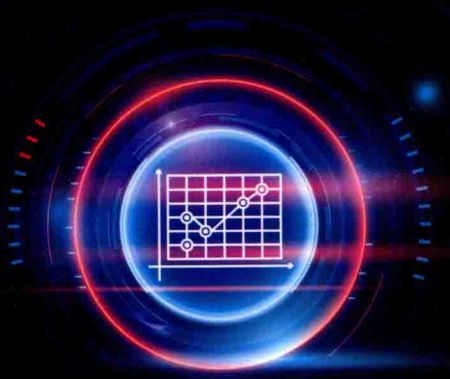
大数据分析与计算

Big Data Analytics and Computation

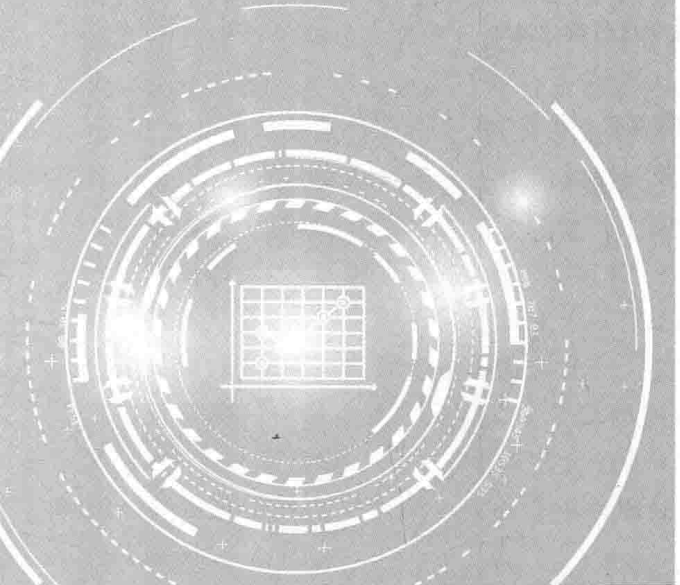


◎ 汤羽 林迪 范爱华 吴薇薇 编著

清華大學出版社

大数据分析与计算

Big Data Analytics and Computation



◎ 汤羽 林通 范爱华 吴薇薇 编著

清華大學出版社
北京

内 容 简 介

大数据应用已成为行业热点和产业发展新增长点，数据科学与计算技术也是最新的前沿领域，其中，大数据计算分析提供了核心的技术支撑。本书从大数据计算系统的三个层次对数据模型、处理算法、计算模型与架构、开发技术标准等内容进行了综合性的介绍，重点阐述了各类数据分析算法和 MapReduce，图并行计算，交互式处理，流计算，内存计算等计算架构。本书适合作为数据科学与大数据计算技术、计算机科学与技术、互联网应用系统、物联网工程等专业相关课程的教材。

图书在版编目(CIP)数据

大数据分析与计算/汤羽等编著. —北京：清华大学出版社，2018
ISBN 978-7-302-48586-5

Ⅰ. ①大…　Ⅱ. ①汤…　Ⅲ. ①数据处理－高等学校－教材　Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2017)第 316141 号

责任编辑：贾　斌　薛　阳
封面设计：刘　键
责任校对：梁　毅
责任印制：王静怡

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：北京鑫海金澳胶印有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：31.25　　**字　　数**：759 千字
版　　次：2018 年 3 月第 1 版　　**印　　次**：2018 年 3 月第1次印刷
印　　数：1～2000
定　　价：89.00 元

产品编号：070307-01

前言

FOREWORD

大数据(Big Data)已被视为硬件、软件、网络之外的第四种计算资源,随着各类大数据应用的兴起,大数据的采集、存储、建模及计算处理已成为分布式计算领域的热门研究课题,也引起产业界极大的兴趣和关注。大数据的计算处理不仅涉及各类数据分析挖掘算法,其计算系统的性能更多依赖于计算模型与计算架构。目前,比较一致的看法是大数据计算系统大致可分为三个层次:数据存储层、数据处理层和数据应用层。数据存储层提供海量数据存储架构与数据访问界面;数据处理层提供对数据分析算法和计算模型的支持;数据应用层则包含各种基于大数据计算分析的应用软件系统。这三个层面都涉及不同的数据模型、计算架构及开发技术标准,目前主流的有两个主线:以 Google 为代表的商业产品和以 Hadoop 为代表的开源技术。在学习和研究大数据计算技术时,需要对上述计算架构、技术和标准有一个总体的了解,这样才能做到不限于一点而把握全局。

针对国家"互联网+"的战略发展需求,近期国内不少高校新开设了数据科学与大数据计算技术专业,大数据分析与计算成为其主干专业课程,其他如计算机科学与技术、互联网应用系统、物联网工程等专业都需要开设大数据计算课程,因此迫切需要一本对大数据处理与计算有一个较全面的论述、适合高年级本科生或研究生学习的教材,正是基于这种需求,本书作者编著了此书,希望对大数据计算系统的各类分析算法、计算模型、计算架构与开发技术做出一个综合性的介绍与阐述,为大家进一步学习大数据技术及应用开发打下基础。

全书共计 20 章,第 1～3 章介绍大数据计算的概念、计算体系总体架构、技术标准等,让读者建立大数据计算的基本概念;第 4～6 章介绍数据采集方法、数据建模及各类分析算法;第 7～10 章介绍文本数据读取、数据处理与分析、数据可视化技术;第 11 章和第 12 章详细介绍 Hadoop 计算平台,包括 HDFS 分布式文件系统与 MapReduce 计算模型;第 13～16 章具体介绍各类大数据计算模型与架构,包括图并行计算、交互式计算、流计算、内存计算等,其中重点阐述了 Pregel、Hama、Storm、Spark 等计算架构;第 17～20 章则介绍了大数据计算技术在医疗保险系统、互联网电子商务、金融信贷系统等领域的应用。本书包含内容较多、篇幅较长,教师在讲授时可根据自己的需要对章节进行选取裁剪。

汤羽教授负责本书的总体结构及第 1～3 章、第 11 章和第 12 章的撰写,林迪副教授负责第 4～10 章,范爱华副教授负责第 13～16 章,吴薇薇硕士负责第 17～20 章。本书部分图片取自互联网,部分文字也参考了网页内容,作者尽可能将引用链接在参考文献罗列中给出,少部分无法给出引用的,作者在此一并致谢。

大数据计算是一个新兴技术领域且仍在高速发展中，新的概念、方法和技术不断涌现。作者因学识有限，本书必然会存在不足，希望得到学界同仁的批评指正，以利我们改进完善。“业精于勤荒于嬉、行成于思毁于随”，作者愿与科学界同行一起努力在这个领域耕耘。

汤　羽

2017 年 7 月于蓉城

目录

CONTENTS

第1章

绪　论

进入21世纪，人类发现自己正面临着中国唐代诗人李白描述的“黄河之水天上来”(图1-1)的大数据场景：互联网搜索引擎Google每天完成10亿次查询，社交网站Facebook每天处理80亿条信息；在科学领域，2003年基本完成的基因组计划完成了四十多种生物全基因组测序以及3.2×10^9人类基因组碱基对测序，到2006年DNA碱基数目已超过1300亿，目前全世界每年生物数据产出量估计已达10^{15}B(1PB)，且以每三年翻一番的速度增长；在金融领域，中信银行2008年发卡500万张，2010年则翻了一倍，带来了海量数据需要处理；2012年，国家工业和信息化部宣布，中国移动互联网用户已达7.5亿。与此同时，随着智慧城市、物联网等新兴应用模式的发展，各种摄像头、数字标牌、感应装置、检测装置以及嵌入式终端的数量也在急剧增加，有关数据预测显示：作为物联网一个重要组成部分的射频识别(RFID)标签，其销量将从2011年的不到30亿个发展到2021年

图1-1　黄河之水天上来

的2090亿个。在互联网浏览搜索、物联网传感数据、移动终端与GPS系统、以及社交网络等领域，全世界的信息量以每两年翻番的速度增长。据国际研究机构IDC报告：2011年，全球数据量为1.8ZB(1ZB＝10^6PB＝10^9TB＝10^{12}GB)，2015年达到8ZB，2020年将达到35ZB。

1.1 数据与数据科学

1. 数据定义

数据(Data)被看作现实世界中自然现象和人类活动所留下的轨迹[1]。在计算机科学中，数据的定义是指所有能输入到计算机并被计算机程序处理的符号的总称，是具有一定意义的数字、字母、符号和模拟量的统称。《韦伯斯特大词典》(*Merriam-Webster Dictionary*)把数据定义为“用于计算、分析或计划某种事物的事实或信息；由计算机产生或存储的信息(facts or information used usually to calculate, analyze or plan something; information that is produced or stored by a computer)”[2]。事实上，数据的形式多样化，可以表现为数值、文字、图像、音频、视频或其他计算机可以识别和处理的形式，数据来源也可以是社会数据(商业数据、生产数据、系统数据、媒体数据等)、个人数据(社交网络、个人消费)、政府数据(统计数据、人口普查、经济年报等)。人类四千年历史所产生的所有的文明记录，包括历史、文学、艺术、哲学、考古及一切的科学成就，都可以数据的形式存储和保留下来。

2. 数据简史

在人类文明有记载的四千年历史中，人类活动记录从早期古埃及的结绳记事(图1-2)到中国殷商时期(公元前1320—1046年)的甲骨文(图1-3)，从东汉宦官蔡伦(公元61—121年)发明造纸术(图1-4)到北宋布衣毕昇(公元61—121年)发明活字印刷术(图1-5)，文明的记录无不以文字(数据的一种形式)传承下来。

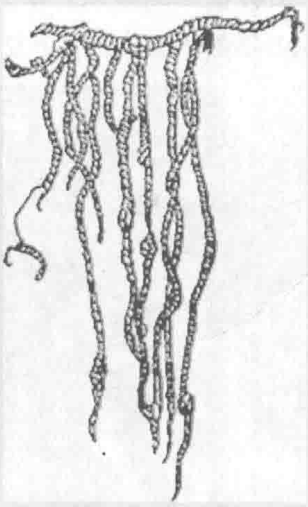


图1-2 结绳记事

图1-3 殷商甲骨文

图 1-4 蔡伦造纸术

图 1-5 毕昇活字印刷术

到了近代，面对日益增长的计算量，人工计算的方式已难以应对，人类进入了机械计算时代。这一时期的代表有：1642 年，法国数学家布莱士·帕斯卡(Blaise Pascal)发明的机械式加减法计算机；1671 年，德国数学家戈特弗里德·莱布尼茨(Gottfried W. Leibniz)制成了第一台能够进行加减乘除运算的机械式计算机；1833 年，英国科学家查尔斯·巴贝奇(Charles Babbage)提出了自动化分析机的设计，第一次提出了机器可执行预先记录在穿孔卡上的指令来完成计算的思路。与之合作的艾达·洛夫莱斯伯爵夫人(Augusta Ada Lovelace)更进一步为巴贝奇机器写出了计算注释笔记，其中包含由机器执行的计算伯努利数的算法和步骤，被认为是人类完成的第一个可执行的计算机程序，艾达也因此作为人类历史上的第一个计算机程序员而被纪念。

进入 20 世纪，1904 年，英国工程师约翰·弗莱明(John A. Fleming)发明了真空二极管；1906 年，美国工程师李·弗雷斯特(Lee de Forest)发明了真空三极管；1947 年，美国贝尔实验室的科学家威廉·肖克利(William B. Shockley)、约翰·巴丁(John Bardeen)和沃尔特·布拉顿(Walter Brattain)组成的研究小组发明了晶体管；1958 年，美国物理学家杰克·基尔比(Jack Kilby)发明了集成电路技术。电子管、晶体管、集成电路技术及随后的超大规模集成电路(VLSI)的诞生使得人类真正进入了电子计算机时代。1946 年 2 月，世界上第一台电子管计算机“埃尼阿克”(ENIAC)在美国宾夕法尼亚大学诞生；1956 年，美国贝尔实验室研制出第一台晶体管计算机 TRADIC；1964 年，美国 IBM 公司推出第一代采用集成电路的电子计算机 IBM360 系列，这以后各种计算能力更为强大、数据处理能力呈爆炸式增长的超级计算机、高端服务器、图形工作站以及计算机集群层出不穷，使得计算机处理数据的能力从早期的 KB、MB 级别达到了今天的 TB 或 PB 量级。2015 年，美国著名社交网站 Facebook 每天需处理 100 亿条消息和 3.5 亿张新图片，而谷歌(Google)每天应对的查询请求达到 30 亿次，后台处理的数据量达到 85TB。

应当注意的是，数据(Data)、信息(Information)、知识(Knowledge)与价值(Value)这 4 个词在信息科学中既相互关联，又具有不同的含义。数据体现的是一种过程、状态或结果的记录，这类记录被数字化后可以被计算机存储和处理。信息则是包含在数据之中的能够为人脑理解和思维推理和结论。例如，“01001000 01100101 01101100 01101100 01101111

00100000 01110111 01101111 01110010 01101100 01100100 00100001"是一串二进制数值，是一组能被计算机识别、存储和处理的数据。经计算机程序识别转换(ASCII 码值字符转换)，我们知道它代表"Hello world!"这样一个字符串，包含向世界问好的特殊信息。更进一步，在计算机编程语言世界，"Hello world!"实际上是一个约定俗成的机器或程序语言启动显示语句，这就上升为知识。最终，如果有人把这一固有的显示方法拿去注册了专利并因此获利，就产生了价值。图 1-6 表征了这一从数据到信息到知识到价值的过程。

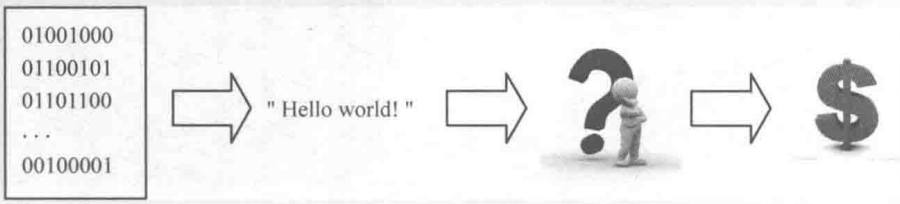


图 1-6 数据-信息-知识-价值的转换过程

3. 数据科学

当信息科学处理的数据发展到 Facebook 和 Google 的数据规模，数据本身(类型、规模、属性、用途等)及相关的大规模数据分析计算技术就形成了一门新的学科领域：数据科学(Data Science)或数据工程(Data Engineering)。早在 1974 年，丹麦计算机科学家、2005 年图灵奖得主彼得·诺尔(Peter Naur)即提出了"数据学"(Datalogy)的概念[3]，但他更多指的是以数据为对象的计算机科学和编程，认为"数据学"是计算机科学的延伸，其研究对象是数值化的数据。事实上，在个人计算机(PC)出现之前的早期的计算机确实是更多地用于处理数据和科学计算，计算机大大加快了数据处理的速度、效率和准确性，但作为计算机运算对象和输出结果的数据本身尚未引起科学家们的特别注意。

1996 年，在日本神户的一个国际会议上第一次正式使用"数据科学"这一名称[4]。1997 年，密歇根大学教授杰夫·吴(Jeff C. Wu)在演讲中提出"统计学＝数据科学?"的观点并建议将统计学改名为数据科学，统计学家改名为数据科学家[5]。2001 年，贝尔实验室科学家威廉·克里富兰(William S. Cleveland)第一次提出数据科学应作为由统计学延伸出来的一个独立研究领域，认为统计学中与数据分析有关的技术内容(区别于概率论)在下面 6 个方面扩展后形成一个新的独立学科"数据科学"(Data Science)[6]。

(1) 多学科研究(Multidisciplinary Investigations)；

(2) 数据模型与分析方法(Models and Methods for Data)；

(3) 数据计算(Computing with Data)；

(4) 数据学教程(Pedagogy)；

(5) 工具评估(Tool Evaluation)；

(6) 理论(Theory)。

在 2002 年和 2003 年，国际科学委员会(International Council for Science)和哥伦比亚大学分别创办了数据科学杂志，为这一学科领域的研究工作发表和交流建立了国际学术平台。大规模数据计算的特点和重要性已引起科学界注意，数据科学或数据处理技术被有些科学家认为将成为一个与计算科学并列的新科学领域。已故著名图灵奖获得者 Jim Gray 在 2007 年的一次演讲中提出，数据密集型科学发现(Data-Intensive Scientific Discovery)将