深入浅出 系列规划教材

深入浅出

大数据

宋智军 编著

清华大学出版社

深入浅出 系列规划教材

大数据

宋智军 编著

清华大学出版社
北京

内 容 简 介

本书坚持以大数据基础和应用为主导的编写原则，理论联系实际，并通过大量实例循序渐进地为读者介绍了进行大数据实践所涉及的各类知识。为了更好地帮助读者在短时间内掌握大数据基础理论知识和实践能力，全书的基础知识介绍清晰，理论联系实际，具有很强的操作性，并提供了大量通过测试可运行的完整实例，这些实例都给出了设计步骤、代码详解及程序运行结果，对于容易出现问题的地方，则以“注”的方式介绍常用的技巧和注意事项。另外本书的配套资料可从清华大学出版社网站（www. tup. com. cn）上下载。

本书可作为计算机专业的本科生和研究生的大数据基础教材，也可作为大数据技术培训、Hadoop应用开发和运行维护人员的必备参考书。

图书在版编目（CIP）数据

深入浅出大数据/宋智军编著. --北京：清华大学出版社，2016
深入浅出系列规划教材
ISBN 978-7-302-42181-8

Ⅰ. ①深…　Ⅱ. ①宋…　Ⅲ. ①数据处理－教材　Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2015)第 272725 号

责任编辑：白立军　薛　阳
封面设计：傅瑞学
责任校对：焦丽丽
责任印制：刘海龙

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
课件下载：http://www. tup. com. cn，010-62795954

印 刷 者：三河市君旺印务有限公司
装 订 者：三河市新茂装订有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：24　　**字　　数**：553 千字
版　　次：2016 年 3 月第 1 版　　**印　　次**：2016 年 3 月第1 次印刷
印　　数：1～2000
定　　价：49.00 元

产品编号：062976-01

丛书序

为什么开发深入浅出系列丛书？

目的是从读者角度写书，开发出高质量的、适合阅读的图书。

“不积跬步，无以至千里；不积小流，无以成江海。”知识的学习是一个逐渐积累的过程，只有坚持系统地学习知识，深入浅出，坚持不懈，持之以恒，才能把一类技术学习好。坚持的动力源于所学内容的趣味性和讲法的新颖性。

计算机课程的学习也有一条隐含的主线，那就是“提出问题→分析问题→建立数学模型→建立计算模型→通过各种平台和工具得到最终正确的结果”，培养计算机专业学生的核心能力是“面向问题求解的能力”。由于目前大学计算机本科生培养计划的特点，以及受教学计划和课程设置的原因，计算机科学与技术专业的本科生很难精通掌握一门程序设计语言或者相关课程。各门课程设置比较孤立，培养的学生综合运用各方面的知识能力方面有欠缺。传统的教学模式以传授知识为主要目的，能力培养没有得到充分的重视。很多教材受教学模式的影响，在编写过程中，偏重概念讲解比较多，而忽略了能力培养。为了突出内容的案例性、解惑性、可读性、自学性，本套书努力在以下方面做好工作。

1. 案例性

所举案例突出与本课程的关系，并且能恰当反映当前知识点。例如，在计算机专业中，很多高校都开设了高等数学、线性代数、概率论，不言而喻，这些课程对于计算机专业的学生来说是非常重要的，但就目前对不少高校而言，这些课程都是由数学系的老师讲授，教材也是由数学系的老师编写，由于学科背景不同和看待问题的角度不同，在这些教材中基本都是纯数学方面的案例，作为计算机系的学生来说，学习这样的教材缺少源动力并且比较乏味，究其原因，很多学生不清楚这些课程与计算机专业的关系是什么。基于此，在编写这方面的教材时，可以把计算机上的案例加入其中，例如，可以把计算机图形学中的三维空间物体图像在屏幕上的伸缩变换、平移变换和旋转变换在矩阵运算中进行举例；可以把双机热备份的案例融入到马尔科夫链的讲解；把密码学的案例融入到大数分解中等。

2. 解惑性

很多教材中的知识讲解注重定义的介绍，而忽略因果性、解释性介绍，往往造成知其然而不知其所以然。下面列举两个例子。

(1) 读者可能对 OSI 参考模型与 TCP/IP 参考模型的概念产生混淆，因为两种模型之

间有很多相似之处。其实，OSI 参考模型是在其协议开发之前设计出来的，也就是说，它不是针对某个协议族设计的，因而更具有通用性。而 TCP/IP 模型是在 TCP/IP 协议栈出现后出现的，也就是说，TCP/IP 模型是针对 TCP/IP 协议栈的，并且与 TCP/IP 协议栈非常吻合。但是必须注意，TCP/IP 模型描述其他协议栈并不合适，因为它具有很强的针对性。说到这里读者可能更迷惑了，既然 OSI 参考模型没有在数据通信中占有主导地位，那为什么还花费这么大的篇幅来描述它呢？其实，虽然 OSI 参考模型在协议实现方面存在很多不足，但是，OSI 参考模型在计算机网络的发展过程中起到了非常重要的作用，并且，它对未来计算机网络的标准化、规范化的发展有很重要的指导意义。

(2) 再例如，在介绍原码、反码和补码时，往往只给出其定义和举例表示，而对最后为什么在计算机中采取补码表示数值？浮点数在计算机中是如何表示的？字节类型、短整型、整型、长整型、浮点数的范围是如何确定的？下面我们来回答这些问题(以 8 位数为例)，原码不能直接运算，并且 0 的原码有+0 和-0 两种形式，即 00000000 和 10000000，这样肯定是不行的，如果根据原码计算设计相应的门电路，由于要判断符号位，设计的复杂度会大大增加，不合算；为了解决原码不能直接运算的缺点，人们提出了反码的概念，但是 0 的反码还是有+0 和-0 两种形式，即 00000000 和 11111111，这样是不行的，因为计算机在计算过程中，不能判断遇到 0 是+0 还是-0；而补码解决了 0 表示的唯一性问题，即不会存在+0 和-0，因为+0 是 00000000，它的补码是 00000000，-0 是 10000000，它的反码是 11111111，再加 1 就得到其补码是 100000000，舍去溢出量就是 00000000。知道了计算机中数用补码表示和 0 的唯一性问题后，就可以确定数据类型表示的取值范围了，仍以字节类型为例，一个字节共 8 位，有 00000000～11111111 共 256 种结果，由于 1 位表示符号位，7 位表示数据位，正数的补码好说，其范围从 00000000～01111111，即 0～127；负数的补码为 10000000～11111111，其中，11111111 为-1 的补码，10000001 为-127 的补码，那么到底 10000000 表示什么最合适呢？8 位二进制数中，最小数的补码形式为 10000000；它的数值绝对值应该是各位取反再加 1，即为 01111111+1=10000000=128，又因为是负数，所以是-128，即其取值范围是-128～127。

3. 可读性

图书的内容要深入浅出，使人爱看、易懂。一本书要做到可读性好，必须做到“善用比喻，实例为王”。什么是深入浅出？就是把复杂的事物简单地描述明白。把简单事情复杂化的是哲学家，而把复杂的问题简单化的是科学家。编写教材时要以科学家的眼光去编写，把难懂的定义，要通过图形或者举例进行解释，这样能达到事半功倍的效果。例如，在数据库中，第一范式、第二范式、第三范式、BC 范式的概念非常抽象，很难理解，但是，如果以一个教务系统中的学生表、课程表、教师表之间的关系为例进行讲解，从而引出范式的概念，学生会比较容易接受。再例如，在生物学中，如果纯粹地讲解各个器官的功能会比较乏味，但是如果提出一个问题，如人的体温为什么是 37℃？以此为引子引出各个器官的功能效果要好得多。再例如，在讲解数据结构课程时，由于定义多，表示抽象，这样达不到很好的教学效果，可以考虑在讲解数据结构及其操作时用程序给予实现，让学生看到直接的操作结果，如压栈和出栈操作，可以把 PUSH()和 POP()操作实现，这样效果会好

很多，并且会激发学生的学习兴趣。

4. 自学性

一本书如果适合自学学习，对其语言要求比较高。写作风格不能枯燥无味，让人看一眼就拒人千里之外，而应该是风趣、幽默，重要知识点多举实际应用的案例，说明它们在实际生活中的应用，应该有画龙点睛的说明和知识背景介绍，对其应用需要注意哪些问题等都要有提示等。

一书在手，从第一页开始的起点到最后一页的终点，如何使读者能快乐地阅读下去并获得知识？这是非常重要的问题。在数学上，两点之间的最短距离是直线。但在知识的传播中，使读者感到“阻力最小”的书才是好书。如同自然界中没有直流的河流一样，河水在重力的作用下一定沿着阻力最小的路径向前进。知识的传播与此相同，最有效的传播方式是传播起来损耗最小，阅读起来没有阻力。

是为序。

欢迎老师投稿：bailj@tup.tsinghua.edu.cn。

2014 年 12 月 15 日

前言

随着互联网的快速发展以及云计算、物联网和移动互联网等新一代信息技术的广泛应用，全球数据总量规模呈指数级增长。根据麦肯锡全球研究院(MGI)预测：目前，全球数据总量规模已达到ZB(泽字节)级别；预计到2020年，人类所产生的数据量将超过40ZB(4×10^{13}GB)。因此，新一轮的信息消费热潮已爆发，大数据时代已经悄然来到人们身边。

大数据无疑是当前社会发展的热点，如何借助大数据技术来改变企业信息化建设的现状，如何利用大数据进行商业模式、业务模式及经营模式的创新和变革，是当下迫切需要解决的问题。这就要求我们掌握好大数据技术，为企业信息化建设添砖加瓦。由于大数据技术对个人的综合能力要求较高，而对于初学者来说还不知道如何进入大数据殿堂。本书为读者打开了大数据领域的大门，帮助初学者建立大数据的基本概念、大数据思维、大数据基础技术。在读者掌握了这些基础知识之后，再结合大量的代码实例对各个知识点进行深入浅出的讲解，使读者可以在掌握大数据各项技术的基础上，结合实际应用项目进行大数据实践。全书共分为10章，每章的内容简介如下。

第1章：帮助读者更好地认识和了解大数据发展历程，大数据的基本概念及特征，大数据与传统数据在分析和处理方式上的区别，大数据的价值，大数据安全与隐私保护方面的内容。

第2章：介绍要进行大数据实践所涉及的一些关键技术，如大数据采集与预处理技术、大数据存储与管理技术、大数据分析与挖掘技术、大数据应用与展现技术。

第3章：由于Hadoop已成为企业大数据应用的事实标准，因此本章内容以Hadoop为基础，介绍基于Hadoop的大数据生态系统的发展、架构以及构建过程等内容。

第4～9章：分别讲解Hadoop生态系统中各个组件的基本原理、体系构架，以及具体实例等内容。

第10章：介绍如何结合行业特点，利用开源的大数据相关产品，进行大数据产品的设计、开发、部署等过程，并以互联网和智慧交通行业的大数据应用案例为基础，来为读者说明大数据如何展开行业应用以及大数据在行业应用的价值。

另外，本书最后还添加了几个附录，如Hadoop默认端口及作用、Hadoop 1.0与Hadoop 2.0属性名称变化对比、HBase和Hive默认配置说明。本书配套的附件主要由Demo、Hadoop-2.6.0-API、HadoopWorksapces、Softwares和VMwareHadoop文件夹组成。其中，Demo目录中包括HDFS、HBase、YARN、Mahout等Hadoop组件的代码示例；Hadoop-2.6.0-API目录中的index.html可查看Hadoop所有类及功能说明，帮助读

者进行 Hadoop 二次开发；HadoopWorksapces 目录中的每个文件夹都是一个 Eclipse 项目工程，读者将项目导入到 Eclipse 中查看和运行工程；Softwares 目录中包括 Hadoop 生态系统的基本组件、Eclipse IDE 及 Hadoop 的 Eclipse 插件等，方便读者进行搭建大数据平台及二次开发；VMwareHadoop 目录为 Hadoop 集群中的 Master1. Hadoop、Master2. Hadoop、Slave1. Hadoop、Slave2. Hadoop 和 Slave3. Hadoop 节点，读者可通过 VMware 工具直接导入查看已搭建好的 Hadoop 集群。本书的附件可从清华大学出版社官网进行下载。

本书由宋智军编著，同时感谢中国电子科技集团公司第二十八研究所的夏耘、张春晖、高翔、刘文、常庆龙、许家尧、曹博琦、丁旻等给予的帮助和支持。同时非常感谢清华大学出版社的广大员工，他们为本书的选题策划、编辑加工和出版发行付出了辛勤的劳动。由于编者水平有限，加之时间仓促，书中难免有疏漏和不足之处，恳请专家和广大读者指正。

宋智军

目 录

第1章 大数据概述

大数据是继云计算、物联网之后信息技术产业领域的又一重大技术革新。大数据让人们以一种新的数据处理模式对结构化、半结构化以及非结构化的海量数据进行分析，从而获得更强的决策力和洞察力。本章内容旨在帮助读者更好地认识和了解大数据发展历程，大数据的基本概念及特征，大数据与传统数据在分析和处理方式上的区别，大数据现状及发展趋势等。

1.1 大数据发展历程

大数据的概念并不是突然出现的，而是IT技术发展到一定阶段的必然产物。以下是大数据发展过程中一些具有里程碑意义的事件，以及属于大数据概念进化历程中的一些“第一次”或“新发现”等。

早在2008年9月，国际顶级期刊 *Nature* 就推出了 *Big Data* 专刊[1]，并邀请一些研究人员和企业家预测大数据所带来的革新。同年，计算社区联盟(Computing Community Consortium)发表了报告 *Big-Data Computing: Creating revolutionary breakthroughs in commerce, science, and society*[2]，阐述了在数据驱动的研究背景下，解决大数据问题所需的技术以及在商业、科研和社会领域所面临的一些挑战。

2011年2月，国际顶级期刊 *Science* 推出 *Dealing with Data* 专刊[3]，主要围绕着科学研究中大数据的问题展开讨论，专题中的文章既强调了数据洪流所带来的挑战，也强调了如果人们能够更好地组织和访问数据，那么所能抓住和实现的机遇。从而说明大数据对于科学研究的重要性。全球知名的咨询公司麦肯锡(McKinsey)在同年5月份发布了一份关于大数据的详尽报告 *Big data: The next frontier for innovation, competition, and productivity*[4]，对大数据的影响、关键技术和应用领域等进行了详细的分析。AMD公司在同年6月份主办的IDC白皮书 *Big Data: What It is and Why You Should Care*[5]中，强调了大数据可以用来从海量数据中高效地提取价值，使高速的采集、发现、分析数据成为可能。

2012年1月，达沃斯世界经济论坛特别针对大数据发布了 *Big Data, Big Impact: New Possibilities for International Development* 报告[6]，该报告重点关注了个人产生的移动数据与其他数据的融合与利用，以及在新的数据产生方式下，如何更好地利用数据来产生良好的社会效益。同年3月，美国二十多位数据管理领域的知名专家从专业的研究

角度出发，经过约三个月的深入研讨，撰写并发布了 *Challenges and Opportunities with Big Data* 白皮书[7]，文章阐述了大数据的产生，分析了大数据处理流水线的各个阶段，指出了其中的诸多技术挑战，并提供了重要的解决思路。与此同时，美国政府发布了 *Big Data Research and Development Initiative*[8]，旨在通过推动和改善与大数据相关的收集、组织和分析工具及技术，提升从海量和复杂的数据集中获取知识和洞察分析能力。美国将大数据作为国家级的战略，其在经济社会发展中的重要地位可见一斑。同年 7 月，联合国的创新倡议项目 Global Pulse 发布了 *Big Data for Development: Opportunities & Challenges* 白皮书[9]，指出大数据促社会发展，对于全世界是一个历史性的机遇，可以利用大数据造福人类。

进入 2013 年，大数据已成为热门话题，并在越来越多的领域当中逐渐得到广泛的应用。实力雄厚的传统 IT 企业及互联网巨头已通过对大数据的存储、挖掘分析、大数据治理等方面进入到大数据领域中掘金。大数据咨询服务 Big Data Group 通过评估上百家提供大数据服务的公司绘制了大数据产业生态地图 2013 版，将大数据产业生态划分为三个层次：大数据应用、大数据基础设施和大数据技术，如图 1.1 所示。

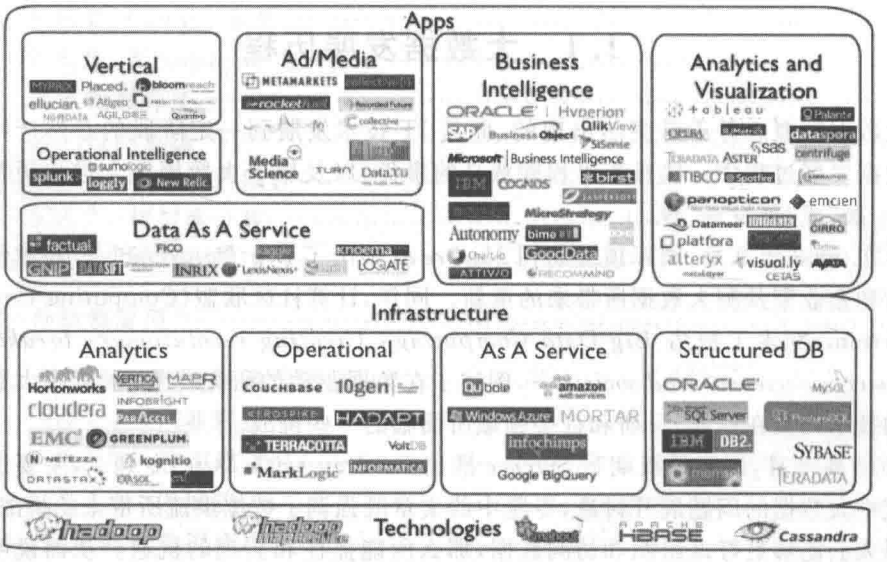

图 1.1 2013 年大数据产业地图

（图片来源：http://www.bigdatalandscape.com/）

1. 大数据应用

目前大数据发展处于初级阶段，大数据应用进程也相对缓慢，主要是互联网领先，其他领域仍在探索中。从图 1.1 中可以看到，一些公司开发出了大数据通用应用，例如大数据可视化和分析工具、大数据商业智能工具或数据服务等。还有一些大数据公司开发出了面向行业用户的垂直应用。未来大数据还将在更多行业得到广泛应用，例如医疗、能源、电信运营、制造业、金融、零售业等。

2. 大数据基础设施

大数据的基础设施不只是简单的物理基础服务器、存储设备等，主要是指大数据平台 PaaS 层的基础设施，如数据采集、存储、数据集成、数据并行处理和数据分析等基础的平台层能力。

3. 大数据技术

大数据技术包括数据采集、数据存取、数据处理、统计分析、数据挖掘、模型预测、结果呈现等技术。目前，Hadoop 已经确立了其作为大数据生态系统基石的地位。Hadoop 框架是由 Java 实现的，它可以对分布式环境下的大数据以一种可靠、高效、可伸缩的方式处理。

到目前为止，大数据市场仍处于初级阶段，也是形成大数据市场竞争格局的关键时期，企业的大数据应用开始从概念验证和实验走向真正的商业化道路，如 IBM、Oracle、EMC、SAP 等国际 IT 巨头提供满足客户需求的大数据解决方案，并推出一体化的集成设备等。

1.2　大数据的定义及特征

1.2.1　大数据定义

大数据是一个涵盖多种技术的概念，它的诞生和发展原动力最初来自于互联网的快速发展。然而，对于大数据的定义至今都没有一个被业界广泛采纳的明确定义，可谓是仁者见仁，智者见智。下面给出一些具有代表性的大数据定义。

(1) 麦肯锡全球研究所在其报告 *Big data: The next frontier for innovation, competition, and productivity* 中给出的大数据定义是：大数据指的是大小超出常规的数据库工具获取、存储、管理和分析能力的数据集[4]。但它同时强调，并不是说一定要超过特定 TB 值的数据集才能算是大数据。

(2) 在维基百科中关于大数据的定义为[10]：大数据指的是所涉及的资料量规模巨大到无法透过目前主流软件工具，在合理时间内达到撷取、管理、处理、并整理成为帮助企业经营决策更积极目的的资讯。笔者认为，这并不是一个精确的定义，因为无法确定主流软件工具的范围，并且可接受时间也是个概略的描述。

(3) 互联网数据中心(IDC)从大数据的 4 个特征来定义[11]，即海量的数据规模(Volume)、数据处理的快速性(Velocity)、多样的数据类型(Variety)、数据价值密度低(Value)，即所谓的“4V”特性。然而，IBM 认为大数据还应该具有其真实性(Veracity)[12]。

(4) 全球性的信息技术研究和顾问公司 Gartner 给出了这样的定义[13]：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。