



# 1

## 第1章 数据技术概述



### 导读

在当今这个信息爆炸的时代，数据已经成为我们学习和生活中不可或缺的宝贵资源。从社交媒体上的点赞和评论，到学校的成绩报告和学习进度分析，我们无时无刻不被数据包围。而数据技术的发展和运用，已经成为促进我们学习和成长的关键力量。

### 学习目标

1. 理解数据的重要性，包括其在不同领域中的应用和影响。
2. 了解数据技术的定义，包括其在数据获取、分析和可视化方面的作用。
3. 掌握数据技术的系统框架，包括数据获取、数据分析和数据可视化等环节。
4. 熟悉数据技术的基础原理，理解其背后的核心概念和方法。
5. 熟悉常用的数据技术工具，能够灵活运用它们进行数据处理和分析。
6. 理解数据技术在不同领域（如商业、科学、医疗等领域）中的应用。



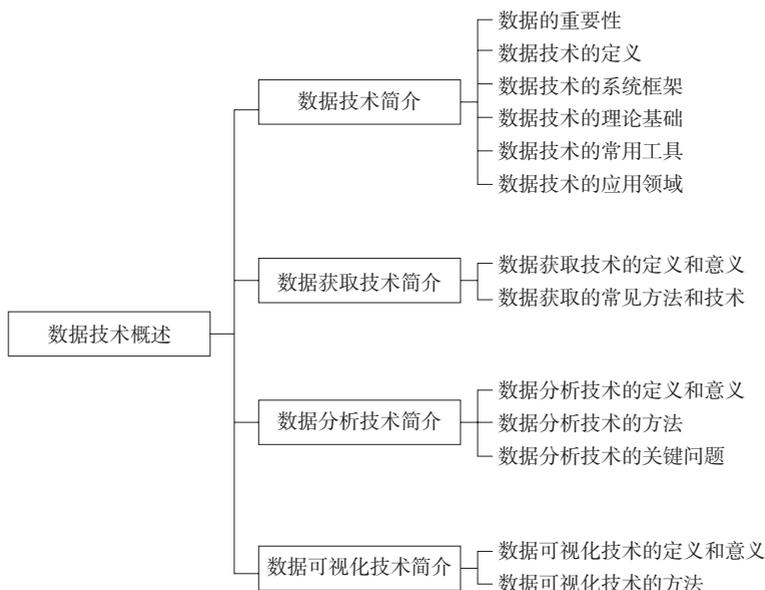
## 重点与难点

1. 数据获取技术的关键问题，包括数据质量问题、采样、抽样和样本问题、时间和空间限制问题等，需要学会解决这些问题以确保获取可靠的数据。

2. 数据分析技术的关键问题，数据安全与隐私问题、分析算法选择问题和业务理解与需求分析问题，需要学会有效地处理和分析大规模的数据。

3. 数据可视化技术的关键问题，选择合适的可视化形式、信息过载、跨平台兼容等，需要学会将数据转化为直观且有意义的可视化展示。

## 知识导图



## 1.1 数据技术简介

随着数据的爆炸式增长，软件的处理重心由以流程控制为核心转向以数据价值挖掘为核心，数据技术在趋势预测、个性化推荐、事务的相关性分析等方面有着极其广泛的应用。在全球范围内，数据技术研究和计划正日益受到重视，各国和国际组织都在积极制定和实施自己的数据战略，以确保他们在数据驱动的创新和进步中保持竞争力。国内很

多地区也出台了相应的大数据发展战略。学术界和产业界针对大数据发展的迅猛需求，展开了大数据相关技术的研究，引发了大数据的标准化需求，各个标准化组织也纷纷提出了大数据标准。

数据技术是利用科学方法、流程、算法和系统从数据中提取价值的跨学科领域。数据技术综合利用一系列技能（包括统计学、计算机科学和业务知识）来分析来自网络、智能手机、客户、传感器和其他来源收集的数据。数据技术的目的是揭示趋势并产生见解，企业可以利用这些见解做出更好的决策并推出更多创新产品和服务。数据是创新的基石，但是只有数据科学家利用数据技术从数据中收集信息，然后采取行动，才能实现数据的价值。接下来将从数据的重要性、数据技术的定义、数据技术框架等方面对数据技术进行讲解。

### 1.1.1 数据的重要性

数据是对现实世界进行观察、测量和记录的结果，它蕴含着巨大的价值和潜力。数据可以帮助我们认识事物的本质和规律，指导我们的决策和行动。无论是个人还是组织，都需要依靠数据来获取洞察、发现机会，并做出明智的选择，数据价值丰富多样如图 1-1 所示。

数据在现代社会中具有非常重要的作用，接下来将从以下 5 个方面讲解数据的重要性。

#### 1. 改善决策制定

数据可以用于支持决策制定，帮助企业 and 组织做出更明智的决策。通过分析不同来源的数据，可以发现潜在趋势、模式和感兴趣的关系，为企业和组织提供基于事实的决策依据。这种基于数据的决策过程不仅能够显著降低错误决策的风险，还能有效减少资源的不必要消耗，提升整体运作效率。

#### 2. 提高生产效率

通过对数据进行分析可以帮助企业和组织发现并解决生产过程中的瓶颈和问题，改进流程并提高生产效率。例如，生产线上的传感器技术可以收集实时数据，监测设备运行状况，能够及时发现设备故障并组织人员进行修理，从而减少停机时间，提高生产效率。

#### 3. 洞察客户行为和提供个性化服务

数据分析可以帮助企业和组织了解客户需求和购买行为等方面的信息，并根据这些信息提供个性化的服务。例如，根据客户购买历史和浏览行为，企业可以提供个性化的商品

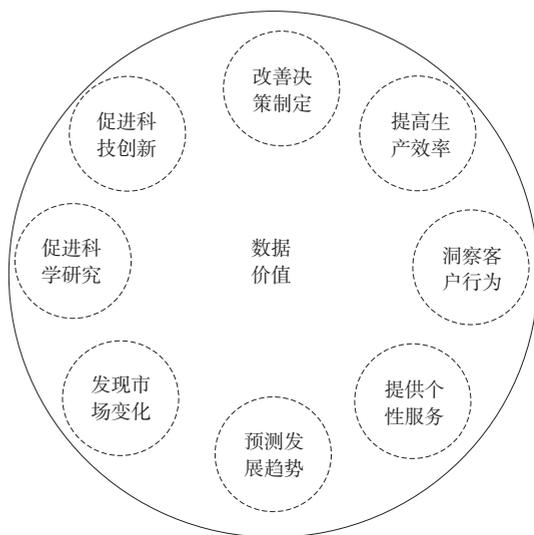


图 1-1 数据价值



推荐，从而提高客户的满意度和忠诚度。

#### 4. 预测发展趋势和发现市场变化

数据分析可以帮助企业和组织预测发展趋势和发现市场变化，以便及时调整战略和业务规划。例如，通过收集和分析市场数据，企业可以了解不同地区和行业的需求和趋势，从而调整产品或服务的规划，并开拓新的市场。

#### 5. 促进科学研究科技和创新

数据对于科学研究和技术创新具有重要意义。通过对数据的分析和挖掘，科学家们可以发现新的规律和关系，促进学术研究的发展。此外，数据还可以为创新和发明提供灵感，帮助企业和组织开发出新的产品和服务，从而促进社会 and 经济发展。

数据在现代社会中非常重要，可以用于改善决策制定、提高生产效率、洞察客户行为提供个性化服务、预测发展趋势和促进科技创新等方面，为企业和组织提供更多的机遇和挑战。

### 1.1.2 数据技术的定义

数据技术（Data Technology）由马云在世界互联网大会中演讲时正式提出，与信息技术（Information Technology）相对应。

在浙江乌镇举办的首届世界互联网大会上，众多业界领袖共同探讨了数据技术的重要性。与会者普遍认为，我们正从信息技术时代迈向数据技术时代。在这一转变中，技术的应用不再局限于自我控制和管理，而是更加注重服务大众和激发生产力。

数据技术的核心在于其能够为企业提供强大的决策支持。通过分析来自不同渠道的数据，企业和组织能够洞察市场趋势、识别模式，并发现潜在的商业机会。这不仅有助于降低决策失误的风险，还能优化资源配置，提高运营效率，从而增强企业的市场竞争力。

随着云计算和大数据技术的兴起，数据的价值被进一步放大。云计算为大数据分析提供了必要的计算能力和存储资源，使得企业能够更高效地处理和分析海量数据。这种结合不仅提高了数据处理的速度，也降低了成本，使得数据技术成为推动创新和增长的关键因素。

在当今时代，企业的成功越来越依赖于其对数据技术的掌握和应用。信息技术和数据技术虽然密切相关，但它们在理念和应用上存在显著差异。数据技术强调的是利他主义精神，即通过帮助他人成功来实现自身的成功。这种思想的转变标志着从信息技术时代到数据技术时代的过渡，其中最关键的是帮助他人成功。

随着互联网和云时代的深入发展，数据已成为企业和组织关注的焦点。数据技术的应用，包括大数据平台和相关工具，正在不断推动着各行各业的创新和转型。企业必须认识到，数据不仅是资产，更是推动未来发展的引擎。

数据技术涵盖了数据的收集、处理、分析和应用等关键环节。“数据技术”并非凭空创造，而是随着互联网的发展和市场需求的演变而自然形成的。这门技术已成为推动社会

进步和创新的重要力量。从学术概念来说，数据技术是一种可以将各种信息（无论信息的载体是什么）转化为计算机可以识别的语言进行加工、存储、分析、传递的技术。互联网行业从门户网站时代到搜索引擎时代，再到移动社交网络时代，直至今天的自媒体时代，数字化早已存在于企业的系统中。这些系统包括前端、数据中心、信息系统以及后台等，部分企业甚至会定制开源或半开源的系统，以便日后随着企业的发展增加相应的模块。这种数字化的初级形态为企业发展打下了坚实基础，而数据技术则成为推动企业发展的主要动力。

数据技术包含数据获取、存储、处理、分析和可视化等方面的技术，如图 1-2 所示。

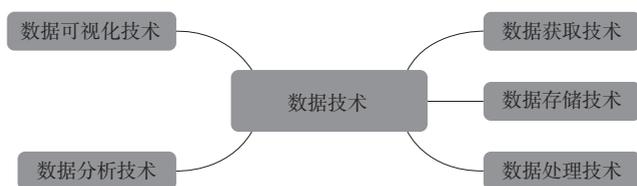


图 1-2 数据技术

数据技术的发展让我们能够更好地处理和分析海量数据，并从中获取有益的信息和见解。数据获取、存储、处理、分析和可视化等方面的技术不断完善和深化，为各行各业带来了全新的机遇。随着人工智能、大数据和云计算等领域的不断创新，数据技术必将在未来发挥越来越重要的作用，成为推动科技进步和社会发展的重要引擎。因此，对于数据技术的探索和应用，需要我们保持持续的关注和学习，以更好地应对日益增长的数据需求和挑战。

### 1.1.3 数据技术的系统框架

数据技术的系统框架是指用于处理、存储、管理和分析数据的整体架构，图 1-3 展示了完整的数据技术系统框架。一个完善的数据技术系统框架能够帮助组织更好地利用数据，从而提高业务决策的效率和准确性。下面对图 1-3 所示的数据技术系统框架的构成部分进行简单的介绍。

#### 1. 需求分析层

在这一层，主要的任务是理解业务需求，明确需要收集和分析哪些数据，以及为了支持业务目标而需要开展的数据分析和挖掘工作。这个层次涉及与业务部门的沟通及协调，以确保数据技术系统的建设与业务需求相匹配。

#### 2. 数据获取层

这一层负责从各种数据源中收集数据，包括传统的数据库、日志文件、第三方数据接口、传感器等。数据获取的方式包括批量导入、实时流式数据采集等。

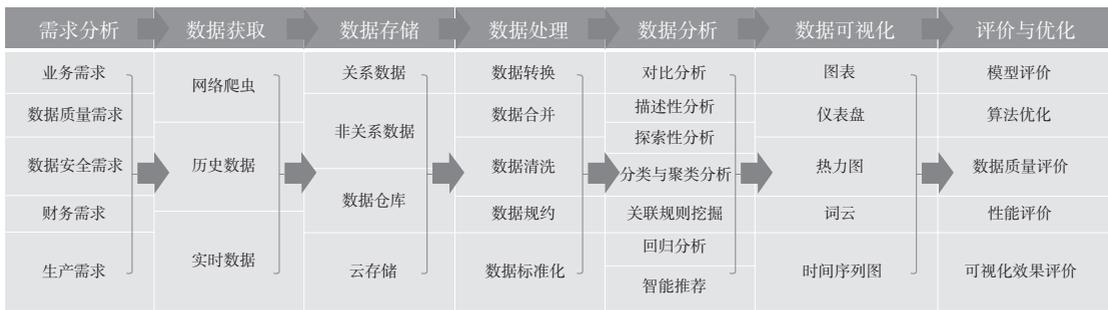


图 1-3 数据技术系统框架

### 3. 数据存储层

在这一层，数据经过处理后被存储在适当的存储介质中，以便随时访问和查询。常见的数据存储技术包括关系型数据库、NoSQL 数据库、数据仓库、云存储等。

### 4. 数据处理层

数据处理层对采集到的数据进行清洗、转换、集成和存储。这一层包括数据清洗、ETL（抽取、转换、装载）等过程，确保数据质量和一致性。

### 5. 数据分析层

在数据处理完成后，数据分析层负责利用清洗和转换后的数据进行进一步的分析和挖掘，以获得有价值的信息。这可能涉及使用统计分析、机器学习、数据挖掘等技术手段。

### 6. 数据可视化层

数据可视化层通过图表、仪表盘、热力图、词云和时间序列图等多种形式，将分析结果直观地呈现给用户，助力业务决策和行动。这一层次的工作关键在于将数据转化为具体的业务影响。可视化工具在此扮演着至关重要的角色，它们帮助创建直观的报表、仪表盘和数据可视化，使业务人员能够轻松理解数据。

### 7. 评价与优化层

这一层的核心职责是评估并优化数据处理、分析及可视化流程，以提升系统性能和效率。工作内容涵盖模型评价、算法优化、数据质量评价、性能评价和可视化效果评价。对分析模型的准确性和适用性进行评价，确保算法能够高效处理数据并提供准确的洞察。通过这些综合措施，我们能够不断提升数据处理的质量和效率，为决策提供强有力的数据支持。

## 1.1.4 数据技术的理论基础

数据技术的理论基础涉及多个领域的知识，包括数学、统计学、计算机科学、信息科学等，这些领域的知识为数据技术的进一步发展和应用提供了扎实的理论基础，图 1-4 展示了数据技术的理论基础与研究领域。

数据技术的理论基础主要涉及统计学、算法、数据库与数据仓库等技术，接下来就从以下几个方面对数据技术的理论基础进行阐述。



图 1-4 数据技术的理论基础与研究领域

(1) 统计学。统计学是数据技术的理论基础之一，它提供了对数据进行分析和解释的工具。统计学中的基本概念和原理，如描述性统计、推论性统计、假设检验等，为数据科学家提供了对数据进行深入探索和理解的方法。

(2) 算法。算法是解决特定问题的步骤和方法。在数据技术中，算法用于处理和分析数据，提取有用的信息。常见的算法包括聚类算法、分类算法、回归算法等，这些算法可以帮助数据科学家对数据进行分类、预测和优化。

(3) 数据库与数据仓库技术。数据库和数据仓库技术是存储和管理数据的重要技术，它们提供了高效的数据存储、查询和访问功能，支持对大量数据进行处理和分析。关系型数据库和非关系型数据库是常见的数据库类型，而数据仓库则是一个专门用于存储和分析大量数据的系统。

(4) 数据处理与分析技术。数据处理和分析技术是数据技术的核心之一，它涉及对原始数据的收集、清洗、转换和建模等过程，以提取有用的信息。在数据处理和分析过程中，数据科学家可以使用各种工具和技术，如 Python、R 语言、Excel 等，以及各种数据处理和分析库，如 Pandas、NumPy 等。

(5) 数据可视化技术。数据可视化技术是将大量数据转化为图形或图像的过程，以便更直观地理解和分析数据。数据可视化技术可以帮助人们更好地理解数据和发现其中的模式和趋势。常见的数据可视化工具包括 Tableau、Power BI、D3.js 等。

(6) 机器学习与深度学习。机器学习和深度学习是近年来发展迅速的领域，它们为数据技术提供了强大的工具。机器学习是一种让计算机自动从大量数据中学习规律和模式的技术，而深度学习则是一种基于神经网络的机器学习方法。在数据技术中，机器学习和深度学习用于提高数据处理和分析的准确性和效率。

(7) 自然语言处理技术。自然语言处理技术是使计算机理解和处理人类语言的技术。在数据技术中，自然语言处理技术用于文本数据的分析和处理，如文本分类、情感分析、信息提取等。常见的自然语言处理工具和技术包括词袋模型、TF-IDF、词嵌入等。



数据技术的理论基础是多学科知识的融合，并不局限于数学、统计学、计算机科学和信息科学。这些理论基础为数据技术的发展提供了坚实的支撑，使得数据科学家能够运用统计学原理深入分析数据，利用算法提取信息，通过数据库技术高效管理数据，以及运用数据分析和分析技术、数据可视化技术等技术，从海量数据中挖掘出有价值的知识和洞见。

## 1.1.5 数据技术的常用工具

数据技术的常用工具包括数据库管理系统（如 MySQL、Oracle）、数据处理工具（如 Python 的 Pandas、R 语言的 Tidyverse）、数据可视化工具（如 Tableau、Power BI）、机器学习和数据挖掘工具（如 Scikit-learn、TensorFlow）以及商业智能工具（如 QlikView）。这些工具对于数据的存储、清洗、转换、分析和可视化至关重要，它们使数据管理更加高效，并为人们提供了精确的洞察力和决策支持。接下来将从数据获取工具、数据分析工具和数据可视化工具三方面来讲解数据技术常用工具。

### 1.1.5.1 数据获取工具

数据获取工具是用于从数据库、Web 网站、API 接口、日志文件等数据源中自动收集和提取数据的软件工具。常见的数据获取工具包括爬虫工具、数据库提取工具、API 工具、日志分析工具、OCR 工具，以及数据采集工具。选择合适的工具可以帮助人们高效地获取各种来源的数据，并为后续的数据处理和分析提供准确的数据基础。以下是几种常见的数据获取工具。

#### 1. 爬虫工具

爬虫工具是用于从 Web 网站中自动抓取数据的工具。通过指定抓取规则和目标网站，可以让爬虫工具自动化地收集数据并存储在本地或远程数据库中。常见的爬虫工具包括 Scrapy、BeautifulSoup 等。

#### 2. 数据库提取工具

数据库提取工具是用于从关系型数据库中提取数据的工具。数据库提取工具可以连接到目标数据库，并根据预设的查询条件和过滤器提取数据，还可以对数据进行转换和处理后将其导出为常见的数据格式，如 CSV、Excel 等。常见的数据库提取工具包括 SQL Server Management Studio、MySQL Workbench 等。

#### 3. API 工具

API 工具是用于调用 Web API 接口并获取数据的工具。很多网站和服务都提供了 API 接口，例如 Twitter、Google Maps 等，开发人员可以使用 API 工具设置请求参数，然后调用 API 接口并获取返回的数据。常见的 API 工具包括 Postman 等。

#### 4. 日志分析工具

日志分析工具是用于从服务器日志文件中提取数据的工具。服务器日志文件包含了服

务器请求、错误和访问信息等，这些数据以文本格式保存在服务器上。日志分析工具可以对日志文件进行解析，并将解析后的数据进行存储和可视化展示。常见的日志分析工具包括 ELK Stack（Elasticsearch、Logstash、Kibana）等。

### 5. 数据采集工具

数据采集工具是用于从多个数据源中收集和整合数据的工具。数据采集工具可以设置抓取规则和过滤条件，然后定期从数据源中收集数据并将其存储到本地或远程数据库中。常见的数据采集工具包括八爪鱼数据采集器、后羿数据采集器、Parsehub 数据采集器等。

（1）八爪鱼数据采集器。八爪鱼数据采集器（以下简称“八爪鱼”）是一款知名的数据采集工具，它提供了强大而灵活的功能，可帮助用户从各种数据源中高效地采集、提取和整合数据。八爪鱼具有以下几个主要特点：

①多种数据采集方式。八爪鱼支持多种数据采集方式，包括网页抓取、API 接口调用、数据库提取、文件下载等。用户可以根据需要选择适合的采集方式，从不同的数据源中获取所需的数据。

②强大的可视化操作界面。八爪鱼提供了直观、友好的可视化操作界面，让用户通过简单的拖动和配置就能完成数据采集任务。用户无须编写复杂的代码，即可快速构建和管理数据采集项目。

③智能的数据解析和处理能力。八爪鱼内置了智能的数据解析引擎，能够自动识别和提取网页、API 接口返回的结构化数据。同时，它还提供了丰富的数据处理功能，如数据清洗、筛选、转换等，可以对采集到的数据进行进一步处理和优化。

④多种输出格式和目标。八爪鱼支持将采集到的数据导出为多种格式，包括 Excel、CSV、数据库等。此外，八爪鱼还可以将数据直接上传到云端或存储到其他目标系统，方便用户后续的数据分析和应用。

⑤强大的扩展性和定制性。八爪鱼提供了丰富的插件和扩展接口，用户可以根据自己的需求进行二次开发和定制。通过编写自定义脚本或添加自定义模块，可以完成更加复杂和个性化的数据采集任务。

八爪鱼是一款功能强大、操作简便的数据采集工具，能够帮助用户快速、准确地从各种数据源中采集所需的数据，为数据分析和应用提供重要的支持。图 1-5 展示了八爪鱼数据采集器的操作界面。

（2）后羿数据采集器。后羿数据采集器是一款强大的数据采集软件，支持跨平台操作。后羿数据采集器具备以下几个主要特点：

①它支持 Linux、Windows 和 Mac 三大操作系统，可以直接在官网上免费下载。

②后羿数据采集器把采集工作分为智能模式和流程图模式两种类型。智能模式可以自动分析网页结构、智能识别网页内容、简化操作流程，适用于简单的网页。流程图模式则允许用户利用后羿数据采集器提供的各种控件，模拟编程语言中的各种条件控制语句，进



图 1-5 八爪鱼数据采集器的操作界面

而模拟真人浏览网页的各种爬取数据的行为。

③后羿数据采集器还提供了丰富的采集功能，如定时采集、自动导出、文件下载、加速引擎、按钮启动和导出、Webhook、RESTful API、智能识别 SKU 和大图等，无论是采集稳定性还是采集效率，都能够满足个人、团队和企业的采集需求。

④后羿数据采集器具有强大的数据导出功能，可以无限制地导出采集到的数据，这是非常难得的。

⑤后羿数据采集器还提供了云端账号功能，用户可以在后羿数据采集器的云端服务器上加密保存所有采集任务设置和运行的数据，无须担心采集任务丢失。

⑥后羿数据采集器对账号没有设置终端绑定限制，用户切换终端时采集任务也会同步更新，任务管理非常方便快捷。

总的来说，后羿数据采集器是一款功能强大、操作简便的数据采集软件，适用于各种平台和需求。无论是个人的数据采集需求，还是团队和企业的的天数据采集任务，后羿数据采集器都可以提供全面的支持和高效的解决方案。图 1-6 展示了后羿数据采集器的操作界面。

(3) Parsehub 数据采集器。Parsehub 数据采集器是一款强大的网页数据抽取工具，旨在帮助用户从网页中自动化地提取结构化数据。它提供了用户友好的界面和智能的数据解析功能，可以轻松应对各种复杂的数据抽取需求。图 1-7 展示了 Parsehub 数据采集器的操作界面。

### 1.1.5.2 数据分析工具

数据分析工具包括各种用于探索、清理和分析数据的软件和工具。其中，Excel、R 语言和 Python 是比较受欢迎的工具有，它们具有强大的数据处理和数据分析功能，能帮助用户快速创建各种图表和图形。此外，Power BI、Fine BI 和 SAS 也是广受欢迎的工具有，它们分别由微软、帆软和赛仕公司开发。其他如 KNIME 和 RapidMiner 则是开

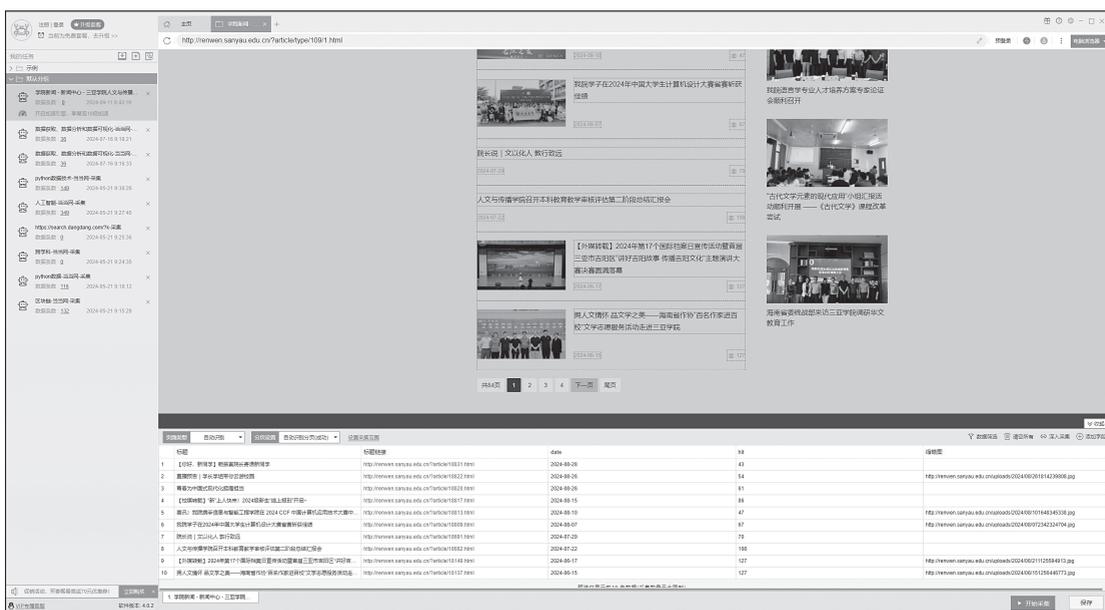


图 1-6 后羿数据采集器的操作界面

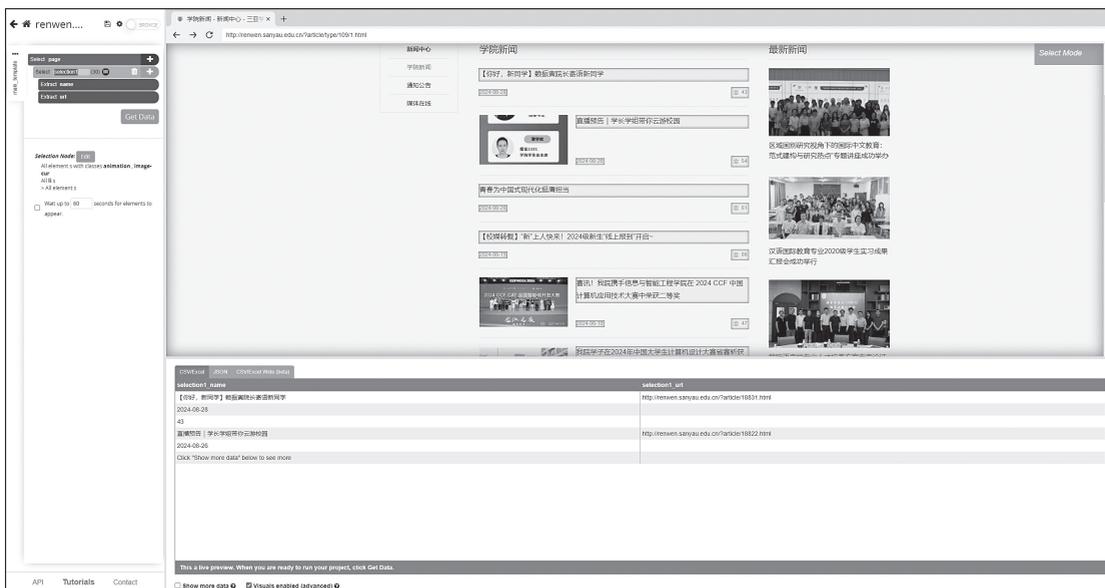


图 1-7 Parsehub 数据采集器操作界面

源的数据分析工具，具有强大的数据处理和分析功能。这些工具的选择应基于实际的需求和数据分析的精度。接下来简单介绍一下这些数据分析工具。

### 1. Excel 数据分析工具

Excel 是一款被广泛使用的数据分析工具，它提供了许多功能强大的数据工具，包括数据可视化、数据清理、数据分析、数据挖掘等。以下是一些常用的 Excel 数据分析工具：



(1) 数据透视表。数据透视表是 Excel 中最重要的数据分析工具之一，它可以快速地汇总、过滤、组合和分析大量数据。通过数据透视表，用户可以把数据按照不同的维度进行组合和分组，以便更好地了解数据的分布和关系。

(2) 数据透视图。数据透视图是基于数据透视表而生成的，它是一种图表形式的数据透视表，可以更形象地展示数据的分布和关系。通过数据透视图，用户可以把数据按照不同的维度进行组合和分组，以便更好地了解数据的走势和关系。

(3) 模拟运算表。模拟运算表是 Excel 中一个非常实用的功能，它可以帮助用户模拟和预测数据的走势和关系。通过模拟运算表，用户可以输入不同的假设值，并查看它们对数据的影响。

(4) 回归分析。回归分析是 Excel 中一个非常强大的功能，它可以帮助用户分析数据之间的关系，并预测未来的走势。通过回归分析，用户可以确定数据之间的相关性和趋势，以便更好地了解数据的规律和特点。图 1-8 展示了常用的 Excel 数据分析工具。



图 1-8 常用的 Excel 数据分析工具

## 2. Python 数据分析工具

Python 是一款流行的编程语言，它的数据分析工具广泛应用于数据挖掘、机器学习、统计学等领域。

Python 数据分析工具可以帮助用户快速、高效地处理和分析大量数据，从而更好地理解 and 掌握 Python 中数据的特征和规律。对于数据分析师来说，主要应掌握 Python 中基础语法和数据科学的模块，主要包括 Numpy、Pandas 以及机器学习的库。图 1-9 展示了常用的 Python 数据分析工具。

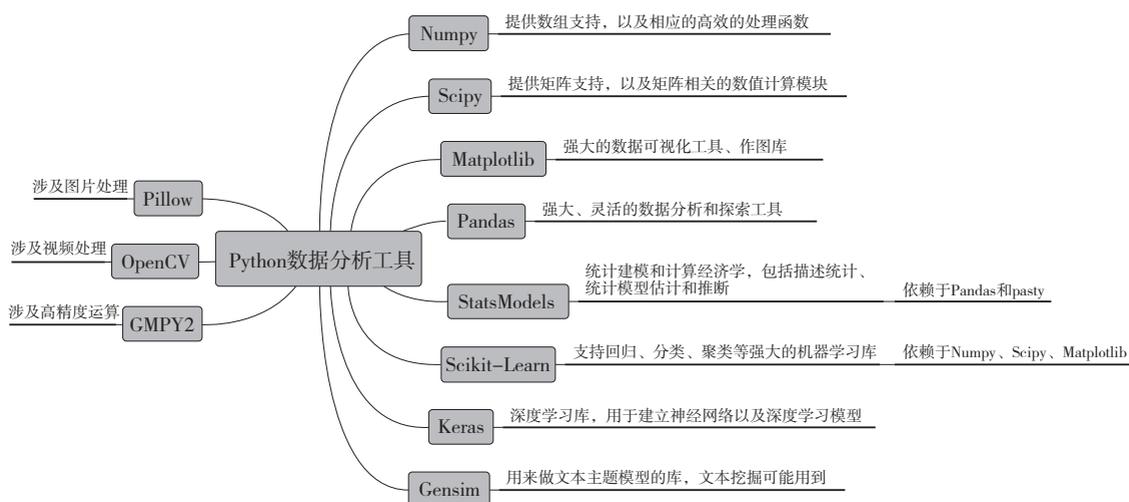


图 1-9 常用的 Python 数据分析工具

### 3. Power BI 数据分析工具

Power BI 是由 Microsoft 开发的一款商业智能分析和可视化工具，旨在帮助用户从各种数据源中提取、分析和可视化数据，并将数据转化为有意义的见解和决策支持。它具有强大的数据整合、可视化和分享功能，适用于个人用户和企业组织。

Power BI 是一款功能强大、易于使用的商业智能工具，可帮助用户从各种数据源中提取、分析和可视化数据。它提供了丰富的数据整合和建模功能、多样化的数据可视化选项以及方便的数据共享和协作功能，可满足个人用户和企业组织对数据分析和决策支持的需求。图 1-10 展示了 Power BI 的功能。

#### 1.1.5.3 数据可视化工具

数据可视化工具是用于将数据转化成易懂、有意义的图表、图形或仪表盘的软件工具，常见的数据可视化工具包括 Tableau、Power BI、QlikView/Qlik Sense、D3.js、Plotly 等。这些工具提供了丰富多样的可视化选项，支持实时数据更新和自动化报告，并帮助用户更好地探索和呈现数据，揭示数据中的模式、趋势和见解，从而更好地支持决策和业务发展。由于前面已经介绍了 Power BI 工具，接下来主要介绍 Tableau、D3.js 和 QlikView/Qlik Sense 数据可视化工具。

##### 1. Tableau 可视化工具

Tableau 是一款领先的数据可视化和商业智能工具，旨在帮助用户从各种数据源中提取、分析和展示数据，并将数据转化为有意义的见解和决策支持。它具有直观易用的界面、强大的数据整合和可视化功能，适用于个人用户和企业组织。

Tableau 提供了丰富的数据整合和连接功能、多样化的可视化选项以及高性能的数据查



图 1-10 Power BI 的功能

询和分析能力。通过 Tableau，用户可以轻松地各种数据中获取见解，并与其他人共享和讨论分析结果。图 1-11 展示了 Tableau 的操作界面。

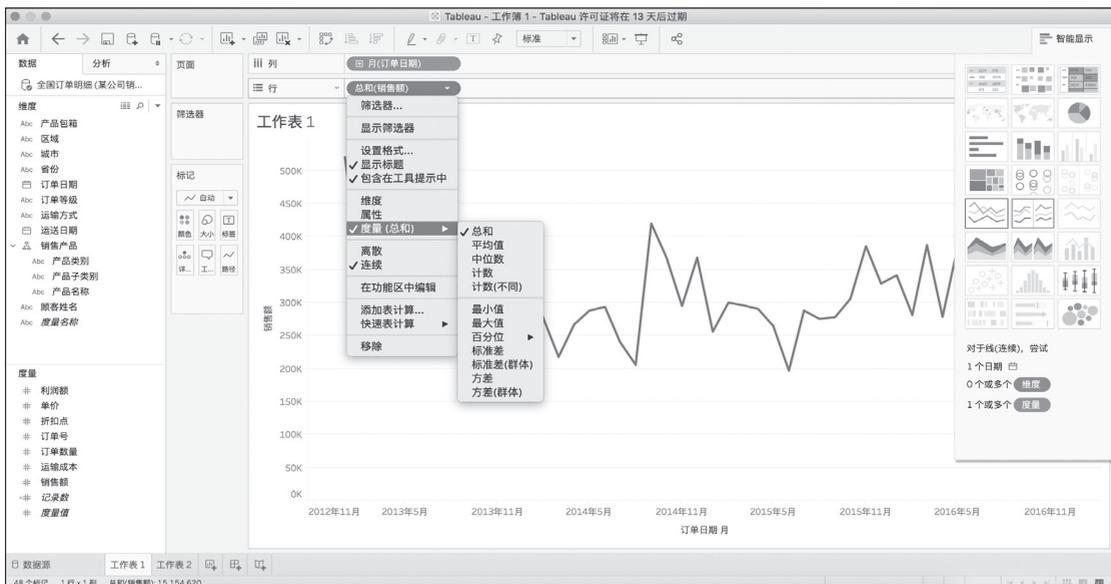


图 1-11 Tableau 的操作界面

## 2. D3.js 可视化工具

D3.js (Data-Driven Documents) 是一款基于 JavaScript 的开源数据可视化库, 它提供了强大、灵活的功能, 使用户能够创建高度定制化和交互式的数据可视化应用。

D3.js 提供了丰富多样的图表类型和可视化选项, 并支持数据绑定、数据转换、交互和动画效果等功能。通过 D3.js, 用户可以根据自己的需求和创意, 创建高度定制化、交互式 and 动态的数据可视化应用。图 1-12 展示了使用 D3.js 制作的数据可视化效果。

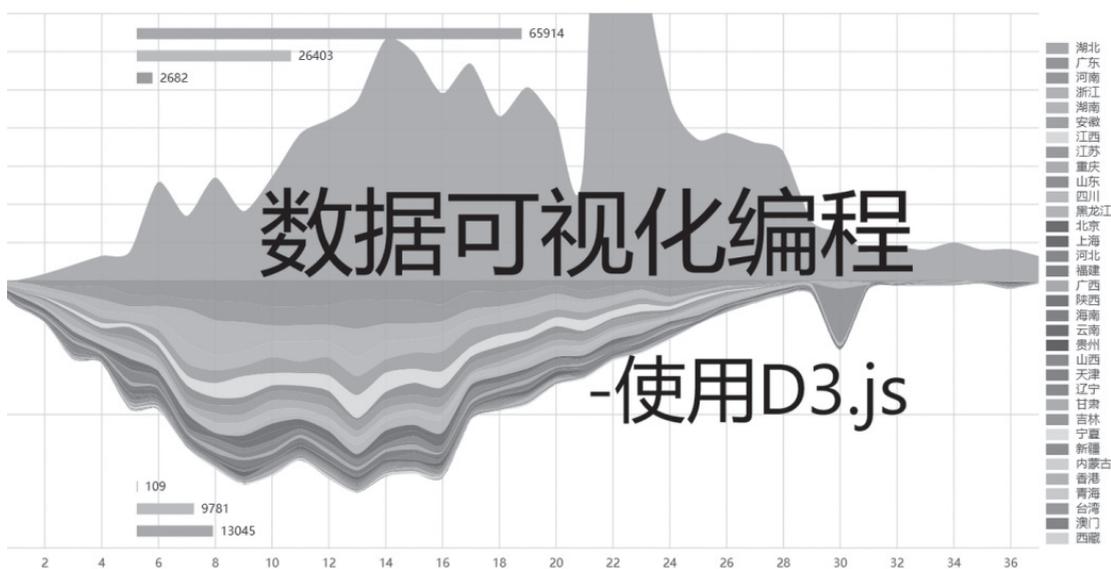


图 1-12 使用 D3.js 制作的数据可视化效果

### 1.1.6 数据技术的应用领域

数据技术在现代社会中的应用领域非常广泛。在商业领域, 数据技术可以用于市场营销、客户关系管理、供应链管理、金融分析和销售预测等方面, 帮助企业做出更精准、高效的商业决策。在医疗领域, 数据技术可以用于电子病历、个性化治疗、药物研发和生物信息学等方面, 提升医疗服务质量和管理患者信息。在政府领域, 数据技术可以用于公共管理、决策支持、城市规划、环境保护、交通管理和公共卫生等方面, 帮助政府做出准确、高效的决策。在社交媒体领域, 数据技术可以支持推荐算法、广告投放和社交网络分析, 为用户提供个性化的服务和推荐。此外, 数据技术还可以用于能源、交通、环境保护和农业等各种领域。随着数据技术的发展, 其应用领域将继续扩大, 也将继续推动人类社会不断进步。图 1-13 显示了数据技术应用领域。



图 1-13 数据技术应用领域

## 1.2 数据获取技术简介

数据获取是数据技术中非常重要的一个环节，它涉及从多种来源收集、采集和提取数据。数据获取的方法和技术因数据类型、数据来源、数据目的等而异。本节将对数据获取技术进行简单介绍。

### 1.2.1 数据获取技术的定义和意义

数据获取技术可以定义为从各种来源、渠道或系统中获取数据的技术。这些数据可以是结构化数据（如数据库、电子表格等），也可以是非结构化数据（如文本、图像、音频等）。数据获取技术的意义可以概括为以下 5 点，如图 1-14 所示。



图 1-14 数据获取技术的意义

(1) 支持决策制定。数据获取为决策者提供了获取信息和深入了解问题的途径。通过收集相关数据，决策者可以更好地了解当前状况、发现问题、识别趋势和模式，从而做出明智的决策。

(2) 提供洞察和分析基础。数据获取是进行数据分析和挖掘的基础。收集到的数据可以用于分析、统计和建模，揭示隐藏在数据背后的模式、关联和规律，以求得更深刻的洞察和理解。

(3) 支持业务发展和创新。数据获取可以帮助企业和组织了解客户需求、市场趋势和竞争环境等因素，为业务发展和创新提供依据和方向。通过对大量数据的分析，可以发现新的商机、改进产品或服务，并提供更优质的用户体验。

(4) 改善效率和效益。数据获取有助于改善工作流程、业务运营和提高资源利用效率。通过对数据的分析，可以发现瓶颈、识别问题，并采取相应的措施进行优化和改进，提高工作效率、降低成本，并取得更好的效益。

(5) 支持科学研究和社会发展。数据获取是科学研究的基础。从不同领域获取的数据可以被用于学术研究、社会调查和政策制定等方面，为社会发展提供科学依据。

数据获取技术的重要性不言而喻，它是现代社会中各行业和领域的基础。有效的数据获取技术能够帮助组织和个人获取、整合和分析各种类型的数据，从而为决策制定、业务优化、创新发展提供有力支持。无论是商业领域的市场调研和客户洞察，医疗领域的疾病监测和药物研发，政府领域的交通管理和决策支持，还是科研领域的实验数据和信息检索，都离不开高效的数据获取技术。因此，数据获取技术的重要性在于为我们提供了深入了解现实世界、发现问题和机遇，并做出明智决策的基础信息。

## 1.2.2 数据获取的常见方法和技术

数据获取的常见方法和技术有多种，包括网络爬虫、传感器技术、数据库和 API 以及人工输入和扫描等，如图 1-15 所示。下面将重点介绍各种方法的特点、优缺点以及适用场景。

### 1. 网络爬虫

网络爬虫是一种自动化程序，能够按照预定的规则和算法在互联网上自动浏览并抓取信息，然后将这些信息整理存储或进行进一步的处理分析。网络爬虫通常被用来收集互联网上的数据，比如搜索引擎通过网络爬虫来建立网页索引，研究人员利用网络爬虫来获取特定领域的数据进行分析。网络爬虫的基本流程包括以下几个步骤：首先，确定要爬取的目标网站或特定网页；其次，通过发送 HTTP 请求，获取目标网页的 HTML 源代码；再次，解析 HTML 源代码，提取出需要的数据和链接；从次，将提取的数据进行存储或进一步处理；同时，将提取到的链接加入待爬取队列中；最后，不断循环执行以上步骤，直至完成对所有目标页面的爬取。在整个流程中，还需要考虑反爬机制、并发控制和数据清洗等问题，以确保爬取的效率和数据的准确性。图 1-16 展示了网络爬虫的基本流程。

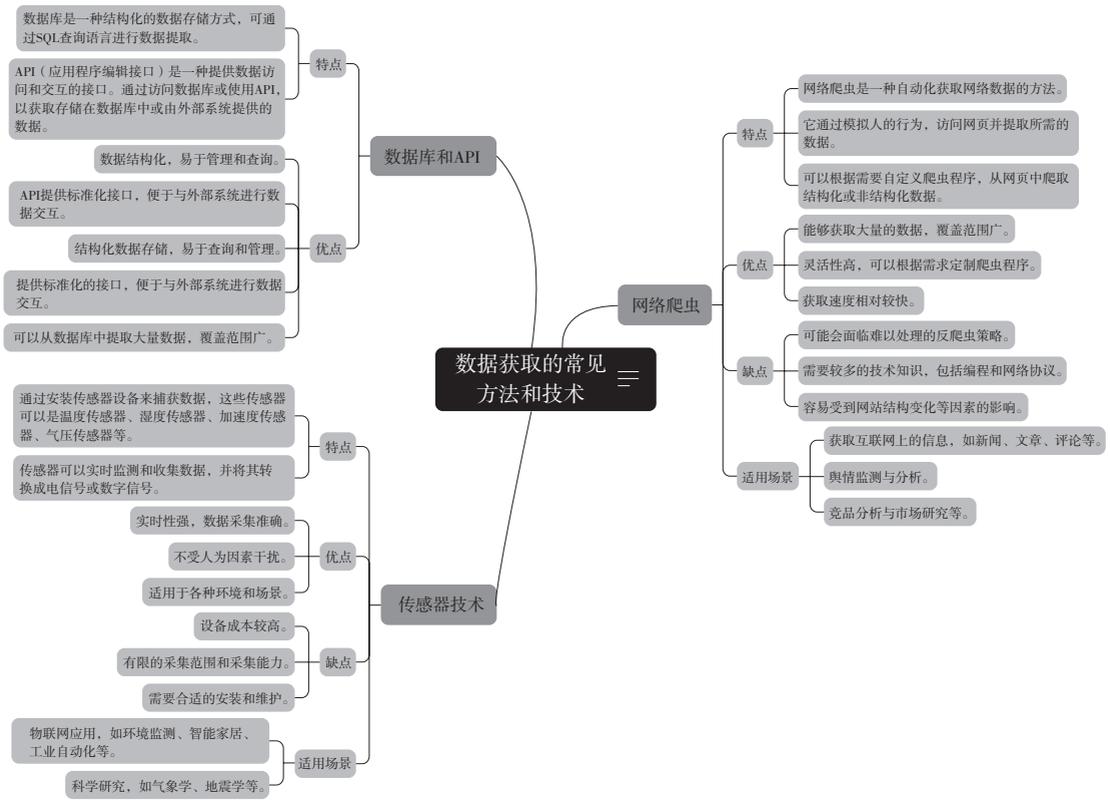


图 1-15 数据获取的常见方法和技术

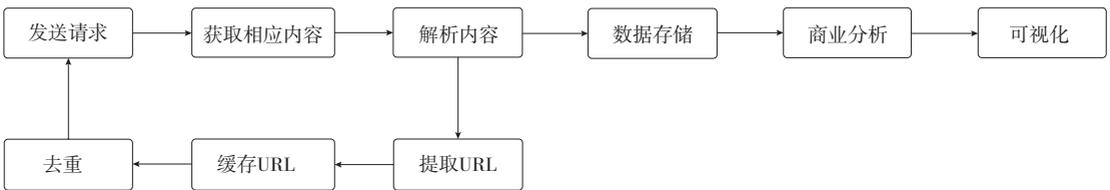


图 1-16 网络爬虫的基本流程

## 2. 传感器技术

传感器技术通过安装传感器设备来捕获各种数据。传感器是一种专门用于检测和响应特定物理量或现象的装置，如温度、湿度、压力、光照等。传感器设备将感知到的信号转换为可量化的电信号或数字信号，并通过接口与计算机系统或其他设备进行通信。这些传感器可被广泛用于工业、农业、医疗、环境监测等领域，实时获取和监测数据，为决策制定、控制系统优化，以及提高效率和安全性提供重要的支持。图 1-17 展示了传感器技术采集数据的原理。

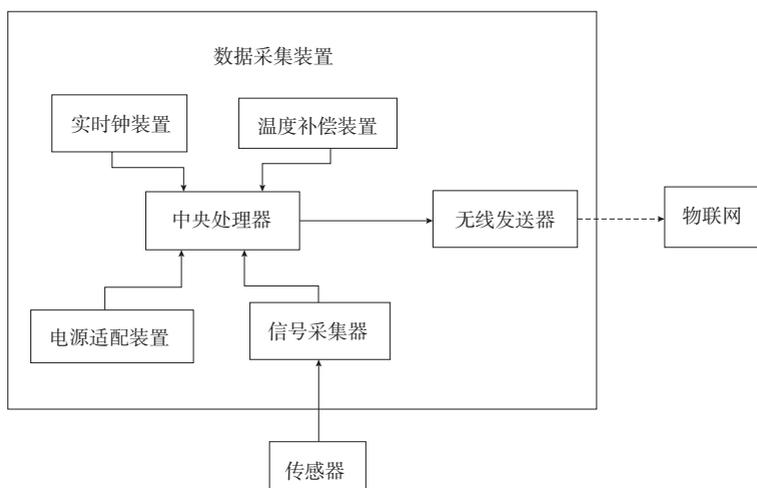


图 1-17 传感器技术采集数据的原理

### 3. 数据库和 API

数据库和 API 都是常用的数据获取方式。数据库是一个结构化的数据存储系统，可以按照一定的规则和格式存储、管理和检索数据。用户可以使用 SQL 等查询语言从数据库中提取所需的数据。而 API（应用程序编程接口）是一种软件工具，允许不同应用程序之间进行数据交互。通过 API，开发者可以请求远程服务器上的数据，并以特定格式（如 JSON 或 XML）返回所需数据。数据库适用于存储和管理大量结构化数据，而 API 则更加灵活，可以从外部系统或服务中获取数据，以满足特定的需求。无论是使用数据库还是 API，都能够有效地获取和处理数据，以供应用程序或用户使用。

### 4. 人工输入和扫描

人工输入和扫描是常见的数据获取方式。人工输入指的是通过人工手动录入信息到计算机系统中，例如通过键盘输入文本。而扫描则是利用扫描设备将纸质文件或图片转换为数字化形式，例如将纸质文件扫描成 PDF 文档或将照片扫描成数字图像文件。人工输入和扫描这两种方式都能够将现实世界中的信息转化为数字化的数据，为后续的存储、处理和分析奠定了基础。

不同的数据获取方法和技术各有特点和适用场景。选择何种数据获取方法取决于需要获取的数据类型、规模和实时性要求，以及自身的技术能力和资源限制。

不同的数据获取方法和技术适用于不同的应用需求，网络爬虫适合从互联网上收集大规模数据进行分析和挖掘，传感器技术则适用于实时监测环境参数并进行自动化控制，数据库和 API 适合于存储管理和获取结构化数据，而人工输入和扫描则适用于将纸质文件转换为数字化形式。根据具体的应用场景和需求，可以选择合适的的数据获取技术来满足不同的目的，从而更好地支持决策制定、系统优化和效率提升。



## 1.3 数据分析技术简介

数据分析技术是在已经获取的数据流或信息流中，寻找匹配关键词或关键短语的技术。数据分析的目的是把隐没在一大批看起来杂乱无章的数据中的信息集中、萃取和提炼出来，以找出所研究对象的内在规律。在实际应用中，数据分析可帮助人们作判断，以便采取适当行动。例如 J. 开普勒通过分析行星角位置的观测数据，找出了行星运动规律。又如，一个企业的领导人要通过市场调查，分析所得数据以判定市场动向，从而制定合适的生产及销售计划。

### 1.3.1 数据分析技术的定义和意义

数据分析技术可以分为狭义的数据分析和广义的数据分析两个层面，涵盖了从简单的数据探索到复杂的模型选择和推断分析等多种多样的任务和目的。狭义的数据分析主要指探索性数据分析，它在尽量少的先验假定下处理数据，通过表格、摘要、图示等直观的手段，探索数据的结构，并检测对于某种指定模型是否有重大偏离。探索性数据分析可以作为进一步分析的基础，也可以对数据做出非正式的解释，对实验方案进行调整，并重做实验。而广义的数据分析则包括了探索性数据分析在内，并进一步涉及模型选定分析和推断分析。模型选定分析是在探索性数据分析的基础上提出可能的模型，并通过进一步的分析从中挑选一定的模型，例如确定模型的形式和估算模型的参数等。而推断分析则使用数理统计方法对所挑选的模型的可靠程度和精确程度进行推断，例如检验所定模型的可用性，评估模型的精确程度等。

数据分析的意义在于通过对数据的挖掘和分析，提供有力的决策支持、优化业务流程、预测和预防风险，并发现新的机会和洞察，从而帮助企业 and 组织取得商业上的成功和竞争优势（见图 1-18）。以下是数据分析的几个重要意义。

#### 1. 发现洞察和解释现象

数据分析可以揭示数据中的隐藏信息和规律，帮助人们理解各种现象。例如，通过对市场销售数据的分析，可以发现产品的受欢迎程度、消费者的购买偏好等。

#### 2. 支持决策制定

数据分析为决策提供了有力的支持。通过数据分析，可以得出客观的结论和预测，帮

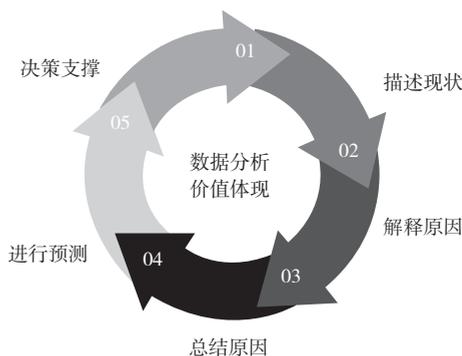


图 1-18 数据分析价值体现

助决策者做出明智的商业决策。例如，在市场营销中，通过对市场调研数据的分析，可以帮助决策者确定目标市场、优化产品定位和制定推广策略。

### 3. 优化业务流程和效率

数据分析可以帮助企业识别和解决业务流程中的瓶颈和问题。例如，通过对运营数据的分析，可以帮助企业和组织找到生产过程中的瓶颈，优化资源分配，提高生产效率。

### 4. 预测和预防风险

数据分析可以基于历史数据进行预测，帮助企业和组织预测未来的趋势和风险。例如，在金融领域，通过对市场数据和客户行为数据的分析，可以帮助企业和组织预测市场变化和风险，从而制定相应的风险管理策略。

### 5. 支持创新和发现新机会

数据分析可以揭示市场和消费者的新机会和潜在需求。例如，通过对市场和用户数据的分析，可以发现新的产品创新方向和市场细分，从而帮助企业和组织在竞争中保持竞争优势。

随着技术的不断进步和数据的不断增长，数据分析将发挥越来越重要的作用。因此，应该充分利用数据分析的价值，将其应用于各个领域，为企业和组织带来更大的发展和竞争优势。

## 1.3.2 数据分析技术的方法

数据分析技术是指利用各种工具和技术对数据进行处理、挖掘和解释的方法。随着大数据时代的到来，数据分析技术变得越来越重要，它不仅可以帮助企业和组织发现商业机会、降低风险，还可以促进科学研究和社会发展。数据分析技术的方法和技术多种多样，图 1-19 展示了几种常见的数据分析方法。

接下来就简单地阐述一下分组分析法、矩阵分析法、平均分析法、统计分析法、文本分析法和对比分析法 6 种常见的数据分析方法。

#### 1. 分组分析法

分组分析法就是将数据按照某种规则或特定因素分成不同的组，然后对不同组之间的差异进行分析。这种方法可以帮助我们更好地理解数据的变化趋势，从而优化业务流程。

#### 2. 矩阵分析法

矩阵分析法就是将数据按照二维矩阵的方式呈现，通过对比不同行列的变化趋势，分

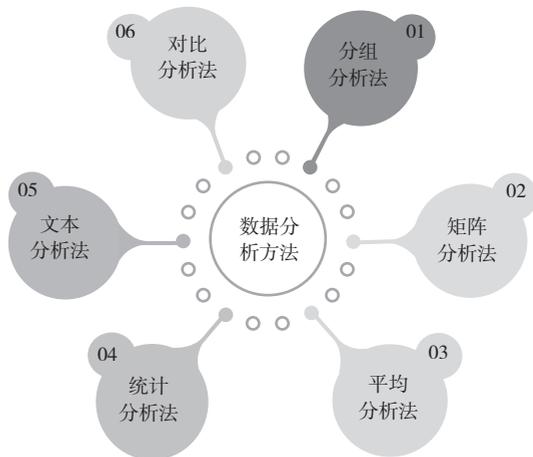


图 1-19 数据分析方法



析不同因素之间的相关性和影响程度。这种方法适用于多维度数据分析，可以帮助我们更好地理解数据之间的关系和相互作用。

### 3. 平均分析法

平均分析法就是将数据按照一定的时间范围或空间范围进行平均，以便更好地了解数据的总体情况和趋势，从而帮助我们做出更准确的预测和决策。

### 4. 统计分析法

统计分析法就是使用数学统计方法对数据进行分析，以获取总体趋势、标准差、方差等统计指标，进而进行预测和推断。这种方法可以帮助我们更好地了解数据的变化规律和潜在趋势。

### 5. 文本分析法

文本分析法就是对文本数据进行挖掘和分析，以获取文本中隐藏的信息和规律。这种方法可以帮助我们更好地了解用户需求和反馈，从而优化产品设计和改进服务。

### 6. 对比分析法

对比分析法就是将不同时间、不同地区或不同对象的数据进行对比，以便更好地了解它们之间的差异和共性。这种方法可以帮助我们更好地了解业务流程的变化趋势和变化原因，从而进行优化和改进。

数据分析技术提供了多种方法和工具，帮助我们更好地理解数据，发现潜在的机会和问题，并支持优化业务流程和决策制定。无论是分组分析法、矩阵分析法、平均分析法、统计分析法、文本分析法还是对比分析法，都为我们提供了不同的视角和手段来揭示数据中的信息和规律。通过这些方法，我们可以深入挖掘数据的价值，为企业和社会带来更好的发展。

## 1.3.3 数据分析技术的关键问题

数据分析技术在现代社会中有着广泛的应用，但在实际应用中仍面临多方面的挑战。数据分析技术的关键问题包括数据质量、数据安全和隐私、分析算法选择、业务理解和需求分析、数据共享等。数据质量是数据分析的基础，需要对数据进行清洗、去重、处理和转换等操作。解决上述关键问题，对于提高数据分析技术的性能和实用性至关重要。在实际应用过程中，数据分析技术面临着以下一些关键问题。

### 1. 数据质量问题

数据质量是数据分析的基础，如果原始数据不准确或不完整，则无法得出真正有价值的结论。因此，需要对数据进行清洗、去重、处理和转换等操作，以确保数据质量。

### 2. 数据安全和隐私问题

在数据分析过程中涉及大量的敏感数据和个人隐私信息，这需要对数据进行安全保护。同时，也需要确保数据使用符合相关法律和规定。

### 3. 分析算法选择问题

不同类型的数据和分析目标需要使用不同的分析算法，如聚类、分类、回归等。因此，在选择分析算法时需要结合数据的特点和需求进行选择。

### 4. 业务理解 and 需求分析问题

在数据分析前，需要对业务领域有充分的了解和认识，并通过需求分析确定分析目标和方法、数据来源和处理方法等，以保证数据分析的有效性和准确性。

### 5. 数据共享问题

在大数据时代，数据共享已成为发展的重点。但是不同的数据管理系统往往存在着互相独立、不兼容的问题，这就使得数据共享变得十分困难。数据分析技术需要解决数据共享问题，以实现跨平台、跨系统的数据互操作与数据共享。

## 1.4 数据可视化技术简介

数据可视化技术是将数据通过图表、图形、地图等方式进行可视化展示，以便人们更直观地理解数据特征和关系，能够帮助人们更好地分析数据、传达信息、探索数据和支持决策，提高数据分析的效率和质量。

### 1.4.1 数据可视化技术的定义和意义

数据可视化技术是指使用图表、图形、地图等可视化方式将数据呈现出来的一种技术。它通过形象、直观的图像将复杂的数据信息转化为可被人们轻松理解的形式，从而帮助人们更好地探索数据、发现规律和趋势、传达信息并做出决策。在进行数据可视化时可以根据不同的要求选择不同的图表类型，图 1-20 展示了数据可视化图表选择思维。

数据可视化不仅仅是简单地将数据转化为图表的过程，更是一种重要的数据分析和传播工具。数据可视化的意义主要体现在以下几个方面。

#### 1. 帮理解解和发现趋势

通过图表和图形的形式呈现数据，可以帮助人们更直观地理解数据之间的关系、趋势和模式，从而更容易地发现数据中的规律和趋势。例如，使用折线图可以清晰地展示销售额随时间的变化趋势，帮助业务团队了解市场趋势和产品销售情况。

#### 2. 提高决策效率

将复杂的数据转化为可视化形式，可以帮助决策者更快速地理解数据，并做出基于数据的决策。例如，使用仪表盘或可交互的图表，管理层可以实时监控关键指标的变化情

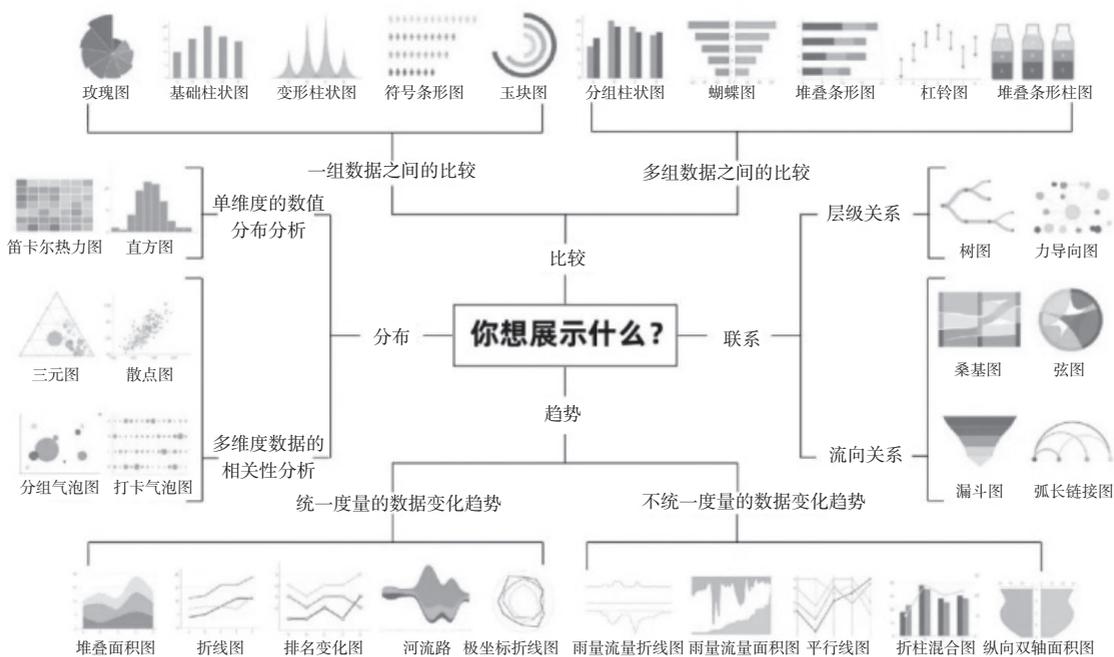


图 1-20 数据可视化图表选择思维

况，及时做出调整和决策。

### 3. 有效沟通与共享

通过可视化图表，人们可以更清晰地向他人展示数据分析的结果，从而更有效地进行沟通 and 共享。这对于团队内部协作和与外部利益相关者的沟通都非常重要。可视化图表可以帮助消除信息传递中的歧义，使得沟通更加清晰和准确。

### 4. 识别异常和问题

数据可视化可以帮助人们更容易地发现数据中的异常值、离群点或潜在的问题，从而及时采取相应措施进行调整和改进。例如，在质量控制过程中，使用控制图可以很容易地观察到偏离正常范围的数据点，从而及时发现生产线上的问题。

### 5. 推动数据驱动决策

数据可视化可以促进组织和团队更加注重数据驱动的决策，使决策更加客观和科学。此外，数据可视化可以帮助各个层面的决策者更深入地了解业务情况，从而更好地利用数据制定决策。

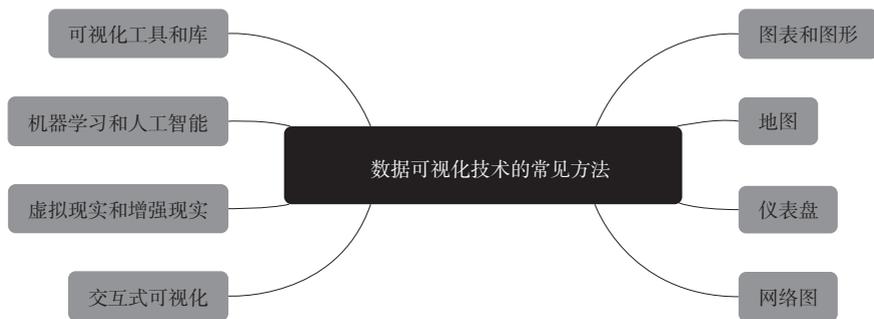
### 6. 增强记忆和理解

人们对于视觉信息的记忆更为深刻，通过可视化呈现数据的方式可以帮助人们更好地记忆和理解数据。相比于纯文字或数字的呈现方式，图表和图形更具有吸引力和记忆性，能够更好地传达信息和概念。

数据可视化是一种非常重要的数据分析和传播工具。它可以帮助人们更好地理解数据、做出决策并推动业务发展，并在很多方面都具有重要的意义。随着数据分析技术的不断发展和普及，数据可视化也将继续发挥重要作用，成为人们更好地利用数据实现创新和进步的重要手段。

## 1.4.2 数据可视化技术的方法

数据可视化技术的方法多种多样，可以根据具体需求和场景选择适合的方法和工具来实现数据的可视化展示。在选择使用哪种方法时，需要根据数据的类型、分析目标以及受众等因素进行综合考虑。图 1-21 展示了数据可视化技术的常见方法。



接下来将对一些数据可视化技术的常见方法进行讲解。

### 1. 图表和图形

图表和图形主要是指折线图、柱状图、饼图等。例如，在销售数据分析中，使用折线图可以清晰展示不同产品销售额随时间的变化趋势。

### 2. 地图

地图主要是指热力图、气泡地图、区域地图等。例如，通过热力图可以直观展示全球温室气体排放情况，帮助政府和环保组织制定相关政策。

### 3. 仪表盘

仪表盘是指集成了多个图表和指标的可视化工具。例如，一家电商公司可以利用仪表盘实时监控网站流量、销售额和用户行为数据，从而及时调整营销策略。

### 4. 网络图

网络图用于展示复杂关系和连接的数据。例如，社交媒体公司可以利用网络图展示用户之间的关注关系，以优化推荐算法和社交互动体验。



### 5. 交互式可视化

交互式可视化用户可以与可视化图表进行互动。例如，在线教育平台可以提供交互式学习大数据可视化课程，让学生通过操作图表加深对数据分析的理解。

### 6. 虚拟现实和增强现实

虚拟现实和增强现实通过相应技术，以更直观、沉浸的方式呈现数据。例如，在医疗领域，医生可以利用增强现实技术在手术前模拟患者的内部器官结构，以提高手术准确性。

### 7. 机器学习和人工智能

机器学习和人工智能可应用于数据可视化中，如自动聚类、异常检测、预测分析。例如，智能城市项目可以利用机器学习技术对城市交通数据进行预测分析，以优化交通管理和规划。

### 8. 可视化工具和库

可视化工具和库主要包括 Tableau、Power BI、D3.js、Matplotlib、ggplot2 等。例如，一家市场营销公司可以利用 Tableau 创建精美的报告和可视化图表，以便向客户展示市场分析和营销策略。

## 本章小结

本章首先介绍了数据技术的入门知识，包括数据的重要性、数据技术的定义、数据技术的系统框架、数据技术的理论基础，数据技术的常用工具和应用领域；其次介绍了数据获取技术的定义和意义、数据获取技术的方法、数据技术获取的关键问题；再次介绍了数据分析技术的定义和意义、数据分析技术的方法、数据分析技术的关键问题；最后介绍了数据可视化技术的定义和意义、数据可视化技术的方法、数据可视化技术的关键问题。通过本章的学习，帮助读者对数据技术有一个初步的了解，为后续的学习和应用打下坚实的基础。

## 即测即练



## 复习思考题

- 数据获取是指 ( )。
  - 从互联网上下载数据
  - 将数据存储于数据库中
  - 收集和提取数据以供进一步处理
  - 使用图表展示数据分析结果
- 下列哪种数据获取方法不需要使用 API? ( )
  - 网络爬虫
  - 数据库查询
  - 文件导入
  - 实时数据流
- 数据可视化的目的是 ( )。
  - 展示数据模型的准确性
  - 增加数据存储的效率
  - 提供对数据的洞察和理解
  - 加密和保护数据的安全性
- 下列哪个工具适合用于创建交互式数据可视化? ( )
  - Excel
  - Matplotlib
  - Power BI
  - SAS
- 数据挖掘是指 ( )。
  - 从数据库中提取有用信息
  - 对数据进行可视化展示
  - 使用机器学习算法发现模式和关联规则
  - 对数据进行分类和预测
- 下列哪种数据获取方法适合用于处理结构化数据? ( )
  - 网络爬虫
  - API 调用
  - 文件导入
  - 实时数据流
- 数据预处理包括以下哪些步骤? ( )
  - 数据清洗
  - 特征选择
  - 数据变换
  - 所有选项都正确
- 数据聚合是指 ( )。
  - 将多个数据集合并为一个
  - 将数据划分为不同的组别
  - 将数据转化为统一的格式
  - 将数据按照时间顺序排序
- 下列哪个工具适合用于处理大规模文本数据? ( )
  - Hadoop
  - Excel
  - Tableau
  - R 语言
- 下列哪种数据获取方法适合用于处理非结构化数据? ( )
  - 网络爬虫
  - API 调用



- C. 文件导入
- D. 实时数据流
11. 数据标准化是指 ( )。
- A. 将数据转化为 0 和 1 之间的值
- B. 将数据转化为正态分布
- C. 将数据按照一定比例缩放
- D. 将数据按照特定规则重新编码
12. 在数据分析中, 下列哪种方法适用于对数据进行关联分析? ( )
- A. 决策树算法
- B. K-means 聚类算法
- C. 关联规则挖掘
- D. 主成分分析