An Innovative Dynamic Method of Multiple Users Log Analysis using ELK Stack

Hiral M. Patel[1] [0000-0002-8232-8303] and Vatsala D. Patel[2]

[1,2]Department of Computer Engineering, SAL Institute of Diploma Studies, Ahmedabad 380060, India
hirvinpatel@gmail.com , vatsala5883@gmail.com

Abstract. Numerous different types of data, including raw data, metadata, logs, and metrics, were created by the systems, equipment, and devices in the most recent studies. The ELK i.e. Elasticsearch, Logstash and Kibana stack can be used to aggregate logs from applications and infrastructure. Summarizing and analyzing logs is of great value. Delivering innovative products and features faster is one of the main goals of the new world of cloud applications. To enable faster innovation in a data-driven digital economy, it is imperative to extract value from machine data by aggregating and analyzing logs, metrics, and events. Using ELK stack we can analyze logs and create visualizations for application and infrastructure monitoring, faster troubleshooting, security analytics, and much more purposes. Currently, we can manage to monitor any number of devices, apps, or infrastructures thanks to the ELK stack. To use ELK across numerous users, one must always visualize the data specific to each session (user) as well as set up ELK to operate in a highly dynamic environment for both current registered users and newly added monitoring entities. So, utilizing the ELK stack, we have developed a novel method of managing multiple users' logs on a dynamic basis.

Keywords: Elastic search, logstash, kibana, Data Visualization, Logs

1. Introduction

In the current era of cloud and AI, Industry cannot afford a single second of downtime or poor performance for their applications or IT infrastructure. It damages brand value as well as some time revenue loss too. Due to which log analysis and monitoring comes into picture which continuously keep gathering data for these applications and infrastructure to make sure its smooth performance. Now-a-days there is an increased use of monitoring data analytics based applications. In such applications to store, visualize and analyze log based data ELK stack [1] is the best open source solution. There are many log analysis tools are available in market like Splunk, SumoLogic, SolarWinds/Loggly, Graylog, Humio and many more. But ELK is an open source which is helpful in initial use for organizations on tighter budgets. Before discussing about ELK stack, let us first understand the basic of log, log analysis and its necessity.

1.1 What is log?

A log is an impression of any action performed by the client. Logs contain large measure of data put away inside them. If a highly productive environment is running then a huge measure of logs would be delivered each second. It isn't attainable to physically peruse line by line and thus there is a prerequisite of a tool which can automate the process. Normally, a log line may incorporate fields like IP Address, Timestamp, zone, http technique, client request, used protocol, status code, execution and response time, memory consumed or custom fields. There are numerous different sorts of logs, for example, application logs, error logs, server logs, proxy error logs, system log, gateway logs, database logs and other custom logs [2].

1.2 What is log analysis?

Inside the tech industry, logs are common. These logs originate from a wide range of devices, including network equipment, web and application servers, IoT devices, etc. Log analysis is a system for looking at and comprehending logs in order to gain insightful information. Therefore, firms may research their logs effectively to gain statistics from them that they otherwise wouldn't be able to. They might then benefit from this knowledge by improving a variety of processes, not only their decision- making process [3], but also a wide range of other processes.

1.3 Why bothers with log analysis?

There are numerous reasons why businesses conduct log analysis. Problem-solving is the primary goal of appearing log analysis. Log analysis aids in real-time problem detection and error correction to prevent damage. Concerns about compliance and security dominate the list of reasons to undertake log analysis. Consequently, understanding and responding to safety occurrences as well as data breaches is
the second reason why enterprises should be concerned about log analysis in the context of safety. Finding proof of a crime or hack, enabling information recovery after disasters, watching the actions of a malicious actor, and many more reasons are behind log analysis.

This work aimed at the set up ELK to operate in a highly dynamic environment for both current registered users and newly added monitoring entities.

The structure of this paper is as follows. Section 2 explains the ELK stack's underlying theory. The related work done by researchers around the world is presented in Section 3. The comparison of static and dynamic analysis is the main topic of Section 4. The research technique is displayed in Section 5. The experimental design and findings are represented in Sections 6 and 7. Finally, Section 8 brings the paper to the end.

2. Background Theory

Monitoring and Data Analysis systems require gathering a very large amount of data in a regular manner. It is very difficult to handle such a large amount of data by traditional Relational Database systems. In such a scenario ELK comes into picture which can easily handle bulk data which get generated in very short duration.

ELK stack is consists of three open source products: Elasticsearch, Logstash, and Kibana.

Let’s have a look at brief overview of all three components of ELK Stack as shown in Figure 1.

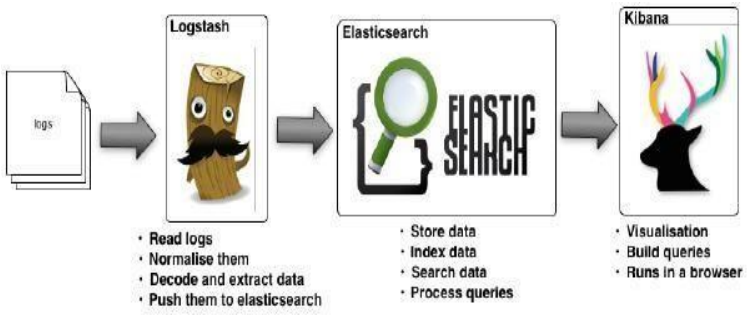

Fig..1.ELK Components [4]

2.1 ElasticSearch (E)

Elasticsearch is a log based database. It is NoSQL database and has its own query structure to search and retrieve data. It stores data/logs in the form of JSON documents and indexes. It is based on the Lucene search engine. ElasticSearch supports a bunch of REST APIs to execute calls to fetch data from storage data.

It also offers advanced queries to perform detailed analysis and stores all the data centrally. It is also helpful for executing a quick search of the documents. Elasticsearch works with JSON record documents. Using an internal structure, it may parse the data almost continuously in order to find the information required.

Elastic search's core ideas are indexing and mapping.

Indexing
It is a collection of different report types and archive attributes. In order to enhanceperformance, Record also makes use of the concept of shards.

Multiple Indices, which consequently contain a variety of Types, can be found in an ElasticSearch bunch. These types include a lot of Records and each record has Properties.

Mapping
It specifies the fields for documents of a particular kind as well as the data types andhow Elasticsearch should index and store the fields [1].

2.2 Logstash (L)

Logstash is the data collection pipeline tool. It collects data from different sources (Devices/Applications/Infrastructure) and feeds that data into the Elasticsearch.
Logstash can unify data from different sources and normalize the data into your desired form. It filters and cleans all your data for analytics and visualization. Logstash consists of three components:

Input: It collects logs received from different sources through different networkchannels (tcp/udp).
Filters: It consists of the set of instructions to filter data received from sources.
Output: It passes filtered data to elastic search indexes for storage and analytics.

2.3 Kibana(K)

Kibana is used for visualization of stored data (documents) in ElasticSearch. Kibana also offers many options by which users can manage stored data. Kibana dashboard offers various interactive diagrams, geospatial data, and graphs to visualize complex queries as well as data search.

Kibana is also useful to search, view and manage data stored in indexes. Kibana helps you to perform advanced data analysis and visualize your data in a variety of tables, charts, and maps.

In summary, following are the key features of ELK that makes it a powerful, flexible, and scalable platform for log analysis, making it a popular choice for organizations of all sizes.

Scalability: ELK is highly scalable and can handle large amounts of data. Elasticsearch, the search and analytics engine behind ELK, is built to handle huge amounts of data in real-time, making it an ideal choice for large-scale log analysis.

Flexibility: ELK is a modular system, which means that you can choose which components you want to use and how you want to configure them. This makes it flexible and adaptable to different use cases and requirements.

Integration: ELK integrates with a wide range of data sources and tools, making it easy to analyze logs from different applications and systems. It also supports various plugins, which can extend its functionality.

Visualization: Kibana, the visualization component of ELK, offers a wide range of visualization options, making it easy to create dashboards and reports. You can also customize the visualizations to suit your needs.

3. Related works

In order to examine log data generated by neutron experiments completely, K. Moriyama et al. [5] employed the ELK stack in their research. They claim that the ELK stack makes it possible to skillfully and effectively dissect log data, learn about information investigation and exploratory planning, as well as instrument activity.

The effectiveness of the open source ELK stack against the commercial Splunk solution has been compared by Sung Jun Sonet et al. [6]. With the experiment, they were able to demonstrate that the ELK stack can be an effective, entry-level security log analysis solution with comparable performance to expensive commercial products.

The system for cyber attack detection and analysis that is suited to small and mid-size businesses was created by R. Stoleriu et al. [7]. They have integrated the ELK stack with other open-source Threat Intelligence platforms and a strong security technology stack. To conduct real-time searches, they have integrated the ELK stack with the malware information sharing platform (MISP). To find anomalies in the network traffic, they also employed machine learning algorithms in Elasticsearch.

The ELK Stack network log system (NetFlow Log) has been built by Yang, CT et al. [8] to visually examine log data and show a number of network attack behaviour characteristics for additional investigation.

P. Sankar et al.[9] , showed how to use an ELK stack to analyze, process and drawinferences from logs obtained from social media networks like Twitter.

To address the web application security concerns of the campus network, J. Gong et al. [10] created a log security threat monitoring platform based on ELK. The platform can efficiently identify security threats in the campus network, visual association analysis is practical, and it can efficiently increase the effectiveness of operation and maintenance as well as the speed of security attack investigation, according to system operation results.

In a context of meteorological observation, Eugenio S. Almeida et al.[11] employed the ELK stack to index sensor data and metadata. They choose relevant time windows, capture, transform, enrich, store, index, and generate graphs that are integrated in a dashboard for combined display and analysis using the ELK stack.

The architecture model of a log management system using ELK Stack with Ceph was suggested by Yang C. T. et al. [12] in order to provide a secure network, strong Wi-Fi signal, and suitable backup data mechanisms. They claimed that by using the analyses'

findings, they might identify cyber attacks and respond appropriately.

Using the open source stack ELK, Prakash et al.[13] carried out the geo identification of users based on the access logs. They've also said that the ELK stack may be used totrack fraudulent behavior, security flaws, and it appears promising for monitoring the Internet of Things, which is the future.

4. Static vs. Dynamic Log Management

Let's now concentrate on the reasons why it is necessary to make ELK capable of operating in a completely dynamic environment for ongoing registered users by outlining the drawbacks of the static method and how they are resolved by using the dynamic approach.

Static Approach:

-It requires creating input/output configuration files manually.
-It is not triggering instantly and requires more time to get initial output at customerdashboard.
-Code syntax possibility is more.

Dynamic Approach:

-It does not require any manual intervention.
-All required configurations get created dynamically and instantly.
-Output will be generated instantly as total process is automated.
-Code syntax possibility is much less.

5. Methodology

The ELK ecosystem as shown in figure 2 comprises of three parts as referenced previously. Logs are caught from various servers these logs are gathered utilizing different logstash specialists and given to the logstash for indexing and sorting. Different logstash modules are utilized to drive the information into the logstash specialist. These logstash specialists pass the indexed data to the elasticsearch for searching and query processing. Client can surrender their inquiries to elasticsearch utilizing kibana through HTTP request and in the wake of handling the query the JSON response is made and given to kibana. This outcome would be shown on kibana. Result is essentially as graphical portrayal, for example, pie diagrams, bar charts, line charts, histogram because of which perusing and comprehension of information becomes simpler for the end clients. Various outcomes can be consolidated together to frame dashboards.

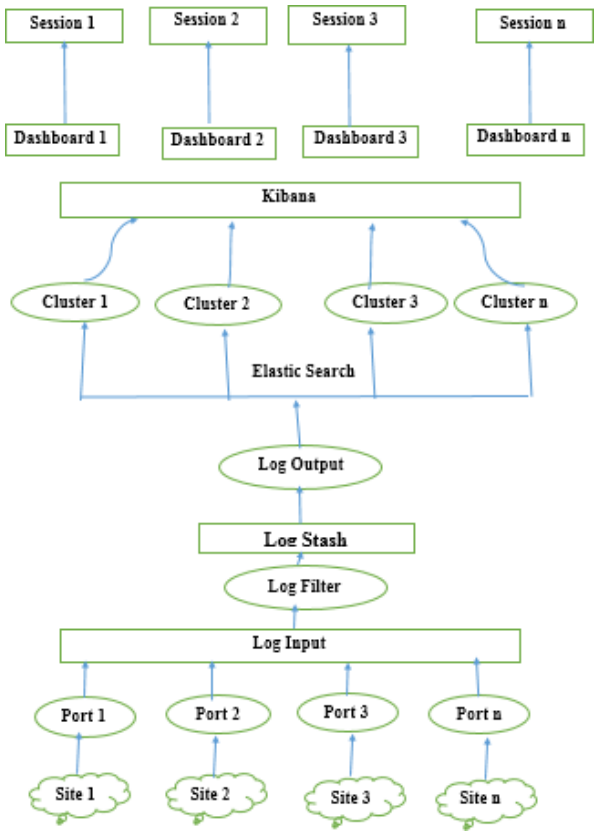

Fig.2. ELK Ecosystem

To use ELK stack between multiple users/Infrastructures with live applications requires customization at all three component levels. It needs customization in the whole data process from start point to end point.
We are listing customization from data start point (logstash) to end point (Kibana Visualization). We are using syslog for example and explanation.

Logstash

As mentioned, Logstash has 3 main components. To make parallel data flow alongmultiple sessions requires Logstash to start multiple data channels.

Structure of syslog-input.conf

```
input {
tcp {
port => 1514 type => syslog
}
udp {
port => 1514 type => syslog
}
}
```

- In this configuration Logstash has opened port 1514 for both tcp and udp data channels to collect data.

- In a multi-user (session) environment to collect data for different users Logstash needs to open multiple ports (one port per user(session)) to create individual data channels for each user.

Structure of syslog-output.conf

```
output {
elasticsearch {
hosts => ["localhost: 9200"]
index => "syslog-%{+yyyy.MM.dd}"
}
}
```

- As mentioned in the structure output file defines elastic search node addressand index.
- In a multi user (session) environment to push user wise data it need to bepushed in per user unique index format.

Implementation at Application

- Per user account a port needs to be reserved along with logstash nodeaddress.
- For the same user input as well output files need to be created dynamicallywith that bind port.

Example Input File:

```
input {
tcp {
port => 2021 type => syslog
}
udp {
port => 2021 type => syslog

}
}
```

Example Output File:

```
output {
elasticsearch {
hosts => ["192.168.1.51:9200"]
index => "user-syslog-%{+yyyy.MM.dd}"
}
}
```

For output file elastic search node addresses need to bind with the user account by defining a unique index pattern per user.

Elastic Search

- Logstash Output file will define user wise index pattern for parallel datachannels.
- Accordingly indexes will be created and those indexes will follow rules defined for the index pattern.

Kibana

- According to the index being created dynamically dashboard need to becreated.
- Dashboard will be mapped with a specific index.
- Dashboard can be displayed inside any other application inside an iframe orcan be shared.

6. Experimental Setup

Using the ELK stack, we created a solution for huge volume log analysis. As stated in Table I, we chose the hardware and software environment for the ELK stack for this investigation. According to hardware specifications, the CPU is a 3.40 GHz, 4-core processor. There are 16GB of memory and Ubuntu 22.04 installed. Elastic Search version 8.5.3, Logstash version 8.5.3, Kibana version 8.5.3, and Syslog Plug-in version: 3.6.0 is all current software versions.

Table 1. ELK Stack Installation Specification

Hardware specification	Software Specification
CPU – 4 core processor i7-4770 CPU @ 3.40GHz Memory : 16GB OS : Ubuntu 22.04	Elasticsearch-8.5.3.tar.gz Logstash-8.5.3.tar.gz Kibana-8.5.3-linux- x86_64.tar.gz Syslog Plugin-plugin_syslog-3.6.0.zip

7. Results

Here in this section we are representing some of the results of our implementation.
Let see first, how the syslog file looks. Figure 3 depicts the portion of the syslog file containing various fields like id, version, type, timestamp and many more. Original Syslog is in JSON format.


```
{
  "_index": "xyz-syslog-2022.09.19",
  "_type": "_doc",
  "_id": "y5gxWIMBzYD-FGDYgj8b",
  "_version": 1,
  "_score": 1,
  "_ignored": [
    "message.keyword"
  ],
  "_source": {
    "host": "14.143.172.226",
    "@timestamp": "2022-09-19T23:59:43.005Z",
```


Fig. 3. Syslog File Sample

Below figure 4 display syslog captured for a particular user account. Time duration is changeable to view and analyze logs received during specific duration.

Fig. 4. Syslog for user account

8. The Future Scope of Research Work

The future scope of research work in making dynamic pipeline in ELK is vast, and there are several areas where further investigation could lead to significant improvements in this field. Here are some possible future research directions:

- **Automation and orchestration:** One of the major challenges in managing dynamic pipelines is automation and orchestration. As the number of sources and pipelines grows, it becomes increasingly difficult to manage them manually. Future research could focus on developing more sophisticated automation and orchestration tools to manage dynamic pipelines more efficiently.
- **Machine learning-based approaches:** Machine learning can be used to improve the performance of dynamic pipelines. For example, anomaly detection algorithms can be used to identify unusual patterns in log data and trigger the creation of new pipelines dynamically. Future research could explore the potential of machine learning-based approaches to make dynamic pipelines more effective.
- **Integration with cloud-based services:** With the increasing adoption of cloud-based services, there is a need for dynamic pipelines that can integrate with these services seamlessly. Future research could focus on developing tools and technologies that enable dynamic pipelines to integrate with cloud-based services more effectively.
- **Security and privacy:** With the increasing concern about data privacy and security, there is a need to ensure that dynamic pipelines are secure and compliant with data protection regulations. Future research could focus on developing tools and techniques that enable dynamic pipelines to be more secure and privacy-compliant.

Overall, the future of research work in making dynamic pipeline in ELK is exciting, and there are many opportunities for innovation and improvement in this field. By addressing the challenges mentioned above and exploring new approaches, we can make dynamic pipelines in ELK more effective, efficient, and secure.

9. Conclusion

Log management and analysis are crucial for maintaining a reliable system. Easily combining and analyzing logs from many sources is a terrific method to encourage continuous improvement. The open source ELK Stack platform is the finest and quickest approach to create log analysis solutions. In this paper, we demonstrate how to leverage the ELK stack to manage numerous users' logs on a dynamic basis. When compared to expensive commercial products, the ELK stack can be a powerful beginning log analysis tool with acceptable performance.

References

1. ELK Stack, https://www.elastic.co/products
2. Sumitra Purushottam Pundlik, Bhupendra Moharil et. Al.,"Real Time Generalized Log File Management and Analysis using Pattern Matching and Dynamic Clustering" ,International Journal of Computer Applications 91(16):1-6, DOI: 10.5120/15962-5320, April 2014
3. What is Log Analysis and Why Do You Need It? A comprehensive Guide,https://www.xplg.com/what-is-log-analysis-and-why-do-you-need-it/
4. AbdelkaderLahmadi, Frédéric Beck. Powering Monitoring Analytics with ELK stack. 9th International Conference on Autonomous Infrastructure, Management and Security (AIMS 2015), Jun

2015,Ghent, Belgium. 9th International Conference on Autonomous Infrastructure, Management and Security (AIMS 2015), 2015. hal- 01212015

5. K. Moriyama , T. Nakatani , Y. Yasu , H. Ohshita and T. Seya,"development of status analysis system based on elk stack at j-parc mlf", 16th Int. Conf. on Accelerator and Large Experimental Control Systems, ICALEPCS2017, Barcelona, Spain JACoW Publishing, doi:10.18429/JACoW-ICALEPCS2017-THPHA033, ISBN: 978-3-95450-193-9

6. Sung Jun Son and Youngmi Kwon "Performance of ELK Stack and Commercial System in Security Log Analysis", IEEE 13th Malaysia International Conference on Communications (MICC), 28-30 Nov. 2017

7. R. Stoleriu, A. Puncioiu and I. Bica, "Cyber Attacks Detection Using Open Source ELK Stack," 2021 13th International Conference on Electronics, Computers and Artificial Intelligence (ECAI),2021, pp. 1-6, doi:10.1109/ECAI52376.2021.9515120.

8. Yang, CT., Chan, YW., Liu, JC. et al. Cyberattacks detection and analysis in a network log system using XGBoost with ELK stack. Soft Comput 26, 5143–5157 (2022). https://doi.org/10.1007/s00500-022-06954-8

9. P. Sankar, D. E. George and A. S. N. S, "Social media monitoring using ELK Stack," *2022 IEEE International Conference on Signal Processing, Informatics,Communication and Energy Systems (SPICES)*, 2022, pp. 231-235, doi: 10.1109/SPICES52834.2022.9774273.

10. J. Gong and S. Ling, "Security Analysis of University Web Log Based on Elk," 2021 International Conference on Intelligent Computing, Automation and Applications (ICAA), 2021, pp. 694-699, doi: 10.1109/ICAA53760.2021.00125.

11. Eugenio S. Almeida, Ivo Koga , Marcio A. A. Santana, Patricia L. O.,Guimaraes, Luciana M. Sugawara and Tero Eklin," Exploratory study of the ELK stack for meteorological observation system data analysis ", In Journal of Computational Interdisciplinary Science, 8(3),pp. 131-14,2017.

12. Yang, C. T., Kristiani, E., Wang, Y. T., Min, G., Lai, C. H., & Jiang, W. J. (2020). On construction of a network log management system using ELK Stack with Ceph. The Journal of Supercomputing, 76(8), 6344-6360.

13. Prakash, T., Kakkar, M., & Patel, K. (2016, January). Geo-identification of web users through logs using ELK stack. In 2016 6th international conference-cloud system and big data engineering (confluence) (pp. 606-610). IEEE.