高等教育规划教材

免费提供
电子教案

BIG DATA TECHNOLOGY AND APPLICATION

大数据
技术与应用

• 周苏 冯婵璟 王硕苹 等编著 •

机械工业出版社
CHINA MACHINE PRESS

高等教育规划教材

大数据技术与应用

周　苏　冯婵璟　王硕苹　等编著

机械工业出版社

本书针对计算机、信息管理和其他相关专业学生的发展需求，系统、全面地介绍了大数据技术与应用的基本知识和技能，详细介绍了大数据基础、大数据的行业应用、大数据的基础设施、大数据技术基础、Hadoop 分布式架构、大数据管理、大数据分析、人工智能与机器学习、数据科学与数据科学家、开放数据的时代，以及大数据发展与展望等内容，具有较强的系统性、可读性和实用性。

本书是为高等院校“大数据”相关课程全新设计编写、具有丰富实践特色的主教材，也可供有一定实践经验的软件开发人员和管理人员参考，或作为继续教育的教材。

本书配有授课电子课件，需要的教师可登录 www.cmpedu.com 免费注册，审核通过后下载，或联系编辑索取（QQ：2850823885，电话：010-88379739）。

图书在版编目（CIP）数据

大数据技术与应用 / 周苏等编著. —北京：机械工业出版社，2016.3
高等教育规划教材
ISBN 978-7-111-53304-7

Ⅰ. ①大… Ⅱ. ①周… 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（2016）第 060332 号

机械工业出版社（北京市百万庄大街 22 号 邮政编码 100037）
策划编辑：郝建伟 责任编辑：郝建伟
责任校对：张艳霞 责任印制：乔 宇

北京铭成印刷有限公司印刷

2016 年 4 月第 1 版 • 第 1 次印刷
184mm×260mm • 13.25 印张 • 328 千字
0001－3000 册
标准书号：ISBN 978-7-111-7-53304-7
定价：39.00 元

前　言

由于互联网和信息行业的快速发展，大数据（Big Data）越来越引起人们的关注，已经引发自互联网、云计算之后 IT 行业的又一大颠覆性的技术革命。面对信息的激流，多元化数据的涌现，大数据已经为个人生活、企业经营，甚至国家与社会的发展都带来了机遇和挑战，成为 IT 信息产业中最具潜力的蓝海。人们用大数据来描述和定义信息爆炸时代产生的海量数据，并命名与之相关的技术发展与创新。云计算主要为数据资产提供了保管、访问的场所和渠道，而数据才是真正有价值的资产。企业内部的经营信息、互联网世界中的商品物流信息，以及互联网世界中的人与人交互信息、位置信息等，其数量将远远超越现有企业 IT 架构和基础设施的承载能力，实时性要求也将大大超越现有的计算能力。如何盘活这些数据资产，使其为国家治理、企业决策乃至个人生活服务，是大数据的核心议题，也是云计算内在的灵魂和必然的发展方向。

大数据技术与应用是一门理论性和实践性都很强的课程。在长期的教学实践中，笔者体会到，坚持“因材施教”的重要原则，把实践环节与理论教学相融合，用实践教学促进理论知识的学习，是有效改善教学效果和提高教学水平的重要方法之一。本书的主要特色是：理论联系实际，结合一系列了解和熟悉大数据技术与应用的学习和实践活动，把大数据的相关概念、基础知识和技术技巧融入实践当中，使学生保持浓厚的学习热情，加深对大数据技术的认识、理解和掌握。

本书是为高等院校相关专业开设“大数据”相关课程而全新设计编写、具有丰富实践特色的主教材，也可供有一定实践经验的软件开发人员和管理人员参考，或作为继续教育的教材。

本书针对计算机、信息管理和其他相关专业学生的发展需求，系统、全面地介绍了大数据技术与应用的基本知识和技能，详细介绍了大数据基础、大数据的行业应用、大数据的基础设施、大数据技术基础、Hadoop 分布式架构、大数据管理、大数据分析、人工智能与机器学习、数据科学与数据科学家、开放数据的时代，以及大数据发展与展望等内容，具有较强的系统性、可读性和实用性。

结合课堂教学方法改革的要求，本书设计了全新的课程教学过程，为每章教学内容都有针对性地设计了课后的实验与练习环节，要求和指导学生在课后阅读课文、网络搜索浏览的基础上，延伸阅读，拓展视野，深入理解课程知识内涵。

本课程的教学进度设计体现在“课程教学进度表”中。该表可作为教师授课参考和学生课程学习的概要。

实际授课时，应按照教学大纲编排教学进度，按照教学日历考虑本学期节假日安排，进而确定本课程的教学进度。

本课程的教学评测可以从以下几个方面入手。

（1）将每周的课后实验与思考（10 次）作为平时成绩。

（2）课程实验总结（第 11 章）。

（3）结合平时考勤。

课程教学进度表

（20 —20 学年第 学期）

课程号：_____课程名称：大数据·技术与应用 学分：2 周学时：2
总学时：34 （其中理论学时（课内）：34 （课外）实践学时：（34）
主讲教师：_____

序号	教学日历 周次	章节（或实验、习题课等） 名称与内容	学时	教学方法	课后作业布置
1	1	引言与第 1 章 大数据概述	2	课堂教学	
2	2	第 1 章 大数据概述	2	课堂教学	实验与思考
3	3	第 2 章 大数据的行业应用	2	课堂教学	实验与思考
4	4	第 3 章 大数据的基础设施	2	课堂教学	
5	5	第 3 章 大数据的基础设施	2	课堂教学	实验与思考
6	6	第 4 章 大数据技术基础	2	课堂教学	
7	7	第 4 章 大数据技术基础	2	课堂教学	实验与思考
8	8	第 5 章 Hadoop 分布式架构	2	课堂讨论	
9	9	第 5 章 Hadoop 分布式架构	2	课堂教学	实验与思考
10	10	第 6 章 大数据管理	2	课堂教学	实验与思考
11	11	第 7 章 大数据分析	2	课堂教学	
12	12	第 7 章 大数据分析	2	课堂教学	实验与思考
13	13	第 8 章 人工智能与机器学习	2	课堂教学	实验与思考
14	14	第 9 章 数据科学与数据科学家	2	课堂教学	实验与思考
15	15	第 10 章 开放数据的时代	2	课堂教学	实验与思考
16	16	第 11 章 大数据发展与展望	2	课堂教学	
17	17	（机动）课程总复习，实验总结	2	课堂教学	课程实验总结

填表人（签字）： 日期：
系（教研室）主任（签字）： 日期：

本书配有授课电子课件，需要的教师可登录 www.cmpedu.com 免费注册，审核通过后下载。欢迎教师与作者交流，索取为本书教学配套的相关资料并交流：zhousu@qq.com，QQ：81505050，个人博客：http://blog.sina.com.cn/zhousu58。

本书的编写得到了浙江大学城市学院、浙江省科技人才教育中心、温州安防职业技术学院和浙江商业职业技术学院等多所院校师生的支持，褚赟、蔡锦锦、张丽娜、王文参与了本书的部分编写工作，在此一并表示感谢！

周 苏

目　录

第1章 大数据概述

所谓大数据，从狭义上可定义为：难以用现有的一般技术管理的大量数据的集合。对大量数据进行分析，并从中获得有用的观点，这种做法在一部分研究机构和大企业中早已存在。现在的大数据和过去相比，主要有三点区别：第一，随着社交媒体和传感器网络等的发展，正产生出大量且多样的数据；第二，随着硬件和软件技术的发展，数据的存储和处理成本大幅下降；第三，随着云计算的兴起，大数据的存储和处理环境已经没有必要自行搭建。

通过分析顾客与公司之间的交互数据，可以得到相关交易数据产生的背景信息。目前，网上（线上）交互数据的采集与分析正先行一步，但今后，对线下及 O2O（Online to Offline）交互数据的分析将变得愈发重要。

1.1 什么是大数据

人类的数字世界包括上传到手机中的图像和视频、用于高清电视的数字电影、ATM 机中的银行数据、机场和重要活动的安全录像（比如奥林匹克运动会）、欧洲原子能研究机构（CERN）中大型强子对撞机的亚原子碰撞记录、优步专车的拼车路线记录、通过移动网络传输的微信语音通话，以及用于日常沟通的短信文本等。

根据 IDC[⊖]《数字世界》研究项目的统计，2010 年全球数字世界的规模首次达到了 ZB（1ZB=1 万亿 GB）级别（1.227ZB）；而 2005 年这个数字只有 130EB，基本上 5 年增长了 10 倍。这种爆炸式的增长，意味着到 2020 年，数字世界的规模将达到 40ZB，即 15 年增长 300 倍。如果单就数量而言，40ZB 相当于地球上所有海滩上的沙粒数量的 57 倍。如果用蓝光光盘保存所有这些 40ZB 数据，这些光盘的重量（不包括任何光盘套和光盘盒）将相当于 424 艘尼米兹级航空母舰的重量（满载排水量约 10 万吨），或者相当于世界上每个人拥有 5247GB 的数据。无疑，现在已经进入了“大数据”时代。

和之前的一些 IT 流行语一样，“大数据”也是一个起源于欧美的词汇。在一些以大数据为主题的报告中，经常会引用 2010 年 2 月出版的《经济学家》（The Economist）杂志中一篇题为 The data deluge 的文章。Deluge 的中文意思是“大泛滥、大洪水”“大量”。因此，这篇文章的标题直译出来，就是“数据洪流”或“海量数据”。自这篇文章问世以来，大数据作为热门话题的出镜率便急剧上升，因此可以肯定的是，这篇文章是大数据备受瞩目的一个重大契机。

⊖ IDC：指国际数据公司（International Data Corporation），是全球著名的信息技术、电信行业和消费科技市场咨询、顾问和活动服务专业提供商。

基本知识：字节大小。

字节最小的基本单位是 Byte（B），按照进率 1024（即 2 的十次方）计算，顺序给出如下。

1B = 8bit（位），一个英文字符
1KB = 1024B，一个句子或一段话
1MB = 1024KB，一个 20 页的幻灯片演示文稿或一本小书
1GB = 1024MB，书架上 9m 长的书
1TB = 1024GB，300h 的优质视频、美国国会图书馆存储容量的 1/10
1PB = 1024TB，35 万张数字照片
1EB = 1024PB，1999 年全世界生成的信息的一半
1ZB = 1024EB，暂时无法想象
1YB = 1024ZB
1DB = 1024YB
1NB = 1024DB

2011 年 5 月，美国麦肯锡全球研究院（MGI）发表了一篇名为 Big Data: The Next Frontier for Innovation，Competition and Productivity（大数据：未来创新、竞争、生产力的指向标）的研究报告，“大数据”（big data，见图 1-1）这个关键词便开始沿用至今。不过，最先对如何面对庞大数据这一问题进行剖析的，应该还是《经济学家》杂志中的那篇文章。从 2012 年开始，大数据成了 IT 业界关注度不断提高的关键词之一。

图 1-1　大数据时代

1.1.1　大数据的定义

所谓大数据，是指用现有的一般技术难以管理的大量数据的集合，即所涉及的资料量规模巨大到无法通过目前主流软件工具，在合理时间内实现获取、管理、处理、并使之成为有效的辅助企业经营决策的信息。

所谓“用现有的一般技术难以管理”，是指用目前在企业数据库占据主流地位的关系型数据库无法进行管理的、具有复杂结构的数据。或者也可以说，是指由于数据量的增大，导致对数据的查询（Query）响应时间超出允许范围的庞大数据。

研究机构 Gartner 给出了这样的定义：大数据是需要新的处理模式，才能使用户具有更强的决策力、洞察发现力和流程优化能力，以及海量、高增长率和多样化的信息资产。

麦肯锡说："大数据指的是所涉及的数据集规模已经超过了传统数据库软件获取、存储、管理和分析的能力。这是一个被故意设计成主观性的定义，并且是一个关于多大的数据集才能被认为是大数据的可变定义，即并不定义大于一个特定数字的 TB 才称为大数据。因为随着技术的不断发展，符合大数据标准的数据集容量也会增长；并且定义随不同的行业也有变化，这依赖于在一个特定行业通常使用何种软件和数据集有多大。因此，大数据在今天不同行业中的范围可以从几十 TB 到几 PB。"

如今，"大数据"这一通俗直白、简单朴实的名词，已经成为最火爆的 IT 行业词汇，随之，数据仓库、数据安全、数据分析和数据挖掘等围绕大数据商业价值的利用正逐渐成为行业人士争相追捧的利润焦点，在全球引领了又一轮数据技术革新的浪潮。

1.1.2 用 3V 描述大数据的特征

从字面来看，"大数据"这个词可能会让人觉得只是容量非常大的数据集合而已。但容量只不过是大数据特征的一个方面，如果只拘泥于数据量，就无法深入理解当前围绕大数据所进行的讨论。因为"用现有的一般技术难以管理"这样的状况，并不仅仅是由于数据量增大这一个因素所造成的。

IBM 说："可以用 3 个特征相结合来定义大数据：数量（Volume，或称容量）、种类（Variety，或称多样性）和速度（Velocity），或者就是简单的 3V，即庞大容量、极快速度和种类丰富的数据。"如图 1-2 所示。

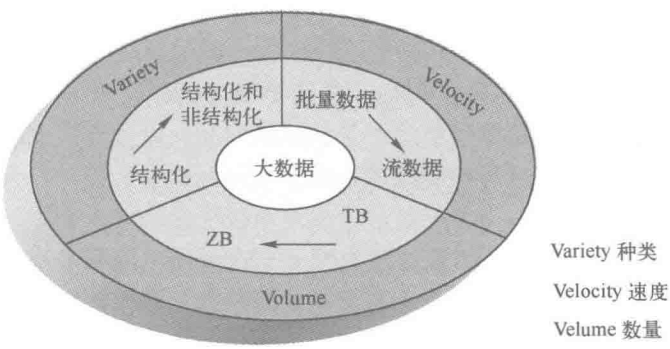

图 1-2 按数量、种类和速度来定义大数据

1. Volume（数量）

用现有技术无法管理的数据量，从现状来看，基本上是指从几十 TB 到几 PB 这样的数量级。当然，随着技术的进步，这个数值也会不断变化。

如今，存储的数据数量正在急剧增长中，存储的事物包括环境数据、财务数据、医疗数据和监控数据等。有关数据量的对话已从 TB 级别转向 PB 级别，并且不可避免地会转向 ZB 级别。可是，随着可供企业使用的数据量的不断增长，可处理、理解和分析的数据的比例却不断下降。

2. Variety（种类、多样性）

随着传感器、智能设备及社交协作技术的激增，企业中的数据也变得更加复杂，因

为它不仅包含传统的关系型数据，还包含来自网页、互联网日志文件、搜索索引、社交媒体论坛、电子邮件、文档、主动和被动系统的传感器数据等原始、半结构化和非结构化数据。

这里的种类是表示所有的数据类型。其中，爆发式增长的一些数据，如互联网上的文本数据、位置信息、传感器数据和视频等，用企业中主流的关系型数据库是很难存储的，它们都属于非结构化数据。

当然，在这些数据中，有一些是过去就一直存在并保存下来的。和过去不同的是，这些大数据并非只是存储起来就够了，还需要对其进行分析，并从中获得有用的信息。例如监控摄像机中的视频数据。近年来，超市、便利店等零售企业几乎都配备了监控摄像机，其最初目的是为了防范盗窃，但现在也出现了使用监控摄像机的视频数据来分析顾客购买行为的案例。

例如，美国高级文具制造商万宝龙（Montblanc）过去是凭经验和直觉来决定商品陈列的布局的，现在尝试利用监控摄像头对顾客在店内的行为进行分析。通过分析监控摄像机的数据，将最想卖出去的商品移动到最容易吸引顾客目光的位置，使得销售额提高了 20%。

美国移动运营商 T-Mobile 也在其全美 1000 家店中安装了带视频分析功能的监控摄像机，可以统计来店人数，还可以追踪顾客在店内的行动路线、在展台前停留的时间，甚至是试用了哪一款手机、试用了多长时间等，对顾客在店内的购买行为进行分析。

3．Velocity（速度）

数据产生和更新的频率也是衡量大数据的一个重要特征。就像所收集和存储的数据量和种类发生了变化一样，生成和处理数据的速度也在变化。不要将速度的概念限定为与数据存储库相关的增长速率，应动态地将此定义应用到数据，即数据流动的速度。有效处理大数据需要在数据变化的过程中对它的数量和种类进行分析，而不只是在它静止后进行分析。

例如，遍布全国的便利店在 24 小时内产生的 POS 机数据，电商网站中由用户访问所产生的网站点击流数据，高峰时达到每秒近万条的微信短文，以及全国公路上安装的交通堵塞探测传感器和路面状况传感器（可检测结冰、积雪等路面状态）等，每天都在产生着庞大的数据。

IBM 在 3V 的基础上又归纳总结了第四个 V——Veracity（真实和准确）。“只有真实而准确的数据才能让对数据的管控和治理真正有意义。随着社交数据、企业内容、交易与应用数据等新数据源的兴起，传统数据源的局限性被打破，企业愈发需要有效的信息治理以确保其真实性和安全性。”

IDC（互联网数据中心）说：“大数据是一个貌似不知道从哪里冒出来的大的动力。但是实际上，大数据并不是新生事物。然而，它确实正在进入主流，并得到重大关注，这是有原因的。廉价的存储、传感器和数据采集技术的快速发展、通过云和虚拟化存储设施增加的信息链路，以及创新软件和分析工具，正在驱动着大数据。大数据不是一个‘事物’，而是一个跨多个信息技术领域的动力/活动。大数据技术描述了新一代的技术和架构，其被设计用于：通过使用高速（Velocity）的采集、发现和/或分析，从超大容量（Volume）的多样（Variety）数据中经济地提取价值（Value）。”

这个定义除了揭示大数据传统的 3V 基本特征，即 Volume（大数据量）、Variety（多样性）和 Velocity（高速）外，还增添了一个新特征——Value（价值）。

一个大数据实现的主要价值可以基于下面 3 个评价准则中的 1 个或多个进行评判。

- 它提供了更有用的信息吗？
- 它改进了信息的精确性吗？
- 它改进了响应的及时性吗？

事实上，大数据，或者说“极限信息”（Extreme Information）具有 12 个维度（象限）。图 1-3 展示了极限信息管理的 3 个层次和 12 个象限。

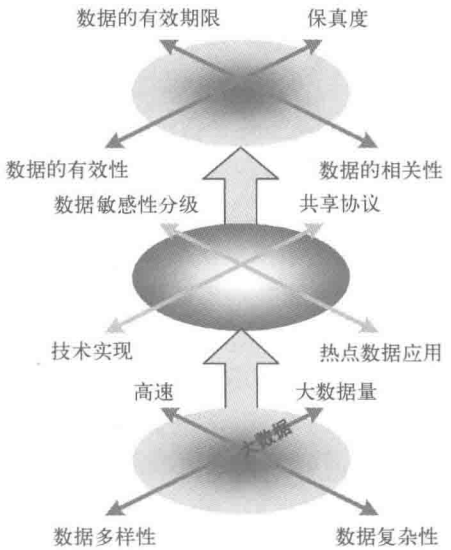

图 1-3　极限信息管理的 3 个层次和 12 个象限

最下面一层“量化指标”指的是大数据的基本特征，即大数据量、多样性和高速，即传统的 3V 概念。另外还加上了“复杂性”（Complexity），包括空间维、时间维等多种数据复杂性。大数据解决方案应首先考虑以这些问题为出发点。然而，解决这 4 个方面的问题只是大数据解决方案的基础，用以支撑起大数据平台，在这之上还有很多问题需要解决。

第二层“访问权限管理和控制”有很多关于访问权限的问题。数据的敏感性是一个很基础的问题，但到现在为止，基于现有的技术和管理手段，还没有对数据的敏感性进行分析的优秀的解决方案。所谓共享协议，即数据将会以什么形式、什么格式和时间点通过什么样的接口实现这些共享和数据的交换，这是大数据的重点问题之一。数据交换的所有方式都是以标准的协议来支持的，因为在大数据时代，数据的来源本身是多样性的，数据的格式甚至是无法管理的，还有很多数据来自企业外部，来自互联网的提供商，到底如何通过这些协议自动将数据放到数据仓库里面来，这种情况下，数据的共享协议是一个很关键的问题。至于热点数据，在大数据时代，数据管理与传统的方式有非常明显的差别。传统的数据管理会把单独的时间点作为一个热点数据，但是在大数据时代，热点数据有可能是并行的多个。这些热点数据之间实际上是有可能有联系的。由于各种事件的相互触发，这些热点数据可能同时出现，而且是相互关联的，甚至是可以预测的。所以说在大数据时代，热点数据的管理也是一个重要话题。

最上面一层“质量管理”也是传统数据管理中非常重要的一个方面。这里面提到的有效性和有效期限，都有明确的技术工具来解决。但到现在为止，在这些方面还是非常依赖传统

的数据仓库工具，而没有专门针对大数据的工具和技术能够解决这些问题。其结果是，大数据应用一方面受制于用户接受的程度，另一方面也受制于技术。现在看来，很多用户仍然必须依赖传统的数据管理的解决方案，而只能拿大数据的技术作为一个前台来做一些预处理。因为它缺少相应的技术和工具的支持。所以，大数据从 12 个象限的角度来说，还只是一个初步，因为里面一些非常基本的问题到现在还没有解决。大数据的形态有很多，现在仍然是雏形阶段。数据的集成，尤其是跨行业、跨不同的部门、跨各种技术能集成起来的机会还是非常少的。

除了业内主流的以大数据 3V 特征为基础的定义外，还有使用 3S 或者 3I 来描述大数据特征的定义。

3S 分别是 Size（大小）、Speed（速度）和 Structure（结构）。实际上，这个维度的特征与 3V 异曲同工，除了用词的不同，并没有太大的差别。

关于大数据的 3I，介绍如下。

1）Ill-defined（定义不明确的）：多个主流的大数据定义都强调了数据的规模需要超过传统方法的处理能力。而随着技术的进步，数据分析的效率不断提高，符合大数据定义的数据规模也会相应地不断变大，因而并没有一个明确的标准。

2）Intimidating（令人生畏的）：从管理大数据到使用正确的工具获取它的价值，利用大数据的过程充满了各种挑战。

3）Immediate（即时的）：数据的价值会随着时间快速衰减。因此，为了保证大数据的可控性，需要通过减少数据收集到获得数据洞察之间的时间，使得大数据成为真正的即时大数据。这意味着能尽快地分析数据对获得竞争优势是至关重要的。

总之，大数据是一个动态的定义，不同行业根据其应用的不同有着不同的理解，其衡量标准也在随着技术的进步而改变。

1.1.3 广义的大数据

前面关于大数据定义的着眼点仅仅在于数据的性质上，因此，将其视为狭义上的定义，并在广义层面上再为大数据下一个定义，如图 1-4 所示。

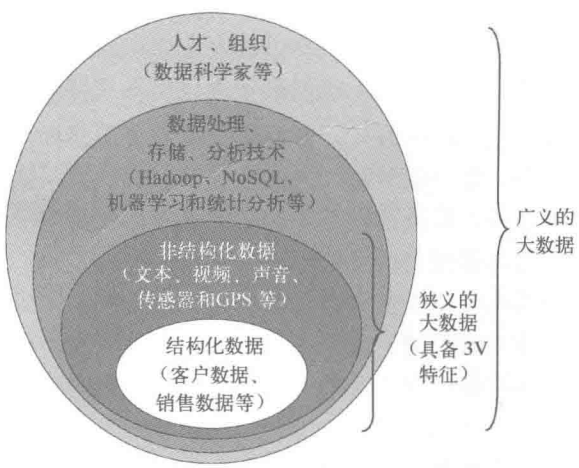

图 1-4 广义的大数据

所谓大数据，是一个综合性概念，它包括因具备 3V（Volume、Variety 和 Velocity）特征而难以进行管理的数据，对这些数据进行存储、处理和分析的技术，以及能够通过分析这些数据获得实用意义和观点的人才和组织。

所谓“存储、处理和分析的技术”，指的是用于大规模数据分布式处理的框架 Hadoop、具备良好扩展性的 NoSQL 数据库，以及机器学习和统计分析等。所谓“能够通过分析这些数据获得实用意义和观点的人才和组织”，指的是目前十分紧俏的“数据科学家”这类人才，以及能够对大数据进行有效运用的组织。

1.2 大数据的结构类型

大数据具有多种形式，从高度结构化的财务数据，到文本文件、多媒体文件和基因定位图的任何数据，都可以称为大数据。数据量大是大数据的一致特征。由于数据自身的复杂性，作为一个必然的结果，处理大数据的首选方法就是在并行计算的环境中进行大规模并行处理（Massively Parallel Processing，MPP），这使得同时发生的并行摄取、并行数据装载和分析成为可能。实际上，大多数的大数据都是非结构化或半结构化的，这需要不同的技术和工具来处理和分析。

大数据最突出的特征是它的结构。图 1-5 显示了几种数据结构类型数据的增长趋势，由图 1-5 可知，未来数据增长的 80%～90%将来自不是结构化的数据类型（半结构化、“准”结构化和非结构化）。

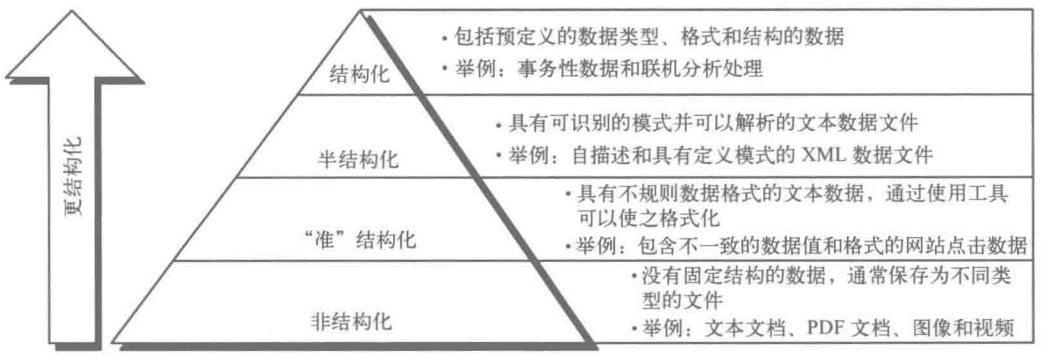

图 1-5 数据增长日益趋向非结构化

虽然图 1-5 显示了 4 种不同的、相分离的数据类型，实际上，有时这些数据类型是可以被混合在一起的。例如，有一个传统的关系数据库管理系统保存着一个软件支持呼叫中心的通话日志，这里有典型的结构化数据，比如日期/时间戳、机器类型、问题类型和操作系统，这些都是在线支持人员通过图形用户界面上的下拉式菜单输入的。另外，还有非结构化数据或半结构化数据，比如自由形式的通话日志信息，这些可能来自包含问题的电子邮件，或者技术问题和解决方案的实际通话描述。另外一种可能是与结构化数据有关的实际通话的语音日志或者音频文字实录。即便是现在，大多数分析人员还无法分析这种通话日志历史数据库中的最普通和高度结构化的数据，因为挖掘文本信息是一项强度很大的工作，并且无法简单地实现自动化。

人们通常最熟悉结构化数据的分析，然而，半结构化数据（XML）、“准”结构化数据

（网站地址字符串）和非结构化数据代表了不同的挑战，需要不同的技术来分析。

1.3 大数据的发展

大数据本身并不是一个新的概念。特别是仅仅从数据量的角度来看的话，大数据在过去就已经存在了。例如，波音的喷气发动机每 30min 就会产生 10TB 的运行信息数据，安装有 4 台发动机的大型客机，每次飞越大西洋就会产生 640TB 的数据。世界各地每天有超过 2.5 万架的飞机在工作，可见其数据量是何等庞大。生物技术领域中的基因组分析，以及以 NASA（美国国家航空航天局）为中心的太空开发领域，从很早就开始使用十分昂贵的高端超级计算机来对庞大的数据进行分析和处理了。

现在和过去的区别之一，就是大数据已经不仅产生于特定领域中，而且还产生于人们每天的日常生活中，微信、Facebook（脸谱）和 Twitter（推特）等社交媒体上的文本数据就是最好的例子。而且，尽管人们无法得到全部数据，但大部分数据可以通过公开的 API（应用程序编程接口）相对容易地进行采集。在 B2C（商家对顾客）企业中，使用文本挖掘（text mining）和情感分析等技术，就可以分析消费者对自家产品的评价。

1.3.1 硬件性价比提高与软件技术进步

计算机性价比的提高，磁盘价格的下降，利用通用服务器对大量数据进行高速处理的软件技术 Hadoop 的诞生，以及随着云计算的兴起，甚至已经无须自行搭建这样的大规模环境——上述这些因素大幅降低了大数据存储和处理的门槛。因此，过去只有像 NASA 这样的研究机构及屈指可数的几家特大企业才能做到对大量数据的深入分析，现在只需极小的成本和时间就可以完成。无论是刚刚创业的公司还是存在多年的公司，也无论是中小企业还是大企业，都可以对大数据进行充分利用。

1. 计算机性价比的提高

承担数据处理任务的计算机，其处理能力遵循摩尔定律，一直在不断进化。所谓摩尔定律，是美国英特尔公司共同创始人之一的高登• 摩尔（Gordon Moore，1929—）于 1965 年提出的一个观点，即“半导体芯片的集成度，大约每 18 个月会翻一番”。从家电卖场中所陈列的计算机规格指标就可以一目了然地看出，现在以同样的价格能够买到的计算机，其处理能力已经和过去不可同日而语了。

2. 磁盘价格的下降

除了 CPU 性能的提高，硬盘等存储器（数据的存储装置）的价格也在明显下降。2000 年的硬盘驱动器平均每 GB 容量的单价约为 16～19 美元，而现在只有 7 美分（换算成人民币的话，就相当于 4～5 角），相当于下降到了 10 年前的 230～270 分之一。

变化的不仅仅是价格，存储器在重量方面也有了巨大进步。1982 年日立公司最早开发的超 1GB 级硬盘驱动器（容量为 1.2GB），重量约为 250lb（约合 113kg）。而现在，32GB 的微型 SD 卡重量却只有 0.5g 左右，技术进步的速度相当惊人。

3. 大规模数据分布式处理技术 Hadoop 的诞生

Hadoop 是一个可以在通用服务器上运行的开源分布式处理软件，它的诞生成为目前大数据浪潮的第一推动力。如果只是结构化数据不断增长，用传统的关系型数据库和数据仓