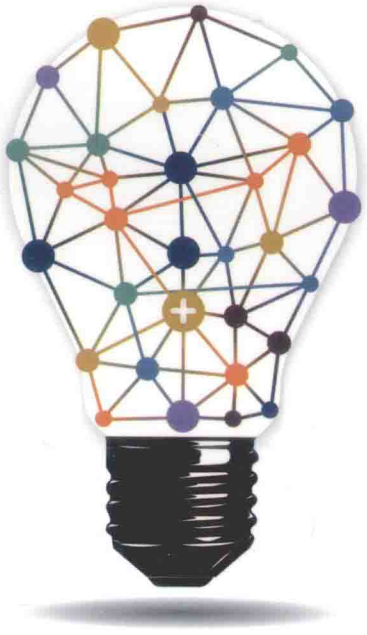
- 本书以理论与实践相结合的方式讲解了大数据技术基础
- 通过三个经典案例详细讲述企业级MapReduce数据处理
- 所有案例均来源于企业，贴近实战，内容充实，拒绝空洞

大数据技术基础

中科普开◎编著

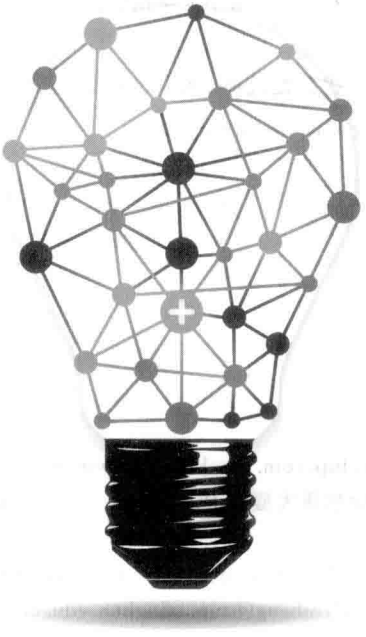


清华大学出版社

本书循序渐进地讲解了Hadoop应用技能
为以后的大数据技术进阶提升夯实基础

大数据技术基础

中科普开⊙编著



清华大学出版社
北京

内 容 简 介

本书的知识架构是在培训了多届学员的基础上总结整理得来的，已经经过了实践的考验，证实了其科学性；本书当中的案例都为企业实际开发的案例，通过学习这些大量的实际案例，帮助学生在进入企业后可以很快融入大数据工作岗位。

本书包括大数据概论、初识 Hadoop、认识 HDFS、HDFS 的运行机制、访问 HDFS、Hadoop I/O 详解、认识 MapReduce 编程模型、MapReduce 应用编程开发、MapReduce 的工作机制与 YARN 平台、MapReduce 高级开发、MapReduce 实例共 11 章内容。

本书既可作为高等院校学习大数据技术的教材，亦可作为广大大数据技术学习者的入门用书。

图书在版编目(CIP)数据

大数据技术基础/中科普开编著. --北京：清华大学出版社，2016
ISBN 978-7-302-43757-4

Ⅰ. ①大…　Ⅱ. ①中…　Ⅲ. ①数据处理　Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2016)第 092661 号

责任编辑：刘翰鹏
封面设计：傅瑞学
责任校对：李　梅
责任印制：沈　露

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，010-62770175-4278
印 装 者：北京嘉实印刷有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　印　　张：16.5　　字　　数：396 千字
版　　次：2016 年 6 月第 1 版　　印　　次：2016 年 6 月第 1 次印刷
印　　数：1～2000
定　　价：39.00 元

产品编号：068768-01

编 委 会

（排名不分先后，按姓氏汉语拼音排序）

吴　钊　湖北文理学院数学与计算机学院院长
夏　英　重庆邮电大学计算机科学与技术学院副院长
向　磊　北京龙诚健康大数据科技有限公司技术总监
杨国为　南京审计大学工学院院长
杨　力　中科普开(北京)科技有限公司大数据教学总监
叶　刚　中科普开(北京)科技有限公司 CEO
叶曲炜　哈尔滨广厦学院信息学院院长
于建江　盐城师范学院信息科学与技术学院院长
张涵诚　百分点科技产品市场总监
张华平　北京理工大学副教授
朱扬清　佛山科学技术学院电子与信息工程学院教授

为什么要写这本书

近年来，大数据(big data)一词越来越多地被提及，人们用它来描述和定义信息爆炸时代产生的海量数据，并命名与之相关的技术发展与创新。它已经上过《纽约时报》、《华尔街日报》的专栏封面，进入美国白宫官网的新闻，现身在国内一些互联网主题的讲座沙龙中，甚至被嗅觉灵敏的国金证券、国泰君安、银河证券等写进了投资推荐报告。最早提出“大数据”时代到来的是全球知名咨询公司麦肯锡。麦肯锡称：“数据，已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素。人们对于海量数据的挖掘和运用，预示着新一波生产率增长和消费者盈余浪潮的到来。”“大数据”在物理学、生物学、环境生态学等领域以及军事、金融、通信等行业存在已有时日，却因为近年来互联网和信息行业的发展而引起人们关注。数据正在迅速膨胀并变大，它决定着企业的未来发展，虽然很多企业可能还没有意识到数据爆炸性增长带来问题的隐患，但是随着时间的推移，人们将越来越多地意识到数据对企业的重要性。

在如今的社会，大数据的应用越来越彰显它的优势，它占领的领域也越来越大，如电子商务、O2O、物流配送等，各种利用大数据进行发展的领域正在协助企业不断地发展新业务和创新运营模式。有了大数据这个概念，对于消费者行为的判断，产品销售量的预测，精确的营销范围以及存货的补给已经得到全面的改善与优化。然而，这些数据的规模是如此庞大，以至于不能用 G 或 T 来衡量。

为了解决这些数据的存储和相关计算问题，就必须构建一个强大且稳定的分布式集群系统作为搜索引擎的基础架构支撑平台，但是对于大多数互联网公司而言，研发这样一个高效性能系统往往要支付高昂的费用。经过多年的发展，如今已形成了以 Hadoop 为核心的大数据生态系统，开创了通用海量数据处理基础架构平台的先河。Hadoop 是一个优秀的分布式计算系统，利用通用的硬件就可以构建一个强大、稳定、简单并且高效的分布式集群计算系统，完全可以满足互联网公司基础架构平台的需求，付出相对低廉的代价就可以轻松处理超大规模的数据。因此，使用 Hadoop 的公司越来越多，具有丰富工作经验的 Hadoop 人才也就越来越供不应求，从而学习和使用 Hadoop 的爱好者和开发者也越来越多，编写这本书也正是为了帮助更多的人学习并掌握 Hadoop 技术，从而推动 Hadoop 技术在中国的推广，进而推动中国信息产业的发展。

读者对象

本书适合以下读者阅读：

(1) 大数据技术的学习者和爱好者；

(2) 有 Java 基础的开发者；

(3) Hadoop 技术开发者；

(4) Hadoop 集群运维开发者;

(5) 分布式系统的相关研发人员。

如何阅读本书

本书分为三个部分。

第一部分为简介。简介部分为第 1 章,主要介绍了大数据的时代背景,从大数据来源到大数据的价值和影响,以及对应用场景和发展前景的介绍,帮助用户明白什么是大数据,大数据是用来干什么的,以及大数据的发展前景是怎样的。大数据的基本概念,首先明白什么是大数据,大数据中数据结构的复杂度,重点明白大数据的四个核心特征,接着了解大数据所使用的技术,最后介绍了一些大数据的应用实例,帮助大家更好地理解大数据、大数据系统,理解其核心设计目标,在系统设计目标的实现过程中,系统还需遵循一定的设计原则。

第二部分为 Hadoop 技术的讲解,包括第 2 章到第 9 章。从认识 Hadoop 开始到正式介绍 Hadoop 的基本应用,通过 HDFS 分布式文件系统和 MapReduce 并行计算模型从理论到实现机制的角度对 Hadoop 计算进行讲解。讲述了 HDFS 的特性和目标、核心设计、体系结构以及 HDFS 中数据流的读写、HA 机制和 Federation 机制,同时重点介绍了 HDFS 的命令行接口和 Java 接口。接着介绍了 Hadoop I/O,讲述了数据的完整性、文件压缩、问价序列化和 Hadoop 文件的数据结构。最后是对 MapReduce 的讲解,由浅入深,讲述了 MapReduce 的编程模型,MapReduce 应用编程开发,包括 MapReduce 的类型格式,Java API 解析,还重点讲述了 MapReduce 的工作机制与 YARN 平台,包括 MapReduce 作业运行机制的剖析、shuffle 和排序、任务的执行、作业调度、YARN 平台的简介和架构。

第三部分为实战部分,包括第 10 章和第 11 章。首先是从几个具体的小实例讲解了简单高效的 MapReduce 编程方式。然后通过最后的 MapReduce 编程实例,带我们进入大数据实战项目,帮助学习者更深入地掌握 Hadoop 技术。

勘误和支持

除本书编委会以外,参加本书编写的工作人员有:毛妍、白高平、赵真。由于本书编写者水平有限,书中难免会出现一些错误或者不准确的地方,恳请读者批评指正,可以将书中遇到的错误和问题发邮件到 mayh@zkpk.org,希望您能提出更多宝贵的意见,期待您的真挚反馈。

中科普开

2016 年 3 月

第1章

大数据概论

本章提要

在这个日新月异发展的社会中，人们发现未知领域的规律主要依赖抽样数据、局部数据和片面数据，甚至无法获得真实数据时只能纯粹依赖经验、理论、假设和价值观去认识世界。因此，人们对世界的认识往往是表面的、肤浅的、简单的、扭曲的或者是无知的。然而大数据时代的来临使人类拥有更多的机会和条件在各个领域更深入地获得和使用全面数据、完整数据和系统数据，深入探索现实世界的规律。大数据的出现帮助商家了解用户、锁定资源、规划生产、做好运营及开展服务。

本章主要从大数据时代背景、大数据基本概念、大数据系统以及大数据与企业等方面，让读者对大数据有初步的认识。

1.1 大数据时代背景

中国庞大的人数和应用市场，其复杂性高并且充满变化，从而成为世界上拥有最复杂的大数据的国家。解决这种由大规模数据引发的问题，探索以大数据为基础的解决方案，是中国产业升级、效率提高的重要手段。因此，解决大数据这一问题不仅提高公司的竞争力，也能提高国家竞争力。

1.1.1 大数据的数据源

近年来，随着信息技术的发展，我国在各个领域产生了海量数据，主要分布如下。

1. 以 BAT 为代表的互联网公司

(1) 阿里巴巴：目前保存的数据量为近百个拍字节(PB)，90%以上是电商数据、交易数据、用户浏览和点击网页数据、购物数据。

(2) 百度：2013 年的数据总量接近一千个拍字节(PB)，主要来自中文网、百度推广、百度日志、UGC，由于占有 70%以上的搜索市场份额从而坐拥庞大的搜索数据。

(3) 腾讯：存储数据经压缩处理后总量在 100PB 左右，数据量月增 10%，主要是大量社交、游戏等领域积累的文本、音频、视频和关系类数据。

2. 电信、金融与保险、电力与石化系统

(1) 电信：包括用户上网记录、通话、信息、地理位置等。运营商拥有的数据量都在 10PB 以上，年度用户数据增长数十拍字节(PB)。

(2) 金融与保险：包括开户信息数据、银行网点和在线交易数据、自身运营的数据等。金融系统每年产生数据达数十拍字节(PB)，保险系统数据量也接近拍字节(PB)级别。

(3) 电力与石化：仅国家电网采集获得的数据总量就达到10个拍字节(PB)级别，石化行业、智能水表等每年产生和保存下来的数据量也达到数十拍字节(PB)级别。

3. 公共安全、医疗、交通领域

(1) 公共安全：在北京，就有50万个监控摄像头，每天采集视频数量约3PB，整个视频监控每年保存下来的数据在数百拍字节(PB)以上。

(2) 医疗卫生：据了解，整个医疗卫生行业一年能够保存下来的数据就可达到数百PB。

(3) 交通：航班往返一次就能产生太字节(TB)级别的海量数据；列车、水陆路运输产生的各种视频、文本类数据，每年保存下来的也达到数十拍字节(PB)。

4. 气象与地理、政务与教育等领域

(1) 气象与地理：中国幅页辽阔，气象局保存的数据为4～5PB，每年约增数百个太字节(TB)，各种地图和地理位置信息每年约增数十太字节(PB)。

(2) 政务与教育：北京市政务数据资源网涵盖旅游、教育、交通、医疗等门类，一年上线公布400余个数据包。政务数据多为结构化数据。

5. 其他行业

线下商业销售、农林牧渔业、线下餐饮、食品、科研、物流运输等行业数据量还处于积累期，整个体积都不算大，多则达到拍字节(PB)级别，少则几百太字节(TB)，甚至只有数十太字节(TB)级别，但增速很快。

1.1.2 大数据的价值和影响

数量巨大、与微观情境相结合的运行记录信息的最终结果就是大数据。尽管运行记录信息不是大数据的全部，但却应该是以后大数据的主流。目前看得到的金融、电信、航空、电商、零售渠道等领域中的大数据，多数也都是运行记录信息。大数据具有采集过程价值未知、力争全面、即时、系统性并发的记录方式，以及主受体统一和大微观的特征，这些特征决定了大数据的价值发挥。

大数据的应用很广泛，解决了大量的日常问题。大数据是利害攸关的，它将重塑人们的生活、工作和思维方式，比其他划时代创新引起的社会信息范围和规模急剧扩大所带来的影响更大。大数据需要人们重新讨论决策、命运和正义的性质。人们的世界观正受到大数据优势的挑战，拥有大数据不但意味着掌握过去，更意味着能够预测未来。因此，大数据给人们带来了巨大的价值和影响。

(1) 全面洞察客户信息。全面分析来自渠道的反馈、社会传媒等多源信息，让每个客户作为个体了解全景。

(2) 提升企业的资源管理：利用实时数据实现预测性维护，并减少故障，推动产品和服务开发。

(3) 数据深度利用。梳理结构化、非结构化、海量历史/实时、地理信息4类数据资源，以企业核心业务及应用为主线实现四类数据资源的关联利用。

(4) 风险及时感知和控制。通过全面数据分析改进风险模型，结合交易流数据实时捕获风险，及时有效地控制。

(5) 辅助智能决策。实时分析所有的运营数据和效果反馈,优化运营流程。利用投资回报率最大程度减少信息技术成本。

(6) 更快和更大规模的产品创新。多源捕获市场反馈,利用海量市场数据和研究数据来快速驱动创新。

1.1.3 大数据技术应用场景

当前,大数据技术的应用涉及各个行业领域。

1. 大数据在金融行业的应用

近年来,随着"互联网金融"概念的兴起,催生了一大批金融、类金融机构转型或布局的服务需求,相关产业服务应运而生。而随着互联网金融向纵深发展,行业竞争日趋白热化,金融、类金融机构在其中的短板日益凸显。为了更好地获得最佳商机,金融行业也步入了大数据时代。

华尔街某公司通过分析全球 3.4 亿微博账户留言来判断民众情绪。人们高兴的时候会买股票,而焦虑的时候会抛售股票,它通过判断全世界高兴的人多还是焦虑的人多来决定公司股票的买入还是卖出。

阿里公司根据在淘宝网上中小企业的交易状况筛选出财务健康和诚信经营的企业,给他们提供贷款,并且不需要这些中小企业的担保。目前阿里公司已放贷款上千亿元,坏账率仅为 0.3%。

2. 大数据在政府的应用

为充分运用大数据的先进理念、技术和资源,加强对我国各地市场主体的服务和监管,推进简政放权和政府职能转变,提高政府治理能力,我国一些省市运用大数据加强对市场主体服务和监管实施方案已然出炉。

3. 大数据在医疗健康的应用

随着医疗卫生信息化建设进程的不断加快,医疗数据的类型和规模也在以前所未有的速度迅猛增长,甚至产生了无法利用目前主流软件工具的现象,这些医疗数据能帮助医改在合理的时间内达到撷取、管理信息并整合成为能够帮助医院进行更积极的经营决策的有用信息。这些具有特殊性、复杂性的庞大的医疗大数据,仅靠个人甚至个别机构来进行搜索,那基本是不可能完成的。

4. 大数据在宏观经济管理领域的应用

IBM 日本分公司建立了一个经济指标预测系统,它从互联网新闻中搜索出能影响制造业的 480 项经济数据,再利用这些数据进行预测,准确度相当高。

印第安纳大学学者利用 Google 提供的心情分析工具,根据用户近千万条短信、微博留言预测琼斯工业指数,准确率高达 87%。

淘宝网建立了"淘宝 CPI",通过采集、编制淘宝网上 390 个类目的热门商品价格来统计 CPI,预测某个时间段的经济走势比国家统计局的 CPI 还提前半个月。

5. 大数据在农业领域的应用

由 Google 前雇员创办 Climate 公司,从美国气象局等数据库中获得几十年的天气数据,各地的降雨、气温和土壤状况及历年农作物产量做成紧凑的图表,从而能够预测美国任一农场下一年的产量。农场主可以去该公司咨询明年种什么能卖出去、能赚钱,说错了该公