Informationssysteme

Sommersemester 2016

Prof. Dr.-Ing. Sebastian Michel
TU Kaiserslautern

smichel@cs.uni-kl.de

Wiederholung: Dritte Normalform

- Intuition: Nicht-Schlüssel Attribut darf kein anderes Nicht-Schlüssel Attribut bestimmen.

Ein Relationenschema $\mathcal{R}$ ist in dritter Normalform, wenn für jede für $\mathcal{R}$ geltende FD der Form $\alpha \rightarrow B$ mit Attribut $B \in \mathcal{R}$ mindestens eine von drei Bedingungen gilt:

1. $B \in \alpha$, d.h. die FD ist **trivial**
2. α is **Superschlüssel** von R
3. B ist **prim**

Eigenschaften:

- 3NF verhindert **partielle und transitive** Abhängigkeiten
- 3NF $\Rightarrow$ 2NF

Boyce-Codd-Normalform

Die Boyce-Codd-Normalform (BCNF) ist nochmals eine **Verschärfung der 3NF**. **Intuition:** Jedes Attribut darf **nur** den gesamten Schlüssel beschreiben und nichts anderes.

Ein Relationenschema $\mathcal{R}$ mit FDs $\mathcal{F}$ ist in BCNF, wenn für jede für $\mathcal{R}$ geltende funktionale Abhängigkeit der Form $\alpha \rightarrow B$ mit Attribut $B \in \mathcal{R}$ mindestens eine von zwei Bedingungen gilt:

1. $B \in \alpha$, d.h. die FD ist **trivial**
2. α ist Superschlüssel von $\mathcal{R}$

- Unterschied zu 3NF: 3. Bedingung fällt weg (B ist prim).
- **Man kann jede Relation verlustlos in BCNF-Relationen zerlegen**
- **Aber: manchmal lässt sich dabei die Abhängigkeitserhaltung nicht erziehlen!**

Städte ist in 3NF, aber nicht in BCNF

- Städte: {[Ort, Bland, Ministerpräsident/in, EW]}
- Geltende FDs:
  1. {Ort, Bland} $\rightarrow$ {EW}
  2. {Bland} $\rightarrow$ {Ministerpräsident/in}
  3. {Ministerpräsident/in} $\rightarrow$ {Bland}
- Kandidatenschlüssel:
  - {Ort, Bland}
  - {Ort, Ministerpräsident/in}
- linke Seite von FD1 ist Superschlüssel (Bedingung 2)
- rechten Seiten von FD2 und FD3 sind prim (Bedingung 3)
- Also: **3NF**
- linke Seiten von FD2 und FD3 sind aber **kein Superschlüssel**
- Also: **nicht in BCNF**
- **D.h.: BCNF verhindert zusätzlich zu 3NF transitive Abhängigkeiten zu Schlüssel-Attributen**
- wer welches Bundesland regiert wird hier also mehrfach abgespeichert

Dekomposition

Man kann grundsätzlich jedes Relationenschema $\mathcal{R}$ mit funktionalen Abhängigkeiten $\mathcal{F}$ so in $\mathcal{R}_1, ..., \mathcal{R}_n$ zerlegen, dass gilt

- $\mathcal{R}_1, ..., \mathcal{R}_n$ ist eine verlustlose Zerlegung von $\mathcal{R}$
- Alle $\mathcal{R}_1, ..., \mathcal{R}_n$ sind in BCNF.

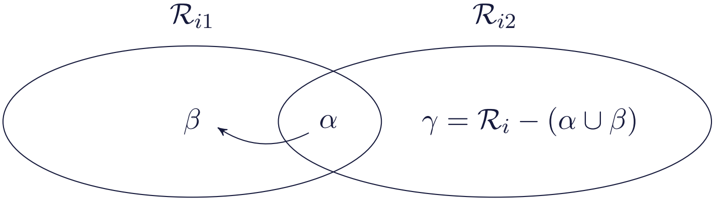
Es kann allerdings nicht immer erreicht werden, dass die Zerlegung $\mathcal{R}_1, ..., \mathcal{R}_n$ abhängigkeitserhaltend ist

Dies ist in der Praxis selten.

Dekompositions-Algorithmus für BCNF

- Starte mit $Z = \{\mathcal{R}\}$
- Solange es noch ein Relationenschame $\mathcal{R}_i \in Z$ gibt, das nicht in BCNF ist, mache folgendes:
  1. Es gibt also eine für $\mathcal{R}_i$ geltende nicht-triviale funktionale Abhängigkeit $\alpha \to \beta$ mit:
     - $\alpha \cap \beta = \emptyset$ (α und β sind disjunkt)
     - $\neg(\alpha \to \mathcal{R}_i)$ (α kein Superschlüssel von $\mathcal{R}_i$)
  2. Finde eine solche FD:
     Man sollte sie so wählen, dass β alle von α funktionalen abhängigen Attribute $B \in (\mathcal{R}_i - \alpha)$ enthält, damit der Dekompositionsalgorithmus möglichst schnell terminiert.
  3. Zerlege $\mathcal{R}_i$ in $\mathcal{R}_{i1} := \alpha \cup \beta$ und $\mathcal{R}_{i2} := \mathcal{R}_i - \beta$
  4. Entferne $\mathcal{R}_i$ aus Z und füge $\mathcal{R}_{i1}$ und $\mathcal{R}_{i2}$ ein:
     $Z := (Z - \mathcal{R}_i) \cup \mathcal{R}_{i1} \cup \mathcal{R}_{i2}$

Veranschaulichung eines Dekompositionsschrittes



Dekomposition der Relation Städte in BCNF-Relationen

- Städte: {[Ort, Bland, Ministerpräsident/in, EW]}
- Geltende FDs:
  1. {Ort, Bland} $\rightarrow$ {EW}
  2. {Bland} $\rightarrow$ {Ministerpräsident/in}
  3. {Ministerpräsident/in} $\rightarrow$ {Bland}

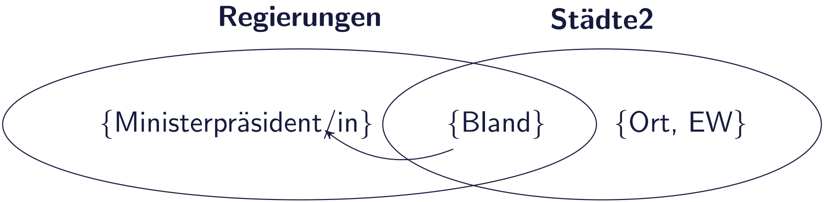
$\mathcal{R}_{i1}$

- **Regierungen:** {[Bland, Ministerpräsident/in]}

$\mathcal{R}_{i2}$

- **Städte:** {[Ort, Bland, EW]}

Veranschaulichung



- Zerlegung ist verlustlos und auch abhängigkeitserhaltend.
- **warum?**

Weiteres Beispiel: Dekomposition des PLZVerzeichnis in BCNF-Relationen

PLZVerzeichnis: {[Straße, Ort, Bland, PLZ]}

1. {Straße, Ort, Bland} → {PLZ} OK
2. {PLZ} → {Ort, Bland} verletzt BCNF **warum?**

Betrachte nun die Zerlegung:

- Orte: {[PLZ, Ort, Bland]}
- Straßen: {[PLZ, Straße]}

Diese Zerlegung

- ist verlustlos, **aber**
- **nicht abhängigkeitserhaltend, warum?**

Mehrwertige Abhängigkeiten

Bislang:

Funktionale Abhängigkeiten der Form

- Ausprägung R
- Seien $\alpha \subseteq \mathcal{R}$ und $\beta \subseteq \mathcal{R}$
- $\alpha \rightarrow \beta$ genau dann, wenn $\forall r, s \in R$ gilt: $r.\alpha = s.\alpha \Rightarrow r.\beta = s.\beta$
- D.h. die α-Werte bestimmen die β-Werte funktional (=eindeutig)

Nun: Mehrwertige Abhängigkeiten (multivalued dependencies)

Notation: $\alpha \twoheadrightarrow \beta$
Falls einem Attributwert α eine **Menge** von β-Werten zugeordnet werden.

Genaue Definition folgt (gleich).

Mehrwertige Abhängigkeiten: Beispiel

Fähigkeiten		
PersNr	Sprache	ProgSprache
3002	griechisch	C
3002	lateinisch	Pascal
3002	griechisch	Pascal
3002	lateinisch	C
3005	deutsch	Ada

Mehrwertige Abhängigkeiten dieser Relation:

- $\{\text{PersNr}\} \twoheadrightarrow \{\text{Sprache}\}$ und
- $\{\text{PersNr}\} \twoheadrightarrow \{\text{ProgSprache}\}$

MVDs führen zu Redundanz und Anomalien

Mehrwertige Abhängigkeiten

R		
A	B	C
a	b	c
a	bb	cc
a	b	cc
a	bb	c

- $A \twoheadrightarrow B$
- $A \twoheadrightarrow C$

Bei zwei Tupeln mit gleichen α-Werten kann man die β-Werte vertauschen und die resultierenden Tupel müssen auch in der Relation sein.

Mehrwertige Abhängigkeiten: Definition 1

	R		
	$\overbrace{A_1,\ldots,A_i}^{\alpha}$	$\overbrace{A_{i+1},\ldots,A_j}^{\beta}$	$\overbrace{A_{j+1},\ldots,A_n}^{\gamma}$
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	$a_1,\ldots,a_i$	$b_{i+1},\ldots,b_j$	$a_{j+1},\ldots,a_n$

$\alpha \twoheadrightarrow \beta$ **gilt genau dann, wenn für jede Ausprägung von R gilt:**

- wenn es zwei Tupel t_1 und t_2 mit gleichen α-Werten gibt, dann muss es auch zwei Tupel t_3 und t_4 geben mit
  - $t_1.\alpha = t_2.\alpha = t_3.\alpha = t_4.\alpha$ (alle α-Werte gleich)
  - $t_3.\beta = t_1.\beta$, $t_4.\beta = t_2.\beta$ (β-Paare gleich)
  - $t_3.\gamma = t_2.\gamma$, $t_4.\gamma = t_1.\gamma$ (γ-Paare **vertauscht**)

Veranschaulichung: Spezialfall

Veranschaulichung für MVD $\alpha \twoheadrightarrow \beta$:

Wenn α, β, γ jeweils nur aus einem Attribut A,B, und C bestehen:

Wenn $\{b_1, ..., b_i\}$ und $\{c_1, ..., c_j\}$ die B bzw. C-Werte für einen bestimmen A-Wert a sind, dann muss die Relation auch die folgenden $(i * j)$ Tupel enthalten:

$$\{a\} \times \{b_1, ..., b_i\} \times \{c_1, ..., c_j\}$$

Mehrwertige Abhängigkeiten: Definition 2

- Eine mehrwertige Abhängigkeit (multivalued dependency, MVD) $\alpha \twoheadrightarrow \beta$ besagt, dass einem Attribut α in $\mathcal{R}$ eine **Menge** von β-Werten zugeordnet werden.
- Wenn die MVD $\alpha \twoheadrightarrow \beta$ in R erfüllt, dann kann es als Erweiterung zu FDs α, β, c geben mit

$$|\pi_\beta(\sigma_{\alpha=c}(R))| > 1$$

- Diese Zuordnung ist **unabhängig** von den restlichen Attributen in $\mathcal{R}$

Mit anderen Worten:

- α bestimmt **nicht nur** einen **einzelnen** Wert (ein singleton)
- genau das ist ja bei einer normalen FD $\alpha \to \beta$ der Fall!
- **sondern** eine Menge von Werten
- diese Wertemenge ist unabhängig von den anderen Attributen in $\gamma = \mathcal{R} - \alpha - \beta$

Natürlich: Jede FD ist auch eine MVD (aber nicht umgekehrt)

Beispiel

Fähigkeiten		
PersNr	Sprache	ProgSprache
3002	griechisch	C
3002	lateinisch	Pascal
3002	griechisch	Pascal
3002	lateinisch	C
3005	deutsch	Ada

$\pi_{PersNr,Sprache}$

Sprachen	
PersNr	Sprache
3002	griechisch
3002	lateinisch
3005	deutsch

$\pi_{PersNr,ProgSprache}$

ProgSprachen	
PersNr	ProgSprache
3002	C
3002	Pascal
3005	Ada

Verlustlose Zerlegung bei MVDs: hinreichende + notwendige Bedingung

Die Zerlegung von $\mathcal{R}$ in $\mathcal{R}_1$ und $\mathcal{R}_2$ ist verlustlos genau dann, wenn

- $\mathcal{R} = \mathcal{R}_1 \cup \mathcal{R}_2$
- **und** mindestens eine von zwei MVDs gilt:
  - $\mathcal{R}_1 \cap \mathcal{R}_2 \twoheadrightarrow \mathcal{R}_1$ **oder**
  - $\mathcal{R}_1 \cap \mathcal{R}_2 \twoheadrightarrow \mathcal{R}_2$

Für unser Beispiel gilt:

- {PersNr, Sprache, ProgrSprache} = {PersNr, Sprache} $\cup$ {PersNr, ProgSprache}
- {PersNr} $\twoheadrightarrow$ {Sprache}
- {PersNr} $\twoheadrightarrow$ {ProgSprache}

D.h. es gelten sogar beide MVDs!

MVDs in Paaren

- **Es gilt zusätzlich:** wenn $\alpha \twoheadrightarrow \beta$, dann gilt immer
  - $\alpha \twoheadrightarrow \gamma$
  - mit $\gamma = \mathcal{R} - \alpha - \beta$
- **D.h. MVDs treten immer als Paare auf**
- Wir könnten MVDs deshalb auch so notieren: $\alpha \twoheadrightarrow \beta | \gamma$

Triviale MVDs ...

- .... sind solche, die von jeder Relationenausprägung R von $\mathcal{R}$ erfüllt werden.
- $\alpha \cup \beta \subseteq \mathcal{R}$
- Eine MVD $\alpha \twoheadrightarrow \beta$ ist trivial genau dann, wenn
  1. $\beta \subseteq \alpha$ oder
  2. $\beta = \mathcal{R} - \alpha$
- **Nur** die Bedingung 1 galt auch für normale FDs.
- Beispiel für Bedingung 2:
  - $\mathcal{R} = \{PersNr, Sprache\}$
  - $\alpha = \{PersNr\}$
  - $\beta = \{Sprache\}$
  - $\mathcal{R} - \alpha = \{PersNr, Sprache\} - \{PersNr\} = \{Sprache\} = \beta \Rightarrow$ MVD ist trivial!

Vierte Normalform

Eine Relation $\mathcal{R}$ ist in 4NF, wenn für jede MVD $\alpha \twoheadrightarrow \beta$ eine der folgenden Bedingungen gilt:

- Die MVD ist trivial **oder**
- α ist Superschlüssel von $\mathcal{R}$

- D.h. 4NF ist sehr ähnlich zu BCNF!
- Unterschied:
  - MVDs statt FDs
  - Definition von "trivial" wurde erweitert.
- 4NF erfüllt $\Rightarrow$ BCNF erfüllt, da jede FD eine MVD ist.

Dekomposition in 4NF

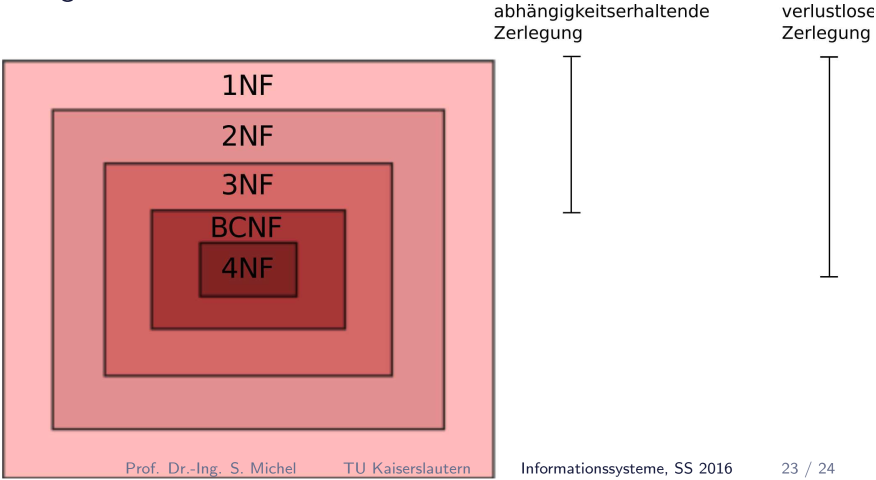
Starte mit der Menge $Z := \{\mathcal{R}\}$,

Solange es noch ein Relationenschema $\mathcal{R}_i$ in Z gibt, dass nicht in 4NF ist, mache folgendes:

- Es gibt also eine für $\mathcal{R}_i$ geltende nicht-triviale MVD $(\alpha \twoheadrightarrow \beta)$, für die gilt:
  - $\alpha \cap \beta = \emptyset$
  - $\neg(\alpha \rightarrow \mathcal{R}_i)$
- Finde eine solche MVD
- Zerlege $\mathcal{R}_i$ in $\mathcal{R}_{i1} := \alpha \cup \beta$ und $\mathcal{R}_{i2} := \mathcal{R}_i - \beta$
- Entferne $\mathcal{R}_i$ aus Z und füge $\mathcal{R}_{i1}$ und $\mathcal{R}_{i2}$ ein, also $Z := (Z - \mathcal{R}_i) \cup \{\mathcal{R}_{i1}\} \cup \{\mathcal{R}_{i2}\}$

Alle Normalformen auf einen Blick

- Die Verlustlosigkeit ist für alle Zerlegungsalgorithmen in alle Normalformen garantiert.
- Die Abhängigkeitserhaltung kann nur bis zur dritten Normalform garantiert werden.



Zusammenfassung Entwurfstheorie

- **Ziel: Bewertung der Güte einer Relation; Vermeidung von Anomalien durch Zerlegung in "bessere" Relationen**
- **Betrachtung von funktionalen Abhängigkeiten zur Identifikation von Redundanz**
- **Korrektheit von Zerlegung** definiert als verlustlos und abhängigkeitsbewahrend.
- **1NF, 2NF, 3NF, BCNF und 4NF betrachtet.**
- **Algorithmen zur Zerlegung/Synthese**