

21世纪高等教育计算机规划教材

COMPUTER

大数据技术原理与应用

——概念、存储、处理、分析与应用

Principles and Applications of Big Data Technology - Big Data
Conception, Storage, Processing, Analysis and Application

林子雨 编著

搭建起通向“大数据知识空间”的桥梁和纽带

构建知识体系、阐明基本原理、引导初级实践、了解相关应用

为读者在大数据领域“深耕细作”奠定基础、指明方向



中国工信出版集团



人民邮电出版社
POSTS & TELECOM PRESS

21世纪高等教育计算机规划教材

COMPUTER

大数据技术原理与应用

——概念、存储、处理、分析与应用

Principles and Applications of Big Data Technology—Big Data
Conception, Storage, Processing, Analysis and Application

■ 林子雨 编著



人民邮电出版社

北京

图书在版编目(CIP)数据

大数据技术原理与应用：概念、存储、处理、分析与应用 / 林子雨编著. — 北京：人民邮电出版社，2015.8

21世纪高等教育计算机规划教材
ISBN 978-7-115-39287-9

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(2015)第114600号

内 容 提 要

本书系统介绍了大数据的相关知识，分为大数据基础篇、大数据存储篇、大数据处理与分析篇、大数据应用篇。全书共13章，内容包含大数据的基本概念、大数据处理架构(Hadoop)、分布式文件系统(HDFS)、分布式数据库(HBase)、NoSQL数据库、云数据库、分布式并行编程模型(MapReduce)、流计算、图计算、数据可视化，以及大数据在互联网、生物医学和物流等领域的应用。本书在Hadoop、HDFS、HBase和MapReduce等重要章节安排了入门级的实践操作，以便读者更好地学习和掌握大数据关键技术。

本书可以作为高等院校计算机、信息管理等相关专业的大数据课程教材，也可供相关技术人员参考。



-
- ◆ 编 著 林子雨
责任编辑 邹文波
执行编辑 吴 婷
责任印制 焦志炜
 - ◆ 人民邮电出版社出版发行 北京市丰台区成寿寺路11号
邮编 100164 电子邮件 315@ptpress.com.cn
网址 <http://www.ptpress.com.cn>
北京中新伟业印刷有限公司印刷
 - ◆ 开本：787×1092 1/16
印张：16.25
字数：423千字

2015年8月第1版

2015年8月北京第1次印刷

定价：45.00元

读者服务热线：(010)81055256 印装质量热线：(010)81055316
反盗版热线：(010)81055315

前言

大数据作为继云计算、物联网之后 IT 行业又一颠覆性的技术，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，都融入了大数据的印迹，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。

大数据时代的到来，迫切需要高校及时建立大数据技术课程体系，为社会培养和输送一大批具备大数据专业素养的高级人才，满足社会对大数据人才日益旺盛的需求。本书定位为大数据技术入门教材，为读者搭建起通向“大数据知识空间”的桥梁和纽带。本书将系统梳理、总结大数据相关技术，介绍大数据技术的基本原理和大数据的主要应用，帮助读者形成对大数据知识体系及其应用领域的轮廓性认识，为读者在大数据领域“深耕细作”奠定基础、指明方向。在本书的基础上，感兴趣的读者可以通过其他诸如《Hadoop 权威指南》等工具书，继续深入学习和实践大数据相关技术。



本书紧紧围绕“构建知识体系，阐明基本原理，引导初级实践，了解相关应用”的指导思想，对大数据知识体系进行系统梳理，做到“有序组织、去粗取精、由浅入深、渐次展开”。本书共分四大部分，包括大数据基础篇、大数据存储篇、大数据处理与分析篇和大数据应用篇。在大数据基础篇中，第1章介绍大数据的基本概念和应用领域，并阐述大数据、云计算和物联网的相互关系；第2章介绍大数据处理架构 Hadoop，由于 Hadoop 已经成为应用最广泛的大数据技术，因此，本书的大数据相关技术主要围绕 Hadoop 展开，包括 Hadoop MapReduce、HDFS 和 HBase，本章

是第3、4、7章的基础。在大数据存储篇中，用4章（第3~6章）的内容介绍了大数据存储相关技术的概念与原理，包括分布式文件系统（HDFS）、分布式数据库（HBase）、NoSQL数据库和云数据库。在大数据处理与分析篇，首先在第7章介绍了大数据处理和分析的核心技术——分布式并行编程模型 MapReduce，然后，在第8章和第9章分别介绍了大数据时代两种新兴的数据分析技术——流计算和图计算，最后在第10章简单介绍了可视化技术。在大数据应用篇，用3章（第11章~第13章）内容介绍了大数据在互联网、生物医学和物流等领域的典型应用。

本书面向高校计算机和信息管理等相关专业的学生，可以作为专业必修课或选修课的教材。在教学过程中，建议安排32学时，16个教学周，每周2学时。每章的具体学时分配如下：第1、2、5、6、8、10、11章每章安排2学时；第3、4、9章每章安排4学时；第7章安排6学时；第12、13章这两章内容由学生自学完成。

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刘颖杰、叶林宝、蔡珉星、李雨倩、谢荣东、罗道文以及本科生黄梓铭、李粲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。

本书官方网站是 <http://dblab.xmu.edu.cn/post/bigdata>，提供教学PPT和相关资料的下载，并接受错误反馈和发布教材勘误信息。

本书在撰写过程中，参考了大量国内外的教材、专著、论文和资料，对大数据知识进行了系统梳理，有选择地把一些重要知识纳入本书。本书也是笔者多年在数据科学领域从事教学、科研、产业方面工作的系统总结。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

厦门大学计算机科学系数据库实验室

2015年3月

作者介绍



林子雨 (1978 -), 男, 博士。厦门大学计算机科学系助理教授, 主要研究领域为大数据、数据库、数据仓库和数据挖掘。

主讲课程:《大数据技术基础》。

个人主页: <http://www.cs.xmu.edu.cn/linziyu>。

E-mail: ziyulin@xmu.edu.cn。数据库实验室网站: <http://dblab.xmu.edu.cn>。

林子雨博士, 厦门大学计算机科学系助理教授, 中国高校首个“数字教师”的提出者和建设者。于 2001 年获得福州大学水利水电专业学士学位, 2005 年获得厦门大学计算机专业硕士学位, 2009 年获得北京大学计算机专业博士学位。主要研究方向为数据库、数据仓库、数据挖掘、大数据和云计算, 发表期刊和会议学术论文多篇。曾作为志愿者翻译了《Google Spanner》、《BigTable》和《Architecture of a Database System》等英文学术资料, 并与广大网友分享, 深受欢迎。2013 年在厦门大学开设了《大数据技术基础》课程, 并因在教学领域的突出贡献和学生的认可, 成为 2013 年度厦门大学教学类奖教金获得者。

目 录

第一篇 大数据基础

第 1 章 大数据概述 2

1.1 大数据时代..... 2	
1.1.1 第三次信息化浪潮..... 2	
1.1.2 信息科技为大数据时代 提供技术支撑..... 3	
1.1.3 数据产生方式的变革促成 大数据时代的来临..... 5	
1.1.4 大数据的发展历程..... 6	
1.2 大数据的概念..... 7	
1.2.1 数据量大..... 7	
1.2.2 数据类型繁多..... 8	
1.2.3 处理速度快..... 8	
1.2.4 价值密度低..... 9	
1.3 大数据的影响..... 9	
1.3.1 大数据对科学研究的影响..... 9	
1.3.2 大数据对思维方式的影响..... 10	
1.3.3 大数据对社会发展的影响..... 11	
1.3.4 大数据对就业市场的影响..... 12	
1.3.5 大数据对人才培养的影响..... 12	
1.4 大数据的应用..... 13	
1.5 大数据关键技术..... 14	
1.6 大数据计算模式..... 14	
1.6.1 批处理计算..... 15	
1.6.2 流计算..... 15	
1.6.3 图计算..... 15	
1.6.4 查询分析计算..... 16	
1.7 大数据产业..... 16	
1.8 大数据与云计算、物联网..... 17	
1.8.1 云计算..... 17	
1.8.2 物联网..... 20	

1.8.3 大数据与云计算、物联网的关系..... 24	
1.9 本章小结..... 25	
1.10 习题..... 25	

第 2 章 大数据处理架构 Hadoop .. 26

2.1 概述..... 26	
2.1.1 Hadoop 简介..... 26	
2.1.2 Hadoop 的发展简史..... 26	
2.1.3 Hadoop 的特性..... 27	
2.1.4 Hadoop 的应用现状..... 27	
2.2 Hadoop 的项目结构..... 28	
2.2.1 Common..... 28	
2.2.2 Avro..... 29	
2.2.3 HDFS..... 29	
2.2.4 HBase..... 29	
2.2.5 MapReduce..... 29	
2.2.6 Zookeeper..... 30	
2.2.7 Hive..... 30	
2.2.8 Pig..... 30	
2.2.9 Sqoop..... 30	
2.2.10 Chukwa..... 30	
2.3 Hadoop 的安装与使用..... 31	
2.3.1 创建 Hadoop 用户..... 31	
2.3.2 Java 的安装..... 31	
2.3.3 SSH 登录权限设置..... 32	
2.3.4 安装单机 Hadoop..... 32	
2.3.5 Hadoop 伪分布式安装..... 33	
2.4 本章小结..... 35	
2.5 习题..... 36	

第二篇 大数据存储

第 3 章 Hadoop 分布式文件系统 .. 40

3.1 分布式文件系统..... 40	
---------------------	--

3.1.1 计算机集群结构.....	40	4.3.6 面向列的存储.....	65
3.1.2 分布式文件系统的结构.....	41	4.4 HBase 的实现原理.....	67
3.1.3 分布式文件系统的设计需求.....	42	4.4.1 HBase 的功能组件.....	67
3.2 HDFS 简介.....	42	4.4.2 表和 Region.....	68
3.3 HDFS 的相关概念.....	43	4.4.3 Region 的定位.....	69
3.3.1 块.....	43	4.5 HBase 运行机制.....	70
3.3.2 名称节点和数据节点.....	44	4.5.1 HBase 系统架构.....	70
3.4 HDFS 体系结构.....	45	4.5.2 Region 服务器的工作原理.....	72
3.4.1 概述.....	45	4.5.3 Store 的工作原理.....	73
3.4.2 HDFS 命名空间管理.....	46	4.5.4 HLog 的工作原理.....	73
3.4.3 通信协议.....	46	4.6 HBase 编程实践.....	74
3.4.4 客户端.....	46	4.6.1 HBase 常用的 Shell 命令.....	74
3.4.5 HDFS 体系结构的局限性.....	46	4.6.2 HBase 常用的 Java API 及 应用实例.....	76
3.5 HDFS 的存储原理.....	46	4.7 本章小结.....	86
3.5.1 冗余数据的保存.....	46	4.8 习题.....	86
3.5.2 数据存取策略.....	47	第 5 章 NoSQL 数据库..... 90	
3.5.3 数据错误与恢复.....	48	5.1 NoSQL 简介.....	90
3.6 HDFS 的数据读写过程.....	49	5.2 NoSQL 兴起的原因.....	91
3.6.1 读数据的过程.....	49	5.2.1 关系数据库无法满足 Web 2.0 的 需求.....	92
3.6.2 写数据的过程.....	50	5.2.2 关系数据库的关键特性在 Web 2.0 时代成为“鸡肋”.....	92
3.7 HDFS 编程实践.....	51	5.3 NoSQL 与关系数据库的比较.....	93
3.7.1 HDFS 常用命令.....	51	5.4 NoSQL 的四大类型.....	95
3.7.2 HDFS 的 Web 界面.....	52	5.4.1 键值数据库.....	96
3.7.3 HDFS 常用 Java API 及应用实例.....	53	5.4.2 列族数据库.....	96
3.8 本章小结.....	56	5.4.3 文档数据库.....	97
3.9 习题.....	57	5.4.4 图形数据库.....	97
第 4 章 分布式数据库 HBase 59		5.5 NoSQL 的三大基石.....	98
4.1 概述.....	59	5.5.1 CAP.....	98
4.1.1 从 BigTable 说起.....	59	5.5.2 BASE.....	100
4.1.2 HBase 简介.....	59	5.5.3 最终一致性.....	101
4.1.3 HBase 与传统关系数据库的 对比分析.....	60	5.6 从 NoSQL 到 NewSQL 数据库.....	102
4.2 HBase 访问接口.....	61	5.7 本章小结.....	104
4.3 HBase 数据模型.....	62	5.8 习题.....	104
4.3.1 数据模型概述.....	62	第 6 章 云数据库 105	
4.3.2 数据模型的相关概念.....	62	6.1 云数据库概述.....	105
4.3.3 数据坐标.....	64		
4.3.4 概念视图.....	64		
4.3.5 物理视图.....	65		

6.1.1 云计算是云数据库兴起的基础.....	105	7.3.3 MapReduce 的具体执行过程.....	136
6.1.2 云数据库的概念.....	106	7.3.4 一个 WordCount 执行过程的实例.....	137
6.1.3 云数据库的特性.....	107	7.4 MapReduce 的具体应用.....	139
6.1.4 云数据库是个性化数据存储需求的理想选择.....	108	7.4.1 MapReduce 在关系代数运算中的应用.....	139
6.1.5 云数据库与其他数据库的关系.....	109	7.4.2 分组与聚合运算.....	140
6.2 云数据库产品.....	110	7.4.3 矩阵-向量乘法.....	140
6.2.1 云数据库厂商概述.....	110	7.4.4 矩阵乘法.....	141
6.2.2 Amazon 的云数据库产品.....	110	7.5 MapReduce 编程实践.....	141
6.2.3 Google 的云数据库产品.....	111	7.5.1 任务要求.....	141
6.2.4 微软的云数据库产品.....	111	7.5.2 编写 Map 处理逻辑.....	142
6.2.5 其他云数据库产品.....	112	7.5.3 编写 Reduce 处理逻辑.....	143
6.3 云数据库系统架构.....	112	7.5.4 编写 main 方法.....	144
6.3.1 UMP 系统概述.....	112	7.5.5 编译打包代码以及运行程序.....	144
6.3.2 UMP 系统架构.....	113	7.6 本章小结.....	146
6.3.3 UMP 系统功能.....	115	7.7 习题.....	147
6.4 云数据库实践.....	118	第 8 章 流计算..... 151	
6.4.1 阿里云 RDS 简介.....	118	8.1 流计算概述.....	151
6.4.2 RDS 中的概念.....	118	8.1.1 静态数据和流数据.....	151
6.4.3 购买和使用 RDS 数据库.....	119	8.1.2 批量计算和实时计算.....	152
6.4.4 将本地数据库迁移到云端 RDS 数据库.....	123	8.1.3 流计算的概念.....	153
6.5 本章小结.....	124	8.1.4 流计算与 Hadoop.....	153
6.6 习题.....	125	8.1.5 流计算框架.....	154
第三篇 大数据处理与分析		8.2 流计算的处理流程.....	154
第 7 章 MapReduce..... 128		8.2.1 概述.....	154
7.1 概述.....	128	8.2.2 数据实时采集.....	155
7.1.1 分布式并行编程.....	128	8.2.3 数据实时计算.....	155
7.1.2 MapReduce 模型简介.....	129	8.2.4 实时查询服务.....	156
7.1.3 Map 和 Reduce 函数.....	129	8.3 流计算的应用.....	156
7.2 MapReduce 的工作流程.....	130	8.3.1 应用场景 1: 实时分析.....	156
7.2.1 工作流程概述.....	130	8.3.2 应用场景 2: 实时交通.....	157
7.2.2 MapReduce 的各个执行阶段.....	131	8.4 开源流计算框架 Storm.....	158
7.2.3 Shuffle 过程详解.....	132	8.4.1 Storm 简介.....	159
7.3 实例分析: WordCount.....	135	8.4.2 Storm 的特点.....	159
7.3.1 WordCount 的程序任务.....	135	8.4.3 Storm 的设计思想.....	160
7.3.2 WordCount 的设计思路.....	136	8.4.4 Storm 的框架设计.....	161
		8.4.5 Storm 实例.....	162
		8.4.6 哪些公司在使用 Storm.....	165

8.5	本章小结	166
8.6	习题	166
第9章 图计算 168		
9.1	图计算简介	168
9.1.1	传统图计算解决方案的 不足之处	168
9.1.2	图计算通用软件	169
9.2	Pregel 简介	169
9.3	Pregel 图计算模型	170
9.3.1	有向图和顶点	170
9.3.2	顶点之间的消息传递	170
9.3.3	Pregel 的计算过程	171
9.3.4	实例	171
9.4	Pregel 的 C++ API	174
9.4.1	消息传递机制	174
9.4.2	Combiner	175
9.4.3	Aggregator	175
9.4.4	拓扑改变	176
9.4.5	输入和输出	176
9.5	Pregel 的体系结构	176
9.5.1	Pregel 的执行过程	177
9.5.2	容错性	178
9.5.3	Worker	179
9.5.4	Master	179
9.5.5	Aggregator	180
9.6	Pregel 的应用实例	180
9.6.1	单源最短路径	180
9.6.2	二分匹配	181
9.7	Pregel 和 MapReduce 实现 PageRank 算法的对比	182
9.7.1	PageRank 算法	183
9.7.2	PageRank 算法在 Pregel 中的 实现	183
9.7.3	PageRank 算法在 MapReduce 中的 实现	184
9.7.4	PageRank 算法在 Pregel 和 MapReduce 中实现的比较	186
9.8	本章小结	187
9.9	习题	187

第10章 数据可视化 189		
10.1	可视化概述	189
10.1.1	什么是数据可视化	189
10.1.2	可视化的发展历程	189
10.1.3	可视化的重要作用	191
10.2	可视化工具	193
10.2.1	入门级工具	193
10.2.2	信息图表工具	194
10.2.3	地图工具	195
10.2.4	时间线工具	196
10.2.5	高级分析工具	196
10.3	可视化典型案例	197
10.3.1	全球黑客活动	197
10.3.2	互联网地图	197
10.3.3	编程语言之间的影响力 关系图	198
10.3.4	百度迁徙	199
10.3.5	世界国家健康与财富之间的 关系	199
10.3.6	3D 可视化互联网地图 APP	199
10.4	本章小结	201
10.5	习题	201

第四篇 大数据应用

第11章 大数据在互联网领域的 应用 204		
11.1	推荐系统概述	204
11.1.1	什么是推荐系统	204
11.1.2	长尾理论	205
11.1.3	推荐方法	205
11.1.4	推荐系统模型	206
11.1.5	推荐系统的应用	206
11.2	协同过滤	207
11.2.1	基于用户的协同过滤	207
11.2.2	基于物品的协同过滤	209
11.2.3	UserCF 算法和 ItemCF 算法的 对比	210

11.3 协同过滤实践.....	211	13.2 大数据在城市管理中的应用.....	229
11.3.1 实践背景.....	211	13.2.1 智能交通.....	230
11.3.2 数据处理.....	211	13.2.2 环保监测.....	231
11.3.3 计算相似度矩阵.....	212	13.2.3 城市规划.....	232
11.3.4 计算推荐结果.....	213	13.2.4 安防领域.....	232
11.3.5 展示推荐结果.....	213	13.3 大数据在金融行业中的应用.....	233
11.4 本章小结.....	214	13.3.1 高频交易.....	233
11.5 习题.....	214	13.3.2 市场情绪分析.....	233
第 12 章 大数据在生物医学		13.3.3 信贷风险分析.....	234
领域的应用.....	215	13.4 大数据在汽车行业中的应用.....	235
12.1 流行病预测.....	215	13.5 大数据在零售行业中的应用.....	236
12.1.1 传统流行病预测机制的不足.....	215	13.5.1 发现关联购买行为.....	236
12.1.2 基于大数据的流行病预测.....	216	13.5.2 客户群体细分.....	236
12.1.3 基于大数据的流行病预测的		13.5.3 供应链管理.....	237
重要作用.....	217	13.6 大数据在餐饮行业中的应用.....	237
12.1.4 案例:百度疾病预测.....	217	13.6.1 餐饮行业拥抱大数据.....	237
12.2 智慧医疗.....	218	13.6.2 餐饮 O2O.....	238
12.3 生物信息学.....	219	13.7 大数据在电信行业中的应用.....	239
12.4 案例:基于大数据的综合健康服务		13.8 大数据在能源行业中的应用.....	240
平台.....	220	13.9 大数据在体育和娱乐领域中的应用.....	241
12.4.1 平台概述.....	220	13.9.1 训练球队.....	241
12.4.2 平台业务架构.....	221	13.9.2 投拍影视作品.....	242
12.4.3 平台技术架构.....	222	13.9.3 预测比赛结果.....	243
12.4.4 平台关键技术.....	223	13.10 大数据在安全领域中的应用.....	243
12.5 本章小结.....	224	13.10.1 大数据与国家安全.....	243
12.6 习题.....	224	13.10.2 应用大数据技术防御网络	
第 13 章 大数据的其他应用.....	225	攻击.....	244
13.1 大数据在物流领域中的应用.....	225	13.10.3 警察应用大数据工具预防	
13.1.1 智能物流的概念.....	225	犯罪.....	245
13.1.2 智能物流的作用.....	226	13.11 大数据在政府领域中的应用.....	246
13.1.3 智能物流的应用.....	226	13.12 大数据在日常生活中的应用.....	246
13.1.4 大数据是智能物流的关键.....	227	13.13 本章小结.....	247
13.1.5 中国智能物流骨干网——菜鸟.....	227	13.14 习题.....	248
		参考文献.....	249

第一篇

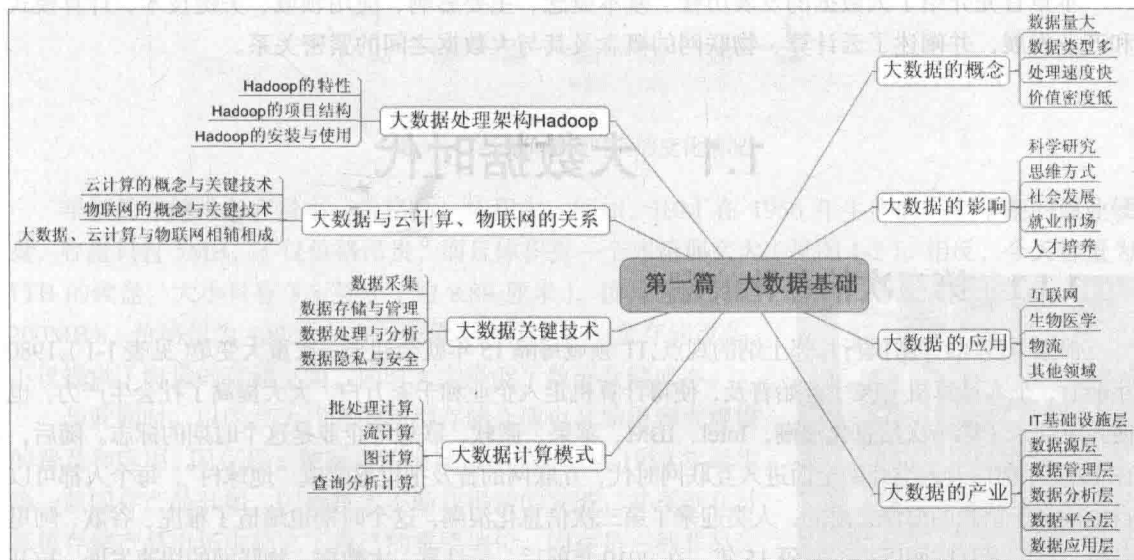
大数据基础

本篇内容

本篇介绍大数据 (Big Data) 的基本概念、影响和应用领域,并阐述大数据、云计算和物联网的相互关系,还介绍了大数据处理架构 Hadoop。由于 Hadoop 已经成为应用最广泛的大数据技术,因此,本书的大数据相关技术主要围绕 Hadoop 展开,包括 Hadoop MapReduce、HDFS 和 HBase。本篇内容是学习后续内容的基础。

本篇包括两章。第 1 章介绍大数据的概念和应用,分析了大数据、云计算和物联网的相互关系;第 2 章介绍大数据处理架构 Hadoop。

知识地图



重点与难点

重点为理解大数据的概念,大数据对科学研究、思维方式和社会发展的影响,以及大数据处理架构 Hadoop。难点为掌握 Hadoop 的安装与使用方法。

第1章

大数据概述

大数据时代悄然来临,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民生活的方方面面。全球范围内,世界各国政府均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加以重点推进。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。

对于一个国家而言,能否紧紧抓住大数据发展机遇,快速形成核心技术和应用参与新一轮的全球化竞争,将直接决定未来若干年世界范围内各国科技力量博弈的格局。大数据专业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,因此,各高等院校非常重视大数据课程的开设,大数据课程已经成为计算机科学与技术专业的重要核心课程。

本章首先介绍了大数据的发展历程、基本概念、主要影响、应用领域、关键技术、计算模式和产业发展,并阐述了云计算、物联网的概念及其与大数据之间的紧密关系。

1.1 大数据时代

1.1.1 第三次信息化浪潮

根据 IBM 前首席执行官郭士纳的观点,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,使得计算机走入企业和千家万户,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

续表

信息化浪潮	发生时间	标志	解决的问题	代表企业
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	将涌现出一批新的市场标杆企业

1.1.2 信息科技为大数据时代提供技术支撑

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（见图 1-1）。

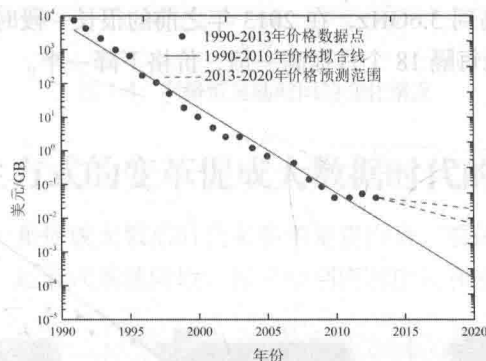


图 1-1 存储价格随时间的变化情况

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 5MB，不仅价格昂贵，而且体积有一个冰箱那么大（见图 1-2）。相反，今天容量为 1TB 的硬盘，大小只有 3.5 英寸（约 8.89 厘米），读写速度达到 200MB/s，价格仅为 400 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。

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 32GB 的 SD 卡，体积只有

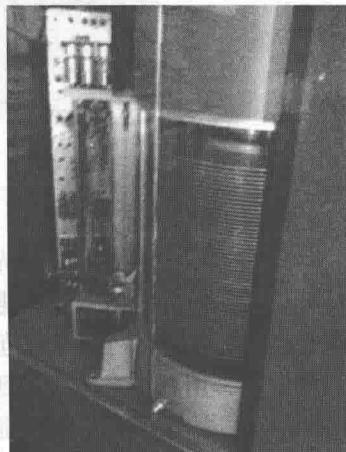


图 1-2 IBM 在 1956 年生产的一个早期的商业硬盘

24mm×32mm×2.1mm, 质量只有 0.5 克。以前 7 200r/min 的硬盘, 一秒钟只有 100 个 IOPS (Input/Output Operations Per Second), 速率只有 50MB/s, 而现在的基于闪存的固态硬盘, 每秒有几万甚至更高的 IOPS, 访问延迟只有几十微秒, 允许我们以更快的速度读写数据。

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, 在存储设备价格高企的年代, 由于考虑到成本问题, 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, 人们开始倾向于把更多的数据保存起来, 以期在未来某个时刻可以用更先进的数据分析工具从中挖掘价值。

2. CPU 处理能力大幅提升

CPU 处理速度的不断提升也是促使数据量不断增加的重要因素。性能不断提升的 CPU, 大大提高了处理数据的能力, 使得我们可以更快地处理不断累积的海量数据。从 20 世纪 80 年代至今, CPU 的制造工艺不断提升, 晶体管数量不断增加 (见图 1-3), 运行频率不断提高, 核心 (Core) 数量逐渐增多, 而同等价格所能获得的 CPU 处理能力也呈几何级数上升。在 30 多年里, CPU 的处理速度已经从 10MHz 提高到 3.6GHz, 在 2013 年之前的很长一段时期, CPU 处理速度的增加一直遵循“摩尔定律”, 性能每隔 18 个月提高一倍, 价格下降一半。

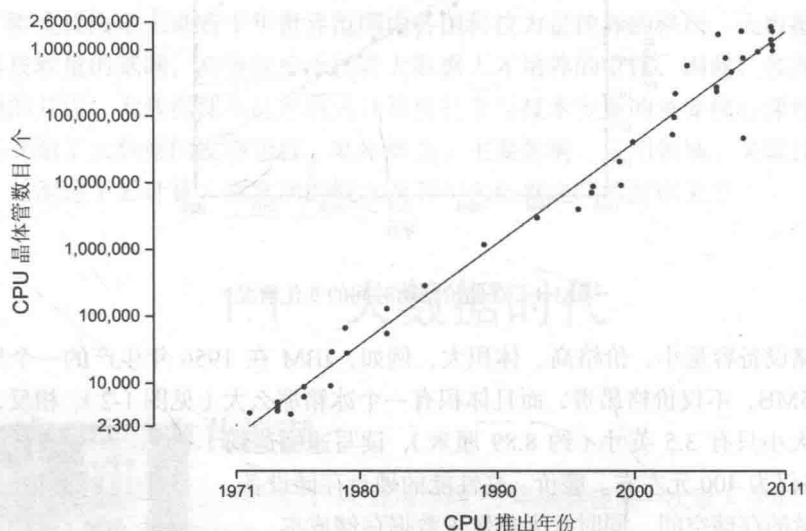


图 1-3 CPU 晶体管数目随时间的变化情况

3. 网络带宽不断增加

1977 年, 世界上第一条光纤通信系统在美国芝加哥市投入商用, 数据传输速率为 45Mbit/s, 从此, 人类社会的信息传输速度不断被刷新。进入 21 世纪, 世界各国更是纷纷加大宽带网络建设力度, 不断扩大网络覆盖范围和传输速度 (见图 1-4)。以我国为例, 截至 2012 年 6 月, 92.6% 的固定宽带用户接入速率达到或超过 2Mbit/s, 国际互联网出口带宽达到 1.48Tbit/s, 是 2005 年的 11.4 倍。与此同时, 移动通信宽带网络迅速发展, 3G 网络基本普及, 4G 网络覆盖范围不断加大, 各种终端设备可以随时随地传输数据。大数据时代, 信息传输不再遭遇网络发展初期的瓶颈和制约。

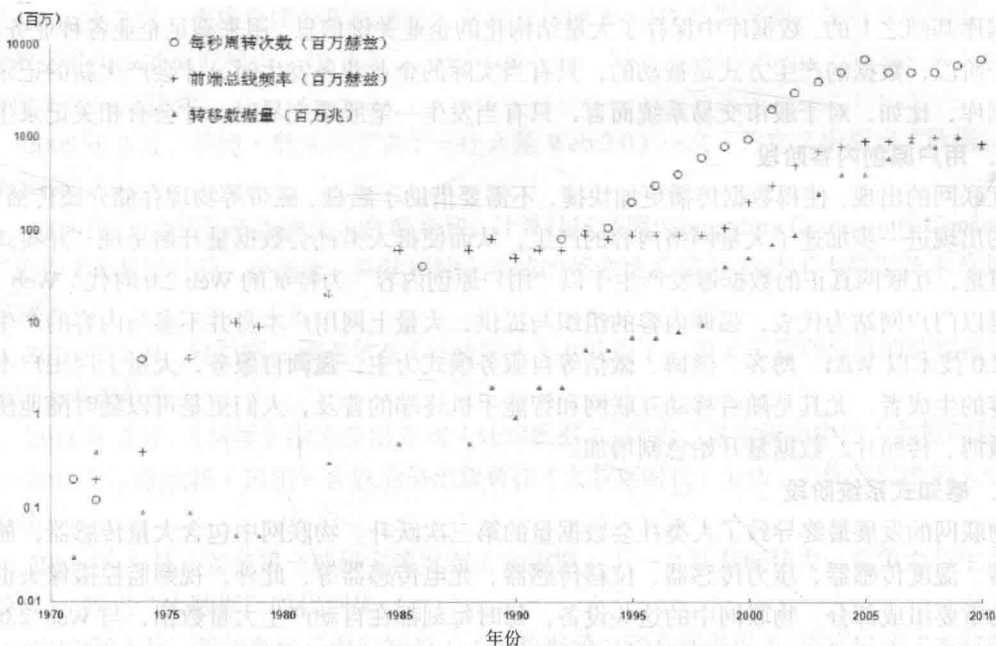


图 1-4 网络带宽随时间的变化情况

1.1.3 数据产生方式的变革促成大数据时代的来临

数据产生方式的变革，是促成大数据时代来临的重要因素。总体而言，人类社会的数据产生方式大致经历了三个阶段：运营式系统阶段、用户原创内容阶段和感知式系统阶段（见图 1-5）。



图 1-5 数据产生方式的变革

1. 运营式系统阶段

人类社会最早大规模管理和使用数据，是从数据库的诞生开始的。大型零售超市销售系统、银行交易系统、股市交易系统、医院医疗系统、企业客户管理系统等大量运营式系统，都是建立

在数据库基础之上的,数据库中保存了大量结构化的企业关键信息,用来满足企业各种业务需求。在这个阶段,数据的产生方式是被动的,只有当实际的企业业务发生时,才会产生新的记录并存入数据库,比如,对于股市交易系统而言,只有当发生一笔股票交易时,才会有相关记录生成。

2. 用户原创内容阶段

互联网的出现,使得数据传播更加快捷,不需要借助于磁盘、磁带等物理存储介质传播数据,网页的出现进一步加速了大量网络内容的产生,从而使得人类社会数据量开始呈现“井喷式”增长。但是,互联网真正的数据爆发产生于以“用户原创内容”为特征的 Web 2.0 时代。Web 1.0 时代主要以门户网站为代表,强调内容的组织与提供,大量上网用户本身并不参与内容的产生。而 Web 2.0 技术以 Wiki、博客、微博、微信等自服务模式为主,强调自服务,大量上网用户本身就是内容的生成者,尤其是随着移动互联网和智能手机终端的普及,人们更是可以随时随地使用手机发微博、传照片,数据量开始急剧增加。

3. 感知式系统阶段

物联网的发展最终导致了人类社会数据量的第三次跃升。物联网中包含大量传感器,如温度传感器、湿度传感器、压力传感器、位移传感器、光电传感器等,此外,视频监控摄像头也是物联网的重要组成部分。物联网中的这些设备,每时每刻都在自动产生大量数据,与 Web 2.0 时代的人工数据产生方式相比,物联网中的自动数据产生方式,将在短时间内生成更密集、更大量的数据,使得人类社会迅速步入“大数据时代”。

1.1.4 大数据的发展历程

从大数据的发展历程来看,总体上可以划分为 3 个重要阶段:萌芽期、成熟期和大规模应用期(见表 1-2)。

表 1-2

大数据发展的 3 个阶段

阶段	时间	内 容
第一阶段:萌芽期	20 世纪 90 年代至 21 世纪初	随着数据挖掘理论和数据库技术的逐步成熟,一批商业智能工具和知识管理技术开始被应用,如数据仓库、专家系统、知识管理系统等
第二阶段:成熟期	21 世纪前十年	Web 2.0 应用迅猛发展,非结构化数据大量产生,传统处理方法难以应对,带动了大数据技术的快速突破,大数据解决方案逐渐走向成熟,形成了并行计算与分布式系统两大核心技术,谷歌的 GFS 和 MapReduce 等大数据技术受到追捧,Hadoop 平台开始大行其道
第三阶段:大规模应用期	2010 年以后	大数据应用渗透各行各业,数据驱动决策,信息社会智能化程度大幅提高

这里简要回顾一下大数据的发展历程。

- 1980 年,著名未来学家阿尔文·托夫勒在《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”。
- 1997 年 10 月,迈克尔·考克斯和大卫·埃尔斯沃思在第八届美国电气和电子工程师协会(IEEE)关于可视化的会议论文集中,发表了《为外存模型可视化而应用控制程序请求页面调度》的文章,这是在美国计算机学会的数字图书馆中第一篇使用“大数据”这一术语的文章。
- 1999 年 10 月,在美国电气和电子工程师协会(IEEE)关于可视化的年会上,设置了名为“自动化或者交互:什么更适合大数据?”的专题讨论小组,探讨大数据问题。