УДК 004.048

Шыхалиев Р.Г.
Институт Информационных Технологий НАНА, Баку, Азербайджан
ramiz@science.az

О МЕТОДАХ СБОРА, ХРАНЕНИЯ И АНАЛИЗА БОЛЬШОГО СЕТЕВОГО ТРАФИКА

Сбор, хранение и анализ сетевого трафика компьютерных сетей (КС) являются основными этапами процесса мониторинга. Однако в современных КС процесс сбора, хранения и анализа полного сетевого трафика представляет собой очень сложную проблему. Так как с ростом скорости и масштаба КС растут и объемы сетевого трафика, который в день может потребовать петабайтов объемов памяти для хранения. Существуют различные методы сбора, хранения и анализа сетевых данных, которые при правильном выборе могут существенно уменьшить объем собранных данных, а следовательно, объем требуемых для анализа и хранения данных. В статье рассматриваются подходы к решению вопросов сбора, хранения и анализа большого сетевого трафика с применения Big Data технологий.

Ключевые слова: *компьютерные сети, мониторинг, сетевой трафик, сбор сетевого трафика, хранение сетевого трафика, анализ сетевого трафика, Big Data технологии.*

Введение

Сегодня компьютерные сети, особенно Интернет, стали глобальной инфраструктурой, обеспечивающей повсеместно доступные, интерактивные и безопасные сервисы. Для того чтобы обеспечить высокий уровень QoS (*Quality of Service*) этих сервисов, необходима эффективная инфраструктура мониторинга. При этом самым подходящим методом для инфраструктуры мониторинга КС является пассивный мониторинг [1], который позволит определить общее состояние и безопасность КС, а также уровень QoS предоставляемых сервисов и т.д. При этом процесс сбора, хранения и анализа сетевых данных является основой пассивного мониторинга и представляет из себя более сложную и важную проблему, в особенности при мониторинге больших КС. Так как с повышением скорости и масштабов КС растет и объем сетевого трафика, который в день может измеряться петабайтами. Вместе с тем для того, чтобы провести всеобъемлющий анализ состояния и безопасности КС, необходимо произвести сбор и хранение всей информации о трафике. Например, при мониторинге безопасности КС сбор всех сетевых пакетов необходим для обнаружения вредоносных активностей или определения вирусных атак (например сетевыми червями) и т.д.

Известно, что при пассивном мониторинге КС собираются большие объемы данных мониторинга, что приводит к возникновению проблем, связанных с их хранением, и уменьшению эффективности анализа. Поэтому, наряду с необходимостью разработки новых методов анализа больших объемов сетевых трафиков, актуальной проблемой является разработка новых подходов для сбора и хранения больших сетевых трафиков для мониторинга КС. При этом очень важным является сокращение размерности признакового пространства сетевого трафика, который используется для мониторинга КС [2]. Исходя из этого, в статье осуществляется анализ подходов к решению вопросов сбора, хранения и анализа большого сетевого трафика с применения Big Data технологий.

Появление технологий Big Data обуславливалось различными факторами [3]. В основном к этим факторам относят объем, многообразие и скорость данных, и при очень больших значениях этих характеристик сбор, хранение и обработка данных с помощью традиционных технологий становится трудным делом. Поэтому для сбора, хранения и обработки больших объемов данных и извлечения из них полезной информации требуются

новые технологии. Для решения этих задач такими фирмами, как IBM [4], Oracle [5], Microsoft [6], SAS [7], SAP [8], HP [9] и др., были предложены различные подходы.

Методы сбора сетевого трафика

Существующие сегодня средства мониторинга осуществляют сбор различных типов и объемов информации. При этом имеются три основных метода сбора трафика, которые имеют различные требования к объему памяти: сбор всех пакетов; сбор сетевого потока и так называемый сбор расширенного потока.

Целью сбора пакетов является сбор всего сетевого трафика, который генерируется компьютерами и устройствами КС, при котором осуществляются сбор и хранение данных заголовка каждого пакета и передаваемой в пакетах информации. Другими словами, производятся сбор, обработка и хранение копии каждого пакета трафика для последующего анализа. Эти собранные данные обеспечивают аналитиков полной информацией о трафике: информацией заголовков пакетов и передаваемой в пакетах информацией. Следовательно, такой метод сбора данных мониторинга может быть наиболее универсальным, так как большой объем информации может интенсивно храниться и обрабатываться [10].

Сетевой поток определяется как множество IP-пакетов, проходящих через точку наблюдения в сети в течение определенного интервала времени. Все пакеты, принадлежащие к определенному потоку, имеют набор общих свойств. Требования к потокам IP-пакетов определены в RFC 3917 [11]. Согласно приведенному определению, сетевой поток представляет из себя потоки сетевых пакетов, для которых выполняются следующие условия:

- происходят в течение одного и того же периода времени;
- имеют один и тот же адрес источника и номер порта;
- имеют один и тот же адрес назначения и номер порта;
- используют один и тот же протокол.

При этом, если не рассматривать передаваемую в пакетах информацию и информацию некоторых полей заголовков пакетов, а также объединить некоторые пакеты, то уменьшится объем данных, что приведет к уменьшению требуемой памяти для хранения потоков. Однако это приведет и к уменьшению качества анализа сетевого трафика [12].

Сбор расширенного потока включает в себя сбор всех пакетов и сетевого потока. При этом к информации потока добавляется информация, взятая непосредственно из заголовков пакетов или из передаваемой в пакетах информации. Вместе с тем расширенный поток также может содержать дополнительную информацию о каком-то внешнем источнике, например, о географическом расположении IP-адресов источника и назначения. Поэтому некоторые решения сбора расширенного потока рассматривают эту информацию как метаданные [13].

Большая часть современных исследований в области сбора сетевого трафика посвящена вопросам сбора пакетов в скоростных сетях с минимальной потерей данных и сжатию данных после сбора, то есть снижению объемов. Например, в работах [14, 15] авторы соответственно обсуждают вопросы преобразования данных для их эффективного хранения и обработки и мониторинга в облаке (*cloud*). В работах [16, 17] авторы предлагают подход к разработке приложений по сбору данных в скоростных сетях, основанный на стандартных аппаратных средствах. А в работе [18] для полного анализа сетевого трафика авторы предлагают метод агрегации потоков.

Методы хранения сетевого трафика

Другой проблемой эффективного анализа сетевого трафика является хранение собранных данных, которые должны сохраняться достаточно долго и надежно, чтобы при необходимости аналитики могли иметь доступ к ним. При этом, в зависимости от места и

способа хранения данных, могут существенно изменяться требуемый объем памяти для хранения, а также возникать проблемы, связанные с администрированием и обслуживанием, и т.д. Вместе с тем данные могут быть сохранены локально в организации, облаке или другом внешнем хранилище и использованы различные способы хранения данных, такие как: файлы (например, лог-файлы); базы данных и их комбинации. Каждый из этих способов имеет собственные аспекты.

Обычно в большинстве организаций КС производят сбор сетевого трафика в нескольких точках. Поэтому очень важен выбор места физического расположения собранных данных. Например, централизованное хранение всех собранных данных в одном месте может упростить управление и анализ данных, однако это требует передачи данных в центр, что приводит к неэффективному использованию полосы пропускания каналов передачи сети. А также такой способ хранения данных нецелесообразен с точки зрения безопасности, так как при компрометации хранилища может произойти несанкционированное изъятие данных. Альтернативой централизованного хранения данных является распределенное хранение данных, однако при таком подходе хранения усложняется процесс анализа данных, а также администрирования и обслуживания. Одним из видов распределенного хранения данных можно считать облачное хранение данных [19, 20], которое может осуществлять также и сбор данных.

Большинство средств сбора сетевого трафика записывают полученные данные в файлы (лог-файлы) и обычно имеют собственные форматы файлов. При этом очень важно знать формат хранимого файла, чтобы без затруднения организовать передачу данных между приложениями сбора и анализа данных, так как большинство из них поддерживают определенные форматы файлов. Однако имеются некоторые общие форматы (например, pcap), которые поддерживаются большинством приложений сбора и анализа данных. Вместе с тем форматы создаваемых файлов могут определить необходимые объемы для хранения файлов. Несмотря на то, что в малых объемах данных различия между форматами несущественны, в больших объемах данных выбор того или иного формата является существенным. Уменьшение объема памяти, необходимого для хранения данных, также может быть достигнуто сжатием данных, которое может сделать хранение и анализ данных более эффективными. Эффективный алгоритм сжатия может не только уменьшить дисковое пространство, необходимое для хранения данных, но также уменьшить время, необходимое для извлечения данных с этого диска. Например, алгоритм сжатия данных lzo1x может уменьшить размер записей примерно на 50% [21].

Некоторые средства сбора данных для хранения данных могут использовать базы данных. Приложения, которые поддерживают только файлы, могут хранить их в базе сами или с помощью приложения анализа данных. Аналитик это может сделать вручную. Вместе с тем при использовании базы данных для хранения сетевого трафика необходимо учитывать ожидаемый размер, количество записей и свойства базы данных, ограничивающих общий размер базы данных, размер записи и т.д. Учитывая то, что реляционные базы данных не так масштабируемы, как при хранении данных в виде файлов, для решения проблемы масштабируемости могут быть использованы NoSQL базы данных [22].

Сетевой трафик как Big Data

Исследования сетевого трафика показали, что он представляет собой сложный динамический процесс и является суперпозицией многих потоков с множественными взаимосвязанными характеристиками, которые генерируются различными протоколами. Во-первых, это трафики, связанные с управлением КС (например, трафик инициализации клиентов, серверный трафик и т.д.), которые генерируются периодически. Во-вторых, это трафики сетевых сервисов, приложений (например, DNS, FTP, запросы WINS, ARP, сеанс NetBIOS, HTTP, P2P, SMTP, POP3, Telnet и т.д.) и протоколов, которые составляют

основную часть сетевого трафика КС [23]. При этом для того, чтобы проанализировать сетевой трафик КС с помощью методов Big Data, необходимо определить, что данные сетевого трафика удовлетворяют характеристикам Big Data. Потому, что для анализа не всех данных могут потребоваться методы Big Data и с помощью традиционных методов анализа может быть проведен достаточно эффективный анализ. Так как сегодня нет единого мнения по принципиальному вопросу о том, насколько большими должны быть данные, чтобы квалифицировать их как Big Data. Поэтому прежде чем анализировать большой сетевой трафик, необходимо определить его характеристики с точки зрения Big Data. То есть, при каких значениях характеристик объема, многообразия и скорости данные сетевого трафика можно считать Big Data. Это очень важная задача, решение которой позволит создать эффективные Big Data модели для анализа больших сетевых трафиков, так как определение значений этих характеристик даст возможность выбрать эффективные Big Data-технологии.

Обычно при мониторинге КС для централизованного сбора и анализа потока данных сетевого трафика используются высокопроизводительные серверы с большой памятью. Однако при мониторинге крупных КС, например общегосударственных, приходится иметь дело с тера- или петабайтами информации. А также при вирусных заражениях (вспышке сетевых червей) или DDoS (Distributed denial of service) появляется необходимость быстро обрабатывать большой объем данных. В таких случаях для того, чтобы проанализировать трафик, за короткое время невозможно вычислить статистику трафика из большого потока данных. Для решения этой проблемы, то есть для уменьшения объема постоянно поступающего потока данных трафика, традиционно используется метод выборки или агрегации [24, 25]. Однако при таких подходах необходимо заранее знать характеристики трафика.

Big Data методы анализа данных

Сегодня в мире Big Data технологии привлекают очень большое внимание и в этой области имеется множество исследований и разработок. К ним можно отнести исследования и разработки в области хранения и обработки данных в большом масштабе, таких, как облачное вычисление (Cloud Computing) [26], MapReduce, Hadoop [27], а также методы анализа данных – машинного обучения и интеллектуального анализа данных (Data Mining). Например, компании Google, Yahoo, Amazon, Facebook разработали и используют платформы кластерных файловых систем и облачных вычислений. Компания Google разработала модель параллельного программирования MapReduce для ранжирования веб-страниц и анализа веб-журналов, которая поддерживает распределенные вычисления и имеет две функции, такие, как map и reduce [28], где функция map обрабатывает пары ключ/значение для генерации набора пар промежуточных ключ/значения, а функция reduce объединяет все промежуточные значения, связанные с одним и тем же промежуточным ключом. В компании Google работают тысячи машин для MapReduce, чтобы обработать большие наборы веб-данных. После того, как Google объявила о разработке модели MapReduce, Yahoo выпустила систему Hadoop [29] для платформы облачных вычислений, которая может легко обрабатывать очень большие файлы с потоковой моделью доступа. А Amazon представляет сервисы облачных вычислений на основе Hadoop, такие, как Elastic Compute Cloud (EC2) или простой сервис хранения (Simple Storage Service (S3)) [30]. Сегодня Facebook также использует Hadoop для анализа данных веб-журнала социальной сети [31].

Сегодня в литературе имеется ряд работ, которые посвящены применению указанных Big Data технологий для мониторинга КС. С помощью этих технологий из огромного количества сетевых данных может быть получена полезная информация, которую раньше без таких технологий невозможно было получить. В работе [32] авторы предлагают метод анализа потока интернет-трафика на основе программного обеспечения MapReduce в рамках

платформы облачных вычислений. В работе [33] авторы представляют систему мониторинга сетевого трафика на основе Hadoop, который выполняет IP, TCP, HTTP и NetFlow анализ терабайтов интернет-трафика. В работе [34] автор обсуждает проблемы классификации Big Data данных с использованием методов геометрического представления-обучения и современных Big Data технологий. В частности, автор рассматривает вопросы комбинирования методов обучения с учителем, представления-обучения, постоянного машинного обучения (machine lifelong learning) и Big Data технологий (например Hadoop, Hive и Cloud) для решения задач классификации сетевого трафика.

Заключение

Сегодня мониторинг КС является одним из основных средств обеспечения их нормальной работы и безопасности. При этом сбор, хранение и анализ сетевого трафика являются основой мониторинга. Для получения полной информации о деятельности КС необходимы постоянный и полный сбор, хранение и анализ сетевого трафика, что позволяет своевременно и эффективно реагировать на отказы в работе и инциденты безопасности. Однако это требует постоянного сбора и хранения большого объема данных мониторинга, что может вызвать необходимость создания огромных объемов памяти, а также снизить эффективность анализа собранных данных. Другой причиной проблемы сбора, хранения и анализа большого объема сетевого трафика, на наш взгляд, является неправильный выбор соответствующих методов сбора и хранения данных в зависимости от задачи мониторинга.

В статье были проанализированы существующие методы сбора, хранения и анализа сетевого трафика, а также проблемы, имеющиеся в этой области. В результате анализа можно сказать, что при сборе сетевого потока или расширенного потока требуется гораздо меньше памяти для хранения. Вместе с тем в условиях чрезмерно большого объема сетевого трафика решать задачу анализа сетевого трафика традиционными методами анализа становится трудно. Для решения этой задачи более подходящим является использование технологий Big Data.

Проведенный в статье анализ методов сбора, хранения и анализа сетевого трафика поможет администраторам КС, национальным провайдерам и т.п. в соответствии с задачей мониторинга выбрать необходимый метод.

Литература

1. Şıxəliyev R.H. Kompüer şəbəkələrinin monitorinqi üsulları və vasitələri haqqında // İnformasiya cəmiyyəti problemləri, 2011, №2, s. 61–70.
2. Шыхалиев Р.Г. Об одном методе сокращения размерности анализируемых признаков сетевых трафиков, используемых для мониторинга компьютерных сетей // Телекоммуникации, 2011, № 6, с. 44–48.
3. Əliquliyev R.M., Hacırəhimova M.Ş. "Big data fenomeni: problemlər və imkanlar // İnformasiya cəmiyyəti problemləri", 2014, №2, s. 3–16.
4. InfoSphere Platform: Big Data Analytics, 2013, www-01.ibm.com/software/
5. Oracle and Big Data: Big Data for the Enterprise, 2013, www.oracle.com
6. Big Data, 2013, www.microsoft.com
7. Big Data – What Is It? 2013, http://www.sas.com/big-data/
8. SAP HANA integrates predictive analytics, text and big data in a single package, 2013, www54.sap.com/
9. Big Data Solutions, 2013www8.hp.com/
10. Bejtlich R. Why Collect Full Content Data?, http://taosecurity.blogspot.com, 2012
11. Quittek J., Zseby T., Claise B., Zander S., RFC 3917: Requirements for IP Flow Information Export (IPFIX). Internet Engineering Task Force, 2004. http://tools.ietf.org/html/rfc3917

12. RFC 7011, Specification of the IP Flow Information Export (IPFIX) Protocol, a standardized network flow format, provides a more technical definition of flow. http://tools.ietf.org/search/rfc7011
13. National Information Standards Organization (NISO). Understanding Metadata. NISO, 2004.
14. Aceto G., Botta A., Pescape A., Westphal C. Efficient Storage and Processing of High-Volume Network Monitoring Data // IEEE Transactions on Network and Service Management, 2013, vol.10, no.2, pp.162–175.
15. Aceto G., Botta A., de Donato W., Pescape A. Cloud Monitoring: A Survey // Computer Networks, 2013, vol.57, no.9, pp.2093–2115.
16. Deri L., Cardigliano A., Fusco F. 10 Gbit Line Rate Packet-to-Disk Using n2disk / Proceedings IEEE INFOCOM, 2013, pp.3399–3404.
17. Banks D. Custom Full Packet Capture System, SANS, 2013.
18. Francois J. State R., Engel T. Aggregated Representations and Metrics for Scalable Flow Analysis / IEEE Conference on Communications and Network Security (CNS), 2013, pp.478–482.
19. Sivashakthi T., Prabakaran N. A Survey on Storage Techniques in Cloud Computing // International Journal of Emerging Technology and Advanced Engineering, 2013, vol.3, no.12, pp.125–128.
20. Spoorthy V., Mamatha M., Santhosh Kumar B. A Survey on Data Storage and Security in Cloud Computing / International Journal of Computer Science and Mobile Computing, 2014, vol.3, no.6, pp.306–313.
21. Software Engineering Institute, Carnegie Mellon University. SiLK FAQ https://tools.netsa.cert.org/silk/faq.html (2014).
22. http://nosql-database.org/
23. Шыхалиев Р.Г. Анализ и классификация сетевого трафика компьютерных сетей // İnformasiya texnologiyaları problemləri, 2010, №2, с.15–23.
24. Hohn N. and Veitch D. Inverting sampled traffic / Proceedings of the 3rd ACM SIGCOMM conference on Internet measurement, 2003, pp.222–233.
25. Duffield N., Lund C. and Thorup M. Properties and prediction of flow statistics from sampled packet streams / Proceeding of the 2nd ACM SIGCOMM Workshop on Internet measurment, 2002, pp.159–171.
26. Carlin S, and Curran K. Cloud Computing Technologies // International Journal of Cloud Computing and Services Science (IJ-CLOSER), 2012, vol.1, no.2, pp.59–65.
27. Hadoop, http://hadoop.apache.org/
28. Dean J., and Ghemawat S. MapReduce: Simplified Data Processing on Large Cluster // Magazine Communications of the ACM, 2008, vol.51 no.1, pp.107–113.
29. https://developer.yahoo.com/hadoop/
30. http://wiki.apache.org/hadoop/AmazonEC2
31. http://borthakur.com/ftp/hadoopmicrosoft.pdf
32. Lee Y., Kang W., Son H. An Internet Traffic Analysis Method with MapReduce / Proceedings of the Network Operations and Management Symposium Workshops (NOMS Wksps), 2010 IEEE/IFIP, 2010, pp.357–361.
33. Lee Y., and Lee Y. Toward Scalable Internet Traffic Measurement and Analysis with Hadoop // ACM SIGCOMM Computer Communication Review, 2013, vol.43, no.1, pp.6–13.
34. Shan S., Big data classification: problems and challenges in network intrusion prediction with machine learning / ACM SIGMETRICS Performance Evaluation Review, 2014, vol.41, no.4, pp.70–73.

UOT 004.048
Şıxəliyev Ramiz H.
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
ramiz@science.az
Böyük şəbəkə trafikinin toplanması, saxlanması və analizi üsulları haqqında
Kompüter şəbəkələrinin (KŞ) şəbəkə trafikinin toplanması, saxlanması və analizi onların monitorinqi prosesinin əsas mərhələləridir. Lakin, müasir KŞ-lərdə tam şəbəkə trafikinin toplanması, saxlanması və analizi prosesi çox mürəkkəb məsələyə çevrilir. Çünki KŞ-lərin sürətinin və miqyasının artması ilə şəbəkə trafikinin həcmi böyüyür və gün ərzində saxlanılması üçün petabaytla ölçülən yaddaş tələb edə bilər. Şəbəkə trafikinin toplanması, saxlanması və analizinin mövcud üsullarının düzgün seçilməsi ilə toplanan verilənlərin və nəticədə tələb olunan yaddaşın həcminin əhəmiyyətli dərəcədə azaldılmasına imkan verər. Bunu nəzərə alaraq şəbəkə trafikinin toplanması, saxlanması və analizi məsələlərinin həllinə yönəlmiş "Big data" texnologiyalarına əsaslanmış yanaşmalara baxılır.
Açar sözlər: *kompüter şəbəkələri, monitorinq, şəbəkə trafiki, şəbəkə trafikinin toplanması, şəbəkə trafikinin saxlanması, şəbəkə trafikinin analizi, "Big data" texnologiyaları, "Big data" analizi üsulları*.

Ramiz H. Shikhaliyev
Institute of Information Technology of ANAS, Baku, Azerbaijan
ramiz@science.az
Methods of collection, storage and analysis of large network traffic
Collection, storage and analysis of the network traffic of computer networks (CNs) are the main stages of the monitoring process. However, in modern CNs, the process of collection, storage and analysis of the entire network traffic is a very complex problem. Whereas, growing speed and scale of CNs increases the volume of network traffic, which may require petabytes of storage capacity a day. There are various methods for collection, storage and analysis of network data; the correct choice of which can significantly reduce the volume of collected data and consequently, the volume required for storage. Hence, the paper describes the approaches to the solution of data collection, storage and analysis of large network traffic with the use of "Big data" technologies.
Keywords*: computer networks, monitoring, network traffic, network traffic collection, network traffic storage, network traffic analysis, network traffic analysis, "Big data" technologies.*