HZ BOOKS
华章科技

大数据思维与应用攻略

王崇骏 编著

现象及感性思辨——溯源大数据
技术及选型思路——剖析大数据
实施及理性思考——实践大数据
机遇及应用思索——拥抱大数据

机械工业出版社
China Machine Press

大数据思维与应用攻略

王崇骏 编著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

大数据思维与应用攻略 / 王崇骏编著 . —北京：机械工业出版社，2016.7

ISBN 978-7-111-54261-2

I. 大… II. 王… III. 数据处理 IV. TP274

中国版本图书馆 CIP 数据核字（2016）第 157469 号

大数据思维与应用攻略

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：余 洁　　责任校对：殷 虹

印　　刷：北京市荣盛彩色印刷有限公司　　版　　次：2016 年 7 月第 1 版第 1 次印刷

开　　本：186mm × 240mm 1/16　　印　　张：25.5

书　　号：ISBN 978-7-111-54261-2　　定　　价：59.00 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

Preface 序

或许还从来没有一个技术名词像“大数据”一样，从其诞生之日起，就得到了“政产学研商用”的一致认同，并引起了哲学家、科学家、技术研究者、工程研发人员的普遍关注。大数据能够受到如此重视的原因可能是：不同的利益主体及不同的工作群体都不约而同地意识到了由“大数据”引发和驱动的思维变革、问题凝练、技术挑战和价值发现。

作为一名大学全职教师，在教学和科研的过程中，我有幸接触、参与和主持了“大数据”相关的课程讲授、项目研发和咨询服务。在教学相长及每一个项目场景的迭代研发过程中，我对“大数据”及数据驱动的项目研发颇有感触。当我尝试把这种感触和心得与学生、同事、学界小同行及项目甲方（有的是软件企业，有的是直接应用方，比如政府、企事业单位等）进行交流的时候，我发现针对不同特点的受众，必须有意地调整相关说辞，否则很难有理论技术及应用情怀方面的共鸣和共振。我想这其中的原因最有可能在于：

1）虽然大家都认为“大数据”是有价值的，但因为价值是极具主观色彩的，所以针对同一个应用场景，不同角色的利益主体出于自身的价值观和心理习惯会持有不同的价值期望。这就意味着，不在同一个价值观体系里的所有讨论，其成效大多会大打折扣。

2）拥抱“大数据”是一种涉及多个专业领域和工种的集体协作行为，具有不同知识背景和方法论的不同个体如何有效地进行协作本身就是一个难题，需要在多边理解与互补的基础上互融互通、相互成就。

出于对“大数据”的敬畏，本书尝试将笔者面向不同群体陈述和演讲的观点加以梳理和规整，希望从不同的视角和层面分析和研判“大数据”这件事。但是“大数据”涉及的内容太多，笔者无法也从来没有想过，仅仅通过一本书就能详尽介绍其方方面面。本书尝试从相对宏观的角度对“大数据”进行介绍，希望通过这种介绍，能够让大家对“大数据”形成初步的印象，而具体的细节则需要通过其他途径深入了解。

本书的整体行文是基于“说些历史、讲些故事、聊些技术、谈些思考”这样的思路展开的。本书共有 13 章，逻辑上分为四篇，分别是：

第一篇　现象及感性思辨：本篇尝试对“数觉→数→数据→大数据”的历史脉络进行梳理，并通过社会各界迎接和拥抱“大数据”的若干事实，厘清几个基本问题：“大数据”是什么？为什么会有“大数据”？“大数据”得到各界关注的原因是什么？社会各界或各个利益主体是如何拥抱“大数据”的？本篇包括3个主要章节，分别是：

- ❑ 第1章　大数据溯源
- ❑ 第2章　大数据现象
- ❑ 第3章　大数据产业

第二篇　技术及选型思路：本篇尝试从技术实现和部署、实施的角度厘清大数据技术流程，并从多个视角和层面阐述各个环节所面临的挑战和机遇，重点叙述不同知识背景的研究群体针对大数据的态度、行动和思维方式。本篇将通过对“数据采集→数据存取→数据建模→知识发现”技术流程的梳理，以及各个环节技术选型的应用提示，厘清几个基本的技术问题：从何处及如何获得数据？将数据存储在何处及如何管理？如何分析数据以探明数据背后的知识和洞见？本篇包括6个主要章节，分别是：

- ❑ 第4章　大数据支撑技术
- ❑ 第5章　数据采集与整合
- ❑ 第6章　数据存储与管理
- ❑ 第7章　数据表示与理解
- ❑ 第8章　数据理解与建模
- ❑ 第9章　知识发现与应用

第三篇　实施及理性思考：本篇尝试从管理策略、价值实现及思维方式三个角度厘清大数据落地应用所涉及的技术和非技术问题，并从多个视角和层面梳理各个环节的要点和细则。此外，本篇将围绕“大数据实施及过程管理→大数据价值及价值评估→大数据思维及价值实现”这一主线，给出各个环节的应用提示，并厘清几个基本问题：“大数据”的价值在哪里以及如何实现？如何部署、实施一个大数据项目？应该以怎样的思维方式和执行策略应对“大数据”的挑战？作为一个数据分析师应该具有哪些情怀？本篇包括3个主要章节，分别是：

- ❑ 第10章　大数据实施
- ❑ 第11章　大数据价值
- ❑ 第12章　大数据思维

第四篇　机遇及应用思索：本篇在对互联网的技术发展脉络及国际经济形势进行梳理的基础上，分析了在“互联网＋”概念被热炒及全民总动员的当代，“大数据”的潜在发展机遇和应用场景，并通过对电子商务、工业4.0、互联网金融这三条主线的扼要描述和分析，厘清几个基本问题：“互联网＋”的本质是什么？究竟是“互联网＋X”还是“X＋互联网”？“互联网＋商务”“互联网＋工业”“互联网＋金融”的本质、门类及潜在机遇有哪些？作为一个

数据分析师，在“互联网 +”的环境下应该具有哪些情怀？本篇包括 1 个主要章节：

❑ 第 13 章　大数据机遇

本书的每一章都围绕某个专题展开叙述，独立成文，读者可以根据自己的兴趣和时间选择性阅读。本书的行文主要有两种字体：以宋体行文的部分主要描述相关理论、技术，以及必要的分析和说明；以楷体行文的部分大多是基于上下文罗列的一些佐证、案例和思考；此外，本书还分别以显式和隐式的方式给出了笔者对于一些技术选型、场景分析及感悟心得的“应用提示”。

本书的初始行文动机是“归纳所见所闻、总结项目经历、独立自主思考”，希望从客观、独立的第三方视角介绍和分析“大数据”及与“大数据”相关的技术观点、执行思路和关键问题。在不影响阅读的情况下，本书对所涉及的公司产品的介绍简明扼要，更多详细信息可参见给出的推荐链接或参考读物。希望通过这样的章节安排及内容梳理，本书能够给“大数据”相关工作者一些参考，比如有意建设大数据项目的创业型公司创始人、企业主或政府部门的主管及信息主管（作为技术丛书和应用手册），有意向数据型公司转型的传统软件企业技术人员，包括市场人员、研发人员和主管（作为应用指南及技术白皮书），处于战略转型期的传统企业主或信息主管（作为技术丛书），大数据项目研发工程技术人员（作为技术丛书及应用指南），普通院校大数据相关专业的研究生和本科生（作为教辅材料），或对大数据有兴趣的读者（作为科普读物）等。

本书的编写及出版得益于诸多前辈、同仁和南京大学计算机科学与技术系及南京大学软件新技术国家重点实验室的各位领导、同事的提携、关心和鼓励。感谢各类项目的资助方以及我所在研究团队“南京大学智能信息处理研究组”的小伙伴，多年来共同努力和协作完成的一个个项目为本书的撰写提供了大量素材，同时也让我有更多的可能性去思考所有这些成功（当然也有失败的）案例背后的大数据逻辑。还要特别感谢“南京大学智能信息处理研究组”的吴骏博士、张雷博士及彭岳、徐鸣、陆恒杨、王楠、李明、王陆霞、夏丽、谭龙海、陈鹏飞、冯艺琳、唐驰、谢璐遥、李永春等同学，在对本书进行通本润色的过程中，他们给予了极大的帮助。本书行文伊始，我就将本书的编写计划、组织结构与南京大学的谢俊元教授、陈家骏教授、郑滔教授进行交流和沟通，三位教授均给予了很多务实的建议，在本书的整个编写过程中，几位教授也在不同的场合、时机关注本书的编写进度。这些对于笔者而言都是莫大的支持和鼓舞，在此一并感谢。本书的编写和出版还离不开机械工业出版社华章公司姚蕾编辑的建议和鼓舞，尤其是行文过程中几乎不间断的支持和鼓励，才使本书得以顺利完稿，再次表示感谢。

由于水平有限及知识面、价值观的狭隘，书中有疏漏和不足之处在所难免，敬请各位专家和读者批评指正。

目　录 *Contents*

第一篇 Part 1

现象及感性思辨

科学技术的不断进步和人类需求的持续膨胀，就如两个互相咬合和彼此驱动的齿轮伴随着人类文明进程不断地发展：需求的膨胀不断刺激科学研究的持续进行，而科技的进步驱动着需求的进一步膨胀。目前社会各界普遍热议和关注“大数据”，恰是由于人们对数据价值的期望（需求）与目前对数据处理的研究和技术水平不匹配，并因为各界的普遍价值认同，“大数据”的概念被炒作得近乎神话。本篇尝试对“数觉→数→数据→大数据”历史脉络进行梳理，并通过社会各界迎接和拥抱“大数据”的若干事实，厘清几个基本的问题：“大数据”是什么？为什么会有“大数据”？“大数据”得到各界关注的缘由是什么？社会各界或者各个利益主体是如何拥抱“大数据”的？

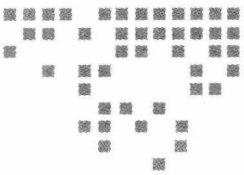
本篇包括3章内容，分别是：

第1章　大数据溯源　尝试梳理在人类文明进程的不断发展过程中，因为人类需求的不断膨胀和科技的不断发展而引发的，或许是必然的“数觉→数据”及“数据→大数据”的历史脉络，并介绍了不同利益主体对“大数据”的理解和定义。

第2章　大数据现象　尝试梳理“大数据”这一概念得到“热炒”的当下，因为不同的利益驱动使然而引发“政产学研商用”各界迎接和拥抱大数据的各类行动举措和思维态度，并分析了大数据这一概念得到多边热议的动机和缘由。

第3章　大数据产业　尝试梳理“大数据”这一概念持续得到多边认同的当下，因为不同的价值期望使然而引发不同利益主体迎接和响应大数据带来的机遇和挑战的各类决策和行动，并给出大数据潜在应用场景的理性评判依据。

【关键字】　大数据，历史脉络，多边热议，产业环境



第 1 章 Chapter 1

大数据溯源

在本章的写作及润色过程中，得到了南京大学计算机科学与技术系及智能信息处理研究组的杨骏元、汤兆亮、王陆霞、李明、唐驰、王姗姗等几位同学的协助，在此表示深深的谢意。

1.1 引言

140 亿年前，宇宙诞生……

46 亿年前，地球诞生……

38 亿年前，简单生命体出现……

1500 万年至 1000 万年前，腊玛古猿出现……

400 万年至 100 万年前，南方古猿出现……

200 万年至 150 万年前，能人出现……

200 万年至 20 万年前，直立人出现……

20 万年至 1 万年前，智人出现……

以上一组简单的数据勾勒出人类在整个历史长河中的进化史轮廓。与动物仅仅通过遗传进化不同，人类在进化过程中发展和演化出了一种非遗传性的继承：通过独一无二且日益发达的文化媒介，将知识和传统留给后代。这种文化传统使得人类以很快的速度和加速度进化并最终成为这个地球的统治者，而遗传进化退居于次要的位置。

这里所说的知识，指的是人类在改造世界的实践中所获得的认识和经验的总结归纳，可以指导解决实践问题的观点、经验、程序等信息。

或许正因为人类有了可以把知识和传统传递给后代的文化媒介，所以通过本身遗传系统所传递的信息也就愈来愈少。动物有许多生存的本领是通过遗传系统直接传递给后代的，而人除了吃喝哭喊之外，绝大部分生存本领只有靠后天学习。因此，发现知识、传递知识和学习知识是人类文明进程中亘古不变的主题。

野中郁次郎和竹内广孝在1995年提出的SECI（Socialization，Externalization，Combination，Internalization）模型专门阐述了知识构建和管理的完整过程：从噪声中分拣出数据，转化为信息，升级为知识，升华为智慧，让信息从庞大无序到分类有序。

数据是指描述事物的符号记录，是构成信息和知识的原始材料，如图形、声音、文字、数、字符和符号等；信息一般指数据所包含的意义，可以使数据所描述事件的不确定性减少。数据、信息和知识的关系可以描述为：数据是信息的载体，信息是知识的载体。知识可以从数据中发掘出来，即知识发现。

知识发现是从数据（库）中识别出有效的、新颖的、潜在有用的以及最终可理解的模式的过程，是将低层数据转换为高层知识的过程。其中的一个重要步骤是数据挖掘，也有很多场合将知识发现与数据挖掘有意无意地混淆。数据挖掘一般是指从大量的数据中通过算法发现隐藏于其中的信息的过程。关于数据挖掘与知识发现的详细介绍，参见9.2.4节，此处不赘述。

数据挖掘通常与计算机科学有关，通过统计、在线分析处理、情报检索、机器学习、专家系统（依靠过去的经验法则）和模式识别等诸多方法来实现上述目标，而机器学习是其中最为重要的基础，其基本的分析方法包括：聚类、分类、关联规则发现、预测等。机器学习是一门多领域交叉学科，专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。关于机器学习的详细介绍参见9.2.3节。

如上所述，作为信息的载体，数据是指描述事物的符号记录，鉴于其承担的功能，数据具有很多天生的特征（数据层特征），比如异构、分布、多态、多模式等。

1）异构性。数据的异构性主要表现在系统异构、模式异构、来源异构三个方面：

①数据源所依赖的业务应用系统、数据库管理系统乃至操作系统之间的不同构成了系统异构性。

②模式异构指的是数据源在存储模式上的不同。存储模式主要包括关系模式、对象模式、对象关系模式和文档嵌套模式等，其中关系模式（关系数据库）为主流存储模式。同时，即便是同一类存储模式，它们的模式结构可能也存在着差异，不同的关系数据管理系统的数据类型等方面并不是完全一致的，如DB2、Oracle、Sybase、Informix、SQL Server、Foxpro等。

③来源异构，即内部数据源和外部数据源之间的异构。

2）分布性。数据的分布性指的是数据的分布特征。一组数据的分布特征可以从以下三个方面进行测度：

①集中趋势测度，如众数、中位数、分位数、均值、几何平均数、截尾均值等。

②离散程度测度，如极差、内距、方差和标准差、离散系数等。

③偏态与峰度测度，如偏态及其测度、峰度及其测度等。

在某种意义上而言，正是由于数据具有分布性，我们才可以挖掘出数据内部的模式，并对数据进行分析以得到某种结果、做出某种策略。

3）多态性。数据的多态性即数据的多样性。一般而言，数据可分为两类：结构化与非结构化数据。在信息社会，信息可以分为两大类：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；而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。数据的多样性在现今世界表现得更加明显，由于数据采集与存储技术的不断进步，生活中所有的数据都可以保存下来，来源的广泛与用途的多样使得数据变得丰富而有价值。

4）多模式性。数据模式指对某一类数据的结构、属性、联系和约束的描述。数据模式是基于选定的数据模型对数据进行“型”方面的刻画，而相应的“实例”则是对数据“值”方面的描述。先有数据模型，才能据其讨论相应数据模式，有了数据模式，就能依据该模式得到相应的实例。正是数据的多模式性，才能够通过构造复杂的数据结构来建立数据之间的内在联系与复杂关系，从而构成数据的全局结构模式。

除了上述数据自身的特点，针对具体的应用场景，还需要考虑其他一些来自数据层的特征，比如数据的质量、数据的活度、数据的厚度、数据的规模等。

1）数据的质量：正所谓“失之毫厘，谬以千里”，数据是否具备可靠性和有效性，直接影响到数据分析过程以及是否能得出正确的结论并做出正确的决策。而在大数据时代，高质量的数据是大数据发挥效能的前提和基础，强大、先进的数据分析技术是大数据发挥效能的重要手段。对大数据进行有效分析的前提是必须要保证数据的质量，专业的数据分析工具只有在高质量的大数据环境中才能提取出隐含的、准确的、有用的信息，企业基于这些高质量分析结果所做出的各项决策才不至于偏离正常轨道；否则，即使数据分析工具再先进，在充满“垃圾”的大数据环境中也只能提取出毫无意义的“垃圾”信息。因此，数据质量在大数据环境下显得尤其重要。

2）数据的活度：数据的活度即数据持续更新的频度，也称为数据的新鲜度。以银行借贷业务和手机通话记录为例，一般来说，我们的工资每月发放一次，而手机通话记录则每天每小时都会产生，详细地描述了一个人与其他人交往的情况，加上个人在互联网中娱乐、购物等记录，形成了一个人生活轨迹的画像。一般数据活度越大，其可参考的价值可能就越高。比如电商平台在夏季向用户推荐商品时，使用用户最新（春夏）的购买数据所做的分析，显然比使用往年（春夏）的购买数据所做的分析更为准确和有意义。数据的这种特性依赖于人类的活动，如果用户发生某类行为的次数越频繁，则该类数据的活度也就越高，反之亦然。根据数据的活度进行分析，可以得出很多有意义的结论，比如根据数据活度将用户分类等。

3）数据的厚度：数据的厚度指的是数据的维度大小以及每个维度的语义丰富程度。一般来讲，一条数据的厚度是不定的，也就是说不同领域中的数据厚度是不相同的，可能很大，也可能很小。如在统计领域中，一条具有完整意义的数据一般有时间、地域和指标三个维度。维度一般是越多越好（往往厚度越大），以一个用户为例，如果你仅仅知道他的姓名、住址、电话等信息，你对他的了解就很有限；如果你知道他的体育爱好（比如打篮球、打乒乓球等）、文化爱好（比如喜欢读史学作品等），你可以对他进行更加有针对性的营销（语义信息

更多），比如推荐NBA球星的球鞋、推荐《史记》和《全球通史》等书籍，这样成功的概率就大。当然在分析数据时，一味地追求高维度分析，也会使得分析结果变差，这是由于维度可能是噪声，或者维度之间存在相关性，或者维度与具体问题不相关。也就是说，数据的厚度不是单纯的维度问题，还包含与其蕴含的语义丰富度。

4）数据的规模：数据一般分为结构化和非结构化数据，一般的数据库属于关系型数据库，以存储结构化数据为主，级别在TB级。随着数据，比如通话详单的积累，数据会变得越来越大，甚至达到PB、EB、ZB级。大数据区别于传统数据的一个显著特性是数据的规模，因此产生了如何进行存储、并行计算、挖掘数据内部模式等话题，进而形成了大数据概念，催生了大数据时代。

阿基米德说过："给我一个支点，我能撬起地球。"仿照类似的语调，微软的史密斯这样说："给我提供一些数据，我就能做一些改变。如果给我提供所有数据，我就能拯救世界。"现在看来，阿基米德的话语虽然很朴素，但是在那个年代能够提出，还是很振奋人心的。虽然我们没有生活在那个年代，不过后面的这句话，我们都是当事人，因为我们现在正生活在这样的大数据时代。

一般意义上，大数据是大到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。此定义或许还不足以完全描述大数据，因此附加了一个普遍认可的4V特征：

1）数据来源多、体量大，描述为"Volume"。这个特征隐含的应用提示是：在具体的部署实施过程中，数据存取必须支撑海量的数据并发访问，并且数据处理的性能必须支持海量吞吐率。

2）数据来源广、类型多，描述为"Variety"。这个特征隐含的应用提示是：在具体的部署实施过程中，数据存取必须支持多格式、多模式的高效管理，以及数据分析手段必须支持多格式、多模式的有效分析等。

3）数据来源速度快，描述为"Velocity"。这个特征隐含的应用提示是：在具体的部署实施过程中，数据采集速度要快、数据存取速度要快、数据分析速度要快，而所有的这些要求都对相关的技术选型、策略定位有很大的影响。

4）数据有用、有价值，描述为"Value"。这个特征隐含的潜台词是：价值密度稀疏。这是因为数据的有用和有价值都是有目标指向的，确切地说，应该是数据对于某个具体的应用而言是有价值的。由于数据量巨大，因此对于某个具体的应用而言，海量的数据中可能只有一部分是有价值的。

显然上述的定义及4V特征的描述只是一个普适的通用描述，在具体的落地实施过程中，不同的公司及团体也会根据不同的价值观和方法论提出其他具体特征。

IBM提出的真实性（"Veracity"）特征是从大数据部署实施过程中数据质量的维度考虑的，IBM认为：真实性是当前企业亟待考虑的重要维度，将促使他们利用数据融合和先进的

数学方法进一步提升数据的质量，从而创造更高的价值。

引发人们对大数据关注的原因有许多，其中之一必定是数据有价值。如果说知识及其获取是一件有价值的事情的话，那么如何实现“数据→信息→知识”就是一个重要的环节。与这个技术流耦合的关键问题还有：数据在哪里，如何获得数据，知识如何体现在具体的应用中以获取其价值等。这些问题比较复杂，因为它们与具体的应用场景、应用模式和商业模式有直接的关系。

本章尝试解释一些更简单的问题，比如：人生来就有数据这个概念吗？数据是如何产生的？如何使用这些数据？为什么数据会变“大”，以及大数据得到如此关注的缘由在哪里？为了有效回答上述问题，本章尝试梳理在人类文明进程的不断发展过程中，因为人类需求的不断膨胀和科技的不断发展而引发的或许是必然的“数觉→数据”及“数据→大数据”的历史脉络，并介绍不同利益主体对大数据的理解和定义，本章下面的结构安排如下：1.2节介绍人类还处于蒙昧阶段就已经具有的“数觉”以及作为地球主宰的生物种类如何从一开始仅有的“数觉”逐步产生出“数”的概念；1.3节介绍因为人类的聪明才智发明和创造的一些计算工具（包括模拟的和数字的），更具有变革意义的是，图灵机及全电子通用计算机的发明直接将人类推进到如今丰富多彩的信息化时代和本文关注的大数据时代；1.4节简单介绍数据的产生和催生大数据时代来临的若干技术和非技术缘由；1.5节简单叙述和回顾我们正生活的大数据时代。

1.2 数觉及数的起源

所谓数觉，指的是在一个小的集合里，增加或者减去一个元素的时候，尽管未曾直接知道增减，也能够辨认到其中有所变化。有研究表明若干种动物具有数觉，而人是否有数觉则是一个难以研究和回答的问题。

有个农场主计划打死一只在望楼里筑巢的乌鸦，试了多次，始终未成功。因为人一走近，乌鸦就离开了巢，飞开了，它栖在远远的树上守着，等到人离开了望楼，才肯飞回巢。这样的试验继续，两个人走进望楼，一个人留着，一个人出来走开了，但是乌鸦并不上当，它等着直到留在望楼里的人也走了出来才罢。试验一连做了几天：两个人，三个人，四个人，都没有成功，五个人的时候成功了。五个人首先都进了望楼，留一个在里面，其他四人走出来，离开了。这次乌鸦却数不清了，它不能辨别四与五，马上就飞回巢了。

实际上，我们所知道的最惊人的例子要算一种叫作“独居蜂”（solitary wasp）的昆虫。这种母蜂在每个巢里下一个卵，并且在巢里面预先储藏了一批活的尺蠖，作为幼虫孵化后的食料。使人吃惊的是，各类独居蜂在每个巢里所放的尺蠖数目都是一定的：有些放五条，有些放十二条，多的甚至于有放二十四条的。最特别的是一种叫作“蜾蠃”的蜂，这种蜂雄的比雌的小得多。母蜂能用神秘的方法辨别孵化出来的幼虫是雄的还是雌的，并且据此相应地分