

Oracle Big Data Handbook

ORACLE

BIG DATA

Oracle

大数据解决方案

规划和实施企业大数据基础架构

Tom Plunkett

[美] Brian Macdonald 等著

Bruce Nelson

许向东 李园花

杨雷 王欣

译

大数据应用与技术丛书

Oracle 大数据解决方案

Tom Plunkett

[美] Brian Macdonald 等著

Bruce Nelson

许向东 李园花

杨 雷 王 欣

译

清华大学出版社

北 京

Tom Plunkett, Brian Macdonald, Bruce Nelson, et al.
Oracle Big Data Handbook
EISBN: 978-0-07-182726-3
Copyright © 2014 by McGraw-Hill Education.

All Rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including without limitation photocopying, recording, taping, or any database, information or retrieval system, without the prior written permission of the publisher.

This authorized Chinese translation edition is jointly published by McGraw-Hill Education and Tsinghua University Press Limited. This edition is authorized for sale in the People's Republic of China only, excluding Hong Kong, Macao SAR and Taiwan.

Copyright © 2014 by McGraw-Hill Education and Tsinghua University Press Limited.

版权所有。未经出版人事先书面许可,对本出版物的任何部分不得以任何方式或途径复制或传播,包括但不限于复印、录制、录音,或通过任何数据库、信息或可检索的系统。

本授权中文简体字翻译版由麦格劳-希尔(亚洲)教育出版公司和清华大学出版社有限公司合作出版。此版本经授权仅限于中华人民共和国境内(不包括中国香港、澳门特别行政区和中国台湾地区)销售发行。

版权©2014 由麦格劳-希尔(亚洲)教育出版公司与清华大学出版社有限公司所有。

北京市版权局著作权合同登记号图字: 01-2013-8905

本书封面贴有 McGraw-Hill Education 公司防伪标签,无标签者不得销售。

版权所有,侵权必究。侵权举报电话: 010-62782989 13701121933

图书在版编目(CIP)数据

Oracle 大数据解决方案 / (美) 普伦凯特(Plunkett, T.), (美) 麦克唐纳(Macdonald, B.),
(美) 纳尔逊(Nelson, B.) 等著; 许向东 等译. —北京: 清华大学出版社, 2015
(大数据应用与技术丛书)

书名原文: Oracle Big Data Handbook

ISBN 978-7-302-38554-6

I. ①O… II. ①普… ②麦… ③纳… ④许… III. ①关系数据库系统 IV. ①TP311.138

中国版本图书馆 CIP 数据核字(2014)第 27648 号

责任编辑: 王 军 韩宏志

封面设计: 牛艳敏

版式设计: 方加青

责任校对: 邱晓玉

责任印制: 宋 林

出版发行: 清华大学出版社

网 址: <http://www.tup.com.cn>, <http://www.wqbook.com>

地 址: 北京清华大学学研大厦 A 座 邮 编: 100084

社 总 机: 010-62770175 邮 购: 010-62786544

投稿与读者服务: 010-62776969, c-service@tup.tsinghua.edu.cn

质 量 反 馈: 010-62772015, zhiliang@tup.tsinghua.edu.cn

印 装 者: 三河市中晟雅豪印务有限公司

经 销: 全国新华书店

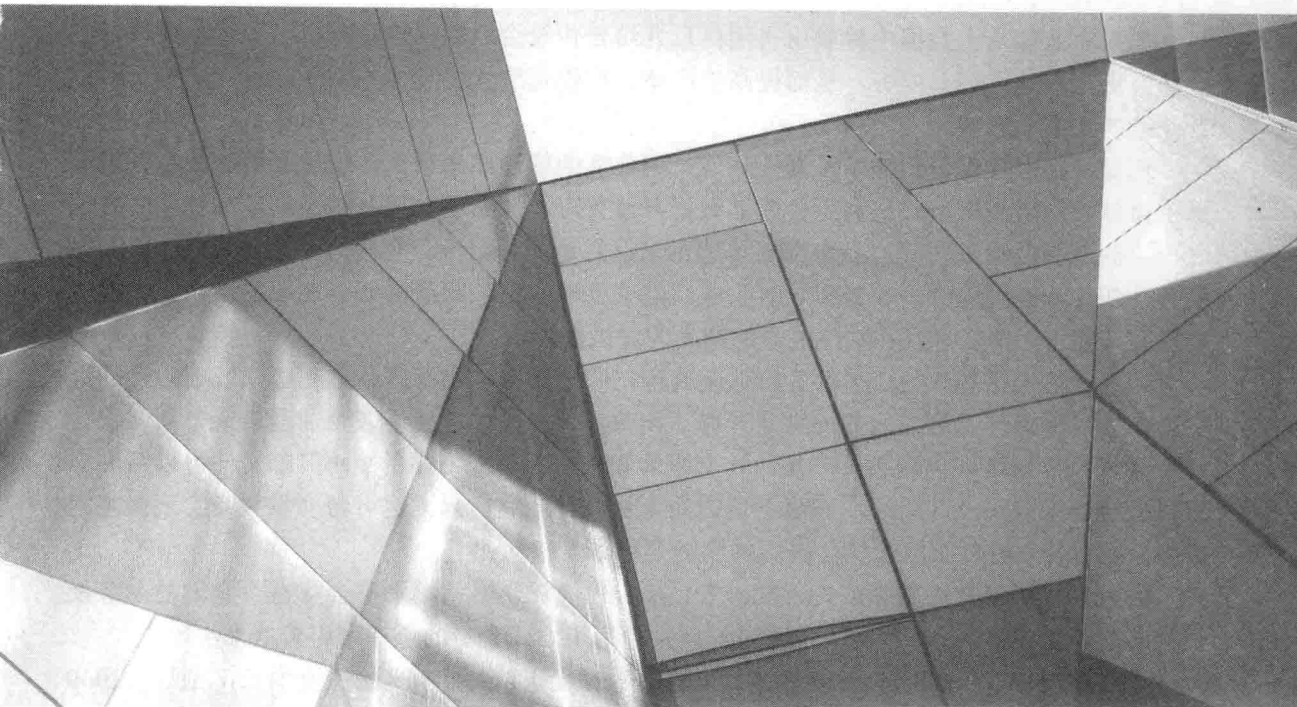
开 本: 185mm×260mm 印 张: 20.5 字 数: 499 千字

版 次: 2015 年 1 月第 1 版 印 次: 2015 年 1 月第 1 次印刷

印 数: 1~3500

定 价: 59.80 元

产品编号: 057240-01



译者序

近年来，“大数据”这一术语颇受关注。越来越多的企业都在考虑将传统企业数据(如CRM系统中的客户信息、事务型ERP数据、网店交易、总账数据等)与机器生成的数据(包括呼叫详情记录(CDR)、Web日志、智能仪表、传感器、设备日志、实时交易数据等)以及社交数据(包括客户反馈、Twitter等微博网站、Facebook等社交媒体、微信平台等)同时用于业务智能分析。

正如我们所看到的，机器生成的数据量远大于传统的业务数据量。例如，一架喷气发动机30分钟内即可生成10TB数据。如果每日的航班超过25 000架次，则仅此一个数据源每日即可生成PB级别的数据量。智能仪表和重型工业设备(如炼油设备和钻井平台)生成的数据量与此类似。

特别是社交媒体的数据流，虽然没有机器生成的数据量大，但会生成对客户关系管理有用的大量观点和关系。即使是在每条微博只有140个字符的情况下，Twitter数据的生成速度(或频率)之高也足以产生大量数据(每天超过8TB)。这些数据的类型和结构更是多种多样(包括文本、图像、声音、文档等)，而这些大量的非结构化数据中往往隐藏着有用的信息。

如果企业能从大数据中提取出有用信息并结合传统企业数据进行分析,那么企业可以更全面、更深入地了解其业务,从而提高生产率、增强竞争优势和加强创新,这一切都会对经营业绩产生重大影响。

例如,在医疗保健服务中,慢性病或长期疾病的管理成本很高。使用家用监测设备来测量生命体征和监测病情只是利用传感器数据来改善患者健康状况以及减少门诊和住院的方法之一。制造企业通过在其产品中部署传感器来返回遥测数据流。有时这是为了提供OnStar等服务,OnStar提供通信、安全和导航服务。但更重要的是,此遥测还会展现使用模式、故障率以及其他可以降低开发和装配成本的产品改进机会。

随着智能电话和其他GPS设备的迅速流行,广告商也能将商店、咖啡店或饭店附近的消费者作为目标群体。这为服务提供商开辟了新的收入来源,并使许多企业得以锁定新客户。通常零售商都知道谁会购买其产品。社交媒体和来自其商业网站的Web日志文件可以帮助零售商了解谁未购买及其原因,而眼下他们尚无从获得此类信息。这可帮助零售商实现微观客户细分、有针对性的市场营销并提高供应链效率。

而且,如果没有大数据,Facebook和LinkedIn等社交网站根本不会存在。它们的业务模式需要个性化Web体验,而这只能通过捕获并使用用户或会员的所有可用数据来实现。

据McKinsey Global Institute估计,全球的数据量正以每年40%的速度增长,预计从2009年到2020年之间将增长44倍。

为从大数据中获得真正的业务价值,企业需要适当的工具来从不同的数据源捕获和组织多种数据类型,并要能结合其所有企业数据轻松分析这些非结构化数据。同时,企业必须改进其IT基础架构,以处理快速生成且类型多样的海量数据,从而实现这些非结构化数据与要进行分析的其他企业数据的有效集成。

也正是为了满足企业对各种大数据的分析处理需求,Oracle在2011年的OpenWorld全球用户大会上,宣布推出大数据机和Oracle Big Data Connectors,成为首家提供全面、集成的大数据解决方案供应商,以满足企业对各种大数据需求的处理和分析。

Oracle大数据战略的核心思想是使企业能通过改进其当前的企业数据架构来引入大数据和提供业务价值。通过改进当前的企业架构,企业可利用Oracle系统无可置疑的可靠性、灵活性和性能来满足大数据需求。

本书涵盖了从Oracle的角度来看与大数据相关的很多主题,包括Oracle大数据机(Oracle Big Data Appliance)、大数据连接器(Big Data Connectors)、数据库云服务器(Exadata)、商务智能云服务器(Exalytics)、Oracle R Enterprise、Oracle NoSQL等。

本书的11位原作者Tom Plunkett、Brian Macdonald和Bruce Nelson等,都是来自Oracle Big Data产品管理和研发部门的技术权威和专家,有着丰富的大数据架构设计、Hadoop、NoSQL、Java、MapReduce、Oracle R Enterprise、Business Intelligence、Data Warehousing、ETL等方面的专业经验。

作为第一本较为全面地介绍Oracle大数据整体解决方案的专业书籍,希望这本汇集了原作者和译者最佳实践经验和技术的中文版译著,能够帮助中国的用户和技术人员更深入地了解Oracle大数据机和产品的精髓。同时欢迎读者提出宝贵意见和建议。



Wei Hu (胡伟)
Vice President of Development,
Oracle Corporation

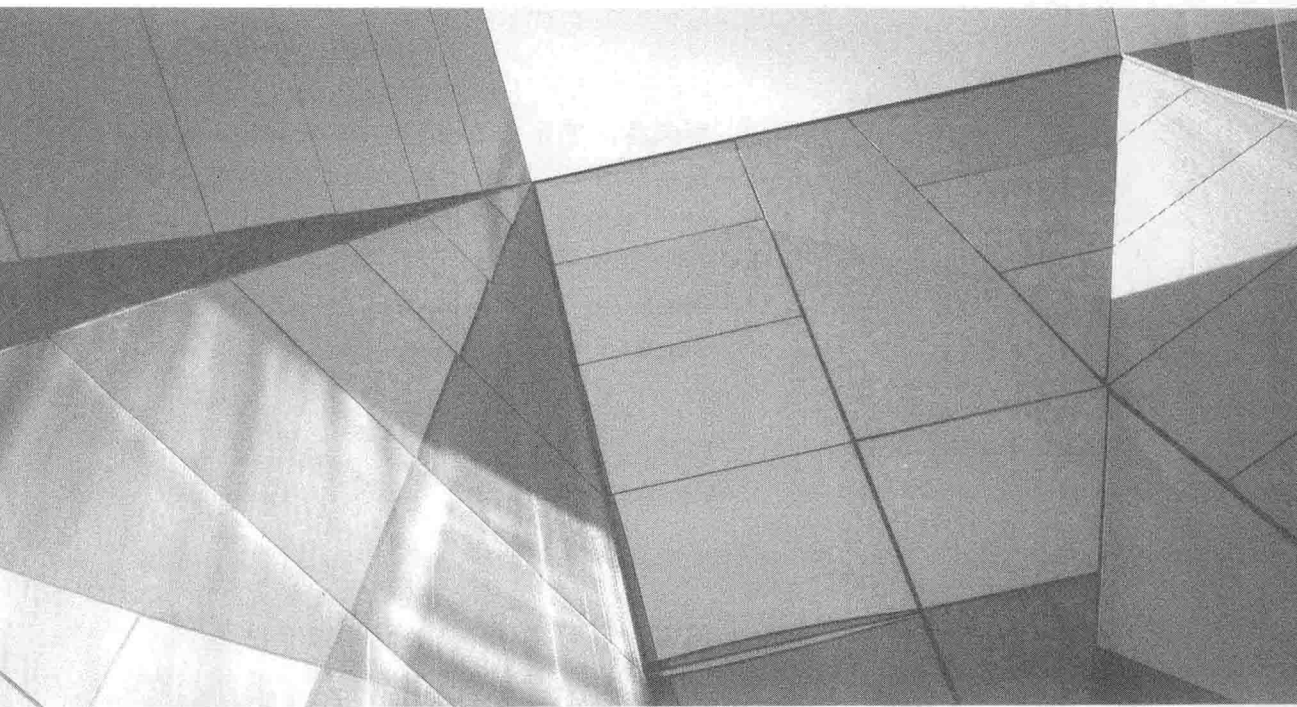


Sally Piao (朴红玉)
Vice President of Server Technology,
Oracle Corporation

致谢

为方便中国读者更好地阅读和理解该书的内容，清华大学出版社与甲骨文(中国)软件系统有限公司技术产品售前部门的4位技术顾问(许向东、李园花、杨雷、王欣)共同合作，顺利地完成了本书的中文翻译。在整个翻译过程中，得到了清华大学出版社王军老师、韩宏志编辑和工作人员的大力帮助和指导，在此深表谢意！

本书的译者均来自甲骨文(中国)软件系统有限公司售前技术产品团队，他们利用业余时间完成了本书的翻译。在翻译过程中，也得到部门领导、同事和家人的大力支持，在此深表谢意！



译者简介



许向东，甲骨文(中国)软件系统有限公司技术产品售前总监，2002年加入Oracle公司，支持过金融、政府、交通、电力等行业的Oracle数据库及相关选件产品的设计、开发、运维管理和培训工作。目前带领团队致力于金融行业的技术产品售前支持工作，包括Oracle Exadata数据库云服务器、OEM企业运维管理平台工具、Oracle大数据一体机、数据仓库、商务智能等的技术支持工作。



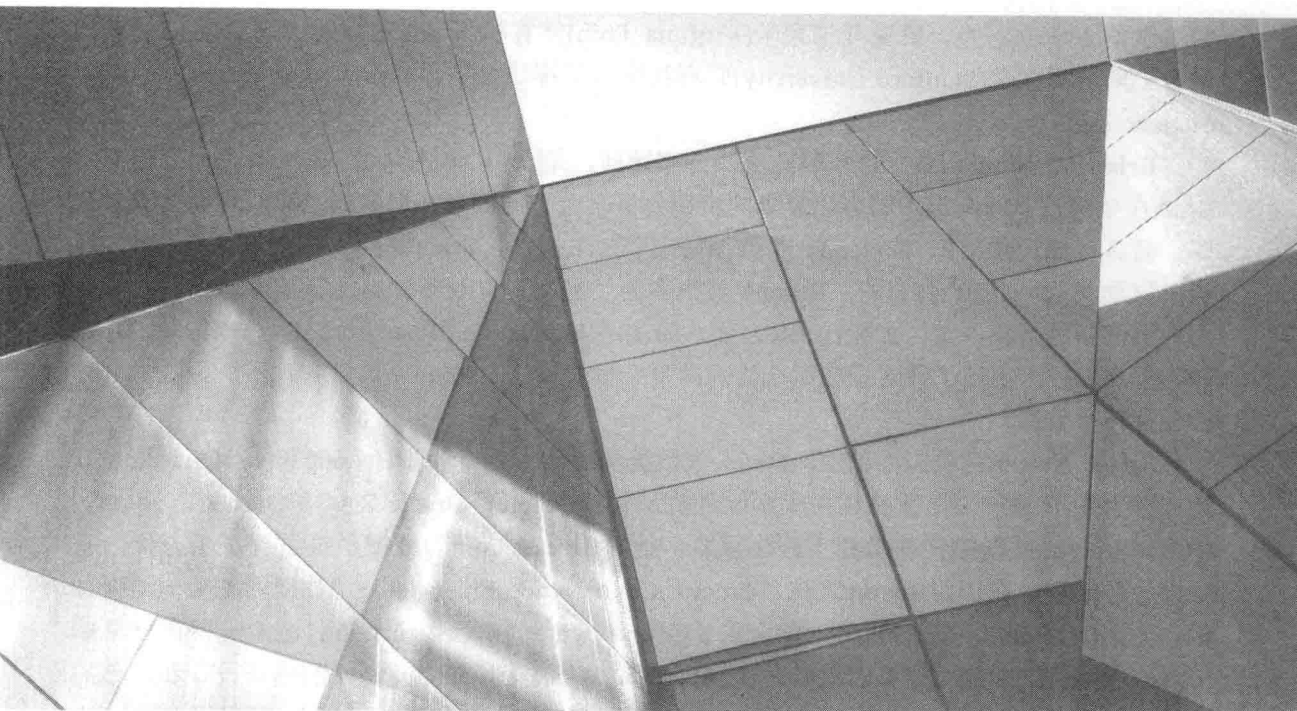
李园花，甲骨文(中国)软件系统有限公司技术产品售前高级顾问，有多年互联网大数据经验，专注于Hadoop大数据处理技术、Oracle NoSQL技术，以及Hadoop平台和关系型数据库(RDBMS)之间的数据集成工具Oracle Big Data Connectors、Oracle大数据一体机产品Oracle Big Data Appliance的技术支持。



杨雷，甲骨文(中国)软件系统有限公司技术产品售前资深顾问，负责金融行业的大数据解决方案和Oracle大数据一体机、Exadata、Oracle Advanced Analytics等产品支持。曾参与港口运输、航空、电力和金融等行业的商务智能项目实施。之前在埃森哲咨询公司工作过3年，有数据仓库、商务智能、OLAP、主数据管理、ETL技术、数据挖掘和Hadoop等设计和实施经验。



王欣，甲骨文(中国)软件系统有限公司技术产品售前高级顾问，拥有10多年的电信行业数据仓库设计、开发和实施经验。目前主要从事Exadata、Oracle大数据一体机、Oracle Advanced Analytics、OBIEE等技术产品和企业信息架构、大数据、数据仓库和商业智能等解决方案的支持工作。曾在Teradata工作多年，参与过中国移动多省经营分析系统建设和集团规范编写。



作者简介

大数据包含很多种技术，因此本书所有的作者都精通各自负责的对应的Oracle大数据产品。该团队在架构设计、大数据、商务智能、Hadoop、Java、MapReduce和平台设计等方面都拥有丰富经验。下面详细介绍每个作者的具体信息。

Tom Plunkett是多本书的主要作者，包括*Oracle Exalogic Elastic Cloud Handbook*。2009年，Tom带领一个团队为美国国防部办公室实施过一个大数据研究项目。2010年，Tom与其他人合作为弗吉尼亚理工大学的计算机科学系讲授了一学期的MapReduce和其他大数据主题的课程。2012年，Tom帮助Frederick国家癌症研究实验室(Frederick National Laboratory for Cancer Research)凭借Oracle大数据一体机在分析基因与癌症亚型之间的关系上赢得多项行业大奖，包括2012年政府大数据解决方案奖(从80多个提名的大数据项目中脱颖而出)，作为卓越创新者入围2013年政府卓越创新决赛，并获2013计算机世界创新奖。除了为Oracle工作外，Tom还担任弗吉尼亚理工大学计算机科学研究生的兼职讲师，并提供远程学习指导。Tom在大数据和云计算领域发表过40多次国际会议演讲。之前，Tom在IBM工作过并实施了Fliesler Meyer专利法。Tom拥有乔治·梅森大学(George Mason University)的文学学士

和法学博士学位、弗吉尼亚理工大学(Virginia Tech)计算机科学的理学硕士，并且已经参加了学习了斯坦福大学(Stanford University)管理科学与工程研究生课程(Management Science and Engineering)。

Brian Macdonald是一位卓越的解决方案顾问，是获得Oracle认证的Oracle企业架构师。Brian在架构设计和满足不同客户需求的实施分析平台方面拥有20多年的经验，包括大数据仓库、商务智能、OLAP、Hadoop、主数据管理和ETL技术。Brian曾在数百家公司工作过，包括多家顶级的世界500强公司。Brian参与很多为了提高现有业务应用的新技术研究，并为客户提供创新的解决方案。加盟Oracle之前，Brian在Information Resources公司工作，用OLAP和数据仓库技术实现了针对贷款组合管理的复杂数学算法。Brian还非常喜欢用分析技术来分析各种体育数据。

Bruce Nelson是美国西部地区Oracle大数据的负责人，专注于Hadoop和NoSQL。他在IT行业拥有超过24年的高性能数据库系统经验，包括Oracle、Oracle RAC和Oracle Exadata。作为Bizrate的数据库管理员和工程化总监，他对Bizrate.com的数据系统进行了全方位的重设计并实施。在加盟Bizrate之前，Bruce是Countrywide Financial公司数据系统架构的副总裁，他把Oracle和Linux的商用硬件引入了Countrywide公司。他还负责过制定一些围绕早期Hadoop和商用的HPCC计算的举措。在Yahoo时，Bruce带领几个开创性的重新设计和迁移的项目，包括Yahoo的高性能Oracle RAC从SAN迁移到NFS上。1993年，Bruce开发了FileNet公司的互联网在线业务，创造了当时全球第一个几千元注册的Web网站之一。他接着创建了第一个数据库驱动交互式客户网站，该网站支持客户在互联网上开放他们的服务票。

Helen Sun是一位大数据的理论家、信息架构师，同时还是通过认证的Oracle企业架构师。Helen拥有15年以上的业务和技术管理经验，涉及财务、医疗保健、市场研究和供应链管理。她主要的专业领域是企业数据管理和信息架构，包括主数据(MDM)、数据整合(Data Integration)、商务智能/数据仓库(BI/DW)和大数据。在Oracle工作期间，Helen协助很多大型企业完成了复杂业务和IT架构的切换。Helen是*Oracle Information Architecture Framework and Development Process*的主要作者，并出版过关于信息管理、数据治理和大数据的白皮书。Helen举办过一些主题研讨会，并曾在在线论坛和甲骨文全球大会(Oracle OpenWorld)上担任过发言人。她还曾在全球大数据开放组会议上发表过大数据的公开演讲。

Mark F. Hornick是一位Oracle数据库高级分析(Oracle Database Advanced Analytics)组的总监，专注于Oracle R Enterprise(ORE)、Oracle R Connector for Hadoop(ORCH)和Oracle R Distribution(ORD)。他同时为Oracle内部与外部的客户在Oracle数据库、Exadata或大数据一体机(Big Data Appliance)使用R的应用提供技术支持和专业建议，同时也会参与一些针对R与SAS的竞争分析和性能测试。Mark是*Java Data Mining: Strategy, Standard, and Practice*一书的合著者。他于1999年Oracle收购Thinking Machines Corp时加入了Oracle数据挖掘技术组。Mark是IOUG商务智能仓库和分析(BIWA)SIG的创始人之一，目前担任Oracle顾问。他在Oracle OpenWorld、Collaborate、BIWA Summit和R用户大会userR上做了很多培训。Mark先后获得Rutgers大学(Rutgers University)的计算机专业学士学位和布朗大学(Brown University)的计算机专业硕士学位。

Keith Laker在Oracle已经工作了15年以上，现在是数据仓库和大数据的高级产品经

理。他拥有丰富的数据、数据仓库和商务智能经验，而且承担多种工作角色，包括售后咨询、客户支持和产品管理。Keith在全球很多Oracle用户大会上发表过演讲，并且交付了很多演讲稿，涉及数据仓库、数据整合、多维建模和大数据等主题。在加盟Oracle之前，Keith在Information Resources公司工作，并为其实施了涉及全球客户范围的OLAP和数据仓库系统。

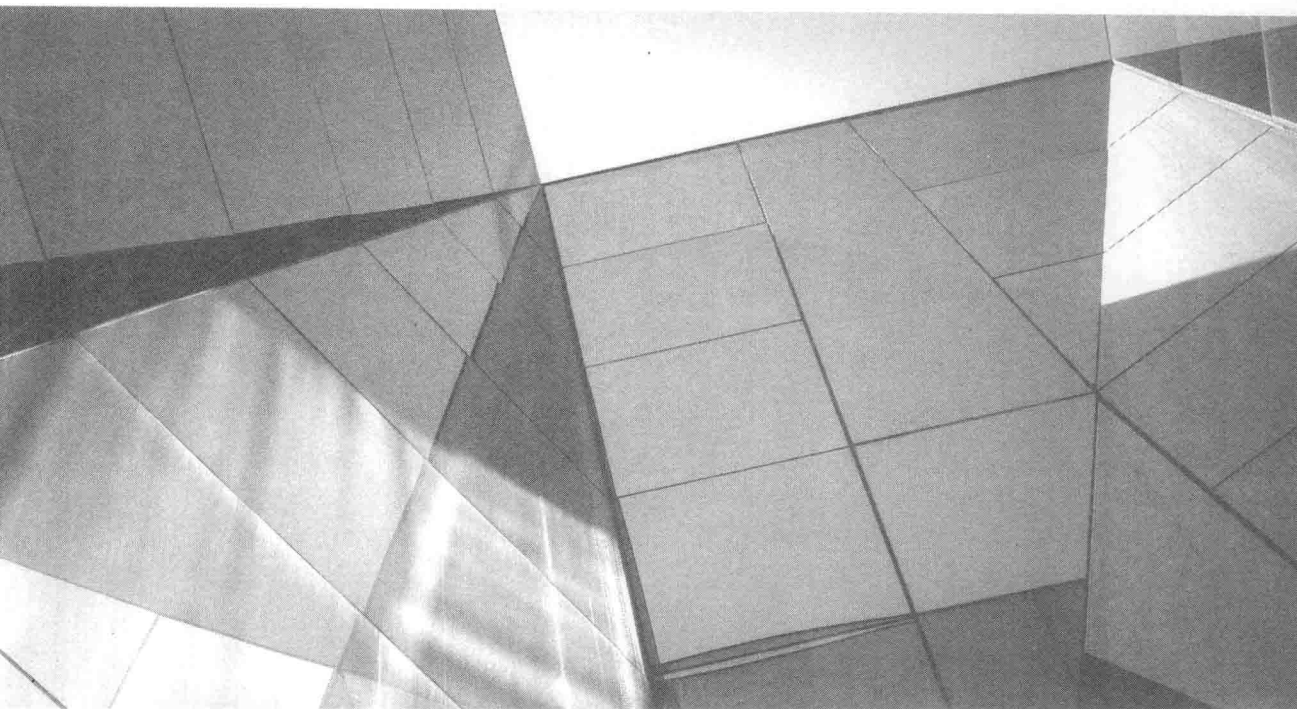
Khader Mohiuddin是Oracle公司位于北美中部地区的大数据负责人。Khader在Oracle工作了6年时间，然后在Sun Microsystems工作6年时间，在Accenture工作过3年，并于2010年再次加入Oracle。他现在的工作职责是利用他20年的企业软件、硬件和行业咨询经验设计合理的大数据解决方案，从而解决客户复杂的业务问题。Khader是一位信息架构专家，在石油天然气、电力、电信、财务和自动化工业等领域拥有丰富的解决方案开发经验，并开发实现了一个涵盖数据仓库、高级分析和高并发在线系统的创造性解决方案。Khader于1996年第一次加入Oracle，供职于Oracle咨询部门位于纽约的分部，在此期间，他为几大主要电信公司解决了复杂数据库系统的扩展问题。他在1997年转到位于加利福尼亚州Redwood Shores的Oracle数据库研发组，为全球范围的客户设计并赢得了很多基准测试。他在2002年作为高级工程师加入Sun Microsystems，负责在Sun平台的内核上优化Oracle数据库和Oracle企业应用。在最初的Exadata架构设计、整合Sun硬件和Oracle数据库技术理论论证阶段，他扮演了关键角色。Khader在埃森哲工作期间，作为一位业务转型架构人员，为世界100强的客户设计了私有云/公有云、DbaaS和IaaS等。他帮助这些公司运用最新的技术实现了业务的高效运营，从而节省了数百万美元的成本。此外，Khader还写过一些白皮书并在一些当地的Hadoop用户大会上发表过演讲。

Debra L. Harding是一位拥有20年以上工作经验的技术专家，她具备丰富的与世界500强一起合作的行业经验，并且这些公司都把他的想法付诸实现。作为Oracle业务架构师和大数据先驱，Debra与决策管理团队一起合作改善情况，降低维护成本和复杂性，从而获得竞争优势。作为一个狂热的旅行家，Debra已经完成了跨悉尼港的大桥攀登，并尽情享受在法国阿尔卑斯山的滑雪时刻，最近她和她的小女儿完成了在阿拉斯加的哈丁冰原上的徒步旅行。然而，Debra最成功的经历却是癌症第4阶段的幸存者，正是这段经历促使她在生命科学行业运用了大数据分析技术来驱动医学的发展。在2008年，仅有2%的人被确诊为患有这种类型的癌症，这就意味着可以被用来收集和研究用于治疗该癌症的信息非常少。Debra认为通过大数据分析技术的能力分析更多的数据集，将会更加事半功倍，科学家们将从那些更小的模式中识别出异常情况，并且大家一起合作来找到更好的办法从而改变生活。

David Segleau是Oracle的产品总监，负责管理Oracle NoSQL数据库、Oracle Berkeley DB和Oracle数据移动服务器(Oracle Database Mobile Server)。他以Sleepycat Software (makers of Berkeley DB)工程学副总裁的身份加入Oracle。他在该行业拥有30年以上的经验，负责和管理技术产品团队。他具备丰富的数据库技术能力，既做过客户，也当过供货商。David做过很多技术公司的管理人员，其中包括Britton-Lee、ShareBase、Teradata、Illustra、Informix、ANTs Software、Sleepycat Software和目前所在的Oracle。David花费了他职业生涯绝大部分时间去开创和创新技术。他非常享受在做产品研发时，会同时考虑客户的业务需求和产品开发的专业知识，因此创造了能解决实际问题的产品和技术。

Gokula Mishra是Oracle公司的副总裁，并负责全球市场的高级分析、大数据团队和IBU。Gokula在业务驱动IT战略、企业信息管理、大数据、数据仓库和商务智能、大数据和高级分析架构以及解决方案的设计、开发和管理实施方面拥有25年以上的经验。他的行业经验涵盖零售、消费品、石油天然气、能源、金融服务、医疗保健和生命科学，以及制造业。他是一位实践家，也是高级分析和大数据、主数据、数据质量和数据管理的领导者。在加入Oracle之前，Gokula曾经服务过Sears Holdings、HP、TransUnion、EW Solutions、Sierra Atlantic、Booz & Co.和ZS Associates。Gokula获得BITS Pilani大学电子和电子工程的工学学士学位(荣誉)，以及西北大学(Northwestern University)计算机科学系的理学硕士学位。

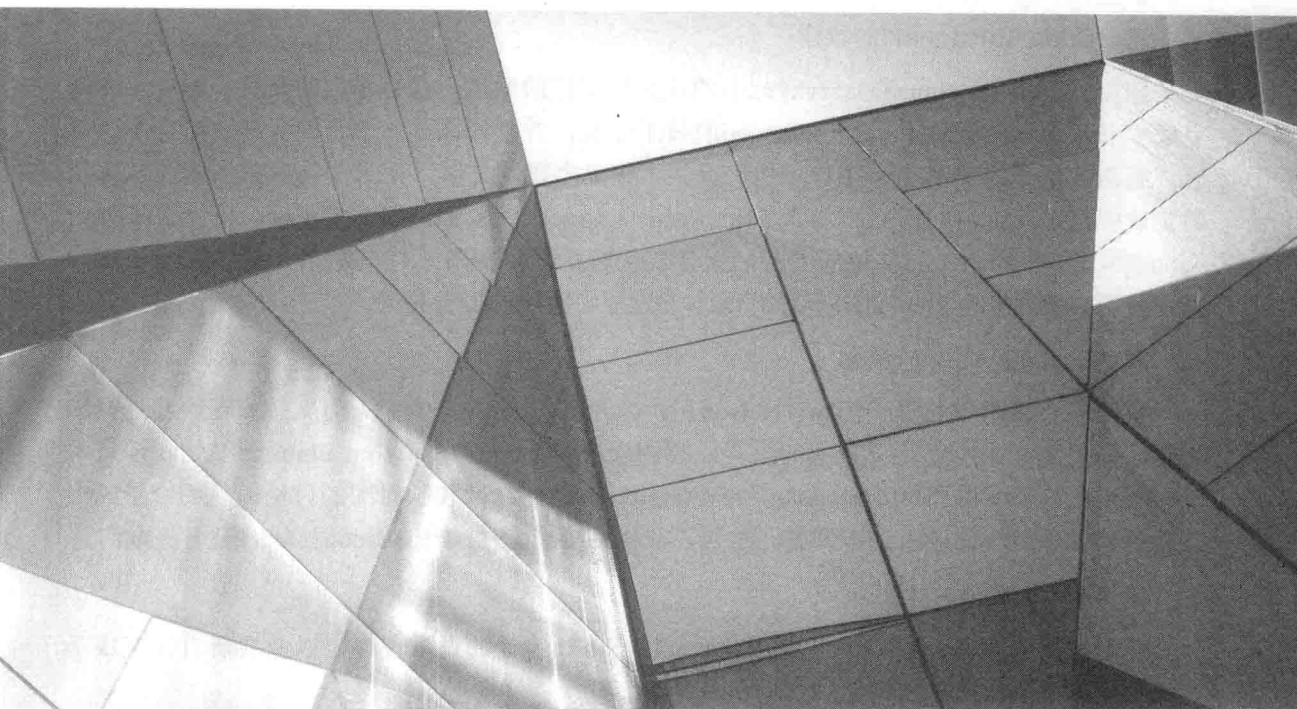
Robert Stackowiak是Oracle公司企业架构和信息架构部门的副总裁。他在Oracle数据仓库、大数据战略、商务智能技术和解决方案方面的经验是全球公认的。他在世界各地的大会上发表商务智能和数据库技术的演讲，并且撰写了很多关于商务智能和数据库技术的书籍和文章，其中包括*Achieving Extreme Performance with Oracle Exadata*(Oracle出版)，五个版本的*Oracle Essentials*(O' Reilly Media)和*Oracle Data Warehousing and Business Intelligence Solutions*(Wiley)。详细信息可以关注Twitter @rstackow。



技术编辑简介

Jean-Pierre Dijcks是Oracle服务器技术部门的高级资深产品经理，专注于研究大数据技术。攻读完荷兰特温特大学(University of Twente)的工业工程专业学位后，Jean-Pierre担任Oracle荷兰区数据仓库顾问，在此期间负责整个欧洲的数据仓库实施。2000年末，他调入Oracle产品开发部，担任Oracle Warehouse Builder的产品经理。2008年之后，Jean-Pierre一直担任Oracle数据库并行计算的产品经理。目前Jean-Pierre的主要工作方向是大数据和Hadoop，包括Oracle大数据一体机(Oracle Big Data Appliance)。

Dan McClary目前是Oracle公司的资深产品经理，主要负责大数据和Hadoop。在加盟Oracle之前，他担任位于加州Palo Alto的Red Robot实验室商务智能的总监。此前，他曾作为访问学者到美国西北大学和霍华德休斯医学研究所，在此期间，研究使用Hadoop技术处理复杂系统中的大型图形问题。McClary博士获得了亚利桑那州立大学(Arizona State University)计算科学系的博士学位，集中研究移动和即席应用网络的自适应优化问题。他拥有亚利桑那州立大学(Arizona State University)计算科学系硕士学位，研究分布式文件系统中的硬实时调度，他是imgSurf的创始人，这是一家生物识别和电子病历领域的公司。



致 谢

为这个项目大家付出了巨大的努力。在此感谢本书所有的作者，他们在自己本职工作之余付出了宝贵时间。同时感谢Oracle和支持本项目所有大数据邮件列表里的成员。感谢技术编辑JP和Dan，他们提供了很多建议和注解。感谢Oracle出版社、McGraw-Hill编辑和制作团队，尤其感谢Paul Carlstroem和Amanda Russell。感谢Doug Cutting、Margo Seltzer、Melliyal Annamalai、Peter Jeffcock和许多为此书提供过建议、反馈意见和材料的人。最后，有太多的人需要单独感谢，我们在此就不再一一提及了。

来自Tom Plunkett的致谢

感谢Laura、Daniel、Daphne、我的父母和我其他的家庭成员，因为为写此书而占用了我陪伴他们的时间。同时也感谢与我合作的作者、技术编辑、我的同事、我的管理团队、我的编辑和每一位为此书做过努力和付出的人。特别感谢Rizwan Jaka、Mark Comishock、Ken Currie、Peter Doolan、Mark C. Johnson、Lauren Farese、Mark A. Johnson和Franco Amalfi，感谢他们对我的支持和鼓励。

来自Brian Macdonald的致谢

感谢Yvette、Julianna和Jazzy对此书的理解和工作的激情。没有你们的支持，我肯定无法完成本书。着重感谢Melliyal Annamalai的技术专长，并且乐此不疲地帮助我解决所有细节问题。感谢我所有的同事和合作者，他们提供了很多宝贵的见解，而且作为我的智囊团从大量的建议中筛选有价值的意见；尤其是Jim Fisher、Bruce Nelson和Marty Gubar。我要感谢Nuge Ajouz和Craig Lockwood多年的支持，促使我仍有机会保持着工作的激情，最终完成了这本书。最后，感谢Tom Plunkett在我创作这本书过程中给予的专业指导。

来自Bruce Nelson的致谢

我要衷心感谢我的妻子Hilary Nelson和我的儿子Zachary，他们大力支持和理解我选择的道路，而且我如此痴迷信息技术的世界。感谢Brian Macdonald和Tom Plunkett鼓励我涉足本书的项目。我还要感谢Jeff Needham和Ed Gasiorowski的帮助和长时间的讨论，他们给予的很多见解和灵感最终都融入到本书里。最后，我们都要感激热衷于Hadoop社区的无名英雄们。

来自Helen Sun的致谢

感谢我丈夫William Smith和儿子Nathaniel Smith给予的真诚理解，以及我所有家庭成员在我写作本书过程中不断给予的支持和理解。

同时也感谢我的父母，我的父亲Huazhi Sun和心爱的已故的母亲Hongyuan Lin，他们始终给我灌输学习的热情和追求成功的精神，这些都不断地推动我追求更高的生活和工作目标。

来自Khader Mohiuddin的致谢

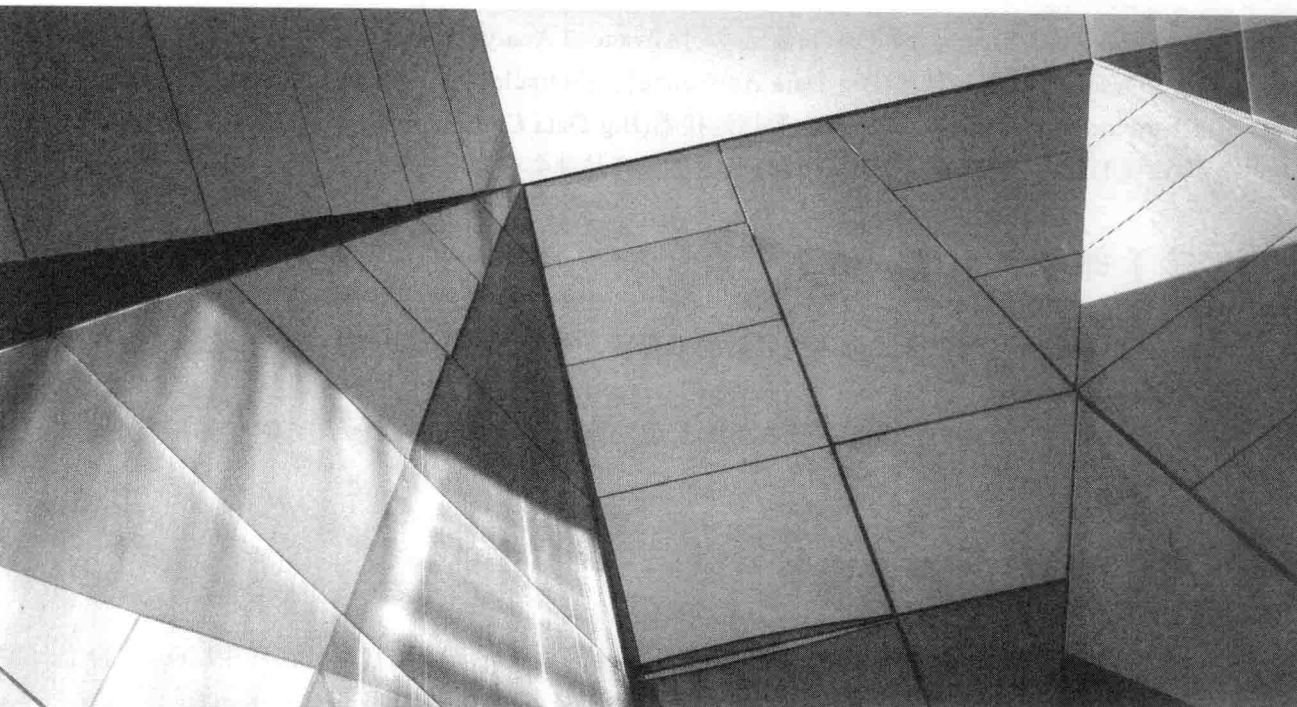
感谢我的父母和兄弟姐妹，他们的辛勤工作为我提供了良好的基础和指导，使我的生活和事业在这个阶段取得成功。感谢我的妻子和孩子包容我在写作这本书上用了大量假期和业余时间。我同时也感谢Jacco Draaijer、Jean-Pierre Dijcks、Dan McClary和Brian Macdonald对我所写材料的技术评阅，同时还要感谢一直鼓励我开拓进取的Tom Plunkett。还要感谢在Cenveo Publishing Services工作的Nidhi Chopra和在McGraw-Hill Professional的Oracle出版社的所有工作人员，他们负责管理并控制在每个时间点上把Oracle工程化系统中最新的创新产品特征加入到本书中。

来自David Segleau的致谢

感谢我的妻子Lynn持续不断的支持我占用了大量的业余时间去书写本书第8章。Lynn同时还是我的第一个审阅人和编辑。非常感谢Tom Plunkett、Dan McClary和J.P. Dijcks提供帮助和技术审阅。最后，感谢整个的Oracle NoSQL数据库工程化系统团队，他们为我提供了这款优秀产品的许多素材。

来自Gokula Mishra的致谢

感谢我的妻子Ratna和孩子们(Seetal、Jyoti和Samir)，他们给予很多支持、鼓励和理解，我因为写作管理章节而占用了大量的周末时间。同时我还要感谢与我一起合作的作者和同事的鼓励和支持。特别感谢Tom Plunkett的帮助和指导；特别感谢J.P. Dijcks和Dan McClary给我所写章节提供的反馈意见和点评；同时还要特别感谢Sunil Soaresfor给予的灵感。



前言

本书描述Oracle的大数据方法论。本书可作为所有对大数据感兴趣人士的参考手册，包括架构师、管理员和开发人员。为使介绍更全面，本书将更多地关注Oracle产品未来的版本。虽然有其他关于大数据和Apache Hadoop的书，但这是第一本专门阐述Oracle大数据方法论的书籍。阅读该技术手册之前，不需要读者事先具备任何Oracle大数据方法论的知识。该书里含有截屏、图和一些相关的照片。

在过去几年里(2010年到2013年)，大数据是最大的技术热点。数据以极快的速度在增长(部分分析师认为有60%的复合年增长率；另外一些人认为增长率超过100%)。相比传统的IT领域，越来越多的信息将要被存储和处理。基于Apache Hadoop的新技术能帮助企业分析不同类型数据。企业基于这些信息在激烈的竞争环境下做出决策时，往往要结合非结构化数据和结构化数据一起来分析，非结构化数据包括Facebook上的职位信息，Twitter上的用户情感信息、传感器、交通信息等，结构化数据则包括在线处理系统和传统的数据仓库。

Oracle最近又宣布了一些重要的新产品，这些产品都与大数据密切相关。Oracle数据库、Oracle Exadata数据库一体机、Exalytics一体机、Endeca Information Discovery、

Oracle NoSQL Database和Oracle数据库的Advanced Analytics Option都具备处理大数据的能力。Oracle大数据一体机(Big Data Appliance)包括Oracle硬件、Oracle NoSQL和Cloudera's Distribution of Hadoop。Oracle大数据连接器(Big Data Connectors)提供从HDFS到Oracle数据库连接的能力。所有这些产品,还有其他更多产品都会在本书中阐述。

第 I 部分: 引言

第1章: 大数据简介 描述大数据的重要性,为什么现在是热门技术,什么地方适合用到Oracle的技术。

第2章: 大数据的价值 描述大数据价值的定义。该章会提到一个深度大数据案例的列表,这些案例在后续章节会被引用。

第 II 部分: 大数据平台

第3章: Apache Hadoop平台 描述Hadoop的软硬件基础和它运行在商用Linux平台上的能力。

第4章: 选择Appliance的理由 提供运行在Oracle大数据一体机(Big Data Appliance)和自建Apache Hadoop集群(在常用硬件之上)的性能对比。该章节提供总拥有成本的计算(TCO)并讨论不同分析类型的影响。

第5章: BDA配置、部署架构和监控 描述配置选项:只选Hadoop、只选NoSQL、选Hadoop和NoSQL、多台BDA以及用BDA rack连接Exadata。同时将描述内存选项,按需分配容量、NameNode节点高可用性和多租户。该章还涵盖管理和监控Hadoop集群的工作。

第6章: 为大数据集成数据仓库和分析基础架构 描述为了扩展访问信息的能力,将大数据环境连接到传统的数据仓库系统。

第7章: BDA连接器 描述连接器是什么以及如何使用它们。该章包括代码示例、配置选项和性能对比。该章包含但不仅限于描述Oracle SQL Connector for HDFS和Oracle Loader for Hadoop的架构、优点/案例。

第8章: Oracle NoSQL数据库 概要介绍NoSQL技术和NoSQL应用,也包含NoSQL用例。该章涵盖Oracle NoSQL Database架构、API、操作功能和选项,还介绍安装、配置和运行。

第 III 部分: 分析信息和制定决策

第9章: 数据库库内分析: 快速交付彰显时间价值 介绍库内分析和预测分析。涵盖Oracle Data Mining、Oracle Data Miner、文本挖掘和数据挖掘算法、Oracle统计功能;介绍ORE、Oracle OLAP、spatial分析、semantic/graph分析以及从库内分析集成到数据转换。

第10章: 使用R分析数据 介绍Open Source R、R工作环境、IDE、来自于Comprehensive