WILEY

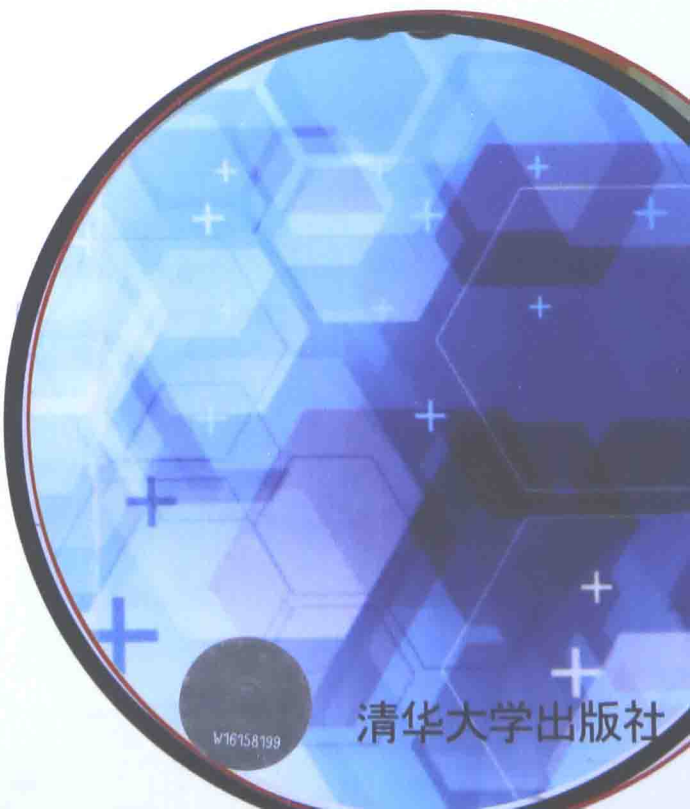
大数据应用与技术丛书

Microsoft Big Data Solutions

微软大数据解决方案

[美] Adam Jorgensen
James Rowland-Jones
John Welch
Dan Clark
Christopher Price
Brian Mitchell
著

王翔 杨道文 译



清华大学出版社

大数据应用与技术从书

微软大数据解决方案

[美] Adam Jorgensen
James Rowland-Jones
John Welch
Dan Clark
Christopher Price
Brian Mitchell
著

王翔 杨道文
译

清华大学出版社

北 京

Microsoft Big Data Solutions
Adam Jorgensen, James Rowland-Jones, John Welch, Dan Clark, Christopher Price, Brian Mitchell
EISBN: 978-1-118-72908-3

图书在版编目(CIP)数据

微软大数据解决方案/(美) 约根森(Jorgensen, A.) 等著；王翔，杨道文 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：Microsoft Big Data Solutions
ISBN 978-7-302-39652-9

Ⅰ. ①微… Ⅱ. ①约… ②王… ③杨… Ⅲ. ①企业管理—数据管理 Ⅳ. ①F270.7

中国版本图书馆 CIP 数据核字(2015)第095451号

责任编辑：王　军　韩宏志
装帧设计：孔祥峰
责任校对：邱晓玉
责任印制：何　芊

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 刷 者：北京鑫丰华彩印有限公司
装 订 者：三河市吉祥印务有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　印　张：20.25　　字　数：493 千字
版　　次：2015 年 5 月第 1 版　　印　次：2015 年 5 月第1 次印刷
印　　数：1～4000
定　　价：58.00 元

产品编号：059784-01

译者序

随着云时代的来临，大数据日益引起关注。从技术角度看，大数据与云计算像是一枚硬币的正反面。大数据必然无法使用单台计算机进行处理，必须依托云计算的分布式处理、分布式数据库和云存储等技术。大数据的战略意义不在于掌握庞大的数据信息，而在于对这些含有意义的数据进行专业化处理并得到有价值的产出。网络日志、社交媒体、机器数据、传感数据等都是大数据来源，大数据对社会生产和生活的影响已经十分显著，无论是银行、电信、铁路、航空，还是军事、政治、工业、商业，基于大数据的决策已成为现代社会各行业运行的基石。

以医疗行业为例，临床医学借助新兴科技的发展，进入了以科学和大数据为基础的现代医学时代。计算机图像处理技术与 X 光、超声、核磁共振技术相结合，促进了基于大数据的新型复杂成像技术的发展；客户服务与医疗设备传感器数据结合，创新地提出了基于大数据分析的预见性故障排查的新型客户服务理念。

大数据需要使用特殊技术进行有效处理。本书作者以清晰的思路、通俗易懂的语言和形象具体的实例为读者讲述了大数据的含义和作用，并重点介绍了微软的大数据平台解决方案。从手把手教导搭建你的第一个大数据环境开始，通过实景模拟将大数据相关技术娓娓道来。从数据存储结构到数据仓库，从 Hadoop、MapReduce 到 Hive、HBase、HCatalog，一个个晦涩的技术名词在你的脑海中逐渐清晰，最后介绍了数据仓库与 Hadoop 的集合，现实生活中如何使用和运营大数据，让你从一个大数据世界的“菜鸟”华丽蜕变为能理解微软平台大数据解决方案的“专家”。

本书适合有志于学习微软大数据解决方案的读者阅读，零基础也可以快速融入大数据的世界中；也适合大数据行业的专家总体回顾微软大数据解决方案的整体架构。

对于这本经典著作，译者力求完整表达出原作者的真实意图，希望为广大读者学习大数据贡献一份绵薄之力。如有任何意见和建议，请不吝指正。

本书全部章节由王翔、杨道文翻译，参与翻译活动的还有欧亚丽、唐亚丽、侯宗明、张晓丽、王宇思、卫莉、张寒、赵伟、曾途、马骁。同时要感谢清华大学出版社的编辑为翻译本书提供的有力支持。没有你们，也就没有本书的顺利出版。

作者简介

Adam Jorgensen，Pragmatic Works 总裁，SQL Server 专业协会(PASS)执行副总裁。在过去 13 年里，在 SQL Server、SharePoint 和分析学方面积累了丰富经验。现在主要帮助企业和高管通过新技术解决方案、管理技巧和财务优化实现业务增长。专注于研究云计算和大数据分析领域的解决方案，使其真正为企业服务。他与妻子 Cristina 共同居住在佛罗里达州的杰克逊维尔。

James Rowland-Jones，The Big Bang Data公司的首席顾问。专注于搭建和交付具有创意、简洁精巧又具有高度扩展性和分析能力的平台。同时还专注于研究融合SQL Server PDW 和Hadoop生态系统两者的大数据仓库解决方案。不管在国际上还是在英国本土，James都是SQL Server社区的忠心拥护者。目前是PASS的董事会成员、SQLBits组委会成员(微软数据平台在欧洲最大的项目)。因多年来服务于社区，在 2008 年获得微软MVP称号。

John Welch，目前在 Pragmatic Works 工作，负责 BI 产品套件的开发，使其解决方案在开发、管理和记录上更加简洁。自 2001 年起，John 就专注于 BI 和数据仓库技术，并且重点关注微软的异构环境产品。因坚持不懈地在 IT 社区分享知识经验，同样获得了微软 MVP 称号。同时还是位 SSAS 大师。John 是位经验丰富的演讲者，在 PASS 会议、微软的商业智能大会、西部软件发展会议(SD West)、软件管理大会(ASM/SM)等会议上做过演讲。参与撰写多本 SQL Server 书籍，如 *Smart Business Intelligence Solutions with Microsoft SQL Server 2008* (微软出版社，2009 年)和 *SQL Server MVP Deep Dives* (Manning 出版社)系列。John 在 http://agilebi.com/jwelch 上撰写了大量关于 BI 和 SQL Server Information Services (SSIS)的博客文章。他积极参与开源项目的开发，帮助 BI 开发者们简化开发过程，包括 ssisUnit(http://ssisunit.codeplex.com)，这是 SSIS 的一个单元测试框架。

Dan Clark，Pragmatic Works 高级 BI 顾问。热衷于学习新的 BI 技术，培训他人如何以最合理的方式实现 BI 技术。Dan 对数据与决策之间的关系尤其感兴趣。在.NET 编程和 BI 开发方面已经出版了一些书籍和大量文章。他定期在开发会议、BI 会议和用户组会议上发表演讲，并且非常喜欢与微软开发者和数据库社区互动。

Christopher Price，佛罗里达州坦帕市的微软高级顾问。拥有南佛罗里达大学的信息系统管理学士学位和工商管理硕士学位。最初是一名 VB 和 Java 程序员，后来成为一名 VB.Net 和 C#软件架构师，现在是一位 BI 顾问。虽仍热衷于软件开发，但目前重点从事 ETL(提取、转换和加载)、数据集成、数据质量、MDM(主数据管理)、SSAS(SQL Server 分析服务器)、SharePoint 以及所有与大数据相关的研究。经常在 SQL 星期六、PASS 峰会、代码阵营、和其他社区活动上发表演讲。除了经常更新博客，还撰写了多本书籍和白皮书，并担任一系列 BI 和大数据书籍的技术编辑。可在 http://bluewatersql.wordpress.com/和 @BluewaterSQL 上阅读到他的博客和推特。

Brian Mitchell，微软大数据专家中心的首席架构师。关注数据仓库/商业智能(DW/BI)解决方案，大部分时间专注于 SQL Server 并行数据仓库(PDW)和 HDInsight。拥有超过 15 年的 Microsoft SQL Server 和商业智能工作经验。获得微软 SQL Server 2008 大师认证。

在 http://brianwmitchell.com 上可以找到他的关于如大数据、SQL Server 并行数据仓库、Microsoft 商业智能等主题的博客。Brian 拥有佛罗里达大学的工商管理硕士学位。在研究 SQL Server 或 Hadoop 之余的闲暇时间里，他喜欢与妻子 Shannon 及孩子们一起游览并欣赏家乡佛罗里达州的美丽风光。

技术编辑简介

Rohit Bakhshi 是 Hortonworks 公司(Apache Hadoop 支持与服务的领先供应商)的产品经理。Hortonworks 创建和发布了 Hortonworks 数据平台(HDP)，一个由 Hadoop 驱动、可同时用于 Window 和 Linux 操作系统平台的 100%的开源数据管理软件。

Rohit负责Window产品线的HDP(Apache Hadoop的核心组件)及其服务平台。他曾与微软合作把Apache Hadoop组件的整个堆栈应用于Windows，使得Windows开发人员和系统管理员可以充分利用Apache Hadoop。

在加盟Hortonworks之前，Rohit是埃森哲技术研发实验室咨询小组的顾问，专注于架构搭建和为世界500强企业客户提供大数据解决方案。

John Hoang 是位于加利福尼亚维耶荷市的 Azure 客户咨询团队(AzureCAT)的高级项目经理。拥有超过 20 年的经验担当不同角色，如开发人员、业务分析师、项目经理等。在为制造业、零售业和医疗保健行业提供软件解决方案方面积累了丰富经验。目前专注于 SQL Server PDW 的研究。在空闲时间喜欢骑自行车、打网球以及与两个孩子玩耍。

Josh Luedeman拥有超过8年的SQL Server经验，目前是Data Structures公司的解决方案架构师，帮助客户更好地使用BI工具和大数据。拥有10年以上的IT经验，在一些世界500强企业、高等院校的重要机构、中小型企业以及创业型企业中担任过应用程序支持、数据库管理和BI等职位。Josh是包括Code On The Beach和SQL Saturdays等相关软件开发和数据方面讨论会的演讲者。他来自纽约的康宁，目前与妻子和儿女居住在佛罗里达州的奥兰多。可通过www.joshluedeman.com、josh@joshluedeman.com、www.linkedin.com/in/joshluedeman和推特@joshluedeman与他取得联系。

Michael Reed，一直致力于为棘手的业务难题设计创新解决方案。在过去 14 年里专注于数据库的开发与体系结构，最近更专注于商业智能和分析。目前作为资深 BI 顾问就职于 Pragmatic Work 公司。他之前担任 Insight and Analytics 的医疗索赔处理处总监职务。在此之前，在微软联机服务部门负责过运营、数据和信息交付中心任务；特别是在 AdCenter 行为目标团队，这是 Microsoft 主要的挖掘社交行为的研究机构，支持 Bing 决策搜索引擎和 BingAds 广告服务。以前，他曾白手起家与他人共同建立起一家价值百万美元的制造企业，他在那里获得许多商业知识和见解并应用到了当前工作中。

致　谢

感谢救世主耶稣基督的所有恩赐，感谢妻子和家人给予的支持和爱。本书作者团队令人难以置信地根据不停变化的平台和市场因素，在这个快速变化的领域中取得了一项伟大成就。在此，特别感谢来自 Hortonworks 和 Microsoft 的技术编辑们的支持。最后，最必不可少的自然是要感谢我们的读者以及大数据专家们，是你们的热情让这样一本著作物有所值。正是有了你们，我们才做到了！

——Adam Jorgensen

通过SQL社区，我结识了很多人。有一些只是泛泛之交，有一些会更深入了解一点，成了朋友。毫无疑问，Adam属于后者。我要感谢他的不仅是让我有机会和他一起合著本书，更重要的是他给予我的友谊，特别是在我们一起经历了PASS峰会的“过山车般的旅程”之后。我要感谢文字编辑Keith Cline，我的技术报告审核专员们：Josh Luedemen (InfoTech)、Michael Reed (PragmaticWorks)和John Hoang (Microsoft Azure CAT)。特别是John多年来的帮助和支持，对我而言是无价的财富。在此，向你表示衷心的谢意。感谢编辑Jennifer Lynn (www.pageoneediting.com)。我知道在这件事上让你辛苦了；值此机会，因为我那些似乎永无止境的各种借口，向你表示我衷心的谢意和歉意。

最后，我要感谢我的非官方评论家，为了给我提供反馈、见解和意见而牺牲了自己的时间。Lara Rubbelke，你天生就是个极具感召力的思想者。非常感谢你抽出宝贵的时间帮助我，使我的想法得以成型，见解得以升华。谢谢你，我忠实的朋友。

——James Rowland-Jones

前　言

如果你正在寻找并渴望知道大数据将对数据世界带来什么样的影响，那么本书就为你而著。与那些动辄几百页让人头疼的长篇累牍的叙述不同，我们通过一种不同方式来阐明你需要大数据，每个人都在做这件事情，而你一定要做得更“酷”一些。

作者团队希望创造出一些东西，它能成为你想脱离现有的关系型世界时的首选资源，不仅为你提供了前进的发展蓝图，还提供了实践经验而不需要你再去四处查找操作条款。大数据的新颖性和复杂性决定了在阐述细节时必须更详尽，而本书做到了！

我们的重点是确保你可以轻松过渡到使用这些工具和技术，因为你需要做的事情我们都曾经历过。也许你的老板参加完一个会议后走到你面前说：“我们需要大数据解决方案。”当你问他想要解决什么问题时，他回答不了，但他却非常清楚大数据解决方案对企业的重要性。此时，你就得承担起让这些大数据由梦想变为现实的责任。

通常，当有数据仓库或数据立方体(cube)需求时，需要通过培训课程和花费很长时间在网上搜寻相关信息，同时这些信息令你感到如此陌生。你将了解到大数据真的是很大——这绝不是双关语。它可以做大事，解决大问题，是一个庞大的含有工具和平台的生态系统。尽管如此，也像其他多数生态系统一样(RDBMS、编程语言、移动化和云)，最基本的也只有那么几样东西。一旦能掌握这些最基本的东西，当需要使用更高级的工具或自动化操作时，你会深深被这些基础知识所带来的结果而震惊。

我们的团队

我们组建了一支强大的国际作家团队以确保在正确的主题上传播卓越的观点和知识(稍后将讨论这些内容)。这些主题包括：

(1) 大数据、Hadoop、NoSQL 和关键行业知识的快捷概述

(2) 人们正试图解决的关键问题以及如何识别这些问题

(3) 在 Microsoft 环境中交付大数据

(4) 选择工具和平台

(5) 安装、配置和管理
(6) 存储和管理大数据
(7) 使用数据、添加数据结构和清理数据
(8) 大数据与 SQL Server 结合
(9) 大数据分析
(10) 云端的工作方式
(11) 案例学习以及现实世界的应用
(12) 在崭新的世界中让你的机构取得进步

这支团队的成员来自以下不同的机构：Pragmatic Works 公司，它是一家全球领先的信息服务、软件和培训机构；微软研究院；微软咨询服务部；Azure 客户咨询团队；以及其他一些在这个不断扩展的领域中产生着巨大影响的行业厂商。

不开玩笑

大数据如潮水般汹涌而至，在24个月之内你的环境就将拥有这些解决方案，而你应该提前准备好。本书旨在帮助你完成从关系型数据到更“进化的”数据世界视野的实用技巧的过渡。这包括处理那些并不非常适合表状结构的数据的解决方案，某些情况下，这些数据与你小心翼翼地维护了许多年的数据一样重要，或许更加重要。

同样，你将学到许多新的术语，作为一门技术课程，它简直就像一门词汇课程一样。

本书读者对象

本书面向数据开发人员、超级用户以及希望理解大数据技术将如何影响他们的世界以及如何在新的生态系统中恰当地采用解决方案的管理人员。读者需要对数据系统有基本的理解，并且拥有学习新技术和新技巧的热情。一些数据库或应用程序解决方案的开发经验将有助于理解一些高级领域的内容。

使用本书的先决条件

我们已将本书设计为广泛使用云资源，因此作为读者，需要有一台能可靠访问互联网的新型电脑，PC 或 Mac 都可以。此外，你将希望能够安装作者建议的额外的程序和工具，因此请确保你对正在使用的机器有恰当的权限。不同的章节将用到不同的工具和数据集，因此请按这些章节中的作者指示来得到最大化的操作体验。某些章节要求拥有对 SQL Server 数据库的访问权限，假如你希望建立内部环境，那么推荐使用 Hyper-V、VMWare 或 VirtualBox 之类的虚拟化技术。

章节内容概述

现在我们将浏览一下本书中的所有章节，并探讨你将在每一章中学到哪些内容。

第 1 章：行业需求与解决方案

没有涵盖生态系统的历史、起源和使用案例的大数据书籍是不完整的。本书同样需要探讨行业参与者和平台，其他著作会占用 5 或 6 章的篇幅来烦琐地讲述这些信息，但是我们更高效地完成了这部分内容，让你有更多时间接触那些更有趣的内容。

第 2 章：Microsoft 大数据解决方法

Microsoft 环境下的部署和传统的 UNIX 或 Linux 环境下的部署有些许区别。当我们感觉这种方法能让数百万 Window 管理员、开发人员和超级用户更容易理解时，我们就选择了这种方法。在著书之前就对许多人做过调查，最终发现压倒性地需要一个偏重于 Windows 的解决方案来帮助拥有最多人数的企业用户接触这门新技术。

第 3 章：配置首个大数据环境

在该章中将开始配置大数据环境。

第 4 章： HDFS、Hive、HBase 和 HCatalog

这些是一些关键的数据和元数据技术，我们将确保你理解使用每一个的正确时机以及如何发挥其最大性能。

第 5 章：HDFS 的数据存储与管理

分布式文件系统对于大部分读者而言可能是一个新概念，因此我们将完整地介绍 Hadoop 的这个核心组件并确保你准备好使用这个不可思议的功能来进行设计。

第 6 章：添加 Hive 结构

由于将经常使用 Hive，因此我们需要更深入地钻研它。在该章中让我们一起来一探究竟，确保你理解了有效地使用 Hive 所需的命令和逻辑。

第 7 章：使用 HBase 和 HCatalog 来扩展功能

处理大型表和元数据需要用到一些新的工具和技术。HBase 和 HCatalog 将有助于你控制这些类型的挑战，我们将让你明白如何使用它们。准备好迎接“大”数据吧！

第 8 章：使用 SSIS、Pig 和 Sqoop 进行有效的大数据 ETL

我们不得不加载数据，没有人能提出比我们的 ETL 专家作者更好的方法了。和他们一起使用熟悉和喜欢的工具以及一些新工具，快速有效地加载数据。

第 9 章：使用 Pig 和 Hive 进行数据研究和高级数据清理

现在我们已经安装、配置、管理并加载过一些数据，让我们使用新的工具和平台来研

究和清理数据。

第 10 章：数据仓库与 Hadoop 整合

SQL Server 和商业智能能在很大程度上适应大数据。大部分时间它们是一前一后地工作。我们将展示使用每种解决方案的时机以及它们在扩大和扩展的解决方案中是如何协同工作的。

第 11 章：使用 Windows BI 呈现大数据

现在我们已经有了分析结果，那么要如何将它们形象化地展示给我们的用户？我们有相关的新工具吗？我们会使用我们熟悉的工具吗？当然！让我们一起来做，这样我们可以明白如何将这些解决方案结合起来为我们的用户和客户实现最好的结果。

第 12 章：大数据分析

你已经听说过分析，这一章包括高级统计学分析、社会情绪分析、预测、建模以及其他很多内容！

第 13 章：大数据与云

你需要在数据中心拥有大量的服务器才能完成本书中的事项吗？当然不是！我们可使用灵活的、可伸缩的方式在云端完成这些事项。

第 14 章：现实生活中的大数据

其他公司在这个生态环境中是如何取得成功或遭遇失败的。我们将看看一些最有代表意义的成功和失败案例，并理解为什么会产生这种结果，那样我们就可以复制成功的经验或避免失败的结果。

第 15 章：创建和执行大数据计划

我们应该如何利用已经完成的东西并将其变为现实？该章将帮助你制订大数据规划。

第 16 章：运营的大数据管理

好像其他关键系统一样，管理这些技术并将它们整合进现有的基础架构中需要进行规划并仔细执行。让我们一起来完成这个规划吧！

本书特色

本书使用的以下特色段落和图标有助于将你的注意力转移到本书中一些最重要或最有用的信息上。

警告：

看到这个时一定要警觉，当因一些特定步骤没有被正确执行而造成损害时，就会看到这样一个旁白。

提示：

这些旁白包含一些快速提示，讲述如何简单地执行手头的任务。

注意：

这些旁白包含可能很重要的额外信息，包括能让特定项目开发更简单的视频链接和在线材料。

示例标题

这些旁白是关于当前话题或相关话题的一些更深入介绍。

目　录

第 III 部分　存储并管理大数据