

第3章 数 据

【案例引入】 人们生活在一个被数据包围的世界。从智能手环记录的一串步数,到流媒体播放的电影,万物皆可化为数据。然而,数据在计算机中究竟是什么样呢?

当写下人类习以为常的十进制 42 或 3.14 时,机器却只认 0 与 1。数与数制的转换,以及数值型数据在底层如何被精妙表示与运算处理便是这座数字大厦的基石。世界并非只有冷冰冰的数字,一段动人的文字、一幅绚丽的画卷、一首悠扬的乐曲……这些非数值信息又是如何通过编码规则,被翻译成 0 和 1 的比特流在芯片中流淌的?

随着信息爆炸,数据如潮水般涌来,如何告别杂乱无章,让数据得以高效存储、查询与更新?这便是数据管理的使命。如今,人类已迈入以拍字节(PB)、艾字节(EB)衡量的大数据时代,海量、多源的数据洪流中蕴藏着洞见未来的无限密码。

从最底层的 0 与 1,到非数值的编码转换,再到井然有序的数据管理与波澜壮阔的大数据海洋,数据究竟如何被刻画与驾驭?下面,继续探寻数据世界的底层逻辑。



思维导图

3.1 数与记数制

数据是事实或观察的结果,是对客观事物的逻辑归纳,是表示客观事物的未经加工的原始素材,是对事实、概念或指令的一种特殊表达形式,可以用人工方式或自动化装置进行通信、翻译转换或加工处理。

在计算机科学中,数据是指所有能输入计算机并被计算机程序处理的符号、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、模拟量等的统称。

计算机中的数据一般包含以下两类。

(1) 数值型数据。数值型数据是具有特定值的一类数据,可用于表示数量的多少,可比较大小。

(2) 非数值型数据。非数值型数据是不具有特定值的数据,不用于表示数量的多少,不能直接比较大小,但可以进行排序等操作。

信息是指能够提供知识或描述事实、过程、客观事物的内容。

数据是对客观事物的记录、描述或观察结果,是事实的体现。数字是数据的一种形式,数据是信息的一种载体。数据通过对客观事物的逻辑归纳和分析而产生,并为信息的传递和理解提供基础。

数据与信息的关系如下。

(1) 数据是反映事物属性的记录,是信息的具体表现形式。

(2) 数据是信息的载体,信息是有意义的数。

(3) 数据经过加工处理后称为信息,而信息需要经过数字化变成数据才能存储和处理。

例如:

星球出现的位置和时间等是数据。

星球运动的规律是信息。

当与人告别时,可以通过文本写“再见”,通过语音说“Bye bye”,通过图像表情包“挥手”等不同的数据形式表达相同的信息内容。

3.1.1 常用的记数制

现代计算机内部采用二进制符号进行信息编码。

在现代计算机系统中,信息的表示和处理是通过数字化编码实现的。数字化编码是一种将各种类型的信息转换成计算机能够理解和处理的格式的过程。这一过程涉及将信息转换为一系列由基本的符号集合按照特定的规则组合而成的数字代码。

数字化编码的核心在于简洁性和高效性。它使用少量的基本符号(例如二进制系统中的“0”和“1”)的有序组合表示复杂的信息。这种编码方式不仅简化了信息的存储和处理,还使计算机能够以统一的方式处理文字、图像、音频和视频等多种类型的数据。

编码过程的两大基本要素如下。

(1) 基本的符号集合。这是构成编码的最小单位,例如在二进制系统中,基本符号集合仅包含“0”和“1”。这些符号是构成更复杂信息的基础。

(2) 符号的组合规则。这些规则定义了如何将基本符号组合起来以表示不同的信息。例如,在二进制系统中,每个“0”或“1”称为 1 比特(bit,b),8 比特组成 1 字节(byte,B),而字节可以组合成更长的序列表示特定的信息。

现代计算机采用二进制符号进行信息编码的原因是二进制简单可行、容易实现,运算规则简单,适合逻辑运算。二进制系统具有易于实现电子化处理的优点,二进制的每个符号(“0”或“1”)可以通过电子设备中的两种不同状态(如电压的高或低)表示,这使得计算机能够通过简单的电路处理复杂的计算和逻辑操作。

为了适应不同种类的信息,计算机系统使用了多种编码标准和协议。例如,文本信息通常使用 ASCII(美国信息交换标准代码)或 Unicode 编码,而图像和音频数据则可能使用 JPEG、PNG、MP3 等专门的格式进行编码。这些标准和协议确保了信息在不同的计算机系统和应用之间能够准确无误地传输和处理。

记数制是一种表示及计算数的方法。例如 R 进制有以下要素。

(1) 组成部分。 R 进制数由 R 个数码组成。

(2) 运算规则。运算规则是指记数制的进位和借位规则。例如加法的逢 R 进 1,减法的借 1 当 R 。

(3) 基数。基数是指一个记数制所包含的数字符号的个数(用 R 表示)。

(4) 位值(位权)。由位置决定的值称为位值或位权,用基数 R 的 i 次幂(R^i)表示。

1. 非进位记数制

非进位记数制表示数值大小的数码和它在数值中所处的位置无关。

结绳记事是一种古老的信息记录方式,通过在绳子上打结来记录事件或传递信息。这种记事方式在古代社会中被广泛使用,尤其在没有书写系统或书写材料稀缺的地区。图 3.1 所示的结绳记事方式有助于理解这种原始的记录方法。

罗马记数法是一种历史悠久的数字表示系统,它使用特定的符号表示数值,如图 3.2 所示。罗马数字包括 I(1)、V(5)、X(10)、L(50)、C(100)、D(500)和 M(1000)。罗马记数法

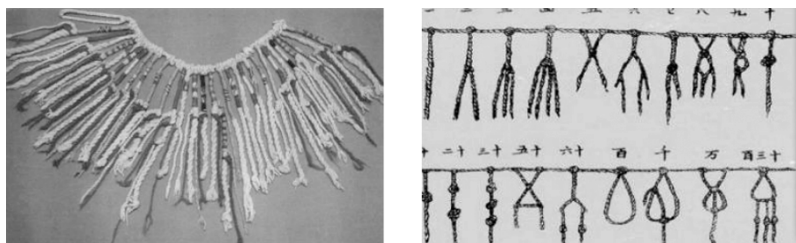


图 3.1 结绳记事

的规则如下：当相同的数字并列时，表示这些数字的数值相加；当不同的数字并列时，若小的数字位于大的数字的右边，则表示相加；若小的数字位于大的数字的左边，则表示从大的数字中减去小的数字的值。例如，罗马数字Ⅳ表示 4(即 5－1)，而Ⅵ表示 6(即 5＋1)。

1 I	11 XI	50 L
2 II	12 XII	100 C
3 III	13 XIII	500 D
4 IV	14 XIV	1000 M
5 V	15 XV	
6 VI	16 XVI	
7 VII	17 XVII	
8 VIII	18 XVIII	
9 IX	19 XIX	
10 X	20 XX	

图 3.2 罗马记数法

2. 进位记数制

在一个数中，同一个数码处在不同的位置，所表示的数值大小是不相同的。

例如，十进制数 8852.67 的位权展开式如图 3.3 所示。

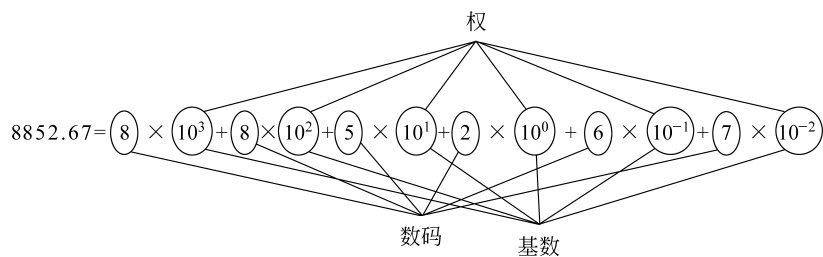


图 3.3 8852.67 的位权展开式

二进制数字系统是最基础的数字系统之一，其基数为 2，意味着在该系统中，每个位置上的数字只能取 0 或 1 这两个值。这种系统因其简洁性，在计算机科学和数字电路设计中占据着核心地位。在二进制系统中，记数规则遵循“逢二进位”的原则，即每当一个位置上的数字从 1 增加到 2 时，该位置归零，并在左边相邻的位置上增加 1。例如，二进制数从 0 增加到 1，再增加到 10，然后是 11，接着是 100，以此类推。

除二进制系统外，还有其他几种常见的进位记数制，它们在日常生活和不同领域中有着

广泛的应用。

(1) 八进制数字系统。八进制数字系统的基数为 8, 每位的数字可以取 0~7 的任意值。记数规则是“逢八进位”。八进制系统在早期计算机系统中较为常见, 因为它比二进制更紧凑, 易于人类阅读和理解。

(2) 十进制数字系统。十进制数字系统是人们日常生活中最熟悉的数字系统, 基数为 10, 每位的数字可以取 0~9 的任意值。记数规则是“逢十进位”。十进制系统基于人类有 10 个手指的事实, 在历史上被广泛采用。

(3) 十六进制数字系统。十六进制数字系统基数为 16, 每位的数字可以取 0~15 的任意值。十六进制系统在计算机科学中非常有用, 因为它可以非常方便地表示二进制数。十六进制数中的 10~15 分别用 A~F 表示。例如, 二进制数 1111 在十六进制中表示为 F。

了解这些不同的进位记数制对于学习计算机科学和相关技术领域非常重要, 这是因为它们有助于更好地理解数字信息在计算机内部的表示和处理方式。

常见的进位记数制如表 3.1 所示。

表 3.1 常见的进位记数制

记数制	十进制	二进制	八进制	十六进制
基数	10	2	8	16
数码	0, 1, 2, ..., 9	0, 1	0, 1, 2, ..., 7	0, 1, 2, ..., 9, A, B, ..., F
位权	10^i	2^i	8^i	16^i
规则	逢十进一	逢二进一	逢八进一	逢十六进一
角标	D(decimal)	B(binary)	O(octal)	H(hexadecimal)

4 种记数制的对应关系如表 3.2 所示。

表 3.2 4 种记数制

十进制	二进制	八进制	十六进制	十进制	二进制	八进制	十六进制
0	0000	0	0	8	1000	10	8
1	0001	1	1	9	1001	11	9
2	0010	2	2	10	1010	12	A
3	0011	3	3	11	1011	13	B
4	0100	4	4	12	1100	14	C
5	0101	5	5	13	1101	15	D
6	0110	6	6	14	1110	16	E
7	0111	7	7	15	1111	17	F

在计算机科学中, 不同的进位记数制有着各自的特点和应用场景。以下是几种常见的进位记数制的应用场景。

(1) 二进制数。由于二进制数只有 0 和 1 两个数码, 所以非常适合计算机硬件的实现。

这是因为只需要两种状态(开或关,高电平或低电平),从而简化了电子设备的设计。二进制数在书写和阅读时较为冗长,尤其是对于较大的数值。

(2) 十进制数。十进制数是人类日常生活中最常用的记数系统,它基于 10 个不同的数码(0~9)。尽管十进制数符合人的自然习惯,便于理解和计算,但计算机无法直接识别十进制数,需要通过转换才能处理。

(3) 十六进制数。十六进制数基于 16 个数码(0~9 和 A~F),它与二进制数有着紧密的联系,因为每 4 位二进制数可以直接转换为一个十六进制数。十六进制数的书写和阅读相对方便,且转换过程简单,因此在计算机编程和硬件设计中广泛使用。

常见的信息量单位如下。

(1) 比特(bit,b)。计算机中最小的信息量单位。信息熵公式中对数 $\log$ 若以 2 为底,则计算出来的信息熵就以比特为单位。一位二进制数码能表示的信息量称为 1 比特。它是构成信息的最小信息量单位;一个二进制位只能表示 0 或 1 两种状态。

(2) 字节(byte,B)。计算机中最基本的信息量单位,多用于计算存储容量和传输容量。8 个连续的二进制位称为 1 字节,即 $1B=8b$ 。为了方便标识和计算,更大的容量单位如千字节(KB)、兆字节(MB)、吉字节(GB)、太字节(TB)、拍字节(PB)等也被广泛使用。

(3) 字(word,W)。计算机进行数据处理时,一次存取、加工和传送的二进制位组称为“一字”。CPU 在单位时间内一次所能处理的二进制位组称为字,一字的长度称为字长。字长是计算机系统结构中的一个重要的性能指标,例如 32 位处理器、64 位处理器。

3.1.2 不同记数制之间的转换

不同记数制之间的转换规则如下。

(1) 非十进制转十进制。将非十进制数的每一位都乘以其位权(基数的幂),然后将结果相加得到十进制数。

(2) 十进制转非十进制。整数部分使用“除基数取余数”法,即不断除以基数(二进制为 2,八进制为 8,十六进制为 16),并将余数倒序排列;小数部分使用“乘基数取整数”法,即不断乘以基数,并取结果的整数部分,重复此过程直到小数部分为零或达到所需的精度。

(3) 二进制与八进制互转。将二进制数以小数点为中心向左右每 3 位分为一组,然后将每组 3 位二进制数转换为对应的八进制数。

(4) 二进制与十六进制互转。将二进制数以小数点为中心向左右每 4 位分为一组,然后将每组 4 位二进制数转换为对应的十六进制数。

(5) 八进制与十六进制互转。通常通过先将八进制或十六进制数转换为二进制数,然后转换为另一种进制实现。

3.1.3 二进制数的运算

1. 二进制数的基本运算

(1) 加法。二进制加法遵循与十进制加法类似的规则,但进位发生在 2 而不是 10。具体规则如下:

$$0 + 0 = 0$$

$$0 + 1 = 1$$

$$1 + 0 = 1$$

$$1 + 1 = 10 \text{ (计算结果中的 1 进位到下一位)}$$

(2) 减法。二进制减法也类似于十进制减法,但借位发生在 2 而不是 10。具体规则如下:

$$0 - 0 = 0$$

$$1 - 0 = 1$$

$$1 - 1 = 0$$

$$0 - 1 = 1 \text{ (这里需要从高位借 1)}$$

(3) 乘法。二进制乘法与十进制乘法类似,但更为简单,因为只涉及 0 和 1 的乘法。乘法规则如下:

$$0 \times 0 = 0$$

$$0 \times 1 = 0$$

$$1 \times 0 = 0$$

$$1 \times 1 = 1$$

(4) 除法。二进制除法与十进制除法类似,但同样更为简单。除法规则如下:除以 0 是未定义的,除以 1 得到原数,除以自身得到 1,除以大于自身的数得到 0。

2. 二进制的逻辑运算

计算机中的逻辑运算包括非(NOT)、与(AND)、或(OR)、异或(XOR)。这些运算对于处理二进制数据和执行布尔逻辑至关重要。

(1) 非(NOT)。非(NOT)运算是一种单目运算,它将一个二进制位取反,即若输入是 1,则输出是 0;若输入是 0,则输出是 1。

(2) 与(AND)。与(AND)运算是一种双目运算,它对两个二进制位进行比较,只有当两个输入都是 1 时,输出才是 1;否则,输出为 0。

(3) 或(OR)。或(OR)运算也是一种双目运算,它对两个二进制位进行比较,只要有一个输入是 1,输出就是 1;只有当两个输入都是 0 时,输出才是 0。

(4) 异或(XOR)。异或(XOR)运算同样是一种双目运算,它比较两个二进制位,若两个输入不同,则输出为 1;若两个输入相同,则输出为 0。

逻辑运算在计算机科学中有着广泛的应用,包括程序中的条件判断、数字电路设计、数据加密和解密、计算机图形学中的像素操作等。

通过掌握这些基本的二进制运算和逻辑运算,有助于更好地理解计算机如何处理信息,为进一步学习计算机科学的高级概念打下坚实的基础。

3.2 数值型数据的表示与处理

在计算机系统中,数值型数据是信息处理的基础。数值型数据可以分为两大类:定点数和浮点数。这两类数值在计算机中的表示方法和处理方式各具特色,适用于不同类型的应用场景。

3.2.1 定点数的表示

定点数是一种在计算机中表示数值的方法,小数点的位置固定不变。根据小数点的位置,定点数可以分为定点整数和定点小数两种类型。定点整数的小数点位于最低有效位之后,这意味着所有位都用于表示整数部分。而定点小数的小数点位于最高有效位之前,这意味着最高位用于表示符号,其余位则用于表示小数部分。

在定点数的表示中,通常使用二进制数表示数值。例如,一个 8 位的定点整数可以表示的范围是 $-128 \sim 127$ (使用补码表示法)。例如,数值 53 的二进制表示为 00110101,数值 -35 的二进制补码表示为 11000101。在补码表示法中,最高位(最左边的位)用作符号位,0 表示正数,1 表示负数。剩余的位则用于表示数值的绝对值。

定点数的优点是运算简单、速度快,不需要复杂的硬件支持。然而,定点数的表示范围有限,当数值超出其表示范围时,就会发生溢出。例如,一个 8 位的定点整数无法准确表示大于 127 或小于 -128 的整数。此外,定点数在处理小数时也存在精度问题,这是因为小数点后的位数是固定的,无法表示任意精度的小数。

3.2.2 浮点数的表示

浮点数在计算机中用于表示实数,能够表示非常大或非常小的数值。浮点数的表示基于科学记数法,数值被表示为一个尾数(mantissa)和一个指数(exponent)的乘积。

在 IEEE 754 标准中,一个浮点数通常由符号位、指数位和尾数位 3 部分组成。符号位表示数值的正负,指数位表示数值的范围(即科学记数法中的指数部分),尾数位表示数值的精度(即科学记数法中的尾数部分)。

例如,一个 32 位的单精度浮点数包含 1 位符号位、8 位指数位和 23 位尾数位。指数位通常使用偏移量表示法(偏移量为 127),以允许表示正指数和负指数。尾数位表示小数点后的有效数字,经过规范化处理(即尾数的最高位为 1,不显式表示)和舍入处理后得到。例如,数值 3.14159 可用 IEEE 754 单精度浮点数表示为符号位为 0(正数),指数位为 128(因为 $2^1 < 3.14159 < 2^2$,所以指数为 1,加上偏移量 127 后得到 128),尾数位表示 0.14159(经过规范化处理和舍入处理后的二进制形式)。

浮点数的优点是能够表示非常大或非常小的数值,并且具有较宽的动态范围。此外,浮点数还可以表示小数,并且具有较高的精度。浮点数的运算比定点数复杂得多,需要专门的硬件支持。同时,浮点数运算存在精度问题,特别是在进行多次运算时可能会累积误差。例如,连续进行多次浮点数加法运算可能会导致最终结果与实际值产生微小偏差。

在实际应用中,选择定点数还是浮点数取决于具体问题的需求。例如,在金融计算中,由于需要高精度的算术运算和避免溢出问题,因此定点数被广泛使用。而在科学计算和图形处理中,由于需要表示非常大或非常小的数值以及处理小数问题,所以首选浮点数。

3.3 非数值信息编码

3.3.1 文字的处理与表示

字符是非数值型数据的基础,字符与字符串是计算机中用得最多的非数值型数据。为

了使计算机能够识别和处理字符,必须对字符按一定规则进行二进制编码,使得每个字母都有唯一的编码;此外,文本中的数字和标点符号也都需要编码。

ASCII 是由美国国家标准学会(ANSI)制定的标准单字节字符编码方案。它最初是美国国家标准,用于不同的计算机在相互通信时作为共同遵守的西文字符编码标准。后来被国际标准化组织(ISO)定为国际标准,适用于所有拉丁文字字母。ASCII 码使用 7 位或 8 位二进制数表示 128 或 256 种可能的字符。这些字符包括英文字母、数字、标点符号和一些特殊符号。

汉字编码使用的统一码(Unicode)是一种以满足跨语言、跨平台进行文本转换和处理的要求为目的而设计的计算机字符编码。它为每种语言中的每个字符都设定了统一且唯一的二进制编码。统一码标准使用 16 位或 32 位二进制数表示字符(分别对应 UTF-16 和 UTF-32 编码方案),但最常用的 UTF-8 编码方案则使用 8 位序列编码字符,可以表示任意统一码字符,并且与 ASCII 码兼容。UTF-8 编码方案使用 1~4 字节表示一个字符,具体取决于字符的统一码的值。

3.3.2 图形图像表示与处理

位图(bitmap)是由像素(pixel)组成的图像。每个像素都有自己的颜色和位置信息。位图的构成主要包括以下几个要素:像素、分辨率、色彩深度和图像大小。像素是位图的基本单位,每个像素通常包含一个或多个颜色通道的信息(如红色、绿色和蓝色通道)。分辨率是指图像中每英寸(1 英寸为 25.4mm)包含的像素数(DPI),它决定了图像的清晰度或打印质量。色彩深度是指每个像素可以表示的颜色数量(即位深度),它决定了图像色彩的丰富程度。图像大小是指图像的宽度和高度(以像素为单位)。

矢量图(vector graphic)是一种使用数学公式描述图形的图像格式。与位图不同,矢量图使用线段、曲线、多边形等几何元素表示图像,而不是像素的集合。这些几何元素由数学公式定义,因此可以无限放大或缩小而不会失真。常见的矢量图格式有 WMF、AI、SWF、SVG 等。矢量图在处理图形缩放、旋转和变形等操作时具有优势,这是因为它是基于数学公式而不是像素进行计算的。

3.3.3 音频表示与处理

音频实际上是模拟量,是连续的声音信号。为了将模拟音频信号转化为数字音频信号以便计算机进行处理和存储,需要进行采样。采样是将连续的时间信号离散化的过程,它按照一定的时间间隔对模拟音频信号进行取值,并将这些取值量化为数字值。采样率是指每秒采样的次数(以赫兹为单位),可决定数字音频信号的频率范围和音质。采样位数是指每个采样点的量化精度(即位深度),它决定了数字音频信号的动态范围和信噪比。常见的数字音频格式有 WAV、MP3、AAC 等。

视频是单幅图像按时间顺序连续呈现,是典型的动态数据类型。动态视频的基础是前面讨论过的静态单幅图像(帧)。可以利用帧与帧之间的相关性压缩动态视频数据。常见的视频压缩标准有 H.264/AVC、HEVC 等。这些标准通过去除帧与帧之间的冗余信息以减少数据量,从而提高视频存储和传输的效率。

3.4 数据管理

数据管理是对不同类型的数据进行收集、整理、组织、存储、加工、传输、检索的过程。它是计算机的一个重要应用领域。

数据管理的目的有两个。第一个目的是从大量原始的数据中抽取、推导出有价值的信息,然后利用这些信息作为行动和决策的依据;第二个目的是借助计算机科学地保存和管理复杂的、大量的数据,以便人们能够方便而充分地利用这些信息资源。

3.4.1 数据的分类

按结构模式不同,可以将数据划分为结构化数据、半结构化数据和非结构化数据。

(1) 结构化数据。这类数据具有严格的结构模式,通常以二维表格形式存在,易于使用关系数据库进行存储和管理。例如,关系数据库中的数据表、Excel 电子表格、SQL 数据库中的数据等。结构化数据的特点包括明确的数据模型和结构、易于进行标准查询和分析,以及数据的高度组织和格式化。

(2) 半结构化数据。半结构化数据具有基本固定的结构模式,但不完全符合关系数据模型的严谨结构。这类数据适用于数据库集成,其结构模式通常与数据本身融合在一起,数据自身可描述结构信息。例如,HTML 文档、JSON、XML 和一些 NoSQL 数据库等。半结构化数据的特点是,数据不符合严格的数据模型但包含标签或其他标记以分隔数据元素;数据结构灵活,介于结构化数据和非结构化数据之间,可以存储复杂的数据。

(3) 非结构化数据。非结构化数据没有固定的结构模式,格式多样,包括办公文档、文本、图片、图像、音频、视频等信息。这类数据通常直接进行整体存储,且一般存储为二进制格式。非结构化数据的特点是,没有固定的格式或结构,难以用传统数据库和数据分析工具处理;通常需要高级处理方法,例如自然语言处理或图像分析。

3.4.2 数据库管理基础

数据管理的发展经历了人工管理、文件系统和数据库系统 3 个阶段。

(1) 人工管理阶段。在计算机硬件和软件尚未成熟之前,数据不保存,数据和程序不具有独立性,数据不能共享。用户将应用程序和数据一起输入内存,处理完毕后数据随程序一起释放。

(2) 文件系统阶段。随着计算机技术的发展,数据开始长期保存在磁盘上,程序与数据具有一定的独立性,但数据的独立性仍然较低。数据共享性差,存在数据冗余和不一致的问题。

(3) 数据库系统阶段。引入数据库技术后,计算机系统能够有组织地、动态地存储大量相关数据,并提供数据处理和信息资源共享的便利手段。数据库系统由硬件系统、数据库集合、数据库管理系统、相关软件支持以及相关人员组成,如图 3.4 所示。

(4) 硬件系统。硬件系统包括 CPU、内存、外存、输入设备、输出设备等,要求有足够的内存储器 and 外存储器容量,以及网络通信设备的支持(对于网络数据库系统)。

(5) 数据库集合。数据库集合是存储在计算机外存储器中的结构化相关数据的集合,不仅包含描述事物本身的数据,还包括数据之间的联系。

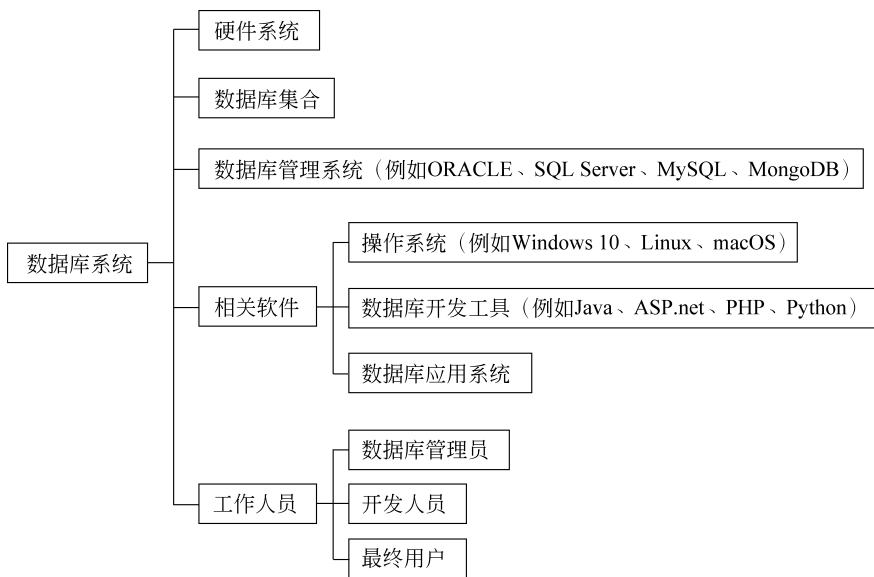


图 3.4 数据库系统

(6) 数据库管理系统。数据库管理系统运行在操作系统中,提供数据的建立、使用和维护,以及数据的安全性和完整性控制。

(7) 相关的支持软件。相关的支持软件包括操作系统、数据库开发工具、数据库应用系统等。数据库应用系统是根据具体需求开发的软件,例如图书管理系统、学籍管理系统等。数据库开发工具是开发人员编写应用系统所使用的软件平台。

(8) 相关人员。相关人员包括数据库管理员(DBA)、开发人员和最终用户。数据库管理员负责数据库的规划、设计、建立、管理和维护;开发人员负责应用系统的需求分析和功能实现;最终用户通过运行数据库应用系统处理数据。

通过数据库系统,可实现数据的高效管理,支持复杂的数据操作和信息共享,是现代数据管理的核心技术。

3.5 大 数 据

3.5.1 大数据的概念与特征

在 21 世纪的科技浪潮中,计算机技术特别是互联网和移动技术的迅猛发展,导致了数据规模的爆炸性增长,随之而来的是大数据概念的诞生。大数据又称巨量数据、海量数据或大资料,是指无法在合理时间内通过传统方法或计算机技术捕捉、管理和处理的数据集合。这些数据集合需要新的处理模式,以便提供更强的决策支持、洞察力和流程优化能力。大数据通常具有以下特征。

(1) 体量(volume)。大数据的数据存储量巨大,数据增量庞大。数据规模的庞大是大数据最显著的特点。随着云计算等技术的进步,数据量持续增长,已经从吉字节(GB)、太字节(TB)扩展到拍字节(PB)、艾字节(EB)、泽字节(ZB)为单位进行计量。例如,雅虎公司每月处理的查询超过 17.5 亿条,为此运行着 4 万多台服务器,分布在 19 个集群中,存储总量

约为 600PB。而沃尔玛每天需要处理来自几千家商店超过 40PB 的交易数据。通过这些数据的分析,企业可以制定产品需求策略,对每周 2.5 亿客户的需求进行预测。

(2) 速度(velocity)。大数据的产生和处理速度极快。数据可以通过社交媒体、定位系统等应用快速、大量地产生。同时,数据的处理速度也必须极快,只有快速、及时地处理数据,才能更有效地利用所得信息。数据处理速度可以用太字节每秒(TB/s)或拍字节每秒(PB/s)衡量,仅依靠单一网络连接,很难达到这样的速度,因此数据通常通过多个网络连接同时传输。

(3) 多样性(variety)。大数据的多样性主要体现在数据格式和来源的广泛性。大数据产生的数据种类繁多,包括结构化、半结构化和非结构化数据,甚至不完整和错误数据。数据来源多种多样,例如网页日志、电子邮件、传感器、智能手机等。数据的格式多样,包括文本、图像、视频、数字和音频等。

(4) 价值(value)。大数据的价值密度低,这意味着虽然数据量庞大,但其中具有利用价值的信息并不多。因此,需要通过特定的技术进行处理和进一步挖掘,以提取最有用的信息并加以利用。

(5) 真实性(veracity)。大数据中包含未知数量的不准确数据,这些数据会对分析和决策的准确性产生负面影响。通过应用大数据技术,在保证数据真实性的同时提高数据质量,可使数据更好地服务人们的需求。

随着大数据技术的不断进步,企业和研究机构正在开发新的工具和算法应对挑战,从而更有效地管理和分析大数据,以期在竞争激烈的市场中获得优势。

3.5.2 大数据的处理流程

大数据的处理流程通常涵盖 6 个关键步骤:数据采集、数据清洗、数据预处理、数据存储、数据分析与挖掘、数据可视化。

1. 数据采集

大数据处理的起点是数据采集。研究和分析大数据的前提条件是拥有大数据。获取大数据的方式有两种:自行采集或汇聚数据,或者使用他人已整理好的数据。大数据采集通常借助 ETL(extract-transform-load)工具,将分散的、异构数据源中的数据(例如关系数据、平面数据以及其他非结构化数据)抽取到临时文件或数据库中。当前的中大型项目多采用微服务架构进行分布式部署,因此数据采集需要在多台服务器上完成,并且不能干扰正常业务运行。为了满足这一需求,出现了多种日志收集工具,例如 Flume、Logstash、Kibana 等,它们通过简单配置即可实现复杂的数据收集和聚合。

2. 数据清洗

采集到的数据中往往包含大量重复或无用信息,这是因为数据来源多样,数据源越多,复杂性越高。数据源可能存在问题,例如数据缺失、异常值,数据集成时还会面临格式不统一、单位不一致、数据冗余等问题。此时需要对数据进行清洗和预处理,将不同来源的数据整合成适合分析算法和工具读取的统一格式,例如去重、异常值处理和数据归一化等,随后将这些数据存储到大型分布式数据库或存储集群中。

数据清洗和预处理工作一般包括数据集成、数据清洗以及数据预处理。

数据集成是将来自多个数据源的数据合并为统一的数据结构或表格的过程。若涉及结构

化和非结构化数据,则需在集成时统一转化为结构化数据,并存储于同一个数据库或表格中。

数据清洗的目标是通过清理特殊字符、转换数据格式、统一数据概念、调整数据类型、去除冗余样本等操作确保原始数据的准确性,避免后续特征工程或建模过程中出现问题。

3. 数据预处理

数据预处理是对缺失值、异常值及数据转换等问题的处理过程。大部分机器学习模型无法处理缺失值或对异常值敏感,因此预处理有助于提升模型准确性。例如,在反欺诈模型中,异常值可能是关键特征。反欺诈的核心在于发现与正常行为相悖的少数样本,这些样本在变量上表现为偏离正常范围的异常值。因此,是否需要处理异常值取决于建模目标。

4. 数据存储

大数据的数据形式通常是半结构化(例如日志数据),甚至是非结构化(例如视频、音频数据)。为了解决海量半结构化和非结构化数据的存储问题,衍生出 HDFS、KFS、GFS 等分布式文件系统。这些系统支持结构化、半结构化和非结构化数据的存储,并可通过增加设备实现横向扩展。

研究高性能、高扩展性、自适应、高可靠性、高可用性的大数据存储与管理机制是大数据技术的重要方向之一,主要研究内容包括云存储系统的持久存储机制、压缩列存储数据的查询优化、实时流数据、大规模图数据、视频数据存储机制等。

5. 数据分析与挖掘

数据分析是指运用适当的统计方法对收集的海量数据进行统计处理,提取有价值的信息,形成结论,并对资料进行深入研究、归纳和总结。在实际应用中,数据分析可辅助决策,帮助采取相应措施。数据本身并无价值,但通过分析方法,数据变得意义非凡。大数据处理的核心环节就是数据分析,通常分为批处理和流处理两种模式。

批处理和流处理均有各自的适用场景。对于时间要求不高或硬件资源有限的场景,适合批处理;而对于时间敏感、实时性要求高的场景,适合流处理。随着服务器硬件成本下降以及人们对实时性需求的提高,流处理逐渐普及,应用于股票价格预测和电商运营数据分析等领域。

6. 数据可视化

大数据分析的用户群体包括专家和普通用户,但无论是谁,对大数据分析的基本需求都离不开数据的可视化分析。数据的可视化分析能够直观地展示大数据特点,易于理解,清晰明了。

3.5.3 大数据技术的应用

1. 个性化的交通信息服务

随着交通数据环境的不断完善,基于大数据技术的交通信息服务产品不断涌现,为城市交通出行提供了多样化、个性化的信息服务。这些服务包括在线打车平台、打车软件等,能够根据用户需求和实时交通状况提供最优路线建议和出行方案。

2. 交通诱导信息服务

交通诱导信息服务分为出行前和出行中两个阶段。出行前诱导通过计算机、手机、车载导航终端等设备向用户提供出行所需信息;出行中诱导是根据实时交通状况动态调整诱导信息,为用户提供动态的路线规划。传统的诱导信息方式包括交警疏导、可变信息交通标

志、信息发布、交通广播等。随着移动通信技术的发展,用户已经可以通过移动应用获取实时诱导信息,进一步提升出行效率。

3. 现代城市物流服务

大数据技术在现代城市物流服务中发挥着重要作用。通过分析巨额交易量产生的数据,物流需求得到精确预测和高效管理,从而优化了物流配送路线和货物分发过程。

4. 公共交通出行服务

公共交通出行信息的接收方式分为定点接收和移动接收。定点接收主要通过公交电子站牌为候车乘客提供线路信息和车辆到站信息;移动接收则通过手机等智能移动终端中的公交查询应用,根据乘客的出行目的地和当前位置提供最佳乘坐方案、换乘信息及预计出行时间等。

5. 个性化营销

大数据技术能够根据用户的行为进行评估,对用户群进行细分,并为每个群体量身定制营销策略,其中不仅包括用户过去的购买行为,还包括用户在社交媒体上发布的信息,为零售商的主动营销提供了有力支持。

6. 库存管理

大数据技术在库存管理中的应用体现在其高速分析能力,能够快速为零售企业提供策略决策支持。这种分析不仅基于传统的销售数据,还包括网络数据、供应商数据和实时销售量数据,从而帮助企业在调整库存和进货策略时赢得更多时间和空间。

7. 客户意见收集

大数据技术改变了传统零售方式中商品质量评估的局限性。零售商现在可以通过政府公开信息和网络销售的顾客反馈,提取并分析商品的用户体验,从而鉴别出劣质和问题商品,不仅为零售商提供了进货参考,减少了退换货,也提高了顾客满意度。

8. 疫情预测预警

疫情预测预警是利用健康医疗大数据,通过建立统计分析和数学模型,挖掘传染病发生、发展和流行的规律,实现对疫情的预测和预警。

9. 职业病防治

国务院办公厅印发的《国家职业病防治规划(2016—2020年)》强调了信息化建设在职业病防治中的重要性。职业病防治大数据来源多样化,包括手工录入材料、电子报告数据、文档图片、微博、论坛网络信息等,具有大数据的典型特征。利用互联网、大数据、云计算等技术,可以有效进行职业病防治工作。

10. 医院管理

随着大数据时代的到来,健康医疗大数据对医院未来发展的影响日益显著,为医院管理决策提供了前所未有的潜力与空间。医疗系统能够获取更加多样和完整的数据,为医院管理者提供全局性的数据支持。

11. 运动员训练

大数据技术可以用于确定运动员的训练方向和方式,通过分析大量训练数据,帮助教练员制定更为科学和个性化的训练计划。

12. 体育竞赛中的战术安排

在垒球、篮球等体育竞赛中,大数据技术可用于战术安排。通过分析对手的比赛数据和

自身队伍的表现,教练团队可以制定出更有效的比赛策略。

13. 体育节目制作

在体育节目制作中,大数据技术可为解说员提供更加详尽、准确的比赛信息。除了比赛的比分和球员姓名外,控球率、进球率、抢断次数、射门次数等统计数据都可以通过大数据提供,极大地丰富了体育节目的内容和观赏性。

3.5.4 大数据思维

维克托·迈尔-舍恩伯格在《大数据时代》一书中,通过众多案例阐述了一个核心观点:随着大数据时代的到来,人们需要采用一种全新的思维方式挖掘大数据的潜在价值。什么是大数据思维呢?迈尔-舍恩伯格认为,需要全部数据样本,而不是抽样;关注效率而不是精确度;关注相关性而不是因果关系。

(1) 全部数据样本的原理。在大数据时代,人们不再依赖于传统的抽样调查,而是尽可能地使用全部数据样本。与抽样调查相比,这种方法提供了更高的真实性和可靠性。通过分析大量的数据,人们能够更加深入地理解现象背后的本质,揭示事物的内在规律。数据量越大,人们对事物真实性的理解就越深刻。

(2) 效率原理。大数据分析能够快速指出事物的可能性,这使企业能够基于分析结果迅速做出决策,采取行动,从而抢占市场先机,提高工作效率。在竞争激烈的商业环境中,这种快速反应能力至关重要。

(3) 相关性原理。大数据思维强调相关性而非传统的因果关系,从根本上改变了人们处理数据的方式。在大数据的框架下,人们不再追求找到复杂的因果关系,而是关注识别和利用数据之间的相关性。这种思维模式使人们能够直接了解“是什么”,而无须深入探究“为什么”。

通过这些原则,大数据思维为人们提供了一种全新的视角,帮助人们更好地理解 and 利用大数据,从而在商业决策、科学研究、社会治理等多个领域实现突破和创新。

习 题 3

1. 在向他人解释“什么是数据”时用什么进行比喻比较合适?试结合“数据是信息的载体”这句话,设计一个生动的比喻。

2. 计算机采用二进制,而人类习惯十进制。假设一下,若人类有 12 根手指,则今天的记数系统可能会变成什么样?会对计算机的设计产生什么影响?

3. 数据分为结构化、半结构化和非结构化。分别举出一个日常生活中的例子,并说明“管理”这类数据时会面临什么挑战。

4. 大数据思维强调“关注相关性而不是因果关系”。试举一个生活中的例子说明“知道相关性已经很有用,不一定需要知道原因”。

5. 图片分为位图(由像素组成)和矢量图(由数学公式描述)两种。若要设计一个无论怎么放大都不会模糊的标识,应该用哪种格式?如果要存储一张照片,又该用哪种格式?试说明理由。

6. 是否同意“数据量越大,我们对事物真实性的理解就越深刻”这个观点?试举例说明可能存在的问题或反例。