



Hurtownie danych - przegląd technologii

Robert Wrembel
Politechnika Poznańska
Instytut Informatyki
Robert.Wrembel@cs.put.poznan.pl
www.cs.put.poznan.pl/rwrembel



Modelowanie hurtowni danych

- ⇒ Model wielowymiarowy
- ⇒ Implementacja ROLAP
- ⇒ Implementacja MOLAP
- ⇒ Modelowanie ROLAP
- ⇒ Zarządzanie modyfikacjami danych
 - Slowly Changing Dimensions



Dane przechowywane w HD

- ⇒ Dane bieżące
- ⇒ Dane historyczne



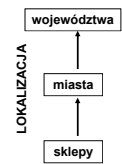
Kategorie analizowanych danych

⇒ Fakty

- dane podlegające analizie
 - sprzedaż, rozmowy telefoniczne
 - charakteryzowane ilościowo za pomocą **miar**
 - **liczba** sprzedanych sztuk towaru, **czas** trwania rozmowy

⇒ Wymiary

- ustalają kontekst analizy
 - sprzedaż czekolady (**produkt**) w Auchan (**sklep**) w poszczególnych miesiącach roku (**czas**)
 - składają się z **poziomów**, które tworzą hierarchię



ROLAP, MOLAP, HOLAP

- ⇒ Model relacyjny (ROLAP)
 - schemat gwiazdy (ang. star schema)
 - schemat płatka śniegu (ang. snowflake schema)
 - schemat konstelacji faktów (ang. fact constellation schema)
 - schemat gwiazda-płatek śniegu (ang. starflake schema)
- ⇒ Model wielowymiarowy (MOLAP, MDOLAP)
- ⇒ Model hybrydowy (HOLAP)

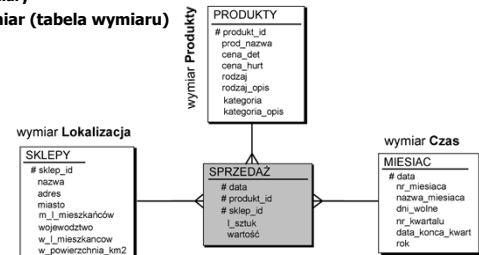


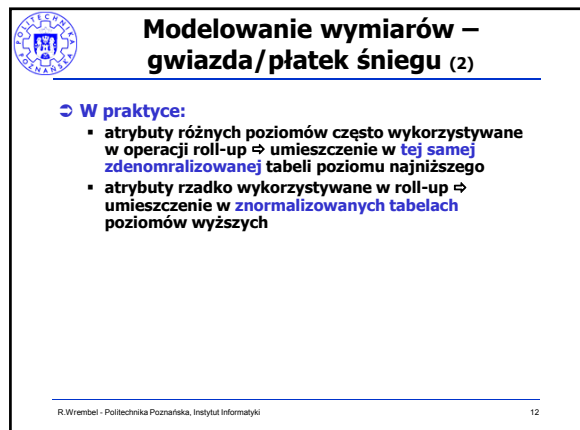
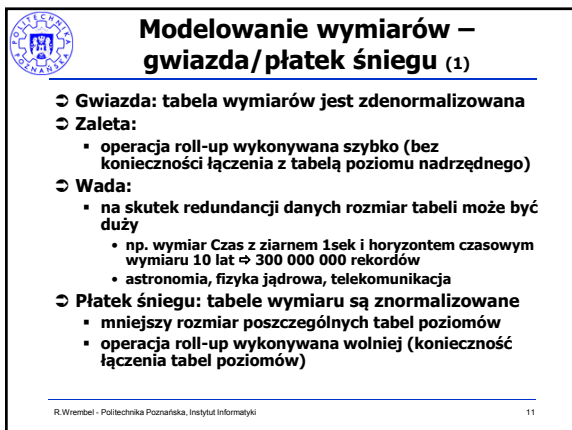
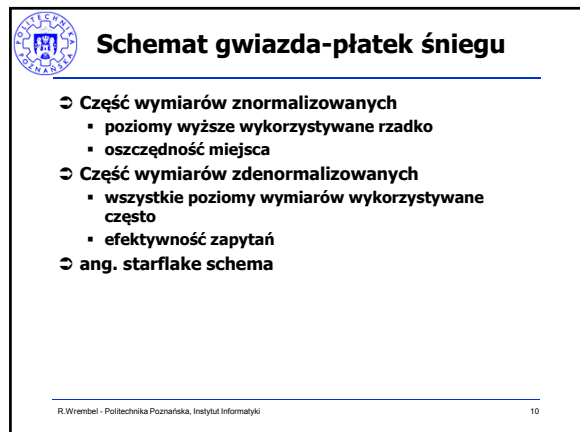
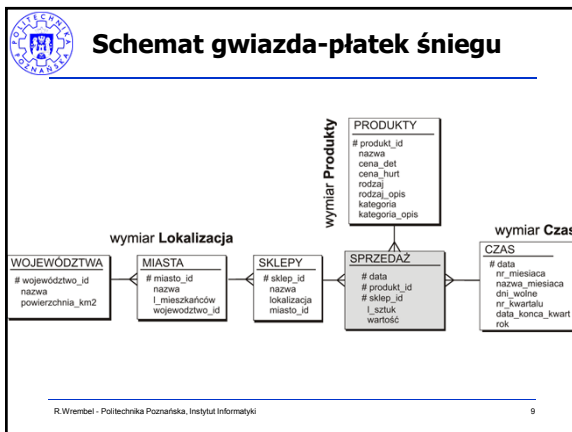
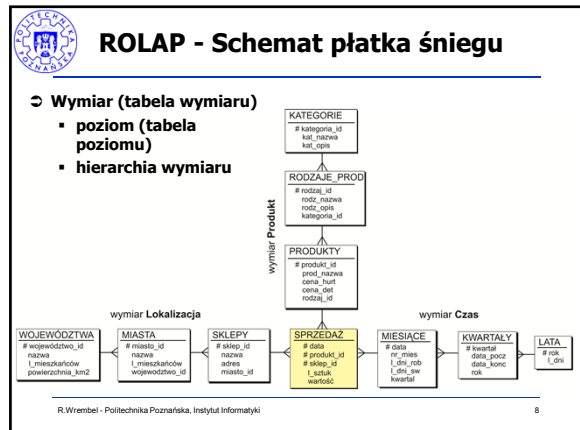
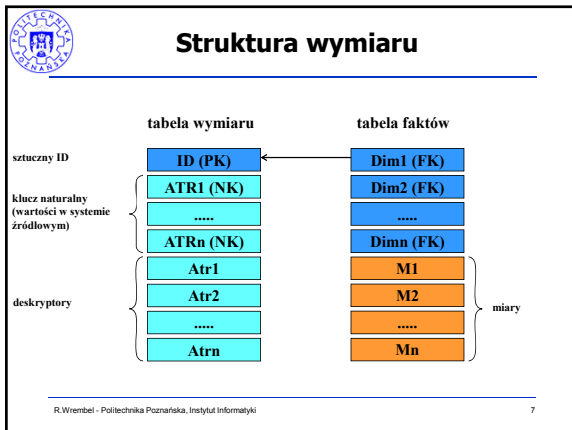
ROLAP - Schemat gwiazdy

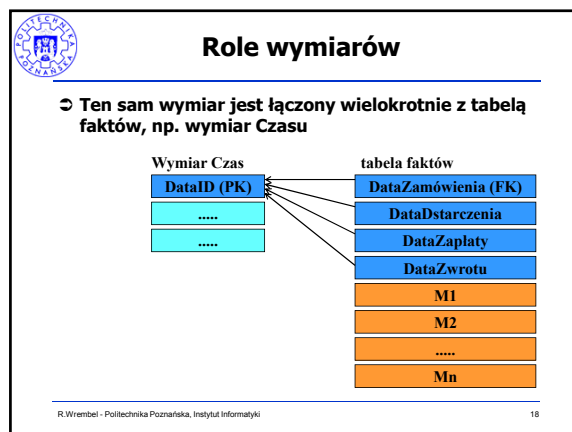
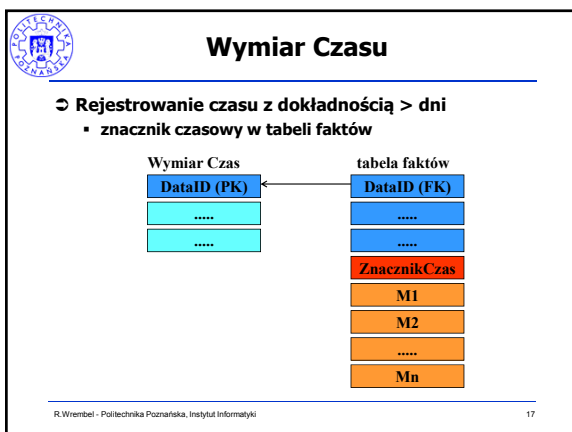
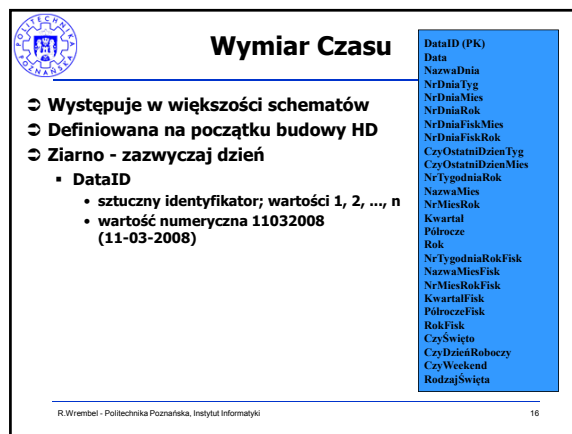
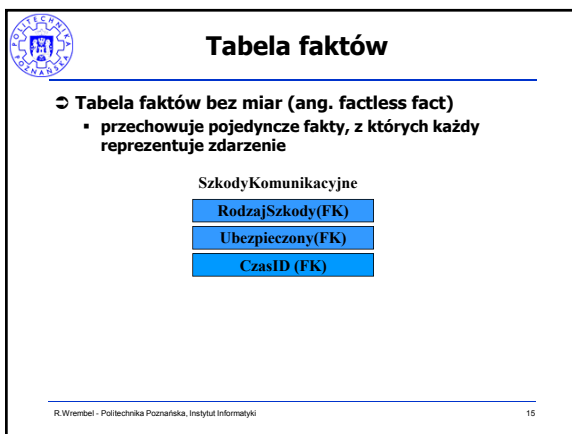
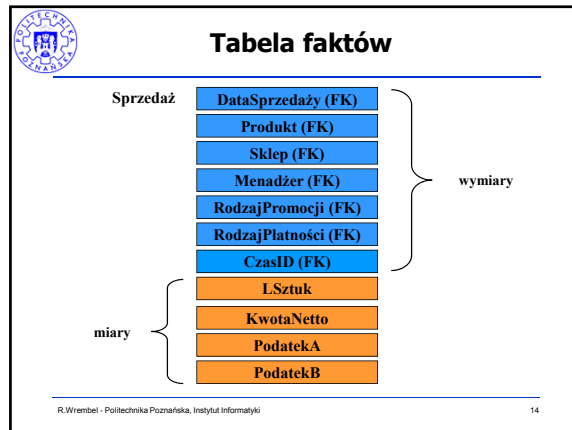
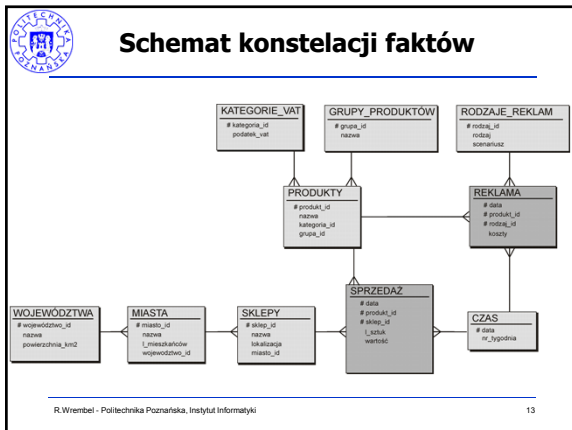
⇒ Fakty (tabela faktów)

- miary

⇒ Wymiar (tabela wymiaru)







Role wymiarów

⇒ Wymiar w tabeli faktów (fact dimension)

R.Wrembel - Politechnika Poznańska, Instytut Informatyki 19

Generowanie sztucznych ID

⇒ Baza danych

- wyzwalacz - pogarsza efektywność HD i ETL
- sekwencja - rozwiązanie akceptowalne

⇒ Sposób generowania

- wartości numeryczne - zadowalająca efektywność
- wartości znakowe
 - konkatenacja klucza naturalnego i znacznika czasowego (data i czas utworzenia w systemie źródłowym lub wczytania do HD)
 - niska efektywność
 - duży rozmiar wartości (kilkanaście B)

R.Wrembel - Politechnika Poznańska, Instytut Informatyki 20

Sztuczne ID

⇒ Numeryczne

- brak semantyki
- chronologia wstawiania wartości reprezentowana kolejnymi wartościami
- odizolowanie wartości w HD od zmian wartości danych w źródłach
- dobrze wspierają integrację danych z wielu źródeł
- dobrze wspierają historię zmian instancji wymiaru

R.Wrembel - Politechnika Poznańska, Instytut Informatyki 21

Agregacja (1)

⇒ Ścieżka agregacji

R.Wrembel - Politechnika Poznańska, Instytut Informatyki 22

Agregacja (2)

⇒ Agregowalność (summarizability)

- możliwość agregowania wartości na poziomie wyższym (np. Miasto) w oparciu o agregaty z poziomu niższego (np. Sklep)

⇒ Kryteria poprawnej agregowalności

- rozsączność zbioru instancji
 - związek 1:M pomiędzy poziomem nadrzędnym i podrzędnym
 - instancja poziomu podrzędnego jest powiązana tylko z jedną instancją poziomu nadrzędnego
- kompletność
 - wszystkie instancje poziomów należą do instancji wymiaru
 - każda instancja poziomu podrzędnego jest powiązana z instancją poziomu nadrzędnego
- właściwa funkcja agregująca
 - zastosowanie właściwej funkcji agregującej dla danego typu miary

R.Wrembel - Politechnika Poznańska, Instytut Informatyki 23

Agregacja (3)

⇒ Typy miar (kryterium agregowalności)

- addytywne (additive, flow, rate) ⇒ zapewniają poprawną agregowalność dla **wszystkich** wymiarów
 - np. liczba sprzedanych sztuk towaru ⇒ możliwe agregowanie w wymiarze czasu, klienta, produktu, sklepu, dostawcy, ...
- częściowo addytywne (semiadditive, stock, level) ⇒ zapewniają poprawną agregowalność dla **niektórych** wymiarów
 - np. liczba sztuk towaru w magazynie ⇒ możliwe agregowanie w wymiarze magazynu (Magazyn → Miasto → Region), agregowanie w wymiarze czasu daje nieinterpretowalne wyniki
- nieaddytywne (nonadditive, value-per-unit) ⇒ nie podają interpretowalnych wartości zagregowanych w **żadnym** z wymiarów
 - np. cena netto, kurs wymiany, kurs akcji dla SUM

R.Wrembel - Politechnika Poznańska, Instytut Informatyki 24

Agregacja (4)

⇒ Typy miar (kryterium własności funkcji agregującej)

- **dystrybutywne (distributive)**
 - niezależne wyniki działania funkcji **F** na **n** podzbiórach zbioru danych **Z** są agregowane do jednego wyniku, który jest identyczny z wynikiem działania **F** na całym zbiorze **Z**
 - agregat na poziomie wyższym może być obliczony na podstawie agregatów na poziomie niższym
 - count, min, max, sum
- **algebraiczne (algebraic)**
 - obliczana na podstawie wyników działania funkcji dystrybutywnych
 - avg, stdev
- **holistyczne (holistic)**
 - do obliczenia wyników potrzebne są wszystkie dane elementarne
 - median, rank

R.Wrembel - Politechnika Poznańska, Instytut Informatyki 25

Hierarchie wymiarów

⇒ Malinowski E., Zimanyi E.: *Advanced Data Warehouse Design. From Conventional to Spatial and Temporal Applications*. Springer Verlag, 2008

R.Wrembel - Politechnika Poznańska, Instytut Informatyki 26

Hier. zrównoważona

⇒ Posiada jedną ścieżkę w schemacie

⇒ Związek 1:M obustronnie obowiązkowy pomiędzy poziomem nadrzędnym, a podrzędnym

R.Wrembel - Politechnika Poznańska, Instytut Informatyki 27

Hier. niezrównoważona (1)

⇒ Posiada jedną ścieżkę w schemacie

⇒ Związek 1:M pomiędzy poziomem nadrzędnym, a podrzędnym, opcjonalny od strony poziomu nadrzędnego

R.Wrembel - Politechnika Poznańska, Instytut Informatyki 28

Hier. niezrównoważona (2)

⇒ Szczególny przypadek ⇒ hierarchia niezrównoważona rekursywna

R.Wrembel - Politechnika Poznańska, Instytut Informatyki 29

Hier. z generalizacją

⇒ Poziomy wymiaru są powiązane związkami generalizacji (nadtyp-podtyp)

⇒ Podtypy mogą posiadać własne hierarchie

⇒ Na poziomie schematu istnieje wiele ścieżek współdzielących przynajmniej poziom najniższy

⇒ Część poziomów może być specyficzna dla ścieżki

⇒ Część poziomów może być wspólna dla kilku ścieżek

⇒ Każda instancja poziomu należy tylko do **jednej** ścieżki

R.Wrembel - Politechnika Poznańska, Instytut Informatyki 30

Hier. z generalizacją

Promocja
 #IDPromocji
 * Forma
 * Budżet
 * Koordynator
 ...

Stacyjna
 * Nośnik
 * Lokalizacja
 * Koszt
 ...

Dynamiczna
 * Rodzaj
 * Czas_trwania
 * Częstotliwość
 * Kanal
 ...

R.Wrembel - Politechnika Poznańska, Instytut Informatyki 31

Hier. z generalizacją

Promocja
 #IDPromocji
 * Budżet
 ...

Stacyjna
 * Nośnik
 * Lokalizacja
 * Koszt
 ...

Dynamiczna
 * Rodzaj
 * Czas_trwania
 * Częstotliwość
 * Kanal
 ...

Forma
 #IDFormy
 * Nazwa
 * Opis
 * Koordynator
 ...

- Specjalizacja częściowa \Leftrightarrow hierarchia niezrównoważona
- Specjalizacja nierozłączna \Leftrightarrow problem z obliczeniem agregatów (agregowanie w razy tych samych wartości)
 - specjalizowana procedura agregująca
 - zlikwidowanie specjalizacji nierozłącznej w hierarchii wymiaru

R.Wrembel - Politechnika Poznańska, Instytut Informatyki 32

Hier. ścisła

\Rightarrow Na poziomie schematu wymiaru istnieją wyłącznie związki 1:M pomiędzy poziomem wyższym (1), a niższym (M)

R.Wrembel - Politechnika Poznańska, Instytut Informatyki 33

Hier. nieściśla (1)

\Rightarrow Na poziomie schematu wymiaru istnieją przynajmniej jeden związek M:N pomiędzy poziomami

\Rightarrow Problem agregowania w razy tego samego produktu

- sprzedaż elektroniki \Leftrightarrow iPhone liczony 3 razy

R.Wrembel - Politechnika Poznańska, Instytut Informatyki 34

Hier. nieściśla (2)

\Rightarrow Rozwiązania

- wprowadzenie nowej kategorii, np. smartphone: {iPhone}
- wybór jednej kategorii (głównej) do której należy produkt w ścieżce agregacji
- dystrybucja wartości miar dla produktu pomiędzy kategorie
 - przyjęta oddzielnie, np. telefon 40%, odtwarzacz MP4 20%, PDA 40%
 - równomierna: wartość miary podzielona przez liczbę kategorii do których należy produkt, np. 1/3 dla iPhone
- transformacja nieściśłej hierarchii do ścisłej

R.Wrembel - Politechnika Poznańska, Instytut Informatyki 35

Hier. nieściśla (3)

\Rightarrow Transformacja nieściśła \Leftrightarrow ścisła

- możliwa jeśli znamy dokładną dystrybucję wartości miary, np. ile pracownik zarabia w konkretnym projekcie

R.Wrembel - Politechnika Poznańska, Instytut Informatyki 36

Hier. nieściła (4)

↪ **Problem**

- zliczenie pracowników jednostek

Projekt	Pracownik	Wynagrodzenie
P1	Prac1	80
P1	Prac2	50
P2	Prac1	80
P2	Prac2	30
P2	Prac3	40
P3	Prac4	60
P3	Prac5	20
P3	Prac2	20
P4	Prac1	30
P4	Prac3	50
P4	Prac5	40

J1 (3) J2 (5)

R.Wrembel - Politechnika Poznańska, Instytut Informatyki 37

Hier. alternatywna (1)

↪ Złożona z hierarchii prostych współdzielących przynajmniej poziom najniższy
 ↪ Ścieżki analizy nie są wyłączne
 ↪ Agregacja każdą ścieżką dla poziomów współdzielonych daje identyczne wyniki, np. sprzedaż roczna, miesięczna

R.Wrembel - Politechnika Poznańska, Instytut Informatyki 38

Hier. alternatywna (2)

↪ Instancja wymiaru jest grafem ⇔ instancja podrzędna jest powiązana z **kilkoma** instancjami nadrzędnymi, a każda z instancji nadrzędnych należy do innego poziomu
 ↪ Każda ścieżka agregacji prowadzi do tych samych instancji poziomów najniższego i najwyższego

R.Wrembel - Politechnika Poznańska, Instytut Informatyki 39

Hier. alternatywna (3)

↪ **Problem**

- jednoczesna analiza danych wzdłuż obu ścieżek agregacji może prowadzić do źle interpretowanych wyników
- np. sprzedaż w 1 kwartale 2009 i 2 półroczu 2009 ⇒ brak danych
- rozwiązanie: wybór jednej ścieżki agregacji w danej analizie

R.Wrembel - Politechnika Poznańska, Instytut Informatyki 40

Hier. równoległe (1)

↪ Niezależne (parallel independent hierarchies) ⇔ przypadek standardowy

- brak współdzielenia poziomów pomiędzy hierarchiami
- każda hierarchia stanowi inne kryterium analizy, np. Produkt, Klient, Czas

R.Wrembel - Politechnika Poznańska, Instytut Informatyki 41

Hier. równoległe (2)

↪ Zależne (parallel dependent hierarchies)

- współdzielenie poziomów pomiędzy hierarchiami
- każda hierarchia stanowi inne kryterium analizy

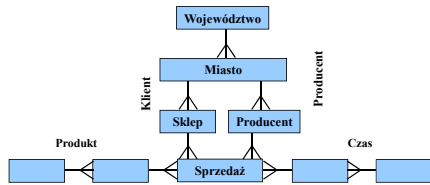
↪ analiza danych w kontekście obu hierarchii, np. wskaźniki sprzedaży w poszczególnych miastach danego producenta

R.Wrembel - Politechnika Poznańska, Instytut Informatyki 42



Hier. równoległe (3)

- Agregacja danych na poziomach wspólnych, np. miasto daje różne wyniki dla różnych ścieżek agregacji



Praktyka

- ⇒ W praktyce maksymalna liczba wymiarów to 15-18

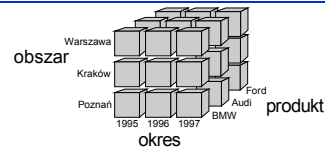


Narzędzia do modelowania

- ⇒ DB2 Data Warehouse Center,
- ⇒ Sybase Warehouse Studio,
- ⇒ Microsoft Data Warehousing Framework,
- ⇒ SAP Data warehouse management,
- ⇒ NCR Teradata Warehouse Builder,
- ⇒ Oracle Designer6i/9i



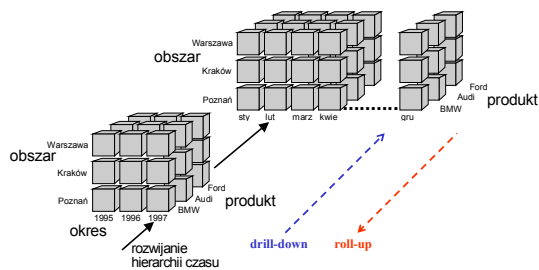
Model wielowymiarowy



- ⇒ "Wielowymiarowa kostka" (data cube, hypercube)
- ⇒ Operacje
 - drill-down / roll-up
 - slice, dice
 - rotate (pivot)
 - drill-across
 - drill-through

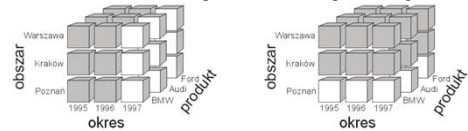


Drill-down / roll-up

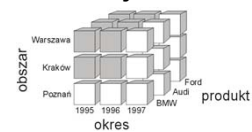


Slice, dice

- ⇒ Slice ⇒ warunki selekcji nałożone na jeden wymiar



- ⇒ Dice ⇒ warunki selekcji nałożone n wymiarów





Implementacja MOLAP

- ↻ Tablica wielowymiarowa
- ↻ Tablica haszowa (SQL Server)
- ↻ BLOB (Oracle)
- ↻ Quad tree
- ↻ K-D tree



HOLAP

- ↻ Dane elementarne i "słabo zagregowane" → ROLAP
- ↻ Dane zagregowane (tematyczne HD) → MOLAP

Modelowanie ROLAP



Problematyka

- ↻ Zidentyfikowanie faktów
- ↻ Zidentyfikowanie kluczowych wymiarów
- ↻ Zaprojektowanie tabel faktów
- ↻ Zaprojektowanie tabel wymiarów



ROLAP – zidentyfikowanie faktów

- ↻ Zidentyfikowanie kluczowych typów transakcji w systemie produkcyjnym (realizują kluczowe akcje/operacje w obszarze działania przedsiębiorstwa)
 - **handel**: transakcje sprzedaży
 - **bankowość**: kursy walut, operacje na rachunkach
 - **giełda**: wahania kursów akcji, operacje giełdowe
 - **ubezpieczenia**: wykupienie polisy, zmiana warunków polisy, zgłoszenie szkody, wypłacenie odszkodowania
 - **telekomunikacja**: zrealizowanie rozmowy przez abonenta, podłączenie telefonu, zawarcie umowy, zmiana abonamentu, płatności za abonament
 - **opieka zdrowotna**: przyjęcie pacjenta do szpitala, forma leczenia, wynik leczenia



ROLAP – zidentyfikowanie wymiarów

- ↻ Zidentyfikowanie kluczowych wymiarów dla faktów (określenie kontekstu analizy faktów)
 - **handel**:
 - analiza **sprzedaży** w poszczególnych **miastach** i **okresach** czasowych
 - **bankowość**:
 - **wahania** kursów **walut** w poszczególnych **dniach**
 - analiza przyrostu **liczby** nowych rachunków w poszczególnych **miesiącach** z podziałem na **rodzaje rachunków**
 - **giełda**:
 - **wahania** kursów akcji poszczególnych **firm** w poszczególnych **dniach**
 - **liczba** zawartych **transakcji** kupna lub sprzedaży w jednostce **czasu** i łączne **kwoty** tych operacji
 - **ubezpieczenia**:
 - analiza przyrostu/spadku **liczby** polis poszczególnych **rodzajów** w **miastach** w poszczególnych **miesiącach**
 - **telekomunikacja**:
 - analiza rozkładu **czasu rozmów** poszczególnych **klientów** w czasie **doby**

ROLAP – projektowanie schematu tabeli faktów (1)

- ⇒ Poziom szczegółowości informacji ⇒ rozmiar tabeli faktów
 - rejestrowanie kwoty zakupu pojedynczego produktu
 - rejestrowanie sumarycznej kwoty zakupu całego koszka
 - rejestrowanie sumarycznej kwoty zakupu w miesiącu
- ⇒ Horyzont czasowy danych
 - jak długo przechowywać informacje na najwyższym poziomie szczegółowości?
 - opracowanie strategii agregowania danych starszych
 - raporty roczne ⇒ najczęściej wystarczają agregaty sumujące fakty z dokładnością do miesiąca
 - raporty agregujące dane sprzed kilku lat ⇒ najczęściej wystarczają agregaty sumujące fakty z dokładnością do miesiąca, kwartału, lub roku

R.Wrembel - Politechnika Poznańska, Instytut Informatyki 61

ROLAP – projektowanie schematu tabeli faktów (2)

- ⇒ Właściwy zbiór atrybutów tabeli faktów
- ⇒ Usunięcie zbędnych atrybutów ⇒ rozmiar tabeli faktów
 - czy atrybut wnosi nową/niezbędną wiedzę o fakcie?
 - czy wartość atrybutu można wyliczyć?
- ⇒ Minimalizacja rozmiarów atrybutów
 - przykład: telekomunikacja
 - tabela wymiaru Abonenci zawiera $8 \cdot 10^6$ abonentów
 - każdy abonent dzwoni średnio 2 razy dziennie
 - roczny horyzont czasowy tabeli faktów
 - zmniejszenie długości rekordów tabeli faktów o 10B ⇒ zyskujemy 54GB

R.Wrembel - Politechnika Poznańska, Instytut Informatyki 62

ROLAP – klucze podstawowe i obce

- ⇒ Klucze naturalne
 - nr rejestracyjny pojazdu, VIN, nr rachunku, NIP, PESEL
- ⇒ Klucze sztuczne – generowane automatycznie przez system
 - nr klienta, id produktu, nr transakcji
- ⇒ Połączenie tabeli wymiaru i faktów za pomocą klucza podstawowego-obcego
 - pokaż liczbę szkód pojazdu o numerze rejestracyjnym xxx w ostatnim roku ⇒ zapytanie wyłącznie do tabeli faktów
- ⇒ Jeśli wartość klucza podstawowego może się zmienić ⇒ wysoki koszt uaktualnienia faktów
 - sytuacja mało prawdopodobna
 - uwaga: nr rachunku!

R.Wrembel - Politechnika Poznańska, Instytut Informatyki 63

Reprezentowanie czasu w tabeli faktów (1)

- ⇒ Sztuczny identyfikator (data_id)
 - konieczność łączenia z tabelą wymiaru czasu
- ⇒ Naturalny identyfikator (data, timestamp) – składowanie fizycznej daty
 - sposób bardziej efektywny
 - większość analiz wykonuje się w wymiarze czasu
 - zapytanie nie zawiera połączenia z tabelą wymiaru czasu
- ⇒ Składowanie przesunięcia czasowego
- ⇒ Składowanie zakresów dat

R.Wrembel - Politechnika Poznańska, Instytut Informatyki 64

Reprezentowanie czasu w tabeli faktów (2)

- ⇒ Składowanie przesunięcia czasowego
 - partycjonowane tabele faktów

Platnosci_styczen2004			Platnosci_luty2004		
klient_id	kwota	nr_dnia	klient_id	kwota	nr_dnia
100	57.60	1	100	57.60	1
203	123.90	1	203	123.90	2
4005	99.00	2	4005	99.00	3
205	79.40	3	205	79.40	4
111	205.90	3	111	205.90	5
5008	432.00	13	5008	432.00	28
23	332.40	14	23	332.40	29
567	87.00	31	567	87.00	29

R.Wrembel - Politechnika Poznańska, Instytut Informatyki 65

Reprezentowanie czasu w tabeli faktów (3)

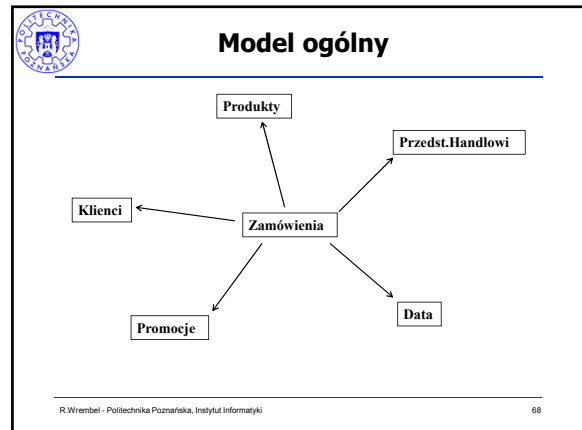
- ⇒ Składowanie przesunięcia czasowego
- ⇒ Wada:
 - konieczność konwersji daty z zapytania użytkownika do postaci przesunięcia czasowego
 - perspektywa
- ⇒ Zaleta:
 - podział dużej tabeli na mniejsze, z których każda może być adresowana w zapytaniu niezależnie
 - mniejszy rozmiar atrybutu reprezentującego datę (1B)

R.Wrembel - Politechnika Poznańska, Instytut Informatyki 66

Reprezentowanie czasu w tabeli faktów (4)

- Składowanie zakresów dat
 - np. atrybuty: data_od, data_do
 - stan magazynu supermarketu
 - sprzedaż w okresie
- Wada: bardziej złożone zapytanie testujące warunki początku i końca okresu
- Zaleta: rozszerzenie zakresu ważności rekordu poprzez zmodyfikowanie wartości data_do
 - np. liczba produktów w magazynie nie ulega zmianie w danym dniu ⇒ zmodyfikuj wartość data_do dla tego produktu

R.Wrembel - Politechnika Poznańska, Instytut Informatyki 67



Wymiary wolnozmiennie - problem

- Nowe instancje wymiaru
 - np. nowe sklepy, produkty, taryfy
- Modyfikacje wartości atrybutów
 - np. zmiana ceny brutto produktu, zmiana widełek w taryfikatorze mandatów
- Zmiana struktury instancji wymiaru
 - np. zmiana klasyfikacji produktu do innej grupy
 - zmiana struktury organizacyjnej (województwa, WE PP)
- Konieczność uaktualniania instancji wymiarów
- Rozwiązanie
 - Ralph Kimball ⇒ **Slowly Changing Dimensions**, typ 1, 2, 3

R.Wrembel - Politechnika Poznańska, Instytut Informatyki 69

SCD Typ 1

- Nadpisanie wartości
 - przed zmianą

prodPK	prodID	nazwa	brutto
10	KAJ10	kajzerka	1,3
 - po zmianie

prodPK	prodID	nazwa	brutto
10	KAJ10	kajzerka	1,5
- Brak historii zmian ⇒ utrata danych historycznych
- Niemożliwość porównania stanu sprzed modyfikacji i aktualnego

R.Wrembel - Politechnika Poznańska, Instytut Informatyki 70

SCD Typ 2

- Nowy rekord
 - przed zmianą

prodPK	prodID	nazwa	brutto
10	KAJ10	kajzerka	1,3
 - po zmianie

prodPK	prodID	nazwa	brutto
10	KAJ10	kajzerka	1,3
91	KAJ10	kajzerka	1,5
- możliwość rozszerzenia o atrybut "ważne do" lub 2 atrybuty daty ważności

R.Wrembel - Politechnika Poznańska, Instytut Informatyki 71

SCD Typ 3

- Dodatkowy atrybut
 - przed zmianą

prodPK	prodID	nazwa	brutto
10	KAJ10	kajzerka	1,3
 - po zmianie

prodPK	prodID	nazwa	brutto_akt	brutto_poprz
10	KAJ10	kajzerka	1,5	1,3
- Pamiętana jest aktualna i poprzednia wartość
- Niemożliwość porównania dowolnego stanu sprzed modyfikacji i aktualnego

R.Wrembel - Politechnika Poznańska, Instytut Informatyki 72



Ćwiczenie - operator aukcji internetowych

- ⇒ Wymagane analizy
- ⇒ Liczba otwartych aukcji w danym okresie
- ⇒ Liczba zakończonych aukcji w danym okresie
- ⇒ Ranking sprzedawców/kupujących z punktu widzenia opinii
- ⇒ Ranking sprzedawców/kupujących z punktu widzenia kwot sprzedaży
- ⇒ Porównanie zysków operatora w kolejnych okresach (miesiąc, kwartał, półrocze, rok)
- ⇒ Porównanie liczby zakończonych aukcji w poszczególnych okresach z podziałem na kraje