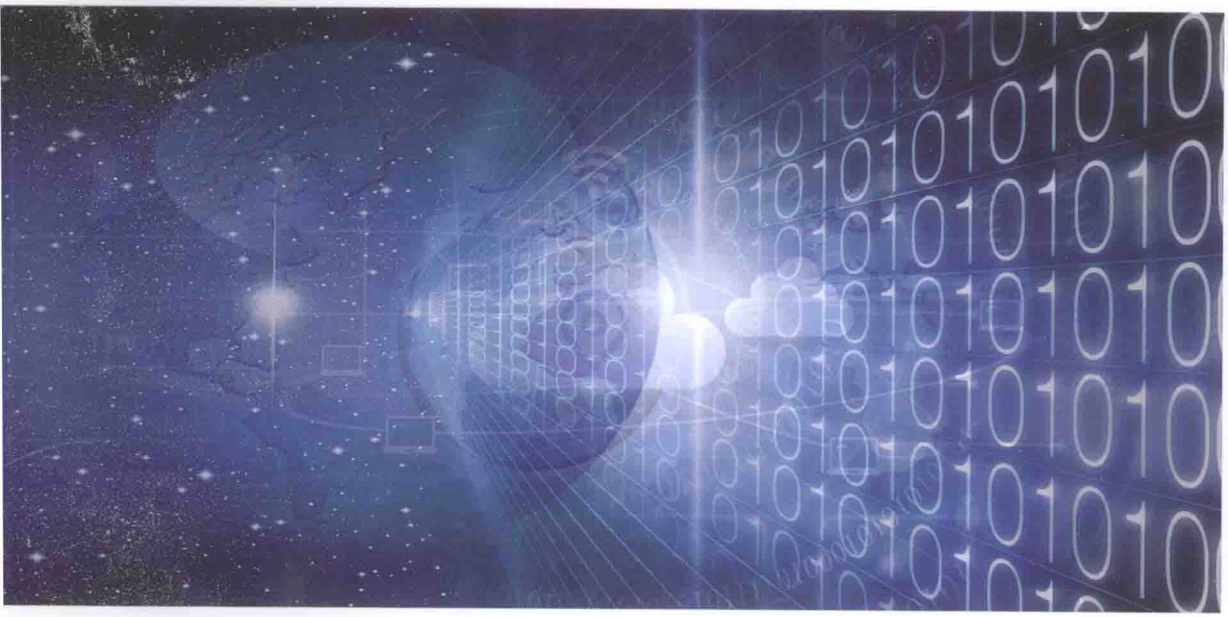
高级大数据人才培养丛书

刘鹏教授主编的权威教材《云计算》和《大数据》的实践动手篇

大数据实验手册

BIG DATA

刘　鹏 ◎ 主编



中国工信出版集团

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

高级大数据人才培养丛书

大数据实验手册

主　编　刘　鹏

副主编　叶晓江　朱光耀　杨震宇

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

内 容 简 介

本书是中国大数据专家委员会刘鹏教授团队的心血之作。针对高校大数据相关专业实践教学以及个人提升大数据动手能力的需求，带领大数据研发团队，经过反复实践、提炼和验证形成本书。本书主要内容包括 HDFS 实验、YARN 实验、MapReduce 实验、Hive 实验、Spark 实验、ZooKeeper 实验、HBase 实验、Storm 实验、MongoDB 实验、LevelDB 实验、Mahout 实验和综合实战等。每个实验呈现详细的实验目的、实验内容、实验原理和实验流程。在线大数据实验平台（https://bd.cstor.cn）或 BDRack 大数据实验一体机可为全部实验提供完整的支撑。

“让学习变得轻松”是本书的初衷。本书填补了大数据教学过程的缺失环节，可培养学生实操动手和自主设计的能力。本书适合作为相关专业本科和研究生的实验手册，也可作为高职高专学校选做的实验教学内容，同时还可作为大数据从业人员的自学书籍。

图书在版编目（CIP）数据

大数据实验手册 / 刘鹏主编. —北京：电子工业出版社，2017.6
（高级大数据人才培养丛书）
ISBN 978-7-121-31618-0

Ⅰ. ①大… Ⅱ. ①刘… Ⅲ. ①数据处理—手册 Ⅳ. ①TP274-62

中国版本图书馆 CIP 数据核字（2017）第 107714 号

策划编辑：董亚峰
责任编辑：董亚峰　　文字编辑：徐　烨
印　　刷：北京京科印刷有限公司
装　　订：北京京科印刷有限公司
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：787×1092　1/16　印张：18.25　字数：427 千字
版　　次：2017 年 6 月第 1 版
印　　次：2017 年 6 月第 1 次印刷
定　　价：45.00 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）88254888，88258888。

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。

本书咨询联系方式：（010）88254694。

编 写 组

主　编：　刘　鹏

副主编：　叶晓江　朱光耀　杨震宇

编　委：　戎新堃　苏泽月　吴荣荣

沈大为　谢　超　方龙双

武郑浩　朱纪光　张　燕

刘　明　孔　坎　石　胡

董广明

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井喷式发展。全球多家研究机构统计数据显示，大数据产业将迎来发展黄金期：IDC 预计，大数据和分析市场将从 2016 年的 1300 亿美元增长到 2020 年的 2030 亿美元以上；中国报告大厅发布的大数据行业报告数据也说明，自 2017 年起,我国大数据产业将迎来发展黄金期，未来 2～3 年的市场规模增长率将保持在 35%左右。

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问题的凸显。麦肯锡预测，每年数据科学专业的应届毕业生将增加 7%，然而仅高质量项目对于专业数据科学家的需求每年就会增加 12%，完全供不应求。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。

以贵州大学为例，其首届大数据专业研究生就业率就达到 100%，可以说“一抢而空”。急切的人才需求直接催热了大数据专业，国家教育部正式设立“数据科学与大数据技术”本科新专业。目前已经有两批共计 35 所大学获批，包括北京大学、中南大学、对外经济贸易大学、中国人民大学、北京邮电大学、复旦大学等。估计 2018 年会有几百所高校获批。

不过，就目前而言，在大数据人才培养和大数据课程建设方面，大部分高校仍然处于起步阶段，需要探索的还有很多。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，尚未形成完善的大数据人才培养和课程体系，院校缺“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，开展大数据教学科研工作缺“原材料”。

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》教材也正式面世。

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（第一版）》、2011 年出版了《云计算（第二版）》、2015 年出版了《云计算（第三版）》，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》教材也成为国内高校的首选教材，在 CNKI 公布的高被引图书名单中，对于 2010 年以来出版的所有图书，《云计算（第一版）》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，我也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等职位。

近几年，面对日益突出的大数据发展难题，我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，我面向全国院校陆续举办多期大数据师资培训班。2016 年末至今，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，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的问题而开发的大数据实验平台，正在为越来越多高校的教学科研带去方便：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、西京学院、郑州升达经贸管理学院、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。该平台也以云服务的方式在线提供（大数据实验平台，https://bd.cstor.cn），帮助师生通过自学，用一个月左右成为大数据动手的高手。

同时，为了解决缺乏权威大数据教材的问题，我所负责的南京大数据研究院，联合金陵科技学院、河南大学、云创大数据、中国地震局等多家单位，历时两年，编著出版了适合本科教学的《大数据》《大数据库》《大数据实验手册》等教材。另外，《数据挖掘》《虚拟化与容器》《大数据可视化》《深度学习》等本科教材也将于近期出版。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、高专院校则更偏向于技术性和技能训练，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，我们还联合多家高职院校专家准备了《云计算基础》《大数据基础》《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，目前也已经陆续进入定稿出版阶段。

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，与时俱进且日臻完美，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　教授
于南京大数据研究院

前 言

教材是体现教学内容和教学方法的知识载体，是教师授课和学生学习的重要参考资料，直接关系到教学质量和人才培养目标的实现，在教学过程中占据十分重要的地位。特别是在大数据教学中，除了理论学习外，实验尤为重要。对于大数据专业毕业生而言，拥有实际操作技能与工作经验俨然成为了其入职薪酬的加分项。以 Hadoop 开发工程师为例，Hadoop 入门月薪可达 8 千元，而具有 2~3 年工作经验的 Hadoop 人才年薪则可达到 30-50 万元。所以，大数据实验与实训直接关系到学生们的职业前景，重要性可见一斑。

然而，对于大数据实验而言，各大高校在开设课程的过程中却遇到了诸多问题。首先，大数据专业处于起步阶段，人才培养课程体系缺乏系统性，大数据教学资源匮乏，可配置和指导实验环境的专业师资不足；其次，教学过程中缺乏相应的实训项目，只有理论教育，难以培养实用型人才，存在专业学习与实际应用脱轨的情况；最后，缺乏相应的基础实验环境，无法为每一个学生都提供一套实验集群。

针对大数据实验课程建设的三大难题，我们的大数据研发团队通过长期的研究，经过反复的验证，推出了《大数据实验手册》这本教材。本教材紧扣应用型人才培养需求，本着“有用、够用、实用”的原则，在某些知识点上做了适当的扩充和提高，在突出重点、有效化解难点方面做了认真考虑和合理安排。教材打破纸上谈兵的传统模式，设计了大量的大数据实验项目，使纸质教材的实际功能辐射到学生实际操作中，引导学生对教材某些内容与观点进行探究。

本教材以实战方式进行编写，一是为了推动大数据人才培养和应用成果转化，使本书成为全国高校首选实验教材；二是为了从社会发展与高校教材发展的关系出发，寻求适应新世纪“创新人才”培养目标的新思路。同时，我们的团队开发了大数据实验平台和大数据实验一体机，可提升高校信息化管理水平和实验项目研究水平，为高校大数据课程提供基础实验环境和实验数据。

本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　教授

于南京大数据研究院

2017 年 6 月 6 日

目 录

实验一　大数据实验一体机基础操作

1.1　实验目的

1．熟悉大数据实验一体机并了解如何搭建集群；
2．熟悉 Linux 基本命令；
3．掌握 vi 编辑器的使用；
4．了解 SSH 免密登录的原理以及为何需要配置 SSH 免密登录；
5．掌握如何配置 SSH 免密登录；
6．掌握 Java 基本命令；
7．熟悉集成开发软件 Eclipse 的安装和使用。

1.2　实验要求

本次实验完成后，要求学生能够：
1．使用大数据实验一体机搭建自己的集群；
2．通过 SSH 工具登录集群服务器；
3．实现每台服务器相互之间的免密登录；
4．通过 vi 编辑器编写 Java 程序；
5．通过 Java 命令编译和运行编写的 Java 程序；
6．通过 jar 命令打包编写的 Java 程序；
7．安装 Eclipse 并在其中编写 Java 程序。

1.3　实验原理

1.3.1　大数据实验一体机

随着移动互联网、云计算、物联网的快速发展，特别是智能手机端博客、社交网络、位置服务（LBS）等信息发布方式的不断涌现，数据正以前所未有的速度不断增长和累积，大数据时代已经来到。

在海量数据面前，大数据人才无疑是其中最关键环节之一。然而，不论国内外，大数据人才却相当稀缺。例如，当前我国数据人才缺口高达 150 万，而在未来 5～10 年，随着市场规模不断增加，这一缺口还将不断加大。