“十二五”
国家重点图书出版规划项目

SAS WILEY

大数据分析
点“数”成金

BIG DATA ANALYTICS
TURNING BIG DATA INTO BIG MONEY

[美] Frank J. Ohlhorst 著
王伟军 刘凯 杨光 译
Jenny Sun 审校

人民邮电出版社
POSTS & TELECOM PRESS

“十二五”
国家重点图书出版规划项目

SAS WILEY

大数据分析
点“数”成金

BIG DATA ANALYTICS
TURNING BIG DATA INTO BIG MONEY

[美] Frank J. Ohlhorst 著
王伟军 刘凯 杨光 译
Jenny Sun 审校

人民邮电出版社
北京

图书在版编目（CIP）数据

大数据分析 : 点“数”成金 / (美) 奥尔霍斯特(Ohlhorst,F.J.) 著 ; 王伟军, 刘凯, 杨光译. -- 北京 : 人民邮电出版社, 2013.9
书名原文: Big data analytics: turning big data into big money
ISBN 978-7-115-32452-8

Ⅰ. ①大… Ⅱ. ①奥… ②王… ③刘… ④杨… Ⅲ. ①统计分析一研究 Ⅳ. ①O212.1

中国版本图书馆CIP数据核字(2013)第150261号

版 权 声 明

◆ 著　[美] Frank J.Ohlhorst
译　王伟军　刘　凯　杨　光
审　校　Jenny Sun
责任编辑　杨海玲
责任印制　程彦红　杨林杰
◆ 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
大厂聚鑫印刷有限责任公司印刷
◆ 开本：700×1000　1/16
印张：9.75
字数：130 千字　2013 年 9 月第 1 版
印数：1-4 000 册　2013 年 9 月河北第 1 次印刷
著作权合同登记号　图字：01-2013-3653 号

定价：35.00 元

读者服务热线：(010)67132692　印装质量热线：(010)67129223
反盗版热线：(010)67171154

内容提要

全书以大数据分析的商业及金融价值为主线，对大数据这一崭新领域进行了深入探索。本书以大数据技术的基本概念和演进历程开篇，随后详细介绍了不同类型的数据源及其对企业的重要意义、企业投资大数据技术的成功商业案例、有效利用数据集的必备技能，解析了打造大数据分析系统所需的存储、加工、软件平台及其他构成要素，海量数据的安全工具和技术，系统潜在风险及其规避方法，以及怎样对大数据进行分析并从中提取有价值的可用信息，并详实阐述了如何将大数据融入企业文化等问题。

本书不但向读者阐明了大数据分析的重要性，更淋漓尽致地展现了大数据分析的具体过程，从而帮助企业提升智能化水平、解决实际问题、提升利润空间、提高生产率并发现更多的商业机会。本书适合对数据处理、数据挖掘、数据分析感兴趣的企业决策者、技术人员等阅读。

推荐序

信息革命、信息化、信息社会、信息时代……这些当下耳熟能详的词汇，流行的关键在于“信息”二字。借助现代信息技术，规模化、定制化、高效地生产、采集、处理、存储、传播、利用信息，是信息社会区别于农业社会和工业社会的主要特征，也是信息社会变得更为聪明、更为智慧的要诀。

众所周知，信息革命本质上是一场信息数字化革命。缺少数字化信息，便无从发挥计算机出色的处理能力，也就无法展现现代信息通信技术的强大优势。信息的原始物理形态虽然多种多样，但在计算机内都表现为一系列 “0”和“1”组合所代表的数据。计算机中的信息处理，便表现为对数据的处理。因此，数据的采集、存储、处理、分析利用和传播的水平，也就代表了人类开发利用信息资源的能力和水平。

正是“数字化”开启了信息时代的一扇崭新的大门，为人类开发利用信息资源带来了新模式、新技术和新方法。1946 年世界上第一台电子计算机诞生至今已有 60 多年，这期间数据处理技术大约经历了三个主要发展阶段。

第一个阶段是 20 世纪 60 至 80 年代，这期间数据开发利用的主要方式是数据库（Data Base）。各行各业开发了大量功能各异的信息化应用系统，典型代表就是企业操作层面的数据处理系统和管理层面的管理信息系统。这些信息化应用系统的主要任务是采集各种业务流程中的数据和信息，采用自动或半自动的方式进行组织并按照特定格式存储到数据库中，以支持业务活动的运行和管理。

20 世纪 90 年代以来，迅速发展起来的商业智能（Business Intelligence）成为第二阶段的主要特征。商业智能形成了“关系数据库→数据仓库→数据挖掘→数据可视化”的一整套技术和产业链，极大地推动操作与管理层对数据资源的开发利用。商业智能的主要目标是对决策信息资源的充分利用，不仅使信息成为决策之基，而且利用数据挖掘及可视化工具，使决策活动更为“智能”。当然，无论数据库技术还是商业智能技术，处理对象都是结构化数据。

二十一世纪以来“大数据”开始受到人们广泛关注，这标志着数据处理技术

迈入第三阶段——一个更充分、更彻底、更高效开发利用数据资源的新阶段。人类社会数据量指数级的增长宣告着大数据时代的来临，这是全球信息化快速发展的必然结果。尽管大数据中蕴含着极为丰富的业务知识和有价值的决策信息，但如何利用这些数据来发现新识、创造新价值，已成为科学技术和经济社会发展面临的一个新挑战。

大数据以“超大量级”和“非结构化”为主要特征。在大数据面前，从前的数据库技术和商业智能技术都显得“心有余而力不足”。因此，大数据处理、分析和利用等问题便迅速提上了信息化发展的议事日程。这是数据资源开发利用向高端演进的一个标志，也是当代科学和技术发展的一个前沿，甚至有可能成为一个新兴学科——“数据科学”（Data Science）的核心内涵。

Frank Ohlhorst 所著的这本书，是目前众多介绍大数据概念、方法和应用中较有特点的一本著作。Ohlhorst 是一位知名技术专家，且长期担任企业 IT 咨询顾问，具有丰富的信息技术实践经验。全书以大数据分析的商业及金融价值为主线，对大数据这一崭新的领域进行了全面的探索，不但向读者阐明了大数据分析的重要性，更淋漓尽致地展现了大数据分析的具体过程。与其他同一主题的书籍相比，本书的引人之处在于作者分享了大数据分析在诸多行业的应用经验，以及如何在商务视角下对数据进行挖掘，怎样从数据中获取价值形成竞争优势。

大数据时代的来临，正在唤醒国人对于数据资源开发利用的重视。当务之急，不是一哄而上炒作大数据的概念，而是脚踏实地的研究，不断思考并总结如何提高我国数据资源的开发利用水平，加快缩小我国与发达国家的信息化差距，特别是在数据资源利用效率和效果方面的差距。王伟军教授等人在百忙之中翻译了 Ohlhorst 的这本书，切合我国当前大数据技术发展和应用需求，可谓相当适时。本书抛砖引玉，不仅能为有志利用大数据技术提升企业利润、提高服务水平进而获取持续竞争优势的各位企业家带来指导与启发，对我国大数据及数据科学教学与科研水平的提高，乃至大数据分析技术在我国更为广泛的推广，也将起到积极的促进作用。

是以为序。

国家信息化专家咨询委员会常务副主任

2013 年 6 月 28 日于北京

译者序

大数据时代就这么轰轰烈烈地到来了，不仅引得实业界、金融界和政界的青睐，更是强势地打入畅销书的阵营。大数据的书虽已不少，不过众口总是难调：有的读者关注大数据的具体技术、有的关注大数据的实践应用、有的关注大数据的管理方案、还有的关注大数据背景下特定行业面临的机遇与挑战等。然而，不论最后的落脚点究竟是什么，没有人能回避这些问题：大数据是什么？它能为我们带来哪些价值？本书就是介绍大数据基本知识及技术体系的入门书。

全书内容充实，涉及大数据的方方面面，而且行文流畅、语言优美。在翻译过程中，我们确立了非常严苛的翻译标准：不仅要做到忠于原文、保证句子和段落通顺，更要尽力展示原作的文采。值得赞许的是，以企业管理者为写作视角，这是本书区别于其他大数据书籍的根本区别：包含技术又不拘泥于技术，而是借助技术解决现实的管理和经营问题。说到底，它关注的不是技术实现的具体细节，而是大数据能够为企业带来哪些价值，以及怎样将这些价值转化成企业的利润。因此，本书特别适合各类企业管理人员阅读，其他想要了解大数据的读者也一定能得到许多收获和启示。

本书的翻译工作凝结着众人的智慧和心血。受人民邮电出版社杨海玲编辑之邀，我们欣然接受了本书的翻译工作。在华中师范大学王伟军教授的协调下，短时间内便组建起一支国际化的翻译的团队。全书由王伟军教授组织翻译并统稿，王伟军教授与刘凯博士合译了本书的第 1~9 章，来自澳大利亚昆士兰大学的杨光同学翻译了本书的第 10 章及附录，加拿大邮政公司高级分析和模型经理 Jenny Sun 对全书进行审校。

感谢杨海玲编辑慧眼如炬及充分的信任，感谢王伟军教授的统筹管理，感谢杨光为本书付出的无私奉献，也感谢 Jenny 数十年的业务沉淀及严谨细致的专业态度。我谨代表所有译者，为读者道出翻译完成后我们的一个共同心愿：愿此书能够成为您大数据探索之路上的益友。

此书翻译得到国家自然科学基金项目“基于用户偏好感知的 SaaS 服务选择优化研究”（项目编号：71271099），工信部信息化推进司软科学项目“大数据时代的信息化发展及相关公共政策研究”，以及湖北省自然科学基金项目“基于云计算的知识集成与服务研究”（项目编号： 2011CDA116）的支持，在此表示感谢。

译者在翻译的过程中，对原书存在的一些明显错误进行了修改。由于时间仓促，加之水平有限，书中的疏忽与错译之处，恳请读者朋友批评并指正。

刘 凯

2013 年 7 月

前　言

何为数据？这个问题看似简单，但在数据演绎的基础上，“数据”的定义可谓包罗万象。从“信息记录”到“世间万物”，不论是传感器记录的机器信息，还是人们拍下的照片，或是科学家记录的宇宙事件，我们所感知的一切都可以被归纳为“数据”。换言之，一切皆可谓数据。然而，人们捕捉、保护数据的技术能力依旧有限，数据的记录和保护一直都是难题。

人类的大脑的存储量大约为 2.5 PB（或者 100 万 GB[①]）。试想一下，假如你的大脑是一部电视里的数码摄像机，2.5 PB 足以装下 300 万小时的电视节目。电视机要持续不断地运行 300 年才能用光这些存储空间。目前可用的数据存储技术力不从心，由此产生了一项技术分支，即所谓的“大数据”。这项技术正在以指数级速度迅速发展。

现如今，企业记录的信息越来越多，随着信息（或数据）的剧增，消耗的存储空间越来越大，数据管理的难度也越来越高，大数据应运而生。至于为何要记录如此海量的信息，原因不尽相同。有些是为了服从管理规定，有些是为了保护交易，而更多的时候它仅仅是备份策略（Backup Strategy）的一部分，甚至它只是为了造福我们的子孙后代。

不论出于何种目的，存储数据的确耗费了大量的时间和金钱。其中最大的挑战莫过于企业如何才能继续负担海量数据存储的成本？值得庆幸的是，有些人已经想出了应对存储问题的技术方法，能够从“垃圾”中获取价值，变废为宝。这个过程就是所谓的“大数据分析”。

事实上，大数据分析的观念并不新鲜。企业使用 BI 工具已长达几十年之久，科学家们通过研究数据集来揭示普遍规律也已经很多年了。但是，收集的数据规模依然在变化，我们手里掌握的数据越丰富，能够从中推断的信息就越多。

---

① 1 PB=1 024 TB=1 024×1 024 GB=1 048 576 GB≈100 万 GB。

现在的挑战在于如何用更有意义、更可行的方式挖掘数据价值，探索数据源，推动智能化发展，进而帮助人们制定决策、找出关系、解决问题、增加收益、提高生产力，甚至是改善生活质量。

很重要的一点是要有“大想法”，也就是大数据分析。

此书将会探索大数据的真正概念、分析数据的方法以及诠释分析结论能够带来的收益。

第 1 章 探讨大数据分析的渊源、相关技术的发展，并阐释价值获取背后的概念。

第 2 章 探究不同类型的数据源，并解释为何这些数据源对于想要从数据集中寻求价值的企业来说意义重大。

第 3 章 帮助那些想要利用大数据分析的人构建商业案例，刺激大数据技术投资；为了能够从数据集中成功地提取智能、获取价值，需要哪些必备技能。

第 4 章 引入分析团队的概念，大数据团队需要哪些必备技能，以及如何将大数据与公司文化相融合。

第 5 章 协助大数据分析获取数据源，涵盖各种类型的公共数据和私有数据，以及识别可用于分析的不同类型的数据。

第 6 章 解决存储、处理能力以及平台问题，构成大数据分析系统需要哪些要素。

第 7 章 安全、合规以及审计的重要性——这些工具或技术能够在安全的前提下确保大数据源的分析可用性。

第 8 章 回顾大数据进化史，展望大数据在长期或短期内将会实现的改变，在大数据演变的过程中，这些变化将惠及更多的企业。

第 9 章 探讨大数据分析的最佳实践，涵盖一些增强大数据易用性的关键概念，对可能的陷阱提出警示，以及应如何避免这些陷阱。

第 10 章 讨论数据管道的概念，在分析过程中大数据如何运动并转化为可用的信息以辅助价值挖掘。

有时候，关于某项技术最全面、最优质的信息来自该技术的推广者，他们出于利润和增长的目的创作了白皮书。白皮书是为了推广某项技术，教育或知会潜在的消费者，并逐步鼓励消费者购买商家的产品。

也就是说，看待白皮书最好持保留态度。尽管如此，白皮书依然是研究技术、挖掘重大教育价值的优质“教材”。考虑到这一点，笔者将下列白皮书附在附录，作为补充材料供大家参考。如果你也期待找到大数据解决方案，敬请翻阅本书附录中的“Apache Hadoop 的 MapR 发行版”和“高可用性：无单点故障”（均出自 MapR 技术公司）。

致谢

在我看来，一本书的写作需要充分的时间、足够的耐心和持久的动力，三者缺一不可。尽管困难无处不在，有时甚至会让我迷失方向，可我依旧热衷于利用数据和模式分析来揭露数据背后隐藏的秘密。当我努力从复杂的数据分析中取得一点点突破，或者从完全不相关的数据集中分析出一点点结论时，所有的汗水都显得那么微不足道了。

这本书的创作灵感来自与 John Wiley 出版社编辑 Timothy Burgard 的一次简短交流。他联系到我并建议我在原有大数据文章的基础上写一本关于大数据的书。Timothy 认为高层管理者和那些真正从事数据分析的人非常缺乏全面解读大数据的资料，他相信我完全能够胜任这项具有挑战性的任务。正是在 Timothy 的鼓励下，我开始沿着这样的思路创作一本大数据的书。

此外，我还要郑重地感谢 John Wiley 出版社的开发编辑——Stacey Rivera。他给了我中肯的建议和莫大的鼓舞，是他顶住了压力给我不断前进的动力，使我有幸在遍布荆棘的写作之路上笔耕不辍。

本书的写作好比一次漫长的旅程。旅程中，我有幸结识了众多良师益友，是他们帮我形成自己的思路，启发我在海量数据之羹中摸索工艺之萃，并享受探求趋势和价值的乐趣。我还要感谢许多大数据领域的公司，他们无意中帮助我形成了书中数据价值的理念。数十家企业通过不懈努力宣传和推广大数据蕴含的价值，这些宣传推广进一步开拓了市场，促进社会各界开展对大数据的讨论，这些都激励着我最终完成本书的写作。

写书是一件既费时又费力的事情。身为自由作家，面对着诸多的写作任务，我常常窘迫于时间匆匆。所以我要感谢帮助过我的编辑们，没有他们的理解和通融，我永远无法妥善安排时间写就任何作品。我还要特别感谢 Mike Vizard、Ed Scannell、Mike Fratto、Mark Fontecchio、James Allen Miller 和 Cameron Sturdevant。

当被问到谁给予我的鼓励和支持最为重要时，我想，我的妻子 Carol 的付出无人能及。她理解写书会占用家庭时间，却仍对我全力相助，以成此书。我也要感谢我的孩子们——Connor、Tyler、Sarah 和 Katelyn，他们理解爸爸由于工作而不能常伴左右。拥有如此完美的家庭让我铭记不忘。

目录

第1章

什么是大数据

“大数据”到底是什么？这个概念乍看上去相当模糊，它似乎指的是数量庞大、信息量巨大的数据。尽管这样的描述确实符合“大数据”的字面含义，但它并没有解释清楚大数据到底是什么。

大数据常常被描述成已经大到无法用传统的数据处理工具进行管理和分析的极大的数据集。从网上我们可以得到一个被大多数人所认同的观点：当数据集已经发展到相当大的规模，常规的信息技术已无法有效地处理、适应数据集合的增长和演化时，大数据就应运而生了。换言之，数据集规模已大到难以用传统信息技术进行有效的管理，更不用说从中挖掘价值了。具体来说，难题主要在于数据的采集、存储、检索、共享、分析和数据可视化。

关于大数据到底是什么，绝非三言两语就能解释清楚。这个概念经过演变不仅包含了对数据集规模的描述，还包括数据利用的过程。大数据甚至变成了其他商务概念的代名词，如商业智能（Business Intelligence, BI）、数据分析（Analytics）和数据挖掘（Data Mining）。

“大数据”虽新，可大数据却早已存在。虽然海量的数据规模在最近两年内才形成，但大数据概念却早已在科学界、医学界等组织中萌芽。这些组织对海量数据集进行复杂的分析，并将其运用于药物研制、物理建模及其他研究领域。正是这些渊源为大数据今日的发展奠定了基础。