设计目标

- 关系数据库设计的目标是：
  - BCNF
  - 无损连接
  - 依赖保持
- 如果不能达到这些，也可接受
  - 缺少保持依赖
  - 因3NF引起的冗余
- 除了超码之外，SQL并没有提供直接声明函数依赖的方法。可以通过断言声明FD，但检测代价太大
- 因此即使我们有一个保持依赖的分解，使用SQL，我们也不能有效地检测左部不是码的函数依赖

多值依赖

- 有时属于BCNF的模式仍然未充分规范化
- 考虑数据库

  $classes(course,\ teacher,\ book)$

  定义$(c, t, b) \in classes$，意思是教师t 可以教课程c，而b 是需用于课程c 的教材
- 数据库将为每门课程列出能讲授该课程的教师的集合，以及需用的书的集合(不管谁讲授该课)
  - *course*: *teacher* =1:n，*course*: *book* = 1:n，*teacher*和*book*是多值属性，并且*teacher*和*book*相互独立

多值依赖

course	*teacher*	*book*
database	Avi	DB Concepts
database	Avi	DB system (Ullman)
database	Hank	DB Concepts
database	Hank	DB system (Ullman)
database	Sudarshan	DB Concepts
database	Sudarshan	DB system (Ullman)
operating systems	Avi	OS Concepts
operating systems	Avi	OS system (Shaw)
operating systems	Jim	OS Concepts
operating systems	Jim	OS system (Shaw)

classes

- 由于没有非平凡依赖，(*course*, *teacher*, *book*)是唯一的键，因此该关系模式属于BCNF

多值依赖

- 冗余与插入异常： 如果Sara是能教数据库的新教师，必须插入两条元组

  (database, Sara, DB Concepts)

  (database, Sara, Ullman)

多值依赖

- 因此，最好将*classes* 分解成：

course	*teacher*
database	Avi
database	Hank
database	Sudarshan
operating systems	Avi
operating systems	Jim

teaches

key={course, teacher}

course	*book*
database	DB Concepts
database	DB system (Ullman)
operating systems	OS Concepts
operating systems	OS system (Shaw)

text

我们将看到这两个关系都属于第四范式(4NF)

多值依赖

- 设有关系模式R，令$\alpha \subseteq R$ 及$\beta \subseteq R$。多值依赖

$$\alpha \rightarrow\rightarrow \beta$$

在R 上成立当且仅当在任意合法关系$r(R)$中，对所有满足$t_1[\alpha] = t_2[\alpha]$的元组对$t_1$ 和t_2，必存在元组t_3 和t_4 使得：

$$t_1[\alpha] = t_2[\alpha] = t_3[\alpha] = t_4[\alpha]$$
$$t_3[\beta] = t_1[\beta]$$
$$t_4[\beta] = t_2[\beta]$$
$$t_3[R - \alpha - \beta] = t_2[R - \alpha - \beta]$$
$$t_4[R - \alpha - \beta] = t_1[R - \alpha - \beta]$$

令$R - \alpha - \beta = z$

$t_3[z] = t_2[z]$

$t_4[z] = t_1[z]$

多值依赖

- 用表来表示 $\alpha \rightarrow\rightarrow \beta$

	α	β	$R-\alpha-\beta$
t_1	$a_1 \ldots a_i$	$a_{i+1} \ldots a_j$	$a_{j+1} \ldots a_n$
t_2	$a_1 \ldots a_i$	$b_{i+1} \ldots b_j$	$b_{j+1} \ldots b_n$
t_3	$a_1 \ldots a_i$	$a_{i+1} \ldots a_j$	$b_{j+1} \ldots b_n$
t_4	$a_1 \ldots a_i$	$b_{i+1} \ldots b_j$	$a_{j+1} \ldots a_n$
	t1[ ]=t2[ ] =t3[ ]=t4[ ]	t1[ ]=t3[ ] t2[ ]=t4[ ]	t1[ ]=t4[ ] t2[ ]=t3[ ]

- 如果$\beta \subseteq \alpha$，或$\alpha \cup \beta = R$，则$\alpha \rightarrow\rightarrow \beta$ 是平凡的

多值依赖

- 另一种定义

	α	β	$R-\alpha-\beta$
t_1	$a_1 \dots a_i$	$a_{i+1} \dots a_j$	$a_{j+1} \dots a_n$
t_2	$a_1 \dots a_i$	$b_{i+1} \dots b_j$	$b_{j+1} \dots b_n$
t_3	$a_1 \dots a_i$	$a_{i+1} \dots a_j$	$b_{j+1} \dots b_n$
t_4	$a_1 \dots a_i$	$b_{i+1} \dots b_j$	$a_{j+1} \dots a_n$
	t1[ ]=t2[ ] =t3[ ]	t1[ ]=t3[ ]	t2[ ]=t3[ ]

对任意三个元组都成立

多值依赖

- 例，设关系模式R 的属性集合被分成三个非空子集

  $Y,\ Z,\ W$

  称$Y \rightarrow\rightarrow Z$（$Y$ 多值决定Z）当且仅当对所有可能的关系$r(R)$，若

  $\langle y_1,\ z_1,\ w_1 \rangle \in r$ 及 $\langle y_2,\ z_2,\ w_2 \rangle \in r$

  则

  $\langle y_1,\ z_1,\ w_2 \rangle \in r$ 及 $\langle y_1,\ z_2,\ w_1 \rangle \in r$

- 注意由于Z 和W 的行为完全对称，若$Y \rightarrow\rightarrow W$ 则$Y \rightarrow\rightarrow Z$

多值依赖

- 在前面的例子中：

  $course \rightarrow\rightarrow teacher$

  $course \rightarrow\rightarrow book$

- 上述形式定义表达了这种概念：给定Y（$course$）的特定值，则有一个Z（$teacher$）值的集合和一个W（$book$）值的集合与之相关联，而这两个集合在某种意义上是相互独立的

多值依赖理论

- 根据多值依赖的定义，可导出下列规则：
  - 若 $\alpha \rightarrow \beta$，则$\alpha \rightarrow\rightarrow \beta$，即函数依赖是多值依赖的特例
- D 的闭包D^+ 是D 逻辑蕴含的所有函数依赖和多值依赖的集合
  - 可根据函数依赖与多值依赖的形式定义来从D 计算$D+$
  - 我们只对在实践中较常见的简单多值依赖可用这样的推理
  - 对于复杂的依赖，最好利用一套推理规则来对依赖集合进行推理

4NF

- 关系模式R 关于函数依赖及多值依赖集合D 属于4NF当且仅当对D^+ 中所有形如$\alpha \rightarrow\rightarrow \beta$ 的多值依赖，其中$\alpha \subseteq R$ 且$\beta \subseteq R$，下列条件中至少一个成立：
  - $\alpha \rightarrow\rightarrow \beta$ 是平凡的（即，$\beta \subseteq \alpha$ 或$\alpha \cup \beta = R$）
  - α 是模式R 的超码
- 若关系属于4NF，则它必属于BCNF

4NF分解算法

result: = {*R*};
done := false;
compute D^+;
令D_i 表示D^+ 在R_i 上的限制
while (not *done*)
if (在*result* 中有不属于4NF的模式R_i) then
begin
令$\alpha \rightarrow\rightarrow \beta$ 是在R_i 上成立的一个非平凡多值依赖，它使得$\alpha \rightarrow R_i$ 不属于D_i，并且$\alpha \cap \beta = \phi$;
result := (*result* − R_i) ∪ (R_i − β) ∪ (α, β);
end
else *done*:= true;

R_{i1}　　R_{i2}

注：每个R_i 属于4NF，且分解是无损连接的

4NF

- 例，$R=(A,\ B,\ C,\ G,\ H,\ I)$

  $F=\{A \rightarrow\rightarrow B$

  $B \rightarrow\rightarrow HI$

  $CG \rightarrow\rightarrow H\}$

- R 不属于4NF，因为$A \rightarrow\rightarrow B$，$A$ 不是R 的超码
- 分解：
  - a）$R_1=(A,\ B)$　　（R_1 属于4NF）
  - b）$R_2=(A,\ C,\ G,\ H,\ I)$　　（R_2 不属于4NF）
  - c）$R_{21}=(C,\ G,\ H)$　　（R_{21} 属于4NF）
  - d）$R_{22}=(A,\ C,\ G,\ I)$　　（R_{22} 属于4NF）

更多的范式

- 连接依赖概化了多值依赖
  - 并引出了另一种范式投影-连接范式（PJNF）
- 还有一类更一般化的约束，它引出一种称作域-码范式（domain-key normal form）
- 使用这些一般化的约束的一个实际问题是，它们不仅难以推导，而且也还没有形成一套具有正确有效性和完备性的推理规则用于约束的推导
- 因此，很少使用

数据库设计过程

- 假设给定了模式R
  - R 可能是经转换E-R图到表而生成的
  - R 可能是单个包含所有属性的关系（称为全关系）
  - 规范化将R 分解成较小的关系
  - R 可能是某种特定设计的结果，然后再测试/转换成范式

E-R模型和规范化

- 当我们小心地定义当E-R图，并正确地标识所有实体，则从E-R图生成的关系模式就不需要太多进一步的规范化
- 但是，在现实(不完善的)设计中可能存在从实体的非码属性到其他属性的FD
  - 例如：*employee* 实体具有属性 *department_name* 和*building*，以及FD *department_name* $\rightarrow$ *building*
  - 好的设计应该将*department* 作为实体
- 从联系集的非码属性引出FD是可能的，但极少——多数联系是二元的

为了性能去规范化

- 当我们小心地定义当E-R图，并正确地标识所有实体，则从E-R图生成的关系模式就不需要太多进一步的规范化
- 为提高性能可能使用非规范化的模式
- 例如：假定每次访问一门课程时，所有的先修课都必须和课程信息一起显示，需将*course*和*preq*连接
- 做法1：使用包含*course*以及*preq*的所有上述属性的反规范化关系
  - 查找迅速
  - 额外空间以及额外更新执行时间
  - 程序员的额外编码工作以及更多出错可能性

为了性能去规范化

- 做法2：使用如下定义的物化视图

$$course \bowtie preq$$

  - 优缺点同上，除了程序员的额外编码工作和避免可能的错误

其他设计问题

- 有些数据库设计问题不能被规范化解决
- 应避免的坏的数据库设计：
  - 例1，不用*earnings*(*company-id*, *year*, *amount*)，而是用具有同样模式(*company-id*, *earnings*) 的*earnings-2000*, *earnings-2001*, *earnings-2002*，等等
  - 这属于BCNF，但使得跨年度查询很困难，并且每年都需要新表
  - 例2，*company-year*(*company-id*, *earnings-2000*, *earnings-2001*, *earnings-2002*)
  - 也属于BCNF，但每年都需要添加新的属性
  - 这是一个crosstab的例子，将属性值作为了列名
  - 可用于spreadsheets，以及数据分析工具

总结

- 我们概述了一个将关系分解成BCNF的算法，有一些关系不存在保持依赖的BCNF分解
- 用正则覆盖将关系分解成3NF，它比BCNF的条件弱一些。属于3NF的关系也会含有冗余，但是总存在保持依赖的3NF分解
- 介绍了多值依赖的概念，它指明仅用函数依赖无法指明的约束
- 用多值依赖定义了4NF
- 其他的范式，如PJNF和DKNF，消除了更多细微形式的冗余。但是，它们难以操作而且很少使用
- 数据库设计过程中，要考虑的规范化问题

谢谢！