数据库系统原理

教程：数据库系统概论（第5版）

结合：CMU 15-445/645 INTRO TO DATABASE SYSTEMS

华中科技大学 计算机学院

左琼

第六章　关系数据理论

Principles of Database Systems

第六章 关系数据理论

6.1 问题的提出

6.2 规范化

6.3 数据依赖的公理系统

*6.4 模式的分解

6.5 小结

- 一个好的关系模式应该具备以下4个条件：
  - 尽可能少的数据冗余
  - 没有更新异常
  - 没有插入异常
  - 没有删除异常

6.2.3 范式

- 规范化的基本思路——用于改造关系模式，通过分解关系模式来消除其中不合适的数据依赖，以解决插入异常、删除异常、更新异常和数据冗余问题。
- 范式（Normal Form）：关系数据库的规范化过程中为不同程度的规范化要求设立的不同标准。
- 满足最基本规范化要求的关系模式叫第一范式，
  在第一范式中进一步满足一些要求为第二范式，
  ……
- 通过模式分解将一个低级范式转换为若干个高级范式的过程称作规范化。

$$1NF \supset 2NF \supset 3NF \supset$$

$$BCNF \supset 4NF \supset 5NF$$

第一范式 1NF

- 如果关系模式R，其所有的属性均为简单属性，即每个属性都是不可再分的，则称R属于第一范式，简称1NF，记作R∈1NF。
- 第一范式是最基本的规范形式，满足这个条件的关系称为规范化关系。
- 例：以下关系是一个非规范化的关系。

XH	KM	CJ		
		test1	test2	test3
001	OS	80	78	90
002		85	82	95

S#	C#
S1	{C1，C2，C3}
S2	{C2, C4}

关系DBS中，所有关系模式都必须是1NF。

1NF

- 将非1NF转换为1NF的方法：
  - ① 去掉嵌套属性上层
  - ② 重写行交叉处的值

Sno	Cno	test1	test2	test3
001	OS	80	78	90
002	OS	85	82	95

Sno	Cno
S1	C1
S1	C2
S1	C3
S2	C2
S2	C4

- 例：SCD(学号, 姓名, 年龄, 所在系, 系主任, 课程号, 成绩)，

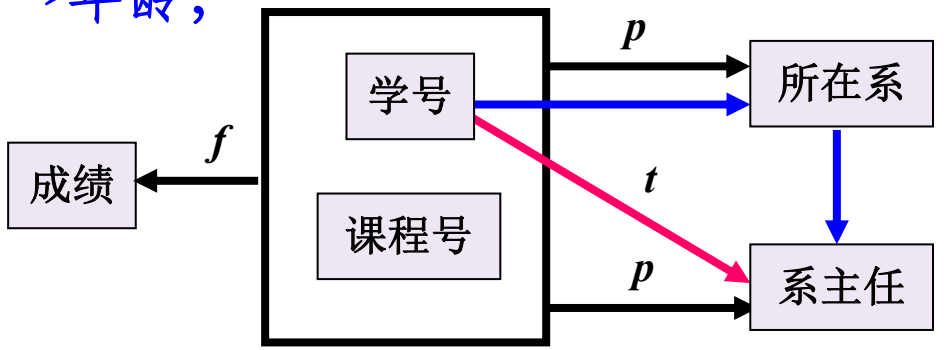
  SCD $\in$ 1NF

- 1NF存在的问题：数据冗余，具有插入异常、删除异常、更新复杂。
- 问题存在原因：——非主属性部分函数依赖于候选码。

1NF

- 例：SCD(学号, 姓名, 年龄, 所在系, 系主任, 课程号, 成绩)

(学号，课程号)→成绩，学号→姓名， 学号→年龄，
学号→所在系，
(学号，课程号)$\xrightarrow{p}$年龄
(学号，课程号)$\xrightarrow{p}$所在系，
(学号，课程号)$\xrightarrow{p}$姓名，
学号$\xrightarrow{t}$系主任，(学号，课程号)$\xrightarrow{p}$系主任



- 解决方法：规范化(投影分解)

消除非主属性对码的部分FD。

- 所有完全FD于码的属性组成一个关系模式。
- 所有部分FD于码的属性组成一个关系模式。

6.2.4 2NF

- 如果关系模式R∈1NF，且每个非主属性都完全函数依赖于R的码，则称R属于第二范式（简称2NF），记作R∈2NF。

例：

解决了哪些问题？

SD

学号 Sno	姓名 Sname	年龄 Age	系别 Dept	系主任 Mname
S1	李勇	18	计算机	王平
S2	刘晨	17	计算机	王平
S3	王敏	18	自控系	刘伟

还有没有问题？

SC

学号 Sno	课程号 Cno	成绩 Score
S1	C1	85
S1	C2	82
S2	C4	78
S3	C2	87

SD ∈ 2NF

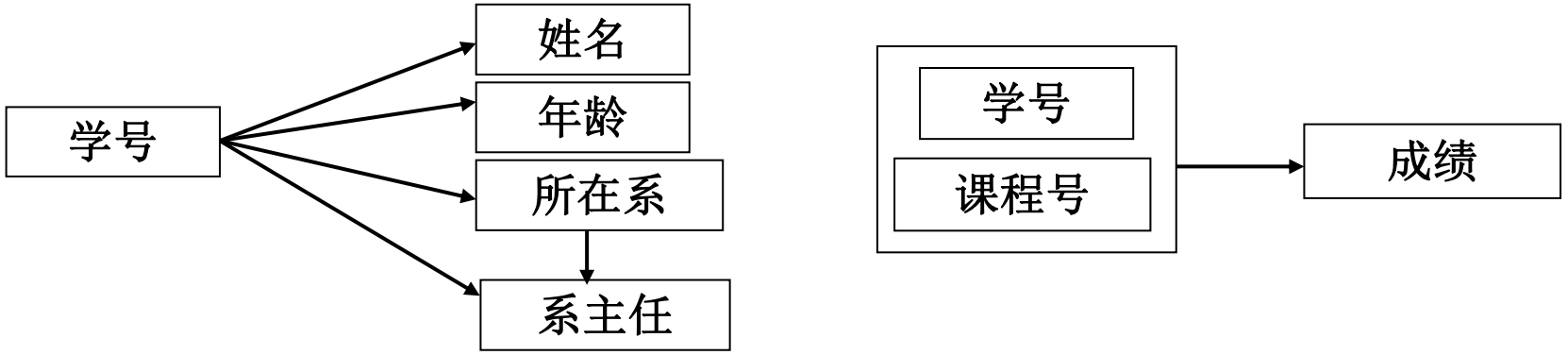
SC ∈ 2NF

即：不能存在仅依赖主关键字一部分的属性，如果存在，那么这个属性和主关键字的这一部分应该分离出来形成一个新的实体，新实体与原实体之间是一对多的关系。**简而言之，第二范式就是属性完全依赖于主键。**

- 推论：若R∈1NF,且其候选码为单个属性,则R∈2NF。 (Why?)

2NF

- SCD分解后，SD和SC的函数依赖分别如下图所示：



1NF→2NF,解决的问题：

- 消除了一些数据冗余，如：姓名、年龄不需要重复存储多次。
- 插入异常部分解决，如：学生的基本信息与选课信息分开插入。
- 删除异常部分解决，如：删除某学生选课信息，而学生其它信息不受影响。

仍然存在的问题：

- 冗余仍然存在，如：系主任。
- 更新复杂，如：更换系主任。
- 插入异常，如：没有学生，系主任无法插入。
- 删除异常，如：删除某系全体学生，删除系主任信息。

症由：非主属性对码的传递依赖:*sno→dept, dept→mname*

解决方法：投影分解，消去非主属性对码的传递依赖。

6.2.5 3NF

- 如果关系模式R(U, F) ∈1NF，其中不存在码X、属性组Y及非主属性Z（Z⊆Y）使得X→Y（Y↛X），Y↛Z成立，则称R属于第三范式（Third Normal Form），简称3NF，记作R∈3NF。
- 3NF建立在2NF之上，它要求所有的非主键列都必需直接依赖于主键，不包括任何传递依赖。
- 若R∈3NF，则每一非主属性既不部分依赖于码也不传递依赖于码。
- 若R ∈ 3NF , 则必有R ∈ 2NF 。
- 采用投影分解法将一个2NF的关系分解为多个3NF的关系，可以在一定程度上解决原2NF关系中存在的插入异常、删除异常、数据冗余度大、修改复杂等问题。

3NF

- 以2NF关系模式SD为例，说明3NF规范化的过程。

  SD(学号, 姓名, 年龄, 所在系, 系主任)

  分析SD的属性，可以将SD分解成如下两个关系：

  - S(学号,姓名,年龄,所在系)，描述学生实体；
  - D(系名, 系主任)，描述系的实体。

  分解后的两个关系S和D，码分别为学号和系名，不存在非主属性对码的传递函数依赖。因此，S∈3NF，D∈3NF。

  - 分解为3NF后：
    - 数据冗余降低。
    - 不存在插入异常、不存在删除异常、不存在更新异常。

3NF

- 推论1: 若R∈2NF，且至多存在一个非主属性，则R∈3NF。
- 推论2: 任何二元关系模式R(A,B)必为3NF。
- 说明:
  - 部分*FD* 和传递*FD* 是冗余及操作异常的重要根源。
  - 3NF不存在 非主属性 对 候选码 的部分*FD* 和传递*FD*。
  - 3NF消去了大部分冗余及操作异常。
  - 但并非所有的3NF都能完全消除冗余及操作异常。

案例：规范化的过程

- 某书店购书情况汇总登记表：

订单号 NO	订户代号 C#	姓名 CN	地址 CA	书号 B#	书名 BN	出版单位 PU	单价 UP	定购数量 QUA
				02164	计算机语言	教育出版社	2.0	50
00001	0253	清华大学 自动化系	北京	01003	BASIC语言	科学出版社	1.1	50
				06372	FORTRAN语言	清华出版社	1.5	30
				01083	电子线路	邮电出版社	1.4	1
0002	0372	陈刚	上海	02954	自动控制原理	国防出版社	0.95	1
				02954	自动控制原理	国防出版社	0.95	20
0003	2234	天津大学 计算机系	天津	02164	计算机语言	教育出版社	2.0	50
				01083	电子线路	邮电出版社	1.4	50
				01003	BASIC语言	科学出版社	1.1	20
0004	2523	北京钢铁学院 图书馆	北京	01003	BASIC语言	科学出版社	1.1	30
				06372	FORTRAN语言	清华出版社	1.5	30

案例：规范化的过程(2)

- 重写行交叉处的值后，关系模式满足1NF的要求。

订单号 NO	订户代号 C#	姓名 CN	地址 CA	书号 B#	书名 BN	出版单位 EU	单价 UP	定购数量 QUA
00001	0253	清华大学自动化系	北京	02164	计算机语言	教育出版社	2.0	50
00001	0253	清华大学自动化系	北京	01003	BASIC语言	科学出版社	1.1	50
00001	0253	清华大学自动化系	北京	06372	FORTRAN语言	清华出版社	1.5	30
0002	0372	陈刚	上海	01083	电子线路	邮电出版社	1.4	1
0002	0372	陈刚	上海	02954	自动控制原理	国防出版社	0.95	1
0003	2234	天津大学计算机系	天津	02954	自动控制原理	国防出版社	0.95	20
0003	2234	天津大学计算机系	天津	02164	计算机语言	教育出版社	2.0	50
0003	2234	天津大学计算机系	天津	01083	电子线路	邮电出版社	1.4	50
0003	2234	天津大学计算机系	天津	01003	BASIC语言	科学出版社	1.1	20
0004	2523	北京钢铁学院图书馆	北京	01003	BASIC语言	科学出版社	1.1	30
0004	2523	北京钢铁学院图书馆	北京	06372	FORTRAN语言	清华出版社	1.5	30

根据分析可以得到一组函数依赖：

F={ NO→C#,C#→CN,C#→CA, B#→BN,B#→EU,B#→UP, (NO,B#) →QUA },

表中（NO,B#）为主码。

案例：规范化的过程(3)

- 消除部分函数依赖，将其分解成三个关系，使每一个非主属性都完全依赖于主码，满足2NF的要求。

订单号 NO	订户代号 C#	姓名 CN	地址 CA
00001	0253	清华大学自动化系	北京
0002	0372	陈刚	上海
0003	2234	天津大学计算机系	天津
0004	2523	北京钢铁学院图书馆	北京

书号 B#	书名 BN	出版单位 EU	单价 UP
02164	计算机语言	教育出版社	2.0
01003	BASIC语言	科学出版社	1.1
06372	FORTRAN语言	清华出版社	1.5
01083	电子线路	邮电出版社	1.4
02954	自动控制原理	国防出版社	0.95

订单号 NO	书号 B#	定购数量 QUA
00001	02164	50
00001	01003	50
00001	06372	30
0002	01083	1
0002	02954	1
0003	02954	20
0003	02164	50
0003	01083	50
0003	01003	20
0004	01003	30
0004	06372	30

各自函数依赖：

F1={NO→C#,C#→CN,C#→CA}，NO为主码。

F2={B#→BN,B#→EU,B#→UP}，B#为主码。

F3={(NO,B#) →QUA }，（NO,B#）为主码。

案例：规范化的过程(4)

- 进一步消除传递函数依赖，满足3NF的要求。

订户代号 C#	姓名 CN	地址 CA
0253	清华大学自	北京
0372	陈刚	上海
2234	天津大学计	天津
2523	北京钢铁学	北京

书号 B#	书名 BN	出版单位 EU	单价 UP
02164	计算机语言	教育出版社	2.0
01003	BASIC语言	科学出版社	1.1
06372	FORTRAN语言	清华出版社	1.5
01083	电子线路	邮电出版社	1.4
02954	自动控制原理	国防出版社	0.95

订单号 NO	书号 B#	定购数量 QUA
00001	02164	50
00001	01003	50
00001	06372	30
0002	01083	1
0002	02954	1
0003	02954	20
0003	02164	50
0003	01083	50
0003	01003	20
0004	01003	30
0004	06372	30

订单号 NO	订户代号 C#
00001	0253
0002	0372
0003	2234
0004	2523

各自函数依赖：

F11={ NO→C# }，NO为主码。

F12={ C#→CN,C#→CA }，C#为主码。

3NF可能存在的问题

R

学号	课程	教师
S1	OS	于得水
S2	OS	马千里
S4	DB	牛得草
S5	DB	牛得草
S5	OS	于得水
……	……	……

- 语义：
  - 每门课可由多个教师讲；
  - 每位教师只讲一门课；
  - 学生一旦选定一门课，就确定了一个固定教师。

解：

- 候选码：（学号，课程），（学号，教师）
- 非主属性：无
- 函数依赖：(学号,课程)→教师，教师→课程
- 思考：R ∈3NF?

∵不存在非主属性，∴R∈3NF。

3NF可能存在的问题

R

学号	课程	教师
S1	OS	于得水
S2	OS	马千里
S4	DB	牛得草
S5	DB	牛得草
S5	OS	于得水
……	……	……

1）数据冗余

多个学生选同一个教师的同一门课时，课程，教师重复。

2）更新异常

某门课程改名，则修改多个元组。

3）插入异常

候选码为（学号，课程）和（学号，教师），则：

教师开设某门课程，尚无学生选修，因为无学号信息，则课程、教师信息不能进入数据库。

4）删除异常

删去学生选课信息，丢失教师授课信息。

3NF可能存在的问题

- 存在上述问题的原因：

  存在 主属性 对 候选码 的不良依赖。

  ∵ 教师→课程 ∴（学号，教师）→课程

- 解决方法：投影分解

  ——消去主属性对候选码的部分fd。

可将R分解为：

学号	教师
S1	于得水
S2	马千里
……	……

教师	课程
于得水	OS
马千里	OS
牛得草	DB
……	……

分解后存在的问题得到解决，新关系模式满足BC范式。

6.2.6 BC范式

- 定义：如果关系模式R(U,F)∈1NF，且对于所有的非平凡函数依赖X→Y，X必包含了R的一个码，则称R属于BC范式（Boyce-Codd Normal Form），记作R∈BCNF。
- 定理：如果R∈BCNF，则R∈3NF。
- 证明：（反证法）

  假设R∈BCNF，且R∉3NF。

  ∵R∉3NF，

  ∴存在码X、属性组Y及非主属性Z（Z⊄Y）使得X→Y（Y↛X），Y→Z成立；

  ∵Y↛X，∴Y不包含码，即非平凡函数依赖Y→Z的决定部分不包含码。

  根据BC范式的定义，R∉BCNF，这与假设相矛盾。

  ∴R∈3NF。（证毕）

课堂练习

- 关系模式R(Sno, Cno, ORDER)表示学生选修课程的成绩排名（假设每门课中没有并列名次），

  则由语义有函数依赖：

  (Sno,Cno)→ORDER，(Cno,ORDER)→Sno，

  请问：R属于3NF吗？ R属于BCNF吗？

BCNF

- BCNF的性质：
  1）所有非主属性都完全依赖于候选码；
  2）所有非主属性都不传递依赖于候选码；
  3）所有主属性都完全依赖于不包含它的候选码；
  4）所有主属性都不传递依赖于候选码。

主属性不依赖于主属性

- 若一个关系达到了第三范式，并且它只有一个候选码，或它的每个候选码都是单属性，则该关系是BCNF。

范式

- 1 NF ⊇ 2 NF ⊇ 3 NF ⊇ BCNF
- 如果一个关系DB中所有关系模式都属于3NF，则已在很大程度上消除了插入异常和删除异常，但仍然可能存在主属性对候选键的部分依赖和传递依赖。
- 如果一个关系数据库中所有关系模式都属于BCNF，那么在函数依赖的范畴内，已经实现了模式的彻底分解，完全消除了产生插入异常和删除异常的根源，而且数据冗余也减少到极小程度。
- 但是除了函数依赖，还存在另一种数据依赖——多值依赖。

6.2.7 多值依赖*

- 1.问题的提出

  设学校中某一门课程C由多个教师T讲授，他们使用相同的一套参考书B，每个教员可以讲授多门课程,每种参考书可以供多门课程使用。

- 考察关系模式：

  Teaching(C, T, B)

  课程C

  教师T

  参考书B

课程C	教员T	参考书B
物理	李勇 王军	普通物理学 光学原理
数学	李勇 张平	数学分析 高等代数
计算数学	张平 周峰	数学分析

多值依赖

- Teaching(C, T, B) 规范化的二维表格：

课程C	教员T	参考书B
物理	李勇	普通物理学
物理	李勇	光学原理
物理	王军	普通物理学
物理	王军	光学原理
数学	李勇	数学分析
数学	李勇	高等代数
数学	张平	数学分析
数学	张平	高等代数
计算数学	张平	数学分析
计算数学	周峰	数学分析

- 分析：
  - 具有唯一候选码(C,T,B)，即全码
  - Teaching∈BCNF
  - Teaching模式是否存在不良特性？
    - 数据冗余
    - 插入异常
    - 删除异常
    - 更新异常
- 症由：多值依赖

多值依赖

2. 定义:

定义6.15(描述型)　设R(U)是一个属性集U上的一个关系模式，X、Y和Z是U的子集，并且Z = U - X - Y，多值依赖 X→→Y成立，当且仅当：

1）对R的任一关系r，r在（X，Z）上的每个值对应一组Y的值；

2）这组值仅仅决定于X值而与Z值无关。

例: Teaching（C, T, B）有

C →→T 和 C →→B

对于C的每一个值，T有一组值与之对应，而不论B取何值

多值依赖

- 不当MVD的关系式引起的弊端及对策:
  - Teaching(C,T,B) 分解为： T(C,T) 和 B(C,B)

课程C	教员T
物理	李勇
物理	王军
数学	李勇
数学	张平
计算数学	张平
计算数学	周峰

课程C	参考书B
物理	普通物理学
物理	光学原理
数学	数学分析
数学	高等代数
计算数学	数学分析

多值依赖 vs. 函数依赖

- 函数依赖是多值依赖的特例，即：若X→Y，则X→→Y。

  此时,s、t在X上的投影相等，则在Y上的投影也必然相等，该投影与s和t在Z = U - X - Y上的投影无关。

- 多值依赖与函数依赖的区别：
  - 函数依赖规定某些元组不能出现在关系中，

    多值依赖要求某种形式的其它元组必须在关系中。
  - 有效性与属性集的范围不同：
    - X → Y的有效性仅决定于X、Y属性集上的值；
    - X →→ Y不仅涉及属性组X和Y，而且涉及U中其余属性Z。
    - 若X→Y在R(U)上成立，则对于任何Y'⊆ Y，均有X→Y'成立；
    - 多值依赖X→→Y若在R(U)上成立，不能断言对于任何Y' ⊂ Y有X→→Y' 成立。

6.2.8 4NF

定义6.10 关系模式R<U，F>∈1NF，如果对于R的每个非平凡多值依赖 $X \rightarrow\rightarrow Y$ $(Y \not\subseteq X)$ ，X都含有候选码，则R∈4NF。

含义：不允许同一表内的多对多关系。

如果R ∈ 4NF， 则R ∈ BCNF 。

[例] 关系模式: Teaching(C,T,B) ∉ 4NF，码为(C, T, B)

存在MVD：C→→T,且C不是候选码

分解成两个关系模式：

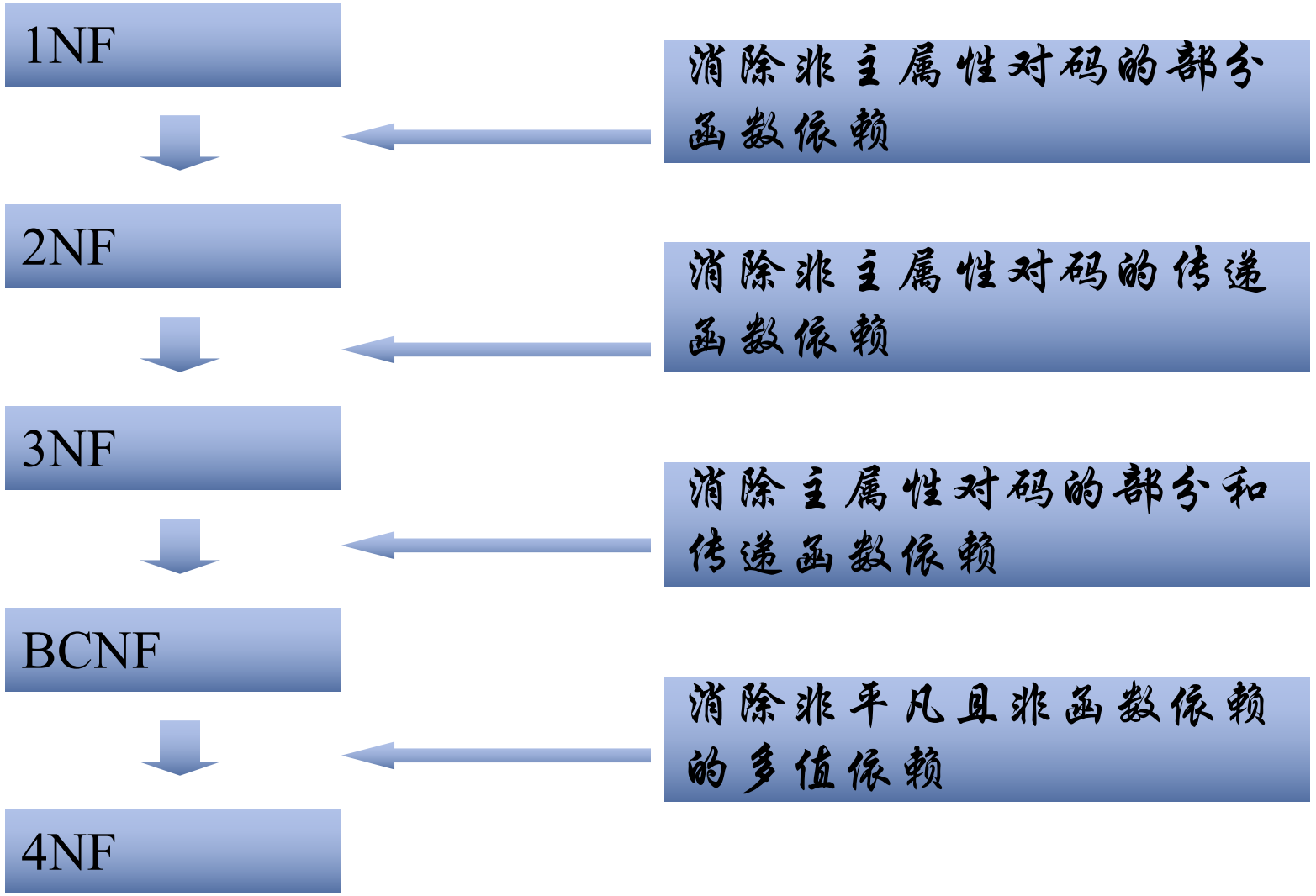
T(C, T) ∈ 4NF (码:CT)

B(C, B) ∈ 4NF (码:CB)

C→→T， C→→B是平凡多值依赖（Z=ϕ）

- 不允许有非平凡且非函数依赖的多值依赖。
- 允许的非平凡多值依赖实际上是函数依赖。

6.2.9 关系规范化小结


1NF
2NF
3NF
BCNF
4NF
消除非主属性对码的部分函数依赖
消除非主属性对码的传递函数依赖
消除主属性对码的部分和传递函数依赖
消除非平凡且非函数依赖的多值依赖

结论

1）3NF必定为2NF和1NF，反之不一定；

2）4NF必为BCNF，BCNF必为3NF，反之不一定；

3）3NF已在很大程度上控制了数据冗余、很大程度上消去了插入和删除操作异常；但，3NF分解仍不够彻底（可能存在主属性对候选码的部分fd和传递fd）；

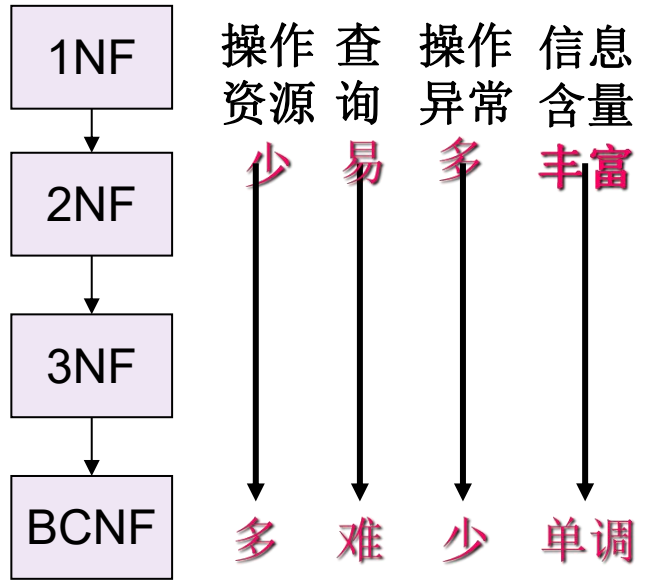
5）在fd范围内，BCNF下已完全消去了插入删除异常；

6）4NF是多值依赖范畴内最高程度的规范化；

7）范式并非越高越好。理论上数据库设计一般应规范到3NF，但实际应用中为减少连接运算，提高查询效率，不一定都达到3NF。

8）适可而止（垃圾，连接运算）；

9）分解不唯一。



关系模式的规范化

反规范化——在数据库中，为了提高查询效率而有意降低关系规范化程度的一种技术。

例如：一般选课数据库的关系模式如下：

class(clno, dept, major)

student(sno, ID-card, sname, sex, birthdate, clno)

course(cno, cname, score, pcno)

sc(sno,cno,grade)

但对于查分系统，为提高查询效率，可设计如下关系模式：

students (sno, ID-card, sname, cno, grade)

关系模式的规范化

思考：

- 任何一个二目关系模式R（A，B）一定属于BCNF吗？一定属于4NF吗？
- 一个全是主属性的关系模式一定可以达到第几范式？
- 一个全码的关系模式一定可以达到第几范式？