

Decizie asistata de calculator. Concepte, metode si tehnici pentru deciziile centrate pe analiza datelor

Acad. Florin Gheorghe FILIP

Vicepresedinte al Academiei Române, Director stiintific al Institutului National pentru Cercetare-Dezvoltare în Informatica

Analiza datelor constituie o etapa initiala importanta în luarea deciziilor. Ea permite identificarea cauzelor care duc la aparitia unei situatii decizionale. Acest al doilea articol al unei serii dedicate prezentarii conceptelor si solutiilor din domeniul deciziei asistate de calculator face o trecere în revista a unor rezultate recente privind magaziiile de date, prelucrarea analitica on-line si mineritul datelor. Acestea reprezinta tehnologii noi si în plina dezvoltare, care au atras atentia în primul rând a marilor firme producatoare de produse informatice.

Cuvinte cheie: *decizie asistata, analiza datelor, magazii de date, mineritul datelor.*

1. Introducere

Într-un articol anterior (Filip, 2000), au fost prezentate câteva metode de asistare a proceselor decizionale în care pozitia centrala era ocupata de intuitia si judecata umana. Tot în acel articol s-a afirmat ca, acele metode, în care caracterul calitativ predomina, necesita de multe ori efectuarea unei analize a datelor. Scopul acestui articol este de a face o prezentare sistematica a principalelor concepte si rezultate în domeniul asistarii cu mijloace informatice a acelor activitati din procesul decizional care presupun analiza datelor. Este vorba de trei concepte si anume: a) *magaziile (sau depozitele) de date (Data ware-house – DW)*; *prelucrarea analitica – on line (On line analytic processing – OLAP)* si c) *mineritul datelor (Data mining – DM)*, denumit uneori *descoperirea cunos-tintelor (knowledge data discovery – KDD)*.

Ratiunile care au condus la alegerea continutului acestui articol privesc atât continuitatea prezentarii cât si, mai ales, efervescenta domeniului abordat. Într-adevar, dupa cum s-a aratat mai sus, metodele si tehnicile care urmeaza a fi prezentate în acest articol au ca scop facilitarea proceselor decizionale în care intuitia si judecata sunt prevalente si au si au uneori nevoie de un suport informational. În ceea ce priveste dezvoltarea puternica a domeniului, în

special în partea a doua a ultimei decade scurse, se observa un interes major si o competitie acerba din partea marilor producatori de tehnologii informatice. Acestea au constientizat limitele sistemelor actuale de gestiune a bazelor de date în contextul schimbarilor petrecute (si oportunitatilor aparute) în modul de administrare a întreprinderilor si afacerilor, care presupune considerarea cât mai operativa a unei multitudini de surse informationale diverse.

În continuare, articolul este organizat dupa cum urmeaza: Capitolul 2 abordeaza problematica magaziei de date. Magazia de date este un concept afirmat la începutul anilor '90 si este definita prin analogie cu depozitele de bunuri industriale. Ca si acestea, magazia de date este alimentata dintr-o multitudine de baze de date care au meniri diverse în scopul combinarii elementelor intrate în magazie într-un "pachet" integrat, care contine elemente verificate. Mai întâi (în subcapitolul 2.1), se evoca o taxonomie vizionara (propusa cu peste doua decenii în urma) a instrumentelor informatice de asistare a deciziilor, urmata de enumerarea tipurilor de date folosite în administrarea si functionarea unei organizatii si de prezentarea necesitatii noului concept. Subcapitolul 2.2 prezinta caracteristicile definitorii generale ale magaziei de date si descrie câteva tipuri particulare.

Capitolul 3 este dedicat sistemelor de prelucrare analitică on – line (OLAP). OLAP este instrumentul informatic care permite realizarea “pachetului” de date plecând de la surse diferite și livrarea informațiilor în forma solicitată pentru luarea deciziei. Se prezintă mai întâi (în subcapitolul 3.1) definiția dată de Codd, care a inspirat, “miscarea OLAP”, tipurile functionale și „testul FASMI”, care poate califica un produs informatic în categoria OLAP. În continuare, subcapitolul 3.2 conține o descriere a principalelor două direcții tehnologice și anume: a) *bazele de date multidimensionale (Multidimensional data bases – MDB)* și *soluțiile relationale OLAP (Relational/OLAP – ROLAP)* denumite uneori “OLAP-pe-relational”. Ideile OLAP au “radacini” vechi și acestea sînt evocate în subcapitolul 3.3.

Capitolul 4 trece în revista definițiile și tehnicile folosite în mineritul datelor.

Capitolul 5, care încheie acest articol, conține unele comentarii privind domeniul instrumentelor de asistare a deciziilor centrate pe date și recomandă câteva surse importante de documentare (în special pe Internet), care sunt în curs de continuă îmbogățire.

Articolul de față are o orientare mai pronunțată decât cel anterior (Filip, 2000) către tehnologiile informatice. El este recomandat în primul rând analiștilor și facilitatorilor proceselor decizionale.

2. Magazii de date

Magazia (sau depozitul) de date reprezintă un nou concept care vizează asistarea proceselor decizionale în contextul noilor cerințe de administrare a organizațiilor folosind analiza a cât mai multe informații provenind din surse diferite.

2.1. Evoluții în asistarea deciziei centrate pe date

Asistarea deciziei centrate pe date are un caracter normativ relativ scăzut. Caracteristica sa esențială constă în furnizarea la timp a informațiilor cerute de decident în activitățile sale bazate în primul rând pe

intuiție și judecată pentru rezolvarea problemei decizionale.

2.1.1. O viziune timpurie

La sfârșitul anilor '70, Alter (1977, 1980), în încercarea sa de a realiza o clasificare a instrumentelor și sistemelor informatice folosite în acea vreme pentru asistarea deciziilor, ajungea la concluzia că, cel mai util criteriu de clasificare este “gradul în care rezultatele furnizate de produsul informatic pot determina *în mod direct* decizia”, sau, altfel spus, “gradul în care ieșirile sistemului implică o acțiune [efectivă]”. Alter distingea șapte clase, pe care le propunea în ordinea crescătoare a gradului de normativitate. Aceste clase variau de la sistemele de tip “sertare cu dosare” (cele mai “pasive”), la modelele computerizate care sugerează direct decizia (cele mai “normative”) în cazul unor situații decizionale bine structurate. Cele șapte clase se grupau în final în două superclase: a) sistemele informatice “orientate către date” și b) sistemele “orientate către modele”. Instrumentele orientate către date îndeplinesc funcții de regasire a datelor și/sau de analiză a datelor. Principalele clase identificate de Alter (1977) sunt prezentate în continuare:

1. *Sistemele de tip “sertare cu dosare”*, care au ca scop automatizarea unui mod de lucru manual și servesc la accesul on-line la anumite date. Ele sunt destinate persoanelor cu sarcini *operationale* (funcționari, șefi de ateliere) care au nevoie de a avea acces imediat la valoarea curentă a anumitor variabile. De exemplu, un șef de atelier, atunci când dorește să lanseze o anumită lucrare pentru a suferi o prelucrare mecanică, are nevoie să cunoască situația curentă a încărcării mașinilor unelte din atelier. În același mod, patronul unui magazin trebuie să cunoască nivelul stocurilor produselor aflate în depozit sau “la raft” pentru a lansa comenzi de aprovizionare.

2. *Sisteme pentru analiza datelor*, care ajută personalul cu sarcini de sprijin al conducerii în analiza fișierelor cu date curente și istorice cu scopul de a produce rapoarte pentru manageri. “Produsele gene-

ralizate” permit analize cu caracter general folosind datele existente într-o baza de date si chiar formularea si programarea unor modele matematice simple. Ca exemple, se pot enumera rapoartele privind stadiul de realizare a unor comenzi de fabricatie, sau analiza eficientei unor investitii.

3. *Sistemele pentru analiza informatiilor*, care asigura accesul la o serie de baze de date destinate asigurarii informationale a deciziilor si la modele de dimensiuni reduse în scopul furnizarii informatiilor specifice pentru rezolvarea unor situatii decizionale particulare. Dupa cum arata Alter (1977), sistemele din aceasta clasa “sunt proiectate sa extraga datele relevante din aplicatiile de prelucrare electronica a datelor (*Electronic data processing – EDP*) proprii ale întreprinderii si sa le completeze cu date externe”. O caracteristica definitorie a unor astfel de sisteme o constituia faptul ca accesul nu trebuia sa fie stânjenit de considerente privind planificarea functionarii eficiente a unui centru de calcul. Dupa cum se va vedea în continuare în acest capitol, sistemele din categoria 3 anticipeaza din punct de vedere functional dezvoltarile tehnologice ale anilor '90 în domeniul magaziilor de date si în realizarea de instrumente de tip OLAP sau EIS (*Executive information systems*).

2.1.2. Tipuri de date

Înainte de a analiza tipurile de date care se acumuleaza într-o organizatie cu o viteza din ce în ce mai mare, se va încerca o foarte sumara clarificare a continutului unor notiuni, care sunt, de multe ori, folosite în mod amestecat. Este vorba de *date*, *informatii* si *cunostinte*. Fara a avea pretentia unei abordari aprofundate, facute din mai multe perspective (teoria comunicarii, stiintele cognitive, stiinta organizarii, stiintele politice, literatura, informatica), precum cea propusa de Rowley (1988), se vor adopta în continuare urmatoarele întelesuri care sunt asemanatoare cu cele propuse de Turban si Aronson (1998) si Targett (1996).

1. *Datele sunt* “urmele” lasate de anumite obiecte (evenimente, stari, situatii sau activitati petrecute sau anticipate, procese de calcul, judecati, opinii etc.) care pot fi reprezentate sub forma unor numere, siruri de caractere, grafice, sunete sau imagini statice (fotografii) sau în miscare (video) în vederea colectarii, memorarii si altor prelucrari ulterioare. Luate ca atare, în contextul unei anumite probleme, datele pot sa nu aiba un înteles care ar putea servi unei activitati decizionale.

2. *Informatiile* sunt acele date care sunt luate în considerare si sunt organizate în asa fel încât au un înteles si o valoare în contextul problemei decizionale si îi sunt de folos utilizatorului pentru rezolvarea sarcinii sale.

3. *Cunostintele* sunt acele date organizate si prelucrate astfel încât favorizeaza înțelegerea unui anumit domeniu cât si identificarea si rezolvarea eficace a problemelor din domeniul respectiv.

În continuare, se va urmări folosirea termenului cel mai potrivit, dar pot apărea situatii în care denumiri încetatenite deja în literatura nu sunt aparent în concordanta cu acceptiunile adoptate mai sus.

Turban si Aronson (1998) identifica mai multe tipuri de date care se acumuleaza într-o organizatie si anume:

1. *Datele interne* se refera la resursele organizatiei (umane, tehnice, financiare etc.) si procesele, evenimentele si activitatile desfasurate în acea organizatie. Într-o organizatie informatizata, aceste date sunt stocate de obicei în diferite baze de date care servesc în aplicatii distincte.

2. *Datele externe* privesc mediul înconjurator (natural, economic, social, de reglementari) în care organizatia își desfasoara activitatea si pot avea surse diferite precum: mijloacele de informare în masa, opiniile comunicate de catre clienti si parteneri, bibliotecile specializate sau generale, Internetul si serviciile diferite de bancile de date.

3. *Datele care apartin unor persoane* angajate în organizatie si pe care acestea le folosesc în interesul organizatiei.

Principalele probleme generale care pot aparea în ceea ce privește conținutul datelor au fost semnalate de Alter (1980) și sunt legate de imprecizia, incompletitudinea, indisponibilitatea sau chiar inexistența (deoarece nu le-a cerut nimeni) acelor date care sunt considerate necesare la un moment dat. Explozia informațională asociată cu multiplicarea și diversificarea surselor au condus în ultima vreme la alte tipuri de complicații precum: opacitate (supraabundența a datelor disponibile asociată cu nerelevanța lor pentru sarcinile de rezolvat), conținut contradictoriu și formate incompatibile datorită platformelor diferite pe care datele se stochează. Acestor complicații li se adaugă problemele de natură tehnologică care vor fi expuse în continuare.

2.1.3. Ratiuni tehnologice

Apariția conceptului de magazie de date este justificată prin limitele *sistemelor de prelucrare on line a tranzacțiilor (On line transaction Processing – OLTP)*. Acestea nu pot furniza cu promptitudinea cerută și în formatul dorit către manageri acele informații care le sunt necesare. De cele mai multe ori, este vorba de informații care servesc la verificarea unor ipoteze înlănțuite considerate în rezolvarea problemelor decizionale. Aceste informații se pot obține (în majoritatea cazurilor) din analiza unor serii de timp diferite, care se găsesc în baze de date distincte. În contrast cu “datele operaționale”, pe care le manipulează sistemele de tip OLTP, cerințele decidenților se numesc în terminologia IBM (1996) “date informaționale”.

Câteva motive pentru o abordare nouă sunt date în continuare (Dhar, Stein, 1997; IBM, 1996; Turban, Aronson, 1998). În primul rând, sistemele de tip OLTP, sunt specializate pe tipuri de probleme, de exemplu conducerea producției, efectuarea de tranzacții comerciale etc. Prin firea lucrurilor, ele au ca misiune principală manipularea rapidă și sigură a datelor și sunt denumite uneori sisteme cu misiune critică, sau “cu timp critic” (“mission/time critical systems”). Pentru a asigura performanțele

cerute și, uneori, din motive istorice sau de securitate, ele nu sunt proiectate pentru a lucra în cooperare cu alte sisteme. În consecință, cu toate progresele tehnologice relative privind “integrarea aplicațiilor” (Filip, Barbat, 1999), realizarea de misiuni suplimentare ca de exemplu cele de furnizare de informații către manageri, altele decât cele originale, poate reprezenta o perturbare inacceptabilă, care poate împiedica asupra performanțelor proceselor conduse cu astfel de sisteme. Alte motive care indică sistemele de tip OLTP ca neadecvate pentru nevoile managerilor sunt legate de: a) limitarea datelor manipulate de aceste sisteme la valorile curente necesare îndeplinirii misiunii lor și b) aspectele de securitate în cazul – din ce în ce mai răspândit al producției globalizate – în care diferitele baze de date sunt distribuite din punct de vedere geografic.

Dacă se analizează problema din punct de vedere al instrumentului informatic care administrează baza de date se pot observa alte complicații. Într-o organizație există “sisteme mostenite” (“legacy systems”) care sunt realizate prin folosirea unei diversități de *sisteme de gestiune a bazelor de date (SGBD)* (Date, 1986; Ulman, 1982). Acestea sunt fie mai vechi (indexat secvențial, de tip rețea, ierarhice), fie mai noi (relazionale, orientate pe obiecte), sau specializate (de exemplu pentru lucrul în timp real, pentru stocarea de date multimediale etc.). La ora actuală predomină SGBD de tip relational, care permit efectuarea de zeci de mii de tranzacții pe minut (1996). SGBD relazionale au o serie de avantaje. Ele permit acumularea unor volume impresionante de date într-un mod neredundant sub forma unor tabele, care pot fi combinate prin anumite operații, bine fundamentate matematic, pentru a obține informațiile dorite. Interogarea poate fi realizată într-un mod flexibil folosind limbajul SQL (“Structured query language”). Cu toate aceste avantaje, folosirea directă a SQL de către un manager de pe nivelurile superioare ale ierarhiei de conducere, necesită din partea acestuia cu-

nostinte si timp pe care acesta nu le are în unele cazuri.

Pentru a ocoli aceste neajunsuri, o solutie intermediara au constituit-o *sistemele de informare a directorilor (Executive information systems- EIS)*. Acestea sunt niste sisteme de tip "front-end" pentru sistemele de tip OLTP si au ca misiune realizarea de operatii de agregare a datelor primite de la sistemele de tip OLTP care sunt stocate în mod intermediar în dispozitivele de memorare ale EIS. Managerul, în loc sa formuleze comenzi de interogare într-un limbaj misterios pentru el (SQL) au posibilitatea sa aleaga dintr-un meniu prestabilit pentru a obtine informatia necesara.

Desi solutiile de tip EIS au reprezentat un pas important înainte în asistarea deciziilor centrate pe date, ele sufera de anumite neajunsuri izvorâte din însasi conceptia de realizare. În primul rând, culegerea de date de la sisteme de tip OLTP ramâne o problema care, uneori, nu este simpla din punct de vedere tehnic. În al doilea rând, solutia EIS sufera de inflexibilitate. Inflexibilitatea se manifesta atât în utilizare (este necesara parcurgerea de catre manager a unor secvente de meniuri prestabilite si nimic mai mult sau mai putin) cât si în întretinere si dezvoltare, situatie în care poate fi necesara reprojectarea de la început pentru a considera cerinte de informare suplimentare. Subiectul EIS va fi dezvoltat într-un articol ulterior.

Cele aratate mai sus cât si limitele atinse de instrumentele de tip sisteme de gestiune a bazelor de date au determinat necesitatea unui concept nou (magazia de date) si, respectiv, a unor instrumente informatice noi (OLAP). Acestea vor fi prezentate în subcapitolele care urmeaza.

2.1.4. O solutie moderna

Solutia pentru asistarea deciziei centrate pe date, adoptata cu entuziasm de furnizori mai mari sau mai mici de produse informatice (Pendse, 2000 c) se bazeaza pe conceptul de magazie de date (DW) si pe instrumentele de prelucrare analitica on-line (OLAP).

Magazia de date este o baza de date specializata care raspunde nevoilor si cerintelor specifice decidentilor aflati pe nivelurile superioare ale ierarhiei de conducere a întreprinderii. Ea cuprinde multe tipuri de date, provenind de la mai multe "surse de aprovizionare", care pot fi sisteme de tip OLTP, calculatoare de proces, retele industriale, biblioteci de carti, fotografii sau CD-ROM, Internet etc. Existenta mai multor tipuri de date diferentiaza magazia de date de noile versiuni 3-D ale foilor de calcul electronic. Acestea au fost inaugurate de 3-D sheet al IBM-Lotus si au fost urmate de solutiile de tip „notebook” si „workbook” ale produselor Quatro Pro for Windows (al lui Corel) si respectiv Excel (al lui Microsoft).

Principalele operatii care se efectueaza asupra datelor primite sunt (Dhar, Stein, 1997):

1. *încarcarea* datelor din diferite surse, care consta în urmatoarele: a) detectarea datelor noi de interes aflate în bazele de date sursa si b) determinarea modului si a locului pentru încarcare. Încarcarea se efectueaza la momente programate, în regim "pe loturi" (batch) pentru a nu perturba functionarea sistemelor furnizoare de tip OLTP;
2. *conversia* datelor din formatul original în cel adoptat în magazia de date;
3. *curatirea* datelor, care cuprinde functii de identificare si corectare a erorilor de conversie si completarea omisiunilor;
4. *transformarea* datelor prin operatii de agregare si rezumare.

Instrumentele de tip OLAP servesc la manipularea datelor într-un mod care extinde si flexibilizeaza functiile si maniera de operare a sistemelor de tip EIS descrise în subcapitolul anterior. Într-un mod intuitiv, functionarea OLAP este sugerata de operatii de tip "feliere si decupare de cubulete" (*Slice and dice* – S&D) a bazei de date în scopul de a permite decidentul sa gaseasca acele informatii care îi permit constatarea unor situatii de interes sau verificarea unor ipoteze. Mai multe detalii despre OLAP vor fi date în capitolul 3.

2.2. Definitii

Conform definitiei lui Inmon (1995), aparent cea mai raspândita si acceptata pâna în prezent, o magazie de date este o colectie integrata de date de tip serii de timp, nevolatile, orientata pe un subiect si care are ca scop sprijinirea proceselor decizionale ale managerilor. Aceasta definitie a celui care este considerat "parintele magaziei de date" (Gray, 1997) va fi detaliata si particularizata în urmatoarele doua subcapitole (2.2.1. si 2.2.).

2.2.1. Caracteristici

O magazie de date este în general un lucru scump. Ea poate costa milioane de dolari si stoca volume de date cuprinse între 50 de Gigaocteti si câtiva Terraocteti (Gray, Watson, 1996). Pentru ca magazia de date sa merite investitia facuta, ea trebuie sa serveasca necesitatile managerilor prin oferirea de raspunsuri rapide si în conformitate cu cerintele de interogare si de prezentare ale acestora. Pentru aceasta, o magazie de date trebuie sa posede o serie de caracteristici generale (Inmon, 1995; Gray, Watson, 1996), care se prezinta mai jos:

1. *Separarea fizica* de bazele de date operationale. Aceasta cerinta are ca scop pe de o parte, furnizarea operativa a informatiilor necesare procesului decizional si pe de alta parte, evitarea perturbării aplicatiilor operative, care trebuie sa satisfaca standardele legate de timpul de raspuns si de siguranta operatiilor.
2. *Orientarea catre un subiect*. Spre deosebire de bazele de date operationale care sunt centrate pe aplicatii sau functii, magazia de date este organizata ca sa vizeze anumite subiecte de interes pentru manager precum: clienti, produse, activitati.
3. *Integrarea fara exceptii a datelor*. Aceasta se traduce prin consistenta (uniformitatea): a) conventiilor folosite în definirea datelor, b) unitatilor de masura utilizate, c) atributelor datelor etc.
4. *Pastrarea seriilor de timp*. Spre deosebire de datele operationale, care sunt valabile numai în momentul accesului, cele continute în DW sunt valabile oricând.

Aceasta se traduce prin urmatoarele trasaturi derivate: a) orizontul de timp de 5-10 ani (fata de maximum câteva zeci de zile în cazul operational), b) structura cheii de acces, care contine elementul "timp" (zi, luna, an) si c) interzicerea modificarii datelor stocate corect la un anumit moment discret de timp.

5. *Nevolatilitatea datelor*. Aceasta se traduce prin reducerea numarului de tipuri de operatii permise în cazul DW la doua: a) încarcarea initiala si b) accesul la date, spre deosebire de cazul operational, în care se efectueaza o multitudine de înserari, stergeri si actualizari la nivelul unei singure înregistrari. Consecintele acestei caracteristici sunt diverse. În primul rând, dispare pericolul potential al actualizarilor eronate. În al doilea rând, se poate realiza proiectarea la nivel fizic pentru a optimiza accesul fara a mai tine cont de cerintele de neredundanta si normalizare. În al treilea rând, dispare necesitatea folosirii unor tehnici complicate pentru asigurarea integritatii datelor si pentru jurnalizare si recuperare cât si pentru deblocare din punct mort.

2.2.2. Tipuri de magazine de date

Termenul de magazie de date este destul de general. Unii autori sau furnizori de produse informatice folosesc diferite denumiri pentru a numi unele cazuri particulare sau componente ale DW. Astfel Turban si Aronson (1998) identifica urmatoarele "componente": a) *baze de date fizice mari* (în care sunt continute datele si metadatele cât si programele de "curatare", organizare, asamblare si preprocesare), b) *magazii de date logice* (care cuprind numai metadatele si programele de organizare, asamblare, preprocesare împreuna cu informatiile necesare pentru a gasi si a avea acces la date indiferent de locul unde sunt stocate) si c) "târguri" sau compartimente de date ("data mart"), care reprezinta subseturi ale unei DW, necesitatile unei parti sau functii a întreprinderii. Târgurile de date împreuna cu legaturile stabilite prin intermediul DW logice constituie componente în dez-

voltarea incrementală a unei DW de întreprindere (Inmon, 1998).

O clasificare recentă și sistematică a tipurilor de magazine de date este propusă de către Firestone (2000), care identifică cinci tipuri de magazine de date. Toate tipurile au în comun faptul că sunt “colecții integrate, nevolatile, de date de tip serie de timp care au ca scop asistarea proceselor decizionale manageriale”. Ceea ce diferențiază tipurile de magazine de date este aria de cuprindere a proceselor decizionale, după cum se arată mai jos:

1. Magazin de date *galactica* (GDW), asistă procesele decizionale manageriale care privesc *oricare și toate* procesele de business și compartimentele întreprinderii cât și întreprinderea luată ca un întreg.
2. Magazin de date *orientată pe un proces de business* (BPDW) asistă procesele decizionale care privesc *oricare și toate* procesele de business și legăturile lor reciproce precum și cu mediul lor înconjurător.
3. Magazin de date *departamentală* (DDW) asistă procesele decizionale care privesc *oricare și toate* compartimentele și interacțiunile lor reciproce precum și cu mediul lor înconjurător.
4. Un târg de date (data mart) de tip *proces de business* (BPDM) asistă procesele decizionale centrate pe un *singur* proces de business.
5. Un târg de date *departamental* (DDM) asistă procesele decizionale centrate pe un singur compartiment.

Legat de această clasificare, Firestone (2000) face câteva observații interesante. În primul rând el consideră că interesul către GDW este în descreștere și că, BPDW și BPDM sunt mai de dorit datorită orientării actuale către considerarea proceselor de business, care nu tin seama de granițele compartimentelor (Hammer, Champy, 1993). În același timp, Firestone (2000) mai observă că simpla reuniune a unor DDM nu constituie o magazie de date deoarece nu asigură suportul pentru deciziile manageriale care privesc procesele care traversează granițele compartimentelor sau interacțiunile dintre compartimente

și pe cele cu mediul înconjurător. În același spirit se arată că, reuniunea unor BPDW nu constituie o magazie de date. Totuși autorul citat nu respinge posibilitatea distribuirii la nivel fizic a unei magazine de date care poate fi caracterizată la nivel logic așa cum s-a arătat la începutul acestui subcapitol în prima clasificare prezentată.

3. Prelucrarea analitică on line

Conform cu Glosarul propus de OLAP Council (1997), “*prelucrarea analitică on line* (OLAP) desemnează o categorie de instrumente software, care permit analizelor, managerilor și directorilor să înțeleagă esența datelor printr-un acces rapid, consistent și interactiv la o mare varietate de vederi posibile ale informațiilor, care au fost obținute prin transformarea datelor primare, astfel încât să reflecte dimensiunile reale ale întreprinderii așa cum o percepe și o înțelege utilizatorul”.

Un *server OLAP* este, conform aceluiași Glosar (OLAP Council, 1997) “un motor de tip multi-utilizator, de mare capacitate pentru manipularea datelor proiectat astfel încât să poată lucra cu structuri de date multidimensionale. O *structură multidimensională* este alcătuită astfel încât orice dată poate fi localizată și accesată prin intersecția coordonatelor sale de-a lungul acelor dimensiuni care definesc amplasarea datei. Proiectarea serverului și a structurii de date sunt realizate astfel încât să se poată realiza atât optimizarea regasirii ad-hoc a informației cât și calcule flexibile folosind formule matematice”.

OLAP și magazin de date sunt complementare. După cum se arată în Cartea albă a Consiliului OLAP (Forsman, 1997), este important să se facă distincția între conceptul de magazie de date și OLAP. În timp ce magazin de date acumulează informații cu caracter tactic într-o bază de date relațională specializată, cu scopul de a furniza răspunsuri la întrebări de tipul “Cine...?” și “Ce...?”, OLAP folosește de obicei o vedere cu mai multe dimensiuni a unor date agregate pentru a putea răspunde la întrebări suplimentare de tipul “De ce...?”

(“Why...?”) si “Ce se întâmpla daca...?” (What if...?”), tipice pentru sistemele de asistare a deciziei.

3.1. Evolutii

Desi, dupa cum arata Pendse (2000 b), radacinile OLAP pot fi identificate cu multi ani în urma în limbajul APL, adevarata lansare a termenului a fost prilejuita de un articol al lui Codd et all. (1993). Cel care este socotit “parintele bazelor de date relationale a aratat ca, datele ar trebui folosite pentru o prelucrare analitica si ca în acest scop bazele de date multidimensionale sunt mai adecvate si, în consecinta, le vor înlocui pe cele relationale.

3.1.1. Functii

Raden (1995) distinge doua tipuri de baza de instrumente OLAP:

1. OLAP în *sens restrâns* care permite interogari interactive ale datelor prin care se poate realiza o analiza constând din *feliere si decupari de cubulete* („slice and dice”) ale bazei de date si din *forari* (“drill-down”) succesive pe niveluri de detalieri, din ce în ce mai pronuntate. Informatiei i se atribuie caracteristica de *multidimensionalitate* si poate fi afisata în tabele încrucisate, pentru care exista instrumente specializate pentru a realiza pivotarea axelor si tabularea încrucisata. Toate operatiile sunt numai de citire.

2. OLAP în *sens larg* permite operatii suplimentare precum: a) actualizare, b) calcule de baza de agregare sau consolidare pe una sau mai multe dimensiuni si c) calcule mai avansate de tipul: productie, analiza de trenduri si chiar optimizare.

3.1.2. Regulile lui Codd si testul FASMI

Cartea alba propusa în 1993 de Codd et all, intitulata “Providing OLAP (On line Analytical Processing) to User - Analyst: An IT Mandate”, la comanda firmei Arbor Software (Pendse, 2000 a), continea 12 reguli, a caror îndeplinire putea califica un produs software ca fiind de tip OLAP. Aceste reguli, au facut epoca la vremea respectiva, desi au existat suspiciuni ca ele

au fost influentate de comanda unui producator de software (Pendse, 2000 a). Regulile defineau urmatoarele caracteristici ale produselor de tip OLAP (Gray, Watson, 1996; Pendse, 2000 a):

R1 Posibilitatea unor *vederi conceptuale multidimensionale* care este asociata cu operatii de feliere si decupare de cuburi (“slice and dice”).

R2 *Transparenta pentru utilizator*, în sensul ca acesta putea nici macar sa nu aiba idee de sursele datelor, care puteau fi oricât de eterogene.

R3 *Accesibilitatea*, care dadea motorului OLAP caracterul de mediator (“middle-ware”) între sursele eterogene de date si front-end-ul OLAP.

R4 *Raportarea consistenta*, care cere ca performantele rapoartelor sa nu fie degradate în masura semnificativa prin cresterea numarului de dimensiuni ale bazei de date.

R5 *Arhitectura de tip client – prestator* (“client – server”), asociata cu un grad ridicat de inteligenta a instrumentului OLAP, astfel încât clienti noi si diversi sa poata fi atasati cu un efort minim de programare.

R6 *Numar generic de dimensiuni echivalente* atât în ceea ce priveste structura cât si operatiile posibile. Aceasta regula este cea mai controversata din tot setul celor 12 reguli. Pendse (2000 a) arata ca, în sistemele aplicative, exista tendinte de a ignora aceasta regula, în timp ce furnizorii de produse informatice declara ca încearca sa o respecte.

R7 *Manipularea automata a unor matrici rare*. Ulterior aceasta a fost extinsa la *ajustarea automata a schemei la nivel fizic* în scopul adaptarii la tipul de model de date, la volumul si la existenta unor pozitii goale în matricea datelor, care poate deveni rara.

R8 *Suport multi-utilizator*, care extindea accesul dincolo de aplicatiile “numai citire” (“read only”) si permitea acces concurrent (inclusiv actualizari) cu facilitati de asigurare a integritatii si securitatii datelor.

R9 *Efectuarea fara restrictii de operatii între dimensiuni care însemna ca, se putea realiza orice fel de operatii folosind datele*

care aveau coordonatele în dimensiuni diferite.

R10 Manipularea intuitiva a datelor, care trebuia realizata prin actiunea directa asupra celulelor dintr-o vedere fara a recurge la miscarea prin meniuri, sau la comenzi multiple.

R11 Raportarea flexibila, care cerea ca aranjamentul axelor în rapoarte sa fie la libera alegere a utilizatorului.

R12 Numarul nelimitat de dimensiuni si niveluri de agregare, fiind acceptate totusi (din motive de capacitate limitata a calculatoarelor) valori de minimum 15 dimensiuni si fiind recomandate valori de 20 de dimensiuni. Dhar si Stein (1997) indica o limita maxima practica de 7 dimensiuni iar, mai de curând, Pendse (2000 a) arata ca sunt putine aplicatii care depasesc cifre care variaza între 8 si 10 dimensiuni.

Pendse (2000 a) arata ca, în 1995, Codd a mai adaugat sase reguli si a restructurat întregul set în patru grupe de *trasaturi* ("features"), dupa cum urmeaza:

1. *Trasaturi de baza*, care sunt numerotate de la F1 la F8. Dintre acestea, trasaturile F1, F2, F3, F5, F6, F7 si F8 corespund în ordine regulilor R1, R10, R3, R5, R2 si respectiv R8. În plus, sunt introduse urmatoarele trasaturi noi:

F4. Extractie pe loturi ("batch") vs. de tip interpretor, care se traduce prin posibilitatea de a folosi fie propria baza de date aranjata special, fie de avea acces "pe viu", direct de la surse externe.

F5. Asistarea modelelor de natura de tip OLAP si anume: a) raportarea statica parametrizata, b) "felierea si decuparea de cubulete cu forarea în adâncime", c) analiza de tip "Ce se întâmpla daca...?" (simulare) si d) model de "urmarire a unui scop" (optimizare).

2. *Trasaturile speciale*, numerotate cu F9 – F12:

F9. Tratarea datelor nenormalizate.

F10. Memorarea rezultatelor OLAP si pas-trarea lor separat de datele sursa.

F11. Extragerea valorilor care lipsesc, interpretata de Pendse (2000 a) prin reali-

zarea distinctiei acestor valori de valorile nule.

F12. Tratarea valorilor care lipsesc, care sunt ignorate de analizorul OLAP, indiferent de provenienta.

Plecând de la constatarea ca, "18, sau chiar 12 reguli sunt prea mult" si încercând sa defineasca, în mod independent de implementare, caracteristicile unei *aplicatii OLAP*, Pendse (2000 a) propune testul FASMI ("Fast Analysis of Shared Multidimensional Information" – *Analiza rapida a informatiei multidimensionale partajate*). Cele patru cuvinte cheie care califica un produs ca fiind conform cu OLAP sunt descrise mai jos.

1. Cuvântul cheie "analiza" cere ca sistemul de tip OLAP sa permita efectuarea de catre utilizator, într-un mod suficient de facil si intuitiv a oricarei analize logice sau statistice care este relevanta pentru aplicatie. Aceasta include, pe lânga functii pre-programate si posibilitatea de definire a unor calcule ad hoc si furnizarea de rapoarte în orice forma dorita fara a face apel la limbaje de programare profesionale.

2. Cuvântul cheie "rapida", indica faptul ca, sistemul trebuie sa furnizeze majoritatii utilizatorilor datele cerute într-un interval de timp pâna la cinci secunde. Pentru analize simple, cerinta de timp de raspuns este de o secunda si, numai în foarte putine cazuri exceptionale, se admite un timp de raspuns mai mare de 20 de secunde.

3. Cuvântul cheie "informatie" se refera la toate datele primare si informatiile care au fost obtinute din acestea în masura în care ele sunt relevante pentru aplicatie. Masura volumului de informatie se refera la cantitatea de informatie care poate fi *manipulata* si nu la capacitatea de *stocare*.

4. Cuvântul cheie "multidimensionala", care reprezinta punctul esential al testului FASMI, arata ca o aplicatie OLAP trebuie sa poata furniza vederi conceptuale ale datelor cu mai multe dimensiuni, fara a fi nevoie sa se specifice a priori numarul acestora, care e dependent de aplicatie.

5. Cuvântul cheie "partajata" implica cerinte privind asigurarea confidentialitatii si

de blocare a actualizarilor concurente când e cazul unor accese multiple de tip “înscrisoare”.

3.2. Tehnologii de implementare

Dupa cum s-a aratat mai înainte cuvântul cheie care definește conceptul de OLAP (produs sau aplicatie) este în mod constant *multidimensionalitatea* vederilor folosite în analiza. În prezent, coexistă două solutii de implementare a multidimensionalității: a) Solutia *MOLAP* (“*Multidimensional OLAP*”), care folosește baze de date multidimensionale (“*Multidimensional Data Base*” – *MDB*), sau baze de date de tip “hipercub” (de fapt, hiperpoliedru) și b) Solutia bazelor de date relationale care asigură multidimensionalitatea pentru analiza, denumite pe scurt *ROLAP* (“*Relational/OLAP*” sau “*OLAP - on - Relational*”).

3.2.1. Solutii de tip ROLAP

O baza de date relationala are două dimensiuni sau axe (coloane și rânduri) și presupune existența unui limbaj de interogare de tip SQL (sequential query language). Asigurarea posibilității de a efectua analize multidimensionale folosind o baza de date relationala se realizează în implementările actuale prin folosirea unei scheme de proiectare intitulată “schema stea”, sau “floarea soarelui” (IBM, 1996; Gray, Watson, 1996; Grady, 2000). Aceasta creează două tipuri de tabele: a) o “tabela de fapte” (care poate avea milioane de rânduri) și conține informațiile care sunt subiectul interogărilor și b) mai multe tabele care conțin datele descriptive privind fiecare dimensiune.

Raden (1998) propune câteva criterii pentru a califica un produs drept ROLAP dintre care amintim:

? existența unui generator SQL, care poate realiza operații de selectare în mai multe treceri;

? posibilitatea de a realiza clasificări nebanale, comparații și calculul procentelor pe clase;

? existența unui mecanism care descrie modelul datelor prin intermediul metadate-

lor și care poate folosi metadatele în timp real pentru a realiza interogări;

? posibilitatea de a partitiona aplicația între clienți, prestatori și un element mediator care gestionează sirurile de apeluri la baza de date.

Primul produs de tip ROLAP, denumit Metaphor, a fost lansat în 1984 pentru aplicațiile de marketing la companiile producătoare de bunuri de consum. Succesul comercial al lui Metaphor (detinut din anul 1991 de către firma IBM) a fost destul de limitat. De atunci o serie de alte produse au fost lansate, ca de exemplu, AXSYS al firmei Advantage, Beacon al firmei Prodea sau MetaCube al lui Stanford Technology Group. De asemenea s-au petrecut o serie de fuziuni și achiziții de firme din domeniu (Raden, 1995; Pendse, 2000 b).

3.2.2. Solutii de tip MOLAP

Operațiile primare și tipice pentru aplicațiile OLAP sunt *felierea și forarea*. Ele presupun o organizare logică a datelor sub forma unui hiper-cub (n – dimensional) format din *celule* în care timpul este o dimensiune obligatorie (Gray, Watson, 1996) în locul tabelelor bidimensionale ale bazelor de date relationale. Raden (1995) identifică trei etape de proiectare a hiper-cubului: a) alegerea subiectelor de interes pentru întreținere (de exemplu: vânzările, reclamațiile, sau capacitățile de producție), b) stabilirea valorilor care trebuie culese (de exemplu volumul lucrărilor, elementele de cost, sau gradul de încărcare și timpii de staționare ai mașinilor și instalațiilor) și c) determinarea granularității datelor sau a nivelului de detaliu. Același autor observă că, spre deosebire de bazele de date relationale, MDB nu au încă un model acceptat de toată lumea și nici o metodă de acces standard cum era SQL în cazul relational.

Primul instrument informatic pentru MDB a fost Express, care a fost gândit la începutul anilor '70, pentru aplicații cu timp partajat în mediul academic. În prezent, Express este detinut din anul 1995 de către firma ORACLE (un mare producător de

instrumente pentru bazele de date relationale) și își mentine conceptele initiale deși prezintă noi soluții de implementare. De atunci, alte produse au fost lansate ca de exemplu: System W al firmei Comshare, sau Essbase al firmei Arbor și Produsul Tensor al firmei Microsoft, care a fost adoptat ca standard industrial chiar înainte a fi lansat. Cel mai nou produs al aceleiași firme (Microsoft), denumit Plato, posedă o arhitectură sofisticată (ROLAP/MOLAP/Hibrid) (Pendse, 2000 b; Raden, 1995). Detalii practice privind implementarea soluțiilor de tip MOLAP sunt date de Thomsen (1997). Pentru completitudine tabloului, vom menționa în trecut dezvoltarea în paralel a soluției Improve a lui IBM-Lotus pentru fișiere de calcul tabelar electronic.

O comparație a soluțiilor de tip ROLAP și MOLAP/MDB, folosind mai multe criterii precum: a) funcționalitatea, b) adecvanta, c) performanțele și d) scalabilitatea împreună cu recomandări de a folosi una sau alta dintre cele două soluții este dată de Raden (1995).

4. Mineritul datelor

Mineritul datelor ("Data Mining" – DM), denumit uneori și *descoperirea cunoștințelor* în bazele de date ("Knowledge Discovery in Data Bases" – KDD) este cea mai recentă tehnologie de analiză a datelor, asociată, alături de OLAP, cu conceptul de magazie de date.

Alte denumiri și înțelesuri cuprinse în conceptul de DM/KDD sunt: "arheologia datelor", "prelucrarea pattern-urilor de date", "culesul informațiilor" sau "dragarea datelor" (Turban, Aronson, 1998).

Radacinile DM/KDD sunt considerate a fi în principal în statistica matematică și în pachetele software folosite în științele sociale (Agosta, 2000) și în inteligența artificială (Moxon, 1996).

4.1. Definiții

Ca și în cazul OLAP, în literatura există mai multe definiții pentru DM/KDD, lucru

absolut explicabil în cazul unei tehnologii noi, aflate într-o mare efervescență.

O definiție mai veche (Frawley, 1991, citat de Mertens et al., 1996) spune că, mineritul datelor constă în "extragerea nu simplă și banală a unor informații potențial utile, implicite și recunoscute dinaintea bazei de date". Gray și Watson (1996) arată că, "mineritul datelor permite analiștilor și managerilor să găsească în magazia de date răspunsuri la problemele întreprinderii, pe care aceștia nici măcar nu și le-au pus". Moxon (1996) adoptă un înțeles mai restrâns al mineritului datelor pe care îl consideră "un set de tehnici folosit în abordarea automatizată a operațiilor de *explorare exhaustivă* a unor baze de date foarte mari și de *aducere la suprafață* a unor relații complexe existente în acele baze de date".

Într-o întreprindere se acumulează de-a lungul timpului foarte multe date. Datele pastrate de multă vreme în fișiere pot conține informații și cunoștințe ascunse care pot servi la bunul mers al întreprinderii, dar luate ca atare, nu au mare utilitate dacă nu sunt însoțite de mecanisme care să permită explorarea lor și înțelegerea fenomenelor care au guvernat funcționarea surselor de date. După cum arată H. Simon, laureatul Premiului Nobel pentru economie, la Conferința Internațională pentru Economie, Management și Tehnologie Informației (desfășurată la Tokio în anul 1992), "Informația nu lipsește. Ceea ce lipsește este timpul managerului de a considera toate informațiile care sunt disponibile".

Tehnologiile de analiză a datelor prezentate mai înainte (EIS, OLAP) pot fi văzute ca jucând rolul unor "ajutoare de verificare" ale ipotezelor formulate de utilizatori (manageri sau analiști) privind relațiile care există între diferitele date acumulate. Cheia succesului constă în intuiția utilizatorului de a formula cele mai adecvate întrebări inițiale și în capacitatea lui de a continua rafinarea analizei pe baza informațiilor primite de la sistemul informatic. Contribuția principală a acestuia constă în

facilitarea procesului de analiza. Aceasta se realiza fie prin usurarea activitatii de formulare a interogarilor si de obtinere a unor rapoarte care permiteau o vizualizare sugestiva (în cazul EIS), fie prin obtinerea rapida a unor raspunsuri prin explorarea multidimensionala a unor ierarhii de date agregate, însoțita eventual de prelucrari statistice (în cazul OLAP).

Dupa cum se poate observa (Fayyad, Uthurusamy, 1996; Moxon, 1996), succesul unei astfel de abordari poate fi afectat de capacitatea limitata a utilizatorului de a intui fenomenele si de îndemânarea sa de a folosi toate posibilitatile oferite de instrumentul informatic. Eficacitatea procesului poate scadea dramatic în cazurile complexe, în care utilizatorul are de a face cu milioane de situatii, fiecare descrisa prin sute de înregistrari.

În cazul DM/KDD, sistemul informatic are ca misiune investigarea exhaustiva a datelor pentru a scoate la lumina clasificari, asocieri, sau alte modele ale datelor, în timp ce analistului (sau managerului) îi revine sarcina de a decide ce sa faca apoi cu aceste rezultate (IBM, 1996).

4.2. Functii

Principalele functii realizate de DM/KDD sunt (Moxon, 1996):

1. *Gruparea* ("clustering"), care este, de cele mai multe ori, prima etapa în analiza bazata pe mineritul datelor. Ea consta în identificarea grupurilor de înregistrari înrudite care pot constitui punctul de plecare pentru explorarea în continuare a unor alte relatii.

2. *Clasificarea*, care este cea mai raspândita tehnica de minerit al datelor, foloseste un set de exemple în scopul realizarii unui model pentru aranjarea în clase a multimii de înregistrari.

Alte functii pot fi: estimarea (o varianta a clasificarii), asocierea si analiza bazata pe secvente.

Mineritul datelor foloseste o serie de algoritmi si tehnici de analiza, care sunt descriși în alte articole ale acestei serii, precum arborii de decizie rationamentul bazat

pe situatii, logica vaga, algoritmi genetici, sau transformarile bazate pe fractali. O lista bogata si la zi de referinte privind solutiile dezvoltate în domeniu este data de Fayyad, Piatetsky – Shapiro si Smith (2000) iar lucrarea recenta a lui Witten, Frank si Gray (1999) descrie o serie de tehnici bazate în special pe instruire.

În ceea ce priveste aplicatiile DM/KDD se pot enumera o serie de realizari (Turban, Aronson, 1998; Cox, 2000) dintre care de interes în contextul acestei lucrari sunt: a) clasificarea clientilor în categorii, b) stabilirea preturilor si pozitionarea produselor pe piata, c) descoperirea de clienti potentiali si pastrarea celor existenti, d) predictia si prevenirea situatiilor de faliment, e) determinarea nivelului de siguranta al portofoliului de comenzi si a masurilor adecvate care trebuie sa fie luate etc.

Toate aceste aplicatii se caracterizeaza prin aceea ca datele sunt sub forma numerica sau a unor siruri de simboluri ordonate într-un mod consistent (de obicei, tabele cu rânduri si coloane în baze de date relationale sau în foi de calcul tabelar electronic). Aceasta permite din capul locului înțelegerea semanticii si a structurii datelor de catre utilizator, în timp ce tehnologiei DM îi revine rolul de a descoperi paternuri prin explorarea exhaustiva a bazei de date. În ultima vreme se constata o noua directie promitatoare de dezvoltare si anume cea denumita "mineritul textelor" cu format liber (Cox, 2000).

Mineritul datelor este o tehnologie "în crestere", care pare sa fie pe punctul de a deveni o "tehnologie cheie". Exista o serie de prototipuri si chiar produse comerciale propuse atât de firme mici si inovatoare cât si de marii producatori (de exemplu, Intelligent Miner al IBM lansat în 1998). Simptomul cel mai semnificativ care poate fi observat este tendinta multor giganti ai pietei tehnologiei informatiei (Microsoft, Oracle, IBM, NCR, Yahoo, Amazon etc.) de a achizitiona produse si firme inovatoare în domeniul DM/KDD (Augusta, 2000). Cu toate acestea, dezvoltarea si afirmarea deplina a domeniului este conditio-

nata de rezolvarea adecvata a unor probleme tehnice (Moxon, 1996) precum: a) sensibilitatea fata de datele "necurate" si b) capacitatea limitata de explicare a rezultatelor în termeni inteligibili pentru om (cum se întâmpla de altfel în cazul sistemelor expert bazate pe reguli).

5. Note si comentarii

În acest articol s-a încercat sa se prezinte, într-un numar rezonabil de pagini, principalele abordari privind instrumentele informatice de asistare a deciziilor bazate pe date. Aceste abordari, care poarta denumiri cumva insolite pentru un decident, precum magazii si târguri de date, minierul datelor, sau sunt referite prin abrevieri misterioase ca de exemplu MOLAP, ROLAP, au cunoscut o dezvoltare impresionanta începând cu mijlocul anilor '90. Aceasta miscare a fost sustinuta în principal de interesul marilor firme de pe piata tehnologiilor informatiei si în mai mica masura de lumea academica. O serie de produse comerciale în domeniul DW, OLAP, DM/KDD elaborate de IBM si partenerii sai sunt mentionate de IBM (1996), iar solutiile furnizate de alte mari firme (NCR, ORACLE, Comshare, Seagate Software) sunt descrise de Watson si Gray (1997).

Cu toate acestea, se pot auzi voci care îndeamna la precautie. Astfel, P. Keen (1997), o mare autoritate si un pionier în domeniul deciziei asistate de calculator, arata ca, "furnizarea de date nu creeaza informatie. Informatia nu conduce automat la cunostinte. Cunostintele nu conduc în mod direct la actiuni. Actiunea în business si impactul sau trebuie sa fie tinta. Este un pericol real ca înmagazinarea de date si domeniul managementului cunostintelor sa treaca usor cu vederea cele de mai sus. Trebuie sa începem cu oamenii, nu cu informatia".

Capitolul nu a abordat unele subiecte care, de multe ori, sunt descrise în literatura în contextul instrumentelor informatice care asista decizia bazata pe analiza datelor precum sistemele centrate pe *date spatiale/geografice* (Geographic Informa-

tion Systems - GIS), sau încercările de folosire a abordarii obiectuale în magazinele de date. Cititorului interesat de aceste subiecte i se recomanda adresele:

1. <http://mis.ucd.ie/iswsdss/sdss.html>,

unde se poate gasi pagina lui P.Keenan referitoare la sisteme de asistare a deciziilor bazate pe date spatiale si

2. <http://www.dkms.com/ODDW2.html>,

unde se poate gasi Cartea alba nr.5 a Consiliului OLAP, din august 1997 privind magazinele de date orientate pe obiecte.

Dupa cum observa Gray (1997), "anul 1996 a fost anul magazinei de date" (în ceea ce priveste publicarea de lucrari în domeniu). Dintre lucrarile aparute în acest an, se recomanda studierea împreuna a cartilor scrise de Inmon (1996), Inmon si Hackarthorn (1996) si Kimball (1996), care dau o imagine aproape completa asupra construirii si utilizarii magaziiilor de date si respectiv, a unor studii de caz din diferite domenii aplicative. O buna prezentare intuitiva a conceptelor DW, OLAP si DM/KDD se gaseste în capitolul 4 al cartii lui Dhar si Stein (1997). Combinarea recenta a DW cu tehnologia web este prezentata de Kimball si Merz (2000) care introduc conceptul promitator de "Data Webhouse".

Câteva adrese Internet recomandate pentru aprofundarea studiului si urmarirea dezvoltarilor din domeniu sunt:

1. <http://dssresources.com/dsstypes/ddss.html>,

care reprezinta o buna "poarta de intrare" pentru materialele disponibile pe Internet în domeniul asistarii deciziei bazate pe date. O buna parte dintre lucrarile referite în acest capitol au putut fi gasite plecând de la aceasta pagina;

2. <http://dssresources.com/dssbook>, care contine capitolul 7 al hipercartii lui Daniel Power (2000) dedicat asistarii deciziei bazate pe date;

3. <http://www.datawarehousing.com/papers.asp>, a lui Data Warehousing Center, de unde se poate începe explorarea unor articole interesante;

4. http://www.dkms.com/white_Papers.htm, de la care se pot explora o serie de Carti

albe în domeniul magaziilor de date, OLAP și mineritului datelor.

Nota. Autorul multumeste domnului profesor Horatiu Dragomirescu pentru o serie de materiale documentare și adrese furnizate.

Bibliografie

- ✦ Agosta, L. (2000). From data to insight: the critical path to data mining, a short history of data mining. **PC/AI**, Sept/Oct, 16-21.
- ✦ Alter, S. (1977). A taxonomy of Decision Support Systems. **Sloan Management Review**, Fall, 39-56.
- ✦ Alter, S. (1980). **Decision Support Systems; Current Practices and Continuing Challenges**. Addison – Wesley, Reading MA.
- ✦ Codd, E.F., S.B. Codd, T.S. Clinch (1993). Beyond decision support. **Computerworld**, 26 July.
- ✦ Cox, E. (2000). Free – form text data mining integrating fuzzy systems, self – organizing neural nets and rule-based knowledge bases. **PC/AI**, Sept.-Oct., 22-25.
- ✦ Date. C.J. (1986). **An Introduction to Database Systems**. Fourth Edition, Addison – Wesley Publishing Co., Reading MA.
- ✦ Dhar, V., R. Stein (1997). **Intelligent Decision Support Methods; the Science of Knowledge Work**. Prentice Hall, Upper Saddle River, New Jersey.
- ✦ DKMS (2000). **White Papers, Reports, Working Papers, and DKMS Briefs**. Executive Information Systems Inc http://www.dkms.com/white_papers.htm.
- ✦ DWC – Data Warehousing Center. **Paper and Articles** (www.datawarehouse.com/papers/).
- ✦ Fayyad, U. (1996). Data mining and knowledge discovery in data bases. **Communications of the ACM, (Special Issue)**, 39 (11).
- ✦ Fayyad, U., G. Piatetsky - Shapiro, P. Smith (2000). Expanded references. The KDD process for extracting useful knowledge from volumes of data (<http://www.research.microsoft.com/research/dataware/CACM-DM-refs/fayyad-intrs.htm>).
- ✦ Filip, F. G. (2000). Decizie asistata de calculator; metode și tehnici de asistare a deciziilor centrate pe judecata umana. **Informatica economica**, IV (3), 10-22.
- ✦ Filip, F.G., B. Barbat (1999). **Informatica aplicata; noi paradigme și aplicatii**. Editura Tehnica, Bucuresti.
- ✦ Firestone, J.M. (1997). **Object Oriented Data Warehousing**. White Paper No. Five. Executive Information Systems, Inc. (<http://www.dlms.com/00DW2.html>).
- ✦ Firestone, J.M. (2000). DKMS Brief No. Six: **Data Warehouses, Data Marts, and Data Warehousing: New Definitions and New Concepts** (<http://www.dkms.com/DWDMED.html>).
- ✦ Forsman, Sarah (1997). **OLAP Council White Paper** (<http://www.olapcouncil.org/research/whtpapco.htm>).
- ✦ Frawley J. F., G. Piatetsky – Shapiro, G.J. Matheus (1991). Knowledge discovery in data bases: an overview. **A.I. Magazine**, 3 (13), 57-70.
- ✦ Grandy, Cheril (2000). **Breakthrough Performance for Star Schemas – A Data Warehousing Solutions. A White Paper**. Dynamic Information Systems Corporation (<http://www.disc.com/dwhpaper.html>).
- ✦ Gray, P. (1997). Data warehousing; review of “Building the Data Warehouse” (by W.H. Inmon). **Journal of Information Systems Management**, 14(1) (<http://dss.cba.uni.edu/books/review1.html>).
- ✦ Gray, P. H.J. Watson (1996). The new DSS: Data Warehouses, OLAP, MDD, and KDD. (<http://hsb.baylor.edu/ramsower/ais.ac.96/papers/graywats.htm>).
- ✦ Hammer, M.J. Champy (1993). **Reengineering the Corporation. A Manifesto for Business Revolution**. Harper Business, New York.
- ✦ IBM(1996). Data Warehousing Concepts for AS/400 (<http://www.as400.ibm.com/db2/dataware.htm>).

- ✦ Inmon, W.H. (1993). **Building Data Warehouse**. John Wiley & Sons. New York.
- ✦ Inmon, W.H. (1995). **What is a Data Warehouse?** (http://www.cait.wvstl.edu/cait/papers/prism/voll_no1/subject/home.html).
- ✦ Inmon, W.H., Claudia Imhoff, R. Sousa (1998). **Corporate Information Factory**. John Wiley & Sons, New York.
- ✦ Inmon, W.H., R.D. Hackathorn (1994). **Using the Data Warehouse**. John Wiley & Sons. New York.
- ✦ Keen, P.G.W. (1997). Let's focus on action not info. **Computerworld**. 17 November. (<http://www2.Computerworld.com/home/print9497.nsf/AII/SL46keen16F02>).
- ✦ Kimball, R. (1996). **The Data Warehouse Toolkit**. John Wiley & Sons. New York.
- ✦ Kimball, R. Merz (2000). **The Data Warehouse Toolkit: Building the Web-Enabled Data Warehouse**. John Wiley & Sons, New York.
- ✦ Mertens, P., J. Hagedorn, M. Fischer, N. Bissantz, M. Haase (1996). Towards active management systems. In P. Humphrey, L. Bannon, A. McCosh, P. Migliarese, J. Ch. Pomerol (Eds.). **Implementing Systems for Management Decisions; Concepts, Methods and Experience**. Chapman & Hall, p.305-325.
- ✦ OLAP Council (1997). **OLAP: On - Line Analytical Processing** (<http://dssresources.com/glossary/olaptrms.html>).
- ✦ Pendse, N. (2000 a) What is OLAP? An analysis of what the increasingly misused OLAP term is supposed to mean. **OLAP Report** (<http://www.olapreport.com/FASMLIHTML>).
- ✦ Pendse, N. (2000 b). The origins of today's OLAP products. **The OLAP Report**. <http://www.olapreport.com/origins.htm>.
- ✦ Pendse, N. (2000 c) What's in a name? The OLAP Report (<http://www.olapreport.com/name.html>).
- ✦ Poe, V. (1996). **Building a Data Warehouse for Decision Support**. Prentice Hall, Upper Saddle River, New Jersey.
- ✦ Power, D. (2000). **Decision Support Systems: Concepts and Resources**. Cedar Falls, I.A.: DSS Resources. COM, prepublication PDF version (<http://dssresources.com/dssbook>).
- ✦ Raden, N. (1995). Data, data everywhere. **Information Web**, Oct.30 (http://www.members.aol.com/nraden/iw_mct01.htm).
- ✦ Rowley, Jennifer (1998). What is information? **Information Services & Use**, 18(4), 243-255.
- ✦ Thomsen, E. (1997). **OLAP Solutions: Building Multidimensional Information Systems**. John Wiley & Sons, New York.
- ✦ Turban, E., J.E. Aronson (1998). **Decision Support Systems and Intelligent Systems**. (Fifth Edition). Prentice Hall, Upper Saddle River, New Jersey.
- ✦ Ulman, J.D. (1982). **Principles of Database Systems**. Computer Science Press, M.D.
- ✦ Watson, H.J., P. Gray (1997). **Decision Support in the Data Warehouse**. Prentice Hall, Englewood Cliffs, New Jersey.
- ✦ Witten, I.H., E. Frank, J. Gray (editors). (2000). **Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations**. Morgan Kaufmann Publishers.