模块 1

初识大数据

学习要点

- 大数据的定义。
- 大数据的分析工具。
- 大数据的应用。
- 大数据的处理过程。

1.1 必备知识

1.1.1 大数据概述

近年来，随着社交网络渗透进人们的生活，人们从其中的数据中观察到更多的人类社会的复杂行为模式。大量的信息汇集、分析的第一手资料，产生了重要的数据资产。这些数据资产产生了巨大的经济价值与社会价值。人类历史迎来第四次革命，大数据的产生使得从前孤立的数据具有关联性，使得人们发现新的机遇，创造新的价值。

1. 大数据的定义

作为全球咨询行业的标杆，麦肯锡公司俨然成为大数据研究的先驱。2011 年，麦肯锡的报告中给出关于大数据的定义：大数据（big data，mega data）或称巨量资料，指的是需要新处理模式才能具有更强的决策力、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。大数据的“大”的界定范畴是动态的，从前的 GB 就是数据类的巨大范畴，但是大数据出现后，在物理、基因等很多领域，TB 级的数据已很普遍，更有 PB，甚至 EB 级也并不

罕见。数据的类型有很多种，其主要分为结构化数据、半结构化数据和非结构化数据。因此，数据量的不断增长及数据类型的多样化，都给大数据系统的存储和计算带来了不小的挑战。

2. 大数据的价值

大数据不等于大量的数据。大数据的三个明显的特征分别是数据量(volume)大，数据实时性(velocity)高，数据多样性(variety)丰富。

量大是大数据的必要条件，但并不充分，因为大数据的数据增长是爆发性的，因此计量单位一般是 PB、EB 甚至 ZB。

大数据具有多变性，它包括结构化数据和非结构化数据，与传统数据相比，大数据呈多样性。人们必须运用大数据的技术从海量的非结构化数据中提炼、分析数据，因此非结构化的数据在数据存储、收集上更为复杂。

大数据的实时性代表这些数据的分析与使用是随时的、实时的，与传统数据不同，动态性是大数据的显著特征。

在传统的数据时代，商业决策的做出主要依靠历史数据与经验总结，不可避免地出现由于信息滞后造成的决策效果不佳，在大数据时代，依据在线的、实时的数据收集与分析，实现精准营销，极大地提高决策实效性。

在大数据时代，随着个人计算机和手机移动端的普及，每个人都在随时随地提供数据。各种各样的行为，如清晨搭车、点击网上商品、刷卡购物、使用手机玩游戏等，都会产生专属于每个人的数据痕迹，然后形成大数据被记录下来，每个人的年龄、性别、消费偏好、喜欢的商品类型、出行习惯等信息都被记录成数据，商家可以提取有效的商业信息，根据客户的习惯和偏好，精准营销。

大数据使每个人从中受益，生物领域的专家在对基因信息、遗传物质的信息等分析的基础上，结合每个人特有的健康数据、身体功能指标、既往病史、过敏史等，得出研究结果。医疗研发机构根据互联网采集的病人数据基础，推进慢性疾病医疗方面的服务，探索慢性疾病的信息管理和新型的医疗方式，同时，互联网借助医疗机构的治疗数据，构建起慢性疾病患者的大数据。

大数据的时代拥有更便捷的方式来甄选有效、真实的数据。大数据的多样性使来自不同数据源、不同维度的数据相互之间产生一定程度的关联性，这种关联性可以通过多种方式交互验证。例如，某厂将生产量少报一半，目的是少报税，但是它的生产电力等各种能耗却超过对应的指标一半，这种虚假数据就能及时被大数据系统甄别。大数据能根据各种关联性的明细数据综合判断出企业真实的盈利能力，并能形成成熟的数据信息，生成更多更有价值的数据。

数据作为现代社会的资源之一，不同于物质性的资源，那些资源缺乏可再生性，无法共享，但是数据资源却能反复使用，并产生不同的价值。这种良性的资源使用，使得大数据能

发生巨大作用，产生出多赢的局面。

大数据因其背后的价值，被喻为新世纪的黄金，被看作新兴起的经济元素，大数据不仅本身可以看作重要的生产要素，其对产品的形成过程也起到至关重要的作用。大数据的主要价值如下。

1）大数据是新时代信息技术的关键支撑

大数据的热潮在全球的盛行，顺应了现代信息技术发展的趋势。互联网时代为大数据的普及和发展打下了坚实的基础，人们能随时通过移动端使用互联网，伴随着物联网、网上购物、交友网站和云计算的兴起，每个人的数据无处不在，且随时随地产生。作为信息技术时代的产物，大数据的应用又反作用于信息技术的发展，促进物联网、云计算等技术的革新，大数据作为融合新时代信息技术的关键支撑，为物联网、云计算等现代信息技术的发展提供了依托的平台。

2）大数据是促进现代社会经济发展的推动力

大数据本身隐含着巨大的经济价值和社会价值。大数据行业的爆发式发展，将带来一批针对大数据市场的新的商业理念、新的营销服务、新的产品和新的技术，推动现代信息产业的发展。在国内的城市建设、民生发展等领域，大数据也起着举足轻重的作用。目前，我国着力推行智慧城市的建设，大数据的应用能将城市中方方面面的数据联合起来，分析提取有效数据，依靠它们做出智慧决策。例如，可以依照不同的时间段，某条道路的车流量，拥堵状况数据分析，来合理设置红绿灯的时间，缓解交通。随着智慧城市在我国的不断建设和完善，大数据在提升地方政务能力和社会管理能力方面发挥着重要作用，使之形成充满各地特色的、新兴的智能领域应用。

大数据帮助企业深度挖掘客户喜好，助力企业智能决策。大数据为企业洞察用户提供了有力的武器，满足企业针对客户的个性化营销需求，为企业做出正确的市场决策提供更多维度。大数据出现以前，市场调查是通过人工方式获取，采用调研和营销实现的，这样的数据具有明显的“人工计划”特征，在市场调查之前，收集数据的样板、调研方式、分析方式和获取数据的目的都有一个清晰的规划，因此，这些数据是“结构化”的。依靠互联网产生的大数据，其来源是互联网用户行为，包括网页检索、页面浏览、网络交易和网络社交行为等，它并不受人工计划，因此数据的产生、分析过程具有不确定性，这样的数据是多维度的，360 度全方位接近用户，从而使决策的依据更科学。

3）大数据将成为科技创新的引擎

在人工数据时代，信息化的滞后，使得大量的数据彼此分离，闲置在各自的系统空间里，技术的落后使传统的信息处理方式无法满足科技发展的需求。新兴的大数据在整合数据、分析数据、存储数据、处理数据、应用数据，解决系统实时性的、并发性的问题，包括云存储、数据价值分析等方面都颠覆了传统。大数据成为各个领域科技创新的引擎。例如，大型家电生产厂家在产品生产线上安装传感器采集数据，这些生产信息的分析和价值挖掘，能实时提高产品合格率。在电力领域，智能电表的数据采集同样发挥着不可忽视的作用，其不仅作

为电费收取的依据，还扮演着判断房屋空置与否的角色，延伸开来，可作为城市房价定位的参考指标。再者，电网所采集的耗电量数据可以判断出该部分地区的商业发展情况。在未来，不论是国家政府，还是金融商业、各个数据集中的领域，大数据将成为各企业和单位提升竞争力、占领市场的核心竞争力，加速企业从“业务驱动”向“数据驱动”转型升级，为企业提高利润，增强实力，研发产品带来新的机遇。

3. 大数据的特点

如图 1-1 所示，大数据具有四大特点：volume（容量），代表海量的数据规模；variety（种类），代表数据类型的多样性；value（价值），代表深度的数据价值；velocity（速度），代表数据流转的迅速与体系的动态性。

1）volume：数据体量巨大

目前，人类社会所生产的印刷材料总和的数据量是 200 PB（1 PB＝2^{10} TB），人类说过的语言的总和数据量大约是 5 EB（1 EB＝2^{10} PB）。数据的体量决定了它背后的信息价值，随着各种移动端的流行和云存储技术的发展，现代社会的人类活动都可以被记录下来，因此产生了海量的数据。发送的微博、自拍的照片、戴的运动手环等通过互联网上传到云端，各种数据聚集到特定地点的存储系统，如政府机构等，形成了体量巨大的数据。

图 1-1　大数据的特点

2）variety：数据类型呈多样性

数据分为结构化数据与非结构化数据两种，而互联网将网络通过各种移动端形成了整体，人们不仅可以通过互联网获取数据，同时也是数据的传播者，相对于过去，以文本为主的结构化数据往往是便于存储的，随着非结构化数据越来越多，如网络小说、拍摄的视频、录制的音频、共享的地理位置等，这些多样性的数据使得对数据处理的能力要求更高。需要对数据进行加工、清洗、分析等步骤，将它们变为易于存储的结构化数据。这需要在海量的数据之间发现它们之间的关联性，把看似毫无关系的数据联系起来，形成有价值的信息。

3）velocity：处理迅速

velocity 是大数据区分于传统数据挖掘的最显著特征，大数据具有实时性。例如，人们出去吃饭，导航餐厅，用移动端的地图查询位置，选择不堵车的路线，还会从网络上查看餐厅的评价如何，吃饭后，也许会拍下食物和餐厅的照片上传到微博，因此，各种网络的链接带来

大量的数据交换，对速度的要求更高，要以实时的方式传达给用户。

4）value：数据价值大

大数据的应用在物联网、云计算、大数据挖掘等技术迅速发展的带动下，呈现出它的过程：把数据源的信号转换为数据，再把大数据加工成信息，通过获取的信息来做决策。因此，大数据价值的挖掘过程就像大浪淘沙，数据的体量越大，相对有价值的数据就越少。

大数据的价值密度实际是比较低的，因为数据采集并非都是及时的，样本的数量有限，数据不完全连续，但是，当数据的体量越来越大时，就能从海量数据中提取有价值的信息，为决策做支撑。

1.1.2　大数据的产生和类型

1. 大数据的产生

早在 3 000 多年前的埃及，人类就用过计数来统计、策划、安排日常的劳动与生活。16 世纪的欧洲，人类通过一些经验数据来总结人文规律。伴随着信息现代化的进步和数字化发展的日新月异，人们已经不再将数据仅仅作为刻度表征，而通过数据对世间万物进行表达和量化，人们通过表现为数据的信息进一步认识世界。数据成为表述世界的通用语言，所有图像、文字、图形、多媒体等都能采用数据形式表达。

不论是早期人类的计数还是后来人类用数据总结规律，通过对数据的研究，进行规律总结，人类对数据的利用推动了人类历史的进程。公元前 3000 年，两河流域生活着苏美尔人，他们建造了繁荣的城镇，发展了农业。步入农业社会的美苏尔人随着人口的增加，遇到了一系列问题：人口越来越多，怎么管理？如何保持社会安稳？该收多少税赋？该种多少小麦？于是苏美尔人发明了一套专门处理大量的数字与数据的符号，如图 1-2 所示。

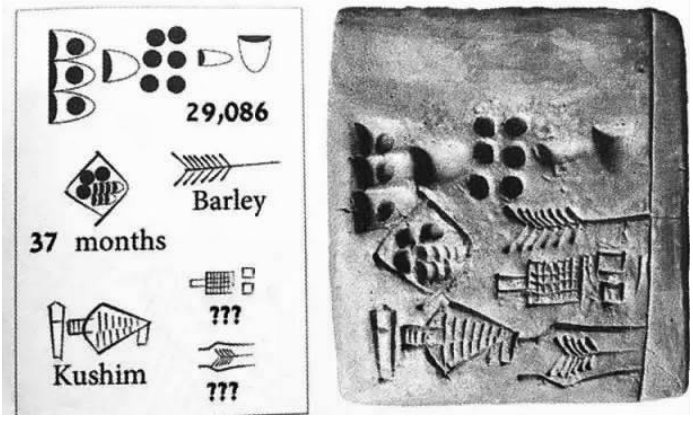

图 1-2　苏美尔人用于统计的符号

这种方式极大地提高了苏美尔人安排生产生活的效率，显示出数据的力量。

步入现代社会，人们日常面临更多、更复杂的问题，迫使数据的归纳和使用方法变得更

为重要。1980 年，未来学家托夫勒在其所著的《第三次浪潮》中提到了“大数据”一词。2011 年，麦肯锡正式定义了大数据的概念。

第一次工业革命以蒸汽机和印刷术为标志，第二次工业革命以内燃机和电信技术为标志，第三次工业革命以核能为标志，而现在的第四次工业革命则以以数据和内容作为核心的互联网为标志。在商业、经济及其他领域中，不论传统行业还是新兴行业，谁率先成功地融合互联网，能够从互联网的大数据中发现隐含的规律，基于数据和分析做出决策，谁就能够抢占先机，占领蓝海。现在人们生活中的各个方面的信息通过互联网被不断地采集、分析、汇总，海量的数据产生了各式各样的信息资产，这些信息资产被称为大数据，其增长迅速，又具有多样性。

大数据时代已经来临，美国在 2012 年成立了“大数据指导委员会”，规划了大数据研究计划。欧盟与日本也相继出台大数据战略规划。2016 年，我国“十三五”规划中将推动大数据的应用纳入其中，国家将加大大数据在工业制造、研发、产业链全流程的应用，鼓励服务业基于大数据分析精准营销，定制服务。

2. 大数据的类型

大数据的数据类型繁多，互联网作为大数据的主要来源，包含了各种数据源，如声音和电影文件、文档、网络日记、元数据、E-mail、表格数据、图像、地理定位数据、文本数据等，其中网页日记为半结构化数据，图片、视频为非结构化数据。

1.1.3 大数据分析工具

1. InfoSphere BigInsights 简介

InfoSphere BigInsights 是由 IBM 公司推出的大数据平台软件，用于处理流数据和持久性数据，将大数据转为大洞察。旨在帮助公司从大量不同范围的数据中挖掘商机并进行分析，因为用传统方法来处理大量数据有些不切实际且难度很大，常常会忽略或丢弃一些数据，如日志记录、点击流、社会媒体数据、新闻摘要、电子传感器输出，甚至是一些事务数据等。为了帮助公司以一种有效的方法从这些数据中获取有价值的信息，InfoSphere BigInsights 提供了无分享硬件集群和内置分析技术。它能透明地分配存储在附加子集群附件的处理器，能有效减少网络信息流通量，提高运行性能。在容错方面，BigInsights 根据管理员指定的参数自动复制多个磁盘上的每一部分数据。该复制操作使 BigInsights 能够通过将工作重定向至别处，自动从磁盘或节点故障中恢复。它是一个可以增强现有分析基础架构的平台，能够对大量的原始数据进行过滤，并将结果与存储在 DBMS 或数据仓库中的结构化数据进行组合。

2. BigQuery 简介

BigQuery 是 Google 推出的一项 Web 服务，该服务让开发者可以使用 Google 的架构来运行 SQL 语句对超级大的数据库进行操作，BigQuery 旨在分析数十亿行近似的数据，使用类

SQL 语法。它并不是完全符合 SQL 数据库的替代，并不适用于交易处理应用。BigQuery 支持分析交互风格。使用 select 命令构建查询，对于任何 SQL 开发者来说应该都很熟悉。

查询语言包括支持标准操作，如 joining、sorting、grouping、内嵌数据结构等。其可以支持统计函数，如 count、sum、average、variance 和 standard deviation(标准偏差)等。

3. 魔镜简介

魔镜为企业提供数据可视化、分析、挖掘的整套解决方案及技术支持，是一款基于 Java 平台开发的可扩展、自助式分析、大数据分析产品。魔镜在垂直方向上采用三层设计：前端为可视化效果引擎，中间层为魔镜探索式数据分析模型引擎，底层对接各种结构化或非结构化数据源。它是由苏州国云数据科技有限公司开发的首款免费大数据可视化分析工具。先后获得了黑马大赛全国百强、国际精英创业周 A 类项目等殊荣，魔镜支持各种数据源，颠覆了传统的 Excel 分析和报表，操作简单方便，自动拖曳建模，是目前功能较为全面的可视化分析平台，拥有国内最大的可视化效果库，支持 500 多种图表，包括列表、饼图、漏斗图、散点图、线图、柱状图、条形图、区域图、气泡图、矩阵、地图、树状图、时间序列相关的图表，还支持树图、社交网络图、3D 图表等多维动态图表类型。

魔镜视觉效果库超大，数据市场开放。这款大数据分析工具已经为超过一万多家企业提供了其行业的大数据解决方案。

魔镜现在有五个版本，即企业基础版、企业标准版、企业高级版、云平台版和 Hadoop 版。

(1)企业基础版：可代替报表工具、传统 BI(商业智能)，适合中小型企业使用，内部使用时可以全公司协同分析。

(2)企业标准版：可实现企业的基础数据分析和数据结果呈现，满足企业的一般数据分析需求。

(3)企业高级版：适合规模较大的公司，建立数据仓库，帮助企业完成数据转型。

(4)云平台版：免费版本，适合接受 SaaS(软件及服务)的企业和个人进行数据分析使用。

(5)Hadoop 版：支持 PB 级大数据计算，实时计算，完美兼容 Spark、HBase 非结构化计算，适合大企业。

1.2　扩展知识

1.2.1　大数据的应用

1. 大数据经典案例

在大数据时代，如何通过大数据的应用来实现企业价值，是很多企业思考的问题，下面

介绍一些利用大数据创造价值的成功案例。

1)啤酒与尿布

全球零售业巨头沃尔玛在对消费者的购物行为分析时发现,男性顾客在购买婴儿尿片时,常常会顺便搭配几瓶啤酒来犒劳自己,于是沃尔玛尝试推出了将啤酒和尿布摆在一起的促销手段。没想到这个举措居然使尿布和啤酒的销量都大幅增加。如今,“啤酒+尿布”的数据分析成果早已成了大数据技术应用的经典案例,被人津津乐道。

2)数据新闻让英国撤军

2010 年 10 月 23 日,《卫报》利用维基解密的数据做了一篇“数据新闻”。将伊拉克战争中所有的人员伤亡情况均标注在地图上。地图上一个红点便代表一次死伤事件,鼠标点击红点后弹出的窗口则有详细的说明:伤亡人数、时间,造成伤亡的具体原因。密布的红点多达 39 万,显得格外触目惊心。一经刊出立即引起全国震动,推动英国最终做出撤出驻伊拉克军队的决定。

3)Google 成功预测冬季流感

2009 年,Google 通过分析 5 000 万条美国人最频繁检索的词汇,将其与美国疾病中心在 2003 年至 2008 年间季节性流感传播时期的数据进行比较,并建立一个特定的数学模型。最终,Google 成功预测了 2009 年冬季流感的传播,甚至可以具体到特定的地区或州。

4)大数据与乔布斯癌症治疗

乔布斯是世界上第一个对自身所有 DNA 和肿瘤 DNA 进行排序的人。为此,他支付了高达几十万美元的费用。他得到的不是样本,而是包括整个基因的数据文档。医生按照其所有基因按需下药,最终这种方式帮助乔布斯延长了好几年的生命。

5)奥巴马大选连任成功

2012 年 11 月,奥巴马大选连任成功的胜利果实也被归功于大数据,因为他的竞选团队进行了大规模与深入的数据挖掘。时代杂志更是断言,依靠直觉与经验进行决策的优势急剧下降,在政治领域,大数据的时代已经到来。各种媒体、论坛、专家铺天盖地的宣传让人们对大数据时代的来临兴奋不已,无数公司和创业者都纷纷跳进了这支狂欢队伍。

6)微软大数据成功预测奥斯卡 21 项大奖

2013 年,微软纽约研究院的经济学家大卫·罗斯柴尔德(David Rothschild)利用大数据成功预测 24 个奥斯卡奖项中的 19 个,成为人们津津乐道的话题。2014 年,罗斯柴尔德再接再厉,成功预测第 86 届奥斯卡金像奖颁奖典礼 24 个奖项中的 21 个,继续向人们展示现代科技的神奇魔力。

2. 大数据的操作实例

1)中信银行信用卡营销

(1)实施背景。中信银行信用卡中心是国内银行业为数不多的几家分行级信用卡专营机构之一,也是国内具有竞争力的股份制商业银行信用卡中心之一。近年来,中信银行信用

卡中心的发卡量迅速增长，2008 年银行向消费者发卡约 500 万张，而这个数字比 2010 年增加了一倍。随着业务的迅猛增长，业务数据规模也急剧膨胀。中信银行信用卡中心无论在数据存储、系统维护等方面，还是在有效地利用客户数据方面，都面临着越来越大的压力。

中信银行信用卡中心迫切需要一个可扩展、高性能的数据仓库解决方案，支持其数据分析战略，提升业务的敏捷性。通过建立以数据仓库为核心的分析平台，实现业务数据集中和整合，以支持多样化和复杂化的数据分析，如卡、账户、客户、交易等主题的业务统计和 OLAP（联机分析处理）多维分析等，提升卡中心的业务效率；通过从数据仓库提取数据，改进和推动有针对性的营销活动。

（2）技术方案。从 2010 年 4 月至 2011 年 5 月，中信银行信用卡中心实施了 EMC Greenplum 数据仓库解决方案。实施 EMC Greenplum 解决方案之后，中信银行信用卡中心实现了近似实时的商业智能（BI）和秒级营销，运营效率得到全面提升。中信银行的大数据应用技术架构图如图 1-3 所示。

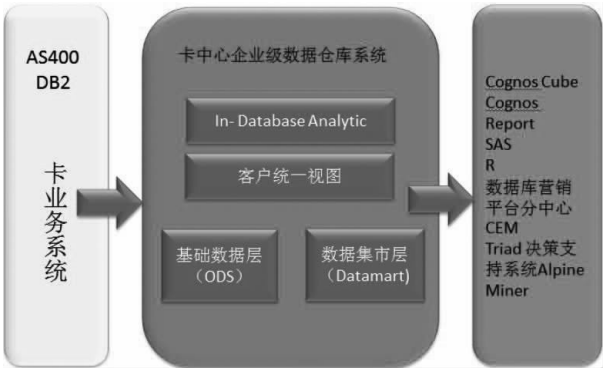

图 1-3　中信银行的大数据应用技术架构图

Greenplum 解决方案的一个核心的功能是，它采用了“无共享”的开放平台的 MPP 架构，此架构是为 BI 和海量数据分析处理而设计的。目前，最普遍的关系数据库管理系统（Oracle 或 Microsoft SQL Server），都是利用共享磁盘架构来实现数据处理的，会牺牲单个查询性能和并行性能。使用 Greenplum 数据库提供的 MPP 架构，数据在多个服务器区段间会自动分区，而各分区拥有并管理整体数据的不同部分，其所有的通信是通过网络互连完成的，没有磁盘级共享或连接，使其成为一个“无共享”架构。Greenplum 数据库提供的 MPP 架构为磁盘的每一个环节提供了一个专门的、独立的高带宽通道，段上的服务器可以以一个完全并行的方式处理每个查询，并根据查询计划在段之间有效地移动数据，因此，相比普通的数据库系统，该系统提供了更高的可扩展性。

（3）效益提升。2011 年，中信银行信用卡中心通过其数据库营销平台进行了 1 286 个宣传活动，每个营销活动配置平均时间从 2 周缩短到 2～3 天，且市场活动中答应客户在刷满一定金额或次数后送给他们的礼品，可以在客户刚好满足条件的那次刷卡后马上获得，实现了秒级营销，而不必等待几个工作日。2011 年的前三个季度，中信银行信用卡中心交易量

增加 65%，比股份制商业银行的平均水平高 14%，比中国所有银行的平均值高 4%。中信银行信用卡中心迄今已为客户进行了 4 000 万次的信用额度调整。中信银行信用卡中心催收管理团队使用了基于数据仓库的 FICO TRIAD 系统后，信用卡不良贷款(NPL)比率同比减少了 0.76%。中信银行信用卡中心电话销售中心将所有外呼营销历史整合到数据仓库，通过对大量历史数据分析后调整营销策略，在上线后的第一个月便实现单位工时创收提升 33%，笔均贷款额提升 18%，目前银行正在开发针对每个产品的营销响应模型，以进一步提升产能。

2)兴业证券客户综合分析管理系统

(1)实施背景。随着我国证券市场的日益规范和成熟，证券公司之间的竞争也日趋激烈。证券公司越来越注重对客户的有效服务及对营业部、经纪人的业绩管理，而现有的 IT 系统通常只是面向业务交易而设计的，随着市场竞争的日益激烈，其越来越不能满足证券公司的决策分析需要。为了提升证券公司的客户服务及精准营销的能力，兴业证券采用大数据技术提升自身客户综合分析管理系统能力。

(2)技术方案。吉贝克信息技术有限公司针对兴业证券所面临的环境，采用数据仓库和数据挖掘技术，自主研发了证券公司客户综合分析管理系统，以满足证券公司日益深化的客户管理需求。客户综合分析管理系统的功能架构图如图 1-4 所示，客户生命周期服务管理的示意图如图 1-5 所示。

围绕客户维护生命周期，在不同的生命周期阶段采用有针对性的方式来降低客户流失率，如主动关怀、客户营销、流失预警挽留、销户挽留等。

(3)效益提升。采用了客户综合分析管理系统之后，数据加载速度明显提升，目前 100 万行数据入库仅需 6～7 秒，10 GB 的数据加载和导出也可以在 5 分钟内完成。同时，数据处理和查询的效率也显著提升，目前每天的数据处理时间基本控制在 2 小时以内。对于日常简单查询，在 50 条并发查询的情况下可以实现 1 秒内完成。对于长时间跨度、多条件的复杂查询，也能在 5 秒内完成。

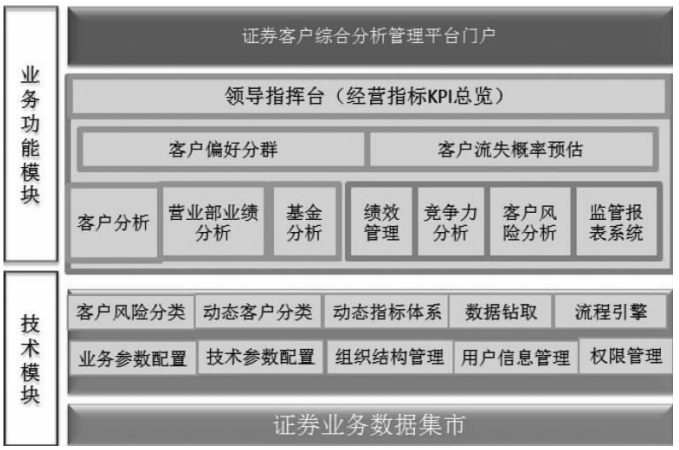

图 1-4　客户综合分析管理系统的功能架构图

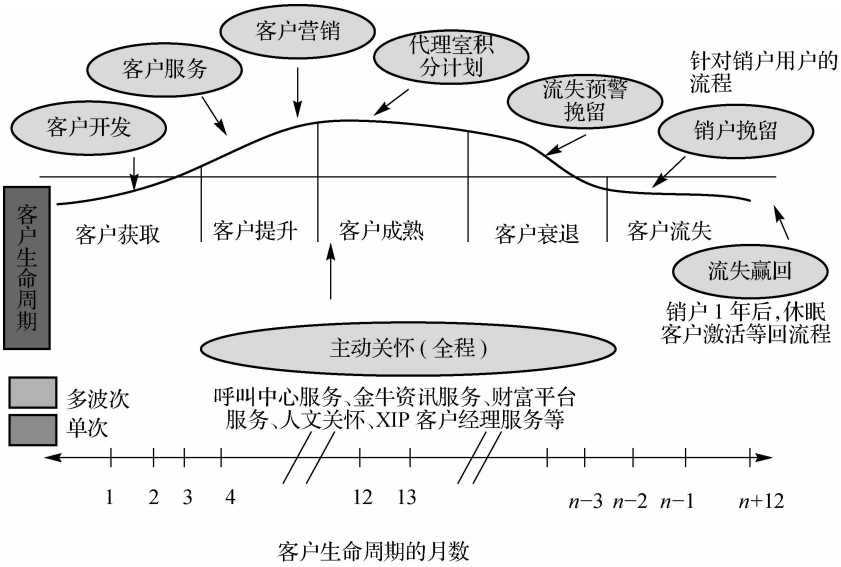

图 1-5　客户生命周期服务管理的示意图

1.2.2　大数据技术概述

1. 大数据的处理过程

大数据的处理过程为大数据的采集—大数据的导入与预处理—大数据的统计与分析—大数据的挖掘。

1）大数据的采集

在“大数据时代”的今天，数据被提到一个前所未有的高度。无论是小企业还是大公司，网上销售还是线下营销都意识到了数据的重要性。随着大数据越来越被重视，数据采集的挑战也变得尤为突出。

很多人不清楚需要搜集什么样的数据，通过什么渠道来搜集数据，还有大部分不清楚搜集整理的数据如何去分析，进而也就不清楚怎么去利用这些数据。所以，很多数据也就仅仅只是数字，无法去转化和为公司利益服务，成为了摆设。

下面介绍三类将数据做成摆设的类型。

（1）重视数据但不清楚如何搜集，这是“被数据”类型，表现为对数据处于模糊了解状态，明确公司和企业做事和计划要靠数据来支撑。由于缺乏专业的相关数据人员，公司该做哪些数据，通过什么渠道来搜集整理，处于一知半解的状态，通过网上学习东拼西凑而成的数据自然就只是摆设了。

（2）了解所需数据但来源不规范，这是“误数据”类型，表现为对数据比较了解，大概明确需要什么数据。同样，由于缺乏专业的数据人员，对于数据的来源和制作并不规范，数据采集也可能存在误差。因此，采集的数据就可能失真，数据价值较小。

(3)会做数据但不会解读分析,这是“低估数据”类型,表现为对数据清楚了解,并有准确的数据来源和较明确的数据需求,但是坐拥金矿却不会利用,只是简单地搜集整理,把数据形成可视化的报表,这种简单化的采集方式使得数据的价值被低估。

了解数据背后的意义,解读数据来为公司和个人创造价值,利用数据来规避可能存在的风险,这些才是数据采集的意义。

数据的采集系统是基于计算机或测试平台的测量系统。常见的采集工具有很多,如麦克风、摄像头等,数据的采集技术应用广泛。

大数据的采集一般分为以下两个层次。

(1)大数据智能感知层:包括传感适配体系、网络通信系统、智能识别体系、数据传感体系和软硬件资源接入体系,用来完成对不同类型的数据结构的智能识别、清洗、接入、信号转换、监控、处理和管理等。

(2)大数据基础支撑层:是一种虚拟的服务器,是大数据服务平台所必需的,提供包含各种类型数据结构的数据库和物联网等支撑环境。

在大数据的采集过程中,现存难点是并发数高,也许存在无数的用户在同时访问同一个页面的情况,在并发数高峰期,访问量峰值高达百万次每分钟,必须在数据库之间进行负载均衡与分片,同时在采集端衔接大量数据库才能支撑。

2)大数据的导入与预处理

要实现对海量数据的有效分析,需要将数据导入集中的分布式数据库或分布式存储集群,之后需要对数据库进行简单的预处理和清洗。如果企业对业务有实时需求,可以在导入时使用 Storm 对数据进行流式计算。

3)大数据的统计与分析

随着技术的更新,大数据分析越来越多地在医疗、建设智慧城市等方面发挥了积极的作用。在商业应用方面,不少企业对大数据分析的需求上升。迫切需要引进专业的数据分析人员,或与大数据分析服务机构合作,以挖掘数据价值,为企业科学的运营决策做支撑。

运用好大数据的统计与分析技术,能协助企业精准定位客户喜好、优化资源配置、定制营销。目前,在发达国家,大数据分析行业已进入蓬勃发展期,专业的数据分析机构和数据分析人员的规模也不断扩大,大数据分析广泛应用于发达国家的各个行业,如电商、金融、零售、通信等领域。

大数据的统计与分析主要利用分布式计算集群或分布式数据库来对数据进行分类和汇总。在企业的实时性需求方面,可以用 Oracle 的 Exadata、EMC 的 Greenplum、基于 MySQL 的列式存储 Infobright 等。对于批处理或半结构化数据的需求,则可以使用 Hadoop。

4)大数据的挖掘

人们需要从海量的数据中发现有用的数据价值,进而将数据价值转化为决策依据,这需要一些合适的工具来进行这项工作,因此产生了大数据的挖掘。数据挖掘是一个新生的、动

态的领域，是人们从数据时代迈入信息时代必不可少的步骤。

人们每天都在搜索引擎进行查询，每天可达数亿次查询，如果人们的查询都被看作一个任务，人们通过关键词描述任务需求，那么日积月累，搜索引擎能从海量的查询中学到什么？这里有一个发现，在海量的查询中，有些查询模式能呈现出大量的知识，而这些知识却不能通过仅仅读取单个人的查询数据得到。例如，百度的飞行时间查询，使用这个搜索项作为航班飞行活动的指示，它呈现出搜索飞行时间相关信息的人数与正在候机的人数之间的联系。其与飞行时间相关的搜索都汇总在一起时，即产生了一种模式。使用这种汇聚的搜索数据，百度的飞行时间能比传统的系统早几个小时或对航班准点率做出评估。这样的实例表示，大数据的挖掘能把数据集转换成信息，帮助人们得到答案。与统计和分析过程相区别的是，大数据的挖掘通常没有预先设定的主题，而是在现有数据的基础上计算，来实现预测的结果，用于满足高级别的分析需求。常见的算法有 Kmeans（用于聚类）、SVM（用于统计）、NaiveBayes（用于分类）等。大数据的挖掘通常使用的算法以单线程为主，因其计算的数据量大。

2. 大数据技术的特征

大数据技术具有以下几个特征。

1）数据进行全面分析

大数据技术的数据分析是全面的，而不是随机抽样进行的。在大数据技术之前，因缺乏对全体样本进行抽取的技术，对待样本的抽取方式，都是从小样本中进行随机抽取。在理论上曾认为，随机抽取的样本，能代表整体样本的多样性，但这种方法费力且费时。在大数据出现后，在云计算和数据库的基础上，大数据技术能获取足够大的样本，并能存储至数据库中。所有的数据都存储在数据仓库中，因此不需要以随机抽样的方法对数据进行分析。获取大数据本身并不是人们最终的目的，如果能用小数据解决人们的疑惑，就不需要使用大数据进行分析。牛顿力学定律、行星定律等都是通过小数据分析发现的，人脑就是通过小数据学习来获取知识的。

2）强化数据的复杂性

对于小数据来说，收集的样本是有限的，因此需要尽可能使保存的数据精准。例如，采用抽样方法时，要求在运算时精准，在 1 万只羊中采取随机抽取的方式，抽取 100 只羊，如果在 100 只羊的样本上计算有误，放大至 1 万只羊，偏差就会扩大，而在这 100 只羊的样本上，产生的偏差是固定的，不会扩大。

小数据注重减少差错以保证质量，大数据更注重数据的复杂性。

在小数据的情况下，为了避免放大时造成的偏差，要求得到样本的精准计算结果，但需要耗费很多的时间，在大数据的情况下，样本等于总体，能迅速获得总体的特点和趋势，这比精准性更为重要。

大数据的算法简单，但比小数据有效，因此对大数据不必要求精准性。

3)重视数据的相关性

变量 A 与变量 B 有关联，变量 A 与变量 B 的变化存在一定的联系，表明两者具有相关性。相关性不代表因果关系，不能说变量 A 是变量 B 变化的原因。

例如，淘宝网运用它的大数据技术算法，根据消费者的历史购买记录或浏览记录来推送给该消费者可能喜欢的商品，这种算法并不能说明该消费者喜欢推送商品的原因，也不能说明消费者如果购买了 A 之后又购买了 B，购买 A 就是购买 B 的原因，只能说购买两者具有相关性，或存在一定的概率。大数据技术知道是“什么”，但不知道“为什么”，在大数据技术下，通过相关性查找数据比小数据时代更便捷、更迅速。

大数据系统依赖相关性，而非因果性，相关性表明发生的可能性，而不是发生的原因，通过大数据技术分析，查询到现象之间的关系，更快、更迅速，而且不易受到偏见的影响。建立起技术分析法的预测是大数据的内在要求。

4)算法复杂度高

大数据是一种综合交叉的科学，具有不同于一般统计学的计算方法，处理海量的数据需要更智能、更简单的操作方法和问题求解方式。因此，对于算法的要求更高，不仅仅是简单算法的集合，而是更复杂。

3. 大数据的关键问题和关键技术

1)大数据的关键问题

大数据的数据源来源广泛，且数据类型呈多样性，数据计算时，读取和分析的数据量大，要求数据服务具有高效性。

(1)半结构化和非结构化的数据处理。在大数据中，结构化数据只占 15%左右，其余的 85%左右都是半结构化和非结构化的数据，它们大量存在于互联网和电子商务等各个领域。如果把系统通过分析数据得到信息的过程称为一次挖掘过程，那么将得到的信息再结合人们的主观知识，如具体的经验、常识、本能、情境知识和用户偏好，而产生“智能知识”的过程称为二次挖掘。从一次挖掘到二次挖掘类似事物“量”到“质”的飞跃。

由于大数据所具有的半结构化和非结构化的特点，经过大数据的一次挖掘后的结构化的“粗糙知识”(潜在模式)产生出一些新的特征。一次挖掘后的结构化粗糙知识可以被主观知识加工处理并转化，生成半结构化和非结构化的智能知识。寻求智能知识是大数据研究的核心价值。

(2)大数据的系统建模与其复杂性。这一问题的突破是将大数据转化为知识的基础和重点。目前，由于大数据的数据个体复杂且随机，这种数据特征将促使大数据形成自己的数学结构，有利于建立并完善大数据的统一理论。现在，研究界倡导发展一种适应大数据交叉应用的、一般性的结构化数据和半结构化、非结构化数据之间的转化原则。管理学的理论将在实现这种一般性原则和建立大数据规律中发挥关键的作用。

实践中的大数据处理问题是非常复杂的，很难运用单一的计算模式，满足各种不同的大

数据计算需求。

大数据的复杂形式促使产生了很多对粗糙知识的量化和评估的相关研究。已知的最优化、数据包络分析、期望理论、管理科学中的效用理论等可以被应用到研究如何将主观知识与二次挖掘过程相融合。这里，人机交互将起到至关重要的作用。

(3)大数据的异构性与决策异构性影响知识发展。大数据本身的复杂性使得传统的数据挖掘理论和技术无法适应大数据的需求。在大数据条件下，管理决策迎来了挑战，即两个异构性问题：数据异构性和决策异构性。传统的管理决策基于对自身的知识构建和过往的业务经验，而数据分析又是管理决策的基础。

大数据改变了传统的管理决策结构的模式。决策结构的变化要求人们去探讨如何通过二次挖掘获得的知识去支撑管理决策。无论大数据带来哪种数据异构性，大数据中的粗糙知识仍可被看作一次挖掘的范畴。通过寻找二次挖掘产生的智能知识来作为数据异构性和决策异构性之间的桥梁是十分必要的。

大数据是具有隐秘规则的“人造森林”，获寻大数据的科学模式是人们的挑战也是机遇，如果人们找到了将非结构化、半结构化数据转化成结构化数据的规则，已知的数据挖掘方法将成为大数据挖掘的工具。

2)大数据的关键技术

大数据的关键性技术主要分为流处理、并行化、可视化和摘要索引四种。

(1)流处理。随着公司的业务处理流程越发复杂，流处理技术已成为大数据的重要处理技术，能满足实时的数据处理需求，随时产生数据流的架构，随时处理。

例如，在传统的方法中，只能计算已经给出具体数据的一组数据的平均值，如果数据是移动的，这样的平均值计算则需要大数据的流处理方法，即创建一个数据流统计集，逐步添加数据块，进行移动平均值计算。

(2)并行化。小数据的存储能力通常不到 10 GB，中数据的存储能力不到 1 TB，大数据的存储则是分布于多台机器上，存储能力多达 PB，在分布式数据条件下，需要在极短的时间内处理数据，需要并行化处理。

(3)可视化。数据可视化分为信息可视化和科学可视化两种。可视化工具是实现可视化的必要手段，常见的可视化工具有以下两类。

①管理决策者或数据分析师可以利用探索性可视化工具找出数据之间的关联性，这是可视化工具的洞察力作用，如 Tableau、TIBCO、QlikeView。

②叙事性可视化工具挖掘数据的方式较为独特。例如，需要用叙事性可视化工具查看某个时间段内某企业的营销数据，可视化格式将预先被创建，数据会按照时间点被逐年显示，并按照设定的条件排序。

(4)摘要索引。摘要索引是加速查询数据的预计算摘要的过程，这个预计算摘要会被预先创建。摘要索引的作用是为将要进行的查询做计划。现在摘要索引尚没有一个明确的规则，但随着大数据技术的发展，这一问题将会得到解决。

思考与练习

一、填空题

1. 数据的类型有很多种，主要分为三种，即_____、_____和_____。

2. 大数据的三个明显特征分别是：_____、_____和_____。

3. 魔镜现在有五个版本，即企业基础版、企业标准版、企业高级版、_____版和Hadoop 版。

二、简答题

1. 简述大数据的定义。

2. 大数据的价值表现在哪几个方面？

3. 大数据的特点有哪些？

4. 大数据的分析工具主要有哪些？