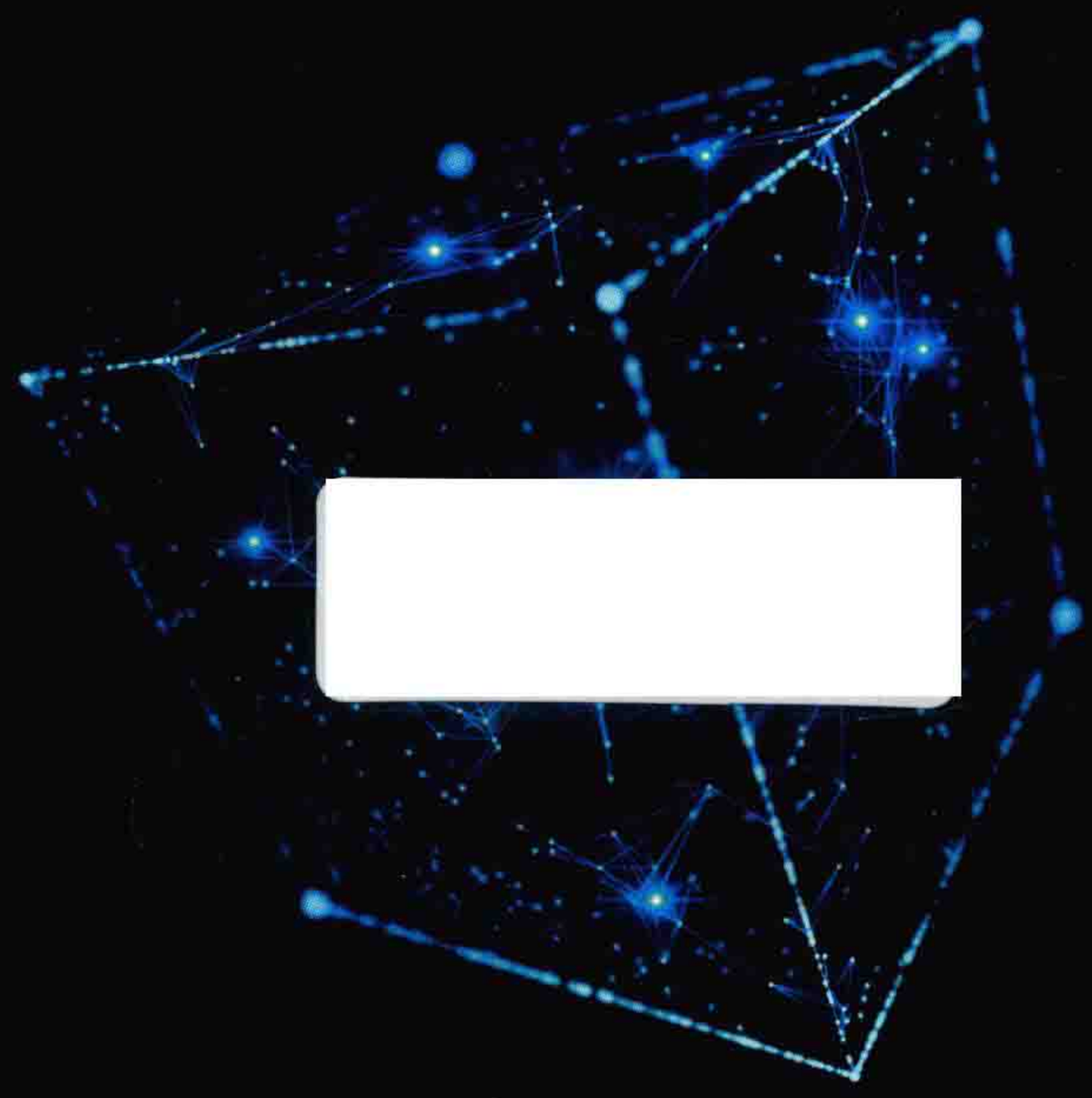
大数据：规划、实施、运维

Big Data
Planning, Deploying, and Operations

◎ 谢朝阳 编著



中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

工业和信息产业科技与教育专著出版资金资助出版

大数据：规划、实施、运维

谢朝阳　编著

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

内 容 简 介

你是不是有这样的困惑：“读了不少关于大数据的书，发现这大数据既可以用于竞选美国总统，又能够预测禽流感，还能卖啤酒和尿不湿，又是围棋高手……大数据好像什么都能干耶！可是咋整呀？大数据多大为大呀？大数据能赚钱不？……唉，怎么还是一头雾水。”本书将为你答疑解惑。

本书将展现作者在国内外大数据第一线的实战经验，面向不同行业的共性诉求来指导读者大数据该怎么做，并阐明大数据发展的误区。本书对大数据，从经济价值、商业模式、框架搭建、数据挖掘、网络布置、安全防护、人员能力和后续运维管理多个维度，以及基础设施、中间件、重点应用等多个层面进行系统阐述，帮助决策者将大数据概念落地，建立起理性的预期、合理的规划，并最终收获满意的经济效益。

企业正面临从传统 IT 转入大数据环境这一不可避免的范式变化，恰好为我国追赶发达国家信息化建设带来了契机。本书以企业共同关注的客户关系管理（CRM）为实例谈大数据落地，利用大数据采集、分析、决策以达到客户关系拓展、精准营销和创新产品的目的，提出一整套从规划到实施再到后续运维的技术路线和策略。并用一个已上线的实例将各部分内容串起来综合展示，以解决大数据热潮中的“老虎吃天，无处下爪”的窘境。这对于大数据的正确理解，企业信息系统的建立，以及相应的商业模式改变都具有实际指导意义。

本书读者对象为大数据产业政策制定者、从业者和分析师，包括：政府及企业 IT 负责人，企业 CIO、架构师、网络与系统管理人员、应用开发人员，高校和研究院所教师、研究人员，高校学生等。

图书在版编目（CIP）数据

大数据：规划、实施、运维 / 谢朝阳编著. —北京：电子工业出版社，2018.5

ISBN 978-7-121-33952-3

Ⅰ. ①大… Ⅱ. ①谢… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2018）第 061843 号

策划编辑：冉 哲

责任编辑：冉 哲

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：787×1 092 1/16 印张：32 字数：810 千字

版 次：2018 年 5 月第 1 版

印 次：2018 年 5 月第 1 次印刷

定 价：98.00 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）88254888，88258888。

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。

本书咨询联系方式：ran@phei.com.cn。

序言1

当前，我国正在实施国家大数据战略。大数据是信息化发展的新阶段，它对经济发展、社会治理、国家管理、人民生活都产生了重大影响。要推动大数据技术产业创新发展，我们要瞄准世界科技前沿，集中优势资源突破大数据核心技术，加快构建自主可控的大数据产业链、价值链和生态系统。大数据的迅速发展，也造成了大数据人才的异常紧缺，为了使我国的技术从业者赶上这波变革的浪潮，有必要迅速地培养起从业人员的大数据实践能力。

我曾为作者的《云计算：规划、实施、运维》作序，当时就觉得要是有一本既有系统工程的方法论同时又具有实用性的关于大数据的书该有多好。现在见到了作者的《大数据：规划、实施、运维》，异常欣喜。

本书对于帮助读者从正确认识大数据的概念，到指导读者真正将大数据实施落地，再到引导读者把握大数据的发展趋势以及大数据与人工智能等的衔接，均有实际的价值。本书贯彻了作者写作的“实”的特色以及“单刀直入”的风格，融入了作者从事大数据一线工作的工程实践经验，对大数据领域内共同关心的问题从规划、实施到运维进行了详尽的、切合实际需求的介绍。本书举例丰富、深入浅出，既有一定的理论高度，又直接贴近一线实战，显得难能可贵。

作者从数据技术的内涵和自己的实践经验出发，提出了大数据的狭义、广义、泛义、伪义的概念，针对当前的一些大数据误区，理清思路，帮助读者建立正确的大数据观念。从大数据产业链入手，正确认识该产业的业态，明确自身的定位和价值点，解惑从业者们共同关心的问题，从而建立起理性的预期与合理的规划。并通过深入剖析大数据落地的规划、实施、运维这三部曲，针对可能遇到的困惑和问题，给出特别需要注意的事项及指导原则，帮助读者在较短的时间内推出成本可控且能满足需求的大数据产品与服务，最终产生经济效益。

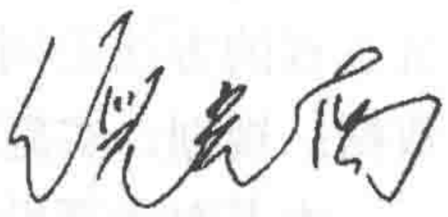
作者沿着几条主线对大数据进行了探讨。首先正如书名所说，大数据的交付就像创作一部大型交响乐，需要遵循规划、实施、运维三部曲。规划篇帮助读者了解自己的数据，以及数据和数据之间的关系，体量的增长趋势。弄清楚了这些，大数据才能规划好，派上用场，并且当变化来临时具有可扩展性。大数据的实施具有相当的复杂性，涉及的技术组件很多，而这些组件本身发展也比较快。实施需要分析在大数据实施过程中所应遵循的一般方法和特别之处，以及大数据关键技术点。最后，作者反复强调了，好的大数据系统所具有的运维特性，即 RASSM（Reliability，Availability，Security，Scalability，Manageability），以及“三分建设，七分运维”的重要性。

作者对大数据是否可以被视为一门科学也进行了讨论，并且严密地论证了大数据的科学性。同时，大数据的处理方法和常人做事一样，遵循的是一个“Work Hard, Work Smart,

Getting Help”的过程。大数据本身也是一个求知的过程，从已知走向未知，经历着“数字—数据—信息—知识—应用—数据”的轮回。期望读者也能抱以科学的态度来研究和学习大数据，同时做好准备，面对未知，既能充分运用大数据的强大能力，又能从容处置预测不准是常态、预测准是概率事件的情况。

作者在美国从事 IT 行业期间经历和参与了数次 IT 行业重要的发展和变迁，其中包括开放系统、因特网和云计算。经中组部“千人计划”引进回国后，作者的工作聚焦在大数据和云计算领域，组建国内首个运营商专业云计算公司，并通过各种途径，包括撰写本书在内，毫无保留地将其在大数据方面的造诣和经验介绍给广大读者，为推进我国大数据的发展做出了许多贡献。

相信本书的内容一定可以在各行各业的大数据项目实践中带给读者很大的裨益。在此，我谨将本书推荐给涉及大数据工作的政府、企业 IT 负责人员，相关 IT 产业的从业者，高校和研究院所的教师、研究员及学生。



中国工程院院士

序言2

大数据正逐渐从新兴的概念向成熟的技术和应用转变。大数据的热度也推动了不少相关著作的发布，多数主要集中于介绍大数据概念和讲述大数据相关故事，这些对于初识大数据的读者来说是必要的。但是大数据经过这几年的发展，仅仅讲故事已经远远不够了，更需要的是进入实践应用的阶段。

本书内容就如书名一样，全面地介绍了大数据技术的方方面面，着重描绘了大数据的规划、实施和运维全生命周期，密切地结合了理论和实践，是一部既适合想要进一步了解大数据的普通读者又适合想要更进一步了解业态的大数据从业人员阅读的著作。像作者的《云计算：规划、实施、运维》一书一样，作者通俗易懂的写作风格，无论是对于专业人士还是普通读者，都是极具价值的。

大数据技术应当为满足企业业务的发展需求而服务，因此，企业在开展大数据建设前，应结合自身的信息化现状，分析现有业务系统和 IT 服务的类型与特征，正确判断在当前情况下自己的应用是否需要大数据技术，进而确定企业中的哪些业务系统或 IT 服务适合实施大数据，以及如何统筹资源来科学合理地开展大数据建设，并确定相关的计划和评价标准。作者就企业做好大数据转型展开了讨论，从评估大数据技术能够给自身带来多大的价值和战略意义，合理平衡“功能—性能—成本”，做好风险管控，来进行合理的规划，使读者首先要了解自己的需求，了解自己在行业中所处的位置，根据自身的实际情况，来应对大数据，而不是为做大数据而做大数据。

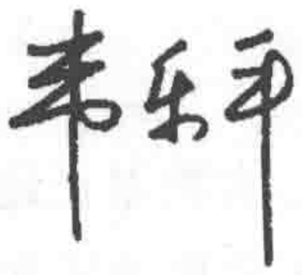
大数据本身就是人类求知的过程，从数据，到信息，到知识，再到知识的运用，而后，优劣有别的运用结果又反馈到数据中，周而复始。大数据的目的在于了解、管理、共享、使用数据，从而服务于工作与生活。进而，由已知预测未知。预测未知，测不准是常态，预测准是概率事件。作者对类似的一些容易混淆的概念进行了澄清。同时，围绕大数据的处理能力、运算能力、应用场景进行了深入浅出的阐述。

作者是由中国电信最先引进的中组部“千人计划”专家，其在国外具有丰富的 IT 大型工程实践经验，曾在世界上最大、要求最严格的数据中心第一线工作，对数据技术从整体架构到细枝末节都有着难得的实践经验。回国后，他在中国电信云计算业务的实践中充分发挥了“亲历者”的特点，为中国电信的云计算发展奠定了很好的基础。大数据建筑在作者倡导的广义云基础之上，本书的写作是作者长期以来在数据技术领域的探索和实践的系统总结。

本书涉猎面广，从多个层次及维度，针对业界对于大数据领域共同关注的问题，以规划、实施和运维三个重要阶段为切入点，对在这三个阶段内可能遇到的问题、困惑和

难点进行了详细介绍、深入分析并给出了指导性的意见和具体建议。

我相信本书的出版一定能对我国大数据的推广、普及和发展应用起到积极的作用，谨将此书推荐给读者，是为序。



中国电信科学技术委员会主任

前言

你是不是有这样的困惑：“读了不少关于大数据的书，发现这大数据既可以用于竞选美国总统，又能够预测禽流感，还能卖啤酒和尿不湿，又是围棋高手……大数据好像什么都能干耶！可是咋整呀？大数据多大为大呀？大数据能赚钱不？……唉，怎么还是一头雾水。”

当你拿到这本书就对了。大数据，大数据，多大算大呢？当所要处理的数据量超过了现有的计算环境的数据处理能力时，就是大数据了。它可以是 ZB、EB、PB、TB 级的，也可以是 GB 级的。当然，如果你的资金足够充裕，可以买得起 TB 级的内存、上百个处理器插槽以及海量的存储设备，那对别人来说是大数据，对你而言可能就只是小数据了。

大数据本身就是人类求知的过程，从数字，到数据，到信息，到知识，再到知识的运用，而后，优劣有别的运用结果又反馈到数据中，周而复始。其实，大数据所面临的场景只有两种：已知和未知。在已知的场景下需要累积大量的样本，或者，在有公认规则的前提下——如棋艺类，按照规则自己生成样本，AlphaZero 就属于这一类。而未知的场景就只能是做预测了。预测究竟能有多准？或许“Most likely”是最保险的答案。大数据既没有预测到美国总统特朗普的当选，也没有准确预测到埃博拉，沃尔玛也从未把啤酒和尿不湿放在一起。预测不准是常态，预测准是概率事件。

国内的 IT 热潮一波接着一波，俨然就像一场场运动。先是云计算，接着又是大数据。各路玩家都想追一下这些时髦热词的风潮，生怕赶不上，纷纷试着寻找将大数据整合到自身 IT 系统中的可能性。而原本的 IT 公司和从业者更是绞尽脑汁地想要在大数据业务中开拓新的市场。媒体对大数据产业未来几年的发展更是持有过热的描述，甚至对 2020 年的大数据产业规模给出了 5 万亿元的惊人估值，充满了 Big Data = IT 的味道。

在此背景下，一大批冠以大数据标题的书籍上架。就当前每年出版的大数据书籍的性质与数量来看，有很多属于通俗类、科普类，以及吸引眼球的读物范畴。有些大数据著作中充满着“正确的废话”，而在真正意义上具有实践价值的内容少而又少。然而，其中并不乏受到热捧的作品。

这也在一定程度上反映出读者的求知心理：希望只需遵循一定的阅读捷径，就能消化掌握相关的技术，成为高手。然而，在阅读完众多所谓的技术类书籍后，读者却并不能收获到预期的效果。要么只模模糊糊地“见森林见不到树木”，要么又好像“摸到了树木见不到森林”，越来越迷茫。

究其原因，这类书籍并未本着科学的理念来传播可用于实践的知识与技术，更多的是为了迎合热点话题，以一种美化的甚至扭曲的形式来对新技术做介绍，缺乏严谨性和实用性，缺乏将技术以“科学知识”的高度进行传授的态度，更少了如何将技术落地到实处的关键内容，甚至很多书是作者为了提升职称和赚取稿酬等目的而拼凑的。当然，

写书也是一门营生，追逐热潮没有错，可是过热的“泡沫来，泡沫往”却并不可取。对新技术的学习应该落到实处，切不可以讹传讹，Be careful with what you read，就是这个意思。

事实上，大数据的应用实情或许并不像许多例子中所描述的那样可以用来当兴奋剂。现阶段对大数据，从概念到应用，连认识都不清晰，更谈不上数据挖掘的深度。此时如果不对大数据有一个严谨客观的传授，可能会使读者在理解上产生谬误、从路线上走偏，甚至当前已经出现了不少对大数据认识的误区。可以发现，众多谈大数据的书籍中反复引用着几个所谓“经典”的例子，其实只不过是作者们的想象，经不起推敲。甚至一些例子所谈论的情况与大数据这个词汇一点关系都没有，譬如廉价机票、啤酒和尿不湿等。

今天再谈大数据，应该先摒弃盲目乐观与炒作的成分。如果还是停留在反复谈论具有吸引眼球效果的数字和示例（如谷歌预测流感、奥巴马竞选总统等）上，谈论便失去了意义。

大数据或大数据技术就是工具。要让工具用得好，首先得用对地方，其次要会正确地使用。

基于以上认识，身为一线的数据从业者，作者深感为大数据从业者提供系统的、正确的知识与观念正当其时。本书即是在此背景下编写的，旨在根据作者个人多年的从业经验和心得，从科学知识的高度出发，一步步帮助读者将大数据变成看得见摸得着的东西，使之有效实施，真正落地成为有用的工具。

除技术层面的内容外，本书立足于大数据的实践和商业价值，从规划、实施到运维来进行阐述。本书在构想与撰写时，遵循了以下原则。

在对象方面，本书兼顾专业化与大众化，且遵循着可以将本书作为研究生课程教材的撰写原则；在知识的深度和广度上，一方面与高校专业教育水准相符合，另一方面也进阶到大数据专业从业者水准。此外，大数据作为当前的 IT 技术热点，也是大众非常想了解的领域。为适应大众读者的需要，也为了使大数据技术可以获得更广泛的推广，本书力求使普通读者也能够理解吸收。因此在取材与撰写时，除在文字上深入浅出外，在用例方面也尽量运用合适的例子把事情说清说透。事实上，本书的大部分内容曾用在作者为华中师范大学和上海交通大学硕士、博士研究生开设的大数据科学应用课程中，收到了良好的反馈。

在内容方面，本书采用将学术性与实用性相结合且更突出实用性的原则。大数据技术可以算作一种理论性的学科技术，需要重视对其所包含理论的探讨。在大数据范畴内，涉及包括统计学、人工智能等在内的各类专业知识，就连大数据这个词本身也是一个含义纷呈、范围甚广、概念抽象的名词。而在大数据技术的另一个层面上，它又是与实践紧密联系的，多数读者希望通过学习大数据书籍来解决最实际的大数据软硬件平台及应用的建设问题，而且大数据这一概念本身也是从实际的数据行业需求中产生出来的。因此，本书在内容上，力求结合理论与实际，既探讨必要的理论知识，给予读者正确的概念，又重视实践的各个环节。

在架构方面，本书采用专门性与普遍性均衡原则。就知识范围而言，大数据技术是多种技术的组合，从单一的需求出发点可以分化到涉及大数据规划、实施、运维全生命

周期的各个不同的细分技术环节。本书内容注重大数据技术中的普通知识与深入的专业技术之间的均衡，以指引有志从事大数据行业的读者，在普通知识之外，找到自己感兴趣的方向。为达到这一目标，本书的编排涉及大数据的各个环节，并对每个环节的各细分方向都做了由浅入深的专题介绍。

所谓 God creates the numbers, men do the rest。自从有人类文明以来就有了数字，进而有了数据，甚至可以说就有了大数据。为什么今天把大数据提到如此的高度呢？这和数据的产生量及相应的处理能力（软的、硬的）是分不开的。中国的智能手机用户数量居全球第一，企业的数量也居全球第一，随着 IT 业的推进和渗透，每时每刻都有海量的数据产生和被保存，这也正是大数据在中国发展的基础。利用好大数据技术，了解数据、管理数据、共享数据、使用数据，可方便人们的日常生活，有助于企业打破信息孤岛，有效地融合各方面的信息，从而为合作伙伴的选择、供应链的管理、目标市场的锁定等提供定量的决策依据。

除论述大数据是什么、能做什么外，更侧重的是怎么做。本书以“客户关系管理（Customer Relationship Management，CRM）”这一企业级应用场景为例，这也是目前大数据应用为数不多的成功案例，深入、细致、完整地展示大数据的各个环节。紧扣如何利用大数据来实现以客户行为来指导销售推送以及生产决策的过程，也就是“推荐系统”，力求使读者能真正将大数据落地于实践。

本书立足于作者所处企业的案例和产品，结合流行的开源软件（Hadoop、Spark 等），实打实地谈大数据，并给出了一手的市场情况以及真实的数据。全书从规划到实施再到运维，系统、全面地帮助读者把握大数据落地的各个环节，了解大数据的全貌。大数据的实践是与业务密切关联的，本书以一个实际的大数据项目为专题，将书中讲述的规划、实施、运维穿针引线，Put it all together，向读者完整展示大数据实践过程，拉近读者与大数据的距离，让大数据理念切实与读者的工作相结合。

在市场环境下，任何技术都要围绕商战的“三匹老马”（价格、质量、服务）及经济社会的三个主要环节（生产、流通、消费）来发展。对于各个企业的大数据活动而言，其目的是寻找一条利用大数据来提高自身业务运作效率、维系现有客户、扩大新客户群的路线，从而达到以大数据促进产业链并实现精准客户管理的效果，做到向数据要效益。直白地说，就是怎样通过多渠道、多维度获取有用的用户消费行为数据，对其进行建模分析，从而做出决策来服务现有的用户，通过给用户推荐其感兴趣的相关产品以达到精准营销，挖掘已有客户的价值。而大数据的高级阶段则是——设计出新的产品。

本书在撰写中秉持以下观点。

1）大数据的定义应该是多层次的。狭义的大数据停留在技术处理的层面；而广义的大数据则包含了大数据产业链的各个环节所提供的产品和服务；泛义的大数据扩展到每个细分的行业大数据中，成为“数据+”；伪义大数据则以营销为目的，虽不可避免地包含了一部分炒作的成分，但也确确实实起到了一定的推广效用，是一股不可低估的市场力量。

2）做好大数据和做成任何一件事情一样，只有三种方法：Work Hard，Work Smart，Getting Help。Work Hard 体现在对处理单元性能的提升上，Work Smart 则是对算法的改

进，Getting Help 是指借助多个处理单元以集群的思维来解决对超大规模数据集的处理。

3）大数据的处理过程可形成一个持续提升的迭代闭环。由原始的数据开始，大数据先将其处理为信息，进而利用算法抽取出其中所蕴含的知识，知识的正确运用可以帮助决策，最终知识的集成和梳理就可以晋升为智慧和文化。而在开展决策实践的过程中，还会产生新的数据，即，数字—数据—信息—知识—应用—数据。因此，上述过程又会进入新的一轮，并不断提升，也就是所谓的波浪式前进、螺旋式上升。

4）大数据并非一次技术的跳跃式飞升。多数 IT 技术领域在相当长的一段时间内并未出现划时代的本质变化，其技术增强点大都集中在计算能力（算力）上，而这种计算能力或者说数据处理能力的增强则又集中体现到了大数据上。因此，如何将大数据的这种数据处理能力结合到具体的业务中，探寻合适的商业模式，是我们讨论大数据时特别值得关注的问题。对于提供的产品和服务，谁买的单、客户/用户是谁、现金流从哪里来到哪里去都不清楚，空谈大数据产业是没有意义的。

5）要认清什么是伪义大数据。透过大数据的炒作层面，理解其对具有海量高速、多样可变特征的多维数据集进行深度挖掘的本质。并且，该本质尚处在发展的早期，对于其中涉及的认知计算、深度学习、人工智能、统计相关性等背后的因果机理，甚至大数据预测中的“测不准”现象都还需要长期研究。因此，当前不应对大数据盲目地崇拜和信任，而要提醒读者保持清醒的头脑。要认识到，大数据只有服务于具体行业，进行融合应用并作为行业驱动，才能获得真正的产值，才是回归到谈论大数据的正途。大数据这一跨界学科，也是多个学科的基础，譬如认知计算、人工智能等，如果不涉及这些方面，大数据的阐述层次是不够的。本书将适当涉及这些内容。

6）要真正体现大数据的 4 个数据特征：Volume（体量大）、Variety（模态多）、Velocity（变化快）、Value（价值高），并且确保大数据的应用不会造成安全隐患，就要时刻理清和把控数据的来源和去向。从统计学的角度看，大数据意味着样本集变得更大了。大数据下的数据来源不再是传统的企业内部单一来源，而应当整合包括商业对手在内的各种数据来源渠道。还可以基于搜索引擎来获取与题目相关的数据，或者来自线下。如果离开了这些数据源的相对的全覆盖、多格式和多维度，大数据很可能就成了数据前面加“个大”而已。

7）当前，IT 对于企业及行业的服务广度和深度正发生着变化，工业 4.0、智能制造、现代服务业等无不体现着 IT 正进入新的时代。如果将传统的 IT 视为 IT 的 1.0 版，那么云计算所引领的对 IT 资源的复用，使得用户的 IT 基础设施的成本大幅降低，这可以算作 IT 的 2.0 版。在基础设施不再成为障碍的前提下，更进一步地，大数据及数据挖掘等技术的发展用以解决数据和业务之间的结合问题，人工智能的研究用以实现机器的自学习问题等，可以说已经将 IT 带入了现代服务业的 3.0 版。当然，这种划分并非绝对。

总的来说，本书的宗旨是帮助大数据从业者从大数据产业链入手，正确地认识该产业的业态，明确自身的定位和价值点，解惑从业者们共同关心的问题，使其建立起理性的预期与合理的规划。并通过深入剖析大数据落地的规划、实施、运维这三部曲，针对可能遇到的困惑和问题，给出特别需要注意的事项及指导原则，帮助读者在较短的时间内推出成本可控且能满足需求的大数据产品与服务，最终产生经济效益。

本书共分为6篇。

第1篇（第1～2章）为大数据导论。简要介绍大数据的基本概念、建设目标和意义，以及与大数据产业链相关的生态圈。另外，随着大数据概念持续被炒热，在对其的认识和理解上存在着各种偏颇，本篇也会对大数据认识上的误区进行讨论。

第2篇（第3～4章）为规划篇。大数据技术应当为满足企业业务的发展需求而服务，因此，企业在开展大数据建设前，应结合自身的信息化现状，分析现有业务系统和IT服务的类型与特征，正确判断在当前情况下自己的应用是否需要大数据技术，进而确定企业中的哪些业务系统或IT服务适合实施大数据，以及如何统筹资源来科学合理地开展大数据建设，并确定相关的计划和评价标准。企业做大数据转型，需要评估大数据技术能够给自身带来多大的价值和战略意义，合理平衡Scope-Schedule-Cost铁三角，做好风险管控，采取各种方式规避和解决可能遇到的问题。通过规划篇，读者首先要了解自己，了解自己所处的位置，根据自身的实际情况，来应对大数据，而不是为做大数据而做大数据。

第3篇（第5～8章）为实施篇。将大数据规划落地，需要选择具体的技术路径，此时主要受Function-Performance-Cost铁三角的制约。大数据的实施具有相当的复杂性，本篇将分析在大数据实施过程中所应遵循的一般方法和特别之处，就大数据实施中的关键技术点依次展开。由于大数据具有多技术交织的特征，因此本篇更偏重于介绍与大数据直接相关的技术，其中包括：以MapReduce为代表的大数据并行计算框架，以大数据生态圈中最具活力的Hadoop为代表的分布式处理系统，大数据存储系统，以及相关的机器学习与人工智能技术等。

第4篇（第9～13章）为运维篇。大数据项目实施完成后运维就开始了。运维是一个持续的过程，包括升级、优化、扩容等。企业需要维护业务运营的持续性，需要采取必要的技术手段和人力资源来保证运维。大数据的运维主要包括4个方面。第一，网络畅通。大数据的海量数据处理与分布式业务流量模型对现有数据中心网络架构提出新的挑战，保障大数据业务网络畅通与稳定需要从SDN等新兴网络技术中寻求解决方案。第二，数据安全。大数据核心价值在于数据分析与利用，在数据采集、存储、挖掘和发布等阶段都需要采取相应的安全技术以保证数据的安全性，大数据平台的安全机制也至关重要。第三，大数据集群一定是会出故障的，数据备份与恢复这个“古老”的看家本领是必不可少的。大数据分布式存储特性使得大数据备份和恢复具有自身特性。第四，高效运维管理。本篇从大数据集群配置管理、集群监控、日志分析等方面展示如何进行大数据环境的运维管理，并从运维服务、运维流程模型、运维人员、自动化与智能运维（AIOps）多方面讨论如何有效进行大数据运维日常工作。

第5篇（第14～16章）为实例篇。本篇以一个类似于Netflix（奈飞）的公司为例，展示其如何建立和运营大数据业务，并通过挖掘客户的行为数据来实现推送营销。这一篇将前面几个篇章的内容综合起来用于实践，从发展思路、产品与服务到赢利模式展现给读者。本篇将围绕着迷你的Netflix——Oracle MoviePlex案例来阐述传统关系数据库怎样与大数据技术紧密结合，数据怎样通过关系数据库或其他方式加载、提取、转换至大数据环境，在大数据环境中，业务部门怎样在数据池中分析和挖掘大数据的价值。本篇

也可用于单独阅读。

第 6 篇（第 17～18 章）介绍明天的大数据。预见大数据的明天会怎样是一件“危险”的事情。在本篇中，作者对大数据未来发展的基本共识做了一些梳理总结和展望。就当前的实情和趋势，分析大数据所面对的挑战，探讨该领域的发展和技术演进方向。随着时间的推移，就像云计算一样，人们已经把它视为常态，“云”字就会消失，大数据中的“大”字也会消失，而成为新常态。

本书末尾是三个附录。

附录 A 详细介绍了如何安装一套可运行的 Hadoop 平台软件，以 Cloudera 发布的开源版本 CDH 作为例子，帮助读者顺利跨出大数据实践的第一步。

附录 B 利用 MATLAB 对美国 21 年航空公司到达和起飞时间的记录数据，展示了大数据的数据处理过程，以使读者更直观地理解 MapReduce 的过程。

附录 C 则从 DeepMind 的 AlphaGo Zero 论文和最新的 AlphaZero 入手，解读人工智能由最初的大量收集棋谱比对，到按照人工输入的简单规则“自己和自己对弈”生产棋谱的过程，并和读者分享一些想法。不要夸张（Make no mistake），一定是规则在先才能自造样本。人工智能，一定是人工在先，才能智能。

作者长期在美国从事 IT 前沿工作，在美期间亲身参与了数次 IT 行业重要的发展和变迁，其中包括开放系统（Open System）、互联网、云计算等。并且，作者在 IT 行业中所任职的美国海关总署、北美索尼、美国 Intel 等政府和世界 500 强企业对 IT 系统的要求是非常苛刻和现代的。作者作为 2011 年中组部“千人计划”特聘专家回国，同年成功组建了首个运营商云计算专业公司，算是在运营商中实现了 IT 1.0 到 IT 2.0 的范式变化，此后，继续投身到了大数据领域。因此，作者非常希望在书中对国内外与大数据相关的数据科学与信息技术，以及工程实践进行全面的论述和比较，为国内政策制定者、企业的 CIO 和 IT 工作者、创业者、投资人在大数据业务开展方面提供务实的、系统化的考量角度和评估方法。并希望能通过本书为政府和企业合理又经济地发展大数据提供有价值的建议，避免低水平或过度的建设。

最后，本书保留了作者在《云计算：规划、实施、运维》（电子工业出版社，2015）中为读者所喜爱的“单刀直入、直奔主题”的风格。所以这本《大数据：规划、实施、运维》可以视为该书的姊妹篇。

从初步构想到最后出版，对本书品质的方方面面，一切都希望能做到尽职尽责。唯成书仓促，难免有诸多缺失甚至偏颇，祈业内先进赐教，以匡正之。

作者要感谢的人很多。

作者首先感谢倪光南院士在百忙之中拨冗为本书作序。感谢中国电信科学技术委员会主任韦乐平先生，我的师长。没有韦先生的指导和帮助，作者回国后会需要更长的时间来适应国内的工作环境。邬贺铨院士、李德毅院士在为《云计算：规划、实施、运维》一书写了推荐评语后，又在百忙之中再次欣然为本书写了推荐评语，对作者的努力给予了很大的鼓励与肯定，在此作者表示诚挚的谢意。还要感谢国家数字化学习工程技术研究中心主任杨宗凯（西安电子科技大学校长）和同仁们对本书所提出的宝贵建议。同时在本书的写作过程中，电子工业出版社的冉哲编辑提供了全程帮助，特别是面对我的英

式中文，耐心、细致、不厌其烦地为我修改，没有冉编辑的帮助，本书难以与读者见面。作者特别感谢以下几位：陈劭力按照作者的思路和录音整理出了最初文稿；陈琪和郑芳交叉校阅了实施篇和运维篇；张彬帮助准备了 Oracle MoviePlex 案例；李昊溟帮助准备了附录；徐小飞和夏晴进行了最后的文字整理。当然，书中的任何瑕疵完全是作者的责任。最后，也是最重要的，作者感谢家人们的支持与付出。

谢朝阳

2018 年

目录