Analysis of the Dynamics of Internet Threats for Corporate Network Web Service

Dmitry Kononov and Sergey Isaev

Institute of Computational Modelling of the Siberian Branch of the Russian Academy of Sciences, Akademgorodok 50/44, Krasnoyarsk, 660125, Russia

Abstract

Analyzing web service logs is an important task to ensure the uninterruptible functioning and security for computer systems. When implementing complicated software systems, it is necessary to pay special attention to collecting, storing, processing, and analyzing logs of various services to identify existing and potential security problems. This paper describes an approach to analyzing the dynamics of web services functioning over two years and identifying security risks, as well as impact of the COVID-19 pandemic on the use of Internet services. Recommendations are given to strengthen the protection of web services and reduce cybersecurity risks.

Keywords

Internet, security, web, threat, log, network, data analysis

1. Introduction

Modern information technologies are used in many areas of economy, including government management systems. The use of web technologies and web systems allows the provision of online services without the need to visit the organization, which is especially important in the case of global pandemics. Also, web services are used in corporate networks of various size, providing access to web mail, private clouds, and other online resources.

It should be noted that since web systems and web services use the Internet for their work, there are risks associated with information security. Ensuring information security is a complex task which includes a set of measures that must be taken to reduce the risks of threats. An important part is the analysis of the activity logs of web services, which allows detecting web attacks and optimizing hardware settings [1]. For an adequate assessment of the threat level, it is necessary to involve computer security experts [2]. In [3], it is shown that threats can increase when using various technologies for the development of web services. It is also necessary to analyze the activity of services to identify infrastructure weaknesses (CPU, memory, disk, and network operations) in order to reduce the consequences of increased loads, including hacker attacks. The paper [4] suggests proactive resource planning using the bandwidth load simulation technology. The analysis of the effectiveness of the protection tools should be made without side effects for the existing infrastructure [5].

2. Related works

Many works are devoted to analyzing logs of various services to identify security problems. In [6], statistical methods are used to analyze system logs to build a system for detecting hidden attacks on the network infrastructure. The authors in [7] use the graph theory to detect early attacks for various

SibDATA 2021: The 2nd Siberian Scientific Workshop on Data Analysis Technologies with Applications 2021, June 25, 2021, Krasnoyarsk, Russia
EMAIL: ddk@icm.krasn.ru (D. Kononov)
ORCID: 0000-0002-8757-5274 (D. Kononov); 0000-0002-6678-0084 (S. Isaev)

services. In [8], a regression analysis using the correlation between the elements of cloud service logs is proposed. Analyzing web server logs allows detecting a wide class of attacks including SQL injection. In similar studies [9], the authors use predefined rules to detect SQL injections and XSS attacks, which limits their use to certain types of attacks. A big threat to the functioning of web services is web spiders, which allow the automatic detection of system weaknesses [10]. As will be shown in this paper, web spiders cause the majority of errors in web services. Various methods are being developed to prevent automated scanning including real-time detection and response [11]. It should be noted that the COVID-19 pandemic has led to a change in the traffic patterns and usage profile of network and cloud infrastructure. The paper [12] analyzes the homogeneity of attacks on popular services during remote work in the COVID-19 pandemic, and identifies a list of countries which are the sources of attacks.

The existing works cover various aspects and methods for analyzing service logs but use short time intervals as data sources, which makes it difficult to assess the dynamics of the ongoing processes. In addition, the analysis is often made only at one level and using one data source, which does not allow assessing the reliability of the results obtained.

In this research, web services and traffic monitoring systems operating in the corporate network of the Krasnoyarsk Science Center (KSC SB RAS, Russia) are studied. The purpose is to analyze the functioning of web services in dynamics over 2 years, identify potential risks and threats, as well as to create recommendations for improving methods and means of ensuring the protection of Internet services. Another goal of this work is to assess the impact of the COVID-19 pandemic on the use of Internet services and their security.

In contrast to the existing studies, multiple data sources are used to extract web services data at the network and application layers of the OSI network model [13]. The analysis is carried out over large time intervals, which makes it possible to assess the dynamics of the web services behavior by hours, days, months, and years. In this paper, the authors consider *a potential attack* to be a request for a non-existent web service entry point or an unauthorized request for the existing entry point according to web traffic logs, and a request for a non-existent service according to Netflow IP traffic logs. This study continues our research on the security of Internet services in the corporate network [14].

3. Data sources

In this paper, we used the following data as data sources for 2019 and 2020: 1) Netflow IP traffic: more than 460 GB, more than 25 billion records; 2) logs from web-services: about 32 GB, more than 128 million records. The analysis was performed using the following software: UNIX CLI tools, GAccess, MaxMind, JSON tools, Python, FlowTools, Microsoft Excel.

4. IP Traffic analysis

To compare the level of activity of web service users, IP traffic data was analyzed using the HTTP and HTTPS protocols (fig. 1). A 7-day average was used to smooth out the activity peaks during the week.

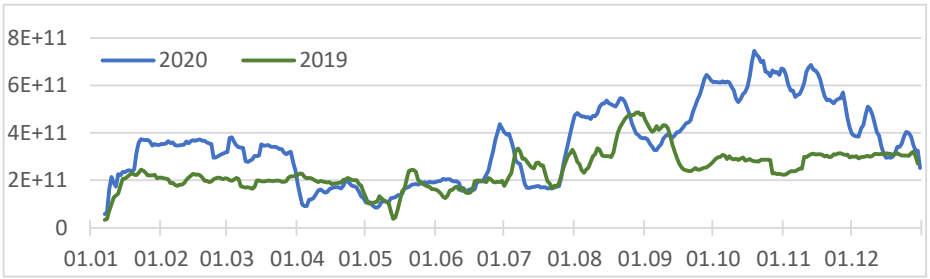

Figure 1: The daily amount of the incoming web services traffic in bytes

The analysis shows a general trend of increasing the activity of using web services: the average daily traffic for 2019 (224 GB) is 1.5 times lower than in 2020 (329 GB), and the correlation is weak (0.38). While in 2019 the activity increases quite smoothly throughout the year with dips during long weekends, in 2020 there is a sharp decrease in the activity by a factor of 2 at the end of March due to the introduction of lockdown and remote work during the COVID-19 pandemic. The activity returns to its previous levels only in the fall and decreases again by the end of the year against the background of the second wave of COVID-19. The analysis of the activity by days of the week (Fig. 2) shows that while the overall activity profile in 2020 remains the same (correlation 0.99), there is an approximately 10% increase in the weekend activity, which is likely due to the active use of remote workplaces. The comparative analysis of the use of the HTTP and HTTPS protocols shows an increase in the portion of the latter (from 86% to 91%), which reduces the level of cyber threats.

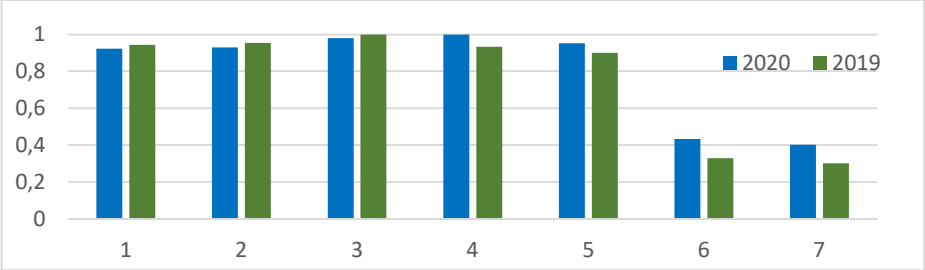

Figure 2: The ratio of the relative use of web services by days of the week (1 – Mon, 7 – Sun)

To analyze the use of web services, correlations for 2019 and 2020 of daily download traffic were calculated using NetFlow IP data (complete data) and web service logs (data from a part of hosts). As the proportion of the host traffic with the available activity logs increased from 30% to 48%, the correlation also increased, indicating that the data is correct, and that these sets can be used together for detailed analysis.

The web usage activity profiles by days of the week based on the IP traffic show little change (correlation 0.84).

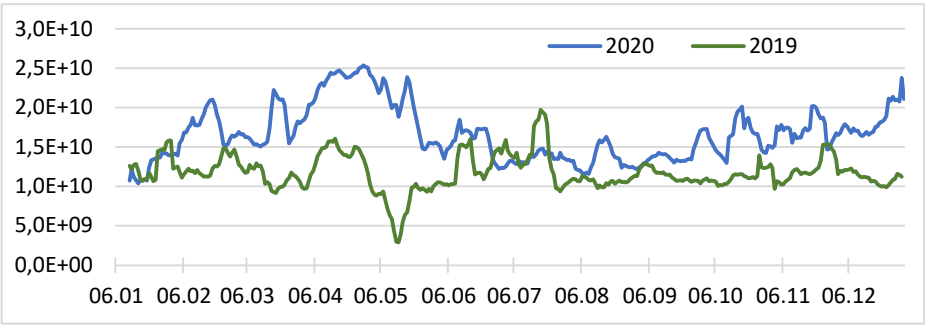

Figure 3: The daily response traffic of the web services in bytes for 2019 and 2020

The annual analysis of the use of the web services of KSC SB RAS in Figure 3 shows a significant increase in the use of its own web services during the transition to remote work in the spring and autumn of 2020.

The analysis of access attempts to non-existent web services of the KSC SB RAS network using the HTTP and HTTPS protocols for 2019 and 2020 was made (Fig. 4). During the analyzed period, there was a smooth increase in access attempts using the HTTPS protocol, which is consistent with the general trends in the use of web services. In 2020, the daily number of attacks increased 1.5 times for HTTP and 2.5 times for HTTPS.

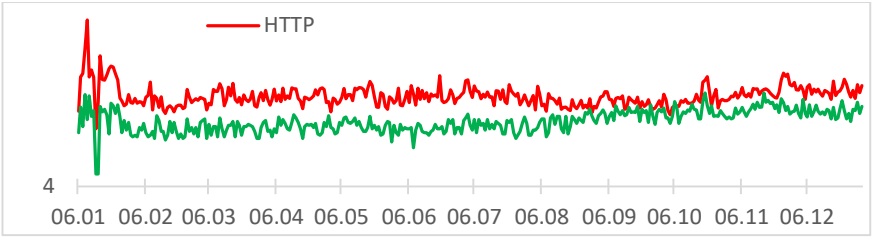

Figure 4: The number of attempts to access the web services in 2020 by days

The standard deviation σ and variation of c_v were calculated for the obtained aggregated data sets:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x^i - \bar{x})^2}{(N-1)}}; c_v = \frac{\sigma}{\bar{x}} \quad (1)$$

For the access attempts via the HTTP protocol, the variation coefficients were 1.76 and 0.65 for 2019 and 2020, respectively, and for the HTTPS protocol: 0.61 and 0.35. Thus, we can conclude that the number of intensive attacks decreased in 2020 as compared to 2019, while the intensity of HTTP attacks remained approximately twice as high. The calculated variation coefficient parameters allow us to build attack detection models, as well as to simulate the normal operation of web services.

We also analyzed the IP traffic data to identify the dynamics of changes in the popularity of individual Internet services (Table 1).

Table 1
Rating of threats to Internet services

Rank in 2020	Protocol name/threat	Port	Proportion in Top 50	Rank change
1	Telnet protocol/Trojan	23	18.3%	0
2	Microsoft-DS Active Directory/threat	445	18.0%	0
3	Microsoft-SQL-Server/threat	1433	7.2%	0
4	Session Initiation Protocol (SIP)	5060	5.9%	7
5	Secure Shell (SSH) Protocol/Trojan	22	4.4%	–3
6	Hypertext Transfer Protocol (HTTP)/threat	80	4.2%	–2
7	Microsoft Terminal Server (RDP)	3389	3.3%	3
8	HTTP Alternate/threat	8080	2.7%	–1
9	Torpark - Onion routing/threat	81	2.5%	–1
10	Personal Agent/threat	5555	2.1%	2
11	iTunes Radio streams, JSON RPC default port	8545	2.0%	–2
12	Hypertext Transfer Protocol over SSL / threat	443	1.9%	3
13	iTunes Radio streams, MikroTik Winbox	8291	1.4%	–8
14	Network Time Protocol / threat	123	1.3%	5
15	Asterisk Manager Interface (VoIP)	5038	1.3%	1

It should be noted that, in general, the set of services used in most attacks and disguised as malware remained unchanged: Telnet and Microsoft-DS Active Directory protocols are by far the leading ones, and can be used to access data on a remote computer. The following protocols significantly changed their position in the rating: Session Initiation Protocol (SIP) – plus 7 positions and iTunes Radio streams – minus 8 positions. The increasing number of the SIP attacks can be explained by the popularity of video conferencing during the COVID-19 pandemic. The fifth position of one of the most attacked, according to security experts, is the SSH protocol which can be explained by an efficiently functioning system for preventing password guessing and blocking hosts on the corporate network edge router.

5. WWW data analysis

This paper also analyzes the activity logs of web resources for 2019 and 2020. The analysis shows the presence of requests and frequency of errors by days of the week, and hours of the day, as well as an increase in the number of requests from 52.5 million (2019) to 76 million (2020) due to the development of web services and an increase in their audience.

In this work, all the web service requests are divided into two groups: legitimate and erroneous according to the HTTP protocol specification [15]. Legitimate requests are executed by web applications and web services in normal mode without causing errors (response code 1XX, 2XX, 3XX). Erroneous requests (or errors), in turn, are divided into two groups: client errors which occur due to an incorrect web client (response code 4XX), and server errors which occur on the server side due to an incorrect client request or internal errors (response code 5XX).

Table 2
Ranking of the countries by the requests in 2019 and 2020

Rank	2019	%	2020	%	Rank change
1	Russia	81.97	Russia	83.53	0
2	United States	7.14	United States	5.57	0
3	Germany	1.91	France	1.89	2
4	Ukraine	1.30	Germany	1.58	–1
5	France	1.19	Ukraine	1.00	–1
6	China	1.08	Netherlands	0.82	3
7	United Kingdom	0.86	Canada	0.77	3
8	Ireland	0.81	United Kingdom	0.74	–1
9	Netherlands	0.56	Unknown	0.66	3
10	Canada	0.28	CN China	0.52	–4

As shown in Table 2, in 2019 and 2020, the first two places in the number of requests belong to Russia and the United States. Russia accounts for more than 80% of all the requests. In 2020, France came third, displacing Germany and Ukraine by one position. The Netherlands and Canada moved 3 positions up. It is noteworthy that the proportion of requests from China decreased by a half, and the position of the country dropped by 4 points. The high positions of the US, France, and Germany can be explained by the presence of many hosting providers in these countries, which are used by web spider owners to scan hosts on the Internet. The most popular browsers are: Chrome – 40% in 2019 and 43% in 2020, Firefox – 14% and 11%, respectively. Web spiders account for 7% of all the requests in 2019 and 6% in 2020 with an error rate of 62% in 2019 and 58% in 2020.

Figure 5 shows the trend graphs of the number of requests in 2019 and 2020. The analysis shows the dependence of the number of requests on holidays when the number of requests decreases. The activity of requests remains high from Monday to Friday, and on Saturday and Sunday there is a decrease of up to 40%, indicating the use of web services mainly on weekdays.

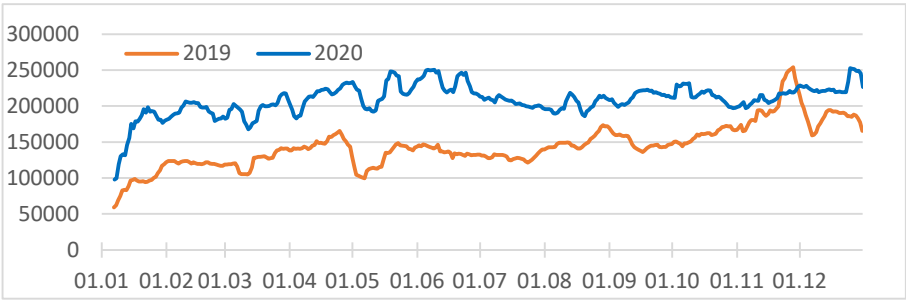

Figure 5: The daily number of the requests for web services in 2019 and 2020

Figure 6 shows trend graphs of the number of errors in 2019 and 2020. The peak values on the graphs indicate the presence of abnormal activity. As mentioned above, most of the errors are caused by the activity of web spiders, which can be divided into three groups: search, research, and malicious. Search spiders belong to search engines (Google, Bing, Yandex) and scan web resources to include pages in search results. Due to the improper configuration of web resources, search engine spiders can follow links that are not public, causing errors. Research spiders belong to public, academic, or commercial organizations which collect data and monitor the Internet. Malicious spiders belong to criminal groups and scan for the known vulnerabilities in web resources, and if they are present, the spiders perform attacks in the form of automatic exploitation of vulnerabilities with the execution of a malicious code on the server. As a rule, this scanning is performed for popular open source content management systems (CMS), online stores, forums, and Internet of Things (IoT) devices.

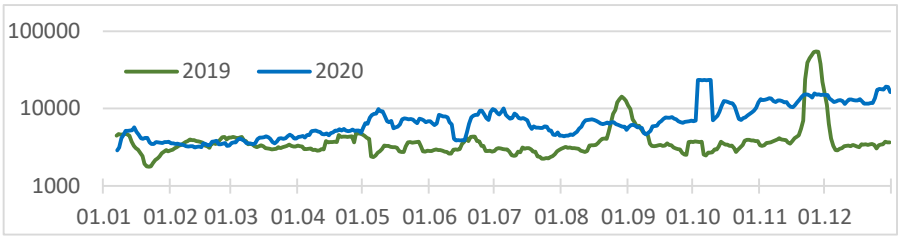

Figure 6: The daily number of errors for web services in 2019 and 2020

Figure 7 shows a graph of the number of requests and errors by hours in 2019 and 2020. As you can see from the graphs, the number of requests per hour increases proportionally due to the increase in the total annual number of requests. The highest activity is observed during working hours from 9:00 to 18:00 (a small dip can be seen at lunchtime at 13:00), and in the evening the activity decreases until 22:00. In the error graph, one can see that for both years there is a rather high number of errors at night, which indicates the presence of constant activity of web spiders and bots performing scanning of web resources. This constant activity remains at about the same level both in 2019 and 2020.

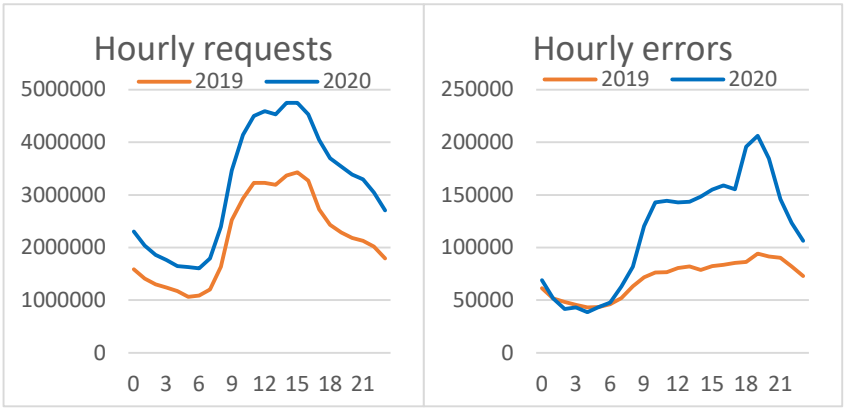

Figure 7: The number of requests and errors by hours

The correlation coefficient between the requests and the errors was calculated: 0.35 (2019), and 0.38 (2020), indicating a weak relationship between the requests and the errors due to the incorrect operation of web services. However, most of the errors are caused by scans and web spider attacks. The correlation coefficient for the requests in 2019 and 2020 is 0.99, and for the errors it is 0.96. As one can see from the graphs, the profile of requests and errors persists in 2019 and 2020. The average number of the errors in 2020 increased by more than 60%. The variation coefficient in 2019 was 1.99, and in 2020 it was 1.01, indicating a decrease in the number of intensive attacks on web resources by about 2 times. This agrees with the above analysis of scans on the HTTP (80) and HTTPS (443) protocols.

6. Recommendations

After the analysis, the following recommendations were formulated to strengthen the security of Internet services. (1) We recommend adding TCP ports from Table 1 to the intrusion detection system and using the calculated standard deviation parameters for different services to distinguish the background port scanning activity from targeted attacks. (2) Web moderators have to regularly update web resources which use popular systems: content management systems (CMS), forums, third-party modules. The study of the constant malicious activity of web spiders shows an increased interest towards vulnerabilities in old versions of these systems. (3) It is necessary to integrate automatic downloading of malicious IP address lists obtained from web resource logs into the threat blocking system on the edge router. This measure will allow blocking hosts not only for web services, but also for the entire range of IP addresses of the autonomous system (AS) when a malicious activity is detected. (4) The most effective way to prevent security threats is to whitelist access to the administrative interfaces of the systems using IP addresses and/or VPN services.

7. Conclusion

In this paper, we analyzed the dynamics of using web-services of the corporate network of Krasnoyarsk Science Center (Russia). The main parameters of the web traffic are revealed; the sources of Internet threats and dynamics of their behavior over 2 years are clarified. The calculated parameters of the distributions allow building models for detecting attacks, as well as for simulating the normal operation mode of the web services. Based on the results, we formulated recommendations to strengthen the security protection of web services, which should minimize cybersecurity risks.

8. References

[1] M. Landauer, F. Skopik, M.W urzenberger, A.Rauber, System log clustering approaches for cyber security applications: A survey, Computers & Security 92 (2020). doi: 10.1016/j.cose.2020.101739.

[2] S. Khan, S. Parkinson, Discovering and utilising expert knowledge from security event logs, Journal of Information Security and Applications 48 (2019). doi: 10.1016/j.jisa.2019.102375.

[3] F. Yilmaz, M. Sridhar, A. Mohanty, etc., A fine-grained classification and security analysis of web-based virtual machine vulnerabilities, Computers & Security 105 (2021). doi: 10.1016/j.cose.2021.102246.

[4] J. Hu, L.Huang, T. Sun, et al., Proactive planning of bandwidth resource using simulation-based what-if predictions for Web services in the cloud, Frontiers of Computer Science 15 (2021) 151201.

[5] M. Wurzenberger, F. Skopik, G.S ettanni, W. Scherrer, Complex log file synthesis for rapid sandbox-benchmarking of security- and computer network analysis tools, Information Systems 60 (2016) 13–33.

[6] Z. Gu, K. Pei, Q. Wang, et al., LEAPS: Detecting camouflaged attacks with statistical learning guided by program analysis, in: 2015 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, 2015, pp. 57–68.

[7] A. Oprea, Z. Li, T. Yen, S. H. Chin, S.Alrwais, Detection of Early-Stage Enterprise Infection by Mining Large-Scale Log Data, in: 2015 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, 2015, pp. 45–56.

[8] M. Farshchi, J.-G. Schneider, I.Weber, J. Grundy, Metric selection and anomaly detection for cloud operations using log and metric correlation analysis, Journal of Systems and Software 137 (2018) 531–549.

[9] M. B. Seyyar, F. O. Catak, E. Gul, Detection of attack-targeted scans from the Apache HTTP Server access logs, Applied Computing and Informatics 14 (2018) 28–36.

[10] T. Tanaka, H. Niibori, S. Li, et al., Bot Detection Model using User Agent and User Behavior for Web Log Analysis, Procedia Computer Science 176 (2020) 1621–1625.

[11] G. Suchacka, A. Cabri, S. Rovetta, F. Masulli, Efficient on-the-fly Web bot detection, Knowledge-Based Systems 223 (2021).
[12] C. Kelly, N. Pitropakis, A. Mylonas, S. McKeown, W.J. Buchanan, A Comparative Analysis of Honeypots on Different Cloud Platforms, Sensors 21 (2021) 2433.
[13] ISO/IEC 7498-1:1994. Open Systems Interconnection: The Basic Model, URL: https://www.iso.org/ru/standard/20269.html.
[14] S. Isaev, D. Kononov, A. Malyshev, Analysis of Internet Service Log Data to Assess the Level of Cyber-threats in the Corporate Network, in: CEUR Workshop Proceedings, volume 2727, 2020, pp. 16–24.
[15] RFC 2068. HTTP/1.1, URL: https://tools.ietf.org/html/rfc2068.