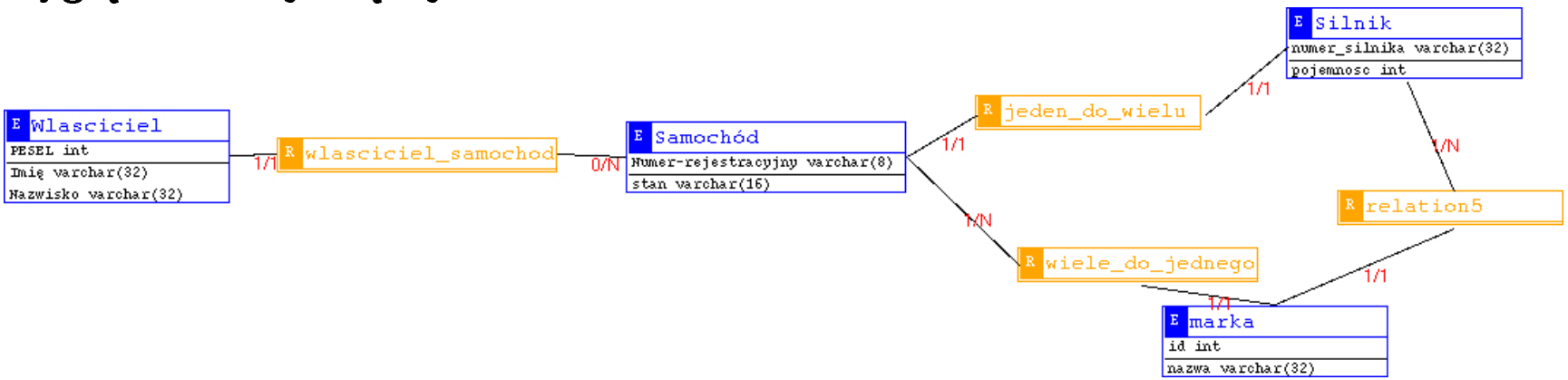
Bazy danych

Jacek Bzdak

April 21, 2009

Wprowadzenie

Wyobraźmy sobie bazę danych przechowującą informację o samochodach i ich właścicielach. Jest ona stworzona w dziale informatyzacji departamentu regulacji ruchu kołowego i rzecznego, a jej schemat wygląda mniej więcej tak:



Ma ona dwa zastosowania:

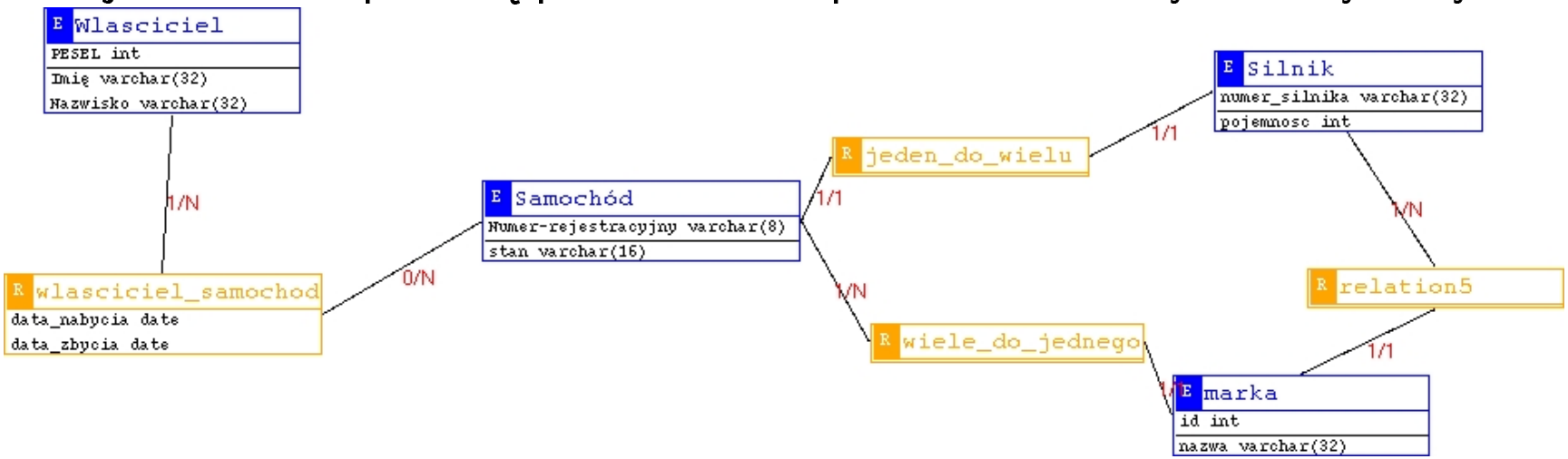
1. Pozwala na naliczenie podatku za posiadanie pojazdu (uzależnionego od pojemności silnika).
2. Pozwala zidentyfikować właściciela samochodu znając jego rejestrację.

Historia kryminalna

Wyobraźmy sobie następującą historię: Zuchwała szajka złodziei dokonuje napadu na muzeum. Zatarli wszystkie ślady – wiadomo tylko jedno: odjechali różowym jaguarem. Komisarz Ryba z właściwą sobie flegmatycznością zwlekał całe trzy dni z wysłaniem zapytania o właściciela samochocu do urzędu regulacji ruchu kołowego i rzecznego. W międzyczasie szajka zbirów zbyła ów samochód na rzecz bogu ducha winnego Jana Kalosza. Baza danych wskadzała jednoznacznie – winnym kradziedzy jest Jan Kalosz. Jedną z właściwości hurtowni danych jest to że dane do niego wprowadzone są niezmienne.

Historii ciag dalszy

Policja nauczona porażką postanowiła wprowadzić zmiany do bazy danych.



Relacja samochodu i właściciela zmieniła się z jeden-do-wiele na wiele-do-wiele oraz owa relacja ma teraz własności: data_zbycia i data_nabycia.Jednak takie rozwiązanie technologiczne ma zasadniczą wadę – komplikuje działanie systemów urzędu regulacji ruchu kołowego i rzecznego. Na poziomie tego systemu nie jest istotne kto był *kiedyś* właścicielem samochodu.

Co to hurtownie danych

Hutrownia danych to system przechowujący dane mający następujące cechy:

niezmienne Po wprowadzeniu danych się nie zmienia Zamiast kasować dane – tworzy się wpis o ich skasowaniu

tematyczność Przechowuje dane grupowane tematycznie Dane są pogrupowane tematami.

historyczność Przechowuje dane historyczne

kanoniczność Model hurtowni danych jest niezależny od modelów danych sysemów które wprowadzają dane. Jeden system może przechowywać dane o płci klienta w polu CHAR(1) o wartości 'K' i 'M' inny może robić to za pomocą odwołania do tabeli słownikowej. Hurtownia danych jest w stanie przetłumaczyć te dane na swój własny model który jest jednolity dla całej organizacji.

Zalety hurtowni

Kwerendy nie obciążają systemów tranzakcyjnych Baza danych może mielić skomplikowany raport nawet 20 minut.

kanoniczność Łatwiej jest odpytywać dane jeśli ich model jest spójny.

spójność Przy wprowadzaniu danych do magazynu bada się ich spójność — więc jest ona zagwarantowana podczas odpytywania.

Kwerendy nie obciążają systemów tranzakcyjnych Baza danych może mielić skomplikowany raport nawet 20 minut.

kanoniczność Łatwiej jest odpytywać dane jeśli ich model jest spójny.

spójność Przy wprowadzaniu danych do magazynu bada się ich spójność — więc jest ona zagwarantowana podczas odpytywania.

1. Życiowy przykład: w jednej z firm gdzie pracowałem był raport który mielił się 20 minut, a że silniki baz dancyh nie wspierają prioryzowania zapytań to cały system w tym czasie leżał. Problem rozwiązano puszczając je codziennie o 4.00 w nocy i keszując wyniki.
2. Jest to szczególnie ważne z okazji tego że schematy hurtowni nie są znormalizowane — o tym potem!

Zalety hurtowni

optymalizacja względem odczytu Schemat bazy danych jest zoptymalizowany do odczytu. Systemy operacyjne organizacji z reguły sprofilowane są względem wydajności operacji tranzakcyjnych – a często opytmalizacje przyśpieszające tranzakcje zwalniają odczyt i *vice verso*.

Użycie przez pracowników nie-technicznych Z reguły schemat hurtowni jest na tyle prosty że osoby bez wykształcenia technicznego mogą samodzielnie pisać zapytania. Ponadto schemat z reguły jest profilowany pod zapytania pisane automatycznie – przez graficzne generatory zapytań.

Zastosowania hurtowni

Podejmowanie decyzji Merytoryczni pracownicy mogą sami pisać zapytania pozwalające im podejmować lepsze decyzje.

Zarządzanie relacjami z klientem Magazyny zawierają całą informacje o danym kliencie — więc pomagają stwierdzić komu dać rabat — a komu nie.

Zarządzanie danymi historycznymi Systemu operacyjnego sieci komórkowej *nie interesuje* jaki plan taryfowy miał klient dwa lata temu. Interesuje go jaki plan ma teraz — żeby mógł wyliczyć wysokość rachunku.

Podejmowanie decyzji Merytoryczni pracownicy mogą sami pisać zapytania pozwalające im podejmować lepsze decyzje.

Zarządzanie relacjami z klientem Magazyny zawierają całą informacje o danym kliencie — więc pomagają stwierdzić komu dać rabat — a komu nie.

Zarządzanie danymi historycznymi Systemu operacyjnego sieci komórkowej *nie interesuje* jaki plan taryfowy miał klient dwa lata temu. Interesuje go jaki plan ma teraz — żeby mógł wyliczyć wysokość rachunku.

1. Natomiast ważne jest żeby takich informacji nie gubić — w ten sposób można badać ewolucję klienta i robić z tego fajne analizy

Przykłady hurtowni

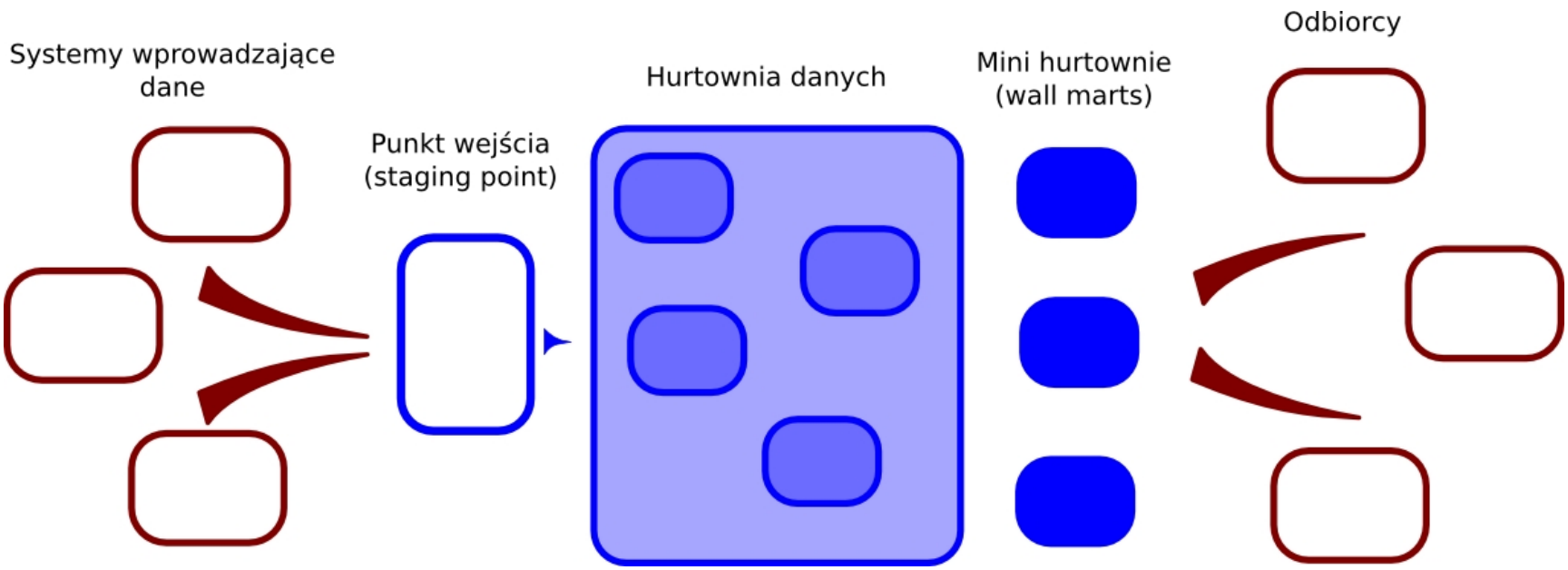
Lista rozmów telefonicznych Pozwala wygenerować billing klienta, podliczyć rachunek, etc.

Historia tranzakcji na koncie bankowym

Historia sprzedarzy

Zarządzanie planami zajęć Podobnie — program studiów się zmienia, natomiast studentom nie można zmieniać go w trakcie studiów.

Architektura



Architektura

punkt wejścia System wprowadzający dane do hurtowni — potrafi tłumaczyć modele danych systemów klienckich na model kanoniczny (albo wymusza użycie modelu kanonicznego). Sprawdza spójność danych. W mojej opinii jest to element niezbędny.

hutrownia Przechowuje i, częściowo, przetwarza dane.

mini markety Wystawiają na świat podzbiory danych z magazynu – na przykład jeśli magazyn zawiera dane o sprzedarzy, jeden z wallmartów może zawierać dane o sprzedarzy u klientów detalicznych.

Punkt wejścia

Może być zaimplementowany różniaście:

Jako reguły walidacji tabel W tym przypadku staging pointa prawie nie ma. Można na przykład sprawdzać czy wartość w polu płeć przyjmuje albo 'K' albo 'M'.

Jako pakiet P-SQL Nie daje się wtedy użytkownikom praw do insertów i updejtów do tabel, a tworzy się procedury PSQL owe które będą sprawdzały poprawność danych.

Jako aplikacja Sama baza danych jako taka nie ma walidacji (albo nie ma ich zbyt wiele). Dane wprowadza się za pośrednictwem aplikacji (na przykład napisanej w JAVIe). Aplikacja może wystawiać różne metody dostępu (GUI, web serwisy, połączenia TCP/IP, RPC).

Może być zaimplementowany różniaście:

Jako reguły walidacji tabel W tym przypadku staging pointa prawie nie ma. Można na przykład sprawdzać czy wartość w polu płeć przyjmuje albo 'K' albo 'M'.

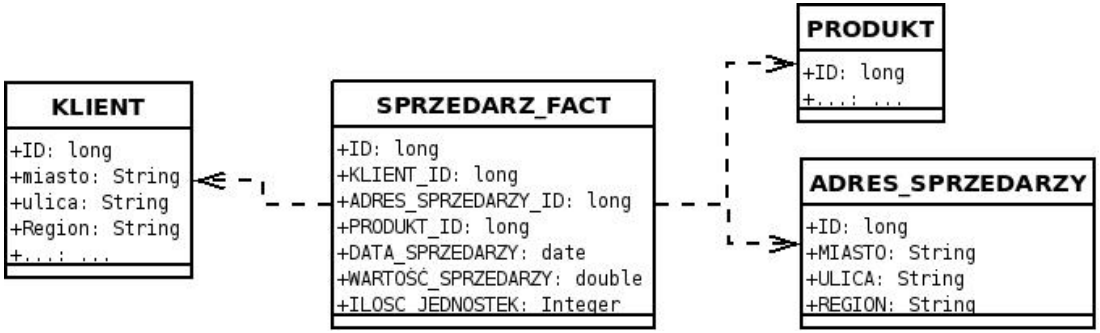
Jako pakiet P-SQL Nie daje się wtedy użytkownikom praw do insertów i updejtów do tabel, a tworzy się procedury PSQL owe które będą sprawdzały poprawność danych.

Jako aplikacja Sama baza danych jako taka nie ma walidacji (albo nie ma ich zbyt wiele). Dane wprowadza się za pośrednictwem aplikacji (na przykład napisanej w JAVIe). Aplikacja może wystawiać różne metody dostępu (GUI, web serwisy, połączenia TCP/IP, RPC).

1. Ja osobiście najbardziej preferuje trzecie rozwiązanie. Pierwsze dwa prowadzą do wprowadzania zbyt dużej ilości logiki do bazy danych. Poza tym — łatwiej się takie walidacje pisze w JAVIE niż SQLu.

Schemat gwiazdy *Star schema*

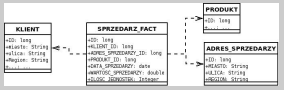
Z lotu ptaka



Dokładniej powiemy o tym później. Na razie intuicja jest taka że:

tabela faktów Tabela zawierająca informacje o zdarzeniu które chcemy rejestrować. Zawiera zarówno informacje atomowe o samym zdarzeniu jak i odniesienia do tabeli wymiarów zawierająde wymiary analityczne zdarzenia.

tabele wymiarów Tabele zawierające dane o umożliwiające analizę zdarzenia. Tabele faktów nie zależą od siebie, oraz nie zawierają odwołań do innych tabel.



Dokładniej powiemy o tym później. Na razie intuicja jest taka że:

tabela faktów Tabela zawierająca informacje o zdarzeniu które chcemy rejestrować. Zawiera zarówno informacje atomowe o samym zdarzeniu jak i odniesienia do tabeli wymiarów zawierającdce wymiary analityczne zdarzenia.

tabele wymiarów Tabele zawierające dane o umożliwiające analizę zdarzenia. Tabele faktów nie zależą od siebie, oraz nie zawierają odwołań do innych tabel.

Dostawcy oprogramowania bazodanowego uznają że hutrownia danych to baza relacyjna, co nie musi być prawdą — hurownia to system mający cechy wymienione na początku prezentacji, to że z reguły robi je się w oparciu o relacyjną bazę danych jest detalem technicznym. W nowszych zastosowaniach datamineing–owych używa rezygnuje się z relacyjnych baz danych (a raczej w ogóle z dyskowych baz danych) na rzecz przetwarzania i przechowywania danych w pamięci RAM. (Zainteresowanych zapraszam na: http://www.amzulsystems.com/ — firma budująca maszyny mające ponad $>$ 500 procesorów i $>$ 500GB RAMu)

Dalej jednak będziemy mówić jak się implementuje hurtownie danych w oparciu o relacujne bazu danych.

Schemat gwiazdy

Normalizacja baz danych

Głównym problemem w zarządzaniu bazami danych jest redundancja (nadmiarowość informacji) tj. jakaś informacja jest w kilku miejscach. Powoduje to problemy z odświerzaniem danych w bazie. 1

Brak wyróżnienia encji Dane tego samego typu są w kilku różnych tabelach. .

Współzależne dane Dane w tabelach zależą od siebie.

Te same dane zawarte są w wielu wierszach tabeli Ta sama informacja jest w kilku wierszach.

Głównym problemem w zarządzaniu bazami danych jest redundancja (nadmiarowość informacji) tj. jakaś informacja jest w kilku miejscach. Powoduje to problemy z odświerzaniem danych w bazie. 1

Brak wyróżnienia encji Dane tego samego typu są w kilku różnych tabelach. .

Współzależne dane Dane w tabelach zależą od siebie.

Te same dane zawarte są w wielu wierszach tabeli Ta sama informacja jest w kilku wierszach.

1. Istnieją formalne definicje stopni normalizacji, ale ja jako dziecko neostrady nie lubie formalizmów powiem więc jak to działa.
2. U nas adres jest w tabeli ADRES_SPRZEDARZY i KLIENT
3. U nas region zależy od adresu
4. Znowu adres: to że Rumiankowa 8A jest w regionie Warszawa jest zawarta w wielu wierszach reprezentujących rózne regiony.
5. Są różne podejścia do denormalizacji - niektórzy piszą: 'Normalize for correctness, denormalize for performance' inni: 'Zdenormalizuj baze danych tylko jeśli masz gwarancję że za pół roku stracisz pracę (kto inny posprząta bałagan).

Schemat gwiazdy

Wady i zalety

Zalety

Szybki!!! Silniki baz danych z reguły mają wspomaganie wyszukiwania po schematach gwiazdy, które działa tak że nie trzeba czytać tabeli faktów aż do momentu w którym zapytanie zwróci wyniki (przeszukuje się indeksy).

Łatwe partycjonowanie Partycjonujemy tylko tabelę faktów (a optymalizator zapytań potrafi zignorować partycje które nie mogą zawierać wyników).

Łatwo pisać zapytania mało joinów

Łatwo zaimplementować mało zmienne wymiary O tym potem

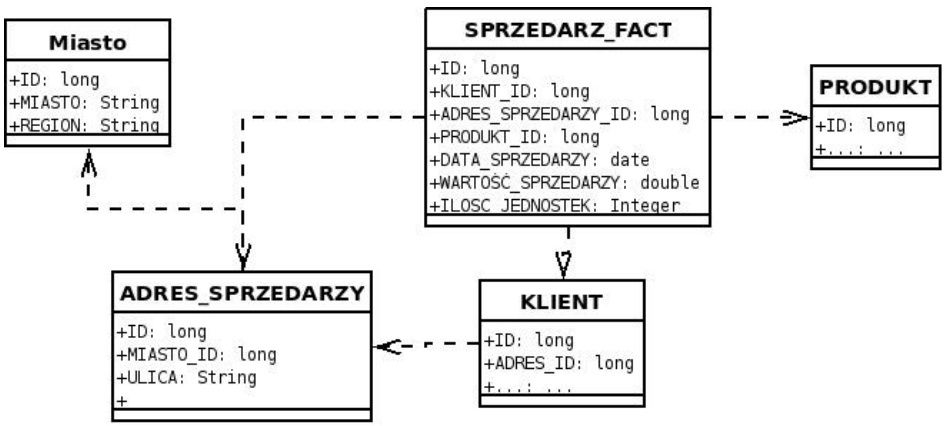
Wady

- Zdenormalizowany .

1. Przed problemami z denormalizacją chroni nas to że dane w magazynie zasadniczo się nie zmieniają. Ale wyobraźmy sobie że chcemy zmienić podział kraju na regiony — wymaga to przeszukania przeszukania DWÓCH dużych tabel, które zawierają powtarzające się wpisy. Nie jest to problem — ale jest to ciężkie do zautomatyzowania. Więc jeśliby się często regionizacja zmieniała warto pomyśleć nad czymś innym

Schemat płatka śniegu *Snowflake schema*

Z lotu ptaka



Normalizacja Tabele faktów są znormalizowane.

Brak zależności między gałęziami

Schemat płatka śniegu *Snowflake schema*

Wady i zalety

Zalety

Oszczędność miejsca Jeśli wymiar zawiera rzadkie (*sparse*, czyli większość kolumn to zawsze NULLE) to można oszczędzić miejsce. (ale — oszczędność może być minimalna bo tabela faktów może zajmować 90% bazy danych.

Łatwy import danych W systemie operacyjnym dane *będą* znormalizowane. Więc nie będzie trzeba przetwarzać danych przy imporcie (mogą mieć tą samą postać). .

Tak myślą użytkownicy Może się okazać że wielowymiarowy model odzwierciedla sposób myślenia użytkowników. .

Narzędzia Są narzędzia dostosowane do współpracy z wielowymiarowymi bazami danych które lepiej radzą sobie z płatkiem śniegu niż gwiazdą. .

Wady

Wolny

Jak się *to* robi w prawdziwym świecie

W praktyce nie używa się czystego schematu gwiazdy i płatka śniegu, ale się je miesza. Część wymiarów jest znormalizowana, część nie.

linki

- Praktyk o hurtowniach – podejście nietechniczne – `http://www.dwinfocenter.org/index.html` (zawiera np. *polityczne* problemy z hurtowniami danych).
- Dlaczego schemat gwiazdy — `http://stackoverflow.com/questions/110032/star-schema-design`
- Gwiazda a platek śniegu `http://oracle-online-help.blogspot.com/2006/11/star-vs-snowflake-schema.html`

Narzędzia

- Prezentacja powstała w LaTeX-ie, z użyciem klasy beamer
- Schematy baz danych powstały w narzędziu ferret i dia
- Obrazek z architekturą bazy danych powstał w inkscapie