国家软件与集成电路公共服务平台信息技术紧缺人才培养工程指定教材

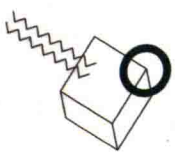
大数据技术与应用丛书

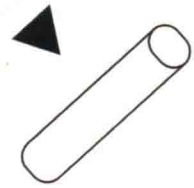
有问题，就找黑马程序员问答精灵！

Hadoop

大数据技术原理与应用

黑马程序员 / 编著





清华大学出版社

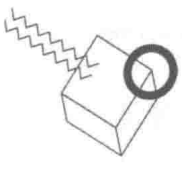
NITE 国家软件与集成电路公共服务平台信息技术紧缺人才培养工程指定教材

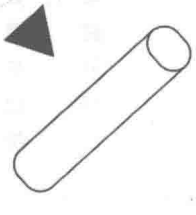
大数据技术与应用丛书

Hadoop

大数据技术原理与应用

黑马程序员 / 编著





清华大学出版社
北 京

内 容 简 介

本书围绕 Hadoop 生态圈相关系统介绍大数据处理架构。全书共 11 章，其中，第 1、2 章主要带领大家认识 Hadoop 以及学会搭建 Hadoop 集群；第 3～5 章讲解分布式文件系统（HDFS）、分布式计算框架 MapReduce 以及分布式协调服务；第 6 章讲解 Hadoop 2.0 新特性，包含 YARN 和高可用特性；第 7～10 章主要讲解 Hadoop 生态圈的相关辅助系统，包括 Hive、Flume、Azkaban 和 Sqoop；第 11 章是一个综合项目——网站流量日志数据分析系统，目的是教会大家如何利用 Hadoop 生态圈技术构建大数据系统架构并进行开发，同时加深对 Hadoop 技术的理解。

本书附有配套视频、源代码、习题、教学设计、教学课件等资源。同时，为了帮助初学者更好地学习本书中的内容，还提供了在线答疑，欢迎读者关注。

本书可作为高等院校本、专科计算机相关专业，信息管理等相关专业的大数据课程教材，也可供相关技术人员参考，是一本适合广大计算机编程爱好者的优秀读物。

图书在版编目（CIP）数据

Hadoop 大数据技术原理与应用/黑马程序员编著. —北京：清华大学出版社，2019
（大数据技术与应用丛书）
ISBN 978-7-302-52440-3

Ⅰ. ①H…　Ⅱ. ①黑…　Ⅲ. ①数据处理软件　Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2019）第 040842 号

责任编辑：袁勤勇　杨　枫
封面设计：韩　冬
责任校对：李建庄
责任印制：丛怀宇

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 刷 者：北京富博印刷有限公司
装 订 者：北京市密云县京文制本装订厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：19　　**字　　数**：464 千字
版　　次：2019 年 5 月第 1 版　　**印　　次**：2019 年 5 月第1次印刷
定　　价：39.00 元

产品编号：083124-01

序言

江苏传智播客教育科技股份有限公司(简称“传智播客”)是一家致力于培养高素质软件开发人才的科技公司。“黑马程序员”是传智播客旗下的高端 IT 教育品牌。

“黑马程序员”的学员多为大学毕业后,想从事 IT 行业,但各方面条件还不成熟的年轻人。“黑马程序员”的学员筛选制度非常严格,包括严格的技术测试、自学能力测试,还包括性格测试、压力测试、品德测试等,以百里挑一的残酷筛选制度确保学员质量,降低企业的用人风险。

自“黑马程序员”成立以来,教学研发团队一直致力于打造精品课程资源,不断在产、学、研三个层面创新自己的执教理念与教学方针,并集中“黑马程序员”的优势力量,有针对性地出版了计算机系列教材 60 多种,制作教学视频数十套,发表各类技术文章数百篇。

“黑马程序员”不仅斥资研发 IT 系列教材,还为高校师生提供以下配套学习资源与服务。

为大学生提供的配套服务

1. 请登录在线平台 http://yx.boxuegu.com,免费获取海量学习资源,还有专业的老师在线为您答疑解惑。

2. 针对高校学生在学习过程中存在的压力等问题,我们还面向大学生量身打造了 IT 技术女神——“播妞”,可提供教材配套源代码和习题答案以及更多 IT 学习资源。同学们可以添加“播妞”微信号 208695827 和“播妞”QQ 号 3231342131,获取学习资源。

“播妞”微信

“播妞”QQ

为教师提供的配套服务

针对高校教学,“黑马程序员”为 IT 系列教材精心设计了“教案+授课资源+考试系

统＋题库＋教学辅助案例”的系列教学资源。高校老师请登录在线平台 http://yx.boxuegu.com 或关注码大牛老师微信/QQ2011168841，获取配套资源，也可以扫描下方二维码，加入专为 IT 教师打造的师资服务平台——“教学好助手”，获取最新教师教学辅助资源的相关动态。

前言

我们生活在一个充满“数据”的时代，刷微信、聊 QQ、网购、旅游、看病等一系列行为无时无刻不在产生新的数据，日积月累形成巨大的数据集，迎来了大数据时代。大数据时代的力量，正在积极地影响着人们生活的方方面面，深刻改变着人类的思维、生产、生活、学习方式，深刻展示了世界发展的前景。

大数据时代，数据的存储与挖掘至关重要。企业在追求高可靠性、高扩展性及高容错性的大数据处理平台的同时还希望能够降低成本，而 Hadoop 为实现这些需求提供了解决方案。这里列举 3 条使用 Hadoop 作为大数据业务的基础原因，具体如下。

(1) Hadoop 底层的分布式文件系统具有高拓展性，通过数据冗余保证数据不丢失和提升计算效率，同时可以存储各种格式的数据。它还有多种计算框架，既可以进行离线计算也可以进行在线实时计算。

(2) Hadoop 是架构在廉价的硬件服务器上，且产品是开源的，供开发者免费使用，开发成本和维护成本都降低很多。

(3) Hadoop 具有成熟的生态圈，有许多辅助系统对数据进行处理。

本书作为大数据技术 Hadoop 的入门教程，最重要又最难的一件事情就是将一些复杂、难以理解的思想和问题简单化，让初学者能够轻松理解并快速掌握。本教材对每个知识点都进行了深入分析，并针对每个知识点精心设计了相关案例，然后模拟这些知识点在实际工作中的运用，真正做到了知识的讲解由浅入深、由易到难。

全书共分为 11 章。

第 1 章主要讲解什么是大数据以及 Hadoop 相关概念。通过本章的学习，读者可对大数据有简单的认识，并了解 Hadoop 生态圈工具及各自的用途。

第 2 章主要讲解 Hadoop 集群的构建。通过本章的学习，读者能掌握 Linux 系统网络配置、独立搭建 Hadoop 开发平台，以及简单操作 Hadoop 系统。

第 3 章主要讲解 Hadoop 分布式文件系统(HDFS)。通过本章的学习，读者可以掌握 HDFS 的架构和工作原理，并能够通过 Shell 接口和 Java API 操作 HDFS。

第 4 章主要讲解 MapReduce 的相关知识。通过本章的学习，初学者可以了解 MapReduce 计算框架的思想并且能够使用 MapReduce 解决实际问题。

第 5 章主要讲解 Zookeeper 分布式协调服务。通过本章的学习，读者能够对 Zookeeper 分布式协调服务有基本的认识，掌握 Zookeeper 内部运行原理，并会通过 Shell 和 Java API 操作 Zookeeper。

第 6 章主要讲解 Hadoop 2.0 的新特性，包括 YARN 资源管理框架和 HDFS 的高可用。其中，YARN 作为资源管理框架，读者需要明白它的体系结构和工作流程；HDFS 的高

可用性能够解决集群的单点故障问题，读者要掌握高可用架构的部署方式，并能独立参考文档搭建高可用的 Hadoop 集群。

第 7 章主要讲解 Hive 的相关知识。读者需要了解 Hive 架构、数据模型、Hive 的安装和管理以及 Hive 的数据操作。这里建议初学者在学习 Hive 时多动手操作 Hive，通过丰富的案例练习，掌握 Hive 的使用。

第 8 章主要讲解 Flume 日志采集系统的基本知识。通过本章的学习，读者应该掌握 Flume 的基本概念、运行机制并且能够掌握 Flume 的安装配置和基本使用。

第 9 章主要讲解 Azkaban 工作流管理器的基本知识。通过本章的学习，读者应该对 Azkaban 有一定的了解，掌握 Azkaban 的部署和使用，并能够使用 Azkaban 进行任务调度管理。

第 10 章主要讲解 Sqoop 数据迁移工具的相关知识。通过本章的学习，读者可以掌握 Sqoop 工作原理，会独立搭建 Sqoop 工具并且能够使用 Sqoop 工具完成常用的数据迁移操作。

第 11 章主要通过开发网站流量日志分析系统来讲解利用 Hadoop 生态体系的技术解决实际问题。通过本章的学习，读者可以了解大数据系统的架构、数据采集、数据预处理、数据仓库的设计、数据分析、数据导出以及最后可视化处理。读者应该熟练掌握系统架构以及业务流程，熟练使用 Hadoop 生态体系相关技术。

致谢

本书的编写和整理工作由传智播客教育科技股份有限公司完成，主要参与人员有吕春林、高美云、石荣新、翟振方、文燕等，全体参编人员在这近一年的编写过程中付出了许多辛勤的汗水，在此表示衷心的感谢。

意见反馈

尽管我们尽了最大的努力，但书中难免会有欠妥之处，欢迎各界专家和读者朋友们来信提出宝贵意见，我们将不胜感激。您在阅读本书时，如果发现任何问题或有不认同之处可以通过电子邮件与我们取得联系。

请发送电子邮件至 itcast_book@vip.sina.com。

黑马程序员

2019 年 3 月于北京

目录

第1章 初识Hadoop

学习目标

- 了解大数据的概念及其特征。
- 熟悉大数据的典型应用。
- 了解 Hadoop 的发展历史及其版本。
- 掌握 Hadoop 的生态体系。

随着近几年计算机技术和互联网的发展，“大数据”这个词被提及得越来越频繁。与此同时，大数据的快速发展无时无刻不在影响着我们的生活，例如，医疗方面，大数据能够帮助医生预测疾病；电商方面，大数据能够向顾客个性化推荐商品；交通方面，大数据会帮助人们选择最佳出行方案。

Hadoop 作为一个能够对大量数据进行分布式处理的软件框架，用户可以利用 Hadoop 生态体系开发和处理海量数据。由于 Hadoop 可靠及高效的处理性能，使得它逐渐成为分析大数据的领先平台。接下来，将深入介绍大数据以及 Hadoop 的相关概念，为后面知识的学习建立概念体系。

1.1 大数据概述

1.1.1 什么是大数据

高速发展的信息时代，新一轮科技革命和变革正在加速推进，技术创新日益成为重塑经济发展模式和促进经济增长的重要驱动力量，而“大数据”无疑是核心推动力。

那么，什么是“大数据”呢？如果从字面意思来看，大数据指的是巨量数据。那么可能有人会问，多大量级的数据才叫大数据？不同的机构或学者有不同的理解，难以有一个非常定量的定义，只能说，大数据的计量单位已经越过 TB 级别发展到 PB、EB、ZB、YB 甚至 BB 级别。

最早提出“大数据”这一概念的是全球知名咨询公司麦肯锡，它是这样定义大数据的：一种规模大到在获取、存储、管理、分析方面大大超出了传统数据库软件工具能力范围的数据集合，具有海量的数据规模、快速的数据流转、多样的数据类型以及价值密度低四大特征。

研究机构 Gartner 是这样定义大数据的：“大数据”是需要新处理模式才能具有更强的决策力、洞察发现力和流转优化能力来适应海量、高增长率和多样化的信息资产。

若从技术角度来看，大数据的战略意义不在于掌握庞大的数据，而在于对这些含有意义

的数据进行专业化处理，换言之，如果把大数据比作一种产业，那么这种产业盈利的关键在于提高对数据的“加工能力”，通过“加工”实现数据的“增值”。

1.1.2 大数据的特征

一般认为，大数据主要具有以下 4 个方面的典型特征，即大量(Volume)、多样(Variety)、高速(Velocity)和价值(Value)，即所谓的“4V”，接下来，通过一张图 1-1 来具体描述。

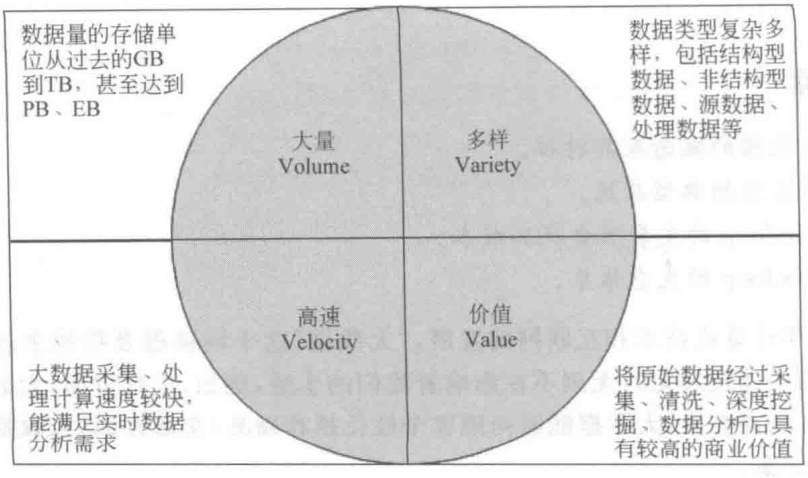


图 1-1 大数据 4V 特征

接下来针对图 1-1 中的 4V 特征进行简要介绍，具体如下。

1. Volume(大量)

大数据的特征首先就是数据规模大。随着互联网、物联网、移动互联技术的发展，人和事物的所有轨迹都可以被记录下来，数据呈现出爆发性增长。数据相关计量单位的换算关系如表 1-1 所示。

表 1-1 单位换算关系

单 位	换算公式	单 位	换算公式
Byte	1Byte=8bit	TB	1TB=1024GB
KB	1KB=1024Byte	PB	1PB=1024TB
MB	1MB=1024KB	EB	1EB=1024PB
GB	1GB=1024MB	ZB	1ZB=1024EB

2. Variety(多样)

数据来源的广泛性，决定了数据形式的多样性。大数据可以分为三类，一是结构化数据，如财务系统数据、信息管理系统数据、医疗系统数据等，其特点是数据间因果关系强；二是非结构化的数据，如视频、图片、音频等，其特点是数据间没有因果关系；三是半结构化数

据，如 HTML 文档、邮件、网页等，其特点是数据间的因果关系弱。有统计显示，目前结构化数据占据整个互联网数据量的 75%以上，而产生价值的大数据，往往是这些非结构化数据。

3. Velocity(高速)

数据的增长速度和处理速度是大数据高速性的重要体现。与以往的报纸、书信等传统数据载体生产传播方式不同，在大数据时代，大数据的交换和传播主要是通过互联网和云计算等方式实现的，其生产和传播数据的速度是非常迅速的。另外，大数据还要求处理数据的响应速度要快，例如，上亿条数据的分析必须在几秒内完成。数据的输入、处理与丢弃必须立刻见效，几乎无延迟。

4. Value(价值)

大数据的核心特征是价值，其实价值密度的高低和数据总量的大小是成反比的，即数据价值密度越高数据总量越小，数据价值密度越低数据总量越大。任何有价值的信息的提取依托的就是海量的基础数据。当然目前大数据背景下有个未解决的问题，如何通过强大的机器算法更迅速地在海量数据中完成数据的价值提纯。

1.1.3　研究大数据的意义

现在的社会是一个高速发展的社会，科技发达，信息流通，人们之间的交流也越来越密切，生活也越来越便捷，大数据就是这个高科技时代的产物。阿里巴巴创办人马云曾经说过，未来的时代将不是 IT 时代，而是 DT 的时代，DT 就是 Data Technology，数据科技，这显示出大数据对于阿里巴巴集团来说是举足轻重的。

有人把数据比喻为蕴藏能量的煤矿。煤炭按照性质有焦煤、无烟煤、肥煤、贫煤等分类，而露天煤矿、深山煤矿的挖掘成本又不一样。与此类似，大数据并不在于“大”，而在于“有用”。数据的价值含量、挖掘成本比数量更为重要。对于很多行业而言，如何利用这些大规模数据，发掘其潜在价值，才是赢得核心竞争力的关键。

研究大数据，最重要的意义是预测。因为数据从根本上讲，是对过去和现在的归纳和总结，其本身不具备趋势和方向性的特征，但是可以应用大数据去了解事物发展的客观规律、了解人类行为，并且能够帮助我们改变过去的思维方式，建立新的数据思维模型，从而对未来进行预测和推测。比如，商业公司对消费者日常的购买行为和使用商品习惯进行汇总和分析，了解到消费者的需求，从而改进已有商品并适时推出新的商品，消费者的购买欲就会提高。知名互联网公司谷歌对其用户每天频繁搜索的词汇进行数据挖掘，从而进行相关的广告推广和商业研究。

大数据的处理技术迫在眉睫，近年来各国政府和全球学术界都掀起了一场大数据技术的革命，众人纷纷积极研究大数据的相关技术。很多国家都把大数据技术研究上升到了国家战略高度，提出了一系列的大数据技术研发计划，从而推动政府机构、学术界、相关行业和各类企业对大数据技术进行探索和研究。

可以说大数据是一种宝贵的战略资源，其潜在价值和增长速度正在改变着人类的工作、生活和思维方式。可以想象，在未来，各行各业都会积极拥抱大数据，积极探索数据挖掘和