

BIG DATA
一本书读懂
大数据

黄颖◎编著

看懂大数据 掌握时代先机

化繁为简说数据

每个人都看得懂的大数据入门书

大数据时代
必备宝典
得数据者
得天下

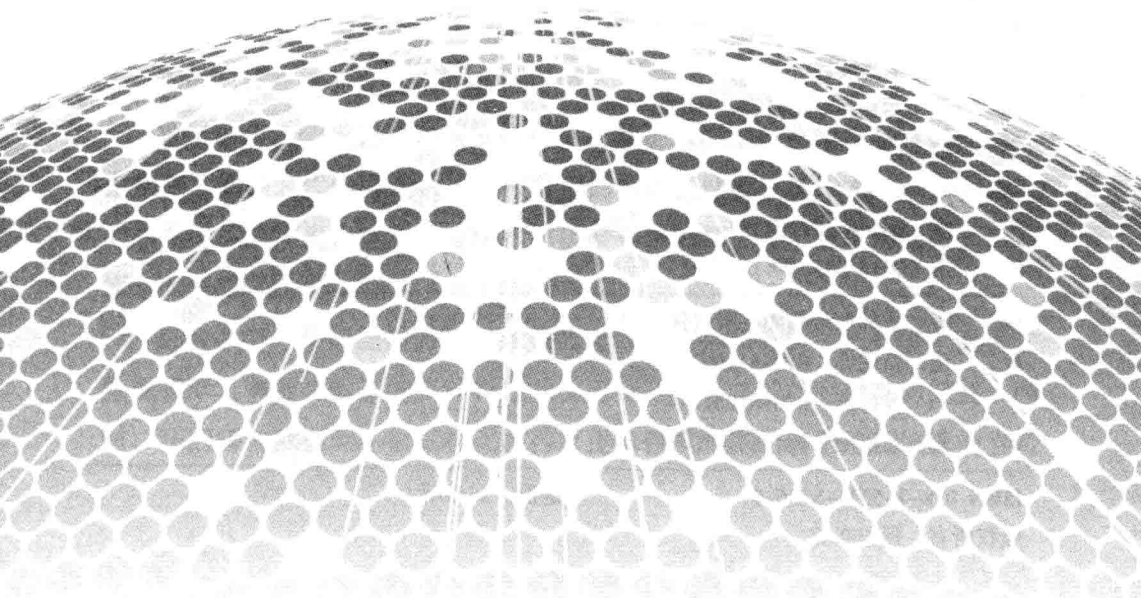


吉林出版集团有限责任公司

BIG DATA

一本书读懂 大数据

黄颖◎编著



图书在版编目 (CIP) 数据

一本书读懂大数据 / 黄颖编著. — 长春 : 吉林出版集团有限责任公司, 2014.11

ISBN 978-7-5534-5736-9

I. ①—… II. ①黄… III. ①商业信息学
IV. ①F713.51

中国版本图书馆CIP数据核字 (2014) 第225942号

一本书读懂大数据

编 者 黄 颖

策划编辑 李异鸣

特约编辑 周乔蒙

责任编辑 齐 琳 王 平

封面设计 上尚装帧

开 本 787mm × 1092mm 1/16

字 数 170千字

印 张 13.25

版 次 2014年12月第1版

印 次 2014年12月第1次印刷

出 版 吉林出版集团有限责任公司

电 话 总编办: 010-63109269

发行部: 010-81282844

印 刷 北京旭丰源印刷技术有限公司

ISBN 978-7-5534-5736-9

定价: 36.00元

版权所有 侵权必究

前 言

世界的万千变化一直超乎人们的预测，自2012年以来，大数据一词成了人类生活的代名词。如今，数据几乎已经渗透到了每一个行业的每一个领域之中，成了不可或缺的生产因素。每一天，互联网都会繁衍出无数的数据，这些内容足以刻满2亿张光碟；而手机客户端发出的帖子和邮件总数可达到3000万亿……如此惊人的数据使得对海量数据的挖掘和分析，成了企业发展的重要内容。大数据的数量大、类型多、时效快、价值密度低的特点，让这个崭新的时代充满了变数和乐趣。

数据迅速地膨胀，让差别细微的算法就足以决定企业的发展方向。很多企业在大数据时代纷纷进行了多种多样的尝试，这是一场伟大的革命，庞大的数据资源的冲击，让商界、学术界等所有领域都开始了量化的流程，积极探索大数据时代的奥秘。

这本书中，我们能够揭开大数据的面纱，挖掘和分析大数据整个流程的重要关卡，掌握大数据的多种特性和价值特征，对其结构有精准的把握。同时，我们将大数据和小数据时代进行对比，让读者更加清晰地认识我们生活

的时代。

电子商务界乘着大数据的狂风，掀起了人们生活的数据风暴。国内外的企业使尽浑身解数，在大数据的海洋里摸爬滚打。从市场来看，阿里巴巴、小米、亚马逊的行动最为迅速、高效。任何企业，只有把握住了大数据时代的机遇、接受大数据时代的挑战，企业才能拥有了缩短发展时间、完成飞跃的筹码。

大数据和企业的生存发展息息相关，企业的管理层需要全面的数据源来确定正确的航向，全面的数据源搜索和分析需要专业人才，专业的人才需要经历商业气息的洗礼。这一切都成了企业在大数据时代直接面临的机遇和巨大挑战。得数据者得天下不再是一句标语，企业的整个商务链条都需要数据的支撑来保驾护航，失之毫厘谬以千里的教训时时刻刻都可能发生。重视大数据、对大数据了解详尽的企业高管才有可能带着企业在正确的路上，获得竞争优势。

进入大数据时代，让数据开口说话将成为司空见惯的事情，本书将从大数据时代的前因后果讲起，全面分析大数据时代的特征、企业实践的案例、大数据的发展方向、未来的机遇和挑战等内容，展现一个客观立体、自由开放的大数据时代。

目 录

第一章 身处数据时代，揭开大数据的面纱

大数据到底是什么？ / 002

“大”是重点，还是“数据”是重点？ / 006

与众不同的大数据 / 008

大数据方式下的云计算 / 010

大数据的奥秘 / 013

当下是大数据发展的最佳时机 / 017

第二章 大数据如此重要，引无数英雄竞折腰

多样的非结构性数据 / 020

大数据的价值发掘 / 023

大数据的结构化、非结构化、半结构化及多结构化 / 026

大数据是扩展性的下一代传统数据 / 028

是什么构成了大数据价值链？ / 030

大数据时代真的来了 / 036

第三章 求挖掘与分析，电子商务与大数据

大数据时代中的电子商务 / 040

亚马逊在大数据时代的实践 / 042

小米手机在大数据时代的实践 / 047

小米手机对“米粉”需求的文化挖掘 / 051

阿里巴巴数据化运营的那些“大招” / 054

大数据中的企业价值及客户价值 / 061

第四章 数据和企业管理，高层更看重大数据

沃尔玛如何用数据构建管理模式 / 064

让大数据进入企业管理 / 066

职业乞丐脑袋里的大数据 / 070

职业经理人与大数据 / 073

企业组织管理不介入大数据，就要被淘汰 / 077

第五章 生活无处不数据，大数据真的能算命？

未来的先兆——大数据 / 084

大数据带来的经营理念的转变 / 088

大数据的舆情服务 / 094

大数据预测你的下一步行动 / 096

数据也会骗人，从人的动作推导数据 / 099

网络数据背后的价值 / 103

第六章 颠覆与重塑思维，大数据与思维革命

大数据时代的综合人才 / 110

飞利浦的大数据营销策略 / 113

阿里小贷的“不可能的任务” / 115

第三方支付业务的另一种思路 / 119

P2P网络借贷动了谁的奶酪 / 121

大数据带来的智能化与柔性化 / 124

生活、工作、思维的颠覆重构 / 128

第七章 得数据者得天下，商业竞争中的大数据

大型公司的垂直一体化趋势 / 136

客户形象的丰富源于对客户的全局理解 / 139

量化奠定了数据化的内核 / 143

文字的数据化进程 / 147

地理位置的数据化构建 / 151

数据化的沟通方式 / 156

企业竞争力的关键——大数据 / 159

第八章 让数据张口说话，管理决策中的大数据

客观数据最具发言权 / 164

挖掘潜力股的数据化进程 / 167

时代因大数据而变革 / 170

大数据时代的风险规避策略 / 174

企业文化的数据化构建 / 177

第九章 更自由，更开放，大数据的机遇和挑战

人机结合的未来发展趋势 / 184

数据时代，引发时代大变革 / 186

数据可以表示世间万物，会带来惊喜 / 191

数据化带来的挑战前所未有 / 194

第一章

身处数据时代，揭开大数据的面纱

科技的迅速发展，互联网金融的兴起和繁荣，把数据推到了所有金融元素的核心位置。越来越多的企业逐渐认识到只有掌握正确的数据并看透数据背后的故事，才能够获得源源不断的财富。大数据时代伴着铿锵有力的节奏引领了世界的新潮流。

大数据到底是什么？

如果要追溯“大数据”这个专业术语最初的出处的话，就必然要提及apache org的开源项目Nutch。在那个时候，大数据的意思是更新网络搜索索引，同时还需要批量处理和分析大量的数据集。谷歌的Map Reduce和Google File System（GFS）发布了之后，大数据的定义中除了涵盖大量数据之外，还包括数据处理的速度。

研究机构Gartner曾给大数据（Big data）下过这样的定义：大数据是一种基于新的处理模式而产生的具有强大的决策力、洞察力以及流程优化能力的多样性的、海量的且增长率高的信息资产。

大数据一词源于英文的“Big Data”一词，以往也有类似的词语，如“信息爆炸”“海量数据”等等似乎都很难去准确描述这个词的具体内涵。麦肯锡全球研究所所做的《大数据：创新、竞争和生产力的下一个前沿》（James, 2011）是这么定义“大数据”的：

大数据通常指的是大小规格超越传统数据库软件工具抓取、存储、管理和分析能力的数据库。这个定义也有很强的主观色彩，因为究竟什么样规格的数据才是大数据，这没有统一的标准，也就是无法确定超过多少TB（1000GB）的数据才是大数据。随着时间的推移和技术的发展，我们必须知

道“大数据”的量会越来越大。还有一点，这定义也会因为部门的差异而发生标准的变化，这和通用的是什么软件以及特定行业数据集的大小有密切的关系。所以，现有各行业的大数据可以是几十TB，也可以是几千TB。

按照EMC的界定，特指的大数据一定是指大型数据集，规模大概在10TB。通过多用户将多个数据集集合在一起，能构成PB的数据量。

在IBM2011IOD大会上，负责IBM软件和硬件两大集团的高级副总裁Steve Mills曾说过：“分析已经成为必要的能力，不再只是一个工具，是一种能让业务流程运转的智慧能力。企业要转化信息的洞察力为行动，而不是仅仅去争取竞争的优势，要将其转换为生存的根本。”

IBM公司概括大数据时有三个V，也就是大量化（Volume），多样化（Variety）和快速化（Velocity），此外它们还针对客户有了“大数据解决方案”的服务。IBM公司对大数据所概括出的三个V，其实也说明大数据潜藏的另一个V——价值（Value）。就这么说的话，大数据确实具备这四个V的基本特征。

大数据的第一个特征是数据的量大。电脑的数据运算和储存单位都是字节（byte），1KB（kilobyte）等于1024B，就是千字节。除此之外还有更高的单位MB（Megabyte兆字节），GB（Gigabyte，吉字节），TB（Trillion byte，太字节）、PB（Pet byte，拍字节），EB（Exabyte，艾字节），ZB（Zetta byte，泽它字节）和YB（Yotta byte，尧字节）。每一级之间的换算关系是1024。到了2009年，几乎每一个美国企业，只要是雇员人数超过1000人的，它的数据存储量大概都超过了200TB，这是十年前沃尔玛公司数据仓库存储量的2倍还多。在不少经济部门当中，企业平均的数据存储量甚至都达到了1PB。2010年欧洲组织的存储总量大概为11EB，这个数字几乎是整个美国数据总量（16EB）的70%。2010年全球企业在硬盘上的数据存储量已经超过了7EB，而在PC和笔记本电脑等设备上的个人存储量也超过了6EB。美国国会

图书馆当时存储的数据大概只是1EB的4000分之一（James，2011）。硬件技术的发展速度远远赶不上数据容量的增长速度，为此数据存储和处理的危机应运而生。巨大数量的数据被处理掉，例如医疗卫生提供商会将它们90%的数据给处理掉（这其中包括几乎所有在手术过程中产生的实时视频和图像资料）。

只不过，大数据不单纯只是大。海量数据存储危机的产生不仅仅是由于数据量爆炸性的增长，还有数据类型的改变带来的，这就是第二个V，多样化。此前的数据库用二维表结构存储方式就可以储存数据，譬如常见的Excel软件中处理的数据，这称为结构化数据。可是现在随着互联网多媒体应用的出现，像是声音、图片和视频等等非结构化的数据所占的比重在日益增多。有统计表明，全世界非结构化数据的增加率是63%，相对而言结构化数据增长率只有32%。2012年，非结构化数据在整个互联网数据中的占比已经超过了75%。

Informatica中国区的首席产品顾问但彬就提到过，大数据里有海量数据的含义，但它又大于海量数据的定义。简单来说，海量数据加上其他复杂类型的数据就是大数据的概念了。但彬还提到，所有交易和交互数据集都属于大数据，它的规模和复杂程度早已在依据合理成本和时限进行捕捉、管理和处理数据集的传统技术的能力之上。

简而言之，三种主要技术趋势汇聚成了大数据：其一是海量交易数据，包括半结构化和非结构化信息，在从ERP应用程序到基于数据仓库应用程序的在线交易处理（OLTP）和分析系统的过程当中总在不断增长。企业很多的数据和业务流程也在不断走向公共和个人云转移，将造成更为复杂的局面。其二是海量交互数据。因为Facebook、Twitter、LinkedIn以及其他更多的社交媒体的兴起，这一部分数据诞生了海量的交互数据，其中涵盖了呼叫详细记录（CDR）、设备和传感器信息、GPS和地理定位映射数据，还有利

用管理文件传输（Manage File Transfer）协议传送的海量图像文件、Web文本和点击流数据、科学信息、电子邮件，等等。其三就是海量数据处理。随着大数据的涌现，已经有很多用于密集型数据处理的架构应运而生，比如Apache Hadoop，它具有开放源码以及在商品硬件群中运行的特性。此外还有能以可靠、高效、可伸缩的方式分布式处理大数据的软件框架Hadoop。它之所以可靠，是因为它能够提前假定计算元素和存储失败，所以它能够维护多个工作数据副本，用并行处理的方式来加快处理能力和速度。Hadoop也是可伸缩的，PB级的数据它也可以处理。另外，Hadoop因为依赖于社区服务器，所以它的成本很低，不论是谁都可以使用。对企业来说，最难的在于如何通过成本效益的方式从Hadoop中存取数据。Hadoop最知名的用户是脸谱。通过Hadoop，像脸谱这一类的网站，也就可以自由地处理海量的数据，同时获得较高的收益。

“大”是重点，还是“数据”是重点？

先来做一个小测验。当阅读开始前，先停下来思考这么一个问题：哪部分是术语“大数据”中最为重要的？是大，还是数据？还是二者都一样重要，或是都一样不重要？花一分钟的时间去思考这个问题。假如已经有了自己的答案，那就开始阅读接下来的内容。

既然答案已经有了，那就来看看哪个是正确的？显然，正确的答案应该是最后一个，事实上在大数据中，“大”和“数据”都不重要。其中最重要的是企业该如何去驾驭这些大数据。对大数据进行分析，以及在此基础上采取的业务改进才是最为关键的。

事实上，大数据本身是没有任何价值可言的。即便是一个人比另一个人拥有更多的数据，这也不代表什么。任何一个数据集，它们或大或小，本身都没有价值可言。如果不懂得如何去使用收集来的数据，那这些数据不会比地下室里的垃圾更有用。要是不投入环境或者是付诸使用的话，数据的意义就不在了。任何大量或是少量的大数据该如何体现自己的威力呢？要怎么去分析这些数据呢，又该如何去洞察或是采取什么样的行动呢？这些数据又要如何来改进业务呢？

很多人在阅读了众多炒作大数据的文章之后就相信之所以大数据要比其

他数据有优势，就在于它的容量大、速度快和多样性，这种说法并不准确。在很多大数据当中，相比以往数据会存在更多毫无价值或是价值很小的数据。一旦大数据被精简到实际需要的容量时，它们所呈现出来的就不是大数据了。事实上这也不重要，无论是它被精简还是保持原本庞大的模样，这些关系都不大，最重要的是处理它的方式。所以说使用数据要比起它的容量更为重要。

大数据庞大的规模并非人们所关注的，包括它们能带来巨大的内在价值也非关注的事实。最大的价值还在于分析的方式，以及采用什么样的方式来改进自己的业务。

在人们阅读一本书的时候，关键点的第一个是大数据的大数据量，并且要承认大数据也是数据中的一种。只不过这并非企业兴奋的理由所在。这些数据使用时的新颖且强大的分析方式才是企业注意力集中的地方。作为社交网络应用的Facebook和微博，都构建了关联普遍用户的行为数据。人们在网络上浏览网页、购买商品、游戏休闲原本是不关联的。当智能手机推广普及之后，网络的行为越来越碎片化了。假设没有一定的关联，就很难去分析和利用这些数据。社交网络提供给用户统一的借口，让无论是玩游戏还是买商品的客户可以轻松地吧碎片化的信息发到网络上。就像是一个用户行为数据连接器的角色一样，微博把所谓网络上用户的行为，完整地关联起来，画出一幅生动的网络生活图景，把用户的偏好、性格、态度等特征真实地反映出来，而这当中就是最为充分的商业机会。

彼此关联的数据价值要远大于孤立的数据。可是在当下数据孤岛是很常见的。个人计算机中的文件一般都会以某种类目来存放，内容和内容之间没太强的联系。企业之间也是如此，很多部门之间都壁垒林立，似乎每个人都愿意去保护自己的数据，从而形成“数据割据”的局面。只要是处在数据孤岛中，大数据所潜在的价值是很难被挖掘出来的。

与众不同的大数据

有别于传统数据源的大数据有不少重要的特征，不是每个大数据源都有这些特征存在，绝大多数的大数据或多或少地都存在一些这样的特征。

第一个特征是大数据的来源往往是机器自动的结果。人工不会干涉到新数据的产生过程，完全是机器自动的结果。如果拿传统数据源进行分析的话，就会发现它们的形成过程中会有人工的痕迹，像是零售业和银行交易、电话呼叫记录、产品发票等等，和某个人做的事情都有关系，无论什么情形，都会有人参与到新数据的形成过程中。可是大数据不是这样产生的，它不会在产生过程中与人互动，像是引擎中内置的传感器，即便没有人干预周围数据也会自动生成。

第二个特征是大数据作为一个全新的数据源，不仅仅是已有数据的收集扩展，比如在互联网中，顾客与银行、零售商之间可以直接在线交易。事实上这种交易方式和传统交易差异不大，不过是换一种渠道而已。企业通过收集网络交易数据就会发现这样情形下的数据和多年来他们得到的传统数据差异不大，不过是数量增加了而已。如果收集的是客户浏览行为的数据，那就会产生本质上全然不同的数据。

上面提到的相同类型数据，不过是数量多了的说法也会因为达到另一个