第 1 章

绪　论

【本章学习目标】

通过学习本章，学生应该能够掌握以下内容。

1. 掌握什么是大数据，对大数据有一个全面、清晰的认知。
2. 了解大数据的产生背景及应用场景。
3. 掌握大数据营销的内涵及其特点。

1.1　什么是大数据

1.1.1　大数据的定义

伴随着信息技术的日新月异，人类社会进入高速发展的时期。科技发达，信息流通，生活越来越便利，人与人之间的交流越来越密切，“大数据”就是信息时代的产物。

“信息”一词在古希腊文中表示“事物的本质”，其英文单词“information”源于拉丁文，语意表示“通知”，中国古代称之为“消息”。作为科学术语，“信息”最早出现在哈特莱于 1928 年撰写的论文《信息传输》中。该文中他首次提出了将信息量化处理的设想。1948 年，信息论创始人香农在他的著名论文《通信的数学理论》中，给出了计算信息熵的数学表达式，用以描述信息源各可能事件发生的不确定性。由此，信息被视为“不确定性的减少”。

在浩瀚的信息海洋中，人们对世界的认知和改造过程就是获取信息、加工信息和发送信息的过程。信息（information）是用文字、数字、符号、语言、图像等介质来表示事件、事物、现象等的内容、数量或特征，从而向人们（或系统）提供关于现实世界新的事实和知识，作为生产、建设、经营、管理、分析和决策的依据。数据（data）是事实或观察的结果，是对客观事物的逻辑归纳，是用于表示客观事物的未经加工的原始素材，它不仅是指狭义上的数字，还可以是具有一定意义的文字、字母、数字符号的组合、图形、图像、视频、音频等，也是客观事物的属性、数量、位置及其相互关系的抽象表示。例如，“0、1、2…”“阴、雨、下降、气温”“学生的档案记录、货物的运输情况”等都是数据。数据经过加工后就成为信息。

在计算机科学中，数据是所有能输入计算机并被计算机程序处理的符号的介质的总

称，是用于输入电子计算机进行处理，具有一定意义的数字、字母、符号或模拟量。计算机存储和处理的对象十分广泛，表示这些对象的数据也随之变得越来越复杂。人工智能、移动互联网、社交网络和物联网等新兴技术，正在通过新的数据形式和来源，增加数据的复杂性。随着大数据时代的来临，越来越多的政府、企业等组织机构开始意识到数据正在成为组织最重要的资产，数据分析能力正在成为组织的核心竞争力。

大数据（big data），又称巨量资料，指的是传统数据处理应用软件不足以处理的大或复杂的数据集。大数据是在信息技术高速发展、数据量爆炸性增长的背景下形成的，从它被提出的那一天起就受到了广泛的关注，不同的学者和机构从不同的角度对大数据进行了阐释。

高德纳咨询研究机构指出，数据增长的挑战和机遇包括数量、速度、多样性等三个维度，因此，“大数据是高容量、高速增长、多样化的信息资产，需要通过经济高效的新型信息处理方式去促成更强的决策能力、洞察力与流程优化的能力”。

麦肯锡全球研究院给出的“大数据”定义是：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、多样的数据类型和低的价值密度等特征。

我国政府于 2015 年颁布了《促进大数据发展行动纲要》（国发〔2015〕50 号），指出大数据是以容量大、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。

可见，大数据的外在表现形式是充斥着大量繁杂信息的数据集合，并且这些繁杂的数据难以被传统的技术手段所存储、解析、应用。而我们所要做的则是通过更先进的技术和手段对其进行核心价值的挖掘，洞察数据之间的关联逻辑，剔除无用的、冗余的信息，找出那些对国家治理、对企业决策、对组织和业务流程、对个人生活方式能够产生巨大影响的高价值信息或知识。因此，大数据的战略意义不在于掌握庞大的数据信息，而在于对数据的专业化分析与应用，从而释放出数据所蕴含的巨大价值。本书认为，大数据只有通过经济高效的新型处理方式才能具有更强的决策力、洞察力和海量、高增长的流程优化能力以及多样化的数据资源。

1.1.2 大数据的特征

2001 年，高德纳咨询公司分析员道格·莱尼用 3V 来描述数据增长的特征，即大量（volume）、多样（variety）和高速（velocity），以此说明日益庞大的电子商务的发展趋势。之后，麦肯锡公司提出大数据的 4V 特征：规模性（volume）、高速性（velocity）、多样性（variety）、价值性（value），即海量的数据规模、快速的数据流转、多样的数据类型和低成本的高价值创造。2013 年，IBM 公司又在 4V 的基础上提出了一个新特征，即真实性（veracity），体现了数据质量。由此，大数据的 5V 特征如图 1-1 所示。

（1）规模性。大数据最明显的特征就是规模大，随着信息技术的发展，数据开始爆发性增长。天文学和基因学是最早产生大数据变革的领域，2000 年，斯隆数字巡天计划启动

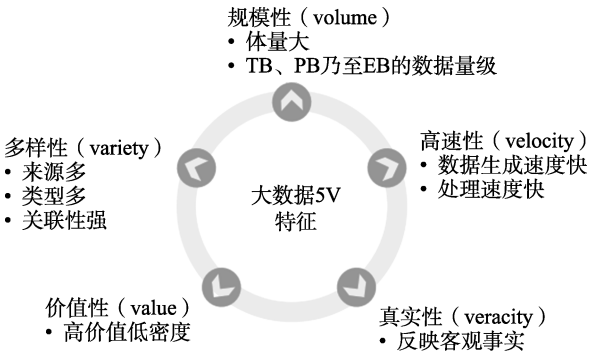

图 1-1　大数据 5V 特征

时，位于新墨西哥州阿帕奇山顶天文台的 2.5 m 口径望远镜，记录到近 200 万个天体的数据，包括 80 多万个星系和 10 多万个类星体的光谱数据，短短几周搜集到的数据比天文学历史上所有数据还要多；2003 年，人类第一次破译人体基因密码时，用了 10 年才完成了 30 亿对碱基对的排序，而在 10 年之后，世界范围内的基因仪 15 分钟就可以完成同样的工作量。马丁·希尔伯特和普里西拉·洛佩兹教授追踪计算了 60 余种模拟和以数字技术为载体的信息数量，于 2011 年在《科学》杂志撰文表明，从 20 世纪 80 年代开始，每隔 40 个月世界上存储的人均科技信息量就会翻倍。据国际数据公司（international data corporation，IDC）发布的白皮书《数据时代 2025》估计，目前全球数据信息总量为 44ZB（44 万亿 GB），并且根据进一步预测，到 2025 年，全球数据信息总量将达到 163ZB（163 亿 GB）。这里的 163ZB 数据是指在一年内全球所有被采集、创建和复制的新数据，其中涵盖了包括各行业的 50 余种内容创建或内容采集设备。数据单位的换算关系如表 1-1 所示。

表 1-1　数据单位换算关系

数据单位	换算关系
Byte	1Byte=8 bit
KB	1KB=1024 Byte
MB	1MB=1024 KB
GB	1GB=1024 MB
TB	1TB=1024 GB
PB	1PB=1024 TB
EB	1EB=1024 PB
ZB	1ZB=1024 EB

（2）高速性。大数据的高速性主要体现在两个方面，一是数据生成速度快。从数据生成速度来看，自工业革命以后，以文字为载体的信息量大约每十年翻一番；1970 年以后，微电子技术的发展使信息量大约每 3 年翻一番；1980 年以来，互联网等技术的应用使全球信息总量每两年就可以翻一番。进入 21 世纪以后，移动互联网、社交媒体、物联网、人工智能等技术的发展和应用，使数据呈现爆炸性增长。二是数据处理速度快。以阿里云为

例，在 2020 年淘宝“双十一”活动期间，11 月 11 日 0 点刚过 26 s，天猫双十一的订单创建峰值就达到 58.3 万笔/s，阿里云又一次扛住全球最大规模流量洪峰。阿里云在双十一期间系统“零死机”，由 Apache Flink 支持的阿里巴巴实时计算平台，在高峰时期每秒处理的数据流量总数为 40 亿条，与 2019 年的 25 亿条相比增势显著。阿里云专有的数据库平台 MaxComute，在 11 月 1 日至 11 日的 11 天购物节期间，每日处理高达 1.7EB 数据，规模相当于处理全球约 70 亿人口、每人 230 张高清照片。

（3）多样性。大数据的多样性首先体现在数据类型多，大体可以分为 3 类。一是结构化数据。通常来说，传统数据属于结构化数据，能够整齐地纳入关系数据库，如财务系统数据，医疗系统数据等。二是非结构化的数据，如文本、音频和视频等，其特点是数据间没有因果关系，它们需要经过额外的预处理操作才能真正提供洞察和支持性元数据。三是半结构化数据，如超文本标记语言（hyper text markup language，HTML）文档、电子邮件、网页、搜索索引、社交媒体论坛、主动和被动系统的传感器原始数据等，其特点是数据间的因果关系弱，数据来源多，不仅产生于组织内部运作的各个环节，而且也来自于组织外部。社交网络（微博、微信、抖音）、移动互联网、各种智能工具，服务工具等，都成为数据的来源。例如，淘宝网近 4 亿名的会员每天产生的商品交易数据约 20TB，脸书约 10 亿的用户每天产生的日志数据超过 300TB；交通领域，北京市交通智能化分析平台上来自路网摄像头/传感器、公交、轨道交通、出租车以及省际客运、旅游、化危运输、停车、租车等运输行业的数据，还有问卷调查和地理信息系统数据，4 万辆浮动车每天产生 2000 万条记录，交通卡刷卡记录每天 1900 万条，手机定位数据每天 1800 万条，出租车运营数据每天 100 万条，电子停车收费系统数据每天 50 万条，定期调查覆盖 8 万户家庭等。三是数据之间的关联性强。挖掘数据之间的关联性非常重要，探索这些形态各异、快慢不一的数据流之间的相关性，是大数据做前人之未做、能前人所不能的机会。大数据不仅是处理巨量数据的利器，而且为处理不同来源、不同格式的多元化数据提供了可能。例如，为了使计算机能够理解人的意图，人类就必须要将需解决的问题的思路、方法和手段通过计算机能够理解的形式告诉计算机，使得计算机能够根据人的指令一步一步工作，完成某种特定的任务。以往，人们只能通过编程这种规范化计算机语言发出指令，但随着自然语言处理技术的发展，人们可以用计算机处理自然语言，实现人与计算机之间基于文本和语音的有效通信。自然语言是一种更复杂、更多样的新型数据来源，它包含诸如省略、指代、更正、重复、强调、倒序等大量的语言现象，还包括噪声、含混不清、口头语和音变等语音现象。借助这项技术，手机可以识别用户的语音信息，并调用自带的各项应用如读短信、询问天气、设置闹钟、安排日程，甚至搜寻餐厅、电影院等生活信息，查看相关评论，根据用户的位置判断、过滤搜寻的结果直接订位、订票。

（4）价值性。大数据的核心特征是价值密度低，由于数据样本不全面、数据采集不及时、数据不连续等，有价值的数据所占的比例很小。随着移动互联网和物联网的广泛应用，信息感知无处不在，信息海量但价值密度较低，最常见的例子就是一天的监控视频，在 24 小时的记录中可能只有几秒钟的时间是有价值的。当然，数据量越大，数据价值密度越低是常见情况，但不是必然情况，如中国银联股份有限公司（以下简称银联）、维萨（VISA）

等清算组织有海量的交易数据，不仅数据量大，而且很有价值。如何通过强大的机器算法更迅速、更精准地完成数据的价值提升，是大数据时代亟待解决的难题。与传统的小数据相比，大数据最大的价值在于，可以从大量不相关的各种类型的数据中，挖掘出对未来趋势与模式预测分析有用的信息。通过对机器学习、人工智能或数据挖掘等方法的深度分析，得到新规律和新知识，并运用于交通、电商、医疗等各个领域，最终达到提高生产率、推进科学研究的效果。

（5）真实性。大数据的真实性指的是与传统的抽样调查相比，大数据反映的内容更加全面、真实，体现的是数据的质量。数据的重要性就在于对决策的支持，数据的规模并不能决定其能否为决策提供帮助，数据的真实性和质量才是获得真知和思路最重要的因素，是制定成功决策最坚实的基础。真实是对大数据的重要要求，也是大数据面临的巨大挑战。即使最优秀的数据清洗方法也无法消除某些数据固有的不可预测性。例如，人的情感和态度、未来的天气变化、宏观经济走势等。在处理这些类型的数据时，数据清洗无法修正这种不确定性，然而，尽管存在不确定性，数据仍然包含宝贵的信息。我们首先要认同、接受大数据的不确定性，并采用科学的方法加以修正或是处理。例如，采取数据融合，即通过综合多个可靠性较低的数据来源进而创建更准确、更真实的数据点，或者通过鲁棒优化技术以及模糊逻辑方法等数学工具最大限度地降低数据的不确定性。

1.1.3　大数据的分类

大数据分类在收集、处理和应用过程中非常重要，往往不同的环节需要理解和掌握不同的分类方式，以便更好地组织、管理、分析和应用大数据。下面介绍几种常见的分类方式。

1. 根据数据集的结构和建索引的难易程度，通常分为 3 类

结构化数据：这类数据最容易整理和搜索，主要包括财务数据、机器日志和人口统计明细等。结构化数据很好理解，类似于 Excel 电子表格中预定义的行列布局。这种结构下的数据很容易分门别类，数据库设计人员和管理人员只需要定义简单的算法就能实现搜索和分析数据。不过，即使结构化数据数量非常大，也不一定称得上“大数据”，因为结构化数据本身比较易于管理，不符合大数据的定义标准。一直以来，数据库都是使用结构化查询语言（structured query language，SQL）编程语言管理结构化数据。SQL 是由 IBM 在 20 世纪 70 年代开发的，旨在帮助开发人员构建和管理当时正逐步兴起的关系型（电子表格式）数据库。

非结构化数据：这类数据包括社交媒体内容、音频文件、图片和开放式客户评论等。这些数据符合大数据定义中大而复杂的要求，也因此这些数据通常很难用标准的行列关系型数据库捕获。如何利用这类大数据是企业不断探索的问题，大多数情况下企业若想搜索、管理或分析大量非结构化数据，只能依靠烦琐的手动流程。毫无疑问，分析和理解这类数据能够为企业带来价值，但是执行成本往往太过高昂。而且，由于耗时太长，分析结果往往还未交付就已经过时。因为无法存储在电子表格或关系型数据库中，所以非结构化数据通常存储在数据湖、数据仓库和非关系型数据库（not only SQL，NoSQL）中。

半结构化数据：顾名思义，是结构化数据和非结构化数据的混合体。电子邮件就是一个很好的例子，因为其中的正文属于非结构化数据，而发件人、收件人、主题和日期等则属于结构化数据。使用地理标记、时间戳或语义标记的设备也可以同时提供结构化数据和非结构化数据。例如，一张未做标识的智能手机图片仍然可以告诉你，这是一张自拍照，以及拍摄的时间和地点。采用人工智能技术的现代数据库不仅能够即时识别不同类型的数据，还能够实时生成算法，有效地管理和分析各种相关的数据集。

这种分类方式近几年特别重要，相关的场景包括：其一，结构化数据是传统数据的主体，而半结构化和非结构化数据是大数据的主体。后者的增长速度比前者快很多，大数据的量这么大，主要是因为半结构化和非结构化数据的增长速度太快。其二，在数据平台设计时，结构化数据用传统的关系数据库便可高效处理，而半结构化和非结构化数据必须用海杜普（Hadoop）等大数据平台。其三，在数据分析和挖掘时，不少工具都要求输入结构化数据，因此半结构化、非结构化数据需要经过清洗、整理、筛选，转换为结构化数据。

2. 从字段类型上分为文本类（string、char、text 等）、数值类（int、float、number 等）、时间类（data、timestamp 等）

文本类数据常用于描述性字段，如姓名、地址、交易摘要等。这类数据不是量化值，不能直接用于四则运算。在使用时，可先对该字段进行标准化处理（比如地址标准化）再进行字符匹配，也可直接模糊匹配。

数值类数据用于描述量化属性，或用于编码，如交易金额、额度、商品数量、积分数、客户评分等都属于量化属性，可直接用于四则运算，是日常计算指标的核心字段。邮编、身份证号码、卡号之类的则属于编码，是对多个枚举值进行有规则编码，可进行四则运算，但无实质业务含义，不少编码都作为维度存在。

时间类数据仅用于描述事件发生的时间，时间是一个非常重要的维度，在业务统计或分析中非常重要。

这种分类方式是最基本的，和很多场景有关。其一，在系统设计时，需要确定每个字段的类型，以便设计数据库结构。其二，在数据清洗时，文本类数据往往很难清洗，而且很多文本类数据也没有清洗的必要，如备注或客户评论。数值类和时间类数据是清洗的重点，这类字段在业务上一般都有明确的取值范围，如年龄必须大于 0。对于不合法的取值，通常用默认值填充。其三，在建立维度模型时，数值类中的编码型字段和时间类字段通常作为维度，数值类中的量化属性作为度量。

3. 从描述事物的角度分为：状态类数据、事件类数据、混合类数据

用数据来描述客观世界，一般可以从两个方面出发。

第一方面是描述客观世界的实体，也即一个个对象，如人、桌子、账户等。这些对象，各有各的特征，不同种类的对象拥有不同的特征。比如，人的特征包括姓名、性别和年龄，桌子的特征包括颜色和材质。对于同一种对象的不同个体，其特征值的不同，如张三男 20 岁、李四女 24 岁。有些特征稳定不变，而另一些则会不断发生变化，如性别一般不变，但账户金额、人的位置则随时可能变化。因此，可以使用一组特征数据来描述每个对象，

这些数据可以随时间发生变化（数据的变化一方面依赖于对象的变化，另一方面依赖于反映到数据上的时间差变化），每个时点的数据反映这个时点对象所处的状态，因此称之为状态类数据。

第二方面是描述客观世界中对象之间的关系，它们是如何互动的，如何发生反应的。我们把这一次次互动或反应记录下来，这类数据称之为事件类数据。比如，客户到商店买了件衣服，这里出现 3 个对象，分别是客户、商店、衣服，3 个对象之间发生了一次交易关系。

混合类数据理论上也属于事件类数据范畴，两者的差别在于，混合类数据所描述的事件发生过程持续较长，记录数据时该事件还没有结束，还将发生变化。比如，订单，从订单生成到结束整个过程需要持续一段时间，首次记录订单数据是在订单生产的时候，订单状态、订单金额后续还可能多次变化。

这种分类方式在数据仓库建模时特别重要。数据仓库需要保存各种历史数据，不同类型的历史数据保存方式差别很大。状态类数据保存历史的方式一般有两种：存储快照或者缓慢变化维度方式。事件类数据一旦发生就已经是历史了，只需直接存储或者按时间分区存储。混合类数据保存历史比较复杂，可以把变化的字段分离出来，按状态类数据保存，剩下不变的则按事件类数据保存，使用时再把两者合并。另一个相关场景就是客户画像，客户画像通常用状态类数据，对于和客户相关的事件类数据和混合类数据，也会转换成和状态类数据相同的形态。

4. 从数据处理的角度分为原始数据、衍生数据

原始数据是指来自上游系统的，没有做过任何加工的数据。虽然会从原始数据中产生大量衍生数据，但还是会保留一份未作任何修改的原始数据，一旦衍生数据发生问题，可以随时从原始数据重新计算。

衍生数据是指通过对原始数据进行加工处理后产生的数据。衍生数据包括各种数据集市、汇总层、宽表、数据分析和挖掘结果等。从衍生目的上来说，可以简单分为两种情况，一种是为提高数据交付效率，数据集市、汇总层、宽表都属于这种情况。另一种是为解决业务问题，数据分析和挖掘结果就属于这种。

这种分类方式主要用在管理数据上，对原始数据的管理和衍生数据的管理有一些差别。原始数据通常只要保留一份，衍生数据却不同，管理形式比较灵活。只要有利于提高数据分析和挖掘效率，产生更大的数据价值，任何形式都可以尝试。比如，为每个业务条线定制个性化数据集市，提高每个业务条线的数据分析效率，虽然不同集市存在大量冗余的数据，但只要能大幅提高分析效率，用空间换时间也未尝不可。

5. 从数据粒度上分为明细数据、汇总数据

通常从业务系统获取的原始数据，是粒度比较小的，包括大量业务细节。比如，客户表中包含每个客户的性别、年龄、姓名等数据，交易表中包含每笔交易的时间、地点、金额等数据。这种数据我们称之为明细数据。明细数据虽然包括了最为丰富的业务细节，但在分析和挖掘时，往往需要进行大量的计算，效率比较低。

为了提高数据分析效率，需要对数据进行预加工，通常按时间维度、地区维度、产品维度等常用维度进行汇总。分析数据时，优先使用汇总数据，如果汇总数据满足不了需求则使用明细数据，以此提高数据使用效率。

这种分类方式的相关场景有两种。一种是在设计数据仓库时，如何对数据进行汇总，按什么方式进行汇总，才能达到使用效率和汇总成本的平衡。另一种是数据分析人员在分析数据时，在明细数据、各种汇总数据之间选择合适的数据，以提高分析效率。

6. 从更新方式上分为批量数据、实时数据

源系统提供数据时，不同的源系统有不同的提供方式，主要可以分为两种方式。 一种是批量方式，这种方式每隔一段时间提供一次，把该时段内所有变化的都提供过来。批量方式时效较低，大部分传统系统都采用 $T+1$ 方式，业务用户最快只能分析到前一天的数据，看前一天的报表。

另一种方式是实时方式，即每当数据发生变化或产生新数据，就会立刻提供过来。这种方式时效快，能有效满足时效要求高的业务，如场景营销。但该方式对技术要求更高，必须保证系统足够稳定，一旦出现数据错误，容易造成较严重的业务影响。

这种分类方式也非常重要，目前有越来越多系统采取该方式提供数据。这对数据处理、数据分析和数据应用产生了巨大的影响。一方面能为业务提供近乎实时的数据和报表支持，实现高时效的业务场景；另一方面也极大地增加了数据架构、数据分析和应用的技术难度。

7. 从数据来源上分为交易数据、社交数据、机器数据

交易数据是世界上发展速度和增长速度最快的数据。例如，一家大型国际零售商每小时处理超过 100 万笔客户交易，还有我们在前面提到的淘宝“双十一”活动所产生的交易数据。大数据平台能够获取时间跨度更大、更海量的结构化交易数据，这样就可以对更广泛的交易数据类型进行分析，不仅仅包括销售时点系统（point of sale，POS）或电子商务购物数据，还包括行为交易数据。例如，Web 服务器记录的互联网点击流数据日志。此外，交易数据越来越多地由半结构化数据组成，包括图片和注释等，使得管理和处理难度不断增加。

社交数据来源于社交媒体评论、发帖、图片以及与日俱增的视频文件，这些广泛存在的非结构数据流为使用文本分析功能进行分析提供了丰富的数据源泉。随着 5G 蜂窝网络的普及，2023 年全球手机视频用户达到 27.2 亿名。虽然社交媒体及其使用趋势瞬息万变、难以预测，但作为数字数据的主要来源，其稳定增长趋势是不会改变的。

机器数据包括功能设备创建或生成的数据，如智能电表、智能温度控制器、工厂机器和连接互联网的家用电器。这些设备可以配置为与互联网络中的其他节点通信，还可以自动向中央服务器传输数据，这样就可以对数据进行分析。物联网设备和机器都配有传感器，能够发送和接收数字数据。物联网传感器能够帮助企业采集和处理来自整个企业的设备、工具和装置的机器数据。从天气和交通传感器到安全监控，全球范围内的数据生成设备正在迅速增多。据互联网数据中心（internet data center，IDC）估计，到 2025 年，全球物联网设备数量将超过 400 亿台，生成的数据量几乎占全球数字数据总量的一半。

1.2　大数据发展历程

1.2.1　大数据产生背景

1. 信息时代的来临

20 世纪中叶，人类迈入了信息时代。随着计算机的发明和应用，信息的载体从“语言”“文字”，进而发展到“数据”。“数据”是客观事物的符号化表示，二进制的发明实现了数据在物理机器中的表达、计算和传输。数据可输入计算机，被计算机程序理解和处理。这样，利用计算机强大的计算能力，人类对数据得以有效的管理与开发利用。

经过近半个世纪的探索，人类以计算机为工具管理数据从依赖“特有程序”到“文件系统”管理，再到“数据库管理系统”，基本实现了数据的快速组织、存储和读取。以计算机为工具管理数据不仅提升了信息描述的精确性，而且扩大了信息传递的广泛性，信息在越来越广阔的空间发挥着越来越重要的作用。

1999 年，“物联网”（internet of things，IoT）的概念被提出。物联网是新一代信息技术的重要组成部分，它是基于互联网、传统电信网等的信息承载体，让所有能够被独立寻址的普通物理对象形成互联互通的网络，是“万物互联”的网络。通过各种信息传感器、射频识别技术、全球定位系统、红外感应器、激光扫描器等各种装置与技术，实时采集任何需要监控、连接、互动的物体或过程，采集其声、光、热、电、力学、化学、生物、位置等各种需要的信息，通过各类可能的网络接入，实现物与物、物与人的泛在连接，实现对物品和过程的智能化感知、识别和管理。智能设备的普及、物联网的广泛应用、存储设备性能的提高，为大数据的产生提供了储存和流通的物质基础。

传统数据基本来源于行业或企业的内部数据，现在则大部分来源于互联网和物联网。传统数据以结构化数据为主，而现在来源于社交网站、电子商务、物联网的数据基本都是非结构化和半结构化的数据。传统数据用关系数据库的管理系统可实现有效的管理与开发，现在数据因其大量、迅速、复杂，大大超出了传统数据库软件工具的能力范围，以至于引发了数据存储与处理的危机。

2. 云计算的兴起

随着互联网规模不断扩大，网络上汇聚的计算资源、存储资源、数据资源和应用资源不断增加，互联网正在从传统意义的通信平台转化为泛在、智能的计算平台。与计算机系统这样的传统计算平台相比，互联网上还没有形成类似计算机操作系统的服务环境，以支持互联网资源的有效管理和综合利用。在传统计算机中已成熟的操作系统技术，已不再能适用于互联网环境，其根本原因在于：互联网资源的自主控制、自治对等、异构多尺度等基本特性，与传统计算机系统的资源特性存在本质上的不同。为了适应互联网资源的基本特性，形成承接互联网资源和互联网应用的一体化服务环境，业界展开了面向互联网计算的虚拟计算环境（internet-based virtual computing environment，iVCE）的研究工作，使用户能够方便、有效地共享和利用开放网络上的资源。

2006 年 8 月，云计算（cloud computing）这个概念首次在搜索引擎会议上提出，成为互联网的第三次革命。狭义上讲，云计算就是一种提供资源的网络，使用者可以随时获取“云”上的资源，按需求量使用，并且可以看成无限扩展的，只要按使用量付费就可以；从广义上说，云计算是与信息技术、软件、互联网相关的一种服务，这种计算资源共享池叫作“云”，云计算把许多计算资源集合起来，通过软件实现自动化管理，只需要很少的人参与，就能让资源被快速提供，如图 1-2 所示。云计算是分布式计算的一种，通过网络“云”将巨大的数据计算处理程序分解成无数个小程序，然后，通过多部服务器组成的系统对这些小程序进行处理和分析，得到结果并返回给用户。

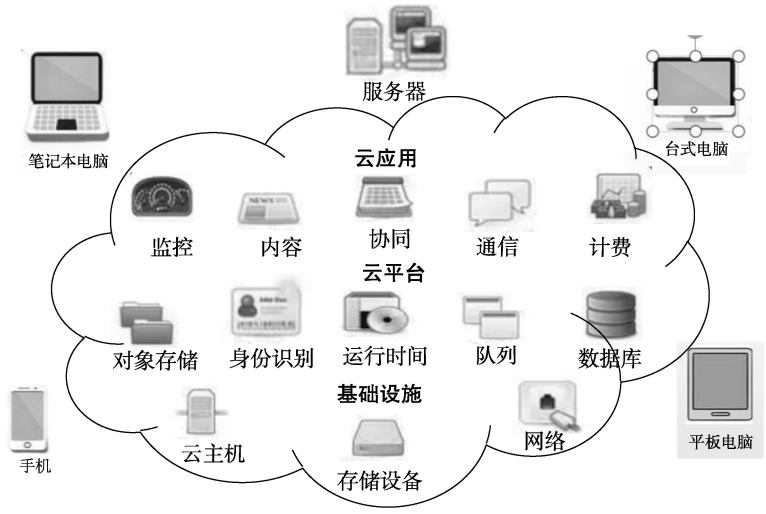

图 1-2　云计算

早期的云计算，就是简单的分布式计算，解决任务分发，并进行计算结果的合并。现阶段所说的云服务，则是分布式计算、效用计算、负载均衡、并行计算、网络存储、热备份冗余和虚拟化等计算机技术混合演进并跃升的结果。云计算也正在成为信息技术产业发展的战略重点，全球的信息技术企业都在纷纷向云计算转型。

云计算是大数据的基础，大数据是云计算的应用，二者密不可分。从应用视角上来看，云计算以新型的资源管理模型，为终端用户提供了组织、共享和管理资源的方式和机制，以支持互联网大数据资源的有效共享和综合利用。从开发视角上来看，云计算是互联网新型应用的软件开发平台，提供了与大数据资源管理模型一致的程序设计模式与运行支撑模式，能方便、快捷地帮助开发人员构造面向互联网的应用系统。从系统视角上来看，云计算包括了支持资源管理模型的程序设计语言，网络延迟探测、支持网络资源按需聚合和协同的虚拟节点、资源聚合管理、资源协同管理、虚拟网络内存、虚拟网络外存和虚拟执行网络等基础服务，以及云计算应用管理与运行支撑环境。

3. 数据资源化的趋势

数据作为数字经济全新的、关键的生产要素，贯穿于数字经济发展的全部流程，与其

他生产要素不断组合迭代，加速交叉融合，引发生产要素多领域、多维度、系统性、革命性的群体突破。数据资源是能够参与社会生产经营活动、可以为使用者或所有者带来经济效益、以电子方式记录的数据。区别数据与数据资源的依据，主要在于数据是否具有使用价值。

数据资源化使无序、混乱的原始数据成为有序、有使用价值的数据资源。数据资源化阶段就是通过对数据的采集、整理、聚合、分析等，形成可采、可见、标准、互通、可信的高质量数据资源。数据资源化是激发数据价值的基础，其本质是提升数据质量、形成数据使用价值的过程。

习近平总书记指出，要“发挥数据的基础资源作用和创新引擎作用”，党的十九届四中全会首次明确数据可作为生产要素按贡献参与分配，《关于新时代加快完善社会主义市场经济体制的意见》首次将数据与技术、人才、土地、资本等要素一起纳入改革范畴，《关于构建更加完善的要素市场化配置体制机制的意见》、十九届五中全会等历次重要会议、文件都将数据要素作为重要内容，为加快数据要素市场发展提供了根本遵循，为数据要素市场发展确定了目标、指明了方向。历史经验表明，每一次经济形态的重大变革，必然催生也必须依赖新的生产要素。如同农业经济时代以劳动和土地、工业经济时代以资本和技术为新的生产要素一样，数字经济时代，数据成为新的关键生产要素。由网络所承载的数据、由数据所萃取的信息、由信息所升华的知识，正在成为企业经营决策的新驱动、商品服务贸易的新内容、社会全面治理的新手段，带来了新的价值增值。

1.2.2　大数据发展的三个阶段

有关大型数据集的起源，最早可追溯至 20 世纪 60 至 70 年代。当时数据世界正处于萌芽阶段，全球第一批数据中心和首个关系数据库便是在那个时代出现的。1980 年，未来学家阿尔文托夫勒在《第三次浪潮》一书中称“大数据是第三次浪潮中最华彩的乐章”。2005 年左右，人们开始意识到用户在使用脸书（Facebook）等社交媒体以及其他在线服务时生成了海量数据。同一年，专为存储和分析大型数据集而开发的开源框架 Hadoop 问世，NoSQL 也在同一时期开始渐渐普及开来。2008 年 8 月，维克托·迈尔·舍恩伯格和肯尼斯·库克耶在《大数据时代》中指出，大数据是对所有数据进行整体分析处理，而不是采用随机分析法，即抽样调查进行分析。2008 年 9 月，《自然》杂志推出了“大数据”封面专栏。

Hadoop 及后来 Spark 等开源框架的问世对于大数据的发展具有重要意义，正是它们降低了数据存储成本，让大数据更易于使用。在随后几年里，大数据数量进一步呈爆炸式增长。时至今日，全世界的“用户”——不仅有人，还有机器——仍在持续生成海量数据。如今，随着物联网的兴起，越来越多的设备接入了互联网，收集了大量的客户使用模式和产品性能数据。同时，机器学习的出现也进一步加速了数据规模的增长。

大数据技术的发展可以按照其特点，分为大数据 1.0、大数据 2.0、大数据 3.0 三个阶段，各阶段的需求驱动和关键技术如图 1-3 所示。

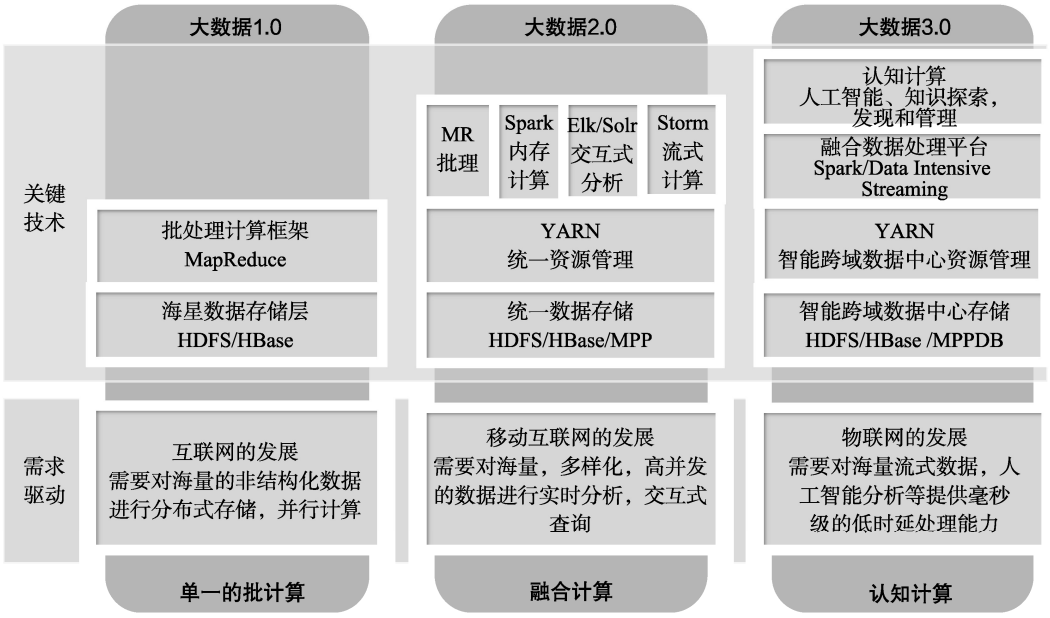

图 1-3　大数据技术发展阶段

大数据 1.0 阶段是在不同分析业务场景下，采用不同技术如基于 Share-Nothing 无共享模式并行计算的大规模并行处理（massively parallel processing，MPP）技术和基于 Hadoop 体系的大数据技术，分别封装成不同产品，完成 TP 级结构化数据和 PB 级非结构化数据的汇聚和分析。此时的大数据平台是以处理海量数据存储、计算及流数据实时计算等场景为主的传统意义上的“大数据平台”基础设施，以 Hadoop、Spark、Hive 等作为大数据基础能力层，在大数据组件上搭建包括数据分析，机器学习程序等抽取、转换、装载方法（extract transformation load method，ETL method）流水线，以及包括数据治理系统、数据仓库系统、数据可视化系统等核心功能。这一时期，硬件投资与软件开发投入量巨大，极大增加了研发的难度、调试部署的周期、运维的复杂度，且架构的缺陷，数据应用开发运行维护的困难，多用户资源隔离的复杂度等原因经常造成“数据孤岛”“应用孤岛”的问题。

大数据 2.0 阶段进入云时代，基于云基础设施（infrastructure），统一部署，但业务依然是分群的，给用户搭建的湖仓平台也依然是分离的架构。数据同步、计算流程以离线为主，固化的数据处理加工，引入实时计算引擎对其中的指标进行实时计算。上游将数据生产到 Spark/Storm 作为消费者进行数据消费，实时计算处理后，将数据写入查询效率较高的存储中，如 Hbase 或 MPP 数据库中，通过新建数据链路，来满足数据的实时计算需求。然而数据链路维护两份，开发成本较高，同时，两个链路的计算结果需要做指标对账，存在业务逻辑、计算逻辑不一致的风险。

大数据 3.0 阶段基于云原生（cloud native）技术，核心要素是容器化轻量化部署、存储和计算的分离、极致和可靠的按需弹性扩展,提供智能跨域数据平台解决方案。在经历了前两阶段的认知与技术发展的铺垫后，大数据迅速渗透各行各业。数据驱动决策，信息社会智能化程度大幅提高，同时出现跨行业、跨领域的数据整合，甚至是全社会的数据整合，

并且可以从各种各样的数据中找到对社会治理、产业发展更有价值的应用。云原生的代表技术包括不可变基础设施、微服务、声明式应用程序编程接口（application programming interface，API）、容器和服务网格，这些技术能够构建容错性好、易于管理和便于观察的松耦合系统。在云原生技术的背景下，各组织在公有云、私有云和混合云等新型动态环境中，构建和运行可弹性扩展的应用，云上资源利用率和应用交付效率得到显著提升。面对业态各异的业务上云以及碎片化的物联网解决方案部署，利用云原生思维和模式，构建基于云原生的物联网平台以及解决方案，势必将加速企业甚至整个社会的数字化转型。

尽管已经出现了相当长的一段时间，但人们对大数据的利用才刚刚开始。今天，云计算进一步释放了大数据的潜力，通过提供真正的弹性与可扩展性，让开发人员能够轻松启动点对点（Ad-Hoc）集群来测试数据子集。此外，图形数据库在大数据领域也变得越来越重要，它们能够以独特的形式展示大量数据，帮助用户更快速执行更为全面的分析。

1.3　大数据技术的应用场景

大数据技术已经在互联网、商业智能、医疗服务、零售业、金融业、电信等领域得到广泛的应用，其典型场景包括气象预测、射频识别（radio frequency identification，RFID）、遥感遥测、天文观测、交通运输、基因组学、生物学、大社会数据分析、国际网络文件处理、搜索引擎索引、军事侦察、金融大数据、医疗大数据、社群网络、医疗记录、照片图像和影像封存、大规模的电子商务等。伴随着各种穿戴设备、物联网和云计算、云存储等技术的发展，数据内容和格式的多样化，数据颗粒度也越来越细，随之出现了分布式存储、分布式计算、流处理等大数据技术，各行业基于多种甚至跨行业的数据源相互关联以探索更多的应用场景，同时更注重面向个体的决策和应用的时效性。

1.3.1　大数据技术在各行业的应用

1. 互联网行业

互联网行业是大数据应用的发源地，其主流应用包括：搜索引擎，电子商务，社交媒体。

（1）搜索引擎。搜索引擎是天然的大数据服务，它连接了人与信息、人与服务，其目的就是更好地理解用户的搜索需求，将信息与用户匹配起来。与此同时，大数据技术也在推动着搜索引擎不断向前演进，呈现了以下 3 个突破。一是智能交互。用户需求更趋于复杂化和个性化，使用语音和图像来表达需求的比例更高，为了不断提升用户体验，图像识别和语音识别技术被开发应用于该领域。二是知识图谱。用户更希望找到答案、加深了解以及发现更多的内容。知识图谱是基于海量的互联网数据，实现这种演变的最为重要的技术之一。三是深度问答。深度问答是一种基于海量互联网数据和深度语义理解的智能系统，它通过海量数据的深层分析和语义理解，并通过搜索和语义匹配技术，提炼出答案信息，对信息进行聚合、提炼，给出最全面、准确的结果。

（2）电子商务。电子商务的大数据应用有以下 3 个方面：精准营销、个性化服务、商

品个性化推荐。精准营销是互联网企业使用大数据技术采集有关客户的各类数据，并通过大数据分析建立“用户画像”来抽象地描述一个用户的信息全貌，从而可以对用户进行个性化推荐、精准营销和广告投放等。电子商务具有提供个性化服务的先天优势，可以通过技术支持实时获得用户的在线记录，并及时为他们提供定制化服务。例如，一个用户想要在天猫上购买一个电视，他可以使用定制功能，在电视机生产以前选择尺寸、边框、清晰度、能耗、颜色、接口等属性，再由厂商组织生产并送货到顾客家中。这样的个性化服务受到了广泛欢迎。商品个性化推荐的功能依托于个性化推荐系统，系统通过分析用户的行为，包括反馈意见、购买记录和社交数据等，以分析和挖掘顾客与商品之间的相关性，从而发现用户的个性化需求、兴趣等，然后将用户感兴趣的信息、产品推荐给用户。个性化推荐系统针对用户特点及兴趣爱好进行商品推荐，能有效地提高电子商务系统的服务能力，从而保留客户。此外，应用大数据分析技术还可以帮助企业预测消费趋势，如区域消费特征，顾客消费习惯，消费者行为，消费热点和影响消费的重要因素等，并根据消费者的习惯提前生产物料和进行物流管理，实现精益生产。

（3）社交媒体。大数据产生的背景离不开脸书、微博、微信、抖音等社交媒体的兴起，人们每天通过这些自媒体传播信息或者沟通交流，由此产生的信息被网络记录下来，社会学家可以在这些数据的基础上分析人类的行为模式、交往方式等。涂尔干计划就是依据个人在社交网络上的数据分析其自杀倾向，通过脸书的行动 App 收集资料，并将受测用户的活动数据传送到一个医疗资料库。收集完成的数据会接受人工智能系统分析，接着利用预测模型来即时监视受测者是否出现一般认为具有伤害性的行为，从而采取预防措施。社交媒体对于消费者行为的分析数据，也被应用于电子商务领域。

2. 电信行业

电信行业因其遍布世界的网络节点而拥有体量巨大的数据资源，单个运营商其手机用户每天产生的话单记录、信令数据、上网日志等数据就可达到 PB 级的数据规模。尽管电信行业利用信息技术采集数据来改善网络运营、提供客户服务已有数十年的历史，但是传统处理技术下电信运营商实际上只能用到其中不足 1%的数据资源。大数据技术应用包括以下几个方面。

（1）网络安全与网络维护。公众电信网络具有开放性的特征，因此网络安全是最为重要的问题之一。在电信领域运用大数据技术可以很好地加强网络维护，解决网络安全问题。运用大数据技术，首先可以构建较为完整的电信网络安全分析体系。通过对通信网络的历史数据进行分析，可以从人为因素和环境因素两个方面对网络安全问题进行评估。在人为因素方面，主要用于防护恶意的偷窃破坏等。在环境因素方面，主要包括自然灾害，动物破坏等各方面的因素。大数据技术还可以建立完善的风险评估体系，并在云盘空间中储存数据测试结果，通过虚拟化技术提取数据，用于明确具体的分析指标，再按照数据分析结果来判断网络安全的风险点，并提出相应的解决措施。大数据技术还可以通过构建分析平台来对数据进行科学有效的处理，从而保障通信网络安全运行。在网络运维方面，采集基站等硬件设备的数据，分析设备负荷状况，生成设备的扩容、优化、质量排查、扩建等建议，达到均衡网络流量的目的。

（2）大数据技术可以提高数据分析能力。当今社会，移动互联网的发展使得接入网络的移动终端设备越来越多，同时也会产生海量的通信数据。伴随着信息技术的发展，如何在海量的数据资源里面快速挑选有价值的数据，并且充分发挥这些数据的利用价值，已经成为通信领域最重要的工作之一。而大数据技术可以快速地对数据进行收集整理和挖掘，并建立模型对这些数据进行分析，根据用户的网络浏览情况来准确地把握用户需求，在完善网络运营商自身业务的同时，更好地为客户提供服务。大数据技术可以在这些海量数据中获得有用的信息，从而挖掘出隐含在数据当中的价值，这对于业务的决策具有重要的指导意义。

（3）丰富数据处理手段。任何平台的数据模型都可以采用大数据技术进行分析处理。目前各电信运营商的工作重心都是围绕在数据中发掘客户喜好开展的。由于大数据技术的先进性和准确性，由此获得的结果较为精准、客观。电信运营商想要推出更加适合客户的产品与服务，就必须利用大数据技术来挖掘客户的需求，这样才能在最大程度上提高经济效益。运营商通过分析用户的话单数据，界定用户属性，分析手机终端的特征，从而形成套餐推荐、终端推荐等决策；根据用户使用的手机软件（mobile app，App）、访问的网页进行更为全面的用户行为分析、用户喜好分析；采集微博等社交网络数据，了解用户对运营商的评价和意见，进行舆情分析。

3. 制造业

智能制造时代的到来，也意味着工业大数据时代的到来。制造业向智能化转型的过程中,将促进工业大数据的广泛应用。工业大数据是指在工业领域信息化应用中所产生的数据，是工业互联网的核心，是工业智能化发展的关键。工业大数据无疑将成为未来提升制造业生产力、竞争力、创新力的关键要素,也是目前全球工业转型必须面对的重要课题。大数据在工业企业的应用主要体现在以下几个方面。

（1）基于数据的产品价值挖掘。通过对产品及相关数据进行二次挖掘，创造新价值。日本的科研人员日前设计出一种新型座椅，能够通过分析相关数据识别主人，以此确保汽车的安全。这种座椅装有 360 个不同类型的感应器，可以收集并分析驾驶者的体重、压力值，甚至坐到座椅上的方式等多种信息，并将它们与车载系统中内置的车主信息进行匹配，以此判断驾驶者是否为车主，从而决定是否开动汽车。实验数据显示，这种车座的识别准确率高达 98%。三一公司的挖掘机指数也是如此，通过在线跟踪销售出去的挖掘机的开工、负荷情况，就能了解全国各地基建情况，进而对宏观经济判断、市场销售布局、金融服务提供调整依据。

（2）提升服务型生产。增加服务在生产（产品）的价值比重，主要体现在两个方向的延伸。一是前向延伸，就是在售前阶段，通过用户参与、个性化设计的方式，吸引、引导和锁定用户。比如，红领西服的服装定制，通过精准的量体裁衣，在其他成衣服装规模关店的市场下，能保持每年 150%的收入和利润增长，每件衣服的成本仅比成衣高 10%。小米手机也属于这一类。二是后向延伸，通过销售的产品建立客户和厂家的联系，产生持续性价值。苹果手机的硬件配置是标准的，但每个苹果手机用户安装的软件是个性化的，苹果公司销售苹果终端产品只是开始，而通过苹果应用程序商店（App Store）建立用户和厂

商的连接，满足用户个性化需求，提供差异性服务，每年创造收入百亿美元才是最终目的。

（3）创新商业模式。一是基于大数据，制造型企业对外能提供什么样的创新性商业服务；二是在工业大数据背景下，能接受什么样的新型的商业服务。最优的情况是，通过提供创新性商业模式能获得更多的客户，发掘更多的蓝海市场，赢取更多的利润；同时通过接受创新性的工业服务，降低了生产成本、经营风险。以通用电气公司（GE）为例，它不销售发动机，而是将发动机租赁给航空公司使用，按照运行时间收取费用，这样 GE 通过引入大数据技术监测发动机运行状态，通过科学诊断和维护提升发动机使用寿命，获得的经济回报高于销售发动机。在接受服务方面，目前国内外有一批企业提供云服务架构的工业大数据平台。包括海尔收购 GE 的白色家电业务的一揽子合作中，GE 的 Predix 工业大数据平台向海尔开放，接入海尔的工厂，提供工业大数据服务。

4. 金融业

1）大数据在银行业中的应用

信贷风险评估。在传统方法中，银行对企业客户的违约风险评估多是基于过往的信贷数据和交易数据等静态数据，这种方式的最大弊端就是缺少前瞻性。因为影响企业违约的重要因素并不仅仅是企业历史的信用情况，还包括行业的整体发展状况和实时的经营情况。而大数据手段的介入使信贷风险评估更趋近于事实。内外部数据资源整合是大数据信贷风险评估的前提。一般来说，商业银行在识别客户需求、估算客户价值、判断客户优劣、预测客户违约可能的过程中，既需要借助银行内部已掌握的客户相关信息，也需要借助外部机构掌握的个人征信信息、客户公共评价信息、商务经营信息、收支消费信息、社会关联信息等。在供应链金融方面，利用大数据技术，银行可以根据企业之间的投资、控股、借贷、担保以及股东和法人之间的关系，形成企业之间的关系图谱，利于关联企业分析及风险控制。知识图谱再通过建立数据之间的关联链接，将碎片化的数据有机地组织起来，让数据更加容易被人和机器理解和处理，并为搜索、挖掘、分析等提供便利。在风控上，银行以核心企业为切入点，将供应链上的多个关键企业作为一个整体。利用交往圈分析模型，持续观察企业间的通信交往数据变化情况，通过与基线数据的对比来洞察异常的交往动态，评估供应链的健康度及为企业贷后风控提供参考依据。

2）大数据在证券行业中的应用

股市行情预测。大数据可以有效拓宽证券企业量化投资的数据维度，帮助企业更精准地了解市场行情。随着大数据的广泛应用、数据规模爆发式增长以及数据分析及处理能力显著提升，量化投资将获取更广阔的数据资源，构建更多元的量化因子，建立更完善的投研模型。证券企业应用大数据对海量个人投资者样本进行持续性跟踪监测，对账本投资收益率、持仓率、资金流动情况等一系列指标进行统计、加权汇总，了解个人投资者交易行为的变化、投资信心的状态与发展趋势、对市场的预期以及当前的风险偏好等，对市场行情进行预测。智能投顾是近年证券公司应用大数据技术匹配客户多样化需求的新尝试之一，目前已经成为财富管理新蓝海。智能投顾业务提供线上的投资顾问服务，能够基于客户的风险偏好、交易行为等个性化数据，采用量化模型，为客户提供低门槛、低费率的个性化财富管理方案。智能投顾在客户资料的收集分析、投资方案的制定、执行以及后续的

维护等步骤上均采用智能系统自动化完成，且具有低门槛、低费率等特点，因此能够为更多的零售客户提供定制化服务。

3）大数据在保险行业中的应用

借助大数据手段，保险企业可以识别诈骗规律，显著提升骗保识别的准确性与及时性。通过建立保险欺诈识别模型，大规模地识别近年来发生的所有赔付事件，从数万条赔付信息中挑出疑似诈骗索赔，再根据疑似诈骗索赔展开调查。此外，保险企业可以结合内部、第三方和社交媒体数据进行早期异常值检测，包括客户的健康状况、财产状况、理赔记录等，及时采取干预措施，减少先期赔付。保险公司还可以通过大数据分析解决现有的风险管理问题。例如，通过智能监控装置搜集驾驶者的行车数据，如行车频率、行车速度、急刹车和急加速频率等；通过社交媒体搜集驾驶者的行为、情绪数据，如在社交媒体的言行、性格情况等；通过医疗系统搜集驾驶者的健康数据。以这些数据为出发点，如果一个人不经常开车，并且开车习惯稳妥谨慎，那么可以针对他的保费报价比平均水平低 30%～40%，这将极大地提高保险产品的竞争力。

5. 政府

作为重要的基础性战略资源，大数据改变了现代政府治理的思维与方式，成为推进政府治理现代化不可或缺的重要力量。实践证明，大数据治理已经成为政府治理的一种客观形态，在很大程度上推动了政府治理的有效运转。

（1）决策模式转变，公共政策的科学性增强。大数据注重事物之间的联系及耦合性，要求政府决策体现系统性、统筹性、全局性。这在一定程度上必然对政府决策模式产生影响。比如，在疫情防控过程中，需要政府借用精准的大数据对海量的人员流动信息精准识别，这是传统状态下政府决策模式所不具备的。一个成熟的政府在制定公共政策过程中必然注重大数据的开发及精准利用，力求用它来推动公共政策的科学合理性。比如，北京、上海、广州等大城市开发建设的交通信息综合应用平台，集道路传感系统、出租车卫星定位系统、实时视频采集系统等多系统信息于一体，不仅可以用来分析实时交通状况，增强交通管控措施的准确性和时效性，而且可以为后续交通设施建设提供大数据支撑，进而提高交通基础设施建设的科学决策水平。

（2）公共服务的精准化。伴随现代社会的来临，人们对生产生活服务的需求日渐呈现出差异化和个体化特征。这就意味着政府在提供公共服务的过程中要朝精准化的方向努力，大数据的精准利用恰恰为服务型政府提供了有力工具。比如，在扶贫开发工作中，地方政府通过居民经济状况核查比对，不仅能够检测出不符合申领救济资助条件的“假贫困户”，而且能够比对出本应享受低保救助的困难户，进而实现救助服务的精准化。

（3）跨部门的协同合作。大数据的精准使用呼唤部门之间的协同配合，进而形成高度集成、密切融合的数据系统。而传统条块分割的职能部门往往使得各部门数据资源分散，难以形成集约性开发和运用，“数据孤岛”现象时常发生。要解决这一问题，首要之处是革除部门本位主义思想，以系统性思维和开放包容的理念对待大数据的采集和运用。在此基础上，借用现代信息技术形成统一的数据标准和格式规范，加快建设一网集成、信息共享的公共数据平台，积极推动信息跨部门、跨区域互通共享、校验核对、深度整合，实现

部门专网与大数据平台的共享交换，从深层次解决“多网并存”“二次录入”等问题。

1.3.2 大数据技术应用中存在的问题

互联网时代，数据已成为社会重要的基础资源，众多组织采用大数据等现代技术来收集和处理数据。大数据的应用，有助于企业改善业务运营并预测行业趋势。然而，若这项技术被恶意利用，没有适当的数据安全策略，就有可能对用户隐私造成重大威胁。因此，必须意识到大数据技术应用存在的安全问题及其负面影响。

（1）分布式系统。大数据解决方案将数据和操作分布在许多系统上，以便更快地进行处理和分析。这种分布式系统可以平衡负载，并避免产生单点故障。然而，这样的系统很容易受到安全威胁，黑客只需攻击一个点就可以渗透到整个网络。因此，网络犯罪分子可以很容易地获取敏感数据并破坏联网系统。

（2）数据访问。大数据系统需要设置访问控制来限制对敏感数据的访问，否则，任何用户都可以访问机密数据，有些用户可能将其用于恶意行为。此外，网络犯罪分子可以侵入与大数据系统相连的系统，以窃取敏感数据。因此，使用大数据的企业需要检查并验证每个用户的身份。如果使用不正确的身份验证方法，则可能会将访问权限授予未经授权的用户或黑客。这种非法访问会危及敏感数据，而这些数据可能会在网上泄露或出售给第三方。

（3）不正确的数据。网络犯罪分子可以通过操纵存储的数据来影响大数据系统的准确性。为此，网络罪犯分子可以创建虚假数据，并将这些数据提供给大数据系统。例如，医疗机构可以使用大数据系统来研究患者的病历，而黑客可以修改此数据以生成不正确的诊断结果。这种有缺陷的结果不容易被发现，企业可能会继续使用不准确的数据。此类网络攻击会严重影响数据完整性和大数据系统的性能。

（4）侵犯隐私权。大数据系统通常包含个人的隐私数据，网络犯罪分子经常攻击大数据系统，以盗取或破坏敏感数据。此类数据泄露已数次发生在脸书这样的全球性社交网络，致使数百万人的敏感数据被盗。类似的机密数据也可能通过在线交易被泄漏。例如，最近有 8.85 亿人次的银行交易、社会保险号和其他机密数据在网上被泄露，这些安全问题均会威胁和侵犯人们的隐私。

（5）云安全问题。大数据系统收集的数据通常存储在云中，这可能是一个潜在的安全威胁。一些知名企业的云数据被发现存在网络罪犯分子入侵的痕迹。如果存储的数据没有加密，或者没有适当的数据安全措施性，就会出现这些问题，使黑客可以轻松访问敏感数据。为了解决这些安全问题，需要加密所有敏感数据，并使用入侵防御系统来检测网络入侵者。除此之外，还可以采用多因素身份验证来对用户进行身份验证。这种认证机制有助于保护敏感数据免受黑客攻击。另外，定期进行安全审计，可以发现现有安全方法中的漏洞。

1.3.3 大数据技术应用的未来趋势

在大数据应用需求的驱动下，计算技术体系正面临重构，从“以计算为中心”向“以

数据为中心”转型，一些基础理论和核心技术问题亟待破解。

数据与应用进一步分离，实现数据要素化。数据一开始是依附于具体应用的，数据库技术的出现使得数据与应用实现了第一次分离。数据存储在数据库中，不再依赖具体的应用而存在。数据要素化的需求将推动数据与应用进一步分离。数据不再依赖于具体的业务场景，而是以独立的形态存在于数据库中，并通过数据服务为不同的业务场景提供服务。

从单域到跨域数据管理，促进数据要素的共享与协同。以数据为中心的计算的核心目标是数据价值的最大化，实现数据要素的高效共享与协同。传统数据局限在单一企业、业务、数据中心等内部，未来大数据管理将从传统的单域模式发展到跨域模式，跨越空间域、管辖域和信任域。除了云迁移的技术优势之外，最明显的可能是共享不再以物理方式存储在企业内部网络中，企业向第三方提供有价值的数据，这些数据用于战略、财务并且合规性，可以简化供应商和消费者的分销流程。从单域单模态分析到多域多模态融合，实现广谱关联计算，对不同来源、不同模态（如文本、图像、音视频等）的数据进行联合分析，从而实现不同来源与不同模态数据之间的信息互补。因此，探究能够跨模态关联、跨时空关联的广谱关联技术是大数据分析处理的一个重要趋势。

数联网成为数字化时代的新型信息基础设施。数联网将形成一套完整的数联网基础软件理论、系统软件架构、关键技术体系，包括针对数联网软件以数据为中心的特点，从复杂网络和复杂系统等复杂性理论出发，研究数联网软件的结构组成、行为模式和外在性质；针对数联网软件的数据传存算一体化需求，采用数据互操作技术和软件定义思想，研究数联网软件运行机理、体系结构与关键机制；针对数联网软件跨层级、跨地域、跨系统运行带来的可靠性、可用性、安全性等质量挑战，以数据驱动为手段，研究数联网环境下保障服务质量与保护质量的原理、机制与方法。

数据湖的应用，流数据和静态数据相统一。随着数字企业的建设和发展，数据湖已经成为企业的一种非常经济的选择。远程工作和混合工作环境的兴起增加了对数据湖的需求，以实现更快、更高效的数据操作。随着企业迁移到云平台并专注于云计算数据湖，他们也将转向将数据仓库与数据湖融合。创建数据仓库是为了针对 SQL 分析进行优化，但是需要一个开放、直接和安全的平台来支持快速增长的新型分析需求和机器学习，最终将使数据湖成为数据的主要存储方式。数据湖的采用将持续到 2022 年及以后，市场规模将从 2020 年的 37.4 亿美元增长到 2026 年的 176.0 亿美元，在 2021—2026 年预测期间的复合年增长率为 29.9%。流数据以及驻留在数据库或数据湖中的数据来源将继续与流媒体和操作系统融合，从而提供更统一的分析。对分布式集群执行分析，并将其他集群上的流数据和操作数据源的结果聚合到一个单一的控制平台中将成为常态。

扩展性优先设计到性能优先设计。数据规模急剧增长，大数据处理需求越来越走向深度价值挖掘，数据处理计算越发密集，数据管理与处理的成本成为大数据管理与处理系统的重要考量因素，传统“扩展性优先”的大数据处理系统设计将会被“以性能优先”的系统设计代替。Spark、Flink 等系统在大数据处理生态系统中的占有率明显体现了这一趋势，图计算（图加速器、图计算框架等）、深度学习框架等领域专用大数据处理系统的崛起也是这一系统设计理念在技术生态上的表现。智能化数据管理、近似计算等新兴管理和处理

方法成为性能优先设计的重要技术手段。

大数据的技术标准制定和以开源社区为核心的软硬件生态系统将成为未来发展的重点。随着大数据在各个领域应用的迅速普及，标准化需求将不断增长，与大数据流动融合、质量评估，以及与行业、领域应用密切相关的大数据标准将成为发展重点。开源社区在大数据软硬件生态建设中的地位不断加强，对开源社区的主导权争夺将成为各国技术、产品和市场竞争的重点。

1.4 大数据营销的内涵及其特点

在大数据时代到来之前，企业营销只能采用传统的营销数据获取和分析方式，包括客户关系管理系统中的客户信息、广告效果、会展等一些线下营销活动的效果。营销数据的来源仅限于消费者某些方面的有限信息，不能提供全局充分的提示和线索。互联网时代社交媒体的普及，为企业的营销管理带来了新型的数据，包括消费者使用网站的数据、地理位置的数据、邮件数据、社交媒体数据等。下面介绍大数据营销的内涵及特点。

1.4.1 大数据营销的内涵

大数据营销是大数据时代的企业借助大数据技术将新型的数据与传统数据进行整合，从而更全面地了解消费者的消费行为信息，创造更深层次并且具有互动关联性的顾客关系的新型的市场营销方式。

传统的营销理念是根据顾客的基本属性，如顾客的性别、年龄、职业和收入等来判断顾客的购买力和产品需求，从而进行市场细分，以及制定相应的产品营销策略，这是一种静态的营销方式。而大数据不仅记录了人们的行为轨迹，还记录了人们的情感与生活习惯，能够精准预测顾客的需求，从而实现以客户生命周期为基准的精准化营销。大数据营销是一个动态的营销过程，包括客户信息收集与处理、客户细分与市场定位、辅助营销决策与营销战略设计、精准的营销服务、营销结果反馈。

1. 客户信息收集与处理

客户数据收集与处理是一个数据准备的过程，是数据分析和挖掘的基础，是搞好精准营销的关键和基础。精准营销所需要的信息内容主要包括描述信息、行为信息和关联信息等 3 大类。描述信息是顾客的基本属性信息，如年龄、性别、职业、收入和联系方式等基本信息；行为信息是顾客的购买行为的特征，通常包括顾客购买产品或服务的类型、消费记录、购买数量、购买频次、退货行为、付款方式、顾客与企业的联络记录，以及顾客的消费偏好等；关联信息是顾客行为的内在心理因素，常用的关联信息包括满意度和忠诚度、对产品与服务的偏好或态度、流失倾向及与企业之间的联络倾向等。

2. 客户细分与市场定位

企业要对不同客户群展开有效的管理并采取差异化的营销手段，就需要区分出不同的客户群。在实际操作中，传统的市场细分变量，如人口因素、地理因素、心理因素等只能

提供较为模糊的客户轮廓，已经难以为精准营销的决策提供可靠的依据。大数据时代，利用大数据技术能在收集的海量非结构化信息中快速筛选出对企业有价值的信息，对客户行为模式与客户价值进行准确判断与分析，帮助企业在众多用户群中筛选出重点客户，精准确定企业的目标客户，从而帮助企业将其有限的资源投入到这少部分的忠诚客户中，以较小的投入获取较大的收益。

3. 辅助营销决策与营销战略设计

在得到基于现有数据的不同客户群特征后，市场人员需要结合企业战略、企业能力、市场环境等因素，在不同的客户群体中寻找可能的商业机会，最终为每个客户群制定个性化的营销战略，每个营销战略都有特定的目标，如获取相似的客户、交叉销售或提升销售量，以及采取措施防止客户流失等。

4. 精准的营销服务

动态的数据追踪可以改善用户体验。企业可以追踪了解用户使用产品的状况，做出适时的提醒。例如，食品是否快到保质期；汽车的使用磨损情况，以及是否需要保养维护等。流式数据使产品“活”起来，企业可以随时根据反馈的数据做出方案，精准预测顾客的需求，提高顾客生活质量。针对潜在的客户或消费者，企业可以通过各种现代化信息传播工具直接与消费者进行一对一的沟通，也可以通过电子邮件将分析得到的相关信息发送给消费者，并追踪消费者的反应。

5. 营销方案设计

在大数据时代，一个好的营销方案可以聚焦到某个目标客户群，甚至精准地根据每一位消费者不同的兴趣与偏好为他们提供专属的市场营销组合方案，包括针对性的产品组合方案、产品价格方案、渠道设计方案、一对一的沟通促销方案，如渠道设计、网络广告的受众购买方式和实时竞价技术、基于位置的促销方式等。

6. 营销结果反馈

营销活动结束后，大数据技术可以帮助企业对营销活动执行过程中收集到的各种数据进行综合分析，从海量数据中挖掘出最有效的企业市场绩效度量，并与企业传统的市场绩效度量方法展开比较，以确立基于新型数据的度量的优越性和价值，从而对营销活动的执行、渠道、产品和广告的有效性进行评估，为下一阶段的营销活动打下良好的基础。

1.4.2　大数据营销的特点

与传统的市场营销方式相比较，大数据营销主要有以下几方面的特点。

（1）客观性：市场调查是传统市场营销管理中常用的一手数据获取手段，这种方式往往因受访者的主观因素产生局限性；而基于大数据技术的营销数据获取，是通过互联网采集大量的行为数据，是消费者行为的真实记录。

（2）多样性：大数据的数据来源通常是多样化的，多平台化的数据采集能使企业对网民行为的刻画更加全面而准确。这些数据来源包含互联网、移动互联网、广电网、智能电

视、可穿戴设备、物联网等数据。

（3）实时性：在互联网时代，消费者的网络消费行为和购买方式是实时动态变化的。在消费者需求点最高时及时采取营销策略非常重要。例如，泰一传媒公司是一家知名的大数据营销企业，它针对实时性这个特点提出了“时间营销策略”，可通过技术手段充分了解消费者的需求及波动规律，并及时响应他们的动态需求，精准投放广告，使消费者在决定购买的“黄金时间”内及时接收到商品广告。

（4）个性化：互联网时代新媒体的涌现，使广告主的营销理念从传统的“媒体导向”向“受众导向”转变。大数据技术可以帮助广告主洞察目标受众身处何方，关注着什么位置的什么屏幕，实现以受众为导向的个性化广告投放。甚至，大数据技术与人工智能技术的结合，可以做到当不同用户关注同一媒体的相同界面时，广告内容有所不同，从而实现“千人千面”的个性化营销。

（5）经济性：与传统营销的广告投放“几乎有一半的广告费被浪费掉”相比，大数据营销在最大程度上，让广告主的投放做到有的放矢，并可根据实时动态的效果反馈，及时对投放策略进行调整，从而降低市场成本。

（6）关联性：大数据营销的一个重要特点在于分析消费者关注的广告与广告之间的关联性，由于大数据在采集过程中可快速得知目标受众关注的内容，以及可获取消费者的位置等信息行为，这些数据的联合使用可让广告的投放过程产生前所未有的关联性。消费者所看到的是，呈现的上一条广告与下一条广告的深度互动。

案例 1-1

华为融合数据湖：加速银行业基于大数据的业务创新

随着全球银行业数字化转型的加快，以及数据驱动战略在全球领先银行的落地实践，融合数据湖已成为越来越多主流银行实现业务创新的首选平台。从中国大量领先银行的融合数据湖实践，到海外众多国家和地区（如马来西亚、新加坡和北欧地区等）主流银行对融合数据湖平台的接纳，华为公司正通过联合业界独立软件供应商伙伴，帮助越来越多的银行迈向以数据驱动业务创新的崭新路径。

长期以来，数据仓库系统一直是企业信息技术架构的重要组成部分，特别是对于银行业这类高度依赖数字技术的传统行业而言，无论是在传统的监管报送，还是在近年来火热的商业智能领域，数据仓库都扮演着越来越重要的角色。传统的数据仓库平台通常其处理能力在数百 GB 到数百 TB 不等，而一个大型现代银行平均每天产生的数据量都高达几 TB 甚至几十 TB，每年的新增数据量则高达 PB 级别；同时，随着银行深入融入客户的场景化生活，每天都会产生大量的非结构化数据，如埋点数据、交易日志、图像和音/视频等，这些都给传统用来处理单一结构化数据类型和有限数据量的数据仓库平台带来了严峻挑战。

通过整合分布式数据仓库平台和大数据处理平台，融合数据湖具备了对结构化数据和非结构化数据的同时处理能力，以及实时数据处理和线下批量处理的能力，并借助分布式线性扩展能力来适应海量数据的处理需求。伴随着金融业务的日趋移动化和线上化，以及客户体验的快速提升，融合数据湖已成为银行构建以客户为中心的场景化金融、实现快速

业务创新的重要依托平台。华为融合数据湖解决方案见图 1-4。

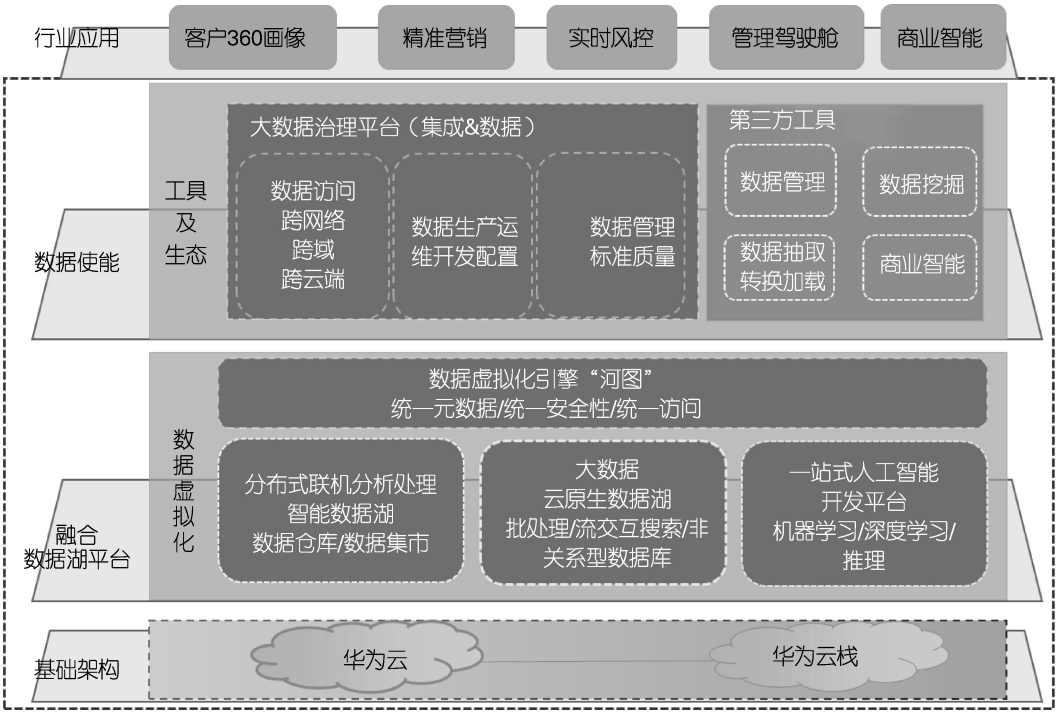

图 1-4　华为融合数据湖解决方案

资料来源：https://www.sohu.com/a/427957532_296821

思考题

1. 什么是大数据？
2. 试阐述大数据的产生背景。
3. 举例说明大数据技术的主要应用场景。
4. 试述大数据营销的内涵及其特点。

案例 1-2：趣多多的愚人节营销

第 1 章扩展阅读

即测即练

自学自测　扫描此码