大数据技术与应用专业规划教材

大数据技术与应用

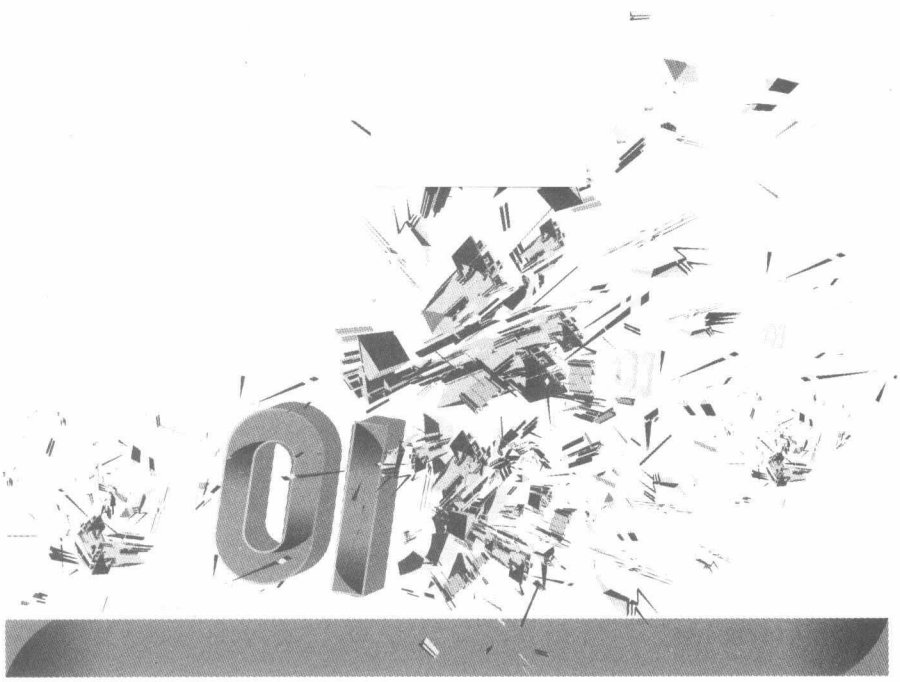
◎ 娄岩 主编

清華大學出版社

大数据技术与应用专业规划教材

大数据技术与应用

◎ 娄岩 主编



清華大學出版社
北京

内 容 简 介

本书是将大数据这一计算机前沿科学和基本应用有机结合的典范教材，全面介绍大数据和相关的基础知识，由浅入深地剖析大数据的分析处理方法和技术手段，突出介绍大数据最新的发展趋势和技术成果。

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。全书每章均包括实际应用案例与关键词注释，方便读者查阅和自学，同时配备习题和参考答案。

本书体系完整、内容丰富、注重应用、前瞻性强、适用性好，并有开放式的课程教学网站(http://www.cmu.edu.cn/computer)提供技术支持。

本书既可以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

大数据技术与应用/娄岩主编. —北京：清华大学出版社，2016
(大数据技术与应用专业规划教材)
ISBN 978-7-302-45181-5

Ⅰ. ①大…　Ⅱ. ①娄…　Ⅲ. ①数据处理　Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2016)第 239578 号

责任编辑：贾　斌　王冰飞
封面设计：刘　键
责任校对：胡伟民
责任印制：沈　露

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课 件 下 载：http://www.tup.com.cn，010-62795954

印 装 者：北京密云胶印厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：10.5　　**字　　数**：262 千字
版　　次：2016 年 11 月第 1 版　　**印　　次**：2016 年 11 月第1 次印刷
印　　数：1～2500
定　　价：32.00 元

产品编号：070869-01

本书编委会

主　　编：娄　岩

副 主 编：郑琳琳　徐东雨

编委成员（按姓氏笔画排列）：

丁　林　马　瑾　刘尚辉　张志常

李　静　庞东兴　曹　阳　霍　妍

FOREWORD 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地朝人们涌来，并将触及到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深刻。

大数据的应用激发了一场思想风暴，也悄然地改变了人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，引起工业、商业、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，各个专业的高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”：一是掌握大数据基本技术与应用，使大数据能够为我所用；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。为此，本书围绕大数据及其相关技术这一主题，采用深入浅出的叙述方式，简明扼要地阐述大数据及其相关最新技术的基本理论、关键技术和实际应用，目的是让广大师生以计算机公共基础课程为知识载体，对大数据在各个领域的应用方法和相关知识有所了解。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技术本身应有的系统性和理论性，又着重体现其在各个领域内的应用性与针对性。本书的一大亮点是每章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。全书每章均包括实际应用案例与关键词注释，方便读者查阅和自学，同时配备习题和参考答案。

全书在内容上共分成11章：第1章大数据概论由娄岩编写，第2章大数据采集及预处理由郑琳琳编写，第3章大数据分析概论由刘尚辉编写，第4章大数据可视化由李静编写，第5章Hadoop概论由马瑾编写，第6章HDFS和Common概论由丁林编写，第7章MapReduce概论由徐东雨编写，第8章NoSQL概论由曹阳编写，第9章Spark概论由庞东兴编写，第10章云计算与大数据由张志常编写，第11章典型大数据解决方案由霍妍编写。

清华大学出版社对本书的出版做了精心策划，充分论证，在此向所有参加编写的同事们及帮助和指导过我们工作的朋友们表示衷心的感谢！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

2016年9月

CONTENTS 目录

第1章

大数据概论

导学

内容与要求

本章主要涉及大数据技术简介、大数据的技术架构、大数据的整体技术、大数据分析4种典型工具及大数据未来发展趋势，以便读者更好地了解什么是大数据技术。

“大数据技术简介”一节包含IT产业的发展简史、大数据的主要来源、数据生成的3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。了解大数据的主要来源，掌握大数据的特点和大数据的处理流程。

“大数据的技术架构”一节介绍4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的整体技术。

“大数据分析的4种典型工具简介”一节介绍的工具包括Hadoop、Spark、Storm和Apache Drill。

“大数据未来发展趋势”一节中简介数据资源化。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、难点

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

大数据(Big Data)指当传统的数据挖掘和处理技术对某些数据无法进行处理时使用的过程。如数据是非结构化,时间敏感或信息量巨大,以至于无法通过关系数据库引擎进行处理的数据。这些类型的数据,需要采用不同的处理方法和实时且具有分布式处理能力的并行硬件设备。

1.1 大数据技术简介

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?本节将一一介绍这些问题。

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,虽然它们的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20世纪80年代后,小型机便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等公司成为了当之无愧的领军者。从20世纪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,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内容所描述的全新的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。人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,人们自己也在不断地制造新的数字信息,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,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。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存储和管理。如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，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。同时个人日常生活的“数字足迹”也大大地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”，数字世界的规模从2006～2011年5年间约膨胀了10倍，如图1-1所示。

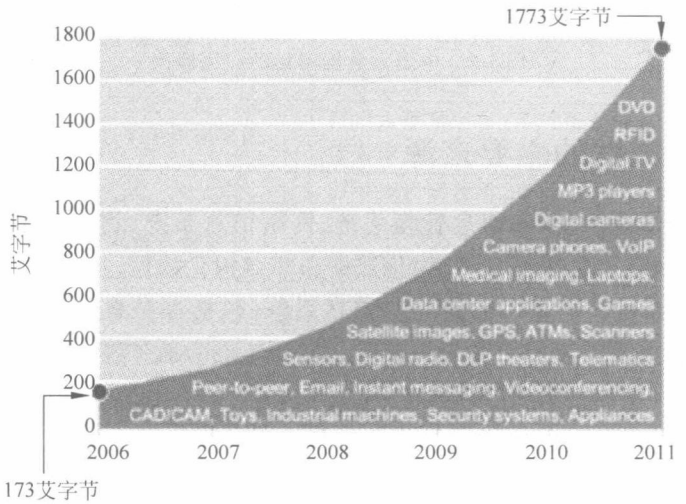


图1-1 2006～2011年全球数字信息的增长

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。

大数据时代已悄然地来到人们身边，并渗透到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，人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。中国大数据市场的应用展望如图1-2所示。

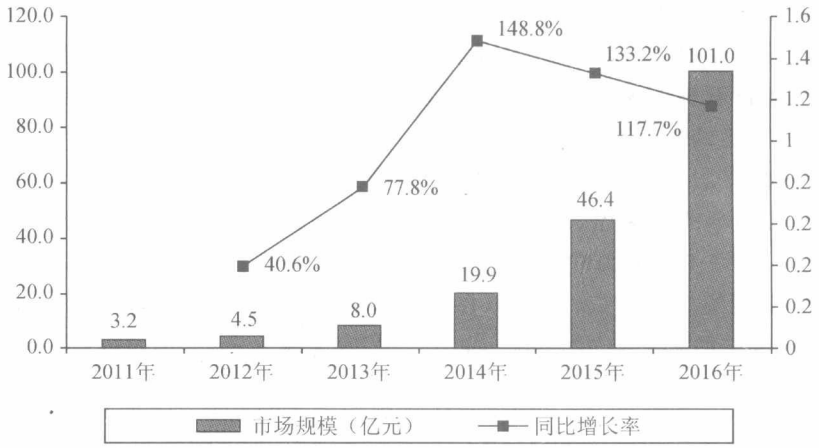


图 1-2 中国大数据市场的应用展望

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即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。

（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，是在互联网基础上的延伸和扩展的网络，其用户端延伸和扩展到了任何物品与物品之间，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、化学和生物等数据信息。

（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 3 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 3 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地加速了数据的产生几率。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，因此数据大量地产生出来，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的改变。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

	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 1-3 所示。

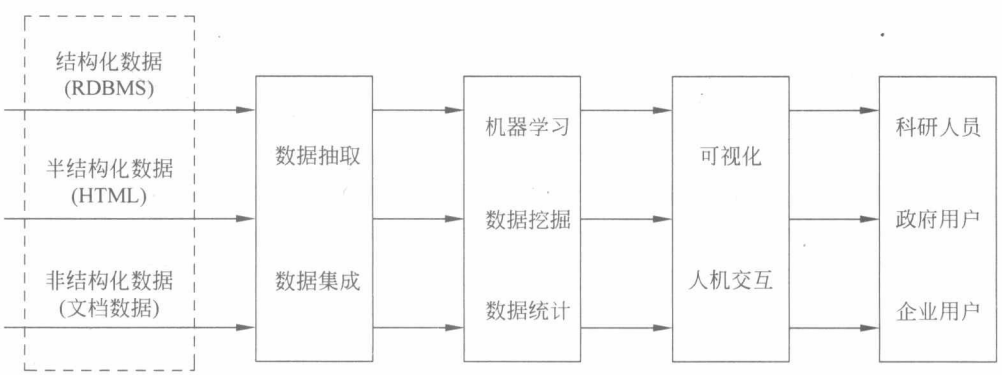


图 1-3 大数据处理的基本流程

1. 数据汲取与集成

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。现有的数据汲取和集成引擎有 3 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，其分析结果可用于更有效的市场营销、得到新的收入机会、更好的客户服务、提高运营效率、竞争优势和其他商业利益。大数据分析是大数据处理流程的核心步骤，通过汲取和集成环节，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 3 个阶段。必须注意的是，旧的阶段仍在不断地发展，如关系数据库的使用。因此 3 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(On-Line Transaction Processing，联机事物处理系统)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，快速地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)：这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，因此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，显然这些数据远远不及在线的商业价值大。

总之，无所遁形的大数据时代已经到来，并快速地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(Radio Frequency identification Devices，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Information Cloud解决方案、IBM战略、Microsoft战略、京东框架结构等，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。