



教育部实用型信息技术人才培养系列教材

大数据概论

陈明 编著



科学出版社

教育部实用型信息技术人才培养系列教材

大数据概论

陈 明 编著

科学出版社

北 京

内 容 简 介

本书主要介绍大数据概论，内容包括大数据概述、科学研究第四范式、分布系统设计的CAP理论、NoSQL数据库、复杂网络、MapReduce分布编程模型、大数据存储、大数据分析、大数据挖掘、大数据可视化、大数据安全、大数据机器学习、大数据推荐技术，以及数据科学与数据思维。全书对上述内容概念性地介绍，语言精练、内容全面。

本书可作为高等院校的大数据入门教材，也可以作为学习大数据的科学技术人员的参考书。

图书在版编目(CIP)数据

大数据概论 / 陈明编著. — 北京: 科学出版社, 2015.1

教育部实用型信息技术人才培养系列教材

ISBN 978-7-03-042467-9

I. ①大… II. ①陈… III. ①数据处理—技术培训—教材 IV. ①TP274

中国版本图书馆CIP数据核字(2014)第268307号

责任编辑: 于海云 张丽花 / 责任校对: 郭瑞芝

责任印制: 霍 兵 / 封面设计: 迷底书装

科 学 出 版 社 出 版

北京东黄城根北街16号

邮政编码: 100717

<http://www.sciencep.com>

三河市骏杰印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2015年1月第一版 开本: 787×1092 1/16

2015年1月第一次印刷 印张: 17 1/2

字数: 414 000

定价: 48.00元

(如有印装质量问题, 我社负责调换)

前 言

需求是科学技术发展的原动力。目前,大数据问题的出现与研究已经成为了计算机科学与技术研究的新热点,并显示出日益强大的吸引力,科学大数据的出现催生了数据密集型知识发现的科学研究第四范式。对于信息领域,大数据带来的不仅是机遇,还有一系列的困难和挑战。目前,大数据技术与应用展现出锐不可当的强大生命力,科学界与企业界寄予无比的厚望。大数据成为继 20 世纪末、21 世纪初互联网蓬勃发展以来的又一轮 IT(信息技术)工业革命。数据本身是无意义的,而通过统计、分类、萃取、特征抽取等一系列技术手段,可以从数据中产生信息与知识。所以说,数据是重要的战略资源,隐含着巨大的经济价值,因此已经引起人类社会的广泛关注和高度重视。有效地组织和利用数据,将对经济发展产生巨大的推动作用。大数据是以大样本或全样本代替抽样、以近似代替准确、以联系代替因果,因此,大数据是对传统的 IT 各领域的挑战,研究大数据意义深远。大数据的出现孕育着前所未有的机遇。对大数据的交换、整合和分析,可以发现新的知识,创造新的价值,带来大知识、大科技、大利润和大发展。

本书概括性地介绍大数据的主要内容,是大数据技术入门的参考书。全书分为 14 章,内容包括大数据概述、科学研究第四范式、分布系统设计的 CAP 理论、NoSQL 数据库、复杂网络、MapReduce 分布编程模型、大数据存储、大数据分析、大数据挖掘、大数据可视化、大数据安全、大数据机器学习、大数据推荐技术,以及数据科学与数据思维。为了便于学习,在许多章节首先介绍传统的结构化数据的方法,然后再介绍非结构化数据的方法。

拥有数据的规模、活性及解释运用的能力将成为综合国力的重要组成部分,对数据的占有和控制将成为陆权、海权、空权之外的另一种国家核心资产。联合国在 2012 年发布了大数据政务白皮书,指出大数据对于联合国和各国政府是一个历史性的机遇,通过使用极为丰富的数据资源,对社会经济进行前所未有的实时分析,帮助政府更好地响应社会和经济运行。以数据为王的大数据时代已经到来,对数据的占有和控制也将成为国家间和企业间新的争夺点。大数据技术的专业人才,特别是数据分析专业复合型人才的稀缺将会影响该市场的发展。

在技术层面上,大数据、海量数据与超大规模数据无本质的区别,它们都是指用传统处理方法无法处理的大量数据。通过对大数据的高速有效处理,可以发现数据中蕴藏的规律与规则,进而为各种关键决策提供依据与指导,正确的预测与决策将导致巨大财富的产生。技术与工具密不可分,目前常用的数据处理技术与工具是小数据处理技术与工具,一些海量数据处理方法与工具是一种过渡性的方法与工具,大数据处理技术与工具的研究是一项有理论意义和实际价值的工作。大数据技术是一门崭新的技术,大数据的出现,对 IT 各领域的传统处理方法提出了新的挑战。

2012 年作者接受了教育部 ITAT 教育工程的大数据处理工程师培训《大数据概论》的编写任务,通过对大数据的论文、著作的学习与实践,编写了本书,在这里,向所有大数据的研究者表示谢意。本书在结构上为积木状,各章内容是独立的概念性论述。

由于作者水平有限,书中不足之处在所难免,敬请读者批评指正。

陈明

2014 年 2 月

目 录

前言

第 1 章 大数据概述	1
1.1 问题的提出	2
1.1.1 电子数据迅速增加	2
1.1.2 数据孕育巨大的经济价值	3
1.1.3 数据是国家的核心资产	4
1.2 大数据的产生源泉	4
1.2.1 互联网世界	5
1.2.2 物理世界	6
1.3 大数据的概念	7
1.3.1 数据容量巨大	7
1.3.2 数据类型多	8
1.3.3 价值密度低	8
1.3.4 数据传播迅速	9
1.3.5 真实性	9
1.4 大数据的特性	9
1.4.1 价值	9
1.4.2 非结构性	9
1.4.3 不完备性	10
1.4.4 时效性	10
1.4.5 安全性	10
1.4.6 可靠性	10
1.5 大数据技术概述	10
1.5.1 大数据技术的主要内容	11
1.5.2 大数据的处理过程	12
1.5.3 大数据技术的特征	13
1.5.4 大数据的关键问题与关键技术	14
1.6 大数据应用趋势	16
1.6.1 大数据细分市场	17
1.6.2 大数据推动企业发展	17
1.6.3 大数据分析的新方法出现	17
1.6.4 大数据与云计算高度融合	17
1.6.5 大数据一体设备陆续出现	17
1.6.6 大数据安全日益重视	18
1.7 大数据应用	18
1.7.1 判断大数据应用成功的指标	18

1.7.2	大数据技术的应用	19
1.8	大数据的展望	22
1.8.1	资源与投入	23
1.8.2	工程技术	23
1.8.3	复杂网络分析	23
1.8.4	涉及众多领域	23
1.8.5	构建大数据生态环境	23
	本章小结	23
第2章	科学研究四种范式	24
2.1	科学研究第一范式	25
2.1.1	科学实验特点	25
2.1.2	科学实验步骤	26
2.1.3	科学实验分类	26
2.1.4	科学实验构成	27
2.1.5	科学实验程序	28
2.1.6	科学研究第一范式使用原则	29
2.2	科学研究第二范式	30
2.2.1	科学理论的特征	30
2.2.2	科学理论的结构	31
2.2.3	科学理论的价值	31
2.2.4	建立科学理论体系的一般方法	32
2.3	科学研究第三范式	33
2.3.1	概述	33
2.3.2	离散模型的模拟	34
2.3.3	连续系统的模拟	35
2.3.4	模拟语言	35
2.4	科学研究第四范式	36
2.4.1	数据密集型计算	36
2.4.2	格雷法则	38
2.4.3	第四范式的核心内容	40
	本章小结	41
第3章	分布系统设计的CAP理论	42
3.1	分布式系统的伸缩性	42
3.1.1	可伸缩性的概念	43
3.1.2	影响横向扩展的主要因素	44
3.2	横向扩展方案	47
3.2.1	可伸缩共享数据库	47
3.2.2	对等复制的横向扩展方案	48
3.2.3	链接服务器和分布式查询	49
3.2.4	分布式分区视图	50

3.2.5 数据依赖型路由的横向扩展	50
3.3 CAP 理论	51
3.3.1 分布系统设计的核心系统需求	51
3.3.2 CAP 定理	53
3.4 BASE 模型	56
3.4.1 三个核心需求分析	56
3.4.2 ACID、BASE 与 CAP 的关系	57
3.4.3 CAP 与延迟	58
3.4.4 CAP 理论的进一步研究	58
3.5 Web 分布式系统设计	60
3.5.1 系统核心需求	60
3.5.2 系统服务	61
3.5.3 冗余	62
3.5.4 分区	62
本章小结	64
第 4 章 NoSQL 数据库	65
4.1 NoSQL 概述	65
4.1.1 非结构化问题	65
4.1.2 NoSQL 的产生	66
4.2 NoSQL 的特点与问题	67
4.2.1 NoSQL 的特点	67
4.2.2 NoSQL 问题	68
4.3 NoSQL 的主要存储方式	69
4.3.1 键值存储方式	69
4.3.2 文档存储方式	72
4.3.3 列存储方式	73
4.3.4 图形存储方式	76
4.3.5 各种典型的存储方式所对应的 NoSQL 数据库	77
4.4 常用的 NoSQL 数据库	78
4.4.1 Cassandra	78
4.4.2 Lucene/Solr	78
4.4.3 Riak	79
4.4.4 CouchDB	79
4.4.5 Neo4J	79
4.4.6 Oracle 的 NoSQL	79
4.4.7 Hadoop 的 HBase	79
4.4.8 Bigtable/ Accumulo/ Hypertable	80
4.4.9 DynamoDB	80
4.4.10 MongoDB	80
本章小结	82

第 5 章 复杂网络	83
5.1 概述	83
5.1.1 复杂网络概念	84
5.1.2 社会网络概述	84
5.1.3 社会计算	86
5.2 社会网络应用	87
5.2.1 知识获取分析	87
5.2.2 知识类型与传递	88
5.2.3 知识创新	89
5.3 社会网络分析	89
5.3.1 社会网络分析概述	89
5.3.2 社会网络分析的原理	90
5.3.3 社会网络分析的特征	90
5.3.4 社会网络分析的常用方法	90
5.4 社会网络中的隐私保护	91
5.4.1 用户隐私面临的威胁	92
5.4.2 身份隐私攻击与保护	93
5.4.3 面向用户关系的攻击及保护	93
5.4.4 万维网用户隐私保护规范	93
5.5 社会感知计算	94
5.5.1 社会感知计算概念	94
5.5.2 社会感知计算的主要内容	94
5.6 人类通信方式	95
5.6.1 通信方式的演化	95
5.6.2 六度分隔理论	96
5.6.3 150 法则	98
5.6.4 唯象理论与唯象方法	98
5.7 社交网站	99
5.7.1 社交网站概述	99
5.7.2 社交网站的作用	99
5.7.3 移动社交网络	100
5.7.4 Web 2.0 网站	101
5.7.5 Web 2.0 开发平台与必备要素	104
5.7.6 Web 3.0 网站	105
本章小结	105
第 6 章 MapReduce 分布编程模型	106
6.1 函数式编程范式	106
6.1.1 函数型语言	106
6.1.2 函数式编程	107
6.2 映射函数与化简函数	108

6.2.1 映射与映射函数	108
6.2.2 化简与化简函数	109
6.3 MapReduce 计算	110
6.4 基于 Hadoop 平台的分布式计算	111
6.4.1 Hadoop 概述	111
6.4.2 分布式系统与 Hadoop	112
6.4.3 SQL 数据库和 Hadoop	113
6.4.4 基于 Hadoop 的分布式计算	114
本章小结	119
第 7 章 大数据存储	120
7.1 大数据存储概述	120
7.1.1 大数据存储模型	121
7.1.2 大数据存储问题	121
7.2 存储方式	122
7.2.1 存储介质	122
7.2.2 直接连接存储	122
7.2.3 网络连接存储	124
7.2.4 存储域网络存储	125
7.2.5 IP-SAN	126
7.2.6 三种存储方式的比较	126
7.3 大数据的存储	127
7.3.1 数据容量问题	127
7.3.2 大图数据	127
7.3.3 分布式存储的架构	129
7.3.4 数据存储管理	130
7.4 数据云存储	132
7.4.1 云存储的意义与问题	133
7.4.2 技术措施	133
7.5 数据存储的可靠性	135
7.5.1 磁盘与磁盘阵列的可靠性	136
7.5.2 文件系统的可靠性	138
本章小结	138
第 8 章 大数据分析	139
8.1 数据分析概述	140
8.1.1 数据分析的概念	140
8.1.2 数据分析的目的与意义	140
8.1.3 数据分析的基本方法	141
8.1.4 数据分析的类型	146
8.1.5 数据分析的步骤	147
8.2 大数据分析基础	147

8.2.1	可视化分析	148
8.2.2	数据挖掘	148
8.2.3	大数据预测分析	148
8.2.4	语义引擎	148
8.2.5	数据质量和数据管理	148
8.2.6	大数据的离线与在线分析	148
8.3	大数据预测分析	149
8.3.1	大数据预测分析关键因素	150
8.3.2	大数据预测分析演进方向	150
8.3.3	大数据预测分析相关问题	151
8.3.4	舆情监测与分析	152
8.3.5	舆情报告图表制作	153
8.4	大数据分析应用	154
8.4.1	为客户提供服务	154
8.4.2	优化业务流程	154
8.4.3	改善生活	155
8.4.4	提高医疗条件	155
8.4.5	提高体育成绩	155
8.4.6	优化机器和设备性能	155
8.4.7	改善安全和执法	155
8.4.8	改进和优化城市	155
8.4.9	金融交易	156
8.4.10	电信业务	156
8.4.11	销售	156
8.5	大数据分析平台与工具	156
8.5.1	大数据分析基础平台	156
8.5.2	大数据分析的工具	158
	本章小结	160
第9章	大数据挖掘	161
9.1	数据挖掘概述	162
9.1.1	数据挖掘的定义	162
9.1.2	数据挖掘的分类	163
9.1.3	数据挖掘的技术	163
9.2	数据挖掘对象与过程	164
9.2.1	数据挖掘对象	164
9.2.2	数据挖掘过程	164
9.2.3	数据挖掘过程工作量	165
9.3	数据挖掘的常用方法	166
9.3.1	神经网络方法	166
9.3.2	遗传算法	166

9.3.3	决策树方法	166
9.3.4	粗集方法	166
9.3.5	覆盖正例排斥反例方法	167
9.3.6	统计分析方法	167
9.3.7	模糊集方法	167
9.4	数据挖掘的几个问题	167
9.4.1	数据挖掘与数据分析的区别	167
9.4.2	数据挖掘与数据仓库	167
9.4.3	数据挖掘和 OLAP 的比较	168
9.4.4	数据挖掘与人工智能	169
9.4.5	软硬件发展对数据挖掘的影响	169
9.4.6	数据挖掘和统计分析的区别	169
9.4.7	Web 挖掘和数据挖掘的区别	170
9.5	关联规则	170
9.5.1	关联规则定义	170
9.5.2	关联规则分类	171
9.5.3	关联规则的挖掘过程	171
9.5.4	关联规则应用	172
9.6	数据挖掘的经典算法	172
9.6.1	Apriori 算法集	173
9.6.2	划分算法	173
9.6.3	FP-树频集算法	173
9.7	大数据挖掘技术	173
9.7.1	大数据挖掘关键技术	174
9.7.2	大数据挖掘策略	176
9.8	大数据挖掘应用	176
9.8.1	市场营销	177
9.8.2	销售矿泉水	178
9.8.3	物流	178
9.8.4	CRM	179
	本章小结	181
第 10 章	大数据可视化	182
10.1	数据可视化技术概述	182
10.1.1	数据可视化技术的产生史	183
10.1.2	数据可视化技术适用范围	183
10.1.3	信息展现方式	183
10.1.4	数据、信息及知识	185
10.1.5	交互式处理	185
10.2	科学可视化	185
10.2.1	科学可视化的概念与过程	186

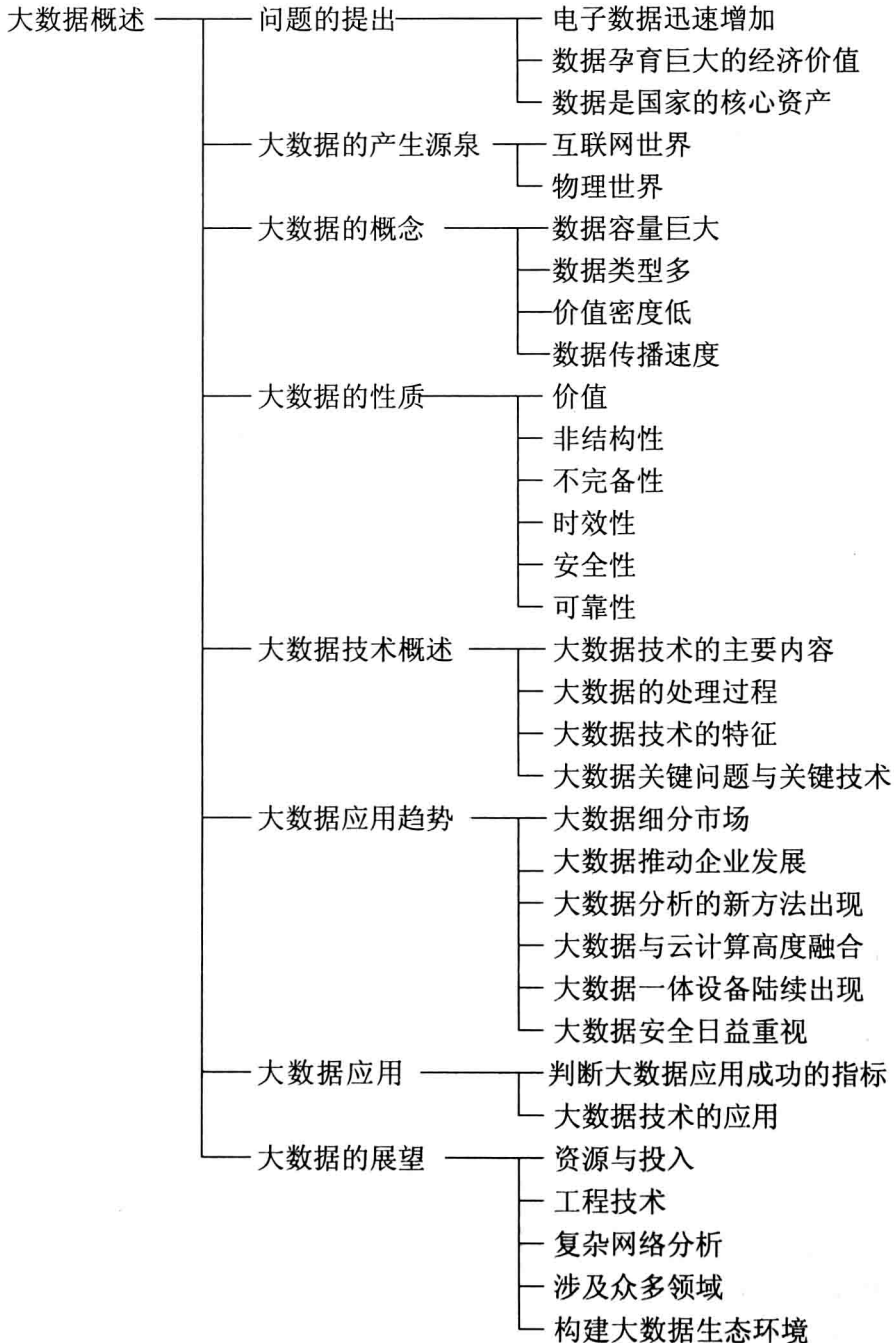
10.2.2	大数据科学可视化的技术	186
10.3	信息可视化	189
10.3.1	信息可视化概念	189
10.3.2	知识发现	190
10.3.3	知识发现工具	191
10.3.4	信息可视化技术的应用	191
10.4	数据可视化应用	192
10.4.1	数据可视化的概念	192
10.4.2	数据可视化技术的特点	192
10.4.3	数据可视化技术的相关概念	193
10.4.4	数据可视化技术的应用	193
10.5	大数据可视分析	194
10.5.1	大数据可视分析的概念	195
10.5.2	大数据可视分析的方法	195
	本章小结	198
第 11 章	大数据安全	199
11.1	数据安全概述	199
11.1.1	数据安全的定义	200
11.1.2	数据处理与存储的安全	200
11.1.3	数据安全的基本特点	200
11.1.4	威胁数据安全的主要因素	201
11.1.5	安全制度与防护技术	202
11.1.6	典型应用	203
11.2	安全措施实现	208
11.2.1	网络分段	208
11.2.2	数据链路层的物理分段	208
11.2.3	VLAN 的划分	208
11.3	电子商务安全	209
11.3.1	计算机网络安全的内容	209
11.3.2	计算机商务交易安全的内容	210
11.4	大数据安全	211
11.4.1	大数据的不安全因素	211
11.4.2	大数据安全的关键问题	212
11.4.3	大数据安全措施	213
11.5	云安全	214
11.5.1	云计算中用户的安全需求	214
11.5.2	威胁模型	215
11.5.3	云安全的支撑技术问题	215
11.5.4	用户数据隐私保护	216
11.5.5	云计算执行环境的可信性	216

11.5.6 资源共享问题	217
本章小结	217
第 12 章 大数据机器学习	218
12.1 机器学习概述	219
12.1.1 机器学习的产生与发展	219
12.1.2 机器学习的概念	219
12.1.3 机器学习理论及研究	220
12.1.4 机器学习系统的结构	221
12.2 机器学习类型	222
12.2.1 基于学习策略的学习分类	222
12.2.2 基于应用领域的学习分类	223
12.2.3 基于综合因素的学习分类	223
12.3 知识表示形式	224
12.4 大数据机器学习	225
12.4.1 大数据机器学习的特点	226
12.4.2 大数据机器学习的评测指标	227
12.5 大数据机器学习的应用	228
12.5.1 基于大数据的空气质量推断	228
12.5.2 人与建筑的关系分析	228
12.5.3 针对全球问题的预测模型	229
12.5.4 全球地表覆盖制图可视化与数据分析	229
本章小结	229
第 13 章 大数据推荐技术	230
13.1 概述	231
13.1.1 推荐系统的产生与发展	231
13.1.2 推荐系统的概念	231
13.2 推荐系统架构	232
13.2.1 用户特征提取模块	232
13.2.2 相关物品检索模块	232
13.2.3 推荐结果排序模块	232
13.3 推荐系统类型	232
13.3.1 基于用户行为数据推荐	232
13.3.2 基于用户标签数据推荐	233
13.3.3 基于上下文信息推荐	233
13.3.4 基于社交网络数据推荐	233
13.4 推荐系统的评判标准	234
13.5 推荐算法	235
13.5.1 基于人口统计学的推荐算法	235
13.5.2 基于内容的推荐算法	235
13.5.3 协同过滤推荐算法	236

13.5.4	混合推荐算法	238
13.6	推荐模式与系统	238
13.6.1	推荐模式	238
13.6.2	下一代推荐系统	239
13.7	大数据推荐技术	240
13.7.1	数据稀疏性	241
13.7.2	大数据推荐系统冷启动	241
13.7.3	多样性与精确性的两难命题	241
13.7.4	增量计算	242
13.7.5	推荐系统的鲁棒性	242
13.7.6	推荐系统效果评估	242
13.7.7	用户行为模式的挖掘和利用	242
13.7.8	用户界面与用户体验	243
13.7.9	多维数据的交叉利用	243
13.7.10	社会推荐	244
13.8	大数据人才推荐系统	244
	本章小结	245
第 14 章	数据科学与数据思维	246
14.1	数据科学概述	246
14.1.1	数据科学定义与信息化过程	246
14.1.2	数据科学研究内容	247
14.1.3	数据科学的研究过程与体系框架	248
14.2	大数据研究方式	249
14.2.1	大数据分析的是全面的数据	249
14.2.2	重视数据的复杂性与弱化精确性	251
14.2.3	关注数据的相关性而非因果关系	251
14.3	数据专家	252
14.3.1	数据科学家	252
14.3.2	数据工程师	254
14.4	数据思维	254
14.4.1	思维的概念与特征	254
14.4.2	思维的形成	256
14.4.3	计算思维	258
14.4.4	网络思维	260
14.4.5	系统思维	263
14.4.6	大数据思维	264
	本章小结	264
	参考文献	265

第 1 章 大数据概述

本章主要内容



需求是科学技术发展的原动力。目前，大数据问题的出现与研究已经成为了计算机科学与技术研究的新热点，并显示出日益强大的吸引力，科学大数据的出现催生了数据密集型知识发现的科学研究第四范式的出现。对于信息领域，大数据带来的不仅是机遇，还有一系列的困难和挑战。目前，大数据技术与应用展现出锐不可当的强大生命力，科学界与企业界寄予无比的厚望。大数据成为继 20 世纪末、21 世纪初互联网蓬勃发展以来的新一轮 IT 工业革命。

1.1 问题的提出

在全世界范围内，以电子方式存储的数据(又简称为电子数据)总量空前巨大。在 2011 年电子数据总量已达到 1.8ZB(1ZB=1024PB)，较 2010 年同期提高超过 1ZB，统计结果表明，每经过 2 年就可以增加 1 倍，预计到 2020 年可达到 35ZB，如图 1-1 所示。面对数据增长的速度迅猛提升，数据量的飞速增加，对大量电子数据的高效存储、高效传输与快速的处理是必须面对的研究问题。

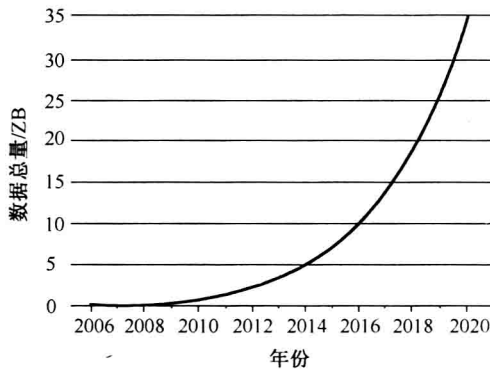


图 1-1 全球数据创建及复制的数据总量预测

1.1.1 电子数据迅速增加

物联网、云计算、移动互联网、车联网、手机、平板电脑、个人计算机(PC)、气候信息、公开的信息，如杂志、报纸和文章、交易记录、网络日志、病历、军事监控、视频和图像、档案及大型电子商务，以及遍布地球各个角落的各种各样的传感器是数据来源或者承载的方式不断更新与发展、大型科学研究设备产生的数据，以及社交媒体的快速发展，构成了大数据持续产生的生态环境。尤其是近年来，随着互联网技术的发展，来自人们的日常生活，特别是来自互联网服务而产生的大量数据迅猛增加。据不完全统计，互联网当前包含 93 亿多个页面，80%~85%的数据是存储在数据库的文本中。互联网一天产生的全部内容可以刻满 1.68 亿张 DVD，发出的邮件有 2940 亿封之多，发出的社区帖子达 200 万个(相当于《时代》杂志 770 年的文字量)，卖出的手机为 37.8 万台，高于全球每天出生的婴儿数量 37.1 万……从数据统计角度来看，电子数据量迅速增加。预计中国数据技术和市场未来 5 年的复合增长率将达 51.4%，其中增长率最高的是存储市场，将达 60.8%，服务器市场的增长率则是 38.3%，远远高于其他产品相关的市场。

1.1.2 数据孕育巨大的经济价值

数据本身是无意义的，而通过统计、分类、萃取、特征抽取等一系列技术手段，可以从数据中产生信息与知识。数据是重要的战略资源，隐含巨大的经济价值，因此已经引起科学界和企业界的高度重视。有效地组织和利用数据，将对经济发展产生巨大的推动作用。大数据出现孕育着前所未有的机遇。对大数据的交换、整合和分析，可以发现新的知识，创造新的价值。

越来越多的企业等机构意识到数据正在成为最重要的资产，数据分析能力正在成为核心竞争力。经过了由 PC 成功转向了软件和服务，而这次将远离服务与咨询，更多地专注于因数据分析而带来的全新业务增长点。数据将成为各行业中决定胜负的根本因素，最终数据将成为人类至关重要的自然资源。各著名的大型公司已经致力于开发自己的大数据处理和存储系统，目前已经到了数据化运营的黄金时期，如何整合这些数据成为未来的关键任务。

在互联网、电信、金融等行业，几乎已经到了数据就是业务本身的地步。物联网、社交网络等新的互联网技术在为人们带来便利的同时，也产生了大量的数据。如何有效地存储和查询这些数据，如何通过数据挖掘，从数据中获得有用的信息，为用户提供好的用户体验，增强企业的竞争力，是一个挑战。研究表明，数字领域存在着 1.8 万亿 GB 的数据，企业数据正在以 55% 的速度逐年增长。目前，两天就能创造出自人类文明诞生以来到 2003 年所产生的数据总量。大数据已经成为重要的时代特征，充分利用大数据可帮助全球个人定位服务提供商增加 1000 亿美元的收入，帮助欧洲公共部门的管理每年提升 2500 亿美元产值，帮助美国医疗保健行业每年提升 3000 亿美元产值，并可帮助美国零售业获得 60% 以上的净利润增长率。由此可见，充分使用大数据和挖掘大数据商业价值将为行业企业带来强大经济效益与竞争力。

大数据既是对信息技术发展的高度抽象和概括，同时也体现了信息技术服务于数据蕴藏的巨大价值。大数据给数据的采集、存储、维护、共享带来了具有研究意义的现象和挑战，但更多的意义是可以处理、分析并使用大量数据，通过这些数据的处理、整合和分析，可以发现新知识、创造新价值，带来大知识、大科学和大发展，逐渐走向创新社会化的新信息时代。

大数据全生命周期可以划分为“数据产生—数据采集—数据传输—数据存储—数据处理—数据分析—数据发布、展示和应用—产生新数据”等阶段。已经形成了大数据的“生产与集聚层—组织与管理层—分析与发现层—应用与服务层”的产业链，而 IT 基础设施为这各环节提供基础支撑。

据统计，2012 年市场规模达到 4.5 亿元，2016 年估计可达到百亿规模，如图 1-2 所示。

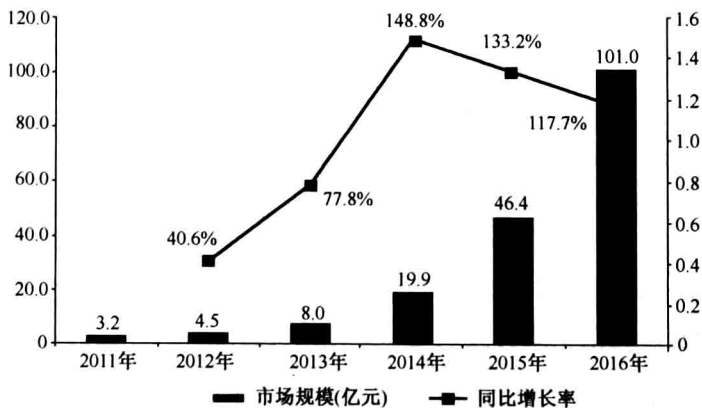


图 1-2 中国大数据应用市场规模与增长