

新工科信息技术基础系列规划教材

大数据导论

——数据思维、数据能力和数据伦理

（通识课版）

林子雨 编著

高等教育出版社

新工科信息技术基础系列规划教材

大数据导论

——数据思维、数据能力和数据伦理

(通识课版)

林子雨 编著

高等教育出版社·北京

内容提要

本书详细阐述了培养具有数据素养的综合型人才所需要的大数据相关知识。本书紧紧围绕通识教育的核心理念,努力培养读者的数据意识、数据思维、数据能力和数据能力。本书介绍了大数据技术、大数据应用、大数据安全、大数据思维、大数据伦理、数据共享、数据开放、大数据交易以及大数据治理等内容。为了避免陷入空洞的理论介绍,本书融入了丰富的案例,这些案例就发生在我们生活的大数据时代,很具有代表性和说服力,能够让读者直观感受相应理论的具体内涵。

本书可作为高等学校非计算机专业(尤其是文科)数据科学通识类课程教材,也可供对大数据感兴趣的读者自学使用。

图书在版编目(CIP)数据

大数据导论:数据思维、数据能力和数据伦理:通识课版/林子雨编著.--北京:高等教育出版社,2020.2

ISBN 978-7-04-053577-8

I. ①大… II. ①林… 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(2020)第014114号

Dashuju Daolun——Shuju Siwei、Shuju Nengli he Shuju Lunli

策划编辑 唐德凯

责任编辑 唐德凯

封面设计 张申申

版式设计 童丹

插图绘制 于博

责任校对 刘娟娟

责任印制 刁毅

出版发行 高等教育出版社

社址 北京市西城区德外大街4号

邮政编码 100120

印刷 中农印务有限公司

开本 787mm×1092mm 1/16

印张 16

字数 390千字

购书热线 010-58581118

咨询电话 400-810-0598

网 址 <http://www.hep.edu.cn>

<http://www.hep.com.cn>

网上订购 <http://www.hepmall.com.cn>

<http://www.hepmall.com>

<http://www.hepmall.cn>

版 次 2020年2月第1版

印 次 2020年2月第1次印刷

定 价 32.00元

本书如有缺页、倒页、脱页等质量问题,请到所购图书销售部门联系调换

版权所有 侵权必究

物料号 53577-00

。前言

作为继云计算、物联网之后 IT 行业又一颠覆性的技术,大数据备受人们关注。大数据无处不在,包括金融、汽车、零售、餐饮、电信、能源、政务、医疗、体育、娱乐等在内的社会各行各业,都融入了大数据,大数据对人类的社会生产和生活必将产生重大而深远的影响。

对于一个国家而言,能否紧紧抓住大数据发展机遇,快速形成以核心技术为主导的应用参与新一轮的全球化竞争,将直接决定未来世界范围内各国科技力量博弈的格局。大数据人才的培养是新一轮科技较量的基础,高等院校承担着大数据人才培养的重任。大数据人才的培养包括两个部分,一个是培养掌握复杂大数据技术知识的专业型人才,另一个是培养具有数据素养的综合型人才。数据素养是信息素养在大数据时代的拓展和延伸,也是在大数据时代的新环境下,对当代大学生提出的新要求。对大学生进行数据素养教育,可以有效提升大学生的数据意识、数据思维和数据能力,使其在大数据时代获得更好的生存和发展空间。

随着全国高校数据科学与大数据技术专业建设的持续推进,在全国从事大数据专业教学的老师的共同努力下,目前已经有了一大批优秀的大数据专业教材,较好地满足了新专业对教材的紧迫需求。但是,在通识导论教材方面,目前还处于稀缺状态。笔者所在的厦门大学数据库实验室,是全国高校知名的大数据教学团队,每年有很多来自全国各地的教师来到厦门大学与我们开展交流,与此同时,我们也积极走出去,到全国各地学习其他高校的大数据课程和教材的建设经验。通过大量的交流,我们了解到,国内高校关于开设大数据通识导论课程的需求,都是苦于缺乏相应的教材。为了让更多高校能够把大数据通识导论课程开设起来,笔者在大量调研学习的基础上编著了本书。

作为通识类课程教材,本书服务于具有数据素养的综合型人才的培养,并非面向大数据专业型人才的培养,因此,本书的读者对象是非计算机专业大学生。本书紧紧围绕通识教育核心理念,努力培养学生的数据意识、数据思维、数据伦理和数据能力。

本书共 11 章,详细阐述了培养具有数据素养的综合型人才所需要的大数据相关知识储备。第 1 章介绍了数据的概念、大数据时代到来的背景、大数据的发展历程、世界各国的大数据发展战略、大数据的概念与影响、大数据的应用以及大数据产业。第 2 章介绍了云计算、物联网、人工智能的概念和应用,并阐述了大数据与云计算、物联网、人工智能之间的紧密关系。第 3 章介绍大数据分析全流程所涉及的各种技术,包括数据采集与预处理、数据存储和管理、数据处理与分析、数据可视化、数据安全和隐私保护等。第 4 章介绍大数据在各大领域的典型应用,包括互联网、生物医学、物流、城市管理、金融、汽车、零售、餐饮、电信、能源、体育、娱乐、安全、政府和日常生活等领域。第 5 章讨论大数据安全问题并给出相关的典型案例。第 6 章介绍大数据时代新的思维方式,包括全样而非抽样、效率而非精确、相关而非因果、以数据为中心、“人人为我,我为人人”等。第 7 章介绍大数据伦理的概念,并给出与大数据伦理相关的典型案例。第 8 章介绍数据共享的意义、实现数据共享所面临的

相关挑战以及推进数据共享应当采取的措施。第9章介绍政府开放数据的理论基础、政府数据开放的重要意义、国外政府开放数据的经验和我国政府开放数据的现状。第10章介绍大数据交易的发展现状、大数据交易平台、大数据交易在发展过程中出现的问题以及推进大数据交易发展的对策。第11章介绍大数据治理的概念、要素和治理模型。

本书面向高校非计算机专业的学生，可以作为通识类课程教材。本书由林子雨执笔。在撰写过程中，厦门大学计算机科学系硕士研究生魏亮、曾冠华、程璐、林哲、郑宛玉、陈杰祥等做了大量辅助性工作。在此，向这些同学的辛勤工作表示衷心的感谢。同时，感谢夏小云老师在书稿校对过程中的辛勤付出。

本书官方网站（扫描下方二维码即可）免费提供了全部配套资源的在线浏览和下载，并接受错误反馈和发布勘误信息。

本书在撰写过程中，参考了大量网络资料、文献和书籍，对相关知识进行了系统梳理，有选择性地把一些重要知识纳入本书。由于笔者能力有限，本书难免存在不足之处，望广大读者不吝赐教。

林子雨
2019年3月



本书官方网站

○ 目 录

第 1 章 大数据概述

1.1 数据	002	1.4.3 法国	014
1.1.1 数据的概念	002	1.4.4 韩国	014
1.1.2 数据类型	002	1.4.5 日本	015
1.1.3 数据组织形式	003	1.4.6 中国	015
1.1.4 数据生命周期	003	1.5 大数据的概念	016
1.1.5 数据的使用	003	1.5.1 数据量大	016
1.1.6 数据的价值	005	1.5.2 数据类型繁多	017
1.1.7 数据爆炸	005	1.5.3 处理速度快	018
1.2 大数据时代	006	1.5.4 价值密度低	018
1.2.1 第三次信息化浪潮	006	1.6 大数据的影响	019
1.2.2 信息科技为大数据时代提供 技术支撑	006	1.6.1 大数据对科学研究的影响	019
1.2.3 数据产生方式的变革促成大数据时代 的来临	009	1.6.2 大数据对社会发展的影响	020
1.3 大数据的发展历程	011	1.6.3 大数据对就业市场的影响	021
1.4 世界各国的大数据发展战略	013	1.6.4 大数据对人才培养的影响	022
1.4.1 美国	013	1.7 大数据的应用	023
1.4.2 英国	014	1.8 大数据产业	025
		1.9 本章小结	026
		习题	026

第 2 章 大数据与云计算、物联网、人工智能

2.1 云计算	030	2.3 大数据与云计算、物联网的 关系	040
2.1.1 云计算的概念	030	2.4 人工智能	042
2.1.2 云计算的服务模式和类型	033	2.4.1 人工智能的概念	042
2.1.3 云计算数据中心	033	2.4.2 人工智能的关键技术	042
2.1.4 云计算的应用	034	2.4.3 人工智能的应用	047
2.1.5 云计算产业	035	2.4.4 人工智能产业	051
2.2 物联网	036	2.5 大数据与人工智能的关系	054
2.2.1 物联网的概念	036	2.6 本章小结	055
2.2.2 物联网的关键技术	037	习题	055
2.2.3 物联网的应用	039		
2.2.4 物联网产业	040		

第3章 大数据技术

3.1 概述	058	3.4.1 数据挖掘和机器学习算法	067
3.2 数据采集与预处理	058	3.4.2 大数据处理与分析技术	068
3.2.1 数据采集的概念	059	3.5 数据可视化	070
3.2.2 数据采集的3大要点	059	3.5.1 数据可视化的概念	070
3.2.3 数据采集的数据源	060	3.5.2 数据可视化的重要作用	071
3.2.4 数据清洗	061	3.5.3 数据可视化案例	073
3.3 数据存储和管理	063	3.6 数据安全和隐私保护	075
3.3.1 传统的数据存储和管理技术	063	3.6.1 数据安全技术	075
3.3.2 大数据时代的数据存储和管理 技术	065	3.6.2 隐私保护技术	076
3.4 数据处理与分析	067	3.7 本章小结	076
		习题	076

第4章 大数据应用

4.1 大数据在互联网领域的应用	080	4.5.2 市场情绪分析	102
4.1.1 推荐系统的概念	080	4.5.3 信贷风险分析	103
4.1.2 长尾理论	081	4.5.4 大数据征信	104
4.1.3 推荐方法	081	4.6 大数据在汽车领域的应用	106
4.1.4 推荐系统模型	082	4.7 大数据在零售领域的应用	107
4.1.5 推荐系统的应用	083	4.7.1 发现关联购买行为	107
4.2 大数据在生物学领域的应用	084	4.7.2 客户群体细分	108
4.2.1 流行病预测	084	4.7.3 供应链管理	109
4.2.2 智慧医疗	087	4.8 大数据在餐饮领域的应用	109
4.2.3 生物信息学	088	4.8.1 餐饮行业拥抱大数据	109
4.2.4 案例：基于大数据的综合健康 服务平台	089	4.8.2 餐饮O2O	109
4.3 大数据在物流领域的应用	093	4.9 大数据在电信领域的应用	111
4.3.1 智能物流的概念	093	4.10 大数据在能源领域的应用	112
4.3.2 智能物流的作用	094	4.11 大数据在体育和娱乐领域的 应用	113
4.3.3 智能物流的应用	094	4.11.1 训练球队	113
4.3.4 大数据是智能物流的关键	095	4.11.2 投拍影视作品	114
4.3.5 中国智能物流骨干网——菜鸟	095	4.11.3 预测比赛结果	115
4.4 大数据在城市管理领域的应用	098	4.12 大数据在安全领域的应用	116
4.4.1 智能交通	098	4.12.1 大数据与国家安全	116
4.4.2 环保监测	099	4.12.2 应用大数据技术防御网络攻击	117
4.4.3 城市规划	100	4.12.3 警察应用大数据工具预防犯罪	117
4.4.4 安防领域	101	4.13 大数据在政府领域的应用	118
4.5 大数据在金融领域的应用	102	4.14 大数据在日常生活中的应用	119
4.5.1 高频交易	102	4.15 本章小结	120

习题	121
----------	-----

第5章 大数据安全

5.1 传统数据安全	124	5.5 大数据保护的基本原则	132
5.2 大数据安全与传统数据安全的 不同	124	5.5.1 数据主权原则	133
5.3 大数据安全问题	125	5.5.2 数据保护原则	133
5.3.1 隐私和个人信息安全问题	125	5.5.3 数据自由流通原则	133
5.3.2 国家安全问题	127	5.5.4 数据安全原则	133
5.4 典型案例	128	5.6 大数据时代数据安全性与隐私 保护的对策	134
5.4.1 棱镜门事件	128	5.7 世界各国保护数据安全的实践	134
5.4.2 维基解密	129	5.7.1 欧盟	135
5.4.3 Facebook 数据滥用事件	129	5.7.2 美国	136
5.4.4 手机应用软件过度采集个人信息	130	5.7.3 英国	137
5.4.5 12306 数据泄露	131	5.7.4 中国	137
5.4.6 免费 WiFi 窃取用户信息	132	5.7.5 其他国家	138
5.4.7 收集个人隐私信息的“探针 盒子”	132	5.8 本章小结	139
		习题	139

第6章 大数据思维

6.1 传统的思维方式	142	营销	148
6.2 大数据时代需要新的思维方式	142	6.4.4 吸烟有害身体健康的法律诉讼	148
6.3 大数据思维方式	143	6.4.5 基于大数据的药品研发	150
6.3.1 全样而非抽样	144	6.4.6 基于大数据的谷歌广告	151
6.3.2 效率而非精确	144	6.4.7 搜索引擎“点击模型”	151
6.3.3 相关而非因果	145	6.4.8 迪士尼 MagicBand 手环	152
6.3.4 以数据为中心	145	6.4.9 谷歌流感趋势预测	153
6.3.5 我为人人，人人为我	146	6.4.10 大数据的简单算法比小数据的 复杂算法更有效	153
6.4 运用大数据思维的具体实例	147	6.4.11 谷歌翻译	153
6.4.1 商品比价网站	147	6.5 本章小结	154
6.4.2 啤酒与尿布	148	习题	154
6.4.3 零售商 Target 的基于大数据的商品			

第7章 大数据伦理

7.1 大数据伦理概念	158	7.2.4 隐性偏差问题	160
7.2 大数据伦理的典型案例	158	7.2.5 魏则西事件	160
7.2.1 徐玉玉事件	159	7.2.6 “信息茧房”问题	161
7.2.2 大麦网撞库事件	159	7.3 大数据的伦理问题	162
7.2.3 大数据“杀熟”	159	7.3.1 隐私泄露问题	162

7.3.2 数据安全问题	164	7.4.7 技术乌托邦的消极影响	169
7.3.3 数字鸿沟问题	164	7.4.8 大数据技术本身的缺陷	169
7.3.4 数据独裁问题	165	7.5 大数据伦理问题的治理	169
7.3.5 数据垄断问题	165	7.5.1 提高保护个人隐私数据的意识	170
7.3.6 数据的真实可靠问题	166	7.5.2 加强大数据伦理规约的构建	170
7.3.7 人的主体地位问题	166	7.5.3 努力实现以技术治理大数据	170
7.4 大数据伦理问题产生的原因	167	7.5.4 完善大数据立法	171
7.4.1 人类社会价值观的转变	167	7.5.5 完善大数据伦理管理机制	171
7.4.2 数据伦理责任主体不明确	167	7.5.6 引导企业坚持责任与利益并重	171
7.4.3 相关主体的利益牵涉	167	7.5.7 努力弘扬共享精神化解数字鸿沟	172
7.4.4 道德规范的缺失	168	7.5.8 倡导跨行业跨部门合作	172
7.4.5 法律体系不健全	168	7.6 本章小结	173
7.4.6 管理机制不完善	168	习题	173

第8章 数据共享

8.1 数据孤岛问题	176	8.5 推进数据共享开放的举措	180
8.1.1 政府的数据孤岛问题	176	8.5.1 政府层面的举措	180
8.1.2 企业的数据孤岛问题	176	8.5.2 企业层面的举措	181
8.2 数据孤岛问题产生的原因	177	8.6 数据共享案例	181
8.2.1 政府数据孤岛的产生原因	177	8.6.1 案例1: 菜鸟物流	181
8.2.2 企业数据孤岛的产生原因	177	8.6.2 案例2: 政府一站式平台 ——i 厦门	182
8.3 消除数据孤岛的重要意义	178	8.6.3 案例3: 浙江打通政府数据, 让群众 最多跑一次	184
8.3.1 对于政府的意义	178	8.7 本章小结	186
8.3.2 对于企业的意义	178	习题	186
8.4 实现数据共享所面临的挑战	179		
8.4.1 政府层面的挑战	179		
8.4.2 企业层面的挑战	179		

第9章 数据开放

9.1 政府开放数据的理论基础	188	9.3.3 政府开放数据有利于社会治理 创新	192
9.1.1 数据资产理论	188	9.4 国外政府开放数据的经验	193
9.1.2 数据权理论	188	9.4.1 概述	193
9.1.3 开放政府理论	189	9.4.2 G8 数据开放原则	194
9.2 政府信息公开与政府数据开放的 联系与区别	190	9.4.3 美国开放数据国家行动计划	195
9.3 政府数据开放的重要意义	191	9.4.4 英国开放数据国家行动计划	195
9.3.1 政府开放数据有利于促进开放透明 政府的形成	191	9.4.5 德国政府开放数据行动	197
9.3.2 政府开放数据有利于创新创业和 经济增长	192	9.4.6 日本政府开放数据行动	197
		9.5 国内政府开放数据	198
		9.5.1 概述	199

9.5.2 我国政府数据开放制度体系	199	9.6 政府数据开放的几点启示	208
9.5.3 当前数据开放存在的主要问题	200	9.7 本章小结	209
9.5.4 各地政府数据开放实践	204	习题	210

第10章 大数据交易

10.1 概述	212	10.4.5 大数据需求不明确,抑制交易市场 发展	220
10.2 大数据交易发展现状	213	10.4.6 大数据交易组织机构定位不清	220
10.3 大数据交易平台	214	10.4.7 用户隐私保护隐患重重	220
10.3.1 交易平台的类型	214	10.4.8 大数据交易专业人才缺乏	221
10.3.2 交易平台的数据来源	215	10.5 推进大数据交易发展的对策	221
10.3.3 交易平台的产品类型	215	10.5.1 加快制定隐私保护相关法律 法规	221
10.3.4 交易平台涉及的主要领域	216	10.5.2 加快推进政府数据开放共享	221
10.3.5 平台的交易规则	216	10.5.3 加快完善市场交易机制	222
10.3.6 交易平台的运营模式	216	10.5.4 加快建立大数据交易监管职能 部门	222
10.3.7 代表性的大数据交易平台	217	10.5.5 加快培育大数据交易人才	222
10.4 大数据交易在发展过程中出现的 问题	218	10.6 本章小结	223
10.4.1 互联网数据马太效应显现	218	习题	223
10.4.2 大数据产权界定不清晰	219		
10.4.3 大数据交易规则和标准缺乏	219		
10.4.4 数据估值定价机制有待完善	219		

第11章 大数据治理

11.1 概述	226	11.3.2 HESA 数据治理模型	234
11.1.1 数据治理的必要性	226	11.3.3 数据治理螺旋模型	234
11.1.2 数据治理的基本概念	226	11.4 大数据治理保障机制	235
11.1.3 数据治理与数据管理的关系	227	11.4.1 大数据治理战略目标	235
11.1.4 大数据治理的基本概念	228	11.4.2 大数据治理组织	236
11.1.5 大数据治理与数据治理的关系	230	11.4.3 制度章程	237
11.1.6 大数据治理的重要意义和作用	231	11.4.4 流程管理	238
11.2 大数据治理的要素	232	11.4.5 技术应用	238
11.3 大数据治理模型	233	11.5 本章小结	239
11.3.1 ISACA 数据治理模型	233	习题	239

参考文献	241
------------	-----

第1章

大数据概述

第1章电子教案

大数据时代的悄然来临，带来了信息技术发展的巨大变革，并深刻影响着社会生产和人们生活的方方面面。世界各国政府高度重视大数据技术的研究和产业发展，纷纷把大数据上升为国家战略并重点推进。企业和学术机构纷纷加大技术、资金和人员投入力度，加强对大数据关键技术的研发与应用，以期在“第三次信息化浪潮”中占得先机、引领市场。大数据已经不是“镜中花、水中月”，它的影响力和作用力正迅速触及社会的每个角落，所到之处，或是颠覆，或是提升，都让人们深切感受到了大数据实实在在的威力。

本章介绍了数据的概念、大数据时代到来的背景、大数据的发展历程、世界各国的大数据发展战略、大数据的概念与影响、大数据的应用以及大数据产业。

1.1 数据

本节介绍数据的概念、数据类型、数据组织形式、数据生命周期、数据的使用、数据的价值性和数据爆炸。

1.1.1 数据的概念

数据是对客观事物的性质、状态以及相互关系等进行记载的物理符号，是可识别的、抽象的。数据和信息是两个不同的概念，信息是较为宏观的概念，它由数据的有序排列组合而成，传达给读者某个概念方法等。而数据则是构成信息的基本单位，离散的数据没有任何实用价值。

数据有很多种类型，比如数字、文字、图像、声音等。随着人类社会信息化进程的加快，在我们日常生产和生活中每天都在不断产生大量的数据。数据已经渗透到当今每一个领域，成为重要的生产要素。对企业而言，从创新到所有决策，数据推动着企业的发展，并使得各级组织的运营更为高效，可以认为，数据将成为每个企业获取核心竞争力的关键因素。数据资源已经和物质资源、人力资源一样，成为国家的重要战略资源，影响着国家和社会的安全、稳定与发展，因此，数据也被称为“未来的石油”（见图 1-1）。

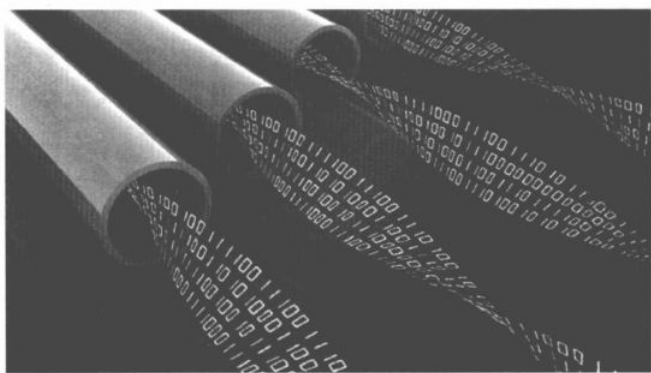


图 1-1 数据是“未来的石油”

1.1.2 数据类型

常见的数据类型包括文本、图片、音频、视频等。

- 文本：文本是一种由若干字符构成的计算机文件，常见格式包括 ASCII、MIME 和 TXT。
- 图片：图片是指由图形、图像等构成的平面媒体。图片的格式很多，大体可以分为点阵图和矢量图两大类，常见的 BMP、JPG 等格式都是点阵图形，而 Flash 动画制作软件所生成的 SWF 格式和 Photoshop 绘图软件所生成的 PSD 等格式的图形属于矢量图形。
- 音频：音频是指存储声音内容的文件，把音频文件用一定的音频程序播放，就可以还

原以前录下的声音。音频文件的格式很多,包括 CD、WAV、MP3、MID、WMA、RM 等。

- 视频:视频是通常指各种动态影像的储存格式,包括 MPEG-4、AVI、DAT、RM、MOV、ASF、WMV、DivX 等。

1.1.3 数据组织形式

计算机系统中的数据组织形式主要有两种,即文件和数据库。

(1) 文件:在计算机系统中,很多数据都是以文件形式存在的,比如文本文件、网页文件、图片文件等。文件的文件名包含主名和扩展名,扩展名用来表示文件的类型,比如文本文档、图片、音频、视频等。在计算机中,文件是由文件系统负责管理的。

(2) 数据库:计算机系统中另一种非常重要的数据组织形式就是数据库,今天,数据库已经成为计算机软件开发的基础和核心,数据库在人力资源管理、固定资产管理、制造业管理、电信管理、销售管理、售票管理、银行管理、股市管理、教学管理、图书馆管理、政务管理等领域发挥着至关重要的作用。从 1968 年 IBM 公司推出第一个大型商用数据库管理系统 IMS 至今,人类社会已经经历了层次数据库、网状数据库、关系数据库和 NoSQL 数据库等多个数据库发展阶段。关系数据库仍然是目前的主流数据库,大多数商业应用系统都构建在关系数据库基础之上。但是,随着 Web 2.0 的兴起,非结构化数据迅速增加,目前人类社会产生的数字内容中有 90% 是非结构化数据,因此,能够更好地支持非结构化数据管理的 NoSQL 数据库应运而生。

1.1.4 数据生命周期

数据都存在生命周期,数据生命周期是指数据从创建、修改、发布利用到归档或销毁的整个过程。在不同的时期内,数据的利用价值也会不同。为了充分发挥存储设备和数据的价值,需要对数据生命周期进行认真分析,在不同阶段对数据采用不同的数据存储管理方式。

数据生命周期管理工作主要包括以下几个方面。

- 对数据进行自动分类,分离出有效的数据,对不同类型数据制定不同的管理策略,并及时清理无用的数据。
- 构建分层的存储系统,满足不同类型的数据在不同生命周期阶段的存储要求,对关键数据进行数据备份保护,将处于生命期末期的数据进行归档并保存到适合长期保存数据的存储设备中。
- 根据不同的数据管理策略,实施自动分层数据管理,即自动把不同生命周期阶段的数据存放在最合适的存储设备上,提高数据的可用性和管理效率。

1.1.5 数据的使用

日常生活中存在着各种各样的数据,那么,如何把数据变得可用呢?

第一步:数据清洗。数据分析的第一步都是数据清洗,即把数据变成一个可用的状态。这个过程需要借助工具去实现数据转换,比如 UNIX 工具 AWK、XML 解析器和机器学习库等。此外,脚本语言,比如 Perl 和 Python,也可以在这个过程中发挥重要的作用。完成数据转换之后,就要开始关注数据的质量。对于来源众多、类型多样的数据而言,数据缺失和语义模

糊等问题是不可避免的，必须采取有效措施解决这些问题。

第二步：数据管理。数据经过清洗以后，被存放到数据库系统中进行管理和使用。从 20 世纪 70 年代到 21 世纪前 10 年，在数据库管理系统中，关系型数据库一直占据主流地位，它以规范化的行和列的形式保存数据，并提供 SQL 语句进行各种查询操作，同时支持事务一致性功能，很好地满足了各种商业应用需求。但是，随着 Web 2.0 应用的不断发展，非结构化数据开始迅速增加，关系型数据库擅长管理结构化数据，对于大规模非结构化数据则显得力不从心，暴露了很多问题。NoSQL 数据库的出现，有效满足了对非结构化数据进行管理的市场需求，使得它得到了非常迅速的发展。

第三步：数据分析。存储数据是为了更好地分析数据，分析数据需要借助于数据挖掘和机器学习算法，也需要使用相关的大数据处理技术。Google 提出了面向大规模数据分析的分布式编程模型 MapReduce，Hadoop 对其进行了开源实现。MapReduce 将复杂的、运行于大规模集群上的并行计算过程高度地抽象到了两个函数——Map 和 Reduce，MapReduce 作业通常会把输入的数据集切分为若干个独立的数据块，由 Map 函数以完全并行的方式处理它们，大大提高了数据分析的速度。此外，构建统计模型对于数据分析也十分重要。统计是数据分析的重要方式，在众多开源的统计分析工具中，R 语言和它的综合类库 CRAN 是十分重要的。为了能够让数据说话，使得分析结果更容易被理解，还需要对分析结果进行可视化。可视化可以帮助人们直观地理解数据。

这里以数据仓库为例演示数据在企业中的使用方法。很多企业为了支持决策分析会构建数据仓库系统，如图 1-2 所示，其中会存放大量的历史数据，这些数据来自不同的数据源，利用抽取、转换、装载（extract transform load, ETL）方法加载到数据仓库中，并且不会发生更新，技术人员可以利用数据挖掘和联机分析处理（online analytical processing, OLAP）工具从静态历史数据中找到对企业有价值的信息。

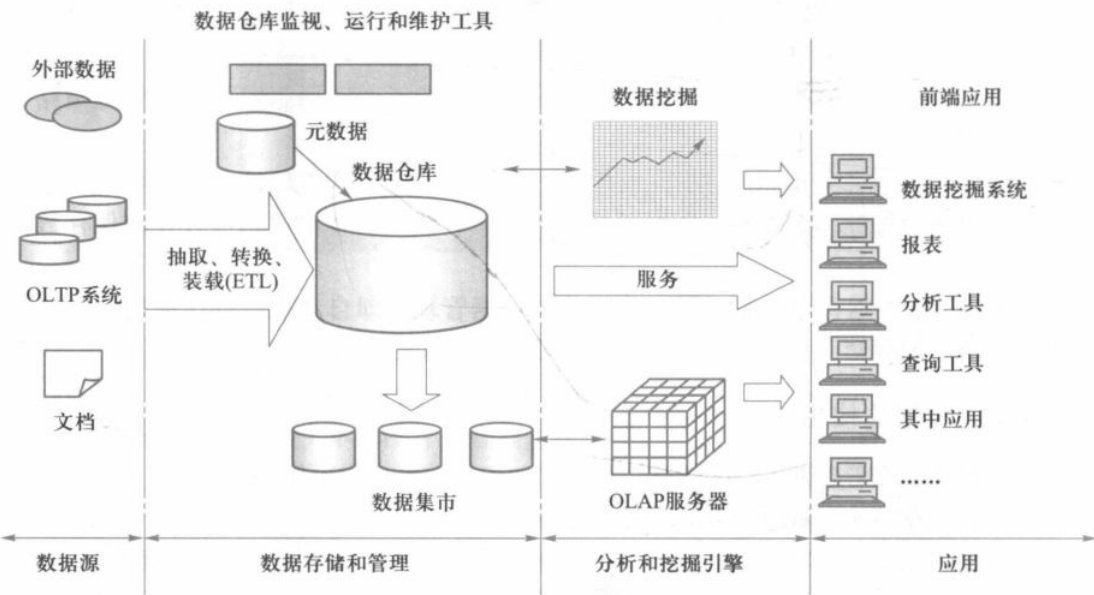


图 1-2 数据仓库体系架构图

1.1.6 数据的价值

数据的根本价值在于可以为人们找出答案。收集数据往往都是为了某个特定的目的,对于数据收集者而言,数据的价值是显而易见的,且是不断被人发现的。在过去,一旦数据发挥了基本作用,往往就会被删除。一方面是由于过去的存储技术落后,人们需要删除旧数据来存储新数据;另一方面则是由于人们没有认识到数据的潜在价值。例如,在淘宝或者京东搜索一件衣服,当输入性别、颜色、布料、款式等关键词之后,消费者很容易就会找到心仪的产品,当购买行为结束之后,这些数据就会被消费者删除。但是,对于这些购物网站,它们会记录和整理这些购买数据,当收集到海量的购买信息后,就可以预测未来流行的产品特征。网络公司可以把这些信息有偿提供给各类生产商,帮助生产商在竞争中脱颖而出,这就是数据价值的再发现。

数据的价值不会因为不断使用而削减,反而会因为不断重组而产生更大的价值。例如,将一个地区的物价、地价走势、高档轿车的销售数量、二手房转手的频率、出租车密度等各种不相关的数据整合到一起,可以更加精准地预测该地区房价走势。这种方式已经被国外很多房地产网站采用。而这些数据,也可以被重新整合。基于数据的价值特性,各类收集来的数据都应当被尽可能长时间地保存下来,同时也应当在一定条件下与全社会分享,并产生价值。因为数据的潜在价值,往往是收集者不可想象的。当今社会,人们已经逐步产生了一种认识,在大数据时代,今天和未来最有价值的商品是数据。目前占有大量数据的谷歌、亚马逊等全球前五大公司,每个季度的利润总和高达数十亿美元,并在继续快速增加,这都是数据价值的最好佐证。因此,要实现大数据时代思维方式的转变,就必须正确认识数据的价值,数据已经具备了资本的属性,可以用来创造经济价值。

1.1.7 数据爆炸

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,全球的数据量增长了100倍,今后的数据量增长速度将更快,我们正生活在一个“数据爆炸”的时代。随着Web 2.0和移动互联网的快速发展,人们已经可以随时随地发布博客、微博、微信、抖音等各种信息。在1分钟内,新浪可以产生2万条微博, Twitter可以产生10万条推文,苹果可以产生4.7万次应用下载,淘宝可以卖出6万件商品,百度可以产生90万次搜索查询, Facebook可以产生600万次浏览量。以后,随着物联网的推广和普及,各种传感器和摄像头将遍布人们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数据。如果把印刷在纸上的文字和图形也看作数据的话,那么,人类历史上第一次数据爆炸发生在造纸术和印刷术发明的时期,而现在,人类社会正经历第二次数据爆炸。各种数据产生速度之快,数量之大,数据爆炸成为大数据时代的鲜明特征。

在数据爆炸的今天,人类一方面对知识充满渴求,另一方面对数据的复杂特征感到困惑。数据爆炸对科学研究提出了更高的要求,需要人类设计出更加灵活高效的数据存储、处理和分析工具,来应对大数据时代的挑战。因此,必将带来云计算、数据仓库、数据挖掘等技术和应用的提升或者根本性改变。在存储效率(存储技术)领域,需要实现低成本的大规模分布式存储;在网络效率(网络技术)方面,需要实现及时响应的用户体验;在数据中心方面,

需要开发更加绿色节能的新一代数据中心，在有效面对大数据处理需求的同时，实现最大化资源利用率、最小化系统能耗的目标。

1.2 大数据时代



第三次信息化浪潮涌动，大数据时代全面开启。人类社会信息科技的发展为大数据时代的到来提供了技术支撑，而数据产生方式的变革是促进大数据时代到来至关重要的因素。

1.2.1 第三次信息化浪潮

根据 IBM 前首席执行官路易斯·格斯特纳的观点，IT 领域每隔 15 年就会迎来一次重大变革，见表 1-1。1980 年前后，个人计算机（PC）开始普及，使得计算机进入企业和千家万户，大大提高了社会生产力，也使人类迎来了第一次信息化浪潮，Intel、IBM、苹果、微软、联想等企业的崛起是这个时期的标志。随后，在 1995 年前后，人类开始全面进入互联网时代，互联网的普及把世界变成“地球村”，每个人都可以自由徜徉于信息的海洋，由此，人类迎来了第二次信息化浪潮，这个时期也缔造了雅虎、谷歌、阿里巴巴、百度等互联网巨头。时隔 15 年，在 2010 年前后，云计算、大数据、物联网的快速发展，拉开了第三次信息化浪潮的大幕，大数据时代已经到来，也必将涌现出一批新的市场标杆企业。

表 1-1 三次信息化浪潮

信息化浪潮	发生时间	标 志	解决的问题	代表企业
第一次浪潮	1980 年前后	个人计算机	信息处理	Intel、AMD、IBM、苹果、微软、联想、戴尔、惠普等
第二次浪潮	1995 年前后	互联网	信息传输	雅虎、谷歌、阿里巴巴、百度、腾讯等
第三次浪潮	2010 年前后	物联网、云计算和大数据	信息爆炸	亚马逊、谷歌、IBM、VMWare、Palantir、Hortonworks、Cloudera、阿里云等

1.2.2 信息科技为大数据时代提供技术支撑

大数据，首先是一场技术革命。毫无疑问，如果没有强大的数据存储、传输和计算等技术能力，缺乏必要的设施设备，大数据的应用也就无从谈起。从这个意义上说，信息科技进步是大数据时代的物质基础。信息科技需要解决信息存储、信息传输和信息处理 3 个核心问题，人类社会在信息科技领域的不断进步，为大数据时代的到来提供了技术支撑。

1. 存储设备容量不断增加

数据被存储在磁盘、磁带、光盘、闪存等各种类型的存储介质中，随着科学技术的不断进步，存储设备制造工艺不断升级，容量大幅增加，速度不断提升，价格却在不断下降，如图 1-3 所示。

早期的存储设备容量小、价格高、体积大，例如，IBM 在 1956 年生产的一个早期的商业硬盘，容量只有 5 MB，不仅价格昂贵，而且体积类似于一个冰箱，如图 1-4 所示。相反，今天容量为 1 TB 的硬盘，体积很小，读写速度可以达到 200 Mb/s，价格仅为 300 元左右。廉价、高性能的硬盘存储设备，不仅提供了海量的存储空间，同时大大降低了数据存储成本。

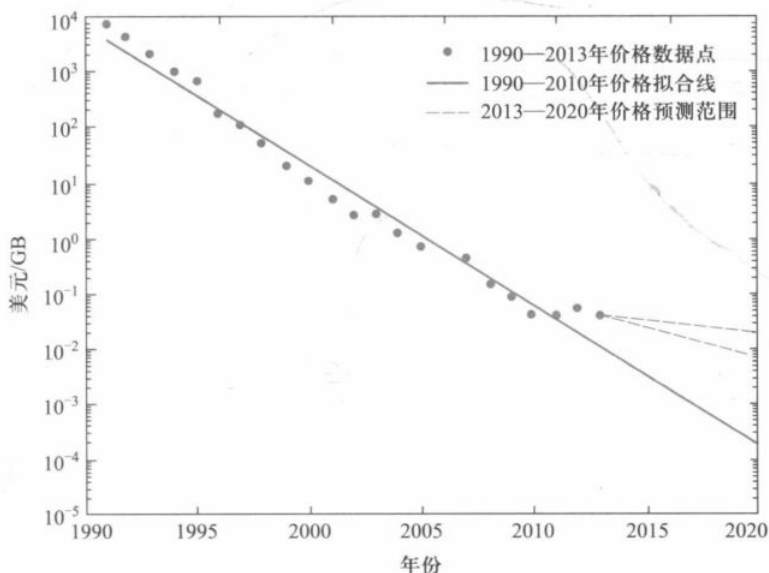


图 1-3 存储价格随时间的变化情况

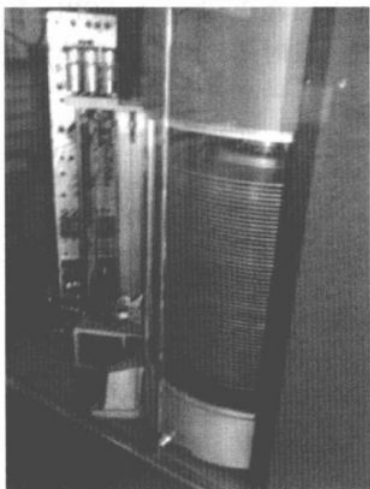


图 1-4 IBM 在 1956 年生产的一个早期的商业硬盘

与此同时，以闪存为代表的新型存储介质也开始得到大规模的普及和应用。闪存是一种新兴的半导体存储器，从 1989 年诞生第一款闪存产品开始，闪存技术不断获得新的突破，并逐渐在计算机存储产品市场中确立了自己的重要地位。闪存是一种非易失性存储器，即使发生断电也不会丢失数据，因此，可以作为永久性存储设备，它具有体积小、质量轻、能耗低、抗震性好等优良特性。

闪存芯片可以被封装制作成 SD 卡、U 盘和固态硬盘等各种存储产品，SD 卡和 U 盘主要用于个人数据存储，固态硬盘则越来越多地应用于企业级数据存储。一个 32 GB 的 SD 卡，体积只有 24 mm×32 mm×2.1 mm，质量只有 0.5 g。以前 7 200 r/min 的硬盘，每秒能进行 100 次读写操作，即 IOPS（input/output operations per second）值为 100，速率只有 50 Mb/s。而现在的基于闪存的固态硬盘，IOPS 值为几万甚至更高，访问延迟只有几十微秒，数据读写速度更快。

总体而言，数据量和存储设备容量二者之间是相辅相成、互相促进的。一方面，随着数据的不断产生，需要存储的数据量不断增加，对存储设备的容量提出了更高的要求，促使存储设备生产商制造更大容量的产品满足市场需求；另一方面，更大容量的存储设备，进一步加快了数据量增长的速度。在存储设备价格高昂的年代，考虑到成本问题，一些不必要或当前不能明显体现价值的数据往往会被丢弃。但是，随着单位存储空间价格的不断降低，人们