

大数据分析

张维冲

华北电力大学法政系

2024年春



授课形式：实战为主、理论为辅；线上+线下混合模式。

授课内容：

- I. 大数据分析基础
- II. 大数据采集和处理
- III. 大数据分析与挖掘
- IV. 数据可视化与解释
- V. 实践项目：政策文本数据
- VI. 实践项目：政务理论数据
- VII. 实践项目：政府舆情数据
- VIII. 实践项目：政府招聘数据
- IX. 实践项目：行管考研数据
- X. 实践项目：多源数据融合

道

法

术

器

考核方式:

- (1) 考勤记录
- (2) 平时成绩 (课堂参与、小组讨论、在线作业等)
- (3) 小组合作完成的大数据分析项目, 需提交项目报告 (或课堂展示)
- (4) 期末测试记录 60% (笔试+上机测试)

分组:

课代表:

技术顾问:

课程介绍

推荐教材:



课堂纪律强调：

- 1.积极参与：**鼓励同学们在课堂上积极参与讨论和活动，对课堂内容提出问题和意见，深化理解和增进交流，提升课程教学质量。
- 2.手机和电子设备：**本课程无特殊说明情况下需常备电脑上课；严禁上课期间使用电子设备进行与课程无关的活动，严禁带耳机，一经发现本课程的平时成绩直接记为0！
- 3.提交作业：**按时提交作业和实践项目报告，如有特殊情况无法按时完成，请提前与教师沟通。
- 4.诚信学习：**严禁任何形式的作弊行为，包括但不限于抄袭、代写和在考试中使用不正当手段获取答案等。



华北电力大学

NORTH CHINA ELECTRIC POWER UNIVERSITY

第1章 大数据分析基础

目 录

1.1

什么是大数据

1.2

大数据的定义

1.3

大数据的结构类型

1.4

大数据应用改变生活

1.5

认识大数据分析

1.6

大数据分析生命周期



PART 01

1.1

什么是大数据

1.1 什么是大数据

信息社会所带来的好处是显而易见的：每个口袋里都揣着一部手机，每台办公桌上都放着一台电脑，每间办公室内都连接到局域网或者互联网。半个世纪以来，随着计算机技术全面和深度地融入社会生活，信息爆炸已经积累到了一个引发变革的程度。它不仅使世界充斥着比以往更多的信息，而且其增长速度也在加快。信息总量的变化还导致了信息形态的变化——**量变引起了质变。**



1.1 什么是大数据

1.1.1 天文学—— 信息爆炸的起源

1.1.2 信息爆炸的 社会

1.1.3 大数据的发展

半个世纪以来，随着计算机技术全面和深度地融入社会生活，信息爆炸已经积累到了一个引发变革的程度。

1.1.1 天文学——信息爆炸的起源

综合观察社会各个方面的变化趋势，我们能真正意识到信息爆炸或者说大数据时代已经到来。以天文学为例，2000年斯隆数字巡天项目（SDSS）启动的时候，位于美国新墨西哥州的望远镜在短短几周内收集到的数据，就比世界天文学历史上总共收集的数据还要多。到了2010年，信息档案已经高达 1.4×2^{42} 字节。



美国斯隆数字巡天望远镜

1.1.1 天文学——信息爆炸的起源

斯隆数字巡天使用阿帕奇山顶天文台的2.5米口径望远镜，计划观测25%的天空，获取超过一百万个天体的多色测光资料和光谱数据。2006年，斯隆数字巡天进入名为SDSS-II的新阶段，进一步探索银河系的结构和组成，而斯隆超新星巡天计划搜寻超新星爆发，以测量宇宙学尺度上的距离。



不过人们认为，在智利帕穹山顶峰LSST天文台投入使用的大型视场全景巡天望远镜（LSST）五天之内就能获得同样多的信息。

智利帕穹山顶峰的LSST全景巡天望远镜

1.1.1 天文学——信息爆炸的起源

LSST巡天望远镜于2015年开始建造，重3吨，32亿像素，它将由189个传感器和接近3吨重的零部件组装完成，可以捕捉半个地球。根据该项目建设的时间表，它将在2020年第一次启动，2022年到2023年开始运行。



1.1.1 天文学——信息爆炸的起源

LSST望远镜的镜头拍摄的一张照片将需要1500块高清电视屏才能充分展示出来，其一年的观测数据将达到600万GB的存储空间。这个数据量相当于用一款800万像素的数码相机每天拍摄80万张照片，连续拍摄一整年。未来，LSST望远镜将绘制数百亿恒星的分布，为科学家提供最佳的光学照片，以前所未有的细节拍摄深空天体图像。科学家能够据此研究星系的形成、追踪潜在威胁的小行星、观测恒星爆炸，研究暗物质和暗能量等。

天体图像



1.1.1 天文学——信息爆炸的起源

LSST有一个很特别的地方，那就是世界上任何一个有电脑的人都可以使用它，这和以前的科学专业设备不同。LSST数据的开放，意味着大家都有机会与科学家分享令人兴奋的探索旅程。LSST可以帮助我们解开宇宙的谜团，对于科学研究具有划时代的重大意义。



1.1.2 信息爆炸的社会

天文学领域 发生的变化在社会各个领域都在发生。

2003年，人类第一次破译人体基因密码的时候，辛苦工作了十年才完成三十亿对碱基对的排序。大约十年之后，世界范围内的基因仪每15分钟就可以完成同样的工作。

在金融领域，美国股市每天的成交量高达70亿股，而其中三分之二的交易都是由建立在数学模型和算法之上的计算机程序自动完成的，这些程序运用海量数据来预测利益和降低风险。



1.1.2 信息爆炸的社会

互联网公司更是被数据淹没了。**谷歌**公司每天要处理超过24拍字节（PB，250字节）的数据，这意味着其每天的数据处理量是美国国家图书馆所有纸质出版物所含数据量的上千倍。**脸书**这个创立不过十来年的公司，每天更新的照片量超过1 000万张，每天人们在网站上点击“喜欢”（Like）按钮或者写评论大约有三十亿次，这就为脸书挖掘用户喜好提供了大量的数据线索。与此同时，谷歌的子公司**YouTube**是世界上最大的视频网站，它每月接待多达8亿的访客，平均每一秒钟就会有一段长度在一小时以上的视频上传。**推特**是美国的一家社交网络及微博客服服务的网站，是互联网上访问量最大的十个网站之一，其消息也被称作“推文”，它被形容为“互联网的短信服务”。推特上的信息量几乎每年翻一番，每天都会发布超过4亿条微博。



1.1.2 信息爆炸的社会

从科学研究到医疗保险，从银行业到互联网，各个领域都在讲述着一个类似的故事，那就是爆发式增长的数据量。这种增长超过了创造机器的速度，甚至超过了人们的想象。那么，我们周围到底有多少数据？增长的速度有多快？许多人试图测量出一个确切的数字。尽管测量的对象和方法有所不同，但他们都获得了不同程度的成功。



1.1.2 信息爆炸的社会

南加利福尼亚大学通信学院的**马丁·希尔伯特**进行了一个比较全面的研究，他试图得出人类所创造、存储和传播的一切信息的确切数目，研究范围不仅包括书籍、图画、电子邮件、照片、音乐、视频（模拟和数字），还包括电子游戏、电话、汽车导航和信件。他还以收视率和收听率为基础，对电视、电台这些广播媒体进行了研究。据他估算，仅在2007年，人类存储的数据就超过了**300艾字节**。下面这个比喻应该可以帮助人们更容易地理解这意味着什么：一部完整的数字电影可以压缩成一个GB的文件，而**一个艾字节相当于10亿GB**，一个泽字节（ZB，270字节）则相当于1024艾字节。总之，这是一个非常庞大的数量。

1.1.2 信息爆炸的社会

有趣的是，在2007年的数据中，只有7%是存储在报纸、书籍、图片等媒介上的模拟数据，其余全部是数字数据。模拟数据也称为**模拟量**，相对于数字量而言，指的是取值范围是连续的变量或者数值，例如声音、图像、温度、压力等。模拟数据一般采用模拟信号，例如用一系列连续变化的电磁波或电压信号来表示。数字数据也称为**数字量**，相对模拟量而言，指的是取值范围是离散的变量或者数值。数字数据采用数字信号，例如用一系列断续变化的电压脉冲（如用恒定的正电压表示二进制数1，用恒定的负电压表示二进制数0）或光脉冲来表示。



1.1.2 信息爆炸的社会

但在不久之前，情况却完全不是这样的。虽然1960年就有了“信息时代”和“数字村镇”的概念，2000年数字存储信息仍只占全球数据量的四分之一，当时，另外四分之三的信息都存储在报纸、胶片、黑胶唱片和盒式磁带这类媒介上。事实上，1986年，世界上约40%的计算能力都在袖珍计算器上运行，那时候，所有个人电脑的处理能力之和还没有所有袖珍计算器处理能力之和。但是因为数字数据的快速增长，整个局势很快就颠倒过来了。按照希尔伯特的说法，数字数据的数量每三年多就会翻一倍。相反，模拟数据的数量则基本上没有增加。



1.1.2 信息爆炸的社会

到2013年，世界上存储的数据达到约1.2泽字节，其中非数字数据只占不到2%。这么大的数据量意味着什么？如果把这些数据全部记在书中，这些书可以覆盖整个美国52次。如果将之存储在只读光盘上，这些光盘可以堆成五堆，每一堆都可以伸展到月球。事情真的在快速发展。人类存储信息量的增长速度比世界经济的增长速度快4倍，而计算机数据处理能力的增长速度则比世界经济的增长速度快9倍。难怪人们会抱怨信息过量，因为每个人都受到了这种极速发展的冲击。



1.1.2 信息爆炸的社会

量变导致质变。物理学和生物学都告诉我们，当改变规模时，事物的状态有时也会发生改变。以专注于把东西变小而不是变大的纳米技术为例，其原理就是当事物到达分子级别时，它的物理性质会发生改变。一旦你知道这些新的性质，就可以用同样的原料来做以前无法做的事情。铜本来是用来导电的物质，但它一旦到达纳米级别就不能在磁场中导电了。银离子具有抗菌性，但当它以分子形式存在时这种性质会消失。同样，当我们增加所利用的数据量时，也就可以做很多在小数据量的基础上无法完成的事情。

大数据的科学价值和社会价值正是体现在这里。一方面，对大数据的掌握程度可以转化为经济价值的来源。另一方面，大数据已经撼动了世界的方方面面，从商业科技到医疗、政府、教育、经济、人文以及社会的其他各个领域。尽管我们还处在大数据时代的初期，但我们的日常生活已经离不开它了。

1.1.3 大数据的发展

如果仅仅从数据量的角度来看，大数据在过去就已经存在了。例如，波音的喷气发动机每30分钟就会产生10TB的运行信息数据，安装有4台发动机的大型客机，每次飞越大西洋就会产生640TB的数据。世界各地每天有超过2.5万架的飞机在工作，可见其数据量是何等庞大。生物技术领域中的基因组分析以及以NASA（美国国家航空航天局）为中心的太空开发领域，从很早就开始使用十分昂贵的高端超级计算机来对庞大的数据进行分析 and 处理了。



1.1.3 大数据的发展

现在和过去的区别之一，就是大数据不仅产生于特定领域，而且还产生于人们的日常生活中，脸书、推特、领英、微信、QQ等社交媒体上的文本数据就是最好的例子。而且，尽管我们无法得到全部数据，但大部分数据可以通过公开的API（应用程序编程接口）相对容易地进行采集。在B2C（商家对顾客）企业中，使用文本挖掘和情感分析等技术，就可以分析消费者对于自家产品的评价。

(1) 硬件性价比提高与软件技术进步。计算机性价比的提高，存储设备价格的下降，利用通用服务器对大量数据进行高速处理的软件技术Hadoop的诞生，这些因素大幅降低了大数据存储和处理的门槛。因此，如今无论是中小企业还是大企业，都可以对大数据进行充分的利用。

1.1.3 大数据的发展

(2) 云计算的普及。随着云计算的兴起，大数据的处理环境现在在很多情况下并不一定要自行搭建了。例如，使用亚马逊的云计算服务EC2和S3，就可以以按用量付费的方式，来使用由计算机集群组成的计算处理环境和大规模数据存储环境。利用这样的云计算环境，即使是资金不太充裕的创业型公司，也可以进行大数据分析。实际上，新的IT创业公司如雨后春笋般不断出现，它们利用云计算环境对大数据进行处理，从而催生出新型的服务。例如提供预测**航班起飞晚点**等“航班预报”服务、对消费电子产品价格走势进行预测等。



1.1.3 大数据的发展

(3) 从交易数据分析到交互数据分析。对从像“卖出了一件商品”、“一位客户解除了合同”这样的交易数据中得到的“点”信息进行统计还不够，我们想要得到的是“为什么卖出了这件商品”、“为什么这个客户离开了”这样的上下文（背景）信息。而这样的信息，需要从与客户之间产生的交互数据信息中来探索。以非结构化数据为中心的大数据分析需求的不断高涨，也正是这种趋势的一个反映。

例如，像阿里巴巴运营电商网站的企业，可以通过网站的点击流数据，追踪用户在网站内的行为，从而对用户从访问网站到最终购买商品的行为路线进行分析。这种点击流数据，正是表现客户与公司网站之间相互作用的一种交互数据。

1.1.3 大数据的发展

对于消费品公司来说，可以通过客户的会员数据、购物记录、呼叫中心通话记录等数据来寻找客户解约的原因。随着“社交化CRM（客户关系管理）”呼声的高涨，越来越多的企业都开始利用微信、推特等社交媒体来提供客户支持服务。这些都是表现与客户之间交流的交互数据，只要推进对这些交互数据的分析，就可以越来越清晰地掌握客户离开的原因。

一般来说，网络数据比真实世界中的数据更容易收集，因此，来自网络的交互数据也得到了越来越多的利用。随着传感器等物态探测技术的发展和普及，在真实世界中对交互数据的利用也将不断推进。进一步讲，今后更为重要的是对连接网络世界和真实世界的交互数据进行分析。



PART 02

1.2

大数据的定义

1.2 大数据的定义

如今，人们不再认为数据是静止和陈旧的。但在以前，一旦完成了收集数据的目的之后，数据就会被认为已经没有用处了。比方说，在飞机降落之后，票价数据就没有用了——设计人员如果没有大数据的理念，就会丢失掉很多有价值的数据。

数据已经成为了一种商业资本，一项重要的经济投入，可以创造新的经济利益。事实上，一旦思维转变过来，数据就能被巧妙地用来激发新产品和新服务。今天，大数据是人们获得新的认知、创造新的价值的源泉，大数据还是改变市场、组织机构以及政府与公民关系的方法。大数据时代对我们的生活和与世界交流的方式都提出了挑战。

1.2 大数据的定义

1.2.1 定义大数据

1.2.2 大数据的3V 特征

1.2.3 广义的大数据

大数据时代对我们的生活和与世界交流的方式都提出了挑战。

1.2.1 定义大数据

所谓**大数据**，狭义上可以定义为：**用现有的一般技术难以管理的大量数据的集合**。这实际上是指用目前在企业数据库占据主流地位的关系型数据库无法进行管理的、具有复杂结构的数据。或者也可以说，是指由于数据量的增大，导致对数据的查询响应时间超出了允许的范围。

研究机构加特纳给出了这样的定义：“大数据是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。”



1.2.1 定义大数据

世界级领先的全球管理咨询公司麦肯锡说：“大数据指的是所涉及的数据集规模已经超过了传统数据库软件获取、存储、营理和分析的能力。这是一个被故意设计成主观性的定义，并且是一个关于多大的数据集才能被认为是大数据的可变定义，即并不定义大于一个特定数字的TB才叫大数据。因为随着技术的不断发展，符合大数据标准的数据集容量也会增长；并且定义随不同的行业也有变化，这依赖于在一个特定行业通常使用何种软件和数据集有多大。因此，**大数据在今天不同行业中的范围可以从几十TB到几PB。**”

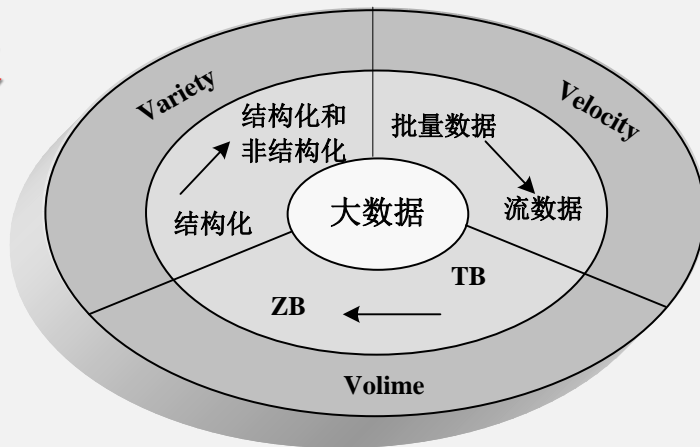
随着“大数据”的出现，数据仓库、数据安全、数据分析、数据挖掘等围绕大数据商业价值的利用正逐渐成为行业人士争相追捧的利润焦点，在全球引领了新一轮数据技术革新的浪潮。

1.2.2 大数据的3V特征

从字面上看，“大数据”这个词可能会让人觉得只是容量非常大的数据集合而已，但容量只不过是大数据特征的一个方面，如果只拘泥于数据量，就无法深入理解当前围绕大数据所进行的讨论。因为“用现有的一般技术难以管理”这样的状况，并不仅仅是由于数据量增大这一个因素所造成的。

IBM说：“**可以用3个特征相结合来定义**

大数据：数量（Volume，或称容量）、种类（Variety，或称多样性）和速度（Velocity），或者就是简单的3V，即庞大容量、极快速度和种类丰富的数据。”



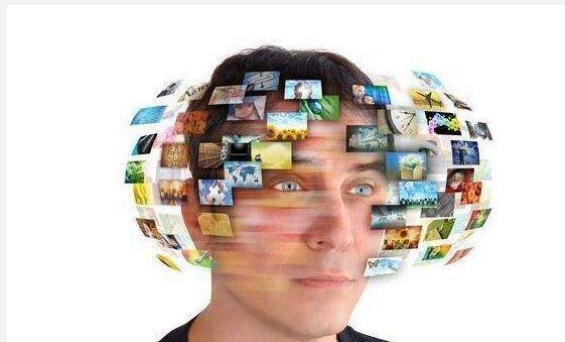
Variety 种类

Velocity 速度

Volume 数量

1.2.2 大数据的3V特征

(1) **Volume (数量)**。用现有技术无法管理的数据量，从现状来看，基本上是指从几十TB到几PB这样的数量级。当然，随着技术的进步，这个数值也会不断变化。如今，存储的数据量在急剧增长中，我们存储所有事物，包括环境数据、财务数据、医疗数据、监控数据等等，数据量不可避免地会转向ZB级别。可是，随着可供企业使用的数据量不断增长，可处理、理解和分析的数据的比例却不断在下降。



1.2.2 大数据的3V特征

(2) **Variety (种类、多样性)**。随着传感器、智能设备以及社交协作技术的激增，企业中的数据也变得更加复杂，因为它不仅包含传统的关系型数据，还包含来自网页、互联网日志文件（包括流数据）、搜索索引、社交媒体、电子邮件、文档、主动和被动系统的传感器数据等原始、半结构化和非结构化数据。

种类表示所有的数据类型。其中，爆发式增长的一些数据，如互联网上的文本数据、位置信息、传感器数据、视频数据等，用目前企业主流的关系型数据库是很难存储的，它们都属于非结构化数据。



1.2.2 大数据的3V特征

当然，这些数据中有些是过去就一直存在并保存下来的。和过去不同的是，除了存储，还需要对这些大数据进行分析，并从中获得有用的信息。例如监控摄像机中的视频数据，超市、便利店等零售企业几乎都配备了监控摄像机，最初目的是为了防范盗窃，但现在也出现了使用视频数据来分析顾客购买行为的案例。

例如，美国高级文具制造商万宝龙过去是凭经验和直觉来决定商品陈列布局的，现在尝试利用监控摄像头对顾客在店内的行为进行分析。通过分析监控摄像数据，将最想卖出去的商品移动到最容易吸引顾客目光的位置，使得销售额提高了20%。

美国移动运营商T-Mobile也在其全美1 000家店中安装了带视频分析功能的监控摄像机，可以统计来店人数，还可以追踪顾客在店内的行动路线、在展台前停留的时间，甚至是试用了哪一款手机、试用了多长时间等，对顾客在店内的购买行为进行分析。

1.2.2 大数据的3V特征

(3) **Velocity (速度)**。数据产生和更新的频率也是衡量大数据的一个重要特征。就像我们收集和存储的数据量和种类发生了变化一样，生成和需要处理数据的速度也在变化。这里，速度的概念不仅是与数据存储相关的增长速率，还应该动态地应用到数据流动的速度上。有效地处理大数据，需要在数据变化的过程中对它的数量和种类执行分析，而不只是在它静止后执行分析。

例如，遍布全国的各种便利店在24小时内产生的POS机数据，电商网站中由用户访问所产生的网站点击流数据，高峰时达到每秒近万条的微信短文，全国公路上安装的交通探测传感器和路面状况传感器（可检测结冰、积雪等路面状态）等，每天都在产生着庞大的数据。

1.2.2 大数据的3V特征

在3V的基础上，IBM又归纳总结了第四个V——**Veracity（真实和准确）**。“只有真实而准确的数据才能让对数据的管控和治理真正有意义。随着新数据源的兴起，传统数据源的局限性被打破，企业愈发需要有效的信息治理以确保其真实性及安全性。”



1.2.2 大数据的3V特征

互联网数据中心IDC说：“大数据是一个貌似不知道从哪里冒出来的大的动力。但是实际上，大数据并不是新生事物。然而，它确实正在进入主流并得到重大关注，这是有原因的。廉价的存储、传感器和数据采集技术的快速发展、通过云和虚拟化存储设施增加的信息链路，以及创新软件和分析工具，正在驱动着大数据。

大数据不是一个‘事物’，而是一个跨多个信息技术领域的动力/活动。大数据技术描述了新一代的技术和架构，其被设计用于：**通过使用高速（Velocity）的采集、发现和/或分析，从超大容量（Volume）的多样（Variety）数据中经济地提取价值（Value）。**”这个定义除了揭示大数据传统的3V基本特征，即大数据量、多样性和高速，还增添了一个新特征：价值。

1.2.2 大数据的3V特征

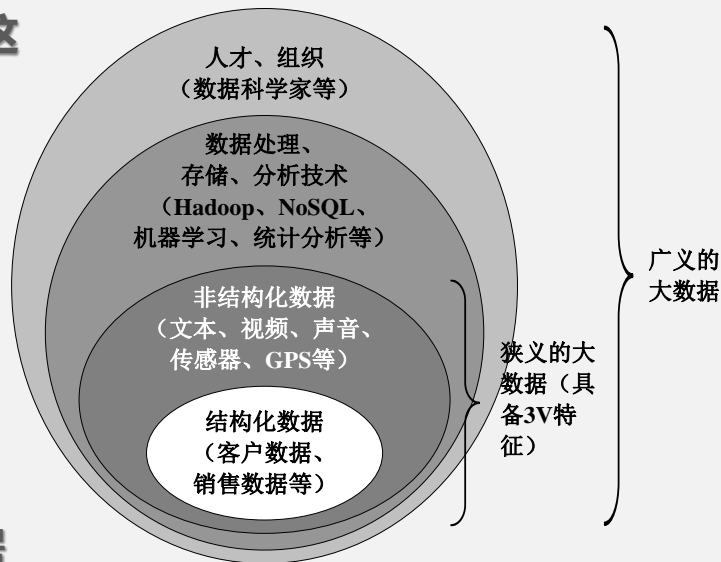
总之，大数据是个动态的定义，不同行业根据其应用的不同有着不同的理解，其衡量标准也在随着技术的进步而改变。



1.2.3 广义的大数据

大数据的狭义定义着眼点在数据的性质上，我们从广义层面上再为大数据下一个定义：

“所谓 ‘大数据’ 是一个综合性概念，它包括因具备3V（Volume/Variety/Velocity，数量/品种/速度）特征而难以进行管理的数据，对这些数据进行存储、处理、分析的技术，以及能够通过分析这些数据获得实用意义和观点的人才和组织。”



广义的大数据

1.2.3 广义的大数据

“存储、处理、分析的技术” 指的是用于大规模数据分布式处理的框架Hadoop、具备良好扩展性的NoSQL数据库，以及机器学习和统计分析等；

“能够通过分析这些数据获得实用意义和观点的人才和组织”，指的是目前十分紧俏的**“数据科学家”** 这类人才以及能够对大数据进行有效运用的组织。



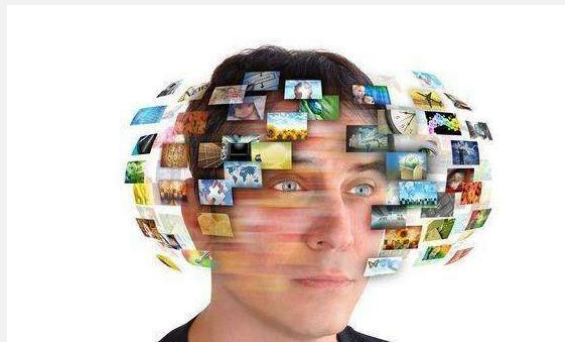
PART 03

1.3

大数据的结构类型

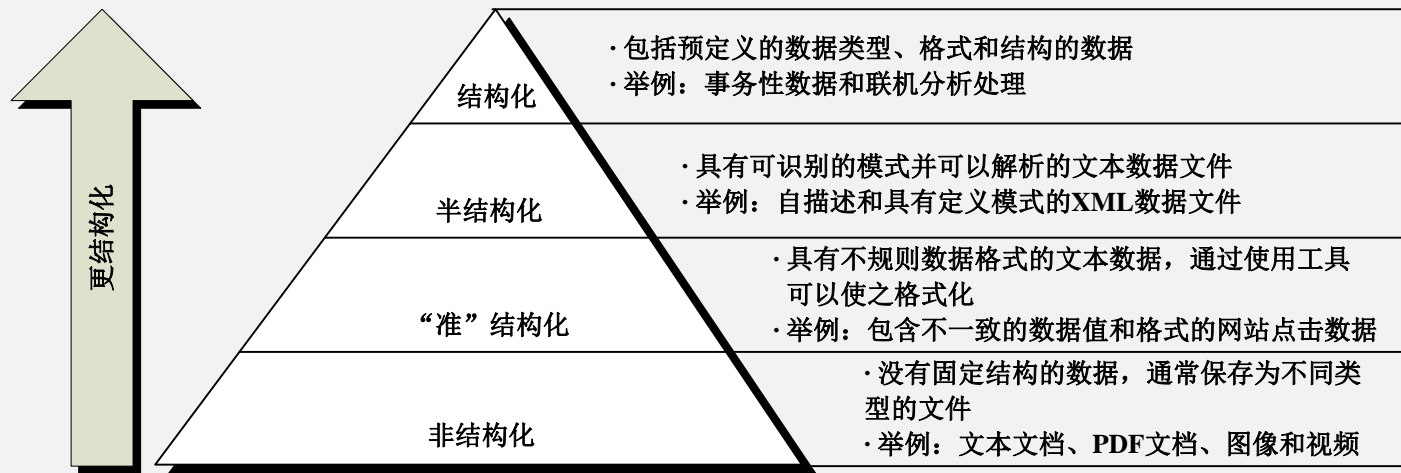
1.3 大数据的结构类型

数据量大是大数据的一致特征。由于数据自身的复杂性，作为一个必然的结果，处理大数据的首选方法是在并行计算的环境中进行大规模并行处理（Massively Parallel Processing, MPP），这使得同时发生的并行摄取、并行数据装载和分析成为可能。实际上，**大多数的大数据都是非结构化或半结构化的，需要不同的技术和工具来处理和分**
析。



1.3 大数据的结构类型

大数据最突出的特征是它的结构。下图显示了几种不同数据结构类型数据的增长趋势，由图可知，未来数据增长的80%~90%将来自于不是结构化的数据类型（半、准和非结构化）。



数据增长日益趋向非结构化

1.3 大数据的结构类型

实际上，有时这4种不同的、相分离的数据类型是可以被混合在一起的。例如，一个传统的关系数据库管理系统保存着一个软件支持呼叫中心的通话日志，这里有典型的结构化数据，比如日期/时间戳、机器类型、问题类型、操作系统，这些都是在线支持人员通过图形用户界面上的下拉式菜单输入的。另外，还有非结构化数据或半结构化数据，比如自由形式的通话日志信息，这些可能来自包含问题的电子邮件，或者技术问题和解决方案的实际通话描述。另外一种可能是与结构化数据有关的实际通话的语音日志或者音频文字实录。即使是现在，大多数分析人员还无法分析这种通话日志历史数据库中的最普通和高度结构化的数据，因为挖掘文本信息是一项强度很大的工作，并且无法简单地实现自动化。

1.3 大数据的结构类型

人们通常最熟悉结构化数据的分析，然而，半结构化数据（XML）、“准”结构化数据（网站地址字符串）和非结构化数据代表了不同的挑战，需要不同的技术来分析。除了三种基本的数据类型以外，还有一种重要的数据类型为元数据。元数据提供了一个数据集的特征和结构信息，这种数据主要由机器生成并且能够添加到数据集中。搜寻元数据对于大数据存储、处理和分析是至关重要的一步，因为它提供了数据系谱信息以及数据处理的起源。元数据的例子包括：

- XML文件中提供作者和创建日期信息的标签；
- 数码照片中提供文件大小和分辨率的属性文件。



PART 04

1.4

大数据应用改变生活

1.4 大数据应用改变生活

事实上人们每天都在体验着大数据应用带来的社会进步，例如QQ、微信、脸书、谷歌搜索、领英以及推特等等，大量数据为我们提供解析，也供我们娱乐。

脸书存储和使用的大数据形式包括用户资料、照片、信息及广告。通过分析这些数据，脸书能更好地理解用户，并判断该向用户呈现何种内容。推特每天处理的推文超过5亿，而数据分析的创业公司Topsy主营推文的实时分析，使用这些数据源在推特及其他平台顶部建立应用程序。谷歌抓取数十亿网页，并拥有大量的其他大数据源。例如谷歌地图包含的海量数据，有实际街道位置，也有卫星图像、街道照片，甚至还有许多建筑的内部图。而领英掌握了数以百万计的在线简历以及人们如何相互联系的信息，它使用所有数据在数百万人当中帮助我们找到想要联系的人。

1.4 大数据应用改变生活

事实上，人们每天都在体验着大数据应用带来的社会进步积累到了一个引发变革的程度。

线上学习其余内容……



PART 05

1.5

认识大数据分析

1.5.1 大数据分析的定义

大数据是一个含义广泛的术语，是如此庞大而复杂的，需要专门设计的硬件和软件工具进行处理的大数据集。这些数据集收集自各种各样的来源：传感器，气象信息，公开信息如杂志、报纸、文章等等。大数据产生的其他例子包括购买交易记录、网络日志、病历、监控、视频和图像档案以及大型电子商务。



1.5.1 大数据分析的定义

传统批处理数据分析的典型场景是这样的：在整个数据集准备好后，在整体中进行统计抽样。然而，出于理解流式数据的需求，大数据可以从批处理转换成实时处理。这些流式数据、数据集不停地积累，并且以时间顺序排序。由于分析结果有存储期（保质期），流式数据强调及时处理，无论是识别向当前客户继续销售的机会，还是在工业环境中发觉异常情况后需要进行干预以保护设备或保证产品质量，时间都是至关重要的。



1.5.1 大数据分析的定义

在不同行业中，那些专门从事行业数据的搜集、对收集的数据进行整理、对整理的数据进行深度分析，并依据数据分析结果做出行业研究、评估和预测的工作被称为**数据分析**。

所谓**大数据分析**，是指用适当的方法对收集来的大量数据进行分析，提取有用信息和形成结论，从而对数据加以详细研究和概括总结的过程。或者，顾名思义，大数据分析是指**对规模巨大的数据进行分析**。大数据分析是**大数据到信息，再到知识的关键步骤**。大数据分析可以分为四个层次，即描述分析、诊断分析、预测分析和规范分析。如果分析者熟悉行业知识、公司业务及流程，对自己的工作内容有一定的了解，比如熟悉行业认知和公司业务背景，这样的分析结果就会有很大使用价值。

1.5.1 大数据分析的定义

大数据分析结合了传统统计分析方法和计算分析方法，在研究大量数据的过程中寻找模式，相关性和其他有用的信息，帮助企业更好地适应变化并做出更明智的决策。

首先，我们要列出搭建数据分析框架的要求，比如确定分析思路就需要用到营销、管理等**理论知识**；另一方面是针对数据分析结论提出有指导意义的分析建议。能够**掌握数据分析基本原理与一些有效的数据分析方法，并能灵活运用到实践中**，这对于开展数据分析起着至关重要的作用。数据分析方法是理论，而数据分析工具就是实现数据分析方法理论的工具，面对越来越庞大的数据，必须依靠强大的数据分析工具帮我们完成数据分析工作。

1.5.1 大数据分析的定义

- (1) 数据分析可以让人们对数据产生更加优质的诠释，而具有预知意义的分析可以让分析员根据可视化分析和数据分析后的结果做出一些**预测性的推断**。
- (2) **大数据的分析与存储和数据的管理**是一些数据分析层面的最佳实践。通过规范的流程和工具对数据进行分析，可以保证一个预先定义好的高质量的分析结果。
- (3) 不管使用者是数据分析领域中的专家还是普通的用户，作为数据分析工具的数据**可视化**可以直观地展示数据，让数据自己表达，让客户得到理想的结果。
- (4) 只有经过分析的数据才能对用户产生重要的**价值**，所以大数据的分析方式在IT领域显得格外重要，是决定最终信息是否有价值的决定性因素。

1.5.2 四种数据分析方法

数据分析是一个通过处理数据，从中发现一些**深层知识、模式、关系或是趋势**的过程，它的总体目标是做出更好的决策。通过数据分析，可以对分析过的数据建立起关系与模式。

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，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，数据分析学还包括分析方法、科学技术、自动化分析工具等。

1.5.2 四种数据分析方法

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，使之有能力处理大量的来自不同信息源的数据。

不同的行业会以不同的方式使用大数据分析工具和技术，例如：

- 在商业组织中，利用大数据的分析结果能降低运营开销，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。



1.5.2 四种数据分析方法

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4类，即**描述性分析、诊断性分析、预测性分析和规范性分析**（见图2-8）。不同的分析类型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，生成的高质量分析结果将加大分析环境的复杂性和开销。每一种分析方法都对业务分析有很大的帮助，同时也应用在数据分析的各个方面。

1.5.2 四种数据分析方法

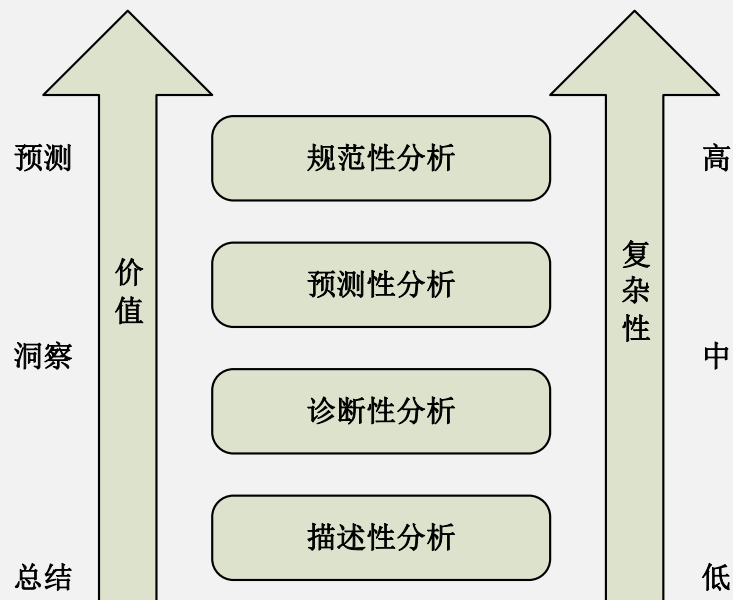
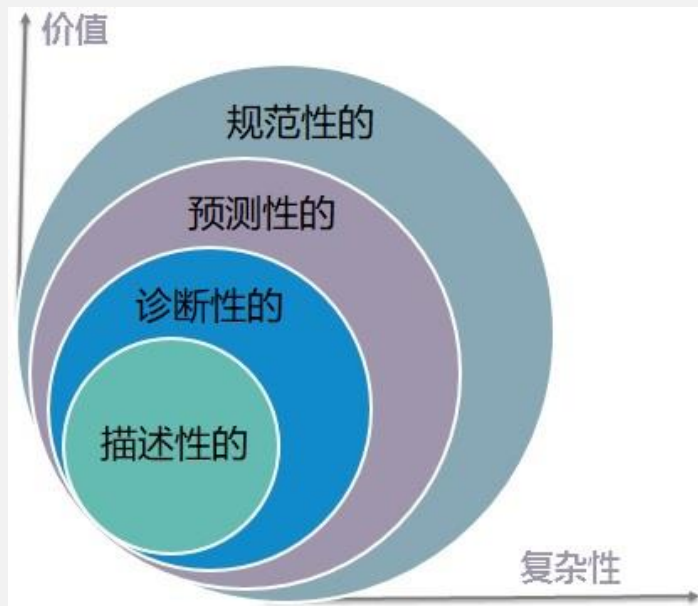


图2-8 四种数据分析方法的价值和复杂性不断提升

一、描述性分析

描述性分析是最常见的分析方法，是探索历史数据并描述发生了什么，是对已经发生的事件进行问答和总结。这一层次包括发现数据规律的聚类、相关规则挖掘、模式发现和描述数据规律的可视化分析，这种方法向数据分析师提供了重要指标和业务的衡量方法。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。例如，每月的营收和损失账

单，分析师可以通过这些账单，获取大量的客户数据。如左图可以明确的看到哪些商品的销售达到了销售量预期。利用可视化工具，能够有效的增强描述型分析所提供的信息。



各产品销售量统计表预警图

一、描述性分析

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。描述性分析提供的价值较低，但也只需要相对基础的训练集。

一、描述性分析

进行描述性分析常常借助OLTP（联机事务处理过程）、CRM（客户关系管理系统）、ERP（企业资源规划系统）等信息系统，经过描述性分析工具的处理生成**即席报表**或者**数据仪表盘**。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CRM或者ERP。

二、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。这类分析通过评估描述性数据，利用诊断分析工具让数据分析师**深入分析数据，钻取数据核心**。其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：

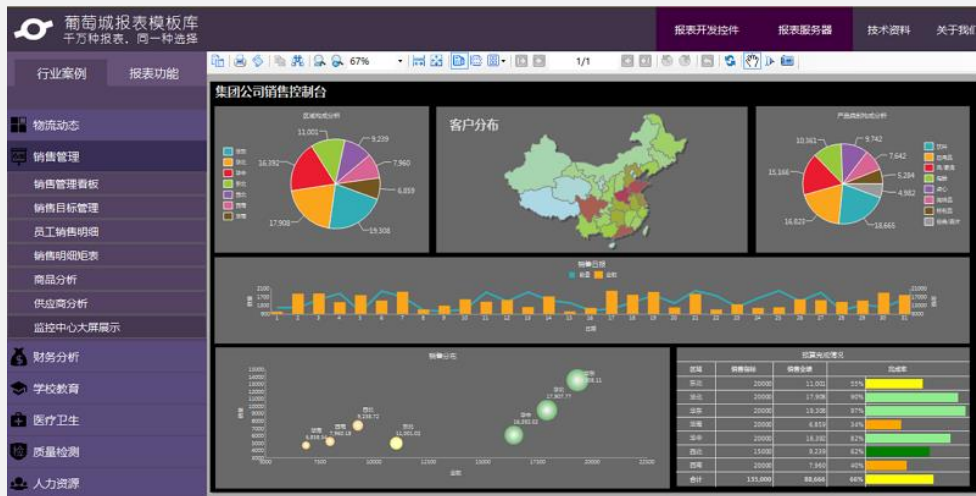
- **为什么**Q2商品比Q1卖得多？
- **为什么**来自东部地区的求助电话比来自西部地区的要多？
- **为什么**最近三个月内病人再入院的比率有所提升？

二、诊断性分析

诊断性分析是基于分析处理系统中的多维数据进行的。与描述性分析相比，诊断性分析的查询处理更加复杂，它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**。诊断性分析常常需要**从不同信息源搜集数据**，并以一种易于进行下钻和上卷分析的结构加以保存。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。

二、诊断性分析

良好设计的BI仪表板能够整合，按照时间序列进行数据读入、特征过滤和钻取数据等功能，以便更好的分析数据。如下图中的“销售控制台”，从图中可以分析出“区域销售构成”、“客户分布情况”、“产品类别构成”和“预算完成情况”等信息。



BI仪表板

三、预测性分析

预测性分析用于预测未来的概率和趋势，例如基于逻辑回归的预测、基于分类器的预测等。预测性分析预测事件未来发生的可能性、预测一个可量化的值，或者是预估事情发生的时间点，这些都可以通过预测模型来完成。通过预测性分析，信息将得到增值，它主要表现在**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过去已经发生的事件的潜在条件是隐式相关的，如果这些潜在的条件改变了，那么用于预测性分析的模型也需要进行更新。

三、预测性分析

预测模型通常会使用各种可变数据来实现预测。数据成员的多样化与预测结果密切相关。在充满不确定性的环境下，预测能够帮助做出更好的决定。预测模型也是很多领域正在使用的重要方法。如下图中的“销售额和销售量”，可以分析出全面的销售量和销售

额基本呈上升趋势，借此可推断下一年的基本销售趋势。



预测基本销售趋势

三、预测性分析

预测性分析提出的问题常常以假设的形式出现，例如：

- 如果消费者错过了一个月的还款，那么他无力偿还贷款的几率有多大？
- 如果以药品B来代替药品A的使用，那么这个病人生存的几率有多大？
- 如果一个消费者购买了商品A和商品B，那么他购买商品C的概率有多大？

预测性分析尝试着基于模式、趋势以及来自于历史数据和当前数据的期望，来预测事件的结果，这将让我们能够分辨风险与机遇。这种类型的分析涉及包含外部数据和内部数据的大数据集以及多种分析方法。与描述性分析和诊断性分析相比，这种分析显得更有价值，同时也要求更加高级的训练集。如图2-12所示，这种工具通常通过提供用户友好的前端接口对潜在的错综复杂的数据进行抽象。

三、预测性分析

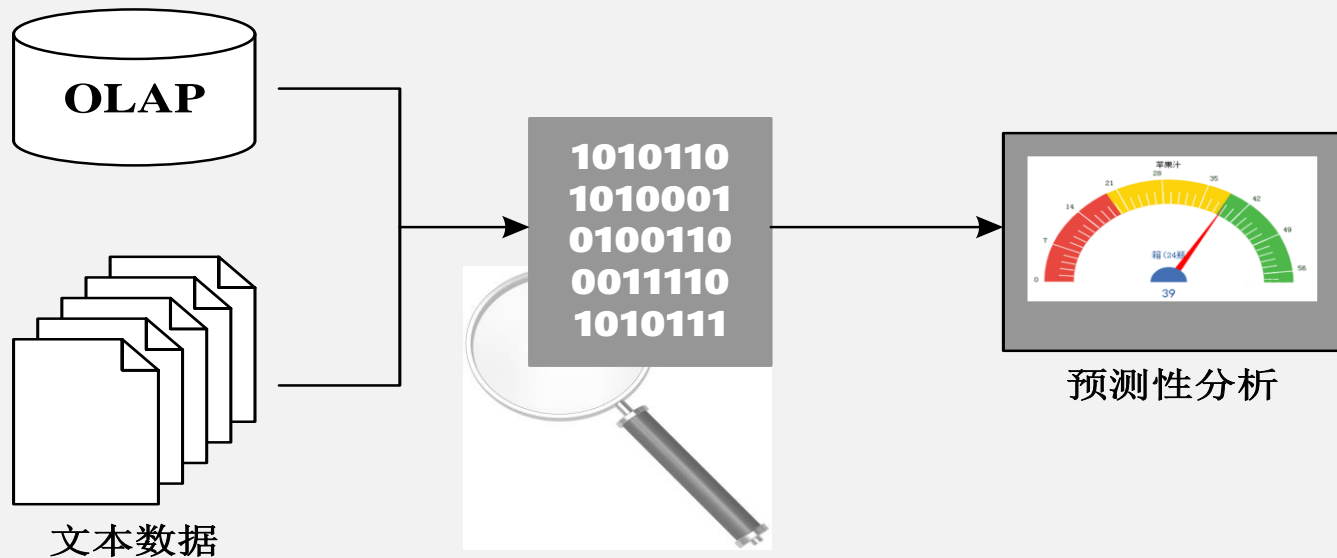


图2-12 预测性分析能够提供用户友好型的前端接口

四、规范性分析

规范性分析建立在预测性分析的结果之上，基于对“发生了什么”、“为什么会发生”和“可能发生什么”的分析，规范需要执行的行动，帮助用户决定应该采取什么措施。规范性分析根据**期望的结果、特定场景、资源以及对过去和当前事件的了解**对未来的决策给出建议，例如基于模拟的复杂系统分析和基于给定约束的优化解生成。规范性分析通常不会单独使用，而是在前面方法都完成之后，最后需要完成的分析方法。它注重的不仅是哪项操作最佳，还包括了其**原因**。换句话说，规范性分析提供了经得起质询的结果，因为它们嵌入了情境理解的元素。因此，这种分析常常用来建立优势或者降低风险。

四、规范性分析

例如，交通规划分析考量了每条路线的距离、每条线路的行驶速度、以及目前的交通管制等方面因素，来帮助选择最好的回家路线。

下面是两个这类问题的样例：

- 这三种药品中，哪一种能提供最好的疗效？
- 何时才是抛售一只股票的最佳时机？

四、规范性分析

规范性分析比其他三种分析的价值都高，同时还要求最高级的训练集，甚至是专门的分析和工具。这种分析将计算大量可能出现的结果，并且推荐出最佳选项。解决方案从解释性的到建议性的均有，同时还能包括各种不同情境的模拟。这种分析能将内部数据与外部数据结合起来。内部数据可能包括当前和过去的销售数据、消费者信息、产品数据和商业规则。外部数据可能包括社交媒体数据、天气情况、政府公文等等。如图2-13所示，规范性分析涉及利用商业规则和大量的内外部数据来模拟事件结果，并且提供最佳的做法。

四、规范性分析

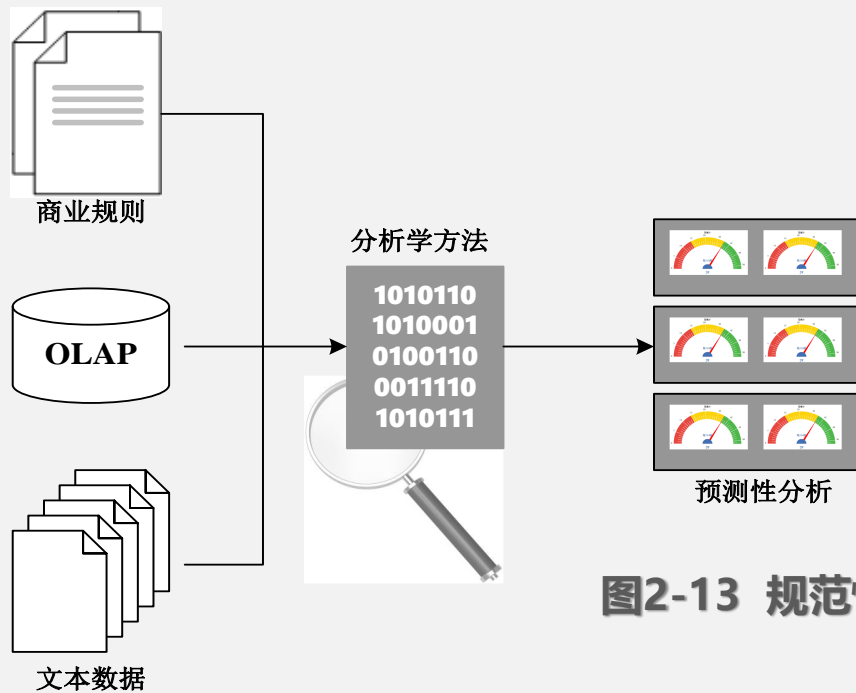


图2-13 规范性分析通过引入商业规则、内部数据以及外部数据来进行深入彻底的分析

补充：定性分析与定量分析

定性分析与定量分析都是一种数据分析技术。其中，**定性分析专注于用语言描述不同数据的质量**。与定量分析相对比，**定性分析涉及分析相对小而深入的样本**。由于样本很小，这些分析结果不能适用于整个数据集，它们也不能测量数值或用于数值比较。例如，冰激凌销量分析可能揭示了五月份销量图不像六月份一样高。分析结果仅仅说明了“不像它一样高”，而并未提供数字偏差。定性分析的结果是描述性的，即用语言对关系的描述。

补充：定性分析与定量分析

定量分析专注于量化从数据中发现的模式和关联。基于统计实践，这项技术涉及分析大量从数据集中所得的观测结果。因为样本容量极大，其结果可以被推广，在整个数据集中都适用。定量分析结果是绝对数值型的，因此可以被用在数值比较上。例如，对于冰激凌销量的定量分析可能发现：温度上升5度，冰激凌销量提升15%。



1.5.3

数据分析的五大
思维方式

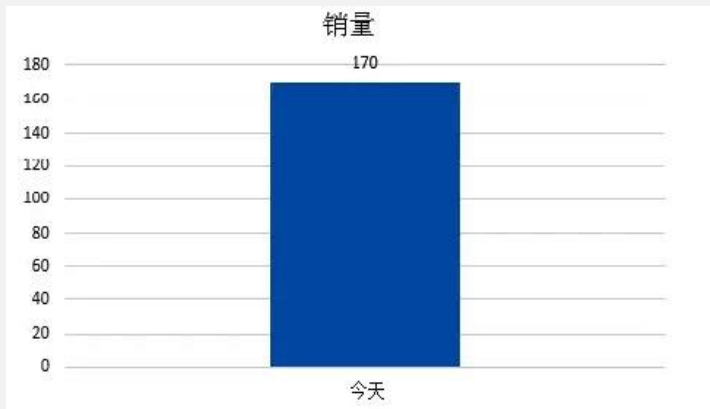
数据可视化的价值在于呈现数据背后的规律，从而帮助使用者提高决策效率与能力。对用户数据的分析是进行可视化系统建设必不可少的一个环节。首先，我们要知道什么叫数据分析。**其实从数据到信息的这个过程就是数据分析**。数据本身并没有什么价值，有价值的是我们从数据中提取出来的信息。其次，我们还要搞清楚数据分析的目的是什么？目的是解决现实中的某个问题或者满足现实中的某个需求。

在这个从数据到信息的过程中，有一些固定的思路，或者称之为**思维方式**。

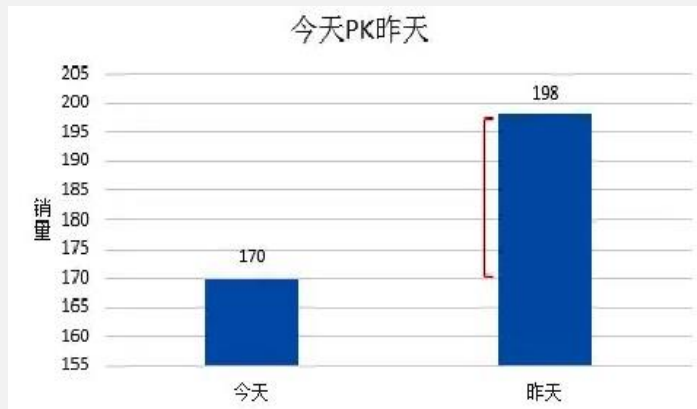
1.5.3

数据分析的五大
思维方式

第一大思维：对照。俗称对比，单独看一个数据是不会有感觉的，必须跟另一个数据做对比才能找到感觉（见下图）。在图中单独看左图无感觉，而右图经过对比就会发现两天的销量实际上差了一大截。



(a)



(b)

1.5.3

数据分析的五大
思维方式

对照是最基本也是最重要的思路，在现实中的应用非常广。比如选款测算、监控店铺数据等，这些过程就是在做“对照”。分析人员拿到数据后，如果数据是独立的，无法进行对比的话，就无法判断，即无法从数据中读取有用的信息。

1.5.3

数据分析的五大
思维方式

第二大思维：拆分。分析这个词的字面理解，就是拆分和解析，可见拆分在数据分析中的重要性。当某个维度可以对比的时候，我们选择对比。在对比后发现问题需要找出原因的时候，或者根本就无法对比的时候，就用到拆分了。

我们来看这样一个场景：运营小美经过对比店铺的数据，发现今天的销售额只有昨天的50%，这个时候，我们再怎么对比销售额这个维度，已经没有意义了。这时需要对销售额这个维度做分解，拆分指标。

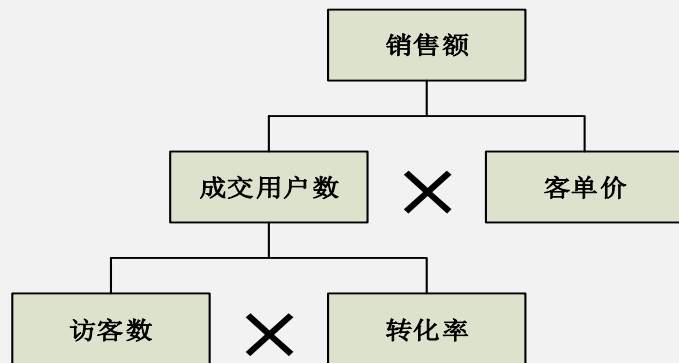
销售额 = 成交用户数 × 客单价

其中成交用户数又等于访客数 × 转化率。

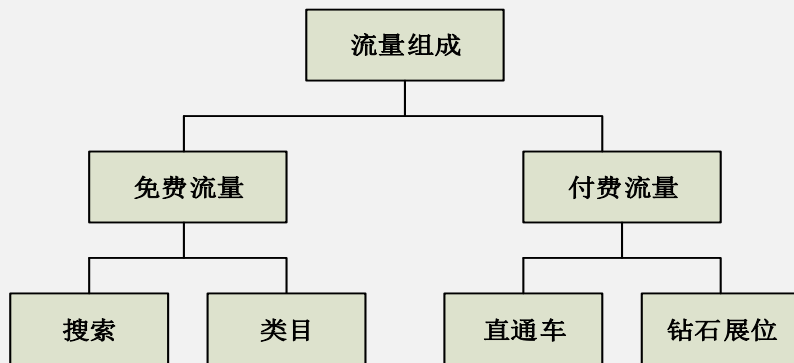
1.5.3

数据分析的五大
思维方式

例如，下图左图是一个指标公式的拆解，右图是对流量的组成成分做的简单分解（还可以分得更细更全）。拆分后的结果相对于拆分前会清晰许多，便于分析查找细节。可见，拆分是分析人员必备的思维之一。



(a)



(b)

1.5.3

数据分析的五大
思维方式

第三大思维：降维。是否有面对一大堆维度的数据却束手无策的经历？当数据维度太多的时候，我们不可能每个维度都拿来分析，可以从一些有关联的指标中筛选出代表的维度即可（见下表 关联指标的维度）。

序号	关键词	搜索人气	搜索指数	占比	点击指数	商城 点击占比	点击率	当前 宝贝数
1	毛呢外套	242,165	1,119,253	58.81%	512,673	30.76%	45.08%	2,448,482
2	毛呢外套（女）	33,285	144,688	7.29%	80,240	48.88%	54.79%	2,448,368
3	韩版毛呢外套	7,460	29,714	1.45%	15,070	21.385%	50.04%	1,035,325
4	小香风毛呢外套	6,400	22,543	1.09%	11,143	22.34%	48.72%	60.258
5	斗篷毛呢外套	5,463	23,443	1.14%	11,328	19.87%	19.87%	108.816

1.5.3

数据分析的五大
思维方式

这么多的维度其实不必每个都分析。我们知道

$$\text{成交用户数} \div \text{访客数} = \text{转化率}$$

当存在这种维度可以通过其他两个维度通过计算转化出来的时候，就可以降维。

成交用户数、访客数和转化率，只要三选二即可。另外，成交用户数*客单价=销售额，这三个也可以三选二。我们一般只关心对我们有用的数据，当有某些维度的数据跟我们的分析无关时，就可以筛选掉，达到“降维”的目的。

1.5.3

数据分析的五大
思维方式

第四大思维：增维。增维和降维是对应的，有降必有增。在当前的维度不能很好地解释我们的问题时，就需要对数据做一个运算，增加多一个指标（见下表 增加指标）。

序号	关键词	搜索人气	搜索指数	占比	点击指数	商城 点击占比	点击率	当前 宝贝数
1	毛呢外套	242,165	1,119,253	58.81%	512,673	30.76%	45.08%	2,448,482
2	毛呢外套（女）	33,285	144,688	7.29%	80,240	48.88%	54.79%	2,448,368
3	韩版毛呢外套	7,460	29,714	1.45%	15,070	21.385%	50.04%	1,035,325
4	小香风毛呢外套	6,400	22,543	1.09%	11,143	22.34%	48.72%	60.258
5	斗篷毛呢外套	5,463	23,443	1.14%	11,328	19.87%	19.87%	108.816

1.5.3

数据分析的五大
思维方式

我们发现一个搜索指数和一个宝贝数，这两个指标一个代表需求，一个代表竞争，有很多人把搜索指数÷宝贝数=倍数，用倍数来代表一个词的竞争度，这种做法就是在增维。增加的维度也称之为“辅助列”。增维和降维是必须对数据的意义有充分的了解后，为了方便我们进行分析，有目的的对数据进行转换运算。

1.5.3

数据分析的五大
思维方式

第五大思维：假说。当我们迷茫的时候，可以应用“假说”。假说是统计学的专业名词，俗称假设。当我们不知道结果，或者有几种选择的时候，那么我们就召唤“假说”，先假设有了结果，然后运用逆向思维。从结果到原因，要有怎么样的因，才能产生这种结果。这有点寻根的味道。那么，我们可以知道，现在满足了多少因，还需要多少因。如果是多选的情况下，就可以通过这种方法来找到最佳路径（决策）。

当然，“假说”的威力不仅仅如此。“假说”可是一匹天马（行空），除了结果可以假设，过程也可以被假设。



PART 06

1.6

大数据分析生命周期

生命周期

1、2

大数据分析生命周期
案例评估

3、4

数据标识
数据获取与过滤

5、6

数据提取
数据特征与清理

7、8

数据聚合与表示
数据分析

9、10

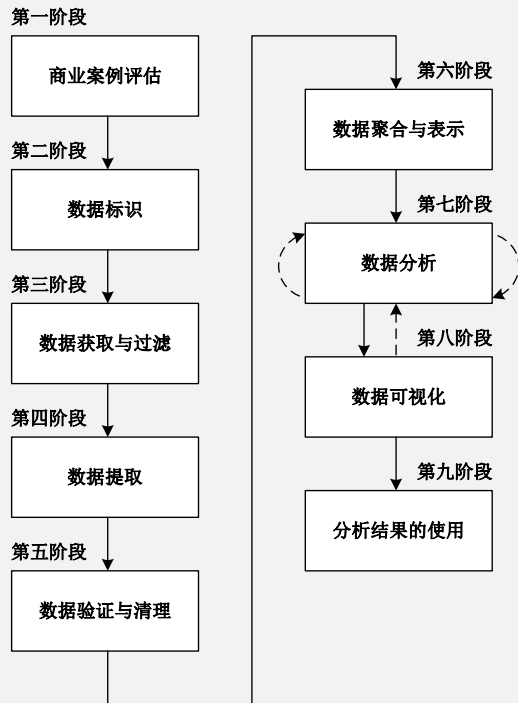
数据可视化
分析结果的使用

1.6.1 大数据分析生命周期

从组织上讲，采用大数据会改变商业分析的途径。大数据分析的生命周期从大数据项目商业案例的创立开始，到保证分析结果部署在组织中并最大化地创造价值时结束。在数据识别、获取、过滤、提取、清理和聚合过程中有许多步骤，这些都是在数据分析之前所必需的。



1.6.1 大数据分析生命周期



由于被处理数据的容量、速率和多样性的特点，大数据分析不同于传统的数据分析。为了处理大数据分析需求的多样性，需要一步步地使用采集、处理、分析和重用数据等方法。大数据分析生命周期可以组织和管理与大数据分析相关的任务和活动。从大数据的采用和规划的角度来看，除了生命周期以外，还必须考虑数据分析团队的培训、教育、工具和人员配备的问题。生命周期的执行需要组织重视培养或者雇佣新的具有相关能力的人。

大数据分析的生命周期可以分为九个阶段。

1.6.2 案例评估

在分析阶段中，每一个大数据分析生命周期都必须起始于一个被很好定义的商业案例，它有着清晰的执行分析的理由、动机和目标，并且应该在着手分析之前就被创建、评估和改进。

商业分析案例的评估能够帮助决策者了解需要使用哪些商业资源，需要面临哪些挑战。另外，在这个环节中详细区分关键绩效指标，能够更好地明确分析结果的评估标准和评估路线。如果关键绩效指标不容易获取，则需要努力使这个分析项目变得**SMART**，即**Specific**（具体的）、**Measurable**（可衡量的）、**Attainable**（可实现的）、**Relevant**（相关的）和**Timely**（及时的）。



1.6.2 案例评估

基于商业案例中记录的商业需求，我们可以确定所定位的商业问题是否是真正的大数据问题。为此，这个商务问题必须直接一个或多个大数据的特点相关。

同样还要注意的，本阶段的另一个结果是确定执行这个分析项目的基本预算。任何如工具、硬件、培训等需要购买的东西都要提前确定，以保证可以对预期投入和最终实现目标所产生的收益进行衡量。比起能够反复使用前期投入的后期迭代，大数据分析生命周期的初始迭代需要在大数据技术、产品和训练上有更多的前期投入。

1.6.3 数据标识

数据标识阶段主要用来标识分析项目所需要的数据集和所需的资源。标识种类众多的数据资源可能会提高找到隐藏模式和相互关系的可能性。例如，为了提供洞察能力，尽可能多地标识出各种类型的相关数据资源非常有用，尤其是当我们探索的目标并不是那么明确的时候。

1.6.3 数据标识

根据分析项目的业务范围和业务问题的性质，我们需要的数据集和它的数据源可能是企业内部和/或企业外部的。在内部数据集的情况下，像是数据集市和操作系统等一系列可供使用的内部资源数据集，往往靠预定义的数据集规范来进行收集和匹配。在外部数据集的情况下，像是数据市场和公开可用的数据集这样的一系列可能的第三方数据集会被收集。一些外部数据的形式则会内嵌到博客和一些基于内容的网站中，这些数据需要通过自动化工具来获取。

1.6.4 数据获取与过滤

在数据获取和过滤阶段，前一阶段标识的数据已经从所有的数据资源中获取，这些数据接下来会被归类并进行自动过滤，以去掉**被污染的数据和对分析对象毫无价值的数据**。根据数据集的类型，数据可能会是档案文件，如购入的第三方数据；可能需要API集成，像是微博、微信上的数据。在许多情况下，我们得到的数据常常是并不相关的数据，特别是外部的非结构化数据，这些数据会在过滤程序中被丢弃。

被定义为“坏”数据的，是其包括遗失或毫无意义的值或是无效的数据类型。但是，被一种分析过程过滤掉的数据集还有可能对于另一种不同类型的分析过程具有价值。因此，**在执行过滤前存储一份原文拷贝是个不错的选择**。为了节省存储空间，可以对原文拷贝进行压缩。

1.6.4 数据获取与过滤

内部数据或外部数据在生成或进入企业边界后都需要继续保存。为了满足批处理分析的要求，数据必须在分析之前存储在磁盘中，而在实时分析之后，数据需要再存储到磁盘中。

元数据会通过自动操作添加到内部和外部的数据资源中来改善分类和查询（见图3-5）。扩充的元数据例子主要包括数据集的大小和结构、资源信息、日期、创建或收集的时间、特定语言的信息等。确保元数据能够被机器读取并传送到数据分析的下一个阶段是至关重要的，它能够帮助我们在大数据分析的生命周期中保留数据的起源信息，保证数据的精确性和高质量。

1.6.4 数据获取与过滤

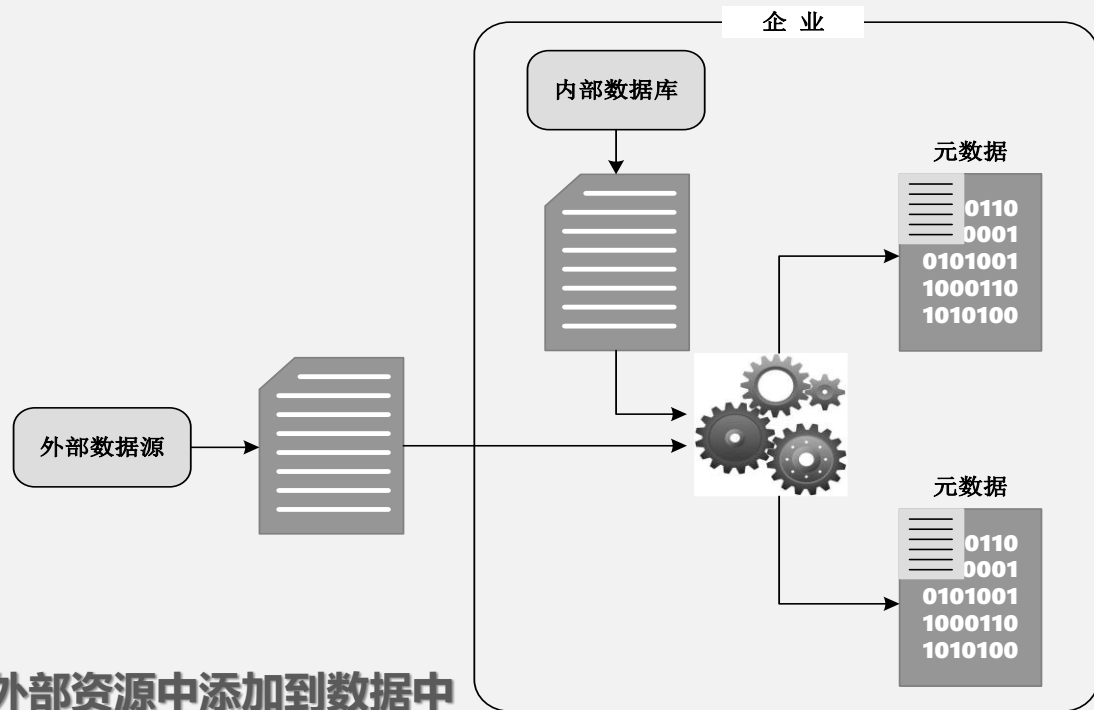


图3-5 元数据从内部资源和外部资源中添加到数据中

1.6.5 数据提取

为分析而输入的一些数据可能会与大数据解决方案产生格式上的不兼容，这样的数据往往来自于外部资源。数据提取阶段主要是要提取不同的数据，并将其转化为大数据解决方案中可用于数据分析的格式。

需要提取和转化的程度取决于分析的类型和大数据解决方案的能力。例如，如果相关的大数据解决方案已经能够直接加工文件，那么从有限的文本数据（如网络服务器日志文件）中提取需要的域，可能就不必要了。类似的，如果大数据解决方案可以直接以本地格式读取文稿的话，对于需要总览整个文稿的文本分析而言，文本的提取过程就会简化许多。

1.6.5 数据提取

下图显示了从没有更多转化需求的XML文档中对注释和内嵌用户ID的提取。

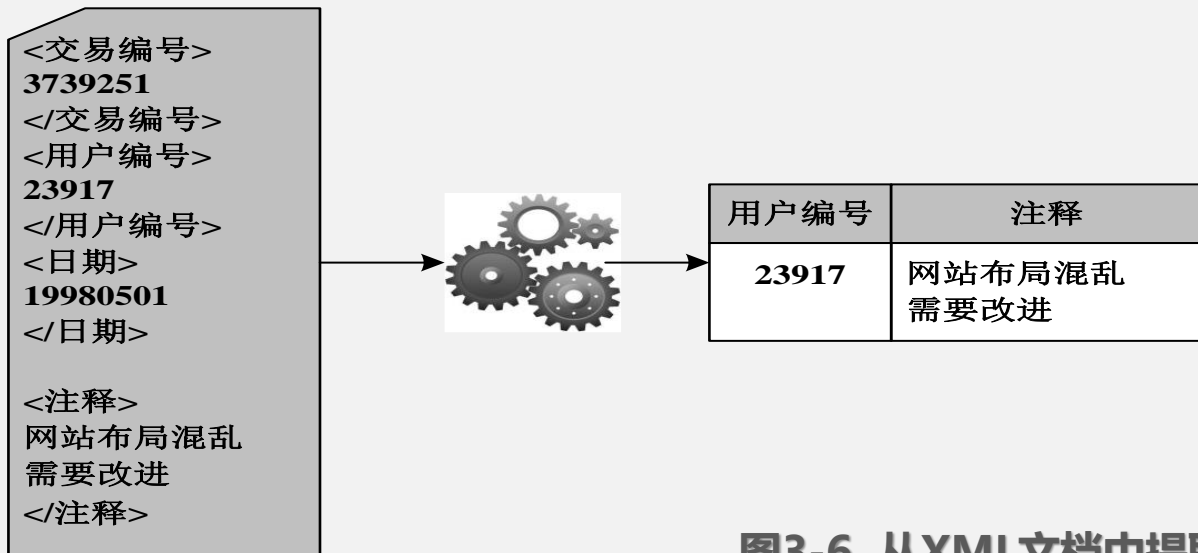


图3-6 从XML文档中提取注释和用户编号

1.6.5 数据提取

图3-7显示了从单个JSON字段中提取用户的经纬度坐标。为了满足大数据解决方案的需求，将数据分为两个不同的域，这就需要做进一步的数据转化。

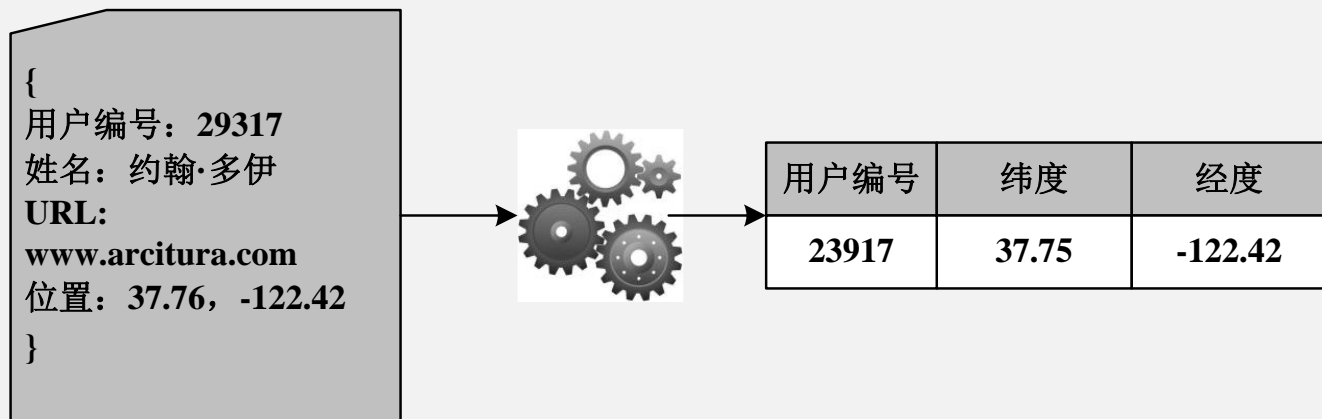


图3-7 从单个JSON文件中提取用户编号和相关信息

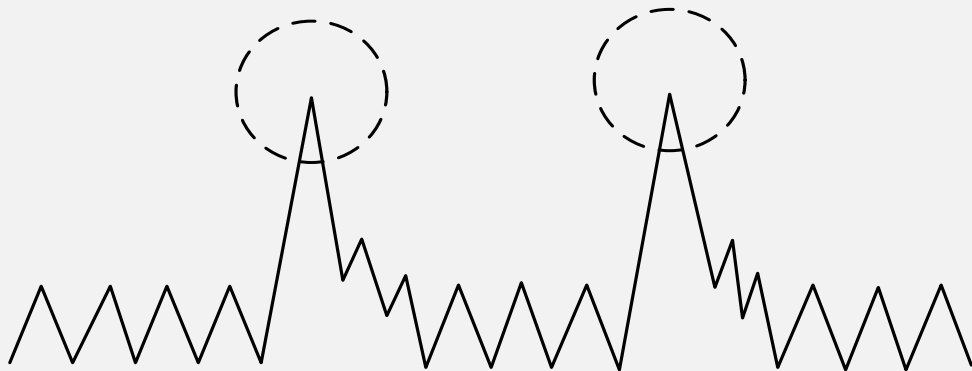
1.6.6 数据验证与清理

无效数据会歪曲和伪造分析的结果。和传统的企业数据那种数据结构被提前定义好、数据也被提前校验的方式不同，大数据分析的数据输入往往没有任何的参考和验证来进行结构化操作，其复杂性会进一步使数据集的验证约束变得困难。

数据验证和清理阶段是为了**整合验证规则并移除已知的无效数据**。大数据经常会从不同的数据集中接收到冗余的数据，这些冗余数据往往会为了整合验证字段、填充无效数据而被用来探索有联系的数据集。数据验证会检验具有内在联系的数据集，填充遗失的有效数据。

1.6.6 数据验证与清理

对于批处理分析，数据验证与抽取可以通过离线ETL（抽取/转换/加载）来执行。对于实时分析，则需要一个更加复杂的在内存中的系统来对从资源中得到的数据进行处理，在确认问题数据的准确性和质量时，来源信息往往扮演着十分重要的角色。有的时候，看起来无效的数据可能在其他隐藏模式和趋势中具有价值，在新的模式中可能有意义。



无效数据的存在造成了一个峰值

1.6.7 数据聚合与表示

数据可以在多个数据集中传播，这要求这些数据通过相同的域被连接在一起，就像日期和ID。在其他情况下，相同的数据域可能会出现在不同的数据集中，如出生日期。无论哪种方式都需要对数据进行核对的方法或者需要确定表示正确值的数据集。

数据聚合和表示阶段是专门为了**将多个数据集进行聚合，从而获得一个统一的视图**。在这个阶段会因为以下情况变得复杂：

- 数据结构——数据格式相同时，数据模型可能不同。
- 语义——在两个不同的数据集中，具有不同标记的值可能表示同样的内容，比如“姓”和“姓氏”。

1.6.7 数据聚合与表示

由大数据解决方案进行标准化的数据结构可以作为一个标准的共同特征被用于一系列的分析技术和项目。这可能需要建立一个像非结构化数据库一样的中央标准分析仓库。

Id	Name

数据集A

+

Id	DOB

数据集B

=

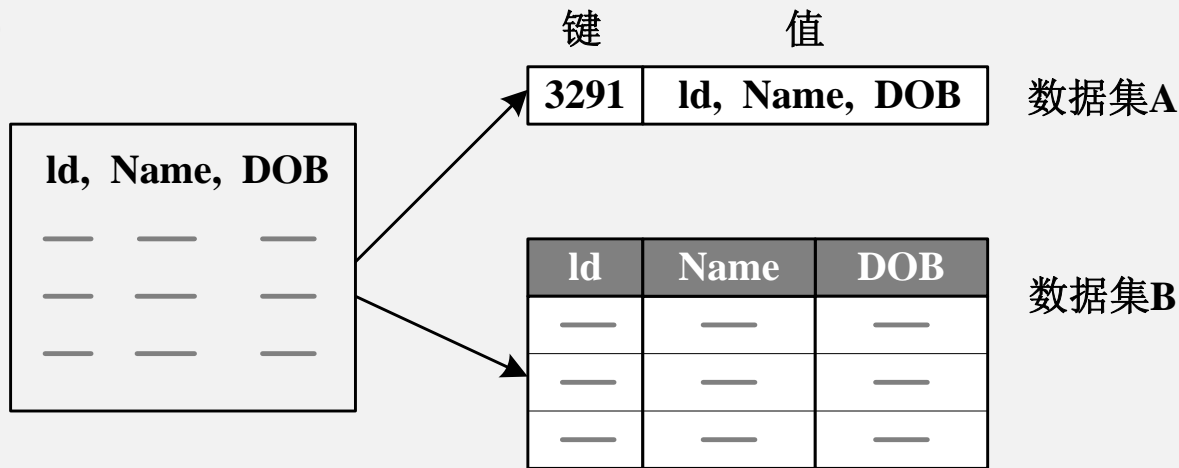
Id	Name	DOB

数据集C

使用ID域聚集两个数据域的简单例子

1.6.7 数据聚合与表示

下图展示了存储在两种不同格式中的相同数据块。数据集A包含所需的数据块，但是由于它是BLOB的一部分而不容易访问。数据集B包含有相同的以列为基础来存存储的数据块，使得每个字段都被单独查询到。数据集A和B能通过大数据解决方案结合起来创建一个标准化的数据结构。



1.6.8 数据分析

数据分析阶段致力于**执行实际的分析任务**，通常会涉及一种或多种类型的数据分析。在这个阶段，数据可以自然迭代，尤其在数据分析是探索性分析的情况下，分析过程会一直重复，直到适当的模式或者相关性被发现。

根据所需的分析结果的类型，这个阶段可以被尽可能地简化为查询数据集以实现用于比较的聚合。另一方面，它可以像结合数据挖掘和复杂统计分析技术来发现各种模式和异常，或是生成一个统计或是数学模型来描述变量关系一样具有挑战性。

1.6.8 数据分析

数据分析可以分为**验证分析**和**探索分析**两类，后者常常与数据挖掘相联系。

验证性数据分析是一种演绎方法，即先提出被调查现象的原因，被提出的原因或者假说称为一个假设。接下来使用数据分析以验证和反驳这个假设，并为这些具体的问题提供明确的答案。我们常常会使用数据采样技术，意料之外的发现或异常经常会被忽略，因为预定的原因是一个假设。

探索性数据分析是一种与数据挖掘紧密结合的归纳法。在这个过程中没有假想的或是预定的假设产生。相反，数据会通过分析探索来发展一种对于现象起因的理解。尽管它可能无法提供明确的答案，但这种方法会提供一个大致方向以便发现模式或异常。

1.6.9 数据可视化

如果只有分析师才能解释数据分析结果的话，那么分析海量数据并发现有用的见解的能力就没有什么价值了。数据可视化阶段致力于使用数据可视化技术和工具，并通过图形表示有效的分析结果。为了从分析中获取价值并在随后拥有从向下一阶段提供反馈的能力，商务用户必须充分理解数据分析的结果。



1.6.9 数据可视化

完成数据可视化阶段得到的结果能够为用户提供执行可视化分析的能力，这能够让用户去发现一些未曾预估到的问题的答案。相同的结果可能会以许多不同的方式来呈现，这会影响最终结果的解释。因此，重要的是保证商务域在相应环境中使用最合适的可视化技术。

另一个必须要记住的方面是：为了让用户了解最终的积累或者汇总结果是如何产生的，提供一种相对简单的统计方法也是至关重要的。

1.6.10 分析结果的使用

大数据分析结果可以用来为商业使用者提供商业决策支持，像是使用图表之类的工具，可以为使用者提供更多使用这些分析结果的机会。在分析结果的使用阶段，致力于确定如何以及在哪里处理分析数据能保证产出更大的价值。

基于要解决的分析问题本身的性质，分析结果很可能会产生对被分析的数据内部一些模式和关系有着新的看法的“模型”。这个模型可能看起来会比较像一些数据公式和规则的集合，它们可以用来改进商业进程的逻辑和应用系统的逻辑，也可以作为新的系统或者软件的基础。

1.6.10 分析结果的使用

在这个阶段常常会被探索的领域主要有以下几种：

- **企业系统的输入**——数据分析的结果可以自动或者手动输入到企业系统中，用来改进系统的行为模式。例如，在线商店可以通过处理用户关系分析结果来改进产品推荐方式。新的模型可以在现有的企业系统或是在新系统的基础上改善操作逻辑。
- **商务进程优化**——在数据分析过程中识别出的模式、关系和异常能够用来改善商务进程。例如作为供应链的一部分整合运输线路。模型也有机会能够改善商务流程逻辑。
- **警报**——数据分析的结果可以作为现有警报的输入或者是新警报的基础。例如，可以创建通过电子邮件或者短信的警报来提醒用户采取纠正措施。

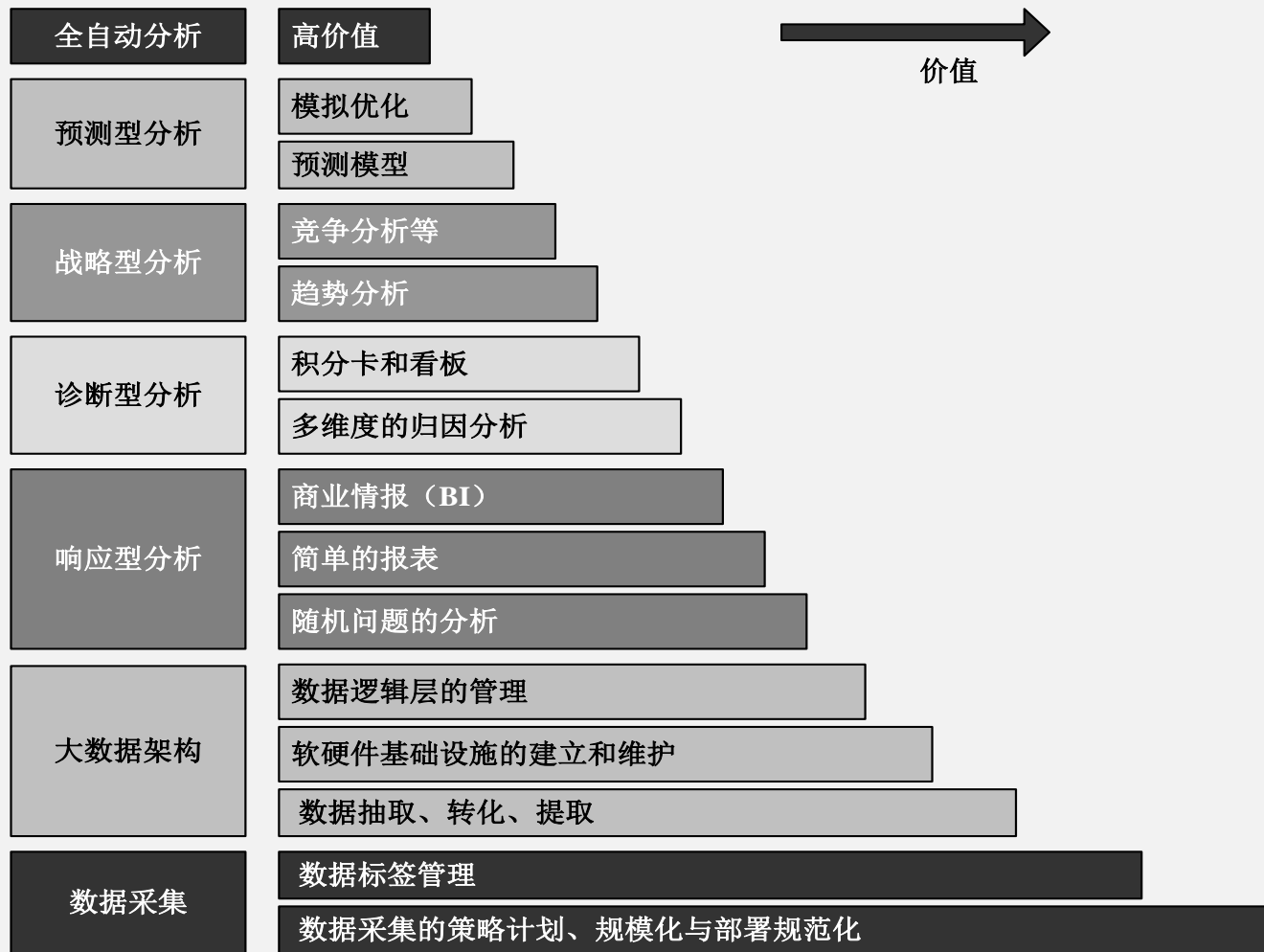


图4-5 数据分析框架



华北电力大学
NORTH CHINA ELECTRIC POWER UNIVERSITY

第1章 大数据分析基础