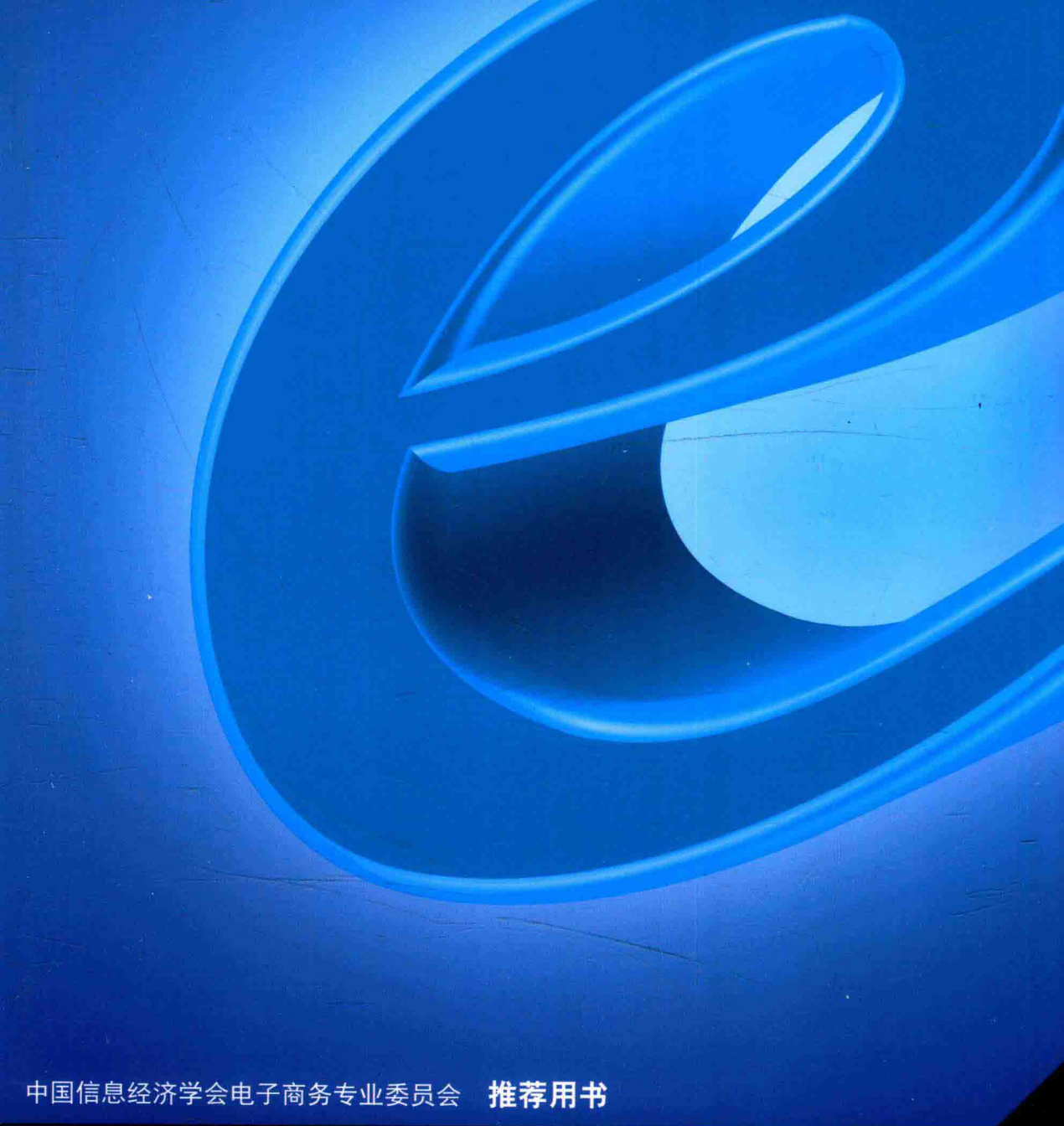

中国信息经济学会电子商务专业委员会 **推荐用书**

高等院校电子商务专业系列教材

大数据分析

王伟军 刘 蕤 周光有 编著

重庆大学出版社

中国信息经济学会电子商务专业委员会 **推荐用书**

高等院校电子商务专业系列教材

大数据分析

王伟军　刘　蕤　周光有　编著

重庆大学出版社

内容提要

本书结合大数据分析实操和商务应用场景，以大数据分析流程为主线，按照“原理、方法、工具和应用”组织内容体系，主要内容包括：大数据生态系统和大数据分析的环境搭建、大数据收集、大数据计算、大数据挖掘、大数据可视化，通过在用户搜索行为分析和个性化推荐系统两个现实场景中的实验，阐述并展示了大数据分析的环境配置和大数据分析的应用实例。

本书以附录形式呈现大数据分析实验环境搭建、Hadoop 组件参数配置，以及大数据分析相关学习资源。此外，我们还制作了配套 PPT 课件、案例、习题、试卷及答案等电子资源，以及实验所用完整数据，方便读者动手实践书中所讲解的实例。

本书适合于电子商务、信息管理与信息系统及相关专业的大学生和研究生学习，以及对大数据分析感兴趣和有志于从事数据分析工作的读者阅读使用。

图书在版编目(CIP)数据

大数据分析/王伟军，刘蕤，周光有编著.—重庆：重庆大学出版社，2017.3

高等院校电子商务专业系列教材

ISBN 978-7-5689-0043-0

Ⅰ.①大… Ⅱ.①王…②刘…③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 177632 号

高等院校电子商务专业系列教材

大数据分析

王伟军 刘 蕤 周光有 编著

策划编辑：尚东亮

责任编辑：李定群 版式设计：尚东亮

责任校对：邬小梅 责任印制：赵 晟

*

重庆大学出版社出版发行

出版人：易树平

社址：重庆市沙坪坝区大学城西路 21 号

邮编：401331

电话：(023) 88617190 88617185(中小学)

传真：(023) 88617186 88617166

网址：http://www.cqup.com.cn

邮箱：fxk@cqup.com.cn (营销中心)

全国新华书店经销

重庆升光电力印务有限公司印刷

*

开本：787mm×1092mm 1/16 印张：14 字数：332 千

2017 年 4 月第 1 版 2017 年 4 月第 1 次印刷

印数：1—3 000

ISBN 978-7-5689-0043-0 定价：39.00 元

高等院校电子商务专业系列教材编委会

顾　问

乌家培　国家信息中心专家委员会名誉主任，中国数量经济学会名誉理事长，中国信息经济学会名誉理事长，博士生导师。

祝家麟　中国计算数学学会常务理事，国家级有突出贡献的中青年专家，重庆市工业与应用数学协会会长，重庆大学原党委书记，教授、博士生导师。

孟卫东　新世纪百千万人才工程国家级人选，全国哲学社会科学领军人才，教育部新世纪优秀人才，首届教育部高等学校电子商务专业教学指导委员会委员，重庆大学副校长、教授、博士生导师。

总主编

李　琪

常务编委（以姓氏笔画为序）

王学东　陈德人　彭丽芳

编　委（以姓氏笔画为序）

于宝琴	王　晔	王伟军	王学东	王喜成	孔伟成
帅青红	司林胜	刘四青	刘业政	孙细明	李　明
李　琪	李志刚	李洪心	李陶深	杨坚争	杨路明
吴明华	张小蒂	张仙锋	张宽海	张耀辉	陈德人
赵紫剑	钟　诚	施敏华	党庆忠	秦立崴	秦成德
谢　康	琚春华	彭丽芳	董晓华	廖成林	熊　励
魏修建					

总 序

重庆大学出版社“高等院校电子商务专业本科系列教材”出版10多年来，受到了全国众多高校师生的广泛关注，并获得了较高的评价和支持。随着国内外电子商务实践发展和理论研究日新月异，以及高校电子商务专业教学改革的深入，促使我们必须把电子商务最新的理论、实践和教学成果尽可能多地反映和充实到教材中来，对教材进行全面修订更新，增补新选题，以适应新的电子商务教学的迫切需要，做到与时俱进。为此，我们于2015年启动了本套教材第3版修订和增加新编教材的工作。

从2010年以来，中国的电子商务进入新的发展阶段：规模发展与规范发展并举。电子商务三流规范发展与中国电子商务法的制定同步进行：①商流：网上销售实名制由国家工商总局负责管理；②金流：非金融支付服务资质管理由中国人民银行总行负责管理；③物流：快递业务规范管理由国家邮政局负责管理；④电子商务立法：中国电子商务法起草工作由全国人大财经委负责组织 。中共中央、国务院及多个部委陆续出台了一系列引导、支持和鼓励发展电子商务的法规和政策，极大地鼓舞了已经从事和将要从事电子商务活动的企业、行业和产业，从而推动了电子商务在我国的稳步发展。特别是李克强总理提出：“互联网+”行动计划以来，电子商务在拉动内需、促进就业和促进创业的作用正空前显现出来。全国从中央到地方多个层面和行业对电子商务的认识逐步提高，电子商务这一先进生产力正在成为我国经济社会新的发动机。

2015年7月28日人民日报报道：全国总创业者1 000余万，大学生占618余万。其中应届毕业生占第一位，回国留学生占第二位，在校大学生占第三位。2016年5月5日，中央电视台新闻报道：全国大学生就业20%由创业带动；全国就业前十大行业中互联网电子商务排名第一。中国的大学正在为中国的崛起提供源源不断的人力支持、智力支持、创新支持和创业支持，互联网、电子商务正成为就业创业的领头羊。

在教育部《普通高等学校本科专业目录(2012年)》中已经把电子商务作为一个专业类给予定义。即在学科门类：12管理学下设1 208电子商务类，120 801电子商务(注：可授管理学、经济学或工学学士学位)。2013年教育部公布了新一届高等学校电子商务类专业教学指导委员会(2013—2017年)，共由39位委员组成，是上一届21名委员的近两倍，主要充实了除教育部直属高校以外的地方和其他部委所属高校的电子商务专家代表。

截至2015年年底，全国已有400多所高校开办电子商务本科专业，1 136所高职院校开办电子商务专科专业，几十所学校有硕士生培养，十几所学校有博士生培养。全国电子商务专业在校生人数达到60多万，规模全球第一，为我国电子商务产业和相关产业发展奠定了坚实的基础。

重庆大学出版社多年来一直致力于高校电子商务教材的策划出版，得到了“全国高校电子商务专业建设协作组”“中国信息经济学会电子商务专业委员会”和“教育部高等学校电子商务类专业教学指导委员会”的大力支持和帮助，于2004年率先推出国内首套“高等院校电子商务专业本科系列教材”，并于2012年修订推出了系列教材的第2版，2015年根据教育部“电子商务类专业教学质量国家标准”和电子商务的最新发展启动了本套教材的第3版修订和选题增补，增加了新编教材14种，集中修订教材10种，电子商务教指委有14名委员参与并担任主编，2016年即将形成一个近30个教材品种、比较科学完善的教材体系。这是特别值得庆贺的事。

我们希望此套教材的第3版修订和新编，能为繁荣我国电子商务教育事业和专业教材市场，支持我国电子商务专业建设和提高电子商务专业人才培养质量发挥更好更大的作用。同时，我们也希望得到同行学者、专家、教师和同学们更好更多的意见和建议，使我们能够不断地提高本套教材的质量。

在此，我谨代表全体编委和工作人员向本套教材的读者和支持者表示由衷的感谢！

总主编　李　琪

2016年5月10日

前 言

自2011年以来，“大数据”迅速席卷全球，并成为继“云计算”“物联网”后的又一个备受关注的热词。2012年3月，美国联邦政府发布《大数据的研究和发展计划》，由美国国家自然基金会（NSF）、卫生健康总署（NIH）、能源部（DOE）、国防部（DOD）等6大部门联合启动大数据技术研发，引发了世界各国的关注。2014年，欧盟发布《数据驱动经济战略》，大数据有望为欧盟恢复经济增长和扩大就业作出贡献。2015年8月，国务院下发《促进大数据发展行动纲要》，要求全面推进我国大数据发展和应用，加快建设数据强国。2016年6月，国务院办公厅印发《关于促进和规范健康医疗大数据应用发展的指导意见》，部署通过“互联网+健康医疗”探索健康医疗大数据应用服务新模式和新业态，构建全国性的健康医疗大数据应用平台，建立起健康医疗大数据产业体系。可以说，大数据的应用已逐步深入我们生活的方方面面，涵盖医疗、交通、金融、教育、体育、零售等各行各业。尤其是企业成为大数据应用的主体，对大数据的利用将成为企业提高核心竞争力和抢占市场先机的关键。大数据正日益对全球生产、流通、分配、消费活动以及经济运行机制、社会生活方式和国家治理能力产生重要影响。深化大数据应用，已成为我国创新发展、推动产业转型升级、提升信息服务水平和政府治理能力现代化的内在需要和必然选择。

大数据应用主要通过大数据分析挖掘技术的实际应用，来获得数据的价值和预见。如美国Target公司通过“怀孕预测指数”预测高中生顾客怀孕的故事和沃尔玛“啤酒和尿不湿”的销售故事早已被人津津乐道。而当人们在微博等社交平台抒情或议论的时候，华尔街分析师正通过网站后台收集人们的记录来分析人们的情绪变化，并据此做出股票投资决策。跨国公司常通过对海量数据的分析，在全球优化供应链、指导采购和生产、制定市场营销策略等。

因此，“大数据分析”技术性强、应用范围广、成效显著，其相关技术及应用仍在迅速发展和深化中。如何编写一本适合电子商务或信息管理与信息系统专业的大数据分析相关教材确实是一个很大的挑战。从现有大数据分析相关教材的内容来看，要么是纯技术性内容体系，没有与商务应用相结合；要么偏重大数据商务应用价值或应用场景分析，缺乏必要的技术支持。我们认为：大数据分析技术是基础与工具，商务应用是本质与核心；要做好大数据商务分析工作，需要掌握基本的大数据分析技术，并能从商务应用需求出发，有目的地收集与管理数据，通过运用大数据分析相关技术和方法，发现不同数据间的数据相关性和潜在规律，获得洞察力，并最终促成决策和行动。

我们将本书定位于方法应用型教材，立足于商务应用环境，完整地阐述大数据分析流程涉及的基本原理、方法技术、操作实例和应用场景。

全书共包括8章，第1章和第2章是概述部分，主要介绍大数据的概念与价值、Hadoop生态系统与Spark生态系统，从理论层面与技术层面搭建大数据系统基本框架；第3章至第6章是分析方法部分，以大数据分析流程为主线，尽可能地结合实例，对大数据收集、大数据计算、大数据挖掘、大数据可视化进行系统性的阐述。在大数据收集一章中，介绍两种实现Hadoop数据收集的开源工具Flume和Kafka；在大数据计算一章中，介绍大数据离线计算框架Mapreduce、交互式计算框架Impala、流式计算框架Storm；在大数据挖掘一章中，梳理机器学习的主要算法，重点讲解利用Mahout、Weka和R语言进行大数据预处理与算法实现；在大数据可视化一章中，介绍Tableau和EChart两种可视化工具的基本功能和应用；第7章和第8章是实验实例部分，通过搜索日志用户行为分析和推荐系统两个应用场景，集中讲解Hadoop的环境配置、Hive的安装部署、使用Hive进行数据处理和用户信息检索行为分析、使用Mahout进行个性化推荐等实用技术与方法。

本书的特点主要体现在以下几个方面：

①本书力求涵盖目前较为成熟的大数据分析方法和工具，兼顾技术的先进性和科学性，完整地阐述了从大数据收集到大数据可视化操作这一分析流程中的实用技术、基本原理；

②本书通过上机实验和应用实例强调大数据分析实战，突出理论与实践相结合，知识与技能并重；

③本书配备有较为丰富的教学资源，每章明确提出学习目标与要求、课后附有复习思考题，以附录形式呈现大数据分析实验环境搭建、Hadoop组件参数配置等内容，并详细介绍大数据分析相关学习资源。本书还提供了配套课件以及实验所用完整数据，方便读者动手实践书中所讲解的实例。

全书由王伟军、刘蕤、周光有讨论并提出编写大纲，负责总体规划、统稿和校对工作。各章节的编写分工如下：第1章（王伟军、余跃、肖海清）、第2章（周光有、李伟卿、宁丹）、第3章（王阳、连宸、张婷婷）、第4章（刘蕤、姜毅）、第5章（刘蕤、李颖、李照东、刘辉）、第6章（池毛毛、侯银秀、张婷婷）、第7章和第8章（王伟军、周光有、黄英辉）。

此书得到国家自然科学基金项目“基于用户偏好感知的Saas服务选择优化研究”（项目编号：71271099）和“基于屏幕视觉热区的网络用户偏好提取及交互式个性化推荐研究”（项目编号：71571084）的支持，在此表示感谢！

在本书编写过程中，参考了国内外大量文献。在此，向所有参考文献的作者表示衷心的感谢！本书的编写是一次有益的探索，大数据应用是一个新兴事物，商务分析的应用实例还不够多，基于数据驱动的企业绩效优化、过程优化管理和运营科学决策的大数据分析还有待深入应用。由于编者水平有限，书中难免存在疏漏，敬请读者提出宝贵意见。

编　者

2017年1月10日

目　录

第 1 章
大数据概述

【本章学习目标与要求】

- 了解大数据的定义及结构特征。
- 熟悉大数据的来源及分类。
- 认识大数据分析的价值及影响。

大数据是以容量大、类型多、存取速度快、应用价值高为主要特征的数据集合，它正快速发展为对数量巨大、来源分散、格式多样的数据进行采集、存储和关联分析，从中发现新知识、创造新价值、提升新能力的新一代信息技术和服务业态。伴随着数据的爆炸性增长，大数据已深入社会的各行各业。但是，只有对大数据进行挖掘分析，才能获取有深度和有价值的信息。因此，大数据的分析方法在大数据发展中就显得尤为重要，可以说大数据分析是关系到最终数据信息是否有价值的决定性因素。首先让我们对大数据的概念、特征与来源，以及大数据分析的价值与影响等有一个基本的了解和认识。

1.1 大数据的背景

“大数据”是一种规模已大到难以用传统信息技术进行有效的管理，大大超出传统数据库软件工具能力范围的数据集合。《2015 年中国大数据发展调查报告》显示，2015 年中国大数据市场规模达到 115.9 亿元，增速达 38%。面对庞大的市场，各大数据企业也纷纷从中寻求商机。

自 2011 年以来，“大数据”迅速席卷全球，并成为继“云计算”“物联网”后的又一个备受关注的热词，关于它的报道和著作也层出不穷。早在 1980 年，著名的未来学家 Alvin Toffler 在其所著的《第三次浪潮》中就将“大数据”称颂为“第三次浪潮的华彩乐章”。2008 年，《Nature》杂志推出名为“Big Data”的封面专栏。2011 年 6 月，著名咨询公司麦肯锡全球研究院（MGI）发布名为《大数据：下一个创新、竞争和生产力的前沿》的研究报告，对大数据的影响、关键技术和应用领域等都进行了详尽的分析，并指出大数据将会是带动未来生产力发展和创新以及消费需求增长的指向标。2012 年 3 月，美国联邦政府发布《大数据的研究和发

展计划》,由美国国家自然基金会(NSF)、卫生健康总署(NIH)、能源部(DOE)、国防部(DOD)等6大部门联合,投资2亿美元启动大数据技术研发,引发了世界各国的关注。2012年7月,联合国在纽约发布了一本关于大数据政务的白皮书《大数据促发展:挑战与机遇》。这本白皮书总结了各国政府如何利用大数据响应社会需求、指导经济运行、更好地为人民服务,并建议成员国建立"脉搏实验室(Pulse Labs)",挖掘大数据的潜在价值。2014年,欧盟发布《数据驱动经济战略》,使大数据有望成为欧盟经济单列行业,为欧盟恢复经济增长和扩大就业作出巨大贡献。从2015年开始,我国政府对互联网、高科技和大数据产业空前重视,并明确表示要开放大数据和促进大数据发展。2015年5月,国务院发布《中国制造2025》,提出"建设重点领域制造业工程数据中心,为企业提供创新知识和工程数据的开放共享服务"。2015年8月,国务院下发《促进大数据发展行动纲要》,要求深入贯彻落实党中央、国务院决策部署,全面推进我国大数据发展和应用,加快建设数据强国。作为新的重要资源,世界各国都在加快大数据战略布局,我国已将大数据战略上升至国家层面。

随着计算机和信息技术的迅猛发展和普及应用,行业应用系统的规模迅速扩大,行业应用所产生的数据呈爆炸性增长。动辄达到数百TB甚至数十至数百PB规模的行业/企业大数据已远远超出了现有传统的计算技术和信息系统的处理能力。根据互联网数据中心(Internet Data Center,IDC)监测,人类产生的数据量正在呈指数级增长,大约每两年翻一番,并且这个速度在2020年之前会继续保持下去。这意味着人类在最近两年产生的数据量相当于之前产生的全部数据量。在百度指数中,输入关键词"大数据",进行检索得到探索"大数据"的整体趋势效果图,如图1-1所示。

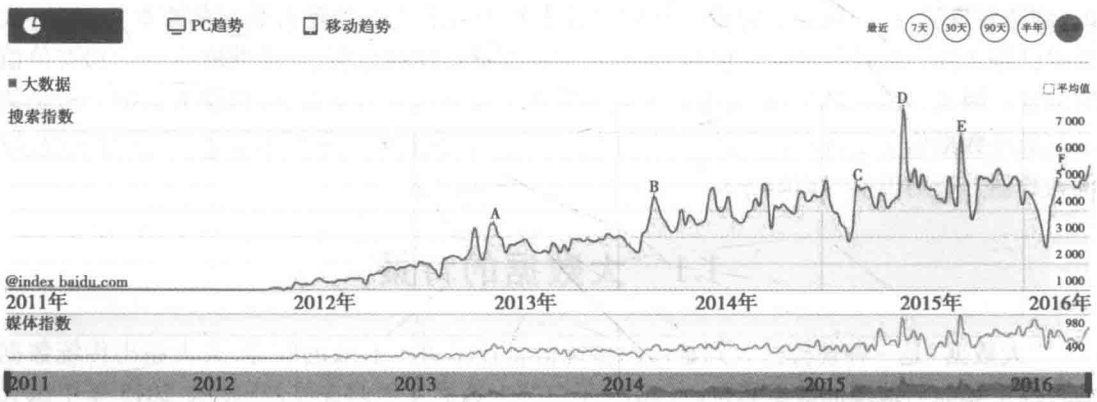

图1-1 百度指数"大数据"整体趋势图

由图1-1可知,"大数据"的搜索量自2011年以来呈现快速增长,"大数据"日趋变成大家耳熟能详的热词,并且未来的数年里,"大数据"的热度可能会持续下去。正因为大数据处理需求的迫切性和重要性,大数据技术已经在全球工业界、学术界和各国政府得到高度关注和重视,全球掀起了一个可与20世纪90年代的信息高速公路相提并论的研究热潮。

大数据的研究和分析应用具有十分重大的意义和价值。被誉为"大数据时代预言家"的维克托·迈尔-舍恩伯格在其《大数据时代》一书中列举了大量翔实的大数据应用案例,并分析、预测了大数据的发展现状和未来趋势,提出了很多重要的观点和发展思路。他认为"大数据开启了一次重大的时代转型",指出大数据将带来巨大的变革,改变我们的生活、工作和

思维方式，改变我们的商业模式，影响我们的经济、政治、科技和社会等各个层面。

1.2 大数据的基本概念

对大量数据进行分析，并从中获得有用观点，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，主要有以下3点区别：

①量大类杂，随着社交媒体和传感器网络等的发展，在我们身边正产生出大量且多样的数据。

②处理成本下降，随着硬件和软件技术的发展，数据的存储、处理成本大幅下降，数据处理环境也已经没有必要自行搭建。

③计算能力增强，随着云计算的兴起，对大数据的存储能力和处理速度大大提高。

1.2.1 大数据的定义

大数据概念的演变不仅包含了对数据集规模的描述，还包括数据利用的过程。大数据最早出现于麦肯锡全球研究院2011年发布的《大数据：下一个创新、竞争和生产力的前沿》研究报告。之后，经Gartner的宣传和2012年维克托·迈尔-舍恩伯格《大数据时代》的出版推广，大数据概念开始风靡全球。虽然大数据已成为社会热议的话题，但是到目前为止，大数据尚无统一的定义，也难以有一个定量的定义。

互联网数据中心(Internet Data Center，IDC)在报告中对大数据进行了描述：大数据是一个看起来似乎来路不明的大的动态过程。但实际上，大数据并不是一个新生事物，虽然它确实正在走向主流和引起广泛的注意。大数据并不是一个实体，而是一个横跨很多IT边界的动态活动。

麦肯锡全球研究院认为，大数据是指大小超过了典型数据库软件工具收集、存储、管理和分析能力的数据集。

Gartner公司认为，大数据就是高容量、高速和高多样化的信息资产，需要新的处理技术来增强决策能力、原理分析和流程优化。

百度百科认为，大数据是指无法在可承受的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、洞察力和流程优化能力来适应海量、高增长率和多样化的信息资产。

维基百科认为，大数据指的是所涉及的资料量规模巨大到无法通过目前主流软件工具，在合理时间内达到撷取、管理、处理并整理成为帮助企业经营决策目的的资讯。

大数据专家李国杰院士提出：大数据是指无法在可容忍的时间内用传统IT技术和软硬件工具对其进行感知、获取、管理、处理和服务的数据集合。

各个定义尽管在具体的表达中对大数据的范围、内涵等描述不一，但存在一个共识，即大数据不是对数据量大小的定量描述，重要的是在种类繁多、数量庞大的多样数据中如何进行快速的信息获取和分析，也就是如何将数据分析为信息，将信息提炼为知识，以知识促成决策和行动的过程。归根到底，大数据的最终意义在于获得洞察力和价值。

1.2.2 大数据的结构特征

关于大数据的结构特征，IBM 提出 3V，即认为大数据具备规模性（Volume）、多样性（Variety）和高速性（Velocity）3 个特征：规模性是指数据量巨大，量级达到 TB 级及 PB 级；多样性是指数据类型繁多，包括结构化数据和非结构化数据；高速性是指数据创建、处理和分析的速度持续在加快。

对大数据概念的探究涉及多个维度，在 IBM 提出 3V 的基础上，人们认为下面 4 个维度最为重要，并将其统称为 4V。

1）海量性（Volume）

大数据都是数量巨大的数据。很多企业都拥有海量数据，数据量都很容易就积累到 TB 级，甚至跃升至 PB 级。

2）多样性（Variety）

大数据冲破结构化数据的局限，不仅包括结构化数据，还覆盖了如文本、音频、视频、点击流、日志文件、地理位置信息等各种类型的半结构化和非结构化数据。

3）精确性（Veracity）

挖掘大数据价值类似沙里淘金，从海量数据中挖掘稀疏但珍贵的信息。如何处理和挖掘海量数据，以使用其价值成为至关重要的问题。

4）时效性（Velocity）

大数据对时效性要求很高，企业必须能够在短时间内高速、流畅地处理源源不断产生或流入企业的海量数据，方能最大化地显现出大数据的商业价值。以往的周、天和小时为单位的运算处理周期，下降到以分、秒为单位。同时，大数据还应被归档存储，以备不时之需。

大数据的 4V 特征为我们进行数据分析指明了方向，然而大数据的复杂性并非只体现在这 4 个维度上，还有其他因素在起作用，这些因素存在于大数据所推动的一系列过程中。在这一系列过程中，需要结合不同的技术和分析方法，才能充分揭示数据源的价值，进而用数据指导行为，促进业务发展。

1.2.3 大数据与云计算

百度公司总裁张亚勤说："云计算和大数据是一个硬币的两面，云计算是大数据的 IT 基础，而大数据是云计算的一个杀手级应用。"的确是这样，在云计算出现以前，数据大都保存在个人计算机和企业服务器中。数据存储量小，且较为分散。在云计算服务器出现后，"大数据"才有了运行轨道，逐渐发挥其真正的价值。

1）云计算与大数据的关系

有人将云计算和大数据形象地比喻为"高速公路"和"汽车"，"高速公路"的建设是为了让"汽车"快速地行使，而"汽车"的大量出现也促使了"高速公路"的快速建设。最著名的实例就是 Google 搜索引擎。面对海量 Web 数据，Google 于 2006 年首先提出云计算的概念。支撑 Google 内部各种"大数据"应用的，正是 Google 公司自行研发的云计算服务器。云计算

和大数据的关系可表现为：

(1)云计算是大数据的IT基础

云计算可构建在不同的基础平台之上，可有效兼容大数据的异构数据源。大部分的大数据环境下都采用可扩展的云存储技术，使存储性能随着存储容量的增加而得到提升。

(2)云计算是大数据成长的驱动力

大数据要面对PB级数据，因此，大数据存储系统必须能够方便、迅速地扩大存储规模，以满足数据增长的需求，在稳定的前提下让软硬件的增设变得透明。如果不以云计算进行挖掘和分析，大数据就只是僵死的数据，没有太大价值。云计算为大数据提供了解决方法。

(3)大数据是云计算的延伸

大数据技术涵盖了从数据的海量存储、处理到应用多方面的技术，包括海量分布式文件系统、并行计算框架、NoSQL数据库、实时流数据处理以及智能分析技术，如模式识别、自然语言理解、应用知识库等。

(4)大数据的信息隐私保护是云计算发展的重要前提

信息产业及服务的健康、快速发展需要安全的环境，大数据挖掘中的隐私保护为其提供了保障。

总而言之，云计算是大数据发展的前提，大数据是云计算的延伸，两者相互促进、相辅相成。大数据应用将消耗大量的计算和存储资源，这推动了云计算的普及，同时云计算的弹性能力为大数据应用程序的执行提供了保障，进而使处理大数据作业变得更为经济。

2)云计算为大数据带来的变化

纵观历史，过去的数据中心无论应用层次还是规模大小都仅仅是停留在过去有限的基础架构之上，采用的是传统精简指令集计算机和传统大型机，各个基础架构之间都相互孤立，没有形成一个统一的有机整体。因此，在这种背景下，数据中心需要向集中大规模共享平台推进，并且数据中心要能实现实时动态扩容，实现自助和自动部署服务。由此云计算、虚拟化和云存储等新IT模式出现并流行起来。云计算的出现为大数据带来了诸多变化。

(1)云计算为大数据提供了弹性扩展

云计算的出现带来了更便宜的分布式运算存储，解决了大数据的海量数据存储问题，使得中小企业也可像亚马逊一样通过云计算来完成大数据分析。

(2)云计算为大数据提供了技术保障

云计算IT资源庞大，分布较为广泛，是异构系统较多的企业及时、准确处理数据的有力方式，甚至是唯一方式。

(3)云资源的建设保障了大数据走向云计算

云资源的建设使原始数据能够迁移到云环境，资源得到了弹性扩展。数据分析集逐步扩大，企业级数据仓库将成为主流，未来还将逐步纳入行业数据、政府公开数据等多源数据。

(4)云计算使大数据逐步“云”化

通过云计算对资源进行自动调度和分配，大数据实现了一个自动部署、自动管理和自动运维的数据中心架构。数据中心逐步过渡到“云”，这其中既包括私有云，也包括公有云。

1.3 大数据的来源及分类

大数据的来源可按照数据产生主体、数据来源行业、数据存储形式进行划分。

1.3.1 按数据产生主体划分

大数据的来源按产生主体划分可分为3类:交易数据、交互数据和观测数据。

①交易数据由企业以及个人在线商品交易时产生,包括企业内部运营与管理数据,企业与企业之间、企业与个人之间以及个人与个人之间的交易数据。这类大数据一般表现为系统关系型数据库中的数据和数据仓库中的数据。

②交互数据是指人产生的大量在线交互数据,主要包括网络用户在线浏览、点击等日志数据,用户生成内容(UGC)的数据,如微博、微信产生的数据,用户评论、留言、短信、电子邮件或者电话投诉等数据。格式包括文本、图片、视频及音频等。

③观测数据是指大量机器、遥感及各类传感器产生的数据,主要包括应用服务器日志数据,科研专业机构产生的数据(如CERN的离子对撞机每秒运行产生的数据高达40 TB),传感器数据(天气、水、智能电网等),图像和视频(摄像头监控数据等),RFID、二维码或条形码扫描数据,北斗导航卫星位置数据和遥感卫星的观测数据,等等。随着物联网和智慧城市的发展,此类数据将呈爆炸式增长,大大超过前两种数据的量级。

对第一类和第二类数据,目前企业特别是互联网企业应用较多,主要应用于挖掘用户消费行为,预测特定需求和整体趋势等。但必须指出的是,第三类数据的应用将越来越重要,在科学研究、行业管理、物联网应用和智慧城市建设中必不可少,它将作为基础性资源创新商业模式和产生新的商业机会。例如,汽车传感数据用于评价司机行为会推动汽车保险业的深刻变革;农业遥感数据可用于农作物估产;北斗位置数据可用于城市交通指挥系统优化等等。

1.3.2 按数据来源行业划分

根据我国一年产生的数据总量以及大致分布情况,我国的大数据来源大体来自于以下行业:

1)以百度、阿里巴巴和腾讯(简称BAT)为代表的互联网公司

阿里巴巴目前保存的数据量为近百PB,同时拥有90%以上的电商数据。百度2013数据总量接近一千个PB,其以70%以上的搜索市场份额坐拥庞大的搜索数据。腾讯的总存储数据量经压缩处理以后在100 PB左右,并且数据量月增约10%,存储了大量的社交、游戏等领域积累的文本、音频、视频和关系型数据。

2)电信、金融、保险、电力、石化系统

电信行业拥有大量的用户上网记录、通话、信息、地理位置等数据,且年度用户数据增长约数十PB。金融与保险行业拥有大量的开户信息数据、银行网点数据、在线交易数据和自身运营的数据。在电力和石化行业,仅国家电网采集获得的数据总量就有10 PB,石油化工、