



陈明◎编著

# 大数据 基础与应用

- ◆ 大数据技术与教育扛鼎之作
- ◆ 全面系统深入讲述大数据



北京师范大学出版集团  
BEIJING NORMAL UNIVERSITY PUBLISHING GROUP  
北京师范大学出版社



陈明◎编著



# 大数据 基础与应用



北京师范大学出版集团  
BEIJING NORMAL UNIVERSITY PUBLISHING GROUP  
北京师范大学出版社

---

图书在版编目 (CIP) 数据

大数据基础与应用/陈明编著. —北京: 北京师范大学出版社, 2016  
ISBN 978-7-303-20018-4

I. ①大… II. ①陈… III. ①数据处理 IV. ①TP274

中国版本图书馆 CIP 数据核字 (2016) 第 005840 号

---

营销中心电话 010-62978190 62979006  
北师大出版社科技与经管分社网 <http://jsws.bnupg.com>  
电子邮箱 [kjgg@bnupg.com](mailto:kjgg@bnupg.com)

---

出版发行: 北京师范大学出版社 [www.bnupg.com](http://www.bnupg.com)  
北京新街口外大街 19 号  
邮政编码: 100875

印刷: 北京京师印务有限公司  
经销: 全国新华书店  
开本: 184 mm×260 mm 1/16  
印张: 21.25  
字数: 504 千字  
版次: 2016 年 3 月第 1 版  
印次: 2016 年 3 月第 1 次印刷  
定价: 42.00 元

---

策划编辑: 李丹	责任编辑: 李丹
美术编辑: 刘超	装帧设计: 刘超
责任校对: 马军令	责任印制: 曲利华

**版权所有 侵权必究**

反盗版、反侵权举报电话: 010-62978190

北京读者服务部电话: 010-62979006-8021

外埠邮购电话: 010-62978190

本书如有印装质量问题, 请与印制管理部联系调换。

印制管理部电话: 010-62979006-8006

# 前 言

需求是科学技术发展的原动力。大数据问题的出现与研究已经成为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第四科学研究范式的出现。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。大数据技术与应用展现出锐不可当的强大生命力，科学界与企业界对其寄予无比的厚望。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。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，可以从数据中提取信息与知识。所以说，数据是重要的战略资源，隐含巨大的经济价值，已经引起科技界和企业界的高度重视。有效地组织和使用数据，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。

本书概括性地介绍了大数据的主要内容，是大数据技术入门的参考书。全书分为 17 章，其中第 1 章是对大数据的简单概述，第 2 章介绍大数据研究的方法论，第 3 章、第 8 章、第 9 章和第 14 章介绍大数据的生态环境，第 4 章、第 5 章、第 6 章、第 7 章、第 10 章、第 11 章、第 12 章、第 13 章、第 15 章和第 16 章介绍大数据技术及应用方法，第 17 章简单介绍数据科学的内容。

大数据的规模、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，帮助政府更好地调整社会和经济运行。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新的争夺点。大数据技术的专业人才，特别是数据分析复合型人才的稀缺将会影响市场的发展。

在技术上，大数据、海量数据与超大规模数据都是指用传统处理方法无法处理的大量数据。通过对大数据进行高速有效的处理，可以发现数据中蕴藏的规律，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适用于小数据处理，一些海量数据处理方法与工具是过渡性的，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。大数据的出现，对 IT 各领域的传统处理方法提出了新的冲击与挑战。大数据技术是一门实践性较强的技术，需要重视工具与应用方法的选择与研究。

高校是培养人才的重要基地，需要分析和定位大数据带来的影响。大数据推动学科发展和学科建设、利用大数据技术整合现有的教学资源为已经到来的“数据革命”培养行业紧缺人才都是教育工作面临和需要解决的问题。

本书在结构上呈积木状，对各章内容进行独立的概念性论述。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

2015年10月

# 目 录

第 1 章 走进大数据时代.....	1	2.1.1 科学实验特点与步骤 .....	23
1.1 应对大数据 .....	2	2.1.2 科学实验构成与分类 .....	24
1.1.1 电子数据迅速增加 .....	2	2.1.3 科学实验程序 .....	25
1.1.2 数据中蕴含的价值 .....	2	2.1.4 科学实验使用原则 .....	25
1.1.3 数据是国家的核心资产 .....	3	2.2 科学研究第二范式 .....	26
1.2 大数据的生态环境 .....	3	2.2.1 科学理论的特征与价值.....	27
1.2.1 互联网世界 .....	4	2.2.2 科学理论的结构与体系建立 方法 .....	27
1.2.2 物理世界 .....	5	2.3 科学研究第三范式 .....	28
1.3 大数据的概念 .....	6	2.3.1 系统模拟发展过程 .....	29
1.3.1 数据容量 .....	6	2.3.2 系统模拟基本方法 .....	29
1.3.2 数据类型 .....	8	2.3.3 系统模拟语言 .....	30
1.3.3 价值密度 .....	8	2.4 科学研究第四范式 .....	30
1.3.4 速度 .....	8	2.4.1 数据密集型计算 .....	31
1.3.5 真实性 .....	8	2.4.2 格雷法则 .....	32
1.4 大数据的性质 .....	8	2.4.3 核心内容 .....	35
1.4.1 非结构性 .....	8	小结 .....	36
1.4.2 不完备性 .....	9	第 3 章 分布系统设计的 CAP 理论.....	37
1.4.3 时效性 .....	9	3.1 分布式系统的伸缩性 .....	38
1.4.4 安全性 .....	10	3.1.1 可伸缩性的概念 .....	38
1.4.5 可靠性 .....	10	3.1.2 影响横向扩展的主要因素.....	39
1.5 大数据技术概述 .....	10	3.2 横向扩展方案 .....	42
1.5.1 大数据处理的全过程 .....	11	3.2.1 可伸缩共享数据库 .....	42
1.5.2 大数据技术的特征 .....	13	3.2.2 对等复制的横向扩展方案.....	43
1.5.3 大数据的关键问题与 关键技术 .....	14	3.2.3 链接服务器和分布式查询.....	44
1.6 大数据应用 .....	16	3.2.4 分布式分区视图 .....	45
1.6.1 大数据应用趋势 .....	16	3.2.5 数据依赖型路由的横向扩展 ...	46
1.6.2 大数据应用评价与应用实例 ...	17	3.3 CAP 理论 .....	47
小结 .....	21	3.3.1 分布系统设计的核心 系统需求 .....	47
第 2 章 科学研究范式 .....	22	3.3.2 CAP 定理.....	49
2.1 科学研究第一范式 .....	23		

3.4	BASE 模型	53	4.6.3	150 法则	78
3.4.1	三个核心需求分析	53	4.6.4	唯象理论与唯象方法	79
3.4.2	ACID、BASE 与 CAP 的关系	54	4.7	社交网站	80
3.4.3	CAP 与延迟	55	4.7.1	社交网站作用	80
3.4.4	CAP 理论的进一步研究	55	4.7.2	Web 网站	80
3.5	Web 分布式系统设计	57	小结		81
3.5.1	系统核心需求	57	第 5 章	MapReduce 分布编程模型	82
3.5.2	系统服务	58	5.1	函数式编程范式	82
3.5.3	冗余	59	5.1.1	函数型语言与函数式编程	83
3.5.4	分区	60	5.1.2	函数式编程优点	83
小结		61	5.1.3	函数式编程的特征	84
第 4 章	大数据网络空间	62	5.2	映射函数与化简函数	84
4.1	复杂网络空间概述	63	5.2.1	映射与映射函数	84
4.1.1	复杂网络概念与特征	63	5.2.2	化简与化简函数	85
4.1.2	复杂网络的特性	64	5.3	MapReduce 的体系结构	86
4.2	社会网络	65	5.3.1	MapReduce 计算描述	86
4.2.1	社会网络结构	65	5.3.2	MapReduce 适用情况	88
4.2.2	社会网络理论	66	5.4	基于 Hadoop 平台的分布式计算	88
4.2.3	社会计算	67	5.4.1	Hadoop 发展历程	88
4.2.4	社会网络应用	68	5.4.2	分布式系统与 Hadoop	90
4.3	社会网络分析	69	5.4.3	SQL 数据库和 Hadoop	90
4.3.1	社会网络分析概述	70	5.4.4	基于 Hadoop 的分布式计算	92
4.3.2	社会网络分析的原理	70	小结		98
4.3.3	社会网络分析的特征	71	第 6 章	大数据流式计算	99
4.3.4	社会网络分析的方法	71	6.1	流式数据的概念与特征	99
4.4	社会网络中的隐私保护	72	6.1.1	流式数据的概念	99
4.4.1	用户隐私类型	72	6.1.2	流式数据的特征	100
4.4.2	身份隐私攻击与保护	73	6.2	大数据的计算模式	101
4.4.3	用户关系的攻击及保护	73	6.2.1	大数据批量计算模型	101
4.4.4	万维网用户隐私保护	74	6.2.2	大数据流式计算模型	101
4.5	社会感知计算	74	6.2.3	大数据流式计算与批量计算的比较	103
4.5.1	社会感知计算概念	74	6.3	流式大数据处理工具	104
4.5.2	社会感知计算的内容	75	6.3.1	Storm 系统	104
4.6	人类通信方式	76	6.3.2	S4 系统	107
4.6.1	通信方式的演化	76	6.3.3	Data Freeway and Puma 系统	110
4.6.2	六度分隔理论	77			

6.4 大数据流式计算的应用 .....	111	8.2.1 数据容量问题 .....	139
6.4.1 金融银行业的应用 .....	112	8.2.2 大图数据 .....	140
6.4.2 互联网领域的应用 .....	112	8.2.3 分布式存储的架构 .....	142
6.4.3 物联网领域的应用 .....	113	8.2.4 数据存储管理 .....	143
6.4.4 三种典型应用场景的对比 ...	113	8.3 数据云存储 .....	145
小结 .....	114	8.3.1 云存储的意义与问题 .....	145
<b>第 7 章 大数据搜索技术</b> .....	<b>115</b>	8.3.2 技术措施 .....	146
7.1 搜索引擎概述 .....	116	8.4 数据存储的可靠性 .....	148
7.1.1 搜索引擎的发展过程 .....	116	8.4.1 磁盘与磁盘阵列的可靠性....	148
7.1.2 搜索引擎的定义 .....	117	8.4.2 文件系统的可靠性 .....	151
7.1.3 搜索引擎的组成 .....	117	小结 .....	151
7.1.4 搜索引擎的分类 .....	117	<b>第 9 章 NoSQL 数据库</b> .....	<b>152</b>
7.1.5 搜索引擎的工作过程 .....	120	9.1 NoSQL 概述 .....	153
7.1.6 搜索引擎的评价指标 .....	121	9.1.1 非结构化问题 .....	153
7.2 语义搜索引擎 .....	121	9.1.2 NoSQL 的产生 .....	153
7.2.1 语义与语义搜索引擎的		9.2 NoSQL 的特点与问题 .....	155
概念 .....	121	9.2.1 NoSQL 的特点 .....	155
7.2.2 图谱 .....	122	9.2.2 NoSQL 面对的问题 .....	156
7.2.3 搜索就是回答 .....	123	9.3 NoSQL 的主要存储方式 .....	157
7.2.4 语义搜索引擎的组成 .....	123	9.3.1 键值存储方式 .....	157
7.2.5 基于本体的语义搜索引擎 ...	123	9.3.2 文档存储方式 .....	158
7.3 网站数据对搜索引擎的影响 .....	126	9.3.3 列存储方式 .....	160
7.3.1 垂直网站与垂直搜索 .....	126	9.3.4 图形存储方式 .....	166
7.3.2 私有化的 Web 化数据 .....	127	9.3.5 存储类型对应的 NoSQL	
7.3.3 没有 Web 化的数据 .....	127	数据库 .....	167
7.3.4 大数据流动 .....	128	9.4 常用的 NoSQL 数据库 .....	168
7.4 搜索引擎优化 .....	128	9.4.1 Cassandra .....	168
7.4.1 搜索引擎优化的产生 .....	129	9.4.2 Lucene .....	168
7.4.2 网页级别 .....	129	9.4.3 Riak .....	169
7.4.3 搜索引擎优化的方法 .....	129	9.4.4 CouchDB .....	169
小结 .....	131	9.4.5 Neo4j .....	169
<b>第 8 章 大数据存储</b> .....	<b>132</b>	9.4.6 Oracle 的 NoSQL .....	169
8.1 大数据存储概述 .....	132	9.4.7 Hadoop 的 HBase .....	170
8.1.1 大数据存储模型 .....	133	9.4.8 Bigtable .....	170
8.1.2 大数据存储问题 .....	133	9.4.9 DynamoDB .....	170
8.1.3 存储方式 .....	135	9.4.10 MongoDB .....	170
8.2 大数据的存储技术 .....	139	小结 .....	173



第 10 章 大数据预处理技术 .....	174	11.1.1 数据分析的概念 .....	197
10.1 数据抽取概述 .....	175	11.1.2 数据分析的目的与意义 .....	197
10.1.1 数据抽取的概念与包装器 .....	175	11.1.3 数据分析方法的分类 .....	198
10.1.2 抽取数据的方法 .....	176	11.1.4 数据分析的类型 .....	198
10.2 Web 数据抽取 .....	177	11.1.5 数据分析步骤 .....	199
10.2.1 Web 数据抽取问题的 提出 .....	177	11.2 基本数据分析方法 .....	199
10.2.2 Web 数据抽取的目的与 方法 .....	177	11.2.1 统计方法 .....	199
10.2.3 Web 数据抽取过程 .....	178	11.2.2 指标对比分析法 .....	200
10.3 数据质量与数据清洗 .....	179	11.2.3 分组分析法 .....	201
10.3.1 数据质量 .....	179	11.2.4 综合评价分析法 .....	201
10.3.2 数据清洗的目的 .....	180	11.2.5 指数分析法 .....	201
10.3.3 数据清洗算法衡量标准 .....	180	11.2.6 平衡分析法 .....	201
10.3.4 数据清洗的主要研究成果 .....	181	11.2.7 平滑和滤波 .....	202
10.3.5 数据清洗技术面临的 问题 .....	181	11.2.8 基线与峰值 .....	202
10.4 不符合要求的数据 .....	182	11.3 高级数据分析方法 .....	202
10.4.1 不完整的数据 .....	182	11.3.1 时间数列及动态分析法 .....	202
10.4.2 异常的数据 .....	182	11.3.2 相关分析 .....	203
10.4.3 重复的数据 .....	183	11.3.3 回归分析 .....	203
10.5 数据清洗技术的实现 .....	183	11.3.4 判别分析 .....	204
10.5.1 数据清洗的方法与技巧 .....	184	11.3.5 对应分析 .....	204
10.5.2 数据清洗的实现方式 .....	186	11.3.6 预测分析 .....	204
10.5.3 数据清洗的步骤 .....	187	11.3.7 主成分分析 .....	204
10.5.4 数据清洗的评价标准 .....	188	11.3.8 多维尺度分析 .....	205
10.5.5 常用的数据清洗算法 .....	188	11.3.9 因子分析 .....	205
10.5.6 大数据清洗工具 .....	189	11.3.10 方差分析 .....	205
10.6 数据集成 .....	189	11.4 复合技术分析 .....	205
10.6.1 数据集成技术概述 .....	190	11.4.1 快速傅里叶变换 .....	206
10.6.2 数据集成系统的构建 .....	192	11.4.2 分类 .....	206
10.7 数据转换与约简 .....	193	11.4.3 聚类分析 .....	206
10.7.1 数据转换 .....	193	11.5 大数据分析基础 .....	207
10.7.2 数据约简 .....	194	11.5.1 可视化分析 .....	207
小结 .....	195	11.5.2 数据挖掘 .....	207
第 11 章 大数据分析 .....	196	11.5.3 预测性分析 .....	208
11.1 数据分析概述 .....	197	11.5.4 语义引擎 .....	208
		11.5.5 数据质量和数据管理 .....	208
		11.5.6 大数据的离线与在线分析 .....	208
		11.6 大数据预测分析 .....	209
		11.6.1 预测学简介 .....	209

11.6.2 预测原理 .....	211	小结 .....	248
11.6.3 预测的步骤 .....	212	<b>第 13 章 大数据可视化</b> .....	<b>249</b>
11.6.4 预测技术分类 .....	213	13.1 可视化技术概述 .....	249
11.6.5 预测模型及分类 .....	215	13.1.1 可视化技术的产生与发展 ...	250
11.6.6 大数据预测分析要素 .....	215	13.1.2 科学可视化 .....	253
11.6.7 大数据预测分析的演化 .....	216	13.1.3 信息可视化 .....	254
11.6.8 大数据预测分析相关问题 ...	216	13.1.4 数据可视化 .....	257
11.6.9 舆情监测与分析 .....	217	13.2 大数据科学可视化 .....	260
<b>11.7 大数据分析应用</b> .....	<b>220</b>	13.2.1 高可伸缩性的分布式并行 可视化算法 .....	260
11.7.1 为客户提供服务 .....	220	13.2.2 并行图像合成算法 .....	261
11.7.2 优化业务流程 .....	220	13.2.3 并行颗粒跟踪算法 .....	261
11.7.3 改善生活 .....	220	13.2.4 重要信息的提取与显示 .....	262
11.7.4 提高体育成绩 .....	221	13.2.5 原位可视化 .....	263
11.7.5 优化机器和设备性能 .....	221	13.3 大数据可视化分析 .....	264
11.7.6 改善安全和执法 .....	221	13.3.1 大数据可视化分析概念 .....	264
11.7.7 金融交易 .....	221	13.3.2 大数据可视化分析方法 .....	264
11.7.8 电信业务 .....	221	小结 .....	267
11.7.9 销售 .....	222	<b>第 14 章 大数据安全</b> .....	<b>268</b>
<b>11.8 大数据分析平台与工具</b> .....	<b>222</b>	14.1 大数据安全概述 .....	268
11.8.1 大数据分析平台 .....	222	14.1.1 数据安全的定义 .....	269
11.8.2 大数据分析的工具 .....	223	14.1.2 数据处理与存储的安全 .....	269
小结 .....	226	14.1.3 数据安全的基本特点 .....	269
<b>第 12 章 大数据挖掘</b> .....	<b>227</b>	14.1.4 威胁数据安全的主要因素 ...	270
12.1 数据挖掘概述 .....	227	14.1.5 安全制度与防护技术 .....	271
12.1.1 数据挖掘的几个概念 .....	228	14.1.6 应用 .....	273
12.1.2 数据挖掘对象与过程 .....	229	14.2 大数据安全的内容 .....	278
12.1.3 数据挖掘的常用方法 .....	232	14.2.1 大数据的不安全因素 .....	278
12.1.4 数据挖掘的几个问题 .....	234	14.2.2 大数据安全的关键问题 .....	279
12.1.5 数据挖掘的经典算法 .....	237	14.2.3 大数据安全措施 .....	280
12.2 大数据挖掘技术 .....	238	14.3 云安全 .....	281
12.2.1 大数据挖掘关键技术 .....	239	14.3.1 云计算中用户的安全需求 ...	281
12.2.2 大数据挖掘策略 .....	240	14.3.2 威胁模型 .....	282
12.3 大数据挖掘应用 .....	243	14.3.3 云安全的支撑技术 .....	282
12.3.1 市场营销 .....	243	14.3.4 用户数据隐私保护 .....	283
12.3.2 销售 .....	244	14.3.5 云计算执行环境的 可信性 .....	283
12.3.3 物流 .....	245		
12.3.4 CRM .....	246		

14.3.6 资源共享问题 .....	284	16.3.3 多样性与精确性的 两难命题 .....	310
小结 .....	284	16.3.4 大数据处理与增量 计算问题 .....	311
<b>第 15 章 大数据机器学习</b> .....	<b>285</b>	16.3.5 推荐系统的脆弱性问题.....	311
15.1 机器学习概述 .....	285	16.3.6 推荐系统效果评估 .....	311
15.1.1 机器学习的产生与发展 .....	285	16.3.7 用户行为模式的挖掘和 利用 .....	312
15.1.2 机器学习类型 .....	288	16.3.8 用户界面与用户体验.....	312
15.1.3 知识表示形式 .....	291	16.3.9 多维数据的交叉利用.....	313
15.2 大数据机器学习的特点与 评测指标 .....	292	16.3.10 社会推荐 .....	313
15.2.1 大数据机器学习的特点 .....	293	<b>16.4 大数据人才推荐系统</b> .....	<b>314</b>
15.2.2 大数据机器学习的 评测指标 .....	294	小结 .....	315
15.3 大数据机器学习的应用 .....	295	<b>第 17 章 数据科学与数据思维</b> .....	<b>316</b>
15.3.1 基于大数据的空气 质量推断 .....	295	17.1 数据科学概述 .....	316
15.3.2 人与建筑的关系分析 .....	296	17.1.1 数据科学的定义与信息化 过程 .....	316
15.3.3 针对全球问题的预测模型 .....	296	17.1.2 数据科学的研究内容.....	317
15.3.4 全球地表覆盖制图可视化与 数据分析 .....	296	17.1.3 数据科学的研究过程与 体系框架 .....	318
小结 .....	297	17.2 大数据研究方式 .....	319
<b>第 16 章 大数据推荐技术</b> .....	<b>298</b>	17.2.1 大数据分析的是全面的 数据 .....	320
16.1 推荐技术概述 .....	298	17.2.2 重视数据的复杂性与弱化 精确性 .....	321
16.1.1 推荐系统的产生与发展 .....	299	17.2.3 关注数据的相关性而 非因果关系 .....	321
16.1.2 推荐系统的概念 .....	299	17.3 数据专家 .....	322
16.1.3 推荐系统架构 .....	300	17.3.1 数据科学家 .....	322
16.1.4 推荐系统类型 .....	300	17.3.2 数据处理工程师 .....	325
16.1.5 推荐系统的评判标准 .....	302	17.3.3 大数据思维 .....	325
16.2 推荐算法与推荐模式 .....	303	小结 .....	326
16.2.1 推荐算法 .....	303	<b>参考文献</b> .....	<b>327</b>
16.2.2 推荐模式 .....	306		
16.2.3 下一代推荐系统 .....	308		
16.3 大数据推荐技术的挑战 .....	310		
16.3.1 数据稀疏性问题 .....	310		
16.3.2 大数据冷启动 .....	310		

# 第 1 章 走进大数据时代



## 本章主要内容



需求是科学技术发展的原动力，大数据中蕴含着巨大而重要的价值，大数据的研究已经成为计算机科学与技术研究的新热点，并显示出日益强大的生命力和吸引力。科学大数据催生了数据密集型知识发现的第四科学研究范式的出现。大数据的出现，不仅带来了机遇，也带来了困难和挑战。无论是科学界，还是企业界，都对大数据所带来的巨大冲击寄予厚望，大数据是继 20 世纪末、21 世纪初互联网蓬勃发展以来的新一轮新的 IT 工业革命。

## 1.1 应对大数据

在全球范围内，以电子方式存储的数据（简称为电子数据）总量空前巨大。2011年，电子数据总量达到1.8ZB，比2010年同期提高了1ZB，统计结果表明，每经过两年就可以增加一倍，预计到2020年可达到35ZB，其预测曲线如图1-1所示。面对数据增长速度飞快地提升，数据量狂增，对大量电子数据的高效存储、高效传输与快速处理成为必须解决的问题。

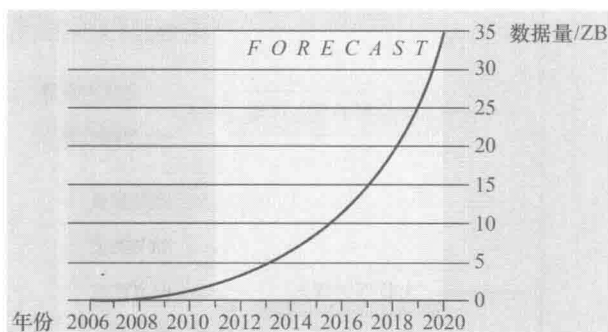


图 1-1 全球数据总量的预测示意图

### 1.1.1 电子数据迅速增加

云计算、物联网、移动互联网、手机、平板电脑和PC机中的数据，各种公开的数据、大型电子商务数据、遍布全球的各种传感器数据、工业蓬勃发展产生的工业数据、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基本生态环境。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+X服务的数据迅猛增加。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，发出的电子邮件有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。从统计角度来看，由于电子数据量迅速增加，预计中国数据技术和服务市场未来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 数据中蕴含的价值

数据本身是无意义的，但是，有效地组织和使用数据，即对大数据进行交换、整合和分析，可以发现新的知识、创造新的价值、推动科技大发展，凸显了数据是重要的战略资源。

据统计，2012年中国市场规模达到4.5亿元，2016年估计可达到上百亿元的规模，如图1-2所示。

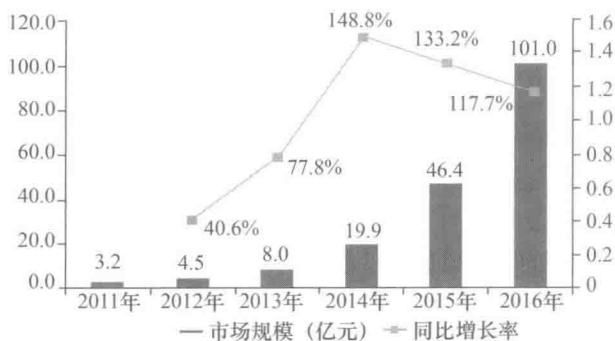


图 1-2 中国大数据应用市场规模与增长的预测

### 1.1.3 数据是国家的核心资产

一个国家拥有数据的规模及分析获取价值的能力将成为综合国力的重要组成部分，存储与控制的数据将成为国家陆权、海权和空权之外的另一个核心资产。

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来说是一个历史性的机遇，通过使用丰富的数据资源，对社会经济进行前所未有的实时分析，能够帮助政府更好地响应社会和经济运行。对数据的占有和控制也将成为国家之间和企业之间的新的争夺点。社会需要大量大数据技术的专业人才，特别是数据分析复合型人才，对这一专业人才的培养，将是一项十分重要的任务。

## 1.2 大数据的生态环境

大数据是人类活动的产物，它来自人们改造客观世界的过程中，是生产与生活在网络空间的投影。信息爆炸是对信息快速发展的一种逼真的描述，形容信息发展的速度如同爆炸一般席卷整个空间。在 20 世纪 40~50 年代，信息爆炸主要指的是科学文献的快速增长。而经过 50 年的发展，到 20 世纪 90 年代，由于计算机和通信技术的广泛应用，信息爆炸主要指的是所有社会信息快速增长，包括正式交流过程和非正式交流过程所产生的电子式的和非电子式的信息。而到 21 世纪的今天，信息爆炸是由于数据洪流的产生和发展所造成的。在技术方面，新型的硬件与数据中心、分布式计算、云计算、高性能计算、大容量数据存储与处理技术、社会化网络、移动终端设备、多样化的数据采集方式使大数据的产生和记录成为可能。在用户方面，日益人性化的用户界面、信息行为模式等都容易作为数据量化而被记录，用户既可以成为数据的制造者，又可以成为数据的使用者。可以看出，随着云计算、物联网计算和移动计算的发展，世界上所产生的新数据，包括位置、状态、思考、过程和行动等数据都能够汇入数据洪流，互联网的广泛应用，尤其是“互联网+”的出现，促进了数据洪流的发展。

归纳起来，大数据主要来自互联网世界与物理世界。

## 1.2.1 互联网世界

大数据是计算机和互联网相结合的产物，计算机实现了数据的数字化，互联网实现了数据的网络化，两者结合起来之后，赋予了大数据强大的生命力。随着互联网如同空气、水、电一样无处不在地渗透人们的工作和生活，以及移动互联网、物联网、可穿戴联网设备的普及，新的数据正在以指数级加速产生，目前世界上 90% 的数据是互联网出现之后迅速产生的。来自互联网的网络大数据是指“人、机、物”三元世界在网络空间（Cyberspace）中交互、融合所产生并可在互联网上获得的大数据，网络大数据的规模和复杂度的增长超出了硬件能力增长的摩尔定律。

大数据来自人类社会，尤其是互联网的发展为数据的存储、传输与应用创造了基础与环境。依据基于唯象假设的六度分隔理论而建立的社交网络服务（Social Network Service, SNS），以认识朋友的朋友为基础，扩展自己的人脉。基于 Web 2.0 交互网站建立的社交网络，用户既是网站信息的使用者，也是网站信息的制作者。社交网站记录人们之间的交互，搜索引擎记录人们的搜索行为和搜索结果，电子商务网站记录人们购买商品的喜好，微博网站记录人们所产生的即时的想法和意见，图片视频分享网站记录人们的视觉观察，百科全书网站记录人们对抽象概念的认识，幻灯片分享网站记录人们的各种正式和非正式的演讲发言，机构知识库和期刊记录学术研究成果等。归纳起来，来自互联网的数据可以划分为下述几种类型。

### 1. 视频图像

视频图像是大数据的主要来源之一，电影、电视节目可以产生大量的视频图像，各种室内外的视频摄像头昼夜不停地产生巨量的视频图像。视频图像以每秒几十帧的速度连续记录运动着的物体，一个小时的标准清晰视频经过压缩后，所需的存储空间为 GB 数量级，对于高清晰度视频所需的存储空间就更大了。

### 2. 图片与照片

图片与照片也是大数据的主要来源之一，截至 2011 年 9 月，用户向脸书（Facebook，美国的一个社会网络服务网站）上传了 1400 亿张以上的照片。如果拍摄者为了保存拍摄时的原始文件，平均每张照片大小为 1MB，则这些照片的总数据量约为  $1.4 \times 10^{12} \times 1\text{MB} = 140\text{PB}$ ，如果单台服务器磁盘容量为 10TB，则存储这些照片需要 14000 台服务器，而且这些上传的照片仅仅是人们拍摄到的照片的很少一部分。此外，许多遥感系统 24 小时不停地拍摄并产生大量照片。

### 3. 音频

DVD 光盘采用了双声道 16 位采样，采样频率为 44.1kHz，可达到多媒体欣赏水平。如果某音乐剧的时间为 5.5min，计算其占用的存储容量为：

$$\begin{aligned}\text{存储容量} &= (\text{采样频率} \times \text{采样位数} \times \text{声道数} \times \text{时间}) / 8 \\ &= (44.1 \times 1000 \times 16 \times 2 \times 5.5 \times 60) / 8 \\ &\approx 55.5\text{MB}\end{aligned}$$

## 4. 日志

网络设备、系统及服务程序等，在运作时都会产生 log 的事件记录。每一行日志都记载着日期、时间、使用者及动作等相关操作的描述。Windows 网络操作系统设有各种各样的日志文件，如应用程序日志、安全日志、系统日志、Scheduler 服务日志、FTP 日志、WWW 日志、DNS 服务器日志等，这些根据系统开启的服务的不同而有所不同。用户在系统上进行一些操作时，这些日志文件通常记录了用户操作的一些相关内容，这些内容对系统安全工作人员相当有用。例如，有人对系统进行了 IPC 探测，系统就会在安全日志里迅速地记下探测者探测时所用的 IP、时间、用户名等，用 FTP 探测后，就会在 FTP 日志中记下 IP、时间、探测所用的用户名等。

网站日志记录了用户对网站的访问，电信日志记录了用户拨打和接听电话的信息，假设有 5 亿用户，每个用户每天呼入呼出 10 次，每条日志占用 400B，并且需要保存 5 年，则数据总量为  $5 \times 10 \times 365 \times 400 \times 5 \text{ Byte} \approx 3.65\text{PB}$ 。

## 5. 网页

网页是构成网站的基本元素，是承载各种网站应用的平台。通俗地说，网站就是由网页组成的，如果只有域名和虚拟主机而没有制作任何网页，客户仍旧无法访问网站。网页要通过网页浏览器来阅读。文字与图片是构成一个网页的两个最基本的元素。可以简单地理解为：文字就是网页的内容，图片就是网页的美观描述。除此之外，网页的元素还包括动画、音乐、程序等。

网页分为静态网页和动态网页。静态网页的内容是预先确定的，并存储在 Web 服务器或者本地计算机、服务器之上，动态网页取决于用户提供的参数，并根据存储在数据库中的网站上的数据而创建。通俗地讲，静态页是照片，每个人看都是一样的，而动态页则是镜子，不同的人（不同的参数）看都不相同。

网页中的主要元素有感知信息、互动媒体和内部信息等。感知信息主要包括文本、图像、动画、声音、视频、表格、导航栏、交互式表单等。互动媒体主要包括交互式文本、互动插图、按钮、超链接等。内部信息主要包括注释，通过超链接链接到某文件、元数据与语义的元信息，字符集信息，文件类型描述，样式信息和脚本等。

网页内容丰富，数据量巨大，每个网页有 25KB 数据，则一万亿个网页的数据总量约为 25PB。

## 1.2.2 物理世界

来自物理世界的大数据又被称为科学大数据，科学大数据主要来自大型国际实验：跨实验室、单一实验室或个人观察实验所得到的科学实验数据或传感数据。最早提出大数据概念的学科是天文学和基因学，这两个学科从诞生之日起就依赖于基于海量数据的分析方法。由于科学实验是科技人员设计的，数据采集和数据处理也是事先设计的，所以不管是检索还是模式识别，都有科学规律可循。例如希格斯粒子，又称为“上帝粒子”的寻找，



采用了大型强子对撞机实验。这是一个典型的基于大数据的科学实验，至少要在 1 万亿个事例中才可能找出一个希格斯粒子。从这一实验可以看出，科学实验的大数据处理是整个实验的一个预定步骤，这是一个有规律的设计，发现有价值的信息可在预料之中。大型强子对撞机每秒生成的数据量约为 1PB。建设中的下一代巨型射电望远镜阵每天生成的数据量大约在 1EB。波音发动机上的传感器每小时产生 20TB 左右的数据量。

随着科研人员获取数据方法与手段的变化，科研活动产生的数据量激增，科学研究已成为数据密集型活动。科研数据因其数据规模大、类型复杂多样、分析处理方法复杂等特征，已成为大数据的一个典型代表。大数据所带来的新的科学研究方法反映了未来科学的行为研究方式，数据密集型科学研究将成为科学研究的普遍范式。

利用互联网可以将所有的科学大数据与文献联系在一起，创建一个文献与数据能够交互操作的系统，即在线科学数据系统，如图 1-3 所示。

对于在线科学数据，由于各个领域互相交叉，不可避免地需要使用其他领域的的数据。利用互联网能够将所有文献与数据集成在一起，可以实现从文献计算到数据的整合。这样可以提高科技信息的检索速度，进而大幅度地提高生产力。也就是说，在线阅读某人的论文时，可以查看他们的原始数据，甚至可以重新分析，也可以在查看某些数据时查看所有关于这一数据的文献。



图 1-3 在线科学数据系统示意图

## 1.3 大数据的概念

大数据是指数据规模大，尤其是指因为数据形式多样性、非结构化特征明显，导致数据存储、处理和挖掘异常困难的那类数据集。大数据增长快速，类型繁多，如文本、图像、视频等。大数据处理包含数千万个文档、数百万张照片或者工程设计图的数据集，如何快速访问数据成为核心挑战。无法用常规的软件工具捕捉与处理。

通常将大数据归纳为 5 个“V”：Volume（数据容量）、Variety（数据类型）、Viscosity（价值密度）、Velocity（速度）、Veracity（真实性），如图 1-4 所示。

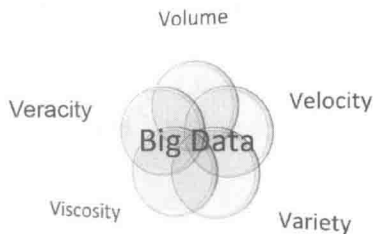


图 1-4 大数据的 5 个“V”

### 1.3.1 数据容量

Volume 代表大数据的数据量巨大，存储容量单位的定义如表 1-1 所示。