

Hurtownie danych — wprowadzenie

Tomasz Traczyk

ttraczyk@ia.pw.edu.pl



Instytut Automatyki i Informatyki Stosowanej
Politechniki Warszawskiej



Hurtownie danych — wprowadzenie

Wstęp

- Co to jest hurtownia danych?
- Analiza wielowymiarowa
- Przetwarzanie analityczne OLAP
- Podsumowanie

Hurtownie danych — wprowadzenie

2

Motywacja

Wymagania zarządzania organizacją

- Nadążanie za potrzebami klientów
- Dotrzymywanie kroku konkurencji
- Szybkie i celne decyzje

Do podejmowania decyzji potrzeba

- Aktualnych danych o organizacji
- Analiz obejmujących całość organizacji
- Systemów przetwarzania analitycznego

Hurtownie danych — wprowadzenie

3

Dlaczego istniejące systemy nie są wystarczające

Operacyjne bazy danych

- Rozproszone
 - wiele systemów informatycznych
 - dane niejednolite
 - systemy nie zintegrowane
- Heterogeniczne
 - pochodzenie o dwidu producentów
 - różne bazy danych i systemy zapisu
 - różne formaty i modele danych
 - zróżnicowane systemy identyfikatorów
- W niewłaściwym układzie
 - układ danych dostosowany do efektywnego wprowadzania i modyfikacji
 - układy z reguły nie sprzyja sprawnej analizie danych
- Brak historii
 - dane odzwierciedlają stan aktualny
 - dane historyczne są niewykorzystywane dla krótkiego okresu

Potrzeby przetwarzania analitycznego

- System zintegrowany
- System jednorodny
- Układ danych dostosowany do efektywnej analizy
- Dostępność danych historycznych z długiego okresu do analiz i porównań

Hurtownie danych — wprowadzenie

4

OLTP a przetwarzanie analityczne

Systemy OLTP

(On-line Transaction Processing)

- Duża wydajność wprowadzania i modyfikacji danych
- Sprawna obsługa wielkiej liczby niewielkich transakcji
- Efektywna obsługa wielu użytkowników

Potrzeby przetwarzania analitycznego

- Duża efektywność wyzyskiwania danych
- Odczyt dużych wolumenów danych (duże i długotrwałe „transakcje” analityczne)
- Sprawne agregowanie danych

Hurtownie danych — wprowadzenie

5

Potrzeby i rozwiązanie

Potrzeby

- Scalenie danych z różnych źródeł
- Efektywne udostępnianie aktualnych danych do analizy
- Przechowywanie danych historycznych i zagregowanych

Rozwiązanie?

HURTOWNIA DANYCH

Hurtownie danych — wprowadzenie

6

- Wstęp
- Co to jest hurtownia danych?**
- Analiza wielowymiarowa
- Przetwarzanie analityczne OLAP
- Podsumowanie

7

Hurtownia danych

Co to jest hurtownia danych?

Hurtownia danych (magazyn danych, data warehouse) jest wydzieloną centralną bazą danych zbierającą informacje służące do zarządzania organizacją

Cechy hurtowni danych

- Scentralizowana baza danych
 - informacje dostępne w jednym miejscu
- Oddzielona od baz operacyjnych
 - może mieć budowę dostosowaną do swych specyficznych zadań
- Scala informacji z wielu źródeł
 - ujednoczony model i format
- Jest zorientowana tematycznie
 - zbiera tylko informacje przydatne do analiz w zakresie przewidzianym dla hurtowni
- Przechowuje dane historyczne
 - dane mogą dotyczyć długiego okresu
- Utrzymuje wielką ilość informacji
 - zbiera dane z wielu źródeł
 - przechowuje i historię
- Agreguje informacje
 - dla sprawności analiz przechowuje wyniki obliczeń

8

Dane w hurtowni

- Dane elementarne
 - kopie aktualnych danych źródłowych z baz operacyjnych
 - odpowiednio przetworzone (np. ujednoczone)
- Zmaterializowane agregaty
 - wyliczone wartości obliczeń (sumy, średnie itp.)
 - w różnych przekrojach (np. w jednostkach czasu i w podziale terytorialnym)
 - na różnych stopniach agregacji (np. sumy dzienne, miesięczne i roczne)
- Dane historyczne
 - dane elementarne i/lub agregaty dotyczące przeszłości
 - przechowywane przez długi okres
- Metadane — informacje słownikowe opisujące:
 - strukturę hurtowni danych
 - strukturę źródłowych baz danych i sposób pozyskiwania tych danych
 - sposób wyliczania agregatów

9

Cykl życia danych w hurtowni

- Ładowanie i scalanie
 - dane okresowo ładowane z baz operacyjnych
 - w czasie ładowania scalenie i ujednoczenie danych
- Agregacja
 - wyliczanie zmaterializowanych agregatów
- Przeniesienie do danych historycznych
 - stare dane elementarne są przechowywane danych jako:
 - historyczne dane elementarne
 - agregaty
 - agregaty typu *rolling summary*
- Usuwanie
 - nie jest operacją typową dla hurtowni

10

Składnice danych

Powody tworzenia

- Sposób organizacji (np. agregacji) danych optymalny dla różnych użytkowników może się różnić
- Przechowywanie wszystkich ujęć danych w centralnej hurtowni jest
 - nieefektywne
 - niepożądane ze względów organizacyjnych

Rozwiązanie

- Tworzenie mniejszych składnic danych (*data marts*, minihurtowni)
 - wyspecjalizowanych do obsługi określonej grupy użytkowników (np. wydziału)
 - czerpiących dane z centralnej składnicy danych

11

Hurtownia a składnice danych

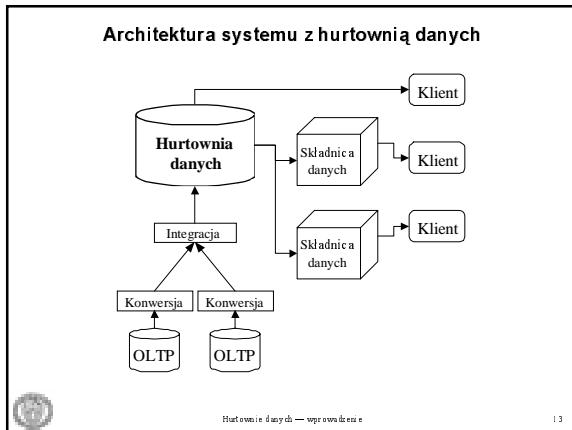
Hurtownia danych (data warehouse)

- Przeznaczona dla całej organizacji
- Niezależna od zastosowania
 - obejmując wszystkie przewidywane analizy
- Scentralizowana
 - jedna hurtownia w organizacji
 - gromadzi dane ze wszystkich baz operacyjnych
- Zawiera dane historyczne
- Przechowuje dane mało z agregowane
 - jedynie podstawowe, powszechnie wykorzystywane agregaty
- Przechowuje dane mało z denormalizowane
 - dane nie zawierają wielu powtórzeń
- Ma wiele źródeł danych
 - dane operacyjne z wielu baz
 - inne dane z zewnętrzne
- Typową operacją jest dodawanie danych
 - modyfikacje i usuwanie zdarza się rzadko

Składnice danych (data marts)

- Przeznaczone dla określonych użytkowników (np. wydziałów)
- Specyficzne dla zastosowania
 - budowa inna w każdym z wydziałów
 - dostosowana do prowadzonych analiz
- Dane w różnych składnicach powtarzają się
- Dane silnie zagregowane
 - przechowywane są wyniki wielu wyliczeń
 - szczególnie dostosowane do potrzeb prowadzonych analiz
- Dane silnie zdenormalizowane
 - zawierają liczne powtórzenia
 - struktura danych jest optymalizowana pod kątem szybkości dokonywania analiz
 - koszty zapisu danych są mało istotne
- Ma kilka źródeł danych
 - najczęściej tylko jedno źródło — centralna hurtownia danych
- Może być wymagana podatność danych na modyfikacje
 - analizy typu *what-if*

12



- ### Wykorzystanie hurtowni danych
- Przetwarzanie analityczne OLAP (*On-Line Analytical Processing*)
 - tworzenie analiz przydatnych w zarządzaniu organizacją
 - najczęściej bazuje na analizie wielowymiarowej
 - Eksploracja danych (drażenie danych, *data mining*)
 - automatyczne pozyskiwanie wiedzy z baz danych
 - zadaniem jest odkrywanie w danych wcześniej nieznanych zależności
 - stosowane są techniki sztucznej inteligencji
- Hurtownia danych — wprowadzenie 14

- Wstęp
 - Co to jest hurtownia danych?
 - Analiza wielowymiarowa
 - Przetwarzanie analityczne OLAP
 - Podsumowanie
- Hurtownia danych — wprowadzenie 15

Dane wielowymiarowe

Fakty i miary

- Fakty — elementarne komórki danych
- Miary — liczbowe atrybuty opisujące fakty

Przykład:
sprzedaż opisana ilością i wartością sprzedanego towaru

Wymiary

- Fakty są przedstawione w funkcji wielu wymiarów
- Wymiary są opisane wartościami dyskretnymi
- Wartości opisujące wymiary mogą tworzyć hierarchie

Przykłady:

- czas (dni, miesiące, kwartały, lata)
- produkt (typ, rodzaj)
- Klient (jednostkowy, miłoścowość, region)

Hurtownia danych — wprowadzenie 16

Prezentacja tekstowa danych wielowymiarowych

Grupy lamiające

- Wymiary oznaczone przez widzę

Produkt	Klient	Czas	Sprzedaż
Zielony	Abacki	Marzec	120
	Abacki	Kwiecień	14
	Babacki	Marzec	103
	Babacki	Kwiecień	30
Niebieski	Abacki	Marzec	20
	Abacki	Kwiecień	85
	Babacki	Marzec	35
	Babacki	Kwiecień	79

Tabela przestawna

- Dwa wymiary widoczne w tabeli
- Pozostałe wymiary opisują stronę

Produkt: Zielony		Czas	
Klient	Marzec	Kwiecień	
Abacki	120	14	
Babacki	103	30	

Produkt: Niebieski		Czas	
Klient	Marzec	Kwiecień	
Abacki	20	85	
Babacki	35	79	

Hurtownia danych — wprowadzenie 17

Typowe operacje wielowymiarowe

Obrót

- zmiana perspektywy oglądania danych
- zamiana wymiarów na prezentacji

Produkt:	Zielony	Czas	
Klient	Marzec	Kwiecień	
Abacki	120	14	
Babacki	103	30	

Czas:	Marzec	Produkt	
Klient	Zielony	Niebieski	
Abacki	120	20	
Babacki	103	35	

Hurtownia danych — wprowadzenie 18

Typowe operacje wielowymiarowe (2)

Selekcja

- Wybór interesujących elementów wymiarów
- Pozostałe elementy są pomijane

Produkt: Zielony				
Klient	Czas	Luty	Marzec	Kwiecień
Abacki	63	120	14	
Babacki	48	103	30	
Cabacki	56	98	44	

Produkt: Zielony			
Klient	Czas	Marzec	Kwiecień
Abacki	120	14	
Babacki	103	30	

Hurtownie danych — wprowadzenie 19

Typowe operacje wielowymiarowe (3)

Projekcja

- Zmniejszenie liczby wymiarów
- Zaprezentowanie danych z agregacją względem usuniętych wymiarów

Produkt: Zielony				
Klient	Czas	Luty	Marzec	Kwiecień
Abacki	2	3	2	
Babacki	6	7	2	
Cabacki	1	4	5	

Produkt: Zielony			
Klient	Czas	Niebieski	Czerwony
Abacki	7	10	15
Babacki	15	28	17
Cabacki	10	31	21

Hurtownie danych — wprowadzenie 20

Typowe operacje wielowymiarowe (4)

Wycinanie (slice and dice)

- Redukcja liczby wymiarów:
 - wybór określonych elementów usuwanych wymiarów
 - projekcja danych dotyczących ww elementów na pozostawiane wymiary
- Na części dane prowadzi do widoku dwuwymiarowego płaszczyzny

Produkt: Zielony				
Klient	Czas	Luty	Marzec	Kwiecień
Abacki	2	3	2	
Babacki	6	7	2	
Cabacki	1	4	5	

Produkt: Zielony			
Klient	Czas	Niebieski	Czerwony
Abacki	7	8	12
Babacki	9	24	15
Cabacki	9	27	17

Hurtownie danych — wprowadzenie 21

Typowe operacje wielowymiarowe (5)

Ranking

- U porządkowanie elementów w wymiarze
 - według miary
 - według agregatu miary

Hurtownie danych — wprowadzenie 22

Typowe operacje wielowymiarowe (6)

Zwijanie (roll-up) i rozwijanie (drill-down)

- Zwijanie
 - naviagacja w górę hierarchii wymiaru (uogólnienie)
 - agregacja miar
- Rozwijanie
 - naviagacja w dół hierarchii wymiaru (szczegółowienie)
 - dezagregacja miar

Klient: Abacki			
Produkt	Czas	Marzec	Kwiecień
Niebieskie	N2	4	1
N1	1	9	
Zielone	Z2	6	3
Z1	2	2	5

Klient: Abacki			
Produkt	Czas	Marzec	Kwiecień
Niebieskie	5	10	
Zielone	8	8	

Hurtownie danych — wprowadzenie 23

Typowe struktury danych wielowymiarowych

Gwiazdista (starnet)

Płatek śniegu (snowflake)

Hurtownie danych — wprowadzenie 24

Dane wielowymiarowe w bazach relacyjnych

Reprezentacja danych

- Wymiary rozproszone między wiele tabel
 - osobna tabela dla każdego poziomu hierarchii wymiaru
- W tabeli faktów bardzo dużo wierszy
 - osobny wiersz dla każdej kombinacji wymiarów z najwyższych poziomów hierarchii
- Dodatkowe tabele dla agregatów

Zapytania do struktur gwiazdzących

- Kłopotliwa budowa zapytań
 - wiele tabel
 - skomplikowane wielopoziomowe złączenia
- Problemy z wydajnością
 - bardzo duża objętość danych
 - złożone zapytania

Przykład:

```
SELECT e.region, r.rodzaj_produkту, SUM(wartosc_sprzedazy)
FROM sprzedaz f, sprzedawca s, oddzial o, region e, produkt p, rodzaj r
WHERE f.id_sprzedawcy=s.id_sprzedawcy AND f.id_produkту=p.id_produkту
AND s.id_oddzialu=o.id_oddzialu AND o.id_regionu=e.id_regionu
AND p.id_rodzaju=r.id_rodzaju
GROUP BY e.region, r.rodzaj_produkту
```



Wstęp

Co to jest hurtownia danych?

Analiza wielowymiarowa

Przetwarzanie analityczne OLAP

Podsumowanie



Co to jest OLAP?

Cechy OLAP (On-Line Analytical Processing)

- Efektywne analizowanie wielkiej ilości danych
- Szybkie realizowanie zapytań i obliczeń, umożliwiające interaktywną analizę iteracyjną
- Wykonywanie różnorodnych obliczeń
 - agregacji
 - operacji wielowymiarowych
 - obliczeń statystycznych, macierzyowych
 - prognozowania, analizy trendów itp.
- Łatwe tworzenie różnych form prezentacji wyników analizy
 - raportów
 - arkuszy kalkulacyjnych
 - wykresów itd.
 - stron WWW
- Prezentacja danych niezależna od sposobu ich przechowywania



ROLAP i MOLAP

ROLAP (Relational OLAP)

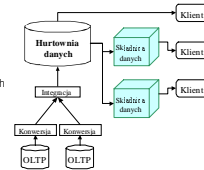
- Zdolność do przechowywania wielkiej objętości danych (rzędu terabajtów)
- Złożone struktury danych
 - konieczność relacyjnego odzwierciedlenia zależności wielowymiarowych
- Problemy z wydajnością
 - niedostosowanie struktur relacyjnych do analizy wielowymiarowej
- Wzrost kosztów i czasu modyfikacji danych

MOLAP (Multidimensional OLAP)

- 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 wymagać może przebudowy struktury wielowymiarowej

Rozwiązania mieszane

- Centralna hurtownia danych w bazie relacyjnej
- Serwery wielowymiarowe w wyspecjalizowanych składowiskach danych
- Automatyczny przepływ danych



Oprogramowanie do budowy aplikacji OLAP

Cele narzędzi OLAP

- Efektywna interaktywna analiza danych
- Możliwość tworzenia *ad hoc*
 - zapytań
 - raportów
 - prezentacji
 - stron WWW
- Łatwe tworzenie i modyfikacja aplikacji
- Możliwość wykorzystania przez osoby bez przygotowania informatycznego
- Przedstawianie danych w języku
 - zrozumiałym dla biznesmenów
 - niezależnym od wewnętrznej reprezentacji danych

Cechy narzędzi OLAP

- Przyjazny interfejs, np. przypominający arkusz kalkulacyjny lub przeglądarkę danych i browserów
- Możliwość składania aplikacji z gotowych bloków
- Duże możliwości graficznego z obrazowania wyników
- Wbudowane mechanizmy analizy wielowymiarowej
- Możliwość zaskupienia gotowych aplikacji
 - rozwiązujących typowe problemy
 - zdatnych do dalszej rozbudowy



Wstęp

Co to jest hurtownia danych?

Analiza wielowymiarowa

Przetwarzanie analityczne OLAP

Podsumowanie



Trudności

Technologiczne

- Problemy technologii ROLAP
 - ograniczona wydajność analizy
- Problemy technologii MOLAP
 - ograniczona objętość danych
- Problemy obu technologii
 - duże wymagania co do sprzętu i oprogramowania
 - problemy z pozyskiwaniem danych

Inne

- Organizacyjne
 - z pozyskaniem wiedzy o celach hurtowni
 - z pozyskaniem wiedzy o danych źródłowych
 - z ustaleniem odpowiedzialności za prawidłowość danych zasilających
- Konceptyjne
 - z budową modelu
 - obejmujące go wszystkie potrzeby
 - dać go się zrealizować w oparciu o dostępne dane źródłowe
 - z danymi historycznymi
 - z innymi modelami struktury danych
- Psychologiczne
 - projektantów — z odjęciem od myślenia w kategoriach OLTP
 - menedżerów — o przywróceniu zaufania do analizy z hurtowni albo nadmierne do nich zaufanie
- Finansowe
 - budowa hurtowni jest przedsięwzięciem drożym (miliony \$)

Nie można kupić gotowych hurtowni danych

Hurtownie danych — wprowadzenie

31

Korzyści

- Wykorzystanie danych gromadzonych w organizacji do zarządzania
- Integracja i centralizacja danych
- Szybkie i precyzyjne analizy
- Oddzielenie zadań analitycznych od operacyjnych

Ryzyko

- Inwestycyjne
 - duże koszty
 - niepewny zwrot (ale z nadzieją na bardzo dużą stopę zwrotu)
- Biznesowe
 - błędne analizy mogą prowadzić do fatalnych decyzji

Hurtownie danych — wprowadzenie

32

Hurtownie danych w Polsce

Dostawcy

- Obecni niemal w wszyscy wiodący dostawcy technologii
- Liczne firmy (w tym firmy o znacznej renomie) oferują budowanie hurtowni

Użytkownicy

- Banki i ubezpieczenia
- Duże firmy i przedstawicielstwa handlowe
- Telekomunikacja
- Energetyka
- Przemysł

Hurtownie danych — wprowadzenie

33

Podsumowanie

Za

- Możliwość podejmowania trafnych decyzji w oparciu o analizę aktualnych danych

Przeciw

- Spore ryzyko w przypadku niepowodzenia projektu lub błędów systemu

Wynik

- Większość dużych organizacji decyduje się na tworzenie hurtowni danych
- Wartość rynku hurtowni danych na świecie ocenia się na:
 - 1,5 mld \$ w 1996
 - 6,9 mld \$ w 1999
- Rynek ten szybko rozwija się także w Polsce

Hurtownie danych — wprowadzenie

34

Hurtownie danych — wprowadzenie

- Wstęp
- Co to jest hurtownia danych?
- Analiza wielowymiarowa
- Przetwarzanie analityczne OLAP
- Problemy budowy hurtowni danych
- Podsumowanie

Hurtownie danych — wprowadzenie

35