大数据分析得天下　　大数据应用赢未来

大数据

技术及应用教程

李联宁　编著

Big Data

清华大学出版社

大数据

技术及应用教程

李联宁　编著

清华大学出版社
北京

内 容 简 介

本书详细介绍了大数据技术的基础理论和最新主流前沿技术，全书共分为10章，分别介绍我们目前面临的数字化信息社会的大数据时代、大数据技术基本概念、云计算网络、大数据采集与预处理、大数据存储、计算模式与处理系统、查询显示与交互、大数据分析与数据挖掘、隐私与安全、大数据技术发展前景，同时包括行业案例研究(银行、保险、证券、金融行业)，典型系统与相关大数据分析实例。

本书主要作为高等院校计算机专业、信息管理专业、经济类专业、管理类专业相关本科生和研究生专业基础课的教材，也可以作为干部培训、职业技术教育以及职业培训机构的云计算与大数据分析技术的专业训练教材。对从事云计算与大数据分析工作的财政金融、政府管理、计算机网络、软件工程的方面的管理与工程技术人员也有学习参考价值。

图书在版编目(CIP)数据

大数据技术及应用教程/李联宁编著. —北京：清华大学出版社，2016
ISBN 978-7-302-44561-6

Ⅰ. ①大…　Ⅱ. ①李…　Ⅲ. ①数据处理—教材　Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2016)第174858号

责任编辑：白立军　李　晔
封面设计：杨玉兰
责任校对：李建庄
责任印制：宋　林

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 装 者：北京密云胶印厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张：**23.75　　**字　　数：**576千字
版　　次：2016年10月第1版　　**印　　次：**2016年10月第1次印刷
印　　数：1～2000
定　　价：49.00元

产品编号：069155-01

前言

本书试图在介绍大数据技术的理论基础上对大数据分析最新前沿技术做全面详细介绍，给出实际案例及行业解决方案，达到技术全面、案例教学及工程实用的目的。

本书主要分为4个部分，共10章，分别按大数据的技术架构分层次详细讲述涉及大数据分析系统的各类相关技术：

第一部分　大数据基础知识，简单介绍我们目前面临的数字化时代与信息社会的状况，大数据的定义和特点、大数据技术基础、大数据的社会价值、大数据的商业应用、大数据的基础架构、云计算网络的技术层次、典型的云计算网络平台，包括第1章“大数据技术基本概念”和第2章“基础架构——云计算网络”；

第二部分　大数据理论与技术，介绍涉及大数据分析的基本理论与技术基础，按照技术层次分别介绍大数据采集与预处理、大数据存储、大数据计算模式与处理系统、大数据查询、显示与交互、大数据分析与数据挖掘、大数据隐私与安全，包括第3章到第8章的内容；

第三部分　为行业案例研究，以银行、保险、证券、金融行业为例，介绍涉及大数据分析的理论与技术方法在具体行业中的应用，包括第9章“行业案例研究”；

第四部分　大数据技术发展前景，介绍大数据引发的新一代信息技术变革浪潮、大数据各个过程的最新技术与发展前景，包括第10章大数据技术发展前景。

本书主要作为高等院校计算机专业、信息管理与信息系统专业、经济类专业、管理类专业相关专业本科生和研究生专业基础课的教材，安排课时为48课时(3学分)。如课时缩减，可在概要叙述第一部分的基础上，主要讲解第二部分第3章到第8章的内容，并安排学生在课外自主阅读每章节后的案例及第9章“行业案例研究”。第10章“大数据技术发展前景”仅作参考性讲解。

本书的特点是紧扣实践应用需求，全面讲述讲云计算与大数据分析实用技术，提供了大量的实际案例、数据分析适用技术。内容新颖、用表格和结构图直观描述知识并力图反映最新主流技术。

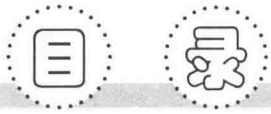
每一章在讲解相关理论外，还讲解了最新前沿技术。各章都附有案例、习题以帮助读者学习理解和实际工程应用。为方便教师教学，附有全套教学PPT课件、教学大纲、教学计划以便教师使用。

本书由李联宁教授编著，在本书编写过程中，编者参考了国内外大量的云计算网络与大数据分析技术的书刊及文献资料，主要参考书籍及研究论文在书后“参考文献”中

列出。但由于大量来自网络的资料未能详尽标注作者及文献资料来源，疏漏之处在所难免，在此一并对书刊文献、科技论文的作者表示感谢。如有遗漏，恳请相应书刊文献作者及时告知，将在书籍再版时列入。如发现本书有错误或不妥之处，恳请广大读者不吝赐教。

编　者

2016年8月



第一部分　大数据基础

第二部分 大数据技术

第三部分 大数据分析案例

第四部分 大数据技术现状及发展展望

第一部分

大数据基础

第 1 章　大数据技术基本概念

第1章　大数据技术基本概念

当今，信息技术为人类步入智能社会开启了大门，带动了互联网、物联网、电子商务、现代物流、网络金融等现代服务业发展，催生了车联网、智能电网、新能源、智能交通、智能城市、高端装备制造等新兴产业发展。现代信息技术正成为各行各业运营和发展的引擎。但这个引擎正面临着大数据这个巨大的考验。各种业务数据正以几何级数的形式爆发，其格式、收集、储存、检索、分析、应用等诸多问题，不再能以传统的信息处理技术加以解决，对人类实现数字社会、网络社会和智能社会带来了极大的障碍。

大数据的出现将影响各行各业以及每个人生活。以下十个事实会让你相信，每个人都必须注意大数据：

(1) 全球数据的90%产生于过去2年内。

(2) 当前数据产生的速度非常快，以今天的数据生产速度，我们可以在2天内生产出2003年以前的所有数据。

(3) 行业内获取并且存储的数据量每1.2年就会翻一番。

(4) 到2020年，全球数据量将由现在的3.2ZB变为40ZB(1ZB＝1024EB，1EB＝1024PB，1PB＝1024TB)。

(5) 仅Google一家搜索引擎，每秒就处理4万次搜索查询，一天之内更是超过35亿次。

(6) 最近的统计报告显示，我们每分钟在Facebook上贡献180万次赞，上传20万张照片。与此同时，我们每分钟还发送2.04亿封邮件，发送27.8万个推文。

(7) 每分钟大约有100小时的视频被传上类似YouTube这样的视频网站。更有趣的是，要花费15年才能看完一天之内被传到YouTube上的全部视频。

(8) AT&T被认为是能够用单一数据库存储最多数据量的数据中心。

(9) 在美国，很多新的IT工作将被创造出来以处理即将到来的大数据工程潮，而每个这样的职位都将需要3个额外职位的支持，这将会带来总计600万个新增工作岗位。

(10) 全球每分钟会新增570个网站。这一统计数字至关重要，也具有颠覆性。

预测是：数据以及数据分析能力正与日俱增，未来五年，无论何等规模的企业都将使用某种形式的数据分析来影响其商业运作。

1.1　数据

数据(data)是对客观事物的逻辑归纳，用符号、字母等方式对客观事物进行直观描述。数据是进行各种统计、计算、科学研究或技术设计等所依据的数值，是表达知识的字符的集合。数据是信息的表现形式。数据可以是连续的值，例如声音，称为模拟数据；也

可以是不连续(离散)的值,例如成绩,称为数字数据。

1.1.1 数据的单位

数据最小的基本单位是 bit,按顺序给出所有单位:bit、Byte、KB、MB、GB、TB、PB、EB、ZB、YB、BB、NB、DB。

它们按照进率 1024(2 的十次方)来计算:

1Byte = 8bit
1KB = 1024Bytes = 8192bit
1MB = 1024KB = 1 048 576Bytes
1GB = 1024MB = 1 048 576KB
1TB = 1024GB = 1 048 576MB
1PB = 1024TB = 1 048 576GB
1EB = 1024PB = 1 048 576TB
1ZB = 1024EB = 1 048 576PB
1YB = 1024ZB = 1 048 576EB
1BB = 1024YB = 1 048 576ZB
1NB = 1024BB = 1 048 576YB
1DB = 1024NB = 1 048 576BB

1.1.2 数据与信息的关系

数据是一种未经加工的原始资料。数字、文字、符号、图像都是数据。数据是客观对象的表示,而信息则是数据内涵的意义,是数据的内容和解释。综上所述,数据就是指能够客观反映事实的数字和资料。

信息与数据的关系是:信息与数据是不可分离的,数据是信息的表达,信息是数据的内涵。数据本身并没有意义数据只有对实体行为产生影响时才成为信息。

1.1.3 数据的分类

在信息社会,信息可以划分为两大类:一类信息能够用数据或统一的结构加以表示,我们称之为结构化数据,如数字、符号;另一类信息无法用数字或统一的结构表示,如文本、图像、声音、网页等,我们称之为非结构化数据。结构化数据属于非结构化数据的一部分,是非结构化数据的特例。

1. 结构化数据

结构化信息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。我们通常接触的,包括生产、业务、交易、客户信息等方面的记录都属于结构化信息。

结构化数据简单来说就是存储在结构化数据库里的数据,可以用二维表结构来逻辑表达实现的数据。结合到典型场景中更容易理解,比如企业 ERP、财务系统;医疗 HIS 数

据库；教育一卡通；政府行政审批；其他核心数据库等。这些应用需要包括高速存储应用需求、数据备份需求、数据共享需求以及数据容灾需求。

2. 非结构化数据

不方便用数据库二维逻辑表来表现的数据即称为非结构化数据，包括所有格式的办公文档、文本、图片、标准通用标记语言下的子集 XML、HTML、各类报表、图像和音频/视频信息等等。

所谓非结构化数据库，是指数据库的变长记录由若干不可重复和可重复的字段组成，而每个字段又可由若干不可重复和可重复的子字段组成。用它不仅可以处理结构化数据（如数字、符号等信息）而且更适合处理非结构化数据（全文文本、图像、声音、影视、超媒体等信息）。简单地说，非结构化数据库就是字段可变的数据库。

非结构化 Web 数据库主要是针对非结构化数据而产生的，与以往流行的关系数据库相比，其最大区别在于它突破了关系数据库结构定义不易改变和数据定长的限制，支持重复字段、子字段以及变长字段并实现了对变长数据和重复字段进行处理和数据项的变长存储管理，在处理连续信息（包括全文信息）和非结构化信息（包括各种多媒体信息）中有着传统关系型数据库所无法比拟的优势。

3. 半结构化数据

所谓半结构化数据，就是介于完全结构化数据（如关系型数据库、面向对象数据库中的数据）和完全无结构的数据（如声音、图像文件等）之间的数据，HTML 文档就属于半结构化数据。它一般是自描述的，数据的结构和内容混在一起，没有明显的区分。

4. 各类数据的区别

结构化数据：行数据，存储在数据库里，可以用二维表结构来逻辑表达实现的数据。

非结构化数据：包括所有格式的办公文档、文本、图片、XML、HTML、各类报表、图像和音频/视频信息等等。

半结构化数据：介于完全结构化数据和完全无结构的数据之间的数据，它一般是自描述的，数据的结构和内容混在一起。

1）数据模型

各类数据的数据模型和基本特征如下：

结构化数据：二维表（关系型）。

半结构化数据：树、图。

非结构化数据：无。

2）关系型数据库系统 RMDBS 的数据模型

RMDBS 的数据模型包括网状数据模型、层次数据模型、关系型。

3）不同类型数据的形成过程

结构化数据：先有结构，再有数据。

半结构化数据：先有数据，再有结构。

5. 互联网信息分类

互联网上出现的海量信息，同样分为结构化、半结构化和非结构化三种。