第一章

导　论

本部分内容简要介绍大数据的基本概念、大数据带来的变革、大数据的应用场景和人才需求以及大数据分析的一般流程。

第一节　大　数　据

大数据（Big Data）这个词最早可以追溯到 1997 年，Michael Cox 和 David Ellsworth 在一篇论文中使用了“Big Data”这个词。虽然那时还没有一个明确的定义。真正使“Big Data”这个词为人所广泛知晓的是 2001 年 Doug Laney 在一篇研究报告中提出了大数据的 3V 模型——Volume（大量）、Velocity（高速）、Variety（多样性），这也奠定了大数据概念的基础。

从 2008 年开始，随着互联网的发展，大数据开始广泛应用于网络搜索、社交网络分析等领域。“Big Data”一词也越来越频繁地出现在各种报道和文章中。2011 年，James Manyika 等发布了 *Big Data: the next frontier for innovation, competition, and productivity* 研究报告，系统阐述了大数据给全球经济带来的机遇和影响，使大数据成为当时的热词。2012 年，Gartner 正式将大数据纳入 IT 词汇，也使大数据的概念被进一步确立和明晰。除了 3V 模型，还出现了更多描述大数据的特征，如 Veracity（可信度）。

大数据概念的提出经历了一个逐步发展和丰富的过程，从最初的提法到影响力更大的模型，随着时代变迁不断拓展外延和完善内涵。目前大数据通常指数据量规模巨大到无法利用传统主流数据分析软件和工具，在合理时间内进行采集、管理、处理成为帮助企业经营决策的数据信息。大数据具有如下几个特点：

1. 规模大（Volume）

国际数据公司（IDC）报告：2018 年全世界产生的新数据为 33ZB，中国产生 7.6ZB，美国产生 6.9ZB，到 2025 年全世界产生的新数据将增至 48.6ZB。现在许多系统每天产生的数据达到 PB 级别。这要求能够进行海量并行计算，对 PB 级、EB 级的数据进行高效处理。数据量级换算关系如下：

8bit=1Byte 一字节

1024B=1KB(KiloByte)千字节

1024KB=1MB(MegaByte)兆字节

1024MB=1GB(GigaByte)吉字节

1024GB=1TB(TeraByte)太字节

1024TB=1PB(PetaByte)拍字节

1024PB=1EB(ExaByte)艾字节

1024EB=1ZB(ZetaByte)泽字节

1024ZB=1YB(YottaByte)尧字节
1024YB=1BB(BrontoByte)珀字节
1024BB=1NB(NonaByte)诺字节
1024NB=1DB(DoggaByte)刀字节

2. 类型多（Variety）

大数据来自各种不同的源头，有以文本为主的结构化数据，有半结构化的网页数据，还有大量网络日志、音频、视频、图片、地理位置等非结构化数据。要想从这些数据中提炼有价值的信息，这要求相关技术能高效地综合处理多种类型的数据。

3. 流转速度快（Velocity）

大数据的生成速度非常快，需要实时处理和分析以及快速获取有用信息。像网站日志、传感器等都在以每秒、每毫秒的速度生成数据，大数据处理在很多领域对系统响应速度有极高的要求。Jeff Johnson 在他 2010 年出版的 *Designing with the Mind in Mind* 一书中首次提出“1 秒定律”，在网络应用尤其是网页加载速度方面，1 秒定律显得尤为重要：

（1）0.1 秒的响应时间给用户的感觉是系统反馈立即出现。

（2）1 秒的响应时间不会打断用户的思路。

（3）10 秒的响应时间会让用户的注意力分散。

（4）但如果页面加载时间超过 10 秒，用户就会觉得系统工作缓慢，想离开。

1 秒定律要求要在秒级时间范围内给出分析结果，否则数据就会失去价值。

4. 价值密度低（Value）

大数据规模大，但信息含量低，如何通过机器算法更迅速有效地完成数据价值“提纯”是大数据背景下的关键问题。

5. 可信度低（Veracity）

大数据中包含的数据质量参差不齐，需要通过一些方法对数据进行清洗和验证，来确保数据的准确性和真实性，以及模型算法的稳健性。

只有满足规模大、类型多、流转速度快，以及处理不确定数据等要求，才能有效实现对大数据的分析利用。这也是大数据技术取得进展的基本动力所在。

第二节　金融大数据

金融行业运行所产生的大数据就是金融大数据（financial big data）。金融行业是从事金融服务的特殊行业，天然具有“海量客户和大数据”的特点。金融行业包括银行、证券、保险、信托和租赁等。这些金融行业中的金融机构在从事金融服务的过程中，产生了大量的客户信息、产品档案资料、报价数据、市场行情数据和客户交易数据等。这些数据记录了金融机构业务活动过程，蕴含着对金融机构未来开展业务活动具有重大指导价值的信息，是金融机构重要的数据资产。对这些数据的开发和利用，可以重塑金融服务的商业模式、创新运营方式，极大提高金融服务的效率。

从产品本质上看，任何金融产品包括各类存款、贷款、基金、债券、股票、期货、期权、衍生品等，都是围绕资金流转约定交易双方权利与义务的一种契约，契约的存续期从隔夜到 30 年及以上不等。记录这些产品及交易就会产生大量数据。此外，金融机构的业务管理和行业监管也需要大量统计汇总报表，这也会产生大量管理和监管数据。从金融产品本质特性看，金融行业天然就是产生大数据的行业，因此也是大数据相关技术优先受益的行业。

从金融活动实际运行来看，金融行业是一个高度依赖信息的行业。金融活动的有效运行需要大量的信息支持。例如，银行在进行贷款决策时，需要了解借款人的信用状况、财务状况、经营情况等信息，以便评估其还款能力和贷款风险。证券公司在投资决策时，需要对市场趋势、公司财务、行业情况等信息进行深入分析，以评估投资价值，确定投资策略和风险控制措施。因此，获得信息的数量、质量和效率对金融活动运行至关重要。大数据相关技术可以解决大数据存储、管理和使用问题，高效组织不同来源的数据，快速分析、提取数据中所蕴含的价值，用于指导金融业务活动。

具体来说，金融大数据相关技术的运用对金融机构产生以下影响。

1. 节约成本，提高效率

金融大数据技术的运用使得金融机构能够更高效地处理和分析数据，从而降低数据管理成本。例如，通过使用大数据存储技术，金融机构可以将海量的客户数据、交易数据等存储在低成本、高效率的数据中心，避免数据冗余和重复存储。此外，金融大数据技术还可以帮助金融机构实现自动化数据采集、处理和分析，减少人工干预和错误，提高数据准确性。

2. 挖掘数据价值，促进金融业务创新

金融大数据技术可以帮助金融机构挖掘新的业务机会，促进创新发展。通过对市场数据、行业数据进行深入分析，金融机构可以及时掌握市场动态和行业趋势，开发出更符合市场需求的产品和服务。例如，保险公司可以利用大数据技术分析车辆行驶数据、天气数据等，开发出更精准的车险产品和服务。

3. 保障合规，提高风险控制能力

金融大数据技术可以实现全面的数据采集、数据脱敏与脱密，搭建风险评估模型，帮助金融机构更准确地评估风险，提高风险控制能力。通过对客户数据进行深入分析，金融机构可以获取客户的更多信息，如信用状况、消费习惯等，从而更准确地评估客户的信用风险。此外，金融大数据技术还可以帮助金融机构监测市场动态、识别异常交易等，及时发现潜在的风险隐患。

4. 优化业务流程，提高服务质量，提升客户服务体验

金融大数据技术可以帮助金融机构更好地了解客户需求，优化客户服务体验。通过对客户的行为数据、偏好数据等进行深入分析，金融机构可以为客户提供更个性化、更精准的产品和服务。例如，银行可以根据客户的消费习惯和偏好推荐信用卡产品、贷款产品等，促进创新，提升服务广度，提高客户满意度和忠诚度。

第三节　大数据带来的变革

随着大数据时代的到来，人类社会正经历着前所未有的变革。大数据不仅改变了人们看待世界的方式，还对人类的思维方式、行为方式及探索世界的方式产生了深刻影响。

1. 深刻影响和改变着人类的思维方式、行为方式以及探索世界的方式

在思维方式上，人类将更加依赖数据思维，根据数据进行判断和决策，而不是仅仅依靠主观经验或直觉。与此同时，基于算法和数据模型进行决策也会扮演更重要的角色，逐渐取代依赖个人经验的决策方式。

另一个重大转变是思维方式从传统的小样本推断逻辑向依赖海量大数据的发现逻辑转型。大数据时代，人们有机会从海量数据中发现隐藏的模式和规律，得到更为准确和全面的结果。同时，大数据时代，追求确定的因果关系变得更加困难，而发现变量之间的相关性和规律则更具价值。这也将推动人类思维从传统的因果关系推断逻辑，向更多依赖变量相关性分析和概率统计推断的方式转型。

可以说，大数据将重塑人类的认知模式，思维方式将更严谨、开放和全面，这也将是一场新的科学发现革命。但我们也要保持理性，注意大数据应用的负面影响（如隐私泄露、数据错误、数据偏见、模式误用等），以更智慧的方式使用这一新生力量。

2. 导致商业营销变革和业务创新

大数据带来营销手段的变革。企业传统的营销手段往往依赖广告、促销和人员推销等手段，但这些方式在大数据的背景下显得捉襟见肘。通过收集和分析大量数据，企业可以更好地了解客户的行为习惯、购买偏好和需求，从而提供更加个性化的产品和服务，实现精准营销，提高客户满意度。

大数据赋能业务创新，通过发掘数据新价值，企业可以拓展新型服务方式与渠道，并提供持续优化的能力。例如，汽车座椅传感器将收集到的座椅压力数据用于汽车防盗系统，通过检测座椅上的压力变化来判断是否有人非法进入车辆。此外，还可以识别驾驶员的坐姿，用于自动驾驶技术，以确保驾驶员和乘客的安全。

3. 带来商业决策变革

首先，商业决策基础从过去依靠个人经验，转向现在更多依赖数据分析。通过分析大量的数据，人们可以看到隐藏在数据背后的模式和趋势，减少主观经验的局限，从而做出更准确和科学的决策。

其次，决策视野从过去的局部优化，拓展到基于大数据的整体决策。大数据通过整合多源异构数据，提供了反映全局业务情况及客户全景画像的信息，让决策者可以审视全局，也启发了交叉视角，实现了决策的全局最优。

再次，决策方式从传统的人工经验转变为运用智能算法和模型，提高决策的科学性和准确性；同时，大数据也加快了决策的速度。传统的决策方式往往存在滞后的现象，这意味着在情况发生变化时，决策者可能无法及时做出反应。然而，通过大数据分析，人们可以近乎实时地支持决策，使决策者能够快速响应变化的情况。

最后，降低决策错误的成本。大数据支撑的决策可以快速进行试错，一个决策方案失败后可以及时获得反馈，并迅速止损，避免错误扩大，快速调整推出新方案，降低风险。

4. 逐渐改变公共部门的管理与服务方式

随着大数据技术的发展，公共部门正逐渐从传统的经验治理转向数据驱动治理，从事后被动应对转向提前预判和主动出击。例如在传染病防控中，过去依靠病例报告和流行病学调查，现在可以通过互联网搜索数据、社交媒体等大数据实时监测传染病早期信号，提前预测疫情发展，及早防控；在自然灾害防治中，通过大量气象数据，利用模型提前几天预测可能的暴雨洪水，做好防范准备；在社会治安管理中，通过大量犯罪数据分析，发现犯罪规律，在高危地区和高危时间提前部署警力资源进行重点预防。在金融安全方面，通过大量交易数据分析，发现异常交易模式和交易行为，提高反欺诈和反洗钱的能力，同时帮助公安部门监测和预防犯罪活动。在城市交通管理中，通过收集和分析城市中各种传感器、设备和系统产生的数据，优化交通路线规划，提高交通效率，提升城市居民的生活质量。

同时，大数据还可以提升政务透明度，推动政务数据的开放和共享，加强公众对政府行为的监督。公众可以通过查看相关数据，了解政府决策和行动的依据，促进公共事务的透明与公正。总的来说，大数据改变了公共部门的管理与服务方式，使其能够更加高效、准确地应对各种挑战，并提供更好的公共服务。

第四节　大数据应用场景

大数据在现代社会中的应用场景已经非常广泛，以下是一些常见的大数据分析应用场景。

（1）市场营销：通过大数据分析，企业可以了解客户的偏好和行为，以制定更有效的市场营销策略，提高销售额。

（2）健康医疗：利用大数据分析技术，可以对大量医疗数据进行分析，从而更好地了解疾病的流行病学特征和趋势，辅助医生进行更准确的诊断和治疗，制定更好的治疗方案，实现精准医疗，提高医疗保健质量。

（3）金融服务：金融机构可以通过大数据分析技术对市场数据、交易数据、用户行为等进行分析，提高风险管理能力，预测市场趋势和风险，优化投资组合和贷款决策等。

（4）物流与运输：通过大数据分析，可以更好地了解物流和运输的供应链和运营，提高效率和降低成本，同时也能更好地预测和应对交通拥堵、交通事故等情况。

（5）政府管理：政府可以通过大数据分析技术对社会经济、环境、人口等方面进行分析，以更好地制定政策、优化资源分配和公共服务，提高政府管理水平和效率。

（6）社交网络：社交媒体和网络平台可以通过大数据分析用户的兴趣和偏好，提供更好的个性化推荐和服务，同时也可以通过数据分析监控和应对虚假信息、网络暴力等问题。

（7）电商行业：大数据在电商领域的应用主要体现在用户画像、精准营销、价格预测等方面。例如，电商平台可以根据用户的购买行为、浏览记录等数据，为用户推荐相关的产品；同时，通过对市场数据的分析，预测未来价格走势，制定合理的价格策略。

（8）制造业：制造业可以利用大数据分析技术对生产过程中的各种数据进行分析，从而提高生产效率、优化供应链、降低成本等。

这只是大数据应用场景的一部分，随着数据技术和算法的不断发展，大数据分析将在更多的领域发挥作用。

第五节　大数据分析的类型

根据分析目的的不同，可以将大数据分析划分为以下几种类型：

1. 描述性分析

描述性分析主要回答“发生了什么”的问题。这种类型的分析是通过统计分析来描述或总结数据总体特征和分布情况，以帮助使用者理解数据，以及数据的关联性和依赖关系。常用的指标有平均值、中位数、众数、方差等统计指标。例如，描述性分析可以显示一组员工的销售额分布以及每位员工的平均销售额。

2. 诊断性分析

通常是在描述性分析确定“发生了什么”的基础上，诊断性分析回答“为什么会发生这种情况”。假设描述性分析显示医院中患者异常涌入。进一步深入研究数据可能会发现这些患者中多数都有特定病毒的症状。这种诊断分析可以帮助医生确定传染因子（“为什么”）导致患者涌入。它可以帮助企业更好地了解客户行为，改善业务流程，提高客户满意度，并最终提高企业的效率和利润。诊断性分析常用的方法有：假设检验、相关性分析、关联分析、判别分析、回归分析、聚类分析、时间序列分析等。

3. 预测性分析

预测性分析回答的问题是“未来会发生什么”。描述性分析和诊断性分析已经描绘出当前情况的特征和规律，并确定问题产生的原因。预测性分析旨在通过已有的数据预测未来的趋势和结果。例如，已经研究了某产品的历史销售情况，发现它在每年的 9 月和 10 月期间销量最好，从而预测来年会出现类似的高点。预测性分析的常用方法有：时间序列预测、回归模型预测、决策树预测、神经网络预测等。

4. 规范性分析

规范性分析采用从前三种类型的分析中收集的所有见解，并使用它们来形成公司应如何行动的建议。使用之前的示例，这种类型的分析可能会建议一个市场计划，以便在高销售额月份的成功基础上再接再厉，并在较低的月份寻求新的增长机会。规范性分析回答了“应该怎么做”这个问题。

规范性分析是**数据驱动决策**概念发挥作用的地方。数据驱动决策（data-driven decision-making，DDDM），可以定义为基于事实、数据和指标，而不是直觉、情感或观察做出战略业务决策的过程。这听起来可能很明显，但在实践中，并非所有组织都能实施数据驱动决策。根据麦肯锡的调查，数据驱动型公司更善于获取新客户、保持客户忠诚度和实现高于平均水平的盈利。

第六节 大数据分析方法与实现工具

一、大数据分析方法

随着金融业务的不断扩张和数据的不断增长，金融领域中的数据量已经达到了海量级别，其中包括交易数据、市场数据、客户数据、企业数据、经济数据等。对这些数据进行有效的分析和利用，可以帮助金融机构更好地了解市场和客户，优化业务流程和提升业务效率，提高风险控制能力，推动金融业的创新和发展，发挥海量数据的价值。

在实践中，常用的大数据分析方法有以下几种。

1. 数据挖掘

数据挖掘（data mining）是一种用于探索、分析和发现数据中隐藏关系和模式的方法，它通过使用相关算法和统计模型，在大型数据集中挖掘出有价值的信息和知识。这些信息包括数据集中的模式、异常和相关性，可以用来预测结果并做出更好的决策。数据挖掘可以应用于客户分类、信用风险评估、投资组合优化等方面，帮助金融机构预测市场变化、识别潜在风险等。以下是一些例子：

（1）营销：数据挖掘可以帮助探索大型数据库，改善市场细分、客户忠诚度、交叉销售和促销策略。

（2）零售：数据挖掘可以帮助分析客户行为、偏好、趋势和销售模式，以优化库存、定价和促销。

（3）银行业：数据挖掘可以帮助检测欺诈、评估信用风险、管理客户关系和优化现金流。

（4）医学：数据挖掘可以帮助诊断疾病、发现新药、分析基因组数据并提高医疗保健质量。

（5）电视和广播：数据挖掘有助于个性化内容推荐、衡量受众评分和偏好，并优化广告收入。

2. 机器学习

机器学习（machine learning）是人工智能（AI）的一个重要分支，是使用一定的算法从数据中学习模型并预测新的数据结果的技术，也是完成 AI 大部分工作的核心技术。机器学习在各个领域都有很多应用。以下是一些例子：

（1）推荐系统：机器学习算法可以分析用户的历史购买记录、搜索历史和浏览行为，从而生成个性化的产品推荐。这种方法在电商网站、音乐流媒体平台和视频网站上广泛应用。

（2）垃圾邮件过滤：机器学习算法可以学习正常邮件和垃圾邮件的特征，从而自动将垃圾邮件从收件箱中过滤出来。这种技术广泛应用在电子邮件服务中。

（3）人脸识别：机器学习算法可以通过分析人脸的图像来识别人的身份。这种技术广泛应用在安全监控、社交媒体和手机生物识别中。

（4）语音识别：机器学习算法可以将人类语音转换成文本。这种技术广泛应用在智能助手、手机操作系统和汽车系统中。

（5）自动驾驶：机器学习算法可以通过分析摄像头和传感器数据来识别道路上的物体，规划行驶路径并控制车辆。

（6）智能家居助理：机器学习可以帮助理解自然语言命令、执行任务、回答问题并从用户反馈中学习。

在金融领域，机器学习可以用于自动化交易、风险评估、客户分类等方面，提高金融机构的业务效率和准确性。

3. 数据可视化

数据可视化（data visualization）是将原始数据转换为图形、图表、图像等视觉元素进行表达和呈现的过程，以帮助用户更好地了解数据中的基本趋势和模式。

4. 自然语言处理

自然语言处理（natural language processing，NLP）是一种基于计算机科学和语言学的交叉学科，能够帮助计算机理解人类语言。在金融领域，自然语言处理可以用于分析金融新闻、舆情信息，对市场情绪进行情感分析，识别市场热点等。

5. 时间序列分析

时间序列分析（time series analysis）是一种统计学方法，能够用于分析时间序列数据的规律和趋势。在金融领域，时间序列分析可以用于股票价格预测、货币市场波动分析、风险评估等方面。

二、金融大数据分析实现工具

金融大数据分析常用的工具有很多，以下是其中的一些：

（1）R 语言：R 语言是一种广泛用于数据分析和统计建模的编程语言。R 语言具有丰富的数据分析和可视化功能，以及大量的数据处理和统计学习包，是金融大数据分析的主要工具之一。

（2）Python：Python 是一种通用编程语言，也是金融大数据分析的主要工具之一。Python 具有丰富的科学计算和数据分析库，例如 NumPy、SciPy、pandas 和 scikit-learn 等，能够支持大规模数据处理和机器学习算法。

（3）SQL：SQL（structured query language）是一种用于管理关系数据库的标准化语言。金融机构通常有海量的数据需要处理，SQL 可以帮助分析师有效地查询和分析大规模数据。

（4）Hadoop：Hadoop 是一个分布式计算框架，可以支持海量数据的存储和处理。Hadoop 提供了分布式存储和计算功能，能够帮助分析师在大规模数据集上进行数据挖掘和分析。

（5）Spark：Spark 是一个基于内存计算的大数据处理框架，能够支持快速的数据处理和机器学习算法。Spark 提供了丰富的机器学习和图形处理库，也能够与 Hadoop 等其他大数据处理工具集成使用。

（6）Excel：Excel 是一种电子表格软件，也是金融数据分析中最常用的工具之一。Excel 可以帮助分析师进行数据可视化、数据透视和统计分析等操作，同时也提供了一些数据分

析插件。

以上这些工具是金融大数据分析中最常用的一些工具，分析师可以根据自己的需求和熟练程度选择适合自己的工具。

第七节　金融大数据分析一般流程

金融大数据分析是利用大数据分析的相关技术对大量金融数据进行分析，以发现数据中隐藏的模式和趋势，得出有意义见解的过程。它可用于识别潜在的机会或风险，以及深入了解客户行为和偏好，更好更快地制定金融业务决策。从具体实践来看，金融大数据分析一般流程如图 1-1 所示，每个阶段都有其特定的任务和目标，以确保最终的分析结果准确、可靠。

（1）定义问题：金融机构面临的核心业务问题，是金融大数据分析的出发点和落脚点。

（2）数据采集：数据采集主要是围绕业务问题，采集与业务问题相关的所有多源异构数据，包括结构化数据（账务数据、报表等）和非结构化数据（文本、媒体等），为后续分析提供数据基础。

（3）数据清洗：数据清洗主要是处理缺失值、异常值、重复数据、错误数据等，确保数据质量。

（4）特征预处理：特征预处理是进行特征选择和特征构造，主要目的是满足建模需要，提高模型性能，并改善模型解释性和鲁棒性。

（5）数据选择：根据目标问题的特点和性质，选择合适的算法模型，进行训练和优化，以得到稳健的结果。

（6）模型评估：通过报表、可视化等方式对关键分析结果进行呈现，帮助人们理解数据和模型表现。

（7）模型应用：将建立的模型部署到生产环境，用于支持业务决策。这个阶段需要将模型应用到实际业务中，通过实践来检验模型的准确性和可靠性。同时，也需要根据实际应用情况进行相应的调整和完善，以不断提高模型的性能和效果。

通过金融大数据分析，金融机构能够更好地了解客户的需求和市场的发展趋势，从而制定更加科学、合理的业务决策，实现持续发展和稳定增长。

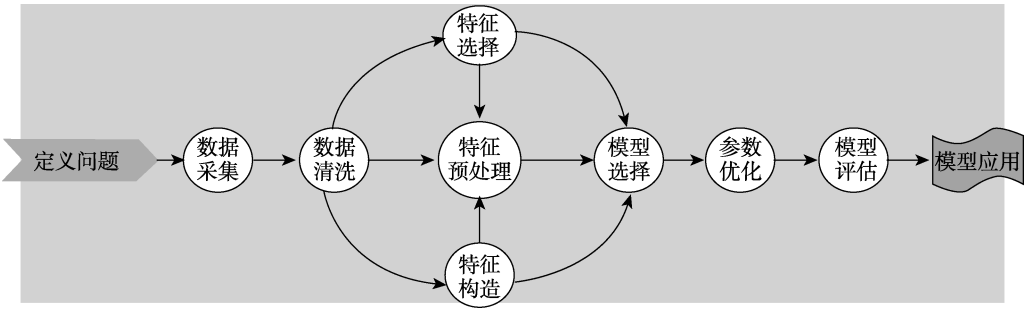

图 1-1　金融大数据分析一般流程

通过上述内容的介绍，可以看出金融大数据分析需要掌握以下几方面的知识：

（1）数学和统计学，这是大数据分析的基础，可以帮助理解和运用各种算法和模型；

（2）编程语言，如 R 语言、Python 等，这是大数据分析的工具，可以实现各种功能和逻辑；

（3）业务知识，这是大数据分析的应用，可以帮助理解数据并解决各种业务问题和需求。

本章关键词

大数据　金融大数据　大数据分析方法

思考题

1. 大数据分析的类型有哪些？
2. 常用的金融大数据分析实现工具有哪些？
3. 如何理解金融大数据分析的一般流程？

即测即练

自学自测  扫描此码