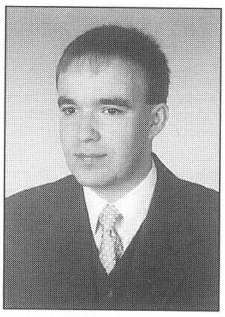
Bartosz JĘDRZEJEC, Krzysztof ŚWIDER
POLITECHNIKA RZESZOWSKA, KATEDRA INFORMATYKI I AUTOMATYKI

Wielowymiarowa analiza danych w trybie czasu rzeczywistego

Mgr inż. Bartosz JĘDRZEJEC

Studia na wydziale Wydziale Elektrotechniki i Informatyki Politechniki Rzeszowskiej ukończył w 2001 roku. Pracuje na stanowisku asystenta w Katedrze Informatyki i Automatyki Politechniki Rzeszowskiej. Jego główną dziedziną zainteresowań są hurtownie danych, analiza danych i pozyskiwanie wiedzy z baz danych.

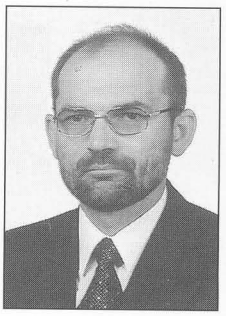
e-mail: bartoszj@prz-rzeszow.pl



Dr inż. Krzysztof ŚWIDER

Dyplom magistra inżyniera uzyskał w 1981 roku na Wydziale Systemów Sterowania Politechniki Kijowskiej, a stopień doktora nauk technicznych w 1993 roku na Wydziale Elektroniki Politechniki Warszawskiej. Zajmuje się problematyką zarządzania dużymi zbiorami danych, metodami integracji danych pochodzących z niejednorodnych źródeł oraz pozyskiwaniem wiedzy z baz danych.

e-mail: kswider@prz-rzeszow.pl



Streszczenie

W pracy rozważono systemy wielowymiarowej analizy danych wykorzystujące technologie hurtowni danych oraz przetwarzania analitycznego on-line (OLAP). Szczególną uwagę zwrócono na specyfikę zastosowań dla danych zmieniających się z dużą częstotliwością (real-time OLAP). Jako przykład aplikacji przedstawiono możliwości analizy OLAP w trybie czasu rzeczywistego oferowane przez MS SQL Server Analysis Services.

Abstract

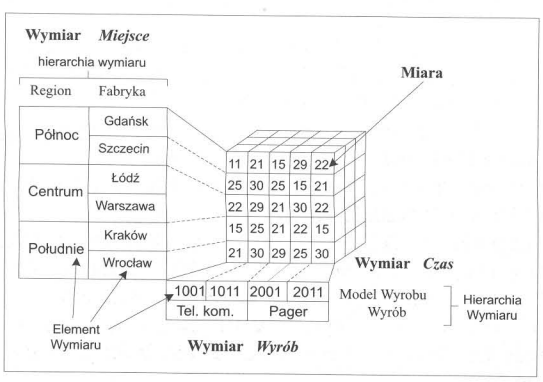
The paper concerns multi-dimensional data analysis based on data warehousing and On-Line Analytical Processing (OLAP) technologies. Especially we focused on applications with relative high frequency of data changes (real-time OLAP). As an illustrative example, the real-time OLAP functionalities of MS SQL Server Analysis Services are presented.

Słowa kluczowe: analiza danych, hurtownia danych, OLAP, tryb czasu rzeczywistego.

Keywords: data analysis, data warehouse, OLAP, real-time mode.

1. Wstęp

Istotnym problemem w systemach automatycznego przetwarzania danych jest ich wydajność, rozumiana jako możliwość wykorzystania dużych baz danych, przechowujących złożone struktury, przy zachowaniu zadowalającego czasu odpowiedzi. Powszechnie stosowana architektura klient-serwer zapewnia dostęp do wyspecjalizowanych serwerów, które są optymalizowane pod kątem konkretnych potrzeb użytkowników. Aplikacje takie jak programy do analizy rynku i prognozowania zjawisk finansowych oraz inne systemy wspomagające podejmowanie decyzji często wymagają modeli danych zorientowanych na zapytania o charakterze wielowymiarowym. Struktury wielowymiarowe tworzą tzw. kostki danych (data cubes) (rys.1) [1,2].



Rys. 1. Przykładowa struktura kostki danych
Fig. 1. An example structure of data cube

Umieszczone wewnątrz kostki wartości reprezentują dane ilościowe, oznaczające np. liczbę sprzedanych produktów należących do analizowanej grupy. Wielkości te są określane jako miary i „układane" według istotnych cech semantycznych, takich jak np.: data i miejsce transakcji, rodzaj produktu, forma transakcji itp., nazywanych wymiarami. Wymiary tworzą zazwyczaj hierarchie szczegółowości wykorzystywane w analizie. Na przykład dla wymiaru Czas można określić hierarchię: rok-kwartał-miesiąc-dzień.

Wykorzystanie kostek danych jest typowe dla aplikacji zaliczanych do klasy OLAP (On-Line Analytical Processing - analityczne przetwarzanie on-line). Wielowymiarowy model danych można stosunkowo łatwo modyfikować; np. kostka danych może zostać rozszerzona o kolejny wymiar. Aplikacje OLAP umożliwiają ponadto wykonywanie operacji specyficznych dla danych wielowymiarowych, takich jak: obrót (spojrzenie na dane z różnych stron kostki), przekrój (ograniczanie zakresu wymiarów) oraz tzw. rozwijanie i zwijanie wymiaru w ramach hierarchii (rys.2).

Sprzedaż (tys.szt.)		Telefon kom.		Pager	
		1001	1011	2001	2011
Południe	Kraków	33	12	8	12
	Wrocław	45	34	20	23

Rozwijanie wymiaru Miejsce

Sprzedaż (tys.szt.)		Telefon kom.		Pager	
		1001	1011	2001	2011
Kraków	Sklep 1	20	8	6	7
	Sklep 2	13	4	2	5

Sprzedaż (tys.szt.)		2002			
		Kw. 1	Kw. 2	Kw. 3	Kw. 4
Południe	Kraków	78	45	34	56
	Wrocław	90	67	87	91

Zwijanie wymiaru Czas

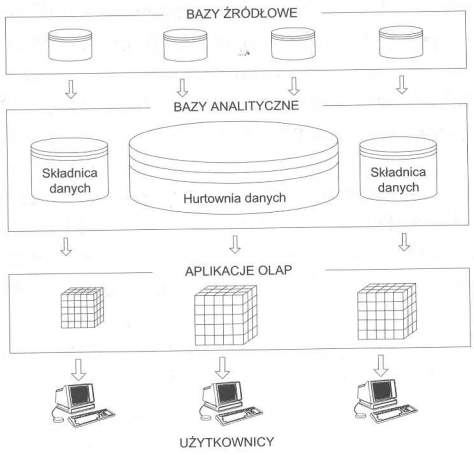
Sprzedaż (tys.szt.)		Kwartał 1		
		Styczeń	Luty	Marzec
Południe	Kraków	30	26	22
	Wrocław	28	30	32

Rys. 2. Rozwijanie i zwijanie wymiarów w ramach hierarchii pozwala odpowiednio zwiększać lub zmniejszać stopień szczegółowości widzenia danych.
Fig. 2. Drill-down and roll-up are the operations for moving the view down and up along the dimensional hierarchy levels.

Ponadto dla danych zorganizowanych wielowymiarowo jest możliwe osiąganie większej efektywności w realizacji zapytań, niż dla modeli relacyjnych. Uzyskuje się to poprzez wstępną agregację danych i zapisanie odpowiednich podsumowań w bazie danych. Ze względu na sposób przechowywania analizowanych danych systemy OLAP zostały podzielone na kategorie, w zależności od architektury bazy danych, na której pracują. W tym aspekcie wyróżnia się trzy typy OLAP: wielowymiarowy (MOLAP), relacyjny (ROLAP) oraz hybrydowy (HOLAP) [3].

Dane dla aplikacji OLAP są często gromadzone w hurtowniach danych (data warehouses) oraz składnicach danych (data marts) przystosowanych do przechowywania dużych ilości danych do analizy (rys. 3). Hurtownie obejmują zwykle globalne potrzeby informacyjne systemu, podczas gdy składnice danych przechowują dane w pewnym zakresie tematycznym, np. dla konkretnego działu firmy.

Dane do hurtowni i składnic danych są pobierane z baz operacyjnych i poddawane niezbędnym przekształceniom. Zasadniczym celem wstępnej obróbki danych jest ich oczyszczenie, uzyskanie spójności, uzupełnienie brakujących wartości itp., tak, aby mogły być później wykorzystane do budowy aplikacji OLAP udostępnianych użytkownikom.



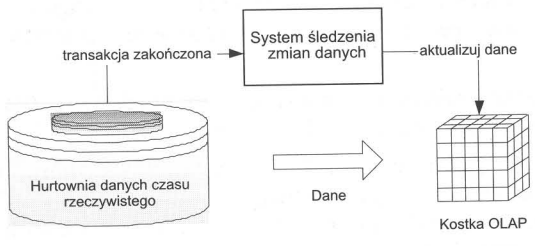
Rys. 3. Hurtownie danych i OLAP
Fig. 3. Data warehouses and OLAP

Typowym schematem danych dla aplikacji OLAP jest schemat gwiazdy. Jest to logiczna struktura składająca się z usytuowanej centralnie tabeli danych zawierającej jedną lub więcej wielkości mierzalnych (miary) oraz tabel definiujących wymiary. W praktyce tablice faktów i wymiarów są ze sobą powiązane za pomocą mechanizmu kluczy obcych, stosowanego powszechnie w relacyjnych bazach danych.

2. Analiza w trybie czasu rzeczywistego

Podstawą analiz w typowych zastosowaniach ekonomicznych OLAP są zwykle dane historyczne, które są periodycznie zbierane z operacyjnych baz danych oraz przetwarzane i agregowane. Cykl, w zależności od zastosowań, może się wahać od 1 miesiąca do 1 dnia [4]. Rosnące wymagania wobec aktualności uzyskiwanych wyników spowodowały zapotrzebowanie na systemy analityczne działające w trybie czasu rzeczywistego. Chodzi o uwzględnianie na bieżąco zmian takich wielkości jak np. kursy akcji na giełdzie czy kursy walut na podstawie danych z ostatnich kilkunastu minut. Wykorzystanie systemów tej klasy jest możliwe także w innych dziedzinach, np. firma Federal Express wykorzystuje hurtownie danych czasu rzeczywistego i analizę danych do podejmowania decyzji w zarządzaniu transportem, tak, aby przesyłki docierały w określonym czasie. Podjęcie odpowiedniej decyzji jest uzależnione od stopnia aktualności danych i możliwości ich interpretacji w jak najszybszym czasie [5].

W rozbudowanych systemach może wystąpić pewna ilość atrybutów, dla których dane są pobierane na bieżąco, poddawane ewentualnej filtracji i zapisywane w hurtowni danych. W tym przypadku stosuje się zazwyczaj system śledzenia zmian danych, który reaguje na sygnał zakończenia wykonywania transakcji w hurtowni, wysyłając sygnał potrzeby aktualizacji danych dla kostek OLAP (rys. 4).



Rys. 4. Hurtownia danych i analiza OLAP w trybie czasu rzeczywistego
Fig. 4. Real-time data warehouse and real-time OLAP analysis

Jednym z problemów napotykanych przy wdrożeniu OLAP czasu rzeczywistego jest częstotliwość uaktualniania danych. Powinno ono odbywać się bez przerywania możliwości dostępu do systemu, po każdej zmianie wartości atrybutu. W tradycyjnych systemach OLAP uaktualniania danych dokonuje się w czasie, gdy system jest niewykorzystywany np. w nocy czy w dni świąteczne. W przypadku procesów, w których uaktualnianie danych powinno być natychmiastowe i ciągłe, może nastąpić nadmierne obciążenie bazy, a co za tym idzie - uniemożliwienie prowadzenia analiz. Innym problemem, jaki napotykają twórcy systemów OLAP tej klasy jest aktualizacja danych. Ponieważ większość danych w procesach czasu rzeczywistego jest powiązana z atrybutem określającym czas, stąd system aktualizacji danych dla aplikacji OLAP musi mieć możliwość efektywnego przekształcania danych czasowych zapisywanych hierarchicznie. Istotną kwestią staje się także zadawanie zapytań OLAP, które w systemach tradycyjnych są przewidziane dla danych nie zmieniających się, tj. danych historycznych. W systemach, gdzie dane zmieniają się z dużą częstotliwością, muszą istnieć odpowiednie zabezpieczenia zapewniające poprawność zapytań. Z reguły stosuje się odseparowanie części transakcyjnej systemu od części związanej z hurtowniami danych dla potrzeb analizy. Dzieje się tak z powodu znacznej złożoności zapytań analitycznych. Zapytania te nie są wystarczająco wydajne, szczególnie, jeśli w tym samym czasie wykonywanych jest wiele innych zadań takich jak: wstawianie, aktualizacja oraz usuwanie danych [6,7].

3. Opcja real-time OLAP w MS SQL Server

Coraz większe zapotrzebowanie na systemy przetwarzania i analizowania danych szybkozmiennych wymusiło na producentach oprogramowania opracowanie nowych lub dostosowanie istniejących produktów do potrzeb systemów wymagających analizy w trybie czasu rzeczywistego. Jednym z produktów oferujących tej klasy narzędzie analityczne jest MS SQL Server Analysis Services firmy Microsoft. Produkt ten udostępnia kilka mechanizmów i obiektów wspierających analizę danych zmieniających się z dużą częstotliwością. Wykonywanie analiz w trybie czasu rzeczywistego wymaga utworzenia tzw. wymiaru czasu rzeczywistego lub kostki czasu rzeczywistego. Wymiary czasu rzeczywistego są zoptymalizowane pod względem częstej ich aktualizacji, co oznacza możliwość dokonywania zmian bez potrzeby dodatkowego przetwarzania wymiaru lub kostki zawierającej ten wymiar. Przetwarzanie takie blokowałoby na pewien czas dostęp aplikacji analitycznych do wymiarów i kostek czasu rzeczywistego.

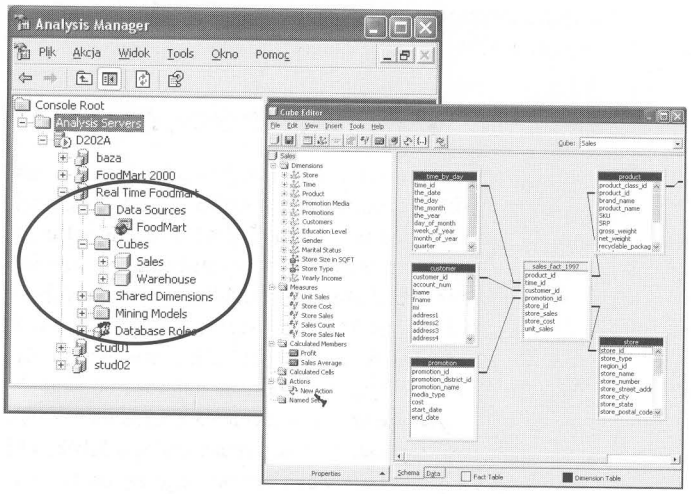
Wprawdzie wymiary czasu rzeczywistego umożliwiają nieprzerwany dostęp do aktualnych danych, lecz odbywa się to kosztem dłuższego czasu przetwarzania zapytań. Stąd wymiary tego typu mogą być zapisywane tylko w obrębie partycji ROLAP.

Agregacje danych nie są wtedy tworzone, ponieważ częste ich odświeżanie mogłoby spowodować znaczny spadek wydajności, a w skrajnych przypadkach, nawet zablokowanie dostępu do danych. Niestety brak agregacji spowalnia z kolei wykonywanie zapytań analitycznych, a co za tym idzie, powoduje mniejszą efektywność systemu. Aby rozwiązać ten problem, generowane są perspektywy indeksowane dla przechowywania agregacji. Dla każdej agregacji jest tworzona oddzielna perspektywa, przechowywana wewnątrz partycji przystosowanej do aktualizacji danych w czasie rzeczywistym.

Dla każdego obiektu wspierającego aktualizacje w czasie rzeczywistym system śledzenia w MS SQL Server tworzy zdarzenie powiadamiania dla tabel powiązanych z tym obiektem. Uruchomiony jest także wątek nasłuchujący, który odbiera informacje o takich zdarzeniach. Zdarzenie powiadamiania wywoływane jest dopiero po zatwierdzeniu całej transakcji. Jest to dość istotne, ponieważ w celu zoptymalizowania wykonywania zapytań do bazy danych istnieje możliwość grupowania wzajemnie nie wykluczających się operacji w pojedynczą transakcję. W przypadku otrzymania informacji o zdarzeniu z tabeli bazy danych, wątek nasłuchujący powiadamia serwer analiz o dezaktualizacji danych w pamięci podręcznej serwera dla obiektów zależnych od tej tabeli. Serwer analiz

aktualizuje odpowiednie informacje, a dopiero po takiej aktualizacji żądanie danych z aplikacji może zostać zrealizowane. W przypadku zapytań o dane, porównywana jest najpierw pamięć podręczna aplikacji (klienta) z pamięcią podręczną serwera analiz. W przypadku zgodności wersji, aplikacja pobiera dane z komputera lokalnego bez potrzeby komunikacji z serwerem. W przypadku stwierdzenia rozbieżności, dane w pamięci podręcznej klienta są oznaczane jako nieaktualne i pozostają takimi, dopóki nie zostaną zsynchronizowane z serwerem. Synchronizacja danych pomiędzy serwerem a klientem odbywa się periodycznie w ustalonych momentach, nawet, gdy nie zostało zadane żadne zapytanie. Pozwala to unikać ewentualnego problemu z dużą ilością danych do aktualizacji w przypadku niezgodności.

Przyjęte rozwiązania są ukierunkowane na zwiększenie szybkości działania systemu. Pamiętać jednak należy, że system, który ma działać sprawnie i wydajnie, musi zostać zaplanowany w sposób rozsądny i przemyślany. Zwiększając ilość elementów powiązanych z danymi czasu rzeczywistego, należy liczyć się ze spadkiem wydajności, dlatego zalecane jest, aby obiektów typu real-time używać tylko w przypadkach uzasadnionych [8].



Rys. 5 Konsola zarządzania aplikacjami analitycznymi w MS SQL Server
Fig. 5 Analysis Manager console in MS SQL Server environment

Rysunek 5 przedstawia wygląd konsoli operatora narzędzi Analysis Services w systemie MS SQL Server. W szczególności umożliwia ona tworzenie i zarządzanie kostkami OLAP, zarówno dla analiz typowych, jak też takich, które mają być prowadzone w czasie rzeczywistym. Podczas konstruowania kostek OLAP, użytkownik może ustalić sposób, w jaki będą przechowywane dane, dokonując wyboru odpowiedniego typu partycji. Ponadto jest możliwe określenie ilości agregacji, a także zadeklarowanie utworzenia wymiarów typu real-time.

4. Wsparcie dla aplikacji OLAP

Po utworzeniu kostki można analizować dane przy pomocy wbudowanego narzędzia BrowseData oraz arkusza kalkulacyjnego MS Excel. W tym drugim przypadku można używać tabel i wykresów przestawnych, które umożliwiają wykonywanie typowych operacji dla danych wielowymiarowych oraz odpowiednią wizualizację wyników.

Analizę można prowadzić w sposób zdalny, dzięki możliwości podłączenia się do serwera OLAP poprzez usługę PivotTable, pełniącą rolę sterownika, zarządzającego połączeniem pomiędzy klientem a serwerem [9]. Do podstawowych korzyści uzyskanych w wyniku zastosowania PivotTable należą m.in.: większa efektywność systemu, możliwa dzięki ograniczeniu ruchu w sieci oraz dostęp do serwera analiz dla aplikacji internetowych.

Dość typowym problemem, z jakim mogą zetknąć się użytkownicy, jest pilna potrzeba wykonania analiz danych wielowymiarowych w czasie, gdy są oni odłączeni od sieci przedsiębiorstwa (np. przemieszczają się z komputerem przenośnym). W takich sytuacjach zwykle wystarcza dostęp do określonych fragmentów danych wielowymiarowych. Usługa PivotTable zapewnia możliwość zdefiniowania tylko pewnych fragmentów kostek OLAP, zapisu ich na komputerze użytkownika zdalnego i późniejszą analizę danych, gdy użytkownik ten znajdzie się poza zasięgiem sieci przedsiębiorstwa. Dane są aktualizowane po ponownym połączeniu z serwerem OLAP.

Okresową aktualizację danych można przeprowadzać przy pomocy usługi Data Transformation Services (DTS) wbudowanej w SQL Server. Dzięki mechanizmowi tzw. pakietów, mamy możliwość zdefiniowania operacji przetwarzania, przesyłania i transformacji danych z różnych źródeł oraz aktualizacji danych OLAP. Pakiet DTS jest opisem pracy, która ma być wykonana jako część procesu transformacji. Każdy pakiet definiuje jedno lub więcej zadań do wykonania w skoordynowanym ciągu, przy czym może on być realizowany w sposób cykliczny, dzięki wbudowanemu terminarzowi zadań. Pakiet DTS może być utworzony przy użyciu interfejsu graficznego lub języka programowania [3].

5. Podsumowanie

Systemy informacyjne dla przedsiębiorstw charakteryzuje w ostatnich latach gwałtowny wzrost rozmiarów własnych danych operacyjnych oraz zwiększone zapotrzebowanie na dane zewnętrzne. Coraz bardziej skuteczne mechanizmy gromadzenia i przetwarzania danych historycznych, szczególnie w przypadku dużych firm, umożliwiają prowadzenie różnorakich analiz w celach wspomagania decyzji. W niektórych zastosowaniach istotną rolę odgrywa czas, w jakim nowe dane mogą być dostępne dla celów analizy, stąd zrozumiałym wydaje się powstanie przedstawionych w tej pracy narzędzi analitycznych, określanych jako real-time OLAP. Systemy takie stanowią rozszerzenie dość powszechnie stosowanej obecnie technologii hurtowni danych i przetwarzania analitycznego on-line. Chociaż znajdują się w fazie rozwoju, to należy przypuszczać, że w niedalekiej przyszłości uda się pokonać ich obecne ograniczenia i będą stale zwiększać zarówno swoje możliwości jak też efektywność działania.

6. Literatura

[1] C. Ballard, D. Herreman, D. Schau, R. Bell, E. Kim, A. Valencic: Data Modeling Techniques for Data Warehousing. IBM Corporation, 1998. Dostępne via: http://www.redbooks.ibm.com .

[2] R. Elmasri, S. B. Navathe: Fundamentals of Database Systems, Addison-Wesley, Vancouver, Canada 2000.

[3] B. Jędrzejec, R. Bembenik: Hurtownie danych i technologia OLAP w Microsoft SQL Server 7.0. Praca magist. PRz, 2001.

[4] J. Kiviniemi: Opportunities of OLAP in Industrial Applications, Research Report TTE1-3-98, VTT Information Technology, Espoo, Finland, December 1998.

[5] C. Hall: The move to real-time data warehousing and business intelligence, Business Intelligence Advisor, January 2001.

[6] J. Langseth: Real-Time Data Warehousing: Challenges and Solutions, DSSResources.COM, 02/08/2004. Dostępne via: http://dssresources.com/papers/features/langseth/langseth02082004.html.

[7] R. Basu: Challenges of Real-Time Data Warehousing, DM Direct Special Report, November 11, 2003 Issue.

[8] Dennis Kennedy: The Reality of Real-time OLAP, Microsoft Corporation. Dostępne via: http://msdn.microsoft.com/library/default.asp?url=/library/en-us/dnsql2k/html/sql_real-timeolap.asp.

[9] Microsoft SQL Server 7.0 Resource Kit, Microsoft Press, 1999.

Tytuł: A Real-Time Mode in Multidimensional Data Analysis

Artykuł recenzowany