商业分析

Business Analytics

商业数据存储和管理

阮光册◎编著

华东师范大学出版社

商业分析

Business Analytics

商业数据存储和管理

阮光册◎编著

华东师范大学出版社

图书在版编目（CIP）数据

商业数据存储和管理/阮光册编著. —上海：华东师范大学出版社，2016.1
（商业分析丛书）
ISBN 978-7-5675-4613-4

Ⅰ.①商… Ⅱ.①阮… Ⅲ.①商业信息—数据管理 Ⅳ.①F713.51

中国版本图书馆 CIP 数据核字（2016）第 016971 号

商业数据存储和管理

编　　著　阮光册
策划组稿　孙小帆
项目编辑　孙小帆
特约审读　金　天
装帧设计　卢晓红　俞　越

出版发行　华东师范大学出版社
社　　址　上海市中山北路 3663 号　邮编 200062
网　　址　www.ecnupress.com.cn
电　　话　021-60821666　行政传真 021-62572105
客服电话　021-62865537　门市（邮购）电话 021-62869887
地　　址　上海市中山北路 3663 号华东师范大学校内先锋路口
网　　店　http://hdsdcbs.tmall.com/

印 刷 者　常熟高专印刷有限公司
开　　本　787×1092　16 开
印　　张　15
字　　数　278 千字
版　　次　2016 年 4 月第 1 版
印　　次　2016 年 4 月第 1 次
书　　号　ISBN 978-7-5675-4613-4/F·351
定　　价　39.00 元

出 版 人　王　焰

本书简介

目标：

通过对商业数据存储领域相关知识点的介绍，帮助阅读者建立商业数据管理（Business Data Management）的整体知识框架，初步掌握商业数据存储与管理的思维，即商业环境中的网络化存储、云存储、商业数据架构、数据仓库、数据质量、数据安全等知识，并对商业数据管理常用的工具有一定了解和认识。

内容组织：

本书分为基础编、分析编、提高编、工具编和展望编五部分。

基础编介绍了数据对企业的重要性（第 1 章）、数据存储的基本方法（第 2 章）和云存储的相关知识（第 3 章）。通过这 3 章帮助大家理解数据存储与管理的一般知识。

分析编是对商业数据管理的相关内容介绍，包括商业数据集成（第 4 章）、数据备份与恢复（第 5 章）、商业数据架构管理（第 6 章）以及数据仓库（第 7 章），在商业数据存储的基础之上，对商业数据管理的方法、思路进行了分析。

提高编是在数据管理的基础上，重点强调了数据质量管理（第 8 章）和数据安全管理（第 9 章）的问题。

工具编主要介绍了商业数据存储与管理的主要工具（第 10 章），描述了这些工具的使用方法和特点。

展望编阐述了商业数据存储与管理的发展趋势（第 11 章），就大数据、移动数据、数据智能管理、数据中心等进行介绍。

体例特点：

本书在概括性介绍的基础之上，辅以案例介绍，主要目的是构建一个相对完整的商业数据存储与管理的学习框架。每章结尾部分均有本章小结，有助于帮助阅读者进一步总结和思考。

目录

第五编　展望编　215

第一编 基础编

1 信息：企业的资产

一个企业如果没有认识到管理数据和信息资产如同管理有形资产一样极其重要，那么它在新经济时代将无法生存。 ——汤姆·比德斯，2001

1.1 数据在企业中的地位

数据和信息是21世纪的经济命脉。数据如何影响企业？数据能否给企业带来利润？这是每一个企业在面对数据时都会面临的问题，可以通过一个在《大数据时代》中非常著名的大数据在企业中的实际应用案例来回答这些问题。

2003年，奥伦·埃齐奥尼(Oren Etzioni)准备乘坐从西雅图到洛杉矶的飞机去参加弟弟的婚礼。他知道飞机票越早预订越便宜，于是他提前几个月在网上预订了一张去洛杉矶的机票。在飞机上，埃齐奥尼好奇地问邻座乘客花了多少钱购买机票。当得知那个人的机票虽然比他买得更晚，但是票价却比他便宜得多时，他感到非常气愤。于是，他又询问了另外几个乘客，结果发现大家买的票都比他的便宜。对大多数人来说，这种被敲竹杠的感觉也许会随着他们走下飞机而消失。然而，埃齐奥尼是美国最有名的计算机专家之一，从担任华盛顿大学人工智能项目的负责人开始，他创立了许多在今天看来非常典型的大数据公司，而那时候还没有人提出“大数据”这个概念。

1994年，埃齐奥尼参与创建了最早的互联网搜索引擎MetaCrawler，该引擎后来被InfoSpace公司收购；他参与创立了第一个大型比价网站Netbot，后来把它卖给了Excite公司；他创立的从文本中挖掘信息的公司ClearForest则被路透社收购了。在他眼中，世界就是由一系列的大数据问题组成，而且他认为他有能力解决这些问题。作为哈佛大学首届计算机科学专业的本科毕业生，自1986年毕业以来，他也一直致力于解决这些问题。

飞机着陆之后，埃齐奥尼下定决心要帮助人们开发一个系统，用来推测当前网页上的机票价格是否合理。作为一种商品，同一架飞机上每个座位的价格本来不应该有差别，但实际上的价格

却千差万别，其中缘由只有航空公司自己清楚。

埃齐奥尼表示，他不需要去解开机票价格差异的奥秘，他要做的仅仅是预测当前的机票价格在未来一段时间内会上涨还是下降。这个想法是可行的，但操作起来并不是那么简单。这个系统需要分析所有特定航线机票的销售价格并确定票价与提前购买天数的关系。如果一张机票的平均价格呈下降趋势，系统就会帮助用户做出稍后再购票的明智选择。反过来，如果一张机票的平均价格呈上涨趋势，系统就会提醒用户立刻购买该机票。简言之，这是埃齐奥尼针对 9 000 米高空开发的一个加强版的信息预测系统。这确实是一个浩大的计算机科学项目。不过，这个项目是可行的。于是，埃齐奥尼开始着手启动这个项目。埃齐奥尼创立了一个预测系统，它帮助虚拟乘客节省了很多钱。这个预测系统建立在 41 天内价格波动产生的 12 000 个价格样本基础之上，而这些信息都是从一个旅游网站上搜集来的。这个预测系统不能说明原因，只能推测会发生什么。也就是说，它不知道是哪些因素导致了机票价格的波动。机票降价是因为很多座位没卖掉、季节性原因，还是所谓的周六晚上不出门，它都不知道，只知道利用其他航班的数据来预测未来机票价格的走势。“买还是不买，这是一个问题。”埃齐奥尼沉思着。最终，他给这个研究项目取了一个非常贴切的名字叫“哈姆雷特”。这个小项目后来得到风险投资基金支持，并逐渐发展成为一家科技创业公司，名为 Farecast。通过预测机票价格的走势以及增降幅度，Farecast 票价预测工具能帮助消费者抓住最佳购买时机，而在此之前还没有其他网站能让消费者获得这些信息。这个系统为了保障自身的透明度，会把对机票价格走势预测的可信度标示出来，供消费者参考。系统的运转需要海量数据的支持。为了提高预测的准确性，埃齐奥尼找到了一个行业机票预订数据库。有了这个数据库，系统进行预测时，预测的结果就可以基于美国商业航空产业中，每一条航线上每一架飞机内的每一个座位一年内的综合票价记录而得出。如今，Farecast 已经拥有 2 000 亿条飞行数据记录。利用这种方法，Farecast 为消费者节省了一大笔钱。

Farecast 是大数据公司的一个缩影，也代表了当今世界发展的趋势。五年或者十年之前，奥伦·埃齐奥尼认为成立这样的公司是不可能的。那时候，所需要的计算机处理能力和存储能力相对来说太昂贵了！而技术上的突破是这一切得以发生的主要原因，同时也有一些细微而重要的改变正在发生，特别是人们关于如何使用数据的理念。

1.1.1 数据与信息

长期以来，人们认为资产是指企业过去的交易或事项形成的，由企业拥有或控制的，预期会给企业带来经济利益的资源，包括有形和无形资产。现在，数据已经被公认为是企业的资产了。离开了高质量的数据，企业难以在激烈的竞争中生存。今天，各个组织都依赖于它们的数据资产

从而做出更明智和更有效的决策。市场领导者正在利用他们手中的数据资产，通过丰富的客户资料，信息的创新使用和高校的运营来获得竞争优势。企业通过使用数据，以提供更好的产品和服务，降低成本，控制风险。政府、教育机构以及非营利组织也需要高质量的数据来指导其日常运营、战术和战略活动。随着企业对数据需求的不断增长，以及企业对数据的依赖性不断增强，人们可以越来越清楚地评估数据资产的商业价值。

目前，世界上的数据以每年10—20亿字节的速度飞速增长，然而，面对许多重要决定，我们在已知信息和做决策所需的信息之间存在巨大差距。这种差距可能对企业的经营效益和盈利能力包括战略决策产生深远影响。

(1) 什么是数据?

数据是指某一目标定性、定量描述的原始资料，包括数字、文字、符号、图形、图像以及它们能够转换成的数据等形式。数据本身无特定含义，只是记录事物的性质、形态、数量特征的抽象符号。数据是描述现实事物的符号记录，是用物理符号记录下来的可以识别的信息。

(2) 什么是信息?

信息的概念是由 N. Wiener 在 1948 年发表的《控制论》中提出的。他认为“信息这个名称的内容就是我们将事物调节为外界所了解时而与外界交换来的东西”。Wiener 的定义揭示了信息的功能和范围。但他将交换的东西都定义为信息是不准确的，因为还有物质和能量的交换。

然后，信息的概念不断得到发展，人们对信息所提出的概念不下百种，一些比较有名的定义如下：

① 信息就是信息，既不是物质也不是能量(Wiener，1948)；

② 信息是用来消除不确定性的东西(Shannon，1948)；

③ 信息是事物之间的差异(G. Longo，1975)；

④ 信息是一种场(Eepr，1971)；

⑤ 信息是负熵(Brillouin，1956)；

⑥ 信息是对某种事物的预报(《广辞苑》)。

可以看出，不同的角度对信息的定义很不同。哲学家研究信息的最一般特性，认为信息是物质存在的普遍形式，是与“原型世界”对应的“信息世界”；语言学家对“信息”作词语解释，将“信息”与“音讯”、“消息”一样看待；计算机科学的研究者们研究信息语义形式化问题，即对符号表达式与它的内涵之间的关系做研究，所以在研究信息处理的语言学家眼中，信息是“符号”，而所谓“语义学”就是对符号表达式与它的内涵之间的研究；社会学家考虑信息的载体，通常把“信息”说成“消息”；通信专家则对信息的度量、量化以及编码进行论述，在描述过程中，常以“信号”代表“信息”。

(3) 从对信息认识的层次上来说，可以将其分为三个层次：

① 本体论层次：信息是事物存在的方式和运动状态的表现形式。这是最高的层次，也是普遍的层次，无条件约束的层次。在这个层次上定义的信息是最广义的信息，适用范围最广。引入一个约束条件就会把最高层次变为次高层次的定义，由于加了约束，次高层次的适用范围就要比最高层次的窄，引入的条件越多，定义的层次就越低，所定义的信息适用范围就越窄。

② 认识论层次：信息是主体所感知或表达的事物存在的方式和运动状态。

③ 负熵论层次：信息是用来减少随机不确定性的东西。

数据和信息的区别在于，信息是对数据加工处理后得到的有用数据。信息是数据的语义解释，是数据的内涵。信息以数据的形式表现出来，并为人们所理解接受。

1.1.2 商业信息

从广义角度讲，商业信息是指能够反映商业经济活动情况，同商品交换和管理有关的各种消息、数据、情报和资料的统称。商业信息的范畴不但包括直接反映商业购销和市场供求变化及供应运动的信息，还包括各种影响市场供求变化的有关情况的信息，如自然灾害、政治事件等会影响当年或来年市场商品可供量的增减、购买力投向变化等，有关这些方面的情况变化也可纳入商业信息的范围。

从狭义角度看，商业信息是指直接反映商品买卖活动的特征、变化等情况的各种消息、情报、资料的统称。

(1) 商业信息的特点

商业信息是社会生产、交换、消费等经济活动必不可少的信息，它除了具有一般信息共有的可传递性、可记存性、可复制性、可共享性等特点外，还具有多变性、零散性和实用性等。

① 多变性。商业信息的多变性是它不同于其他信息的最突出的特征，主要表现为以下三点：一是商品价格信息瞬息万变，而且不同商品之间的比价也不断变化；二是商品的供求关系处在不断变动之中。由于大量新的企业参与市场竞争，某些商品往往很快由短缺转为过剩，畅销商品与滞销商品的位次也不断调整；三是商品的品种、品牌不断增多，同一种商品的更新换代周期越来越短，使用功能也不断趋向于复合多元。

② 零散性。商品信息的零散性与商品生产的分散性和商品信息传播的多渠道、无序化密切相关。主要表现为以下三点：一是商品生产多以分散的企业或企业集团为单位，为占领市场，企业只注重商品信息的及时发布而缺乏累积性，造成商品信息满天飞的局面；二是商品信息经过各种社会传播渠道传播时，虽经过一定的整合，但仍然无法从根本上改变其分散的状态；三是在以

商品销售为目的的信息传播活动中，良莠不齐，存在片面、无序、虚假宣传的现象。

③ 实用性。商品信息的实用性与商品信息的功能密切相关。它主要表现在以下几个方面：一是沟通社会生产、流通、消费等环节的联系，促使其出现良性循环；二是贴近大众生活，有广泛的共享性，提高经济活动的透明度；三是服务于不同用户的需求。如企业可以据此了解竞争对手的生产情况、商品营销策略、价格与服务措施、商品的市场占有率等，从而有针对性地组织生产销售，使自己在竞争中占据有利地位。

(2) 商业信息与商业数据的区别

商业信息是人们对与商业活动相关的事物及其变化规律的认识。这些认识被某种载体记录，进而加工、利用和传播，使其具有更高的利用价值。商业数据是记录商业信息的载体。商业数据可以以文字、数值、票据、凭证、文件、图表、多媒体等多种方式对商业信息进行记录。

在商业信息系统中，商业信息以某种方式被记录下来，由此产生商业数据，所以商业信息是对商业数据的解释，是有一定含义且具有价值的数据。对商业数据进行计算加工，得到新的商业数据，这些新的商业数据可为进一步的管理、决策提供依据，是具有更高价值的商业信息，经过这样螺旋式的上升，使商业信息管理成为商品流通企业不断发展的推动力，这也正是管理商业信息的最终目的。由此可见，商业信息是事物的本质，商业数据是管理商业信息的媒介。商业信息与商业数据在运动中相互依存、相互转化，所以在使用中有时对二者并不进行严格区分。

1.1.3 信息资产

信息资产是由企业拥有或者控制的能够为企业未来带来经济利益的信息资源。其本质是信息作为一种经济资源参与企业的经济活动，减少和消除了企业经济活动中的风险，为企业的管理控制和科学决策提供合理依据，并预期给企业带来经济利益。

企业取得的信息资产正式投入生产经营运作以后，由于合理利用了现有的生产资源，充分减少了企业的生产经营费用，大大提高了产品的质量，提高了产品的产销量，实现了更多的实际收入，从而提高了企业的经营利润。

(1) 信息资产的分类

根据我国大多数企业目前的经营状况，信息资产按内容大致可分为四大类：科学技术信息资产、市场信息资产、生产信息资产和外部宏观信息资产。

① 科学技术信息资产。科学类信息资产是第一大类，是指企业在生产经营和科学实验等创新过程中，所发明创造的高新技术和技术诀窍而形成的一种产权形式。主要包括专利权、版权、技术机密、计算机软件等。

② 市场信息资产。第二大类是市场类信息资产，它指一个企业通过其所拥有的与市场相关联的信息资产而可能获得的未来经济利益，主要包括品牌、客户关系和合同等。

③ 生产信息资产。生产类信息资产是指企业在日常生产经营活动中的各种生产情况记录形成的信息，这类信息资产对企业成本核算和成本控制有极其重要的作用。包括原材料信息、加工信息、存储信息和传输信息等。

④ 外部宏观信息资产。外部宏观信息则是指企业针对其所生存的宏观环境进行分析所获取的信息，包括社会发展信息、政策法规信息和技术经济信息等。

(2)信息资产的特征

信息资产是企业拥有和控制的一项特殊资产，既具有一般物质资产的特征，又兼有无形资产和信息资源的双重特征。概括来说，信息资产特有的特征主要表现在以下几个方面：

① 信息资产的首要特征是共享性，即使用的非排他性。信息资产的共享性来自信息本身的非占有性。信息资产持有者不会因为传递信息而失去它们，信息资产的获得者取得信息资产也不以其持有者失去信息资产为必要前提，二者在信息使用上不存在竞争关系，因而信息资产可以被反复交换、反复使用。随着市场经济的发展和政府作用的不断增强，信息的共享性已经受到一定程度的限制，但是即使是这种受知识产权保护的信息资产在一定的范围内和一定的条件下也具有非排他性。这是因为，信息资产持有者可以通过收取使用费的方式允许他人使用而自己并不一定失去该信息资产。再者，由于社会知识产权制度只能对知识产权进行有限保护，使得信息资产不是被永久独占使用，而是在保护期结束之后会被全社会无偿使用，从而更加表现出其非排他性。此外，也会出现其他组织不遵守知识产权制度而窃取、仿制、传播信息资产的行为，使得信息资产在一定程度上难以被独占。

② 信息资产具有高附加值。信息资产一旦被企业应用，就能创造出巨大的潜在价值，其所产生的经济利益将不可估量。信息资产带来经济利益的规模，通常不是受到生产规模的约束，而是受到市场规模的约束。比如计算机软件，一旦研究与开发成功并投入市场，用于承载该软件的磁盘的边际成本微乎其微，生产、销售越多，其所产生的利润就越高。

③ 信息资产具有高风险性。信息资产的高风险性源自信息资产使用的高附加值和传播的低成本性。在激烈竞争的市场环境中，信息资产的安全问题至关重要。一般来说，信息资产经常处于公共的介质中或处于流动状态，这就使信息资产的复制成本较低，从而导致企业拥有和控制的信息资产的安全性很差。没有安全保障的信息资产，谈不上资产价值。

④ 信息资产具有强烈的时效性。信息资产比其他任何资产更加具有时效性。首先，对于一些流动性极强的信息资产来说，比如市场类信息资产，如果不能在最恰当的时机加以开发利用，

机会就会稍纵即逝，此后再对该信息资产进行利用，也不可能达到最好的效果，甚至可能会完全失去效用；其次，即使对于流动性相对较弱的信息资产，比如受到产权保护的科学技术类信息资产，其效用的发挥同样具有时效性。一般而言，科学技术类信息资产的产权保护期限都在10年以上，其时效性表现在，一旦过了保护期，这些信息资产不再受到保护而充分显示其共享性的特征；第三，在科学技术飞速发展的今天，技术的更新换代越来越频繁，其生命周期越来越短，即使在保护期内，信息资产也可能由于技术进步而很快被取代或者淘汰。

(3) 信息资产和日常生活中资产的区别

信息资产与物理资产的基本区别是，信息资产是动态变化的，而物理资产是固定不变的。信息资产在许多方面表现出动态特征，从信息以运行数据(客户账户、业务交易等)的形式产生开始，直到在各种业务功能和过程中最终的应用(ERP、CRM、商业智能)。IT界为信息生命周期的每一个阶段推出了许多单一性的产品。这些产品分别用于解决生命周期中某个方面的问题，包括信息的生成、处理、分布、存档、检索和处置。某一种信息资产在生命周期的每一个阶段各有其价值，一般来说它的价值会随着生命周期的进展而增加。

1.2 信息存储

商业世界依赖于快速可靠的信息访问，这对企业成功十分重要。设计信息处理的商业应用包括机票预订、电话收费系统、电子交易、ATM机、产品设计、存货管理、邮件存档、门户网站、专利记录、信用卡以及全球资本市场等。

信息对商业的日益增长的重要性大大增加了对数据管理和保护的挑战性。商业机构需要管理的数据信息正驱动着各种策略的产生，使其在数据生命周期内，根据数据的价值来分类和创建数据管理规则。这些策略不仅可以在商业上带来经济和监管利益，而且有利于组织在操作上的管理。

信息存储是指将经过加工整理序化后的信息按照一定的格式和顺序存储在特定的载体中的一种信息活动。目的是为了便于信息管理者和信息用户快速、准确地识别、定位和检索信息。信息的储存是信息系统的重要方面，如果没有信息储存，就不能充分利用已收集、加工所得的信息，同时还要耗资、耗人、耗物来组织信息的重新收集、加工。有了信息储存，就可以保证随用随取，为单位信息的多功能利用创造条件，从而大大降低费用。

1.2.1 存储虚拟化技术

随着计算机内信息量的不断增加，以往直连式的本地存储系统已无法满足业务数据的海量