Broadview
www.broadview.com.cn

大数据丛书

大数据挑战与NoSQL数据库技术

陆嘉恒 编著

BIG DATA

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

大数据丛书

大数据挑战与NoSQL数据库技术

陆嘉恒 编著

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

内 容 简 介

本书共分为三部分。理论篇重点介绍大数据时代下数据处理的基本理论及相关处理技术，并引入NoSQL 数据库；系统篇主要介绍了各种类型 NoSQL 数据库的基本知识；应用篇对国内外几家知名公司在利用 NoSQL 数据库处理海量数据方面的实践做了阐述。

本书对大数据时代面临的挑战，以及 NoSQL 数据库的基本知识做了清晰的阐述，有助于读者整理思路，了解需求，并更有针对性、有选择地深入学习相关知识。

图书在版编目（CIP）数据

大数据挑战与 NoSQL 数据库技术 / 陆嘉恒编著. —北京：电子工业出版社，2013.4
（大数据丛书）
ISBN 978-7-121-19660-7

Ⅰ. ①大… Ⅱ. ①陆… Ⅲ. ①数据处理②数据库系统 Ⅳ. ①TP274 ②TP311.13

中国版本图书馆 CIP 数据核字(2013)第 035612 号

责任编辑：许　艳
印　　刷：三河市鑫金马印装有限公司
装　　订：三河市鑫金马印装有限公司
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：27.5　字数：446.8 千字
印　　次：2013 年 4 月第 1 次印刷
印　　数：4000 册　定价：79.00 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。
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。
服务热线：（010）88258888。

前　言

为什么写本书

计算机技术已经深刻地影响了我们的工作、学习和生活。大数据及 NoSQL 技术是现今 IT 领域最炙手可热的话题，其发展非常迅速，潜力巨大，悄然改变着整个行业的面貌。随着 Web 2.0 技术的发展，微博、社交网络、电子商务、生物工程等的不断发展，各领域数据呈现爆炸式的增长，传统关系型数据库显得越来越力不从心。NoSQL 数据库技术的出现为当前面临的问题提供了新的解决方案，它摒弃了传统关系型数据库 ACID 的特性，采用分布式多节点的方式，更加适合大数据的存储和管理。

政府和高校都十分重视对大数据及 NoSQL 技术的研究和投入；在产业界，各大 IT 公司也在投入大量的资源研究和开发相关的 NoSQL 产品，与之相应的新兴技术和产品正在不断涌现。这一切都极大地推动了 NoSQL 技术的发展。

大数据处理和 NoSQL 技术涉及的内容繁多，目前不同公司也有不同的 NoSQL 数据库产品，而且某一产品往往是为特定的应用而设计的，并不一定能够适用于所有的场景。很多人在学习的初始阶段需要进行大量的摸索和实践，然而目前这方面系统的参考资料却非常少。为了便于所有想了解和掌握 NoSQL 技术的朋友学习，并在学习的过程中少走弯路，笔者将自己在该领域的经验和知识的积累凝聚在本书，希望能够推动大数据处理及 NoSQL 相关技术在国内的发展。

本书面向的读者

在编写本书时，我们力图使不同背景和职业的读者都能从其中获益。

如果你是专业技术人员，本书将带领你快速地进入大数据处理及 NoSQL 的世界，全面掌握 NoSQL 及其相关技术，帮助你使用 NoSQL 技术解决面临的问题或提供必要的参考。

如果你是高等院校计算机及相关专业的学生，本书为你在课堂之外了解最新的 IT 打开一扇窗，帮助你拓宽视野，完善知识结构，为迎接未来的挑战做好知识储备。

在学习本书之前，应具有如下的基础：

- 有一定的 Linux 操作系统的基础知识。
- 有较好的编程基础和阅读代码的能力。
- 对数据库知识有一定的了解。

如何阅读本书

本书一共包括 16 章，分为三个部分。其中第一部分为理论篇，包括：大数据产生的背景、数据一致性理论、数据存储模型、数据分区与防治策略、海量数据处理方法、数据复制与容错技术、数据压缩技术和数据缓存技术。此部分重点从理论上介绍、分析大数据管理过程中遇到的各方面问题。第二部分为系统篇，包括：键值数据库、列存数据库、文档数据库、图存数据库、基于 Hadoop 的数据库管理系统、NoSQL 数据库以及分布式缓存系统。该部分以理论篇为基础，根据数据存储模型对数据库类型进行划分，每一部分以具体开源数据库为实例进行介绍，涉及系统的架构、安装以及使用等方面知识，力图使读者对 NoSQL 数据库有具体的认识。第三部分为应用篇，包括企业应用以及经验总结和对未来的展望。该部分介绍企业如何使用 NoSQL 数据库解决自身遇到的问题。

在阅读本书时，读者可以先系统地学习理论篇的知识，对海量数据处理方法有一个很好的理解，在此基础之上，读者可以对后面的章节进行选择性的学习。本书涉及内容较多，从开源数据库方面讲，包括了 Dynamo、Redis、Voldemort、Cassandra、Hypertable、CouchDB、MongoDB、Neo4j、GraphDB、OrientDB、HBase、Hive、Pig、MySQL Cluster、VolteDB、MS-Velocity、Memcached 等将近 20 个数据库。因此，建议读者可以重点学习感兴趣或有一定需求的数据库系统。当然，如果时间允许，还是建议读者系统地学习本书的内容。

另外，在系统篇的学习过程中，建议读者能够一边阅读，一边根据书中的指导进行实践，亲自实践本书中所给出的编程范例。

致谢

在本书的编写过程中，还有很多 NoSQL 领域的实践者和研究者为本书做了大量的工作，他们是张林林、许翔、程明、王海涌、顾向楠、吴少辉、杨宁、杨华、吴梦迪、任乔意、於洋、张轩等，在此特别感谢。

在线资源及勘误

本书官方网站为：http://datasearch.ruc.edu.cn/NoSQL/。本书的勘误、讨论以及相关资料等

都会在该网站上发布和更新。

在本书的撰写和相关技术的研究中，尽管笔者投入了大量的精力，付出了艰辛的努力，然而受知识水平所限，错误和疏漏之处在所难免，恳请大家批评指正。如果有任何问题和建议，可发送邮件至 jiahenglu@gmail.com 或 jiahenglu@ruc.edu.cn。

陆嘉恒

目　录

理 论 篇

系 统 篇

应 用 篇

第1章 概 论

“这是最好的时代，这是最坏的时代；这是智慧的时代，这是愚蠢的时代；这是信仰的时期，这是怀疑的时期；这是光明的季节，这是黑暗的季节；这是希望之春，这是绝望之冬；人们面前什么都有，人们面前一无所有；人们正在直登天堂，人们正在直下地狱。”

——狄更斯《双城记》

对于数据管理界来说，这是一个充满挑战的时代。急速增长的数据让人们焦头烂额，传统关系型数据库在扩展性方面的瓶颈让人们无所适从：如何存储大数据，如何处理大数据，如何挖掘大数据，大数据已经成为数据管理界的新挑战。这又是一个充满机遇的时代，新的系统孕育而出，百花齐放，它们“标新立异”，它们“独树一帜”，它们在数据模型、事务处理等方面采取不同的策略解决海量数据带来的问题。这注定是一段不平凡的岁月。