Elasticsearch 8 for Developers

2nd Edition

A beginner's guide to indexing, analyzing, searching, and aggregating data

Anurag Srivastava

www.bpbonline.com

First published: 2021
Second published: 2024

Published by BPB Online
WeWork
119 Marylebone Road
London NW1 5PU

UK | UAE | INDIA | SINGAPORE

ISBN 978-93-55519-825

www.bpbonline.com

Dedicated to

My beloved parents:

Sri Virendra Nath Srivastava

Smt Kiran Srivastava

&

My wife ***Chanchal***

and

kids ***Anvit*** *&* ***Aditri***

About the Author

Anurag Srivastava boasts over 16 years of experience in the IT industry, marked by successful collaborations with various service-based companies on government and private projects. He excelled in pivotal roles encompassing development and technical management, showcasing his adaptability and expertise in achieving project success.

Anurag's accomplishments extend to the realm of literature, with several highly regarded books to his name. His publications are widely respected for their comprehensive and practical approach to Kibana and Elasticsearch.

In addition to his writing achievements, Anurag has delivered corporate training programs focusing on Elasticsearch, Kibana, ELK, and Cumulocity IoT, enabling organizations to leverage these technologies effectively.

With a diverse skill set, strong leadership qualities, and a passion for innovation, Anurag Srivastava brings substantial value to IT initiatives. He remains committed to delivering exceptional results while staying abreast of industry trends and advancements. Anurag is a trusted expert in the IT landscape, poised to contribute to technological excellence.

About the Reviewer

Jeeva is the Co-Founder of Epsil Technologies Private Limited, a company that specializes in providing top-quality support and solutions for Elasticsearch and its ecosystem. He has assembled a close-knit group of highly skilled Elasticsearch engineers, each with a range of certifications, and has brought them under the Epsil umbrella.

He has been supporting other companies in harnessing the power of Elastic through various use cases, including enterprise search systems, monitoring solutions, cluster optimizations, and version upgrades.

He possesses extensive experience in designing, creating, and troubleshooting Elasticsearch clusters in various environments, including on-premises, cloud, and **Elastic Cloud Enterprise** (**ECE**) setups.

Thanks to his wealth of experience, he not only comprehensively understands Elastic (and the ELK stack) but also takes great pleasure in staying well-informed about the latest versions and features within the Elastic Stack.

He has also been awarded Bronze Elastic Contributor for 2022 honouring his contributions towards the product and its community.

His world revolves around his beloved wife, Sandhiya, and their two exceptional children, Dhiyara and Aadhira, and he treasures every moment they share together.

Acknowledgement

I would like to express my deepest gratitude to my family and friends for their unwavering support and encouragement throughout the process of writing this book. Special thanks to my parents, my wife Chanchal, and my kids, Anvit and Aditri, for their patience and understanding during this journey.

I am also grateful to BPB Publications for their guidance and expertise in bringing this book to fruition. It was a long journey of revising this book, with valuable participation and collaboration of reviewers, technical experts, and editors.

I would also like to acknowledge the valuable contributions of my colleagues and co-workers, who have during many years of my working in the tech industry, taught me so much and provided valuable feedback on my work.

Finally, I would like to thank all the readers who have taken an interest in my book and for their support in making it a reality. Your encouragement has been invaluable.

Preface

This book is tailored towards a wide range of professionals, including developers, architects, database administrators, DevOps engineers, and other readers who are interested in efficiently learning Elasticsearch and how to apply it in their applications - whether new or existing. It particularly benefits those who wish to work with data using Elasticsearch.

It is recommended that readers have basic computer knowledge, as well as a familiarity with JSON and REST, before starting this book. Additionally, this book covers Elasticsearch alongside an introduction to other tools within the Elastic Stack.

No prior knowledge of Elasticsearch is necessary, as this book is designed to start with the basics and gradually progress to advanced topics in a practical and understandable way. With this approach, anyone can easily grasp the concepts presented in this book.

By the end of this book, you will have a deep understanding of Elasticsearch and be able to use it to manage and extract valuable insights from vast amounts of data. This book provides a comprehensive guide to managing data using Elasticsearch, making it an essential resource for developers, data analysts, and anyone else who works with data. I hope you will find this book informative and helpful.

Chapter 1: Getting Started with Elasticsearch - It provides an overview of Elasticsearch and its features. This chapter by introducing the concept of search and analytics and why they are important in today's data-driven world. It then goes on to explain how Elasticsearch works. Various use cases where Elasticsearch can be applied, and its benefits are covered. It concludes with a brief history of Elasticsearch and its evolution over time. By the end of this chapter, readers will have a good understanding of what Elasticsearch is, its purpose, and how it fits into the broader context of data analytics.

Chapter 2: Installing Elasticsearch - It provides a comprehensive guide to installing and configuring Elasticsearch on various operating systems. The prerequisites needed to install Elasticsearch, installation methods, and how to configure Elasticsearch for optimal performance are covered. Moreover, the instructions on how to verify the installation and configuration are included. Readers will have a working installation of Elasticsearch on their chosen platform.

Chapter 3: Elastic Stack: The Ecosystem of Elasticsearch - It provides an overview of the Elastic Stack, which includes Kibana, Logstash, and Beats. It explains how these components work together to provide a complete data analytics solution. Readers will learn about the role of each component and how they can be used to create powerful dashboards, visualize data, and ingest data from various sources.

Chapter 4: Preparing Data for Indexing - Here, we delve into the various steps in preparing data for indexing. This includes exploring different types of analyzers, normalizers, tokenizers, token filters, and character filters that can be used in Elasticsearch to preprocess the data. We will go through practical examples of how to apply these techniques to different types of data sources, such as text files, web logs, and structured data from databases. You will have a solid understanding of optimizing your data for efficient and effective indexing in Elasticsearch.

Chapter 5: Importing Data into Elasticsearch - You will learn how to import data from various sources like relational databases, CSV files, and more into Elasticsearch. You will also learn how to transform and preprocess the data using Logstash and Beats, two important components of the Elastic Stack. Additionally, you will explore how to handle errors and monitor the ingestion process for efficient data processing.

Chapter 6: Index Management: Creating, Updating, and Deleting Elasticsearch Indices - It covers the fundamental aspects of Elasticsearch index management, including how to create, update, and delete indices. It explains the different data types and mappings and how to define and manage these elements. Additionally, it explores techniques for index maintenance, including configuring shard allocation and implementing index lifecycle management policies.

Chapter 7: Search Capabilities: Mastering Query DSL and Search Techniques - We will learn about the search capabilities of Elasticsearch. You will understand Query DSL and different search techniques that Elasticsearch offers. You will learn to write complex queries to retrieve specific information from the indexed data. Additionally, you will learn about search optimization techniques such as pagination, sorting, and highlighting.

Chapter 8: Handling Geo with Elasticsearch - This chapter focuses on how Elasticsearch can be used for geospatial search and analysis. It covers geospatial data types, geo queries, filtering by location, geospatial aggregations, and geospatial mapping.

Chapter 9: Analyzing Data with Elasticsearch Aggregations - We will learn how to use Elasticsearch aggregations to analyze and summarize your data. Aggregations allow you to perform complex calculations and generate insights from your data, such as finding the average value, the maximum or minimum value, or the most common value for a particular field. We will cover different aggregations like metric, bucket, and pipeline aggregations and how they can be combined to get even more powerful insights.

Chapter 10: Performance Tuning - We will learn about optimizing Elasticsearch performance for large-scale data. You will explore hardware and network considerations, memory management, shard allocation, and indexing performance. You will also learn about various tools and techniques to monitor and diagnose performance issues and how to configure Elasticsearch to scale horizontally.

Chapter 11: Administration: Managing Elasticsearch Clusters - We will learn about the administration of Elasticsearch clusters, including managing and scaling Elasticsearch clusters. Topics covered include cluster management, node management, shard allocation, and scaling Elasticsearch clusters. You will also learn about backup and restore strategies, security features, and monitoring Elasticsearch clusters.

Coloured Images

Please follow the link to download the
Coloured Images of the book:

https://rebrand.ly/m12niro

We have code bundles from our rich catalogue of books and videos available at **https://github.com/bpbpublications**. Check them out!

Errata

We take immense pride in our work at BPB Publications and follow best practices to ensure the accuracy of our content to provide with an indulging reading experience to our subscribers. Our readers are our mirrors, and we use their inputs to reflect and improve upon human errors, if any, that may have occurred during the publishing processes involved. To let us maintain the quality and help us reach out to any readers who might be having difficulties due to any unforeseen errors, please write to us at :

errata@bpbonline.com

Your support, suggestions and feedbacks are highly appreciated by the BPB Publications' Family.

Piracy

If you come across any illegal copies of our works in any form on the internet, we would be grateful if you would provide us with the location address or website name. Please contact us at **business@bpbonline.com** with a link to the material.

If you are interested in becoming an author

If there is a topic that you have expertise in, and you are interested in either writing or contributing to a book, please visit **www.bpbonline.com**. We have worked with thousands of developers and tech professionals, just like you, to help them share their insights with the global tech community. You can make a general application, apply for a specific hot topic that we are recruiting an author for, or submit your own idea.

Reviews

Please leave a review. Once you have read and used this book, why not leave a review on the site that you purchased it from? Potential readers can then see and use your unbiased opinion to make purchase decisions. We at BPB can understand what you think about our products, and our authors can see your feedback on their book. Thank you!

For more information about BPB, please visit **www.bpbonline.com**.

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

https://discord.bpbonline.com

Table of Contents

Chapter 1
Getting Started with Elasticsearch

Introduction

In this chapter, we will provide an in-depth introduction to Elasticsearch. We will start by discussing the benefits of using Elasticsearch and how it can help businesses achieve their data management goals. From there, we will delve into what Elasticsearch is and how it leverages the powerful search engine Lucene to provide fast and scalable search capabilities. To lay a solid foundation for understanding Elasticsearch, we will cover some basic concepts such as nodes, clusters, documents, indices, and shards. These concepts are essential for understanding how Elasticsearch stores and organizes data for efficient search and retrieval.

We will then explore some of the key use cases for Elasticsearch, including data search, logging and analysis, application and system performance monitoring, and data visualization. These use cases highlight the versatility of Elasticsearch and demonstrate its potential to provide insights and valuable information across a wide range of industries and applications. Additionally, we will discuss the various Elasticsearch clients available for developers, such as Java, PHP, Perl, Python, .NET, and JavaScript. These clients enable developers to leverage Elasticsearch's search capabilities in their preferred programming language and ecosystem.

Finally, we will discuss how to use Elasticsearch as a primary data source, secondary data source, or as a stand-alone system. We will provide guidance on how to make informed decisions about incorporating Elasticsearch into your data architecture based on your specific business needs and technical requirements.

Structure

In this chapter, we will discuss the following topics:

- Introduction to data search
- What is Elasticsearch, and why is it important for search and analytics
- Overview of Elasticsearch architecture and components
- Applications and use cases for Elasticsearch
- Different Elasticsearch clients and their usage scenarios

Objectives

This chapter provides an overview of Elasticsearch and its features. It starts by introducing the concept of search and analytics and why they are important in today's data-driven world. It then goes on to explain how to utilize different Elasticsearch clients effectively.

Introduction to data search

In the modern world, the exponential growth of digitized data from various sources like smart devices, IoT sensors, and online transactions presents a significant challenge. One of the major challenges is converting unstructured data into a structured form to streamline the data storage process. However, the real challenge lies in searching the stored data for relevant information. Traditional data storage systems like RDBMS are not suitable for text search due to their complex SQL query writing process and search inefficiency, even after applying all required indexes. In contrast, Elasticsearch, a search engine built on top of Lucene, offers a sophisticated search mechanism with search relevancy, data aggregation, and many other benefits not available in RDBMS systems. Therefore, understanding the importance of search and how Elasticsearch can help streamline data storage and search is crucial for any organization dealing with large amounts of data.

Search is a critical component of modern-day applications as it enables users to quickly and accurately find the information they need. Whether a blog site, e-commerce platform, or any other application dealing with large volumes of

data, a search mechanism is essential to provide users with relevant results. The importance of providing quick and accurate search results cannot be overstated, as users are more likely to abandon an application that does not meet their search expectations. Therefore, optimizing search performance is crucial to ensure a positive user experience and retain user engagement.

Apart from providing relevant and speedy search results, there are other critical aspects of search that need to be considered, such as search relevance, data aggregation, and analysis. These aspects can be effectively addressed by Elasticsearch, which is a powerful and scalable search engine capable of handling a variety of data types and sources. By leveraging Elasticsearch's capabilities, applications can provide users with fast, accurate, and relevant search results, making it a critical tool for modern-day search-oriented applications.

The importance of search functionality cannot be overstated, as it enables users to quickly and accurately find the information or products they are seeking. In addition to providing a quick response time with relevant results, there are several other aspects of the search that must be considered, such as:

- **Search suggestion**: An effective search system should suggest potential search terms as soon as a user starts typing, allowing for quick and efficient search queries.
- **Fuzzy data searching**: The system should also be able to suggest relevant results even if the user misspells a search term or uses a synonym.
- **Derivative search**: A high-quality search system should recognize derivatives of search terms, such as plural or singular versions, to provide the most comprehensive results.
- **Data aggregation**: The system should support data aggregation to display additional options and filters to users, such as price range, ratings, brands, and other relevant information.
- **Relevant results**: Search results should be displayed in order of relevance, taking into account factors such as search term frequency, recency, and user behavior.
- **Advanced filters**: Users should be able to apply advanced filters to their search results, such as screen resolution, RAM capacity, color, and other relevant criteria.
- **Quick response time**: A search system should provide search results quickly, within a matter of seconds, to ensure a smooth user experience and avoid user frustration.

By considering these aspects, developers can create effective search systems that provide quick and accurate results to users, ultimately leading to increased user engagement and satisfaction.

What is Elasticsearch, and why is it important for search and analytics

Elasticsearch was created by *Shay Banon*, the founder of Elastic, a company that develops and supports Elasticsearch. Elasticsearch is open-source software that can be run on a single server or distributed across hundreds of servers to handle petabytes of data without any issue. Elasticsearch is a powerful search engine that is used to search for relevant data from a large data store.

In the current information age, the amount of data is growing exponentially due to digitization and the emergence of new data sources like smart devices, IoT sensors, and online transactions. These data can be structured or unstructured, device-specific or time-series data, and come from different sources, which makes it difficult to search through them manually. To overcome these challenges, Elasticsearch provides a distributed, scalable, and document-oriented search engine that is built on top of the Lucene library. Lucene is a high-performance search engine library that provides fast and efficient search results. However, it requires complex Java code to use and is not easily distributable across multiple nodes.

Elasticsearch encapsulates the complexities of Lucene and provides REST APIs that allow users to interact with Elasticsearch in a more user-friendly way. Elasticsearch also provides support for multiple programming languages through language clients, so users can code in their preferred language and still interact with Elasticsearch. Additionally, Elasticsearch can be interacted with using the command-line tool cURL.

In summary, Elasticsearch is a powerful search and analytics engine that provides fast and efficient search capabilities on large volumes of data, making it a vital tool for organizations looking to derive insights and value from their data.

Overview of Elasticsearch architecture and components

Elasticsearch is designed with a distributed architecture that allows it to handle large amounts of data across multiple nodes. It is composed of several components

that work together to provide a scalable and highly available search and analytics platform.

Node

In Elasticsearch, a node refers to a discrete running instance of the search engine. Elasticsearch is composed of one or more nodes, which are instances of the Elasticsearch server. For instance, in a cluster of 10 servers running Elasticsearch, each server would be considered a node. In some use cases, a single node cluster of Elasticsearch may suffice for non-production environments. However, as data size increases, the need for additional nodes arises to horizontally scale the cluster, which also provides fault tolerance. Through knowledge of other nodes within the cluster, a node can transfer client requests to the appropriate node. It is worth noting that nodes can take on various roles, including data nodes that store and execute queries, master nodes that manage cluster-wide operations, and coordinating nodes that forward requests to the appropriate nodes. Each node runs independently and communicates with other nodes to form a cluster. Nodes can be added or removed from a cluster dynamically without affecting the overall system. Nodes can be of different types:

Master-eligible node

In Elasticsearch 8, the master-eligible node is responsible for managing the cluster state, including adding or removing nodes, allocating shards to nodes, and maintaining the health of the cluster. It is recommended to have at least three master-eligible nodes in the cluster to ensure high availability and avoid split-brain situations*.

Dedicated master-eligible node

A dedicated master-eligible node is a node in an Elasticsearch cluster that is configured to be eligible for the role of a master node but is not tasked with any other responsibilities, such as storing data or processing search requests. The purpose of a dedicated master-eligible node is to improve the stability and reliability of the cluster by allowing it to elect a dedicated node to perform the tasks of a master node.

To configure a master-eligible node in Elasticsearch 8, you need to set the following options in the **`elasticsearch.yml`** configuration file:

```
node.roles: [ master ]
```