一本书读懂大数据（每个人都看得懂的大数据入门书）

作者：黄颖

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.一… II.黄… III.商业信息学 IV.F713.51

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

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

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

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群。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把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极端，成为最新的数据，比如说传统读电表都是人工方式，也就是说自动读取用电数据的智能电表所产生的数据就是类型相同，不过是数量增加了。不过这种数据在某种程度上也能成为一种有别于人工读取的数据，应用更为深层次的分析方式，这样一来它们就可以称作是新的数据源。

第三个特征是大数据中的大多数设计并非友好。实际上这些数据并未经过设计。就拿社交媒体网站上的文本流举例，用户不一定会被要求用标准的语序、语法和词汇表。人们的信息一经发布，社交平台就能够获得数据。这些不太规范的数据处理起来还是有一定困难的。在设计之初，大多数的传统数据都尽量要友好一些，就比如收集交易信息的系统最早生成数据会以整洁或是预先规范的方式来操作，这样形成的数据就更有利于加载和使用。还有一部分原因是由于要对空间进行高效利用，以避免出现空间不够的局面。

大数据有时候还会是凌乱和丑陋的。通常最开始传统数据就已经被严格地定义。每一比特的数据都存在重要的价值，这是必需的。一般大数据源一开始不会被严格定义，这和存储空间的开销越来越微乎其微有关，必须对各种有用的信息进行收集。所以说大数据分析的时候，各种凌乱丑陋的数据都有可能遇见。

最后的特征是海量数据并非有大量价值。实际的数据很多都是毫无价值的。在一篇网页日志当中，非常重要的数据就包含其中，当然也有好多没价值的数据也在其中。很有必要从中提炼最有价值的部分。定义传统数据源的起初就要求数据是百分百有用。这是因为可扩展性受到了限制，所以如果有没价值的信息在当中的话代价会非常昂贵。除了最初定义的有数据记录的格式外，数据内容和价值也被定义和约束了。当下存储空间的问题已经不存在了。大数据所收集的是所有的信息，然后再去解决这些冗余信息所带来的问题。只有这样才会不遗漏所有的信息，与此同时在分析数据时的麻烦也会让人头疼不已。

大数据方式下的云计算

消费者会觉得大数据和云计算很无聊，可是对于Delphix来说却是一座宝藏，因为它正在利用这种技术进行敏捷数据管理。

Delphix不需要部署冗余的基础设施在自己的敏捷数据管理解决方案之上，还能同时提升流程的速度。客

户因此能更为快捷地完成交付使用。其实敏捷数据管理就是企业数据库内虚拟化数据，再提高数据库驱动型应用的开发敏捷性质，因此使数据库和应用管理都发生大的改变。企业的数据库被Delphix放到了云上，再通过数据同步和虚拟化技术交给适当的人最恰当的数据。Delphix宣称有了应用交付解决方案后，应用项目的进度会提升5倍之多，成本会减少90%，事实上2010年Delphix面世后的销售增长率达到了300%。

成立于2010年的Delphix，2012年6月它的C轮融资就完成了2500万美元。这一次融资的领投是Jafco Ventures，投资人中还有Greylock Partners。迄今为止Delphix总融资金额高达4550万美元。公司依赖其“敏捷数据”拿到了超额认购。企业数据库的数据在“敏捷数据”的虚拟化作用下，增强了数据驱动应用的敏捷性，经济数据库和应用管理速度也提升了。

不少企业都把自己的目标设定为借由一个强大的平台来实现品牌推广，可是很多社交网站的数据还是找不到可行的商业模式，因为预期真正得以实现的不多。不过社交数据公司在不断发展壮大，可以想见不远的将来社交网站的影响力利用问题不会再是遥远的梦想。

像是纽约的SumAll公司期望就是要带给每个客户“小而美”的数据。SumAll所提供的平台在于提供给中小企业实时的数据服务，利用桌面、iPhone和安卓系统来访问，可以看到很多可视性的大量数据，也就更便于阅读和观看。SumAll在和Shopify、PayPal和Magento合作电子商务和支付系统的时候，用户点击几下就能完成账户的集成工作。SumAll对于实时数据的分析很快速，再为用户提供一个如社交媒体式的“新闻订阅”一样的简洁分析和见解。SumAll还会为客户提供深入挖掘税收、发货和出售量的服务，甚至连对客户依照不同标准的排序分析也可以完成。

2011年11月成立的SumAll，在2012年6月著名风险投资公司Battery Ventures牵头联合Wellington Partners、Matrix Partners和General Catalyst Partners为SumAll投资了150万美元的种子期融资。SumAll到2012年12月对外宣布获得了600万美元的A轮融资，还是Battery Ventures联合Wellington Patners对其进行投资。目前设在纽约总部的公司有25名员工。

还有Ngdata公司，企业用户和他们的消费者通过它们能够进行一对一的营销模式提供和得到最好的建议和产品。Ngdata曾推出过一个产品Lily集成了内外部的结构化和非结构化的数据。Lily还可以用人工智能拍照工具对消费者的习惯和爱好进行记录。正在快速成长的大数据市场，对企业的价值越来越大了，企业对市场的评估和行为的预判都要通过这些数据分析。ING的投资总监Tom Bousmans说过，消费者所产生的数据有上亿个，企业都可以通过这些来了解用户需求，彼此间还有个性和动态的互动。

成立于2009年的Ngdata的员工现有20名，它们还有类似Wibidata和Spire这样的竞争对手。Ngdata与竞争对手的不同在于它能够提供企业与消费者实现互动的数据解决方案，不仅是单纯专注在大批量数据分析之上。2012年10月Ngdata获得了250万美元的融资。这一次融资的资金主要来自ING、Sniper investment、Plug and Play Ventures等投资机构和一些天使投资人，这份资金将帮助Ngdata推广个性化产品线的拓展，并在纽约和旧金山专门为美国客户设立服务办公室。

Attivio的创始人Ali Riaz觉得企业用户每发送一条查询请求的时候，得到的信息都是具有洞察性的，绝非罗列出来的链接或是一张简单的图表。它回答的问题不仅是“是什么”还有“为什么”，就比如销售量下降是市场需求下降还是销售人员表现不够突出造成的。

任何一家企业要做的工作都是市场营销。近几年社会化媒体的兴起，让营销业者的注意力都集中在了数字营销之上，不过对于这个领域营销人员还欠缺有效的分析。Good Data公司正是瞅准这一商机，开始为营销人员提供集成服务，让他们可以利用微博等社交网络平台进行深度的分析。

大数据的奥秘

事实上并不是说大数据的处理就有多困难。收集一些数据，企业的分析专家团队就可以开始进行数据价值的探索。企业要做的就是要让分析专家团队最近地去接触那些数据，接下来的工作才是开始进行分析探索。要相信分析专家和数据科学家们都会很好地做好他们应该完成的工作。