

EXPERT'S VOICE IN OPEN SOURCE

Architecting, Developing,
and Administering MongoDB

MongoDB实战

架构、开发与管理的

[美] Shakuntala Gupta Edward
Navin Sabharwal
蒲成

著
译



Apress®

清华大学出版社

MongoDB 实战

架构、开发与管理

[美] Shakuntala Gupta Edward 著
Navin Sabharwal 译
蒲 成

清华大学出版社

北 京

Shakuntala Gupta Edward, Navin Sabharwal
Practical MongoDB: Architecting, Developing, and Administering MongoDB
EISBN: 978-1-4842-0648-5
Original English language edition published by Apress Media. Copyright © 2015 by Shakuntala Gupta Edward and Navin Sabharwal. Simplified Chinese-language edition Copyright © 2016 by Tsinghua University Press. All rights reserved.

本书中文简体字版由 Apress 出版公司授权清华大学出版社出版。未经出版者书面许可，不得以任何方式复制或抄袭本书内容。

北京市版权局著作权合同登记号 图字：01-2016-8570

本书封面贴有清华大学出版社防伪标签，无标签者不得销售。
版权所有，侵权必究。侵权举报电话：010-62782989 13701121933

图书在版编目(CIP)数据
MongoDB 实战 架构、开发与管理 / (美) 夏琨塔拉·古普塔·爱德华 等著；蒲成译. —北京：清华大学出版社，2017
书名原文：Practical MongoDB:Architecting, Developing,and Administering MongoDB
ISBN 978-7-302-45673-5
I. ①M… II. ①夏… ②蒲… III. ①关系数据库系统 IV. ①TP311.138
中国版本图书馆 CIP 数据核字(2016)第 281183 号

责任编辑：王 军 于 平
装帧设计：牛静敏
责任校对：成凤进
责任印制：李红英

出版发行：清华大学出版社
网 址：http://www.tup.com.cn, http://www.wqbook.com
地 址：北京清华大学学研大厦 A 座 邮 编：100084
社 总 机：010-62770175 邮 购：010-62786544
投稿与读者服务：010-62776969, c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015, zhiliang@tup.tsinghua.edu.cn

印 刷 者：北京鑫丰华彩印有限公司
装 订 者：三河市溧源装订厂
经 销：全国新华书店
开 本：185mm×260mm 印 张：16.25 字 数：375 千字
版 次：2017 年 1 月第 1 版 印 次：2017 年 1 月第 1 次印刷
印 数：1~3500
定 价：49.80 元

产品编号：071606-01

译者序

随着以用户为中心的社交类网站的兴起，尤其是在当前飞速发展的移动互联时代背景下，传统的关系数据库在应对海量的信息，特别是超大规模和高并发的社交网络类型应用所带来的爆炸式数据时已经显得力不从心，暴露了很多难以克服的问题，而非关系型数据库则由于其本身的特点能够快速适应这些应用场景，因而得到了非常迅速的发展。

虽然 NoSQL 的流行不过短短数年时间，但不可否认的是，在持续不断的版本迭代之下，现在的 NoSQL 系统已经更加成熟、稳定。在这众多的 NoSQL 工具中，MongoDB 可以说是其中的佼佼者，它可以为大数据建立快速、可扩展的存储库，从而满足日新月异的应用场景需求。

MongoDB 是一个基于分布式文件存储的数据库。它的基础开发语言是 C++。其目的在于为基于互联网的应用提供可扩展的高性能数据存储解决方案。对于大型互联网公司以及处在飞速发展之中的互联网公司来说，以 MongoDB 为代表的 NoSQL 数据库产品正逐渐成为其无法绕过的必备部署工具之一。

实际上，MongoDB 是一个介于关系型数据库和非关系型数据库之间的产品，它是非关系型数据库中功能最丰富、最像关系型数据库的一个。它支持松散的数据结构，即基于 JSON 的 BSON 格式，因而在实际使用中，MongoDB 可以存储较为复杂的数据类型。它的特点是高性能、易部署、易使用，并且存储数据非常方便。不过，其最大的优势在于所支持的查询语言非常强大，其语法有点类似于面向对象的查询语言，几乎可以实现类似关系型数据库单表查询的绝大部分功能，此外还支持对数据建立索引。

本书的两位作者都是数据库和大数据应用领域经验丰富的实践专家，他们在数据分析应用领域的见解自然是独一无二且领先于行业的。相信在阅读完本书并且充分吸收书中的理念之后，读者将对 MongoDB 有一个全面了解，从而帮助读者深刻理解 MongoDB 在现实环境中的应用与实践。本书使用了大量贴合实际的示例，结合作者的亲身实践以及多年的丰富经验来介绍 MongoDB 的概念与应用原则。重点在于让读者理解其价值和实现的机制及原理，而不是空谈枯燥乏味的理论知识点。

本书提供了使用 MongoDB 平台进行架构、开发与部署应用程序的清晰指导与实践示例。数据库开发人员、架构师和管理员将在本书中找到涵盖 MongoDB 平台所有知识点的有用信息，以及如何将它用于实践的内容。当下，互联网、特别是移动互联网正处于高速发展时期，已有的工具要跟上时代的步伐就必须不断地推陈出新，MongoDB 的开发团队也在不知疲倦地对 MongoDB 进行更新和调整，使其能跟上时代发展的脚步。作

为了一本 MongoDB 实践领域的权威书籍，本书能让你快速、专业地掌握以 MongoDB 为基础的重点知识，从而构建高效、高并发的应用。

在此要特别感谢清华大学出版社的编辑们，在本书翻译过程中他们提供了颇有助益的帮助，没有其热情付出，本书将难以付梓。

本书全部章节由蒲成翻译，参与本次翻译的还有何东武、李鹏、林超、刘洋洋、茆永锋、潘丽臣、王滨、申成龙、王佳、负书谦。由于译者水平有限，难免会出现一些错误或翻译不准确的地方，如果有读者能够指出并勘正，译者将不胜感激。

译 者

作者简介

Shakuntala Gupta Edward 从 10 年前就开始使用数据库技术。她的经验涵盖了 SQL Server、Oracle 数据库、Analytics 平台以及大数据技术，例如 MongoDB、Cassandra 和 SAP HANA。

Shakuntala 是一位才华横溢的架构师，擅长于利用各种数据库技术为各种业务领域构建产品和解决方案。

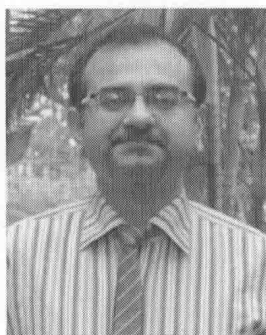
Shakuntala 一直在参与利用大数据技术 MongoDB 和 Cassandra 来开发产品和解决方案的工作。Shakuntala 拥有计算机应用的硕士学位。

Navin Sabharwal 是一位创新者、思想领袖、作者、顾问。他专注于报表与分析领域，包括 SQL Server、Oracle、MySql 在内的 RDBMS 技术以及包括 Hadoop、MongoDB 和 SAP HANA 在内的大数据技术。Navin 一直在使用大数据技术为 IT 服务管理、产品开发、云计算、云生命周期管理以及社交网络产品开发领域构建产品和服务。

Navin 已经构建出有良好商机的屡获殊荣的产品和解决方案，并且在各个领域都取得了大量专利，比如 IT 服务、评估引擎、排名算法、容量规划引擎以及知识管理。

Navin 还著有以下书籍：*Cloud Computing First Steps*(CreateSpace 出版，ISBN#: 978-1478130086)、*Apache Cloudstack Cloud Computing*(Packt Publishing 出版，ISBN#: 978-1782160106)、*Cloud Capacity Management*(Apress 出版，ISBN #: 978-1430249238)。Navin 拥有信息技术方面的硕士学位，并且是经认证的项目管理专家(Certified Project Management Professional)。

技术审校者简介



Prasoon Kumar 是一位经验丰富的技术专家和培训师，他在构建世界一流的软件产品方面拥有超过 18 年的丰富经验。他在 1993 年的印度工程学系入学考试(IIT-JEE)中从 10 万名考生中脱颖而出，该入学考试可以说是世界上最具挑战性和竞争性的考试。

他现在主要居住在班加罗尔，与印度和美国硅谷的像 MongoDB、Justdial、Avaya 等这样的公司都有广泛接触。他使用 MySQL、MongoDB、HBase、Apache Solr、Elasticsearch、PHP、Node.JS 来为复杂应用程序构建可扩展的后端。

他担当推广用于横向扩展数据存储的混合持久化以及 NoSQL 解决方案的角色。他帮助印度的大型电子商务、FSI、医疗卫生和出版公司解决了文档管理、高流量网站的扩展性需求问题。他已经为 TB 量级的数据存储做过调整、优化、备份、恢复、迁移和升级。他在 <http://prasoonk.wordpress.com> 上发表了关于技术、商业、黑客马拉松以及创业精神的博文。



Sundar Rajan Raman 是一位大数据架构师，他目前供职于美国银行。他拥有位于印度锡尔杰尔的国立技术学院的技术学士学位。在他供职于 AT&T、新加坡电信、德意志银行期间，他逐渐成为经验丰富的 Java 和 J2EE 编程人员。他是一位消息传递平台的专家，在 Sonic MQ、Websphere MQ、TIBCO 方面拥有丰富的经验，并且具有这些产品各自的认证。他当前主要专注于大数据技术。目前他正在使用 Hadoop 及其生态系统，比如 Pig、HIVE、Oozie 和 Storm、Spark 等。他为 AT&T 架构了一个基于 MongoDB

的分析引擎。



致 谢

要特别感谢那些帮助编著本书的人，Rajeev Pratap Singh 和 Amit Agrawal 为书中的代码片段提供了帮助，Dheeraj Raghav 为本书的内容设计提供了创意。

万分感谢 Stuti Awasthi，他促成了本书并且为本书提供了灵感。

本书作者要感谢大数据技术和开源社区的创建者，因为他们提供了如此强大的工具和技术用于编码，并且使轻易快速解决真实业务问题的产品和解决方案得以实现。

前言

如今，数据仓库作为一个行业已经存在很多年了。关系型数据库被用于存储数据已经几十年了，同时 SQL 已经成为实际上的与 RDBMS 交互的语言。随着社交网络、物联网以及互联网上巨量的非结构化数据的涌现，数据存储、处理以及分析的需求正逐渐爆发。传统的 RDBMS 系统和存储技术并非旨在处理各种各样海量的数据。

因此，大数据技术诞生了，如今它推动着各个互联网规模公司及其巨量数据的发展。像 Facebook、Twitter、Google 以及雅虎这样的公司正在利用大数据技术提供互联网规模的产品和服务，它们能够支持数百万的用户。

本书将帮助读者理解大数据技术、其出现的背景、需求，然后我们将介绍与使用 MongoDB 架构解决方案有关的深层技术观点。本书将让读者能够理解适合使用大数据技术的关键用例，也会为读者提供关于应该在何处小心使用大数据技术或者结合传统 RDBMS 技术来提供灵活解决方案的指导。

顺着本书的内容结构阅读，我们旨在提供关于学习 MongoDB 和使用 MongoDB 创建应用程序及解决方案的分步指南。

我们衷心希望我们的读者能够享受到阅读本书的乐趣，就像我们享受了编写本书的乐趣一样。

本书内容

- 作为一本指南，将帮助读者领会大数据技术中的各种专业术语并且牢牢掌握大数据的各个方面。
- 作为一本指南，将帮助读者理解 NoSQL 和基于文档的数据库，以及它们与传统的关系型数据库有多么不同。
- 提供了使用 MongoDB 架构解决方案的见解，还提供了 MongoDB 作为一个工具所受限制的信息。
- 系统地介绍了 MongoDB 的架构、开发、管理和数据模型。
- 引用了示例，以便让用户轻松地开始学习该技术。

阅读本书你需要做的准备

MongoDB 支持大多数主流平台。

可以从 MongoDB 下载页面(<http://www.mongodb.org/downloads/>)上下载 MongoDB 最新稳定的正式版本。

在本书中，我们将专注于在 64 位 Windows 平台上使用 MongoDB，并且在许多地方也引用了如何使用在 Linux 上运行的 MongoDB 的参考。

我们将使用 64 位的 Windows 2008 R2 和 Linux 系统来提供安装过程的示例。

本书读者对象

对于编程人员、大数据架构师、应用程序架构师、技术爱好者、学生、解决方案专家以及那些希望选择合适的大数据产品来满足其需求的人来说，本书将会很有意义。

本书介绍了与大数据、NoSQL 以及在 MongoDB 上架构和开发的详细信息有关的内容。因此它为使用 MongoDB 的开发人员、架构师和运营团队提供了用例。

目 录

第 1 章 大数据	1	2.2.1 定义	14
1.1 入门指南	1	2.2.2 NoSQL 简史	14
1.2 大数据	3	2.3 ACID 对比 BASE	15
1.3 大数据源	4	2.3.1 CAP 定理	15
1.4 大数据的三个 V	5	2.3.2 BASE	16
1.4.1 数量	6	2.4 NoSQL 的优缺点	17
1.4.2 多样性	6	2.4.1 NoSQL 的优点	17
1.4.3 速率	7	2.4.2 NoSQL 的缺点	18
1.5 大数据的使用	7	2.5 SQL 与 NoSQL 数据库的 对比	18
1.5.1 可见性	8	2.6 NoSQL 数据库的种类	21
1.5.2 发现和分析信息	8	2.7 本章小结	22
1.5.3 市场细分和产品定制	8	第 3 章 MongoDB 介绍	23
1.5.4 协助决策	8	3.1 历史	23
1.5.5 创新	8	3.2 MongoDB 设计原则	24
1.6 大数据的挑战	9	3.2.1 高速、可扩展性与敏捷性	24
1.6.1 政策与程序	9	3.2.2 非关系型方法	24
1.6.2 访问数据	9	3.2.3 基于 JSON 的文档存储	25
1.6.3 技术与技能	9	3.2.4 性能与功能对比	25
1.7 传统系统与大数据	10	3.2.5 随处都能运行数据库	25
1.7.1 大数据的结构	10	3.3 与 SQL 的对比	26
1.7.2 数据存储	10	3.4 本章小结	26
1.7.3 数据处理	10	第 4 章 MongoDB 数据模型	27
1.8 大数据技术	10	4.1 数据模型	27
1.9 本章小结	11	4.1.1 JSON 和 BSON	28
第 2 章 NoSQL	13	4.1.2 标识符(_id)	29
2.1 SQL	13	4.1.3 固定集合	30
2.2 NoSQL	13		

4.2 多态模式..... 30

4.2.1 面向对象编程..... 30

4.2.2 模式演化..... 31

4.3 本章小结..... 32

第 5 章 MongoDB-安装与配置..... 33

5.1 选择你的版本..... 33

5.2 在 Linux 上安装 MongoDB 33

5.2.1 使用仓储进行安装..... 34

5.2.2 手动安装..... 34

5.3 在 Windows 上安装 MongoDB..... 35

5.4 运行 MongoDB..... 35

5.4.1 先决条件..... 35

5.4.2 开启服务..... 36

5.5 验证安装结果..... 36

5.6 MongoDB Shell 36

5.7 保障部署安全..... 37

5.7.1 使用身份验证和授权..... 37

5.7.2 控制网络访问..... 42

5.8 使用 MongoDB 云管理器进行配置..... 46

5.9 本章小结..... 50

第 6 章 使用 MongoDB Shell..... 51

6.1 基本查询..... 51

6.1.1 创建和插入..... 56

6.1.2 显式创建集合..... 58

6.1.3 使用循环插入文档..... 58

6.1.4 通过显式指定_id进行插入..... 59

6.1.5 更新..... 59

6.1.6 删除..... 61

6.1.7 读取..... 62

6.1.8 使用索引..... 68

6.2 进阶介绍..... 78

6.2.1 使用条件操作符..... 78

6.2.2 正则表达式..... 80

6.2.3 MapReduce 81

6.2.4 aggregate()..... 83

6.3 设计应用程序的数据模型 84

6.3.1 关系型数据模型与标准化..... 84

6.3.2 MongoDB 文档数据模型方法..... 86

6.4 本章小结..... 93

第 7 章 MongoDB 架构..... 95

7.1 核心程序..... 95

7.1.1 mongod..... 95

7.1.2 mongo 95

7.1.3 mongos..... 96

7.2 MongoDB 工具..... 96

7.3 独立部署..... 96

7.4 复制..... 97

7.4.1 主/从复制..... 97

7.4.2 副本集..... 98

7.4.3 实现带有副本集的高级群集..... 115

7.5 分片..... 124

7.5.1 分片组件..... 126

7.5.2 数据分发过程..... 127

7.5.3 数据平衡过程..... 130

7.5.4 操作..... 133

7.5.5 实现分片..... 134

7.5.6 控制集合分布 (基于标签分片)..... 142

7.5.7 在将数据导入到分片环境时 要记住的要点..... 152

7.5.8 监控分片..... 153

7.5.9 监控配置服务器..... 153

7.6 生产环境群集架构..... 153

7.6.1 场景 1..... 154

7.6.2 场景 2..... 155

7.6.3 场景 3..... 156

7.6.4 场景 4..... 157

7.7 本章小结..... 158

第 8 章 MongoDB 阐释.....	159	9.4.6 识别和修复集合级别的数据.....	204
8.1 数据存储引擎.....	159	9.5 监控 MongoDB.....	205
8.2 (与 MMAPv1 相关的)数据文件.....	161	9.5.1 mongostat.....	205
8.3 (与 WiredTiger 相关的)数据文件.....	168	9.5.2 mongod 网络接口.....	206
8.4 读取和写入.....	170	9.5.3 第三方插件.....	206
8.5 使用日志时如何写入数据.....	172	9.5.4 MongoDB 云管理器.....	206
8.6 GridFS——MongoDB 文件系统.....	176	9.6 本章小结.....	212
8.6.1 GridFS 的基本原理.....	177	第 10 章 MongoDB 用例.....	213
8.6.2 GridFS 的底层机制.....	177	10.1 用例 1——性能监控.....	213
8.6.3 使用 GridFS.....	179	10.1.1 模式设计.....	213
8.7 索引.....	182	10.1.2 操作.....	214
8.7.1 索引类型.....	183	10.1.3 分片.....	218
8.7.2 行为和限制.....	188	10.1.4 管理数据.....	219
8.8 本章小结.....	189	10.2 用例 2——社交网络.....	220
第 9 章 管理 MongoDB.....	191	10.2.1 模式设计.....	220
9.1 管理工具.....	191	10.2.2 操作.....	222
9.1.1 mongo.....	191	10.2.3 分片.....	225
9.1.2 第三方管理工具.....	191	10.3 本章小结.....	225
9.2 备份和恢复.....	191	第 11 章 MongoDB 使用限制.....	227
9.2.1 数据文件备份.....	192	11.1 MongoDB 的空间过大(对于 MMAPv1 而言).....	227
9.2.2 mongodump 和 mongorestore.....	192	11.2 内存问题(对于 MMAPv1 而言).....	228
9.2.3 fsync 和锁.....	196	11.3 32 位与 64 位对比.....	228
9.2.4 从备份.....	198	11.4 BSON 文档.....	228
9.3 导入和导出.....	198	11.5 命名空间使用限制.....	229
9.3.1 mongoimport.....	198	11.6 索引使用限制.....	229
9.3.2 mongoexport.....	199	11.7 固定集合使用限制——固定集合中文档的最大数量.....	229
9.4 管理服务器.....	199	11.8 分片使用限制.....	229
9.4.1 启动一台服务器.....	199	11.8.1 及早分片以避免出现问题.....	230
9.4.2 停止服务器运行.....	200	11.8.2 不能更新分片键.....	230
9.4.3 浏览日志文件.....	200	11.8.3 分片集合使用限制.....	230
9.4.4 服务器状态.....	201		
9.4.5 识别和修复 MongoDB.....	203		

- 11.8.4 选择合适的分片键..... 230
- 11.9 安全性限制.....230
 - 11.9.1 默认情况下没有身份验证..... 230
 - 11.9.2 与 MongoDB 的交互通信没有被加密..... 231
- 11.10 写入和读取限制.....231
 - 11.10.1 大小写敏感的查询 231
 - 11.10.2 类型敏感的字段..... 231
 - 11.10.3 没有联结..... 231
 - 11.10.4 事务..... 231
- 11.11 MongoDB 不适用的范围...232
- 11.12 本章小结.....232

- 第 12 章 MongoDB 的最佳实践.....233
 - 12.1 部署.....233
 - 12.1.1 MongoDB 网站的硬件配置建议..... 234
 - 12.1.2 要注意的一些要点..... 235
 - 12.2 编码.....235
 - 12.3 应用程序响应时间优化.....238
 - 12.4 数据安全性.....238
 - 12.5 管理.....239
 - 12.6 复制延迟.....239
 - 12.7 分片.....240
 - 12.8 监控.....240
 - 12.9 本章小结.....241

“大数据是用于描述海量的、具有各种结构并且高速生成的数据的一个术语。这类数据对用于存储和处理数据的传统 RDBMS 系统提出了挑战。大数据为处理和存储数据的新途径铺平了道路。”

在本章中，我们将探讨大数据基础、来源以及挑战，将介绍大数据的三个 V(数量(volume)、速率(velocity)和多样性(variety))以及在面临处理大数据时传统技术所受到的限制。

1.1 入门指南

大数据与云、社交、分析以及移动性一样，都是当前信息技术世界中的流行语。供应给大众使用的互联网和电子设备，其数量每一天都在增长。尤其是，智能手机、社交网络站点以及像平板电脑和传感器这样的其他数据生成设备都在导致数据爆炸性增长。数据是从各种来源中生成的，具有各种格式，例如视频、文本、语音、日志文件以及图片。一个高清(HD)视频的每一秒会生成比单页文本多 2000 倍的字节。

思考一下在 Facebook 公司网站上发表的与该公司有关的以下统计数据：

(1) 2015 年 6 月平均每天有 9 亿 6 千 8 百万日活跃用户。2015 年 6 月平均每天有 8 亿 4 千 4 百万移动端日活跃用户。

(2) 截至 2015 年 6 月 30 日，有 14 亿 9 千万月活跃用户。截至 2015 年 6 月 30 日，有 13 亿 1 千万移动端月活跃用户。

(3) 截至 2013 年 5 月，每天会生成 45 亿个点赞，这一数据相对于 2012 年 8 月增长了 67%。

图 1-1 描述了 Twitter 的统计数据。

这里有另一个示例：想象一下像去看一场电影这样的简单事件所能生成的数据量。你首先要在电影评论网站上搜索一部电影，阅读关于该电影的评论，并且提交查询。你可能会发表关于该电影的推文或者在 Facebook 上发布打算去看该电影的照片。在去影院的途中，你的 GPS 系统会追踪你的路线并且生成数据。

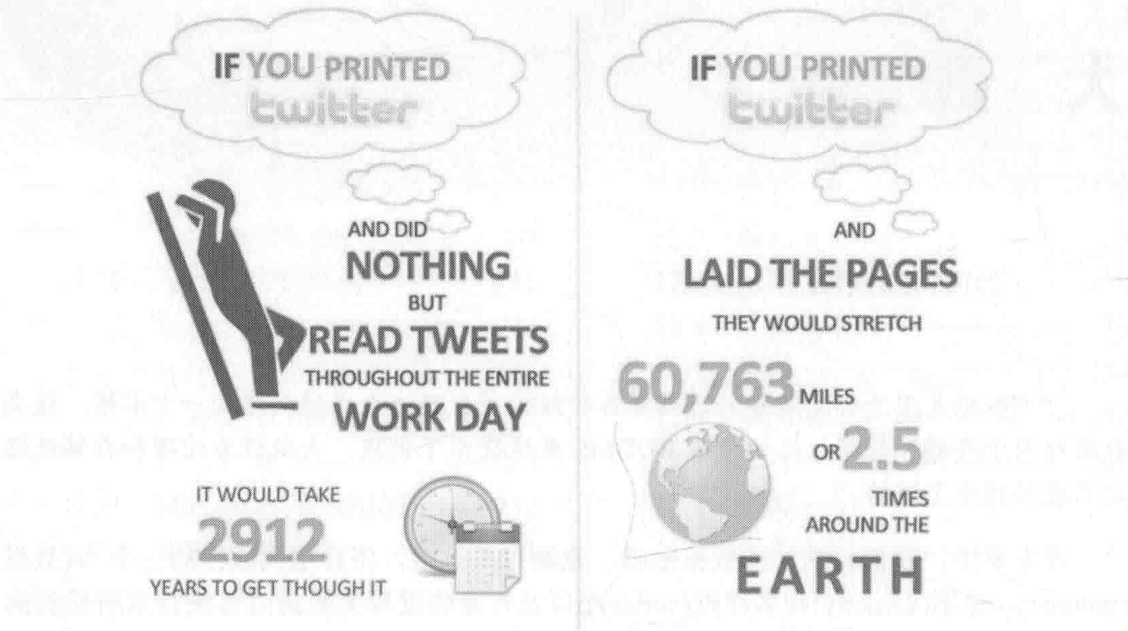


图 1-1 如果将 Twitter 印刷出版.....

你会发现这样的场景：智能手机、社交网络站点及其他媒体都在创造数据洪流以便这些公司能够处理和存储。当数据的大小对典型软件工具捕获、处理、存储和管理数据的能力提出挑战时，我们就面临大数据的处理。图 1-2 以图形化的方式定义了大数据。

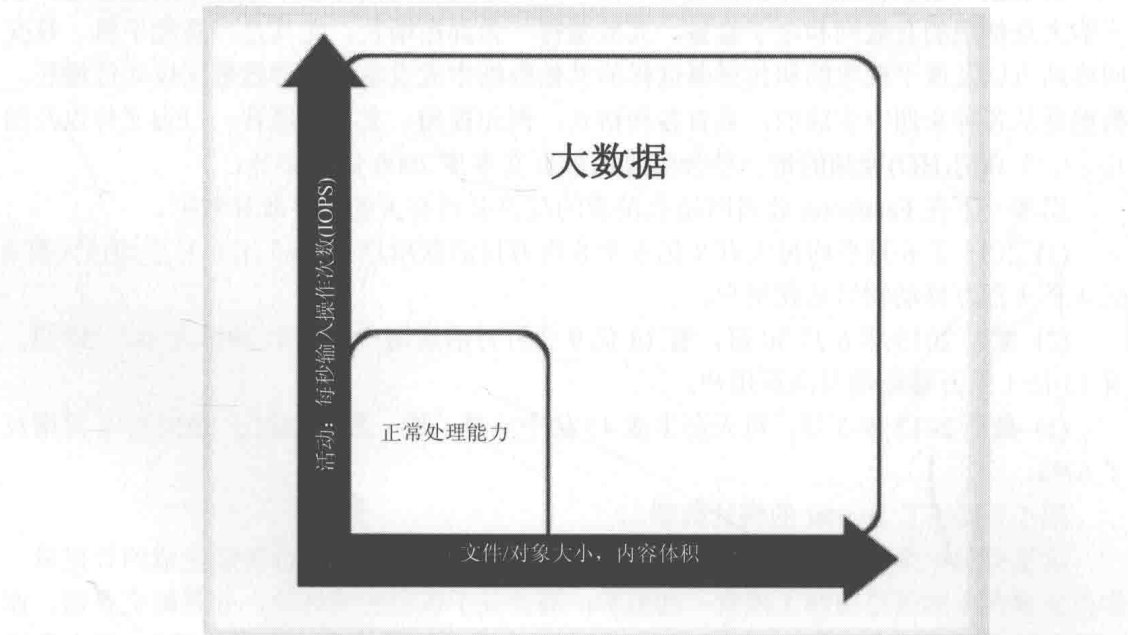


图 1-2 大数据的定义

1.2 大数据

大数据是具有高容量的、高速生成的并且具有多种样式的数据。我们来看看大数据的一些实际情况。

大数据的一些实际情况

世界各地的各种研究团队已经对所产生的大量数据进行了分析。例如，IDC 的分析表明，一年中(2007 年)所产生的数字数据量要比整个世界用于存储它的总体容量还大，这意味着没有办法存储所产生的所有数据。另外，数据产生的速度很快就会超过数据存储能力扩充的速度。

后面几节涵盖了来自 MGI(Mckinsey Global Institute, 麦肯锡全球研究所)于 2011 年 5 月发布的报告(www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation)中的见解。该研究明确地表明，大数据的商业和经济潜力及其广阔的影响是很重要的问题，商业领袖和政策制定者必须应对该问题。

1. 大数据的大小因行业而异

大数据的增长是在每一个行业中都被观测到的一种现象。MGI 估算出，2010 年世界各地的企业使用了超过 7 艾字节(exabyte)的增额硬盘驱动数据存储能力；有意思的是，其中近 80%的数据看起来都是在其他地方被存储过的重复数据。MGI 还估算出，截至 2009 年，美国经济中涉及的所有行业几乎每家公司平均都有至少 200 太字节(terabyte)的存储数据，并且许多行业中每家公司平均都有超过 1 拍字节(petabyte)的存储数据。

有些行业展现出了远高于其他行业的数据强度水平；在这里，数据强度指的是该行业中跨公司/企业所累积的数据的平均量，这表明这些公司/企业拥有更多的从大数据中获得价值的潜力。

金融服务行业，其中包括银行、投资以及安全服务，都是高度交易导向的；根据法规，它们也被要求存储数据。该分析表明，从平均值来看它们中的每家企业存储了最大的数字数据量。

通信和媒体公司、公用事业以及政府，这些领域中的每家企业或组织同样存储了大量的数字数据，这似乎反映出，像这样的实体具有大量的操作和多媒体数据。

离散型和流程式制造业具有以字节方式存储的最高水平的聚合数据。不过，这些行业在强度方面的排名要远低得多，因为它们都被划分成了大量的企业。

2. 大数据的类型因行业而异

该 MGI 研究还表明，数据存储的类型也会因行业而异。例如，零售业和批发业、政府的行政管理部门以及金融服务都会产生大量的文本和数值数据，其中包括客户数据、交易信息以及数学建模和模拟。像制造、医疗卫生、媒体和通信这样的行业都要负责处理和存储较高比例的多媒体数据。而形如 X 光、CT 和其他扫描的图片数据在医疗卫生