

# Relacyjne bazy danych. Normalizacja i problem nadmierności danych.

Robert A. Kłopotek

[r.klopotek@uksw.edu.pl](mailto:r.klopotek@uksw.edu.pl)

Wydział Matematyczno-Przyrodniczy. Szkoła Nauk Ścisłych, UKSW

# Relacyjne bazy danych

- ▶ Stworzone przez E.F. Codd w 1970 i rozwijane przez ponad 20 lat
- ▶ Bazujące na matematycznej teorii precyzji mającej na celu:
  - ▶ Zwiększenie produktywności wytwarzania oprogramowania
  - ▶ Unikanie niejednoznaczności poprzez przygotowania precyzyjnych specyfikacji
  - ▶ Zapewnienie zwiększonej niezależności pomiędzy programami oraz ich danymi

# Baza danych (raz jeszcze)

- ▶ Definicja: zbiór struktur danych służący do organizacji i przechowywania danych
- ▶ Główna koncepcja - relacja, jest zdefiniowana poprzez następujące cechy
  - ▶ Każda relacja (tabela) posiada unikalną nazwę w ramach bazy danych
  - ▶ Każda kolumna posiada unikalną nazwę w ramach relacji
  - ▶ Wszystkie wartości w kolumnie są tego samego typu
  - ▶ Porządek kolumn w tabeli nie jest istotny (choć w praktyce jest istotny podczas wstawiania implicite - implicit insert statement)

# Relacja

## ► Pozostałe cechy relacji

- Każdy rekord w relacji musi być inny tzn. nie może być wielu takich samych rekordów
- Porządek rekordów nie jest istotny
- Dopuszczalne są tylko wartości atomowe dla każdej z kolumn. Nie dozwolone są zbiory wartości.

| Budynek | Piętro | Sala |
|---------|--------|------|
| WOY/21  | 0      | 033  |
| WOY/21  | 1      | 134  |
| ...     |        | ...  |
| WOY/12  | 1      | 1221 |

# Klucze główne

- ▶ Nie może być wielu powtórzeń tego samego rekordu => musi istnieć zbiór kolumn takich, że ich wartości są unikalne dla całej tabeli
- ▶ Taki zbiór nazywamy kluczem głównym (PK - primary key)

| Budynek | Piętro | Sala |
|---------|--------|------|
| WOY/21  | 0      | 033  |
| WOY/21  | 1      | 134  |
| ...     |        | ...  |
| WOY/12  | 1      | 1221 |



klucz główny (PK - primary key)

# Klucze kandydujące

| Imię | Nazwisko | PESEL       | NIP           |
|------|----------|-------------|---------------|
| Jan  | Kowalski | 45012300561 | 521-125-33-12 |
| Anna | Nowak    | 75011202561 | 521-125-33-14 |
| ...  | ...      | ...         | ...           |
|      |          | 88122300561 | 521-125-01-22 |

klucz główny

klucz kandydujący

# Klucze złożone

- ▶ Klucz złożony - klucz składający się z więcej niż jednej kolumny
- ▶ Ponieważ klucz musi identyfikować rekord w sposób jednoznaczny, więc wartości dla kolumn klucza muszą być określone
- ▶ W poprzednim przykładzie: PESEL i NIP nie mogą być puste

| Budynek | Piętro | Sala |
|---------|--------|------|
| WOY/21  | 0      | 033  |
| WOY/21  | 1      | 134  |
| ...     |        | ...  |
| WOY/12  | 1      | 1221 |

złożony klucz główny

# Wartość NULL

- ▶ Jest używana do określenia/anotowania wartości „nieznanej” (unknown) lub „niedostępnej” (NA - not available)
- ▶ Przykład: data sprzedaży produktów, które nie zostały jeszcze sprzedane będzie ustawiona na NULL. Nie ma innej prawidłowej daty, którą można uzyc w tym kontekście.
- ▶ ALE:
  - ▶ Użycie wartości NULL zmienia logikę operacji na danych. Każde porównanie z wartością NULL zwróci fałsz.

# NULL i porównania

- ▶ Wyobraźmy sobie następujący przykład
  - ▶ Niech będzie 25 sprzedawców, którzy wygenerowali przychód w wysokości co najmniej 1000 zł
  - ▶ Niech będzie 15 sprzedawców, którzy wygenerowali przychód w wysokości nie przekraczającej 1000 zł
  - ▶ Co jeśli są sprzedawcy, którzy mają przychód równy NULL?
  - ▶ Oni nie są uwzględnieni w żadnej z wcześniejszych grup!
  - ▶ **Nie możemy tu używać już logiki dwuwartościowej prawda/fałsz!**

# NULL - wskazówki

- ▶ NULL to nie jest puste pole np. 0 lub ,'
- ▶ NULL nie powinien być używany kiedy wymagamy pustej wartości
- ▶ Kolumny nie powinny akceptować wartości NULL o ile nie ma dobrego uzasadnienia

| Imię | Nazwisko | PESEL       | NIP           |
|------|----------|-------------|---------------|
| Jan  | Kowalski | 45012300561 | 521-125-33-12 |
| Anna | Nowak    | 75011202561 | 521-125-33-14 |
| ...  | ...      | ...         | ...           |
|      |          | 88122300561 | 521-125-01-22 |

Wszystkie te kolumny nie powinny pozwalać na wprowadzanie wartości NULL

# Projekt bazy danych

- ▶ Główne założenie - systemy bazodanowe jak każdy inny system powinien być traktowany jako inwestycja. System komputerowy może generować przychód lub zmniejszać koszty poprzez
  - ▶ Dostarczanie informacji dla bardziej efektywnych/skutecznych decyzji, tzn. generowanie większego przychodu lub minimalizacji kosztów
  - ▶ Optymalizację procesów biznesowych, tzn. minimalizować koszty przeprowadzania tego procesu
- ▶ Wśród wyboru projektów (oraz metod uzasadnienia)
  - ▶ ROI (Return on Investment)
  - ▶ NPV (Net Present Value) - metoda formalna obliczania przychodu wygenerowanego z projektu oprogramowania

Powyższe czynniki grają główną rolę podczas ewaluacji zysków z projektu.

# Koncepcja projektu bazy danych

## - krok 1

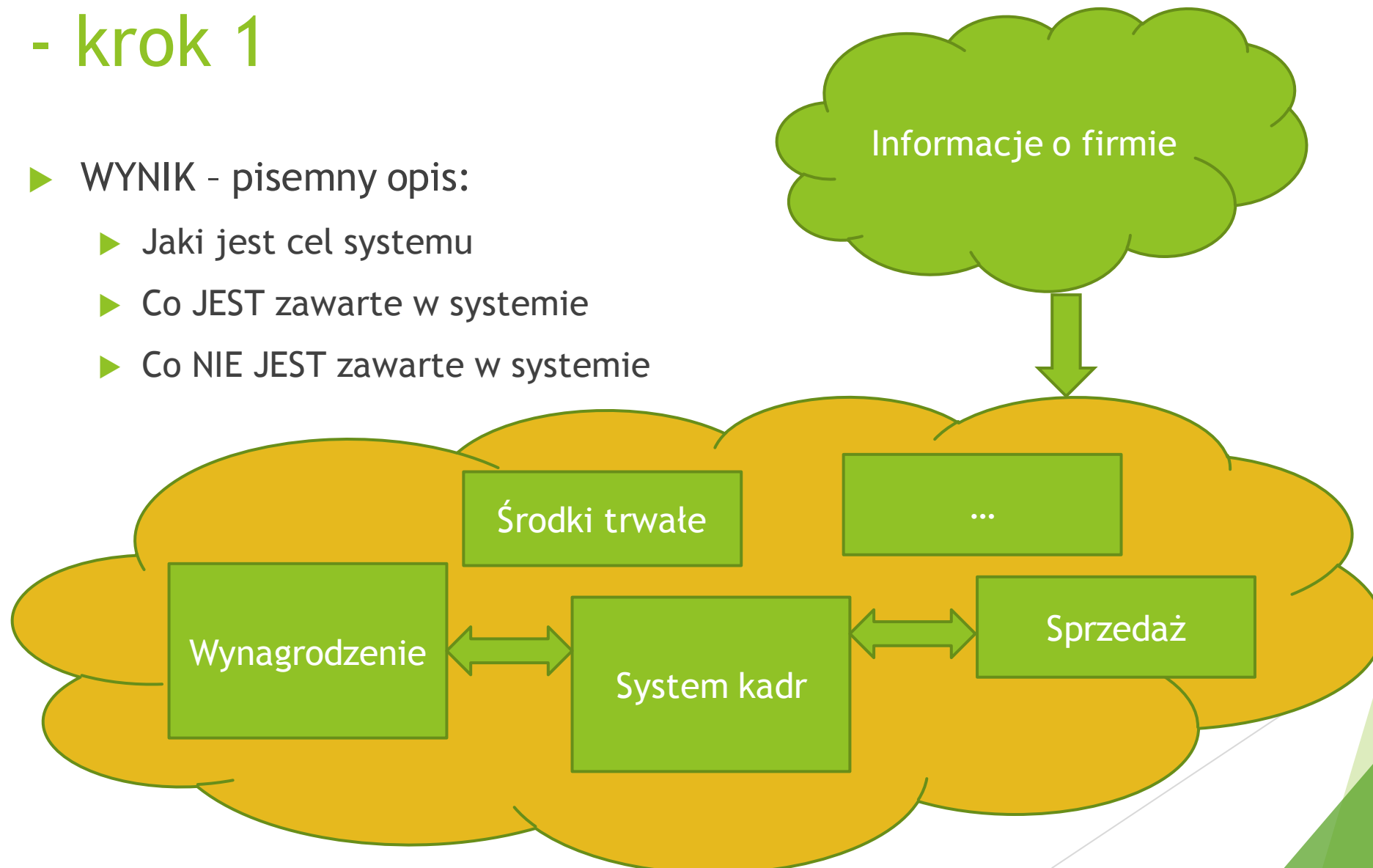
- ▶ CEL: projekt relacji relacyjnego systemu bazodanowego dla analizowanego problemu
- ▶ KROK 1:
  - ▶ Identyfikacja zakresu systemu oraz ograniczeń systemu
  - ▶ Przykład: system dla zarządzania działem kadr nie zawierałaby listy środków trwałych w organizacji/firmie

# Koncepcja projektu bazy danych

## - krok 1

► WYNIK - pisemny opis:

- Jaki jest cel systemu
- Co JEST zawarte w systemie
- Co NIE JEST zawarte w systemie



# Koncepcja projektu bazy danych

## - krok 2

### ► KROK 2:

- Identyfikacja istotnych procesów biznesowych, który będzie wspierany przez system
- Pamiętaj: te procesy muszą być zoptymalizowane przy użyciu oprogramowania
- Przykład:
  - Proces generowania miesięcznych raportów sprzedaży
  - Dzięki systemowi bazodanowemu:
    - Nie jest potrzebne przygotowywanie ręczne (zredukowane koszty procesu)
    - Informacja jest dostępna na bieżąco, więc mogą być podejmowane lepsze decyzje

# Koncepcja projektu bazy danych

## - krok 3

### ► KROK 3:

- Określenie zasad biznesowych oraz regulacji

- Przykład:

- Jakie informacje mają być drukowane na fakturach?
- Kto może zmieniać wynagrodzenia?
- Czy potrzebna jest podwójna autoryzacja?
- ...

# Koncepcja projektu bazy danych

## - krok 4

### ► KROK 4:

- Identyfikacja obiektów/encji w organizacji/firmie
- Porady i wskazówki
  - Użyj podejść odgórnego (top-bottom approach) lub oddolnego (bottom-up approach) aby zidentyfikować encje takie jak: faktury, produkty, dokumenty magazynowe, itp.
  - Używaj określeń z danej firmy (wspólny słownik pojęć), zidentyfikuj „obiekty pracowników” tzn. obiekty, które są znane pracownikom (przyszłym użytkownikom) takie jak: faktura, zamówienie, dokument zakupu, ...
- WYNIK: lista encji (możliwe późniejsze relacje znane jako tabele)

# Koncepcja projektu bazy danych

## - krok 5

### ▶ KROK 5 - normalizacja:

- ▶ Jest to proces projektowania relacji tak, aby
  - ▶ Eliminować powtórzenia w tabelach
  - ▶ Unikać możliwych anomalii podczas modyfikacji
  - ▶ Przedefiniować encje oraz ich atrybuty
- ▶ Procedura normalizacji ma wpływ na atrybuty oraz ilość tabel

# Normalizacja - 1NF

- ▶ Pierwsza forma normalna - 1NF:
  - ▶ Encja jest w pierwszej formie normalnej, jeśli jest relacją, czyli
    - ▶ Nie ma powtórzeń atrybutów
    - ▶ Nieatomowe wartości muszą być wyeliminowane
    - ▶ Muszą być obecne klucze główne oraz muszą być dobrze określone
    - ▶ ...

# Procedura 1NF

Sale

| Budynek | Piętro | Wyposażenie                                                            | Sala |
|---------|--------|------------------------------------------------------------------------|------|
| WOY/21  | 0      | <ul style="list-style-type: none"><li>Krzeseł</li><li>Stolik</li></ul> | 033  |
| WOY/21  | 1      |                                                                        | 134  |
| ...     |        |                                                                        | ...  |
| WOY/12  | 1      |                                                                        | 1221 |

klucz główny (PK - primary key)

Sale

| Budynek | Piętro | Sala |
|---------|--------|------|
| WOY/21  | 0      | 033  |
| WOY/21  | 1      | 134  |
| ...     |        | ...  |
| WOY/12  | 1      | 1221 |

Wyposażenie

| Budynek | Zasób   | Sala |
|---------|---------|------|
| WOY/21  | Krzeseł | 033  |
| WOY/21  | Stolik  | 033  |

klucz główny (PK - primary key)

# Normalizacja - 2NF

- ▶ Druga forma normalna - 2NF:
  - ▶ Encja jest w drugiej formie normalnej, jeśli:
    - ▶ Jest w 1NF
    - ▶ Atrybuty, które nie są kluczem zależą od całego klucza głównego
  - ▶ Kontrprzykład
    - ▶ Tabela zamówień zawierająca: customer\_id, order\_id, customer\_city, customer\_country. Jeśli (customer\_id, order\_id) jest kluczem głównym (PK), to customer\_city zależy tylko od części klucza głównego (customer\_id).

# Normalizacja - 2NF

## ► Kontrprzykład - dyskusja

- Jeśli zasada 2NF jest naruszona wtedy część każdego rekordu nie zależy tylko od wpisu. W tym przypadku zależy tylko od klienta, a nie zależy od zamówienia.
- Jako konsekwencja tego, wiele rekordów zawiera tą samą informację, np. adres danego klienta (nadmiar)
- Nadmiar może doprowadzić do niejednoznaczności - co się stanie jak zrobimy aktualizację adresu klienta tylko w jednym rekordzie opisującym klienta?

# Normalizacja - 3NF

- ▶ Trzecia forma normalna - 3NF:
  - ▶ Encja jest w 3 formie normalnej, jeśli:
    - ▶ Jest w 2NF
    - ▶ Atrybuty, które nie są kluczami, są niezależne od siebie. Zależą jedynie od klucza głównego.
  - ▶ Kontrprzykład
    - ▶ Tabela zamówień zawiera: `customer_id`, `order_id`, `customer_city`, `customer_country`, `employee_id`, `employee_address`.
    - ▶ Podczas gdy `employee_id` zależy tylko od zamówienia to `employee_address` stwarza problem.

# Normalizacja - 3NF

## ► Kontrprzykład - dyskusja

- Jeśli zasada 3NF jest naruszona wtedy część każdego rekordu nie zależy tylko i wyłącznie od wpisu. W tym przypadku `employee_adres` zależy raczej od pracownika niż od zamówienia.
- Jako konsekwencja tego, wiele rekordów zawiera tą samą informację, np. adres danego pracownika (nadmiar)
- Nadmiar może doprowadzić do niejednoznaczności - co się stanie jak zrobimy aktualizację adresu pracownika tylko w jednym rekordzie opisującym zamówienie?

# Normalizacja - 4NF

- ▶ MVD (multivalued dependency) - zależność multiwartościowa, występuje gdy wartość atrybutu A jest zawsze powiązana z tym samym zbiorem wartości z atrybutu B
- ▶ Encja jest w 4NF, jeśli:
  - ▶ Jest w 3NF
  - ▶ Nie ma nakładających się na siebie kluczy kandydujących, czyli klucze kandydujące nie zawierają atrybutów klucza głównego
  - ▶ Nie ma niezależnych multiwartościowo atrybutów w encji, np. atrybut A determinuje możliwe wartości atrybutu B, A determinuje możliwe wartości atrybutu C, ale atrybuty C i B są niezależne)