TURING 图灵程序设计丛书

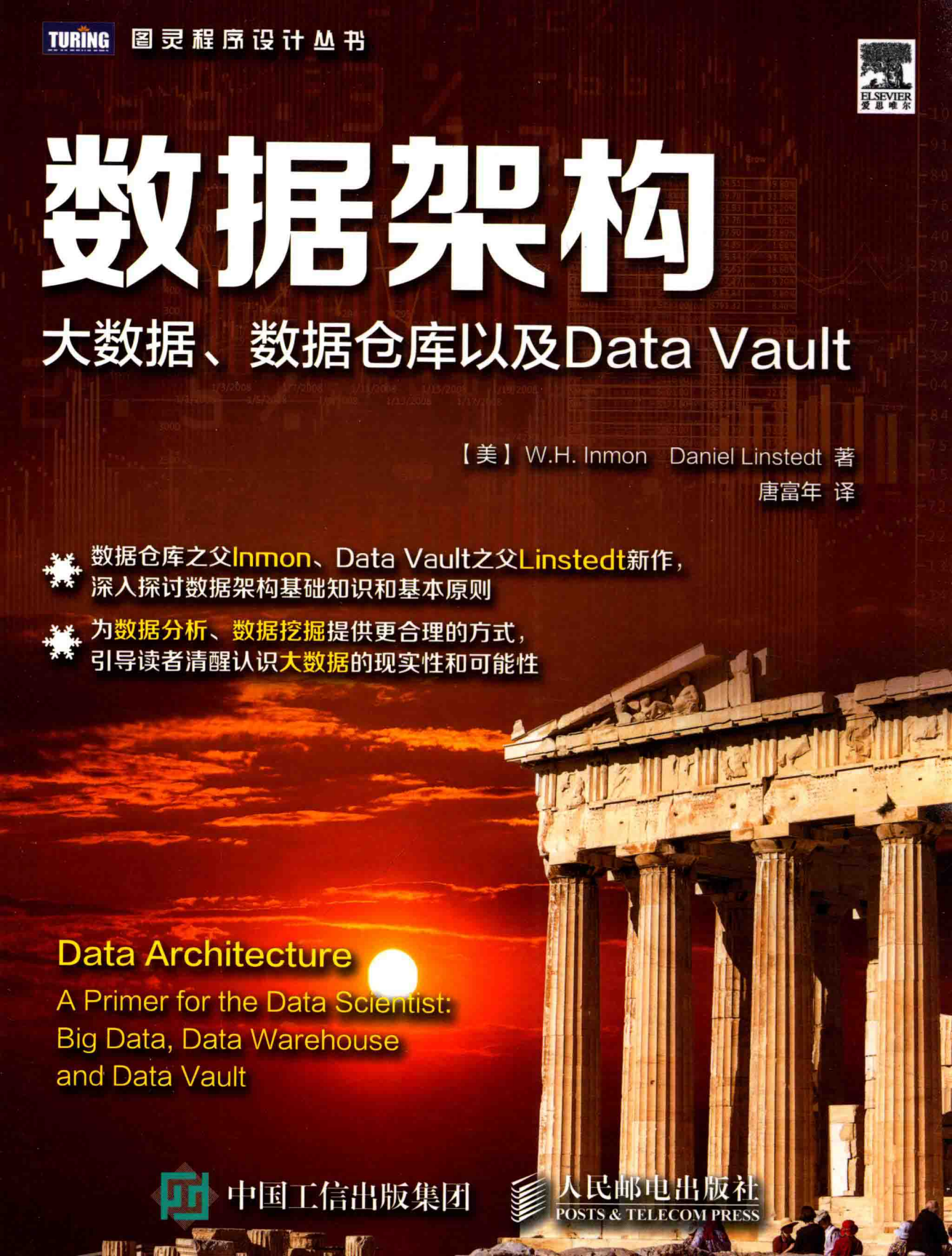
ELSEVIER 爱思唯尔

数据架构

大数据、数据仓库以及Data Vault

【美】W.H. Inmon Daniel Linstedt 著
唐富年 译

- 数据仓库之父Inmon、Data Vault之父Linstedt新作，深入探讨数据架构基础知识和基本原则
- 为数据分析、数据挖掘提供更合理的方式，引导读者清醒认识大数据的现实性和可能性



Data Architecture
A Primer for the Data Scientist:
Big Data, Data Warehouse
and Data Vault

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

TURING 图灵程序设计丛书

数据架构

大数据、数据仓库以及Data Vault

【美】W.H. Inmon　Daniel Linstedt 著
唐富年 译

Data Architecture

A Primer for the Data Scientist:
Big Data, Data Warehouse
and Data Vault

人民邮电出版社
北　京

图书在版编目（CIP）数据

数据架构 : 大数据、数据仓库以及Data Vault / (美) 威廉.H·英蒙 (W.H. Inmon) , (美) 丹尼尔·林斯泰特 (Daniel Linstedt) 著 ; 唐富年译. -- 北京 : 人民邮电出版社, 2017.1
（图灵程序设计丛书）
ISBN 978-7-115-43843-0

Ⅰ. ①数… Ⅱ. ①威… ②丹… ③唐… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2016)第254063号

内 容 提 要

本书是数据仓库之父 Inmon 的新作，探讨数据的架构和如何在现有系统中最有效地利用数据。本书的主题涵盖企业数据、大数据、数据仓库、Data Vault、业务系统和架构。主要内容包括：在分析和大数据之间建立关联，如何利用现有信息系统，如何导出重复型数据和非重复型数据，大数据以及使用大数据的商业价值，等等。

本书的读者对象包括大数据架构师、数据科学家以及从事数据分析和研究的科研人员。

◆ 著 [美] W.H. Inmon Daniel Linstedt
译 唐富年
责任编辑 朱 巍
执行编辑 杨 琳 赵瑞琳
责任印制 彭志环

◆ 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
北京昌平百善印刷厂印刷

◆ 开本：800×1000 1/16
印张：18.25
字数：442千字 2017年1月第1版
印数：1-4 000册 2017年1月北京第1次印刷

著作权合同登记号 图字：01-2015-2831号

定价：69.00元

读者服务热线：(010)51095186转600 印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广字第 8052 号

译 者 序

学习大师的著作通常令人满怀景仰，而翻译大师的著作又往往让人惴惴不安。Inmon 的著作总是将复杂的技术讲解得通俗易懂，体现出清晰的知识脉络，阐述观点的视角也非常独到。“授人以鱼，不如授之以渔。”这本书讲述的是原理、架构和方法论，颇有授人以“捕鱼之术”的味道。本书有三个比较重要的关键词：数据架构、大数据和 Data Vault。对于工程技术人员、管理人员（包括行政管理人员和信息管理人员）以及从事各种数据分析和研究的科研人员而言，本书绝对是一本不可错过的好书。

从本质上讲，数据架构与建筑架构并无二致。没有良好定义的架构，就难以支撑起数据的捕获、计算、分析和管理运维等各个环节，更不用说管理和使用海量数据了。为什么我们的数据总是难以集成和交换？为什么我们的信息系统总是不够可靠，生命周期是那么短暂？为什么我们难以从数据中分析挖掘出业务价值？关键就在于我们在数据架构设计上投入的精力太少，总是草草地完成（甚至是跳过）设计阶段的工作，急匆匆地进入实施阶段，而忽略了数据的本质特性。

“大数据”的概念出现之后，在一种急功近利的狂躁情绪的牵引下，在商业包装和媒体炒作的推动下，在信息化的很多角落里，很多人正在试图将原来的“小垃圾桶”换成新的“大垃圾桶”；但是，真正从大数据技术中获益的人要远少于宣传大数据的人，而且“大数据”这个词实际上正面临着滥用的危险。在我们的各种数据标准尚不够完善之时，在我们的数据架构仍然存在短板之时，我们的大数据走不了多远。在静下心来读完这本书之后，相信你对此会有更为深刻的体会，不会再被各种有关大数据的华丽辞藻和神话传说迷住双眼。

Data Vault 是本书的核心内容，蕴含着 Inmon 等人对数据仓库这门技术在大数据环境下如何发展和走向成熟的思考。在本书翻译之初，我曾经信誓旦旦地对编辑说，要在翻译工作结束之后为 Data Vault 这个英文词组找一个对应的汉语词组。遗憾的是，在全部翻译工作完成之后仍然未能如愿。我曾经试图将 Data Vault 翻译成“数据仓”“数据宝库”“数据仓储”“数据库所”等，但是又觉得这其中的每一个都有不妥。Data Vault 的内涵比数据仓库丰富得多，也更加雄心勃勃。就我的理解来说，如果将企业视为一个封闭世界，那么 Data Vault 所面向的就是这个世界穹顶之下的所有数据。为了避免混淆和误导，在找到一个足够准确的词组之前，我觉得还是不作翻译为好。

虽然已经竭尽绵力，但是译文仍难免有错误和疏漏之处，还望读者海涵。感谢图灵公司的各位编辑为本书付出的心血与汗水。

唐富年

2016 年 3 月于济南

前　言

不久前有一段卡通视频非常流行，它从不同的视角展示了一架飞机。从防御装备的视角来看，整架飞机都采用了重型装甲。从武器装备的视角来看，飞机到处都配有火炮和火箭弹。从轰炸的视角来看，飞机携带了各种各样的炸弹。从飞行员的视角来看，该飞机造型优美且机动性良好。从工程师的视角来看，飞机上配置了各种各样的部件、按钮和小装置。

上述各个视角之间存在的问题在于，它们完全不同而且彼此不相称。到了最后，飞机其实是各个视角相互妥协的产物。在最终的实际产品中，每一个视角的优化都不能以牺牲其他视角为代价。

数据的情况与之非常类似：不同的人群对于数据有着不同的看法。有些群体需要处理海量数据；有些群体希望能够以近乎瞬时的速度在线访问详细数据；有些群体希望拥有严格控制完整性的数据；而有些群体则只关心自己的“个人”数据，希望能够使用计算机轻松快捷地创建和处理自己的数据版本。

每个群体都有自己的视角，都在自己的世界里有合乎情理的观点。不过数据无法同时满足所有的视角和所有需要。

数据很复杂，本身涉及很多方面，也有很多种用途。

本书旨在围绕数据展开研究，探索较为宽泛的数据架构问题。本书试图展现组织或企业中所有的数据用途和视角。此外，本书试图以一种合理、公平的方式来平衡所有对数据的需求和看待数据的视角。

本书首先介绍了企业中看待数据的最主流视角。为此，首先要明白企业数据存在广泛的多样性。要想有效地使用数据，组织就必须根据不同的情况来处理数据。

有些书是讲“如何做”的书，例如手册；有些书是讲故事的书，例如小说和非小说文学；还有些书是纯粹逃避现实的娱乐性书籍。与它们不同，本书是一本描述性的书，是一本讲“是什么”的书，是一本关于大而复杂的架构的书。形形色色的数据就像马赛克一样，而各个组织的数据都是不同的。本书首先从一个比较高的架构层次讲述数据，然后深入到清晰、易于理解的细节，确保你明白本书所要讲述的内容。

现在，关于数据有很多令人混淆的说法（只要有电脑就会存在这样的情况），而其中大部分是由技术供应商引起的。技术供应商并不会提出荒唐和毫无依据的说法，但是他们很容易渲染和夸大自己的案例。最糟糕的是，技术供应商还可能会有“近视”的毛病，并深受其害。在对数据的认识方面，技术供应商很容易管中窥豹。他们很可能向人们呈现这样一种对世界的看法：自己

的技术在现在或者未来是唯一的；而这并不是现实。这种由技术供应商引起的严重“近视”会造成很大的混乱。

有关大数据的说法很容易让人们在理解大数据的现实性和可能性时迷失方向。本书着眼于大数据是如何适用于决策领域的。本书从如下几个重要的视角进行思考：当前企业是如何进行决策的，企业应该如何进行决策，以及在大数据条件下如何进行决策。

本书主要涵盖了以下几个主题。

- **企业数据**

  企业数据是指整个企业的信息全景。在企业中有很多种不同类型的数据。本书展示了一种数据视角，并且在很高的层次上阐述了如何在企业决策过程中使用（或者不使用）数据。

- **大数据**

  讲述了大数据是什么，以及它能够如何增强企业的决策。大数据有几种不同的定义。本书采用了一种非常务实的大数据观点，然后讨论了它的一些突出特点。大数据最明显却并未被技术供应商所提起的一个特征是重复型大数据和非重复型大数据之间的差异性。重复型大数据和非重复型大数据之间深刻的差别也称作“分界线”。本书之所以值得购买，正是因为通过阅读本书你可以很容易地理解这条“分界线”，而且本书对企业决策能力也有所启示。

- **数据仓库**

  数据仓库面向企业数据完整性方面的需求。总有一天，企业会开始领悟到这样的事实：拥有数据和拥有可信的数据并不是一回事。他们醒悟之后意识到了“数据完整性”的意义。这个时候，企业级数据仓库（enterprise data warehouse，EDW）诞生了。有了 EDW，企业可以利用其中的基础数据制定重要、可信的决策。在 EDW 出现之前，企业已经有了大量的数据，但这些并不是可信的数据。

- **Data Vault**

  Data Vault 面向管理随时间推移而发生数据变更的需求。数据仓库会随着时间推移而不断演化，这最终形成了一种名为 Data Vault 的学科和结构。不论过去还是现在，都有多种原因采用 Data Vault 作为具有完整性需求的系统的主干。

- **业务系统**

  业务系统面向企业日常业务运作方面的需求。由于管理超大规模数据量和数据完整性方面的需求，需要一些能够运行和增强组织日常业务的系统（今后也一直需要）。

- **架构**

  架构是指如何以一种整体而内聚的方式将不同类型的数据和不同类型的数据需求组织到一起。认识企业中各种数据视角的不同需求是一回事，而设想如何以一种整体而内聚的方式将不同类型的数据组织到一起则是另外一回事。

通过阅读本书，你会了解如何将企业中所有形式的数据组合到一起。本书旨在提供一个关于企业全部数据的高层次、全方位的视图，并且介绍如何使不同的数据形式以建设性的方式相互协作。

本书面向管理人员、架构师、业务人员和技术人员。所有参与企业决策的人都会从本书中受益。对本书特别感兴趣的人群是数据科学家。对于一名数据科学家来说，本书就像一本地图册，标绘出了世界上不同的大洲和海洋。数据科学家再也不需要去摸索着认识一个被认为是“平的”的世界，也不需要通过反复的艰苦探索来完成对岛屿和大陆形状的认知。

很多年前，当我还是耶鲁大学一年级的学生时，Ernest Lockridge 博士是我的英语老师。他讲授的是英语作文课，也是我唯一上过的英语作文课。那时候我和 Ernest Lockridge 博士都不知道这今后会对我有什么样的影响。后来我撰写了 53 本书，我由衷感谢他对我的指导和启示。如果我没有记错（毕竟过了这么多年），Ernest Lockridge 博士是第一位称呼我为“Inmon 先生”的人。这一直印在我的脑海里，直到今天，久久不能忘怀。

我终生感谢 Ernest Lockridge 博士。

WHI/DL

2014 年 3 月 25 日

目　录

第1章 企业数据

1.1 企业数据

如今，人们在处理数据的时候很容易迷失。数据有很多种不同的类型，而且每个类型的数据都有其自身的风格和特质。产品、供应商和应用程序都变得过于专注自己所处的特定世界，忽视了用更加宽广的视野来考察如何将各种事物组合成一个整体。因此，后退一步用更宽广的视野看待数据，经常能够获得更为恰当的观点。

1.1.1 企业的全体数据

试想一下企业里所能找到的所有数据。图1.1.1简要描述了企业中全体数据的情况。

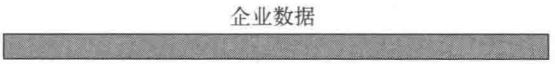


图 1.1.1

这里的全体数据包括与企业中各类型数据相关的所有事项。

进一步细分企业中的全体数据有很多种方式。一种细分方式（但是肯定不是唯一方式）是将全体数据划分为结构化数据和非结构化数据，如图1.1.2所示。

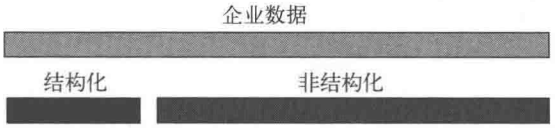


图 1.1.2

结构化数据是一种可预见、经常出现的数据格式。通常，结构化数据包括记录、属性、键和索引等，可以通过数据库管理系统（database management system，DBMS）进行管理。结构化数据是定义良好的、可预测的，并且可通过复杂的基础设施进行管理。通常，结构化环境中的大多数数据单元都可以很快地进行定位。

相反，非结构化数据是不可预见的，而且没有可以被计算机识别的结构。访问非结构化数据通常很不方便，想要查找给定的数据单元，就必须顺序搜索（解析）长串的数据。非结构化数据有很多种形式和变体。最常见的非结构化数据的表现形式也许就是文本了。然而无论如何，文本都不是非结构化数据的唯一形式。

1.1.2　非结构化数据的划分

非结构化数据可以进一步划分成两种基本的数据形式：重复型非结构化数据和非重复型非结构化数据。与企业数据的划分一样，非结构化数据的细分方式也有很多种。这里给出的只是其中一种细分非结构化数据的方法。图1.1.3展现了非结构化数据的这一细分方法。

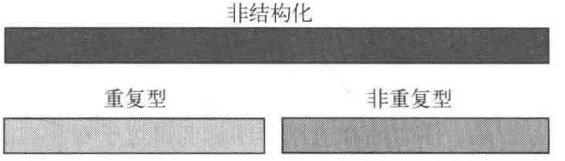


图　1.1.3

重复型非结构化数据是指以同样的结构甚至同样的形态出现多次的数据。通常，重复型数据会出现很多很多次。重复型数据的结构与之前的记录看起来完全一样或者大致相同。没有用于管理重复型非结构化数据内容的大型复杂基础设施。

非重复型非结构化数据是指记录截然不同的数据。通常，每个非重复型的记录都与其他记录明显不同。

企业数据类型的划分有多种不同的体现。参见图1.1.4中所示的数据。

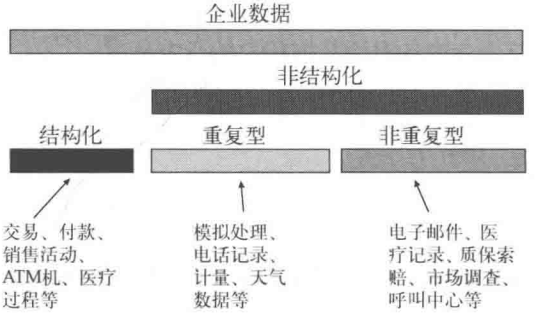


图　1.1.4

结构化数据通常是交易的副产品。每当一次销售完成时，每当银行账户有取款操作时，每当有人在ATM机上办理业务时，每当发送一份账单时，都会产生一条交易记录。交易记录最终会形成一条条结构化的记录。

重复型非结构化数据则有所不同。非结构化的重复记录通常是机器间交互所产生的记录，例如对即将离开生产过程的产品进行模拟验证，或者对消费者的能源用量进行计量等。就拿计量来说，在读取计量读数时，会产生大量在形式和内容上重复的记录。

非重复型非结构化信息与重复型非结构化记录有着根本性的不同。对于非重复型非结构化记录而言，它们无论在形式还是内容上都很少重复或者根本不重复。非重复型非结构化信息的例子有电子邮件、呼叫中心对话和市场调查等。当你查看一封电子邮件时，会有很大概率发现数据库中的下一封邮件与前一封邮件是极为不同的。对呼叫中心信息、质保索赔、市场调查等数据来说也是如此。

1.1.3 业务相关性

重复型非结构化数据和非重复型非结构化数据在很多方面都有着极为不同的特征，其中一方面就是业务相关性。在重复型非结构化数据中，通常只有很少的记录具备真正的业务价值。然而，非重复型非结构化数据则有很大比例与业务相关。

这两种数据在业务相关性方面的不同如图1.1.5所示。

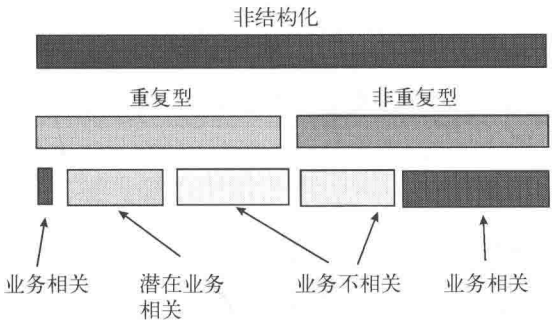


图 1.1.5

非结构化重复型数据中只有很小比例是与业务相关的。例如，可以设想一下每天数以百万计的电话呼叫；政府只对其中的极小一部分感兴趣。此外，还可以设想一下生产控制信息；几乎所有生产记录都不会引起人们的兴趣，只有极少数除外（通常是当测量参数超过某个阈值时）。一般情况下，对重复型非结构化的记录而言，还存在一些虽然并不能直接或马上引起人们兴趣但是却存在潜在价值的记录。

对于非重复型非结构化数据而言，人们不感兴趣的记录就没那么多了。尽管其中有垃圾信息和停用词，但是除了这两种类别的信息之外，几乎其他所有的非结构化非重复型数据都是人们感兴趣的。

1.1.4 大数据

值得注意的是，企业中的大数据包括重复型非结构化数据和非重复型非结构化数据，如

图1.1.6所示。

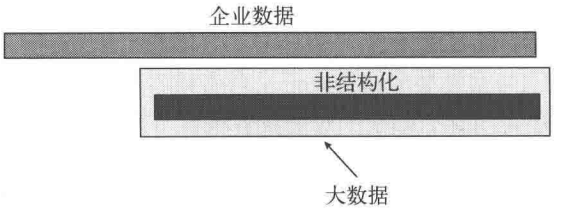


图　1.1.6

1.1.5　分界线

一开始，对于非结构化数据的两种类型（重复型非结构化数据和非重复型非结构化数据），我们可能认为它们之间的差别是难以预料、微不足道的。实际上，这两种非结构化数据类型之间的差异并非微不足道。因为这两种非结构化数据类型存在深刻差异，所以它们之间存在一条明显的分界线。

图1.1.7展现了分割两种非结构化数据类型的分界线。

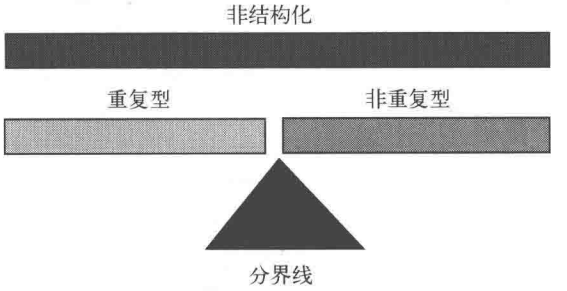


图　1.1.7

之所以用这条分界线划分非结构化数据的两种类型，是因为在分界线一边的数据是以一种方式处理的，而在分界线另一边的数据则是以另一种完全不同的方式处理的。实际上，在分界线两边的数据也可能完全不同。

按照数据处理方式进行划分的原因是，重复型非结构化数据几乎完全是通过一个管理Hadoop的固定设施来处理的。对于重复型非结构化数据而言，其重点完全集中在对大数据管理器（例如Hadoop）中的数据进行访问、监视、显示、分析和可视化。

非重复型非结构化数据的重点则几乎完全集中在文本消歧上。这里的重点在于消歧的类型、输出的重新格式化、数据的上下文分析和数据的标准化等。

该分界线值得注意的一点是，围绕分界线两边不同类型的数据形成的学科也是完全不同的。文本消歧与访问和分析Hadoop中的数据是两个极为不同的课题。正是因为这两个领域存在极大差异，可以说这两个领域属于完全不同的范畴，之间毫无关系。

可以用一个比喻来说明管理Hadoop和管理文本消歧这两个领域有多么不同。管理Hadoop就像生物医学领域，而文本消歧领域就像竞技骑牛领域。这两个领域截然不同，二者之间根本没有可比性。研究生物医学领域的人完全不知道骑着一头野牛是什么感觉，而擅长骑野牛的骑牛士与生产新药所需的规程格格不入。

图1.1.8描绘了这两个领域之间的差别。

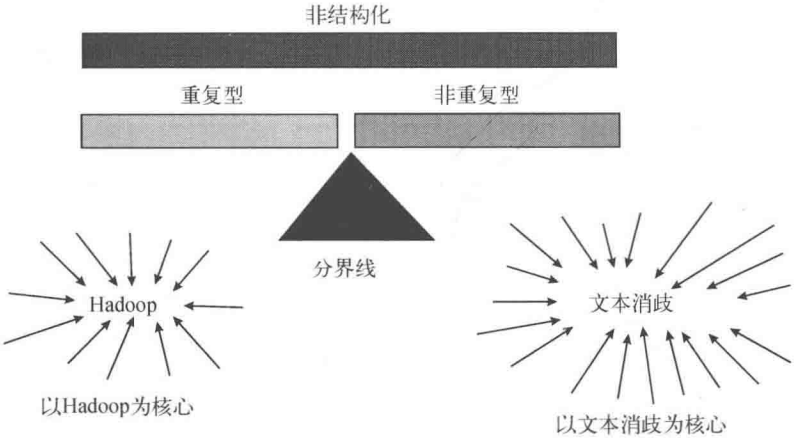


图 1.1.8

1.1.6 大陆分水岭

与非结构化数据分界线相似的另一条分界线是北美大陆分水岭（如图1.1.9所示）。大陆分水岭一侧的降水会流向大西洋方向，而另一侧的降水则流向完全不同的太平洋方向。

图 1.1.9