大数据丛书系列之十

总主编◎曾　羽　龙奋杰

大数据安全与隐私

DASHUJU
ANQUAN YU YINSI

主　编◎姚剑波　杨朝琼

电子科技大学出版社

大数据丛书系列之十

总主编◎曾　羽　龙奋杰

大数据安全与隐私

DASHUJU
ANQUAN YU YINSI

主　编◎姚剑波　杨朝琼

电子科技大学出版社

图书在版编目(CIP)数据

大数据安全与隐私 / 姚剑波, 杨朝琼主编. -- 成都: 电子科技大学出版社, 2017.7
ISBN 978-7-5647-4819-7

Ⅰ. ①大… Ⅱ. ①姚… ②杨… Ⅲ. ①数据处理 - 安全技术 - 研究 Ⅳ. ①TP274

中国版本图书馆CIP数据核字（2017）第177084号

大数据安全与隐私

姚剑波　杨朝琼　主编

策划编辑　杨仪玮　李燕芩
责任编辑　李燕芩

出版发行　电子科技大学出版社
　　成都市一环路东一段159号电子信息产业大厦　邮编 610051
主　　页　www.uestcp.com.cn
服务电话　028-83203399
邮购电话　028-83201495

印　　刷　成都市火炬印务有限公司
成品尺寸　165mm × 240mm
印　　张　20
字　　数　370千字
版　　次　2017年7月第一版
印　　次　2017年7月第一次印刷
书　　号　ISBN 978-7-5647-4819-7
定　　价　65.00元

前　言

21世纪是信息科学与技术极速发展的时代，信息成为一种重要的战略资源，信息的获取、存储、处理及安全保障能力成为一个国家的综合国力的重要组成部分。数据开放要走之路，让不同领域的数据真正流动、融合起来，才能释放大数据的价值。但是，原始的数据形式通常包含个人的敏感信息，发布这些数据会侵犯个人的隐私。现今各行各业收集数据的能力大大提升，随之为基于知识和信息的决策提供了广泛的机会。在利益或者规章的驱动下，不同的群体之间都有数据交换和共享的需求。然而，数据的收集、发布和分析（挖掘）要面对的一个重要问题是隐私泄露和信息安全，即在前网络时代，隐私在法律、政府、组织、个人的多重保护下，是相对安全的，而网络的出现，令现实社会中个人隐私权的有关问题延伸到了网络空间，由于网络社会的开放特征，使得个人隐私面临着严重的威胁。

数据发布的现行做法主要依赖于相关的政策法规和去标识符的简单处理。这种简易处理方法可能导致大量数据缺乏足够多的保护。为此，相关专家和科技工作者正在积极开展相关研究，开发数据发布隐私保护（PPDP）的有效方法，寻找保护数据隐私的同时提高数据的实用性的平衡。数据隐私保护已经成为一个新兴的、非常热门的研究领域，并且针对不同的数据发布场景提出了许多方法。

本书主要介绍新兴的数据隐私保护研究领域的产生背景、基础知识（当前隐私问题、隐私法律、隐私保护模型、数据匿名化、统计数据库、隐私保护数据分析、社交网络隐私等）、隐私保护技术、实现方法、商业应用、最新研究成果和进展。研究数据实际发布过程中遇到的挑战，并对今后的研究方向提出建议。

本文引用了参考资料中的大量内容，在此，感谢参考资料的所有作者，他们的工作和研究成果给了我极大帮助和启发，是他们刻苦钻研和辛勤工作的成果成就了大数据安全与隐私这片天地。

未尽事宜，敬请谅解！

目 录

第1章 大数据时代 ……1
1.1 大数据汹涌来临 ……2
1.1.1 什么是大数据 ……3
1.1.2 大数据无处不在 ……3
1.1.3 大数据的特点 ……4
1.1.4 大数据的应用 ……5
1.2 大数据推手 ……7
1.2.1 摩尔定律：铸造数据滋生的利器 ……7
1.2.2 吉尔德定律：大带宽支撑大数据 ……9
1.2.3 麦特卡夫定律：大数据价值是用户创造的 ……11
1.3 大数据安全隐私 ……13
1.3.1 数据产生环节 ……14
1.3.2 数据获取及传输环节 ……15
1.3.3 数据存储环节 ……16
1.3.4 数据分析及应用环节 ……17
第2章 基础设施安全 ……19
2.1 大数据与云计算 ……19
2.1.1 云计算的特征 ……19
2.1.2 大数据和云计算的关系 ……19
2.2 基础设施安全：网络层面 ……22
2.2.1 确保数据的保密性和完整性 ……22
2.2.2 确保适当的访问控制 ……23
2.2.3 确保面向互联网资源的可用性 ……24
2.2.4 用域替换已建立的网络区域及层面模型 ……25
2.2.5 网络层减灾 ……26

2.3 基础设施安全：主机层面 ……27
2.3.1 SaaS和PaaS的主机安全 ……28
2.3.2 IaaS的主机安全 ……28
2.3.3 虚拟化软件安全 ……29
2.3.4 虚拟服务器的安全 ……30
2.4 基础设施安全：应用层面 ……32
2.4.1 应用级安全威胁 ……32
2.4.2 终端用户的安全 ……33
2.4.3 云计算的网络应用程序安全 ……34

第3章 数据安全与存储 ……40

3.1 数据安全 ……40
3.2 降低数据安全的风险 ……43
3.3 提供商数据及其安全 ……44
3.4 存储 ……44
3.4.1 保密性 ……44
3.4.2 完整性 ……46
3.4.3 可用性 ……47

第4章 密码学基础 ……49

4.1 引言 ……49
4.2 最基本的概念 ……50
4.2.1 一般的保密通信系统 ……50
4.2.2 明文 ……50
4.2.3 密文 ……52
4.2.4 密本 ……52
4.2.5 密表 ……53
4.2.6 密钥 ……53
4.2.7 密码体制 ……54
4.2.8 解密和密码分析 ……55
4.2.9 密码算法 ……57
4.2.10 密码体制（算法）的设计准则 ……58
4.2.11 密码学 ……59
4.2.12 密码学格言 ……60
4.2.13 香农的保密通信理论 ……61

4.3 富于想象的古典密码术 ……64
4.3.1 英语的统计特性 ……64
4.3.2 单表代替体制 ……65
4.3.3 多表代替体制 ……67
4.4 近代密码术 ……68
4.5 现代密码学 ……70
4.5.1 算法复杂性理论基础知识介绍 ……70
4.5.2 计算上保密的密码体制 ……71
4.5.3 密码体制分类 ……72
4.5.4 分组密码体制 ……72
4.5.5 序列密码体制 ……83
4.5.6 秘密密钥密码体制 ……90
4.5.7 公开密钥密码体制 ……90
4.5.8 散列函数 ……92
4.6 至关重要的密钥管理 ……95
4.6.1 密钥管理的基本要求 ……95
4.6.2 密钥的意义 ……96
4.6.3 密钥种类及作用 ……96
4.6.4 密钥的分层结构 ……97
4.6.5 密钥管理 ……97
4.6.6 密钥自动分发技术 ……98
4.6.7 密钥托管技术 ……101
4.6.8 密钥管理基础设施（KMI/PKI） ……102

第5章 网络安全保密技术 ……103

5.1 引言 ……103
5.2 通信保密技术类型 ……103
5.2.1 话音保密通信 ……103
5.2.2 数据保密通信 ……112
5.2.3 图像保密通信 ……113
5.3 网络化面临的严重威胁 ……118
5.3.1 日益紧迫的信息战威胁 ……118
5.3.2 被动攻击 ……120
5.3.3 主动攻击 ……121
5.3.4 内部人员攻击 ……122

5.4 网络安全保密的基本要求 ……122
5.4.1 保密性要求 ……123
5.4.2 完整性要求 ……123
5.4.3 可用性要求 ……123
5.4.4 可认证性要求 ……123
5.4.5 不可抵赖性要求 ……124
5.4.6 实时性要求 ……124
5.4.7 可控性要求 ……124
5.5 网络安全保密的基本模式 ……125
5.5.1 信息网络的标准分层模型 ……125
5.5.2 网络各层中的安全服务 ……125
5.5.3 链路层保密原理 ……127
5.5.4 网络层保密原理 ……127
5.5.5 传输层保密原理 ……128
5.5.6 应用层保密原理 ……128
5.5.7 多网络互连的安全保密模式 ……129
5.6 密钥管理技术和网管安全 ……132
5.6.1 基于对称密钥密码体制的密钥管理技术 ……132
3.6.2 基于公开密钥密码体制的密钥管理技术 ……133
5.6.3 系统化的密钥管理实施要点 ……136
5.7 网管系统的安全 ……137
5.7.1 网管系统的结构特点 ……137
5.7.2 网管系统的安全保密需求 ……138
5.7.3 网管系统的安全保密实施要点 ……140

第6章 信息系统安全 ……143

6.1 引言 ……143
6.2 升级换代的信息安全体系 ……143
6.2.1 通信保密时代 ……144
6.2.2 信息系统安全时代 ……144
6.2.3 信息保障时代 ……144
6.3 以纵深防御为基础的信息保障 ……145
6.3.1 纵深防御 ……145
6.3.2 保卫网络和基础设施 ……148
6.3.3 保卫边界和外部连接 ……149

6.3.4 保卫计算环境 ……150
6.3.5 支持性安全基础设施 ……150
6.4 信息系统安全保障体系 ……151
6.4.1 信息系统安全保障体系框架 ……151
6.4.2 信息系统安全保障体系的属性 ……151
6.4.3 信息系统安全保障体系的发展 ……153
6.5 信息系统安全工程 ……153
6.5.1 信息系统安全工程过程 ……155
6.5.2 发掘信息保护需求 ……156
6.5.3 确定系统安全要求 ……159
6.5.4 设计系统安全体系结构 ……160
6.5.5 开发详细安全设计 ……160
6.5.6 实现系统安全 ……162
6.5.7 评估信息保护的有效性 ……162
6.6 信息系统安全风险分析与评估 ……162
6.6.1 风险分析和评估的若干问题 ……163
6.6.2 风险分析的方法 ……164
6.6.3 风险分析与评估的要素及程序 ……166
6.6.4 风险评估与等级保护 ……168
6.7 信息系统安全等级保护 ……169
6.7.1 我国信息系统等级保护的政策 ……169
6.7.2 等级保护的进展 ……169
6.7.3 等级保护的问题 ……170
6.8 信息系统安全服务 ……171
6.8.1 信息系统安全服务概述 ……171
6.8.2 信息安全服务的发展 ……172
6.9 信息安全技术 ……173
6.9.1 安全数据隔离与交换 ……173
6.9.2 理想的安全基础软硬件 ……175
6.9.3 定期的漏洞扫描和风险评估 ……175
6.9.4 快捷的灾难恢复 ……176
6.9.5 智能的动态安全管理 ……176

第7章 信任、匿名和隐私 ……179

7.1 信任 ……179

7.1.1 什么是信任模型 ……179
7.1.2 信任模型是怎样工作的 ……180
7.1.3 信任可能会在哪里出错 ……186
7.1.4 信任为什么难以定义 ……188
7.1.5 我们学到了什么 ……189
7.2 PKI系统 ……189
7.2.1 密码学的起源 ……190
7.2.2 PKI系统概述 ……191
7.2.3 PKI系统的组成部分 ……192
7.2.4 PKI系统的过程 ……194
7.2.5 PKI系统的现状和未来前景 ……195
7.3 电子通信中的隐私 ……197
7.3.1 针对第三方的保密性 ……197
7.3.2 保护隐私免受通信参与方的侵犯 ……204
7.3.3 对私人电子空间的侵犯 ……208
7.3.4 在其他要求与隐私之间进行权衡 ……212
7.3.5 隐私的结构 ……213
7.4 数据内容安全 ……214
7.4.1 数字内容安全：需求和挑战 ……215
7.4.2 内容保护技术 ……217

第8章 大数据隐私 ……222

8.1 隐私保护的理论基础 ……222
8.1.1 隐私的定义 ……222
8.1.2 隐私分类 ……222
8.1.3 隐私的度量 ……223
8.1.4 隐私主要威胁 ……224
8.1.5 隐私泄露的原因和表现形式 ……226
8.1.6 隐私数据安全的基本要求 ……231
8.1.7 信息度量和隐私保护原则 ……232
8.1.8 社交网络的隐私保护 ……233
8.2 隐私技术 ……235
8.3 隐私保护机制 ……238
8.4 隐私法律法规 ……238
8.4.1 法律和监管的内涵 ……238

8.4.2 国际的法律法规 ……239
8.4.3 国内的法律法规 ……243
第9章 隐私保护策略 ……246
9.1 法律层面的网络隐私保护 ……247
9.1.1 欧盟关于网络隐私保护的法律法规 ……248
9.1.2 英国对于网络隐私保护的法律法规 ……250
9.1.3 德国对于网络隐私保护的法律法规 ……250
9.1.4 法律层面的网络隐私保护策略分析 ……251
9.2 管理层面的网络隐私保护 ……252
9.2.1 行业自律模式 ……252
9.2.2 管理层面的网络隐私保护策略分析 ……254
9.3 个人层面的网络隐私保护 ……254
9.3.1 提高个人防范意识 ……254
9.3.2 保护个人在线隐私技巧 ……255
9.4 我国网络隐私保护策略及存在的问题 ……258
9.4.1 我国网络隐私保护策略 ……258
9.4.2 我国网络隐私存在的问题 ……263
9.5 我国移动电商的展望 ……264
9.6 大数据与用户信息安全 ……265
第10章 大数据隐私技术 ……266
10.1 匿名技术 ……266
10.1.1 k-匿名 ……271
10.1.2 l-多样性 ……273
10.1.3 t-closeness ……274
10.1.4 （X，Y）-匿名模型 ……275
10.2 差分隐私技术 ……275
10.3 数据清洗 ……279
10.3.1 数据清洗国内外研究现状 ……279
10.3.2 数据清洗的定义与对象 ……280
10.3.3 数据清洗基本原理与框架模型 ……282
10.3.4 数据清洗算法与工具 ……284
10.3.5 数据清洗经验分享 ……287
10.4 随机化技术 ……290

10.4.1 随机扰动 ……290
10.4.2 随机化应答 ……291
10.5 安全多方计算 ……292
10.5.1 基本概念和数学模型 ……292
10.5.2 安全多方计算理论的特点 ……293
10.5.3 安全多方计算理论的应用领域 ……293
10.5.4 安全多方计算理论的基础协议 ……294
10.5.5 安全多方计算理论研究进展 ……295
10.6 访问控制技术 ……296
10.6.1 安全模型 ……297
10.6.2 访问控制策略 ……299
10.6.3 访问控制的实现 ……300
10.6.4 访问控制与授权 ……301
10.6.5 访问控制与审计 ……301
10.7 希波克拉底数据库 ……302
第11章 大数据隐私前沿 ……303
11.1 匿名技术 ……303
11.1.1 匿名隐私保护的主要研究方向 ……303
11.1.2 隐私保护数据发布研究展望 ……303
11.2 差分隐私 ……304
11.3 数据清洗 ……304
11.4 安全多方计算 ……306
参考资料 ……308

第1章 大数据时代

2008年9月4日，《自然》刊登了一个名为“Big Data”的专辑，提出了大数据概念。该专辑对如何研究PB级容量的大数据流，以及目前正在制定的、用以最为充分地利用海量数据的全新策略进行了探讨。

2012年1月，在瑞士达沃斯世界经济论坛上，大数据成为主题之一。论坛发布了《大数据，大影响：国际发展的新机会》的报告，宣称数据就像货币和黄金一样，已经成为一种新的经济资产。

报告指出，随着计算机、GPS设备、手机和医疗设备的普及应用，信息洪流如约而至。每天产生的在线或移动交易、社交媒体流和GPS位置信息高达2.5 EB，全球进入了大数据时代。

大数据在金融服务、健康、教育、农业、医疗等多个领域的应用潜力巨大。在金融服务领域，特别是在移动金融服务领域，应用大数据的前提条件包括完善的监管机制、消费者权益保护、市场竞争力、市场催化剂、终端用户授权与访问等，而移动理财收集的数据能清楚地反映出不同地域的消费构成和习惯等；在教育领域，从移动增值服务中得到的数据可以用于改变对教育的理解，让人们接受更多有针对性的重要信息的传播；在健康领域，通过移动设备收集的信息，有助于医疗卫生机构了解人口健康趋势和阻止疾病爆发，个人电子健康记录可以发送个人的连续治疗情况；在农业领域，农产品的移动支付、采购投入和补贴数据能帮助政府更好地预测粮食生产趋势，通过一些鼓励措施，保证粮食存储，减少糟蹋浪费。

报告还指出：要理解大数据的生态系统，缩小信息鸿沟，洞察大数据应用带来的收益。同时，大数据发展面临诸多挑战，如隐私与安全、数据个性化、数据共享激励、人力资本等。应对这些挑战需要采取新方法，政府是大数据发展的催化剂。大数据的繁荣和发展离不开各国政府的支持和配合。

2012年2月13日，《纽约时报》网站刊载文章称，“大数据时代”已经降临，在这一领域拥有专长的人士正面临许多机会。文章指出，“大数据”正在对每个领域产生影响。举例来说，在商业、经济及其他领域中，决策行为将日益基于数据和分析，而并非基于经验和直觉；而在公共卫生、经济发展和经济预测等领域中，“大数据”的预见能力也已经崭露头角。

随着云计算的突破，“大数据”成了时下最火热的IT行业的词汇和商业焦点。大数据技术的战略意义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，在于提高对数据的“加工能力”，通过“加工”实现数据的“增值”。“这是一场革命，庞大的数据资源使得各个领域开始了量化进程，无论学术界、商界还是政府，所有领域都将开始这种进程。”哈佛大学社会学教授加里·金如是说。随着大数据应用的爆发性增长，它已经衍生出了自己独特的架构，而且也直接推动了存储、网络以及计算技术的发展。

大数据是一个很好的视角和工具。从资本角度来看，什么样的公司有价值，什么样的公司没有价值，从其拥有的数据规模、数据的活性和这家公司能运用、解释数据的能力，就可以看出这家公司的核心竞争力。而这几个能力正是资本关注的热点，也是经济增长的蓝海。移动互联网与社交网络兴起将大数据带入新的征程，互联网营销将在行为分析的基础上向个性化时代过渡。创业公司应用“大数据”告诉商家什么是正确的时间，谁是正确的用户，什么是应该发表的正确内容等，这正好切中了商家的需求。社交网络产生了海量用户以及实时和完整的数据，同时社交网络也记录了用户群体的情绪，通过深入挖掘这些数据来了解用户，然后将这些分析后的数据信息推给需要的商家。大数据时代网民和消费者的界限正在消弭，企业的疆界变得模糊，数据成为核心的资产，并将深刻影响企业的业务模式，甚至重构其文化和组织。因此，大数据对国家治理模式、对企业的决策、组织和业务流程、对个人生活方式都将产生巨大的影响。落花流水春去也，如果不能利用大数据更加贴近消费者、深刻理解需求、高效分析信息并做出预判，所有传统的产品公司都只能沦为新型用户平台级公司的附庸，其衰落不是管理能扭转的。价格的涨落是许多企业的坟墓，而掌握大数据，将拈花微笑，成为获得成功的利器。谷歌、亚马逊、阿里巴巴等拥有海量数据的企业，会跨界搅局，笑傲江湖，成为创造新财富的引领者。数据这一目前尚未列入资产项目表中的价值，可由Facebook的市值看出，它的固定资产微不足道，而其总市值在千亿美元之上。

事实上，全球互联网巨头都已意识到了“大数据”时代，据IDC预测，到2020年全球将总共拥有35ZB的数据量，而麦肯锡则预测未来大数据产品在三大行业的应用就将产生7千亿美元的潜在市场，给信息服务行业也包括媒体行业开拓了一个新的黄金时代。

1.1 大数据汹涌来临

2013年3月1日，工业和信息化部电信研究院在北京召开了2013年ICT深度观察大型报告会暨移动互联网白皮书、中国通信产业十大关键词发布

会。电信研究院在会议上发布了《移动互联网白皮书（2013）》以及“2012年中国通信产业十大关键词”的评选结果，云计算、智能终端、TD-LTE、宽带中国、移动互联网、物联网、网络与信息安全、微博、大数据、微信入选十大关键词。

近年来，大数据的概念受到了学术界、商界甚至政府的热捧，一时间大数据无处不在，这让同时代其他的IT技术相形见绌，无地自容。数据正在迅速膨胀并变大，它决定着企业的未来发展，虽然目前企业可能还没有意识到数据爆炸性增长带来的问题隐患，但是随着时间的推移，人们将越来越重视数据的作用。正如《纽约时报》2012年2月的一篇专栏中所称：“大数据”时代已经降临，在商业、经济及其他领域中，决策将日益基于数据和分析作出，而不是基于经验和直觉。

1.1.1 什么是大数据

“大数据”的概念起源于2008年9月《自然》（Nature）杂志刊登的名为“Big Data”的专题，继而迅速得到了科学、计算机、经济等不同领域专家的响应。由于其成因复杂，对大数据目前没有公认的定义，不同的研究人员从不同领域对大数据进行了定义，下面列出三个不同角度对大数据的定义。

（1）Kusnetzky Dan在What is“Big Data”一文中提出，大数据是指所涉及的数据量规模巨大，无法通过人工在合理时间内截取、管理、处理并整理成为人类所能解读的信息。

（2）维克托·迈尔一舍恩伯格、肯尼斯·库克耶在《大数据时代》一书中把大数据看成一种方法，即不用随机分析法（抽样调查）这样的捷径，而采用所有数据的方法。

（3）“大数据”研究机构Gartner的报告指出，“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的海量、高增长率和多样化的信息资产。

这三种定义中，第一种定义更强调处理能力，第二种定义更强调应用方法，第三种定义更侧重应用价值。

1.1.2 大数据无处不在

现实生活中的数据有多大呢？根据统计，在2006年，个人用户刚刚迈进TB时代，这一年全球共产生了约180EB=180×1018字节的数据；在2011年，达到了1. 8ZB=1.8×1021字节。有市场研究机构预测：到2020年，整个世界的数据总量将会增长44倍。你也许会好奇为何会产生如此庞大的数据，下面我们举几个现实中的大数据例子。

· 社交网络

由于数据来自所有用户的参与，社交网络中的数据量非常大，而且增长非常迅速。例如，新浪微博在晚高峰的时候1秒产生的数据达到100条以上。如果把脸书（Facebook）中的社交网络看成图，在2012年这个图已经达到了超过8亿个顶点，平均每个点的度超过130，每天增加的数据量达到500TB。

· 科学仪器

科学仪器获取了非常巨大的数据，比如说中国遥感国家重点实验室采集的中国大陆地表信息，每个月产生4TB数据。中国天文观测站用LAMOST每年观测到的数据达到3. 65TB，美国NASA中心每年获取超过125TB的数据，英国Sanger中心2002年就已经收集了20TB的数据，并且以每年4倍的速度增长。

· 移动通信

我们每天使用的手机产生了非常巨大的数据，中国移动每年产生的记录超过300TB。

· 传感数据

传感器持续检测环境信息并不断返回结果，产生了巨大的数据。以波音787为例，其每一个飞行来回可产生TB级的数据，美国每个月收集360万次飞行记录；监视所有飞机中的25 000个引擎，每个引擎一天产生588GB的数据。风力发电机装有测量风速、螺距、油温等多种传感器，每隔几毫秒测一次，用于检测叶片、变速箱、变频器等的磨损程度，一个具有500个风机的风场一年会产生2PB的数据。

· 医疗数据

美国著名医疗保健公司InSiteOne平均每年获取2.1PB的放射影像数据，英国每年产生300TB乳腺癌数据，在美国相应的数据量达到2.6PB。哈尔滨医科大学第一附属医院每年通过各类医疗仪器搜集的数据超过30TB。

· 商务数据

生活中的每次刷卡，在超市或者网络中购买的每件商品都产生相应的数据。淘宝网站每天有超过数千万笔交易，单日数据产生量超过50TB。为了有效使用商务大数据，沃尔玛建立了包含PB级数据的数据仓库，Bestbuy建立了包含TB级数据的数据仓库。

1.1.3 大数据的特点

1. 规模性（Volume，耗费大量存储、计算资源）

大数据之“大”，体现在数据的存储和计算均需耗费海量规模的资源

上：美国宇航局收集和处理的气候观察、模拟数据达到32PB；谷歌公司索引的网页总数超过1万亿；FICO的信用卡欺诈检测系统保护全世界超过18亿个活跃信用卡账户。

2. 高速性（Velocity，增长迅速、急需实时处理）

大数据的另一特点在于速度快：大型强子对撞机实验设备中包含了15亿个传感器，平均每秒收集超过4亿条实验数据；每秒超过3万次用户查询提交到谷歌，3万条微博被新浪用户撰写。而在感知、传输、决策、控制这一闭环控制过程中的计算，对数据实时处理有着极高的要求，通过传统数据库查询方式得到的“当前结果”很可能已经没有价值，只有最新的数据才有价值。

3. 多样性（Variety，来源广泛、形式多样）

在大数据背景下，数据在来源和形式上的多样性愈加凸显：除大量以非结构化形式存在的文本数据，也存在位置、图片、音频、视频等信息。除信息形式的多元化，信息的来源也表现出多样性：从网络日志、物联网、移动设备、传感器到基因图谱、医疗影像、天体运行轨迹、交通物流数据等。大数据中的多样性已经超越了数据管理中的异构数据库，其不仅仅是模式或模型的不一样，甚至数据本身的存在形式也完全不同，比如说存在文本、多媒体数据，也存在仪器采集来的完全是数字的数据，以及用户产生的用户行为的数据，这些数据有各种各样的存在形式，这些形式导致处理技术的差异，因此需要新的处理技术。

4. 价值稀疏性（Value，价值总量大、知识密度低）

大数据以其高价值吸引了广泛关注。据全球著名咨询公司麦肯锡报告：“如果能够有效地利用大数据来提高效率和质量，预计美国医疗行业每年通过数据获得的潜在价值可超过3000亿美元，能够使美国医疗卫生支出降低8%。”虽然大数据价值高，但是知识密度非常低。谷歌公司首席经济学家Hal Varian指出“数据是广泛可用的，所缺乏的是从中提取出知识的能力”；IBM副总裁兼CTO Dietrich表示“可以利用Twitter数据获得用户对某个产品的评价，但是往往上百万条记录中只有很小的一部分真正讨论这款产品”。

只有经过高度分析的大数据才可以产生新的价值，需要设计能够适应上述特征的大数据处理算法来处理数据。

1.1.4 大数据的应用

大数据在许多方面有着广泛的应用，甚至说达到了无处不在的程度。

1. 预测

2013年2月19日，微软研究院的David Rothschil博士带领的大数据分析