Mianxiang Dashuju Yingyong
De Shuju Caiji Jishu Yanjiu

面向大数据应用的数据采集技术研究

韦鹏程 颜蓓 陈美成 著

中国原子能出版社
China Atomic Energy Press

面向大数据应用的数据采集技术研究

韦鹏程　颜　蓓　陈美成　著

中国原子能出版社
China Atomic Energy Press

图书在版编目（CIP）数据

面向大数据应用的数据采集技术研究 / 韦鹏程，颜蓓，陈美成著. -- 北京 : 中国原子能出版社，2019.12
ISBN 978-7-5221-0364-8

Ⅰ. ①面… Ⅱ. ①韦… ②颜… ③陈… Ⅲ. ①数据采集－研究 Ⅳ. ①TP274

中国版本图书馆CIP数据核字（2019）第288542号

内容简介

本书属于大数据采集方面的著作，主要由大数据概念、大数据采集重要性、大数据采集的特点和相关技术、大数据采集架构、大数据迁移技术、大数据采集实例等部分组成，全书以大数据采集为研究对象，对面向大数据应用的数据采集技术从数据采集、数据预处理、数据分析及数据可视化的综合应用作出分析，并提出大数据采集技术的相关要点。对大数据采集研究者以及从事大数据相关工作的工程技术人员具有学习和参考价值。

面向大数据应用的数据采集技术研究

出版发行　中国原子能出版社（北京市海淀区阜成路43号　100048）
责任编辑　高树超
装帧设计　河北优盛文化传播有限公司
责任校对　冯莲凤
责任印制　潘玉玲
印　　刷　三河市华晨印务有限公司
开　　本　710 mm×1000 mm　1/16
印　　张　17
字　　数　320千字
版　　次　2019年12月第1版　　2019年12月第1次印刷
书　　号　ISBN 978-7-5221-0364-8
定　　价　67.00元

发行电话：010-68452845　版权所有　侵权必究

此为试读，需要完整PDF请访问：www.ertongbook.com

前 言

数据是和土地、资本、人力资源齐头并进的主要生产要素，其在社会、经济、科学研究等方面颠覆了人们探索世界的方法，驱动了产业间的融合与分立。

大数据是用来描述数据规模巨大、数据类型复杂的数据集的，它本身蕴含着丰富的价值。例如，金融行业将企业和个人的一些信用记录、消费记录、客户点击数据集、客户刷卡、存取款、电子银行转账、微信评论等行为数据组合为金融大数据，他们利用大数据技术为客户推荐产品，利用客户行为数据设计满足客户需求的金融产品，利用金融行业全局数据了解业务运营的薄弱点并加快内部数据的处理速度，利用决策树技术进行抵押贷款管理，利用数据分析报告实施产业信贷风险控制，利用客户社交行为记录实施信用卡反欺诈，依据客户消费习惯、地理位置、销售时间进行推荐（精准营销）。不仅仅金融行业，政府部门也会根据大数据分析结果来做预算，企业也会根据大数据进行市场策略调整。

回想 20 世纪科学技术的发展，计算机是对人类经济建设和生活最有影响力的发明。尤其是自 20 世纪 70 年代以来，微处理器的问世促使微型计算机技术迅速发展和应用，在世界范围内掀起了一场新的技术革命。作为微型计算机应用技术的一个重要分支——大数据采集集传感器、信号采集与转换、计算机等技术于一体，是获取信息的重要工具和手段。随着微型计算机的应用与普及，它在科学研究、生产过程等领域中发挥着越来越重要的作用。在科学研究中应用大数据与数据采集，将提高人们对各种瞬态现象进行研究的能力；在生产过程中应用大数据与数据采集，将能迅速地对各种工艺参数进行采集，为计算机控制提供必需的信息，从而实现生产过程的自动控制。因此，大数据与数据采集是机电一体化、智能化仪器仪表、自动控制、计算机应用、机械设计制造及其自动化、农业机械化与自动化等专业的学生和相关专业的工程技术人员必备的专业知识。

该专著由重庆第二师范学院数学与信息工程学院韦鹏程、颜蓓、陈美成三位教师共同完成，并得到儿童大数据重庆市工程实验室、交互式教育电子重庆市工程技术研究中心、重庆市计算机科学与技术重点学科、重庆市计算机科学与技术一流专业、重庆市高校“儿童教育大数据分析关键技术及其应用研究”创新研究群体、重庆市教育委员会科学技术研究计划重点项目资助（NO.KJZD-K201801601）和教育部学校规

划建设发展中心重庆第二师范学院儿童研究院课题项目（CRIKT201902）的支持！

本书主要讲述大数据与数据采集的基本理论、基本概念，数据采集器件的工作原理、性能和使用，数据采集系统硬件和软件的设计方法，目的是帮助读者在实际应用中能正确、合理地设计数据采集系统。

本书有三个主要特点。

1. 系统性

本书对大数据与数据采集技术从整体上进行论述，既讲述大数据与数据采集技术的基本理论、基本概念，又讲述工程上的应用；既涉及硬件设计的知识，又涉及软件设计的知识。

2. 实用性

本书写作的指导思想是以实用为前提，将理论与应用紧密地结合起来；在语言描述上力求简明扼要、通俗易懂；在内容组织上注意知识的完整性，突出重点，并提供了大量的插图和表格。另外，书中还附有大量的应用实例和程序。其中，大部分系作者多年来科研工作的经验总结，并在实际工作中得到了应用和验证，可供读者在开发数据采集系统时参考引用，相信对读者会有很大的帮助。

3. 要点清晰

本书强调基本理论、基本概念，突出软件与硬件的结合，着重介绍设计方法，加强实际应用。作者在写作过程中注意将国内外的新技术、新原理和新方法融入本书。

本专著是由重庆第二师范学院数学与信息工程学院韦鹏程、颜蓓、陈美成三位教师共同完成，并得到儿童大数据重庆市工程实验室、交互式教育电子重庆市工程技术研究中心、重庆市计算机科学与技术重点学科、重庆市计算机科学与技术一流专业、重庆市高校“儿童教育大数据分析关键技术及其应用研究”创新研究群体、重庆市教育委员会科学技术研究计划重点项目资助（KJZD-K201801601）和教育部学校规划建设发展中心重庆第二师范学院儿童研究院课题项目（CRIKT201902）的支持！

目 录

第一章　大数据时代与数据采集

第一节　什么是大数据

大数据不是一项单一的技术，而是一个概念，是一套技术，是一个生态圈。大数据技术的专业术语多达几十个，记录了大数据从炒作到成熟并进入主流应用的过程。数据科学家、预测分析、开放政府数据，都属于大数据的范畴。政府和企业希望从自己的数据中获得更多的信息，软件厂商希望将“大数据解决方案”融入公司的产品之中。在大数据软件公司的助推下，政府和企业已经有能力利用廉价的服务器、开源技术和云计算来进行开销不大的大数据部署了。

对于什么是大数据，不同的研究机构从不同的角度给出了不同的定义。Gartner认为：“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。”麦肯锡认为：“大数据指的是大小超出常规的数据库工具获取、存储、管理和分析能力的数据集。”但它同时强调：“并不是说一定要超过特定 TB 值的数据集才能算是大数据。”维基百科的定义为：“大数据是指无法在可承受的时间范围内用常规软件工具进行捕捉、管理和处理的数据集合。”IDG 认为：“大数据一般会涉及 2 种或 2 种以上数据形式，它要收集超过 100 TB 的数据，并且是高速实时数据流；或者是从小数据开始，但数据每年会增长 60% 以上。”

从客户的角度来看，大数据技术的战略意义不在于拥有多么庞大的数据信息，而在于对这些有意义的大数据进行专业化处理，从中获得商业价值。例如，以色列已经把所有政府部门的视频都整合到一个大数据管理平台上，并在这个平台上开发了一套智慧安防系统。在这个系统上，只要把某一个人的脸或人的主要特征数据输入系统，就能从海量的监控记录中查出同那个人相关的视频片段，并自动变成一个有时间顺序的视频片段。

随着以云计算、大数据、物联网等为代表的新一代信息技术的发展和应用，世界经济进入了大转型时代，主要发达国家以及国内发达省市都紧盯、紧跟这一轮产业变革，试图抢占未来经济发展的先机。大数据是一种产业，这种产业实现盈利的关键在于提高对数据的“加工能力”，通过“加工”实现数据的“增值”，完成“数据变现”。这种加工能力体现在技术上就是大数据分析。简言之，从各种类型的数据中快速获得有价值信息的能力，就是大数据技术。大数据最核心的技术就在于对海量数据进行采集、存储、管理和分析。

第二节　大数据的特征

大数据具有“4V”特征，即Volume（数据体量大）、Variety（数据类型繁多）、Velocity（数据产生的速度快）、Value（数据价值密度低）。

Volume指的是数据体量巨大。例如，一家三甲医院的影像数据（包括CT、B超、X光片、胃镜、肠镜等）可能就是几百个TB，全国的医疗影像数据超过PB级别，接近EB级别。全球数据已进入ZB时代，互联网数据中心（Internet Data Center，简称IDC）预计2020年全球数据量为40 ZB。

Variety指的是数据类型繁多。数据可分为结构化数据、半结构化数据和非结构化数据。结构化数据，即行数据，存储在数据库里，可以用二维表结构来逻辑表达数据，如企业财务系统、医疗HIS数据库、环境监测数据、政府行政审批等。非结构化数据，一般存储在文件系统上，如视频、音频、图片、图像、文档、文本等，典型案例有医疗影像系统、教育视频点播、公安视频监控、国土GIS、广电多媒体资源管理系统等应用。半结构化数据是介于完全结构化数据（如关系型数据库、面向对象数据库中的数据）和完全无结构的数据（如声音、图像等文件）之间的数据，如邮件、HTML、报表等，典型场景如邮件系统、教学资源库、档案系统等。非结构化与半结构化数据的增长速率大于结构化数据，超过80%的数据是非结构化数据。IDC的报告显示，目前在大数据的1.8×10^{12} GB容量中，非结构化数据占80%～90%，并且到2020年将以44倍的发展速度增加。非结构化数据比例不断升高，这些数据中蕴含着巨大的价值。

Velocity是指大数据往往以数据流的形式动态、快速地产生，具有很强的时效性。数据自身的状态与价值也往往随时空变化而发生演变（这些数据往往包括空间

此为试读，需要完整PDF请访问：www.ertongbook.com

维、时间维等多种数据），如环境监测中的水质和空气质量数据、高速路卡口的视频监测数据等。

Value 是指数据已经成为一类新型资产，蕴藏着巨大价值。大数据的价值密度低，需要通过专业的技术手段进行挖掘。只有对其进行正确、准确的分析，才会带来很高的价值回报。例如，从电视机顶盒的频道切换数据、各大电视台的分析数据中可以准确判断观众的喜好，以推出更加符合观众口味的节目。

并非总是有数百个 TB 才算得上是大数据。根据实际使用情况，有时候数百个 GB 的数据也可称为大数据，这主要看它的其他维度，也就是速度或者时间维度。假如能在 1 s 之内分析处理 300 GB 的数据，而通常情况下却需要花费 1 h，那么这种巨大变化所带来的结果就会极大地增加价值。所谓大数据技术，就是至少实现这四个判据（特征）中的几个。

第三节　大数据分析

大数据平台可以存储所有类型的数据，包括从简单的文件存储到不强调一致性的非关系型数据库存储。得益于自身基础设计理念，大数据平台可以无限扩展。如果大数据平台在云端运行维护，那么它的灵活性将更强。从概念上讲，存储数据是大数据应用中最易于实现的部分。

当大数据平台上存储了足够多的数据后，我们该怎样对其加以利用呢？分析大数据并将分析结果应用于决策中才是最重要的事情。预测分析（predictive analytics）是大数据分析领域中的一个常用模式，它通过分析采集的数据预测未来的行为或趋势。它根据事物的过去和现在估计未来，根据已知预测未知，从而减少对未来事物认识的不确定性，以便指导我们的决策行动，减少决策的盲目性。在大数据分析领域，预测分析常常与预测模型、机器学习和数据挖掘有关。对于一个政府部门而言，通过预测分析可以精准把握政府工作的重点。例如，云升科技帮助湖州市公安局分析来自各个渠道的海量群众诉求，预测下个月的警务工作热点，从而帮助湖州市公安局合理安排警力，最终实现民意引领警务。美国的医疗决策支持系统基于预测分析判断某些人得某些疾病的风险，并基于当前的健康状态给出最正确的医疗决定。国内的很多金融企业通过预测分析实现业务的风险控制。例如，某银行通过分析其客户的消费数据和基本数据，从而预测该客户的信用卡和贷款的偿还能力。环保部门用数据决策，利

用环保大数据综合研判，制定环境政策措施，预警环境风险，提高环境综合治理科学化水平。

除了预测分析，还需要关联分析。关联分析的目的在于找出数据之间内在的联系。例如，购物篮分析，即消费者常常会同时购买哪些产品（如游泳裤、防晒霜），从而有助于商家进行捆绑销售。

第四节　大数据国家战略

斯诺登效应导致政府和企业对信息产品的安全性非常关注。政府和大型企业已经将关注的重点转移到开源软件上，因为开源软件被认为是更加透明和安全的。这是以 Hadoop 为代表的开源软件在国内发光发热的大机遇。另外，2015 年 8 月 19 日，国务院常务会议通过了《关于促进大数据发展的行动纲要》（以下简称《行动纲要》）。会议强调，一要推动政府信息系统和公共数据互联共享，消除信息孤岛，加快整合各类政府信息平台，避免重复建设和数据打架；二要顺应潮流引导支持大数据产业发展，以企业为主体、以市场为导向，加大政策支持，着力营造宽松公平的环境，建立市场化应用机制，深化大数据在各行业的创新应用，催生新业态、新模式，形成与需求紧密结合的大数据产品体系，使开放的大数据成为促进创业创新的新动力；三要强化信息安全保障，完善产业标准体系，依法依规打击数据滥用、侵犯隐私等行为，让各类主体公平分享大数据带来的技术、制度和创新红利。这个《行动纲要》明确指出了推动政府大数据开放、共享和安全的重要性。

2016 年 3 月 17 日发布的《国民经济和社会发展第十三个五年规划纲要》中指出，实施国家大数据战略，把大数据作为基础性战略资源，全面实施促进大数据发展行动，加快推动数据资源共享开放和开发应用，助力产业转型升级和社会治理创新。加快政府数据开放共享，全面推进重点领域大数据高效采集、有效整合，深化政府数据和社会数据关联分析、融合利用，提高宏观调控、市场监管、社会治理和公共服务精准性和有效性，依托政府数据统一共享交换平台，加快推进跨部门数据资源共享共用。加快建设国家政府数据统一开放平台，推动政府信息系统和公共数据互联开放共享。制定政府数据共享开放目录，依法推进数据资源向社会开放。统筹布局建设国家大数据平台、数据中心等基础设施。

一、政府大数据的价值

对大数据企业来说，目前遇到的发展机遇已经十分清楚。《行动纲要》的三个关键词的出发点和落脚点都指向政府大数据，即关于政府大数据的开放共享、关于政府大数据的研究与应用、关于政府大数据的示范效应。那么我们不禁要问，政府大数据的价值究竟何在？为什么政府大数据更有价值？这就要从数量和质量两个层面说起。

就数量而言，或许有人会问，政府数据的数量能比得过 BAT 吗？表面上看，百度、阿里和腾讯都分别拥有数以亿计的用户量，但这与政府大数据相比，不是一个量级，可谓小巫见大巫。 阿里巴巴的数据容量在 100 PB 左右，而仅一个北京市政府就拥有几百个 PB 的数据容量，相当于几个阿里巴巴。这还仅仅是一个北京市政府。中国有几百个城市、几千个行政县。当前，中央和省级政务部门的电子政务覆盖率已经达到 70%。粗略估算，全国政府大数据加起来至少也该有数百甚至上千个阿里巴巴数据的容量。

至于政府大数据的质量，我们也可以通过和 BAT 对比来说明。比如，百度拥有庞大的用户搜索记录，但这些数据较为单一，不进行关联应用毫无价值；腾讯的优势在于拥有数亿的 QQ 和微信用户量以及更庞大的社交数据，但这些数据目前仅局限于营销应用；阿里的交易数据似乎价值更高，但也只是局限在电商领域以及外延应用。换句话说，BAT 这三家企业的短板共同点在于数据种类的单一化程度较高。政府大数据不同，它涉及工商、税务、司法、交通、医疗、教育、通信、金融、地理、气象、房产、保险、农业、环境等领域，数据的种类繁多、关联性强、统计规格较为统一，便于应用处理。政府的数据事关百姓生活的方方面面，数据的利用价值也最高。

二、政府大数据的应用场景

各地政府都非常关注政府大数据。例如，为推进浙江省信息资源的整合开放和大数据产业发展，浙江省政府在 2015 年成立了浙江省数据管理中心。浙江省数据管理中心的职责是拟定并组织实施大数据发展规划和政策措施；研究制定数据资源采集、应用、共享等标准规范；统筹推进大数据基础设施建设、管理；组织协调大数据资源归集整合、共享开放，推进大数据应用；组织协调大数据信息安全保障体系建设。浙江省政府成立大数据发展领导小组，负责大数据发展的政策制定及相关整合工作、数据开放共享等顶层设计。

我国对与政务相关的政府大数据产业需求是非常明显的。由于全国各地区信息化

发展的水平差异较大，政务信息化建设也存在着明显的区域差异性特征。《行动纲要》提出，“要推动政府信息系统和公共数据互联共享，消除信息孤岛，加快整合各类政府信息平台，避免重复建设和数据打架”，这主要就是针对政务数据的开放共享平台而言的。很多地市正在兴建公共信息服务平台，对全市政府资源数据进行集中存储和统一管理。一些省市已经设立了大数据管理局。

除了上述的智慧政务外，我国政府大数据还主要应用于以下领域：智慧城市、公共服务、医疗、教育、交通、环境保护、能源等。这些领域大多涉及国民生活和城镇化进程。截至 2016 年初，我国的智慧城市试点已达 193 个，而公开宣布建设智慧城市的城市超过 400 个，投资总规模高达 5 000 亿元。智慧城市的概念包含了智慧政务、智慧能源、智慧交通、智慧医疗、智慧环保等多领域的应用，而这些都要依托大数据。大数据产业是“智慧”的源泉，是智慧城市的推手。

有专家曾将目前政府大数据的发展现状看作城市建设自来水管道系统时期。因为每个城市只建一套自来水供水系统，不可能建第二套，所以快速布局政府大数据云平台和大数据管理平台是大数据基础建设的第一步，也是赢得政府大数据市场的第一步。至于自来水（平台上所管理的数据）问题，那就是基础建设后顺理成章的事了。还有，大数据基础建设是大数据行业发展的前期环节，等到基础环节铺设完善，自来水得以顺畅流通了，政府大数据的价值才真正爆发出来，这就进入大数据商业应用的时期了。大数据管道建设涉及设计和技术等十分专业的工作，政府的策略是请专业公司提供大数据管理平台，并以此作为大数据管道和基石。

第五节　大数据产业链分析

大数据不仅是一个热门词汇，而且代表着一个欣欣向荣的产业。《行动纲要》从国家大数据发展战略全局的高度提出了我国大数据发展的顶层设计，将大数据定为驱动经济增长和社会进步的重要国家战略基础资源。我国的大数据产业是一片广阔的蓝海，从与每个人密不可分的健康医疗到金融个人征信，其正以飞快的速度融入社会经济发展的方方面面。据分析预测，大数据应用将在国内十多个领域有很大的发展，涵盖万亿市场。最精练的总结正如马云所提出来的——“未来最重要的能源不是石油，而是数据”。并且，大数据产业的核心是推动数据资源的共享与开放，单一的公司是很难发展的，因为在数据领域中单一公司所需要获得的数据要由大量的公司提供。只

有大数据资源开放共享，才能更好地推动大众创业、万众创新。此外，作为云计算、大数据基础的数据中心耗能巨大，绿色数据中心技术将成为全球数据产业的生命线。

全国各地都在布局大数据产业。例如，2014 年武汉市政府出台《武汉市大数据产业发展行动计划（2014—2018 年）》，通过构建“2+7+N”的大数据产业发展格局，以“中国·武汉光谷”为核心，全面推进武汉市大数据产业发展战略，重点发展左岭大数据产业园等多个产业基地，形成丰富的大数据资源聚集地和完善的产业链，建成国内领先、国际知名的大数据产业和数据资源聚集“洼地”。到 2018 年，实现武汉市大数据产业产值规模 2 000 亿元，带动相关产业新增销售收入过万亿元。大数据成为武汉市经济社会发展的新引擎。

一、技术分析

从技术实施的层面上，我们把整个大数据产业链（或大数据市场）分为以下四个层面，如表 1-1 所示。

表 1-1　大数据产业的四个层面

序　号	内　容
1	大数据应用（政府、金融、运营商、互联网等）、大数据交易、大数据运营
2	大数据分析工具（数据处理、数据挖掘、可视化、模型预测）
3	基础软件平台（数据采集、内容管理、数据库）
4	基础设施（计算、存储和网格）

由表 1-1 可以看出，最底层是同硬件相关的基础设施层；最上层是同行业相关的大数据应用层，它需要行业的专业知识，使用大数据技术来实施。在有些研究机构中，中间的两层被认为是一层。我们认为，基础软件平台完成数据的汇聚，形成企业的大数据管理层（国外也有人把它叫做数据湖，Data Lake）。在实施了这一层之后，企业或政府单位的数据已经在一个统一的平台上了，完成了“数据即服务”的基础平台，实现了全域的数据层。这好比生活中做饭，我们已经把油盐酱醋、蔬菜、肉等调料和食材都放在冰箱里了。延续上面的比喻，那么大数据分析工具就是做饭的菜刀、锅、搅拌器等工具。工具的好坏决定了数据处理和挖掘的效率和结果。大数据分析工具市场是一个竞争化的市场，既有一些新创立的小企业的参与，也有一些类似于谷

歌、微软、IBM 等行业龙头的参与。大数据分析工具的需求是否还会一直持续下去，是否会进入一个成熟阶段是值得我们观察的。

从 2015 年下半年开始，随着数据的积累、大数据的逐渐深入，越来越多的行业客户（如银行、政府相关职能部门）了解了大数据的价值，也清楚了大数据适用的边界。我们已经明显感觉市场在逐渐成熟，开始有正规的独立大数据项目招标了，这意味着行业客户已经成熟。随着明确、稳定的需求的出现，整个大数据的商业模式越来越清晰了，大数据市场进入了一个新的阶段。

二、角色分析

整个大数据产业链有四种角色：数据提供商、算法提供商、数据优化提供商和应用提供商。

（一）数据提供商

数据提供商一般都拥有某种入口资源（如运营商、电商等），经过了数年，甚至十数年的积累，形成了在某一领域、某一行业独特的数据资源优势。数据提供商可以将数据提供给第三方使用，从而将资源优势转化成实际的收益。由于分工的细化，数据提供商未必自己去做产业链的其他角色。当然，随着数据成本的日益增加，数据将逐渐汇聚到几家巨头手中，而形成以几家数据巨头为中心，数家各领域、各行业垄断企业为补充的格局。

数据提供商领域依旧处于市场初期。目前没有任何一家数据提供商可以提供所有维度的数据，每家都只拥有部分数据。现在最时髦的各家的“用户画像”也只是盲人摸象，距离真相还有一定的距离。一些大数据企业为在某一个行业占据一定地位，苦练内功，成为该行业的大数据应用服务提供商，从而间接地成为数据提供商（因为数据还是属于购买其应用的客户，所以还是替客户操办业务的数据提供商）。

（二）算法提供商

算法提供商虽然没有数据，但具有丰富的行业经验和背景，可以为客户提供很好的算法服务。目前各个行业都有一些独立的第三方算法服务提供商。算法提供商将会随着行业应用的深化不断地强化自身在行业中的优势，对后来者筑起壁垒。而且随着行业经验的积累，算法提供商是最容易成为应用服务提供商的，也可能会被应用服务提供商所取代，不会以单独的形式存在。随着行业应用的深入，每个行业也逐渐会形成几家独大的格局。由算法提供商演变的应用提供商势必会给后来的单纯算法提供商造成很大的壁垒。所以，单纯的算法提供商在未来几年内可能会逐渐淡出。

（三）数据优化提供商

数据优化提供商也没有数据，它需要从数据提供方购买数据（或者由需求方提供数据），然后按照需求方的要求，对数据进行整理、优化，交付给甲方。至于甲方如何使用，它并不介入。数据优化提供商既没有足够的数据资源，又没有算法提供商强大的算法和行业洞察能力，所以只能做些低附加值的技术劳务输出。虽然数据优化服务提供商低端，但在整个产业链里还不容易被取代。随着产业链的日益成熟、分工的日益细化，数据优化服务提供商可能作为一个环节独立存在，而不是作为数据提供商的一环。这一角色，需要精通各种大数据的模型、算法，也需要了解不同数据的特点，从而可以根据用户的需求，为用户“优化”出符合他们需求的数据。

（四）应用提供商

这一角色又叫解决方案提供商，是离客户最近的一个环节，也是最能体现价值的一个环节。对客户而言，他并不关心大数据到底有多大，数据是否足够优化，算法是否足够科学。他关心的是，能否为他解决实际的问题。从这一点上来看，应用提供商颇似一个系统集成商。它需要根据用户的实际需求，去判断需要准备什么样的数据，需要采用什么样 11 的算法，需要将数据如何优化，以便达到最优的效果，帮助客户解决什么样的实际问题。应用提供商需要清楚地知道哪些是大数据能做到的，哪些是大数据做不到的。大数据不是万能的，他需要懂得约束客户的需求和预期。

按照上节中的技术分析，应用提供商可按照技术层次细分为多个角色。应用提供商是大数据市场最关键的角色。数据终究是原材料，能否做出一桌好菜，还要看厨师的手艺。对行业的洞察力和经验，就是对火候的掌握，就是厨师的手艺。

大数据不但有用，而且确确实实可以赚钱。数据作为未来企业的战略资源，的确有着毋庸置疑的重要性，但不至于没有数据，就寸步难行，还没到那种“得数据者得天下”的地步。

上面将整个大数据产业链划分成了四种角色。要想在大数据市场上立足，需要先明白自己属于哪个角色。清楚了自己的身份，清楚自己在产业链的位置。继而沿着自己既定的发展方向坚定不移地走下去。接下来要做的就是积累；不断地积累和优化，不断地进步，争取做到各自领域的领头羊。想在大数据市场上谋有席之地，最终靠的还是实力。

今天大数据公司应该做的就是两件事：数据和能力。对于一个大数据公司，你要么有数据，要么有管理和处理数据的能力。没有数据这个生产材料，肯定无法做大数据运营：如果没有足够的驾驭数据的能力，做不出客户满意的效果，也终将会被市场

此为试读，需要完整PDF请访问：www.ertongbook.com