Normálne formy na základe
funkčných a spojovacích závislostí

a) 2NF, 3NF, HBCNF
b) Spojovacia závislosť
c) ETNF

Normalizácia je proces, ktorý vedie k návrhu, v ktorom sa minimalizuje **redundancia, nadbytočnosť** v databázovej schéme. Nedávno bola definovaná [1] nová normálna forma **ETNF** - *essential tuple normal form*, v ktorej každá **n-tica** je **nevyhnutná** (essential tuple). Vďaka nej, na zabezpečenie n-tíc s minimálnou redundanciou nie je potrebná 4NF a 5NF: ETNF je slabšia ako 5NF ale v eliminácii nadbytočných n-tíc je rovnako efektívna.

Najprv na základe funkčných závislostí definujeme druhú, tretiu a Heath-Boyce-Codd - ovu normálnu formu, ktoré ilustrujeme príkladmi. Potom po definícii **spojovacej závislosti** sa pozrieme podrobnejšie na vlastnosti HBCNF a na záver zadefinujeme **tri typy n-tíc** a ETNF.

a) 2NF, 3NF, HBCNF

Uvažujme relačnú schému $R \equiv R(\mathscr{A}, \mathscr{F})$, $\mathscr{A} = \{A_1,...,A_n\}$. Vieme, že A_i je kľúčový atribút, ak je prvkom (kompozitného) kandidátneho kľúča. Jeho zovšeobecnením je nadkľúč.

Pripomíname, že množina atribútov X je **podkľúč (**PK) / **nadkľúč (**NK), ak X je podmnožinou / nadmnožinou nejakého, aspoň jedného kandidátneho kľúča, $X \subseteq KK_i$ / $KK_i \subseteq X$.

Poznamenáme, že

- každý kľúč je NK, ale naopak to neplatí (napr. NK={W,K}, kde $W \not\subset \mathscr{A}$);
- $\mathscr{A}$ je NK;
- nulová množina je PK pre ľubovoľnú relačnú schému *R*.

Definícia 1

Uvažujme relačnú schému $R(\mathscr{A}, \mathscr{F})$. Ak pre **ľubovoľnú** (aj implicitnú) **netriviálnu** funkčnú závislosť

$$X \rightarrow Y,$$

vyplývajúcu z $\mathscr{F}$, kde $X \subset \mathscr{A}$, $Y \subset \mathscr{A}$ platí

jedna vlastnosť z troch	**jedna** vlastnosť z dvoch	že
- X je nadkľúč - NK, - Y je podkľúč – $Y \subseteq KK_i$, - X nie je podkľúč – $X \not\subseteq KK_j$, potom a len potom je *R* v **2NF**.	- X je nadkľúč, - Y je podkľúč – $Y \subseteq KK_i$, potom a len potom je *R* v **3NF**.	- X je nadkľúč, potom a len potom je *R* v **HBCNF**.

Môže sa v danej definícii X rovnať $\mathscr{A}$, teda X$\subseteq \mathscr{A}$?

Poznamenáme, že z daných normálnych foriem pochopiteľne 2NF je najslabšia, lebo pre jej splnenie stačí platnosť jednej vlastnosti z troch.

a) Príklady.

Uvažujme štyri príklady s tromi atribútmi (v DB teórii AB je množina zjednotení atribútov A, B)

1) $R = R(\mathscr{A}, \mathscr{F})$, kde $\mathscr{A}$ ={A,B,C} a $\mathscr{F}$ ={A→B};
2) $R = R(\mathscr{A}, \mathscr{F})$, kde $\mathscr{A}$ ={A,B,C} a $\mathscr{F}$ ={A→B, B→C};
3) $R = R(\mathscr{A}, \mathscr{F})$, kde $\mathscr{A}$ ={A,B,C} a $\mathscr{F}$ ={AC→B, B→C};
4) $R = R(\mathscr{A}, \mathscr{F})$, kde $\mathscr{A}$ ={A,B,C} a $\mathscr{F}$ ={A→BC }.

Dokážme, že

- prvý príklad nie je v 2N;
- druhý príklad je v 2NF, ale nie je v 3NF;
- tretí príklad je v 3NF, ale nie je v HBCNF;
- štvrtý príklad je v HBCNF.

jedna vlastnosť z troch	jedna vlastnosť z dvoch	že
- X je nadkľúč, - Y je podkľúč – Y $\subseteq$ KK$_i$, - X nie je podkľúč – X $\not\subseteq$ KK$_j$, potom *R* je v **2NF**.	- X je nadkľúč, - Y je podkľúč – Y $\subseteq$ KK$_i$, potom R je v **3NF**.	- X je nadkľúč, potom R je v **HBCNF**.

Dôkaz:
NO - aspoň pre jednu FZ neplati ani jedna vlastnosť; OK - pre všetky FZi platí aspoň jedna vl.

- 2NF NO: z FZi A→B atr. A nie je NK; $\{AC\}^+ = \mathscr{A}$ ale pre FZ A→B neplatí B⊆AC ani A⊈AC (pretože A⊆AC, teda A je podkľúč, neplatí A⊈AC);

- 2NF OK: $\{A\}^+ = \mathscr{A}$ a pre A→B platí A je NK a pre B→C platí B ⊈ A; 3NF NO: pre B→C atr. B nie je NK ani C⊆A;

- 3NF OK: $\{AC\}^+ = \mathscr{A}$ a pre AC→B AC je NK a pre B→C C⊂AC; HBCNF NO: B nie je NK;

- HBCNF OK: $\{A\}^+ = \mathscr{A}$.

b) Príklad. Na cvičení.
Nech $R \equiv R(\mathscr{A}, \mathscr{F})$, kde $\mathscr{A}$ ={A,B,C,D}, $\mathscr{F}$ ={ACD → B, AC→ D, AC→B, D→ C}.
2a) Skontrolujte, že $\{AC\}^+ = \mathscr{A}$ a preto AC je KK. Existuje ďalší dvojatribútový KK?
2b) Bude *R* v HBCNF?
2c) Bude *R* v 3NF?

Zdôrazňujeme, že HBCNF musí platiť pre ľubovoľnú explicitnú alebo implicitnú FZ X→Y v *R* a $\mathscr{F}$ obsahuje explicitné FZi. Nižšie uvidíme, že závislosti je možné zaviesť aj cez iné obmedzenia, konštrukcie, ako napr. SZi $\mathscr{J}$, z ktorých potom sa dajú vydedukovať implicitné FZi. Preto formuláciu "FZ vyplývajúca z $\mathscr{F}$ " treba rozšíriť na "FZ vyplývajúca z $\mathscr{F}$ a $\mathscr{J}$ " .

b) Spojovacia závislosť

V databázovej teórii okrem

- **funkčnej závislosti FZ** sa rozlišuje aj
- viachodnotová závislosť (Multy valued dependency MVD) a
- **spojovacia závislosť SZ** (Join dependency JD).

Definícia 4NF a 5NF sa zakladajú práve na MVD a SZ [1]. Nižšie po definícii SZi uvedieme iba definíciu HBCNF a ETNF, z ktorých posledná sa explicitne neopiera o SZ, definíciu 4NM a 5NM nie a ukážeme, ako sú HBCNF a ETNF prepojené.

Tu iba poznamenáme, že z ETNF vyplýva 4NF. Preto definíciu MVD, potrebnú pre 4NF, neuvádzame (navyše MVD je možné vyjadriť aj pomocou SZi).

Relačná schéma *R* je trojica, pozostávajúca z konečnej množiny atribútov a množiny obmedzení, ktoré špecifikujú *R* pomocou množiny FZí $\mathscr{F}$ a SZi $\mathscr{J}$

$$R(\mathscr{A},\mathscr{F},\mathscr{J}).$$

Relácia r je inštanciou relačnej schémy R ak

- $\mathscr{A}$ je množina atribútov **r** a
- **r** vyhovuje obmedzeniam, zadaných pomocou závislostí $\mathscr{F}$ a $\mathscr{J}$.

Je dobre známe, že množina FZí môže mať za následok (imply) inú FZ. To isté platí aj pre SZi. Hovoríme, že **implicitné** FZi a SZi **logicky** vyplývajú z **explicitne** zadaných FZí a SZí.

Príklad 1: Uvažujme $R = R(\mathscr{A},\mathscr{F})$, kde $\mathscr{F}$ ={X→Y, Y→Z}. Potom X→Y, Y→Z sú explicitné FZi a X→Z je implicitná FZ.

Definícia 2. Nech $S_1 \cup \ldots \cup S_k = \mathscr{A}$ **.** Hovoríme, že relácia **r** ako inštancia R *splňa* **spojovaciu závislosť** (SZ)

$$\bowtie\{S_1,\ldots,S_k\},$$

ak vždy, keď $t_1,\ldots,t_k$ sú ľubovoľné n-tice relácie **r**, a existuje n-tica t (nie nutne z **r**) taká, že $t[S_i]= t_i[S_i]$ pre každý *komponent* S_i, potom t **je n-tica r**.

Poznamenáme, že namiesto $\bowtie\{S_1,\ldots,S_k\}$ sa používa aj označenie $*\{S_1,\ldots,S_k\}$.

Názov (*k-komponentná*) *spojovacia závislosť* vyjadruje skutočnosť, že

relácia **r** vyhovuje, podlieha **SZi** $\bowtie\{S_1,\ldots,S_k\}$ vtedy a len vtedy, keď **r** je **bezstratové spojenie** (lossless join) jej **projekcií** **r**$[S_1]$, ..., **r**$[S_k]$,

$$\mathbf{r} \text{ vyhovuje SZi } \bowtie\{S_1,\ldots,S_k\} \quad \Leftrightarrow \quad \mathbf{r} = \bowtie_{i=1}^{k} \mathbf{r}[S_i]$$

kde $\mathbf{r} = \bowtie_{i=1}^{k} \mathbf{r}[S_i]$ znamená, že r možno znovu vytvoriť spojením viacerých (menších) relácií, z ktorých každá má podmnožinu atribútov r, teda spojením projekcií **r**$[S_1]$, ..., **r**$[S_k]$.

Poznamenáme, že medzi k komponentami SZi musí existovať k-1 spoločných atribútov, potrebných na k-1 spojení.

SZ sa vyjadruje aj slovami: **r** sa **rozkladá bezstratovo** na jej **projekcie** (decomposes losslessly, nonloss decomposition).

Triviálna SZ (nad $\mathscr{A}$) je taká závislosť, ktorá platí pre každú reláciu s atribútmi $\mathscr{A}$. Ľahko nahliadneme, že SZ je triviálna vtedy a len vtedy, keď niektorý jej komponent je množinou všetkých atribútov.

Príklad 2: Uvažujme $R\{\mathscr{A}, \mathscr{J}\}$ s

$$\mathscr{A} = \{D, S, P\}, \quad \mathscr{J} = \bowtie\{DS, SP, PD\},$$

kde D-dodávateľ, S-súčiastka, P-projekt.

Spojovacia závislosť $\mathscr{J}$ znamená:

- ak dodávateľ d dodáva súčiastku s – DS (pre niektorý projekt p),
- a súčiastka s je dodávaná do projektu p – SP (niektorým dodávateľom d),
- a projekt p je zásobovaný dodávateľom d – PD (niektorou súčiastkou s),

potom dodávateľ d zásobuje súčiastkou s projekt p
$\Leftrightarrow$ dodávateľ d dodáva súčiastku s do projektu p.

Príklad 3. Uvažujme relačnú schému $R(\mathscr{A}, \mathscr{F}, \mathscr{J})$, kde

$$\mathscr{A} = \{A,B,C\}, \ \mathscr{F} = \{AB \to C\} \text{ a } \mathscr{J} = \bowtie\{AB, AC\}.$$

Dokážme, že $A \to C$ je (implicitná) FZ.

Nech **r** je inštancia *R*. Predpokladajme, že dve n-tice

$$\begin{cases} t_1 = [a, b_1, c_1], \\ t_2 = [a, b_2, c_2], \end{cases}$$

sú z **r**. Nech AB→C je FZ. Pochopiteľne iba z nej nevyplýva A→C, ako to dokazujú dané dve n-tice. Uvažujme tretiu n-ticu

$$t = [a, b_1, c_2],$$

ktorá samozrejme nemusí byť z $\boldsymbol{r}$ (prečo?). Ale zo SZ-ti ⋈{AB, AC} dostaneme, ako to čoskoro ukážeme, že t musí byť z $\boldsymbol{r}$, čo sa kvôli AB→C môže stať iba ak $c_1 = c_2$ a preto A→C.

Teda ostalo nám ukázať, že $t \in \boldsymbol{r}$. Ak ⋈{AB, AC} je SZ, kde AB ∪ AC = {A,B,C} = $\mathscr{A}$, potom podľa definície SZi z dvoch rovností $t[a, b_1] = t_1[a, b_1]$ a $t[a, c_2] = t_2[a, c_2]$, teda z $t(AB) = t_1(AB)$ a $t(AC) = t_2(AC)$ dostaneme, že $t \in \boldsymbol{r}$.

Veta 1 (Heath). Nech X a Y sú podmnožiny $\mathscr{A}$ a nech Z= $\mathscr{A}$ \(X∪Y). Potom nasledujúce tvrdenia sú ekvivalentné

- X→Y,
- platí SZ ⋈{XY, XZ} a zároveň XZ je nadkľúč, tzn. XZ→Y.

Poznámka:

- pôvodne Heath dokázal iba to, že z FZi X→Y vyplýva SZ ⋈{XY, XZ};
- ako sme to vyššie poznamenali, SZ ⋈{XY, XZ} (s XY∪XZ= $\mathscr{A}$) znamená bezstratové spojenie

  r[XY] ⋈ r[XZ] = r.

Definícia 3. Relačná schéma R je v Heath-Boyce-Codd normálnej forme (**HBCNF**), ak **každá** explicitná alebo implicitná FZ v R logicky vyplýva z kľúčov R.

Veta 2. Relačná schéma R je v HBCNF vtedy a len vtedy, ak pre každú explicitnú alebo implicitnú netriviálnu FZ X→Y v R determinant X je nutne nadkľúč.

Pozri definíciu 1, teda: Def.1 ⇔ Def.3

Príklad 4 (pokračovanie). Ukážme, že v príklade 3 *R* nie je v HBCNF.

Z poslednej vety (či z Def.1) vidíme, že k určeniu, či *R* je v HBCNF, musíme uvažovať aj implicitné FZi. V príklade 3 sme ukázali, že A→C je netriviálna implicitná FZ. Otázka znie, bude A (nad)kľúč? Nebude, lebo A→B neplatí.

Poznamenáme, že v tomto príklade implicitná FZ A→C zohráva dôležitú úlohu. Keby sme nekorektne uvažovali iba explicitné FZi, teda iba AB→C, potom by sme došli nesprávnemu záveru, že *R* je v HBCNF.

Príklad 5. Uvažujme relačnú schému $R(\mathscr{A}, \mathscr{F}, \mathscr{J})$, kde

$\mathscr{A}$ ={A,B,C,D}, $\mathscr{F}$ ={ A→BCD, BC→AD } a $\mathscr{J}$ ={⋈{ABC, CD, BD}}.

Ukážeme pomocou poslednej vety, že R je v HBCNF. Nech X→Y je ľubovoľná netriviálna FZ, kde X je ľubovoľná podmnožina $\mathscr{A}$. Máme ukázať, že determinant X je nadkľúč. Pretože dve FZi $\mathscr{F}$ s determinantami A a BC sú (nad)kľúče, stačí ukázať, že **X obsahuje A alebo BC**.

Stačí rozlišovať štyri prípady: X obsahuje jeden, dva alebo tri atribúty (4 prečo nie?), alebo X je prázdna množina. Ukážeme iba prípad, keď X má tri atribúty. V tomto prípade ľahko nahliadneme, že potom X buď obsahuje A, alebo je BCD. Potom ale X obsahuje buď A alebo BC, ktoré sú kľúče (v ďalších troch prípadoch dôjdeme k rovnakému záveru). Preto X je nadkľúč.

Príklad 6. Dokážte pr. 5 pre prípad, keď X má dva atribúty. Na cvičení.

c) ETNF

Nielen **FZ** môže logicky vyplývať z iných závislostí, ale aj **n-tica** z iných n-tíc. Nech R je relačná schéma, S je množina n-tíc a t je n-tica. Hovoríme, že t **logicky** vyplýva **z S** (vzhľadom na R), ak každá inštancia R, ktorá obsahuje všetky n-tice S, obsahuje aj t.

Teraz uvedieme tri typy n-tíc z hľadiska nadbytočnosti, z ktorých sa v definícii ETNF používa posledný.

Definícia. Nech je R($\mathscr{A}$) relačná schéma, **r** relácia ako inštancia R a t n-tica z **r**. Potom

- n-tica t z r je **čiastočne redundantná/nadbytočná** v **r** (vzhľadom na R) ak
  a) existuje n-tica t***** z **r** taká, že t ≠ t*****,
  b) existuje X ⊆ $\mathscr{A}$, že t[X] = t*****[X] a zároveň
  c) existuje netriviálna FZ X→A, kde A∈ $\mathscr{A}$ (a A ⊈ X).

  (Dôvod, prečo sa hovorí, že n-tica t je **čiastočne** redundantná, je nasledujúci. Pretože t[X] = t*****[X] a v relácii **r** platí FZ X→A, dostaneme t[A] = t*****[A]. Intuitívne teda informácia o tom, že ktoré hodnoty A sú spojené s ktorými hodnotami X, je daná aj s t a t***** a nielen s FZ X→A.)

  Veta 3. Nech R je ľubovoľná schéma. Potom R je v HBCNF vtedy a len vtedy, ak žiadna inštancia R nemá čiastočne redundantnú n-ticu.

- n-tica t z r je **plne redundantná** v **r** (vzhľadom na R), ak
  existuje množina S n-tíc z **r** taká, že
  - t∉S a
  - t logicky vyplýva z S vzhľadom na R.

  (**Plne** odkazuje na skutočnosť, že **celá** n-tica je **nadbytočná**.
  Intuitívne n-tica t je plne nadbytočná, ak jej prítomnosť v r logicky vyplýva z ďalších existujúcich n-tic z r.)

  Poznamenáme, že kým čiastočná redundancia súvisí s HBCNF, plná redundancia súvisí s anomáliami vymazávania, pozri [1].

- n-tica je **nevyhnutná** (essential), ak **nie je** ani čiastočne ani plne redundantná.

Príklad 7. Uvažujme príklad 2, teda schému $R\{\mathscr{A}, \mathscr{J}\}$ s

$$\mathscr{A} = \{D, S, P\}, \quad \mathscr{J} = \bowtie\{DS, SP, PD\}$$

a inštanciu r relačnej schémy R. Ukážme, že n-tica (d, s, p) z r je **plne redundantná**, teda informácia o (d, s, p) je v r dvakrát.

Skutočne. Nech relácia r je inštanciou triedy R (takže r má atribúty D, S a P a r spĺňa SZ $\mathscr{J}$). Predpokladajme, že r má n-tice (d, s, p'), (d', s, p) a (d, s', p) a možno aj iné n-tice a že $d \neq d'$, $s \neq s'$ a $p \neq p'$. Zo SZi $\mathscr{J} = \bowtie\{DS, SP, PD\}$ vyplýva ($t[DS]=t_1[DS]$, …), že r, ako inštancia R, musí obsahovať aj n-ticu (d,s,p), ktorá nie je prvkom S={(d, s, p'), (d', s, p), (d, s', p)}. Preto v súlade s definíciou (d, s, p) je plne redundantná. Takže v relácii r informácia o n-tici (d, s, p) je zastúpená dvakrát:

- po prvé, explicitne, pomocou samotnej n-tice (d, s, p) z r a
- po druhé, implicitne, pomocou n-tíc (d, s, p'), (d', s, p) a (d, s', p) a spojovacej závislosti $\mathscr{J}$.

Príklad 8 (rozšírenie príkladu 2 o $\mathscr{F}$): Uvažujte $R\{\mathscr{A}, \mathscr{F}, \mathscr{J}\}$ s

$$\mathscr{A} = \{D, S, P\}, \ \mathscr{F} = \{ DS \rightarrow P \} \ \text{a} \ \mathscr{J} = \bowtie\{DS, SP, PD\},$$

kde D-dodávateľ, S-súčiastka, P-projekt. Ukážte, že

a) žiadna inštancia R nemá **čiastočne redundantnú** n-ticu
b) R je v HBCNF.

Na cvičení:

Definícia. Relačná schéma R je v normálnej forme nevyhnutných n-tíc (essential tuple normal form - **ETNF**), ak každá n-tica ľubovoľnej inštancie R je nevyhnutná.

Z nasledujúcich dvoch viet, ktoré spájajú HBCNF a ETNF, sa prvá opiera o komponent SZi a druhá o kľúč.

Veta 4 ⇔. Nech relačná schéma R je špecifikovaná iba s FZ-mi a SZ-mi. Potom R je v ETNF vtedy a len vtedy, ak je v HBCNF a niektorý **komponent** každej explicitnej SZ R je **nadkľúč**.

Veta 5 ⇒. Nech relačná schéma R je špecifikovaná iba s FZ-mi a SZ-mi a je v HBCNF. Ak niektorý **kľúč** sa skladá iba z **jedného** atribútu, potom R je v ETNF.

Príklad 9 (pokračovanie príkladu 5).

$R = R(\mathscr{A}, \mathscr{F}, \mathscr{J})$: $\mathscr{A}$ ={A,B,C,D}, $\mathscr{F}$ ={ A→BCD, BC→AD } a $\mathscr{J}$ ={⋈{ABC, CD, BD}}.

Ukážme, že R je v ETNF.

Pretože na základe príkladu 5 R je v HBCNF, vďaka poslednej vety 5 R bude aj v ETNF, lebo z dvoch nadkľúčov A a BC, prvý sa skladá iba z jedného atribútu.

Príklad 10 Uvažujme pr. 8. Ak vieme dodatočne, že žiadna inštancia R nemá plne redundantnú n-ticu, ukážte, že R je v ETNF. Na cvičení.

[1] Darwen, Date, Fagin, A normal form for preventing redundant tuples in relational databases, 2012

Chriss Date

A Relational Model - set theory and logic.

"**Learn the relational model first, SQL second.**"

1a. Learn the relational model, by reading the right books and/or attending the right courses.

2. Go out and get your hands dirty working on a real project for a year or three.

1b. Come back and read those books and/or attend those courses again.

Finally, read either or both of Ted Codd's first two papers every year.

- "Derivability, Redundancy, and Consistency of Relations Stored in Large **Data Banks**", IBM Research Report RJ599, August 19th, 1969, reprinted in ACM SIGMOD Record 38, No. 1 (March 2009)

- "A **Relational Model** of Data for Large Shared Data Banks", CACM 13, No. 6, June 1970,

Poznámka. Zo SZi ⋈ môže vyplývať aj FZ aj redundantnosť n-tice

SZ ⋈ sa požíva v prikladoch 3. a 7. na logický dôkaz, že

a) teoretická n-tica $t = [a, b1, c2]$

b) konkrétna n-tica t = [d, s, p] z r

musí byť v r.

Preto ...

a) $c1 = c2$ a A→C

b) t je plne redundantná.