教育部高等学校计算机类专业教学指导委员会-华为ICT产学合作项目

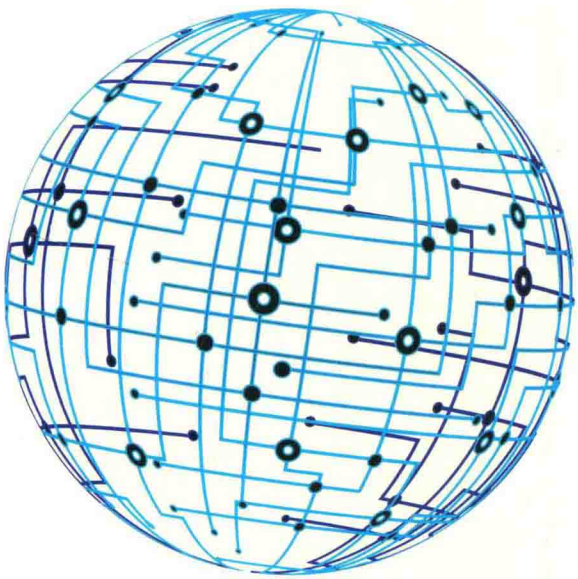
数据科学与大数据技术专业系列规划教材

华为信息与网络技术学院指定教材

# 大数据技术基础

薛志东 ◉ 主编

吕泽华 陈长清 黄浩 ◉ 副主编



**系统**、**完整**的数据科学与大数据技术**专业解决方案**

**名校名师**打造**大数据领域**精品力作

**全面掌握**大数据技术概况

讲解**Hadoop**生态圈**平台**、**工具**与**技术**

中国工信出版集团

人民邮电出版社 POSTS & TELECOM PRESS

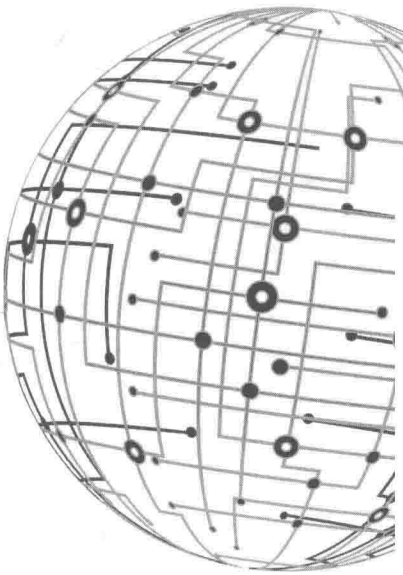
教育部高等学校计算机
产学合作项目

数据科学与大数据技术专业系列规划教材

华为信息与网络技术学院指定教材

# 大数据技术基础

薛志东 ◉ 主编
吕泽华 陈长清 黄浩 ◉ 副主编





人民邮电出版社
北京

**图书在版编目（CIP）数据**

大数据技术基础 / 薛志东主编. -- 北京 : 人民邮电出版社, 2018.8
数据科学与大数据技术专业系列规划教材
ISBN 978-7-115-48307-2

Ⅰ. ①大… Ⅱ. ①薛… Ⅲ. ①数据处理软件 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2018)第097274号

## 内 容 提 要

本书系统、全面地介绍了大数据技术的基础知识，期望读者通过对本书的学习和实践了解大数据技术的概貌，掌握 Hadoop 生态圈大数据技术中最为基础和关键的知识。本书主要内容包括大数据概述、大数据软件基础、大数据存储技术、MapReduce 分布式编程、数据采集与预处理、数据仓库与联机分析处理、大数据分析与挖掘技术、Spark 分布式内存计算框架、数据可视化技术、大数据安全。

本书可作为数据科学与大数据、软件工程、计算机科学与技术等专业的大数据概论课程的教材，也可供大数据工程技术人员阅读使用。

◆ 主　　编　薛志东
副 主 编　吕泽华　陈长清　黄　浩
策划编辑　戴思俊
责任编辑　邹文波
责任印制　焦志炜

◆ 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
北京鑫正大印刷有限公司印刷

◆ 开本：787×1092　1/16
印张：20　　2018 年 8 月第 1 版
字数：526 千字　　2018 年 8 月北京第 1 次印刷

定价：55.00 元

读者服务热线：(010)81055256　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字 20170147 号

教育部高等学校计算机类专业教学指导委员会-华为 ICT 产学合作项目
数据科学与大数据技术专业系列规划教材

# 编　委　会

# 丛书序一 PREFACE

毫无疑问，我们正处在一个新时代。新一轮科技革命和产业变革正在加速推进，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”无疑是第一核心推动力。

当前，发展大数据已经成为国家战略，大数据在引领经济社会发展中的新引擎作用更加突显。大数据重塑了传统产业的结构和形态，催生了众多的新产业、新业态、新模式，推动了共享经济的蓬勃发展，也给我们的衣食住行带来根本改变。同时，大数据是带动国家竞争力整体跃升和跨越式发展的巨大推动力，已成为全球科技和产业竞争的重要制高点。可以大胆预测，未来，大数据将会进一步激起全球科技和产业发展浪潮，进一步渗透到我们国计民生的各个领域，其发展扩张势不可挡。可以说，我们处在一个“大数据”时代。

大数据不仅仅是单一的技术发展领域和战略新兴产业，它还涉及科技、社会、伦理等诸多方面。发展大数据是一个复杂的系统工程，需要科技界、教育界和产业界等社会各界的广泛参与和通力合作，需要我们以更加开放的心态，以进步发展的理念，积极主动适应大数据时代所带来的深刻变革。总体而言，从全面协调可持续健康发展的角度，推动大数据发展需要注重以下五个方面的辩证统一和统筹兼顾。

一是要注重“长与短结合”。所谓“长”就是要目标长远，要注重制定大数据发展的顶层设计和中长期发展规划，明确发展方向和总体目标；所谓“短”就是要着眼当前，注重短期收益，从实处着手，快速起效，并形成效益反哺的良性循环。

二是要注重“快与慢结合”。所谓“快”就是要注重发挥新一代信息技术产业爆炸性增长的特点，发展大数据要时不我待，以实际应用需求为牵引加快推进，力争快速占领大数据技术和产业制高点；所谓“慢”就是防止急功近利，欲速而不达，要注重夯实大数据发展的基础，着重积累发展大数据基础理论与核心共性关键技术，培养行业领域发展中的大数据思维，潜心培育大数据专业人才。

三是要注重“高与低结合”。所谓“高”就是要打造大数据创新发展高地，要结合国家重大战略需求和国民经济主战场核心需求，部署高端大数据公共服务平台，组织开展国家级大数据重大示范工程，提升国民经济重点领域和标志性行业的大数据技术水平和应用能力；所谓“低”就是要坚持“润物细无声”，推进大数据在各行各业和民生领域的广泛应用，推进大数据发展的广度和深度。

四是要注重“内与外结合”。所谓“内”就是要向内深度挖掘和深入研究大数据作为一门学科领域的深刻技术内涵，构建和完善大数据发展的完整理论体系和技术支撑体系；所谓“外”就是要加强开放创新，由于大数据涉及众多学科领域和产业行业门类，也涉及国家、社会、个人等诸多问题，因此，需要推动国际国内科技界、产业界的深入合作和各级政府广泛参与，共同研究制定标准规范，推动大数据与人工智能、云计算、物联网、网络安全等信息技术领域的协同发展，促进数据科学与计算机科学、基础科学和各种应用科学的深度融合。

五是要注重“开与闭结合”。所谓“开”就是要坚持开放共享，要鼓励打破现有体制机制障碍，推动政府建立完善开放共享的大数据平台，加强科研机构、企业间技术交流和合作，推动大数据资源高效利用，打破数据壁垒，普惠数据服务，缩小数据鸿沟，破除数据孤岛；所谓“闭”就是要形成价值链生态闭环，充分发挥大数据发展中技术驱动与需求牵引的双引擎作用，积极运用市场机制，形成技术创新链、产业发展链和资金服务链协同发展的态势，构建大数据产业良性发展的闭环生态圈。

总之，推动大数据的创新发展，已经成为了新时代的新诉求。刚刚闭幕的党的十九大更是明确提出要推动大数据、人工智能等信息技术产业与实体经济深度融合，培育新增长点，为建设网络强国、数字中国、智慧社会形成新动能。这一指导思想为我们未来发展大数据技术和产业指明了前进方向，提供了根本遵循。

习近平总书记多次强调“人才是创新的根基”“创新驱动实质上是人才驱动”。绘制大数据发展的宏伟蓝图迫切需要创新人才培养体制机制的支撑。因此，需要把高端人才队伍建设作为大数据技术和产业发展的重中之重，需要进一步完善大数据教育体系，加强人才储备和梯队建设，将以大数据为代表的新兴产业发展对人才的创新性、实践性需求渗透融入人才培养各个环节，加快形成我国大数据人才高地。

国家有关部门“与时俱进，因时施策”。近期，国务院办公厅正式印发《关于深化产教融合的若干意见》，推进人才和人力资源供给侧结构性改革，以适应创新驱动发展战略的新形势、新任务、新要求。教育部高等学校计算机类专业教学指导委员会、华为公司和人民邮电出版社组织编写的《教育部高等学校计算机类专业教学指导委员会-华为 ICT 产学合作项目——数据科学与大数据技术专业系列规划教材》的出版发行，就是落实国务院文件精神，深化教育供给

侧结构性改革的积极探索和实践。它是国内第一套成专业课程体系规划的数据科学与大数据技术专业系列教材，作者均来自国内一流高校，且具有丰富的大数据教学、科研、实践经验。它的出版发行，对完善大数据人才培养体系，加强人才储备和梯队建设，推进贯通大数据理论、方法、技术、产品与应用等的复合型人才培养，完善大数据领域学科布局，推动大数据领域学科建设具有重要意义。同时，本次产教融合的成功经验，对其他学科领域的人才培养也具有重要的参考价值。

我们有理由相信，在国家战略指引下，在社会各界的广泛参与和推动下，我国的大数据技术和产业发展一定会有光明的未来。

是为序。

中国科学院院士　郑志明

2018 年 4 月 16 日

# 丛书序二 PREFACE

在500年前的大航海时代，哥伦布发现了新大陆，麦哲伦实现了环球航行，全球各大洲从此连接了起来，人类文明的进程得以推进。今天，在云计算、大数据、物联网、人工智能等新技术推动下，人类开启了智能时代。

面对这个以“万物感知、万物互联、万物智能”为特征的智能时代，“数字化转型”已是企业寻求突破和创新的必由之路，数字化带来的海量数据成为企业乃至整个社会最重要的核心资产。大数据已上升为国家战略，成为推动经济社会发展的新引擎，如何获取、存储、分析、应用这些大数据将是这个时代最热门的话题。

国家大数据战略和企业数字化转型成功的关键是培养多层次的大数据人才，然而，根据计世资讯的研究，2018年中国大数据领域的人才缺口将超过150万人，人才短缺已成为制约产业发展的突出问题。

2018年初，华为公司提出新的愿景与使命，即“把数字世界带入每个人、每个家庭、每个组织，构建万物互联的智能世界”，它承载了华为公司的历史使命和社会责任。华为企业BG将长期坚持“平台+生态”战略，协同生态伙伴，共同为行业客户打造云计算、大数据、物联网和传统ICT技术高度融合的数字化转型平台。

人才生态建设是支撑“平台+生态”战略的核心基石，是保持产业链活力和持续增长的根本，华为以ICT产业长期积累的技术、知识、经验和成功实践为基础，持续投入，构建ICT人才生态良性发展的使能平台，打造全球有影响力的ICT人才认证标准。面对未来人才的挑战，华为坚持与全球广大院校、伙伴加强合作，打造引领未来的ICT人才生态，助力行业数字化转型。

一套好的教材是人才培养的基础，也是教学质量的重要保障。本套教材的出版，是华为在大数据人才培养领域的重要举措，是华为集合产业与教育界的高端智力，全力奉献的结晶和成果。在此，让我对本套教材的各位作者表示由衷的感谢！此外，我们还要特别感谢教育部高等学校计算机类专业教学指导委员会副主任、北京大学陈钟教授以及秘书长、北京航空航天大学马殿富教授，没有你们的努力和推动，本套教材无法成型！

同学们、朋友们，翻过这篇序言，开启学习旅程，祝愿在大数据的海洋里，尽情展示你们的才华，实现你们的梦想！

华为公司董事、企业BG总裁　阎力大

2018年5月

# 前言 FOREWORD

大数据已经进入我们社会生活的各个层面，学习、使用大数据成为社会各行各业的共识。掌握大数据技术成为数据科学、计算机科学与技术、软件工程、管理科学与工程等相关领域大数据工作者的一种内在要求。

我们希望本书能结合大学教学的实际情况，向学生介绍大数据技术的基础知识，帮助学生了解大数据技术的概貌。主要内容安排如下。

第 1 章　大数据概述。在介绍目前主流大数据技术前，本章概括介绍了诸如分布式、虚拟化与云计算、数据库与数据仓库等与大数据技术密切相关的概念。

第 2 章　大数据软件基础。考虑到大学授课的特点，本章把在前序课程中可能忽视的 Linux 基础操作、Java 基础和 SQL 语法等与后续大数据实践相关的重点知识作为大数据软件技术基础进行了补充，避免因为学生基础知识的不足而导致学习困难等方面的问题。此外，本章还介绍了如何安装 Linux 集群，为后续章节的内容做铺垫。

第 3 章　大数据存储技术。重点介绍 Hadoop 分布式文件系统 HDFS 以及常见的 NoSQL 数据库，并对 Hadoop 和 HBase 的安装配置及 API 开发进行了介绍。

第 4 章　MapReduce 分布式编程。重点介绍 Hadoop 的 MapReduce 编程及其基本原理。

第 5 章　数据采集与预处理。重点介绍大数据采集与传输数据的工具，包括 Flume、Sqoop 和 Kafka。

第 6 章　数据仓库与联机分析处理。本章首先讨论被业界广泛接受的数据仓库的概念和定义，研究应用于数据仓库和 OLAP 的多维数据模型——数据立方体，然后详细介绍基于 Hadoop 平台的数据仓库工具与相应的联机分析技术，包括 Hive、Kylin 及 Superset 等。

第 7 章　大数据分析与挖掘技术。本章对数据挖掘与分析的基本原理进行讨论，并对 Hadoop 家族中的重要成员——Mahout 进行介绍，描述其在具体应用中的使用方法。

第 8 章　Spark 分布式内存计算框架。本章立足于实战，重点介绍 Spark 的编程模型和 RDD 统一抽象模型、Spark 的工作和调度机制以及以 Spark 为核心衍生的生态系统——SparkSQL、流式计算、机器学习、图计算等，最后对 Zeppelin 数据分析工具进行简要介绍。

第 9 章　数据可视化技术。本章首先简单介绍数据可视化的发展历史、可视化工具分类，然后重点结合 ECharts 介绍 Web 可视化组件生成方法，并给出 JavaWeb 开发与相关大数据组件的数据集成，以展现数据可视化结果。

第 10 章　大数据安全。本章首先介绍大数据安全的挑战与对策，然后结合企业界成熟的华为公司大数据技术安全解决方案，对大数据基础设施安全、安全管理技术、安全分析、隐私保护等内容进行了介绍。

本书的编写得益于华中科技大学软件学院数据科学中心师生的共同努力，其中薛志东负责本书的策划并主要编写了第 2 章、第 3 章、第 4 章、第 5 章和第 9 章；陈长清主要编写了第 1 章；吕泽华主要编写了第 6 章、第 7 章和第 8 章；黄浩主要编写了第 10 章。此外，姚益阳、杜海朋、董英豪、卢璟祥、张双双、邹小威、张学清、郭映中、汪元也参加了本书部分内容的编写工作。曾辉、余晨晨、奉俊丰参加了本书部分代码的整理工作。

在本书的编写过程中，编者参考、引用了华为技术有限公司 ICT 学院提供的资料、相关技术的官方文档和大量互联网资源，在此向有关单位、作者表示感谢，并尽量在参考文献部分一一列出，若有遗漏和不妥之处，敬请相关作者指正。

感谢华为技术有限公司刘洁、张志峰，华中科技大学软件学院陈传波教授、肖来元教授、沈刚教授，以及陈维亚博士、区士颀博士、石强博士对图书编写工作予以的支持与帮助。

由于时间仓促，编者水平有限，书中难免存在不足之处，敬请读者批评指正。

编者

2018 年 5 月于华中科大软件学院

# 目 录 CONTENTS


**第 1 章 大数据概述 ····· 1**

1.1 大数据的相关概念 ····· 2
1.2 大数据处理的基础技术 ····· 4
1.2.1 大数据处理流程 ····· 4
1.2.2 分布式计算 ····· 5
1.2.3 分布式文件系统 ····· 6
1.2.4 分布式数据库 ····· 7
1.2.5 数据库与数据仓库 ····· 8
1.2.6 云计算与虚拟化技术 ····· 8
1.2.7 虚拟化产品介绍 ····· 9
1.3 流行的大数据技术 ····· 12
1.4 大数据解决方案 ····· 17
1.5 大数据发展现状和趋势 ····· 19
1.5.1 大数据现状分析 ····· 19
1.5.2 大数据发展趋势 ····· 21
1.6 教学建议及教辅资料 ····· 22
习题 ····· 23

**第 2 章 大数据软件基础 ····· 24**

2.1 Linux 基础 ····· 25
2.1.1 Linux 简介 ····· 25
2.1.2 Linux 基本操作 ····· 25
2.1.3 网络配置管理 ····· 29
2.1.4 其他常用网络命令 ····· 32
2.2 Java 基础 ····· 34
2.2.1 面向对象与泛型 ····· 34
2.2.2 集合类 ····· 36
2.2.3 内部类与匿名类 ····· 37
2.2.4 反射 ····· 38
2.3 SQL 语言基础 ····· 39
2.4 在 VirtualBox 上安装 Linux 集群 ····· 41
2.4.1 master 节点的安装 ····· 41
2.4.2 配置 Virtualbox 网络及虚拟机网卡 ····· 49
2.4.3 slave 节点的安装与配置 ····· 51
2.4.4 Java 环境的安装 ····· 51
2.4.5 MySQL 服务 ····· 52
2.4.6 SSH 免密钥登录 ····· 53
2.4.7 配置时钟同步 ····· 55
习题 ····· 56

**第 3 章 大数据存储技术 ····· 57**

3.1 理解 HDFS 分布式文件系统 ····· 58
3.1.1 HDFS 简介 ····· 58
3.1.2 HDFS 的体系结构 ····· 59
3.1.3 HDFS 中的数据流 ····· 62
3.2 NoSQL 数据库 ····· 66
3.2.1 键值数据库 Redis ····· 66
3.2.2 列存储数据库 HBase ····· 68
3.2.3 文档数据库 MongoDB ····· 71
3.2.4 图数据库 Neo4j ····· 73
3.3 Hadoop 的安装与配置 ····· 74
3.3.1 Hadoop 的配置部署 ····· 75
3.3.2 启动 Hadoop 集群 ····· 79
3.4 HDFS 文件管理 ····· 82
3.4.1 命令行访问 HDFS ····· 82
3.4.2 使用 Java API 访问 HDFS ····· 84
3.5 HBase 的安装与配置 ····· 88
3.5.1 解压并安装 HBase ····· 88
3.5.2 配置 HBase ····· 88
3.6 HBase 的使用 ····· 91
3.6.1 HBase-shell ····· 91
3.6.2 Java API ····· 94
习题 ····· 96

## 第4章 MapReduce分布式编程……97

4.1 MapReduce编程概述……98
4.2 MapReduce编程示例……98
4.2.1 词频统计程序示例……99
4.2.2 MapReduce编译与运行……101
4.3 深入理解MapReduce程序的运行过程……102
4.4 MapReduce任务调度框架……104
4.4.1 经典MapReduce任务调度模型……104
4.4.2 YARN框架原理及运行机制……105
4.5 MapReduce的数据类型与输入/输出格式……107
4.5.1 MapReduce的数据类型……107
4.5.2 MapReduce的文件输入/输出格式……107
4.6 MapReduce编程实例……111
4.6.1 视频类型统计……111
4.6.2 查询TOP10用户上传的视频列表……113
习题……118

## 第5章 数据采集与预处理……119

5.1 流数据采集工具Flume……120
5.1.1 Flume的安装……121
5.1.2 Flume的配置与运行……122
5.1.3 Flume源……124
5.1.4 Flume槽……127
5.1.5 通道、拦截器与处理器……129
5.2 数据传输工具Sqoop……130
5.2.1 Sqoop的安装……131
5.2.2 Sqoop的配置与运行……131
5.2.3 Sqoop实例……132
5.2.4 Sqoop导入过程……135
5.2.5 Sqoop导出过程……136
5.3 数据接入工具Kafka……136
5.3.1 Kafka的安装与配置……138
5.3.2 Kafka消息生产者……140
5.3.3 Kafka消息消费者……140
5.3.4 Kafka核心特性……141
习题……142

## 第6章 数据仓库与联机分析处理……143

6.1 数据仓库……144
6.1.1 数据仓库的概念……144
6.1.2 数据仓库与操作性数据库的区别……144
6.1.3 数据仓库的体系结构……145
6.2 多维数据模型……146
6.2.1 数据立方体……146
6.2.2 数据模型……147
6.2.3 多维数据模型中的OLAP操作……150
6.3 Hive……153
6.3.1 Hive简介……153
6.3.2 Hive的安装与配置……154
6.3.3 Hive使用……156
6.3.4 Hive导入数据实例……161
6.4 Kylin……164
6.4.1 Kylin简介……164
6.4.2 Kylin的安装与配置……165
6.4.3 Kylin的使用……168
6.5 Superset……175
6.5.1 Superset简介……175
6.5.2 Superset的安装与配置……175
6.5.3 Superset的使用……177
习题……186

## 第7章 大数据分析与挖掘技术……187

7.1 概述……188
7.1.1 数据挖掘简介……188

7.1.2 Mahout 的安装与配置……189
7.2 推荐……192
7.2.1 推荐的定义与评估……192
7.2.2 Mahout 中的常见推荐算法……194
7.2.3 对 GroupLens 数据集进行推荐与评价……196
7.3 聚类……198
7.3.1 聚类的基本概念……198
7.3.2 常见的 Mahout 数据结构……199
7.3.3 几种聚类算法……200
7.3.4 聚类应用实例……202
7.4 分类……206
7.4.1 分类的基本概念……206
7.4.2 Mahout 中一些常见的训练分类器算法……208
7.4.3 应用实例：使用 SGD 训练分类器对新闻分类……210
习题……213
第 8 章 Spark 分布式内存计算框架……214
8.1 Spark 简介……215
8.2 Spark 的编程模型……216
8.2.1 核心数据结构 RDD……216
8.2.2 RDD 上的操作……216
8.2.3 RDD 的持久化……218
8.2.4 RDD 计算工作流……218
8.3 Spark 的调度机制……219
8.3.1 Spark 分布式架构……219
8.3.2 Spark 应用执行流程……220
8.3.3 Spark 调度与任务分配……222
8.4 Spark 应用案例……225
8.4.1 Spark Shell……225
8.4.2 单词计数……227
8.4.3 统计用户的视频上传数……229
8.4.4 查询 Top100 用户的上传视频列表……230
8.5 Spark 生态圈其他技术……233
8.5.1 Spark SQL……233
8.5.2 Spark Streaming……235
8.5.3 MLlib……236
8.5.4 GraphX……242
8.6 Zeppelin：交互式分析 Spark 数据……243
8.6.1 Zeppelin 简介……243
8.6.2 安装和启动……244
8.6.3 在 Zeppelin 中处理 YouTube 数据……244
习题……246
第 9 章 数据可视化技术……247
9.1 数据可视化概述……248
9.2 数据可视化工具……249
9.2.1 桌面可视化技术……249
9.2.2 OLAP 可视化工具……251
9.2.3 Web 可视化技术……251
9.3 可视化组件与 ECharts 示例……253
9.3.1 ECharts 使用准备……253
9.3.2 ECharts 示例……254
9.4 与大数据平台集成……268
9.4.1 获取对 Hive 数据库的连接……268
9.4.2 通过 Java 调用 Hive 提供的 API 操作数据……269
9.4.3 将数据提交到 Web 页面进行数据可视化……271
习题……272
第 10 章 大数据安全……273
10.1 大数据安全的挑战与对策……274
10.1.1 大数据安全与隐私的挑战……274
10.1.2 数据加密技术……275
10.1.3 大数据安全保障体系……275
10.1.4 华为大数据安全解决方案……276
10.2 基础设施安全……277

10.2.1 认证技术……278
10.2.2 访问控制……279
10.2.3 公钥基础设施……281
10.2.4 华为大数据平台……281
10.3 数据管理安全……285
10.3.1 数据溯源……285
10.3.2 数字水印……285
10.3.3 策略管理……287
10.3.4 完整性保护……287
10.3.5 数据脱敏……288
10.4 安全分析……290
10.4.1 大数据安全分析架构……290
10.4.2 大数据防 DDoS 攻击……292
10.4.3 攻击可视化与安全业务定制……297
10.5 隐私保护……298
10.5.1 隐私保护面临的挑战……298
10.5.2 内容关联密钥……298
10.5.3 华为大数据隐私保护方案……300
习题……302
附录 《大数据技术基础》配套实验课程方案简介……303
参考文献……304

# 01 第1章 大数据概述

大数据技术已经进入我们社会生活的各个层面，我们不仅在消费大数据，也在源源不断地产生大数据。数以亿计的移动互联网用户将位置、微博、朋友圈、打车、外卖、邮件、网购、社交等信息源源不断地上传到服务商的服务器上，而这些服务商也非常乐意为用户保存各种信息，因为他们意识到了这些数据的价值。与此同时，各行各业都受到大数据的影响，涌现出了诸如工业大数据、金融大数据、环境大数据、医疗健康大数据、教育大数据等，人们开始结合行业与领域的特点和优势，通过大数据技术进行领域改进和升级。在实践中，人们也逐渐对大数据的概念、价值、范围有了清醒的认识，应对各种需求的大数据技术也逐渐走向成熟，大数据处理技术体系也越来越完备。

本章主要介绍大数据的基本概念、相关技术和目前的应用现状。

# 1.1 大数据的相关概念

大数据是指在一定时间内无法用常规软件工具对其内容进行抓取、处理、分析和管理的数据集合。大数据一般会涉及两种以上的数据形式，数据量通常是 100TB 以上的高速、实时数据流，或者从每年增长速度快的小数据开始。

1. 大数据的特征

大数据有 4 个特性，简称 4V：Volume、Variety、Velocity、Value，如图 1.1 所示。


图 1.1 大数据的 4V 特征

（1）Volume（规模性）：大数据的特征首先体现为“数据量大”，存储单位从过去的 GB 到 TB，直至 PB、EB。随着网络及信息技术的高速发展，数据开始爆发性增长。社交网络、移动网络、各种智能终端等，都成为数据的来源，企业也面临着数据量的大规模增长，IDC 的一份报告预测称，到 2020 年，全球数据量将扩大 50 倍。此外，各种意想不到的来源都能产生数据。

（2）Variety（多样性）：一个普遍观点认为，人们使用互联网搜索是形成数据多样性的主要原因，这一看法部分正确。大数据大体可分为三类：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。

（3）Velocity（高速性）：数据被创建和移动的速度快。在网络时代，通过高速的计算机和服务器，创建实时数据流已成为流行趋势。企业不仅需要了解如何快速创建数据，还必须知道如何快速处理、分析并返回给用户，以满足他们的实时需求。

（4）Value（价值性）：相比于传统的小数据，大数据最大的价值在于通过从大量不相关的各种类型的数据中，挖掘出对未来趋势与模式预测分析有价值的数据，并通过机器学习方法、人工智能方法或数据挖掘方法进行深度分析，发现新规律和新知识，并运用于农业、金融、医疗等各个领域，从而最终达到改善社会治理、提高生产效率、推进科学研究的效果。

2. 大数据的构成

大数据分为结构化数据、非结构化数据和半结构化数据三种，如图 1.2 所示。结构化数据是指信息经过分析后可分解成多个互相关联的组成部分，各组成部分间有明确的层次结构，其使用和维护通过数据库进行管理，并有一定的操作规范。通常，信息系统涉及生产、业务、交易、客户等方面的数据，采用结构化方式存储。一般来讲，结构化数据只占全部数据的 20%以内，但是就是这 20%以内的数据浓缩了很久以来企业各个方面的数据需求，发展也已经成熟。而无法完全数字化的文档文件、图片、图纸资料、缩微胶片等信息就属于非结构化数据，非结构化数据中往往存在大量的有价值的信息，特别是随着移动互联网、物联网的发展，非结构化数据正以成倍速度快速增长。

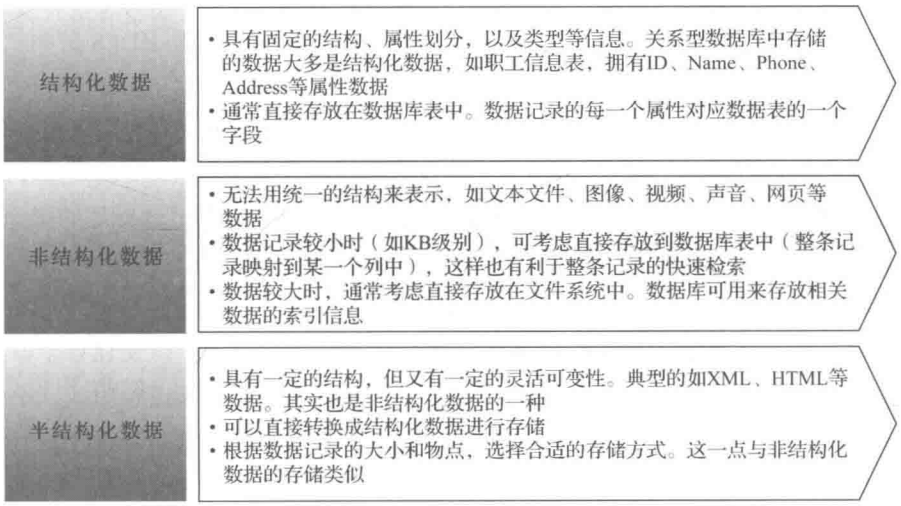



图 1.2 三种数据结构的简单总结

（1）结构化数据

结构化数据是由二维表结构来逻辑表达和实现的数据，也称作行数据，严格地遵循数据格式与长度规范，有固定的结构、属性划分和类型等信息，主要通过关系型数据库进行存储和管理，数据记录的每一个属性对应数据表的一个字段。

（2）非结构化数据

与结构化数据相对的是不适于由数据库二维表来表现的非结构化数据，包括所有格式的办公文档、各类报表、图片和音频、视频信息等。在数据较小的情况下，可以使用关系型数据库将其直接存储在数据库表的多值字段和变长字段中；若数据较大，则存放在文件系统中，数据库则用于存放相关文件的索引信息。这种方法广泛应用于全文检索和各种多媒体信息处理领域。

（3）半结构化数据

半结构化数据既具有一定的结构，又灵活多变，其实也是非结构化数据的一种。和普通纯文本、图片等相比，半结构化数据具有一定的结构性，但和具有严格理论模型的关系数据库的数据相比，其结构又不固定。如员工简历，处理这类数据可以通过信息抽取、转换等步骤，将其转化为半结构化数据，采用 XML、HTML 等形式表达；或者根据数据的大小，采用非结构化数据存储方式，结合关系数据存储。

随着大数据技术的发展，对非结构化数据的处理越来越重要。据 IDC 的一项调查报告显示，