# Aktywne hurtownie danych

Mikolaj Morzy

Instytut Informatyki Politechniki Poznanskiej
ul. Piotrowo 3A, 60-965 Poznan
Mikolaj.Morzy@cs.put.poznan.pl

**Abstract.** Tradycyjne hurtownie danych staly sie nieodlacznym skladnikiem infrastruktury informatycznej wielu przedsiebiorstw. Wraz z gwaltownie nastepujacym postepem technicznym pojawiaja sie jednak wyzwania, którym tradycyjna architektura hurtowni danych nie moze sprostac. W szczególnosci, tradycyjne hurtownie danych wspieraja dzialalnosc przedsiebiorstwa na szczeblu strategicznym, lecz nie oferuja taktycznego wsparcia podejmowania decyzji na szczeblu operacyjnym. Aby sprostac tym wymaganiom, zaproponowano nowy model informatycznej infrastruktury przedsiebiorstwa, zwany aktywna hurtownia danych. Aktywna hurtownia danych umozliwia współdzielenie danych skladowanych w operacyjnych bazach danych z danymi przechowywanymi w hurtowniach danych i stanowi istotny komponent umozliwiajacy integracje aplikacji dzialajacych w przedsiebiorstwie. W niniejszym artykule przedstawiamy pojecie aktywnej hurtowni danych i wskazujemy na czynniki, które decyduja o powodzeniu wdrozenia takiego rozwiazania. Omawiamy równiez korzysci, jakie moga zostac odniesione przez przedsiebiorstwo w efekcie wdrozenia aktywnej hurtowni danych. Dyskusje ilustrujemy przykladami wdrozen, które zakonczyly sie sukcesem. Artykul zamyka dyskusja na temat integracji aplikacji przedsiebiorstwa i roli, jaka w tym procesie odgrywa aktywna hurtownia danych.

## Wprowadzenie

### Model przetwarzania analitycznego on-line

Sposób, w jaki uzytkownik korzysta z systemu komputerowego, nazywamy modelem przetwarzania. W dziedzinie baz danych najbardziej rozpowszechnionym modelem przetwarzania jest *przetwarzanie transakcji w trybie on-line* (ang. on-line transaction processing, OLTP). Model ten charakteryzuje sie zalozeniem, ze w systemie wykonuja sie wspólbieznie krótkie transakcje, z których kazda dokonuje odczytu lub zapisu niewielkiej ilosci danych, zazwyczaj kilku rekordów. Model OLTP zaklada równiez, ze równolegle wykonuja sie setki lub tysiace transakcji o podobnej charakterystyce. W systemach implementujacych taki model przetwarzania kluczowe znaczenie maja zagadnienia takie jak zapewnienie spójnosci danych, integralnosc danych, maksymalizacja współbieznosci przetwarzania, zapewnienie odpowiedniego poziomu izolacji miedzy transakcjami, czy udostepnienie wydajnych mechanizmów wycofywania lub odtwarzania transakcji, które ulegly uszkodzeniu. Model OLTP dobrze charakteryzuje przetwarzanie typowe dla systemów sluzacych do obslugi biezacej dzialalnosci przedsiebiorstwa, takich jak systemy ksiegowe, finansowe, kadrowe, czy magazynowe.

Współczesne przedsiebiorstwa wymagaja jednak bardziej zaawansowanych narzedzi informatycznych aby radzic sobie w obliczu ostrej konkurencji. Jednymi z najbardziej uzytecznych narzedzi sa systemy wspomagania decyzji. Systemy te umozliwiaja analize i wykorzystanie ogromnych ilosci danych gromadzonych przez przedsiebiorstwa. Wiekszosc przedsiebiorstw, organizacji, instytucji publicznych, czy osrodków naukowych gromadzi ogromne ilosci danych. Dane te, przechowywane w heterogenicznych bazach danych, arkuszach kalkulacyjnych lub dokumentach tekstowych, sa dodatkowo rozproszone geograficznie i niespójne ze wzgledu na lokalizacje narodowe. Mimo, ze ich integracja jest zadaniem trudnym, niemniej jednak jest oplacalna. Wynika to z faktu, ze zgromadzone archiwa zawieraja w sobie bardzo duza ilosc wartosciowej wiedzy, która moze byc wykorzystana do wspierania dzialalnosci przedsiebiorstwa, w szczególnosci do wspomagania decyzji. Informacje zawarte w zgromadzonych danych opisuja trendy, anomalie, regularnosci, korelacje i uzyteczne wzorce, np. wzorce zachowan klientów, wzorce okresowych wahan sprzedazy, itp. Aby umozliwic wykorzystanie danych gromadzonych przez przedsiebiorstwo, zaproponowano nowy model przetwarzania, nazwany *przetwarzaniem analitycznym on-line* (ang. on-line analytical processing) i opracowano nowa architekture systemu baz danych, nazwana hurtownia danych.

Model przetwarzania analitycznego on-line zaklada, ze w systemie wykonuje sie wspólbieznie niewiele transakcji (zazwyczaj kilka do kilkunastu), przy czym kazda transakcja posiada dwie charakterystyczne cechy: dokonuje tylko i wylacznie odczytów danych i przetwarza jednorazowo bardzo duzy wolumen danych rzedu tysiecy lub milionów rekordów. Wydajnosc przetwarzania mierzy sie nie, jak w przypadku modelu OLTP, liczba transakcji w jednostce czasu, lecz czasem wykonania pojedynczej transakcji. Poniewaz transakcje dokonuja

tylko i wylacznie odczytów danych, stad zagadnienia kluczowe dla modelu OLTP, np. synchronizacja wspólbieznego wykonania zbioru transakcji, w modelu OLAP staja sie nieistotne. Z drugiej strony, dla efektywnosci przetwarzania wazne staja sie optymalizacja zapytan, efektywne struktury dostepu do danych, wstepna agregacja i materializacja danych.

**Hurtownia danych**

Model przetwarzania OLAP stal sie podstawa hurtowni danych. Hurtownia danych to „(…) zorientowana tematycznie, zintegrowana, zmienna w czasie i nieulotna kolekcja danych wspierajaca proces wspomagania decyzji" (W.H.Inmon). Hurtownia danych jest zorientowana tematycznie, co oznacza, ze dane przechowywane w hurtowni danych stanowia wyczerpujacy opis pewnego fragmentu rzeczywistosci, w której dziala dane przedsiebiorstwo. Zazwyczaj perspektywa danych przechowywanych w hurtowni danych rózni sie znaczaco od perspektywy danych przechowywanych w operacyjnej bazie danych. Dla przykladu, dane w systemie transakcyjnym banku przedstawiaja dzialalnosc banku jako sekwencje transakcji, podczas gdy hurtownia danych banku jest zorientowana na klientów i ich konta, lokaty, kredyty, itp. Integracja hurtowni danych polega na tym, ze dane przechowywane w hurtowni pochodza z wielu potencjalnie heterogenicznych zródel danych. Integracja danych z wszystkich galezi dzialania przedsiebiorstwa jest kluczem do osiagniecia wysokiej jakosci hurtowni danych. Dane w przedsiebiorstwie sa przechowywane w wielu izolowanych systemach, dzialajacych na róznych platformach i stworzonych w róznych technologiach. Integracja polega na ich scaleniu, ujednoliceniu formatów i nazewnictwa, konwersji danych, itp. Zmiennosc w czasie spowodowana jest tym, ze hurtownia danych zawiera dane historyczne. Hurtownie danych mozna postrzegac jako zbiór migawek, z których kazda przedstawia globalny stan informacji obecnej wewnatrz przedsiebiorstwa w danym momencie. Horyzont czasowy jest bardzo istotna skladowa hurtowni danych, umozliwiajaca analize trendów i wzorców zmiennych w czasie oraz ewolucje rzeczywistosci, w której dziala dane przedsiebiorstwo. Wreszcie nieulotnosc danych zgromadzonych w hurtowni danych powoduje, ze dane które zostaly wprowadzone do hurtowni danych nigdy nie ulegaja zadnym modyfikacjom. Najczesciej jedynymi dozwolonymi modyfikacjami zawartosci hurtowni danych jest odswiezenie hurtowni (wprowadzenie nowych danych z systemów zródlowych) i archiwizacja [1,4].

Hurtownie danych sa narzedziem skierowanym przede wszystkim do analityków, strategów i decydentów. Umozliwiaja im analize duzych ilosci danych i okreslanie trendów, potwierdzanie lub obalanie hipotez, identyfikacje zasobów (klientów, produktów lub uslug) przynoszacych wysoki dochód lub wysokie straty. Analiza zawartosci hurtowni danych najczesciej odbywa sie na podstawie zapytan formulowanych przez uzytkownika, gdzie poszczególne zapytania analizuja dane atomowe zgromadzone w tzw. *tabeli faktów* (ang. fact table) w kontekscie powiazanych z faktami wymiarów. Fakty reprezentuja podstawowe „cegielki" skladajace sie na analizowana rzeczywistosc (np. pojedynczy fakt moze reprezentowac sprzedaz produktu, podpisanie umowy z klientem, lub wykonanie polaczenia telefonicznego) i zazwyczaj przechowuja mierzalna informacje o danym fakcie, zwana *miara* (ang. measure). Przykladem miary moze byc cena, liczba sprzedanych produktów, lub czas polaczenia telefonicznego. Wymiary, przechowywane w *tabelach wymiarów* (ang. dimension tables) sluza do opisywania faktów. Przykladami wymiarów sa: czas, klient, produkt, czy abonament telefoniczny. Wymiary moga tworzyc hierarchie umozliwiajace na analize danych z podzialem na miasta, powiaty, województwa (hierarchia wymiaru opisujacego lokalizacje) lub z podzialem na miesiace, kwartaly i lata (hierarchia wymiaru opisujacego czas). Zapytania kierowane przez uzytkowników do hurtowni danych dokonuja wyliczania zlozonych agregatów wykorzystujac do tego zdefiniowane wymiary. Przykladem zapytania analitycznego moze byc zadanie wyliczenia sredniej sprzedazy danej grupy produktów z podzialem na lokalizacje sklepu, kategorie klientów i miesiace. Zapytania takie sa formulowane przez analityków, menedzerów i decydentów w celu budowania strategicznej wizji dzialalnosci przedsiebiorstwa. Warto tutaj zauwazyc, ze grono uzytkowników tradycyjnej hurtowni danych jest bardzo ograniczone. Co wiecej, wzorce odkryte w hurtowni danych moglyby wspierac podejmowanie decyzji taktycznych w przedsiebiorstwie, jesli istnialaby mozliwosc udostepnienia tych danych uzytkownikom koncowym (sprzedawcom, urzednikom, dostawcom, itp.). Wspólczesnie dostepne hurtownie danych nie posiadaja takich mozliwosci. Wlasnie ten fakt stal sie zródlem krytyki wielu specjalistów i spowodowal opracowanie nowej architektury aktywnych hurtowni danych.

## Operacyjna skladnica danych

Podstawowym problemem informatycznym z jakim borykaja sie wspólczesne przedsiebiorstwa jest brak zintegrowanego i aktualnego obrazu danych wykorzystywanych przez dane przedsiebiorstwo. Jak wskazano wczesniej, tradycyjna hurtownia danych tylko czesciowo rozwiazuje ten problem. Po pierwsze, dane przechowywane w hurtowni nigdy nie sa aktualne i w zaleznosci od procedury odswiezania hurtowni moga

reprezentowac migawke sprzed wielu dni lub tygodni. Po drugie, dane w hurtowni czesto sa juz przetworzone do postaci wymaganej przez narzedzia analityczne i raportujace. Taka transformacja moze za soba pociagac uogólnienie danych i utrate informacji o danych atomowych. Koniecznosc posiadania zintegrowanych i aktualnych danych rodzi potrzebe budowania *operacyjnej skladnicy danych* (ang. operational data store, ODS).

**Koniecznosc wyodrebnienia operacyjnej skladnicy danych**

Hurtownia danych przedsiebiorstwa jest czescia modulu *Inteligencji Biznesowej* (ang. Business Intelligence). Oprócz samej hurtowni danych na modul Inteligencji Biznesowej skladaja sie aplikacje analityczne i raportujace, oraz *tematyczne hurtownie danych* (ang. data marts). Aplikacje analityczne komunikuja sie z hurtownia danych i umozliwiaja analitykom i decydentom wglad w statystyki, podsumowania, informacje demograficzne, prognozy, itp. Informacje uzyskane za pomoca aplikacji analitycznych sluza do tworzenia raportów, zestawien i podsumowan. Problem polega na tym, ze aktualna infrastruktura nie umozliwia wykorzystania wiedzy pochodzacej z analizy hurtowni danych do podjecia decyzji taktycznych, czyli takich, które sa podejmowane w aplikacjach operacyjnych. Aplikacje operacyjne stanowia komponenty *modulu Zarzadzania* (ang. Business Management). Komponenty te umozliwiaja przedsiebiorstwu wykonywanie swej misji na podstawie wiarygodnych danych i prognoz. Jakosc uslug swiadczonych przez przedsiebiorstwo, a zatem jakosc procesów wykonywanych za pomoca aplikacji operacyjnych, w duzej mierze zalezy od jakosci danych dostarczanych do aplikacji operacyjnych, a ta z kolei zalezy od stopnia integracji modulu Zarzadzania z modulem Inteligencji Biznesowej. Modul Zarzadzania jest wykorzystywany przez personel przedsiebiorstwa majacy bezposrednia stycznosc z klientem, aplikacje wchodzace w sklad tego modulu musza miec do dyspozycji pelna, zintegrowana, aktualna i wzbogacona wiedze o kliencie. Nalezy zauwazyc, ze informacje dostarczane do aplikacji operacyjnych musza stanowic polaczenie aktualnych i szczegółowych danych tradycyjnie przechowywanych w operacyjnych bazach danych z wynikami analiz (np. danymi demograficznymi) powstalych w hurtowni danych. Architektura umozliwiajaca polaczenie aplikacji operacyjnych z aplikacjami analitycznymi i stosujaca zaawansowana integracje danych zwana jest *korporacyjna fabryka informacji* (ang. Corporate Information Factory, CIF) i stanowi bardzo obiecujaca propozycje, która w przyszlosci bez watpienia bedzie odgrywala istotna role w infrastrukturze informatycznej nowoczesnych przedsiebiorstw.

Operacyjna skladnica danych, podobnie jak hurtownia danych, jest zorientowana tematycznie. Oznacza to, ze dane przechowywane w operacyjnej skladnicy danych przedstawiaja zintegrowany fragment rzeczywistosci w której dziala przedsiebiorstwo i udostepniaja wglad do danego wycinka danych z perspektywy calego przedsiebiorstwa, a nie tylko wybranego dzialu (jak dzieje sie w przypadku zwyklej operacyjnej bazy danych). Dla przykladu, operacyjna skladnica danych zorientowana na klienta moze zawierac, poza szczegółowymi danymi o kliencie, dane o wszystkich ostatnich interakcjach klienta z przedsiebiorstwem (listy zakupionych produktów, stan rozliczen z klientem, informacje o kontaktach telefonicznych klienta z dzialem obslugi, itp.). Dane przechowywane w operacyjnej skladnicy danych sa tez zintegrowane i stanowia spójny katalog danych. Kontynuujac przyklad, operacyjna skladnica danych zawiera wszystkie informacje o danym kliencie zebrane w calym przedsiebiorstwie. Dzieki temu spójny katalog danych o kliencie jest wykorzystywany we wszystkich punktach stycznosci klienta z przedsiebiorstwem i moze byc wykorzystywany przez caly personel który wchodzi w interakcje z danym klientem. W przeciwienstwie do hurtowni danych operacyjna skladnica danych nie zawiera horyzontu czasowego. Dane przechowywane w OSD sa zawsze najswiezsze, podobnie jak w tradycyjnej operacyjnej bazie danych. Oczywiscie, nic nie stoi na przeszkodzie aby w ramach spójnego katalogu integrujacego dane o konkretnym kliencie zawrzec równiez pewna ilosc danych historycznych (np. poprzedni adres, poprzedni telefon kontaktowy, transakcje z ostatniego tygodnia) jesli istnieja aplikacje operacyjne wykorzystujace takie dane. Generalnie jednak rzecz ujmujac, operacyjna skladnica danych jest ahistoryczna. Poniewaz dane prezentowane w operacyjnej skladnicy danych sa nieustannie aktualizowane, sa one równiez ulotne (to kolejna istotna róznica miedzy OSD a hurtownia danych). Wszystkie aktualizacje dokonywane w operacyjnych bazach danych musza byc jak najszybciej propagowane do OSD, aby zapewnic wszystkim uzytkownikom OSD aktualny widok danych. Wreszcie OSD zawiera dane szczegółowe, bez jakichkolwiek wstepnie wyliczonych agregatów i podsumowan. W przeciwienstwie do hurtowni danych, szczegółowe dane przechowywane w OSD maja charakter dynamiczny i czesto ulegaja modyfikacjom.

**Odswiezane operacyjnej skladnicy danych**

Jedna z najwazniejszych róznic wystepujacych miedzy operacyjna skladnica danych a tradycyjna hurtownia danych jest czestotliwosc odswiezania. Tradycyjna hurtownia danych, odswiezana za pomoca metod ETL (Extract-Transform-Load) lub ELT (Extract-Load-Transform), jest najczesciej odswiezana okresowo, w momentach bezczynnosci. W zaleznosci od domeny zastosowania hurtownie danych sa odswiezane w cyklach

dziennych, tygodniowych lub miesiecznych. Operacyjne skladnice danych mozna sklasyfikowac ze wzgledu na czestotliwosc odswiezania w nastepujacy sposób:

- Klasa I: dane wprowadzone do aplikacji operacyjnej musza zostac propagowane do operacyjnej skladnicy danych natychmiast, w przeciagu kilku sekund. Przykladem takich danych moga byc dane nowego klienta, które podaje on za pomoca formularza na stronie internetowej przedsiebiorstwa.
- Klasa II: dane moga zostac wprowadzone do operacyjnej skladnicy danych w sposób asynchroniczny, lecz czas odswiezenia nie powinien wyniesc wiecej niz godzine. Przykladem takich danych sa podsumowania opisujace profil zakupów klienta. Takie dane sa wykorzystywane m.in. do wygenerowania automatycznych rekomendacji dla danego klienta i w takim celu wystarczajacy jest przyblizony profil danego klienta.
- Klasa III: dane nalezace do tej klasy moga byc odswiezane w cyklach dziennych. Dotyczy to w szczególnosci danych pochodzacych z automatycznych systemów odkrywania wiedzy, np. dane o wzorcach zakupów, na podstawie których buduje sie automatyczne rekomendacje, nie zmieniaja sie na tyle czesto, aby konieczne bylo ich nieustanne odswiezanie.
- Klasa IV: do tej klasy naleza dane pochodzace z analiz przeprowadzanych w hurtowni danych. Czestotliwosc odswiezania tych danych zalezy od czestotliwosci odswiezania hurtowni danych. Przykladami takich danych moga byc uaktualnienia statystyk opisujacych grupy klientów lub zmieniajace sie trendy sprzedazy. Fakt, iz dane te sa uaktualniane stosunkowo rzadko nie oznacza bynajmniej, ze dane te sa mniej istotne. Z analizy hurtowni danych moze wynikac np. model opisujacy prawdopodobienstwo tego, ze dany klient zrezygnuje z uslug przedsiebiorstwa na rzecz konkurencji. Znajomosc i jakosc tego modelu jest kluczowa w pracy personelu bezposrednio kontaktujacego sie z klientem.

W celu dokladnego zobrazowania idei operacyjnej skladnicy danych rozwazmy nastepujacy przyklad. Klient loguje sie do sklepu internetowego. System identyfikuje klienta i odczytuje wyniki analizy zachowan klienta. Na wynik takiej analizy skladaja sie rózne czynniki, miedzy innymi historia nawigacji po sklepie, ogladane artykuly, dokonane poprzednio zakupy, dane demograficzne klienta, informacje z centrum obslugi nt. pytan zadawanych przez klienta, itp. Powyzsze dane, zintegrowane i polaczone, pozwalaja przedsiebiorstwu na zbudowanie dokladnego profilu klienta obejmujacego jego zainteresowania, zwyczaje oraz preferowane metody zakupu i zaplaty. Na bazie tych informacji system moze wygenerowac oferte promocyjna, która bedzie idealnie dostosowana do wymagan danego klienta. Istotne jest to, ze caly proces analizy nie moze zachodzic w czasie rzeczywistym, poniewaz system nie ma na to czasu. Od momentu przylaczenia sie klienta do momentu wyswietlenia mu oferty nie moze uplynac wiecej niz 5-6 sekund. Taki czas jest wysoce niewystarczajacy aby przeanalizowac historie nawigacji, demografie klienta, historie jego zakupów, itd. Wszystkie te dane musza juz byc obecne w operacyjnej skladnicy danych i podczas logowania sie klienta odpowiedni profil musi byc tylko zlokalizowany. Na podstawie profilu system powinien maksymalnie szybko zaproponowac oferowany produkt promocyjny. Przeciwienstwem takiego podejscia jest metoda czesto wykorzystywana przez róznych sprzedawców i zwana pogardliwie „spray and pray", a polegajaca na zarzuceniu wszystkich klientów licznym zbiorem takich samych ofert i oczekiwaniu, ze nikly procent klientów zainteresuje sie niektórymi produktami.

## Aktywna hurtownia danych

*Aktywna hurtownia danych* (ang. active data warehouse) to srodowisko zapewniajace zintegrowane skladowanie spójnych danych na potrzeby zarówno strategicznego wspomagania decyzji jak i taktycznych aplikacji operacyjnych. Innymi slowy, aktywna hurtownia danych to technologia umozliwiajaca polaczenie w ramach jednego produktu zarówno tradycyjnej hurtowni danych, jak i operacyjnej skladnicy danych. Z racji swego dualnego charakteru aktywna hurtownia danych posiada pewne interesujace wlasnosci, które zostaly podsumowane ponizej.

Aktywna hurtownia danych nie moze ograniczac sie do efektywnego wsparcia jednego modelu przetwarzania. Poniewaz na obciazenie aktywnej hurtowni danych sklada sie mieszanka zlozonych zapytan analitycznych (model OLAP) i krótkich, szybkich transakcji pochodzacych z aplikacji operacyjnych (model OLTP), architektura aktywnej hurtowni danych musi spelniac zupelnie nowe, czasem przeciwstawne wymagania dotyczace efektywnosci przetwarzania, skalowalnosci i niezawodnosci. Przykladem sprzecznych wymagan jest wymaganie dotyczace szybkosci odpowiedzi. W danym momencie w systemie aktywnej hurtowni danych moze znajdowac sie kosztowne i czasochlonne zapytanie wyliczajace wspólczynniki demograficzne dla jakiejs grupy klientów. W tym samym czasie z aplikacji operacyjnej do systemu moze zostac wprowadzone zapytanie wyszukujace najlepsza oferte dla konkretnego klienta. Poniewaz drugie zapytanie ma wyzszy priorytet (prawdopodobnie osoba wprowadzajaca to zapytanie wlasnie rozmawia z klientem), pierwsze zapytanie powinno zostac wstrzymane lub przelaczone na nizszy priorytet w celu zagwarantowania jak najszybszej odpowiedzi na drugie zapytanie. Z drugiej strony, mocne obciazenie ze strony aplikacji operacyjnych moze oznaczac calkowite zablokowanie analitycznych zdolnosci aktywnej hurtowni danych.

Zazwyczaj rozmiar hurtowni danych jest wiekszy niz rozmiar operacyjnej bazy danych. Poniewaz aktywna hurtownia danych laczy ze soba dane tradycyjnie przechowywane w hurtowni danych z danymi operacyjnymi, skalowalnosc takiego rozwiazania staje sie istotnym problemem. Architektura aktywnej hurtowni danych musi zapewniac efektywne wykonywanie wspólbieznych zapytan dzialajacych na bardzo duzych wolumenach danych. Jednak najistotniejsza róznica miedzy hurtownia danych a systemem operacyjnym jest zagadnienie dostepnosci. Tradycyjne hurtownie danych nie musza byc dostepne 24×7×365. Aktywna hurtownia danych musi równiez wspierac dzialalnosc operacyjna przedsiebiorstwa. Ze wzgledu na gwaltowny wzrost internetowego kanalu dzialalnosci przedsiebiorstw (przede wszystkim chodzi o sprzedaz, ale nie tylko), infrastruktura informatyczna lezaca u podstaw funkcjonowania aplikacji operacyjnych musi byc dostepna zawsze. Oznacza to, ze aktywna hurtownia danych nie moze miec zadnych przestojów, nie posiada okresów bezczynnosci przeznaczanych tradycyjnie na odswiezanie, oraz musi byc w 100% niezawodna.

Innym czynnikiem, który istotnie rózni aktywna hurtownie danych od tradycyjnej hurtowni danych jest kwestia swiezosci danych. Aktywna hurtownia danych musi dzialac na najnowszych danych, co powoduje, ze mechanizmy dostarczania danych do hurtowni danych musza ulec zmianie. Tradycyjny model asynchronicznej propagacji zmian do hurtowni danych nie zdaje egzaminu. W idealnym przypadku proces odswiezania jest ciagly i zapewnia calkowita synchronizacje danych w aktywnej hurtowni danych. Jak wspomniano wczesniej, wymagany stopien swiezosci danych zalezy od konkretnej aplikacji. Nalezy jednak przyjac, ze aktywna hurtownia danych wymaga w ogólnosci danych o bardzo wysokim stopniu swiezosci [3].

**Zyski wynikajace z aktywnej hurtowni danych**

Implementacja aktywnej hurtowni danych moze zaowocowac wieloma istotnymi zyskami. Architektura aktywnej hurtowni danych charakteryzuje sie skalowalnoscia, dostepnoscia, oraz efektywnoscia. Obecnosc aktywnej hurtowni danych eliminuje nadmiarowosc danych, poniewaz aktywna hurtownia danych dostarcza jednego spójnego obrazu wszystkich danych jakimi dysponuje przedsiebiorstwo. Takie zjawisko nazywa sie *jedyna wersja prawdy* (ang. single version of truth) i ma niebagatelne znaczenie praktyczne. Dzieki brakowi redundancji danych przedsiebiorstwo unika podejmowania sprzecznych decyzji dotyczacych tego samego klienta, nie zmusza klienta do wielokrotnego podawania tych samych danych, unika nieprzyjemnych sytuacji i nieporozumien wynikajacych z nadmiarowosci danych. Istnienie jednej wersji prawdy w znaczacy sposób zwieksza tez zaufanie miedzy klientem i przedsiebiorstwem.

W tradycyjnej hurtowni danych wyniki analiz sa dostepne tylko analitykom, menedzerom i decydentom. Historia hurtowni danych wskazuje jasno, ze wyposazenie decydentów w narzedzia analityczne znaczaco poprawia jakosc podejmowanych przez nich decyzji. Aktywna hurtownia danych pozwala na propagacje tego zjawiska w dól hierarchii przedsiebiorstwa, do szeregowych pracowników. Oczywiscie, zamiast udostepniac analizy i opracowania globalne, uzyskuja oni istotnie wzbogacone dane dotyczace aktualnie obslugiwanego klienta. Podobnie jak w przypadku decydentów wyzszego szczebla, pracownicy nizszych szczebli sa w stanie podjac lepsze decyzje jesli sa w posiadaniu lepszych danych.

Aktywna hurtownia danych minimalizuje równiez czas opóznienia miedzy podjeciem decyzji strategicznej i taktycznej. Wartosc podjetej akcji szybko maleje wraz z czasem. Jesli przedsiebiorstwo jest w stanie zareagowac na jakies zdarzenie natychmiast po jego wystapieniu, wartosc takiej reakcji jest duzo wyzsza niz wartosc pózniejszej reakcji. Przykladowo, zaoferowanie klientowi towaru po promocyjnej cenie w momencie, gdy klient znajduje sie przy kasie sklepowej ma duzo wieksza wartosc (i nizsza cene) niz zaoferowanie tego samego produktu w katalogu wyslanym do klienta. Wynika stad, ze oprócz szybkosci wprowadzenia danych do aktywnej hurtowni danych nie mniej istotna jest szybkosc uzyskania sprzezenia zwrotnego, czyli wsparcie decyzji taktycznej która odbywa sie w punkcie stycznosci z klientem. Aktywna hurtownia danych znaczaco zmniejsza opóznienie dostarczenia decyzji taktycznej, poniewaz hurtownia danych bedaca zródlem decyzji i aplikacja operacyjna bedaca konsumentem decyzji współdziela to samo srodowisko.

Do istotnych zalet aktywnej hurtowni danych zaliczyc równiez nalezy automatyzacje procesów zwiazanych z dzialaniem aplikacji operacyjnych. Poniewaz w srodowisku aktywnej hurtowni danych aplikacje operacyjne dysponuja danymi duzo wyzszej jakosci niz ma to miejsce w tradycyjnym srodowisku, czesc procesów moze zostac w pelni zautomatyzowana. Dzieki temu aplikacje operacyjne moga zostac wyposazone w znacznie wieksza ilosc „inteligencji", bardziej polegajac na danych dostarczonych przez aktywna hurtownie danych niz na arbitralnych decyzjach uzytkownika.

**Od czego zalezy sukces?**

Implementacja aktywnej hurtowni danych jest trudnym i wymagajacym zadaniem. Mozna pokusic sie o wskazanie istotnych czynników, które decyduja o sukcesie tego przedsiewziecia. Pierwszym czynnikiem, który nalezy wziac pod uwage, jest zapewnienie wsparcia ze strony kadry zarzadzajacej. Nalezy zdawac sobie sprawe

z tego, ze implementacja hurtowni danych oznacza daleko idace zmiany w polityce funkcjonowania przedsiebiorstwa. Takie zmiany moga napotkac na opór, w szczególnosci ze strony szeregowych pracowników, których taka zmiana najbardziej dotyczy. Stad zyskiwanie zaufania kierownictwa i przyszlych uzytkowników ma kluczowe znaczenie dla powodzenia calego projektu. Drugim czynnikiem jest edukowanie uzytkowników i promowanie otwartej polityki informacyjnej. W ten sposób unika sie sytuacji, w której dzial IT staje sie waskim gardlem ograniczajacym uzytkownikom dostep do informacji zawartej w hurtowni danych. Kolejnym czynnikiem jest uproszczenie architektury. Aktywna hurtownia danych zazwyczaj jest implementowana w przedsiebiorstwach posiadajacych dluga historie wdrozen róznych produktów informatycznych. Aby aktywna hurtownia danych nie okazala sie kolejnym, nieudanym wdrozeniem, wazne jest podejscie totalne: usuniecie nalecialosci i starych systemów, jesli ich funkcjonalnosc moze byc w latwy sposób zapewniona przez aktywna hurtownie danych. Doswiadczenie wskazuje, ze ten krok równiez czesto napotyka na znaczny opór uzytkowników, którzy zycza sobie, aby nowo wdrazany produkt w niczym nie zmienial procedur, których sie nauczyli i do których zdazyli sie przyzwyczaic. Ostatni czynnik nie jest charakterystyczny tylko dla aktywnych hurtowni danych, ale odnosi sie do kazdego nowego produktu informatycznego. Ze wzgledu na bardzo dynamiczny charakter dziedziny jaka jest informatyka wybór stabilnej platformy i wiarygodnego dostawcy oprogramowania nabiera szczególnego znaczenia w przypadku duzego i kosztownego wdrozenia, a do takich bez watpienia nalezy aktywacja hurtowni danych przedsiebiorstwa.

**Wyzwania technologiczne**

Poniewaz, jak juz wczesniej wielokrotnie wspominano, aktywne hurtownie danych to pomysl nowy, warto poswiecic nieco uwagi analizie wyzwan natury technologicznej przed którymi staja przedsiebiorstwa chcace zaimplementowac to rozwiazanie. Pierwsze wyzwanie jest zwiazane z rozmiarem aktywnej hurtowni danych. Wspólczesne hurtownie danych z latwoscia osiagaja rozmiary rzedu terabajta danych. Zgodnie z najnowszym rankingiem opublikowanym przez Winter Corp. najwieksza funkcjonujaca hurtownia danych nalezy do Land Registry, angielskiej agencji rzadowej zarzadzajacej nieruchomosciami i gruntami w Anglii i Walii i posiada rozmiar 18.3 terabajta. Kolejne miejsca w rankingu zajmuja BT Group (11.7 terabajta) i United Parcel Service (9.0 terabajta). Takie rozmiary swiadcza nie tylko o ilosci danych skladowanych w hurtowni danych, ale równiez daja pojecie o stopniu skomplikowania schematu takiej hurtowni. Nierzadko na schemat skladaja sie setki i tysiace tabel, tysiace wiezów referencyjnych i indeksów, ogromna liczba perspektyw i migawek. Efektywne zarzadzanie tak ogromna baza danych stanowi wielkie wyzwanie i powazny problem. Drugi czynnik technologiczny zwiazany jest z dziedzina zastosowan aktywnych hurtowni danych. Twórcy hurtowi musza byc przygotowani na to, ze gwaltowny wzrost liczby uzytkowników pociaga za soba lawinowy wzrost liczby wymagan stawianych aktywnej hurtowni danych. Uzytkownicy, w tym przypadku bardziej zróznicowani niz uzytkownicy tradycyjnej hurtowni danych, formuluja rózne, czasami sprzeczne, wymagania odnosnie funkcjonalnosci hurtowni danych. Zaspokojenie wszystkich oczekiwan uzytkowników moze nie okazac sie zadaniem trywialnym. Wreszcie trzecie wyzwanie zwiazane jest z przeniesieniem ciezaru wykorzystania aktywnej hurtowni danych z biur analityków (back-office) do biurek pracowników bezposrednio obslugujacych klientów (front-office). Ta zmiana pociaga za soba koniecznosc zapewnienia wsparcia informatycznego duzo wiekszej liczbie uzytkowników którzy sa skupieni na biezacej dzialalnosci przedsiebiorstwa. Wsparcie musi tez byc udzielane szybciej, poniewaz od szybkosci i jakosci dzialania uzytkowników operacyjnych zalezy zadowolenie klienta.

## Przyklady aktywnej hurtowni danych

Aby lepiej zobrazowac pojecie aktywnej hurtowni danych i wyraznie zaznaczyc, czym ta architektura rózni sie od tradycyjnie stosowanych rozwiazan, ponizej zamieszczono analize dwóch przykladów wdrozenia aktywnej hurtowni danych, które zakonczyly sie pelnym sukcesem.

**3M**

3M to firma o ponad stuletniej historii. Dzis warta 16 miliardów dolarów, spólka ma swoje przedstawicielstwa w ponad 60 krajach i obsluguje klientów z 200 krajów, sprzedajac produkty z katalogu liczacego ponad pól miliona pozycji. W latach 90-tych 3M posiadala 40 oddzielnych departamentów i 60 narodowych oddzialów. Kazda jednostka samodzielnie zbierala informacje o klientach, produktach, dostawcach, oraz sprzedazy. Istnialy pewne aplikacje o charakterze analitycznym, ale pozwalaly one na niewielka integracje informacji z kilku polaczonych jednostek. Nie istnial zaden globalny widok danych przedsiebiorstwa, a kazda jednostka obslugiwala swoich

klientów tak, jak gdyby byla niezalezna firma. Na to nakladaly sie niespójnosc i niekompletnosc danych, skutecznie uniemozliwiajace utworzenie globalnej perspektywy przedsiebiorstwa.

W 1996 firma zainicjowala ambitny projekt, którego efektem bylo powstanie aktywnej hurtowni danych. Celem projektu bylo zbudowanie spójnego globalnego obrazu kazdego produktu, klienta i partnera handlowego. Rewolucyjna zmiana dotknela takze strukture firmy, na miejsce 10 departamentów odpowiadajacych technologiom oferowanym przez 3M powolano 6 departamentów odpowiadajacych docelowym grupom klientów. Cala informacja o kazdym indywidualnym kliencie zostala scentralizowana i wprowadzona do *globalnej hurtowni danych przedsiebiorstwa* (ang. global enterprise data warehouse). Utworzona aktywna hurtownia danych zostala otwarta w 1997 roku i od tego czasu stala sie podstawowym narzedziem dzialania korporacji. Z technologicznego punktu widzenia najtrudniejszym zadaniem podczas konstrukcji aktywnej hurtowni danych okazala sie integracja ogromnych wolumenów danych z setek niezaleznych systemów i dostarczenie wyników do geograficznie rozproszonej grupy uzytkowników liczacej tysiace uzytkowników.

Dzis aktywna hurtownia danych umozliwia powstawanie nowych aplikacji, zarówno operacyjnych jak i analitycznych. Aplikacje zarzadzajace planowaniem zapotrzebowania i zamówien, aplikacje logistyczne, magazynowe, systemy sprzedazy, narzedzia analizujace rentownosc, narzedzia do przewidywania trendów, udostepnianie zasobów klientom i kontrahentom, kazdy rodzaj aplikacji moze powstac na bazie aktywnej hurtowni danych, poniewaz przedstawia ona soba jedyna i aktualna wersje prawdy. Utworzone aplikacje znacznie zwiekszyly produktywnosc i efektywnosc procesów biznesowych, uproscily wspólprace z kontrahentami, umozliwily stworzenie zaawansowanego interfejsu webowego dla klientów. Posród bezposrednich korzysci przedstawiciele 3M wymieniaja zwiekszenie sprzedazy wiazanych oraz budowanie aktywnych zwiazków z klientami, których zaufanie do firmy znaczaco wzroslo. Sama tylko konsolidacja rozproszonych dzialów IT zaowocowala zmniejszeniem nadmiarowych kosztów, zlikwidowaniem konkurencyjnych projektów, oraz przyspieszeniem dzialania. Ocenia sie, ze w przypadku dzialów IT calej korporacji oszczednosci wyniosly 100 milionów dolarów. Imponujaco wyglada tez liczba opisujaca oszczednosci 3M wynikajace z automatyzacji logistyki, szacuje sie ja na miliard dolarów w przeciagu calego zycia projektu. Wszystkie operacje sprzedazy sa przetwarzane przez jeden centralny system, który zapewnia kazdemu z krajowych oddzialów firmy wlasny, indywidualny widok danych, troszczac sie o konieczne transformacje (np. przeliczanie walut). Nawet takie zadania jak drukowanie etykiet na paczki wysylane do klientów poczta sa obslugiwane przez aktywna hurtownie danych. Dane przechowywane w aktywnej hurtowni danych sa na najwyzszym poziomie szczególowosci. Odswiezanie danych odbywa sie w niektórych podsystemach w sposób ciagly (tak jest w przypadku obslugi zamówien), lecz niektóre podsystemy stosuja odswiezanie wsadowe z czestotliwoscia co dwie godziny (zamówienia kierowane do fabryk).

**Ford**

Ford Motor Company to druga co do wielkosci samochodowa firma swiata. Liczy 350 000 pracowników a sprzedaz siega 160 milionów dolarów rocznie. Bezposrednio z firma kontaktuje sie 2000 dostawców, 6000 przedstawicieli, firma zarzadza tez 8 centrami dystrybucji w których pracuje 4000 pracowników. Centra dystrybucji zajmuja sie zaopatrzeniem 50 milionów samochodów w 200 000 róznych rodzajów czesci. Zarzadzaniem dostawami czesci zajmuje sie aplikacja IMAS (Inventory Management and Alerting System). Aplikacja ta byla poczatkowo pomyslana jako proste narzedzie raportujace, którego zadaniem bylo informowanie uzytkowników o ciezarówkach przewozacych czesci, które z jakiegos powodu byly opóznione. Ewolucja wymagan sprawila, ze aplikacja ta zaczela coraz bardziej zmierzac w kierunku aktywnej hurtowni danych. Dzis IMAS to operacyjna skladnica danych posiadajaca rozlegle mozliwosci operacyjne i analityczne. System co noc mierzy poziom zaopatrzenia kazdego centrum dystrybucji w czesci i sledzi geograficzne polozenie kazdego samochodu dostawczego. Analiza tych danych pozwala na codzienne monitorowanie i modyfikowanie przeplywu czesci i dynamiczna zmiane priorytetów i punktów docelowych poszczególnych samochodów dostawczych. W ten sposób ciezarówka opuszczajaca fabryke w Detroit z najnizszym priorytetem po dotarciu do Seattle moze posiadac juz pierwszenstwo rozladunku, jezeli wiezione przez nia czesci sa aktualnie potrzebne. IMAS pozwala nie tylko na monitorowanie samochodów dostawczych, ale równiez wspomaga prace magazynierów w centrach zaopatrzenia. Kazda czesc jest skanowana i system natychmiast podpowiada pracownikowi, gdzie dana czesc powinna sie znalezc. Nalezy zauwazyc, ze jest to ekstremalny przyklad rozszerzenia grona uzytkowników aktywnej hurtowni danych, poniewaz w tym przypadku uzytkownikami sa pracownicy fizyczni zajmujacy sie zaopatrzeniem magazynu w czesci. Z technologicznego punktu widzenia dwa wyzwania wydaja sie najbardziej interesujace: zapewnienie natychmiastowej odpowiedzi na zapytanie skierowane do aktywnej hurtowni danych (natychmiast po zeskanowaniu kodu kreskowego czesci system musi odpowiedziec, w którym magazynie i na której pólce polozyc dana czesc) oraz zmiana czestotliwosci odswiezania hurtowni danych z cyklu miesiecznego na dzienny.

## Integracja aplikacji przesiebiorstwa

Zaawansowana architektura aktywnej hurtowni danych wymaga daleko idacej integracji aktualnie istniejacych systemów analitycznych z systemami operacyjnymi. *Integracja aplikacji przedsiebiorstwa* (ang. Enterprise Application Integration, EAI) zapewnia poprawna i efektywna propagacje danych miedzy tymi systemami i dostarcza platforme, na której integracja heterogenicznych systemów moze odbywac sie mozliwie najmniejszym kosztem. Integracja aplikacji przedsiebiorstwa oznacza zapewnienie nieograniczonej wymiany informacji pomiedzy wszystkimi komponentami skladajacymi sie na infrastrukture informatyczna przedsiebiorstwa [5]. Taka integracja jest bardzo trudnym zadaniem, wymaga bowiem stworzenia infrastruktury technicznej, która umozliwi plynne polaczenie procesów biznesowych, platform programowych i sprzetowych, czy standardów i architektur informatycznych. W tym artykule koncentrujemy sie na roli, jaka gra integracja aplikacji przedsiebiorstwa w stworzeniu aktywnej hurtowni danych.

Aktywna hurtownia danych wymaga ciaglego odswiezania strumieniem nowych danych pochodzacych z systemów operacyjnych. Proces odswiezania nie jest trywialny i wymaga efektywnych narzedzi informatycznych. Jednym z takich narzedzi jest wlasnie integracja aplikacji przedsiebiorstwa. Poprawnie zaimplementowana, umozliwia na odswiezanie zawartosci aktywnej hurtowni danych w czasie rzeczywistym lub zblizonym do rzeczywistego. Z drugiej strony, integracja aplikacji przedsiebiorstwa umozliwia takze polaczenie w druga strone, tzn. dostarczanie wyników analiz przeprowadzonych w komponencie analitycznym do aplikacji operacyjnych. Taka infrastruktura zapewnia sprzezenie zwrotne w procesie podejmowania decyzji: decyzje sa podejmowane na podstawie integralnych i aktualnych danych, podjeta decyzja jest natychmiast transmitowana do aplikacji operacyjnej i na jej podstawie podejmowana jest bezzwlocznie akcja. Komunikacja miedzy zródlami danych operacyjnych, aplikacjami analitycznymi i aplikacjami operacyjnymi odbywa sie za pomoca specjalnej *magistrali komunikatów* (ang. EAI message bus).

Skuteczna implementacja integracji aplikacji przedsiebiorstwa wymaga spelnienia kilku warunków. Dane transmitowane magistrala pomiedzy poszczególnymi komponentami maja czesto charakter poufny, system przesylania wiadomosci musi wiec charakteryzowac sie bezpieczenstwem, niezawodnoscia, wydajnoscia i skalowalnoscia. Komunikujace sie ze soba moduly musza byc oczywiscie polaczone siecia komputerowa. Poniewaz przedsiebiorstwa implementujace takie rozwiazanie to najczesciej duze, rozproszone geograficznie organizmy, integracja aplikacji przedsiebiorstwa musi brac pod uwage mozliwosci i zagrozenia zwiazane z wykorzystaniem sieci rozleglych do komunikacji miedzy poszczególnymi komponentami architektury. Magistrala przesylajaca komunikaty musi posiadac zdolnosc do zarzadzania przeplywem komunikatów, co pociaga za soba umiejetnosc konwersji komunikatów. Integrowane moduly moga byc wysoce heterogeniczne, a koszty ich modyfikacji wysokie, zatem sposób wlaczania danego modulu do magistrali wymiany komunikatów musi wiazac sie z mozliwie najmniejszymi kosztami. W idealnym przypadku wlaczenie danego modulu do magistrali wymiany komunikatów nie powinno sie wiazac z zadna ingerencja w kod aplikacji.

Technologie umozliwiajace zbudowanie architektury integrujacej aplikacje przedsiebiorstwa mozna podzielic na nastepujace klasy [2]. W ogólnosci kazda kolejna klasa stanowi bardziej zaawansowana implementacje oferujaca zwiekszone mozliwosci.

- Integracja na poziomie danych (data-level integration): w tym przypadku integracja miedzy aplikacjami osiagana jest poprzez współdzielenie danych. Dane sa konwertowane na format odpowiedni dla kazdej aplikacji i ladowane za pomoca tradycyjnych technik ETL i ELT. Taka formula integracji nie pociaga za soba praktycznie zadnych kosztów zwiazanych z przystosowaniem istniejacych aplikacji. Istotne jest, aby system zapewnial okna bezczynnosci umozliwiajace przeslanie danych miedzy poszczególnymi komponentami. Wada rozwiazania jest to, ze dane w zadnym z integrowanych systemów nigdy nie sa w pelni aktualne.
- Integracja na poziomie komunikatów (message-level integration): takie rozwiazanie umozliwia współdzielenie danych w czasie rzeczywistym. Aplikacje komunikuja sie ze soba wykorzystujac oprogramowanie do przesylania wiadomosci (IBM MQSeries, Microsoft MSMQ), które umozliwia zarówno komunikacje typu point-to-point, jak i subskrypcje kolejek wiadomosci. Wada tego rozwiazania jest to, ze komunikujace sie komponenty musza byc swiadome istnienia magistrali wiadomosci, tzn. kazdy komponent musi implementowac interfejsy do wysylania i odbierania komunikatów. To z kolei moze pociagac za soba koniecznosc ingerencji w kod istniejacych juz aplikacji.
- Integracja na poziomie procesów (process-level integration): jest to rozszerzenie wczesniejszego rozwiazania o dodanie zdolnosci do *zarzadzania przeplywem pracy* (ang. workflow management) i przeplywem procesów miedzy komunikujacymi sie komponentami. Rozwiazanie to, mimo ze wymagajace najwiekszych inwestycji, zapewnia tez najbardziej zaawansowana i daleko idaca integracje aplikacji przedsiebiorstwa.

Stworzenie infrasturktury umozliwiajacej integracje aplikacji przedsiebiorstwa jest pierwszym krokiem do stworzenia aktywnej hurtowni danych. Oczywiscie, sama tylko mozliwosc współdzielenia danych miedzy systemami dzialajacymi w przedsiebiorstwie to jeszcze nie wszystko. Konieczne jest zapewnienie sprzezenia zwrotnego miedzy aplikacjami analitycznymi i operacyjnymi, wzbogacenie aplikacji operacyjnych o mozliwosc

korzystania z decyzji generowanych w aplikacjach analitycznych, i wiele innych. Niemniej jednak, integracja aplikacji przedsiebiorstwa stanowi warunek konieczny stworzenia aktywnej hurtowni danych.

## Podsumowanie

W niniejszym artykule przestawiono pojecie operacyjnej skladnicy danych. Wskazano cechy odrózniajace ja od zarówno tradycyjnej hurtowni danych, jak i operacyjnej bazy danych. Omówione zostaly metody odswiezania operacyjnej skladnicy danych. Operacyjna skladnica danych jest podstawowym komponentem aktywnej hurtowni danych. Pojecie aktywnej hurtowni danych zostalo przedstawione szczególowo, wraz z prezentacja korzysci wynikajacych z zaimplementowania tego narzedzia informatycznego. Kolejno przedstawiono czynniki wplywajace na udana implementacje oraz podkreslono wyzwania technologiczne wynikajace z wdrozenia aktywnej hurtowni danych. Dwa przyklady wdrozen aktywnej hurtowni danych mialy na celu przekonanie czytelnika o niebagatelnych zyskach plynacych z implementacji aktywnej hurtowni danych. Artykul konczy dyskusja nad metodami umozliwienia integracji aplikacji przedsiebiorstwa, która to integracja jest niezbedna do stworzenia srodowiska pozwalajacego na wdrozenie aktywnej hurtowni danych.

## Literatura


[1] H. Garcia-Molina, J.D. Ullman, J. Widom, Database System Implementation, Prentice Hall, 2000
[2] B.Gold-Bernstein, EAI Market Segmentation, EAI Journal, Jul/Aug 1999
[3] R.Hackathorn, Value Proposition for the Active Data Warehousing: Achieving the Intelligent Enterprise, Feb 2002
[4] M. Jarke, M. Lenzerini, Y. Vassiliou, P. Vassiliadis, Fundamentals of Data Warehouses, Springer-Verlag Berlin Heidelberg, 2000
[5] D.Linthicum, Enterprise Application Integration, Addison-Wesley, 1999