HZ BOOKS
华章IT

大数据算法

王宏志◎编著
（哈尔滨工业大学）

BIG DATA ALGORITHMS

机械工业出版社
China Machine Press

BIG DATA ALGORITHMS

大数据算法

王宏志◎编著

（哈尔滨工业大学）

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

大数据算法 / 王宏志编著 . —北京：机械工业出版社，2015.7
（大数据技术丛书）

ISBN 978-7-111-50849-6

I. 大… II. 王… III. 数据处理 – 算法分析 IV. TP274

中国版本图书馆 CIP 数据核字（2015）第 156885 号

大数据算法是大数据得以有效应用的基础，也是有志于从事大数据以及相关领域工作必须学习的课程。本书由从事大数据研究的专家撰写，系统地介绍了大数据算法设计与分析的理论、方法和技术。本书共分为 10 章，第 1 章概述大数据算法，第 2 章介绍时间亚线性算法，第 3 章介绍空间亚线性算法，第 4 章概述外存算法，第 5 章介绍大数据外存查找结构，第 6 章讲授外存图数据算法，第 7 章概述 MapReduce 算法，第 8 章通过一系列例子讲授 MapReduce 算法，第 9 章介绍超越 MapReduce 的算法设计方法，第 10 章讨论众包算法。

本书适合作为计算机科学、大数据等专业本科生、研究生教材，也可供从事大数据相关工作的工程技术人员参考。

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：谢晓芳　　责任校对：董纪丽

印　　刷：中国电影出版社印刷厂　　版　　次：2015 年 7 月第 1 版第 1 次印刷

开　　本：185mm × 260mm　1/16　　印　　张：15.5

书　　号：ISBN 978-7-111-50849-6　　定　　价：49.00 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）88378991　88361066　　投稿热线：（010）88379604

购书热线：（010）68326294　88379649　68995259　　读者信箱：hzjsj@hzbook.com

前言

本书的缘起

“大数据”在今天成为一个非常时尚的概念，其影响已经远远超过了计算机学科本身，甚至影响到了自然科学、社会科学、人文科学等。由于其深远的影响和广泛的应用，大数据一直得到IT从业人员的重视，他们对大数据相关理论、技术的学习有着强烈的需求。

“算法设计与分析”是计算机科学的重要主题，进行大数据计算，“算法设计与分析”是必不可少的步骤，可以说，算法设计是“大数据落地”的关键之一。然而，虽然在今天的书店里，关于大数据的书籍数不胜数，但真正从“算法设计与分析”角度关注大数据的书却很少。究其原因，当前“大数据算法”的知识体系还远不完备，因为“大数据”是计算机学科的增长点之一，“大数据算法”的内涵和外延也不断发生着变化，而且大数据上算法设计与分析得到的知识驳杂，难以梳理出一个明晰的知识体系。而大数据不同方面的从业人员，对“大数据算法”的理解也不尽相同。作者曾经调研过国内外和“大数据算法”相关的课程，其教学内容的差异非常大。

因而，笔者写了本书，作为一种勇敢的尝试，试图兼顾深度和广度来介绍“大数据算法”。其缘起有三。

其一，笔者从本科加入了李建中教授领导的哈尔滨工业大学数据库研究中心，留校工作到现在。随着“数据”在计算机学科扮演的角色日益重要，中心的名字经历了“数据库研究中心”到“知识与数据工程研究中心”到“海量数据计算研究中心”到“国际大数据研究中心”的变化，并且一直是围绕“数据”的计算开展研究。在中心良好的学术氛围下，笔者进行了十几年“数据”计算的研究，也一直在思考“数据为中心的计算到底需要何种特别的算法设计技术”这一问题，有一些不成熟的心得，希望与读者分享。

其二，机械工业出版社王彬编辑在2013年全国大数据会议上邀请笔者写一本和“大数据”、“算法”相关的书，促使笔者去思考和学习，试图梳理出一条“大数据算法”的脉络。

其三，在网易云课堂的孙志岗总监的鼓动下，笔者在2014年开设了自己的第一门MOOC课程“大数据算法”，2014年夏季学期笔者在哈尔滨工业大学作为全校选修课也开设了“大数据算法”这门课程，这督促着笔者不得不从教学内容到教学方法上去思考如何

表述“大数据算法”。在教学过程中，很多学习这门课程的学生询问教材的事情，很遗憾，笔者只能提供一个参考文献列表，而无法推荐教材，这也促使笔者撰写这样一本书。

本书的特点

本书对大数据计算中涉及的算法设计与分析技术进行了介绍，针对大数据对算法的要求，主要涉及四个方面：亚线性算法、外存算法、并行算法和众包算法。书中给出了多个算法，并对其进行了分析，尽可能使本书适用于各个层次的读者。

书中每一章涉及一类大数据算法设计技术，算法主要用自然语言、伪代码和例子来描述，力图使本书介绍的算法易懂易用。由于为大数据设计算法，在“大数据”上进行实验的成本比较高，因此“算法分析”在“大数据算法”中扮演着更重要的角色，本书也在算法分析方面投入了相当的笔墨。有不同需求的读者可以着重阅读本书不同的部分。

由于“大数据”涉及的内容较广，本书围绕大数据的特点着重介绍大数据算法设计与分析的方法，和大数据分析、大数据系统、大数据编程等书籍具有互补性，可以相互参照进行阅读。

本书适合作为本科生和研究生“大数据”或者“大数据算法”课程的教材，也可以作为“算法设计与分析”等课程的补充教材或课外读物。同时，本书也适合大数据领域从业人员参考。

由于本书是一种新的尝试，涉及的内容非常宽且又是变化迅速，尽管笔者尽全力来写本书(其中的一部分内容甚至来自于2015年发表的文献)，但是由于笔者水平有限，在本书内容的安排、表述、推导等方面的各种不当之处在所难免，敬请读者在阅读本书的过程中，不吝提出宝贵的建议，以改进本书。读者的任何意见和建议请发至邮箱 wangzh@hit.edu.cn。

致使用本书的教师

本书涉及了多方面内容，对于教学而言，本书适用于多门课程的教学，并可以作为“数据结构”、“算法设计与分析”、“数据库系统原理”等课程的补充教材，教师可以从本书中选择适合教学的内容，例如，第5章适合作为“数据库系统原理”这门课“数据库索引”部分的补充教学内容，第4章适合作为“数据结构”这门课“排序”部分的补充教学内容。

针对不同层次的教学可以选择不同的内容。针对本科生或者职业培训的教学可以侧重于算法设计，着重讲授算法本身和算法的应用场景，而对算法分析可以略讲；针对研究生的教学可以在讲算法设计的同时利用更多的时间来讲授算法的分析和推导。

本书每章后包含一些习题，供学生巩固所学内容。

致使用本书的学生

希望本书为学生提供“大数据算法”方面的入门指导，我们尽量让描述通俗易懂，但是一些算法、数据结构或者分析本身比较复杂，有些算法分析远看略显“高冷”，请在阅读时不要畏惧，可以按照相关的证明过程和推理步骤仔细梳理证明的脉络。对于本书涉及的一些可能没有学过的知识，通过“补充知识”部分进行了介绍。

要阅读本书，希望读者有一些算法和程序设计方面的基础，“数据结构”和“算法设计与分析”是本书的先修课程，如果读者没有学过这方面的课程，可以通过阅读《算法导论(原书第3版)》[㊀]如下章节自学相关知识：第1～12章、第15～17章、第18章、第22～24章。本书第2章和第3章涉及一些概率分析知识，如果不需要掌握概率分析的技术而仅读懂本书，本书提供的补充知识足以帮助你理解证明过程；如果希望系统掌握概率分析，可以先阅读一下《概率与计算》[㊁]的第1～6章，奠定概率分析方面的基础，再阅读本书第2章和第3章中的证明。本书第7～9章涉及了并行算法，但并不需要读者具备并行体系结构和并行计算相关的知识，因为当前平台(如Hadoop等)已经提供了足够方便的接口，可以让读者在不具备这些知识的前提下实现数据密集型并行算法。

致使用本书的专业技术人员

本书可以作为一本关于大数据算法的参考手册，供专业技术人员参考。本书各章节具有一定的独立性，读者可以单独查阅感兴趣的主题。

如果读者是一名开发人员，可以根据需要选择本书中的算法进行实现或者以此为参考设计软件当中的新算法。本书提供的伪代码可以很容易地翻译成某种程序设计语言所对应的代码。

在选择和设计算法的过程中，如果需要对算法复杂度有一定了解，本书将可以单独描述的算法复杂度结论以“引理”、“定理”的形式给出，可以直接参考这些结论，而不用详细阅读其证明。

不同类型的大数据应用和本书的不同章节相关。如果应用涉及数据量很大，而内存、计算时间等限制比较严格，请参考本书第2章和第3章；如果应用中数据源源不断到来，必须根据当前接收到的数据进行计算，请参考本书第3章；如果应用中数据存储在外存中，而内存受限，请参考本书第4～6章；如果数据存储在集群中，需要多台计算机并行计算，请参考本书第7～9章；如果应用需要只有人具备的知识，请参考本书第10章。

㊀ 该书由机械工业出版社出版，ISBN：978-7-111-40701-1。——编辑注

㊁ 该书由机械工业出版社出版，ISBN：978-7-111-20805-1。——编辑注

致谢

本书的成书感谢哈尔滨工业大学的李建中教授、高宏教授以及国际大数据研究中心诸位同事的指导和建议，以及在专业上的帮助。

在本书的撰写过程中，哈尔滨工业大学的李可利、张美范、毛运东、王鑫鹏、孙芳媛、周剑、李明达、马钰、田家源、徐扬、张笑影、甘小楚、郭欣彤、李宁宁等同学在资料搜集、整理、文本校对、制图等多个方面提供了帮助和支持，在此表示感谢。

非常感谢我的爱人黎玲利博士，感谢她在我撰写这本书的过程中对我的支持。她除了给我爱和家庭的温暖，还阅读了本书全文并给出了许多专业的建议。

在本书的成书过程中我和机械工业出版社保持愉快的合作，感谢机械工业出版社的王彬编辑和朱劼编辑对我的帮助与支持。

还要感谢在哈尔滨工业大学和 MOOC 选修我课程的同学，你们的意见和建议对本书的写作大有裨益。

最后，笔者关于大数据方面的研究和本书的写作得到了国家重点基础研究发展计划(973)(编号：2012CB316200)、国家自然科学基金(编号：61472099)和国家科技支撑计划基金(编号：2015BAH10F00)的部分资助。

王宏志

2015 年 6 月 7 日于哈尔滨

目录

Chapter

第1章 绪 论

1.1 大数据概述

毫无疑问，大数据已经成为一个热门的概念，然而，不同领域(例如商业、系统结构、数据管理等)对这个概念的解读却各不相同。本节我们对大数据的定义、特点和应用进行概述。

1.1.1 什么是大数据

“大数据”的概念起源于2008年9月《自然》(Nature)杂志刊登的名为“Big Data”的专题，继而迅速得到了科学、计算机、经济等不同领域专家的响应。由于其成因复杂，对大数据目前没有公认的定义，不同的研究人员从不同领域对大数据进行了定义，下面列出三个不同角度对大数据的定义。

1) Kusnetzky Dan 在 What is “Big Data?”一文中提出，大数据是指所涉及的数据量规模巨大，无法通过人工在合理时间内截取、管理、处理并整理成为人类所能解读的信息。

2) 维克托·迈尔-舍恩伯格、肯尼斯·库克耶在《大数据时代》一书中把大数据看成一种方法，即不用随机分析法(抽样调查)这样的捷径，而采用所有数据的方法。

3) “大数据”研究机构 Gartner 的报告指出，“大数据”是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。

这三种定义中，第一种定义更强调处理能力，第二种定义更强调应用方法，第三种定义更侧重应用价值。本书的主题是“大数据算法”，因而更加侧重于第一种定义，即“规模巨大，无法通过人工来处理”。

1.1.2 无处不在的大数据

现实生活中的数据有多大呢？根据统计，在2006年，个人用户刚刚迈进TB时代，这一年全球共产生了约180EB=180×10^{18}字节的数据；在2011年，达到了1.8ZB=1.8×10^{21}字节。有市场研究机构预测：

到2020年，整个世界的数据总量将会增长44倍。你也许会好奇为何会产生如此庞大的数据，下面我们举几个现实中的大数据例子。

- **社交网络**　由于数据来自所有用户的参与，社交网络中的数据量非常大，而且增长非常迅速。例如，新浪微博在晚高峰的时候1秒产生的数据达到100条以上。如果把脸书(Facebook)中的社交网络看成图，在2012年这个图已经达到了超过8亿个顶点，平均每个点的度超过130，每天增加的数据量达到500TB。
- **科学仪器**　科学仪器获取了非常巨大的数据，比如说中国遥感国家重点实验室采集的中国大陆地表信息，每个月产生4TB数据。中国天文观测站用LAMOST每年观测到的数据达到3.65TB，美国NASA中心每年获取超过125TB的数据，英国Sanger中心2002年就已经收集了20TB的数据，并且以每年4倍的速度增长。
- **移动通信**　我们每天使用的手机产生了非常巨大的数据，中国移动每年产生的记录超过300TB。
- **传感数据**　传感器持续检测环境信息并不断返回结果，产生了巨大的数据。以波音787为例，其每一个飞行来回可产生TB级的数据，美国每个月收集360万次飞行记录；监视所有飞机中的25 000个引擎，每个引擎一天产生588GB的数据。风力发电机装有测量风速、螺距、油温等多种传感器，每隔几毫秒测一次，用于检测叶片、变速箱、变频器等的磨损程度，一个具有500个风机的风场一年会产生2PB的数据。
- **医疗数据**　美国著名医疗保健公司InSiteOne平均每年获取2.1PB的放射影像数据，英国每年产生300TB乳腺癌数据，在美国相应的数据量达到2.6PB。哈尔滨医科大学第一附属医院每年通过各类医疗仪器搜集的数据超过30TB。
- **商务数据**　生活中的每次刷卡，在超市或者网络中购买的每件商品都产生相应的数据。淘宝网站每天有超过数千万笔交易，单日数据产生量超过50TB。为了有效使用商务大数据，沃尔玛建立了包含PB级数据的数据仓库，Bestbuy建立了包含TB级数据的数据仓库。

补充知识：数据的概念相信读者已经很熟悉，“大数据”重点是大，我们下面看一些关于“大”的定义。

计算机的发展史一直和“大”的定义紧密相连，例如关于硬盘的存储量就经历了一个从KB发展到MB，再发展到TB的过程。英语对“字节”的计数法如下：

1Byte=8bit
1KB=1024Byte
1MB=1024KB=1 048 576Byte
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

汉语计数能力更强一点，可以达到10^{44}，具体的值如下：

千 1000

万 10^4

亿 10^8

兆 10^{12}

京 10^{16}

垓 10^{20}

秭 10^{24}

穰 10^{28}

沟 10^{32}

涧 10^{36}

正 10^{40}

载 10^{44}

1.1.3 大数据的特点

通常用3V或者4V来描述大数据的特点，本小节用4V描述大数据的特点。

1. **规模性**(Volume，耗费大量存储、计算资源)

大数据之“大”，体现在数据的存储和计算均需耗费海量规模的资源上：美国宇航局收集和处理的气候观察、模拟数据达到32PB；谷歌公司索引的网页总数超过1万亿；FICO的信用卡欺诈检测系统保护全世界超过18亿个活跃信用卡账户。

2. **高速性**(Velocity，增长迅速、急需实时处理)

大数据的另一特点在于速度快：大型强子对撞机实验设备中包含了15亿个传感器，平均每秒收集超过4亿条实验数据；每秒超过3万次用户查询提交到谷歌，3万条微博被新浪用户撰写。而在感知、传输、决策、控制这一闭环控制过程中的计算，对数据实时处理有着极高的要求，通过传统数据库查询方式得到的“当前结果”很可能已经没有价值，只有最新的数据才有价值。

3. **多样性**(Variety，来源广泛、形式多样)

在大数据背景下，数据在来源和形式上的多样性愈加凸显：除大量以非结构化形式存在的文本数据，也存在位置、图片、音频、视频等信息。除信息形式的多元化，信息

的来源也表现出多样性：从网络日志、物联网、移动设备、传感器到基因图谱、医疗影像、天体运行轨迹、交通物流数据等。大数据中的多样性已经超越了数据管理中的异构数据库，其不仅仅是模式或模型的不一样，甚至数据本身的存在形式也完全不同，比如说存在文本、多媒体数据，也存在仪器采集来的完全是数字的数据，以及用户产生的用户行为的数据，这些数据有各种各样的存在形式，这些形式导致处理技术的差异，因此需要新的处理技术。

4. **价值稀疏性**(Value，价值总量大、知识密度低)

大数据以其高价值吸引了广泛关注。据全球著名咨询公司麦肯锡报告："如果能够有效地利用大数据来提高效率和质量，预计美国医疗行业每年通过数据获得的潜在价值可超过 3000 亿美元，能够使美国医疗卫生支出降低 8%。"虽然大数据价值高，但是知识密度非常低。谷歌公司首席经济学家 Hal Varian 指出"数据是广泛可用的，所缺乏的是从中提取出知识的能力"；IBM 副总裁兼 CTO Dietrich 表示"可以利用 Twitter 数据获得用户对某个产品的评价，但是往往上百万条记录中只有很小的一部分真正讨论这款产品"。

只有经过高度分析的大数据才可以产生新的价值，需要设计能够适应上述特征的大数据处理算法来处理数据。

1.1.4　大数据的应用

大数据在许多方面有着广泛的应用，甚至说达到了无处不在的程度，本小节将讨论若干大数据的典型应用。

1. **预测**

2013 年 2 月 19 日，微软研究院的 David Rothschil 博士带领的大数据分析团队通过分析入围影片相关数据，预测出 2013 年各项奥斯卡大奖的最终归属，成功命中除最佳导演奖(华裔导演李安获得)外的 13 项大奖。

《纽约时报》FiveThirtyEight 的博客作者和统计学家 Nate Silver 预测：奥巴马有超过 80%的机会赢得周二的大选(后来提升到 90.9%)；David Rothschild 带领的分析团队，在 2012 年使用一个通用的数据驱动型模型，预测了美国 50 个州和哥伦比亚特区共计 51 个选区中 50 个地区的选举结果，准确率高于 98%。

日本国内有一个网站，你只要打开这个网站用自己的 Twitter 账号登录，就可以在短时间内通过数万条 Twitter 找出可能感冒的人，并对过去的感冒情况和今日的感冒情况进行分析(以及统计目前发烧以及嗓子痛的患者数量)。另外该程序还会结合气温和湿度的变化来预测将来感冒的流行情况，并开发了一个"易感冒日历"。通过这个服务，人们就能知道身边有多少人有感冒的症状，并提前做好预防。

2. **推荐**

商务信息推荐和我们每天的生活息息相关，用户在淘宝、京东、卓越等电子商务网站上购物的时候，网站会为我们推荐相关的商品，这些推荐来自大数据。商家采集了大量的用户行为信息，包括购买、浏览、评价等，根据这些行为信息预测当前使用这个网

站的用户下一步可能有哪些行为，再根据预测的结果来给用户推荐他最需要的商品，从而提高用户的购买效率。推荐是很多网站的重要盈利模式，借助推荐技术，大数据能够为电子商务带来价值。

3. 商业情报分析

为了对营销情况进行有效分析，沃尔玛建立了PB级的数据仓库，使得在线完成购物率提高了10%到15%。连锁超市特易购(Tesco PLC)在数据仓库中搜集了700多万个冰箱的数据，通过对这些数据的分析，能够全面监控冰箱状况，并且根据监控和预测的结果，对这些冰箱进行主动维修，从而降低能耗。还有一些案例，比如说有一家牛排店，通过分析Twitter大数据知道哪些人可能是常客，根据客户以往的订单，推测出其所乘的航班，然后派出一位身着燕尾服的侍者为客户提供晚餐，通过这样的服务吸引了越来越多的熟客。

4. 科学研究

今天的科学研究已经超越了牛顿的时代。从历史上看，第谷积攒了大量的天文数据，开普勒通过数据的分析得到了天体三大运动定律，当时计算靠手工进行，需要人工分析，缺少计算机这样有效的计算工具，如果当年有大数据的处理方法的话，开普勒三大运动定律可能更早出来。今天大量的科学仪器产生了海量的数据，这样的数据量已经不是人拿纸拿笔就能分析的，而是需要强大的数据处理能力。今天，由于大数据的支持，科学研究由假设驱动转向基于探索的科学方法，过去设问“我应该设计什么样的实验来验证这个假设?”，现在设问“从这些数据中我能够看到什么?”和“如果把其他领域的数据融合进来，能够发现什么?”，数据密集型科学发现被称为“科学研究的第四范式”。以美国能源部为例，其提出了基于大数据科学研究的支持计划，包括生物和环境的研究计划、大气辐射测量气候的研究计划以及系统生物学的知识库对微生物和植物环境这些功能群落的识别。

补充知识：科学研究的范式

第一范式：几千年前，也就是亚里士多德的时代，科学研究是基于经验的，用于描述自然现象。

第二范式：数百年前，也就是牛顿的时代，科学研究是基于理论研究的，着眼于建立数学模型并进行推广。

第三范式：几十年前，开始了基于计算的科学研究，通过强大的能力，得以模拟复杂的自然现象。

第四范式：也叫作eScience，基于数据探索的科学研究，利用仪器获取数据或者利用模拟器生成数据，再利用软件进行处理，将知识或信息存储在计算机中，科学家利用数据管理技术和统计方法进行科学发现。

1.2 大数据算法

这一节我们概述大数据算法。

1.2.1　大数据上求解问题的过程

首先我们看一看在大数据上问题求解的过程。我们面对的是一个计算问题，也就是说我们要用计算机来处理一个问题。

拿到一个计算问题之后，首先需要判定这个问题是否可以用计算机进行计算，如果学习过可计算性理论，就可以了解有许多问题计算机是无法计算的，比如判断一个程序是否有死循环，或者是否存在能够杀所有病毒的软件，这些问题都是计算机解决不了的。从“可计算”的角度来看，大数据上的判定问题和普通的判定问题是一样的，也就是说，如果还是用我们今天的电子计算机模型，即图灵机模型，在小数据上不可计算的问题，在大数据上肯定也不可计算。计算模型的计算能力是一样的，只不过是算得快慢的问题。

那么，大数据上的计算问题与传统的计算问题有什么本质区别呢？

第一个不同之处是数据量，就是说处理的数据量要比传统的数据量大。第二个不同之处是有资源约束，就是说数据量可能很大，但是能真正用来处理数据的资源是有限的，这个资源包括 CPU、内存、磁盘、计算所消耗的能量。第三个不同之处是对计算时间存在约束，大数据有很强的实时性，最简单的一个例子是基于无线传感网的森林防火，如果能在几秒之内自动发现有火情发生，这个信息是非常有价值的，如果三天之后才发现火情，树都烧完了，这个信息就没有价值，所以说大数据上的计算问题需要有一个时间约束，即到底需要多长时间得到计算结果才是有价值的。判定能否在给定数据量的数据上，在计算资源存在约束的条件下，在时间约束内完成计算任务，是大数据上计算的可行性问题，需要计算复杂性理论来解决，然而，当前面向大数据上的计算复杂性理论研究还刚刚开始，有大量的问题需要解决。

注意：在本书中，有的算法可能很简单，寥寥几行就结束了，然而后面的分析却长达几页。这本书花更大的精力讲授算法分析，是因为在大数据上进行算法设计的时候，要先分析清楚这个算法是否适用于大数据的情况，然后才能使用。

本书讨论的主要内容是大数据上算法的设计与分析，即设计大数据上的算法并且加以分析。

特别值得说明的一点是，对于大数据上的算法，算法分析显得尤为重要，这是为什么呢？对于小数据上的算法可以通过实验的方法来测试性能，实验可以很快得到结果，但是在大数据上，实验就不是那么简单了，经常需要成千上万的机器才能够得出结果。为了避免耗费如此高的计算成本，大数据上的算法分析就十分重要了。

经过算法设计与分析，得到了算法。接着需要用计算机语言来实现算法，得到的是一些程序模块，下一步用这些程序模块构建软件系统。这些软件系统需要相应的平台来实现，比如常说的 Hadoop、SparK 都是实现软件系统的平台。

上面的叙述可以归纳为图 1-1，从中可以看到，大数据算法与分析在整个大数据问题求解过程中扮演着一个核心的角色，因而本书将对此重点介绍。

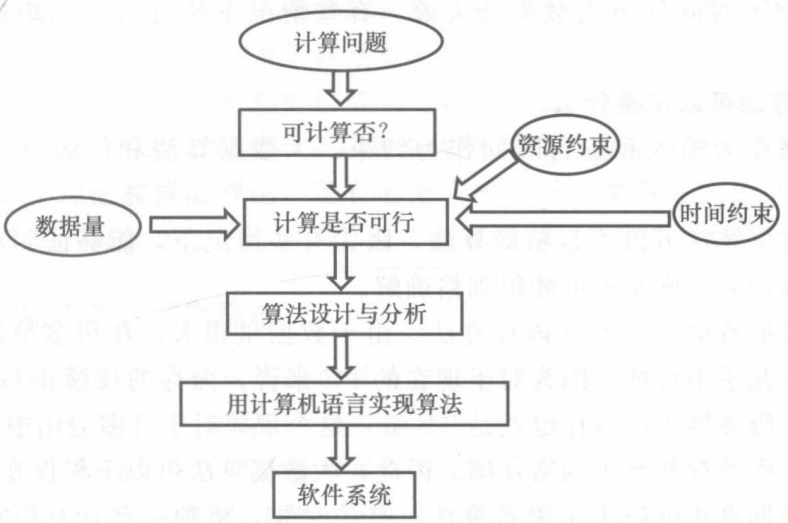

图 1-1 大数据上解决计算问题的过程

1.2.2 大数据算法的定义

1. 大数据算法是什么

根据大数据上的计算过程可以定义大数据算法的概念，如定义 1-1 所示。

定义 1-1(大数据算法) 在给定的资源约束下，以大数据为输入，在给定时间约束内可以计算出给定问题结果的算法。

这个定义和传统的算法有一样的地方，首先大数据算法也是一个算法，有输入有输出；而且算法必须是可行的，也必须是机械执行的计算步骤。

补充知识：算法的定义

定义 A-1(计算) 可由一个给定计算模型机械地执行的规则或计算步骤序列称为该计算模型的一个计算。

定义 A-2(算法) 算法是一个满足下列条件的计算：

1) 有穷性/终止性：有限步内必须停止；

2) 确定性：每一步都是严格定义和确定的动作；

3) 可行性：每一个动作都能够被精确地机械执行；

4) 输入：有一个满足给定约束条件的输入；

5) 输出：满足给定约束条件的结果。

第一个不同之处是，大数据算法是有资源约束的，这意味着资源不是无限的，可能在 100KB 数据上可行的算法在 100MB 的数据上不可行，最常见的一个错误是内存溢出。这意味着进行大数据处理的内存资源不足，因此在大数据算法的设计过程中，资源是一个必须考虑的约束。第二个不同之处是，大数据算法以大数据为输入，而不是以传统数据的小规模为输入。第三个不同之处是，大数据算法需要在时间约束之内产生结果，因