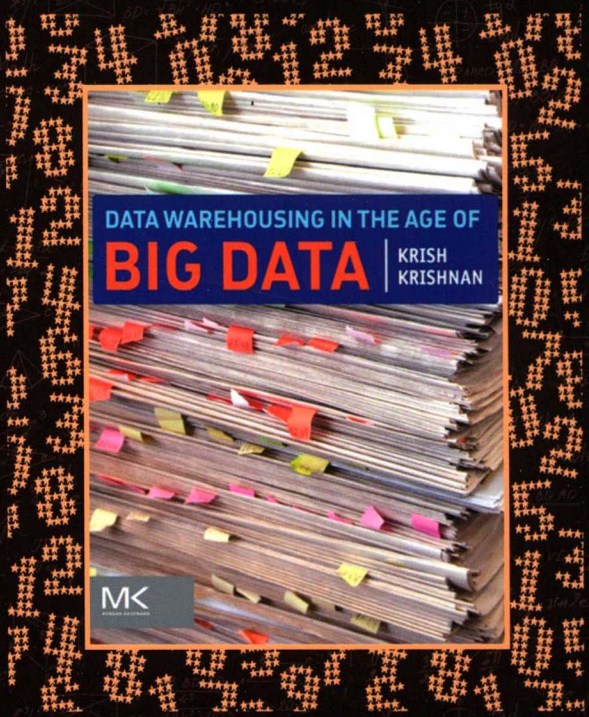
ELSEVIER
HZ BOOKS
华章IT

数据科学与工程技术丛书

# 大数据与数据仓库

## 集成、架构与管理

[美] 克里什·克里希南（Krish Krishnan）著
邢春晓 张勇 张桂刚 译




DATA WAREHOUSING IN THE AGE OF
BIG DATA

机械工业出版社
China Machine Press

数据科学与工程技术丛书

DATA WAREHOUSING IN THE AGE OF

# BIG DATA

# 大数据与数据仓库

## 集成、架构与管理



[美] 克里希纳·克里桑南（Krish Krishnan）著
陈春晓 张勇 张桂刚 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

大数据与数据仓库：集成、架构与管理 /（美）克里什·克里希南（Krish Krishnan）著；邢春晓，张勇，张桂刚译 . —北京：机械工业出版社，2018.4
（数据科学与工程技术丛书）
书名原文：Data Warehousing in the Age of Big Data

ISBN 978-7-111-59482-6

I. 大… II. ① 克… ② 邢… ③ 张… ④ 张… III. 数据库系统 IV. TP311.13

中国版本图书馆 CIP 数据核字（2018）第 056562 号

本书版权登记号：图字 01-2013-7854

Elsevier
Elsevier (Singapore) Pte Ltd.
3 Killiney Road, #08-01 Winsland House I, Singapore 239519
Tel: (65) 6349-0200; Fax: (65) 6733-1817

Data Warehousing in the Age of Big Data
Krish Krishnan

ISBN-13: 978-0-12-405891-0



《大数据与数据仓库：集成、架构与管理》（邢春晓 张勇 张桂刚 译）
ISBN: 978-7-111-59482-6



# 大数据与数据仓库：集成、架构与管理

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：李秋荣
印　　刷：北京市荣盛彩色印刷有限公司　　版　　次：2018 年 4 月第 1 版第 1 次印刷
开　　本：185mm×260mm 1/16　　印　　张：17.75
书　　号：ISBN 978-7-111-59482-6　　定　　价：79.00 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

# 译 者 序

数据仓库最早用于为企业决策提供所有类型数据支撑的大型数据集合。随着大数据时代的到来，数据量越来越大，对数据的处理速度越来越快，同时数据的价值密度也变得越来越小。大数据时代下具有 4V 特征的数据环境中，数据仓库又会变得怎样呢？本书为回答这一问题做了详细的分析和呈现。

目前对数据仓库进行论述的著作都没有明确指出大数据这一特殊时代的特点以及面临的问题和相应的解决方法，而本书将数据仓库与大数据两者进行了有机融合与统一，从而使得数据仓库技术符合大数据这一特定环境的需求。同时，也为新型企业基于大数据 – 数据仓库的管理决策提供了有效的途径。

本书逻辑清晰，内容较为全面，具有很强的适用性。全书由三大部分组成，分别为大数据、数据仓库以及构建大数据 – 数据仓库。第一部分主要包含：大数据简介，使用大数据，大数据处理架构，大数据技术简介，以及大数据驱动的商业价值。第二部分主要包括：再论数据仓库，数据仓库的再造，数据仓库中的工作负载管理，以及应用到数据仓库的新技术。第三部分主要包括：大数据和数据仓库的集成，大数据的数据驱动架构，大数据的信息管理和生命周期，大数据分析、可视化和数据科学家，以及实施大数据 – 数据仓库的现实情况。值得一提的是，本书还对客户案例研究与建设医疗保健信息工厂两个案例做了分析，有助于读者更好地理解本书。

本书的翻译工作主要由如下人员完成。清华大学信息技术研究院邢春晓研究员负责译稿的审校工作，清华大学信息技术研究院张勇副研究员和中国科学院自动化研究所张桂刚副教授负责本书的翻译工作。

# 前　言

Web 2.0 改变了我们的生活和工作方式，比如开展业务、与客户沟通、与朋友和家人共享信息、用业务收入和客户花销份额来衡量成功，以及定义品牌管理。最重要的是，它创造了一种独一无二的生财之道。无论是安排度假地点、购买最新型的电视、更换移动服务供应商，还是想要为郊游买最好的食材，你都可以通过互联网查看顾客的评论和读者的推荐。同样，在个人生活中，你可以使用 Facebook、YouTube、iTunes、Instagram 和 Flickr 分享你喜欢的音乐、电影、照片和视频。

当今，企业所提供的产品和服务的个性化为消费者创造了许多机会，同时也大大促进了数据量增大、数据格式（品种）增多和数据生产速度加快。数据的关键价值是，当我们使用地理和人口学数据建模来创建关于相似人群的个性、行为和影响的聚类时，能够找到在数据中隐含的智慧。

向服务的个性化和以客户为中心的商业模式进行转变形成了三个不同的趋势。

- **众包**。这是 Jeff Howe 于 2006 年在《连线》杂志上提出的术语。众包是在当今世界使用协同智能研究人类行为的过程。信息管理和个人层次上的推荐共享共同形成了业界的趋势。
  - 众包已演变成一个强有力的工具。它现在在商业上有很多用途，例如寻找有竞争力的研究、客户情感分析和因果分析等。同时还部署了其他的分析模型，例如协作过滤、推荐和机器学习算法。
  - 众包的最佳案例之一是当时身为参议员的奥巴马在 2008 年的总统候选人提名竞选中筹款。通过使用互联网和社交媒体作为一种个性化的联系渠道，他在筹措资金方面明显超过了其他候选人，从而能够进行有效的竞选。
- **社交媒体分析**。今天的消费者依靠的数据和信息是通过社交媒体渠道获得的，而这些数据和信息又依赖于将这个平台作为其“个人决策支持平台”的广大用户所做出的个人决策。这使得更多的人利用社交媒体作为与客户、合作伙伴和供应商直接和间接的沟通渠道。今天，如果你没有使用社交媒体，那么你是过时的，尤其是与 90 后和新千年的客户群相比。
  - 如何度量你的社交媒体渠道和沟通策略的有效性？这表明你从哪里开始实施一项社交媒体分析战略。该战略应从两个角度进行度量，包括从内向外和从外向内。在这一领域一个企业的成熟和演变往往需要经过多个阶段。在现在的新闻和互联网上，你会发现一些使用该策略成功进行业务变革的例子。
- **游戏化**。今天的另一个热门趋势是在企业内外使用游戏化策略来吸引员工、客户、潜在客户和任何对你的业务和服务感兴趣的人。

- 游戏化基于博弈理论和统计模型的组合，在对“长尾”得到的结果进行建模时，这已成为一种非常有效的工具。它也是在 Web 2.0 时代由 Chris Andersson 定义的术语，他还专门就这个主题写了一本书。
- 这一策略最好的例子是在 2012 年美国总统大选中，竞选策略专家使用博弈论和统计模型寻找目标选民，而且非常有效。奥巴马总统的竞选专门使用这种组合作为一种有效的和颠覆性的策略，从个人层次在候选人和选民之间创建了很多需要的连接。

从 Web 2.0 的观点来看，所有趋势、理论和成果的共同主线可以归结为两点。

- 使用 Web 2.0 平台所需要的数据量远远大于现今企业所用到的。
- 在计算的历史上，使用统计模型和分析的需求比以往更加强烈。

这两个事实已经被 Facebook、Groupon、Google、Yahoo、Apple 和其他财富 500 强公司证明是成功的。

数据带来了如下问题：如何计算海量和多样的数据，以及如何应对数据体量。这是 Google、Facebook 和 Yahoo 清楚展示的方式；前者创造了一种新的计算模型，该模型基于文件系统和一种叫作 MapReduce 的编程语言。MapReduce 扩展了搜索引擎的能力，能够同时处理多个查询。2002 年，架构师 Doug Cutting 和 Mike Carafella 正在做开源搜索引擎项目 Nutch，这促使他们基于 Google 模型来对底层架构进行建模。这也使得 Nutch 成为一个开源的顶级 Apache 项目。该项目于 2006 年被 Yahoo 所采用，称为 Hadoop。在过去的几年中 Hadoop 成就了大量的公司，这些公司有商业化的解决方案，同时将相应功能回馈给基础的开源项目，这是一种真正基于协作的软件和框架开发。

另外一项技术也演化为一个强大的平台，即 NoSQL（Not only SQL）运动。该平台基于 Eric Brewer 在 2002 年提出的 CAP 定理。根据 CAP 定理，一个数据库不能在任何一个时间点满足 ACID 兼容的所有规则，同时又是可扩展的和灵活的。然而，在一致性、可用性和分区容忍性三个基本性质中，一个数据库可以满足三个性质中的两个，从而创建可扩展的分布式结构，该架构可以演变成满足水平方向上缩放的可扩展性要求并提供更高的吞吐量，因为在这种环境中计算和存储是非常接近的，同时是一个允许多种一致性级别的分布式架构。

Facebook 是 NoSQL 架构的最早提倡者之一，因为他们要解决用户的可扩展性和可用性要求，其用户量仅次于中国和印度的人口。Cassandra 是一个流行的数据库，在 Facebook 经历了很长时间的开发和使用（现在由于更大的可扩展性需求，它已经被 Facebook 抛弃）。许多其他公司把它与 Hadoop 以及其他传统的 RDBMS 解决方案一起使用。它仍然是一个顶级的 Apache 项目，并且正在添加更多的功能。

随着这些新技术和商业模式的出现，也出现了大量噪音，并导致了混乱。这些趋势或噪音之一是“数据仓库的死亡”，这在全球都带来了严重影响，因为企业已不只投入数以百万计的美元来搭建这种决策支持平台，而且基于其输出开发了若干下游系统。

作为传统的数据仓库领域和大数据领域中都有经验的数据实践者和咨询师，我开始在数据仓储研究所（TDWI）教授课程，在许多国际峰会和其他会议中谈论大数据和数据仓库，以消除数据仓库的“死亡”所带来的恐怖。在过去四年中，在全球关于这个话题展开了大量讨论之后，我决定写这本书并讨论大数据。包括谁使用大数据，它是如何

影响数据仓库世界的，以及数据分析的未来，更重要的是，下一代数据库仓库的概念以及它是如何构建的。

坦白地说，我们将继续构建和使用数据仓库，而且它将仍然是“单一版本的事实”，但我们将不再使用 RDBMS 作为数据仓库和分析的平台。在写这本书的时候，我们看到每隔几个月，有时是几周，Hadoop、MapReduce 和 NoSQL 就会发生变化，新功能就会浮出水面。人们正在设计和搭建这些架构，它们可以处理大型和复杂的数据，能够在批处理环境中有效处理数据，但是比起关系数据库管理系统在实时和交互能力方面比较有限。该架构的最终状态将是这些架构的异构组合，以共同创建一个强大和巨大的决策支持架构，这个系统的名称依然是数据仓库。

在读这本书时，你会发现三个不同的部分。第一部分讨论大数据，包括大数据技术及来自早期实践者的用例。第二部分介绍数据仓库、它失败的原因、新的架构选项、工作负载、工作负载驱动的架构，以及大数据和数据仓库的集成技术。第三部分涉及数据治理、数据可视化、信息生命周期管理、数据科学家，以及适合大数据的数据仓库。附录包括来自供应商的实现和一个关于如何建立医疗保健信息工厂的案例研究。

本书的总体目标是帮助你了解大数据和数据仓库的复杂层次，同时为你提供关于如何有效使用所有这些技术和架构来设计下一代数据仓库的信息。

下面描述各章的内容和全书组织结构，为你提供阅读路线图。在逐章阅读时，这些章节结合起来就会为你提供简洁而深入的理解。

## 第一部分：大数据

第 1 章的重点是让你彻底理解大数据。我们避免使用流行词，探讨了新兴的大数据领域和它对企业的重要性。

第 2 章的重点是大数据隐含的复杂之处（即三个 V——体量、速度以及多样性和多义性），如何处理这些特点，以及在这些主题域有哪些隐藏的陷阱。

第 3 章重点讨论需要或者设计什么架构以进行大数据处理，还讨论了算法级的方法、分类系统、集群和其他内容。

第 4 章重点讨论的是为解决大规模数据处理，核心技术是如何演化的。这些技术包括 Hadoop 及其生态系统、NoSQL 数据库和其他技术。这一章对于这些技术的介绍是极其浓缩的，建议你进一步阅读有关这些主题的核心书目。

第 5 章论述在现实生活中不同公司利用大数据实现价值的各种用例。这些用例涉及 B2B、B2C、C2C 等场景，该章还介绍在每个场景中是如何定义和实现价值的。

## 第二部分：数据仓库

第 6 章重点追溯数据库仓库的起源以及这些年来的演化。该章讨论早期版本的缺陷所导致的数据仓库的失败，以及如何识别和避免这些缺陷。

第 7 章主要介绍如何以及为什么要现代化数据仓库架构。这将为你提供概念上的思想以及实现上的一些选项。

第 8 章重点介绍工作负载，及其在数据仓库和大数据领域中的真正含义，理解工作负载的重要性，以及基于工作负载如何创建数据仓库的架构。对于任何数据管理解决方案来说，这都是其未来架构最重要和最关键的方面之一。

第 9 章重点讨论那些已持续应用到企业中的新兴技术，特别是在处理数据库仓库的性能和可扩展性方面。该章还讨论数据仓库一体机、云计算、数据虚拟化和内存计算。

## 第三部分：构建大数据 – 数据仓库

第 10 章重点介绍将数据仓库与大数据集成的方法和相关的技术，这些技术的采用基于公司的数据类型、当前演化状态和现有技术。

第 11 章重点讨论在大数据领域中通过部署有效的 MDM 和元数据策略来创建数据驱动的架构。它强调对数据管理的这两大支柱的需求，特别是在大数据领域。该章还讨论语义层和基于语义网的方法。

第 12 章的重点是管理大数据的生命周期，包括哪些数据是基本的，在处理前和处理后如何以及在哪里保存数据。还将讨论企业大数据中如果不实现一个鲁棒的 ILM 策略会带来哪些问题。

第 13 章涉及使用大数据的最终目标，也就是提供强大的可视化，分析大数据，最重要的是，新兴的数据科学家的角色。这里的目标是为你提供关于这些主题的概念性的想法以及它们如何影响整体的大数据策略。

第 14 章着重介绍在财富 500 强企业的下一代数据仓库的实际实施中的最终架构。目的是当你的企业演化到新的数据领域后，为你提供一些面向未来的想法。

## 附录

附录 A 展示具体的客户案例研究。

附录 B 给出建设医疗保健信息工厂的案例研究。

# 致　谢

本书的出版离不开太多人的支持，我要感谢他们在本书的出版过程中给予的支持和帮助。

首先，我要感谢我的妻子和两个儿子一直以来的支持，他们牺牲了很多周末、假期、看电影、参加学校活动和社会活动的时间（这些我都缺席了），给我提供了考虑如何组织本书内容的黄金时间。没有他们的帮助和支持，我将永远没有机会写作这本书。

接下来，感谢我亲爱的朋友、导师兼商业伙伴 Bill Inmon，他的支持和鼓励让我坚持在做一份全职咨询工作的同时写这本书——工作在一家初创公司的同时，又做着无数的事情。谢谢 Bill 总是及时提供帮助。

我要特别感谢三个最好的朋友（我很幸运，有这些家伙作为我的朋友）：Todd Nash、Hans Hultgren 和 Shankar Radhakrishnan。他们花了数小时的时间（晚上、周末和在飞机上）审查草稿，并对每一步都提供反馈。没有他们不懈的努力，我们将不能及时完成这本书。他们令我倍感惊喜，他们是这段令人难以置信的旅程中无法用语言描述的一部分。他们是令人难以置信的，很高兴能有这样的朋友，与他们在一起的时光总是很特别。

仅仅有一堆理论是没法写成一本书的——本书获得了很多业内资深人士以供应商支持的方式提供的帮助。IBM 的 Glenn Zimmerman，他对我的所有案例研究都非常支持，甚至帮我把它们改写成 Word 格式。Teradata 的 Kim Dossey，谢谢他提供的所有支持并快速提供了许多案例研究。Cloudera 的 Alan 和 Kate，他们两个毫不犹豫地让我选择尽可能多的案例进行研究。Klout 的案例研究要感谢微软的 SQL Server 团队。我想郑重地感谢所有这些供应商的鼎力支持和热心帮助。

我还要感谢在这个征途中以鼓励的言语来支持我的一些朋友：Paul Kautza, Philip Russom, Dave Stodder, Dave Wells, Claudia Imhoff, Jill Dyche, Mark Madsen, Jonathan Seidman, Kylie Clement, Dave Nielsen, Shawn Rogers, John Myers, John O'Brien, William McKinght, Robert Eve, Tony Shaw，以及 John Onder。

最后一点也很重要，感谢 Andrea Dierna、Heather Scherer 编辑和整个 Morgan Kauffman 团队在这一过程中的所有帮助、指导和支持——如果没有他们，这一切都不可能实现。

谢谢各位！

# 作者简介

Krish Krishnan 是高性能数据仓库解决方案和非结构化数据的战略、架构和实现等方面全球公认的专家。他是一位很受欢迎且有远见的数据仓库思想领导者和实践者，是世界顶尖的战略和架构咨询师之一。Krish 也是全世界大数据相关会议上独立的分析人员和演讲者，且在数据仓储研究所（TDWI）讲授这一主题。Krish 和其他的专家正在帮助推动下一代数据仓储的行业成熟度，侧重于大数据、语义技术、众包、分析和平台工程。

Krish 是 Sixth Sense Advisors 公司的创始人兼 CEO，提供行业分析服务，覆盖数据仓库、分析、云计算、社交媒体和商务智能。Krish 还与其合伙人的组织一起在全球各地提供战略和创新咨询服务。

# 目 录

## 第二部分 数据仓库

第一部分

# 大 数 据

# 第 1 章
# 大数据简介

## 1.1 引言

自“因特网”产生以来，吸引现代计算行业注意力的最大的现象是“大数据”。这两个词首次同时出现在麦肯锡公司同主题的论文中，而其基本定义则来自 Gartner 公司的 Doug Laney。

今天的“大数据”受欢迎的根本原因是因为技术平台的同时出现，这些平台有能力处理多种格式和结构的数据，而且不用担心传统的系统和数据库平台的限制。

## 1.2 大数据

数据是表示信息或知识最原始的格式。在计算领域，我们通常认为数据是有组织的值的行和列，这些值代表一个或者多个实体及其属性。在使用电脑帮助处理计算或信息管理的时代之前的很长一段时间，数据是由古希腊人随着计数和贸易的到来而发明的。简单来说，就是把值分配给数字，然后使用这些数字来标记货币价值、人口、日历、税收以及许多和历史有关的实例。这些实例提供了大量证据来证明人类思维在数据和知识的获取与管理方面的魅力。

根据卡内基·梅隆大学的一系列研究，信息或数据管理的必要过程是组织、采集、存储、检索和管理数据。从不同的过程采集的数据使得我们在处理输出结果时，能够理解和满足需求，从而做出决策。这种管理行为是赫伯特·西蒙有限的理性观点[⊖]或当人类思维应用到数据管理时有限的视野的基本主题。对于决策行为和管理行为的争论的提出是很有意义的，因为我们限制了数据建模过程和应用算法的应用中的数据，而且一直以来都致力数据内的离散关系而不是整体。

然而，在现实中，决策总是超越用于辅助处理的传统系统。例如，患者的治疗和管理不局限于计算机和程序。但是现在在医院里，对于每一个患者，均可通过非结构化的数据集成技术和算法，以电子方式收集和处理由医生、护士、实验室技术人员、急救人员和医

⊖ March, J. G., & Simon, H. A. (1958) Organizations (http://www.amazon.com/Organizations-James-G-March/dp/063118631X).