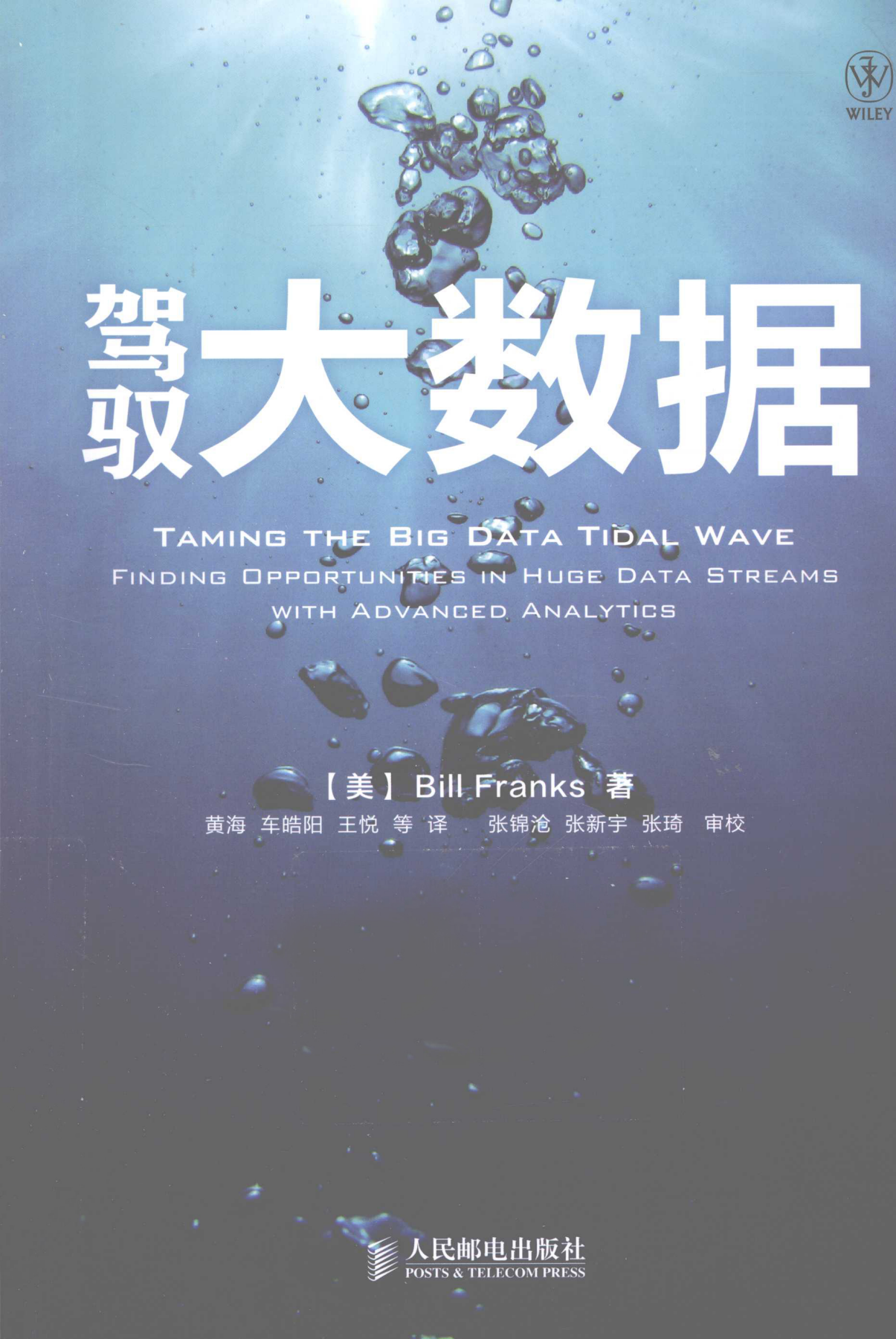

驾驭大数据

TAMING THE BIG DATA TIDAL WAVE

FINDING OPPORTUNITIES IN HUGE DATA STREAMS WITH ADVANCED ANALYTICS

【美】Bill Franks 著

黄海 车皓阳 王悦 等 译　张锦沧 张新宇 张琦 审校

人民邮电出版社
POSTS & TELECOM PRESS

WILEY

驾驭大数据

TAMING THE BIG DATA TIDAL WAVE

FINDING OPPORTUNITIES IN HUGE DATA STREAMS WITH ADVANCED ANALYTICS

【美】Bill Franks 著

黄海 车皓阳 王悦 等 译　　张锦沧 张新宇 张琦 审校

人民邮电出版社

北京

图书在版编目（CIP）数据

驾驭大数据 /（美）弗兰克斯（Franks, B.）著；黄海等译. -- 北京：人民邮电出版社，2013.1
ISBN 978-7-115-30480-3

Ⅰ. ①驾… Ⅱ. ①弗… ②黄… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2012)第298152号

版权声明

驾驭大数据

◆ 著　[美] Bill Franks
译　黄　海　车皓阳　王　悦　等
审　校　张锦沧　张新宇　张　琦
责任编辑　杨海玲
执行编辑　赵　越

◆ 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
北京天宇星印刷厂印刷

◆ 开本：700×1000　1/16
印张：16.75　2013 年 1 月第 1 版
字数：246 千字　2013 年 1 月北京第 1 次印刷
著作权合同登记号　图字：01-2012-7920 号

ISBN 978-7-115-30480-3

定价：49.00 元

读者服务热线：(010)67132692　印装质量热线：(010)67129223
反盗版热线：(010)67171154
广告经营许可证：京崇工商广字第 0021 号

内容提要

本书提供了处理大数据和在企业中培养创新和探索文化所需的工具、流程和方法，描绘了一个易于实施的行动计划，以帮助企业发现新的商业机会，实现新的业务流程，并做出更明智的决策。

本书重点介绍了如何驾驭大数据浪潮，并详细地介绍了什么是大数据，大数据为什么重要，以及如何应用大数据。本书还从具体实用的角度，介绍了用于分析和操作大数据的工具、技术和方法；以及从人才和企业文化的角度，介绍了如何使分析专家、分析团队以及所需的分析原则更加高效，如何通过分析创新中心使得分析更加具有创造力，以及如何改变分析文化。

本书适合对数据处理、数据挖掘、数据分析感兴趣的技术人员和决策者阅读。

对本书的赞誉

这本书关注了它应该关注的地方，主要关注大数据的有效分析，而不是大数据管理（BDM）。它从数据讲起，并进一步讲到如何制定决策，如何创建卓越的分析中心，以及如何建立分析文化。你将可以发现关于大数据管理的一些话题，但是，大量的内容是关于如何创建、组织、补充、执行那些使用数据作为输入的分析活动。

——Thomas H. Davenport，国际数据分析研究所联合创始人、研发总监

这是一本一站式手册，任何想要了解大数据是什么，以及如何通过高级分析流程和方法驾驭大数据的人都应该阅读它。Bill Franks 深刻理解了如何创建一个完整的、意在竞争中获得优势的分析生态系统，并在本书中对其进行了详细描述。

——Stuart Aitken，美国 dunnhumby 公司首席执行官

在《驾驭大数据》中，Bill Franks 很好地介绍了可以产生新商业价值的大数据和分析类型，而这些价值将从正在被大数据浪潮冲击的企业所掌握的新型大数据源中获得。这本书很容易阅读，在每章的末尾都有“本章小结”来帮助你进行总结。这本书还避免使用过于专业的技术术语，但本书绝不是一本轻量级书籍。在这本很棒的大数据入门书中，Bill 为分析创新和从现在开始做大数据分析提供了强大的案例。

——James Taylor，Decision Management Solutions 公司首席执行官

Dicision Management Systems: A Practical Guide to Using Business Rules and Predictive Analytics 作者

如果你想要了解为什么在许多行业中，大数据都可以产生商业价值，那么，这本书将为你提供多个视角和多种答案——从高科技，到数据科学，到业务用户和流程等。在我整个分析的研究和教学生涯中，我从没遇到过这样一本书，能将信息技术与公司业务以如此简洁的方式结合到一起。我推荐任何与大数据有交集的人都阅读本书。

——Diego Klabjan，美国西北大学教授

Master of Science in Analytics Program 负责人

Bill Franks 以一种寓教于乐的方式来讨论这个复杂的主题。他为从业人员和新手们提供了他对大数据最真实的理解和远见卓识，这使得本书成为了一本重要的读物，任何分析领域的新手和从业人员都可以通过本书向分析行业的领导者学习。Franks 跨多个行业的见解，以及他对大数据的驾驭，都证明了他是带领你进入大数据分析领域最好的领路人。

——Richard Maltsbarger，美国劳氏公司高级战略副总裁

驾驭未来的价值发现之旅

这不仅是数据爆炸的时代，更是一个大数据爆发的时代。面对大数据的激流，多元化数据的大量涌现，大数据已经为个人生活、企业经营，甚至国家和社会都带来了机遇和影响。

大数据的技术和市场正在快速发展，而驾驭大数据的呼声则一浪高过一浪。随着大数据所蕴含价值的激情释放，使得大数据已经成为 IT 信息产业中最具潜力的蓝海。但是，面对各种不同的大数据工具和解决方案，到底哪些才是技术核心，并能够带来真正的价值？

本书作者 Bill Franks 先生是 Teradata 天睿公司首席分析官，他将自己和 Teradata 在数据分析领域的知识和经验进行了总结，并带领我们迈上了大数据价值的发现之旅。我很荣幸率先阅读了本书的中文版，并郑重推荐给大家，同大家一起分享数据价值极致演绎的心得体会。

大数据的核心

麻省理工学院管理评论在“通往价值的新道路”研究报告中，总结了“顶尖绩效的公司使用正确分析挖掘方法和工具的使用率，与绩效较低的公司相比，高出了 5 倍。”美国全国保险公司客户管理副总裁 Kathy. Koontz 女士指出：“重要的不是数据，而是如何使用数据。企业必须改变它们的经营方式，学会从数据中洞察事实并做出反应，否则数据整理得再有条理，也没什么价值。”政府或企事业单位对于数据的驾驭，从最基本的获取，到整合、治理、分析、探索、汲取智能、采取精确的行动，这种全程能力的建立已经比以往任何时候更为重要。

所以，数据的核心是发现价值，而驾驭数据的核心是分析。我想强调一点，过去所谓“得数据者得天下”的说法，只是说明了“获取”数据的重要；然而，立身于大数据时代的我们，应该更加专注于数据的核心价值，如何转化和激发

它的潜能，赋予它新的生命，创造出更多的业务提升机会，这才是真正的重点所在。

IDC 调研显示，中国的大数据市场未来 5 年将以 51.4%的速度增长。正如书中所言“今天的大数据并非明天的大数据”，帮助政府和企业掌握驾驭大数据的能力就是帮助它们赢得未来。Teradata 天睿公司在帮助政府和企业进行大数据分析的过程中，倡导使用已经过无数次验证的 IDA 方法论，即通过对信息的整合（Integration）、探索（Discovery），并使其转化成行动（Action），最终帮助用户建立制胜未来的核心竞争力。

大数据的挑战和趋势

随着大数据浪潮的加速到来，未来 5 年将成为大数据的全面发展期，将出现产业链的整体繁荣。如何在大数据浪潮的洗礼中确保技术架构、人才、政府和企业战略以及商业模式能够“逐浪潮头”，将更需要积极主动地选择适合的技术、方法论、解决方案和发展策略等。

环顾整个市场，我们在某些领域取得了突破性发展，但仍然面临着大量挑战。例如，研发分析各种多元结构化数据的高效技术，提高大数据分析的易用性，让大数据分析技术实现“开箱即用”，使得数据分析成为政府和企业建立核心竞争力的关键途径。技术创新永无止境，面对快速增长的大数据，我们还需要处理“更大的数据”，激活“各种渠道、各种结构、过去、现在甚至未来的数据”的更大价值。

驾驭大数据就是驾驭未来

本书作者 Bill Franks 先生奉献出自己的智慧、见解和实践经验，帮助武装我们的思想和技能。

无论你是首席技术官、首席信息官和首席营销官，还是想成为更加优秀的业务分析师，本书将告诉你如何整合数据、探索数据，并转化为行动，并最终带来业务价值。书中不仅介绍了分析流程的演进、方法论、分析团队的组建，还有对建立分析文化的深入探究。我相信本书将成为大家应对大数据来袭的最佳工具

书，成为你驾驭未来的技术指南，帮助你成为赢得蓝海的真正王者。

最后，我要感谢本书的原著作者 Bill Franks，感谢几位先期读者在百忙中为本书写下真知灼见的书评，感谢为中文版出版做出贡献的人民邮电出版社的领导、编审和各位译者，感谢 Teradata 天睿公司的技术和市场团队付出的日日夜夜，请相信你们的努力将会在我们的数据价值发现之旅中绽放精彩。

辛儿伦

Teradata 天睿公司大中华区首席执行官

2012 年 12 月 12 日

序言

无论你是否喜欢，大量的数据都会在不久的将来涌入你的生活。也许它现在已经出现在你的生活中了，也许你已经与它们打了一段时间交道——例如，试图解决这些数据的存储问题以便后续的访问，处理错误和缺陷，或者将这些数据进行结构化分类。或许你现在准备通过分析庞大的数据集提炼出一些有价值的数据，进而从中得到一些关于你的客户、业务或者你的企业所处商业环境的信息。或许你还没有到这一步，但是你已经意识到了数据管理的重要性。

无论你属于上述哪种情况，你都找对了地方。正如 Bill Franks 所说，在不久的将来，不仅会有大数据，还会有许多关于大数据的书籍。但是，我觉得这本书不同于其他的大数据书籍。首先，该书是这个领域的先驱者。最重要的是，它与其他书籍侧重的内容有所不同。

很多大数据的书籍侧重于大数据管理：如何将大数据存储到数据库或者数据仓库中，或者如何将非结构化数据进行结构化和分类。如果你发现自己阅读到了很多关于 Hadoop、MapReduce 或者其他关于数据仓库方法的内容，那么你可能已经遇到了，或正在寻找一本“大数据管理（BDM）”的书籍。

当然，大数据管理是一项重要的工作。无论你有多少何种质量的数据，如果你不能将它们按照某种便于访问和分析的格式存储到一个环境中，那么你就无法体现出这些数据的价值。

但仅仅是大数据管理方面的知识还不能让你走得更远。为了让这些任意大小的数据变得有价值，你不得不自己分析和操作这些大数据。正如传统的数据库管理工具不能自动地分析来自传统系统的交易数据一样，Hadoop 和 MapReduce 也不能自动解释来自网站、基因图谱、图像分析或者其他大数据源的数据的含义。即使在大数据时代到来之前，许多从事数据管理多年（甚至是几十年）的组织也没能从它们的数据中获取到便于分析和决策的有价值信息。

在我看来，这本书将重点放对了地方。它主要是关于大数据的有效分析，而

不是大数据管理本身。它从数据开始，所有的内容均围绕如何做整体决策，如何构建卓越的数据分析中心，以及如何构建数据分析文化等主题。你也会发现一些大数据管理中提到的内容，但该书内容的主体仍是关于如何利用输入数据生成、组织、配置和执行数据分析。

或许你还没有意识到，分析在今天的商业领域中是一个很热门的话题。这本书将主要围绕公司如何利用分析进行竞争，我在该领域的著作和论文一直是我所有著作中最热门的内容。关于分析的会议也在各地不断涌现。大的咨询公司，例如，Accenture、Deloitte 和 IBM 已经在该领域积累了大量经验。许多公司、公共服务部门甚至非营利机构都已经将分析作为一个优先的战略。现在人们对大数据非常感兴趣，但是重点仍应该放在如何组织这些数据并使得它们便于分析，进而影响决策和行动。

Bill Franks 独创地将讨论重点放在大数据和分析的交集上。与其他数据仓库和数据应用供应商相比，他所在的公司 Teradata，在数据分析及从中提取商业价值的领域，一直都表现出了最高的专注程度。尽管 Teradata 最被人们熟知的是其企业数据仓库工具，但是这些年来，它也提供了一系列的分析应用工具。

在过去的一些年中，Teradata 为了开发面向大数据的高度可扩展的分析工具，已经和领先的数据分析软件供应商 SAS 建立了紧密的联系。这些工具通常是数据仓库环境的嵌入式分析工具，并针对大量数据分析应用，例如，实时欺诈检测和大规模客户购买倾向评分。Bill Franks 是 Teradata 的首席分析专家，因此有机会了解大规模分析和库内处理的理念和专业知识。如果讨论这个主题，可能没有比 Bill Franks 更好的人选了。

那么，本书还提供了哪些特别有趣且重要的内容呢？以下是关于本书重点的简要介绍。

- 第 1 章概述了大数据的相关概念，还解释了“数据的大小并不总是最重要的”这个观点。事实上，在整本书中，Franks 指出了许多大数据其实并没有用，如何过滤掉无效的数据才是真正重要的。
- 第 3 章是对大数据源的综述，将大数据源进行了创造性和有价值的分类，且非常全面。该书第 2 章介绍了网络数据及其分析，对希望了解

在线用户行为的企业和个人会很有帮助。这部分内容绝不仅仅是一般的面向网页分析的报表。

- 第 4 章致力于介绍分析可扩展性的演进，这部分内容为您提供了一个大数据和分析技术平台的全新视角。可以肯定的是，你在其他地方都未曾看到过这部分的内容。该章也讲述了最新的技术，例如，MapReduce，并讨论了大部分大数据分析工作都需要一个混合的环境。
- 该书包含了一部分关于如何生成和管理分析数据环境的最新内容，这也是在其他地方看不到的内容。如果你想要了解最新的关于“分析沙箱”和“企业分析数据集”内容（这对我来讲也是全新的内容，但是现在我知道了它们是什么以及它们的重要性），那么你可以在第 5 章中找到答案。本章还包含了一些关于对管理系统和处理流程进行建模和评分的重要信息。
- 第 6 章讨论了目前常用分析软件工具的类型，包含开源包 R。虽然很难找到关于这些不同分析环境优缺点的评价，但是本章中你将读到这些分析。最后，本章讨论了一些组合和简易分析的方法，以便于像我这样的非技术人员理解。
- 该书的第三部分从技术角度给出了在分析中和企业管理方面的建议。同时，选取的角度也是很合理的。例如，我特别喜欢第 7 章中关于制定决策和发现问题的部分。许多分析专家进行分析时都没有考虑一个更大的问题——这些问题是如何产生的。
- 近来有人问我，关于分析文化内容的描述是否超出了本书的范畴。我回答说，在我读 Franks 所写的第四部分之前，我并不知道这个问题的答案。他将分析文化和创新文化联系在了一起，这一点我非常喜欢，并且以前从未见到过此类内容。

尽管这本书并没有避开技术话题，但它以一种直接和解释性的方式对它们进行了描述。这使得本书适合更广泛的读者，包括那些技术背景有限的读者。Franks 使用数据可视化工具的论述借以概括整本书的基调和视角：“简单即是最好的。仅当必要时，再把它变得复杂。”

如果您的企业打算进行分析工作——毫无疑问你将需要解决很多在这本书

中所涉及的问题。即使你不是一个技术人员，你也需要熟悉一些关于构建企业分析能力所涉及的内容。如果你是一个技术人员，你将学习到分析中人性化的一面。如果你正在书店或者通过“搜索本书内容”浏览本书的前言部分，那么买下这本书吧。如果你已经买了这本书，那就赶快行动起来，阅读它吧！

Thomas H. Davenport
信息、技术与管理领域杰出教授，美国巴布森学院
联合创始人、研发总监，国际数据分析研究所

前言

你收到一封邮件，邮件中提供了一套个人电脑的报价。而你几个小时前刚刚在这家零售商的网站上搜索过电脑的信息，似乎它们已经读出了你的想法……当你驱车前往这家商店购买这套个人电脑时，你路过了一家咖啡店，你看到了这家咖啡店的一条折扣信息。你获知由于你刚来到这片区域，你可以在未来 20 分钟内享受 10%的折扣……

在你享用咖啡的时候，你收到了一家制造商关于某产品的道歉，而你昨天刚刚在你的 Facebook 主页和这家公司的网站上抱怨了它们的产品……

最后，当你回到家之后，你又收到了一条关于购买你最喜欢的在线视频游戏升级装备的信息。有了这些装备，你才能顺利通过某些曾经苦苦挣扎的关卡……

听起来很疯狂吗？难道这些事情只有在很远的未来才发生吗？不，这些场景都是我们今天可能见到的！大数据、高级分析、大数据分析，似乎今天你已经逃脱不了这些术语了。无论在哪里，你都会听到人们在讨论大数据和高级分析，看到关于它们的文章或是宣传推销它们。好了，现在你也可以将这本书加入关于它们的讨论中了。

什么是真实的，什么是炒作？这些关注可能会使你怀疑大数据分析是一种炒作，而非真实的东西。尽管在过去的几年曾经有不少被炒作的概念，然而就分析能力和处理海量数据而言，我们确实处在一个转型的年代。如果你肯花一些时间来理清并过滤掉那些有时被媒体过分炒作的部分，你会发现大数据背后有一些非常真实和强大的东西。随着时间的推移，大数据分析会使企业和消费者都获益，而收益带来的兴奋和期待又会继续引发更多的炒作。

大数据是下一波新数据源的浪潮，并会驱动分析在商业、政府及教育界的下一次革新。这些革新将有可能快速改变企业审视它们自身业务的方式。大数据分析可以促成更加明智的决策，在某些情况下，促成这些决策的方式将明显不同于今天。它带来的很多洞察在今天看起来都像是在做梦。你会看到，征服大数据的

需求和一直以来征服新数据源的需求在很大程度上是一致的。然而，大数据的额外规模必须使用新的工具、技术、方法和流程。传统的分析方法已经不再适用于新的环境，我们有必要使用高级分析将商业界带入更高的层次。这就是这本书要讲的内容。

“驾驭大数据”并不只是本书的书名，而是下一个十年中，决定哪些商业活动将振兴，而哪些商业活动将消亡的决定性因素。准备主动接受大数据，企业可以通过驾驭大数据浪潮而取得成功，而不是遭受大数据浪潮连绵不断的冲击。你需要了解些什么？你如何为征服大数据做准备？你如何从大数据中获得振奋人心的分析结果？坐下来，找一个舒服的姿势，准备好发现大数据的秘密！

读者对象

这些年来有无数关于高级分析的书籍问世，最近也开始有关于大数据的书籍出现。本书是从一个与其他书籍不同的角度来看大数据的，主要帮助读者理解什么是大数据，如何通过分析来利用大数据，以及在如今的大数据环境中，如何处理世界范围内的高级分析生态系统的创新和变革。大部分读者都将发现这本书有价值且充满趣味。无论你是分析专家，还是使用分析结果的企业家，或者只是对大数据和高级分析感兴趣的人，这本书都有适合你阅读的内容。

本书并不会深入介绍所涉及主题的技术细节。本书的技术高度刚刚能够让读者从高层次来理解其所讨论的概念。本书的目的是使读者可以理解，并开始运用这些概念，以及帮助他们认识在哪些方面还需要更加深入的研究。这本书更像是一本手册而非教科书，完全可以被非技术人员理解和掌握。同时，那些对这些主题已经有深入了解的读者，也可以从本书的一些讨论中获得一些技术方面更深层次的启示。

内容提要

本书由四部分组成，每一部分都从一个方面来介绍如何驾驭大数据浪潮。第一部分将介绍什么是大数据，大数据为什么重要，以及如何应用大数据。第二部分集中介绍那些能够用于分析和操作大数据的工具、技术和方法。第三部分介绍

如何使分析专家、分析团队以及所需的分析原则更加高效。第四部分将前三部分结合在一起，重点介绍了如何通过分析创新中心使得分析更加有创造力，以及如何改变分析文化。以下是关于各章节所涉及内容的详细提纲。

第一部分　大数据的兴起

第一部分重点介绍了什么是大数据，大数据为什么重要，以及分析大数据可以带来什么好处。本部分覆盖了 10 种类型的大数据源，以及如何利用这些资源来帮助企业提高其业务水平。如果读者拿起这本书时，还不知道什么是大数据，以及大数据的应用有多么广泛，那么第一部分会帮助你了解这部分内容。

第 1 章　什么是大数据，大数据为什么重要

本章首先介绍了大数据的背景知识，以及大数据到底是关于什么的。然后给出了一些企业如何利用大数据的案例。如果读者想要帮助自己的企业驾驭大数据浪潮，那么请首先理解本章所讲的内容。

第 2 章　网络数据：原始的大数据

如今，或许应用最为广泛并为人们所熟知的大数据源是从网站上收集来的详细数据。用户浏览互联网所产生的日志信息，是等待分析和挖掘的信息宝库。不同行业的企业都将从它们网站上收集到的详细用户信息整合到它们的企业业务分析中。本章将探索这些数据是如何增强和改变一系列业务决策的。

第 3 章　典型大数据源及其价值

在本章中，我们将从高层次来探索 9 种大数据源。其目的是介绍每种数据源，并讨论每种数据源在商业中的应用和启示。一些本质相同的技术应用在不同的行业中，以产生多种大数据源，这个趋势已经越来越明显。另外，不同的行业可以利用一些相同的大数据源，大数据并非只能用于某些狭窄的领域。

第二部分　驾驭大数据：技术、流程以及方法

第二部分将集中介绍用于驾驭大数据的技术、流程以及方法。这些年取得的

重大进展增加了这 3 个方面的可扩展性。企业不能继续依赖外部的方法和专家来保持它们在大数据世界中的竞争力。本书的这一部分将是技术性最强的一部分，但仍然可以被绝大多数的读者所理解和接受。读完这些章节后，读者将熟悉他们今后进入大数据分析领域时可能遇到的一系列概念。

第 4 章　分析可扩展性的演进

在每一个时期，数据的高速增长使得当时最具可扩展性的工具也只能疲于应付。在大数据出现之前，传统的高级分析方法已经到达了它们的瓶颈。如今，传统的方法已经不再适用。本章将讨论分析和数据环境的融合、海量并行处理（MPP）体系、云、网格计算，以及 MapReduce 技术。这些技术增强了可扩展性，并且在大数据分析中扮演着重要角色。

第 5 章　分析流程的演进

为了更好地利用被极大增强的可扩展性，分析流程也需要进行升级。本章将首先概述如何利用分析沙箱为分析专家提供一个可扩展的环境，从而建立高级分析流程。然后，我们将介绍企业分析数据库如何帮助在创建分析数据时，获得更高的一致性并减小风险，同时提高分析专家的生产效率。本章最后将探讨如何使用嵌入式评分过程将高级分析流程部署和转移到用户端和应用端。

第 6 章　分析工具和方法的演进

本章将介绍一些高级分析方法演进的过程，以及这些改进将如何继续改变分析专家完成工作和处理大数据的方式。讨论的主题将包括可视化图形界面、单点分析解决方案、开源工具，以及数据可视化工具的演进。本章也讲述了分析专家将如何改变他们建模的方法，以便更好地利用可用资源。讨论的主题包括组合模型、简易模型以及文本分析。

第三部分　驾驭大数据：人和方法

第三部分重点讨论驾驭大数据的人和他们所属的团队，以及确保他们能够提供优质分析的方法。如何提供优质的分析，包括大数据分析，其关键因素是找到合适的人来掌舵，并且他们能够遵循正确的分析原则。读完这 3 章后，读者将了