Normalizacja schematów logicznych

PROJEKTOWANIE SCHEMATÓW BAZ DANYCH

Problem

Dany jest zbiór atrybutów $U = \{A_1, A_2, ..., A_n\}$ oraz informacja o zależnościach między tymi atrybutami - (*FD*, *MVD*).

Projektowanie logicznej struktury relacyjnej bazy danych polega na definiowaniu schematu logicznego bazy danych *B*, to jest , na definiowaniu schematów relacji tworzących tę bazę danych w taki sposób, aby schematy te posiadały pożądane własności ułatwiające eksploatację bazy danych.

Dostawcy

Nazwisko	Adres	Towar	Cena
Dziubasik	Czajcza 5	Deska	10
Dziubasik	Czajcza 5	Śrubka 1	0,2
Dziubasik	Czajcza 5	Śrubka 2	O,3
...	...	...	...
Nowak	Krucza 10	Deska	9
Nowak	Krucza 10	Śrubka1	0,25

Adresy_dostawców

Nazwisko	Adres
Dziubasik	Czajcza 5
Dziubasik	Czajcza 5
Dziubasik	Czajcza 5
...	...
Nowak	Krucza10
Nowak	Krucza10

Dostawy

Nazwisko	Towar	Cena
Dziubasik	Deska	10
Dziubasik	Śrubka1	0,2
Dziubasik	Śrubka2	0,3
...	...	...
Nowak	Deska	9
Nowak	Śrubka1	0,25

Zapytanie

Podaj adresy dostawców dostarczających towary: ”deska”, ”śrubka 1”, ...

Cechy relacji Dostawca

- redundacja danych -problem spójności danych
- anomalia wprowadzania danych
- anomalia usuwania danych
- anomalia uaktualniania danych
- dekompozycja bez utraty informacji

Pożyczki

Oddział	Nr.konta	Kwota	Nazwisko
III PKO	17	1	Nowak
III PKO	23	1,5	Kowalski
V NBP	5	3	Zimoch
V NBP	10	1,5	Kusiak
I PKO	77	3	Puc
BGŻ	17	1	Walusiak
BGŻ	21	1,3	Mijał

Konta pożyczkowe

Odział	Nr_konta	Kwota
III PKO	17	1
III PKO	23	1.5
V NBP	5	3
V NBP	10	1.5
I PKO	77	3
BGŻ	17	1
BGŻ	21	1.3

Pożyczkobiorcy

Kwota	Nazwisko
1	Nowak
1.5	Kowalski
3	Zimoch
1.5	Kusiak
3	Puc
1	Walusiak
1.3	Mijał

Odtworzenie relacji początkowej:

Konta-pożyczkowe ⋈ *Pożyczkobiorcy* (wg atrybutu połączeniowego *Kwota*)

Pożyczki

Oddział	Nr_konta	Kwota	Nazwisko
III PKO	17	1	Nowak
III PKO	17	1	Walusiak
III PKO	23	1.5	Kowalski
III PKO	23	1.5	Kusiak
V NBP	5	3	Zimoch
V NBP	5	3	Puc
V NBP	10	1.5	Kusiak
V NBP	10	1.5	Kowalski
I PKO	77	3	Puc
I PKO	77	3	Zimoch
BGŻ	17	1	Walusiak
BGŻ	17	1	Nowak
BGŻ	21	1.3	Mijał

- Dekompozycja relacji z utratą informacji
- Dekompozycja relacji bez utraty informacji

Zależności funkcyjne

Niech schemat bazy danych posiada *n* atrybutów $A_1, A_2, ..., A_n$. Atrybuty te tworzą tzw. uniwersalny schemat relacji $R = A_1, A_2, ..., A_n$.

Zależność funkcyjna (FD)

Dana jest relacja *r* o schemacie *R*. *X*,*Y* są podzbiorami atrybutów *R*. W schemacie relacji *R*, X wyznacza funkcyjnie *Y*, lub *Y* jest funkcyjnie zależny od *X*, co zapisujemy $X \rightarrow Y$, wtedy i tylko wtedy, jeżeli dla dwóch dowolnych krotek t_1, t_2 takich, że $t_1[X] = t_2[X]$ zachodzi zawsze $t_1[Y] = t_2[Y]$.

Zależność funkcyjna określa zależność pomiędzy atrybutami. Jest to własność semantyczna, która musi być spełniona dla dowolnych wartości krotek relacji. Relacje które spełniają nałożone zależności funkcyjne nazywamy instancjami legalnymi. Zależność funkcyjna jest własnością schematu relacji *R*, a nie konkretnego wystąpienia relacji.

1. Nazwisko → Adres
2. {Nazwisko, Towar} → Cena

$t_1[X] = t_2[X]$ *i* $X \rightarrow Y$, to zachodzi zawsze $t_1[Y] = t_2[Y]$.

NORMALIZACJA

Wprowadzenie

Proces normalizacji relacji można traktować jako proces, podczas którego schematy relacji posiadające pewne niepożądane cechy są dekomponowane na mniejsze schematy relacji o pożądanych własnościach.

Proces normalizacji musi posiadać trzy dodatkowe własności:

- **Własność zachowania atrybutów** - żaden atrybut nie zostanie zagubiony w trakcie procesu normalizacji
- **Własność zachowania informacji** - dekompozycja relacji nie prowadzi do utraty informacji
- **Własność zachowania zależności** - wszystkie zależności funkcyjne są reprezentowane w pojedynczych schematach relacji

Nadkluczem schematu relacji $R = \{A_1, A_2, ..., A_n\}$ nazywamy zbiór atrybutów $S \subseteq R$, który jednoznacznie identyfikuje wszystkie krotki relacji r o schemacie R. Innymi słowy, w żadnej relacji r o schemacie R nie istnieją dwie krotki t_1, t_2 takie, że $t_1[S] = t_2[S]$.

Kluczem K schematu relacji R nazywamy *minimalny nadklucz*, to znaczy taki, że nie istnieje $K' \subset K$ będące *nadkluczem* schematu R.

Klucze potencjalne (ang. candidate keys)

- klucz podstawowy (ang. primary key)
- klucze drugorzędne (ang. secondary keys)

Atrybuty:

- atrybuty podstawowe: atrybut X jest podstawowy w schemacie R jeżeli należy do któregokolwiek z kluczy schematu R;
- atrybuty wtórne: atrybut X jest wtórny w schemacie R jeżeli nie należy do żadnego z kluczy schematu R.

Pierwsza postać normalna (1NF)

Schemat relacji R znajduje się w pierwszej postaci normalnej(*1NF*), jeżeli wartości atrybutów są atomowe (niepodzielne).

Płeć

Imię	**Płeć**
Piotr,Jan,Tomasz	męska
Janina,Anna,Maria	żeńska

Relacja Płeć w 1NF

Imię	**Płeć**
Piotr	męska
Jan	męska
Tomasz	męska
Janina	żeńska
Anna	żeńska
Maria	żeńska

Pierwsza postać normalna zabrania definiowania złożonych atrybutów, które są wielowartościowe. Relacje ,które dopuszczają definiowanie takich złożonych atrybutów nazywamy **relacjami zagnieżdżonymi** (ang. nested relations). W relacjach zagnieżdżonych każda krotka może zawierać inną relację.

Pracownicy (idPrac, Nazwisko, {Projekty (nr, godziny)})

Pracownicy

IdPrac	**Nazwisko**	**Projekty**	
		Nr	**Godziny**
12345678	Nowak	1 2	32.5 7.5
66543723	Dziubasik	3	40.5
43432266	Tarzan	1 2	20.0 20.0
33333333	Morzy	2 3 10 20	10.0 10.0 10.0 10.0

Pracownicy ***Uczestnictwo***

IdPrac	**Nazwisko**

IdPrac	**Nr**	**Godziny**

Druga postać normalna (2NF)

Pełna zależność funkcyjna

Zbiór atrybutów Y jest w pełni funkcyjnie zależny od zbioru atrybutów X w schemacie R, jeżeli $X \rightarrow Y$ i nie istnieje podzbiór $X' \subset X$ taki, że $X' \rightarrow Y$.

Zbiór atrybutów Y jest częściowo funkcyjnie zależny od zbioru atrybutów X w schemacie R, jeżeli $X \rightarrow Y$ i istnieje podzbiór $X' \subset X$ taki, że $X' \rightarrow Y$.

Druga postać normalna

Dana relacja r o schemacie R jest w drugiej postaci normalnej (2NF), jeżeli żaden atrybut wtórny tej relacji nie jest częściowo funkcyjnie zależny od żadnego z kluczy relacji r .

II wersja 2NF

Dana relacja r o schemacie R jest w drugiej postaci normalnej (2NF), jeżeli każdy atrybut wtórny tej relacji jest w pełni funkcyjnie zależny od klucza podstawowego relacji r .

Uczestnictwo

IdPrac	NrProj	Funkcja	Nazwisko	NazwaProj	Lokalizacja

fd1: {IdPrac, NrProj} $\rightarrow$ *Funkcja*
fd2: {IdPrac, NrProj} $\rightarrow$ *Nazwisko*
fd3: {IdPrac, NrProj} $\rightarrow$ *NazwaProj*
fd4: {IdPrac, NrProj} $\rightarrow$ *Lokalizacja*
fd5: {IdPrac} $\rightarrow$ *Nazwisko*
fd6: {NrProj} $\rightarrow$ *NazwaProj*
fd7: {NrProj} $\rightarrow$ *Lokalizacja*

Zależności *fd2*, *fd3*, *fd4* są zależnościami niepełnymi

Uczestnictwo'

IdPrac	NrProj	Funkcja

fd1: {IdPrac, NrProj} $\rightarrow$ *Funkcja*

Pracownicy

IdPrac	ENAME

fd5: {IdPrac} $\rightarrow$ *Nazwisko*

Projekty

NrProj	NazwaProj	Lokalizacja

fd6: {NrProj} $\rightarrow$ *NazwaProj*
fd7: {NrProj} $\rightarrow$ *Lokalizacja*

$\{fd1, fd2, fd3, fd4, fd5, fd6, fd7\}^+ \equiv \{fd1, fd5, fd6, fd7\}^+$
bo:
fd1 $\Rightarrow$ *fd2*, *fd3*, *fd4*, zgodnie z regułą poszerzenia

Trzecia postać normalna (3NF)

Pracownicy-PP

Nazwisko	**Instytut**	**Wydział**
Brzeziński	I.Informatyki	Elektryczny
Morzy	I.Informatyki	Elektryczny
Koszlajda	I.Informatyki	Elektryczny
Królikowski	I.Informatyki	Elektryczny
...	...	...
Babij	ElektroEnerg.	Elektryczny
Kordus	ElektroEnerg.	Elektryczny
Sroczan	ElwktroEnerg.	Elektryczny

Klucz: *Nazwisko*

Zależności funkcyjne:

Nazwisko → *Instytut*
Nazwisko → *Wydział*
Instytut → *Wydział*

Przechodnia zależność funkcyjna

Zbiór atrybutów *Y* jest przechodnio funkcyjnie zależny od zbioru atrybutów *X* w schemacie *R*, jeżeli X → Y i istnieje zbiór atrybutów Z, nie będący podzbiorem żadnego klucza schematu R taki, że zachodzi X → Z i Z → Y.

Zależność funkcyjna *X* → *Y* jest zależnością przechodnią jeżeli istnieje podzbiór atrybutów *Z* taki, że zachodzi *X* → *Z*, *Z* → *Y* i nie zachodzi *Z* → *X* lub *Y* → *Z*.

Trzecia postać normalna

Dana relacja *r* o schemacie *R* jest w trzeciej postaci normalnej (***3NF***), jeżeli dla każdej zależności funkcyjnej *X* → *A* w *R* spełniony jest jeden z następujących warunków:

(a) *X* jest nadkluczem schematu *R*, lub
(b) *A* jest atrybutem podstawowym schematu *R*.

II wersja 3NF

Dana relacja *r* o schemacie *R* jest w trzeciej postaci normalnej (*3NF*), jeżeli jest w drugiej postaci normalnej i żaden atrybut wtórny nie jest przechodnio zależny od odstawowego klucza schematu *R*.

Pracownicy-PP-1

Nazwisko	**Instytut**
Brzeziński	I.Informatyki
Morzy	I.Informatyki
Koszlajda	I.Informatyki
Królikowski	I.Informatyki
...	...
Babij	ElektroEnerg.
Kordus	ElektroEnerg.
Sroczan	ElektroEnerg.

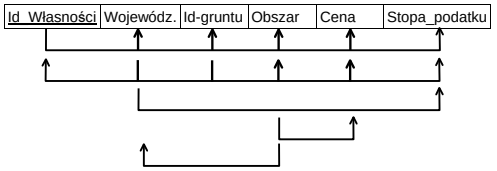
Pracownicy-PP-2

Instytut	**Wydział**
I.Informatyki	Elektryczny
...	...
ElektroEnerg.	Elektryczny

Postać normalna Boyce-Codd (BCNF)

Postać normalna *Boyce-Codd'a* stanowi warunek dostateczny 3NF, ale nie konieczny.

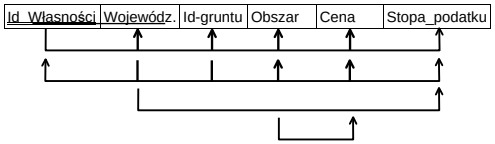
GRUNTY



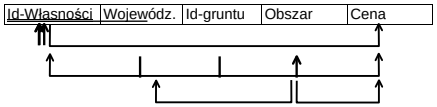
Załóżmy, że w relacji *Grunty* mamy tylko dwa województwa. Co więcej, załóżmy, że działki w pierwszym województwie mają rozmiar 0.5, 0.6, 0.7 h; natomiast działki w drugim województwie mają obszar 1, 1.2, 1.4 h. Ta informacja może być powielona w tysiącach krotek relacji *Grunty* oraz, po dekompozycji, w relacji *Grunty-1A*.

Relacja *Grunty* jest nadal w trzeciej postaci normalnej (*Wojewódz*. jest atrybutem podstawowym)

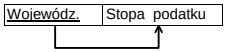
Grunty



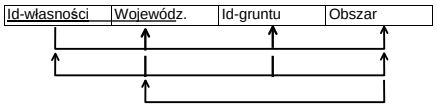
Grunty-1



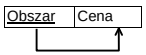
GRUNTY-2



GRUNTY-1A



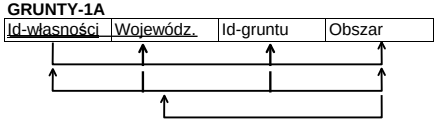
GRUNTY-1B



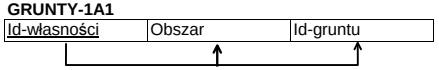
Postać normalna Boyca-Codd

Schemat relacji *R* jest w postaci *BCNF*, jeżeli dla każdej zależności funkcyjnej *X* → *A* w *R*, *X* jest nadkluczem schematu *R*.

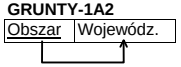
GRUNTY-1A



GRUNTY-1A1



GRUNTY-1A2



ZALEŻNOŚCI WIELOWARTOŚCIOWE

Loty

Lot	**Dzień_tygodnia**	**Typ_samolotu**
106	poniedziałek	134
106	czwartek	154
106	poniedziałek	154
106	czwartek	134
206	środa	747
206	piątek	767
206	środa	767
206	piątek	747

Języki

Nazwisko	**Język_obcy**	**Język_programowania**
Nowak	angielski	Basic
Nowak	włoski	Fortran
Nowak	angielski	Fortran
Nowak	włoski	Basic
Nowak	czeski	Basic
Nowak	czeski	Fortran

Modyfikacja relacji z zależnościami wielowartościowymi

Lot 106 będzie dodatkowo odbywał się w Środę i na tę linię wprowadzamy, dodatkowo, nowy typ samolotu – 104.

Loty

Lot	Dzień-tygodnia	Typ-samolotu
106	poniedziałek	134
106	czwartek	154
106	poniedziałek	154
106	czwartek	134
106	poniedziałek	104
106	czwartek	104
106	środa	134
106	środa	154
106	środa	104

Dekompozycja

Lot-1

Lot	Dzień-tygodnia
106	poniedziałek
106	czwartek
206	środa
206	piątek
...	...
106	środa

Lot-2

Lot	Typ-samolotu
106	134
106	154
206	747
206	767
...	...
106	104

Język-1

Nazwisko	Język_obcy
Nowak	angielski
Nowak	włoski
Nowak	czeski

Język-2

Nazwisko	Język_prog
Nowak	Basic
Nowak	Fortran

Zależności wielowartościowe

Zależności wielowartościowe są konsekwencją wymagań pierwszej postaci normalnej, która nie dopuszcza, aby krotki zawierały atrybuty wielowartościowe.

Zależność wielowartościowa występuje w relacji r(R) nie dlatego, że na skutek zbiegu okoliczności tak ułożyły się wartości krotek, lecz występuje ona dla dowolnej relacji r o schemacie R dlatego, że odzwierciedla ona ogólną prawidłowość modelowanej rzeczywistości.

Lot→→Dzień-tygodnia

Lot→→Typ-samolotu

Nazwisko→→Język-obcy

Nazwisko→→Język-programowania

Wystąpienie zależności wielowartościowej $X \rightarrow\rightarrow Y$ w relacji o schemacie $R = XYZ$ wyraża dwa fakty:

- Związek pomiędzy zbiorami atrybutów X i Y;
- Niezależność zbiorów atrybutów Y, Z. Zbiory te są związane ze sobą pośrednio poprzez zbiór atrybutów X.

Lot-3

Lot	Dzień-tygodnia	Typ-samolotu
106	poniedziałek	134
106	czwartek	154
106	czwartek	134
206	środa	747
206	piątek	767

Definicja własności zależności wielowartościowych

Niech R oznacza schemat relacji, natomiast X, Y są rozłącznymi zbiorami atrybutów schematu R i $Z = R – (XY)$.

Relacja $r(R)$ spełnia zależność wielowartościową $X \rightarrow\rightarrow Y$, jeżeli dla dwóch dowolnych krotek t_1 i t_2 z $r(R)$ takich, że $t_1[X] = t_2[X]$, zawsze istnieją w $r(\mathbf{R})$ krotki t_3, t_4 takie, że spełnione są następujące warunki:

- $t_1[X] = t_2[X] = t_3[X] = t_4[X]$
- $t_3[Y] = t_1[Y]$ i $t_4[Y] = t_2[Y]$
- $t_3[R - X – Y] = t_2[R – X – Y]$ i
- $t_4[R – X – Y] = t_1[R – X – Y]$

Z symetrii powyższej definicji wynika, że jeżeli w relacji $r(R)$ zachodzi $X \rightarrow\rightarrow Y$, to zachodzi również:

$$X \rightarrow\rightarrow [R – X – Y].$$

Ponieważ $R – X – Y = Z$, to powyższy fakt zapisujemy czasami w postaci: $X \rightarrow\rightarrow Y / Z$.

Trywialna zależność wielowartościowa

Zależność wielowartościowa $X \rightarrow\rightarrow Y$ w relacji $r(R)$ nazywamy zależnością trywialną, jeżeli

- zbiór Y jest podzbiorem X, lub
- $X \cup Y = R$

Zależność nazywamy trywialną, gdyż jest ona spełniona dla dowolnej instancji r schematu R.

Czwarta postać normalna(4NF)

Relacja r o schemacie R jest w czwartej postaci normalnej (4NF) względem zbioru zależności wielowartościowych *MVD* jeżeli jest ona w *3NF* i dla każdej zależności wielowartościowej $X \rightarrow\rightarrow Y \in MVD$ zależność ta jest trywialna lub X jest nadkluczem schematu R.

Dekompozycja relacji na pod-relacje bez utraty informacji

1. Dekompozycja na podrelacje w 3NF

Dana jest relacja r o schemacie R, i dany jest zbiór F zależności funkcyjnych dla R. Niech relacje r_1 i r_2 o schematach, odpowiednio, R_1 i R_2, oznaczają dekompozycję relacji $r(R)$. Dekompozycja ta jest dekompozycją bez utraty informacji, jeżeli co najmniej jedna z poniższych zależności funkcyjnych jest spełniona:

- $R_1 \cap R_2 \rightarrow R_1$
- $R_1 \cap R_2 \rightarrow R_2$

2. Dekompozycja na pod-relacje w 4NF

Dana jest relacja r o schemacie R. Niech relacje r_1 i r_2 o schematach, odpowiednio, R_1 i R_2, oznaczają dekompozycję relacji $r(R)$. Dekompozycja ta jest dekompozycją bez utraty informacji, jeżeli co najmniej jedna z poniższych zależności wielowartościowych jest spełniona:

- $R_1 \cap R_2 \rightarrow\rightarrow (R_1 - R_2)$
- $R_1 \cap R_2 \rightarrow\rightarrow (R_2 - R_1)$

Zależności połączeniowe

Agenci

Agent	Firma	Produkt
Kulczyk	Volkswagen	samochody
Kulczyk	Volkswagen	ciężarówki
Kulczyk	Audi	samochody
Kulczyk	Audi	ciężarówki
Nowak	Ford	samochody
Nowak	Ford	ciężarówki
Misiek	Nissan	samochody

R1

Agent	Firma
Kulczyk	Volkswagen
Kulczyk	Audi
Nowak	Ford
Misiek	Nissan

R2

Agent	Produkt
Kulczyk	samochody
Kulczyk	ciężarówki
Nowak	samochody
Nowak	ciężarówki
Misiek	samochody

R3

Firma	Produkt
Volkswagen	samochody
Volkswagen	ciężarówki
Audi	samochody
Audi	ciężarówki
Ford	samochody
Ford	ciężarówki
Nissan	samochody
Nissan	ciężarówki

Zależności połączeniowe

Niech: $R = \{R_1, R_2, \ldots, R_p\}$

oznacza zbiór schematów relacji, zdefiniowanych nad zbiorem atrybutów:

$$U = \{A_1, A_2, \ldots, A_n\},$$

takich że:

$$R_1 \cup R_2 \cup \ldots \cup R_p = U.$$

Mówimy, że relacja $r(U)$ spełnia zależność połączeniową, oznaczoną przez ***JD*** $[R_1, \ldots, R_p]$, jeżeli można ją zdekomponować bez utraty informacji na pod-relacje:

$$r_1(R_1), r_2(R_2), \ldots, r_p(R_p).$$

Zachodzi wówczas następujący warunek:

$$r(U) = r_1(R_1) \bowtie r_2(R_2) \bowtie \ldots \bowtie r_p(R_p).$$

Zależność połączeniowa $JD[R_1, R_2, \ldots, R_p]$ jest **trywialna,** jeżeli jeden ze schematów R_i, $i = 1,2,\ldots,p$, jest równy R.

Piąta postać normalna (5NF)

Schemat relacji R jest w piątej postaci normalnej (*5NF* lub *PJNF*), jeżeli dla każdej zależności połączeniowej *JD* w schemacie R zachodzi:

- zależność ta jest trywialna;
- każdy podschemat R_i, $i = 1,2,\ldots,p$ jest nadkluczem schematu R