Hurtownie danych i systemy wspomagania decyzji

Olaf Morawski
Hewlett-Packard Polska Sp. z o.o.,
ul. Szturmowa 2A, 02-678 Warszawa

Poniższy tekst opisuje architekturę systemów wspomagania decyzji, z uwzględnieniem głównych cech hurtowni danych i specjalizowanych narzędzi analitycznych.

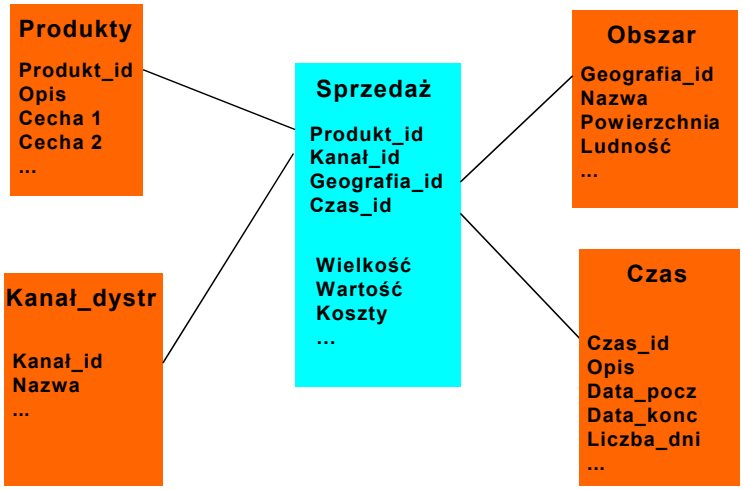
Wstęp

Przedsiębiorstwa muszą odpowiadać sobie na pytania o to jak są postrzegane na rynku, jak zmienia się w czasie sprzedaż produktów i usług lub, którzy klienci przynoszą największy zysk. W celu odpowiedzi na te, zasadnicze dla firmy jako całości, pytania rozwinięto obok finansowej - rachunkowość zarządczą. Dzięki innej perspektywie patrzenia na dane księgowe, wyłaniał się całościowy obraz firmy przedstawionej w kategoriach ekonomicznych. Taki obraz umożliwiał właściwe planowanie i podjęcie działań stosownych dla dalszego wzrostu firmy. Okazało się jednak, że dla uzyskania takiego obrazu, konieczne jest stworzenie zupełnie nowych systemów informatycznych dedykowanych informacji zarządczej. Systemy takie nazwano systemami wspomagania decyzji, a bazy danych gromadzące dane dla tych systemów – hurtowniami danych.

Technologia i architektura hurtowni danych

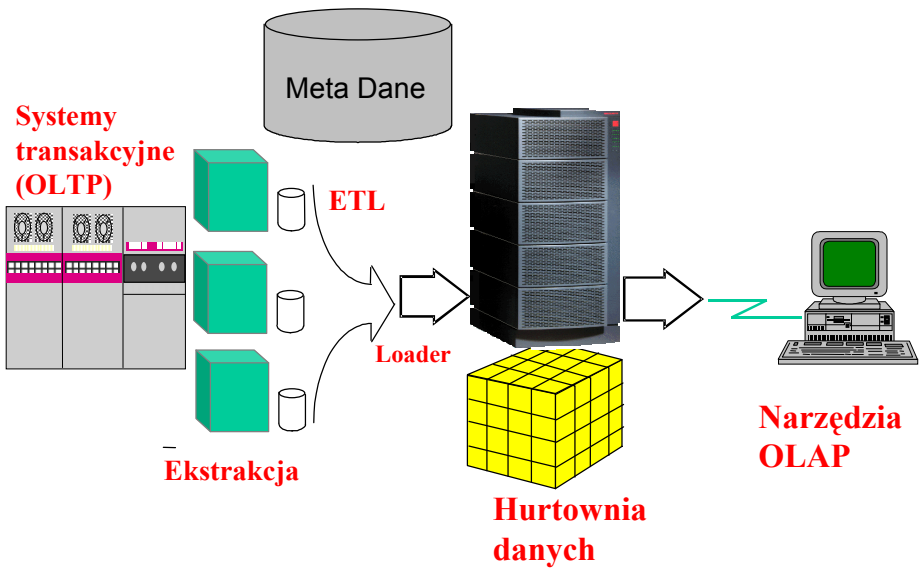
Hurtownia danych jest, zgodnie z definicją, skonsolidowanym repozytorium danych historycznych. Repozytorium zawiera dane historyczne, nie podlegające już zmianom (bo nie zmieniamy danych o sprzedaży sprzed np. roku). Dane są skonsolidowane, co rzutuje na architekturę bazy danych – jest ona znacznie mniej złożona niż w przypadku baz OLTP (On Line Transactional Processing – baz danych dla przetwarzania transakcyjnego). Ponadto dane są zorganizowane w strukturę wielowymiarową, w której fakty (liczby przechowywane w bazie) są zależne od wielu parametrów (nazywanych wymiarami). W najprostszym przypadku baza ma strukturę gwiazdy, w której (por. rys. 1) dane dotyczące sprzedaży (wielkość, wartość, …) gromadzone są w centralnej tabeli faktów, natomiast parametry (wymiary) od których sprzedaż zależy znajdują się w mniejszych tabelach takich jak np. produkt, geografia czy czas. Elementy wymiarów (dla geografii np. województwa) mogą posiadać cechy charakterystyczne (np. rozmiar, liczba ludności), które będą wykorzystywane w analizach. Schemat gwiazdy odpowiada strukturze najbardziej zdenormalizowanej, w innym schemacie - płatka śniegu – pojawia się normalizacja tabel wymiarów.

Zdenormalizowana struktura bazy hurtowni danych powstała w wyniku procesu poszukiwania modelu logicznego bazy, który umożliwiałby wydajne, złożone i wszechstronne raportowanie. Próby usprawnienia raportowania i prowadzenia analiz w oparciu o tradycyjne bazy OLTP poniosły fiasko. Kiedy okazało się, że istnieją fundamentalne przyczyny dla których generowanie złożonych raportów w oparciu o systemy OLTP nie może być wydajne, zaczęto zastanawiać się nad nowym sposobem gromadzenia i udostępniania danych. Poszukiwania rozwiązania tego problemu doprowadziły do koncepcji hurtowni danych i nowego – zdenormalizowanego - modelu danych.



Rysunek 1. Struktura gwiazdy

Organizacja danych w strukturze wielowymiarowej nie jest jedynym wyróżnikiem hurtowni danych. Ważnym elementem tych systemów są procedury ekstrakcji, czyszczenia, transformacji i ładowania danych do bazy (*ang. Extract, Transformation, Load – ETL*). Procedury ekstrakcji danych z systemów OLTP uruchamiane są w czasie minimalnego obciążenia tych systemów. Dane wyekstrahowane są następnie weryfikowane względem reguł i danych słownikowych przechowywanych w repozytorium metadanych (rys. 2), przekształcane do pożądanej w hurtowni postaci i następnie ładowane do bazy. Dzięki procedurom ETL dane w hurtowni charakteryzują się wysoką jakością, przewyższającą znacznie jakość danych systemów OLTP. Za prosty przykład niech posłuży *deduplikacja* – procedura usuwająca powtórzenia danych: w systemach OLTP panowie Jan Kowalski i Jan Piotr Kowalski mogą być różnymi klientami, mimo iż mieszkają pod tym samym adresem, w hurtowni (dzięki procesowi *deduplikacji*) zostaną zidentyfikowani jako ta sama osoba umożliwiając tym samym rzeczywistą analizę zachowań klienta. Zauważmy, że prowadzenie ekstrakcji danych z systemów transakcyjnych w godzinach nocnych oraz umieszczenie hurtowni danych na oddzielnym serwerze odciążyło systemy OLTP umożliwiając efektywne ich wykorzystanie do zadań im dedykowanych.



Rysunek 2. Zasilanie hurtowni danymi i dostarczanie analiz użytkownikom końcowym.

Kolejnym ważnym elementem stosowanym w hurtowniach danych są narzędzia analityczne. Aby uniknąć konieczności budowania dla kolejnych raportów coraz to nowych programów stworzono narzędzia analityczne umożliwiające tworzenie w bardzo prosty sposób, praktycznie tylko za pomocą myszy, nawet bardzo złożonych raportów. Stworzono klasę narzędzi pozwalających prowadzić nawet skomplikowane analizy w trybie *on-line* (On Line Analitycal Processing - OLAP). Systemy klasy OLAP przełamały wszystkie ograniczenia systemów raportujących z baz OLTP umożliwiając prowadzenie różnego rodzaju analiz biznesowych także na bardzo dużych bazach danych. Powszechnie zaczęto wykorzystywać te nowe możliwości dla wsparcia biznesu oferując mu wszechstronne raportowanie, elastyczne planowanie oraz zawansowane analizy finansowe, sprzedaży, klientów i rynku. Nazwy takie jak Systemy Wspomagania Decyzji (DSS – Decision Support Systems) czy MIS (Management Information Systems) stały się znane kadrze kierowniczej wielu firm, a korzystanie z tych systemów – powszechną praktyką.

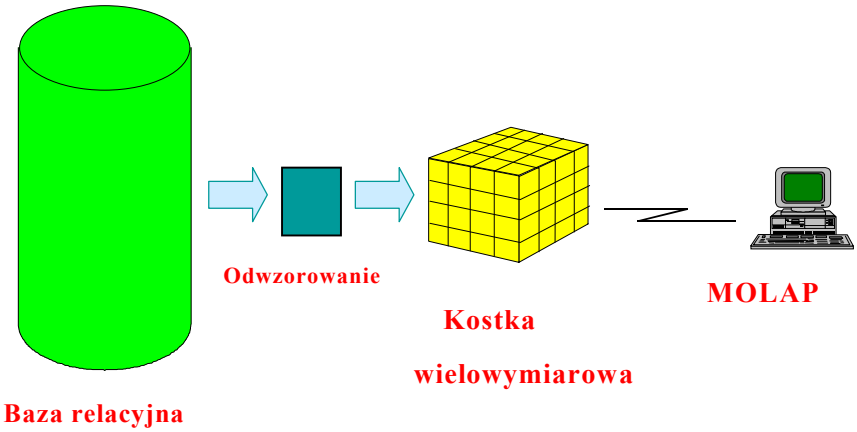
Od samego początku hurtownie danych realizowane były w oparciu o bazy danych dwojakiego rodzaju: relacyjne i wielowymiarowe. Relacyjne bazy danych wymyślono w 1966 roku i obecnie są bardzo szeroko stosowane do wielu zadań. Informacja w bazach relacyjnych gromadzona jest w tabelach (relacjach) połączonych ze sobą więzami integralności. Najbardziej znanymi przedstawicielami tych baz są Oracle, DB2, Informix, MS SQL Server czy Sybase. Omawiana powyżej struktura gwiazdy jest przykładem modelu danych bazy relacyjnej. W takiej strukturze fakty są funkcją wielu zmiennych (wymiarów), i jest to jedna z bardzo wielu możliwych struktur danych bazy relacyjnej. W bazie wielowymiarowej dane są przechowywane z definicji w postaci faktów – funkcji kilku zmiennych, i jest to podstawowa forma gromadzenia danych w tej bazie. Przedstawicielami baz wielowymiarowych są np. Oracle Express czy Essbase (Hyperion).

Zarówno relacyjne i wielowymiarowe baz danych posiadają zalety i wady, i są stosowane w nieco różnych dziedzinach. Relacyjne bazy danych mogą gromadzić olbrzymie ilości danych osiągając rozmiar wielu Terabajtów. Bazy wielowymiarowe natomiast mają ograniczenie na wielkość i przeważnie nie przekraczają rozmiarem kilkudziesięciu Gigabajtów. Oferują natomiast strukturę danych przygotowaną do prowadzenia wszechstronnych i złożonych raportów na podstawie różnorakich przecięć przez dane wielowymiarowe. Z kolei bazy relacyjne, początkowo przynajmniej, nie posiadały zaawansowanych narzędzi analitycznych.

Dla każdego typu baz danych opracowano specjalne narzędzia analityczne klasy OLAP. Do baz wielowymiarowych stworzono narzędzia typu MOLAP (Multidimensional OLAP), do baz relacyjnych ROLAP (Relational OLAP). Przedstawicielami narzędzi MOLAP są na przykład Oracle Sales Analyzer a ROLAP – Oracle Discoverer, Business Objects czy DSS Agent (MicroStrategy).

Dla połączenia zalet baz danych obu typów często proponuje się architekturę hybrydową wykorzystującą obie bazy jednocześnie. Na rysunku 3 przedstawiono taką strukturę, w której dane gromadzone są w bazie relacyjnej a następnie odwzorowywane (przez aplikację mapującą) na strukturę wielowymiarowej bazy danych (zaznaczonej symbolicznie kostką). Baza wielowymiarowa nie zawiera faktów a tylko elementy wymiarów, jej rozmiar jest zatem bardzo mały, nawet jeśli baza relacyjna jest wielkości Terabajtów. Do analiz stosowane jest narzędzie analityczne klasy MOLAP. W czasie analizy zapytanie z narzędzia MOLAP jest przesyłane do wielowymiarowej bazy danych, gdzie generowane jest dynamiczne zapytanie SQL, które jest przesyłane przez warstwę odwzorowującą do bazy relacyjnej. Wynik

zapytania powraca do narzędzia analitycznego tę samą drogą (oczywiście w odwrotnej do zapytania kolejności).



Rysunek 3. Architektura hybrydowa hurtowni danych. Dane gromadzone są w bazie relacyjnej, odwzorowane do struktury wielowymiarowej i poprzez nią udostępniane do analizy.

Narzędzia analityczne można podzielić pod względem złożoności prowadzonych analiz na trzy grupy:

(i) proste narzędzia raportowe służące tworzeniu powielanych raportów wykorzystywanych przez szerokie rzesze użytkowników biznesowych. Narzędzia te umożliwiają utworzenie tabelarycznych lub graficznych raportów szeroko dostępnych poprzez sieć korporacyjną. Raporty są odświeżane przy każdym uzupełnieniu hurtowni o nowe dane. Służą głównie prezentacji wybranych wskaźników i dlatego są często nazywane *raportami standardowymi*.

(ii) narzędzia klasy OLAP służące tworzeniu dowolnych, różnych raportów (*ad-hoc*). Narzędzia OLAPowe umożliwiają tworzenie przekrojów przez wielowymiarowe kostki danych. Takie przekroje pozwalają na odkrywanie zależności pomiędzy miarami i elementami wymiarów, na przykład wykrycie, który region jest odpowiedzialny za spadek sprzedaży. Narzędzia tej klasy są wykorzystywane przez analityków biznesowych dla ustalania przyczyn zdarzeń biznesowych (wzrost/spadek sprzedaży, skuteczność kampanii reklamowej czy promocji, itp.) oraz śledzenia trendów.

(iii) zaawansowane narzędzia drążenia i eksploracji danych (*ang. Data Mining*) służące do automatycznego znajdowania związków między danymi. Narzędzia klasy DataMining wykorzystują wiele wyrafinowanych technik takich jak na przykład sieci neuronowe, drzewa decyzyjne, sieci Bayesa, algorytmy genetyczne, clustering czy regresja. Narzędzia tej klasy są wykorzystywane przez analityków między innymi do segmentacji bazy klientów, prognozowania, pozycjonowania produktu na rynku, a także do wykrywania oszustw w czasie rzeczywistym. DataMining, choć oferuje automatyczne generowanie wyników, wymaga dobrego ich zrozumienia (w celu uniknięcia pułapek) i dlatego prowadzony jest zwykle przez zaawansowanych analityków często posiadających doktorat z matematyki czy nauk przyrodniczych. Dla ułatwienia i usystematyzowania analiz drążenia danych opracowano w 1996 roku metodykę CRISP DM (*Cross-Industry Standard Process for Data Mining*), w której określono sposób prowadzenia analizy od zrozumienia zadań biznesowych i dostępnych danych poprzez

przygotowanie danych i modelowanie aż po oszacowanie poprawności modelu i jego wdrożenie do eksploatacji. Metodyka ta jest dziś wspierana przez praktycznie wszystkich wytwórców oprogramowania klasy DataMining.

Zastosowania hurtowni danych

Hurtownie danych są dziś stosowane w licznych firmach praktycznie wszystkich działów gospodarki. Dzięki analizom danych zawartych w hurtowniach, firmy mogą podjąć działania mające na celu zwiększenie sprzedaży, zmniejszenie kosztów własnych, lepszą obsługę klienta, ograniczenie kosztów promocji i szereg innych.

W bankowości od początku hurtownie danych budowane były z myślą o wspomaganiu zarządzania i ułatwieniu:

- Oceny sytuacji finansowej oddziałów i planowania rozwoju,
- Badania zyskowności produktów i usług oraz kształtowania ich portfela,
- Analizy kredytowej i szacowaniu ryzyka,
- Analizy płatności, należności i zaległości.

Dla powyższych celów budowano tzw. hurtownie finansowe. W okresie ostatnich lat tworzone są w bankach hurtownie danych klientów. Bank dysponując wszechstronnymi danymi o każdym ze swych klientów może zaoferować lepsze usługi i produkty. Jest to szczególnie ważne dziś, gdy konkurencja w sektorze bankowym jest ogromna. Bankom opłaca się tworzyć olbrzymie hurtownie danych klientów i, współdziałające z nimi, systemy 'Zarządzania Relacjami z klientami' (*ang. CRM – Customer Relationship Management*), aby podnosząc jakość usług sprostać konkurencji i zatrzymać klientów banku. Dzięki zaawansowanym analizom danych klientów bank może dokonać segmentacji klientów – podziału klientów na grupy o np. różnej zyskowności. Segmentacja zezwala na zróżnicowane traktowanie potrzeb różnych klientów – przedstawianie dodatkowej oferty najbardziej zyskownym klientom i kontakt z nimi poprzez dedykowanych opiekunów jak również ograniczenie kontaktów z klientami najmniej zyskownymi do minimum i, znaczną dzięki temu, redukcję kosztów ich obsługi. Możliwe jest też prowadzenie ukierunkowanego marketingu - np. wysyłanie listów tylko do wybranych klientów, planowanie kontaktów, oferowanie 'produktów niszowych' dla wybranych, małych grup klientów, uzgadnianie marketingu szczebla banku z marketingiem oddziału itp. Ogólnie zastosowanie hurtowni danych, w których zgromadzono dane o klientach pochodzące z różnych systemów informatycznych poszczególnych produktów czy usług bankowych oferuje następujące możliwości dla banku:

- zahamowanie odpływu klientów,
- zwiększony napływ nowych klientów,
- zwiększone obroty dla wielu produktów,
- zmniejszenie złych długów,
- ograniczenie oszustw,
- zmniejszenie kosztów marketingowych ('inteligentne' kampanie, itp.).

W sektorze ubezpieczeń, podobnie jak w bankowości, hurtownie danych okazały się bardzo przydatne i w sposób wymierny wspomagają podejmowanie decyzji biznesowych. Hurtownie danych, w których zgromadzono dane o klientach pochodzące z różnych systemów informatycznych poszczególnych produktów czy usług ubezpieczeniowych umożliwiają:

- zwiększenie zysku z istniejących polis poprzez
  - ograniczenie ryzyka,

  - ograniczenie fałszerstw,
  - ustanowienie stawek zapewniających odpowiedni zysk,
  - ograniczenie kosztów marketingowych i sprzedaży związanej z produktami (agenci, niezależni akwizytorzy),
- wprowadzenie na rynek nowych produktów i przejęcie części rynku od innych instytucji w sektorach ubezpieczeń emerytalnych, zdrowotnych, majątkowych, kształcenia i opiekuńczych (dzięki dokładnej znajomości potrzeb klientów).

W telekomunikacji hurtownie danych wykorzystujące dane bilingowe umożliwiają między innymi segmentację klientów na grupy w różny sposób korzystających z usług operatora. To pozwala na ustanowienie taryf dedykowanych specjalnie dla tych grup. Ponadto, poprzez obliczanie tzw. wskaźnika 'churn' można określić, którzy klienci noszą się z zamiarem zrezygnowania z usług firmy telekomunikacyjnej. Aby zatrzymać takich klientów, firma może proponować nowe stawki taryf, specjalne promocje itp. Obserwacja zachowań klientów może (u operatorów telefonii bezprzewodowej) wskazać także obszary niepoprawnego działania sieci. Ostatnio podejmuje się próby wykrywania w czasie rzeczywistym oszustw. Do tego celu stosuje się bardzo zaawansowane techniki analityczne typu Data Mining z wykorzystaniem sztucznej inteligencji. (Podobne działanie podejmuje się obecnie w bankach dla wykrywania oszustw na kartach kredytowych.)

We wszystkich sektorach rynku istotna jest opieka nad najlepszymi klientami. Jak pokazują statystyki utrata nawet ułamka (rzędu 5%) najlepszych klientów przekłada się na znaczące (rzędu kilkunastu – kilkudziesięciu procent) straty finansowe. Dotyczy to zarówno banków jak firm telekomunikacyjnych czy ubezpieczeniowych.

W handlu hurtownie danych stały się istotnym narzędziem wspomagającym sprzedaż, marketing, promocje czy nawet sposób wystawiania towarów w sklepie. Dzięki analizie danych dziesiątek i setek tysięcy transakcji dla tysięcy produktów oferowanych w supermarketach (jest to tzw. analiza koszyka) można dobrze określić preferencje klientów i korelacje pomiędzy produktami. Umożliwia to handlowcom wyjście na przeciw oczekiwaniom klientów i zwiększenie sprzedaży. Dzięki takiej analizie można też np. ograniczyć zbędne promocje towarów, które i tak dobrze się sprzedają oraz wycofać lub ograniczyć stany towarów 'wolno rotujących' na półkach. Wszystkie te akcje przekładają się na wymierne wyniki finansowe, znacznie przewyższające koszty wdrożenia hurtowni danych.