Декомпозиция и нормализация

Създаване на добра схема

Една схема се състои от структура и ограничения. От гледна точка само на структурата е трудно да се каже, коя схема е най-добра. Как да се групират атрибутите на един обект? В общия случай семантиката определя избора. Общите и не много ясни семантични понятия трябва да бъдат преобразувани (представени) като синтактични свойства, удобни за формални преобразувания. Схемата може да бъде развивана в три посоки:

- Представяне – то трябва да отговаря на естествените връзки между атрибутите или обектите така, че всички естествени отношения представени чрез ограничения да бъдат отразени в схемата.
- Без излишък – всички ограничения, които могат да бъдат изведени от други, не трябва да са представени в схемата, защото дублирането на информацията създава объркване и аномалии.
- Разделяне – отделните единици за разделени и не взаимодействат една с друга.

С тези 3 критерия схемите могат да бъдат оценявани и променяни за да се дефинират ограниченията, наложени на данните. Функционалните зависимости (ФЗ) позволяват да се оценяват схемите

Декомпозиция

Трябва да се получи "добра" схема, като се отстрани излишъка и се осигури добро логическо разделяне на данните.

Оригиналната схема може да се дефинира по два начина:

- Дадени са всички релации и атрибути;
- Дадено е множество от релации и множество от ФЗ между атрибутите им.

Тези 2 случая изглеждат различни, но благодарение на теорията на ФЗ се свеждат до един..

Можем да си представим, че всички атрибути са включени в една универсална релация U по такъв начин, че всички схеми и релации са проекции на U Защо да не запазим U като релация със цялата съдържаща се информация? Този подход вкарва опасност от аномалии, които се избягват при декомпозицята.

Нека е дадена универсалната релация:

FIRM (N_employee, N_department, chef, contract_type)

или

HOSPITAL(N_patient, Name, Ward, Doctor)

- аномалия при обновяване – промяната на шефа на отдела изисква промяната на шефа на всеки чиновник от този отдел и на типа на договора, в който отдела е включен. Това води до преправяне на цялата база, повишена цена и риск за загуба на кохерентността.
- аномалия при изтриване – когато се изтрие и последния чиновник от даден отдел, се губи ицялата информация за отдела, която съществува само чрез неговите чиновници. А самият отдел може да продължи да съществува, но да се назанчат нови чиновници.
- аномалия при вмъкване – ако nul стойности не са позволени за даден атрибут, то стойности трябва да бъдат задавани, макар и да не са още известни, или трябва да се въвеждат многократно едни и същи стойности с риск да се направят грешки.
- излишък – типът на договара и шефът се повтарят n пъти, което води до неефективно ползване на паметта.

Декомпозицията използва 2 елементарни операции: проекция и съединение. Една декомпозиция на релацията R(A1,..., An) е заместването ù с множеството релации R1,…, Rn, получени чрез проекции така, че R et R1 * R2 * **...** * Rn имат една и съща схема.

Нека е дадена релацията DOCTOR

R	N°	name	age	address	speciality
	1	A	30	x1	S1
	2	B	30	x2	S2
	3	C	30	x3	S2
	4	D	40	x4	S1

R1	âge	speciality
	30	S1
	30	S2
	40	S1

R2	N	name	addr
	1	A	x1
	2	B	x2
	3	C	x3
	4	D	x4

R3	N	age
	1	30
	2	30
	3	30
	4	40

R'=R1*R2*R3	N°	name	address	age	speciality
	1	A	x1	30	S1
	1	A	x1	30	S2
	2	B	x2	30	S1
	2	B	x2	30	S2
	3	C	x3	30	S1
	3	C	x3	30	S2
	4	D	x4	40	S1

ако, освен това, всяка за всяка реализация R е изпълненено R = R1 * R2 * **...** *Rn, казваме че това е една **декомпозиция без загуба на информация.** Например ако е в сила ограничението, всички възрасти да са различни, то R=R1*R2*R3.

Ключ и нормални форми

Декомпозицията на една схема използва понятието функционална зависимост. Ключ е подмножество на атрибутите на релацията R(A1, A2,..., An), за което

X—>AlA2...An

и не съществува друго подмножество $Y \subset X$ за което Y —> Al A2 ... An.

Първа нормална форма

Една релация е в първа нормална форма, ако всичките атрибути съдържат атомарна стойност (*базирана върху прост домен*) и няма повтарящи се атрибути.

Тази дефиниция позволява да се избягват домени със съставни стойности.

Пример: EMPLOYEE (No, name, Child (firstname, age))

No	NAME	CHILD	
		FIRSTNAME	AGE
500	DUPONT	ANDRE	10
501	DURAND	JEAN	11
501	DURAND	PIERRE	12
510	LEFEBVRE	PAUL	13
510	LEFEBVRE	JACQUES	14

Child е съставна група и има излишък на информация. Освен това, No вече не е ключ, защото се повтаря в повече от един кортеж. За да стане ключ трябва да се добавят атрибути, с което се

усложнява схемата. Едно решение е да се направи декомпозиция, като се включи ключа в две нови релации и се премахне повтарящатата се група.

EMPLOYEE	**No**	**NAME**
	500	DUPONT
	501	DURAND
	510	LEFEBVRE

CHILDREN	**No**	**FirstName**	**age**
	500	André	10
	501	Jean	11
	501	Pierre	12
	510	Paul	13
	510	Jacques	14

Пример 2: MOVIE(No, Name, Director, Actor1, Actor2, Actor3)

MOVIE	**No**	**Name**	**Director**	**Actor1**	**Actor2**	**Actor3**
	1	The Silence of the Lambs	Jonathan Demme	Jodie Foster	Anthony Hopkins	
	2	The Sixth Sense	M. Night Shyamalan	Bruce Willis		

В този случай има повтарящи се атрибути, което води до следните неудобства:

- Не могат да се запишат повече от 3-ма актьори или пък ако са по-малко, ще останат празни полета.
- Трудно се търси, в кои филми играе даден актьор.

Декомпозицията в този случай е следната:

MOVIE(No, Name, Director) и CAST(No, Actor)

MOVIE	**No**	**Name**	**Director**
	1	The Silence of the Lambs	Jonathan Demme
	2	The Sixth Sense	M. Night Shyamalan

CAST	**No**	**Actor**
	1	Jodie Foster
	1	Anthony Hopkins
	2	Bruce Willis

Втора нормална форма

Определение: Ще казваме, че Y напълно зависи от X, ако $X \rightarrow Y$ и няма подмножество $Z \subset X$, за което $Z \rightarrow Y$, т.е. Y зависи от целия X, но не зависи от никоя негова част.

Една релация R е във 2NF ако и само ако:

- Тя е в 1 NF
- Всички неключови атрибути са напълно зависими от ключа на R.

Пример : STOCK (part, warehouse, quant, address)

Атрибутът **address** зависи само от атрибута **warehouse.** За всеки кортеж съдържащ името на склад има излишък на информация за адреса на този склад.

В случай на смяна на адреса, за всеки кортеж, който се отнася до части складирани в този склад трябва да се обновява адреса. Освен това ако в даден склад няма части, няма и да съществува адреса на склада.

Релацията не е във 2NF и трябва да бъде декомпозирана:

STOCK (part, warehouse, quant)

LOCAL(<u>warehouse</u>, address) , която е в 2 NF

Трета нормална форма

Една релация е в 3NF, ако и само ако:

- Тя е във 2NF
- *Всички атрибути не принадлежащи на един ключ не зависят от неключов атрибут.*

Пример: PERSONAL(<u>employee</u>, name, FirstName, Service, Address)

Address зависи са мо от **Service**.

Декомпозира се на **: PERSONAL(<u>employee</u>, name, FirstName, Service)**

LOCAL(<u>Service</u>, Address)

Ако ключът е първичен второто правило може да се изкаже по следния начин: Всички атрибути не принадлежащи на ключа не зависят *транзитивно* от него.

На практика 3NF е достатъчна за изграждането на една база. Тя е много важна, тъй като всяка релация има най-малко една декомпозиция до 3NF със следните свойства:

- Тя запазва ФЗ (оригиналните ФЗ могат да се извлекат и от новите релации)
- Тя е без загуба на информация.

Пример : CAR(<u>No</u>, Make, Model, Power, Color) приема 2 тъпа декомпозиция до 3 NF

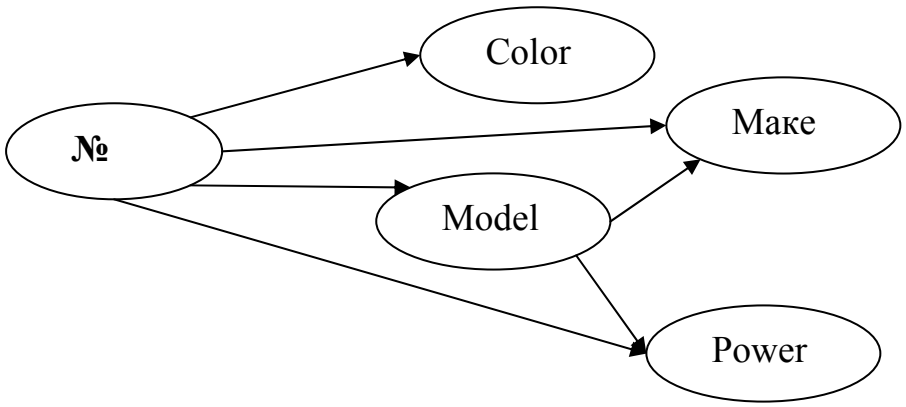
R1(<u>No</u>, Model, Color)

R2 (<u>Model</u> ,Make, Power) която запазва ФЗ.

R'1(<u>No</u>, Model)

R'2(<u>No</u>, Power, Color)

R'3(<u>Model</u> ,Make) която не ги запазва, защото липсва ФЗ **Model→ Power**

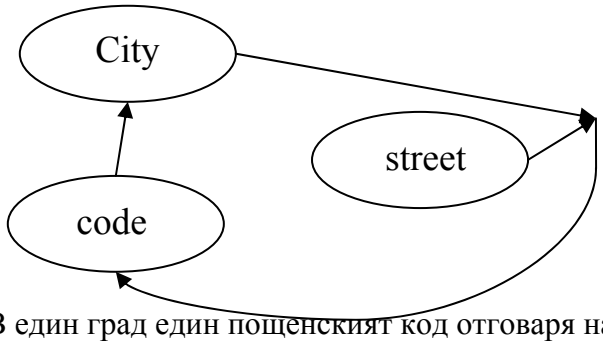


Нормална форма на BOYCE-CODD (BCNF)

Една релация е в BCNF, ако и само ако един ключ определя атрибут само в елементарни ФЗ (елиминират се частичните и транзитивните ФЗ). АСко една релация е в BCNF, то тя е и в 3NF.

BCNF не запазва функционалните зависимости. Да разгледаме пример, в който съществуват ФЗ от вида AB→C и C→A

Нека POSTCODE(code, <u>City, street)</u>



В един град един пощенският код отговаря на няколко улици.

POSTCODE	code	City	street
	59650	VA	Gambetta
	59650	VA	Jean-Jaurès

Декомпозицията е : CODE CITY (code, City)

CODE STREET(code, street)

Втората ФЗ се губи, но може да се възстанови чрез съединение по атрибута **code.**

Пример 2:

В един университет се ицка информация за записванията на студентите по дисциплини и назначаването на асистенти за лабораторни упражнениея. Нека са в сила следните правила:

- В един курс(дисциплина) се записват много студенти.
- Всеки студент може да запище в повече курсове
- Всеки асистент се назаначава само за един курс
- Всеки студент има единсвен асистент за даден курс.

COURSE_STUD_TUT	**CourseNum**	**Student**	**Tutor**
	ENG101	Jones	Clark
	ENG101	Grayson	Chen
	ENG101	Samara	Chen
	MAT350	Grayson	Powers
	MAT350	Jones	O'Shea
	MAT350	Berg	Powers

В релацията има два възможни ключа – (**CourseNum, Student)** и (**Student, Tutor**). Но ако вторият е приет за първичен ключ, релацията не би била във 2NF, защото **CourseNum** зависи само от **Tutor** и затова не може да бъде кандадат-ключ.
Асистентите не могат да бъдат определени преди записването на студентите, защото **Student** е част от ключа. Това е причината за декомпозицията на Boyce-Codd (вж. по-долу).

INSCRIPTION	**CourseNum**	**Student**
	ENG101	Jones
	ENG101	Grayson
	ENG101	Samara
	MAT350	Grayson
	MAT350	Jones
	MAT350	Berg

TUTORIAL	**Student**	**Tutor**
	Jones	Clark
	Grayson	Chen
	Samara	Chen
	Grayson	Powers
	Jones	O'Shea
	Berg	Powers

TUTORS	**CourseNum**	**Tutor**
	ENG101	Clark
	ENG101	Chen
	MAT350	Powers
	MAT350	O'Shea

Многозначни зависимости и четвърта нормална форма

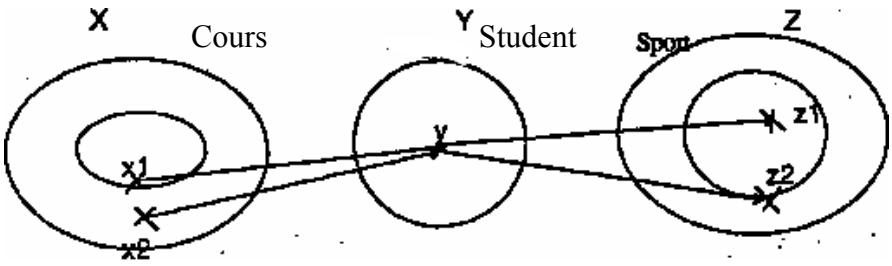
Не винаги BCNF е достатъчна, за да елиминира аномалиите по обновяване и излишъка.

STUDENTS	N	COURS	SPORT
	100	BD	Tennis
	100	BD	Football
	200	BD	Swimming
	200	AN	Swimming

Тази релация е в 3NF, но все още има излишък, а в същото време няма декомпозиция, тъй като няма функционални зависимости. Това е слабост на концепцията за еднозначната ФЗ, която не взима предвид взаимозависимостта между атрибути като COURS и SPORT. Тук ще въведем понятието многозначна зависимост (МЗ).

Нека R(A1, Ai, An) е релация, а X, Z са подмножества на ключа. Ще казваме, че X —» Z или че съществува мулти определяне на Z по отношение на, ако за стойностите на Y, съществува множество от стойности на X, независимо от другите атрибути (множество Z) на релацията.

Една МЗ характеризира взаимозависимостта на 2 множества от стойности на атрибутите X и Z, корелирани от трети атрибут Y. С други думи, всяка стойност на Y определя едно множество от стойности на X и друго множество от стойности на Z, но последните два атрибута не зависят един от друг.



.

$$X \twoheadrightarrow Z : x1yz1 \text{ et } x2yz2 \in R \Rightarrow x1yz2 \text{ et } x2yz1 \in R$$

Вижда се,че ФЗ е частен случай на МЗ. Съществуват следните аксиоми.

ДОПЪЛНЕНИЕ : X—» Y ==> X—»R-X-Y

МУЛТИ-УВЕЛИЧЕНИЕ : ако (X —» Y) и W е множество от атрибути на R, то XW—»YW

ПСЕВДОТРАНЗИТИВНОСТ : (X —» Y) и (Y —» Z) ==> X —»Z-Y

ПОВТОРЕНИЕ : $X \rightarrow Y$ ==> X —» Y

КОВАРИАНТНОСТ : X --» Y и $Z \subseteq Y$ и съществува $W \subseteq R$ с$W \cap Y = \varnothing$ и $W \rightarrow Z$, то $X \rightarrow Z$

изчисленията използват аксиомите на :

ОБЕДИНЕНИЕ : (X -—» Y) и (Y —» Z) ==> X —»YZ.

СЕЧЕНИЕ : X—» Y и X—»Z==>X —» $Y \cap Z$.

РАЗЛИКА: X—» Y и X—»Z==> X—» Y-Z и X —» Z-Y

Една елементарна ДЗ X —» Y на R е такава, че

- Y не е празно и не се пресича с X.
- R не съдържа друга ДЗ X' —» Y' такава, че X' ∈ X и Y' ∈ Y.

Четвърта нормална форма

Една релация е в 4NF ако и само ако единствените елементарни ДЗ са тези, при които един ключ определя един атрибут,

STUDENT (num, cours, sport) не е в 4NF, защото ключът е множество от атрибути. Съществуват следните елементарни ДЗ между атрибутите принадлежащи на ключа:

num -—» cours – R1(num, cours)

num —» sport– R2(num, sport).

Тези 2 релации са в 4NF. Една релация в 4NF е в BCNF и следователно в 3NF.

Зависимост от съединение и пета нормална форма

Декомпозицията до 4NF не е достатъчна да се елиминират аномалиите, защото съществува още излишък

R1	STUDENT	COURS	PROF
	X	CL	Z
	X	CL	T
	X	SIO	T
	Y	SIO	T

Тази релация н е в 4NF, но е съществува ДЗ:

(X CL Z) и (X SIO T) =/=> (X CLT) ∈ R и (X SIO Z) ∈ R

STUDENT —» COURS не е вярно, защото (X SIO Z) не съществува и може въобще да не съществува.

COURS —» PROF също, защото (Y CL Z) не съществува.

PROF —» STUDENT също, защото (Y SIO T) не съществува.

Soient les 3 projections de R :

R1	STUDENT	COURS
	X	CL
	X	SIO
	Y	CL

R2	STUDENT	PROF
	X	Z
	X	T
	Y	T

R3	COURS	PROF
	CL	Z
	CL	T
	SIO	T

В R има излишък (X CL съществува 2 пъти), но съединенията R1 R2, R2 R3 или R1 R3 не дават R. Опитът за декомпозиция на 2 релации (X —» Y ===> (XY) и(YZ)) е неуспешен, защото Y и Z са независими от X. Съществуват релации, които не могат да бъдат декомпозирани до 2, до N релации.

ако *(a,b) ∈ R1*

(a, *c) ∈ R2* ===> (a, *b,* c) ∈ *R*

(b,c) ∈ R3

то : R = R1 * R2 * R3. Това се нарича зависимост от съединение (отбелязана с *), на която ДЗ са частен случай, защото:

нека е дадена R(X, Y, Z) с X —» Y и X —» Z. Тогава, зависимостта от съединение *(XY, XZ) е удовлетворена.

ДЗ се използват, за да изразят взаимозависимостта между 2 атрибута. Зависимостите от съединение изразяват взаимозависимостите на повече атрибути и представляват по-широко понятие от ДЗ.

Пета нормална форма.

Една релация е в 5NF ако всички зависимости от съединение включват кандидат-ключовете на R. Нека е дадена R(A1, A2, A3, A4), като A1 и A2 са кандидат ключове. Декомпозиция без загуба е:

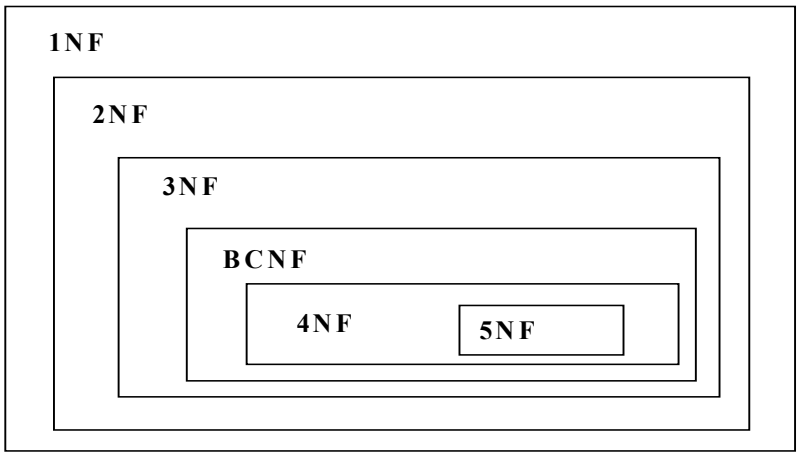
*(A1A2, A1A3, A1A4) или още:

*(A1A2, A2A3, A3A4)

R : *(student cours, cours prof, student prof) трябва да бъде декомпозирана до R1, R2, R3. В 5NF няма повече риск от аномалии на обновяване, изтриване и вмъкване.

Релация в 5NF не може да се декомпозира без загуба на информация. Следователно няма повече декомпозиции.

Отношението между различните схеми на нормализация може да се изрази графично:



Забележка: Процесът на нормализация има за цел да премахне междурелационните връзки, и следователно да увеличи ефективността и да избегне аномалиите при съхранение. Но това създава една фрагментация, която затруднява и забавя правенето на справки (увеличава се броят на съединенията). Следователно е за предпочитане да се използва база с по-ниска степен на нормализация, особено ако се разполага със средства за управление на аномалиите, както е в случая на складовете от данни.

Пример

Изследваната област са полетите осигурени от различни самолетни компании.

Данните са моделирани чрез различни атрибути по следния начин:

- Имената на компаниите чрез атрибута C
- номерата на полетите чрез атрибута V
- номерата на самолетите чрез атрибута A
- моделите на самолетите чрез атрибута M
- капацитета на самолетите чрез атрибута K
- имената на пилотите чрез атрибута P

– часовете на излитане чрез атрибута H
– дните на заминаване за полета чрез атрибута J
– начален град на полета чрез атрибута D
– краен град на полета чрез атрибута E

Изучаването на реалния обект показа следните ФЗ:

(1) Всеки модел самол има само един капацитет на местата
$M \rightarrow K$

(2) Всеки самолет е от един модел и принадлежи на една компания:
$A \rightarrow M, C$

(3) Всеки полет заминава има само един час на заминаване и едни и същи начални и крайни град.
$V \leftrightarrow H, D, E$

(4) За ждаден ден полетът е осигурен с един пилот и определен самолет :
$J, V \rightarrow P, A$

(5) За даден ден и час един пилот лети по точно определен маршрут с точно определен самолет:
$J, H, P \rightarrow D, E, A$

(6) За даден ден и час един самолет лети по точно определен маршрут с точно определен пилот:
$J, H, A \rightarrow D, E, P$

Универсална релация:

R (C, A, V, M, K, P, H, J, D, E) с кандидат ключове [VJ], [HDEJ], [PJH], [AJH]

От ФЗ (3) R не е в 2NF :

R1(V, H, D, E) avec sur-clés [V] et [HDE]

R2(C, A, V, M, K, P, J) с кандидат-ключове [VJ]

От ФЗ (2) R2 не е в 3NF

R21(V, J, A, P) [VJ]

R22(A, C, M, K) [A]

От ФЗ (1) не е в 3NF

R221(A, C, M) [A]

R222(M, K) [M]

Нормализираната схема е:

R1(<u>V</u>, H, D, E)

R21(<u>V, J,</u> A, P)

R221(<u>A</u>, C, M)

R222(<u>M</u>, K)

Предимства и недостатъци на релационния модел

Предимства

a1 – Простота за потребителя
a2 – Независимост на потребителя от логическата и физическата структура, както и от методите за достъп до данните. Понятието файл съществува само за администратора и е грижа на СУБД.
a3 – Мощност и еднотипност на представянето : математическата теория позволява точно и алгоритмично проектиране на схемата.
a4 – Мощност на осигуряване на сигурност на данните: контрол според съдържанието, структурата и контекста.
a5 – Съществуване на непроцедурен интерфейс за неинформатици.
a6 – Бурно развитие на комерсиални СУБД и създаването на 4-то поколение SQL и QBE.

Недостатъци

i1 – Необходимост от мощна СУБД
i2 – Известна загуба на логическа независимост при нормализацията.