MAGAZIILE DE DATE – SUPORT INFORMATIC PENTRU LUAREA DECIZIILOR

Dr.hab., conf. Univ. V. Cotelea
Academia de Studii Economice din Moldova

1. DESCRIERE GENERALĂ

Odată cu apariţia calculatoarelor, organizaţiile utilizează datele recuperate din sistemele informatice pentru a-şi satisface necesităţile informaţionale. Unele software-uri oferă accesul direct la datele conţinute în sistemele operaţionale. Altele extrag datele din bazele de date pentru a le combina în diverse forme nestructurate, ţinând seama de necesităţile în informaţii ale utilizatorilor. Ambele metode au evoluat pe parcursul timpului şi acum organizaţiile gestionează datele necurăţate şi inconsistente, în baza cărora, în majoritatea cazurilor, se iau decizii importante.

Este recunoscut faptul că, pentru o creştere a eficienţei întreprinderii, este nevoie de cea mai bună utilizare a resurselor informatice care există în interiorul şi exteriorul ei. Cu toate că sistemele informatice au început să fie implementate cu mulţi ani în urmă, încă nu se poate spune că există o utilizare eficientă a acestora.

Motivul principal îl constituie modul în care au evoluat calculatoarele, tehnologiile informaţionale şi sistemele. Majoritatea organizaţiilor tind să obţină informaţii calitative, dar atingerea acestui obiectiv depinde de arhitectura sistemelor, atât a hardware-ului, cât şi software-ului.

Sistemele care susţin procesele de gestiune şi luare a deciziilor sunt esenţial diferite de sistemele convenţionale, de prelucrare a tranzacţiilor sau care se mai numesc sisteme operaţionale, adică, aplicaţiile tipice, într-o întreprindere, pot fi clasificate în două mulţimi mari:

- aplicaţiile de business, aplicaţiile ce susţin businessul de fiecare zi, care garantează funcţionarea întreprinderii;
- aplicaţiile despre business, aplicaţiile ce analizează businessul, care ajută la interpretarea a ceea ce se va întâmpla şi la luarea deciziilor asupra strategiilor de viitor.

O arhitectură de date adecvată, care ar suporta ambele tipuri de aplicaţii, se bazează pe două medii de baze de date: bazele de date operaţionale (pentru a susţine aplicaţiile de business) şi bazele de date pentru asistarea deciziilor (pentru a susţine aplicaţiile despre business).

Înmagazinările de date (*data warehouses*), actualmente sunt, centrul de atenţie al întreprinderilor şi instituţiilor mari, deoarece ele asigură un mediu în care sunt mai bine utilizate datele gestionate de diverse aplicaţii operaţionale.

O înmagazinare de date este o colecţie de date, în care datele organizaţiei sunt integrate şi care se utilizează în calitate de suport pentru procesul de luare a deciziilor manageriale. Deşi diverse organizaţii şi persoane fizice au reuşit să înţeleagă tehnicile unei înmagazinări de date, experienţa a demonstrat că există încă multe dificultăţi în această direcţie.

Evident că reunirea elementelor de date corespunzătoare, din diverse izvoare de aplicare într-un mediu integral centralizat, simplifică problema de acces la date şi, în consecinţă, accelerează procesul de analiză şi consultare.

Aplicaţiile pentru susţinerea deciziilor bazate pe o înmagazinare de date pot face mai practică şi mai uşoară exploatarea datelor şi o mai înaltă eficienţă a afacerilor, care nu se obţine când se utilizează numai datele provenite din aplicaţiile operaţionale, în care datele se obţin, realizând procese independente şi, de multe ori, mai complexe.

Magaziile de date se creează pe datele extrase dintr-una sau mai multe baze de date orientate spre aplicaţiile operaţionale. Datele extrase sunt transformate (cu scopul de a fi eliminate inconsistenţele) şi generalizate, dacă este nevoie, şi apoi încărcate în înmagazinarea de date. Procesele de transformare, creare a variantelor în timp, generalizare, agregare şi combinare a datelor extrase, asigură crearea unui mediu pentru accesul la datele organizaţiei. Această tehnică nouă dă persoanelor fizice posibilitatea de acces la toate nivelele întreprinderii, de a lua decizii cu o mai mare siguranţă şi responsabilitate.

Noile tehnologii informaţionale, în cadrul unui mediu de înmagazinare, oferă organizaţiei posibilitatea de folosire optimală a datelor în calitate de ingredient-cheie pentru un proces eficient de luare a deciziilor. Organizaţiile pot scoate

avantaje din resursele lor de informații în realizarea operațiilor de afaceri, dar, pentru aceasta, trebuie să fie considerate strategiile tehnologice speciale de implementare a unei arhitecturi complete a unei magazii de date [1].

2. CONCEPTUL „*MAGAZIE DE DATE*”

O magazie de date este, în general, o bază de date dedicată. Ea înglobează majoritatea datelor din sistemul operațional al întreprinderii şi este separată de operațiile acestuia. Înmagazinarea de date facilitează utilizarea aplicațiilor, organizarea şi păstrarea datelor necesare pentru prelucrarea analitică, prelucrarea datelor pe o perspectivă mai largă de timp.

O magazie de date se poate caracteriza prin modul în care datele de afaceri păstrate în înmagazinarea de date diferă de datele operaționale utilizate în aplicațiile de producție (figura 1).

Baza de date operaționale	Magazia de date
Date operaționale	Date de afaceri pentru informare
Orientate pe aplicație	Orientate pe subiect
Actuale	Actuale + istorice
Detaliate	Detaliate + agregate
Schimbare continuă	Stabile

Figura 1. Deosebirea tipurilor de date.

Evident că datele, într-o magazie de date, vin, în majoritatea cazurilor, din mediul operațional. Înmagazinarea de date este întotdeauna o stocare de date transformate şi separate fizic de datele mediului operațional.

William H. Inmon, pionier în elaborarea magaziilor de date, emite următoarea definiție a magaziei de date [2]:

Definiția 1. ***O magazie de date*** este o colecție de date orientată spre subiect, integrată, variabilă în timp şi nevolatilă, care are drept obiectiv susținerea procesului de luare a deciziilor.

Această definiție are două supoziții implicite: un data warehouse este separat fizic de sistemul operațional şi susține date agregate şi tranzacții de date (atomice), care sunt separate de baza de date utilizată în regim OLTP.

Astfel, o înmagazinare de date reprezintă o bază de date ce conține date extrase din mediul de producție al întreprinderii, care sunt selectate şi epurate, optimizate pentru procesul de consultare, şi nu pentru procesul de tranzacții. Magaziile de date presupun consolidarea datelor din mai multe surse, fie cele depozitate în bazele de date relaționale, fie datele provenite din tabelele electronice, documente textuale etc.

În acord cu alt specialist în domeniul înmagazinărilor de date, Richard Hacathom, obiectivul unei magazii de date constă în “formarea unei imagini unice a realității de business”:

Definiția 2.…sistemele data warehouse cuprind o mulțime de programe ce extrag datele din mediul operațional al întreprinderii, o bază de date care le păstrează şi sisteme ce furnizează aceste date utilizatorilor săi.

Există numeroase definiții ale magaziilor de date, primele concentrându-se asupra caracteristicilor datelor păstrate în magazia de date. Definițiile alternative extind domeniul de valabilitate al definiției înmagazinării datelor, pentru a include prelucrarea asociată accesării datelor, de la resursele inițiale până la livrarea acestora către organele de decizie.

Oricare ar fi definiția, scopul suprem al înmagazinării datelor rezidă în integrarea datelor generale din întreaga întreprindere într-o singură magazie, de la care utilizatorii pot lansa interogări, elabora rapoarte şi efectua analize. Magazia de date reprezintă un mediu de susținere a deciziilor, care preia datele stocate în diverse surse operaționale, le organizează şi le face disponibile organelor de decizie din cadrul întreprinderii. Pe scurt, o magazie de date constituie o tehnologie de administrare şi analiză a datelor.

Precum s-a afirmat, principalele caracteristici ale unei magazii de date sunt:

- orientarea spre subiect (temă);
- datele sunt integrate;
- datele sunt istorice, adică variabile în timp;
- datele sunt nevolatile.

2.1. Orientarea spre temă

Prima caracteristică a magaziei de date constă în faptul că datele ei se clasifică, ținând cont de aspectele ce prezintă interes pentru întreprindere, adică, datele sunt organizate pe teme.

Mediul operațional este proiectat în jurul aplicațiilor şi funcțiilor, ca prestări, economii şi depozite într-o instituție bancară. De exemplu, o aplicație de intrare a ordinelor poate accesa date despre clienți, articole şi conturi. Baza de date îmbină aceste elemente într-o structură ce satisface necesitățile aplicației.

Mediul unui data warehouse este organizat în jurul unor subiecte precum client, vânzător, articol şi activitate. De exemplu, pentru un fabricant,

acestea pot fi clienţii, articolele, furnizorii sau vânzătorii. Pentru o universitate, pot fi studenţii, cursurile sau profesorii. Pentru un spital, pot fi pacienţii, personalul medical, medicamentele etc. Alinierea în jurul temelor afectează proiectarea şi implementarea datelor găsite în magazia de date.

Aplicaţiile sunt asociate cu proiectarea bazei de date şi a proceselor. Înmagazinarea datelor se focalizează pe modelarea datelor şi proiectarea bazei de date, şi proiectarea proceselor (în forma sa clasică) nu este separată de acest mediu.

Diferenţa dintre orientarea pe procese şi funcţii (ale aplicaţiilor) de orientarea pe teme constă în conţinutul datelor la nivelul detaliat. Într-o magazie de date, se exclud datele ce nu vor fi utilizate de sistemele de susţinere a deciziilor, în timp ce datele orientate pe aplicaţii sunt datele îndreptate spre satisfacerea imediată a cerinţelor funcţionale şi a proceselor, care pot fi utilizate (sau nu) şi pentru susţinerea deciziilor.

Altă deosebire importantă este legătura dintre date. Datele operaţionale menţin o legătură continuă între două sau mai multe relaţii bazate pe o regulă comercială aflată în vigoare. Înmagazinările de date măsoară o perioadă de timp în care sunt multe legături. Multe din regulile comerciale (şi legăturile corespunzătoare de date) se reprezintă într-o magazie de date între două sau mai multe relaţii.

Se poate menţiona că interesul organizării pe teme constă în faptul că devine posibilă realizarea analizelor pe subiecte, trecând prin structurile funcţionale şi organizaţionale ale întreprinderii. Această orientare permite, de asemenea, petrecerea analizelor prin iteraţie, subiect după subiect.

Integrarea într-o structură unică este indisolubilă de evitarea duplicatele datelor ce se referă la mai multe subiecte. Uneori, în practică, există şi **piețe de date** (*datamarts*), adică, magazia de date este fragmentată în mai multe baze, care suportă orientarea spre temă.

2.2. Date integrate

Cea mai importantă caracteristică a mediului de înmagazinare a datelor constă în faptul că datele dintr-o magazie de date sunt integrate întotdeauna.

Integrarea datelor se prezintă prin multe aspecte: convenţii de nume consistente, unităţi de măsură uniformă a variabilelor, codificări de structuri consistente, atribute fizice ale datelor consistente, izvoare etc.

Pe parcursul anilor, proiectanţii diferitelor sisteme luau decizii specifice asupra modului în care trebuie să fie construită o aplicaţie. Stilurile şi proiectele personalizate pot fi reflectate în mai multe moduri. Ele se deosebesc prin codificare, prin structurile cheilor, prin caracteristicile lor fizice, prin convenţiile de denumiri etc. Posibilitatea unui grup de proiectanţi de aplicaţii de a crea aplicaţii inconsistente este fabuloasă. În continuare, sunt prezentate cele mai importante deosebiri ce pot fi în formele în care se proiectează aplicaţiile.

2.2.1. Codificarea

Proiectanţii de aplicaţii pot codifica atributul *Sex* în diverse moduri. Unii proiectanţi reprezintă atributul *Sex* prin valorile "*M*" şi "*F*", alţii prin "*1*" şi "*0*", alţii prin "*X*" şi "*Y*", şi inclusiv prin "*masculin*" şi "*feminin*".

Nu importă cum va sosi atributul *Sex* în magazia de date. Probabil, valorile "*M*" şi "*F*" sunt cele mai potrivite. E important faptul că atunci când atributul *Sex* vine din diverse izvoare, trebuie să sosească în magazie în stare uniform integrată. Prin urmare, când atributul *Sex* este încărcat într-o magazie de date dintr-o aplicaţie şi în magazie se păstrează în formatul "*M*" şi "*F*", datele trebuie convertite în formatul respectiv.

2.2.2. Măsurarea atributelor

Proiectanţii folosesc unităţi de măsură diferite pentru a prezenta, de exemplu, mărimea ţevilor. Unii proiectanţi păstrează datele despre ţevi în centimetri, alţii în inche-uri, alţii în milioane de picioare, cuburi pe secundă, iar alţii în yarzi.

La definirea unităţilor de măsură a atributelor, transformarea respectivă traduce diverse unităţi de măsură utilizate în diferite baze de date într-o măsură standard comună. Astfel, oricare ar fi izvorul când datele despre ţevi sunt trecute într-o magazie de date trebuie să fie măsurate în acelaşi mod, cu aceeaşi unitate de măsură.

2.2.3. Convenţii de denumire

Acelaşi element, în diverse aplicaţii, frecvent, este referit prin diferite nume. Procesul de transformare asigură faptul că se va utiliza numele preferabil utilizatorului. De exemplu, acelaşi concept *balanţă,* într-o aplicaţie, poate fi numit *balanţă activă*, în alta *balanţă curentă*, în a treia *flux în casă* etc.

2.2.4. Izvoare multiple

Acelaşi element poate fi obţinut din mai multe surse. În acest caz, procesul de transformare trebuie să asigure că este utilizat izvorul potrivit, documentat şi motivat pentru a fi depozitat în magazie.

După cum urmează din exemplele aduse, integrarea afectează aproape toate aspectele proiectului - caracteristicile fizice ale datelor, alternativa de a avea mai mult de o sursă de date, problema de standardizare a denumirilor inconsistente, formate de date inconsistente etc.

Oricare ar fi forma de proiectare, rezultatul trebuie să fie acelaşi – datele trebuie să fie păstrate în magazia de date, într-un model global acceptabil şi singular, chiar dacă sistemele operaționale subiacente păstrează datele în mod diferit. Or, când analistul sistemului de asistare a deciziilor lucrează cu magazia de date, el trebuie să se concentreze asupra utilizării datelor ce se găsesc depozitate, dar nu să fie preocupat de confiabilitatea şi consistența datelor.

2.3. Date istorice sau variabile în timp

Datele istorice sunt necesare pentru a urmări în timp evoluția diferitelor valori ale indicatorilor supuşi analizei. Astfel, un timp referențial trebuie să fie asociat cu datele pentru a permite identificarea pe durata valorilor precise.

Datele asociate unui anumit moment sunt solicitate dintr-o magazie de date. Această principială caracteristică a datelor este foarte diferită de cea a datelor din mediul operațional. În mediul operațional, sunt cerute datele asociate momentului de accedere. Cu alte cuvinte, în mediul operațional, când se accede la unitatea de date, se speră că valorile cerute reflectă starea actuală a domeniului de interes.

Deoarece, într-o magazie de date, datele solicitate pot fi asociate oricărui moment de timp (adică, nu "chiar acum"), datele din magazie se numesc ***variabile în timp***.

Datele istorice sunt puțin utilizate în procesarea operațională. Datele din magazie, în schimb, trebuie să includă date istorice pentru a fi utilizate la identificarea şi evaluarea tendințelor.

Datele istorice sau variabile în timp sunt caracterizate în diverse moduri:

- În primul rând, datele sunt păstrate o perioadă îndelungată de timp – de la cinci la zece ani. Perioada de timp reprezentată în mediul operațional este mult mai scurtă – de la valori actuale la şaizeci - nouăzeci de zile.

Aplicațiile care au performanțe bune şi sunt disponibile pentru procesarea tranzacțiilor trebuie să aducă o cantitate minimă de date şi să posede un oarecare grad de flexibilitate. Datorită proiectării rigide a aplicației, aplicațiile operaționale au un termen de viață scurt.

- În al doilea rând, într-o magazie de date, datele sunt variabile în timp datorită structurii cheii. Orice cheie într-o magazie de date conține, explicit sau implicit, un element de timp cum ar fi ziua, săptămâna, luna etc.

Elementul care reprezintă timpul este, aproape întotdeauna, parte componentă a cheii concatenate. Ocazional, timpul poate exista implicit, ca în cazul când fişierul complet este duplicat la sfârşitul lunii sau trimestrial.

- În al treilea rând, datele unei magazii de date, odată înregistrate corect, nu mai pot fi actualizate. Pentru toate scopurile practice, datele unei magazii de date reprezintă o serie largă de viziuni instantanee (snapshots).

Bineînțeles, dacă snapshoturile datelor sunt luate incorect, atunci pot exista schimbări. Dacă snapshoturile sunt făcute adecvat, ele nu sunt modificate niciodată. În unele cazuri, este neetică şi chiar ilegală modificarea snapshoturilor din magazie. În schimb, datele operaționale care trebuie să reprezinte realitatea la momentul accesării, sunt actualizate în acord cu schimbarea domeniului de interes.

2.4. Date nevolatile

Datele, într-o magazie de date, sunt nevolatile, deoarece datele nu sunt reactualizate în timp real, ci sunt reîmprospătate de către sistemele operaționale în intervale regulate de timp. Datele noi sunt adăugate întotdeauna, mai degrabă, ca un supliment al bazei de date, decât ca o înlocuire. Baza de date absoarbe continuu aceste date noi, integrându-le pe rând cu datele anterioare. Cu alte cuvinte, aceeaşi cerere lansată la diferite momente de timp, asupra aceloraşi date întotdeauna trebuie să returneze aceleaşi rezultate.

Astfel, datele într-o magazie sunt utile numai când sunt stabile, adică datele nu sunt suprimate. Perspectiva mai mare, esențială pentru o analiza şi luare a deciziilor, necesită o bază de date stabilă.

Datele din sistemele operaționale se schimbă din moment în moment, adică actualizarea (inserarea, suprimarea şi modificarea) în mediul operațional se face regulamentar, înregistrare cu înregistrare. În schimb, manipularea datelor într-o magazie de date este mult mai simplă. Există două tipuri de operații: încărcarea inițială a datelor şi accesarea acestora. Nu există actualizări de date (în sensul general de actualizare) în magazie.

Există unele consecințe ale acestei deosebiri importante între prelucrarea operațională şi prelucrarea în înmagazinările de date. La nivelul de proiectare în magazii, nu trebuie ținut cont de

anomaliile de actualizare, deoarece nu are loc actualizarea datelor. Aceasta înseamnă că, la nivelul fizic de proiectare, se poate optimiza accesul la date, în particular, cu utilizarea normalizării şi denormalizării fizice.

O altă consecinţă a simplicităţii operaţiei înmagazinării de date constituie tehnologia subiacentă, utilizată pentru a trimite datele în magazia de date. Suportarea actualizării înregistrare cu înregistrare în mod on line (cum este frecvent în cazul prelucrării operaţionale) cere ca tehnologia să posede un fundament mai complex, dar sub o faţadă simplă.

Sursa aproape a tuturor datelor unei magazii de date este mediul operaţional. La o simplă viziune, se poate crede că există o redundanţă masivă de date între datele mediilor. Bineînţeles, prima impresie a multor persoane se centrează spre redundanţa mare de date între mediul operaţional şi mediul unei magazii de date. Această concluzie este superficială şi demonstrează o carenţă în înţelegerea a ceea ce apare într-o bază de date. De fapt, există o redundanţă minimă de date între ambele medii, dacă se ţine cont de următoarele:

- Datele sunt filtrate când trec din mediul operaţional în magazie. Există multe date ce nu vor ieşi din mediul operaţional. Numai datele ce realmente sunt necesare vor intra în magazia de date.
- Termenul de păstrare a datelor este foarte diferit de la un mediu la altul. Datele din mediul operaţional sunt mai recente decât cele din magazia de date. Din perspectivele de timp unic, există o suprapoziţie între mediile operaţional şi magazia de date.
- Magazia de date conţine un rezumat de date, care nu se găseşte în mediul operaţional.
- Datele suportă o transformare serioasă când trec în magazia de date. Cea mai mare parte din date se schimbă semnificativ când sunt selectate şi mutate în magazia de date. Altfel spus, majoritatea datelor se modifică fizic când se mută în magazie. Din punct de vedere al integrării nu sunt aceleaşi date care rezidă în mediul operaţional.

Din perspectiva acestor factori, redundanţa datelor între cele două medii este o stare rară.

3. STRUCTURA UNEI MAGAZII DE DATE

Magaziile de date dispun de o structură distinctă. Există diverse niveluri de sinteză şi detaliere ce delimitează magazia de date. Structura unei magazii de date este prezentată în figura 2.

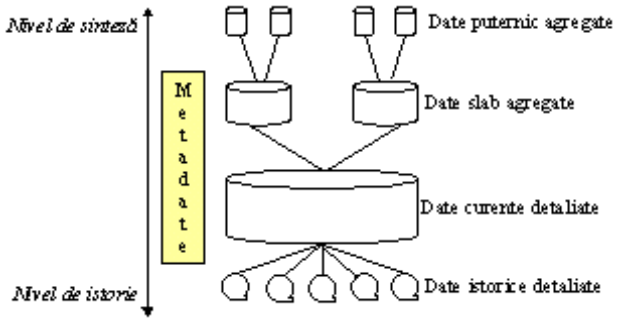

Figura 2. Structura datelor într-o magazie de date

O magazie de date conţine cinci clase de date, organizate conform unei axe de istorie şi unei axe de sinteză:

- datele curente detaliate;
- datele istorice (mai vechi) detaliate;
- datele slab agregate;
- datele puternic agregate;
- metadatele.

O problemă-cheie ce trebuie soluţionată în proiectarea unei magazii de date constă în granularitatea datelor. Datele primitive, precum cele păstrate în sistemele de prelucrare a tranzacţiilor, de obicei, sunt cele mai detaliate.

Agregarea sau sumarea este un proces de intrare a datelor primitive sau detalii de date pentru păstrarea mai compactă, într-o formă utilă pentru a fi analizate sau utilizate de aplicaţii specifice. Agregarea datelor presupune selectarea, filtrarea, combinarea, reorganizarea şi manipularea datelor detaliate sau atomice, datelor pentru producerea categoriilor predeterminate şi specifice, totalurilor şi comparaţiilor.

3.1. Date curente detaliate

Interesul cel mai mare îl prezintă datele curente. Datele curente detaliate reflectă cele mai recente evenimente. Aceste date sunt voluminoase şi pot ocupa multă memorie pentru păstrare, dar reprezintă, în acelaşi timp, cel mai de jos nivel de granularitate. Deseori, ele sunt, pur şi simplu, o replică a bazei de date tranzacţionale curente, care, mai întâi, este curăţată şi apoi stocată. Însă nu toate câmpurile păstrate în sistemele tranzacţionale pot fi mutate în magazie. Trebuie menţionat că, deşi sunt referite drept curente, aceste date sunt actualizate o singură dată, în momentul în care sunt trecute în magazia de date.

Datele curente detaliate aproape întotdeauna se păstrează pe disc, care este uşor de accesat, cu toate că administrarea lor este costisitoare şi complexă.

3.2. Date vechi detaliate

Datele vechi se păstrează într-o formă masivă de stocare. Ele nu sunt frecvent accesate şi sunt păstrate la un nivel de detaliere consistent cu datele curente detaliate. Deşi sunt recuperabile în formă detaliată, timpul de acces este mai mare.

3.3. Date slab agregate

Datele slab agregate provin dintr-un nivel jos de detaliere precum nivelul datelor curente detaliate. Acest nivel al magaziei de date aproape întotdeauna se păstrează pe disc.

Experienţa arată că agregarea datelor în modul care anticipează aplicaţiile îmbunătăţeşte sensibilitatea şi utilizarea magaziei de date. Din punctul de vedere al proiectantului, pentru construirea acestui nivel, sunt necesare două decizii: selectarea atributelor şi selectarea unităţilor de timp pentru agregare.

Ambele probleme implică schimbări în care calculele nu trebuie să fie realizate în mod repetat, dar este necesar mult spaţiu de păstrare. Este evident că atributele şi combinaţiile de atribute ce sunt frecvent utilizate în cereri trebuie să fie agregate, în timp ce cele rar utilizate – nu. Odată ce atributele sunt selectate, prima problemă care survine este determinarea frecvenţei cu care fiecare atribut va fi agregat.

3.4. Date puternic agregate

Următorul nivel de date, întâlnite într-o magazie de date, sunt datele cu un grad înalt de generalizare. Aceste date sunt compacte şi uşor accesibile.

Întotdeauna, unele date, în particular cele solicitate de personalul administrativ superior al unei întreprinderi, trebuie să fie disponibile în formă compactă şi trebuie să fie uşor accesibile. Aceste date, de obicei, includ datele ce sunt consultate în mod repetat.

Acest nivel va dispune de capacitatea de menţinere a datelor agregate o perioadă lungă de timp, conform tendinţelor prestabilite. Odată cu stocarea datelor cu un grad înalt de agregare, timpul de răspuns va fi esenţial redus.

3.5. Metadate

Componentul final al structurii datelor, într-o magazie de date, este cel al metadatelor. Metadatele se situează într-o dimensiune diferită de alte date ale magaziei de date, deoarece conţinutul acestora nu este luat direct din mediul operaţional.

Metadatele joacă un rol foarte important şi sunt utilizate în calitate de:

- directoriu pentru a ajuta analistul în stocarea conţinutului magaziei;
- ghid pentru maparea datelor din mediul operaţional în magazia de date;
- ghid al algoritmilor utilizaţi pentru agregarea datelor curente detaliate în date slab agregate şi agregarea acestora în datele puternic agregate etc.

Metadatele joacă un rol mult mai important într-un mediu data warehouse decât într-un mediu operaţional clasic.

Într-o bază de date relaţională, metadatele sunt o reprezentare a obiectelor definite în baza de date – în special, ca definiţii ale relaţiilor, atributelor, bazei de date, viziunilor şi altor obiecte. Într-o magazie de date, se referă la ceea ce defineşte un obiect data warehouse, aşa ca relaţie, atribut, cerere, raport, regulă de afaceri sau algoritm de transformare.

Înţelegerea acestor definiţii este esenţială pentru toate aspectele procesului de elaborare a unei magazii de date. Gestiunea magaziei de date presupune controlul rigid al tuturor proceselor de la elaborarea programelor, care extrag datele din izvoarele sistemului operaţional, la transformarea unei colecţii de date într-un obiect al magaziei de date. Magazia de date este utilă, numai dacă aduce un avantaj competitiv, adică, dacă datele transformate pentru umplerea sau stocarea datelor pot fi utilizate pentru a răspunde cererilor de afaceri.

Metadatele reprezintă o hartă sau o schemă a acestor date. Structura metadatelor diferă pentru fiecare proces, deoarece scopul acestora este diferit. Aceasta înseamnă că, în cadrul magaziei de date, există mai multe copii ale metadatelor, care descriu acelaşi articol de date. În plus, majoritatea instrumentelor comerciale de administrare a copiilor şi acces la date al utilizatorilor finali utilizează propriile versiuni de metadate. Mai exact, instrumentele de administrare a copiilor utilizează metadatele pentru a înţelege regulile de corespondenţă care trebuie aplicate pentru a transforma datele-sursă într-o formă comună. Instrumentele de acces ale utilizatorilor finali utilizează metadatele pentru a înţelege cum să construiască o interogare. Administrarea metadatelor în cadrul magaziei de date reprezintă o sarcină foarte complexă, care nu trebuie subestimată.

4. ARHITECTURA UNEI MAGAZII DE DATE

Unul din motivele pentru care numărul de magazii de date elaborate creşte rapid constă în faptul că această tehnologie, realmente, este foarte înțeleasă. De fapt, o înmagazinare de date poate reprezenta mai bine structura amplă a unei întreprinderi pentru administrarea datelor din cadrul acesteia. Pentru a înțelege cum interacționează componentele unei

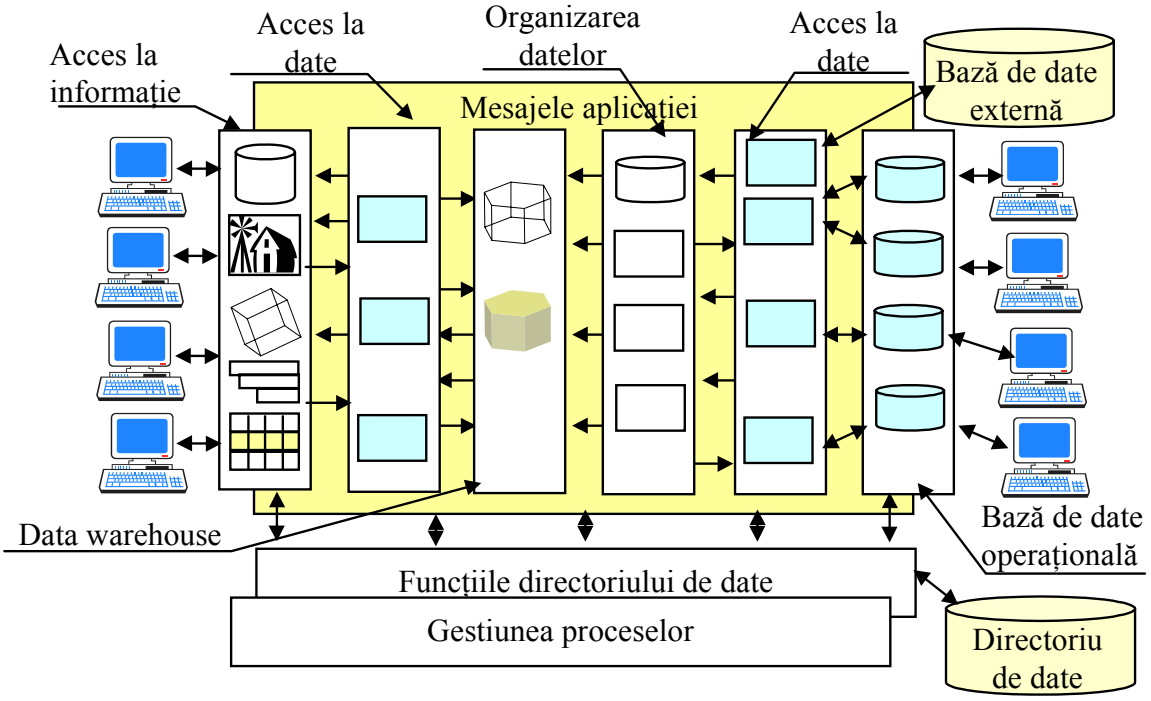

Figura 3. Arhitectura tipică a unui data warehouse.

magazii de date, se trece în revistă arhitectura tipică reprezentată în figura 3.

4.1. Elementele constituente

O arhitectură a unei magazii de date (*Dates Warehouse Architecture*) constituie o formă de reprezentare a structurii de date, comunicații, procese şi prezentări, care există pentru utilizatorii finali ce dispun de un calculator în cadrul întreprinderii.

Arhitectura este constituită dintr-o mulțime de niveluri interconectate:

- baza de date operaționale - nivelul *baza de date externă*;
- nivelul *accesul la informații*;
- nivelul *accesul la date*;
- nivelul *directoriul de date* (metadatele);
- nivelul *gestiunea proceselor*;
- nivelul *mesajele aplicației*;
- nivelul *data warehouse*;
- nivelul *organizarea datelor*.

4.1.1. Baza de date operaționale – nivelul *baza de date externe*

Sistemele operaționale prelucrează datele pentru a susține necesitățile operaționale ale întreprinderii. Pentru aceasta, se creează o bază de date operaționale istorice, care asigură o structură de prelucrare eficientă pentru un număr relativ mic de tranzacții comerciale bine definite.

Dar, din cauza tehnicilor limitate ale sistemelor operaționale, bazele de date proiectate pentru suportarea acestor sisteme au dificultăți la accesarea datelor pentru altă gestiune sau scopuri de procesare.

Această dificultate de accesare a datelor operaționale este amplificată şi de faptul că multe dintre aceste sisteme au o vechime de 10 – 15 ani. Vârsta unor asemenea sisteme face că tehnologiile de acces la datele disponibile pentru a obține date operaționale sunt, de asemenea, vechi.

Este cert că un scop al magaziei de date este eliberarea datelor ce se păstrează într-o bază de date operațională şi combinarea lor cu datele din alt flux de date, în general, extern.

Tot mai mult, organizațiile mari cer date adiționale din baze de date externe. Aceste date includ tendințe demografice, econometrice, achizitive şi competitive. Internetul, numit, de asemenea, "*information superhighway*", oferă accesul la multe surse de date în fiecare zi.

4.1.2. Nivelul *accesul la informații*

Nivelul *accesul la informații* al arhitecturii unei magazii de date este nivelul de care utilizatorul final este legat direct. În particular, el reprezintă instrumentele pe care utilizatorul final, de obicei, le utilizează zi de zi. De exemplu, *Excel*, *Lotus* 1-2-3. *Focus*, *Acces*, SAS etc.

Acest nivel include, de asemenea, hardware-ul şi software-ul implicate în afişarea informațiilor pe ecran şi emiterea rapoartelor pentru imprimare, foilor de calcul, graficelor şi diagramelor pentru analiză şi prezentare. Pe parcursul a două decenii, acest nivel s-a lărgit enorm, în special pentru utilizatorii finali, care au trecut de la PC-urile monoutilizator la PC-urile în rețea.

Actualmente, există instrumente mult mai sofisticate pentru manipularea, analiza şi prezentarea datelor. Cu toate acestea, există probleme ce țin de tratarea şi convertirea datelor, ca să poată fi recolectate, păstrate şi făcute transparente pentru instrumentele şi utilizatorii finali. Una din cheile soluționării acestor probleme ar fi elaborarea unui limbaj de date comun ce ar putea fi folosit prin toată întreprinderea.

4.1.3. Nivelul *accesul la date*

Nivelul *accesul la date* al arhitecturii unei magazii de date este implicat, împreună cu nivelul *accesul la informații,* în conversarea cu nivelul operațional. Pe piața mondială, astăzi, limbajul de date comun este SQL. Inițial, SQL a fost elaborat de IBM ca un limbaj de cereri, dar, în ultimii douăzeci ani, a devenit un standard al schimbului de date.

Unul din progresele-cheie ale ultimilor ani a fost elaborarea unei serii de "filtre" de acces la date, precum EDA/SQL. Ele servesc la accederea aproape la toate sistemele de gestiune ale bazelor de date şi sistemele de fişiere de date. Aceste filtre permit instrumentelor de acces accesarea, de asemenea, a datelor păstrate cu sistemele de gestiune a bazelor de date de o vechime de douăzeci de ani.

Nivelul *accesul la date* nu numai conectează SGBD-uri diferite şi sisteme de fişiere asupra aceluiaşi hardware, dar şi **celor fabricate şi protocoale de** rețea. Una din strategiile-cheie ale înmagazinării de date este asigurarea utilizatorilor finali cu "un acces la datele universale".

Accesul la date universale semnifică că, din punct de vedere tehnic cel puțin, utilizatorii finali nu țin cont de instrumentele de acces la date sau locație. Ei trebuie să fie capabili să acceseze orice sau toate datele întreprinderii care sunt necesare acestora pentru a-şi îndeplini obligațiunile de serviciu.

Nivelul *accesul la date,* atunci, este responsabil de interfața dintre instrumentele de acces la informație şi la bazele de date relaționale. În unele cazuri, aceasta este tot de ce are nevoie utilizatorul final. Cu toate acestea, în general, organizațiile elaborează un plan mult mai sofisticat pentru a susține tehnologiile de înmagazinare a datelor.

4.1.4. Nivelul *directoriul de date (metadatele)*

Pentru a asigura accesul la datele universale, este absolut necesară menținerea unui directoriu de date sau repozitoriu de metadate. Metadatele sunt date despre datele din întreprindere.

Astfel, descrierea unei înregistrări într-un program COBOL sunt metadate. De asemenea, metadate este sentința DIMENSION într-un program Fortran sau sentința de creare în limbajul SQL:

Pentru a avea o magazie de date totalmente funcțională este necesar de a avea o varietate de metadate disponibile, informații asupra viziunilor de date ale utilizatorilor finali şi informații asupra bazelor de date operaționale. Ideal, utilizatorii finali trebuie să acceadă datele din magazia de date (sau din bazele de date operaționale), fără a avea cunoştințe unde datele rezidă sau forma în care sunt păstrate.

4.1.5. Nivelul *gestiunea proceselor*

Nivelul *gestiunea proceselor* este responsabil de programarea diferitelor sarcini ce trebuie realizate pentru construirea şi menținerea magaziei de date şi datelor din directoriul de date. Acest nivel poate depinde de multe procese (proceduri) ce trebuie să existe pentru menținerea magaziei de date.

4.1.6. Nivelul *mesajele aplicației*

Nivelul *mesajele aplicației* ține de transportarea datelor în rețeaua întreprinderii. Mesajul aplicației este referit, de asemenea, ca "subproduct", dar poate implica numai protocoale

de rețea. Poate fi utilizat, de exemplu, pentru a izola aplicațiile operaționale sau strategice, pornind de la formatul de date exact, recolecta tranzacții sau mesaje şi a le furniza unei locații sigure într-un timp sigur.

4.1.7. Nivelul *data warehouse (fizic)*

Data warehouse-ul este nivelul, unde apar datele curente, utilizate, în principal, pentru uz strategic. În unele cazuri, se poate gândi despre magazia de date simplu ca despre o viziune logică sau virtuală a datelor. În multe exemple, magazia de date poate să nu implice stocări de date.

Într-o magazie de date fizică, copiile, în unele cazuri, multe copii de date operaționale sau externe, sunt realmente păstrate într-o formă flexibilă şi uşor accesibilă. Din ce în ce mai mult, magaziile de date sunt construite pe platforme client-server, dar, în general, sunt stocate pe mainframeuri.

4.1.8. Nivelul *organizarea datelor*

Componenta finală a arhitecturii unei magazii de date este *organizarea datelor*. Se numeşte acest nivel, de asemenea, gestiunea copiei sau replicării, dar, de fapt, include toate procesele necesare, cum ar fi selectarea, editarea, rezumarea, combinarea şi introducerea datelor în magazie, dar şi procesele de accesare a datelor din bazele de date operaționale sau externe.

Organizarea datelor implică frecvent o programare complexă, dar tot mai mult, sunt create instrumente de înmagazinare pentru a facilita acest proces. Organizarea include, de asemenea, programe de analiză a calității datelor şi filtre ce identifică modele şi structura datelor din datele operaționale existente.

4.2. Procese şi fluxuri de date în magazia de date

În figura 4 sunt arătate principalele operații şi fluxurile de date ce au loc într-un mediu de înmagazinare a datelor. Precum se vede, înmagazinarea de date se concentrează asupra administrării operațiilor şi a patru fluxuri de date asociate, şi anume fluxul intern, fluxul ascendent, fluxul descendent şi fluxul extern.

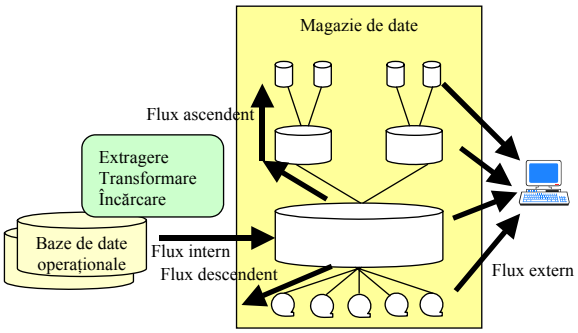

Figura4. Fluxuri de date într-o magazie de date.

4.2.1. Fluxul intern

Datele gestionate de sistemele operaționale reprezintă izvorul principal de date pentru construirea unei magazii de date.

Datele operaționale sunt organizate în fişiere indexate (UFAS,VSAM), în baze de date de tip rețea sau ierarhice (I-D-S/II, IMS, IDMS) sau sub gestiunea sistemelor de baze de date relaționale (DB2, Oracle, Informix etc.)

Fluxul intern se referă la prelucrarea datelor din sistemele-sursă, pentru a le încărca în magazia de date. Întrucât datele-sursă sunt generate mai ales de către sistemele operaționale, adică sistemele OLTP, este necesar ca acestea să fie reconstruite pentru scopurile magaziei de date. Procesele asociate fluxului intern cuprind: extragerea, transformarea şi încărcarea datelor.

Definiția 3. ***Fluxul intern*** reprezintă procesele asociate extragerii, transformării şi încărcării datelor din sistemele-sursă în magazia de date.

Astfel, sunt necesare instrumente de gestiune a datelor şi extragerea lor din baze de date şi/sau fişiere tranzacționale. Atunci când se extrag date din sistemele-sursă, este important să existe garanția că acestea se află într-o stare coerentă. Complexitatea procesului de extragere este determinată de gradul de acordare dintre sistemele-sursă.

Odată ce datele sunt extrase, de regulă, ele sunt încărcate într-un depozit temporar, pentru a fi curățate şi verificate din punctul de vedere al coerenței. Adică datele sunt transformate şi încărcate în magazia de date.

Datele din bazele de date operaționale, proiectate pentru suportarea diferitelor aplicații de producție, deseori, au formate şi structuri diferite. Transformarea sau integrarea datelor presupune luarea datelor din diverse baze de date operaționale şi maparea în date cerute pentru magazie.

Aceleaşi elemente de date, dacă sunt folosite de aplicații diferite sau administrate de diferite SGBD-uri, pot fi definite, utilizând nume de elemente inconsistente şi pot poseda formate

inconsistente şi/sau pot fi codificate în moduri diferite. Toate aceste inconsistenţe trebuie rezolvate înainte ca elementele de date să fie acumulate şi stocate în magazia de date.

Evident că atât procesele fluxului intern, cât şi celorlalte fluxuri necesită crearea metadatelor. Unii cercetători afirmă că chiar există *metafluxul*, adică procesele asociate administrării metadatelor. Metadatele (adică, datele despre date) descriu conţinutul magaziei de date. Metadatele constau din definiţiile elementelor din magazie, sistemele de date-sursă. În calitate de date, metadatele se integrează şi se transformă înainte de a fi stocate.

4.2.2. Fluxul extern

Utilizatorii accesează magazia de date, folosind instrumente ce includ sisteme GUI (*Graphical User Interface*).

Utilizatorilor le pot fi oferite diverse tipuri de instrumente. Acestea pot include software-uri de consultare, generare a rapoartelor, procesare analitică on line, instrumente data/visual mining etc., în funcţie de categoria utilizatorilor şi cerinţele particulare ale acestora. Cu toate acestea, un singur instrument nu poate satisface toate cerinţele, prin urmare, este necesară integrarea unei serii de instrumente.

Definiţia 4. ***Fluxul extern*** constă în procesele asociate punerii la dispoziţie a datelor pentru utilizatorii finali.

Fluxul extern are loc acolo unde valoarea reală a înmagazinării datelor este percepută de către organizaţie. Există două activităţi-cheie implicate în fluxul extern:

- *accesarea*, care se referă la satisfacerea cererilor utilizatorilor finali privind datele de care au nevoie;
- *livrarea*, care e preocupă de livrarea activă a datelor către staţiile de lucru ale utilizatorilor finali.

Magaziile de date, care conţin date agregate, pun la dispoziţie un număr distinct de surse de date, pentru a răspunde unei interogări specifice – inclusiv înseşi datele detaliate şi orice număr de grupuri care satisfac cerinţele informaţionale ale interogării.

În funcţie de aplicaţie, utilizarea unei magazii de date poate fi extinsă prin capacitatea de accesare a datelor externe. De exemplu, datele accesibile on line prin serviciile computerului sau via Internet, pot fi disponibile utilizatorilor magaziei de date.

4.2.3. Fluxul ascendant

Definiţia 5. ***Fluxul ascendent*** reprezintă procesele asociate adăugării unei valori datelor din magazia de date, prin agregarea, împachetarea şi distribuirea datelor.

Activităţile asociate fluxului ascendent cuprind:

- ***agregarea datelor***, prin selecţia, proiectarea, uniunea şi gruparea datelor relaţionale în cadrul unor viziuni, care sunt mai convenabile şi mai utile pentru utilizatorii finali; agregarea se întinde dincolo de operaţiile relaţionale simple, pentru a implica o analiză statistică sofisticată, cuprinzând identificarea tendinţelor, comasarea şi realizarea de eşantioane de date;
- ***împachetarea datelor***, prin transformarea datelor detaliate sau agregate în formate mai utile, cum ar fi foile de calcul tabelar, documentele de tip text, diagramele, reprezentările grafice, bazele de date personale şi animaţia;
- ***distribuirea datelor*** în grupuri adecvate, pentru a mări disponibilitatea şi accesibilitatea acestora.

În timp ce se adaugă valoare datelor, este necesar să se acorde atenţie şi susţinerii cerinţelor privind performanţele magaziei de date, ca şi minimizării continue a costurilor operaţionale. În esenţă aceste cerinţe împing proiectarea în direcţii opuse. Astfel, administratorul bazei de date trebuie să identifice cel mai adecvat proiect, care îndeplineşte toate cerinţele, ceea ce, adeseori, necesită anumite compromisuri.

4.2.4. Fuxul descendent

Definiţia 6. ***Fluxul descendent*** reprezintă procesele asociate arhivării şi copierii de siguranţă a datelor din magazia de date.

Arhivarea datelor vechi, istorice joacă un rol important în menţinerea eficacităţii şi performanţelor magaziei de date, prin transferarea datelor mai vechi, cu valori limitate, într-o arhivă de stocare pe benzi magnetice sau discurile optice.

Fluxul descendent de date include procesele de asigurare că starea curentă a magaziei de date poate fi reconstruită, ca urmare a unei pierderi de date sau defecţiuni software sau hardware. Datele arhivate trebuie stocate într-un mod care să permită restabilirea acestora în magazia de date, atunci când este necesar.

Astfel, fluxul de date din figura 4 este normal şi prezis într-o magazie de date. Datele intră în magazia de date din mediul operaţional (există mici excepţii de la această regulă). La intrare în magazia

de date, datele merg spre nivelul de date curente detaliate. Ele rămân aici şi se utilizează până apare unul din următoarele evenimente:

- sunt eliminate;
- sunt agregate;
- sunt arhivate.

Apoi odată cu procesul de dezactualizare din magazia de date, datele curente detaliate se mută la nivelul *date vechi*, asociate cu timpul. Procesul de agregare utilizează datele curente detaliate pentru a calcula datele în formă slab sau puternic agregate.

Există mici excepţii de la fluxul prezentat. Dar, în general, pentru majoritatea datelor întâlnite într-o magazie, fluxul de date este precum s-a prezentat.

4.2.5. Platforma unei magazii de date

Platforma unei magazii de date este aproape întotdeauna un server de baze de date relaţionale. Când se manipulează volume foarte mari de date, poate fi cerută o configuraţie în bloc de servere UNIX cu multiprocesor simetric (SMP) sau un server cu procesor paralel masiv (MPP) specializat.

Extrasele de date integrate şi transformate sunt încărcate în magazia de date. Una din cele mai populare SGBDR-uri disponibile pentru înmagazinări de date asupra platformei UNIX (SMP şi MPP), de obicei, este Teradata. Alegerea platformei este importantă. Magazia de date va creşte şi trebuie să cuprindă cerinţele peste 3-5 ani.

Multe organizaţii cer sau nu alegerea unei platforme din diverse raţionamente: "Sistemul X este utilizat de compania concurentă" sau "Sistemul Y este deja disponibil în baza sistemului operaţional UNIX, care, deja, este folosit. Una din cele mai mari erori, pe care organizaţiile le comit la selectarea platformei, este că ele presupun că sistemul (hardware-ul sau SGBD-ul) se va dezvolta odată cu datele.

4.2.6. Evoluţia magaziei de date

Construirea unei magazii de date este o sarcină complexă. Nu se recomandă a înţelege elaborarea unui data warehouse pentru o întreprindere ca o construire a unui proiect ordinar. Mai bine, se recomandă ca cerinţele unei serii de faze să fie elaborate şi implementate în modele consecutive, care ar constitui un proces de implementare gradual şi iterativ.

Nu există nicio organizaţie care a realizat elaborarea magaziei sale de date, dintr-un singur pas. Multe întreprinderi, însă, au atins acest scop după o elaborare graduală, pas cu pas. Paşii precedenţi evoluează împreună cu materia ce trebuie adăugată.

Datele dintr-o magazie de date nu sunt volatile şi reprezintă un depozit de date de o singură lectură (în general). Dar pot fi adăugate elemente noi conform unei baze regulare, pentru ca conţinutul să urmeze evoluţia datelor în baza de date-sursă atât în conţinut, cât şi în timp.

Una din problemele menţinerii unei magazii de date constă în găsirea metodelor de identificare a datelor noi sau de modificări în bazele de date operaţionale. Unele căi de identificare includ inserarea componentei data/timp în înregistrările bazei de date şi crearea copiilor de înregistrări actualizate şi a copiilor datelor înregistrărilor tranzacţiilor sau ale bazelor de date zilnice.

Aceste elemente de date noi sau modificate sunt extrase, integrate, transformate şi adăugate la magazia de date la anumite perioade programate. Când se adaugă noi date, datele vechi sunt eliminate. De exemplu, dacă detaliile unui subiect particular se păstrează 5 ani, când se adaugă datele ultima săptămână, prima săptămâna este eliminată.

Bibliografie

1. Gray Paul, Watson Hugh J*. Present and Future Directions in Data Warehousing. The Data Base for Advances in Information Systems. Summer 1998, Vol. 29, Nr 3, pp. 83-90.*

2. Inmon W. H. *The Data Warehouse and Data Mining. Communications of the ACM. 1996, Vol. 39, Nr 11, pp. 49-57.*

Recomandat spre publicare: 14.11.2012