TURING 图灵程序设计丛书

Cassandra

权威指南

Cassandra: The Definitive Guide

[美] *Eben Hewitt* 著
Jonathan Ellis 序
王旭 译

O'REILLY®

人民邮电出版社
POSTS & TELECOM PRESS

TURING 图灵程序设计丛书

Cassandra权威指南

Cassandra: The Definitive Guide

[美] Eben Hewitt 著
Jonathan Ellis 序
王 旭 译

O'REILLY®
Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo
O'Reilly Media, Inc.授权人民邮电出版社出版

人民邮电出版社
北 京

图书在版编目（CIP）数据

Cassandra权威指南 / (美) 休伊特 (Hewitt,E.) 著;
王旭译. -- 北京 : 人民邮电出版社, 2011.8
(图灵程序设计丛书)
书名原文: Cassandra : The Definitive Guide
ISBN 978-7-115-25854-0

Ⅰ. ①C… Ⅱ. ①休… ②王… Ⅲ. ①关系数据库－数据库管理系统 Ⅳ. ①TP311.138

中国版本图书馆CIP数据核字(2011)第129426号

内 容 提 要

本书是一本广受好评的 Cassandra 图书。与传统的关系型数据库不同，Cassandra 是一种开源的分布式存储系统。书中介绍了它无中心架构、高可用、无缝扩展等引人注目的特点，讲述了如何安装、配置 Cassandra 及如何在其上运行实例，还介绍了对它的监控、维护和性能调优手段，同时还涉及了 Cassandra 相关的集成工具 Hadoop 及其类似的其他 NoSQL 数据库。

本书适合数据库开发人员与网站开发者阅读。

图灵程序设计丛书

Cassandra 权威指南

◆ 著 [美] Eben Hewitt
序 [美] Jonathan Ellis
译 王 旭
责任编辑 傅志红
执行编辑 李 盼

◆ 人民邮电出版社出版发行 北京市崇文区夕照寺街14号
邮编 100061 电子邮件 315@ptpress.com.cn
网址 http://www.ptpress.com.cn
北京艺辉印刷有限公司印刷

◆ 开本：800×1000 1/16
印张：19
字数：394千字 2011年8月第1版
印数：1－3 500册 2011年8月北京第1次印刷
著作权合同登记号 图字：01-2011-0812号
ISBN 978-7-115-25854-0

定价：59.00元

读者服务热线：(010)51095186转604 印装质量热线：(010)67129223
反盗版热线：(010)67171154

O'Reilly Media, Inc.介绍

O'Reilly Media 通过图书、杂志在线服务、调查研究和会议等方式传播创新知识。自 1978 年开始，O'Reilly 一直都是前沿发展的见证者和推动者。超级极客们正在开创着未来，而我们关注真正重要的技术趋势——通过放大那些“细微的信号”来刺激社会对新科技的应用。作为技术社区中活跃的参与者，O'Reilly 的发展充满了对创新的倡导、创造和发扬光大。

O'Reilly 为软件开发人员带来革命性的“动物书”，创建第一个商业网站（GNN）；组织了影响深远的开放源代码峰会，以至于开源软件运动以此命名；创立了 Make 杂志，从而成为 DIY 革命的主要先锋；公司一如既往地通过多种形式缔结信息与人的纽带。O'Reilly 的会议和峰会聚集了众多超级极客和高瞻远瞩的商业领袖，共同描绘出开创新产业的革命性思想。作为技术人士获取信息的选择，O'Reilly 现在还将先锋专家的知识传递给普通的计算机用户。无论是通过书籍出版，在线服务或者面授课程，每一项 O'Reilly 的产品都反映了公司不可动摇的理念——信息是激发创新的力量。

业界评论

“O'Reilly Radar 博客有口皆碑。”

——Wired

“O'Reilly 凭借一系列（真希望当初我也想到了）非凡想法建立了数百万美元的业务。”

——Business 2.0

“O’Reilly Conference 是聚集关键思想领袖的绝对典范。”

——CRN

“一本 O’Reilly 的书就代表一个有用、有前途、需要学习的主题。”

——Irish Times

“Tim 是位特立独行的商人，他不光放眼于最长远、最广阔的视野并且切实地按照 Yogi Berra 的建议去做了：‘如果你在路上遇到岔路口，走小路（岔路）。’回顾过去 Tim 似乎每一次都选择了小路，而且有几次都是一闪即逝的机会，尽管大路也不错。”

——Linux Journal

译者序

对于一位分布式存储系统的开发者，Cassandra 无疑是非常引人注目的，它的无中心架构、高可用性、无缝扩展等继承自亚马逊 Dynamo 的特质，相对于其他主从架构的 NoSQL 系统更加简洁，也更具有美感。

我从 2010 年初开始关注这个系统，并翻译过几篇 Cassandra 相关的文章，还引起一些读者热烈的讨论。2010 年底，当刘江老师为本书寻找译者时，我按捺不住，毛遂自荐，并随后在 2011 年 1 月中下旬开始了本书的翻译工作。我用了三个月的业余时间，终于在 4 月份完成了译稿。因为 Cassandra 仍在快速开发中，翻译时我也尽力争取快一些，以便能让中文版出版时不至于落伍。

本书对 Cassandra 的概念、架构、配置、使用进行了全面的介绍，非常详尽，而且给出了很多参考信息。对于希望了解 Cassandra、评估 Cassandra 是否是适合自己的应用，以及开始着手在 Cassandra 上进行应用开发的人都是不错的读物。当然，如果想参与 Cassandra 的开发或做更深入的工作，还需要直接通过源代码来获取更详尽的信息。

在翻译中，我尽力使用已有的、被广泛接受的名词或术语，对于一些译法没有被广泛接受的术语，在不产生歧义的前提下，我会选择一个自以为恰当的词，有时还会给出英文，以避免读者不能将代码和本书给出的名词对应上。还有一些名词尚没有贴切的中文译法，或是译出容易产生歧义，或是国内开发者已习惯使用英文，这时我在翻译中保留了英文原文。这些选择都以帮助理解、避免歧义为首要考虑。

本书的翻译工作得到了很多朋友和网友的关注，希望没有让他们久等。我的同事郭磊涛，作为数据库和 HBase 的专家、Cassandra 用户，在本书的翻译过程中给予了很多

有益的帮助。感谢现在 CSDN 的刘江老师，给我这个机会把 Cassandra 介绍给大家。当然，还要感谢图灵的编辑杨海玲、傅志红，还有李松峰在本书的翻译过程中做了大量的细心工作。

希望本书的翻译出版能对读者进入 NoSQL 的世界、开始自己的 Cassandra 应用有些许的帮助。

序

Cassandra 是 Facebook 于 2008 年 7 月开源的项目。它最早的版本主要是由一位亚马逊前雇员和一位微软的工程师写成的。这个系统受到了亚马逊前卫的键 / 值存储系统 Dynamo 的巨大影响。Cassandra 实现了 Dynamo 风格的副本复制模型和没有单点失效的架构，但增加了更为强大的“列族”数据模型。

当年 12 月，在 Rackspace 要求我帮他们建立一个可扩展的数据库的时候，我加入到这个项目之中。那是个很好的时机，因为今天所有重要的开源可扩展数据库在那时都有了，可以做做比较。尽管最初 Cassandra 只有一个主要的应用案例，但它的底层架构是最强大的，于是，我致力于改进代码，同时建立一个社区。

之后，Cassandra 被接纳为 Apache 的孵化器项目，并于 2010 年 3 月毕业成为顶级项目。此时它已经成为了一个真实的开源软件的成功案例，Rackspace、Digg、Twitter 等公司都成了忠实的用户，他们不愿意从零开始写自己的数据库，但却希望一起来构建一个更优秀的系统。

今天的 Cassandra 已经远不止是当初那个（现在也还在）用来驱动 Facebook 的收件箱搜索的系统了，按照 Tony Bain 的说法，它已经成为了“事务处理性能的不二赢家”，而且在可靠性和可扩展性方面具有显赫的声誉。

随着 Cassandra 逐渐成熟并获得了更多的主流用户，我们显然有为它提供商业支持的需要，于是，Matt Pfeil 和我在 2010 年 4 月共同创立了 Riptano。帮助推动 Cassandra 的应用具有丰富的回报，特别是可以看到更多的还没有被公开讨论过的应用。

另一个需求就是一本关于 Cassandra 的书。和很多开源项目一样，Cassandra 的文档一直就是一个弱项。而且即使是文档最终得到了改善，一本这样的书仍然会非常有用。

感谢 Eben 来承担这项集艺术与科学于一身的艰巨任务，讲解 Cassandra 的开发与部署。读者朋友现在有机会可以有条理地学习这些新概念了。

——Jonathan Ellis

Apache Cassandra 项目主席、Riptano 联合创始人

前言

选择Apache Cassandra

Apache Cassandra是一个自由、开源的分布式数据存储系统，与传统的关系型数据库管理系统截然不同。

Cassandra在2009年1月成为了Apache基金会的一个孵化器项目。不久，以Apache Cassandra项目主席Jonathan Ellis为首的开发者们发布了Cassandra 0.3，随后稳定不断地发布新的小版本。虽然Cassandra在本书完成时仍然没有达到1.0发布版本，但已经被互联网领域的很多巨头使用在了生产系统之中，他们包括Facebook、Twitter、Cisco、Rackspace、Digg、Cloudkick、Reddit等。

因为它非常出色的技术特性，Cassandra已经变得非常受欢迎了。它具有持久性、无缝扩展性、可调的一致性。它的写操作非常快，可以存储上百TB数据，而且是无中心的和对称的，所以不会有单点失效。它还是高度可用的，提供了无schema的数据模型。

目标读者

本书适用于各类读者。它对以下读者都会非常有用。

- 大规模、高容量网站的开发者，比如Web 2.0的社交应用。
- 需要理解这个高性能、无中心、弹性数据存储系统的应用架构师或数据架构师。
- 希望理解如何实现容错、最终一致的数据存储系统的标准关系型数据库系统管理员或开发者。

- 希望了解 Cassandra 的优势（和不足）以及其他相关的列数据库，以帮助进行技术路线选择的管理者。
- 正在进行 Cassandra 或其他非关系型数据库相关项目的学生、分析师或研究员。

本书是一本技术指南。从某种意义上说，Cassandra 代表了一种对数据的新的思考。在过去的 15 ~ 20 年间，很多合格的职业开发者都在使用纯粹的关系型或是面向对象的术语来描述他们的数据。Cassandra 的数据模型与此非常不同，起先可能很难吸引你，特别是对于数据库（应该）是什么已经有了先入为主的概念的人，更是如此。

使用 Cassandra 并不意味着你必须成为一个 Java 开发者。不过，Cassandra 是用 Java 开发的，所以若要深入分析源代码，你需要对 Java 语言有更坚实的理解。虽然不一定需要懂得 Java，但 Java 可以帮助你更好地了解异常、学会如何编译源码以及使用一些流行的客户端。本书中的很多例子都是用 Java 写成的。尽管如此，因为 Cassandra 使用了语言中立的 RPC 接口，所以你可以使用多种语言来开发 Cassandra 应用，包括 C#、Scala、Python 以及 Ruby 等。

最后，本书假设读者已经了解了 Web 是如何工作的，能够使用集成开发环境，并对数据驱动的应用的典型问题有某些了解。你可能是一个经验丰富的开发者或管理员，但是对于在 Cassandra 的世界里使用到的工具可能偶尔也不是非常熟悉。比如 Cassandra 使用 Apache Ivy 进行编译，而用一个流行的客户端（Hector）使用 Git 进行版本管理。当我感到你可能需要自己进行一些设置才能运行一个例子的时候，我会尽量予以说明。

本书的结构

本书把每章合理地设计为一个个独立的指南。因为本书是介绍 Cassandra 的，读者们可能背景各异，而且技术变化很快，所以这么处理非常重要。借用一个软件界的说法，我希望本书能够有点儿“模块化”。如果你是一个 Cassandra 新人，那么可以按照顺序阅读；而如果你已经有所了解，不需要介绍了，那么也可以在后面的章节里找到有价值的内容，把它们当做独立的指南来看。

本书的具体结构是这样的。

- 第 1 章 Cassandra 概况
  这一章介绍了 Cassandra，并讨论了它与众不同的特质、优势和目前的用户。
- 第 2 章 安装 Cassandra
  在这一章中，作者会带你在不同平台上安装 Cassandra。

- 第 3 章 Cassandra 的数据模型
  这里，我们介绍了 Cassandra 的数据模型以了解 Cassandra 中的列、超级列、行都是什么。我们特别介绍了 Cassandra 和传统的关系型数据库之间的差别。

- 第 4 章 应用实例
  这一章给出了一个完整可用的例子，将一个大家熟悉的领域中的应用实例从关系模型迁移到了 Cassandra 的数据模型之上。

- 第 5 章 Cassandra 的架构
  这一章会帮你理解在 Cassandra 进行读写操作时，到底都发生了什么，这个数据库是如何做到它的那些特点的，比如持久性和高可用性。我们深入到底层来了解一些更复杂的内部工作机制，比如 gossip 协议、提示移交、读时修复、Merkle 树等。

- 第 6 章 配置 Cassandra
  这一章介绍了如何设置分区器、副本放置策略和 snitch。我们配置了一个集群，了解不同配置选项对于集群的影响。

- 第 7 章 读写数据
  这是我们一直期待的时刻。这里介绍了 Cassandra 模型在查询和更新数据时与传统关系型数据库的不同，然后还使用 API 进行了操作。

- 第 8 章 客户端
  第三方开发者为 Cassandra 开发了很多不同的客户端，支持多种语言，包括 Java、C#、Ruby、Python 等，对 Cassandra 的底层 API 进行了再次抽象。我们会帮你从整体上了解这些客户端，这样你就可以选择一个适合自己的了。

- 第 9 章 监控
  一旦集群已经配置好并开始运行了，就需要监控它的利用率、内存占用和线程状况，了解它的日常行为。Cassandra 内建了丰富的 Java 管理扩展（JMX）接口，我们可以监控所有这些信息，甚至更多。

- 第 10 章 维护
  通过服务器自带的一些工具，可以更简单地进行很多 Cassandra 集群的日常维护工作。我们会看到如何退服一个节点，对集群进行负载均衡，获取统计信息以及进行其他日常维护操作任务。

- 第 11 章 性能调优
  Cassandra 的一个最值得一提的特性就是它的速度——非常地快。但有很多东西，包括内存设置、数据存储、硬件选择、缓存和缓冲区大小等，都需要进一步调优，从中获得更高的性能。

- 第 12 章 集成 Hadoop
  这一章由 Jeremy Hanna 写作。在这一章我们会把 Cassandra 放到一个更大的背景中，学习如何将它与 Hadoop 集成在一起，Hadoop 是 Google 的 Map/Reduce 算法目前一个十分流行的实现。

- 附录
  很多新的数据库都在今日海量数据的需求之下应运而生了，有的从“无 schema”模型中获益，有的支持更新的一些趋势，如语义网络。这里我们把 Cassandra 放到各种流行的非关系型数据库背景之中，分别了解面向文档的数据库、分布式哈希表、图数据库等，来更好地理解 Cassandra 所提供的东西。

- 词汇表
  理解一些确实很新的东西是相当困难的，Cassandra 中有些名词对于关系型应用的开发者和 DBA 来说可能非常陌生，我编写了一个词汇表，来方便大家阅读本书。如果某个概念让你不知所云，可以翻到词汇表来了解诸如 Merkle 树、向量时钟、提示移交、读时修复和其他生僻的名词。

本书针对 Cassandra 0.6 和 0.7 写成。项目组正在努力开发 Cassandra，新的小版本和修订版本会不断释出。在可能的地方，我会尽量解释版本间的不同，不过你在阅读时可能已经用上了一个更新的版本，有些实现因此会有所不同。

更多信息

如果你需要了解关于 Cassandra 的更多信息，获得最近的更新，可以访问本书网站：http://www.cassandraguide.com。

在 Twitter 上关注我（@ebenhewitt）也是个好主意。

格式约定

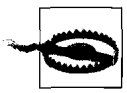
本书使用了如下排版约定。

- 楷体
  用于标记新名词。

- 等宽字体
  用于程序代码，在段落中用于表示程序的组成部分，如变量或函数名、数据库、数据类型、环境变量、语句、关键字。

- 等宽粗体
  命令或是其他应该由用户输入的内容。
- 等宽斜体
  应该由用户提供的或由上下文确定的值。

这个图标表明一个提示、建议或一般注记。



这个图标表示一个警告或警示。

使用示例代码

本书用于帮助你完成工作。通常，你可以在程序或文档中使用本书提供的代码。除非你在重新发布我们的大量代码，否则不需要联系我们来获得许可。比如，在程序中使用本书代码的一些片段是无需我们许可的。但是出售或再分发 O'Reilly 的图书示例光盘显然是需要授权的。引用本书或引用示例代码来回答问题是不需要授权的，但使用本书的大量示例代码作为你的产品的文档是需要授权的。

我们乐于见到你在使用时声明引用信息，但不强制要求。引用信息通常包括书名、作者、出版社和 ISBN。比如："*Cassandra: The Definitive Guide* by Eben Hewitt. Copyright 2011 Eben Hewitt, 978-1-449-39041-9."

如果你认为对示例代码的使用需要授权，请通过这个邮箱联系我们 permissions@oreilly.com。

Safari®在线图书

Safari 在线图书是应需而变的数字图书馆。它能够让你非常轻松地搜索 7500 多种技术性和创新性参考书以及视频，以便快速地找到需要的答案。

订阅后就可以访问在线图书馆内的所有页面和视频。可以在手机或其他移动设备上阅读。还能在新书上市之前抢先阅读，也能够看到还在创作中的书稿并向作者反馈意见。复制粘贴代码示例、放入收藏夹、下载部分章节、标记关键点、做笔记甚至

打印页面等有用的功能可以节省大量时间。

这本书也在其中。欲访问本书的英文版电子版，或者由 O'Reilly 或其他出版社出版的相关图书，请到 http://my.safaribooksonline.com 免费注册。

我们的联系方式

请把对本书的评论和问题发给出版社。

美国：

O'Reilly Media, Inc.
1005 Gravenstein Highway North
Sebastopol, CA 95472

中国：

北京市西城区西直门南大街 2 号成铭大厦 C 座 807 室（100035）
奥莱利技术咨询（北京）有限公司

O'Reilly 的每一本书都有专属网页，你可以在那儿找到关于本书的相关信息，包括勘误表、示例代码以及其他的信息。本书的网站地址是：

http://www.oreilly.com/catalog/0636920010852/

中文书

http://www.oreilly.com.cn/index.php?func=book & isbn=

对于本书的评论和技术性的问题，请发送电子邮件到：

bookquestions@oreilly.com

关于本书的更多信息、会议、资源中心和网络，请访问以下网站：

http://www.oreilly.com
http://www.oreilly.com.cn

致谢

在此，我希望感谢很多帮助我们完成本书的优秀人士。

感谢 Jeremy Hanna 写作了 Hadoop 的章节，也感谢他如此地易于合作。

感谢本书的技术审校者们。特别是 Stu Hood 独到的批注，极大提高了本书的质量。

Robert Schneider 和 Gary Dusbabek 也给出了很有见地的审稿意见。

感谢 Jonathan Ellis，谢谢他为本书作序。

感谢我的编辑 Mike Loukides，旧金山晚餐时优雅的健谈者。

感谢 Rain Fletcher，他为本书提供了支持，一直鼓励着本书的写作。

我的灵感来自很多了不起的 Cassandra 开发者。向编写出这个优美而强大的数据库的开发者们致敬。

还要一如既往地感谢 Alison Brown，他阅读了手稿、给我提示，并确保我的写作时间，没有他，这本书就不会诞生。

目录