大数据导论

Big Data Fundamentals

李建伟　主编

北京邮电大学出版社
www.buptpress.com

策 划 人：马晓仟
责任编辑：马晓仟
封面设计：七星博纳

大数据导论

Big Data Fundamentals

策划中心
电　话：010-62285935
E-mail：2449868465@qq.com

ISBN 978-7-5635-5881-0

定价：42.00元

大数据导论

李建伟　主编

北京邮电大学出版社
www.buptpress.com

内容简介

本书系统地介绍了大数据技术的基础知识。本书实战环节的知识是在大数据培训的基础上总结提炼出来的，案例都为企业实际开发中的案例，所以内容的科学性和有效性已经被证实过，期望读者通过对本书的学习和对本书案例的实践，理解大数据技术的概念和原理，掌握 Hadoop 大数据技术中最基础和最重要的知识和实践。

本书的主要内容包括大数据的概念及价值，Hadoop2.0 介绍，分布式文件系统 HDFS 的原理、常用命令操作和编程实践，分布式计算框架 MapReduce 的原理、基础编程和高级编程，分布式资源管理系统 YARN 平台，分布式锁服务 ZooKeeper，Hadoop 高可用集群搭建和 Hadoop 实战项目。

本书可作为高等院校成人教育数据科学与大数据技术、计算机科学与技术和软件工程等专业的大数据课程教材，也可作为相关技术人员的参考书。

图书在版编目(CIP)数据

大数据导论 / 李建伟主编. -- 北京：北京邮电大学出版社，2019.9
ISBN 978-7-5635-5881-0

Ⅰ. ①大… Ⅱ. ①李… 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 204847 号

书　　名：大数据导论
作　　者：李建伟
责任编辑：马晓仟
出版发行：北京邮电大学出版社
社　　址：北京市海淀区西土城路 10 号(邮编：100876)
发 行 部：电话：010-62282185　传真：010-62283578
E-mail：publish@bupt.edu.cn
经　　销：各地新华书店
印　　刷：保定市中画美凯印刷有限公司
开　　本：787 mm×1 092 mm　1/16
印　　张：16.25
字　　数：422 千字
版　　次：2019 年 9 月第 1 版　2019 年 9 月第 1 次印刷

ISBN 978-7-5635-5881-0　　定价：42.00 元

前　言

当今社会是一个高速发展的社会，科技发达，信息流通，人们之间的交流越来越密切，生活也越来越方便，大数据就是这个高科技时代的产物。阿里巴巴创办人马云在一次演讲中提到，未来的时代将不是IT时代，而是DT时代，DT就是Data Technology，即数据科技。

大数据技术的迅速发展得益于计算速度越来越快、存储成本越来越低和人工智能越来越能理解数据。因此，“用数据说话”“让数据发声”已成为人类认知世界的一种全新方法。

大数据技术起源于行业应用发展，其发展速度领先于高校的人才培养速度。所以，目前大数据领域的人才缺口非常大。据2017年数联寻英发布的《大数据人才报告》显示，目前全国的大数据人才仅46万人，未来3～5年内大数据人才的缺口将高达150万人。在百度、阿里巴巴、腾讯、今日头条、美团和滴滴等大型互联网企业发布的招聘职位中，大数据相关岗位占比已经超过60%。

在如此巨大的人才需求和国家政策的鼓励下，全国普通高等院校、高职高专院校等纷纷启动大数据人才培养计划。但是，数据科学与大数据技术专业的建设面临很多困难。2016年2月，教育部公布新增“数据科学与大数据技术”专业，北京大学、对外经济贸易大学、中南大学成为首批获批高校。2017年3月，教育部公布第二批“数据科学与大数据技术”专业获批的32所高校。国内培养大数据人才的院校大多都处于起步阶段，院校普遍缺少对口专业的教师和教材，缺少完善的培养计划，缺少大数据实验平台，缺少开展大数据教学的海量数据。

成人教育可以被理解为一个以成人的方式指导教育的过程。成人教育是指有别于普通全日制教学形式的教育形式。从教育学的观点看，当我们说一个人进入成人期意味着其认知能力及学习能力均达到了成熟水平，如能够运用经验去认知周围的事物，能够自导学习的过程，等等。成人继续教育学历有4种主要形式，分别是高等教育自学考试（自考）、网络教育（远程教育）、成人高考（学习形式有脱产、业余、函授）、开放大学（原广播电视大学现代远程开放教育）。

由于成人教育的特征与普通高等教育不同，所以，成人教育的教学内容也不同于普通高等教育，以大数据技术为例，目前市场上有很多与大数据技术基础相关的图书都是以普通高等教育的学生和社会上需要培训的人员为对象的，很少以成人教育的学生为对象，这些图书对大数据的讲解主要存在以下两种问题。

① 对大数据技术的内容介绍追求大而全，例如，介绍Hadoop、数据采集、数据挖掘算法和工具、NoSQL数据库、Spark技术、数据可视化等几乎所有的大数据技术，对这些技术主要介绍基本概念和原理，内容不够深入和具体，学生只能简单地了解这些大数据技术，不能较深入地理解和掌握其中的某一项技术。

② 有些图书是在大数据培训的基础上总结整理出来的，这些书的内容比较偏向实践，内容范围控制得不错，不追求大而全的内容讲解，内容介绍比较深入，与实际案例的结合也比较多。但是，这类图书太注重实践，对相关理论的介绍偏少，不利于学生对整个大数据知识体系

的了解，影响学生对知识的扩展。

本书以成人教育的学生为对象，总结吸取以上列举图书的优点，并结合成人学历教育学生“边工作边学习”的特点，注重理论与实战的结合。本书的内容聚焦于大数据技术的基础知识和实践，由浅入深地对“Hadoop2.0”的概念、原理和技术进行全面而详细的介绍，主要内容包括大数据概述、Hadoop介绍、分布式文件系统HDFS、访问HDFS的常用接口、分布式计算框架MapReduce、MapReduce基础编程、分布式资源管理系统YARN、分布式锁服务ZooKeeper和Hadoop高可用集群的搭建，并结合目前企业的实际开发应用，引入真实的MapReduce高级编程案例。学生通过学习本书，将真正掌握Hadoop2.0技术的概念、原理、编程和部署，并为以后Hadoop生态圈相关技术的学习打下坚实的基础。

本书的另外一个特点是考虑成人学生的基础薄弱，起点参差不齐，特意在第2章中加入“Hadoop依赖的技术基础”内容，把与学习Hadoop密切相关的先修基础知识，如Java编程基础、Web可视化技术、关系数据库和Linux基础等知识，进行了详细的补充讲解，以避免因学生基础知识不足而导致学习困难等方面的问题，为后续章节的内容学习做铺垫。

本书的编写得益于北京邮电大学网络教育学院的大力支持，苏占玖、高大永两位老师为本书提供了部分内容，张文辉和王晓军老师对本书提出了很多宝贵的意见。另外，在本书编写过程中，北京思开教育科技有限公司的郎伟提供了部分实战编程内容。最后，本书还参考了相关技术的官方文档和大量的互联网资源，并尽量在参考文献部分一一列出，若有遗漏和不妥之处，敬请相关作者指正。在此，向有关单位、作者表示由衷的感谢。

由于编者水平有限，加之时间仓促，书中难免存在不足之处，敬请读者批评指正。请大家将遇到的错误和问题发邮件到jwli321@126.com。希望您能提出宝贵的意见，期待您的真挚反馈。

李建伟

2019年5月13日

目　录

第 1 章　大数据概述

现在的社会是一个科技与信息高速发展的社会，人们之间的交流越来越密切，生活也越来越方便，大数据技术已经不知不觉地渗入人们生活的方方面面，人们不仅生产大数据，同时也在使用大数据。

阿里巴巴创办人马云在一次演讲中提到，未来的时代将不是 IT 时代，而是 DT 时代，DT 就是 Data Technology，数据科技，表明了大数据对于阿里巴巴集团来说举足轻重。随着大数据价值逐渐被发现，传统的互联网公司将从 IT 科技公司转变为大数据技术公司，大数据技术将成为 IT 企业的核心技术。

有人把数据比喻为蕴藏能量的煤矿。大数据并不在“大”，而在于“有用”。价值含量、挖掘成本比数量更为重要。对于很多行业而言，如何利用这些大规模数据是赢得竞争的关键。

本章主要介绍大数据的概念、价值、应用场景和相关技术，并分析大数据与云计算和人工智能之间的区别与联系。

1.1　大数据概念及价值

随着移动互联网、移动终端和数据传感器的出现，数据正以超乎想象的速度快速增长。近几年，数据量已经从太字节级别跃升到拍字节乃至泽字节级别。

根据有“互联网女皇”之称的玛丽·米克尔发布的 2019 年互联网趋势报告，2018 年中国移动互联网数据流量同比增长 189%，增速在逐年加快。

图 1-1　中国移动互联网数据流量走势图

2011 年 5 月，麦肯锡全球研究院发布《大数据：下一代具有创新力、竞争力与生产力的前沿领域》，提出“大数据”时代的到来。各国政府也相继出台了一系列促进大数据产业发展的政

策。例如，2012 年 3 月，美国奥巴马政府发布了《大数据研究和发展倡议》，正式启动“大数据发展计划”，大数据上升为美国国家发展战略。2014 年 5 月，美国政府发布 2014 年全球“大数据”白皮书——《大数据：抓住机遇、守护价值》，报告鼓励使用数据来推动社会进步。2015 年 8 月，国务院印发《促进大数据发展行动纲要》系统部署我国大数据发展工作，加快建设数据强国。2015 年 9 月 18 日我国在贵州省启动我国首个大数据综合试验区的建设工作，力争通过 3～5 年的努力，将贵州大数据综合试验区建设成为全国数据汇聚应用新高地、综合治理示范区、产业发展聚集区、创业创新首选地、政策创新先行区。2016 年 3 月 17 日，《中华人民共和国国民经济和社会发展第十三个五年规划纲要》发布，其中第二十七章“实施国家大数据战略”提出：把大数据作为基础性战略资源，全面实施促进大数据发展行动，加快推动数据资源共享开放和开发应用，助力产业转型升级和社会治理创新。具体包括：加快政府数据开放共享、促进大数据产业健康发展。

通过百度关于“大数据”关键词的搜索指数，可以看出，大数据从 2012 年开始逐渐被大家关注，在 2017 年和 2018 年达到了最高关注度，目前，搜索指数已经逐渐趋向平稳，如图 1-2 所示。

图 1-2　百度“大数据”关键词搜索趋势图

大数据目前有多个定义，以下为其中的一些。

百度百科的定义是，大数据(Big Data)，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”研究机构 Gartner 给出了这样的定义：“大数据”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来适应海量、高增长率和多样化的信息资产。麦肯锡全球研究所给出的定义是：一种规模大到在获取、存储、管理、分析方面大大超出了传统数据库软件工具能力范围的数据集合，具有海量的数据规模、快速的数据流转、多样的数据类型和价值密度低四大特征。

在维克托·迈尔-舍恩伯格及肯尼斯·库克耶编写的《大数据时代》一书中大数据指不用

随机分析法(抽样调查)这样的捷径,而采用所有数据进行分析处理。

业界通常用 4V 来概括大数据的特征:Volume(大量)、Variety(多样)、Value(低价值密度)、Velocity(速度快、时效高)。

(1) 数据量大(Volume)

第 1 个特征是数据量大。大数据的起始计量单位至少是拍字节(约 1 000 个太字节)、艾字节(约 100 万个太字节)或泽字节(约 10 亿个太字节)。截至目前,人类生产的所有印刷材料的数据量是 200 PB(1 PB=1 024 TB),而历史上全人类说过的所有的话的数据量大约是 5 EB(1 EB=1 024 PB)。当前,典型个人计算机硬盘的容量为太字节量级,而一些大企业的数据量已经接近艾字节量级。截止到 2018 年,互联网用户数已达到 39 亿人,据 Statista 2018 年最新统计数据显示,2018 年全球连接设备的数量将超过 230 亿人。微信发布的 2018 数据报告中显示:月活跃用户人数超 10.8 亿人,每天发出 450 亿次信息。

(2) 类型繁多(Variety)

第 2 个特征是数据类型繁多。数据的格式是多样化的,如文字、图片、视频、音频、地理位置信息等,数据也可以有不同的来源,如传感器、互联网等。这种类型的多样性也让数据被分为结构化数据和非结构化数据。相对于以往便于存储的以文本为主的结构化数据,非结构化数据越来越多,包括网络日志、音频、视频、图片、地理位置信息等。这些多类型的数据对数据的处理能力提出了更高要求。

(3) 价值密度低(Value)

第 3 个特征是数据价值密度相对较低。随着物联网的广泛应用,信息感知无处不在,信息海量,但价值密度较低,例如监控视频,一部 1 小时的视频,在连续不间断的监控中,有用数据可能仅有一二秒。如何通过强大的机器算法更迅速地完成数据的价值"提纯"成为目前大数据背景下亟待解决的难题。

(4) 速度快、时效高(Velocity)

第 4 个特征是处理速度快,时效性要求高。这是大数据区分于传统数据挖掘最显著的特征。根据互联网数据中心(IDC)的"数字宇宙"的报告,预计到 2020 年,全球数据使用量将达到 35 ZB。在如此海量的数据面前,处理数据的效率就是企业的生命。

另外,数据具有一定的时效性,是不停变化的,数据量可以随时间逐渐增大,也可在空间上不断移动变化的数据。如果采集到的数据不经过流转,最终会过期作废。客户的体验在分秒级别,海量的数据,带来的第 1 个问题就是大大延长了各类报表生成的时间,我们能否在极端的时间内提取最有价值的信息呢?数据在 1 秒内得不到流转处理,就会给客户带来较差的使用体验,若数据处理软件达不到"秒"处理,所带来的商业价值就会大打折扣。

大数据技术的价值不在于掌握庞大的数据信息,而在于对这些含有意义的数据进行专业化处理。换而言之,如果把大数据比作一种产业,那么这种产业实现盈利的关键,就在于提高对数据的"加工能力",通过"加工"实现数据的"增值"。既有的 IT 技术架构和路线,已经无法高效处理如此海量的数据,而对于相关组织来说,如果投入巨额财力而采集的信息无法通过及时处理得到有效信息,那将是得不偿失的。可以说,大数据时代对人类的数据驾驭能力提出了新的挑战,也为人们获得更为深刻、全面的洞察能力提供了前所未有的空间与潜力。

1.2 大数据数据源

关于大数据的来源，普遍认为互联网及物联网是产生并承载大数据的基地。主要通过各种数据传感器、数据库、网站、移动 App 等产生大量的结构化和非结构化数据。互联网公司是天生的大数据公司，在搜索、社交、媒体、交易等各自核心业务领域，积累并持续产生海量数据。例如，百度公司数据总量超过了千拍字节级别，数据涵盖了中文网页、百度日志、用户生产的内容(UGC)、百度推广等多个部分，并占有国内 70%以上的搜索市场份额。阿里巴巴公司目前保存的数量超过百拍字节级别，其中 90%以上是电商数据、交易数据、用户浏览数据。腾讯公司保存的数据总量超过百拍字节级别，主要是社交和游戏数据。物联网设备每时每刻都在采集数据，设备数量和数据量都与日俱增，Statista 2018 年统计显示，2015～2025 年全球连接设备的数量将从 15 亿个增加到 750 亿个。这两类数据资源作为大数据金矿，正在不断产生各类应用数据。

此外，还有一些行业大数据，如电信、金融与保险、电力与石化、制造业、医疗、教育和交通运输等行业的大数据。这些行业的企事业单位在业务中也积累了许多数据。例如，电信行业包括用户上网记录、通话、信息、地理位置信息等，运营商拥有的数据量都在 10 PB 以上；国家电网采集获得的数据总量就达到 10 PB 级别；列车、水陆路运输产生的各种视频、文本类数据，每年大约几十拍字节级；金融系统每年产生数据达到几十拍字节级；整个医疗卫生行业一年能够保存下来的数据有数百拍字节级。从严格意义上讲，这些数据资源比较分散，还算不上大数据，但对商业应用而言，却是最易获得和比较容易加工处理的数据资源，也是当前在国内比较常见的应用资源。

还有一类是政府部门掌握的数据资源，如公共安全、政务、气象与地理、人口与文化等数据。例如，北京市有 50 多万个监控摄像头，每天采集的视频数据量约 3 PB，整个视频监控每年保存下来的数据有近千拍字节级；中国幅员辽阔，气象局保存的气象数据为 5 PB，各种地图和地理位置信息每年约为几十拍字节。这些数据普遍认为质量好、价值高，但开放程度低。国务院印发的《促进大数据发展行动纲要》中部署三方面主要任务，其中首要任务就是要加快政府数据开放共享，推动资源整合，提升治理能力。大力推动政府部门数据共享，稳步推动公共数据资源开放，统筹规划大数据基础设施建设，支持宏观调控科学化，推动政府治理精准化，推进商事服务便捷化，促进安全保障高效化，加快民生服务普惠化。

数据从哪里来是我们评价大数据应用的第 1 个关注点。一是要看这个应用是否真有数据支撑，数据资源是否可持续，来源渠道是否可控，数据安全和隐私保护方面是否有隐患。二是要看这个应用的数据资源质量如何，是“富矿”还是“贫矿”，能否保障这个应用的实效。对于来自自身业务的数据资源，具有较好的可控性，数据质量一般也有保证，但数据覆盖范围可能有限，需要借助其他资源渠道。对于从互联网抓取的数据，技术能力是关键，既要有能力获得足够大的量，又要有能力筛选出有用的内容。对于从第三方获取的数据，需要特别关注数据交易的稳定性。数据从哪里来是分析大数据应用的起点，如果一个应用没有可靠的数据来源，再好、再高超的数据分析技术都是无本之木。

1.3　大数据技术应用场景

大数据技术的应用已经渗透各行各业，如医疗、金融、餐饮、电商、农业、交通、教育、体育、环保、食品和政务等领域。下面将重点介绍几个行业中大数据应用的场景。

1. 大数据在医疗领域的应用

医疗行业很早就遇到了海量数据和非结构化数据的挑战，而近年来很多国家都在积极推进医疗信息化发展，这使得很多医疗机构有资金来做大数据分析。

医疗行业拥有大量的病例、病理报告、治愈方案、药物报告等，如果这些数据可以被整理和应用将会极大地帮助医生和病人，疾病的治疗将变得更加精准和高效。

如果未来基因技术发展成熟，可以根据病人的基因序列特点进行分类，建立医疗行业的病人分类数据库。在医生诊断病人时可以参考病人的疾病特征、化验报告和检测报告，参考疾病数据库来快速帮助病人确诊，明确定位疾病。在制定治疗方案时，医生可以依据病人的基因特点，调取相似基因、年龄、人种、身体情况的有效治疗方案，制定出适合病人的治疗方案，帮助更多人及时进行治疗。同时这些数据也有利于医药行业开发出更加有效的药物和医疗器械。

除此之外，利用大数据技术还可以实现流行病预测，如 Google 流感趋势(Google Flu Trends)便是利用搜索关键词预测禽流感的散布情况。

2. 大数据在零售和电商行业的应用

首先，零售行业可以利用大数据技术进行精准营销。例如，商家可以根据客户消费喜好和趋势，进行商品的精准营销，降低营销成本，当客户购买商品以后，再依据客户购买的产品，为客户提供可能购买的其他产品，扩大销售额。其次，零售行业可以通过大数据掌握未来消费趋势，有利于热销商品的进货管理和过季商品的处理。再次，零售行业的数据对于产品生产厂家是非常宝贵的，零售商的数据信息将会有助于资源的有效利用，降低产能过剩，厂商依据零售商的信息按实际需求进行生产，减少不必要的生产浪费。最后，零售行业还可以根据市场需求和库存情况实时定价，例如，梅西百货根据需求和库存的情况，基于 SAS 的系统对多达 7 300 万种货品进行实时调价。

电商是最早利用大数据进行精准营销的行业，除了精准营销，电商可以依据客户消费习惯来提前为客户备货，并利用便利店作为货物中转点，在客户下单 15 分钟内将货物送上门，提高客户体验。马云的菜鸟网络宣称的 24 小时完成在中国境内的送货，以及刘强东宣传未来京东将在 15 分钟完成送货上门都是基于客户消费习惯的大数据分析和预测。

电商可以利用其交易数据和现金流数据，为其生态圈内的商户提供基于现金流的小额贷款，电商也可以将此数据提供给银行，同银行合作为中小企业提供信贷支持。未来，电商还可以利用大数据预测流行趋势、消费趋势、地域消费特点、客户消费习惯、各种消费行为的相关度、消费热点、影响消费的重要因素等。

3. 大数据在金融行业的应用

大数据在金融行业应用范围较广。例如，花旗银行利用 IBM 沃森计算机为财富管理客户推荐产品；美国银行利用客户点击数据集为客户提供特色服务，如有竞争的信用额度；招商银行通过客户刷卡、存取款、电子银行转账、微信评论等行为对数据进行分析，每周给客户发送针对性广告信息，里面有客户可能感兴趣的产品和优惠信息。大数据在金融行业的应用可以总

结为以下5个方面。

① 精准营销：依据客户消费习惯、地理位置、消费时间进行推荐。

② 风险管控：依据客户消费和现金流提供信用评级或融资支持，利用客户社交行为记录实施信用卡反欺诈。

③ 决策支持：利用决策树技术进行抵押贷款管理，利用数据分析报告实施产业信贷风险控制。

④ 效率提升：利用金融行业全局数据了解业务运营薄弱点，利用大数据技术加快内部数据处理速度。

⑤ 产品设计：利用大数据计算技术为财富客户推荐产品，利用客户行为数据设计满足客户需求的金融产品。

4. 大数据在交通出行领域的应用

交通作为人类行为的重要组成和重要条件之一，对于大数据的感知也是最急迫的。近年来，我国的智能交通已实现了快速发展，许多技术手段都达到了国际领先水平。但是，问题和困境也非常突出，从各个城市的发展状况来看，智能交通的潜在价值还没有得到有效挖掘，对交通信息的感知和收集有限，对存在于各个管理系统中的海量的数据无法共享运用、有效分析，对交通态势的研判预测乏力，很难满足公众对交通信息服务的需求。

目前，交通领域的大数据应用主要体现在两个方面，一方面可以利用大数据传感器收集的数据来了解车辆通行密度，合理进行道路规划包括单行线路规划。另一方面可以利用大数据来实现即时信号灯调度，提高已有线路运行能力。科学地安排信号灯是一个复杂的系统工程，必须利用大数据计算平台才能计算出一个较为合理的方案。科学的信号灯安排将会将已有道路的通行能力提高30%～40%。例如，2018年11月在乌镇召开的第五届世界互联网大会人工智能论坛上，百度董事长李彦宏发表演讲，称从北京海淀区开始，百度将接管海淀区的所有红绿灯，并称以后将使人们等待红绿灯的时间减少30%～40%。在美国，政府依据某一路段的交通事故信息来增设信号灯，降低了50%以上的交通事故率。机场的航班起降依靠大数据将会提高航班管理的效率，航空公司利用大数据可以提高上座率，降低运行成本。铁路利用大数据可以有效安排客运和货运列车，提高效率、降低成本。

5. 大数据在教育领域的应用

美国新媒体联盟(NMC)与北京师范大学智慧学习研究院合作的《2016新媒体联盟中国基础教育技术展望：地平线项目区域报告》指出，大数据学习分析技术将在未来两至三年成为极具影响力的教育技术，并表明有效运用学习分析技术可以设计更好的教学活动，让学生积极主动地参与学习，准确定位处于危险中的学生群体，评估预测影响学生成绩的因素。

利用大数据分析方法可以对学生的在线学习数据进行全面的收集、测量和分析，理解与优化教学过程及其情境，为教学决策、学业预警提供支持，真正实现个性化学习，提高教学效果，这是大数据学习分析在教育领域的价值所在。例如，美国的Knewton就是一家利用大数据技术提供个性化教育的公司，它利用适配学习技术，通过数据收集、推断和建议三部曲来提供个性化的教学。国内的松鼠AI利用大数据和人工智能技术为接受基础教育(小学、初中)和高中教育的学生提供自适应个性化教学，让每个学生都清楚自己的潜力，了解自己的强项和弱项。

大数据还可以帮助家长和教师甄别孩子的学习差距和有效的学习方法。比如，美国的麦格劳-希尔教育出版集团就开发出了一种预测评估工具，帮助学生评估他们已有的知识和达标

测验所需程度的差距，进而指出学生有待提高的地方。评估工具可以让教师跟踪学生学习情况，从而找到学生的学习特点和方法。有些学生适合按部就班，有些则更适合图式信息和整合信息的非线性学习。这些都可以通过大数据搜集和分析很快识别出来，从而为教育教学提供坚实的依据。

未来，大数据在教育领域的应用主要集中在自适应个性化学习、英语语音测评、教育机器人、智能陪练、分级阅读等几个方向。

6. 大数据在制造业的应用

利用大数据推动信息化和工业化深度融合，研究推动大数据在研发设计、生产制造、经营管理、市场营销、售后服务等产业链各环节的应用，研发面向不同行业、不同环节的大数据分析应用平台，选择典型企业、重点行业、重点地区开展工业企业大数据应用项目试点，积极推动制造业网络化和智能化。最近几年，从国家到地方政府，日益重视大数据在制造业特别是高端智能制造领域的应用，如《中国制造 2025》的发布。从这个意义上来说，大数据在制造业将发挥巨大潜力，释放更大空间。未来，利用工业大数据将提升制造业水平，主要集中在产品故障诊断与预测、分析工艺流程、改进生产工艺、优化生产过程能耗、工业供应链分析与优化、生产计划与排程等方面。

1.4　大数据处理流程及技术

大数据处理流程如图 1-3 所示，主要包括数据收集、数据预处理、数据存储、数据处理与分析、数据展示/数据可视化等环节，每一个数据处理环节都会对大数据质量产生影响。通常，一个好的大数据产品要有大量的数据规模、快速的数据处理能力、精确的数据分析与预测能力、优秀的可视化图表以及简练易懂的结果解释，下面将分别介绍大数据处理流程及相关的主要技术。

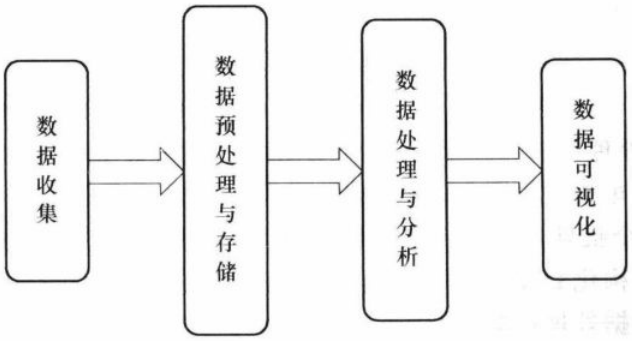

图 1-3　大数据处理流程图

1. 数据收集

大数据的采集指利用多个数据库来接收发自客户端（Web、App 或者传感器形式等）的数据，并且用户可以通过这些数据库来进行简单的查询和处理工作，另外，大数据的采集不是抽样调查，它强调数据尽可能完整和全面，尽量保证每一个数据精确有用。

在大数据的采集过程中，其主要特点和挑战是并发数高，因为同时有可能会有成千上万的用户来进行访问和操作，比如火车票售票网站和淘宝，它们并发的访问量在峰值时达到上百万