Hurtownie danych

dr hab. Maciej Zakrzewicz
Politechnika Poznańska
Instytut Informatyki

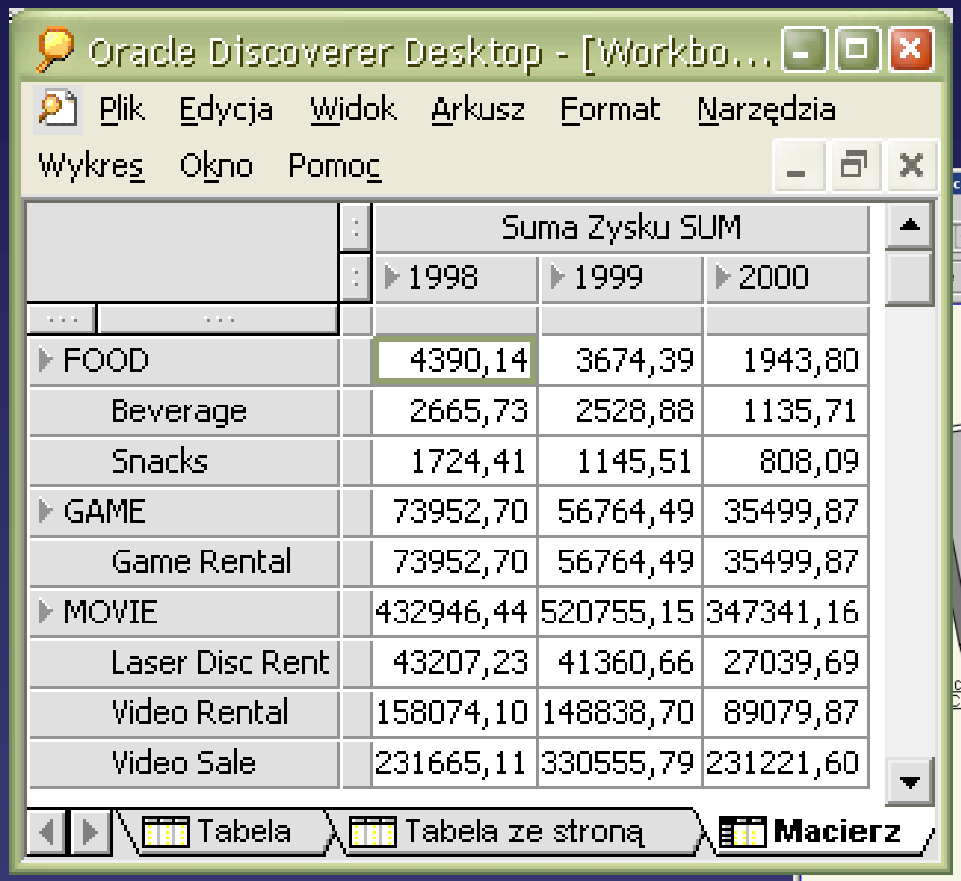
Plan wykładu

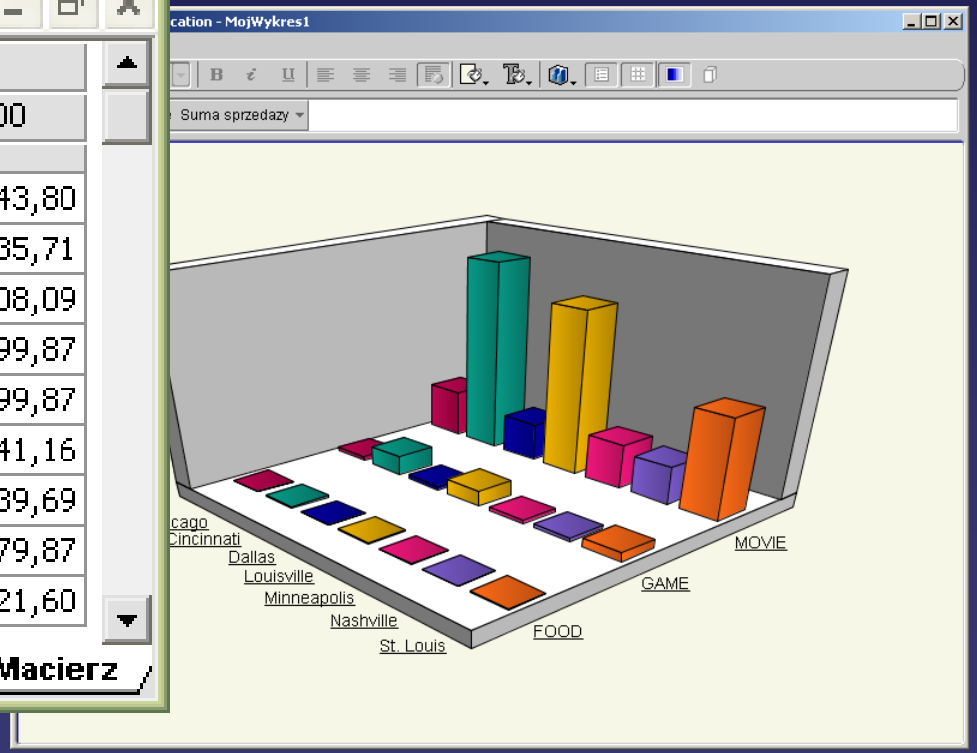
- Wprowadzenie do Business Intelligence (BI)
- Hurtownia danych
- Zasilanie hurtowni danych (ETL)
- Przetwarzanie OLAP
- Projektowanie hurtowni danych
- Podsumowanie

Business Intelligence

- Technologia informatyczna służąca przekształcaniu dużych wolumenów danych w informacje, a następnie przekształcaniu tych informacji w wiedzę
- Adresowana do pracowników szczebla kierowniczego, wspomagająca podejmowanie ich decyzji
- Stawia drastyczne wymagania wydajnościowe,przede wszystkim z powodu ogromnych rozmiarów danych, które podlegają przetwarzaniu
- Skupiona wokół technologii hurtowni danych

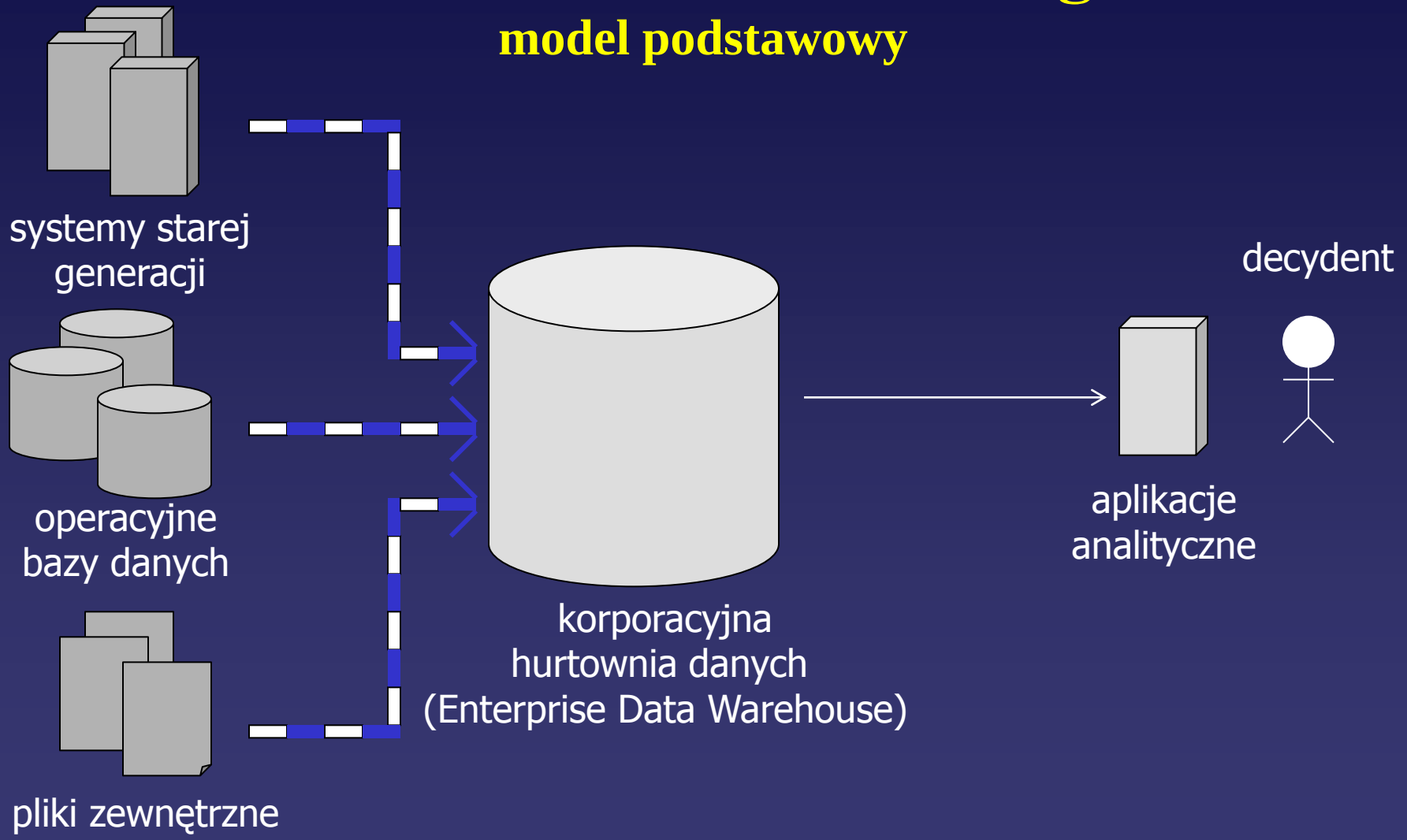
Business Intelligence - przykłady zastosowań





Środowisko Business Intelligence

model podstawowy



Środowisko Business Intelligence

z obszarem składowania

Środowisko Business Intelligence

z obszarem składowania i składnicami danych

Hurtownia danych - definicja

- „Hurtownia danych to tematyczna baza danych, która trwale przechowuje zintegrowane dane opisane wymiarem czasu” [Inmon96]
- „Tematyczna baza danych” – dane dotyczą głównych obszarów działalności przedsiębiorstwa
- „trwale przechowuje” – dane nie są zmieniane ani usuwane; hurtownia danych ma charakter przyrostowy
- „zintegrowane dane” – dane dotyczące tego samego podmiotu stanowią całość
- „opisane wymiarem czasu” – dane opisują zdarzenia historyczne, a nie tylko stan aktualny

Porównanie hurtowni danych z systemami klasycznymi

Cecha	System klasyczny	Hurtownia danych
czas odpowiedzi aplikacji	ułamki sekundy – sekundy	sekundy – godziny
wykonywane operacje	DML	select
czasowy zakres danych	30-60 dni	2-10 lat
organizacja danych	według aplikacji	tematyczna
rozmiar	małe – duże	duże – wielkie
intensywność operacji dyskowych	mała – średnia	wielka

Porównanie hurtowni danych ze składnicami danych

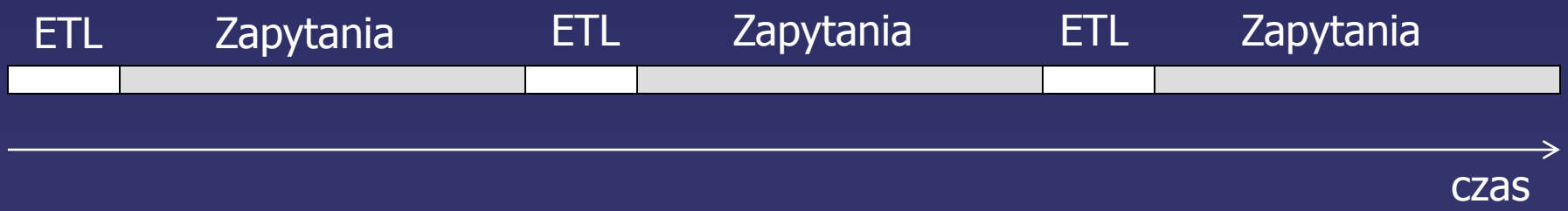
Cecha	Hurtownia danych	Składnica danych
zasięg wykorzystywania	przedsiębiorstwo	wydział
zakres tematyczny	wielotematyczna	jednotematyczna
liczba źródeł danych	wiele	1 – kilka
czas implementacji i wdrożenia	miesiące – lata	miesiące

ETL: Extraction, Transformation, Loading

- Ekstrakcja: odczyt źródłowych danych z operacyjnych baz danych, systemów starej generacji, plików zewnętrznych
- Transformacja: łączenie danych, ich weryfikacja, walidacja, czyszczenie i znakowanie czasowe
- Wczytywanie: wprowadzanie danych do docelowej hurtowni danych
- Realizacja ETL jest najtrudniejszym zadaniem implementacji hurtowni danych (pochłania nawet 70% czasu)

Dwa tryby pracy hurtowni danych

- Ładowanie danych
  - zwykle wykonywane w regularnych odstępach czasu, w porze niskiej aktywności użytkowników
- Realizacja zapytań analitycznych
  - podstawowy rodzaj obciążenia systemu hurtowni danych



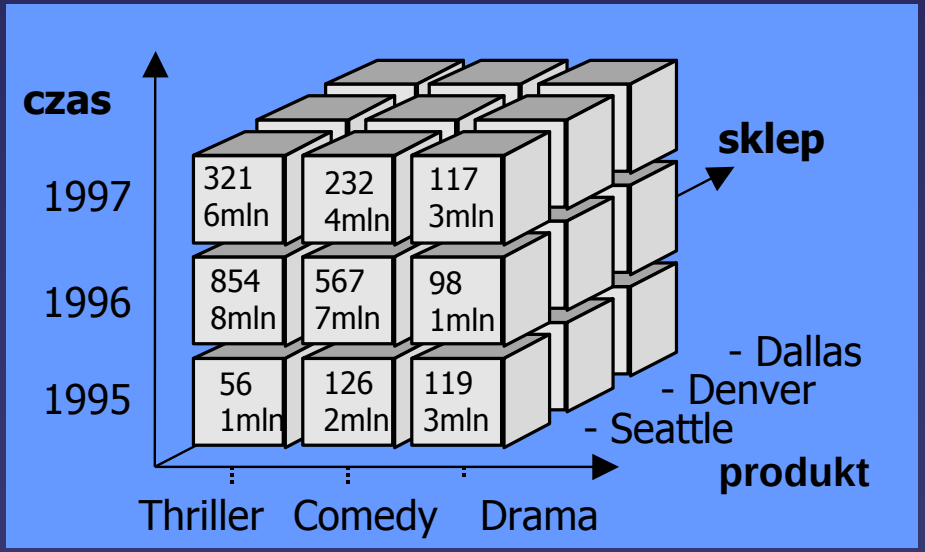
Zmienność charakterystyki obciążenia systemu hurtowni danych to poważny problem konfiguracyjny

On-Line Analytical Processing (OLAP)

- W środowisku BI konstruuje się aplikacje, które w wydajny sposób realizują złożone analizy danych
- Umożliwiają analizę danych w trybie ad-hoc
- Użytkownicy określają swoje żądania, co do analizy danych, a system bezpośrednio je realizuje
- Przetwarzanie OLAP (ang. On-Line Analytical Processing), w odróżnieniu od klasycznego przetwarzania transakcyjnego OLTP (ang. On-Line Transaction Processing)
- Bezpośrednio wspomagają pracę pracowników szczebla kierowniczego

Wielowymiarowy model danych

- Dane na potrzeby przetwarzania OLAP są w naturalny sposób przedstawiane w postaci wielowymiarowej
- Logiczne kostki stanowią sposób organizacji miar mających te same wymiary

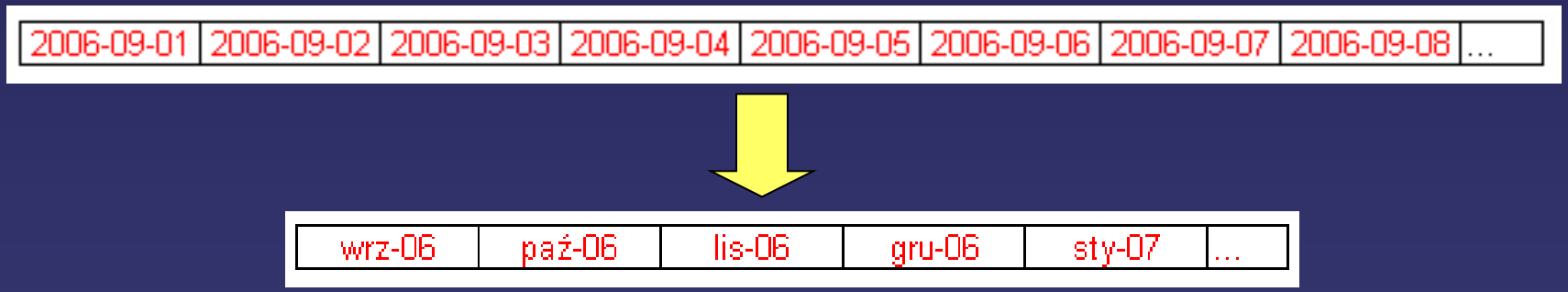


Miary i wymiary

- Fakty (ang. facts), miary (ang. measures)
  - Wartości ciągłe, numeryczne
  - Typowe miary: wartość sprzedaży, koszt, zysk
- Wymiary (ang. dimensions)
  - Wartości dyskretne, niezmienne lub rzadko zmienne
  - Nadają znaczenie danym (miarom, faktom)
  - Typowe wymiary: klient, czas, produkt, sklep
  - Hierarchie (ang. hierarchies) umożliwiają organizację danych na różnych poziomach agregacji
  - Poziom (ang. level) reprezentuje pozycję w hierarchii
  - Atrybuty (ang. attributes) dostarczają dodatkowych informacji o danych, np. kolor, smak, dzień tygodnia

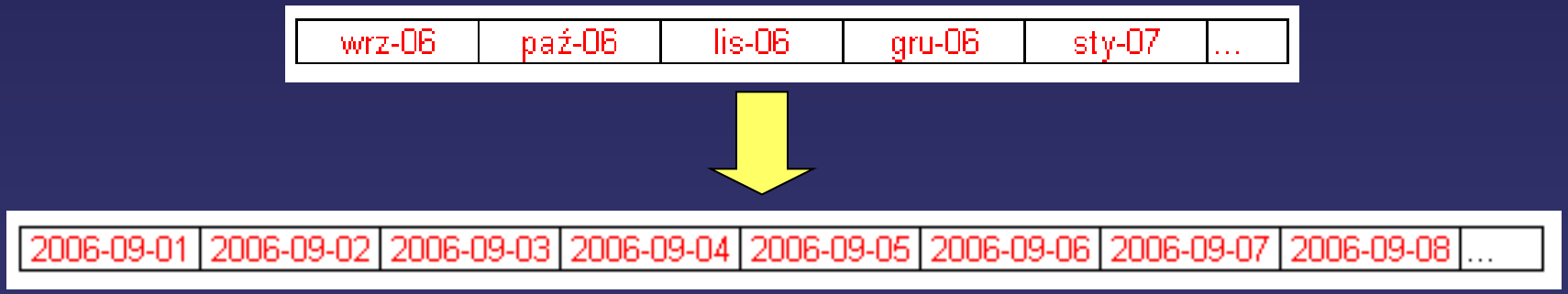
Agregacja, konsolidacja, zwijanie wymiaru

- Przejście na ogólniejszy poziom wymiaru
- Wybór operacji agregującej miarę, np. suma, średnia
- Aggregation, consolidation, roll-up



Rozwijanie wymiaru

- Przejście na bardziej szczegółowy poziom wymiaru
- Roll-down, drill-down



Selekcja

- Wybór fragmentu danych poprzez zawężenie wartości wymiarów
- Slicing and dicing

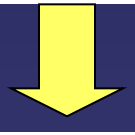
	2003	2004	2005	2006
Poznań	2 342,22 zł	4 543,00 zł	3 432,12 zł	5 200,00 zł
Warszawa	3 432,10 zł	3 223,00 zł	6 445,00 zł	4 980,30 zł
Kraków	2 322,00 zł	232,00 zł	1 290,50 zł	3 400,00 zł
Wrocław	1 781,00 zł	1 234,00 zł	3 440,00 zł	3 100,00 zł

	2005	2006
Poznań	3 432,12 zł	5 200,00 zł
Warszawa	6 445,00 zł	4 980,30 zł

Obrót

- Zamiana miejscami wymiarów
- Wymiana wyświetlanych wymiarów
- Pivot

	2003	2004	2005	2006
Poznań	2 342,22 zł	4 543,00 zł	3 432,12 zł	5 200,00 zł
Warszawa	3 432,10 zł	3 223,00 zł	6 445,00 zł	4 980,30 zł
Kraków	2 322,00 zł	232,00 zł	1 290,50 zł	3 400,00 zł
Wrocław	1 781,00 zł	1 234,00 zł	3 440,00 zł	3 100,00 zł



	Poznań	Warszawa	Kraków	Wrocław
2003	2 342,22 zł	3 432,10 zł	2 322,00 zł	1 781,00 zł
2004	4 543,00 zł	3 223,00 zł	232,00 zł	1 234,00 zł
2005	3 432,12 zł	6 445,00 zł	1 290,50 zł	3 440,00 zł
2006	5 200,00 zł	4 980,30 zł	3 400,00 zł	3 100,00 zł

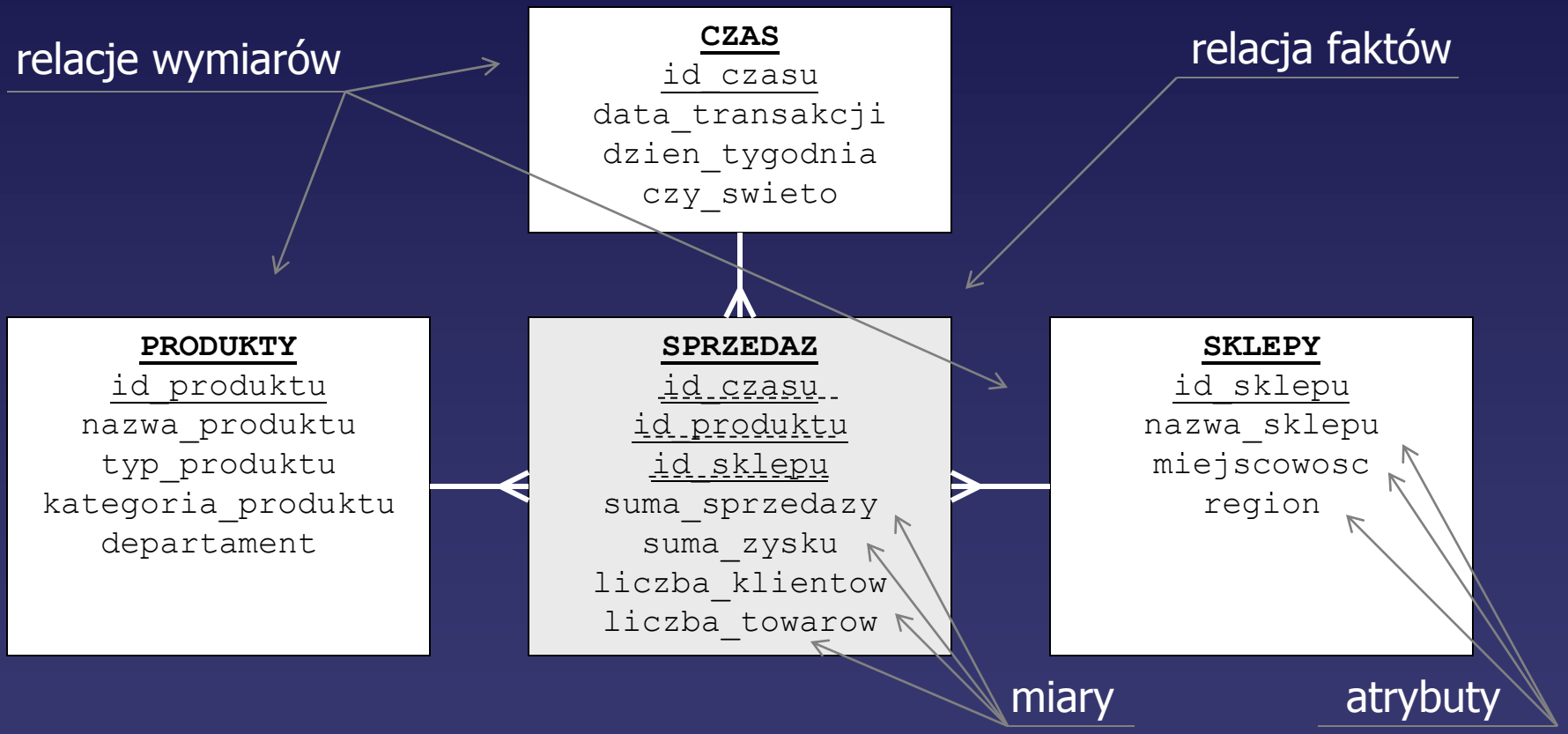
Kroki technicznej implementacji systemu Business Intelligence

- Analiza wymagań
- Projekt logiczny hurtowni danych
- Implementacja struktur fizycznych hurtowni danych
- Implementacja oprogramowania ETL
- Realizacja aplikacji analitycznych
- Strojenie hurtowni danych

Implementacje logicznego wielowymiarowego modelu danych

- Relacyjna implementacja modelu (ROLAP)
  - Powiązane ze sobą relacje: relacje faktów i wymiarów
  - Schematy logiczne:
    - Schemat gwiazdy (ang. star schema)
    - Schemat płatka śniegu (ang. snowflake schema)
    - Konstelacja faktów (ang. fact constellation)
  - Materializowane perspektywy dla wartości agregowanych
- Wielowymiarowa reprezentacja modelu (MOLAP)
  - Dane fizycznie składowane w postaci wielowymiarowej

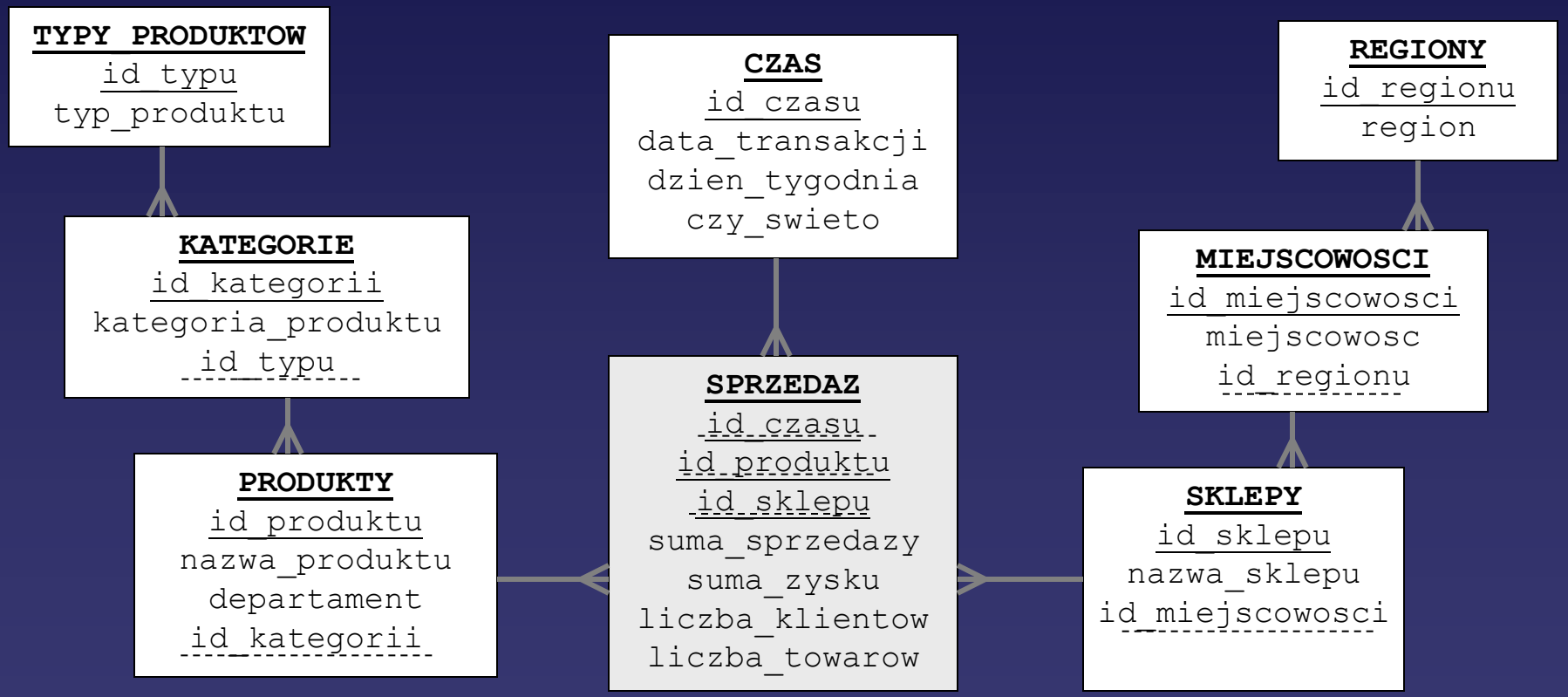
Schemat gwiazdy



Charakterystyka schematu gwiazdy

- Centralna relacja faktów
- Zdenormalizowane relacje wymiarów (1NF)
- Relacja faktów połączona z relacjami wymiarów poprzez klucze główne i klucze obce
- Prosta struktura
- Duża efektywność zapytań ze względu na niewielką liczbę połączeń relacji
- Stosunkowo długi czas ładowania danych do relacji wymiarów ze względu na denormalizację
- Dominująca struktura dla hurtowni danych, wspierana przez wiele narzędzi

Schemat płatka śniegu

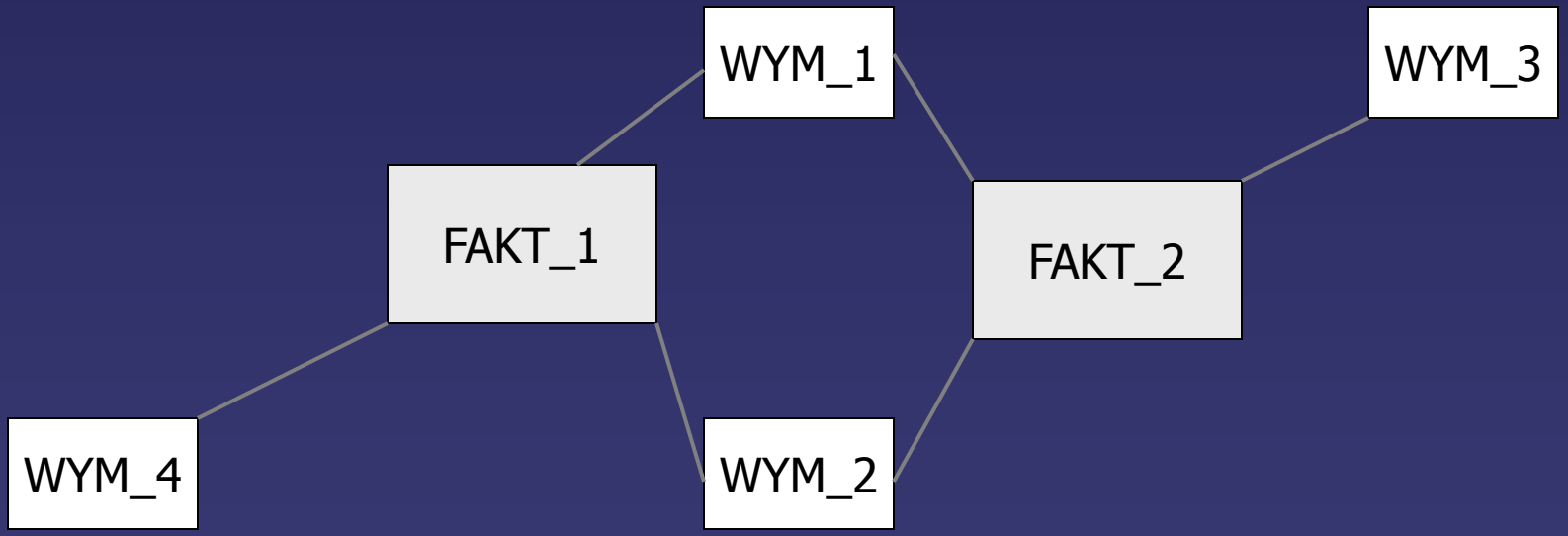


Charakterystyka schematu płatka śniegu

- Centralna relacja faktów
- Znormalizowane relacje wymiarów (3NF)
- Spadek wydajności zapytań w porównaniu ze schematem gwiazdy ze względu na większą liczbę połączeń relacji
- Struktura łatwiejsza do modyfikacji
- Krótki czas ładowania danych do relacji (mniejszy rozmiar relacji)
- Wykorzystywany rzadziej niż schemat gwiazdy, gdyż efektywność zapytań jest ważniejsza niż efektywność ładowania danych do relacji wymiarów

Konstelacja faktów

- Schemat stanowiący kombinację schematów gwiazd współdzielących niektóre wymiary
  - Różne relacje faktów mogą odwoływać się do różnych poziomów danego wymiaru



Charakterystyka relacji faktów i wymiarów

- Relacja faktów:
  - Zawiera numeryczne miary
  - Posiada wieloatrybutowy klucz główny złożony z kluczy obcych odwołujących się do wymiarów
  - Największy rozmiar spośród relacji tworzących gwiazdę
    - Typowo zawiera ponad 90% danych
  - Jej rozmiar szybko się powiększa

Charakterystyka relacji faktów i wymiarów

- Relacje wymiarów:
  - Zawierają atrybuty opisowe
  - Nadają znaczenie faktom
    - Definiują „przestrzeń faktów”
  - Zawierają dane stosunkowo statyczne
    - Podlegają zmianom np. pojawianie się nowych klientów, produktów

Technologie dla hurtowni danych

- Rozszerzenia języka SQL
- Uniwersalne narzędzia ETL
- Partycjonowanie relacji
- Indeksy bitmapowe
- Narzędzia implementacji aplikacji analitycznych

Podsumowanie

- Cechy środowisk Business Intelligence
- Modele danych
- Zagadnienia wydajności

Literatura uzupełniająca

Matthias Jarke, Maurizio Lenzerini, Yannis Vassiliou:

"Hurtownie danych - podstawy organizacji i funkcjonowania"

Chris Todman:

"Projektowanie hurtowni danych"