普通高等教育“十三五”规划教材
高等院校数据科学与大数据专业“互联网+”创新规划教材

大数据导论

主编 | 王道平　陈　华

扫一扫联系客服

电子课件

北京大学出版社
PEKING UNIVERSITY PRESS

普通高等教育“十三五”规划教材
高等院校数据科学与大数据专业“互联网+”创新规划教材

大数据导论

主　编　王道平　陈　华

北京大学出版社
PEKING UNIVERSITY PRESS

内 容 简 介

本书系统地介绍了大数据技术与应用的基础知识，详细阐述了大数据的采集、存储、处理、分析和可视化等相关内容，并且讲述了大数据在金融、互联网、生物医学等领域的应用，同时剖析了大数据环境下的隐私和安全问题。

本书既可以作为高等院校计算机科学与技术、软件工程及相关专业课程的教材，也可以供系统分析师、系统架构师、软件开发工程师、项目经理及学习大数据技术的读者阅读和参考。

图书在版编目(CIP)数据

大数据导论/王道平，陈华主编. —北京：北京大学出版社，2019.9
高等院校数据科学与大数据专业“互联网+”创新规划教材
ISBN 978-7-301-30665-9

Ⅰ.①大… Ⅱ.①王… ②陈… 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 181016 号

书　　名	大数据导论 DA SHUJU DAOLUN
著作责任者	王道平　陈　华　主编
策划编辑	程志强
责任编辑	程志强　孙　丹
数字编辑	刘　蓉
标准书号	ISBN 978-7-301-30665-9
出版发行	北京大学出版社
地　　址	北京市海淀区成府路 205 号　100871
网　　址	http://www.pup.cn　新浪微博：@北京大学出版社
电子信箱	pup_6@163.com
电　　话	邮购部 010-62752015　发行部 010-62750672　编辑部 010-62750667
印 刷 者	北京富生印刷厂
经 销 者	新华书店
	787 毫米×1092 毫米　16 开本　10.75 印张　258 千字
	2019 年 9 月第 1 版　2019 年 9 月第 1 次印刷
定　　价	39.00 元

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有，侵权必究
举报电话：010-62752024　电子信箱：fd@pup.pku.edu.cn
图书如有印装质量问题，请与出版部联系，电话：010-62756370

此为试读，需要完整PDF请访问：www.ertongbook.com

前　言

随着互联网和信息技术的快速发展，大数据逐渐成为人们关注的焦点。我国高度重视大数据的发展与应用，目前已出台了多份国家级文件，涵盖金融业、物流业和制造业等多个行业及政务公开、审计、简政放权等多个重要领域。全球知名管理咨询公司——麦肯锡称："大数据已经渗透到当今每个行业和业务职能领域，成为重要的生产因素。"在当前大数据浪潮的猛烈冲击下，人们需要充实和完善原有的知识结构，掌握全新的科技技能，从而充分利用大数据，发挥其潜在的价值。

本书系统地介绍了大数据理论和技术，详细阐述了大数据从采集到可视化整个过程的相关内容，讲述了大数据在不同领域的应用及所面临的问题与挑战。

本书内容分 4 个部分共 8 章。

第 1 部分为大数据的相关概述（第 1 章）。主要介绍大数据的背景、概念、特征和结构类型，大数据的关键技术（包括大数据采集、预处理、分析和存储等），以及大数据的发展和应用。

第 2 部分为大数据的相关技术（第 2～6 章）。第 2 章介绍系统日志采集和网络数据采集等大数据采集技术，数据清洗、集成、变换和归约等预处理技术，数据仓库的概念、组成和数据模型，以及 ETL 技术。第 3 章介绍传统存储技术（包括 DAS、NAS 和 SAN 技术）、分布式存储和云存储技术。第 4 章介绍大数据处理与计算，分别对 Hadoop、Spark 和 Storm 处理框架进行阐述，并介绍 HDFS、MapReduce 和 YARN。第 5 章介绍描述性分析、探索性分析和验证性分析等大数据分析类型，大数据分析常用的方法（包括回归分析、关联分析、分类和聚类），以及大数据分析的工具。第 6 章介绍大数据可视化的概念、起源和作用，基于图形、像素和平行坐标法等的大数据可视化技术，以及大数据可视化的常用工具。

第 3 部分为大数据的相关应用（第 7 章）。通过结合案例，介绍大数据在金融、互联网、生物医学、物流、汽车等领域的应用，并且分析了大数据对人们日常生活的重要价值和作用。

第 4 部分为大数据的隐私与安全（第 8 章）。介绍大数据隐私与安全的定义、影响因素和分类，并分别从存储、应用和管理方面对大数据隐私与安全的防护策略进行阐述，同时介绍了相应的防护技术。

本书由北京科技大学王道平和陈华担任主编，负责设计全书结构、草拟写作提纲、组织编写工作和最后统稿。参加本书编写工作的人员还有李明芳、李锋、蒋中杨、宋雨情、徐良越、李小燕、张博卿等。

在编写本书的过程中，编者参阅了大量书籍和相关资料，在此对各位作者表示真诚的谢意！本书的出版得到了北京大学出版社的大力支持，在此一并表示衷心的感谢！由于编者水平有限，书中难免存在错误和疏漏之处，恳请广大读者批评斧正。

编　者

2019 年 7 月

【资源索引】

【电气信息类教材】

目　录

第1章 大数据概述

本章教学要点

知识要点	掌握程度	相关知识
大数据的背景	了解	互联网的三次浪潮、大数据的变革思维
大数据的概念	掌握	大数据的定义、数据存储单位
大数据的特征	掌握	大数据的4V特征
大数据的结构类型	熟悉	结构化、半结构化、准结构化和非结构化数据
大数据的关键技术	熟悉	大数据采集、预处理、储存与管理等技术
大数据的核心产业链	了解	生态商业角色的构成、生态商业模式的分析
大数据的发展	了解	大数据的发展态势及其所面临的挑战
大数据的应用	熟悉	大数据在金融、互联网等领域的应用

信息技术为人类步入人工智能社会开启了大门，带动了互联网、物联网、电子商务和网络金融等现代服务业的发展，催生了新能源、智慧交通、智慧城市和高端装备制造等新兴产业。与此同时，各种业务数据呈爆炸式增长，大数据时代已来临，传统的信息处理技术已难以满足其收集、存储、分析和应用的需求。世界各国均高度重视大数据技术的研究与发展，以期在“互联网第三次浪潮”中占得先机、引领市场。

1.1 大数据的背景

20世纪80年代以来，互联网经历了三次浪潮，相继解决了信息处理、信息传输和信息爆炸三个方面的诸多问题，促使思科、微软、亚马逊和科大讯飞等行业标杆企业的诞生，人类由此进入了大数据时代。

1.1.1 互联网的三次浪潮

根据国际商业机器公司(IMB)前CEO郭士纳的观点，IT领域每隔若干年就会迎来一次重大变革，见表1.1。

表 1.1 互联网的三次浪潮

互联网浪潮	发生时间	解决的问题	代表企业
第一次浪潮	20 世纪 90 年代	信息处理	思科、斯普林特、惠普、太阳微系统、微软和苹果等
第二次浪潮	21 世纪初	信息传输	谷歌、亚马逊、eBay、腾讯和阿里巴巴等
第三次浪潮	2010 年前后	信息爆炸	科大讯飞、百度和滴滴出行等

20 世纪 90 年代，个人计算机进入千家万户，为网络世界的到来打下了坚实的基础，人类迎来了互联网的第一次浪潮。在这个阶段，思科、斯普林特、惠普、太阳微系统、微软、苹果等公司创造的硬件、软件和网络成为人们与互联网联通的工具。21 世纪初，谷歌等搜索引擎的出现，方便了人们探索网络世界中的海量信息。亚马逊和 eBay 在互联网上推出了一站式购物模式，电子商务应运而生，社交网络此时也进入了成熟期，互联网的第二次浪潮席卷而来。2010 年前后，云计算、物联网、大数据的快速发展，拉开了互联网的第三次浪潮的序幕，大数据时代已经到来，以科大讯飞为代表的标杆企业不断涌现。

1.1.2 大数据的变革思维

大数据是人们获得新的认知、创造新的价值的源泉，是改变市场、组织机构及政府与公民关系的方法。维克托认为大数据的核心就是预测，这个核心代表着分析信息时的三个转变。第一个转变是在大数据时代需要分析更多的数据，有时甚至要处理与某个特别现象相关的所有数据，而不再依赖于随机采样；第二个转变是研究数据如此之多，以至于不再热衷于追求精确度；第三个转变由前两个转变促成，即不再热衷于寻找因果关系。

最初，需要处理的信息量过大，已经超出了一般计算机在处理数据时所能使用的内存量，因此工程师们改进处理数据的工具，促使了新的处理技术的诞生，如谷歌公司的 MapReduce 和开源 Hadoop 平台，使得人们可以处理的数据量大大增加。更重要的是，数据不再需要用传统的数据库表格来整齐地排列。这是传统数据库结构化查询语言(Structured Query Language，SQL)的要求，而非关系型数据库(NoSQL)没有这些要求，于是可以消除僵化的层次结构的一致性技术就出现了。同时，因为互联网公司可以收集到大量有价值的数据，所以成为了最新处理技术的领衔实践者。

以前，一旦完成了收集数据的工作，数据就会被认为没有太大的价值了。例如，在飞机降落之后，票价数据就失去了“价值”，能够反映重要通勤信息的数据被工作人员“自作主张”地丢弃了。也就是说，如果没有大数据的理念，很多有价值的数据就会丢失。如今，人们认为数据不再是静止和陈旧的了。数据已经成为一种商业资本、一项重要的经济投入，可以创造新的经济利益。事实上，一旦思维转变过来，数据就能被巧妙地用来激发新产品和新服务。

1.2 大数据简介

大数据不等同于数据量大的数据，它是具有一定价值的资源，确切地说，它可以为人类带来经济效益和社会效益。大数据类型繁多、处理速度快，但价值密度低，很多数

此为试读，需要完整PDF请访问：www.ertongbook.com

据无法直接使用，甚至没有分析价值。除了结构化的数据，大数据更多是半结构化、准结构化和非结构化的，这对大数据的处理和分析工作提出了很高的技术要求。

1.2.1 大数据的概念

【大数据的定义】

从经济学的角度看，大数据是经过系统整理的储存在现实或虚拟空间中，能够提供一定价值的信息资源。从会计学的层面看，这些信息资源是大数据企业或大数据研究机构通过合法交易取得的能够拥有或控制并可以带来经济利益的资产。从海量的数据规模来看，根据统计，全球 IP 流量达到 1EB 所需的时间在 2001 年是 1 年，而在 2013 年仅为 1 天，到 2016 年则仅为半天。全球新产生的数据年增 40%，信息总量每两年即可翻番。2012 年 IDC 和 EMC 联合发布的《2020 年的数字宇宙》报告指出，2011 年全球数据总量已达到 1.87ZB，如果用 DVD 光盘存储这些数据，则这些光盘排起来的长度达 8×10^5km。数据存储单位及其换算关系见表 1.2。

表 1.2 数据存储单位及其换算关系

单　位	换 算 关 系
B(Byte，字节)	1B=8bit
KB(Kilobyte，千字节)	1KB=1024B
MB(Megabyte，兆字节)	1MB=1024KB
GB(Gigabyte，吉字节)	1GB=1024MB
TB(Trillionbyte，太字节)	1TB=1024GB
PB(Petabyte，拍字节)	1PB=1024TB
EB(Exabyte，艾字节)	1EB=1024PB
ZB(Zettabyte，泽字节)	1ZB=1024EB

【大数据存储单位的换算】

大数据并不仅仅是指海量数据，更多的是指这些数据都是非结构化的、残缺的、无法用传统方法进行处理的。也正是因为应用了大数据技术，谷歌才能比政府的公共卫生部门早两周时间预告 2009 年甲型 H1N1 流感的暴发。也就是说，大数据需要量化并进行不断的开发、分析和应用。所谓量化是指从错综复杂的数据中不断地提取和整理，把现象转变成可以分析应用的形式。

1.2.2 大数据的特征

关于“大数据的特征是什么”这个问题，学术界比较认可大数据的 4V 说法：数据量大(Volume)、数据类型繁多(Variety)、处理速度快(Velocity)和价值密度低(Value)。

1. 数据量大

人类进入信息社会以后，数据以自然方式增长，其产生不以人的意志为转移。从 1986 年到 2010 年的 20 多年时间里，全球数据的数量增长了 100 倍，今后数据的增长速度会更快。预计到 2020 年，全球将拥有 35ZB 的数据量，与 2010 年相比增长近 30 倍。随着 Web 2.0 和移动互联网的迅速发展，人们已经可以随时随地发布包括博客、微博和

微信在内的各种信息。物联网也得到了飞速发展，各种传感器和摄像头几乎遍布工作和生活的各个角落，这些设备每时每刻都在自动产生大量的数据。

2. 数据类型繁多

大数据的来源众多，科学研究和 Web 应用等领域都在源源不断地生成新的数据。生物大数据、交通大数据、医疗大数据、电信大数据、电力大数据和金融大数据等呈现出“井喷式”增长，所涉及的数据数量十分巨大，已经从 TB 级跃升到 PB 级，这些数据往往被归类为结构化数据、半结构化数据和非结构化数据。与以往的结构化数据为主导地位的局面不同，如今的数据多为非结构化数据，包括网络日志、社交网络信息和地理位置信息等，对数据的处理提出了巨大的挑战。

传统的数据主要储存在关系型数据库中，但是在 Web 2.0 等应用领域中，越来越多的数据开始被储存在 NoSQL 数据库中，这就必然要求在集成的过程中进行数据转换，而这种转换过程是非常复杂和难以管理的。传统的联机分析处理(Online Analytical Processing，OLAP)和商务智能工具大多面向结构化数据，而在大数据时代，商业软件必须是用户友好的且支持非结构化数据分析的，这样才能具有广阔的应用场景。

3. 处理速度快

大数据的处理速度非常快，各种数据基本上实时在线，并能够进行快速的处理、传送和存储，以便全面反映对象的当下情况。在数据量非常庞大的情况下也能够做到数据的实时处理，可以从各种类型的数据中快速获得高价值的信息。以谷歌的 Dremel 为例，它是一种可扩展的、交互式的实时查询系统，用于嵌套数据的分析，通过结合多级树状执行过程和列式数据结构，能做到几秒内完成对万亿张表的聚合查询，也能扩展到成千上万的 CPU 上，满足谷歌众多用户操作 PB 级数据的需求，并且可以在 2～3s 内完成 PB 级数据的查询。

4. 价值密度低

大数据的价值密度相对较低，需要做很多的工作才能挖掘出有价值的信息。随着互联网和物联网的广泛应用，信息感知无处不在，在数据的海洋中不断寻找才能“淘”出一些有价值的东西，可谓“沙里淘金”。以监控视频为例，一天的记录中可能只有几秒是有价值的，但是为了安保工作的顺利进行，不得不投入大量的资金来购买各种设备，耗费大量的电能和存储空间以保存不断更新的监控数据。

有人把数据比喻为蕴藏能量的煤矿，煤炭按照性质不同分为焦煤、无烟煤、肥煤、贫煤等，而露天煤矿、深山煤矿的挖掘成本也不同。与此类似，大数据并不在“大”，而在于“有用”，价值含量、挖掘成本比数量更重要。对于很多行业而言，如何利用好大数据已成为赢得竞争的关键。

【大数据的组成部分】

1.2.3 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等，都可以称为大数据。由于数据自身的复杂性，处理大数据的首选方法是在并行计算的环境中进行大规模并行处理，这使得同

时进行并行摄取、数据装载和数据分析成为可能。多数的大数据是非结构化或半结构化的，就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 1.1 显示了四种不同结构类型的数据的增长趋势。可知，未来增长的 80%～90%的数据来自不是结构化的数据(半结构化数据、准结构化数据和非结构化数据)。虽然图 1.1 显示了 4 种不相分离的数据类型，但有时这些数据类型是可以被混合在一起的。例如，某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，其中包括典型的结构化数据，如日期/时间戳、机器类型、问题类型、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菜单输入的；非结构化数据或半结构化数据，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、技术问题和解决方案的实际通话描述、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。即使是现在，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。

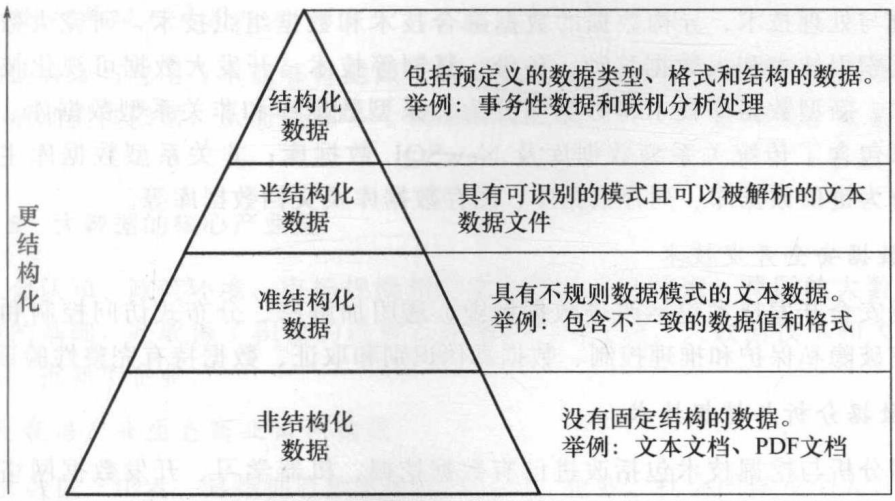

图 1.1 四种不同结构类型的数据的增长趋势

1.2.4 大数据的关键技术

【大数据的关键技术】

大数据的关键技术一般包括大数据采集技术、大数据预处理技术、大数据存储与管理技术、大数据安全开发技术、大数据分析与挖掘技术及大数据展现与应用技术等。

1. 大数据采集技术

数据采集是指通过 RFID 射频、传感器、社交网络交互及移动互联网等方式获得的结构化、半结构化、准结构化和非结构化的海量数据，是大数据知识服务模型的根本。大数据采集一般分为智能感知层和基础支撑层。智能感知层主要包括数据传感体系、网络通信体系、传感适配体系、智能识别体系及软硬件资源接入系统，实现对海量数据的智能化识别、定位、跟踪、接入、传输、信号转换、监控、初步处理和管理等；基础支撑层提供大数据服务平台所需的虚拟服务器、数据库及物联网资源等基础支撑环境。

2. 大数据预处理技术

大数据预处理主要完成对已接收数据的抽取、清洗等操作。

(1) 抽取：因获取的数据可能具有多种结构和类型，将复杂的数据转化为单一的或者便于处理的构型，以达到快速分析、处理的目的。

(2) 清洗：由于在海量数据中，数据并不全是有价值的，有些数据与所需内容无关，有些数据则是完全错误的干扰项，因此要对数据进行“去噪”，从而提取有效数据。

3. 大数据存储与管理技术

大数据存储与管理就是用存储器把采集到的数据存储起来，建立相应的数据库，并进行管理和调用。大数据存储与管理技术重点解决复杂结构化、半结构化、非结构化数据的管理与处理；主要解决大数据的存储、表示、处理、可靠性和有效传输等几个关键问题；开发可靠的分布式文件系统(Distributed File System，DFS)、能效优化的存储、计算融入存储、大数据的去冗余及高效低成本的大数据存储技术，突破分布式非关系型大数据管理与处理技术、异构数据的数据融合技术和数据组织技术，研究大数据建模技术、大数据索引技术和大数据移动、备份、复制等技术，开发大数据可视化技术和新型数据库技术。新型数据库技术将数据库分为关系型数据库和非关系型数据库。其中，关系型数据库包含了传统关系型数据库及 NewSQL 数据库；非关系型数据库主要指 No-SQL，又分为键值数据库、列存数据库、图存数据库及文档数据库等。

4. 大数据安全开发技术

大数据安全开发技术包括改进数据销毁、透明加解密、分布式访问控制和数据审计等技术，突破隐私保护和推理控制、数据真伪识别和取证、数据持有完整性验证等技术。

5. 大数据分析与挖掘技术

大数据分析与挖掘技术包括改进已有数据挖掘、机器学习、开发数据网络挖掘、特异群组挖掘和图挖掘等新型数据挖掘技术，突破基于对象的数据连接、相似性连接等大数据融合技术和用户兴趣分析、网络行为分析、情感语义分析等面向领域的大数据挖掘技术。

数据挖掘就是从大量的、不完全的、有噪声的、模糊的和随机的实际应用数据中提取出隐含在其中的，人们事先不知道但又潜在有用的信息和知识的过程。数据挖掘涉及的技术方法很多：根据挖掘任务可分为分类或预测模型发现、数据总结、聚类、关联规则发现、序列模式发现、依赖关系或依赖模型发现、异常和趋势发现等；根据挖掘对象可分为关系数据库、面向对象数据库、空间数据库、时态数据库、文本数据库、多媒体数据库、异质数据库、遗产数据库；根据挖掘方法可粗分为机器学习方法、统计方法、神经网络方法和数据库方法，机器学习方法又可细分为归纳学习方法、基于范例学习方法和遗传算法等，统计方法可细分为回归分析(多元回归、自回归等)、判别分析(贝叶斯判别、费歇尔判别、非参数判别等)、聚类分析(系统聚类、动态聚类等)和探索性分析(主元分析法、相关分析法等)等，神经网络方法可细分为前向神经网络(BP 算法等)和自组织神经网络(自组织特征映射、竞争学习等)等，数据库方法可细分为多维数据分析法、

OLAP 方法、面向属性的归纳方法。

从挖掘任务和挖掘方法的角度，数据挖掘着重突破以下几个方面。

(1) 可视化分析。无论是对普通用户还是数据分析专家，数据可视化都是最基本的功能。数据图像化可以让数据“说话”，让用户直观地看到结果。

(2) 数据挖掘算法。图像化是将机器语言翻译给人们看，而数据挖掘算法用的是机器语言，通过分割、集群、孤立点分析，可以精炼数据、挖掘价值。要求数据挖掘算法能处理大量的数据，同时应具备很高的处理速度。

(3) 预测性分析。预测性分析可以使分析师根据图像化分析和数据挖掘的结果作出前瞻性判断。

(4) 语义引擎。语义引擎需要设计足够的智能以从数据中主动地提取信息。语言处理技术包括机器翻译、情感分析、舆情分析、智能输入和问答系统等。数据质量与管理是管理的最佳实践，通过标准化流程和机器对数据进行处理，可以确保获得达到预设质量目标的分析结果。

6. 大数据展现与应用技术

大数据展现与应用技术能够将隐藏于海量数据中的信息和知识挖掘出来，为人类的社会经济活动提供依据，从而提高各个领域的运行效率，大大提高整个社会经济的集约化程度。

1.2.5 大数据的核心产业链

在社会认知、政策环境、市场规模和产业支撑能力等方面，我国的大数据产业已经具备一定的基础，并取得了积极的进展，大数据资源建设、大数据技术和大数据应用领域涌现出一批新型企业。

1. 大数据产业生态商业角色构成

(1)大数据产出者。是指拥有数据的政府机构、企业、社会团体及个人，属于大数据产业链上的基础角色，包括数据源提供者、数据流通平台提供者和数据 API 提供者。目前，我国大数据产出者包括政府管理部门、企业数据源提供商、互联网数据源提供商、物联网数据源提供商、移动通信数据源提供商、提供数据流通平台服务和数据 API 服务的第三方数据服务企业、社会团体或者个人等。

(2)大数据产品提供者。是指提供直接应用于大数据产品的企业，包括提供大数据应用软件、大数据基础软件、大数据相关硬件产品的企业。大数据应用软件产品提供者如提供整体解决方案的综合技术服务商，它们在大数据计算基础设施上(与云结合)，从简单文件存储的空间租售模式逐步扩展到提供数据聚合平台，进而扩展到为客户提供分析业务的服务。大数据基础软件提供者搭建大数据平台，提供相关大数据技术支持、云存储和数据安全等，在某些垂直行业或者区域掌握大数据的入口与出口，并能对一些数据进行采集、整合和汇集，包括传统的 IT 企业、设备商及新兴的云服务相关企业。大数据相关硬件产品提供者提供大数据采集、接入、存储、传输、安全等硬件产品和设备。

(3)大数据服务提供者。是指以大数据为核心资源、以大数据应用为主业开展商业经营的企业，包括大数据应用服务提供者、大数据分析服务提供者、大数据基础设施服务

提供者。这类企业处于大数据产业链的下游，通过挖掘隐藏在大数据中的价值，不断推动大数据产业链中各个环节的发展和成熟化。从某种角度上说，正是此类公司创造了大数据的真正价值，大数据应用服务提供者基于大数据技术，对外提供大数据服务；大数据分析服务提供者提供技术服务支持、技术(方法、商业等)咨询，或者为企业提供类似数据科学家的咨询服务；大数据基础设施服务提供者提供面向大数据技术与服务提供者的培训、咨询和推广等的基础且通用的服务。

2. 大数据产业生态商业模式分析

大数据产业拥有多元化的商业模式，并在此基础上扩展和衍生，具体包含数据买卖模式、信息服务模式、第三方数据服务模式、融合服务模式和软硬件销售模式。

(1)数据买卖模式。是指企业直接通过买卖数据取得收入。此类模式的主体是大数据经营商，业务核心是大数据的交易，发展的原动力是大数据的重复利用。这种公司具有很强大的大数据技术能力。多数情况下，大数据技术主要用于自身的运作，如通过经营大数据交易平台和大数据 API 开发盈利的互联网企业。

(2)信息服务模式。是指企业通过分析隐含在信息服务中的大数据获取利润。这类企业往往具备多种技能，甚至同时具有大数据提供者、技术提供者和服务提供者的能力。这类企业既包括传统的信息技术服务和软件服务企业，也包括咨询、审计、财务和金融等非传统意义上的 IT 企业。信息服务模式是最能表现出大数据核心产业和衍生产业相互融合的一种模式。

(3)第三方数据服务模式。是指企业既不是数据的提供者，也不是数据服务的应用者，而是专注通过提供第三方数据服务取得收入者。其主体为数据中间商，本身不具有创造数据的能力，从各种地方搜集数据进行整合，通过搭建或提供数据交易平台，从数据中提取有用信息进行交易，从而获取利润。

(4)融合服务模式。有很多企业将隐含在传统产品及服务中的数据挖掘出来以取得收入，就是在应用融合服务模式，这其中既包括提供信息服务的咨询、审计、财务等企业，也包括利用大数据在产业链上下游提供金融、物流等服务而获取利润的制造业企业。

(5)软硬件销售模式。是指各类大数据产业链企业通过直接销售服务和产品的方式获取利润。对于大数据硬件提供者和大数据基础设施服务提供者来说，软硬件销售模式是他们主要的盈利方式。

1.3 大数据的发展和应用

随着互联网的发展，大数据走进了人们生活的各个角落。世界各国都在抢抓布局，不断加大扶持力度，全球大数据的市场规模保持高速增长的态势。我国也紧跟大数据的发展趋势，大数据迅速成为我国社会各领域关注的热点，地区大数据发展格局初步形成，但同时面临着部分领域较热、数据开放发展滞后和制度建设不完善等亟待解决的问题。

1.3.1 大数据的发展态势

在 2016 年 7 月 Gartner 公司发布的新兴技术成熟度曲线中，往年备受关注的大数据

及相关技术概念并没有出现。“这些从曲线中消失的技术依然是关键，只是不再是‘新兴’的技术”，Gartner公司如此解释。随着大数据相关的基础设施、产业应用和理论体系的发展与完善，大数据越来越被各界所了解，而不像原来仅是少数科技极客眼中的“新领域”。目前，大数据以爆炸式的发展速度迅速蔓延至各行各业。总体来看，大数据进入了从概念推广到应用落地的关键转折期。

1. 大数据全球战略布局全面升级

发达国家期望通过建立大数据竞争优势，巩固其在该领域的领先地位。美国作为大数据发展的策源地和创新的引领者，最早正式发布国家大数据战略。美国政府在2012年3月发布了《大数据研究和发展倡议》（*Big Data Research and Development Initiative*），将大数据提升为一种战略性资源应用在科研、工程、教育与国家安全上。该倡议一出台便得到多个联邦部门和机构的响应。随后，美国政府又在2016年5月发布《联邦大数据研究与开发战略计划》，围绕人类科学、数据共享和隐私安全等七个关键领域部署推进大数据建设的相关计划。

之后全球各国家、组织纷纷在大数据战略推进方面积极行动。以欧洲联盟（简称欧盟）为例，其在2011年发布《开放数据：创新、增长和透明治理的引擎》后，又出台了《数据驱动经济战略》，着力开展对开放数据、云计算和数据价值链等关键领域的研究。澳大利亚、英国、日本和韩国等国家也相继推出大数据战略。澳大利亚政府于2011年5月和2013年8月先后发布《国家数字经济战略报告》（*National Digital Economy Strategy*）与《澳大利亚公共服务大数据战略》（*Australian Public Service Big Data Strategy*），为国家大数据战略发展确立了基本原则与政策指导。英国的大数据战略注重强化数据分析能力，其商务、创新和技能部在2013年10月发布了《英国数据能力发展战略规划》，对数据能力的定义和优化进行了系统的研究和指导，以大数据分析为突破点，提高国家和社会的大数据研究应用水平。日本于2012年7月发布了《面向2020年的ICT综合战略》，又于2013年出台新IT战略——《创建最顶尖IT国家宣言》，以大数据应用开发为主要战略方向，通过新技术革命带动IT产业与传统产业的协调发展，助力地区联动、民本高效、安全开放的高水平信息社会建设。同处亚洲地区的韩国也积极推行了“创意经济”计划，以孵化信息通信技术与融合领域有潜力的新兴企业和项目为抓手，推动互联网相关产业的发展。早在2011年，韩国科学技术研究院就曾提出“大数据中心战略”及“构建英特尔综合数据库”等计划，设计大数据未来发展路线。2013年，韩国政府又率先宣布建设首个对社会公众开放的全行业数据中心。

对比世界各国的大数据发展战略，可以发现三个共同点：一是政府全力推动，同时引导市场力量共同推进大数据发展；二是推动大数据在政用、商用和民用领域的全产业链覆盖；三是重视数据资源开放和管理的同时，全力抓好数据安全问题。

【实施国家大数据战略，加快建设数字中国】

2. 我国加快构建大数据战略体系

我国紧跟大数据的发展趋势，在短短几年内，大数据迅速成为我国社会各领域关注的热点。我国政府高度重视将大数据作为一种前瞻领域的战略意义，

此为试读，需要完整PDF请访问：www.ertongbook.com