

数据建模概念

首先要知道数据库技术所指的“三个世界”是什么——现实世界、信息世界和数据世界。针对这三个世界，数据库又有与之对应的概念，分别是“实体模型”、“概念模型”和“数据模型”。

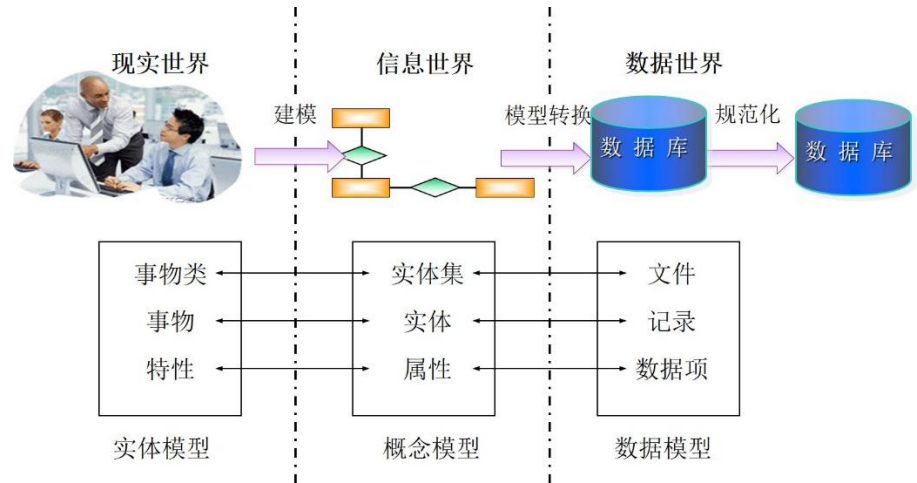


图 3-1 数据库技术领域的“三个世界”及与之对应的“三个模型”

实体模型（Entity Model）是现实世界的事务类的集合，有事务类以及所包括事物和性质；概念模型（Conceptual Model）把现实世界转换为信息世界的模型，有实体集以及所包括的实体和属性，可以使用例如实体联系（Entity Relationship, E-R）模型进行描述；数据模型（Data Model）是把信息世界转化为数据世界使用的模型，也就是数据在计算机系统里的存储形式，有文件以及所包括的记录和数据项。本教材采用的是关系模型。

数据模型除了关系模型外，常见的还有层次和网状模型。

数据建模是指根据用户的实体模型需求，设计出概念模型，此阶段为逻辑结构设计阶段。E-R 模型表示的概念结构模型独立于任何一种数据模型，并独立于任何一个具体的数据库管理系统。

概念结构设计阶段设计的 E-R 模型将转换为关系数据模型二维表结构。数据建模后，就可以在某一具体的数据库管理系统上，设计出数据库以及表的结构和建立真实可用的数据库和表。

3.1.1 现实世界数据化过程

将显示世界存在的客观事物进行数据化，要经历从“现实世界”到“信息世界”、再到“数据世界”的三个阶段。

首先将现实世界中客观存在的事物及它们所具有的特性抽象为信息世界的实体和属性；然后使用实体联系（E-R）图表示实体、属性和实体间的联系；最后再将 E-R 图转换为数据世界中的关系。

E-R 图转换为关系模型后，为了保证数据完整性，还需对关系规范化，得到最大数据共享和最小数据冗余的关系模型。

3.1.2 实体模型

实体是现实世界中存在的并可相互区别的事物或概念。实体模型对应的是“现实世界”。

实体可以是具体的人、事和物，也可以是抽象的概念或联系，其中的一个具体事物称为实例。这些事物的集合称为事物类，是由具有相似类型的事物，具有共同的性质。如图书馆管理系统中的“书籍”是一个事务类；“读者”也是一个事务类。除此之外，还有业务活动，例如书籍和读者之间发生“借阅”现象。

一个事物可以拥有多种特性来形容自己。如“书籍”，具有书名、作者、出版社等共同的性质；“读者”，具有姓名、性别、电话等共同的性质

此阶段是分析项目涉及的业务活动和数据的使用情况，弄清所用数据的种类、范围、数量以及在业务活动中的存储情况，确定用户对数据库系统的使用要求和各种约束条件等，

3.1.3 概念模型

概念模型是现实世界在人们头脑中的反映，是对客观事物及其联系的一种抽象描述，从而产生概念模型。概念模型对应的是“信息世界”，是按用户的观点对数据库建模。

概念模型是现实世界到数据世界必然经过的中间层次，使用相关信息化概念和技术对现实世界进行抽象；抽象的结果可立即转化为相应的数据模型，在具体的计算机系统上进行实施部署。

在概念模型下，有如下术语：

实体 (Entity)：是现实世界中的一个对象或概念，如图书馆管理系统中的书籍和读者等。

属性 (Attribute)：是进一步描述实体的某个具体性质，例如书籍的书名、作者等。如果某个属性或者属性组合的值能唯一地标识出实体集中一行，则可以选择该属性或者属性组合作为实体标识符。

联系 (Relationship)：表示实体之间存在的相互作用，例如表示书籍实体与读者实体之间读者借阅的联系。

具有相同属性的实体的集合称为实体集。在同一实体集中，每个实体的属性及其值域是相同的，但可能取不同的值。

3.1.4 联系类型

实体不是孤立的，实体与实体之间有着相互的联系。实体间的联系分为 3 种联系类型，1 对 1 (1:1)、1 对多 (1:n) 和多对多 (m:n)。可以使用实体联系 (E-R) 图表示实体与实体之间的联系。

一对一联系 (1:1)：设 A、B 为两个实体集。若 A 中的每个实体至多和 B 中的一个实体有联系，反过来，B 中的每个实体至多和 A 中的一个实体有联系，称 A 对 B 或 B 对 A 是 1:1 联系。注意，1:1 联系不一定都是一一对应的关系，可能存在着无对应。

一对多联系 (1:n)：如果 A 实体集中的每个实体可以和 B 中的几个实体有联系，而 B 中的每个实体至多和 A 中的一个实体有联系，那么 A 对 B 属于 1:n 联系。

多对多联系 (m:n)：若实体集 A 中的每个实体可与 B 中的多个实体有联系，反过来，B 中的每个实体也可以与 A 中的多个实体有联系，称 A 对 B 或 B 对 A 是 m:n 联系。

3.1.5 实体联系（E-R）图

通常使用实体联系（E-R）图描述现实世界的信息结构，并且可以将 E-R 图转换成为数据库管理系统（Database Management System, DBMS）能支持的数据模型。

E-R 图有 4 个要素

- （1）矩形：表示实体，矩形内要标注实体的名称；
- （2）菱形：表示实体之间的联系，菱形内标注联系的名称；
- （3）椭圆：表示实体和联系所具有的属性，椭圆内标注属性的名称。如果属性较多，也可以将实体与其相应的属性另外单独用列表表示。
- （4）连线：用来连接实体与实体所具有的属性、联系与联系所具有的属性，以及实体与联系。连线上要标出联系类型。

对于复杂的系统，E-R 图设计通常都应经过以下两个阶段：

- （1）针对每一用户画出该用户信息的局部 E-R 图，确定该用户视图的实体、属性和联系。需注意的是：能作为属性的就不要作为实体，这有利于 E-R 图的简化。
- （2）综合局部 E-R 图，生成总体 E-R 图。在综合过程中，同名实体只能出现一次，还要去掉不必要的联系，以便消除冗余。一般来说，从总体 E-R 图必须能导出原来的所有局部视图，包括实体、属性和联系。

3.1.6 数据模型

数据模型是数据特征的抽象，描述了系统的静态特征、动态行为和约束条件，为数据库系统的信息表示与操作提供了一个抽象的框架。

数据模型对应的是“数据世界”。数据模型，按计算机系统的观点对数据建模。常见的数据模型种类有网状模型、层次模型、关系模型和面向对象数据模型。

数据模型的文件、记录和数据项，是数据库系统真实保存数据的方式。分别对应着概念模型的实体集、实体和属性。

数据模型三要素：数据结构、数据操作和完整性约束。

数据结构：用于描述系统的静态特征。

数据操作：指对数据库中数据允许执行的操作的集合，用语言描述系统的动态特征。数据模型必须定义操作的确切含义、操作符号、操作规则以及实现操作的方法。

完整性约束：是完整性规则的集合，指给定的数据模型中数据及其联系所具有的制约和依存规则。

3.1.7 关系模型

关系模型是应用最广的数据模型，是用一组关系来描述数据库。通俗地说，关系就是一张表格，用“二维表”来表示数据之间的联系。

关系数据库是将数据表示为表的集合，通过建立简单表之间的联系来定义结构的一种数据库。

关系模型术语：

关系（Relation），数据库可以由多个表组成，表与表之间可以以不同的方式相互关联。

元组（Tuple），表中的行称为元组。一行是一个元组，对应存储文件中的一个记录值。

属性（Attribute），表中的列称为属性，每一列有一个属性名。

域（Domain），表中的垂直列，也称为属性。属性的取值范围，即不同元组对同一属性

的取值所限定的范围,称为域的规格。

分量: 元组的某个属性值。在一个关系数据库中, 它是一个操作原子, 即关系数据库在做任何操作的时候, 属性是“不可分的”。

码: 表中可以唯一确定一个元组的某个属性 (或者属性组), 如果这样的码有不只一个, 都叫候选码, 从候选码中挑一个出来做主属性, 它就叫主码。

全码: 如果一个码包含了所有的属性, 这个码就是全码。

主属性: 一个属性只要在任何候选码中出现过, 这个属性就是主属性。

非主属性: 与上面相反, 没有在任何候选码中出现过, 这个属性就是非主属性。

外码: 一个属性 (或属性组), 它不是码, 但是它别的表的码, 它就是外码。

关系模式 (Relation Schema): 对关系的描述, 记为 $R(A_1, A_2, \dots, A_n)$, 是由关系名和一组属性构成。

关系数据库术语:

表 (Table), 是一种按行与列排列的相关信息的逻辑组, 类似于工作单表。

字段 (Field), 数据库表中的每一列称作一个字段。表是由其包含的各种字段定义的, 每个字段描述了它所含有的数据。创建一个数据库时, 须为每个字段分配一个数据类型、最大长度和其它属性。字段可包含各种字符、数字甚至图形。

记录 (Record), 各个有关的信息存放在表的行, 被称为记录。一般来说, 数据库表创建时任意两个记录都不能相同。

数据库 (Database), 多个相关联的数据表的集合。