

云计算与虚拟化技术丛书

Mastering Elasticsearch
Second Edition

深入理解 Elasticsearch

(原书第2版)

[美]

拉斐尔·酷奇 (Rafał Kuć) 著
马雷克·罗戈任斯基 (Marek Rogoziński)

张世武 余洪淼 商旦 译



机械工业出版社
China Machine Press

图书在版编目 (CIP) 数据

深入理解 Elasticsearch (原书第 2 版) / (美) 拉斐尔·酷奇 (Rafał Kuć) 等著; 张世武等译. —北京: 机械工业出版社, 2017.5

(云计算与虚拟化技术丛书)

书名原文: Mastering Elasticsearch, Second Edition

ISBN 978-7-111-56825-4

I. 深… II. ①拉… ②张… III. 互联网络—情报检索 IV. G254.928

中国版本图书馆 CIP 数据核字 (2017) 第 087852 号

本书版权登记号: 图字: 01-2016-6257

Rafał Kuć, Marek Rogoziński: Mastering Elasticsearch, Second Edition (ISBN: 978-1-78355-379-2).

Copyright © 2015 Packt Publishing. First published in the English language under the title “Mastering Elasticsearch, Second Edition”.

All rights reserved.

Chinese simplified language edition published by China Machine Press.

Copyright © 2017 by China Machine Press.

本书中文简体字版由 Packt Publishing 授权机械工业出版社独家出版。未经出版者书面许可, 不得以任何方式复制或抄袭本书内容。

深入理解 Elasticsearch (原书第 2 版)

出版发行: 机械工业出版社 (北京市西城区百万庄大街 22 号 邮政编码: 100037)

责任编辑: 张梦玲

责任校对: 殷虹

印刷: 三河市宏图印务有限公司

版次: 2017 年 5 月第 1 版第 1 次印刷

开本: 186mm × 240mm 1/16

印张: 20

书号: ISBN 978-7-111-56825-4

定价: 79.00 元

凡购本书, 如有缺页、倒页、脱页, 由本社发行部调换

客服热线: (010) 88379426 88361066

投稿热线: (010) 88379604

购书热线: (010) 68326294 88379649 68995259

读者信箱: hzit@hzbook.com

版权所有·侵权必究

封底无防伪标均为盗版

本书法律顾问: 北京大成律师事务所 韩光 / 邹晓东

华章IT
HZBOOKS | Information Technology



den39/07

随着互联网时代的来临，人类面临着前所未有的信息过载问题。为了方便人们从海量信息中快速精准地检索感兴趣的内容，Web 搜索引擎应运而生。在互联网发展早期，数据量比较小，单机索引就能支撑起一个完整应用。在这个时期，Apache Lucene 凭借其精巧的代码设计、优异的性能、丰富的查询接口，以及众多衍生搜索产品（如 Apache Solr、Nutch 等），在开源搜索领域大放异彩。随着互联网的发展，数据量快速膨胀，此时对搜索引擎提出了分布式、准实时、高容错、可扩展、易于交互等诸多进阶要求。基于 Lucene 的简单二次开发已经不能满足日常搜索需求，Elasticsearch 的诞生很好地满足了上述大数据时代的搜索产品需求。

Elasticsearch 是一款基于 Apache Lucene 的开源搜索引擎产品，最早发布于 2010 年。之后 Elasticsearch 的开发团队成立了专门的商业公司，继续对其进行开发并提供服务和技术支持。Elasticsearch 具有开源、分布式、准实时、RESTful、便于二次开发等特点，代码实现精巧，系统稳定可靠，已经被国内外众多知名组织和公司广泛采用。

本书内容丰富，不仅深入介绍了 Apache Lucene 的评分机制、查询 DSL、底层索引控制，还详细介绍了 Elasticsearch 的分布式索引机制、用户体验的改善、系统管理及性能优化，以及自定义插件的开发。本书文笔优雅，辅以大量翔实的实例，能帮助读者快速提高 Elasticsearch 应用水平。需要提醒读者的是，本书的目标读者是 Elasticsearch 的中高级用户，如果读者对 Elasticsearch 的基础概念（如 Mapping、Type）缺乏了解的话，建议先阅读作者另一本针对初学者的书《Elasticsearch Server》（目前已经出了第 2 版）。

相对于第 1 版，本书重新组织了多个章节的内容，删减了部分对读者来说意义不大的内容，如 Java API 的使用，突出了 Elasticsearch 使用过程中的进阶技能，比如查询优化、用户体验改善、Elasticsearch 节点集群管理等。

本书译文经过精心组织，结合了译者对 Elasticsearch 的使用经验，并参考了 IBM、微软、百度、腾讯等众多知名企业的业界专业人士的意见。本书翻译团队由拥有丰富 Elasticsearch 实践经验的搜索算法专家、架构师和开发者组成。其中，张世武负责第 3、5、6 章的翻译和校对工作，余洪森负责第 7、8、9 章的翻译和校对，商旦负责第 1、2、4 章的翻译及全书校对。在

本书翻译过程中，翻译团队经过多次讨论、审校，力求翻译准确、优雅。由于本书涉及很多新概念，业界尚无统一术语，另外译者水平有限，难免会出现一些问题，欢迎广大读者及业内同行批评指正。

翻译团队的所有成员在精力和时间都很有限的情况下，能够四个月如一日，顶着日常工作和生活的压力，不计回报地通力合作，为一本书的翻译付出极大的热忱，实在难得。这里要感谢翻译团队家人们的理解和默默支持，让我们一路走到今天。

在本书的翻译过程中，译者张世武得到所在单位成都数联铭品科技有限公司（BBD）的大力支持。BBD 是一家行业领先的金融大数据公司，Elasticsearch 在 BBD 得到了广泛的应用。感谢曾途、尹康、丁国栋、宋开发、刘荀、田志伟、何宏靖等领导对本书出版给予的关怀与支持。技术部及兄弟部门同事提出了很多宝贵建议，感谢以下同事的支持：李想、王尧、丁明会、范东来、刘世林、赵龙、王振宇、何耀、刘兆强、金涛、黄光虎、范从俊、许尧、张伟、邢杰、丁永强、程柳航、何俊、唐国海、刘兴洋、苏印、刘俊杰、闫俊杰、张学锋、徐胜、吕司君、吴德文、邹俊、何正开、刘龙均、郭羽凌、吴勇、杨熹岑、曾昌强、秦德强、谭亚军、魏犇、席爱龄、沈思丞、田曲、陈珊、胡馨月、申秋艳、马小思、高翔、葛相宇、唐斌、苏柏钰、魏林玮、沈思成、韩远、夏阳、杨建蓉、孙健、蒲雪芹、何晓宁、薛双凯、张冬冬。

同样需要感谢机械工业出版社的编辑团队，感谢他们提供的这次翻译机会。尤其需要感谢张梦玲女士，没有她的督导和支持，本书很难如期翻译校对完毕。还要感谢出版社相关岗位的工作人员，离开他们的支持，本书也无法很快与读者见面。

再次感谢！

译者

2017 年 3 月

About the Author 作者简介

Rafał Kuć 是一个很有天赋的团队领袖以及软件开发人员。他现在担任 Sematext 集团公司的咨询专家及软件工程师，专注于开源技术，如 Apache Lucene、Solr、Elasticsearch、Hadoop stack 等。他拥有超过 13 年的软件研发经验，涉及领域广阔，从银行软件到电子商务产品，主要侧重于 Java 平台。除此之外，他对能提高研发效率的任意工具或编程语言都抱有极高的热情。同时他也是 solr.pl 网站的创始人之一，该网站致力于帮助人们解决 Solr、Lucene 相关问题。他还是世界范围内各种会议热邀的演讲嘉宾，之前受邀出席过 Lucene Eurocon、Berlin Buzzwords、ApacheCon、Lucene Revolution、DevOps Days 等会议。

早在 2002 年 Rafał 就接触到 Lucene，一开始他并不喜欢这个开源产品，然而在 2003 年，当他再次使用 Lucene 时，他改变了自己的看法，并看到了搜索技术的巨大潜力。随后 Solr 诞生了。Rafał 于 2010 年开始使用 Elasticsearch。Rafał 目前主要着眼于 Lucene、Solr、Elasticsearch 和信息检索等方面的研究。

Rafał 是《Solr 3.1 Cookbook》一书及其后续版本《Solr 4.0 Cookbook》以及再版作品《Solr Cookbook, Third Edition》的作者。他也是《Elasticsearch Server》第 1 版和第 2 版、《Mastering Elasticsearch》第 1 版的作者，这些书均由 Packt Publishing 出版。

随着 Elasticsearch 的快速发展，我和 Marek 考虑针对本书第 1 版做一次更新。这本书并不适合所有读者，不过在第 1 版中我们没有做出足够的强调。我们把本书第 1 版作为《Elasticsearch Server》的升级版。本书第 1 版的定位也是如此：此刻你手捧的这本书是对《Elasticsearch Server, Second Edition》的扩展和续作。正因为如此，本书可以聚焦在一些高级主题上，比如如何选择恰当的查询方式、如何扩展 Elasticsearch、大量的评分细节和示例、过滤器内部机制、新增的聚集器、不同文档关系处理方式的优劣对比，等等。期望在读完本书后，读者能够更好地理解 Elasticsearch 和 Apache Lucene 底层架构的各种细节，这将有助于他们更方便快捷地获取所需知识。

在此，我想感谢我的家庭，当我在电脑屏幕前全身心投入本书写作的那些日日夜夜里，他们表现出极大的耐心，是我最坚强的后盾。

同样也要感谢 Sematext 所有的同事们，尤其是 Otis，感谢他为我付出的时间，并让我深刻认识到 Sematext 是一个非常适合我的公司。

最后，非常诚挚地感谢 Elasticsearch、Lucene 项目的所有创建者和开发者，感谢他们杰出的工作和对开源项目的热情。没有他们，就没有本书的诞生；没有他们，开源搜索引擎就不会有现在的这种活力。

再次感谢你们！

——Rafał Kuć

Marek Rogoziński 是一个有着超过 10 年经验的软件架构师和咨询师。其技术专长涉及基于开源搜索引擎（如 Solr、Elasticsearch 等）的解决方案及大数据分析技术（Hadoop、HBase、Twitter Storm 等）。

他也是 solr.pl 网站的联合创始人之一，该网站致力于提供 Solr 和 Lucene 相关资讯和教程。他同时也是 Packt Publishing 出版的《Elasticsearch Server》第 1 版和第 2 版、《Mastering Elasticsearch》第 1 版^①的作者之一。

目前，他是 ZenCard 公司的首席技术官和主架构师。ZenCard 公司提供实时处理和分析大规模交易事务的相关工具，可自动化、匿名识别所有零售渠道（移动电子商务、电子商务、实体店）的零售客户，给零售商提供客户留存和忠诚度运营的工具。

这已经是我们完成的第 4 本关于 Elasticsearch 的书了。在编写本书时，我一再沉醉于 Elasticsearch 的快速进化。我们一直在被标记为“试验中”和“正在开发中”的各种特性之间求取平衡，也不得不冒着最终代码变更，甚至某些有趣的功能被废弃的风险写作相关内容。本书第 2 版重写了很多内容，并引入不少 Elasticsearch 的新特性。不过，这些更新也是有代价的：我们剔除了一些对读者来说不是那么重要的信息。在本书中，我们还试图为读者介绍一些 Elasticsearch 相关的扩展主题。不过请注意，Elasticsearch 的完整生态系统和 ELK 开发栈（Elasticsearch、Logstash 和 Kibana）或者 Hadoop 集成，都需要精心撰写单独的书籍加以介绍。

现在，致谢时间到了。

感谢 Elasticsearch、Lucene 及所有相关产品的创造者们。

同时也要感谢本书的出版团队。尤其是感谢帮助检查错误、校稿、消除表达歧义的伙伴们。非常感谢给我们留言评论和提出建设性意见的读者们。发现有人认可我们的工作令我感到非常惊喜，并深受鼓舞。谢谢你们！

最后，感谢所有支持我和理解我写作的朋友们。

——Marek Rogoziński

① 《Mastering Elasticsearch》的中文版《深入理解 Elasticsearch》已由机械工业出版社出版，书号为 978-7-111-52416-8，本书为《Mastering Elasticsearch, Second Edition》的中文版。——编辑注

About the Reviewer 评审者简介

Hüseyin Akdoğan 的软件职业生涯起源于 GwBasic 编程语言。他先后掌握了 QuickBasic 和 Visual Basic 语言，并用 VB 开发了许多应用程序。2000 年，他迈入了 Web 和 PHP 的世界，后来又穿插使用 Java 语言进行开发。从 2005 年起，他一边进行咨询和培训活动，一边继续使用 Java EE 技术开发企业应用程序。他擅长 JavaServer Faces、Spring Frameworks，以及诸如 NoSQL 和 Elasticsearch 等大数据技术。除此之外，他还在钻研其他大数据相关技术。

Julien Duponchelle 是一位毕业于欧洲理工学院的法国工程师。他致力于开发提升 IT 团队工作效率的各种工具，并在其职业生涯中贡献了多个开源软件项目。

自从他加入 ETNA（一个法国 IT 职业教育学校）开始涉足教育领域起，Julien 以首席支持工程师的身份协助过好几个创业公司进行工作，并参与了多个重大项目的募资活动，包括 Plizy 和 Youboox。

感谢我的女友 Maëlig，感谢她在我评审本书和开发其他开源项目的无数个夜晚的善意支持和极大的耐心。

——Julien Duponchelle

Marcelo Ochoa 现任教于阿根廷布宜诺斯艾利斯省中部国立大学精确科学与自然科学学院的系统实验室，也是 Scotas.com 公司的 CTO，该公司致力于提供基于 Solr 和 Oracle 的准实时搜索解决方案。在高校任职的同时，他也参与一些与 Oracle、大数据相关的外部项目。他参与过的 Oracle 相关项目有 Oracle 手册文档翻译、多媒体培训等。其技术背景涉及数据库、网络、Web、Java 等。在 XML 领域，他参与过 Apache Cocoon 中的 DB Generator，开源项目 DBPrism、DBPrism CMS，基于 Oracle JVM Directory 的 Lucene-Oracle 集成方案，Restlet.org 项目中的 Oracle XDB Restlet Adapter（一个能在基于数据库驻存的 JVM 内部生成本地 REST Web 服务的解决方案）等项目或模块的开发，并因此为业界所熟知。

从 2006 年开始，他参与了 Oracle ACE 计划，这是 Oracle 公司官方推出的一个计划，旨在认可和奖励 Oracle 技术社区中技术娴熟并愿意分享他们的知识、经验的成员为该社区所做

的贡献。

Marcelo Ochoa 还是《Oracle Database Programming Using Java and Web Services》（Kuassi Mensah，由 Digital Press 出版）和《Professional XML Databases》（Kevin Williams，由 Wrox Press 出版）两书的合著者，同时也是 Packt 出版社的多部书籍，如《Apache Solr 4 Cookbook》《Elasticsearch Server》等的评审者。

欢迎来到 Elasticsearch 的世界并阅读本书第 2 版。通过阅读本书，我们将带领你接触与 Elasticsearch 紧密相关的各种话题。请注意，本书不是为初学者写的。笔者将本书作为《Elasticsearch Server, Second Edition》的续作和姊妹篇。相对于《Elasticsearch Server》，本书涵盖了很多新知识，不过你偶尔也可以在本书中发现一些引自《Elasticsearch Server》的内容。

本书将探讨与 Elasticsearch 和 Lucene 相关的多个不同主题。首先，我们以介绍 Lucene 和 Elasticsearch 的基本概念作为开始，带领读者认识 Elasticsearch 提供的众多查询方式。在这里，将涉及和查询相关的不同主题，比如结果过滤以及如何为特定场景选择合适的查询方式。显然，Elasticsearch 不仅仅只有查询功能。因此，本书还将介绍 Elasticsearch 新加入的聚集功能，以及众多能够赋予被索引数据意义的特性，并设法提供更佳的用户查询体验。

对大多数用户来说，查询和数据分析是 Elasticsearch 最吸引人的部分，不过这些还不是我们想要探索的全部内容。因此，本书在涉及索引架构时还会试图跟读者探讨一些额外话题，比如如何选择合适的分片数和副本数，如何调整分片分配行为等。当谈论 Elasticsearch 和 Lucene 之间的关系时，我们还将介绍不同的打分算法、算法之间的差异、如何选择合适的存储机制，以及为什么需要做此选择。

最后，我们还将触及 Elasticsearch 的管理功能，将探讨发现和恢复模块，以及对人类友好的 Cat API。Cat API 可以帮助我们快速获取相关的运维信息，它的返回数据组织成一种大多数人都易于阅读的格式，无需进行 JSON 解析。我们还将认识和使用部落节点，它能够为我们提供在多个节点间联合查询的能力。

因为本书的书名，我们无法忽略与性能相关的话题，所以我们决定用整整一章来探讨性能。我们谈论了文档取值及其相关改进，还介绍了垃圾回收器的工作方式，以及在垃圾回收器未能如我们期望般工作时可以做些什么。最后，探讨了如何扩展 Elasticsearch 以应对高索引量和查询量的场景。

和本书第 1 版一样，我们决定以开发 Elasticsearch 插件的话题作为本书结尾。我们将展示如何构建 Apache Maven 项目，并开发两个不同类型的插件——自定义 REST 操作插件和自定义

分析插件。

假如你在读完某些主题后对其产生浓厚的兴趣，那么这本书就是适合你的。希望你在读完后能够喜欢这本书。

本书主要内容

第 1 章先介绍 Apache Lucene 的工作方式，再介绍 Elasticsearch 的基本概念，并演示 Elasticsearch 内部是如何工作的。

第 2 章描述 Lucene 评分过程，为什么要进行查询改写，什么是查询模板以及如何使用查询模板。除此之外，还介绍了过滤器的使用，以及如何为特定场景选择合适的查询方式。

第 3 章描述了查询二次评分、多匹配控制，并介绍了用于做查询分析的各种聚合类型。关键词项聚合和最优词项聚合可以根据所含内容片段对文档进行归类。除此之外，还介绍了 Elasticsearch 的 parent-child 文档关系处理，并提供了在 Elasticsearch 中使用脚本的相关知识。

第 4 章覆盖了有关用户体验提升的相关话题。本章介绍了查询建议 (suggester)，它能帮助修正查询中的拼写错误并构建高效的自动完成 (autocomplete) 机制。除此之外，通过实际的案例展示如何通过使用不同查询类型和 Elasticsearch 的其他功能来提高查询相关性。

第 5 章介绍了以下技术：如何选择合适的分片及副本数，路由是如何工作的，索引分片机制是如何工作的以及如何影响分片行为。同时介绍了什么是查询执行偏好，以及它是如何影响查询执行的。

第 6 章描述如何修改 Lucene 评分以及如何选择备用的评分算法。本章也介绍了 Elasticsearch 的准实时搜索和索引，事务日志的使用，理解索引的段合并，以及如何调整段合并来适应应用场景。在本章最后，还将介绍 Elasticsearch 的缓存机制和请求打断器，以避免出现内存用尽的故障。

第 7 章介绍了什么是发现、网关、恢复模块，如何配置这些模块，以及有哪些令人心烦的疑难点。还介绍了什么是 Cat API，如何把数据备份到各种云服务上（比如亚马逊的 AWS 和微软的 Azure），以及如何从云服务上恢复数据。最后还介绍了如何使用部落节点进行联盟搜索。

第 8 章覆盖了与 Elasticsearch 性能相关的各种主题，从使用文档取值来优化字段数据缓存的内存使用，到 JVM 垃圾回收器的工作原理，再到查询基准测试，最后到如何扩展 Elasticsearch 以适应更高的索引量和查询量场景。

第 9 章通过演示如何开发你自己的 REST 操作插件和查询语言分析插件来介绍 Elasticsearch 的插件开发。

阅读本书的必备资源

本书写作时采用了 Elasticsearch 的 1.4.x 版本，所有的范例代码应该能在该版本下正常运行。除此之外，读者需要一个能发送 HTTP 请求的命令行工具，例如 `curl`，该工具在绝大多数操作系统上是可用的。请记住，本书的所有范例都使用了 `curl`。如果读者想使用其他工具，请注意检查请求的格式，以保证你所选择的工具能正确解析它。

除此之外，为了运行第 9 章的范例，需要读者的机器上已安装了 JDK，并且需要一个编辑器来开发相关代码（或者类似 Eclipse 的 Java IDE）。另外，还要求使用 Apache Maven 进行代码的管理与构建。

本书的目标读者

本书的目标读者是那些对 Elasticsearch 基本概念已经很熟悉但是又想深入了解其本身，同时也对 Apache Lucene、JVM 垃圾收集感兴趣的 Elasticsearch 用户和发烧友。除此之外，想了解如何改进查询相关性、如何使用 Elasticsearch Java API、如何编写自定义插件的读者，也会发现本书的趣味性和实用性。

如果你是 Elasticsearch 的初学者，连查询和索引这些基本概念都不熟悉，那么你会发现本书的绝大多数章节难以理解，因为这些内容假定读者已经有相关背景知识。如果是这种情况，建议参考 Packt 出版社出版的另一本关于 Elasticsearch 的图书——《Elasticsearch Server, Second Edition》。

读者反馈

欢迎读者的任何反馈。请让我们知道你对本书的看法——喜欢或者不喜欢哪些内容。读者的反馈对我们开发一些大众能够受益的课题是非常重要的。

一般性的反馈可直接发送 e-mail 至 feedback@packtpub.com，请在邮件标题中提及相关书名。

如果你觉得自己在某个话题上有专长，并有兴趣撰写一本书或做一些贡献，请浏览 www.packtpub.com/authors 上的作者指南。

客户支持

现在，你已经成为 Packt 图书尊贵的读者了。我们提供了各种服务来帮助你从本次购书行为中获得最大价值。

范例代码下载

如果读者通过 <http://www.packtpub.com> 账号购买了 Packt 图书，可直接在本网站下载范例代码。如果读者在其他地方购买了 Packt 图书，可登录 <http://www.packtpub.com/support>，并注册账号，我们将通过 e-mail 发送代码给你。

勘误

尽管我们已经尽了最大努力来保障图书内容的准确性，错误还是难以避免的。如果读者发现了任何错误，不论是文字部分还是代码部分，请报告我们，我们将不胜感激。你的这种行为能帮助其他读者免受误导，并能帮助我们提高图书后续版本的质量。发现任何错误，请登录 <http://www.packtpub.com/submit-errata>，选择相应的书名，单击 errata submissionform 并在表单中输入错误的详细信息。一旦所提交的错误被确认，它会被上传至我们的网站，或者被添加到现有的勘误表里。可登录 <http://www.packtpub.com/support> 选择具体的书名，查看现有的勘误表。

译者序

作者简介

评审者简介

前言

第 1 章 Elasticsearch 简介..... 1

1.1 Apache Lucene 简介..... 1

1.1.1 熟悉 Lucene 2

1.1.2 Lucene 的总体架构 2

1.1.3 分析数据 4

1.1.4 Lucene 查询语言 5

1.2 何为 Elasticsearch 8

1.2.1 Elasticsearch 的基本概念 8

1.2.2 Elasticsearch 架构背后的关键 概念 10

1.2.3 Elasticsearch 的工作流程 10

1.3 在线书店示例 14

1.4 小结 17

第 2 章 查询 DSL 进阶..... 18

2.1 Apache Lucene 默认评分公式解释 18

2.1.1 何时文档被匹配上 19

2.1.2 TF/IDF 评分公式 19

2.1.3 Elasticsearch 如何看评分 21

2.1.4 一个例子 21

2.2 查询改写 24

2.2.1 前缀查询示例 24

2.2.2 回到 Apache Lucene 26

2.2.3 查询改写的属性 28

2.3 查询模板 30

2.3.1 引入查询模板 31

2.3.2 Mustache 模板引擎 33

2.3.3 把查询模板保存到文件 35

2.4 过滤器的使用及作用原理 36

2.4.1 过滤及查询相关性 36

2.4.2 过滤器的工作原理 40

2.4.3 性能考量 41

2.4.4 后置过滤和过滤查询 42

2.4.5 选择正确的过滤方式 44

2.5 选择正确的查询方式 45

2.5.1 查询方式分类 45

2.5.2 使用示例 50

2.6 小结 65

第 3 章 不只是文本搜索 66

3.1 查询二次评分 66

3.1.1 什么是查询二次评分.....	67	4.1.3 suggester.....	121
3.1.2 一个查询例子.....	67	4.2 改善查询相关性.....	142
3.1.3 二次评分查询的结构.....	67	4.2.1 数据.....	142
3.1.4 二次评分参数.....	70	4.2.2 改善相关性的探索之旅.....	145
3.1.5 总结.....	70	4.3 小结.....	157
3.2 多匹配控制.....	71	第5章 分布式索引架构.....	159
3.3 重要词项聚合.....	78	5.1 选择合适的分片和副本数.....	159
3.3.1 一个例子.....	79	5.1.1 分片和过度分配.....	160
3.3.2 选择重要词项.....	81	5.1.2 一个过度分配的正面例子.....	161
3.3.3 多值分析.....	81	5.1.3 多分片与多索引.....	161
3.3.4 额外的配置.....	84	5.1.4 副本.....	161
3.3.5 使用限制.....	89	5.2 路由.....	162
3.4 文档分组.....	89	5.2.1 分片和数据.....	162
3.4.1 top_hits 聚合.....	90	5.2.2 测试路由功能.....	162
3.4.2 一个例子.....	90	5.2.3 索引时使用路由.....	166
3.5 文档关系.....	95	5.2.4 别名.....	169
3.5.1 对象类型.....	95	5.2.5 多个路由值.....	169
3.5.2 嵌套文档.....	98	5.3 调整默认分片的分配行为.....	170
3.5.3 parent-child 关系.....	99	5.3.1 部署意识.....	171
3.5.4 其他解决方案.....	102	5.3.2 过滤.....	173
3.6 Elasticsearch 各版本中脚本的变化.....	102	5.3.3 运行时更新分配策略.....	174
3.6.1 脚本变迁.....	102	5.3.4 确定每个节点允许的总分片数.....	175
3.6.2 Groovy 简单介绍.....	103	5.3.5 确定每个物理机器允许的 总分片数.....	175
3.6.3 全文检索中的脚本.....	108	5.4 查询执行偏好.....	179
3.6.4 Lucene 表达式.....	115	5.5 小结.....	181
3.7 小结.....	118	第6章 底层索引控制.....	182
第4章 改善用户搜索体验.....	119	6.1 改变 Apache Lucene 的评分方式.....	182
4.1 改正用户拼写错误.....	119	6.1.1 可用的相似度模型.....	183
4.1.1 测试数据.....	120		
4.1.2 深入技术细节.....	121		

6.1.2	为每字段配置相似度模型	183
6.1.3	相似度模型配置	184
6.1.4	选择默认的相似度模型	185
6.2	选择适当的目录实现——store 模块	188
6.3	准实时、提交、更新及事务日志	191
6.3.1	索引更新及更新提交	192
6.3.2	事务日志	193
6.3.3	准实时读取	194
6.4	控制索引合并	195
6.4.1	选择正确的合并策略	196
6.4.2	合并策略配置	197
6.4.3	调度	199
6.5	关于 I/O 调节	200
6.5.1	控制 I/O 节流	200
6.5.2	配置	200
6.6	理解 Elasticsearch 缓存	202
6.6.1	过滤器缓存	203
6.6.2	字段数据缓存	204
6.6.3	查询分片缓存	212
6.6.4	使用 circuit breaker	213
6.6.5	清除缓存	214
6.7	小结	215

第 7 章 管理 Elasticsearch

7.1	发现和恢复模块	216
7.1.1	发现模块的配置	217
7.1.2	主节点	218
7.1.3	网关和恢复模块的配置	223
7.1.4	索引恢复 API	226
7.2	使用人类友好的 Cat API	229
7.2.1	基础知识	230

7.2.2	使用 Cat API	231
7.2.3	一些例子	232
7.3	备份	232
7.4	联盟搜索	236
7.4.1	测试用的集群	236
7.4.2	建立部落节点	237
7.4.3	通过部落节点读取数据	238
7.4.4	通过部落节点写入数据	239
7.4.5	处理索引冲突	240
7.4.6	屏蔽写操作	241
7.5	小结	242

第 8 章 提高性能

8.1	使用 doc values 来优化查询	243
8.1.1	字段缓存存在的问题	244
8.1.2	使用 doc values 的例子	245
8.2	了解垃圾回收器	247
8.2.1	Java 内存	248
8.2.2	解决垃圾回收问题	249
8.2.3	在类 UNIX 系统上避免内存 交换	254
8.3	对查询做基准测试	255
8.3.1	为基准测试配置集群	256
8.3.2	进行基准测试	256
8.3.3	控制运行中的基准测试	259
8.4	热点线程	261
8.4.1	热点线程的使用说明	261
8.4.2	热点线程 API 的响应	262
8.5	扩展 Elasticsearch	263
8.5.1	垂直扩展	263
8.5.2	水平扩展	264

8.5.3 在高负载的场景下使用 Elasticsearch	271
8.6 小结	283
第 9 章 开发 Elasticsearch 插件	284
9.1 创建 Maven 项目	284
9.2 了解基本知识	285
9.2.1 Maven Java 项目的结构	285
9.2.2 POM 的理念	285
9.2.3 执行构建过程	286
9.2.4 引入 Maven 装配插件	287
9.3 创建自定义 REST 行为	289
9.3.1 设定	289
9.3.2 实现细节	289
9.4 创建自定义分析插件	295
9.4.1 实现细节	295
9.4.2 测试自定义分析插件	302
9.5 小结	304