

“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



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 (C I P) 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
字数: 410 千字

2018 年 1 月第 1 版

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



1.1 大数据的概念与价值

1.2 大数据的关键技术

1.3 大数据产业

1.4 大数据应用场景

1.5 本章总结

练习题

目 录

第 1 章 大数据概述	0
1.1 大数据的概念与价值	2
1.1.1 什么是大数据	2
1.1.2 大数据的来源	4
1.1.3 大数据有什么价值	5
1.1.4 如何挖掘企业大数据的价值	6
1.2 大数据的关键技术	7
1.2.1 大数据采集、预处理与存储管理	7
1.2.2 大数据分析 with 挖掘	8
1.2.3 数据可视化	9
1.3 大数据产业	9
1.3.1 数据提供	9
1.3.2 技术提供	9
1.3.3 服务提供	10
1.4 大数据应用场景	10
1.5 本章总结	11
练习题	12
第 2 章 Hadoop 大数据处理平台	14
2.1 Hadoop 平台概述	16
2.1.1 Hadoop 简介	16
2.1.2 Hadoop 的特性	17
2.1.3 Hadoop 应用现状	17
2.1.4 Hadoop 版本及相关平台	18
2.2 Hadoop 生态系统	18
2.2.1 Hadoop 存储系统 (HDFS&HBase)	18
2.2.2 Hadoop 计算框架 (MapReduce&YARN)	19
2.2.3 Hadoop 数据仓库 (Hive)	20
2.2.4 Hadoop 数据转换与日志处理 (Sqoop&Flume)	20

2.2.5	Hadoop 应用协调与工作流 (ZooKeeper&Oozie)	20
2.2.6	大数据安全技术 (Kerberos&LDAP)	21
2.2.7	大数据即时查询与搜索 (Impala&Solr)	21
2.2.8	大数据消息订阅 (Kafka)	21
2.3	Hadoop 安装部署	22
2.3.1	Hadoop 规划部署	22
2.3.2	Hadoop 的安装方式	23
2.4	华为 FusionInsight HD 安装部署	26
2.4.1	FusionInsight HD 简介	26
2.4.2	FusionInsight HD 集成设计	28
2.4.3	FusionInsight HD 安装部署	33
2.4.4	FusionInsight HD 重要参数配置	41
2.5	本章总结	42
	练习题	43
第 3 章	大数据存储技术 (HDFS)	44
3.1	概述	46
3.1.1	分布式文件系统的概念与作用	47
3.1.2	HDFS 概述	47
3.2	HDFS 的相关概念	48
3.2.1	块	48
3.2.2	NameNode	49
3.2.3	Secondary NameNode	50
3.2.4	DataNode	51
3.3	HDFS 体系架构与原理	52
3.3.1	HDFS 体系架构	52
3.3.2	HDFS 的高可用机制	52
3.3.3	HDFS 的目录结构	54
3.3.4	HDFS 的数据读写过程	57
3.4	HDFS 接口及其在 FusionInsight HD 编程中的实践	58
3.4.1	HDFS 常用 Shell 命令	59
3.4.2	HDFS 的 Web 界面	60
3.4.3	HDFS 的 Java 接口及应用实例	62
3.5	本章总结	67
	练习题	67

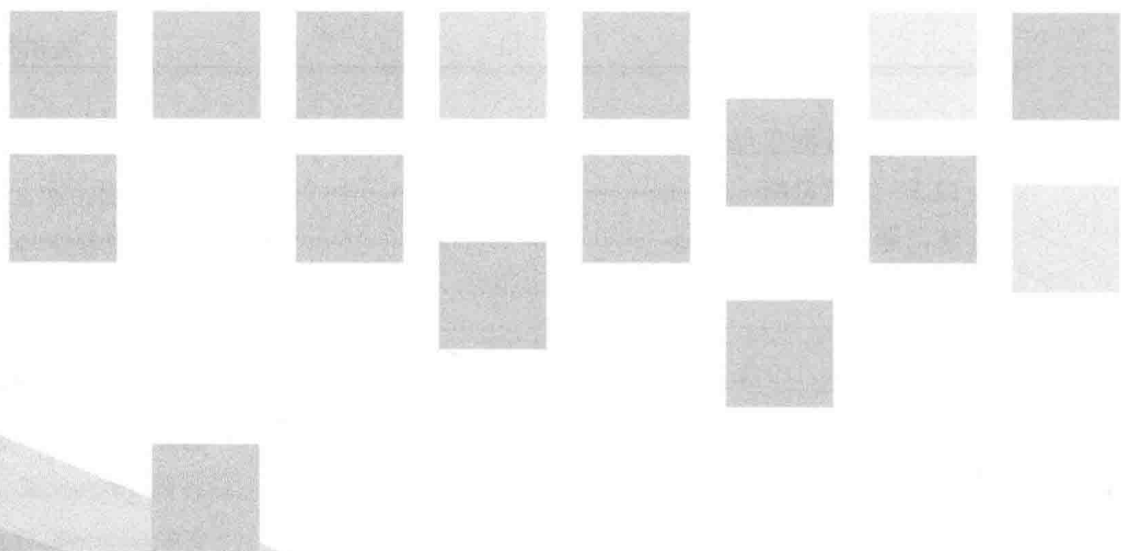
第 4 章 大数据离线计算框架 (MapReduce & YARN)	70
4.1 MapReduce 技术原理	72
4.1.1 MapReduce 概述	73
4.1.2 Map 函数与 Reduce 函数	73
4.2 YARN 技术原理	74
4.2.1 YARN 的概述与应用	74
4.2.2 YARN 的架构	75
4.2.3 MapReduce 的计算过程	76
4.2.4 YARN 的资源调度	78
4.3 FusionInsight HD 中 MapReduce 的应用	78
4.3.1 WordCount 实例分析	78
4.3.2 MapReduce 编程实践	79
4.4 本章总结	85
练习题	86
第 5 章 大数据数据库 (HBase)	88
5.1 HBase 概述	90
5.1.1 HBase 简介	90
5.1.2 HBase 与关系型数据库的区别	91
5.1.3 HBase 的应用场景	92
5.2 HBase 的架构原理	92
5.2.1 HBase 的数据模型	92
5.2.2 表和 Region	93
5.2.3 HBase 的系统架构与功能组件	94
5.2.4 HBase 的读写流程	96
5.2.5 HBase 的 Compaction 过程	97
5.3 FusionInsight HD 中 HBase 的编程实践	98
5.3.1 FusionInsight HD 中 HBase 的常用参数配置	98
5.3.2 HBase 的常用 Shell 命令	100
5.3.3 HBase 常用的 Java API 及应用实例	103
5.4 本章总结	118
练习题	118
第 6 章 大数据数据仓库 (Hive)	120
6.1 Hive 概述	122
6.1.1 Hive 简介和应用	122

6.1.2	Hive 的特性	123
6.1.3	Hive 与传统数据仓库的区别	124
6.2	Hive 的架构和数据存储	124
6.2.1	Hive 的架构原理	124
6.2.2	Hive 的数据存储模型	127
6.2.3	HiveQL 编程	128
6.3	FusionInsight HD 中 Hive 应用实践	132
6.3.1	FusionInsight HD 中 Hive 的常用参数配置	132
6.3.2	加载数据到 Hive	133
6.3.3	使用 HiveQL 进行数据分析	135
6.4	本章总结	139
	练习题	139
第 7 章	大数据数据转换 (Sqoop 与 Loader)	142
7.1	Sqoop 概述	144
7.1.1	Sqoop 简介与应用	145
7.1.2	Sqoop 的功能与特性	145
7.1.3	Sqoop 与传统 ETL 的区别	146
7.2	FusionInsight HD 中 Loader 的应用实践	146
7.2.1	FusionInsight HD 中 Loader 与 Sqoop 的对比	147
7.2.2	FusionInsight HD 中 Loader 的参数配置	148
7.2.3	使用 Loader 进行数据转换	149
7.2.4	Loader 的常用 Shell 命令	150
7.2.5	Loader 应用实践	152
7.3	本章总结	153
	练习题	154
第 8 章	大数据日志处理 (Flume)	156
8.1	Flume 概述	158
8.1.1	Flume 简介与应用	158
8.1.2	Flume 的功能与特性	161
8.1.3	Flume 与其他主流开源日志收集系统的区别	162
8.2	FusionInsight HD 中 Flume 的应用实践	162
8.2.1	FusionInsight HD 中 Flume 的常用参数配置	163
8.2.2	Flume 常用的 Shell 命令	164
8.2.3	Flume 与 Kafka 结合进行日志处理	165

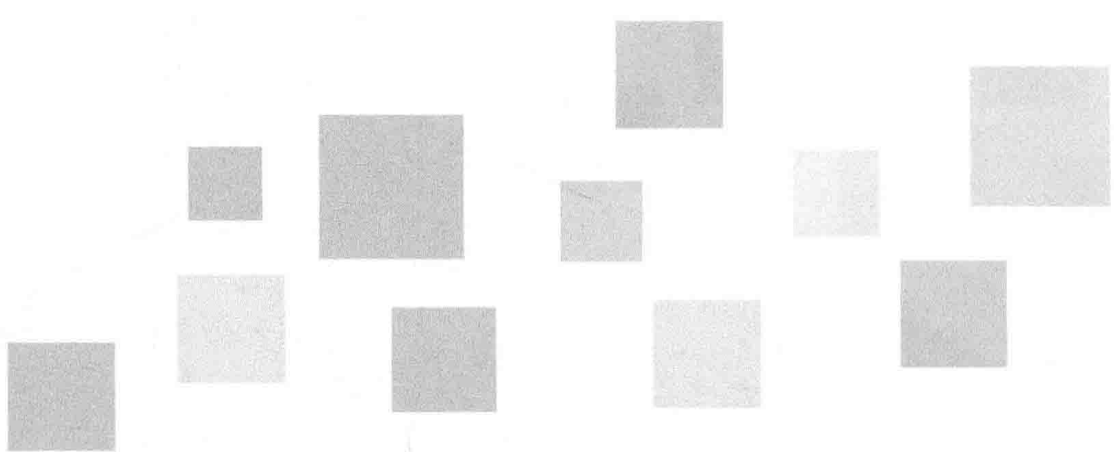
8.3 本章总结	168
练习题	169
第 9 章 大数据实时计算框架 (Spark)	170
9.1 Spark 概述	172
9.1.1 Spark 的概述与应用	173
9.1.2 Scala 语言介绍	174
9.1.3 Spark 生态系统组件	174
9.1.4 Spark 与 Hadoop 的对比	175
9.2 Spark 技术架构	176
9.2.1 Spark 的运行原理	176
9.2.2 RDD 概念与原理	177
9.2.3 Spark 的三种部署方式	181
9.2.4 使用开发工具测试 Spark	182
9.3 FusionInsight HD 中 Spark 应用实践	183
9.3.1 运行 Spark Shell	183
9.3.2 进行 Spark RDD 操作	184
9.3.3 使用 Spark 客户端工具运行 Spark 程序	185
9.4 Spark Streaming	188
9.4.1 Spark Streaming 的设计思想	188
9.4.2 Spark Streaming 的应用实例	189
9.5 Spark SQL	191
9.5.1 Spark SQL 的功能	191
9.5.2 FusionInsight HD 中 Spark SQL 的应用实例	192
9.6 Spark MLlib	193
9.6.1 机器学习简介	193
9.6.2 Spark MLlib 的功能	194
9.7 Spark GraphX	194
9.7.1 图计算简介	194
9.7.2 Spark GraphX 功能简介	195
9.8 本章总结	195
练习题	196
第 10 章 大数据流计算	198
10.1 流计算概述	200
10.1.1 静态数据和流数据	201

10.1.2	流计算的概念	201
10.1.3	MapReduce 和流计算	202
10.1.4	流计算框架	202
10.2	流计算的处理流程	203
10.2.1	数据实时采集	203
10.2.2	数据实时计算	203
10.2.3	数据实时查询	203
10.3	Streaming 流计算	204
10.3.1	Streaming 简介	204
10.3.2	Streaming 的特点	206
10.3.3	Streaming 在 FusionInsight HD 上的应用实践	208
10.3.4	Spark Streaming 与 Streaming 的差异	212
10.4	本章总结	213
	练习题	213
第 11 章	数据可视化	216
11.1	可视化概述	218
11.1.1	数据可视化简介	219
11.1.2	数据可视化的重要性	219
11.1.3	可视化的发展历程	219
11.1.4	数据可视化的过程	221
11.2	可视化工具	222
11.2.1	入门级工具 (Excel)	222
11.2.2	普通工具 (R 语言)	222
11.2.3	高级工具 (Tableau 和 QlikView)	223
11.3	可视化的典型应用	223
11.3.1	可视化在医学上的应用	223
11.3.2	可视化在工程中的应用	224
11.3.3	可视化在互联网的应用	225
11.4	本章总结	225
	练习题	226
第 12 章	大数据行业应用	228
12.1	大数据在金融行业的应用	230
12.2	大数据在电信行业的应用	232
12.3	大数据在公安系统的应用	236

12.4 大数据在互联网行业的应用·····	237
12.5 本章总结·····	237
练习题·····	238
术语表·····	240
参考文献·····	252



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 and 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