Ramiz H. Shikhaliyev
Institute of Information Technology of ANAS, Baku, Azerbaijan
ramiz@science.az

DOI: 10.25045/jpit.v07.i2.06

ABOUT THE METHODS OF COLLECTING, STORING AND ANALYZING BIG NETWORK TRAFFIC

Collecting and storing network traffic of computer networks (CNs) is one of the major stages of the monitoring process. However, collecting and maintaining full network traffic in the modern CNs is a very complex problem. With rising speed and scale of the CNs and network traffic size, petabytes of storage might be needed for a day. There are various methods for network data collecting and storing. Their correct choice can significantly reduce collected data size and, respectively, the required storage size. The article examines the issues of network data collection and storage with the use of Big Data technology.

Keywords: *computer network, monitoring, network traffic, network traffic collection, network traffic storage, analysis of network traffic, Big Data technology*

Introduction

Today, computer networks (CNs), especially the Internet, have become a global infrastructure for interactive, secure and everywhere accessible services. In order to provide high QoS (Quality of Service), effective monitoring infrastructure needed. In this case, the most proper method for the infrastructure of the CNs monitoring is a passive monitoring [1]. Passive monitoring finds the general condition of the CNs and safety, as well as the QoS level of provided services, etc. For this, it is necessary to constantly collect, store and analyze the network traffic, i.e. to constantly carry out the collection, storage and analysis of large data scale. However, this is a very difficult task, especially during the monitoring of large CNs. Since, with rising speed and scale of the CNs and network traffic size, a day petabytes storage maybe needed. However, to conduct a comprehensive analysis of the condition and safety of the CNs, it is necessary to make the collection and storage of all traffic data. For example, at the CNs security monitoring collecting all network packets is needed to detect malicious activity or viral attacks (e.g., worms), etc.

It is known that, while passive monitoring of the CNs the large size of monitoring data collected. This leads to the problems associated with their storage, which cuts the efficiency of the analysis. Therefore, along with the need to develop new methods for analyzing large network traffic, the real problem is the development of new approaches to collect and store large network traffic for the CNs monitoring. Thus, reduction of network traffic feature dimension space, which is used to monitor the CNs [2] is very important. The article examines the issues of network data collection and storage with the use of Big Data technology.

The advent of Big Data technology depends on various conditional factors [3]. Basically, these factors include the volume, variety and velocity of data, and the collection, storage and processing of data with traditional technologies become difficult. Therefore, the collection, storage and processing of large amounts of data and extracting useful information from them requires new technologies. To meet these challenges, companies such as IBM [4], Oracle [5], Microsoft [6], the SAS [7], SAP [8], HP [9] etc., offer a variety of approaches.

Network traffic collection methods

Today, available monitoring tools collect different types and size of information. At the same time, there are three main methods of traffic collection, which have different requirements to memory size: all packets' collection; network flow collection; and so-called enhanced flow collection.

The aim of all packets' collection is the collection of all network traffic, which is generated by computers and devices of the CNs, while each packet header and transmitted information is collected and stored for further analysis. This is collected data provides analysts with complete information about the traffic, i.e. about packet headers and transmitted information. Therefore, this monitoring data collection method may be the most multipurpose, since a large size of information can be intensively stored and processed [10].

Network flow is defined as IP-packets passing through the observation point in the network during a certain time interval. All packets belonging to a particular flow have a set of common properties. Requirements to IP-packet streams are defined in RFC 3917 [11] and according to the given definition the network stream is a set of streams of network packages for which the following conditions are provided:

– occur during the same time period;

– have the same source address and port number;

– have the same destination address and port number;

– use the same protocol.

Thus, if not to consider information transferred in packages and information of some fields heading of packages and to unite some packages, then the data size is decreased, which leads to the reduction of the required memory for storage of flows. However, it leads to the reduction of quality of the analysis of network traffic [12].

Collection of the expanded flow includes collecting all packets and network flow. Thus, information of the packets headers or information transferred in packages is added to flow information. At the same time, the expanded flow may also contain more other information about any external sources such as the geographic location of the source and destination IP address. Therefore, in some expanded flow collecting solutions, this information is considered as metadata [13].

Most of current research in the field of collecting network traffic is devoted to the questions of collecting packages in high-speed networks with minimum, data loss and to the compression of data after collecting that is to decrease in size. For example, in [14, 15], the authors discuss questions of transformation of data for their effective storage and processing and monitoring in a cloud respectively. In works [16, 17], the authors offer approach to applications programming on data collection in high-speed networks based on standard hardware. And in work [18] for the full analysis of network traffic the authors offer a method of aggregation of flows.

Network traffic storage methods

Other problem of the effective network traffic analysis is a storage of collected data, which has to be stored long enough and reliably, and analytics may use them when needed. At the same time, depending on the place and the way of data storage, the required memory size for storage can be much definitely changed. Also, the problems can be arisen connected with the administration and service, etc. However, data can be stored locally in the organization, in a cloud or other external storage, and various ways of data storing can be used, such as: files (e.g., logs); database, and their combinations. Each of these methods has its own aspects.

Usually, in most organizations CNs collects network traffic at several points. Therefore, it is very important to choose the place of the physical place of collected data. For example, the centralized storage of all collected data in one place can simplify the management and analysis of data, but it requires a data transfer to the center, resulting in inefficient use of bandwidth network transmission channels. And this way of storage is impractical in terms of data security, since the unauthorized removal of data can occur when compromising storage. An alternative to the centralized data storage is the distributed data storage, but in this approach is complicated the process of data analysis, as well as administration and maintenance. One of distributed data storage types is cloud storage [19, 210], which can also implement data collection.

Most network traffic collecting tools record the obtained data in the files (log files) and usually have their own file formats. It is very important to know the format of the stored file to easily arrange the transfer of data between data collection and analysis applications, as the majority of them support certain file formats. However, there are some common formats (e.g., pcap), which are supported by most data collection and analysis applications. Thus, the format of the file may define the necessary size for file storage. Despite the fact that differences between the formats in small data are insignificant, but in large size data the choice of this or that format is essential. The reduction of memory size required to store the data can also be achieved by data compression, which can make the data storage and analysis more efficient. An efficient compression algorithm can not only cut the required disk space for data storage, but also reduce the time required to retrieve data from the disc. For example, lzo1x data compression algorithm can reduce the size of the records to about 50% [21].

Some data collection tools for storage data can use the database. Applications that support only files can store its own database, or using data mining application and the analyst can do it manually. However, when using a database to store network traffic, it is necessary to consider the expected size and the number of records and database properties limiting total size of the database and record size, etc. Considering that relational databases aren't scaled as at data storage in the form of files, then NoSQL databases, such as Hadoop may be used to solve of the scalability problem [22].

Network traffic as Big Data

Network traffic studies have shown that it is a complex and dynamic process is a superposition of many fluxes interlinked with multiple characteristics, which are generated by various protocols. Firstly, this traffic is associated with the CS control (for example, customer traffic initialization, the server traffic, etc.), which are generated periodically. Second, it is the traffic of network services, applications (e.g. DNS, FTP, WINS requests, ARP, NetBIOS session, HTTP, P2P, SMTP, POP3, Telnet, etc.) and protocols that make up the bulk of the network traffic CN [23]. Thus, in order to adopt the network traffic to the CN by methods Big Data, necessary to determine the data network traffic characteristics satisfy Big Data. Because, the analysis of data may not be required, and by the methods of Big Data sufficiently efficient analysis can be performed by conventional analysis techniques. Since there is no consensus today on the fundamental question of how big should the data to qualify as Big Data. Therefore, before analyzing large network traffic, you must define its characteristics in terms of Big Data. That is, for any value of volume characteristics, variety and velocity of data network traffic can be considered a Big Data. This is a very important task, which will allow to create effective Big Datamodeli to analyze large network traffic, as the determination of the values of these characteristics will make it possible to select an effective Big Data-technology.

Normally, high-performance servers with large memory is used when monitoring the CN for centralized data collection and analysis of network traffic flow. However, while monitoring the large CN, such as national, have to deal with tera- or petabytes of information. Moreover, in viral infections (outbreak of worms) or DDoS (Distributed denial of service), there is a need to quickly process large amounts of data. In such cases, in order to analyze the traffic in a short time can not be calculated traffic statistics from a large data stream. To solve this problem, i.e. to reduce the volume of traffic steadily supplied data stream, sampling method or aggregation is traditionally used, [24, 25]. However, such approaches need to know the traffic characteristics in advance.

Big Data methods of data analysis

Today, the world's Big Data technology attracted a lot of attention, and there is a lot of research and development in this area. These include research and development in the field of data storage and processing on a large scale, such as cloud computing (Cloud Computing) [26], MapReduce, Hadoop [27], as well as the methods of data analysis - Machine Learning and Data Mining (Data Mining). For example, Google's, Yahoo, Amazon, Facebook, and use the platform developed cluster file systems and cloud computing. Google has developed a parallel programming model MapReduce for ranking web pages and analyze web logs that supports distributed computing and has two functions, such as map and reduce [28], where the map function handles the key / value pair to generate a set of pairs of intermediate key / value, and reduce function unites all intermediate values associated with the same intermediate key. Google has thousands of machines that are working for MapReduce, to handle large Web data sets. After Google announced the development of MapReduce model, Yahoo released Hadoop system [29] for the cloud computing platform that can easily handle very large files to streaming access model. And Amazon is cloud computing services based on the Hadoop, such as the Elastic Compute Cloud (EC2) or a simple storage service (Simple Storage Service (S3)) [30]. Today, Facebook is also using Hadoop for analyzing web log data of social networks [31].

Today, there is a growing body of literature, devoted to the use of the Big Data technologies for the monitoring of the CN. With the help of these technologies useful information can be obtained from the huge amounts of network data, which was previously impossible without these technologies get. In [32] the authors propose a method of analysis flow of Internet traffic based on the MapReduce software within the cloud computing platform. In [33], the authors present the network traffic monitoring system based on Hadoop, which performs IP, TCP, HTTP, and NetFlow analysis of terabytes of Internet traffic. In [34] the author discusses the problems of Big Data classification data using the methods of geometrical representation, training and advanced Big Data technologies. In particular, the author examines the combination of teaching methods with the teacher, presentation, training, continuous learning machine (machine lifelong learning), and Big Data technologies (such as Hadoop, Hive and Cloud) to meet the challenges of network traffic classification.

Conclusion

Today, network traffic monitoring of the CNs is one of the main tools to ensure their proper operation and safety, and data collection and storage is the basis of monitoring. For complete information on the activities of the CNs, the constant and complete collection and storage of network traffic is needed, which allows prompt and effective response to refusals and security incidents. However, this needs a constant collection and storage of large size of monitoring data that may need excessive storage size, which reduces the effectiveness of the analysis of the collected data. Another reason for the problem of collecting and storing large size of network traffic, in our opinion, is improper choice of proper data collection and storage methods.

The article analyzes the existing methods for collection and storage of network traffic, as well as problems existing in this field. As a result of the analysis, we can say that collection of the network flow or enhanced flow requires much less memory for the storage. However, to solve the problem of network traffic analysis with traditional methods of analysis in an extremely high volume of network traffic, it becomes difficult. To solve this problem, it is rather suitable to use a more appropriate technologies Big Data.

This analysis can help CNs administrators to select the desired method according to the monitoring task.

References

1. Shikhaliyev R.H. About the methods and tools for Computer networks monitoring //Problems of Information Society, 2011, No2, pp.61-70,.
2. Shikhaliyev R.G. About the method for reducing the dimension of the analyzed features of network traffic used to monitor computer networks // Telecommunications. - No6. pp. 44-48, 2011.
3. Alguliyev R.M., Hajirahimova M.S. “Big data phenomenon: Challenges and Opportunities” // Problems of Information Society, 2014, No2, pp. 3-16.
4. InfoSphere Platform: Big Data Analytics, 2013, http://www-01.ibm.com/software/
5. Oracle and Big Data: Big Data for the Enterprise, 2013, http://www.oracle.com/
6. Big Data, 2013, http://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, http://www54.sap.com/
9. Big Data Solutions, 2013, http://www8.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. Shikhaliyev R.G. Analysis and classification of network traffic of computer networks // Problems of Information Society, 2010, No2, pp.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.