Hurtownie danych, Eksploracja danych

dr inż. Artur Gramacki
a.gramacki@iie.uz.zgora.pl

Instytut Informatyki i Elektroniki
Uniwersytet Zielonogórski

Business Intelligence (BI)

- Na początek tłumaczenie
  - inteligencja biznesowa (fatalnie!)
  - analityka biznesowa (lepiej?)
  - usługi biznesowe (lepiej?)
  - przetwarzanie analityczne (lepiej?)
  - inne?

Business Intelligence (BI)

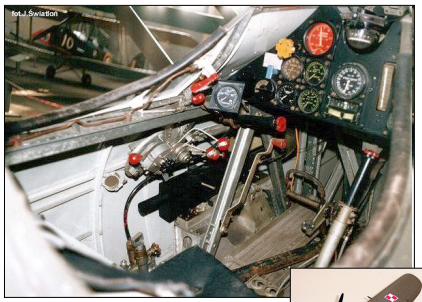
- Zbiór programów i technologii informatycznych umożliwiających zbieranie i przetwarzanie danych w informacje a informacji w użyteczną wiedzę w celu ułatwienia użytkownikowi podejmowanie lepszych decyzji biznesowych
  - dane → informacje → wiedza → decyzje biznesowe
- Przeznaczone dla pracowników „od dyrektora wzwyż”
- Efektywne eksploatowanie BI jest bardzo mocno uzależnione od utworzenia hurtowni danych
  - wielkie ilości danych, drastyczne wymagania wydajnościowe, specyfika wykonywanych zadań (innych nich w klasycznych bazach danych)

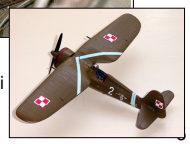
Mnogość pojęć z okolic BI

- ERP, planowanie zasobów przedsiębiorstwa (ang. Enterprise Resource Planning)
- CRM, zarządzanie relacjami z klientami (ang. Customer Relationship Management)
- EIS, systemy informowania kierownictwa (ang. Executive Information Systems)
- DSS, systemy wspomagania decyzji (ang. Decision Support Systems)
- MIS, systemy informacyjne zarządzania (ang. Management Information Systems)
- BPD, odkrywanie procesów biznesowych (ang. Business Process Discovery)
- IBP, zintegrowane planowanie buznesu (ang. Integrated business planning)
- i wiele, wiele innych …

Pożądany (?) efekt finalny

- Panel z najważniejszymi informacjami biznesowymi
  - BI dashboard
  - modne ostatnio określenie: kokpit informacyjny





(**PZL P.11**, zwany „jedenastką” – polski samolot myśliwski konstrukcji inżyniera Zygmunta Puławskiego z okresu przed II wojną światową)

Kokpitów ci u nas dostatek ☺

Systemy przetwarzanie analitycznego

- Wymagania
  - dostęp do kompletnych i aktualnych danych dotyczących całego przedsiębiorstwa (firmy, organizacji)
- Cel
  - trafne podejmowanie decyzji biznesowych
  - możliwość łatwego tworzenia zbiorczych analiz dotyczących całości przedsiębiorstwa
  - odkrywanie na bazie posiadanych danych nowych, nieznanych przedtem, informacji (eksploracja danych)
- Cechy
  - duża wydajność pracy z ogromnymi ilościami danych
  - łatwe scalanie danych „detalicznych” do postaci „hurtowej”
  - łatwy dostęp do danych historycznych

Technologie wspierające BI

- BI = OLTP + OLAP + DW + DM + ETL + inne
- OLTP OnLine Transaction Processing
- OLAP OnLine Analytical Processing
- DW Data Warehouse
- DM Data Mining
- ETL Extraction Transformation Loading
- inne

Systemy OLTP

- OLTP – OnLine Transaction Processing
  - OLTP zwane są często operacyjnymi bazami danych
  - przede wszystkim mają w bezpieczny sposób gromadzić bieżące dane (jeżeli też historyczne, to niezbyt „odległe”)
    - kompleksowe zabezpieczenia przed: utratą danych, utratą spójności danych, nieautoryzowanym dostępem
  - zwykle przetwarzają wielką liczbę stosunkowo krótkich transakcji
    - INSERT, UPDATE, DELETE
  - wielodostęp (często w reżimie 24 –7–365)
  - wymagany szybki dostęp do bieżących danych
    - dla wszystkich pracujących w danej chwili użytkowników
  - też praca w trybie wsadowym (np. nocne aktualizacje danych)
  - najczęściej bazują na klasycznych relacyjnych bazy danych
    - zwykle wysoce znormalizowane (przynajmniej 3 PN)

Systemy OLAP

- OLAP – OnLine Analytical Processing
  - OLAP zwane są często analitycznymi bazami danych
  - tworzone przede wszystkim w celu efektywnego analizowania wielkich ilości danych
    - podsumowania, raportowanie, prognozowanie, ocena, itp.
  - gromadzą w zasadzie dane tylko historyczne
    - zmiany w danych wyłącznie sporadyczne (np. usuwanie zauważonych błędów)
    - dane są często uzupełniane ale bardzo rzadko kasowane
    - głównie odczyt danych (SELECT)
    - niewielka liczba ale za to długich transakcji
  - najczęściej bazują na hurtowniach danych
    - zwykle słabo znormalizowane, dane silnie zagregowane
  - umiarkowany wielodostęp

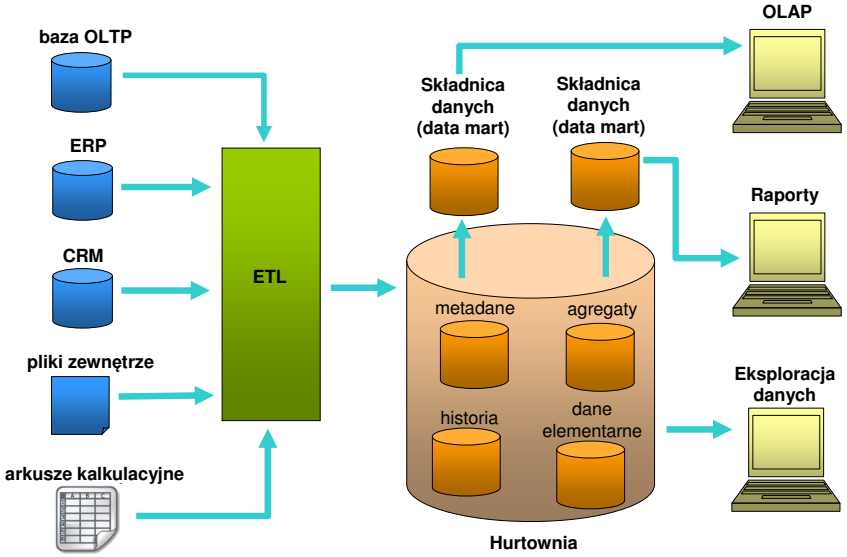
OLAP vs. OLTP

Element / cecha	OLAP	OLTP
Źródło danych	Dane skonsolidowane, zwykle na bazie innych systemów OLTP	Dane operacyjne na bazie typowych baz relacyjnych
Zakres czasowy	Dane wieloletnie	Dane z ostatnich tygodni, miesięcy
Cel posiadania	Analizy, planowanie, wspieranie w podejmowaniu decyzji biznesowych	Bieżące przetwarzanie danych zapewniające właściwe działanie przedsiębiorstwa
Dla kogo	Kadra zarządzająca	Pracownicy niższego szczebla
Wykonywane operacje	Cykliczne odświeżanie danych (zwykle wstawianie, rzadko kasowanie)	Dużo operacji wstawiania, modyfikacji i kasowania danych
Rodzaj zapytań	Złożone zapytania, odwołujące się do danych zagregowanych	Dużo prostych zapytań odwołujących się do wielu tabel i widoków relacyjnych
Czas odpowiedzi	Może być bardzo długi (nawet godziny)	Bardzo szybki (maksymalnie kilka sekund, lepiej mniej)
Wymagane zasoby	Wielkie (dużo danych wstępnie zagregowanych, dużo danych historycznych, dużo indeksów)	Relatywnie małe (gdy nie przechowujemy zbyt dużo danych historycznych)
Model danych	Dane zdenormalizowane, typowe dla HD układy (gwiazda, płatek śniegu)	Relacyjny (mocno znormalizowany), wiele tabel, widoków

Hurtownie danych (Data Warehouse)

- **W.H. Inmon**, 1992, *Building the Data Warehouse*
- „A warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process”
  - tematyczna baza danych
  - dane zintegrowane
  - (dane historyczne) opisane wymiarem czasu
  - trwałe
  - do analiz biznesowych
- Więc HD to: tematyczna baza zintegrowanych danych historycznych, opisanych wymiarem czasu, przeznaczona głównie do wykonywania analiz biznesowych

Typowe architektury HD



Hurtownie danych – rodzaje danych

- Dane elementarne
  - skopiowane z baz operacyjnych, odpowiednio „wyczyszczone”
- Agregaty
  - np. wyliczone średnie i podsumowania w różnych przekrojach
    - np. dla poszczególnych miesięcy, obszarów, lat
  - zmaterializowane
  - różny stopień przetworzenia
- Dane historyczne
  - dane z „dalekiej przeszłości”, elementarne jak i po agregacji
  - usuwane tylko wyjątkowo
- Metadane
  - słowniki, opisy struktury HD, opisy pochodzenia danych, algorytmy wykonywanych agregacji

Eksploracja danych (Data Mining)

- Wywodzi się ze statystki
  - nauka badająca prawidłowości występujące w zjawiskach o charakterze masowym
  - wnioskowanie statystyczne, model statystyczny, weryfikacja hipotez
  - wiemy mniej więcej czego szukamy w danych
- Eksploracja danych
  - idzie dalej niż klasyczna statystyka
  - odkrywanie wiedzy z danych
  - zwykle nie wiemy dokładnie, czego szukamy w danych, trudno jest sformułować precyzyjną hipotezę (klasyczny przykład: grupowanie danych)
  - gromadzenie danych a wyciąganie z nich sensownych wniosków (eksploracja danych) to dwa bardzo różne zagadnienia!

Procesy ETL

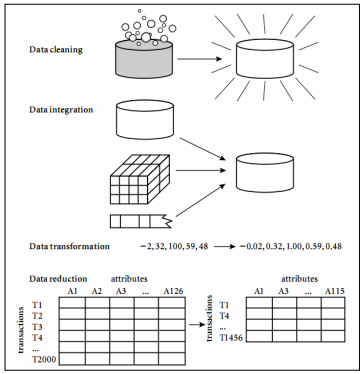
- ETL – Extraction Transformation Loading
- Ekstrakcja (pozyskanie) danych ze źródeł zewnętrznych
  - zwykle z baz OLTP
  - często z „płaskich” plików tekstowych, arkuszy kalkulacyjnych oraz plików XML-owych
- Transformacja
  - wstępna obróbka danych (czyszczenie, integracja, transformacja, redukcja)
- Ładowanie
  - wprowadzenie pozyskanych danych do hurtowni
- projekt ETL-a jest bardzo trudnym etapem implementacji HD. Ocenia się, że pochłania ok. 70% czasu!

Procesy ETL

- 2 etapy
  - projektowanie procesów ETL
    - wymagane bardzo dokładne poznanie danych źródłowych
    - wykorzystanie różnych narzędzi wspomagających
    - ciągłe kontrolowanie poprawności (struktury z danymi źródłowymi mogą podlegać zmianom)
  - działanie procesów ETL
    - mocno obciążają system
    - inny charakter obciążenia systemu niż w czasie realizacji zapytań analitycznych
      - dodatkowa trudność dla administratora !
    - zwykle procesy ETL „puszcza" się w czasie małej aktywności hurtowni danych (np. w nocy, w każdą niedzielę)
    - ciągłe kontrolowanie poprawności (struktury z danymi źródłowymi mogą podlegać zmianom)

Wstępna obróbka danych

- Podstawowe czynności do wykonania:
  - czyszczenie danych
    - poprawianie „brudnych" danych
    - detekcja punktów oddalonych
    - analiza brakujących danych
  - integracja (scalanie) danych
  - transformacja (konwersja) danych
  - redukcja danych
    - redukcja wymiarowości
    - redukcja liczności



Jiawei Han, Micheline Kamber: Data Mining: Concepts and Techniques, Morgan Kaufman, 2000

Czyszczenie danych

Poprawianie „brudnych" danych

- Tzw. literówki, „czeskie błędy"
  - lubuskie, lubelskie, lubuksie
- Niejednoznaczny format daty i daty nieistniejące
  - 01/05/05, 01-05-2005, 13/13/13, 20-20-2010
- Różne format liczb
  - 1000, 1E3, 2,45, 2.45
- Wiele wariantów zapisu tych samych danych
  - woj. lubuskie, województwo lubuskie, lubuskie
  - Uniwersytet Zielonogórski, Uniw. Ziel., UZ
  - lubuskie, LUBUSKIE, Lubuskie
- Niezgodności merytoryczne
  - np. w polu nazwisko wpisano nazwę firmy
- Dublowanie się danych
  - np. nazwa jednej firmy pojawia się w wielu miejscach

Czyszczenie danych

Poprawianie „brudnych" danych

- Różne oznaczenia
  - K, M, kobieta, mężczyzna,
  - 0, 1, true, false
- Różne jednostki miar, wag, walut
  - kg, gramy, funty, litry, pinty, PLN, EUR, USD
- Brak wymaganej unikalności
  - nieistniejący numer NIP
  - zdublowane numery NIP
- Niejednoznaczności
  - Robert Nowak, Zielona Góra = Robert Nowak, Nowa Sól
- Kolejność danych
  - A. Gramacki, Gramacki Artur, Gramacki A.
- ... oraz wiele, wiele innych ☹

Czyszczenie danych
Detekcja punktów oddalonych

- Zwykle dane bardzo różniące się od innych (kilka rzędów wielkości) są wynikiem jakiś błędów, np. w pomiarach
  - „zwykle" to nie znaczy oczywiście „zawsze"!
- Zignorowanie problemu → błędne wyniki
- Problemy:
  - uznanie jakiegoś punktu za oddalony, mimo że nim nie jest
  - punkt nie zostanie uznany za oddalony, gdy takim w rzeczywistości jest
- Metody:
  - oparte o najprzeróżniejsze testy i miary statystyczne
  - wizualna inspekcja danych (praktycznie tylko dla 2D/3D)

Czyszczenie danych
Co robić z brakującymi danymi?

- Możliwe podejścia:
  - usunąć rekordy z brakującymi danymi
  - w miejsce brakujących danych wpisać wartość „najbardziej odpowiednią", tzw. imputacja
- Metody zastępowania brakujących danych przez:
  - pewną wartością stałą określoną przez analityka
  - pewną wartością wynikającą z analizy zależności pomiędzy poszczególnymi przypadkami (rekordami)
    - wstawienie wartości średniej, mediany, wartości modalnej itp.
  - pewną wartością wynikająca z analizy zależności pomiędzy poszczególnymi atrybutami (zmiennymi)
    - np. bazując na dopasowaniu modelu regresji do danych
  - inne metody
    - np. tzw. wielokrotna imputacja

Integracja danych

- Cel: połączyć dane z wielu, często bardzo różniących się od siebie, źródeł danych w jedną hurtownię danych
  - chyba najtrudniejszy etap tworzenia HD
- Źródła danych
  - „duże" bazy relacyjne (np. Oracle)
  - „małe" bazy relacyjne (np. MS Access)
  - inne hurtownie danych
  - pliki tekstowe, w tym XML-owe
  - arkusze kalkulacyjne
  - pliki binarne (chyba dość rzadko)

Transformacja danych
Normalizacja oraz skalowanie

- Spostrzeżenie: wartości poszczególnych atrybutów często różnią się od siebie skalą oraz rzędem wielkości
- Skutek: wiele wykonywanych obliczeń może być bądź niestabilnych numerycznie, bądź też bardzo duże wartości mogą całkowicie zdominować te bardzo małe
  - co więc robić? – skalować i normalizować
- Podstawowe metody:
  - normalizacja min-max (często też zwana normalizacją 0-1). Dane po normalizacji będą należeć do przedziału od 0 do 1
  - standaryzacja. Dane po normalizacji będą miały średnią równą 0 i wariancję równą 1
  - centrowanie wokół wartości zero

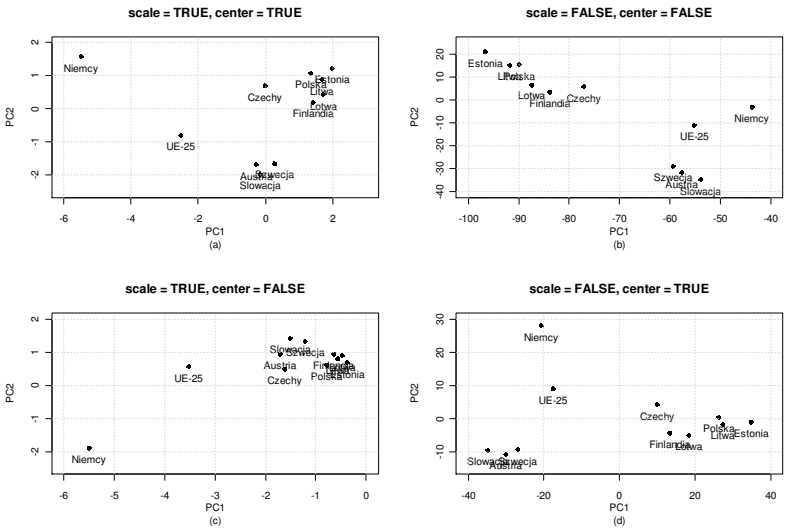
Redukcja wymiarowości – przykł. 1 (1/2)

- Redukcji wymiarowości to proces transformacji danych wielowymiarowych (w sensie dużej ilości atrybutów) do przestrzeni o sensownie mniejszym wymiarze
  - z przyczyn czysto praktycznych zwykle redukujemy do 2D lub 3D
- Czy można coś powiedzieć o podobieństwie krajów?

	X1	X2	X3	X4	X5	X6	X7	X8
UE-25	51,3	0,7	21,4	5,4	3,8	4,0	4,7	8,7
Austria	49,5	1,3	43,5	1,6	0,4	0,8	0,5	2,4
Czechy	76,4	0,1	10,2	0,1	2,8	5,6	0	4,8
Estonia	99,0	0	0,3	0,7	0	0	0	0
Finlandia	82,7	0	14,7	0,2	0,5	0	0	1,9
Litwa	92,9	0	5,0	0	0,3	1,4	0,4	0
Łotwa	86,9	0	12,5	0,2	0,3	0,1	0	0
Niemcy	41,3	2,2	10,1	14,0	8,6	13,1	0,8	9,9
Polska	91,2	0	4,1	0,3	1,2	2,6	0,2	0,4

X1 – biomasa stała, X2 – energia promieniowania słonecznego, X3 – energia wody, X4 – energia wiatru, X5 – biogaz, X6 – biopaliwa, X7 – energia geotermalna, X8 – odpady komunalne

Redukcja wymiarowości – przykł. 1 (2/2)



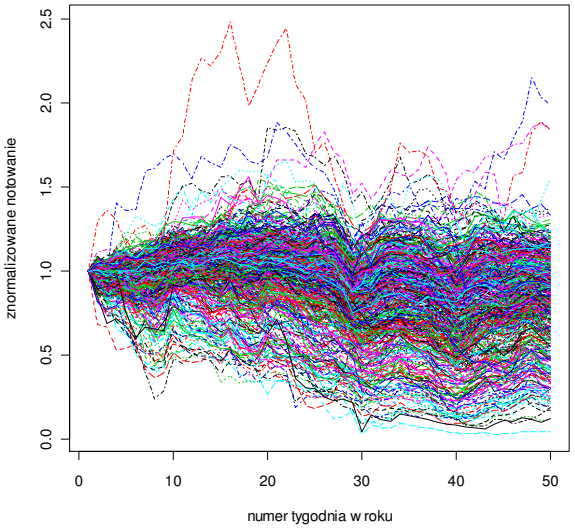
Redukcja wymiarowości – przykł. 2 (1/4)

- Fragment notowań 498 firm z indeksu SP500 w okresie 52 tygodni

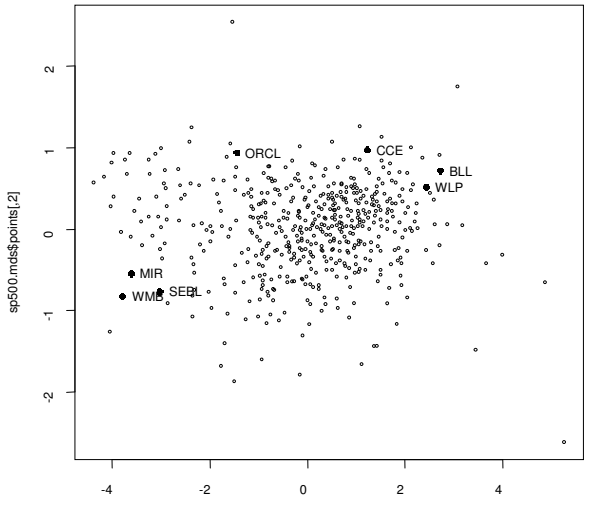
Id firmy	Skrót	A0	A1	A2	...	A50	A51
1	A	1	0.947	0.512	...	0.525	0.869
...	...	...	...	...	...	...	...
342	ORCL	1	1.053	1.067	...	0.73	0.689
...	...	...	...	...	...	...	...
498	ZMH	1	1.013	1.056	...	1.33	1.328

- Zadanie: pogrupować firmy na te, których notowania w analizowanym okresie rosły, były w miarę równe oraz spadały

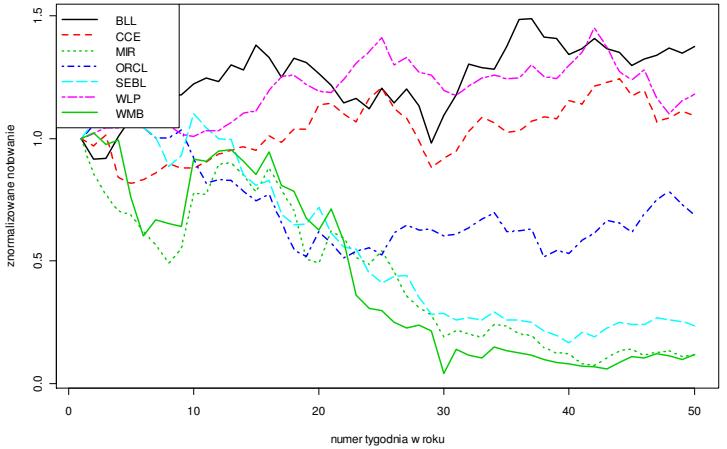
Redukcja wymiarowości – przykł. 2 (2/4)



Redukcja wymiarowości – przykł. 2 (3/4)



Redukcja wymiarowości – przykł. 2 (4/4)

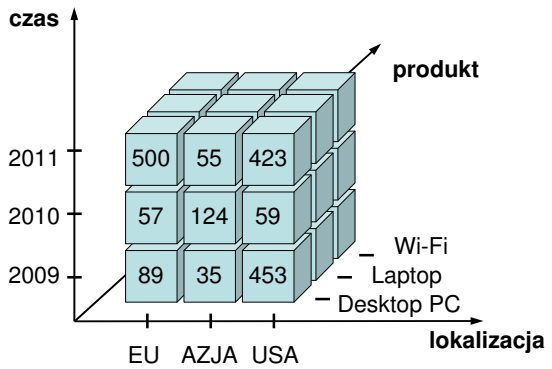


Redukcja liczności

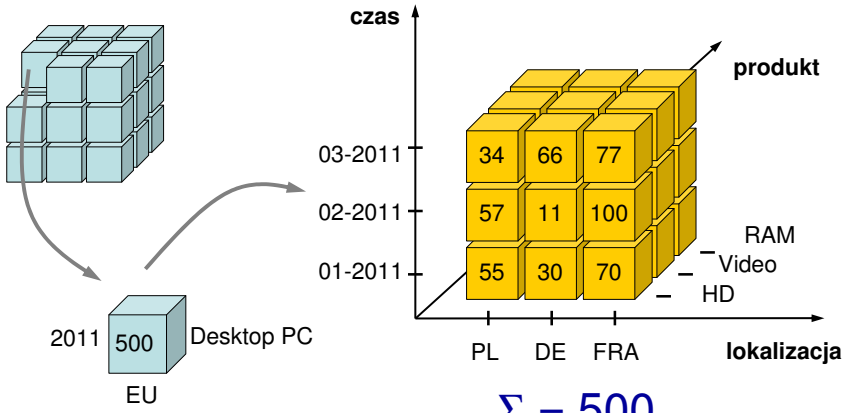
- Redukcji liczności to proces wyboru z posiadanego dużego zbioru danych pewnego podzbioru, zwanego próbą reprezentatywną
  - w terminologii bazodanowej zbiorem danych będzie relacja
  - w terminologii statystycznej zbiorem danych będzie populacja
  - właściwy wybór próby reprezentatywnej jest warunkiem koniecznym dla poprawnego wyciągania wniosków o tej populacji !
- Rozróżnienie:
  - populacja w naukach społecznych, ekonomicznych – niemal zawsze nieznana
  - populacja w bazach danych – zawsze znana
- Nie mylić z redukcją wymiarowości
- Inna nazwa: próbkowanie

Modelowanie hurtowni danych

- Model wielowymiarowy
  - uogólnienie na więcej niż 3 wymiary oczywiste



Model wielowymiarowy



$\Sigma = 500$

Model wielowymiarowy

- Ilustracje z opracowania:
  - „Modele danych i ewolucja systemów baz danych”, Krzysztof Dembczyński, Politechnika Poznańska

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

Model wielowymiarowy

Takie same tabele dla Chicago, Nowego Jorku i Toronto:

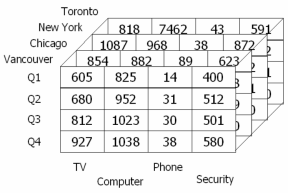
605	825	14	400
680	952	31	512
812	1023	30	501
927	1038	38	580

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:



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 Computer Phone Security

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

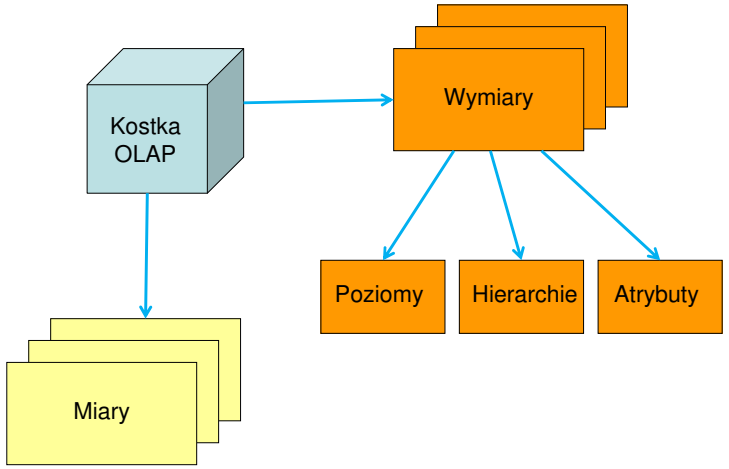
Rodzaje danych wielowymiarowych

- Fakty (ang. facts)
  - elementarne komórki z danymi
  - zwykle zawierają też miary (ang. measures)
  - miary są to zwykle wartości numeryczne ciągłe
    - wielkość sprzedaży, zysk, strata, ilość towaru
    - mogą być składowane (ang. stored measures) albo też być wyliczane on-line w czasie wydawania przez użytkownika zapytania (ang. calculated / derived measures)

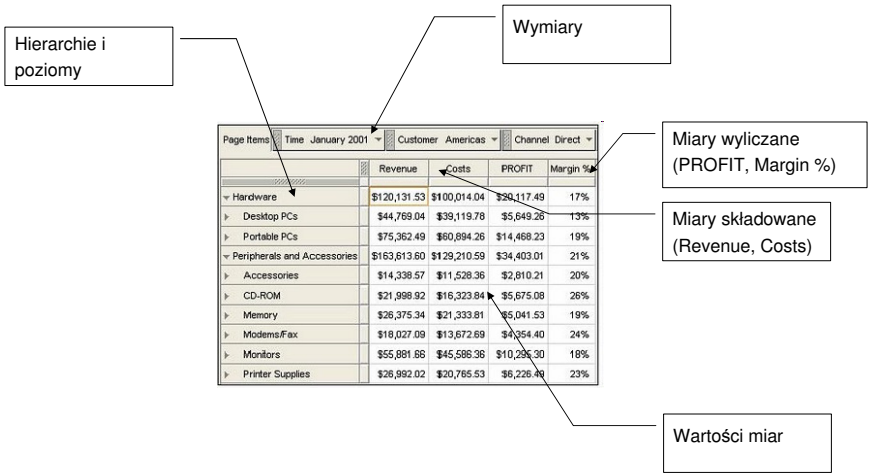
Rodzaje danych wielowymiarowych

- Wymiary (ang. dimensions)
  - nadają znaczenie faktom i miarom
  - zwykle wartości dyskretne, które nadają znaczenie faktom
    - np. produkt, lokalizacja, czas (rys. na poprzedniej stronie)
  - mogą tworzyć hierarchie, hierarchie pozwalają agregować dane
  - elementy w hierarchii są na określanych poziomach (ang. levels)
    - rok → kwartały → miesiące → tygodnie → dni
    - towar → kategoria → podkategoria
    - uczelnia → wydzial → instytut → zakład
  - atrybuty (ang. attributes) dostarczają dodatkowych informacji o danych
    - np. kolor, smak, rozmiar
    - np.”jaki kolor kurtek najlepiej sprzedawała się w lecie 2010?”

Rodzaje danych wielowymiarowych



Typowy raport analityczny

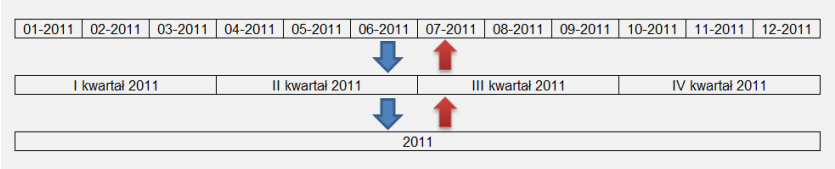


Operacje wielowymiarowe

- Zwijanie, rozwijanie (ang. *roll-up*, *roll-down*)
  - inne nazwy tych operacji to drążenie w górę, drążenie w dół (ang. *drill-up*, *drill-down*)
- Obracanie (ang. *rotating*)
- Wycinanie (ang. *sliping and dicing*)
- Filtrowanie (ang. *filtering*)
- Wyznaczanie rankingu (ang. *ranking*)

Operacje wielowymiarowe

- Zwijanie (rozwijanie)
  - przejście na poziom bardziej ogólny (szczegółowy)
  - inna nazwa: agregacja (de-agregacja)



Operacje wielowymiarowe

- Obracanie
  - wymiary zamieniamy miejscami

Kolumna1	2005	2006	2007	2008	2009	2010	2011
Zielona Góra	21 858,38 zł	21 939,12 zł	47 952,72 zł	53 986,24 zł	47 689,27 zł	74 610,71 zł	56 919,40 zł
Poznań	45 387,68 zł	74 232,79 zł	6 146,71 zł	98 505,45 zł	49 436,76 zł	90 574,28 zł	14 304,67 zł
Wrocław	74 837,31 zł	85 346,91 zł	38 194,06 zł	42 979,83 zł	78 078,34 zł	81 408,35 zł	22 206,96 zł
Warszawa	85 771,88 zł	18 860,00 zł	27 156,20 zł	84 832,30 zł	62 402,15 zł	25 328,67 zł	91 096,50 zł
Szczecin	37 130,31 zł	13 449,90 zł	90 394,53 zł	93 272,45 zł	97 202,66 zł	15 376,21 zł	53 082,48 zł
Gdańsk	26 229,71 zł	82 070,17 zł	33 971,94 zł	44 662,52 zł	35 448,55 zł	76 730,24 zł	30 859,55 zł

Kolumna1	Zielona Góra	Poznań	Wrocław	Warszawa	Szczecin	Gdańsk
2005	21 858,38 zł	45 387,68 zł	74 837,31 zł	85 771,88 zł	37 130,31 zł	26 229,71 zł
2006	21 939,12 zł	74 232,79 zł	85 346,91 zł	18 860,00 zł	13 449,90 zł	82 070,17 zł
2007	47 952,72 zł	6 146,71 zł	38 194,06 zł	27 156,20 zł	90 394,53 zł	33 971,94 zł
2008	53 986,24 zł	98 505,45 zł	42 979,83 zł	84 832,30 zł	93 272,45 zł	44 662,52 zł
2009	47 689,27 zł	49 436,76 zł	78 078,34 zł	62 402,15 zł	97 202,66 zł	35 448,55 zł
2010	74 610,71 zł	90 574,28 zł	81 408,35 zł	25 328,67 zł	15 376,21 zł	76 730,24 zł
2011	56 919,40 zł	14 304,67 zł	22 206,96 zł	91 096,50 zł	53 082,48 zł	30 859,55 zł

Operacje wielowymiarowe

- Wycinanie
  - wybór fragmentu danych poprzez wybór tylko niektórych wartości wymiarów
  - też zmniejszenie liczby wymiarów

Kolumna1	2007	2010
Zielona Góra	47 952,72 zł	74 610,71 zł
Poznań	6 146,71 zł	90 574,28 zł
Wrocław	38 194,06 zł	81 408,35 zł
Warszawa	27 156,20 zł	25 328,67 zł
Szczecin	90 394,53 zł	15 376,21 zł
Gdańsk	33 971,94 zł	76 730,24 zł

Kolumna1	2005	2006	2007
Zielona Góra	21 858,38 zł	21 939,12 zł	47 952,72 zł
Poznań	45 387,68 zł	74 232,79 zł	6 146,71 zł
Wrocław	74 837,31 zł	85 346,91 zł	38 194,06 zł
Warszawa	85 771,88 zł	18 860,00 zł	27 156,20 zł

Operacje wielowymiarowe

- Filtrowanie
  - usuwamy dane, które w danym momencie nie są dla nas interesujące

Kolumna1	2005	2006	2007	2008	2009	2010	2011
Zielona Góra	21 858,38 zł	21 939,12 zł	47 952,72 zł	53 986,24 zł	47 689,27 zł	74 610,71 zł	56 919,40 zł
Poznań	45 387,68 zł	74 232,79 zł	6 146,71 zł	98 505,45 zł	49 436,76 zł	90 574,28 zł	14 304,67 zł
Wrocław	74 837,31 zł	85 346,91 zł	38 194,06 zł	42 979,83 zł	78 078,34 zł	81 408,35 zł	22 206,96 zł
Warszawa	85 771,88 zł	18 860,00 zł	27 156,20 zł	84 832,30 zł	62 402,15 zł	25 328,67 zł	91 096,50 zł
Szczecin	37 130,31 zł	13 449,90 zł	90 394,53 zł	93 272,45 zł	97 202,66 zł	15 376,21 zł	53 082,48 zł
Gdańsk	26 229,71 zł	82 070,17 zł	33 971,94 zł	44 662,52 zł	35 448,55 zł	76 730,24 zł	30 859,55 zł

przychód <	50 000,00 zł

Operacje wielowymiarowe

- Wyznaczanie rankingu
  - „kto ma największe przychody?”

Kolumna1	2005	2006	2007	2008	2009	2010	2011
Zielona Góra	21 858,38 zł	21 939,12 zł	47 952,72 zł	53 986,24 zł	47 689,27 zł	74 610,71 zł	56 919,40 zł
Poznań	45 387,68 zł	74 232,79 zł	6 146,71 zł	98 505,45 zł	49 436,76 zł	90 574,28 zł	14 304,67 zł
Wrocław	74 837,31 zł	85 346,91 zł	38 194,06 zł	42 979,83 zł	78 078,34 zł	81 408,35 zł	22 206,96 zł
Warszawa	85 771,88 zł	18 860,00 zł	27 156,20 zł	84 832,30 zł	62 402,15 zł	25 328,67 zł	91 096,50 zł
Szczecin	37 130,31 zł	13 449,90 zł	90 394,53 zł	93 272,45 zł	97 202,66 zł	15 376,21 zł	53 082,48 zł
Gdańsk	26 229,71 zł	82 070,17 zł	33 971,94 zł	44 662,52 zł	35 448,55 zł	76 730,24 zł	30 859,55 zł

Tabele przestawne (Pivot Table)

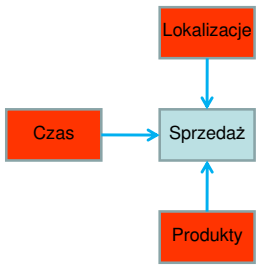
Region	Miasto	Data wysyłki	Jednostka	Cena
Zachód	ZG	01-01-2011	kg	12
Zachód	ZG	01-06-2011	kg	55
Zachód	POZ	12-05-2011	litr	77
Zachód	POZ	11-11-2011	litr	98
Zachód	SZCZ	23-08-2011	litr	44
Zachód	SZCZ	23-08-2011	litr	22
Wschód	WAW	23-08-2011	litr	45
Wschód	WAW	23-08-2011	litr	95
Wschód	LUB	11-11-2011	szt	46
Południe	KRA	01-01-2011	szt	111
Południe	KRA	01-06-2011	szt	564
Południe	KAT	11-12-2011	szt	333
Południe	KAT	23-08-2011	szt	87
Południe	KAT	01-05-2011	szt	5
Północ	SZCZ	06-10-2011	szt	
Północ	SZCZ	11-11-2011	szt	
Północ	SZCZ	23-08-2011	m3	
Północ	GDA	23-08-2011	m3	
Północ	GDA	23-08-2011	m3	

Suma z Cena	Etykiety kolumn	
Etykiety wierszy	szt	Suma końcowa
Południe	1100	1100
KAT	425	425
KRA	675	675
Północ	80	80
SZCZ	80	80
Suma końcowa	1180	1180

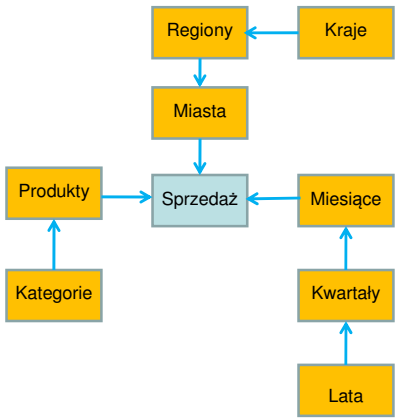
Suma z Cena	Etykiety kolumn				
Etykiety wierszy	kg	litr	m3	szt	Suma końcowa
Południe				1100	1100
KAT				425	425
KRA				675	675
Północ			226	80	306
GDA			163		163
SZCZ			63	80	143
Wschód		140		46	186
LUB				46	46
WAW		140			140
Zachód	67	241			308
POZ		175			175
SZCZ		66			66
ZG	67				67
Suma końcowa	67	381	226	1226	1900

Typowe schematy logiczne

- Gwiazda

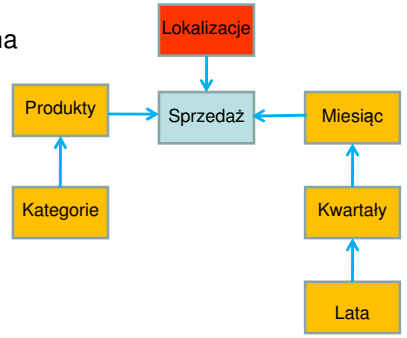


Płatek śniegu



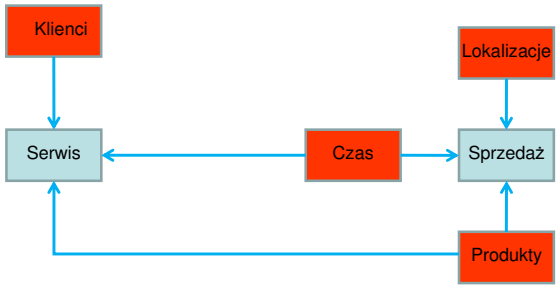
Typowe schematy logiczne

- Schematy mieszane (gwiazda – płatek śniegu)
  - część wymiarów zdenormalizowana
    - potrzeba więcej miejsca, jednak zapytania wykonują się bardziej efektywnie
  - część wymiarów znormalizowana
    - oszczędzamy miejsce, jednak zapytania wykonują się mniej efektywnie



Typowe schematy logiczne

- Konstelacja faktów



Schemat gwiazdy

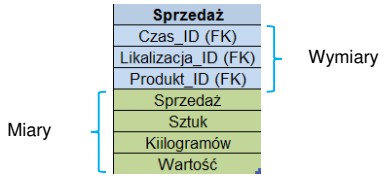
- Prosta struktura
- Niewielka ilość złączeń tabel …
  - … więc prostsze i szybciej działające zapytania
- Sporo danych dubluje się …
  - … ale zwykle wymiary nie przechowują, w stosunku do faktów, zbyt dużo danych
- Efektywność ładowania danych do wymiarów słaba, ze względu na konieczność dokonania denormalizacji …
  - … gdyż dane do HD trafiają zwykle z wysoce znormalizowanych tabel relacyjnych
- Mimo powyższych wad, tak struktura dominuje w HD i jest wspierana przez różne narzędzia

Schemat płatka śniegu

- Złożona struktura
- Dość duża ilość złączeń tabel …
  - … więc zapytania są bardziej złożone i są mniej efektywne (konieczność złączania wielu tabel)
- Dane w niewielkim stopniu dublują się
- Ładowanie danych do HD szybsze …
  - … gdyż struktura płatka śniegu jest bliższa strukturze baz źródłowych
- Struktura rzadziej stosowana we praktyce niż struktura gwiazdy …
  - … gdyż efektywność ładowania danych jest mniej ważna, niż efektywność wykonywanych zapytań

Tabele faktów

- Posiadające zarówno wymiary jak i miary
- Często tabela faktów nie posiada klucza głównego
- Tabele faktów bez miar
  - przechowuje pojedyncze zdarzenia, np., że w danym dniu, w danej lokalizacji, sprzedano dany produkt
  - wydaje się, że tabela faktów bez miar ma mniejsze zastosowanie w praktyce niż wersja z miarami



Tabele wymiarów

- Nadają znaczenie faktom i miarom
- Tabela z wymiarem czasu jest szczególnie ważna
  - pozwala dokonywać agregacji danych z różną ziarnistością (np. sprzedaż tygodniowa, kwartalna, roczna, tylko w piątki, tylko w dni świąteczne itd.)

Tabele z demonstracyjnego modelu SH w ORACLE

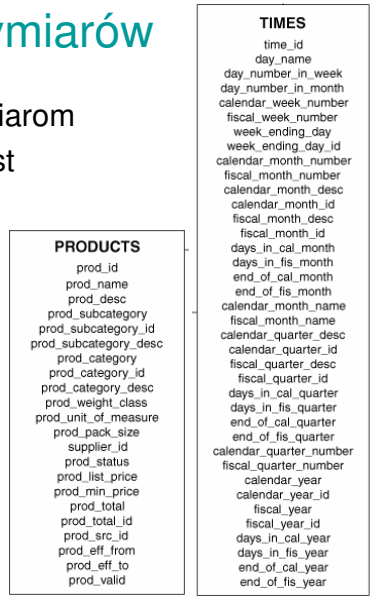
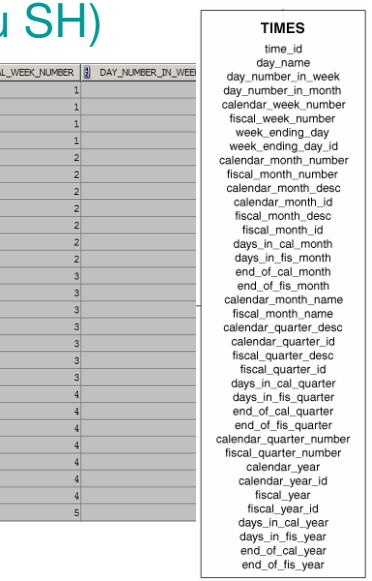


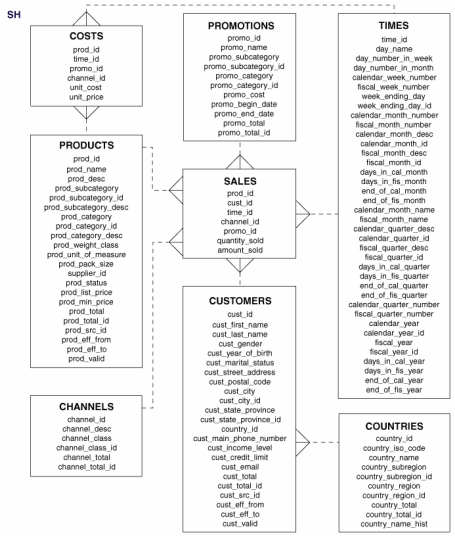
Tabela wymiaru dla czasu (z modelu SH)

	TIME_ID	DAY_NAME	WEEK_ENDING_DAY	FISCAL_MONTH_NAME	FISCAL_WEEK_NUMBER	DAY_NUMBER_IN_WEE
1	01-01-1998	Thursday	04-01-1998	January	1	
2	02-01-1998	Friday	04-01-1998	January	1	
3	03-01-1998	Saturday	04-01-1998	January	1	
4	04-01-1998	Sunday	04-01-1998	January	1	
5	05-01-1998	Monday	11-01-1998	January	2	
6	06-01-1998	Tuesday	11-01-1998	January	2	
7	07-01-1998	Wednesday	11-01-1998	January	2	
8	08-01-1998	Thursday	11-01-1998	January	2	
9	09-01-1998	Friday	11-01-1998	January	2	
10	10-01-1998	Saturday	11-01-1998	January	2	
11	11-01-1998	Sunday	11-01-1998	January	2	
12	12-01-1998	Monday	18-01-1998	January	3	
13	13-01-1998	Tuesday	18-01-1998	January	3	
14	14-01-1998	Wednesday	18-01-1998	January	3	
15	15-01-1998	Thursday	18-01-1998	January	3	
16	16-01-1998	Friday	18-01-1998	January	3	
17	17-01-1998	Saturday	18-01-1998	January	3	
18	18-01-1998	Sunday	18-01-1998	January	3	
19	19-01-1998	Monday	25-01-1998	January	4	
20	20-01-1998	Tuesday	25-01-1998	January	4	
21	21-01-1998	Wednesday	25-01-1998	January	4	
22	22-01-1998	Thursday	25-01-1998	January	4	
23	23-01-1998	Friday	25-01-1998	January	4	
24	24-01-1998	Saturday	25-01-1998	January	4	
25	25-01-1998	Sunday	25-01-1998	January	4	
26	26-01-1998	Monday	01-02-1998	February	5	

1826 rekordów

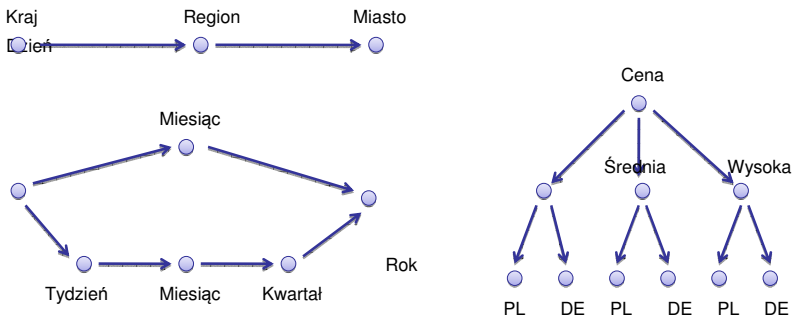


Schemat demonstracyjny SH (ORACLE)

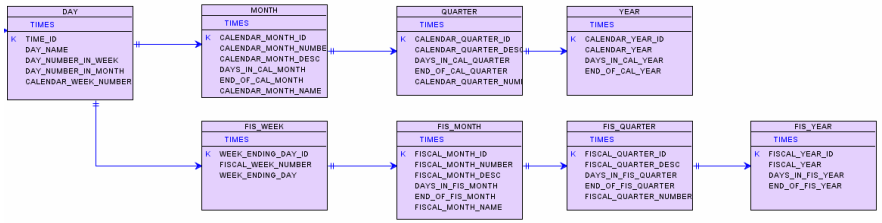


Hierarchie wymiarów

- Pozwalają dokonywać agregacji danych na różnych poziomach

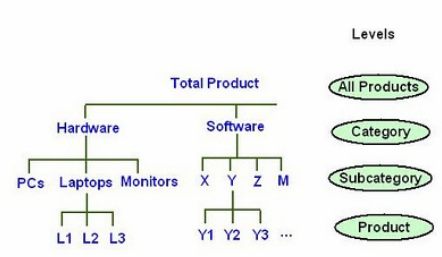


Hierarchie wymiarów, przykład



Poziomy wymiarów

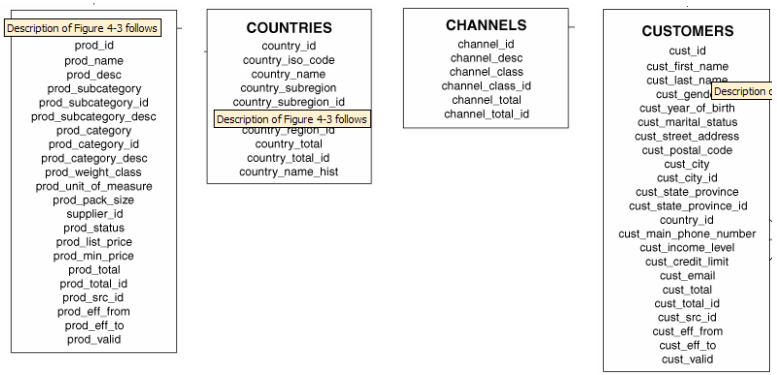
- Poziom reprezentuje pozycje w hierarchii wymiarów



Levels				Attributes	
Total	Category	Subcategory	Product	Manager	Color
All Products	Clothing	Women's	Red shirt	Mary	Red
		Men's	Green pants	John	Green
		Kids'	Red pants	John	Red
			Blue shirt	Mary	Blue
	Bedding	Sheets	Yellow sheets	Karl	Yellow
			Red sheets	Karl	Red
		Pillows	White pillows	Bruce	White

Atrybuty wymiarów

- Dostarczają dodatkowych informacji



Tabele z demonstracyjnego modelu SH w ORACLE

Klucze główne i klucze obce

- Tabele wymiarów
  - PK jako klucze naturalne (np. numer faktury, num. rej. pojazdu) lub
  - PK jako klucze sztuczne (ang. surrogate keys)
    - np. klient_id, produkt_id, transakcja_id
    - generowane automatycznie przez system (np. MySQL: klauzula AUTO INCREMENT, Oracle: użycie sekwencji)
    - klucz sztuczny jest zwykle lepszy niż naturalny (bo jest zwykle mniejszy/krótszy, brak konkretnej interpretacji itd)
  - FK tylko w układzie płatka śniegu oraz układzie mieszanym
- Tabele faktów
  - PK występują raczej rzadko, jeżeli już jest, to klucz główny składa się z kolumn będących kluczami obcymi do tabel wymiarów plus ew. dodatkowe kolumny
  - FK jako referencje do odpowiednich kolumn w tabelach wymiarów

Zmienność danych

- Zawartość tabeli faktów zmienia się bardzo znacznie (dodawane są nowe rekordy)
- Zawartości poszczególnych tabel wymiarów są względnie stabilne
  - nowe produkty pojawiają się stosunkowo rzadko
  - nowe sklepy otwierane są jeszcze rzadziej
  - nowe województwa? raz na dekadę lub rzadziej
- Stosunkowo często zmieniają się wartości w tabelach wymiarów
  - zmiana kategoryzacji produktu
  - zmiana cen produktów
  - zmiana adresów klientów
- Rozwiązanie
  - tzw. wymiary wolnozmienne (ang. Slowly Changing Dimensions, SCD)
  - typ 1, typ 2, typ 3

Wymiary wolnozmienne

- SCD, Typ 1
  - nadpisanie wartości
  - konsekwencja: utrata danych historycznych ze wszelkimi tego konsekwencjami

przed

prod_id	nazwa	cena
1	mleko	2,50

po

prod_id	nazwa	cena
1	mleko	3,20

- SCD, Typ 2
  - tworzenie nowych rekordów, stare pozostają
  - konsekwencja: w pewnym sensie dane dublują się

przed

prod_id	nazwa	cena
1	mleko	2,50

po

prod_id	nazwa	cena
1	mleko	2,50
1	mleko	3,20

- SCD, Typ 3
  - tworzymy w tabelach wymiarów nowe kolumny

przed

prod_id	nazwa	cena
1	mleko	2,50

po

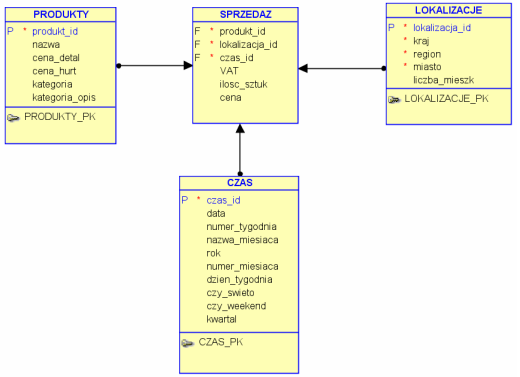
prod_id	nazwa	cena_przed	cena_po
1	mleko	2,50	3,20

Implementacja modelu wielowymiarowego

- ROLAP (Relational OLAP)
  - stworzone w relacyjnej bazie danych
  - mogą być bardzo wielkie (rzędu TB)
  - w typowych zadaniach realizowanych przez HD (agregowanie danych) mogą być niezbyt wydajne
- MOLAP (Multidimensional OLAP)
  - stworzone z wykorzystaniem specjalnych serwerów wielowymiarowych
  - są zwykle mniejsze niż ROLAP (rzędu GB)
  - wydajne operacje typowe dla HD
- HOLAP (Hybrid OLAP)
  - połączenie dwóch powyższych

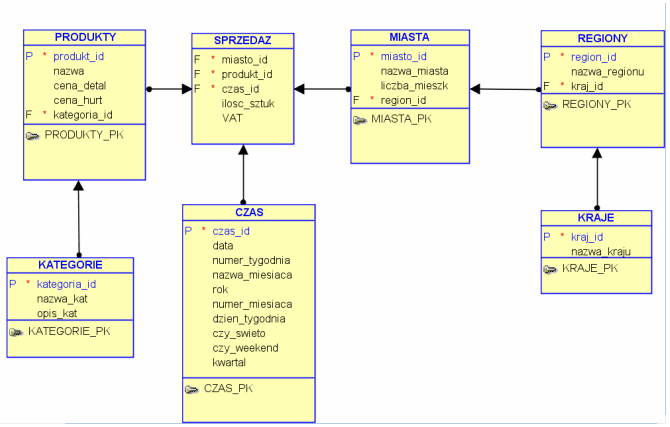
Typowe schematy logiczne

- Gwiazda, implementacja w modelu relacyjnym

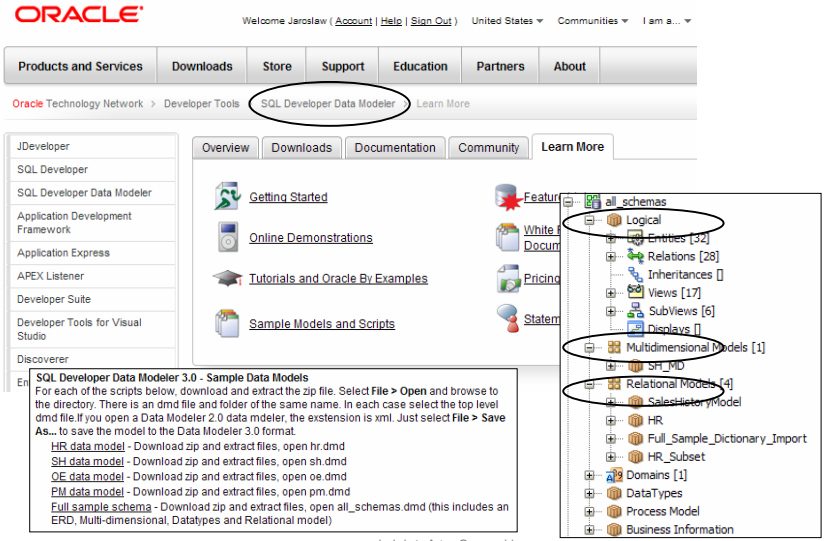


Typowe schematy logiczne

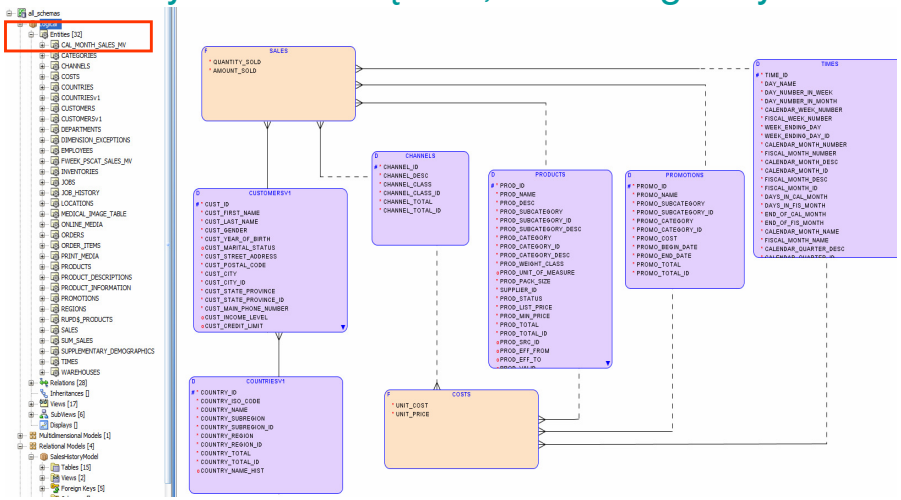
- Gwiazda – płatek śniegu, implementacja w modelu relacyjnym



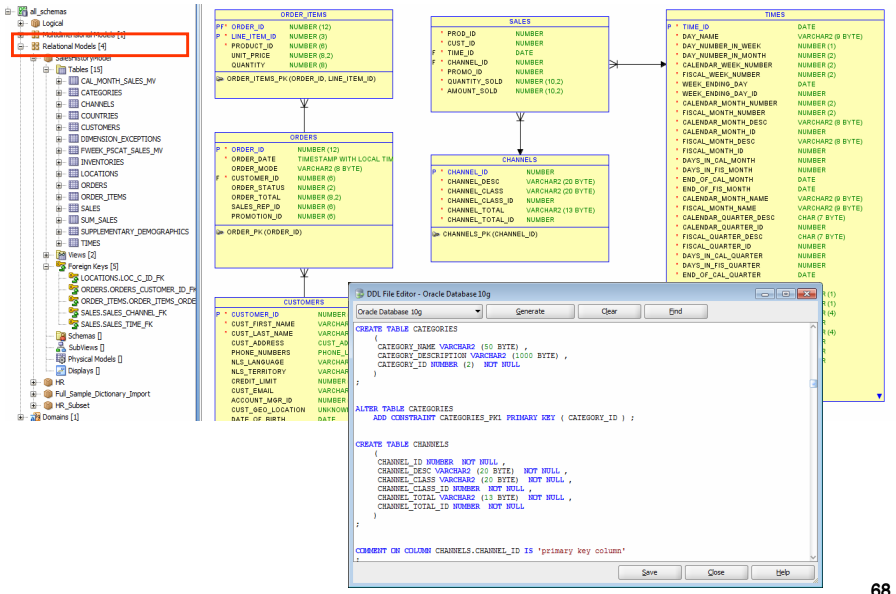
Wybrane narzędzia do modelowania



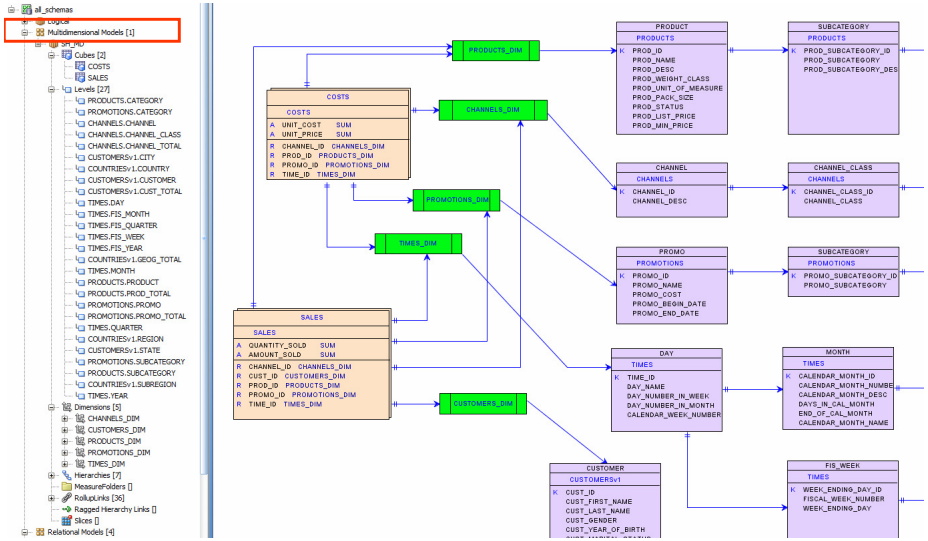
Wybrane narzędzia, model logiczny



Wybrane narzędzia, model relacyjny



Wybrane narzędzia, model wielowymiarowy



Wybrane narzędzia

- Wiodący produkt Baza danych i jej rozszerzenia
  - Data Options