大数据

理论与工程实践

THEORY AND ENGINEERING PRACTICE

BIG DATA

陆 晟 刘振川 汪关盛等 编著

国际数据管理协会（DAMA）
中国分会主席

胡本立 作序推荐

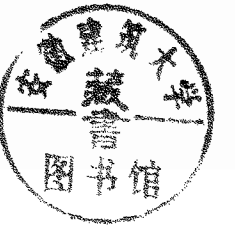
中国工信出版集团

人民邮电出版社
POSTS & TELECOM PRESS

大数据
理论与工程实践

陆 晟 刘振川 汪关盛等 编著

BIG DATA
THEORY AND ENGINEERING PRACTICE



人民邮电出版社
北京

图书在版编目（CIP）数据

大数据理论与工程实践 / 陆晟等 编著. -- 北京 : 人民邮电出版社, 2018.12
ISBN 978-7-115-49683-6

Ⅰ. ①大… Ⅱ. ①陆… Ⅲ. ①数据处理－研究 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2018)第231227号

◆编　　著　陆　晟　刘振川　汪关盛　等
责任编辑　朱玉芬　鄂卫华
责任印制　周昇亮
◆人民邮电出版社出版发行　　北京市丰台区成寿寺路 11 号
邮编 100164　　电子邮件 315@ptpress.com.cn
网址 http://www.ptpress.com.cn
大厂聚鑫印刷有限责任公司印刷
◆开本：720×960　1/16
印张：19.5　　2018 年 12 月第 1 版
字数：195 千字　　2018 年 12 月河北第 1 次印刷

定　价：58.00 元

读者服务热线：（010）81055522　印装质量热线：（010）81055316
反盗版热线：（010）81055315
广告经营许可证：京东工商广登字20170147号

推荐序

汪关盛先生邀请我为本书写序，粗粗翻阅后，我的第一个反应是，我可能没有足够的经验和知识来写；接着，细看之下，就被它的内容吸引，书中有许多我希望了解但不知道去哪里找的信息和知识。关于本书的定位、写作过程和阅读对象等，作者团队已经做了很好的介绍，我在此就跟大家简单分享一下我对本书独到之处的一些粗浅看法。

与传统数据处理相比，大数据由于其属性和量级的不同，处理起来也需要遵循不同的理论和采用不同的手段。本书对数据的收集、存储和处理，CPU 及网络等资源的分配和同步等做了全面和详细的介绍，是一本关于大数据理论和工程实践的不多见的好书，内容比我读过的其他讨论大数据技术的书要更广泛和深入。本书有助于读者了解大数据从蓝图设计到工程落地需要考虑和到位的各层技术。

在阅读和学习的过程中，我觉得本书有以下几个特点。

1. 与现在许多为吸引眼球起个名头大的标题而缺乏实质内容的作品相比，本书的做法正好相反。书中各章节的标题看似很普通，但下面包含的内容却极为丰富，体现了作者对大数据理论和工程问题了解的深度。作者在各章节中引用了一些原创和权威资料，同时适当配置了一些程序作为例子，使我感动于他们的专业精神和为此付出的大量努力。

2. 大数据的工程理论和实施技术十分复杂，本书进行了系统的讲述。对工程的每一步、每一层均有详细介绍但内容间并不孤立，一环扣一环，上下文有机关联，从大数据的应用到配套的软、硬底层基础，一气呵成。不少技术书往往就技术论技术，本书能结合应用和应用的需求谈技术，也是它的独到和可贵之处。

3. 本书把大数据的工程理论和实践与人工智能结合起来讨论。我一直希望能把传统的数据处理与大数据、人工智能关联理解和融合，这本书的内容和设计对我有所启发。虽然这三个领域各有各的侧重点，但是最终，业务拓展、企业运营和市场开拓一定都需要基于数据的应用和技术，而不管它们需要及处理的数据类型或属性是否相同。这本书为理解大数据、大数据处理及人工智能如何互联互通搭建了一个桥梁。

数据行业经过多年的发展，已成为当前数字经济的主要部分。同时，如所有专业和行业的发展过程一样，它必然会发展出更细和更专门的子领域。我觉得这本书的出版可以加强从事各数据行业子领域的专业人士间的沟通和了解，对整个数据行业的协同发展也有很强的理论和现实意义。

胡本立

国际数据管理协会（DAMA）中国分会主席、世界银行前首席技术官

2018 年 10 月 13 日于华盛顿

序言

2017年年初，我参加北大组织的大数据人才交流论坛时，无论是从会上的嘉宾发言中还是从会下同仁的交流中，都很容易得出一个结论，那就是大数据人才是非常非常匮乏的，各大企业、院校及组织都不得不用互相挖角的方法寻找相关人才。2017年，大数据已经发展了十多年，相关书籍也是汗牛充栋，因此，出现这种人才匮乏的状态，确实有些令人疑惑。仔细想想，也许是以下这些原因导致的：首先，大数据的应用领域扩展太快，人才的培养速度跟不上；其次，大数据技能的学习周期比较长，进入门槛较高；最后，大数据本质上是一种工程应用，在不同领域哪怕使用相同算法，或者在相同领域使用不同数据，算法都需要一个调试和优化的过程，这就要求学习者领悟原理，而不能简单地照着葫芦画瓢，而领悟原理的要求又和工程应用的实用性需求有一定差异。

到了那年四五月份的时候，和刘振川先生、甘智峰博士讨论后，我们都觉得可以把我们多年相关工作和经验总结一下，写一本比纯粹的工程应用更理论一些、比纯粹的理论介绍更实用一些的书。这样的书面向大数据工程师，帮助受过基本训练的工程师开发出系统，达到实用目的。由于我们三人知识面有局限，便又邀请了周翊博士和金淬博士加入，他们在各自的领域都有很丰富的实战经验。

晚些时候，我又认识了在国内大数据领域做过很多工作和进行过投资的

潘磊先生。潘总又给我介绍了国际数据管理协会（DAMA）中国分会资深顾问汪关盛先生，还有母润坤先生。通过和他们的沟通，我们才意识到我们原本的计划是不完备的。我们一直关注数据处理，可是在实际应用中，很多时候面临的不是如何处理已有的数据，而是如何管理和治理已有的数据。“数据过多就相当于没有数据”，这句话不仅仅指我们需要用算法发现大量数据背后的价值，同时也指我们需要去芜存菁，从更有价值的数据中以更小的代价发现更高的价值。从事大数据行业一段时间的人都会有两个感受：很多时候数据源比算法有价值，获得好的数据源总能得出有价值的结论；事后再看大数据分析出的结论，往往发现那些结论很直观。这些都体现出数据治理和项目实施管理的价值。汪总和母总为我们的计划补上了最后的拼图。

经过了差不多一年半的努力，我们终于完成了规划的小目标，结果发现好像已经错过了大数据图书的热卖期。后来，人民邮电出版社的缪永合先生对我们的努力给予了认可，并支持我们把这本书出版发行。我们在此也感谢本书的编辑团队。

书中关于高速缓存、集群总线、资源调度、用户画像和广告投放的实用内容都来自刘振川先生的实践。第 6 章数据治理的内容则来自汪关盛先生的长期经验。第 7 章大数据在人工智能领域的应用是周翊博士、甘智峰博士和金淳博士的专长，他们分别贡献了语音部分、视觉部分和博弈部分的内容。母润坤先生则将他多年来实际的大数据处理和实施方法总结在了第 1 章的相应部分。

由于作者团队一直在第一线工作，理论基础研究相对比较薄弱，为了让更多读者有更深入的收获，我们梳理、借鉴了很多经典的论文和网络资源。书中也对引用和借鉴的资料标明了来源，在此对相关资料的著作者表示感谢！

研究与实践都还在不断发展中，诚挚希望有关专家与专业人士给予宝贵的意见和建议，共同推动大数据事业的快速发展！

谢谢！

陆　晟

前言

大数据是近年来炙手可热的一个词汇。无论是国家还是企业，都希望从大数据产业的发展中获益，而科学家、工程师们也希望在这个新兴的行业中获得较高的回报。因此，市面上大数据相关的书籍也快速丰富了起来，从概述类的书到具体介绍某项技术的书，应有尽有。而本书则从工程实践和基础理论角度讲述大数据的应用，为不同的大数据应用场景提供了思路。

目前，在实际应用中，人们往往通过架设 Hadoop，以及基于 Hadoop 生态的各种系统来满足大数据应用需求。然而，不是所有的大数据应用都适合用 Hadoop 的数据存储方式、系统架构和计算模型。例如，对于高实时性要求或者高并发的应用场景，Hadoop 就不适合，因此出现了许多基于 Hadoop 生态的扩展，以解决某些特定类型的问题。

近年来，大数据技术一直处于高速发展中，很多两年前非常流行的技术逐渐淡出或者销声匿迹了。作为大数据业务的开创者和领头羊，Google 公司从未停止过对技术的改进甚至颠覆，例如将数据存储从 GFS 发展到了 BigTable，也推出了 Dremel 和 Pregel 等新的计算框架。这是因为 Google 的工程师了解需求，也知道这些需求背后的技术原理，懂得根据需求权衡和选择最适合特定需求的技术路线和方案；而不是只有榔头这一个工具，导致看任何问题都像是钉子，而解决问题的手段也只有敲击这一项。

本书不是大数据技术手册，也不是某种具体技术的说明；而是面对具体应用场景时的技术考虑和权衡。在实际应用中，各类大数据应用方案没有优劣之分，只有适合或不适合的差异。甚至大部分情况下，任何选择都需要付出代价，而针对这种收益和代价的衡量及评估才是本书所关注的。此外，书中也会出现一些具体的示例代码，作者提供这些示例代码，希望体现其背后的原理，即使某段代码采用了特定的语言和系统，也不代表在该场景下推荐使用该语言及语言所依赖的系统。

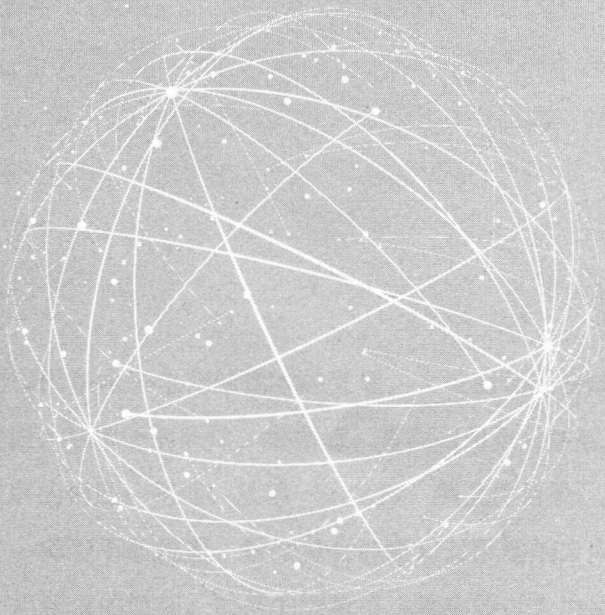
本书通过探讨技术原理，帮助读者选择合适的工具，或者自行开发适合自己应用场景的工具，无论这个工具是榔头还是钻子，是刨子还是螺丝刀，甚至是目前还不存在的某种类型的工具。作者团队衷心希望本书能为国内大数据企业建立自己的技术特色和技术优势贡献微薄之力。

本书目标读者群：主要面向架构师，或者是有具体大数据问题需要解决的工程师；也适合从零开始搭建大数据架构，或者需要将现有的非大数据的需求修改成大数据方案的读者和相关专业学习者。同时，对于那些实际上正从事大数据相关工作而自己并不清楚这一点的个人或企业，本书也能给你们带来启发。

非本书的目标读者群：希望通过教科书式学习从而掌握大数据的某项具体技术的读者；希望通过一本书就知道大数据是什么，从而可以找到一份大数据工作的人士。

本书作者都长期从事大数据相关的工作，对于很多具体的技术有自己的看法和独到见解，也真正踩过很多坑。由于应用场景的不同，作者对于技术的理解和认识也可能存在差异。我们希望这本书的推出能够抛砖引玉，涌现出更多精彩著作。

目录



第1章

概述

随着数据的编码和电子化存储技术的发展，大数据现已变成了一种被广泛运用的技术手段。从单一的照片、到相册、再到相册集、然后到家庭相册、再到千千万万家庭的相册；同理，从单一的文件、到文件目录、到文件系统、再到磁盘阵列……随着不断积累，数据总会朝着与日俱增的方向发展。而随着使用人数和使用场景的增加，数据的增量很快就会超过人力所能处理的范畴。个人拍摄的照片尚可自行处理和筛选，而无处不在的监控就不可能再用人工方式全面地查看了。因此，在一定程度上，各种问题最终都会转化成大数据问题。

关于大数据意义和作用的文章和著作有很多了，例如，吴军博士在《智能时代》一书中列举了大量生动的例子，我们就不再重复。我们写作本书的目的，是为了说明在工程上使用大数据时的各种具体考量。

大数据处理的特征

随着数据日积月累，需求的应用场景也会越来越丰富。那么，大数据到底是如何被处理的呢？对很多人来说，大数据只是一个概念，而工程师面对的却是待解决的实际问题。他们需要解决这些问题，至于是不是用大数据的方式，一开始未必就能确定。也许他们一开始并没有意识到需要用大数据。

当他们发现：我的天啊！数据怎么这么多！我的程序跑个基本处理竟然要五个小时！这时，就该大数据出马了。

当你发现，需要解决的问题具备几个共同特征，那么这个问题就可以运用大数据手段去解决。也就是说，这个问题基本上就可以算是大数据问题了。

我们总结了需要利用大数据技术手段处理的数据的三大特征。

第一，数据量大。至于数据量大到什么程度才算大数据，并不存在统一的硬性标准。在不同的历史时期和软硬件条件下，数据量标准也是不同的。但不管怎么说，当数据量大到用一台处理器处理不过来、多到用单一存储设备难以存下时，就需要采用大数据手段了。

第二，数据一般带有时间属性。对有些数据来说，时间是主要属性，例如，在某个时刻的设备状态监控信息。而对另外一些数据来说，虽然时间不是最重要的属性，但也是属性之一，例如，某首歌曲或者某部电影，虽然大家关注的是其内容，但是它们同时也具有产生和被使用的时间属性。

第三，数据一般具有多个属性维度。单一属性的数据虽然可能量也很大，但是从处理和分析的角度来看，数据往往可以被分为很多详细的属性，而这些属性之间的关联和关系才是最有价值的。例如，监控视频包含的也许都是单一的图像数据，而需要被处理的常常是这些图像被分析之前的元数据以及被分析之后的详细数据。例如，采集视频的时间和采集时的地理位置、图像的分辨率是元数据，而图像分析之后得到的人数、天气情况、是否存在需要关注的异常事件等，就属于含有更详细的维度的信息。

IBM 公司提出大数据有 5V 特征，分别是大量（Volume）、高速（Velocity）、真实（Veracity）、多样（Variety）和低价值密度（Value），它们可以用来说明大数据的数据量大、需要的处理速度快、对数据质量的追求高，同时数据的来源往往很不同，以及价值密度的高低与数据总量的大小成反比等特性。此外，还有人认为大数据的特征是体量大、可分析的维度多、数据

完备性重要，以及数据不能够用传统方式处理。[①] 这些特性分析和理解当然是没错的，但从事物的不同角度看，关注的重点、可以进行的分类和得到的结论会不同，因此本书中提出的三项大数据特性更多关注的是大数据项目的实施属性，所以我们也称之为大数据处理的三大特征。

基本处理模型

大数据技术是一种帮助数据实现价值的技术手段。挖掘出数据中的价值，才是大数据的应用目标。大数据技术虽然是新兴的数据处理技术，但它与传统的数据仓库等技术相比，数据处理的核心模型并没有发生多大的变化。以前做过传统的数据仓库管理等工作的人转行做大数据，就会发现后者仅仅是处理步骤对应的技术产生了变化。

传统的数据类问题的解决可以分为四个基本步骤：数据采集、数据存储、数据分析和数据使用。前三个步骤都很直接，而所谓数据使用则有不同的表现形式：可能是用图表对数据进行展示；也可能是利用分析结果做出某种决策；还有可能带来另一轮的采集、存储、分析、使用过程，即在前一轮分析的基础上对结果进行新一轮处理。以前文提到的监控视频数据为例，第一轮采集的数据可能是视频流本身。这些视频数据和元数据（例如采集时间、采集地点）需要被保存下来，然后根据不同需求做出不同的分析，例如分析其中车辆的信息、车牌号码、是否违章等。至于这些数据的分析结果，可以是按时间统计的车辆通行量的图表；也可以是提交给交通管理部门的违章信息；还可以根据不同时间和不同位置的通行情况进一步分析车辆，从而画出车辆

① 这四项特性来自吴军博士所著《智能时代》一书的第二章，其中关于多维度的解释同本书的观点不同。本书强调的是数据存在多维度属性，吴博士强调的是数据可以被多维度分析。