Datamodellering og normalisering

Sidste gang: Modelleringssprog ODL og E/R og principper for oversættelse til Rel.Alg.

I dag: **Normalisering** $\sim$ at forbedre relationelt design.

Splitte relationer op i mindre relationer, så

1. Der opnås “kvalitetsegenskaber”
2. Informationen bevares

Notation for i dag:
$A \sim A_1, A_2, \ldots, A_{n_1}$
$B \sim B_1, B_2, \ldots, B_{n_2}$
$C \sim C_1, C_2, \ldots, C_{n_3}$
hvor $A_1, A_2, \ldots, B_1, B_2, \ldots, C_1, C_2, \ldots$ er attributter.

Princip i normalisering

Dekomponér $R(A, B, C)$ til $\pi_{A,B}(R)$ og $\pi_{A,C}(R)$ så

$$\pi_{A,B}(R) \bowtie \pi_{A,C}(R) = R$$

OBS:

- $\bowtie$ giver potentielt mange flere tupler,
- ergo giver dekomposition potentielt meget færre tupler.

De gode spørgsmål:

- Hvilke egenskaber ønsker vi at opnå/undgå?
- Hvordan kan A, B og C vælges så betingelsen er overholdt??

Problemet redundans

- Samme info. er gentaget potentielt mange steder
- komplicerede constraints som skal overholdes/gennemtvinges
- (alternativt: risiko for "inkonsistente" databaser)
- komplicerede transaktioner, svært at programmere.

"Blive-væk"-problemet

Hvis givet information altid og kun er hæftet på de tupler, som bruger den, f.eks.:

R:

C	Postnr	By
$\cdots$	$\cdots$	$\cdots$
$\cdots$	2630	Taastrup
$\cdots$	$\cdots$	$\cdots$

Hvis der pludselig ikke er "$\cdots$"er i Taastrup, så forsvinder al info. om sammenhæng 2630 $\leftrightarrow$ Taastrup!
Lægger op til: Postnr og By indtastes hver gang IC checker for inkonsistens (ikke ét Postnr. med to Byer).

Bemærk her:

$$\pi_{C,\mathrm{Postnr}}(R) \bowtie \pi_{\mathrm{Postnr},\mathrm{By}}(R) = R$$

Bemærk yderligere:

Lad P = den offielle danske postnummerliste, dvs.

$$\pi_{\mathrm{Postnr},\mathrm{By}}(R) \subseteq P$$

og vi har til enhver tid

$$\pi_{C,\mathrm{Postnr}}(R) \bowtie P = R$$

Smart, ikk'?

Funktionel afhængighed, kilde til redundans + mulighed for dekomposition

Definition: En *funktionel afhængighed* for relation(sskema) $R(A, B, C)$,

$$A \to B$$

er den begrænsning, at hvis

$$\langle a, b_1, c_1 \rangle \in R$$

$$\langle a, b_2, c_2 \rangle \in R$$

da er

$$b_1 = b_2$$

Eksempel

R:

C	Postnr	By
$\cdots$	$\cdots$	$\cdots$
$\cdots$	2630	Taastrup
$\cdots$	$\cdots$	$\cdots$

Vi må forvente

$$\text{Postnr} \to \text{By}$$

men ikke den anden vej.

Generel observation

For relation $R(A, B, C)$ med constraint:

> $\pi_{A,B}(R) \subseteq R_0$ for en eller anden ***statisk*** (=uforanderlig)
> relation $R_0(A, B)$ med $A \to B$

har vi for ***enhver instans af*** R at:

$$R_0 \bowtie \pi_{A,C}(R) = R.$$

Dvs. R_0 kan oprettes én gang for alle og ikke opdateres sidenhen;
kun $\pi_{A,C}(R)$ skal opdateres.

Definition: En mgd. attributter A er en *nøgle* for relation $R(A, B)$ såfremt $A \to B$ og A er minimal.

Definition: En *supernøgle* for relation R er en mgd. attributter, som indeholder en nøgle.

Dvs. som en nøgle, men ikke nødvendigvis minimal.

Eksempel

Fornavn	Efternavn	Gade/vej	Nr	Postnr	By
Katrine	Petersen	Strandvejen	1	8000	Århus C
Peter	Jensen	Skolegade	7	8000	Århus C
Hans	Jensen	Jomfru Ane Gade	3	9000	Ålborg
Børge	Børgesen	Markvej	1	5772	Kværndrup
Peter	Hansen	Snevej	312	3900	Nuuk
Petrine	Hansen	Snevej	312	3900	Nuuk
Peter	Hansen	Morbærvej	8	4200	Slagelse

Forventet nøgle: {Fornavn, Efternavn, Gade/vej, Nr, Postnr}

Er Postnr. en nøgle? nej, men mindre vi antager højst én agent pr. Postnr!!!!!!!

Antages højst en agent pr. fysisk gade/vej, så er {Gade/vej, Postnr} en nøgle.

Eksempler på supernøgle ved "højst en agent pr. fysisk gade/vej":

{Gade/vej, Postnr, By}, {Efternavn, Gade/vej, Nr, Postnr},
{Fornavn, Efternavn, Gade/vej, Nr, Postnr}

Eksempel

P:

Postnr	By
...	...
2630	Taastrup
...	...

Her forventes Postnr. at være nøgle.

{Postnr, By} er supernøgle.

Læs selv følgende i bogen:

- 3.6 om hvordan man udfra givne funktionelle afhængigheder for R kan regne sig frem til *alle* funktionelle afhængigheder for R, f.eks.:

  Hvis $A \rightarrow B$ *og* $B \rightarrow C$, *så* $A \rightarrow C$

- 3.7.5 om hvordan man projicerer funktionelle afhængigheder ud på delrelationer.

Boyce-Codd normal form, BCNF

Definition: En relation R er i BCNF såfremt for enhver ikke-triviel funktionel afhængighed $A \to B$, at A er en supernøgle for R.

Dvs. A indeholder en nøgle.

Eksempel: Antag $R(A, B, C)$ i BCNF og $A \to B$.

- A indeholder en nøgle.
- Dvs. given værdi a for A bestemmer højst én tupel $\langle a, b, c \rangle \in R$;
- Skriv tuplen som $\langle a, F(a), c \rangle \in R$.
- Koblingen $a \mapsto F(a)$ er registreret højst én gang i R.
- Ergo ingen redundans i den anledning; dekomposition uinteressant.

Hvorfor "supernøgle" i def.? Fordi der står "for enhver"!

Bemærk, at hvis $A \to B$, gælder også $AX \to B$ for vilkårlig attribut X

Eksempel: Hvis Postnr $\to$ By, gælder også Postnr, Hårfarve $\to$ By.

Hvordan opnås BCNF?

Lad R være en relation og $\mathbf{R}$ være en variabel, som holder en mængde relationer; til start $\mathbf{R} := \{R\}$.

1. Hvis alle relationer i $\mathbf{R}$ er BCNF er vi færdige, ellers find $S \in \mathbf{R}$ som ikke er BCNF.
2. Lad $A \to B$ være ikke-triviel funktionel afhængighed i S så A ikke er supernøgle; antag S har skema $S(A, B, C)$.
3. $\mathbf{R} := \mathbf{R} \backslash \{S\} \cup \{\pi_{AB}(S), \pi_{AC}(S)\}$
4. Gå til 1.

Heuristik: Vælg A så lille som mulig, B så stor som mulig.

Overvej: Er nogen af de resulterende relationer i $\mathbf{R}$ delmængder af statiske relationer i stil med (Postnr, By)??

Et problem ved BCNF, som motiverer 3NF

Funktionel afhængighed er en constraint, som skal opretholdes ved opdatering.

OBS: I visse tilfælde kan en "oprindelig" funktionel afhængighed for relation ikke beskrives som funktionelle afhængigheder for de afledte relationer!

Ønske kan være: "Bevare" funktionel afhængighed ved at undlade dekomposition til BCNF. Prisen: en smule redundans.

Eksempel (fra bogen, afs. 3.7.7): Bookings(title, theater, city)

$$\text{theater} \rightarrow \text{city} \qquad\qquad \text{title city} \rightarrow \text{theater}$$

Der kan observeres to mulige nøgler:

$$\{\text{title, city}\} \qquad\qquad \{\text{theater, title}\}$$

Eksempel på ok-instans:

title	theater	city
The Net	Park	Menlo Park
The Net	DriveIn	SF
Start Wars	DriveIn	SF
Star Wars	Guild	Menlo Park

OBS: Indeholder redundans "DriveIn → SF" optræder to gange!
OBS: Ikke BCNF da "theater" i "theater → city" ikke er supernøgle!

Ergo dekomponere ud i $\pi_{\text{theater,city}}(\text{Bookings})$ og $\pi_{\text{theater,title}}(\text{Bookings})$.

Men betragt følgende instanser af de afledte:

Bookings_1:

theater	city
Guild	Menlo Park
Park	Menlo Park

Bookings_2:

theater	title
Guild	The net
Park	The net

Bemærk overskridelse af "title city → theater".
Pointen: Dette kan ikke checkes ved "afledte" funk.afh. på de to delrel.
Formulering af funk.afh. bliver et relationelt udtryk som indeholder

$\text{Bookings}_1 \bowtie \text{Bookings}_2$.

Så hvorfor dekomponere overhovedet?!?!?

3NF

Slække på BCNF, så vi undgår problem med at check af funk.afh. kræver udregning af "$\bowtie$".

Definition: En relation R er i 3NF såfremt for enhver ikke-triviel funktionel afhænighed $A \rightarrow B$, at A er en supernøgle for R eller B indgår i en nøgle.

Påstand: Det virker!

Eksemplet fra før: Bookings(title, theater, city)

$$\text{theater} \rightarrow \text{city} \qquad\qquad \text{title city} \rightarrow \text{theater}$$

Nøgler:

$$\{\text{title, city}\} \qquad\qquad \{\text{theater, title}\}$$

Funk.afh. "theater $\rightarrow$ city" er BCNF-problematisk, men 3NF-acceptabel!

Ergo vil 3NF-tilhængeren vælge ikke at dekomponere.

NB: Hvis vi kan ***optimere*** check af funktionel afhængighed så "$\bowtie$" undgås, svækker det argument for 3NF.

(Måske når vi i kurset til "simplification" af integrity constraints, jvf. Nicolas, 1982, Quian, 1988, o.m.a)

Flerværdiafhængigheder

Problem: Redundans som skyldes at Rel.Alg. ikke tillader attributter at have mængder som attributter.

"Flerværdiafhængigheder havde været et specialtilfælde af funktionelle afhængigheder hvis Rel.Alg. havde tilladt mængdeatributter"

Opstår f.eks. ved at oversætte ODL's mdg. attributter eller ODL+E/R's sammenhænge til Rel.Alg., når der er mere end én mængde — *Eller som en egenskab, man ikke var klar over.*

Eksempel

Antag en "mængdeattributrelation" med følgende "tupler":

$$\left\langle 1, \left\{ \begin{array}{c} \text{kold} \\ \text{varm} \end{array} \right\}, \left\{ \begin{array}{c} \text{vanille} \\ \text{banan} \\ \text{citron} \end{array} \right\} \right\rangle \qquad \left\langle 2, \left\{ \begin{array}{c} \text{lunken} \\ \text{hundehold} \end{array} \right\}, \left\{ \begin{array}{c} \text{pistacie} \\ \text{chokolade} \end{array} \right\} \right\rangle$$

Udfoldet til "Rel.Alg.-relation" får vi følgende R, som er i BCNF;
i "$BC \rightarrow A$" er "BC" en nøgle!

R:

A	B	C
1	kold	vanille
1	kold	banan
1	kold	citron
1	varm	vanille
1	varm	banan
1	varm	citron
2	lunken	pistacie
2	lunken	chokolade
2	hundekold	pistacie
2	hundekold	chokolade

Multiværdiafhængighed $A \twoheadrightarrow B$:
Til $A = 1$ hører mgd. af B-værdier
$B_{A=1} = \{kold, varm\}$.
Til $A = 2$ hører mgd. af B-værdier
$B_{A=2} = \{lunken, hundekoldt\}$.

Multiværdiafhængighed $A \twoheadrightarrow C$:
Til $A = 1$ hører mgd. af C-værdier
$C_{A=1} = \{vanille, banan, citron\}$.
Til $A = 2$ hører mgd. af C-værdier
$C_{A=2} = \{pistacie, chokolade\}$.

Flerværdiafhængigheder, fortsat

Eksemplet

R:

A	B	C
1	kold	vanille
1	kold	banan
1	kold	citron
1	varm	vanille
1	varm	banan
1	varm	citron
2	lunken	pistacie
2	lunken	chokolade
2	hundekold	pistacie
2	hundekold	chokolade

Multiværdiafhængighed $A \twoheadrightarrow B$:
Til $A = 1$ hører mgd. af B-værdier
$B_{A=1} = \{kold, varm\}$.
Til $A = 2$ hører mgd. af B-værdier
$B_{A=2} = \{lunken, hundekoldt\}$.

Multiværdiafhængighed $A \twoheadrightarrow C$:
Til $A = 1$ hører mgd. af C-værdier
$C_{A=1} = \{vanille, banan, citron\}$.
Til $A = 2$ hører mgd. af C-værdier
$C_{A=2} = \{pistacie, chokolade\}$.

Notation:
$B_{A=a,C=c}$ mgd. af b hvor $\langle a, b, c\rangle \in R$; $B_{A=a}$ fælles værdi for $B_{A=a,C=?}$
$C_{A=a,B=b}$ mgd. af c hvor $\langle a, b, c\rangle \in R$; $C_{A=a}$ fælles værdi for $C_{A=a,B=?}$

Karakteristik af f.v.a. $A \twoheadrightarrow B$ (alternativ til bogen)

For alle A-værdier a og
alle C-værdier c, c' med $B_{A=a,C=c}$ og $B_{A=a,C=c'}$ ikke tomme,
er $B_{A=a,C=c} = B_{A=a,C=c'} = B_{A=a}$
og
alle B-værdier b, b' med $C_{A=a,B=b}$ og $C_{A=a,B=b'}$ ikke tomme,
er $C_{A=a,B=b} = C_{A=a,B=b'} = C_{A=a}$.

Altså: $R = \{a_1\} \times B_{A=a_1} \times C_{A=a_1} \cup \{a_2\} \times B_{A=a_2} \times C_{A=a_2} \cup \cdots$

Vi har altså: $A \twoheadrightarrow B$ hviss $A \twoheadrightarrow C$.

Bogens def. af flerværdiafhængighed $A \twoheadrightarrow B$

For alle a og alle

$$\langle a, b_1, c_1 \rangle, \langle a, b_2, c_2 \rangle \in R,$$

gælder

$$\langle a, b_1, c_2 \rangle \in R.$$

(og ved symmetri

$$\langle a, b_2, c_1 \rangle \in R).$$

Eksempel

R:

A	B	C
1	kold	vanille
1	kold	banan
1	kold	citron
1	varm	vanille
1	varm	banan
1	varm	citron
...	...	...

Lad $\langle a, b_1, c_1 \rangle = \langle 1, kold, vanille \rangle$
$\langle a, b_2, c_2 \rangle = \langle 1, varm, citron \rangle$
og dermed
$\langle a, b_1, c_2 \rangle = \langle 1, kold, citron \rangle$
$\langle a, b_2, c_1 \rangle = \langle 1, varm, vanille \rangle$

Afskaffe redundans ved flerværdiafhængighed: Dekomposition

$$\pi_{A,B}(R) \bowtie \pi_{A,C}(R) = R$$

Eksemplet fra før:

$\pi_{A,B}(R)$:

A	B
1	kold
1	varm
2	lunken
2	hundekoldt

$\pi_{A,C}(R)$:

A	C
1	vanille
1	banan
1	citron
2	pistacie
2	chokolade

Generelle tilfælde: Dekomposition skal måske gentages.

Fjerde normalform, 4NF

"Ingen redundans pga. interessant multiværdiafhængighed"

Definition: En relation R er i 4NF såfremt for enhver ikke-triviel multiværdiafhængighed $A \twoheadrightarrow B$, at A er en supernøgle for R.

Eksempel: Antag $R(A, B, C)$ i 4NF og $A \twoheadrightarrow B$.

- A indeholder en nøgle.
- Dvs. given værdi a for A bestemmer højst én tupel $\langle a, b, c \rangle \in R$.
- Dvs. $B_{A=a}$ indeholder nul eller et element.
- Dvs. $A \twoheadrightarrow B$ er blot en $A \rightarrow B$.
- og som ved BCNF, ingen redundans i den anledning; dekomposition uinteressant.

Ikke-triviel? Ingen overlap mellem A og B; der er attr. udenfor A og B.

Opsummering: *Lyn*kursus i design af relationsskemaer, funktionelle (m.v.) afhængigheder, normalformer, normaliseringsprocedurer, osv.

- Formål: Undgå redundans og hvad deraf følger
  Andre kvaliteter: Statiske relationer,
  hensyn til check af funktionelle afhængigheder,

- Normalformer: 3NF $\Leftarrow$ BCNF $\Leftarrow$ 4NF
  (betegnelser historiske; se i bogen om 1NF, 2NF; søg i litteratur for 5NF).

- Bringe relation over i normalform ved dekomposition $R_1 \bowtie R_2 = R$; evt. $R_1^+ \bowtie R_2 = R$

- Egenskaber ved relationer: Funktionelle afhængigheder, nøgler, flerværdiafhængigheder