第一章　大数据技术及应用

随着信息技术的飞速发展，人类社会进入了数字信息时代。获取和掌握信息的能力已成为衡量一个国家实力强弱的标志。一切信息因需求不同其效益也不同，而一切有益信息都是从大量数据中分析出来的。海量数据又随着时间持续产生，不断流动，进而扩散形成大数据。大数据不仅用来描述数据的量非常巨大，还突出强调处理数据的速度，所以，大数据成为数据分析领域的前沿技术。“大数据”可能带来的巨大价值正渐渐被人们认可，它通过技术的创新与发展，以及数据的全面感知、收集、分析、共享，为人们提供了一种全新的看待世界的方法。

第一节　数　　据

从计算机科学的角度，数据（data）是所能输入计算机并被计算机程序处理的符号的总称，是具有一定意义的数字、字母符号和模拟量的统称。在计算机科学之外，我们可以更加抽象地定义数据，如人们通过观察现实世界中的自然现象、人类活动，都可以形成数据。

计算机最初的设计目的就是用于数据的处理，但计算机需要将数据表示0和1的二进制形式，用一个或若干个字节（byte，B）表示，一个字节等于8个二进制位（bit），每个位表示0或1。因此，计算机对数据的处理首先需要对数据进行表示和编码，从而衍生出不同的数据类型。

对于数字，可以编码成二进制形式。例如，十进制数的10，在计算机中会用二进制表示为1010。同样，对于负数、小数，在计算机内部也会有不同的编码方式。

对于文本数据，通常计算机会采用ASCII码将其编码为一个整数。如字符A就会编码为整数32；同样，对于汉字或其他特殊符号，也对应不同的编码体系，如《信息交换用汉字编码字符集》（GB 2312—1980）会将一个汉字编码为连续的两个字节。

有时可能需要用更加复杂的数据结构（如向量、矩阵）来表达一个复杂的状态。例如，表达地图上的位置信息，就需要用到二维坐标。

表示一个实体的不同方面，会用到不同的数据。例如，描述一个学生，可能会包括姓名、性别、年龄等多种属性，每种属性都需相应类型的数据来刻画。有时，如果连续观察一

个实体在一段时间的状态变化，就可以得到一个时间序列数据，例如，用于检测城市空气质量中细颗粒物（PM2.5）含量的传感器，每隔5分钟会产生一个监测数据，这些数据就形成了一个PM2.5随时间的变化情况。根据数据所刻画的过程、状态和结果的特点，数据可以划分为不同的类型。按照数据是否有强的结构模式，可将数据划分为结构化数据、半结构化数据和非结构化数据。在数据的处理过程中，会根据数据的不同类型，选择不同的数据管理方法和处理技术。

一、结构化数据

结构化数据是指具有较强的结构模式，可使用关系型数据库表示和存储的数据。结构化数据通常表现为一组二维形式的数据集，一行表示一个实体的信息，每一行的不同属性表示实体的某一方面，每一行数据具有相同的属性。这类数据本质上是“先有结构，后有数据”。

二、半结构化数据

半结构化数据是一种弱化的结构化数据形式，它并不符合关系型数据模型的要求，但仍有明确的数据大纲，包含相关的标记，用来分割实体以及实体的属性。这类数据中的结构特征相对容易获取和发现，通常采用XL、JSON等标记语言来表示。

三、非结构化数据

人们日常生活中接触的大多数数据都属于非结构化数据。这类数据没有固定的数据结构，或难以发现统一的数据结构。各种存储在文本文件中的系统日志、文档、图像、音频、视频等数据都属于非结构化数据，如图1-1所示。

图1-1　非结构化数据

在得到数据的同时，往往也能够得到或分析关于数据本身的一些信息。例如，描述学生的例子是指几个人的数据集，每个人有编号、姓名、年龄、性别等属性，而这些属性本身也

有不同的数据类别，需要按照不同的方式进行编码。这些信息是描述一个数据集本身特征的数据，通常称为元数据（metadata）。元数据是描述数据的数据，机器可读的元数据可以帮助计算机自动地对一组数据进行处理。同时还要注重对数据进行归一处理[①]。

此外，在计算机中为了方便数据的组织，可以将数据以文件的方式保存起来。相同的数据表示，可以按照不同的具体格式组织在文件中。例如，一个表格数据，可以按照行来顺序写入文件，也可以按照列来顺序写入文件。针对不同目的设计的存储系统（如文件系统、关系数据库等），会专门选择最适合这类数据的存储方式，以更好地利用存储空间，并可以加速对数据的访问。在计算机中，文件系统也帮助管理大量文件，以及管理文件名，创建用户，设置读写权限，创建时间等产生的元数据。

第二节　大数据的内涵和外延

数据要被计算机处理，首先需要编码成计算能够接受的二进制格式。在前面我们提到可以用字节作为衡量数量大小的基本单位，每个字节代表8个二进制位，因此，一个字节可以表示0～255种不同的状态。字节可以看成是表示信息的“基本单位”。由于硬件设计的原因，计算机处理信息通常以2的整数倍作为处理边界，最接近于1000的2的整数倍是1024，因此，在计算机中大多采用这些标准单位前缀与字节的组合表示数据量，各数据单位之间的换算关系如表1-1所示。

表1-1　数据单位之间的换算关系

单　位	换算关系
byte（字节）	1byte=8bit
KB(Kilobyte 千字节)	1KB=1024byte
MB(megabyte，兆字节)	1MB=1024KB
GB(gigabyte，吉字节)	1GB=1024MB
TB(Trillionbyte，太字节)	1TB=1024GB
PB(petabyte，派字节)	1PB=1024TB
EB(exabyte，艾字节)	1EB=1024PB
ZB(zettabyte，泽字节)	1ZB=1024EB

那么，数据要多大才算“大数据”？是否还有其他区分数据和大数据的标准呢？下面来了解一下大数据的概念是如何被提出来的，大数据到底有什么特征等。

① 刘云霞.数据预处理 数据归约的统计方法研究及应用[M].厦门：厦门大学出版社，2011：33.

一、大数据时代的驱动力

近年来，随着互联网技术的发展及移动互联网、物联网等技术的广泛应用，人、机、物三元对象进入深度融合时代，网络信息空间反映了人类社会与物理世界的复杂联系。网络信息空间的数据与人类活动密切相关；网络信息空间的规模以指数级增长，呈高度复杂化趋势。换句话说，人们进入了一个数据爆炸的大数据时代。

这一轮数据增长的一个推动力是大量信息传感设备的出现，以及快速发展的物联网技术及应用，这使得大量物理世界的状态被获取存储下来。例如，随着新一代数据采集与传输设备在民用客机上的应用，2011年空客A50的飞机监控参数达到40万个，波音787飞机的监控参数达到15万个，极大改善了对机载系统及发动机运行状态的监控能力。今天随着我国城市化的发展，城市中部署的大量交通治安摄像头进行联网，由此汇聚的数据量将更加惊人。

数据增长的另一个重要的推动力量来自快速发展的互联网和移动互联网。互联网上汇聚了数十亿网民，用户产生的数据量十分巨大；移动互联网使用户更紧密地融入网络世界中。据近几年的统计显示，中国用户平均每天花费在各类移动应用上的时间达到了31亿小时，用户通过使用这些移动应用产生了大量行为数据和内容数据。以社会网络应用——微博为例，2017年，我国微博月活跃用户就达到3.92亿，每天发布微博超过2亿条（根据新浪微博的2017年财报），其中图片和小视频的数量达到2000万。按每条微博170B，每张图片或小视频1MB计算，仅微博一类应用一天产生的数据就高达19TB。互联网搜索类业务需要检索互联网的网站内容，平均每天需要扫描处理的数据量甚至达到了100PB量级。

IDC（international data corporation，国际数据公司）曾做过一项统计，该统计的主要数据为人类产生并存储下来的数据，截至2009年，该数据已有0.8ZB。截至2013年，该数据就已经超过了4.4ZB，并且这些数据的增长速度是逐渐加快的。IDC还预测到，到2025年，这些数据有可能达到163ZB。

在这些数据的基础上，当人们研究一个现象或问题时，就有了基于数据形成的对现实世界的理解。与传统的统计学类似，通常需要通过精心设计的传感器或各类移动互联网应用去对现实世界进行抽样。但与传统统计学不同的是，人们有可能通过获得更接近于全样的抽样，形成一个客观世界的实体和现象在计算机能够处理的信息世界中的一个数字映像。例如，在智能制造系统中，有数字孪生（digital twin）的概念。美国国防部最早提出在数字空间里建立真实飞机的模型，并通过传感器实现飞机真实状态的完全同步。这样，每次飞行后，就可以基于数字模型的现有情况和过往载荷，及时分析评估飞机是否需要维修，能否承载下次任务载荷等。因此，如何利用已经获得汇聚的数据，以及如何精巧地设计新的数据获取方式，构建一个能够足够精确反映客观世界的实体、现象和行为特征的数字映像，进而在这一数字映像之上，对客观世界的实体、现象和行为特征进行推演，是许多实际应用领域数据增长的内生动力。

然而，随着数据总量的快速增长，以及越来越多的数据分析任务的出现，针对大数据的获取、存储、传输、处理等能力都面临新的技术挑战。如果数据不能存储下来并及时分析处理，大数据就无法产生具有时效性的价值，因此，拥有真实数据以及对数据的实时处理能力，才能够从大量无序的数据中获取价值，也才会具有大数据时代的核心竞争力。

二、大数据的概念和特征

虽然世界都在时刻关注着大数据，但是关于大数据的具体概念，实际上还没有一个官方的定义。麦肯锡全球研究机构（McKinsey global institute）给出的大数据定义，综合了“现有技术无法处理”和“数据特征”定义。他们认为数据是指大小超过经典数据库软件工具收集、存储、管理和分析能力的数据集。这一定义是站在经典数据库的处理能力的基础上看待大数据的。维基百科对“大数据”的解读是：“大数据”或称巨量数据、海量数据、大资料，指的是所涉及的数据量规模巨大到无法通过人工在合理时间内达到截取、管理、处理，并整理成为人类所能解读的信息。美国国家标准与技术研究院（national institute of standards and technology，NIST）认为，大数据是用来描述在我们网络的、数字的、遍布传感器的、信息驱动的、世界中呈现出的数据泛滥的常用词语。大量数据资源为解决以前不可能解决的问题带来了可能性。

目前通常认为大数据具有4V特征，即规模庞大（volume）、种类繁多（variety）、处理速度快（velocity）和价值巨大但价值密度低（value）。

1. 规模宏大

分析数据集目前所拥有的计算能力以及存储能力，大数据是具有规模庞大的特点的。在刚刚出现大数据这一概念时，人们认为PB级的数据就可以看作是“大数据”。但是事实上，这种观点并不是完全正确的，其原因主要包括两点：一点是因为当数据的存储技术和计算的技术得到了发展时，当在互联网上生成的数据增多时，当通过传感器所获得的数据增多时，是会影响判断是否为“大数据”的依据的；另一点在于一些数据具有大数据的特点，但是这些数据却不属于PB级，这种情况下我们也不能说这些数据就不是大数据。这种具有庞大规模的大数据，对于传输数据、存储数据以及分析数据等方面，都带来了不小的挑战。

2. 种类繁多

大数据拥有繁多的种类，主要表现在两个方面：一方面是指在同一情境中，大数据集中拥有多种不同种类的数据，包括结构化数据、非结构化数据以及半结构化数据等；另一方面是指在同一种类的数据中，其数据的结构模式是多样的。举例来说，用于处理城市交通数据的应用，其拥有的数据类型就包括结构化数据类型、半结构化数据类型以及非结构化数据类型三种，其中结构化数据类型指的是车辆注册的数据信息、驾驶人的数据信息以及城市交通道路的信息等；半结构化数据类型指的是不同类型的文档数据；非结构化数据类型指的是摄像头所记录下的各种数据信息等。对于大数据的处理工作之所以比较复杂，主要是因为数据所具有的异构性，而这种异构性就是因这些数据类型的多样性而形成的，因此也对于数据处理能力有着极高要求。

3. 处理速度快

大数据时代的数据产生速度非常迅速。在Web 2.0应用领域，在1分钟内，新浪可以产生2万条微博，Twitter可以产生10万条推文，苹果可以下载4.7万次应用，淘宝可以卖出6万件商品，人人网可以发生30万次访问，百度可以产生90万次搜索查询，Facebook可以产生600万次浏览量。大名鼎鼎的大型强子对撞机（LHC），大约每秒产生6亿次的碰撞，每秒生成约700MB的数据，有成千上万台计算机分析这些碰撞。

大数据时代的很多应用，都需要基于快速生成的数据给出实时分析结果，用于指导生产和生活实践，因此，数据处理和分析的速度通常要达到秒级响应，这一点和传统的数据挖掘技术有着本质的不同，后者通常不要求给出实时分析结果。

为了实现快速分析海量数据的目的，新兴的大数据分析技术通常采用集群处理和独特的内部设计。以谷歌公司的Dremel为例，它是一种可扩展的、交互式的实时查询系统，用于只读嵌套数据的分析，通过结合多级树状执行过程和列式数据结构，它能做到几秒内完成对万亿张表的聚合查询，系统可以扩展到成千上万的CPU上，满足谷歌上万用户操作PB级数据的需求，并且可以在2～3秒内完成PB级别数据的查询。

4. 价值巨大但价值密度低

分析大数据所具有的价值可以发现，其蕴含的价值是巨大的，但是这种价值的密度并不高。该价值源于大数据中所包含的隐含知识上，因为隐含知识是具有高价值的，所以大数据才蕴含巨大的价值，并且在许多方面都能体现出大数据所具有的价值，例如关联和假设检验。这种隐含知识在表面上并不能被发现，首先需要对大数据进行分析，其次在分析出的各类无序数据之间建立起一种关联，最后才能得到这种隐含知识。大数据的数据集是在不断增长的，但是数据所蕴含的价值并没有随之而增长，因为这些增长的数据并不都是有价值的，这类数据通常被称作无用数据，大量的无用数据将有价值的数据掩盖了起来，因此，大数据的价值密度比较低。关于大数据的计算，其最主要的一个问题，就是怎样才能在无用数据中找到有价值的数据，其价值密度又该怎样进行度量。

在此基础上，还有一些学者在大数据的4V特征基础上增加了其他提法，形成大数据的5V特征。例如，前面提到的BM就从获取的数据质量的角度，将真实性或准确性（veracity）作为大数据的特征，着重说明大数据面临的数据质量挑战。从互联网或传感器获得的关于真实世界和人类行为的数据中，可能存在各类噪声、误差，甚至是虚假、错误的数据，有些情况下也会有数据缺失。数据的真实性，则强调数据的质量是大数据价值发挥的关键。

其实，无论是4V还是5V，都是从特性的角度刻画数据集本身的一些特征。这些特征对发现事实，揭示规律并预测未来，提出了新的挑战，并将对已有计算模式、理论和方法产生深远的影响。

三、大数据带来的思维模式改变

大数据给传统的数据带来了三个思维模式的改变。

（一）采样与全样：尽可能收集全面而完整的数据

在统计方法中，由于数据不容易获取，数据分析的主要手段是进行随机采样分析，该手段成功应用到了人口普查、商品质量监管等领域。然而随机采样的成功依赖于采样的绝对随机性，而实现绝对随机性非常困难，只要采样过程中出现任何偏见，都会使分析结果产生偏离。即使有了最优采样的标准与方法，在大数据时代由于数据的来源非常多，需要全面地考虑采样的范围，因此找到最优采样的标准非常困难。同时，随机采样的数据方法具有

确定性，即针对特定的问题进行数据的随机采样，一旦问题变化，采样的数据就不再可用。随机采样也受到数据变化的影响，一旦数据发生变化，就需要重新采样。

大数据不仅是数据量大，更体现在“全”上。当有条件和方法获取到海量信息时，随机采样的方法和意义就大幅降低了。确实，各类传感器、网络爬虫、系统日志等方式使人们拥有了大数据。存储资源、计算资源价格的大幅降低以及云计算技术的飞速发展，不仅使大公司的存储能力和计算能力大幅提升，也使中小企业有了一定的大数据处理与分析的能力。

（二）精确与非精确：宁愿放弃数据的精确性也要尽可能收集更多的数据

对小数据而言，由于收集的信息较少，对数据基本要求是数据尽量精确、无错误。特别是在进行随机抽样时，少量错误将可能导致错误的无限放大，从而影响数据的准确性。同时，正由于数据量小，才有可能保证数据的精确性。因此，数据的精确性是人们追求的目标。

然而，对于大数据，保持数据的精确性几乎是不可能的。一方面，大数据通常源于不同领域产生的多个数据源，当由大数据产生所需信息时，通常会出现多源数据之间的不一致性。同时，也由于数据通过传感器、网络爬虫等形式获取时经常会产生数据丢失，因而使数据不完整。虽然目前有方法和技术来进行数据清洗，试图保证数据的精确性，然而这不仅耗费巨大，而且保证所有数据都是精确的几乎是不可能的。因此，大数据无法实现精确性。

另一方面，保持数据的精确性并不是必需的。经验表明，有时牺牲数据的精确性而未得更广泛来源的数据，反而可以通过数据集间的关联提高数据分析结果的精确性。例如，Facebook、微博、新闻网站、旅游网站等通常允许用户对网站的图片、新闻、游记等打标签。每个用户打的标签并没有精确的分类标，也没有对错，完全从用户的感受出发。这些标签达到几十亿的规模，却能让用户更容易找到自己所需的信息。

（三）因果与关联：基于归纳得到的关联关系与逻辑推理的因果关系同样具有价值

通常人们对数据进行分析从而预测某事会发生，其中基于因果关系分析和关联关系分析进行预测是常用的方法。

在大数据时代，对于已经获取到的大量数据广泛采用的方法是使用关联关系来进行推测。经验表明，在大数据时代，由于因果关系的严格性使数据量的增加并不一定有利于得到因果关系，反而关联关系更容易得到。例如，通过观察可以发现打伞和下雨之间存在关联关系，这样，当看到窗外所有人都打着伞，那么就可以推测窗外在下雨，在这个过程中，我们并不在意到底是打伞导致了下雨，还是下雨导致了打伞。目前，基于关联关系分析的预测被广泛应用于各类推荐任务上。例如，著名的“啤酒加尿布”例子，并没有得到男性顾客买啤酒一定会买尿布或买尿布一定会买啤酒的结论，而是得到了啤酒和尿布之间的关联关系。同样，2009年谷歌的科研人员在《自然》杂志撰文，通过对每日超过30亿次的用户搜索请求及网页数据的挖掘分析，在甲型H1N1流感暴发的几周前

就预测出流感传播，也是利用了搜索关键词和流感发病率之间的关联而非因果关系。通常，数据中能够发现的更多是关联关系，因果关系的判断和分析需要由领域专家的参与才能完成。

当然，重视关联关系并不否定探寻因果关系的重要性。事实上，也有很多研究在探索如何从数据中获得因果关系。医学上利用典型的“双对比试验”来判断药物对疾病的作用；智能工业互联网应用中，需要了解究竟是哪个因素与产品优良率之间存在因果关系，这些都是典型的基于实验数据推断因果关系，进而推动应用的例子。因此，在大数据中，关联关系与因果关系同样具有应用价值。

四、大数据的作用和意义

如今的世界已经进入了全球信息化的时代，并且这种信息化还在不断地快速发展着，而大数据在这样的情境下逐渐变成一种带有战略性质的最基础的资源，并且在任何国家都占有重要地位，同时各个国家的科技创新也是在大数据的基础上进行的。对于大数据的开发和利用不仅能为国家提供商业价值，还能提供社会价值，同时对改变科学研究的模式也具有推动的作用。大数据存在的意义表现在多个方面，包括对国家经济发展和国家经济的安全具有战略性的意义以及长远性的意义等，同时，大数据也可以看作是一种竞争优势，存在于国家之间的竞争中。在我国，要想提升大数据的质量、规模以及使用大数据的技术水平，最先要做的就是将我国自身的规模优势充分地利用起来，并发挥大数据中数据的价值，从而稳固其具备的战略作用。

（一）在经济方面，大数据成为推动经济转型发展的新动力

大数据的出现对于科学技术、社会人才、物质资源以及资金都具有深远的影响，同时受到影响的还有社会上工作的组织模式，对于生产组织方式的创新具有推动作用。社会中的各类生产要素因大数据的产生，可以通过网络化实现共享，通过集约化实现整合，通过协作化实现开发，并得到充分的利用，从而使生产方式和社会上的经济运行机制得到了转变，都不再是传统的模式，经济运行的水平和运行的效率也都因此得到了提升。为实现互联网中各新型领域业务的增值创新，实现相关企业核心价值的提升，就需要利用大数据使各商业的模式得到创新发展，并促进新业态的出现。在信息产业格局中，大数据的影响力度是极其深远的，同大数据相关的产业也成了一个全新的经济增长点。

有数据显示，2016 年全球大数据业市场规模为 1403 亿美元。例如，阿里巴巴凭借其电子商务平台的大量交易数据，提前 8 ~ 9 个月预测出 2008 年的金融危机；百度通过对超过 4 亿用户的搜索请求及交互数据的挖掘分析，建立用户行为分析模型，在提供个性化智能搜索和内容推荐的同时，取得了中国互联网搜索市场的领先地位；以共享单车、各类专车等城市出行领域为代表的共享经济应用，改善了供需匹配，促进了资源的有效利用。而大数据在传统工业和制造业领域的应用则有助于帮助制造企业打通产业链，延伸产品的价值链条，并支持产品有更快的升级迭代和更好的个性化服务。

（二）在社会方面，大数据成为提升政府治理能力的新途径，社会安全保障的新领地

一些数据的关联关系是没有办法通过传统的技术方式展现出来的，最有效的方法就是通过使用大数据将其展示出来。对于政府数据，通过对大数据的应用，可以使政府的数据得到共享，使社会中各项事业的数据和资源得到整合，关于政府的整体数据，其分析能力也有所提升，同时，大数据的应用也是一种新的、用于处理比较复杂的社会问题的方式。有了大数据，就可以建立一种新的管理机制，即“用数据说话、用数据决策、用数据管理、用数据创新”，通过数据的利用还可以完成许多科学决策，对于政府的管理以及社会的治理都能起到更新的作用，从而推进政府的治理能力步入现代化，同时，在大数据的基础上创造出一种新型的政府，即廉洁的政府、法治的政府、服务型的政府以及创新型的政府，且该政府是符合中国特色社会主义事业发展道路的，也是符合社会主义市场经济体制的。

可以说，有了百度或谷歌，就可以分析掌握用户的浏览习惯；有了淘宝或亚马逊，就可以分析掌握用户的购物习惯；有了新浪微博或Twitter，就可以了解用户的思维习惯及其对社会的认知。而且对微博等网络信息大数据的挖掘，能够及时反映经济社会动态与情绪，预警重大、突发和敏感事件（如流行病暴发、群体异常行为等），协助提高社会公共服务的应对能力，对维护国家安全和社会稳定具有重大意义。

（三）在科研方面，大数据成为科学研究的新途径

借助对大数据的分析研究，能够发现医学、物理、经济和社会等领域的新现象，揭示自然与社会中的新规律，并预测未来趋势，这使基于数据的探索（data exploration）成为科学发现继实验/经验（empirical）、理论（theory）、计算（computational）之后的“第四种范式”。数据密集型科学探索与第三种范式（计算密集型仿真与模拟）都是信息技术支撑的科学发现方式，但最大的不同在于计算范式是先提出可能的理论，再收集数据，然后通过计算仿真进行理论验证；而数据密集科学探索则是先通过各种信息获取技术获得大量已知数据，然后通过分析和计算寻找其中的关联与因果关系，从而得出之前未知的理论。正在兴起的环境应用科学、基于全球数据共享的天文观测、下一代传感器网络与地球科学、脑科学与大脑神经回路突破，这些都是在快速成长和发展的交叉学科方向，也是大数据用于科学研究和发现的很好实例。同时，这些科学研究的新需求，也在催生传感、网络、存储、计算等信息技术的突破，以及以数据为中心的获取、传输、管理、分析和可视化技术的进步。2010年《经济学人》周刊发表封面文章，也提出了“数据泛滥（data deluge）为科研带来新机遇”的观点。而《自然》《科学》相继出版了*Big Data*和*Dealing with Data*的专刊。许多国际著名期刊和会议均专门研究大数据的相关问题，在国际上引发了新一轮科研热潮。

大数据对于社会、经济以及技术的研究和发展都具有巨大的价值，正是因为这种价值，使世界上众多的发达国家对大数据格外关注，并在研究的过程中使用了许多人力和财力。这些国家还专门制定了相关的政策法规，以推动大数据产生得到进一步的发展。美国曾提出大数据研发计划，用英文表示为big sata & sevelopment initiative，该计划是奥巴马于2012年3月提出的，并为大数据的研发提供了2亿美元的研究资金，大数据研究也因此进入了国家战略的层面。欧盟于2012年的6月，也为大数据的研究提供了资金上的支持，其研

究的方式是从大科学(big science)问题研究的角度出发,找到新的更加科学的研究方法,并对超级计算以及用于开发大规模数据的平台给予充分的支持。

对大数据高度关注的国家还有英国,其主要是通过利用大数据中公开数据的商业潜力,并对该潜力进行开发,从而为国家提供一条用于创新发展的新道路,并推动可持续发展政策的进一步实施,在2012年5月,英国还为此特地创建了一个开放数据研究所,用英文表示为The Open Data Institute,缩写为ODI。值得一提的是,该研究并具有非营利性质。同时,英国将这种大数据看作是一种战略性的技术,并设立了相关的战略规划,即《英国数据能力发展战略规划》。该战略规划主要是由英国的商务部、创新部门以及技能部门共同制定的。

在同一年开始关注大数据发展的国家还有澳大利亚,并同样为大数据制订了一份相关的战略计划,其目的是有效地利用大数据对公共行业的服务进行改革,并制定更加适宜的公共政策,为维护公民的隐私权限,该战略计划名为《公共服务大数据战略》,发布战略的主体为政府的信息管理办公室(AGIMO),发布的时间为2013年8月,澳大利亚也因该战略计划,使得他们对大数据处理和利用达到世界领先水平。日本对于大数据也是有所关注的,并在2013年发布了《创建最尖端IT国家宣言》,在该宣言中日本把公共数据和大数据的开放以及发展工作作为国家在2013—2020年最为重要的国家战略,并提出只有有效应用大数据,才能使日本获得更高的竞争力。

对于大数据所带来的机遇,在我国也有同样的认识,并为大数据的研究工作专门部署了相关的计划,即《国家中长期科学和技术发展规划纲要(2006—2020年)》,我国的国家战略需求也因该计划增添了信息处理和挖掘知识的内容。之后我国的科技部门发布了《"十五"国家科技计划信息技术领域2013年度备选项目征集指南》,在该指南中最重要的内容就是关于大数据的研究工作。随着大数据研究工作的展开,我国开始向数据强国的方向迈进。关于数据强国的概念,是在2015年9月国务院发布的《促进大数据发展行动纲要》中第一次正式提出的,并在同年的10月,在党的十八届五中全会上,大数据正式成为我国的国家战略。之后在2016年,关于国家大数据的战略,在《中华人民共和国国民经济和社会发展第十三个五年规划纲要》中,将大数据作为一种带有战略性质的基础资源,在推动大数据发展的过程中,实现数据资源的共享,并且充分利用数据资源帮助社会中各类产业的转型或升级,同时辅助国家对社会开展有效的治理工作。

大数据已成为关系国家经济发展、社会安全和科技进步的重要战略资源,是国际竞争的焦点和制高点。开展大数据计算的基础研究,推动大数据的技术和应用,提升我国在相关领域的自主创新能力和核心竞争力,对推动经济转型,提升社会治理,增强我国科技竞争力具有至关重要的意义。

第三节　大数据带来的影响

大数据所具有的影响是多方面的,主要包括对人们思维方式的影响、对科学研究产生的影响,以及对社会发展产生的影响等。对思维方式产生影响的原因在于大数据所具有的

三个特征，即“全样而非抽样、效率而非精确、相关而非因果”，这是完全不同于传统思维方式的三大特征。对科学研究产生影响的原因在于，出现大数据之前，人们所做的各项研究只有3种范式，即实验、理论和计算；在出现大数据之后，增加了一种范式（即数据）。对社会发展产生影响的原因在于，在大数据的基础上所做出的决策已经变成了一种全新的用于决策的方式，并且在利用大数据的过程中，使各行各业都充分使用了信息技术。对于大数据的开发利用，就是在社会中创造新的先进技术以及建设新的应用。

在就业市场方面，大数据的兴起使得数据科学家成为热门人才；在人才培养方面，大数据的兴起将在很大程度上改变我国高校信息技术相关专业的现有教学和科研体制。

一、大数据对科学研究的影响

图灵奖获得者、著名数据库专家吉姆·格博士观察并总结认为，人类自古以来在科学研究上先后历经了实验、理论、计算和数据4种范式。在最初的科学研究阶段，人类采用实验解决一些科学问题，著名的比萨斜塔实验就是一个典型实例，如图1-2所示。

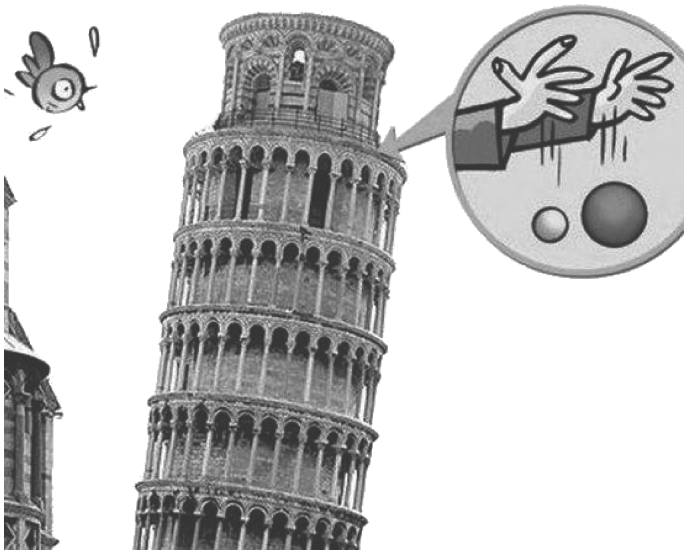

图1-2　比萨斜塔实验

关于物体下落的速度和物体重量之间的关系，亚里士多德认为二者之间应成正比例的关系，并提出了相关的理论学说，这一学说在当时的时代背景下得到了长达1900年的支持，直到1590年，伽利略在比萨斜塔上做了关于此方面的实验，并得出了两个不同重量的铁球依旧能同时落地的结论，才将亚里士多德提出的学说验证为一个错误的结论。

实验科学的研究会受到当时实验条件的限制，难以完成对自然现象更精确的理解。随着科学的进步，人类开始采用各种数学、几何、物理等理论构建问题模型和解决方案。例如，牛顿第一定律、第二定律、第三定律构成了牛顿力学的完整体系，奠定了经典力学的概念基础，它的广泛传播和应用对人们的生活和思想产生了重大影响，在很大程度上推动了人类社会的发展与进步。

随着1946年人类历史上第一台计算机NIAC的诞生，人类社会开始步入计算机时代，科学研究也进入了一个以“计算”为中心的全新时期。在实际应用中，计算科学主要用于对各个科学问题进行计算机模拟和其他形式的计算。通过设计算法并编写相应程序输入计算机运行，人类可以借助计算机的高速运算能力来解决各种问题。计算机具有存储容量大、运算速度快、精度高、可重复执行等特点，是科学研究的利器，推动了人类社会的飞速发展。

随着数据的不断累积，其宝贵价值日益得到体现，物联网和云计算的出现，更是促成了事物发展从量变到质变的转变，使人类社会开启了全新的大数据时代。此时，计算机不仅能进行模拟仿真，还能进行分析总结。在大数据环境下，一切将以数据为中心，从数据中发现问题并解决问题，真正体现数据的价值。大数据将成为科学工作者的宝藏，从数据中可以挖掘未知模式和有价值的信息，服务于生产和生活，推动科技创新和社会进步。虽然第三范式、第四范式都是利用计算机进行计算的，但是二者还是有本质的区别；在第三范式中，一般是先提出可能的理论，再收集数据，然后通过计算来验证；而在第四范式中，则是先有了大量已知的数据，然后通过计算得出之前未知的理论。

二、大数据对思维方式的影响

关于大数据对思维方式的影响在《大数据时代：生活、工作与思维的大变革》中有过相关的内容，该书是由维克托·迈尔·舍恩伯格编写的，书中强调了思维方式的三种转变，关于这三种转变的具体内容，下面将一一进行详细论述。

（一）全样而非抽样

全样和非抽样主要体现在对数据的处理和分析上。以前由于没有先进的处理数据的能力，因此，在对数据分析研究时，都是通过对数据抽样的方式进行研究，具体的操作方法是在全集的数据之中选取一部分数据，这部分数据就是样本数据，通过对样本数据展开研究和分析，从而判断全集数据具有怎样的特征。从规模上来看，全集数据的规模要大于样本数据的规模，因此，在使用样本数据时，要保证研究分析所要付出的代价是可以被控制的。在现在的大数据时代背景下，存在分布式文件系统以及分布式数据库技术，通过这两项内容，使存储数据的能力得到提升，而分布式并行编程框架，则提升了处理数据的能力，因此，在对数据进行处理和分析时，不再需要通过抽样来完成，并且在短时间内就可以得到分析数据的结果。

（二）效率而非精确

在以前的数据处理中，所使用的方法是抽样分析的方法，由于这种方法是通过部分来分析整体，因此要求对于数据的分析处理要做到十分精确，如果在样本中存在一个小小的误差，在全集数据中，这种误差被放大，从而变成一个很大的误差。想要保证存在误差的全集数据依旧可以被应用，就需要保证分析抽样数据的精确度。而在大数据的时代，对于数据的分析方法不再是抽样的方式，而是直接研究分析全集数据，在分析过程中所得到的误差并不会有被放大的情况发生，因此，这个时候对于数据的处理不再着重要求提高精确度，

而是提高对于数据分析处理的效率。对于数据分析的处理要求在几秒钟内就能得到实时的结果，这是充分发挥数据价值的一项重要内容，也是如今分析数据的一项重要的核心内容。

（三）相关而非因果

过去，数据分析的目的，一方面是解释事情背后的发展机理，例如，一个大型超市在某个地区的连锁店在某个时期内净利润下降很多，这就需要部门对相关销售数据进行详细分析并找出发生问题的原因；另一方面是用于预测未来可能发生的事件，例如，通过实时分析微博数据，当发现人们对雾霾的讨论明显增加时，就可以建议销售部门增加口罩的进货量，因为人们关注雾霾的一个直接结果是想要购买口罩来保障自己身体健康。无论是哪个目的，其实都反映了一种“因果关系”。但是，在大数据时代，因果关系不再那么重要，人们转而追求“相关性”而非“因果性”。例如，在淘宝网购物时，当人们购买了一个汽车防盗锁以后，淘网还会自动提示用户，与其购买相同物品的其他客户还购买了汽车坐垫。也就是说淘宝网只会告诉用户“购买汽车防盗锁”“购买汽车坐垫”之间存在相关性，但是并不告诉用户为什么其他客户购买了汽车防盗锁以后还会购买汽车坐垫。

三、大数据对社会发展的影响

大数据对于社会产生的影响主要表现在几个方面：一是在大数据的基础上所做出的决策已经变成了一种全新的用于决策的方式；二是利用大数据的过程，实际上就是各类行业充分使用信息技术的过程；三是对于大数据的开发和利用，实际上就是在社会中创造新的先进技术，并建设新的应用。

（一）大数据决策成为一种新的决策方式

根据数据制定决策，并非大数据时代所特有的方式。从20世纪90年代开始，数据仓库和商务智能工具就开始大量用企业决策。发展到今天，数据仓库已经是一个集成的信息存储仓库，既具备批量和周期性的数据加载能力，也具备数据变化的实时探测、传播和加载能力，并能结合历史数据和实时数据实现查询分析和自动规则触发，从而提供对战略决策（如宏观决策和长远规划等）和战术决策（如实时营销和个性化服务等）的双重支持。但是，数据仓库以关系型数据库为基础，无论是在数据类型还是在数据量方面都存在较大的限制。

如今最新的一种决策方式就是大数据决策，这种决策所针对的数据具有种类多且非结构化的特点，这种决策也是当前使用最为频繁的一种决策方式。例如，在政府部门就可以在“舆情分析”中使用大数据技术，并通过对多种不同来源的数据所开展的综合性分析，得到信息中的真实数据，将信息中所包含的隐藏内容发掘出来，并以此来推断事物未来的发展趋势，在帮助政府下决策的同时，对突发事件提供有效的应对措施。

（二）大数据应用促进信息技术与各行业的深度融合

之前就有相关的专家提出，大数据的存在对于社会中任何一个行业所具有的业务功能

都会产生影响，尤其是在互联网、交通、服务、银行等行业中，由于大数据在不断累积，从而推动信息技术在这些行业领域中得到应用，并为行业的发展提供了新方向。例如，在物流行业中，通过大数据的分析，可以帮助快递公司选择运输成本最低的路线进行运输；在股票投资领域，通过大数据的分析，可以帮助投资者选择获得最多利益的股票投资方式；在零售行业，通过大数据的分析，可以帮助商户准确地找到目标群体等。总体而言，存在大数据的地方，就会对人们的生产活动以及生活产生深远的影响。

（三）大数据开发推动新技术和新应用的不断涌现

之所以对大数据方面的新技术不断进行开发，是因为社会对于使用大数据的需要。为了满足不同的应用需求，就需要开发相关的大数据技术，并将其充分地利用起来，对于大数据技术的使用也是发挥数据价值的过程，并且这种应用将会不断取代通过人工进行判断的应用。举个例子来说，以前的保险公司在对客户提供关于汽车方面的保险时，需要先通过车主的信息，人工对不同的客户所属的类别进行划分，再根据客户的车险次数，向其提供比较合适的保险方案，同时，对于客户来说，所有的保险公司都是用这种方式提供服务，因此，客户选择哪一家保险公司进行投保差别并不大。但是当大数据出现后，保险公司的商业模式发生了改变，能够充分利用大数据信息，得到更多关于客户车辆信息的细节内容，就可以为客户提供具有针对性和个性化的保险方案，从而提高该公司在行业内的竞争力度，客户对于保险公司也有了更多的选择。

四、大数据对就业市场的影响

大数据的兴起使数据科学家成为热门人才。2010年，高科技劳动力市场上还很难见到数据科学家的头衔，但此后，数据科学家逐渐发展成为市场上最热门的职位之一，具有广阔的发展前景，并代表着未来的发展方向。

互联网企业和零售、金融类企业都在积极争夺大数据人才，数据科学家成为大数据时代最紧缺的人才。据麦肯锡公司预测，有大数据专家估算过，5年内国内的大数据人才缺口会达到130万，以大数据应用较多的互联网金融为例，这一行业每年增速达到4倍，届时，仅互联网金融需要的大数据人才就是现在需求的4倍以上。与此同时，大数据人才的薪资水平也在“水涨船高”，根据第四届贵州人才博览会发布的《全国大数据人才需求指数报告》显示，2016年2月，贵阳大数据人才月薪已逼近8000元。

根据中桥调研咨询2013年7月针对中因市场的一次调研结果显示，中国用户目前还主要局限在结构化数据分析方面，尚未进入对半结构化和非结构化数据进行分析，以及捕捉新的市场空间的阶段。但是，大数据包含了大量的非结构化数据，未来将会产生大量针对非结构化数据分析的市场需求，因此，未来中国市场对掌握大数据分析专业技能的数据科学家的需求会逐年递增。

尽管有少数人认为未来有更多的数据会用自动化处理，会逐步降低对数据科学家的需求，但是仍然有更多的人认为，随着数据科学家给企业所带来的商业价值的日益体现，市场对数据科学家的需求会越发旺盛。

五、大数据对人才培养的影响

在我国一些设有信息技术相关专业的高校中，大数据的出现对于学校的教学模式、教学内容以及科研等方面都产生了较大的影响，这种影响主要表现在两个方面：一方面是对于数据科学家人才的培养，另一方面是培养这类人才所需要的环境。从数据科学家的能力来看，他们需要掌握的能力包括数学能力、编程能力、统计能力以及机器学习能力等，如果将这类人才进行归类，他们应属于复合型人才类型。目前，关于数据科学，只在一些相关的专业中能够学到一部分的知识内容，还没有一个专门的数据科学专业，因此，对于这类复合型人才的培养还存在一些缺陷。对于数据科学家来说，最重要的就是有应用大数据的环境，只有在真正的大数据环境中开展实践活动或进行学习，才能真正有效地掌握大数据，并且想要发掘出数据中所蕴含的具有价值的信息内容，就需要将业务需求同技术背景相结合。但是我国的许多高校，目前仍没有大规模的基础数据，对于业务需求也并没有太多的认识。

鉴于上述两个原因，目前国内的数据科学家人才并不是由高校培养的，而主要是在企业实际应用环境中通过边工作边学习的方式不断成长起来的，其中，互联网领域集中了大多数的数据科学家。在未来5～10年，市场对数据科学家的需求会日益增加，不仅互联网企业需要数据科学家，类似金融、电信这样的传统企业在大数据项目中也需要数据科学家。由于高校目前尚未具备大量培养数据科学家的基础和能力，传统企业很可能会从互联网行业“挖墙脚”，以此来满足企业发展对数据分析人才的需求，继而造成用人成本高昂，制约企业的成长壮大。因此，高校应该继承“养人才、服务社会”的理念，充分发挥科研和教学综合优势，培养一大批具备数据分析基础能力的数据科学家，以有效缓解数据科学家的市场缺口，为促进经济社会发展做出更大贡献。目前，国内很多高校开始设立大数据专业或者开设大数据课程，加快推进大数据人才培养体系的建立。2014年，中国科学院大学开设了首个“大数据技术与应用”专业，面向科研发展及产业实践，培养信息技术与行业需求结合的复合型大数据人才；同样是在2014年，清华大学成立了数据科学研究院，推出了多学科交叉培养的大数据硕士项目。2015年10月，复旦大学大数据学院成立，在数学、统计学、计算机、生命科学、医学、经济学、社会学、传播学等多学科交叉融合的基础上，聚焦大数据学科建设、研究应用和复合型人才培养；2016年9月，华东师范大学数据科学与工程学院成立，新设置的本科专业“数据科学与工程”是华东师范大学除“计算机科学与技术”“软件工程”以外，第三个与计算机相关的本科专业。此外，厦门大学于2013年开始在研究生层面开设大数据课程，并建设了国内高校首个大数据课程公共服务平台。

高校培养数据科学家需要采取“两条腿走路”的策略，即“引进来”“走出去”。所谓“引进来”，是指高校要加强与企业的紧密合作，从企业引进相关数据，为学生搭建起接近企业应用实际的、仿真的大数据战略环境，让学生有机会理解企业业务需求和数据形式，为开展数据分析奠定基础；同时，从企业引进具有丰富实战经验的高级人才，承担起数据科学家相关课程教学任务，切实提高教学质量、水平和实用性。所谓“走出去”，是指积极鼓励和引导学生走出校园，进入互联网、金融、电信等具备大数据应用环境的企业去开展实践活动，同时努力加强产、学、研合作，创造条件让高校教师参与到企业大数据项目中，实现理论知识与实际应用的深层次融合，锻炼高校教师的大数据实战能力，为更好地培养数据科学家奠

定基础。

在课程体系的设计上，高校应该打破学术界限，设置跨院系跨学科的“组合课程”，由来自计算机、数学、统计等不同院系的教师构建联合教学师资力量，多方合作，共同培养具备大数据分析基础能力的数据科学家，使其全面掌握包括数学、统计学、数据分析、商业分析和自然语言处理等在内的系统知识，具有独立获取知识的能力，并具有较强的实践能力和创新意识。

第四节　大数据技术的应用

大数据已成为现代社会的流行语，大数据技术给生产生活带来了天翻地覆的变化，带来了时代的变革。然而，实际上很多人对大数据的应用模糊不清。下面将从大数据应用案例来介绍最真实的大数据故事，以及大数据在生活当中实际应用的情况。

一、电视媒体

对于体育爱好者而言，追踪电视播放的最新运动赛事几乎是一件不可能的事情，因为有超过百个赛事在8000多个电视频道播出。

而现在市面上开发了一种可追踪所有运动赛事的应用程序RUWT，它可以在iOS设备、Android设备及Web浏览器上使用，它通过不断地分析运动数据流，让体育迷知道他们应该转换到哪个台看喜欢的节目，并使他们能够在比赛中进行投票。

RUWT程序的主要功能就是将各类赛事进行评分和排名，其评分的依据主要在于赛事的激烈程度，观众可以依据这些排名选择观看精彩的赛事，并进入该赛事所在的频道。

二、医疗行业

沃森技术是用于分析和预测医疗保健内容的一项使用IBM的技术，使用该技术的第一个客户是Seton Healthcare，通过该技术可以获得更多关于患者的临床医疗信息，再加上经过大数据的处理，有利于对患者的信息进行分析。例如，在多伦多的一家医院中，对早产儿的数据信息读取的速度要大于每秒3000次，并对这些数据信息进行分析，医院根据分析的结果，可以较早地了解到存在问题的早产儿，再根据问题的具体情况制定相应的解决措施，保证婴儿不会因为早产而夭折。

一些创业者还会借助于大数据开发许多的产品。一些健康类的应用，其数据的收集主要来自社交网络，在未来，这些应用非常有可能会被医生使用于对患者的诊断中，也会使诊断的结果更准确。例如，关于药品的用量，可能不再是每天固定的次数和每次固定的数量，而是通过检测患者血液中的药剂含量，在药剂吸收和代谢完成之后，应用会自动提醒患者进行再次服药。

美国的一家处方药管理服务公司Express Scripts，该公司所掌握的处方有1.4亿个，所掌握的信息包括1亿的美国人及65000家药店。该公司正是利用数据上的优势，开始利用一系列比较复杂模型对各类药品进行检测，并判断该药品是否为虚假药品，同时，该模型还可用于提醒人们停止用药。该公司既能对潜在问题进行识别，还能利用数据信息对问题进行解决，尤其以前出现过的问题。Express Scripts对于医生开出的处方进行分析，可以判断出处方中的药物属于哪一类，同时，还能够记录每一位医生所拥有的患者对其进行的评价。而一位医生是不是值得信赖，就可以根据该医生是否有红色旗帜的标志来进行判断。

三、保险行业

从技术创新的角度来说，保险行业并不具备引领的作用，但是美国的一家保险公司MetLife依旧为建立一个全新的系统投入了3亿美元，该公司设计的第一款产品所使用的应用程序为MongoDB，在该程序中包含了70多个数据，并且这些数据属于遗留系，通过对数据的合并，形成一个单一的记录，而客户的信息也被该程序存放在了一起。该程序需要通过两个数据中心的六台服务器来保证其运行，并且已经存储了24TB的数据信息。所有MetLife的美国客户的信息都被存储在了该程序中，并且这些数据信息不断再进行实时的更新，只要有新的数据被输入进来，就会被立即保存进系统。

虽然大部分的疾病都可以通过服用药物进行治疗并取得一定的效果，但是在大数据的时代，依旧希望可以通过一些干预项目对患者的健康状况进行调整，而寻找愿意参与干预项目的患者以及专注于该项目的医生并不是一件容易的事情。率先有所举措的是安泰保险，为了完成尝试，该保险公司选择了102名患有代谢综合征的患者，其实验的最终目的是降低患者的发病率。其实验的方法是先将患者三年内的化验结果以及理赔事件进行扫描，之后结合患者检测试验的结果，组合成一个治疗方案，该治疗方案的个性化程度较高，并对危险因素和重点的治疗方案进行评估，最后给出相关的治疗建议，例如，服用他汀类的药物以及减重2.27千克左右等，以此来降低患者在以后的10年之内50％的发病概率。

四、职业篮球赛

专业的篮球队对于赛事的分析主要是通过对数据的收集来完成的，但是，对于整理数据和分析数据背后的意义是存在困难的。Krossover公司就努力地通过数据的分析结果，找到球队可以赢得比赛的关键，或者找到可以在比赛中获得较高分数的方法。其主要方法就是分解篮球队教练上传的比赛视频，教练可以在分解后的视频中找到自己想要的信息，例如一些统计数据、在比赛中球员的一些表现等。对于比赛视频的分析，就是在对所有可量化的数据进行分析。

五、能源行业

在欧洲已经出现了智能电表，实际上，这是智能电网的终端。在德国，除了一些自用电

的家庭，大部分的家庭都会装有太阳能发电板，通过利用太阳能完成发电，为了鼓励这一行为，供电公司还会将多余的电力买走。电网每隔5分钟或者10分钟，就会通过电网收集用电的数据，从而预测每个家庭具体的用电习惯，并推测出在未来的几个月中，整个电网所需要的电量是多少，再根据预测的结果向供电公司购买电量，而购买电量的价格会随着购买的时间发生变化，越是提前购买所需要的价格就越便宜，因此，有了预测之后，可以减少购买电量所需要的成本。在风能源中，也同样可以用到大数据，在维斯塔斯风力系统中寻找最适合安装风力涡轮机的位置，以及风电场所在的位置，主要是通过分析气象数据来完成的，而分析的方法则是通过BigInsights软件以及IBM超级计算机来进行的。同时，分析工作所需要的时间，通过使用大数据而减少了。

六、公路交通

在洛杉矶，交通拥堵的情况十分严重，美国政府为了改善这一情况，同施乐公司一同在I-10和I-110州际公路上建立了一条收费快速通道，并利用大数据引导司机行驶在该通道上，其维持正常交通秩序的办法是通过动态定价以及使用Express Lanes等方法来实现的。纳泰什·曼尼科作为施乐公司的手下技术执行官，设计了一款高占用收费系统，该系统规定，司机想要驾驶在热车道上，就需要将车速控制在每小时72.4千米左右，而当交通出现拥堵的情况时，收费标准会发生改变，将私家车支付的价格提升，从而减少私家车在该车道上的行驶，并提供给类似于公共汽车或者大巴车等对车道具有高占用率的行驶车辆。

施乐公司还开发了另一个项目叫Express Park，该项目主要用于帮助司机寻找停车的场地以及停车所要交的费用是多少，因此，该项目最重要的设计，就是保证用户收到的数据是实时的。

七、汽车制造

人们对于汽车制造的流程依旧认为是依靠各类的生产装配流水线上以及不同类型的制造机器来完成的。在美国的福特公司，对于汽车的制造，从研发设计的阶段就早已开始对大数据进行应用了。例如，关于SUV车后行李箱车门的问题，福特的开发团队就采用手动还是电动进行了详细的分析，并发现有很多人都十分在意这个问题。经过分析发现，不同的方式都有其自己的优缺点，如果采用手动的方式，其缺点是不够方便和智能；而采用电动的方式，其缺点是车门的开启十分有限。

八、零售业

有一家在专业时装领域比较领先的销售商，其向客户提供服务的方式是依靠当地的商店以及网络等，如果该销售商想向客户提供差异化的服务，就需要从社交网站上收集到用户的社交信息，并了解更多的关于化妆品的营销模式，进而发现具有价值的客户一共有两类：一类为高消费者，另一类为高影响者。想让客户为销售商进行宣传，可以通过为他们提

供免费化妆的服务来实现。以上的这种方案，就是在将交易数据与交互数据融合之后所得到的。

这家销售商为了提高自己的服务水平，并使其更具有目标性，其主要使用的是Informatica技术。该技术最主要的作用就是将客户的主数据进行补充，并辅助销售商了解自己的客户在店内的情况和商品之间具有怎样的互动，再通过对得到的数据信息进行分析，根据其分析的结果，为商品的销售、商品的摆放以及销售的价格等方面提出建议。

这种方式在某零售企业中已经有所使用，不仅使该零售企业的存货下降了17%，使自有品牌的商品所占有的利润比例增加，还保证了该企业所占有的市场份额。

九、电子邮件

MailChimp为用户提供的核心服务就是电子邮件服务，使用该邮件的用户大约有300万人次，在一年内，其提供服务的邮件数量高达350亿封，而该平台最大的价值就在于处理这些邮件数据，并对这些数据进行分析。在MailChimp中有一项名为Wavelength的服务，提供该服务的主要目的在于帮助用户了解自己发出的邮件信息，如果有类似的信息内容，也会为用户展示出来，简单来说就是在向用户说明其他的用户都订阅了哪些邮件、查看了什么样的邮件以及浏览了哪些链接等。不同的邮件地址之间所产生的互动都会被记录在公司的数据库之中，其存储的位置就在Wavelength中，而上述的服务就是这样实现的。在MailChimp中存在的另一个功能是Ecommerce360，该功能主要辅助用户进行跟踪点击，并通过转换的方式实现。

十、基于Hadoop平台的电信客服数据的处理与分析

通信运营商每时每刻会产生大量的通信数据，例如通话记录、短信记录、采集记录、第三方服务资费等众多信息。数据量巨大，除了要满足用户的实时查询和展示之外，还需要定时定期地对已有数据进行离线的分析处理。例如，当日话单、月度话单、年度话单、通话详情等，项目以此为背景。模拟电信客服产生的通信日志数据，使用Flume采集数据到Kafka集群，然后提供给HBase消费，并通过协处理器对HBase进行优化。将HBase数据导入MySQL当中，并从多个维度对通话记录进行分析，最后通过数据可视化技术，以图形、图表的形式展示给用户。

第五节　大数据相关技术

数据本身没有价值，我们需要对数据进行加工处理，让数据产生价值。对数据进行加工处理的技术就是大数据技术，从大数据使用流程上来说，是指伴随着大数据的采集、存储、分析和应用的相关技术，是一系列使用非传统的工具来对大量的结构化、半结构化和非

结构化数据进行处理，从而获得分析和预测结果的一系列数据处理和分析技术。

本节重点介绍大数据分析全流程所涉及的各种技术，包括数据采集、数据清洗、数据存储和管理、数据处理与分析、数据可视化等。

一、数据采集

数据采集是大数据产业的基石，如果没有数据采集，大数据中的数据也就失去了来源，数据价值也就无从谈起。数据采集又称“数据获取”，是数据分析的入口，它通过各种技术手段把外部各种数据源产生的数据进行采集并加以利用。在数据大爆炸的时代，被采集的数据类型是复杂多样的，包括结构化数据、半结构化数据、非结构化数据。

数据采集的主要数据源包括传感器数据、互联网数据、日志数据、数据库数据等。

1. 传感器数据

传感器是一种检测装置，能感受到被测量的信息，并能将感受到的信息按一定规律变换成电信号或其他形式的信息输出，以满足信息的传输、处理、存储、显示、记录和控制等要求。传感器数据是指由传感设备收集和测量的数据。在工业或农业生产中经常用到的有超声波传感器、温度传感器、湿度传感器、气体传感器、气体报警器、压力传感器等。

2. 互联网数据

互联数据采集通过网络爬虫或网站公开API等方式从网站上获取数据信息。该方法可以将非结构化数据从网页中抽取出来，将其存储为统一的本地数据文件，并以结构化的方式存储。它支持图片、音频、视频等文件或附件的采集，附件与正文可以自动关联。网络爬虫示意图如图1-3所示。

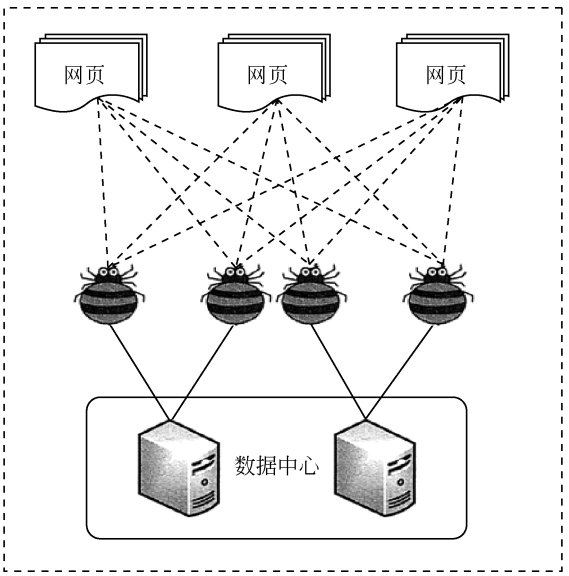

图1-3　网络爬虫示意图

3. 日志数据

许多公司的业务平台每天都会产生大量的日志数据。对于这些日志数据，我们可以从

中得到很多有价值的信息。通过对这些日志数据进行日志采集、收集，然后进行数据分析，挖掘公司业务平台日志数据中的潜在价值，从而为公司决策和公司后台服务器平台性能评估提供可靠的数据保证。

系统日志采集系统做的事情就是收集日志数据提供离线和在线的实时分析使用。目前常用的开源日志收集系统有Apache Flume、Scribe等。其中，Apache Flume是一个分布式、可靠、可用的服务，用于高效地收集、聚合和移动大量的日志数据，它具有基于流式数据流的简单灵活的架构，其可靠性机制和许多故障转移和恢复机制，使Apache Flume具有强大的容错能力。

4. 数据库数据

一些企业会使用传统的关系型数据库MySQL和Oracle等来存储数据，除此之外，Redis和MongoDB这样的NoSQL数据库也常用于数据的存储。企业每时每刻产生的业务数据，以数据库记录的形式被直接写入数据库中。

通过数据库采集系统直接与企业业务后台服务器结合，将企业业务后台数据抽取、转换、加载到企业数据仓库中，以供后续的商务智能分析使用。

二、数据清洗

数据清洗就是把“脏”的数据“洗掉”，这是发现并纠正数据文件中可识别的错误的最后一道程序，包括检查数据一致性，处理无效值和缺失值等。因为数据仓库中的数据是面向某一主题的数据集合，这些数据从多个业务系统中抽取而来并包含历史数据，这样就避免不了有的数据是错误数据，有的数据相互之间有冲突，这些错误的或有冲突的数据显然不是我们想要的，因此称为“脏数据”。我们要按照一定的规则把“脏数据”从大数据中“洗掉”，这就是数据清洗。而数据清洗的任务是过滤那些不符合要求的数据，将过滤的结果交给业务主管部门，确认是过滤掉，还是由业务单位修正之后再进行抽取。不符合要求的数据主要有不完整的数据、错误的数据、重复的数据三大类。数据清洗可以通过专业的数据清洗工具完成，常用的数据清洗工具有DataPipeline、Kettle、Talend、Informatica、Datax、Oracle Goldengate等。

三、数据存储和管理

采集后的数据需要存储起来，才能方便后续对数据的持续使用。常用的大数据存储主要以下三种。

（一）分布式文件系统

计算机通过文件系统管理、存储数据，而信息爆炸时代中人们可以获取的数据成指数级别的增长，单纯通过增加硬盘个数来扩展计算机文件系统的存储方式，在容量大小、容量增长速度、数据备份、数据安全等方面的表现都差强人意。分布式文件系统可以有效解决数据的存储和管理难题：将固定于某个地点的某个文件系统，扩展到任意多个地点/多个文

件系统，众多的节点组成一个文件系统网络。每个节点可以分布在不同的地点，通过网络进行节点间的通信和数据传输。人们在使用分布式文件系统时，无须关心数据是存储在哪个节点上，或者是从哪个节点从获取的，只需要像使用本地文件系统一样管理和存储文件系统中的数据。分布式文件存储如图1-4所示。

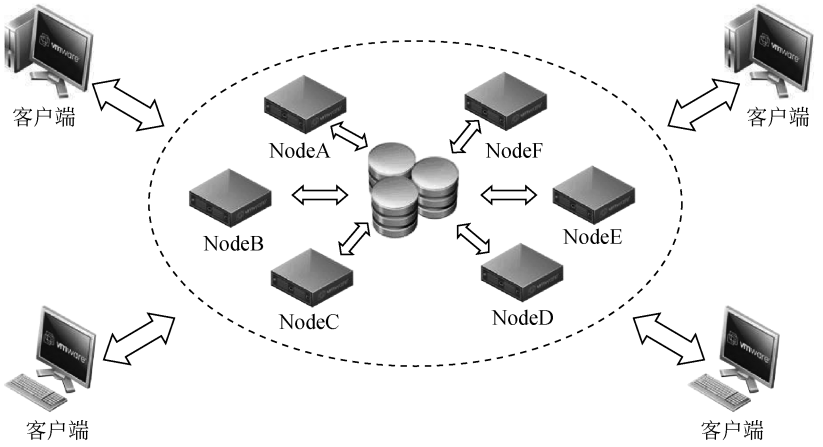

图1-4　分布式文件存储示意图

(二) NoSQL数据库

传统关系型数据库在密集型数据的处理及应用方面显得力不从心，主要表现在灵活性差、扩展性差、性能差等方面。最近出现的一些存储系统摒弃了传统关系型数据库管理系统的设计思想，转而采用不同的解决方案来满足扩展性方面的需求。这些没有固定数据模式并且可以水平扩展的系统现在统称为NoSQL，这里的NoSQL指的是“not only SQL”，即对关系型SQL数据系统的补充。

相对于关系型数据库，NoSQL数据存储管理系统的主要优势有以下几点。

（1）避免不必要的复杂性。关系型数据库提供各种各样的特性和强一致性，但是许多特性只能在某些特定的应用中使用，大部分功能很少被使用。NoSQL系统则提供较少的功能来提高数据系统的性能。

（2）高吞吐量。一些NoSQL数据系统的吞吐量比传统关系型数据库管理系统要高很多，如Google使用MapReduce每天可处理20PB存储在Bigtable中的数据。

（3）高水平扩展能力和低端硬件集群。NoSQL数据系统能够很好地进行水平扩展，与关系型数据库集群方法不同，这种扩展不需要很大的代价。而基于低端硬件的设计理念为采用NoSQL数据系统的用户节省了很多硬件上的开销。

（4）避免了昂贵的对象—关系映射。许多NoSQL系统能够存储数据对象，这就避免了数据库中关系模型和程序中对象模型相互转化的代价。

(三) NewSQL数据库

虽然NoSQL数据库具有高可用性和可扩展性，但它放弃了传统SQL的强事务保证和关系模型，不保证强一致性。这对于普通应用没问题，但还是有不少像金融机构一样的企业级应用有强一致性的需求。NewSQL提供了与NoSQL相同的可扩展性，而且仍基于关系

模型，还保留了极其成熟的SQL作为查询语言，保证了ACID事务的特性。目前主流的NewSQL数据库包括VoltDB、ClustrixDB、MemSQL、ScaleDB、TiDB。

四、数据处理与分析

常用的数据处理与分析技术主要有批处理计算、流计算、图计算、查询分析计算。

（一）批处理计算

批处理计算主要解决针对大规模数据的批量处理，也是日常数据分析工作中常见的一类数据处理需求。MapReduce是最具有代表性和影响力的大数据批处理技术，可以并行执行大规模数据处理任务，用于大规模数据集（大于1TB）的并行运算。MapReduce极大地方便了分布编程工作，它将复杂的、运行于大规模集群上的并行计算过程高度地抽象成了两个函数——Map和Reduce。这样编程人员在不会分布式并行编程的情况下，可以很容易地将自己的程序运行在分布式系统上，完成海量数据集的计算。MapReduce运行过程如图1-5所示。

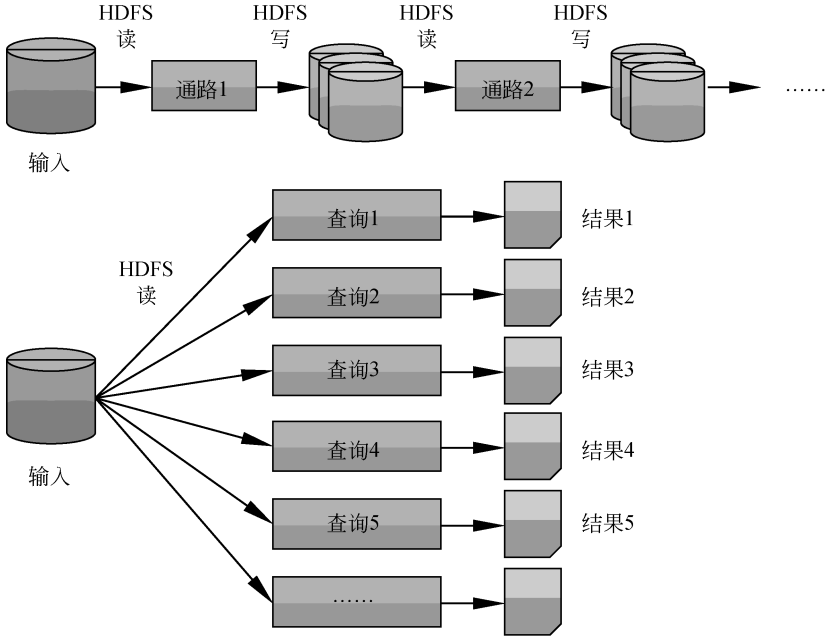

图1-5　MapReduce运行过程

Spark是一个针对超大数据集合的低延迟的集群分布式计算系统，比MapReduce快很多。Spark启用了内存分布数据集，除了能够提供交互式查询外，还可以优化迭代工作负载。在MapReduce中，数据流从一个稳定的来源，进行一系列加工处理后，流出到一个稳定的文件系统（如HDFS）。而对于Spark而言，则使用内存替代HDFS或本地磁盘来存储中间结果，因此，Spark要比MapReduce的速度快很多。Spark运行过程如图1-6所示。

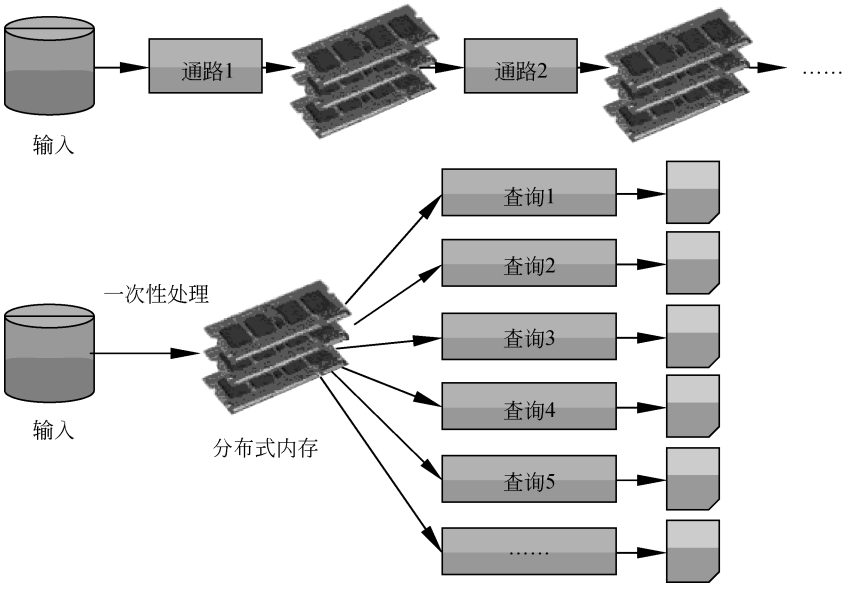

图 1-6　Spark 运行过程

（二）流计算

近年来，在 Web 应用、网络监控、传感监测等领域，兴起了一种新的数据密集型应用——流数据，即数据以大量、快速、时变的流形式持续到达。流计算可以实时获取来自不同数据源的海量数据，经过实时分析处理，获得有价值的信息。流计算过程如图 1-7 所示。

流计算秉承一个基本理念，即数据的价值随着时间的流逝而降低，如用户点击流。因此，当事件出现时就应该立即进行处理，而不是缓存起来进行批量处理。为了及时处理流数据，就需要一个低延迟、可扩展、高可靠的处理引擎。

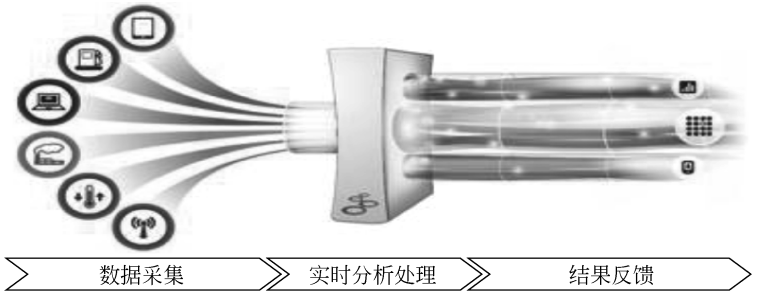

图 1-7　流计算示意图

目前业内已涌现出了许多的流计算框架与平台，第一类是商业级的流计算平台，包括 IBM InfoSphere Streams 和 IBM StreamBase 等；第二类是开源流计算框架，包括 Twitter Storm（免费、开源的分布式实时计算系统，可简单、高效、可靠地处理大量的流数据）、Yahoo S4（开源流计算平台，是通用、分布式、可扩展、分区容错、可插拔的流式系统）；第三类是公司为支持自身业务开发的流计算框架，如 Facebook 使用 Puma 和 HBase 相结合来处理实时数据，百度开发了通用实时流计算系统 DStream，淘宝开发了通用流数据实时计算系统——银河流数据处理平台。

（三）图计算

在大数据时代，许多大数据都是以大规模图或网络的形式呈现，如社交网络、传染病传播途径、交通事故对路网的影响等，此外，许多非图结构的大数据，也常常会被转换为图模型后进行分析。MapReduce作为单输入、两阶段、粗粒度数据并行的分布式计算框架，在表达多迭代、稀疏结构和细粒度数据时，往往显得力不从心，不适合用来解决大规模图计算问题。因此，针对大型图的计算，需要采用图计算模式，目前已经出现了不少相关图计算的产品。Pregel是Google提出的大规模分布式图计算平台，专门用来解决网页链接分析、社交数据挖掘等实际应用中涉及的大规模分布式图计算问题。

（四）查询分析计算

针对超大规模数据的存储管理和查询分析，需要提供实时或准时的响应，才能很好地满足企业经营管理需求。Dremel是Google的“交互式”数据分析系统，可以组建成规模上千的集群，处理PB级别的数据。MapReduce处理一个数据需要分钟级的时间。作为MapReduce的发起人，Google开发了Dremel，将处理时间缩短到秒级，作为MapReduce的有力补充。最近Apache计划推出Dremel的开源实现Drill，将Dremel的技术又推到了浪尖上。

五、数据可视化

在大数据时代，人们面对海量数据，有时难免显得无所适从。一方面，数据复杂繁多，各种不同类型的数据大量涌来，庞大的数据量已经大超出了人们的处理能力，在日益紧张的工作中已经不允许人们在阅读和理解数据上花费大量时间；另一方面，人类大脑无法从堆积如山的数据中快速发现核心问题，必须有一种高效的方式来刻画和呈现数据所反映的本质问题。要解决这个问题，就需要数据可视化，它通过丰富的视觉效果，把数据以直观、生动、易理解的方式呈现给用户，可以有效提升数据分析的效率和效果。

数据可视化是让用户直观了解数据潜藏的重要信息，有助于帮助用户理解并分析数据。常用的可视化工具主要有四类。

（一）在线可视化工具

在线可视化工具主要有镝数、花火等，优点是图表种类丰富、类型新颖、配色年轻化，还提供了一些十分酷炫动态图表，操作也比较简单，很多新媒体都在用；缺点是数据保密性不够。

（二）编程可视化工具

编程可视化工具主要有E-charts、D3、ggplot、Matplotlib、pandas、plt等，优点是可以制作大型数据集和交互动画的图表，高端、大气、上档次，可视化效果非常好；缺点是需要有编程基础，门槛较高。

（三）商业智能工具

在商业智能工具方面，如国内比较知名的FineBI等，是专业的大数据 BI 和分析平台，主要为企业提供一站式商业智能解决方案，用它们做数据分析和可视化驾驶舱不需要写代码，而且操作比较方便；缺点是目前市场上的大部分BI都收费，不过FineBI个人版免费，这一点算是比较人性化。

（四）基础可视化工具

基础可视化工具主要指Excel，优点是通用、易用、实用，傻瓜式操作，基本上人人都会，使用成本较低，同时还有基于Excel开发了图表插件Thinkcell Chart、Zebra Bi，国内开发的Easyshu，都可以高效地制作出商业图表；缺点是Excel本身主要制作常规性的图表，很多特殊图表无法实现，功能强大的图表插件价格不菲。

习　　题

一、选择题

1. 当前大数据技术的基础是由(　　)公司首先提出的。

A.微软　　B.百度　　C.谷歌　　D.阿里巴巴

2. 大数据的起源是(　　)。

A.金融　　B.电信　　C.互联网　　D.公共管理

3. 大数据最显著的特征是(　　)。

A.数据规模大　　B.数据类型多样

C.数据处理速度快　　D.数据价值密度高

4. 下列关于舍恩伯格对大数据特点的说法中，错误的是(　　)。

A.数据规模大　　B.数据类型多样

C.数据处理速度快　　D.数据价值密度高

5. 当前社会中，最为突出的大数据环境是(　　)。

A.综合国力　　B.物联网　　C.自然资源　　D.互联网

二、简答题

1. 简述大数据技术的特点。

2. 简述科学研究的第一至第四范式。