GDAŃSK UNIVERSITY
OF TECHNOLOGY

Normalizacja baz danych

Marek Kulawiak, Krzysztof Goczyła

Wydział Elektroniki, Telekomunikacji i Informatyki

GDAŃSK UNIVERSITY OF TECHNOLOGY

Pojęcie normalizacji

- Normalizacja relacyjnych baz danych polega na modyfikacji struktury jej tabel w celu zlikwidowania nadmiarowości danych
- Redundancja danych może być przyczyną różnych **anomalii**:
  - **Anomalia aktualizacji** (*Update Anomaly*) zachodzi wtedy, gdy aktualizacja pewnej informacji musi być dokonywana w wielu miejscach jednocześnie, aby baza danych nie utraciła integralności
  - **Anomalia usuwania** (*Deletion Anomaly*) zachodzi wtedy, gdy usunięcie pewnych atrybutów powoduje jednocześnie utratę innych atrybutów z danej tabeli
  - **Anomalia wstawiania** (*Insertion Anomaly*) zachodzi wtedy, gdy dodanie nowego atrybutu wymaga istnienia innego atrybutu w bazie danych
- Normalizacja bazy danych dokonywana jest poprzez dostosowanie jej tabel do zestawu reguł zwanych postaciami normalnymi (ang. *Normal Forms*)
- W uproszczeniu, postacie normalne można rozumieć jako *dobre praktyki projektowania relacyjnych baz danych*

GDAŃSK UNIVERSITY OF TECHNOLOGY

Postacie normalne

- 1NF - wszystkie elementy są atomowe i żadne wiersze się nie powtarzają
- 2NF - jw., a oprócz tego atrybuty niekluczowe nie mogą być zależne od części klucza złożonego
- 3NF - jw., a dodatkowo wszystkie atrybuty niekluczowe zależą od klucza wyłącznie bezpośrednio (nie przechodnio)
- BCNF - jw., a do tego każdy atrybut, od którego w pełni zależy inny atrybut, jest kluczem kandydującym
- 4NF - jw., a także nie występują zależności wielowartościowe

GDAŃSK UNIVERSITY OF TECHNOLOGY

Pierwsza postać normalna (1NF)

- Aby dana tabela była w tzw. *pierwszej postaci normalnej* (ang. *First Normal Form*, w skrócie 1NF), musi ona spełniać następujące reguły:
  - Każdy jej element powinien reprezentować pojedynczą wartość (np. tylko jeden numer telefonu)
  - Nie mogą w niej występować dwa jednakowe wiersze
  - Wartości w danej kolumnie powinny być z tej samej domeny (tzn. powinny być tego samego typu/rodzaju)
  - Nie ma żadnego warunku narzucającego kolejność jej wierszy i kolumn

GDAŃSK UNIVERSITY OF TECHNOLOGY

Baza programistów (wariant I)

PROGRAMMERS

WorkerID	Name	ProgLang1	ProgLang2
1	Anon One	C++	Java
2	Anon Two	Javascript	Python

- Przechowywanie informacji o językach programowania w dwóch kolumnach
  - Kolumny *ProgLang1* i *ProgLang2* mają tę samą dziedzinę
  - Ograniczenie do dwóch języków na programistę
  - Niepotrzebne komplikacje przy wyszukiwaniu programistów konkretnego języka
  - Potencjalne narzucenie kolejności kolumn (nazwy języków posortowane alfabetycznie)

GDAŃSK UNIVERSITY OF TECHNOLOGY

Baza programistów (wariant II)

PROGRAMMERS

WorkerID	Name	ProgLang
1	Anon One	C++, Java
2	Anon Two	JavaScript, Python

- Przechowywanie informacji o wszystkich językach programowania w pojedynczej kolumnie
  - Naruszenie zasady o atomowości
  - Trudności w wydobyciu właściwych informacji
  - Przy założeniu że dziedziną dla *ProgLang* jest zestaw języków programowania, tabela teoretycznie jest w 1NF

GDAŃSK UNIVERSITY OF TECHNOLOGY

Baza programistów (1NF)

PROGRAMMERS

WorkerID	Name	ProgLang
1	Anon One	C++
1	Anon One	Java
2	Anon Two	JavaScript
2	Anon Two	Python

- Przechowywanie informacji o językach programowania w większej liczbie wierszy
  - Tabela jest w 1NF
  - Redundancja danych (powtarzanie wartości w kolumnie *Name*)

GDAŃSK UNIVERSITY OF TECHNOLOGY

Baza programistów (1NF)

PROGRAMMERS

WorkerID	Name
1	Anon One
2	Anon Two

PROG_LANGUAGES

WorkerID	ProgLang
1	C++
1	Java
2	JavaScript
2	Python

- Podział danych na dwie tabele
  - Obie tabele są w 1NF
  - Powtarzane są jedynie wartości w kolumnie *WorkerID* tabeli *PROG_LANGUAGES* (powtarzanie liczb zamiast ciągów znaków)

GDAŃSK UNIVERSITY OF TECHNOLOGY

Druga postać normalna (2NF)

- Aby tabela była była w *drugiej postaci normalnej* (2NF), musi ona spełniać dwie reguły:
  - musi ona być w pierwszej postaci normalnej (1NF)
  - wszystkie jej atrybuty niekluczowe muszą być w pełni zależne od każdego z kluczy kandydujących (jeśli w tabeli występują klucze złożone, to atrybuty niekluczowe muszą być zależne od całych kluczy, a nie tylko ich części)

GDAŃSK UNIVERSITY OF TECHNOLOGY

Baza kursantów

PARTICIPANTS

IDENT	NAME	CITY	INHAB	COURSE	GRADE
P1	Collins	London	8000000	English	A
P1	Collins	London	8000000	Geography	C
P1	Collins	London	8000000	Logics	A
P2	Jones	Glasgow	400000	Geography	B
P2	Jones	Glasgow	400000	Databases	C
P3	Rodin	Aberdeen	400000	Physics	B
P4	Thatcher	London	8000000	Logics	A
P4	Thatcher	London	8000000	Chemistry	C
P5	Biggs	Bristol	800000	Databases	A
P5	Biggs	Bristol	800000	English	A
P5	Biggs	Bristol	800000	Biology	A

GDAŃSK UNIVERSITY OF TECHNOLOGY

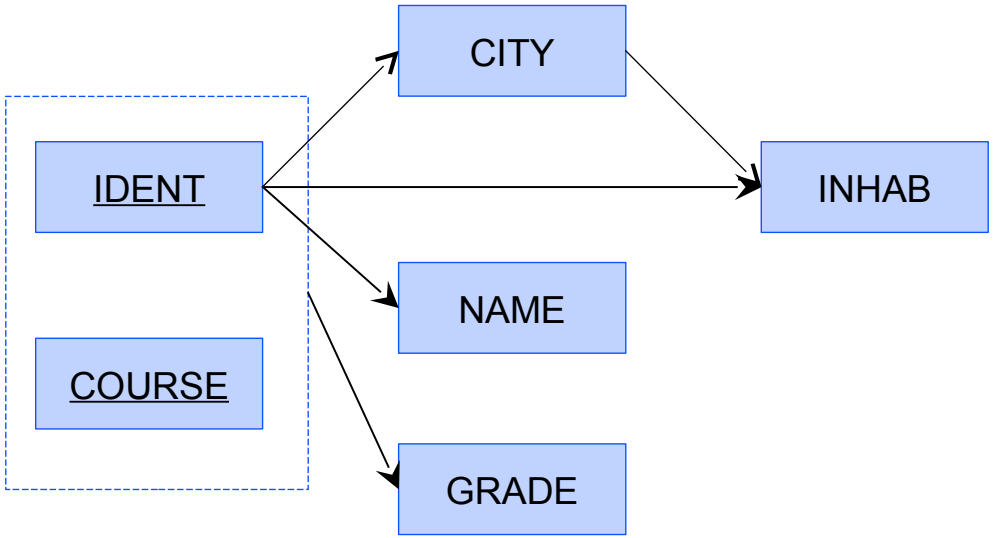
Baza kursantów

- Nadmiarowość danych w tabeli PARTICIPANTS może prowadzić do różnych problemów:
  - Aktualizacja liczby mieszkańców danego miasta musi być dokonywana w wielu różnych wierszach jednocześnie, aby baza nie utraciła integralności (**anomalia aktualizacji**)
  - W niektórych przypadkach, usunięcie informacji o danym kursancie może jednocześnie spowodować usunięcie informacji o liczbie mieszkańców miasta z którego on pochodzi (**anomalia usuwania**)
  - Wstawienie informacji o nowym kursie ukończonym przez danego kursanta wymaga powtórzenia informacji dotyczących miasta z którego on pochodzi (**anomalia wstawiania**)

GDAŃSK UNIVERSITY OF TECHNOLOGY

Zależności funkcyjne

- Zależność funkcyjna (ang. *functional dependency*):
  - Atrybut Y zależy funkcyjnie od atrybutu X (X→Y), gdy dla danej wartości X występuje dokładnie jedna wartość Y
- Pełna zależność funkcyjna:
  - Atrybut Y zależy w pełni funkcyjnie od atrybutu X (X→Y), gdy Y zależy funkcyjnie od całego X i nie zależy funkcyjnie od elementów X



Zależności funkcyjne:

IDENT → CITY

IDENT → NAME

IDENT → INHAB

CITY → INHAB

Pełna zależność funkcyjna:

(IDENT, COURSE) → GRADE

GDAŃSK UNIVERSITY OF TECHNOLOGY

Baza kursantów (2NF)

PART_DATA

IDENT	NAME	CITY	INHAB
P1	Collins	London	8000000
P2	Jones	Glasgow	400000
P3	Rodin	Aberdeen	400000
P4	Thatcher	London	8000000
P5	Biggs	Bristol	800000

PART_COURSE

IDENT	COURSE	GRADE
P1	English	A
P1	Geography	C
P1	Logics	A
P2	Geography	B
P2	Databases	C
P3	Physics	B
P4	Logics	A
P4	Chemistry	C
P5	Databases	A
P5	English	A
P5	Biology	A

- Podział danych na dwie tabele
  - Wyeliminowanie niepełnej zależności funkcyjnej

GDAŃSK UNIVERSITY OF TECHNOLOGY

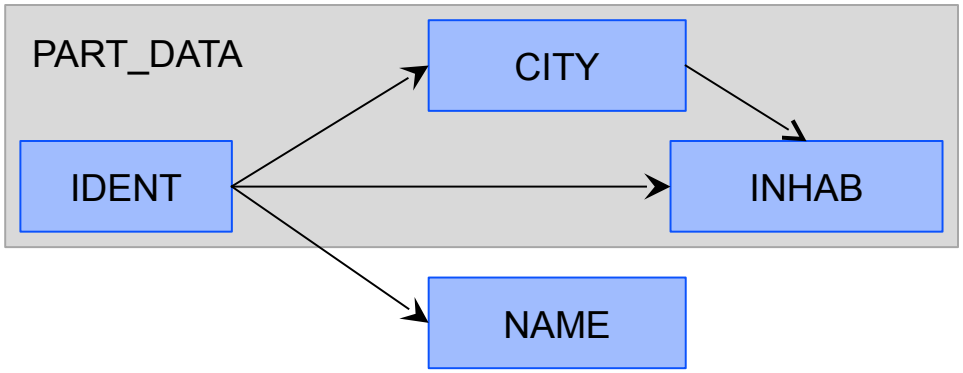
Trzecia postać normalna (3NF)

- Aby tabela była w *trzeciej postaci normalnej* (3NF), musi ona spełniać następujące reguły:
  - musi ona być w drugiej postaci normalnej (2NF), zatem musi ona spełniać także reguły 1NF
  - żaden atrybut niekluczowy nie może zależeć przechodnio od żadnego z kluczy kandydujących (tj. wszystkie atrybuty wtórne zależą bezpośrednio od wszystkich kluczy kandydujących)

GDAŃSK UNIVERSITY OF TECHNOLOGY

Zależności przechodnie

- Przechodnia zależność funkcyjna:
  - Atrybut Y zależy funkcyjnie od atrybutu X za pośrednictwem innego atrybutu Z



Zależność przechodnia:

IDENT → CITY → INHAB

GDAŃSK UNIVERSITY OF TECHNOLOGY

Baza kursantów (3NF)

PART_ID

IDENT	NAME	CITY
P1	Collins	London
P2	Jones	Glasgow
P3	Rodin	Aberdeen
P4	Thatcher	London
P5	Biggs	Bristol

CITIES

CITY	INHAB
London	8000000
Glasgow	400000
Aberdeen	400000
Bristol	800000

PART_COURSE

IDENT	COURSE	GRADE
P1	English	A
P1	Geography	C
P1	Logics	A
P2	Geography	B
P2	Databases	C
P3	Physics	B
P4	Logics	A
P4	Chemistry	C
P5	Databases	A
P5	English	A
P5	Biology	A

GDAŃSK UNIVERSITY OF TECHNOLOGY

Postać normalna Boyce'a Codda

- Aby tabela była w *postaci normalnej Boyce'a Codda* (BCNF), musi ona spełniać następujące reguły:
  - musi ona być w trzeciej postaci normalnej (3NF)
  - każdy atrybut w tabeli, od którego w pełni funkcyjnie zależy inny atrybut, musi być kluczem kandydującym

GDAŃSK UNIVERSITY OF TECHNOLOGY

Czwarta postać normalna (4NF)

- Aby tabela była w *czwartej postaci normalnej* (4NF), musi ona spełniać następujące reguły:
  - musi ona być w trzeciej postaci normalnej (3NF) lub w postaci normalnej Boyce'a Codda (BCNF)
  - nie może ona zawierać zależności wielowartościowych

GDAŃSK UNIVERSITY OF TECHNOLOGY

Zależności wielowartościowe

PERSONS (1)

SSN	LANGUAGE	SPORT
P1	English	football
P1	Finnish	football
P1	French	football
P1	English	skiing
P1	Finnish	skiing
P1	French	skiing

PERSONS (2)

SSN	LANGUAGE	SPORT
P1	English	football
P1	Finnish	football
P1	French	skiing

- Obie tabele są w 3NF
  - W pierwszej tabeli występuje znaczna redundancja
  - Druga tabela jest podatna na anomalie
- Zależność wielowartościowa:
  - Dla każdej wartości X istnieje zbiór możliwych wartości Y i zbiór ten nie zależy od Z ($X \rightarrow\!\!\rightarrow Y$)
- SSN $\rightarrow\!\!\rightarrow$ LANGUAGE
  - (znajomość języków nie zależy od uprawianych sportów)
- SSN $\rightarrow\!\!\rightarrow$ SPORT
  - (uprawiane sporty nie zależą od znajomości języków)

GDAŃSK UNIVERSITY OF TECHNOLOGY

Przykład bazy w 4NF

P_LANG

SSN	LANGUAGE
P1	English
P1	Finnish
P1	French

P_SPORT

SSN	SPORT
P1	football
P1	skiing

- Problem zależności wielowartościowych, podobnie jak pozostałe problemy, został rozwiązany drogą dekompozycji na dwie osobne tabele