“十三五”
国家重点图书出版规划项目
ICT认证系列丛书

HUAWEI
Information
Network
Academy

华为信息与网络技术学院指定教材

大数据原理与技术

黄史浩 / 编著

LEADING NEW ICT

非外借

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

“十三五”
国家重点图书出版规划项目
ICT认证系列丛书

华为信息与网络技术学院指定教材

大数据原理与技术

黄史浩 / 编著

人民邮电出版社
北京

图书在版编目（CIP）数据

大数据原理与技术 / 黄史浩编著. — 北京 : 人民邮电出版社, 2018.1(2018.3重印)
（ICT认证系列丛书）
ISBN 978-7-115-45871-1

Ⅰ. ①大… Ⅱ. ①黄… Ⅲ. ①数据处理－教材 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2017)第279983号

内 容 提 要

本书是华为ICT学院大数据技术官方教材，旨在帮助学生进一步学习大数据的基本概念、技术原理，以及大数据平台的搭建和使用。

本书从大数据的概念和特征开始讲起，首先让读者对大数据有一个感性的认识；接下来结合大数据平台的各个功能模块，详细介绍大数据的存储、处理、分析、可视化等原理和操作；最后对大数据在各种行业中的应用加以叙述，让读者更加充分地感受到大数据应用的价值。

除华为ICT学院的学生之外，本书同样适合正在备考HCNA-Big Data认证，或者正在参加HCNA-Big Data技术培训的学员进行阅读和参考。其他有志进入ICT行业的人员和大数据技术爱好者也可以通过阅读本书，加深自己对大数据技术的理解。

◆ 编　　著　黄史浩
责任编辑　李　静
责任印制　彭志环

◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路11号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
固安县铭成印刷有限公司印刷

◆ 开本：787×1092　1/16
印张：17　　2018年1月第1版
字数：410千字　　2018年3月河北第2次印刷

定价：56.00元

读者服务热线：(010)81055488　印装质量热线：(010)81055316
反盗版热线：(010)81055315

序

物联网、云计算、大数据、人工智能等新技术的兴起，推动着社会的数字化演进。全球正在从“人人互联”，发展至“万物互联”，未来二三十年，人类社会将演变成以“万物感知、万物互联、万物智能”为特征的智能社会。

新兴技术快速渗透并推动企业加速数字化转型，企业业务应用系统趋于横向贯通，数据趋于融合互联，ICT 正在成为企业新一代公共基础设施和创新引擎，成为企业的核心生产系统。据华为 GIV（全球 ICT 产业愿景）预测，到 2025 年，全球的连接数将达到 1000 亿，85%的企业应用上云，100%的企业会连接云服务，工业智能的普及率将超过 20%。数字化发展为各行业带来的纵深影响远超出想象。

作为企业数字化转型中的关键使能者，ICT 人才将站在更新的高度，以更为全局的视角审视整个行业，并依靠新思想、新技术驱动行业发展。因此，企业对于融合型 ICT 人才需求也更为迫切。未来 5 年，华为所领导的全球 ICT 产业生态系统对人才的需求将超过 80 万。华为积累了 20 余年的 ICT 人才培养经验，对 ICT 行业发展现状及趋势有着深刻的理解。面对数字化转型背景下的企业 ICT 人才短缺，华为致力于构建良性的 ICT 人才生态。2013 年，华为开始与高校合作，共同制定 ICT 人才培养计划，设立华为信息与网络技术学院（简称华为 ICT 学院），依据企业对 ICT 人才的新需求，将物联网、云计算、大数据等新技术和最佳实践经验融入到课程与教学中。华为希望通过校企合作，让大学生在校园内就能掌握新技术，并积累实践经验，促使他们快速成长为有应用能力、会复合创新、能动态成长的融合型人才。

教材是知识传递、人才培养的重要载体，华为聚合技术专家、高校教师倾心打造 ICT 学院系列精品教材，希望帮助大学生快速完成知识积累，奠定坚实的理论基础，助力同学们更好地开启 ICT 职业道路，奔向更美好的未来。

亲爱的同学们，面对新时代对 ICT 人才的呼唤，请抓住历史机遇，拥抱精彩的 ICT 时代，书写未来职业的光荣与梦想吧！华为，将始终与你同行！

前　言

随着技术的发展，数据的产生、传输和存储变得越来越容易。人类社会产生的信息也越来越多地被数据化。这些海量、详尽的数据让人们可以更客观、全面地探索和研究世界。数据已经成为一种重要的生产要素。大小超出了典型数据库软件的采集、存储、管理和分析等能力的数据集被称为大数据，大数据有海量的数据（Volume）、快速的数据处理（Velocity）、多样的数据类型（Variety）和低价值密度（Value）四大特征，统称“4V”。大数据蕴藏着巨大的价值，对大数据的运用和价值挖掘会给社会和企业带来新的机遇和变革。

本书主要内容

第 1 章　大数据概述

人类社会产生的数据与日俱增，涉及社会的方方面面。数据已经成为一种重要的生产要素，这些海量的数据被称为“大数据”。大数据蕴藏着巨大的价值，对大数据的运用和价值挖掘会给社会和企业带来新的机遇和变革。本章讲述了大数据的基本概念、关键技术及其产业模式。

第 2 章　Hadoop 大数据处理平台

目前大数据标准开源软件为 Hadoop。Hadoop 是 Apache 基金会开发的分布式计算平台。它可以在大规模计算机集群中提供海量数据的处理能力。由于其良好的性能，Hadoop 大数据处理平台在大数据企业中应用广泛。本章对 Hadoop 大数据处理平台做了详细介绍。

第 3 章　大数据存储技术（HDFS）

Hadoop 分布式文件系统（Hadoop Distributed File System，HDFS）是 Hadoop 的两大核心之一。它是使用 Java 实现的、分布式的、可横向扩展的分布式文件系统，是对谷歌分布式文件系统（Google File System，GFS）的开源实现。本章对 HDFS 做了详细介绍。

第 4 章　大数据离线计算框架（MapReduce&YARN）

Hadoop 另一个核心组件是 MapReduce。Hadoop MapReduce 能在由大量的普通配置的计算机组成的集群上处理超大数据集，具有易于编程、扩展性高和容错性高的特点。

除了两大核心 HDFS 和 MapReduce 之外，Hadoop 还有其他组件为其提供丰富的功能。

Hadoop 中的资源管理调度系统被称为 YARN（Yet Another Resource Negotiator，另一种资源协调者）。它是一个通用的资源管理模块，可以为上层应用提供统一的资源管理和调度。本书第 4 章对 MapReduce 和 YARN 做了详细介绍。

第 5 章　大数据数据库（HBase）

Hadoop 中除了基本的文件系统，还提供数据库和数据仓库，方便用户对数据进行处理。

Hadoop 中使用的数据库为 HBase。HBase 是基于谷歌 Bigtable 开发的开源分布式数据库，具有高可靠、高性能、高伸缩、面向列等特点。HBase 运行在 HDFS 上。它主要用来存储非结构化和半结构化数据。第 5 章对 HBase 做了详细介绍。

第 6 章　大数据数据仓库（Hive）

Hive 是基于 Hadoop 的数据仓库软件。它可以用来进行数据提取转化加载（ETL），在 Hadoop 中存储、查询和分析大规模数据。本书第 6 章对 Hive 做了详细介绍。

第 7 章　大数据数据转换（Sqoop 与 Loader）

为了方便外界存储与 Hadoop 平台之间的数据传输，Hadoop 提供了高效传输批量数据的工具 Sqoop。Sqoop 可用于将数据从外部结构化数据存储导入 Hadoop 平台；也可用于从 Hadoop 中提取数据，并将其导出到外部结构化数据存储。本章对 Sqoop 做了详细介绍。

第 8 章　大数据日志处理（Flume）

此外，Hadoop 还提供组件做日志收集处理。Flume 是一个分布式、高可靠和高可用的海量日志聚合系统。它支持从多种数据源收集数据，在对数据进行简单处理后，将数据写到数据接收方（可定制）。本章对 Flume 做了详细介绍。

第 9 章　大数据实时计算框架（Spark）

Spark 在 2009 年诞生于 UC Berkeley AMP Lab（加州大学伯克利分校的 AMP 实验室），是使用内存计算的开源大数据并行计算框架。它提供了强大的技术栈，可以应对复杂的大数据处理场景，包括 SQL 查询、机器学习、图形计算和流式计算等方面。第 9 章对 Spark 做了详细介绍。

第 10 章　大数据流计算

Hadoop 可通过 Spark 和 Storm 进行流计算。流计算是一种由事件触发、持续作用、低延迟的计算方式。它可以很好地对流数据进行实时分析处理，捕捉到可能有用的信息。本书第 10 章对流计算做了详细介绍。

第 11 章　数据可视化

为了更直观地展示大数据的分析处理结果，我们还需要使用数据可视化技术。常见的数据可视化工具有 Excel、R 语言、Tableau 和 QlikView。这部分内容会在第 11 章中

做详细介绍。

第 12 章　大数据行业应用

目前，大数据技术在金融、医疗、制造业、能源、互联网、政府公共事业、媒体、零售等领域已经得到了广泛的应用，在社会以及企业的发展上起到了重要的作用。本书第 12 章介绍了大数据在金融行业、电信行业、公安系统以及互联网行业的应用案例，以便读者对大数据的应用现状有更直观的了解。

配套资源

华为 ICT 学院为本书从理论到实战提供了贴合实际应用的定制化教学学习资源，成为华为 ICT 学院之后即可获取使用如下配套资源。

- 实验手册：教材配套实验材料，助力读者动手能力的提升，以实验促进读者对理论知识的理解。
- 视频讲解：教材配套重点知识讲解小视频，帮助读者更好地理解书中的重点、难点，相关视频可到华为 ICT 学院官方网站进行观看。
- 授课 PPT：教材配套授课材料，方便高校授课，提升教师备课效率。
- 综合实训：教材配套实训课程，还原真实项目，提升读者应对实际项目的能力。

关于本书读者

本书定位是华为 ICT 学院大数据技术官方教材，本书适合以下几类读者。

- 华为 ICT 学院的学生。
- 各大高校 ICT 专业领域学生。
- 正在学习 HCNA-Big Data 课程的学员和正在备考 HCNA-Big Data 认证的考生。
- 有志于进入 ICT 行业的初学者。
- 大数据技术爱好者。

本书作者

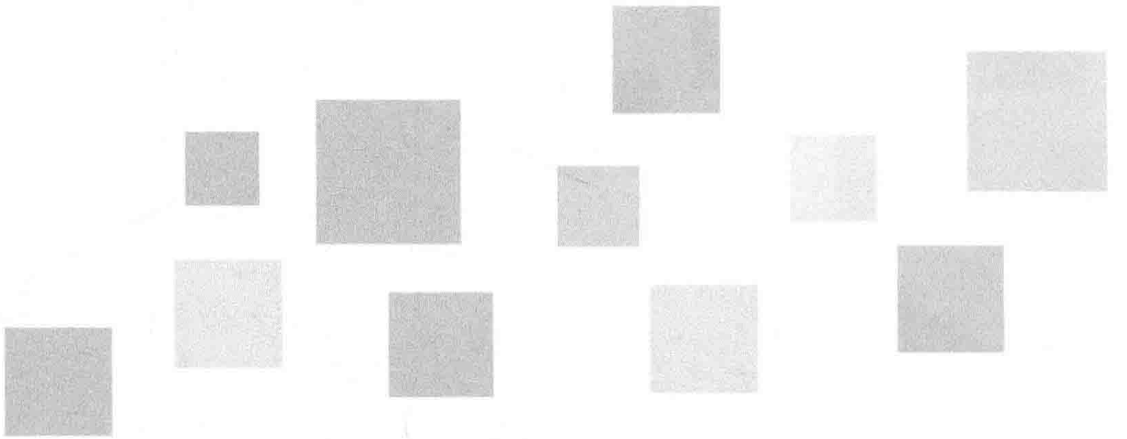
编著：黄史浩

编委人员（排名不分先后）：林业灿、钱兴会、张文博、张粤磊

技术审校（排名不分先后）：高冬冬、贾云涛、刘洁、刘洋、鄢华、张博、张亮、张志峰、傅开宏、鲁戈

目　录

第1章
大数据概述



随着信息技术的发展，硬件成本不断降低，网络带宽获得大幅提升，这些都为大量数据的传输和存储提供了基础；智能终端的普及及物联网的发展让更多的人和物链接到互联网中，成为数据的生产者；电子商务、社交网络、共享经济的发展也让人类活动踪迹越来越多地发生在网络上并以数据的形式记录下来。人类社会活动产生的数据与日俱增，涉及社会的方方面面。数据已经成为一种重要的生产要素，这些海量的数据被称为“大数据”。大数据蕴藏着巨大的价值，对大数据的运用和价值挖掘会给社会和企业带来新的机遇和变革。

本章讲述大数据的基本概念、关键技术以及产业模式。

学习目标

- 理解大数据的概念。
- 了解大数据的关键技术。
- 了解大数据的产业模式。
- 了解大数据的应用。

1.1 大数据的概念与价值

1.1.1 什么是大数据

信息技术咨询研究与顾问咨询公司 Gartner 给大数据做出了这样的定义：大数据是