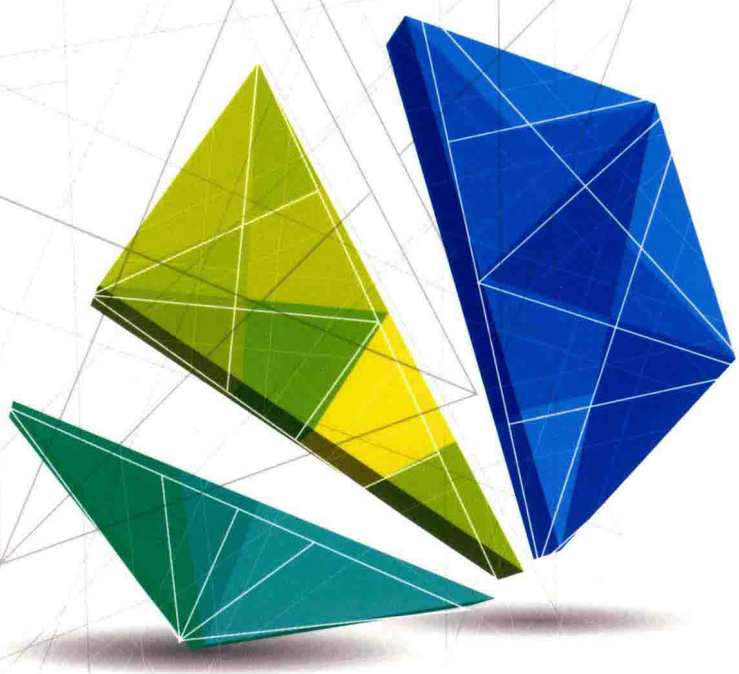
高等学校大数据技术与应用规划教材

大数据应用基础

DASHUJU YINGYONG JICHU

娄 岩 主编



中国铁道出版社
CHINA RAILWAY PUBLISHING HOUSE

技术与应用规划教材

大数据应用基础

主　编　娄　岩
副主编　徐东雨
编　委　郑琳琳　刘尚辉　李　静
　　马　瑾　丁　林　曹　阳
　　庞东兴　张志常　霍　妍

中国铁道出版社
CHINA RAILWAY PUBLISHING HOUSE

内 容 简 介

本书是将大数据基本理论与基本应用有机结合的教材，按照定义、特征、技术流程和典型案例分析的方式编写，抽丝剥茧，由易到难，有助于读者理解和掌握大数据技术。

本书的一大亮点是每章中都使用图表对大数据与传统数据处理方式进行对比。另外，本书注重启发式的学习策略，便于读者理解和掌握。全书在每一章均附有实际应用案例与关键词注释，方便读者查阅和自学，同时配备了习题和参考答案。

本书适合作为普通高校大数据技术的基础教材，也可以作为职业培训教育及相关技术人员的参考用书。

图书在版编目(CIP)数据

大数据应用基础/娄岩主编. —北京：中国铁道出版社，2018.10

高等学校大数据技术与应用规划教材

ISBN 978-7-113-24854-3

Ⅰ.①大… Ⅱ.①娄… 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(2018)第 235759 号

书　　名：大数据应用基础

作　　者：娄　岩　主编

策　　划：周海燕　　**读者热线：(010) 63550836**

责任编辑：周海燕　徐盼欣

封面设计：穆　丽

责任校对：张玉华

责任印制：郭向伟

出版发行：中国铁道出版社（100054，北京市西城区右安门西街 8 号）

网　　址：http://www.tdpress.com/51eds/

印　　刷：三河市宏盛印务有限公司

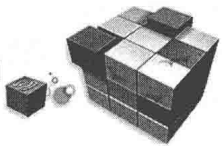
版　　次：2018 年 10 月第 1 版　2018 年 10 月第 1 次印刷

开　　本：787 mm × 1 092 mm　1/16　印张：10.5　字数：232 千

书　　号：ISBN 978-7-113-24854-3

定　　价：32.00 元

前 言



习近平总书记在党的十九大报告中提出要“推动互联网、大数据、人工智能和实体经济深度融合”，强调“贯彻新发展理念，建设现代化经济体系”。大数据、VR（虚拟现实）、AR（增强现实）和人工智能等信息技术必将为社会发展和时代进步注入新的生机和血液。

为此，本书围绕大数据应用，从理论、相关技术和实际应用三个层面进行简明扼要的阐述，目的是让广大师生对大数据的应用方法和相关知识有所了解，更好地把握科学发展的方向。

大数据技术教学在中国医科大学已经连续开展五年，已经成为大学计算机教育的重要组成部分。为国家培养了一批掌握最新 IT 发展动态和技能的医学人才，同时也积累了一定的教学经验。

在编写原则上，本书注重知识的系统性、针对性、理论性和应用性。本书倡导启发式的学习策略，通过案例启发学生的学习兴趣，检验其学习效果，提高其学习能力。

本书内容包括 12 章：第 1 章大数据概论主要讲解了大数据技术概念、架构、整体技术；第 2 章大数据采集及预处理主要讲解了大数据采集的概念、数据来源和技术方法；第 3 章大数据分析概论主要讲解了大数据分析的方法、流程、主要技术；第 4 章大数据可视化主要讲解了大数据可视化的过程和可视化工具 Tableau；第 5 章 Hadoop 概论主要讲解了 Hadoop 的架构；第 6 章 HDFS 和 Common 概论主要讲解了 HDFS 的体系结构、工作原理和 Common 模块；第 7 章 MapReduce 概论主要讲解了 MapReduce 的架构、原理和工作流程；第 8 章 NoSQL 概论主要讲解了 NoSQL 的基本知识和典型工具；第 9 章 Spark 概论主要讲解了 Spark 生态系统的组成；第 10 章云计算与大数据主要讲解了云计算的服务模式、部署模式；第 11 章典型大数据解决方案主要讲解了各种大数据解决方案；第 12 章大数据应用案例分析（医疗领域）主要讲解了大数据在医疗领域的应用案例。

本书由娄岩任主编，由徐东雨任副主编，郑琳琳、刘尚辉、李静、马瑾、丁林、曹阳、庞东兴、张志常、霍妍参与编写。具体编写分工如下：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章由丁林编写，第7章由徐东雨编写，第8章由曹阳编写，第9章由庞东兴编写，第10章由张志常编写，第11章、第12章由霍妍编写。

中国铁道出版社对本书的出版做了充分论证，精心策划。在此向所有参加编写的同事们、帮助和指导过我们工作的朋友们和参考文献的作者前辈们表示衷心的感谢！

由于编者水平有限，加之时间仓促，书中难免存在疏漏之处，恳请广大读者批评斧正！

娄　岩

2018年6月

目 录

第1章 大数据概论

>>>导学

【内容与要求】

本章主要对大数据的技术架构、大数据的整体技术、大数据分析的四种典型工具以及大数据未来发展趋势进行介绍，使读者更好地了解什么是大数据技术。

“大数据技术简介”一节介绍IT产业的发展简史、大数据的主要来源、数据生成的三种主要方式、大数据的特点、大数据的处理流程、大数据的数据格式、基本特征和应用领域。

“大数据的技术架构”一节介绍四层堆栈式技术架构，包括基础层、管理层、分析层和应用层。

“大数据的整体技术”一节介绍数据采集、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。

“大数据分析的四种典型工具简介”一节介绍Hadoop、Spark、Storm和Apache Drill。

“大数据未来发展趋势”一节介绍数据资源化，随着大数据应用的发展，大数据资源成为重要的战略资源，数据成为新的战略制高点。

【重点与难点】

本章的重点是了解大数据的特点、特征和大数据未来发展趋势；本章的难点是了解大数据技术架构和整体技术。

大数据(Big Data)是指无法在一定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，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。

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 大数据技术简介

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”。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，如云式处理、分布式数据库、云存储与虚拟化技术等。

大数据是继物联网之后IT产业又一次颠覆性的技术变革，其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。

1.1.1 IT产业的发展简史

可以说，IT产业的每一个发展阶段都是由新兴的IT供应商主导的，虽然起因可能是由于军事方面或科学发展的需要。它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。

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。而80年代后，小型机如雨后春笋般涌现出来，为首的公司包括DEC、IBM、Data General、Wang、Prime等。

到了20世纪90年代，IT产业进入了微处理器或个人计算机阶段，Microsoft（微软）、Intel、IBM和Apple等公司成为当之无愧的领军者。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10亿，这一阶段由Cisco、Google、Oracle、EMC、Salesforce.com等公司领导。局域网、互联网和物联网等的发展方兴未艾。IT产业的下一个阶段，也就是本书将介绍的全新的IT变革还没有被正式命名，人们更愿意称其为云计算/大数据阶段。

众所周知，目前数字信息每天在无线电波、电话电路和计算机电缆等媒介中川流不息。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，我们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，都会产生新的数字信息；通过电子邮件把照片发给朋友和家人，同样制造了许多数字信息。不过，没人知道这些流式数字信息有多少，增加速度有多快，以及其激增意味着什么。

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。调查结果强调，现在人类应该也必须合理调整数据存储和管理。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。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，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,94% 以上的数字信息都是半结构化或非结构化数据。在各组织和企业中,它们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、数字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,数字世界的容量和膨胀速度前所未有。同时,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、视频、移动电话、数码照相机和在线信用卡交易等多种方式,每个人的日常生活都在被“数字化”。

大数据快速增长的原因之一是智能设备的普及,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% 。非结构化信息增长的另一个原因是由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。目前,全球手机用户共拥有 50 亿台手机,其中 20 亿台为智能手机,相当于 20 世纪 80 年代 20 亿台 IBM 的大型机在消费者手里。

大数据正在以不可阻拦的磅礴气势,与当代同样具有革命意义的最新科技进步(如虚拟现实技术、增强现实技术、纳米技术、生物工程、移动平台应用等)一起,揭开人类新世纪的序幕。

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。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。

1.1.2 大数据的主要来源

大数据的来源非常广泛,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,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。

(3)物联网系统:物联网是新一代信息技术,其核心和基础仍然是互联网,是在互联网基础上延伸和扩展的网络,其用户端延伸和扩展到了任何物品与物品之间,以进行信息交换和通信,而其具体实现是通过传感技术获取外界的物理、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。

1.1.3 数据生成的三种主要方式

从数据库技术诞生以来，产生数据的方式主要有三种。

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，数据随业务系统运行而产生，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了自己的行为，如实时发送照片、邮件和其他信息，使之进入了社交、移动时代。大量移动终端设备的出现，使用户不仅主动提交自己的行为，还和自己的社交圈进行了实时互动，因此产生了大量的数据，且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生成方式得以彻底改变。例如，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。

1.1.4 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

对比分类	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，不同数据之间添加的数据整合难度较大	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

1.1.5 大数据的处理流程

大数据的处理流程可以定义为在适合工具的辅助下，对不同结构的数据源进行汲取和集成，并将结果按照一定的标准统一存储，再利用合适的数据分析技术对其进行分析，最后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前的用户。大数据处理的基本流程如图1-1所示。

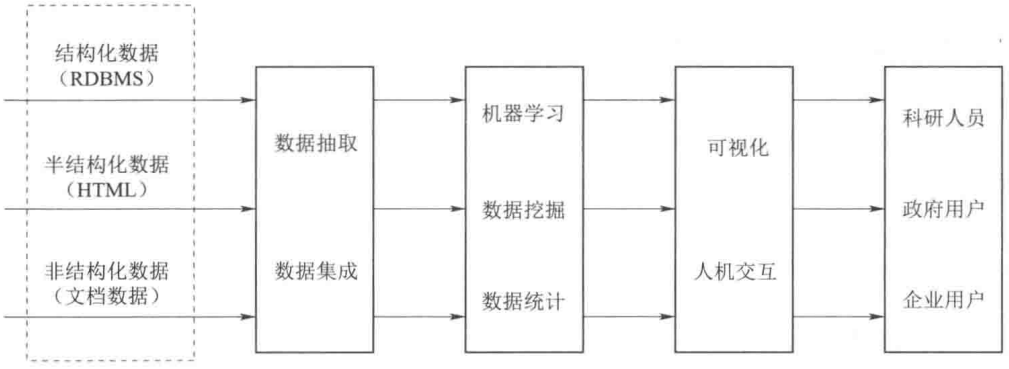


图 1-1 大数据处理的基本流程

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. 大数据分析

大数据分析是研究大型数据集的过程，其中包含各种各样的数据类型。大数据能够揭示隐藏的信息模式、未知事物的相关性、市场趋势、客户偏好和其他有用的商业信息。大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，从不同结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. 数据可视化

数据可视化主要是指借助于图形化手段，清晰有效地传达与沟通信息。数据可视化技术的基本思想，是将数据库中每一个数据项作为单个图元元素表示，大量的数据集合构成数据图像，同时将数据的各个属性值以多维数据的形式表示，可以从不同的维度观察数据，从而对数据进行更深入的观察和分析。而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结果。

1.1.6 大数据的数据格式

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，如关系数据库的使用，因此三种数据结构类型一直存在，只是在不同阶段，其中一种结构类型主导其他结构。

(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到或看到)。许多大数据都是非结构化的,其庞大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。

1.1.7 大数据的基本特征

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"4V1O"的特征,具体如下:

(1)数据量大(Volume)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 100 TB。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,其编码方式、数据格式、应用特征等多个方面都存在差异性,多信息源并发形成了大量的异构数据。

(3)数据价值密度化(Value)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。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)速度快,时效高(Velocity)。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。例如好大夫在线,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。如果把它们放在磁盘中或者是离线的,则显然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,如何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。

1.1.8 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 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、生物和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，其响应时间、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，对此将在后续章节中讨论。

1.2 大数据的技术架构

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。

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。大数据一般采用四层堆栈技术架构，如图 1-2 所示。

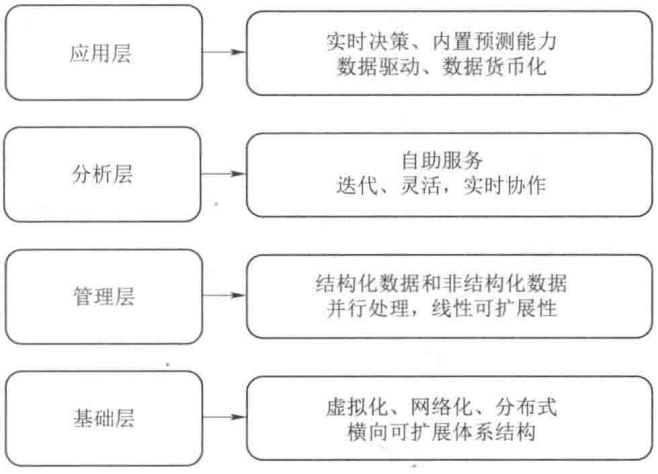


图 1-2 四层堆栈式技术架构

1. 基础层

第一层作为整个大数据技术架构基础的底层，也是基础层。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。

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领悟。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

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。它是传统方法和新的解决途径的完美结合。

(1)数据采集：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：主要指自然语言处理(Natural Language Processing，NLP)，它是研究人与计算机交互的语言问题的一门学科。

(5)统计分析：包括假设检验、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、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。

(6)数据挖掘：相对传统的数据挖掘，大数据挖掘需要挑战一些新技术，如通过分布式计算、内存计算和列存储等技术来处理大数据量情况的计算。前端展示分析和挖掘过程类似，唯一不同的是后台的高性能计算能力。

(7)模型预测：包括预测模型、机器学习、建模仿真等。

(8)结果呈现：包括云计算、标签云、关系图等。