OPTIMIZATION OF BIG DATA LOADING ON THE WEB THROUGH THE ELASTICSEARCH ENGINE

Andriavelonera Anselme A.[1], Rivosoaniaina Alain N.[1], Mahatody Thomas[2], Manantsoa Victor[2].

[1]Laboratory for Mathematical and Computer Applied to the Development Systems, University of Fianarantsoa, Madagascar,
[2]Professor on the University of Fianarantsoa, Madagascar,

ABSTRACT

Information retrieval is an issue that has gained prominence since the beginning of the era of digitalization. As the volume of data becomes increasingly large, it has become essential to be able to efficiently retrieve information. Together with the evolution of web technology, the number of users sharing online information by means of the search engine is growing faster. This article presents the optimization of big data loading by exploiting the power of the database management system "Elasticsearch". In this document, we present an approach to synchronize the traditional data processing method with the elasticsearch engine to improve the processing environment.

KEYWORDS

Elasticsearch, optimization, Database, data processing, big data.

1. INTRODUCTION

There are 2.5 trillion bytes of raw data per hour, structured or unstructured, according to IBM's 2014 annual report. In 2020, the estimate rises to more than 40 Zettabytes in the space of 6 years [1]. In our time, this explosion of data on the web has not stopped due to the most popular social networks. Users are the most active content creators without worrying about impact. Data from business systems processes and data generated by the Internet of Things are also major sources of big data. This offers new opportunities and new challenges to exploit the vast expanses of data generated on the web. The change is so great that even the IT tools used over the past decade are no longer able to meet new challenges.

1.1. Objective and Problematic of the Research

Faced with this new challenge, web search engines have become familiar and essential for Internet users. Their use has become commonplace in very diverse situations of daily life, whether in a professional or private context [2]. Engines sometimes work with unstructured or semi-structured data, which is why accessing useful information is increasingly difficult. There is a question whether the search engine system should have distributed processing capabilities, to accommodate the growth and improvement of the system's ability to process information. Guru's work in 2022 presents the reduction in time required to analyze big data. And they proposed an algorithm to optimize the extraction of big data in the Hadoop cluster. Although, this work was

David C. Wyld et al. (Eds): BDAP, AIS, NLPA - 2023
pp. 01-10, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.131101

limited on the evaluation of the time of sorting input data and to the study on the occupation of the used disk space [3]. Since the advent of new technologies and architectures such as "Big Data", the data has exploded. ElasticSearch is one of the tools based on the RESTfull search engine, designed for cloud computing. We need a search server available in real time and a simple multi-tenant cloud solution. Elasticsearch provides real-time, stable, reliable, fast, easy-to-install search for data support using indexes over http. It was designed to meet all these needs. How can we harness the environmental power of the Elasticsearch engine for processing large volumes of data?

1.2. Contribution

In this article, first, we propose the use of Elasticsearch to optimize data loading during real-time playback. Elasticsearch which is a complete search engine allowing full-text search. This type of technology is necessary when it comes to recovering unstructured data. In the context of optimization, the first aspect is to improve the processing time. The latter should be collected and processed before being loaded into the search engine. This step may take some time, depending on the type and size of the data we want to consult.

Secondly, we will address the relevance of Elasticsearch, which is a difficult question because it is specific to each individual.

2. STATE OF ART

In this section, we are going to address the search and visualization methods of information, the most commonly used search methods and the implementation of the ELK stack in Bigdata.

2.1. Finding and Viewing Information

The objective of information retrieval is to obtain relevant information to meet the needs of the user. The traditional approach has evolved over time to help users achieve the goal interactively, allowing them to participate in information-seeking processes [4].

Information seeking theory explains the processes associated with people seeking information [5]. An important part of this theory is the concept of informational scent, which describes a mechanism by which people interpret the information available to them to decide where to direct their efforts. An interactive image retrieval interface can be designed to enable information scent by providing visual cues that provide the researcher with information about the composition of the image collection and the ability to predict the results of their interaction with the image search interface [6]. Joachim realized the extensive inquiries in his article which assesses the perception of different aspects of research ethics. It showed with current situations and with an increased ethical responsibility, not only of the institutions, but above all, of the researchers themselves [7]. Information visualization deals with the presentation and communication of abstract data using graphical representations. A visual representation can be processed faster than a similar amount of text, using the human ability to perceive, interpret, and make sense of visual stimuli [8]. Even though precision may be lost when visualizing information, it allows similarities and relationships between a large amount of data to be assessed with minimal effort. In some of the previous research and investigations, the Internet and social networks resort to the need to constantly retrieve, analyze, synthesize, control and visualize relevant information. This has been constructive for data mining because of the vast social media platforms and applications of the social media that are then filed into web blogs, social networking sites, forums, podcasts, multimedia platforms, rating and review bookmarking sites and avatar-based virtual reality ones sites [9].

The means required to monitor the operation of data mining can be tedious due to the multiplicity of networked sites and the quantity and complexity of the data. A more critical reference is given to the set of data and metadata that have not been systematically addressed in the text mining literature [10]. An effective IoT-based data visualization framework, has been proposed to enhance leakage risk, multiple data source analysis and data quality management for business intelligence in corporate finance. The analysis results show 5 ms less response time and better revenue analysis with 29.42% improvement over existing models, which proves the reliability of the proposed framework [11]. Data mining is applicable to a wide range of domains on the Web, where large amounts of data are identified among unknown or hidden information patterns, depending on their availability. Therefore, this indicates that the data mining methods used on the Web are usually called Web mining. Similarly, those used in text are called textmining, while those used in libraries are called bibliomining [12]. A normal data mining process contains an interactive sequence of steps initiated by integrations of raw data obtained from various sources and then subjected to in-depth analyzes of duplicate or inconsistent data. Filtration and aggregation techniques are then incorporated for summary data extraction. Knowledge is acquired as the user obtains detailed information about data mining [9].

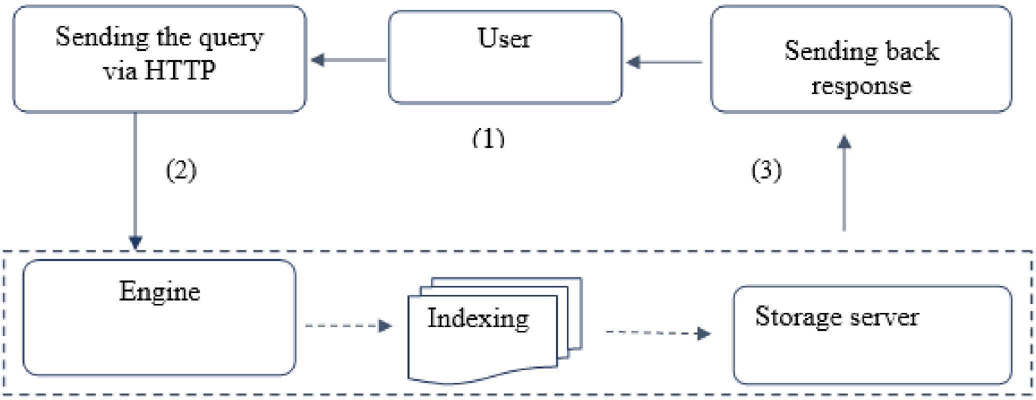

Figure 1. Research and information visualization on the web

(1) The user send his/her information request to internet via the HTTP query
(2) Then, the navigator transfers the query to the search engine. The latter effects the treatment and the data indexing to the server.
(3) The server sends back the risponse to the requester

2.2. Search Engines for Big Data Processing

The importance of data in the web world be best described as the Big Data approach. With millions of downloads, there are several search engine platforms that provide big data visualization:

2.2.1. Search Engine: Splunk

In his paper Jon Stearley, the author describes experiment with the application of the Splunk log analysis tool as a means of combining both data and people [14]. Splunk's search language, search, macros and sub-searches reduce tedious hours to seconds of simplicity, and its tags, saved searches and dashboards provide both operational information and collaborative vehicles.

Splunk's operation takes the flat as input, connecting it to a database via a connector, and then retrieving that data over the connection to extract events to add to the engine. Splunk Enterprise indexes the data that makes up the IT infrastructure to get data from websites, application, servers, databases, operating systems. Search is the primary way users browse data and save the search as a report to feed the dashboard. Most modern systems produce many diverse data sets. The lack of tools to efficiently store and correlate various data sets makes data mining for analytical insights tedious. Therefore, implementing the potentiality of Splunk to manage a semi-structured database proves to be useful for indexing, searching, and analyzing massive data sets [15]. Another work to show the effectiveness of splunk as a global operational data intelligence tool. It highlighted its usefulness for software developers, site reliability engineers, DevOps engineers, data scientists, security professionals and business analysts [16].

2.2.2 Search engine: Graylog

Graylog is a solution that uses the Elasticsearch document indexing engine and Mongodb as a database to store configuration and some other parameters. According to Roxana Daniela SUSTIC, Graylog is easy to maintain. It meets the needs of a centralized management of logs, and more suitable for security and alerting [17].

Graylog is a system that allows you to centralize, monitor and analyse logs, thus, logs from different servers and applications can all be viewed in on place. Emre GÜL used Graylog as a log management tool to pre-identify attacks against the system [18]. This application is preferred for high performance during log analysis.

In terms of functioning, it contains plugins, content packs, GELF libraries and more content created by Graylog developers and community members.

2.2.2. Search Engine: Elasticsearch

One of the first works by Bai in 2017, to use ElasticSearch to present a method for searching big data in real time. It used it to improve queries on conventional databases to make them faster before sending them to the server [19]. In 2019, Zhanglong Wang's work presents a shard optimization strategy of elasticsearch by data and performance analysis. They found the better performances, fragments are placed in nodes with better performance which are evaluated by linear weighting method and optimized load balancing strategy will migrate hot data to make the cluster load balanced [13]. This can reduce latency during the search. Here the search compatibility by providing users with fast search while ensuring high quality. The design and development of a system indexing group searches of scientific documents [20]. While Vidhya and al has implemented the ElasticSearch server to create a search engine to search, retrieve and download search documents stored in databases using the Django framework where the indexing process is performed by ElasticSearch [21]. Another paper discovered the challenge in allows users to quickly navigate to files in their own datasets as well as identify relevant files in shared or public datasets. The desirable qualities of searches is favorable due to the architecture of Elasticsearch [22]. In their work, Divya and al, have argued that the right way to evolve Lucene's capabilities is to move away from tools like SOLR and use a tool that is fully designed to work with terabytes of data [23]. Elasticsearch engine is a very reliable search engine and can provide relevant search results. It is also illustrated that it is able to handle very large amounts of data with short processing times.

Table 1: Summary of Search Engines

Search engine / Features	Splunk	Graylog	Elasticsearch
Dashboard	✓	✓	✓
Data collection	✓	✓	✓
Indexing	✓	Limited	✓
REST API Interface	✓	✓	✓
Out of order data	✓	-	✓
Integration and Plugin	✓	✓	✓
Licence	Paying	Free	Free

We have presented above the most used search engines. They each have dashboard of critical metrics and the ability to efficiently search large data sets. All three use the same indexing method except graylog which is limited. Graylog is designed from the outset for log management. However, graylog and elasticsearch are both open source search engines, while splunk is a paid tool.

2.3. Operation of the ELK Stack While Loading Big Data

To improve big data loading, we leveraged the ELK stack: Elasticsearch, Logstash, and Kibana in order to achieve optimal processing performance.

2.3.1. Elasticsearch

Neel Shah shows us the exploitation of the stack based on the configuration of Elasticsearch to efficiently decompose the analysis of large-scale, real-time text mining [24]. Elasticsearch's distributed search engine allows us to perform data indexing and search faster. The latter gives us advantages on searching for large volumes of data on the Web. Indeed for a search request, instead of examining all the documents, the application only checks the index of previously created and stored documents. These performances of the stack has been experimented with by Kiran Deshpande for small enterprises commercial use [25]. It can also process searches and remove duplicates. According to a study, ElasticSearch is effective at searching and detecting duplicates using indexing [27]. The performance and operation we just presented is validated by Tao Xie in his article. He shows that ELK has an open-source log analysis system. The system makes the data distribution, operation and maintenance better and more ardent, and yields a method and tool for data processing and visual presentation which can assist people to better grasp the information or make auxiliary decisions [30].

2.3.2. Logstash

The Logstash stack improves loading by exploiting the real-time data synchronization system. It provides collection, analysis and storage of log files. Processing at Logsatsh level is organized into one or more pipelines. In each pipeline, one or more plug-ins receive or collect data which are then placed in an internal queue. Logstash integrates, transforms and transfers large volumes of data regardless of the source and format. It allows you to transform unstructured data into structured data and facilitate the whole process of processing and loading. The Logstash stack also offers the ability to develop large-scale logs and network traffic data. A research presents us with a method for conceptualizing and developing a system for acquiring, analyzing, visualizing and correlating logs in real time in order to track and identify key security occurrences that might lead to security vulnerabilities. The author compare the performance of the new framework for big data analysis with log analysis platforms [26]. Modern applications produce large amounts of data in the form of logs and events to facilitate rapid diagnosis and fault mitigation. The inability to ensure data velocity faces many problems, such as difficult integration of heterogeneous data, poor analysis capability and visualization. Vlad-Andrei's work proposes an alternative based on data collected on the operation and maintenance of a certain server presents important information obtained by searching, summarizing and analyzing logs in the form of various graphs. The results show that the system can improve and make the data distribution, operation and maintenance more vivid, and provide a method and tool for data processing and visual presentation, which can help people to better grasp the information or make auxiliary decisions [28].

2.3.3. Kibana

All the data synchronized during the load is visualized in the Kibana web interface. Thanks to using the console, one of Kibana's development tools, we can compose queries and send them to the elasticsearch. We can edit, import and export saved objects directly from Kibana. A registered object can be a search, a visualization or an indexing model. In terms of security, the features in the Elastic Suite grant users the appropriate access rights. IT, operational, and business teams can all work together to those in charge of applications can rely on these features to manage well-meaning users and repel malicious actors. Data stored in elasticsearch is nearly secure. Dessislava Petrova-Antonova's article shows us the monitoring of patients with pneumonia and bronchitis. The author offers a software solution to visually analyze air quality data using the potential of Kibana. The implementation of Elasticsearch with Kibana result in a visualization of the presented data. The paper shows that the proposed solution provides more intuitive perception and valuable information through multi-perspective graphs of air pollution [29].

Table 2: Classification by utility of the ELK stack

ELK **Features**	**ElasticSearch**	**Logstash**	**Kibana**
Dashboard	-	-	✓
Data collection	-	✓	-
Indexation	✓	-	-
Interface REST API	✓	-	✓
Out of order data	-	✓	✓
Integration et Plugin	✓	✓	✓
Licence	Free	Free	Free

Above we have evaluated the performance of each of the stack based on its utilities. Only the Kibana stack has a dashboard that aggregates critical metrics while having the ability to efficiently search big data. The same indexing method is done by ElasticSearch before any visualization. Logstash is designed from the start for log management and log analysis. However, Kibana offers data visualization with a secure overview. All three tools are open source tools.

3. APPROCH

To make good use of the power on search operation in Elasticsearch, we will combine it with existing relational databases. In these cases, we will probably need to ensure that Elasticsearch stays in sync with the data stored in its associated relational database. Next, we'll use both Logstash and Symfony to efficiently copy records and synchronize updates from our database into Elasticsearch.

We have several database, of different structures, inherited with a large data set. We want to do real-time analysis and search of this data on our application developed with Symfony. We have the choice to use Elastic Stack to do the analysis and visualization in real time. To use the ELK stack, we must first migrate the existing data on mysql to Elasticsearch before doing any processing.

The overall architecture of our approach is illustrated in Figure 2 below:

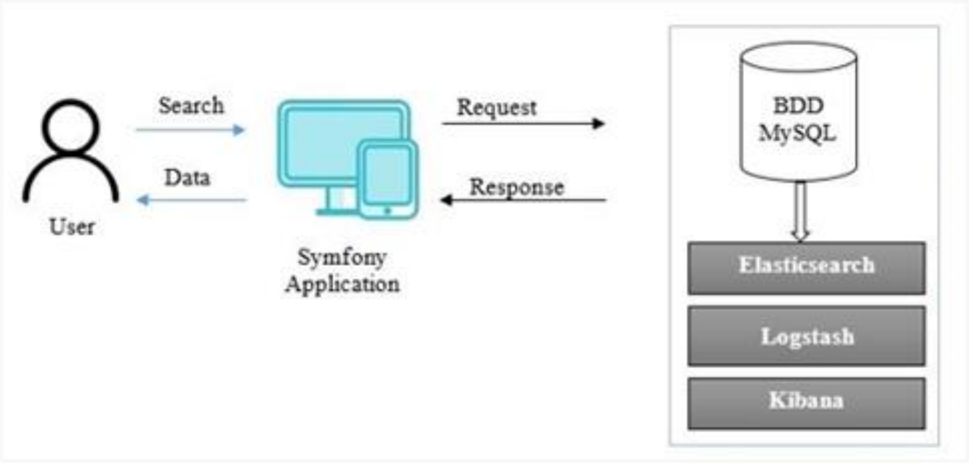

Figure 2: Database management architecture of the approach

First, we will use Logstash and the JDBC with input plug-in to synchronize Elasticsearch with MySQL. The idea is that the Logstash JDBC input plug-in runs a loop that regularly polls for MySQL records inserted or modified since the last iteration of the loop. For this to work properly, the following conditions must be met:

(1) Since MySQL documents are written to Elasticsearch, the “id” field in Elasticsearch must be set immediately to the “id” field from MySQL. This allows the MySQL record and the Elasticsearch document to be mapped directly. When a record is updated in MySQL, the associated document is completely replaced in Elastisearch.

(2) When a record is inserted or updated in MySQL, it must have a field containing the date and time of update or insertion date is more recent than the one received in the previous iteration of the query loop.

If the above conditions are met, Logstash can be configured to regularly query MySQL, for all modified or newly created records and then write them to Elastcsearch.
In the diagram, Logstash executes the configuration file that triggers the predefined query we defined to collect data from our interests in the sequential database. Once the query is run to the JDBC plug-in, this is transmitted to the database, and collects the data, which it will pass to logstash. Depending on the user's needs, the data can be processed and transformed into the desired format. After processing, the processed data is indexed to Elasticsearch.
Detailed System of the architecture:

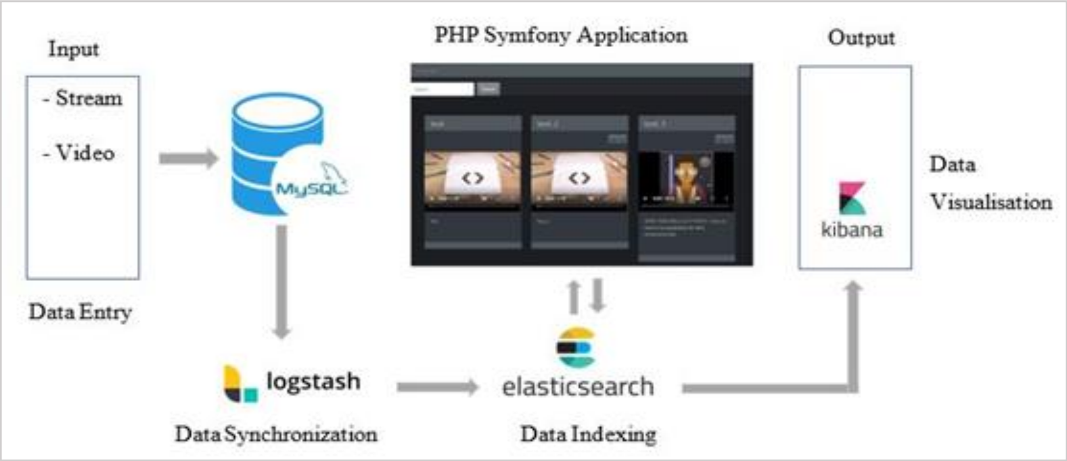

Figure 3: Elasticsearch and MySQL synchronization

To capitalize on Elasticsearch's powerful search capabilities, many organizations deploy it alongside their existing relational databases. In these cases, it is probably necessary to ensure that Elasticsearch stays in sync with the data stored in the associated relational database. In our case, we will use both Logstash and Symfony to efficiently copy records and synchronize updates from our database into Elasticsearch.

4. RESULTS

This section presents the main results.
The following results were obtained:

- Fast access to data: the contents of the document last consulted by users are automatically indexed. This reduces the number of data readings and, consequently, increases the response of search results.
- Handling huge amounts of data compared to traditional SQL. Database management systems that take more than 10 seconds to retrieve the required search query data, Elasticsearch can do it in microseconds (10, to be exact).
- System scalability: as the elasticsearch engine has a distributed architecture, it can handle multiple servers in parallel and petabytes of data. Users no longer have difficulties with managing the complexity of the distributed database management system. This is done automatically.

Minor results,

- Optimization of the number of queries made to retrieve data
- Simplified cache: using the server cache

5. CONCLUSION

In conclusion, many tasks were completed to understand the work required to establish a data pipeline and the difficulties of extracting information. At the beginning of the study, the data processing at the core of our MySQL databases lacked optimization to retrieve big data. Furthermore, the traditional tool was not designed to handle large amounts of very complex data. It was decided that the synchronization of this tool with the ELK stack would bring more improvement during the loading of the data on the web. Implementing both database tools allowed us to manage and track the loading of data into the Logstash pipeline at the same time, and gave us more advantage in terms of reading time through the search engine. Synchronization consumes a lot of hardware resources, but the main thing is to efficiently maintain control of the data flow and loading time. In this paper, we have presented the real time loading and synchronization of big data through the elasticsearch engine; and as perspectives, we intend to continue this work with the management of big database storage services.

REFERENCES

[1] Virginia M. Rometty, 2014, "IBM Annual Report", IBM.

[2] Savolainen R., 1995, "Everyday life information seeking: Approaching information seeking in the context of "way of life", Library & Information Science Research, 17(3), pp. 259-294.

[3] Guru Prasad M S, Naveen Kumar H N, Mohd Asif Shah,Raju K, Santhosh Kumar D & Chandrappa S, 2022, "FOEH: Frequent Pattern Mining Performance Optimization over Large Transactional Data in Extended Hadoop MapReduce", Research square.

[4] Belkin, N. J., 2016, "People, Interacting with Information1". ACM SIGIR Forum, 49, 13–27.

[5] Pirolli, P. & Card, 1999. "Information foraging. Psychological review", 106(4), 643.

[6] Chi, E. H., Pirolli, P., Chen & Pitkow, J. (2001). "Using information scent to model user information needs and actions and the Web". Proceedings of the SIGCHI conference on Human factors in computing systems, 490–497.

[7] Joachim Schöpfel & Otmane Azeroual, 2022, "Les systèmes d'information recherche: un nouvel objet du questionnement éthique", The Research Information Management Systems: A New Object of Ethical Questioning

[8] Ward, M. O., Grinstein, G., & Keim, D. (2015). "Interactive data visualization: Foundations, techniques, and applications". AK Peters/CRC Press.

[9] Agrawal, Divyakant, Bernstein, Philip, Bertino, Elisa, Davidson, S., Dayal, U., Franklin, M. & Widom, J., 2011, "Challenges and opportunities with Big Data", 2011-1, Purdue e-Pubs.

[10] Nour El Houda Ben Chaabene, 2022, "Detection of violent users and threats in social networks", HAL Id

[11] Cuili Shao, Yonggang Yang, Sapna Juneja & Tamizharasi GSeetharam, 2022, "IoT data visualization for business intelligence in corporate finance", Information Processing & Management, ELSEVIER

[12] N. Girija & S.K. Srivatsa, 2006. "A Research Study: Using Data Mining in Knowledge Base Business Strategies". Information Technology Journal, 5: 590-600.

[13] Zhanglong Wang & Yang Pi, 2019, "An Optimization Strategy of Shard on Elasticsearch", 4th International Conference on Automatic Control and Mechatronic Engineering (ACME 2019).

[14] Jon Stearley & Sophia Corwell, Ken Lord, "Bridging the Gaps: Joining Information Sources with Splunk." (SLAML)Sandia National Laboratories Albuquerque, 2010.

[15] Ratna Nayak, Purvi Sankhe, Shruti Mathur, Nishtha Mathur & Neha Patwari, 2019, "Splunk for Big Data Analytics"

[16] Karun Subramanian, 2020, "The Splunk Platform presentation", Apress.

[17] Roxana-Daniela SUSTIC, Alexandra MORARU, Andrei-Bogdan RUS & Virgil DOBROTA, 2022, "Performance evaluation of elk stack versus graylog as open-source log management tools", Acta technica napocensis Electronics and Telecommunications.

[18] Emre GÜL & Ercan Nurcan YILMAZ, 2019, "Log management with open source tools", ISAS, 3ème International Symposium on Innovative Approches in Scientific Studies.

[19] Bai, J., "Feasibility analysis of big log data real time search based on hbase and elasticsearch" Ninth international conference on natural computation (ICNC), 2013.

[20] Franz Frederik Walter Viktor Walter-Tscharf, 2022, "Indexing, clustering, and Search Engine for Documents Utilizing Elasticsearch and Kibana", Lecture Notes on Data Engineering and Communications Technologies (LNDECT, volume 126).

[21] Vidhya, R. & Vadivu, G., 2016, "Research Document Search using Elastic Search". Indian Journal of Science and Technology, 9(37).

[22] Jake Rosenberg, Josue Balandrano Coronel, Joseph Meiring, Sarah Gray & Tracy Brown, 2019, "Leveraging Elasticsearch to Improve Data Discoverability in Science Gateways", ACM Digital Library.

[23] Divya, M. S. & Goyal, S. K., 2013, "ElasticSearch: An advanced and quick search technique to handle voluminous data". Compusoft, 2(6), 171.

[24] Neel Shah, Darryl Willick & Vijay Mago, 2018, "A framework for social media data analytics using Elasticsearch and Kibana", Lecture Notes in Electrical Engineering (LNEE, volume 915)

[25] Kiran Deshpande & Madhuri Rao, 2022, "A Comprehensive Performance Evaluation of Novel Big Data Log Analytic Framework", Lecture Notes in Electrical Engineering book series (LNEE,volume 915).

[26] Sung Jun Fils & Youngmi Kwon, 2022, "Performance of ELK stack and commercial system in security log analysis", IEEE Xplore, ISBN.

[27] Shaik Subhani, Nalamothu Naga Malleswara Rao, 2020, "Improved Elastic Search and Efficient Duplicate Data Detection and Removal Using Ensemble Big Data Algorithms", ISSN 2277-8616.

[28] Vlad-Andrei Zamfir, Mihai Carabas, Costin Carabas & Nicolae Tapus, 2019, "Systems Monitoring and Big Data Analysis Using the Elasticsearch System", IEEE Xplore.

[29] Dessislava Petrova-Antonova, Stefan Baychev, Irena Pavlova & Georgi Pavlov, "Air Quality Visual Analytics with Kibana", IEEE Xplore.

[30] Tao Xie, Mingjiang Zhang, Zhe Wang & Linna Yang, 2022, "Research on intelligent analysis and visualization of big data based on ELK", 5th International Conference on Mechatronics and Computer Technology Engineering (MCTE 2022).

Author

ANDRIAVELONERA Anselme Alexandre, Laboratory for Mathematical and Computer Applied to the Development Systems, University of Fianarantsoa, Madagascar. Ten years of experiences in System and Network Administration. IT Instructor on the studies direction, INSCAE Madagascar