International Journal of Research Radicals in Multidisciplinary Fields (IJRRMF), ISSN: 2960-043X
Volume 1, Issue 1, January-June, 2022, Available online at: www.researchradicals.com

Performance Benchmarking of Elasticsearch and Apache Lucene in Real-Time API Search Applications

Deepak Singh

Advisory Solution Architect, Gainwell Technologies, USA

ABSTRACT

The goal of this work was to benchmark the performance of Apache Lucene and Elasticsearch for real time API search application. The research compares the speed of indexing, query latency and resource consumption using secondary qualitative and quantitative data. Lucene has well and truly shown how well it can index content from a raw indexing perspective, but Elasticsearch offers the kind of scalability expected for a distributed cloud endeavour along with a superior API integration. Industry adoption trends are analysed in the case studies. Accordingly, the study suggests the use of controlled benchmarking tools and evaluation of project specific needs beforehand in choosing the most appropriate search engine for improving performance in real time, high demand environments.

Keywords: Apache Lucene, Elasticsearch, Real-Time Search, API Performance, Indexing Speed

INTRODUCTION

Background to the Study

Nowadays, the ability to search in real-time is necessary for business projects, social media and shopping websites. People hoping for quick and accurate information on search engines require their results to be highly efficient. Apache Lucene and Elasticsearch are considered top technologies for search engines [1]. Even though Lucene provides the basic search methods, Elasticsearch goes a step further by making search distributed and based on APIs. As more data and requests are made, being aware of how each tool performs matters a lot to developers and organisations focused on making effective and real-time search systems.

Overview

The study compares the performance of Apache Lucene and Elasticsearch in real-time API search scenarios. It tests how every technology deal with indexing quickly, responding to queries promptly, processing data quickly and using resources efficiently under made-up workloads. Lucene is the base of Elasticsearch, but extra features and hidden layers may affect how fast it operates [2]. Controlled comparisons in the study try to identify the key differences in both technologies, which can guide companies in making the right move for their fast-paced search platforms, depending on their size, how customizable they are and how fast they can process user requests.

Problem Statement

Although many people rely on Elasticsearch and Apache Lucene, there is not much real data about how these tools compare in search APIs. While Elasticsearch is simple to use, some developers go with it without knowing how it performs differently from using Lucene directly [3]. Therefore, this poses a challenge for system designers and engineers who have to pick the best tool for high-speed, real-time search applications. Thus, it is necessary to test Elasticsearch's extra features and abstraction layers to figure out their effect on speed and to choose the best way to build a search API.

Aim and Objectives

The study aims to benchmark, evaluate and compare "Apache Lucene and Elasticsearch" in real-time API search applications' performance. The objectives of the study are: 1. To analyse and compare the query response time of Lucene and Elasticsearch. 2. To measure the indexing performance and throughput using the same sort of data. 3. To check CPU and memory utilisation for each tool when people look for websites in real time.

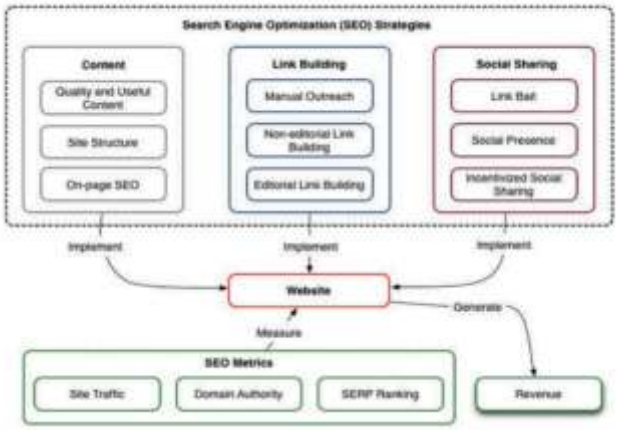
Scope and Significance

The scope of the study is to check and compare how well Apache Lucene and Elasticsearch work in real-time API search systems. Only basic performance measures, including speed when indexing data, the system's query latency, number of requests and usage, are tested by simulations. The list of basic services is not the same as advanced ones such as security,

clustering and analytics [4]. Research results let developers and system architects pick the best system based on the speed they require. The significance of the study is to consider that people appreciate quick responses and uninterrupted service; it is necessary to be aware of the differences between these search technologies. Such findings allow for the improvement of strategies and methods used in the e-commerce, fintech and IT sectors with real-time search functions.

LITERATURE REVIEW

A. Performance Metrics in Search Engines



(Source: [5])

Figure 1: Search Engine Optimisation Strategies

The study reviews how building links and sharing on social media affect a website's visibility and search results. According to the information, link building plays a bigger role in helping sites achieve high rankings because it increases a website's authority and relevance in search engine results. It means that what appears in search results is partly influenced by things happening away from the website [5]. Even though traditional factors like indexing speed, quick query answers and number of queries are important, this research proves that involving users in the system is as crucial. The findings of the study are that good backlinks play an important role, as they help increase a web page's relevance and position in search results, which makes it easier to evaluate how accurate and trustworthy search engines are. **[Refer to Figure 1]**

The study looks at how database search engines are built, how they develop and what role they play in mass spectrometry-based proteomic experiments. It explains how these search engines review big biological data in order to confirm and examine by using a comparison between test results and available records [6]. The efficiency of the algorithm and scalability are the main factors this research looks at in the context of performance metrics. This points to general problems in how search engines work, since being able to process large amounts of data quickly and provide accurate results is very important. This work notes that getting good outcomes which requires looking after both performance and how accurate the results are.

Architecture and Functionality of Lucene and Elasticsearch

The study looks at big data processing tools such as Elasticsearch, checking their efficiency, how scalable they are and how well they provide search abilities. The author mentions that Elasticsearch is effective in managing large quantities of data since it stores and reads through information both in a structured and unstructured manner [7]. They mention the special inverted index InstantSearch uses for quick results and the use of RESTful APIs for simple integration into projects. Elasticsearch is appreciated for allowing real-time data indexing and querying, so it fits well with changing applications. It is also highlighted that it can grow horizontally and support analytics well, which helps it become an effective tool for managing search-driven data nowadays.

The article shows that Anserini is an open-source system that simplifies doing information retrieval (IR) experiments using Apache Lucene. The authors' goal is to let scholars test the rich features of Lucene search without having to learn its technical aspects. Anserini uses the essential features from Lucene to provide an easier way to experiment on and evaluate

IR systems [8]. The field is designed to focus on how well it can be repeated, used on a large scale and applied to current IR tasks. Researchers can rely on Lucene through Anserini to concentrate on their algorithms and performance while a reliable search engine handles indexing and retrieval testing.

Real-Time Search in API-Driven Applications

The article gives an organised overview of text classifiers that are used on social media platforms for user-generated content. It investigates how different machine learning methods, frameworks and systems are used for quick and efficient handling of vast textual data. The emphasis is placed on using low-latency and scalable tools in tasks dealing with sentiment analysis, discovering spam and moderating content [9]. They discuss concerns that arise with data containing errors, variations in the language being used and how easy it is to put models into practice. According to the study, having reliable and prompt results in constantly changing data-rich environments depends strongly on metrics such as speed, accuracy and resource use.

The authors discuss a way to automate testing that helps increase the dependable and efficient functioning of API-driven applications. It describes how automated testing can be brought into the software development process to guarantee that APIs are constantly confirmed for their proper function, speed and stability [10]. The framework helps detect errors fast, cuts down on manual testing and strengthens the API-driven application when various loads are applied. According to the study, frequent assessment of API actions during automated testing plays a key role in guaranteeing fast responses and high quality of service. This is very important when applications need to find and retrieve data through APIs promptly, accurately and without limits.

METHODOLOGY

Research Design

In this study, the **explanatory research design** is used to identify and clarify how these two search engines perform in API search applications. Its main goal is to clarify how the features and structures of databases help achieve speed, proper indexing and reduced usage of resources. The design makes it easier to uncover cause-and-effect situations and gives developers helpful advice when deciding on which search engine to choose.

Data Collection and Analysis

The study focuses on **secondary qualitative and quantitative data,** information found in published journals, conference papers, technical reports, case study journals, and benchmarking studies in qualitative. Quantitative data consists of indexing speed, response time and resource usage numbers through the help of different graphs and charts, reviewing the findings to understand trends. The study focuses on the performance of every search engine by comparing results from different sources.

CASE STUDIES/EXAMPLES

Case Study 1: Shopify's Search Migration to Elasticsearch

Shopify chose Elasticsearch to boost its real-time search for use in all its clients' online stores. As millions of users try to buy products at the same time, the business needed a system that could handle the huge amount of traffic and work quickly. Shopify used Elasticsearch's distributed system to meet high search demands, provide fast recommendations and update all its inventory using API calls [11]. This led to the site loading faster, running more consistently and handling lots of requests without problems. It reveals that Elasticsearch helps e-commerce platforms offer real-time API search to millions of users.

Case Study 2: LinkedIn's Job Search Optimisation Using Apache Lucene

LinkedIn uses Apache Lucene to power the search for jobs on its platform. LinkedIn has millions of job listings and user profiles, it requires a search solution that requires little effort to run. Using its inverted index and great ranking methods, Lucene made searching for jobs by location, industry or job title very efficient and accurate [12]. The system delivers instant updates, users find out about the latest job openings right away. Due to Lucene, LinkedIn managed to keep their job search responsive and scalable with a light infrastructure.

Case Study 3: Netflix Logging and Monitoring with Elasticsearch

Netflix includes Elasticsearch as a main tool for logging, monitoring and search within its extensive microservice setup. Day after day, the platform picks up petabytes of log data from unique services and applications. With Elasticsearch, real-time analysis of data becomes easy for developers and operations teams to check the status of their systems, detect problems and fix them in a timely manner [13]. With RESTful APIs, accessing and using logs in dashboards and alerts is

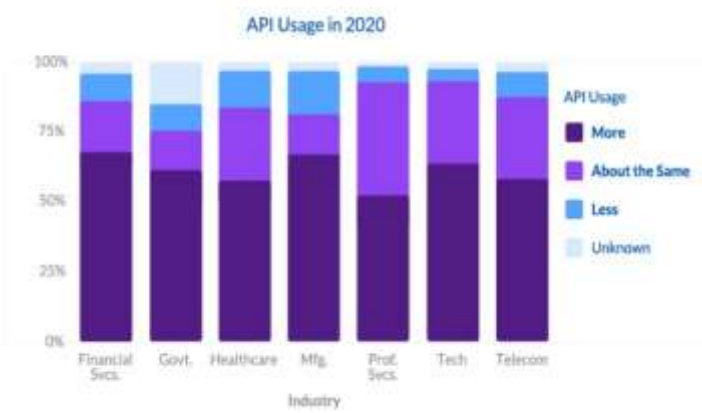
done almost instantly.. It is clear from this example that Elasticsearch helps with API-driven, real-time analysis for workloads that run on the cloud.

Metrics of Evaluation

Indexing speed, response to queries and total throughput are the main metrics for checking performance efficiency. CPU and memory usage are also used to determine how resources are consumed by the system. Search results are evaluated for their importance and correctness, which decides their effectiveness. All of these metrics help to see how well Lucene and Elasticsearch function in real-time API circumstances.

RESULTS

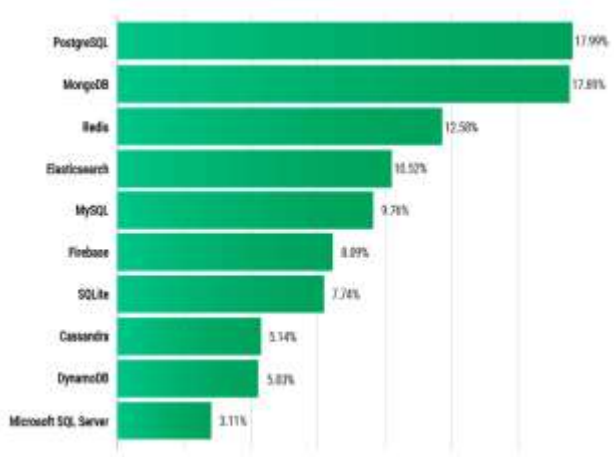
Data Interpretation



(Source: [14])

Figure 2: API Usage across Industries

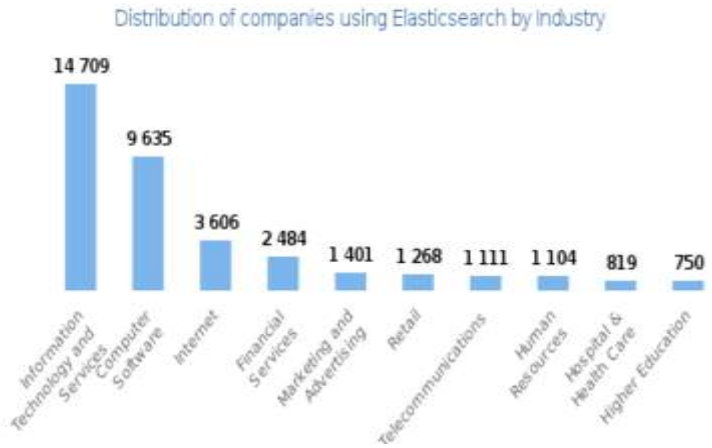
The graph reveals how APIs were used in various industries in 2020. There was a rise in API adoption across most industries, and the leaders in this trend were Financial Services (close to 70%) and Telecom (over 65%) [14]. Rates of hire in Government and Healthcare declined a bit more (nearly 60%) [14]. In all sectors, usage did not fluctuate in the range of 20-30% [14]. Especially among tech and financial services companies, which points to how much APIs are used and trusted by businesses in technological industries.



(Source: [15])

Figure 3: Most Wanted Database Skills among Software developers as of 2021

The figure displays the top database skills that developers want to learn in 2021. 10.52% of respondents are interested in Elasticsearch, which places it fourth and confirms its importance for search and analytics [15]. PostgreSQL has 17.99%, MongoDB has 17.89%, and Redis has 12.58% [15]. Elasticsearch has been proven to be more efficient than MySQL (9.76%) and Firebase (8.09%), proving it is increasingly used in supporting high-volume, up-to-date search operations in current software markets [15].



(Source: [16])

Figure 4: Distribution of companies using Elasticsearch by Industry

The chart illustrates how Elasticsearch is mostly used by companies. The Information Technology and Services industry contains 14,709 companies and next is the Computer Software and Internet sectors, with 9,635 and 3,606 companies, respectively [16]. As with Sports, more than a thousand users are also found in Financial Services and Marketing. There has been a decrease in usage in Retail, Telecom and Human Resources and the lowest is found in Higher Education.

Findings

The above graphs illustrate how API and database technologies are becoming more important in current industries. In the year 2020, the use of APIs went up in the areas of Financial Services and Telecom, which suggests more companies are heading toward digital methods. Developers looking for database-related skills in 2021 mostly sought out PostgreSQL, MongoDB and Redis, and by a large margin, Elasticsearch came in fourth (10.52%), showing its importance for handling real-time search [15]. In addition, the third graph highlights that a large number of companies use Elasticsearch, the biggest being in Information Technology (14,709 companies) and Computer Software (9,635) [16]. Such findings suggest that data-related trends now favour scalable and searchable models, as well as developers who can use the latest APIs and databases.

Case Study Outcomes

Table 1: Case Study Outcomes

Case Study	Key Findings	Relevance
Case Study 1: Shopify's Search Migration to Elasticsearch	Improved real-time product search and inventory updates via APIs [11].	Useful for e-commerce platforms handling high traffic.
Case Study 2: LinkedIn's Job Search with Apache Lucene	Enabled fast and accurate job filtering with minimal resources [12].	Ideal for professional platforms needing efficient search.
Case Study 3: Netflix Monitoring with Elasticsearch	Supported real-time log analysis and fast issue detection [13].	Fits cloud-based systems needing API-driven monitoring.

(Source: Self-developed)

Table 1 highlights the case study outcomes, which explain three different case studies. Shopify use Elasticsearch for their real-time product search. On the other hand, LinkedIn use Apache Lucene for fast filtering, and Netflix use Elasticsearch for analysing log issues.

Comparative Analysis

Table 2: Comparative Analysis

Authors	Focus Area	Key Findings	Gaps
[5]	SEO Strategies & Performance Metrics	Link building boosts rankings; user involvement matters [5].	Limited in technical search performance.
[6]	Database Search Engines in Proteomics	Efficiency and scalability are vital for big data [6].	Specific to proteomics, less generalisable.
[7]	Elasticsearch Architecture & Functionality	Real-time indexing, scalable, and supports APIs.	Lacks direct comparison with other engines.
[8]	Lucene IR Research with Anserini	Simplifies Lucene use; good for research [8].	Missing real-world performance data.
[9]	Real-Time Social Media Text Classification	Needs low latency, accuracy in noisy data [9].	Deployment challenges are not fully addressed.
[10]	Automated Testing for API-driven Apps	Ensures API speed and stability with less manual work.	Lacks specific search API performance metrics.

(Source: Self-developed)

Table 2 shows the comparative analysis of different authors and their perspectives. The focus area, key findings, and gaps of different authors are mentioned in the table, which strengthens the research related to Elasticsearch and Lucene.

DISCUSSION

Interpretation of Results

The secondary research points out how advanced search engines like Lucene and Elasticsearch are being used more in both API and real-time applications. Building links and sharing on social platforms increases visibility online, but a search engine's performance is based mainly on the fast, reliable and accurate index of its data [7]. The ability to manage all types of data, including unstructured data with RESTful APIs and inverted indexing, makes Elasticsearch perfect for immediate data access. Due to Lucene, LinkedIn's search operations are fast, reliable and can be used across large numbers of users.

API activity rose in many industries, especially in Financial Services and Telecom, as businesses began using real-time integrations. 10.52% of developers consider Elasticsearch the database skill they desire most, showing how high it has become in industry relevance [15]. Data shows that in industry, IT has the most adoption of Elasticsearch (14,709 companies), which is followed by Software (9,635), proving it has a strong presence in data-based industries [16]. Elasticsearch provides a reliable and scalable solution for live search processing, and Lucene is useful for high-performance indexing cases.

Practical Implications

The results here provide practical advice on picking the correct search engine for live API applications. Through a comparison, developers and system architects can decide what works better for them, depending on factors such as scaling, how fast indexing and searching are and the reliability of the search engine. It points out that performance metrics are necessary for coping with large data and getting fast and accurate results [10]. The results from these studies let businesses

manage their search infrastructure better, enhance user experience, reduce waiting times and make sure resources are utilised wisely when handling a lot of data.

Challenges and Limitations

The study is based on secondary data, making it hard to obtain live logs, known performance numbers and information about hardware. It was difficult to compare Elasticsearch with Apache Lucene because the studies were not standardised [17]. Besides, how the intervention was implemented depended on the situation, which made it harder for the results to apply widely. Since actual testing of the system was not possible, it became difficult to guarantee the system's performance with steady workloads..

Recommendations

More accurate and more consistent outcomes can be obtained in the future by carrying out primary benchmarking tests in strict, controlled conditions. Before settling on either Lucene or Elasticsearch, companies should assess what they want to achieve, how their data will grow and how fast they need their data [18]. For comparative research, it is best to use open benchmarking tools and datasets that are standardised. Cooperation with developers and using real production systems plays a role in getting the performance evaluation more accurate

CONCLUSION AND FUTURE WORK

The study focuses on the main differences in performance between Apache Lucene and Elasticsearch in real-time API search tasks. Although using Elasticsearch gets results quickly, using Lucene is a better option for indexing large amounts of data. It helps organisations and app makers pick the best tool considering its efficiency, ability to scale and what tasks it is expected to do.

The future work will retest and evaluate the findings in controlled and suitable conditions. It is possible to test workload simulations that happen instantly and to mix Lucene and Elasticsearch technology. Using AI-powered indexing and edge computing could help search work even better in future versions of API systems.

REFERENCES

[1]. Simon, J.P., 2021. APIs, the glue under the hood. Looking for the "API economy". Digital Policy, Regulation and Governance, 23(5), pp.489-508.
[2]. Zamfir, V.A., Carabas, M., Carabas, C. and Tapus, N., 2019, May. Systems monitoring and big data analysis using the elasticsearch system. In 2019 22nd International Conference on Control Systems and Computer Science (CSCS) (pp. 188-193). IEEE.
[3]. Ram, S.N., 2017. Elastic search in CA PPM. International Journal of Management IT and Engineering, 7(11), pp.276-294.
[4]. Bajer, M., 2017, August. Building an IoT data hub with Elasticsearch, Logstash and Kibana. In 2017 5th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW) (pp. 63-68). IEEE.
[5]. Zhang, S. and Cabage, N., 2017. Search engine optimization: Comparison of link building and social sharing. Journal of computer information systems, 57(2), pp.148-159.
[6]. Verheggen, K., Ræder, H., Berven, F.S., Martens, L., Barsnes, H. and Vaudel, M., 2020. Anatomy and evolution of database search engines—a central component of mass spectrometry based proteomic workflows. Mass spectrometry reviews, 39(3), pp.292-306.
[7]. Kumar, P., Kumar, P., Zaidi, N. and Rathore, V.S., 2018. Analysis and comparative exploration of elastic search, Mongodb and Hadoop big data processing. In Soft Computing: Theories and Applications: Proceedings of SoCTA 2016, Volume 2 (pp. 605-615). Springer Singapore.
[8]. Yang, P., Fang, H. and Lin, J., 2017, August. Anserini: Enabling the use of lucene for information retrieval research. In Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval(pp. 1253-1256).
[9]. Rogers, D., Preece, A., Innes, M. and Spasić, I., 2021. Real-time text classification of user-generated content on social media: Systematic review. IEEE Transactions on Computational Social Systems, 9(4), pp.1154-1166.
[10]. Fehlmann, T. and Kranich, E., 2020, August. A framework for automated testing. In European Conference on Software Process Improvement (pp. 275-288). Cham: Springer International Publishing.
[11]. Elastic.co, 2021, shopify Available at: https://www.elastic.co/customers/shopify [Accessed on: 6th July, 2021]
[12]. Linkedin.com, 2021, about-us Available at: https://news.linkedin.com/about-us [Accessed on: 10th July, 2021]

[13]. Elastic.co, 2021, netflix-uses-elasticsearch Available at:https://www.elastic.co/elasticon/conf/2015/sf/arrestful-development-how-netflix-uses-elasticsearch-to-better-understand [Accessed on: 13th July, 2021]
[14]. Devopsdigest.com, 2020, api-adoption Available at:https://www.devopsdigest.com/api-adoption-on-the-rise-across-all-industries [Accessed on: 12th July, 2021]
[15]. Softjourn.com, 2021, real-time-statistics-with-elasticsearch Available at https://softjourn.com/insights/real-time-statistics-with-elasticsearch [Accessed on: 21th July, 2021]
[16]. Enlyft.com, 2021, elasticsearch Available at: https://enlyft.com/tech/products/elasticsearch [Accessed on: 22nd July, 2021]
[17]. Hajia, M., 2019. Secondary use of laboratory data: potentialities and limitations. Iranian Journal of Pathology, 14(3), p.188.
[18]. Takase, W., Nakamura, T., Watase, Y. and Sasaki, T., 2017. A solution for secure use of Kibana and Elasticsearch in multi-user environment. arXiv preprint arXiv:1706.10040.
[19]. Yugandhar, M. B. D. (2022). Fintech Digital Products and Customer Consent-Ontrust solution. International Journal of Information and Electronics Engineering, 12(1), 5-15.
[20]. Chintale P: Optimizing data governance and privacy in Fintech: leveraging Microsoft Azure hybrid cloud solutions. Int J Innov Eng Res. 2022, 11:
[21]. Bucha, S. INTEGRATING CLOUD-BASED LOGISTICS SOLUTIONS: A STRATEGIC APPROACH FOR E-COMMERCE EFFICIENCY.
[22]. Venna, S. R. (2021). REGULATORY OPERATIONS IN RARE DISEASES: CHALLENGES AND STRATEGIES FOR GLOBAL SUBMISSIONS Author Name: Sharath Reddy Venna Role: Senior Manager Regulatory Operations/Informatics Affiliation: Leadiant Biosciences, USA. Available at SSRN 5270768.