

Przetwarzanie danych w magazynach danych

Tadeusz Morzy

Politechnika Poznańska, Instytut Informatyki
60-965 Poznań, Piotrowo 3A
morzy@put.poznan.pl

Plan wystąpienie

- Systemy przetwarzania transakcyjnego
- Magazyn (hurtownia) danych
- Model przetwarzania analitycznego OLAP
- Wielowymiarowy model danych
- Schematy pojęciowe magazynów danych
- Typy magazynów danych
- Architektury fizyczne magazynów danych
- Efektywność przetwarzania danych
- Wnioski i uwagi końcowe

Systemy przetwarzania transakcyjnego

- Celem systemów przetwarzania transakcyjnego (OLTP) jest usprawnienie bieżącej działalności operacyjnej przedsiębiorstwa
- Komercyjnie dostępne systemy OLTP (systemy zarządzania bazami danych SZBD) dostarczają efektywnych rozwiązań dla:
 - efektywnego i bezpiecznego przechowywania danych,
 - transakcyjnego odtwarzania danych,
 - optymalizacji dostępu do danych,
 - zarządzania współbieżnością.

Przetwarzanie analityczne

- Systemy OLTP charakteryzują się krótkimi i prostymi transakcjami, które operują na niewielkiej części danych przechowywanych w bazie danych
- Miarą oceny działania systemu OLTP jest **przepustowość transakcji**
- Systemy OLTP nie wspomagają procesów analizy danych, gdyż w znacznie mniejszym stopniu wspomagają operacje agregacji danych, wykonywania podsumowań czy też optymalizacji złożonych zapytań formułowanych ad hoc

Przetwarzanie analityczne

- Potrzeba przetwarzania analitycznego danych:
 - analiza działalności przedsiębiorstwa
 - analiza trendów i anomalii
 - zarządzanie przedsiębiorstwem
 - opracowywanie strategii marketingowej
 - analiza rentowności inwestycji, itp.
- Aplikacje analityczne wymagają:
 - integracji danych
 - złożonej analizy danych
 - eksploracji danych

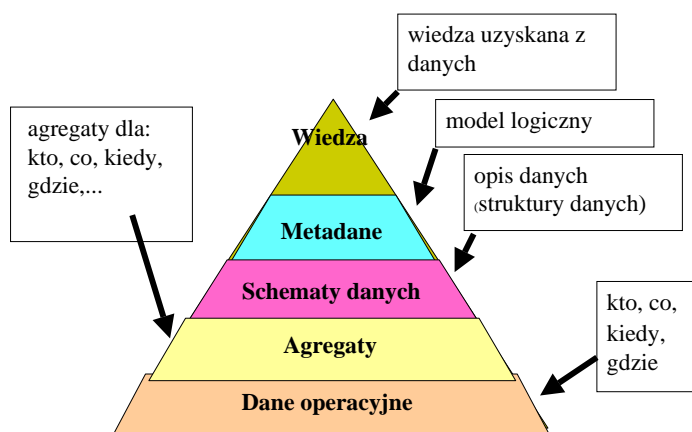
Przykładowe aplikacje analityczne

- Przykłady:
 - bankowość (np. identyfikacja czynników ryzyka wskazujących, którzy klienci gwarantują bezpieczne spłacanie udzielonego kredytu),
 - rynki finansowe (np. identyfikacja trendów w zakresie akcji spółek giełdowych),
 - telekomunikacja (np. identyfikacja klientów zainteresowanych nowymi usługami i nowymi warunkami współpracy z firmą),
 - medycyna (np. analiza efektywności procedur leczenia pacjentów)

Pytania

- Dane przechowywane w bazie danych zawierają olbrzymią ilość potencjalnie użytecznej wiedzy, która może zostać użyta w procesie podejmowania decyzji strategicznych dotyczących działalności przedsiębiorstwa:
 - Czym różnią się klienci supermarketu w Poznaniu i Warszawie?
 - Jakie oddziały supermarketu miały „anormalną” sprzedaż w pierwszym kwartale 2002 r?
 - Jakie produkty miały największą dynamikę sprzedaży w roku 2001?
 - Jakie produkty klienci supermarketu kupują najczęściej razem?

Architektura danych



Magazyn (hurtownia) danych

Magazyn danych jest „(...) zorientowaną tematycznie, zintegrowaną, zmienną w czasie i trwałą, kolekcją (bazą) danych zaprojektowaną i zaimplementowaną dla potrzeb wspomagania podejmowania decyzji, w której dane odnoszą się do określonej chwili czasowej”

-- (W. H. Inmon, *Building the Data Warehouse*, QED Tech. Pub. Group, 1992)

Magazyn danych

- **Zorientowany tematycznie** - struktura danych w magazynie danych jest zorganizowana odpowiednio do podstawowego obszaru działalności danego przedsiębiorstwa: klienci, typy ubezpieczeń, polisy, konta, żądania wypłat, itp.
- **Zintegrowany** - magazyn danych musi zawierać możliwie pełny zbiór danych opisujących działalność danego przedsiębiorstwa; dane opisujące działalność przedsiębiorstwa są najczęściej rozproszone niezbędna staje się **integracja danych** z wielu heterogenicznych źródeł.

Magazyn danych

- **Trwały** - dane operacyjne są regularnie aktualizowane i zmieniane; magazyny danych są natomiast **trwale** - po załadowaniu danych do magazynu, dane nie są z magazynu usuwane. Po dezaktualizacji dane są archiwizowane
- **Zmienny w czasie** - horyzont czasowy magazynu danych jest znacząco większy niż horyzont czasowy operacyjnych baz danych. Magazyny danych przechowują całą historię danych (czyli zbiór migawek zrobionych w pewnych odstępach czasowych) i czas stanowi zawsze jeden z podstawowych elementów składowych magazynu danych

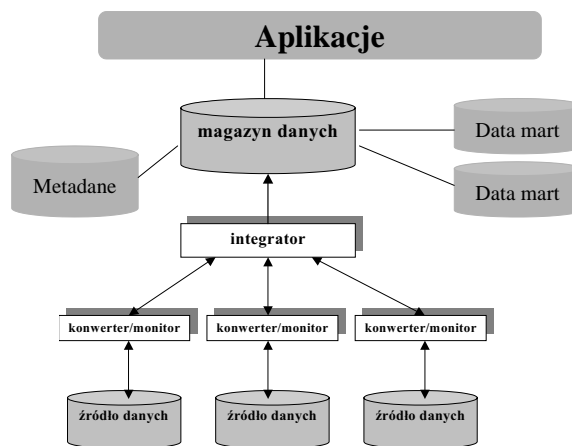
Dlaczego niezależny magazyn danych

- **Względy funkcjonalne:**
 - brakujące dane: systemy DCS wymagają danych historycznych, których systemy OLTP nie przechowują
 - integracja danych: systemy DCS wymagają integracji i agregacji danych z różnych heterogenicznych źródeł
 - jakość danych: różne źródła stosują różne reprezentacje danych, kody, formaty, nazewnictwo

Dlaczego niezależny magazyn danych

- **Względy efektywnościowe:**
 - Złożone zapytania OLAP znacząco obniżają efektywność przetwarzania transakcyjnego
 - Specjalne struktury danych, metody dostępu, materializowane perspektywy i agregaty, specjalne metody implementacji operacji wspierających wielowymiarowy model danych
 - Brak mechanizmu transakcji, zarządzania współbieżnością, odtwarzania po awarii

Architektura



Implementacja i pielęgnacja magazynu

- Ekstrakcja danych
- Transformacja danych
- Czyszczenie danych
- Integracja danych
- Ładowanie danych
- Monitorowanie zmian
- Odświeżanie danych
- Metadane i ich repozytorium

Implementacja i pielęgnacja magazynu (1)

- **Ekstrakcja danych**
pobieranie danych ze źródeł danych (bramki, standardowe interfejsy, procedury własne, mechanizm replikacji)
- **Konwersja danych**
transformowanie danych z formatu wykorzystywanego w źródle, do formatu wykorzystywanego w magazynie

Implementacja i pielęgnacja magazynu (2)

- **Czyszczenie danych**

proces ten ma na celu zapewnienie jakości i poprawności danych w magazynie (dane z wielu źródeł będą zawierały błędy i anomalie: niespójne długości pól, niespójne opisy atrybutów, różne formaty danych, wartości puste, naruszone ograniczenia integralnościowe; źródłem niespójności są często pola opcjonalne)

- **Metody czyszczenia danych**

- **Migracja danych:** proste reguły transformacji danych, np. „zastąp słowo customer słowem klient”
- **Czyszczenie specjalne:** wykorzystanie wiedzy przedmiotowej do czyszczenia danych (np. kody pocztowe)
- **Śledzenie danych:** wykorzystanie technik eksploracji danych do czyszczenia danych (detect outliers)

Implementacja i pielęgnacja magazynu (3)

- **Ładowanie danych** - ładowanie danych pociąga za sobą dodatkowe przetwarzanie: sprawdzanie ograniczeń integralnościowych, sortowanie, podsumowywanie, budowanie indeksów, itp..
- **Metody ładowanie:**
 - Wsadowe
 - Inkrementalne
- **Problemy:**
 - monitorowanie stanu ładowania, wstrzymanie ładowania, zmiana ziarna ładowania, anulowanie aktualizacji
 - ładowanie sekwencyjne/równoległe
 - restart po awarii
 - wsadowe/inkrementalne

Implementacja i pielęgnacja magazynu (4)

- **Monitorowanie zmian**

monitorowanie zmian zachodzących w źródłach danych, istotnych z punktu widzenia magazynu danych

- mechanizm wyzwalaczy (trigger) DBMS
 - analiza pliku log (analiza zawartości dziennika)
 - mechanizm replikacji danych
 - procedury własne (tzw. legacy systems)
 - polling (zapytania do źródeł)
- Zmiany w danych źródłowych są propagowane do magazynu danych podczas procesu odświeżania

Implementacja i pielęgnacja magazynu (4)

- **Odświeżanie danych** - proces propagowania zmian zachodzących w źródłach danych do magazynu
- Kiedy odświeżać:
 - odświeżanie natychmiastowe
 - okresowe
 - zależnie od źródła (np. w momencie dostępu przez użytkownika)
- W jaki sposób odświeżać:
 - ładowanie danych (full loading)
 - odświeżanie inkrementalne
- Mechanizm pielęgnacji replik
 - transfer danych
 - transfer transakcji

Metadane

- Dane o danych
- Stanowią integralną część magazynu danych
- Określają znaczenie i kontekst informacji zawartej w magazynie danych
- Jakie dane są dostępne, gdzie są zlokalizowane, oraz w jaki sposób są dostępne
- Metadane są przechowywane w różnej postaci: arkusze kalkulacyjne, CASE, dokumenty tekstowe

Repozytorium metadanych

- **metadane fizyczne:** lista źródłowych baz danych i opis ich zawartości, opisy i charakterystyki bramek między bazami źródłowymi a magazynem, schemat magazynu danych, definicje perspektyw i danych wyliczalnych, opisy wymiarów i hierarchii, zbiór predefiniowanych zapytań i raportów, lokalizacja tematycznych hurtowni danych, indeksy i reguły partycjonowania danych
- **metadane logiczne:** reguły biznesowe, podstawowe pojęcia i definicje, procedury postępowania, logiczne definicje tablic i atrybutów magazynu danych, odwzorowanie danych operacyjnych na struktury magazynu danych)

Repozytorium metadanych

- **metadane operacyjne:** reguły ekstrakcji, czyszczenia, transformacji, korekcji danych źródłowych, zasady odświeżania danych, dane szczegółowe i dane wyprowadzalne
- **metadane historyczne:** zmiany zachodzące w środowisku magazynu danych, informacja dotycząca aliasów
- **metadane administracyjne:** bezpieczeństwo magazynu, autoryzacja użytkowników, prawa dostępu do poszczególnych komponentów magazynu, profile użytkowników i profile grup użytkowników
- **metadane personalizacyjne:** reguły obliczania pewnych agregatów dla określonych użytkowników końcowych lub grup użytkowników

OLAP

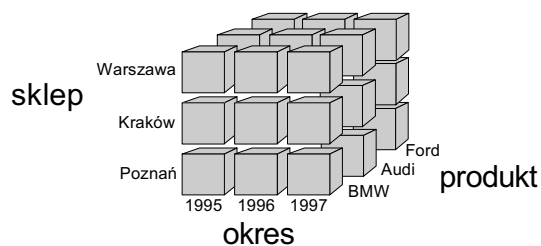
- **Przetwarzanie analityczne on-line** (ang. On-Line Analytical Processing OLAP), ma za zadanie wspieranie procesów analizy magazynów danych
- Analiza magazynu polega na obliczaniu agregatów dla zadanych „wymiarów” magazynu
- Logiczny model danych:
 - struktury danych, która opisują logiczną organizację danych i sposób, w jaki dane są postrzegane przez użytkowników,
 - zbioru operatorów umożliwiających wyszukiwanie i modyfikowanie danych, oraz
 - ograniczeń integralnościowych, specyfikujących poprawność danych

Wielowymiarowy model danych

- Podstawowy model logiczny dla MDD/OLAP
- Dane są postrzegane przez użytkowników w postaci **wielowymiarowej perspektywy** (tzw. kostki OLAP)
- Obiektem analizy w modelu MDD jest zbiór **miar numerycznych** nazywanych **faktami**
- **Fakt** opisuje pojedyncze zdarzenie, o którym chcemy przechowywać informację w magazynie danych
- Fakt jest daną ilościową (numeryczną) reprezentującą jednostkę aktywności biznesowej przedsiębiorstwa, np. sprzedaż produktów, średnia ocena studenta, ilość gości hotelowych, zysk, wartość produktu krajowego, itp.

Wielowymiarowy model danych

- Wartość każdej miary zależy od zbioru wymiarów
- W modelu MDD, miara jest reprezentowana jako punkt w wielowymiarowej przestrzeni wymiarów

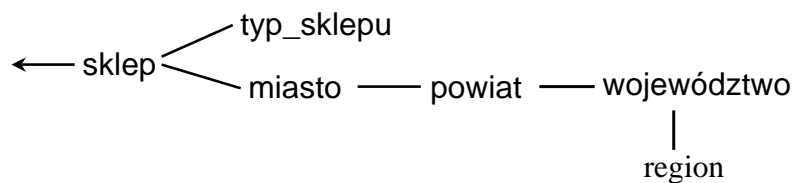


Wielowymiarowy model danych

- Każdy wymiar jest opisany **zbiorem atrybutów**

Sklep (Id_sklep, sklep, adres, miasto, powiat, województwo, region, typ_sklepu, telefon, szef)

- Atrybuty wymiaru mogą tworzyć **hierarchię wymiaru**



Operacje modelu MDD

- **Agregacja** – łączna sprzedaż dla poszczególnych sklepów w poszczególnych latach
- **Pivoting – wyznaczenie punktu centralnego**: wskazanie miary i wybranie 2 wymiarów, w których ma ona być reprezentowana (sprzedaż dla sklepów w poszczególnych latach)
- **Roll-up – zwijanie**: dla wskazanego wymiaru nawigacja w górę hierarchii wymiaru w celu prezentacji większych agregatów
- **Drill-down – rozwijanie**: nawigacja wzdłuż hierarchii danego wymiaru w celu rozbicia agregatu na agregaty składowe

Operacje modelu MDD

- **Slice_and_dice – wycinanie**: operacja redukcji liczby wymiarów, tj. projekcja danych na wybranym podzbiore wymiarów dla wybranych wartości innych wymiarów
- **Rotating - obracanie**: umożliwia prezentowanie danych w różnych układach
- **Ranking** – wybór pierwszych n elementów
- Nowe operatory:
 - Pull – utwórz nowy wymiar z istniejących elementów
 - Destroy – usuń wymiar
 - Restrict – usuń wartości z kostki
 - Join – połącz informacje z dwóch kostek

Ograniczenia integralnościowe

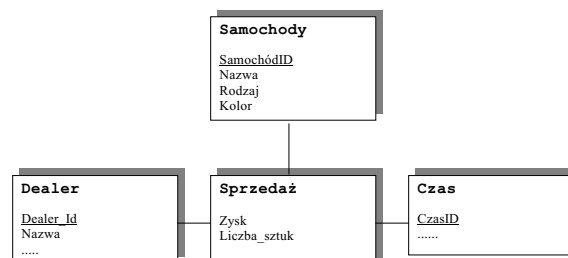
- **Ograniczenia integralnościowe pojedynczej kostki danych** (ang. intra cube constraints)
związane z definicjami zależności pomiędzy atrybutami wymiarów, wymiarami, wymiarami a miarami, oraz hierarchiami wymiarów
- **Ograniczenia integralnościowe pomiędzy kostkami danych** (ang. inter cube constraints)
określają związki pomiędzy dwoma lub więcej kostkami danych, tj. związki pomiędzy wymiarami dwóch kostek, miarami kostek, miarą jednej kostki a wymiarami innej kostki, itp..

Projektowanie schematów pojęciowych magazynów danych

- Do zaprojektowania schematu pojęciowego można wykorzystać dowolny z modeli pojęciowych wykorzystywanych do projektowania schematów pojęciowych baz danych
- Schemat pojęciowy magazynu danych powinien:
 - koncentrować się na podstawowych pojęciach i dziedzinach aktywności danego przedsiębiorstwa
 - powinien być łatwo transformowalny do wielowymiarowego modelu danych
- Podstawowe struktury schematów pojęciowych – schemat gwiazdy, płątka śniegu, konstelacji faktów

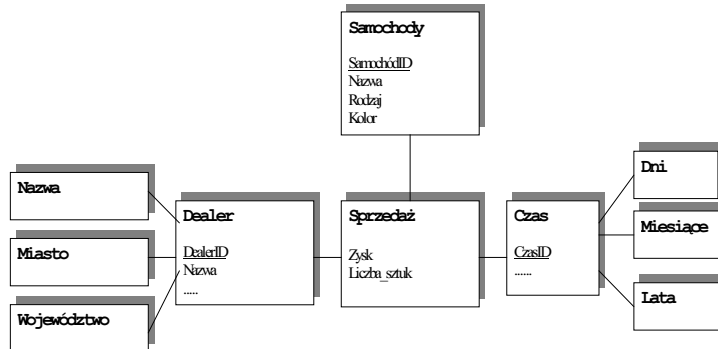
Struktura gwiazdy

- *Struktura gwiazdy* (ang. star schema) - centralna encja opisuje podstawową miarę (zbiór miar), która jest powiązana z encjami wymiarów



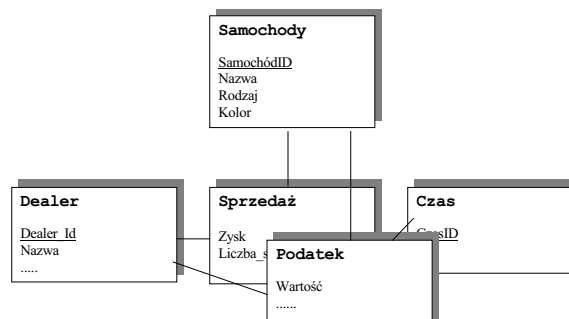
Struktura płatka śniegu

- *Struktura płatka śniegu* (ang. snowflake schema) - zmodyfikowana wersja struktury gwiazdy, w której explicite zamodelowane są hierarchie wymiarów



Struktura konstelacji faktów

- *Struktura konstelacji faktów* (ang. fact constellation schema) - zbiór encji faktów współdzieli zbiór encji wymiarów, choć niekoniecznie na tym samym poziomie hierarchii tych wymiarów



Typy magazynów danych

- W jaki sposób wielowymiarowy model danych jest przechowywany i przetwarzany w magazynie danych?

Dwa podejścia (zależnie od modelu danych)

1. Magazyn danych wykorzystujący *model relacyjny*, nazywany również **ROLAP** (ang. Relational OLAP)
2. Magazyn danych wykorzystujący *model wielowymiarowy*, nazywany również **MOLAP** (ang. Multidimensional OLAP)

ROLAP

- Dane są przechowywane w specjalizowanych relacjach
- Schemat logiczny magazynu ROLAP odpowiada strukturze schematu pojęciowego (centralna relacja faktów powiązana kluczami obcymi z odpowiednimi relacjami wymiarów)
- W przypadku schematu płatka śniegu, relacje wymiarów są znormalizowane - wyodrębnienia hierarchii wymiaru
- Charakteryzuje się dużą skalowalnością i elastycznością
- W stosunku do magazynów typu MOLAP cechują się niższą efektywnością przetwarzania danych

MOLAP

- Dane przechowywane w specjalizowanych *wielowymiarowych tablicach* (ang. multidimensional arrays) zwanych też *kostkami danych* (ang. data cubes)
- Pozycja komórki wielowymiarowej tablicy jest wyznaczona przez kombinację wartości odpowiednich wymiarów
- Tablice zawierają również wstępnie przetworzone, tj. zagregowane dane
- Kostki danych są tworzone przed rozpoczęciem przetwarzania i mają charakter statyczny
- Charakteryzują się wysoką efektywnością wielowymiarowego przetwarzania danych, jednakże, w stosunku do magazynów typu ROLAP, cechują się gorszą skalowalnością i elastycznością

Efektywność magazynów danych

- W celu poprawy efektywności działania magazynów danych stosuje się wiele technik:
 - materializowanie agregatów,
 - przetwarzanie równoległe,
 - partycjonowanie danych
 - indeksowanie danych

Indeksowanie danych

- Indeksowanie danych polega na łączeniu wartości indeksowanego atrybutu z adresami fizycznych bloków dyskowych, w których przechowywane są rekordy o danej wartości
- Poprawiają znacząco czas dostępu do danych
- Magazyn danych jest statyczny (dominują odczyty)
- Definiuj indeksy na kluczu podstawowym i kluczach obcych – zawsze!
- Nowe typy indeksów:
 - Indeks bitmapowy
 - Indeks połączeniowy

Indeks bitmapowy

- Dla każdej unikalnej wartości atrybutu jest przechowywana *mapa bitowa*
- Każdy bit mapy odpowiada jednej krotce relacji R
- Dla mapy $A='w'$ bit n przyjmuje wartość jeden, jeśli atrybut A krotki o numerze n przyjmuje wartość 'w', w przeciwnym przypadku bit n przyjmuje wartość zero
- Indeks bitmapowy jest zbiorem map bitowych
- Indeks bitmapowy posiada strukturę B–drzewa, w którego liściach zamiast adresów rekordów są przechowywane mapy bitowe

Indeks bitmapowy

Sprzedaż			kolor	
klientID	marka	Kolor	zielony	niebieski
1010	Fiat	zielony	1	0
1020	BMW	niebieski	0	1
1030	Fiat	zielony	1	0
1040	Audi	zielony	1	0
1050	Volvo	zielony	1	0
1060	Fiat	niebieski	0	1
1070	Ford	niebieski	0	1
1080	Opel	zielony	1	0
1090	Opel	niebieski	0	1
1100	Ford	zielony	1	0

Indeks połączeniowy

- **Indeks połączeniowy** (ang. join index) łączy z sobą krotki z różnych relacji posiadające tę samą wartość atrybutu połączeniowego (jest więc strukturą zawierającą zmaterializowane połączenie wielu relacji)
- Indeks połączeniowy posiada strukturę B–drzewa zbudowanego na atrybucie połączeniowym relacji
- Dla magazynu danych o strukturze gwiazdy indeks połączeniowy wiąże krotki relacji wymiaru (lub wymiarów) z krotkami relacji faktów
- **Bitmapowy indeks połączeniowy** (ang. bit–mapped join index) - w liściach zamiast adresów krotek znajdują się mapy bitowe opisujące krotki łączonych relacji

Indeks połączeniowy

product	id	name	price	jindex
	p1	bolt	10	r1,r3,r5,r6
	p2	nut	5	r2,r4

sale	rld	prold	storeld	date	amt
	r1	p1	c1	1	12
	r2	p2	c1	1	11
	r3	p1	c3	1	50
	r4	p2	c2	1	8
	r5	p1	c1	2	44
	r6	p1	c2	2	4

Materializacja perspektyw

- Wstępne przeprowadzenie obliczeń i zmaterializowanie otrzymanych wyników w magazynie danych w celu ich późniejszego wykorzystania
- Materializacja agregatów oraz perspektyw
- Dwa zasadnicze pytania:
 - (1) które z agregatów materializować, a które agregaty pozostawić do obliczeń w trybie on-line,
 - (2) w jaki sposób pielęgnować zmaterializowane agregaty (ponowne obliczanie agregatów, inkrementalna pielęgnacja agregatów)
- Czy materializować pośrednie wyniki obliczeń (nie tylko agregaty), np. wyniki niektórych operacji połączeń, które są wspólne dla wielu agregatów?

Selekcja i pielęgnacja materializowanych perspektyw

- Redukcja czasu odpowiedzi i zajętości pamięci
- Wybór perspektyw, które należy zmaterializować, zależy od charakterystyki obciążenia, częstości określonych zapytań, kosztu przechowywania i aktualizacji perspektyw
- Zaproponowano w literaturze szereg heurystyk
- Dane są aktualizowane (w ciągu roku wzrastają dwukrotnie)
- W jaki sposób pielęgnować:
 - Wylizanie od początku
 - Pielęgnacja inkrementalna
- Materializacja wyników pośrednich (deficytowy instytut)

Przetwarzanie równoległe

- *Przetwarzanie równoległe* (ang. parallel processing) polega na rozbiciu złożonych operacji na mniejsze, które następnie są wykonywane równoległe, np. na wielu procesorach lub komputerach
- Równoległe przetwarzanie zapytań, sortowanie danych, operacje odczytu i zapisu na dysk, budowa relacji i indeksów, ładowanie danych do magazynu danych

Partycjonowanie danych

- **Partycjonowanie danych** (ang. data partitioning) polega na automatycznym rozpraszaniu danych (pochodzących z jednej lub wielu relacji) na wielu dyskach, znajdujących się w tym samym lub wielu węzłach (komputerach) sieci
- Zyski:
 - (1) bardzo kosztowne operacje wejścia/wyjścia, mogą być wykonywane równolegle,
 - (2) równoważone jest obciążenie dysków,
 - (3) polecenia SQL mogą być wykonywane równolegle, np. tworzenie relacji i indeksów, wykonywanie zapytań,
 - (4) wzrasta bezpieczeństwo danych w przypadku awarii sprzętu,
 - (5) wzrasta szybkość tworzenia kopii zapasowych magazynu danych i szybkość odtwarzania danych po awarii.

Wnioski

- Magazyn danych jest nie jest produktem ani też aplikacją
- Jest to architektura przetwarzania danych opracowana z myślą o budowie systemów wspomagania podejmowania decyzji
- **Jakie problemy pozostają nadal nierozwiązane lub wymagają nowych rozwiązań w zakresie technologii magazynów danych?**
- **Problem aktualizacji wymiarów i ewolucji schematu magazynu danych** – temporalne i wielowersyjne magazyny danych
- Narzędzia i techniki akwizycji danych (czyszczenie danych, rozwiązywanie niespójności danych)

Wnioski

- Optymalizacja zapytań
- Algorytmy selekcji i pielęgnacji materializowanych perspektyw
- Narzędzi do zarządzania metadanymi
- Technikami odtwarzania magazynu danych po awarii w czasie procesu ładowania i odświeżania danych
- Technikami automatycznego archiwizowania danych w momencie ich dezaktualizacji

Technologia magazynów danych jest ciągle jeszcze technologią na etapie rozwoju