Hurtownie danych – czyli jak zapewnić dostęp do wiedzy tkwiącej w danych

Andrzej Ptasznik
Warszawska Wyższa Szkoła Informatyki
aptaszni@wwsi.edu.pl

Streszczenie

Przedmiotem wykładu są podstawy teorii hurtowni danych i aspekty ich wykorzystania. W pierwszej części zostaną omówione podstawowe cechy systemów OLTP (ang. *On-Line Transaction Processing*) oraz systemów OLAP (*ang. On-Line Analitycal Processing*). Omówione zostaną podstawowe pojęcia i przykłady projektów hurtowni danych. Przedstawione zostaną podstawowe zagadnienia związane z integracją danych oraz pojęcie analitycznej kostki wielowymiarowej. Zaprezentowane zostaną elementy technologii usług analitycznych i ich znaczenie w systemach typu *Business Intelligence*. W części końcowej wykładu omówione zostaną krótko podstawowe pojęcia związane z eksploracją danych (ang. *Data Mining*).

Spis treści

1 WPROWADZENIE

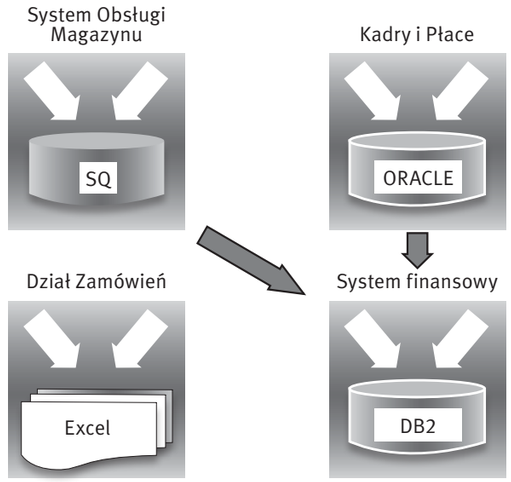
Burzliwy rozwój technologii informatycznych, a w szczególności baz danych, spowodował, że w każdej firmie czy instytucji gromadzone są różne dane na różnych etapach działalności. Bardzo często dane są gromadzone w różnorodny sposób – od plików tekstowych poprzez arkusze kalkulacyjne do baz danych. W okresie początkowego rozwoju systemy informatyczne wspomagające działalność firm koncentrowały się na wsparciu działalności operacyjnej. Powstawały rozmaite systemy ukierunkowane na konkretny aspekt działania, przykładowo:

- wystawianie faktur;
- obsługa magazynu;
- systemy kadrowe;
- systemy księgowe;
- obsługa klientów.

Zwykle systemy takie nie były z sobą w żaden sposób powiązane i tworzyli je różni producenci w odmiennych technologiach. Stosowanie technologii informatycznych w codziennej działalności firm i instytucji było związane z gromadzeniem danych na potrzeby konkretnego typu działania. Dane zbierane w różnych systemach, oprócz wspomagania codziennych działań, były wykorzystywane także do celów raportowania i informowania kierownictwa. Istniały jednak podstawowe problemy takiej działalności:

- dane po pewnym czasie stawały się niepotrzebne, ponieważ obsługa działalności codziennej nie musiała korzystać z danych historycznych (w systemie obsługi magazynu istotny był aktualny stan towaru w magazynie, a nie jaki był ten stan w zeszłym roku) – często w tego typu systemach usuwano starsze dane;
- wielokrotnie przetrzymywano te same dane w rożnych formatach;
- przetwarzanie danych na potrzeby inne niż wsparcie działalności codziennej znacząco wpływało na wydajność tych systemów.

Na rysunku 1 przedstawiony został schemat organizacji instytucji z wykorzystaniem różnych systemów informatycznych.



Rysunek 1.
Przykładowa organizacja firmy z wykorzystanie różnych systemów informatycznych

Duże ilości gromadzonych danych stają się kopalnią wiedzy, która może zostać wykorzystana do właściwego kierowania firmą i osiągnięcia przewagi konkurencyjnej na rynku.

2 SYSTEMY OLTP I OLAP

Tradycyjne systemy baz danych ukierunkowane są na realizację wielu małych i prostych zapytań i mają zapewnić wsparcie dla realizacji codziennych działań pracowników danej firmy lub instytucji. Dla tego typu systemów Edgar Frank „Ted" Codd (brytyjski informatyk, znany przede wszystkim ze swojego wkładu do rozwoju teorii relacyjnych baz danych) wprowadził pojęcie systemów **OLTP** (ang. *On-Line Transaction Processing*) i zdefiniował zbiór zasad, które powinny spełniać systemy tego typu. Podstawowe cechy systemów OLTP:

- przechowywane dane są zorientowane procesowo, np. wystawione faktury, otrzymane zamówienia, złożone reklamacje, wykonane przelewy itp.;
- stosunkowo niewielkie rozmiary baz danych (kilka gigabajtów);
- przechowywane są dane bieżące bez konieczności gromadzenia danych historycznych;
- realizowana jest duża liczba w miarę prostych zapytań;
- przechowywane są dane elementarne;
- realizowane są operacje wstawiania, modyfikowania i usuwania danych.

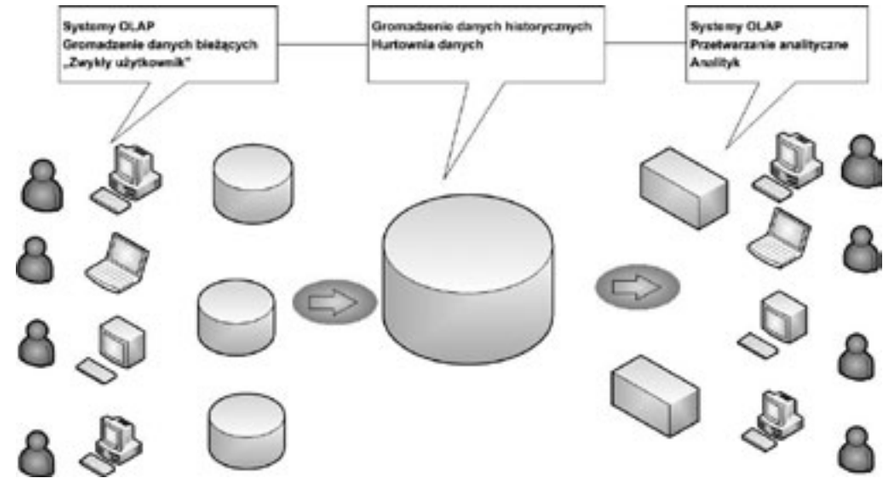
Zbiory danych tworzone w systemach OLTP stają się przydatne do pozyskiwania dodatkowych informacji potrzebnych kierownictwu firmy do podejmowania decyzji. Pojawiają się tu jednak pewne problemy:

- w ramach jednej firmy może istnieć wiele systemów typu OLTP;
- realizowanie dodatkowych czynności w ramach systemu OLTP wpływa na jego wydajność, tym bardziej że pozyskiwanie danych analitycznych wymaga wykonywania złożonych zapytań operujących na dużej liczbie danych;
- klasyczne zapytania SQL dostarczają danych w postaci dwuwymiarowych tabel, co często jest niewystarczające dla tego typu zastosowań.

Rozwiązaniem tych problemów stała się koncepcja wydzielonych systemów informatycznych świadczących usługi analityczne. Wspomniany wyżej Edgar Codd nazwał systemy tego typu **OLAP** (ang. *On-Line Analitycal Processing*) i również dla tych systemów sformułował zbiór zasad, które powinny spełniać. Podstawowe cechy systemów OLAP:

- przechowywane dane są zorientowane tematycznie, np. sprzedaż produktów, stany zapasów, wydatki, akcje promocyjne itp.;
- ogromne ilości gromadzonych danych (rzędu wielu terabajtów);
- przechowywane są dane bieżące i historyczne;
- realizowane są bardzo złożone zapytania operujące na wielkiej grupie danych;
- przechowywane są dane elementarne i zagregowane (sumy, średnie itp.);
- wykonywane są głównie operacje dopisywania nowych danych – praktycznie nie wykonuje się operacji modyfikowania danych.

Elementem łączącym systemy OLTP i OLAP są wyspecjalizowane bazy danych, gromadzące w specjalnie zaprojektowanych strukturach dane historyczne zwane **hurtowniami danych** (ang. *Data Warehouse*). Na rysunku 2 przedstawiono schemat architektury systemów OLTP i OLAP z hurtownią danych. Pokazuje on w sposób symboliczny ideę centralnej zbiornicy danych łączącej systemy OLTP i systemy OLAP.



Rysunek 2.
Schemat architektury powiązania systemów OLTP i OLAP

3 PODSTAWY HURTOWNI DANYCH

Potrzeba analizy danych dotyczących bieżącej i przyszłej działalności organizacji była podstawowym impulsem do powstania nowych systemów informatycznych. Analiza taka stanowi podstawę do podejmowania decyzji dotyczących zarządzania przedsiębiorstwem i wspomagania podejmowania decyzji. Istniejące dotychczas systemy informatyczne (głównie klasy OLTP) nie mogą dostarczyć potrzebnych danych, gdyż są oparte na operacyjnych bazach danych realizujących codzienne procesy, mogą być rozproszone (dane znajdują się w wielu różnych źródłach), niejednorodne, a często nie są z sobą powiązane. Struktury danych są dostosowane do działań operacyjnych, dane są poddawane operacjom modyfikacji. W operacyjnych bazach danych przechowuje się dane odzwierciedlające jedynie aktualny stan lub najnowszą historię, tymczasem do analiz i porównań potrzebne są długookresowe dane historyczne. Rozwiązaniem tego problemu okazała się **hurtownia danych**. Hurtownia danych jest wydzieloną centralną bazą danych zbierającą informacje służące do zarządzania organizacją. Jest ona odizolowana od baz operacyjnych, a jej struktura i użyte do jej budowy narzędzia powinny być zoptymalizowane pod kątem przetwarzania analitycznego. Prostą, najczęściej cytowaną, definicję pojęcia hurtowni danych zaproponował William H. Inmon (jeden z czołowych teoretyków hurtowni danych i systemów OLAP – autor książki *Building the Data Warehouse,* Wiley & Sons, New Jork 1996).

> *Hurtownia danych to zbiór zintegrowanych, nieulotnych, ukierunkowanych baz danych, wykorzystywanych w systemach wspomagania decyzji.*

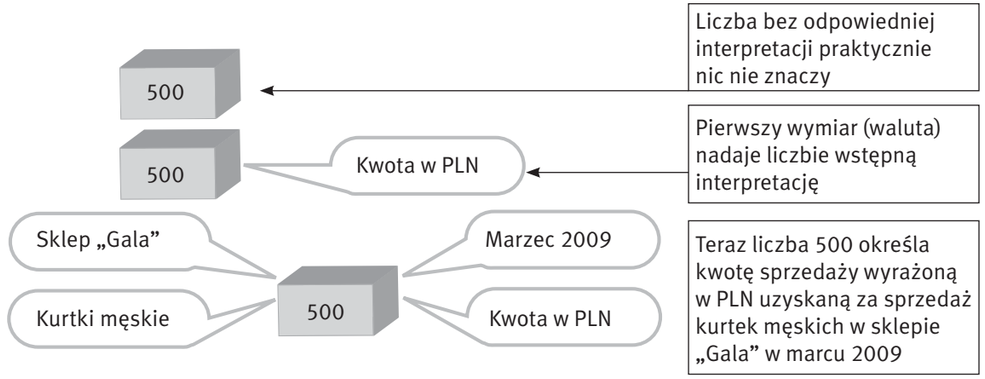
Podstawowe cechy hurtowni danych:

- **Jest scentralizowaną bazą danych** – gromadzi dane z różnych źródeł i przechowuje je w specjalnie zaprojektowanych strukturach.
- **Jest oddzielona od baz operacyjnych** – tym samym operacje wykonywane na danych gromadzonych w hurtowniach nie wpływają na wydajność operacji realizowanych w systemach OLTP.
- **Scala informacje z wielu źródeł** – ponieważ dane dotyczące jednego procesu mogą być w konkretnej firmie tworzone i przechowywane w różnych bazach danych lub nawet w plikach czy arkuszach kalkulacyjnych.

- **Jest zorientowana tematycznie** – gromadzi dane opisujące różne aspekty działalności firmy.
- **Przechowuje dane historyczne** – hurtownie mają niezaspokojony „apetyt" na dane, im dłuższa historia przechowywanych danych, tym większe możliwości analizy.
- **Utrzymuje wielką liczbę informacji** – w hurtowniach danych praktycznie nie wykonuje się operacji usuwania danych, czyli suma danych tylko rośnie wraz z dostarczaniem nowych porcji danych.
- **Agreguje informacje** – z punktu widzenia analizy najczęściej interesują nas podsumowania, obliczenia średnich i inne działania matematyczne wykonywane na grupach danych.

Najczęściej hurtownie danych są tworzone jako bazy relacyjne, w których są projektowane tabele faktów i tabele wymiarów. **Fakt** to pojedyncze zdarzenie będące podstawą analiz (np. sprzedaż produktów, udzielone kredyty itp.). Fakty są opisane przez wymiary i miary. **Miara** to wartość liczbowa dowiązana do danego faktu, np. kwota sprzedaży, liczba sztuk, a **wymiar** to cecha opisująca dany fakt, np. data, klient, produkt, lokalizacja. Dodatkowo, wymiary zawierają atrybuty, które są cechami wymiaru, np. dla wymiaru czas atrybutami mogą być miesiąc, kwartał i rok. Istotę pojęć miar i wymiarów omówimy na przykładzie. Podstawowymi elementami gromadzonymi w hurtowniach są wartości liczbowe, czyli miary pewnych faktów.

Jak pokazano na rysunku 3, wymiary są cechami opisującymi wartość miar, czyli nadają wartościom liczbowym odpowiedni sens. Najczęściej stosowanym wzorcem przy projektowaniu hurtowni jest tak zwany **schemat gwiazdy**. Na rysunku 4 przedstawiono przykładowy projekt hurtowni danych opisujący sprzedaż samochodów.

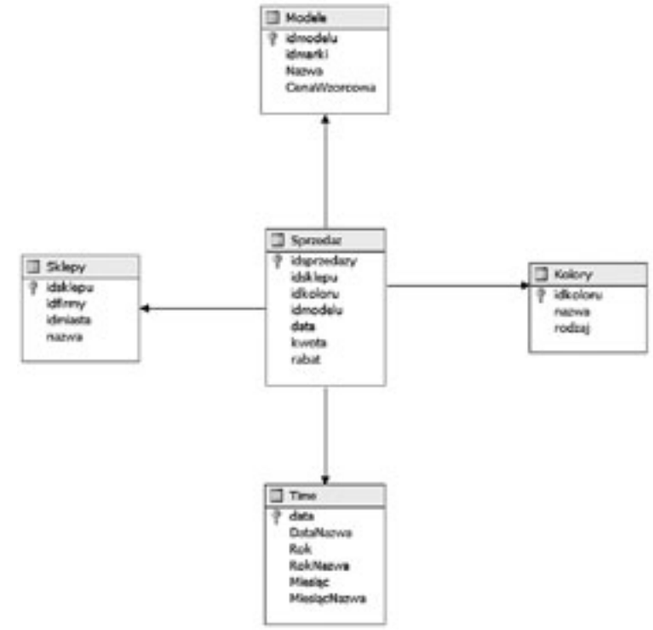


Rysunek 3.
Interpretacja miary

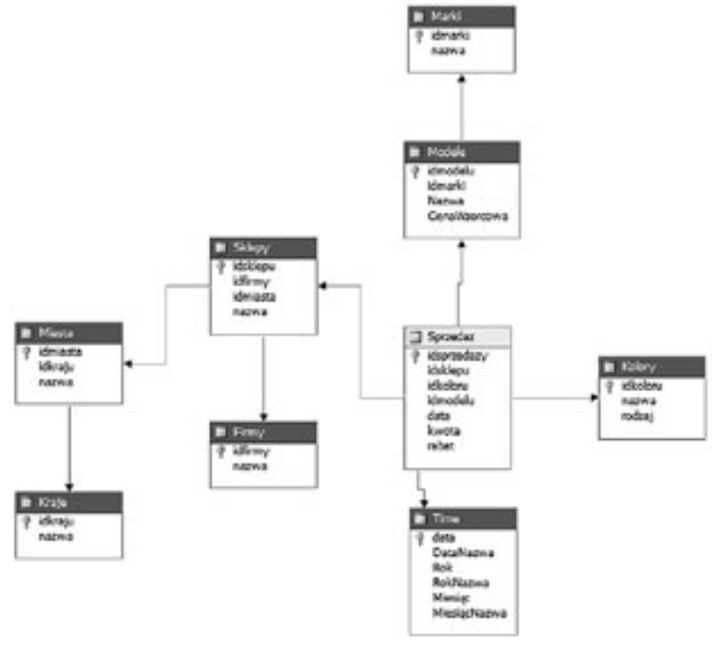
Centralną tabelą jest tabela o nazwie Sprzedaz, w której są zapisywane fakty opisujące kwoty uzyskane za sprzedaż samochodów. Tabela faktów łączy się z czterema tabelami opisującymi różne wymiary (kolor, model, sklep i czas). Połączenia tabel wymiarów z tabelą faktów są realizowane za pomocą odpowiednich kluczy obcych.

Do podstawowych cech schematu gwiazdy należy zaliczyć:
- prostą strukturę, dzięki czemu schemat jest łatwy do zrozumienia;
- dużą efektywność zapytań ze względu na niewielką liczbę połączeń tabel;
- dominującą strukturę dla hurtowni danych, wspieraną przez wiele narzędzi.

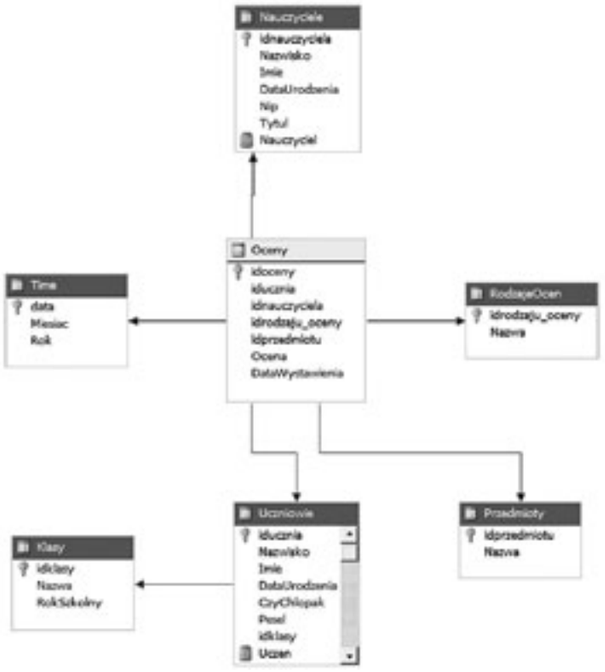
Rozwinięciem schematu gwiazdy jest schemat **płatka śniegu**, który występuje wtedy, gdy wymiary są powiązane z innymi tabelami. Na rysunku 5 przedstawiono przykładowy projekt hurtowni w schemacie płatka śniegu, który jest rozszerzeniem projektu z rysunku 4.



Rysunek 4.
Przykładowy projekt hurtowni danych w schemacie gwiazdy



Rysunek 5.
Przykładowy projekt hurtowni danych w schemacie płatka śniegu



Rysunek 6.
Przykładowy projekt hurtowni dla wystawianych ocen w szkołach

Podstawowe cechy schematu płatka śniegu:
- spadek wydajności zapytań w porównaniu ze schematem gwiazdy ze względu na większą liczbę połączeń tabel;
- struktura łatwiejsza do modyfikacji;
- wykorzystywany rzadziej niż schemat gwiazdy, gdyż efektywność zapytań jest ważniejsza niż efektywność ładowania danych do tabel wymiarów.

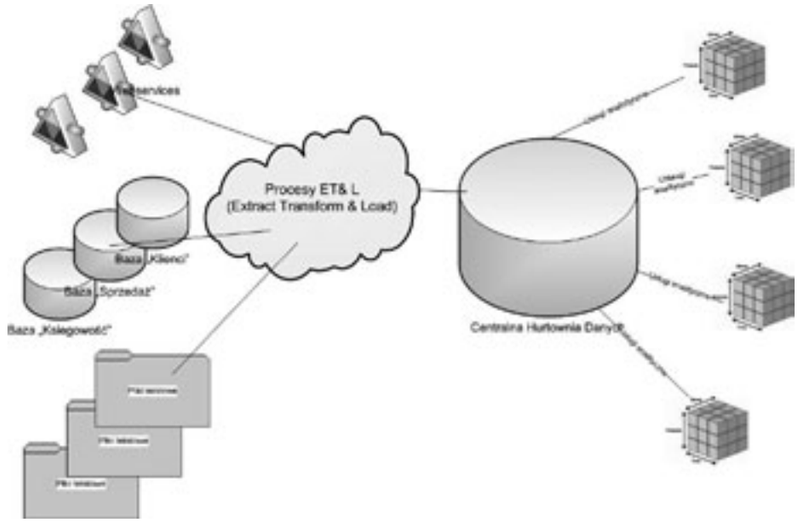
Hurtownie danych stanowią podstawowe źródło zasilające procesy analizy danych. Przedstawione przykłady projektów hurtowni są jedynie wycinkiem, gdyż w rzeczywistości hurtownie składają się z wielu podobnych struktur danych, opisujących różne fakty i korzystających z różnych wymiarów. Na rysunku 6 został przedstawiony jeszcze jeden przykład projektu struktury hurtowni danych, w którym faktami są oceny wystawione uczniom. Każda ocena jest charakteryzowana przez :
- datę jej wystawienia – wymiar Time;
- ucznia, który otrzymał ocenę – wymiar Uczniowie, który jest dodatkowo opisywany przez wymiar Klasy,
- nauczyciela, który ocenę wystawił – wymiar Nauczyciele;
- przedmiot, z którego ocena została wystawiona – wymiar Przedmioty;
- rodzaj wystawionej oceny – wymiar RodzajeOcen.

Tworzenie hurtowni danych dla jednej szkoły wydaje się niecelowe ze względu na stosunkowo niewielką liczbę danych, ale można sobie wyobrazić istnienie takiej hurtowni w skali kraju i wtedy stanowiłaby podstawę do analizy skuteczności nauczania.

Nie jesteśmy w stanie w ramach tego wykładu omówić wszystkich aspektów tworzenia hurtowni danych, gdyż są to zagadnienia złożone i praktycznie każdy projekt ma swoją specyfikę i może wyglądać zupełnie inaczej w zależności od swojego przeznaczenia i założeń, jakie dana firma przyjęła przy realizacji. Przestawione zasady stanowią punkt wyjścia przy realizowaniu konkretnego projektu.

4 PROBLEMY INTEGRACJI DANYCH

Hurtownie danych są zasilane danymi pobieranymi z systemów OLTP, które mogą być wykonane w różnych technologiach, oraz innych źródeł danych dostępnych w konkretnej firmie. Na bazie hurtowni są realizowane różne zadania analityczne. Na rysunku 7 został przedstawiony przykładowy schemat architektury takiego systemu.

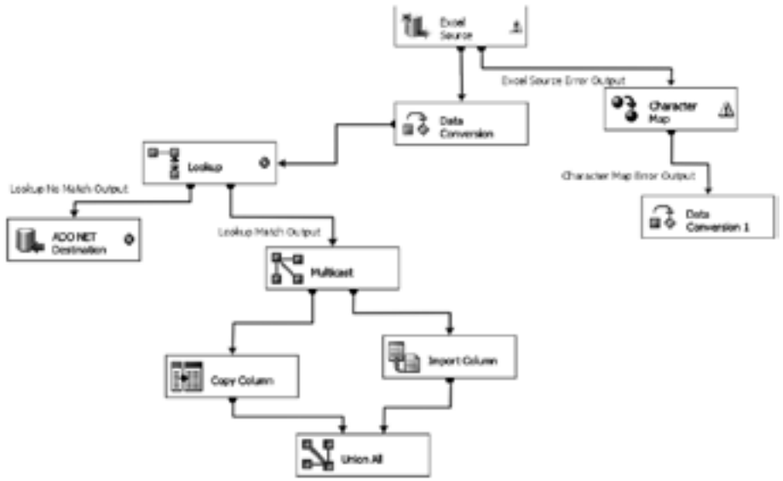


Rysunek 7.
Architektura otoczenia hurtowni danych

Przedstawiony na rysunku 7 schemat pokazuje warstwę (nazwaną procesami ET&L), która występuje pomiędzy systemami OLTP i innymi źródłami danych a hurtownią danych. Problemy związane z pozyskiwaniem danych dla hurtowni są jednymi z najtrudniejszych zadań przy jej tworzeniu. W ramach warstwy ET&L (ang. *Extract Transform & Load* – pobierz, przekształć i zapisz) są realizowane następujące zadania:
- standaryzacja danych – ponieważ dane pobierane mogą być z wielu różnego typu źródeł, to należy doprowadzić je do jednakowej postaci;
- konwersja typów danych – różne systemy mogą w inny sposób zapisywać dane i dlatego należy je doprowadzić do tego samego typu;
- transformacje danych – dane w systemach roboczych mogą być przechowywane w innej postaci niż postać ich zaprojektowana w hurtowni, dlatego należy je odpowiednio przekształcić;
- agregacja danych – w hurtowniach nie musimy zapisywać każdej elementarnej danej z systemów operacyjnych, a jedynie pewne zbiorcze wartości;
- integracja danych z różnych źródeł – dane tego samego rodzaju z punktu widzenia hurtowni (np. opis klienta) mogą być zapisywane w różnych źródłach danych i przed zapisaniem w hurtowni należy je odpowiednio powiązać;

- czyszczenie danych i kontrola poprawności – ponieważ w systemach operacyjnych mogą być przechowywane dane błędne, dlatego przed zapisaniem w hurtowni należy je sprawdzić i usunąć;
- dodatkowe przekształcenia, np. przeliczenie wartości różnych walut.

Zadania warstwy ET&L są wspierane przez różne technologie, w ramach których projektuje się i programuje działanie odpowiednich procesów. Na rysunku 8 został przedstawiony przykładowy fragment schematu procesu ET&L wykonany w MS SQL Server 2008 Integration Services.

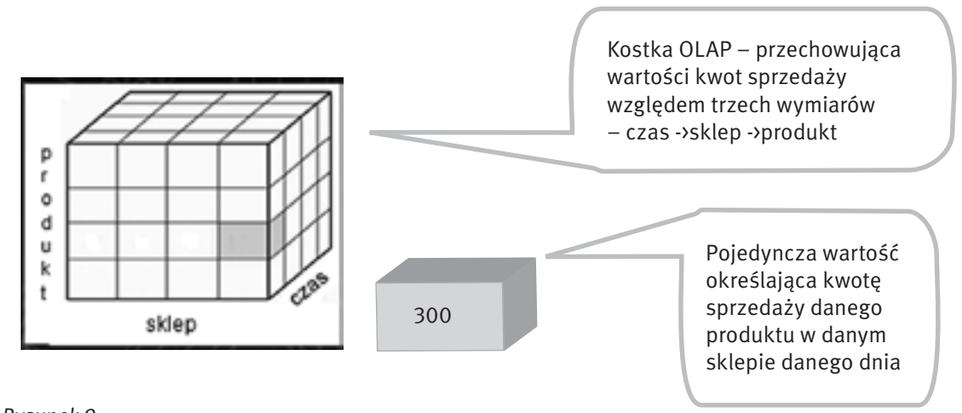


Rysunek 8.
Przykładowy pakiet usługi MS SQL Server 2008 Integration Services

Technologia MS SQL Server 2008 Integration Services umożliwia definiowanie złożonych procesów pozyskiwania, przekształcania i zapisywania danych z rożnych źródeł. Projektowany schemat przetwarzania prezentowany jest za pomocą ikon opisujących różne etapy i zadania procesu.

5 KOSTKA WIELOWYMIAROWA

Hurtownie danych stanowią punkt wyjścia do realizacji usług analitycznych. Najczęściej stosowanym elementem usług analitycznych jest **wielowymiarowa kostka OLAP**, która przechowuje dane w sposób bardziej przypominający wielowymiarowe arkusze kalkulacyjne niż tradycyjną, relacyjną bazę danych. Kostka umożliwia wyświetlanie i oglądanie danych z różnych punktów widzenia. Do jej budowy potrzeba dowolnego źródła danych opartego na tabelach relacyjnych – oznacza to, że najczęściej kostki wielowymiarowe buduje się w oparciu o hurtownie danych. Kostka składa się z miar, wymiarów oraz poziomów i jest zoptymalizowana pod kątem szybkiego i bezpiecznego dostępu do danych wielowymiarowych. **Miary** to wskaźniki numeryczne (ile?), natomiast **wymiary** reprezentują dane opisowe (kto? co? kiedy? gdzie? jak?). Wymiary są pogrupowane za pomocą **poziomów**, które odzwierciedlają hierarchię i umożliwiają użytkownikom zwiększanie lub zmniejszanie poziomu szczegółowości analizowanego wymiaru. Jak widać, kostka OLAP oparta jest na tych samych pojęciach (miary i wymiary) co schematy hurtowni danych. Trudno graficznie zaprezentować strukturę wielowymiarową – dlatego najczęściej kostka jest pokazywana w postaci sześcianu, czyli kostki złożonej z trzech wymiarów.

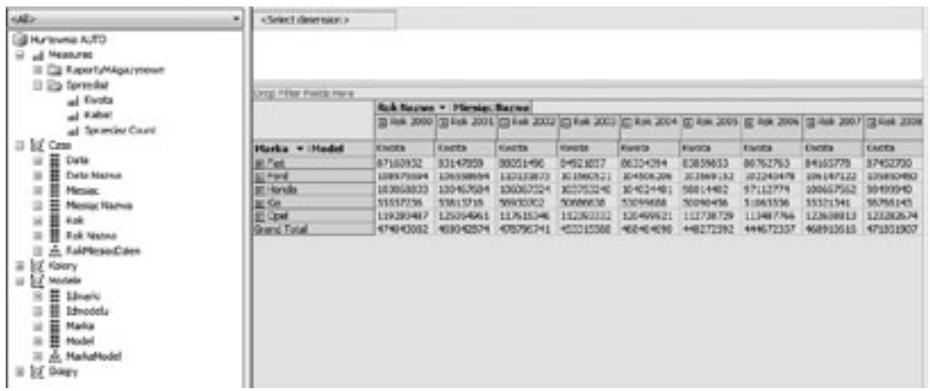


Rysunek 9.
Kostka OLAP

Podczas analizy z wykorzystaniem kostek wielowymiarowych, dane są poddawane typowym operacjom, do których zaliczamy m.in.:

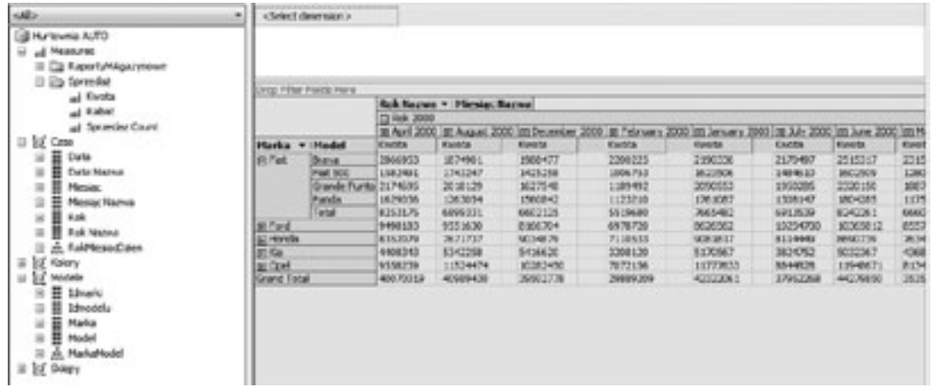
- **zwijanie** – podnoszenie poziomu agregacji, czyli uogólnianie danych;
- **rozwijanie** – zmniejszanie poziomu agregacji, dane stają się bardziej szczegółowe;
- **selekcję** – wybór interesujących elementów wymiarów;
- **projekcję** – zmniejszanie liczby wymiarów.

Obsługę tworzenia i eksploatacji kostek wielowymiarowych wspierają różne technologie, między innymi MS SQL Server 2008 Analysis Services. Na rysunku 10 przedstawiono przykładowe zestawienie na bazie kostki OLAP, opisującej sprzedaż samochodów. Zestawienie pokazuje wartość sprzedaży poszczególnych marek samochodów w kolejnych latach.



Rysunek 10.
Przykładowe zestawienie zbiorcze na bazie kostki OLAP

Kolejne zestawienie na rysunku 11 pokazuje elementy uszczegółowienia, polegające na rozbiciu kwot rocznych na poszczególne miesiące oraz rozbiciu kwot sprzedaży marki Fiat na poszczególne modele.



Rysunek 11.
Przykładowe zestawienie zbiorcze na bazie kostki OLAP z elementami uszczegółowienia

Do obsługi i pozyskiwania danych z kostek wielowymiarowych istnieje specjalny język MDX (ang. ***M**ulti **D**imensional e**X**pressions* – wyrażenia wielowymiarowe) – opis tego języka wykracza zdecydowanie poza ramy naszego wykładu. Wielowymiarowe kostki OLAP są przechowywane w specjalistycznych strukturach zoptymalizowanych pod kątem szybkości pozyskiwania danych.

6 SYSTEMY BUSINESS INTELLIGENCE

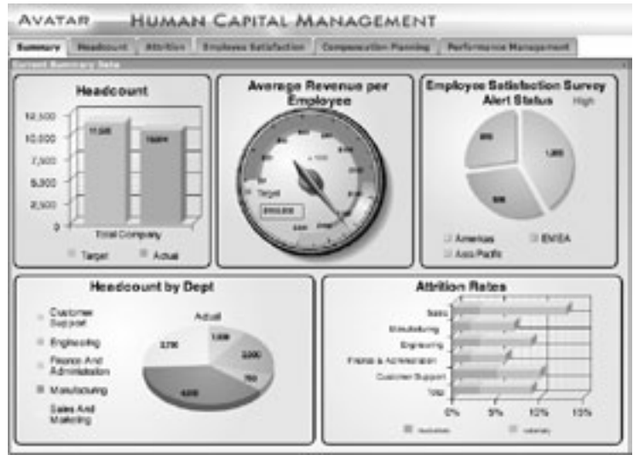
Business Intelligence (**BI**) – **analityka biznesowa** – jest pojęciem bardzo szerokim. Do dzisiaj nie istnieje powszechnie przyjmowana definicja systemów tej klasy. Najbardziej ogólnie można przedstawić je jako proces przekształcania danych w informacje, a informacji w wiedzę, która może być wykorzystana do zwiększenia konkurencyjności przedsiębiorstwa. Systemy BI są mocno uzależnione od utworzenia hurtowni danych, które umożliwiają ujednolicenie i powiązanie danych zgromadzonych z różnorodnych systemów informatycznych przedsiębiorstwa. Utworzenie hurtowni danych zwalnia systemy transakcyjne od tworzenia raportów i umożliwia równoczesne korzystanie z różnych systemów BI. System BI opiera się na następującej koncepcji:

- generuje standardowe raporty lub wylicza kluczowe wskaźniki efektywności działania przedsiębiorstwa (ang. *Key Performance Indicators*);
- na podstawie standardowych raportów i wskaźników stawia się hipotezy;
- postawione hipotezy weryfikuje się poprzez wykonywanie szczegółowych analiz danych z wykorzystaniem różnego rodzaju narzędzi analitycznych (np. OLAP, *Data Mining*).

Najczęściej spotykane odmiany systemów zaliczanych do BI to:

- EIS – systemy powiadamiania kierownictwa (ang. *Executive Information Systems*);
- DSS – systemy wspomagania decyzji (ang. *Decision Support Systems*);
- MIS – systemy wspomagania zarządzania (ang. *Management Information Systems*);
- GIS – systemy informacji geograficznej (ang. *Geographic Information Systems*).

Systemy BI są narzędziem dla menedżerów i specjalistów zajmujących się analizami i strategią. Dla menedżerów niższych szczebli, którzy oczekują informacji o aktualnym stanie procesów, przeznaczone są rozwiązania *Business Activity Monitoring* (**BAM**), umożliwiające przetwarzanie napływających na bieżąco danych. Techniki prezentacyjne są dobierane odpowiednio do potrzeb użytkownika. Jednym ze sposobów prezentowania wyników wstępnej analizy i sygnalizowania przekroczenia założonych wartości w działalności firmy jest koncepcja **kokpitu menedżera**. Idea kokpitu jest taka, aby bardzo szybko informować menedżera o wartościach podstawowych wskaźników oraz sygnalizować niekorzystne zjawiska zachodzące w jego dziedzinie odpowiedzialności. Do graficznej prezentacji takich faktów są używane proste gadżety (wskaźniki, sygnalizatory świetlne, liczniki). Elementy kokpitu powinny dać ogólny obraz procesów zachodzących w firmie. Na rysunku 12 został pokazany przykładowy kokpit menedżera.



Rysunek 12.
Przykładowa postać kokpitu menedżera [źródło: http://xelfin.pl/galeria/Galerie/1/]

Jeżeli z obrazu wskaźników kokpitu wynika problem, to należy uruchomić inne, przeważnie bardziej złożone procesy analizy.

7 EKSPLORACJA DANYCH

Eksploracja danych (spotyka się również określenie drążenie danych, pozyskiwanie wiedzy, wydobywanie danych, ekstrakcja danych) (ang. *Data Mining*) – jest jednym z etapów procesu, który bywa nazywany **odkrywaniem wiedzy z baz danych** (ang. *Knowledge Discovery in Databases*, **KDD**). Idea eksploracji danych jest oparta na wykorzystaniu komputerów i ogromnych zbiorów danych do znajdowania ukrytych dla człowieka prawidłowości w danych zgromadzonych w hurtowniach danych. Istnieje wiele technik eksploracji danych, które są oparte na zaawansowanej statystyce (statystyczna analiza wielowymiarowa) oraz technikach i metodach wywodzących się z obszaru badań nad sztuczną inteligencją. Główne przykłady stosowanych rozwiązań to:

- wizualizacje na wykresach;
- metody statystyczne;
- sieci neuronowe;
- metody uczenia maszynowego;
- metody ewolucyjne;
- logika rozmyta;
- zbiory przybliżone.

Motywację dla rozpatrywania tego typu narzędzi stanowi ciągły wzrost technicznych możliwości gromadzenia i analizy danych, w których ukryte są potencjalnie cenne informacje dopełniające wiedzę. Zastosowanie technik KDD daje szczególnie dobre wyniki w nowych dziedzinach, gdzie tak zwana wiedza ekspercka jest jeszcze w dużej mierze niepełna i nieugruntowana. Do takich dziedzin można przykładowo zaliczyć:

- analizę różnych aspektów ruchu internetowego;
- marketing z wykorzystaniem Internetu;
- rozpoznawanie obrazu, pisma, mowy itd.;
- wspomaganie diagnostyki medycznej;
- badania genetyczne;
- analizę historii operacji bankowych i zapobieganie wyłudzeniom;
- optymalizację działań (związanych z systemami CRM) zajmujących się zarządzaniem relacjami z klientami.

Proces odkrywania wiedzy z danych przebiega według poniższego schematu:

- **Zrozumienie dziedziny problemu** – złożoność danych, a także problemów stawianych przy okazji ich analizy, coraz częściej nie umożliwia natychmiastowego sformułowania pytań, na które użytkownik chce uzyskać odpowiedź. Trzeba dobrze zrozumieć problem, dla rozwiązania którego chcemy stosować techniki KDD.
- **Budowa roboczego zbioru danych** – określenie, z jakich zasobów danych będziemy korzystać w procesie KDD.
- **Oczyszczenie, przekształcanie i redukcja danych** – istotę tego problemu omówiliśmy w rozdziale poświęconym integracji danych.
- **Eksploracja danych** – realizacja procesu odkrywania wiedzy przy użyciu bardzo różnorodnych technik, opartych na statystyce, sztucznej inteligencji, czy też odwołujących się do metod uczenia maszynowego.

Rysunek 13.
Przykładowa postać kokpitu menedżera [źródło: www.shopfloorreporting.com]

Podstawowym problemem procesów odkrywania wiedzy tkwiącej w danych jest to, że różnych regularności jest w danych praktycznie nieskończenie wiele, zaś dla użytkownika interesujące będą tylko niektóre z nich i to w różnym stopniu. Osiągnięcie dobrych wyników w procesie eksploracji danych jest uzależnione nie tylko od danych i wykorzystywanych technologii, ale przede wszystkim od wiedzy i zaangażowania analityków wykonujących te zadania. Przykładowe postaci zobrazowań wyników, które można uzyskiwać w procesie eksploracji danych przedstawiono na rysunku 13.

Techniki i metody eksploracji danych są w stadium ciągłego rozwoju i należy się spodziewać nowych rozwiązań w tym zakresie.

PODSUMOWANIE

Hurtownie danych są wydzielonymi, specjalizowanymi bazami danych, przeznaczonymi do wspomagania usług analitycznych. Wdrożenie hurtowni danych może dostarczyć firmie wiele korzyści:

- **Odciążenie systemów transakcyjnych** – przygotowanie analiz i zestawień nie obciąża już systemów transakcyjnych, które mogą obsługiwać bieżące operacje. Zasilenie hurtowni danymi z systemów źródłowych wykonywane jest automatycznie i najczęściej odbywa się w cyklu dziennym, z reguły w nocy, gdy użytkownicy nie korzystają z systemu.
- **Poprawa jakości analizowanych danych** – analizując dane w hurtowni danych na zagregowanym poziomie dużo łatwiej wychwycić pewne nieprawidłowości w systemach źródłowych. W hurtowni danych bardzo dobrze widać np., czy koszty są przypisane do odpowiednich nośników, czy wszyscy klienci są przypisani do regionów sprzedaży lub handlowców itd.
- **Przechowywanie danych o długim horyzoncie czasowym** – dzięki temu, że w hurtowni danych mamy łatwy dostęp do danych wieloletnich możemy wykonywać bardziej trafne prognozy, czy też doszukiwać się określonych trendów.
- **Łączenie danych pochodzących z różnych systemów transakcyjnych –** hurtownia danych może pobrać dane z praktycznie każdego źródła danych. Dane te są następnie porządkowywane i dokonywana jest unifikacja pojęć i mierników. Dzięki temu możliwe staje się porównanie niejednorodnych danych.
- **Udostępnienie danych dla wszystkich potrzebujących –** w hurtowni danych możemy zdefiniować poszczególnym użytkownikom uprawnienia do odpowiedniego wycinka danych. Przy pomocy narzędzi analitycznych i wizualizacji danych, użytkownicy mogą wykonywać na ich bazie różne zestawienia, raporty i analizy.

LITERATURA

1. Hand D., Mannila H., Smyth P., *Eksploracja danych*, WNT, Warszawa 2002
2. Jarke M., Lenzerini M., Vassiliou Z., Vassiliadis P., *Hurtownie danych. Podstawa organizacji funkcjonowania*, WSiP, Warszawa 2003
3. Poe V., Klauer P., Brobst S., *Tworzenie hurtowni danych*, WNT, Warszawa 2000
4. Surma J., *Business Intelligence*. *Systemy wspomagania decyzji biznesowych*, WN PWN, Warszawa 2009
5. Todman Ch., *Projektowanie hurtowni danych*, WNT, Warszawa 2003