7th International Scientific Conference
Technics and Informatics in Education
Faculty of Technical Sciences, Čačak, Serbia, 25-27th May 2018
Session 2: IT Education and Practice
UDC: 378:004.42

Using Web Server Log Files for Analysis and Improvements Related to Study Programs

Predrag Stolić [1*], Danijela Milošević [2]
[1] University in Belgrade, Technical faculty in Bor, Serbia
[2] University in Kragujevac, Faculty of technical sciences Čačak, Serbia
[*] pstolic@tfbor.bg.ac.rs

Abstract: *Almost every computer system records data about the corresponding events in the system through the so-called log data. Log data can provide meaningful information and knowledge about different aspects of system use, but placed in the appropriate context, log data also provides knowledge beyond the boundaries of the system in which they are generated. The paper shows how log data obtained from one web server can be used to analyze and improve study programs. The entire flow of log data transformation from the sources to the final results applicable for further work is shown. The paper points to a part of the potentials that log records have and how this potential can be used within the framework of higher education work. A concrete solution is presented based on the use of an appropriate infrastructure powered by Elastic Stack solution.*

Keywords: *Elastic Stack, higher education, log, study program, web server*

1. INTRODUCTION

Although the log is the term commonly used in computer and related sciences, this term dates from the time before the beginning of using computers in the form it is known to us nowadays. For example, if we look at The Illustrated Dictionary Oxford, we can find, among other definitions, that the log is "a record of events occurring during and affecting the voyage of a ship or aircraft" [1]. Basically, similar explanation ("a record of events") has been taken by computer science. But if we look deeper in computer systems, faced with their complexity, logs are more than simple records of events because each event represents individual phenomenon within an environment that usually includes an attempt to change the state [2]. Accordingly, log is described as a set of records of individual occurrences in an environment that has changed or attempted to change a previous state. Logs (log files) provide vital information about various types of behavior [3].

Using of log files in computer systems today have a wider application than one observed in some traditional sense. List of solutions based on approaches which involved some type of log files is potentially unlimited. The domain of education also is following this trend and becomes one of the areas where a large increase in the use of solutions based on the use of log files is recorded.

Zheng, He, Ma, Xue, Li and Dong in [4], analyzing one e-learning system, pointed out the complexity of working with log files in modern educational environment and discussed potential solution and infrastructure to overcome that complexity using Big Data approach and techniques.

Takahashi, Asahi, Suzuki, Kawasumi and Kameya in [5] also used log data from one e-learning system which is realized using cloud principles. Extracted knowledge from log files is used for analysis of self-learning styles as potential improvement of system aimed to support self-learning from home. Similar approach, based on extraction of knowledge about learning styles from log data, is done by Umezawa, Aramoto, Kobayashi, Ishida, Nakazawa and Hirasawa in [6] in order to improve the implementation of the educational approach called "flipped classroom".

References [7], [8] and [9] show how log data can be used in area of higher education especially in overcoming the "language barrier" problem related to mobility of students. Provided vital information from log data, authors increased chances of getting a job in Japan for their international students that have a lack of knowledge of the Japanese language.

Log data which originally come from non-educational systems also can be used for the findings that can improve various segments of education.

Authors in [10] used log data for analysis of user influences in social networks and authors in [11] done some opposite process, construct the social network from chat between users. In both cases

some conclusions obtained from those log data can be used for generating principles for interaction between users in educational systems.

Li, Zhao, Wang, Ma and Liu in [12] used log data from one commercial business system for analyzing and finding patterns in user behavior and based on that to predict further purchases on online shopping system. Similar approach can be obtained for example in e-learning systems for prediction of next course which user will be involved in.

In this paper log data which are used are originated from classical web server of educational institution and findings from these data will be used for some analysis related to existing study programs.

2. OBTAINING DATASETS

As already mentioned above, log data, which are used in further explanations and analysis, are provided from single web server. This server belongs to academic institution and serve eleven web sites related to that institution: one official site of the institution, four sites of study programs, three conference sites, two journal sites and one site for presentation of R&D activities. All sites are operating under Apache HTTP Server 2.4 and all sites use common log files.

Access logs will be observed because this type of log records show all requests processed by the server. No log rotations are provided by the server. Data are recorded using format string referred as the Combined Log Format [13] represented as:

```
LogFormat "%h %l %u %t \"%r\" %>s %b \"%{Referer}i\" \"%{User-agent}i\"" combined
CustomLog log/access_log combine
```

which means that in access log are recorded IP, identity, user, time, request, status, size, Referer and User-Agent HTTP request header. This set of data is enough for type of analysis provided in this paper.

In this paper two generated access.log files are used. Log files are shown in table 1. As we see from corresponding table we have total of over 8,5 million recorded log messages which is quite impressive number of records for not heavy-load sites of mentioned academic institutions.

Bhole, Adinarayana and Shenoy in [14] highlighted that the first phase in work with every dataset must be so called "data cleaning" which implies procedures of removing irrelevant and redundancy data in datasets. Same approach is used in this paper in aim of achieving better performances during working with datasets and obtaining more precise results from datasets. Results of datasets cleaning are shown in table 2.

As it shown in tables 1 and 2, initial state was about 8,5 million recorded log messages and after cleaning process over both datasets total amount

Table 1. *Used access.log files*

access.log	I	II
Logging started	22 June 2017 09:47:04	8 December 2017 08:04:37
Logging ended	8 December 2017 08:01:52	23 February 2018 06:46:34
Number of recorded log messages	6.192.378	2.349.682
Size of file	1,5 GB	591,1 MB

Table 2. *Used access.log files after cleaning process*

access.log	I	II
Started number of log messages	6.192.378	2.349.682
Number of removed log messages	3.348.180	982.098
Percentage of removed log messages	54,07 %	41,8 %
Number of log messages after cleaning process	2.844.198	1.367.584

of valid log messages for further processing is reduced to about 4,2 million log messages which represents overall improvement of over 50 percent. This significant reducing also made reducing in time needed for whole other sets of operations during analysis so we could say that we will use resources during data processing in some optimal and efficient way.

3. INFRASTRUCTURE FOR DATA ANALYSIS

There are many possible solutions for realization of log analytics platforms. One proposed solution is described in [15], which offers use of layered log analytics architecture. Mentioned solution is good choice for dealing with logs in Enterprise Architecture (EA) business solutions, but in case which presented in this paper authors are tried to find less complex and more narrowed solution which can handle Apache log data described above.

After comparison of few solutions, which can be found on market nowadays, it was decided that

Elastic Stack product family would be used and complete analysis of log files are done using this product family consisted of Logstash, Elasticsearch and Kibana in this case [16].

Using the recommendations of the software vendors, Elastic Stack software is used on laboratory network consisted of three HP Proliant ML310 G5p servers (one server based on quad core Intel Xeon X3330 with 8 GB RAM and two servers based on dual core Intel Xeon E3120 with 4 GB RAM, all servers equipped with standard SATA 7200 rpm hard drives) connected via MikroTik routerboard. All servers worked under Linux Fedora Server 27 operating system and MikroTik routerboard worked under RouterOS 6.42.

Deployment of hardware components is shown in the figure 1.

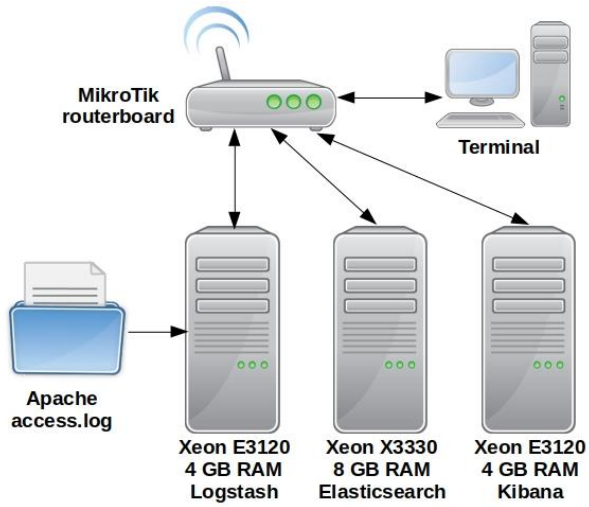

Figure 1. *Deployment of hardware components*

The abstraction of data flows is graphically shown in the figure 2.

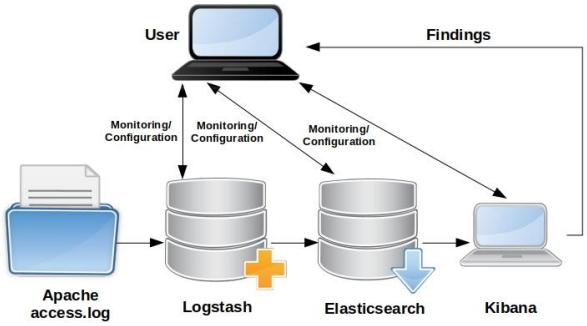

Figure 2. *Data flows*

As we see from figure 2, desired log data, which are previously cleaned as mentioned above, are first collected and processed by Logstash which represents server-side data processing pipeline. Log data are then transformed using built-in grok filter which primary function is to transform and structure non-structured data from log file. Filtered data then are sent to Elasticsearch for further processing.

Key analysis processes are happened within Elasticsearch which provided three main components for dealing with filtered data: search, analyze and store. Using Elasticsearch capabilities we are in position to reveal some knowledge from large amount of log data which cannot be detected on some other way. Elasticsearch can be characterized as one big solver for various types of doubts, issues and events which represents priceless value in our case of answering some important questions about existing study programs.

Elasticsearch data are in raw format so it must be processed further for interpretation and presentation in some understandable way. For that reason data from Elasticsearch are sent to Kibana which has two main goals. First, Kibana is powerful visualization tool for all data which are processed by Elasticsearch. Kibana provided so called dashboards which can obtained results of data processing in one accurate, precise and appealing way. Also, Kibana provided GUI for configuring and monitoring of Elastic Stack as it shown on figure 3. With this type of data management users can improve performances of various parts of system and at the same time secure the quality of the whole process.

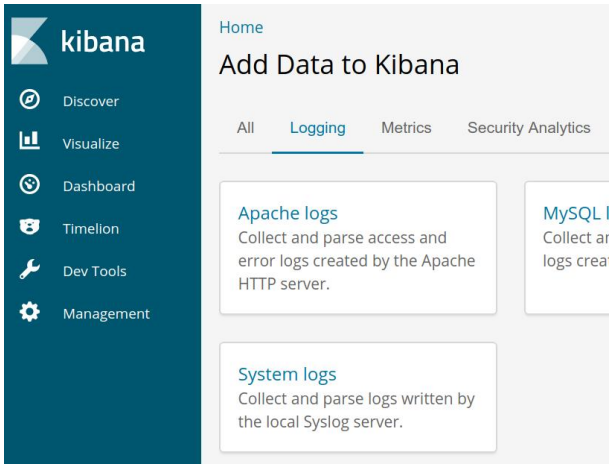

Figure 3. *Configuring Elastic Stack using Kibana*

Final product of all operations over log data through whole Elastic Stack components are findings which provided solutions for single or set of problems.

In this paper we used configuration based on small laboratory network described above and presented on figure 1, but we must mention that the whole process can be done also using single computer. Authors also tested this type of implementation which are realized on a computer based on Intel i5-6400 CPU with 32 GB RAM and equipped with SSD and SATA 7200 rpm HDD. All services of Elastic Stack are successful implemented and configured on this single machine and all data are processed, but if we compare performances of both solutions, better performances are achieved using small laboratory network so that solution is used further.

Also, in this research, log files that are used are final and locked without further recording of any log message. But for various purposes this whole process can be implemented in real-time

configuration. Basically, concept is the same. Only difference is that we must, before Logstash, add Filebeat component which will collected all log messages from log file in real-time and processed that log messages further to Logstash. Rest of the described process is the same.

Authors did not work with real-time data in this research, so base point in this paper will be log file attached to Logstash without presence of Filebeat.

4. FINDINGS

4.1. Search

First method which we provided using Elastic Stack over the cleaned data from log files is search. Previously, already is mentioned that in these data are provided log messages for eleven sites, so the task was to see distribution of visits among these sites. Of course we will be mainly focused on institution official site and study programs sites.

For search we will use Referer field because this field from log data provides to us the most accurate data. Of course, the addresses of all sites were well known during investigation so we could searching using these parameters. Results are shown in table 3.

We will not discuss results related to sites of conferences, journals and R&D. Only we will say here that occurrence of zero hits for the site of Conference III is related to the fact that the site was not operative during creation of log file I and become operative during creation of log file II so zero result is expected occurrence for log file I.

We will discuss results related to official institution site and sites of study programs. As we see from the results, we have two observed facts. First we see good results for official site and site of study program I and very poor results for the rest of study program sites (only 2-3 percentage). And second is occurrence in log file II that site of the study program I has more hits than official site.

4.2. Analysis

References [17] and [18] gave some basic instructions about a way for interpreting results from log files. But analysis and interpretation of log files is some state of art and can vary significantly from case to case. In this case, our entering point was search results. In some other cases entering point can be something else.

Search method gave to us some key guidelines and knowledge about environment, but to get the real answers and to get key findings, search was only one stage of larger process. Guided with search results, further analysis are provided.

First part of analysis was related to our efforts to understand why three sites of study programs had a very poor results and to define is it anomaly or there are some key facts related to these study programs.

Table 3. *Results of provided search over log files*

access.log	I		II	
	HITS	**%**	**HITS**	**%**
Institution official site	1.546.591	*54,38*	589.666	*43,12*
Study program I	925.561	*32,54*	606.582	*44,35*
Study program II	57.412	*2,02*	28.426	*2,08*
Study program III	70.450	*2,48*	34.717	*2,54*
Study program IV	*91.993*	*3,23*	*52.682*	*3,85*
Conference I	75.007	2,64	12.623	0,92
Conference II	16.929	0,60	6.145	0,45
Conference III	0	0	9.505	0,70
Journal I	21.821	0,77	12.303	0,90
Journal II	33.662	1,18	4.441	0,32
R&D site	4.772	0,17	10.494	0,77

In analysis we used almost all fields from Apache log messages, primarily time, request, status, size and Referer.

First we analyzed status codes from log messages to find are there any massive errors during visits to these sites. Most of log messages were with status code 200 (OK) so we concluded that the most requests were legitimate and without errors.

After that we made further analysis combined fields Referer and request to see distribution among visiting pages and objects per each site.

For study program sites with poor hits results major visiting pages and objects are schedules of various types (mostly classes and exams) with percentage of about 71 percent of all visiting pages. Opposite, for study program site with high hits rate, that percentage was about 34 percent of all visiting pages. Also users which visit this site mostly use site for access to some kind of teaching materials (videos, presentations, books and similar), about 52 percent of all visiting pages. From these facts we concluded that the poor results for some sites and very good result for one site on the other side

are direct consequence of presence of teaching materials on one side and major absence on the other.

Using this finding, further analysis of concrete study programs are made. Absence of teaching materials on sites with poor rates are direct consequence of very old literature which are used as teaching materials on these study programs. Some of literature are limited available and most of them not have digital interpretations or digital equivalents which could be presented to students online.

As a step for the improvement of these study programs, it is proposed to innovate literature with a special emphasis on creating digital content that will be available to students online on a 24/7 basis.

Special analysis were also made for mentioned occurrence in log file II to find what led to unusual situation that site of the study program I has more hits than official site of the institution.

Also as previously, first we analyzed status codes in pursuit for possible errors and again most of log messages were with status code 200 (OK). Among different analysis which not gave correct answer to us, the analysis which primarily involved time distribution using time field gave us precious insights.

During the comparison of obtained results for official site and obtained results for site of study program I, overlap was observed. Almost at the same time less hits were done on the official site and more hits were done on the site of study program I. After the notice of period which this were happen, further analysis are made especially for that period of time. Using adequate fields there are noticed that users more often visits teaching materials on the study program site than usually. Analyzing the timeline of events which are occurred in that period, we found that in that period were a legitimate terms for taking exams, so students study and prepare exams and using much more teaching materials on site than usually. According to that, this occurrence is treated as legitimate.

5. CONCLUSION

The research presented in this paper shows the possibility of transforming data contained in log files into knowledge that can contribute in improving the environment in which higher education takes place, or in this case, improvements related to study programs. Log data which were used to achieve described findings are obtained from classic web server instead common practice in similar cases when log data are obtained mostly from some specialized learning systems. As we saw, extracted knowledge from those data are efficiently made space for future actions on the improvement of the study programs.

But this presented solutions is not without some limitations.

We already mentioned that more accurate approach and results are achieved when we are dealing with those data in real-time. So the best solution for this sort of analysis of log data will be based on some sort of streaming log data and parsing them further in real time. Of course, this kind of approach requires additional investments in infrastructure which will be made exclusively for these purposes, so there is a potential issues in realization for institutions with smaller sources of income.

Also more accurate findings can be obtained when the results which are extracted from these log data are compared with other statistical indicators. For example in our case a higher degree of accuracy can be achieved if we in our observations involved exact number of students per study programs, per semester and per subject. This approach need various integrations of presented system which are dealing with log data from different aspects with various parts of information systems on faculties, universities, local administration and similar. Sometimes these integrations are very complex, demanding and expensive and require a lot of resources.

However, the prediction is that in near future we will be introduced with various systems based on use of different types of log data and we will made large amount of decisions based on insights and findings revealed from those systems.

REFERENCES

[1] Group of authors (2002). *The Illustrated Dictionary Oxford.* Novi Sad, Serbia: Mladinska knjiga nova.

[2] Chuvakin, A., Schmidt, K. & Phillips, C. (2013). *Logging and Log Management*. Waltham, Massachusetts: Elsevier.

[3] Chhajed, S. (2015). *Learning ELK Stack*. Birmingham, England: Packt Publishing.

[4] Zheng, Q., He, H., Ma, T., Xue, N., Li, B. & Dong, B. (2014). *Big Log Analysis for E-learning Ecosystem.* 2014 IEEE 11th International Conference on e-Business Engineering (ICEBE), 258-263. doi:10.1109/ICEBE.2014.51.

[5] Takahashi, T., Asahi, K., Suzuki, H., Kawasumi, M. & Kameya, Y. (2015). *A Cloud Education Environment to Support Self-Learning at Home - Analysis of Self-Learning Styles from Log Data.* 2015 IIAI 4th International Congress on Advanced Applied Informatics (IIAI-AAI), 437-440. doi:10.1109/IIAI-AAI.2015.213

[6] Umezawa, K., Aramoto, M., Kobayashi, M., Ishida, T., Nakazawa, M. & Hirasawa, S. (2015). *An Effective Flipped Classroom Based on Log Information of Self-Study.* 2015 3rd International Conference on Applied Computing and Information Technology/2nd

International Conference on Computational Science and Intelligence (ACIT-CSI), 248-253. doi:10.1109/ACIT-CSI.2015.52

[7] Uosaki, N., Ogata, H., Mouri, K. & Lkhagvasuren, E. (2015). *Career Support for International Students in Japan Using Ubiquitous Learning Log System.* 2015 IEEE 15th International Conference on Advanced Learning Technologies (ICALT), 78-82. doi:10.1109/ICALT.2015.19

[8] Uosaki, N., Kiyota, M., Mouri, K. & Ogata, H. (2016). *Career Support for International Students in Japan Using Learning Log System with eBook.* 2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), 1205-1206. doi:10.1109/IIAI-AAI.2016.224

[9] Uosaki, N., Kiyota, M., Mouri, K., Ogata, H. & Choyekh, M. (2016). *Onomatopoeia Learning Support for Japanese Language Learners Using Ubiquitous Learning Log System with eBook.* 2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT), 347-348. doi:10.1109/ICALT.2016.98

[10] Wang, H., Liu, S., Yu, H. & Lu, Y. (2014). *A Method to Measure User Influence in Social Network Based on Process Log.* 2014 IEEE 11th International Conference on e-Business Engineering (ICEBE), 338-343. doi:10.1109/ICEBE.2014.65

[11] Tavassoli, S., Moessner, M. & Zweig, K.A. (2014). *Constructing social networks from semi-structured chat-log data.* 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 146-149. doi:10.1109/ASONAM.2014.6921575

[12] Li, D., Zhao, G., Wang, Z., Ma, W. & Liu, Y. (2015). *A Method of Purchase Prediction Based on User Behavior Log.* 2015 IEEE International Conference on Data Mining Workshop (ICDMW), 1031-1039. doi:10.1109/ICDMW.2015.179

[13] Apache HTTP Server Project (2018). *Apache HTTP Server Version 2.4 Documentation: Users' Guide – Log Files*. The Apache Software Foundation. Accessed in April 2018 at https://httpd.apache.org/docs/2.4/logs.html

[14] Bhole, A., Adinarayana, B. & Shenoy, S. (2015). *Log analytics on cloud using pattern recognition a practical perspective to cloud based approach.* 2015 International Conference on Green Computing and Internet of Things (ICGCIoT), 699-703. doi:10.1109/ICGCIoT.2015.7380553

[15] Delic, K. & Riley J. (2015). *SLA : Smart log analytics*, 2015 XXV International Conference on Information, Communication and Automation Technologies (ICAT), 1-3. doi:10.1109/ICAT.2015.7340517

[16] The Open Source Elastic Stack (2018). Accessed in April 2018 at https://www.elastic.co/products

[17] Nimbalkar, P., Mulwad, V., Puranik, N., Joshi, A. & Finin, T. (2016). *Semantic Interpretation of Structured Log Files.* 2016 IEEE 17th International Conference on Information Reuse and Integration (IRI), 549-555. doi:10.1109/IRI.2016.81

[18] Zhu, J., He, P., Fu, Q., Zhang, H., Lyu, M. & Zhang, D. (2015). *Learning to Log: Helping Developers Make Informed Logging Decisions.* 2015 IEEE/ACM 37th IEEE International Conference on Software Engineering (ICSE), 415-425. doi:10.1109/ICSE.2015.60