FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

# Datenbanksysteme I
# Datenbanken und Informationssysteme

Prof. Dr. Viktor Leis

WS 2019/2020

Professur für Datenbanken und Informationssysteme

# Relationale Entwurfstheorie

# Relationale Entwurfstheorie

- Wir haben bereits gesehen wir man mit Hilfe des ER-Modells zu einem relationalen Schema gelangt
- Nun kommt die Feinabstimmung dieses Schemas
- Normalformen dienen dazu die “Güte” eines Schemas formal zu bewerten

# Anomalien

- Ein schlecht entworfenes Schema führt zu folgenden Anomalien
  - Updateanomalien: bei Änderungen *eines* Fakts müssen *viele* Tupel angefasst werden
  - Einfügeanomalien: beim Einfügen eines Tupels können viele Werte nicht angegeben werden (im schlimmsten Fall sind nicht alle Schlüsselattributwerte bekannt $\Rightarrow$ Tupel kann nicht eingefügt werden)
  - Löschanomalien: beim Löschen eines Tupels geht mehr Information verloren als beabsichtigt

# Beispiel

- Ein Finanzberater speichert seine Daten in einem relationalen DBMS
- Es gibt (M)akler, ihre (B)üros, (I)nvestoren, (A)ktien, (D)ividenden und die (Q)uantität einer Aktie, die ein Investor hält
- Alles steht in einer großen Relation:

Finanz(M, B, I, A, D, Q)

# Beispiel(2)

Finanz

| M | B | I | A | D | Q |
|---|---|---|---|---|---|
| Müller | 103 | Schmidt | IBM | 1.20 | 50 |
| Müller | 103 | Schmidt | Merck | 1.00 | 30 |
| Müller | 103 | Schmidt | Coca Cola | 0.00 | 100 |
| Jung | 214 | Wilhelm | BASF | 0.80 | 80 |
| Jung | 214 | Wilhelm | Merck | 1.00 | 140 |
| Jung | 214 | Wilhelm | IBM | 1.20 | 30 |
| ... | ... | ... | ... | ... | ... |

- Was passiert, wenn Müller von 103 nach 145 umzieht?
- Was passiert, wenn Meier als Makler neu anfängt, aber noch keine Kunden betreut?
- Was passiert, wenn Wilhelm alle Aktien verkauft und aussteigt?

# Näher hingeschaut

- Die Qualität eines relationalen Schemas kann formal überprüft werden
- Auf den ersten Blick enthält die Relation `Finanz` sehr viel redundante Daten:
  - 103 ist Müllers Büro
  - IBM zahlt eine Dividende von 1.20 auf jede Aktie
  - Der Berater von Wilhelm ist Jung
- Wo kommt diese Redundanz her?
- Bestimmte Attributwerte legen andere Attributwerte fest, es gibt *funktionale Abhängigkeiten*

# Funktionale Abhängigkeiten

- Z.B. gibt es eine funktionale Abhängigkeit (kurz FD, für functional dependency) zwischen `Makler` und `Büro`: Makler $\rightarrow$ Büro
- Formale Definition einer FD
  - $\alpha$ und $\beta$ sind Attributmengen eines relationalen Schemas $\mathcal{R}$
  - Es gibt eine FD $\alpha \rightarrow \beta$, gdw. für *alle* Instanzen von $\mathcal{R}$ gilt: für alle Paare von Tupeln $r, t \in R$ gilt $r.\alpha = t.\alpha \Rightarrow r.\beta = t.\beta$

# Funktionale Abhängigkeiten(2)

- Es ist unrealistisch, alle möglichen Instanzen von $\mathcal{R}$ in der Praxis zu prüfen
- FDs müssen aus dem Hintergrundwissen über die Anwendung abgeleitet werden
- Welche anderen FDs gibt es in `Finanz`?

# Funktionale Abhängigkeiten(3)

- $\mathcal{F}_{\mathcal{R}} = \{M \to B, A \to D, I \to M, IA \to Q\}$
- Was machen wir jetzt mit diesen FDs?
- Man kann einen Schlüssel für $\mathcal{R}$ *berechnen*!

# Schlüsselberechnung

- Eigenschaften eines Schlüssels $\kappa$:
  1. $\kappa \subseteq \mathcal{R}$
  2. $\kappa \rightarrow \mathcal{R}$
  3. Es gibt kein $\kappa' \subset \kappa$ so dass $\kappa' \rightarrow \mathcal{R}$
- Eigenschaft 2. wird Vollständigkeit genannt, Eigenschaft 3. Minimalität
- $\kappa$ wird auch *Kandidatenschlüssel* genannt (es kann mehr als ein $\kappa$ geben, das obige Eigenschaften erfüllt)
- Ein $\kappa$ wird als *Primärschlüssel* gewählt
- Wenn nur Eigenschaften 1. und 2. gelten, wird $\kappa$ *Superschlüssel* genannt

# Schlüsselberechnung(2)

- Ist `IA` ein Schlüssel von `Finanz`?
- Eigenschaft 1: $\surd$
- Für die Überprüfung von 2. und 3. brauchen wir weitere Konzepte der Relationentheorie

# Herleitung weiterer FDs

- Aus einer Menge $\mathcal{F}$ von FDs sind weitere FDs herleitbar
- $\mathcal{F}^+$, die Menge aller aus $\mathcal{F}$ herleitbaren FDs, wird Hülle (closure) von $\mathcal{F}$ genannt
- Es gibt Inferenzregeln, die *Armstrong Axiome*, zum Herleiten weiterer FDs (siehe Buch)
- Da die Anwendung der Armstrong Axiome für die Schlüsselfindung etwas aufwendig ist, benutzt man meistens Attributhüllen.

# Attributhülle

- Die *Attributhülle* $AH(\alpha)$ einer Attributmenge $\alpha$ ist die Menge alle Attribute aus $\mathcal{R}$ die funktional von $\alpha$ abhängen
- Es gibt einen einfachen Algorithmus zur Bestimmung von $AH(\alpha)$:

AH := $\alpha$
while (AH ändert sich noch) do
    for each FD $\beta \rightarrow \gamma$ in $\mathcal{F}_{\mathcal{R}}$ do
        if ($\beta \subseteq AH$)
        then $AH := AH \cup \gamma$

# Attributhülle(2)

- $\mathcal{F}_{\mathcal{R}} = \{M \to B, A \to D, I \to M, IA \to Q\}$
- $AH(IA) = MBIAQD$
  $\Rightarrow IA$ ist Superschlüssel
- $AH(I) = IMB$
  $AH(A) = AD$
  $\Rightarrow IA$ ist auch Schlüssel

# Schemaqualitätsprüfung

- Es gibt *Normalformen* (NF) die etwas über die Qualität eines Schemas aussagen
- Wir betrachten 1NF, 2NF, 3NF, BCNF, 4NF
- Es gibt noch höhere, die aber mehr von theoretischem Interesse sind

# Erste Normalform (1NF)

- Ein relationales Schema ist in 1NF, gdw. alle Attribute nur atomare Werte annehmen

Nicht in 1NF:

Eltern

| Mutter | Vater | Kinder |
|---|---|---|
| Marie | Hans | {Ines, David} |
| ... | ... | ... |

In 1NF:

Eltern

| Mutter | Vater | Kind |
|---|---|---|
| Marie | Hans | Ines |
| Marie | Hans | David |
| ... | ... | ... |

# Zweite Normalform (2NF)

- Ein Relationenschema ist in 2NF, gdw. es in 1NF ist und jedes Nichtschlüsselattribut (NSA) voll funktional von jedem Schlüssel abhängt
- $\beta$ hängt voll funktional von $\alpha$ ab ($\alpha \overset{\bullet}{\rightarrow} \beta$), gdw. $\alpha \rightarrow \beta$ und es existiert kein $\alpha' \subset \alpha$, so dass $\alpha' \rightarrow \beta$
- Jedes Nichtschlüsselattribut ist von allen ganzen Schlüsseln abhängig
- Verstoß gegen 2NF deutet darauf hin, dass verschiedene Beziehungen (zwischen Entitäten) in einer Relation gemischt wurden

# Verletzung 2NF

- `Finanz` ist in 1NF, aber nicht in 2NF, da *IA* Schlüssel, aber $I \rightarrow M$
- Intuition: Redundanz im Makler Attribut

Finanz

| M | B | I | A | D | Q |
|---|---|---|---|---|---|
| Müller | 103 | Schmidt | IBM | 1.20 | 50 |
| Müller | 103 | Schmidt | Merck | 1.00 | 30 |
| Müller | 103 | Schmidt | Coca Cola | 0.00 | 100 |
| Jung | 214 | Wilhelm | BASF | 0.80 | 80 |
| Jung | 214 | Wilhelm | Merck | 1.00 | 140 |
| Jung | 214 | Wilhelm | IBM | 1.20 | 30 |
| ... | ... | ... | ... | ... | ... |

# Dritte Normalform (3NF)

- Selbst bei Erfüllung von 2NF können immer noch Redundanzen im Schema enthalten sein (durch transitive Abhängigkeiten)
- Beispiel transitive Abhängigkeit: $S \rightarrow X$, $X \rightarrow Y$
- Ein Relationenschema ist in 3NF, gdw. für jede FD $\alpha \rightarrow \beta$ mindestens eine der folgenden Eigenschaften gilt:
  - $\alpha \rightarrow \beta$ ist trivial, d.h., $\beta \subseteq \alpha$
  - $\alpha$ ist ein Superschlüssel
  - Jedes Attribut in $\beta$ ist in einem Schlüssel enthalten
- Auf diese Weise werden transitive Abhängigkeiten vermieden (in denen NSAe über andere NSAe vom Schlüssel abhängen)

# Boyce-Codd Normalform (BCNF)

- Weitere Verschärfung ist BCNF, hier dürfen alle Attribute nur noch direkt vom Schlüssel abhängen
- Ein Relationenschema ist in BCNF, gdw. für jede FD $\alpha \rightarrow \beta$ mindestens eine der folgenden Eigenschaften gilt:
  - $\alpha \rightarrow \beta$ ist trivial, d.h., $\beta \subseteq \alpha$
  - $\alpha$ ist ein Superschlüssel
- Für weitere Beschreibungen von Redundanzen müssen wir über FDs hinausgehen

# Mehrwertige Abhängigkeiten

| Fähigkeiten | | |
|---|---|---|
| PersNr | Sprache | ProgSprache |
| 3002 | Englisch | C |
| 3002 | Deutsch | C |
| 3002 | Englisch | Java |
| 3002 | Deutsch | Java |
| 3005 | Englisch | C |
| 3005 | Deutsch | C |

- In diesem Schema gibt es keine FDs, trotzdem haben wir Redundanz

# Mehrwertige Abhängigkeiten(2)

- Jemand der fünf Programmiersprachen und vier Sprachen beherrscht, benötigt 20 Tupel zur Speicherung dieser Information!
- Hier werden voneinander unabhängige Konzepte miteinander vermischt
- Kompaktere Speicherung möglich:

| SprachFähigkeiten | |
|---|---|
| PersNr | Sprache |
| 3002 | Englisch |
| 3002 | Deutsch |
| 3005 | Englisch |
| 3005 | Deutsch |

| ProgFähigkeiten | |
|---|---|
| PersNr | ProgSprache |
| 3002 | C |
| 3002 | Java |
| 3005 | C |

# Mehrwertige Abhängigkeiten(3)

- Es gibt *mehrwertige Abhängigkeiten* (MVDs: multivalued dependencies) in *Fähigkeiten*: PersNr $\twoheadrightarrow$ Sprache und PersNr $\twoheadrightarrow$ ProgSprache
- Formale Definition: $\alpha, \beta, \gamma \subseteq \mathcal{R}$ (mit $\alpha \cup \beta \cup \gamma = \mathcal{R}$); $\alpha \twoheadrightarrow \beta$ gilt, gdw. für jede Instanz $R$ gilt: für jedes Paar von Tupeln $t_1, t_2$ in $R$ mit $t_1.\alpha = t_2.\alpha$ existiert ein $t_3 \in R$ mit $t_3.\alpha = t_1.\alpha$, $t_3.\beta = t_1.\beta$ und $t_3.\gamma = t_2.\gamma$
- Umgangssprachlich bedeutet dies, dass für alle Tupel mit gleichem Wert für $\alpha$ alle $\beta, \gamma$-Kombinationen vorkommen

# Mehrwertige Abhängigkeiten(4)

- MVDs sind eine Verallgemeinerung von FDs, d.h. jede FD ist eine MVD (aber nicht unbedingt umgekehrt)
- Ähnlich den Armstrong Axiomen gibt es auch für MVDs Herleitungsregeln (sollen hier aber nicht im Detail besprochen werden)

# Vierte Normalform

- In der vierten Normalform (4NF) werden die Eigenschaften der BCNF auf MVDs ausgeweitet
- Ein Relationenschema ist in 4NF, gdw. für jede MVD $\alpha \twoheadrightarrow \beta$ mindestens eine der folgenden Eigenschaften gilt:
  - $\alpha \twoheadrightarrow \beta$ ist trivial, d.h., $\beta \subseteq \alpha$ **ODER** $\alpha \cup \beta = \mathcal{R}$
  - $\alpha$ ist ein Superschlüssel

# Normalformen

- Eine höhere NF schließt alle niedrigeren mit ein:
  $4NF \subset BCNF \subset 3NF \subset 2NF \subset 1NF$
- Je höher die NF, desto besser das Schema, d.h. desto weniger Redundanzen
- Was macht man, wenn die Qualität des momentanen Schemas nicht gut genug ist?
- Man überführt Schema in eine höhere Normalform und zwar mit Hilfe von Zerlegungen

# Zerlegung von Relationen

- Ein Schema $\mathcal{R}$ wird in die Teilschemata $\mathcal{R}_1, \ldots, \mathcal{R}_n$ zerlegt, mit $\mathcal{R}_i \subset \mathcal{R}$ für $1 \leq i \leq n$
- Eine Zerlegung sollte zwei Eigenschaften haben:
  - Verlustlosigkeit: die in der ursprünglichen Instanz $R$ enthaltene Information muss aus den Instanzen $R_1, \ldots R_n$ rekonstruierbar sein (für alle möglichen Instanzen $R$)
  - Abhängigkeitsbewahrung: alle FDs in $\mathcal{F}_\mathcal{R}$ sollten in den $\mathcal{F}_{\mathcal{R}_1}, \ldots, \mathcal{F}_{\mathcal{R}_n}$ bewahrt bleiben

# Verlustlosigkeit

- Zerlegung von $\mathcal{R}$ in $\mathcal{R}_1$ und $\mathcal{R}_2$ (mit $\mathcal{R} = \mathcal{R}_1 \cup \mathcal{R}_2$ und $R_1 = \pi_{\mathcal{R}_1}(R)$, $R_2 = \pi_{\mathcal{R}_2}(R)$)
- Die Zerlegung ist verlustlos, wenn für jede mögliche Instanz $R$ von $\mathcal{R}$ gilt:

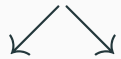
$$R = R_1 \bowtie R_2$$

- Hinreichende Bedingung für verlustlose Zerlegung:
  - $(\mathcal{R}_1 \cap \mathcal{R}_2) \rightarrow \mathcal{R}_1 \in \mathcal{F}_{\mathcal{R}}^+$ oder
  - $(\mathcal{R}_1 \cap \mathcal{R}_2) \rightarrow \mathcal{R}_2 \in \mathcal{F}_{\mathcal{R}}^+$

# Beispiel für Verlust

Finanz

| M | B | I | A | D | Q |
|---|---|---|---|---|---|
| Müller | 103 | Schmidt | IBM | 1.20 | 50 |
| Müller | 103 | Schmidt | Merck | 1.00 | 30 |
| Jung | 214 | Wilhelm | Merck | 1.00 | 140 |
| Jung | 214 | Wilhelm | IBM | 1.20 | 30 |



Finanz1

| M | B | A |
|---|---|---|
| Müller | 103 | IBM |
| Müller | 103 | Merck |
| Jung | 214 | Merck |
| Jung | 214 | IBM |

Finanz2

| I | A | D | Q |
|---|---|---|---|
| Schmidt | IBM | 1.20 | 50 |
| Schmidt | Merck | 1.00 | 30 |
| Wilhelm | Merck | 1.00 | 140 |
| Wilhelm | IBM | 1.20 | 30 |

# Beispiel für Verlust(2)

Finanz1 ⋈ Finanz2

Finanz

| M | B | I | A | D | Q |
|---|---|---|---|---|---|
| Müller | 103 | Schmidt | IBM | 1.20 | 50 |
| Müller | 103 | Schmidt | Merck | 1.00 | 30 |
| Jung | 214 | Wilhelm | Merck | 1.00 | 140 |
| Jung | 214 | Wilhelm | IBM | 1.20 | 30 |
| Müller | 103 | Wilhelm | IBM | 1.20 | 50 |
| Müller | 103 | Wilhelm | Merck | 1.00 | 30 |
| Jung | 214 | Schmidt | Merck | 1.00 | 140 |
| Jung | 214 | Schmidt | IBM | 1.20 | 30 |

# Abhängigkeitsbewahrung

- Zerlegung von $\mathcal{R}$ in $\mathcal{R}_1, \ldots, \mathcal{R}_n$
- Zerlegung ist abhängigkeitsbewahrend, wenn
  $\mathcal{F}_{\mathcal{R}} \equiv (\mathcal{F}_{\mathcal{R}_1} \cup \cdots \cup \mathcal{F}_{\mathcal{R}_n})$ bzw.
  $\mathcal{F}^+_{\mathcal{R}} = (\mathcal{F}_{\mathcal{R}_1} \cup \cdots \cup \mathcal{F}_{\mathcal{R}_n})^+$
- Im vorherigen Beispiel geht die FD $I \rightarrow M$ verloren, ist also auch nicht abhängigkeitsbewahrend

# Zerlegungsalgorithmen

- Zerlegung kann automatisch durchgeführt werden
- Wichtigster Algorithmus: 3NF-Synthesealgorithmus
  - Zerlegt ein Schema verlustlos und abhängigkeitsbewahrend in 3NF
  - Braucht als Eingabe allerdings eine redundanzfreie Menge von FDs (kanonische Überdeckung, siehe Buch)

# Beispiel

- Anwendung des 3NF-Synthesealgorithmus auf `Finanz` mit $\mathcal{F}_{\mathcal{R}} = \{M \to B, A \to D, I \to M, IA \to Q\}$
- Ergebnis:
  - $\mathcal{R}_M(M, B)$
  - $\mathcal{R}_I(I, M)$
  - $\mathcal{R}_A(A, D)$
  - $\mathcal{R}_{IA}(I, A, Q)$
- Was ist mit den Anomalien vom Anfang des Kapitels passiert? Sie sind verschwunden!

# Beispiel(2)

Finanz

| M | B | I | A | D | Q |
|---|---|---|---|---|---|
| Müller | 103 | Schmidt | IBM | 1.20 | 50 |
| Müller | 103 | Schmidt | Merck | 1.00 | 30 |
| Müller | 103 | Schmidt | Coca Cola | 0.00 | 100 |
| Jung | 214 | Wilhelm | BASF | 0.80 | 80 |
| Jung | 214 | Wilhelm | Merck | 1.00 | 140 |
| Jung | 214 | Wilhelm | IBM | 1.20 | 30 |

# Beispiel (3)

Makler

| M | B |
|---|---|
| Müller | 103 |
| Müller | 103 |
| Müller | 103 |
| Jung | 214 |
| Jung | 214 |
| Jung | 214 |

Dividende

| A | D |
|---|---|
| IBM | 1.20 |
| Merck | 1.00 |
| Coca Cola | 0.00 |
| BASF | 0.80 |
| Merck | 1.00 |
| IBM | 1.20 |

Finanz

| M | I | A | Q |
|---|---|---|---|
| Müller | Schmidt | IBM | 50 |
| Müller | Schmidt | Merck | 30 |
| Müller | Schmidt | Coca Cola | 100 |
| Jung | Wilhelm | BASF | 80 |
| Jung | Wilhelm | Merck | 140 |
| Jung | Wilhelm | IBM | 30 |

# Beispiel (4)

Makler

| M | B |
|---|---|
| Müller | 103 |
| Müller | 103 |
| Müller | 103 |
| Jung | 214 |
| Jung | 214 |
| Jung | 214 |

Dividende

| A | D |
|---|---|
| IBM | 1.20 |
| Merck | 1.00 |
| Coca Cola | 0.00 |
| BASF | 0.80 |
| Merck | 1.00 |
| IBM | 1.20 |

Berater

| M | I |
|---|---|
| Müller | Schmidt |
| Müller | Schmidt |
| Müller | Schmidt |
| Jung | Wilhelm |
| Jung | Wilhelm |
| Jung | Wilhelm |

Depot

| I | A | Q |
|---|---|---|
| Schmidt | IBM | 50 |
| Schmidt | Merck | 30 |
| Schmidt | Coca Cola | 100 |
| Wilhelm | BASF | 80 |
| Wilhelm | Merck | 140 |
| Wilhelm | IBM | 30 |

# Zerlegung in BCNF und 4NF

- Es gibt noch weitere Zerlegungsalgorithmen für BCNF und 4NF
- Problem: es gibt Schemata, die nicht abhängigkeitsbewahrend in BCNF oder 4NF zerlegt werden können
- Meistens gibt man sich mit 3NF zufrieden, weiterer Grund: höhere Normalformen bevorzugen Updateoperationen vor Anfrageoperationen

# Zusammenfassung

- Mit Hilfe von Normalformen kann die Qualität eines Schemas bestimmt werden
- Es gibt Algorithmen, die ein Schema normalisieren können
- Mit etwas Erfahrung erhält man durch ein ER-Modellierung normalisiertes Schema
- In Ausnahmefällen kann es (z.B. aus Performancegründen) vorteilhaft das Schema teilweise zu denormalisieren