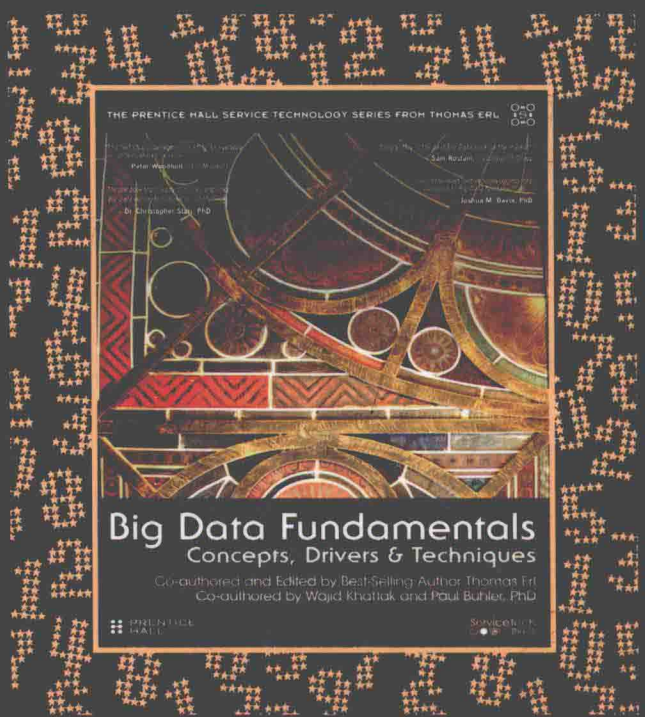
HZ BOOKS
华章IT
Pearson
数据科学与工程技术丛书

大数据导论

托马斯·埃尔（Thomas Erl）
[美] 瓦吉德·哈塔克（Wajid Khattak） 著
保罗·布勒（Paul Buhler）

彭智勇 杨先娣 译



BIG DATA FUNDAMENTALS

CONCEPTS, DRIVERS AND TECHNIQUES

机械工业出版社
China Machine Press

数据科学与工程技术丛书

BIG DATA FUNDAMENTALS
CONCEPTS, DRIVERS AND TECHNIQUES

大数据导论

托马斯·埃尔（Thomas Erl）
[美] 瓦吉德·哈塔克（Wajid Khattak） 著
保罗·布勒（Paul Buhler）

彭智勇 杨先娣 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

大数据导论 /（美）托马斯·埃尔（Thomas Erl）等著；彭智勇，杨先娣译．—北京：机械工业出版社，2017.5
（数据科学与工程技术丛书）
书名原文：Big Data Fundamentals: Concepts, Drivers and Techniques

ISBN 978-7-111-56577-2

I. 大… II. ① 托… ② 彭… ③ 杨… III. 数据处理 IV. TP274

中国版本图书馆 CIP 数据核字（2017）第 076608 号

本书版权登记号：图字：01-2016-3781

本书是面向商业和技术专业人员的大数据权威指南，清楚地介绍了大数据相关的概念、理论、术语与基础技术，并使用真实连贯的商业案例以及简单的图表，帮助读者更清晰地理解大数据技术。

本书可作为高等院校相关专业“大数据基础”、“大数据导论”等课程的教材，也可供从事大数据相关工作的技术人员、管理人员和所有对大数据感兴趣的人士阅读。

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：唐晓琳　　责任校对：殷　虹
印　　刷：北京文昌阁彩色印刷有限责任公司　　版　　次：2017 年 5 月第 1 版第 1 次印刷
开　　本：185mm × 260mm　1/16　　印　　张：11.75
书　　号：ISBN 978-7-111-56577-2　　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

译 者 序

现今，“大数据”已经成为全球科技界和企业界关注的热点。数据为王的时代已经到来，各行各业高度关注大数据的研究和应用。企业关注的重点从追求计算机的计算速度转变为追求大数据处理能力，从以软件编程为主转变为以数据为中心。在云计算技术和海量数据存储技术的助力下，大数据已经成为当前学术界、工业界的热点和焦点。大数据的出现将会对社会各个领域产生深刻影响。从公司战略到产业生态，从学术研究到生产实践，从城镇管理到国家治理，都将发生本质的变化，大数据将成为时代变革的力量。“用数据来说话、用数据来管理、用数据来决策、用数据来创新”的文化氛围与时代特征愈发鲜明。大数据时代需要一大批具备大数据知识的专业人才，他们应能有效地将数据科学和各行各业的应用相结合，推动新技术和新应用的发展。因此，掌握大数据核心技术且拥有专业领域知识的人才储备成为国家大数据战略布局的重中之重。

在本书中，IT 畅销书作者 Thomas Erl 和他的团队清楚地解释了关键的大数据概念、理论和术语，以及基本的大数据技术和方法。本书分两部分：第一部分主要从商业相关问题的讨论引出大数据的驱动力，解释了如何通过大数据推动企业的发展，介绍了大数据的应用背景和基本概念；第二部分主要是大数据技术相关问题的讨论，重点介绍了大数据的存储技术和分析方法。本书的特色在于每一章后都有案例学习，用一家大型的保险公司 ETI 对大数据的应用案例贯穿始终，为相关章节的知识应用提供了现实场景，以加深读者对大数据实际应用的认识。另外，本书大量应用了简单的图表说明。这些都使得本书非常实用且通俗易懂，因此，本书特别适合作为了解大数据基本知识和相关技术的入门教材，也可以作为高校的通识课教材来使用。

在本书翻译过程中，武汉大学计算机学院的刘歆文、李卓、史成良、陈洪洋、贺潇雅、万言历、陈昊等同学做了大量辅助性工作，在此，向这些同学的辛勤工作表示衷心的感谢。

由于译者能力有限，译稿难免存在疏漏及不足之处，望广大读者不吝赐教。

互联网（IoE）。目前他的研究兴趣是通过权衡响应式设计原则与基于目标的执行方式，减少业务策略与流程执行之间的差距。

作为 Modus21 的首席科学家，Paul Buhler 博士根据当前业务架构与流程执行框架的发展趋势调整企业的战略布局。目前，他还是查尔斯顿学院的合作教授，负责本科生与硕士生计算机科学课程的教学工作。Paul Buhler 博士在南卡罗来纳大学获得计算机工程博士学位，在约翰霍普金斯大学获得计算机科学硕士学位，在塞特多大学获得计算机科学学士学位。

目　录

第一部分

大数据基础

大数据具有改变企业性质的能力。事实上，有很多公司仅仅依靠着能够提出一些深刻的见解而存在，而这些见解只有通过大数据才能实现。第一部分的四章主要从商业的角度阐述了大数据的基本要素。企业需要理解大数据，不仅仅与技术相关，也与如何通过这些技术推动公司的发展相关。

第一部分由如下 4 章组成：

- 第 1 章主要介绍一些关键性的概念和术语，定义了大数据技术中的许多基本元素，并且阐述了大数据处理复杂的商业中蕴含的深层知识的能力。同样，第 1 章也介绍了辨别大数据的数据集的许多特征，并且定义了很多能够作为大数据分析技术的主体的数据类型。
- 第 2 章旨在解答以下问题：作为一种市场经济和商业世界潜在变化的结果，企业为什么应该使用大数据技术？大数据本身跟企业转型没有关联，但是，当一个企业能依靠自身的洞察力的时候，大数据技术能激发企业内部的革新。
- 第 3 章阐释了大数据技术不仅仅是简单的“普通商业活动”，在选择使用大数据技术时，必须要有许多商业性和技术性的考虑。这一点对企业提出了要求：企业能在大数据技术的帮助下接触到外部数据的影响，但同时意味着企业需要控制、管理这些数据。此外，大数据分析的生命周期还提出了不同的数据分析操作的要求。

- 第4章检验了目前能接触到的企业级数据仓库和大数据商务智能的方法。然后扩展了这个思路，表明大数据存储技术以及分析数据资源可以与企业绩效监控工具相结合，以此来强化企业的分析能力，同时深化由商务智能提供的深层知识。

商业的内部数据往往没有办法体现出商业的全部价值——在这个前提下，正确地利用大数据将成为战略性主动权的一部分。换句话说，大数据不仅仅是关于可以被现有技术解决的数据管理的问题，更是关于一些商务的问题，而为它们提供解决方案的技术需要支持大数据的数据集。因此，第一部分的商业相关问题的讨论将为第二部分技术相关问题的讨论奠定基础。

1～2

第 1 章

理解大数据

大数据是一门专注于对大量的、频繁产生于不同信息源的数据进行存储、处理和分析的学科。当传统的数据分析、处理和存储技术手段无法满足当前需求的时候，大数据的实践解决方案就显得尤为重要。具体地说，大数据能满足许多不同的需求，例如，将多个没有联系的数据集结合在一起，或是处理大量非结构化的数据，抑或是从时间敏感的行为中获取隐藏的信息等。

虽然大数据看起来像是一门新兴的学科，却已有多年的发展历史。对大型数据集的管理与分析是一个存在已久的问题——从利用劳动密集方法进行早期人口普查的工作，到计算保险收费背后的精算学科，都涉及这个方面的问题，大数据就由此发展起来。

作为对传统的基于统计学分析方法的优化，大数据加入了更加新的技术，利用计算资源和方法的优势来执行分析算法。在当今数据集持续地扩大化、扩宽化、复杂化和数据流化的背景之下，这种优化十分重要。自《圣经》时代以来，统计学方法一直在告诉我们通过抽样调查的手段能够粗略地测量人口。但计算机科学目前的发展使我们完全有能力处理那样庞大的数据集，因此抽样调查的手法正在逐渐“失宠”。

对于大数据的数据集的分析是一项综合数学、统计学、计算机科学等多项专业学科的跨学科工作。这种多学科、多观点的混合，常常会使人对大数据及大数据分析这门学科所涵盖的内容产生疑问，每个人都会有不同的见解。大数据问题所涵盖的内容范围也会随着软硬件技术的更新而变化。这是因为我们在定义大数据的时候考虑了数据特征对于数据解决方案本身的影响。比如 30 年前，1GB 的数据就称

得上是大数据，而且我们还会为这份数据专门申请计算资源，而如今，1GB 的数据十分常见，面向消费者的设备就能对其进行快速的存储、转移、复制或者其他处理。

3~4

大数据时代下的企业数据，常常通过各种应用、传感器以及外部资源聚集到企业的数据集中。这些数据经过大数据解决方案的处理后，能够直接应用于企业，或者添加到数据仓库中丰富现有的数据。这种大数据解决方案处理的结果，将会给我们带来许多深层知识和益处，例如：

- 运营优化
- 可实践的知识
- 新市场的发现
- 精确的预测
- 故障和欺诈的检测
- 详细的信息记录
- 优化的决策
- 科学的新发现

显然，大数据的应用面和潜在优势十分广阔。然而，在何时选用大数据分析手段的问题上，还有大量的问题需要考虑。当然，我们需要去理解这些存在的问题，并与大数据的优势进行权衡，最终才能做出一个合理的决策并提出合适的解决方案。这些内容我们将在第二部分单独讨论。

1.1 概念与术语

作为开端，我们首先要定义几个基本概念和术语，以便大家理解。

1.1.1 数据集

我们把一组或者一个集合的相关联的数据称作数据集。数据集中的每一个成员数据，都应与数据集中的其他成员拥有相同的特征或者属性。以下是一些数据集的例子：

- 存储在一个文本文件中的推文（tweet）
- 一个文件夹中的图像文件
- 存储在一个 CSV 格式文件中的从数据库中提取出来的行数据

❑ 存储在一个 XML 文件中的历史气象观测数据

图 1.1 中显示了三种不同数据格式的数据集。

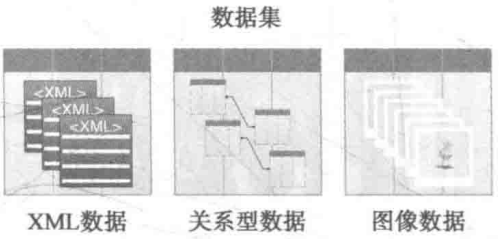


图 1.1　数据集可以有多种不同的格式

1.1.2　数据分析

数据分析是一个通过处理数据，从数据中发现一些深层知识、模式、关系或是趋势的过程。数据分析的总体目标是做出更好的决策。举个简单的例子，通过分析冰淇淋的销售额数据，发现一天中冰淇淋甜筒的销量与当天气温的关系。这个分析结果可以帮助商店根据天气预报来决定每天应该订购多少冰淇淋。通过数据分析，我们可以对分析过的数据建立起关系与模式。图 1.2 显示了代表数据分析的符号。

图 1.2　用于表示数据分析的符号

1.1.3　数据分析学

数据分析学是一个包含数据分析，且比数据分析更为宽泛的概念。数据分析学这门学科涵盖了对整个数据生命周期的管理，而数据生命周期包含了数据收集、数据清理、数据组织、数据分析、数据存储以及数据管理等过程。此外，数据分析学还涵盖了分析方法、科学技术、自动化分析工具等。在大数据环境下，数据分析学发展了数据分析在高度可扩展的、大量分布式技术和框架中的应用，使之有能力处

6 理大量的来自不同信息源的数据。图 1.3 显示了代表数据分析学的符号。

图 1.3 用于表示数据分析学的符号

大数据分析（学）的生命周期通常会对大量非结构化且未经处理过的数据进行识别、获取、准备和分析等操作，从这些数据中提取出能够作为模式识别的输入，或者加入现有的企业数据库的有效信息。

不同的行业会以不同的方式使用大数据分析工具和技术。以下述三者为例：

- 在商业组织中，利用大数据的分析结果能降低运营开销，还有助于优化决策。
- 在科研领域，大数据分析能够确认一个现象的起因，并且能基于此提出更为精确的预测。
- 在服务业领域，比如公众行业，大数据分析有助于人们以更低的开销提供更好的服务。

大数据分析使得决策有了科学基础，现在做决策可以基于实际的数据而不仅仅依赖于过去的经验或者直觉。根据分析结果的不同，我们大致可以将分析归为以下 4 类：

- 描述性分析
- 诊断性分析
- 预测性分析
- 规范性分析

不同的分析类型将需要不同的技术和分析算法。这意味着在传递多种类型的分

析结果的时候，可能会有大量不同的数据、存储、处理要求。如图 1.4 所示，生成高质量的分析结果将加大分析环境的复杂性和开销。 7

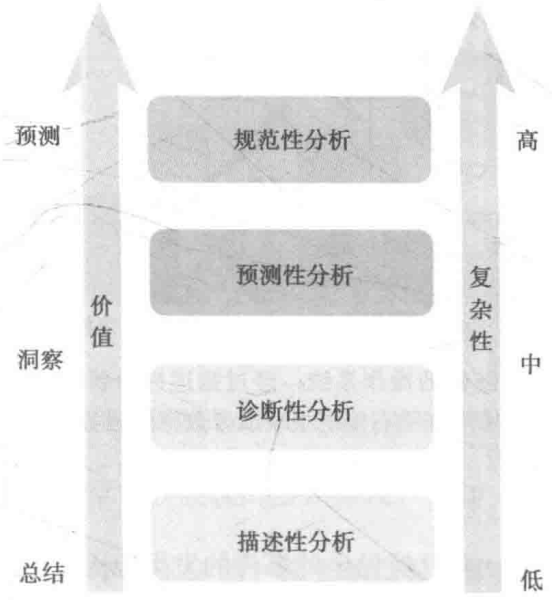


图 1.4　从描述性分析到规范性分析，价值和复杂性都在不断提升

1. 描述性分析

描述性分析往往是对已经发生的事件进行问答和总结。这种形式的分析需要将数据置于生成信息的上下文中考虑。

相关问题可能包括：

- 过去 12 个月的销售量如何？
- 根据事件严重程度和地理位置分类，收到的求助电话的数量如何？
- 每一位销售经理的月销售额是多少？

据估计，生成的分析结果 80% 都是自然可描述的。描述性分析提供了较低的价值，但也只需要相对基础的训练集。

如图 1.5 所示，进行描述性分析常常借助即席报表和仪表板（dashboard）。报表常常是静态的，并且是以数据表格或图表形式呈现的历史数据。查询处理往往基于企业内部存储的可操作数据，例如客户关系管理系统（CRM）或者企业资源规划系统（ERP）。 8

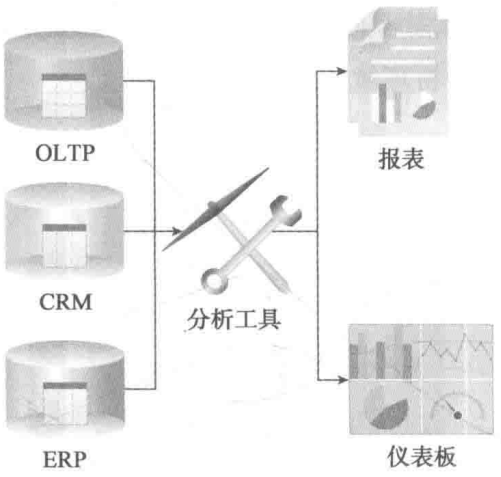


图 1.5 图左侧的操作系统，经过描述性分析工具的处理，能够生成图右侧的报表或者数据仪表板

2. 诊断性分析

诊断性分析旨在寻求一个已经发生的事件的发生原因。这类分析的目标是通过获取一些与事件相关的信息来回答有关的问题，最后得出事件发生的原因。

相关的问题可能包括：

- 为什么 Q2 商品比 Q1 卖得多？
- 为什么来自东部地区的求助电话比来自西部地区的要多？
- 为什么最近三个月内病人再入院的比率有所提升？

诊断性分析比描述性分析提供了更加有价值的信息，但同时也要求更加高级的训练集。如图 1.6 所示，诊断性分析常常需要从不同的信息源搜集数据，并将它们以一种易于进行下钻和上卷分析的结构加以保存。而诊断性分析的结果可以由交互式可视化界面显示，让用户能够清晰地了解模式与趋势。诊断性分析是基于分析处理系统中的多维数据进行的，而且，与描述性分析相比，它的查询处理更加复杂。

3. 预测性分析

预测性分析常在需要预测一个事件的结果时使用。通过预测性分析，信息将得到增值，这种增值主要表现在信息之间是如何相关的。这种相关性的强度和重要性构成了基于过去事件对未来进行预测的模型的基础。这些用于预测性分析的模型与