

Hurtownie danych

Tomasz Traczyk

Referat przedstawia jeden z „gorących tematów” współczesnej informatyki. Opisano zadania, które spowodowały powstanie i rozwój hurtowni danych. Przedstawiono definicję hurtowni i typową architekturę systemu informacyjnego z hurtownią danych. Pokazano różnice między operacyjnymi bazami danych a hurtownią danych. Wyróżniono rolę centralnej hurtowni danych i wydziałowych składnic danych. Zaprezentowano zagadnienie analizy wielowymiarowej i typowe operacje na danych wielowymiarowych. Przedstawiono przykłady danych wielowymiarowych i typowe dla analizy wielowymiarowej struktury danych i operacje. Porównano najczęściej spotykane rodzaje systemów OLAP (*On-line Analytical Processing*): systemy relacyjne i wielowymiarowe. Zasygnalizowano specyficzne dla hurtowni danych problemy projektowania i implementacji.

Wprowadzenie

Po co hurtownie danych

Wiele organizacji i przedsiębiorstw napotyka w swym rozwoju barierę polegającą na braku szybkiego dostępu do aktualnej informacji dotyczącej strategicznych dla organizacji obszarów działania. Paradoksalnie, wszystkie dane źródłowe niezbędne w procesie decyzyjnym są w organizacjach zbierane i przechowywane w formie elektronicznej. Jednak rozproszenie danych uniemożliwia ich efektywne wykorzystanie do zarządzania.

Dla sprawnego zarządzania organizacją potrzeba, by:

- dane zgromadzone w organizacji mogły być wykorzystywane w procesie decyzyjnym;
- istniała możliwość tworzenia analiz obejmujących całość organizacji.

Postulowana jest zatem możliwość efektywnego przetwarzania analitycznego danych dotyczących organizacji.

Zgromadzone w istniejących systemach informatycznych dane operacyjne nie nadają się do wymienionych wyżej zadań gdyż są:

- rozproszone — w organizacji istnieje zwykle wiele systemów informatycznych służących do różnych celów, dane są w nich rozproszone i niejednorodne, a systemy często nie są zintegrowane ani nawet połączone;
- heterogeniczne — systemy informatyczne w organizacji pochodzą najczęściej od wielu producentów, przechowują dane w różnych bazach danych i innych systemach zapisu, dane są w różnych formatach, a ich budowa opiera się na różnych modelach danych; dodatkowy problem stanowią zróżnicowane systemy identyfikatorów;
- w niewłaściwym układzie — układ danych jest dostosowany do potrzeb operacyjnych, na ogół więc dane są przechowywane w sposób umożliwiający ich maksymalnie efektywne dopisywanie i modyfikacje (np. brak zapisanych danych zagregowanych, jak sumy czy średnie); układ taki z reguły nie sprzyja sprawnej analizie danych;
- bez historii — w operacyjnych bazach danych często przechowuje się dane odzwierciedlające jedynie stan aktualny; jeśli nawet przechowywane są dane historyczne, to zwykle jedynie dla krótkiego okresu (np. od początku roku); tymczasem do analiz i porównań mogą być potrzebne dane z długiego okresu.

Oprogramowanie zarządzające danymi operacyjnymi także nie nadaje się dobrze do przetwarzania analitycznego, gdyż jest zoptymalizowane do innych celów. Są to na ogół systemy przetwarzania transakcji OLTP (*On-line Transaction Processing*), które konstruowano z myślą o maksymalnej wydajności operacji wprowadzania i modyfikacji danych. Najczęściej systemy te sprawnie obsługują

wielką liczbę stosunkowo niewielkich transakcji wykonywanych równocześnie przez wielu użytkowników. Tymczasem w przetwarzaniu analitycznym wielodostęp i efektywna modyfikacja danych są problemami drugorzędnymi. Istotna jest natomiast efektywność operacji wyszukiwania, odczytu i agregowania bardzo dużych wolumenów danych.

Cechy danych operacyjnych zgromadzonych w organizacji sprawiają, że bezpośrednia analiza tych danych do celów zarządzania jest bardzo trudna i nieefektywna albo w ogóle niemożliwa. Jednocześnie potrzeba wykorzystywania najnowszych danych do zarządzania staje się bardzo ważna, zwłaszcza dla organizacji, które muszą stawiać czoło konkurencji.

Przykład

W dużej fabryce produkującej podzespoły elektroniczne dane operacyjne obsługiwane są przez kilka różnych systemów:

- zarządzanie produkcją, sprzedaż i gospodarka magazynowa są obsługiwane przez pakiet klasy MRP II z własną bazą danych;
- systemy księgowo, kadrowe i płacowe są zbudowane w oparciu o relacyjną bazę danych średniej klasy;
- planowanie sprzedaży i produkcji odbywa się za pomocą arkuszy kalkulacyjnych.

Przedsiębiorstwo to produkuje kilkadziesiąt odmian wyrobu, różniących się parametrami technologicznymi, ale w różnych systemach informatycznych stosowane są różne systemy oznaczeń do identyfikacji wyrobu.

Żaden z tych systemów nie daje środków do przeprowadzania analiz potrzebnych osobom zarządzającym fabryką. By móc przeprowadzać efektywne analizy, zwłaszcza zaś porównania planowania i wykonania zadań, firma ta buduje hurtownię danych.

Jak rozwiązać problem

Ponieważ istniejące w organizacjach systemy informatyczne obsługujące dane operacyjne nie umożliwiają efektywnego pozyskiwania na bieżąco danych dla zarządzania, konieczne stało się znalezienie rozwiązania, które pozwoliłoby:

- scalić dane z różnych źródeł;
- efektywnie udostępniać aktualne dane do analiz;
- przechowywać dane historyczne.

Okazało się, że cechy te można uzyskać budując specjalną scentralizowaną bazę — hurtownię danych i gromadząc w niej dane z różnych systemów działających w organizacji.

Co to jest hurtownia danych

Hurtownia danych (magazyn danych, *data warehouse*) jest wydzieloną centralną bazą danych zbierającą informacje służące do zarządzania organizacją. Baza ta jest odizolowana od baz operacyjnych, a jej struktura i użyte do jej budowy narzędzia powinny być zoptymalizowane pod kątem przetwarzania analitycznego. W hurtowni są gromadzone dane pozyskiwane okresowo z systemów obsługujących dane operacyjne.

Cechy hurtowni danych

Hurtownia, mając za zadanie efektywną obsługę przetwarzania analitycznego, musi mieć odpowiednie do tego cechy:

- jest scentralizowaną bazą danych — dzięki temu wszystkie potrzebne informacje, bez względu na to w której z operacyjnych baz danych powstały, są zgromadzone i dostępne w jednym miejscu;
- jest oddzielona od baz operacyjnych — a przez to może mieć inną budowę, dostosowaną do swych specyficznych zadań;
- scala informację z wielu źródeł — zbieranie danych z wielu systemów informatycznych w organizacji jest związane z ujednoceniem sposobu ich zapisu, identyfikatorów itp.; ładowanie danych najczęściej odbywa się okresowo;

- jest zorientowana tematycznie — hurtownia nie zbiera oczywiście wszystkich danych z całej organizacji, a jedynie te dane, które będą przydatne do sporządzania analiz w przewidywanym dla hurtowni zakresie;
- przechowuje dane historyczne — by umożliwić analizy porównawcze; dane te mogą dotyczyć długich okresów czasu;
- utrzymuje wielką ilość informacji — hurtownia zbiera dane szczegółowe z wielu systemów i przechowuje ich wersje historyczne, co szybko prowadzi do powstania bardzo dużej bazy danych;
- agreguje informację — ponieważ objętość danych w hurtowni jest bardzo duża, wyliczanie informacji zagregowanej jest czasochłonne; aby umożliwić efektywne analizy w hurtowni przechowuje się wyliczone wyniki agregacji (tzw. zmateriałizowane agregaty).

Dane w hurtowniach

W hurtowniach danych przechowuje się dane różnych rodzajów:

- elementarne — kopie aktualnych danych źródłowych pozyskanych z baz operacyjnych i odpowiednio przetworzonych (np. ujednoliconych);
- zmateriałizowane agregaty — czyli wyliczone wartości obliczeń (sumy, średnie itp.), w różnych przekrojach (np. sumy wartości sprzedaży w jednostkach czasu i w podziale terytorialnym) i na różnych stopniach agregacji (np. sumy dzienne, miesięczne i roczne);
- historyczne — dane elementarne i/lub agregaty dotyczące przeszłości;
- metadane — informacje słownikowe, opisujące strukturę hurtowni danych i źródłowych baz danych, z których pozyskuje się dane do hurtowni, oraz sposób wyliczania danych zagregowanych.

Cykl życia danych w hurtowni wygląda zwykle następująco:

- ładowanie i scalanie — dane są okresowo (np. raz dziennie) ładowane z baz operacyjnych; w czasie ładowania dokonywane jest scalenie i ujednolicenie danych, tzn. konwersja typów i formatów, przetłumaczenie identyfikatorów, przekształcenie do innego modelu danych;
- agregacja — od razu w czasie ładowania albo zaraz po nim dokonuje się wyliczenia zmateriałizowanych agregatów;
- przeniesienie do danych historycznych — zanim załadowana zostanie nowa wersja danych elementarnych, dotychczasowe dane muszą być zapamiętane jako historyczne; nie są one jednak przenoszone do osobnego archiwum, ale są na ogół przechowywane w tej samej hurtowni danych, by możliwe było sprawne dokonywanie porównań i przekrojów czasowych;
- usuwanie — nie jest operacją typową dla hurtowni, jest przeprowadzane rzadko albo w ogóle nigdy; usuwane mogą być dane historyczne tak stare, że już nie są wykorzystywane; kasowanie danych może oczywiście mieć miejsce także przy przebudowie hurtowni, np. gdy zmieniono jej zadania.

Ilość informacji zgromadzonej w typowej hurtowni stale rośnie. Dane elementarne są w zasadzie jedynie dopisywane, nie są potem na ogół modyfikowane ani usuwane. Sposób pracy hurtowni jest więc rzeczywiście zupełnie inny od działania baz operacyjnych. Nawet w niewielkiej organizacji hurtownia danych szybko osiąga bardzo duże rozmiary.

Hurtownie i składnice danych

Do gromadzenia danych dotyczących całej organizacji konieczne jest stworzenie centralnej hurtowni danych. W hurtowni tej przechowuje się pozyskane z innych systemów dane elementarne niezbędne do tworzenia potrzebnych analiz. O ile zestaw danych elementarnych do bardzo wielu różnych analiz zwykle jest podobny (np. szczegółowe dane o produkcji i sprzedaży), to pożądany sposób agregacji danych silnie zależy od rodzaju prowadzonych analiz. Poszczególne wydziały organizacji mogą potrzebować różnych danych zagregowanych w odmienny sposób.

Przechowywanie wszystkich danych i agregatów zaspokajających potrzeby wszystkich odbiorców w jednej centralnej hurtowni jest często nieefektywne i niepożądane ze względów organizacyjnych. Dlatego tworzy się mniejsze, wyspecjalizowane składnice danych (*data marts*), zwykle tworzone dla wydziałów organizacji.

Między centralną hurtownią a wydziałowymi składnicami danych zachodzą liczne różnice, zarówno co do ich roli i miejsca w architekturze całego systemu informatycznego, jak i co do zasad budowy:

- **Hurtownia danych (*data warehouse*)**
 - jest niezależna od zastosowania — gromadzi dane elementarne pokrywające potrzeby wszystkich przewidywanych analiz;
 - jest scentralizowana — jedna hurtownia w organizacji gromadzi dane ze wszystkich baz operacyjnych;
 - jest przeznaczona do wykorzystania w całej organizacji;
 - zawiera dane historyczne;
 - przechowuje dane mało zagregowane — przechowywane są jedynie podstawowe, powszechnie wykorzystywane agregaty;
 - przechowuje dane mało zdenormalizowane, tzn. nie zawierające wielu powtórzeń;
 - ma wiele źródeł danych: dane operacyjne z wielu baz i inne dane zewnętrzne;
 - typową operacją jest dodawanie danych, natomiast modyfikacje i usuwanie zdarzają się rzadko.
- **Składnice danych (*data marts*)**
 - są specyficzne dla zastosowania — ich budowa jest inna w każdym z wydziałów, dostosowana do prowadzonych analiz;
 - są przeznaczone dla określonych użytkowników (np. wydziałów);
 - dane w różnych składnicach powtarzają się, choć mogą istnieć w różnych układach;
 - dane są silnie zagregowane — przechowywane są wyniki wielu wyliczeń, dostosowane do potrzeb prowadzonych analiz;
 - dane są często silnie zdenormalizowane, tzn. zawierają liczne powtórzenia — struktura danych jest bowiem optymalizowana pod kątem szybkości dokonywania analiz, zaś koszty zapisu danych (powiększane przez redundancję) są mało istotne;
 - mają niewiele źródeł danych, a najczęściej tylko jedno — centralną hurtownię danych;
 - może być wymagana podatność danych na modyfikacje, np. w przypadku prowadzenia analiz typu *what-if*.

Przykładową architekturę systemu z hurtownią i składnicami danych przedstawia Rysunek 1.

Wykorzystanie hurtowni danych

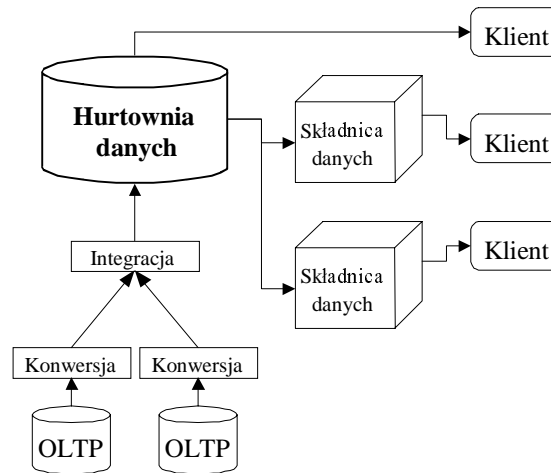
Dane zgromadzone w hurtowniach i składnicach danych są zwykle wykorzystywane przez menedżerów — użytkowników systemów wspomaganie decyzji. Systemy takie wykonują różnego rodzaju analizy, wśród których szczególną rolę odgrywają:

- przetwarzanie analityczne OLAP (*On-line Analytical Processing*), dokładniej opisane w dalszych częściach referatu;
- eksploracja danych (*data mining*) — czyli automatyczne pozyskiwanie wiedzy z baz danych; zadaniem jest tu odkrycie w danych wcześniej nieznanymi zależności; na ogół stosowane są do tego techniki sztucznej inteligencji¹.

Analiza wielowymiarowa

Dane gromadzone w hurtowniach danych najczęściej mają charakter wielowymiarowy. Jest to związane z potrzebą prowadzenia analiz wpływu wielu różnych czynników na zjawiska zachodzące w organizacji.

¹ Bardziej szczegółowe przedstawienie tego zagadnienia wykracza poza ramy niniejszego referatu.



Rysunek 1. Przykładowa architektura systemu z hurtownią i składnicami danych

Dane wielowymiarowe

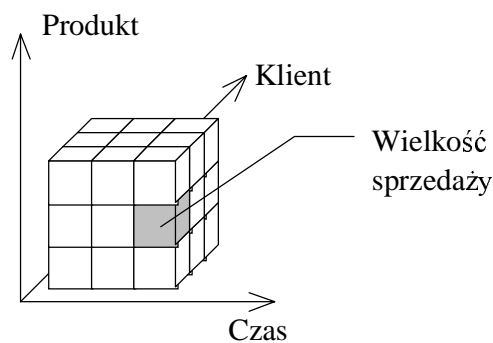
Struktura wielowymiarowa przedstawia elementarne komórki danych, tzw. fakty, w funkcji wielu niezależnych czynników, zwanych wymiarami.

Wymiary są opisane wartościami dyskretnymi, które mogą tworzyć hierarchie. Typowe wymiary to np.:

- czas (np. w dniach, miesiącach, kwartałach, latach);
- produkt (np. typ i rodzaj);
- jednostka organizacyjna (np. wydział, oddział) lub terytorialna (np. gminna, wojewódzka).

Fakty są opisane atrybutami liczbowymi, tzw. miarami. Najbardziej typowym faktem jest wielkość sprzedaży, której miarami są np. ilość sprzedanego towaru i jego wartość.

Dane wielowymiarowe można sobie wyobrazić jako kostkę umieszczoną w przestrzeni wymiarów, co pokazuje Rysunek 2.



Rysunek 2. Przykładowe dane wielowymiarowe

W czasie analizy dane wielowymiarowe są poddawane pewnym typowym dla nich operacjom, takim jak:

- obracanie, czyli zmiana perspektywy oglądania danych — w przypadku dwóch wymiarów jest to po prostu zamiana ich miejscami;
- selekcja, czyli wybór interesujących elementów wymiarów — pozostałe elementy są pomijane;
- projekcja, czyli zmniejszenie liczby wymiarów i zaprezentowanie danych w pozostałych wymiarach — prezentowane są dane zagregowane względem usuniętych wymiarów;

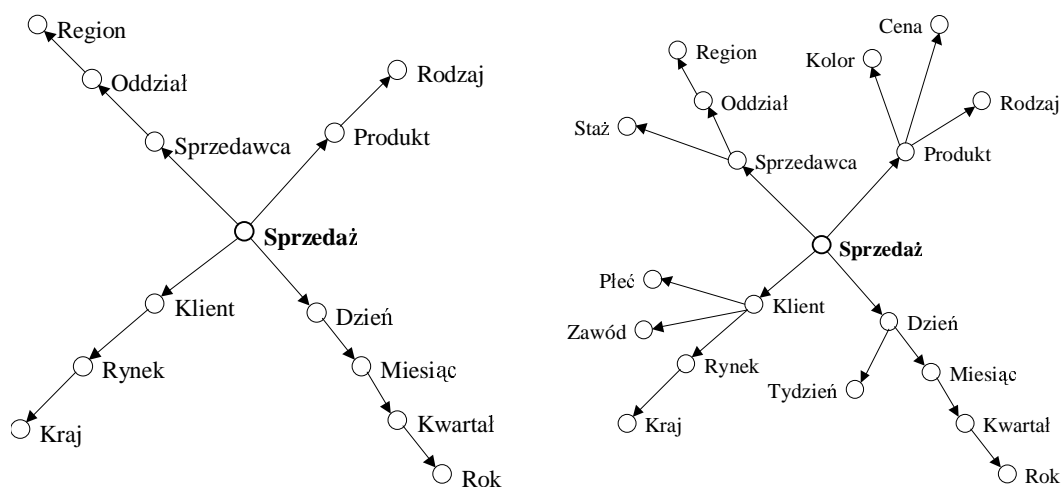
- wycinanie (*slice and dice*), czyli wykonanie projekcji, ale jedynie dla wybranych elementów pozostawianych wymiarów — jest to zatem połączenie selekcji i projekcji;
- ranking czyli uszeregowanie elementów wymiaru wg wzrostu miary lub jej agregatu;
- zwijanie (*roll-up*) i rozwijanie (*drill-down*) czyli nawigacja po hierarchii wymiaru; zwijanie łączy się z agregacją miar (np. prezentowaniem sum odpowiadających coraz ogólniejszym poziomom w hierarchii wymiaru), zaś rozwijanie — z dezagregacją.

Typowe struktury wielowymiarowe

Fakty stanowią centralny punkt struktury danych wielowymiarowych, są one powiązane związkami z wymiarami. W zależności od typu hierarchicznej budowy wymiarów struktura taka może przybrać jedną z dwóch form:

- struktura gwiazdzista (*starnet*) — gdy wymiary tworzą proste (liniowe) hierarchie;
- struktura „płatek śniegu” (*snowflake*) — gdy hierarchie wymiarów mają postać drzew.

Te typowe struktury danych wielowymiarowych przedstawia Rysunek 3.



Rysunek 3. Typowe struktury danych wielowymiarowych: gwiazdzista i „płatek śniegu”

Przetwarzanie analityczne

Przetwarzanie analityczne OLAP (*On-line Analytical Processing*) musi mieć następujące cechy:

- efektywne analizowanie wielkiej ilości danych w środowisku wielodostępnym;
- prezentacja danych niezależna od sposobu ich przechowywania;
- szybkie realizowanie zapytań i obliczeń, umożliwiające interaktywną analizę iteracyjną;
- wykonywanie różnorodnych obliczeń: agregacji, operacji wielowymiarowych, obliczeń statystycznych, macierzowych, prognozowania, analizy trendów itp.;
- łatwe tworzenie różnych form prezentacji wyników analizy: raportów, arkuszy kalkulacyjnych, wykresów itd.

Wymaganiom tym mogą sprostać jedynie odpowiednie serwery danych i specjalne narzędzia do budowy aplikacji analitycznych.

Serwery OLAP

Spotykane na rynku systemy OLAP można podzielić na dwie grupy, różniące się rodzajem zastosowanego do ich konstrukcji serwera danych.

- ROLAP (*Relational OLAP*) to systemy zbudowane w oparciu o relacyjną bazę danych, cechuje je:
 - zdolność do przechowywania wielkiej objętości danych (rzędu terabajtów);

- złożone struktury danych, wynikające z konieczności relacyjnego odwzorowania zależności wielowymiarowych;
- problemy z wydajnością, wiążące się z niedostosowaniem struktur relacyjnych do analizy wielowymiarowej;
- względnie łatwa modyfikacja danych, co wynika z zastosowanego oprogramowania i struktur danych.

Systemy takie, ze względu na możliwość przechowywania ogromnej ilości danych, są często stosowane do budowy centralnych hurtowni danych.

- MOLAP (*Multidimensional OLAP*) są to systemy zbudowane z wykorzystaniem specjalizowanych serwerów wielowymiarowych, cechują je:
 - mniejsze możliwości przechowywania danych (rzędu gigabajtów);
 - naturalna reprezentacja struktur wielowymiarowych;
 - duża wydajność analizy wielowymiarowej;
 - mechanizmy kompresji dla danych rzadkich;
 - znacznie trudniejsza modyfikacja danych — zmiana danych wymaga bowiem często przebudowy struktury wielowymiarowej;

Systemy takie są stosowane dla wydziałowych składnic danych albo jako struktury pomocnicze współpracujące z relacyjną hurtownią danych. Mogą też być używane samodzielnie gdy ilość gromadzonych danych nie jest bardzo duża.

Ponieważ oba rozwiązania mają przeciwstawne zalety i wady, dobrym rozwiązaniem jest ich połączenie, tak aby można było przechowywać wielkie objętości danych i prowadzić efektywnie analizę wielowymiarową.

Rozwiązanie takie może polegać na zastosowaniu relacyjnej bazy danych do przechowywania całego zbioru danych elementarnych i historycznych, zaś użyciu serwerów wielowymiarowych do przechowywania, agregowania i przetwarzania roboczych podzbiorów danych. Najlepiej, jeśli przepływ danych między serwerem relacyjnym i wielowymiarowym odbywa się automatycznie, w zależności od żądanych analiz; serwer wielowymiarowy stanowi wówczas rodzaj inteligentnego bufora danych.

Oprogramowanie do budowy aplikacji OLAP

W szybko zmieniającym się świecie bardzo trudno jest przewidzieć, jakiego typu obliczenia analityczne będą w przyszłości potrzebne we wspomaganie decyzji. Niezbędne jest zatem posłużenie się takimi narzędziami, które dają możliwość łatwego tworzenia *ad hoc* zapytań, obliczeń i prezentacji.

Narzędzia służące do budowania i przeprowadzania analiz muszą przy tym być narzędziami przyjaznymi, gdyż ich użytkownikami nie mają być programiści, lecz specjaliści z dziedziny zarządzania.

Dlatego do budowania aplikacji typu OLAP nie używa się na ogół typowych dla przetwarzania OLTP narzędzi do tworzenia aplikacji. Stosuje się łatwe w obsłudze i wydajne narzędzia specjalizowane. Typowe cechy tych narzędzi to:

- przyjazny interfejs, np. przypominający arkusz kalkulacyjny lub przeglądarkę danych (*browser*);
- możliwość składania aplikacji z gotowych bloków;
- duże możliwości graficznego zobrazowania wyników;
- wbudowane mechanizmy analizy wielowymiarowej.

Niektórzy producenci tego typu narzędzi dostarczają także gotowe aplikacje, rozwiązujące typowe problemy (np. analizę sprzedaży). Zawsze jednak użytkownik musi mieć możliwość samodzielnego rozbudowania takiej aplikacji o nowe analizy.

Budowanie hurtowni danych

Proces projektowania i budowania hurtowni danych nie jest niestety łatwy i z reguły wiąże się ze znacznymi kosztami. Dlatego jest to przedsięwzięcie które powinno być starannie zaplanowane i prowadzone. Gotowa hurtownia staje się dla organizacji aplikacją typu *mission critical*, gdyż od pra-

widliwości jej działania zależy trafność strategicznych decyzji w zarządzaniu. Niepowodzenia w tworzeniu hurtowni mogą być zatem szczególnie kosztowne.

Problemy budowy hurtowni

Podstawowe problemy napotymane przy budowie hurtowni mają charakter:

- koncepcyjny — problemy z budową modelu obejmującego wszystkie potrzeby i dającego się zrealizować w oparciu o dostępne dane źródłowe; trudności z danymi historycznymi (zwłaszcza gdy struktura danych źródłowych zmienia się z czasem) itp.;
- organizacyjny — problemy z pozyskaniem wiedzy o celach hurtowni, o danych źródłowych, trudności z ustaleniem odpowiedzialności za prawidłowość danych zasilających hurtownię itp.;
- psychologiczny — trudności projektantów z odejściem od myślenia w kategoriach OLTP, opory menedżerów przed zaufaniem analizom z hurtowni albo nadmierne do nich zaufanie itd.;
- technologiczny;
- finansowy.

Pewne problemy technologiczne właściwe są dla systemów relacyjnych (ROLAP).

- Problemy z zapytaniami gwiazdzistymi — analiza w strukturze gwiazdzistej lub *snowflake* wymaga zapytań zawierających złączenia bardzo wielu tabel, co przy wielkiej objętości danych może prowadzić do nieakceptowalnych czasów reakcji.
- Problemy z wyszukiwaniem — np. stosowane zwykle w systemach relacyjnych indeksy typu *B-tree* nie zapewniają dostatecznej wydajności przy operacjach wielowymiarowych.

Wystarczającą wydajność systemów ROLAP próbuje się osiągnąć wyposażając serwery relacyjne w środki takie jak równoległe przetwarzanie zapytań, indeksy bitowe oraz złączeniowe, grona, specjalne sposoby optymalizacji zapytań gwiazdzistych.

Problemy właściwe dla serwerów wielowymiarowych (MOLAP) są przede wszystkim związane z ograniczeniami objętości danych składowanych w tych serwerach.

Inne problemy technologiczne są niezależne od użytego modelu danych.

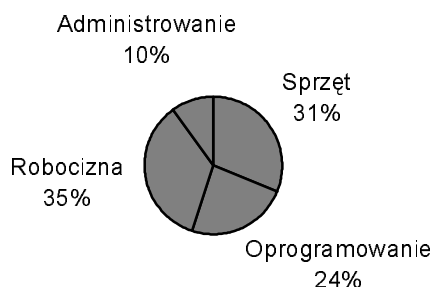
- Konieczność efektywnego partycjonowania danych — wobec swej wielkiej objętości dane muszą być przechowywane na wielu nośnikach fizycznych. Podział powinien zapewniać jak najlepszą wydajność systemu.
- Problemy z ładowaniem danych:
 - heterogeniczność źródeł — dane muszą być odczytane z wielu różnych systemów i doprowadzone do wspólnej postaci; budowa odpowiednich pomostów może stanowić prawdziwe wyzwanie dla programistów;
 - wydajność ładowania danych — efektywne ładowanie powinno być przyrostowe, tj. dotyczyć tylko danych nowych lub zmienionych; jest to jednak trudno zrealizować bez ingerencji w systemy zarządzające danymi źródłowymi, co często jest niemożliwe.
- Spójność transakcji analitycznych — w odróżnieniu od transakcji w systemach OLTP, operacje analityczne zwykle trwają stosunkowo długo (minuty, godziny). Jeśli dane w hurtowni w tym czasie zmieniają się, to analiza może być błędna — oparta na niespójnym obrazie danych. System OLAP powinien zapewniać spójność takich długotrwałych transakcji analitycznych, co może być dość kosztowne (wymaga przechowywania wersji danych z chwili rozpoczęcia obliczeń).

Powodem problemów finansowych związanych z budową hurtowni jest przede wszystkim wielka ilość gromadzonych danych. Pociąga to za sobą konieczność:

- użycia wysokowydajnego sprzętu;
- zakupu dysków o wielkich pojemnościach;
- zastosowania oprogramowania najwyższej klasy;
- zatrudnienia wysoko wykwalifikowanych administratorów.

Same koszty projektu i realizacji hurtowni też mogą być znaczące, zwłaszcza ważyć tu mogą koszty programowania pomostów zasilających hurtownię i koszty przeprowadzenia testów.

Udział poszczególnych składników w kosztach typowej dużej hurtowni danych przedstawia Rysunek 4.



Rysunek 4. Typowy podział kosztów tworzenia hurtowni danych (wg Gartner Group)

Proces tworzenia hurtowni danych

Projektowanie hurtowni danych winno być poprzedzone starannym planowaniem i rozpoznaniem potrzeb organizacji. Istotne jest też precyzyjne zdefiniowanie architektury systemu. Poważny projekt powinien być poprzedzony wdrożeniem projektu pilotowego. Zalecane jest budowanie hurtowni w tzw. cyklu spiralnym, w którym całość zadania dzieli się na części i realizuje je kolejno, przy czym w czasie realizacji kolejnej części udoskonala się te, które były stworzone wcześniej. Takie podejście pozwala na stosunkowo szybkie uruchomienie pierwszych fragmentów systemu i wczesne zebranie doświadczeń, które pozwolą lepiej zaprojektować kolejne części.

W projektowaniu hurtowni danych szczególnie ważna jest staranna dokumentacja, gdyż hurtownia będzie używana w dłuższym okresie czasu, a aplikacje analityczne będą tworzone przez wielu użytkowników. Zaleca się zbudowanie specjalnych słowników danych (metadanych), które obejmują zarówno dane zgromadzone w hurtowni jak dane źródłowe. Szczególnie godne polecenia jest używanie narzędzi CASE do analizy i projektowania hurtowni — repozytorium CASE może wtedy pełnić rolę metadanych.

Zakończenie

Konieczność stałego analizowania aktualnych danych przez zarządzających organizacjami stała się przyczyną powstania i rozwoju koncepcji hurtowni danych oraz przetwarzania analitycznego.

Choć budowa hurtowni danych nie jest przedsięwzięciem prostym i wiąże się z ryzykiem oraz znacznymi kosztami, to jednak potrzeba dostępu do danych jest na tyle paląca, że aż 90% dużych przedsiębiorstw (USA, 1996 r.) decyduje się na tworzenie hurtowni danych. Wartość rynku hurtowni danych oceniano w roku 1996 na 1,5 mld dolarów, przewiduje się zaś jego wzrost do 6,9 mld dolarów w roku 1999 (wg Gartner Group).

Informacja o autorze:

dr inż. Tomasz Traczyk jest adiunktem w Instytucie Automatyki i Informatyki Stosowanej Politechniki Warszawskiej

e-mail: ttraczyk@ia.pw.edu.pl

URL: <http://www.ia.pw.edu.pl/~ttraczyk/>