HZ BOOKS 华章IT

云计算与虚拟化技术丛书

[PACKT] PUBLISHING

Mastering ElasticSearch

深入理解 ElasticSearch

[美] 拉斐尔·酷奇（Rafał Kuć）
马雷克·罗戈任斯基（Marek Rogoziński） 著
张世武 余洪淼 商旦 译

资深软件开发专家、架构师撰写，系统且深入阐释ElasticSearch涉及的工具、方法、原则和最佳实践

深入剖析ElasticSearch应用过程中遇到的各个层面的问题，涉及分布式索引机制、系统监控及性能优化、用户体验改善、Java API应用，以及自定义插件开发

机械工业出版社
China Machine Press

云计算与虚拟化技术丛书

Mastering ElasticSearch

深入理解 ElasticSearch

[美] 拉斐尔·酷奇（Rafał Kuć）
马雷克·罗戈任斯基（Marek Rogoziński）著
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/（美）酷奇（Kuć, R.），（美）罗戈任斯基（Rogoziński, M.）著；张世武，余洪淼，商旦译 . —北京：机械工业出版社，2016.1
（云计算与虚拟化技术丛书）
书名原文：Mastering ElasticSearch

ISBN 978-7-111-52416-8

I. 深… II. ①酷… ②罗… ③张… ④余… ⑤商… III. 互联网络－情报检索 IV. ①G354.4 ②TP391.3

中国版本图书馆 CIP 数据核字（2015）第 309541 号

本书版权登记号：图字：01-2014-2032

深入理解 ElasticSearch

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：秦　健　　责任校对：董纪丽
印　　刷：北京文昌阁彩色印刷有限责任公司　　版　　次：2016 年 1 月第 1 版第 1 次印刷
开　　本：186mm × 240mm　1/16　　印　　张：16.75
书　　号：ISBN 978-7-111-52416-8　　定　　价：69.00 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

The Translator's Words 译 者 序

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数据中快速精准地检索感兴趣的信息，Web 搜索引擎应运而生。在互联网发展的早期，数据量比较小，单机索引就能支撑一个完整的应用。此时 Apache Lucene 凭借其精巧的代码设计、优异的性能、丰富的查询接口，以及众多的衍生搜索产品（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、易于交互等诸多要求。基于 Lucene 的简单二次开发已经满足不了日常的搜索需求，ElasticSearch 的诞生则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 的开发团队成了专门的商业公司，持续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，而且介绍了 ElasticSearch 的分布式索引机制、系统监控及性能优化、用户体验的改善、Java API 的使用，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 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 的基础概念诸如 Mapping、Types 等缺乏了解的话，可先阅读作者的另外一本针对初学者的书籍《ElasticSearch Server》。

本书的译文经过精心组织，结合了译者的 ElasticSearch 使用经验，并参考了 IBM、微软、百度、腾讯等多位业界专业人士的意见。其中，张世武负责第 1 ～ 3 章的翻译及全书的审校，余洪淼负责第 4 ～ 5 章的翻译，商旦负责第 6 ～ 9 章的翻译。在本书交稿之前，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，难免会出现一些翻译问题。欢迎广大读者朋友及业内同行批评指正。

前　言 *Preface*

欢迎来到 ElasticSearch 的世界。通过阅读本书，我们将带你接触与 ElasticSearch 紧密相关的各种话题。本书会从介绍 Apache Lucene 及 ElasticSearch 的基本概念开始。即使读者熟悉这些知识，简略的介绍也是很有必要的，掌握背景知识对于全面理解集群构建、索引文档、搜索这些操作背后到底发生了什么至关重要。

之后，读者将学习 Lucene 的评分过程是如何工作的，如何影响评分，以及如何让 ElasticSearch 选择不同的评分算法。本书也将介绍什么是查询重写以及进行查询重写的原因。除此之外，本书还将介绍如何修改查询来影响 ElasticSearch 的缓存功能以及如何最大限度地使用缓存。

接着你将学习索引控制的相关知识：如何通过设置不同的倒排表格式（posting format）来改变索引字段的写入模式；索引的段合并机制和段合并的重要性，以及如何调整段合并来适应应用场景；深入探讨索引分片（shard）的分配机制、路由机制，以及当数据量、查询量日渐增长时的应对策略。

当然本书也不会遗漏垃圾收集的相关内容，包括垃圾收集的工作原理、触发时间以及如何调整垃圾收集的行为。此外，本书也将涉及 ElasticSearch 状态诊断的介绍，例如，描述系统段合并状况，ElasticSearch 在高级 API 背后是如何工作以及如何限制 I/O 操作的。然而，本书并不仅限于讨论 ElasticSearch 的底层机制，同时也涵盖了如何改进用户搜索体验，例如处理拼写检查，高效地输入自动提示以及如何改进查询等内容。

除了前面介绍的那些，本书还将指导读者熟悉 ElasticSearch 的 Java API，并演示它的使用方法，其中不仅包含 CRUD（增删查改）等基本功能，同时也包含集群、索引的维护与操作等高级功能。最后，读者将通过开发一个用于数据索引的自定义 river 插件，以及一个在检索期和索引期用于数据分析的自定义分析插件来深入了解 ElasticSearch 的扩展机制。

本书主要内容

第 1 章介绍 Apache Lucene 的工作方式，以及 ElasticSearch 的基本概念，并演示 Elastic-

Search 的内部工作机制。

第 2 章描述 Lucene 评分过程是如何工作的，为什么要进行查询重写，以及查询二次评分（rescore）是如何工作的。除此之外，还将介绍 ElasticSearch 的批处理 API，以及如何使用过滤器（filter）来优化查询。

第 3 章描述如何修改 Lucene 评分，并使用不同的倒排索引格式来改变索引字段的结构。此外还会介绍 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。

第 4 章介绍以下技术：如何选择恰当的索引分片及复制（replicas）数量，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。同时还介绍 ElasticSearch 如何进行系统初始配置，以及当数据量和查询量急剧增长时如何调整系统配置。

第 5 章介绍如何为具体应用选择正确的目录（directory）实现，什么是发现（Discovery）、网关（Gateway）、恢复（Recovery）模块，如何配置这些模块，以及有哪些令人困扰的疑难点。最后介绍如何通过 ElasticSearch 来查看索引段信息，以及如何进行 ElasticSearch 缓存机制的调优。

第 6 章介绍 JVM 垃圾收集的工作原理和重要意义，以及如何对它进行调优。同时还介绍如何控制 ElasticSearch 的 I/O 操作数量，什么是预热器（warmer）以及如何使用它，最后介绍如何诊断 ElasticSearch 中的问题。

第 7 章介绍查询建议（suggester），它能帮助修正查询中的拼写错误以及构建高效的自动完成（autocomplete）机制。除此之外，将通过实际的案例展示如何使用不同查询类型和 ElasticSearch 的其他功能来提高查询相关性。

第 8 章覆盖 ElasticSearch 的 Java API，不仅包括一些基本 API，诸如连接到 ElasticSearch 集群、单条索引或批量索引、检索文档等，而且涵盖 ElasticSearch 暴露的一些用于控制集群的 API。

第 9 章通过演示如何开发你自己的河流（river）和语言处理（language）插件来介绍 ElasticSearch 的插件开发。

阅读本书的必备资源

本书基于 ElasticSearch 0.90.x 版本，所有范例代码均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，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从而保证所选择的工具能正确解析它。

除此之外，为了运行第 8 章和第 9 章的范例，要求已安装 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）。书中这两章都使用 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虽然熟悉 ElasticSearch 基本概念但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，对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具备了相关背景知识。这种情况下，建议参考 Packt 出版社上一本关于 ElasticSearch 的图书《ElasticSearch Server》。

客户支持

亲爱的读者，请随时浏览 http://www.elasticsearchserverbook.com，这里列出了本书最新的勘误表，以及相关的扩展阅读。

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。如果你采用了其他购买方式，可登录 http://www.packtpub.com/support 并注册账号，我们将通过 E-mail 将代码发送给你。

Acknowledgements 致　谢

Rafał Kuć 的致谢

本书正是我在完成《ElasticSearch Server》一书以后的下一个写作目标。幸运的是，我顺利实现了这个目标。我并不想逐一介绍所有主题，而是精选了一部分来阐述和分享我所了解的知识。与《ElasticSearch Server》类似，我也不会在本书中囊括所有的主题，毕竟很多小细节并不是那么重要（这依赖具体的使用案例），因此会忽略这部分内容。尽管如此，我还是希望读者能轻松获取所有 ElasticSearch、Apache Lucene 的相关知识细节，并能轻松快速地掌握感兴趣的知识。

在此，我想感谢我的家庭，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，他们是我最坚强的后盾。

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

最后，非常诚挚地感谢所有 ElasticSearch、Lucene 项目的创建者和开发者，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这种活力。再次感谢！

Marek Rogoziński 的致谢

像往常一样，撰写本书是件非常艰巨的任务。这本书不仅涉及更多的高级话题，同时 ElasticSearch 的代码也在随时改进。ElasticSearch 的开发速度并不会变缓，可以毫不夸张地说，每天都会有新东西呈现。请记住，本书是前一本著作的补充和延续，因此，这意味着我们会忽略上一本著作中已经涉及的内容，并补充该书遗漏的内容。现在看看你是否会成功吧！感谢大家。

感谢 ElasticSearch、Lucene 及所有相关产品的创建者。

同时也要感谢本书的写作和出版团队。尤其要感谢帮助检查错误、校稿、消除表达歧义的伙伴们。

最后，感谢在本书写作期间给予我坚定支持的所有的朋友。

About the Authors 作者简介

Rafał Kuć 是一个很有天资的团队领袖及软件开发人员，现任 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 和 Hadoop stack 等，拥有超过 11 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。他主要侧重于 Java 平台，但对能提高研发效率的任何其他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 和 Lucene 的相关问题。他还是世界范围内各种会议热邀的演讲嘉宾，曾受邀出席过 Lucene Eurocon、Berlin Buzzwords、ApacheCon、Lucene Revolution 等会议。

Rafał 最早于 2002 年接触 Lucene，一开始他并不喜欢这个开源产品，然而在 2003 年再次使用 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，目前主要关注 Lucene、Solr、ElasticSearch 和信息检索等方面。

Rafał 是《 Solr 3.1 Cookbook 》一书及其后续版本《 Solr 4.0 Cookbook 》的作者，同时也是 Packt Publishing 出版的所有版本的《 ElasticSearch Server 》的合著者之一。

Marek Rogoziński 是一个有着 10 多年经验的软件架构师和咨询师，专注基于开源搜索引擎（如 Solr、ElasticSearch 等）的解决方案和大数据分析技术（Hadoop、HBase、Twitter Storm 等）。

他是 solr.pl 网站的联合创始人之一，该网站致力于提供 Solr 和 Lucene 的相关资讯，同时他也是 Packt Publishing 出版的《 ElasticSearch Server 》的作者之一。

Marek Rogoziński 还是一家提供流式大数据处理和分析产品的公司的 CTO。

评审者简介 *About the Reviewers*

Ravindra Bharathi 有着 10 多年的软件工业从业经验，涉及多个领域，如教育、数字媒体营销 / 广告、企业级搜索、能源管理系统等。兴趣涉及基于搜索的应用软件，包括数据的可视化、插件定制、数据报表等。个人博客地址：http://ravindrabharathi.blogspot.com。

感谢我的妻子 Vidya，感谢她对我事业的所有默默付出。

Surendra Mohan 目前在一个印度知名软件咨询公司担任 Drupal 咨询师和架构师。他在加入该公司之前，曾在印度的一些跨国公司服务，并担任过各种角色，如程序员、技术 Leader、项目 Leader、项目经理、解决方案架构师、服务发布负责人等。他拥有 9 年左右的 Web 技术研发经验，涉及媒体、娱乐、房地产、旅游、出版、在线学习、企业级架构等多个领域。他是知名的演讲者，技术涉及 Drupal、开源产品、PHP、Moodle 等。同时他也是印度孟买各种 Drupal 科技会议、活动的组织者和发布者。

Surendra Mohan 也是一些书籍的评审者，这些书包括《 Drupal 7 Multi Site Configuration 》《 Drupal Search Engine Optimization 》《 Building e-commerce Sites with Drupal Commerce Cookbook 》。除了做技术评审之外，他还撰写了一本关于 Apache Solr 的著作。

感谢我的家人和朋友们，正是他们对我的不懈支持和鼓励，我才能保质保量完成我的图书评审工作。

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。他在高校任职的同时，也参与了一些与 Oracle、大数据相关的外部项目。其中 Oracle 相关项目有：Oracle 手册文档翻译、多媒体培训等。技术背景涉及数

据库、Web、Java 等。在 XML 领域，他因为参与 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而为业界所熟知。

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和经验的成员为该社区所做的贡献。

目　录 *Contents*

第 1 章 Chapter 1

ElasticSearch 简介

我们希望读者通过阅读本书能获取和拓展关于 ElasticSearch 的基本知识，并假设读者已经知道如何使用 ElasticSearch 进行单次或批量索引创建，如何发送请求检索感兴趣的文档，如何使用过滤器缩减检索返回文档的数量，以及使用切面 / 聚合（faceting/aggregation）机制来计算数据的一些统计量。不过，在接触 ElasticSearch 提供的各种令人激动的功能之前，仍然希望读者能对 Apache Lucene 有一个快速了解，因为 ElasticSearch 使用开源全文检索库 Lucene 进行索引和搜索，此外，我们还希望读者能了解 ElasticSearch 的一些基础概念，以及为了加快学习进程，牢记这些基础知识，当然，这并不难掌握。同时，我们也需要确保读者能按 ElasticSearch 所需要的那样正确理解 Lucene。本章主要涵盖以下内容：

- Apache Lucene 是什么。
- Lucene 的整体架构。
- 文本分析过程是如何实现的。
- Apache Lucene 的查询语言及其使用方法。
- ElasticSearch 的基本概念。
- ElasticSearch 内部是如何通信的。

1.1 Apache Lucene 简介

为了全面理解 ElasticSearch 的工作原理，尤其是索引和查询处理环节，对 Apache Lucene 的理解显得至关重要。揭开 ElasticSearch 神秘的面纱，你会发现它在内部不仅使用 Apache Lucene 创建索引，同时也使用 Apache Lucene 进行搜索。因此，在接下来的内容中，我们将展示 Apache Lucene 的基本概念，特别是针对那些从未使用过 Lucene 的读者们。