BAZY DANYCH

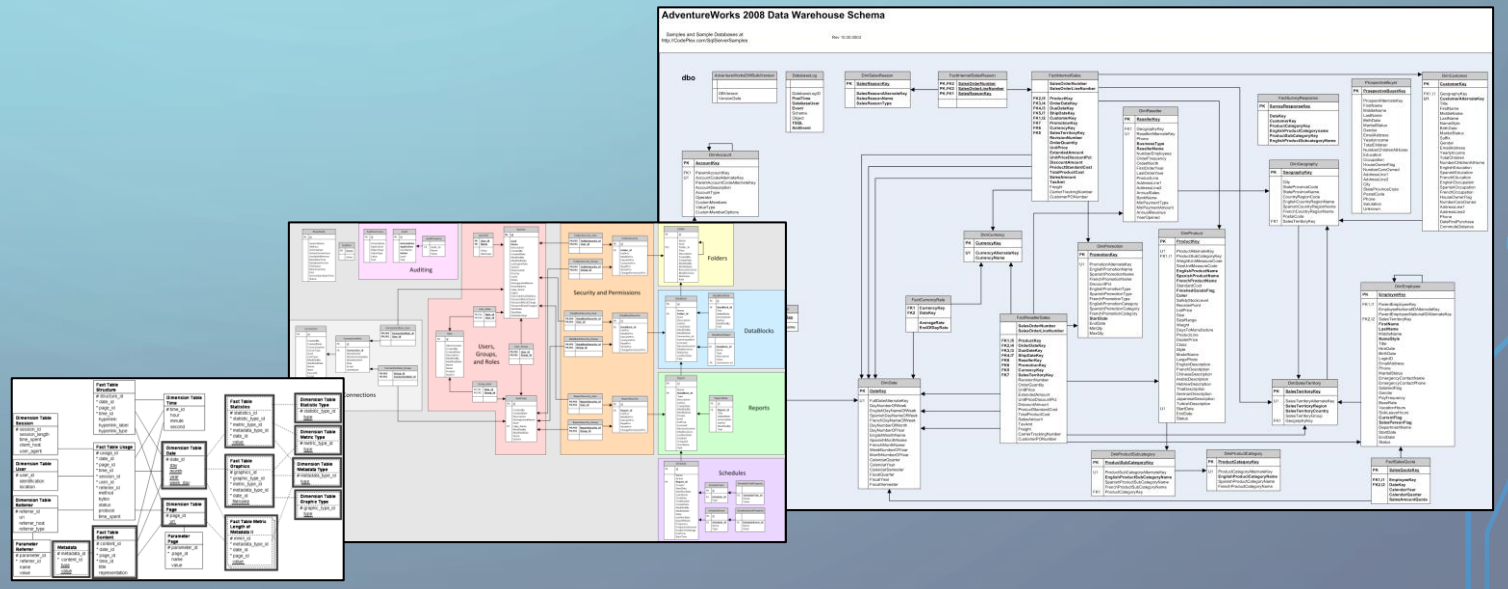
WYKŁAD IV

HURTOWNIE DANYCH

MOTYWACJA

Duża liczba rekordów: 10^6-10^{12} w przypadku baz danych o obiektach niebieskich,

Duża liczba atrybutów (cechy, pomiary, kolumny): Setki zmiennych opisujących medyczne pomiary pacjentów.



MOTYWACJA

Jak zdobyć użyteczną wiedzę (informację) z dużych baz danych?

- ✓ Hurtownie danych - zbieranie danych (w czasie rzeczywistym)
- ✓ OLAP - przetwarzanie analityczne
- ✓ Data mining - odkrywanie interesującej wiedzy (reguł, regularności, wzorców, modeli) z dużych zbiorów danych.

BUSINESS INTELLIGENCE

Proces przekształcania danych w informacje, a informacji w wiedzę, w celu wspomagania procesu podejmowania decyzji biznesowych. Dziedzina ta obejmuje aplikacje i technologie służące gromadzeniu i analizie danych.

BUSINESS INTELLIGENCE

- **DW** - Hurtownie danych – ładowanie, przetwarzanie
- **Data mining** - Eksploracja danych, drążenie danych
- **OLAP** - Online Analytical Processing
- Czyszczenie danych i zarządzanie jakością danych
- **MIS** (Management Information Systems) - Systemy Informowania Kierownictwa
- Raportowanie - Wizualizacja informacji i panele dla kierownictwa
- Prognozowanie, finanse i budżetowanie
- Statystyki i techniczna analiza danych
- **CRM** (Customer Relationship Management) – Zarządzanie Relacjami z Klientami
- **DSS** (Decision Support Systems) – systemy wspomagania decyzji

HURTOWNIA DANYCH

Hurtownia danych (ang. *data warehouse*), według twórcy tej koncepcji W.H. Inmona, to zintegrowany, tematyczny, zmienny w czasie zbiór danych dla wspomagania procesów podejmowania decyzji zarządczych (rok 1996).

HURTOWNIA DANYCH

Podstawowe cele, z powodu których buduje się hurtownie danych, to:

- Przetwarzanie analityczne danych (On-Line Analytical Processing, OLAP) – odpowiednie kwerendy (zapytania SQL) pozwalają na wykonywanie zestawień statystycznych, wykresów i raportów, podsumowujących znaczne ilości danych.
- Wspomaganie decyzji (Decision Support, DS) - wykonywanie bardziej złożonych analiz, symulacji scenariuszy biznesowych itd. Wspomaganie podejmowania decyzji można łączyć również z bardziej zaawansowanymi i zautomatyzowanymi projektami (Knowledge Discovery in Databases, Business Intelligence).
- Centralizacja danych - gromadzenie szczegółowych danych napływających z różnych źródeł, często związanych z bazami OLTP, często przetwarzanych i integrowanych przy użyciu narzędzi Extract Transform Load, celem udostępniania szerokiego zakresu danych dla poszczególnych hurtowni tematycznych, narzędzi OLAP czy też narzędzi Data Mining. W takim rozumieniu, hurtownia danych jest centralnym punktem dla infrastruktury danych w przedsiębiorstwie czy zastosowaniu, zwanej Corporate Information Factory.
- Archiwizacja - wykonywana ze względu na wymagania prawne (niektóre instytucje zobowiązane są do przechowywania pewnych danych), gdzie szybki dostęp do danych poprzez SQL jest jednak wciąż ważny.

HURTOWNIA DANYCH

W celu zapewnienia użytkownikom, a przede wszystkim decydentom szybkiego dostępu do danych hurtownia danych powinna spełniać kilka warunków:

- dane muszą być gromadzone w taki sposób, by istota informacji nie została utracona (chociażby ze względów biznesowych),
- dostęp do informacji powinien odbywać się na przejrzystych zasadach, zaś zasilanie systemu aktualizowanymi danymi musi odbywać się w czasie, który nie podważy sensu istnienia systemu wspomagającego decyzję,
- czas oczekiwania na dane powinien być krótszy od czasu, niezbędnego do podjęcia decyzji,
- użytkownicy hurtowni danych powinni mieć możliwość samodzielnego przeprowadzania złożonych analiz bez pomocy informatyków,
- koszt systemu nie może być większy od zysków jakie są osiągane w wyniku jego wdrożenia.

HURTOWNIA DANYCH

Hurtownia danych to dedykowany systemem baz danych, który w odróżnieniu od systemu transakcyjnego (bazującego na danych operacyjnych) charakteryzuje się następującymi cechami:

- dane są utrzymywane w dużo dłuższym horyzoncie czasowym,
- utrzymywanie danych jest optymalizowane pod kątem odpowiadania na złożone zapytania pochodzące od analityków i zarządzających,
- dane są pozyskiwane z różnorodnych źródeł przy czym zapewniona jest ich jednolitość.

HURTOWNIA DANYCH A BAZA DANYCH

Hurtownie danych i bazy danych (mimo zarządzania przez serwer SQL) odróżnia:

- Przeznaczenie - bazy relacyjne służą do informowania o bieżącej sytuacji, hurtownie danych pomagają w planowaniu. Na przykład program służący do sprzedaży w firmie będzie korzystał z relacyjnej bazy danych przetwarzając na bieżąco aktualne dane, natomiast hurtownia danych będzie wykorzystywana do analizy sprzedaży w jakimś przedziale czasu.
- Sposób aktualizacji - w bazach relacyjnych to użytkownicy na bieżąco modyfikują dane, hurtownie danych są aktualizowane automatycznie.
- Bazy relacyjne zawierają tylko aktualną wersję danych np. bieżący stan konta, natomiast w hurtowniach danych zapisywana jest cała historia zmian.
- Poziom szczegółowości odczytywanych danych. W bazach danych zapisywane są poszczególne operacje, użytkowników hurtowni danych interesują najczęściej ogólne dane np. sprzedaż towaru X w ciągu ostatniego roku.

HURTOWNIA DANYCH A BAZA DANYCH

Bazy danych a hurtownie danych

CECHA	APLIKACJE OPERACYJNE	HURTOWNIE DANYCH
ZASILANIE DANYMI	RÓWNOMIERNY NAPŁYW WIELU KRÓTKICH TRANSAKCJI	POJEDYNCZE, BARDZO DUŻE TRANSAKCJE PODCZAS ŁADOWANIA DANYCH
WZORZEC DOSTĘPU DO DANYCH	DOBRZE OKREŚLONY PRZEZ UŻYTKOWNIKA	MAŁO PRZEWIDYWALNY, ZALEŻNY OD BIEŻĄCYCH POTRZEB UŻYTKOWNIKA
TRYB DOSTĘPU DO DANYCH PRZEZ UŻYTKOWNIKÓW	ZAPIS, ODCZYT, MODYFIKACJA	TYLKO ODCZYT. Z REGUŁY DUZE ILOŚCI DANYCH W JEDNEJ KWERENDZIE
HORYZONT CZASOWY PRZECHOWYWANIA DANYCH	DANE BIEŻĄCE	PEŁNA HISTORIA

HURTOWNIA DANYCH

ZALETY:

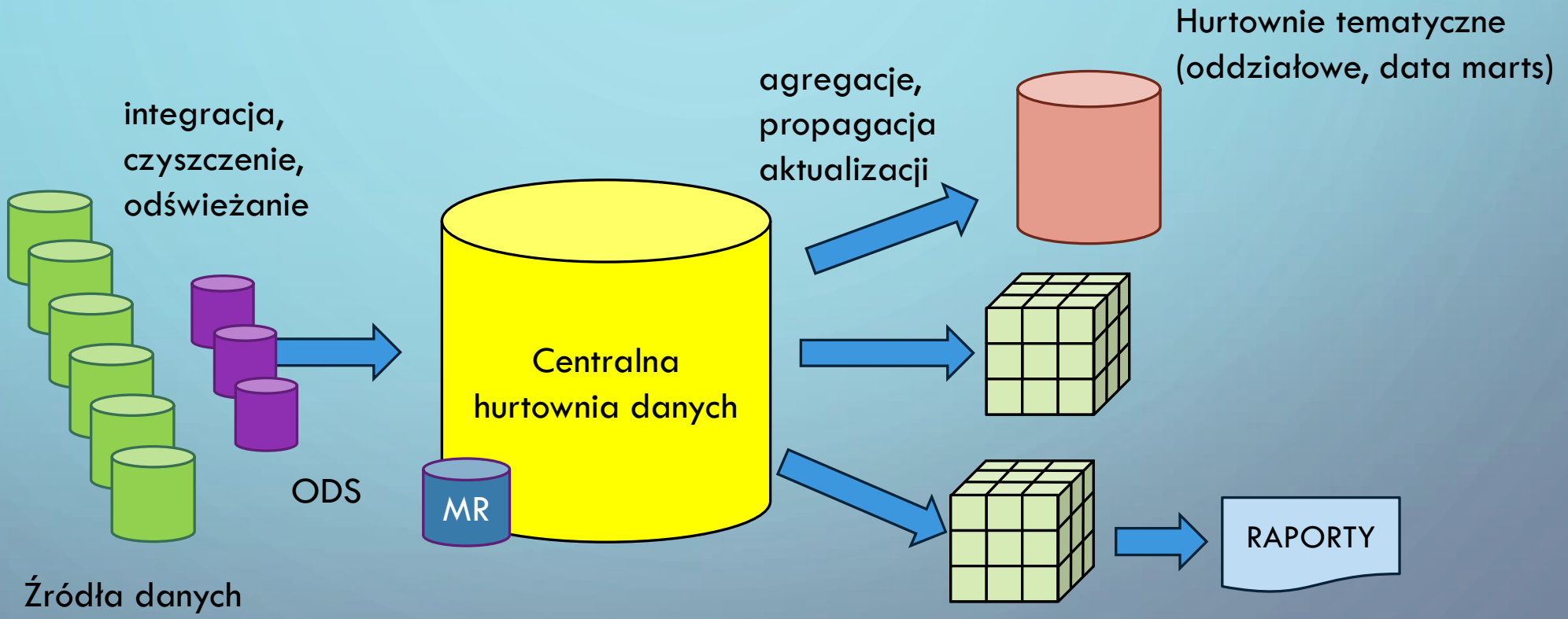
- Dane zintegrowane (wspólna struktura i wartości),
- Szybkość dostępu do danych,
- Niezależność od awarii źródeł.

WADY:

- Redundancja danych,
- Odświeżanie danych.

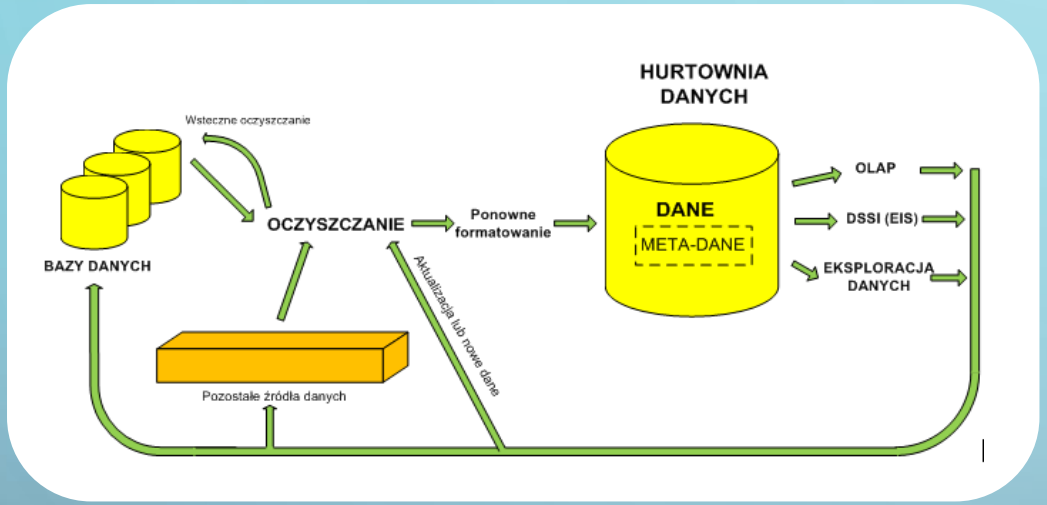
HURTOWNIA DANYCH

Architektura hurtowni danych:



HURTOWNIA DANYCH

Cykl życia danych w hurtowni:



DANE Z APLIKACJI OPERACYJNYCH ORAZ DANE ARCHIWALNE

EKSTRAKCJA

CZYSZCZENIE

TRANSFORMACJA

MAGAZYNOWANIE

AGREGACJA

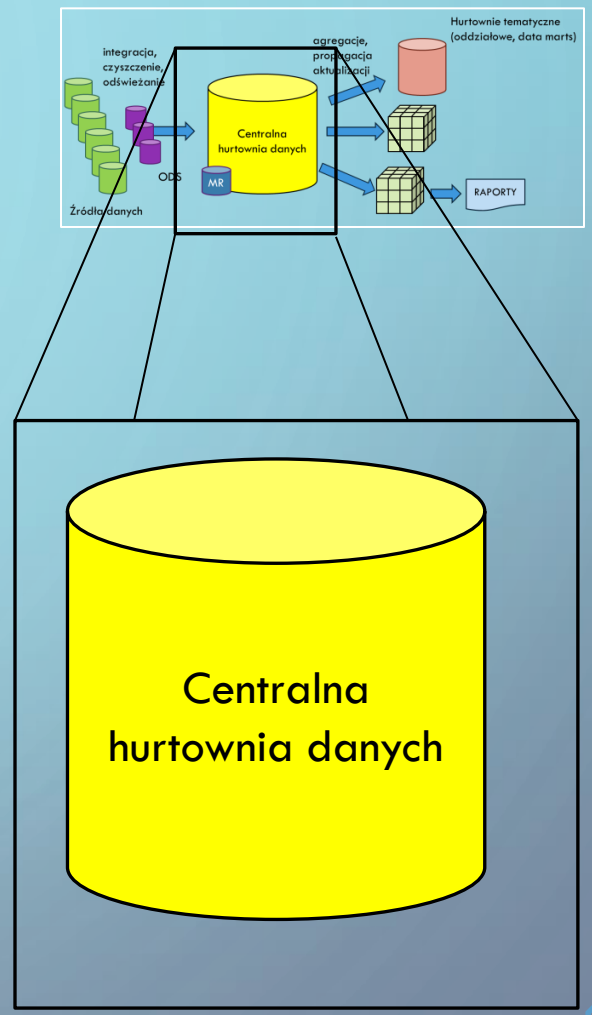
HURTOWNIA DANYCH

DOSTĘP DO ZGROMADZONYCH DANYCH

HURTOWNIA DANYCH

Elementy hurtowni danych

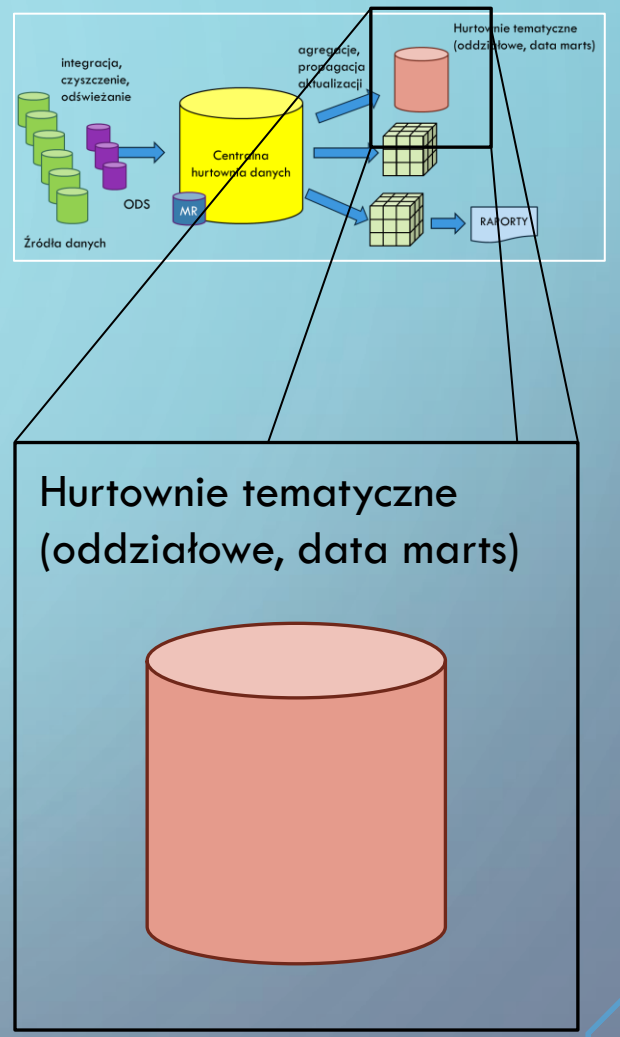
Centralna hurtownia danych (podstawowa, korporacyjna). stanowi podstawowe miejsce przechowywania nieulotnej informacji gromadzonej ze źródeł, jak też częściowych podsumowań przydatnych w zadaniach typu OLAP i we wspomaganiu decyzji. Rejestruje historię źródła danych i jest cyklicznie (w czasie aktualizacji) uzupełniana o nowe informacje dotyczące aktualnego stanu źródła danych. Hurtownia ta spełnia także funkcje archiwalne.



HURTOWNIA DANYCH

Elementy hurtowni danych

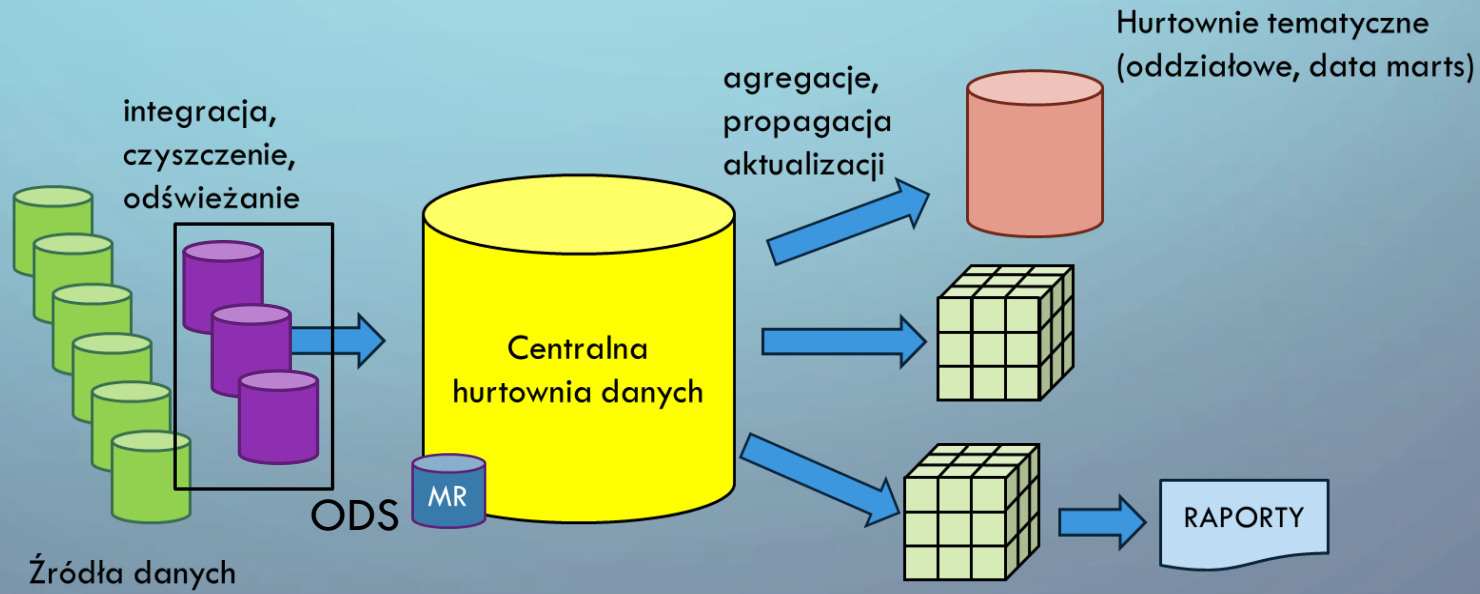
Hurtownie lokalne tworzone są na potrzeby użytkowników (działów analitycznych) i zawierają wyselekcjonowane dane w postaci silnie zagregowanej. Dzięki nim możliwa jest szybka prezentacja podsumowań wykorzystywanych w zarządzaniu, planowaniu długoterminowym, analizach historycznych, analizach trendów, przetwarzaniu informacji i analizach zintegrowanych. Lokalne hurtownie danych nazywane są hurtowniami tematycznymi (data marts, hurtownie oddziałowe). Ze względu na mniejszy rozmiar i możliwość pracy lokalnej, hurtownie tematyczne pozwalają na sprawniejsze operowanie danymi. Mogą być zaimplementowane jako relacyjne bazy danych lub specjalne struktury wielowymiarowe.



HURTOWNIA DANYCH

Elementy hurtowni danych

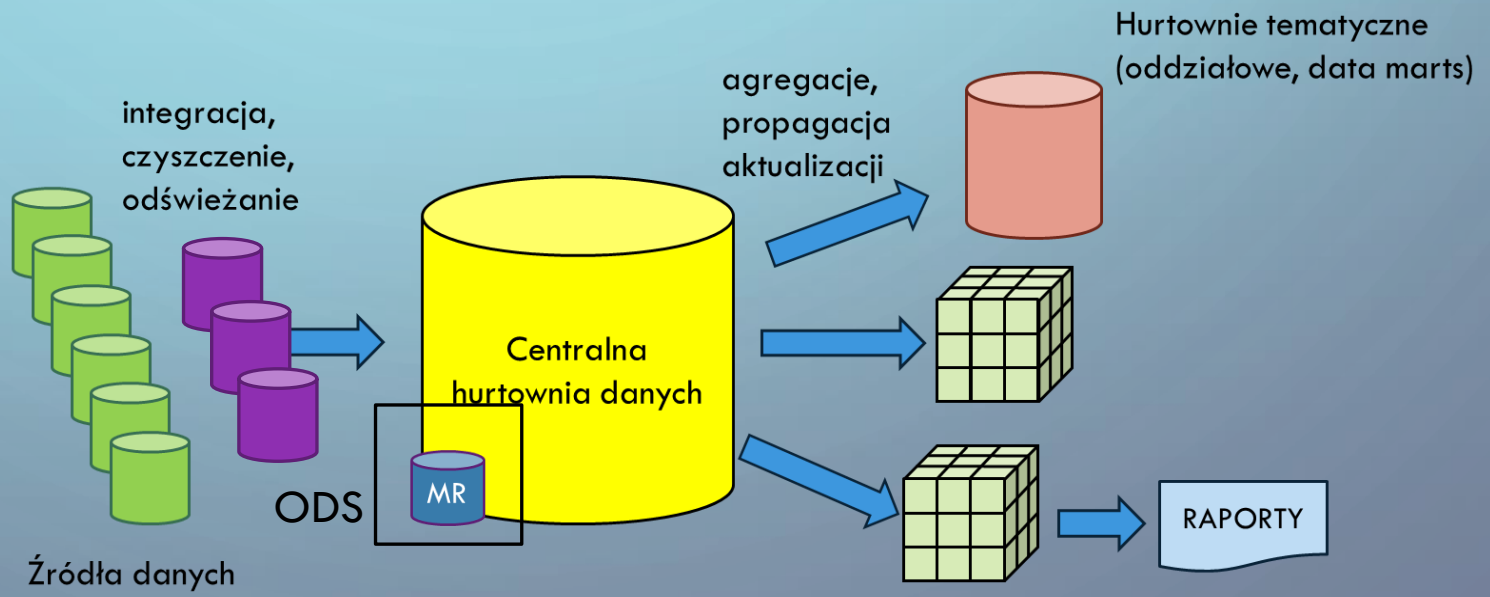
Zdarza się, iż między warstwą danych źródłowych, a globalną hurtownią danych wprowadza się warstwę pośrednią, zwaną magazynem danych operacyjnych (operational data store, ODS). Warstwa ta przechowuje zwykle wyniki transformacji, integracji i agregacji danych ze źródeł i sama stanowi bezpośrednie źródło zasilające globalną hurtownię danych.



HURTOWNIA DANYCH

Elementy hurtowni danych

Dodatkowym i bardzo ważnym elementem systemu hurtowni danych jest baza metadanych (metadata repository). W założeniu ta struktura ma przechowywać aktualny i historyczny schemat fizyczny, logiczny i pojęciowy hurtowni, w tym procesów ekstrakcji, transformacji, agregacji, czyszczenia i przechowywania informacji, a także historię użycia danych.



PROJEKTOWANIE HURTOWNIA DANYCH

Projektowanie hurtowni danych polega na stworzeniu modelu pojęciowego, logicznego i fizycznego hurtowni. Modelowanie na tych trzech poziomach dotyczy wszystkich elementów hurtowni danych - centralnej hurtowni, procesów ETL, hurtowni tematycznych itp. Poziomy te można scharakteryzować następująco:

- Model pojęciowy – to opis struktury, zawartości i przeznaczenia hurtowni danych przeprowadzony na poziomie pojęciowym,
- Model logiczny - to opis odwołujący się do elementów logicznych baz danych i procesów hurtowni,
- Model fizyczny to opis parametrów mających na celu optymalizację działania hurtowni danych.

PROJEKTOWANIE HURTOWNIA DANYCH

Z punktu widzenia przyjętej metody postępowania, wyróżniamy:

- projektowanie wstępujące (od szczegółu do ogółu), w ramach którego najpierw tworzone są projekty związane z poszczególnymi źródłami danych, działami przedsiębiorstwa, potrzebami użytkowników itp., a następnie projekty te scalane są w jeden projekt ogólny;
- projektowanie zstępujące, w ramach którego rozpoczynamy od stworzenia modelu przedsiębiorstwa na poziomie pojęciowym, by następnie stopniowo przejść do projektu integracji potrzebnych danych źródłowych.

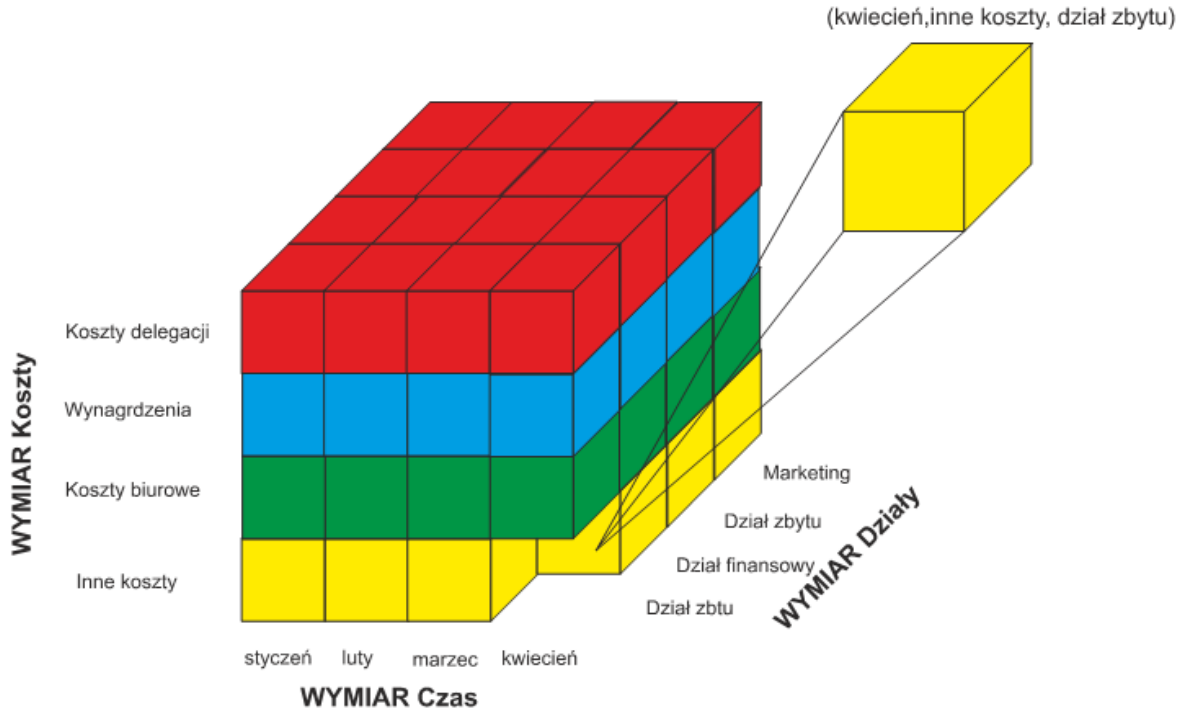
PRZYGOTOWANIE STRUKTUR

Proces przygotowania struktur danych dla mechanizmów analitycznych obejmuje kilka kroków:

- zdefiniowanie źródła danych,
- zdefiniowanie widoku źródła danych
- zdefiniowanie kostki analitycznej,
- zarządzanie kostką analityczną.

KOSTKA ANALITYCZNA

Jedną z bardzo szeroko stosowanych metod analizy danych jest wykorzystanie tzw. kostek analitycznych (ang. Cube). Kostki analityczne są jedynymi obiektami baz analitycznych bezpośrednio dostępnymi dla użytkowników. Łączą one miary z tabel faktów ze zbudowanymi na podstawie tabel wymiarów wymiarami w jeden obiekt - kostkę analityczną.



PRZETWARZANIE OLTP

Reprezentuje tradycyjne podejście do przetwarzania danych umieszczonych w bazie danych (transakcyjne przetwarzanie danych w którym operacje dokonywane są na produkcyjnych bazach danych).

Podstawowe cechy systemów OLTP to:

- wykonywanie dużej liczby prostych zapytań pochodzących od wielu użytkowników (nierzadko są to setki zapytań na sekundę),
- system bazodanowy powinien być zoptymalizowany pod kątem szybkiego wyszukiwania danych,
- częste operacje dodawania, usuwania i modyfikacji pojedynczych rekordów,
- wymagany natychmiastowy dostęp do aktualnych informacji.

PRZETWARZANIE OLAP

Reprezentuje podejście w którym możliwa jest szybka analiza danych i wygenerowanie raportów przeznaczonych do kierownictwa, analityków i administratorów. Sposób przechowywania danych przypomina bardziej wielowymiarowe arkusze kalkulacyjne niż tradycyjną, relacyjną bazę danych.

Podstawowe cechy systemów OLAP to:

- niewielka liczba zapytań, lecz dotyczących wielkich ilości danych (podsumowania itp., mogą to być zapytania zadawane raz na kilka minut przez kilku czy kilkuset użytkowników)
- systemy te zasadniczo tylko odczytują informację z bazy; jeśli system OLAP jest logicznie oddzielony od baz transakcyjnych, to informacje są cyklicznie uzupełniane (dodawanie dużych grup nowych rekordów)
- nie zakładamy pełnej aktualności informacji: dane mogą być dostępne z opóźnieniem (najlepiej znanym z góry, np. jednodniowym), a same obliczenia mogą trwać od sekund do wielu godzin.

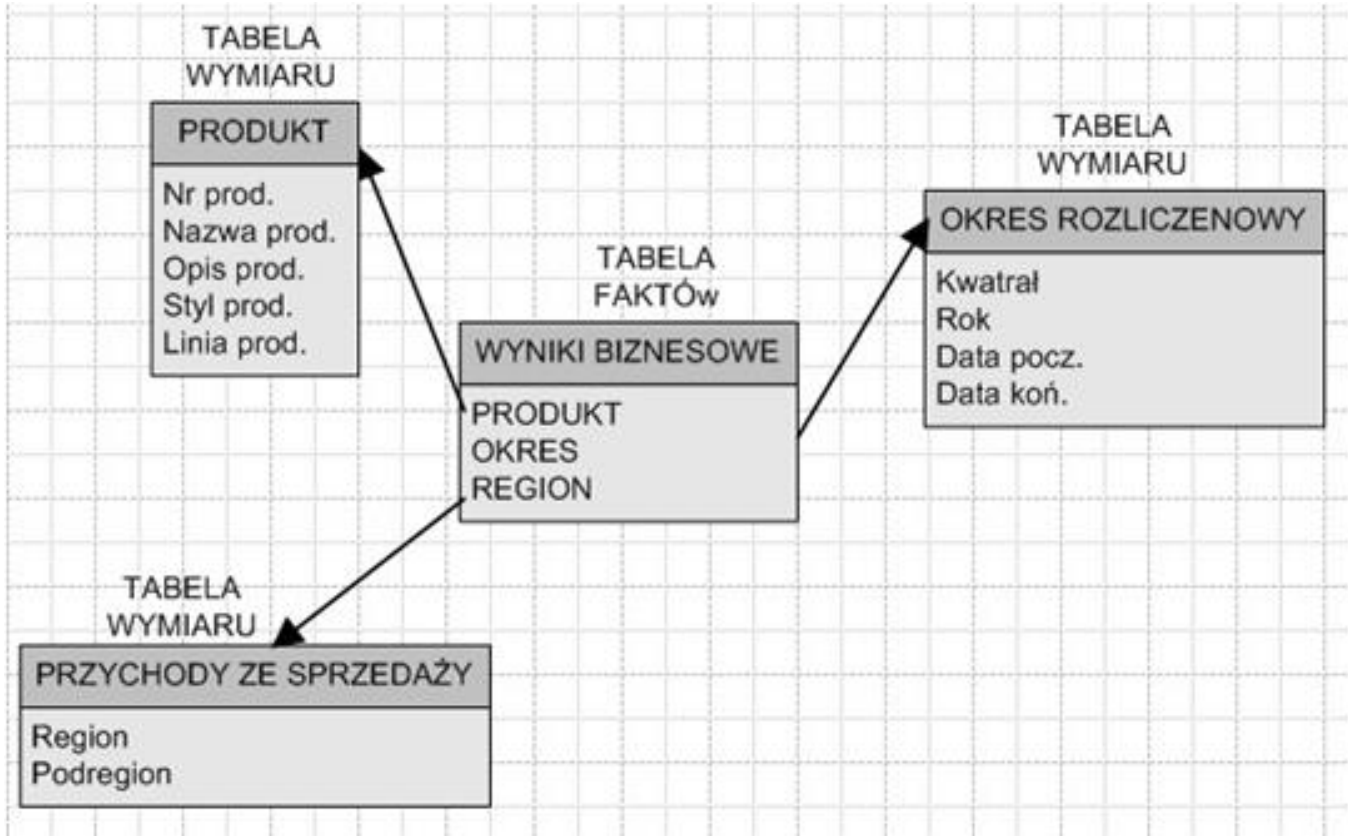
MODELE OLAP

- Model relacyjny (ROLAP)
  - schemat gwiazdy (ang. starschema)
  - schemat płatka śniegu (ang. snowflake schema)
  - schemat konstelacji faktów (ang. fact constellation schema)
  - schemat gwiazda-płatek śniegu (ang. starflake schema)
- Model wielowymiarowy (MOLAP, MDOLAP)
- Model hybrydowy (HOLAP)

MODELE OLAP

Schemat gwiazdy

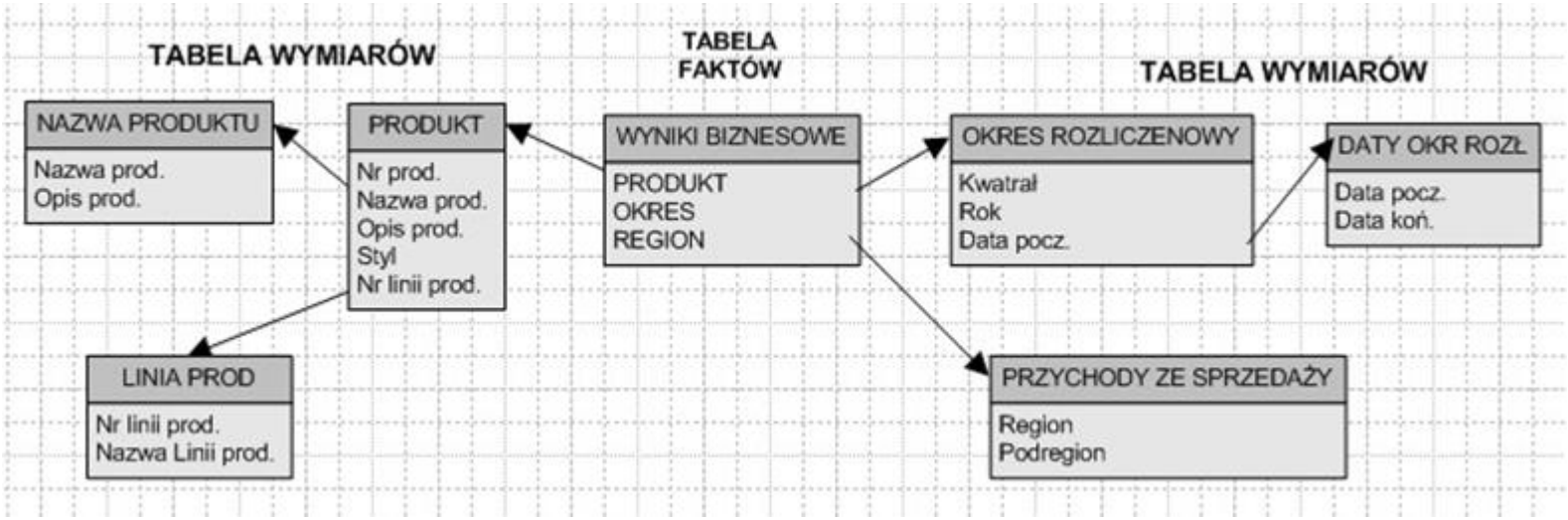
(ang. starschema)



MODELE OLAP

Schemat płatka śniegu

(ang. snowflake schema)



WIELOWYMIAROWY MODEL DANYCH

Wielowymiarowy model danych ma za zadanie przyspieszenie typowych operacji podsumowujących OLAP, i znajduje zastosowanie zwłaszcza w hurtowniach tematycznych (przyśpieszenie takich zadań analitycznych jak podsumowania pewnych wielkości liczbowych, ilości towaru, kwot pieniędzy itp., w rozbiciu na pewne kategorie, często w różnych momentach czasu).

W celu zminimalizowania operacji na wieloterabajtowych danych źródłowych, część agregacji może być policzona zawczasu i przechowywana w postaci wielowymiarowych tabel, tzw. kostek danych.

PRZETWARZANIE OLAP

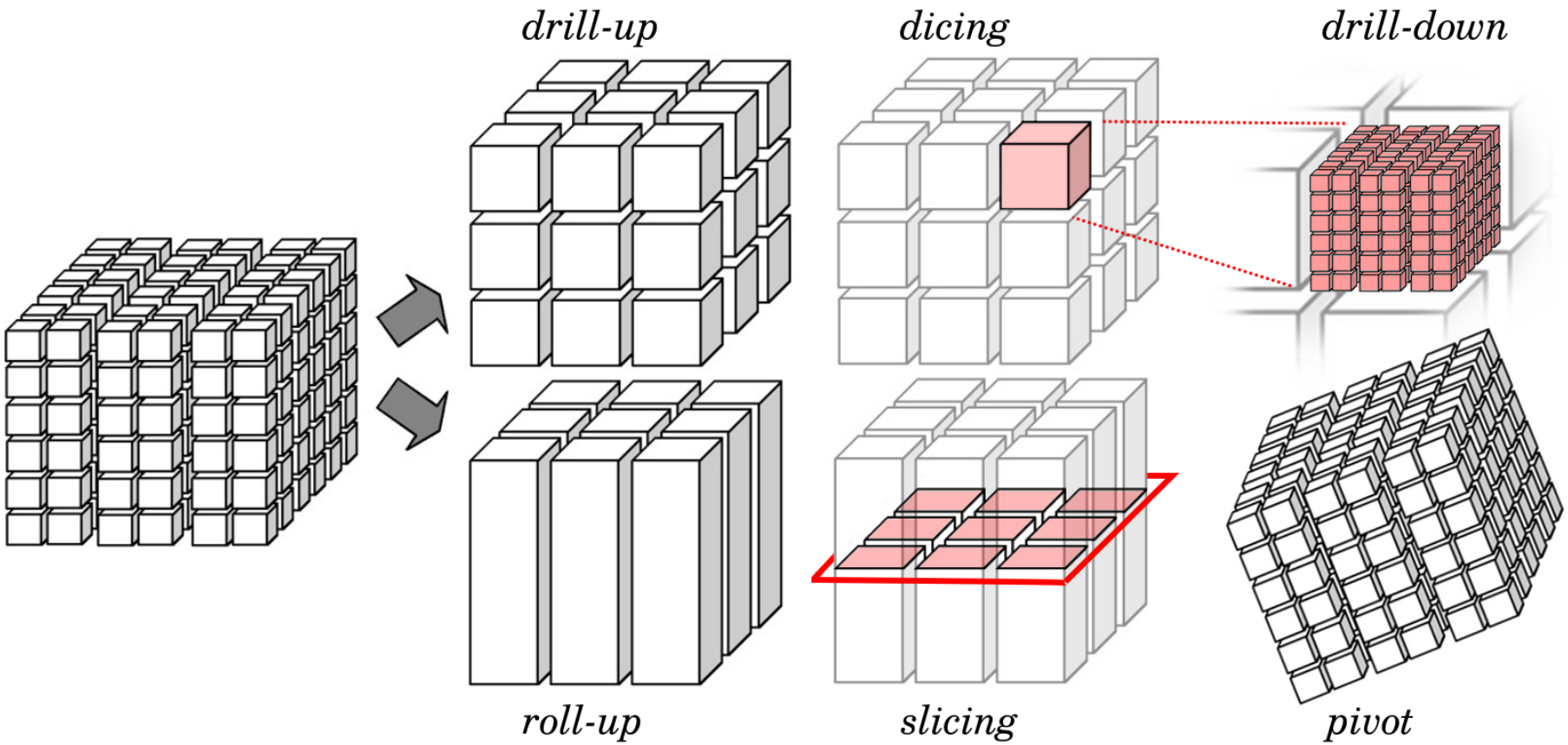
Podstawą modelu OLAP są

- Fakty - informacje podlegające analizie,
- Wymiary - ustalają kontekst analizy, składają się z poziomów, które tworzą hierarchię,
- Wartości miar.

OPERACJE OLAP

- Roll up (drill-up): podsumowanie
  - Przejście do wyższego poziomu w hierarchii lub redukcja wymiarów
- Drill down (roll down): rozwinięcie (odwrotnie do roll-up)
  - Przejście do niższego poziomu w hierarchii lub wprowadzanie nowych wymiarów
- Slice and dice:
  - Rzut i selekcja
- Pivot (rotate):
  - Zmiana orientację kostki, wizualizacja,
- INNE: drill across, drill through

OPERACJE OLAP



NARZĘDZIA

- DB2 DataWarehouse Center,
- Sybase WarehouseStudio,
- Microsoft DataWarehousing Framework,
- SAP Datawarehouse management,
- NCRTeradata Warehouse Builder,
- Oracle Designer6i/9i

DZIĘKUJĘ ZA UWAGĘ