A Survey on Campus Network Log

Derong Zhou[a, *]

Network Information Center, Sichuan Minzu College, Kangding, Sichuan, China

[a]zhoudr0310@163.com

Keywords: Campus network, Log collection, Log storage, Hadoop, Log analysis.

Abstract. The original data about the operation status of IT infrastructure is stored in the log. It can not only reflect the operation status of the whole system but also provide the basis for the management and optimization of IT operation and security analysis. This paper has summarized the infrastructure of campus network in the universities. The main log source of campus network has been also discussed. The methods of collecting the log and mainstream protocol for log collection are analyzed, based on which the technologies for log storage and analysis will be discussed deeply. Moreover, this paper has analyzed and summarized the popular research topics and flaws in the technologies.

1. Introduction

With the fast development of internet, the computer network has been exerting a strong influence on various industries in the society. In order to adapt to the new situation, many universities have made efforts to build the campus network. The campus network is an important platform for teaching, research and informatization management. It can integrate various systems, such as teaching, research, management and office, into a coherent whole, thus realizing the information sharing and interaction between different systems. Hence, it can play a key role in ensuring the normal operation of the university.

In order to ensure the safe and stable operation of the campus network and various systems, many universities have invested heavily in purchasing a variety of security-enhancing equipment and software, such as firewall, IDS and IPS. But they have failed to get the desired results. That is mainly because the role of log system in the management of campus network has been neglected. At present, the main network equipment used in the campus network includes the router, switch, and firewall and intrusion detection. The mainstream operating systems have included Windows, Linux and Unix. A variety of application systems can be executed on those systems. During the operation of the hardware and software systems, a large quantity of logs will be generated in the source. The trace left by the users and intruders will be recorded in a wide variety of logs with different formats. It can directly reflect the operation status of the whole system of campus network. The comprehensive analysis of various equipment and software systems must be carried out so as to identify the abnormality in the campus network. It can also effectively monitor, analyze and remove the security hazard in the network. The platform devoted to collecting log information must be built to analyze and process the information. In this way, the problems existing in the operation of campus network will be solved effectively. Hence, it is quite necessary to research the log in the campus network.

2. Source analysis of campus network log

2.1 Introduction of college campus network infrastructure

College campus network typically adopts layering architecture of core layer, distribution layer and access layer, with 10G switchboard as core, distribution using Gigabit internet, and 100M to desktop. It is divided into teaching, office, dorm, library, network management subnets based on function; is divided into large, medium and small sized campus network based on quantity of users. College campus network has a huge number of user, various net application and large traffic, its outlet method typically uses multiple links of CERNET, China Telecom, China Mobile, etc., ensuring quality of internetworking.

When constructing network infrastructure hardware environment, focus on supplying network service externally, typically constructing multiple basic applications such as WWW, DNS, email system, teaching management system, office automatic system, library management system, asset management system, financial management system, recruitment and employment management system, video request, campus data center, united port platform, ID certification platform, etc. and management system.

In order to fully play the supporting role of campus network in education and teaching, scientific research and management, etc., every college attaches high importance to construct an advanced, highly applicable, highly safe and highly expansive campus network.

2.2 Source of campus network log

College campus network mainly consists of network infrastructure and network service and application. System heterogeneous of network, diversity of equipment, complexity of software environment, various infrastructure equipment and software can all produce log. The source of campus network log is mainly the following:

2.2.1 Network equipment log

Network equipment mainly consists of routing exchange equipment, firewall, invasion monitoring system, UPS system with network function, etc. The log system in network equipment is an important function, quickly knowing and diagnosing relevant problems by looking up logs in the switchboard, router and other network equipment. Due to difference in manufacturer and standard of equipment, there is different formats when producing log.

2.2.2 System log

System log includes Windows system log and Linux system log. Windows and Linux are the most common operating systems in campus network, and their safe operation is the basis for campus network to provide service. Windows system log refers to events produced in operation of every component in windows operating system, mainly including critical problems in operation of various drivers, operation of various components of operating system. Produced system log can be read through Windows event viewer or third-party work. Linux system log mainly consists of logon time log, process statistic log and error log, and log record is mainly text document, being read through system tool.

2.2.3 Application service log

Application service log refers to record of critical event produced by various applications in system operation. Campus network mainly provides Web service, FTP service, domain name service, database service, etc., with corresponding application system Apache, FTP, BIND, DHCP, IIS, oracle database management system, etc., every application software will produce numerous log information closely related to operational behavior.

3. The technologies of log collection

3.1 Log collection

Log collection is reliant on the infrastructure and services in the campus network. Common equipment is capable of realizing the system of log record, which means that the log can be recorded as required. Log collection can be mainly classified as single-machine deployment or distributed deployment. The centralized mode is widely adopted for the ease of log management. The distributed deployment of log collection can be seen in the figure 1.

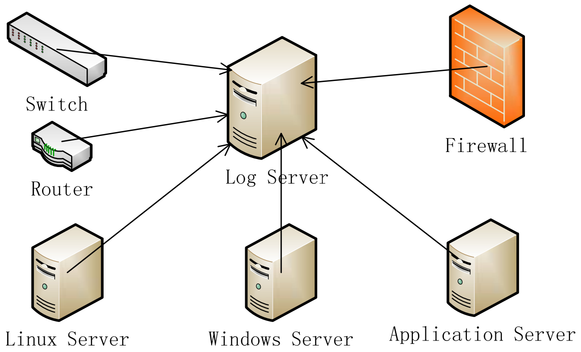

Figure 1 distributed deployment of log collection

3.2 The protocol of log source

Different log protocols must be adopted for different environments of log source. The mainstream protocols include Syslog, SNMP, Windows log system and special dedicated protocol. The methods of collecting log data in the network management include text-typed collection, Syslog collection, SNMP Trap collection and other collections (such as serial collection).

3.2.1 Syslog

As a standard log protocol, Syslog has been widely used in the computer system, especially for the network operation and security management. Syslog was first proposed by Eric Allman. It is a comprehensive log record system. It is mainly used to record the log information in the equipment, view the log record anytime and check the operation status of the system. The events recorded by Syslog protocol can be stored in different systems. They can be also transmitted between different equipment of Syslog protocol through the network. The most fundamental principle of Syslog is that the strict mutual coordination between sender and receiver of the event-related information is not required.

The protocol is realized in the log server software called Syslog. UDP514 port has been adopted. The receiver of log source will send the log information in the form of UDP data message. Syslog can realize the centralized management of the log. The log information sent from other equipment can be also written into the specified files or stored in the database for the sake of management and response analysis.

3.2.2 SNMP

SNMP (Simple Network Management Protocol) is proposed by IETF. It has become an effective standard protocol in the field of network management. SNMP has involved the manager, the agent of managed devices, management information bank and the object of management protocol. The manager is the host computer responsible for carrying out the management process and sending the management instruction. The agent of managed devices is the program running on the managed devices. It is responsible for executing the management instruction of the manager. The management information bank is comprised of various objects managed by the agent of managed devices. SNMP has provided four categories of management operations, namely get, get-next, set and trap. The topology can be created through the SNMP agent. The monitoring and management of managed devices can be realized through SNMP message.

3.2.3 Windows log files

Windows log system is the log recording mechanism in the Windows operating system. It has included the application log, system log and security log. The filename extension is .evt. It is stored under the file folder of %System Root%\system32\config by default. The special system authority must be needed for the modification. The third-party software can be used to send the log collected elsewhere for the ease of log processing.

4. The technologies of log storage

Storage mode of log mainly depends on preservation strategy of log. As log manager, log preservation strategy is often made based on storage type, size, cost, retrieval speed, filing and destroying time of log data. Storage mode of log first considers preservation time and space requirement, log is store in below 5 modes currently.

4.1 File format storage

Normally, log will be stored in the form of text, binary, compressed file. Text file is the most widely used format, with advantage of spending less CPU and i/o resource when writing file, ease of long time storage and management, convenience to read and use. Binary format is machine readable log file produced by application, needing special tool to read, not suitable for long team storage. Compressed file is a kind of log compression mechanism for realizing rapid log access and saving disk space, when compressing log file, choose suitable compression format standard and its support of multi-platform for convenient storage and use of log.

4.2 Database mode storage

The largest advantage of using database to store log lies in usability, rapid inquiry and retrieval with standard SQL, easy development of log inquiry and analysis tool with development languages. Main deficiency of database storing log is too high expenditure of database reading and writing brought by huge quantity of data, challenge of database storage optimization brought by big data; risk such as database failure and data loss, etc. In order to resolve problems brought by huge quantity of log data, we can preserve original log data in log source, store important log items into database to quickly analyze and use.

4.3 Hadoop storage

Hadoop is currently popular distributive computing framework, widely used in such fields as log analysis, data mining, etc. Hadoop system is typically a colony organized by commercial PCs operating Linux platform, and the colony consists of multiple nodes, of which at least one is master node and multiple are slave nodes. In order to expand level computing capacity and storage space of the system, nodes can be added to colony at any time based on actual situation. Hadoop system can make search request distribute at every colony node, realizing rapid lookup, processing and retrieval of result, when log data increases quickly, it can realize elasticity of system. Hadoop system had high error tolerance by copying data between colony nodes.

4.4 Elasticsearch storage

Elasticsearch is a Lucene based open source search engine, featuring stability, reliability and fastness, typically used in construction of large scale log storage and analysis system. Elasticsearch colony system can realize distributive real-time file storage, storing every field into the index to make it retrieved; has good horizontal expansion capacity, expanded to hundreds of servers to process structural or nonstructural data of PB level. Meanwhile, Elasticsearch supports plug-in mechanism including word, synchronization, Hadoop and visualization plug-ins, etc., convenient to log analysis and processing.

5. The technologies of log analysis

5.1 Traditional single-machine log analysis

In the case of small data scale, the log is mostly stored and analyzed on the single machine. Various system tools can be used for simple log analysis, such as awk, grep, sort and join under the Linux. If complicated logic for log analysis is required, various scripting languages must be used for analysis. For example, shell script, python and Perl should be used for programming and analysis. As the log data increases, the database should be used for log storage. SQL is capable of finishing most of statistical analysis in a simple and rapid manner. Moreover, the structural storage of the database can facilitate the mining of log data.

5.2 Large-scale distributed log analysis

When the log data increases constantly, there will be a more urgent need for the large-scale log analysis in terms of time and performance. The distributed technology is quite a good choice.

Hadoop is a distributed system under Apache foundation. It has consisted of Hadoop Distributed Files System (HDFS), MapReduce computing framework and HBase. After years of developemnt, Hadoop technology has become quite mature. The cluster comprised of hundreds of machines can operate in a stable manner. It can also support the data storage and processing above the level of PB. In most cases, Hadoop is used for log analysis. First of all, the log should be stored in HDFS. Some components must be selected for log analysis. HBase is a NoSQL distributed database adopting the column storage. It is capable of making simple K-V query and data analysis. MapReduce API can be used to write the program for log analysis. Hence, a deep understanding of MapReduce must be required. Hive is a sub-item under Hadoop item. It can provide the data for operating Hadoop by means of SQL and the codes for executing MapReduce. To conclude, it is a good choice from the angle of improving the performance and lowering the technical requirement.

6. Conclusion

In a word, united log storage platform can be built in college campus network by choosing optimized technical scheme. Research on log collection and analysis is necessary, especially for operation and maintenance, safety management of campus network, it has clear guiding role and important reference.

Acknowledgements

This research was financially supported by the natural research project of the Education Department of Sichuan Province (NO: 16ZA0361), this article is one of the phased objectives of the project.

References

[1] Jin Lei, Xie Li.Internet network security.Computer engineering and design 2003, 24(2): 19-22.
[2] Oliner A, Ganapathi A, Xu W. Advances and Challenges in Log Analysis. Communications of the Acm, 2012, 55(2): 55-61.
[3] Mauro D R, Schmidt K J. Essential SNMP,Second Edition, pp. 19-72, 2016.
[4] Peikari C, Chuvakin A. Security Warrior, pp. 161,2004.
[5] Schmidt, Kevin, Chuvakin, et al. Logging and Log Management, pp.79-91, 2012.
[6] Lonvick C. The BSD Syslog Protocol. RFC 3164 (Informational). 2001.
[7] Shvachko K, Kuang H, Radia S, et al. The Hadoop Distributed File System.MASS Storage Systems and Technologies. IEEE, 2010:1-10.
[8] Cheng Miao, Chen Huaping. Weblog Mining Based on Hadoop. Computer engineering, 2011, 37(11):37-39.
[9] Li Yang, LV Jiake. Research on Log Data Real-Time Based on Hadoop and Storm. Journal of Southwest China Normal University (Natural Science Edition) 2017, 42(4): 119-126.
[10] Bai, Jun. Feasibility analysis of big log data real time search based on Hbase and ElasticSearch. International Conference on Natural Computation IEEE, 2013: 1166-1170.
[11] Divya M S, Goyal S K. ElasticSearch An advanced and quick search technique to handle voluminous data. Compusoft International Journal of Advanced Computer Technology, 2013, 2(6).