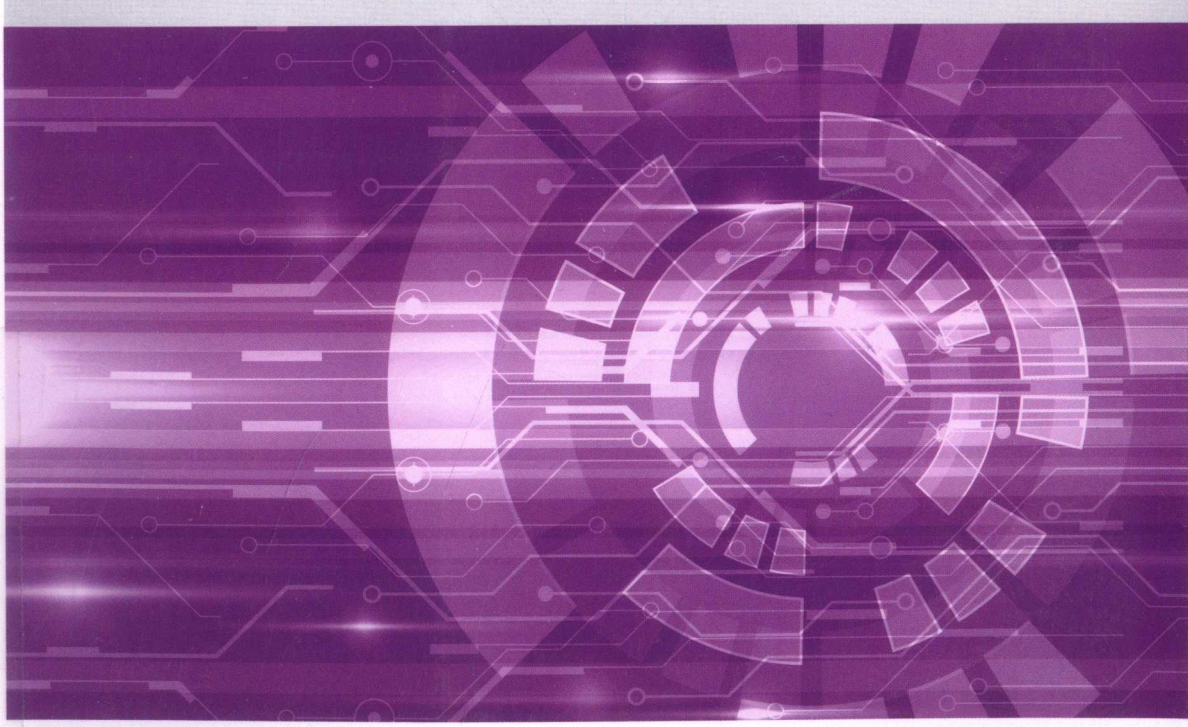
• 大数据应用人才培养系列教材 •

大数据实践

■ 总主编◎刘　鹏　张　燕　■ 主编◎袁晓东　■ 副主编◎黄必栋



清华大学出版社

大数据应用人才培养系列教材

大数据实践

总主编　刘　鹏　张　燕
主　编　袁晓东
副主编　黄必栋

清华大学出版社
北　京

内 容 简 介

本书内容涵盖了目前使用最为广泛的大数据处理系统 Hadoop 生态圈中的几大核心软件系统：分布式大数据处理系统 Hadoop、数据库 HBase、数据仓库工具 Hive、内存大数据计算框架 Spark 和 Spark SQL，详细介绍了它们的架构、工作原理、部署方法、常用配置、常用操作命令、SQL 引擎等内容。本书对上述几大系统的安装部署方式给出了详细步骤，常用命令也都有具体示例介绍，是一本实操性很强的工具书，能帮助初学者快速掌握这几款常用的大数据处理系统。

本书以浅显易懂的语言风格和图文并茂的操作示例引领读者迈入大数据实践之门，可以作为培养应用型人才的课程教材，也可作为相关开发人员的自学教材和参考手册。

图书在版编目（CIP）数据

大数据实践/袁晓东主编. —北京：清华大学出版社，2018
（大数据应用人才培养系列教材）
ISBN 978-7-302-49425-6

Ⅰ. ①大…　Ⅱ. ①袁…　Ⅲ. ①数据处理-技术培训-教材　Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2018）第 015141 号

责任编辑：贾小红
封面设计：刘　超
版式设计：魏　远
责任校对：王　颖
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
印 装 者：北京密云胶印厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：14.75　　**字　　数**：261 千字
版　　次：2018 年 6 月第 1 版　　**印　　次**：2018 年 6 月第 1 次印刷
印　　数：1～2500
定　　价：58.00 元

产品编号：076250-01

编写委员会

总主编　刘　鹏　张　燕

主　编　袁晓东

副主编　黄必栋

参　编　廖若飞　张爱民

总 序

短短几年间，大数据就以一日千里的发展速度，快速实现了从概念到落地，直接带动了相关产业的井喷式发展。数据采集、数据存储、数据挖掘、数据分析等大数据技术在越来越多的行业中得到应用，随之而来的就是大数据人才缺口问题的凸显。根据《人民日报》的报道，未来3~5年，中国需要180万数据人才，但目前只有约30万人，人才缺口达到150万之多。

大数据是一门实践性很强的学科，在其金字塔形的人才资源模型中，数据科学家居于塔尖位置，然而该领域对于经验丰富的数据科学家需求相对有限，反而是对大数据底层设计、数据清洗、数据挖掘及大数据安全等相关人才的需求急剧上升，可以说占据了大数据人才需求的80%以上。比如数据清洗、数据挖掘等相关职位，需要源源不断的大量专业人才。

迫切的人才需求直接催热了相应的大数据应用专业。2018年1月18日，教育部公布了“大数据技术与应用”专业备案和审批结果，已有270所高职院校申报开设“大数据技术与应用”专业，其中共有208所职业院校获批了“大数据技术与应用”专业。随着大数据的深入发展，未来几年申请与获批该专业的职业院校数量仍将持续走高。同时，对于国家教育部正式设立的“数据科学与大数据技术”本科新专业，在已获批的35所大学之外，2017年申请院校也高达263所。

即使如此，就目前而言，在大数据人才培养和大数据课程建设方面，大部分专科院校仍然处于起步阶段，需要探索的问题还有很多。首先，大数据是个新生事物，懂大数据的老师少之又少，院校缺“人”；其次，院校尚未形成完善的大数据人才培养和课程体系，缺乏“机制”；再次，大数据实验需要为每位学生提供集群计算机，院校缺“机器”；最后，院校没有海量数据，开展大数据教学实验工作缺少“原材料”。

对于注重实操的大数据技术与应用专业专科建设而言，需要重点面向网络爬虫、大数据分析、大数据开发、大数据可视化、大数据运维工程师的工作岗位，帮助学生掌握大数据技术与应用专业必备知识，使其具备大数据采集、存储、清洗、分析、开发及系统维护的专业能力和技

能，成为能够服务区域经济的发展型、创新型或复合型技术技能人才。无论是缺“人”、缺“机制”、缺“机器”，还是缺少“原材料”，最终都难以培养出合格的大数据人才。

其实，早在网格计算和云计算兴起时，我国科技工作者就曾遇到过类似的挑战，我有幸参与了这些问题的解决过程。为了解决网格计算问题，我在清华大学读博期间，于 2001 年创办了中国网格信息中转站网站，每天花几个小时收集和分享有价值的资料给学术界，此后我也多次筹办和主持全国性的网格计算学术会议，进行信息传递与知识分享。2002 年，我与其他专家合作的《网格计算》教材正式面世。

2008 年，当云计算开始萌芽之时，我创办了中国云计算网站（chinacloud.cn）（在各大搜索引擎“云计算”关键词中排名第一），2010 年出版了《云计算（第 1 版）》、2011 年出版了《云计算（第 2 版）》、2015 年出版了《云计算（第 3 版）》，每一版都花费了大量成本制作并免费分享对应的几十个教学 PPT。目前，这些 PPT 的下载总量达到了几百万次之多。同时，《云计算》一书也成为国内高校的优秀教材，在中国知网公布的高被引图书名单中，《云计算》在自动化和计算机领域排名全国第一。

除了资料分享，在 2010 年，我们在南京组织了全国高校云计算师资培训班，培养了国内第一批云计算老师，并通过与华为、中兴、360 等知名企业合作，输出云计算技术，培养云计算研发人才。这些工作获得了大家的认可与好评，此后我接连担任了工信部云计算研究中心专家、中国云计算专家委员会云存储组组长、中国大数据应用联盟人工智能专家委员会主任等。

近几年，面对日益突出的大数据发展难题，我们也正在尝试使用此前类似的办法去应对这些挑战。为了解决大数据技术资料缺乏和交流不够通透的问题，我们于 2013 年创办了中国大数据网站（thebigdata.cn），投入大量的人力进行日常维护，该网站目前已经在各大搜索引擎的“大数据”关键词排名中位居第一；为了解决大数据师资匮乏的问题，我们面向全国院校陆续举办多期大数据师资培训班，致力于解决“缺人”的问题。

2016 年年末至今，我们已在南京多次举办全国高校/高职/中职大数据免费培训班，基于《大数据》《大数据实验手册》以及云创大数据提供的大数据实验平台，帮助到场老师们跑通了 Hadoop、Spark 等多个大数据实验，使他们跨过了“从理论到实践，从知道到用过”的门槛。2017

年5月，我们还举办了全国千所高校大数据师资免费讲习班，盛况空前。

其中，为了解决大数据实验难问题而开发的大数据实验平台，正在为越来越多的高校教学科研带来方便，帮助解决“缺机器”与“缺原材料”的问题。2016年，我带领云创大数据（www.cstor.cn，股票代码：835305）的科研人员，应用Docker容器技术，成功开发了BDRack大数据实验一体机，它打破了虚拟化技术的性能瓶颈，可以为每一位参加实验的人员虚拟出Hadoop集群、Spark集群、Storm集群等，自带实验所需数据，并准备了详细的实验手册（包含42个大数据实验）、PPT和实验过程视频，可以开展大数据管理、大数据挖掘等各类实验，并可进行精确营销、信用分析等多种实战演练。

目前，大数据实验平台已经在郑州大学、成都理工大学、金陵科技学院、天津农学院、西京学院、郑州升达经贸管理学院、信阳师范学院、镇江高等职业技术学校等多所院校成功应用，并广受校方好评。同时，该平台以云服务的方式在线提供（大数据实验平台：https://bd.cstor.cn），实验更是增至85个，师生通过自学，可用一个月时间成为大数据实验动手的高手。此外，面对席卷而来的人工智能浪潮，我们团队推出的AIRack人工智能实验平台、DeepRack深度学习一体机以及dServer人工智能服务器等系列应用，一举解决了人工智能实验环境搭建困难、缺乏实验指导与实验数据等问题，目前已经在清华大学、南京大学、南京农业大学、西安科技大学等高校投入使用。

在大数据教学中，本科院校的实践教学应更加系统性，偏向新技术的应用，且对工程实践能力要求更高。而高职高专院校及应用型本科则更偏向于技术和技能训练，理论以够用为主，学生将主要从事数据清洗和运维方面的工作。基于此，我们联合多家高职院校专家准备了《云计算导论》《大数据导论》《数据挖掘基础》《R语言》《数据清洗》《大数据系统运维》《大数据实践》系列教材，帮助解决“机制”欠缺的问题。

此外，我们也将继续在中国大数据（thebigdata.cn）和中国云计算（chinacloud.cn）等网站免费提供配套PPT和其他资料。同时，持续开放大数据实验平台（https://bd.cstor.cn）、免费的物联网大数据托管平台万物云（wanwuyun.com）和环境大数据免费分享平台环境云（envicloud.cn），使资源与数据随手可得，让大数据学习变得更加轻松。

在此，特别感谢我的硕士导师谢希仁教授和博士导师李三立院士。谢希仁教授所著的《计算机网络》已经更新到第7版，与时俱进日臻完美，时时提醒学生要以这样的标准来写书。李三立院士是留苏博士，为

我国计算机事业做出了杰出贡献，曾任国家攀登计划项目首席科学家。他的严谨治学带出了一大批杰出的学生。

本丛书是集体智慧的结晶，在此谨向付出辛勤劳动的各位作者致敬！书中难免会有不当之处，请读者不吝赐教。我的邮箱：gloud@126.com，微信公众号：刘鹏看未来（lpoutlook）。

刘　鹏

于南京大数据研究院

2018 年 5 月

前 言

近年来信息技术迅速发展，互联网、移动、云计算、物联网等技术不断渗透到人们的生活和各行业中，影响和改变着传统的生活方式与工作方式。普及的移动设备、随处部署的物联网设备、互联网后台服务、云计算中心时刻都在产生大量的数据，由此产生了数据的爆炸式增长。企业现在要处理的数据无论是从规模还是从产生速度上都远远超过了以前，传统的数据处理技术已无法适应当前需求。大数据处理技术因此诞生并迅速发展，一方面满足了传统的数据处理需求，另一方面利用大数据技术挖掘出的有价值信息促进了信息技术的应用和发展。

大数据技术最初发展于互联网搜索引擎公司，如 Google、YAHOO! 等，这些公司要检索海量的互联网数据，对大数据处理有着实际的需求。Google 公司于 2003 年发表了分布式文件系统论文，于 2004 年发表了 MapReduce 数据处理框架论文，把 Google 的大数据处理方法和系统公开了。随后基于这两篇论文的开源项目 Hadoop 诞生了，并在 2006 年发布了 0.1.0 版本。YAHOO!公司最初尝试了 Hadoop，在 2006 年部署了 300 台机器的集群，并且逐步扩大集群规模。由于使用 Hadoop 处理大数据非常有效，并且 Hadoop 是开源软件，可以使用普通的机器搭建集群，不少公司开始使用 Hadoop。从 2007 年的 3 家公司到 2008 年的 20 家公司，使用 Hadoop 的公司越来越多，包括 YAHOO!、Facebook、腾讯、阿里巴巴等。其中不少公司还参与到 Hadoop 开源项目中，截至 2011 年，Facebook、LinkedIn、eBay、IBM 集体贡献了 20 万行代码。大公司使用并参与改进 Hadoop，使得 Hadoop 项目迅速发展，功能逐渐丰富，性能不断提高，稳定性得到了增强，逐渐发展为大数据处理的主流工具和框架之一。

在 Hadoop 的应用中人们发现，基于 MapReduce 的数据处理框架存在着性能瓶颈，不适合响应性能要求高的数据处理。而 Hadoop 生态圈中的另一个分布式计算框架 Spark 能够较好地解决这个问题。Spark 诞生于加州大学伯克利分校的 AMP 实验室，最初的目标是进行迭代计算，适用于机器学习等领域（当时 Hadoop 数据处理框架的目标是进行数据批处理），后来发展为既适合数据批处理又适合迭代计算的并行处理框架。Spark 的发展非常迅速，2010 年开源；2013 年贡献给 Apache 基

金会；2014 年成为 Apache 基金会顶级项目，且项目活跃，版本更新快。Spark 和 Hadoop 框架类似，都使用普通机器搭建集群，并且兼容 Hadoop 的分布式文件系统和 HBase 数据库。不同的是，Spark 充分利用了内存资源，并且提供了比 MapReduce 更加灵活和丰富的计算框架。使用 Spark 处理大数据，响应时间更快，编程语言丰富（支持 Java、Scala、Python、R 语言），数据处理效率高。随着 Spark 的不断发展，Spark 自己也形成了庞大的生态圈，包括数据存储、计算框架、结构化数据处理、机器学习、流式处理等重要模块，成为主流的大数据处理工具和框架之一。Spark 并非是 Hadoop 的替代，而是与 Hadoop 取长补短，相互兼容，各自适用于不同需求的数据处理和计算。

本书介绍了目前大数据处理的两套主流框架 Hadoop 和 Spark，包括 Hadoop 分布式文件系统、MapReduce 计算框架、HBase 数据库、Hive 结构化数据处理模块、Spark 计算框架和 Spark SQL 结构化数据处理模块。这些模块都是生态圈中重要的基本模块，模块间存在着依赖关系，如 Hive 中使用到了 MapReduce 计算框架、Spark 计算框架中使用到了 Hadoop 文件系统等。书中按照顺序由浅入深地介绍了各模块的系统原理、部署方法、配置方法、基本操作等内容。本书侧重于实践操作，通过实践学习大数据技术，在使用大数据工具的过程中使读者逐步了解大数据处理的基本概念、方法和步骤，强化实际操作能力，为进一步学习其他大数据技术打下良好的基础。

本书第 1 章和第 2 章由廖若飞编写，第 3 章由袁晓东编写，第 4 章由张爱民编写，第 5 章和第 6 章由黄必栋编写。本书编写过程中得到了刘鹏教授和清华大学出版社王莉、徐瑞鸿编辑的大力支持和悉心指导，在此深表感谢！虽然在完稿前我们反复审查校对，力求做到内容清晰无误、便于学习理解，但疏漏和不完善之处仍在所难免，恳请读者批评指正，不吝赐教。

袁晓东

2018 年 5 月

目 录

第1章

大数据概述

随着社交网络、电子商务、移动互联网等行业的发展，以及云计算、物联网等技术的兴起，数据正以前所未有的速度不断地增长和累积，传统关系数据库的存储能力、处理能力、处理速度、处理效率受到极大的挑战，大数据时代已经来临。工业界、学术界和政府机构都已经开始密切关注大数据领域，并对其产生浓厚的兴趣。市面上关于大数据库的开源和商用系统已经很多；百度学术上近三年来关于大数据的研究文章有 12 万余篇；我国在“十三五”规划（2016—2020 年）中提出：“实施国家大数据战略，推进数据资源开放共享”。作为“‘十三五’十四项大战略”之一的“国家大数据战略”，我国《大数据产业“十三五”发展规划》也正在紧张制定中。“十三五”期间，大数据领域必将迎来建设高峰和投资良机。从全球范围看，大数据主要应用在教育、交通、消费、电力、能源、大健康以及金融等七大重点领域，大数据的应用价值预计在 32 200 亿～53 900 亿美元。

本章先简要介绍了传统关系型数据库的概念和关系型数据库的优点，继而给出了大数据库的定义，分析了大数据库的类型，并结合具体实例介绍了大数据库的应用场景。通过本章的学习，读者可以对大数据库有基本的认识。

1.1 从数据库到大数据库

1.1.1 关系型数据库

传统数据库一般是指关系型数据库，它借助于集合代数等数学概念

和方法来处理数据库中的数据。现实世界中的各种实体以及实体之间的各种联系均用关系模型来表示。现如今业界虽然对此模型有一些批评意见，但它还是数据存储的传统标准。标准数据查询语言 SQL 就是一种基于关系数据库的语言，这种语言执行对应关系型数据库中数据的检索和操作。主流的关系型数据库有 Oracle、SQL Server、MySQL、DB2、SyBase 等。

关系型数据库的优点：

（1）容易理解。关系型数据库使用实体来表示现实世界中的事物，使用属性表示实体的特征，使用二维表来描述逻辑世界的概念，相对于网状、层次等其他模型更容易理解。

（2）使用方便。基本通用的结构化查询语言（SQL）使得关系型数据库的操作十分方便。

（3）易于维护。完整性（实体完整性、参照完整性和用户定义的完整性）支持大大降低了数据冗余和数据不一致的概率。

关系型数据库存在的问题：

（1）难以满足高并发读写需求。网站的用户多，多用户并发操作非常频繁，往往达到每秒上万次读写请求，对于传统关系型数据库来说，磁盘 I/O 是一个很大的瓶颈。

（2）难以满足海量数据的高效率读写需求。网站每天产生的数据量是巨大的，对于关系型数据库来说，在多张包含海量数据的表中关联查询，效率非常低。

（3）扩展性差。在大型应用项目中，数据库是最难进行横向扩展的，当一个应用系统的用户量和访问量与日俱增的时候，数据库很难通过简单增加硬件和服务节点来扩展性能和提高负载能力。对于很多需要提供 24 小时不间断服务的网站来说，对数据库系统进行升级和扩展是非常痛苦的事情，往往需要停机维护和数据迁移。

1.1.2 大数据库

传统处理海量数据（数据仓库）的思路是采用高性能计算机，比如小型机、大型机。如果一台服务器不够用，就把几台服务器连起来，部署分布式数据库，不过这种扩展性也只能达到几台～十几台的级别，扩展性差，成本高。大数据系统放弃磁盘阵列而使用本地硬盘作为存储，通过增加文件副本的方式解决可靠性的问题，存储成本大大降低。分布式计算框架的支持，将计算任务分担到普通的服务器上。从软件层面来解决很多硬件问题，比如单块硬盘故障不影响整个集群的使用、使用普

通服务器搭建集群等。这些新的理念极大地推动了大数据行业的发展。

Hadoop 是大数据系统的典型代表。Hadoop 底层的分布式文件系统具有高拓展性，通过一定数据冗余策略保证数据不丢失并且能提高计算效率，还可以存储各种格式的数据。同时其还支持多种计算框架，既有离线计算，又有在线实时计算，还有内存计算。Hadoop 生态圈中的 Hive 应用的主要场景就是离线分析，HBase 是实时计算的代表，Spark 则是内存大数据计算框架。

大数据是指无法在一定时间内用常规软件工具对其内容进行分析处理的数据集合。大数据技术是指从各种各样类型的数据中，快速获得有价值信息的能力。适用于大数据的技术，包括大规模并行处理数据库、数据挖掘、分布式文件系统、分布式数据库、云计算平台、互联网和可扩展的存储系统等。本书中把以 NoSQL（Not Only SQL）为代表的用于存储、管理、分析海量数据的系统称为大数据库，把大数据库及其依赖的软件环境称为大数据库系统。

具体来说，大数据具有以下四个基本特征：

（1）数据体量巨大。一般指 TB 以及 PB 级的数据。

（2）数据类型多样。比如图片、视频、音频、地理位置信息等。

（3）处理速度快。

（4）价值密度低。

NoSQL 是遵循 CAP 理论和 BASE 原则的典型。CAP 理论可简单描述为：一个分布式系统不能同时满足一致性（Consistency）、可用性（Availability）和分区容错性（Partition Tolerance）这三个需求，最多只能同时满足两个。因此，大部分 key-value 数据库系统都会根据自己的设计目的进行相应的选择。BASE 原则是指 Basically Available（基本可用）、Soft State（软状态）和 Eventually Consistent（最终一致性）。基本可用是指分布式系统在出现不可预知故障的时候，允许损失部分可用性；软状态和硬状态相对，是指允许系统中的数据存在中间状态，并认为该中间状态的存在不会影响系统的整体可用性，即允许系统在不同节点的数据副本之间进行数据同步的过程存在延时；最终一致性强调的是系统中所有的数据副本，在经过一段时间的同步后，最终能够达到一个一致的状态。

在性能上，NoSQL 数据存储系统都具有传统关系数据库所不能满足的特性，是面向应用需求而提出的各具特色的产品。在设计上，它们都关注对数据高并发地读写和对海量数据的存储，并具有很好的灵活性和性能。它们都支持自由的模式定义方式，可实现海量数据的快速访问。灵