

第 4 章 数据库设计和规范化理论

4.1 学习目标

- 了解规范化理论、范式的概念
- 掌握判断关系模式范式等级的方法
- 了解数据库的设计方法和设计步骤,以及各个阶段的作用和任务
- 掌握概念结构设计阶段将需求分析的结果转化为 E-R 模型的方法
- 掌握逻辑结构设计阶段将 E-R 图转换成关系模式的转换内容与转换原则
- 掌握数据模型的优化方法

4.2 关系数据库规范化理论

4.2.1 基本知识

1. 问题引入

前面已经介绍了关系模型、关系数据库的基本概念。关系数据库是由实体与实体之间联系的关系集合构成的。关系数据库设计理论所要研究的就是针对某一个具体问题,如何构造一个适合于它的数据模式,即构造几个关系模式,每个关系模式应该由哪些属性组成等。

下面通过实例来说明采用不同的数据库模式将产生不同的效果。

例如,某学校要建立一个数据库以描述学生选修课程的情况。由现实世界的已知事实可以得到如下对应关系:每一名学生可以选修多门课程,每一门课程可以被多名学生所选修;每一名学生选修一门课程都会有一个成绩。

针对上述情况可能设计出以下两种关系模式。

(1)只产生一个关系模式

学生选课关系模式(学号,姓名,性别,年龄,所在系别,课程号,课程名,教师名,学分,成绩)。

(2)产生三个关系模式

学生关系模式(学号,姓名,性别,年龄,所在系别);

课程关系模式(课程号,课程名,教师名,学分);

选课关系模式(学号,课程号,成绩)。

比较分析这两种关系模式,发现第一种设计方法可能带来如下问题:

(1)数据冗余

当每一个学生选修多门课程的时候,这个学生的姓名、性别、年龄和所在系别是被重复存储的,这种重复存储是毫无意义的,浪费了大量的存储器资源,是数据冗余。

(2)修改异常

由于数据冗余,当修改某些属性(如学生的年龄)时,可能有一部分相关元组被修改,而另一部分相关元组没有被修改(同一名学生可能对应两个年龄),这就造成了数据的不一致性。

(3)插入异常

第一个关系模式中的主码是(学号,课程号)的属性组合,假如要插入刚入学的大一新生的信息,学号为20120001,姓名为周一,男,17岁,计算机系;由于新生刚入学还未选课,选修课程号为空。此时,则无法将这条信息插入学生选课关系模式中。因为在插入数据时,主码是不允许为空的,而这时主码的一部分(课程号)为空,因而导致学生信息插入不成功。

(4)删除异常

如果只有张三同学选修了 DataBase 课程,那么在张三同学毕业离校的时候,学校在删除张三同学基本信息的同时,也将 DataBase 这门课程的基本信息彻底删除了,丢失了应该保存的课程信息。

由于存在上述问题,显然第一种设计不是好的关系模式。第二种设计方法就不存在上述问题,消除了数据冗余,消除了修改、插入、删除异常。但这种方法也有自己的缺点,查询效率太低。

在关系模式的多种组合中选取一个好的关系模式的集合作为数据库模式,将会直接影响到整个数据库系统。那么,什么样的关系模式是相对较好的呢?人们通常依据规范化理论进行判断。

2. 函数依赖

数据依赖是一个关系内部属性与属性之间的一种约束关系,这种约束关系是通过属性间值的相等与否体现出来的数据间的相互关系。数据依赖有多种类型,常用的数据依赖有函数依赖和多值依赖,其中函数依赖是最重要也是最基本的一种数据依赖。

(1)函数依赖的定义

函数依赖普遍地存在于现实生活中,它反映属性或属性组合之间相互依存、相互制约的关系。

函数依赖的定义为:设 $R(U)$ 是属性集 U 上的关系模式, X 与 Y 是 U 的子集, r 是 $R(U)$ 的任意一个可能的关系(即一个二维表)。如果对于 r 中的任意两个元组(即两个记录,或两行数据) t 和 s ,由 $t[X]=s[X]$ 导致 $t[Y]=s[Y]$,则称 X 函数决定 Y ,或称 Y 函数依赖于 X ,记作 $X \rightarrow Y$ 。

函数依赖的相关术语和记号如下:

若 $X \rightarrow Y$,则称 X 为决定因素。

若 $X \rightarrow Y, Y \rightarrow X$, 则记作 $X \leftrightarrow Y$ 。

函数依赖是语义范畴的概念, 需要根据语义来确定一个函数依赖。例如, 在学生的关系模式中“学生姓名 \rightarrow 所在系别”这个函数依赖只有在学生没有重名的条件下才成立。如果允许有相同的学生姓名存在, 则所在系别就不再函数依赖于学生姓名了。

(2) 函数依赖的分类

关系数据库中函数依赖主要有以下几类:

① 平凡函数依赖和非平凡函数依赖

设 $R(U)$ 是属性集 U 上的关系模式, 若对于任何 $X, Y \in U$, 有 $X \rightarrow Y$ 且 Y 不包含于 X , 则称 $X \rightarrow Y$ 是非平凡的函数依赖。反之, 如果 Y 包含于 X , 则称 $X \rightarrow Y$ 是平凡的函数依赖。

例如, 在学生关系模式(学号, 姓名, 年龄, 性别, 所在系别)中,

学号 \rightarrow 性别, (学号, 姓名) \rightarrow 年龄, 均为非平凡函数依赖。

(学号, 姓名) \rightarrow 姓名, 为平凡函数依赖。

若不特别声明, 一般总是讨论非平凡的函数依赖。

② 完全函数依赖和部分函数依赖

设 $R(U)$ 是属性集 U 上的关系模式, 如果 $X \rightarrow Y$, 并且对于 X 的任何一个真子集 X' , 都不存在 $X' \rightarrow Y$, 则称 $X \rightarrow Y$ 是一个完全函数依赖, 即 Y 完全函数依赖于 X , 记作 $X \xrightarrow{F} Y$ 。

反之, 如果存在 $X' \rightarrow Y$ 成立, 则称 $X \rightarrow Y$ 是一个部分函数依赖, 即 Y 部分函数依赖于 X , 记作 $X \xrightarrow{P} Y$ 。

例如, 在学生关系模式(学号, 姓名, 年龄, 性别, 所在系别)中, (学号, 姓名) \rightarrow 年龄, 为部分函数依赖。因为(学号, 姓名)属性组合中存在真子集学号, 使得“学号 \rightarrow 年龄”也成立, 所以它是部分函数依赖。学号 \rightarrow 年龄, 为完全函数依赖。

在选课关系模式(学号, 课程号, 成绩)中, (学号, 课程号) \rightarrow 成绩, 为完全函数依赖。

③ 传递函数依赖

设 $R(U)$ 是属性集 U 上的关系模式, 如果 $X \rightarrow Y, Y \rightarrow Z$, 并且不存在 $Y \rightarrow X$, 则称 $X \rightarrow Z$ 是一个传递函数依赖, 即 Z 传递函数依赖于 X 。

例如, 在职工关系模式(职工编号, 姓名, 所在车间, 车间主任)中, 职工编号 \rightarrow 所在车间, 所在车间 \rightarrow 车间主任, 并且不存在所在车间 \rightarrow 职工编号, 则车间主任传递函数依赖于职工编号。

注意上述定义中的条件不存在 $Y \rightarrow X$ 。如果不加上这一限制, 当 $X \rightarrow Y$ 时允许 $Y \rightarrow X$, 则 $X \leftrightarrow Y$ 。而在 $X \leftrightarrow Y$ 的条件下, $Y \rightarrow Z$ 就等于 $X \rightarrow Z$ 。这样 X 就直接函数决定 Z , 而不是通过 Y 传递决定 Z 了, 即非传递函数依赖。

3. 范式

在关系数据库中, 关系模式设计的好坏取决于它的函数依赖是否满足特定的要求。满足特定要求的模式称为范式, 满足不同程度要求的为不同范式。

1971~1972年, E. F. Codd 首先提出了规范化理论, 系统地提出了第一范式(简称 1NF)、第二范式(简称 2NF)和第三范式(简称 3NF)的概念。1974年, E. F. Codd 和 Boyce 又共同提出了一个新的范式, 即 BCNF(Boyce-Codd Normal Form, 修正的第三范式)。1976年, Fagin 提出了第四范式, 后来又有人提出了第五范式。

一般地, 关系模式 R 为第 x 范式就可以写成 $R \in xNF$ 。各种范式之间的联系为

$5NF \subset 4NF \subset BCNF \subset 3NF \subset 2NF \subset 1NF$ 。

通过关系模式分解,可以将一个低一级范式的关系模式转换为若干个高一级范式的关系模式的集合,这种过程就叫规范化。

(1) 第一范式

如果关系模式 R 中的每一个属性都是不可分解的,则称 R 属于第一范式,记作 $R \in 1NF$ 。

例如,设关系模式 R (系别名称,高级职称人数)表示某学校系别的基本信息,假设系别信息状况如表 4.1 所示。

表 4.1 系别基本信息表

系别名称	高级职称人数	
	教授人数	副教授人数
计算机系	6	10
日语系	3	5
英语系	4	8

从表 4.1 中可以看出,“高级职称人数”属性是可以分解的,所以 R 不满足 $1NF$ 。

解决问题的办法是:将“高级职称人数”属性拆开,形成关系模式 R_1 (系别名称、教授人数、副教授人数)。形式如表 4.2 所示。显然,此时关系模式 R_1 中的每一个属性列都是不可再分的,所以 $R_1 \in 1NF$ 。

表 4.2 分解后的系别基本信息表

系别名称	教授人数	副教授人数
计算机系	6	10
日语系	3	5
英语系	4	8

第一范式是对关系模式最起码的要求。不满足第一范式的数据库模式不能称为关系数据库,但是满足第一范式的关系模式并不一定是一个好的关系模式。

(2) 第二范式

如果关系模式 $R \in 1NF$,且每一个非主属性都完全函数依赖于候选码,则称 R 属于第二范式,记作 $R \in 2NF$ 。

例如,设关系模式 R (仓库号,设备号,数量,地点)表示仓库设备的存储情况。候选码是(仓库号,设备号)属性组合,由于关系模式 R 中的每一个属性都不可再分,所以 $R \in 1NF$ 。因为非主属性“数量”完全函数依赖于候选码。非主属性“地点”部分函数依赖于候选码。即有(仓库号,设备号)→地点,仓库号→地点,所以 R 不满足 $2NF$ 。

关系模式 R 中存在异常,比如某一个仓库只有一种设备,当这种设备被移走后,在删除此设备信息的同时将这个仓库的信息也删除了。

解决问题的办法是:用投影分解把关系模式 R 分解为两个关系模式。将部分函数依赖关系的决定方属性和非主属性从关系模式中提出,单独构成一个关系模式;将余下属性加上码(仍要保留部分函数依赖的决定方属性)构成另一关系模式。

按照上述方法分解,将关系模式 R 分解为 R_1 (仓库号,设备号,数量)和 R_2 (仓库号,地点)两个关系模式。此时, R_1 和 R_2 均属于第二范式。

(3)第三范式

如果关系模式 $R \in 2NF$,且每一个非主属性都不传递函数依赖于候选码,则称 R 属于第三范式,记作 $R \in 3NF$ 。

例如,设关系模式 R (仓库号,仓库面积,所在城市,所在省)表示不同仓库在各省市分布情况。候选码是仓库号,由于关系模式 R 中的每一个属性都不可再分,所以 $R \in 1NF$ 。又因为 R 中每一个非主属性都完全函数依赖于候选码,所以 $R \in 2NF$ 。又因为函数依赖有仓库号 \rightarrow 所在城市,所在城市 \rightarrow 所在省,所以仓库号 \rightarrow 所在省, R 中存在传递函数依赖,所以 R 不满足 3NF。

关系模式 R 中存在异常,比如要在辽宁省大连市设立一个仓库,此时想先存入有关所在城市的信息,但由于没有仓库号,主码为空,则插入是失败的。

解决问题的办法是:用投影分解。把关系模式 R 分解为两个关系模式,将传递函数依赖的属性分解出来,消除传递函数依赖。

按照上述方法分解,将关系模式 R 分解为 R_1 (仓库号,仓库面积,所在城市)和 R_2 (所在城市,所在省)两个关系模式。此时, R_1 和 R_2 均属于第三范式。

(4)BCNF(修正的第三范式)

如果关系模式 $R \in 3NF$,且没有一个属性是部分函数依赖或传递函数依赖于候选码,则称 R 属于修正的第三范式,记作 $R \in BCNF$ 。

关系模式的规范化过程是通过对关系模式的分解来实现的。规范化的过程如图 4.1 所示。

规范化的基本思想是,逐步消除数据依赖中不合理的部分,使每一个关系模式更趋于完美。但并不是范式越高越好,范式越高,模式分解的越多,我们在进行数据查询的时候往往要进行许多张表的连接,系统开销较大,查询效率较低。所以,在进行关系模式规范化的过程中,关系模式一般分解到 3NF 就认为是比较好的了。

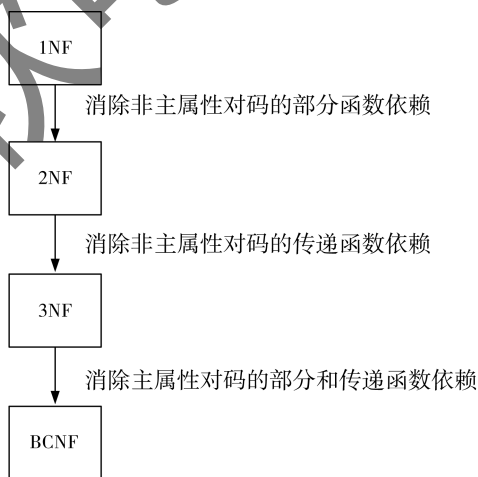


图 4.1 规范化过程

4.2.2 任务实践

任务 1:如果关系模式中存在数据依赖问题,可能导致数据库出现插入异常,这是指什么?

任务解析:

数据库逻辑结构设计不好,可能会导致数据冗余、修改异常、插入异常和删除异常。所谓插入异常是指执行插入操作时,原本应该插入的数据无法插入的情形。

任务 2:如果关系模式中存在数据依赖问题,可能导致数据库出现删除异常,这是指什么?

任务解析:

数据库逻辑结构设计不好,可能会导致数据冗余、修改异常、插入异常和删除异常。所谓删除异常是指执行删除操作时,将不应该删除的数据删除的情形。

任务 3:判断“关系模式的分解是唯一的”这句话是否正确。

任务解析:

对关系进行规范化的主要方法就是对关系模式进行分解。由于选择函数依赖的先后顺序不同或者求解的函数依赖最小集不同,都会使关系模式的分解结果不同,所以对于一个关系模式的分解可以是不唯一的。

任务 4:规范化理论对数据库设计有什么指导意义?

任务解析:

(1)在需求分析时,用数据依赖的概念来分析和表示各个数据项之间的联系。

(2)在概念结构设计阶段,以规范化理论为工具来构造实体类型和联系类型,设计 E-R 模型并对其进行优化。

(3)在逻辑结构设计阶段,使用规范化理论来设计模式和评价模式。

任务 5:设有关系 R,如表 4.3 所示。

表 4.3

关系 R

课程号	课程名称	授课教师名	职称
C1	操作系统	周一	教授
C2	软件工程	吴二	讲师
C3	高等数学	张三	副教授
C4	计算机导论	周一	教授
C5	法律基础	李四	讲师

(1)试问关系 R 为第几范式?为什么?

(2)关系 R 中是否存在删除操作异常?若存在,则说明是在什么情况下发生的。

(3)将关系 R 分解为高一范式,分解后的关系如何解决分解前可能存在的删除操作的异常问题。

任务解析:

(1)关系 R 是 2NF。

因为关系 R 的候选码是课程号,并且“课程号→授课教师名”,“授课教师名→课程号”不成立,“授课教师名→职称”,所以存在非主属性职称对候选码课程号的传递函数依赖,因此关系 R 不是 3NF。又因为不存在非主属性对候选码的部分函数依赖,所以关系 R 满足 2NF。

(2)关系 R 中存在删除操作异常。当删除某门课程信息时,会将不该删除的教师信息也一起删除了。

(3)将关系 R 分解为高一级范式如表 4.4 和表 4.5 所示。

表 4.4

关系 R1

课程号	课程名称	授课教师名
C1	操作系统	周一
C2	软件工程	吴二
C3	高等数学	张三
C4	计算机导论	周一
C5	法律基础	李四

表 4.5

关系 R2

授课教师名	职称
周一	教授
吴二	讲师
张三	副教授
李四	讲师

分解后,若删除某门课程信息时,只需对关系 R1 操作,教师信息在关系 R2 中仍然保留,不会丢失教师的个人信息。

任务 6: 在银行管理系统的数据库中,有一关系模式为 R(BNO, SSNO, BNAME, ADDRESS, CITY, SNAME, SEX, AGE, ACCOUNT),其中属性分别表示银行编号,身份证号,银行名称,银行所在地点,银行所在城市,顾客姓名,性别,年龄,账户号(假设一个人在一个银行只能开一个账户)。写出该关系模式的主码,并判断此关系模式是否满足 3NF,若不满足请对其进行规范化,以达到 3NF。

任务解析:

该关系模式 R 的主码为(BNO, SSNO)。由于关系模式 R 中的每个分量都是不可再分的数据项,所以 R 满足 1NF。关系模式 R 中存在以下函数依赖:

(BNO, SSNO) → BNAME,	BNO → BNAME,
(BNO, SSNO) → ADDRESS,	BNO → ADDRESS,
(BNO, SSNO) → CITY,	ADDRESS → CITY,
(BNO, SSNO) → ACCOUNT,	BNO → CITY,
(BNO, SSNO) → SNAME,	SSNO → SNAME,
(BNO, SSNO) → SEX,	SSNO → SEX,
(BNO, SSNO) → AGE,	SSNO → AGE,

首先,关系模式 R 满足 1NF,但存在部分函数依赖,所以,R 不满足 2NF,将其分解为:

$R_1(BNO, SSNO, ACCOUNT) \in 2NF;$

$R_2(BNO, BNAME, ADDRESS, CITY) \in 2NF;$

$R_3(SSNO, SNAME, SEX, AGE) \in 2NF;$

其次,关系模式 R1、R3 均已满足第三范式,但关系模式 R2 存在传递函数依赖,R2 不满足

第三范式,将 R2 分解为:

$R4(BNO, BNAME, ADDRESS) \in 3NF;$

$R5(ADDRESS, CITY) \in 3NF;$

最后, R1、R3、R4、R5 满足第三范式,总结为:

$R1(BNO, SSNO, ACCOUNT);$

$R3(SSNO, SNAME, SEX, AGE);$

$R4(BNO, BNAME, ADDRESS);$

$R5(ADDRESS, CITY)。$

4.2.3 思考与练习

1. 下列函数依赖中,()属于平凡函数依赖。

A. $(X, Y) \rightarrow Z$

B. $(X, Y) \rightarrow Y$

C. $X \rightarrow Z$

D. $Z \rightarrow Y$

2. 下面关于函数依赖的叙述中,不正确的是()。

A. 若 $X \rightarrow Y, X \rightarrow Z$, 则 $X \rightarrow YZ$

B. 若 $X \rightarrow Y, WY \rightarrow Z$, 则 $XW \rightarrow Z$

C. 若 $XY \rightarrow Z$, 则 $X \rightarrow Z, Y \rightarrow Z$

D. 若 $X \rightarrow Y$, 则 $XZ \rightarrow YZ$

3. X、Y 是关系模式 R 上的两个属性集,当 X、Y 之间具有 1 对多联系时,则存在的函数依赖是()。

A. $Y \rightarrow X$

B. $X \rightarrow Y$

C. $X \leftrightarrow Y$

D. $X \rightarrow Y, Y \rightarrow X$ 都有可能

4. 已知关系模式 $R(A, B, C, D, E)$, 其上的函数依赖集合 $FD = \{A \rightarrow C, E \rightarrow A, BE \rightarrow D\}$, 该关系模式的候选键是()。

A. AE

B. BE

C. CE

D. DE

5. 已知关系模式 $R(A, B, C, D)$, 其上的函数依赖集合 $FD = \{B \rightarrow A, AD \rightarrow C\}$, 该关系模式 R 最高属于()。

A. 1NF

B. 2NF

C. 3NF

D. BCNF

6. 简述 2NF 和 3NF 的关系。

7. 在某一商业集团数据库中,有一个关系模式为 R(商店编号,商品编号,库存数量,部门编号,部门经理)。这些数据有下列语义:

(1) 每个商店的每种商品只在一个部门销售;

(2) 每个商店的每个部门只有一个部门经理;

(3) 每个商店的每种商品只有一个库存数量。

请回答下列问题:

(1) 根据上述语义,写出关系模式 R 的基本函数依赖。

- (2) 写出关系模式 R 的候选码。
- (3) 试问关系模式 R 最高已经达到第几范式? 给出理由。
- (4) 如果关系模式 R 不满足 3NF, 请将 R 规范化到 3NF。

4.3 数据库设计概述

4.3.1 基本知识

什么是数据库设计呢? 具体地说, 数据库设计是要在一个给定的应用环境中, 通过合理的逻辑设计和有效的物理设计, 构造较优的数据库模式, 建立数据库及其应用系统, 能够有效地存储和管理数据, 满足用户的各种信息需求。因此, 数据库设计是数据库在应用领域的主要研究课题。

1. 数据库设计方法

采用合理的数据库设计方法, 可以确保数据库系统的设计质量, 降低系统运行后的维护代价。数据库设计是涉及多学科的综合性的技术, 也是一项庞大的软件开发工程。因此, 一个从事数据库设计的专业人员应该具备多方面的专业技术和知识。除了具备计算机科学的基础知识之外, 还必须了解软件工程的原理, 掌握程序设计的技巧; 具备数据库的基本知识和数据库设计技术, 同时还必须具备应用领域的专业知识, 才能设计出符合具体应用领域要求的数据库应用系统。

早期数据库设计主要采用手工与经验相结合的方法。设计的质量往往与设计人员的经验与水平有直接的关系, 设计质量难以保证。经常是数据库运行一段时间后又出现各种各样的问题, 需要进行修改或重新设计, 大大增加了后期维护的负担。所以人们努力探索, 通过运用软件工程的思想和方法, 提出了各种数据库设计方法, 以及各种设计准则和规程, 这些都属于规范设计方法。例如:

- (1) 关系模式的设计方法
- (2) 新奥尔良(New Orleans)方法
- (3) 基于 E-R 模型的数据库设计方法
- (4) 3NF(第三范式)的设计方法
- (5) 基于抽象语法规则的设计方法
- (6) 计算机辅助数据库设计方法

这些数据库设计方法中比较著名的是新奥尔良(New Orleans)方法, 它将数据库设计分为四个阶段: 需求分析(分析用户要求)、概念设计(信息分析和定义)、逻辑设计(设计实现)和物理设计(物理数据库设计)。

2. 数据库设计的基本步骤

从数据库应用系统设计和开发的全过程来考虑, 一般将数据库设计的步骤分为七个阶段: 系统规划、需求分析、概念结构设计、逻辑结构设计、物理设计、实施、运行和维护。

- (1) 系统规划阶段

系统规划阶段是确定数据库系统在整个企业管理系统中的地位,确定系统的范围,确定开发工作所需的资源(人员、硬件和软件),确定项目进度,估算软件开发的成本及系统可能达到的效益。

(2)需求分析阶段

需求分析阶段是整个设计过程的基础,是最困难、最耗费时间的一个阶段。这一阶段要求计算机人员(系统分析员)和用户共同收集数据库所需要的信息内容,以及用户对处理的要求,并加以规格化和分析,以书面形式确定下来,作为以后验证系统的依据。

(3)概念结构设计阶段

概念结构设计阶段是整个数据库设计的关键。它通过对用户需求进行综合、归纳与抽象,形成一个独立于具体 DBMS 的概念模型,可以用 E-R 图来表示。

(4)逻辑结构设计阶段

逻辑结构设计阶段是将概念结构(E-R图)转换为某个 DBMS 所支持的数据模型,并对其进行优化。

(5)物理设计阶段

数据库的物理设计阶段是为逻辑数据模型选取一个最适合应用环境的物理结构(包括存储结构和存取方法)。

(6)实施阶段

数据库设计人员运用 DBMS 提供的数据库语言及其宿主语言,根据逻辑设计和物理设计的结果建立数据库,编制与调试应用程序,组织数据入库,并进行试运行。

(7)运行和维护阶段

数据库应用系统经过试运行后即可投入正式运行。在数据库系统运行过程中必须不断地对其进行评价、调整与修改。

数据库设计的七个阶段的划分目前尚无统一的标准,各阶段间相互连接,而且常常需要回溯修正。

4.3.2 任务实践

任务:数据库设计的基本步骤有哪些?

任务解析:

一般将数据库设计的基本步骤分为七个阶段:

- (1)系统规划阶段
- (2)需求分析阶段
- (3)概念结构设计阶段
- (4)逻辑结构设计阶段
- (5)物理结构设计阶段
- (6)数据库的实施阶段
- (7)数据库的运行和维护阶段

4.3.3 思考与练习

1. 下列描述正确的是()。

- A. 采用合理的数据库设计方法,可以确保数据库系统的设计质量,降低系统运行后的维护代价
- B. 数据库设计是涉及多学科的综合性的技术,也是一项庞大的软件开发工程
- C. 早期数据库设计主要采用手工与经验相结合的方法,设计质量难以保证
- D. 以上三项均正确
2. 下列描述错误的是()。
- A. 需求分析阶段是整个设计过程的基础,是最困难、最耗费时间的一个阶段
- B. 概念结构设计阶段是整个数据库设计的关键
- C. 逻辑结构设计阶段是将概念结构(E-R图)转换为某个DBMS所支持的数据模型
- D. 物理设计阶段是为概念结构模型选取一个最适合应用环境的物理结构(包括存储结构和存取方法)

4.4 系统规划阶段

4.4.1 基本知识

1. 系统规划的任务

系统规划阶段的主要任务就是进行系统的必要性和可行性分析。包括明确应用系统的基本功能,划分数据库支持的范围;规划人力资源调配;拟定设备配置方案;选择合适的操作系统、DBMS和其他软件;设备配置方案要在使用要求、系统性能、购置成本和维护代价各方面综合权衡;对系统的开发、运行、维护的成本做出估算;预测系统效益的期望值;拟定开发进度计划,还要对现行工作模式如何向新系统过渡做出具体安排。

2. 系统规划的成果

规划阶段的工作成果是写出详尽的可行性分析报告和数据库应用系统规划书。内容应包括:系统的定位及其功能、数据资源及数据处理能力、人力资源调配、设备配置方案、开发成本估算、开发进度计划等。

可行性分析报告和数据库应用系统规划书经审定立项后,成为后续开发工作的总纲。

4.4.2 任务实践

任务:判断“系统规划阶段可有可无”这句话是否正确。

任务解析:

在近些年来的数据库设计过程当中,把系统规划阶段放到数据库设计的第一步骤,也说明了系统规划阶段在数据库设计过程中的必要性。规划阶段的主要任务是进行建立数据库的必要性及可行性分析。如系统调查(即对企业全面调查,画出组织层次图,以了解企业组织结构)、可行性分析、确定DBS(数据库系统)的总目标和制定项目开发计划。如果,系统规划阶段通不过,那么后续的数据库设计的其他步骤也就不需要进行,所以,系统规划阶段是必须要有的。题目中的描述是错误的。

4.4.3 思考与练习

系统可行性分析报告在以下哪个阶段产生()。

- A. 系统规划阶段
- B. 需求分析阶段
- C. 数据库实施阶段
- D. 数据库运行和维护阶段

4.5 需求分析

4.5.1 基本知识

1. 需求分析的任务

需求分析是整个数据库设计过程中最重要的步骤之一,是后续各阶段的基础。需求分析的主要任务是通过详细调查所要处理的对象,包括某个组织、某个部门、某个企业的业务管理等,充分了解原手工或原计算机系统的工作状况以及工作流程,明确用户的各种需求,生成业务流程图和数据流图,然后在此基础上确定新系统的功能,并撰写系统说明书。新系统不能只按当前应用需求来设计数据库,必须充分考虑今后可能的扩充和改变。

在需求分析阶段,从多方面对整个组织进行调查、收集和分析各项应用对信息和处理两方面的需求。需求分析的重点是调查、收集和分析用户对数据管理中的信息要求、处理要求、安全性与完整性要求。信息要求是指用户需要从数据库中获得信息的内容与性质。由信息要求可以导出数据要求,即在数据库中需要存储哪些数据。处理要求是指用户要完成什么处理功能,对处理的响应时间有什么要求,处理方式是批处理还是联机处理。新系统的功能必须能够满足用户的多种需求。

2. 需求分析的步骤

调查、收集和分析用户要求的具体步骤如下:

(1) 调查组织机构情况

调查这个组织由哪些部门组成,各部门担当的职责是什么。

(2) 调查各部门的业务活动情况

调查各部门所需输入和使用的数据,如何加工处理这些数据,输出什么信息,输出到哪个部门,输出结果的格式等。

(3) 协助用户明确对新系统的各种要求

进一步明确用户对数据管理中的信息要求、处理要求、安全性与完整性要求。

(4) 确定新系统的边界

确定哪些功能由计算机完成或将来准备让计算机完成,哪些功能由人工完成。由计算机完成的功能就是新系统应该实现的功能。

3. 需求分析的调查方法

根据不同的问题和条件,调查方法也可以不同。常用的调查方法有以下几种。

(1) 跟班作业

通过亲身参加业务工作来了解业务活动的情况,这种方法可以比较准确地了解用户的需求,但比较耗费时间。

(2) 开调查会

通过与用户座谈的方式来了解业务活动情况及用户需求。

(3) 请专人介绍

通过邀请熟悉业务的专业人士来了解业务活动情况。

(4) 询问

对调查中的某些问题,可以找专人询问。

(5) 设计调查表请用户填写

如果调查表设计合理,这种方法易于用户接受并且会很有效。

(6) 查阅记录

查阅与原系统有关的数据记录,包括原始的单据、报表等。

当需求分析完成后,最终产生阶段性的成果:系统需求说明书,包括数据流图、数据字典、数据表格、系统功能结构图和必要的说明。

4. 数据流图

数据流图(Data Flow Diagram, 简记为 DFD) 是用图形方式来表达系统的逻辑功能,以及数据在系统内部的逻辑流向和逻辑变换过程。任何一个系统都可以抽象为图 4.2 所示的数据流图形式。

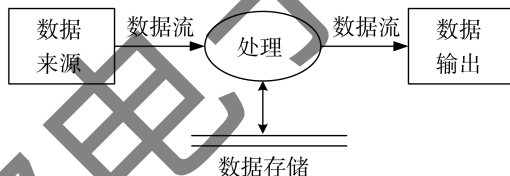


图 4.2 数据流图

(1) 数据流图的基本符号

→: 箭头, 表示数据流;

□: 方框, 表示数据的源点或终点;

○: 圆或椭圆, 表示加工或处理;

=: 双杠, 表示数据存储。

①数据流: 是数据在系统内传播的路径, 因此由一组成成分固定的数据组成。例如订票单由旅客姓名、年龄、单位、身份证号、日期、目的地等数据项组成。由于数据流是流动中的数据, 所以必须有流向, 除了与数据存储之间的数据流不用命名外, 数据流应该用名词或名词短语命名。

②数据源点或终点: 代表系统之外的实体, 可以是人、物或其他软件系统。

③对数据的加工(处理): 是对数据进行处理的单元, 它接收一定的数据输入, 对其进行处理, 并产生输出。

④数据存储: 表示信息的静态存储, 可以代表文件、文件的一部分、数据库的元素等。

(2) 在画数据流图时须注意的原则

①一个加工的输出数据流不应与输入数据流同名, 即使它们的组成成分相同。

②保持数据守恒,即一个加工的所有输出数据流中的数据必须能从该加工的输入数据流中直接获得。

③每个加工必须既有输入数据流,又有输出数据流。

④所有的数据流必须以一个加工开始,或以一个加工结束。

(3)数据流图的实例

如图4.3是一个飞机机票预订系统的数据流图,它反映的功能是:旅行社把预订机票的旅客信息(姓名、年龄、性别、身份证号码、旅行时间、目的地等)输入机票预订系统。系统为旅客安排航班,打印出取票通知单(附有应交的账款)。旅客在飞机起飞的前一天凭取票通知单交款取票,系统检验无误,输出机票给旅客。

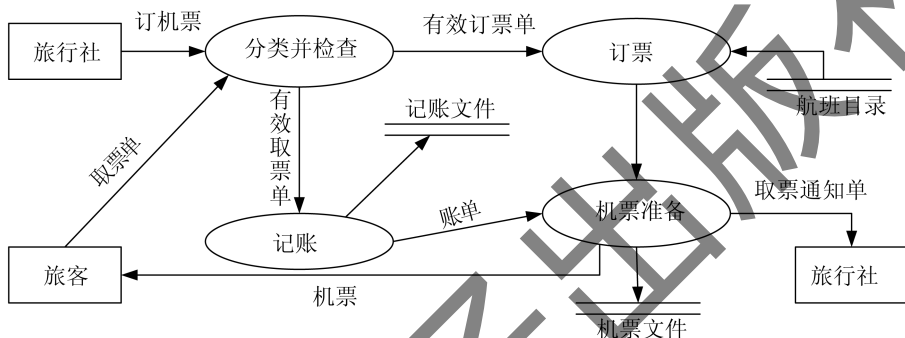


图 4.3 飞机机票预订系统

5. 数据字典

数据字典是系统中各类数据描述的集合,是对数据流图中包含的所有元素的定义的集合。

数据存放于物理数据库中,由数据库管理系统进行管理。数据字典有助于对这些数据进一步管理和控制,为设计人员和数据库管理员在数据库设计、实现和运行阶段控制有关数据提供一定的依据。

数据字典通常包括数据项、数据结构、数据流、数据存储和处理过程五个部分。

(1)数据项是数据的最小组成单位,是不可再分的数据单位。包括项名、含义说明、别名、数据类型、长度、取值范围、与其他数据项的逻辑关系等。

描述:

数据项名:学号

数据项含义:唯一标识每个学生

别名:学生编号

类型:字符型

长度:10

取值范围:0000000000~9999999999

取值含义:前四位标识该学生所在的年级,后六位按顺序编号。

与其他数据项的逻辑关系:该项等于另两项之和。

数据项之间的联系:根据语义写出数据项之间的数据依赖。

(2)数据结构反映了数据之间的组合关系。一个数据结构可以由若干个数据项组成,也可以由

若干个数据结构组成,或由若干个数据项和数据结构混合组成。包括数据结构名、说明、组成等。

描述:

数据结构名:学生

含义说明:是学籍管理子系统的主体数据结构,定义了一个学生的相关信息。

组成:学号,姓名,年龄,性别,所在系,年级

(3)数据流是数据结构在系统内传输的路径。包括数据流名、说明、数据流来源、数据流去向、组成、平均流量、高峰期流量等。

描述:

数据流名:体检结果

说明:学生参加体格检查的最终结果

数据流来源:体检(说明该数据流来自哪个过程)

数据流去向:批准(说明该数据流将到哪个过程去)

组成:身高,体重,视力,血压……

平均流量:单位时间内传输的次数

高峰期流量:最高时期的数据流量

(4)数据存储说明数据流中需要存储的数据,包括数据存储名、说明、流入数据流、流出数据流、组成、数据量、存取频度、存取方式等。

描述:

数据存储名:学生登记表

说明:记录学生的基本信息

输入数据流:指数据来源,如:报到时填的表。

输出数据流:指数据去向,如:学生基本情况表。

组成:数据结构或数据项,如:学号,姓名,年龄,性别,所在系,年级,专业等。

数据量:如每年 5000 张。

存取频度:指每小时或每天或每周存取几次、每次存取多少数据等信息。

存取方式:指是批处理,还是联机处理;是检索,还是更新;是顺序检索,还是随机检索。

(5)处理过程的具体处理逻辑通常用判定表或判定树来描述。包括处理过程名、说明、输入数据流、输出数据流、处理简要说明等。

描述:

处理过程名:分配宿舍

说明:为所有新生分配宿舍

输入数据流:如学生、宿舍

输出数据流:宿舍安排

处理:在新生报到后,为所有新生分配宿舍,要求相同性别的学生才可以居住在同一个房间里,一个人只能有一间宿舍,每个人的居住面积大于等于 3 平方米,安排新生宿舍的处理时间不得超过 20 分钟。

4.5.2 任务实践

任务 1:判断“需求分析阶段的主要任务是产生数据存储结构和存储方法,然后在此基础上

确定新系统的功能,并撰写需求说明书”这句话是否正确。

任务解析:

需求分析阶段的主要任务是通过详细调查所要处理的对象,包括某个组织、某个部门、某个企业的业务管理等,充分了解原手工或原计算机系统的工作状况以及工作流程,明确用户的各种需求,产生数据流图和数据字典,然后在此基础上确定新系统的功能,并撰写需求说明书。所以,题目中描述是错误的。

任务2:简述需求分析常用的调查方法及其调查目的。

任务解析:

需求分析常用的调查方法有以下几种:

- (1)跟班作业。获得业务活动情况和用户需求,特点是用户需求准确但比较费时。
- (2)开调查会。获得不同业务之间的联系信息,特点是与会人员能相互启发。
- (3)请专人介绍。通过邀请熟悉业务的专业人士来了解业务活动情况。
- (4)询问。对调查中的某些问题,可以找人询问,了解该用户业务范围的需求。
- (5)设计调查表请用户填写。获得设计人员关心的用户需求问题。特点是调查目的准确,但效果依赖于调查表设计的质量。
- (6)查阅记录。获得具体的业务细节。

4.5.3 思考与练习

1. 下列不属于需求分析阶段工作的是()。
A. 分析用户活动
B. 建立 E-R 图
C. 建立数据字典
D. 建立数据流图
2. 在数据库设计的需求分析阶段,业务流程一般采用()表示。
A. E-R 图
B. 数据流图
C. 程序结构图
D. 程序框图
3. 简述数据字典包括的五个部分的意义。

4.6 概念结构设计

4.6.1 基本知识

1. 概念结构设计方法

概念结构设计是整个数据库设计的关键,其主要任务是在需求分析阶段产生的需求说明书的基础上,按照特定的方法把它们抽象为一个不依赖于任何具体机器的数据模型,即概念模型。

概念结构的设计方法通常有以下四种。

自顶向下:先定义全局概念结构 E-R 模型的框架,再逐步细化。

自底向上:先定义各局部应用的概念结构 E-R 模型,然后将它们集成,得到全局概念结构 E-R 模型。

逐步扩张:先定义最重要的核心概念 E-R 模型,然后向外扩充,以滚雪球的方式逐步生成其他概念结构 E-R 模型,直至总体概念结构。

混合策略:该方法采用自顶向下和自底向上相结合的方法,先自顶向下定义全局框架,再以它为骨架集成自底向上方法中设计的各个局部概念结构。

其中最经常采用的策略是自底向上方法,即自顶向下地进行需求分析,然后再自底向上地设计概念结构。主要步骤包括进行数据抽象,设计局部概念结构,将局部概念结构合并成全局概念结构,并进行优化。

2. E-R 设计方法的介绍

描述概念模型的有力工具是 E-R 模型。有关 E-R 模型的基本概念已经在第一章介绍过了,下面将用 E-R 模型来描述概念结构。

(1) E-R 方法的基本术语

E-R 方法是“实体-联系方法”(Entity-Relationship Approach)的简称。它是描述现实世界概念结构模型的有效方法。用 E-R 方法建立的概念结构模型称为 E-R 模型,或称为 E-R 图。

E-R 图的三要素是实体、属性和联系。

①实体:用矩形框表示,框内标注实体名称,如图 4.4 所示。



图 4.4 实体

②属性:用椭圆形框表示,框内标注属性名称,如图 4.5 所示。

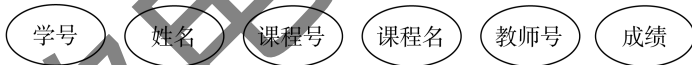


图 4.5 属性

③联系:用菱形框表示,框内标注实体之间的关系。有 1:1, 1:n 和 m:n 三种联系类型。例如系主任领导系,学生选修课程,教师讲授课程,工人生产产品,这里“领导”、“选修”、“讲授”、“生产”表示实体之间的联系,可以作为联系名称。联系用菱形框表示,框内标注联系名称,如图 4.6 所示。

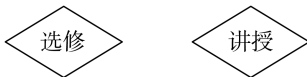


图 4.6 联系

(2) E-R 图的表示

在 E-R 图的描述中,用矩形表示实体,用椭圆表示属性,用菱形表示联系。在各框图内标注它们的名称,它们之间用无向线连接,表示联系时需在线上标明属于哪种类型的联系,如图 4.7 所示。

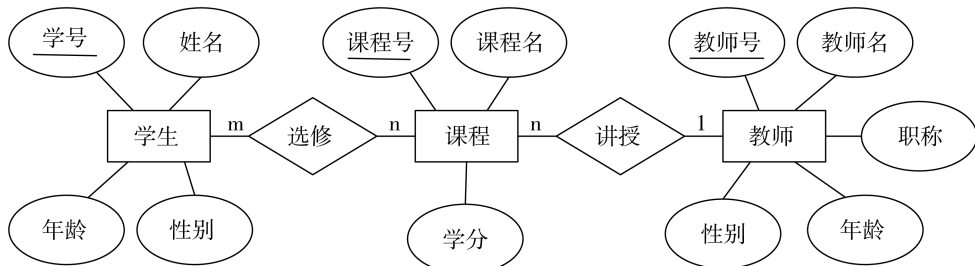
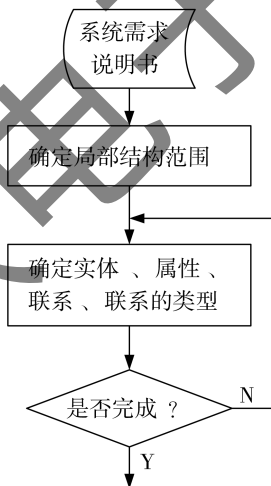


图 4.7 E-R 图的表示

采用 E-R 方法进行概念结构设计, 可以按照局部概念结构设计阶段和全局概念结构设计阶段两步进行, 在全局概念结构设计的过程中要不断进行概念结构的优化。

3. 局部概念结构设计

概念结构设计首先要根据需求分析得到的结果(数据流图、数据字典等)对现实世界进行抽象, 设计各个局部 E-R 模型。在系统需求分析阶段, 得到了多层数据流图、数据字典和系统分析报告。建立局部 E-R 模型, 就是根据系统的具体情况, 在多层的数据流图中选择一个适当层次的数据流图, 作为设计局部 E-R 图的出发点, 让这组图中每一部分对应一个局部应用。在前面选好的某一层次的数据流图中, 每个局部应用都对应了一组数据流图, 局部应用所涉及的数据存储在数据字典中。现在就是要将这些数据从数据字典中抽取出来, 参照数据流图, 确定每个局部应用包含哪些实体, 这些实体又包含哪些属性, 以及实体之间的联系及其联系类型。局部 E-R 模型设计的步骤如图 4.8 所示。



进入“全局 E-R 模型设计”

图 4.8 局部 E-R 模型设计的步骤

例 4.1 以工厂管理为例, 描述局部 E-R 图的设计。从技术科获知, 每种产品由多种零件组成, 每种零件可用在不同的产品上, 每种产品由一定数量的零件组成。从供应科获知, 每种零件使用多种材料制成, 每种材料也可应用在不同的零件上, 每种零件在使用材料上有一个使用量; 每个仓库可以存放多种材料, 每种材料只能放在一个仓库里, 每个仓库存放材料有一个库存量。

根据 E-R 图的建立过程:

第一步, 确定实体类型。

产品、零件、材料和仓库四个实体类型。

第二步,确定联系类型。

产品和零件之间是 $m:n$ 组成的联系,零件和材料之间是 $m:n$ 使用的联系,仓库和材料之间是 $1:m$ 存放的联系。

第三步,确定实体类型和联系类型的属性。

在技术科中,产品实体的属性有:产品号、产品名、性能参数等。

在技术科中,零件实体的属性有:零件号、零件名、价格等。

在供应科中,零件实体的属性有:零件号、规格等。

在供应科中,材料实体的属性有:材料号、价格等。

在供应科中,仓库实体的属性有:仓库号、仓库名、地址等。

产品和零件之间 $m:n$ 组成的联系属性是零件数,零件和材料之间 $m:n$ 使用的联系属性是使用量,仓库和材料之间 $1:m$ 存放的联系属性是库存量。

第四步,根据实体类型和联系类型画出局部 E-R 图,如图 4.9 和图 4.10 所示。

第五步,用下划线标注出实体标识符。

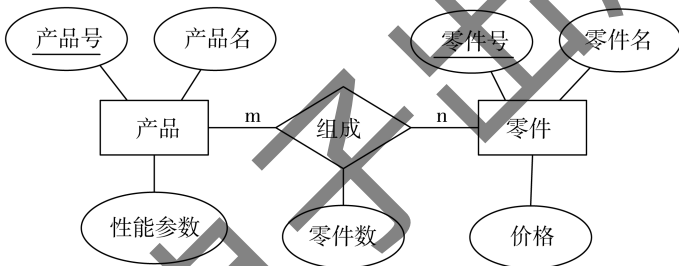


图 4.9 技术科的局部 E-R 图

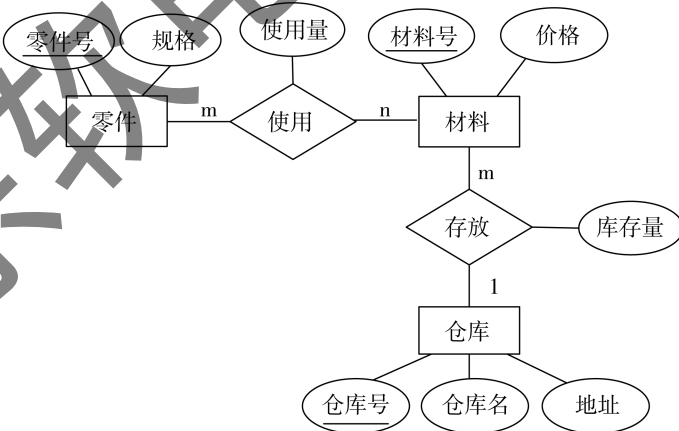


图 4.10 供应科的局部 E-R 图

4. 全局概念结构设计

全局概念结构设计的实质是把局部概念结构设计中所有的局部概念模型统一起来,形成一个完整的系统模型。全局 E-R 模型的设计过程如图 4.11 所示。

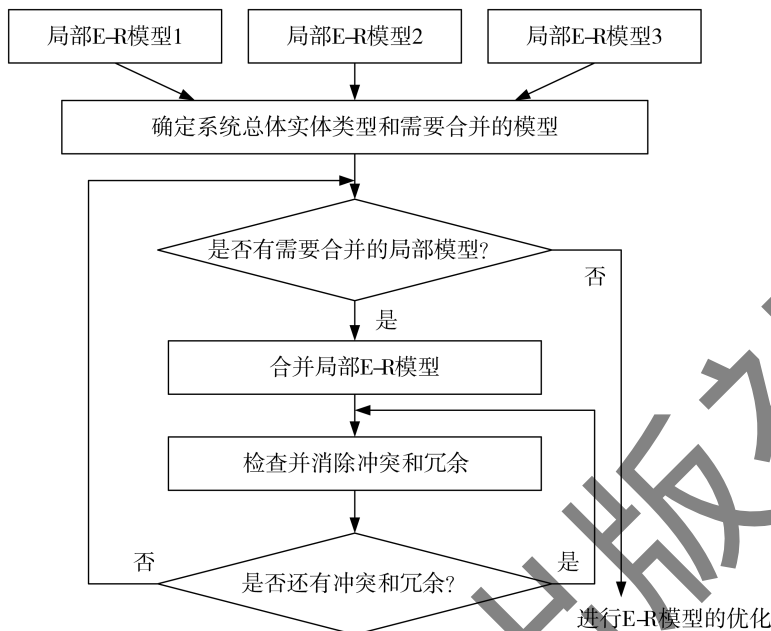


图 4.11 全局 E-R 模型的设计步骤

全局 E-R 模型的建立过程如下：

(1) 合并

将局部概念模型整理合并成全局概念模型。

①先找出具有相同实体的两个 E-R 图。

②以该相同实体为基准进行合并。

③如果还有相同实体的 E-R 图,再次合并。

④这样一直下去,直到所有的具有相同实体的局部 E-R 图都被合并,从而得到全局的 E-R 图。

(2) 消除冲突

解决各种局部 E-R 图之间的冲突问题,生成初步 E-R 图。

①属性冲突

属性值的类型、取值范围及取值单位不一致造成的冲突。如:生日和年龄,厘米和米,学生编号的方式等。

②结构冲突

如在某局部 E-R 图中系主任是属性,而在另一个局部 E-R 图中系主任是实体等。

③命名(实体、属性、联系)冲突

同名异义:教室和宿舍均称为房间;

异名同义:如教材和课本。

将例 4.1 中技术科和供应科的两个局部 E-R 图合并成全局 E-R 图。

图 4.12 是工厂管理中的两个局部 E-R 图按照相同的实体“零件”合并后得到的全局 E-R 图。

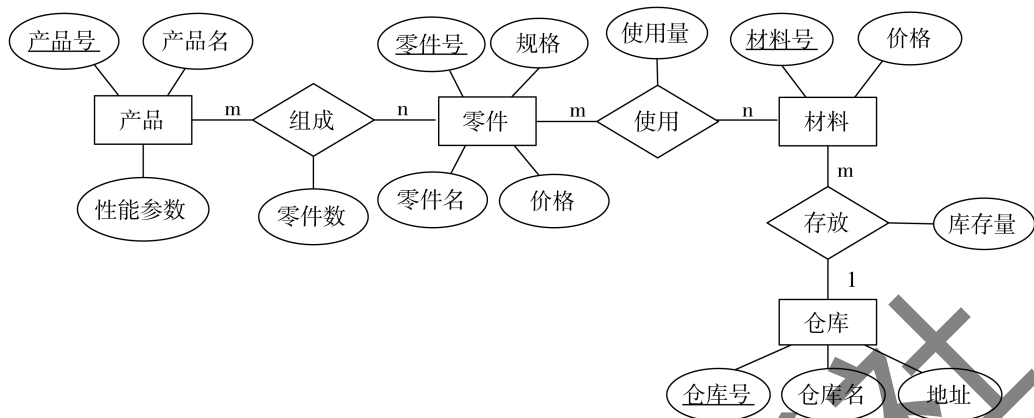


图 4.12 工厂管理的全局 E-R 图

按照上面的方法将各个局部 E-R 模型合并后就得到一个初步的全局 E-R 模型,之所以这样称呼是因为其中可能存在冗余的数据和冗余的联系等。因此,在得到初步的全局 E-R 模型后,还应当进一步检查 E-R 图中是否存在冗余,如果存在冗余则一般应设法将其消除。

一个好的全局 E-R 模型除了能准确、全面地反映用户功能外,还应满足下列条件:实体类型的个数尽可能少、实体类型所含属性的个数尽可能少、实体间联系的冗余最小。模型优化的目的是消除不必要的冗余,使其保持最小冗余度。

优化全局 E-R 模型的几个原则:

- 实体类型的合并,如图 4.13 系主任和系。

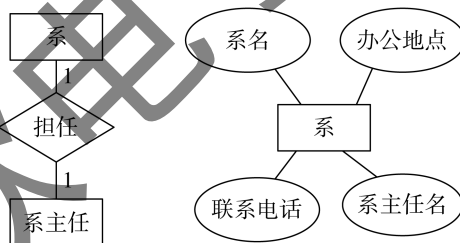


图 4.13 系主任和系实体类型的合并

- 冗余属性的消除,如生日和年龄。
- 冗余联系的消除。

将图 4.14 和图 4.15 所示的两个局部 E-R 图合并成一个全局 E-R 图,并进行优化。

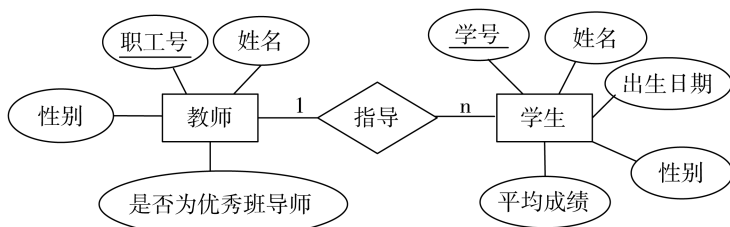


图 4.14 班导师工作局部 E-R 图

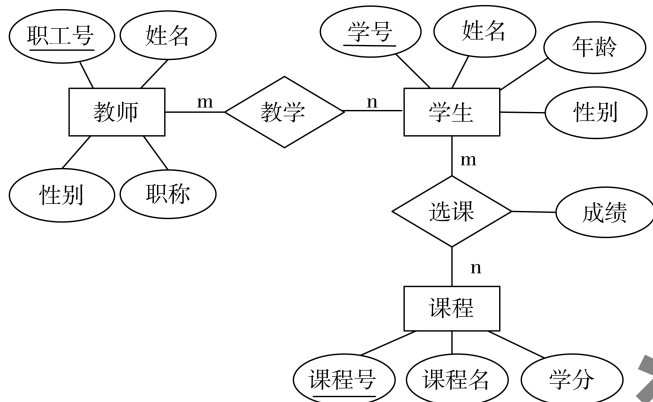


图 4.15 教学活动局部 E-R 图

局部 E-R 模型设计完成之后,根据全局 E-R 模型的建立步骤,将上述两个局部 E-R 模型合并成全局 E-R 模型,如图 4.16 所示。

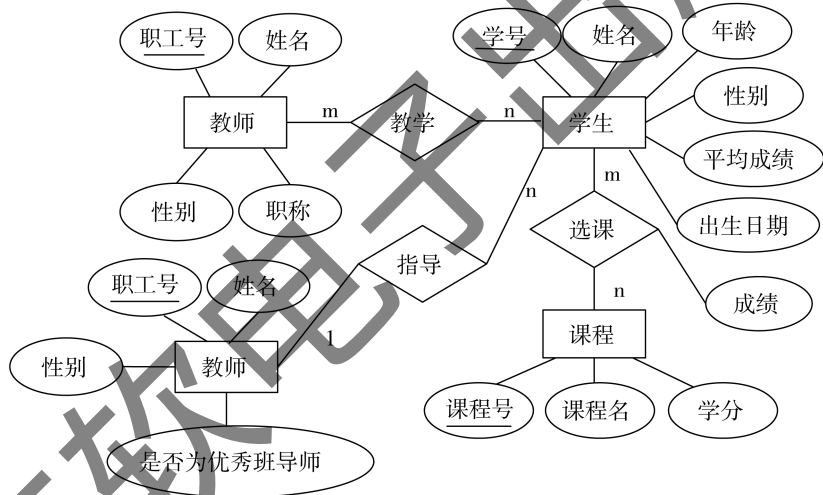


图 4.16 合并后的全局 E-R 图

全局概念结构不仅要支持所有的局部 E-R 模型,而且要合理地表示一个完整、一致的数据库概念结构。由于各个局部应用不同,通常由不同的设计人员进行局部 E-R 图设计,因此,各局部 E-R 图不可避免地会有许多不一致的地方,我们称之为冲突。

在图 4.16 合并的 E-R 模型中存在冲突。实体存在冗余,教学和指导两个联系存在冗余,年龄和出生日期两个属性存在冗余,成绩和平均成绩两个属性存在冗余。

合并局部 E-R 图时并不能简单地将各个局部 E-R 图画到一起,而必须消除各个局部 E-R 图中的不一致,使合并后的全局概念结构不仅支持所有的局部 E-R 模型,而且必须是一个能为全系统中所有用户共同理解和接受的完整的概念模型。上例中对全局 E-R 模型进行优化,消除冗余,得到优化后的概念模型如图 4.17 所示。

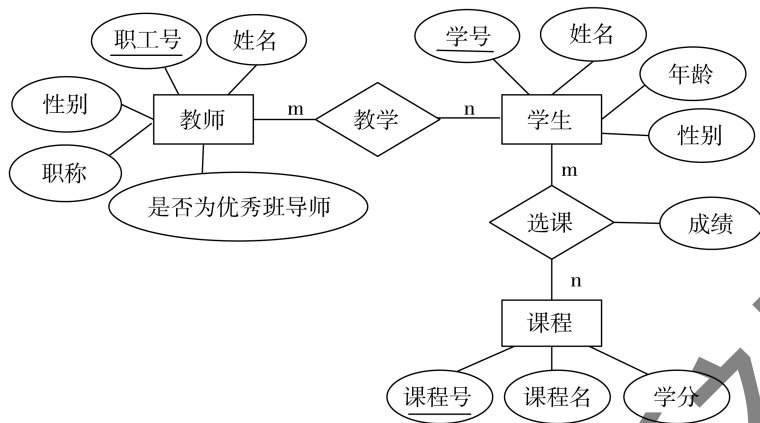


图 4.17 优化后的全局 E-R 图

4.6.2 任务实践

任务 1: E-R 图中的联系可以与几个实体有关。

任务解析：

一个实体内部可以有联系，实体和实体之间也可以有联系，所以联系可以涉及 1 个或 1 个以上的实体。

任务 2: 什么是 E-R 图？构成 E-R 图的基本要素是什么？

任务解析：

E-R 图为实体-联系图，提供了表示实体、属性和联系的方法，用来描述现实世界的概念模型。

构成 E-R 图的基本要素是实体、属性和联系，其表示方法为：

实体：用矩形表示，矩形框内写明实体名；

属性：用椭圆形表示，并用无向边将其与相应的实体连接起来；

联系：用菱形表示，菱形框内写明联系名，并用无向边分别与有关实体连接起来，同时在无向边旁标上联系的类型（1:1, 1:n 或 m:n）。

任务 3: 一个图书借阅信息管理系统有如下信息：

每一个借书人可以借阅多本图书，每一本图书可以被多个借书人借阅；借书人每借阅一本图书都有一个借书日期和还书日期；每一个出版社可以出版多本图书，每一本图书只能在一个出版社出版。

其中，借书人的属性有：借书证号、姓名、单位、电话；图书的属性有：图书编号、书名、位置；出版社的属性有：出版社名、地址、邮编、电话。

根据需求画出 E-R 图，并在 E-R 图中注明实体的属性、联系的类型以及实体的码。

任务解析：

E-R 图如图 4.18 所示：

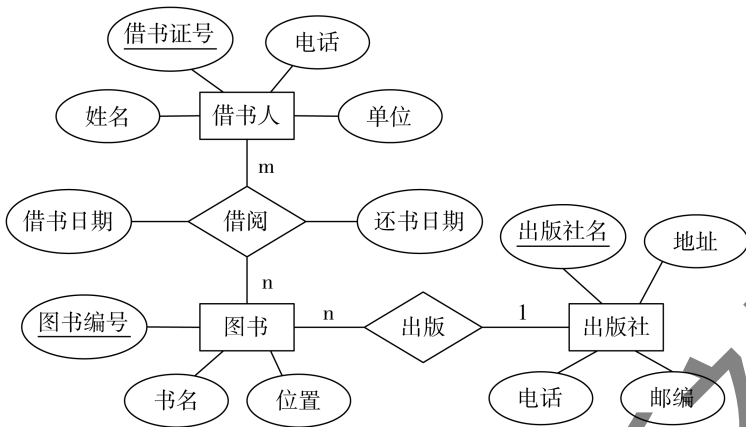


图 4.18 图书借阅 E-R 图

4.6.3 思考与练习

1. 在数据库设计中,建立 E-R 模型属于下列哪个阶段()。
A. 需求分析阶段
B. 概念结构设计阶段
C. 逻辑结构设计阶段
D. 物理结构设计阶段

2. 下列不属于全局 E-R 模型优化时要达到的目的是()。

- A. 实体类型的个数尽可能少
- B. 实体类型所含属性的个数尽可能少
- C. 实体间联系的冗余最小
- D. 实体完整性和参照完整性

3. 简述合并局部 E-R 模型的主要步骤。

4. 某动物园管理中心有如下信息:

动物园有多个笼舍,每个笼舍只属于这一个动物园;每个笼舍可以安置一种动物,每种动物住在一个笼舍里;每名饲养员喂养多种动物,每种动物可由多名饲养员喂养,饲养员每次喂养动物时有一个喂养时间;每个游客可以观赏多种动物,每种动物可以供多名游客观赏。

其中,动物园的属性有:动物园名、地点、电话;笼舍的属性有:笼舍编号、规模、位置;动物的属性有:动物编号、动物名称、产地、所属科目;饲养员的属性有:饲养员编号、姓名、年龄、职位;游客的属性有:身份证号、姓名、性别。

根据需求画出 E-R 图,并在 E-R 图中注明实体的属性、联系的类型以及实体的码。

4.7 逻辑结构设计

4.7.1 基本知识

概念结构设计所得的 E-R 模型是对用户需求的一种抽象的表达形式,它独立于任何一种具体的数据模型,因而也不能为任何一个具体的 DBMS 所支持。为了能够建立起最终的物理

系统,还需要将概念结构进一步转化为某一 DBMS 所支持的数据模型,然后根据逻辑设计的准则、数据的语义约束、规范化理论等对数据模型进行适当的调整和优化,形成合理的全局逻辑结构,并设计出用户子模式。这就是数据库逻辑设计所要完成的任务。

1. 逻辑结构设计的步骤

由于各种 DBMS 产品一般都有许多限制,提供不同的环境与工具,因此,逻辑设计分为如下几步:

- (1) 将概念模型向一般关系、网状和层次模型转化;
- (2) 将得到的一般关系、网状和层次模型向特定的 DBMS 产品所支持的数据模型转化;
- (3) 依据应用的需求和具体的 DBMS 的特征进行调整和完善。

数据库逻辑结构的设计过程如图 4.19 所示。

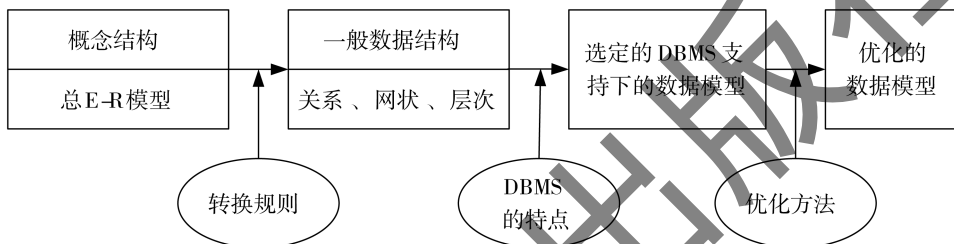


图 4.19 逻辑结构设计的过程

某些早期设计的应用系统中还在使用网状或层次数据模型,而新设计的数据库应用系统都普遍采用支持关系数据模型的 RDBMS,所以这里只介绍 E-R 图向关系数据模型的转换原则与方法。

2. E-R 图向关系模型的转换原则

关系模型的逻辑结构是一组关系模式的集合,而 E-R 图则是由实体、实体的属性和实体之间的联系三个要素组成的。所以将 E-R 图转换为关系模型实际上就是要将实体、实体的属性和实体之间的联系转化为相应的关系模式,下面具体介绍转换的规则。

(1) 一个实体类型转换为一个关系模式。实体的属性就是关系的属性,实体的码就是关系的码。

例 4.2 将图 4.20 中学生实体和课程实体分别转换成两个关系模式。



图 4.20 学生和课程实体

学生实体和课程实体分别转换成如下两个关系模式:

学生关系模式(学号,姓名,年龄,性别),学号为关系模式的主码。

课程关系模式(课程号,课程名,学分),课程号为关系模式的主码。

(2) 一个 $m:n$ 联系转换为一个独立的关系模式。与该联系相连的各实体的码以及联系本身的属性均转换为关系的属性。而关系的码为各实体码的组合。

例 4.3 将图 4.21 中学生选课 E-R 模型转换为相应的关系模式。

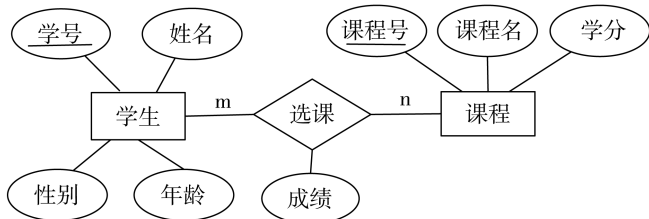


图 4.21 学生选课 E-R 图

将上述 E-R 模型转换为相应的关系模式,先将学生和课程两个实体转换为关系模式,再将这两个实体间的联系转换为关系模式,如下:

学生关系模式(学号,姓名,年龄,性别),学号为关系模式的主码。

课程关系模式(课程号,课程名,学分),课程号为关系模式的主码。

选课关系模式(学号,课程号,成绩),学号和课程号的组合码为关系模式的主码。

(3)一个 1:n 联系可以转换为一个独立的关系模式,也可以与 n 端对应的关系模式合并。如果转换为一个独立的关系模式,则与该联系相连的各实体的码以及联系本身的属性均转换为关系的属性,而该关系的码为 n 端实体的码。如果与 n 端对应的关系模式合并,则只需要将联系本身的属性和 1 端实体的码加入到 n 端对应的关系模式中即可。

例 4.4 将图 4.22 中班导师指导学生的 E-R 模型转换为相应的关系模式。

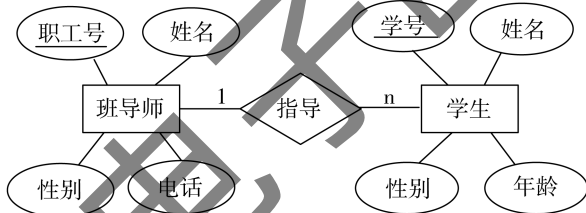


图 4.22 学生指导 E-R 图

将上述 E-R 模型转换为相应的关系模式,先将班导师和学生两个实体转换为关系模式,再将这两个实体间的联系转换为关系模式,如下:

方法一:产生独立的关系模式

学生关系模式(学号,姓名,年龄,性别),学号为关系模式的主码。

班导师关系模式(职工号,姓名,性别,电话),职工号为关系模式的主码。

指导关系模式(学号,职工号),学号为关系模式的主码。

方法二:与 n 端对应的关系模式合并

学生关系模式(学号,姓名,年龄,性别,职工号),学号为关系模式的主码。

班导师关系模式(职工号,姓名,性别,电话),职工号为关系模式的主码。

(4)一个 1:1 联系可以转换为一个独立的关系模式,也可以与任意一端对应的关系模式合并。如果转换为一个独立的关系模式,则与该联系相连的各实体的码以及联系本身的属性均转换为关系的属性,每个实体的码均是该关系的候选码。如果与某一端对应的关系模式合并,则需要在该关系模式的属性中加入另一个关系模式的码和联系本身的属性。

例 4.5 将图 4.23 中的 E-R 模型转换为相应的关系模式。

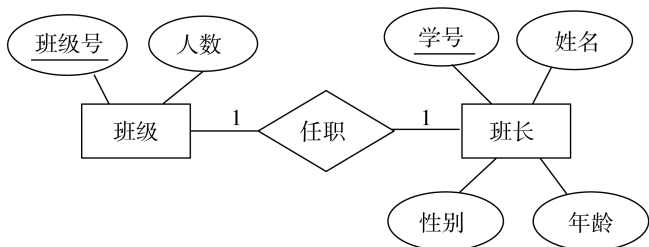


图 4.23 班长任职 E-R 图

将上述 E-R 模型转换为相应的关系模式,先将班级和班长两个实体转换为关系模式,再将这两个实体间的联系转换为关系模式,如下:

方法一:产生独立的关系模式

班级关系模式(班级号, 人数), 班级号为关系模式的主码。

班长关系模式(学号, 姓名, 性别, 年龄), 学号为关系模式的主码。

任职关系模式(班级号, 学号), 班级号为关系模式的主码, 也可以选学号作为关系模式的主码。

方法二:与任意一端对应的关系模式合并

班级关系模式(班级号, 人数), 班级号为关系模式的主码。

班长关系模式(学号, 姓名, 性别, 年龄, 班级号), 学号为关系模式的主码。

或者

班级关系模式(班级号, 人数, 学号), 班级号为关系模式的主码。

班长关系模式(学号, 姓名, 性别, 年龄), 学号为关系模式的主码。

(5) 三个或三个以上实体间的一个多元联系转换为一个关系模式。与该多元联系相连的各实体的码以及联系本身的属性均转换为关系的属性。而关系的码为各实体码的组合。

(6) 同一实体集的实体间的联系, 即自联系, 也可按上述 1:1、1:n 和 m:n 三种情况分别处理。

(7) 具有相同码的关系模式可合并。

例 4.6 每个工厂生产多种产品, 且每种产品可以在多个工厂中生产, 每个工厂按照固定的计划数量生产产品; 每个工厂聘用多名职工, 且每个职工只能在一个工厂工作, 工厂聘用职工有聘用期和工资。工厂的属性有工厂编号、厂名、地址, 产品的属性有产品编号、产品名、规格, 职工的属性有职工号、姓名。

(1) 根据需求画出 E-R 图, 并在 E-R 图中注明实体的属性、联系的类型以及实体标识符。

(2) 将 E-R 图转换成关系模式, 并用下划线标出每个关系模式的主码。

具体解题步骤如下:

(1) 根据题意, 建立全局 E-R 图如图 4.24 所示:

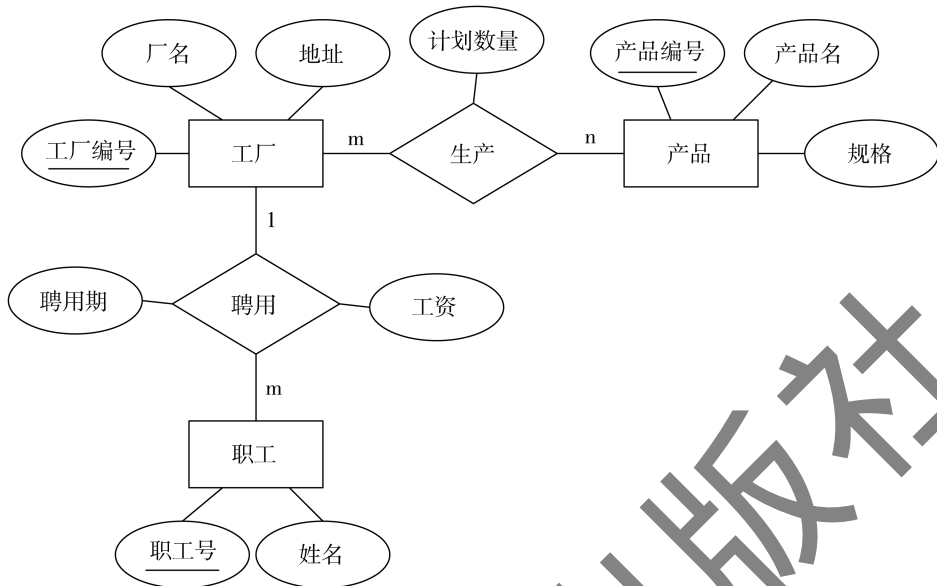


图 4.24 工厂管理 E-R 图

(2)将上述 E-R 图转换成相应的关系模式为:

工厂(工厂编号, 厂名, 地址)

产品(产品编号, 产品名, 规格)

职工(职工号, 姓名, 工厂编号, 工资, 聘用期)

生产(工厂编号, 产品编号, 计划数量)

3. 数据模型的优化

关系数据库逻辑设计的结果不是唯一的。在逻辑结构设计的基础上,根据需要对设计结构进行适当的调整和完善,以提高系统的性能。为了进一步提高数据库应用系统的性能,通常以规范化理论为指导,还应该适当地修改、调整数据模型的结构,这就是数据模型的优化。

数据模型的优化方法为:

(1)确定数据依赖。

(2)对于各个关系模式之间的数据依赖进行极小化处理,消除冗余的联系。

(3)按照数据依赖的理论对关系模式逐一进行分析,考查是否存在部分函数依赖、传递函数依赖、多值依赖等,确定各关系模式分别属于第几范式。

(4)按照需求分析阶段得到的各种应用对数据处理的要求,分析对于这样的应用环境这些模式是否合适,确定是否要对它们进行合并或分解。

(5)对关系模式进行必要的分解。

如果一个关系模式的属性特别多,就应该考虑是否可以对这个关系进行垂直分解。如果有些属性是经常访问的,而有些属性是很少访问的,则应该把它们分解为两个关系模式。如果一个关系的数据量特别大,就应该考虑是否可以水平分解。例如一个论坛中,如果设计时把会员发的主帖和跟帖设计为一个关系,则在帖子量非常大的情况下,就应该考虑把它们分开了。因为显示的主帖是经常查询的,而跟帖则是在打开某个主帖的情况下才查询。又如手机号管理软件,可以考虑按省份或其他方式进行水平分解。

4.7.2 任务实践

任务 1: 在 E-R 模型中, 如果有 5 个不同的实体集, 存在 2 个 1 : n 联系和 3 个 m : n 联系, 根据 E-R 模型转换为关系模型的规则, 该 E-R 图转换为关系模式的数目至少是多少个。

A. 5 个

B. 7 个

C. 8 个

D. 10 个

任务解析:

在 E-R 图向关系模型的转换原则中, 5 个不同的实体集要转换成 5 个不同的关系模式, 2 个 1 : n 联系可以转换为 2 个独立的关系模式, 也可以与 n 端对应的关系模式合并, 3 个 m : n 联系需要转换为 3 个独立的关系模式, 那么该 E-R 图转换为关系模式的数目至少是 $5+3=8$ 个。

任务 2: 某影院管理中心有如下信息:

影院内有多个放映厅, 每个放映厅只属于一个影院; 每个放映厅可以放映多部影片, 每部影片可以在不同的放映厅放映, 每部电影在放映厅放映时有放映时间; 每个观众可以观看多部影片, 每部影片也可以被多名观众观赏, 每个观众观赏影片时都有观看时间。

其中, 影院的属性有: 影院名、地址、电话; 放映厅的属性有: 厅名、规模; 电影的属性有: 许可证号、电影名、类型、时长; 观众的属性有: 身份证号、姓名、年龄。

(1) 根据需求画出 E-R 图, 并在 E-R 图中注明实体的属性、联系的类型以及实体的码。

(2) 将 E-R 图转换成关系模式, 并用下划线标出每个关系模式的主码。

任务解析:

(1) E-R 图如图 4.25 所示:

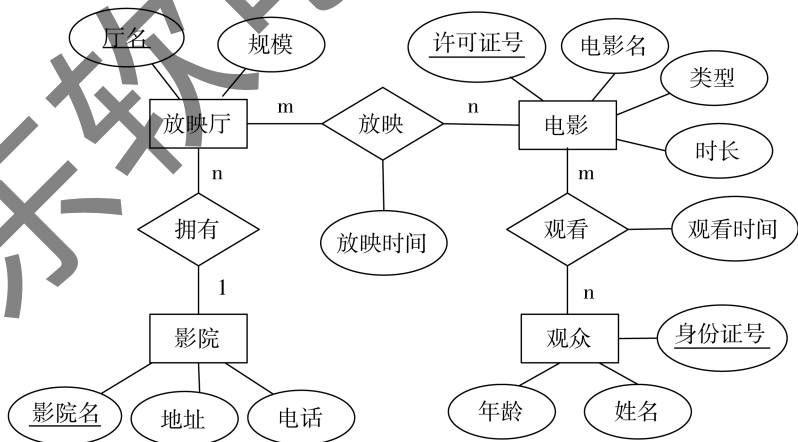


图 4.25 影院管理 E-R 图

(2) 关系模式如下:

影院(影院名, 地址, 电话)

放映厅(厅名, 规模, 影院名)

电影(许可证号, 电影名, 类型, 时长)

观众(身份证号,姓名,年龄)

放映(厅名,许可证号,放映时间)

观看(身份证号,许可证号,观看时间)

4.7.3 思考与练习

1. 下面有关 E-R 模型向关系模型转换的叙述中,不正确的是()。

A. 一个实体类型转换为一个关系模式

B. 一个 1:1 联系可以转换为一个独立的关系模式,也可以与联系的任意一端实体所对应的关系模式合并

C. 一个 1:m 联系可以转换为一个独立的关系模式,也可以与联系的任意一端实体所对应的关系模式合并

D. 一个 m:n 联系转换为一个独立的关系模式

2. 下述哪一条不是由于关系模式设计不当而引起的()。

A. 数据冗余

B. 插入异常

C. 更新异常

D. 丢失修改

3. 什么是数据库的逻辑结构设计? 试述其设计步骤。

4. 简述 E-R 图向关系模型的转换规则。

5. 假设要建立一个企业数据库,该企业有多个下属单位,信息如下:

每一单位有多个职工,一个职工仅隶属于一个单位,且一个职工仅在一个工程中工作,但一个工程中有很多职工参加工作,有多个供应商为各个工程供应一定数量的不同设备。

其中,单位的属性有:单位名、电话;职工的属性有:职工号、姓名、性别;设备的属性有:设备号、设备名、产地;供应商的属性有:供应商名、电话;工程的属性有:工程名、地点。

(1) 根据需求画出 E-R 图,并在 E-R 图中注明实体的属性、联系的类型以及实体的码。

(2) 将 E-R 图转换成关系模式,并用下划线标出每个关系模式的主码。

4.8 物理结构设计

4.8.1 基本知识

数据库物理设计阶段的任务是根据具体计算机系统(DBMS 和硬件等)的特点,为给定的数据库模型确定合理的存储结构和存取方法。所谓的“合理”主要有两个含义:一个是要使设计出的物理数据库占用较少的存储空间,另一个对数据库的操作具有尽可能高的速度。

数据库的物理设计通常分为两步:首先,确定数据库的物理结构,在关系数据库中主要指存取方法和存储结构。其次,对物理结构进行评价,评价的内容是系统的时间和空间效率。

1. 确定数据库的物理结构

(1) 确定数据的存储结构

确定数据库存储结构时要综合考虑存取时间、存储空间利用率和维护代价三方面的因素。这三个方面常常是相互矛盾的,例如消除一切冗余数据虽然能够节约存储空间,但往往会导致检索代价的增加,因此必须进行权衡,选择一个折中方案。

(2) 设计数据的存取路径

在关系数据库中,选择存取路径主要是指确定如何建立索引。例如,应把哪些域作为次码建立次索引,建立单码索引还是组合索引,建立多少个为合适,是否建立聚集索引等。

(3) 确定数据的存放位置

为了提高系统性能,数据应该根据应用情况将易变部分与稳定部分、经常存取部分和存取频率较低部分分开存放。

(4) 确定系统配置

DBMS 产品一般都提供了一些存储分配参数,供设计人员和 DBA 对数据库进行物理优化。初始情况下,系统都为这些变量赋予了合理的缺省值。但是这些值不一定适合每一种应用环境,在进行物理设计时,需要重新对这些变量赋值以改善系统的性能。

2. 评价物理结构

数据库物理设计过程中需要对时间效率、空间效率、维护代价和各种用户要求进行权衡,其结果可以产生多种方案,数据库设计人员必须对这些方案进行细致的评价,从中选择一个较优的方案作为数据库的物理结构。

评价物理数据库的方法完全依赖于所选用的 DBMS,主要是从定量估算各种方案的存储空间、存取时间和维护代价入手,对估算结果进行权衡、比较,选择出一个较优的合理的物理结构。如果该结构不符合用户需求,则需要修改设计。

4.8.2 任务实践

任务:对数据库的物理结构设计优劣评价的重点是什么?

任务解析:

数据库物理设计过程中需要对时间效率、空间效率、维护代价和各种用户要求进行权衡,对数据库的物理结构设计优劣评价的重点是时间和空间的效率问题。

4.8.3 思考与练习

1. 数据库设计中,确定数据库存储结构,即确定关系、索引、聚簇、日志、备份等数据的存储安排和存储结构,这是数据库设计的()。

- | | |
|-----------|-----------|
| A. 需求分析阶段 | B. 逻辑设计阶段 |
| C. 概念设计阶段 | D. 物理设计阶段 |

2. 下列不属于数据库物理设计阶段应考虑的问题是()。

- A. 存取方法的选择
 - B. 索引与入口设计
 - C. 确定数据的存放位置
 - D. 用户子模式设计
3. 简述数据库物理结构设计的内容和步骤。

4.9 数据库的实施

4.9.1 基本知识

在进行概念结构设计、逻辑结构设计和物理结构设计之后,设计者对目标系统的结构、功能已经分析得较为清楚了,但这还只是停留在文档阶段。数据库系统设计的根本目的,是为用户提供一个能够实际运行的系统,并保证该系统的稳定和高效。数据库的实施主要包括以下工作:用数据定义语言(DDL)定义数据库结构,组织数据入库,编制与调试应用程序,数据库试运行四个部分。

1. 定义数据库结构

确定了数据库的逻辑结构与物理结构后,就可以用所选用的 DBMS 提供的数据库定义语言(DDL)来严格描述数据库结构。

2. 数据装载

数据库结构建立好后,就可以向数据库中装载数据了。组织数据入库是数据库实施阶段最主要的工作。对于数据量不是很大的小型系统,可以用人工方式完成数据的入库,其步骤为:

(1) 筛选数据

需要装入数据库中的数据通常都分散在各个部门的数据文件或原始凭证中,所以首先必须把需要入库的数据筛选出来。

(2) 转换数据格式

筛选出来的需要入库的数据,其格式往往不符合数据库要求,还需要进行转换。这种转换有时可能很复杂。

(3) 输入数据

将转换好的数据输入计算机中。

(4) 校验数据

检查输入的数据是否有误。对于中大型系统,由于数据量极大,用人工方式组织数据入库将会耗费大量的人力和物力,而且很难保证数据的正确性。因此应该设计一个数据输入子系统由计算机辅助数据的入库工作。

3. 编制与调试应用程序

数据库应用程序的设计应该与数据库设计并行进行。在数据库实施阶段,当数据库结构建

立好后,就可以开始编制与调试数据库的应用程序,也就是说,编制与调试应用程序是与组织数据入库同步进行的。调试应用程序时由于数据入库尚未完成,可先使用模拟数据。

4. 数据库试运行

应用程序调试完成,并且已有一小部分数据入库后,就可以开始数据库的试运行。数据库试运行也称为联合调试,其主要工作包括:

(1) 功能测试

即实际运行应用程序,执行对数据库的各种操作,测试应用程序的各种功能。

(2) 性能测试

即测量系统的性能指标,分析是否符合设计目标。

由于在数据库设计阶段,设计者对数据库的评价多是在简化了的环境条件下进行的,因此设计结果未必是最佳的。在试运行阶段,除了对应用程序做进一步的测试之外,重点执行对数据库的各种操作,实际测量系统的各种性能,检测是否达到设计要求。如果在数据库试运行时,所产生的实际结果不理想,则应回过头来修改物理结构,甚至修改逻辑结构。

4.9.2 任务实践

任务:简述数据库实施阶段的主要工作。

任务解析:

数据库实施阶段的主要工作包括:用数据定义语言(DDL)定义数据库结构,组织数据入库,编制与调试应用程序,数据库试运行四个部分。

4.9.3 思考与练习

数据库实施的任务是()。

- A. 对系统的开发、运行、维护的成本做出估算
- B. 确定合理的存储结构和存取方法
- C. 定义数据库结构,组织数据入库,编制与调试应用程序和数据库试运行
- D. 转储和恢复,安全性与完整性控制,性能的监督、分析和改进,重组织和重构

4.10 数据库的运行和维护

4.10.1 基本知识

数据库试运行结果符合设计目标后,数据库就可以真正投入运行了。数据库投入运行标志着开发任务的基本完成和维护工作的开始,并不意味着设计过程的终结,由于应用环境在不断变化,数据库运行过程中物理存储也会不断变化,对数据库设计进行评价、调整、修改等维护工

- A. 只要数据库正式投入运行,就标志着数据库设计工作的结束
 - B. 数据库的维护工作就是维持数据库系统的正常运行
 - C. 数据库的维护工作就是发现错误,修改错误
 - D. 数据库正式投入运行标志着数据库运行和维护工作的开始
3. 数据库的重组组织不要求修改应用程序,也不改变数据的()。
- A. 输入输出
 - B. 数据库的关联顺序
 - C. 存储模式
 - D. 逻辑结构和物理结构

4.11 小结

本章首先介绍了关系数据库规范化理论,给出了函数依赖和范式的相关定义,通过实例给出如何规范关系模型和保证数据完整性。

其次介绍了数据库设计的基础知识,给出了数据库设计的方法和具体步骤。详细介绍了系统规划阶段、需求分析阶段、概念结构设计阶段、逻辑结构设计阶段、物理结构设计阶段、数据库的实施阶段、数据库的运行和维护阶段的目标、方法和应该注意的事项。

其中最重要的两个环节是概念结构设计阶段和逻辑结构设计阶段。重点介绍了在概念结构设计阶段中,将需求分析的结果转化为 E-R 模型的方法;在逻辑结构设计阶段中,将 E-R 图转换成关系模式的转换内容与转换原则。