Broadview
www.broadview.com.cn

Elasticsearch

实战与原理解析

牛冬 编著

初学者快速上手，构建搜索引擎全景
洞悉Elasticsearch生态，建立知识网络

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

Elasticsearch

实战与原理解析

牛冬 编著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

本书基于Elasticsearch 7.X版本编写，内容由浅入深，先教会初学者使用，再介绍背后的原理。本书共分为三大部分，分别是Elasticsearch前传、Elasticsearch实战、Elasticsearch生态。Elasticsearch前传部分主要介绍搜索技术发展史和基本知识，并介绍搜索引擎技术原理，为读者构建搜索引擎全景。Elasticsearch实战部分主要介绍Elasticsearch的核心概念和架构设计，并重点介绍客户端、文档、搜索和索引等实战内容，待读者能上手实战后，再介绍这些内容的背后实现原理和关联知识，为读者构建知识网络。Elasticsearch生态部分主要介绍插件的使用和管理，以及Elastic Stack生态圈。

本书适合有一定基础知识的初、中级Elasticsearch学习者阅读。

图书在版编目（CIP）数据

Elasticsearch 实战与原理解析 / 牛冬编著. —北京：电子工业出版社，2020.3
ISBN 978-7-121-38380-9

Ⅰ. ①E… Ⅱ. ①牛… Ⅲ. ①搜索引擎—程序设计 Ⅳ. ①TP391.3

中国版本图书馆 CIP 数据核字(2020)第 021836 号

责任编辑：安 娜
印　　刷：三河市君旺印务有限公司
装　　订：三河市君旺印务有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编：100036
开　　本：787×980　1/16　印张：26.5　字数：554 千字
版　　次：2020 年 3 月第 1 版
印　　次：2020 年 3 月第 1 次印刷
定　　价：109.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819，faq@phei.com.cn。

序

在信息大爆炸的当下，信息过载已成为越来越多的人的负担。

随着 5G 时代的到来，物联网和智慧城市将会随处可见，随之而来的是信息会更加复杂和庞大。如何挣脱信息的束缚，高效地找到自己需要的信息呢？答案就是搜索引擎，即借助搜索引擎来寻找我们想要的信息！

本书介绍的搜索引擎是 Elasticsearch——一个开源的搜索引擎。

目前，Elasticsearch 的功能已不局限于搜索，它还在不断地丰富和完善自己的生态。在 API 接口层面，除基本的数据索引和数据搜索外，Elasticsearch 还提供了 Elasticsearch 服务监控接口、推荐相关接口，以及机器学习相关接口。

本书目的

与追求知识点全部覆盖但都泛泛而谈的书不同，本书聚焦初学者的学习和实战需要，将初学者接触 Elasticsearch 从 0 到 1 过程中的必备知识点讲透。只有学透了基础知识，再学习更多的有关 Elasticsearch 的知识才成为可能。

这一点笔者在培训 Elasticsearch 初学者时深有体会。因此，本书重点结合笔者在 Elasticsearch 上的沉淀、实战、培训和 Elasticsearch 最新版本内容，帮助 Elasticsearch 初学者点破这层窗户纸！

正如王阳明在《传习录》中谈为学之道时所言：“殊不知私欲日生，如地上尘，一日不扫便又有一层。着实用功，便见道无终穷，愈探愈深，必使精白无一毫不彻方可。”

对于知识与近代和现代高速发展的经济之间的关系，管理学大师德鲁克有一段精辟论述。他认为二者的关系可以分为三个发展阶段，即工业革命、生产力革命、管理革命。所谓工业革命，指的是知识应用于生产工具、生产流程和产品创新；所谓生产力革命，指的是知识以及被赋予的含义开始被应用于工作中；所谓管理革命，指的是知识正被用于知识本身。而管理革命的核心在于连接。在知识领域，连接意味着知识点关联。

很多人无法有效地将相似或关联的知识点进行关联，所以更谈不上构建网状知识体系。

因此，在本书行文过程中，笔者会基于自己构建的知识体系向读者进行必要的体系输出，

力求帮助读者在快速上手的同时，构建搜索引擎全景，洞悉 Elasticsearch 生态，建立关联知识网络。

本书基于 Elasticsearch 7.X 系列版本编写，内容由浅入深，先让初学者会用、能用，再介绍背后的原理。这种方式在笔者主导过的 Elasticsearch 技术培训中效果较好。

本书结构

本书分为三大部分，分别是 Elasticsearch 前传、Elasticsearch 实战和 Elasticsearch 生态。

Elasticsearch 前传部分主要介绍搜索技术发展史和基本知识，并介绍搜索引擎技术原理，为读者构建搜索引擎全景。在技术发展史上，我们能看见多久的历史，就能看见多远的未来！

Elasticsearch 实战部分主要介绍 Elasticsearch 的核心概念和架构设计，并重点介绍客户端、文档、搜索、索引等实战内容，待读者能上手实战后，再介绍这些内容的背后实现原理和关联知识，为读者构建知识网络。

Elasticsearch 生态部分主要介绍插件的使用和管理，以及 Elastic Stack 生态圈。

本书特色

特色 1：基于 Elasticsearch 7.X 系列版本编写。

特色 2：聚焦初学者学习和实战需要，不求知识点全部覆盖，但求必备知识透彻易懂。

特色 3：让初学者快速上手的同时，帮助他们构建搜索引擎全景、洞悉 Elasticsearch 生态、建立关联知识网络。

特色 4：由浅入深，先让初学者会用，再介绍背后的原理。

在本书编写过程中，Elasticsearch仍在升级版本，因此书中难免有理解和实践不足之处。“卑辞俚语，不揣谫陋”，欢迎读者和笔者交流学习，共同进步。

牛冬

2019 年 12 月

目　录

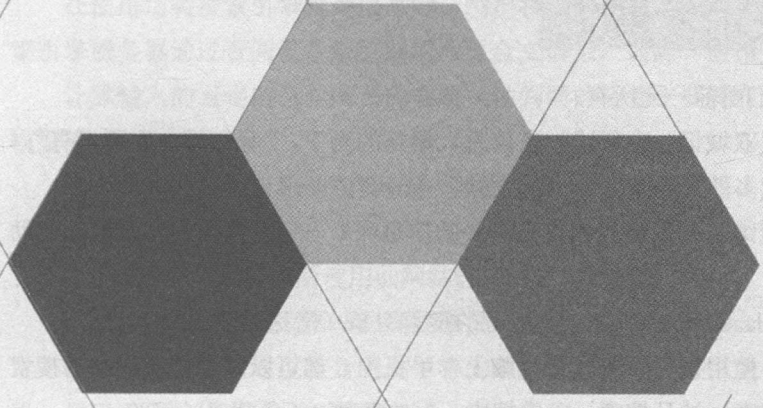
第一部分　Elasticsearch 前传

第二部分 Elasticsearch 实战

第三部分　Elasticsearch 生态

第一部分

Elasticsearch前传



第 1 章
搜索技术发展史

人事有代谢
往来成古今

1.1 正说搜索技术发展史

“我们面前无所不有，我们面前一无所有。”

正如查尔斯·狄更斯在《双城记》中所述。在信息大爆炸的当下，“我们面前无所不有”；而个人信息过载已成为越来越多的人的负担，“我们面前一无所有”。

如何挣脱过载的信息的束缚，高效地找到自己需要的信息呢？——答案是搜索引擎，借助搜索引擎来实现！

本书介绍的搜索引擎是 Elasticsearch——一个开源的搜索引擎（简称 ES）。

我们每天都在某种场景下使用搜索引擎，在电脑上、手机上，都可以找到自己惯用的搜索引擎，比如百度搜索、搜狗搜索、神马搜索、谷歌搜索、360 搜索、头条搜索，等等。

那么，搜索引擎是什么呢，它是如何发展到今天的样子呢？本章就介绍搜索技术发展，让我们沿着技术发展的脉络更深刻地认识搜索技术。

宏观而言，搜索引擎的发展经历了五个阶段和两大分类。五个阶段分别是 FTP 文件检索阶段、分类目录导航阶段、文本相关性检索阶段、网页链接分析阶段和用户意图识别阶段。具体情况汇总如下。

FTP 文件检索阶段

该阶段的搜索引擎只检索多个 FTP 服务器上存储的文件，代表作是 Archie。用户搜索文

件时需输入精确的文件名来搜索查找，搜索引擎会告诉用户从哪一个 FTP 地址可以下载被搜索的文件。

分类目录导航阶段

该阶段的搜索引擎就是一个导航网站，网站中都是网址的分类陈列，用户在互联网上常用的网址在这里一应俱全。

在使用该类搜索引擎时，用户需要从各个分类目录里找到自己想要的网址，单击其网站链接后进入相应的网站。

直到今天，这类搜索引擎依然不过时，我们常用的网站如好 123、搜狗浏览器主页、UC 导航等均是这类导航页面。

文本相关性检索阶段

随着互联网内容的不断丰富，网页的内容和形态也越来越多样化，页面中开始出现内容可能与网页地址和网页标题大相径庭的情况。

为了解决这个问题，搜索引擎引入全文搜索技术，来保证搜索引擎检索到的网页标题与网页全文内容强一致，摒弃了单纯依靠网页标题和网页地址来判断网页内容的方法。

在使用这类搜索引擎查询信息时，用户将输入的查询信息提交给搜索引擎后台服务器，搜索引擎服务器通过查阅已经索引好的网页全文信息，返回一些相关程度高的页面信息。

计算输入的查询信息与网页内容相关性判断的模型主要有布尔模型、概率模型、向量空间模型等。

这个阶段的搜索引擎的主要代表作是 Alta Vista、Excite 等。

网页链接分析阶段

这个阶段的搜索引擎所使用的网站链接形式与当前基本相同。在该阶段，外部链接表示推荐。

因此，通过计算每个网站的推荐链接的数量，就可以判断一个网站的流行性和重要性。

于是，搜索引擎通过结合网页内容的重要性和相似程度来改善搜索的信息质量。在这一阶段，搜索引擎的代表作是谷歌搜索。

这种模式是谷歌首创的，并且大获成功，随之引起了学术界和其他商业搜索引擎的极度关注和效仿。目前，网页链接分析算法及其改进优化的版本在主流搜索引擎中大行其道。

用户意图识别阶段

这个阶段的搜索引擎以用户为中心作为设计的初心，搜索引擎力求理解每一位用户的真正搜索诉求，力求做到千人千面，追求个性化识别和反馈。

在使用这类搜索引擎时，即便是同一个查询的请求关键词，不同的用户可能也会得到不同的查询结果。比如输入的是“小米”，那么一个想要购买小米电子设备的用户和一个想要购买

小米食用的用户，他们的搜索意图显然天壤之别，因而得到不同的搜索结果是顺理成章的事情。不光是不同用户之间，同一个用户搜索同样的关键词也会因时因地的不同而有所差异。比如当用户在搜索引擎上首次输入“TAL”时，可能是想查找 TAL 股票代码对应的好未来公司的网站；当用户在好未来的办公区内搜索“TAL”时，有可能是想查看 TAL 股票代码的实时股价。

其实在这两个案例背后，搜索引擎都在致力于解决同一个问题，即怎样才能通过输入的简短的关键词来判断用户的真正查询诉求。这也是我们将其归类为用户意图识别的原因。这一阶段的搜索引擎典型代表就是百度。

在搜索引擎技术不断演进的过程中，为了更好地识别及满足用户的搜索需求，更多的新技术也在不断引入，如 AI 技术、地理位置信息、用户画像等。

两大分类是指站内搜索和站外搜索。

站外搜索就是全网搜索，现在主流的搜索引擎基本都是全网搜索，如谷歌、百度。随着技术的发展，搜索领域的生态圈搜索形态不断扩大。以谷歌为代表的搜索引擎推出了整合搜索、个人化搜索、实时搜索、地图服务、线上文件编辑、网站统计、浏览器、网管工具、超大容量电子邮件、即时通信等。百度上线了百度百科、百度知道、百度贴吧等服务，这些服务中嵌入了文字搜索、语音搜索、图像搜索、地图搜索等搜索形态。

站内搜索近几年发展比较迅猛，各大网站平台纷纷上线了站内搜索，如 SNS 平台中的微博、人人网等，如电商平台中的京东、饿了么、淘宝、美团等。

另外，区块链内容搜索是近两年新的站内搜索形式，如比特币区块链的搜索内容在比特币公链上，但比特币公链的节点所在地域却是分布式的，和常见的站内搜索大相径庭，如图 1-1 所示。

图 1-1

在未来，搜索引擎的发展会是什么样的呢？我们不妨畅想一下。随着 5G 时代的到来，物

联网和智慧城市将会随处可见；AR/VR 技术会更加成熟，设备更加普及和便宜。与之对应的，除现在的文字搜索、语音搜索、图像搜索外，还会出现 AR/VR 搜索等搜索形态。

在 5G 的加持下，搜索引擎的搜索效率会更高；物联网和区块链中设备和信息搜索也会更加普遍，而搜索引擎的商业模式也可能随之升级，广告的效果可能会更好。

1.2　Elasticsearch 简介

Elasticsearch 是一个分布式、可扩展、近实时的高性能搜索与数据分析引擎。

Elasticsearch 提供了搜集、分析、存储数据三大功能，其主要特点有：分布式、零配置、易装易用、自动发现、索引自动分片、索引副本机制、RESTful 风格接口、多数据源和自动搜索负载等。

Elasticsearch 并非从零起步，而是站在巨人的肩膀上。Elasticsearch 基于 Java 编写，其内部使用 Lucene 做索引与搜索。通过进一步封装 Lucene，向开发人员屏蔽了 Lucene 的复杂性。开发人员无须深入了解检索的相关知识来理解它是如何工作的，只需使用一套简单一致的 RESTful API 即可，从此全文搜索变得简单。

除此之外，Elasticsearch 还解决了检索相关数据、返回统计结果、响应速度等相关的问题。因此，Elasticsearch 能做到分布式环境下的实时文档存储和实时分析搜索。实时存储的文档，每个字段都可以被索引与搜索。

最令人惊喜的是，Elasticsearch 能胜任上成百上千个服务节点的分布式扩展，支持 PB 级别的结构化或者非结构化海量数据的处理。

2019 年 4 月 10 日，Elasticsearch 发布了 7.0 版本。该版本的重要特性包含引入内存断路器、引入 Elasticsearch 的全新集群协调层——Zen2、支持更快的前 k 个查询、引入 Function score 2.0 等。

其中内存断路器可以更精准地检测出无法处理的请求，并防止它们使单个节点不稳定；Zen2 是 Elasticsearch 的全新集群协调层，提高了可靠性、性能和用户体验，使 Elasticsearch 变得更快、更安全，并更易于使用。

1.3　Lucene 简介

Lucene 是一个免费、开源、高性能、纯 Java 编写的全文检索引擎。

在业务开发场景中，Lucene 几乎适用任何需要全文检索的场景。因此，应普遍的搜索开发需求，各种编程语言的 Lucene 版本不断涌现。目前，Lucene 先后发展出了 C++、C#、Perl 和 Python 等语言的版本，Lucene 逐渐成为开源代码中最好的全文检索引擎工具包。

2005 年，Lucene 升级成为 Apache 顶级项目。

Lucene 包含大量相关项目，核心项目有 Lucene Core、Solr 和 PyLucene。

需要指出的是，Lucene 仅仅是一个工具包，它并非一个完整的全文检索引擎，这和 Lucene 的初衷相关。Lucene 主要为软件开发人员提供一个简单易用的工具包，主要提供倒排索引的查询结构，以方便软件开发人员在其业务系统中实现全文检索的功能。这也是我们常说全文检索引擎主要是 Solr 和 Elasticsearch 的原因，虽然二者均是以 Lucene 为基础建立的。

Lucene 作为一个全文检索引擎工具包，具有如下突出优点。

索引文件格式独立于应用平台

Lucene 定义了一套以 8 位字节为基础的索引文件格式，使得兼容系统或者不同平台的应用能够共享建立的索引文件。

索引速度快

在传统全文检索引擎的倒排索引的基础上，实现了分块索引，能够针对新的文件建立小文件索引，提升索引速度。然后通过与原有索引的合并，达到优化的目的。

简单易学

优秀的面向对象的系统架构，降低了 Lucene 扩展的学习难度，方便扩充新功能。

跨语言

设计了独立于语言和文件格式的文本分析接口，索引器通过接收 Token 流完成索引文件的创立，用户扩展新的语言和文件格式，只需实现文本分析的接口即可。

强大的查询引擎

Lucene 默认实现了一套强大的查询引擎，用户无须自己编写代码即可通过系统获得强大的查询能力。Lucene 默认实现了布尔操作、模糊查询、分组查询等。

Lucene 的主要模块有 Analysis 模块、Index 模块、Store 模块、QueryParser 模块、Search 模块和 Similarity 模块，各模块的功能分别汇总如下。

① Analysis 模块：主要负责词法分析及语言处理，也就是我们常说的分词，通过该模块可最终形成存储或者搜索的最小单元 Term。

② Index 模块：主要负责索引的创建工作。

③ Store 模块：主要负责索引的读和写，主要是对文件的一些操作，其主要目的是抽象出和平台文件系统无关的存储。