10. Бази от данни. Нормални форми

→ Нормални форми

?Нормализация? на данните се нарича процеса на разделяне на данните в множество свързани помежду си таблици. Оптимизацията на проекта на базата от данни включва процеса на нормализация. Нормализацията на данните осигурява организиране на данните по такъв начин, че:

- Актуализирането на някой елемент от данните ще изисква в общия случай действия само на едно място
- Изтриването на определен ел. от данните няма да доведе до нежелана загуба на други данни.

Целта на нормализацията е да се декомпозират таблиците в по-малки таблици, дефинирани така, че да се избегне излишно дублиране на данни. След разделянето на данните в отделни по-малки и добре структурирани таблици, се дефинират релационни връзки м/у тях. За да се получи нормализация
1NF → 2NF → 3NF → Бойс-Код → 4NF.

→ Проектиране на схемите на релационните БД

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 БД. Важни цели на този процес са да запази цялостта на данните, както и да нормализира получените релации.

Това да е при нормалните форми

?Нормализацията? на релационна база от данни е процесът на преструктуриране 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 извършването ѝ са синтез, т.е. създаване на нов релационен дизайн, и декомпозиция на вече съществуващ такъв.

Теорията на релационното проектиране се състои от следните основни понятия:

- Таблици и уникалност
- Външни ключове и домейни
- Релационни връзки
- Нормализация на данните
- Правила за запазване на цялостността на данните.

→ Аномалии

Аномалии наричаме различни недостатъци на една релационна схема. Съществуват аномалии на излишество, обновяване, добавяне и отстраняване.

Пример:

Дадена е релационната схема Библиотека (Биб#, Адрес, Книга, Брой). Всяка библиотека е представена с номера (Биб#) и адреса си (Адрес), а всяка книга – с името си (Книга) и бр. екземпляри (Брой). Всяка библиотека е еднозначно определена от номера си, а една и съща книга (заглавие) може да е налична в няколко библиотеки. Така получаваме следните аномалии на рел. схема:

(1) Аномалия на излишество – за всяка книга от една и съща библиотека ще се повтаря адресът на библиотеката.

(2) Аномалия при обновяване – От (1) следва, че при промяна на адреса на библиотеката, той трябва да се промени за всички книги от нея. Така остава възможност в някой от кортежите да се изпусне промяната ⇒ противоречивост на данните в БД.

(3) Аномалия при добавяне – При създаване на нова библиотека адресът ѝ не може да бъде въведен, докато не се посочи поне една книга от нея.

(4) Аномалията при изтриването – Ако се наложи временно всички книгите от дадена библиотека да се преместят в друга, то всички кортежи за първата библиотека ще бъдат премахнати от релацията и така адресът ѝ ще бъде изгубен.

→ Ключове

Всяка релация е мн-во от неповтарящи се n-торки (кортежи). Следователно има ключ, който определя еднозначно кортежите. Възможно е една релация да има повече от един ключ.

Ако едно мн-во от атрибути A е ключ за релацията R, то всяко мн-во X, съдържащо A, също идентифицира еднозначно елементите на R. Затова към всеки ключ имаме изискване за минималност:

Ако K е ключ на релацията R, тогава:

① $\forall r_1 \in R, r_2 \in R$ е в сила $r_1[K] \neq r_2[K]$

② Не съществува подмножество на K, за което ① остава в сила.

Един от всички възможни ключове избираме за първичен ключ.

Външният ключ K^* за R е атрибут или списък от атр., който не е първичен ключ за R, но е такъв за някаква релация Q.

→ Ограничения

Ограниченията дефинират правила за стойностите, допустими за съответните атрибути, и представляват механизъм за налагане на цялост. Основните ограничения в релационния модел са следните: (Накратко не позволяват въвеждане на невалидни данни)

① UNIQUE и PRIMARY KEY – не допускат да се въведе стойност, дублираща съществуваща

② Стойностите на атрибутите от първичния ключ не могат да бъдат нулеви;

③ CHECK – не допуска да се въведе стойност, която не отг. на определени условия (Age > 10, length(EGN) = 10);

④ FOREIGN KEY – налага връзка м/у данните в две релации – за всяка ненулева стойност на атрибут ВК трябва да съществува ключов атрибут от др. релация с тази стойност

⑤ NOT NULL – не допуска стойността на атрибута да е NULL (т.е. да не са въведени данни)

→ Функционални зависимости

Деф.: Нека R е релация. Ако кои да е два кортежа имат едни и същи стойности за атрибутите $A_1 A_2 A_3 \dots A_n$ и от това следва, че имат една и съща стойност и за атр. B, то казваме, че $A_1, \dots A_n$ функционално определят B (или че B зависи функ. от $A_1, \dots A_n$) и записваме $A_1, \dots A_n \to B$. Ако $A_1 \dots A_n$ определят функционално атрибутите $B_1, \dots B_k$ записваме $A_1, \dots A_n \to B_1 \dots B_k$.

Деф.: Казваме, че мн-вото $K = \{A_1, A_2, \dots, A_n\}$ е ключ на релацията R, ако:

1. Мн-вото K функционално определя всички атр. на R
2. За нито едно подмножество на K не е изпълнено 1.

Ако K удовлетворява 1, но не и 2, то K е суперключ.

Видове ФЗ:

1. Тривиални – $A \to B$, ако $B \subseteq A$
2. Нетривиални – $A \to B$, ако $B \not\subseteq A$
3. Напълно нетривиални – $A \to B$, ако $A \cap B = \emptyset$

→ Аксиоми на Армстронг

Аксиомите на Армстронг са съвкупност от правила или аксиоми за извод, чрез които от дадено множество ФЗ се получават нови ФЗ.

1. Рефлексивност – ако $\{B_1, \dots B_m\} \subseteq \{A_1, A_2 \dots A_n\}$, то $A_1 A_2 \dots A_n \to B_1 B_2 \dots B_m$; ??

2. Разширяемост (augmentation) – ако $A_1 \dots A_n \to B_1 \dots B_m$, то $A_1 \dots A_n C_1 C_2 \dots C_k \to B_1 \dots B_m C_1 C_2 \dots C_k$ за всяко мн-во $\{C_1, C_2 \dots C_k\}$

3. Транзитивност – ако $A_1 A_2 \dots A_n \to B_1 B_2 \dots B_m$ и $B_1 B_2 \dots B_m \to C_1 C_2 \dots C_k$ то $A_1 A_2 \dots A_n \to C_1 C_2 \dots C_k$

→ Първа, Втора, трета нормална форма, нормална форма на Бойс-Код (НФБК)

Може в началото

<u>Нормализация</u> е процес на преобразуването на релационни схеми, при които се прилагат ограничения, елиминиращи нежелани св-ва. Нормализацията води до логически устойчива структура на данните, която е лесна за разбиране и проста за поддържане. Могат да се получат различни нива на нормализация, критериите, за които се наричат нормални форми.
(Критериите, определящи нивата на нормализация се наричат нормални форми)

→ Първа нормална форма (1NF)

Една релация е в 1NF, ако:

① Всички атрибути са с атомарна стойност (т.е. ако домейните им са атомарни (прости);

→ Втора нормална форма (2NF)

Една релация е ~~в~~ в 2NF, ако:

① R е в 1NF

② Всеки непървичен атрибут е функционално зависим от атрибутите на първичния ключ, но не и от негово подмножество.

алтернативно:

Дефинираме:

Казваме, че Y напълно зависи от X, ако $X \to Y$ и $\nexists Z \subset X$, $Z \to Y$, т.е. Y зависи от цялото мн-во X, но не и от отделни негови елементи.

② Всички неключови атрибути са напълно зависими от ключа на R

→ Трета нормална форма (3NF)

Една релация е в 3NF, ако и само ако:

① R е в 2NF

② За всяка нетривиална ФЗ в R или лявата страна е суперключ, или дясната е първичен (ключов) атрибут.

(т.е. Всички неключови атрибути не зависят функционално от други неключови атрибути)

→ НФБК

Една релация R е в НФБК, ако и само ако:

1. R е в 3NF
2. За всяка нетривиална ФЗ в R лявата страна е суперключ.

(Неключов атр. не може да определя ключов)

→ Многозначни зависимости ???

Многозначна зависимост (МЗ) (multivalued dependency, MVD) е твърдение за релация R, което означава, че при фиксиране на стойностите за определени атрибути, стойностите на точно определени други атрибути са НЗ?? от стойностите на всички останали атрибути в R. МЗ $X \twoheadrightarrow Y$ утвърждава, че ако два кортежа в R съвпадат по всички атрибути от X, техните компоненти от атрибути на Y могат да бъдат разменени и резултатът ще бъде два нови кортежа, които са от R ~~кортежи~~

Деф.: $A_1A_2..A_n \twoheadrightarrow B_1B_2..B_m$ е МЗ в R, ако $\forall t, u \in R$:

$t[A_1A_2..A_n] = u[A_1,A_2...A_n]$, $\exists v \in R$, за който:

- $v[A_1A_2...A_n] = t[A_1A_2..A_n] = u[A_1A_2...A_n]$
- $v[B_1B_2...B_m] = t[B_1B_2...B_m]$
- $v[C_1C_2..C_k] = u[C_1C_2..C_k]$

където $C_1C_2..C_k$ са всички атр. от R без $(A_1A_2..A_n \cup B_1...B_m)$

Анди да не би да е по-добре??

Една релация може да има излишества, но въпреки това да е в НФБК – напр. ако няма ФЗ и всички атрибути образуват ключ, но има МЗ.

→ Аксиоми на МЗ

- МЗ $A_1A_2...A_n \twoheadrightarrow B_1B_2...B_m$ се нарича тривиална, ако $B_1B_2...B_m \subseteq A_1A_2...A_n$ или $(A_1A_2..A_n \cup B_1B_2..B_m)$ съдържа всички атрибути на R
- МЗ $A_1A_2...A_n \twoheadrightarrow B_1B_2...B_m$ се нарича нетривиална, ако нито един от атрибутите $B_1B_2...B_m$ не съвпада с $A_1A_2...A_n$ и не всички атрибути на R принадлежат на $(A_1A_2...A_n \cup B_1B_2...B_m)$

а МЗ важат следните правила/аксиоми:
① Транзитивно правило – ако $A_1A_2...A_n \twoheadrightarrow B_1B_2...B_m$ и $B_1B_2...B_m \twoheadrightarrow C_1C_2...C_k$ то $A_1A_2...A_n \twoheadrightarrow C_1C_2...C_k$
② Правило на обединението – ако $A_1A_2...A_n \twoheadrightarrow B_1B_2...B_m$ и $A_1A_2...A_n \twoheadrightarrow C_1C_2...C_k$ то $A_1A_2...A_n \twoheadrightarrow (B_1...B_m \cup C_1...C_k)$
③ Правило на допълнението – ако $A_1A_2...A_n \twoheadrightarrow B_1B_2...B_m$, то $A_1...A_n \twoheadrightarrow C_1C_2...C_k$, където $C_1, C_2...C_k$ е множество от всички атрибути на R с изкл. на $(A_1...A_n \cup B_1...B_m)$
④ Правило FD-IS-AN-MVD ??? (всяко ФЗ е МЗ) – ако $A_1A_2...A_n \rightarrow B_1B_2...B_m$, то $A_1A_2...A_n \twoheadrightarrow B_1B_2...B_m$

→ Съединение без загуба, декомпозиция на релации

Декомпозиция на релацията $R(A_1, A_2, ...A_n)$ е заместването ѝ с мн-во релации $R_1, R_2, ... R_n$, получени чрез проекция така, че R и $R_1 \cup R_2 \cup ... \cup R_n$ имат една и съща схема.

Нека $R(A_1, A_2, ..., A_n)$ е релация и R се декомпозира на две релации $S(B_1, B_2 ... B_m)$ и $T(C_1, C_2, ..., C_k)$. Казваме, че декомпозиция без загуба,
- ако $R = S \bowtie T$ ($\bowtie$ е релация на естествено съединение от рел. алгебра)
или
- ⇔ за R е изпълнена поне една от следните ФЗ – $S \cap T \rightarrow S$ или $S \cap T \rightarrow T$

Следващото ме съмнява да трябва??? Алтернативна деф.??
Нека $R(A_1, A_2, ... A_n)$ е релация, за която е изпълнено множеството от ФЗ F, и R се декомпозира на 2 релации $S(B_1, B_2, .., B_m)$ и $T(C_1, C_2 ... C_k)$. Казваме, че декомпозицията е със съединение без загуба на ФЗ, ако $F_1 \cup F_2 = F$, където F_1 и F_2 са проекциите на F съответно в/у S и T.

Стратегия за декомпозиция – не ни трябва!

→ Четвърта нормална форма 4NF

Излишество, породено от МЗ, не може да се отстрани с привеждане в НФБК. Затова се въвежда 4NF, която елиминира всички МЗ и всички ФЗ, нарушаващи НФБК.

Една релация R е в 4NF, ако за всяка нетривиална МЗ $A_1A_2...A_n \twoheadrightarrow B_1...B_m$ е изпълнено, че $A_1A_2...A_n$ е суперключ. и ако е в НФБК

Тези май не трябва:

Декомпозиция към 4NF: Ако $X \twoheadrightarrow Y$ нарушава 4NF правим следната декомпозиция:

1. R_1 е с атрибути X, Y
2. R_2 е с атрибути всички без $X \cup Y$