如何设计数据库——范式与逆范式

数据库系统上机

计算机学院&网络空间安全学院 乜鹏

https://dbis.nankai.edu.cn/2019/0417/c12139a128118/page.htm

声明：上机课程的内容偏向举例，通俗化，一些术语的准确定义请查看理论课程。

投票 最多可选3项

南开大学 Nankai University

下图的学生表中，存在哪些问题？

学号	姓名	性别	课程号	课程名	成绩	课程号	课程名	成绩	居住地	邮编
2012001	张三	男	5	数学	88	4	英语	78	北京	100000
2012002	李四	女	5	数学	69	4	英语	83	上海	200000
2012003	王五	男	5	数学	52	4	英语	79	北京	100000
2012004	马六	女	5	数学	58	4	英语	81	上海	200000
2012005	田七	男	5	数学	92	4	英语	58	天津	300000

学生表

A 数据冗余

B 插入异常

C 删除异常

D 更新异常

E 我觉得挺好的啊，简单直观

允公允能 日新月异 NANKAI UNIVERSITY

问题

学号	姓名	性别	课程号	课程名	成绩	课程号	课程名	成绩	居住地	邮编
2012001	张三	男	5	数学	88	4	英语	78	北京	100000
2012002	李四	女	5	数学	69	4	英语	83	上海	200000
2012003	王五	男	5	数学	52	4	英语	79	北京	100000
2012004	马六	女	5	数学	58	4	英语	81	上海	200000
2012005	田七	男	5	数学	92	4	英语	58	天津	300000

学生表

- 数据冗余
  - 课程名称，居住地，邮编等多次出现在表中
- 插入异常
  - 如果新来了一个学生，还没有参加考试，则将产生多列空值。空值在检索操作存储等方面都有不好的影响。
  - 如果没有新的学生，则无法添加新的居住地。

问题

学号	姓名	性别	课程号	课程名	成绩	课程号	课程名	成绩	居住地	邮编
2012001	张三	男	5	数学	88	4	英语	78	北京	100000
2012002	李四	女	5	数学	69	4	英语	83	上海	200000
2012003	王五	男	5	数学	52	4	英语	79	北京	100000
2012004	马六	女	5	数学	58	4	英语	81	上海	200000
2012005	田七	男	5	数学	92	4	英语	58	天津	300000

学生表

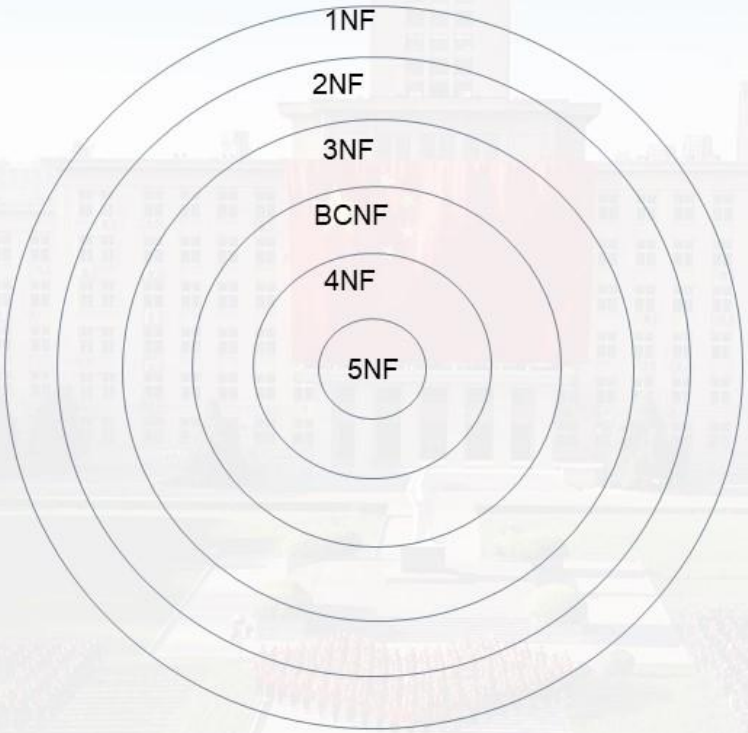
- 删除异常
  - 如果想删除一个居住地，则需要将张三同学的记录全部删掉。这显然是不合理的
- 更新异常
  - 如果想更新北京的邮编，则需要对所有记录都进行操作

关系模式规范化

南開大學 Nankai University

- 规范化是通过最小化数据冗余来提升数据库设计质量的过程，规范化是基于函数依赖以及一系列范式定义的。
- 范式（Normal Form）是符合某一级别的关系模式的集合。目前主要有六种范式，即**第一范式**（1NF）、**第二范式**（2NF）、**第三范式**（3NF）、**BC范式**（BCNF）、**第四范式**（4NF）和**第五范式**（5NF）。
- 其中，比较常用的为1NF, 2NF, 3NF.

范式间的关系



第一范式(1NF)

南開大學 Nankai University

- 确保数据表中每列（字段）的原子性。
- 如果数据表中每个字段都是不可再分的最小数据单元，则满足1NF。
- 1NF是关系模式的基本要求，不满足1NF的数据库模式不能成为关系数据库。
- 1NF的属性不可分包括：
  - 某个属性不能有多个值
  - 不能有重复的属性

投票 最多可选3项

南开大学 Nankai University

下图的学生表是否满足1NF?

学号	姓名	性别	课程号	课程名	成绩	课程号	课程名	成绩	居住地	邮编
2012001	张三	男	5	数学	88	4	英语	78	北京	100000
2012002	李四	女	5	数学	69	4	英语	83	上海	200000
2012003	王五	男	5	数学	52	4	英语	79	北京	100000
2012004	马六	女	5	数学	58	4	英语	81	上海	200000
2012005	田七	男	5	数学	92	4	英语	58	天津	300000

学生表

A 满足

B 不满足

C 忘记了1NF是啥了

第一范式(1NF)

学号	姓名	性别	课程号	课程名	成绩	课程号	课程名	成绩	居住地	邮编
2012001	张三	男	5	数学	88	4	英语	78	北京	100000
2012002	李四	女	5	数学	69	4	英语	83	上海	200000
2012003	王五	男	5	数学	52	4	英语	79	北京	100000
2012004	马六	女	5	数学	58	4	英语	81	上海	200000
2012005	田七	男	5	数学	92	4	英语	58	天津	300000

学生表

学号	姓名	性别	课程号	课程名	成绩	居住地	邮编
2012001	张三	男	5	数学	88	北京	100000
2012002	李四	女	5	数学	69	上海	200000
2012003	王五	男	5	数学	52	北京	100000
2012004	马六	女	5	数学	58	上海	200000
2012005	田七	男	5	数学	92	天津	300000
2012001	张三	男	4	英语	78	北京	100000
2012002	李四	女	4	英语	83	上海	200000
2012003	王五	男	4	英语	79	北京	100000
2012004	马六	女	4	英语	81	上海	200000
2012005	田七	男	4	英语	58	天津	300000

学生表

课程号，课程名重复出现

异常问题缓解

- 数据冗余(存在)
  - 每个学生姓名学号出现两次；课程名重复5次。
- 更新异常(存在)
  - 如果张三名字写错，要改为张叁。需要修改两条记录。
- 插入异常(存在)
  - 新增一个课程，但是没有人选择，则无法插入。
- 删除异常(存在)
  - 删除课程，将会顺带把选课的学生一同删掉。
  - 删除成绩，将会把学生选择相关课程的记录删除掉。

第二范式(2NF)

南開大學 Nankai University

- 在1NF的基础上更进一步，目标是确保表中的每列都和主键相关（依赖性）。如果一个关系满足1NF，并且除了主键之外的其他列，都依赖于该主键，则满足2NF。
- 简而言之，1张表只描述一件事情（非主键列是否完全依赖于主键，还是依赖于主键的一部分）
- 例如：订单表只描述订单相关的信息，所以所有字段都必须与订单id相关。产品表只描述产品相关的信息，所以所有字段都必须与产品id相关。

第二范式(2NF)

学号	姓名	性别	课程号	课程名	成绩	居住地	邮编
2012001	张三	男	5	数学	88	北京	100000
2012002	李四	女	5	数学	69	上海	200000
2012003	王五	男	5	数学	52	北京	100000
2012004	马六	女	5	数学	58	上海	200000
2012005	田七	男	5	数学	92	天津	300000

- 将学生(学号，姓名，性别，课程号，课程名，成绩，居住地，邮编)中存在的非主属性对键的部分函数依赖消除后，可以降低和消除异常问题。
- 成绩完全依赖于(学号，课程号)
- 课程名完全依赖于(课程号)
- (姓名，性别，居住地，邮编)完全依赖于(学号)

修改表结构

南開大學 Nankai University

学号	姓名	性别	居住地	邮编
2012001	张三	男	北京	100000
2012002	李四	女	上海	200000
2012003	王五	男	北京	100000
2012004	马六	女	上海	200000
2012005	田七	男	天津	300000

学生表

课程号	课程名
5	数学
4	英语

课程表

学号	课程号	成绩
2012001	5	88
2012002	5	69
2012003	5	52
2012004	5	58
2012005	5	92
2012001	4	78
2012002	4	83
2012003	4	79
2012004	4	81
2012005	4	58

成绩表

异常问题缓解

南開大學 Nankai University

- 数据冗余(减轻但仍存在)
  - 课程名等信息不再重复出现
  - 居住地名称，邮编等信息冗余
- 更新异常(减轻但仍存在)
  - 如果某地区邮编录入有误，则需要更新多条记录
- 插入异常(减轻但仍存在)
  - 没有新增学生则无法插入新的居住地。
- 删除异常(减轻但仍存在)
  - 删除居住地"北京"，则相应的学生也会被删掉

学号	姓名	性别	居住地	邮编
2012001	张三	男	北京	100000
2012002	李四	女	上海	200000
2012003	王五	男	北京	100000
2012004	马六	女	上海	200000
2012005	田七	男	天津	300000

学生表

课程号	课程名
5	数学
4	英语

课程表

学号	课程号	成绩
2012001	5	88
2012002	5	69
2012003	5	52
2012004	5	58
2012005	5	92
2012001	4	78
2012002	4	83
2012003	4	79
2012004	4	81
2012005	4	58

成绩表

第三范式(3NF)

南開大學 Nankai University

- 满足第三范式（3NF）必须先满足第二范式（2NF）。第三范式（3NF）要求一个数据库表中不包含已在其它表中已包含的非主关键字信息。
- 用外键做表的关联（非主键列是直接依赖于主键，还是直接依赖于非主键列）
- 例如：订单表中需要有客户相关信息，在分离出客户表之后，订单表中只需要有一个用户id即可（外键），而不能有其他的客户信息。因为其他的客户信息直接关联于用户id，而不是直接与订单id直接相关。

第3范式

学号	姓名	性别	居住地
2012001	张三	男	北京
2012002	李四	女	上海
2012003	王五	男	北京
2012004	马六	女	上海
2012005	田七	男	天津

学生表

课程号	课程名
5	数学
4	英语

课程表

学号	课程号	成绩
2012001	5	88
2012002	5	69
2012003	5	52
2012004	5	58
2012005	5	92
2012001	4	78
2012002	4	83
2012003	4	79
2012004	4	81
2012005	4	58

成绩表

居住地	邮编
北京	100000
上海	200000
天津	300000

居住地表

异常问题缓解

- 数据冗余 降低
  - 每个信息都只存储一次
- 更新异常不再
  - 修改任何属性都仅于其记录相关。
- 插入异常不再
  - 新增居住地信息不再需要添加新的学生
- 删除异常不再
  - 删除北京的所有学生并不会将北京的邮编信息一起删除

继续优化

- 我们可以将作为外键的字段所占的空间减少，即减少重复项的空间占用。
- 例如在本例子中，我们可以使用地址号(address_id)来唯一标识居住地。这样，在学生表中，存储的居住地信息，可以使用address_id来代替居住地名称，从而再次减少空间的占用。

ID	城市	邮编
1	北京	100000
2	天津	200000

投票 最多可选1项

南开大学 Nankai University

请问以后在设计数据库时，你会怎么做？

A 严格依据三范式理论

B 结合三范式理论和其他范式理论

C 视情况而定

允公允能 日新月异 NANKAI UNIVERSITY

按照范式来设计能万无一失?

- 范式化设计目标的主要目的就是"减少不必要的更新"。但事事都具有两面性，在对数据库进行范式设计的时候也不可避免的带来了一些副作用。
- 一个完全范式化设计的数据库经常会面临"查询缓慢"这个问题。数据库越范式化，就需要Join越多的表
- 例如在本例中，如果想要知道张三居住地的邮编，就必须Join学生表和地址表。想要知道张三的数学成绩，需要join学生表，课程表和成绩表。

投票 最多可选1项

南开大学 Nankai University

A

学号	姓名	国籍ID

自增ID	国家名称

B

学号	姓名	国家名称

A A更好

B B更好

C 视情况而定

允公允能 日新月异 NANKAI UNIVERSITY

逆范式

南開大學 Nankai University

- 逆范式就是打破范式的规定，通过增加冗余或者重复的数据来提高数据库的性能。
- 有的时候，我们会针对**不可改变**的数据，进行逆范式设计。例如：用户信息中的国籍，由于国家信息(eg: 代码，简称)等改变的几率很小，所以无需在用户信息表中新增列来关联国家信息表，而是直接将相关的国家信息作为用户的信息表中的列来存储。从而优化查询效率。
- 但是，逆范式设计将会降低更新的速度，所以对于数据库不需要经常更新，而频繁检索的场景更加合适。

冗余设计

- 在设计数据库时，某一字段属于一个表，但它又同时出现在另一个或多个表，且表示的意义完全相同，那么这个字段就是一个冗余字段。
- 关系数据库中的数据冗余主要是指关系数据库中同一信息数据的重复存贮。数据冗余浪费了宝贵的资源，应尽量减少。但关系数据库中为实现一些功能有些数据冗余是必需的。

冗余设计

- 在设计数据库时，某一字段属于一个表，但它又同时出现在另一个或多个表，且表示的意义完全相同，那么这个字段就是一个冗余字段。
- 关系数据库中的数据冗余主要是指关系数据库中同一信息数据的重复存贮。数据冗余浪费了宝贵的资源，应尽量减少。但关系数据库中为实现一些功能有些数据冗余是必需的。

冗余设计

南開大學 Nankai University

- 必需的数据冗余主要用于以下用途：
- (1)数据间建立联系，如两表间通过共同属性建立联系；
- (2)数据恢复，如建立备份文件以备正式文件被破坏时恢复；
- (3)数据核查，如设立数据校验位可以检查数据在存贮、传输等过程中的改变；
- (4)数据使用的便利，如为了查看数据的直观，使用数据的方便、高效。
- (5)减少数据通讯开销，如分布式数据库在不同场地重复。

投票 最多可选1项

南开大学 Nankai University

A

ID	用户名	个人信息	...

ID	评论内容	用户表ID	...

B

ID	用户名	个人信息	...

ID	评论内容	用户表ID	用户名	用户等级	...

A A更好

B B更好

C 视情况而定

允公允能 日新月异 NANKAI UNIVERSITY

冗余设计

南開大學 Nankai University

- 冗余设计：提高查询效率，减少关联查询。单表查询比关联查询速度要快；
- 劣势：浪费存储空间；需要多处维护冗余字段的更新；违背了数据库设计范式理论。

- 范式理论：要求数据库设计逻辑清晰、关系明确；连接查询；
- 劣势：如果一个表存在几十万条记录，也只能进行关联检索。

设计原则

南開大學 Nankai University

- 目前要进行一个关系型数据库设计，我们有两种选择：

- 1、尽量遵循范式理论的规约，尽可能少的冗余字段，让数据库设计看起来精致、优雅、让人心醉。
- 2、合理的加入冗余字段这个润滑剂，减少join，让数据库执行性能更高更快。