Bazy danych
i systemy zarządzania bazami

Podstawy formalne, elementy teorii, operacje algebraiczne, SQL, normalizacja, projektowanie i analiza

Antoni Ligęza

Program wykładów

1. Wprowadzenie do problematyki baz danych.
2. Matematyczne podstawy relacyjnych baz danych.
3. Relacyjny model danych.
4. Podstawowe operacje algebraiczne w bazach danych.
5. Specjalizowane operacje algebraiczne w bazach danych.
6. Podstawy języka SQL.
7. Realizacja operacji algebry relacji w języku SQL.
8. Diagramy związków encji (ERD).
9. Diagramy przepływu danych (DFD).
10. Zależności funkcyjne w bazach danych.
11. Rozkład relacji i normalizacja.
12. Elementy projektowania relacyjnych baz danych (*case-study*).
13. Datalog, dedukcyjne bazy danych, Prolog a bazy danych.
14. Problemy i ograniczenia; tendencje rozwojowe baz danych.
15. Elementy analizy i weryfikacja własności.

©Antoni Ligęza

Wprowadzenie do problematyki baz danych

Podstawowe informacje o bazach danych

Bazy danych (BD) służą do przechowywania dużych zbiorów danych w formie elektronicznej; zapewniają szybki i wygodny dostęp do danych oraz możliwość ich przetwarzania. W myśl nieformalnej definicji, *baza danych* (ang. *Data Base*, *Database*, *DB*) jest zbiorem danych (zazwyczaj o dużych rozmiarach), posiadającym określoną struktrę wewnętrzną, przechowywanych przez pewien, zazwyczaj długi, okres czasu na specjalnym nośniku. Bazy danych tworzy się w celu efektywnego wykorzystywania dużych ilości danych.

Do zapewnienia obsługi bazy danych (rozumianej jako zbiór danych) wykorzystuje się *system zarządzania bazą danych* (SZBD) (ang. *Database Management System*, *DBMS*). System taki zapewnia możliwość tworzenia (projektowania) nowych baz danych, komunikowania się z bazą danych w celu wyszukiwania informacji, oraz administrowania danymi i dostępem do nich.

Systemy baz danych stanowią istotny obszar zastosowań systemów informatycznych; spotykane są tam, gdzie istnieje potrzeba zarządzania i zapewnienia efektywnego dostępu do dużych zbiorów danych. Typowe przykładowe zastosowania obejmują:

- systemy obsługi bibliotek,
- systemy rezerwacji miejsc (lotniczych, w pociągach, etc.),
- systemy bankowe,
- systemy wspomagające działanie przedsiębiorstw (kadry, płace, gospodarka materiałowa, kontrahenci, ewidencja zakupów i sprzedaży, etc.),
- systemy ewidencji ludności i pochodne (ZUS, urzędy skarbowe, rejestracja poborowych, etc.).

©Antoni Ligęza

Bazy danych – podstawowe definicje

Informacja i dane

Podstawowa terminologia informatyczne w języku polskim zdefiniowana jest w normie PN-ISO/IEC 2382-1:1996. Zgodnie z powyższą normą obowiązują m.in następujące definicje:

Definicja 1 Informacja (01.01.01): *Wiedza dotycząca obiektów, takich jak fakty, zdarzenia, przedmioty, procesy lub idee, zawierająca koncepcje, która w określonym kontekście ma określone znaczenie.*

Definicja 2 Dane (01.01.02): *Reprezentacja* informacji *mająca interpretację, właściwa do komunikowania się, interpretacji lub przetwarzania.*

Definicja 3 Przetwarzanie danych; automatyczne przetwarzanie danych (01.01.06): *Uporządkowane wykonywanie* operacji *na* danych.

Przykładem są operacje arytmetyczne lub operacje logiczne, łączenie lub sortowanie danych, asemblowanie lub kompilowanie programów, lub operacje na tekście, takie jak redagowanie, sortowanie, łączenie, zapamiętywanie, wyszukiwanie, wyświetlanie, lub drukowanie. Terminu *przetwarzanie danych* nie należy używać jako synonimu *przetwarzania informacji*, które jest pojęciem o szerszym znaczeniu i może obejmować również operacje takie jak przesyłanie danych i automatyzację prac biurowych.

Bazy danych

Definicja 4 Baza danych (01.08.05): *Zbiór* danych *zorganizowany zgodnie z pojęciową strukturą opisującą charakterystyki tych danych oraz związki między ich elementami, stosowany w jednym lub wielu zastosowaniach.*

©Antoni Ligęza

Bazy danych – analiza definicji

Komentarz do definicji bazy danych

Zwróćmy uwagę na najbardziej istotne elementy definicji bazy danych:

- **Zbiór;** *Baza danych* to *zbiór danych*, przy czym przeważnie jest to *duży* zbiór zawierający istotne dane (dane posiadające wartość z punktu widzenia użytkowanika, dane użyteczne),
- **Struktura;** Zbiór ten jest zorganizowany zgodnie z pewną *strukturą pojęciową*, a więc zbiór ten musi posiadać *strukturę wewnętrzną*, zgodną z wybranym *modelem danych*,
- **Charakterystyki;** Struktura danych musi pozwalać na reprezentowanie istotnych dla użytkownika *charakterystyk danych*, a zatem obiekty reprezentowane w bazie danych będą opisywane za pomocą wartości wybranych własności (wartości *atrybutów*),
- **Związki;** Struktura bazy danych musi zapewniać reprezentację *związków* (powiązań) pomiędzy danymi i/lub ich elementami,
- **Użyteczność;** Dane muszą być użyteczne, znajdować zastosowanie praktyczne, przy czym *te same dane* mogą być stosowane w *jednym* lub w *wielu* zastosowaniach, t.j. mogą być wykorzystywane przez wielu użytkowników.

Ponadto przyjmuje się, że struktura bazy danych powinna:

- zapewniać **selektywny dostęp** do danych,
- umożliwiać zapis danych na **zewnętrznym nośniku,**
- **System Zarządzania Bazą Danych;** z uwagi na specjalną strukturę zbioru bazy danych konieczny jest SZBD do jej obsługi.

©Antoni Ligęza

System Zarządzania Bazą Danych

Definicja SZBD

Baza danych stanowi jedynie odpowiednio skonstruowany *zbiór danych*; wykorzystanie tych danych jest możliwe dzięki narzędziom pozwalającym na efektywne *zarządzanie danymi*.

Definicja 5 Zarządzanie danymi (01.08.02): *Funkcje zapewniające* dostęp *do* danych, *wykonujące lub kontrolujące czynność* zapamiętywania *danych oraz sterujące operacjami* wejścia-wyjścia *w* systemie przetwarzania danych.

System Zarządzania Bazą Danych to odpowiednio zorganizowane, specjalizowane oprogramowanie narzędziowe pozwalające na realizację istotnych dla użytkownika operacji na danych.

Operacje na danych

Podstawowe operacje na danych obejmują:

- wprowadzanie danych,
- zapamiętywanie i przechowywanie danych,
- wyszukiwanie i prezentację danych,
- dodawanie i usuwanie danych,
- aktualizację danych,
- przetwarzanie arytmetyczne, statystyczne, algebraiczne i logiczne,
- operacje teoriomnogościowe i operacje algebry relacji,
- kodowanie i dekodowanie danych.

©Antoni Ligęza

Zadania Systemu Zarządzania Bazą Danych

Zadania SZDB

1. Umożliwienie *projektowania* i *implementacji* nowej bazy danych przy użycie narzędzi i języka definicji danych (ang. *Data Definition Language, DDL*).
2. Umożliwienie *selektywnego dostępu* do danych za pomocą języka zapytań i tworzonych w nim *kwerend*.
3. Umożliwienie wykonywania określonych operacji na danych przy pomocy języka operowania na danych (ang. *Data Manipulation Language, DML*).
4. Obsługa przechowywania dużych zbiorów danych, zapewnienie *niezawodności* oraz *efektywności*.
5. Zapewnienie *integralności* danych (wewnętrznej i zewnętrznej; lokalnej i globalnej).
6. Ochrona dostępu do danych, zapewnienie różnych obszarów i poziomów dostępu (hasła, przypisywanie uprawnień; *ochrona i poufność* danych).
7. Zapewnienie dostępu dla wielu użytkowników (wielodostępu) oraz synchronizacja dostępu w przypadku dostępu współbieżnego.
8. Zapewnienie możliwości komunikacji z innymi systemami.
9. Dostarczanie opisu i dokumentacji (schematu i struktury).
10. Optymalizacja pracy (minimalizacja czasu dostępu lub obsługi żądań, optymalizacja dostępu dla poszczególnych użytkowników (perspektywy), optymalizacja gospodarki zasobami i organizacji bazy danych.

©Antoni Ligęza

Celowość budowy baz danych

Cele realizowane przez bazy danych

1. **Centralizacja:** Skupienie danych w jednym systemie bazy danych (umożliwia pełny dostęp do wszystkich danych).
2. **Selektywny dostęp:** Zapewnienie łatwości wyszukiwania i dostępu.
3. **Jednokrotna reprezentacja:** Zmniejszenie redundancji (nadmiarowości), unikanie powtórzeń.
4. **Zachowanie spójności:** ułatwienie zapewnienia spójności (wewnętrznej i zewnętrznej) dzięki odpowiedniej strukturze i narzędziom oraz *metodologii projektowania.*
5. **Standaryzacja:** Zapewnienie jednolitego formatu danych, ułatwienie wymiany i dostępu.
6. **Wielodostęp:** Umożliwienie użytkowania tych samych danych przez wielu użytkowników.
7. **Perspektywy:** Umożliwienie użytkownikom korzystania z danych w dogodnej dla nich postaci.
8. **Dostęp równoległy:** Umożliwienie jednoczenego operowania na tych samych danych, zapewnienie synchronizacji, blokowania dostępu, poprawnej realizacji operacji.
9. **Niezależność danych:** Niezależność *fizyczn*a (od sposobu przechowywania i dostępu) oraz *logiczna* (od aplikacji).
10. **Optymalizacja:** Optymalizacja zajętości pamięci i dostępu, optymalizacja przetwarzania (efektywność, wygoda).

©Antoni Ligęza

Modele baz danych, języki, narzędzia

Modele baz danych

Znane są różne typy baz danych; ich klasyfikacja wynika z przyjętego *modelu reprezentacji danych*, a więc z wewnętrznej struktury:

- systemy plików,
- hierarchiczne bazy danych,
- sieciowe bazy danych,
- bazy danych multimedialne,
- **relacyjne bazy danych**,
- obiektowe bazy danych,
- dedukcyjne bazy danych,
- dynamiczne, temporalne, reaktywne

Języki zapytań

Najczęściej spotykane języki zapytań obejmują:

- *Query by Example* (QBE), szablony (formularze, WWW),
- *Structured Query Language* (SQL), języki algebraiczne,
- języki predykatowe (o zmiennych atrybutowych i krotkowych),
- DATALOG (PROLOG bez termów).

Systemy zarządzania bazami danych

Najbardziej powszechne są systemy relacyjnych baz danych: ORACLE, INFORMIX, SYBASE. W pakiecie MS Office: ACCESS. Inne sytemy to: DBase II, III, IV, FoxBase, Clipper, Paradox.Ostatnio: PostgreSQL, ADABAS D.

©Antoni Ligęza

Użytkownicy baz danych

Rodzaje, uprawnienia i zadania użytkowników

Istnieją różne rodzaje użytkowników baz danych:

- **użytkownicy zwykli z prawem odczytu danych:** użytkownicy ci mają prawo do wyszukiwania i odczytu informacji; dostęp może być warunkowany *poziomem uprawnień*,
- **użytkownicy zwykli z prawem odczytu i modyfikacji danych:** użytkownicy ci mają prawo odczytu i modyfikacji danych (kasowania, modyfikacji, wprowadzania); dostęp może być warunkowany *poziomem uprawnień*,
- **administratorzy baz danych:** zarządzają dostępem do baz danych i przydzielają uprawnienia użytkownikom zwykłym; dbają o ochronę dostępu i bezpieczeństwo (replikacja danych), zarządzają konfiguracją systemu (logiczną, fizyczną),
- **projektanci baz danych:** projektują schemat bazy danych (tabele, zapytania, relacje, formularze, raporty), definiują perspektywy użytkowników, określają strukturę aplikacji, interfejsy użytkowników,
- **analitycy (baz) danych:** badają własności danych i wyznaczają charakterystyki danych, wykrywają zależności i cechy jakościowe, realizują wspomaganie decyzji,
- **inżynierowie wiedzy:** wydobywają wiedzę z danych (*data mining*, *knowledge discovery*, *rule induction*),
- **Kierownicy, managerowie, dyrektorzy:** To co chcą, chociaż nie zawsze wiedzą czego chcą i co jest możliwe do realizacji.

©Antoni Ligęza

Podstawowe założenia RBD – Postulaty Codda

1. **Postulat informacyjny**. Dane są reprezentowane *jedynie* porzez wartości atrybutów w wierszach tabel.
2. **Postulat dostępu**. Każda wartość jest dostępna poprzez podanie tabeli, atrybutu i klucza.
3. **Postulat dotyczący wartości NULL**. Dostępna jest specjalna wartość NULL dla reprezentacji wartości nieokreślonej, inna od wszystkich, i podlegająca prztwarzaniu.
4. **Postulat dotyczący katalogu**. Struktura bazy danych jest dostępna w katalogu będącym relacyjną bazą danych.
5. **Postulat języka danych**. System musi dostarczać pełnego języka przetwarzania danych (interakcja, aplikacje,definicje, przetwarzanie).
6. **Postulat modyfikowalności perspektyw**. System musi umożliwiać modyfikowanie perpektyw, o ile jest ono semantycznie realizowalne.
7. **Postulat modyfikowalności danych**. System musi umożliwiać operacje modyfikacji danych (INSERT, UPDATE, DELETE).
8. **Postulat fizycznej niezależności danych**. Zmiany fizycznej reprezentacji danych i organizacji dostępu nie wpływają na aplikacje.
9. **Postulat logicznej niezależności danych**. Zmiany wartości w tabelach nie wpływają na aplikacje.
10. **Postulat niezależności więzów spójności**. Więzy spójnści są definiowalne w bazie i nie zależą od aplikacji.
11. **Postulat niezależności dystrybucyjnej**. Działanie aplikacji nie zależy od modyfikacji dystrybucji bazy.
12. **Postulat bezpieczeństwa względem operacji niskiego poziomu**. Operacje niskiego poziomu (na poziomie rekordu) nie mogą naruszać modelu relacyjnego i więzów spójności.

©Antoni Ligęza

Idea relacyjnego modelu danych

Reprezentacja tablicowa

R	A_1	A_2	$\ldots$	A_j	$\ldots$	A_n
e_1	$d_{1,1}$	$d_{1,2}$	$\ldots$	$d_{1,j}$	$\ldots$	$d_{1,n}$
e_2	$d_{2,1}$	$d_{2,2}$	$\ldots$	$d_{2,j}$	$\ldots$	$d_{2,n}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
e_i	$d_{i,1}$	$d_{i,2}$	$\ldots$	$d_{i,j}$	$\ldots$	$d_{i,n}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
e_m	$d_{m,1}$	$d_{m,2}$	$\ldots$	$d_{m,j}$	$\ldots$	$d_{m,n}$

(1)

Projekt schematu tabeli

Atrybut	Typ	Opis
A_1	$type\_1$	opis atrybutu A_1
A_2	$type\_2$	opis atrybutu A_2
$\vdots$	$\vdots$	$\vdots$
A_j	$type\_j$	opis atrybutu A_j
$\vdots$	$\vdots$	$\vdots$
A_n	$type\_n$	opis atrybutu A_n

(2)

©Antoni Ligęza

Przykład tabel relacyjnej bazy danych

Pracownicy

ID_prac	Nazwisko	Imię	Data ur	Stanowisko	Dział	Stawka
MT101	Abacki	Adam	61-01-01	robotnik	P10	550,00 zł
MT102	Abakowski	Alojzy	61-01-02	robotnik	P10	574, 00 zł
MT103	Adamski	Antoni	61-01-03	robotnik	P20	1275,00 zł
KT101	Aron	Antonina	61-01-03	robotnik	P10	575,00 zł
MU101	Batman	Bogusław	67-02-13	kierownik	P30	1224,00 zł
KU101	Celińska	Mirosława	69-03-08	analityk	F10	975,00 zł
MV101	Dioniziak	Dariusz	71-10-17	v-prezes	V	3000,00 zł

Działy

ID_działu	Nazwa	Lokalizacja	Opis
P10	Produkcyjny 10	LB101	produkcja kasztanów
P20	Prodykcyjny 20	LB202	produkcja opakowań
P30	Produkcyjny 30	ZA303	produkcja sznurka
P40	Produkcyjny 40	ZB404	produkcja tajna
F10	Finansowy 10	LA101	prowadzenie finansów
K10	Kadry 10	LA102	sprawy kadrowe
V	VIP-y	LA007	kierownictwo

Płace

ID_prac	Miesiąc	Dni przepracowane	Wynagrodzenie
MT101	styczeń	13	7150,00 zł
MT101	luty	7	3850,00 zł
MV101	styczeń	33	*****

©Antoni Ligęza

Przykład: tablica decyzyjna dla optyków

Number	Age	Spectacle	Astigmatic	Tear p.r.	Decision
1	y	m	y	n	H
2	y	n	y	n	H
3	p	m	y	n	H
4	q	m	y	n	H
5	y	m	n	n	S
6	y	n	n	n	S
7	p	m	n	n	S
8	p	n	n	n	S
9	q	n	n	n	S
10	y	m	n	r	N
11	y	m	y	r	N
12	y	n	n	r	N
13	y	n	y	r	N
14	p	m	n	r	N
15	p	m	y	r	N
16	p	n	n	r	N
17	p	n	y	r	N
18	p	n	y	n	N
19	q	m	n	r	N
20	q	m	n	n	N
21	q	m	y	r	N
22	q	n	n	r	N
23	q	n	y	r	N
24	q	n	y	n	N

©Antoni Ligęza

Uwagi o realizacji struktury baz danych

Architektura klient-serwer

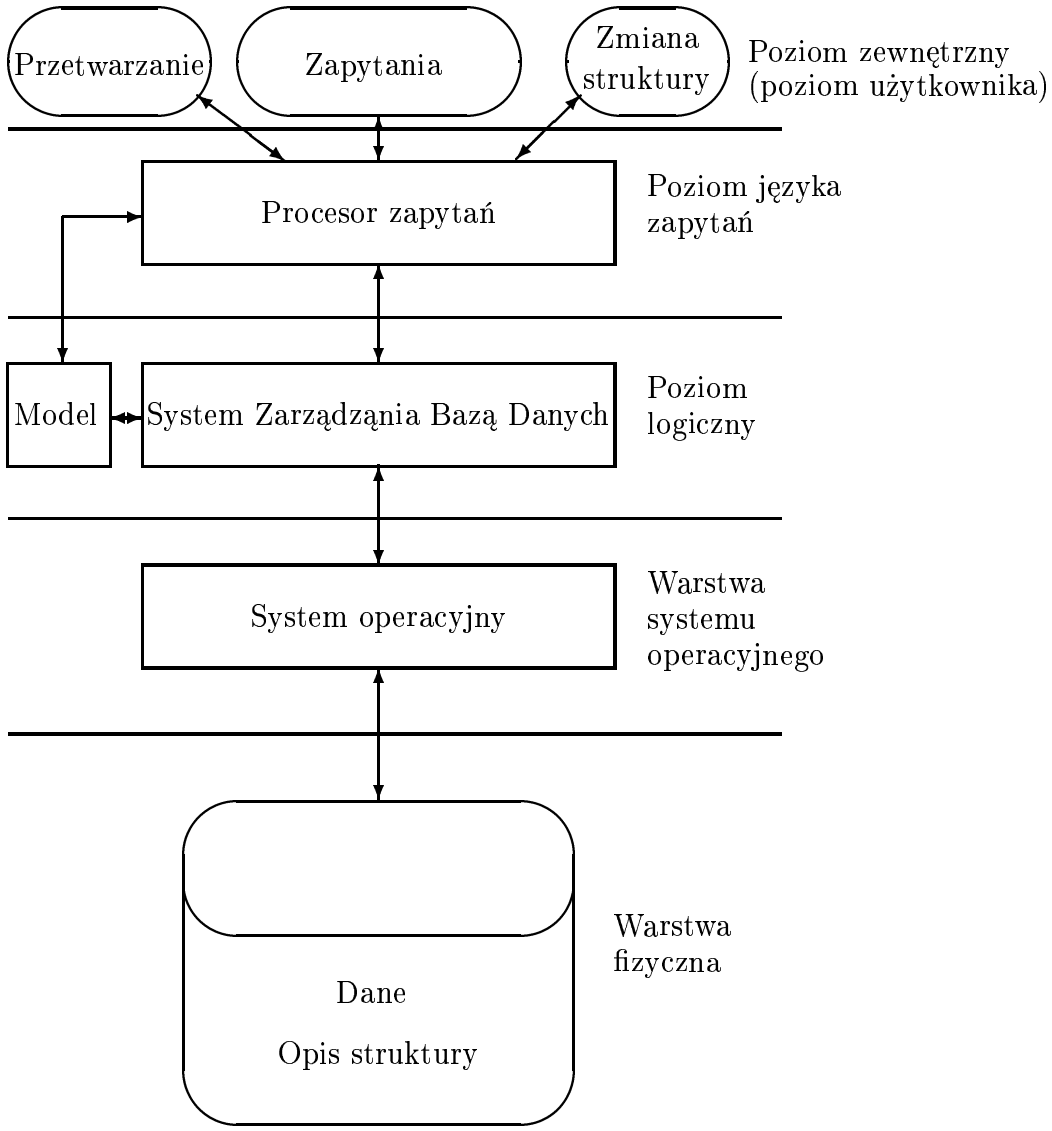
Bazy danych często implementowane są w architekturze *klient-serwer*. Serwer bazy danych zainstalowany jest zazwyczaj na dużym komputerze o znacznej mocy obliczeniowej. Na serwerze zainstalowany jest SZBD oraz sama baza danych. Do jednego serwera może być podłączonych wiele komputerów (terminali) klienckich. Na komputerach tych zainstalowane jest zwykle jedynie oprogramowanie realizujące graficzny interfejs użytkownika oraz oprogramowanie realizujące połączenie z serwerem. Cechy charakterystyczne takiego sytemu są następujące:

- **Przetwarzanie danych po stronie serwera**. Główne funkcje systemu związane z *przetwarzaniem danych* oraz zarządzaniem bazą danych realizowane są po stronie serwera.
- **Organizacja dostępu – serwer**. SZBD zainstalowany na serwerze jest odpowiedzialny za organizację i synchronizację dostępu do danych (wielodostęp, dostęp równoczesny) oraz za ochronę spójności danych.
- **Ochrona danych – serwer**. SZBD zainstalowany na serwerze jest odpowiedzialny za ochronę danych i kontrolę dostępu do nich.
- **Interfejs użytkownika – klient**. System zainstalowany po stronie klienta dostarcza intefejsu użytkownika (graficznego, semigraficznego, tekstowego) oraz jest odpowiedzialny za kontakt z użytkowniekiem, przyjmowanie jego poleceń, wykonanie lub przekazanie tych zleceń do serwera, przyjęcie i interpretację wyników.

Polecenia do serwera formułowane są w specjalizowanym języku zapytań; najczęściej jest to SQL. Mechanizm komunikacji najczęściej oparty jest o tzw. *(ODBC)Open Database Connectivity* .

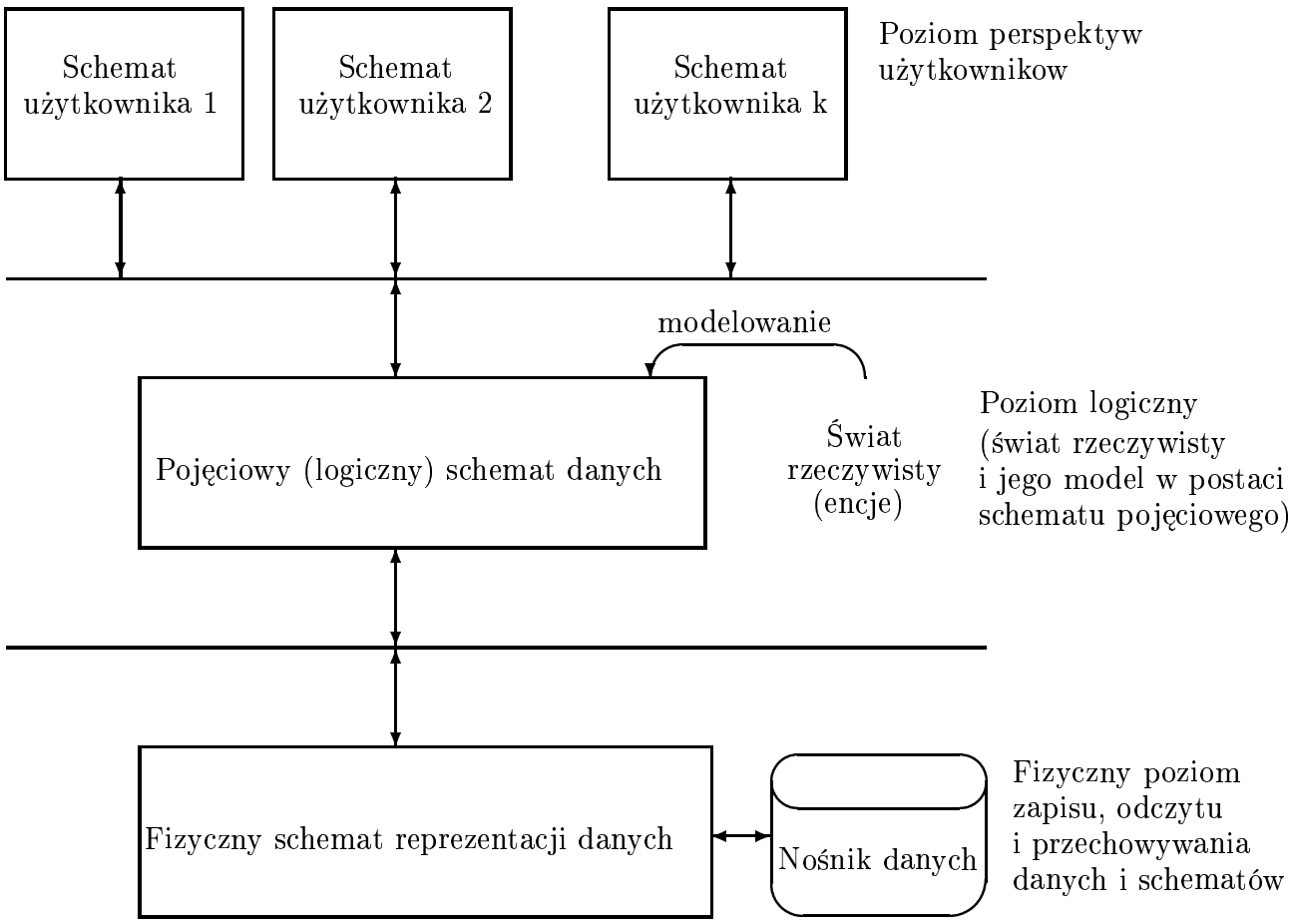
©Antoni Ligęza

Struktura warstwowa systemu bazy danych



©Antoni Ligęza

Poziomy reprezentacji danych



©Antoni Ligęza

Przykładowe typy danych: ACCESS

- **Tekst** – dane typu tekstowego, typ wybierany standardowo; wszelkie napisy, również te zawierające cyfry. Maksymalny rozmiar pola wynosi 255 znaków,
- **Memo|Nota** – dane typu notatnikowego; dowolny tekst o długości do 64 kB, nie może być indeksowany, nie podlega przetwarzaniu (edycja w oknie edytora: `Shft + F2`),
- **Liczba** – dane typu numerycznego (liczbowego); stosowany do wszelkich danych podlegających operacjom arytmetycznym (obliczenia), standardowy rozmiar pola od 1 do 8 bajtów,
- **Data/Godzina** – dane typu daty lub czasu, rozmiar pola 8 bajtów,
- **Waluta** – specjalny typ danych dla reprezentacji i przetwarzania danych wyrażających kwoty (przy obliczeniach stosowana jest arytmetyka stałoprzecunkowa zapobiegająca powstawaniu błędów zaokrągleń; rozmiar pola 8 bajtów,
- **Autonumer** – typ licznikowy; dane tego typu są generowane automatycznie (przyrostowo (+1) lub losowo) jako liczby całkowite długie,
- **Tak/Nie** – dane typu logicznego; pola te służą do przechowywania wartości *prawdy* lub *fałszu* (w ACCESS-ie odpowiednio -1 i 0),
- **Obiekt OLE** – typ specjalny, umożliwia przypisanie do pola obiektu osadzanego lub dołączanego (tekstu, obrazu, dźwięku); maksymalny rozmiar pola do 1 gigabajta,
- **Hiperłącze** – typ specjalny pozwala zapisać adresy plików i dokumentów w formacie UNC lub URL (np. strony WWW),
- **Kreator odnośników** – typ specjalny, służy do tworzenia pola listy lub pola kombi.

©Antoni Ligęza

Formaty liczb i dat: ACCESS

- **Bajt** – liczba całkowita, rozmiar 1 bajt; obejmuje liczby naturalne od 0 do 255,
- **Liczba całkowita** – liczba całkowita (krótka), romiar 2 bajty; obejmuje liczby całkowite od -32 768 do + 32 767,
- **Liczba całkowita długa** – liczba całkowita długa, rozmiar 4 bajty; obejmuje liczby całkowite od - 2 147 483 648 do + 2 147 483 647,
- **Pojedyncza precyzja** – liczby rzeczywiste pojedynczej precyzji (krótkie), rozmiar 4 bajty; obejmuje liczby rzeczywiste od -3,402823E38 do +3,402823E38 (sześć cyfr znaczących po przecinku, łącznie siedem cyfr),
- **Podwójna precyzja** – liczby rzeczywiste podwójnej precyzji (długie), rozmiar 8 bajtów; obejmuje liczby rzeczywiste od -1,7976931349E308 do +1,7976931349E308 (dziesięć cyfr znaczących po przecinku),
- **ID replikacji** – globalny jednoznaczny identyfikator, rozmiar 16 bajtów (unikalny kod przypisywany replikom rekordów lub tabel w bazie danych, tzw. GUID *Globalny Unikatowy Identyfikator Danych*); ten format liczbowy działa tylko z typem *autonumer*.

Data ogólna (21-10-97 17:23:35),
Pełna data (Środa 19 października 1997),
Data średnia (19 X 97),
Data krótka (19-10-97),
Godzina pełna (17:29:33),
Godzina średnia (05:34 PM),
Godzina krótka (17:34).
Dzięki istnieniu wielu symboli pomocniczych można zdefiniować niemal dowolny wymagany format wyświetlania daty (vide: Help).

©Antoni Ligęza