Об Одной Модели Анализа Данных Большого Сетевого Трафика

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

***Аннотация* — сегодня сложность конфигурации компьютерных сетей (КС) продолжает расти. Кроме того, объем трафика, передаваемого по этим сетям, также увеличивается. При этом анализ трафика является перспективным методом для обеспечения эффективности работы и безопасности КС. В статье исследованы возможности применения Big Data-технологий для анализа больших сетевых трафиков. В результате анализа существующих методов анализа Big Data предложена модель анализа большого сетевого трафика КС.**

Ключевые слова — компьютернве сети, мониторинг, большой сетевой трафик, анализ сетевого трафика, Big Data- технологии, методы анализа Big Data.

I. ВВЕДЕНИЕ

Сегодня компьютерные сети играют фундаментальную роль в социально-экономической инфраструктуре общества. Вместе с тем их масштаб и сложность постоянно растут. А также растет и количество сервисов, предоставляемых КС, особенно в сети Интернет, которая стала очень популярной. Широко стали использоваться CDN (Content Delivery Networks) и cloud-сервисы, а также значительно вырос объем потокового видео и т.д. В результате этого намного вырос объем общего трафика КС. В таких условиях для обеспечения эффективности работы и безопасности КС необходим постоянный мониторинг. Потому что постоянный мониторинг даст необходимую информацию, которая позволит планировать и управлять КС, а также обеспечить их безопасность.

Обычно мониторинг трафика КС осуществляется централизованно, что намного усложняет их постоянный мониторинг. Это прежде всего связано с необходимостью анализа очень большого объема гетерогенного и высокоскоростного трафика, а также постоянного обновления данных мониторинга. При этом анализ слишком большого объема данных требует соответствующих ресурсов хранения и вычисления. Однако, несмотря на высокую производительность традиционных методов и приложений анализа данных, они становятся бесполезными.

Целью статьи является решение указанной проблемы с помощью технологий Big Data. Термином Big Data обозначаются большие и сложные наборы данных, которые трудно обрабатывать с помощью традиционных методов или приложений [1]. Big Data определяется тремя характеристиками данных: объемом, многообразием и скоростью, и когда их значения становятся очень большими, то современные методы и приложения не могут справиться с хранением и обработкой данных. В этом контексте анализ сетевого трафика является проблемой Big Data. При этом для анализа сетевого трафика требуются высокоэффективные Big Data-методы и приложения, которые позволили бы решить проблему в реальном масштабе времени. Такой поход к анализу сетевого трафика позволил бы заблаговременно предупредить о предстоящих неисправностях и проблемах, а также об угрозах безопасности КС. Вместе с тем использование методов Big Data для анализа сетевого трафика позволит эффективно осуществлять сбор всевозможных данных о сетевом трафике и в полной мере оценить состояние всей КС.

II. 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 [2]. 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 проанализировать 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 выбирать 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, КС приходится иметь дело с тера- или петабайтами информации. А также при вирусных заражениях (вспышке сетевых червей) или DDoS (Distributed denial of service) атаке появляется необходимость быстро обработать большой объем данных. В таких случаях, 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 [3, 4]. Однако при таких подходах необходимо заранее знать характеристики трафика.

III. 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, такие, как облачное вычисление (Cloud computing) [5], MapReduce, Hadoop [6], а также методы анализа данных – методы машинного обучения и интеллектуального анализа данных (Data Mining). Например, компании Google, Yahoo, Amazon, Facebook разработали и используют платформы кластерных файловых систем и облачных вычислений. Google разработал модель программирования MapReduce для ранжирования веб-страниц и анализа веб-журналов, которая поддерживает распределенные вычисления и имеет две функции, такие, как отображение (map) и уменьшение (reduce) размеров больших наборов данных до кластеров [7]. В фирме Google работают тысячи машин для MapReduce, чтобы обработать большие наборы веб-данных. После того как Google объявила о разработке модели MapReduce, фирма Yahoo выпустила систему Hadoop [8] для платформы облачных вычислений, которая может легко обрабатывать очень большие файлы с потоковой моделью доступа. А компания Amazon предоставляет сервисы облачных вычислений на основе Hadoop, такие, как Elastic Compute Cloud (EC2) или простой сервис хранения (Simple Storage Service (S3)) [9]. Сегодня Facebook также использует Hadoop для анализа данных веб-журналов социальной сети [10].

Сегодня в литературе имеется ряд работ, посвященных 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 [11] авторы предлагают метод анализа потока интернет-трафика на основе программного обеспечения MapReduce в рамках платформы облачных вычислений. В работе [12] авторы представляют систему мониторинга сетевого трафика на основе Hadoop, который выполняет IP, TCP, HTTP и NetFlow анализ терабайтов интернет-трафика. В работе [13] 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, методов представления-обучения, методов постоянного машинного обучения (machine lifelong learning) и Big Data-технологий (например, Hadoop, Hive и Cloud) для решения задач классификации сетевого трафика.

IV. МОДЕЛЬ АНАЛИЗА БОЛЬШОГО СЕТЕВОГО ТРАФИКА

На основе анализа приведенных в предыдущем разделе методов Big Data можно сделать вывод о том, что для решения нашей задачи более подходящей технологией является облачное вычисление. Использование технологии облачного вычисления позволит построить большую и гибкую сетевую инфраструктуру хранения данных, которая может адаптивно изменяться в зависимости от требований обработки Big Data.

В качестве модели анализа большого сетевого трафика предлагается интегрированная модель, которая состоит из трех блоков (рис. 1): системы регистрации (записи) сетевого трафика (Network Traffic Recording System (NTRS); системы хранения облачных вычислений (Cloud Computing Storage System (CCSS)) и системы машинного обучения (Machine Learning System (MLS)). Такая интегрированная модель повысит эффективность обработки большого сетевого трафика.

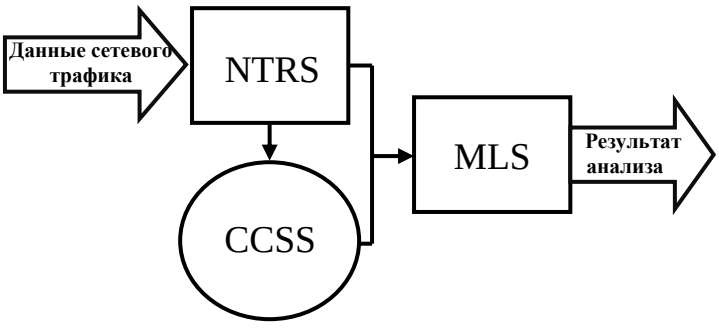

Рис.1. Модель анализа большого сетевого трафика

В предложенной модели блок NTRS, для реализации которого в литературе имеются различные методы [14, 15], получает данные трафика непосредственно из сети или из лог-файлов и записывает их в CCSS. В зависимости от поставленной задачи мониторинга блок MLS анализирует

данные, получаемые из блока CCSS или непосредственно из блока NTRS. В MLS в качестве метода обучения предлагается использовать Machine Lifelong Learning (ML3) [16, 17, 18] метод.

ЗАКЛЮЧЕНИЕ

В условиях постоянного роста объема и сложности трафика очень трудно обеспечить эффективность управления и безопасности КС. При этом постоянный мониторинг, одной из основных функций которого является анализ сетевого трафика, даст необходимую информацию, которая позволит планировать и управлять КС, а также обеспечивать безопасность. Однако в условиях чрезмерно большого объема сетевого трафика решать эту задачу традиционными методами анализа становится все труднее.

Для решения этой задачи в статье предложено использовать технологии Big Data. Для этого была выбрана модель анализа большого сетевого трафика, в которой интегрируются системы регистрации (записи) сетевого трафика (Network Traffic Recording System (NTRS), хранения облачных вычислений (Cloud Computing Storage System (CCSS)) и машинного обучения (Machine Learning System (MLS)).

На наш взгляд, предложенная модель позволит эффективно осуществлять сбор всевозможных данных сетевого трафика, анализировать и в результате этого в полном мере оценивать состояние всей КС.

ЛИТЕРАТУРА

[1] Zikopoulos P. C., Eaton C., et al., Understanding big data – Analytics for enterprise class Hadoop and streaming data, McGraw-Hill, 2012.

[2] Шыхалиев Р.Г. Анализ и классификация сетевого трафика компьютерных сетей // İnformasiya texnologiyaları problemləri, №2, 2010, с. 15–23.

[3] Hohn N. and Veitch D. Inverting sampled traffic, In Proceedings of the 3rd ACM SIGCOMM conference on Internet measurement, 2003, pp. 222–233.

[4] Duffield N., Lund C. and Thorup M., Properties and prediction of flow statistics from sampled packet streams, Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment. New York,

[5] Carlin S and Curran K.. Cloud Computing Technologies. International Journal of Cloud Computing and Services Science (IJ-CLOSER) 1.2: 59–65, 2012.

[6] Hadoop, http://hadoop.apache.org/

[7] Dean J. and Ghemawat S., MapReduce: Simplified Data Processing on Large Cluster, OSDI, 2004. NY, USA: ACM, 2002, pp. 159–171.

[8] https://developer.yahoo.com/hadoop/

[9] http://wiki.apache.org/hadoop/AmazonEC2

[10] http://borthakur.com/ftp/hadoopmicrosoft.pdf

[11] Youngseok Lee, Wonchul Kang, Hyeongu Son, An Internet Traffic Analysis Method with MapReduce, Proceedings of the Network Operations and Management Symposium Workshops (NOMS Wksps), 2010 IEEE/IFIP, 19-23 April 2010, pp. 357 – 361.

[12] Yeonhee Lee, and Youngseok Lee, Toward Scalable Internet Traffic Measurement and Analysis with Hadoop, ACM SIGCOMM Computer Communication Review, vol. 43, num. 1, 2013, pp. 6–13.

[13] Suthaharan Shan, Big data classification: problems and challenges in network intrusion prediction with machine learning, ACM SIGMETRICS Performance Evaluation Review, vol. 41 issue 4, 2014, pp. 70–73.

[14] Papadogiannakis A., Polychronakis M., Markatos Evangelos P. Long-term Raw Network Traffic Recording using Fixed-size Storage, Proceedings of the 2010 IEEE International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems, 2010, pp. 101–110.

[15] Stefan Kornexl, Vern Paxson, et al., Building a time machine for efficient recording and retrieval of high-volume network traffic, Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement, 2005, pp. 23–23.

[16] Qiang Y., Big data, lifelong machine learning and transfer learning, Proceedings of the sixth ACM international conference on Web search and data mining, 2013, pp. 505–506.

[17] Daniel L. Silver and Ryan Poirier, Requirements for Machine Lifelong Learning, Proceedings of the 2nd international work-conference on the Interplay Between Natural and Artificial Computation, Spain, June 18-21, 2007, Part I, pp. 313–319.

[18] Daniel L. Silver, Qiang Y and Lianghao Li, Lifelong Machine Learning Systems: Beyond Learning Algorithms, AAAI Spring Symposium, 2013, pp. 49–55.