10. Бази от данни. Нормални форми.

Анотация: Проектиране на схемите на релационните бази от данни. Аномалии, ограничения, ключове. Нормални форми. Функционални зависимости, аксиоми на Армстронг. Първа, втора, трета нормална форма, нормална форма на Бойс-Код. Многозначни зависимости; аксиоми на функционалните и многозначните зависимости; съединение без загуба; четвърта нормална форма.
Примерни задачи: Привеждане на схема на базата от данни (при зададени функционални зависимости) към зададена нормална форма.

1. Проектиране на схемите на релационните бази от данни

Транслацията на даннов модел (например същност-връзка) на база на данни до съвкупност от релации, над чиито атрибути са дефинирани зависимости, се нарича проектиране на схема на релационна база от данни. Важни цели на този процес са да запази цялостта на данните, както и да нормализира получените релации.

Нормализацията на релационна база от данни е процесът на реструктуриране на релациите и зависимостите на техните атрибути в нормални форми, които целят **минимизация на дублирането на данните**, както и запазването на тяхната **цялост**. Възможни начини за извърпването ѝ са синтез, т.е. създаване на нов релационен дизайн, и декомпозиция на вече съществуващ такъв (по-често използван).

2. Функционални зависимости

Дефиниция (ФЗ). Нека $R(A_1, \ldots, A_n)$, $n \in \mathbb{N}$ е релация. Функционална зависимост (ФЗ) за релацията R наричаме твърдение от вида $A_i \to A_j$, където A_i, A_j са множества атрибути на R, при което $\forall t, u \in R$ е изпълнено, че ако t и u съвпадат по атрибутите A_i на R, то те съвпадат и по атрибутите A_j на R. Четем A_i еднозначно определя A_j.

Правило за разделяне. Може да разделяме атрибутите от дясната страна на ФЗ, така че само един атрибут да се появява от дясно за всяка ФЗ. Например ФЗ $A_1, A_2 \to B_1, B_2$ може да се раздели на $A_1, A_2 \to B_2$ и $A_1, A_2 \to B_2$.

Правило за комбиниране. Може да комбинираме множество от ФЗ с едни и същи леви страни до една ФЗ с комбинирани десни страни на първоначалните ФЗ. Например ФЗ $A_1, A_2 \to B_1$ и $A_1, A_2 \to B_2$ можем да комбинираме в $A_1, A_2 \to B_1, B_2$.

Забележка. правилата за разделяне и комбиниране са приложими само за десните страни на ФЗ.

Дефиниция (Тривиална ФЗ). Нека $A_1, \ldots, A_n \to B_1, \ldots, B_m$, $n, m \in \mathbb{N}$ е ФЗ. Тогава казваме, че тя е тривиална ФЗ ако $\{B_1, \ldots, B_m\} \subseteq \{A_1, \ldots, A_n\}$. Всяка ФЗ, която не е тривиална се нарича нетривиална ФЗ.

3. Ключове

Дефиниция (Ключ). Множество от един или повече атрибути $\{A_1, \ldots, A_n\}$ на дадена релация R наричаме ключ на R ако:

1. $\{A_1, \ldots, A_n\}$ функционално определя всички останали атрибути на R.
2. $\{A_1, \ldots, A_n\}$ е минимално по включване, т.е. всяко друго подмножество, което определя функционално всички останали атрибути на R ще съдъжа в себе си $\{A_1, \ldots, A_n\}$.

Пример. Нека $R(A_1, A_2, B_1, B_2)$ е релация и $A_1, A_2 \to B_1$ и $A_1, A_2 \to B_2$ са ФЗ. Тогава имаме, че $A_1, A_2 \to A_1, A_2, B_1, B_2$ и следователно $\{A_1, A_2\}$ е ключ на R.

Забележка. За дадена релация може да съществуват повече от един ключ. Те се наричат кандидат ключове. За всяка релация се избира един ключ, който се нарича **първичен ключ**.

Дефиниция (Суперключ). Нека R е релация и K е ключ на R. Множеството от атрибути $A : K \subseteq A$ се нарича **суперключ** на R.

4. Покритие на ФЗ

От множеството от ФЗ, които са в сила в релация R, може да определим кандидат ключовете за дадена релация. Това се прави чрез намирането на покритието на всички атрибути и комбинация от атрибути за релацията R.

Дефиниция (Покритие). Нека $R(A_1, \ldots, A_n), n \in \mathbb{N}$ е релация, S е множество от функционални зависимости в сила за R и $T = \{B_1, \ldots, B_m\} \subseteq \{A_1, \ldots, A_n\}$. Покритие на множеството от атрибути T наричаме множеството $T^+ \subseteq \{A_1, \ldots, A_n\}$, такова че $\forall A_i \in T^+ \exists$ ФЗ $B_1, \ldots, B_m \to A_i$, която може да бъде изведена от S.

Алгоритъм за намиране на покритие на множество от атрибути $T = \{B_1, \ldots, B_m\}$ за релацията $R(A_1, \ldots, A_n)$ и множество от ФЗ S:

1. Нека $T^+ := T$;
2. $\forall$ ФЗ от вида $C_1, \ldots, C_k \to D_1, \ldots, D_p$ такива че $\{C_1, \ldots, C_k\} \subseteq T^+$ и $\{D_1, \ldots, D_p\} \nsubseteq T^+$ актуализираме $T^+ := T^+ \cup \{D_1, \ldots, D_p\}$;
3. Повтаряме стъпка 2 до момента, в който вече не може да добавяме нови атрибути в T^+. Понеже множеството T^+ може само да нараства, а броят на атрибутите е краен, то след краен брой итерации ще стигнем до момент, в който няма да може да добавим повече атрибути в T^+ и следователно алгоритъмът ще приключи.
4. Множеството T^+ е търсеното покритие.

4.1. Аксиоми на Армстронг

Аксиомите на Армстронг са правила, по които може да генерираме нови ФЗ, които също са правила в съответната релация. Нека R е релация и T е множество от атрибути в R. Тогава основните аксиоми на Армстронг са :

- **Рефлексивност**. Ако $U \subseteq T$, то $T \to U$;
- **Умножение**. Ако A е атрибут в R, $U \subseteq R$ и $T \to U$, то $T \cup \{A\} \to U \cup \{A\}$;
- **Транзитивност**. Ако $U, P \subseteq R$, T и $U \to P$ следва, че $T \to P$.

5. Ограничения

Дефиниция (Ограничения). Ограниченията са механизъм за налагане на цялост върху данните чрез дефиниране на правила, на които трябва да отговарят стойностите на конкретни атрибути. Накратко не позволяват въвеждането на невалидни данни. Основните ограничения са:

- Уникалност на множество от атрибути – два кортежа не могат да имат едни и същи стойности на атрибутите от множеството. По дефиниция първичните ключове са

уникални. В SQL обичайно се имплементира чре ключовите думи **UNIQUE** и **PRIMARY_KEY**.
- Задаване на домейн от възможни стойности на множество от атрибути – имплементират се чрез предикати. В SQL обичайно се имплементира посредством ключовите думи **CHECK** и **NOT_NULL**.
- Външен ключ – представлява множество от атрибути, такова, че за всяка ненулева стойност на атрибут, който е външен ключ, трябва да съществува ключов атрибут от друга релация, който да съдържа същата стойност. В SQL обичайно се имплементира чрез ключовата дума **FOREIGN_KEY**.

6. Аномалии

При проектиране на релационни схеми, могат да бъдат получени следните аномалии:

- **Излишества** – повторение на информация без това да бъде необходимо;
- **Аномалии при обновяване** – получава се когато при промяна на данни в един кортеж не се обновяват свързаните с него кортежи и може да се стигне до неконсистентна информация в следствие на прекъсната заявка;
- **Аномалии при добавяне** – получава се когато се опитваме да добавим информация, но тя не може да бъде добавена освен по половинчат начин. Случва се при релация с транзитивни ФЗ.
- **Аномалии при изтриване** – получава се когато се опитаме да изтрием информация, но това не може да се случи без премахването на друга важна информация. Случва се при релации с транзитивни ФЗ.

6.1. Избягване на излишествата

Начин да избегнем излишествата е чрез декомпозиране на релацията. Декомпозирането е равносилно на разделяне на атрибутите ѝ в няколко нови релации.

Пример. Нека $R(A_1, A_2, B_1, B_2)$ е релация. Тогава може да декомпозираме релацията R до две релации $R_1(A_1, A_2)$ и $R_2(B_1, B_2)$, такива че:

- множеството от атрибутите на R, да бъде равно на обединението от множествата на атрибутите на R_1 и R_2,
- кортежите в R_1 са проекция по атрибутите A_1 и A_2 на релацията R,
- кортежите в R_2 са проекция по атрибутите B_1 и B_2 на релацията R,

където проекция по атрибути означава, че в новополучените релации взимаме само стойностите на кортежите на R в тези атрибути. Интересно е, че при проекция може да се получат два или повече еднакви кортежи от R. Тогава щом R_1 и R_2 са релации, т.е. множества, значи остава само един от тези кортежи и по този начин се премахва дубликацията на информация.

7. Нормални форми

Първа нормална форма (1НФ). Казваме, че релацията R се намира в 1НФ, когато всички компоненти в кортежите на R имат атомарен домейн.

Втора нормална форма (2НФ). Казваме, че релацията R се намира във 2НФ, когато тя се намира в 1НФ и всеки непървичен атрибут е функционално зависим от атрибутите на първичния ключ, но не и от негово подмножество.

Трета нормална форма (3НФ). Казваме, че релацията R се намира в 3НФ, когато тя се намира във 2НФ и за всяка нетривиална ФЗ, която е в сила за R, или лявата част е

суперключ за R, или всеки атрибут от дясната част е част от някой ключ на R – **не** задължително същия. Казано по прост начин – няма транзитивни ФЗ.

Нормална форма на Бойс-Код (НФБК). Казваме, че една релация R се намира в НФБК, когато тя се намира в 3НФ и за всяка нетривиална ФЗ, която е в сила в R е изпълнено, че лявата ѝ част е суперключ на R.

Твърдение. Ако една релация R се намира в НФБК, то в нея може да възникнат аномалии свързани с ФЗ.

8. Съединение без загуба

Дефиниция (Съединение без загуба). Нека $R(A_1, \ldots, A_n), n \in \mathbb{N}$ е релация и R се декомпозира на релациите $R_1, \ldots, R_m, m \in \mathbb{N}$. Казваме, че декомпозицията е със съединение без загуба, ако $R = R_1 \bowtie \ldots \bowtie R_m$, където $\bowtie$ е релацията на естествено съединение (natural join) от релационната алгебра.

Твърдение. Декомпозицията на релацията $R(A_1, \ldots, A_n)$ на две релации $S(B_1, \ldots, B_m)$ и $T(C_1, \ldots, C_k)$ за $n, m, k \in \mathbb{N}$ е съединение без загуба $\Leftrightarrow$ за R е изпълнена поне една от следните функционални зависимости $\{B_1, \ldots, B_m\} \cap \{C_1, \ldots, C_k\} \to \{B_1, \ldots, B_m\}$ или $\{B_1, \ldots, B_m\} \cap \{C_1, \ldots, C_k\} \to \{C_1, \ldots, C_k\}$.

Алтернативна дефиниция (Съединение без загуба). Нека $R(A_1, \ldots, A_n), n \in \mathbb{N}$ е релация, за която множеството от ФЗ е F. Нека R се декомпозира на релациите $R_1, \ldots, R_m, m \in \mathbb{N}$, чиито множества от ФЗ са съответно $F_1, \ldots, F_m$, като те са съответно проекции на F върху $R_1, \ldots, R_m$. Казваме, че декомпозицията е със съединение без загуба на функционалните зависимости когато $F_1 \cup \ldots \cup F_m = F$.

9. Алгоритъм за свеждане на релация в НФБК (BCNF)

Нека R_0 е релация с множество от функционални зависимости S_0. Прилагаме следните стъпки, за да получим декомпозиция, всяка релация от която е в НФБК.

1. Задаваме $R \leftarrow R_0$, $S \leftarrow S_0$;
2. Проверяваме дали има нарушения на НФБК. Ако не, връшаме $\{R\}$;
3. Избираме една ФЗ от S, която нарушава НФБК – нека тя бъде от вида $X \to Y$. Намираме X^+. Създаваме две релации – $R_1(X^+)$ и $R_2\left(X \cup (R \backslash X^+)\right)$. С други думи, първата релация съдържа X и всичко, което е функционално определено от тях, а втората релация – X и всичко, което **не** е функционално определено от X;
4. Проектираме S в R_1 и R_2 – получаваме S_1 и S_2;
5. Рекурсивно прилагаме стъпка 2 за двойките (R_1, S_1), (R_2, S_2). Обединяваме получените множества релации.

За стъпка 3 използваме алгоритъма за затваряне на множество атрибути в оригиналната релация за всяко подмножество на декомпозираната релация. Започваме от най-малките подмножества и спираме при получаване на минимален базис.

Твърдение. Декомпозицията получена от алгоритъма за свеждане на релация от 3НФ до НФБК е съединение без загуба. Следователно в примера $R = R_1 \bowtie R_2$.

10. Многозначни зависимости

Дефиниция (Многозначни функционална зависимост – МФЗ). Нека R е релация с множество от атрибути $\{A_1, \ldots, A_n, B_1, \ldots, B_m, C_1, \ldots, C_k\}$, $n, m, k \in \mathbb{N}$. Казваме, че $A_1, \ldots, A_n$ определя многозначно $C_1, \ldots, C_k \Leftrightarrow$ ако за всеки два кортежа $U, T \in R$, съвпадащи по атрибутите $A_1, \ldots, A_m$, $\exists$ кортеж $V \in R$ такъв, че:

- Съвпада с кортежите U и T по атрибутите $A_1, \ldots, A_n$
- Съвпада с кортежа U по атрибутите $B_1, \ldots, B_m$
- Съвпада с кортежа T по атрибутите $C_1, \ldots, C_k$

Т.е. ако между два случайни кортежа с еднакви стойности по $A_1, \ldots, A_n$ разменим стойностите на атрибутите $C_1, \ldots, C_k$ ще получим същата релация. Означаваме като $A_1, \ldots, A_n \twoheadrightarrow C_1, \ldots, C_k$. Пример за релация, където $C \twoheadrightarrow B$ е следната:

A	B	C
a_1	b_2	c_3
a_1	b_3	c_2
a_1	b_2	c_2
a_3	b_2	c_1
a_3	b_2	c_3

Дефиниция (Тривиална МФЗ). МФЗ $A_1, \ldots, A_n \twoheadrightarrow B_1, \ldots, B_m$ в релацията R наричаме тривиална когато $\{B_1, \ldots, B_m\} \subseteq \{A_1, \ldots, A_n\}$.

Правило за транзитивност. Ако $A \twoheadrightarrow B$ и $B \twoheadrightarrow C$, то $A \twoheadrightarrow C$.

Нови правила:

- Всяка ФЗ е МФЗ, т.е. ако $A \rightarrow B$, то $A \twoheadrightarrow B$.
- **(Допълнение)** Ако $X \twoheadrightarrow Y$, то $X \twoheadrightarrow R \backslash (X \cup Y)$

Забележка. Правилото за разделяне и комбиниране не е в сила за МФЗ, т.е. ако $A \twoheadrightarrow B_1$, B_2, не значи, че $A \twoheadrightarrow B_1$ и $A \twoheadrightarrow B_2$ и обратното.

11. Четвърта нормална форма

Дефиниция (4НФ). Казваме, че ралацията R е в четвърта нормална форма (4НФ), ако тя е в НФБК и за всяка нетривиална МФЗ $A \twoheadrightarrow B$, която е в сила за релацията R и за която A и B са множества от атрибути на R, е изпълнено, че A е суперключ на релацията R.

Аномалиите, които се получават при МФЗ, могат да бъдат предотвратени чрез декомпозиция в 4НФ. В 4НФ нетривиалните МФЗ, които нарушават 4НФ, се елиминират подобно на нетривиалните ФЗ при НФБК. В резултат на това декомпозираната релация няма да има излишества.

Алгоритъмът за декомпозиране от НФБК в 4НФ е аналогичен на този за декомпозиране в НФБК, с разликата, че се декомпозира по МФЗ. Когато декомпозираме следвайки този алгоритъм е гарантирано, че декомпозицията е съединение без загуба.