

大数据和大分析

段云峰 编著

15年大数据系统建设、运营和应用的经验，
国内大型数据分析系统的全景解剖和案例分享！



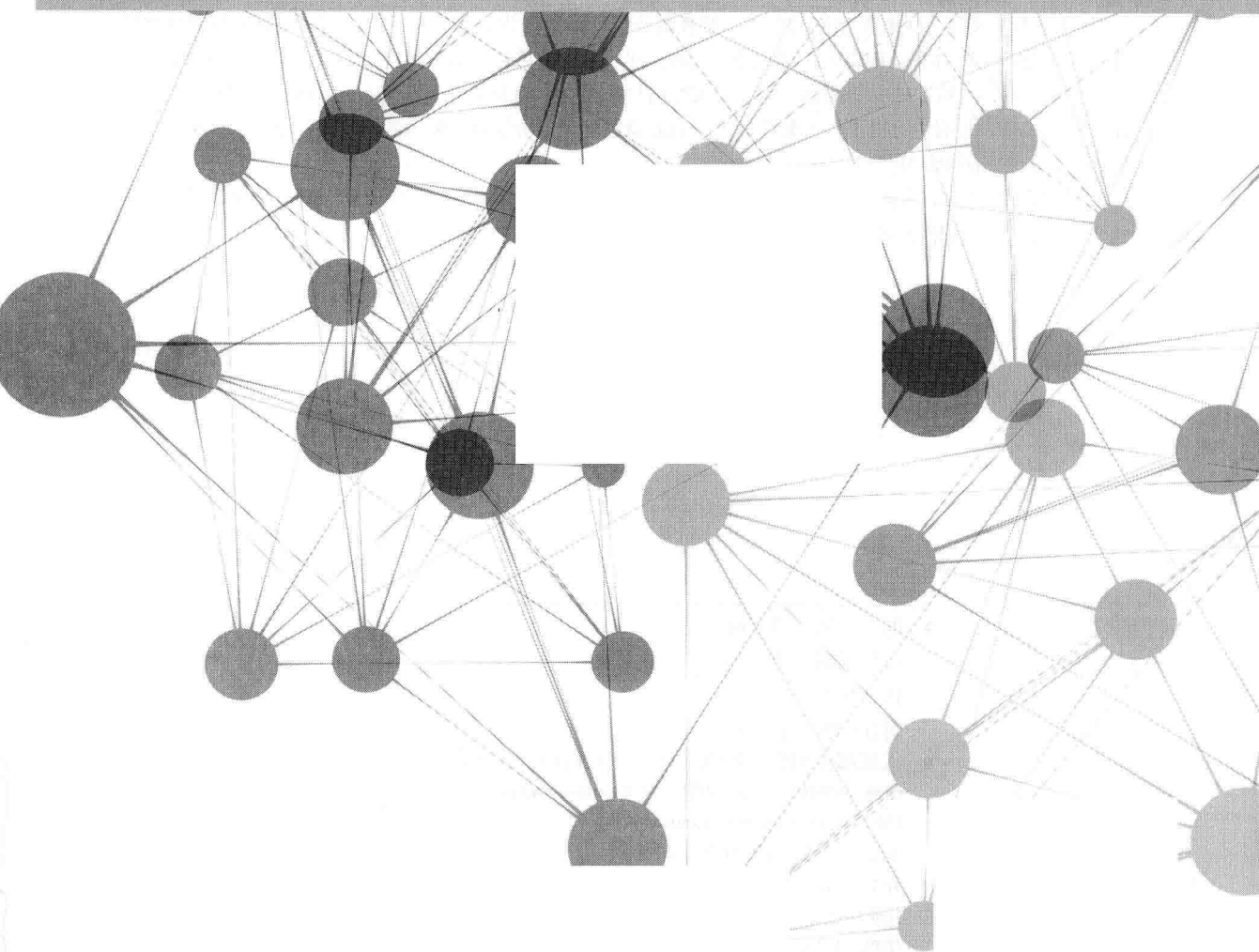
中国工信出版集团



人民邮电出版社
POSTS & TELECOM PRESS

大数据和大分析

段云峰 编著



人民邮电出版社

北京

图书在版编目 (C I P) 数据

大数据和大分析 / 段云峰编著. — 北京 : 人民邮电出版社, 2015. 10
ISBN 978-7-115-40259-2

I. ①大… II. ①段… III. ①数据处理 IV.
①TP274

中国版本图书馆CIP数据核字(2015)第196806号

内 容 提 要

本书首先阐述了大数据出现的背景,解释数据资产、数据驱动等基本概念,剖析数据分析的重要性,介绍了大分析的内容和范围以及关键点等;其次,阐述了大数据建设的基本内容、有关应用领域等,涉及收集、存储、标准、技术选择等内容;辨析了大数据与数据仓库的关系,重点介绍了数据质量、安全管理等数据管控的内容;最后,给出了与大数据建设配套的营销管理分析等方面的内容和经验,分享了如何引入互联网思维、开辟新视野的理念。

本书适合电信、金融、互联网等各个行业的大数据相关从业者,包括企业管理者、开发工程师、系统建设者、业务应用者和运营人员参考阅读。



-
- ◆ 编 著 段云峰
责任编辑 李 静
执行编辑 乔永真
责任印制 彭志环
 - ◆ 人民邮电出版社出版发行 北京市丰台区成寿寺路 11 号
邮编 100164 电子邮件 315@ptpress.com.cn
网址 <http://www.ptpress.com.cn>
北京市艺辉印刷有限公司印刷
 - ◆ 开本: 787×1092 1/16
印张: 19.5
字数: 437 字
-

定价: 68.00 元

读者服务热线: (010)81055488 印装质量热线: (010)81055316
反盗版热线: (010)81055315

前言

最近几年，随着大数据概念的引入和发展，人们逐步意识到大数据是重要的数据资产，能够带来更大的价值。但如何应用这些大数据，挖掘大数据的价值，是所有企业必然面临的问题。

目前，国内大部分企业还处于大数据积累的初级阶段，业界的焦点也聚集在如何借助云计算技术存储大数据。在如何广泛应用这些数据方面，国内依然缺乏经验。为了突出对大数据的应用，作者首次提出了“大分析”的概念，以引起业界对于大数据应用方面的重视。通过各种大分析应用，可升级企业量化管理水平，并最终推动“理性社会”的构筑。

笔者在大数据领域积累了 15 年的经验，实际承担了大型数据仓库的设计、建设和应用推广工作，并将云计算技术引入大数据系统中，在企业精细化管理、精准营销等方面积累了大量的经验。通过本书，笔者分享了在电信和互联网等不同行业的大数据设计、建设和运营的经验。基于多年的实践工作，笔者给出了很多新的观点，列出了大数据系统中各种关键点和注意事项。

大数据需要借鉴数据仓库的管理经验，通过混搭架构，满足大分析的技术支持需求。在对外提供数据服务的同时，也要服务于企业内部的管理活动，服务于一线市场，体现大分析渗透到企业“每个毛孔”的理念。大分析系统与传统 IT 系统的建设有很大的不同，笔者分享了大分析系统建设、应用过程中的各种经验，并分享了大分析在企业客户分析、产品分析、营销分析等内部各个领域应用的真实案例。

在编写本书的过程中，笔者得到了很多领导和同仁的帮助和指导，主要包括魏春晖、刘虹、陶涛、江勇、张航友（四川）、谢志崇（福建）、杨旭（广东）、秦晓飞（山西）、王新印（山东）、易剑光（河北）、汪新勇（广东）、陈刚（华为）、赵懿敏（亚信）等，在此一并表示感谢。

作者

2015 年 8 月

目 录

第1章 背景 //1

1.1 大数据的引出和影响 //3

1.1.1 “大数据，大商机” //3

1.1.2 “数据资产”的引出 //6

1.1.3 数据量庞大 //9

1.1.4 数据结构复杂 //10

1.1.5 数据价值有待挖掘 //11

1.1.6 “数据驱动”的变革 //12

1.1.7 互联网发展中的“数联网” //15

1.2 为何需要大分析 //16

1.2.1 数据价值评估 //16

1.2.2 “数据资产”变现问题 //18

1.2.3 大分析的技术基础 //20

1.2.4 大分析面临的问题 //26

1.3 大分析的应用案例 //29

1.3.1 新的“啤酒和尿布” //29

1.3.2 KPI信息地图 //30

1.3.3 “大数据、超细分、微营销” //32

1.4 小结 //34

第2章 大数据基础 //35

2.1 大数据的基本理念 //36

2.1.1 概念和定义探索 //36

2.1.2 大数据的技术基础 //37

2.1.3 没有大分析，大数据就是大垃圾 //38

2.1.4 大数据如何借鉴“数据仓库”的经验 //38

2.1.5 企业级数据中心 //41

2.2 大数据与数据仓库的关系 //42

2.2.1 大数据扩展数据仓库理论架构 //42

2.2.2 大数据继承数据仓库数据管理的经验 //43

2.2.3 大数据开启了非结构化数据的处理 //43

2.2.4 大数据要借鉴数据仓库的生态圈 //43

2.2.5 大数据应继承数据分析技术 //44

2.2.6 与数据库的关系 //44

2.2.7 数据仓库借鉴大数据的营销模式 //44

2.3 大数据的基本特点 //45

2.3.1 “4V”特点 //45

2.3.2 大分析角度的大数据特征 //45

2.4 大数据的价值和意义 //46

2.4.1 围绕客户信息，提供全方位服务 //46

2.4.2 构筑“虚拟团队”，提升团队管理水平 //46

2.4.3 让“智慧城市”“智能交通”等变为可能 //47

2.4.4 构筑“理性社会”终于成为可能 //47

2.4.5 中国前所未有的“弯道超车”机遇 //47

2.5 大数据的问题和挑战 //48

2.5.1 数据质量问题越发突出 //48

2.5.2 数据分析技术尚缺实质突破 //48

2.5.3 大数据应用水平需要逐步演进、逐步深化 //48

2.5.4 大数据技术架构面临突破 //49

2.5.5	数据理念与国外仍然相距甚远	//49
2.5.6	大数据是一项系统工程	//49
2.6	小结	//50
第3章	大数据的管理	//51
3.1	数据如何收集	//52
3.1.1	能获取哪些数据	//52
3.1.2	基于数据价值, 决定数据的收集、存放策略	//53
3.1.3	没有应用时, 是否收集数据	//53
3.2	数据的标准	//53
3.2.1	数据接口	//53
3.2.2	数据模型	//55
3.3	大数据的 ETL 过程	//57
3.4	大数据如何存储	//58
3.4.1	数据库 / 数据仓库	//58
3.4.2	分布式文件系统 (HDFS)	//59
3.4.3	混搭模式	//63
3.4.4	Hive / Hbase 等	//63
3.4.5	MPP	//65
3.5	数据如何估值和计费	//65
3.5.1	什么数据最好卖	//66
3.5.2	市场价格	//66
3.5.3	数据的开放	//67
3.6	大数据的“数据资产”管理	//67
3.7	数据如何保障安全	//68
3.8	小结	//71
第4章	大数据的技术架构	//73
4.1	大数据处理架构	//74
4.1.1	大数据处理层级和域	//75
4.1.2	哪些计算适合并行	//78
4.2	为何是混搭架构	//78

- 4.2.1 大数据混搭架构的利弊分析 //79
- 4.2.2 架构是否去 IOE //80
- 4.2.3 大数据混搭架构实例 //80
- 4.3 数据集市的模式 //83
 - 4.3.1 数据沙盒模式 //85
 - 4.3.2 贴近角色的平台及应用 //85
 - 4.3.3 文件集市 //89
- 4.4 数据管控模块 //89
 - 4.4.1 元数据 //89
 - 4.4.2 数据质量 //91
- 4.5 大数据的“爬虫”技术 //93
 - 4.5.1 定制爬虫 Nutch //94
 - 4.5.2 分词技术——庖丁分词 //94
 - 4.5.3 索引及全文检索——Splunk //95
 - 4.5.4 上网数据解析流程 //97
- 4.6 大数据安全管理框架 //99
 - 4.6.1 安全管控技术架构 //99
 - 4.6.2 管理制度建设 //101
 - 4.6.3 去隐私化技术举例 //103
- 4.7 小结 //108
- 第5章 大数据的数据质量管控 //111
 - 5.1 数据质量概念 //113
 - 5.1.1 基本概念 //113
 - 5.1.2 大数据就不考虑质量了吗 //117
 - 5.2 元数据 //118
 - 5.2.1 数据的数据 //118
 - 5.2.2 元数据的 CWM 标准 //120
 - 5.2.3 元数据分类 //122
 - 5.3 数据质量管控 //123

5.3.1	数据质量管控目标	//123
5.3.2	数据质量子系统架构	//124
5.4	如何建立数据质量管理制度	//129
5.4.1	数据质量分工管理流程	//129
5.4.2	实时监控和告警	//136
5.5	数据质量管控产品的客户体验	//139
5.5.1	降低技术门槛	//139
5.5.2	产品的手机 App 化	//140
5.6	小结	//141
第 6 章	大数据如何带来大分析	//143
6.1	没有应用的数据是垃圾数据	//145
6.1.1	应用的广度	//145
6.1.2	应用的深度	//145
6.1.3	应用的实时性 / 融合性	//146
6.2	大分析 (BA) 的概念	//147
6.2.1	“大分析”的发展变化	//147
6.2.2	大分析的“群众路线”	//150
6.3	大分析 (BA) 的产品开发	//151
6.3.1	基于数据分析, 解决实际问题	//152
6.3.2	BA 产品和分析工具产品的区隔	//154
6.3.3	自助分析	//158
6.3.4	导航式分析	//161
6.4	应用推广问题	//170
6.4.1	市场的“冬天”就是大分析的“春天”	//170
6.4.2	为何需要应用推广	//170
6.4.3	如何证明分析的独特价值	//171
6.4.4	如何解决员工实际的问题	//172
6.4.5	电信和互联网行业应用推广对比	//172
6.4.6	不同的企业用不同的推广方法	//172

6.5 大分析的“闭环”问题 //173

6.5.1 分析和执行的闭环执行 //173

6.5.2 营销流程的设计——营销管理平台 //173

6.6 小结 //176

第7章 大分析应用案例 //179

7.1 大分析的应用阶段 //180

7.2 非结构化数据的分析 //181

7.2.1 客户投诉内容分析 //181

7.2.2 互联网舆情监控分析 //183

7.3 客户分析 //184

7.3.1 客户生命周期 //184

7.3.2 客户内容属性 //197

7.3.3 客户行为属性 //203

7.3.4 客户新业务分析 //218

7.3.5 客户满意度分析 //220

7.3.6 客户流失分析 //223

7.4 产品分析 //234

7.4.1 建设背景和目的 //235

7.4.2 整体流程 //235

7.4.3 建设中涉及的模型 //235

7.5 内容营销分析 //240

7.5.1 手机视频产品营销触发 //240

7.5.2 手机视频内容分析 //240

7.5.3 目标客户筛选和细分 //240

7.5.4 方案设计 //241

7.5.5 营销脚本设计 //242

7.5.6 营销方式选择 //243

7.5.7 营销方式使用效果 //243

7.5.8 效益评估 //244

7.6	网格化管理分析	//245
7.7	社会渠道欺诈分析	//247
7.7.1	概述	//247
7.7.2	模型方法	//248
7.7.3	模型定义	//250
7.7.4	业务应用	//255
7.7.5	优化方法	//256
第8章	大数据中的互联网思维	//257
8.1	互联网思维介绍	//258
8.1.1	九大特征	//258
8.1.2	大数据为何需要互联网思维	//260
8.1.3	大数据如何借助互联网思维	//260
8.2	BA 产品定义	//260
8.3	BA 产品的客户体验	//263
8.3.1	客户是谁	//263
8.3.2	客户的体验是什么	//264
8.3.3	提升客户体验的展示形式	//266
8.3.4	提升客户体验的解决问题能力	//270
8.4	BA 产品迭代开发	//280
8.4.1	如何构建共享方式	//280
8.4.2	提供 API 接口	//282
8.4.3	提供类似 App Store 开发环境	//284
8.5	BA 产品的“客户入口”把控	//286
8.5.1	抓住客户的入口	//286
8.5.2	让客户参与 BA 产品开发	//287
8.5.3	BA 产品的内部客户营销	//287
第9章	大数据的管理架构及探索	//289
9.1	BA 产品生态圈的建设	//290
9.1.1	生态圈组成	//290

9.1.2 生态圈盈利模式 //292

9.1.3 生态圈的“共赢” //293

9.2 管理架构举例 //294

9.2.1 互联网企业架构 //294

9.2.2 运营公司的架构 //295

9.2.3 架构的特点分析 //295

9.3 人才的培养 //296

9.3.1 知识结构要求 //296

9.3.2 交际（团队）能力要求 //297

9.3.3 耐压能力要求 //297

9.4 团队的构建及激励 //298

9.4.1 技术人员转型业务人员 //298

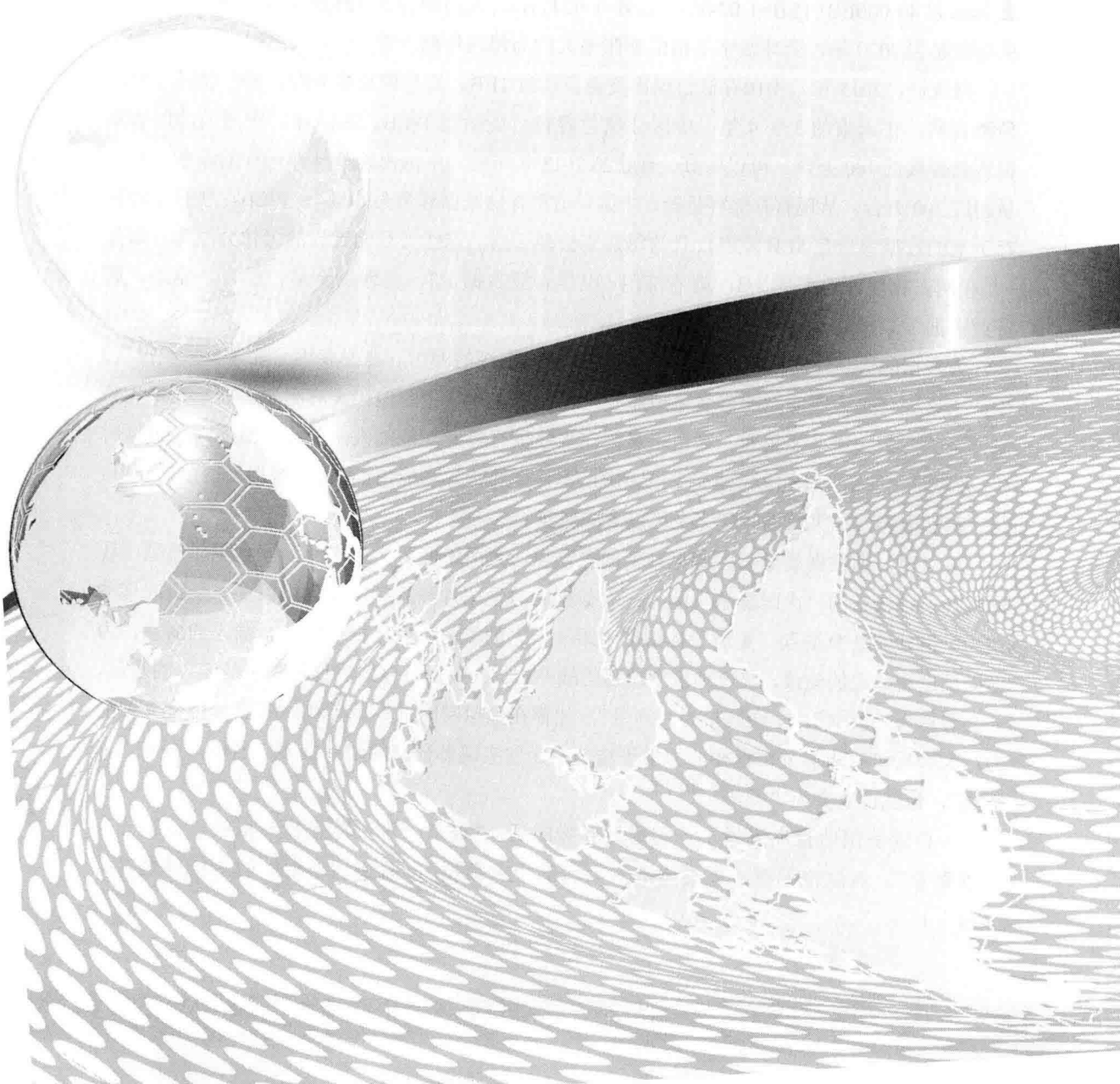
9.4.2 业务经验的培养和积累 //298

9.4.3 待遇激励 //298

第10章 后记 //299

第1章

背景



IDC 和 EMC 联合发布的《2020 年的数字宇宙》报告预测：到 2020 年，全球数字宇宙将会膨胀到 40 000EB (1EB=1 024PB, 1PB=1 024TB)，人均约为 5 200GB 以上。该报告指出，从现在起到 2020 年，全球数字宇宙的膨胀率大约为每两年翻一番。

据统计，2013 年，中国存储市场出货容量超过 1EB。在存储总量方面，IDC 曾经发布的预测表明，在未来的 3 ~ 4 年，中国存储总容量可能达到 18EB。2013 年，中国内地服务器销售总数接近 100 万台。可以估算，截止到 2013 年年底，中国内地整体在运行的服务器总数量超过 300 万台。从现有存储容量看，中国目前可存储数据容量为 8EB ~ 10EB，现有可以保存下来的数据容量在 5EB 左右，且约每两年会翻一倍。这些被存储数据的大体分布为：媒体 / 互联网占据现有容量的 1/3，政府部门 / 电信企业占据 1/3，其他的金融、教育、制造、服务业占据 1/3。

大数据的浪潮扑面而来，没有人可以躲过这次浪潮的冲击。在好莱坞大片《少数派报告》中，未来警局以“罪前”的罪名逮捕罪犯，也就是在即将犯罪但还没有犯罪之前，先把人抓来，至于怎么知道谁要犯罪，则由三名躺在水池里有特异功能的人来判定。然而在大数据时代，要搜寻“罪前”犯可以说是轻而易举的，不再需要特异功能，而是靠大数据模型分析即可。这个科幻电影，在大数据时代完全可以变成现实。

可以说，大数据就像《西游记》里的孙悟空，无所不能！大数据通过预测犯罪模型，让你具备了孙悟空的“火眼金睛”能力；大数据的海量图片和视频，让你可以在家里瞬间移动到几千里外的某个景点，实现了孙悟空“跟斗云”的能力；大数据可以分析敌人的弱点，为你提供进攻敌人的建议，实现孙悟空超强的战斗力；大数据可以给出你各种“伪装”的建议，一会装扮为“美女”，一会装扮为“帅哥”，让你在虚拟网络世界里实现孙悟空“72 变”的能力；大数据可以为你分析棍棒的力学等技巧，让你实现孙悟空的超级棍术。所以，有了“大数据”，你就可以变为“孙悟空”！

一位前美国情报高层说了一句耐人寻味的话：“要从一堆稻草里找一根针，你得先有一堆稻草”。而这堆稻草，就是本书所要讲的“大数据”，而找到那根针的方法就是“大分析”！

1.1 大数据的引出和影响

过去的几十年里，为何在互联网出现了这么多年之后，会出现大数据的概念呢？原因很简单，互联网生产了“前所未有”的海量数据，很多还是非结构化的，而云计算使得对这些海量数据的处理成为可能。在当今时代，每个人每天都在无时无刻地生产着数据，打电话、浏览网页、购物、发微信等，都会产生海量的数据。美国互联网数据中心指出，互联网上的数据每年将增长 50%，每两年就将翻一番，而目前世界上 90% 以上的数据是最近几年才产生的。那么为什么会引出“大数据”的概念呢？

首先，有了海量大数据。据 IDC 的分析，2012 年的数字化内容增长到 2.7ZB，较 2011 年增长 48%，至 2015 年，数字化内容将以火箭速度逼近 8ZB。在大数据方面，IDC 预测，超过 90% 的数据将是非结构化数据（例如图像、视频、MP3 音乐文件，以及其他基于社交媒体的文件和在 Web 上进行的工作）。这解决了之前“巧妇难为无米之炊”的缺少数据的状况。

其次，海量数据可以存储了。因为数据存储成本大为降低。在 1980 年的时候，要储存一个高清电影，它的成本大概相当于 100 万美金。而到 2010 年的时候，它的成本相当于 5 美分，而到今天有可能只要 2 美分。

第三个原因是大批消费者，尤其是年青人已经离不开互联网了，其工作、学习、娱乐、消费都会在互联网上产生“痕迹数据”，这些痕迹数据对于各个领域都具有重要意义。政府可以分析出市民的出行规律，用于智能交通服务；学者可以瞬间扩展自己的学生圈，推介自己的观点；电商可以分析出客户的喜好，及时推荐客户需要的产品；交友网站可以分析客户的朋友圈信息，进行广告营销；门户网站可以监控舆论热点，吸引眼球。

所以，当人们在网络上“生活”时，网络里就存储了海量的数据，配合云计算的处理能力，这些数据能够带来巨大的商业价值。这些因素驱动了“大数据”概念的产生和兴起。

1.1.1 “大数据，大商机”

1. 大数据带来的大商机

大数据带来了很多前所未有的商机。例如，腾讯的易迅网电商平台会辨析每个登录客户的身份，分析其曾经购买的商品，判断客户的喜好，并据此分析推荐客户登录页面的广告位内容，提高客户的点击率。而每个成功的点击，都会给腾讯带来广告收入，腾讯 2014 年在广告方面的收入有 90 亿美元。由此可见，大数据分析在提升客户广告点击率方面的商业价值。

大数据带来的商机不仅体现在广告领域。在金融行业，通过收集客户的数据并进行分析，能够对客户的“征信”程度（即客户信誉度）进行评估。

这种“征信”评估不仅可以降低贷款风险，而且可以用于交友等多方面。例如：阿里巴巴基于互联网的芝麻信用数据，涵盖了信用卡还款、网购、转账、理财、水电气缴费、租房信息、住址搬迁历史、社交关系等各方面的信息。芝麻分数为 350 ~ 950 分，分值越高，信用越好。用支付宝钱包，客户除了能看到自己的信用评分外，还能看到信用历史、行为偏好、履约能力、身份特质、人脉关系等 5 个维度的信息。芝麻分数越高，在租车、住酒店时可以不用再交押金，网购时可以先试后买，办理出国签证时不用再办存款证明，贷款时可以更快地得到批复，甚至可以拿到比别人更低的利率，由此产生的商业机会前途广阔。

在教育领域，也可以实现一直所倡导的“因材施教”。基于学生的日常行为数据，可以分析学生的心理特点，了解学生的喜好，判断学生喜欢的学习模式，从而适配合适的教育方案。例如：通过分析大学生“小明”的上网数据，可以看出其关心“大数据学习”内容；通过分析其在朋友圈的交流内容，可以判断其性格特点（比较内向、严谨）等；通过选课记录，可以判断其喜欢的讲师及类型（通俗易懂型还是学术研究型等）。由此可以探索适合不同性格特点，选择不同教程的教育模式。例如让其选择网络上最流行的涂子沛大数据教材和教学视频。甚至在小明的大学教育阶段，都可以基于大数据分析进行远程教育，制定上课计划，网上寻找教材，网上完成课程学分、论文编写等工作。这就撬开了庞大的网络“个性化教育”的金矿。与以往不同的是“因材施教”，是基于学生特点、独家制定个性化的教育培训方案。

在大数据时代，工业设备、汽车、电表、水表、街头摄像头等每天都能产生海量数据，而这些数据并没有被纳入传统的数据处理、数据分析领域。这些新增领域产生的数据源，同样可以产生很高的应用价值，催生丰富的商业应用模式。

所以，在互联网大数据时代，很多传统的商业模式都将受到新的洗礼，大数据提供了新的商机，让一切变得“皆有可能”！

企业数据犹如一座“金矿”，蕴藏着大量的价值。国内互联网公司已经率先行动开始改变自身的商业模式，通过自身与外部数据聚合，实现大数据“变现”（转变为“现金”）模式。

互联网主要企业大数据“变现”对比如表 1-1 所示。

表 1-1 互联网企业大数据应用对比

企业名称	大数据资源	大数据战略	大数据“变现模式”
百度	互联网入口数据；即时需求数据；公共网页数据	在后台提供云计算能力，加上百度这些年积累的丰富的中文数据与搜索需求，支撑开发者研发、发布各种 App	百度广告联盟；百度指数；百度定制报告
阿里巴巴	整个互联网价值最高的电商流量数据；交易数据；信用数据；社交数据（微博和陌陌）	“数据分享平台”的战略，为客户提供多元化或个性化服务，通过这些不同类型的服务平台，来完成自身对于大数据的收集以及完善，实现“跨界”合作创新的目的	淘宝广告交易平台（Taobao Ad Network & Exchange, Tanx）；淘宝广告联盟；小微企业金融服务；数据交易集市
腾讯	腾讯拥有大量基于客户明确 ID 的行为数据：社交数据，交易数据	“强化大社交网络”，公布了其社会化营销平台，宣称揭开了“大数据”转向广告层面变现的序幕	智能推荐；后端数据整合统一向前台开放；游戏广告