Hurtownie danych

Wielowymiarowy model danych - warstwa fizyczna

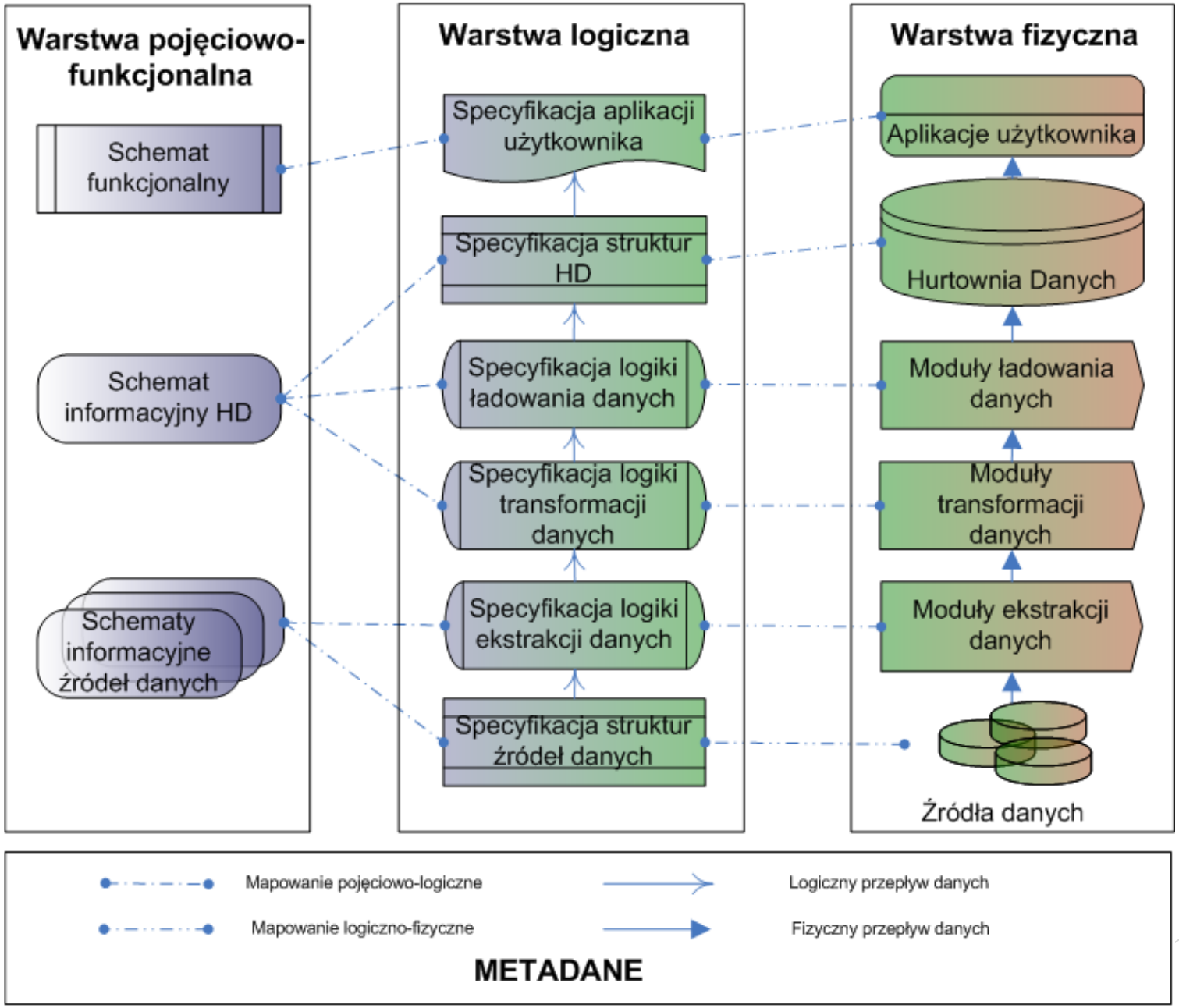
dr inż. Bernadetta Maleszka

Modelowanie hurtowni danych
przypomnienie

- Model biznesowy
  - Efekt analizy strategicznej
  - Identyfikacja miar i wymiarów dla poszczególnych procesów biznesowych
- Model logiczny (wymiarowy)
  - Model abstrakcyjny, konceptualny
  - Encje i atrybuty (reprezentowane w modelu relacyjnym jako tabele i powiązania między nimi)

Modelowanie hurtowni danych
przypomnienie

- Model fizyczny
  - Wybór sposobu składowania danych
  - Formaty danych
  - Strategie partycjonowania
  - Wybór indeksów
  - Wybór materializowanych perspektyw



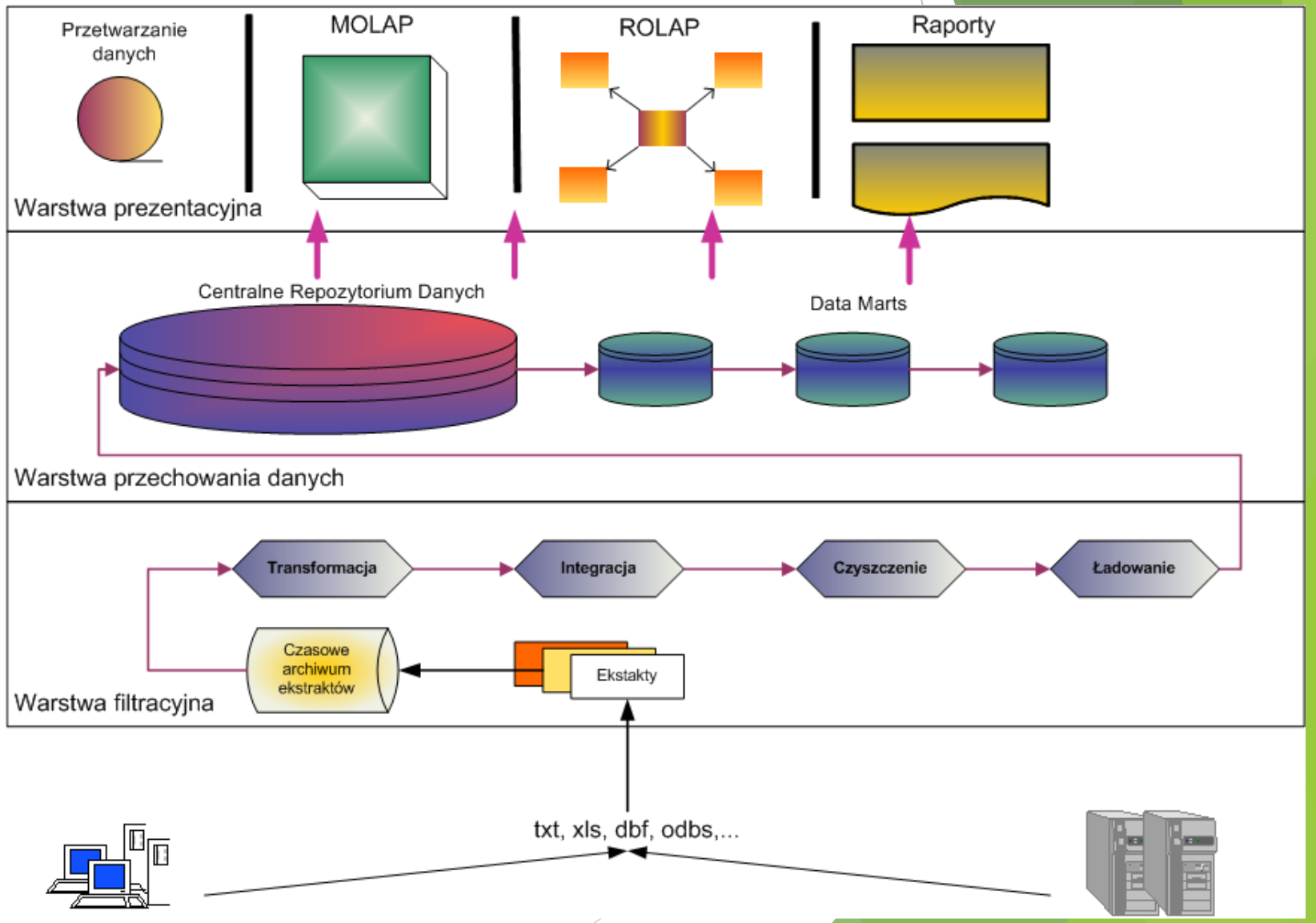
Opracowane na podstawie M.Pajer „Metody i technologie budowy…”

Fizyczny projekt i rozwój bazy danych

- projektowanie bazy danych
- identyfikacja kluczy
- przygotowanie strategii agregacji danych
- tworzenie strategii indeksowania
- przygotowanie strategii podziału (partycjonowania)
- planowanie pojemności (strategia gromadzenia)
- tworzenie obiektów bazy danych

Warstwa fizyczna

- aplikacje użytkownika
- przechowywanie danych
- filtracja danych
  - transformacja
  - ładowanie
- źródła danych
  - ekstrakcja



Ekstrakcja ze źródeł danych

- zintegrowany proces pozyskania danych z różnych źródeł
- dostępność danych
- wydajność procesu ekstrakcji (bez zakłócania funkcjonowania systemów źródłowych)
- weryfikacja poprawności danych
- automatyzacja procesu

Filtracja

- Odczyt danych z systemów źródłowych
  - identyfikacja nadejścia danych i weryfikacja ich źródła
  - weryfikacja kompletności danych
  - wczytanie do struktur pomocniczych
  - przechowywanie danych źródłowych do momentu zakończenia procesu ładowania
  - dokumentacja procedury (logi operacji, listy kontrolne)

- Transformacja
  - konwersja danych ze struktur tymczasowych (ang. TSA – Temporary Storage Area)
    - konwersja typów i wartości
    - kategoryzacja wartości
    - sprawdzenie poprawności

Filtracja

- Integracja
  - relacje pomiędzy danymi z różnych źródeł
  - agregacja danych
- Czyszczenie
  - sprawdzenie poprawności merytorycznej danych z różnych źródeł
  - reguły kontrolne, np. sumy krzyżowe, zgodność sum, kompletność relacji
  - reguły naprawcze dla wykrytych niespójności
- Ładowanie danych
  - przepisanie danych z TSA do stałych struktur HD – CRD (ang. Central Repository of Data)
  - działania sterowane metadanymi procesu

Przechowywanie danych - Centralne repozytorium danych

- Implementacja relacyjna
  - model gwiazdy, płatka śniegu, konstelacji faktów
  - perspektywy zmaterializowane
  - partycjonowanie danych

- Implementacja wielowymiarowa
  - wielowymiarowe kostki
  - selekcja i wycinanie (ang. slice & dice)
  - zwijanie (ang. roll-up, drill-up)
  - drążenie (ang. drill-down, drill-through)
  - obracanie (ang. pivoting)

Przechowywanie danych – Data Marts

- Hurtownie tematyczne – kopia wybranych danych z CRD
- Mniejsze dane -> lepsza wydajność, możliwość rozproszenia, częściej model wielowymiarowy

- Zakres wyodrębnionych danych definiowany przez potrzeby użytkowników:
  - sprawozdawcze
  - raportowe
  - monitorowanie poziomu realizacji planu
  - analityczne (analiza statystyczna lub eksploracja danych)
  - CRM
  - itp..

Widoki/Perspektywy zmaterializowane

- Problem:
  - Duża tabela faktów + mniejsze wymiary
  - Optymalizacja odczytu, a złączenia kosztowne
- Cel:
  - Przyspieszenie odczytu
- Rozwiązanie
  - Unikanie wielokrotnego wykonywania tych samych operacji
  - Widok zmaterializowany

Widok zmaterializowany

```
CREATE MATERIALIZED VIEW Nazwa_widoku AS
SELECT … FROM … WHERE Warunek
```

Zapytania do widoku:

```
SELECT … FROM Nazwa_widoku WHERE P1;
SELECT … FROM Nazwa_widoku WHERE P2;
...
SELECT … FROM Nazwa_widoku WHERE PN;
```

Widoki zmaterializowane

- Korzyści:
  - Dla zwykłego widoku – niejawne wykonanie dla każdego zapytania
  - Wyniki widoku zmaterializowanego przechowywane są fizycznie w systemie
  - Można użyć tabeli tymczasowej
- Ograniczenia:
  - Niebezpieczeństwo niespójności danych (opóźnienie dla automatycznego odświeżania)
  - Wyniki oparte na nieaktualnych danych
  - Czy w HD te aspekty są ważne?

Optymalizacja z wykorzystaniem widoków zmaterializowanych

- Perspektywa zmaterializowana przechowuje wyniki czasochłonnych zapytań analitycznych
- Wydajne, jeśli odpowiedź na zapytanie identyczne lub podobne do zapytania definiującego perspektywę
- Dane w tabelach źródłowych perspektywy nie ulegają modyfikacji

- Sposoby:
  - przepisywanie zapytań
  - indeksowanie

Indeksowanie

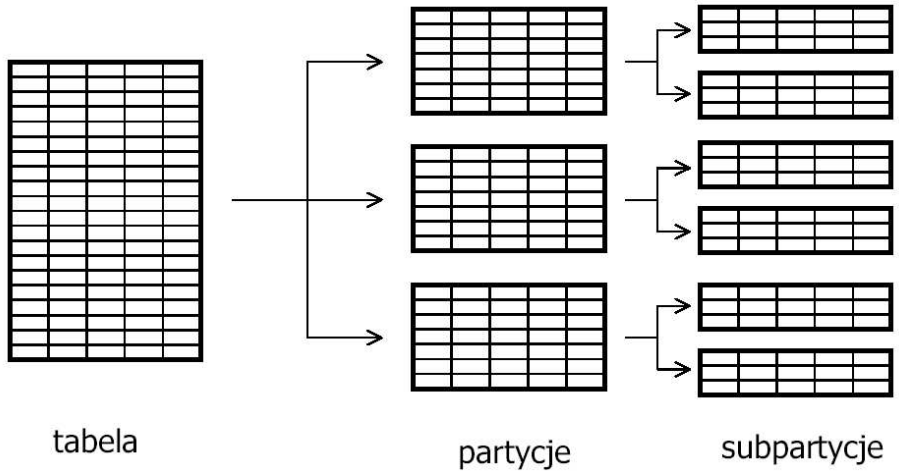
- Indeks połączeniowy (ang. join index)
  - łączy z sobą rekordy z różnych tabel posiadające tę samą wartość atrybutu połączeniowego
  - posiada strukturę B–drzewa zbudowanego na atrybucie połączeniowym tabeli
  - liście indeksu zawierają wspólne wartości atrybutu połączeniowego tabel wraz z listami adresów rekordów w każdej z łączonych tabel
- Indeks bitmapowy (ang. bitmap index)
  - wykorzystanie pojedynczych bitów do zapamiętania informacji o tym, że dana wartość atrybutu występuje w określonym rekordzie tabeli
  - mapa bitowa reprezentująca każdą unikalną wartość atrybutu
  - każdy bit mapy odpowiada jednemu rekordowi w tabeli
  - zbiór map bitowych dla danego atrybutu
  - B-drzewo z mapami bitowymi w liściach

Indeksowanie

- Bitmapowy indeks połączeniowy (ang. bitmap join index)
  - połączenie indeksu połączeniowego i bitmapowego
  - tworzenie tylu map bitowych ile jest wartości unikalnych atrybutu
  - każda mapa opisuje rekordy z tabeli
  - struktura B-drzewa, w którego liściach znajdują się mapy bitowe opisujące łączone rekordy

Partycjonowanie

- Fizyczny podział danych na niewielkie, łatwe w zarządzaniu podzbiory, nazywane partycjami
- Każda partycja stanowi odrębny segment w bazie danych
- Partycje mogą być opcjonalnie dzielone na subpartycje
- Partycjonowanie umożliwia równoległą realizację poleceń DML



Metody partycjonowania

Partycjonowanie zakresowe

- podział według przynależności wartości kolumny-klucza do predefiniowanych przedziałów

Partycjonowanie haszowe

- podział według wartości funkcji haszowej (modulo) wyliczanej dla kolumny-klucza

Partycjonowanie wg listy

- podział według przynależności wartości kolumny-klucza do predefiniowanych list wartości

Metody partycjonowania

Partycjonowanie dwupoziomowe zakresowo-haszowe

- rozdział rekordów na partycje wg zakresów, a następnie na subpartycje wg wartości funkcji haszowej

Partycjonowanie dwupoziomowe zakresowo-listowe

- rozdział rekordów na partycje wg zakresów, a następnie na subpartycje wg przynależności do list wartości

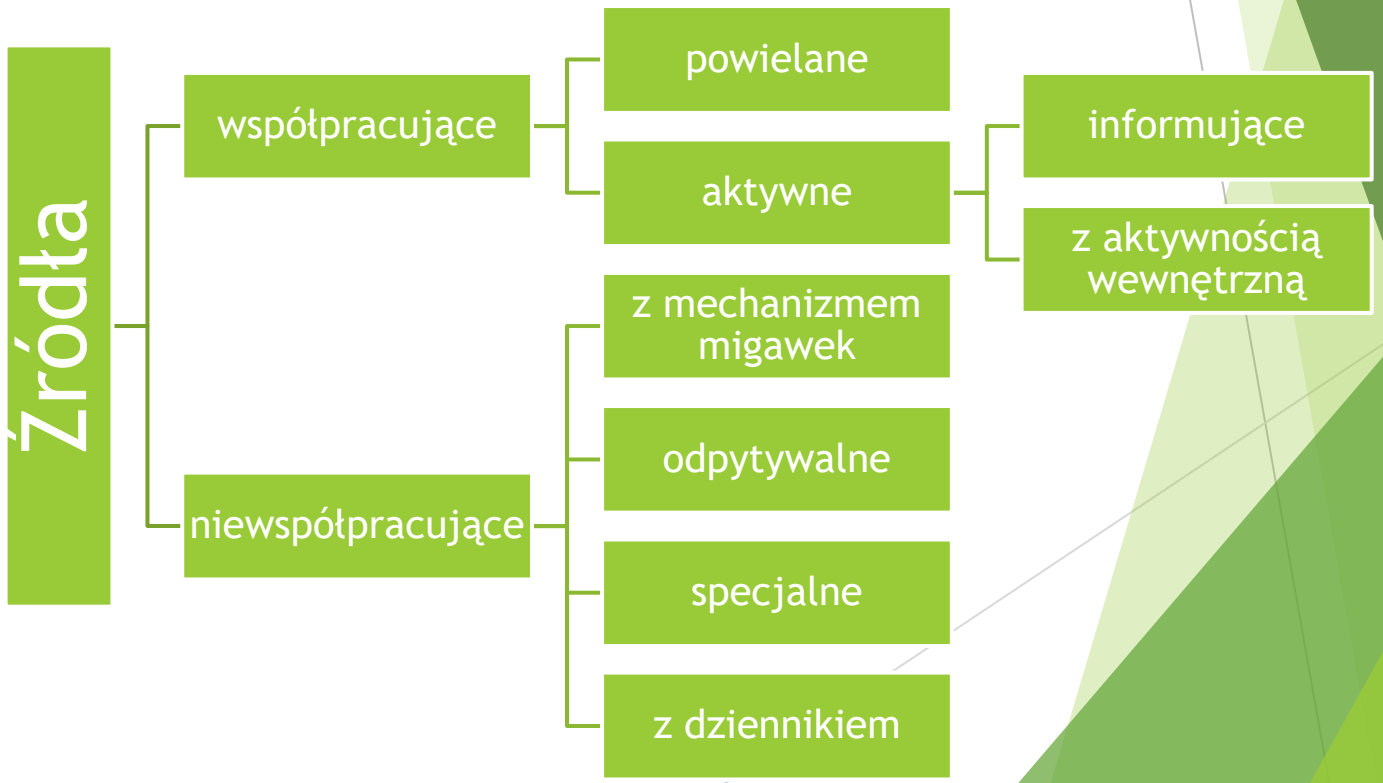
Zalety partycjonowania

- zrównoleglenie operacji dostępu do dysku, zapytań SQL do różnych partycji
- równoważenie obciążenie dysków
- przyspieszenie działania poprzez zapytania do konkretnej partycji
- bezpieczeństwo danych w razie awarii sprzętowych
- po awarii niezbędne odtworzenie partycji, a nie całej tabeli

- Wady?

Odświeżanie danych

- Zapewnienie zgodności danych w hurtowni z danymi źródłowymi
- Wykrywanie zmian w danych źródłowych:
  - monitor zmian
- Klasyfikacja źródeł danych:



Propagacja aktualizacji

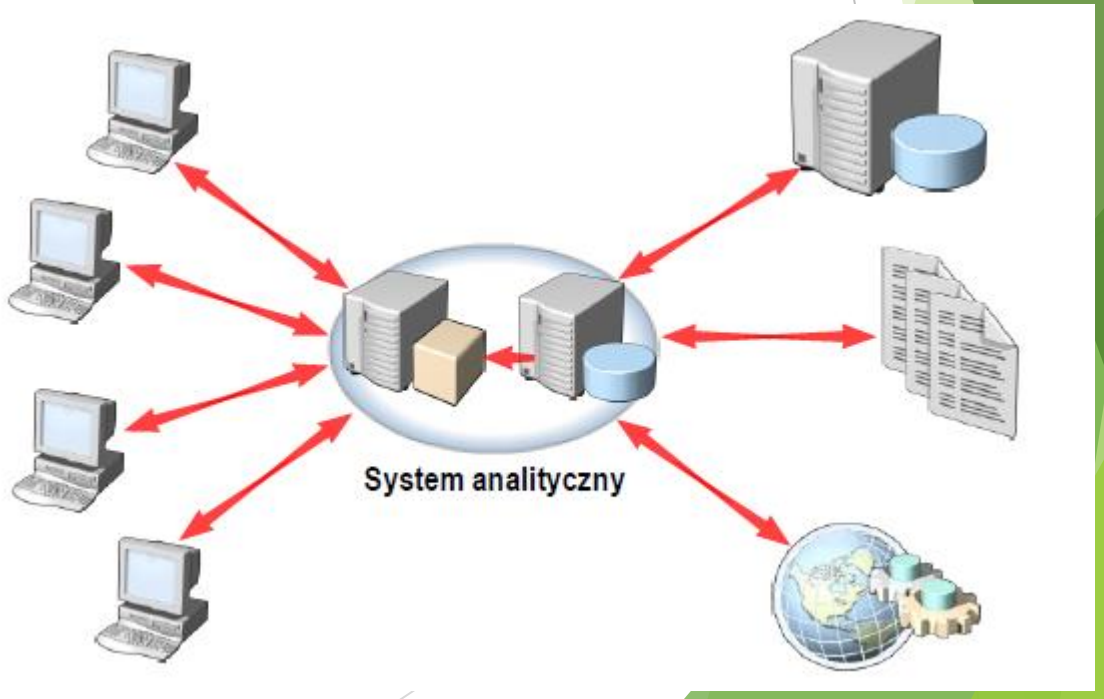
- Nanoszenie zmian na wszelkie potworzone materializacje np. agregacje, hurtownie tematyczne, czy kostki OLAP
- Strategie:
  - Aktualizacja opóźniona (na żądanie, przy pierwszym użyciu po zmianie danych w hurtowni):
    - dłużej trwa pierwsze zapytanie,
    - nie musimy odświeżać tych perspektyw, których nie użyjemy.
  - Aktualizacja natychmiastowa (podczas odświeżania hurtowni):
    - dłużej trwa wsadowe przetwarzanie procesu aktualizacji,
    - przerzucamy kosztowne procesy na godziny nocne,
    - część aktualizacji może okazać się zbędna.

Prezentacja

- Mechanizmy dostępu i uprawnień użytkowników
- Wybór najlepszego rodzaju OLAP:
  - ROLAP
  - MOLAP
  - HOLAP
- Interfejs
- Raporty i zestawienia statyczne i dynamiczne

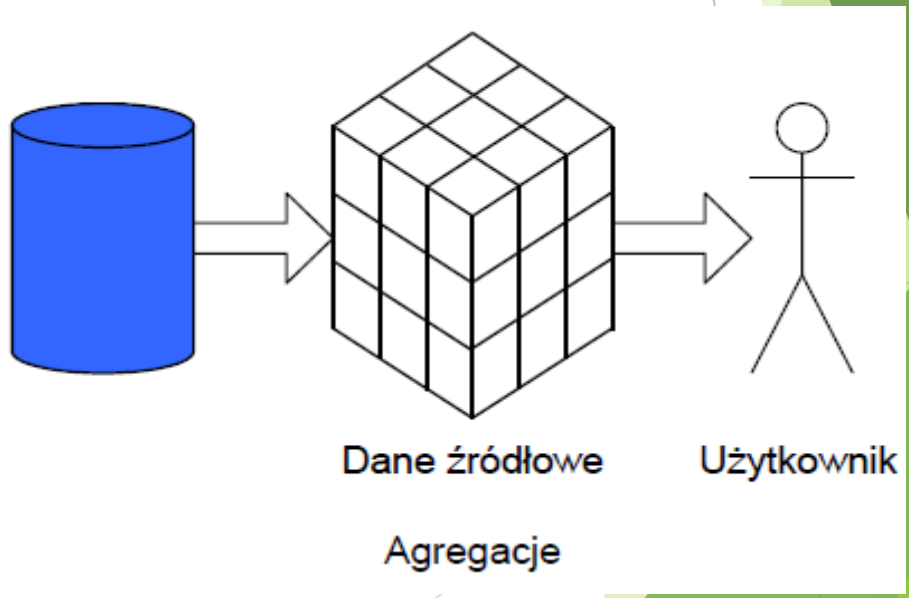
Typy OLAP

- On-Line Analytical Processing – przetwarzanie informacji z wykorzystaniem wielowymiarowej bazy danych
- bezpośrednie przetwarzanie analityczne
  - wielowymiarowe
    (MOLAP, ang. multidimensional OLAP)
  - relacyjne
    (ROLAP, ang. relational OLAP)
  - hybrydowe
    (HOLAP, ang. hybrid OLAP)



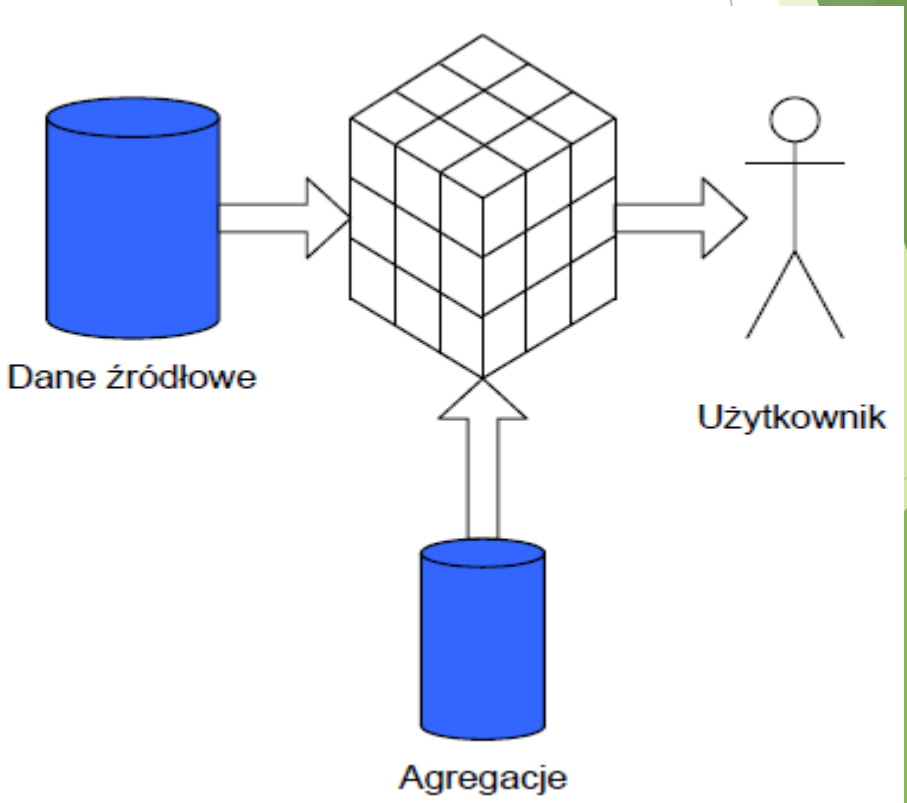
MOLAP

- wielowymiarowe OLAP
- wielowymiarowy format danych i agregacji
- zapytania wykonują się szybko
- wymaga największej przestrzeni na dysku



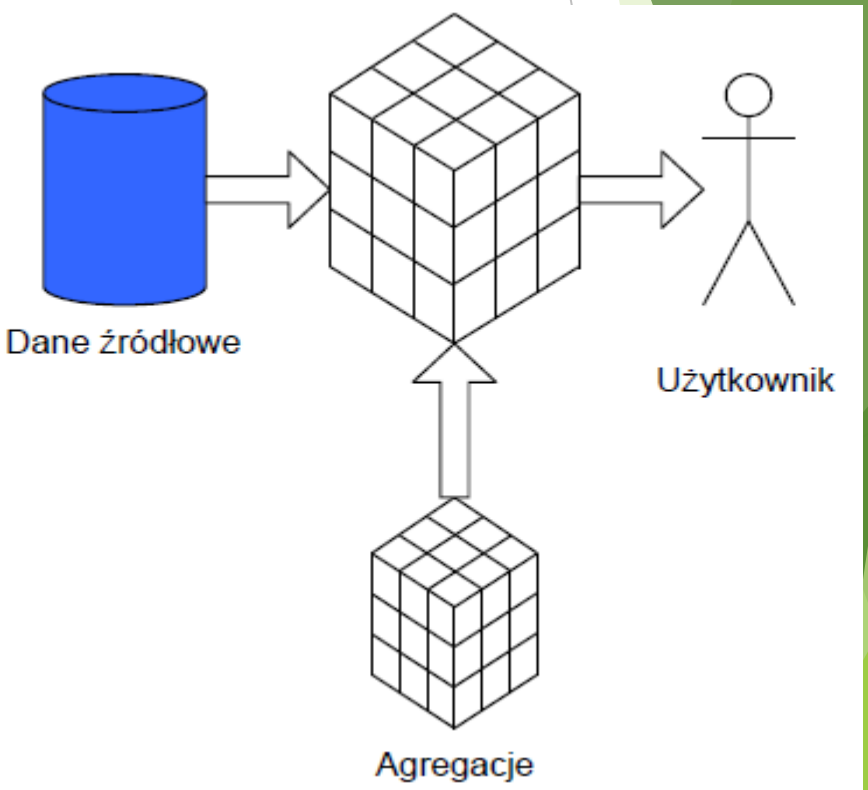
ROLAP

- dane i agregacje są przechowywane w RSZBD
- najwolniejsze odpowiedzi na zapytania
- zwykle najwolniejsze przetwarzanie
- można tworzyć indeksowane widoki
- można tworzyć dodatkowe tabele do przechowywania agregacji podczas ETL
- najbardziej użyteczny tryb dla dużej liczby danych
- wspomaga rozwiązania real-time OLAP

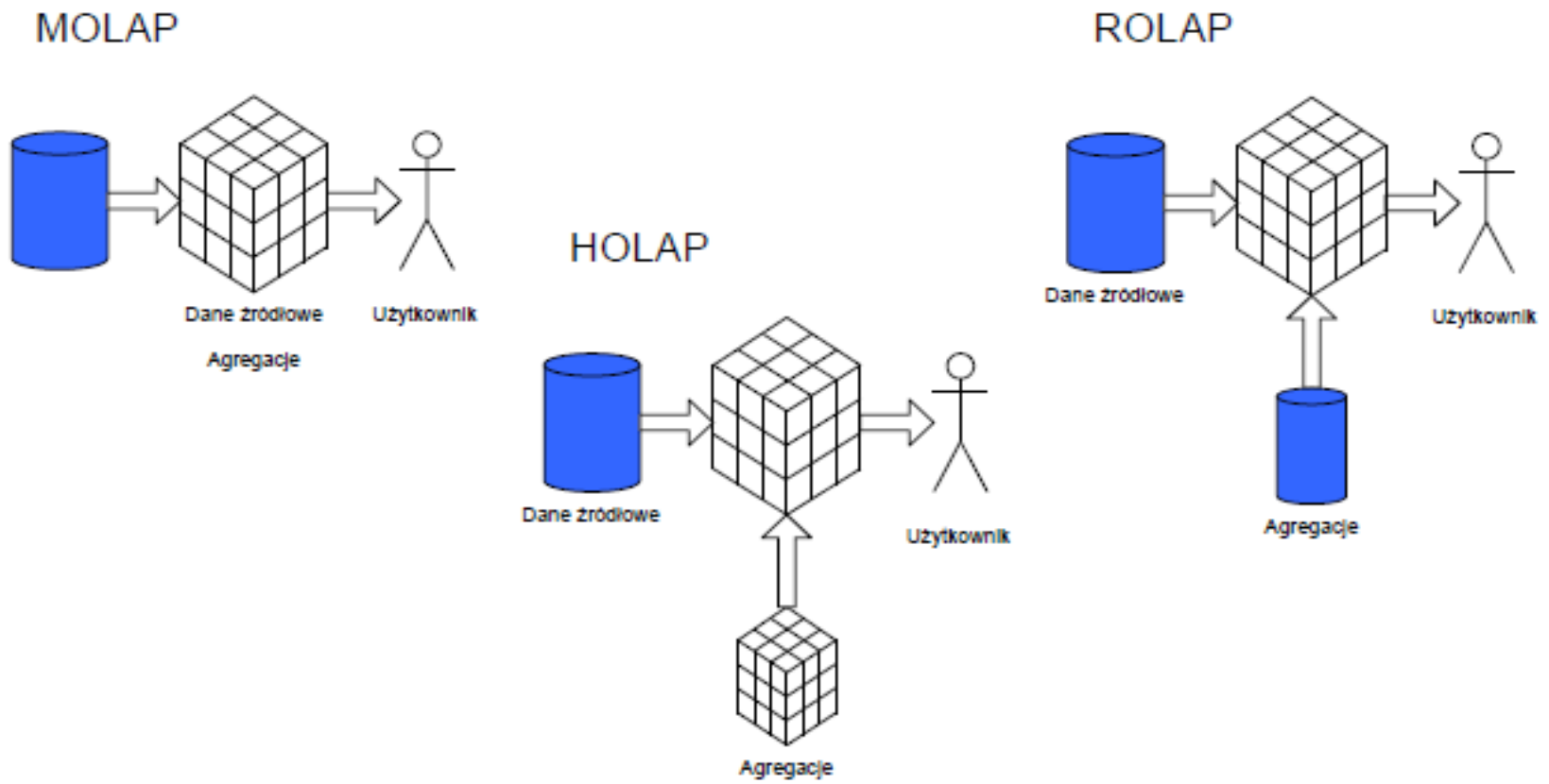


HOLAP

- dane zarządzane przez RSZDB
- agregacje tworzone w formacie wielowymiarowym
- dobry wybór, gdy przestrzeń na dysku jest wąskim gardem
- dobra wydajność dla częstych odwołań do agregacji



Typy OLAP


MOLAP
Dane źródłowe
Użytkownik
Agregacje
HOLAP
Dane źródłowe
Użytkownik
Agregacje
ROLAP
Dane źródłowe
Użytkownik
Agregacje

Typy systemów OLAP – podsumowanie

Typ	MOLAP	ROLAP	HOLAP
Miejsce przechowywania danych	baza wielowymiarowa	baza relacyjna	baza relacyjna
Miejsce przechowywania agregacji	baza wielowymiarowa	baza relacyjna	baza wielowymiarowa
Czas przetwarzania (*)	krótki	bardzo długi	krótki
Wymagane miejsce (*) (rozmiar)	średnio	dużo	mało
Opóźnienie w odświeżaniu danych (*)	wysokie (wymagane procesowanie kostki)	niskie	średnie
Czas odpowiedzi na zapytanie (*)	krótki	długi	średni

(*) w porównaniu do pozostałych typów

Inne typy OLAP

- Desktop OLAP (DOLAP) – systemy niewielkiej, „osobistej” skali
- Real-time OLAP (RTOLAP) – systemy czasu rzeczywistego
- Web-based OLAP (WOLAP) – systemy dostępne w publicznej sieci

Hurtownie danych

Dziękuję za uwagę

dr inż. Bernadetta Maleszka