Next-Generation Database Technology

Magazyny danych i technologia OLAP

Opracował Z. Królikowski na podstawie materiałów T. Morzego, T.Koszlajdy, M. Matysiaka, R. Wrembela

Literatura:

1. T. Koszlajda, *Technologia magazynów danych*, w: Materiały II Kraj. Szkoły PLOUG'97, Zakopane.
2. M. Matysiak, *Technologia OLAP*, w: Materiały II Krajowej Szkoły PLOUG'97, Zakopane.
3. R. Wrembel, *Dane hurtowo*, Informatyka, nr.10, 1998
4. T. Morzy, *Eksploracja danych a bazy danych*, Materiały III Krajowej Szkoły PLOUG'98, Zakopane.
5. Chaudhuri S., U. Dayal, *An Overview of Data Warehousing and OLAP Technology*, SIGMOD Record, Vol. 26, No. 1, March 1997.
6. Codd E.F., S.B. Codd, C.T. Salley, *Providing to User-Analysts: An IT Mandate*, Arbor Software's web site, http://www.arborsoft.com/OLAP.html.
7. Widom J., *Research Problems in Data Warehousing*, Proceedings 4th Intern. CIKM Conference, 1995.
8. Http:// www.olapcouncil.org

Informatyzacja firm, instytucji i innych jednostek organizacyjnych powinna realizować **dwa podstawowe cele**:

- **Usprawnienie pracy pojedynczego pracownika**
- **Racjonalizacja działania całych firm**

Ad 1) ***Usprawnienie pracy pojedynczego pracownika***: sprzedawcy, magazyniera, księgowego lub urzędnika - poprzez automatyzację realizowanych przez nich wybranych, rutynowych działań.

- Przykłady takich działań:
  - wprowadzanie zamówień, wydawanie lub przyjmowanie towaru, realizacja sprzedaży, rezerwacja miejsc lub operacja przelewu na kontach bankowych.
- Działania te charakteryzuje ściśle określona procedura postępowania i cykliczna powtarzalność

Ad 2) ***Racjonalizacja działania całych firm*** - w wyniku wspomagania decyzji kadry zarządzającej - przez dostarczenie danych analitycznych opisujących bieżący stan i historię działania danej firmy.

- **Programowe narzędzia analityczne** - udostępnianie informacji statystycznych o bieżącym stanie firmy, występujących trendach itp.

  Korzyści: trafniejsze decyzje o strategicznym znaczeniu dla rozwoju danego przedsiębiorstwa.

Sposób w jaki użytkownik korzysta z bazy danych (w jaki realizuje do niej dostęp) nazywamy **modelem przetwarzania**

Aplikacje operacyjne systemu informatycznego

- **Cel**: wspomaganie pracy pojedynczych pracowników
- **Charakterystyka**:

- proste przetwarzanie,
- działania na niewielkich zbiorach danych szczegółowych,
- realizacja prostych operacji odczytu, wstawiania, modyfikacji i usuwania danych.

- Modelem przetwarzania właściwym dla tej kategorii aplikacji jest tak zwane ***przetwarzanie transakcyjne*** (ang. ***On-line Transaction Processing*** - ***OLTP***).
- **Główne cele tej technologii**:
  - zapewnienie spójności danych,
  - wysoka wydajność systemów pracujących w środowisku wielodostępnym,
- Krytycznym parametrem efektywnościowym takich systemów jest ich przepustowość, mierzona liczbą transakcji w jednostce czasu.

Aplikacje analityczne systemu informatycznego

☞ **Cel**: wspomaganie pracy kadry zarządzającej

☞ **Charakterystyka**:

- dużo większa złożoność przetwarzania niż aplikacji operacyjnych
- zorientowanie na wspieranie procesów decyzyjnych (przetwarzanie danych historycznych, zagregowanych i często skonsolidowanych z wielu źródeł danych: relacyjnych i obiektowych baz danych, arkuszy kalkulacyjnych, itp.)
- realizacja złożonych zapytań wymagających dostępu do milionów krotek (tysiące gigabajtów), wielu operacji połączenia, grupowania i agregowania oraz filtrowania danych
- przykłady takich zapytań: *Jaka jest sprzedaży produktów w supermarkecie w kolejnych kwartałach, miesiącach itp. ? Jaka jest sprzedaż produktów z podziałem na rodzaje produktów (AGD, produkty spożywcze, kosmetyki, itp.)*

Aplikacje analityczne - podsumowanie

Modelem przetwarzania właściwym dla tej kategorii aplikacji jest ***przetwarzanie analityczne*** (ang. ***On-line Analytical Processing*** - ***OLAP***) - ma za zadanie wspieranie procesów analizy danych dostarczając narzędzi umożliwiających taką analizę w wielu „wymiarach" definiowanych przez użytkowników (czas, miejsce, klasyfikacja produktów, itp.).

OLAP – weryfikacja hipotez

Analiza danych zgodnie z modelem OLAP, jest całkowicie sterowana przez analityka. Analityk formułuje zapytania i dokonuje analizy danych. Z tego punktu widzenia, OLAP można interpretować jako rozszerzenie standardu SQL o możliwości efektywnego przetwarzania złożonych zapytań zawierających agregaty.

Aplikacje analityczne - podsumowanie

☞ Przetwarzanie w aplikacjach analitycznych:

- operacje odczytu dużych wolumenów danych, przetwarzanych następnie przez złożone funkcje analityczne,
- proces analizy jest całkowicie sterowany przez użytkownika – mówimy o *analizie danych sterowanej zapytaniami* (ang. *query-driven exploration*)
- odpowiedzi na takie zapytania umożliwiają decydentom określenie wąskich gardeł sprzedaży, produktów przynoszących deficyt, itp.

☞ Efektywność takich systemów: mierzona czasem odpowiedzi

Problemy realizacji systemów OLAP

☞ Komercyjnie dostępne systemy transakcyjne (systemy zarządzania bazami danych SZBD) dostarczają efektywnych rozwiązań dla takich problemów jak: efektywne i bezpieczne przechowywanie danych, transakcyjne odtwarzanie danych, dostępność danych, optymalizacja dostępu do danych, zarządzanie współbieżnością.

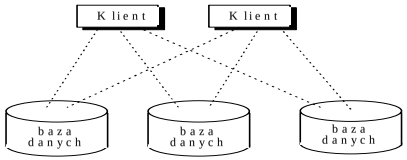
☞ W znacznie mniejszym stopniu systemy te wspomagają operacje agregacji danych, wykonywania pewnych podsumowań czy też optymalizacji złożonych zapytań formułowanych ad hoc.

☞ Systemy te w niewielkim stopniu wspomagają również integrację danych z różnych heterogenicznych źródeł danych.

Problemy realizacji systemów OLAP

- Aby przeprowadzić analizę danych dla wspomagania decyzji, należy dysponować odpowiednimi danymi opisującymi działalność przedsiębiorstwa.
- Bardzo rzadko informacje te są dostępne w jednej bazie danych. Z reguły, są one rozproszone po wielu oddziałowych, rozproszonych geograficznie i heterogenicznych bazach danych.

Typowy stan informatyzacji firm, instytucji: heterogeniczność eksploatowanych systemów - uniemożliwia to bezpośredni dostęp do wszystkich danych określających kondycję danej firmy



Problemy realizacji systemów OLAP

- Stąd, opracowując koncepcję systemu wspomagania podejmowania decyzji należy odpowiedzieć na dwa zasadnicze pytania odnośnie architektury takiego systemu i modelu przetwarzania.

1. Czy analiza powinna mieć charakter rozproszony czy scentralizowany, innymi słowy, czy dane należy zgromadzić i przetwarzać w jednym miejscu w sposób scentralizowany, czy też korzystając z mechanizmu transakcji rozproszonych można przetwarzać dane w sposób rozproszony.
2. Drugie pytanie dotyczy koegzystencji dwóch systemów – systemu bieżącej obsługi działania przedsiębiorstwa oraz systemu wspomagania podejmowania decyzji. Oba systemy operują na tych samych danych, stąd pytanie, czy oba modele OLAP i OLTP mogą współistnieć w tym samym systemie bazy danych, czy też powinny funkcjonować niezależnie.

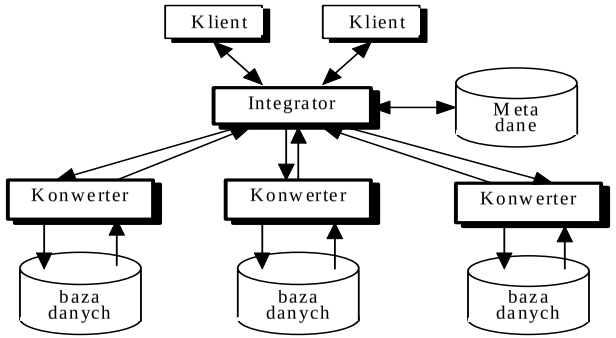
Problemy realizacji systemów OLAP

Problem integracji heterogenicznych i rozproszonych systemów informatycznych

- W ciągu kilku ostatnich lat problem przygotowywania aplikacji realizujących dostęp do heterogenicznych źródeł danych, które są fizycznie rozproszone, zarządzane przez niezależne SZBD, próbowano rozwiązywać na kilka sposobów:

- *konwersja i migracja danych ze starych, zamkniętych systemów do nowych systemów;*
- *wykorzystanie tzw. bramek pomiędzy różnymi systemami baz danych (*ang. *DB gateways)*
- *koncepcja sfederowanych systemów baz danych*

Czy w celu integracji heterogenicznych i rozproszonych systemów informatycznych można wykorzystać koncepcje sfederowanych systemów baz danych ?



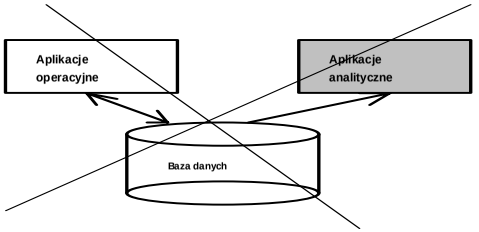
Architektura sfederowanych baz danych

Ocena technologii sfederowanych systemów baz danych

- Idea sfederowanych systemów baz danych nie zakończyła się sukcesem
- Nie powstały żadne rozpowszechnione systemy komercyjne oparte na tej technologii
- Część rozwiązań składających się na tę technologię zostało jednak wykorzystana w produktach komercyjnych i standardach - pomosty i platformy integracyjne:
  - ⇒ *ODBC* (ang. Open Database Connectivity),
  - ⇒ *TUXEDO* i *CORBA* (ang. *Common Object Request Broker Architecture*),
  - ⇒ *DCE* (ang. *Distributed Computing Environment*) i *ODP* (ang. *Open Distributed Processing*).

Problemy realizacji systemów OLAP - cd.

- Odmienność charakterystyki przetwarzania OLTP i OLAP powoduje, że rozwiązania dostępne w standardowych systemach baz danych są nieprzydatne do eksploatacji aplikacji analitycznych
- Równoczesna eksploatacja aplikacji operacyjnych i analitycznych w środowisku tego samego systemu bazy danych, musi prowadzić do niskiej efektywności działania całego systemu informatycznego.



Problemy realizacji systemów OLAP - cd.

Wnioski: analiza powinna mieć charakter scentralizowany, a modele OLAP i OLTP powinny funkcjonować niezależnie.

- Oczywiście, odpowiedź na pytania o architekturę i model przetwarzania jest uzależniona od aktualnego stanu rozwoju technologii informatycznej.
- Ze względu na charakter i pracochłonność obliczeń, częściowo również ze względu na problem autoryzacji dostępu do danych, analiza danych jest aktualnie prowadzona w sposób scentralizowany.
- Wraz z rozwojem sieci komputerowych, wzrostem prędkości transmisji danych, należy się jednak spodziewać przechodzenia od modelu przetwarzania analitycznego scentralizowanego do modelu przetwarzania analitycznego rozproszonego.

Magazyn danych - koncepcja i architektura

W ostatnim czasie prace badawcze i rozwojowe prowadzone w ramach powyższych problemów doprowadziły do opracowania nowego typu relacyjnej bazy danych nazwanego **magazynem danych** (ang. *data warehouse*).

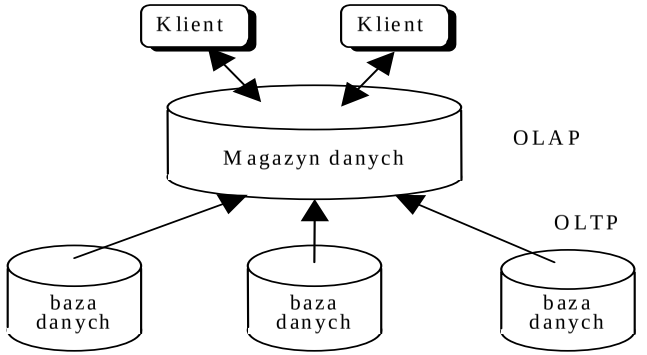
- Magazyny danych, są „*tematycznie zorientowanymi, zintegrowanymi, zmiennymi w czasie, nie ulotnymi zbiorami danych, wykorzystywanymi w organizacjach głównie do przetwarzania analitycznego i podejmowania decyzji*”
- Magazyny danych są niezależne od operacyjnych baz danych, na których działają aplikacje OLTP

Magazyn danych - koncepcja i architektura

☞ Uzasadnienie konieczności budowy magazynów danych dla przetwarzania analitycznego:

1. procesy decyzyjne wymagają danych, na przykład o trendach, których może nie być w operacyjnych bazach danych
2. procesy decyzyjne wymagają dostępu do skonsolidowanych danych pochodzących z wielu heterogenicznych źródeł, które mogą używać niezgodnych formatów danych i niezgodnego kodowania
3. operacje typowe dla systemów OLAP wymagają specjalnego składowania danych, odpowiednich struktur i metod dostępu do danych, których nie stosuje się w tradycyjnych, komercyjnych systemach zarządzania bazami danych (ang. DBMS).

Magazyn danych - rozdzielenie przetwarzania operacyjnego i analitycznego



Koncepcja magazynu danych - cd.

☞ Problem konstrukcji magazynu danych wiąże się z problemem **magazynowania danych** (ang. data warehousing).

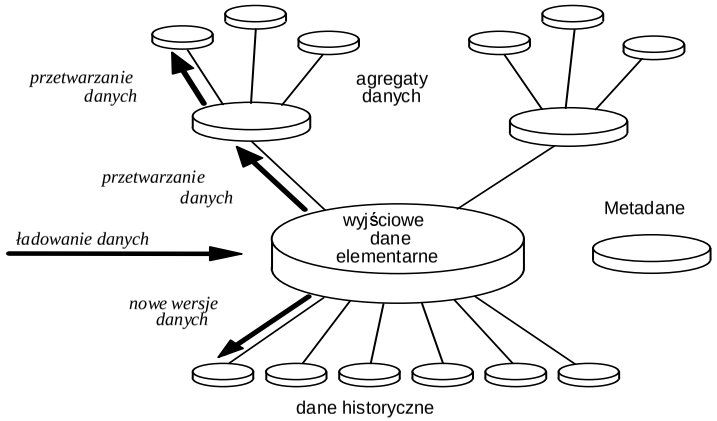
☞ Magazynowanie danych jest procesem zbierania i przetwarzania danych z różnych, heterogenicznych i rozproszonych źródeł danych w celu uzyskania jednolitego obrazu części bądź całości działalności danego przedsiębiorstwa.

Struktura magazynu danych

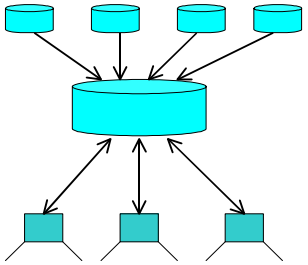
W magazynie danych przechowywane są następujące kategorie danych:

- dane elementarne pozyskane bezpośrednio ze źródłowych heterogenicznych baz danych (wykonanych w różnych technologiach), jak i ze źródeł innych niż bazy danych, np. arkusze kalkulacyjne, dokumenty tekstowe, pliki HTML, multimedia;
- dane historyczne tworzone w momencie pojawiania się nowych wartości już przechowywanych danych;
- dane sumaryczne (zagregowane) o różnym stopniu przetworzenia;
- dane opisujące semantykę, pochodzenie i algorytmy wyznaczania poprzednich trzech typów danych.

Struktura magazynu danych – cd.

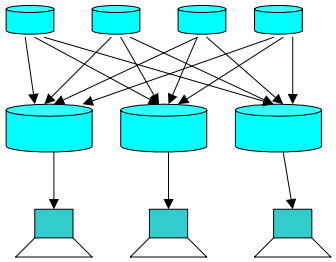


Architektury magazynów danych



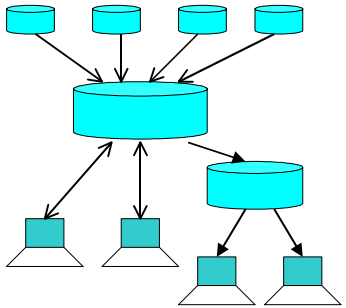
Zakładowy magazyn danych

Architektury magazynów danych



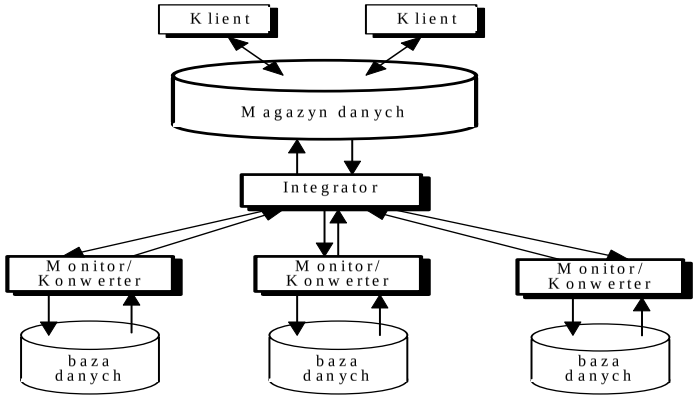
Zbiór niezależnych oddziałowych magazynów danych

Architektury magazynów danych



Zbiór zależnych oddziałowych magazynów danych

Architektura systemu zarządzania magazynem danych



Architektura systemu zarządzania magazynem danych - cd.

Źródła danych: źródłowe bazy danych i źródła inne niż bazy danych

- Formaty fizyczne, logiczne i pojęciowe poszczególnych źródeł danych mogą różnić się między sobą
- Z każdym z takich źródeł jest związana warstwa oprogramowania o nazwie *konwerter / monitor*

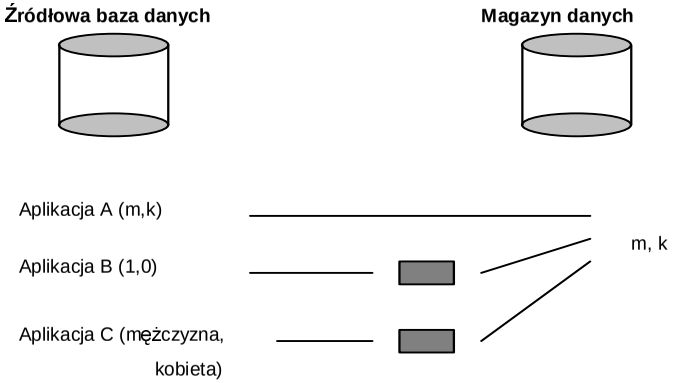
Moduły monitorowania i konwersji danych

- Zadania:
  - automatyczne pozyskiwanie danych z różnych źródłowych baz danych;
  - transformowanie danych z formatu wykorzystywanego w źródle do formatu wykorzystywanego w magazynie - dla każdego modelu danych źródłowych konieczne jest zastosowanie specyficznego modułu *konwertera,*
  - wykrywanie zmian w danych źródłowych i ich przekazywanie do warstwy oprogramowania *integratora* (po uprzedniej konwersji do modelu danych magazynu);

Moduły monitorowania i konwersji danych

- Zadania -cd:
  - sposób wykrywania zmian w danych źródłowych zależy od własności samych źródeł – z tego punktu widzenia, wyróżnia się cztery następujące rodzaje źródeł danych:
    - **aktywne**, tzn. posiadające zaimplementowane mechanizmy wyzwalaczy, które informują *monitor* o zmianach zachodzących w danych źródłowych;
    - **utrzymujące dzienniki operacji** wykonywanych na danych źródłowych - zmiany są wykrywane przez analizę zawartości dziennika przez moduł *monitora;*
    - **umożliwiające wydawanie zapytań** - w celu wykrycia zmian w danych źródłowych, *monitor* okresowo wydaje zapytania do wszystkich źródeł;
    - **wspierające mechanizm migawek** (ang. *snapshot*) - migawka jest programem, który okresowo zapisuje do pliku zawartość źródłowej bazy danych, a zmiany informacji wykrywa się przez porównywanie zawartości kolejnych plików.
  - odfiltrowanie nadmiarowych i błędnych danych;

Potrzeba konwersji i integracji danych



Architektura systemu zarządzania magazynem danych - cd.

Moduł integratora

Dane przechowywane w magazynie danych mogą różnić się schematem pojęciowym od danych przechowywanych w poszczególnych źródłowych bazach danych - zazwyczaj są to dane bardziej przetworzone, na przykład do wartości sumarycznych, średnich itp.

Zadania modułu *integratora*:

☞ Moduł *integratora* jest odpowiedzialny za łączenie danych pochodzących z wielu źródeł i uaktualnianie danych w magazynie - proces ten składa się nie tylko z wpisywania, uaktualniania i usuwania danych, ale również z wstępnego ich przetwarzania (ang. *data scrubbing*), tj. filtrowania, eliminowania duplikatów, usuwania niespójności, obliczania agregatów.

Własności systemu zarządzania magazynem danych

System zarządzania magazynem danych powinien zapewniać:

- **Efektywne przetwarzanie analityczne dużego wolumenu danych**
  - ☞ przyśpieszenie dostępu do wyników analizy danych:
    - *materializacja perspektyw (agregatów)* (ang. materialized view)
    - zastosowanie algorytmów przetwarzania równoległego i parcelacja danych
- **Utrzymywanie i przetwarzanie danych historycznych**
- **Efektywne przetwarzanie danych wielowymiarowych**
  - nowe rodzaje indeksów: indeksy bitmapowe, indeksy połączeniowe, bitmapowe indeksy połączeniowe,
  - nowe algorytmy optymalizacji wykonywania zapytań

Efektywność systemu zarządzania magazynem danych (SZMD) - wybrane problemy

Przetwarzanie równoległe

- Równolegle przetwarza się zapytania, sortuje dane, wykonuje operacje odczytu i zapisu na dysk, buduje tablice i indeksy oraz wczytuje dane do magazynu
- Przetwarzanie równoległe wspierają m.in. systemy zarządzania bazami danych: *Oracle7* i *Oracle8* (Oracle Corporation) , *DB2* (IBM), *OnLine Extended Parallel Server*, *OnLine Dynamic Server* (Informix), *Red Brick Warehouse* (Red Brick), *Sybase IQ* (Sybase)

Efektywność systemu zarządzania magazynem danych (SZMD)- wybrane problemy - cd

Parcelacja danych

Umożliwia automatyczne rozpraszanie danych (pochodzących z jednej lub wielu relacji) na wiele dysków, znajdujących się w tym samym lub wielu węzłach (komputerach) sieci - dzięki podziałowi dużej relacji na mniejsze:

- bardzo kosztowne operacje wejścia/wyjścia, tj. dostępu do dysków mogą być wykonywane równolegle,
- równoważone jest obciążenie dysków,
- polecenia SQL mogą być wykonywane równolegle,
- wzrasta bezpieczeństwo danych w przypadku awarii sprzętu,
- wzrasta szybkość tworzenia kopii zapasowych bazy i szybkość odtwarzania danych po awarii.

Efektywność SZMD - wybrane problemy - cd

Techniki parcelacji danych:

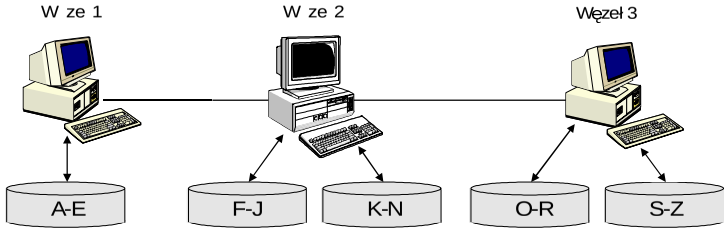
- *round–robin* (ang. round–robin partitioning),
- parcelacja *bazująca na wartości* (ang. range partitioning),
- *haszowa* (ang. hash partitioning),
- *hybrydowa* (ang. hybrid partitioning).

Technika ***round–robin*** umożliwia równomierne rozproszenie danych w węzłach sieci. Przykładowo, jeśli w sieci znajdują się trzy węzły, to pierwsza krotka relacji zostanie umieszczona w węźle pierwszym, druga – w węźle drugim, trzecia krotka – w węźle trzecim, czwarta – znów w węźle pierwszym itp.

Wada: ponieważ dane są rozproszone w sposób przypadkowy, więc odnalezienie żądanych informacji wymaga przeszukania wszystkich węzłów.

Efektywność SZMD - wybrane problemy - cd
Parcelacja danych -cd

Parcelacja ***bazująca na wartości*** - rozmieszczenie danych w sieci zależy od wartości samych danych (np. relacja zawierająca informacje o klientach sieci supermarketów może być podzielona zgodnie z wartością pierwszej litery nazwiska)



Zaleta: ten sposób rozpraszania danych jest efektywny dla zapytań wykorzystujących zakresy wartości w predykatach selekcji, ponieważ umożliwia szybki dostęp do danych z żądanego zakresu, bez potrzeby przeszukiwania wszystkich węzłów.

Efektywność SZMD - wybrane problemy - cd
Parcelacja danych -cd

W ***parcelacji haszowej*** dane są umieszczane w węzłach zgodnie z wartością systemowej funkcji haszowej.

- argumentem wejściowym tej funkcji jest wartość atrybutu, a jej wynikiem – adres węzła, w którym zostanie umieszczona krotka.
- w celu odnalezienia żądanych informacji SZBD wykorzystuję tę samą funkcję haszową.

Zaleta: możliwość automatycznego umieszczania w tym samym węźle krotek pochodzących z różnych, powiązanych z sobą relacji - w ten sposób zwiększa się efektywność wykonywania operacji łączenia krotek, gdyż łączone z sobą krotki znajdują się w tym samym węźle.

Efektywność SZMD - wybrane problemy - cd
Parcelacja danych -cd

Parcelacja hybrydowa umożliwia dwustopniowe rozpraszanie danych.

- w kroku pierwszym dane są umieszczane w poszczególnych węzłach za pomocą parcelacji haszowej;
- w kroku drugim dane są umieszczane na poszczególnych dyskach danego węzła, za pomocą parcelacji bazującej na wartości.

Zaleta: wzrasta równomierność rozproszenia danych i obciążenia węzłów.

Efektywność systemu zarządzania magazynem danych (SZMD)- wybrane problemy - cd

Materializowanie agregatów

☞ Wobec ogromnych rozmiarów magazynów danych, wymóg szybkiej odpowiedzi systemu na złożone zapytanie (np. łączną sprzedaż lodówek) wymaga materializowania agregatów, czyli wyliczenia ich z wyprzedzeniem i zapamiętania w bazie danych, tak aby w chwili otrzymania zapytania zagregowane wartości były już gotowe.

Przetwarzanie w magazynach danych - *Własności danych*

Nazwa klienta	*Adres klienta*	*Telefon*
Alfa	ul. Akacjowa 4	8345-543
Beta	ul. Konwaliowa 8	8665-545
Gamma	ul. Klonowa 34/36	8434-221
Delta	ul. Albańska 8	8665-645

Przykładowa relacja - jeden wymiar

Nawigacja po krotkach relacji:

- wzdłuż tylko jednego wymiaru - wymiaru obiektów lub faktów, o których informacje są przechowywane w relacji
- zbiór identyfikatorów np. klientów - punkty na osi współrzędnych,

Przetwarzanie w magazynach danych - *Własności danych - cd.*

☞ Potrzeba danych wielowymiarowych - Przykład:

- w bazie danych są przechowywane informacje o klientach, towarach i sprzedaży,
- zapytanie: *ile towaru X sprzedano klientowi Y ?*

		Towar		
		Lodówka	Pralka	Zmywarka
Klient	Alfa	20	23	5
	Beta	4	0	24
	Gamma	45	147	35
	Delta	71	12	40

Sprzedaż pokazana w dwóch wymiarach: towary i klienci

Przetwarzanie w magazynach danych - *Własności danych - cd.*

Towar	*Klient*	*Sprzedaż*
Lodówka	Alfa	20
Lodówka	Beta	4
Lodówka	Gamma	45
Lodówka	Delta	71
Pralka	Alfa	23
Pralka	Gamma	147
Pralka	Delta	12
Zmywarka	Alfa	5
Zmywarka	Beta	24
Zmywarka	Delta	40

Sprzedaż towarów dla klientów zapamiętana w 1-wymiarowej relacji

Przetwarzanie w magazynach danych - *Własności danych - cd.*

- Korzyści wynikające ze stosowanie struktur wielowymiarowych do przechowywania informacji:
  - przejrzysta reprezentacja wiedzy,
  - znaczenie efektywnościowe.

Przetwarzanie w magazynach danych - *Materializowanie agregatów*

		Towar			
		Lodówka	Pralka	Zmywarka	*Łącznie:*
Klient	Alfa	20	23	5	**48**
	Beta	4	0	24	**28**
	Gamma	45	147	35	**227**
	Delta	71	12	40	**123**
	Łącznie:	**140**	**182**	**104**	**426**

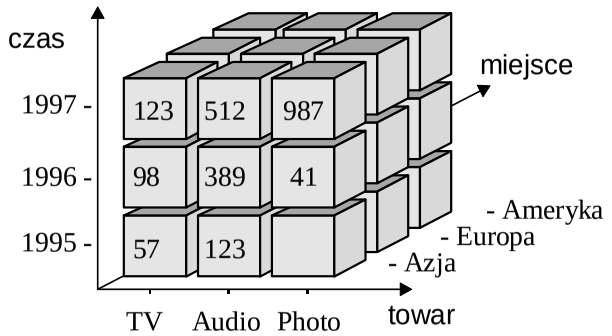
Zmaterializowane agregaty w wielowymiarowej bazie danych

- Agregaty są wyliczane poprzez operacje grupowania dla wskazanych wymiarów

Przetwarzanie w magazynach danych - terminologia

- ***Dana wielowymiarowa*** (ang. *cube*, *multi-dimensional array*) jest zbiorem ***komórek*** danej (ang. *cells*) zlokalizowanych w przestrzeni wielowymiarowej, określonej przez wymiary (ang. *dimension*) danej
- Pojedyncza komórka wyznaczona przez zbiór wartości wymiarów reprezentuje miarę danej w danym punkcie przestrzeni - np. ilości towarów lub obrót ze sprzedaży, są nazywane ***miarą***.
- ***Wymiary danych*** są strukturalnymi i w ogólności złożonymi atrybutami grupującymi elementy (ang. *member*) tego samego typu.
  - Na przykład:
    - *Alfa*, *Beta*, *Gamma*, *Delta*, są elementami wymiaru *Klient*
    - wymiar *czas* jest zbiorem elementów: dekada, rok, kwartał, miesiąc, tydzień, dzień, godzina,
  - W typowych zastosowaniach rolę wymiarów pełnią: *czas*, *lokalizacja*, *typ produktu*

Przetwarzanie w magazynach danych - terminologia - cd.



Przykład danych wielowymiarowych

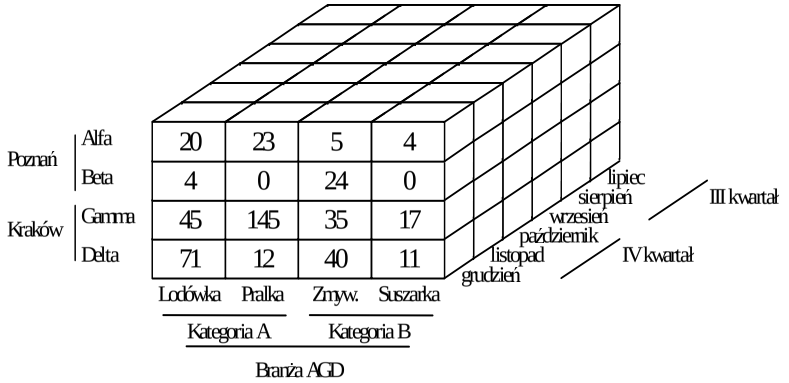
Przetwarzanie w magazynach danych - terminologia - cd.

☞ Wymiary mogą być wewnętrznie złożone i opisane za pomocą wielu atrybutów, a atrybuty mogą pozostawać w pewnych zależnościach, tworząc ***hierarchie atrybutów***

Na przykład:

- *towar*, który jest jednym z wymiarów, może być opisany między innymi kategorią towaru i branżą, do której należy - mamy do czynienia ze złożonym wymiarem, posiadającym trzypoziomową hierarchię atrybutów: towar-kategoria-branża
- *miejsce sprzedaży* - hierarchia klient-miasto-województwo
- *czas sprzedaży* - hierarchia dzień-miesiąc-kwartał-rok

Hierarchie atrybutów - przykład



Hierarchie atrybutów w ramach wymiarów

Projektowanie magazynu danych

Bazy danych wspierające technologię magazynów danych (technologię OLAP) można podzielić na dwa rodzaje, ze względu na wykorzystywane przez nie modele danych.

1. Magazyny *relacyjne*, nazywane również ***ROLAP*** (ang. *Relational OLAP*), wykorzystujące systemy zarządzania relacyjną bazą danych, posiadające dodatkowe mechanizmy efektywnego przetwarzania zapytań typu *OLAP*
2. Magazyny *wielowymiarowe*, nazywane również ***MOLAP*** (ang. *Multi-Dimensional OLAP*), wykorzystujące specjalizowane systemy zarządzania, umożliwiające przechowywanie danych w wielowymiarowych tablicach i wykonywanie operacji *OLAP* zdefiniowanych dla tych struktur danych.

ROLAP

☞ Zwykle schemat takiej hurtowni posiada strukturę **gwiazdy** (ang. *star schema*) lub strukturę bardziej złożoną, przypominającą **płatek śniegu** (ang. *snowflake schema*). W celu skrócenia czasu potrzebnego na wyznaczenie wyników zapytania relacje bazy danych są często denormalizowane, np. zawierają wartości zagregowane, są wynikiem połączenia wielu innych relacji.

☞ Technika projektowania - diagram związków encji

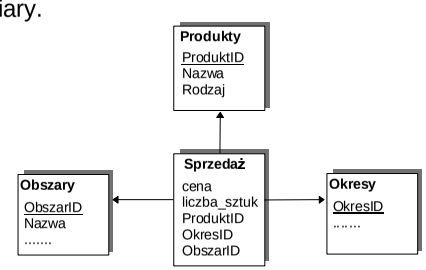
⇒ schemat bazy danych ma strukturę przypominającą gwiazdę - w centrum gwiazdy znajduje się relacja zawierająca dane źródłowe - nazywana relacją faktów, a na około znajdują się relacje odpowiadające wszystkim wymiarom

⇒ poziomy w ramach wymiaru mogą być przechowywane w osobnych relacjach powiązanych wzajemnie związkami typu wiele do jednego.

⇒ każda krotka w relacji faktów, czyli każdy pojedynczy fakt, posiada zbiór kluczy obcych wskazujących na odpowiednie współrzędne w relacjach reprezentujących wymiary.

ROLAP

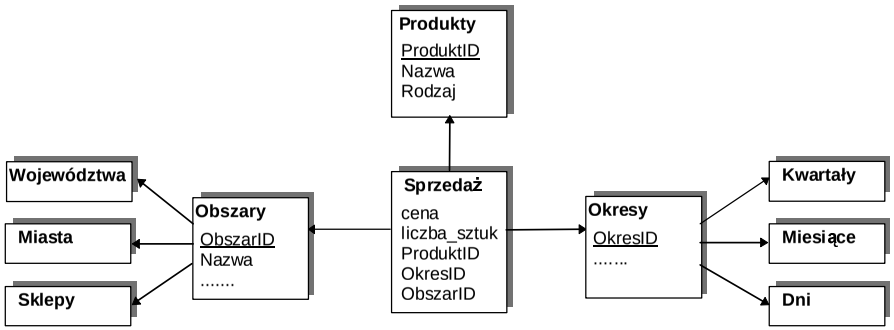
Centralna relacja *Sprzedaż* zawiera informacje o sprzedaży pewnych produktów, w pewnych obszarach geograficznych, w określonym czasie. Relacje *Produkty*, *Obszary* i *Okresy* są *wymiarami*, natomiast relacja *Sprzedaż* jest relacją *faktów* (ang. *fact table*). Atrybuty relacji faktów przechowujące informacje o sprzedaży są *miarami* (ang. *measures*), np. *cena*, *liczba_sztuk*. Relacja faktów – *Sprzedaż* zawiera również atrybuty *ProduktID*, *ObszarID*, *OkresID*, których wartości wskazują na odpowiednie wymiary.



Schemat gwiazdy

ROLAP

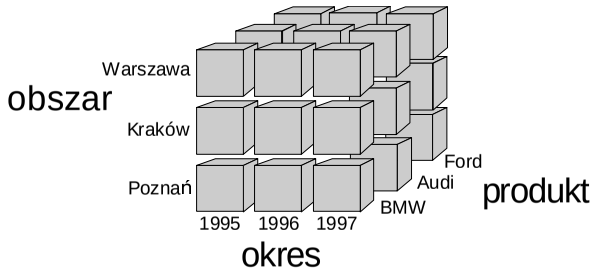
Jeśli wymiary tworzą hierarchie, to schemat hurtowni danych ma często postać płatka śniegu.



Schemat płatka śniegu

MOLAP

Struktury danych *MOLAP* (ang. *multidimensional arrays, datacubes*) zawierają dane wstępnie przetworzone (m.in. zagregowane) pochodzące z wielu źródeł.



Tablica trójwymiarowa, zawierająca trzy wymiary: *obszar*, *okres* i *produkt* oraz zagregowane informacje o sprzedaży samochodów w poszczególnych latach, w wybranych miastach

MOLAP

Analizę danych wielowymiarowych wspomagają specjalne operatory, do których należą:

- ⇒ wyznaczanie punktu centralnego (ang. *pivoting*),
- ⇒ nawigacja w górę lub w dół (rozwijanie (ang. *drill-down*), zwijanie (ang. *roll-up* lub *drill-up*)),
- ⇒ obracanie (ang. *rotating*),
- ⇒ projekcja (wycinanie) (ang. *slice and dice*),
- ⇒ wyznaczanie rankingu (ang. *ranking*).

MOLAP - operacje

Wyznaczanie punktu centralnego (ang. *pivoting*)

⇒ Operacja ta polega na wskazaniu miary i określeniu wymiarów, w których wybrana miara będzie prezentowana. Przykładowo, w wymiarze produktu reprezentującego samochód marki BMW i wymiarze obszaru reprezentującego sklepy województwa poznańskiego może być prezentowana liczba sprzedanych samochodów.

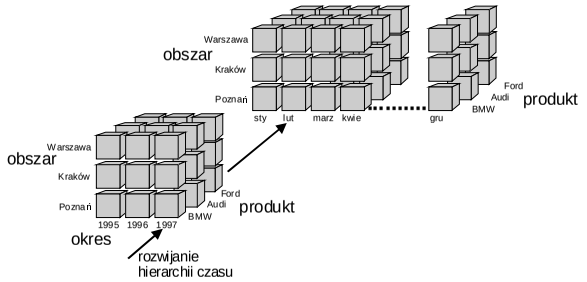
Rozwijanie (ang. *drillng down*)

⇒ Rozwijanie polega na zagłębianiu się w hierarchię danego wymiaru w celu przeprowadzenia bardziej szczegółowej analizy danych. Jako przykład rozważmy informacje o sprzedaży samochodów marek BMW, Audi i Ford, w latach 95, 96 i 97, w poszczególnych miastach.

MOLAP - operacje

Rozwijanie (ang. *drillng down*)

⇒ W celu dokonania analizy sprzedaży w poszczególnych miesiącach roku 97 należy rozwinąć hierarchię reprezentującą czas, tj. rok 97. Analiza sprzedaży w poszczególnych dniach wybranego miesiąca będzie możliwa po rozwinięciu hierarchii reprezentującej ten miesiąc.

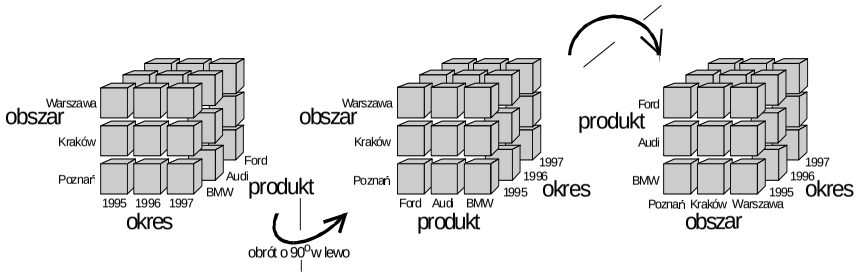


Operacja rozwijania hierarchii wymiaru

MOLAP - operacje

Obracanie (ang. *rotating*);

☞ Operacja obracania umożliwia prezentowanie danych w różnych układach. Celem jej jest zwiększenie czytelności analizowanych informacji.

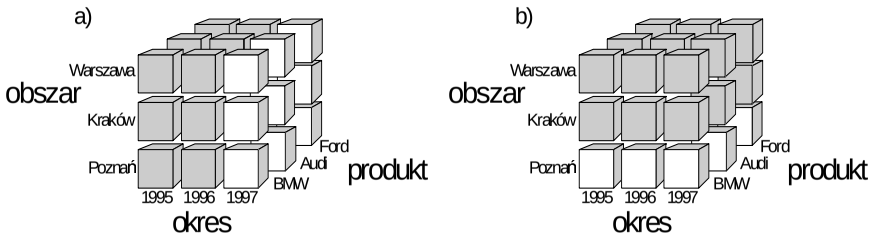


Operacja obracania

MOLAP - operacje

Wycinanie (ang. *slicing and dicing*);

☞ Operacja ta umożliwia zawężenie analizowanych danych do wybranych wymiarów, a w ramach każdego z wymiarów – zawężenie analizy do konkretnych jego wartości.



Wycinanie danych w różnych wymiarach

MOLAP - operacje

Zwijanie (ang. *rolling up*)

☞ Zwijanie jest operacją odwrotną do rozwijania i polega na nawigowaniu w górę hierarchii danego wymiaru. Dzięki tej operacji można przeprowadzać analizę danych zagregowanych na wyższym poziomie hierarchii wymiarów.

Obliczanie rankingu (ang. *ranking*);

Operacja ta umożliwia uporządkowanie informacji w danym wymiarze, zgodnie z wartościami wybranych miar (w kolejności malejącej lub narastającej). Przykładowo, w wymiarze roku 97 można uporządkować marki samochodów zgodnie z narastającym porządkiem liczby sprzedanych egzemplarzy.

Produkty komercyjne

Obecnie wiele wiodących firm w dziedzinie baz danych oferuje serwery wspierające technologię hurtowni danych. Są to:

⇒ *Oracle7*, *Oracle8* i *Oracle Express Server* – Oracle Corporation,
⇒ *DB2* – IBM,
⇒ *Sybase IQ* – Sybase, Inc.,
⇒ *OnLine Dynamic Server*, *OnLine Extended Parallel Server* i *OnLine Workgroup Server* – Informix Software, Inc.,
⇒ *Red Brick Warehouse* – Red Brick Systems, Inc.,
⇒ *Teradata* – NCR, *Adabas C* i *Adabas D* – Software AG,
⇒ *Essbase* – Arbor Software Corporation.

Produkty komercyjne - Oracle

ORACLE for Warehouse Technology

- moduły równolegle ładujące dane do magazynu,
- zestaw bram SQL do relacyjnych i nie-relacyjnych baz danych,
- moduły wspomagające asynchroniczną replikację danych z operacyjnych baz danych do magazynu (periodyczna lub „*event driven*”,

Typy danych w magazynie:

1. sformatowane (ang. *record-oriented*)
2. tekstowe,
3. przestrzenne (wielowymiarowe),
4. dane multimedialne.

Produkty komercyjne - Oracle

Techniki specjalne wspomagające typy danych w magazynie

1.Dane sformatowane:

- optymalizacja zapytań z uwzględnieniem produktu kartezjańskiego,
- równoległe wykonywanie zapytań (*Oracle Parallel Query Option*), w tym:
- tworzenie agregatów: *create table as <subquery>*
- indeksy bitmapowe.

2. Dane tekstowe (Oracle TextServer)

- przykłady: relacje prasowe, raporty roczne, kontrakty, pisma, itp.
- indeksy bitmapowe - bit *i* jest ustawiony jeśli słowo występuje w dokumencie,
- dokumenty są przechowywane w postaci skompresowanej,
- wyszukiwanie kontekstowe - moduł *ConText* (parser języka naturalnego i sieci semantyczne).

Produkty komercyjne - Oracle

Techniki specjalne wspomagające typy danych w magazynie

3. *Dane przestrzenne (wielowymiarowe)*

- integracja informacji geograficznych z danymi operacyjnymi (technika HHCode).

4. *Dane multimedialne*

- w magazynie: reklamy produktów firmy, wywiady - *Oracle Media Server*

Produkty komercyjne - Oracle

Narzędzia firmy Oracle wspomagające technologię magazynów danych:

- **Data Mart Suite** – tematyczne magazyny danych poświęcone wybranym zagadnieniom działalności przedsiębiorstwa (system analizy sprzedaży w sieci supermarketów, analiza wydatków i informacje o użytkownikach kart kredytowych w banku, informacje o długości rozmów i rodzajach połączeń operatorów telefonii komórkowej, itd.)
- **Oracle Discoverer** – (ostatnia wersja – *Discoverer Viewer for Web*) należy do rodziny narzędzi Oracle przeznaczonych do wspomagania decyzji. Jest interaktywnym, łatwym w obsłudze programem do przeszukiwania baz i magazynów danych (zapytania *ad-hoc*), tworzenia raportów, wykresów oraz stron internetowych. Udostępnia użytkownikom na różnych poziomach organizacji informacje wyszukane w *ROLAP* (w tym, analiza wielowymiarowa) i systemach *OLTP*.

Produkty komercyjne - Oracle

Narzędzia firmy Oracle wspomagające technologię magazynów danych:

- **Oracle Express Server i Relational Access Manager** (RAM) – Express Server jest serwerem magazynu *MOLAP*, zasilającym aplikacje analityczne. RAM łączy aplikacje Express'a z danymi w magazynie.
- **Oracle IRI ExpressView** - moduł analizy i ekstrapolacji danych w magazynie (w tym analiza typu „*what-if*")
- **Express Objects** – jest to obiektowo zorientowane narzędzie służące tworzeniu i rozwijaniu aplikacji *OLAP* w architekturze *klient/serwer*. Umożliwia tworzenie aplikacji w trybie graficznym jak również przy pomocy narzędzi programistycznych.