大数据分析5个重要阶段：
数据获取和记录
数据抽取、清洗和注记
数据集成、聚集和表示
数据分析和建模
数据解释

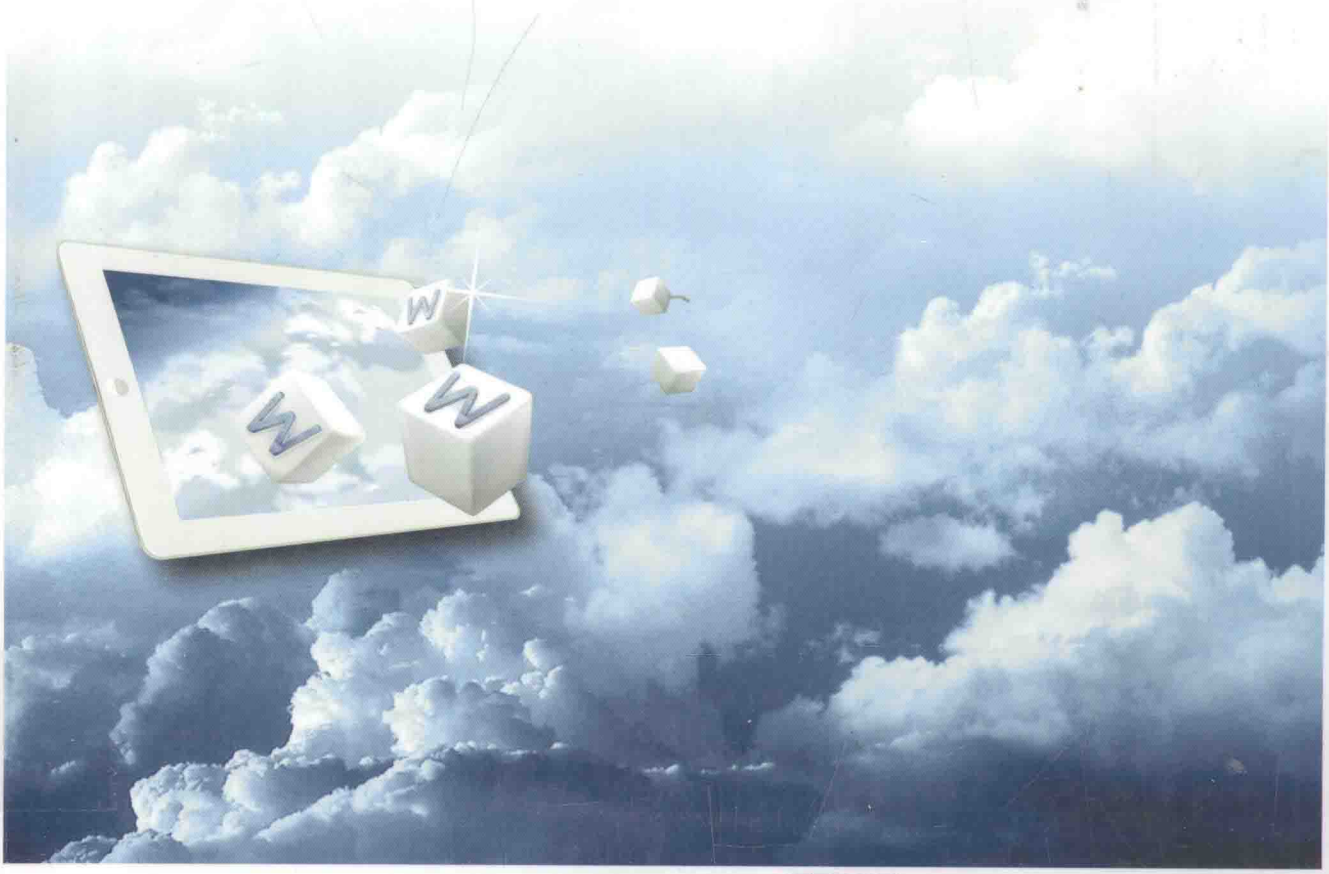
大数据分析3个应用实例：
股票市场预测系统
海量视频检索系统
HDFS云文件系统

大数据分析技术：
Hadoop
HDFS
HBase
MapReduce

实战大数据

怎么做大数据分析 大数据分析怎么应用在业务系统上

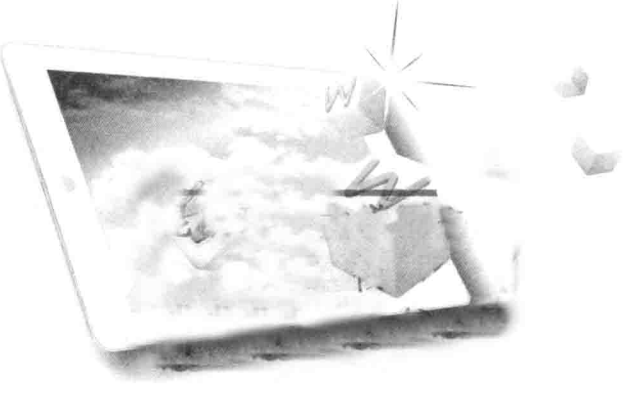
鲍亮 李倩 编著



清华大学出版社

实战大数据

鲍亮 李倩 编著



清华大学出版社
北京

内 容 简 介

“数据是重要资产”已成为大家的共识，众多公司都在争相分析、挖掘大数据背后的信息资源。本书在此背景下，对目前大数据及其相关技术的发展进行总结，理论联系实践，既不缺乏理论深度又具有实用价值。

本书共12章，内容包括大数据的概念、特点、发展历史，数据获取与存储，数据抽取和清洗，数据集成，数据的查询、分析与建模，异构数据采集，文档的存储与检索，异种数据的统一访问与转换，基于微博的股票市场预测系统实例，海量视频检索系统实例，HDFS云文件系统实例。

本书适合大数据技术初学者、大数据从业人员和研究人员，也可以作为高等院校相关专业师生的教学参考书。

图书在版编目（CIP）数据

实战大数据 / 鲍亮，李倩编著. - 北京：清华大学出版社，2014

ISBN 978-7-302-34866-5

I. ①实… II. ①鲍… ②李… III. ①数据处理－研究 IV. ①TP274

中国版本图书馆CIP数据核字（2013）第310963号

责任编辑：夏非彼

封面设计：王　翔

责任校对：闫秀华

责任印制：宋　林

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 装 者：河北新华第一印刷有限责任公司

经　　销：全国新华书店

开　　本：190mm×260mm　　**印　　张**：33.75　　**字　　数**：860千字

版　　次：2014年3月第1版　　**印　　次**：2014年3月第1次印刷

印　　数：1～3500

定　　价：79.00元

产品编号：050201-01

前　言

大数据时代已经到来，大数据处理已经成为当今信息处理的热点研究内容。不同于大规模数据，大数据具有自身鲜明的 4V 特征：Volume（规模性）、Variety（多样性）、Velocity（高速性）和 Veracity（真实性）。大数据不仅规模大，更需要采取新的数据思维来应对，其必然导致理论和技术上的革新。因此，大数据分析也被认为是继实验、理论和计算之后的科学研究第四范式。大数据的出现必将颠覆传统的数据管理方式，在数据来源、数据处理方式和数据思维方面都会对其带来革命性的变化。

2013 年初，美国计算机协会数据库专家委员会联合研究界、产业界和政府部门的相关研究人员，发布了大数据研究白皮书，提出了大数据分析的 5 个重要阶段：数据获取和记录，数据抽取、清洗和注记，数据集成、聚集和表示，数据分析和建模，数据解释。在这 5 个阶段中需要考虑数据的异构性、规模、时效性、复杂性和隐私问题。本书以此为提纲进行内容组织，首先介绍了 5 个阶段中相关的科学与技术问题，然后以实际案例的形式详细介绍了数据采集、数据存储与检索、数据处理、数据访问与转换 4 个大数据领域的重要问题，最后以股票市场预测系统、海量视频检索系统和云文件系统 3 个大数据实际应用系统为例详细介绍如何进行问题分析、数据建模以及系统的设计与实现。本书强调理论联系实际，重点在于介绍如何利用现有技术解决实际的大数据问题。

目前市场上以大数据为主题的书籍较多，但经过作者调研，未见以“利用现有技术解决大数据问题”为主题的大数据实战类书籍。本书编写团队核心成员自 2010 年起陆续承担了一些与大数据采集、存储、处理、分析、挖掘和检索方面的研究与应用开发工作，具有丰富的项目实践经验。这些实际项目经验形成了本书最为核心的第 6~12 章的内容。通过项目实战，我们积累了一些解决大数据问题的宝贵经验，对大数据的核心技术有了较为深刻的理解，认为有必要将自己的经验和认识整理出来，以满足广大读者利用现有技术解决大数据实际问题的迫切需求与心情，这也是书名的由来。

本书适合不同层次的读者阅读，建议读者根据自己的兴趣和目的有选择性地阅读：希望了解大数据相关的基础理论与技术的读者，可以重点阅读第 1~5 章；对于大数据领域的初学者，可以重点阅读第 1~9 章；对于已经掌握大数据基础理论，具有一定的技术基础，想解决实际

大数据问题的读者，可以重点阅读第 10~12 章。

除封面署名的作者之外，参与编写的还有李江、张翔、杨阳、王贺、刘凯、王学良、张静、周文琳、刘晓静、张艳华、王炎楠、黄鹏、高小青。还需要感谢阚传奇、蒋帆的大力帮助，感谢我的导师陈平教授在大数据科学研究方面对我的启发与悉心指导。

由于大数据涉及的学科面很广，研究问题纷繁复杂，相关资料目前还比较少，加之作者水平有限，时间紧迫，书中难免存在错误与不当，恳请读者批评指正。建议和意见请发至作者邮箱 baoliang@mail.xidian.edu.cn。

编者

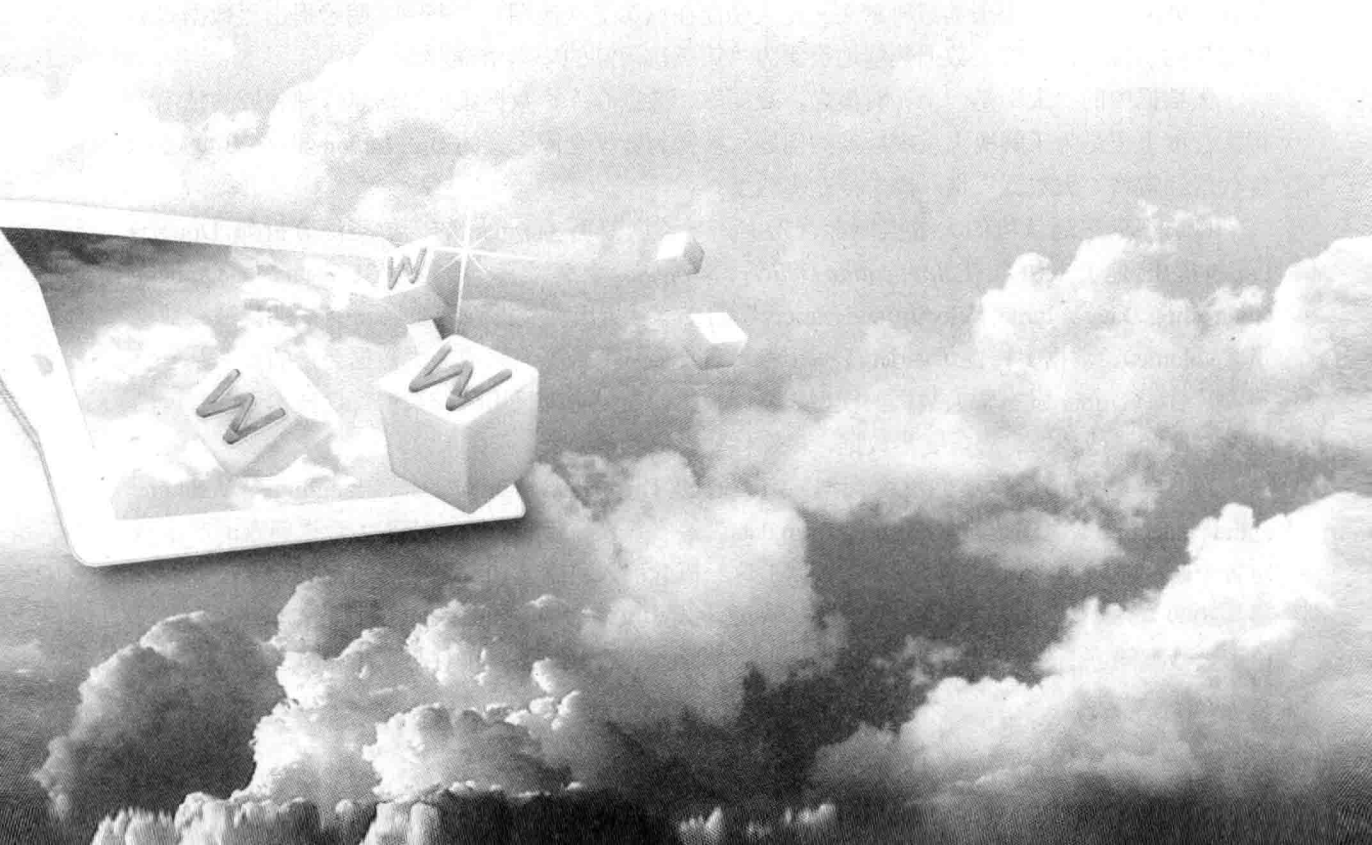
2013 年 12 月

目　录

第一篇　大数据基础篇

第一篇

大数据基础篇



第 1 章
大数据介绍

IT 行业总不乏新鲜的主题，而大数据正当其兴，被业界热情传诵。“数据是重要资产”这一概念已成为大家的共识，众多公司争相分析、挖掘大数据背后的重要资源。为了帮助读者理解大数据的来龙去脉，本章将从大数据的历史与发展、大数据的定义、大数据的研究内容、大数据问题在国内外政府、公司和大学的研究现状等方面进行论述，为这一新兴概念勾勒出一个雏形。

1.1 大数据相关概念

1.1.1 大数据的历史

大数据（Big Data）目前已经成为 IT 领域最为流行的词汇，其实它并不是一个全新的概念。早在 1980 年，著名未来学家阿尔文·托夫勒便在《第三次浪潮》一书中，明确提出“数据就是财富”这一观点，并将大数据热情地赞颂为“第三次浪潮的华彩乐章”。

大数据中的“大”是一个相对概念，数据库、数据仓库、数据集市等信息管理领域的技术，很大程度上也是为了解决大规模数据的问题。被誉为数据仓库之父的 Bill Inmon 早在 20 世纪 90 年代就经常将“大数据”这一概念挂在嘴边了。

目前得到广泛认可的大数据概念首先由知名咨询公司 Gartner 的一位资深分析师 Douglas Laney 提出。他于 2001 年在 *Application Delivery Strategies* 上撰写了一篇名为“3D Data Management: Controlling Data Volume, Velocity, and Variety”的文章，指出大数据管理面临三个 V 的挑战：数据量（Volume）、数据多样性（Variety）、高速（Velocity）。“3V”后来成为大数据公认的三个基本特征。随后，Gartner 发布了大数据的模型，强调大数据需要管理采用传统数据管理技术无法管理的数据，比如微博数据、海量交易数据、多媒体数据，等等。

2008 年 9 月，自然杂志推出《大数据》专刊，通过“The next Google”、“Data wrangling”、“Welcome to the petacentre”、“Distilling meaning from data”等多篇文章，全方位介绍了大数据问题的产生及对各个研究领域的影响，首次将“大数据”这一概念引入科学家和研究人员的视野。

2009 年 8 月，Adam Jacobs 在 *ACM Queue* 上发表文章“The Pathologies of Big Data”，文章讨论了大数据问题的起源、发展与现状，指出“大数据”这一概念是相对的，并提出应该考虑为什

么会出现“大数据”这一现象、“大数据”产生的很大一部分原因是数据录入更加容易等观点。

2011 年 2 月 11 日的《科学》杂志专门推出《数据处理》（*Dealing with Data*）专刊，对大数据现象在科学领域的现状进行了全面分析。该专刊首先联合《科学》杂志的兄弟期刊 *Science Signaling*、*Science Translational Medicine* 和 *Science Careers*，展开了对各科学领域研究数据规模急剧增大情况下各种问题的调研，问题包括“研究数据的规模”、“研究数据如何存储”，等等。随后，该专刊发表多篇文章，对天文学、气象学、生态学、神经科学、信号处理、社会科学、生物学等多个学科的大数据问题进行了解释和阐述，内容涵盖数据采集、分析、处理、挖掘和可视化等多个方面。

2011 年 5 月，麦肯锡全球研究院发表 *Big data: The next frontier for innovation, competition, and productivity* 白皮书，指出企业正在面临海量的交易数据、顾客信息、供货商信息和运营数据等，需要对这些数据进行管理与挖掘。在物联网环境下，传感器、智能手机、工业设备等都在产生海量数据。互联网中的多媒体数据量也在以指数级上升，如何处理这些数据，为用户提供有用的信息，成为需要考虑的重要问题。

2011 年 5 月 26 日，经济学人发表“Building with big data”指出在数据极度膨胀的时代，要掌握数据的分析与处理能力，成为数据的主人，而不要成为数据的奴隶。

2012 年 2 月 11 日，纽约时报发表“The Age of Big Data”，向大众宣传大数据时代的到来。

2012 年 3 月 22 日，奥巴马宣布以 2 亿美元投资大数据领域，在次日的电话会议上，美国政府将数据定义为“未来的新石油”，美国政府认识到了一个国家拥有数据的规模、活性及解释运用的能力将成为综合国力的重要组成部分，未来对数据的占有和控制甚至将成为继陆权、海权、空权之外的另一种国家核心资产。

2012 年 7 月 10 日，联合国在纽约总部发布了一份大数据政务白皮书，总结了各国政府如何利用大数据更好地服务和保护人民。

1.1.2　大数据的定义

1. 维基百科的定义

大数据是指其大小或复杂性无法通过现有常用的软件工具，以合理的成本并在可接受的时限内对其进行捕获、管理和处理的数据集。这些困难包括数据的收入、存储、搜索、共享、分析和可视化。

2. Garnter 的定义

Garnter 咨询公司关注大数据的三个量化指标：数据量、数据种类和处理速度。一般企业所面对的数据管理管理的是数据库、结构化数据，以及所能预先安装好的管理软件所带来的数据。大数据管理的往往是我们无法管理的数据，比如来自企业外部，微博、社交网站和多媒体等各种载体的数据。

数据多样性将是大数据的一个重点。它意味着未来数据的产生将更加方便、快捷，无所不在。

数据种类随着物联网等技术的不断兴起而飞速增加，特别是以多媒体数据为代表的非结构化数据迅速增加，为大数据的分析与处理带来了很大难度。处理速度与企业 CIO 关注的系统性能不是等同的关系。这里的速度指的是从数据产生到最终针对数据产生决策的速度，包括存储的过程、计算的过程、系统模型和以什么方式提交出最后的结果。因此，速度不仅是计算能力和存储性能的问题，还要考虑数据管理、数据保护等方面的响应与处理速度。在大数据问题中，速度往往是性命攸关的。比如对于灾难的预测，当灾难发生时，要很快对灾难发生的程度、影响的区域范围、对长远的影响等量化出来。这是大数据很典型的应用，如果短时间内没有计算出来，那么数据就没用了。

另一方面，Garnter 认为在越来越大的数据集上工作能够得到更大的好处，大数据的数据增长挑战（或机遇）是三维立体的：不断增长的数据量、不断增加的速率（数据 I/O 的速度）和不断增加的种类（数据类型、数据源）。而传统的存储技术难以应对大数据处理的三大挑战。

- 挑战一：不断增长的数据量。在大数据背景下，数据通常是不能删除的，这是企业的宝贵的财富，因此数据将不断积累增长。与此同时，增长有加速的趋势，经常会超出人们预计或规划，从而对信息系统带来了极大的挑战。信息中心需要管理 TB 级甚至 PB 级数据。要为这些数据提供存储、保护和使用的方案，IT 系统需要不断地做相应升级或重构，需要投入大量人力物力。
- 挑战二：多格式数据。海量数据包括了越来越多不同格式的数据，这些不同格式的数据也需要不同的处理方法。从简单的电子邮件、数据日志和信用卡记录，再到仪器收集到的科学研究数据、医疗数据、财务数据以及丰富的媒体数据（包括照片、音乐、视频等），都具有这个特点。比如视频文件格式就非常多，有各软件厂商的厂商标准的格式，工业标准组织的工业标准格式。各种格式在当前高清化的趋势下，数据粒度更小，处理更精细，更复杂的格式还不断出现，造成单一文件的体积成倍增加，从而要求处理速度也成倍增加。
- 挑战三：性能。速度是指数据从客户端到处理器和存储的移动速度，涉及终端数据处理能力、数据流访问和交付、服务器计算处理能力以及后端存储的吞吐能力。速度意味着要求数据必须以多快的频率被处理。大数据处理需要不同于交易类应用的速度，通常其对带宽的要求比 IO 操作的速度更重要。

3. IBM 对大数据的定义

IBM 专门开辟了大数据专栏，从大数据的定义、大数据处理平台等多个方面对大数据问题及解决方案进行了阐述。

Big data spans three dimensions: Volume, Velocity and Variety.

Volume: Enterprises are awash with ever-growing data of all types, easily amassing terabytes—even petabytes—of information.

Velocity: Sometimes 2 minutes is too late. For time-sensitive processes such as catching fraud, big data must be used as it streams into your enterprise in order to maximize its value.

Variety: Big data is any type of data - structured and unstructured data such as text, sensor data, audio, video, click streams, log files and more. New insights are found when analyzing these data types together.