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（1EB＝2^{10}PB，1PB＝2^{10}TB，1TB＝2^{10}GB）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

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，极大地提高了决策实效性。

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、遗传物质的信息等进行分析的基础上，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指标的一半，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。

大数据因其背后的巨大价值，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着至关重要的作用。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,拥堵状况等数据分析,来合理设置红绿灯的时间,缓解交通拥堵。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府政务能力和社会管理能力方面发挥着重要作用,使之形成具有各地特色的、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°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,还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,人类说过的语言总和的数据

量大约是 5 EB。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、戴的运动手环等包含的数据信息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

数据主要分为结构化数据、半结构化数据与非结构化数据三种，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，即大数据具有实时性。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、数据挖掘等技术迅速发展的带动下，呈现出它的完整过程：把数据源的信号转换为数据，再把数据加工成信息，通过获取的信息做决策。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到有价值的信息，为决策提供支撑。

1.1.2　大数据的产生

早在 3 000 多年前的埃及，人类就利用计数来统计、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。步入农业社会的苏美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。

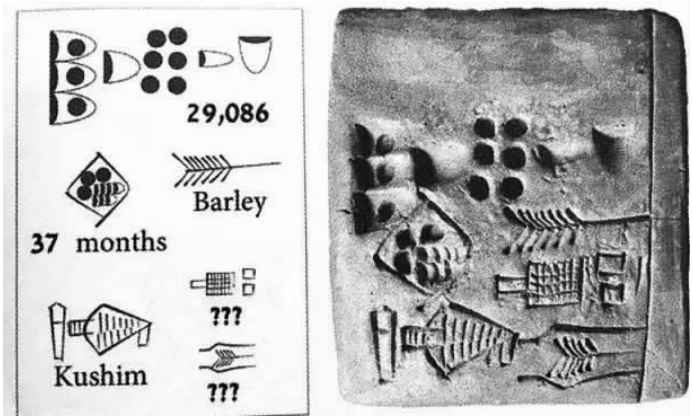

图 1-2　苏美尔人用于统计的符号

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。

1.1.3　大数据分析工具

1. Smartbi

Smartbi 是由广州思迈特软件有限公司生产的核心产品。“思迈特商业智能与大数据分

析软件”是企业级商业智能和大数据分析平台，可以满足用户在企业级报表、数据可视化分析、自助探索分析、数据挖掘建模、AI 智能分析等大数据分析领域的需求。Smartbi 作为国产的商业智能与大数据分析产品，针对国内用户普遍的本土性需求有更好的设计弹性和适应性，能够更好地服务国内的数据分析用户。Smartbi 具有一站式数据服务，全面系统运维保障，超大数据量梳理，一体化数据建模能力，助力企业构建数据文化，领先的增强分析能力的特点。

Smartbi 为了更好地满足所有用户的各种数据分析应用需求，将产品分为专业版（Professional）、企业版（Enterprise）、旗舰版（Eagle）和嵌入版（Embedded）。

- 专业版：面向有深度实施需求的用户。
- 企业版：面向需升级商业智能（business intelligence，BI）工具，匹配其数据平台建设的用户。
- 旗舰版：推行“数据文化”，通过强管控、全自助和真共享实现企业级自助数据门户，满足用户管理协同和社交协同的需求，面向需全面数据化运营和决策的用户。
- 嵌入版：提供二次开发接口嵌入 BI，面向 ISV 厂商。

2. Apache Drill

为了帮助企业用户寻找更为有效、加快 Hadoop 数据查询的方法，Apache 软件基金会发起了一项名为“Drill”的开源项目。Drill 将有助于 Hadoop 用户实现更快查询海量数据集的目的。Apache Drill 是一个引擎，可以连接到许多不同的数据源，并为它们提供 SQL 接口。它不仅是遍历任何复杂事物 SQL 的界面，而且是功能强大的界面，其中包括对许多内置函数和窗口函数的支持。Apache Drill 在基于 SQL 的数据分析和商业智能上引入了 JSON 文件模型，这使得用户能查询固定架构，演化架构，以及各种格式和数据存储中的模式无关（schema-free）数据。该体系架构中关系查询引擎和数据库的构建是有先决条件的，即假设所有数据都有一个简单的静态架构。

Apache Drill 的架构是独一无二的。它是唯一一个支持复杂和无模式数据的柱状执行引擎（columnar execution engine），也是唯一一个能在查询执行期间进行数据驱动查询（和重新编译，也称之为 schema discovery）的执行引擎。这些独一无二的性能使得 Apache Drill 在 JSON 文件模式下能实现记录断点性能（record－breaking performance）。

3. Tableau

Tableau 公司，是由斯坦福大学的三位校友 Christian Chabot（首席执行官）、Chris Stole（开发总监）以及 Pat Hanrahan（首席科学家）于 2003 年在远离硅谷的西雅图注册成立的。Tableau 是一款免费的数据可视化工具，具有高度的灵活性和动态性，可以制作图表、图形，绘制地图；不仅支持个人使用，还允许团队协作同步完成绘制；操作简单，用户可以直接将数据拖动到系统中进行操作。

Tableau 简单、易用、快速，一方面归功于产生自斯坦福大学的突破性技术。Tableau 是集复杂的计算机图形学、人机交互和高性能的数据库系统于一身的跨领域技术，其中最耀眼的莫过于 VizQL 可视化查询语言和混合数据架构。另一方面在于 Tableau 专注于处理最简

单的结构化数据，即已整理好的数据——Excel、数据库等，结构化数据处理在技术上难度较低，这就使得 Tableau 有精力在快速、简单和可视化上做出更多改进。Tableau 包含 Tableau Desktop，Tableau Online，Tableau Server，Tableau Mobile，Tableau Public，Tableau Reader 等产品。

1.2 扩展知识

1.2.1 大数据的应用

1. 大数据经典案例

1）医疗健康

医疗健康大数据的应用为医疗行业带来了宝贵的价值。实际上，大数据的一些应用已经有效地提高了私营和公共医疗服务，更好地帮助患者摆脱病患和协助医生做出准确的诊断。大数据分析可以通过提供决策支持工具，降低医疗行业的高成本来支持运营服务的优化，从而彻底改变传统的医疗模式。以下是医疗领域中一些具体的大数据应用。

（1）大数据分析帮助健康机构检测哪些部门需要被重新配备，能够有助于实时评估和监测服务质量、医疗单位的绩效以及人力资源和医疗设备的需求，从而提供更好的医疗健康服务，同时减少医院不必要的开支。

（2）使医生和患者更好地了解并掌握疾病演变，支持医生的决策。如大量的病毒和 DNA 的信息来源通过数据分析有助于了解疾病演变，这些数据分析有助于医生和研究人员找到预防遗传和遗传性疾病的新方法，从而进一步帮助医生有效地诊断患者的病况。然后，将患者的历史手术结果与患者当前的症状或历史记录进行分析，这样的相互关系有助于根据患者资料找到最合适的干预措施和治疗方法，从而支持医生的决策。

（3）提供医疗服务的用户化。一些医疗项目实时收集和分析患者的反馈意见，以提高他们的满意度。例如，实时医疗数据可以监测病人的病情，以适应药物剂量或根据分析的症状给出医疗建议。一些项目使用智能传感器连接到智能手机或血糖仪，目标是在线监测和实时监测患者的症状（血糖水平、心脏跳动等）。如果有紧急情况和症状信息会立即发送给医生，以便根据患者的新症状调整医疗方案。一般来说，医疗数据分析可以提高患者的生活质量，同时为医生提供有价值的治疗和手术方面的信息。

（4）预测性大数据模型可以分析来自私营和公共医院的临床数据，从而预测疾病的情况，防止流行病蔓延。这些模型根据受影响的地区和人口症状能够检测出与人口健康有关的严重症状，决策者能够通过这些检测结果建立有效的预防计划，并阻止流行病蔓延。

2019 年 12 月，湖北省武汉市出现新型冠状病毒肺炎（简称“新冠肺炎”），随后疫情在全国范围内暴发。相关学者对此次疫情扩散趋势做了大量研究，但基于模型的估算普遍存在高估传染系数和感染人群的问题。基于此，利用大数据回溯新冠肺炎在全国扩散的趋势和

传染系数，从数据上论证了中国政府对于疫情扩散强有力的控制能力。基于大数据开发的软件程序可以精确查找确诊、疑似患者所乘坐的车次，以及与确诊或疑似患者的距离，由此可对潜在感染人员进行排查和隔离，对疫情的防控和排查起到关键性作用。为了防止春节后人员流动对于病毒扩散的影响，我国政府宣布延长春节假期。我国政府强有力的执行能力最大限度地减少了新冠肺炎的扩散，对阻止国内及世界范围的疫情扩散做出积极贡献。

2）金融行业

自从有了大数据，金融服务行业便迅速发展信息体系结构，其中，访问、分析和管理海量数据的能力对提高业务效率和性能至关重要。大数据的出现，使得银行的盈利能力一直在上升，特别是在世界各地经济条件好的地方，银行通过进入新的市场和服务领域来开发新的收入来源。随着客户数量的增加，这会显著影响组织提供的服务水平。现有的数据分析实践简化了银行和其他金融服务机构的监督和评估流程，包括大量客户数据，如个人和安全信息。但是在大数据的帮助下，银行现在可以使用这些信息来实时跟踪客户行为，提供任何特定时刻所需的确切资源类型。这种实时评估反过来会提升整体绩效和盈利能力，从而推动组织进一步进入成长周期。

利用大数据技术提高客户在商业银行业务方面的经验，可以帮助其增加以利息为基础的收费。许多较大型的金融机构都倾向于扩大理财投资组合，以确保风险较低且收费一致。差异化的服务，交叉销售和向上销售的举措，以及扩展到全球新兴的财富管理市场正在上升。大数据分析和用户信息管理在确保分析策略得到正确执行方面起着核心作用。

金融服务机构将继续通过更高的运营效率，更好的风险管理以及改善的客户亲密度来关注收入增长和更高的利润率。这样的知识使得企业能够适应和增强他们的产品、服务和策略（如实时的有针对性的广告宣传）。因此，可以增加顾客的满意度，扩大利润，增强竞争力。例如，Facebook、Google、Amazon 收集和出售有关网络用户行为、反馈、评论和在线交易的信息。信用卡公司（如 Equifax 和 TransUnion）也是这样做的，以增加利润，并提高他们的服务质量。此外，多种通信技术的迅猛发展以及众多实体（如企业子公司，合作伙伴，供应商和在线客户）之间的高度互联互通，带来了基于大数据实时共享和货币化的新商业模式。

实际上，银行和其他金融机构可以从大数据高级分析中获得三个主要方面的优化：客户体验的优化，操作运营的优化以及员工敬业度的优化。

（1）客户体验优化。关注客户的需求是非常重要的，因为如今的客户对他们与银行或信用社的互动方式抱有很高的期望。他们的购买旅程非常复杂且非线性，因此金融机构必须能够仔细了解客户的偏好和动机。为了实现客户的 360°视图，一系列客户快照已经不够了。公司需要一个中央数据中心，将客户与品牌的所有交互结合在一起，包括基本的个人数据、交易历史、浏览历史记录、使用服务等。根据麦肯锡公司的说法，使用数据做出更好的营销决策可以将营销生产力提高 15%～20%，考虑到平均每年 1 万亿美元的全球营销支出，这个数字可能高达 2000 亿美元。以数据为基础的分析可以帮助金融行业了解客户并创建客户细分。这种信息收集和评估需要对组织基础设施进行额外投资，并通过跨组织使多个职能部门人员之间的投入和协调一致。

（2）操作运营优化。虽然大数据已经在金融的很多领域得到了应用，但除了一些早期的

采用者之外，风险管理还没有打开它的力量。大数据技术可以提高风险模型的预测能力，通过提供更多的自动化流程，更精确的预测系统以及更少的失败风险，以指数方式提高系统响应时间和有效性，提供更广泛的风险覆盖范围，并显著节约成本。风险团队几乎可以实时从各种来源获得更准确的风险情报。大数据在金融风险管理方面的很多领域都可以应用和带来价值，包括欺诈管理、信用管理、市场和商业贷款、操作风险和综合风险管理等方面。例如，启用大数据的系统可以检测欺诈信号，使用机器学习实时分析这些信号，并准确预测非法用户的交易。大数据提供了与财务风险相关的不同因素和领域的全球视野的能力。

(3)员工敬业度优化。对于大数据受到的所有关注，许多公司倾向于忘记一个潜在的因素，可能会对他们的业务产生巨大影响，这种因素就是员工体验。如果做得对，它可以帮助追踪、分析和分享员工绩效指标。将大数据分析应用于员工绩效有助于识别并确认绩效最好的员工，也可以认识到挣扎或不快乐的员工。这些工具允许公司查看实时数据，而不仅是基于人类记忆的年度评论。当拥有正确的工具和分析时，可以衡量一切，包括个人表现、团队精神、部门之间的互动以及整个公司的文化。当数据与客户指标相关时，也可以使员工花更少的时间在手动流程上，而更专注于更高级的任务。

3)其他应用领域

零售企业收集的数据量(如大数据量级从 TB 上升到 ZB，数据的维度也在运营数据、交易数据、用户数据的基础上，增加了交互数据直到大数据)继续迅速增长，特别是由于在网上或电子商务上进行的业务的易用性、可用性和普及程度日益提高。通过收集到的大量有关销售和客户购物历史的数据，零售数据挖掘有助于识别顾客行为，发现顾客购物模式和趋势，提高顾客服务质量，获得更好的顾客忠诚度和满意度，提高商品消费率，从而可能分析设计出更有效的货物运输和分销政策，降低商业成本。

为了加强大数据领域的研究和开发，一些国家的政府已经在实时分析多种动态或静态信息的来源(例如，日志、历史事件、公共和私人监控摄像机、社交网络上的公民评论、在线交易、GPS 数据和移动通信)。他们也利用了许多政府信息通信技术的数据，目标是发现有价值的数据信息、模式和相关性，或者建立预测模型，使政府能够优化战略，增强公民的公共服务；另一个重要的目标是确保连续的监督和监测，以保护公民和减轻犯罪的影响。

大数据智能交通系统的出现改善了城市交通管理，为智能交通的发展提供了新趋势。智能交通系统通过收集实时交通数据，可以识别当前的交通运行状况、交通流状况，并可以预测未来的交通流量，然后发布一些最新的实时交通信息，帮助驾驶者选择最佳路线，能够做到对移动车辆进行精确的管理、监控。同时，智能交通系统还具有改善交通条件，减少交通拥挤和管理费用，高可靠性，提高交通安全和不受天气条件影响等优点。

在互联网与电子商务行业中，大数据和相关技术对传统的网络发展带来巨大影响。例如，通过收集互联网用户的地理分布数据、搜索短语实时数据、购物浏览行为数据以及兴趣爱好社交数据等不同的互联网用户数据，就可以实现地理定位，通过用户个性化需求导向、个性偏好导向和关系导向等方式，实现精准化、个性化的网络营销。

在旅游业中，已经有一些大数据旅游模型，这些模型改进了旅游活动，更好地为旅游者提供服务。例如，更好地了解游客的行为，发现其偏好和需求，监测游客的地理位置、活动和

背景。同时，可以根据游客的偏好、在线行为和地理位置向游客推荐实时的酒店、餐馆和活动。一个旅游推荐系统就是基于广泛的大数据分析和可视化工具的结合，其中包括：对旅游活动历史模式的分析；实时分析当前旅游活动、偏好、配置文件和网站访问情况；跟踪旅游地理位置；监测其他参数，如天气状况和交通拥堵情况，以此来实时建立个性化推荐。

水利资源中，自动化的传感器和监测系统提供大量的实时流量数据。例如，灌溉系统中的自动化传感器在分秒中产生各种有关气候（温度、辐射、风速和湿度）、作物（作物高度、植物密度、叶面积指数等）和土壤（含水量、渗透等）的数据和其他可能在多个小时才能产生的数据。这些数据可以被存储和分析，以调节自动化灌溉水源的开启或关闭。传感器产生的数据需要实时处理，以便立即采取行动。然而，使用实时数据开发和验证模型是一项艰巨的任务。

2. 大数据的操作实例

当今，数据已经渗透到每一个行业和业务领域，成为重要的生产因素。人们对于海量数据的挖掘和运用，预示着新一波生产率增长和消费者盈余浪潮的到来。中国的保险销售模式正在酝酿新的变革，互联网、大数据时代的到来给金融业造成的革命性、颠覆性的变化正在发酵，对保险业数据驾驭能力提出了新的挑战，也为保险业的大发展提供了前所未有的空间和潜力。

1）保险业深入挖掘大数据应用潜质

目前，大多数保险企业都已经认识到大数据改善决策流程和业务成效的潜能，但却不知道该如何入手，部分企业在大数据的时代浪潮下积极探索，成为先行者。2010 年，阳光保险集团建成数据挖掘系统，这在保险行业是第一家。利用该系统，阳光保险集团开展了许多保险大数据智慧应用的项目，获得了一些成果，同时培养出了国内保险行业的第一批数据挖掘师。

大数据应用的关键是理念。思维转变过来，数据就能被巧妙地用来激发新产品和新型服务。举一个利用与不利用数据，结果相去甚远的例子：淘宝现有一种运费保险，即淘宝买家退货时产生的退货运费原本由买家承担，如果买家购买了运费保险，退货运费则由保险公司来承担。这种购买的结果是保险公司经营亏损很严重，直接导致它们不愿意再发展和扩大运费保险。运费保险真的必然亏损吗？答案是“否”。

保险公司设计了一套大数据智慧应用的解决方案：因为退货发生的概率，跟买家的习惯、卖家的习惯、商品的品种、商品的价值和淘宝的促销活动等都有关系，所以，使用以上种种数据，应用数据挖掘的方法，建立退货发生的概率模型，植入系统就可以在每一笔交易发生时，给出不同的保险费率，使运费保险的收取与退货发生的概率相匹配，这样运费保险就不会亏损了。

在此基础上，保险公司才有可能通过运费保险扩大客户覆盖面。由严重亏损到成本控制得当并获取客户，靠的就是通过分析，挖掘大数据所提供的价值，吸引客户。

2）大数据网络保险时代来临

大数据发展的障碍，在于数据的流动性和可获取性，而网络完美地解决了这个问题。通

过网络对大数据进行收集、发布、分析、预测会使决策更为精准，释放更多数据的隐藏价值。与传统保险方式相比，网络保险具有降低保险公司和保险中介机构运营成本，拓展保险公司和保险中介机构业务范围，新型营销手段，有价值的交互式交流工具，提供较高水平的信息服务，为客户提供便捷工具，使客户享受个性化服务，降低保险公司风险，更有效地保护客户隐私以及虚拟化的交易方式等特性。

从产品设计角度来说，大数据时代下的网络保险能最大程度地满足不同客户的个性化需求，网络保险能优化客户的体验，大数据能根据客户需求设计出真正让客户满意的产品和服务，两者结合则完全是以客户为中心的。

从大数据时代的网络销售优势来看，一是大数据时代保险网销具有最广泛的客户群，有最大的发展潜力。二是互联网具有信息量大，传播速度快，透明度高的特点，交易双方信息更为对称。通过建立新型的自动式网络服务系统，客户足不出户就可以方便快捷地从保险公司的服务系统上获取公司背景到具体保险产品的详细情况，还可以自由地选择所需要的保险公司及险种，并进行对比，获得低价、高效服务。三是节省费用，降低成本。通过网络销售保险或提供服务，保险公司只需支付低廉的网络服务费，从而降低房租、佣金、薪资、印刷费、交通费和通信费等成本的支出。四是数据管理方面的天然优势。保险市场专业化的深入，经营水平的提高，服务品质的提升，都要建立在对数据尤其对客户消费数据的深入挖掘和分析的基础之上。

可见，大数据时代下的网络保险有利于推动营销体制改革。多年来，我国一直以保险代理人作为保险推销体系的主体重点发展，在寿险推销方面形成了以寿险营销员为主体的寿险营销体系。但是，目前这种体制还存在较为突出的问题。因客户缺乏与保险公司的直接交流，会导致营销人员为急于获取保单而一味夸大投保的益处，隐瞒不足之处，给保险公司带来极大的道德风险，为保险业的长远发展埋下隐患。而且，保险营销人员素质良莠不齐，又会给保险公司带来极大的业务风险。此外，现有营销机制还存在效率低下的弊端。

因此，在大数据时代下发展网络保险，可以快速便捷地进行信息收集、发布，完美地实现大数法则的精致应用，为公众提供低成本，高效率的保险服务。

3）网络保险需多项配套支持

（1）财政支持。在推进保险公司的信息化进程中，政府可采取诸如信息技术方面的投资部分抵消税收，税前可以预留部分资金用于信息技术改造等一系列措施，激励和推进大数据网络保险信息化进程。

（2）培育网络保险集市。网络保险集市就是在网络上提供一个场所，使客户能在这里找到大量的保险公司，方便了解各个公司的基本信息或查询各个保险公司的某一险种的有关信息，并对该险种的优劣进行对比分析，选择最佳的公司进行投保。网络保险集市不仅会给客户带来方便，同时也会扩大保险公司的影响和业务量。因此，保险公司应在保监会和保险协会的组织下，全力支持并在网络保险集市上展示自己，进一步推动我国网络保险集市的发展。

（3）建设大数据中心。大数据中心需要保监会和保险行业进行战略性的顶层设计。首先是与我国标准化数据管理中心进行合作，制定出保险业数据标准化的制度。其次是通过5～10年的时间逐步完成行业数据标准化建设。同时设计出非线性能融合关系数据，并能进一步扩展的数据

库。然后是设计柔性的框架和接口。通过以上步骤逐步完成我国保险业大数据中心的建设。

(4)开发适合的险种。利用网络收集数据形成大数据，利用大数法则设计客户需求的产品，通过网络销售产品，并根据客户反馈进一步修正产品，实现开发与销售完美互动。

(5)吸纳优秀人才和对已有员工进行在职教育。许多保险公司有一个规定，即无论是管理人员还是技术人员都必须完成一定的保险任务。似乎这条规定能为公司增加一点业务量，但是它无形之中就把一些优秀的保险管理人员和技术人员拒之门外。大数据时代需要一流的管理人才和技术人才，必须破除这条不成文的规定。同时还应该重视对已有员工进行保险专业知识、外语知识和信息技术知识再教育，通过再教育提高公司员工综合素质。

(6)责任与自由并举的信息管理。调查显示，66%的被调查者最关心投保后支付保费的转账安全性。消费者对于网络消费的顾虑心理主要集中在对网上交易安全和个人隐私保护的担忧上。因此，网络保险应格外注重网络安全，实现责任与自由的矛盾的和谐统一。

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，无论是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，还有大部分人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。公司和企业明确做事和计划要靠数据来支撑，但由于缺乏专业的相关数据人员，公司对该做哪些数据，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进行支撑。

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。大数据的挖掘是一个新生的、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。

人们每天都在用搜索引擎进行查询，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 K-means（用于聚类）、SVM（用于统计）、Naive Bayes（用于分类）等。大数据的挖掘因其计算的数据量大，通常使用的算法以单线程为主。

2. 大数据技术的特征

大数据技术具有以下几个特征。

1）对数据进行全面分析

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，并能将其存储至数据库中。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,而是更复杂的算法。

3.大数据的关键问题和关键性技术

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,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从“量变”到“质质”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性的作用。

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。

大数据的复杂形式催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）大数据的关键性技术

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 级。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. 大数据的特征分别是_____、_____、_____和_____。

2. 大数据的处理过程有_____、_____、_____和_____。

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. 大数据的分析工具主要有哪些？

模块 2

Hadoop 基础

学习要点

- Hadoop 与大数据的关系。
- HDFS 原理。
- Hadoop MapReduce 原理。
- Hadoop 的应用。

2.1 必 备 知 识

2.1.1 Hadoop 概述

Hadoop 是 Apache 软件基金会旗下的一个开源分布式计算平台。以 Hadoop 分布式文件系统(Hadoop distributed file system,HDFS)和 MapReduce(Google MapReduce 的开源实现)为核心的 Hadoop 为用户提供了系统底层细节透明的分布式基础架构。HDFS 具有高容错性、高伸缩性等优点,允许用户将 Hadoop 部署在价格低廉的硬件上,形成分布式系统,为海量的数据提供了存储方法;MapReduce 分布式编程模型允许用户在不了解分布式系统底层细节的情况下开发并行应用程序,为海量的数据提供了计算方法。所以,用户可以利用 Hadoop 轻松地组织计算机资源,从而搭建自己的分布式计算平台,并可以充分利用集群的计算和存储能力,完成海量数据的处理。经过业界和学术界长达 10 年的研究和开发,

目前的 Hadoop 1.0.1 已经趋于完善，在实际的数据处理和分析任务中担当着不可替代的角色。

Hadoop 本质上起源于 Google 的集群系统，Google 的数据中心使用廉价 Linux PC 组成集群，运行各种应用，即使是分布式开发新手也可以迅速使用 Google 的基础设施。如今广义的 Hadoop 已经包括 Hadoop 本身和基于 Hadoop 的开源项目，并已经形成了完备的 Hadoop 生态链系统。

Hadoop 有以下几个特点。

1. Hadoop 是一个框架

很多初学者在学习 Hadoop 时，对 Hadoop 的本质并不十分了解，Hadoop 其实是由一系列的软件库组成的框架。这些软件库也称为功能模块，它们各自负责 Hadoop 的一部分功能，其中最主要的是 Common、HDFS 和 YARN。Common 提供远程调用 RPC、序列化机制，HDFS 负责数据的存储，YARN 则负责统一资源调度和管理等。

从字面上来说，Hadoop 没有任何实际的意义。Hadoop 不是缩写，而是一个虚构的名字。Hadoop 的创建者 Doug Cutting 这样解释 Hadoop 这一名称的来历："这个名字是我的孩子给一头吃饱了的棕黄色大象取的。我的命名标准是简短，容易发音和拼写，没有太多含义，并且不会被用于别处，小孩子是这方面的高手。"Hadoop 的 Logo 如图 2-1 所示，欢快的棕黄色大象如今已深入人心。

图 2-1　Hadoop 的 Logo

2. Hadoop 适合处理大规模数据

这是 Hadoop 一个非常重要的优点，Hadoop 处理海量数据的能力十分可观，并能够实现分布式存储和分布式计算，有统一的资源管理和调度平台，扩展能力十分优秀。2008 年，Hadoop 打破 297 秒的世界纪录，成为最快的 TB 级数据排序系统，用时仅 209 秒。

3. Hadoop 被部署在一个集群上

承载 Hadoop 的物理实体是一个物理的集群。集群指一组通过网络互联的计算机，集群里的每一台计算机称为一个节点。Hadoop 被部署在集群之上，对外提供服务。当节点数量足够多时，故障将成为一种常态而不是异常现象，Hadoop 在设计之初就将故障的发生作为常态进行考虑，数据的灾备及应用的容错对于用户来说都是透明的，用户得到的只是一个提供高可用服务的集群。

2.1.2 Hadoop 的发展史

Hadoop 原本来自于 Google 一款名为 MapReduce 的编程模型包。Google 的 MapReduce 框架可以把一个应用程序分解为许多并行计算指令，跨大量的计算节点运行巨大的数据集。使用该框架的一个典型例子就是在网络数据上运行的搜索算法。Hadoop 最初只与网页索引有关，后来迅速发展成为分析大数据的领先平台。

Hadoop 的源头是 Apache Nutch，该项目始于 2002 年，是 Apache Lucene 的子项目之一。Nutch 的设计目标是构建一个大型的全网搜索引擎，包括网页抓取、索引、查询等功能，但随着抓取网页数量的增加，其遇到了严重的可扩展性问题，不能解决数十亿网页的存储和索引问题。之后，Google 发表的两篇论文为该问题提供了可行的解决方案。一篇是 2003 年发表的关于 Google 分布式文件系统（GFS）的论文，该论文描述了 Google 搜索引擎网页相关数据的存储架构，该结构可以解决 Nutch 遇到的网页抓取和索引过程中超大文件存储需求的问题。由于 Google 未开放源代码，Nutch 项目组便根据论文完成了一个开源实现（Nutch 的分布式文件系统 NDFS）。另一篇是 2004 年 Google 在“操作系统设计与实现”（Operating System Design and Implementation，OSDI）会议上公开发表的 *MapReduce: Simplified Data Processing on Large Clusters*（《MapReduce：简化大规模集群上的数据处理》）论文，该论文描述了 Google 内部最重要的分布式计算框架 MapReduce 的设计方法，该框架可用于处理海量网页的索引问题。之后，受到启发的 Doug Cutting 等人开始尝试实现 MapReduce 计算框架，并将它与 NDFS（nutch distributed file system）结合，用于支持 Nutch 引擎的主要算法。由于 NDFS 和 MapReduce 在 Nutch 引擎中有着良好的应用，所以它们于 2006 年 2 月被分离出来，成为一套完整而独立的软件，并命名为 Hadoop。到 2008 年年初，Hadoop 已成为 Apache 的顶级项目，包含众多子项目。现在的 Hadoop 1.0.1 版本已经发展成为包含 HDFS、MapReduce 子项目，与 Pig、ZooKeeper、Hive、HBase 等项目相关的大型应用工程。

2.1.3 Hadoop 的优势

Hadoop 是一个能够让用户轻松架构和使用的分布式计算平台。用户可以轻松地在 Hadoop 上开发、运行处理海量数据的应用程序。它主要有以下几个优点。

1. 高可靠性

Hadoop 按位存储和处理数据的能力值得人们信赖。

2. 高扩展性

Hadoop 是在可用的计算机集簇间分配数据完成计算任务的，这些集簇可以方便地扩展到数以千计的节点中。

3. 高效性

Hadoop 能够在节点之间动态地移动数据，以保证各个节点的动态平衡，因此其处理速度非常快。

4. 高容错性

Hadoop 能够自动保存数据的多份副本，并能够自动将失败的任务重新分配。Hadoop 带有用 Java 语言编写的框架，因此运行在 Linux 生产平台上是非常理想的，Hadoop 的应用程序也可以使用其他语言编写，如 C++等。

2.1.4 HDFS 的原理

1. HDFS 简介

HDFS 是基于流数据模式访问和处理超大文件的需求而开发的，是一个分布式文件系统。它是 Google 的 GFS 提出之后出现的另外一种文件系统。它有一定高度的容错性，且提供给了高吞吐量的数据访问，非常适合应用在大规模数据集上。

(1)HDFS 的优点。

①处理超大文件。超大文件通常是指百 MB、甚至数百 MB 大小的文件。但是，目前在实际应用中，HDFS 已经能用来存储管理 PB 级的数据了。在雅虎，Hadoop 集群也已经扩展到了 4 000 个节点。

②流式数据访问。HDFS 的设计建立在“一次写入、多次读写”任务的基础上。这意味着一个数据集一旦由数据源生成，就会被复制分发到不同的存储节点中，然后响应各种各样的数据分析任务请求。在多数情况下，分析任务都会涉及数据集中的大部分数据，对 HDFS 来说，请求读取整个数据集要比读取一条记录更加高效。

③运行于廉价的商用机器集群上。Hadoop 设计对应急需求比较低，只需运行在低廉的商用硬件集群上，而无需运行在昂贵的高可用性机器上。廉价的商用机也就意味着大型集群中出现节点故障情况的概率非常高。遇到了上述故障时，HDFS 被设计成能够继续运行且不让用户察觉到明显的中断。

(2)HDFS 的缺点。正是出于以上种种考虑，我们会发现，HDFS 在处理一些特定问题时不但没有优势，反而存在很多局限性。它的局限性及应对策略如下。

①不适合低延迟数据访问。如果要处理一些用户要求时间比较短的低延迟应用请求，则 HDFS 不适合。HDFS 是为了处理大型数据集分析任务，主要是为达到高的数据吞吐量而设计的，这就要求以高延迟作为代价。

改进策略：对于那些有低延迟要求的应用程序，HBase 是一个更好的选择，通过上层数据管理项目尽可能地弥补这个不足，在性能上有了很大的提升，它的口号是 goes real time。使用缓存或多个 Master 设计可以降低 Client 的数据请求压力，以减少延迟。

②无法高效存储大量的小文件。小文件是指文件大小小于 HDFS 上 Block 大小的文件。小文件会给 Hadoop 的扩展性和性能带来严重问题。当 Hadoop 处理很多小文件时，由于 FileInputFormat 不会对小文件进行划分，因而每一个小文件都会被当作一个 Split 并分配一个 Map 任务，导致效率低下。

例如，1 个 1 GB 的文件，会被划分成 16 个 64 MB 的 Split，并分配 16 个 Map 任务处理，而 10 000 个 100 KB 的文件会被 10 000 个 Map 任务处理。

改进策略：要想让 HDFS 能处理好小文件，有不少方法。例如，利用 SequenceFile、MapFile、Har 等方式归档小文件，这个方法的原理就是把小文件归档起来管理，HBase 就是基于此的。

③不支持多用户写入及任意修改文件。在 HDFS 的一个文件中只有一个写入者，且写操作只能在文件末尾完成，即只能执行追加操作，目前 HDFS 还不支持多个用户对同一文件的写操作及在文件任意位置进行修改。

2. HDFS 架构

一个完整的 HDFS 运行在一些节点之上，这些节点运行着不同类型的守护进程，如 NameNode、DataNode、Secondary NameNode 等，不同类型的节点相互配合，相互协作，在集群中扮演了不同的角色，一起构成了 HDFS。

如图 2-2 所示，在一个典型的 HDFS 架构中，有一个 NameNode、一个 Secondary NameNode 和至少一个 DataNode，而 HDFS 客户端数量并没有限制。所有的数据均存放在运行 DataNode 进程的节点的块(block)里。

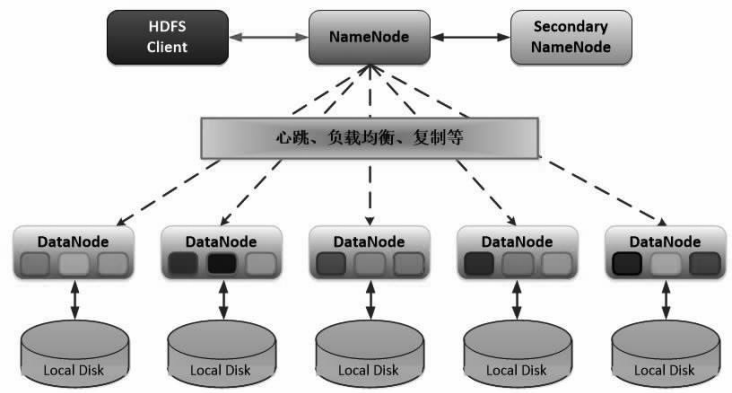

图 2-2 HDFS 架构

(1)HDFS Client(HDFS 客户端)。HDFS 客户端是指用户和 HDFS 交互的手段，HDFS 提供了非常多的客户端，包括命令行接口、Java API、Thrift 接口、C 语言库、用户空间文件系统等，模块 3 中将详细介绍这部分内容。

(2)NameNode(元数据节点)。元数据节点 NameNode 是管理者，一个 Hadoop 集群只有一个 NameNode 节点，是一个通常在 HDFS 实例中的单独机器上运行的软件。NameNode 主要负责 HDFS 文件系统的管理工作，具体包括命名空间管理(namespace)和文件块管理。NameNode 决定是否将文件映射到 DataNode 的复制块上。对于最常见的 3 个复制块，第一个复制块存储在同一个机架的不同节点上，最后一个复制块存储在不同机架的某个节点上。

NameNode 是 HDFS 的大脑，它维护着整个文件系统的目录树及目录树里所有的文件和目录，这些信息以两种文件存储在本地文件中：一种是命名空间镜像，也称为文件系统镜像(file system image，FSImage)，即 HDFS 元数据的完整快照，每次 NameNode 启动时，默

认会加载最新的命名空间镜像；另一种是命名空间镜像的编辑日志(edit log)。

(3)Secondary NameNode(第二名字节点)。第二名字节点是用于定期合并命名空间镜像和命名空间镜像的编辑日志的辅助守护进程。每个 HDFS 集群都有一个 Secondary NameNode，在生产环境下，一般 Secondary NameNode 也会单独运行在一台服务器上。

FSImage 文件其实是文件系统元数据的一个永久性检查点，但并非每一个写操作都会更新这个文件，因为 FSImage 是一个大型文件，如果频繁地执行写操作，会使系统运行极为缓慢。解决方案是：NameNode 只改动内容预写日志(WAL)，即写入命名空间镜像的编辑日志；随着时间的推移，编辑日志会变得越来越大，那么一旦发生故障，将会花费非常多的时间来回滚操作，所以就像传统的关系型数据库一样，需要定期地合并 FSImage 文件和编辑日志。如果由 NameNode 来进行合并操作，那么 NameNode 在为集群提供服务时可能无法提供足够的资源，为了彻底解决这一问题，Secondary NameNode 应运而生。NameNode 和 Secondary NameNode 的交互如图 2-3 所示。

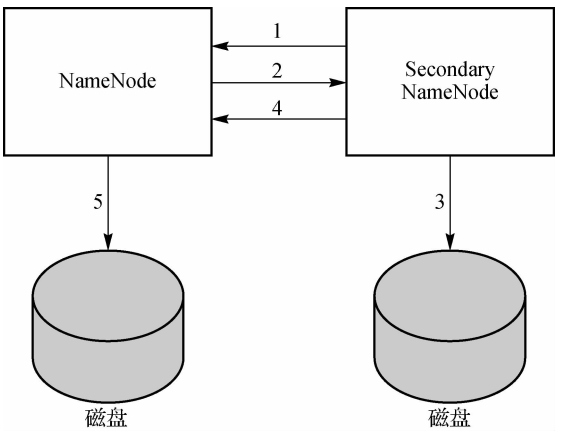

图 2-3　NameNode 与 Secondary NameNode 的交互

①Secondary NameNode 引导 NameNode 滚动更新编辑日志文件，并开始将新的内容写入 EditLog. new。

②Secondary NameNode 将 NameNode 的 FSImage 和编辑日志文件复制到本地的检查点目录。

③Secondary NameNode 载入 FSImage 文件，回放编辑日志，将其合并到 FSImage，将新的 FSImage 文件压缩后写入磁盘。

④Secondary NameNode 将新的 FSImage 文件送回 NameNode，NameNode 在接收新的 FSImage 后，直接加载并应用该文件。

⑤NameNode 将 EditLog. new 更名为 EditLog。

默认情况下，该过程每小时发生一次，当 NameNode 的编辑日志文件达到默认的 64 MB 时也会被触发。

从名称上来看，初学者会以为当 NameNode 出现故障时，Secondary NameNode 会自动成为新的 NameNode，也就是 NameNode 的“热备”。通过上面的介绍，我们清楚地认识到这

是错误的。

(4)DataNode(数据节点)。数据节点是 HDFS 主从架构中的从角色的扮演者,它在 NameNode 的指导下完成 I/O 任务。如前文所述,存放在 HDFS 的文件都是由 HDFS 的块组成的,所有的块都存放于 DataNode 节点。实际上,对于 DataNode 所在的节点来说,块就是一个普通的文件,可以在 DataNode 存放块的目录下[默认是 $(dfs. data. dir)/current]查看,块的文件名为 blk. blkID。

DataNode 会不断地向 NameNode 报告。初始化时,每个 DataNode 将当前存储的块告知 NameNode,在集群正常工作时,DataNode 仍会不断地更新 NameNode,为其提供本地修改的相关信息,同时接收来自 NameNode 的指令,创建、移动或者删除本地磁盘上的数据块。

(5)块。每个磁盘都有默认的数据块大小,这是磁盘进行数据读/写的最小单位,而文件系统也有文件块的概念,如 ext3、ext2 等。文件系统的块大小只能是磁盘块大小的整数倍,磁盘块的大小一般是 512 B,文件系统的块大小一般为几千字节。例如,ext3 的文件块大小为 4 096 B,Windows 的文件块大小为 4 096 B。用户在使用文件系统对文件进行读取或写入时,完全不知道块的细节,这些对于用户来说是透明的。

HDFS 同样也有块的概念,但是 HDFS 的块比一般文件系统的块大得多,默认为 64 MB,并可以随着实际需要而变化,配置项为 hdfs-site. xml 文件中的 dfs. block. size 项。与单一文件系统相似,HDFS 上的文件也被划分为块大小的多个分块,它是 HDFS 存储处理的最小单元。

例如,某个文件 data. txt 大小为 150 MB,如果此时 HDFS 的块大小没有经过配置,默认为 64 MB,那么该文件在 HDFS 中存储的情况如图 2-4 所示。

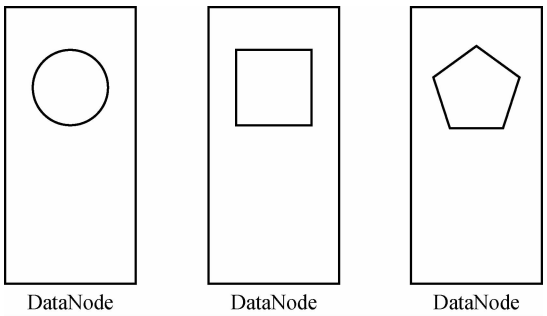

图 2-4 data. txt 在 HDFS 中存储的情况

圆形为保存该文件的第一个块,大小为 64 MB;方形为保存文件的第二个块,大小为 64 MB,五边形为保存文件的第三个块,大小为 22 MB。与其他文件系统不同的是,HDFS 小于一个块大小的文件不会占据整个块的空间,所以第三个块的大小为 22 MB,而不是 64 MB。

HDFS 中的块如此大的目的是为了最小化寻址开销。如果块设置得足够大,从磁盘传输数据的时间可以明显大于定位这个块开始位置所需的时间。这样,传输一个由多个块组

成的文件的时间取决于磁盘传输的效率。得益于磁盘传输速率的提升，块的大小可以被设置为 128 MB 甚至更大。

在 hdfs-site. xml 文件中，还有一项配置为 dfs. relication，该项配置为每个 HDFS 的块在 Hadoop 集群中保存的份数，值越高，冗余性越好，占用存储也越多，默认为 3，即有 2 份冗余。如果设置为 2，那么该文件在 HDFS 中存储的情况如图 2-5 所示。

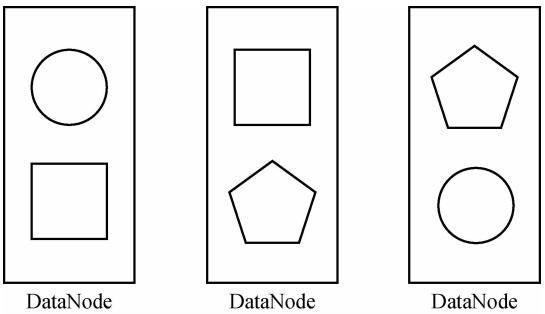

图 2-5　hdfs-site. xml 在 HDFS 中存储的情况

使用块的好处如下。

①可以保存比存储节点单一磁盘大的文件。块的设计实际上就是对文件进行分片，分片可以保存在集群的任意节点，从而使文件存储跨越了磁盘甚至机器的限制。例如，data. txt 文件被切为 3 个块，并存放于 3 个 DataNode 之中。

②简化存储子系统。将存储子系统控制单元设置为块，可简化存储管理，并且也实现了元数据和数据的分开管理和存储。

③容错性高。这是块非常重要的一点，如果将 dfs. relication 设置为 2，如图 2-5 所示，那么任意一个块损坏，都不会影响数据的完整性，用户在读取文件时，并不会察觉到异常。之后，集群会将损坏的块的副本从其他候选节点复制到集群中能正常工作的节点，从而使副本数回到配置的水平。

3. HDFS 容错

如何使文件系统能够容忍节点故障且不丢失任何数据，这就是接下来要介绍的内容，即 HDFS 的容错机制。

1）心跳机制

在 NameNode 和 DataNode 之间维持心跳检测，当网络故障之类的原因导致 DataNode 发出的心跳包没有被 NameNode 正常收到时，NameNode 就不会将任何新的 I/O 操作派发给那个 DataNode，该 DataNode 上的数据被认为是无效的，因此 NameNode 会检测是否有文件块的副本数目小于设置值；如果小于，就自动开始复制新的副本并分发到其他 DataNode 节点。

2）检测文件块的完整性

HDFS 会记录每个新创建文件的所有块的校验和。当以后检索这些文件时或者从某个节点获取块时，HDFS 会首先确认校验和是否一致；如果不一致，会从其他 DataNode 节点上

获取该块的副本。

3)集群的负载均衡

节点的失效或增加可能导致数据分布不均匀,当某个 DataNode 节点的空闲空间大于一个临界值时,HDFS 会自动从其他 DataNode 迁移数据过来。

4)NameNode 上的 FSImage 和编辑日志文件

NameNode 上的 FSImage 和编辑日志文件是 HDFS 的核心数据结构,如果这些文件损坏了,HDFS 将失效。因此,NameNode 由 Secondary NameNode 定期备份 FSImage 和编辑日志文件,NameNode 在 Hadoop 中可能存在单点故障,当 NameNode 出现机器故障时,手动干预是必需的。

5)文件的删除

删除并不是马上从 NameNode 移除命名空间,而是存放在/trash 目录,随时可恢复,直到超过设置时间才被正式移除。设置的时间由 hdfs-site. xml 文件的配置项 fs. trash. interval 决定,单位为秒。

2.1.5 Hadoop MapReduce 的原理

1. MapReduce 简介

MapReduce 源于谷歌的一篇论文,它充分借鉴了分而治之的思想,将一个数据处理过程拆分为主要的 Map(映射)与 Reduce(化简)两步。这样,即使用户不懂分布式计算框架的内部运行机制,只要能用 Map 和 Reduce 的思想描述清楚要处理的问题,即编写 map 和 reduce 函数,就能轻松地实现问题的分布式计算,并在 Hadoop 上运行。MapReduce 的编程具有以下优点。

(1)开发简单。得益于 MapReduce 的编程模型,用户可以不用考虑进程间通信、套接字编程,无须非常高深的技巧,只需要实现一些非常简单的逻辑,其他交由 MapReduce 计算框架去完成,大大降低了分布式程序的编写难度。

(2)可扩展性强。同 HDFS 一样,当集群资源不能满足计算需求时,可以通过增加节点的方式达到线性扩展集群的目的。

(3)容错性强。对于节点故障导致的作业失败,MapReduce 计算框架会自动将作业安排到健康节点重新执行,直到任务完成,而这些对于用户来说都是透明的。

2. MapReduce 架构

MapReduce 采用了 Master/Slave(M/S)架构。它主要由 Client、JobTracker、TaskTracker 和 Task 几个组件组成,如图 2-6 所示。

(1)Client。用户编写的 MapReduce 程序通过 Client 提交到 JobTracker 端;同时,用户可通过 Client 提供的一些接口查看作业运行状态。在 Hadoop 内部用作业(Job)表示 MapReduce 程序。一个 MapReduce 程序可对应若干个作业,而每个作业会被分解成若干个 Map/Reduce 任务。

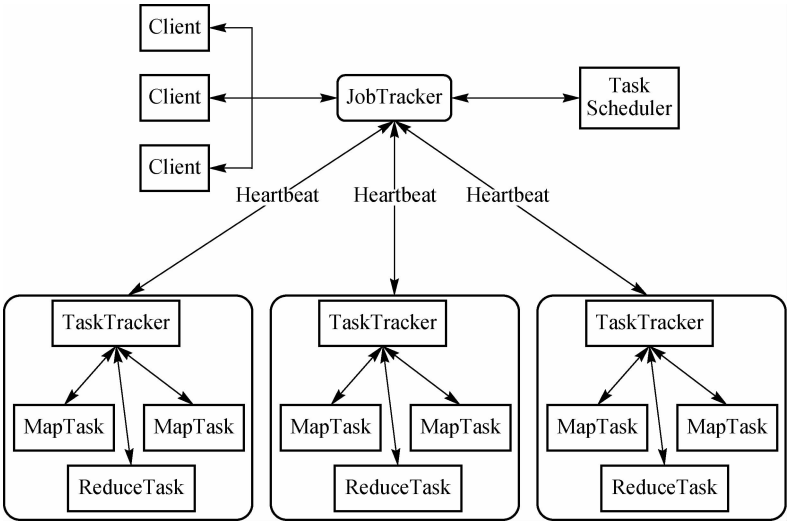

图 2-6　MapReduce 架构

(2)JobTracker。JobTracker 主要负责资源监控和作业调度。JobTracker 监控所有 TaskTracker 与 Job 的健康状况，一旦发现失败情况，其会将相应的任务转移到其他节点；同时，JobTracker 会跟踪任务的执行进度、资源使用量等信息，并将这些信息告诉任务调度器，而调度器会在资源出现空闲时，选择合适的任务使用这些资源。在 Hadoop 中，任务调度器是一个可插拔的模块，用户可以根据自己的需要设计相应的调度器。

(3)TaskTracker。TaskTracker 会周期性地通过 Heartbeat 将本节点上资源的使用情况和任务的运行进度汇报给 JobTracker，同时接收 JobTracker 发送过来的命令并执行相应的操作(启动新任务、杀死任务等)。TaskTracker 使用 Slot 等量划分本节点上的资源量。Slot 代表计算资源(CPU、内存等)。一个 Task 获取到一个 Slot 后才有机会运行，而 Hadoop 调度器的作用就是将各个 TaskTracker 上的空闲 Slot 分配给 Task 使用。Slot 分为 MapSlot 和 ReduceSlot 两种，分别供 MapTask 和 ReduceTask 使用。TaskTracker 通过 Slot 数目(可配置参数)限定 Task 的并发度。

(4)Task。Task 分为 MapTask 和 ReduceTask 两种，均由 TaskTracker 启动。HDFS 以固定大小的 Block 为基本单位存储数据，而对于 MapReduce 而言，其处理单位是 Split。有关 Split 的具体内容和 MapReduce 的操作将在模块 4 中深入展开，这里不再详细描述。

3. MapReduce 在数据处理方面的缺点

(1)不适应事务/单一请求处理。MapReduce 绝对是一个离线批处理系统，对于批处理数据应用得很好。MapReduce(不论是 Google 的，还是 Hadoop 的)是用于处理不适合传统数据库的海量数据的理想技术，但它又不适合事物/单一请求处理。

(2)性能问题。想想 N 个 Map 实例产生 M 个输出文件，每个最后由不同的 Reduce 实例处理，这些文件写到运行 Map 实例机器的本地硬盘。如果 N 是 1 000，M 是 500，Map 阶

段产生 500 000 个本地文件。当 Reduce 阶段开始，500 个 Reduce 实例每个需要读入 1 000 个文件，并用类似 FTP 协议把它要的输入文件从 Map 实例运行的节点上获取过来。假如同时有数量级为 100 的 Reduce 实例运行，那么 2 个或 2 个以上的 Reduce 实例同时访问同一个 Map 节点来获取输入文件是不可避免的，即导致大量的硬盘查找，使有效的硬盘运转速度至少降低 20%。

(3)不适合一般 Web 应用。大部分 Web 应用，只是对数据进行简单的访问，每次请求处理所耗费的资源其实非常小，它的问题是高并发，所以要采用负载均衡技术来分担负载。只有在特殊情况下才可能用到 MR，如创建索引、进行数据分析等。

2.2 扩展知识

谈到 Hadoop 应用，它在为搜索引擎提供动力或为广告商提供用户行为分析的平台方面显然最为知名。此外，还有在线旅游、移动数据、电子商务、能源发现、能源节省、基础设施管理、图像处理、欺诈检测、IT 安全和医疗保健等多个应用领域，Hadoop 显然比人们预想得更加富有生命力。

2.2.1 Hadoop 在全球最大超市业者 Wal-Mart 的应用

Wal-Mart(沃尔玛)虽然十年前就投入了在线电子商务，但在线销售的营收远远落后于 Amazon(亚马逊)。后来，Wal-Mart 决定采用 Hadoop 来分析顾客搜寻商品的行为及用户透过搜索引擎寻找到 Wal-Mart 网站的关键词，利用对这些关键词的分析结果发掘顾客需求，以规划下一季商品的促销策略，并进一步分析顾客在 Facebook、Twitter 等社交网站上对商品的讨论，甚至 Wal-Mart 能比父亲更快知道女儿怀孕的消息，并主动寄送相关商品的促销邮件，可以说是比竞争对手提前一步发现顾客。

2.2.2 Hadoop 在 Visa 的应用

Visa 公司拥有一个全球最大的付费网络系统 VisaNet，用于信用卡付款验证。2009 年时，Visa 公司每天就要处理 1.3 亿次授权交易和 140 万台 ATM 的联机存取。为了降低信用卡诈骗、盗领事件的损失，Visa 公司得分析每一笔事务数据，来找出可疑的交易。虽然每笔交易的数据记录只有短短 200 位，但每天 VisaNet 要处理全球上亿笔交易，2 年累积的资料多达 36 TB，过去光是要分析 5 亿个用户账号之间的关联，就得等 1 个月才能得到结果。所以，Visa 在 2009 年导入了 Hadoop，建置了 2 套 Hadoop 丛集(每套不到 50 个节点)，让分析时间从 1 个月缩短到 13 分钟，更快速地找出了可疑交易，也能更快地对银行提出预警，甚至能及时阻止诈骗交易。

2.2.3 Hadoop 在百度的应用

百度公司作为全球最大的中文搜索引擎公司，提供基于搜索引擎的各种产品，包括以网络搜索为主的功能性搜索，以贴吧为主的社区搜索，针对区域、行业的垂直搜索，音乐搜索等，几乎覆盖了中文网络世界中所有的搜索需求。因此，百度网站对海量数据处理的要求是比较高的，要在线下对数据进行分析，还要在规定的时间内处理完并反馈在平台上。在百度，Hadoop 主要应用于以下几个方面。

(1)数据挖掘与分析。

(2)日志分析平台。

(3)数据仓库系统。

(4)推荐引擎系统。

(5)用户行为分析系统。

虽然 Hadoop 对百度提供了很大的帮助，但是百度在使用 Hadoop 时也遇到了以下一些问题。

(1)MapReduce 的效率问题。

(2)HDFS 的效率和可靠性问题。

(3)内存使用的问题。

(4)作业调度的问题。

(5)性能提升的问题。

(6)健壮性的问题。

(7)Streaming 局限性的问题。

(8)用户认证的问题。

因此，百度为了更好地用 Hadoop 进行数据处理，在以下几个方面做了改进和调整。

(1)调整 MapReduce 策略。

①限制作业处于运行状态的任务数。

②调整预测执行策略，控制预测执行量，一些任务不需要预测执行。

③根据节点内存状况进行调度。

④平衡中间结果输出，通过压缩处理减少 I/O 负担。

(2)改进 HDFS 的效率和功能。

①权限控制。在 PB 级数据量的集群上数据应该是共享的，这样分析起来比较容易，但是需要对权限进行限制。

②让分区与节点独立。这样，一个分区坏掉后，节点上的其他分区还可以正常使用。

③修改 DFS Client 选取块副本位置的策略，增加功能，使 DFS Client 选取块时跳过出错的 DataNode。

④解决 VFS(virtual file system)的 POSIX(portable operating system interface of UNIX)兼容性问题。

(3)修改 Speculation 的执行策略。

①采用速率倒数替代速率,防止数据分布不均时经常不能启动预测执行情况的发生。

②增加任务时,必须达到某个百分比后才能启动预测执行的限制,解决 Reduce 运行等待 Map 数据的时间问题。

③只有一个 Map 和 Reduce 时,可以直接启动预测执行。

(4)对资源使用进行控制。

①对应用物理内存进行控制。内存使用过多会导致操作系统跳过一些任务,百度通过修改 Linux 内核对进程使用的物理内存进行独立的限制,若超过阈值则终止进程。

②分组调度计算资源,实现存储共享、独立计算,在 Hadoop 中运行的进程是不可抢占的。

③在大块文件系统中,x86 平台下一个页的大小是 4 KB。如果页较小,管理的数据就会很多,会增加数据操作的代价并影响计算效率,因此需要增加页的大小。

百度在 2006 年就开始关注 Hadoop 并开始调研和使用,在 2012 年,其总的集群规模达到 10 个,单集群超过 2 800 台机器节点,Hadoop 机器总数有上万台,总的存储容量超过 100 PB,已经使用的超过 74 PB,每天提交的作业数目也有数千个之多,每天的输入数据量已经超过 7 500 PB,输出数据量超过 1 700 TB。同时,百度在 Hadoop 的基础上还开发了自己的日志分析平台、数据仓库系统和统一的 C++编程接口,并对 Hadoop 进行深度改造,开发了 Hadoop C++扩展 HCE 系统。

2.3 实　　训

安装 Hadoop 的第一步,就是根据实际情况选择最合适的 Hadoop 版本。目前,由于 Hadoop 飞速发展,功能更新和错误修复在不断地迭代着,因而版本特别多,显得有些杂乱。结合想要的功能和稳定性,这里选择 CDH5,该版本是目前生产环境中装机量最大的版本之一,涵盖了所有 Hadoop 的主要功能和模块,稳定且有很多有用的新特性。其下载地址为 https://archive.cloudera.com/cdh5/cdh/5/hadoop-3.2.0.tar.gz。

Hadoop 的运行环境有以下两种。

(1)Windows 环境。虽然目前 Hadoop 社区已经支持 Windows,但由于 Windows 操作系统本身不适合作为服务器操作系统,所以本书不介绍 Windows 下 Hadoop 的安装方式。

(2)Linux 环境。Hadoop 的最佳运行环境无疑是世界上最成功的开源操作系统 Linux。Linux 的发行版本众多,常见的有 CentOS、Ubuntu、RedHat 等。本书选择的是 CentOS。

1. 安装虚拟机

(1)进入 VMware 的安装向导,如图 2-7 所示。

(2)选择自定义安装,如图 2-8 所示。

图 2-7　新建虚拟机向导 1

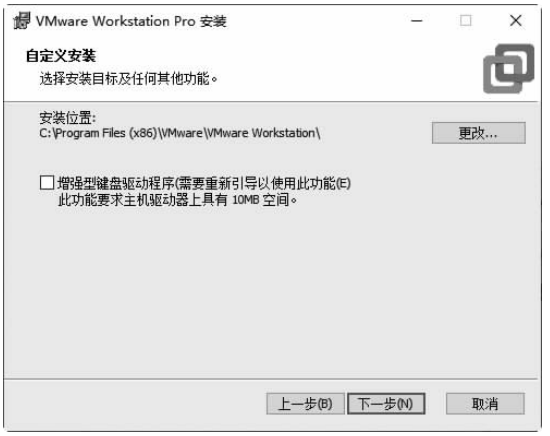

图 2-8　自定义安装

(3)设置用户体验，如图 2-9 所示。

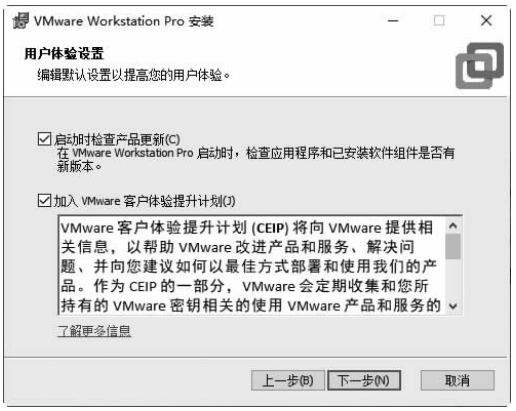

图 2-9　设置用户体验

(4)创建快捷方式，如图 2-10 所示。

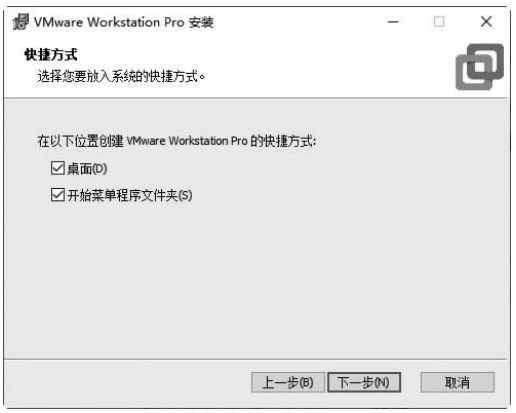

图 2-10　创建快捷方式