21世纪高等院校教材

大数据技术与应用

主　编　余以胜
副主编　刘芷欣　丁慧鸽　张文君

科学出版社

21世纪高等院校教材

大数据技术与应用

主　编　余以胜

副主编　刘芷欣　丁慧鸽　张文君

科学出版社

北　京

内 容 简 介

本书在对大数据理论和技术进行系统、深入研究的基础上，首次将大数据技术方法和行业应用相结合，形成了基础技术篇（上篇）和行业应用篇（下篇）两个部分。其中上篇介绍了大数据的起源、思想、特点和价值，以及大数据关键技术、应用思路和应用关键问题；下篇分别对大数据产业链、大数据＋工业行业、大数据＋金融行业、大数据＋零售行业、大数据＋医疗行业、大数据＋电信行业等多个典型应用行业进行了分析，最后提出中国大数据产业发展前景及趋势。

本书结构合理、内容丰富，具有较强的理论性、科学性、系统性和实用性。既可作为高等院校计算机科学、管理学、情报学等专业的参考书，也可供广大信息工作者、科研人员和管理人员阅读与使用。

图书在版编目（CIP）数据

大数据技术与应用 / 余以胜主编.—北京：科学出版社，2019.11
21 世纪高等院校教材
ISBN 978-7-03-062405-5

Ⅰ. ①大… Ⅱ. ①余… Ⅲ. ①数据处理–高等学校–教材 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2019）第 210689 号

责任编辑：王京苏 / 责任校对：王丹妮
责任印制：张 伟 / 封面设计：蓝正设计

科学出版社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

北京建宏印刷有限公司 印刷

科学出版社发行 各地新华书店经销
*
2019 年 11 月第 一 版 开本：787×1092 1/16
2019 年 11 月第一次印刷 印张：12 1/2
字数：300 000

定价：52.00 元

（如有印装质量问题，我社负责调换）

前　言

大数据（big data）或称巨量资料，是指所涉及的资料量规模巨大到无法通过目前的主流软件工具，在合理时间内撷取、管理、处理并整理成为帮助企业经营决策的资料。

随着移动互联网、移动终端和数据传感器的出现，数据正以超出人们想象的速度快速增长。近几年，数据量已经从太字节（tera byte，TB）级别跃升到拍字节（peta byte，PB）乃至泽字节（zetta byte，ZB）级别，2018 年全球数据总量达 19.4 ZB。目前全球数据的增长速度在每年 25%左右，以此推算，到 2020 年，全球的数据总量将达到 80 ZB。

同时大数据解决方案不断成熟，各领域大数据应用全面展开，为大数据发展带来强劲动力。2017 年全球大数据市场规模达到 721 亿美元，有行业专家预测 2017～2021 年大数据行业年均复合增长率约为 40.98%，2021 年全球大数据市场规模将达到 2840 亿美元左右。我国大数据处理技术虽处于发展的起步阶段，但各地发展大数据的积极性较高，行业应用得到快速推广，市场规模增速明显。2017 年中国大数据市场规模达到 324 亿元，未来几年，在广大现有和新兴细分市场中，大数据市场仍将呈现强劲的增长势头，预计到 2021 年，我国大数据市场规模将突破 900 亿元。

目前，中国的大数据主要应用于金融、医疗、电信、交通、物流、环保、旅游和电商等细分行业，2016 年，《国家发展改革委办公厅关于组织实施促进大数据发展重大工程的通知》印发后，国务院办公厅和国家各部委先后推出大数据发展意见和方案，大数据政策从全面、总体规划逐渐朝各大产业、各细分领域延伸，大数据产业发展也在逐步从理论研究走向实际应用之路。

余以胜

2019 年 1 月 10 日

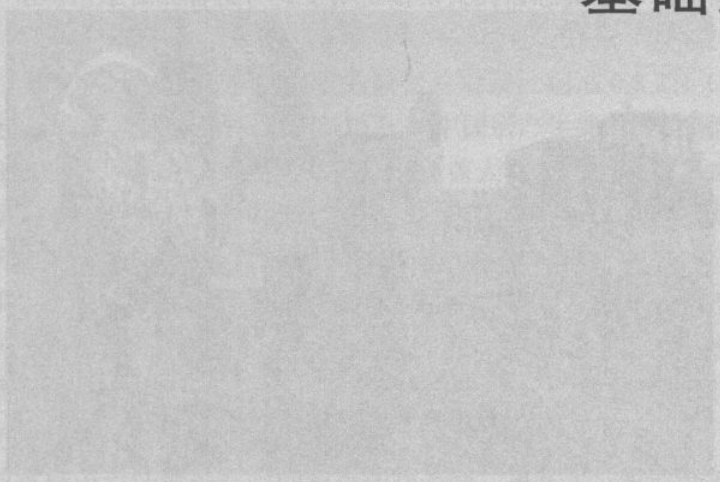
目　录

上篇　基础技术篇

下篇 行业应用篇

上篇

基础技术篇



引　例

利用大数据打造精准农业

美国农业正在采用大数据和互联网方法提升农业生产的效率和效益，以 1%的农业人口维持庞大的农业生产体系，农产品不仅满足美国本土需要，而且大量出口。

罗德尼·席林（Rodney Schilling）是美国伊利诺伊州的一个农场主，他和父亲二人经营着 1300 英亩（1 英亩≈6.07 亩）田地。他的父亲已经 83 岁了，地里的活全靠席林自己上阵，即便在农忙时节，他也不用雇工，其最好的帮手是农场里的那几台农业机械。这些机器普遍高大，一台喷药机完全张开“臂膀”，翼展达 36 米。更重要的是，这些“大家伙”还很有“头脑”——驾驶室里配备的全球卫星导航系统和自动驾驶系统。即使在下田作业时，席林也远没有传统农民那么辛苦，只要他愿意，完全可以坐在驾驶座上，一边喝着咖啡，一边用平板电脑浏览新闻，机器会按照设定的路线工作，施肥、打药完全自动化，哪些地方打过，哪些地方没打，绝对不会弄混，导航系统上都显示得清清楚楚。

图引　数据辅助精准农业

在席林的平板电脑里，安装了气象数据软件。他把农场的坐标和相关信息通过软件上传，即可获得农场范围内的实时天气信息，如温度、湿度、风力、雨水等，这些信息可以帮助他判断每个地块的播种、收获、耕作时间。大数据让农民开始用移动设备管理农场，可以掌握实时的土壤湿度、环境温度和作物状况等信息，大幅度提高了管理的精确性。大多数时候，席林会把平板电脑带在身边，内置的 APP 软件会提醒他何时适宜下地查看，何时打药或施肥，以及提供实时的和未来几天的天气数据。

在美国，像席林这样“劳作”的农场主越来越多。农业生产模式正在从机械化向信息化转变，以精准为特征的农业，正在让种植变得更加容易。

第一章

大数据概述

第一节　大数据时代来临

自20世纪80年代以来，人们就开始尝试在网上进行交易。然而，由于互联网的匿名性与早期第三方监管的缺失，网上交易产生了许多投机行为。随着第三方惩罚机制和声誉机制的建立与完善，网上交易环境逐渐得到优化，越来越多的人愿意在互联网上进行交易事宜，交易所产生的网络数据也不再是人们上网环节产生的副产品。每天在互联网上所产生的交易数据已成为维系社会经济事业的关键纽带，一些业内人士看到了其中的商机，开始运用统计学工具对这些数据进行分析。随着社会的发展，人们的生活水平不断提高，数字化与信息化的普及，与工作生活相关的信息类型和规模都以前所未有的高速不断增长。根据ZDNet（至顶网）发布的《数据中心2013：硬件重构与软件定义》年度技术报告，仅在中国，2013年所产生的数据总量就已超过0.8 ZB（1 ZB=2^{30} TB），相当于2009年全球的数据总量。预计到2020年，中国所产生的数据总量将超过8.5 ZB，相当于2013年的10倍。今天，大数据已深深地覆盖人类经济社会的各方各面，数据从一类简单的处理对象逐渐转变为一种基础资源，如何有效地对其进行开发与管理成为一个具有前瞻性的问题。

2012年8月，美国总统大选正进行得如火如荼。出人意料的是，奥巴马总统的数据团队要求他去一家叫作Reddit的新闻网站回答问题。对许多人来讲，Reddit是一个陌生的名字，总统的高级助手们也不例外。但是来自数据团队的回答却非常简单："因为我们需要动员的一些人，经常在Reddit上。"

这仅是选战过程中一件毫不起眼的数据决策案例。事实上，奥巴马的数据团队非常神秘、低调，但其触角又无处不在，他们被内部人士戏称为"核编码"。他们创建了单一的巨大系统，可以将民调专家、筹款人、选战一线员工、消费者数据库及从"摇摆州"民主党主要选民档案的社会化媒体联系人与手机联系人那里得到的所有数据都聚合到一块。这个组合起来的巨大数据不仅让竞选团队能够发现选民并获取他们的注意，还能让数据处理团队去做一些测试，看哪些类型的人有可能被某种特定的事情所打动或说服。

这个数据库在奥巴马总统大选的筹集资金、广告投放、活动安排等方面发挥了难以替代的作用。以筹集资金为例，在2012年8月，所有人都认为无法完成筹集10亿美元的目标。但是数据团队发现参与"快速捐献"计划的人，捐出的资金是其他捐献者的

4 倍。于是该计划被大规模推广，最终完成了筹集 10 亿美元的目标。

与其依赖于外部媒体顾问来决定广告应该在哪里出现，数据团队觉得不如将他们的购买决策建立在内部大数据库上。“我们可以通过一些真的很复杂的模型，精准定位选民。比如说，迈阿密戴德 35 岁以下的女性选民，如何定位？”一个官员说。结果是，竞选团队买了一些非传统类剧集（如《混乱之子》《行尸走肉》《23 号公寓的坏女孩》）之间的广告时间，而回避了和地方新闻挨着的广告时间。奥巴马团队 2012 年的广告购买效率比 2008 年高了多少呢？芝加哥方面有一个数字：电视广告效率提高了 14%。

数据团队每天晚上都运行 66 000 次选举，次日清晨，数据处理结果告诉竞选团队赢得这些州的机会在哪儿，从而合理调配资源。基于大数据的模拟竞选，可以推算出奥巴马在每个“摇摆州”的胜算，进而采取相对应的活动。

决策者们坐在一间密室里，一边抽雪茄，一边说：“我们总是在《60 分钟》节目上投广告的时代已经结束。在政治领域，大数据的时代已经到来。”

第二节　什么是大数据

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。但这种定义并不够直观和严谨，大数据其实并不是一个新鲜事物，早在 20 世纪 80 年代伊始，被称为“最有影响力的未来学家”的阿尔文·托夫勒（Alvin Toffler）就指出，对大量数据的处理与分析将成为第三次浪潮中最精彩的篇章。最近几年大数据更频繁、更迅速地进入人们的视野，刷新着大众的互联网思维，虽然人们已经对大数据并不感到陌生，但是很少有人真正地探究过隐藏在大数据背后的神秘王国。

提及“大数据”这一概念，很多人只能从数据量上去模糊地感知，其实大数据离我们一点儿也不远。2015 年在亚马逊（Amazon）每秒会产生 72.9 笔购物订单，在 YouTube（优兔）每分钟上传的视频总时长达 20 小时，谷歌（Google）平均每天处理 24 PB 数据量，Facebook（脸谱网）用户超过 10 亿，每月上传近 75 亿张照片，每天生成近 300 TB 日志数据。文字成了数据，机械的物理状态成了数据，人们所处的地理位置成了数据，甚至人与人之间的互动信息也成了数据。据艾力·西格尔估计，全球人类每天都会增加 2.5 万亿字节的数据。

大数据的起源虽然要归功于互联网与电子商务，但大数据最大的应用前景却在传统产业。一是因为几乎所有传统产业都在互联网化，二是因为传统产业仍占据国内生产总值（gross domestic product，GDP）的绝大部分份额。作为传统产业，在互联网时代，应用大数据可以直接获取消费者对产品的反馈，与消费者有了真正意义上的互动沟通。在大数据时代，企业的核心还是在于做更好的产品，提供更好的用户体验。大数据在传统产业的应用其实就是“互联网＋”的一个重要组成部分，如何利用好大数据，对传统行业的升级转型以及管理营销都是一个巨大的机遇和挑战。

简单来讲，大数据需要有大量能互相连接的数据（不管是自己的还是购买、交换别人的），它们在一个大数据计算平台上（或者是能互通的各个数据节点上），有相同的数据标准能正确地关联［如 ETL（extract-transform-load，抽取—转换—加载）数据标准］，

通过大数据相关处理技术（如算法、引擎、机器学习），形成自动化、智能化的大数据产品或者业务，进而形成大数据采集、反馈的闭环，自动智能地指导人类的活动、工业制造、社会发展等。

一、大数据计算提高数据处理效率，增加人类认知盈余

大数据技术就像其他的技术革命一样，是从效率提升入手的。大数据技术平台的出现提升了数据处理效率，其效率是呈几何级数提升的，过去需要几天或更多时间处理的数据，现在可能在几分钟之内就会完成，大数据的高效计算能力，为人类节省了更多的时间。我们都知道效率提升是人类社会进步的典型标志，可以推断大数据技术将带领人类社会进入另外一个阶段。通过大数据计算节省下来的时间，人们可以去消费、娱乐和创造，未来大数据计算将释放人类社会巨大的产能，增加人类认知盈余，帮助人类更好地改造世界。

二、大数据通过全局的数据让人类了解事物背后的真相

相对于过去样本代替全体的统计方法，大数据将使用全局的数据，其统计出来的结果更为精确，更接近事物真相，帮助科学家了解事物背后的真相。大数据带来的统计结果将纠正过去人们对事物错误的认识，改变过去人类社会发展中已有的某些结论，并带来全新的认知，有利于政府、企业、科学家对过去人类社会的各种历史行为真正原因的了解。大数据统计将纠正样本统计误差，为统计结论不断纠错。

三、大数据有助于人类了解事物发展的客观规律，利于科学决策

大数据收集了全局和准确的数据，通过大数据可以了解事物发展过程中的真相，通过数据分析出人类社会的发展规律、自然界的发展规律。利用大数据提供的分析结果来归纳和演绎事物的发展规律，通过掌握事物发展规律来帮助人们进行科学决策，大数据时代的精准营销就是典型的应用。

四、大数据提供了同事物的连接，利于客观了解人类行为

在没有大数据之前，我们了解人类行为的数据往往来源于一些被动的调查表格及滞后的统计数据，拥有大数据技术之后，大量的传感器如手机 APP（application）、摄像头、分享的图片和视频等让我们更加客观地了解人类的行为。大数据技术连接了人类行为，通过大数据将人类的行为数据收集起来，经过一定的分析后来统计人类行为，可以帮助我们了解人类的行为。可以说大数据的一个重要作用就是将人类行为的数据进行收集分析，了解人类行为的特点，为数据价值的商业运用提供基础资产。

五、大数据改变过去的经验思维，帮助人们建立数据思维

人类社会的发展一直都在依赖着数据，如各国的文明演化、农业规划、工业发展、军事战役及政治事件等，尤其是大数据出现之后，我们将会面对着海量的数据，多种维度的数据，如行为的数据、情绪的数据、实时的数据。这些数据是过去没有被了解到的，通过大数据计算和分析技术，人们将会得到不同的事物真相、不同的事物发展规律。依靠大数据提供的数据分析报告，人们将会发现决定一件事、判断一件事、了解一件事不再变得困难。各国政府和企业将借助大数据来了解民众需求，抛弃过去的经验思维和惯性思维，掌握客观规律，跳出利用历史数据预测未来的困境。

第三节　大数据的起源

一、信息科技进步

如果把信息科技的不断进步看成世界万物持续数字化的过程，则会理出一条清晰的主线。信息科技具有三个最核心和基础的能力：信息处理、信息存储和信息传递，几十年来这三个能力的飞速进步，是人类科技史上最为激动人心的事情之一。

现代意义上计算机的发明，归功于军事上的需要。1946 年 2 月 14 日，由美国军方订制的世界上第一台电子计算机——“电子数字积分计算机”在美国宾夕法尼亚大学问世，主要是为了满足计算弹道需要而研制的。“电子计算机”的称谓的确名副其实，其最初的目的就是更迅速地进行大量数学运算。

数学一直是计算机学科的基础，尤其是离散数学，奠定了计算机学科的理论基础。人们把“计算机之父”的桂冠戴在两位数学家的头上，分别是艾伦·图灵和冯·诺依曼。迄今为止，人们都把图灵机作为现代智能类工具的鼻祖。美国计算机协会（Association of Computing Machinery，ACM）于 1966 年设立图灵奖，专门奖励那些对计算机科学研究与推动计算机技术发展有卓越贡献的杰出科学家。它被公认为计算机界的诺贝尔奖。以图灵的姓氏命名的图灵机是一个二进制计算的抽象理论模型，并不是计算机的工程设计。冯·诺依曼则被公认为现代计算机（工程实现）的鼻祖，他领导的小组提出了完善的计算机设计报告。

1965 年，戈登·摩尔（Gordon Moore）——英特尔（Intel）公司的创始人之一，准备了一个关于计算机存储器发展趋势的报告。在他开始绘制数据时，发现了一个惊人的趋势：每个新芯片大体上包含上一代芯片两倍的容量，每个芯片的产生都是在前一个芯片产生后的 18～24 个月内。如果这个趋势继续的话，计算机的计算能力相对于时间周期将呈指数式上升。简而言之，“芯片上可容纳的晶体管数目，每隔 18 个月左右便会增加一倍，性能也将提升一倍”。后来人们发现这不仅适用于对存储器芯片的描述，也精确地说明了计算能力和磁盘存储容量的发展，于是，摩尔定律成为许多工业进行性能预测的基础，主宰了信息产业的发展。

在摩尔定律的指引下，信息产业周期性地推出新的计算机，操作系统和计算能力均

不断提高。工业界和个人都不断地升级计算机设备，从而推动信息产业的迅速进步。每当英特尔公司开发出计算能力更强的芯片，微软公司就会适时推出功能更强大、操作更方便的操作系统。当人们采用了微软公司的新操作系统后，就会发现系统运行的速度变慢，不得不升级硬件设备。每当计算机产业发展放缓，硬件生产商就会翘首企盼微软公司新的操作系统，带动客户新一轮的升级换机热潮。这种循环持续上演了 40 余年。这段波澜壮阔的历史，使信息处理和储存能力获得成千上万倍的提升。

1977 年，世界上第一个光纤通信系统在美国芝加哥市投入商用，速率为 45 Mbit/s，自此，拉开了信息传输能力大幅跃升的序幕。有人甚至将光纤传输带宽的增长规律称为超摩尔定律，认为带宽的增长速度比芯片性能提升的速度还要快。

二、存储介质大幅降价

存储器（memory）是计算机系统中的记忆设备，用来存放程序和数据。计算机中的全部信息，包括输入的原始数据、计算机程序、中间运行结果和最终运行结果都保存在存储器中，它根据控制器指定的位置存入和取出信息。自世界上第一台计算机问世以来，计算机的存储器件也在不断地发展更新，从一开始的汞延迟线、磁带、磁鼓、磁芯，到现在的半导体存储器、磁盘、光盘、纳米存储等，无不体现着科学技术的快速发展。

事实上，存储的价格从 20 世纪 60 年代 1 万美元 1 MB，降到现在的 1 美分 1 GB 的水平，其价差高达亿倍。在线实时观看高清电影，在十几年前还是难以想象的，现在却已是习以为常。网络的接入方式也从有线连接向高速无线连接的方式转变。毫无疑问，网络带宽和大规模存储技术的高速持续发展，为大数据时代提供了廉价的存储和传输服务。

三、互联网诞生

互联网的出现，在科技史上可以比肩“火”与“电”的发明，这个伟大的发明同样是由军事目的驱动的。计算机在军方应用得越广泛，计算机上保存的军事机密就越多。人们担心如果保存重要军事机密数据的主要计算机被摧毁的话，很可能就会输掉整个战争，于是，推动计算机之间互相传递数据并互为备份的通信机制被提上日程。1969 年，美国国防部高级研究计划局（Advanced Research Projects Agency，ARPA）把分属于不同大学的 4 台计算机互相连接起来，组成了一个分组交换网 ARPANET（Advanced Research Projects Agency network，阿帕网）。一年后阿帕网扩大到 15 个节点，1973 年，阿帕网跨越大西洋利用卫星技术与英国、挪威实现连接，扩展到世界范围，这就是最早的互联网雏形。

互联网把每个人桌面上的计算机连接起来，改变了人们的生活，成为人们获取各类数据的首要渠道。通过互联网获取数据的模式可以被简单地抽象为“请求”加“响应”的模式，理解这种获取信息的方式，有助于理解大数据的价值。

四、网上的“脚印”

用收音机听广播，或者用电视机看电视节目，都是“广播”加“接收”的模式。不管有没有电视机在接收信号，广播塔总是在发送电视节目信号。随时打开电视机，随时就能收看电视节目。在“广播”加“接收”模式中，广播塔是不知道有谁在接收节目的，如图 1-1 所示。

图 1-1 “广播”加“接收”模式

“请求”加“响应”模式则不同，如果客户端（所有接入互联网的设备、软件等）不主动要求，服务器端是不会发送任何数据的，如图 1-2 所示。互联网应用协议基本上都是这种模式。当然也有“广播”加“接收”模式的协议，但是不常用。每一次访问请求其实就是一次鼠标点击操作，服务器的日志中忠实地记录下了每个人访问的时间、请求的命令、访问的网址等数据。这些访问记录就像人们在雪地上行走留下的脚印一样，“脚印”连成一串，构成了人们在互联网上的“行为轨迹”。想一想猎人是怎样通过追踪脚印捕获猎物的，就会明白这些“轨迹”中蕴含着巨大的价值。所以各类服务器上的日志就是一种非常重要的大数据类型。

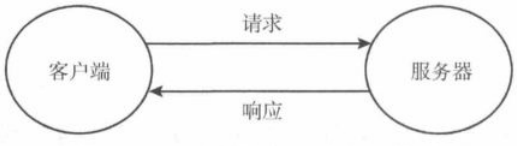


图 1-2 “请求”加“响应”模式

曾经有制衣公司想要调查顾客的购买意愿。需要统计顾客拿起了哪件衣服，试穿了哪件衣服，在专卖店逗留了多长时间。这就需要安装摄像头，要选样本，可能花费上亿元的资金，要想省钱的话其结果可能会失去参考价值，但如果在网上做同样的事情，成本就近乎为零。在淘宝网或者京东商城的主页上，每一个网页都相当于一家店铺，打开这个网页就等于进入了店铺；点击了衣服，相当于顾客拿起衣服仔细端详；把衣服放到收藏夹，可以理解为试穿；实体店中的顾客行为几乎被完整地映射到网页上。不同的是，互联网忠实地记录下顾客在“店”里停留的时间、关心的品类；此外，顾客和销售员的对话、顾客与顾客之间的对话，也被忠实地记录、保存。互联网企业做与那家制衣公司同样的调查，成本近乎为零。

因为互联网的内在机理，互联网成为大规模接近消费者、最理解消费者的工具和平

台，互联网没有删除键，人们在互联网上的一言一行都被忠实地记录。古代皇帝身边总有一位兢兢业业的史官，随身携带纸笔，记下皇帝的起居作息、一言一行。互联网就像每个人的“史官”，它从不知疲倦、事不分大小，悉心而精准地记录着一切。事实上，这位“史官”记录的就是人们的数字化生活。

五、分布式存储技术

1959 年，美国计算机科学家克里斯托弗·斯特雷奇（Christopher Strachey）发表了一篇名为 *Time sharing in large fast computers*（大型高速计算机中的时间共享）的虚拟化论文，虚拟化是今天云计算基础架构的基石。

2004 年，谷歌发布分布式计算系统 Map Reduce 论文。云计算生态系统 Hadoop 就是谷歌集群系统的一个开源项目总称，主要由分布式存储系统 HDFS（Hadoop distributed file system，Hadoop 分布式文件系统）、分布式计算系统 Map Reduce 和分布式数据库 HBase 组成，其中 HDFS 是 Google file system（GFS）的开源实现；Map Reduce 是 Google Map Reduce 的开源实现；HBase 是 Google Big Table 的开源实现。

以谷歌 Map Reduce 为代表的分布式存储计算系统的出现，极大幅度地提升了计算能力同时降低了成本，使得人类获得的计算性能摆脱了硬件摩尔定律的限制，以恐怖的速度发展。

云计算再一次改变了数据的存储和访问方式。在云计算出现之前，数据大多分散保存在每个人的个人计算机中、每家企业的服务器中。云计算，尤其是公用云计算，把所有的数据集中存储到“数据中心”，即所谓的“云端”，用户可通过浏览器或者专用应用程序来访问。

一些大型的网站，通过提供基于“云”的服务，积累大量的数据，成为事实上的“数据中心”。数据是这些大型网站最为核心的资产。它们不惜花费高昂的费用、付出巨大的努力来保管这些数据，以便加快用户的访问速度。谷歌甚至购买了单独的水力发电站，为其庞大的数据中心提供充足的电力。一些公开资料显示，谷歌在全球分布着 36 个数据中心，图 1-3 所示为谷歌数据中心内一景。

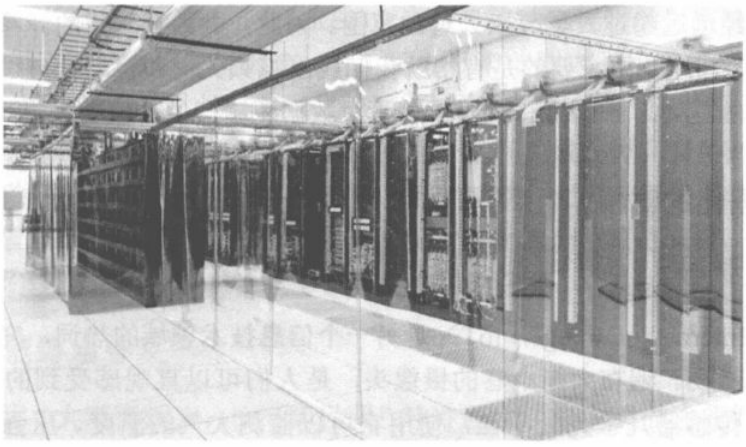


图 1-3 谷歌数据中心内一景

近几年国内兴起了建设云计算基地的风潮，客观上为大数据的诞生准备了必备的储存空间和访问渠道。各大银行、电信运营商、大型互联网公司、政府各个部委都拥有各自的数据中心。银行、电信运营商、互联网公司绝大部分已经实现全国级的数据集中工作。

云计算是大数据诞生的前提和必要条件，没有云计算，就缺少了集中采集数据和存储数据的商业基础。云计算为大数据提供了存储空间和访问渠道，“没有大数据的云计算，就是房地产的代名词”（易欢欢），大数据则是云计算的灵魂和必然的升级方向。云计算确实可以称为一场信息技术领域内的革命，甚至对社会也必将产生革命性的影响，但是它却并不是一场技术革命，云计算在本质上是一场 IT（information technology，信息技术）产品/服务消费方式的变革，云计算中一个广为宣传的核心技术——虚拟化软件早在 20 世纪 60 年代就已经被应用在 IBM 的大型主机中了。

六、机器学习

机器学习（machine learning，ML）是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。其专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构，使之不断改善自身的性能。

它是人工智能的核心，是使计算机具有智能的根本途径，其应用遍及人工智能的各个领域，它主要使用归纳、综合而不是演绎。

这样看机器学习似乎是一门很高深的学科，光是定义我们就看得云里雾里的。但是只要仔细看，我们就能发现，机器学习所需要的知识都是人类发展了很久的数学与计算机知识，也就是说机器学习是过去的计算机技术发展到一定时期的产物。

其实机器学习早在 20 世纪 60 年代就被提出，但是直到近年才兴起，原因就是当年计算机根本没有足够的运算量去实现机器学习的思想。但是云计算的出现解决了这个问题，因此机器学习也就终于能够在人类的历史舞台上成为其中一个主角。

在大数据时代的现在，全球最大视频网站上，每 60 秒就有超过 150 小时长度的视频上传，这已经远远超过人类可以处理的范围。因此，大数据时代的计算机除了要有充足的存储和计算能力，更重要的是有帮助我们找出视频关键和自动处理过去只有人类才能处理的事务的能力。云计算和机器学习是大数据能够产生价值的关键。

七、数据化的时代

（一）物联网

物联网（Internet of things，IoT）是另一个信息技术领域的热词，究其本质是传感器技术进步的产物。遍布大街小巷的摄像头，是人们可以直观感受到的一种物联网形态。事实上，传感器几乎无处不在，使用它可以监测大气的温度、压强、风力，监测