财经商贸类专业创新型“互联网+”精品教材

财务大数据分析

主　编　陆培中　李　强
程　竞

扫描二维码
共享立体资源

财务大数据分析

主编　陆培中　李　强　程　竞

北京出版集团
北京出版社

北京出版集团
北京出版社

图书在版编目（CIP）数据

财务大数据分析 / 陆培中，李强，程竞主编 . —北京 : 北京出版社，2023.1

ISBN 978-7-200-17804-3

Ⅰ. ①财… Ⅱ. ①陆… ②李… ③程… Ⅲ. ①财务管理—数据处理—高等学校—教材 Ⅳ. ① F275

中国国家版本馆 CIP 数据核字（2023）第 009558 号

财务大数据分析

CAIWU DASHUJU FENXI

主　　编：陆培中　李 强　程 竞
出　　版：北京出版集团
　　北 京 出 版 社
地　　址：北京北三环中路 6 号
邮　　编：100120
网　　址：www.bph.com.cn
总 发 行：北京出版集团
经　　销：新华书店
印　　刷：定州启航印刷有限公司
版 印 次：2023 年 1 月第 1 版　2023 年 1 月第 1 次印刷
成品尺寸：185 毫米 ×260 毫米
印　　张：18.5
字　　数：394 千字
书　　号：ISBN 978-7-200-17804-3
定　　价：55.00 元

教材意见建议接收方式：010-58572162　邮箱：jiaocai@bphg.com.cn
如有印装质量问题，由本社负责调换
质量监督电话：010-82685218　010-58572162　010-58572393

目 录

项目一　财务大数据认知

» 学习目标

知识目标

- 掌握大数据的定义及特征；
- 掌握财务大数据的概念及特征；
- 了解大数据分析五步法方法论；
- 掌握数据收集、数据清洗、数据挖掘等大数据工具操作使用；
- 掌握撰写报告的基本要点。

能力目标

- 能够理解财务大数据的应用场景；
- 能够操作使用数据收集、数据清洗、数据挖掘与分析等大数据工具；
- 能够利用大数据分析五步法方法论进行项目分析。

素养目标

- 具有良好的政治素养与道德修养、高度社会责任感和敬业精神，掌握财务大数据的基本理论和基本技能，不断提升自我学习能力和接受新知识的能力。

项目引例

交通运输部的数据显示，2023 年的春运客流量约有 20.95 亿人次，大运量的背后是强大的运力支撑。铁路部门科学的调度，充分运用 12306 大数据分析掌握客流规律，精准实施“一日一图”，科学调配运力，充分有效释放铁路运能。人员到站后，怎么有序尽快疏解，又是一道极具挑战的难题。春运高峰期，广州南站 10 分钟内需要疏解 25 000 名旅客，面对如此庞大的客流量，广州智慧交通系统通过与铁路部门开行计划、售票数据对接，动态调集站场周边的出租车和网约车，加大公交和地铁开行密度，让旅客进站、出站的速度更快捷，大幅度提高全国铁路网的运输效率的同时，还充分解决了旅客的出行需求。种种事实表明，信息技术的快速发展，尤其是大数据在各行各业的广泛应用，极大地影响了我们的日常生活。那么，什么是大数据？大数据有哪些特征和应用场景？从哪里获取相关的大数据？如何采集高质量的数据？带着这些问题我们来学习项目一。

任务一 大数据认知

平台简介

一、大数据的概念

基于数据特征的视角，可将大数据定义为无法在一定时间内用传统数据库软件工具对其内容进行采集、存储、管理和分析的数据集合，该数据集合巨大以至于无法通过目前主流软件工具，在合理时间内达到撷取、管理、处理并整理成帮助企业经营决策的数据。

从描述数据的系统过程，可将大数据定义为那些需要新处理方法才能通过数据体现出更强的决策力、洞察力和流程优化能力的海量、高增长率和多样化的信息资产。

综合来看，大数据中的“大”不仅仅是指数据量的积累，其意义指向是要实现由量的积累到实现“大”的质的变化。大数据中的数据不是传统意义上的数据，这些数据因集合而产生意义价值，具有可观的利用前景。要基于这些大数据产生价值和效能，就必然要求这些数据之间存在意义和结构上的关联。大数据不是“死”数据，而是“活”数据，不是“假”数据，而是“真”数据，是必须予以应用并产生实际效用的数据。

二、大数据的特征

大数据区别于普通数据的四个特征为：数据量巨大（Volume）、数据种类多（Variety）、低密度高价值（Value）、实时处理（Velocity），也称为“4V”特征。

（一）数据量巨大（Volume）

大数据通常指 10 TB 规模以上的数据量。之所以产生如此巨大的数据量，一是由于使用各种仪器设备，使我们能够感知到更多的事物，这些事物的部分甚至全部数据可以

被存储；二是由于使用通信工具，使人们能够全时段地联系，机器—机器（Machine to Machine，简称为 M2M）方式的出现，使得交流的数据量成倍增长；三是由于集成电路价格降低，使很多物品有了智能的成分。

（二）数据种类多（Variety）

随着传感器种类的增多以及智能设备、社交网络等的流行，数据类型也变得更加复杂，不仅包括传统的关系数据类型，还包括以网页、视频、音频、E-Mail、文档等形式存在的未加工的、半结构化的和非结构化的数据。

（三）低密度高价值（Value）

大数据背后隐藏着极高的经济意义和经济价值，但是，大数据的价值深藏于浩瀚的数据当中，需要多来源数据的参照、关联、对比分析，需要独到的思维、高超的技术，挖掘大数据的价值就类似于沙里淘金。大数据的巨大价值来自其超前预测能力和真实性。

（四）实时处理（Velocity）

大数据具有数据增长速度快、处理速度快、时效性要求高的特点。Velocity 是大数据区别于传统数据的显著特征，大数据时代，快速从海量数据中挖掘出用户所需的信息需要强大的信息技术作支撑。例如，淘宝“双 11”促销时，销量、销售金额、订单量等实时信息展示，智慧搜索引擎能将几分钟前的新闻推送给用户，电子商务个性化推荐算法要求实时根据用户搜索或购买结果完成商品推荐等。

三、大数据发展简史

（一）20 世纪，大数据初步发展阶段

（1）最早提出“大数据”时代已经到来的机构是全球知名咨询公司麦肯锡。

（2）1980 年，著名未来学家阿尔文·托夫勒便在《第三次浪潮》一书中，将大数据热情地赞颂为“第三次浪潮的华彩乐章”。

（二）进入 21 世纪，大数据进入快速发展阶段

（1）2005 年，Hadoop 项目诞生，后因技术高效性，被 Apache Software Foundation 公司引入成为开元应用。

（2）2008 年末，“大数据”得到部分美国知名计算机科学研究人员认可，《自然》杂志专刊提出 Big Data 概念。

（3）2012 年，美国第一家大数据软件公司上市，联合国出台大数据白皮书，阿里巴巴全面推进“数据分享平台”战略，大数据价值得到进一步挖掘。

（4）2015 年，国务院正式印发《促进大数据发展行动纲要》，标志着大数据正式上升为我国国家战略。

（5）2016 年，大数据“十三五”规划出台，推动大数据在工业研发、制造、产业链全流程及服务业的发展。

（6）2017 年，工信部发布《大数据产业发展规划 2016—2020 年》，进一步明确了促进我国大数据产业发展的主要任务、重大工程和保障措施。十九大报告指出，加快建设制造强国，加快发展先进制造业，推动互联网、大数据、人工智能和实体经济深度融合。

（7）2020 年，大数据被正式列为新型生产要素。

（8）2021 年，国家《“十四五”发展规划》中指出，完善大数据标准体系建设。

四、大数据的类型

大数据不但数量巨大，而且数据类型较多。按照不同的分类标准，大数据可分为不同的类别。

（一）按照数据结构分类

按照数据结构可以将大数据划分为结构化数据、半结构化数据、非结构化数据三类。

1. 结构化数据

传统的数据大多是结构化数据，即使在大数据时代，结构化数据也是非常重要的数据类型之一。企业信息系统中的数据都是结构化数据。结构化数据具有统一的数据结构和规范的数据访问和处理方法。

如表 1–1–1 所示的数据是典型的结构化数据表现形式——关系数据模型。关系数据模型用二维表来表示数据，二维表由若干列组成，如表 1–1–1 所示的二维表由员工编码、部门、员工姓名、职位和基本工资等列组成，表中的行是二维表的数据，数据行由列的若干取值构成。

表 1–1–1　关系数据模型

员工编码	部门	员工姓名	职位	基本工资
000101	总裁办	吴弘易	总裁	150，000.00
000102	总裁办	张诚毅	副总裁	120，000.00
000103	总裁办	施新河	副总裁	120，000.00
000104	研发管理部	吴雅玲	研发总监	50，000.00
000105	研发管理部	陈晓东	质量总监	40，000.00
000106	研发管理部	许冬冬	研发助理	6，000.00
000107	研发管理部	林怡航	UI 界面设计师	18，000.00
000108	研发一部	蔡以周	产品经理	35，000.00
000109	研发一部	尹诗晴	需求分析师	25，000.00

2. 非结构化数据

与结构化数据相比，非结构化数据是指不能采用预先定义好的数据模型或者没有以一个预先定义的方式来组织的数据。常见的非结构化数据有声音、图像、视频等。非结

构化数据库是针对非结构化数据的存储和处理而产生的新型数据库，与传统关系数据库不同的是，它突破了数据固定长度的限制，支持采用重复字段、子字段和变长字段的应用，从而实现了对变长数据和重复字段进行存储和管理。

3. 半结构化数据

半结构化数据是介于结构化数据和非结构化数据之间的数据，互联网中的 XML 文件、HTML 文件就属于半结构化数据。半结构化数据一般是自描述的，数据的结构和内容混在一起，没有明显的区分。与结构化数据和非结构化数据相比，半结构化数据的格式更接近于结构化数据，但其结构变化又很大，因此，半结构化数据常需要采用非结构化数据的处理方式来管理数据。实际上，结构化、半结构化以及非结构化数据之间的不同，只不过是根据数据的格式划分的。

（二）按照产生主体分类

大数据按照产生主体可以划分为企业数据、机器数据和社会化数据三类。

1. 企业数据

企业数据主要指来自企业信息系统的数据，包括 ERP 系统中的数据、CRM 系统中的数据，以及其他企业业务系统中和企业运营等有关的数据。目前，企业数据仍然是应用最多的数据源之一。

2. 机器数据

机器数据是来自软硬件设备自动产生的数据，大多机器数据都是最原始的数据类型。机器数据包括日志文件、呼叫记录以及设备日志等。在大数据中，机器数据是增长比较快的一种数据，其所占的份额也比较大。在现代企业机构中，不管是什么规模都会产生巨大的机器数据，怎样管理机器数据、如何在万千数据中利用机器数据创造业务，是现代企业急需解决的一大问题。

3. 社会化数据

用户在媒体中分享自己的信息或评论他人的信息被称为社会化数据。社会化数据主要来自用户的行为记录、社交网络及反馈数据等。与静态数据相比，社会化数据更具备实时性和流动性的特点。

随着网络的流行，社交软件的大量使用，用户的登录和访问都会产生巨大的数据，如网络上的评论、视频、图片、个人信息资料等，这些数据都隐含了巨大的商用价值。

五、大数据算法

（一）大数据算法定义

在给定的资源约束下，以大数据为输入，在给定时间约束内可以生成满足给定约束结果的算法。其中的时间约束，不同研究和业务的要求不同。如科学研究可能允许几个月的计算时间，但搜索引擎和个性化推荐要求几分钟甚至几秒钟内计算出结果。

（二）大数据在挖掘领域的经典算法

1. C4.5 算法

C4.5 算法是机器学习算法中的一种分类决策树算法。它是决策树（决策树也就是做决策的节点间的组织方式像一棵树，其实是一个倒树）的核心算法。

2. K–Means 算法

这是一个聚类算法。它把 n 个对象根据它们的属性分为 k 个分割（$k < n$）。它与处理混合正态分布的最大期望算法很相似，因为他们都试图找到数据中自然聚类的中心。它假设对象属性来自空间向量，并且目标是使各个群组内部的均方误差总和最小。

3. Support Vector Machine（支持向量机），简称 SV 机（论文中一般简称 SVM）

它是一种监督式学习的方法，它广泛地应用于统计分类以及回归分析中。支持向量机将向量映射到一个更高维的空间里，在这个空间里建立有一个最大间隔超平面。在分开数据的超平面的两边建有两个互相平行的超平面，分隔超平面使两个平行超平面的距离最大化。假定平行超平面间的距离或差距越大，分类器的总误差越小。一个极好的指南是 C.J.C Burges 的《模式识别支持向量机指南》。van der Walt 和 Barnard 将支持向量机和其他分类器进行了比较。

4. Apriori 算法

这是一种最有影响的挖掘布尔关联规则频繁项集的算法。其核心是基于两阶段频集思想的递推算法。该关联规则在分类上属于单维、单层、布尔关联规则。在这里，所有支持度大于最小支持度的项集称为频繁项集，简称频集。

5. 最大期望（EM）算法

在统计计算中，最大期望（EM，Expectation–Maximization）算法是在概率（probabilistic）模型中寻找参数最大似然估计的算法，其中概率模型依赖于无法观测的隐藏变量（Latent Variable）。最大期望经常用在机器学习和计算机视觉的数据集聚（Data Clustering）领域。

6. PageRank

PageRank 是 Google 算法的重要内容。PageRank 根据网站的外部链接和内部链接的数量和质量，衡量网站的价值。PageRank 背后的概念是每个到页面的链接都是对该页面的一次投票，被链接得越多，就意味着被其他网站投票越多。这个就是所谓的“链接流行度”——衡量多少人愿意将他们的网站和你的网站挂钩。

7. Adaboost 算法

这是一种迭代算法，其核心思想是针对同一个训练集训练不同的分类器（弱分类器），然后把这些弱分类器集合起来，构成一个更强的最终分类器（强分类器）。其算法本身是通过改变数据分布来实现的，它根据每次训练集之中每个样本的分类是否正确，以及上次的总体分类的准确率，来确定每个样本的权值。将修改过权值的新数据集送给下层分类器进行训练，最后将每次训练得到的分类器融合起来，作为最后的决策分类器。

8. K 最近邻（KNN）分类算法

它是一个理论上比较成熟的方法，也是最简单的机器学习算法之一。该方法的思路是：如果一个样本在特征空间中的 k 个最相似（即特征空间中最邻近）的样本中的大多数属于某一个类别，则该样本也属于这个类别。

9. 朴素贝叶斯模型（Naive Bayesian Model，NBC）

朴素贝叶斯模型发源于古典数学理论，有着坚实的数学基础，以及稳定的分类效率。同时，NBC 模型所需估计的参数很少，对缺失数据不太敏感，算法也比较简单。理论上，NBC 模型与其他分类方法相比具有最小的误差率。但是实际上并非总是如此，这是因为 NBC 模型假设属性之间相互独立，这个假设在实际应用中往往是不成立的，这给 NBC 模型的正确分类带来了一定影响。在属性个数比较多或者属性之间相关性较大时，NBC 模型的分类效率比不上决策树模型。而在属性相关性较小时，NBC 模型的性能最为良好。

10. 分类与回归树（CART：Classification and Regression Trees）

在分类树下面有两个关键的思想：第一个思想是关于递归地划分自变量空间的想法；第二个思想是用验证数据进行剪枝。

任务二　大数据在财务领域中的应用

一、财务大数据的概念及特征

（一）财务大数据的概念

传统财务数据主要以财务报告数据为主，包括资产负债表、利润表、现金流量表、股东权益变动表以及报表附注等相关的财务数据。大数据给企业带来了更大的风险与挑战，大数据不仅扩大了企业财务数据的范畴，还对企业财务数据的处理、分析及反馈提出了更高的要求。财务大数据除了涵盖传统的财务报告数据之外，还包含宏观数据、行业数据，以及企业供应链等相关数据；同时，财务大数据的数据类型除了结构化数据之外，还包括非结构化数据和半结构化数据。

（二）财务大数据的特征

随着大数据时代的来临，企业财务管理不再仅仅局限于财务自身领域的一隅之地，而是渗透到企业的各个领域，如研发、生产、人力资源、销售等。可以说大数据时代的来临扩大了财务管理的影响力和作用范围，财务部门从原本的单纯的财务管理活动向数据的收集整理、处理分析方向转变。

具体而言，相比于传统财务数据，财务大数据的特征主要体现在以下四个方面。

1. 数据来源的广度与深度发生改变

大数据时代下，财务管理的管理范围被极大地扩大。除了原来的管理范围之外，大

数据下的财务管理还管理着很多非财务数据，包括销售信息、研发信息以及人力资源信息等。这是财务管理数据来源在广度上发生的变化。

财务管理数据来源在深度上发生的变化是财务管理数据由原来的结构化数据向非结构化数据、半结构化数据转变。结构化财务数据是由传统的运营系统产生的，这部分数据大多存储在关系型数据库中；非结构化和半结构化财务数据的来源较为广泛，比如，来自传感器的各种数据、移动电话的 GPS 定位数据、实时交易信息、行情数据信息、用户的网络单击量、顾客的搜索路径、浏览记录、购买记录等。在开展财务管理的过程中，非结构化和半结构化财务数据直接影响了财务数据的构成。

2. 数据处理由原来的集中式向分布式转变

大数据时代，不但企业数据量呈现出指数化增长趋势，而且企业数据分析处理的时效性要求也更高，传统的财务处理方式已不能满足大数据下的企业财务管理之需。大数据下的财务数据处理需要由原来的集中式计算结构，转为分布式或者扁平式的计算结构。

目前主流的分布式计算系统分别为 Hadoop、Storm 和 Spark 三种。Hadoop 可以轻松地集成结构化、半结构化甚至非结构化数据集。Storm 是分布式实时计算系统，它以全内存计算方式处理源源不断流进来的消息，处理之后再将结果写入到某个存储。而 Spark 则是基于内存计算的开源集群计算系统，能够更快速地进行数据分析。这三种计算架构在财务数据的处理方面各有优势，同时也有自身的劣势。在选择财务数据计算架构时，企业应根据自身具体情况进行判别。

3. 数据分析从数据仓库向深度学习进行转变

财务数据分析工作是企业在信息管理方面的重要内容。早期的会计电算化主要是面向操作型的，从会计凭证、账簿到报表都没有可靠的历史数据来源，自然也就不能将财务信息转换为可用的决策信息。随着信息处理技术的应用，企业可以利用新的技术实现财务数据的联机分享，还可利用统计运算方法和人工智能技术对数据仓库进行横向和纵向的分析，从而将大量的原始数据转化为对企业有用的信息，提高企业决策的科学性和可操作性。

例如，苏宁电器构建了 ERP 系统，其中在物流系统中将库存商品基础数据（包括产品编号、名称、规格型号，计划单价）、商家基本数据（包括商家编号、名称、地址、电话、邮编、银行账号等）与财务信息系统中的数据进行连接；资金流系统中的保理、保险、银行客户的基本数据、支付结算方式编码、货币编码、利率编码等与财务信息系统中的数据进行共享。这些措施在一定程度上使苏宁实现了财务数据共享和深度分析。

4. 数据输出形式由图表化转向可视化

在以前的财务数据输出工作中，企业大多采用图表的形式来报告企业财务信息，如财务报表等。在大数据背景下，企业改变了以往的信息输出形式，将复杂的财务数据转化为直观的图形。更进一步地，企业可以综合采用图形、表格和视频等方式将数据作可视化呈现，从而更好地将信息传达给企业内部及外部的信息使用者，为企业决策提供数

据支持。

例如，社交网络中的语音、图像、视频、日志文件等都是可视化的财务数据输出形式。1 号店、淘宝商城等电商就记录或搜集了网上交易量、顾客感知、品牌意识、产品购买和社会互动等行为数据，以可理解的图形、图片等方式直观呈现出企业在不同时间轴上财务数据的变化趋势。

二、大数据在财务领域中的典型应用场景

大数据场景应用本质上是数据的业务应用场景，是数据和数据分析在企业经营活动中的具体表现。财务大数据的典型应用场景包括财务分析、资金管理、全面预算、成本管控、投资决策等。

（一）财务分析

大数据时代，财务分析数据的来源除了内部财务账表以货币计量的结构化数据外，还有各类非结构化数据、业务数据等，并且可用的外部数据也越来越多。大数据时代的财务分析偏重于相关分析，即从某一相关事物的变化去分析另一相关事物是否发生变化，如没有变化或者变化不合常规，再分析其影响因素，以解释没有变化或者变化不合常规是否合理。比如，由于收入变化了，因此分析利润是否发生变化，如果利润没有变化或者变化不合常规，那么再分析成本、费用是否发生变化，并通过分析成本、费用变化是否合理来判断利润没有变化或变化不合常规是否合理。

（二）资金管理

资金管理是大型企业集团财务管理的核心内容，对企业战略发展和风险控制有重要的影响。大数据的出现也影响着资金管理的工作方式，原有的资金管理流程也会随之改变。

例如，一笔资金支付业务，原来的流程可能是业务部门提出资金需求，财务部门进行账务处理，然后流转到出纳。出纳制单后，再通过企业内部的审核流程，最终在银行付款。财务分析人员可能在周或月度结束后，从财务系统中取得数据，然后对本公司支付用途进行统计分析。而在大数据时代，业务部门和财务部门几乎能在同时进行处理。财务记账也不再需要拿到银行流水单再进行账务处理，而事后的统计分析工作也可以在支付的同时就得以进行。大数据简化了原来的流程，缩短了业务处理时间。

同时，大数据打破了原有的工作边界，资金管理不再只是关注资金的信息，而是要扩大范围，将企业内部各个职能部门都考虑在内，甚至包含上下游企业、竞争对手等，从而实现全流程、信息一体化的工作平台。

（三）全面预算

财务大数据环境下，全面预算依赖的数据类型不但包括传统预算中的财务数据，而且还包括音频、视频、地理位置、天气以及温度等非结构化数据，通过对这些数据的分析可以提升全面预算的准确性。

例如，在编制采购预算时，可以深入分析大数据中隐藏的信息，科学选择原材料供

应商；同时，还可以评价下级部门采购预算是否合理，以便更好地编制企业全面预算。与此同时，由于大数据使传统的自上而下传递预算任务的顺序发生改变，自下而上的预算审批顺序也因此发生变化，从而使得全面预算编制周期明显缩短。此外，在编制资金预算时，依托大数据分析，管理者能够判断预算资金是否合理，以防各部门虚报或瞒报预算资金。

（四）成本管理

成本管理是企业内部控制中最重要的部分，贯穿于企业经营的各个环节，成本管理有利于降低成本，提高经济效益。企业要获取更高的净利润，需要对生产成本和人力成本等多方面进行管控。传统成本管理更偏重于产品的生产成本管理和生产过程管理，相对忽视了其他诸如产品开发、采购、销售等过程的成本管理。

在大数据时代下，财务管理人员能够及时采集企业生产制造成本、流通销售成本等各种类型的数据，并将这些海量数据应用于企业成本控制系统，通过准确汇集、分配成本、分析企业成本费用的构成因素，区分不同产品的利润贡献程度并进行全方位的比较与选择，从而为企业进行有效的成本管理提供科学的决策依据。

（五）投资决策

财务大数据的应用给企业的投资决策提供了海量的可供决策的数据，从而支撑企业制定相对合理且科学的投资决策，提升企业投资决策效率和效果。

一方面，企业可建立专门的大数据收集平台，针对决策相关的数据进行收集、处理与提取，以提升数据获取的准确性、相关性与及时性；然后，构建大数据云计算平台，实时对大数据进行分析；接着，利用数据挖掘功能对信息与结果之间的相关性进行分析；最后，根据分析结果对较大概率能获得收益的项目进行投资。

另一方面，企业也可通过建立量化投资模型帮助决策者处理海量数据，使决策者能够在短时间内对影响投资结果的因素进行多角度的分析，包括经济周期、市场、未来预期、盈利能力、心理因素等，进而根据模型分析结果做出投资决策，大大提高投资效率。企业也可通过大数据建立数学模型对不同的风险因素进行组合分析，使其能在较短时间内迅速识别潜在的风险并进行精确的量化分析，进而实现对投资项目的风险控制。

三、大数据应用于财务场景的关键技术

从本质上看，大数据技术就是从类型各异或内容庞大的数据中能够快速有效地获取有价值的信息并加以分析。大数据应用于财务场景的关键技术主要有数据采集与预处理、数据存储、数据分析与挖掘以及数据的呈现与应用。

（一）大数据采集

根据数据源的不同，大数据采集的方法也不同。大数据采集方法有以下四大类。

1. 从数据库中采集

传统企业会使用 MySQL、Microsoft SQL Server 或 Oracle 等关系型数据库来存储数据。而随着大数据时代的到来，Redis、MongoDB 和 HBase 等 NoSQL 数据库也常用于

数据的存储。企业可在采集端部署大量数据库，以支持完成大数据的采集工作。

2. 从系统日志中采集

系统日志采集主要是收集公司业务平台日常产生的大量日志数据，供离线和在线的大数据分析系统使用。

3. 从网络数据中采集

网络数据采集是指通过网络爬虫或网站公开应用程序编程接口（Application Programming Interface，简写为 API）等方式从网站上获取数据信息的过程。这种方式可将网络中的非结构化数据、半结构化数据从网页中提取出来，存储在本地的存储系统中。

4. 感知设备数据采集

感知设备数据采集是指通过传感器、摄像头和其他智能终端自动采集信号、图片或录像来获取数据。

（二）数据预处理

数据预处理与集成就是对已经采集到的数据进行适当的处理或清洗去噪，之后再进一步集成存储。

数据预处理技术主要有数据清理、数据集成和数据变换。其中数据清理可以将一些噪声数据和异常的数据剔除，同时纠正可能存在的数据不一致情况。数据集成是将来自不同数据源的数据合并在一起，从而形成一致的数据存储。数据变换则是将数据转换成能支持数据分析模型的形式，以使数据分析结果更准确、更有意义。

（三）数据分析与挖掘

经过数据采集和预处理后，便可进入数据分析与挖掘的环节。数据分析与挖掘的目的是从一大批看似杂乱无章的数据中把有用的信息提炼出来，从而找出所研究对象的内在规律。在实际应用中，数据分析与挖掘可帮助人们做出判断，以便采取适当行动。

常见的数据分析与挖掘方法有聚类分析、时间序列分析、关联分析、回归分析、支持向量机、决策树等。

（四）大数据可视化

面对海量的数据，如何将其清晰明朗地展现给用户是大数据处理所面临的巨大挑战。虽然对于大数据处理来讲，数据分析与挖掘才是其主要的核心所在，但是数据使用者所关心的却通常是数据展示的结果。由于大数据在进行结果分析的时候会存在海量或关联关系极为复杂等特点，因此，如何通过图形化、图像化以及动画化等技术和方法展示大数据显得尤为重要。

可视化技术不仅能够迅速且有效地简化与提炼数据，还能让用户从复杂的数据中更快、更好地获取新的发现。在大数据时代，利用形象的图形向用户展示结果已经成了最理想的一种数据展示方式。

任务三 大数据分析方法论

一、大数据分析方法论的种类

大数据分析方法论主要包括两类，一类是统计分析方法论：描述统计、假设检验、相关分析、方差分析、回归分析、聚类分析、判别分析、主成分与因子分析、时间序列分析、决策树等；一类是营销管理常用分析方法论：SWOT、4P、PEST、SMART、5W2H、User behavior 等。

（一）统计分析方法论

1. 描述统计（Descriptive statistics）

描述统计是通过图表或数学方法，对数据资料进行整理、分析，并对数据的分布状态、数字特征和随机变量之间关系进行估计和描述的方法。目的是描述数据特征，找出数据的基本规律。描述统计分为集中趋势分析和离散趋势分析。

（1）集中趋势分析：数据的集中趋势分析是用来反映数据的一般水平，常用的指标有平均值、中位数和众数等。各指标的具体意义如下：

平均值：是衡量数据的中心位置的重要指标，反映了一些数据必然性的特点，包括算术平均值、加权算术平均值、调和平均值和几何平均值。

中位数：是另外一种反映数据的中心位置的指标，其确定方法是将所有数据以由小到大的顺序排列，位于中央的数据值就是中位数。

众数：是指在数据中发生频率最高的数据值。如果各个数据之间的差异程度较小，用平均值就有较好的代表性；而如果数据之间的差异程度较大，特别是有个别的极端值的情况，用中位数或众数有较好的代表性。

（2）离散趋势分析：离散趋势分析主要是用来反映数据之间的差异程度，常用的指标有平均差、方差和标准差。方差是标准差的平方根据不同的数据类型有不同的计算方法。

2. 假设检验

假设检验是数理统计学中根据一定假设条件由样本推断总体的一种方法。具体做法是：根据问题的需要对所研究的总体作某种假设，记作 H_0；选取合适的统计量，这个统计量的选取要使得在假设 H_0 成立时，其分布为已知；由实测的样本，计算出统计量的值，并根据预先给定的显著性水平进行检验，作出拒绝或接受假设 H_0 的判断。常用的假设检验方法有 $u-$ 检验法、$t-$ 检验法、x^2- 检验法（卡方检验）、$f-$ 检验法、秩和检验（rank sum test，又称顺序和检验）等。

3. 相关分析

相关分析是研究现象之间是否存在某种依存关系，并对具体有依存关系的现象探讨其相关方向以及相关程度，是研究随机变量之间的相关关系的一种统计方法。常见的

有线性相关分析、偏相关分析和距离分析。相关分析与回归分析在实际应用中有密切关系。然而在回归分析中，所关心的是一个随机变量 Y 对另一个（或一组）随机变量 X 的依赖关系的函数形式。而在相关分析中，所讨论的变量的地位一样，分析侧重于随机变量之间的种种相关特征。例如，以 X、Y 分别记小学生的数学与语文成绩，感兴趣的是二者的关系如何，而不在于由 X 去预测 Y。

4. 方差分析（Analysis of Variance，简称 ANOVA）

方差分析又称“变异数分析”或“F 检验”，是 R.A.Fisher 发明的，用于两个及两个以上样本均数差别的显著性检验。由于各种因素的影响，研究所得的数据呈现波动状。造成波动的原因可分成两类，一类是不可控的随机因素，另一类是研究中施加的对结果形成影响的可控因素。方差分析是从观测变量的方差入手，研究诸多控制变量中哪些变量是对观测变量有显著影响的变量。

5. 回归分析

回归主要的种类有线性回归、曲线回归、二元 logistic 回归、多元 logistic 回归。回归分析的应用是非常广泛的，统计软件包使各种回归方法计算十分方便。

一般来说，回归分析是通过规定因变量和自变量来确定变量之间的因果关系，建立回归模型并根据实测数据来求解模型的各个参数，然后评价回归模型是否能够很好地拟合实测数据；如果能够很好地拟合，则可以根据自变量作进一步预测。

6. 聚类分析

聚类主要解决的是“物以类聚、人以群分”：如以收入分群，高收入者 VS 低收入者；如按职场分群，职场精英 VS 职场小白；等等。

聚类的方法层出不穷，基于用户间彼此距离的长短来对用户进行聚类划分的方法依然是当前最流行的方法。大致的思路是这样的：首先确定选择哪些指标对用户进行聚类；然后在选择的指标上计算用户彼此间的距离，距离的计算公式很多，最常用的就是直线距离（把选择的指标当作维度，用户在每个指标下都有相应的取值，可以看作多维空间中的一个点，用户彼此间的距离就可理解为两者之间的直线距离）；最后聚类方法把彼此距离比较短的用户聚为一类，类与类之间的距离相对比较长。常用的算法 k－means、分层、FCM 等。

7. 判别分析

从已知的各种分类情况中总结规律（训练出判别函数），当新样品进入时，判断其与判别函数之间的相似程度（概率最大、距离最近、离差最小等判别准则）。常用判别方法有最大似然法、距离判别法、Fisher 判别法、Bayes 判别法、逐步判别法等。注意事项：①判别分析的基本条件：分组类型在两组以上，解释变量必须是可测的。②每个解释变量不能是其他解释变量的线性组合（如出现多重共线性情况时，判别权重会出现问题）。③各解释变量之间服从多元正态分布（不符合时，可使用 Logistic 回归替代），且各组解释变量的协方差矩阵相等（各组协方差矩阵有显著差异时，判别函数不相同）。相对而言，即使判别函数违反上述适用条件，也很稳健，对结果影响不大。

主要应用领域：对客户进行信用预测，寻找潜在客户（是否为消费者，公司是否成功，学生是否被录用，等等），临床上用于鉴别诊断。

8. 主成分与因子分析

主成分分析基本原理：利用降维（线性变换）的思想，在损失很少信息的前提下把多个指标转化为几个综合指标（主成分），即每个主成分都是原始变量的线性组合，且各个主成分之间互不相关，使得主成分比原始变量具有某些更优越的性能（主成分必须保留原始变量 90% 以上的信息），从而达到简化系统结构，抓住问题实质的目的。

因子分析基本原理：利用降维的思想，由研究原始变量相关矩阵内部的依赖关系出发，将变量表示成各因子的线性组合，从而把一些具有错综复杂关系的变量归结为少数几个综合因子（因子分析是主成分的推广，相对于主成分分析，更倾向于描述原始变量之间的相关关系）。

9. 时间序列分析

经典的统计分析都假定数据序列具有独立性，而时间序列分析则侧重研究数据序列的互相依赖关系。后者实际上是对离散指标的随机过程的统计分析，所以又可看作是随机过程统计的一个组成部分。例如，记录了某地区第一个月、第二个月……第 N 个月的降雨量，利用时间序列分析方法，可以对未来各月的降雨量进行预报。

10. 决策树（Decision Tree）

决策树是在已知各种情况发生概率的基础上，通过构成决策树来求取净现值的期望值大于等于零的概率，评价项目风险，判断其可行性的决策分析方法，是直观运用概率分析的一种图解法。由于这种决策分支画成图形很像一棵树的枝干，故称决策树。在机器学习中，决策树是一个预测模型，它代表的是对象属性与对象值之间的一种映射关系。

（二）营销管理常用分析方法论

（1）SWOT 分析。SWOT 即分析项目的优势、劣势、机会、威胁。

（2）4P 分析。4P 即分析产品、价格、促销、渠道。

（3）PEST 分析。PEST 即从政治、经济、社会、技术四个方面进行分析。

（4）SMART 分析。SMART 即分析项目的明确性、可量化、可实现、相关联、时效性。

（5）5W2H 分析。5W2H 分别是 WHY、WHAT、WHERE、WHEN、WHO、HOW、HOW MUCH。

① WHY——为什么？为什么要这么做？理由何在？原因是什么？

② WHAT——是什么？目的是什么？做什么工作？

③ WHERE——何处？在哪里做？从哪里入手？

④ WHEN——何时？什么时间完成？什么时机最适宜？

⑤ WHO——谁？由谁来承担？谁来完成？谁负责？

⑥ HOW ——怎么做？如何提高效率？如何实施？方法怎样？

⑦ HOW MUCH——多少？做到什么程度？数量如何？质量水平如何？费用产出如何？

（6）User behavior 分析。User behavior 即用户行为分析，从认知、熟悉、试用、实用、忠诚几方面进行分析。

二、大数据项目分析流程

大数据项目分析共有五个步骤，如图 1-3-1 所示。

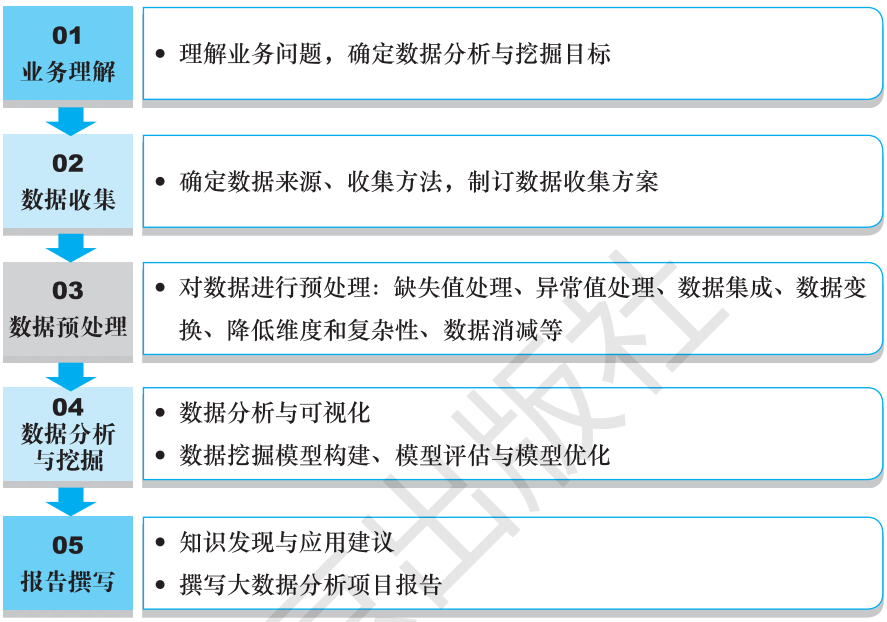

图 1-3-1　大数据项目分析流程

（一）业务理解

业务理解是根据业务背景信息，发现问题并清晰定义问题，确定数据分析与挖掘的目标。背景信息包括全面系统地掌握业务背景资料、熟悉相关行业知识和业务运作逻辑、分析业务背景中的具体问题进行归纳总结，并找出主要矛盾。问题分析包括根据已有的知识、经验分析问题、绩效目标的达成有哪些衡量指标、问题产生的可能性假设及验证。

（二）数据收集

数据收集是每个数据分析项目的第二个步骤。在数据分析的道路上，数据采集是重中之重。数据采集的质量直接决定了后续的分析是否准确。

1. 数据收集的含义

数据收集是指采用合适的方法和工具收集相应的数据，以供分析问题所用。

2. 大数据收集的数据源

从大数据的来源看，我们可以把大数据分为来自组织机构内的内部数据和来自组织机构外的外部数据。

（1）内部数据。内部数据是指来自企业自身日常经营管理中收集、整理的数据，主要有生产数据、库存数据、订单数据、电子商务数据、销售数据、客户关系管理数据，等等，随着企业自动化设备的大量启用，机器和传感器会产生越来越多的数据。内部数据具有较好的可控性，数据质量一般也有保证，但数据覆盖范围可能有限，需要借助其他资源渠道。

在内部数据中，财务数据是最主要的数据之一。财务数据是各类信息的综合集成，涉及人、财、物方方面面。财务人员作为数据的处理、计量、分析和报告者，理应在大数据分析中发挥着不可替代的关键作用。企业内部财务数据无非就是由资产负债表、利润表、现金流量表及所有者权益变动表共同构成的数据集合，是对企业经营状况、财务成果及资金运作的综合概括和高度反映，与财务人员后续的工作核算管理、成本费用管理、财务报表分析管理息息相关。

（2）外部数据。外部数据是指来源于企业外部，如互联网数据、其他供应商供应的付费数据、网络爬虫采集的数据等。对于互联网公开信息来说，互联网是数据的海洋，是获取各种数据的主要途径。例如，国家统计数据，各地方政府公开数据，上市公司的年报、季报，研究机构的调研报告，以及各种信息平台提供的零散数据等；对于供应商提供的付费数据来说，随着数据需求的加大，市场上催生了一些产品化数据交易平台，提供多领域的付费数据资源，可以按需购买使用；对于分析者自行通过网络采集软件，如通过爬虫软件，按照设定好的规则自动抓取互联网上的信息、程序，具有很好的内容收集作用。

典型的公开网络数据如：

① 在国家统计局“数据查询”栏目，可以查询到海量国际、国内统计数据。如图1–3–2所示。

图 1–3–2　国家统计局“数据查询”网页

② 国务院发展研究中心、中国证券监督管理委员会、上海证券交易所、深圳证券交

易所等网站等也提供有丰富的经济数据。

③ 网站分析类数据，如百度指数、Google 趋势、360 指数、腾讯云分析等。

④ 电商数据，如阿里价格指数、淘宝魔方、京东智圈、淘宝排行榜等。

⑤ 数据分析机构提供的数据，如艾瑞、埃森哲、德勤、国际数据等。

3. 数据收集的途径及方法

（1）网络爬取，即利用程序语言或数据采集器对特定网站、数据类型进行爬取，如 Python 爬取、后羿采集器爬取等。

（2）数据调用，包括企业信息管理系统数据调用、其他数据库数据调用、外部采购数据调用。

（3）网络搜索，包括社交网络数据搜索、专业网站如电商网站和证交所等数据搜索、政府部门和第三方数据收集采用。

（4）数据填报，包括常规报表和定制报表上报、企业经营管理活动中需要的各类填报数据收集。

（5）调查数据，基于调查、访谈等收集数据。

（三）数据预处理

数据预处理是指对收集的数据整理成能分析的相对统一的数据格式。数据预处理的方法包括数据清洗、数据集成、数据转换、数据归约（数据消减），其中最常见的方法为数据清洗，就是检测数据中存在的错误、重复和不一致，剔除或者改正它们以提高数据的质量。

（四）数据分析、挖掘与可视化

1. 数据分析

运用适当的统计分析方法对收集来的大量数据进行分析，提取有用信息形成结论而对数据加以详细研究和概括总结的过程。

2. 数据挖掘

数据挖掘（Data Mining）是指在大量数据中，提取隐含在其中却不被人所知，但又是潜在有用的信息和知识的过程。是一个用数据发现问题、解决问题的学科，通常通过对数据的探索、处理、分析或建模实现，以通过数据分析来识别趋势和模式，建立关系来解决业务问题。

3. 数据可视化

利用计算机图形学和图像处理技术，将数据转换为图形或图像在屏幕上显示出来，并利用数据分析和开发工具发现其中未知信息进行各种交互处理的理论、方法和技术。

数据可视化理论模型如图 1－3－3 所示。

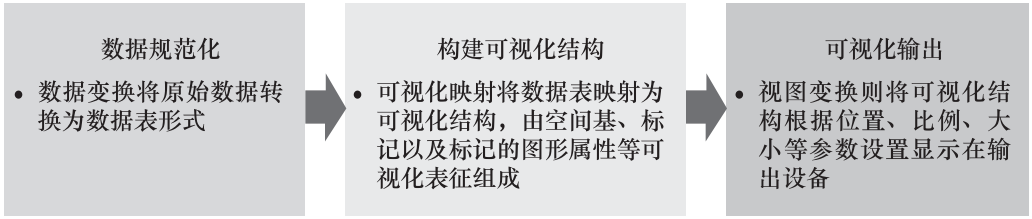

图 1–3–3　数据可视化理论模型

数据可视化的特征包括可视性、交互性、多维性。可视性是指数据可以用图像、曲线、二维图形、3D 和动画等显示，以视觉效果来加强用户对数据的感知能力；交互性是指允许用户选择感兴趣的内容，或者改变数据的展示形式，更好地促进用户和数据之间的互动；多维性是指对数据相关的多个变量或多个属性进行标识，可根据每一维的量值来进行显示、组合、排序与分类。

（五）报告撰写

大数据分析项目报告主要包括标题、目录、前言、正文、结论 5 部分内容。

1. 标题

标题是一份报告的文眼，它反映了全篇报告的主旨。

2. 目录

目录体现数据分析报告的逻辑关系、整体结构。

3. 前言

前言包括报告的目的和背景，阐述现状或者存在的问题，需要解决什么问题，运用了什么分析思路、分析方法和模型，给出总结性的结论或者效果，给出数据来源。

4. 正文

正文要求逻辑性强、层次结构清晰、分析结论明确等。需要进行可视化图形分析、挖掘分析等呈现正确解读的结论。

5. 分析结论

呈现数据：分析的总体结果，对结果进行解释与说明，并提出合理建议或改善策略。

课程思政专栏

贯彻新发展理念，推进高质量发展

【关键词】创新发展、高质量发展、伟大事业

【案例内容】党的二十大报告指出，新时代十年来，国内生产总值从五十四万亿元增长到一百一十四万亿元，我国经济总量占世界经济的比重达百分之十八点五，提高七点二个百分点，稳居世界第二位；人均国内生产总值从

三万九千八百元增加到八万一千元。谷物总产量稳居世界首位，十四亿多人的粮食安全、能源安全得到有效保障。城镇化率提高十一点六个百分点，达到百分之六十四点七。制造业规模、外汇储备稳居世界第一。建成世界最大的高速铁路网、高速公路网，机场港口、水利、能源、信息等基础设施建设取得重大成就。加快推进科技自立自强，全社会研发经费支出从一万亿元增加到二万八千亿元，居世界第二位，研发人员总量居世界首位。基础研究和原始创新不断加强，一些关键核心技术实现突破，战略性新兴产业发展壮大，载人航天、探月探火、深海深地探测、超级计算机、卫星导航、量子信息、核电技术、新能源技术、大飞机制造、生物医药等取得重大成果，进入创新型国家行列。

2023 年 2 月 23 日，国新办就“深入实施创新驱动发展战略，加快建设科技强国”举行新闻发布会，从发布会上获悉，我国全球创新指数排名从 2012 年的第 34 位升至 2022 年的第 11 位，开启了实现高水平科技自立自强、建设科技强国的新阶段。

科技赋能发展，创新决胜未来。党的十八大以来，在以习近平同志为核心的党中央坚强领导下，我国把科技自立自强作为国家发展的战略支撑，持续深入实施创新驱动发展战略，大力建设创新型国家和科技强国。十年风雨征程，十年跨越发展，折射出中国创新能力持续攀升的勃勃态势。

【启示】党的十八大以来，以习近平同志为核心的党中央把创新作为引领发展的第一动力，我国各项事业发生历史性、整体性、格局性变化，这个十年，创新驱动发展战略在神州大地落地生根，自立自强交出精彩答卷，成功迈入创新型国家行列。党中央提出并贯彻新发展理念，着力推进高质量发展，推动构建新发展格局，实施供给侧结构性改革，制定一系列具有全局性意义的区域重大战略，我国经济实力实现历史性跃升。

项目小结

本项目主要介绍了大数据的概念、特征、类型、发展简史及算法，财务大数据的概念、特征、典型应用场景及关键技术，大数据分析方法论的种类及大数据项目分析的步骤。详细论述了财务大数据的应用场景及大数据项目分析的步骤。