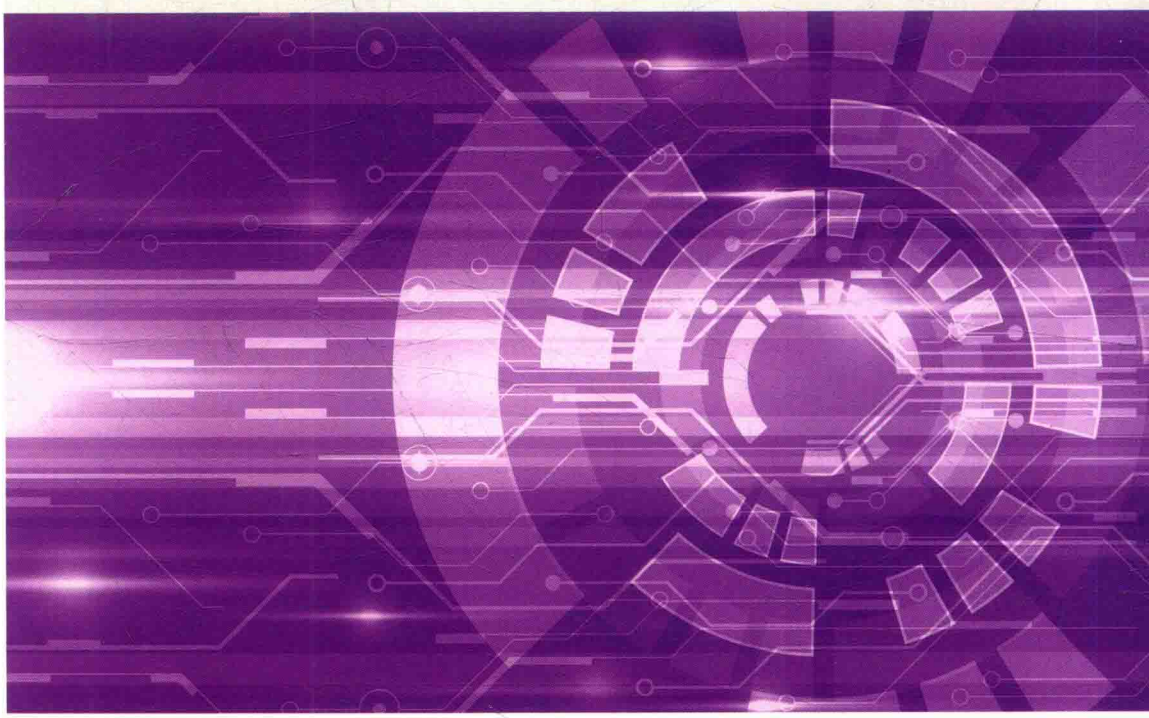
• 大数据应用人才培养系列教材 •

大数据系统运维

■ 总主编◎刘　鹏　张　燕 ■ 主编◎姜才康 ■ 副主编◎陶建辉



清华大学出版社

大数据应用人才培养系列教材

大数据系统运维

总主编　刘　鹏　张　燕
主　编　姜才康
副主编　陶建辉

清华大学出版社
北　京

内 容 简 介

本书是大数据应用人才培养系列教材中的一册，讲解了大数据系统运行维护过程中的各个主要阶段及其任务，包括配置管理、系统管理、故障管理、性能管理、安全管理、高可用性管理、应用变更管理、升级管理及服务资源管理，内容全面且翔实，兼具基础理论知识与运维实践经验，特别是重点介绍了大数据系统的运维特点及运维技能，以保障大数据系统的稳定可靠运行，更好地支撑大数据的商业应用价值。

本书具有很强的系统性和实践指导性，可以作为培养应用型人才的课程教材，也同样适合于有意从事IT系统运维工作的广大从业者和爱好者作为参考书。

图书在版编目（CIP）数据

大数据系统运维/姜才康主编. —北京：清华大学出版社，2018
（大数据应用人才培养系列教材）
ISBN 978-7-302-49326-6

Ⅰ. ①大… Ⅱ. ①姜… 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004244号

责任编辑：贾小红
封面设计：刘　超
版式设计：魏　远
责任校对：马子杰
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：13.5　　字　　数：310千字
版　　次：2018年4月第1版　　印　　次：2018年4月第1次印刷
印　　数：1～2500
定　　价：48.00元

产品编号：075142-01

总主编简介

刘　鹏

教授，清华大学博士毕业。

现任中国大数据技术与应用联盟副理事长、中国大数据应用联盟人工智能专家委员会主任、中国云计算专家咨询委员会专家委员、工业和信息化部云计算研究中心专家。

张　燕

博士，教授，金陵科技学院副校长。

江苏省计算机学会常务理事、江苏省人工智能学会常务理事、江苏省农学会智慧农业分会理事长。先后主持市厅级以上科研项目14项，其中教学改革试验和教学研究项目7项，发表研究论文20多篇，合著专著1部，主编教材5部。

本书主编简介

姜才康

华东计算所硕士毕业，现任中国外汇交易中心工程运行部总经理。长期从事银行间市场（含外汇市场、货币市场、债券市场、衍生品市场）的系统设计开发、系统运维、标准制定等工作。

编写委员会

总主编　刘　鹏　张　燕
主　编　姜才康
副主编　陶建辉
参　编　夏志江　朱　辉　何　玮

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。

大数据是一个实践性很强的学科，在其呈现金字塔型的人才资源模型中，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科人才。

急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“大数据技术与应用”专业备案和审批结果，已有270所高职院校申报开展“大数据技术与应用”专业，其中共有208所职业院校获批“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，除已获批的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发展的（发展型或创新性或复合型）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 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、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。

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，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部署应用，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），实验更是增至 85 个，帮助师生通过自学，用一个月左右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 3 月

前 言

随着信息技术，尤其是互联网技术的迅速发展，各种新技术应用不断渗透到人们的生活中，影响并改变着传统的生活和工作方式。现代社会高度依赖计算机提供的相关服务，人们的一举一动，几乎都在触发计算机的计算，直接或者间接产生大量数据。现今，大数据已广为人知，被认为是信息时代的“新石油”。据不完全统计，大数据量呈现出每两年翻一倍的爆炸性增长态势，隐藏着巨大的机会和价值，并将给社会带来诸多变革和发展，已引起学界、政界以及产业界的广泛关注，各行业已纷纷建立起大数据处理系统，通过对数据的分析和挖据，为经济、社会甚至国防安全等提供帮助。

大数据的“大”包含几个维度：数据量大、种类多、价值密度低和增长速度快等。传统的集中式系统处理方式存在性能不达标、经济成本高等问题，正因为如此，分布式系统成为大数据系统的主流发展方向。谷歌三篇论文（Google File System、MapReduce、Bigtable）的公开发表是大数据技术的一个关键引爆点，开启了使用一般性能的服务器搭建大批量数据处理系统的新趋势。

时至今日，大数据技术的生态圈已经越来越庞大，目前比较流行的应用主要是 Hadoop、Spark 和 Elastic Search，绝大多数的大数据系统是基于这 3 个技术进行开发的，以这些技术为主题的大数据开发书籍也非常普及。但是开发只是系统整个生命周期的一部分，要想系统稳定运行，真正发挥价值，还需要后期的运维管理。根据笔者多年开发和运维的工作经验来看，运维工作也具有很大的挑战性，既要满足业务快速上线，也要保证系统的安全可用，非常强调实践和经验。基于大数据系统的运维工作，还需要考虑其服务器数量多、数据存储量大、开源技术多和新技术稳定性有待提高等特点，一些传统运维工作的服务器管理、备份管理、升级管理和性能调优等工作，都需要针对大数据技术的特点进行相应的改变与调整。

受清华大学出版社之邀，结合大数据系统的特点，笔者从运维　视角进行阐述，编写一本大数据运维的教材，既能弥补这一方面的空白，也是对自己工作的总结，为大数据行业的发展尽自己的一点绵薄之力。

本书从运维工作的分类出发，对每种运维工作都进行了由浅入深的

介绍。配置管理是整个运维工作的基础和核心，没有配置管理，就如同没有地图在复杂的城市道路中行走一样，随时可能迷失方向；同时，在配置管理章节介绍大数据技术的运维管理工具，掌握这些工具能有效地提高工作效率。系统管理、故障管理、变更管理和升级管理是基础性的，也是日常性的运维工作；安全管理、性能管理、服务资源管理和高可用管理则在运维工作中相对比较高阶，也是比较复杂的内容；而且系统运维注重强调标准、流程和制度。本书侧重理论和实践的结合，考虑到以青年学生为主的读者，其对相关概念接触不多，书中在概念阐述上会占有一定篇幅，帮助读者能更好地理解和融会贯通，若学生对书上的一些名词或术语感到比较陌生，则可通过翻阅书后的名词解释进一步理解。本书也安排了专门章节来详细介绍运维的关键技术和工具，希望读者能按照课本内容完成相关实验或者练习，达到学以致用的效果。

本书由姜才康拟定大纲并通稿，其中第 1 章配置管理、第 7 章应用变更管理和第 8 章升级管理由夏志江编写，第 2 章系统管理及日常巡检和第 9 章服务资源管理由姜才康编写，第 3 章故障管理和第 6 章高可用性管理由朱辉编写，第 4 章性能管理由陶建辉编写，第 5 章安全管理由何玮编写。本书在编写过程中受到清华大学出版社的大力支持和刘鹏教授的悉心指导，也得到中国外汇交易中心领导、同事以及其他老师的支持与帮助，在此深表感谢！虽然在完稿前我们反复检查校对，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！

姜才康

于中国外汇交易中心

2018 年 2 月

目 录

第1章

配置管理

配置管理（CM，Configuration Management）是通过技术或行政手段对软件产品及其开发过程和生命周期进行控制、规范的一系列措施。配置管理的目标是记录软件产品的演化过程，确保软件开发者在软件生命周期中各个阶段都能得到精确的产品配置。

随着软件系统的日益复杂化和用户需求、软件更新的频繁化，配置管理逐渐成为软件生命周期中的重要控制过程，在软件开发过程中扮演着越来越重要的角色。一个好的配置管理过程能覆盖软件开发和维护的各个方面，同时对软件开发过程的宏观管理（即项目管理），也有重要的支持作用。良好的配置管理能使软件开发过程有更好的可预测性，使软件系统具有可重复性，使用户和主管部门对软件质量和开发小组有更强的信心。

ITIL 即 IT 基础架构库（Information Technology Infrastructure Library），由英国政府部门 CCTA（Central Computing and Telecommunications Agency）在 20 世纪 80 年代末制定，现由英国商务部 OGC（Office of Government Commerce）负责管理，主要适用于 IT 服务管理（ITSM）。ITIL 为企业的 IT 服务管理实践提供了一个客观、严谨、可量化的标准和规范。在 ITIL 体系中，配置管理作为一项基础流程支撑着其他 4 项流程（事件管理、问题管理、变更管理和发布管理）。配置项作为配置管理中的基本单元，其颗粒度可以根据具体的实践灵活地细化，既有系统级抽象的配置项，也有由具体的软件或者硬件信息构成的配置项单元。由配置管理数据库（CMDB）统一储存配置项以及不同配置项之间的关联关系。配置管理数据库随着变更管理流程的进行而更新配置项信息，结合发布管