Broadview
www.broadview.com.cn

大数据

技术与应用实践指南

Big Data | Technology and Application Practice

赵刚◎著

中国工程院院士 倪光南 倾情作序

雷万云 | 毛新生 | 段永朝 | 安晖 联合力荐

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

大数据

技术与应用实践指南

Big Data Technology and Application Practice

赵刚◎著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

大数据是互联网、移动应用、社交网络和物联网等技术发展的必然趋势，大数据应用成为当前最为热门的信息技术应用领域。本书由浅入深，首先概述性地分析了大数据的发展背景、基本概念，从业务的角度分析了大数据应用的主要业务价值和业务需求，在此基础上介绍大数据的技术架构和关键技术，结合应用实践，详细阐述了传统信息系统与大数据平台的整合策略，大数据应用实践的流程和方法，并介绍了主要的大数据应用产品和解决方案。最后，对大数据面临的挑战和未来的趋势进行了展望。

本书既具有技术深度，又具有很强的可操作性，提供了一个系统性、架构性的大数据应用实践指南，纲要性地指导大数据应用实践，推动大数据技术在各个行业的广泛应用。

图书在版编目（CIP）数据

大数据：技术与应用实践指南 / 赵刚著. —北京：电子工业出版社，2013.10
ISBN 978-7-121-21560-5

Ⅰ. ①大… Ⅱ. ①赵… Ⅲ. ①数据处理－指南 Ⅳ. ①TP274-62

中国版本图书馆 CIP 数据核字（2013）第 228258 号

策划编辑：董　英
责任编辑：付　睿
印　　刷：北京中新伟业印刷有限公司
装　　订：北京中新伟业印刷有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编：100036
开　　本：787×980　1/16　　印张：18.25　　字数：366 千字
印　　次：2013 年 10 月第 1 次印刷
印　　数：4000 册　　定价：59.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线：（010）88258888。

序

随着新一代信息技术的发展和应用，尤其是互联网、物联网、移动互联网、社交网络等技术的发展，我们正在进入一个大数据的时代。从大数据的理念到 Hadoop 开发技术，介绍大数据的书刊纷纷出现，但很多读者看了后可能仍感到不解渴，究其原因是这些书刊没有为读者构建一座连接宏观的理念和深奥的技术细节之间的桥梁，而有关大数据系统性应用实践的书籍则更是凤毛麟角。为此，我向大家推荐这本书，它从大数据技术应用的角度切入，建立了大数据业务价值与技术架构之间的映射关系，内容丰富，条理清晰，深入浅出，繁简适度，使读者能够系统地了解大数据的技术应用体系。

大数据从数据挖掘、商业智能发展而来，是信息技术发展的必然产物。国家“十二五”规划要大力发展包括新一代信息技术在内的战略性新兴产业，大数据就是新一代信息技术的重要领域。它不仅是一次技术领域的革新，因此不仅技术人员必须了解它、研究它、运用它，而且它还将推动企业创新和社会变革，因此各行各业的人员都必须重视它、发展它、推动它。

大数据应用不能一蹴而就，必须遵循科学的方法循序渐进。无论是从业务的角度还是从技术的角度，要将大数据应用讲清楚都不大容易，尤其是要使非本领域的专家能对大数据有一个全面的了解更非易事。为了帮助读者建立起对大数据应用全面、系统的认识，而不只是知道一些零散的技术或服务术语，作者站在系统论的高度对大数据应用做了高度的概括，涵盖大数据的基本概念、业务需求、技术架构、应用集成、实践方法、产业链和制度保障等七个方面，也构成了本书的七个章节。这种结构化、系统化的思想贯穿全书，成为本书的一大特色。这不仅对一般读者，而且对与大数据有关的管理人员和技术人员，都有帮助，使他们可以全面深刻地理解和把握复杂的大数据。

作者提出了大数据应用的业务流程，分析了行业共性业务需求和个性业务需求，并且详细阐述了满足这些业务需求的大数据技术，也介绍了新的大数据技术和现有技术架构的整合。大数据在一些互联网公司有了很好的应用，其他行业也在关注大数据。本书列举出一些实例，给出了大数据应用的流程和方法论，强调了大数据对商业社会的巨大的变革力

量。虽然大数据还是一个新事物，开始时人们难免对其有所怀疑，不敢贸然使用，但越来越多的“吃螃蟹者”已经证明大数据能创造重大的社会效益和经济效益。在当前这场大数据引领的变革浪潮面前，我们应当直面挑战、勇于创新，大胆地应用大数据技术。实际上，在激烈的市场竞争中，不创新的风险往往比创新的风险更大。

本书对大数据的写作高屋建瓴、深入浅出，这与作者的背景是分不开的。赵刚博士一直在中国电子信息产业发展研究院从事信息技术应用研究、咨询和实践工作，承担了多项信息技术战略规划和应用实施项目，有丰富的企业级信息架构的规划和建设经验。2013年，又创办了北京赛智时代信息技术咨询公司，致力于企业级大数据技术的应用咨询和实施工作，发布了银行、保险、电子商务等行业大数据应用研究报告，在大数据应用领域做了很多工作。作者从事产业研究、信息化咨询和信息系统集成的多重背景和学术造诣，使作者能把大数据的业务需求、技术架构和产业链分析在一本书中上下呼应、融会贯通地阐述清晰。

作者在本书最后提出，大数据是中国国内企业迎头赶上的大好机会。我们相信，越来越多的中国大数据公司将会用自己的创新实践证明这一点，中国完全有可能乘大数据的变革之机实现中国信息产业的跨越式发展。

综上所述，本书可以为一切想了解大数据技术应用、建设大数据企业级应用架构、享受大数据分析之美的读者提供一把开启大数据世界的钥匙，即使是对大数据有所研究的人士，本书系统性的视角也可以使他们了解全局、开阔思路，本书具有很高的参考价值。

中国工程院院士　倪光南

前　言

随着互联网、移动互联网、社交网络、物联网、云计算等新一代信息技术的应用和推广，人类产生的数据成倍增长，数据种类繁多，数据在宽带网络中高速流动，数据的待开发价值越来越大，我们已经进入了大数据时代！短短两三年，大数据的理念已经深入人心，大数据的技术也层出不穷，但大数据技术的应用才刚刚开始。本书把阐述的视角放在了大数据的技术应用上，通过分析大数据应用的关键成功因素，希望为政府、行业和企业的大数据技术开发和应用人员提供一本框架性和系统性的技术与应用实践指南。

全书共分为 7 章。

第 1 章是大数据的概念和发展背景，回顾大数据理念和技术的发展历程，梳理大数据发展脉络，并从大数据的体量、数据类型、速度和潜在价值等 4 个特征定义大数据。

大数据的技术应用是为了实现业务的价值，所以第 2 章分析大数据应用的业务需求，梳理企业级大数据应用的业务流程，剖析大数据应用对于组织的业务价值，并深入分析互联网、零售、金融、电信、能源等 9 个行业的大数据应用需求，总结企业级大数据应用的客户分析、绩效分析和风险分析等共性需求。

第 3 章阐述大数据应用的总体架构和关键技术。总体架构分析基于 Apache 开源的大数据平台总体架构参考模型，涵盖了大数据处理、大数据存储、大数据访问、大数据调度、大数据分析展现、大数据与传统数据库连接、大数据管理、安全和备份恢复框架等技术，它能够为企业建设大数据应用平台提供框架参考。基于这一架构，本章进一步详细介绍了大数据存储和处理、大数据查询分析、大数据高级分析和可视化等 3 个方面的关键技术。Hadoop 是大数据技术的内核，本章详细介绍了 Hadoop 三大核心技术，即分布式文件系统 HDFS、分布式计算框架 MapReduce、分布式数据库 HBase 的技术原理、技术构成和应用示例，也介绍了 Hadoop 之外的内存计算、流计算等框架。大数据查询和分析技术介绍了 SQL on Hadoop 技术，包括 Hive、Impala 等技术。大数据高级分析和可视化技术也是大数据的关键技术，本章总体阐述了大数据挖掘与高级分析的算法和技术，对非结构化复杂数据分析、预测分析和开源的 R 语言进行了重点介绍，并介绍了大数据可视化的一些工具。

第 4 章阐述大数据技术应用与企业级应用系统的整合策略。现有企业级数据分析是以关系型数据库为基础的，建立了涵盖网络、存储、服务器、虚拟化、云计算和信息安全等

方面的企业 IT 架构，大数据技术的企业级应用需要实现与这些技术的高效整合，构建新一代的企业级应用架构。本章分别介绍了大数据传输、集成和流程化管理，大数据与存储架构的整合，大数据对网络架构的发展，大数据与虚拟化技术的整合，云计算平台上的大数据云，以及大数据与信息安全等 6 个方面的内容。

第 5 章介绍了大数据企业级应用的实践方法论和应用案例。大数据应用的实践方法论阐述了业务需求定义、现状分析、架构规划和设计、技术切入与实施，以及试用、评估和推广等大数据应用的开发流程。对亚马逊、雅虎、淘宝等互联网企业应用案例的分析，则试图给大数据技术应用实践提供技术细节和实施规模的参考。

第 6 章介绍了大数据应用的主流商业解决方案，首先介绍大数据产业链上的主要厂商，并进一步介绍了 9 家主流厂商的解决方案。

第 7 章是对大数据应用中未来挑战和发展趋势的分析。主要讨论了隐私保护、技术标准、大数据治理等应用发展中的关键挑战和应对策略，最后预测了大数据应用下商业生活的发展趋势。

全书以某商业银行基于大数据的客户分析为案例，便于读者根据案例所阐述的应用场景，结合自身需求学习和掌握大数据技术的应用。

本书的写作最大程度地得益于从事大数据技术研发、应用和研究的社区、业界同仁和爱好者。作者起的作用仅仅是穿针引线，将大数据技术应用开拓者们分享的研究和应用心得总结起来，希望有助于更多技术研发、应用人员和爱好者系统地学习和应用大数据，本书也提供了这些成果的链接，读者可以更加深入地去学习和研究。当然，本书基于作者在信息化领域多年的研究、咨询和系统集成的实践经验，也基于作者所创立的北京赛智时代信息技术咨询有限公司（www.CIOManage.com）在大数据领域的研究成果。本书引用了 CIOManage（赛智时代）的《2013 年中国大数据应用价值研究报告》的很多研究成果。感谢所有为大数据技术应用而努力的同仁们！

本书付梓之际，作者诚惶诚恐，大数据技术远未成熟，大数据技术应用也刚刚拉开帷幕，这样一本技术应用实践指南一定存在诸多问题。但技术应用本来就是一个不断改进和优化的过程，希望我和读者在共同学习和应用的过程，逐步总结出更为精确和实用的经验。欢迎读者与我交流，联系信息如下。

- 微博：http://weibo.com/blogbot
- 博士博客：http://blog.sina.com.cn/blogbot
- 邮箱：blogbot@sina.com

赵刚

2013 年 7 月 29 日于北京嘉铭园

目　录

第 1 章　大数据的概念和发展背景 1

1.1　大数据的发展背景 1
1.2　大数据的概念和特征 4
1.2.1　大数据的概念 4
1.2.2　大数据的特征 4
1.3　大数据的产生 5
1.3.1　数据产生由企业内部向企业外部扩展 5
1.3.2　数据产生从 Web 1.0 向 Web 2.0、从互联网向移动互联网扩展 6
1.3.3　数据产生从计算机/互联网（IT）向物联网（IOT）扩展 7
1.4　数据的量级 7
1.4.1　数据大小的量级 7
1.4.2　大数据的量级 8
1.5　大量不同的数据类型 8
1.5.1　按照数据结构分类 9
1.5.2　按照产生主体分类 12
1.5.3　按照数据作用方式分类 13
1.6　大数据的速度 14
1.7　大数据的潜在价值 14
1.8　大数据的挑战 15
1.8.1　业务视角不同带来的挑战 15
1.8.2　技术架构不同带来的挑战 15
1.8.3　管理策略不同带来的挑战 16

第 2 章　大数据应用的业务需求 17

2.1　大数据应用的业务流程 17
2.1.1　产生数据 17
2.1.2　聚集数据 18
2.1.3　分析数据 19
2.1.4　利用数据 19
2.2　大数据应用的业务价值 19
2.2.1　发现大数据的潜在价值 20
2.2.2　实现大数据整合创新的价值 20
2.2.3　新领域再利用的价值 21
2.3　各行业大数据应用的个性需求 21
2.3.1　互联网与电子商务行业 21
2.3.2　零售业 27
2.3.3　金融业 28
2.3.4　政府 32
2.3.5　医疗业 34
2.3.6　能源业 36
2.3.7　制造业 37
2.3.8　电信运营业 39
2.3.9　交通物流业 41
2.4　企业级大数据应用的共性需求 42
2.4.1　客户分析 42
2.4.2　绩效分析 46
2.4.3　欺诈和风险评估 48
2.5　以银行客户分析为例，分析一个大数据的应用场景 49

第 3 章　大数据应用的总体架构和关键技术 51

3.1　总体架构 51
3.1.1　业务目标 51
3.1.2　架构设计原则 52
3.1.3　总体架构参考模型 55

3.1.4 总体架构的特点 58
3.2 大数据存储和处理技术 59
3.2.1 Hadoop：分布式存储和计算平台 59
3.2.2 Hadoop 之 HDFS：分布式文件系统 65
3.2.3 Hadoop 之 MapReduce：分布式计算框架 72
3.2.4 Hadoop 之 NoSQL：分布式数据库 98
3.2.5 Hadoop 之外的大数据计算技术 113
3.3 大数据查询和分析技术：SQL on Hadoop 126
3.3.1 Hive：基本的 Hadoop 查询和分析 127
3.3.2 Hive 2.0：Hive 的优化和升级 137
3.3.3 实时互动的 SQL：Impala 和 drill 140
3.3.4 基于 PostgreSQL 的 SQL on Hadoop 146
3.4 大数据高级分析和可视化技术 147
3.4.1 传统数据仓库与联机分析处理技术 147
3.4.2 大数据对传统分析的挑战 150
3.4.3 大数据挖掘与高级分析 150
3.4.4 大数据挖掘与高级分析库：Mahout 155
3.4.5 非结构化复杂数据分析 156
3.4.6 实时预测分析 163
3.4.7 开源可视化工具：R 语言 170
3.4.8 可视化技术 178
3.5 以银行客户分析为例的大数据的技术环境部署 187
3.5.1 银行客户大数据应用体系架构 187
3.5.2 技术环境安装与配置 189

第 4 章 大数据与企业级应用的整合策略 202

4.1 大数据传输、整合和流程管理平台 203
4.1.1 数据传输 203
4.1.2 数据整合 209
4.1.3 流程管理 211
4.2 大数据与存储架构的整合 215
4.2.1 传统存储架构比较 215

4.2.2 大数据平台的存储架构的选择 216
4.2.3 集群存储的发展 217
4.2.4 基于 HDFS 的集群存储 219
4.2.5 固态硬盘（SSD）对内存计算的支持 221
4.3 大数据与网络架构的发展 221
4.4 大数据与虚拟化技术的整合 227
4.5 在云计算平台上的大数据云 229
4.6 大数据与信息安全 231
4.7 以银行客户分析为例，分析一个大数据的平台整合 234

第 5 章 大数据应用的实践方法与案例 235

5.1 实践方法论 235
5.1.1 业务需求定义 235
5.1.2 数据应用现状分析与标杆比较 237
5.1.3 大数据应用架构规划和设计 238
5.1.4 大数据技术切入与实施 239
5.1.5 大数据试用和评估 240
5.1.6 大数据应用推广 241
5.2 应用案例 241
5.2.1 亚马逊 241
5.2.2 雅虎 242
5.2.3 淘宝网 242
5.2.4 Facebook 243
5.3 以银行客户分析为例的实施案例分析 244
5.3.1 银行基于大数据的客户分析的业务需求 244
5.3.2 银行基于大数据的客户分析的现状与标杆比较 245
5.3.3 银行基于大数据的客户分析的应用架构规划与设计 246
5.3.4 银行基于大数据的数据分析的实施、试点和推广 247

第 6 章 大数据应用的主流解决方案 248

6.1 产业链 248
6.1.1 国际上的大数据生态环境 248

6.1.2 国内产业链主要力量 251
6.2 主流厂商解决方案 252
6.2.1 Cloundera 252
6.2.2 Hortonworks 254
6.2.3 MapR 254
6.2.4 IBM 255
6.2.5 Oracle 257
6.2.6 EMC 258
6.2.7 Intel 259
6.2.8 SAP 260
6.2.9 Teradata 262

第 7 章 大数据应用的未来挑战和趋势 263

7.1 隐私保护 263
7.1.1 法律保护 264
7.1.2 技术保护 266
7.1.3 理念革新 267
7.2 技术标准 268
7.2.1 ISO 标准化进展 268
7.2.2 评价基准和基准测试 269
7.2.3 标准套件 273
7.3 大数据治理 273
7.3.1 数据治理框架 274
7.3.2 数据质量管理 274
7.3.3 大数据的组织、角色和责任 276
7.4 适应商业社会的未来趋势 277
7.4.1 从产品推销向数据营销的转变 277
7.4.2 从流程驱动到分析驱动的转变 277
7.4.3 从私有资源到公共服务的转变 278

第 1 章

大数据的概念和发展背景

本章阐述大数据的概念、发展背景和内涵等。

1.1 大数据的发展背景

在 20 世纪 90 年代后期，当气象学家在做气象地图分析、物理学家在建立大物理仿真模型、生物学家在建立基因图谱的分析过程中，由于数据量巨大，他们已经不能再用传统的计算技术来完成这些任务时，大数据的概念在这些科学研究领域首先被提出来。面对大量科学数据在获取、存储、搜索、共享和分析中遇到的技术难题，一些新的分布式计算技术陆续被研究和开发出来。

2008 年，随着互联网和电子商务的快速发展，当雅虎、谷歌等大型互联网和电子商务公司不能用传统手段解决他们的业务问题时，大数据的理念和技术被他们实际应用。他们遇到的共性问题是，处理的数据量通常很大（那时是 PB 级，1 个 PB 的数据相当于 50% 的全美学术研究图书馆藏书资讯内容），数据的种类很多（文档、日志、博客、视频等），数据的流动速度很快（包括流文件数据、传感器数据和移动设备的数据的快速流动）。而

且，这些数据经常是不完备甚至是不可理解的（需要从预测分析中推演出来）。大数据的新技术和新架构正是在这种背景下被不断开发出来的，以有效地解决这些现实的互联网数据处理问题。

2010 年，全球进入 Web 2.0 时代，Twitter（推特）、Facebook（脸书）、博客、微博、微信等社交网络将人类带入自媒体时代，互联网数据快速激增。随着苹果、三星等智能手机的普及，移动互联网时代也已经到来，移动设备所产生的数据海量般地涌入网络。为了实现更加智能的应用，物联网技术也逐步被推广，随之而来的是更多实时获取的视频、音频、电子标签（RFID）、传感器等数据也被联入互联网，数据量进一步暴增。根据美国市场调查公司 IDC 的预测[1]，人类产生的数据量正在呈指数级增长，大约每两年翻一番，这个速度在 2020 年之前会继续保持下去。全球在 2010 年正式进入 ZB 时代（1 个 ZB 的数据相当于全世界海滩上的沙子数量的总和），预计到 2020 年，全球将总共拥有 35ZB 的数据量。这意味着人类在最近两年产生的数据量相当于之前产生的全部数据量。人类真正进入了一个数据的世界，大数据技术有了用武之地，大数据技术和应用空前繁荣起来。

2011 年，全球著名战略咨询公司麦肯锡的全球研究院（MGI）发布了《大数据：创新、竞争和生产力的下一个新领域》研究报告[2]，这份报告分析了数字数据和文档的爆发式增长的状态，阐述了处理这些数据能够释放出的潜在价值，分析了大数据相关的经济活动和业务价值链。这篇报告在商业界引起极大的关注，为大数据从技术领域进入商业领域吹响了号角。

2012 年 3 月 29 日奥巴马政府以“大数据是一个大生意（Big Data is a Big Deal）”[3]为题发布新闻（如图 1-1 所示），宣布投资 2 亿美元启动“大数据研究和发展计划”，涉及美国国家科学基金、美国国防部等 6 个联邦政府部门，大力推动和改善与大数据相关的收集、组织和分析工具及技术，以推进从大量的、复杂的数据集合中获取知识和洞见的能力。美国政府认为大数据技术事关美国国家安全、科学和研究的步伐。

---

[1] 2010 年 IDC 提供给 EMC 的报告，见 http://www.emc.com/about/news/press/2010/20100504-01.htm。

[2] http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation。

[3] http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal。

图 1-1　美国白宫发布的大数据新闻

2012 年 5 月，联合国发布了一份大数据白皮书[4]，总结了各国政府如何利用大数据更好地服务公民，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，联合国还探讨了如何利用包括社交网络在内的大数据资源造福人类。

2012 年 12 月“世界经济论坛”发布《大数据，大影响》报告[5]，阐述大数据为国际发展带来的新的商业机会，建议各国与工业界、学术界、非营利性机构与管理者一起利用大数据所创造的机会。

2012 年以来，大数据成为全球投资界最青睐的领域之一，IBM 公司通过并购数据仓库厂商 Netezza、软件厂商 InfoSphere BigInsights 和 Streams 等来增强自己在大数据处理上的实力；EMC 公司陆续收购 Greenplum（Pivotal）、VMWare、Isilo 等公司，展开大数据和云计算产业的战略布局；惠普公司通过并购 3PAR、Autonomy、Vertica 等公司实现了大数据产业链的全覆盖。业界主要的信息技术巨头都纷纷推出大数据产品和服务，力图抢占市场先机。

2012 年以来，国内互联网企业和运营商率先启动大数据技术的研发和应用，如新浪、淘宝、百度、腾讯、中国移动、中国联通、京东商城等企业纷纷启动了大数据试点应用项

[4] http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-GlobalPulseMay2012.pdf。

[5] http://www3.weforum.org/docs/WEF_TC_MFS_BigDataBigImpact_Briefing_2012.pdf。

目，推进大数据应用。

2013 年，第 4 期《求是》杂志刊登中国工程院邬贺铨院士的《大数据时代的机遇与挑战》[6]一文，阐述中国科技界对大数据的重视，郭华东、李国杰、倪光南、怀进鹏等院士也纷纷撰文阐述大数据的战略意义。工信部软件服务业司陈伟司长指出："可以理解，大数据是云计算、物联网、移动互联网、智慧城市等新技术、新模式发展的必然产物"，表明产业主管部门对大数据发展的高度关注。

1.2 大数据的概念和特征

1.2.1 大数据的概念

大数据是指无法在一定时间内用传统数据库软件工具对其内容进行抓取、管理和处理的数据集合（引自维基百科[7]）。

这个定义并不严谨，但这是各种学术和应用领域最广泛引用的一个定义，如果接着以大数据的四个特征作为补充，就能给出一个较为清晰的大数据的概念。

1.2.2 大数据的特征

大数据有四个主要特征。

1. Volume：数据量巨大

体量大是大数据区分于传统数据最显著的特征。一般关系型数据库处理的数据量在 TB 级，大数据所处理的数据量通常在 PB 级以上。

2. Variety：数据类型多

大数据所处理的计算机数据类型早已不是单一的文本形式或者结构化数据库中的表，

[6] 邬贺铨，《大数据时代的机遇与挑战》，《求是》杂志，2013 年 2 月。

[7] http://en.wikipedia.org/wiki/Big_data。

它包括订单、日志、BLOG、微博、音频、视频等各种复杂结构的数据。

3. Velocity：数据流动快

速度是大数据区分于传统数据的重要特征。在海量数据面前，需要实时分析获取需要的信息，处理数据的效率就是组织的生命。

4. Value：数据潜在价值大

在研究和技术开发领域，上述三个特征已经足够表征大数据的特点。但在商业应用领域，第四个特征就显得非常关键！投入如此巨大的研究和技术开发的努力，就是因为大家都洞察到了大数据的潜在的巨大价值。如何通过强大的机器学习和高级分析更迅速地完成数据的价值“提纯”，挖掘出大数据的潜在价值，这是目前大数据应用背景下亟待解决的难题。

1.3　大数据的产生

大量数据的产生是计算机和网络通信技术（ICT）广泛应用的必然结果，特别是互联网、云计算、移动互联网、物联网、社交网络等新一代信息技术的发展，起到了催化剂的作用，它带来了数据产生的四大变化：一是数据产生由企业内部向企业外部扩展；二是数据产生由 Web 1.0 向 Web 2.0 扩展；三是数据产生由互联网向移动互联网扩展；四是数据产生由计算机/互联网（IT）向物联网（IOT）扩展。这 4 个变化，让数据产生源头成倍地增长，数据量也大幅度地快速增长。

1.3.1　数据产生由企业内部向企业外部扩展

在企业内部的企业资源计划（ERP）、办公自动化（OA）等业务、管理和决策分析系统所产生的数据，主要存储在关系型数据库中。内部数据是企业内最成熟并且被熟知的数据。这些数据已经通过多年的 ERP、主数据管理（MDM）、数据仓库（DW）、商业智能（BI）和其他相关应用积累，实现了内部数据的收集、集成、结构化和标准化处理，可以为企业决策提供分析报表和商业智能。