Informatyzacja przedsiębiorstw

Izabela Szczęch

Politechnika Poznańska

Plan wykładu

Elementy Business Intelligence

- Model wielowymiarowy
- Modelowanie hurtowni danych – podstawowe schematy logiczne
- Operacje na modelu wielowymiarowym
- Implementacje modelu wielowymiarowego

Przetwarzanie analityczne OLAP

- Aplikacje analityczne są zorientowane na wspieranie procesów decyzyjnych poprzez wykonywanie zaawansowanych analiz, wspomagających zarządzanie przedsiębiorstwem, np.
  - analiza trendów sprzedaży
  - analiza nakładów reklamowych i zysków
  - analiza ruchu telefonicznego

Przykładowe zapytania analityczne:

- *Ile sprzedano samochodów w Wielkopolsce w zeszłym roku*?
- *Ile sprzedano samochodów osobowych w Poznaniu w ostatnich 10 latach*?

Wielowymiarowy model danych

Z reguły jest wiele różnych wymiarów,
według których można analizować pewien zbiór danych.
Ta złożona perspektywa, czyli wielowymiarowy obraz pojęciowy,
wydaje się być sposobem, w jaki większość ludzi interesu z natury
widzi swoje przedsiębiorstwo.

E. F. Codd, 1993

Wielowymiarowy model danych

			YEAR		
COUNTRY	REGION_N	Dane	1995	1996	Suma całkowita
US	Central	Średnia: sredni_dochod	119 313,12 zł	118 997,20 zł	119 154,37 zł
		Suma: ilosc_zakupu	20 750,00 zł	20 667,00 zł	41 417,00 zł
		Suma: TKW	228 966,00 zł	224 361,00 zł	453 327,00 zł
		Suma: REVENUE	664 352,00 zł	652 498,00 zł	1 316 850,00 zł
	East	Średnia: sredni_dochod	121 614,49 zł	121 764,85 zł	121 690,13 zł
		Suma: ilosc_zakupu	13 542,00 zł	13 808,00 zł	27 350,00 zł
		Suma: TKW	145 434,00 zł	149 316,00 zł	294 750,00 zł
		Suma: REVENUE	428 851,00 zł	441 447,00 zł	870 298,00 zł
	South	Średnia: sredni_dochod	118 505,26 zł	118 401,87 zł	118 453,28 zł
		Suma: ilosc_zakupu	12 740,00 zł	12 921,00 zł	25 661,00 zł
		Suma: TKW	138 687,00 zł	139 483,00 zł	278 170,00 zł
		Suma: REVENUE	408 561,00 zł	409 674,00 zł	818 235,00 zł
	West	Średnia: sredni_dochod	119 923,54 zł	120 099,47 zł	120 012,44 zł
		Suma: ilosc_zakupu	5 491,00 zł	5 591,00 zł	11 082,00 zł
		Suma: TKW	29 984,00 zł	30 300,00 zł	60 284,00 zł
		Suma: REVENUE	68 872,00 zł	70 030,00 zł	138 902,00 zł
US - Średnia: sredni_dochod			119 789,02 zł	119 713,86 zł	119 751,20 zł
US - Suma: ilosc_zakupu			52 523,00 zł	52 987,00 zł	105 510,00 zł
US - Suma: TKW			543 071,00 zł	543 460,00 zł	1 086 531,00 zł
US - Suma: REVENUE			1 570 636,00 zł	1 573 649,00 zł	3 144 285,00 zł
Średnia: sredni_dochod, Razem			119 789,02 zł	119 713,86 zł	119 751,20 zł
Suma: ilosc_zakupu, Razem			52 523,00 zł	52 987,00 zł	105 510,00 zł
Suma: TKW, Razem			543 071,00 zł	543 460,00 zł	1 086 531,00 zł
Suma: REVENUE, Razem			1 570 636,00 zł	1 573 649,00 zł	3 144 285,00 zł

Wielowymiarowy model danych

- Dane w hurtowniach danych są zorganizowane w postaci tzw. **modelu wielowymiarowego** (ang. multidimensional data model), w którym wyróżnia się dwie podstawowe kategorie danych:
  - fakty
  - wymiary

Wielowymiarowy model danych

- **Fakty** (ang. facts) reprezentują informacje podlegające analizie, np. fakt sprzedaży produktu, fakt wykonania rozmowy telefonicznej, fakt ubezpieczenia pojazdu

- Fakty są charakteryzowane ilościowo za pomocą cech zwanych **miarami** (ang. measures). Przykładowo, miarą jest liczba zakupionych produktów, czas trwania rozmowy, kwota ubezpieczenia

Wielowymiarowy model danych

- **Wymiary** (ang. dimensions) ustalają kontekst analizy. Przykładowo analiza sprzedaży czekolady (produkt) w Auchan (lokalizacja) w poszczególnych miesiącach roku (czas) jest dokonywana w wymiarze *Produktu, Lokalizacji i Czasu*

- Wymiary składają się z poziomów, które tworzą **hierarchię**
- Zależności hierarchiczne między poziomami tworzą tzw. **strukturę wymiaru**

`wymiar LOKALIZACJA`

kraj ⟶ województwo ⟶ miasto ⟶ sklep

`wymiar PRODUKT`

kategoria_produktu ⟶ nazwa_produktu

Podstawowe schematy logiczne hurtowni danych

Trzy podstawowe schematy logiczne hurtowni danych:

- schemat gwiazdy
- schemat płatka śniegu
- schemat konstelacji faktów (wielokrotnych tabel faktów)

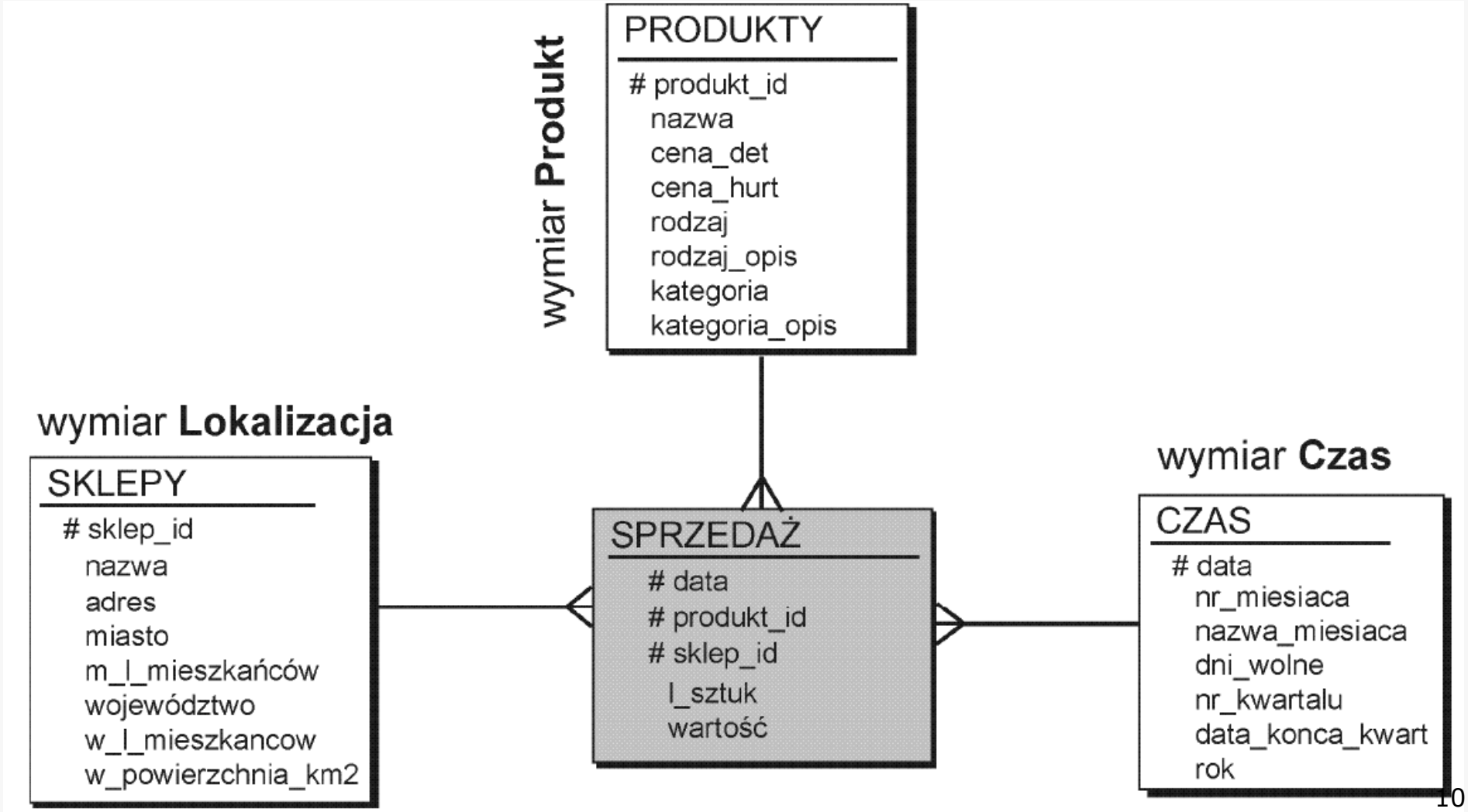
Schemat gwiazdy: pojedyncza tablica (faktów) w centrum połączona z wieloma tablicami wymiarów

Schemat płatka śniegu: rozwinięcie schematu gwiazdy poprzez normalizację relacji wymiarów

Schemat konstelacji faktów: wiele tablic faktów współdzieli tablice wymiarów

Schemat gwiazdy

- **Schemat gwiazdy:** pojedyncza tablica faktów w centrum połączona z wieloma tablicami wymiarów



Tablica faktów

- Każda krotka zawiera mierzalną wartość opisującą analizowany proces
- Każda krotka zawiera klucze obce do tablic wymiarów oraz kolumny numeryczne miar
- Każdy nowy, mierzalny fakt jest do niej zapisywany
- Analizie podlegają agregowane wartości miar
- Miary zależą od zbioru wymiarów, np. liczba sprzedanych sztuk (*l_sztuk*) zależy od produktu, sklepu, miesiąca itp.

Tablica wymiarów

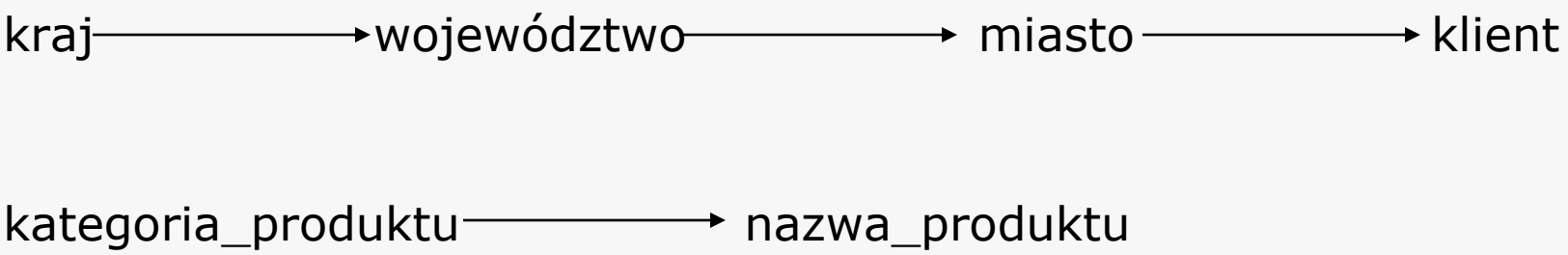
- Każda tablica wymiaru odpowiada obiektowi ze świata rzeczywistego: sklep, produkt, klient, dział, itp.
- Tablica wymiaru zawiera dużo atrybutów opisowych
- Zapisane wartości są stosunkowo statyczne, rzadko się zmieniają (stosunkowo rzadko, w porównaniu z wpisami do tabeli faktów, powstaje np. nowy sklep)
- Zawartość tabeli wymiaru służy do filtrowania i grupowania wyników
- Tablica wymiarów opisuje fakty zapisane w tablicy faktów

Tablica faktów vs. tablica wymiarów

- **Tablica faktów:**
  - wąska
  - długa (bardzo dużo krotek)
  - krotki opisane są za pomocą atrybutów numerycznych (miar)
  - dynamiczna (rośnie z czasem)
- **Tablica wymiaru:**
  - szeroka
  - raczej krótka
  - opisowa
  - statyczna
- **Fakty zawierają liczby, a wymiary etykiety**

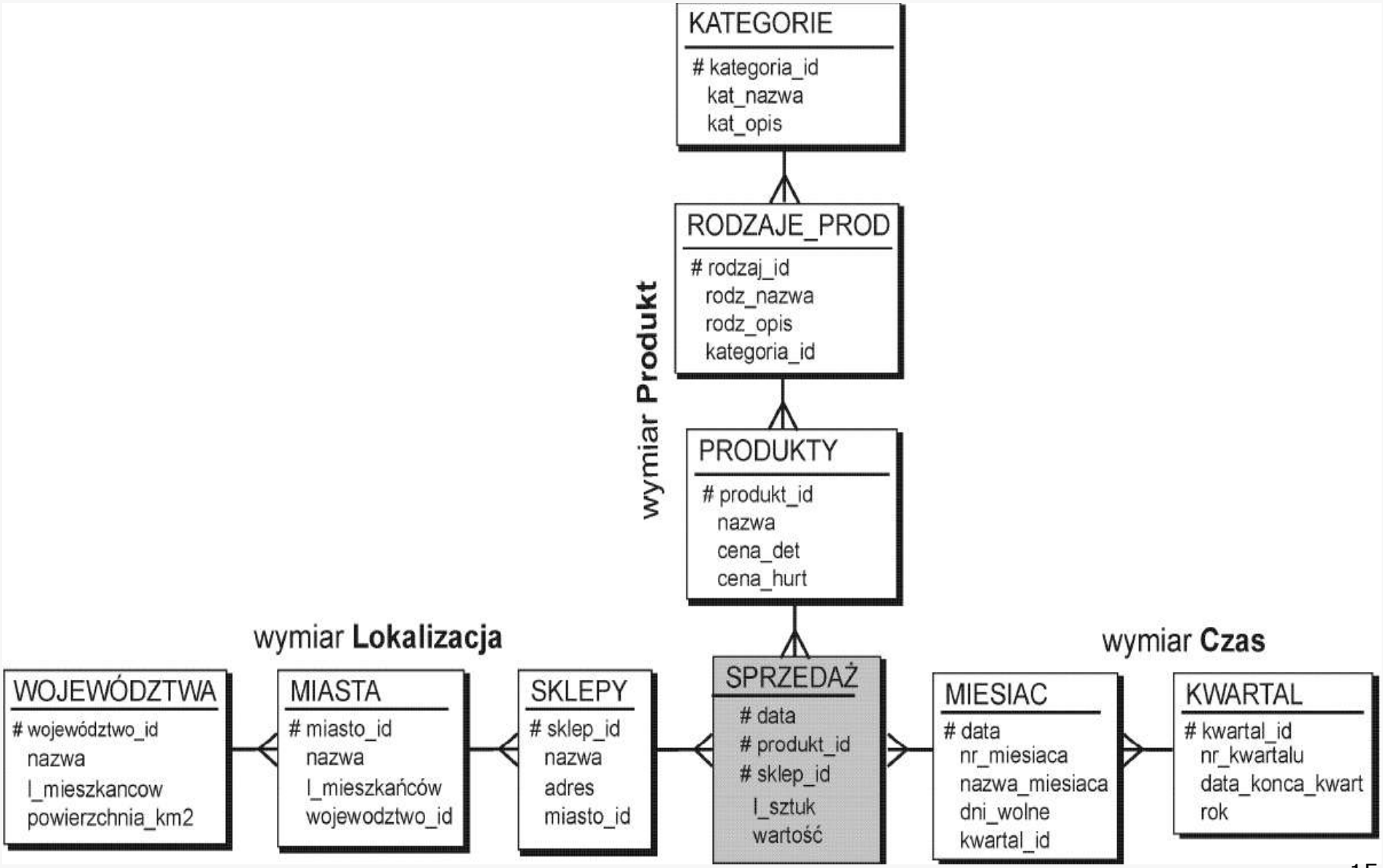
Hierarchie wymiarów

- Dla każdego wymiaru, zbiór opisujących go atrybutów może być ułożony w hierarchiczną strukturę



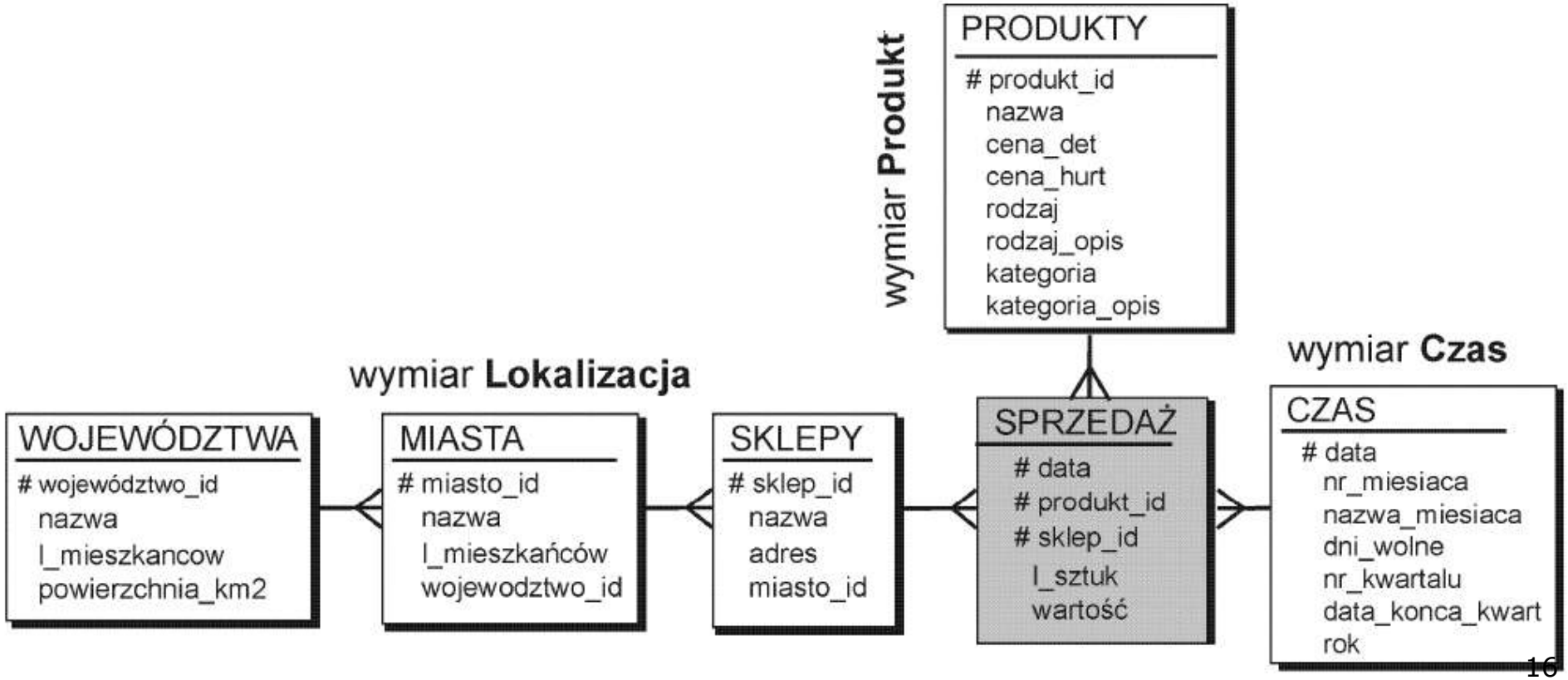
- Tworząc hierarchie wymiarów, normalizujemy wymiary
- Jeśli wymiary są znormalizowane (spełniają przynajmniej 3 postać normalną), wówczas schemat hurtowni danych ma postać **płatka śniegu**

Schemat płatka śniegu



Schemat gwiazda - płatek śniegu

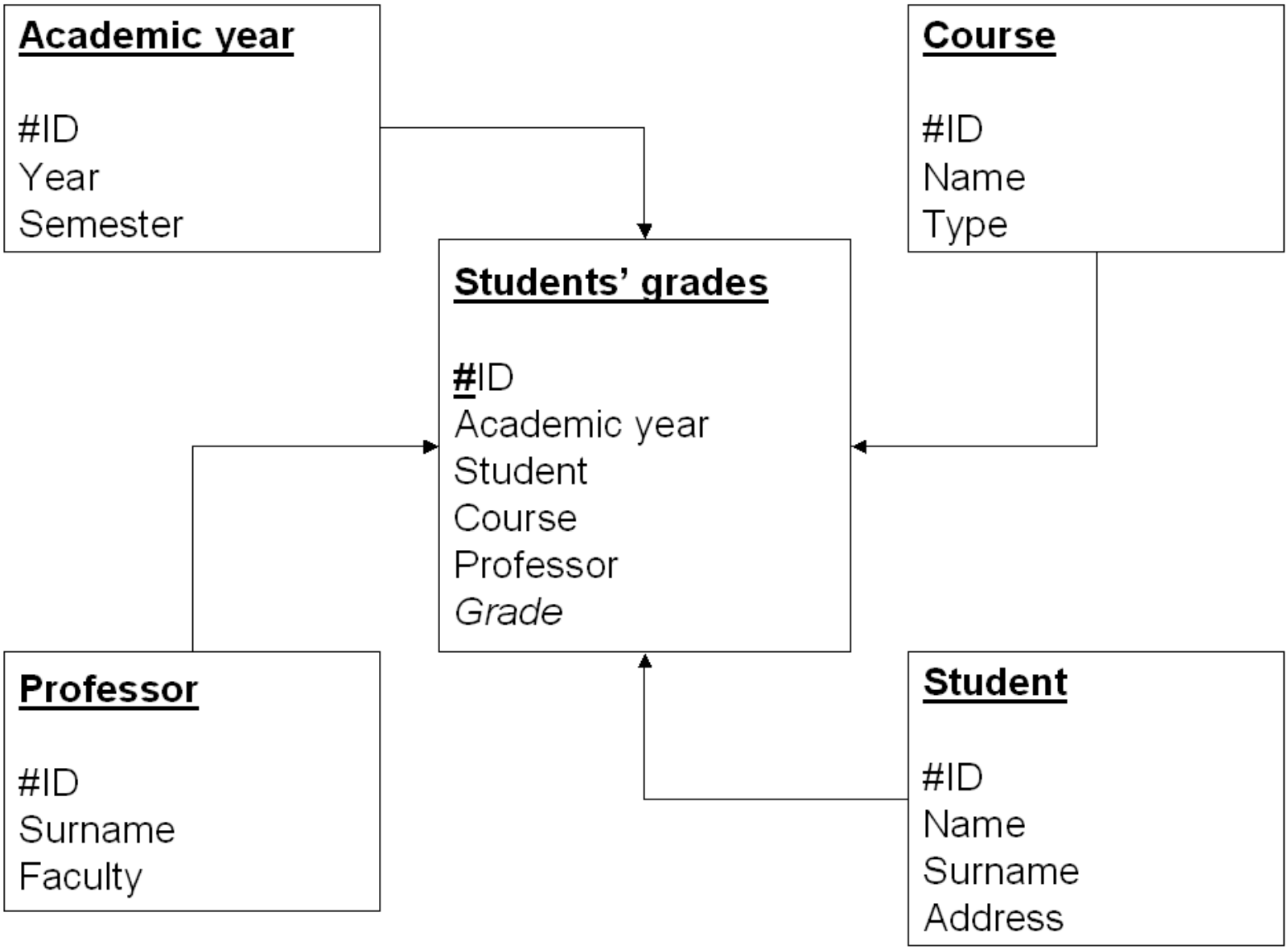
- Schemat, w którym część wymiarów ma postać znormalizowaną, a część ma postać zdenormalizowaną nazywa się schematem **gwiazdy – płatka śniegu**



Schemat gwiazdy/płatka śniegu - zadanie

- **Dla jednostki akademickiej zaprojektuj hurtownię danych zorientowaną na analizę ocen studentów**
  - Zaproponuj relację faktów i wymiary
  - Zaproponuj schemat gwiazdy
  - Zaproponuj hierarchię wymiarów

Schemat gwiazdy/płatka śniegu - zadanie



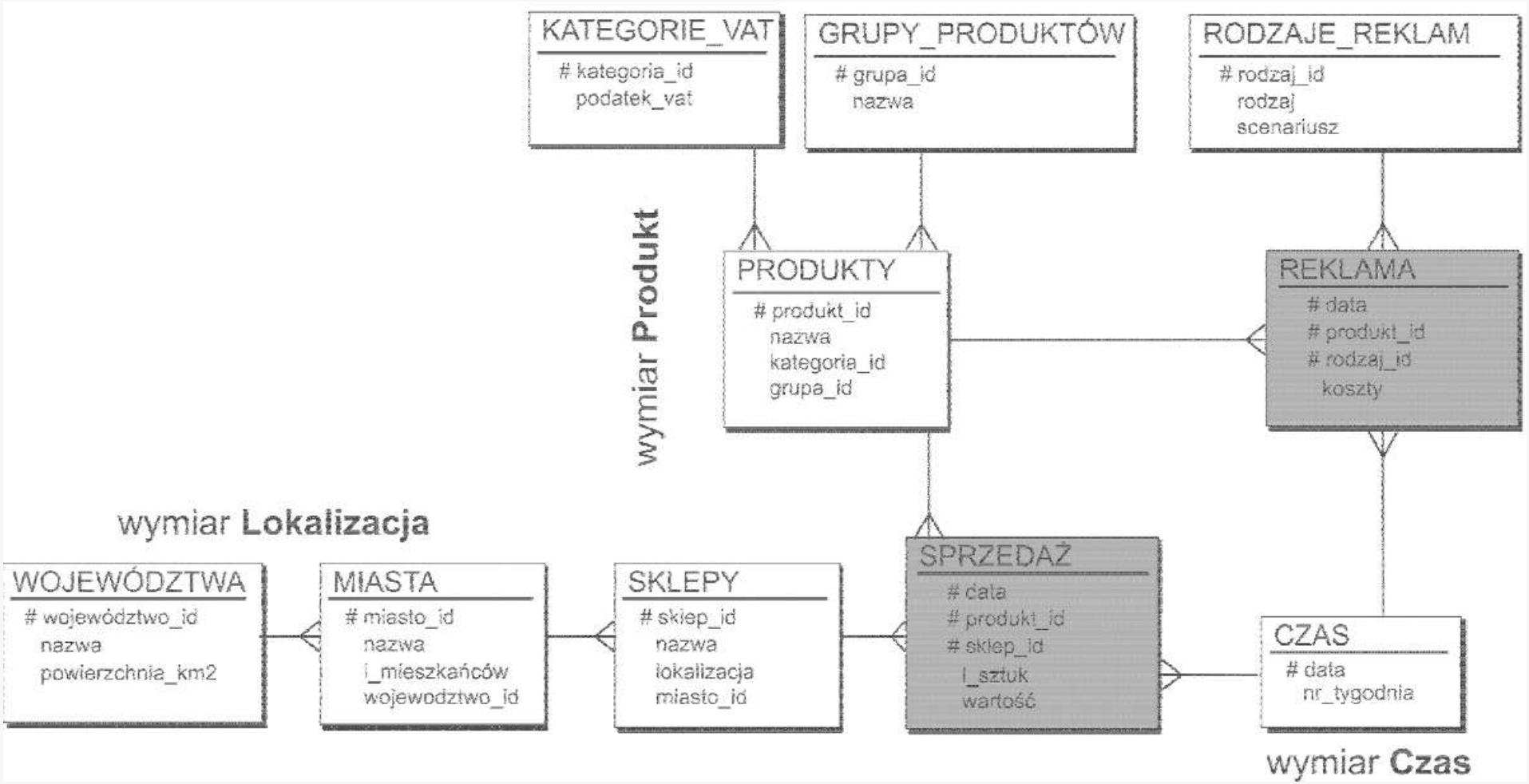
Schemat konstelacji faktów

- **Schemat konstelacji faktów**

  (czyli schemat wielokrotnych tablic faktów):

  - wiele tablic faktów współdzieli relacje wymiarów
  - takie schematy pojawiają się przy projektowaniu hurtowni danych dla dużych i złożonych problemów

Schemat konstelacji faktów



Etapy modelowania koncepcyjnego hurtowni danych

Cztery podstawowe etapy modelowania koncepcyjnego hurtowni danych:

- wybór procesu biznesowego do zamodelowania (np. sprzedaż)
- zdefiniowanie ziarna (rozdzielczości) procesu biznesowego (np. transakcja w sklepie identyfikowana przez skaner przy kasie)
- wybór wymiarów znajdujących się w każdej krotce tablicy faktów (np. lokalizacja sklepu, produkt, data, pora dnia, rodzaj promocji, itp.)
- identyfikacja miar, które wypełnią każdą krotkę tablicy faktów (np. liczba sprzedanych sztuk, łączna wartość sprzedaży)

Wybór procesu biznesowego do zamodelowania

- Powinna to być naturalna aktywność przedsiębiorstwa wspomagana przez operacyjne źródło (np. zakup surowców, zamówienia, dystrybucja, sprzedaż)
- Proces nie może być mylony z działem lub funkcją administracyjną
- Wybór procesu powinien być zależny od jego złożoności, czasu i budżetu przeznaczonego na projekt

Zdefiniowanie ziarna (rozdzielczości) procesu biznesowego

- Należy dokładnie określić znaczenie pojedynczej krotki tabeli faktów
- Rozdzielczość określa jak bardzo szczegółowe dane chcemy przechowywać w hurtowni danych np.
  - transakcja w sklepie identyfikowana przez skaner przy kasie
  - codzienna migawka poziomu inwentarza dla każdego produktu w hurtowni
  - miesięczna migawka dla każdego konta bankowego
- Im większa rozdzielczość, tym większy rozmiar i szybsze rozrastanie się hurtowni danych
- Im mniejsza rozdzielczość, tym mniej dokładny proces wspomagania decyzji

Wybór wymiarów

- Zdefiniowanie opisu danych będących wynikiem procesu biznesowego
- Szczegółowy opis ziarna zdefiniowanego w poprzednim kroku
  np. dla transakcji w sklepie może to być wymiar lokalizacji, produktu, daty, pory dnia, rodzaj promocji, itp.
- Rozdzielczość tabeli faktów determinuje rozdzielczość tabel wymiarów
- Jeżeli dowolny wymiar występuje w dwóch tabelach faktów, muszą to być dokładnie takie same wymiary lub jeden z wymiarów jest podzbiorem drugiego

Identyfikacja miar

- Zdefiniowanie tego, co chcemy zmierzyć
- Każda miara (jak również krotka i wymiar) w tabeli faktów muszą być na tym samym poziomie szczegółowości
- Miary powinny być numeryczne, najlepiej addytywne, ewentualnie częściowo-addytywna (liczba jabłek i pomarańczy)

Planowanie hurtowni dla całego przedsiębiorstwa

- **Korporacyjna hurtownia danych** (ang. Data Warehouse)
  - „Odpytywalne” źródło danych o przedsiębiorstwie
  - Suma logiczna wszystkich składowych hurtowni tematycznych
- **Tematyczna hurtownia danych** (ang. Data Mart)
  - Logiczna część składowa korporacyjnej hurtowni danych
  - Zawężenie hurtowni korporacyjnej do pojedynczego procesu biznesowego lub grupy powiązanych ze sobą procesów skierowanych do konkretnej grupy biznesowej użytkowników
  - Tabele wymiarów lub faktów współdzielone pomiędzy różnymi hurtowniami tematycznymi muszą mieć jedną definicję obowiązującą w całej hurtowni korporacyjnej (conformed dimensions, facts)

Macierz procesów biznesowych i wymiarów

- W procesie projektowania korporacyjnej hurtowni pomocne jest zastosowanie macierzy identyfikującej:
  - procesy biznesowe
  - wymiary

- Przecięcia w macierzy wskazują, które procesy biznesowe korzystają z których wymiarów

Macierz procesów biznesowych i wymiarów

	Data	Klient	Numer telefonu	Plan taryfowy	Kanał sprzedaży	Linia serwisowa #	Producent	Organizacja	Pracownik	Produkt	Rodzaj usterki
Rachunki miesięczne	x	x	x	x	x						
Naprawy	x	x				x	x	x	x	x	x
Zakupy	x	x		x			x	x			

Fragment macierzy procesów i wymiarów dla firmy telekomunikacyjnej

Współdzielone wymiary

- Wymiary współdzielone przez różne tablice faktów muszą utrzymywać tę samą definicję we wszystkich hurtowniach tematycznych, które z nich korzystają
- Odpowiednie zaprojektowanie, zbudowanie i utrzymanie współdzielonych wymiarów to bardzo istotny aspekt pracy nad hurtownia korporacyjną
- **Współdzielenie wymiarów pozwala na:**
  - oszczędzanie fizycznego miejsca na dysku (nie składujemy redundantnych tabel)
  - spójną i jednoznaczną interpretację atrybutów znajdujących się w wymiarach, a co za tym idzie spójną interpretację wszelkich podsumowań w różnych hurtowniach tematycznych

Dodatkowe aspekty modelowania hurtowni

Dane szczegółowe vs. próbkowanie

- Zastępując dane szczegółowe przez reprezentatywna próbkę, można znacząco zmniejszyć wolumen danych pamiętanych w relacji faktów
- Reszta danych jest przechowywana wówczas jako dzienne lub tygodniowe agregaty
- Metoda próbkowania nie nadaje się do hurtowni, w których wymagana jest szczegółowa znajomość wszystkich faktów

Modelowanie wymiaru czasu

- Wymiar czasu jest nieodłącznym wymiarem w projekcie logicznym
- Atrybuty w wymiarze czasu:
  - konkretny czas (klucz główny)
  - dzień miesiąca, dzień tygodnia, weekend
  - 24-godzinny dzień pracy
  - święto publiczne
  - tydzień roku
  - miesiąc, nazwa miesiąca, kwartał, rok
  - rok finansowy
- Brak konieczności wykorzystywania funkcji czasowych (mniejszy koszt obliczeń)
- Możliwość stosowania indeksów do wymiaru czasu

Sztuczne klucze główne vs. klucze naturalne (np. PESEL)

- Klucz sztuczny (ang. surrogate key) może być krótszy, co może poprawić wydajność
- Łatwiejsza obsługa wyjątkowych przypadków (np. brak konkretnych danych – w takim przypadku lepiej jest dodać specyficzny wiersz w relacji wymiaru: “wartość nieznana”)
- Brak nadinterpretacji wartości klucza
- Odporność na zmianę znaczenia klucza naturalnego
- Odporność na ponowne wykorzystanie dawnej wartości klucza naturalnego

Zdegenerowane wymiary

- Niektóre wymiary mają raczej znaczenie **identyfikatora** niż interesujących atrybutów
  np. w hurtowni danych dla sklepu detalicznego `ID_TRANSAKCJI` jest jedynie unikalnym identyfikatorem pozwalającym na połączenie produktów zakupionych w jednym koszyku
- Możliwe podejścia:
  - nie brać pod uwagę `ID_TRANSAKCJI` podczas tworzenia hurtowni
  - utworzyć z `ID_TRANSAKCJI` **zdegenerowany wymiar** (ang. degenerate dimension):
    - nie jest tworzona osobna tablica
    - identyfikator jest bezpośrednio wprowadzany do tabeli faktów
    - możliwa jest analiza np. wielkości koszyka

Agregaty

- Bezpośrednia analiza relacji faktów nie pozwala na wyciąganie wniosków na temat ogólnych prawidłowości i regularności w danych
- Analizy powinny koncentrować się wokół podsumowań i zestawień tworzonych np. na poziomie całej grupy klientów, a zatem ogromnych wolumenów danych
- Umieszczenie w hurtowni tzw. **relacji zbiorczych – agregatów** zwiększa efektywność wykonywania zapytań podsumowujących i zbiorczych (analitycznych)
- **Agregacja** to wykonanie wstępnych obliczeń, materializowanie danych zbiorczych w celu późniejszego ich wykorzystania

Agregaty

- Pożądane cechy agregatów:
  - zapewnienie zauważalnego wzrostu szybkości działania hurtowni (zwiększenie efektywności wykonywania zapytań analitycznych)
  - materializowanie jedynie najczęściej wykorzystywanych danych, by nie powiększać nadmiernie rozmiaru hurtowni
  - struktura agregatów musi pozostać przeźroczysta dla końcowych użytkowników i projektantów aplikacji

Wolno zmieniające się wymiary

- Tablice faktów zmieniają się dużo dynamiczniej niż tablice wymiarów
  - nowe transakcje (fakty) w sposób ciągły dodawane są do relacji faktów
  - nowe produkty, sklepy pojawiają się znacznie rzadziej

- Co jednak robić gdy wartości atrybutów wymiarów ulegną zmianie, np.:
  - klient zmienia adres
  - reforma administracyjna
  - zmiana kategoryzacji produktu

Wolno zmieniające się wymiary

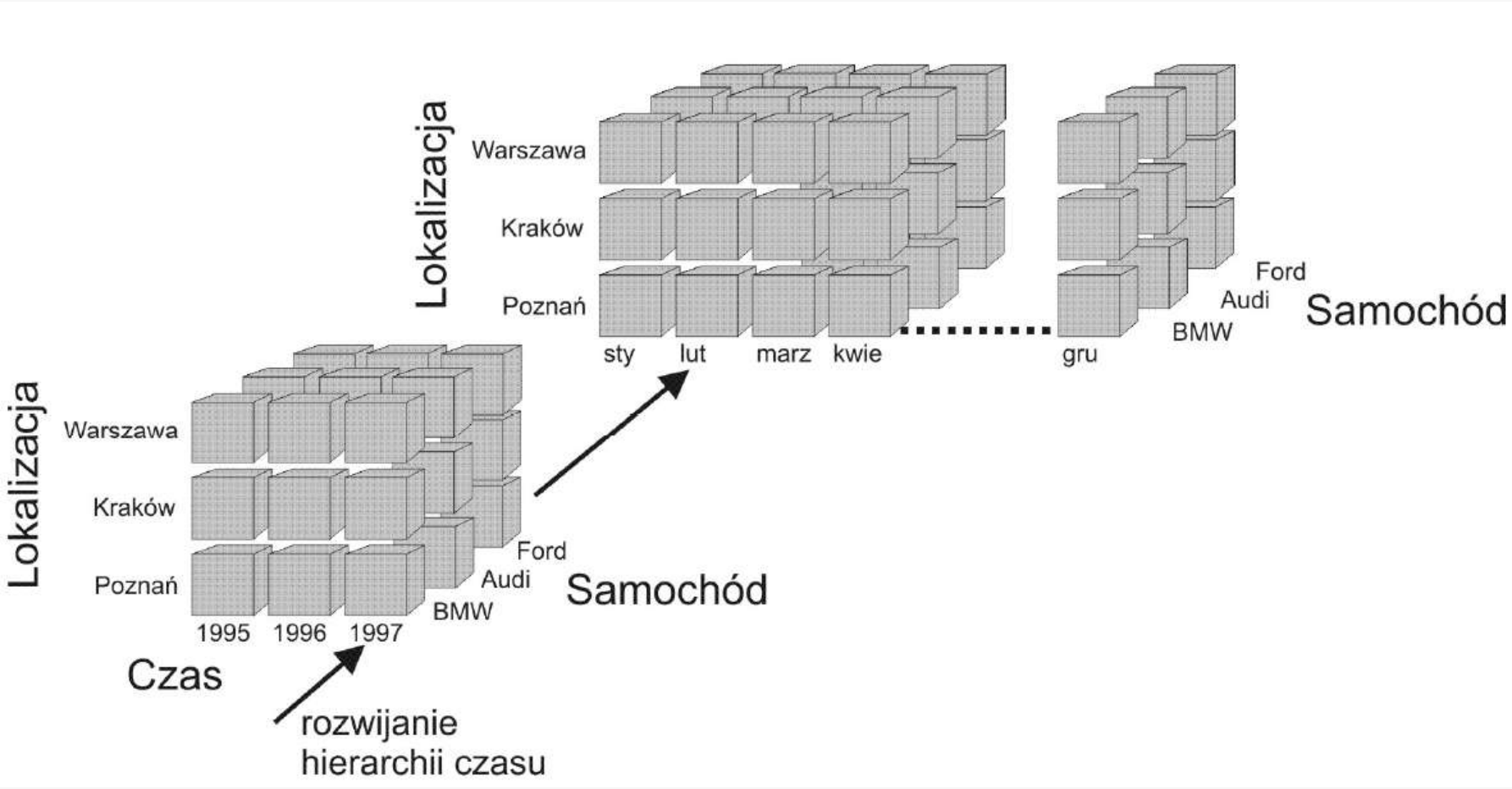
- Możliwe rozwiązania (przykłady!):
  - nadpisywanie starej wartości
  - tworzenie nowych rekordów ze zmienioną wartością
  - tworzenie nowych atrybutów zawierających nowe wartości

Operacje na modelu wielowymiarowym

Podstawowe operacje na modelu wielowymiarowym

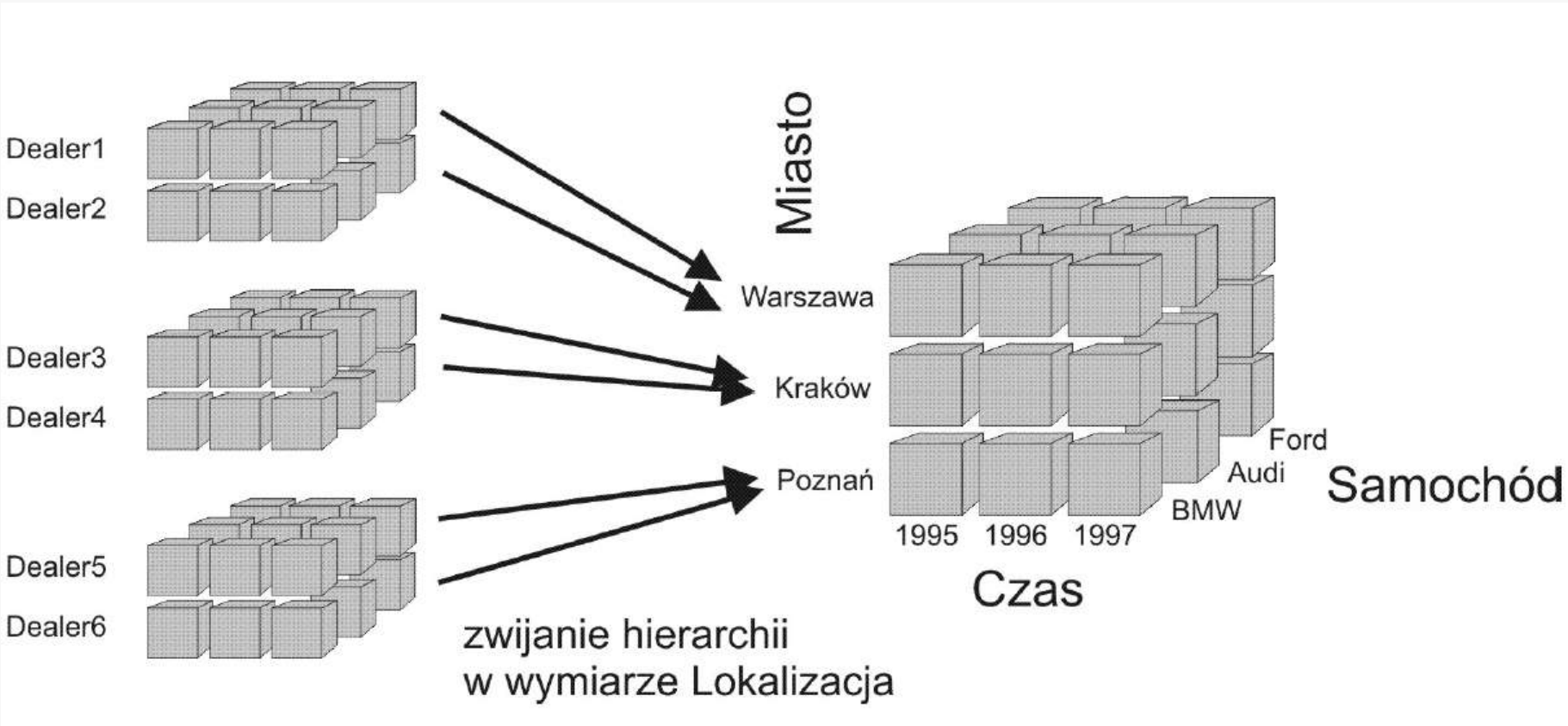
Analizę danych wielowymiarowych wspomagają specjalne operatory:

- **rozwijanie** (ang. drill-down) - polega na zagłębianiu się w hierarchię danego wymiaru w celu przeprowadzenia bardziej szczegółowej analizy



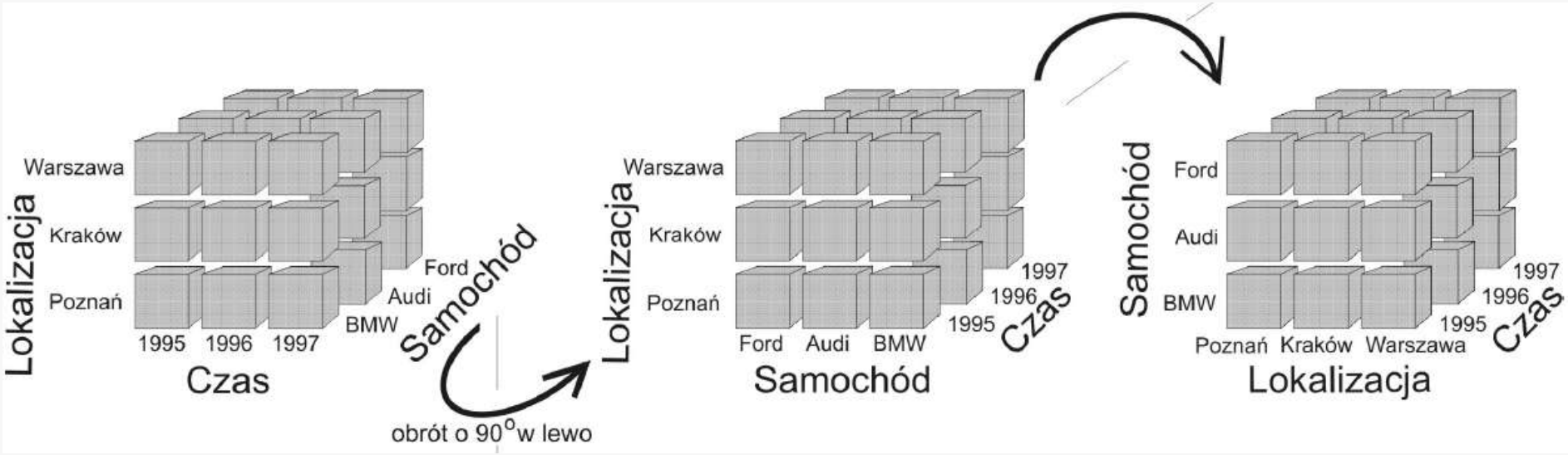
Podstawowe operacje na modelu wielowymiarowym

- **zwijanie**, agregowanie (ang. roll-up) - operacja odwrotna do rozwijania, polega na nawigowaniu w górę hierarchii danego wymiaru, by przeprowadzać analizę danych zagregowanych na wyższym poziomie hierarchii wymiarów



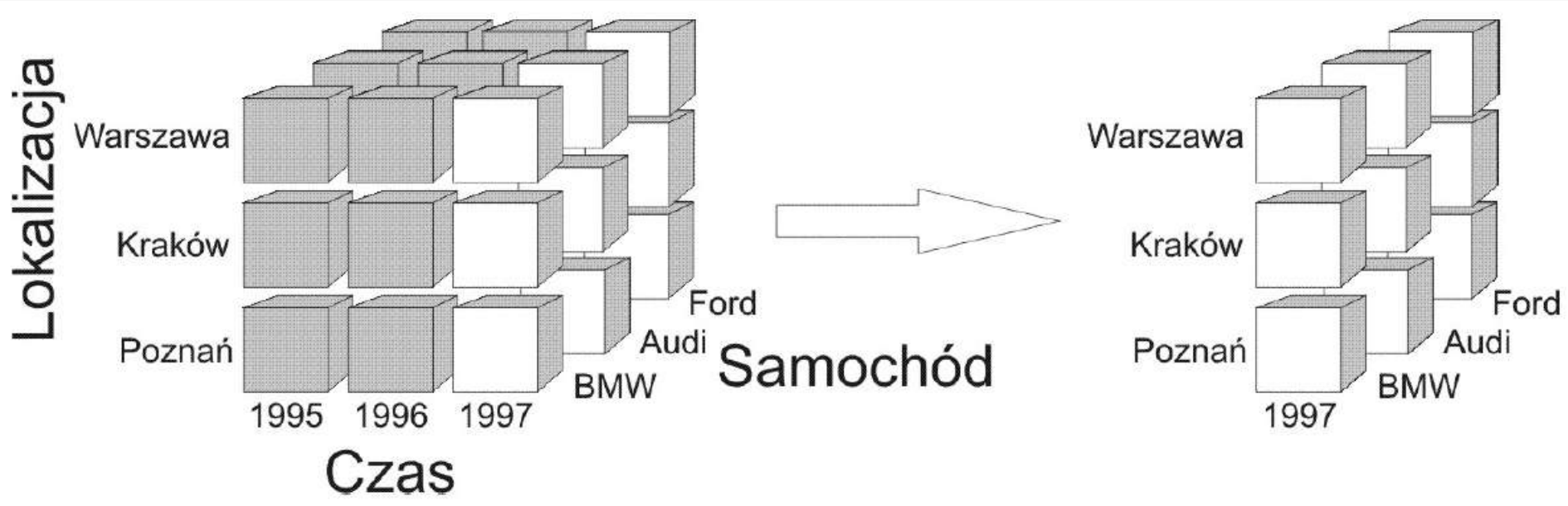
Podstawowe operacje na modelu wielowymiarowym

- **obracanie** (ang. rotating, pivoting) - umożliwia prezentowanie danych w różnych układach, w celu zwiększenia czytelności analizowanych danych



Podstawowe operacje na modelu wielowymiarowym

- **wycinanie** danych w różnych wymiarach (ang. slicing, dicing) – polega na wyborze określonych fragmentów danej wielowymiarowej w celu zawężenie analizowanych danych do wybranych wymiarów, a w ramach każdego z wymiarów – zawężenie analizy do konkretnych jego wartości



Podstawowe operacje na modelu wielowymiarowym

- **filtrowanie** (ang. screening, filtering) - polega na ograniczeniu zakresu analizowanych danych na podstawie zadanego kryterium, np. analiza sprzedaży tylko tych artykułów, których sprzedaż przekraczała 10 000 zł
- **wyznaczanie rankingu** (ang. ranking) – polega na poukładaniu elementów danego wymiaru w malejącej lub rosnącej wartości agregatów,
  np. dla miary *wielkości sprzedaży* można ułożyć elementy wymiaru Samochód zgodnie z malejącą wartością sprzedaży

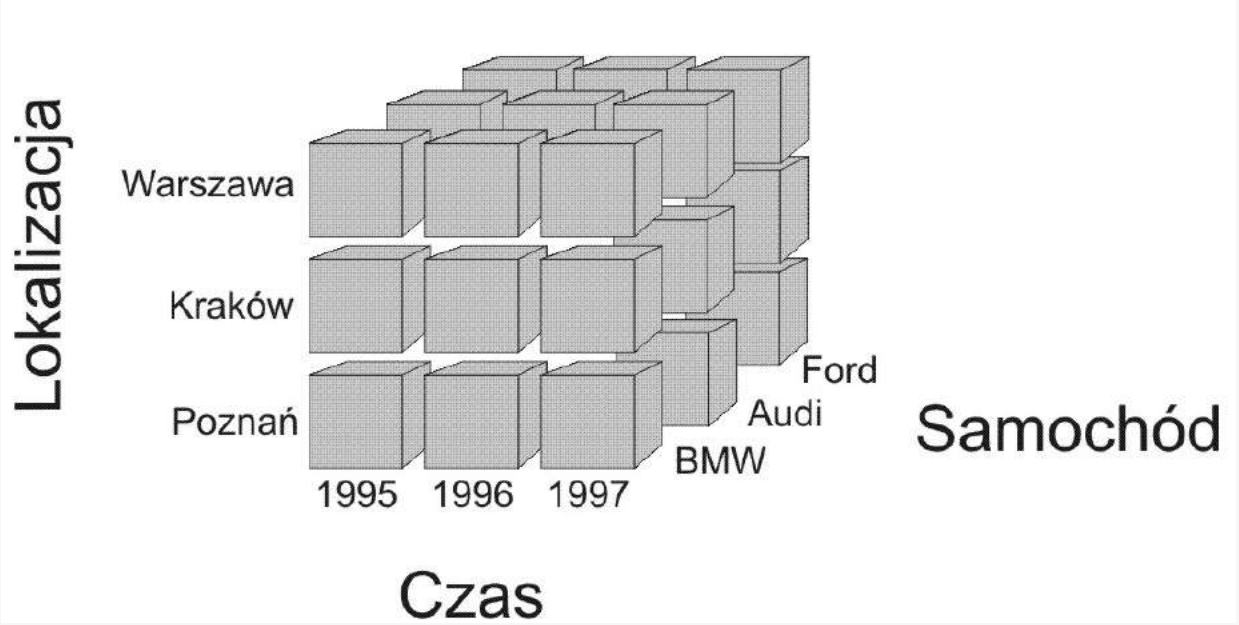
Implementacje modelu wielowymiarowego

Implementacje modelu wielowymiarowego

- **ROLAP** (ang. relational OLAP) - implementacja w serwerach relacyjnych
  - fakty przechowywane w tabelach faktów
  - wymiary przechowywane w tabelach wymiarów
- **MOLAP** (ang. multidimensional OLAP) - implementacja w serwerach wielowymiarowych
  - dane przechowywane w wielowymiarowych tabelach (ang. data cubes), zwanych potocznie kostkami
- **HOLAP** (ang. hybrid OLAP) - implementacja hybrydowa (relacyjno-wielowymiarowa)
  - dane elementarne przechowywane w tabelach
  - dane zagregowane przechowywane w kostkach

Implementacja MOLAP

- Przykładowa wielowymiarowa tablica zawierająca trzy wymiary: *Lokalizacja, Czas i Samochód*
- *Komórki tablicy zawierają* zagregowane informacje o sprzedaży wybranych samochodów (BMW, Audi, Ford) w poszczególnych latach (1995, 1996, 1997), w wybranych miastach (Poznań, Kraków, Warszawa)



Implementacja MOLAP

- Od strony implementacyjnej dane wielowymiarowe mogą być przechowywane nie tylko w wielowymiarowych tablicach. Możliwe jest wykorzystanie również np.:
  - dużych obiektów binarnych (BLOB)
  - tablicy haszowych
  - struktur drzewiastych takich jak Quad-drzewa czy K-D-drzewa

Literatura

- R. Kimball, M. Ross,
  The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling,
  John Wiley & Sons 2002
- M. Jarke, M. Lenzerini, Y. Vassiliou, P. Vassiliadis,
  Hurtownie danych. Podstawy organizacji i funkcjonowania,
  Wydawnictwa Szkolne i Pedagogiczne 2003
- Z. Królikowski,
  Hurtownie danych: logiczne i fizyczne struktury danych,
  Wydawnictwo Politechniki Poznańskiej 2007
- Materiały e-learningowe z przedmiotu „**Zaawansowane systemy baz danych**” (http://wazniak.mimuw.edu.pl/)