

# Modele danych i ewolucja systemów baz danych

Izabela Szczęch  
Krzysztof Dembczyński

Instytut Informatyki  
Zakład Inteligentnych Systemów Wspomagania Decyzji  
Politechnika Poznańska

Technologie wytwarzania oprogramowania  
Studia magisterskie, semestr II  
Semestr zimowy 2009/10

- 1 Zarządzanie danymi
- 2 Systemy zarządzania plikami
- 3 Operacyjne systemy baz danych
- 4 Analityczne systemy baz danych
- 5 Podsumowanie

- 1 Zarządzanie danymi
- 2 Systemy zarządzania plikami
- 3 Operacyjne systemy baz danych
- 4 Analityczne systemy baz danych
- 5 Podsumowanie

**Dane** są jednym z najważniejszych zasobów organizacyjnych.

Dane muszą być **zarządzane** tak samo jak inne ważne **aktywa** i **majątek**.

Większość organizacji nie jest w stanie **przeżyć** bez dobrych jakościowo danych na temat wewnętrznych operacji i zewnętrznego środowiska.

# Podejścia do zarządzania danymi

***Podejście tradycyjne:*** system zarządzania plikami

***Podejście nowoczesne:*** systemy zarządzania bazami danych

**Operacyjne systemy baz danych** wspierają zadania biurowe, administracyjne i organizacyjne.

**Analityczne systemy baz danych** wspierają kadrę kierowniczą, menadżerów, analityków w podejmowaniu decyzji.

# Podejścia do zarządzania danymi

**Internet** jako baza danych:

- **Hipermedialne** bazy danych,
- **Wyszukiwanie informacji** w zasobach Internetu.

**Temporalne i przestrzenne bazy danych**

**Obiektowe i multimedialne bazy danych**

**Bazy danych plików XML**

# Plan wykładu

- 1 Zarządzanie danymi
- 2 Systemy zarządzania plikami**
- 3 Operacyjne systemy baz danych
- 4 Analityczne systemy baz danych
- 5 Podsumowanie

**System zarządzania plikami** jest oprogramowaniem tworzącym, usuwającym i manipulującym plikami.

Pliki o **różnej konstrukcji** są wykorzystywane w **różnych celach**.

Każdy plik jest wykorzystywany **niezależnie**.

**Wiele wad** jednak bardzo często używany :)



## Przykład

Uczelnia posiada niezależne pliki z ocenami i danymi osobowymi dla każdego przedmiotu.

Zmiana adresu studenta wymusza zmianę w każdym pliku.

**Course:**  
**Computer Science**

**#Student**  
First name  
Surname  
Address  
Grade

**Course:**  
**Data mining**

**#Student**  
First name  
Surname  
Address  
Grade

**Course:**  
**Database Systems**

**#Student**  
First name  
Surname  
Address  
Grade

## Wady systemów zarządzania plikami:

- redundancja danych,
- brak spójności danych,
- brak niezależności danych,
- brak struktury danych.

# Plan wykładu

- 1 Zarządzanie danymi
- 2 Systemy zarządzania plikami
- 3 Operacyjne systemy baz danych**
- 4 Analityczne systemy baz danych
- 5 Podsumowanie

# Operacyjne systemy baz danych

**Cel:** wspomaganie pracowników w codziennej pracy, by polepszyć produktywności; przetwarzanie danych biurowych (operacyjnych).

Systemy operacyjne często kojarzone są z **przetwarzaniem transakcji na bieżąco** (ang. **On-line transaction processing - OLTP**).

**Główne zadania:** przetwarzanie dużej liczby współbieżnych transakcji, zapewnienie spójności danych.

**Transakcja** jest atomową jednostką przetwarzania, która jest przeprowadzona całkowicie lub wcale.

Transakcje przenoszą bazę danych z **jednego stanu spójnego** do **następne stanu spójnego**.

Przeniesienie pieniędzy z lokaty klienta banku na konto operacji bieżących

Przykładowa transakcja może składać się z następujących trzech operacji:

- Obniżenie wartości na lokacie pieniężnej,
- Podniesienie wartości na koncie operacji bieżących,
- Zapisanie transakcji w pliku logu.

# Zalety operacyjnych systemów zarządzania danymi

## Ścisłość:

- Możliwość przechowywania dużych wolumenów danych w dobrze zdefiniowanym, łatwym do utrzymania formacie.

## Prawie całkowity brak redundancji danych:

- Informacja występuje zazwyczaj raz, w jednym miejscu,
- Ta sama informacja może być dostępna dla różnych użytkowników.

## Spójność danych:

- Dane są dokładne, spójne, uaktualnione,
- Każda operacja aktualizacji jest dokonywana w jednym konkretnym miejscu.

# Zalety operacyjnych systemów zarządzania danymi

## Niezależność danych:

- Dane są niezależne od aplikacji, które z nich korzystają,
- Dane są składowane w jednej wspólnej bazie danych, a nie w oddzielnych plikach wykorzystywanych przez aplikacje,

## Elastyczny i abstrakcyjny dostęp do danych:

- Dane mogą być dostępne na wiele sposobów, z różnych perspektyw, w zależności od użytkownika i aplikacji,
- Istnieją abstrakcyjne języki zarządzania danymi,
- Dane są niezależne logicznie i fizycznie.

## Wydajność baz danych:

- Systemy zarządzania baz danych pozwalające na elastyczny dostęp do danych, mogą optymalizować wydajność przetwarzania.

# Zalety operacyjnych systemów zarządzania danymi

## Bezpieczeństwo danych:

- Dostęp do danych wymaga autentykacji i autoryzacji,
- Prawo dostępu do danych jest zarządzane,
- Zabezpieczenia przed utratą danych i awariami.

## Współdzielenie danych:

- W tym samym czasie z bazy danych może korzystać wiele osób.



## Wady operacyjnych systemów zarządzania danymi

- **Koszt systemu:** instalacja, utrzymanie oraz zakup oprogramowania i odpowiedniego sprzętu może okazać się bardzo kosztowne,
- **Niebezpieczeństwa:** istnieje niebezpieczeństwo nieautoryzowanego dostępu do danych lub zniszczenia bazy danych.

# Modele danych w zastosowaniach OLTP

**Model danych** określa dostęp i sposób zapisu danych.

**Logiczny model danych** – z punktu widzenia użytkownika

**Fizyczny model danych** – z punktu widzenia komputera

Model danych obejmuje:

- typy danych, związki pomiędzy danymi i ograniczenia na nie nałożone,
- zbiór operacji służący do definiowania, wyszukiwania i uaktualniania bazy danych.

## Podstawowe modele danych w zastosowaniach OLTP:

- Hierarchiczny,
- Sieciowy,
- Relacyjny,
- Post-relacyjny.

## Hierarchiczny model danych:

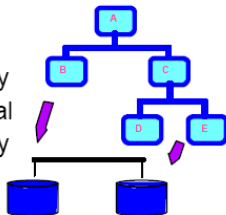
- najstarszy model danych, którego początki sięgają lat 60-tych,
- najstarszy hierarchiczny system bazy danych, Information Management System (IMS) stworzony i rozwijany przez IBM'a, powstał w celu organizacji i przechowywania informacji w projekcie Apollo,
- bardzo efektywne przetwarzanie danych,
- wydajność systemu jest przewidywalna ponieważ wszystkie ścieżki dostępu są znane,
- niestety nie jest tak elastyczny i łatwy do zrozumienia jak model relacyjny,
- **przykłady:** Systemy rezerwacji lotniczej, Serwery LDAP,  
...

## IMS is a Database Management System

- A Database is a collection of interrelated data items, stored once and organized in a form for easy retrieval.
- A Database Management System is a collection of programs for storing organizing, selecting, modifying, and extracting data from a database.

**IMS DB** is organized hierarchically

- To optimize storage and retrieval
- To ensure integrity and recovery



© IBM Corporation 2004

## Cechy hierarchicznego modelu danych:

- dane zorganizowane są poziomami,
- model reprezentuje strukturę drzewa (lub lasu, czyli zbioru drzew), która jest analogiczna do struktur organizacyjnych np. przedsiębiorstw,
- dostęp do danych jest zapewniony poprzez zdefiniowane ścieżki,
- połączenia RODZIC-POTOMEK; każdy potomek ma tylko jednego rodzica, rodzic może posiadać wiele potomków = relacja JEDEN-DO-WIELU,
- istnieje problem z reprezentacją relacji WIELE-DO-WIELE.

## Schemat bazy danych uczelni

<b>Faculty</b>	ID	Name	Address
----------------	----	------	---------

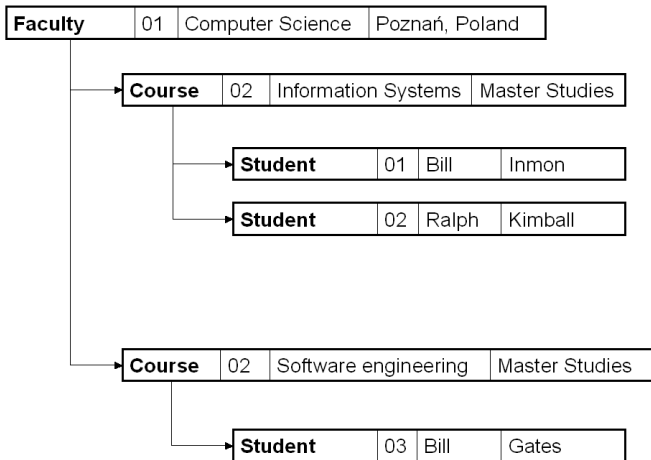


<b>Course</b>	ID	Name	Type of Study
---------------	----	------	---------------



<b>Student</b>	ID	Name	Surname
----------------	----	------	---------

# Baza danych uczelni





## Cechy hierarchicznego modelu danych:

- hierarchie są łatwe do składowania – model logiczny jest podobny do modelu fizycznego (hierarchie są zachowane poprzez wskaźniki),
- duża wydajność i prędkość działania oraz łatwa optymalizacja zapytań (wszystkie ścieżki dostępu są znane, dlatego predykcja wydajności jest bardzo prosta)
- operacje aktualizacji są proste do przeprowadzenia,
- **Trudności:** dane mogą być redundantne, co może prowadzić do niespójności (braku integralności).

Zapytania standardowe są przetwarzane efektywnie:

Zapytanie z góry na dół

Podaj liczbę studentów na wydziale Informatyki i Zarządzania.

Istnieje **problem** wydajnościowy z zapytaniami złożonymi:

Zapytanie z dołu do góry

Co studiuje student Kazimierz Dorn?

Zapytanie wymagające złożonej nawigacji

Znajdź wszystkie wydziały, na których jest kierunek Informatyka.

## Sieciowy model danych:

- w 1971, konferencja Data Systems Languages (CODASYL) zdefiniowała sieciowy model danych,
- podobny do hierarchicznego, jednak z mniejszą liczbą ograniczeń: *potomek* może mieć więcej *rodziców* (model WIELE-DO-WIELE),
- możliwość przeglądania danych w różnych kierunkach,
- wiele możliwych ścieżek dostępu do danych,
- duża wydajność,
- niestety charakteryzuje się dużą złożonością i trudnością utrzymania,
- model oparty jest o teorię mnogości,
- zbiór w modelu jest określony za pomocą rekordu właściciela (nazwa zbioru) oraz rekordów będących elementami zbioru,
- **przykłady:** CA-IDMS, COBOL, ...

## Relacyjny model danych:

- relacyjna baza danych postrzegana jest jako zbiór **relacji** bądź **tabel**,
- model relacyjny dotyczy wyłącznie zagadnień **logicznych**, a nie **fizycznych**,
- systemy relacyjne nie są tak efektywne jak hierarchiczne i sieciowe,
- pomiędzy warstwą logiczną i fizyczną jest miejsce na oprogramowanie **optymalizujące** wykonywanie zapytań,
- model relacyjny jest oparty na matematycznym pojęciu relacji,
- zapytania są deklaratywne,
- nie ma predefiniowanych ścieżek dostępu do danych,
- możliwość różnorodnego spojrzenia na dane,

## Relacyjny model danych:

- podstawową strukturą danych w modelu relacyjnym jest **relacja** reprezentowana w postaci **dwuwymiarowej tablicy**,
- relacja jest podzbiorem **iloczynu kartezyjskiego** dziedzin atrybutów,
- relacja składa się z **krotek** i **atrybutów**,
- wszystkie krotki relacji są różne (nie ma duplikatów),
- atrybuty w relacji są różne,
- dziedzina atrybutu określa dostępne wartości,
- kolejność krotek i atrybutów nie ma znaczenia,
- wartości atrybutów są atomowe,
- prawie całkowity brak redundancji i spójność danych (normalizacja),
- **przykłady**: Oracle, IBM DB2, Microsoft SQL Server, MaxDB, MySQL, PostgreSQL, ...

## Reguły integralności

Każda **reguła integralności** musi być uzależniona od konkretnej bazy danych.

Przykłady reguł integralności:

- Pensja musi być większa od 0
- Nazwisko musi być znane
- Etaty muszą pochodzić z listy dostępnych wartości
- itd.

Model relacyjny obejmuje trzy **ogólne cechy integralności**:

- klucze kandydujące (i główne),
- klucze obce,
- dziedziny – integralność atrybutu, mówi, że każdy atrybut musi spełniać następujący warunek: wartości atrybutu są pobierane z odpowiedniej dziedziny.

### Ważne!

Określenie dziedziny jest najprostszym mechanizmem sprawdzania integralności danych (np. przy porównywaniu wartości różnych atrybutów), jednak jest to element najslabiej wspierany przez istniejące relacyjne systemy baz danych.

**Klucze kandydujące** zapewniają podstawowy **mechanizm adresowania na poziomie krotki** w systemie relacyjnym. Oznacza to, że jedynym zagwarantowanym przez system sposobem dotarcia do określonej krotki jest droga przez *wartość jakiegoś klucza kandydującego*.

Wartość klucza obcego stanowi **referencję** do krotki zawierającej wartość odpowiadającego mu klucza kandydującego (krotki **docelowej**).

Problem zapewnienia, żeby baza danych nie zawierała żadnych niedopuszczalnych wartości klucza obcego, nazywa się problemem **integralności referencyjnej** (referential integrity).

Warunek, aby wartości danego klucza obcego zgadzały się z wartościami odpowiadającego mu klucza kandydującego jest znany jako **więzy referencyjne** (referential constraint).



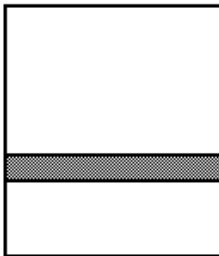
## Operatory relacyjne:

- selekcja (restrykcja),
- projekcja (rzut),
- iloczyn kartezjański,
- suma,
- przecięcie,
- różnica,
- łączenie,
- iloraz.

Operatory relacyjne operują na poziomie **zbiorów**.

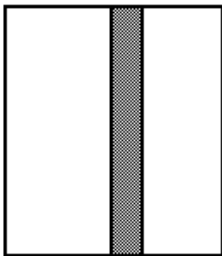
**Selekcja** (czasami nazywana restrykcją) daje w wyniku relację składającą się ze wszystkich krotek ze wskazanej relacji, które spełniają określone warunki.

## Selekcja

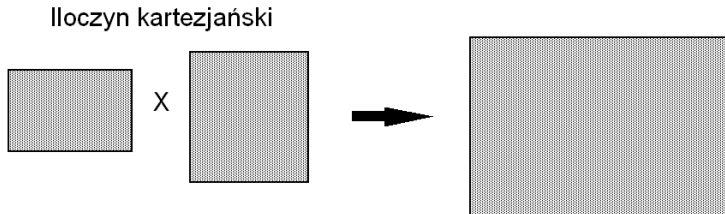


**Projekcja** (rzut) daje w wyniku relację złożoną z tych wszystkich krotek, która pozostały jako krotki danej relacji po usunięciu z niej wskazanych atrybutów.

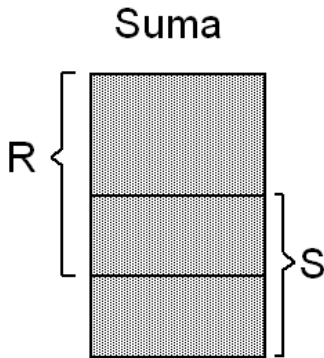
## Projekcja



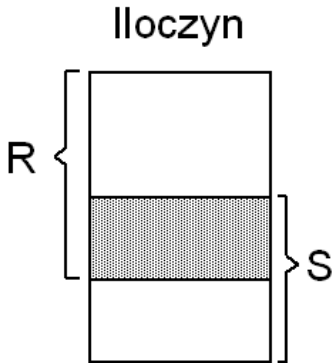
**Iloczyn kartezjański** daje relację składającą się ze wszystkich krotek, będących kombinacją dwóch krotek, po jednej z każdej wskazanej relacji.



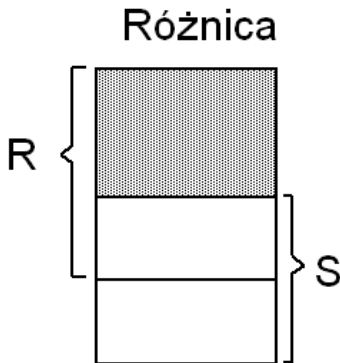
**Suma** daje w wyniku relację składającą się ze wszystkich krotek, występujących w jednej lub obu wskazanych relacjach.



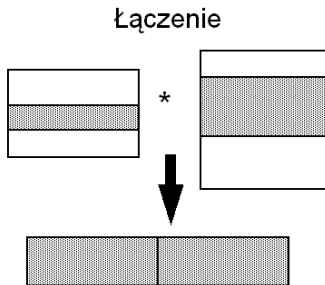
**Przecięcie** daje w rezultacie relację składającą się ze wszystkich krotek, występujących w obu wskazanych relacjach.



**Różnica** daje w wyniku relację składającą się ze wszystkich krotek, występujących w pierwszej relacji i nie występujących w drugiej wskazanej relacji.



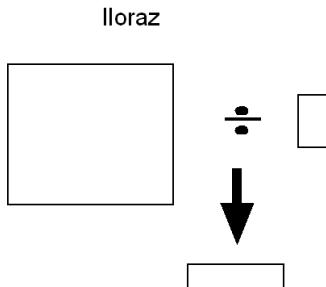
**Łączenie** daje w wyniku relację składającą się ze wszystkich możliwych krotek, które są kombinacjami dwu krotek, po jednej z każdej ze wskazanych relacji, takich że dwie krotki dające wkład do którejkolwiek kombinacji mają tę samą wartość wspólnego atrybutu (lub atrybutów) tych dwu relacji.





**Iloraz** bierze dwie relacje, jedną binarną, a drugą unarną i daje w wyniku relację składającą się ze wszystkich wartości jednego atrybutu relacji binarnej, które zgadzają się (pod względem wartości tego drugiego atrybutu) ze wszystkimi wartościami relacji unarnej.

**Przykład:** podaj klientów, którzy kupują wyposażenie b2 i b3.



T1	A	B
	a1	b1
	a1	b2
	a1	b3
	a2	b1
	a2	b3
	a3	b2
	a3	b3
	a3	b4
	a4	b1

T2	B
	b2
	b3

T3	A
	a1
	a3

A: Customer Number

B: Car's Option ID

$T3 = T1 / T2$

.

## Ważne!!!

Wynik dowolnej operacji jest obiektem tego samego rodzaju co wejście (wszystkie są **relacjami**), zatem *wynik dowolnej operacji może stanowić wejście innej* – jest to własność **domknięcia**.

Pozwala to na tworzenie **zagnieżdżonych wyrażeń relacyjnych**

## Domknięcie

```
SELECT * FROM  
(SELECT * FROM (SELECT * FROM Relacja);
```

# Historia języka SQL:

- język SQL został opracowany w laboratoriach IBM w latach 70'tych,
- w 1986 SQL stał się oficjalnym standardem wspieranym przez ISO i ANSI (standard został opisany na 100 stronach),
- kolejny standard: SQL89 (120 stron),
- SQL92 (około 600 stron) – aka SQL2,
- SQL99 (około 2200 stron) – aka SQL3,
- SQL:2003,
- SQL:2006,
- SQL:2008.

## Główne cechy podejścia relacyjnego wpływające na jego popularność:

- podstawy teoretyczne (algebra i rachunek relacyjny),
- domkniętość systemów relacyjnych,
- podejście abstrakcyjne do przechowywania, wyszukiwania i uaktualniania danych,
- rzeczywista niezależność aplikacji od danych,
- model zapewniający integralność danych,
- standard języka SQL (jednak często krytykowany).

## Główne cechy podejścia relacyjnego wpływające na jego popularność:

- podstawy teoretyczne (algebra i rachunek relacyjny),
- domkniętość systemów relacyjnych,
- podejście abstrakcyjne do przechowywania, wyszukiwania i uaktualniania danych,
- rzeczywista niezależność aplikacji od danych,
- model zapewniający integralność danych,
- standard języka SQL (jednak często krytykowany).

## Postrelacyjny (obiekto-relacyjny) model danych:

- związany z rozszerzeniem języka SQL – **SQL99** lub **SQL3**,
- przechowywanie typów złożonych: multimedialnych, przestrzennych, temporalnych,
- typy danych definiowane przez użytkownika,
- typy referencyjne,
- kolekcje (np. tablice),
- wsparcie dla dużych obiektów,

## Postrelacyjny (obiektowo-relacyjny) model danych:

- hierarchie relacji,
- możliwość korzystania z SQL'a jako samodzielnego języka aplikacji,
- wyzwalacze,
- składowane procedury i funkcje definiowane przez użytkownika,
- zapytania rekursywne,
- operacje OLAP'owe.

# Plan wykładu

- 1 Zarządzanie danymi
- 2 Systemy zarządzania plikami
- 3 Operacyjne systemy baz danych
- 4 Analityczne systemy baz danych**
- 5 Podsumowanie



## Analityczne systemy baz danych

**Cel:** wspomaganie decydentów przy podejmowaniu szybszych i lepszych decyzji – systemy wspomagania decyzji

## Systemy baz danych w analizie danych i wspomaganiu decyzji:

- Integracja systemów analizy danych z systemami baz danych: SAS, WEKA współpracują z SZBD; Oracle i DB2 są wzbogacane o narzędzia analityczne; SQL99
- Analiza bardzo dużych zbiorów danych
- Analiza danych zgromadzonych w operacyjnych bazach danych
- Tworzenie hurtowni danych
- Statystyczne systemy baz danych
- Elastyczne i eksploracyjne zapytania do baz danych
- Eksploracja danych
- Eksploracja zasobów Internetu

## Hurtownie danych i technologia OLAP:

- **Hurtownia danych** służy do magazynowania informacji z różnych źródeł w celu dostarczenia spójnego źródła danych dla zapytań wspomagających decyzje.
- **OLAP (On-Line Analytical Processing)** – przetwarzanie analityczne na bieżąco  
**Zadanie:** efektywne wielowymiarowe przetwarzanie ogromnej ilości danych
- W większości organizacji dane biznesowe są dostępne na miejscu – bardzo dużo, gdzieś, w pewnej nieokreślonej formie ...
- Dane są dostępne, ale nie **informacja (wiedza)** – brak odpowiedniej informacji w odpowiednim czasie

**Systemy wspomagania decyzji** (DSS – Decision Support Systems) mają na celu przyspieszanie podejmowania lepszych decyzji.

Idea systemów wspomagania decyzji powstała dużo wcześniej niż zaawansowane systemy zarządzania bazami danych.

Ich zadaniem jest dostarczanie informacji ludziom podejmującym decyzje.

Uzyskane informacje wzbogacają wiedzę decydentów, wspomagając ich w podejmowaniu decyzji dotyczących działań taktycznych i strategicznych.

Najczęściej zadaniem systemu wspomagania decyzji jest udzielenie rzeczowej odpowiedzi na pytania postawione przez użytkownika:

#### Zapytanie:

Dlaczego moja sprzedaż nie osiąga wymaganego poziomu?

Powyższe pytanie jest trudne do realizacji przez system komputerowy (może kiedyś . . . :)

Na pewno można skierować zapytania następującego typu:

#### Zapytania:

- Ile sprzedano samochodów w Wielkopolsce podczas ostatniego roku?
- Ile sprzedano samochodów osobowych w Poznaniu w ostatnich 10 latach?

VIEWTABLE: Sashelp.Prdsal2

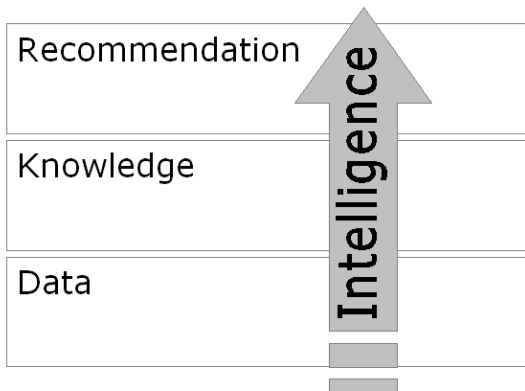
	Country	State/Province	County	Actual Sales	Predicted Sales	Product Type	Product	Year
1	U.S.A.	California		\$987.36	\$692.24	FURNITURE	SOFA	1998
2	U.S.A.	California		\$1,782.96	\$568.48	FURNITURE	SOFA	1998
3	U.S.A.	California		\$32.64	\$16.32	FURNITURE	SOFA	1998
4	U.S.A.	California		\$1,825.12	\$756.16	FURNITURE	SOFA	1998
5	U.S.A.	California		\$750.72	\$723.52	FURNITURE	SOFA	1998
6	U.S.A.	California		\$2,426.24	\$2,428.96	FURNITURE	SOFA	1998
7	U.S.A.	California		\$1,791.12	\$2,250.80	FURNITURE	SOFA	1998
8	U.S.A.	California		\$2,282.08	\$350.88	FURNITURE	SOFA	1998
9	U.S.A.	California		\$2,518.72	\$1,736.72	FURNITURE	SOFA	1998
10	U.S.A.	California		\$1,436.16	\$2,167.84	FURNITURE	SOFA	1998
11	U.S.A.	California		\$2,214.72	\$622.56	FURNITURE	SOFA	1998

Country = U.S.A.

Year		1998				
		Actual Sales				Predicted
State/Province		Total Number of Nonmissing Values	Average	Percent of Sum	Percent of Total Number	Percent of
California		288	\$1,760.83	18.46	8.33	18.85
Colorado		288	\$1,053.72	11.04	8.33	10.95
Florida		288	\$975.11	10.22	8.33	10.45
Illinois	Adams	288	\$76.41	19.47	20.00	20.41
	Cook	288	\$77.35	19.71	20.00	20.50
	Fayette	288	\$75.13	19.15	20.00	19.31
	McLean	288	\$84.68	21.58	20.00	20.12
	Winnebago	288	\$78.80	20.08	20.00	19.67
New York		288	\$1,706.87	17.89	8.33	17.31
North Carolina		288	\$1,053.82	11.05	8.33	10.74
Texas		288	\$1,601.29	16.78	8.33	16.25
Washington		288	\$996.57	10.45	8.33	11.07

- Hurtownia danych może być podstawą DSS
- OLAP jest częścią systemów wspomagania decyzji
- Eksploracja danych (ang. Data Mining) jest silnym, o dużej wydajności narzędziem analizy danych w systemach DSS
- Wielokryterialna analiza decyzji

## Inteligentne systemy wspomagania decyzji:





## Porównanie OLTP i OLAP

Kryterium	OLTP	OLAP
Użytkownicy	Urzędnicy	Decydenci
Funkcja	Codzienne operacje	Wspomaganie decyzji
Projekt bazy danych	Zorientowane na aplikacje	Zorientowane na temat
Dane	Bieżące, aktualne, szczegółowe, płaskie, relacyjne, wyodrębnione	Historyczne, sumowane, wielowymiarowe, zintegrowane
Używanie	Powtarzalne	Ad-hoc
Dostęp	Odczyt/zapis	Wiele przeszukiwań
Jednostka pracy	Transakcje	Złożone zapytania
Liczba krotek w operacji	Rzędu 10	Rzędu miliona
Użytkownicy	Tysiące	Setki
Rozmiar bazy danych	100 MB-GB	100 GB-TB
Metryka	Wydażność transakcji	Odpowiedź na zapytanie

## Co to jest hurtownia danych?

### Definicja 1 (Bill Inmon)

Jest to ukierunkowana, zintegrowana, czasowa, nieulotna kolekcja danych wspomagająca proces wspomaganie decyzji

### Definicja 2

Kolekcja danych wykorzystywana do wspomaganie decyzji

### Definicja 3

Baza danych wspomagająca podejmowanie decyzji odseparowana od operacyjnej bazy danych

Dwa ważne nazwiska: **Bill Inmon** i **Ralph Kimball**

## ● **Ukierunkowana**

- Ukierunkowana na dobrze zdefiniowany cel biznesowy przedsiębiorstwa
- Ukierunkowanie inne niż operacyjna baza danych

## ● **Zintegrowana**

- Usunięte niespójności w zbieranych danych (konwencje nazewnictwa, kodowania pomiędzy różnymi źródłami danych)
- Różne (heterogeniczne) źródła danych
- Konwersja i integracja przenoszonych danych

- **Czasowa**

- Horyzont czasowy jest dłuższy niż w przypadku operacyjnej bazy danych
- Hurtownia danych zawsze zawiera elementy związane z czasem

- **Nieulotna**

- Dane operacyjne są regularnie uaktualniane
- W hurtowniach danych dane są **doładowywane**
- W hurtowniach danych nie ma uaktualniania danych w tradycyjnym znaczeniu

## Wydajność i separowalność hurtowni danych:

- specjalna organizacja danych, metody dostępu i implementacja metod jest wymagana do wspomagania złożonych, wielowymiarowych zapytań,
- złożone zapytania mogłyby obniżyć wydajność transakcji operacyjnych,
- kontrola współbieżności oraz moduły odzyskiwania są różne dla OLTP i OLAP,
- wspomaganie decyzji wymaga danych historycznych, które nie są przechowywane w operacyjnych bazach danych,
- systemy wspomagania decyzji operują na agregacjach danych z różnych źródeł,
- różne źródła przechowują dane w niespójnej postaci.

## **Zalety systemów hurtowni danych:**

- Wysoka wydajność zapytań
- Zapytania są niewidoczne poza hurtownią
- Brak ingerencji w dane operacyjne
- Możliwość pracy w przypadku braku dostępu do źródła danych
- Wspieranie specjalnych rodzajów zapytań
- Dodatkowe informacje udostępniane przez hurtownie danych

## Podstawowe modele danych:

- *Hierarchiczny*,
- Relacyjny,
- *Post-relacyjny*,
- **Wielowymiarowy.**

# Przejście z modelu relacyjnego do wielowymiarowego:

Students' grades	ID	Academic year	Student	Course	Professor	Grade
	01	D1	S4	C1	P1	3.9
	02	D1	S4	C1	P2	4.0
	03	D1	S5	C2	P1	4.4
	04	D1	S5	C2	P2	4.4
	01	D2	S4	C1	P1	3.5
	02	D2	S4	C1	P2	4.0
	03	D2	S5	C2	P1	4.1
	04	D2	S5	C2	P2	4.0
	05	D3	S6	C1	P1	3.6
	06	D3	S6	C1	P2	3.9
	07	D3	S7	C2	P3	4.8

Avg(Grade)  
by Academic year  
and Professor

Academic year/ Professor	P1	P2	P3
D1	4.1	4.2	
D2	3.8	4.0	
D3	3.6	3.9	4.8

Komórki na przecięciu wierszy i kolumn reprezentują zagregowane wartości atrybutu *Grade*.



VIEWTABLE: Sashelp.Prdsal2

	Country	State/Province	County	Actual Sales	Predicted Sales	Product Type	Product	Year
1	U.S.A.	California		\$987.36	\$692.24	FURNITURE	SOFA	1998
2	U.S.A.	California		\$1,782.96	\$568.48	FURNITURE	SOFA	1998
3	U.S.A.	California		\$32.64	\$16.32	FURNITURE	SOFA	1998
4	U.S.A.	California		\$1,825.12	\$756.16	FURNITURE	SOFA	1998
5	U.S.A.	California		\$750.72	\$723.52	FURNITURE	SOFA	1998
6	U.S.A.	California		\$2,426.24	\$2,428.96	FURNITURE	SOFA	1998
7	U.S.A.	California		\$1,791.12	\$2,250.80	FURNITURE	SOFA	1998
8	U.S.A.	California		\$2,282.08	\$350.88	FURNITURE	SOFA	1998
9	U.S.A.	California		\$2,518.72	\$1,736.72	FURNITURE	SOFA	1998
10	U.S.A.	California		\$1,436.16	\$2,167.84	FURNITURE	SOFA	1998
11	U.S.A.	California		\$2,214.72	\$622.56	FURNITURE	SOFA	1998

Country = U.S.A.

Year		1998				
		Actual Sales				Predicted
State/Province		Total Number of Nonmissing Values	Average	Percent of Sum	Percent of Total Number	Percent of
California		288	\$1,760.83	18.46	8.33	18.85
Colorado		288	\$1,053.72	11.04	8.33	10.95
Florida		288	\$975.11	10.22	8.33	10.45
Illinois	Adams	288	\$76.41	19.47	20.00	20.41
	Cook	288	\$77.35	19.71	20.00	20.50
	Fayette	288	\$75.13	19.15	20.00	19.31
	McLean	288	\$84.68	21.58	20.00	20.12
	Winnebago	288	\$78.80	20.08	20.00	19.67
New York		288	\$1,706.87	17.89	8.33	17.31
North Carolina		288	\$1,053.82	11.05	8.33	10.74
Texas		288	\$1,601.29	16.78	8.33	16.25
Washington		288	\$996.57	10.45	8.33	11.07

## Wielowymiarowy model danych: sprzedaż produktów RTV/AGD

Location: Vancouver				
Time (quarters)	Items			
	TV	Computer	Phone	Security
Q1	605	825	14	400
Q2	680	952	31	512
Q3	812	1023	30	501
Q4	927	1038	38	580

Takie same tabele dla Chicago, Nowego Jorku i Toronto:

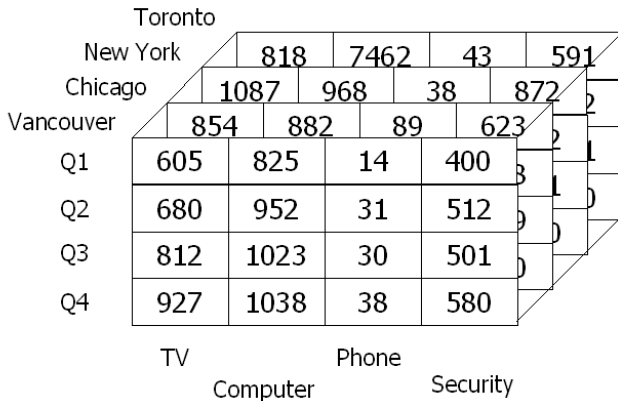
605	825	14	400
680	952	31	512
812	1023	30	501
927	1038	38	580

1087	968	38	872
1130	1024	41	925
1034	1048	45	1002
1142	1091	52	984

854	882	89	623
943	890	64	698
1023	924	59	789
1129	992	63	870

818	746	43	591
894	769	52	682
940	795	58	728
978	864	59	784

## Kostka wielowymiarowa:



	Toronto				
	New York	818	7462	43	591
	Chicago	1087	968	38	872
	Vancouver	854	882	89	623
Q1		605	825	14	400
Q2		680	952	31	512
Q3		812	1023	30	501
Q4		927	1038	38	580
		TV	Computer	Phone	Security

Możliwa jest większa liczba wymiarów.

## Różne poziomy agregacji:

- Sprzedaż(czas, produkt, \*)

Q1	3364	3421	184	2486
Q2	3647	3635	188	2817
Q3	3809	3790	186	3020
Q4	4176	3985	212	3218

TV

Phone

Computer

Security

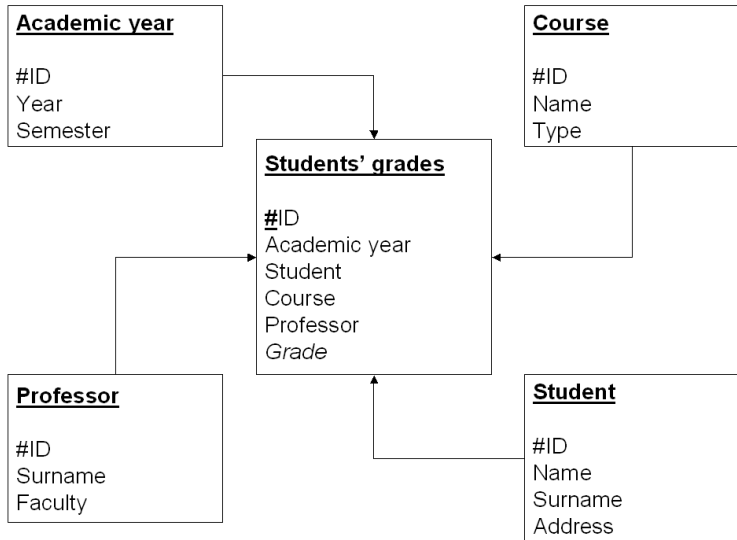
- Sprzedaż(czas, \*, \*); Sprzedaż(\*, \*, \*)

## Operacje w wielowymiarowym modelu danych:

- Roll up – sumowanie danych wzdłuż hierarchii wymiaru (miasto → województwo)
- Drill down – w drugą stronę (województwo → miasto)
- Slice and dice – selekcja i projekcja wymiarów
- Pivot – zamiana wyświetlanych wymiarów
- Inne – np. tworzenie rankingów, średnie ruchome, itp.

Toronto				
New York	818	7462	43	591
Chicago	1087	968	38	872
Vancouver	854	882	89	623
Q1	605	825	14	400
Q2	680	952	31	512
Q3	812	1023	30	501
Q4	927	1038	38	580
	TV	Computer	Phone	Security

# Schemat gwiazdy:



# Plan wykładu

- 1 Zarządzanie danymi
- 2 Systemy zarządzania plikami
- 3 Operacyjne systemy baz danych
- 4 Analityczne systemy baz danych
- 5 Podsumowanie**



## Podsumowanie

- Od systemu plików, przez systemy operacyjne, do systemów analitycznych . . . – nie oznacza to, że systemy plików i systemy operacyjne nie są dalej rozwijane,
- Istnieje wyraźna różnica pomiędzy systemami operacyjnymi i analitycznymi: koncepcyjna i technologiczna,
- Powstało i powstają ciągle nowe modele danych ukierunkowane na specyficzne zastosowania.