*, 2016, pp. 919–933.
12. Marchal S., Xiuyan Jiang et al*. Proc. of the 3rd IEEE Intern. Congress of Big Data*, 2014, pp. 56–63.
13. Klimov S.M., Sychyov M.P., Astrakhov A.V. *Protivodeystviye komp'yuternym atakam. Metodicheskiye osnovy* (Counteraction to the Computer Attacks. Methodical Bases), Moscow, 2013, 108 p. (in Russ.)
14. Faure A. *Perception et recon naissance des formes*, Paris, Editests, 1985, 286 p.
15. Mazin A.V., Klochko O.S. *Naukoyemkiye tekhnologii v priboro- i mashinostroyenii i razvitiye innovatsionnoy deyatel'nosti v vuze* (Science-Intensive Technologies in Instrument Engineering and Development of Innovative Activities in the University), Materials of the All-Russian Scientific and Technical Conference, 2014, vol. 3, pp. 71–75. (in Russ.)
16. Kotenko I.V., Polubelova O.V., Sayenko I.B., Chechulin A.A. *Highly available systems*, 2012, no. 2(8), pp. 100–108. (in Russ.)

Data on author

Nickolay A. Komashinsky — Post-Graduate Student; St. Petersburg Institute for Informatics and Automation of the RAS, Laboratory of Cyber-Security Problems; Junior Scientist; E-mail: nckkm@ya.ru

For citation: Komashinsky N. A. Combining Hadoop and Snort technologies for detection of network attacks. *Journal of Instrument Engineering*. 2018. Vol. 61, N 11. P. 1005—1011 (in Russian).

DOI: 10.17586/0021-3454-2018-61-11-1005-1011