

第1章

大数据概述

当今是一个信息化、数字化的社会,互联网、物联网、人工智能(AI)和云计算等技术的迅猛发展,海量数据也成为一种新的自然资源,亟待人们对其合理、高效、充分地利用,使之能够给人们的生活和工作带来更大的效益和价值。总之,大数据时代的来临,带来了信息技术发展的巨大变革,并深刻影响着社会生产和人们生活的各个方面。大数据是国家基础性战略资源,也是经济社会发展的重要基础。本章主要介绍大数据的基本概念,关键技术简介和应用领域等。

1.1 数据和大数据

1.1.1 数据定义

数据(data)是对客观事物的符号表示。数据是指某一目标定性、定量描述的原始资料,包括数字、文字、符号、图形、图像以及它们能转换成的各种形式。数据是用以承载信息的物理符号,其本身并没有意义。数据在计算机科学中是指所有能输入计算机中并被计算机程序处理的符号总称。它可以是文字、图形、图像或声音。数据的基本单元称为数据项,可以按目的来组织数据结构。数据的格式往往与计算机系统有关,并随承载它的物理设备的形式而改变。

1.1.2 数据与信息/知识/智慧的关系

数据是现实世界的记录。信息源自数据,信息是有意义的信息,是数据的内涵。知识则是在信息的基础上,通过人类实践后得到的对客观世界的总结。智慧是基于知识和智能的总结,其中知识是一切智能行为的基础,而智能是获取知识并运用知识求解问题的能力,是头脑中思维活动的具体体现。智能是指个体对客观事物进行合理分析和判断后自适应地对变化的环境进行响应的一种能力。从图 1-1 所示的 DIKW 金字塔(data information knowledge wisdom pyramid)可以看出,从“数据”到“智慧”的认识转变过程,也是“从部分到整体,并描述过去、认识现在、预测未来”的过程。

1.1.3 大数据的定义

“大数据”的概念起源于 2008 年 9 月《自然》(*Nature*)杂志刊登的名为“Big Data”的专

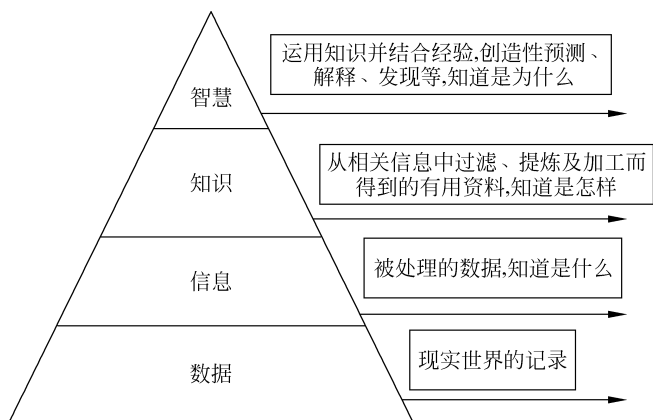


图 1-1 DIKW 金字塔

题,继而迅速得到科学、计算机、经济等不同领域专家的响应。不同于传统数据,大数据结构复杂,并已在 21 世纪成为一个热门的话题,现在大数据的重要性已得到普遍认可,但不同领域对大数据的解读却不相同,目前对大数据还没有公认的定义。下面主要从三个不同角度对大数据进行解读。

定义 1: 强调处理能力。认为大数据是指所涉及的数据量规模巨大,无法通过人工在合理时间内截取、管理、处理并整理成为人类所能解读的信息。

定义 2: 强调应用方法。认为利用先进技术可以从大数据中获取信息。把大数据看成一种方法,即不用随机分析法(抽样调查)的捷径,而采用所有数据的方法。

定义 3: 侧重应用价值。认为大数据是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力来适应海量、高增长率和多样化的信息资产。

综上所述,大数据相对传统数据是一个新概念,也是一种新技术,一种针对企业问题的解决方案,一种对数据形态的诠释;大数据时代对人类的数据驾驭能力提出了新的挑战,也为人们获得更为深刻、全面的洞察能力提供了前所未有的空间与潜力。

1.2 数据科学

1.2.1 关于数据科学

数据科学是一门以“数据时代”,尤其是以“大数据时代”面临的新挑战、新机会、新思维和新方法为核心内容的,包括新的理论、方法、模型、技术、平台、工具、应用和最佳实践在内的一整套知识体系。数据科学是一门以实现“从数据到信息”“从数据到知识”和“从数据到智慧”的转化为主要研究目的,以“数据驱动”“数据业务化”“数据洞见”“数据产品研发”和“数据生态系统的建设”为主要研究任务的独立学科。它是一门将“现实世界”映射到“数据世界”之后,在“数据层次”上研究“现实世界”的问题,并根据“数据世界”的分析结果,对“现实世界”进行预测、洞见、解释或决策的新兴科学。数据科学是以“数据”尤其是“大数据”为研究对象,并以数据统计、机器学习、数据可视化等为理论基础,主要研究数据预处理、数据管理、数据计算等内容的交叉学科。

1.2.2 数据科学的基本流程

从整体来看,数据科学所涉及的基本流程如图 1-2 所示,主要包括数字化、数据处理、探索性分析、数据分析与洞见、结果展现和提供数据产品。

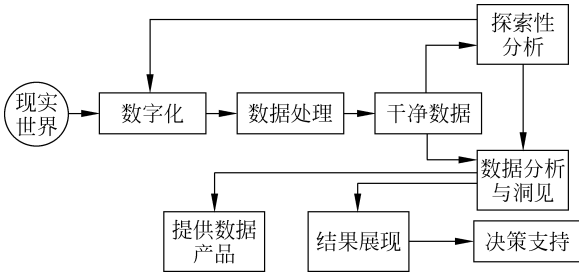


图 1-2 数据科学的基本流程

- (1) 数字化：从现实世界中收集原始数据。
- (2) 数据处理：将原始数据转换为“干净数据”。
- (3) 探索性分析：在无(或较少)先验假设的前提下,采用作图、制表、方程拟合、比较特征量等手段,初步探索数据的结构和规律,为数据分析提供依据和参考。
- (4) 数据分析与洞见：根据“干净数据”本身的特点和“探索性分析”的结合,设计/选择/应用具体的机器学习算法/统计模型进行数据分析。
- (5) 结果展现：在机器学习算法/统计模型的设计与应用的基础上,采用数据可视化、故事描述等方法将数据分析的结果展示给最终用户,提供“决策支持”。
- (6) 提供数据产品：在机器学习算法/统计模型的设计与应用的基础上,还可以进一步将“干净数据”转换成各种“数据产品”,并提供给“现实世界”进行交易与消费。

1.2.3 数据科学家常用的工具

由于大数据具有体量巨大、价值密度低的特点,大数据计算的核心问题就是如何从海量数据中挖掘出更有价值的信息。根据流数据和存档数据的特点,研发适合各自特点的大数据计算能力来支持各种业务需求。其中最突出的需求包括针对存档数据的统计分析、模式分析和可视化;针对流数据的连续分析与连续响应(相关计算、分析将在后续章节介绍)。

从国内外数据科学家岗位的招聘要求及著名数据科学家的访谈结果,可归纳整理出数据科学家常用的工具,见表 1-1。

表 1-1 数据科学家常用的工具

分 类	具体软件/工具
数据科学语言工具	R、Python、Clojure、Haskell、Scala
非关系数据库工具	NoSQL、MongoDB、Couchbase、Cassandra
传统数据库和数据仓库工具	SQL、RDMS、DW、OLAP
支持大数据计算的工具	Hadoop、HDFS、MapReduce、Spark、Storm
支持大数据管理、存储和查询的工具	HBase、Pig、Hive、Impala、Casecalog

续表

分 类	具体软件/工具
支持数据采集、聚合或传递的工具	Webscraper、Flume Avro、Sqoop、Hume
支持数据挖掘的工具	Weka、Klime、RapidMiner、SciPy、Pandas
支持数据可视化的工具	Ggplot2、D3、JS、Tableu、Shiny、Flare、Gephi
数据统计分析工具	SAS、SPSS、Matlab
⋮	⋮

1.3 数据结构

在计算机科学中,数据结构是一种数据组织、管理和存储的格式,它们相互之间存在一种或多种特定关系的数据元素的集合。通常情况下,精心选择的数据结构可以给计算机带来更高的运行或存储效率。数据结构往往与高效的检索算法和索引技术相关。数据结构研究的是数据的逻辑结构和数据的物理结构,以及它们之间的相互关系,主要包含三个方面的内容,即数据的逻辑结构、数据的存储结构和数据的操作,只有这三个方面的内容完全相同,才能成为完全相同的数据结构。依据数据结构特点,常见的有结构化数据、半结构化数据和非结构化数据,大数据以非结构化数据为主。

1.3.1 结构化数据

结构化数据通常存储在数据库中,是可以用二维表结构来表现的数据,是指任何一列数据不可以再细分,并且任何一列数据都具有相同的数据类型。所有关系数据库(如 Access、SQL Server、MySQL、Oracle、DB2 等)中的数据/信息全部为结构化数据。目前对结构化数据进行分析已是一种成熟的技术。

1.3.2 半结构化数据

半结构化数据类型的格式一般较为规范,都是纯文本数据,可以通过某种特定的方式解析得到每项数据。最常见的半结构化数据是日志数据、采用 XML 与 JSON 等格式的数据,每条记录可能都会有预先定义的规范,但是每条记录包含的信息可能不尽相同;也可能会有不同的字段数,包含不同的字段名、字段类型或者嵌套格式等。这类数据一般都是以纯文本的格式输出,管理维护相对而言较为方便。但是,在需要使用这些数据(如进行采集、查询、分析数据)时,应对它们的格式进行相应的转换或解码。

1.3.3 非结构化数据

非结构化数据是指不能用二维表结构来表现的数据,这类数据没有固定的标准格式,无法直接解析出其相应的值,如各种格式的办公文档、图片、图像、音频和视频等,也就是说非结构化数据不容易收集和管理,甚至无法直接查询和分析,所以对这类数据需要使用一些不同的处理方式。非结构化数据分析是一项还在不断发展的新技术。

结构化数据和非结构化数据的区别如表 1-2 所示。

表 1-2 结构化数据和非结构化数据的区别

比 较 内 容	结构化数据	非结构化数据
是否在关系数据库中显示	可以显示在行、列以及关系数据库中	不可以显示在行、列以及关系数据库中
数据格式	数字、日期和字符串	图片、音频、视频、文字处理文档、邮件、数据表
数据量	约占 20%的企业数据	约占 80%的企业数据
存储空间	需要更少存储空间	需要更多存储空间
数据管理	使用遗留解决方案保护和管理数据更方便	难以用遗留解决方案保护和管理数据
分析工具	针对结构化数据的分析工具已处于发展成熟阶段	用于挖掘非结构化数据的分析工具正处于萌芽和发展阶段

1.4 大数据特征及对科学研究的影响

自 21 世纪初至今,已有众多学者对大数据的概念和特点做过总结。认为大数据特征从 3V[体量(volume)、速度(velocity)、多样性(variety)],到 4V[大量(great volume)、高速(rapid generation velocity)、多变(various modalities)、等价值低密度(huge value but very low density)],再到 5V[4V+准确性(veracity)],大数据给人们带来了发现新价值的新机会,尤其为人们提供了巨大的潜在价值,对科学研究产生了较大影响。

1.4.1 大数据的数据特征

对大数据的理解应包含三个方面,即数据特征、技术特征与应用特征。虽然大数据的定义不统一,但大数据特征可以概括为多样化、高速性、体量大、价值密度低等。

(1) **数据类型复杂,也叫多样化。**大数据的数据种类多,复杂度高,包括网络日志、音频、视频、图片、地理位置信息等。大数据有不同的格式,有结构化的关系数据,有半结构化的网页数据,还有非结构化的视频、音频等数据。这些非结构化数据广泛存在于社交网络、物联网、电子商务中,其增长速度比结构化数据快得多。多类型的异构数据对数据的处理能力提出了更高要求。

(2) **数据产生速度快,也叫高速性。**大数据时代获得数据的速度迅速提高,需要频繁采集、处理并输出数据。数据存在时效性,需要快速处理,实时分析并得到结果。

(3) **数据体量巨大。**大数据时代需要采集、处理、传输的数据量大,数据的大小决定所考虑的数据价值和潜在的信息价值。大数据时代处理 PB 级的数据是比较常见的情况。

(4) **数据价值密度低。**从海量价值密度低的数据中挖掘出具有高价值的数据。大数据由于体量不断增大,单位数据的价值密度在不断降低,然而数据的整体价值在提高。根据互联网数据中心(IDC)调研报告预测,大数据技术与服务市场需求年增长率达 40%以上,并且是整个 IT 与通信产业增长率的数倍。通过对大数据进行分析,找出其中潜在的商业价值,探讨其发展趋势和主要驱动因素将会产生巨大的商业利润。

1.4.2 大数据对科学研究的影响

1. 促进了科学研究第四范式的产生

大数据的产生和信息技术领域提出的面向数据的概念同时改变了科学研究的模式。2007年,图灵奖得主吉姆·格雷提出了数据密集型科学“第四范式”。他将大数据科研从第三范式中分离出来单独作为一种科研范式。他认为利用海量的数据可以为科学研究和知识发现提供除实验、理论研究、计算之外的第四种重要方法,即大数据的研究从计算机模拟中分离出来,独立作为一种科研范式。单独分离出来的原因是大数据的研究方式不同于基于数字模型的传统研究方式。科学研究的四种范式的发展历程(图 1-3)同样也反映了从面向计算走向面向数据,即“数据思维”。

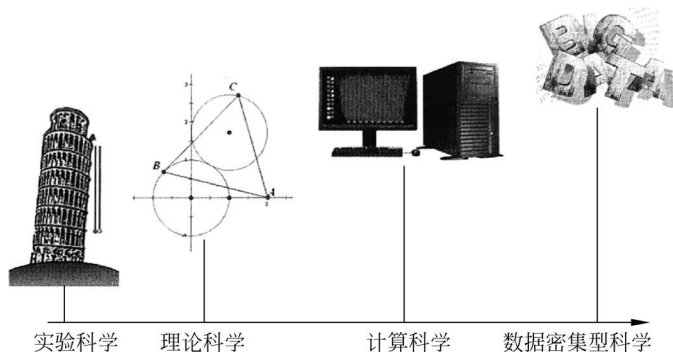


图 1-3 科学研究的四种范式

科学研究的四种范式如图 1-3 所示。第一范式是实验科学,人类早期知识的发现主要依赖于经验、观察和实验,需要的计算和产生的数据很少。人类在这一时期对宇宙的认识都是这样形成的,例如,伽利略通过在比萨斜塔扔下两个大小不一的铁球来证明自由落体定律,人类在那个时代获取知识的方式是原始而朴素的。

当人类知识积累到一定程度后,知识逐渐形成了理论体系,这时人类的科学研究进入第二范式——理论科学,通过理论研究发现知识。例如,牛顿的力学体系、麦克斯韦的电磁场理论等,人类可以利用这些理论体系去预测自然并获取新的知识。这时对计算和数据的需求已经萌生,人类已可以依赖这些理论发现新的行星,如海王星的发现不是通过观测,而是通过计算实现的,即所谓笔尖上发现的行星。

计算机的出现为人类发现新的知识提供了重要的工具,这时人类的科学研究进入第三范式——计算科学,通过计算发现知识。这个时代正好对应于面向计算的时代,可以在某些具有完善理论体系的领域利用计算机仿真计算来进行研究。这时计算机的作用主要是计算,如人类利用仿真计算可以实现模拟核爆炸这样的复杂研究。

随着庞大数据的出现,人类逐步进入面向数据的时代,这时人类的科学研究进入第四范式——数据密集型科学,通过数据研究发现知识。利用海量数据加上高速计算发现新的知识,是数据密集型的科学发现。PB 级的数据使人们在没有模型和假设的前提下也能分析数据,只有将有相互关系的数据丢进巨大计算机集群中,统计分析算法才可能发现过去的科学方法发现不了的新模式、新知识,甚至新规律。此时,计算和数据的关系在面向数据

时代变得十分紧密,也使计算和数据的协作问题成为巨大的技术挑战。例如,谷歌(Google)的广告优化配置及 2016 年李世石在围棋挑战中输给机器人 AlphaGo,都是依据第四范式实现的。人类从依靠自身判断作决定到依靠数据作决定的转变,也是大数据作出的巨大贡献之一。

2. 促进了交叉学科发展

目前,社会科学、自然科学和人文科学只是学术建制意义上的区分,它们之间已经有了密切的联系,不再像以前那样孤立。20 世纪 50 年代之前,社会科学与自然科学相对独立,基本没有跨学科交叉研究。70 年代,随着计算科学的发展,人们开始注意到经济与社会系统中的复杂现象。以圣菲研究所为代表的研究机构,开创了复杂性科学这一全新的领域。为了研究复杂性现象,他们提出了“复杂自适应系统”理论,该理论将计算机作为从事复杂性研究的最基本工具,用计算机模拟相互关联的繁杂网络,观察复杂性事物自适应系统的涌现行为。进入 21 世纪之后,涌现出用户参与的万维网(Web)应用程序,包括微博、论坛、社交网络等,它们被统称为社会媒体,导致大数据的产生。由于大量用户提供了海量的社会行为数据,人们在数据获取方式及理念上发生了重大的变化。早期人们对事物的认知受限于获取、分析数据的能力,较多地使用采样方式,以少量的数据来近似描述事物的全貌,通过采样方式获取的部分样本,可能分析得到的数据与实际的数据存在相反的结论。因此,为了让分析的结果具有更高的准确性,必须获取大量的数据,从接近事物本身的数据开始着手,从更多的细节来解释事物本身所具有的特征。所以,使用计算机模拟测试和验证社会经济政策的效果,成为公共政策领域的迫切需求。各种迹象表明,继物理计算和生物计算之后,社会计算已成为科学计算研发的新焦点,并产生新的方向和领域。

1.4.3 大数据时代的新理念

大数据是信息技术发展的必然产物,更是信息化进程的新阶段,其发展推动了数字经济的形成与繁荣。根据 IBM 前首席执行官郭士纳的观点,IT 领域每隔 15 年就会迎来一次重大变革,从 20 世纪 80 年代以来的三次信息化浪潮中可以看到上述观点(表 1-3)。我们可以认为第三次信息化浪潮涌动,是大数据时代的全面到来。人类社会信息科技的发展为大数据时代的到来提供了技术支撑,而数据产生方式的变革是促进大数据时代到来至关重要的因素。

表 1-3 三次信息化浪潮

信息化浪潮	发 生 时 间	标 志	解决的问题	代表的企业
第一次信息化浪潮	1980 年前后	个人计算机	信息处理	英特尔、ADM、IBM、苹果、微软、联想、戴尔、惠普等
第二次信息化浪潮	1995 年前后	互联网	信息传输	雅虎、谷歌、阿里巴巴、百度、腾讯等
第三次信息化浪潮	2010 年前后	大数据、云计算和物联网	信息爆炸	亚马逊、谷歌、IBM、VMware、Palantir、Hortonworks、Cloudera、阿里云等

大数据时代的到来改变了人们的生活方式、思维模式和研究范式。从传统数据到大数据的重大变化归纳为表 1-4。

表 1-4 传统数据与大数据比较

比 较 项 目	传 统 数 据	大 数 据
研究范式	第三范式	第四范式
数据产生方式	被动采集数据	主动生成数据
数据源	数据源获取较为孤立,不同数据之间添加的数据整合难度较大	利用大数据技术,通过分布式技术、分布式文件系统、分布式数据库等对多个数据源获取的数据进行整合处理
数据重要性	数据资源	数据资产
方法论	基于知识	基于数据
数据分析	统计学	数据科学
智能计算	简单算法	复杂算法
管理目标	业务数据化	数据业务化
决策方式	目标驱动	数据驱动
产业竞合关系	以战略为中心	以数据为中心
数据复杂性	不接受复杂性	接受复杂性
数据处理模式	小众参与	大众参与
思维方式	抽样、精确、因果	全样而非抽样,效率而非精确,相关而非因果

在大数据时代,各行各业的数据都在迅速增长,“大数据”正改变着人们所处的时代,对社会发展产生了方方面面的影响。例如,在大数据时代,用户会越来越多地依赖网络和各种“云端”工具提供的信息做出行为选择。这有利于提升人们的生活质量,降低个人在群体中所面临的风险。再如,一些网络公司可以通过对大量游客飞行数据的搜索和运算,预测各大航空公司每张机票的平均价格走势,从而为游客出行提供选择。

此外,大数据激发内需的剧增,引发产业的巨变。生产者具有自身的价值,而消费者则是价值的意义所在。有意义的东西才会有价值,消费者如果不认同,商品就卖不出去,价值就实现不了;消费者如果认同,商品能够卖出去,价值就得以实现。大数据可以帮助生产者从消费者角度分析意义所在,从而帮助生产者实现更多的价值。例如,大数据帮助航空公司节省运营成本;帮助电信企业实现售后服务质量提升,帮助保险企业识别欺诈骗保行为,帮助快递公司监测分析运输车辆的故障险情以提前预警维修,帮助电力公司有效识别预警,掌握即将发生故障的设备等。

1.5 大数据产生方式及来源

1.5.1 大数据的发展历程

大数据的发展历程总体上可以划分为三个重要阶段,即萌芽期、成熟期和大规模应用期。每个阶段的大致时间及内容归纳见表 1-5。

表 1-5 大数据发展的三个重要阶段

阶 段	时 期	内 容
第一阶段：萌芽期	20 世纪 90 年代— 21 世纪初	随着数据挖掘理论和数据库技术的逐步成熟，一批商业智能工具和知识管理技术开始被应用，如数据仓库、专家系统、知识管理系统等
第二阶段：成熟期	21 世纪最初 10 年	Web 2.0 应用迅猛发展，非结构化数据大量产生，传统处理方法难以应付，带动了大数据技术的快速突破，大数据解决方案逐渐走向成熟，形成了并行计算与分布式系统两大核心计算，谷歌的 CFS (colossus file system) 和分布式计算框架 (MapReduce) 等大数据技术受到重视，Hadoop 平台开始盛行
第三阶段：大规模应用期	2010 年之后	大数据应用渗透各行各业，数据驱动决策，信息社会智能化程度大幅度提高

1.5.2 大数据产生方式

第 1.1 节介绍过，数据与信息既有区别又有联系。随着人类社会信息化进程的加快，在日常生产和生活中每天都会产生大量的数据，如商业网站、政务系统、零售系统、办公系统、自动化生产系统等，每时每刻都在产生数据。数据已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素。从创新到所有决策，数据推动着企业的发展，并使得各级组织的运营更为高效。数据产生方式的变革是促进大数据时代来临的重要因素。从大数据的发展历程来看，大数据产生的方式大致经历了三个阶段，即运营式系统阶段、用户原创内容阶段、感知式系统阶段，或称为被动式数据生成、主动式数据生成、感知式数据生成。

(1) 被动式数据生成是由于数据库技术的产生。数据库技术使得数据的保存和管理变得简单，业务系统在运行时产生的数据可以直接保存到数据库中，由于数据是随业务系统运行而产生的，该阶段所产生的数据是被动的。

(2) 主动式数据生成是由于万维网的发明与发展而兴起的。20 世纪 80 年代末出现万维网，万维网给全球信息的交流和传播带来了革命性的变化，实现人们获取信息的不同方式。万维网不同于互联网，它只是互联网所能提供的服务之一，是靠互联网运行的一项服务。万维网一般使用超文本传输协议。物联网的诞生，使得移动互联网的发展大大加速了数据的产生。例如，人们可以通过手机等移动终端，随时随地产生数据。用户数据不但大量增加，而且用户还主动提交了自己的行为，使之进入了社交、移动时代。大量移动终端设备的出现，使用户不仅主动提交自己的行为，还和自己的社交圈进行了实时互动，因此数据大量地产生出来，且具有极其强烈的传播性。这样生产的数据是主动式的。

(3) 感知式数据生成是物联网的飞速发展所致的。随着物联网的发展，数据生成方式得以彻底改变。例如，遍布在城市各个角落的摄像头等数据采集设备源源不断地自动采集并生成数据。

1.5.3 大数据的来源

大数据的产生首先源于互联网企业对日益增长的网络数据分析的需求，例如：20 世纪 80 年代的典型代表是雅虎 (Yahoo)，它最先使用“分类目录”数据库；90 年代的典型代表是

谷歌,它开始运用算法分析用户搜索信息,以满足用户的实际需求;21 世纪初的典型代表是脸书(Facebook),它不仅满足用户的实际需求,而且在用户产生内容的同时,还创造新的需求。2010 年之后,互联网社会化拉开序幕,社交网站出现,海量的视频、图片、文本、短消息及社会间关系信息数据的分析需求出现。大数据快速增长的原因之一是智能设备的普及,使大数据的来源非常多。

(1) 从数据类型来看,大数据有结构化数据、半结构化数据和非结构化数据。

(2) 根据系统来分,大数据来自信息管理系统、网络信息系统、物联网系统、科学实验系统等。

① **信息管理系统**:企业内部使用的信息系统包括办公自动化系统、业务管理系统等。信息管理系统主要通过用户输入和系统二次加工的方式产生数据,其产生的大数据大多数为结构化数据,通常存储在数据库中,一般为关系型数据库。

② **网络信息系统**:基于网络运行的信息系统即网络信息系统是大数据产生的重要方式,如电子商务系统,社交媒体、搜索引擎等都是常见的网络信息系统。网络信息系统产生的大数据多为半结构化或非结构化的数据,在本质上,网络信息系统是信息管理系统的延伸,其专属于某个领域的应用,具备某个特定的目的。因此,网络信息系统有着更独特的应用。

③ **物联网系统**:物联网是新一代信息技术,其核心和基础仍是互联网,是在互联网基础上延伸和扩展的网络,其用户端延伸和扩展到了任何物品与物品之间进行信息交换和通信,而其具体实现是通过传感技术获取外界的物理、化学、生物等数据信息。

④ **科学实验系统**:科学实验系统主要用于科学技术研究,可以由真实的实验产生数据,也可以通过模拟方式获取仿真数据。

(3) 从数据源及获取设备来看,大数据分为以下四类。

① **由科学装置获取的实验探测大数据**。一个著名的例子是位于贵州省平塘县的天眼 FAST 传感器产生的数据,这类数据源的特点是由特定领域的探测仪器产生,具有很强的专业性。

② **由传感网络获取的大数据**。在城市空间中随处可见的摄像头将道路的环境信息以图像或者视频的方式记录下来,形成海量的街景数据。此外,物联网技术迅速发展,使得每一种带有电子标签的物体均可以定位于物联网,这些物体产生的电子记录同样构成了大数据集。

③ **由城市移动设备产生的数据源**。在城市,如车载全球定位系统(GPS)导航数据、手机基站位置数据、用户携带手机或驾驶车辆移动产生的带有时空标记记录的数据。

④ **由社交媒体和社交网络上传的大数据**。随着移动互联网的发展,人们通过社交媒体如微博、“X”平台(即 Twitter)等分享自己的生活,表达自己的观点。在这个过程中,人和地的属性信息通过地理标签与地理位置关联起来,形成同时兼备丰富语义信息和位置信息的数据集。

(4) 若从动态情况来考虑,数据又可分为静态数据和动态数据(又称流数据)。

① **静态数据**:不随时间发生变化的数据。例如,许多企业为了支持决策分析而构建数据仓库系统,其中存放的大量历史数据就是静态数据。这些数据来自不同的数据源,利用工具加载到数据仓库中,并且不会发生更新,技术人员可以利用数据挖掘等工具从这些静