Malaysian Journal of Computing, 3 (1): 37–53, 2018

eISSN: 2600-8238 online

VISUALIZING WEB SERVER LOGS INSIGHTS WITH ELASTIC STACK– A CASE STUDY OF UMMAIL'S ACCESS LOGS

Harni Yusnidar Muhammad[1] and Jasni Mohamad Zain[2]

Faculty of Computer and Mathematical Sciences, UiTM Shah Alam, Selangor, Malaysia
[1]harniyusnidar@gmail.com, [2]jasni@tmsk.uitm.edu.my

ABSTRACT

One of the most significant information resources that often overlooked and it is mostly owned by the modern organization today is logs data. Likewise, logs data analytics is practised in many industries for different purposes, including website/system performance improvement, web development, information architecture, web-based campaigns/programs, network traffic monitoring, e-commerce optimization, marketing/advertising, etc. Many tools or approaches are available for this purpose, some are proprietary and some are open source. Studying the nature of these tools in finding the suitable and the right log analyzer in order to perform log analytics economically, efficiently and effectively will give advantages to the organization towards utilizing the primary source of information for identifying the system threats and problems that occur in the system at any time through Visualizing Insights of source using Elastic Stack. These kinds of threats and problems which existed in the system can be identified by analyzing the log file and finding the patterns for possible suspicious behaviour. A case study of UMMAIL's access logs is proposed to visualise web server logs. The system administrator's concern can then be furnished with an appropriate infographics representation regarding these security threats and problems in the system, which are generated after the log files, are analysed. Based on this signs the administrator can take appropriate actions.

Keywords*: Data analytics and visualization, Visualizing Insights, infographics, data exploration, Elastic stack*

1. Introduction

Today, in many circles of human activities, massive sets of data are collected and stored on a daily basis. Virtually most activities using digital device generates streams of data which increases at a faster rate that has surrounded us. Normally in many situations, the filtration of data for later uses is not implemented as early as storing the data (Keim Daniel & Mansmann, 2010). This situation is majorly encouraged over modernized and technologies advancement in digital worlds such as the Cloud Computing, Machine Learning, Augmented Reality (AR) and virtual reality (VR), Automation, Big Data, Physical-Digital Integration, Internet of Things and Smart Home, and Everything-On-Demand.

The basic approach to discovering knowledge and behaviour of Web users logged in web server files is acknowledged as log analysis or log normalization. Log analysis also can be achieved via data mining refer to the other term uses by data scientist for data normalization. From the IT people's point of view, it is imperative to understand the way user use a site, navigate through online web pages and make successful processes, which may lead to better organizational service and efficient business decisions. There are many activities

involve within log analysis processes such as collect, parse, aggregate, filter, store, query, visualize, alert, summary & conclude, analyze, and perhaps act on them accordingly.

There is no doubt that in present digital age, analytics tools are still a challenge for many of us. Most analytics solutions offer complexity and rigidity, according to surveys conducted by behavioural analytics provider, Interana, there are at least seven indications of the complexity in analytics spaces including too much coding, hard to use cause by the complicated interface, slowness in query speed, not flexible enough, too expensive, can't be scaled and only a few among employees are armed with the tools (Maria Minsker, 2017). Paper & Iyer (2017) state that, "there are various data analytics tools available for any size and sectors of organizations including retailing, healthcare, education, etc. being used; also can be tricky when there are so many options". This statement considered an important point about challenges when need to decide which analytics tools that are really suited to adopt since there are too many visualization tools exist either open source or enterprise versions, but most of these tools do not provide a comprehensive functionality to produce the desired new pattern.

As a rule, most moderate-sized organizations will maintain a daily log size below 500 MB which requires the administrators/operators to play around with the existing log's rule or make housekeeping like transferring/removing, backup or deleting the log files on a scheduled basis. At this point, by overlooks to the access logs of University of Malaya Email system (UMMAIL), obviously, the analytical tool is needed to "uncover vitally important information from basic to advanced knowledge". Indeed, it is highly likely that UMMAIL's administrator needs an analytic tool as no such tool in place to monitor and advancing the precautionary measure of the system security vulnerabilities especially identifying a higher number of 404's error pages, how many suspicious 'Sign In' authentication attempts for example logins at strange hours, obvious scanning activities, and trials of retrieving the Forgot Password facility by non-provisioned users (Interview Transcript, 2017).

For this reason, University of Malaya Email (UMMAIL) web server logs are used as the main source of information in this study. Here efforts have been made to find the suitable logs analyzer tools to perform log analysis efficiently, effectively and economically. Initially, the study begins by exploring various text-based visualization tools as an alternative solution to the problems mentioned earlier. Henceforth, having considered many facets of robust analytics tool including capabilities in capturing, parsing, storing, searching, analyzing, visualizing, sharing, and transferring make it imperative to visualize insightful information about web security vulnerabilities from enormous log data into graphical representations that assisting in knowledge oriented decision making.

Visualization is more than just a static graphical form, it is solely facilitating in communicating with information data, understanding via presentation, discovering through exploration furthermore influencing for awareness in order to gain adequate responses (Bishop, Pettit, Sheth, & Sharma, 2013). Most importantly, McInerny et al. (2014) had emphasized from a user perspective, the positiveness of visualization expressed in static quantifiable information, diagrams & pictures, maps or interactive technology; then producible through digital media (printer) should promote better understanding and finally aid in making a data-informed decision.

The remaining of this paper is organized as follows: in section 2, present related work which concern web Server Log Files, Comparative Study of Log Analysis Domains and Elastic Stack Technology a Worthwhile Solutions. In section 3 discuss research framework and its related tasks. The developed framework and implemented algorithm are presented in section 4 with reported results.

2. Literature Review

2.1 Web Server Log Files

The Web server log files refer to log files storing the user's activities including success, errors, and lack of response according to web access and reside on the web server. Web servers are the richest and the most common source of data having simple plain text allowing administrators to characterize the audience and the pattern of their server usage using log analyzer tool. Dynamically, web servers generate and update four types of log files (Al-Asadi & Obaid, 2016; Goel & Jha, 2013) listed in Table 1:

Table 1. Types of Web Server Logs

Types of log files	**Actions**	**Format**	**Extracted Knowledge**
Access log	Records all usersrequest processed by server and information about users	[Wed Oct 11 14:32:52 2000] [error] [Client 127.0.0.1] client denied by server configuration: /export/home/live/ap/htdocs/test	• Users' profiles • Frequent patterns • Bandwidth usage
Error log	List of errors for users request made by server	[Wed Oct 11 14:32:52 2000] [error] [Client 127.0.0.1] client denied by server configuration: /export/home/live/ap/htdocs/test	• Types of errors • Generated errors IP address. • Date and time of error occurred
Agent log	Records user browsers and browsers version	"Mozilla/4.0 (compatible; MSIE 4.01; Windows NT)"	• Agent version • Operating system used
Referrer log	Records information about link and redirects visitor to Site	"http://www.google.com/search?q= keyword", "/page.html"	• Browser used • Keywords. • Redirect link content

2.2 Comparative Study of Log Analysis Domains

Since the existence of the first computers scientist and engineers, application logging and system diagnostic become a popular field for many researchers to dive into the details of log analysis and visualization. This section provides the overview with some of the most widely used tools on log analysis domains. There are a lot of alternatives tools for data visualization, for example, Graylog2, Nxlog, Octopussy, Logscape, ELSA, LOGanalyzer, Logalyzer, Logwatcher, logHound, log report, Logsurfer, PHP-Syslog-NG, etc. compared to Splunk that have been used for a long time in analyzing log data (Smith, 2015). In fact, these tools are just some of them with several drawbacks where some wouldn't work, slow, had a terrible user interface and most of them had a similar relational database backend for storage and retrieval like MYSQL or PostgreSQL.

In another work, Jayathilake (2012) has presented an in-depth analysis of current log analysis domain. The log management tools should be empowering customers via the true of log information. In many situations, log files always been referred in troubleshooting for investigation. However, most log analysis tools still have a lot to deliver to proven its capabilities in the field of security and evaluation. The study also admitted several log analysis challenges, as well as lack of standards, support of structured analysis, log file format, evolve with the product, amount of logged information, and consistency issues. Some of the works described in this research provide a list of commercial log analysis tools namely

Splunk, LogRhythm, ArcSight Logger, Loggly, Loglogic, AWStats and SecureVue highlighting their key features accordingly. Henceforth, M. Fedorov et al. (2017) and Chainourov (2017) have mainly concentrated on Splunk for the control system and management.

In the research carried out by Agrawal and Makwana (2015), there is various log management and analysis tools, the focus of open source log analysis can provide full features and reliability in a more affordable way. A comparative study randomly lists some common log management tools and features of Splunk, Graylog2 and LOGalyze. On the contrary, a study of logs analyzer by Fuente et al. (2015), aim to unify and manage data collection by providing a comprehensive study on Logstash and Fluentd. Comparison of both tools in terms of features and capabilities tells that Fluentd insists on simplicity, versatility and robustness whereas Logstash focuses on flexibility and interoperability.

Dealing with a large volume of computer-generated logs involves log collection, centralize aggregation, long-term retention, log analysis, log searching and reporting. However, most of the log management products require technical expertise to process and interpret logs data. A study of performance testing was done by comparing three log analytics systems namely Graylog2, ELK, and ELSA (Prakash & Patel, 2016). In the same way, various web log analyzer tools namely Webalizer, Piwik, Open Web Analytics, Deep Log Analyzer, Fire Stats, Go Access, Web Forensic, AW Log Analyzer, WebLog Expert, and Google Analytic were listed based on its core features such as languages, current stable version, specialization, strength, including some criticism (Brian Jackson, 2017; Čegan & Filip, 2017; Kumar & Thakur, 2017; Shakti & Garg, 2017; Valency Networks, 2016).

2.3 Elastic (ELK) Stack Technology a Worthwhile Solutions

Obviously, there is a large set of weblog analyzer software are available in the market. Among this large set of available tools, it is a time consuming and tedious task to find the suitable tools to perform the analysis economically, efficiently and effectively. To make a proper choice of infrastructure, an extensive investigation reported in (Bodó & Kouˇ, 2013; Fuente et al., 2015; Hasani, Jakimovski, Kon-popovska, & Velinov, 2015; Iuhasz, Pop, & Drăgan, 2016; Koga & Almeida, 2016; Lu et al., 2017; Mateik et al., 2017; Mikula, Adamová, Adam, Chudoba, & Švec, 2016; Mitra, 2016; Prakash & Patel, 2016) was presented. In all papers, presents the prominent use cases of Elastic Stack among 10 relevant articles published from 2013 to 2017 has identified 10 categories in capturing of relevant knowledge from logs analysis using Elastic stack technologies: geo-identification, SQL Queries analytics, monitoring mixed-language application, monitoring multiple big data frameworks, file transfer, log management, Elastic stack versus Splunk, grid site monitoring, network security, and satellite data analytics (Jayathilake, 2012).

The flexibility comes into existence to modify the infrastructure when needed, by adding various other components like Hadoop or scale up/down by adding (duplicate, triplicate, etc.) some of the existing components based on adjusted proposed infrastructure. The ELK Stack is the combination of Elasticsearch, Logstash, and Kibana that is used specifically in log analytics. Logstash ships log data to Elasticsearch, which indexes the information in a searchable data store. Kibana then takes the data store and shows the information in a graphical format for log analysis. Study and experiments are motivated by the need to use such technologies and infrastructures for analysis of web server log have contributed to the realization of this research project.

3. Research Methodology

General Research Framework involved six phases starts with Data Analysis which containing the stages of Information Gathering, Identify Problem Statement and Defined Project Objectives. The purpose of this stage is for gathering all related information about the research. The second phase is Requirement Analysis which might involve system and data

source selection that is used for this research. The third phase is the Experimental Design where the overall process of the research will be designed. Followed by the fourth stage is the Implementation Phase where the installation & configuration and development of the proposed solution require a sequence of steps to be completed. The fifth stage, the Result Analysis phase is to elaborate the results and findings of the research. Finally, the Documentation Phase completed the overall process of research methodology. Completing these phases accordingly will ensure the smoothness of research being done.

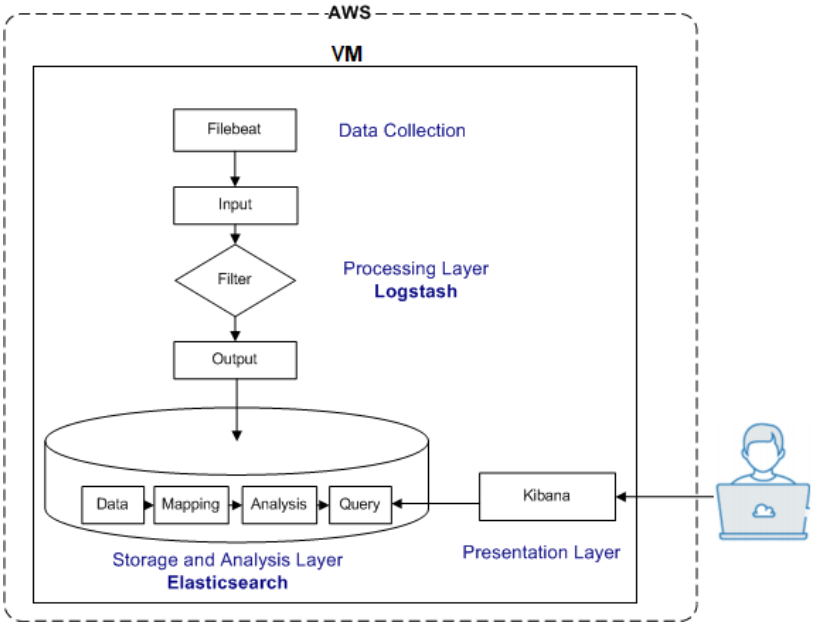

Figure 1. Conceptual Design

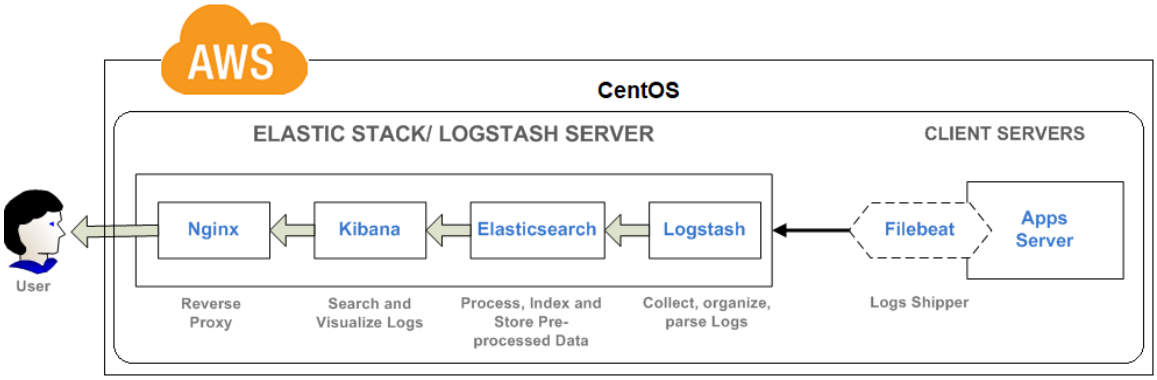

Figure 2. Infrastructure Design

Logstash: Responsible for collecting, organizing and parsing the incoming logs data into the system
Elasticsearch: The place where all of the outputs (logs) in diverse format from Logstash will be stored **Kibana**: Web-based graphical user interface (GUI) for analytics and visualizing logs

4. Results and Analysis

Several significant queries related to General statistics, Activity statistics, Access statistics, Status statistics and Browsers and Operating System statistics needs to be created to be used later for creating Visualize/Graphs for further analysis of indexed log data in the Elastic stack.

General statistics

In this section, the researcher gets the general information of web usage details that includes hits, visitors (Total Unique IPs), and volume (total bytes). Table 3 enlists all the general information related to a system.

Table 3. General statistics obtained after analyzing web logs

Summary	
Hits	
Page Views	4, 226, 683
Visitors	
Total Unique IPs	45, 173
Volume	
Total Bytes	620,261

Elastic stack also provides another feature of statistics according to hourly and daily basis. In addition to the table report, it also provides a graphical chart that helps determine at which hour system was receiving maximum hits of that day along. This information will give a clear picture to the system owner to provide better service by taking an appropriate action such as increasing the number of servers, avoiding any interruption during that period, planning for system maintenance and so on.

Table 4. Hourly Activity Statistics

@timestamp per 30 minutes	No. Accessed
6:30	183
7:00	634
7:30	839
8:00	1,260
8:30	1,202
9:00	2,872
9:30	3,903
10:00	2,876
10:30	2,564
11:00	4,148
11:30	2,586
12:00	3,582
12:30	2,030
13:00	1,550
13:30	1,741
14:00	3,211
14:30	2,789
15:00	3,073
15:30	3,299

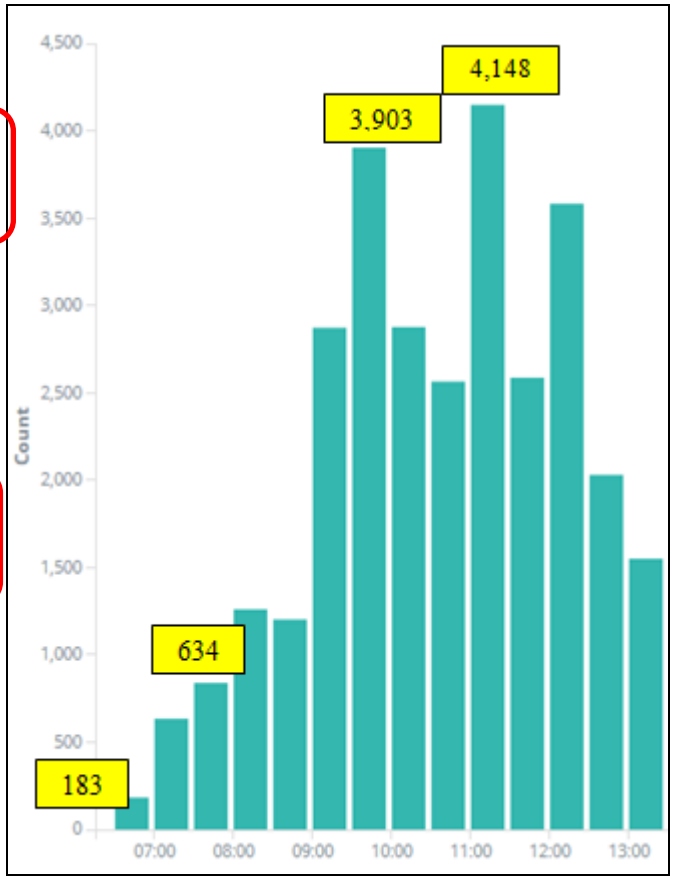

Figure 4. Bar Chart – Hourly Activity Statistics

According to the Table 4 and Figure 4 above, the system is hit a maximum at 10:00 hrs and 11:00 hrs and is least accessed at 6:30 hrs and 8:00 hrs.

Access statistics

Starts with the basic statistics such as access statistic where not only determine the page with maximum hits fairly providing the navigational behaviour of users and thus aid in planning to restructure the website to suit the increasing interests of users. Access statistic reveals which site/page have accessed the maximum amongst all the possible site/page available on a website. The Kibana were listings four types of PHP pages that are accessed by UMMAIL users, um_gvalidate.php, mail_script/google.php, mail_script/process_response.php, and logout.php dated on Dec 5th 2017.

Figure 5 shows four of PHP pages with two of the pages are mail_script/google.php and mail_script/process_response.php each contributed almost to the same percentage approximately 41%, followed by um_gvalidate.php about 37.30% and the last is logout.php around 4%. Assumptions can be made from the small portion of logout.php where on the Dec 5th 2017, the UMMAIL system was accessed by many users and most of them are not practising the "logged out" from the system. Throwing them into the "logged out" screen right away seems like a great idea in theory as it helps the user reinforce their memory and as a reminder to them not to be the target for hackers, thus giving the impression that something went wrong with their UMMAIL account. But this precautionary measure can be annoying to most of UMMAIL users.

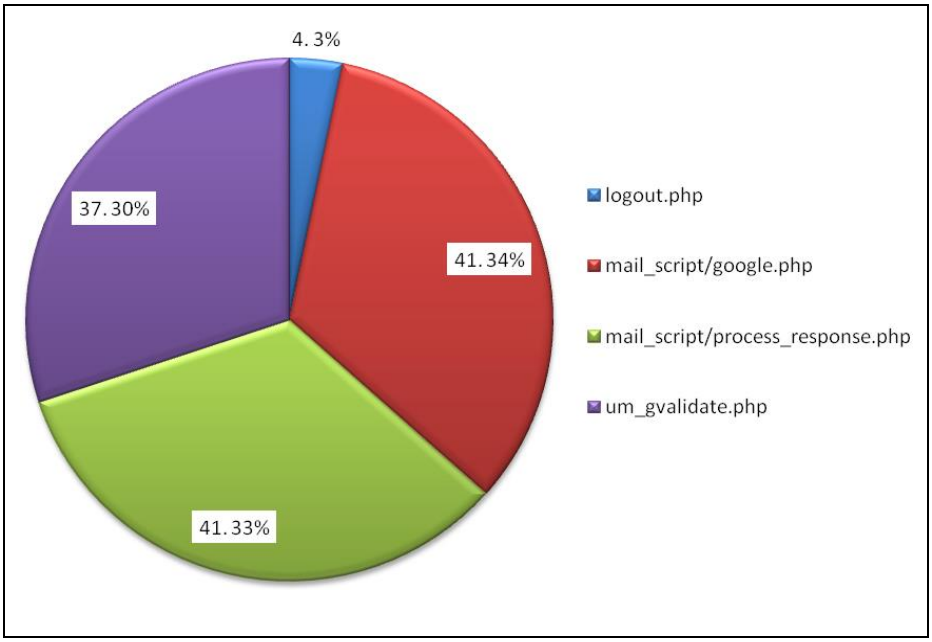

Figure 5. Pie Chart – The Top Accessed *.PHP Pages Statistics

(i) Hit Access on the Map Statistics

One of the most popular Filters available in the Elastic stack is GeoIP. The GeoIP uses a client IP address to detect the location including the country code and name, region, area code, city, zip code, and latitude & longitude to be drawn on the map representation as depicted in Figure 6. In general, GeoIP information will give the rough concept of the geographical location of system's users to an administrator. Always put in mind, client IP address can be enterprise IP and just endpoints in the network edge.

In this case, the result obtained from this research showing the location of users accessing the system from all around the world. Considering the email system of the University of Malaya is accessible to UM staff only, so the distribution of hit/access against the system can be seen on the map by location on average is from Malaysia with the maximum hits of 51K to 64K.

However, there is news about Google mail and Microsoft Online Services blocked/banned in certain countries such announced by Iran's telecommunications agency that it would be suspending Google's email services permanently, saying it would roll out its own national email service. This issue also happened to China and Burma. Besides, the United States also banned countries such Cuba, Syria, Iran, Sudan and North Korea from accessing Google mail and Microsoft Online Services which means that our staff from US embargoed countries may not be able to access UMMAIL from their home country. Though, the front page of UM email system still can be accessed but not the inbox at those restricted countries.

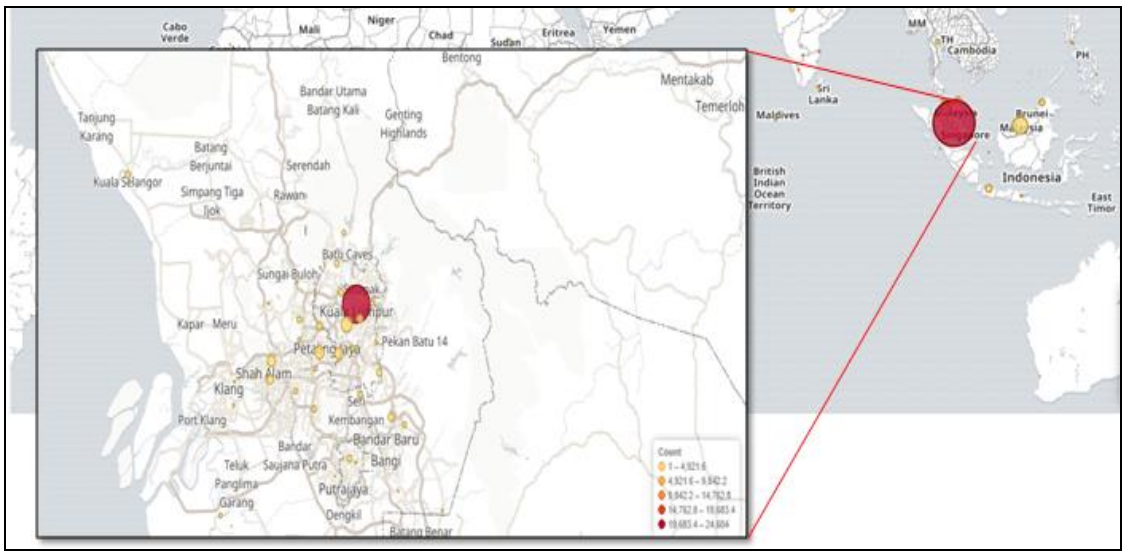

Figure 6. Map – Distribution of user's geolocation

(ii) Client Access by Remote IP

This part of the analysis can be considered the most important as it not only ascertains which page is hit the maximum number of times rather it also provides an idea of the navigational behavior of users and thus help in restructuring the website to suit the increasing demands of system's users. Figure 7 show which page was accessed the most amongst the web pages available in the system.

Among top 5 URLs/pages that the most accessed, the focus on the um_gvalidate.php page because the verification of client needs to be done here before the user can access the email account. Uncertainties arise on this page as there are many attempts to verify the "Login/Sign In" process by the same IP address. Here, the number of attempts per client according to IP address is 192.168.200.1 (4,370), followed by IP 219.93.25.66 (1,869), then IP 103.18.2.229 (1,476), IP 103.18.2.238 (1,205), and lastly IP address 10.34.100.247 (774). In this case, two possibilities have taken place, firstly users keep trying input incorrect username/password or secondly, someone is trying to hack into the system.

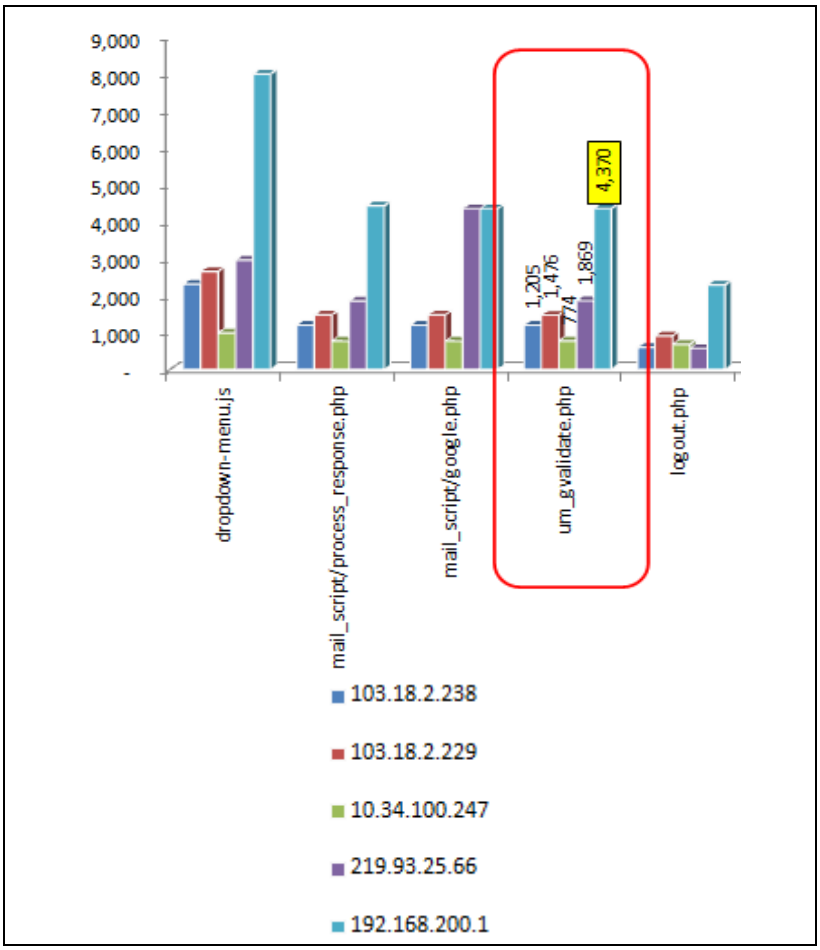

Figure 7. Top 5 URLs/pages

Status statistics

The status code that the server sends back to the client is known as HTTP response/status code. In Figure 8, showing the response code returns to the client while accessing the UMMAIL system over a month (December 1st, 2017 to December 27th, 2017). Moreover, the simple analysis can be made from the figure above, where system administrator knows the system was behaving normally since no server error was reported based on the following codes 200-OK, 206-Partial Content, 302-Found, 404-Not Found and 416-Unsupported Media Type.

There are 51% of users request is positively answered by the server denoted by response code '200-OK' while 18% received '302-Found' response code. However, 32% of the response represents the '404-Not Found' which need to fix. The rest of the response code denoted with '206-Partial Content' and '416-Unsupported Media Type' is considered not significant for this time because it returns 0%. Deeper analysis regarding 4xx or 5xx response code will be described in subsection 4.2. The details on the analysis of both client and server errors response code are obtained after binding the above errors with specified remote IP addresses.

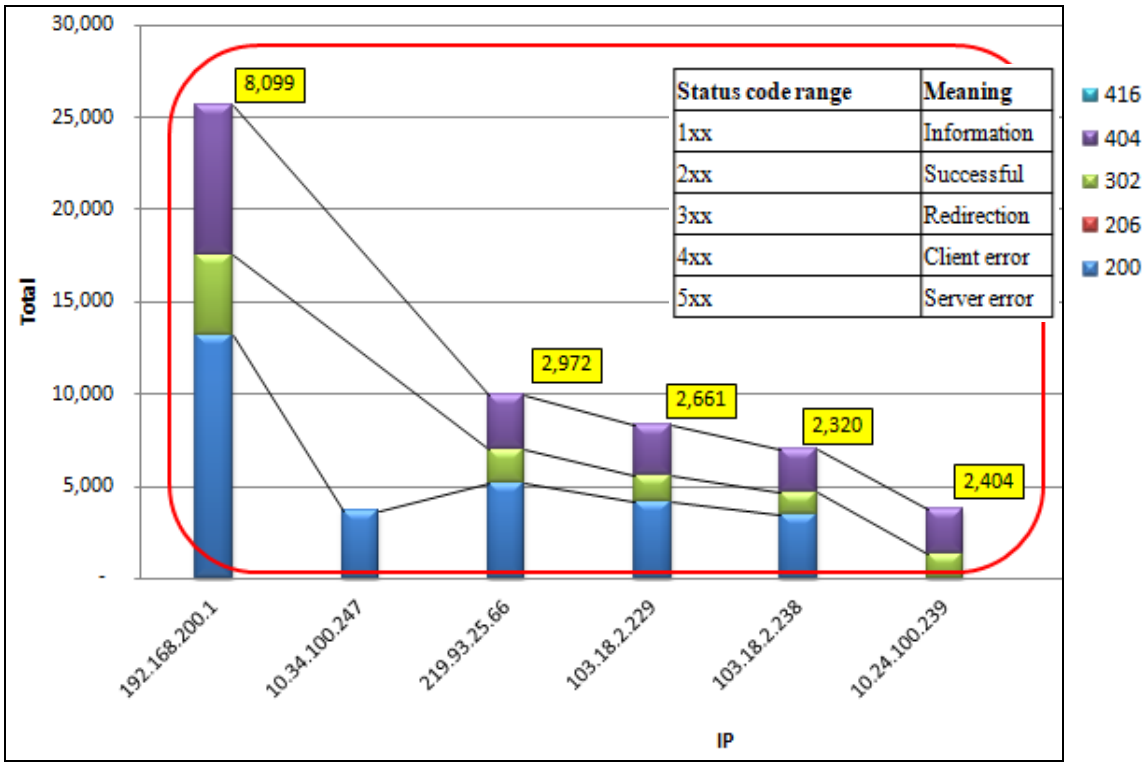

Figure 8. Bar Chart – The Response_Code return while accessing the website

Browsers and Operating System statistics

The Elastic stack tool also helps the system owner in examining which browser is mostly preferred by the users so that the system can be made compatible with that particular web browser. It also gives a tabular description of the total number of hits provided by that particular web browser. Not only browser but it also lists the most preferred Operating System by the users.

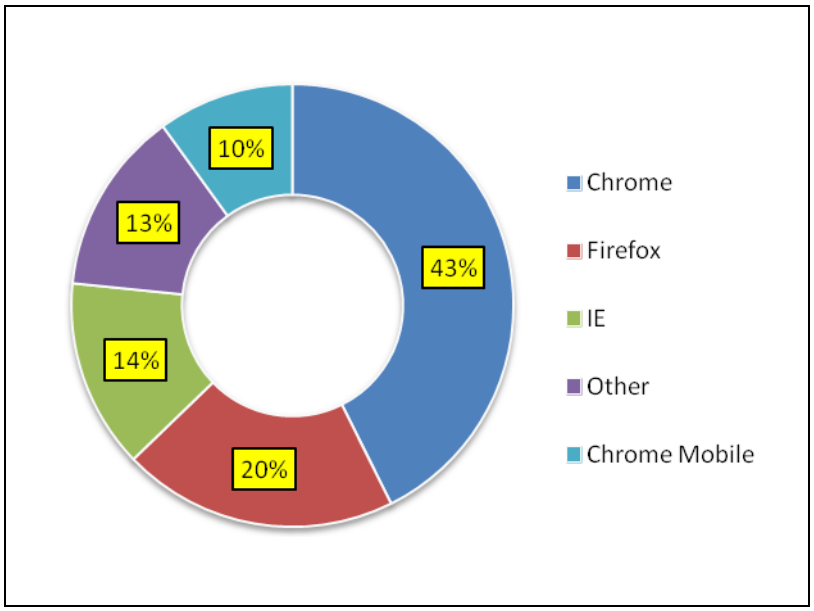

Figure 9. Pie Chart – Showing the most used web browsers/User Agents

Table 5. The most used web browsers/User Agents and Operating System

User Agent	Count	Percentage (%)	Operating System	Count	Percentage (%)
Chrome	86,152	42.64	Windows 10	32,384	39.48
			Windows 7	29,808	36.34
			Windows 8.1	12,739	15.53
			Linux	3,594	4.38
			Mac OS X	3,511	4.28
Firefox	40,723	20.15	Windows 10	20,453	51.05
			Windows 7	15,793	39.42
			Windows 8.1	2,713	6.77
			Windows XP	770	1.92
			Mac OS X	337	0.84
IE	27,925	13.82	Windows XP	14,270	51.19
			Windows 7	12,193	43.74
			Windows 10	854	3.06
			Windows 8.1	473	1.7
			Windows 8	86	0.31
Other	27,110	13.42	Windows	25,367	93.71
			Other	1,448	5.35
			Windows 10	196	0.72
			Android	37	0
			Windows 8	22	0.08
Chrome Mobile	20,158	9.98	Android	20,158	100

Table 5 and the pie charts in Figure 9 above captured the user agents used by UMMAIL users to access the system as at December 7th, 2017, showing that Google Chrome 43% and Firefox 20%, both are highly preferred user agents. The others versions of the Web browser are Internet Explorer 14%, Chrome Mobile 10% and unidentified version of browsers are about 13%. A simple assumption can be made from this insight is that most of the users still preferred access their UMMAIL using Desktop/Notebook (client) rather than Android (mobile).

4.1 Various Security Vulnerabilities

Common Types of 404's Errors

Consider one of the most obvious signs of website/system hacked is a Higher volume of 404 error pages. In addition, the 404 errors or genuine linking errors, are types of errors that reveal information about the types of security attempts being made against the system application/website considering that hackers are constantly using this exposed information via error messages by taking advantage of well-known application website vulnerabilities such UMMAIL system.

The importance of tracking these 404 errors is to identify any mistakes that have been made in internal web links as well as possible to identify external web pages that have posted an incorrect link to our sites. In these cases, the ideal solution to the system admin that the bad links need correction, but as an alternative, the administrator could consider creating a 301-Moved permanently to redirect page to retain the traffic. The following are the two

common types of 404's that researcher have seen encountered by the robots in UMMAIL situations, first is Robots.txt and second is Sitemap.xml.

(i) Robots.txt

The researcher spends a lot of time at focusing on the common 404's Robots. After all, it is a robots.txt file provides robots instructions on what they can and cannot crawl on the website. However, for UMMAIL site, they don't have a robots.txt file because this file is optional, but most robots check for this program before accessing the site to know the rule. Thus, the robot will get a 404 error when they attempt to access the robots.txt on UMMAIL site.

In other words, UMMAIL admin doesn't want the robotic visitors exploring broken pages when they could instead be exploring the working pages on your website. It is recommended for UMMAIL to have their own robots file to guide robots around the website, ensuring they access the right files, notwithstanding admin also can use this file for site maintenance such as mirroring and checking for broken links. Also, be careful with this file; do not simply put/hide information in this file since it is traversed by external robots/bots.

(ii) Sitemap.xml

A Sitemap.xml is an XML file that lists all the URLs/pages contained on your website. Many search engines, like Google, finds the information of all pages on your site using sitemap file to find all the pages and, from that information, have a better idea of what pages to show people searching the web. UMMAIL system, don't have an XML Sitemap file and when Google or any other search engines attempt to access that file, they will reach a 404 error page.

Creating a sitemap.xml file for UMMAIL will reduce or avoid the 404, also helping for marketing efforts by helping such search engines find more of the pages on your site. There are several tools to do so, including XML-Sitemaps.com.

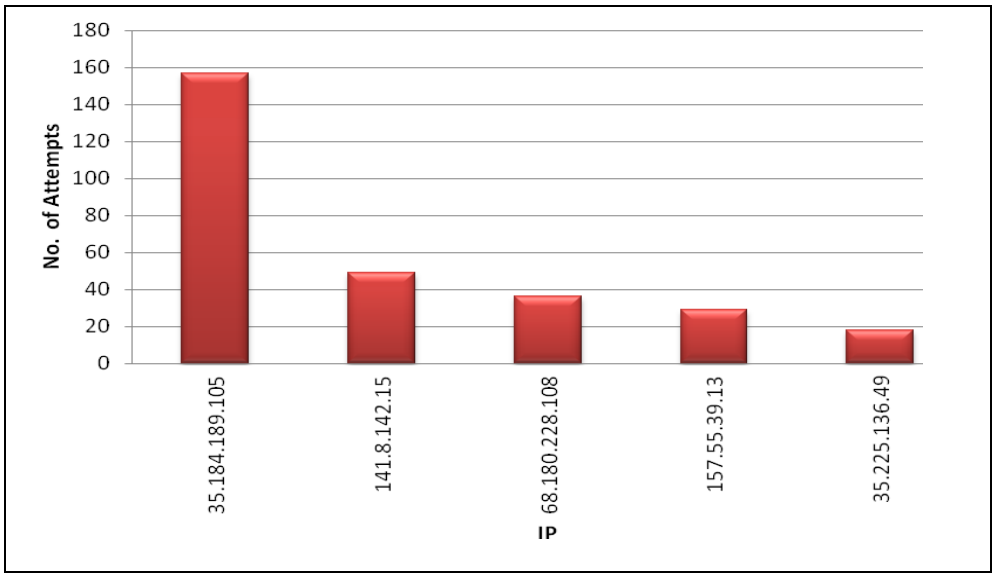

Figure 10. Bar Chart – Common 404s robots encounter

Figure 10 shows the result of common 404s robots encounter errors. The log files contain the data from December 1st, 2017 to December 27th, 2017 were analyzed and managed to capture 5 clients/IPs with the highest 404 error range from 18 to 157. It can be seen within half an hour from 05:05 am to 05:28 am on the same day (03/Dec/2017), there are 157 attempts by IP 35.184.189. The attempted is continuous, followed by the IP 141.8.142.15 of 49 attempts within 2 days (03/12/2017 and 10/12/2017). There are 36 attempts by within 5

consecutive days (06/12/2017 and 11/12/2017-15/12/2017). This attempted decrease, by observing the last two IPs of 157.55.39.13 and 35.225.136.49 respectively with 29 attempts for 2 days (10/12/2017-11/12/2017) and 28 trials for 3 hours in a day (12/Dec/2017).

'Sign In' Attempts

Here are some assumptions can be made by the system administrator for Authentication Logs:

1) Successful login by " non-provisioned" user
2) Logins at strange hours (or during vacation)
3) Obvious scanning activity (i.e., 2+ different users, 1 IP)
4) Multiple failures followed by a success
5) The GeoIP data (where is the user coming from?)
6) Authentication from non-client device (i.e., web server to web server)

In this section reveals the top IPs accessing validation pages with valued hits. The findings from subsection 4.1 above have been referred for deeper analysis to identifying hotlink URL accessed by remote IP. Note that the um_gvalidate.php are identified were accessed more than once from 15 IPs with maximum hits is 4,370 (27%) while the least hits were 466 (3%) as illustrated in Figure 11.

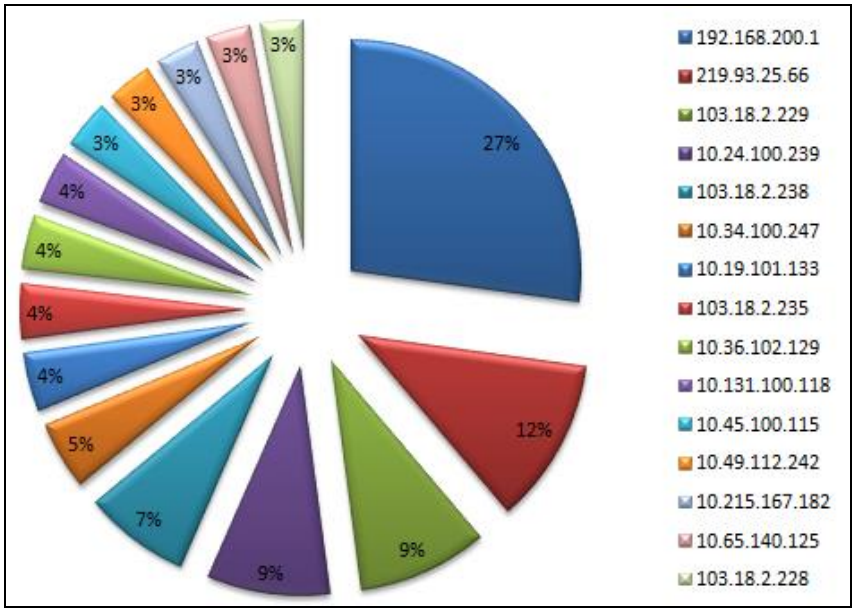

Figure 11. Pie Chart – Top 15 IPs hits validation URL

Before users can access the inbox of their email accounts, the system requires users to input the login credential. After user clicks the 'Sign In' button, the system needs to validate user's input credential through **um_gvalidate.php** in order to allow them to access the inbox. As the UMMAIL system applied server-side validation approach, the information/credential provided by users is sent to the server and validated using server-side languages. If the validation fails, then the server response back to the client, the page displayed the web form refreshed and a feedback is shown similar to Figure 12. For the case of IP number **192.168.200.1** with total hits **4,370**, the system administrator has taken action by blocking this IP at the server firewall level to end this client from continuous access to the system.

This extraordinary circumstance tells the system administrator that it is not supposed to happen because the system has given a clear feedback in Figure 12 to the user if he/she fails to access the system or after he/she is able to access the system.

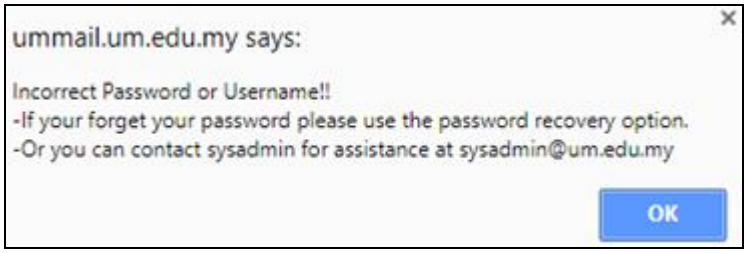

Figure 12. Feedback from server to client when user validation fails

On the other hand, this validation upon submit method will result in a slow response from the server. However, server-side validation is secure because it will work even if JavaScript is turned off in the browser and it can't be easily bypassed by malicious users. Furthermore, the UMMAIL system does not implement client-side validation, due to its main drawback is that it relies on JavaScript. Through this analysis, it is recommended that such critical system like UMMAIL system, should combine server-side and client-side methods for the best of these two approaches where merging fast response, more secure validation especially with client-side on (required fields, correct format and confirmation fields) and better user experience.

Attempts to Change Password through 'Forgot Password' Facility

The statistics obtained the top 5 IPs attempt to change the password from Elastic stack related to Forgot Password recovery action by system's user are shown in Table 6 and illustrated in Figure 13 below. About 55% (70 attempts) of the users indicated by IP 10.17.100.101 tried to change the password. The second and the third highest attempts are 16% (20 attempts) and 11% (14 attempts), both from IP 77.177.92.20 and IP 35.184.189.105. There are about 10% (13 attempts) and the other is 8% (10 attempts) contributed by the IP 40.77.167.77 and 34.192.116.178.

All these attempts were made by accessing https://ummail.um.edu.my/forgotpasswd/forgot_password.php facility. Figure 13 shows the illustration of results obtained in percentage. The system owner/administrator can make two assumptions about this situation either:

- the client/host completely forgot the e-mail password, or
- the client/host attempts to break that page

These facts convince the system owner to ensure that any content being indexed by Google (or other search engines) actually needs to be indexed. Typically, passwords, database connection strings (Conn.php, Myconn.php, etc.), and other sensitive files are not meant to be stored in an accessible, crawlable portion of your website. If it needs to be there, put it in a directory with HTTP authentication and blacklist that directory using robots.txt.

Table 6. Top 5 IPs Attempt to Change Password

Remote IP	Total
10.17.100.101	70
77.177.92.20	20
35.184.189.105	14
40.77.167.77	13
34.192.116.178	10

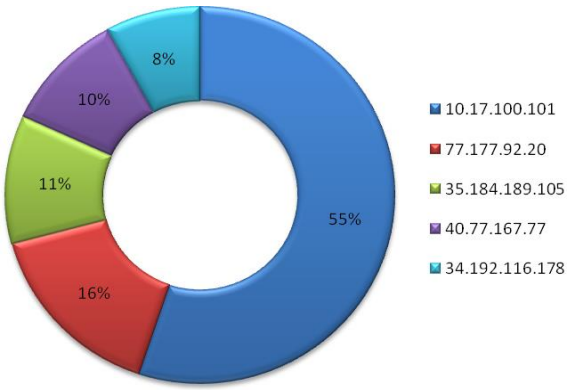

Figure 13. Doughnut Chart – Top 5 IPs Attempt to Change Password

5. Conclusion

In this study, the complete analysis of web server log files was done using Elastic stack. The results obtained shall definitely facilitate people like System Administrator, Website Designers, Website Maintainers, Website Analysts, and Developers to manage their System in searching the problems such as corrupted/broken links. This work will also encourage the staff to find an effective solution to that problem.

From the comparative study and Elastic stack use cases, the Elastic (ELK) stack are seen dominating an analytic space today as being a major competitor to several famous analytic tools like Splunk. It is highly useful as its analysis reports provide the system administrator with enough information for better understanding of the user requirements and preferences and thus leads to improved decision making about the system design and structure.

References

Agrawal, K., & Makwana, H. (2015). Data Analysis and Reporting using Different Log Management Tools. *International Journal of Computer Science and Mobile Computing*, *47*(7), 224–229.

Al-Asadi, T. A., & Obaid, A. J. (2016). Discovering similar user navigation behavior in web log data. *International Journal of Applied Engineering Research*, *11*(16), 8797–8805.

Badshah, G., Liew S. C., Zain, J.M**.,** Mushtaq Ali (2016). Watermark Compression in Medical Image Watermarking Using Lempel-Ziv-Welch (LZW) Lossless Compression Technique. *Journal of Digital Imaging*, 29(2), 216-225. DOI 10.1007/s10278-015-9822-4

Bishop, I. D., Pettit, C. J., Sheth, F., & Sharma, S. (2013). Evaluation of data visualisation options for land-use policy and decision making in response to climate change. *Environment and Planning B: Planning and Design*, *40*(2), 213–233. https://doi.org/10.1068/b38159

Bodó, R., & Kouˇ, D. (2013). Efficient Management of System Logs using a Cloud.

Brian Jackson. (2017). Top 10+ Log Analysis Tools - Making Data-Driven Decisions. Retrieved November 9, 2017, from https://www.keycdn.com/blog/log-analysis-tools/

Čegan, L., & Filip, P. (2017). Webalyt : Open Web Analytics Platform.

Chainourov, B. (2017). *Log Analysis Using Splunk Hadoop Connect*. NAVAL POSTGRADUATE SCHOOL. Retrieved from https://calhoun.nps.edu/bitstream/handle/10945/55581/17Jun_Chainourov_Boulat.docx.pdf?sequence=1

Fuente, A. D. D., Andreassen, O. O., & Charrondière. (2015). Monitoring Mixed-Language Applications with Elasticstash, Logstash and Kibana (ELK). *Proceedings of ICALEPCS2015*, 786–789.

Goel, N., & Jha, C. K. (2013). Analyzing Users Behavior from Web Access Logs using Automated Log Analyzer Tool. *International Journal of Computer Applications*, *62*(2), 975–8887. https://doi.org/10.5120/10054-4643

Hasani, Z., Jakimovski, B., Kon-popovska, M., & Velinov, G. (2015). Real Time Analytic of SQL Queries Based on Log Analytic, 78–87.

Iuhasz, G., Pop, D., & Drăgan, I. (2016). Architecture of a scalable platform for monitoring multiple big data frameworks. *Scalable Computing*, *17*(4), 313–321. https://doi.org/10.12694/scpe.v17i4.1203

Jayakumar, V., & Alagarsamy, K. (2013). Analysing Server Log File using Web Log Expert in Web Data Mining. *International Journal of Science, Environment and Technology*, *2*(5), 1008–1016.

Jayathilake, D. (2012). Towards structured log analysis. *JCSSE 2012 - 9th International Joint Conference on Computer Science and Software Engineering*, 259–264. https://doi.org/10.1109/JCSSE.2012.6261962

Keim Daniel, K. J., & Mansmann, G. rey E. and F. (2010). Mastering the Information Age Solving Problems with Visual Analytics. *Mastering the Information Age Solving Problems with Visual Analytics*, 57–86. https://doi.org/10.1016/j.procs.2011.12.035

Koga, I., & Almeida, E. S. (2016). File Transfer Log Analysis : A meteorological data center case study, *111*, 1–13. https://doi.org/10.6062/jcis.2016.07.03.0111

Kumar, V., & Thakur, R. S. (2017). A Brief Investigation on Web Usage Mining Tools (WUM). *ISSN*, *2*(1), 1–11. https://doi.org/10.21276/sjeat.2017.2.1.1

Liew, S.C. , Liew, S.W. and Zain, J.M**.,** (2010)**.** Reversible Medical Image Watermarking For Tamper Detection And Recovery With Run Length Encoding Compression, *World Academy of Science, Engineering and Technology (WASET)*, **Issue 72**, December 2010, pp. 799-803.

Lu, C., Zeng, H., Liu, J., Zhang, R., Chen, Y., & Yao, Y. (2017). Network Security Log Analysis System Based on ELK, (Cece), 554–559.

M. Fedorov, P. Adams, G. Brunton, B. Fishler, M.Flegel, K. Wilhelmsen, & R. Wilson. (2017). Leveraging Splunk for Control System Monitoring and Management. *ICALEPCS 2017*, 1–7. Retrieved from https://scholar.googleusercontent.com/scholar?q=cache:CbAw1NEck74J:scholar.google.com/+logs+analytics+tools+&hl=en&as_sdt=0,5&as_ylo=2017&as_yhi=2017

Maria Minsker. (2017). Analytics Tools Are Still a Challenge for Many - eMarketer. Retrieved December 24, 2017, from https://www.emarketer.com/Article/Analytics-Tools-Still-Challenge-Many/1015983

Mateik, D., Mital, R., Buonaiuto, N. L., Louie, M., Kief, C., & Aarestad, J. (2017). Using Big Data Technologies for Satellite Data Analytics. *AIAA SPACE and Astronautics Forum and Exposition*, 1–10. https://doi.org/10.2514/6.2017-5161

Mikula, A., Adamová, D., Adam, M., Chudoba, J., & Švec, J. (2016). Grid Site Monitoring

and Log Processing using ELK, 54–61.

Mitra, M. J. (2016). The Rise of Elastic Stack The Rise of Elastic Stack, (November). https://doi.org/10.13140/RG.2.2.17596.03203

Prakash, T., & Patel, K. (2016). Geo-Identification of Web Users through Logs using ELK Stack, 606–610.

Purnami, S. W., Zain, J.M**.**, Tutut Heriawan, "An alternative algorithm for classification large categorical dataset: k-mode clustering reduced support vector machine", *International Journal of Database Theory and Application (IJDTA)*, **Vol. 4, No. 1**, March 2011, pp. 19-29

Shakti, M. K., & Garg, L. (2017). Web Log Analyzer Tools: A Comparative Study to Analyze User Behavior, 17–24. https://doi.org/978-1-5090-3519-9/17/$31.00@2017 IEEE

Smith, G. (2015). Log Analysis with the ELK Stack (Elasticsearch, Logstash and Kibana). Retrieved from https://www.linuxfestnorthwest.org/sites/default/files/slides/Log Analysis with the ELK Stack.pdf

Umarani, J., Silambarasi, A., & Thangaraju, G. (2016). Web Usage Mining Based Analysis of Web Site Using Web Log Expert Tool. *International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)*, *23*(5), 33–36.

Valency Networks. (2016). Security Log Analyzer Tools | Pune Mumbai Hyderabad Delhi Bangalore India | Valency Networks. Retrieved October 17, 2017, from http://www.valencynetworks.com/articles/top-10-web-log-analyzers.html