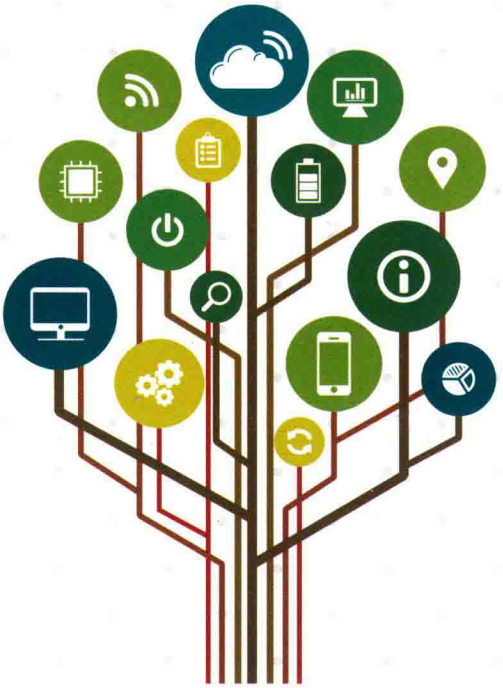
大数据技术与应用专业规划教材
教育部-阿里云产学合作专业综合改革项目规划教材

大数据基础及应用

◎ 吕云翔 钟巧灵 衣志昊 编著

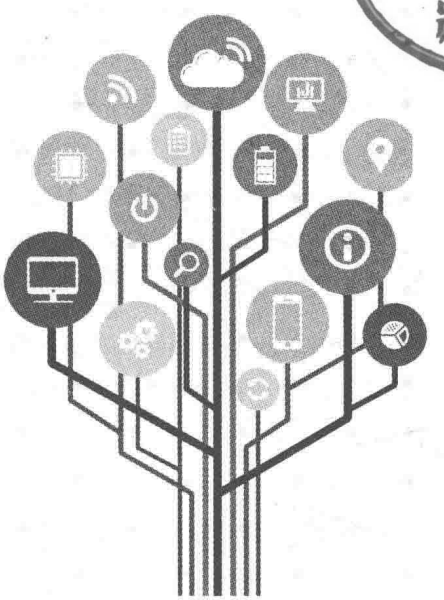


清華大學出版社

大数据技术与应用专业规划教材
教育部-阿里云产学合作专业综合改革项目规划教材

大数据基础及应用

◎ 吕云翔 钟巧灵 衣志昊 编著



清华大学出版社
北京

内 容 简 介

本书从大数据的基本概念开始，由浅入深地领会大数据的精髓。本书除了讲述必要的大数据理论之外，还通过大数据实践来讲述大数据技术的应用，包括如何运用阿里云大数据计算平台分析和解决实际问题，很好地体现了大数据理论与实践的有机结合。

本书分为三大部分，分别是大数据概述及基础、大数据处理和大数据分析与应用。其中，大数据概述及基础部分重点介绍数据组织、重要数据结构、大数据协同技术以及大数据存储技术等内容；大数据处理部分重点介绍大数据处理框架，包括大数据批处理和流处理框架等内容；大数据分析与应用部分重点介绍数据分析技术和机器学习的相关内容，以及如何利用阿里云的数加平台进行基本的大数据开发工作。

本书既可以作为高等院校计算机科学、软件工程及相关专业“大数据”课程的教材，也可以供系统分析师、系统架构师、软件开发工程师和项目经理，以及其他准备或正在学习大数据技术的读者（包括参加计算机等级考试或相关专业自学考试的人员）阅读和参考。

图书在版编目(CIP)数据

大数据基础及应用/吕云翔等编著. —北京：清华大学出版社，2017
（大数据技术与应用专业规划教材）
ISBN 978-7-302-46691-8

Ⅰ. ①大…　Ⅱ. ①吕…　Ⅲ. 数据处理　Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2017)第 038743 号

责任编辑：魏江江　王冰飞
封面设计：刘　键
责任校对：焦丽丽
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
课 件 下 载：http://www.tup.com.cn，010-62795954

印 装 者：清华大学印刷厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：15.75　　**字　　数**：311 千字
版　　次：2017 年 3 月第 1 版　　**印　　次**：2017 年 3 月第1 次印刷
印　　数：1～2000
定　　价：39.50 元

产品编号：072901-01

PREFACE 序言

DT 时代的数据思维与智能思维

本套云计算大数据丛书出版正值信息科技领域进入新一轮巨变，中国经济面临转型机遇的特殊时期。全球信息科技行业伴随着云计算、大数据、物联网、人工智能的发展即将进入一个泛智能的时代，云计算成为数字经济的基础设施；数据驱动、泛在智能成为各行各业转型升级的基础，不仅传统的 IT 从业人员面临能力升级，大多数在校大学生也面临新一轮知识体系的更新，各个垂直行业面临新一轮的人才升级。新一代人才教育与培训，需要一套产学一体的培训课程体系，这是阿里云愿意投身云计算大数据网络安全人才培养体系的时代背景。云计算、大数据、网络安全不仅关乎网络强国的大使命，也逐步成为各行各业专业人才的“元学科”，会逐步成为高等与职业教育的通识课程，一些发达国家已经在中小学立法普及编程课，已经开始指向这个趋势。“懂云计算，有数据思维，理解智能化”，未来可能是每一个工程技术人员与专业人士的必要素质。

2016 年开始，全球信息科技进入一个新的加速爆发周期，可能发生的大概率事件是：二十年之内，有一半的人类知识工作者会被人工智能替代，有服务能力的机器人会诞生，全世界的产业工人会少于机器人；虚拟现实和增强现实会替代今天的智能手机，变成一个新的入口；各行各业都会需要基于物联网的智能化，“中国制造”会成为广泛意义的“中国智造”。

新一轮科技带来了生活方式的变革、生产方式的变革，还有学习方式的变革，这几个趋势的背后，是云计算作为一种普惠科技的基础设施，大数据成为新能源，智能化成为一种新常识。

2016年，全世界的短视频总量增长了6倍，直播业务在中国增长了10倍，远在偏远小镇的青年可以通过直播做电子商务，转化率可以提升十倍以上。当一个技术的使用成本趋近于零的时候，会带来广泛的社会效应。十年以前的直播只有电视台能做，需要专门的摄像机等设备，而今天的直播只需要一个手机，而且是多对多带互动的。无论是短视频，还是直播，背后都有云计算作为普惠科技的支撑作用，由此带来的，所有与知识传播有关的教育，包括整个内容行业，都会被它改变，随着大数据和人工智能的加入，人类学习的方式交互性会更强，“学习系统”会根据不同人的理解程度做个性化的推荐与辅导。

这意味着知识生产与知识传播方式的根本性转变，这个恰恰是云计算、人工智能等科技与各行各业产生化学反应的交叉点，数据是这个转变的新能源。

在2016年10月，阿里云和法院系统合作，发布了一个面向法律服务的智能应用“法小淘”，通过把数千万份法律判例文本化，“法小淘”智能应用可以为普通老百姓以及初级律师提供“打官司”的咨询服务，根据用户输入的案件信息给出建议，包括推荐合适的律师。貌似与科技远离的法律服务也用上了人工智能，这是垂直行业泛智能化的一个小例子。

中国制造进入智能时代

在工业界，阿里云跟中石化合作，协助他们做了企业的电商平台；与徐工合作，推动工厂基于工业云的智能化；与上汽合作，推出具有智能服务的互联网汽车，都收到积极的市场反馈。中国制造，面临智能化的产业机遇，借助互联网人口和产业布局两大优势成为未来的第一个智能产品制造国。

在接下来的几年，互联网+智能制造的叠加会在很多个垂直领域出现，数据智能与制造业结合，产生“跨界重混”的效果，甚至制造业就不是以制造为主，而是以服务化为主。这个巨大的重构背后依赖云和大数据。也因为这个需求，我们可预见工业企业对云计算大数据人才的需求会越来越强烈。

“创业化生存”与共享经济的兴起

创业化，会成为一种常态，越来越多的年轻人开始告别公司，兴起中的数字经济体都是基于云平台的网络化协作组织；云计算成为共享经济的超级容器，催生新一代创业者和“斜杠青年”。十年以后，或许一半以上的从业者都是“斜杠青年”，今天美国就有数千万人是跨工作、跨公司的“斜杠青年”。

过去十年，云计算使得创业公司的创业门槛降低了10倍，没有云计算，Airbnb、NetFlix、推特、Uber等公司不可能这么快成长壮大，新一代创业者的一个核心能力就是要懂技术，理解数据和算法的价值，缺少技术理解力的创业者将面临更大的同质化压力。一句话，无论是草根创业，还是做一个“斜杠青年”，必要的数据思维是生存本能。

创业化和共享经济的崛起，有赖于云计算作为基础设施，大数据作为新能源的全新范式，新一代创业公司需要大量的科技人才。

在未来的经济环境里，普惠云科技的基础设施化、制造的智能化、软件的泛化以及数据无处不在，是一个大趋势，并且不断向各行各业渗透。本套丛书就是希望在这个普惠科技与各行各业深度融合的时代为下一代科技人才的培养提供更多产业界的经验与实践。

感谢清华大学出版社出版本套云计算与大数据方面的系列教材。感谢各位高校老师的辛苦努力和用心付出，使得本系列教材能够付梓出版。

——阿里云业务总经理　刘松

FOREWORD 前言

互联网技术不断发展，各种技术不断涌现，其中大数据技术已成为一颗闪耀的新星。我们已经处于数据世界，互联网每天产生大量的数据，利用好这些数据可以给我们的生活带来巨大的变化以及提供极大的便利。目前大数据技术受到越来越多的机构的重视，因为大数据技术可以给其创造巨大的利润，其中的典型代表是个性化推荐以及大数据精准营销。

本书在讲述大数据的基本概念、原理与方法的基础上，详细而全面地介绍了可以实际用于大数据实践的各种技能，旨在使学生通过有限课时的学习后，不仅能对大数据技术的基本原理有所认识，而且能够具备基本的大数据技术开发能力以及运用大数据技术解决基本的数据分析问题，理解大数据框架（尤其是阿里云大数据计算平台），在阿里云大数据平台上进行基本的大数据开发工作的能力。

本书分为三大部分，分别是大数据概述及基础、大数据处理和大数据分析与应用。其中，大数据概述及基础部分重点介绍数据组织、重要数据结构、大数据协同技术以及大数据存储技术等内容；大数据处理部分重点介绍大数据处理框架，包括大数据批处理和流处理框架等内容；大数据分析与应用部分重点介绍数据分析技术和机器学习的相关内容，以及如何利用阿里云的数加平台进行基本的大数据开发工作。

本书与其他类似著作的不同之处在于，除了讲述必要的大数据理论之外，还通过大数据实践来讲述大数据技术的应用，包括如何运用阿里云大数据计算平台解决和分析实际的问题，如阿里云 MaxCompute 和 StreamCompute 等。本书的最后一章“大数据实践：基于数加平台的推荐系统”是学生在做课程设计时可供模仿的一个项目，它完整地体现了理论与实践的有机结合。

本书的理论知识的教学安排建议如下。

章　节	内　　容	学　时　数
第 1 章	大数据概念和发展背景	1
第 2 章	大数据系统架构概述	1～2
第 3 章	分布式通信与协同	2～4
第 4 章	大数据存储	4～6
第 5 章	分布式处理	2
第 6 章	Hadoop MapReduce 解析	2～4
第 7 章	Spark 解析	2～4
第 8 章	流计算	2
第 9 章	图计算	2
第 10 章	阿里云大数据计算服务平台	2
第 11 章	集群资源管理与调度	4～6
第 12 章	数据分析	2～4
第 13 章	数据挖掘与机器学习技术	2～4
第 14 章	大数据实践：基于数加平台的推荐系统	4～5

建议理论教学时数：32～48 学时。

建议实验(实践)教学时数：16～32 学时。

教师可以按照自己对大数据的理解适当地删除一些章节，也可以根据教学目标，灵活地调整章节的顺序，增减各章的学时数。

在本书成书的过程中，得到了万昭祎、李旭、苏俊洋以及阿里巴巴的李妹芳等人的大力支持，在此表示衷心的感谢。

由于大数据是一门新兴学科，大数据的教学方法本身还在探索之中，加之我们的水平和能力有限，本书难免有疏漏之处。恳请各位同仁和广大读者给予批评指正，也希望各位能将实践过程中的经验和心得与我们交流(yunxianglu@hotmail.com)。

作　者

2017 年 1 月

目录

CONTENTS

第一部分　大数据概述及基础

第二部分 大数据处理

第三部分 大数据分析与应用

第一部分

大数据概述及基础