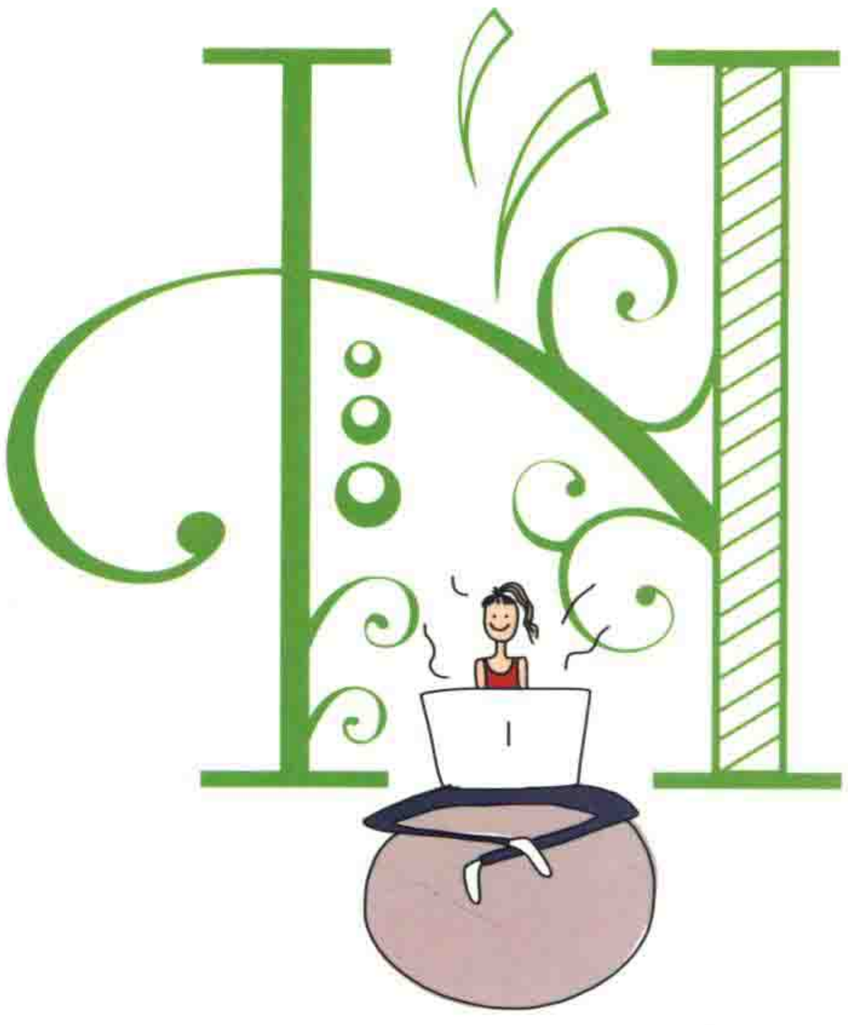
10余位大数据领域资深专家和科研人员10余年大数据挖掘的经验结晶，配套教学材料和上机实验，多所大学选为教材

系统讲解Hadoop、Hive、HBase、Pig、Spark、Oozie等大数据技术，每个知识点都配套实用的企业应用案例，可操作性极强



Hadoop and Big Data Minning

Hadoop与大数据挖掘

张良均 樊哲 位文超 刘名军◎等著

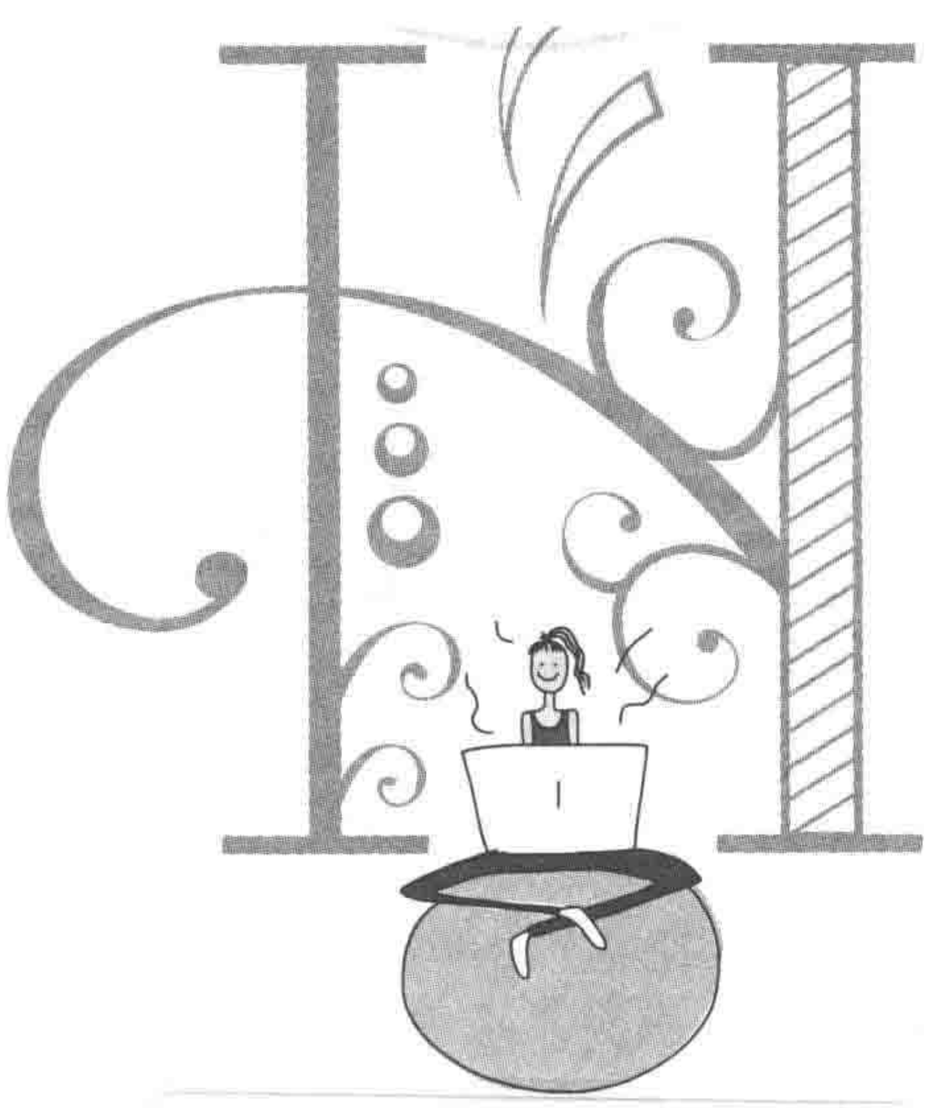
机械工业出版社
China Machine Press

大数据技术丛书

Hadoop and Big Data Minning

Hadoop与大数据挖掘

张良均 樊哲 位文超 刘名军◎等著



机械工业出版社
China Machine Press

图书在版编目（CIP）数据

Hadoop 与大数据挖掘 / 张良均等著 . —北京：机械工业出版社，2017.5
（大数据技术丛书）

ISBN 978-7-111-56787-5

I. H…　II. 张…　III. ① 数据处理软件　② 数据采集　IV. TP274

中国版本图书馆 CIP 数据核字（2017）第 090074 号

Hadoop 与大数据挖掘

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号　邮政编码：100037）
责任编辑：何欣阳　　责任校对：殷　虹
印　　刷：北京文昌阁彩色印刷有限责任公司　　版　　次：2017 年 5 月第 1 版第 1 次印刷
开　　本：186mm × 240mm　1/16　　印　　张：20.75
书　　号：ISBN 978-7-111-56787-5　　定　　价：69.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88379426　88361066　　投稿热线：（010）88379604
购书热线：（010）68326294　88379649　68995259　　读者信箱：hzit@hzbook.com

华章IT
HZBOOKS | Information Technology

Preface 前　言

为什么要写这本书

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。”

早在 2012 年，大数据（big data）一词已经被广泛提起，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。那时就有人预计，从 2013 年至 2020 年，全球数据规模将增长 10 倍，每年产生的数据量将由当时的 4.4 万亿 GB，增长至 44 万亿 GB，每两年翻一番。

既然“大数据”浪潮已经来临，那么与之对应的大数据人才呢？在国外，大数据技术发展正如火如荼，各种方便大家学习的资料、教程应有尽有。但是，在国内，这种资料却是有“门槛”的。其一，这类资料是英文的，对于部分人员来说，阅读是有难度的；其二，这些资料对于初学者或在校生来说，在理论理解上也有一些难度，没有充分的动手实践来协助理解大数据相关技术的原理、架构等；其三，在如何应用大数据技术来解决企业实实在在遇到的大数据相关问题方面，没有很好的资料；其四，对于企业用户来说，如何将大数据技术和数据挖掘技术相结合，对企业大量数据进行挖掘，以挖掘出有价值的信息，也是难点。

作为大数据相关技术，Hadoop 无疑应用很广泛。Hadoop 具有以下优势：高可靠性、高扩展性、高效性、高容错性、低成本、生态系统完善。

一般来说，使用 Hadoop 相关技术可以解决企业相关大数据应用，特别是结合诸如 Mahout、Spark MLlib 等技术，不仅可以对企业相关大数据进行基础分析，还能构建挖掘模型，挖掘企业大数据中有价值的信息。

对于学习大数据相关技术的高校师生来说，本书不仅提供了大数据相关技术的基础讲解及原理、架构分析，还针对这些原理，配备有对应的动手实践章节，帮助读者加深对原

理、架构的认识。同时，在每个模块结束后，书中会有一个相对独立的企业应用案例，帮助读者巩固学到的大数据技术相关知识。

对于企业用户或大数据挖掘开发者来说，特别是对想要了解如何将大数据技术应用到企业大数据项目中的企业用户或者开发者来说，本书也是一份优秀的参考资料。

本书特色

本书提供了大数据相关技术的简介、原理、实践、企业应用等，针对大数据相关技术，如 Hadoop、HBase、Hive、Spark 等，都有专业章节进行介绍，并且针对每一模块都有相应的动手实践，能有效加深读者对大数据相关技术原理、技术实践的理解。书中的挖掘实践篇涉及企业在大数据应用中的所有环节，如数据采集、数据预处理、数据挖掘等，通过案例对整个系统的架构进行了详细分析，对读者有一定实践指导作用。

读者可以从“泰迪杯”全国大学生数据挖掘挑战赛网站（http://www.tipdm.org/tj/865.jhtml）免费下载本书配套的全部数据文件及源程序。另外，为方便教师授课，本书还特意提供了建模阶段的过程数据文件、PPT 课件，有需要的教师可通过热线电话（40068-40020）、企业 QQ（40068-40020）或以下微信公众号咨询获取。

Tip DM

张良均 < 大数据挖掘产品与服务 >

本书适用对象

❑ 开设大数据、大数据挖掘相关课程的高校教师和学生

目前国内不少高校将大数据、大数据挖掘引入本科教学中，在计算机、数学、自动化、电子信息、金融等专业开设了大数据技术相关的课程，但目前针对这一课程的相关教材没有统一，或者使用的教材不利于课堂教学。本书提供了大数据相关技术的简介、原理、实践、企业应用等，能有效帮助高校教师教学；帮助学生学习大数据相关技术原理，进行技术实践，为以后工作打下良好基础。

❑ 大数据开发人员

书中针对大数据相关技术，如 Hadoop、HBase、Hive、Spark 等，都有专业章节进行介绍，并且针对每一模块有相应的动手实践，对初级开发人员有较强指导作用。

❑ 大数据架构师

挖掘实践篇涉及企业在大数据应用中的所有环节，包括数据采集、数据预处理、数据挖掘等方面，通过案例对整个系统的架构进行了详细分析，对大数据架构师有一定的实践指导作用。

❑ 关注大数据挖掘技术的人员

本书不仅包括大数据相关技术的简介及原理分析，还包括大数据相关技术和大数据挖掘相结合的案例分析。对于大数据挖掘技术人员来说，如何应用大数据技术来对大数据进行挖掘是重点和难点，通过学习本书中案例的分析方法，可以将其融入自己的实际工作中。

如何阅读本书

本书主要分为两篇：基础篇和挖掘实战篇。基础篇介绍了大数据相关技术：Hadoop、Hive、HBase、Pig、Spark、Oozie 等。针对每个技术都有相应模块与之对应，首先对该技术的概念、内部原理等进行介绍，使读者对该技术有一个由浅入深的理解；其次在对原理的介绍中会配合相应的动手实践，加深对原理的理解。在每个模块的最后，会有 1～2 个企业案例，主要讲解使用当前模块的技术来解决其中的 1～2 个问题，这样读者不仅对技术的原理、架构有了较深入的了解，同时，对于如何应用该技术也有了一定认识，从而为以后的工作、学习打下良好基础。挖掘实战篇通过对一个大型的企业应用案例的介绍，充分应用基础篇讲解的大数据技术来解决企业应用中遇到的各种问题。本书配套提供了程序代码及数据，读者可通过上机实验，快速掌握书中所介绍的大数据相关技术，获得使用大数据相关技术进行数据挖掘的基本能力。

第一篇是基础篇（第 1～7 章）。第 1 章主要介绍了大数据相关概念，以及大数据相关技术。第 2 章对 Hadoop 进行了介绍，包括概念、原理、架构等，通过动手实践案例帮助读者加深对原理的理解。第 3 章对 Hive 进行了介绍，重点分析了 Hive 的架构及如何与 Hadoop 相结合，同时，引入一个企业案例来分析 Hive 在企业应用中的地位。第 4 章对 HBase 进行了介绍，分析了 HDFS 与 HBase 的异同点、HBase 架构原理、HBase 如何做到支持随机读写等。第 5 章介绍了 Pig，详细分析了 Pig 的实现原理及应用场景，介绍了 Pig Latin，并且通过一个 Pig Latin 的动手实践案例，加深读者对该脚本的理解。第 6 章介绍了 Spark 的基本原理、RDD 实现等，并且对 Scala 进行了简单介绍，使用 Scala 创建 Wordcount 程序，在模块的最后使用 Spark MLlib 完成引入的企业案例中的模型建立环节。第 7 章介绍了 Hadoop 工作流 Oozie，通过动手实际建立 Hadoop MR、Spark、Hive、Pig 的工作流，方便理解企业工作流应用。

第二篇是挖掘实战篇（第 8 章），详细介绍了一个企业级大数据应用项目——法律服务大数据智能推荐系统。通过分析应用背景、构建系统，使读者了解针对系统的每一层应使用什么大数据技术来解决问题。涉及的流程有数据采集、数据预处理、模型构建等，在每一个流程中会进行大数据相关技术实践，运用实际数据来进行分析，使读者切身感受到大

数据技术解决大数据企业应用的魅力。

勘误和支持

除封面署名外，参加本书编写工作的还有周龙、焦正升、许国杰、杨坦、肖刚、刘晓勇等。由于作者的水平有限，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。本书内容的更新将及时在“泰迪杯”全国数据挖掘挑战赛网站（www.tipdm.com）上发布。读者可通过作者微信公众号 TipDM（微信号：TipDataMining）、TipDM 官网（www.tipdm.com）反馈有关问题。也可通过热线电话（40068-40020）或企业 QQ（40068-40020）进行在线咨询。

如果你有更多宝贵意见，欢迎发送邮件至邮箱 13560356095@qq.com，期待能够得到你的真挚反馈。

致谢

本书编写过程中得到了广大企事业单位科研人员的大力支持，在此谨向中国电力科学研究院、广东电力科学研究院、广西电力科学研究院、华南师范大学、广东工业大学、广东技术师范学院、南京中医药大学、华南理工大学、湖南师范大学、韩山师范学院、中山大学、广州泰迪智能科技有限公司等单位给予支持的专家及师生致以深深的谢意。

在本书的编辑和出版过程中还得到了参与“泰迪杯”全国数据挖掘建模竞赛的众多师生及机械工业出版社杨福川老师、李艺编辑的大力帮助与支持，在此一并表示感谢。

张良均

Contents 目　录

前言

第一篇　基础篇

第 1 章　浅谈大数据……2

1.1　大数据概述……3

1.2　大数据平台……4

1.3　本章小结……5

第 2 章　大数据存储与运算利器——Hadoop……6

2.1　Hadoop 概述……6

2.1.1　Hadoop 简介……6

2.1.2　Hadoop 存储——HDFS……8

2.1.3　Hadoop 计算——MapReduce……11

2.1.4　Hadoop 资源管理——YARN……13

2.1.5　Hadoop 生态系统……14

2.2　Hadoop 配置及 IDE 配置……17

2.2.1　准备工作……17

2.2.2　环境配置……18

2.2.3　集群启动关闭与监控……24

2.2.4　动手实践：一键式 Hadoop 集群启动关闭……25

2.2.5　动手实践：Hadoop IDE 配置……26

2.3　Hadoop 集群命令……28

2.3.1　HDFS 常用命令 hdfs dfs……30

2.3.2　动手实践：hdfs dfs 命令实战……31

2.3.3　MapReduce 常用命令 mapred job……32

2.3.4　YARN 常用命令 yarn jar……32

2.3.5　动手实践：运行 MapReduce 任务……33

2.4　Hadoop 编程开发……33

2.4.1　HDFS Java API 操作……33

2.4.2　MapReduce 原理……35

2.4.3　动手实践：编写 Word Count 程序并打包运行……44

2.4.4　MapReduce 组件分析与编程实践……46

2.5　K-Means 算法原理及 Hadoop MapReduce 实现……53

2.5.1　K-Means 算法原理……53

2.5.2　动手实践：K-Means 算法实现……55

2.5.3　Hadoop K-Means 算法实现思路……55

2.5.4 Hadoop K-Means 编程实现……57
2.6 TF-IDF 算法原理及 Hadoop MapReduce 实现……67
2.6.1 TF-IDF 算法原理……67
2.6.2 Hadoop TF-IDF 编程思路……67
2.6.3 Hadoop TF-IDF 编程实现……68
2.7 本章小结……79

第 3 章 大数据查询——Hive……81

3.1 Hive 概述……81
3.1.1 Hive 体系架构……82
3.1.2 Hive 数据类型……86
3.1.3 Hive 安装……87
3.1.4 动手实践：Hive 安装配置……91
3.1.5 动手实践：HiveQL 基础——SQL……91
3.2 HiveQL 语句……93
3.2.1 数据库操作……94
3.2.2 Hive 表定义……94
3.2.3 数据导入……100
3.2.4 数据导出……103
3.2.5 HiveQL 查询……104
3.3 动手实践：基于 Hive 的学生信息查询……108
3.4 基于 Hive 的航空公司客户价值数据预处理及分析……109
3.4.1 背景与挖掘目标……109
3.4.2 分析方法与过程……111
3.5 本章小结……115

第 4 章 大数据快速读写——HBase…116

4.1 HBase 概述……116
4.2 配置 HBase 集群……118
4.2.1 Zookeeper 简介及配置……118
4.2.2 配置 HBase……121
4.2.3 动手实践：HBase 安装及运行…122
4.2.4 动手实践：ZooKeeper 获取 HBase 状态……122
4.3 HBase 原理与架构组件……123
4.3.1 HBase 架构与组件……123
4.3.2 HBase 数据模型……127
4.3.3 读取 / 写入 HBase 数据……128
4.3.4 RowKey 设计原则……129
4.3.5 动手实践：HBase 数据模型验证……131
4.4 HBase Shell 操作……132
4.4.1 HBase 常用 Shell 命令……132
4.4.2 动手实践：HBase Shell 操作…136
4.5 Java API &MapReduce 与 HBase 交互……137
4.5.1 搭建 HBase 开发环境……137
4.5.2 使用 Java API 操作 HBase 表…144
4.5.3 动手实践：HBase Java API 使用……147
4.5.4 MapReduce 与 HBase 交互……147
4.5.5 动手实践：HBase 表导入导出…150
4.6 基于 HBase 的冠字号查询系统……151
4.6.1 案例背景……151
4.6.2 功能指标……151
4.6.3 系统设计……152
4.6.4 动手实践：构建基于 HBase 的冠字号查询系统……162
4.7 本章小结……175

第 5 章 大数据处理——Pig……176

5.1 Pig 概述……176

5.1.1 Pig Latin 简介 177
5.1.2 Pig 数据类型 179
5.1.3 Pig 与 Hive 比较 179
5.2 配置运行 Pig 180
5.2.1 Pig 配置 181
5.2.2 Pig 运行模式 181
5.3 常用 Pig Latin 操作 182
5.3.1 数据加载 182
5.3.2 数据存储 184
5.3.3 Pig 参数替换 185
5.3.4 数据转换 186
5.4 综合实践 194
5.4.1 动手实践：访问统计信息数据处理 194
5.4.2 动手实践：股票交易数据处理 195
5.5 本章小结 196

第 6 章 大数据快速运算与挖掘——Spark 197

6.1 Spark 概述 197
6.2 Spark 安装集群 199
6.2.1 3 种运行模式 199
6.2.2 动手实践：配置 Spark 独立集群 199
6.2.3 3 种运行模式实例 201
6.2.4 动手实践：Spark Streaming 实时日志统计 205
6.2.5 动手实践：Spark 开发环境——Intellij IDEA 配置 207
6.3 Spark 架构与核心原理 212
6.3.1 Spark 架构 212
6.3.2 RDD 原理 213
6.3.3 深入理解 Spark 核心原理 215
6.4 Spark 编程技巧 218
6.4.1 Scala 基础 218
6.4.2 Spark 基础编程 218
6.5 如何学习 Spark MLlib 225
6.5.1 确定应用 227
6.5.2 ALS 算法直观描述 228
6.5.3 编程实现 229
6.5.4 问题解决及模型调优 233
6.6 动手实践：基于 Spark ALS 电影推荐系统 234
6.6.1 动手实践：生成算法包 235
6.6.2 动手实践：完善推荐系统 239
6.7 本章小结 250

第 7 章 大数据工作流——Oozie 252

7.1 Oozie 简介 252
7.2 编译配置并运行 Oozie 253
7.2.1 动手实践：编译 Oozie 253
7.2.2 动手实践：Oozie Server/client 配置 254
7.3 Oozie WorkFlow 实践 257
7.3.1 定义及提交工作流 257
7.3.2 动手实践：MapReduce WorkFlow 定义及调度 260
7.3.3 动手实践：Pig WorkFlow 定义及调度 263
7.3.4 动手实践：Hive WorkFlow 定义及调度 265
7.3.5 动手实践：Spark WorkFlow 定义及调度 267
7.3.6 动手实践：Spark On Yarn 定义及调度 268

7.4 Oozie Coordinator 实践 ····· 270
7.4.1 动手实践：基于时间调度 ····· 270
7.4.2 动手实践：基于数据有效性调度 ····· 273
7.5 本章小结 ····· 275

第二篇 挖掘实战篇

第 8 章 法律服务大数据智能推荐 ····· 278
8.1 背景 ····· 278
8.2 目标 ····· 279
8.3 系统架构及流程 ····· 279
8.4 分析过程及实现 ····· 281
8.4.1 数据传输 ····· 281
8.4.2 数据传输：动手实践 ····· 282
8.4.3 数据探索分析 ····· 283
8.4.4 数据预处理 ····· 292
8.4.5 模型构建 ····· 297
8.5 构建法律服务大数据智能推荐系统 ····· 313
8.5.1 动手实践：构建推荐系统 JavaEE ····· 313
8.5.2 动手实践：Oozie 工作流任务 ··· 317
8.6 本章小结 ····· 322

第一篇 *Part 1*

基 础 篇

Chapter 1

第1章 浅谈大数据

当你早上起床，拿起牙刷刷牙，你是否会想到从拿起牙刷到刷完牙的整个过程中有多少细胞参与其中？这些细胞在参与的过程中会结合周围环境（可能是宏观的天气、温度、气压等，可能是微观的分子、空气中的微生物等），由你的意识控制而产生不同的反映。如果我说结合这些所有的信息，可以预测你接下来的 0.000 000 01 秒的动作，那么，你肯定说，这我也可以预测呀。比如正常情况下，你脚抬起来走路，那么抬起来后，肯定是要落下去的，这算哪门子预测呢？那如果我说可以预测你接下来一个小时的动作呢？甚至一天，一个月，一年呢？其实这也可以勉强说是一个大数据案例了。

听起来有点夸张？

说个大家熟悉的大数据吧。相信很多人都买过股票（或者至少知道买股票这件事情），如果有人可以整合所有信息（包含基本的股票信息：股票涨跌；公司情况：如公司大小、业务等；政策情况：可能政府突然颁布了一个红头文件等），首先肯定这些信息可以被认为是“大数据”，其次对这些“大数据”进行分析建模，如果可以预测股票的涨跌，那么这就是一个实实在在的大数据案例了。

再说一个电影桥段：“赌神”一般都可以预测摇色子的点数或者说摇色子摇到的最大点数，那么在现实情况中，这个可能实现吗？试想这样一个场景：一个人不停地摇色子，然后把摇色子的声音以及最后的点数记录下来，不停地摇，不停地记录，那么就会形成一个巨大的数据集，从而可以使用这个巨大的数据集进行建模，即可以预测色子的点数了。你也可以将这个理解为一个大数据的应用。

现在，你是否已经有点懂“大数据”了？

1.1　大数据概述

来看看所谓官网定义的大数据：大数据（Big data）或称巨量数据、海量数据、大资料，指的是所涉及的数据量规模巨大到无法通过人工或者计算机，在合理的时间内达到截取、管理、处理并整理成为人类所能解读的形式的信息。

看得懂吗？好像也不是那么难以理解。首先，这些数据要够多，即规模巨大；第二，这些数据不能够在合理的时间内被处理并分析，也就意味着，对于一个人来说，如果让他在 1 天内看完 1 万本书，并写相应的书评，那么这 1 万本书对于这个人来说就是大数据；但是，如果让 1 万个人在 1 天内看 1 万本书，并写对应书评，那么其实是可以完成的任务，这样这 1 万本书对于这 1 万个人来说就不是大数据了。

大数据有哪些特点呢？

首先，可以肯定的是数据量比较大，它才能被称为大数据，所以其第一个特点就是数据体量巨大。其次，数据的类型多样也是大数据的一个特征，数据类型不仅指文本形式，更多指的是图片、视频、音频、地理位置信息等多类型的数据，个性化数据占绝大多数。第三，处理速度快也是大数据的一个特征，数据处理遵循“1 秒定律”，可从各种类型的数据中快速获得高价值的信息。最后，大数据具有价值密度低的特点，以视频为例，1 小时的监控视频，在不间断的监控过程中，可能有用的数据仅仅只有一两秒。

生活中大数据有哪些应用呢？

随着大数据的应用越来越广泛，应用的行业也越来越多，我们每天都可以看到大数据的一些新奇的应用，从而帮助人们从中获取到真正有用的价值信息。

（1）理解客户，满足客户服务需求

大数据的应用目前在这个领域是最广为人知的。重点是如何应用大数据更好地了解客户以及他们的爱好和行为。企业非常喜欢搜集社交方面的数据、浏览器的日志、分析文本和传感器的数据，从而更加全面地了解客户。在一般情况下，企业会采用建立数据模型的方式进行预测。

比如美国的著名零售商 Target 就是通过大数据分析得到有价值的信息，精准地预测到客户在什么时候想要小孩。再比如，通过大数据应用，电信公司可以更好地预测出流失的客户，沃尔玛则更加精准地预测出哪个产品会大卖，汽车保险行业会更加了解客户的需求和驾驶水平，外国候选政党也能了解到选民的偏好。

（2）提高医疗水平和研发效率

大数据分析应用的计算能力可以让我们能够在几分钟内解码整个 DNA，并且制定出最新的治疗方案，同时更好地了解和预测疾病。大数据技术目前已经在医疗中应用，如监视早产婴儿和患病婴儿的情况，通过记录和分析婴儿的心跳，对婴儿的身体可能出现的不适症状做出预测，从而更好地救治婴儿。

（3）改善安全和执法

目前来说，大数据已经广泛应用到安全执法的过程当中。想必大家都知道美国安全局已经开始利用大数据打击恐怖主义，甚至监控可疑人的日常生活。而企业则应用大数据技术防御网络攻击，警察应用大数据工具捕捉罪犯，信用卡公司应用大数据工具来检测欺诈性交易等。

（4）改善我们的城市

大数据还被用来改善我们所生活的城市。例如基于城市实时交通信息、利用社交网络和天气数据来优化最新的交通情况。目前很多城市都在进行相关的大数据分析和试点。

（5）金融交易

大数据在金融行业主要是用于金融交易。高频交易（HFT）是大数据应用比较多的领域，其中大数据算法被应用于交易决定。现在很多股权的交易都是利用大数据算法进行的，这些算法越来越多地考虑了社交媒体和网站新闻来决定在未来几秒内是买入还是卖出。

通过上面的描述也可以看出，大数据不只是适用于企业和政府，同样也适用于我们生活当中的每个人。我们可以利用可穿戴装备（如智能手表或者智能手环）生成最新的数据，对热量的消耗以及睡眠模式进行追踪；还可以利用大数据分析来寻找属于我们的爱情，大多数的交友网站就是应用大数据工具来帮助需要的人匹配合适的对象。

1.2 大数据平台

大数据平台有哪些呢？

一般认为大数据平台分为两个方面，硬件平台和软件平台。硬件平台一般如 OpenStack、Amazon 云平台、阿里云计算等，类似这样的平台其实做的是虚拟化，即把多台机器或一台机器虚拟化成一个资源池，然后给成千上万人用，各自租用相应的资源服务等。而软件平台则是大家经常听到的，如 Hadoop、MapReduce、Spark 等，也可以狭义理解为 Hadoop 生态圈，即把多个节点资源（可以是虚拟节点资源）进行整合，作为一个集群对外提供存储和运算分析服务。

Hadoop 生态圈大数据平台，可以大概分为 3 种：Apache Hadoop（原生开源 Hadoop）、Hadoop Distribution（Hadoop 发行版）、Big Data Suite（大数据开发套件）。Apache Hadoop 是原生的，即官网提供的，只包含基本的软件；Hadoop Distribution 是一些软件供应商提供的，具有的功能相对多，这个版本有收费版也有免费版，用户可选；而大数据开发套件则是一些大公司提供的集成方案，提供的功能更多，但是相应的也比较贵。

Apache Hadoop 是开源的，用户可以直接访问或更改代码。它是完全分布式的，配置包含用户权限、访问控制等，再加上多种生态系统软件支持，比较复杂。这里涉及版本不兼容性问题。所以该版本比较适合学习并理解底层细节或 Hadoop 详细配置、调优等。

Hadoop Distribution 版本简化了用户的操作以及开发任务，比如可以一键部署等，而且有配套的生态圈支持以及管理监控功能，如业内广泛使用的 HDP、CDH、MapR 等平台。CDH 是最成型的发行版本，拥有最多的部署案例，而且提供强大的部署、管理和监控工具，其开发公司 Cloudera 贡献了自己的可实时处理大数据的 Impala 项目。HDP 是 100% 开源 Apache Hadoop 的唯一提供商，其开发公司 Hortonworks 开发了很多增强特性并提交至核心主干，并且 Hortonworks 为入门者提供了一个非常好的、易于使用的沙盒。MapR 为了获取更好的性能和易用性而支持本地 UNIX 文件系统而不是 HDFS（使用非开源的组件），并且可以使用本地 UNIX 命令来代替 Hadoop 命令。除此之外，MapR 还凭借诸如快照、镜像或有状态的故障恢复之类的高可用性特性来与其他竞争者相区别。当需要一个简单的学习环境时，就可以选用这个版本，当然，针对一些企业也可以选择这个版本的收费版，也是有很多软件支持的。

Big Data Suite（大数据套件）是建立在 Eclipse 之类的 IDE 之上的，其附加的插件极大地方便了大数据应用的开发。用户可以在自己熟悉的开发环境之内创建、构建并部署大数据服务，并且生成所有的代码，从而做到不用编写、调试、分析和优化 MapReduce 代码。大数据套件提供了图形化的工具来为你的大数据服务进行建模，所有需要的代码都是自动生成的，只需配置某些参数即可实现复杂的大数据作业。当企业用户需要不同的数据源集成、自动代码生成或大数据作业自动图形化调度时，就可以选择使用大数据套件。

1.3　本章小结

通过本章的介绍，相信大家对大数据有了一个比较感性的认识，那接下来学习什么呢？

接下来的内容就是大数据技术涉及的相关技术。在本书中，大数据技术仅指软件层面，比如使用 Hadoop 生态圈软件等，而非硬件平台。这里的硬件平台主要指的是把所有硬件资源整合，使其虚拟化一个资源池的概念，涉及的技术有 OpenStack、亚马逊云平台、阿里云平台等。

在后面的章节中，主要介绍 Hadoop 生态圈的相关技术，如 HDFS、YARN、MapReduce、HBase、Hive、Pig、Spark、Oozie 等。每个章节采用理论加实践的方式，使读者能够在理解相关技术原理的基础上，动手操作，加深理解，做到看完本书就能直接上手实践。

“授人以鱼不如授人以渔”，期望本书能成为愿意学习大数据、愿意加入到大数据开发行列的相关人员的一盏指路明灯，愿读者能乐享其中。