第5章

工业大数据

本章学习目标

- 了解大数据的概念
- 了解工业大数据的概念
- 了解工业大数据处理过程
- 了解工业大数据治理相关概念

5.1 大数据概述

1. 什么是大数据

信息技术的快速发展，引发了数据规模的爆炸式增长，大数据引起了国内外学术界、工业界和政府部门的高度重视，被认为是一种新的非物质生产要素，蕴含巨大的经济和社会价值，并将导致科学研究的深刻变革，对国家的经济发展、社会发展、科学进展具有战略性、全局性和长远性的意义。

大数据(Big Data)指无法在一定时间范围内使用常规软件工具进行捕捉、管理和处理的数据集合。相对于传统的数据分析，大数据是海量数据的集合，它以采集、整理、清洗、存储、挖掘、共享、分析、应用为核心，正广泛地应用在军事、金融、工业、农业、教育、环境保护、通信等各个行业中。

人类利用数据的历史非常悠久，最早可以追溯到数字发明时期，不同文明均掌握了利用数字记录和管理生产生活的能力。从文明之初的“结绳记事”，到文字发明后的“文以载道”，再到近现代科学的“数据建模”，数据一直伴随着人类社会的发展变迁，承载了人类基于数据和信息认识世界的努力和取得的巨大进步。纵观人类利用数据的历史，虽然数据的本质没有变化，但是在制度、技术和经济发展的交织作用下，数据完成了从数字到资产的转变，在这个过程中，数据的规模、价值和影响不断扩大。当前，数据在社会发展中正扮演着愈发重要的作用。从早期仅限于学术研究、军事领域，到后面应用到企业经营活动，再到个人互联网应用，直到云与物联网时代。数据作为一种经济资源和生产要素，是人工智能等新兴技术发展的动力，没有海量的数据积累和应用场景，人工智能很难冲破瓶颈快速发展。

2. 大数据的特点

随着对大数据认识的不断加深，人们认为大数据一般具有4个特征：数据量大、数据类型

繁多、数据产生速度快以及数据价值密度低。

1）数据量大

大数据中的数据量大，就是指海量数据。由于大数据往往是采取全样分析，因此大数据的“大”首先体现在其规模和容量远远超出传统数据的测量尺度。一般的软件工具难以捕捉、存储、管理和分析的数据，通过大数据的云存储技术都能保存下来，形成浩瀚的数据海洋，目前的数据规模已经从 TB 级升级至 PB 级。大数据之“大”还表现在其采集范围和内容的丰富多变，能存入数据库的不仅包含各种具有规律性的数据符号，还囊括了各种如图片、视频、声音等非规则的数据。

2）数据类型繁多

据国际信息技术咨询企业国际数据公司（International Data Corporation，IDC）的调查报告，拜互联网和通信技术近年来迅猛发展所赐，如今的数据类型早已不是单一的文本形式，除了网络日志、音频、视频、图片、地理位置信息等多类型的数据对数据的处理能力提出了更高的要求，数据来源也越来越多样，不仅产生于组织内部运作的各个环节，也来自组织外部的开放数据。其中，内部数据主要包含政府数据（如征信、户籍、犯罪记录等）、企业数据（如阿里巴巴的消费数据、腾讯的社交数据、滴滴的出行数据等）、机构数据（如第三方咨询机构的调查数据）；而开放数据主要包含网站数据和各种 App 终端数据，以及大众媒介数据等。

例如，苹果公司在 iPhone 手机上应用的一项语音控制功能 Siri 就是多样化数据处理的代表。用户可以通过语音、文字输入等方式与 Siri 对话交流，并调用手机自带的各项应用，读短信，询问天气，设置闹钟，安排日程，乃至搜索餐厅、电影院等生活信息，收看相关评论，甚至直接订位、订票，Siri 则会依据用户默认的家庭地址或所在位置判断、过滤搜索的结果。利用射频识别、二维码、智能传感器等感知设备感知获取物体的各类信息。例如，物联网上部署的每个传感器都是一个信息源，不同类别的传感器所捕获的信息内容和信息格式不同。因此，人们可以在物联网上部署海量的多种类型传感器，传感器按一定的频率周期性地采集环境信息，并不断更新数据。

3）数据产生速度快

在数据处理速度方面，有一个著名的“一秒定律”，即要在秒级时间范围内给出分析结果，超出这个时间，数据就失去价值了。大数据是一种以实时数据处理、实时结果导向为特征的解决方案，它的“快”体现在以下两个层面。

（1）数据产生得快。有的数据是爆发式地产生，如欧洲核子研究中心的大型强子对撞机在工作状态下每秒产生 PB 级的数据；有的数据是涓涓细流式地产生，但是由于用户众多，短时间内产生的数据量依然非常庞大，如点击流、日志、论坛、博客、发邮件、射频识别数据、GPS（全球定位系统）位置信息。

（2）数据处理得快。正如水处理系统可以从水库调出水进行处理，也可以处理直接对涌进来的新水流，大数据也有批处理（“静止数据”转换为“正使用数据”）和流处理（“动态数据”转换为“正使用数据”）两种范式，以实现快速的数据处理。

4）数据价值密度低

随着互联网以及物联网的广泛应用，信息感知无处不在，但现实世界所产生的数据中，有价值的数据占比很小。因此，如何结合业务逻辑并通过强大的机器算法挖掘数据价值，是大数据时代最需要解决的问题。以视频为例，一部一小时的视频，在连续不间断监控过程中，可能有用的数据只有一两秒。但是，为了能够得到想要的视频，人们不得不投入大量资金用于购买网络设备、监控设备等。

在大数据时代，由于数据采集得不及时、数据样本不全面、数据可能不连续等，数据可能会失真，但当数据量达到一定规模时，可以通过更多的数据获得更真实全面的反馈。相比于传统的小数据，大数据最大的价值在于从大量不相关的各种类型数据中挖掘出对未来趋势与模式预测分析有价值的数据，并通过机器学习、人工智能或数据挖掘方法深度分析，发现新规律和新知识，并运用于农业、金融、医疗等各个领域，从而最终达到改善社会治理、提高生产效率、推进科学研究的效果。

大数据处理技术在具体的应用方面，可以为国家支柱企业的数据分析和处理提供技术和平台支持，为企业进行数据分析、处理、挖掘，提取出重要的信息和知识，再转化为有用的模型，应用到研究、生产、运营和销售过程中。同时，国家大力倡导“智慧城市”建设，在城市化与信息化融合等背景下，围绕改善民生、增强企业竞争力、促进城市可持续发展等关注点，综合利用物联网、云计算等信息技术手段，结合城市现有信息化基础，融合先进的城市运营服务理念，建立广泛覆盖和深度互联的城市信息网络，对城市的资源、环境、基础设施、产业等多方面要素进行全面感知，并整合构建协同共享的城市信息平台，对信息进行智能处理利用，从而为城市运行和资源配置提供智能响应控制，为政府社会管理和公共服务提供智能决策依据及手段，为企业和个人提供智能信息资源及开放式信息应用平台的综合性区域信息化发展过程。

3. 大数据技术

大数据带来的不仅是机遇，同时也是挑战。传统的数据处理手段已经无法满足大数据的海量实时需求，需要采用新一代的信息技术应对大数据的爆发。人们把大数据技术归纳为以下几类。

1）数据采集

大数据的应用离不开数据采集。数据采集又称为数据获取，是指利用某些装置，从系统外部采集数据并输入系统内部的一个接口。在互联网行业快速发展的今天，数据采集已经被广泛应用于互联网及分布式领域，如摄像头、麦克风以及各类传感器等都是数据采集工具。

数据采集技术是数据处理的必备条件，首先需要有数据采集的手段，只有先把信息收集上来，之后才能应用上层的数据处理技术。数据采集除了各类传感设备等硬件软件设施之外，主要涉及数据的 ETL（采集、转换、加载）过程，能对数据进行清洗、过滤、校验、转换等各种预处理，将有效的数据转换为适合的格式和类型。同时，为了支持多源异构的数据采集和存储访问，还需要设计企业的数据总线，方便企业各个应用和服务之间数据的交换和共享。

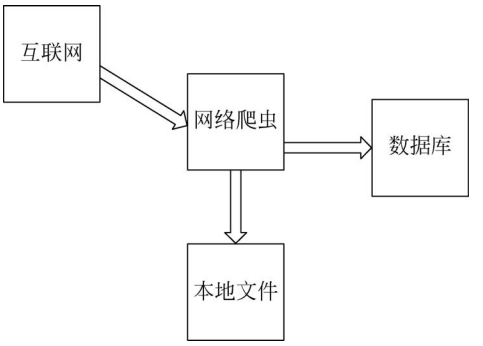

图 5-1　网络爬虫采集流程

区别于小数据采集，大数据采集不再仅仅使用问卷调查、信息系统的数据库取得结构化数据，大数据的来源有很多，主要包括使用网络爬虫获取的网页文本数据、使用日志收集器收集的日志数据、从关系型数据库中获得的数据和由传感器收集到的时空数据等，而对于一些图像和语音数据，则需要高端技术才能使其转换为普通的大数据分析师需要的数据。图 5-1 所示为网络爬虫采集流程。

网络爬虫是一种按照一定的规则，自动地抓取 Web 信息的程序或脚本。网络爬虫可以自动采集所有其能够访问到的页面内容，为搜索引擎和大数据分析提供数据来源。在大数据时代，网络爬虫更是从互联网上采集数据的有力工具。网络爬虫按照系统结构和实现技术，大致可以分为通用网络爬虫（General Purpose Web

Crawler)、聚焦网络爬虫(Focused Web Crawler)、增量式网络爬虫(Incremental Web Crawler)、深层网络爬虫(Deep Web Crawler)。实际的网络爬虫系统通常是几种爬虫技术相结合实现的。

2) 数据存储

如今大数据的火热,带来的第1道障碍就是关于大数据存储的问题。大数据因为规模大、类型多样、新增速度快,所以在存储和计算上都需要技术支持,依靠传统的数据存储和处理工具,已经很难实现高效处理了。

以往的数据存储,主要是基于关系数据库,而关系数据库在面对大数据时,存储设备所能承受的数据量是有上限的,当数据规模达到一定的量级之后,数据检索的速度就会急剧下降,对于后续的数据处理也带来了困难。为了解决这个问题,主流的数据库系统纷纷给出解决方案,如MySQL提供了MySQL proxy组件,实现了对请求的拦截,结合分布式存储技术,从而可以将一张很大的表中的记录拆分到不同的节点上进行查询,对于每个节点,数据量不会很大,从而提高了查询效率。但是实际上,这样的方式没有从根本上解决问题。

目前常见的大数据存储方式主要有分布式存储、NoSQL数据库和云数据库3种。

(1) 分布式存储。分布式存储是相对于集中式存储来说的。在分布式存储出现之前,企业级的存储设备都是集中式存储。所谓集中式存储,从概念上可以看出是具有集中性的,也就是整个存储是集中在一个系统中的。但集中式存储并不是一个单独的设备,是集中在一套系统中的多个设备。在这个存储系统中包含很多组件,除了核心的机头(控制器)、磁盘阵列和交换机等设备外,还有管理设备等辅助设备。

分布式存储最早由Google提出,其目的是通过廉价的服务器提供适用于大规模、高并发场景的Web访问问题。与常见的集中式存储技术不同,分布式存储技术并不是将数据存储在某个或多个特定的节点上,而是通过网络使用企业中的每台机器上的磁盘空间,并将这些分散的存储资源构成一个虚拟的存储设备,数据分散地存储在企业的各个角落。分布式存储目前多借鉴Google的经验,在众多的服务器搭建一个分布式文件系统,再在这个分布式文件系统上实现相关的数据存储业务。图5-2所示为使用Hadoop实现分布式存储。

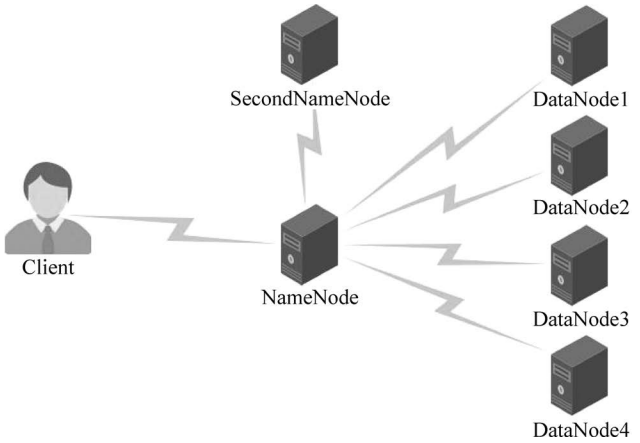

图5-2　使用Hadoop实现分布式存储

(2) NoSQL数据库。传统的关系型数据库采用关系模型作为数据的组织方式,但是随着对数据存储要求的不断提高,在大数据存储中,之前常用的关系型数据库已经无法满足Web 2.0的需求,主要表现为无法满足海量数据的管理需求、无法满足数据高并发的需求、高可扩展性和高可用性的功能太低。在这种情况下,NoSQL数据库应运而生。

NoSQL 数据库又叫作非关系数据库，和数据库管理系统（RDBMS）相比，NoSQL 数据库不使用结构化查询语言（Structured Query Language，SQL）作为查询语言，其存储也可以不需要固定的表模式，用户操作 NoSQL 时通常会避免使用 RDBMS 的 JOIN 操作。NoSQL 数据库一般都具备水平可扩展的特性，并且可以支持超大规模数据存储，灵活的数据模型也可以很好地支持 Web 2.0 应用，还具有强大的横向扩展能力。典型的 NoSQL 数据库种类有键值数据库、列族数据库、文档数据库和图形数据库。值得注意的是，每种类型的数据库都能够解决传统关系数据库无法解决的问题。图 5-3 所示为 Redis 在 Windows 下的运行界面。

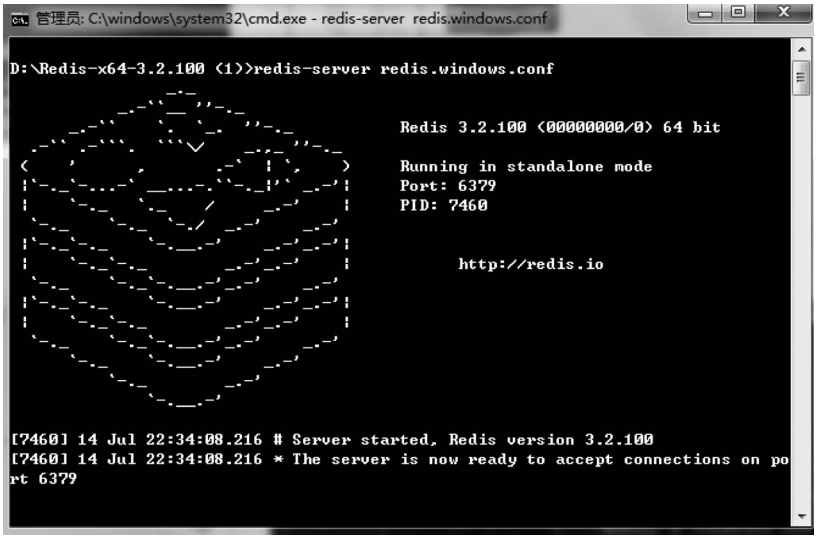

图 5-3　Redis 在 Windows 下的运行界面

Redis 是使用 C 语言开发的一个高性能键值数据库，该数据库可以通过一些键值类型存储数据。Redis 的性能十分优越，可以支持每秒十几万次的读/写操作，其性能超数据库，并且支持集群、分布式、主从同步等配置，还支持一定事务能力。Redis 的出色之处不仅仅是性能，它最大的魅力是支持保存多种数据结构。不过 Redis 的主要缺点是数据库容量受到物理内存的限制，不能用作海量数据的高性能读写，因此 Redis 适合的场景主要局限在较小数据量的高性能操作和运算上。

（3）云数据库。云数据库是指被优化或部署到一个虚拟计算环境中的数据库，是在云计算的大背景下发展起来的一种新兴的共享基础架构的方法，它极大地增强了数据库的存储能力，消除了人员、硬件、软件的重复配置，让软、硬件升级变得更加容易。因此，云数据库具有高可扩展性、高可用性、采用多组形式和支持资源有效分发等特点，可以实现按需付费和按需扩展。

从数据模型的角度来说，云数据库并非一种全新的数据库技术，如云数据库并没有专属于自己的数据模型，它所采用的数据模型可以是关系数据库所使用的关系模型，也可以是 NoSQL 数据库所使用的非关系模型。并且，针对不同的企业，云数据库可以提供不同的服务，如云数据库既可以满足大企业的海量数据存储需求，也可以满足中小企业的低成本数据存储需求，还可以满足企业动态变化的数据存储需求。

3）数据清洗

由于大数据中有更大可能包含各种类型的数据质量问题，这些数据质量问题为大数据的应用带来了困扰，甚至灾难性后果。因此，在大数据分析与应用中，数据清洗是最重要的步骤之一。

在大数据时代，数据清洗通常是指把“脏数据”彻底洗掉。所谓“脏数据”，是指不完整、不

规范、不准确的数据，只有通过数据清洗才能从根本上提高数据质量。数据清洗是发现并纠正数据文件中可识别错误的一道程序，该步骤针对数据审查过程中发现的明显错误值、缺失值、异常值、可疑数据，选用适当方法进行清理，使“脏”数据变为“干净”数据，有利于后续的统计分析得出可靠的结论。当然，数据清洗还包括对重复记录进行删除以及检查数据一致性等。

在数据清洗定义中包含两个重要的概念：原始数据和干净数据。

(1) 原始数据是来自数据源的数据，一般作为数据清洗的输入数据。由于原始数据的来源纷杂，因此不适合直接进行分析。值得注意的是，对于未清洗的数据集，无论尝试什么类型的算法，都无法获得准确的结果。

(2) 干净数据也称为目标数据，即符合数据仓库或上层应用逻辑规格的数据，也是数据清洗过程的结果数据。

因此，数据清洗的目的主要有两个：第一是通过清洗让数据可用；第二是让数据变得更适合进行后续的分析工作。据统计，在大数据项目的实际开发工作中，数据清洗通常占开发过程总时间的 50%～70%。

图 5-4 所示为数据清洗中的异常值检测，图 5-5 所示为检查数据缺失值。

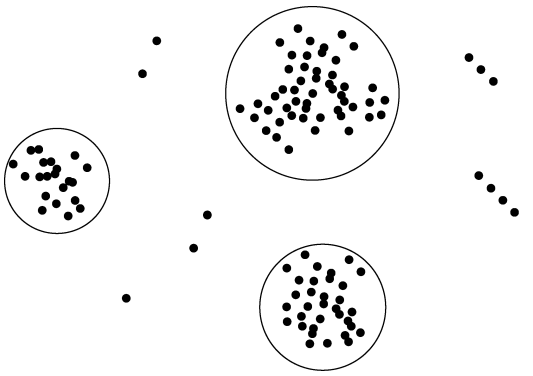

图 5-4　数据清洗中的异常值检测

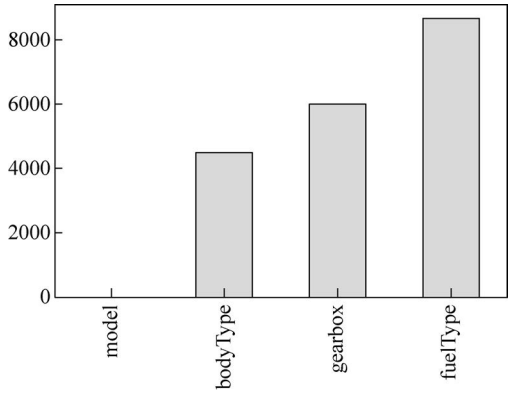

图 5-5　检查数据缺失值

4) 数据计算

面向大数据处理的数据查询、统计、分析、挖掘等需求，产生了大数据计算的不同计算模式，整体上人们把大数据计算分为离线批处理计算、实时交互计算和流计算 3 种。

(1) 离线批处理计算。随着云计算技术的广泛应用和发展，基于开源的 Hadoop 分布式存储系统和 MapReduce 数据处理模式的分析系统也得到了广泛的应用。Hadoop 通过数据分块及自恢复机制，能支持 PB 级的分布式数据存储，以及基于 MapReduce 分布式处理模式对这些数据进行分析和处理。MapReduce 编程模型可以很容易地将多个通用批数据处理任务和操作在大规模集群上并行化，而且有自动化的故障转移功能。MapReduce 编程模型在 Hadoop 这样的开源软件的带动下被广泛采用，应用到 Web 搜索、欺诈检测等各种各样的实际应用中。除了 MapReduce 计算模型之外，以 Swift 为代表的工作流计算模式、以 Pregel 为代表的图计算模式，也都可以处理包含大规模计算任务的应用流程和图算法。Swift 系统作为科学工作流和并行计算之间的桥梁，是一个面向大规模科学和工程工作流的快速、可靠的定义、执行和管理的并行化编程工具。

Hadoop 是一个能够对大量数据进行分布式处理的软件框架，而且是以一种可靠、高效、可伸缩的方式进行处理。通过不断增加廉价的商用服务器提高计算和存储能力，用户可以轻

松地在上面开发和运行处理海量数据的应用程序。

(2) 实时交互计算。当今的实时计算一般都需要针对海量数据进行,除了要满足非实时计算的一些需求(如计算结果准确)以外,实时计算最重要的一个需求是能够实时响应计算结果,一般要求为秒级。实时和交互式计算技术中,Google 的 Dremel 系统表现最为突出。Dremel 是 Google 的交互式数据分析系统,可以组建成规模上千的集群,处理 PB 级别的数据。作为 MapReduce 的发起人,Google 开发了 Dremel 系统,将处理时间缩短到秒级,成为 MapReduce 的有力补充。Dremel 作为 Google Big Query 的 Report 引擎,获得了很大的成功。Spark 是由加州大学伯克利分校 AMP 实验室开发的实时数据分析系统,采用一种与 Hadoop 相似的开源集群计算环境,但是 Spark 在任务调度、工作负载优化方面设计和表现更加优越。Spark 启用了内存分布数据集,除了能够提供交互式查询,还可以优化迭代工作负载。Spark 是在 Scala 语言中实现的,它将 Scala 语言用作其应用程序框架,Spark 和 Scala 语言能够紧密集成,其中的 Scala 可以像操作本地集合对象一样轻松地操作分布式数据集。创建 Spark 可以支持分布式数据集上的迭代作业,是对 Hadoop 的有效补充,支持对数据的快速统计分析。此外,Spark 也可以在 Hadoop 文件系统中并行运行。

(3) 流计算。传统的流式计算系统一般是基于事件机制,所处理的数据量也不大。新型的流处理技术,如 Yahoo 的 S4,主要解决的是高数据率和大数据量的流式处理。S4 是一个通用的、分布式的、可扩展的、部分容错的可插拔平台,开发者可以很容易地在其上开发面向无界不间断流数据处理的应用。Storm 是 Twitter 开源的一个类似于 Hadoop 的实时数据处理框架,这种高可拓展性、能处理高频数据和大规模数据的实时流计算解决方案将应用于实时搜索、高频交易和社交网络上。Storm 可以用来并行处理密集查询,其拓扑结构是一个等待调用信息的分布函数,当它收到一条调用信息后,会对查询进行计算,并返回查询结果。

5) 数据分析与挖掘

数据分析是指用适当的统计分析方法对收集来的大量数据进行分析,为提取有用信息和形成结论而对数据加以详细研究和概括总结的过程。随着大数据时代的来临,大数据分析也应运而生。一般来讲,大数据分析通常是指对规模巨大的数据进行分析,其目的是提取海量数据中的有价值内容,找出内在的规律,从而帮助人们作出最正确的决策。

大数据分析主要有描述性统计分析、探索性数据分析以及验证性数据分析等多种类型。

(1) 描述性统计分析。描述性统计是指运用制表和分类、图形以及计算概括性数据描述数据特征的各项活动。描述性统计分析要对调查总体所有变量的有关数据进行统计性描述,主要包括数据的频数分析、集中趋势分析、离散程度分析、分布以及一些基本的统计图形。

(2) 探索性数据分析。探索性数据分析是指为了形成值得假设的检验而对数据进行分析的一种方法,是对传统统计学假设检验手段的补充。它是对已有的数据(特别是调查或观察得来的原始数据)在尽量少的先验假定下进行探索,通过作图、制表、方程拟合、计算特征量等手段探索数据的结构和规律的一种数据分析方法。特别是在大数据时代,人们面对各种杂乱的“脏数据”,往往不知所措,不知道从哪里开始了解目前拿到手上的数据时,探索性数据分析就非常有效。

从逻辑推理上讲,探索性数据分析属于归纳法,有别于从理论出发的演绎法。因此,探索性数据分析成为大数据分析中不可缺少的一步并且走向前台。

(3) 验证性数据分析。验证性数据分析注重对数据模型和研究假设的验证,侧重于已有假设的证实或证伪。假设检验是根据数据样本所提供的证据,肯定或否定有关总体的声明,一般包含以下流程。

① 提出零假设，以及对应的备择假设。

② 在零假设前提下，推断样本统计量出现的概率（统计量可符合不同的分布，对应不同的概率分布有不同的检验方法）。

③ 设定拒绝零假设的阈值，样本统计量在零假设下出现的概率小于阈值，则拒绝零假设，承认备择假设。

数据挖掘（Data Mining）是指通过大量数据集进行分类的自动化过程，通过数据分析识别趋势和模式，建立关系，解决业务问题。换句话说，数据挖掘是从大量的、不完全的、有噪声的、模糊的、随机的数据中提取隐含在其中的、人们事先不知道的但又是潜在有用的信息和知识的过程。

数据挖掘的基本流程可以总结为以下几个阶段：商业理解、数据理解、数据准备、数据建模、模型评估和模型部署应用，如图 5-6 所示。

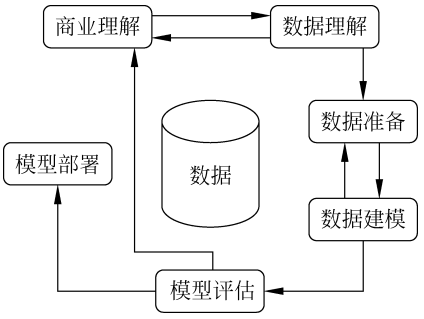

图 5-6　数据挖掘的基本流程

（1）商业理解。商业理解主要是明确业务需求，并根据业务背景进行资源评估，最后确定业务的具体目标。

（2）数据理解（数据探索）。数据理解是对建模分析数据进行先导性的洞察分析，利用绘制图表、计算某些特征量等手段，对样本数据集的结构特征和分布特性进行分析的过程。该步骤有助于选择合适的数据预处理和数据分析技术，它是数据建模的依据。例如，数据探索发现数据稀疏，建模时则选择对稀疏数据支持相对较好的分析方案；如果数据包含文本数据，建模时则需要考虑基于自然语言处理相关技术等。

（3）数据准备（数据预处理）。数据准备是将不规整的业务数据整理为相对规整的建模数据，如数据缺失处理、异常值检测处理等操作。数据的质量决定了模型输出的结果，即数据决定了模型的上限，所以人们需要花大量的时间对数据进行处理。在数据预处理阶段，如果数据存在缺失值情况而导致建模过程混乱甚至无法进行建模，则需要进行缺失值处理，缺失值处理分为删除存在缺失值的记录、对可能值进行插补和不处理 3 种情况；如果建模数据存在数据不均衡情况，则需要考虑数据平衡处理。如果建模分析数据存在量纲、数量级上的差别，则需要进行数据规约处理消除量纲数量级的影响；如果异常数据对分析结果影响巨大，则需要进行异常值检测处理排除影响。

理论上，数据和特征决定了模型的上限，而算法只是逼近这个上限而已，这里的数据指的是经过特征工程得到的数据，因此特征工程是人们进行机器学习必须重视的过程。特征工程的目的是最大限度地从原始数据中提取特征以供算法和模型使用。一般认为特征工程包括特征选择、特征规约、特征生成 3 部分。其中，特征选择在降低模型复杂度、提高模型训练效率、增强模型的准确度方面影响较大；在建模字段繁多的情况下，通过特征规约降低建模数据维度，降低特征共线特性对模型准确度的不利影响，从而提升模型的训练效率；特征生成是在特征维度信息相对单一的情况下为了提升模型准确性能而采取的维度信息扩充的方法体系。

（4）数据建模。数据挖掘的核心阶段是基于既定的数据和分析目标选择适宜的算法模型进行建模训练和迭代优化。数据建模涉及的技术包括机器学习、统计分析、深度学习，相关技术之间没有一个明显的区分界线，且功能互补。值得注意的是，深度学习领域涉及多种模型框架和操作使用技巧，其本身可以作为机器学习的特例，同样适用于机器学习的多个应用场景。深度学习作为一种实现机器学习的技术，往往在数据量大、业务数据指标难以人工提取的情形下发挥着举足轻重的作用，它在图像处理、语音识别、自然语言处理等领域具有其他机器学习

算法无法企及的准确性能。

(5) 模型评估。模型评估是评估所构建的模型是否符合既定的业务目标,有助于发现表达数据的最佳模型和所选模式将来工作的性能如何。模型评估秉承的准则是在满足业务分析目标的前提下优先选择简单化的模型。每个分析场景可以基于多种算法构建多个模型,也可以依据模型优化的方法体系进行模型训练优化,而如何在训练得到的多个模型中选择最优模型,可以选择性能度量作为指标体系,进而基于一定的评估方法择优选择。

(6) 模型部署应用。模型部署及应用是将数据挖掘结果作用于业务过程,即将训练得到的最优模型部署到实际应用中;模型部署后,可使用调度脚本控制数据挖掘模型实现流程化运行。在模型日常运行过程中,可根据实际需求检查模型运行结果是否满足前端业务的实际应用,跟踪模型运行情况定期进行模型结果分析,并适时进行模型优化。

6) 数据可视化

数据可视化在大数据技术中同样至关重要,因为数据最终需要为人们所使用,为生产、运营、规划提供决策支持。选择恰当的、生动直观的展示方式能够帮助人们更好地理解数据及其内涵和关联关系,也能够更有效地解释和运用数据,发挥其价值。在展现方式上,除了传统的报表、图形之外,人们还可以结合现代化的可视化工具及人机交互手段,甚至增强现实技术等实现数据与现实的无缝接口。

与传统的立体建模等特殊技术方法相比,数据可视化所涵盖的技术方法要广泛得多,它是以计算机图形学及图像处理技术为基础,将数据转换为图形或图像形式显示到屏幕上,并进行交互处理的理论、方法和技术。数据可视化涉及计算机视觉、图像处理、计算机辅助设计、计算机图形学等多个领域,并逐渐成为一项研究数据表示、数据综合处理、决策分析等问题的综合技术。

值得注意的是,由于对海量的数据作出有意义的理解非常困难,而许多大数据集中又包含了有价值的数据,因此数据可视化已成为决策者的重要方法。为了利用所有这些数据,许多企业认识到数据可视化的价值在于清晰有效地理解重要信息,使决策者能够理解困难的概念,识别新的模式,并获得数据驱动的洞察力,以便做出更好的决定。确定呈现数据集的最佳方式,并遵循数据可视化最佳实践,对于图形设计人员在创建这些视觉效果时非常重要。特别是在处理非常大的数据集时,开发有张力的表达方式,对于创建既有用又具有视觉吸引力的可视化至关重要。图 5-7 所示为数据可视化中的柱状图表。

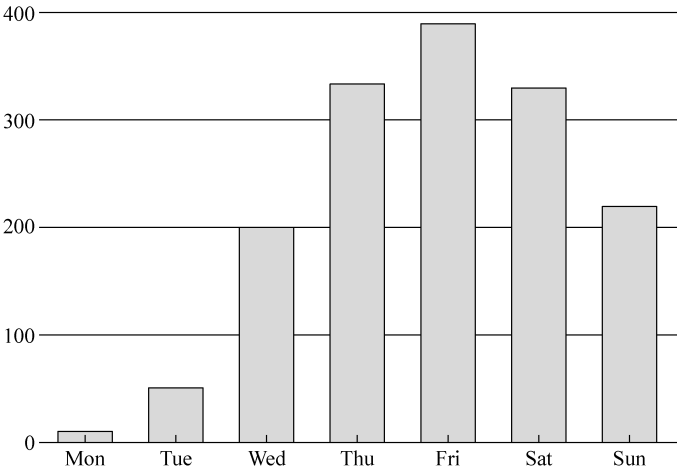

图 5-7 数据可视化中的柱状图表

数据可视化越来越普及，在工业物联网、电信、智慧医疗、智能交通、现代农业等多个行业都有广泛的应用。

（1）金融可视化。在当今互联网金融激烈的竞争下，市场形势瞬息万变，金融行业面临诸多挑战。通过引入数据可视化可以对企业各地日常业务动态实时掌控，对客户数量和借贷金额等数据进行有效监管，帮助企业实现数据实时监控，加强对市场的监督和管理；通过对核心数据多维度的分析和对比，指导公司科学调整运营策略，确定发展方向，不断提高公司风控管理能力和竞争力。例如，金融数据可视化大屏和智能可视化图表直观地跟踪流动性、预算、支出、现金流量和许多其他财务指标；同时，通过实时利用财务数据全面地概述财务信息，帮助企业避免货币危机。

（2）工业可视化。数据可视化在工业生产中有着重要的应用，特别是在智慧工厂中。智慧工厂是在工业一系列科学管理实践的基础上，深度融合自动化技术、信息通信技术和智能科学技术，结合数据、信息和知识建立更具核心竞争力的新一代制造业企业及其生态系统。要实现智慧工厂，将海量工厂数据进行采集、展示与分析的可视化技术不可或缺。

例如，在工业生产中使用可视化平台，与工厂内原有的自动控制系统（主流工控系统）相结合，通过虚拟现实和数据仪表盘等多种展现手段，为大数据时代的工业生产监控和虚拟制造应用，提供效果最优异的可视化解决方案。

又如，在智慧工厂中对厂房和车间进行三维可视化展示，标识内部工作区域，可以帮助工作和管理人员对生产环境有一个全面的了解。工人通过数据展示终端可以实时了解各生产线情况，平台支持异常告警并及时推送告警信息，帮助操作人员迅速应对，保障顺利生产。

（3）农业可视化。近年来，农业物联网、无线网络传输等技术的蓬勃发展，极大地推动了监测数据的海量爆发，农业跨步迈入大数据时代。数据可视化在农业生产中也得到了极大的应用。

例如，在智慧农业中可以利用物联网设备监控农产品生长过程，将数据信息公开透明地展示给消费者，让消费者买得放心、吃得安心。

又如，将可视化技术应用在视频直播、休闲农业的发展当中，这些更加灵活、亲民的应用方式，不但可以给原有的业务增添新的亮点，而且能够让可视农业的新概念更快地得到普及，而这样的一种正向互动，也会为可视农业的发展提供长期动力。

（4）医疗可视化。数据可视化可以帮助医院将之前分散、凌乱的数据加以整合，构建全新的医疗管理体系模型，帮助医院领导快速解决关注的问题，如一些门诊数据、用药数据、疾病数据等。此外，大数据可视化还可以应用于诊断医学以及一些外科手术中的精确建模，通过三维图像的建立帮助医生确定是否进行外科手术或进行何种手术。不仅如此，数据可视化还可以加快临床疾病预防、流行疾病防控等的预测和分析能力。

例如，在医院实施医疗数据可视化系统，在有效展示数据的同时，让数据表达的内容更容易被理解，也能保证信息的有效传递，使医院的医疗信息从简单的医疗业务数据采集与存储发展到对医疗业务数据的共享和交换，并逐步向医疗业务数据的分析与挖掘方向延伸。

（5）教育可视化。在我国对教育科研越来越重视的情况下，可视化教学也逐渐替代传统的教学模式。可视化教学是指在计算机软件和多媒体资料的帮助下，将被感知、被认知、被想象、被推理的事物及其发展变化的形式和过程用仿真化、模拟化、形象化及现实化的方式在教学过程中尽量表现出来。在可视化教学中，知识可视化能帮助学生更好地获取、存储、重组知识，并能将知识迁移应用，促进多元思维的养成，帮助学生更好地关注知识本身的联系和对本质的探求，减少由于教学方式带来的信息损耗，提高有效认知负荷。

（6）交通可视化。城市交通与每个人的生活息息相关，也给现代化城市带来了巨大的难

题，如交通拥堵、空气污染等经济环境问题。随着大数据技术的不断发展，可视分析技术在交通数据分析的过程中扮演着十分重要的角色。

7）数据治理

数据为人类社会带来机遇的同时也带来了风险，围绕数据产权、数据安全和隐私保护的问题也日益突出，并催生了一个全新的命题——数据治理。

数据治理是指从使用零散数据转变为使用统一数据、从具有很少或没有组织流程到企业范围内的综合数据管控、从数据混乱状况到数据井井有条的一个过程。数据治理强调的是一个从混乱到有序的过程。从范围来讲，数据治理涵盖了从前端业务系统、后端业务数据库再到业务终端的数据分析，从源头到终端再回到源头，形成一个闭环负反馈系统。从目的来讲，数据治理就是要对数据的获取、处理和使用进行监督管理。

在数据治理中既包含了企业各种前端数据的输入（企业交易数据、运营数据等），也包含了三方数据（通信数据、客户数据等），甚至还包含了各种采集数据（社交数据、传感数据、图像数据等）。在实施数据治理后，能够为企业带来新的数据价值。随着大数据在各个行业领域应用的不断深入，数据作为基础性战略资源的地位日益凸显，数据标准化、数据确权、数据质量、数据安全、隐私保护、数据流通管控、数据共享开放等问题越来越受到国家、行业、企业各个层面的高度关注，这些内容都属于数据治理的范畴。因此，数据治理的概念就越来越多地受到关注，成为目前大数据产业生态系统中的新热点。

一般来说，数据治理主要包括以下 3 部分工作。

（1）定义数据资产的具体职责和决策权，应用角色分配决策需要执行的确切任务的决策和规范活动。

（2）为数据管理实践制定企业范围的原则、标准、规则和策略。数据的一致性、可信性和准确性对于确保增值决策至关重要。

（3）建立必要的流程，以提供对数据的连续监视和控制实践，并帮助在不同组织职能部门之间执行与数据相关的决策，以及业务用户类别。

因此，数据治理能够有效帮助企业利用数据建立全面的评估体系，实现业务增长；通过数据优化产品，提高运营效率，真正实现数据系统赋能业务系统，提升以客户为中心的数字化体验能力，实现生意的增长。

目前常见的数据治理涉及的领域主要包括数据资产、数据模型、元数据与元数据管理、数据标准、主数据管理、数据质量管理、数据管理生命周期、数据存储、数据交换、数据集成、数据安全、数据服务、数据价值、数据开发和数据仓库。在数据治理时，各领域需要有机结合，如数据标准、元数据、数据质量等几个领域相互协同和依赖。例如，通过数据标准的管理，可以提升数据合法性、合规性，进一步提升数据质量，减少数据生产问题；在元数据管理的基础上，可进行数据生命周期管理，有效控制在线数据规模，提高生产数据访问效率，减少系统资源浪费；通过元数据和数据模型管理，将表、文件等数据资源按主题进行分类，可明确当事人、产品、协议等相关数据的主数据源归属、数据分布情况，有效实施数据分布的规划和治理。因此，数据治理领域是随着企业业务发展而不断变化的，领域之间的关系也需要不断深入挖掘和分布，最终形成一个相互协同与验证的领域网，全方位地提升数据治理成效。

图 5-8 所示为《信息技术服务　治理》（GB/T 34960—2018）系统国家标准的数据治理框架。该数据治理框架比较符合我国企业和政府的组织现状，更加全面和精练地描述了数据治理的工作内容，包含顶层设计、数据治理环境、数据治理域和数据治理过程。

图 5-9 所示为数据治理中的企业业务架构。业务架构是企业治理结构、商业能力与价值

流的正式蓝图，并将企业的业务战略转化为日常运作的渠道。业务架构定义了企业的治理架构（组织结构）、业务能力、业务流程以及业务数据。其中，业务能力定义企业做什么，而业务流程定义企业该怎么做。此外，在具体实施中业务架构还包括企业业务的运营模式、流程体系、组织结构、地域分布等内容，并体现企业大到板块、小到最细粒度的流程环节之间的所有业务逻辑。

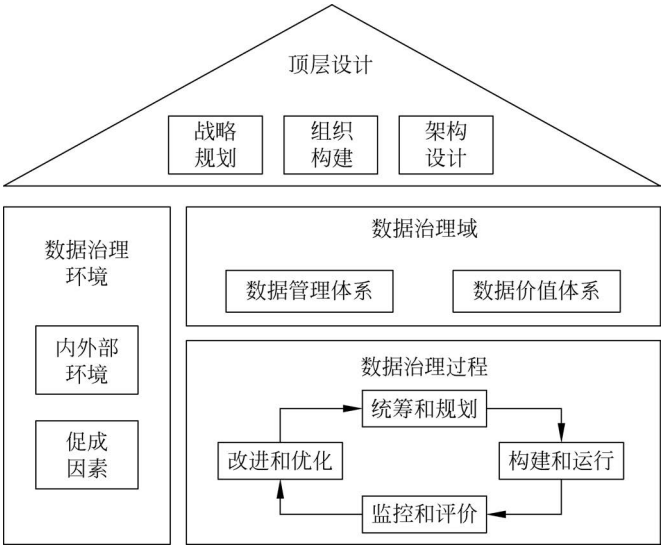

图 5-8 《信息技术服务　治理》(GB/T 34960—2018)系统国家标准的数据治理框架

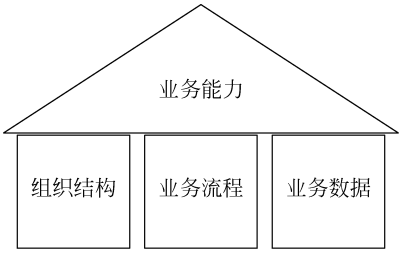

图 5-9　数据治理中的企业业务架构

5.2　工业大数据及其应用

5.2.1　工业大数据概述

1. 认识工业大数据

社会经济的快速发展，信息化和工业化技术不断发展创新，智能制造在工业领域引起了新一轮的工业革命。随着智能制造的发展以及互联网技术的发展，工业大数据作为贯穿整个产品生产的新的要素，在一定程度上推动了智能制造的升级。大数据时代的来临，对工业制造的变革、发展起到了重要的作用。

工业大数据即难以通过传统的分析工具进行有效分析的工业数据的集合，具备明显的大数据的容量大、数据类型多、数据价值高、数据更新快的特性。利用大数据技术有效对工业大数据进行分析，深入挖掘其中的数据价值，才能创造出新的商业价值。通过工业大数据，可以以全方位、数字化的视角对工业的发展进行剖析，对结构化、非结构化的数据进行有效分析，从而建立相应的数据模型，使企业实现智能化的生产制造。

工业大数据是指在工业领域中，围绕典型智能制造模式，从客户需求到销售、订单、计划、研发、设计、工艺、制造、采购、供应、库存、发货和交付、售后服务、运维、报废或回收再制造等整个产品全生命周期各个环节所产生的各类数据及相关技术和应用的总称。工业大数据是工业数据的总称，包括企业信息化数据、工业物联网数据以及外部跨界数据，是工业互联网的核心要素。因此，发展工业大数据，包括工业大数据的理论、技术、产品和保障条件，对于促进工业互联网的蓬勃发展具有重要的价值和意义。

总体来看，工业大数据推动互联网由以服务个人用户消费为主转向以服务生产性应用为主，由此导致产业模式、制造模式和商业模式的重塑。大数据与智能机床、机器人、3D 打印等技术结合，推动了柔性制造、智能制造和网络制造的发展。工业大数据与智能物流、电子商务的联动，进一步加速了工业企业销售模式的变革，如精准营销配送、精准广告推送等。

1）工业大数据的数据来源

工业大数据的数据来源主要有以下 3 类。

第 1 类是生产经营相关业务数据，主要来自传统企业信息化范围，被收集存储在企业信息系统内部，包括传统工业设计和制造类软件、企业资源计划（ERP）、产品生命周期管理（Product Lifecycle Management，PLM）、供应链管理（SCM）、客户关系管理（CRM）和环境管理系统（Environmental Management System，EMS）等。通过这些企业信息系统已累积大量的产品研发数据、生产性数据、经营性数据、客户信息数据、物流供应数据及环境数据，这类数据是工业领域传统的数据资产，在移动互联网等新技术应用环境中正在逐步扩大范围。

第 2 类是设备物联数据，主要指工业生产设备和目标产品在物联网运行模式下实时产生收集的涵盖操作和运行情况、工况状态、环境参数等体现设备和产品运行状态的数据。这类数据是工业大数据新的、增长最快的来源。狭义的工业大数据即指这类数据，即工业设备和产品快速产生的并且存在时间序列差异的大量数据。

第 3 类是外部数据，指与工业企业生产活动和产品相关的企业外部互联网来源数据，如评价企业环境绩效的环境法规、预测产品市场的宏观社会经济数据等。

值得注意的是，近年来，由人产生的数据的比重正逐步降低，企业信息化和工业物联网中机器产生的海量时序数据是工业数据规模变大的主要来源，机器数据的比重将越来越大。

2）工业大数据特征

一般意义上，大数据具有数据量大、数据种类多、商业价值高、处理速度高的特点，在此基础上，工业大数据还有三大特点。

一是多模态。工业大数据形态多样，特别是非结构化数据，这是由工业生产社会化的属性所决定的。生产环节复杂、产业链跨度长、上下游发展程度不均衡、各参与主体任务属性特征差异巨大等因素，导致了数据的多样组织、表达、定义和呈现共同构成多模态特性。

二是实时性强。工业大数据重要的应用场景是实时监测、实时预警、实时控制。在工业生产中，每时每刻都在产生大量数据，如生产机床的转速和能耗、食品加工的温湿度、火力发电机组的燃烧和燃煤消耗、汽车的装备数据、物流车队的位置和速度等。尤其是自工业从社会生产中独立成为一个门类以来，工业生产的数据采集、使用范围就逐步加大。特别是随着信息、电子、数字技术以及传感器、物联网等的发展，一批智能化、高精度、长续航、高性价比的微型传感器面世，以物联网为代表的新一代网络技术在移动数据通信的支持下，能做到任何时间、任何地点采集和传输数据。

三是强关联。工业大数据具有强关联的特点，这个特点尤其重要。工业现场的数据在语义层有复杂的显性和隐性强关联，不同物理变量之间的关系，既有工业机理方面，也有统计分

析方面，不能孤立、局部、片面地看待，否则满足不了工业对于严格性、可靠性和安全性方面的要求。值得注意的是，工业领域行业 Know-How（机理模型）是工业生产的核心。工业机理模型（Model Based）是根据对象、生产过程的内部机制或物质流的传递机理建立起来的精确数学模型。工业机理模型来自工业生产设备，包括飞机、汽车、冶金制造过程的零件模板，以及设备故障诊断、性能优化和远程运维等背后的原理、知识、经验和方法；来自业务流程逻辑，包括 ERP、供应链管理、客户关系管理、生产效能优化等这些业务系统中蕴含着的流程逻辑框架；来自研发工具以及生产工艺中的工艺配方、工艺流程、工艺参数等模型。近代工业无论是信息技术的引入还是自动控制的革新都紧紧围绕工业机理模型在进行。

2. 工业大数据与大数据的关系

工业大数据应用是基于工业数据，运用先进的大数据相关思维、工具、方法，贯穿于工业的设计、工艺、生产、管理、服务等各个环节，使工业系统、工业产品具备描述、诊断、预测、决策、控制等智能化功能模式和结果。工业领域的数据累积到一定量级，超出了传统技术的处理能力，就需要借助大数据技术、方法提升处理能力和效率，大数据技术为工业大数据提供了技术和管理的支撑。

首先，工业大数据可以借鉴大数据的分析流程及技术，实现工业数据采集、处理、存储、分析、可视化。例如，大数据技术应用在工业大数据的集成与存储环节，支撑实现高实时性采集、大数据量存储及快速检索；大数据处理技术的分布式高性能计算能力为海量数据的查询检索、算法处理提供性能保障等。其次，工业制造过程中需要高质量的工业大数据，可以借鉴大数据的治理机制对工业数据资产进行有效治理。图 5-10 所示为使用大数据技术对工业生产进行优化。

图 5-10　使用大数据技术对工业生产进行优化

3. 工业大数据与智能制造

工业大数据是智能制造的关键技术，主要作用是打通物理世界和信息世界，推动生产型制造向服务型制造转型。工业大数据在智能制造中有着广泛的应用前景，在产品市场需求获取、产品研发、制造、运行、服务直至报废回收的产品全生命周期过程中，工业大数据在智能化设计、智能化生产、网络化协同制造、智能化服务、个性化定制等场景都发挥出巨大的作用。图 5-11 所示为工业大数据在智能制造中的应用。

在智能化设计中，通过对产品数据分析，实现自动化设计和数字化仿真优化；在智能化生产过程中，工业大数据技术可以实现在生产制造中的应用，如人机智能交互、工业机器人、制造工艺的仿真优化、数字化控制、状态监测等，提高生产故障预测准确率，综合优化生产效率；在网络化协同制造中，工业大数据技术可以实现智能管理的应用，如产品全生命周期管理、客户关系管理、供应链管理、产供销一体等，通过设备联网与智能控制，达到过程协同与透明化；在智能化服务中，工业大数据通过对产品运行及使用数据的采集、分析和优化，可实现产品智能化及远程维修。同时，工业大数据可以实现智能检测监管的应用，如危险化学品、食品、印染、

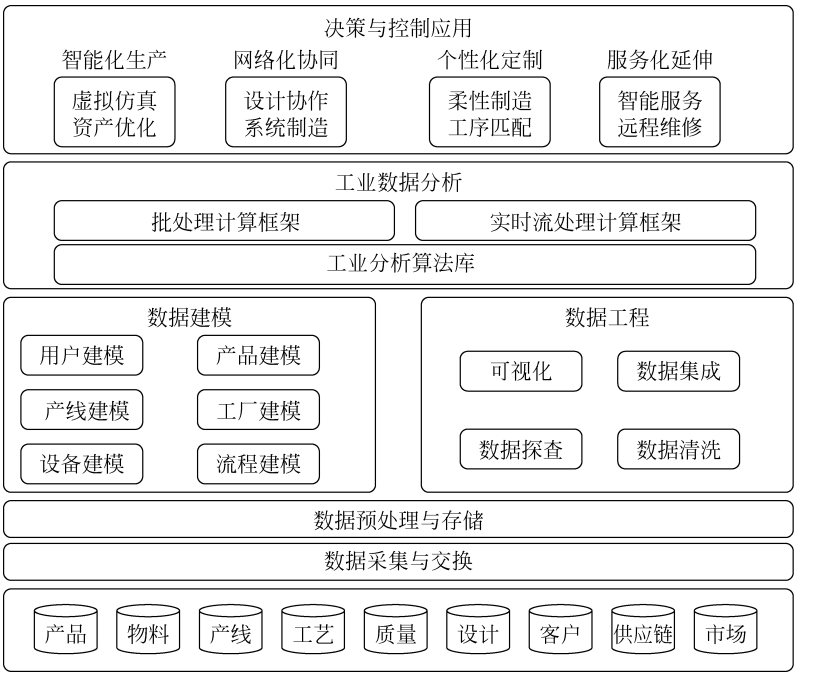

图 5-11　工业大数据在智能制造中的应用

稀土、农药等重点行业智能检测监管应用。此外，通过工业大数据的全流程建模，对数据源进行集成贯通，可以支撑以个性化定制为代表的典型智能制造模式。图 5-12 所示为制造业复杂装备阶段活动示意图。

工业大数据则主要是基于数据集合分析问题。从图 5-12 可以看出，业务活动沿实线部分从上游向下游传递，它主要反映了订单、票据等数据是否正确，这是信息化过程中需要解决的核心问题；虚线主要是反馈部分，通过分析数据集发现业务规律和决策准则，然后反馈给前面的各个环节使用，从而形成数据全生命周期的闭环，这就是信息化和大数据智能化的区别，然而两者又是不可分割的。

4. 工业大数据与工业互联网

当前，大数据已成为业界公认的工业升级的关键技术要素。在“中国制造 2025”的技术路线图中，工业大数据是作为重要突破点来规划的，而在未来的 10 年，以数据为核心构建的智能化体系会成为支撑智能制造和工业互联网的核心动力。

1）基于数字孪生的智慧研发场景应用

如今我国工业正向产业的高价值链环节迈进，工业产品的复杂度和集成度越来越高，设计更改频繁，模型一经修改，改变的内容还会影响到分析测试模型、生产模型、工程图等其他相关模型。利用数字孪生技术进行可视化建模，通过数字化模型的虚拟现实交互、仿真、快速成型，可及早发现设计缺陷，优化产品外形、尺寸和结构，克服以往被动、静态、单纯依赖人的经验的缺点，实现产品制造行业研发设计与生产过程在虚拟空间的实时监控和动态优化，促进制造资源的智能物联及共享协同，并有利于制造知识积累及高效重用。通过基于模型的设计生产一体化协同，缩短产品研制周期缩短，降低产品不良品率，提高生产率。图 5-13 所示为数字孪生模型。

2）基于柔性生产的大规模个性化定制场景

柔性生产是指让系统在制造过程中根据产品加工状况的改变自动进行调整，在原有的自

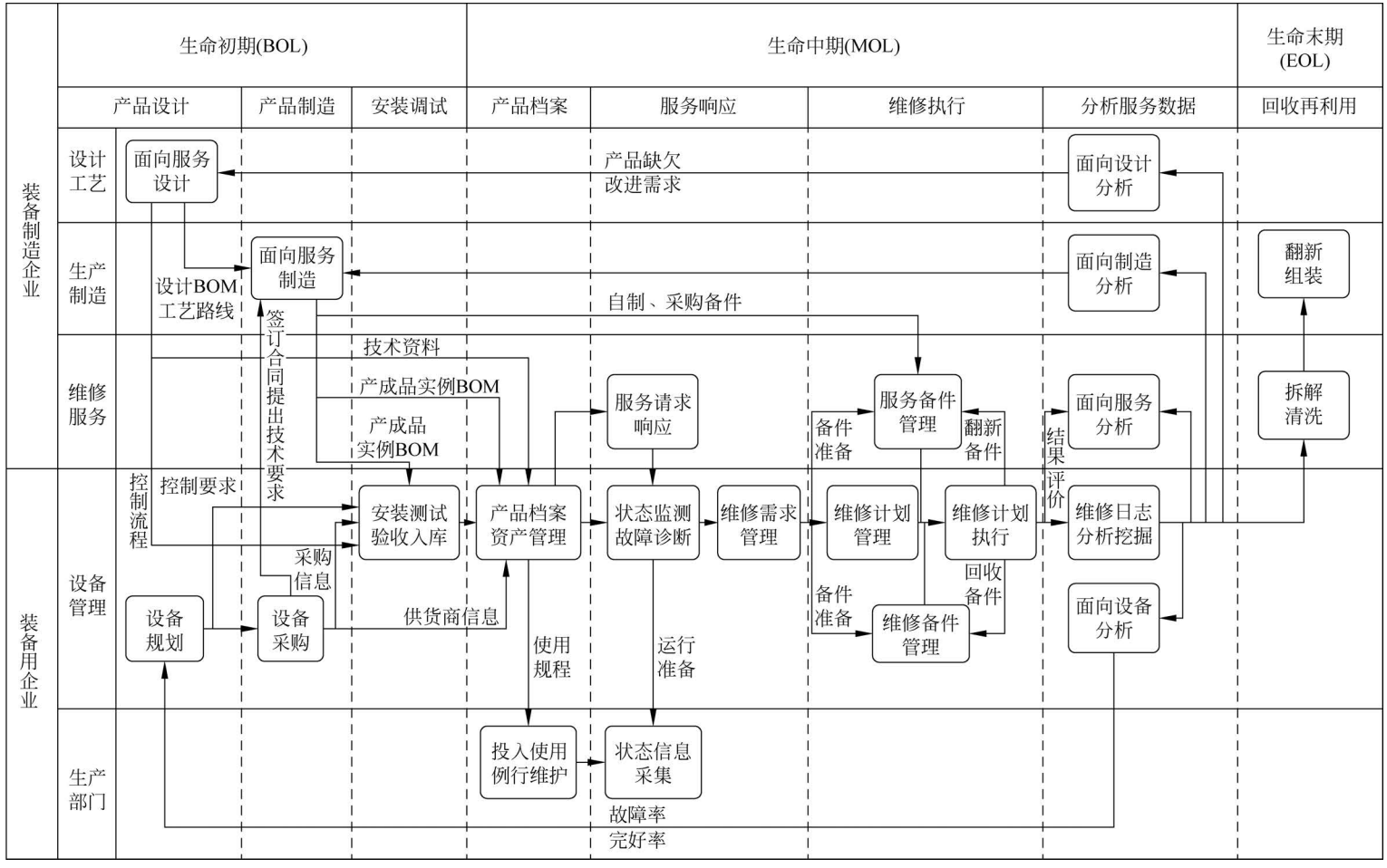

图 5-12　制造业复杂装备阶段活动示意图

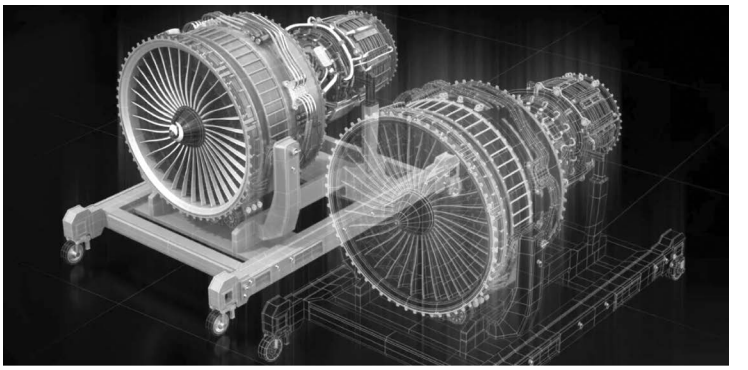

图 5-13　数字孪生模型

动化基础上实现系统的自省功能，实现制造过程的最优智能决策。企业通过外部平台采集客户个性化需求数据，与工业企业生产数据、外部环境数据相融合，将产品的共性特征数据与收集到的客户定制化数据结合转换为个性化的产品模型，并将产品方案、物料清单、工艺方案等数据信息通过制造执行系统快速传递给生产现场，以保证包括样式、颜色、尺寸、物料等在内的产品全生命周期的各个环节都满足个性化定制需求，从而快速生产出符合个性化需求的定制化产品。图 5-14 所示为基于柔性生产的产品定制。

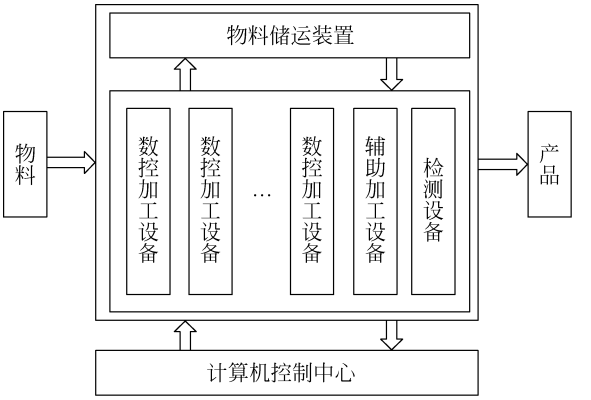

图 5-14　基于柔性生产的产品定制

3）基于产品全生命周期管理的设备预测管理

企业可通过传感器、边缘计算与工业大数据等技术，对产品使用过程中的自身工作状况、周边环境、用户操作行为等数据进行实时采集并连接至云端，在云端进行数据存储、分析与可视化展现，最终实现生产设备在线健康监测、故障诊断预警等服务，提高设备可靠性，延长设备使用寿命。

5. 工业大数据参考架构

工业大数据参考架构以工业过程的业务需求为导向，基于工业系统的业务架构，规划工业大数据的数据、技术和应用（平台）架构，搭建面向多业务领域、贯通多组织和应用层次的工业大数据 IT 架构。

工业大数据架构设计以业务应用需求为先导，将数据作为工业企业核心数据资产之一，与业务流程相互融合，多视图对工业大数据整个业务过程的业务、数据、技术和平台 4 个架构维度进行建模，如图 5-15 所示，实现企业以人流、物流、资金流和信息流等各业务线的顺畅运作。业务架构决定工业大数据的应用目标、价值实现和业务流程模型，树立了工业大数据需求和问

题导向的应用指导思想，既防止企业不重视数据应用、忽略数据资产价值的倾向，同时也防止脱离业务实际需求，避免出现数据处理过载的问题。数据架构实现业务架构所确定的业务模式向数据模型转变，以及业务需求向数据功能的映射。应用（平台）架构以数据架构为基础，建立支撑业务运行的各个业务系统，通过应用系统的集成运行，实现数据的自动化流动。技术架构定义工业大数据应用的主要技术、实现手段和技术途径，实现工业大数据应用的技术标准化，支撑其技术选择、开发技术组件。

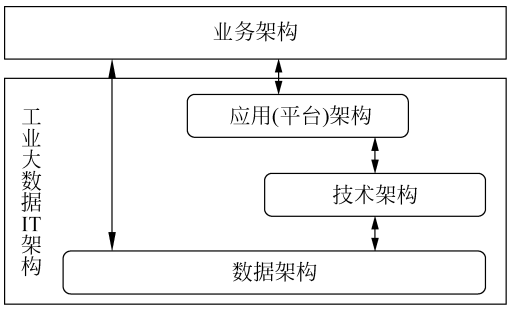

图 5-15　工业大数据架构

5.2.2　工业大数据的应用

工业大数据是提升工业生产效率，降低能耗，转变高耗能、低效率、劳动密集的粗放型生产面貌的必要手段。工业大数据结合数控机床、工业机器人等自动生产设备的使用，并建立从经营到生产系统贯通融合的数据流，做到数据全打通和数据流通不落地，可以提升企业整体生产效率，降低劳动力投入，有效管理并优化各种资源的流转与消耗。

同时，大数据也是实现工业企业从制造向服务转型的关键支撑技术。大数据技术兴起后，诞生了一大批以工业大数据为核心应用方向的企业，推出了一系列智能预测分析解决方案，对生产过程中的不同阶段进行归类，有多个主要的应用方向，分别是预防性维修、生产过程优化、智慧供应链、智能营销以及工业污染与环保监测。

1. 预防性维修

预防性维修主要面向设备的运用环节。工业运维经历了 4 个阶段，目前已经从事后维修逐渐向预防性维修发展。通过实施预防性维修，而不是应对性维修，可以降低设备整个生命周期内的费用，这样大多数的生产设施都有机会大幅提升它们的盈利水平。这有助于优化能源利用，减少设备停机，以及获得在其他方面的提升。

预防性维修主要依赖于数据和建模，主要有两种思路，一种基于机理辨别，对未知对象建立参数估计，进行阶次判定、时域分析、频域分析，或者建立多变量系统，进行线性和非线性、随机或稳定的系统分析等，试图揭示系统的内在规律和运行机理；另一种则是基于人工智能相关的灰度建模思路，利用专家系统、决策树、基于主元分析的聚类算法、支持向量机（Support Vector Machine，SVM）和深度学习等深度学习相关方法，对数据进行分析和预测。

例如，某风电装备企业利用大数据结冰动力模型对风机特征进行动态观测，重点观测和分析风机利用率、环境温度等特征，并对监测和诊断到的早期结冰状况进行及时处理，防止出现严重结冰，由此提高风机运行效率和电网的安全。

又如，为了实时监控发动机的状况，现代民航大多安装了飞机发动机健康管理系统。发动机健康管理系统可以分析由发射系统、信号接收系统、信号分析系统等方式采集的大量数据，从而实现对发动机运行状况的实时监控。

2. 生产过程优化

传统方法的生产过程优化以系统理论的实际应用为主，具有较大的局限性，不能针对具体的问题进行调整优化。而基于大数据的生产过程优化，在制造过程数字化监控的基础上，用大数据、人工智能算法建立模型，研究不同参数变化对设备状态与整体生产过程的影响，并根据实时数据与现场工况动态调优，提供智能设备故障预警、工艺参数优推荐、降低能耗、提升良品

率、提高工作效率等一项或多项功能，对于一些危险生产行业，还能用于控制降低风险，概括起来就是“提质、增效、降耗、控险”。

在具体实现中，目前无所不在的传感器、互联网技术的引入使产品故障实时诊断变为现实，而大数据应用、建模与仿真技术则使预测动态性成为可能。首先，在生产工艺改进方面，在生产过程中使用工业大数据，就能分析整个生产流程，了解每个环节是如何执行的。一旦有某个流程偏离了标准工艺，就会产生一个报警信号，能更快速地发现错误或瓶颈所在，也就能更容易地解决问题。其次，在生产过程中还可以对工业产品的生产过程建立虚拟模型，仿真并优化生产流程，当所有流程和绩效数据都能在系统中重建时，这种透明度将有助于制造商改进其生产流程。最后，在能耗分析方面，在设备生产过程中利用传感器集中监控所有生产流程，能够发现能耗的异常或峰值情形，由此便可在生产过程中优化能源的消耗，对所有流程进行分析将会大大降低能耗。

例如，某生产企业通过对工艺流程中相关参数的数据采集和筛选，利用筛选出的关键参数建立模型，并依据该模型优化实际生产的燃煤消耗，最终达到了能耗优化的目的。

又如，在半导体行业，芯片在生产过程中会经历许多次掺杂、增层、光刻和热处理等复杂的工艺制程，每步都必须达到极其苛刻的物理特性要求，高度自动化的设备在加工产品的同时，也同步生成了庞大的检测结果。如果按照传统的工作模式，人们需要按部就班地分别计算多个过程能力指数，对各项质量特性一一考核，过程十分烦琐。但是，当企业利用大数据质量管理分析平台，除了可以快速地得到一个长长的传统单一指标的过程能力分析报表之外，更重要的是，还可以从同样的大数据集中得到很多崭新的分析结果。

3. 智慧供应链

“智慧供应链”是结合物联网技术和现代供应链管理的理论、方法和技术，在企业中和企业间构建的实现供应链的智能化、网络化和自动化的技术与管理综合集成系统。与传统供应链相比，智慧供应链具有以下几个特征。

(1)智慧供应链的技术渗透性更强。在智慧供应链的语境下，供应链管理和运营者会系统地主动吸收包括物联网、互联网、人工智能等在内的各种现代技术，主动将管理过程适应引入新技术带来的变化。

(2) 智慧供应链可视化、移动化特征更加明显。智慧供应链更倾向于使用可视化的手段表现数据，采用移动化的手段访问数据。

(3) 智慧供应链更人性化。在主动吸收物联网、互联网、人工智能等技术的同时，智慧供应链更加系统地考虑问题，考虑人机系统的协调性，实现人性化的技术和管理系统。

例如，某家电制造企业利用大数据技术对供应链进行优化，改变了传统供应链系统对于固定提前期概念的严重依赖。通过分析相关数据创建更具有弹性的供应链，能够缩短供应周期，使企业获得更大的利润。

又如，某电子商务企业通过大数据提前分析和预测各地商品需求量，从而提高配送和仓储的效能，保证了次日到货的客户体验。

4. 智慧营销

智慧营销是大数据、物联网等信息技术与当代品牌营销领域新思维、新理念、新方法新工具以及人的创造性、创造力、创意智慧融合的产物。面对消费者无时无刻的个性化、碎片化需求，为满足消费者动态需求，建立在工业 4.0、柔性生产与数据供应链基础上的全新营销模式，将消费者纳入企业生产营销环节，实现全面的商业整合。智慧营销是以人为中心、以网络技术为基础、以创意为核心、以内容为依托、以营销为本质目的的消费者个性化营销，实现品牌与实

效的完美结合，将体验、场景、感知、美学等消费者主观认知建立在文化传承、科技迭代、商业利益等企业生态文明之上，最终整合虚拟与现实的当代创新营销理念与技术。

5. 工业污染与环保监测

工业大数据对环保具有巨大价值。目前，我国环境监测体系初步形成，但对于海量数据的运用仍然存在巨大的提升空间。大数据技术的植入，可明显增加环保数据解析的维度，透视众多企业的环境治理状况，开发出多种打击环保违法行为的手段，增强环境监管的效力。例如，企业可在传统人工手动监测的基础上，使用先进技术，创新监测手段，推动开展环境质量连续自动监测和环境污染遥感监测；可以预测排污和预警、监控，并提供关闭排污口的阈值。百度上线“全国污染监测地图”就是一个很好的环保方式，结合开放的环保大数据，百度地图加入了污染检测图层，任何人都可以通过它查看全国及自己所在区域内所有在环保局监控之下的排放机构（包括各类火电厂、国控工业企业和污水处理厂等）的位置信息、机构名称、排放污染源的种类，以及最近一次环保局公布的污染排放达标情况等。

5.3　工业大数据处理过程

5.3.1　工业大数据采集

1. 工业大数据在线采集

实现工业 4.0，需要高度的工业化、自动化基础，是漫长的征程。工业大数据是未来工业在全球市场竞争中发挥优势的关键。无论是德国工业 4.0、美国工业互联网还是“中国制造2025”，各国制造业创新战略的实施基础都是工业大数据的搜集和特征分析，以及以此为未来制造系统搭建的无忧环境。无论智能制造发展到何种程度，数据采集都是生产中最实际最高频的需求，也是工业 4.0 的先决条件。

互联网的数据主要来自互联网用户和服务器等网络设备，主要是大量的文本数据、社交数据以及多媒体数据等，而工业数据主要来源于机器设备数据、工业信息化数据和产业链相关数据。从数据采集的类型上看，不仅要涵盖基础的数据，还将逐步包括半结构化的用户行为数据、网状的社交关系数据、文本或音频类型的用户意见和反馈数据、设备和传感器采集的周期性数据、网络爬虫获取的互联网数据，以及未来越来越多有潜在意义的各类数据。工业数据在线采集方式主要包括以下几种。

（1）海量的 Key-Value 数据。在传感器技术飞速发展的今天，包括光电、热敏、气敏、力敏、磁敏、声敏、湿敏等不同类别的工业传感器在现场得到了大量应用，而且很多时候机器设备的数据大概要到毫秒级的精度才能分析海量的工业数据，因此，这部分数据的特点是每条数据内容很少，但是频率极高。

（2）文档数据，包括工程图纸、仿真数据、设计的 CAD 图纸等，还有大量的传统工程文档。

（3）信息化数据。由工业信息系统产生的数据，一般是通过数据库形式存储的，这部分数据是最好采集的。

（4）接口数据。由已经建成的工业自动化或信息系统提供的接口类型的数据，包括 TXT 格式、JSON 格式、XML 格式等。

（5）视频数据。工业现场会有大量的视频监控设备，这些设备会产生大量的视频数据。

（6）图像数据，包括工业现场各类图像设备拍摄的图片，如巡检人员用手持设备拍摄的设备、环境信息图片。

（7）音频数据，包括语音及声音信息，如操作人员的通话、设备运转的音量等。

(8) 其他数据,包括遥感遥测信息、三维高程信息等。

图 5-16 所示为工业大数据的现场数据采集。该方式属于物联网终端传感器系统的一种,通过装在机器上的无线模块采集指定机器 PLC 工作信息,上传到主机,主机处理数据后上传到云服务器。用户可在手机、平板电脑、计算机上查看机器工作信息,并可以有限度地设置机器工作参数。

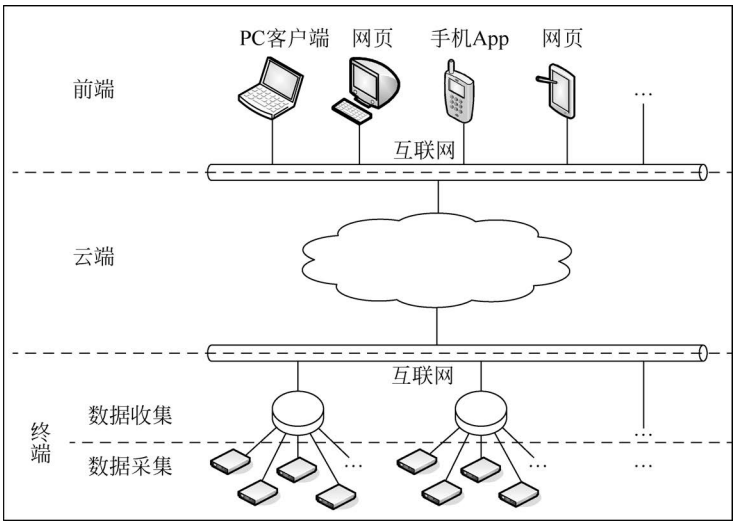

图 5-16 工业大数据的现场数据采集

2. 工业大数据离线采集

离线数据是相对于在线数据而言的。所谓在线数据,就是可以通过数据采集终端直接采集到的数据,如适配器对接设备控制器后直接采集到的数据;反之,离线数据就是不能通过数据采集终端直接采集的数据。例如,一个数控机床加工出来的工件,人们需要知道它的几何尺寸,就需要进行测量。又如,一个化工产品被生产出来,人们要进行化验,看看它的化学成分,就需要用到化验仪器,而化验结果可能显示在屏幕上,也可能是打印出来的。再如,人们需要知道某个仪表上的读数,而这个仪表是一个模拟表,没法对接什么适配器,人们就只好用眼睛去看这个仪表的读数。在制造行业中对上述这些数据进行采集,就是所谓的离线数据采集。

1) 日志数据

在大数据离线采集中,特别是在互联网应用中,不管是采用哪一种采集方式,其基本的数据来源大都是日志数据。例如,许多公司的业务平台每天都会产生大量的日志数据,对于这些日志信息,人们可以得到出很多有价值的数据。尤其对于 Web 应用,日志数据极其重要,它包含用户的访问日志、用户的购买数据或用户的点击日志等。

2) 离线数据采集方式

通常在采集离线数据库数据时,企业可使用 Redis、MongoDB 以及 HBase 等 NoSQL 数据库来完成,通过在采集端部署大量分布式数据库,并在这些数据库之间进行负载均衡和分片完成大数据采集工作。

例如,人们可以先将日志数据采集到 HDFS 中,再进一步使用 MapReduce、Hive 等对数据进行分析,这也是可行的。

又如,处理离线数据,可以使用开源的 Kafka。Kafka 是由 Apache 软件基金会开发的一个开源流处理平台,由 Scala 和 Java 语言编写。Kafka 是一种高吞吐量的分布式发布订阅消息系统,它可以处理消费者规模的网站中的所有动作流数据。互联网关采集到变化的路由信

息，通过 Kafka 的 Producer 将归集后的信息批量传入 Kafka。Kafka 按照接收顺序对归集的信息进行缓存，并加入待消费队列。Kafka 的 Consumer 读取队列信息，并以一定的处理策略将获取的信息更新到数据库，完成数据到数据中心的存储。

值得注意的是，随着工业物联网的快速发展，工业企业在生产经营过程中会采集大量的数据，并进行实时处理，这些数据都是时序的（按时间顺序记录的数据列）。在工业场景中，80％以上的监测数据都是实时数据，且都是带有时间戳并按顺序产生的数据，这些来源于传感器或监控系统的数据被实时地采集并反馈出系统或作业的状态。工业上的实时数据有这些特征：都带有时间戳，并且是按时间顺序生成的；大多为结构化数据；采集频率高、数据量大等。在工业上，通常会使用实时/历史数据库作为核心枢纽，对这些数据进行采集、存储以及查询分析。在工业领域之外，随着移动互联网、物联网、车联网、智能电网等新概念的迅速发展，也形成了对实时数据的分析处理需求，另一种全新架构的解决方案也悄然形成，被称作时序数据库，主要面向互联网场景下海量数据的实时监控和分析需求。

数字化工厂产生的时序数据量是巨大的，处理它有相当的技术挑战。以数控机床加工生产为例，由于工业行业的要求，需要将包括报警在内的各种工况数据存储起来。假设企业每个厂区具有 2000 个监测点，5s 为一个采集周期，全国一共 200 个厂区，这样粗略估算起来每年将产生惊人的几十万亿个数据点。假设每个点有 0.5KB 数据，数据总量将达 PB 级别（如果每台服务器的硬盘容量是 10TB，那么总共需要 100 多台服务器）。这些数据不仅要实时生成，写入存储，还要支持快速查询，实现可视化的展示，帮助管理者分析决策；并且也能够用来做大数据分析，发现深层次的问题，帮助企业节能减排，增加效益。

目前，时序数据处理应用于智慧城市、物联网、车联网、工业互联网领域的过程数据采集、过程控制，并与过程管理建立一个数据链路，属于工业数据治理的新兴领域。例如，工业企业为了监测设备、生产线以及整个系统的运行状态，在各个关键点都配有传感器、采集各种数据。这些数据是周期或准周期产生的，有的采集频率高，有的采集频率低，这些采集的数据一般会发送至服务器，进行汇总并实时处理，对系统的运行进行实时监测或预警。这些时序数据常常被长期保存下来，用以进行离线数据分析。

图 5-17 所示为使用时序数据分析工业生产中的机器异常状况。

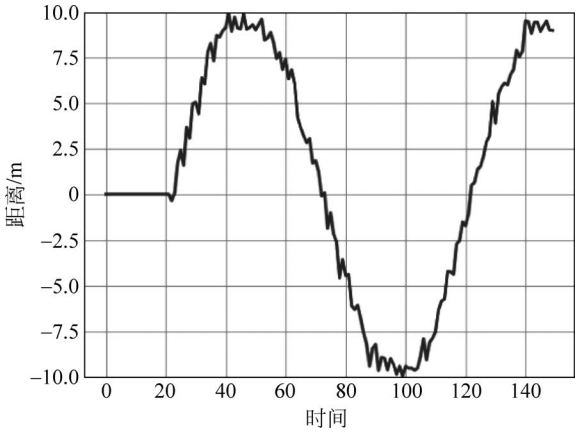

图 5-17　使用时序数据分析工业生产中的机器异常状况

5.3.2　工业大数据预处理

工业过程中产生的数据由于传感器故障、人为操作、系统误差、网络传输乱序等因素的影

响极易出现噪声(异常值)、缺失值以及数据不一致的情况,直接用于数据分析会对模型的精度和可靠性产生严重的负面影响。因此,在工业数据分析之前,需要采用一定的数据预处理技术,如消除数据中的噪声、纠正数据的不一致、删除异常值等,来提高模型鲁棒性。

1. 数据异常处理

异常值也叫作离群值,通常是指采集数据时可能因为技术或物理原因,数据取值超过数据值域范围。值得注意的是,异常值是数据分布的常态,处于特定分布区域或范围之外的数据通常被定义为异常或噪声。异常值通常分为两种:伪异常和真异常。伪异常是由于特定的业务运营动作产生,是正常反映业务的状态,而不是数据本身的异常;真异常不是由于特定的业务运营动作产生,而是数据本身分布异常,即离群值。

异常值会导致某些模型问题。例如,线性回归模型会显得异常值偏离,影响决策树模型的建立。通常,如果我们能找到合理移除异常值的理由,那么将会大大改善模型的表现,但这不意味着是异常值就一定要排除。例如,我们不能因为一个值"特别大"而将其归为异常值而不予以考虑。大数值可能为我们的模型提供重要的信息,这里不展开阐述。总之,在移除任何异常值之前,必须有充分的理由。

处理异常值,首先要识别异常值。目前对于异常值的检测可以通过分析统计数据的散度情况(即数据变异指标)来对数据的总体特征有更进一步的了解。常用的数据变异指标有极差、四分位数间距、均差、标准差、变异系数等。此外,也可以使用 3σ 原则检测异常数据。该方法是指若数据存在正态分布,那么在 3σ 原则下,异常值为一组测定值中与平均值的偏差超过 3 倍标准差的值。如果数据服从正态分布,距离平均值 3σ 之外的值出现的概率为 $P(|x-\mu|>3\sigma)\leqslant 0.003$,属于极个别的小概率事件。图 5-18 所示为 3σ 原则。

此外,箱线图也提供了识别异常值的一个标准:异常值通常被定义为小于 QL−1.5IQR 或大于 QU+1.5IQR 的值。其中,QL 称为下四分位数,表示全部观察值中有四分之一的数据取值比它小;QU 称为上四分位数,表示全部观察值中有四分之一的数据取值比它大;IQR 称为四分位距,是上四分位数 QU 与下四分位数 QL 之差,其间包含了全部观察值的一半。

箱线图依据实际数据绘制,对数据没有任何限制性要求,如服从某种特定的分布形式,它只是真实直观地表现数据分布的本来面貌。另外,箱线图判断异常值的标准以四分位数和四分位距为基础,四分位数具有一定的鲁棒性,多达四分之一的数据可以变得任意远而不会严重扰动四分位数,所以异常值不能对这个标准施加影响。图 5-19 所示为使用箱线图检测异常值。

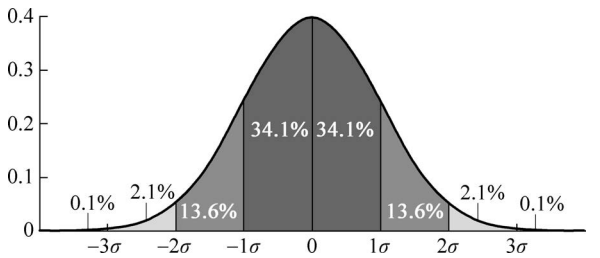

图 5-18 3σ 原则

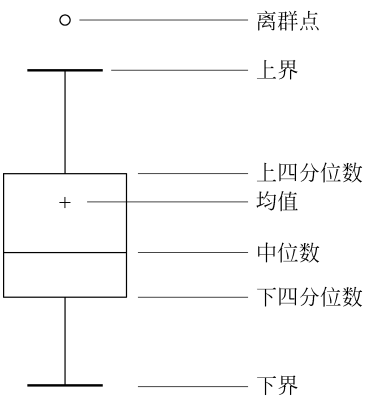

图 5-19 使用箱线图检测异常值

2. 数据缺失处理

现实世界的数据大多都是不完整的，工业大数据更是如此。在数据集中，若某记录的属性值被标记为空白或“-”等，则认为该记录存在缺失值（空值），也常指不完整的数据。造成数据缺失的原因是多种多样的，如空值条件的设置、业务数据的脱密、异常数据的删除等，都会造成一定程度的数据缺失。图 5-20 所示为数据表中的缺失数据。

year	month	day	hour	minute	temp_o	hum_out	pre_out	hum_in	pre_in	y
2019	3	24	6	13	10.7	93	990.5	88	989.6	11.4
2019	3	24	6	14	10.8	94	989.8	88	989.6	11.5
2019	3	24	6	15				88	989.6	11.5
2019	3	24	6	16	10.7	93	990.1	88	989.3	11.5
2019	3	24	6	17	10.7	93	990.2	88	989.6	11.5
2019	3	24	6	18	10.8	93	990.1	88	989.3	11.6
2019	3	24	6	19	10.8	93	990	88	989.3	11.5
2019	3	24	6	20	10.8	93	927.6	88	989.3	11.4
2019	3	24	6	22	10.8	93	989.6	88	989.45	11.3
2019	3	24	6	23	10.75	92	990.1	88	989.5	11.4
2019	3	24	6	24	10.7	92	989.9	88	989.5	11.4
2019	3	24	6	25	10.8	92	989.8	88	989.3	11.4
2019	3	24	6	26	10.8	93	989.6	88	989.3	11.5
2019	3	24	6	27	10.8	92	989.6	88	989.3	11.5
2019	3	24	6	28	10.8	92	989.6	88	989.5	11.5
2019	3	24	6	29				88	989.3	11.5

图 5-20　数据表中的缺失数据

缺失数据在机器学习应用中是比较棘手的问题。首先，不能简单地忽略数据集中缺失的数据值，而是必须以合理的理由处理这类数据，因为大多数算法是不接受缺失数据值的。对于缺失数据的清洗方法较多，如将存在遗漏信息属性值的对象（元组、记录）删除，或者将数据过滤出来，按缺失的内容分别写入不同数据库文件并要求客户或厂商重新提交新数据，要求在规定的时间内补全，补全后才继续写入数据仓库中。有时也可以用一定的值去填充空值，从而使信息表完备化。填充空值通常基于统计学原理，根据初始数据集中其余对象取值的分布情况对一个缺失值进行填充。

处理缺失值可以按照以下 4 个步骤进行。

（1）确定缺失值范围。对每个字段都计算其缺失值比例，然后按照缺失比例和字段重要性，分别制定策略。

（2）对于一些重要性高、缺失率较低的缺失值数据，可根据经验或业务知识估计，也可通过计算进行填补（插值）。常见的缺失值填补方式有线性插值法、拉格朗日插值法、牛顿插值法等。简单地说，插值就是通过离散的数据点求一条经过所有数据点的多项式函数去逼近未知的函数 $f(x)$。例如，某公司收集到了最近一周的气温数据，如图 5-21 所示，但是由于某些原因，星期四的数据丢失了。如果要处理缺失数据，首先，数据工程师将数据转换为坐标点画在坐标图上，并用观察到的 6 天的气温值去填补缺失的那一天的气温值。然后画出一条曲线经过所有点（插值条件），那么这条曲线就体现了 7 天内气温的波动趋势，如图 5-22 所示。最后，人们就可从图中找出星期四的对应的温度估值是 22。

日期	星期一	星期二	星期三	星期四	星期五	星期六	星期日
气温	17	20	21	不知道	23	22	19

图 5-21　最近一周的气温数据

（3）对于指标重要性高、缺失率也高的缺失值数据，需要和取数人员或业务人员沟通，看是否有其他渠道可以取到相关数据，必要时进行重新采集。若无法取得相关数据，则需要对缺失值进行填补。

(4) 对于指标重要性低、缺失率也低的缺失值数据，可只进行简单填充或不作处理；对于指标重要性低、缺失率高的缺失值数据，可备份当前数据，然后直接删掉不需要的字段。

5.3.3 工业大数据建模

1. DIKW 模型

DIKW 模型是一个关于数据(Data)、信息(Information)、知识(Knowledge)、智慧(Wisdom)的模型，如图 5-23 所示。

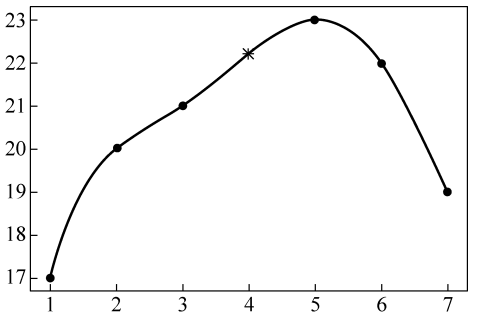

图 5-22 插值填补缺失数据

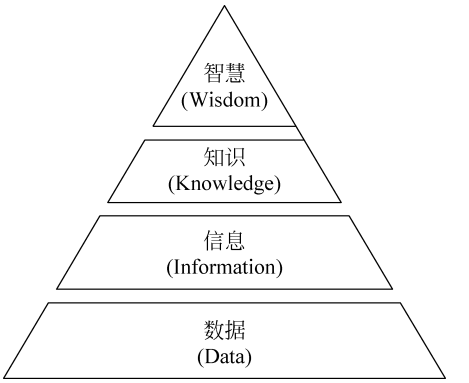

图 5-23 DIKW 模型

DIKW 模型将数据分为 4 个层级，由低到高分别是数据、信息、知识及智慧的体系。

数据是使用约定俗成的关键字，对客观事物的数量、属性、位置及其相互关系进行抽象表示，以适合在特定领域中使用人工或自然的方式进行保存、传递和处理。

信息是具有时效性、有一定含义的、有逻辑的、经过加工处理的对决策有价值的数据流。信息＝数据＋时间＋处理。

通过人们的参与对信息进行归纳、演绎、比较等手段进行挖掘，使其有价值的部分沉淀下来，并与已存在的知识体系相结合，这部分有价值的信息就转变为知识。

智慧是人类基于已有的知识，针对物质世界中运动过程中产生的问题根据获得的信息进行分析、对比、演绎找出解决方案的能力。这种能力运用的结果是将信息的有价值部分挖掘出来并使之成为知识架构的一部分。

DIKW 模型中提及的数据、信息、知识及智慧，它们之间的转化依赖于人们个人的经验、创造力和对内容的理解程度。结合气象数据的例子，就可以直观地了解这个分级模型。

某个观测站观测到某日的最高气温是 35℃，这就是一个数据。数据必须放到相应的环境中一起分析，这样才能了解数据之间的关系，可以分析出问题的根本原因(Root Cause)。例如，每款新机型在交付给航空公司之前都会接受一系列残酷的飞行测试，极端天气测试就是多项严酷的测试之一。该测试的目的是确保飞机的发动机、材料和控制系统能在极端天气条件下正常运行。

综合整月、全年乃至更长时段的气温数据，便能得到这个站点的气温序列，这就是信息。

基于某城市多个观测站的常年观测资料，人们就能够判断当地的气候条件如何，就形成了知识。

如果人们能够对知识进行进一步挖掘分析，利用它提炼出正确的决策，就进一步提升到了智慧。

DIKW 模型将数据、信息、知识、智慧纳入一种金字塔形的层次体系，展现了数据是如何

一步步转化为信息、知识乃至智慧的方式。当系统采集到原始的数据后，通过加工处理得到有逻辑的信息，再通过提炼信息之间的联系获得规则和知识、形成行动的能力和完成任务，最终对各种知识进行归纳和综合形成关注未来不确定性业务的预测能力，这样系统才能真正做到感知、分析、推理、决策、控制功能。

例如，系统通过传感器采集到实时的温度，再把该数据与其他数据关联（如批次、条码、机台、原料、产品质量等级等），同时可以计算生产过程中温度点的各种统计值，这些信息既可以根据已知的知识（工艺要求）进行过程控制，也可以进行相关性分析归纳出模型。

从最底层的观测结果到最高层的智慧，数据就是在这样的阶梯中上升，在递增中产生巨大的价值。知识是人类文明的结晶，涵盖了人类或智能体对世界的各种认知，数据建模的本质是发现知识。但目前工业领域的知识往往是相当丰富的，很少会发现全新的知识。因此，在这种背景下，工业领域的数据建模需要把分析结果和领域知识结合起来，并对已有领域的知识进行深入理解。

2. 知识工程

知识是指人类认识的成果或结晶。常见的知识可分为两种，一种是以书面文字、图表和数学公式加以表述的显性知识，如方案、图纸、源程序等；另一种是技能、经验、诀窍等未被表述的知识，称为隐性知识。

知识工程的研究方向是专家知识的获取、表达和推理过程的系统方法。知识获取研究的问题有对专家或书本知识的理解、认识、选择、抽取、汇集、分类和组织的方法，从已有的知识和实例中产生新知识的机理和方法，检查或保持已获取知识集合的一致性和完全性约束的方法，保证已获取的知识集合无冗余的方法等。知识表示是对知识的一种描述或是关于如何描述事物的一组约定，是一种计算机可以接受的、用于描述知识的数据结构。知识表示的方法有很多，如谓词逻辑表示法、脚本表示法、框架表示法、产生式表示法、函数式表示法、语义网络表示法、状态过程表示法、面向对象表示法等。知识的运用和处理主要包括推理、搜索、知识的管理及维护以及匹配和识别。

工业领域的知识按照其属性可以分为隐性知识、显性知识以及工业大数据三大类，并通过知识之间互相作用、互相转化，应用到企业创新业务活动中。例如，可以通过学习、理解、综合、观察、模仿、感知、试错、实践、试验、数据收集、归纳、分析、总结等方法完成知识之间的转化与关联，获得与掌握工业技术（知识）。这一过程中，人脑作为主要载体，使工业技术（知识）被掌握、被理解并应用到工业过程。

3. 工业建模基础

数据建模的本质是根据一部分能够获得的数据获得另一部分不统一直接获得的数据。不失一般性，某个工业对象可以用函数 $Y=F(X)$ 描述，在这里 F 是一个固定的映射，输入 X 则可计算 Y。不过在现实中，X 往往是无法准确获得的。这时，人们要设法在可以得到的数据中，寻找一些与 X 相关的变量，如 Z。于是，现实的数据模型往往就变成 $Y=H(Z)$。

例如，某厂发现一种材料的合格率与生产这种材料的班组有关。事实上，合格率与某个工艺参数有关，不同班组采用的工艺参数不一样。但每个班组采用的参数不同，也没有记录。所以，人们看到的是合格率与班组有关。在这个例子中，工艺参数就是 X，而班组就是 Z。

又如，人们经常发现材料的性能与生产的季节相关。本质上，材料的性能与生产材料时的温度、湿度、空气流动的综合情况有关。在这里，温度、湿度、空气流动情况是人们需要的 X，而生产季节就是 Z。

因此，对于复杂的工业建模过程，充分利用知识领域是成功的前提。不过，需要注意的是，

在工业领域，由于应用场景的不同以及数据采集条件的不断变化，模型的误差可能会变得很大，而这些变化会对人们的建模过程产生深刻的影响。

1）复杂模型的建立

在工业数据建模中最常见的困难是部分数据无法获得。对此，一般的解决方法是从可以获得的数据中找到一些与之相关的数据，再用间接的手段确定模型。例如，人们可以把输入 X 分成两部分：可以准确获得的记为 X_1，难以准确获得的记为 X_2。而为了获得 X_2，人们可以考虑以下几类相关数据，分别记为 Z_1、Z_2 和 Z_3。用公式描述为

$$Y=G(X_1,Z)$$

其中，$Z=(Z_1,Z_2,Z_3)$，表示建模时可供选择的数据；G 表示工业实际建立的模型。

2）工业大数据模型的理解

事实上，工业大数据的方法早已出现在前人的实践中。这类方法的基本思路就是找一个类似的做法，在此基础上进行修订。例如，冶炼钢水前，需要给出合适的工艺参数。计算过程涉及很多参数，不容易算对，解决这个问题的思路是先从历史数据中找类似的成功案例，以此为基础，根据案例炉与本炉次的参数差异进行修正。工业大数据的根本优势是数据的质量好。质量好的一个方面就是数据分布范围大，覆盖了各种可能发生的情况。这就是所谓的“样本等于全体”。在这样的前提下，就总能从历史上找到类似的案例。所以，大数据的本质优势是数据来源全面，而不是数量多到什么程度。如果数据存储得足够久，场景存储得足够多，新问题就会越来越少，这类方法就容易走向实用了。以设备故障诊断为例，针对单台设备研究问题时，故障样本就少，甚至每次都不一样。但是，如果把成千上万台设备的信息收集起来，情况就不一样了，每次出现问题，都容易在历史数据库中找到类似的案例。这时，人们研究的重点往往是如何利用理论的指导，更加准确地寻找类似案例，更加准确地修正。

不过，值得注意的是，在工业数据建模中，变量的选取是非常重要的。变量的选择不同，最终的模型就不一样，其中一个重要的差别是模型的精度和适用范围不同。对于科学理论模型，模型的精度高往往意味着适用范围大，而现实的模型则不一定。从这种意义上说，模型精度未必越高越好。有些人开发的模型精度比较高，却不能得到生产厂的认可。背后的原因是模型在生产稳定时精度很高，在生产不稳定时精度较低。由于多数时间的生产是稳定的，模型的平均精度往往较高。但是，生产稳定时，工人对模型没有需求；工人对模型有需求时，往往是生产不稳定的时候。

理论上讲，许多工业过程都可以用科学公式描述，但现实的影响因素太多。化工、冶金等行业的一种典型的现象是在同一个生产过程中同时存在着几十种化学反应。每种化学反应都可以用简单的化学反应方程来描述。但反应之间互相影响，许多参数会动态变化、无法准确确定，整体的化学反应过程就很难准确描述。对于这样的情况，传统的办法很难建立准确的模型。除了本身复杂外，一个重要的原因就是许多干扰是不可见的。这就会对模型的验证带来巨大的困难。在大数据的背景下，解决这类困难成为可能。一个重要的原因是当数据量足够大时，随机干扰是可以通过平均的方法滤除的，这相当于数据的精度可以大大提高。同时，大数据还可能为人们提供较好的样本分布，有助于复杂问题的解耦，即把复杂的、变量多的模型简化为若干变量数目少的简单模型。

4. 工业大数据的参考模型 CRISP-DM

CRISP-DM 模型是欧盟起草的跨行业数据挖掘标准流程（Cross-Industry Standard Process for Data Mining）的简称。这个标准以数据为中心，将相关工作分为业务理解、数据理解、数据准备、构建模型、模型评估、模型部署 6 个基本的步骤，如图 5-24 所示。在该模型中，

相关步骤不是顺次完成，而是存在多处循环和反复。在业务理解和数据理解之间、数据准备和构建模型之间，都存在反复的过程。这意味着这两对过程是在交替深入的过程中进行的，更大的一次反复出现在模型评估之后。

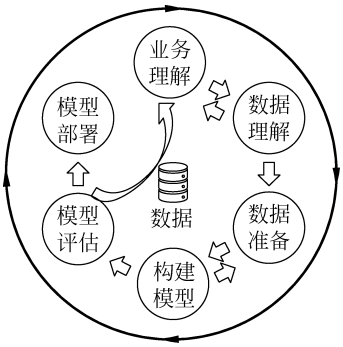

图 5-24　CRISP-DM 模型

1）业务理解

该阶段的目标是明确业务需求和数据分析的目标，将模糊的用户需求转化为明确的分析问题，必须清晰到计划采取什么手段、解决什么问题，要将每个分析问题细化为明确的数学问题，同时基于业务理解制定分析项目的评估方案。

2）数据理解

该阶段的目标是建立数据和业务的关联关系，从数据的角度去深度地解读业务，包括发现数据的内部属性、探测引起兴趣的子集形成隐含信息的假设、识别数据的质量问题、对数据进行可视化探索等。

3）数据准备

该阶段的目标是为数据的建模分析提供干净、有效的输入数据源。首先，基于业务目标筛选有效数据，筛选的数据能够表征业务问题的关键影响因素；其次，对数据的质量进行检查和处理，处理数据的缺失情况、异常情况等；最后，对数据进行归约、集成、变换等，输出建模可用的数据源。

4）构建模型

该阶段是基于业务和数据的理解，选择合适的算法和建模工具，对数据中的规律进行固化、提取，最后输出数据分析模型。首先，基于业务经验、数据建模经验，对业务问题进行逻辑化描述，探索解决问题的算法，反复迭代选择一个最优算法方案；其次，基于输入数据加工关键因子的特征变量，作为建模输入变量，建立有效可靠的数据模型。

5）模型评估

该阶段首先从业务的角度评估模型的精度问题，是否能够满足现有业务的要求；其次分析模型的中影响因子的完备性，为模型的下一步迭代指明优化路径；最后考查模型的假设条件，是否满足实际落地的条件，对模型的部署进行可行性验证。

6）模型部署

在该阶段中，首先，要基于分析目标，制定模型的使用方案和部署方案，并提前为模型的部署做好环境的准备工作；其次，针对模型部署过程中出现的质量问题、运行问题、精度问题等提前做好预备方案；最后，基于模型试运行后的结果，制定模型的持续优化方案。

值得注意的是，在实际工作中，不能单纯只通过数据理解工业对象及相关业务，而是要结合一定的专业领域知识，才能理解数据的含义。业务理解是数据理解的基础，是数据理解的起点；反过来，离开数据，人们对对象的理解将会是粗糙的、模糊的，不利于对系统和业务的精准控制和优化。所以，数据理解支撑对业务理解的深化。

5. 工业大数据建模应用

工业大数据的建模要求用数理逻辑严格地定义业务问题。由于工业生产过程中本身受到各种机理约束条件的限制，利用历史过程数据定义问题边界往往达不到工业的生产要求，因此，人们往往需要采用数据驱动＋模型驱动＋场景部署的多轮驱动方式，实现数据和机理的深度融合，解决实际的工业问题。图 5-25 所示为工业大数据建模的常见流程，其中数据场景化也称为数据场景化分析，它并非只是简单地基于对业务场景的数据分析，而是建构于数字化时

代企业 IT 新架构之上，以数据为基础的应用。数据场景化分析通过从数据和计算层级中实时接入的有用数据，基于丰富的业务模型开展数据的应用，使数据赋能企业业务和经营。不是所有业务场景都需要场景化分析。企业可基于对业务场景的深刻理解和对业务痛点的清晰洞察，选择一个或多个场景开展场景化分析，并随着业务的开展随时调整或拓展场景化分析的领域。在实际应用中，场景化分析将日常的执行和长周期的前瞻性规划连接在一起，可以实现按日、按月、按季、按年的上下横纵协同，通过滚动和整合的计划方法进行市场目标、财务目标、库存目标、服务目标和生产目标等的适时合理调整，提高企业整体的运营效率。

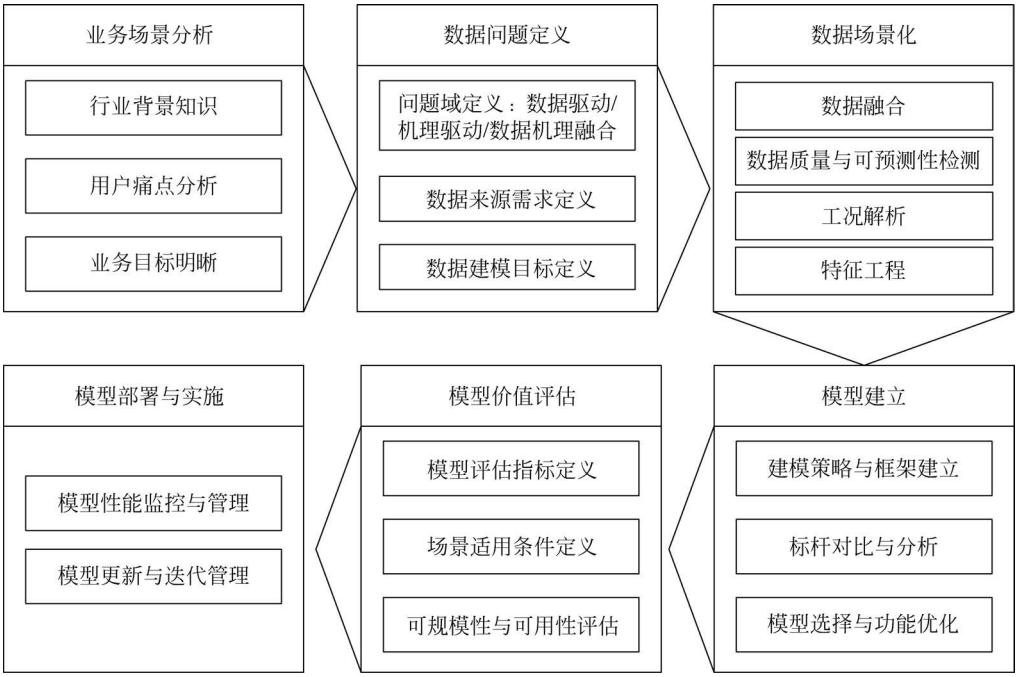

图 5-25　工业大数据建模的常见流程

图 5-26 所示为数控车床寿命预测模型。该模型设备部件为主轴，设备名称为数控车床，通过建立模型预测其寿命，并通过可视化图表显示模型参数，如图 5-27 所示。

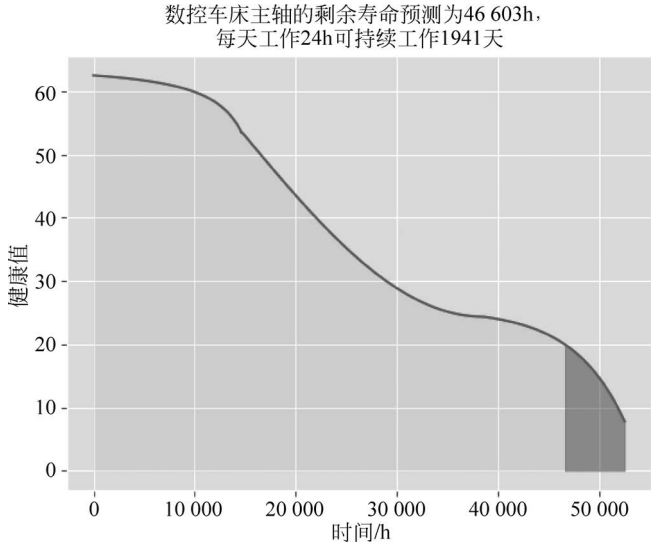

图 5-26　数控车床寿命预测模型

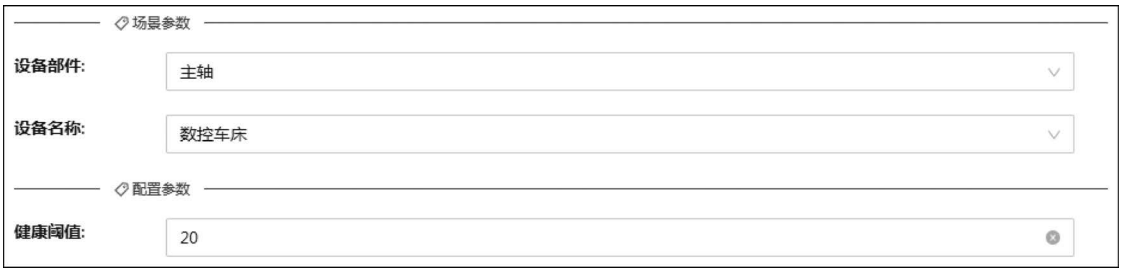

图 5-27　数控车床寿命预测模型参数

5.3.4　工业大数据分析

1. 认识工业大数据分析

工业大数据分析是利用统计学分析技术、机器学习技术、信号处理技术等技术手段，结合业务知识对工业过程中产生的数据进行处理、计算、分析并提取其中有价值的信息、规律的过程。

工业大数据分析的直接目的是获得业务活动所需各种的知识，贯通大数据技术与大数据应用之间的桥梁，支撑企业生产、经营、研发、服务等各项活动的精细化，促进企业转型升级。当代大数据处理技术的价值在于技术进步，同时也是因为技术进步，使大数据成为商业中有价值的核心驱动因素。作为智能制造的核心环节，工业大数据分析已经被多数制造企业所认知并接受。图 5-28 所示为大数据分析在工业中的应用。

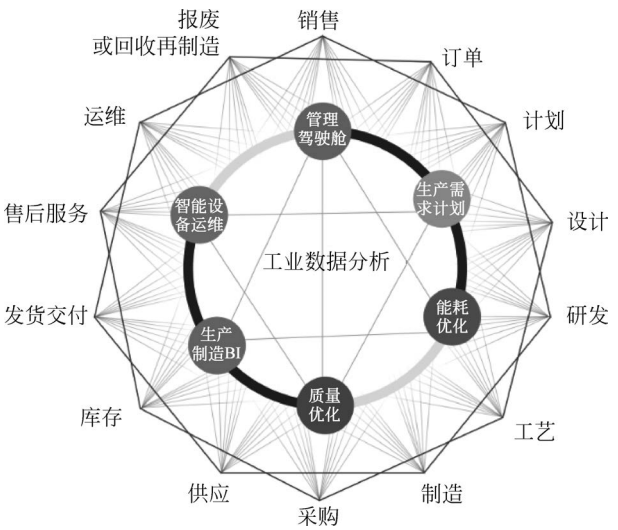

图 5-28　大数据分析在工业中的应用

2. 工业大数据分析的类型

工业大数据分析常见的类型可分为描述类、诊断类、预测类、决策类和控制类等。

1）描述类

描述类主要利用报表、可视化等技术，汇总展现工业互联网各个子系统的状态，使操作管理人员可以在一个仪表盘(Dashboard)上总览全局状态。此类应用一般不给出明确的决策建议，完全依靠人作出决策。

2）诊断类

诊断类通过采集工业生产过程相关的设备物理参数、工作状态数据、性能数据及其环境数据等，评估工业系统生产设备等运行状态并预测其未来健康状况，主要利用规则引擎、归因分

析等，对工业系统中的故障给出告警并提示故障可能的原因，辅助人工决策。

3）预测类

预测类通过对系统历史数据的分析挖掘，预测系统的未来行为。预测类主要是利用逻辑回归、决策树等预测未来系统状态，并给出建议。

4）决策类

决策类通过对影响决策的数据进行分析与挖掘，发现决策相关的结构与规律，主要是利用随机森林、决策树等多种机器学习算法，提出生产调度、经营管理与优化方面的决策建议。

5）控制类

控制类根据确定的规则，直接通过数据分析产生行动指令，控制生产系统采取行动。该类分类主要目前应用在智能制造业中。

3. 机器学习

机器学习是一门涉及多领域的交叉学科，其包含高等数学、统计学、概率论、凸分析和逼近论等多门学科。机器学习的研究方法通常是根据生理学、认知科学等对人类学习机理的了解，建立人类学习过程的计算模型或认识模型，发展各种学习理论和学习方法，研究通用的学习算法并进行理论上的分析，建立面向任务的具有特定应用的学习系统。

通俗地讲，就是机器学习让机器实现学习的过程，让机器拥有学习的能力，从而改善系统自身的性能。让机器具备人工智能的前提就是需要用一定量的数据集对机器进行“训练”。对于机器而言，这里的“学习”指的是从数据中学习，从数据中产生“模型”的算法，即“学习算法”。有了学习算法，只要把经验数据提供给它，它就能够基于这些数据产生模型，在面对新的情况时，模型能够提供相应的判断，进行预测。因此，机器学习实质上是基于数据集的，通过对数据集的研究，找出数据集中数据之间的联系和数据的真实含义。

在机器学习中，首先要输入大量数据，并根据需要训练模型，再对训练后的模型进行应用，以判断算法的准确性，如图 5-29 所示。

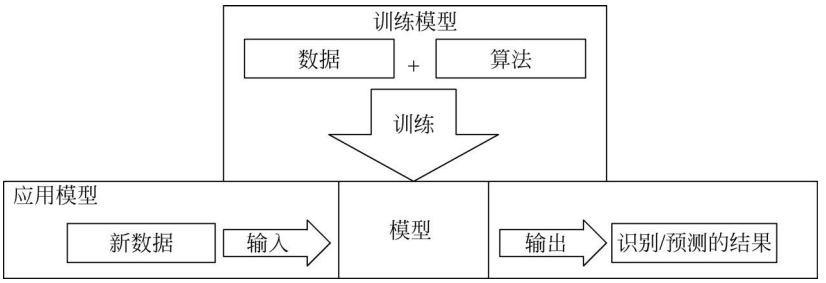

图 5-29　机器学习的模型

机器学习应用广泛，无论是在工业领域还是商用领域，都有机器学习算法施展的机会。近年来，机器学习的研究与应用在国内越来越受重视。机器学习已经广泛应用于语音识别、图像识别、数据挖掘等领域。

1）机器学习的分类

机器学习可分为监督学习、无监督学习和半监督学习。

（1）监督学习。监督学习是指利用一组已知类别的样本调整分类器的参数，使其达到所要求性能的过程，也称为监督训练或有教师学习。

（2）无监督学习。无监督学习和监督学习是一个相对的概念。在监督学习的过程中，人们需要给训练数据打上标签，这是必不可少的一步。而无监督学习中，就不再需要提前对数据进行

人工标记。所以，无监督学习常常被用于数据挖掘，用于在大量无标签数据中发现一些信息。

（3）半监督学习。半监督学习突破了传统方法只考虑一种样本类型的局限，综合利用有标签与无标签样本，是在监督学习和无监督学习的基础上进行的研究。

2）机器学习与流程工业大数据建模

流程工业也称为过程工业，是指通过物理变化和/或化学变化进行生产的行业。典型的流程工业包括石油、化工、冶金、造纸、医药、食品等行业。常见的流程工业系统模型为具有高度非线性的代数、微分方程混合组成的数学模型，并且在模型中包含大量的过程参数和高维状态变量，且高度交联、耦合。因此，现代流程工业具有操作可调、工艺灵活、产品多样、控制系统化等诸多特点。

蓬勃发展的大数据时代对流程工业产生了巨大影响，为实现智能制造提供了前所未有的机遇，这种新的生产方式不仅要求机器能够帮助人类减轻繁重的体力劳动，还要能有效地承担智力劳动，实现自主创新。在现代流程工业中，可以收集和存储越来越多的蕴藏有价值信息的数据。通过利用数据，数据分析和机器学习可以帮助感知环境、发现知识，并自动智能地作出决策。目前，机器学习已经成为工业大数据制造领域的一个热点话题，为处理和分析机器数据提供了许多有用的工具。

在当前的大数据时代，数据分析和机器学习在流程工业中得到了越来越广泛的应用，这些方法渗透到流程工业的各个层次，既包括在过程监控和软测量等底层控制回路中的应用，也包括最优控制和顶层决策等应用。前者的目的是帮助工程人员更好地监测和操作过程，识别过程的关键变化，而不是直接作出决策；相反，最优控制和顶层决策会对工业生产过程造成直接的影响。

工业大数据建模任务一般可以分为无监督学习和监督学习。在无监督学习中，通过建立描述性模型描绘输入数据中的隐藏结构，主要用于描述过程数据的分布，在此基础上实现过程监控。监督学习主要建立输入与输出之间的函数映射，包括回归和分类，因此输出的预测精度是一个关键的因素。在工业生产过程中，快速采样的过程变量多被用于关键质量变量的软测量建模与预报。近年来，表示学习或特征学习得到了越来越多的关注，其要点在于需要在构建模型时紧密结合特定领域的知识。这样，模型的可解释性能够得到显著增强，从而进一步提高模型性能。表示学习的一个具体例子是具有分片线性的神经网络在计算机视觉中的广泛应用。由于图形的抽象特征具有局部不变性，即分片线性，因此将特定领域的知识抽象为分片线性单元有助于提高模型的性能。

4. 特征工程

在信息化时代，数据已经成为现代企业的重要的资产。任何智能系统都需要由数据驱动。这些系统的核心，都是由一个或多个基于某种数据学习的方法或算法，如机器学习、深度学习或统计方法，系统通过分析和利用数据而生成知识，并以此提供决策支持。算法无法直接利用原始数据，而是需要从原始数据中提取有意义的特征，然后人们才能理解和使用数据。

特征是建立在原始数据基础之上的特定表示，是一个单独的可测量的属性，通常由数据集中的列表述。对于一个通用的二维数据集，每个数值由一行表示，特征用列表示，所有数据形成一个二维矩阵，这就是特征集。直接在原始数据之上构建模型是很困难的，无法直接获得期望的结果，需要进行数据准备，对原始数据进行预处理和分析，从中提炼出有意义的属性或特征，这就是特征工程。特征工程是将原始数据转换为特征的过程，这些特征可以更好地描述这些数据和潜在问题，并且利用它们建立的模型，在未知数据上的表现性能可以达到或接近最佳性能，从而提高模型的准确性。图 5-30 所示为特征工程的重要性，在工业大数据中一个好的解决方案来源于对业务的深入理解和对数据的细致分析。

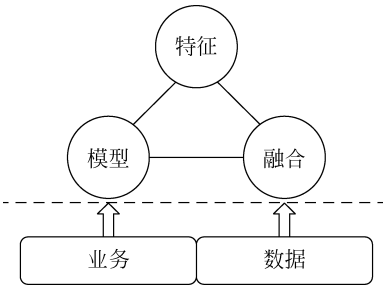

图 5-30　特征工程的重要性

5. 工业大数据分析的常见算法

1）聚类分析

聚类分析是指对一批没有标出类别的样本（可以看作数据框中的一行数据）按照样本之间的相似度进行分类，将相似的归为一类，不相似的归为另一类的过程。这里的相似度指样本特征之间的相似程度。把整个样本集的特征向量看作分布在特征空间中的一些点，点与点之间的距离即可作为相似度的测量依据，也就是将特征空间中距离较近的观察样本归为一类。两个样本的距离越近，相似度就越大。通俗地讲，聚类分析最终的目标就是实现“物以类聚，人以群分”。将样本的群体按照相似性和相异性进行不同群组的划分，经过划分后，每个群组内部各个对象间的相似度会很高，而在不同群组的样本彼此间将具有较高的相异度。

聚类分析适用于很多不同类型的数据集合，很多研究领域，如数学、计算机科学、统计学、生物学和经济学等，都对聚类分析的发展和应用起到了推动作用。

聚类分析实现的一般步骤为根据已知数据（一批观察个体的许多观测指标），按照一定的数学公式计算各观察个体或变量（指标）之间亲疏关系的统计量（距离或相关系数等），根据某种准则（最短距离法、最长距离法、中间距离法、重心法等），使同一类内的差别较小，而类与类之间的差别较大，最终将观察个体或变量分为若干类。

（1）计算样本间的距离。假设每个样本（看作数据框的行属性）有 p 个变量（看作数据框不同的列属性），则每个样本都可以看作 p 维空间中的一个点，n 个样本就是 p 维空间中的 n 个点，则第 i 个样本与第 j 个样本之间的距离记为 d_{ij}。设 $x_i=(x_{i1},x_{i2},\cdots,x_{ip})'$ 和 $x_j=(x_{j1},x_{j2},\cdots,x_{jp})'$ 分别是第 i 个和 j 个样本的观测值，则二者之间的常用距离有以下几种。

① 欧氏（Euclidean）距离，用公式表示为

$$d_{ij}=\sqrt{\sum_{k=1}^{p}(x_{ik}-x_{jk})^2}$$

② 绝对值距离，用公式表示为

$$d_{ij}=\sum_{k=1}^{p}|x_{ik}-x_{jk}|$$

③ 切贝雪夫（Chebychev）距离，用公式表示为

$$d_{ij}=\max_{k=1}^{p}|x_{ik}-x_{jk}|$$

④ 闵氏（Minkowski）距离，用公式表示为

$$d_{ij}=\left(\sum_{k=1}^{p}|x_{ik}-x_{jk}|^{p}\right)^{\frac{1}{p}},\quad p>0$$

⑤ 兰氏（Lance & Williams）距离，用公式表示为

$$d_{ij}(L)=\frac{1}{p}\sum_{k=1}^{p}\frac{|x_{ik}-x_{jk}|}{x_{ik}+x_{jk}},\quad x_{ij}>0$$

⑥ 马氏（Mahalanobis）距离，这是印度著名统计学家马哈拉诺比斯（P. C. Mahalanobis）所定义的一种距离，用公式表示为

$$d_{ij}^2=(x_i-x_j)'\boldsymbol{\Sigma}^{-1}(x_i-x_j)$$

其中，$\boldsymbol{\Sigma}$ 为观测变量之间的协方差矩阵。

（2）K-Means 聚类。K-Means 聚类也称为动态聚类、逐步聚类、迭代聚类、K-均值聚类，

快速聚类，适用于大型数据。K-Means 聚类中的 K 代表类簇的个数，Means 代表类簇内数据对象之间的均值(这种均值是一种对类簇中心的描述)，因此，K-Means 算法又称为 K-均值算法。K-Means 聚类是一种基于划分的聚类算法，以距离作为数据对象间相似性度量的标准，即数据对象间的距离越小，它们的相似性越高，它们就越有可能在同一个类簇。数据对象间距离的计算有很多种，K-Means 聚类通常采用欧氏距离计算数据对象间的距离。

首先导入一组具有 n 个对象的数据集，给出聚类个数 K，K-Means 聚类的思想可描述如下。

① 首先初始化 K 个类簇中心；

② 根据欧氏距离计算各个数据对象到聚类中心的距离，把数据对象划分至距离其最近的聚类中心所在的类簇；

③ 根据所得类簇，更新类簇中心；

④ 继续计算各个数据对象到聚类中心的距离，把数据对象划分至距离其最近的聚类中心所在的类簇；

⑤ 根据所得类簇，继续更新类簇中心，一直迭代，循环步骤②～步骤④直到达到最大迭代次数，或者两次迭代的差值小于某一阈值时，迭代终止；

⑥ 得到最终聚类结果。

图 5-31 所示为 K-Means 聚类算法的实现，将数据聚为 3 类，并以不同的颜色区分。

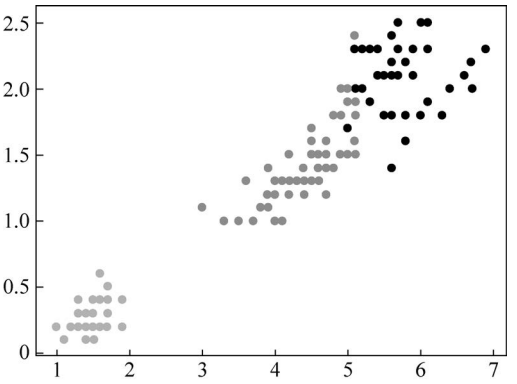

图 5-31　K-Means 聚类算法的实现

在工业生产中，聚类算法往往应用于工艺优化，如对车间生产历史数据进行聚类分析，得到工艺参数与产品质量、能耗水平的影响关系，从而提升制造水平；对生产过程和设备使用过程中异常点进行聚类，为设备潜在性能提升提供依据。

例如，在数控机床中，主轴是最核心的部件，在加工中担任重要的角色。主轴发生故障时，因为故障原因难以确定，导致维修时间过长，降低了机器生产的效率。所以，对故障进行聚类识别进而缩短设备的维修时间，可以提高设备工作效率，从而提高工厂产量。模型使用收集的主轴发生不同类型的故障时产生的特征组成的多维向量作为输入，利用故障聚类算法得到故障的分类结果。

2）降维算法分析

在互联网大数据场景下，人们经常需要面对高维数据，在对这些数据进行分析和可视化时，人们通常会面对高维这个障碍。在数据挖掘和建模的过程中，高维数据也同样带来大的计算量，占据更多的资源，而且许多变量之间可能存在相关性，从而增加了分析与建模的复杂性。人们希望找到一种方法，在对数据完成降维压缩的同时，尽量减少信息损失。由于各变量之间存在一定的相关关系，因此可以考虑将关系紧密的变量转换为尽可能少的新变量，使这些新变量两两不相关，那么就可以用较少的综合指标分别代表存在于各个变量中的各类信息。降维就是这样一类算法。数据降维，一方面可以解决“维数灾难”，缓解“信息丰富、知识贫乏”的现状，降低复杂度；另一方面可以更好地认识和理解数据。

常用的降维算法有主成分分析和因子分析。

图 5-32 所示为数据集在三维特征空间中的分布，图 5-33 所示为数据集在二维特征空间中的分布。

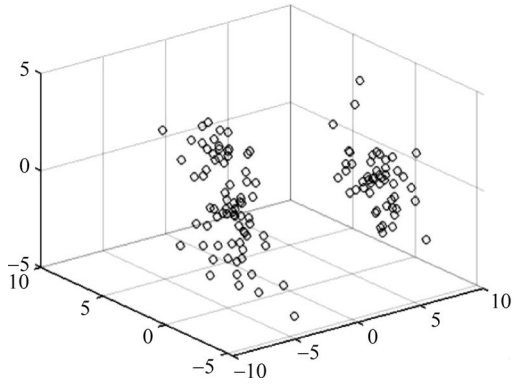

图 5-32　数据集在三维特征空间中的分布

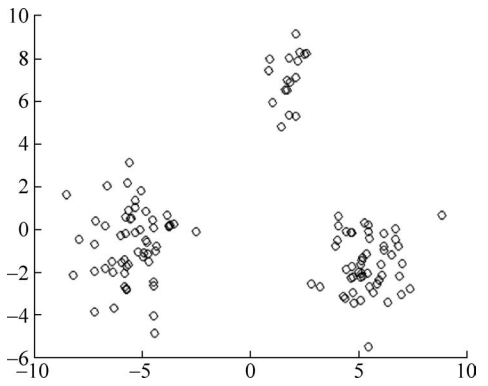

图 5-33　数据集在二维特征空间中的分布

(1) 主成分分析(Principal Component Analysis,PCA)是利用降维的思想,在保持数据信息丢失最少的原则下,对高维的变量空间进行降维,利用正交变换把一系列可能线性相关的变量转换为一组线性不相关的新变量,即在众多变量中找出少数几个综合指标(原始变量的线性组合),并且这几个综合指标将尽可能多地保留原来指标的信息,且这些综合指标互不相关。这些综合指标就称为主成分。

主成分组合之后新变量数据的含义不同于原有数据,但包含了原有数据的大部分特征,并且具有较低的维度,便于后续进一步分析。

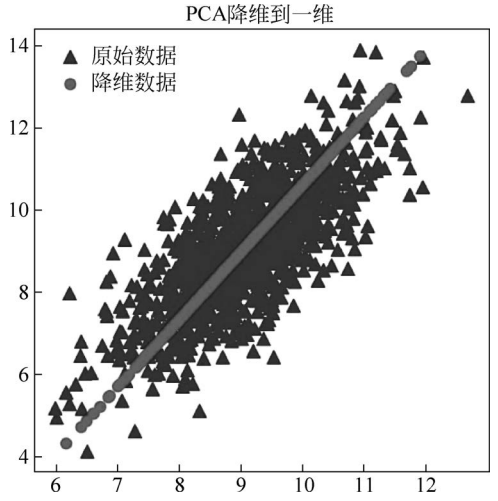

图 5-34　使用主成分分析对数据降维

图 5-34 所示为使用主成分分析对数据降维。

(2) 因子分析是从假设出发,假设所有自变量 x 出现的原因是背后存在一个潜变量 f(也就是所说的因子),在这个因子的作用下,x 可以被观察到。因子分析是通过研究变量间的相关系数矩阵,把这些变量间错综复杂的关系归结为少数几个综合因子,并据此对变量进行分类的一种统计分析方法。因子分析就是将原始变量转变为新的因子,这些因子之间的相关程度较低,而因子内部的变量相关程度较高。例如,一个学生考试,数学、化学、物理都考了满分,那么认为这个学生理性思维较强,理性思维就是一个因子,在这个因子的作用下,偏理科的成绩才会那么高。

因子分析法的主要目的,一是进行结构的探索,在变量之间存在高度相关性时,我们希望用较少的因子概括其信息;二是把原始变量转换为因子得分后,使用因子得分进行其他分析,从而简化数据,如聚类分析、回归分析等;三是通过每个因子得分计算出综合得分,对分析对象进行综合评价。

例如,在工业系统中常常通过部署传感器监控和收集系统运行的数据。传感器读数通常会有噪声,而且噪声的维度比常规数据高得多。因此,可以采用自动编码器去除噪声,降低维度。

3) SVM

支持向量机(Support Vector Machine,SVM)是一种支持线性分类和非线性分类的二元分类算法,经过演进现在也支持多元分类,目前被广泛地应用在回归以及分类当中。SVM 于

1963年由瓦普尼克等提出，解决了传统方法中遇到的问题，可以很好地解决非线性、小样本和高维的问题，并且根据实践检验，SVM在这些方面都表现出了良好的性能。在实际应用中，支持向量机不仅可用于二分类，也可用于多分类。支持向量机在垃圾邮件处理、图像特征提取及分类、空气质量预测等多方面领域都有应用。因此，支持向量机已成为机器学习领域中的不可缺少的一部分。

支持向量机主要分为线性可分支持向量机、线性不可分支持向量机和非线性支持向量机这三大类。线性可分支持向量机指在二维平面内可以用一条线清晰地分开两个数据集；线性不可分支持向量机指在二维平面内用一条线分开两个数据集时会出现误判点；非线性支持向量机指用一条线分开两个数据集时会出现大量误判点，此时需要采取非线性映射将二维平面扩展为三维立体，然后寻找一个平面清晰地切开数据集。

支持向量机的原理可以简单地描述为对样本数据进行分类，实际是对决策函数进行求解。首先，要找到分类问题中的最大分类间隔，然后确定最优分类超平面，并将分类问题转化为二次规划问题进行求解。图5-35所示为线性可分支持向量机，关于超平面的定义如下。

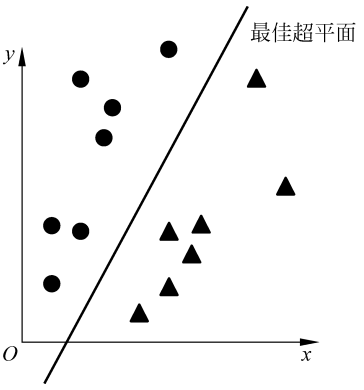

图5-35　线性可分支持向量机

(1) 在二维空间中，两类点被一条直线完全分开叫作线性可分。线性可分严格的数学定义如下。

D_1 和 D_2 为 n 维欧氏空间中的两个点集。如果存在 n 维向量 $\boldsymbol{w}$ 和实数 b，使得所有属于 D_1 的点 $\boldsymbol{x}_1$ 都有 $\boldsymbol{w}\boldsymbol{x}_1+b>0$，所有属于 D_2 的点 $\boldsymbol{x}_2$ 都有 $\boldsymbol{w}\boldsymbol{x}_2+b<0$，则称 D_1 和 D_2 线性可分。

(2) 从二维扩展到多维空间中时，将 D_1 和 D_2 完全正确地划分开的 $\boldsymbol{w}\boldsymbol{x}+b=0$ 就成为一个超平面。

(3) 为了使这个超平面更具鲁棒性，需要寻找最佳超平面，即以最大间隔把两类样本分开的超平面，也称为最大间隔超平面。其特点是两类样本分别分布在该超平面的两侧，两侧距离超平面最近的样本点到超平面的距离被最大化了。

在工业生产中，可以通过训练和操作支持向量机，分析产品内部缺陷检测的性能。

4) 决策树算法

决策树是应用最广的归纳推理算法之一，它是一种逼近离散值函数的方法，对噪声数据有很好的健壮性且能够学习析取表达式。决策树算法搜索一个完整表示的假设空间，从而避免了受限假设空间的不足，决策树学习的归纳偏置是优先选择较小的树。

通过决策树学习到的函数被表示为一棵决策树，学习得到的决策树也能再被表示为多个决策树选择的规则以提高可读性。决策树算法是最流行的归纳推理算法之一，已经被成功地应用到从学习医疗诊断到学习评估贷款申请的信用风险等的广阔应用领域中。

一个典型的决策树示例如图5-36所示，用于预测贷款用户是否具有偿还贷款的能力。贷款用户主要具备是否拥有房产、是否结婚和平均月收入这3个属性。每个内部节点都表示一个属性条件判断，叶子节点表示贷款用户是否具有偿还能力。例如，用户甲没有房产，没有结婚，月收入5000元。通过决策树的根节点判断，用户甲符合右边分支(拥有房产为“否”)；再判断是否结婚，用户甲符合左边分支(是否结婚为“否”)；然后判断月收入是否大于4000元，用户甲符合左边分支(月收入大于4000元)，该用户落在“可以偿还”的叶子节点上，所以预测用户甲具备偿还贷款的能力。

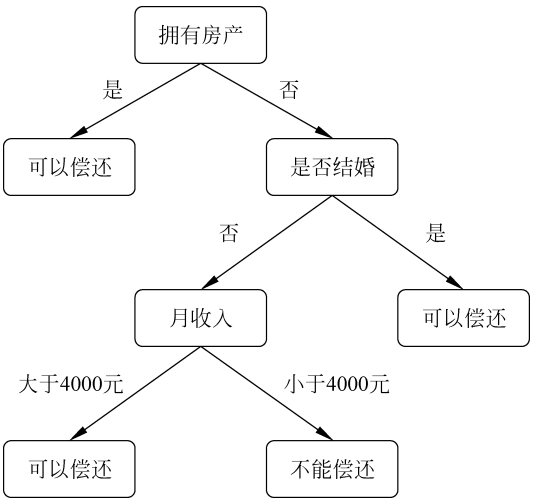

图 5-36 决策树示例

决策树通过把实例从根节点排列(Sort)到某个叶子节点来分类实例,叶子节点即为实例所属的分类。树上的每个节点说明对实例的某个属性(Attribute)的测试,并且该节点的每个后继分支对应于该属性的一个可能值。分类实例的方法是从这棵树的根节点开始,测试这个节点指定的属性,然后按照给定实例的该属性值对应的树枝向下移动,这个过程在以新节点为根的子树上重复。

决策树是附加概率结果的一个树状的决策图,是直观地运用统计概率分析的图法。机器学习中的决策树是一个预测模型,它表示对象属性和对象值之间的一种映射,树中的每个节点表示对象属性的判断条件,其分支表示符合节点条件的对象。树的叶子节点表示对象所属的预测结果。

决策树算法可以应用于很多领域,如根据地理位置预测产品的需求量、根据疾病分类患者、根据起因分类设备故障、根据拖欠支付的可能性分类贷款申请。对于这些问题,核心任务都是要把样例分类到各可能的离散值对应的类别中,因此这些问题经常被称为分类问题。

例如,在工业制造中,机床最核心的问题就是刀具问题。实际上,刀具之于机床就如同牙齿之于人类,只有在刀具发生问题前感知到,才能第一时间去修复。如果在问题发生之后再去修复则意义不大,也会给企业生产造成难以挽回的损失。某公司通过控制器收集了不同机床运行的数据,包括电流、电压等,可以使用决策树算法建立数据模型,预测设备使用多久后会发生故障。

5) 关联规则算法

关联规则算法是一种很重要的数据挖掘的知识模式。1993 年,Agrawal 等率先提出关联规则的概念,关联规则是数据中一种简单但具有很大实际意义的规则。关联规则算法常用来描述数据之间相关关系的描述型模式,挖掘关联规则的算法和聚类算法类似,属于无监督学习的方法。

关联规则是反映一个事物与其他事物之间的相互依存性和关联性,是数据挖掘的一个重要技术,用于从大量数据中挖掘出有价值的数据项之间的相关关系。

关联规则的定义是:两个不相交的非空集合 X、Y,如果有 $X \rightarrow Y$,就说 $X \rightarrow Y$ 是一条关联规则。其中,X 表示的是两个互斥事件,X 称为前因(Antecedent),Y 称为后果(Consequent),上述关联规则表示 X 会导致 Y。关联规则的强度用支持度(Support)和置信度(Confidence)

描述。其中，支持度表示 X 和 Y 同时出现的概率，置信度表示 X 和 Y 同时出现的概率与 X 出现概率的比值。支持度和置信度越高，说明规则越强，关联规则挖掘就是挖掘出满足一定强度的规则。例如，在商场的购物数据中，常常可以看到多种物品同时出现，这背后隐藏着联合销售或打包销售的商机，在大数据分析中的关联规则分析（Association Rule Analysis）就是为了发掘购物数据背后的商机而诞生的。

由关联规则定义可知，任意事务中的两个项集，都可以通过算法挖掘出关联规则，只不过挖掘出的关联规则在属性值上不尽相同。如前所述，在关联规则中通常用支持度和置信度这两个属性值直接描述关联规则的性质。在挖掘关联规则过程中，如果不考虑支持度和置信度阈值，就会从数据库中寻找到无穷多的关联规则。但实际生活中，需要有实际意义的关联规则体现数据隐含的规律。因此，为了更好地挖掘出有实际意义的关联规则，需要为这两个值事先设定一个最小值，即最小支持度和最小置信度。挖掘出的关联规则必须满足最小支持度和最小置信度，通常情况下把同时满足这两个要求的规则称为强关联规则。

关联规则挖掘的过程主要包含：第 1 阶段，必须从数据集中找到所有频繁项集；第 2 阶段，再从这些频繁项集中产生强关联规则。挖掘的第 1 阶段必须要在原始数据集中进行，目的是找出所有频繁项集。当某一项目出现的频率相对于其他项目而言是“频繁”的，就将其称为频繁项集。项目组出现的频率称为支持度，以包含 A 与 B 两个项目的 2-项集为例，可以求得包含$\{A,B\}$项目组的支持度，若支持度大于最小支持度阈值，则$\{A,B\}$称为频繁项集，一个满足最小支持度的 K-项集被称为频繁 K-项集。关联规则的第 2 阶段是产生强关联规则，利用第 1 阶段所得到的频繁项集产生规则，在最小置信度阈值下，若一个规则的置信度满足最小置信度，则称为强关联规则。

用于挖掘关联规则的主要算法有 Apriori 算法、FP-Growth 算法和基于划分的关联规则算法。

（1）Apriori 算法。关联规则问题是数据挖掘领域的一个最基本、最重要的问题，其可以通俗地理解为两个或多个项之间的描述。由于生活中很多事物的联系并不能精确地表示，于是出现了以概率统计为基础的经典算法，Apriori 算法就是其中最具影响力的算法。Apriori 算法是以两阶段频集思想递推算法为核心的，把所有满足最小支持度阈值的项集称为频繁项集，简称为频集。Apriori 算法是最有影响力的挖掘布尔关联规则频繁项集的算法，挖掘出的关联规则属于单维、单层、布尔型的关联规则。Apriori 算法的基本思想是：首先在原始数据集找出所有频繁项集，这些项集的频繁性至少满足事先定义的最小支持度阈值。然后使用第 1 步寻找到的频繁项集生成关联规则，剔除其中不满足最小置信度阈值的关联规则，剩下的关联规则就是同时满足最小支持度和最小置信度阈值的强关联规则。

（2）FP-Growth 算法。Apriori 算法虽然简单准确，但因其需要多次迭代生成大量的候选项集，所以在效率上存在一定缺陷，Han 等提出了一种利用频繁模式树（FP-Tree）进行频繁模式挖掘的 FP-Growth 算法，这种算法不会产生候选项集。算法在第 1 遍扫描之后，先将数据库中的频繁项集生成为一棵频繁模式树，并且保留数据之间的关联信息，再将这棵频繁模式树分化为若干个条件库，其中每个库都有一个长度为 1 的频繁项集与之对应，最后再分别挖掘这些条件库寻找频繁项集。该算法使用的是一种典型的“分而治之”的策略。如果原始数据量很大，可以使用划分的方法，使一棵庞大的频繁模式树同样可以放入主存储器中。FP-Growth 算法不但具有 Apriori 算法的准确性和良好的适应性，同时还有效地解决了 Apriori 算法存在的效率缺陷。

（3）基于划分的关联规则算法。基于划分的关联规则算法先从逻辑上将数据库分为几个互不相交的分块，每次只对一个分块的数据进行独立分析，生成分块中所有频繁项集，然后把

所有分块中产生的频繁项集汇总，得到可能的频繁项集，最后计算这些项集在整个数据库中的支持度，一次生成所有频繁项集。在划分时要限制分块的大小，至少要保证每个分块都能成功地放入主存储器中。因为每个局部频繁项集都能保证在某个分块中是频繁的，所以算法的正确性得以保证。划分算法是可以高度并行的，可以为每个分块都分配一个独立的处理器用于生成频繁项集。当一个循环结束后，寻找到了每个分块的局部的频繁项集，处理器之间就会以通信的方式产生全局候选项集，即可能的频繁项集。然而，在实际应用中，通信过程和每个独立处理器生成频繁项集的时间差异往往是限制算法执行效率的主要瓶颈。

随着关联规则挖掘技术的不断进步，关联规则已经在各行各业中广泛应用，如国内外的知名电商、银行的理财服务等都从关联规则算法中受益。电商网站分析用户的购买信息，挖掘出其中潜在的关联规则，然后根据关联规则的指导设置相应的交叉销售，即购买一件商品时推荐一些类似的商品，或者将多个具有强相关的商品进行捆绑销售。金融行业企业中，基于挖掘出的关联规则，银行可以成功地预测客户需求，改善自身营销方式，为客户提供合适的理财产品。例如，在 ATM 机或手机 App 应用上根据客户的行为信息，宣传银行的相应产品供用户了解，推动产品的购买量。目前关联规则挖掘的应用正进一步向医疗等领域扩展。

6）朴素贝叶斯算法

贝叶斯算法是统计模型决策中的一个基本方法，其基本思想是已知条件概率密度参数表达式和先验概率，利用贝叶斯公式转换为后验概率，再根据后验概率大小进行决策分类。贝叶斯是一种使用先验概率进行处理的模型，其最后的预测结果就是具有最大概率的那个类，在概率的计算中，贝叶斯算法是一个很重要的算法。

在贝叶斯分类过程中，属性的选择对分类结果很重要，用不同的属性分类出来的结果会有差别。朴素贝叶斯的一个特点是条件独立性，也就是说，在使用朴素贝叶斯算法进行分类时，不考虑属性之间的任何联系，可以将问题简单化。贝叶斯算法主要用于计算概率以完成分类及预测等问题，如新闻、文本以及病人等各种情况的分类及预测等。

朴素贝叶斯算法是最常用的一种贝叶斯算法，它是基于贝叶斯公式建立的，朴素贝叶斯算法计算式为

$$P(A \mid B)=\frac{P(B \mid A)P(A)}{P(B)}$$

其中，A 和 B 是两个事件；$P(A)$ 为没有前提条件时 A 发生的概率，其结果是一个常数；$P(B)$ 为没有前提条件时 B 发生的概率，其结果同样是一个常数；$P(A|B)$ 为 B 确定已经发生时 A 发生的概率。朴素贝叶斯公式的“朴素”二字是基于一种假定——所有特征都是独立的，只有满足了这个假定才能使用朴素贝叶斯算法。朴素贝叶斯模型原理可以概率为当一个样本有可能属于多个类别时，简单地选择其中概率最大的那个。

朴素贝叶斯算法主要用于分类问题，如新闻分类、文本分类、病人分类、邮件分类等。例如，在企业中如果已经收集了大量垃圾邮件和非垃圾邮件，则可以使用朴素贝叶斯算法过滤垃圾邮件。此外，在工厂生产中还可以使用并行高斯分布朴素贝叶斯分类算法处理大规模连续型数据。

企业的核心问题是解决和提高资源配置效率。大数据支撑企业决策，就是将正确的数据在正确的时间以正确的方式传递给正确的人和机器。通过工业大数据分析可以建立从局部到全局、从建模到决策的层级化数据分析，发现数据中隐藏的规律，形成可视化图表，预测和分析未知错误和潜在问题，以帮助企业实现智能决策、智能诊断、智能调度、智能预测以及智能设计等。图 5-37 所示为工业大数据分析在工业大数据中的作用。

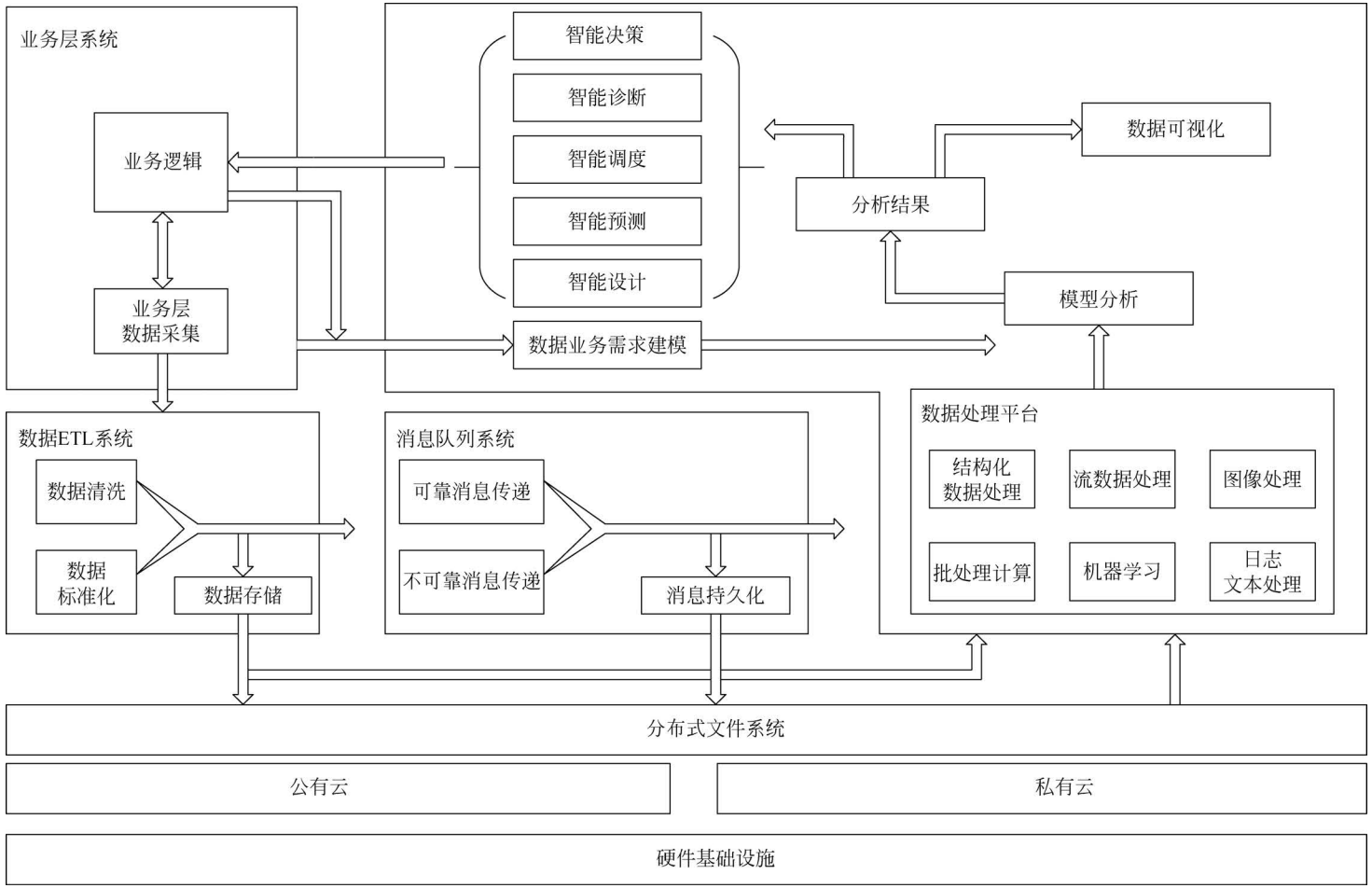

图 5-37　工业大数据分析在工业大数据中的作用

5.3.5 工业大数据可视化

让大数据有意义，使之更贴近大多数人，最重要的手段之一就是数据可视化。通过增加数据可视化使用，企业能够发现其追求的价值。例如，通过三维可视化技术将整个工厂环境和生产设备进行三维呈现，对整个生产过程进行虚拟仿真，结合不断进步的物联网技术和监控技术，真正帮助企业从数字化生产迈向智慧工厂。图 5-38 和图 5-39 所示为三维可视化技术在工业生产中的应用。

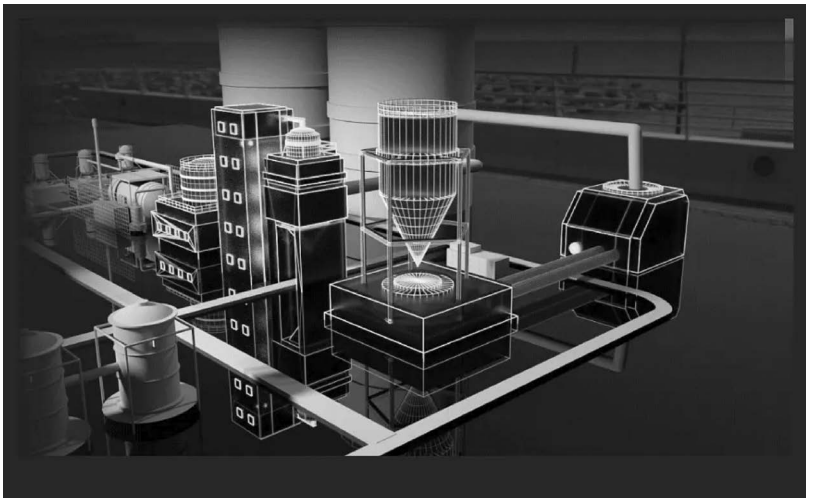

图 5-38 三维可视化技术在工业生产中的应用(1)

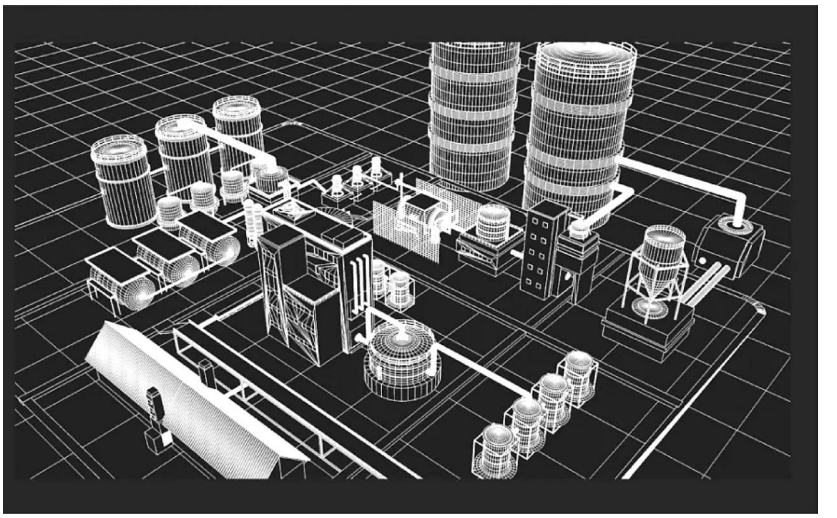

图 5-39 三维可视化技术在工业生产中的应用(2)

此外，在工业可视化中，大屏数据可视化也是常见的实施方案。大屏数据可视化是以大屏为主要展示载体的数据可视化设计。大屏具有面积大、可展示信息多的特点，因此可以通过酷炫的动画效果、色彩丰富的可视化设计给人留下震撼的观感印象，营造仪式感。此外，设计团队或部门决策层也可通过关键信息大屏共享的方式进行讨论和决策，因此在数据分析监测中也常常使用到大屏数据可视化技术。图 5-40 所示为云计算服务监控大屏。

图 5-40　云计算服务监控大屏

5.4　工业大数据治理

5.4.1　工业大数据治理概述

1. 认识工业大数据治理

1）工业大数据治理简介

工业大数据是工业互联网价值实现的核心要素，工业互联网以数据为核心要素实现全要素、全产业链、全价值链的全面连接，同时以数据驱动实现从感知控制到决策优化的闭环反馈。大数据治理可以为业务提供持续的、可度量的价值。工业界 IBM 数据治理委员会给数据治理的定义如下：数据治理是一组流程，用来改变组织行为，利用和保护企业数据，将其作为一种战略资产。

数据治理是指从使用零散数据变为使用统一数据、从具有很少或没有组织流程到企业范围内的综合数据管控、从数据混乱状况到数据井井有条的一个过程。数据治理强调的是一个从混乱到有序的过程。从范围来讲，数据治理涵盖了从前端业务系统、后端业务数据库再到业务终端的数据分析，从源头到终端再回到源头，形成一个闭环负反馈系统。具体来讲，数据治理就是以服务组织战略目标为基本原则，通过组织成员的协同努力、流程制度的制定，以及数据资产的梳理、采集清洗、结构化存储、可视化管理和多维度分析，实现数据资产价值获取、业务模式创新和经营风险控制的过程。

2）工业大数据治理发展

在企业发展初期，数据研发模式一般紧贴业务的发展而演变，数据体系也是基于业务单元垂直建立，不同的垂直化业务带来不同的烟囱式的体系。但随着企业的发展，一方面，数据规模在快速膨胀，垂直业务单元也越来越多；另一方面，基于大数据的业务所需要的数据不仅仅是某个垂直单元的，使用数据类型繁多的数据才能具备核心竞争力。跨垂直单元的数据建设接踵而至，混乱的数据调用和复制、重复建设带来的资源浪费、数据指标定义不同而带来的歧义、数据使用门槛越来越高等，这些问题日益凸显，解决这些问题促进企业发展迫在眉睫。因

此，使用底层平台（数据库）数据存储也经历了不同的发展阶段，从层次、网状模型到关系模型，从单机到集群，从单体架构到集群架构，从线下到云端，以此满足对数据的承载能力、使用特点差异。

针对以上情况，作为数据使用的上层建筑，数据治理逐渐受到企业的高度关注。这主要是因为：一方面，数据的多源、异构、价值差异等特点导致复杂度提高；另一方面，数据价值正在被更多的企业所关注。如何在企业内部用统一视角看待数据，让数据在企业中存好用好，发挥出更大价值，是企业数字化转型必然面临的问题。数据治理正是解决这一问题的利器。过去，数据治理往往在高价值数据集中且规范程度较高的企业（如金融业）受到重视，但现在更多的企业（包括互联网）也重视数据治理的建设。

2. 工业大数据治理的主要环节

数据治理不仅需要完善的保障机制，还需要理解具体的治理内容，如企业数据该如何进行规范、元数据又该如何管理、每个过程需要哪些系统或工具进行配合，这些问题都是数据治理过程中最实际的问题，也是最复杂的问题。因此，数据治理是专注于将数据作为企业的商业资产进行应用和管理的一套管理机制，它能够消除数据的不一致性，建立规范的数据应用标准，提高组织的数据质量，实现数据广泛共享，并能够将数据作为组织的宝贵资产应用于业务、管理、战略决策中，发挥数据资产的商业价值。

1）数据资产

随着大数据时代的来临，对数据的重视提到了前所未有的高度，“数据即资产”已经被广泛认可。数据就像企业的根基，是各企业尚待发掘的财富，即将被企业广泛应用。数据资产可定义为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、以物理或电子的方式记录的数据资源，如文件资料、电子数据等。不过，值得注意的是，在企业中，并非所有数据都构成数据资产，数据资产是能够为企业产生价值的数据资源。因此，只有那些能够给企业带来可预期经济收益的数据资源才能够被称为数据资产。而数据治理正是一门将数据视为一项企业资产的学科，是针对数据管理的质量控制规范，它将严密性和纪律性植入企业的数据管理、利用、优化和保护过程中，并涉及以企业资产的形式对数据进行优化、保护和利用的决策权利。

2）数据模型

数据模型是数据治理中的重要部分。理想的数据模型应该具有非冗余、稳定、一致、易用等特征。逻辑数据模型能涵盖整个集团的业务范围，以一种清晰的表达方式记录跟踪集团单位的重要数据元素及其变动，并利用它们之间各种可能的限制条件和关系表达重要的业务规则。为了满足将来不同的应用分析需要，数据模型必须在设计过程中保持统一的业务定义，逻辑数据模型的设计应该能够支持最细粒度的详细数据的存储，以支持各种可能的分析查询。同时，保障逻辑数据模型能够最大程度上减少冗余，并保障结构具有足够的灵活性和扩展性。

3）数据质量管理

数据质量管理已经成为企业数据治理的有机组成部分，完善的数据质量管理是保障各项数据治理工作能够得到有效落实，达到数据准确、完整的目标，并能够提供有效的增值服务的重要基础。高质量的数据是企业进行分析决策、业务发展规划的重要基础，只有建立完整的数据质量体系，才能有效提升数据整体质量，从而更好地为客户服务，提供更为精准的决策分析数据。

4）数据存储

企业只有对数据进行合理的存储，有效地提高数据的共享程度，才能尽可能地降低数据冗余带来的存储成本。数据存储作为大数据的核心环节之一，可以理解为方便对既定数据内容

进行归档、整理和共享的过程。

(1) 分布式文件系统。分布式文件系统是由多个网络节点组成的向上层应用提供统一的文件服务的文件系统。分布式文件系统中的每个节点可以分布在不同的地点,通过网络进行节点间的通信和数据传输。分布式文件系统中的文件在物理上可能被分散存储在不同的节点上,在逻辑上仍然是一个完整的文件。使用分布式文件系统时,无须关心数据存储在哪个节点上,只需要像本地文件系统一样管理和存储文件系统的数据。

(2) 文档存储。文档存储支持对结构化数据的访问,不同于关系模型的是,文档存储没有强制的架构。事实上,文档存储以封包键值对的方式进行存储。在这种情况下,应用对要检索的封包采取一些约定,或者利用存储引擎的能力将不同的文档划分成不同的集合,以管理数据。

(3) 列式存储。列式存储将数据按行排序、按列存储,将相同字段的数据作为一个列族聚合存储。当只查询少数列族数据时,列式数据库可以减少读取数据量,缩短数据装载和读入读出的时间,提高数据处理效率。按列存储还可以承载更大的数据量,获得高效的垂直数据压缩能力,降低数据存储开销。

(4) 键值存储。键值存储即 Key-Value 存储,简称为 KV 存储,它是 NoSQL 存储的一种方式。它的数据按照键值对的形式进行组织、索引和存储。KV 存储非常适合不涉及过多数据关系和业务关系的业务数据,同时能有效减少读写磁盘的次数,比 SQL 数据库存储拥有更好的读写性能。键值存储一般不提供事务处理机制。

(5) 图形数据库。图形数据库主要用于存储事物及事物之间的相关关系,这些事物整体上呈现复杂的网络关系,可以简单地称之为图形数据。使用传统的关系数据库技术已经无法很好地满足超大量图形数据的存储、查询等需求,如上百万或上千万个节点的图形关系,而图形数据库采用不同的技术能很好地解决图形数据的查询、遍历、求最短路径等需求。在图形数据库领域,有不同的图模型映射这些网络关系,如超图模型,以及包含节点、关系及属性信息的属性图模型等。图形数据库可用于对真实世界的各种对象进行建模,如社交图谱,以反映这些事物之间的相互关系。

(6) 关系数据库。关系模型是最传统的数据存储模型,它使用记录(由元组组成)按行进行存储,记录存储在表中,表由架构界定。表中的每列都有名称和类型,表中的所有记录都要符合表的定义。SQL 是专门的查询语言,提供相应的语法查找符合条件的记录,如表连接(Join)。表连接可以基于表之间的关系在多表之间查询记录。表中的记录可以被创建和删除,记录中的字段也可以单独更新。关系数据库通常提供事务处理机制,为涉及多条记录的自动化处理提供了解决方案。

5) 数据交换

数据交换是企业进行数据交互和共享的基础,合理的数据交换体系有助于企业提高数据共享程度和数据流转时效。从功能上讲,数据交换用于实现不同机构、不同系统之间数据或文件的传输和共享,提高信息资源的利用率,保证了分布在异构系统之间信息的互联互通,完成数据的收集、集中、处理、分发、加载、传输,构造统一的数据及文件的传输交换。在实施中,企业一般会对系统间数据的交换规则制定一些原则,如对接口、文件的命名、内容进行明确,规范系统间、系统与外部机构间的数据交换规则,指导数据交换工作有序进行。建立统一的数据交换系统,一方面可以提高数据共享的时效性,另一方面也可以精确掌握数据的流向。

6) 数据集成

数据集成是把不同来源、格式、特点性质的数据在逻辑上或物理上有机地集中,从而为企业提供全面的数据共享。数据集成的核心任务是要将互相关联的异构数据源集成到一起,使

用户能够以透明的方式访问这些数据资源。因此，数据集成可对数据进行清洗、转换、整合、模型管理等处理工作，它既可以用于问题数据的修正，也可以用于为数据应用提供可靠的数据模型。值得注意的是，在企业中并不是所有地方都要数据治理，数据治理只出现在需要干净数据、需要直观数据呈现的场景中。而数据集成正是把不同来源、格式、特点性质的数据在逻辑上或物理上有机地集中，从而为企业提供全面的数据共享。

7）数据服务

数据的管理和治理是为了更好地利用数据，是数据应用的基础。企业应该以数据为根本，以业务为导向，通过对大数据的集中、整合、挖掘和共享，实现对多样化、海量数据的快速处理及价值挖掘，利用大数据技术支持产品快速创新，提升以客户为中心的精准营销和差异化客户服务能力，增强风险防控实时性、前瞻性和系统性，推动业务管理向信息化、精细化转型，全面支持信息化和数字化的建设。

8）数据安全

企业的重要且敏感数据大部分集中在应用系统中，如客户的联络信息、资产信息等，如果不慎泄露，不仅给客户带来损失，也会给企业自身带来不利的声誉影响，因此数据安全在数据管理和治理过程中是相当重要的。数据安全主要提供数据加密、脱敏、模糊化处理、账号监控等各种数据安全策略，确保数据在使用过程中有恰当的认证、授权、访问和审计等措施。

5.4.2 工业大数据治理核心内容

1. 主数据与元数据管理

1）主数据

主数据是用来描述企业核心业务实体的数据，它是具有高业务价值的、可以在企业内跨越各个业务部门被重复使用的数据，并且存在于多个异构的应用系统中。主数据可以涵盖很多方面，除了常见的客户主数据之外，不同行业的客户还可能拥有其他各种类型的主数据。例如，对于电信行业客户，电信运营商提供的各种服务可以形成其产品主数据；对于航空业客户，航线、航班是其企业主数据的一种。对于某个企业的不同业务部门，其主数据也不同，如市场销售部门关心客户信息，产品研发部门关心产品编号、产品分类等产品信息，人事部门关心员工结构、部门层次关系等信息。

（1）主数据管理。主数据通常需要在整个企业范围内保持一致性（Consistent）、完整性（Complete）、可控性（Controlled），为了达成这一目标，就需要进行主数据管理（Master Data Management，MDM）。集成、共享、数据质量、数据治理是主数据管理的四大要素。主数据管理要做的就是从企业的多个业务系统中整合最核心的、最需要共享的数据（主数据），集中进行数据的清洗和丰富，并且以服务的方式把统一的、完整的、准确的、具有权威性的主数据分发给全企业范围内需要使用这些数据的操作型应用和分析型应用，具体包括各个业务系统、业务流程和决策支持系统等。

一方面，MDM 可以保障主数据的规范性和唯一性。按规则和流程规范管理主数据，如规定主数据名称要使用营业执照上的名称、社会统一信用代码等条件校验，系统内编码唯一，主数据要经流程审核后方能生效等。另一方面，MDM 使主数据能够集中管理。主数据全部在 MDM 中产生或受控，保障来源唯一从而避免歧义。同时，MDM 能够把主数据分发给相关系统，也可以接收外部系统产生的主数据，经处理后再分发出去。

（2）主数据管理平台。主数据是企业最基础、最核心的数据，企业的一切业务基本都是基于主数据开展的，是企业最重要的数据资产。所以，主数据管理也使企业数据治理成为最核心部分。

为了更好地管理主数据，企业常常需要建设主数据管理平台，该平台从功能上主要包括主数据模型、主数据编码、主数据管理、主数据清洗、主数据质量、主数据集成等。

① 主数据模型：提供主数据的建模功能，管理主数据的逻辑模型和物理模型以及各类主数据模板。

② 主数据编码：编码功能是主数据产品的初级形态，也是主数据产品的核心能力，平台应当支持各种形式主数据的编码，提供数据编码申请、审批、集成等服务。

③ 主数据管理：主要提供主数据的增、删、改、查功能。

④ 主数据清洗：主要包括主数据的采集、转换、清理、装载等功能。

⑤ 主数据质量：主要提供主数据质量从质量问题发现到质量问题处理的闭环管理功能。

⑥ 主数据集成：主要提供主数据采集和分发服务，完成与企业其他异构系统的对接。

2）元数据

元数据是描述企业数据的相关数据（包括对数据的业务、结构、定义、存储、安全等各方面对数据的描述），一般是指在 IT 系统建设过程中所产生的有关数据定义、目标定义、转换规则等相关的关键数据，在数据治理中具有重要的地位。元数据不仅表示数据的类型、名称、值等信息，它可以理解为是一组用来描述数据的信息组/数据组，该信息组/数据组中的一切数据、信息都描述或反映了某个数据的某方面特征，则该信息组/数据组可称为一个元数据。例如，元数据可以为数据说明其元素或属性（名称、大小、数据类型等）、结构（长度、字段、数据列）、相关数据（位于何处、如何联系、拥有者）。

（1）元数据管理模型。元数据管理是构建企业信息单一视图的重要组成部分，元数据管理可以保证在整个企业范围内跨业务竖井协调和重用主数据。元数据管理不会创建新的数据或新的数据纵向结构，而是提供一种方法使企业能够有效地管理分布在整个信息供应链中的各种主数据（由信息供应链各业务系统产生）。元数据管理一直比较困难，一个很重要的原因就是缺乏统一的标准。在这种情况下，各公司的元数据管理解决方案各不相同。近几年，随着元数据联盟（Meta Data Coalition，MDC）的开放信息模型（Open Information Model，OIM）和对象管理组织（Object Management Group，OMG）的公共仓库模型（Common Warehouse Model，CWM）标准的逐渐完善，以及 MDC 和 OMG 组织的合并，为数据仓库厂商提供了统一的标准，从而为元数据管理铺平了道路。

（2）元数据集成体系结构。元数据集成体系结构涉及多个概念，如元模型、元-元模型、公共仓库元模型等。值得注意的是，统一完整的元数据管理，特别是清晰的主题域划分、完善的元模型和元-元模型有利于更好地管理主数据。

① 元模型。元模型（Meta Model）也就是模型的模型（或元-元数据），是用来描述元数据的模型。元模型的使用目的在于：识别资源；评价资源；追踪资源在使用过程中的变化；简单、高效地管理大量网络化数据；实现信息资源的有效发现、查找、一体化组织和对使用资源的有效管理。

② 元-元模型。元-元模型就是元模型的模型，有时也被称为本体，是模型驱动的元数据集成体系结构的基础，其定义了描述元模型的语言，规定元模型必须依照一定的形式化规则建立，以便所有软件工具都能够对其进行理解。

③ 公共仓库元模型。公共仓库元模型是被 OMG 采纳的数据仓库和业务分析领域元数据交换开放式行业标准，在数据仓库和业务分析领域为元数据定义公共的元模型和基于可扩展标记语言（Extensible Markup Language，XML）的元数据交换（XML Metadata Interchange，XMI）。CWM 作为一个标准的接口，可以帮助分布式、异构环境中的数据仓库工具与数据仓库平台和

数据仓库元数据存储库之间轻松实现数据仓库和业务分析元数据交换。CWM 提供一个框架为数据源、数据目标、转换、分析、流程和操作等创建和管理元数据，并提供元数据使用的世系信息。因此，CWM 实际上就是一个元数据交换的标准，为各种数据仓库产品提出的一个标准。

2. 数据质量与数据管理

数据无处不在，它贯穿整个数据生命周期，为企业决策提供了可靠的基础支撑，是企业成功的关键。在大数据时代，随着企业数据规模的不断扩大，数据数量的不断增加以及数据来源复杂性的不断变化，为了能够充分地利用数据价值，企业需要对数据进行管理。

1）ISO 8000 数据质量标准

ISO 8000 数据质量标准是针对数据质量制定的国际标准化组织标准，致力于管理数据质量，具体来说，包括规范和管理数据质量活动、数据质量原则、数据质量术语、数据质量特征（标准）和数据质量测试。根据 ISO 8000 数据质量标准的要求，数据质量高低程度由系统数据与明确定义的数据要求进行对比而得到。通过 ISO 8000 的标准规范，可以保证用户在满足决策需求和数据质量的基础上，在整个产品或服务的周期内高质量地交换、分享和存储数据，从而保证用户可以依托获取的数据高效地作出最优化的安全决策。

通过将 ISO 8000 标准应用于组织内部，可以对组织内数据进行规范化整合和管理，对各个部门的数据进行统一识别和管理，从组织的整体层面进行资源与信息的协调管理，从而降低因为信息沟通不畅带来的运营成本。此外，如果在合作公司之间或整个行业采用 ISO 8000 标准，数据或信息将会更有可用性。例如，在医疗卫生领域，各个医疗机构的信息系统不能很好地兼容，导致同一病人在不同医院的信息无法快速共享和传递。通过全国范围内应用 ISO 8000 数据质量标准，可以将病历信息与特定信息系统分离，病历的所有信息可以独立于医疗信息系统存在，并可被任意一个根据 ISO 8000 数据质量标准的信息系统读取，患者可以更加自主地选择就医医院，而不用担心由于自身的健康信息缺失导致的医疗误判。

2）数据质量管理

数据价值的成功发掘必须依托于高质量的数据，唯有准确、完整、一致的数据才有使用价值。因此，需要从多维度分析数据的质量，如偏移量、非空检查、值域检查、规范性检查、重复性检查、关联关系检查、离群值检查、波动检查等。需要注意的是，优秀的数据质量模型的设计必须依赖于对业务的深刻理解，在技术上也推荐使用大数据相关技术保障检测性能和降低对业务系统的性能影响，如 Hadoop、MapReduce、HBase 等。

数据质量管理是指对数据从计划、获取、存储、共享、维护、应用、消亡生命周期的每个阶段中可能引发的各类数据质量问题，进行识别、度量、监控、预警等一系列管理活动，并通过改善和提高组织的管理水平使数据质量获得进一步提高。数据质量管理是企业数据治理一个重要的组成部分，企业数据治理的所有工作都是围绕提升数据质量目标而开展的。

不过，值得注意的是，在数据治理方面，不论是国际的还是国内的，人们能找到很多数据治理成熟度评估模型这样的理论框架，作为企业实施的指引。而说到数据质量管理的方法论，其实业内还没有一套科学、完整的数据质量管理的体系。因为数据质量管理不单纯是一个概念，不单纯是一项技术，也不单纯是一个系统，更不单纯是一套管理流程，数据质量管理是一个集方法论、技术、业务和管理为一体的解决方案。通过有效的数据质量控制手段，进行数据的管理和控制，消除数据质量问题，进而提升企业数据变现的能力。

3）数据周期管理

数据生命周期从数据规划开始，中间是一个包括设计、创建、处理、部署、应用、监控、存档、销毁这几个阶段并不断循环的过程。企业的数据质量管理应贯穿数据生命周期的全过程，覆盖数

据标准的规划设计、数据的建模、数据质量的监控、数据问题诊断、数据清洗、优化完善等各方面。

以典型的工业生产设备资产为例，如图 5-41 所示，其全生命周期一般包括 6 个环节：设计、采购、安装、运行、维护和报废。从设备设计、采购开始，直至设备运行、维护、报废进行全生命周期管理；将基建期图纸、采购、资料信息记录到设备台账中，实现对设计数据、采购数据、施工数据、安装数据、调试数据等后期移交和设备系统生产运维所需要的完整数据平滑过渡，实现基建、生产一体化，提高企业资产利用率和企业投资回报率。同时，结合成本管理、财务管理，既实现对资产过程管控，更实现对资产价值的管理。

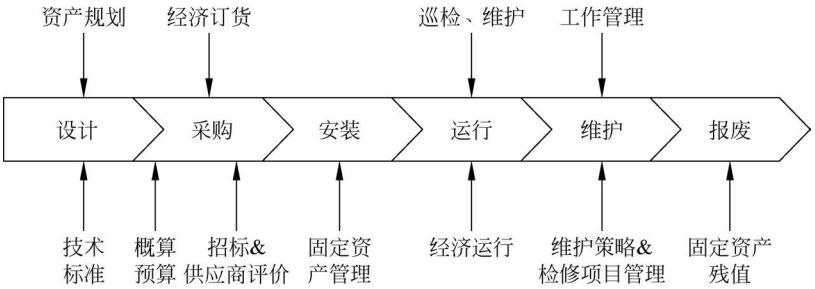

图 5-41　设备资产生命周期

在数据全周期管理中，最重要的几方面为数据规划、数据设计、数据创建和数据使用。

（1）数据规划。从企业战略的角度不断完善企业数据模型的规划，把数据质量管理融入企业战略中，建立数据治理体系，并融入企业文化。

（2）数据设计。推动数据标准化制定和贯彻执行，根据数据标准化要求统一建模管理，统一数据分类、数据编码、数据存储结构，为数据的集成、交换、共享、应用奠定基础。

（3）数据创建。利用数据模型保证数据结构完整、一致，执行数据标准、规范数据维护过程，加入数据质量检查，从源头系统保证数据的正确性、完整性、唯一性。

（4）数据使用。利用元数据监控数据使用；利用数据标准保证数据正确；利用数据质量检查加工正确。元数据提供各系统统一的数据模型进行使用，监控数据的来源去向，提供全息的数据地图支持；企业从技术、管理、业务 3 方面进行规范，严格执行数据标准，保证数据输入端的正确性；数据质量提供了事前预防、事中预警、事后补救的 3 方面措施，形成完整的数据治理体系。

要做好数据质量的管理，应抓住影响数据质量的关键因素，设置质量管理点或质量控制点，从数据的源头抓起，从根本上解决数据质量问题。在企业的数据治理中，数据质量管理必须识别相应产品规范或用户需求中的质量信息，在元数据、质量评价报告中形成正确的质量描述，并且这些规范上的质量结果均要为“合格”。

3. 数据标准

标准是指为了在一定的范围内获得最佳秩序，经协商一致制定并由公认机构批准，共同使用的和重复使用的一种规范性文件。数据标准是指对数据的表达、格式及定义的一致约定，包括数据业务属性、技术属性和管理属性的统一定义。其中，业务属性包括中文名称、业务定义、业务规则等；技术属性包括数据类型、数据格式等；管理属性包括数据定义者、数据管理者等。因此，对于数据标准的定义，通俗地讲，就是给数据一个统一的定义，让各系统的使用人员对同一指标的理解是一样的。

数据标准对于企业来说是非常重要的。因为大数据时代数据应用分析项目特别多，如果数据本身存在非常严重的问题，如数据统计口径不统一、数据质量参差不齐、数据标准不统一

等，往往会影响到项目正常交付，甚至后续数据应用和战略决策。在整个项目实施过程中，应用系统之间需要上传下达、信息共享、集成整合、协同工作。如果没有数据标准，会严重影响企业的正常运行。因此，在大数据行业中，对数据全生命周期进行规范化管理，可以从根本上解决诸多的数据问题。

1）数据标准的分类

数据标准是进行数据标准化、消除数据业务歧义的主要参考依据。数据标准的分类是从更有利于数据标准的编制、查询、落地和维护的角度进行考虑的。数据标准一般包括 3 个要素：标准分类、标准信息项（标准内容）和相关公共代码（如国别代码、邮政编码）。数据标准通常可分为基础类数据标准和指标类数据标准。

2）数据标准管理

数据标准管理是指数据标准的制定和实施的一系列活动，关键活动具体如下。

（1）理解数据标准化需求。

（2）构建数据标准体系和规范。

（3）规划制定数据标准化的实施路线和方案。

（4）制定数据标准管理办法和实施流程要求。

（5）建设数据标准管理工具，推动数据标准的执行落地。

（6）评估数据标准化工作的开展情况。

数据标准管理的目标是通过统一的数据标准制定和发布，结合制度约束、系统控制等手段，实现大数据平台数据的完整性、有效性、一致性、规范性、开放性和共享性管理，为数据资产管理活动提供参考依据。

3）数据标准建设的好处

通过数据标准的建设，可以有效消除数据跨系统的非一致性，从根源上解决数据定义和使用的不一致问题，为企业数据建设带来诸多好处。

（1）数据标准的统一制定与管理，可保证数据定义和使用的一致性，促进企业级单一数据视图的形成，促进信息资源共享。

（2）通过评估已有系统标准建设情况，可及时发现现有系统标准的问题，支撑系统改造，减少数据转换，促进系统集成，提高数据质量。

（3）数据标准可作为新建系统参考依据，为企业系统建设整体规划打好基础，减少系统建设工作量，保障新建系统完全符合标准。

4. 数据治理框架

要实现工业大数据治理，数据治理框架必不可少。目前国内外常见的数据治理框架有国际标准化组织 ISO 38500 治理框架、国际数据管理协会数据治理框架、国际数据治理研究所数据治理框架、IBM 数据治理框架、DCMM 数据治理框架以及 ISACA 数据治理框架等。

1）国际标准化组织 ISO 38500 治理框架

国际标准化组织于 2008 年推出第 1 个 IT 治理的国际标准——ISO 38500，它的出现标志着 IT 治理从概念模糊的探讨阶段进入了正确认识的发展阶段，而且也标志着信息化正式进入 IT 治理时代。ISO 38500 提出了 IT 治理框架（包括目标、原则和模型），并认为该框架同样适用于数据治理领域。

在目标方面，ISO 38500 认为 IT 治理的目标就是促进组织高效、合理地利用 IT。在原则方面，ISO 38500 定义了 IT 治理的 6 个基本原则：职责、策略、采购、绩效、符合和人员行为，这些原则阐述了指导决策的推荐行为，每个原则描述了应该采取的措施，但并未说明如何、何时

及由谁实施这些原则。在模型方面，ISO 38500 认为组织的领导者应重点关注 3 项核心任务：一是评估现在和将来的 IT 利用情况；二是对治理准备和实施的方针和计划作出指导；三是建立“评估→指导→监控”的循环模型。

2）国际数据管理协会数据治理框架

国际数据管理协会(DAMA International)成立于 1988 年，借助其丰富的数据管理经验，提出了最为完整的数据治理体系。

DAMA 数据治理的核心逻辑可以概括如下：在商业驱动因素下，从数据治理的输入端(Input)，到主要的活动(Activities)，再到主要的交付成果。在此过程中，需要首先明确数据治理过程对供应方、参与方与消费者的影响，并在每个数据治理的模块上，都认真地思考商业价值导向与目标导向，最终才形成可以实施的数据治理的可行方案。因此，尽管数据治理的 DAMA 体系非常复杂，但商业价值驱动目标导向是 DAMA 体系的最大特点。理解数据治理的商业驱动，有利于在数据治理时保证正确的方向，使数据治理真正服务于企业的经营，服务于企业市场竞争能力的提升，从而使数据化转型不能只为转型而转型，必须服务于企业战略。

此外，DAMA 认为数据治理是对数据资产管理行使权力和控制，包括规划、监控和执行。它还对数据治理和 IT 治理进行了区分：IT 治理的对象是 IT 投资、IT 应用组合和 IT 项目组合，而数据治理的对象是数据。

3）国际数据治理研究所数据治理框架

国际数据治理研究所(Data Governance Institute，DGI)认为数据治理不同于 IT 治理，应建立独立的数据治理理论体系。DGI 认为数据治理指的是对数据相关事宜的决策制定与权利控制，具体来说，数据治理是处理信息和实施决策的一个系统，即根据约定模型实施决策，包括实施者、实施步骤、实施时间、实施情境以及实施途径与方法。因此，DGI 从组织、规则、流程 3 个层面总结了数据治理的十大关键要素，创新地提出了 DGI 数据治理框架。DGI 数据治理框架以一种非常直观的方式，展示了 10 个基本组件间的逻辑关系，形成了一个从方法到实施的自成一体的完整系统。组件按职能划分为 3 组：规则与协同工作规范、人员与组织结构、流程。

DGI 数据治理框架以其简单、明了、目的清晰著称，在实施的过程中以数据治理的价值判断其实施的效果，并形成关键的管理闭环，是一种可以操作的、实际可行的数据治理框架。

4）IBM 数据治理框架

IBM 可能是最先提出数据治理概念的公司。基于其非凡的管理咨询与 IT 咨询的经验，同时也基于其大数据平台的开发，IBM 提出了数据治理统一流程理论(The IBM Data Governance Unified Process)。这个数据治理流程由 14 个步骤组成，具体包含定义业务问题、获取高层支持、执行成熟度评估、创建路线图、建立组织蓝图、创建数据字典、理解数据、创建元数据存储库、定义度量指标、主数据治理、治理分析、管理安全和隐私、治理信息生命周期、度量结果。

IBM 的数据治理流程是一个操作流程和项目导向的流程，最终形成了一次数据治理的闭环。值得注意的是，IBM 的数据治理流程拥有 InfoSphere Business Glossary 与 IBM InfoSphere Discovery 工具，能够把数据管理的深层次问题揭示出来，方便企业进行大数据配置方案的选择，从而使其数据治理方案能够彻底落地实施。

5）DCMM 数据治理框架

数据管理能力成熟度评估模型(Data Management Capability Maturity Assessment Model，DCMM)是我国首个数据管理领域国家标准。与欧美国家相比，在数据管理领域，我国一直缺乏完善的数据管理成熟度体系的研究，DCMM 填补了这一空白，为国内组织的数据管

理能力建设和发展提供了方向性指导。DCMM 国家标准结合数据生命周期管理各个阶段的特征，按照组织、制度、流程、技术对数据管理能力进行了分析、总结，提炼出组织数据管理的八大过程域（数据战略、数据治理、数据架构、数据应用、数据安全、数据质量管理、数据标准、数据生命周期），并对每项能力域进行了二级过程项（28 个过程项）和发展等级的划分（5 个等级）以及相关功能介绍和评定指标（441 项指标）的制定。

6）ISACA 数据治理框架

ISACA（Information System Audit and Control Association）是国际信息系统审计和控制协会的简称。ISACA 制定的 COBIT（Control Objectives for Information and Related Technology）是 IT 治理的一个开放性标准，该标准目前已成为国际上公认的最先进、最权威的信息技术管理和控制的标准。COBIT 标准体系已在世界 100 多个国家的重要组织与企业中运用，指导这些组织有效利用信息资源管理与信息相关的风险。

ISACA 数据治理框架从企业愿景和使命、策略与目标、商业利益和具体目标出发，通过对治理过程中人的因素、业务流程的因素和技术的因素进行融合和规范，提升数据管理的规范性、标准化、合规性，保证数据质量。ISACA 认为，要实现数据治理的目标，企业应在人力、物力、财力给予相应的支持，同时进行全员数据治理的相关培训和培养，通过管理指标的约束和企业文化的培养双重作用，使相关人员具备数据思维和数据意识，是企业数据治理成功落地的关键。

5.5 本章小结

（1）相对于传统的数据分析，大数据是海量数据的集合，它以采集、整理、存储、挖掘、共享、分析、应用、清洗为核心，正广泛地应用于军事、金融、工业、农业、教育、环境保护、通信等各个行业中。

（2）随着对大数据认识的不断加深，人们认为大数据一般具有 4 个特征：数据量大、数据类型繁多、数据产生速度快以及数据价值密度低。

（3）工业大数据是指在工业领域中，围绕典型智能制造模式，从客户需求到销售、订单、计划、研发、设计、工艺、制造、采购、供应、库存、发货和交付、售后服务、运维、报废或回收再制造等整个产品全生命周期各个环节所产生的各类数据及相关技术和应用的总称。

（4）工业大数据是智能制造的关键技术，主要作用是打通物理世界和信息世界，推动生产型制造向服务型制造转型。

（5）工业领域的数据累积到一定量级，超出了传统技术的处理能力，就需要借助大数据技术、方法提升处理能力和效率，大数据技术为工业大数据提供了技术和管理的支撑。

（6）工业大数据是工业互联网价值实现的核心要素，工业互联网以数据为核心要素实现全要素、全产业链、全价值链的全面连接，同时以数据驱动实现从感知控制到决策优化的闭环反馈。大数据治理的核心是为业务提供持续的、可度量的价值。

扫一扫

自测题

习题 5

（1）什么是大数据？

（2）请阐述大数据的特征。

（3）什么是工业大数据？

（4）什么是工业大数据建模？

（5）什么是机器学习？

（6）什么是工业大数据治理？