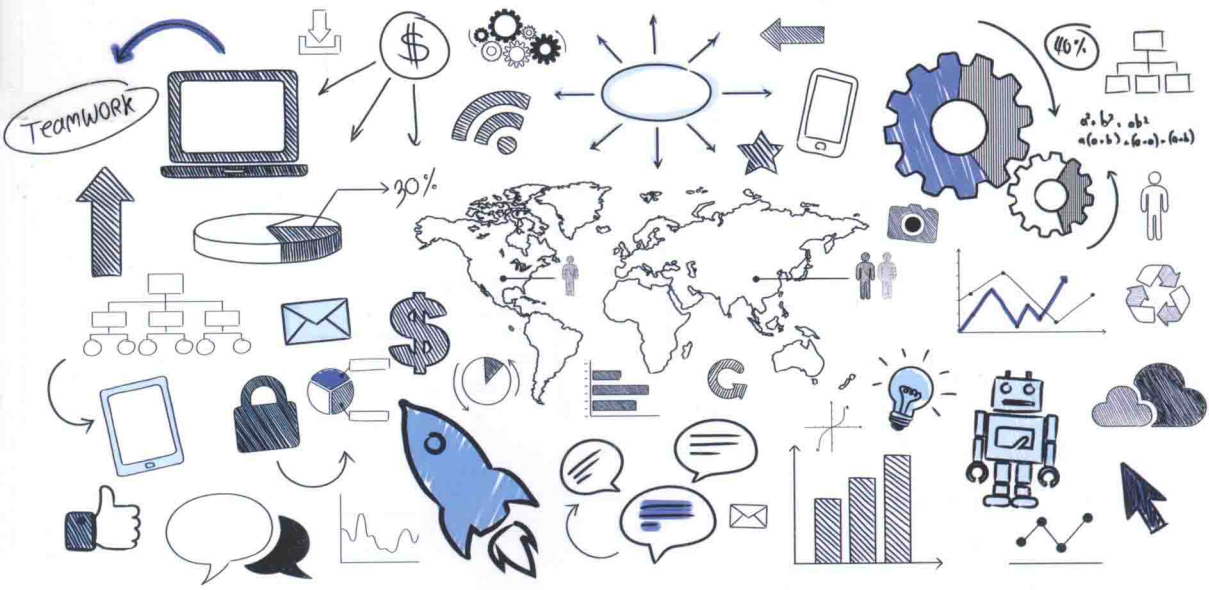

BIG DATA领跑大数据时代

你也许对“大数据”已经有所耳闻，但是你是否真的了解“大数据”的奥秘？你是否了解大数据对当今商业的影响？你是否洞察到了大数据背后企业经营实质的变革？

你是否能真正识别数据的盲点和噪音、抓住了海量信息中的核心信息……就让我为你解开心中的这些困惑。

驾驭人生未来的利器

领跑大数据时代 · 引领时代大变革

孙向杰◎编著

辽海出版社

BIG DATA领跑大数据时代

未来的利器
·引领时代大变革

孙向杰◎编著

辽海出版社

图书在版编目(CIP)数据

领跑大数据时代 / 孙向杰编著. — 沈阳 : 辽海出版社，2015.12

ISBN 978-7-5451-3596-1

Ⅰ. ①领… Ⅱ. ①孙… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2015)第 294458 号

责任编辑：丁　雁
封面设计：孙希前
责任校对：晓　云

出 版 者：辽海出版社
地址：沈阳市和平区十一纬路 29 号
邮编：110003
电话：024-23284381
E-mail：dszbs@mail.lnpgc.com.cn
http：//www.lhph.com.cn
印 刷 者：北京毅峰迅捷印刷有限公司
发 行 者：辽海出版社

幅面尺寸：170mm×240mm
印　　张：15
字　　数：230 千字

出版时间：2016 年 4 月第 1 版
印刷时间：2016 年 4 月第 1 次印刷
定　　价：35.00 元

前言 QIAN YAN

当前，大数据正以一种革命风暴的姿态引发全球关注。阿里巴巴马云指出，“互联网 +”已从 IT 时代到 DT（数字科技）时代，而 DT 是一个数据更充分流动的时代。而且未来大数据会作为一种资产存在，并将诞生一个万亿级别的交易市场。

有人将大数据比作“原油”，其实大数据挖掘才是大数据的核心。据公开数据显示，2013 年中国产生的数据总量超过0.8 ZB，相当于装满 8 亿个容量为 1 TB 的移动硬盘。如果不具备挖掘能力，如此海量的数据只能处于休眠状态。大数据通过数据挖掘技术，将海量数据进行归纳、建模、分析，找到数据中的关联关系，从而得出事情发生的可能性。打个比方，大数据会告诉商家客户喜欢什么，甚至可以精确到每一位客户的喜好。同时，大数据挖掘还需要众多高性能计算机同时承担数据存储、数据处理、数据挖掘的工作，这便是云计算。大数据挖掘必须依托云计算的分布式处理、分布式数据库和云存储、虚拟化技术，才可以快速实现。

有这样一个故事。

2003 年，奥伦·埃齐奥尼准备乘坐从西雅图到洛杉矶的飞机去参加弟弟的婚礼。他知道飞机票越早预订越便宜，于是他在这个大喜日子来临之前的几个月，就在网上预订了一张去洛杉矶的机票。在飞机上，埃齐奥尼好奇

地问邻座的乘客花了多少钱购买机票。当得知虽然那个人的机票比他买得更晚，但是票价却比他便宜得多时，他感到非常气愤。于是，他又询问了另外几个乘客，结果发现大家买的票居然都比他的便宜。

对大多数人来说，这种被敲竹杠的感觉也许会随着他们走下飞机而消失。然而，埃齐奥尼是美国最有名的计算机专家之一，从他担任华盛顿大学人工智能项目的负责人开始，他创立了许多在今天看来非常典型的大数据公司，而那时候还没有人提出“大数据”这个概念。

飞机着陆之后，埃齐奥尼下定决心要帮助人们开发一个系统，用来推测当前网页上的机票价格是否合理。作为一种商品，同一架飞机上每个座位的价格本来不应该有差别。但实际上，价格却千差万别，其中缘由只有航空公司自己清楚。

埃齐奥尼表示，他不需要去解开机票价格差异的奥秘，他要做的仅仅是预测当前的机票价格在未来一段时间内会上涨还是下降。于是，埃齐奥尼开始着手启动这个项目。

埃齐奥尼创立了一个预测系统，它帮助虚拟的乘客节省了很多钱。这个预测系统建立在41天内价格波动产生的12000个价格样本基础之上，而这些信息都是从一个旅游网站上搜集来的。这个预测系统并不能说明原因，只能推测会发生什么。也就是说，它不知道是哪些因素导致了机票价格的波动。机票降价是因为很多没卖掉的座位、季节性原因，还是所谓的周六晚上不出门，它都不知道。这个系统只知道利用其他航班的数据来预测未来机票价格的走势。“买还是不买，这是一个问题。”埃齐奥尼沉思着。他给这个研究项目取了一个非常贴切的名字，叫“哈姆雷特”。

这个系统为了保障自身的透明度，会把对机票价格走势预测的可信度标示出来，供消费者参考。系统的运转需要海量数据的支持。为了提高预测的准确性，埃齐奥尼找到了一个行业机票预订数据库。有了这个数据库，系统进行预测时，预测的结果就可以基于美国商业航空产业中，每一条航线上每

一架飞机内的每一个座位一年内的综合票价记录而得出。这就是大数据的魅力。

本书首先介绍了大数据时代的特征，可以帮助你对大数据及其价值有一个概括性的了解和认识。其次，你将知道如何培养、挖掘、处理数据，使数据为自己创造更大价值。最后，介绍了大数据在企业决策、运营管理、金融投资等方面的实际应用。内容简单实用，特别适合初级读者阅读。

目录 MU LU

第三章　大数据引导大创业

第四章　大数据面临的难题

第五章　大数据创造大价值

第六章　大数据促进大变革

第七章　大数据演绎大未来

第八章　大数据驱动大营销

第九章　大数据带来大应用

第一章

大数据到底是什么

对于“大数据”研究机构 Gartner Group(高德纳咨询公司)给出了这样的定义。大数据是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。

大数据的“前世今生”

“大数据”是什么？要回答这个问题首先要看看数据是怎样产生的。

在信息化时代里，我们每个人都在贡献数据。上网、打电话、发短信、听歌、拍照片、发帖子、看视频，都会产生数据，就像涓涓细流汇聚成江河湖海，“大数据”出现了。

近年来，数据大爆炸的速度快得惊人。马云曾感慨地说：“大家还没搞清PC的时候，移动互联网来了，还没搞清移动互联网的时候，大数据时代来了。”

大数据时代来得太快，以至于人们对大数据的定义都有N多种。

按照美国国家标准与技术研究院发布的研究报告的定义：“大数据是用来描述在我们网络的、数字的、遍布传感器的、信息驱动的世界中呈现出数据泛滥的常用词语。大量数据资源为解决以前不可能解决的问题带来了可能性。”

按照业界权威高德纳咨询公司的定义：“大数据是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。”

根据百度百科词条的定义：“大数据，或称巨量资料，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、处理并整理成为帮助企业经营决策更积极目的的资讯。”

“大数据”到底有多大？目前通行说法，“大数据”至少要达到PB量级。其中，1 PB = 22.3万张DVD光盘的容量，相当于800个人类大脑记忆总量，或90个人身体细胞数目总和。1 PB的MP3歌曲可以连续播放2000年。

美国互联网数据中心指出，互联网上的数据每年将增长50%，每两年便将翻一番，而目前世界上90%以上的数据是最近几年才产生的。全世界的工业设备、汽车、电表等设备上有着无数的数码传感器，随时测量和传递着有关位置、运动、震动、温度、湿度乃至空气中化学物质的变化，也产生着海量的数据信息。

早在1980年，著名未来学家阿尔文·托夫勒便在《第三次浪潮》一书中，将大数据赞颂为“第三次浪潮的华彩乐章”。不过，大约从2009年开始，“163大数据”才成为互联网信息技术行业的流行词汇。

如今，大数据技术可以帮助人们做很多以前做不到的事情。比如，国外某警察局利用大数据预测犯罪的发生几率，可以精确到街区500平方英尺的范围内，有针对性地预防，从而使该地区犯罪率明显下降；某统计学家利用大数据预测总统选举结果；某大学利用手机定位数据和交通数据建立城市规划等。

大数据时代，海量的数据已经成为一种“矿藏”。据测算，三年前，2011年全球大数据产值51亿美元；预计三年后，2017年全球大数据产值将达到534亿美元。目前大量“掘金者”在数据的海洋里挖掘、采集、提炼、分析，从而得出有价值的信息提供给政务的、商务的以及各个领域的买家，从而形成了大数据产业生态圈。

“书同文，车同轨。”任何新兴产业要健康发展，首先要尽快建立大家共同遵守的标准。目前国内外大数据标准化工作已经起步。全国信标委已经对标准化工作进行梳理，从基础、技术、产品、应用等不同角度进行分析，形成了大数据标准体系框架，并发布了《大数据标准化白皮书》。

大数据的四个来源

当今世界，大数据无处不在，它影响到了我们的工作、生活和学习，并将继续施加更大的影响。

大数据用于描述这样的数据组，其规模超出了日常软件在可容忍期限内获取、管理和加工数据的能力。一些网络技术领先的公司持续地投资于昂贵的大数据技术，成效显著。大数据使得创新型公司变成了经营新方法的率先接受者，经营更为成功。通过大数据的分析挖掘，公司可以发现新的经营模式，对工艺加以改进。例如，在获悉消费者行为后，可以将发现用于某些改变，如降低成本或增加销售，就会产生价值。在任意大的数据组中应用统计方法可以发现有用信息，将这些信息商业化即可获益。

大数据时代一切在变，应对之策是改变一切。经营方式发生了变化——制定决策变得与开展行动深度融合；运用信息的方式发生了变化——从处在经营的边缘变成了处于所有方面的中心；技术发生了变化——从批处理到实时处理，从分割到融为一体；人们工作的方式发生了变化——从在命令和控制模式下运作到在合作环境下负责自己的信息和交互应用。

根据麦肯锡全球研究所的分析，利用大数据在各行各业能产生显著的财务价值。美国健康护理利用大数据每年产出 3000 亿美元，年劳动生产率提高 0.7%；欧洲公共管理每年价值 2500 亿欧元，年劳动生产率提高0.5%；全球个人定位数据服务提供商收益 1000 多亿美元，为终端用户提供高达 7000 亿美元的价值；美国零售业净收益可增长 6%，年劳动生产率提高0.5% ~1%；制造业可节省 50% 的产品开发和装配成本，营运资本下降 7%。

当今大数据的来源除了专业研究机构产生大量的数据外（欧州核子研究

组织（CERN）的离子对撞机每秒运行产生的数据高达40 TB），与企业经营相关的大数据可以划分为四个来源：

1. 越来越多的机器配备了连续测量和报告运行情况的装置。几年前，跟踪遥测发动机运行仅限于价值数百万美元的航天飞机。现在，汽车生产商在车辆中配置了监视器，连续提供车辆机械系统整体运行情况。一旦数据可得，公司将千方百计从中渔利。这些机器传感数据属于大数据的范围。

2. 计算机产生的数据可能包含着关于因特网和其他使用者行动和行为的有趣信息，从而提供了对他们的愿望和需求潜在的有用认识。

3. 使用者自身产生的数据、信息，人们通过电邮、短信、微博等产生的文本信息。

4. 至今最大的数据是音频、视频和符号数据。这些数据结构松散，数量巨大，很难从中挖掘有意义的结论和有用的信息。

大型以 Internet 为核心的公司，如 Amazon，Google，eBay，Twitter 和 Facebook 正使用后三类海量信息认识消费行为，预测特定需求和整体趋势。第一类数据可能产生较少的业务，但可以推动某些经营模式实质变革。例如，汽车传感数据用于评价司机行为会推动汽车保险业的深刻变革。

大数据改变了所有行业全部公司的经营方式。从对市场的理解到如何挖掘经营信息，大数据能洞察每项转变。一个致力于收集和分析大数据的行业业已形成，对现有公司产生了深刻影响。据有关调查，有 10% 的公司认为在过去的五年中，大数据彻底改变了它们的经营方式；46% 的公司认同大数据是其决策的一项重要支持因素。

大数据的四个特征

大数据（Big Data）是指“无法用现有的软件工具提取、存储、搜索、共享、分析和处理的海量的、复杂的数据集合”。业界通常用四个 V（即 Volume、Variety、Value、Velocity）来概括大数据的特征。

1. 数据体量巨大（Volume）。截至目前，人类生产的所有印刷材料的数据量是200 PB（1 PB = 210 TB），而历史上全人类说过的所有的话的数据量大约是5 EB（1 EB = 210 PB）。当前，典型个人计算机硬盘的容量为 TB 量级，而一些大企业的数据量已经接近 EB 量级。

2. 数据类型繁多（Variety）。这种类型的多样性也让数据被分为结构化数据和非结构化数据。相对于以往便于存储的以文本为主的结构化数据，非结构化数据越来越多，包括网络日志、音频、视频、图片、地理位置信息等，这些多类型的数据对数据的处理能力提出了更高要求。

3. 价值密度低（Value）。价值密度的高低与数据总量的大小成反比。以视频为例，一部1小时的视频，在连续不间断的监控中，有用数据可能仅有一两秒。如何通过强大的机器算法更迅速地完成数据的价值“提纯”成为目前大数据背景下亟待解决的难题。

4. 处理速度快（Velocity）。这是大数据区分于传统数据挖掘的最显著特征。根据国际数据公司（IDC）的“数字宇宙”的报告，预计到2020年，全球数据使用量将达到35.2 ZB。在如此海量的数据面前，处理数据的效率就是企业的生命。

根据麦肯锡旗下研究部门麦肯锡全球学会2011年发布的一份报告显示，预计美国需要14万~19万名拥有“深度分析”专长的工作者，以及150万名