Folosirea data warehouse în asistarea activitatii de retail

partea I

Conf.dr. ing. Mariana MOCANU
Catedra de Calculatoare, Universitatea "Politehnica" Bucuresti

The paper presents some basic aspects regarding the use of data warehouse in the retail activity. The main steps in designing a data warehouse for a retail system are presented.
Keywords: *Data warehouse, data cubes, information access, retail, multidimensional modeling, star schema*.

Caracteristicile de baza ale unui data warehouse

Având în vedere ca cerinta de baza pentru orice sistem informatic este aceea de a asigura informarea corecta si în timp util a tuturor factorilor de decizie dintr-o companie, activitatea de realizare a sistemelor informatice este acceptata de majoritatea specialistilor prin prisma utilizarii tehnicii de calcul pe doua planuri si anume în activitatea de fundamentare a deciziilor si în asistarea activitatii curente desfasurate în firma.

Primul plan se caracterizeaza prin existenta unor produse software specializate în analiza alternativelor decizionale, urmarindu-se gasirea celei mai favorabile decizii pentru societatea comerciala, în momentul adoptarii lor. Cel de-al doilea plan, urmareste dezvoltarea de aplicatii informatice de gestiune economica la nivel de activitate sau grup de activitati în cadrul societatii comerciale.

Pentru obtinerea performantelor la nivelul proceselor de afaceri asistate, se impune ca sistemele sa controleze mediul prin primirea datelor, prelucrarea lor si returnarea rezultatelor suficient de rapid pentru a fi în masura sa influenteze functionarea proceselor în timp real. Prelucrarea în timp real cere introducerea imediata în sistem a datelor, a mesajelor transmise de la terminalele sursa. Un numar mare de statii, aflate la distanta si legate în sistem prin intermediul unor echipamente de comunicatie de mare viteza pot lucra simultan: unele actualizeaza fisiere, altele participa la interogari etc.

Sistemele informatice în timp real se bazeaza pe solutii tehnologice avansate: procesoare puternice; memorii de masa cu acces direct; legatura directa între calculatorul electronic si sistemul de comunicatie; terminale diferite adaptate nevoilor; perfectionari ale echipamentelor si canalelor de comunicatie; alocarea spatiilor de memorare si prelucrari preliminarii în retea; realizarea unor tehnici evoluate de programare asigurând producerea software-ului necesar etc..

Acumularea unor mari cantitati de date, necesare functionarii activitatii si luarii corecte a deciziilor nu mai este o problema din punct de vedere tehnic. Acum se pune problema extragerii din multitudinea informatiilor existente a celor mai relevante pentru problema data, într-un timp scurt si cu un cost cât mai mic. Asa s-a nascut ideea structurarii informatiei în depozite de date (Data Warehouse – DW). Depozitele de date reprezinta o cerinta acuta a organizatiilor moderne (fie ele întreprinderi, banci, administratie etc.) si, totodata, o realitate tehnologica pusa în practica din ce în ce mai frecvent. Un sondaj realizat de META Group în 1998 arata ca 90% dintre managerii intervievati intentionau lansarea unor proiecte de implementare a acestui concept. Segmentul de piata legat de depozitele de date are o rata anuala de crestere de cca. 35%.

William Inmon este considerat parintele recontestat a notiunii în interesul ei curent (Inmon detine de altfel trademark-ul termenului Data Warehouse). Viziunea sa despre depozitele de date se concentreaza asupra rolului acestora de baza informationala a deciziei manageriale, pastrând astfel un nivel înalt de generalitate. Un alt nume important este cel al lui Earl Hadden, cel care a enuntat, a fundamentat si a experimentat cu succes o

metodologie riguroasa pentru implementarea rapida a depozitelor de date (90 day winners). O serie de firme comerciale si-au adus la rândul lor contributia la clarificarea, dezvoltarea si popularizarea noii tehnologii. Printre acestea se remarca Software AG, Oracle, Red Brick Systems, Prism Solutions, MicroStrategy, SAP etc.

Din perspectiva economica, globalizarea comertului, ascutirea dramatica a concurentei, scurtarea spectaculoasa a ciclurilor de viata a produselor datorita dinamicii tehnologiei, impunerea unor cerinte calitative extrem de ridicate precum si alte asemenea evolutii, au evidentiat si mai mult valoarea strategica a informatiei. Manipularea operativa a informatiei a impus la rândul ei noi modele manageriale, mai suple si mai eficiente. Nevoia de a raspunde în timp optim cerintelor pietei a condus la descentralizare si la reducerea numarului nivelurilor decizionale, consacrând asa-numitele "ierarhii plate" care se bazeaza pe delegarea puterii decizionale operative catre esalonul managerial secund. Practic, clasicul functionar este pe cale de a fi înlocuit de "lucratorul cu informatii". Atât la nivelul conducerii strategice cât si la nivelul managementului operativ, nevoia de informatie pura, corecta si semnificativa a devenit vitala.

Din perspectiva tehnologica, ultimii ani au adus puterea de calcul la preturi accesibile. Servere paralele bazate pe microprocesoare ieftine rivalizeaza ca putere cu supercalculatoarele, la o fractiune din pretul acestora. Sistemele de baze de date pot exploata la maximum arhitecturile hardware paralele, iar evolutiile spre sisteme deschise permit o conectivitate aproape totala la orice fel de surse de date si interoperabilitate între diverse platforme software/hardware. Mediile de stocare magnetice si optice admit volume de ordinul giga si chiar tera. PC-urile au ajuns si ele la maturitate. Puterea lor este acum suficienta pentru functiile de analiza si prezentare care le sunt rezervate, iar interfetele grafice intuitive le fac accesibile utilizatorilor neprofesionisti, în speta managerilor.

Definitia lui Bill Inmon este extrem de concisa: depozit de date este o colectie de date tematica, integrata, plasata într-un context temporal si permanenta, destinata fundamentarii deciziei manageriale. (*A data warehouse is a subject-oriented, integrated, time-variant and nonvolatile collection of data în support of management's decision making process*.)

Datele din data warehouse provin în principal din datele capturate din sistemul operational, dar mai pot proveni din datele de arhiva (în perioada de constituire a depozitului) precum si din surse externe, cum ar fi baze de date publice. Câteva exemple posibile: *date demografice* (obtinute în urma unui recensamânt), *date statistice* (furnizate de institute specializate), *date de prognoza economica* (furnizate de institutii orientate pe studiul pietei), *date obtinute în urma unor sondaje de opinie* etc. Aceste date pot fi cumparate, pot fi preluate pe baza de abonament sau pot fi date publice gratuite. Toate aceste date au o caracteristica comuna, aceea ca sunt destinate fundamentarii deciziei manageriale. Spre deosebire de colectiile de date utilizate de sistemul operational - orientate spre optimizarea si siguranta procesarii datelor - datele dintr-un depozit de date sunt organizate într-o maniera care sa permita analizarea lor, deci extragerea semnificatiei economice pe care o poarta. Rolul unui DW este de a oferi o imagine coerenta asupra datelor relative la activitatea unei organizatii si a contextului în care acesta actioneaza. Utilizarea acestei colectii poate consta din extragerea unor rapoarte (la cerere sau pe baza unui "abonament" cu o anumita periodicitate), extragerea unor date pentru a fi utilizate de aplicatiile de birotica (programe de calcul tabelar, procesoare de text, programe de prezentare etc.), dar mai ales pentru a fi utilizate de catre aplicatii specializate de analiza. Acestea ar putea fi împartite în doua categorii: instrumente de analiza on-line (*OLAP* - On Line Analytical Processing - aplicatii axate pe analiza multidimensionala) si instrumente pentru "minerit" în date (*data mining* - aplicatii axate pe descoperirea unor sabloane semnificative în colectii de date).

Sistemul operational al unei organizatii tinde mereu sa reflecte realitatea curenta. Astfel, el se afla într-o continua evolutie iar datele pe

care le contine sunt relevante doar pentru momentul în care sunt accesate. Orizontul de timp pe care îl acopera este de regula de 60 pâna la 90 de zile, deoarece dupa acest interval tranzactiile efectuate sunt arhivate, fiind considerate deja de domeniul istoriei, deci neinteresante din perspectiva operativa. Pentru nevoile analizei economice, dimpotriva, informatiile cu caracter istoric sunt esentiale, deoarece ele pun în evidenta tendinte care reprezinta fundamentul unei prognoze corecte. DW se constituie într-un istoric al sistemului operational, constituit dintr-o serie de "instantanee", imagini la diverse momente în timp. Orizontul de timp pe care îl acopera DW este de cel putin cinci ani, ajungând uneori la zece ani, în functie de dinamica evolutiei pietei si, deci, de relevanta datelor cu caracter istoric pentru nevoile analizei.

Din punct de vedere tehnic, acesta implica faptul ca orice înregistrare din DW corespunde unui moment de timp specificat. Orice cheie de acces la informatiile din DW va cuprinde si o componenta temporala.

Esenta aplicatiilor operationale este actualizarea continua a colectiilor de date, actualizare realizata în general pe baza tranzactionala. Orice tranzactie procesata implica inserarea unor noi înregistrari, modificarea sau eventual stergerea altora etc. Cu totul altfel stau lucrurile în cazul DW, unde o astfel de dinamica lipseste. Practic, aici are loc adaugarea periodica a unor date extrase din sistemele operative. Din punctul de vedere al aplicatiilor care folosesc DW, accesul la date este doar pentru citire.

Din punctul de vedere al proiectarii, aceasta diferenta este extrem de importanta. În sistemul operational, o tranzactie trebuie sa duca colectia de date dintr-o stare consistenta într-o alta stare consistenta, iar aceasta implica mecanisme extrem de complexe de mentinere a integritatii datelor, mai ales în situatia sistemelor intens concurentiale: mecanisme de jurnalizare, mecanisme de salvare/restaurare/refacere, mecanisme de detectare a blocarilor circulare (dead lock) etc. În cazul depozitelor de date aceste mecanisme sunt inutile, astfel ca gradul de libertate câstigat poate fi utilizat pentru optimizarea accesului la date prin denormalizare, sumarizare, statistici ale accesarii datelor si reorganizare dinamica a indexarii etc.

Datele operationale sunt orientate pe aplicatii, în sensul ca organizarea lor este optimizata pentru a servi procesului tranzactional, dinamicii sistemului. În contrast, DW este orientat pe subiectele importante ale procesului economic, cum ar fi: clienti, furnizori, produse, activitati.

Un exemplu simplu poate fi edificator: o comanda lansata de un client va fi consemnata de sistemul operational printr-un set de înregistrari care vor contine informatii despre client, informatii despre produsele sau serviciile comandate, informatii despre modul de transport si modul de plata etc. Atentia sistemului tranzactional este orientata catre consistenta cheilor, astfel încât operatia sa pastreze consistenta. Multe dintre datele esentiale din perspectiva operationala (numarul comenzii, pozitiile liniilor în cadrul comenzii etc.) sunt complet lipsite de relevanta din perspectiva informationala.

O consecinta importanta a acestei orientari este *redundanta datelor*. Daca în sistemul operational redundanta este eliminata (prin procesul de normalizare) pentru a evita anomaliile de actualizare, în DW redundanta este creata în mod intentionat (prin denormalizare si sumarizare) pentru a permite un acces tematic mai facil.

Cel mai important aspect al DW si, în cele din urma, ratiunea pentru care acesta este creat îl constituie asigurarea datelor pentru a raspunde nevoilor informationale ale întregii organizatii, asigurând faptul ca rapoartele generate pentru diverse compartimente vor genera aceleasi rezultate. Sistemul operational este de cele mai multe ori format din subsisteme semi-independente, create la momente diferite, de echipe diferite, în maniere diferite, rezultând un amalgam care, desi functional, este imposibil de folosit pentru analiza. Integrarea datelor provenind din sistemul operational si din alte surse se refera la numeroase aspecte:

- conventii unice privind denumirile datelor - în sistemul operational acestea difera de la aplicatie la aplicatie;

- modalitati unice de codificare - este suficient sa ne gândim la nenumaratele variante de a codifica sexul: ('m', 'f'), (0, 1), (true, false) etc.
- sistem de unitati de masura unic;
- sistem stabil de reprezentare fizica a datelor - în aplicatiile tranzactionale este posibil ca aceleasi date sa fie memorate în moduri diverse;
- conventii clare privind modul de reprezentare a datelor calendaristice, a timpului etc.;

Analizând redundanta informatiilor între sistemul operational si cel informational trebuie remarcat faptul ca din punct de vedere functional cele doua sisteme sunt disjuncte. Sistemul operational proceseaza tranzactii în timp ce sistemul informational este exploatat prin interogari. Cerintele sunt diametral opuse. Orice administrator de baze de date cunoaste faptul ca optimizarile vizând siguranta si coerenta datelor, esentiale într-un sistem tranzactional, conduc inevitabil la încetinirea dramatica a interogarilor, cu deosebire a celor ad-hoc, bazate pe criterii neprevazute (acestea sunt cele specifice analizei economice). Reciproc, aceste interogari - implicând de regula volume mari de date si fiind adesea lipsite de suportul unor indexuri prestabilite - pot compromite performantele operatiilor tranzactionale pâna sub limitele acceptabile.

În ceea ce priveste datele propriu-zise, filtrarea datelor la transferul din sistemul operational în cel informational face ca doar datele relevante pentru analiza economica sa treaca acest prag.

Si orizontul temporal al celor doua sisteme este diferit. Exista o suprapunere foarte mica între cele doua sisteme. DW contine si date sumarizate, care nu exista niciodata în sistemele operationale.

La preluarea în DW, datele sunt supuse unor transformari radicale atât din punct de vedere fizic cât si logic. Ca urmare, redundanta datelor între cele doua sisteme are de regula o rata mai mica de 1%. Chiar daca aceasta rata ar fi mult mai mare, valoarea DW este imensa, deoarece ofera factorilor decizionali din organizatie o imagine unica, coerenta si semnificativa asupra datelor relevante din perspectiva analizei economice. Mai mult, instrumente specializate, OLAP, permit utilizatorilor sa exploreze efectiv aceasta baza informationala, fara a avea nevoie de intermedierea unui serviciu specializat.

Gradul de detaliere a datelor de pe diferitele niveluri de comprimare este descris cu ajutorul termenului de *granularitate*. O data cu cresterea comprimarii datelor are loc scaderea granularitatii, ceea ce se reflecta asupra necesarului de spatiu de stocare, a vitezei de prelucrare obtinute precum si asupra flexibilitatii Data Warehouse.

În cadrul procesului de partitionare sau fragmentare a bazei de date din Data Warehouse, se creeaza o serie de partitii mai mici, independente din punct de vedere fizic si care contin baze de date neredundante. Pornind de la aceste baze de date mai mici, procesele de restructurare, reorganizare si siguranta a datelor se realizeaza mult mai usor.

Întregul traseu parcurs de date din sistemul operational în DW este descris de *procesele de migrare*. Procesul de migrare descrie una din cele mai complexe activitati în rândul functiilor de baza ale DW. Datele provenind din sursele interne ale întreprinderii precum si cele provenind din sisteme externe sunt preluate în DW, de regula prin intermediul unor programe de transformare. Datele interne provin aproape exclusiv din sistemele operative. Deoarece în cadrul unei întreprinderi sunt utilizate mai multe sisteme, structurile de date sunt în mare parte diferite, trebuind sa fie supuse unor procese de transformare, pentru a ajunge la un format unitar.

Datele externe pot proveni, de exemplu, de la terte societati comerciale, fiind nevoie si în acest caz de utilizarea programelor de transformare. Migrarea datelor se poate realiza prin:

- *Reîmprospatare*. Într-un anumit interval de timp baza de date din DW este înlocuita cu datele sursa. În cadrul acestui proces nu are loc nici o transformare. Sub aspect tehnic este foarte usor de realizat, DW necontinând însa o istorie a datelor.
- *Actualizare*. Într-un anumit interval de timp baza de date din DW este actualizata corespunzator modificarilor aparute de la ultima actualizare a datelor sursa. Din punct de ve-

dere tehnic procesul este destul de greu realizabil deoarece este dificil de urmarit ce date s-au modificat între timp.
- *Propagare*. Modificari în datele sistemelor sursa se oglindesc automat în baza de date a DW. Tranzactiile din sistemul baza se propaga asadar în DW. Acest proces se poate desfasura sincron sau asincron, tehnic fiind destul de greu realizabil.
Modelul datelor din DW (sub aspect semantic si tehnic), provenienta datelor, informatii despre procesul de migrare, procesul de comprimare, inclusiv despre nivelurile de comprimare a datelor, modelele pentru evaluari si analize precum si informatiile referitoare la aspectele tehnice în cadrul DW sunt stocate într-un sistem de metadate.

Modalitati de acces la informatie
Utilizatorii pot accesa datele continute în DW prin intermediul unor front-end tools si anume:
- *Data query si reporting tools* (ex.: high-volume batch jobs, subject-oriented views);
- *Application development tools* (pentru dezvoltarea de aplicatii proprii);
- *Executive information system tools (EIS);*
- *Data mining tools*. Scopul *data mining* îl constituie filtrarea datelor continute în DW, în vederea pregatirii acestora pentru evaluarea si analiza ulterioara. Cel mai simplu tip de filtre îl reprezinta asa numitele "praguri", care pot fi rigide, variabile sau adaptive. Pragurile rigide genereaza valori fixe, în cazul nerespectarii lor (la obtinerea unei valori mai mari sau mai mici), fiind automat generat un mesaj de atentionare. Pragurile variabile sunt cele a caror valoare depinde de context (de exemplu factorii de sezon). Pragurile adaptive sunt orientate dupa modul de distribuire a deviatiilor valorice astfel ca, de exemplu, o deviatie a cifrei de afaceri cu 10% la nivel de articol trebuie altfel interpretata decât o deviatie cu acelasi procent la nivel de grupe de articoles
- *On-line analytical processing tools (OLAP)*. Baza conceptului OLAP si a analizei de date multidimensionale formeaza "Datacube". Descrierea unei multimi de dimensiuni si a unei multimi de masuri (measures) alcatuiesc împreuna un Datacube. Conceptul OLAP ofera de asemenea posibilitatea de utilizare a "Virtual Cube". Similar cu perspectivele (views) în modelul relational, un cub virtual poate fi alcatuit prin definirea de legaturi între Datacube-uri reale. Pentru utilizator nu exista nici un fel de deosebire la accesul de date din Datacube. Virtual cube nu contine date materializate, fiind creat prin interogari asupra Datacube-urilor.
În cadrul analizelor intermediate cu ajutorul tehnologiei OLAP se deosebesc urmatoarele operatii:
- *Roll-up:* agregarea dimensiunilor pentru comprimarea datelor;
- *Drill-down:* miscare opusa în vederea obtinerii de detalii;
- *Slicing*: Crearea de submultimi prin reducerea dimensiunilor;
- *Dicing*: Crearea de submultimi la nivel global (Sub-cube);
- *Pivot*: Reorganizarea datelor rezultate prin schimbul de dimensiuni.

Bibliografie
1. Alex Berson; Stephen J. Smith: Data Warehousing, Data Minig and Olap; McGraw-Hill, 1998
2. Ion Cozac : Arhitectura bazelor de date, 2002, articol Internet
3. Mariana Mocanu: Data Warehouse Konzepte; UPB, note de curs
4. Mariana Mocanu: Ingineria programarii; UPB, note de curs
5. Simona Pavlov: Data Warehouse pentru sisteme informatice de suport a activitatii de retail; lucrare de dizertatie, ian. 2003, UPB/CPRU
6. *** SAP Business Information Warehouse: A.Seemann, B.Schmalzridt, P. Lehmann; Galileo Press GmbH, Bonn 2001, 1.Auflage
7. *** Multi-Dimensional Modeling with BW, ASAP for BW Accelerator, Business Information Warehouse, Document version 2.0
8. *** EWIS – Business Warehouse, Betriebskonzept, Copyright Ewis, Offenburg 2002