Fundusze Europejskie
Wiedza Edukacja Rozwój

Politechnika Wrocławska

Unia Europejska
Europejski Fundusz Społeczny

„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

Hurtownie danych

Potrzeby transakcyjne i analityczne

Podstawowe pojęcia

dr inż. Bernadetta Maleszka

Dane -> informacje -> wiedza

- Dane, informacje – najważniejsze wartości firmy
- Cele:
  - operacyjne – gromadzenie i przetwarzanie
  - analityczne – uzyskiwanie informacji z danych i podejmowanie decyzji
    - zarządzanie

- Zarządzanie jest sztuką zadawania właściwych pytań i podejmowania odpowiednich działań biznesowych w zależności od otrzymanych odpowiedzi

Dlaczego OLTP nie wystarcza?

- Operacje transakcyjne i analityczne na tej samej bazie:
  - modyfikacje
  - agregacje
- Problemy:
  - potrzeba doświadczenie i umiejętności pracownika
  - wydajność
- Rozwiązanie:
  - separacja danych

Hurtownia danych - definicja

Hurtownia danych to:

- **tematycznie zorientowana**
- **zintegrowana**
- **chronologiczna**
- **trwała**

kolekcja danych do wspomagania procesów podejmowania decyzji

Hurtownia danych

- **Hurtownia danych to:**
  - **tematycznie zorientowana**
  - **zintegrowana** – dane skonsolidowane (łączone, scalane, transformowane) dane pozyskiwane z różnych źródeł (systemów, arkuszy Excel, innych źródeł)
  - **chronologiczna**
  - **trwała**

kolekcja danych do wspomagania procesów podejmowania decyzji

- Dane przechowywane w hurtowni są spójne, zintegrowane, łatwo (technicznie) dostępne i gromadzone na przestrzeni czasu (historyczne).

Hurtownia danych

- „Hurtownia danych to kopia danych transakcyjnych zaprojektowana pod kątem zapytań i raportowania”

Ralph Kimball „The Data Warehouse Toolkit”

- „Hurtownia danych to osobne repozytorium danych, w którym informacja jest przechowywana w formacie odpowiednim dla systemów Business Intelligence i DSS”

SAS Rapid Data Warehousing

Hurtownia danych - procesy

- Hurtownia danych wiąże się z procesem:
  - pozyskiwania,
  - „czyszczenia” (weryfikowania, uzgadniania),
  - przetwarzania danych pozyskiwanych z różnych źródeł do poziomu informacji, która jest zrozumiała i dostępna do odbiorcy biznesowego

- Hurtownia danych stanowi jednolitą platformę informacyjną do wszechstronnych zastosowań w dziedzinie systemów wspomagających zarządzanie

Separacja danych

- Hurtownia – kopia danych przygotowanych do celów analitycznych

- DBMS – OLTP:
  - modyfikacja stanu bazy danych
  - INSERT, UPDATE, DELETE, MERGE

- Hurtownia danych – OLAP:
  - kompleksowe raporty
  - widoki wielowymiarowe
  - konsolidacje

Mity i fakty

- **Mity na temat hurtowni danych (HD):**
  - HD to „bardzo duża” baza danych
  - HD wymaga specjalnych modeli przechowywania danych
  - HD to głównie wyzwania w zakresie technologii
  - Hurtownia danych = Business Intelligence
- *„Koncepcja zarządzania mająca na celu zapewnienie menedżerom informacji o odpowiedniej jakości i w odpowiednim momencie”*

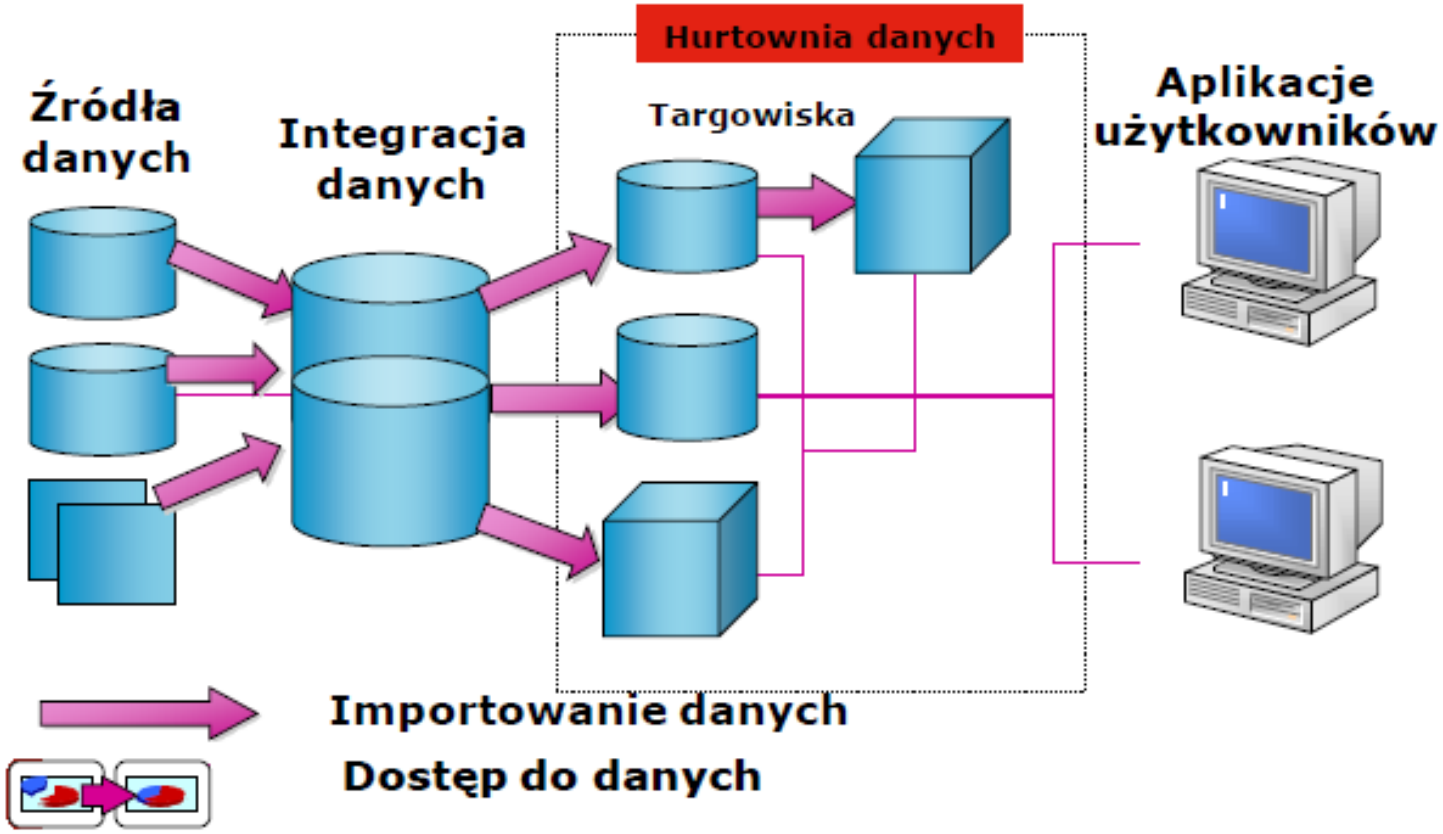
  „Informatyka narzędziem zarządzania w XXI wieku” red. J. Kisielnicki

- **Hurtownia danych to:**
  - Kolekcja danych=f(atrybuty, cel)
  - Model, koncepcja, filozofia zarządzania danymi w przedsiębiorstwie

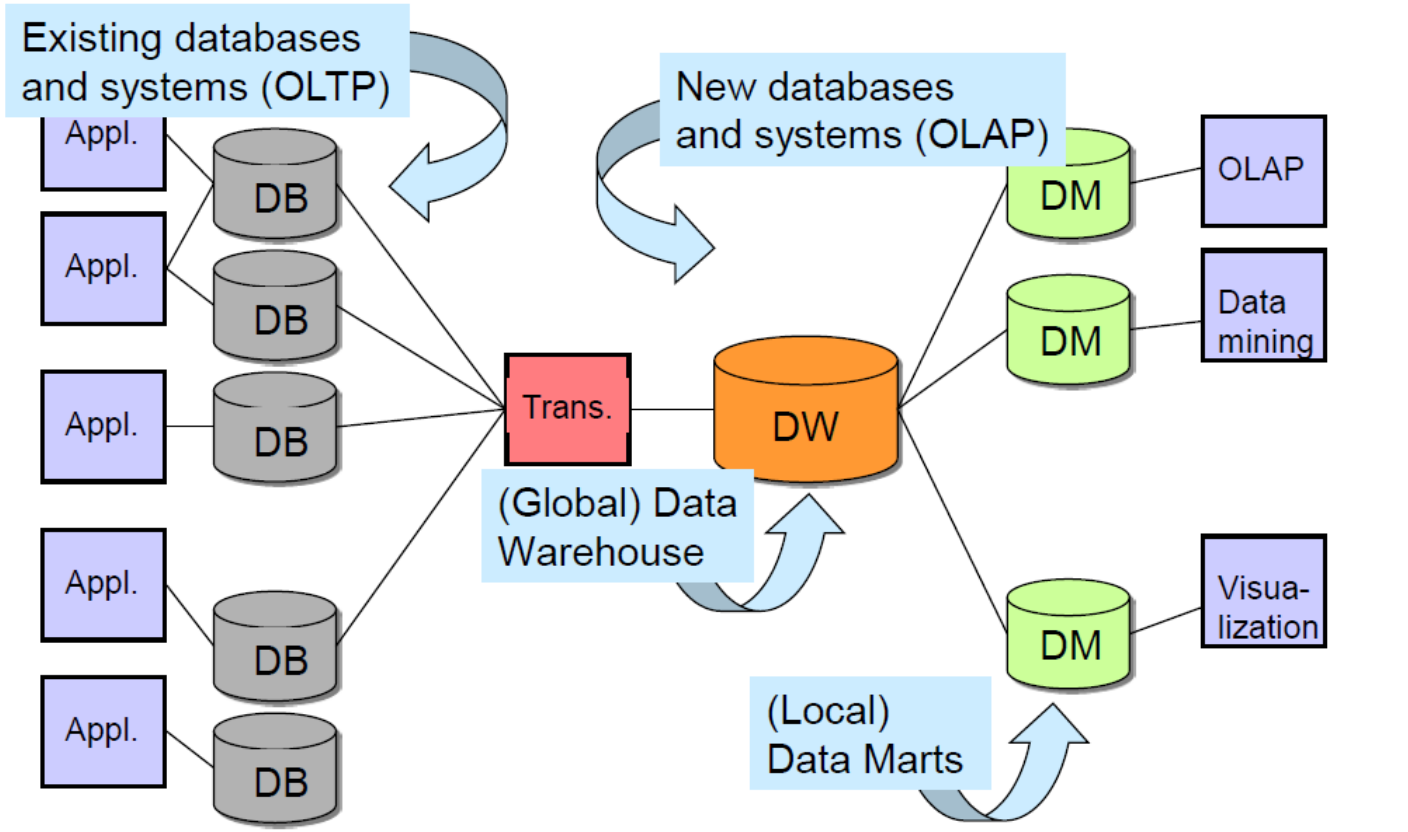
HD vs OLTP

Aspekt	HD	OLTP
Tematycznie zorientowana	Ścisły podział na obszary informacyjne (tematyczne); lokalizacja danych zależy od tematyki	Aplikacje projektowane są wokół procesów oraz funkcji, które wymagają różnych danych
Zintegrowana	Spójność danych, kluczy, synonimy, homonimy, rekordy logiczne...	Jedno źródło, bez integracji
Chronologiczna	Dane są znakowane czasowo, szerszy horyzont czasowy np. 5 lat, historyzacja zmian	Dane są znakowane czasowo w wąskim horyzoncie np. 30-90 dni
Trwała	Dane nie są modyfikowane przez użytkowników, odczyt danych	Transakcje (insert, delete, update)
Cel	Wspomaganie procesów podejmowania decyzji (raczej na poziomach: taktycznym i strategicznym)	Ewidencja transakcji, wspieranie operacyjnego poziomu zarządzania (również podejmowania decyzji na tym poziomie)

Hurtownia danych



Separacja danych



Analogy: (data) producers ↔ warehouse ↔ (data) consumers

Separacja danych

- W celu zapewnienia akceptowalnej jakości usług transakcyjnych i analitycznych dokonano odseparowania danych analitycznych od transakcyjnych

- Efekt:
  - Bazy transakcyjne (dane znormalizowane)
  - Bazy analityczne -> **Hurtownie danych** (Data Warehouse) – dane zdenormalizowane, wstępnie zagregowane

Projektowanie hurtowni

- Opracowanie modelu:
  - pojęciowego – definicje pojęć, opis struktury, zawartości i przeznaczenia
  - logicznego – opis logiczny bazy danych i procesów hurtowni
  - fizycznego – cechy implementacyjne: indeksowanie, partycjonowanie, archiwizacje, itp.

- Projektowanie wstępujące - od szczegółu do ogółu
- Projektowanie zstępujące

Źródła danych

- Dane pochodzące z heterogenicznych źródeł – niezbędne procesy:
  - Integracja
  - Czyszczenie
  - Odświeżanie
- Problemy:
  - Niejednorodność danych
  - Niespójność danych
  - Brak danych
  - …
- Zasilanie hurtowni - proces ETL – extract, transform, load

Cele tworzenia hurtowni danych

- Wykonywanie analiz biznesowych bez ingerencji w systemy transakcyjne
- Wspomaganie decyzji (Decision Support Systems, Knowledge Discovery in Databases)
- Całościowy wgląd w dane firmy
- Dostęp do danych historycznych
- Ujednolicenie posiadanych informacji

Typowe zastosowania

- Analiza trendów i zachowań
- Wykrywanie oszustw
- Ukierunkowany marketing
- Analiza rentowności
- Zapobieganie odejściu klienta
- Zarządzanie zasobami
- Automatyczne generowanie zamówień
- Analiza ryzyka kredytowego
- Długoterminowa ocena wartości klienta

Funkcje i użytkownicy systemów analitycznych

Funkcje systemów analitycznych:

- Zrozumieć (przedsięwzięcie, miary PKI)
- Wspomóc przy podejmowaniu decyzji
- Dostarczyć prognozy
- Wykryć trendy

Użytkownicy systemów analitycznych:

- Dyrekcja i zarząd
- Analitycy, kontrolerzy, planiści
- Pracownicy wiedzy
- Aplikacje

Założenia projektowe

Wymagania:

- Jakie decyzje mają być wspomagane przez HD?
- Kto będzie podejmował decyzje?
- Czy istnieją różnice w znaczeniu miar?
- Jaki jest wymagany poziom interakcji użytkowników?
- Jakie opóźnienia są akceptowalne?

Ograniczenia:

- Użyteczność dla użytkowników
- Koszty wdrożenia i zarządzania
- Zespół projektowy i administracyjny
- Techniczne protokoły i standardy
- Ciągły rozwój

Całkowity koszt posiadania hurtowni

- Planowanie i projektowanie
- Tworzenie i implementacja projektu
- Wdrożenie i zarządzanie
- Zakup i aktualizacja sprzętu
- Zakup lub tworzenie oprogramowania

Pytania

- W jakich warunkach OLTP nie wystarcza?
- Czy w hurtowni gromadzimy i przetwarzamy wszystkie dostępne dane źródłowe? Dlaczego?
- Czy w systemach OLAP ważna jest normalizacja? Dlaczego?
- Z czego wynika trudność projektowania hurtowni danych?

HD - architektura

- Kluczowe własności architektury hurtowni danych:
  - Separacja – analityczne i transakcyjne procesy powinny być niezależnie realizowane bazując na rozłącznych zbiorach danych
  - Skalowalność – architektura sprzętowa i programowa powinna być łatwo modyfikowalna w kontekście zmieniających się wymagań użytkowników

- Podstawa klasyfikacji:
  - liczba warstw
  - liczba ról

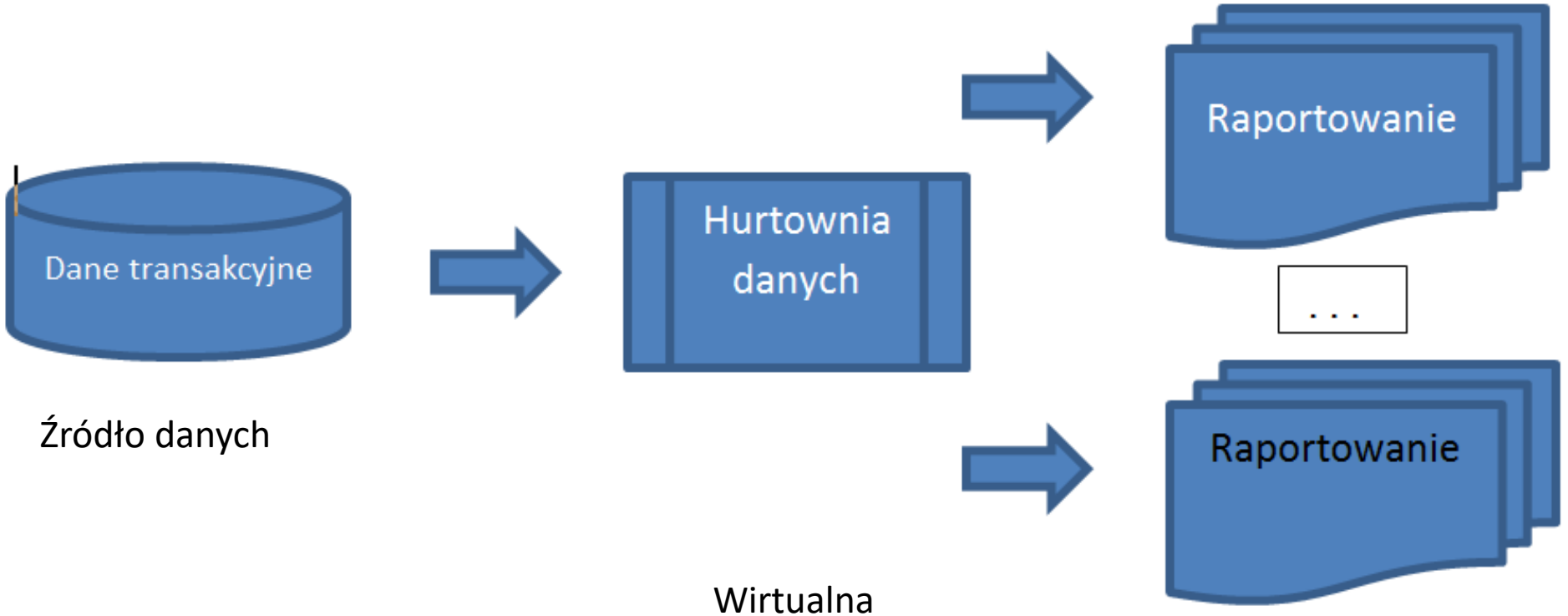
Architektura jednowarstwowa

- stosowana sporadycznie
- raporty ad hoc

- Zalety:
  - brak dodatkowych kosztów budowy i utrzymania HD

- Wady
  - trudności w tworzeniu zaawansowanych raportów
  - ograniczone możliwości analityczne

Architektura jednowarstwowa



Źródło danych

Wirtualna hurtownia danych

Analizy

Wirtualna hurtownia danych

- Implementowana jako widok danych transakcyjnych
- Brak separacji pomiędzy procesami transakcyjnymi i analitycznymi
- Zapytania tworzące informacje analityczne mają wpływ na regularnie realizowane operacje transakcyjne

Wymiar – kontekst analizy biznesowej

- Wymiar jest kontekstem analizy umożliwiając uzyskanie odpowiedzi na następujące kwestie:
  - Kto, co, gdzie, kiedy, dlaczego, jak?
  - Wymiary implementowane są z wykorzystaniem źródeł danych, które zawierają opisowe atrybuty specyficzne dla dziedziny problemowej (biznesu)
  - Atrybuty umożliwiają grupowanie i filtrowanie faktów

Fakt – podstawa oceny

- Fakt jest zdarzeniem, który podlega pomiarowi i jest charakteryzowany za pomocą zbioru miar i ich wartości
- Miary ilościowo charakteryzują zdarzenie w biznesie
- Rekord reprezentujący **fakt** w hurtowni danych **pozostaje w relacji z fizycznie zarejestrowanym zdarzeniem**
- Spójność z zadeklarowaną ziarnistością

Miary

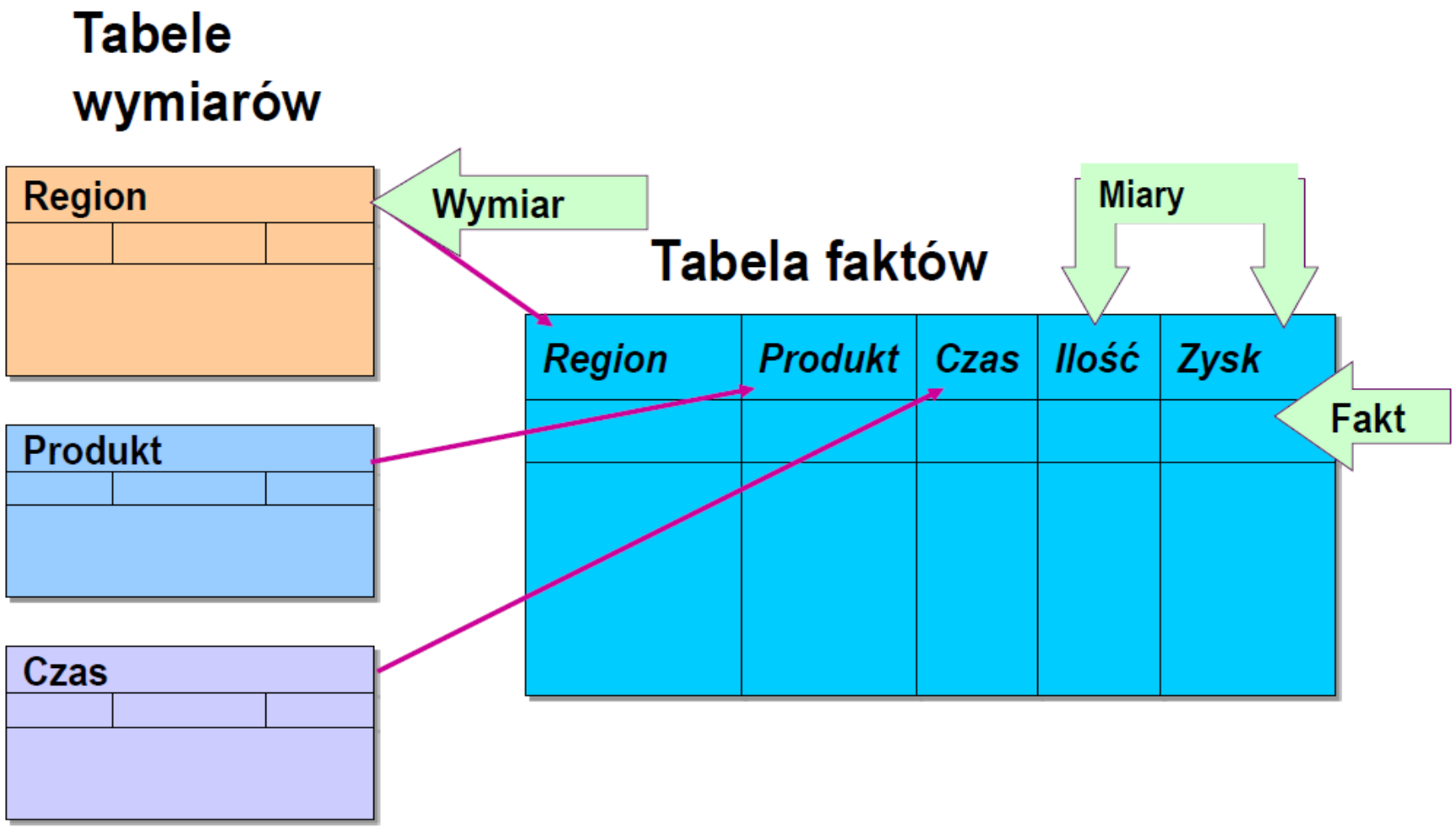
- wartości, które chcemy analizować (wartość sprzedaży, liczba pracowników, zadłużenie, zysk)
- wartości faktów świata rzeczywistego
- liczby (cecha addytywna lub semi-addytywna)
- podlegają ocenie wielowymiarowej

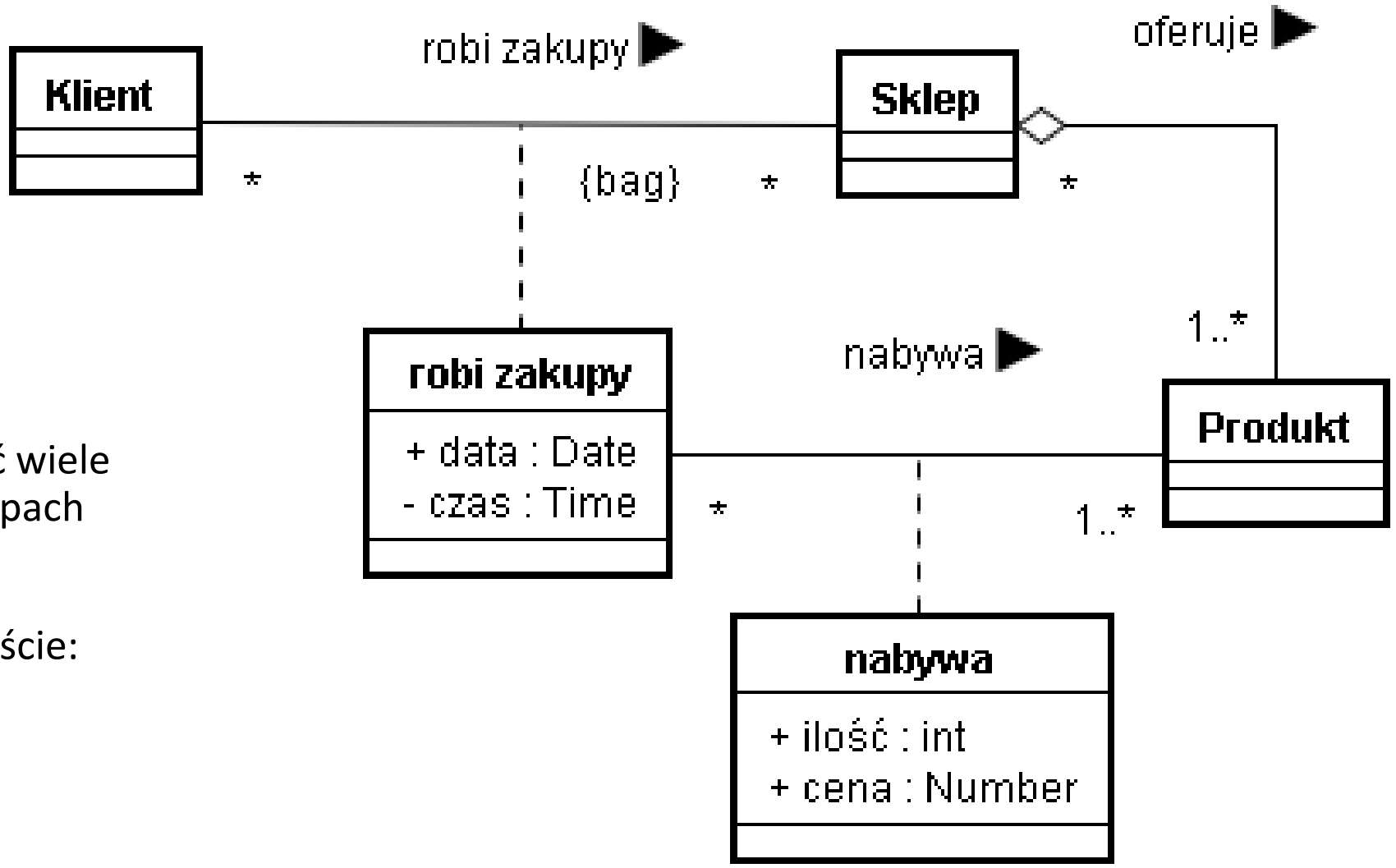
Wymiar

- Zbiór cech (atrybutów) istotnych z punktu widzenia analizy wartości faktów
- Wyznacza kontekst analizy wartości miar (region, produkt)
- Pozwala analizować informacje na różnych poziomach szczegółowości
- Ma charakter tekstowy, opisowy (klienci, regiony, daty).

Model wielowymiarowy

- Modele wielowymiarowe mogą być implementowane:
  - w bazach relacyjnych, lub
  - w bazach wielowymiarowych - analityczne modele OLAP (**O**n**l**ine **A**nalytical Processing Cube) - **kostki**
- Modele reprezentowane w bazie relacyjnej składają się z tabel **faktów** połączonych z tabelami **wymiarów** za pomocą kluczy obcych (tabele faktów) i kluczy głównych (tabele wymiarów)
- **Kostki** zawierają atrybuty wymiarów i faktów oraz wartości atrybutów na różnych poziomach agregacji

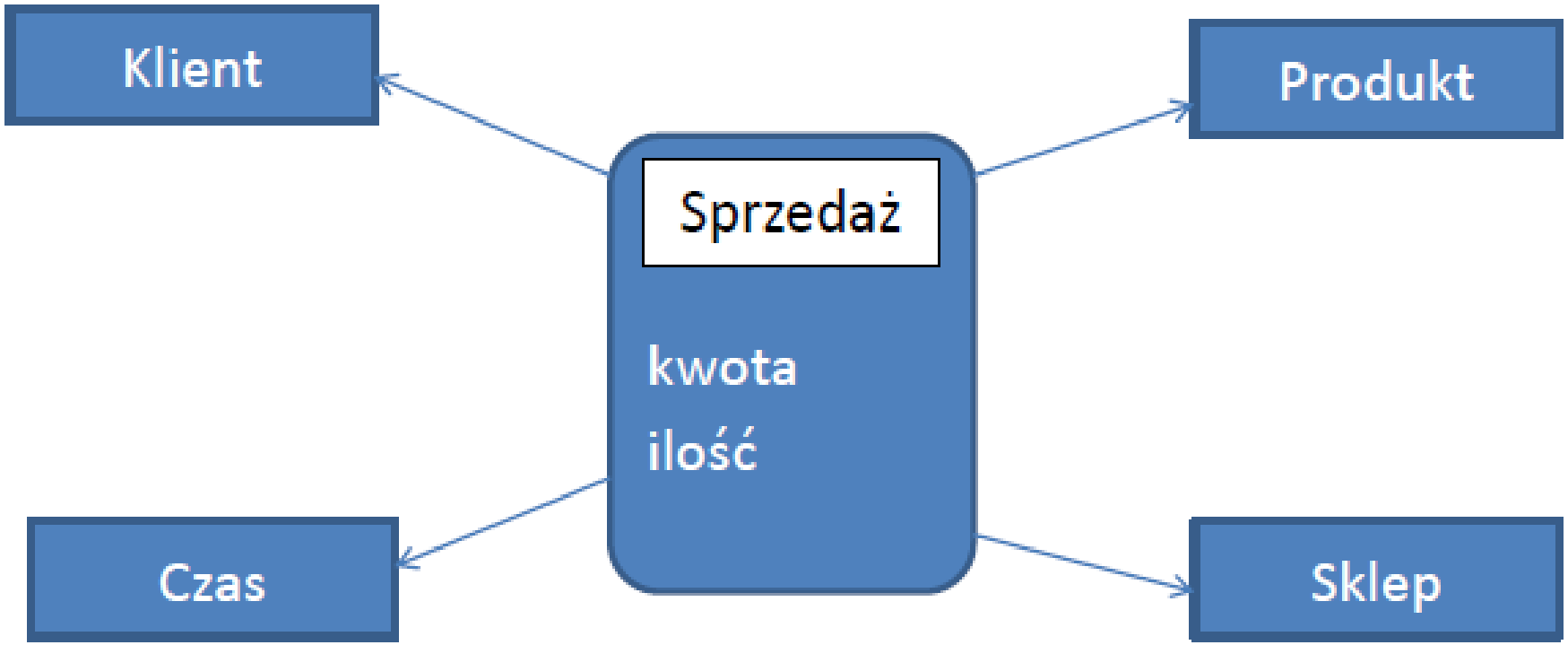

Tabele wymiarów
Region
Produkt
Czas
Wymiar
Miary
Tabela faktów
Region
Produkt
Czas
Ilość
Zysk
Fakt



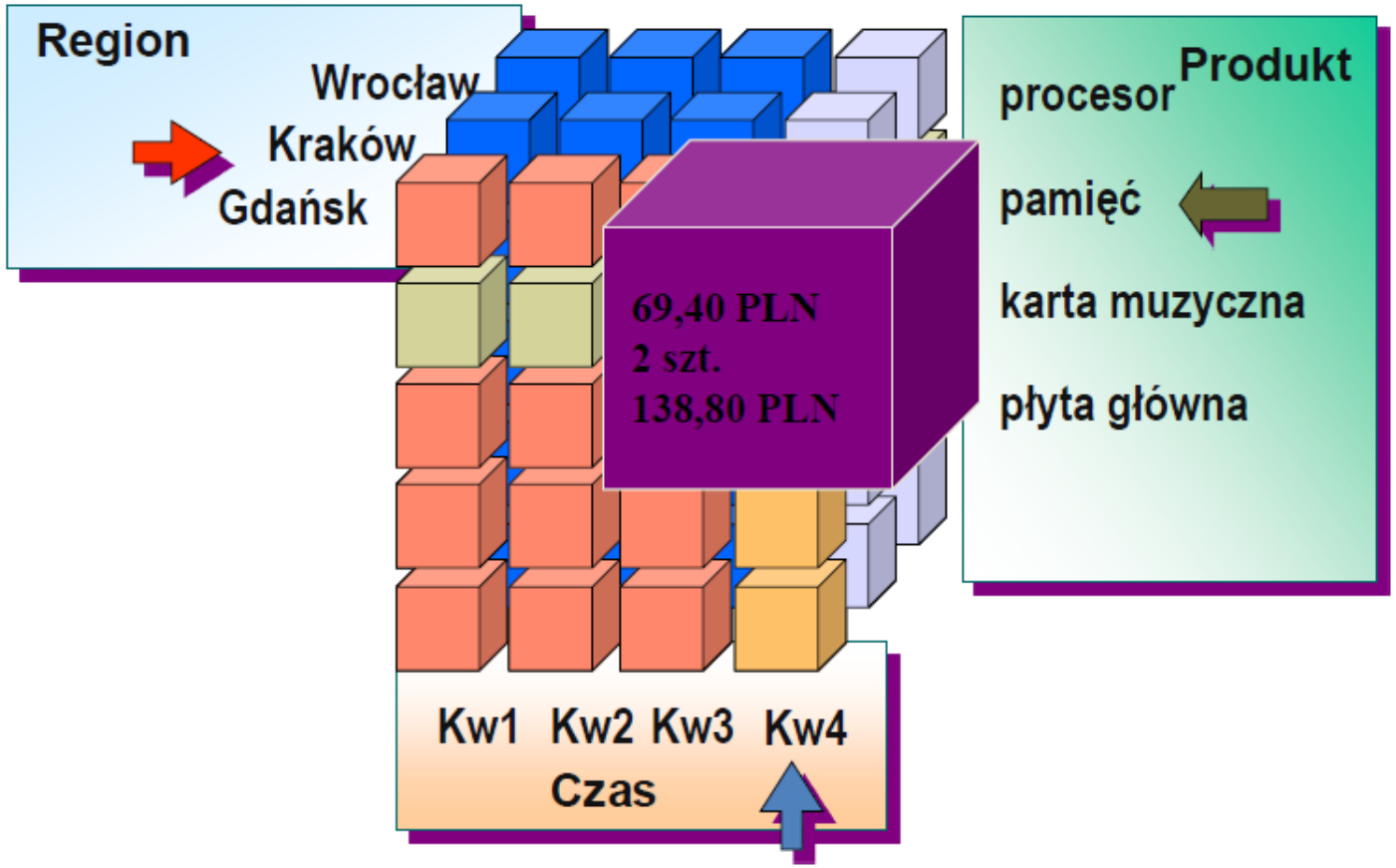
- wielu klientów może kupić wiele produktów w różnych sklepach

- ocena sprzedaży w kontekście:
  - klientów
  - sklepów
  - produktów
  - czasu

Modelowanie konceptualne



Przykład kostki



Hurtownie danych

Dziękuję za uwagę

dr inż. Bernadetta Maleszka