高等教育“十二五”规划教材

中国第一本针对医学院校的大数据应用教材

YIXUE DASHUJU WAJUE YU YINGYONG

医学大数据挖掘与应用

娄 岩 主编

科学出版社

普通高等教育“十二五”规划教材

医学大数据挖掘与应用

娄 岩 主编

刘尚辉 郑琳琳 副主编

科学出版社

北京

内 容 简 介

本书是将大数据这一计算机前沿科学和基础教学有机结合的典范教材，全面介绍了大数据和相关的基础知识，由浅入深地剖析了大数据的分析处理方法和技术手段，突出介绍了其在医学上的实践应用。

全书共 10 章。第 1 章概括介绍了大数据，第 2～4 章介绍了大数据的采集、预处理、建模和可视化方法，第 5、6 章介绍了 Hadoop 技术，第 7 章介绍了 NoSQL 技术，第 8 章介绍了大数据与云计算的关系，第 9 章介绍了大数据解决方案，第 10 章就医学大数据挖掘做了专题介绍。

本书可作为医学高等院校计算机专业大数据分析及应用课程的教材，也可作为相关技术人员的参考用书。

图书在版编目(CIP)数据

医学大数据挖掘与应用/娄岩主编. —北京：科学出版社，2015
（普通高等教育“十二五”规划教材）
ISBN 978-7-03-045492-8

Ⅰ. ①医… Ⅱ. ①娄… Ⅲ. ①医学－数据处理 Ⅳ. ①R319

中国版本图书馆 CIP 数据核字（2015）第 194279 号

责任编辑：吴宏伟 赵宝平 / 责任校对：刘玉靖
责任印制：吕春珉 / 封面设计：东方人华

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
北京路局票据印刷厂印刷
科学出版社发行 各地新华书店经销
*
2015 年 8 月第 一 版 开本：787×1092 1/16
2015 年 8 月第一次印刷 印张：14 1/4
字数：338 000

定价：39.00 元
（如有印装质量问题，我社负责调换〈路局票据〉）

销售部电话 010-62134988 编辑部电话 010-62135120-2005

《医学大数据挖掘与应用》编写委员会

主　编　娄　岩

副主编　刘尚辉　郑琳琳

编　委　（按姓氏笔画排序）

马　瑾　王艳华　刘尚辉　李　丹　李　岩

李　静　张志常　庞东兴　郑琳琳　娄　岩

郭婷婷　徐东雨　曹　阳　霍　妍

前言

IT 产业在其发展历程中，经历过几轮技术浪潮。如今，大数据浪潮正在迅速朝我们涌来，并将触及各个行业和生活的许多方面。大数据浪潮将比之前发生过的浪潮更大、触及面更广，给人们的工作和生活带来的变化和影响更大。

大数据的应用激发了一场思想风暴，也悄然改变了我们的生活方式和思维习惯。大数据正以前所未有的速度颠覆人们探索世界的方法，引起科研、医学、军事等领域的深刻变革。因此，在当前大数据浪潮的猛烈冲击下，包括医学在内的各种专业在校大学生迫切需要充实和完善自己原有的 IT 知识结构，掌握两个“本领”，一是大数据基本技术与应用，使大数据为我所用的本领；二是挖掘数据之间隐藏的规律与关系，使大数据更好地服务于社会发展的本领。为此，本书围绕大数据及其相关技术，采用深入浅出的叙述方式，简明扼要地阐述了大数据和云计算等新兴技术的基本理论、知识内容、关键技术和实际应用，目的是让广大医药院校师生以计算机公共基础课程为知识载体，对大数据在医学领域的应用方法和相关知识有所了解。而将大数据相关课程纳入大学基础教育中，必将引领学生更好地把握时代科学发展的脉搏和历史赋予的机遇。

在编写原则上，本书既维持了大数据相关信息技术本身应有的系统性和理论性，又着重体现其在医学方向的应用性与针对性；本着理论联系实际的教学目的，注重启发式教学策略，同时适合混合式教学模式，便于学生理解和掌握。本书内容涵盖大数据概论、大数据的采集及预处理、数据建模和可视化方法、Hadoop 技术、NoSQL 数据管理技术、大数据与云计算的关系、大数据解决方案以及医学大数据挖掘。由于医学自身的特殊性，在此领域开展大数据挖掘应用教学，意义深远，也有助于培养具有前瞻性的复合型医学人才。

全书共分 10 章，第 1 章由娄岩编写，第 2 章由郑琳琳编写，第 3 章由刘尚辉编写，第 4 章由李静编写，第 5 章由马瑾编写，第 6 章由徐东雨编写，第 7 章由曹阳编写，第 8 章由张志常编写，第 9 章由霍妍编写，第 10 章由庞东兴编写。

本书的编写得到了中国医药教育协会及 463 医院李丹医生的大力支持，在此表示由衷的感谢。由于编者水平有限，加之时间仓促，书中难免存在疏漏之处，恳请广大读者批评指正，以期不断改进。

娄　岩

2015 年 6 月

目录

第1章

1 大数据概论

大数据（Big Data）是继物联网之后IT产业又一次非常大的技术变革。由于互联网发展，科学数据处理、商业智能数据分析等具有海量需求的应用变得越来越普遍，面对如此巨大的数据量，无论从形式还是内容上，都已无法用传统的方式进行采集、存储、操作、管理和分析。全球产生的数据量，仅在2011年就达到1ZB，根据预测，未来十年全球数据存储量将增长50倍。因此无论是从科学研究还是从应用的角度看，大数据应用已经成为信息社会发展的必然。即便如此，大数据也只有针对某个方面的应用时才可称为大数据应用，而找出数据源，确定数据量，选择正确的处理方法，并得出最终结果的过程才有意义。互联网是大数据的载体之一，离开了一定的数据量，大数据就失去了“灵魂”，而避开实际应用，数据量再大也毫无意义。

无论是分析专家还是数据科学家最终都会探索新的、无法想象的庞大数据集，以期发现一些有价值的趋势、形态和解决问题的方法。由于大多数据源是半结构化或非结构化的，因此处理数据不但需要花费很多时间，而且很难找出解决问题的方法。这也是为什么人们很难就大数据给出一个即严格又准确的定义和大数据发展至今也没有建立起一套完整的理论体系的原因。

以企业为例，对企业内部的纷乱数据通过分析进行决策的目的是帮助企业领导者更好地管理企业。一旦人们开始认识到数据的价值，那么驾驭和分析大数据仅仅是现在工作的扩展和延伸。大数据就是互联网发展到现今阶段的一种表象或特征而已，在以云计算为代表的技术创新大幕的衬托下，一些原本很难收集和使用的数据开始容易被利用起来，通过各行各业的不断创新，大数据会逐步为人类创造更多的价值。

学习目标

- 了解大数据的基本概念、特点和技术架构。
- 熟悉大数据整体技术和关键技术。
- 熟悉大数据处理分析的五种典型工具。
- 了解大数据的应用案例和在医疗领域中的应用。
- 了解大数据未来的发展趋势。

1.1 大数据技术概述

大数据已经走进了我们的生活且成为了整个社会的关注热点。大数据究竟是什么？有哪些相关技术？对普通人的生活会有怎样的影响？大数据未来的发展趋势如何？本章将一一介绍这些问题。

1.1.1 大数据的基本概念

早在1980年，著名未来学家阿尔文·托夫勒便在《第三次浪潮》一书中，将大数据热情地赞颂为“第三次浪潮的华彩乐章”。大数据或称巨量资料，指的是所涉及的资料规模大到无法通过当地主流软件和硬件工具，对其进行实时撷取、管理、处理并整理成为帮助企业经营决策的信息。

从技术层面上看，大数据无法用单台计算机进行处理，而必须采用分布式计算架构。其特色在于对海量数据的挖掘，但它又必须依托一些现有的数据处理方法，如云处理、分布式数据库、云存储与虚拟化技术。

互联网是大数据的主要载体之一，可以说没有互联网就没有大数据。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。此外，数据并非单纯指人们在互联网上发布的信息，全世界的工业设备、汽车、电表上有着无数的数码传感器，随时测量和传递着有关位置、运动、震动、温度、湿度乃至空气中化学物质的变化，必然会产生海量的数据信息。

大数据的意义在于可以通过人类日益普及的网络行为附带生成，并被相关部门、企业所采集，蕴涵着数据生产者的真实意图、喜好，其中包括传统结构和非传统结构的数据。

从海量数据中“提纯”出有用的信息，对网络架构和数据处理能力而言无疑是巨大的挑战。在经历了几年的批判、质疑、讨论、炒作之后，人们终于迎来了大数据时代。2012年3月22日，奥巴马政府宣布投资2亿美元拉动大数据相关产业发展，将“大数据战略”上升为国家战略。大数据将成为信息社会未来的“新能源”。

大数据的核心在于为客户从数据中挖掘出蕴藏的价值，而不是软硬件的堆砌。因此，针对不同领域的大数据应用模式、商业模式的研究和探索将是大数据产业健康发展的关键。

1.1.2 IT产业的发展简史

IT产业的发展阶段如图1-1所示，每一个阶段都是由新兴的IT供应商主导的。他们改变了已有的秩序，重新定义了计算机的规范，并为进入IT领域的新纪元铺平了道路。

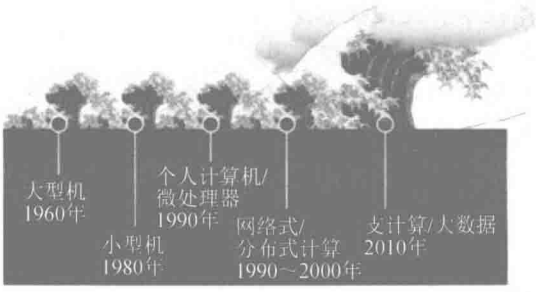

图1-1 IT产业的发展阶段

20世纪60年代和70年代的大型机阶段是以Burroughs、Univac、NCR、Control Data和Honeywell等公司为首的。在步入80年代后，小型机涌现出来，这时为首的公司包括

DEC、IBM、Data General、Wang、Prime 等。

在 20 世纪 90 年代，IT 产业进入了微处理器或个人计算机阶段，领先者为 Microsoft、Intel、IBM 和 Apple 等公司。从 90 年代中期开始，IT 产业进入了网络化阶段。如今，全球在线的人数已经超过了 10 亿，这一阶段由 Cisco、Google、Oracle、EMC、Salesforce.com 等公司领导。IT 产业的下一个阶段还没有正式命名，人们更愿意称其为云计算/大数据阶段。

数字信息每天在无线电波、电话电路和计算机电缆中川流不息。我们周围到处都是数字信息，在高清电视机上看数字信息，在互联网上听数字信息，自己也在不断制造新的数字信息。例如，每次用数码照相机拍照后，都产生新的数字信息；通过电子邮件把照片发给朋友和家人，又制造了更多的数字信息。

不过，没人知道这些流式数字信息有多少，增加速度有多快，其激增意味着什么。

正如中国人在发明文字前就有了阴阳学说，并用其解释包罗万象的宇宙世界一样，西方人用制造、获取和复制的所有 1 和 0 组成了数字世界。人们通过拍摄照片和共享音乐制造了大量的数字信息，而公司则组织和管理对这些数字信息的访问、存储，并为其提供强有力的安全保障。

目前世界上有三种类型的主要模拟数字转换，为数字信息量的增长提供动力和服务：胶片影像拍摄转换为数字影像拍摄、模拟语音转换为数字语音及模拟电视转换为数字电视。从数码照相机、可视电话、医用扫描仪到保安摄像头，全世界有 10 亿多台设备在拍摄影像，这些影像成为数字海洋中最大的组成部分，通过互联网、企业内部网在个人计算机（PC）、服务器及数据中心中复制，通过数字电视和数字投影银幕播放。

2007 年是有史以来人类创造的信息量第一次在理论上超过可用存储空间总量的一年。然而，这并不可怕，调查结果强调现在人类应该也必须合理调整数据的存储和管理。

IDC 和 EMC 都认为数字信息量的增长是因为网络应用的不断增长及人类开始将物理数据转化为数字格式的数据所致。被存储下来的数据从本质上说已经发生了重大的变化，数字化数据总量增长得很快。在 30 多年前，通信行业的数据大部分还是结构化数据。如今，多媒体技术的普及导致非结构化数据（如音乐和视频等的数量）出现爆炸式增长。虽然 30 多年前的一个普通企业用户文件也许表现为数据库中的一排数字，但是如今的类似普通文件可能包含许多数字化图片和文件的影像或者数字化录音内容。现在，92%以上的数字信息都是非结构化数据。在各组织和企业中，非结构化数据占到了所有信息数据总量的 80%以上。

另外，可视化是引起数字世界急速膨胀的主要原因之一。由于数码照相机、数码监控摄像机和数字电视内容的加速增长及信息的大量复制趋势，使得数字世界的容量和膨胀速度超过此前估计。

IDC 的数字世界白皮书指出，个人日常生活的“数字足迹”也大大刺激了数字世界的快速增长。通过互联网及社交网络、电子邮件、移动电话、数码照相机和在线信用卡交易等多种方式，每个人日常生活都在被“数字化”。

大数据快速增长的部分原因归因于智能设备的普及，如传感器、医疗设备及智能建筑（如楼宇和桥梁）。此外，非结构化信息，如文件、电子邮件和视频，将占到未来 10 年新生数据的 90%。非结构化信息的增长部分应归因于高宽带数据的增长，如视频。

用户手中的手机和移动设备是数据量爆炸的一个重要原因。手机用户总数在2015年将超过75亿。

对于地球上每一个普通居民而言，大数据有什么应用价值呢?只要看看周围正在变化的一切，你就可以知道，大数据对每个人的重要性不亚于人类初期对火的使用。大数据让人类对一切事物的认识回归本源，其通过影响经济生活、政治博弈、社会管理、文化教育科研、医疗、保健、休闲等行业，与每个人产生密切的联系。

大数据时代已悄然来到我们身边，并渗透到我们每个人的日常生活之中，谁都无法回避。它提供了光怪陆离的全媒体、难以琢磨的云计算、无法抵御的虚拟仿真环境和随处可在的网络服务。大数据是互联网的产物，即互联网是大数据的载体和平台；同时大数据让互联网生机无限。随着互联网技术的蓬勃发展，我们一定会迎来大数据的智能时代，即大数据技术和生活紧密相连，它再也不仅仅是人们津津乐道的一种时尚，而是成为生活上的向导和助手。

1.1.3 大数据的来源

大数据的来源非常多，如信息管理系统、网络信息系统、物联网系统、科学实验系统等，其数据类型包括结构化数据、半结构化数据和非结构化数据。

1）信息管理系统：企业内部使用的信息系统，包括办公自动化系统、业务管理系统等。信息管理系统主要通过用户输入和系统二次加工的方式产生数据，其产生的大数据大多数为结构化数据，通常存储在数据库中。

2）网络信息系统：基于网络运行的信息系统，网络信息系统是大数据产生的重要方式，如电子商务系统、社交网络、社会媒体、搜索引擎等都是常见的网络信息系统。网络信息系统产生的大数据多为半结构化或非结构化的数据，与信息管理系统的区别在于信息管理系统是内部使用的，不连到外部的公共网络上，而网络信息系统是指在国际互联网上，用以收集、处理、存储、分发信息的相互关联的组件的集合，其作用在于支持组织的决策与控制。在本质上，网络信息系统是信息管理系统的延伸，专属于某个领域的应用，具备某个特定的目的。因此，网络信息系统有着更独特的应用。

3）物联网系统：通过传感器获取外界的物理、化学、生物等数据信息。

4）科学实验系统：主要用于科学技术研究，可以由真实的实验产生数据，也可以通过模拟方式获取仿真数据

1.1.4 大数据的产生过程

从数据库技术诞生以来，产生大数据的方式主要经过了三个发展阶段。

（1）被动式生成数据

数据库技术使得数据的保存和管理变得简单，业务系统在运行时产生的数据可以直接保存到数据库中，由于数据是随业务系统运行而产生的，因此该阶段所产生的数据是被动的。

（2）主动式生成数据

物联网的诞生，使得移动互联网的发展大大加速了数据的产生概率。例如，人们可以

通过手机等移动终端，随时随地产生数据。用户数据不但大量增加，同时用户还主动提交自己的行为，和自己的社交圈进行了实时互动，因此产生出来大量的数据，且其具有极其强烈的传播性。显然如此生成的数据是主动的。

（3）感知式生成数据

物联网的发展使得数据的生成方式得以彻底改变。遍布在城市各个角落的摄像头等数据采集设备源源不断地自动采集并生成数据。

1.1.5 大数据的特点

在大数据背景下，数据的采集、分析、处理较之传统方式有了颠覆性的改变，如表 1-1 所示。

表 1-1 传统数据与大数据的特点比较

项　目	传 统 数 据	大　数　据
数据产生方式	被动采集数据	主动生成数据
数据采集密度	采样密度较低，采样数据有限	利用大数据平台，可对需要分析事件的数据进行密度采样，精确获取事件全局数据
数据源	数据源获取较为孤立，不同数据之间的整合难度较大	利用大数据技术，通过分布式技术、分布式文件系统、分布式数据库等技术对多个数据源获取的数据进行整合处理
数据处理方式	大多采用离线处理方式，对生成的数据集中分析处理，不对实时产生的数据进行分析	较大的数据源、响应时间要求低的应用可以采取批处理方式集中计算；响应时间要求高的实时数据处理采用流处理的方式进行实时计算，并通过对历史数据的分析进行预测分析

1.1.6 大数据处理的基本流程

大数据的处理流程可以定义为在适合工具的辅助下，对异构数据源进行抽取和集成，结果按照一定的标准统一存储，利用合适的数据分析技术对存储的数据进行分析，从中提取有益的知识并利用恰当的方式将结果展示给终端用户。大数据处理的基本流程如图 1-2 所示。

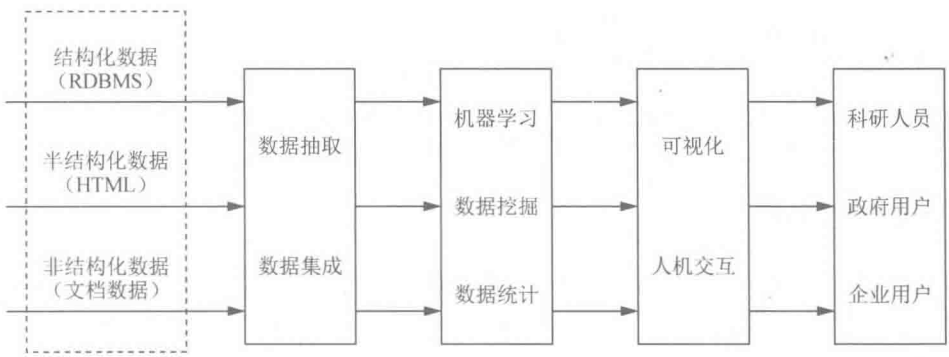

图 1-2 大数据处理的基本流程

1. 数据抽取与集成

由于大数据处理的数据来源类型广泛，而其第一步是对数据进行抽取和集成，从中找出关系和实体，经过关联、聚合等操作，再按照统一的格式对数据进行存储，现有的数据抽取和集成引擎有三种：基于物化或 ETL 方法的引擎、基于中间件的引擎、基于数据流方法的引擎。

2. 数据分析

数据分析是大数据处理流程的核心步骤。通过抽取和集成环节，从异构的数据源中获得用于大数据处理的原始数据，用户根据需求对数据进行分析处理，如数据挖掘、机器学习、数据统计，数据分析可以用于决策支持、商业智能、推荐系统、预测系统等。

3. 数据解释

用户最关心的是数据处理的结果及以何种方式在终端上显示结果，因此采用什么方式展示处理结果非常重要。就目前来看，可视化和人机交互是数据解释的主要技术。使用可视化技术可以将处理结果通过图形方式直观地呈现给用户，如标签云、历史流、空间信息等；人机交互技术可以引导用户对数据进行逐步分析，参与并理解数据分析的结果。

1.1.7 大数据的数据结构类型

从 IT 角度来看，信息结构类型大致经历了三个阶段。必须注意的是，旧的阶段仍在不断发展，因此三种数据结构类型一直存在，只是其中一种结构类型往往主导其他结构。

1）结构化信息：这种信息可以在关系数据库中找到，多年来一直主导着 IT 应用，是关键任务 OLTP 系统业务所依赖的信息。另外，这种信息还可对结构数据库信息进行排序和查询。

2）半结构化信息：包括电子邮件、文字处理文件及大量保存和发布在网络上的信息。半结构化信息是以内容为基础的，可以用于搜索，这也是 Google（谷歌）等搜索引擎存在的理由。

3）非结构化信息：该信息在本质形式上可认为主要是位映射数据。数据必须处于一种可感知的形式中（如可在音频、视频和多媒体文件中被听或看到）。许多大数据都是非结构化的，其庞大的规模和复杂性需要高级分析工具来创建或利用一种更易于人们感知和交互的结构。

1.1.8 大数据的特征

大数据分析常和云计算联系到一起，因为实时的大型数据集分析需要像 MapReduce 那样的框架来向数十、数百或甚至数千的计算机分配工作。简言之，从各种各样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。

大数据呈现出“4V，1O”的特征，具体如下。

1）数据量（Volume）大是大数据的首要特征，包括采集、存储和计算的数据量非常大。

大数据的起始计量单位至少是 100TB。通过各种设备产生的海量数据，其数据规模极为庞大，远大于目前互联网上的信息流量，PB 级别将是常态。

2）多样化（Variety）表示大数据种类和来源多样化，具体表现为网络日志、音频、视频、图片、地理位置信息等多类型的数据，多样化对数据的处理能力提出了更高的要求，编码方式、数据格式、应用特征等多个方面都存在差异性，多信息源并发形成大量的异构数据。

3）数据价值（Value）密度化表示大数据价值密度相对较低，需要很多的过程才能挖掘出来。随着互联网和物联网的广泛应用，信息感知无处不在，信息量大，但价值密度较低。如何结合业务逻辑并通过强大的机器算法挖掘数据价值，是大数据时代最需要解决的问题。

4）速度（Velocity）快、时效高。随着互联网的发展，数据的增长速度非常快，处理速度也较快，时效性要求也更高。例如，搜索引擎要求几分钟前的新闻能够被用户查询到，个性化推荐算法要求实时完成推荐，这些都是大数据区别于传统数据挖掘的显著特征。

5）数据是在线的（On-Line）表示数据必须随时能调用和计算，这是大数据区别于传统数据的最大特征。现在谈到的大数据不仅大，更重要的是数据是在线的，这是互联网高速发展的特点和趋势。例如，对于好大夫在线系统，患者的数据和医生的数据都是实时在线的，这样的数据才有意义。如果把他们放在磁盘中或者是离线的，显然这些数据远远不及在线的商业价值大。

总之，大数据时代已经到来，并快速渗透到每个职能领域，借助大数据持续创新发展，使企业成功转型，具有非凡的意义。

1.1.9 大数据的应用领域

大数据在社会生活的各个领域得到了广泛的应用，如科学计算、金融、社交网络、移动数据、物联网、医疗、网页数据、多媒体、网络日志、RFID（无线射频识别）传感器、社会数据、互联网文本和文件、互联网搜索索引、呼叫详细记录、天文学、大气科学、基因组学、生物，以及其他复杂或跨学科的科研、军事侦察、医疗记录、摄影档案馆视频档案、大规模的电子商务等。不同领域的大数据应用具有不同特点，其相应时间、稳定性、精确性的要求各不相同，解决方案也层出不穷，其中最具代表性的有 Informatica Cloud 解决方案、IBM 战略、Microsoft 战略、京东框架结构等，对此我们将在后续章节中讨论。

1.2 大数据技术架构

各种各样的大数据应用迫切需要新的工具和技术来存储、管理和实现商业价值。新的工具、流程和方法支撑起了新的技术架构，使企业能够建立、操作和管理这些超大规模的数据集和数据存储环境。

在全新的数据增长速度条件下，一切都必须重新评估。这项工作必须从全盘入手，并考虑大数据分析要容纳数据本身，IT 基础架构必须能够以经济的方式存储比以往数量更大、类型更多的数据，此外还必须能适应数据变化的速度。数量如此大的数据难以在当今的网

络连接条件下快速地移动，因此大数据基础架构必须具有分布计算能力，以便能在接近用户的位置进行数据分析，减少跨越网络所引起的延迟。

企业逐渐认识到必须在数据驻留的位置进行分析，提升计算能力，以便为分析工具提供实时响应。考虑到数据速度和数据量，移动数据进行处理是不现实的，然而，计算和分析工具可以被移到数据附近。

另外，云计算模式对大数据的成功至关重要。云模型在从大数据中提取商业价值的同时也在“驯服”它。这种交付模型能为企业提供一种灵活的选择，以实现大数据分析所需的效率、可扩展性、数据便携性和经济性，但仅仅存储和提供数据还不够，必须以新方式合成、分析和关联数据，才能提供商业价值。部分大数据方法要求处理未经建模的数据，因此，可以用毫不相干的数据源比较不同类型的数据和进行模式匹配，从而使大数据的分析能以新视角挖掘企业传统数据，并带来传统上未曾分析过的数据洞察力。基于上述考虑，一般我们可以构建出适合大数据的四层堆栈式技术架构，如图 1-3 所示。

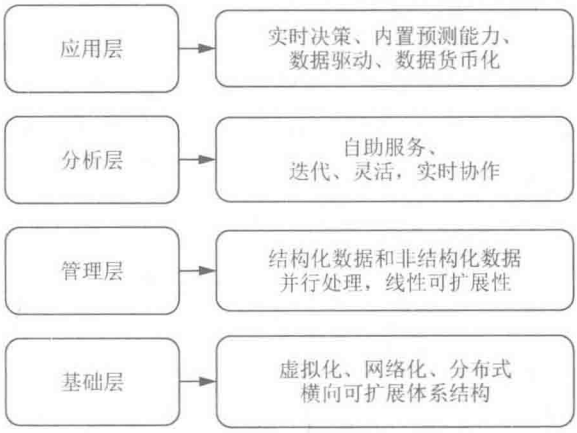

图 1-3　四层堆栈式技术架构

1. 基础层

基础层是整个大数据技术架构基础的最底层。要实现大数据规模的应用，企业需要一个高度自动化的、可横向扩展的存储和计算平台。这个基础设施需要从以前的存储孤岛发展为具有共享能力的高容量存储池。容量、性能和吞吐量必须可以线性扩展。

云模型鼓励访问数据并通过提供弹性资源池来应对大规模问题，解决了如何存储大量数据及如何积聚所需的计算资源来操作数据的问题。在云中，数据跨多个结点调配和分布，使数据更接近需要它的用户，从而缩短响应时间，提高效率。

2. 管理层

大数据要支持在多源数据上做深层次的分析，在技术架构中需要一个管理平台，即管理层，它使结构化和非结构化数据管理为一体，具备实时传送和查询、计算功能。本层既包括数据的存储和管理，也涉及数据的计算。并行化和分布式是大数据管理平台所必须考虑的要素。

3. 分析层

大数据应用需要大数据分析。分析层提供基于统计学的数据挖掘和机器学习算法，用于分析和解释数据集，帮助企业获得深入的数据价值。可扩展性强、使用灵活的大数据分析平台更可成为数据科学家的利器，起到事半功倍的效果。

4. 应用层

大数据的价值体现在帮助企业进行决策和为终端用户提供服务的应用。不同的新型商业需求驱动了不同大数据的应用。反之，大数据应用为企业提供的竞争优势使企业更加重视大数据的价值。新型大数据的应用不断对大数据技术提出新的要求，大数据技术也因此在不断的发展变化中日趋成熟。

1.3 大数据的整体技术和关键技术

大数据需要特殊的技术，以有效地处理在允许时间范围内的大量数据。适用于大数据的技术包括大规模并行处理（MPP）数据库、数据挖掘电网、分布式文件系统、分布式数据库、云计算平台、互联网和可扩展的存储系统。

大数据技术分为整体技术和关键技术两个方面。

1. 整体技术

大数据的整体技术一般包括数据采集、数据存取、基础架构、数据处理、统计分析、数据挖掘、模型预测和结果呈现等。

1）数据采集：ETL 工具负责将分布的、异构数据源中的数据如关系数据、平面数据文件等抽取到临时中间层后进行清洗、转换、集成，最后加载到数据仓库或数据集市中，成为联机分析处理、数据挖掘的基础。

2）数据存取：关系数据库、NoSQL、SQL 等。

3）基础架构：云存储、分布式文件存储等。

4）数据处理：自然语言处理（Natural Language Processing，NLP）是研究人与计算机交互的语言问题的一门学科。处理自然语言的关键是要让计算机“理解”自然语言，所以自然语言处理又称为自然语言理解（Natural Language Understanding，NLU），也称计算语言学（Computational Linguistics）。一方面它是语言信息处理的一个分支，另一方面它是人工智能（Artificial Intelligence，AI）的核心课题之一。

5）统计分析：假设检验、显著性检验、差异分析、相关分析、T 检验、方差分析、卡方分析、偏相关分析、距离分析、回归分析、简单回归分析、多元回归分析、逐步回归、回归预测与残差分析、岭回归、Logistic 回归分析、曲线估计、因子分析、聚类分析、主成分分析、因子分析、快速聚类法与聚类法、判别分析、对应分析、多元对应分析（最优尺度分析）、Bootstrap 技术等。