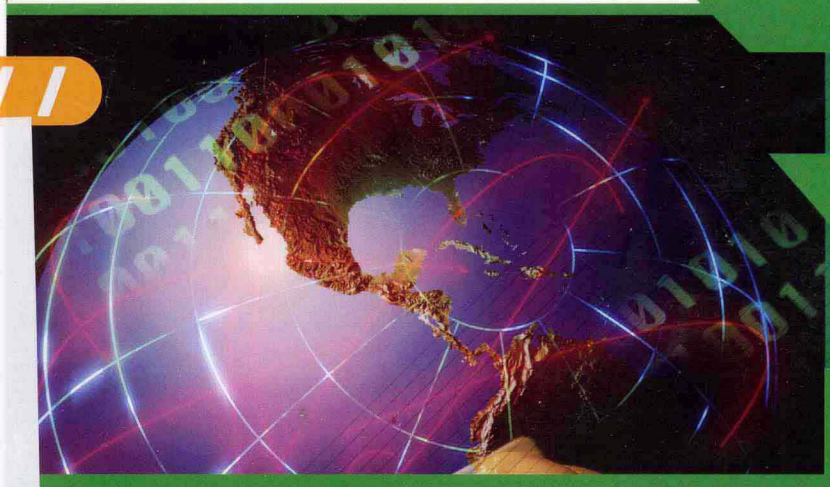
解惑大数据

丁圣勇 樊勇兵 闵世武 编著

BIG DATA



人民邮电出版社
POSTS & TELECOM PRESS

解惑大数据

丁圣勇 樊勇兵 闵世武 编著

BIG DATA

人民邮电出版社
北京

图书在版编目（CIP）数据

解惑大数据 / 丁圣勇，樊勇兵，闵世武编著. -- 北京 : 人民邮电出版社，2013.9
ISBN 978-7-115-32534-1

Ⅰ. ①解… Ⅱ. ①丁… ②樊… ③闵… Ⅲ. ①数据处理－问题解答 Ⅳ. ①TP274-44

中国版本图书馆CIP数据核字(2013)第165372号

内 容 提 要

本书是一本系统介绍大数据的技术普及读物，可帮助读者迅速了解大数据的概况。

全书分为 4 章，共 120 个问题，内容涵盖大数据概念、大数据技术、大数据应用等各个方面。第 1 章为大数据概述篇，介绍了大数据的概念、技术特点及应用方向等；第 2 章为分布式平台篇，介绍了大数据的核心计算平台；第 3 章为分布式数据库篇，介绍大数据中广泛使用的分布式数据库；第 4 章为大数据与数据挖掘篇，介绍经典的数据挖掘算法以及大数据如何实现规模化和并行化处理。

本书可作为对大数据技术感兴趣的读者以及工程技术人员、行业管理人员、大数据系统的设计开发人员的技术参考资料，也可以作为大学本科高年级学生和研究生相关课程的参考书。

◆ 编　　著　丁圣勇　樊勇兵　闵世武
责任编辑　李　静
责任印制　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：8.5　　2013 年 9 月第 1 版
字数：152 千字　　2013 年 9 月河北第 1 次印刷

定价：36.00 元

读者服务热线：(010)67119329　印装质量热线：(010)67129223
反盗版热线：(010)67171154

序

随着全球“大数据”技术和应用的持续升温，有媒体已将2013年称为“大数据”元年。发达国家政府和先进企业纷纷将“大数据”上升为国家战略和企业战略，希望通过挖掘数据资产的价值有效增强自身的竞争优势。“大数据”不但促进了社会科学和自然科学的交融发展，而且也推动着当今社会的思维变革、商业变革和管理变革。

我国已是世界上最大的PC和智能手机市场，拥有全球规模最大的互联网和移动互联网用户群，移动化、消费化、两化融合和智慧城市建设等信息化浪潮正为社会带来了内容丰富多样、颗粒度日趋精细和高增长率的海量数据，海量数据正通过泛在的网络汇集到“云”中，社会需要新的处理模式和技术将这些海量数据变为重要的信息资产，以提升相关的决策能力、洞察能力、流程优化能力和服务能力。

主流的国内外电信运营商已将“大数据”视为战略性技术和潜在发展机遇。一方面，希望通过挖掘对内日益重视自身庞大的宽带信息基础设施服务客户的过程中积累的相当可观的各类宝贵数据，开拓合规的增值应用空间；通过更好地掌握客户行为和需求，助力精准的业务决策，以提升客户的体验和忠诚度。另一方面，电信运营商对外也渴望成为“大数据”产业链上的一个重要环节，通过不断优化和打造自身广泛覆盖的先进智能宽带信息基础设施，为客户提供有竞争力、可扩展、安全合规的数据采集、传输、存储、整合、迁移、灾备、处理和管理等多元服务能力，结合开放自身和伙伴的部分数据，帮助客户更容易地获得所需的数据视图，发展更

多的创新型应用，创造新的细分市场，提升客户的综合竞争能力。

见出以知入，观往以知来，“大数据”正给我们带来无穷的应用畅想。但“大数据”毕竟涉及方方面面的知识，需要我们广博的学习、谨慎的思考、清晰的分辨和有效的应用。为了帮助广大读者了解“大数据”的相关概念、技术和应用方式，本书年轻的作者们根据自身学习工作中的心得编著了这本《解惑大数据》。作为《解惑云计算》的姊妹篇，本书沿袭了前书的编写风格，读者同样可以把本书作为了解“大数据”相关知识的“小百科全书”来参考。

2013 年 6 月 19 日

前言

业界普遍认为，我们已经身处大数据时代，早在2012年，《纽约时报》就刊文称，“大数据时代”已经来临，而国内也有专家认为2013年就是中国“大数据元年”。大数据将带来前所未有的机遇和挑战，各国政府和企业都在加强大数据领域的投入以抢占大数据先机。

大数据之所以得到广泛关注，很重要的原因是大数据将数据本身视为一种资源，将对数据的挖掘能力视为一种核心竞争力，相关的技术能够有效处理日益增长的海量数据，挖掘其中的价值。目前，大数据技术的优势、渗透的领域、发挥的价值已不断得到实践检验，大数据的价值和重要性已得到广泛认同。

然而对于很多人而言，大数据却是一个非常模糊的概念，一方面这是由于大数据涉及很多专业技术元素，包括并行计算技术、数据挖掘技术以及机器学习技术等，普通读者缺少相应的技术基础；另一方面也是因为大数据概念本身较新颖，技术还在不断发展，相应的书籍和参考资料还比较少。

考虑到大数据体系复杂以及很多读者缺少专业的技术基础，本书以设问形式由浅入深地解读大数据的内容，向读者循序渐进地介绍大数据的概念、应用及相关技术。尤其是对电信行业的从业者，本书还结合电信行业的特点探讨大数据在电信行业中的应用。

本书由丁圣勇、樊勇兵进行统筹，主要编著人员有：丁圣勇、樊勇兵、闵世武。

大数据是一个比较新的领域，并且由于时间仓促，作者经验有限，书中难免有疏漏和不当之处，敬请读者批评指正。

致谢

在这里，首先要感谢中国电信广州研究院院长蔡康先生，中国电信集团公司技术部冯明处长，中国电信广州研究院副院长黄勇军先生，中国电信广州研究院数据通信研究部部长唐宏先生、副部长金华敏先生，他们是本书的发起人。

在本书的出版过程中，我们得到了人民邮电出版社的大力支持和朋友周佳新先生、彭芳女士的大量帮助，在此向他（她）们表示衷心感谢！

也要感谢业界的专家、学者及技术人员。在与他们的交流、探讨过程中我们的团队丰富了云计算的知识，加深了对大数据的理解。

还要感谢我们的家人对我们工作的理解与支持！

最后，感谢亲爱的读者您在茫茫书海中选择了本书，希望您能够从中受益。

本书写作组

2013 年 5 月 6 日

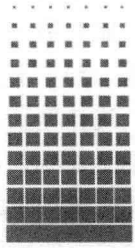
目录

目录

目录

目录

目录



第1章 Chapter 1

大数据概述

Q1. 什么是大数据?

按照维基百科的定义，大数据（Big Data）又称为巨量资料或海量资料，指的是所涉及的资料量规模巨大到无法通过目前主流软件工具，在合理时间内达到撷取、管理、处理并整理成为对企业经营决策具有较高参考价值的资讯。

业界普遍认为，大数据具有 4 个关键特征，分别是海量化（Volume）、多样化（Variety）、快速化（Velocity）和价值化（Value）。

Q2. 大数据的规模如何?

大数据首先是数据量大。在当今信息化时代，全球数据量正以前所未有的速度增长，遍布世界各个角落的传感器、移动设备、在线交易和社交网络每天都要生成上百万兆字节的数据，据 IDC 的“数字宇宙（Digital Universe）”项目统计预测，全球可统计的数据存储量在 2011 年约为 1.8ZB，2015 年将超过 8ZB（1ZB=10^{21} 字节，或等于 1 000EB，1 000 000PB，或者等于大家更加熟悉的 10 亿 TB 的数据）。数据容量增长的速度大大超过了硬件、软件技术的发展速度，以致于引发了数据存储和处理的危机，数据的重要性日益成为人们关注的话题。

从微观而言，数据规模达到亿条数据以上，存储空间超过 Tera Byte 的都可以称为大数据问题。

Q3. 什么是大数据的多样化?

大数据的数据类型具有多样性。海量数据的危机并不单纯是数据量的爆炸性增长，它还涉及数据类型的不断增多。以往传统的数据以结构化数据为主（可

以用二维表结构存储在关系数据库中，如常用的 Excel 软件所处理的数据），但是，现在更多互联网多媒体应用的出现，使诸如图片、声音和视频等非结构化数据占到了很大比重，从而造成了半结构化和非结构化数据类型与数量的井喷。

统计显示，结构化数据的年增长率大概是 32%，而非结构化数据的年增长率则是 63%，目前全世界非结构化数据已占数据总量的 80%以上。非结构化数据的比重越来越大，并显示出其中蕴含着不可小觑的商业价值和经济价值，这就对传统的数据分析处理算法和软件提出了挑战。

Q4. 什么是大数据的快速化?

大数据的快速化是指大数据的处理速度快。快速化是对大数据处理速度的要求。随着经济全球化趋势的形成，生产要素成本不断上升，企业面临的竞争环境越来越严酷。在此情况下，能够及时把握市场动态，迅速洞察产业、市场、经济、消费者需求等各方面的动态，并能快速制定出合理准确的生产、运营、营销策略，就成为企业提高竞争力的关键。而对大数据的快速处理分析，将为企业实时洞察市场变化、迅速做出响应、把握市场先机提供了决策支持。

原先针对传统数据的分析挖掘，主要是实现对过往发展的总结，而通过对大数据的分析，我们将能够快速地、准确地预测业务发展的趋势和方向，实时地调整业务重心，从而实现对未来市场的可预判性。

Q5. 什么是大数据的价值化?

价值是大数据的终极意义所在。随着社会信息化程度的不断提高、数据存储量的不断增加、数据来源和数据类型的不断多样化，对于企业而言，数据正成为企业的新型资产，形成竞争力的重要基础。与曾经广为提倡的“品牌价值化”一样，“数据价值化”已经成为企业提高竞争力的下一个关键点。

然而，大数据的价值虽然巨大，但价值密度却很低，往往需要对海量的数据进行挖掘分析，才能得到真正有用的信息，从而形成用户价值。大数据价值密度低的特性给大数据的分析处理带来挑战。

Q6. 大数据的起源是什么?

随着信息化的不断深入，“大数据”并不是突然产生的概念，而是 IT 技术发展的必然产物。在“大数据”这一概念产生之前，IT 界已经意识到信息和数据的不断持续增长的态势，并提出了“信息爆炸”和“海量数据”等概念。

随着近年来企业信息化的日臻成熟、社会化网络的兴起，以及云计算、移

动互联网和物联网等新一代信息技术的广泛应用，全球数据的增长速度逐步加快。据估算，全球数据正以每年超过 50%的速度爆发式增长。

“大数据”一词首次被提出是在 2011 年有关机构发布的研究报告——《大数据：创新、竞争和生产力的下一个新领域》之中。这份报告研究了数据和文档的状态，同时分析了处理这些数据能够释放出的潜在价值。

此后，全球 IT 巨头纷纷把长期部署的海量数据设备、数据分析、商务智能等硬件、软件与服务以“大数据”这一概念推向战略前沿。实际上，近几年来，IBM、甲骨文、EMC、SAP 等国际 IT 巨头已经花费超过 15 亿美元用于收购相关数据管理和分析厂商，以实现大数据领域的技术整合。大数据的发展历程如图 1-1 所示。

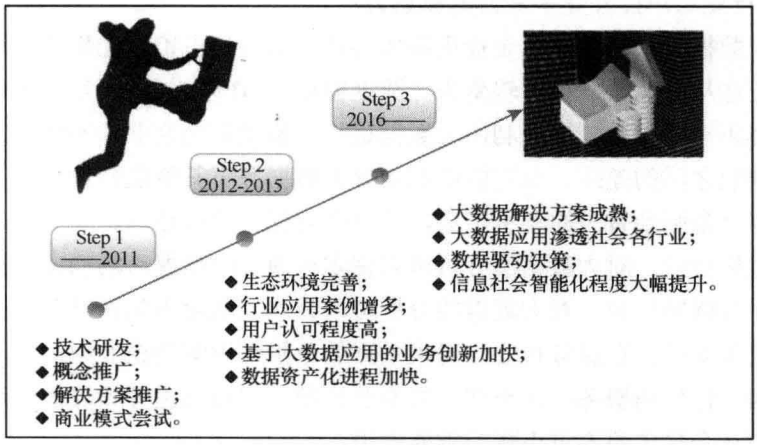


图 1-1　大数据的发展历程

Q7. 大数据带来了哪些机遇?

1. 新一代信息技术融合应用新焦点

对大数据的处理和分析正成为未来新一代信息技术融合应用的核心支撑节点。物联网、移动互联网、数字家庭、社会化网络等都是新一代信息技术具体的应用形态，大数据伴随着这些应用不断增长，云计算为这些海量的、多样化的大数据提供了存储和运算的支撑平台。以大数据为节点，各项新一代信息技术应用产生的信息将不断汇集，并通过对不同来源数据的统一性和综合性的处理、分析与优化，将结果反馈或交叉反馈到物联网、移动互联网、数字家庭、社会化网络等应用中，又将进一步改善使用体验，并创造出巨大的商业价值、经济价值和社会价值。

2．信息产业持续高速增长的新引擎

大数据因其巨大的商业价值和市场需求正成为推动信息产业持续高速增长的新引擎。随着行业用户对大数据价值的认可程度的增加，市场需求将出现井喷，面向大数据市场的新技术、新产品、新服务、新业态会不断涌现，大数据将为信息产业打开一个高速增长的新市场。

在硬件与集成设备领域，大数据面临的有效存储、快速读/写、实时分析等挑战，将对芯片、存储产业产生重要影响，还将催生一体化数据存储处理服务器、内存计算等市场。

在软件与服务领域，大数据中蕴含的巨大价值带来对数据快速处理和分析的迫切需求，将引发数据挖掘和商业智能市场的空前繁荣。

3．行业用户提升竞争能力的新动力

对大数据的利用将成为企业提高核心竞争力并抢占市场先机的关键。企业的决策正在从“业务驱动”转变为“数据驱动”。在未来3～5年，我们将会看到那些真正理解大数据并能利用大数据进行价值挖掘的企业和不懂得大数据价值挖掘企业之间的差距。真正能够利用好大数据并将其价值转化成生产力的企业，必将具备强劲有力的竞争优势，从而成为行业的领导者。

在零售行业，对大数据的分析可以使零售商实时掌握市场动态并迅速做出应对；在互联网行业，对大数据的分析可以为商家制定更加精准有效的营销策略提供决策支持；在服务行业，对大数据的分析可以帮助企业为消费者提供更加及时和个性化的服务；甚至在公共事业领域，大数据也开始发挥促进经济发展、维护社会稳定等不可小觑的重要作用。

Q8. 大数据带来了哪些挑战?

1．数据分析与管理人才紧缺

人才是大数据带来的挑战之一。研究表明，单单在美国，对拥有深厚的海量数据分析（包括机器学习和高级统计分析）技能人才的需求，可能超出目前预测供应量的50%～60%。到2018年，需要新增多达14万～19万名专家。此外，还需要150万名熟悉如何应用海量数据的管理者和分析员。企业必须加大招聘和人才挽留力度，同时大力投入关键数据人员的教育和培训。

2．用户隐私与便利性的冲突

大数据对个人信息获取渠道拓宽的需求引发了另一个重要问题，即隐私和便利性之间的冲突。研究表明，消费者受惠于海量数据，如更低的价格、更符合消费者需求的商品，以及从改善健康状况到提高社会互动顺畅度等生活质量

的提高；但同时，随着个人购买偏好、健康和财务情况的海量数据被收集，人们对隐私的担忧也在增大。

3．数据安全的风险更加凸显

数据安全在大数据时代也同样面临挑战。大数据发展的趋势往往与加大信息开放度、设计新的信息收集设备以及为海量数据的庞大存续和分析需求提供支持的云计算等如影随形。带来的副作用是 IT 基础架构将变得越来越一体化和外向型，从而对数据安全和知识产权构成更大威胁。

Q9. 什么是结构化数据?

结构化数据是指属性固定、能够严格用关系模型刻画的数据，一般存放在关系型数据库中。结构化数据的每个属性一般不能再进一步分解，具有明确的定义。典型的结构化数据，如商品数据可以表示成（商品名称，商品价格，商品产地，保质期）。

Q10. 什么是非结构化数据?

非结构化数据是指没有固定属性结构的数据，数据本身包含的信息不能简单地从不同的属性角度观察。比如一封邮件，包含标题、发送者、正文，但是，要了解这封邮件的内容，必须对正文部分做进一步的分析才能获得，而正文的内容无法通过几个固定属性来进行刻画。

Q11. 大数据的技术特点是什么?

提高数据处理的能力和容量是所有数据处理技术致力解决的问题。理想的数据处理技术在一致性、分区容忍性以及可用性等方面都应该得到很好的满足，但实际上这些特性从实现角度看往往具有冲突，比如对容量的追求会要求数据存放于不同的节点，但一旦位于不同的节点，数据的一致性操作就需要跨节点协作，引入的网络 I/O 延迟以及不可靠性导致操作的效率和成功率相比在一个节点内部都会大大降低，这是大规模数据处理不可避免的问题。大数据技术本质上是放弃了一些数据处理的要求，比如牺牲一致性要求很高的事务处理（多个操作的组合），仅提供最简单的读/写来达到超大规模的存储和访问能力。可以说，简单与大规模是大数据技术的重要特点。

Q12. 大数据有哪些处理模式?

根据数据源的性质以及分析目标不同，数据处理大致分为离线 / 批量和