Hurtownie danych

31 stycznia 2017

Definicja hurtowni danych

Hurtownia danych – wg Williama Inmona – zbiór danych wyróżniający się następującymi cechami

- uporządkowany tematycznie
- zintegrowany
- zawierający wymiar czasowy
- nieulotny
- wspomaga podejmowanie decyzji
- wspomaga przetwarzanie informacji dla celów strategicznych i analitycznych

Przetwarzanie danych

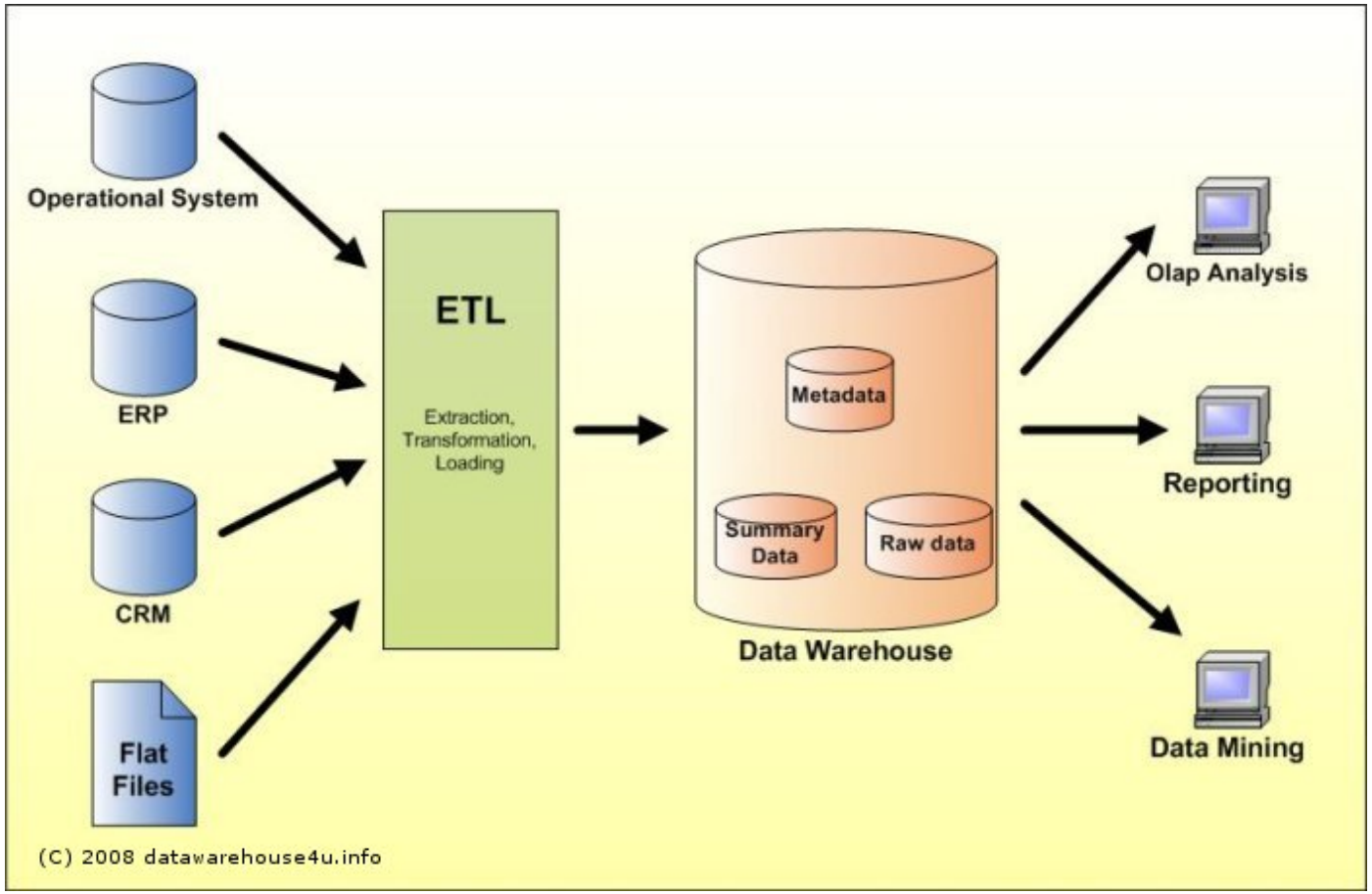
	OLTP OnLine Transaction Processing	OLAP OnLine Analytical Processing
źródło danych	dane opercyjne	dane historyczne
cel	aktualizacje	raportowanie
zastosowania	usługi online, klient-serwer	inteligencja biznesowa, data mining
użytkownicy	klienci, personel	zarząd, marketing
horyzont czasowy	dni, tygodnie, miesiące	lata, dziesiątki lat
model danych	relacyjny	wielowymiarowy
schemat	znormalizowany	gwiazda, płatek śniegu, konstelacja
zapytania	proste	złożone
operacje na danych	szybkie transakcje insert oraz update	cykliczne, długotrwałe importy
optymalizacja	zapisów	odczytu
czas przetwarzania	krótki	długi
objętość	mała	wielka

Projektowanie hurtowni danych

Uwzględniając wymagania jakie są stawiane przed hurtownią danych przy projektowaniu jej schematu należy zwrócić uwagę na dwa aspekty:

- bardzo duża ilść danych: bieżące + historyczne
- osiągnięcie zadawalającego poziomu efektywności zapytań analitycznych

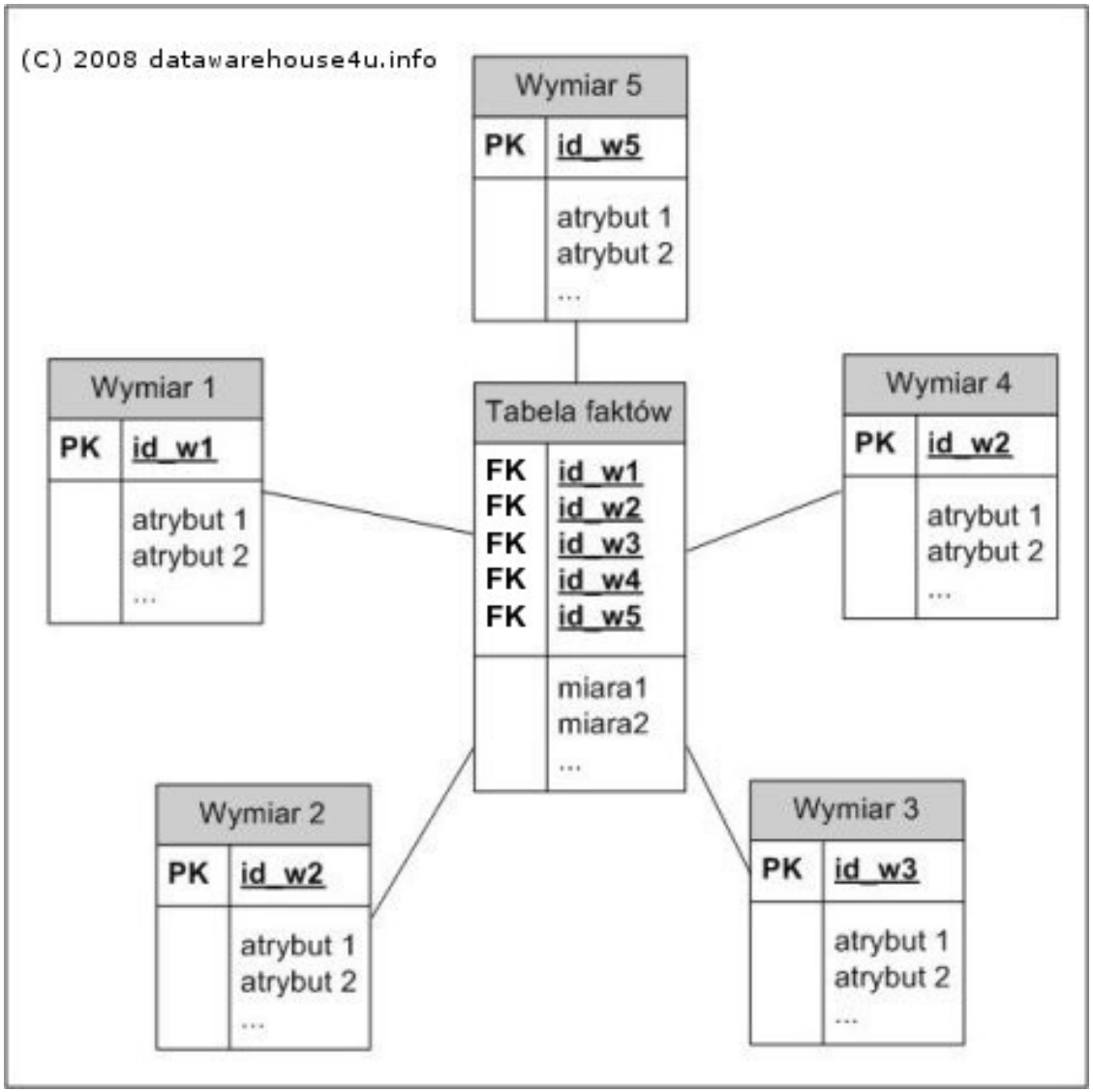
Architektura systemu hurtowni danych



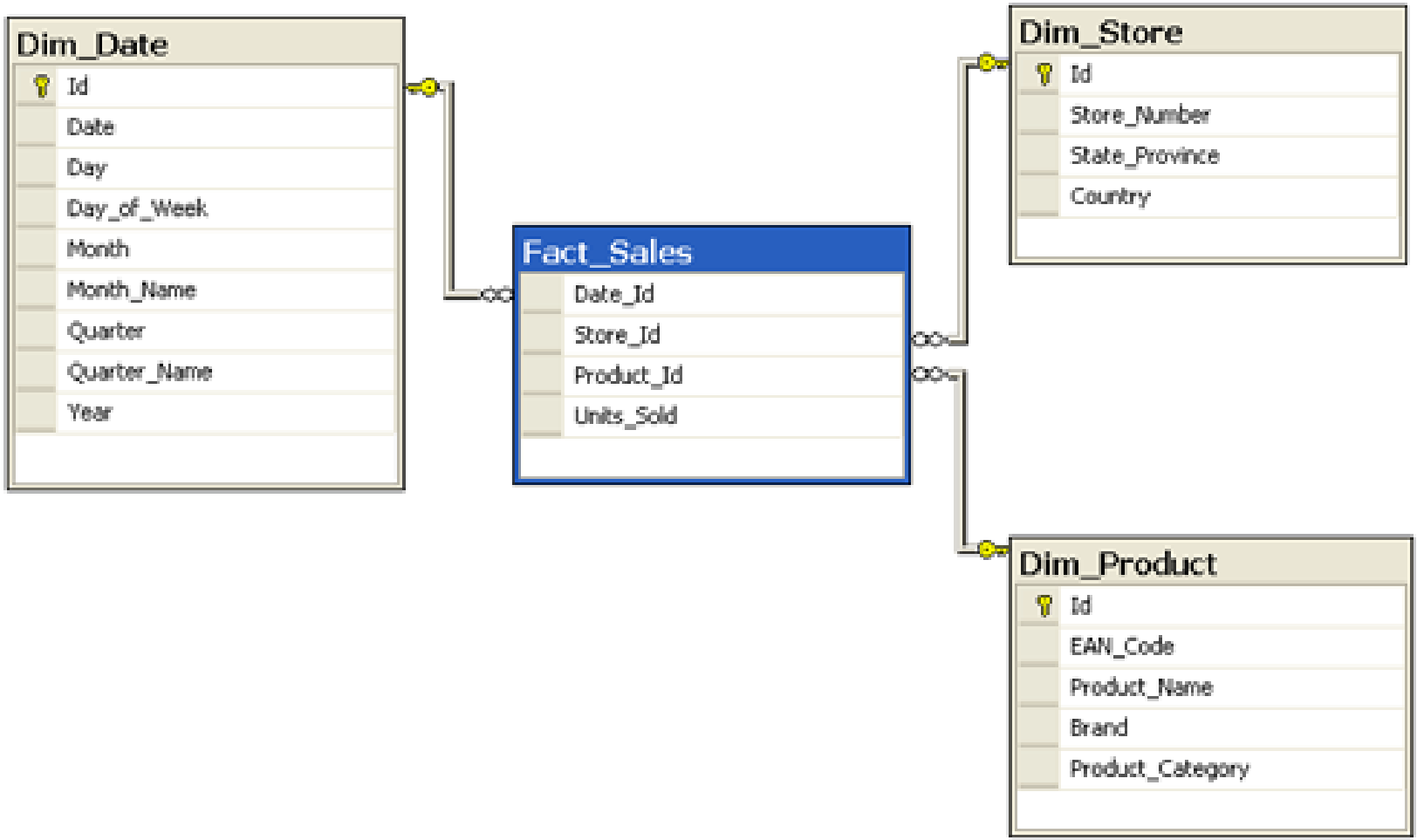
Schemat hurtowni danych

- Schemat gwiazdy (ang. star)
- Schemat płatka śniegu (ang. snowflake)
- Schemat konstelacji faktów (ang. fact constellation albo starflake)

Schemat gwiazdy



Schemat gwiazdy: przykład



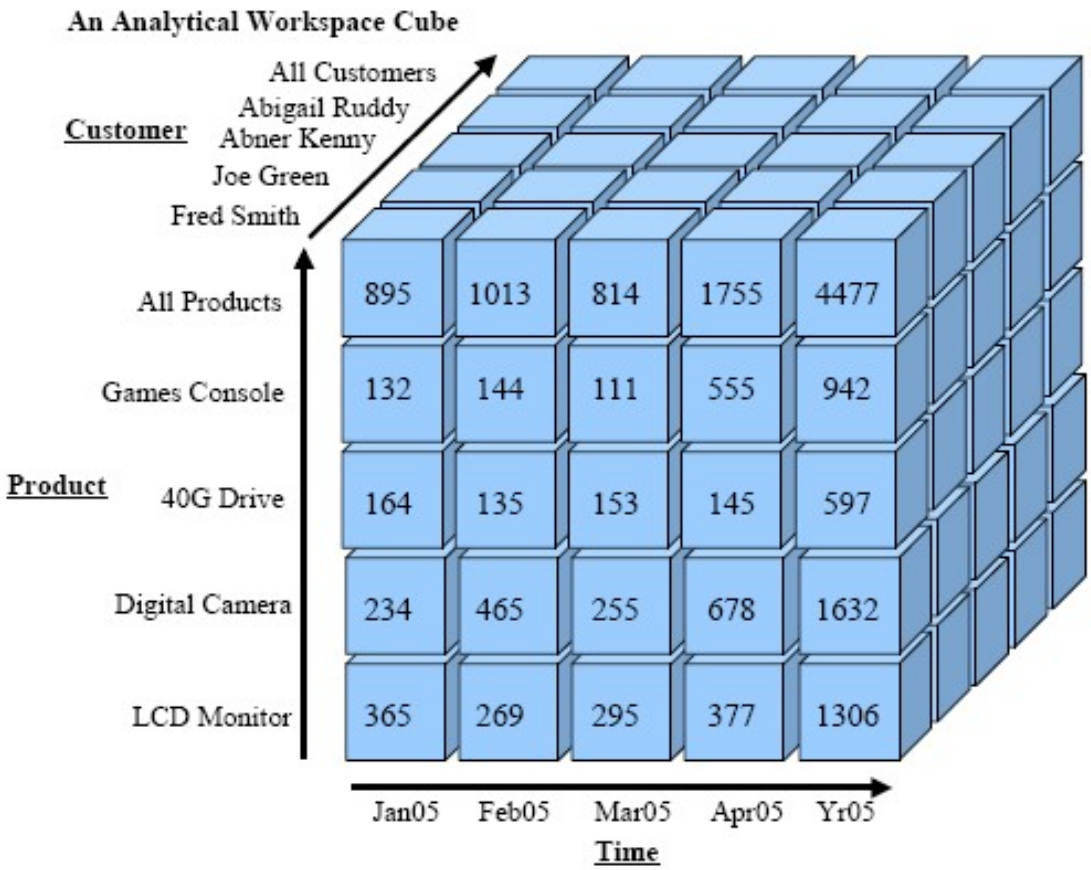
Schemat gwiazdy: cechy

- Prosta struktura
- Duża efektywność zapytań z uwagi na małą liczbę powiązań
- Długi czas ładowania danych do tabel wymiarów z uwagi na denormalizację i w efekcie na redundancję danych
- Tabela faktów składa się z dwóch typów kolumn:
  - miary – wartości numeryczne opisujące dany fakt
  - klucze obce do tabeli wymiarów

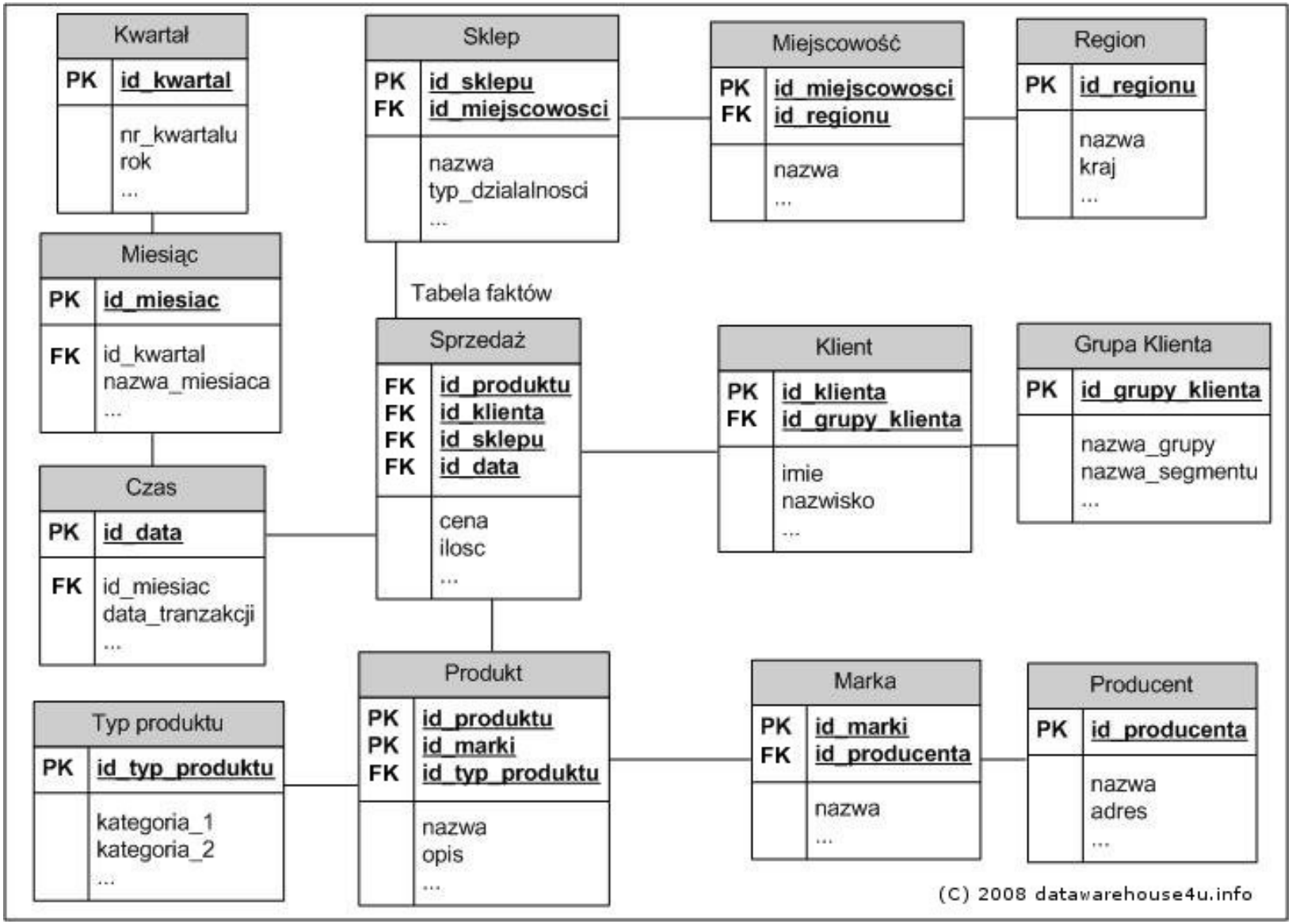
  Wymiary – wartości opisowe danego faktu
- Klucz główny tabeli faktów składa się z jej wszystkich kluczy obcych
- Tabela faktów może zawierać informacje o faktach na poziome detalicznym lub zagregowanym

Dlaczego wymiary?

Wielowymiarowa kostka OLAP



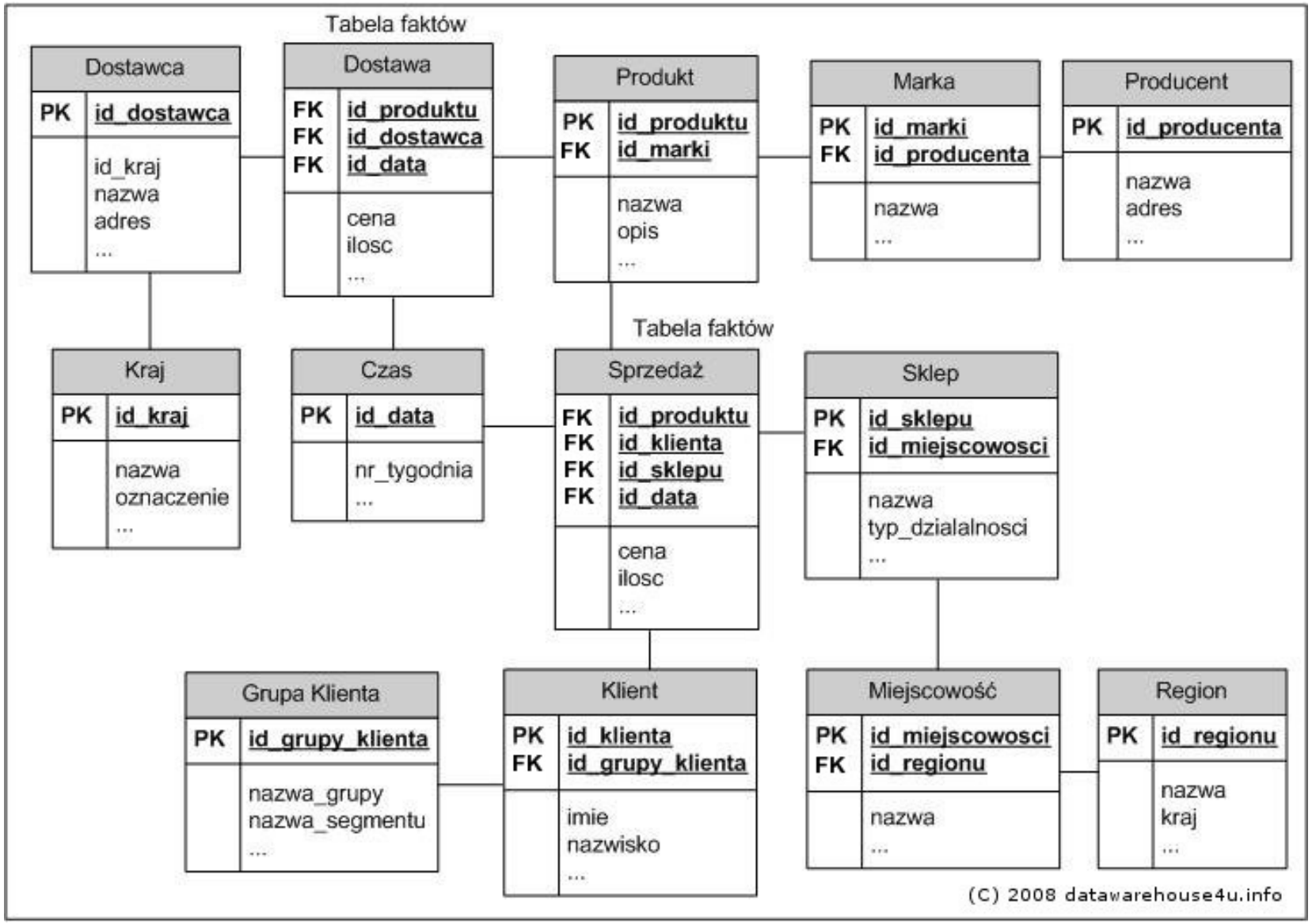
Schemat płatka śniegu



Schemat płatka śniegu: cechy

- Spadek wydajności zapytań w porównaniu ze schematem gwiazdy z uwagi na większą liczbę relacji
- Struktura łatwiejsza do modyfikacji
- Relatywnie krótszy czas ładowania danych do tabel ze względu na mniejszą redundację dzięki normalizacji
- Wykorzystywany rzadziej niż schemat gwiazdy, gdyż efektywność zapytań jest ważniejsza niż efektywnosc ładowania danych

Schemat konstelacji faktów



Schemat konstelacji faktów: cechy

- Rozwiązanie pośrednie między gwiazdą a płatkiem śniegu
- Zwykle występuje więcej niż jedna tabela faktow
- Cześć tabel wymiarów jest znormalizowana a część zdenormalizowana
- Tabele faktów są ze sobą powiązane w relacji $1:1$ lub $1:n$
- Tabele faktów mogą współdzielić te same tabele wymiarów

Factless Fact Table

- Tabela faktów bez faktów nie zawiera miar
- Przekrój wymiarów

Przykład

Obecność studentów na zajęciach jako przekrój 3 wymiarów: *studenci*, *czas*, *zajęcia*. Tabela faktów w tym wypadku złożona jest z trzech kolumn zawierających klucze obce. Pozwala łatwo udzielić odpowiedzi na następujące pytania:

- Ilu studentów było obecnych na danych zajęcich w określonym terminie?
- Na ile zajęć średnio uczęszczał dany student danego dnia?

Junk Dimension

- Nader często zdarzają się sytuacje, gdzie miary są jednobitowe (1/0)
- Przechowywanie danych jednobitowych w tabelach faktów wymaga użycia małych tabel wymiarów
- Jednocześnie ilość danych w tabelach faktów gwałtownie rośnie

Przykład

FACT_TABLE

CUSTOMER_ID
PRODUCT_CD
TXN_ID
STORE_ID
TXN_CODE
COUPON_IND
PREPAY_IND
TXN_AMT

$\longrightarrow$

FACT_TABLE

CUSTOMER_ID
PRODUCT_CD
TXN_ID
STORE_ID
JUNK_ID
TXN_AMT

$\sim$

DIM_JUNK

JUNK_ID	TXN_CODE	COUPON_IND	PREPAY_IND
1	1	Y	Y
2	2	Y	Y
3	3	Y	Y
4	1	Y	N
5	2	Y	N
6	3	Y	N
7	1	N	Y
8	2	N	Y
9	3	N	Y
10	1	N	N
11	2	N	N
12	3	N	N

Silniki hurtowni danych

- ROLAP — Relational OLAP
  - Przetwarzanie dużej ilości danych
  - Naturalne wykorzystanie własności relacyjnych baz danych
  - Niezbyt duża wydajność
  - Ograniczeni narzucone przez model relacyjny
  - IBM DB2 OLAP, Informix MetaCube, MicroStrategy
- MOLAP — Multidimensional OLAP
  - Znakomita wydajność
  - Wylicznia generowane w czasie tworzenia kostki, co pozwala szybko wykonywać skomplikowane operacje
  - Ograniczenia pojemności
  - Oracle Express Server
- HOLAP — Hybrid OLAP
  - Podsumowania w fomie wielowymiarowych kostek
  - Dane szczegółowe w bazie relacyjnej
  - Microsoft SQL Server OLAP Services