第一章

财务大数据分析概述

【学习重点】

1. 了解大数据的概念及特点。
2. 了解常见大数据分析工具。
3. 掌握大数据分析流程。

开篇案例-大数据的典型应用

第一节　财务大数据概述

在传统互联网、物联网、移动互联网等领域，每天都在产生大量的数据，并且每时每刻都在快速增长，这种超大规模、高增长率的数据集被称为大数据(big data)。经济高速发展的时代，科技发达、信息畅通，大数据就是这个时代的产物，它将在众多领域掀起变革的巨浪，未来将是数据科技(data technology，DT)的时代。

一、大数据的定义

随着计算机技术的发展和互联网用户数量的激增，手机已经成为我们工作生活中的重要工具。近年来，大数据已经渗透到各行各业，成为重要的生产力。越来越多的企业通过大数据挖掘优化企业战略，帮助企业寻找新的业务增长点，发现潜在的商业价值，获得更大的利润。那么大数据是如何产生的呢？先来看看2020年我们处理的数据规模。

(1) 我们有约100亿台移动设备在使用，还不包括笔记本电脑和台式机。

(2) 我们每天进行超过10亿次的谷歌、百度搜索。

(3) 全球每天发送大约3000亿封电子邮件。

(4) 全球每天撰写超过2.3亿条推文。

你一定听说过千字节(KB)、兆字节(MB)、吉字节(GB)，甚至是太字节(TB)，这些数据单位是我们在工作生活中可能经常遇到的，它们足以量化我们存储的文件、视频等。然而在未来，这些常见的单位会捉襟见肘，因为仅在2020年底，全世界的数据量已经达到59泽字节(ZB)。这个数字是可观测宇宙中星星数量的50倍。那么到底何为大数据呢？

“大数据”概念最早出现在1980年，由著名的未来学家阿尔文·托夫勒在其著作《第三次浪潮》中首次提出。今天，我们已经能充分感受到大数据的魅力和影响力。我们看到“大数据”一词，会先入为主地认为大数据就是“大量数据”或者“强大的数据”。那么何为大数

据，不同组织从不同角度给出了不同的定义。

亚马逊公司 John Rauser 对大数据的定义为，大数据是任何数据量超过了一台计算机处理能力的数据。

国际数据公司 IDC 对大数据的描述为，大数据一般涉及两种或两种以上的数据形式，指出大数据具有多样性。

维基百科对大数据的定义为，利用常用软件工具获取、管理和处理数据所耗费的时间超过可容忍时间的数据集。大数据对处理速度要求比较快，因为信息有时效性。

Gartner 咨询机构对大数据的定义为，需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。说明了传统方式无法处理大数据，需要新的处理模型，大数据具有更强的决策力、洞察发现力和流程优化能力，需要大数据分析来实现。

麦肯锡全球研究所对大数据的定义为，一种规模大到在获取、存储、管理、分析方面大大超出了传统数据库软件工具能力范围的数据集合，具有海量的数据规模、快速的数据流转、多样的数据类型和价值密度低四大特征。

《国务院关于印发促进大数据发展行动纲要的通知》对大数据的定义为，大数据是以容量大、类型多、存取速度快、应用价值高为主要特征的数据集合，正快速发展为对数量巨大、来源分散、格式多样的数据进行采集、存储和关联分析，从中发现新知识、创造新价值、提升新能力的新一代信息技术和服务业态。

二、大数据的特征

目前普遍认为大数据具有 volume、velocity、variety、veracity 和 value 五个特征，如图 1-1-1 所示。

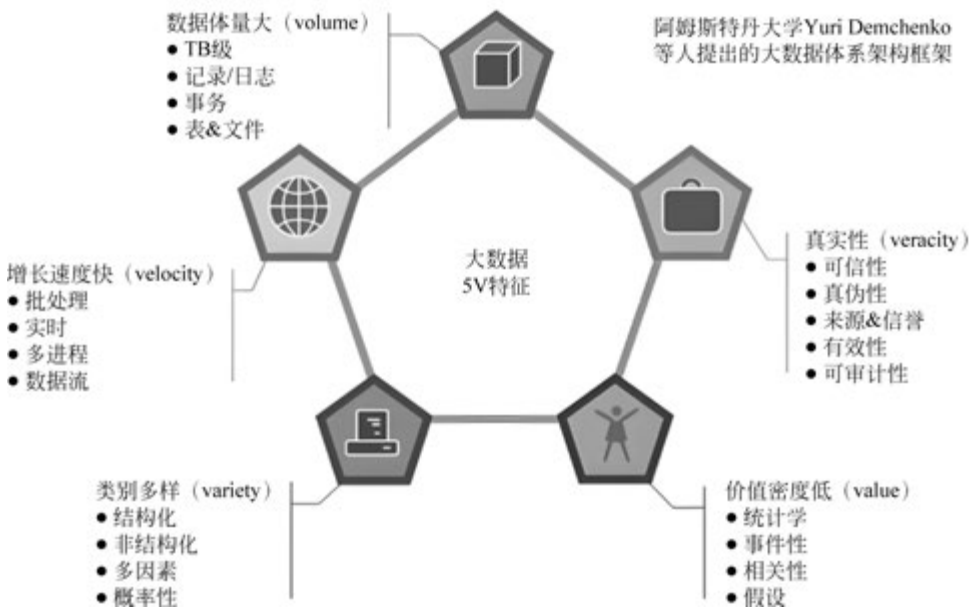

图 1-1-1　大数据的五个特征

1. 数据体量大(volume)

大数据的特征首先是数据规模大。随着互联网、物联网、移动互联网技术的发展,人和事物的所有轨迹都可以被记录下来,使数据呈现爆发性增长。体量大包括数据体量巨大(超过一台计算机存储能力,分布式存储)和数据增量大(每时每刻不断产生巨量的数据)。

2. 增长速度快(velocity)

随着云计算、移动互联网、物联网的飞速发展,数据量迅猛增长。据统计,2020 年全球每秒产生 1.7MB 的数据量,平均每天会生成 2.2EB(23 亿 GB)数据。

3. 类别多样(variety)

现在可以从手机、邮箱、网页、视频等不同的数据源获得数据,这些不同类型的数据大致可分为三类:结构化数据(Excel 表,如银行对账单)、半结构化数据(不是传统行列数据,具有一定规律的文件,如日志)和非结构化数据(无规律数据,如网络文字、图片、视频等)。

另一方面大数据的来源多样,如互联网上的搜索数据、电子商务平台的交易数据、社交平台上的留言数据、物联网上各种传感器收集的数据。

4. 数据应具有真实性(veracity)

数据真实性高是指数据必须符合数据质量、完整性、可信性、准确性要求。由于数据可能从不同来源收集,保证数据的真实性是大数据分析的基础。

5. 价值密度低(value)

价值密度一般与数据总量的大小成反比。如一个视频监控中可能有用的数据仅有一两秒,又如电子商务网站的交易数据,单笔交易数据几乎没有分析的价值,但将大量的交易数据汇聚到一起进行分析,往往能发现一些市场规律。

此外,大数据为了获取事物的全部细节,直接采用原始数据,保留了数据的原貌,这就导致在呈现数据全部细节的同时也引入了大量没有意义甚至错误的信息。因此,相对于特定的应用,大数据关注的非结构化数据具有价值密度低的特点。

三、财务大数据的概念

财务大数据是指在财务领域中产生的,具有海量规模、快速流转和多样类型等特征的数据集合,以及对这些数据进行采集、存储、管理、分析和可视化呈现的一系列技术与方法。

财务大数据分析的数据来源包括企业内部财务系统、企业业务系统以及外部数据。其中企业财务系统提供会计凭证、账簿、财务报表等传统财务数据,以及预算管理、成本控制、资金管理等模块产生的数据。企业业务系统提供如销售、采购、库存、生产等业务部门的系统中与财务相关的数据,例如,销售订单金额、采购发票信息、库存盘点数据等。在进行财务大数据分析时,为了发现问题仅靠内部数据是不够的,还需要借助外部宏观和微观的数据,这些数据包括宏观经济数据、行业数据、市场行情数据、税务法规信息,以及社交媒体上

与企业相关的舆情数据等。

财务大数据蕴含着丰富的信息，通过对其进行深入分析，可以帮助企业管理者了解企业的财务状况、经营成果和现金流量，发现潜在的风险和机会，为决策提供有力支持。目前财务大数据分析被广泛应用于财务分析与决策支持、风险管理、成本控制以及税收筹划等领域。应用最广的是财务分析，通过对大量财务数据和相关业务数据的分析，生成准确、全面的财务分析报告，为企业的战略决策、投资决策、融资决策等提供数据支持。通过上述财务分析指标，企业也可以实现实时监控风险，如偿债能力、流动性风险、信用风险等，通过大数据分析及时发现潜在的风险预警信号，并采取相应的风险控制措施。在成本控制方面，通过大数据技术可以分析成本结构和成本驱动因素，优化采购流程、生产计划和库存管理，降低成本，找出成本节约的空间和优化点。在税收筹划方面，利用大数据分析税务法规和企业的业务数据，进行合理的税务筹划，降低企业的税务负担。分析企业的业务模式和交易结构，寻找符合税收优惠政策的业务场景，合理享受税收优惠。

第二节　财务大数据分析流程及工具

一、财务大数据分析流程

财务大数据来源于互联网、物联网、企业 ERP 系统等，经过大数据处理系统的分析挖掘，产生新的知识用以支撑决策或业务的自动智能化运行。从数据在信息系统中的生命周期看，大数据从数据源经过分析挖掘到最终获得价值一般需要经过 5 个阶段：数据采集、数据预处理、数据存储、数据分析和挖掘、数据可视化，如图 1-2-1 所示。

图 1-2-1　大数据处理流程图

1. 数据采集

数据采集是通过网络爬虫、系统日志采集、传感器等工具从互联网、物联网、交互型社交网络，以及移动互联网等获取多种类型海量数据的过程。

2. 数据预处理

数据预处理是数据分析和挖掘的基础，是对接收数据进行清洗、集成、特征选择、标准化等操作，并最终作为数据分析的输入数据集的过程。

(1) 数据清洗：从现实世界中采集到的数据一般是不完整、有噪声且不一致的。数据清洗过程主要包括数据的默认值处理、噪声数据处理、不一致数据的处理。

(2) 数据集成：数据集成过程是将多个数据源中的数据合并存放到一个一致的数据存储中，其中数据源可包括多个数据库、非关系型(NoSQL)数据、网页及非结构化文本等。

(3) 特征选择：在有限的样本数目下，用大量的特征设计分类器，计算开销大而且分类性能差。选择出重要的特征可以缓解维数(即特征的个数)灾难问题，而去除不相关特征可以降低数据分析任务的难度，通过特征选择可以达到降维、提升模型效果和性能等目标。

(4) 数据标准化：主要包括数据同趋化处理和无量纲化处理两个方面。数据同趋化处理主要解决不同性质数据问题，如定性数据和定量数据如何比较、如何相加；数据无量纲化处理主要解决数据的可比性，如消除各个特征之间数量级的差异。

3. 数据存储

大数据的存储系统需要以极低的成本存储海量数据，还要适应多样化的非结构化数据的管理需求，具备数据格式上的可扩展性。大数据存储的数据库常采用 NoSQL 数据库，大数据存储的文件系统常采用分布式文件系统(Hadoop Distributed File System，HDFS)。

4. 数据分析和挖掘

数据分析是指利用相关数学模型及机器学习算法对数据进行统计、预测和文本分析。数据分析可分为预测性分析、关联分析和可视化分析。数据分析的主要方法有探索性数据分析方法、描述统计法等。

数据挖掘(data mining)是从数据库的大量数据中挖掘出有用的信息，即从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，发现隐含的、规律性的、人们事先未知的，但又潜在有用的并且最终可理解的信息和知识的过程，所挖掘的知识类型包括模型、规律、规则、模式、约束条件等。

5. 数据可视化

数据可视化是通过图形、图表、地图、仪表盘等视觉元素，将抽象的数据转化为直观的可视化形式的过程。其核心目标是通过视觉感知增强人类对数据的理解，揭示数据背后的模式、趋势、异常和关联性。人类大脑处理视觉信息的速度比文字快 6 万倍，图表能帮助用户快速抓住数据的关键特征，如峰值、异常值、分布形态。常见的图表有柱状图、折线图、散点图、饼图、词云图等。

二、财务大数据分析工具

财务大数据分析工具种类繁多，各自具有不同的特点和适用场景。以下是一些常用且重要的大数据分析工具。

1. Excel

作为微软办公套装软件的重要组成部分，Excel 可以进行各种数据的处理、统计分析和辅助决策操作。它广泛应用于管理、统计、财经、金融等众多领域，是数据分析师主要的常

用工具。Excel 的特点是提供了丰富的函数、图表和数据透视表等功能，使得数据分析变得直观且易于操作。

2. Hadoop

Hadoop 是一个能够对大量数据进行分布式处理的软件框架，具有高扩展性和高可靠性。它主要由 HDFS 和 MapReduce 两个核心组件组成，用于存储和处理大数据集。Hadoop 以可靠、高效、可伸缩的方式进行数据处理，通过维护多个工作数据副本确保针对失败节点的重新分布处理。

3. R 和 Python

R 和 Python 都是开源的分析软件，具有强大的数据分析能力和丰富的类库。它们可以用于数据处理、统计分析、机器学习等多个方面。R 语言提供了从基本统计数的计算到各种试验设计的方差分析、相关回归分析，以及多变数分析的多种统计分析过程。Python 则以其简洁清晰的语法和丰富强大的类库而著称，特别适合处理大规模数据和高精度建模。

4. Tableau Software

Tableau Software(简称 Tableau)是一款用于快速分析、可视化并分享信息的工具。它基于斯坦福大学的突破性技术，可以在几分钟内生成美观的图表、坐标图、仪表盘与报告。Tableau 提供了智能建模和数据监控功能，使得数据分析结果更加直观易懂，是数据可视化领域的佼佼者。

5. 其他工具

除上述工具外，SAS、SPSS、Power BI、Spark 等也是常用的大数据分析工具。它们各自具有独特的功能和特点，如 SAS 提供了丰富的统计分析过程，SPSS 则擅长于非线性回归、聚类分析等高级统计分析，Power BI 是微软推出的商业智能工具，Spark 则是一个快速、通用的大数据处理引擎。

综上所述，大数据分析工具种类繁多，选择适合自己需求的工具是关键。无论是 Excel 这样的初级工具，还是 Hadoop、R、Python 这样的高级工具，抑或是 Tableau 这样的可视化工具，都能在不同场景下发挥重要作用。

练习题

拓展阅读-大数据的未来发展趋势

1. 请简述大数据的特征。
2. 请简述财务大数据分析的流程。

自测题

自测题