SISTEME INFORMATICE PENTRU ASISTAREA DECIZIEI BAZATE PE SINTEZA DATELOR. DEPOZITE DE DATE (DATA WAREHOUSE)

Obiective:
- însusirea conceptelor cu privire la sistemele informatice pentru asistarea deciziei bazate pe analiza si sinteza datelor;
- utilizarea tehnologiilor moderne Data Warehousing si On-Line Analytical Processing (OLAP) pentru transformarea datelor în informatii de sinteză;
- însusirea tehnicilor si metodelor de prelucrare multidimensională a datelor.

Concepte cheie: depozite de date (Data Warehouse); prelucrare analitică on-line (OLAP); cubul OLAP; hipercub; bază de date multidimensională.

Modul în care datele sunt retransformate în informatii si apoi în cunostinte este de fapt un proces de valorificare a datelor care se realizează prin sintetizarea si analiza lor si în final prin interpretare. Procesul de sintetizare a datelor presupune centralizarea lor, având în vedere diverse criterii si este utilizat în crearea situatiilor de sinteză necesare informării managerilor ca support pentru luarea deciziilor.

Solutiile oferite de informatică pentru procesul de sintetizare a datelor sunt: programe specifice si dedicate; interogări care dau posibilitatea grupării datelor după criterii stabilite si oferă functii pentru domeniile astfel create; functiile de total si subtotal oferite de generatoarele de rapoarte care permit indicarea ierarhiilor criteriilor de grupare.

În ultimul timp, problema centralizării datelor a rămas aceeasi, însă volumul de date de explorat este imens, ceea ce duce la faptul că metodele clasice să devină ineficiente. De aceea câstigă tot mai mult teren tehnologii moderne ca Data Warehousing (depozitarea datelor) si OLAP (On-Line Analytical Processing) pe măsură ce suporturile soft devin suport de date pentru sistemele tranzactionale.

Tehnologiile de centralizare transformă datele în informatii de sinteză si analiza lor.

Analiza datelor presupune a găsi relatii între datele sintetizate cum ar fi: asocieri, corelatii structurale, cauzale sau functionale. O formă simplă de analiză a datelor este compararea datelor cu date similare, comparare care se face păstrând toate criteriile identice, doar unul singur având valori diferite.

Compararea se face între seturi de date comparabile, iar tehnologiile de comparatie sunt dotate cu tehnici de observare pentru semnalizarea tiparelor, corelatiilor, asocierilor prin similitudini sau sesizează abateri, exceptii. Informatica a venit în întâmpinarea acestor cerinte cu tehnicile de prezentare grafică care transformă informatia cantitativă de informatie calitativă. Au apărut si tehnici de observare analitică a datelor care au la bază teorii matematice prin care datele reale sunt comparate cu date teoretice produse de un model ipotetic.

Dezvoltarea tehnicilor de observare a dus la aparitia tehnicilor de observare automată bazate pe data-driven. Rezultatul unor astfel de tehnici se regăsesc într-un model cu caracter general. Tehnicile de observare analitică a

datelor se regăsesc într-o tehnologie modernă denumită Data Mining (în traducere liberă „Mineritul datelor”).

Rezultatul procesului de observare analitică este obtinerea unor tipare, corelatii si uneori modele din care se pot deduce tendinte sau se poate previziona cu o anumită probabilitate cum vor arăta datele pe o perioadă ulterioară. Modelul permite interpretarea datelor, ce reprezintă un proces cognitiv cu o apreciere generală a situatiei, si identifică probleme, oportunităti sau potentiale cauze de esec.

De remarcat este faptul că interpretarea datelor duce la aparitia de cunostinte noi care se vor cumula la cele deja existente. Instrumentele soft clasice pentru asistarea deciziei au avut ca principal scop asigurarea tehnicilor de analiză, optimizare si simulare, precum si reprezentarea grafică a rezultatelor.

Dintre aceste instrumente se amintesc procesoarele de tabele Lotus si Excel orientate pe volume mici de date, cele referitoare la sistemele de gestiune a bazelor de date Access, Visual Foxpro capabile să lucreze cu volume mari de date cu structură uniformă. Principalul dezavantaj al acestor instrumente clasice este că operează numai asupra acelor date care au o structură prestabilită si provin dintr-o sursă unică. Noile sisteme de asistare a deciziei folosesc tehnici speciale de comasare a datelor stocate în structuri neuniforme, pentru a utiliza informatii implicite care nu sunt specificate în datele existente. Suporturile software de asistare a deciziei oferă utilizatorilor o serie de facilităti cum ar fi: interogarea în limbaj natural, accesul la modele conceptuale, sisteme de gestiune OLAP si servicii de integrare cu alte suporturi soft.

Depozite de date (Data Warehouse)

Necesitatea depozitelor de date este dată de volumul imens de date acumulat în timp de companii. Integrarea acestor date istorice ale companiei într-o structură care să stea la baza luării deciziilor a devenit principala preocupare a noilor tehnologii.

Sistemele de asistare a deciziei care au la bază sinteza si analiza datelor realizează comasarea, sistematizarea, corelarea si gruparea datelor pentru a obtine informatii care să reliefeze factorii care influentează pozitiv sau negativ performanŃele companiei. Ca urmare a obtinerii unor astfel de informatii se poate adopta o strategie de ameliorare a factorilor cu influentă negativă. Obtinerea rezultatelor, sub formă de rapoarte care contin informatii utile factorilor de decizie sunt într-o formă accesibilă si sunt rezultatul tehnicilor speciale de explorare a masivelor de date. Aceste tehnici duc la evidentierea unor corelatii între date, pot face estimări si prognoze precum si să atragă atentia asupra unor disfunctii.

În sinteză tehnicile de exploatare a masivelor de date pot sugera solutii si pot contribui la luarea deciziilor într-o anume situatie

Datele, mai precis structurile de date care fac obiectul sistemelor informatice de asistare a deciziilor sunt denumite depozite de date (Data Warehouse).

Caracteristicile acestor structuri este faptul că ele pot înmagazina volume mari de date preluate din arhive si/sau din bazele de date ale aplicatiilor informatice specifice activitătii curente a întreprinderii (sunt volume de ordin

1012 terabytes). Exploatarea acestor volume uriase de date este asigurată de existenta unor motoare speciale care dau posibilitatea ca masivele să poată fi interogate, precum si existenta unor servicii speciale de analiză on-line a datelor (OLAP). Suporturile software sustin performantele prin transformarea datelor, corelarea si completarea lor, precum si prin crearea dictionarului de date, toate acestea asigurând accesul la structurile primare.

Datele sunt extrase din baze de date eterogene create de sistemele informatice deja existente în companie pe diversele platforme hard si soft. Se poate remarca faptul că datele sunt introduse nu la întâmplare, ci sub controlul unor aplicatii si al SGBD-ului. Acestea asigură prin serviciile de integritate, stocarea si lucrul în condiŃii de siguranŃă maximă. Datele care formează suportul pentru tranzacŃiile primare sunt apoi prelucrate pentru a se obŃine informatiile de sinteză necesare planificării si luării deciziilor si sunt tratate de instrumentele SGBD.

Deoarece exploatarea unui volum enorm de date, pentru a obtine diverse rapoarte, este asigurată de integritatea si coerenta bazei de date, reuniunea tuturor acestor date duce la exploatarea unui mare număr de tabele, la crearea unor multiple legături virtuale si tabele temporare. Acest volum mare de muncă conduce la principalul inconvenient al depozitelor de date si anume timpul mare necesar exploatării lor. Un alt inconvenient îl constituie si aglomerarea motorului bazei de date cu task-uri de centralizare care încetineste astfel tranzactiile curente.

Astfel a apărut necesitatea stocării datelor care sunt dedicate planificării si deciziilor strategice într-un sistem diferit de sistemul operational în asa fel încât functionarea celor două sisteme să se facă fără inconveniente. În depozitul de date se pot stoca atât arhive de date privind activitatea anterioară, cât si date referitoare la tranzactii ulterioare fără ca utilizatorul să poată interveni.

Datele se pot înmagazina pe domenii sau activităti specifice departamentelor unei organizatii în asa numitele magazii de date (Data Marts), separarea lor în acest fel ducând la cresterea performantelor în exploatare. Aceste depozite de date se construiesc de obicei cu tehnologii relationale. Depozitele de date sunt o concentrare de date care organizează, consolidează si centralizează datele din surse eterogene si care vor constitui baza procesărilor analitice atât de necesare proceselor de decizie. Depozitul de date se construieste progresiv adică el permite completări si dezvoltări ulterioare. Pentru a se asigura o calitate sporită a datelor acestea sunt supuse unui proces de curătire si transformare, mentionând si maniera de obtinere a unor date colectate pe baza celor existente, acest proces ducând la micsorarea timpului cerut pentru obtinerea unor rapoarte finale. În depozitele de date se face transformarea codurilor în date explicite, precum si integrarea datelor din nomenclatoare în datele referitoare la tranzactii. Acesta este numit si proces de denormalizare si este caracterizat de faptul că nu modifică integritatea datelor si grăbeste procesul de regăsire. Într-un depozit de date redundanta datelor este permisă.

Diferentele dintre depozitul de date si baza de date sunt următoarele
a. Datele continute de un sistem de prelucrare a tranzactiilor, OLTP (On-Line Transaction Processing) sunt de tip operational, iar datele continute

de un depozit de date sunt specifice asistării deciziilor, sunt date centralizate sau derivate din date operationale, nu se modifică în timp si sunt destinate utilizatorilor finali.
b. În cazul sistemelor tranzactionale, performantele se referă la integritate, confidentialitate, siguranță si timp de răspuns întrucât un număr mare de utilizatori introduc date în sistem, în timp ce în cazul SIAD (deci a depozitelor de date) numărul de utilizatori finali (manageri) este foarte mic.
Astfel si securitatea si siguranta în exploatare nu sunt supuse unor riscuri majore, procedurile de salvare si restaurare fiind mai putin utilizate decât în cazul sistemelor tranzactionale.
c. Datele procesate în sistemele tranzactionale sunt în seturi relativ mici, introduse recent si compact, astfel încât prelucrarea se face destul de rapid. În procesele decizionale, datele necesare acestora sunt în volum mare, stocate dispersat ceea ce duce la o prelucrare mai lentă.
d. Bazele de date construite pentru sisteme tranzactionale sunt proiectate si realizate pe baza unor cerinte cunoscute si certe, modificările care intervin datorită adaptării sistemului la schimbările intervenite reiau anumite faze ale ciclului de viată. Dar odată implementate ele functionează perioade lungi de timp fără modificări. În SIAD cerintele sunt cunoscute doar partial în momentul proiectării si realizării lor, ceea ce obligă depozitul de date să se adapteze din mers cerintelor. De aceea se observă că datele gestionate pentru sisteme tranzactionale sunt privite ca un întreg, pe când cele din depozitele de date sunt organizate pe sectiuni deoarece ele sunt organizate în functie de subiectul de analiză.
e. Sistemele tranzactionale reflectă de obicei fluxul datelor din activităti curente, pe când depozitele de date sunt orientate pe subiecte cum ar fi de exemplu: resurse, produse, clienti, furnizori.

Ciclul de viată al depozitelor de date Depozitul de date (Data Warehouse) este o colectie de date orientate pe subiecte, integrate, corelate în timp si non-volatile care sprijină decizia

Datele care fac obiectul unui depozit sunt integrate în acesta utilizând conventii pentru măsurători, atribute. Structura de care dispune depozitul de date prevede identificarea punctuală a datelor stocate si, mai ales, un acces rapid la ele.

Proiectarea structurii depozitului de date se face prin modelare multidimensională, structura implementându-se ca o bază de date care asigură stocarea unui volum mare de date si un acces rapid la ele, asa numitele baze de date client/server.

Popularea depozitelor de date se face prin preluare din sisteme tranzactionale, dar care vor fi supuse unor procese complexe de transformare care să corespundă structurii depozitului care a fost proiectat. După această etapă, depozitul va putea intra în exploatare pentru a obtine analize si rapoarte.

Etapele enumerate anterior (proiectare, populare, exploatare) sunt asistate de un soft specializat de la browsere si generatoare de rapoarte până la instrumente specifice Data Mining.

În exploatarea curentă a depozitului frecvent vor apare noi cerinte informationale care vor duce neapărat la extinderea structurii, la popularea cu

extensii cuprinzând date istorice, precum si la integrarea noilor date încorporate în aplicatii de analiză. Pe parcursul existentei sale, un depozit de date este incremental si ciclic. Modelarea conceptuală a depozitului de dat În etapa de conceptie a unui depozit de date se folosesc modele dimensionale care grupează datele din tabelele relationale în scheme de tip stea sau fulg de zăpadă. În aceste scheme pot fi regăsite date cantitative cum ar fi cantităti sau valori sau grupate după diverse alte criterii (pe client, pe produs, pe tipuri de servicii etc.). Datele cantitative din bazele de date dimensionale sunt de tip medii, număr de tranzactii, centralizări după anumite caracteristici, totaluri si reprezintă măsuri ale activitătii. Pe de altă parte, criteriile de agregare vor fi denumite dimensiuni. Măsurile identificate prin dimensiuni vor fi stocate întrun tabel relational care este denumit tabel de fapte, iar codurile utilizate sau asociate criteriilor de agregare sunt date de tabelele de tip nomeclator asociate fiind cu tabelele de fapte si în acest fel schema relatională va fi de tip stea. Dacă se reunesc mai multe scheme de tip stea care utilizează aceleasi nomenclatoare formează un model tip constelatie. Dacă nomenclatoarele se pot divide în subnomenclatoare atunci există o dependentă între acestea. De remarcat că pentru acelasi cod pot exista mai multe nomenclatoare alternative. Dacă se integrează aceste subdimensiuni si dimensiuni alternative, se creează o schemă sub formă de fulg de zăpadă.

Schemele de tip stea, fulg de nea sau constelatie sunt modele conceptuale multidimensionale ale depozitelor de date, având ca rol organizarea datelor pe subiecte necesare procesului de decizie. Schema este deschisă (ea se poate modifica pe tot parcursul vietii depozitului de date).

Modul de utilizare a depozitului de date

Depozitele de date contin structuri unice, integrate si cumulative necesare procesului de decizie. Administratorul depozitului de date are ca principală sarcină stabilirea accesului partajat al categoriilor de manageri prin asigurarea de parole si drepturi de acces. Datele din depozit sunt accesate selectiv de manageri în functie de necesitătile acestora. În acest fel se creează colectii specializate pe diverse domenii care se numesc magazii de date (Data Marts). Magaziile de date se pot utiliza si ca structuri intermediare pentru colectarea datelor din surse primare si al căror continut este descărcat periodic în depozitul de date. Depozitele de date pot lua nastere si printr-o stocare exhaustivă a datelor din sistemele tranzactionale în vederea aplicării tehnologiei Data Mining. Utilizarea tehnologiei Data Mining presupune că procesarea datelor se face fără interventia utilizatorilor, în background, iar rezultatele sunt păstrate pentru a fi consultate ulterior la cerere.

Mediul de depozitare al datelor

Mediul în care se construieste si se exploatează un depozit de date contine următoarele elemente: surse de date tranzactionale, instrumente de proiectare dezvoltare, instrument de extractie si transformare a datelor, sistemul de gestiune al bazei de date, instrumente de acces si analiză a datelor si instrumente de administrare

Toate componentele enumerate sunt integrate pe o platformă Microsoft în mediul de lucru Data Warehousing Framework ca si în cazul SQL Server 7.0.

Acest mediu de lucru oferă asistarea proiectării, implementării si administrării depozitelor de date pe durata vietii (existentei) acestuia. Se poate concluziona că Data Warehousing Framework oferă o arhitectură care se poate integra relativ simplu cu produse ce provin de pe alte platforme, asigură servicii de import-export cu validare si transformarea datelor, asigură metadate integrate pentru proiectarea depozitului si gestionează suportul, task-uri si evenimente. Pentru ca un depozit de date să poată fi procesat este necesară existenta unui set specializat de instrumente pentru: descrierea fizică si logică a surselor de date, a depozitelor sau a magaziei de date în care acestea urmează să fie încorporate; validarea, curătirea si transformarea datelor care urmează a fi stocate în depozitul de date; utilizatorii finali, instrumente care permit acestora accesul la datele stocate în depozitul respectiv. Astfel de instrumente sunt specializate pentru medii de dezvoltare a aplicaŃiilor, produse program specializate pe analiza datelor precum si pentru aplicatii personale (individuale).
Abordarea multidimensională a datelor stocate 絜depozite. Definirea si caracterizarea OLAP (On-Line Analytical Processing)
Dacă se analizează tehnologia relatională se observă că cea mai mare parte a problemelor tratate relational sunt în realitate multidimensionale. În modelul relaŃional problemele sunt tratate în tabele care au două dimensiuni: linie si coloană. Problemele reale, care în cea mai mare parte a lor sunt multidimensionale,nu impun limite stocării spatiale a datelor. Astfel, un SGBDR obisnuit nu poate face fată cerintelor de agregări de date, sintetizări, consolidări si proiectii multidimensionale. De aceea, a apărut necesitatea extinderii functionalitătii unui SGBDR prin adăugarea unor componente speciale care să permită modelare si analiză multidimensională (OLAP) si Data Mining.
Noua tehnologie OLAP permite utilizatorilor navigarea rapidă de la o dimensiune la alta si facilităti sporite de obtinere a celor mai detaliate informatii. Tehnologia OLAP se bazează pe 11 principii formulate de Ted Codd (1992). Acestea sunt:
1) abordarea conceptuală multdimensională a datelor;
2) asigurarea unei transparente sporite prin existenta unei arhitecturi deschise a sistemului;
3) accesibilitatea asigurată utilizatorului prin asistarea implicării acestuia în modalitătile tehnice de furnizare a datelor;
4) complexitatea dimensională a analizei oferă performante stabile;
5) utilizarea arhitecturii client-server, unde server-ul are ca scop omogenizarea datelor;
6) posibilitatea de a efectua aceleasi operatii asupra tuturor dimensiunilor si care poartă numele de prelucrare generică a dimensiunilor;
7) gestionarea dinamică a matricilor încrucisate prin facilitatea de a elimina combinariile dimensionale nule, pentru a nu încărca memoria calculatorului;
8) posibilitătile de acces simultan a mai multor utilizatori (multi-user) la aceeasi fază (etapă) de analiză;
9) operatii nerestrictive, ceea ce dă posibilitatea executării fără restrictii a calculelor pentru toate combinările de dimensiuni si niveluri ierarhice;
10) posibilitatea manipulării intuitive a datelor;
11) număr nelimitat de niveluri de agregare si de dimensiuni

OLAP este tehnologia de agregare a datelor stocate în depozite într-o manieră de abordare multidimensională cu facilităti referitoare la accesul la informatii a managerilor în mod interactiv si flexibil. Legătura dintre OLAP si depozitele de date este aceea că OLAP le completează prin transformarea volumului imens de date stocate si gestionat în depozite în informatii utile procesului de decizie. Cele 11 reguli ale lui Codd au fost apoi regrupate într-un test cu 5 reguli denumit FASMI (Fast Analysis Shared Multidimensional Information).
OLAP presupune existenta unor tehnici care permit de la o navigare si selectie simplă a datelor până la analiza detaliată si complexă. Aplicatiile care se rezolvă pe baza acestei tehnologii au la bază analiza rapidă a informatiei multidimensională dispersată în locatii multiple dar accesibile unui mare număr de utilizatori. Pentru utilizarea acestor facilităti, OLAP dispune de eficacitatea bazelor de date multidimensionale si de posibilitatea de a construi alternative pentru diverse probleme de decizie. OLAP presupune că analiza datelor (care pot fi de tip numeric sau statistic) poate fi predefinită de cel care creează aplicatia sau chiar de utilizatorul final.

OLAP se caracterizează prin: perspectiva multidimensională a datelor, capacitatea de calcul intensiv si orientare în timp (time intelligence)

Aspectul multidimensional al datelor este dat de posibilitatea de a integra multiplele aspecte care caracterizează activitatea unei întreprinderi si care sunt considerate din perspective multiple ca: timp, bani, produse. Fiecare dimensiune este definită în genere prin mai multe niveluri ca de exemplu: timpul este divizat în an, trimestre, luni, sezoane; produsul în: categorii, clasă. Conceptul de dimensiune este folosit ca înteles de aspect, dimensiunile fiind independente si cu unităti de măsură specifice dimensiunii respective.

Unitătile de măsură pot constitui criterii de agregare a datelor, iar nivelele unei dimensiuni formează ierarhia care la rândul ei poate constitui criteriu de agregare a datelor. Privite din punct de vedere multidimensional, datele sunt reprezentate în hipercuburi de date, prin extinderea cubului tridimensional la cel n-dimensional.

Pe acest tip de cub se pot efectua calcule prin aplicarea unor algoritmi complecsi asupra datelor structurate în acesta. Acestea implică posibilitatea de adresare multidimensională directă a cuburilor unitare si optimizarea timpului de răspuns. Caracteristica de orientare în timp (time intelligence) presupune flexibilitatea exploatării acestei dimensiuni care este necesară pentru comparatii si aprecieri de valoare în analizele economice. Această dimensiune este luată de obicei din calendarele tranzactiilor economice asa cum se află în bazele de date ale sistemului informatic al companiei. Se pot face astfel grupări pe dimensiuni ca: trimestre, luni, ani, sezoane. Se pot utiliza si dimensiuni speciale cum sunt: perioada curentă, perioada precedentă, aceeasi perioadă din anul..., care trebuie neapărat luate în considerare la proiectarea hipercubului. Bazele de date multidimensionale folosite de OLAP sunt suprapuse depozitelor de date si stochează straturi de date agregate pe diferite criterii ierarhice. De asemenea, aceste baze de date multidimensionale conŃin si date statistice pentru fiecare nivel de agregare.
Modelarea dimensională – cuburi OLAP

Modelarea dimensională presupune conceptualizarea si reprezentarea aspectelor măsurabile ale activitătii studiate în interdependentă cu contextul în care acesta se desfăsoară, aspect identificat prin parametrii activitătii. Legătura dintre valorile înregistrate ale activitătii (valori vânzări, cheltuieli comune, costul produselor) si contextul de desfăsurare al acesteia formează baza numeroaselor rapoarte de sinteză care sunt produse de sistemele tranzacŃionale. Prin modelare dimensională se oferă un model conceptual comun acestor rapoarte si agregarea lor într-o structură uniformă si flexibilă. Totodată se păstrează si legătura cu sursele initiale de date, deci posibilitatea de descompunere a datelor centralizate pe niveluri din ce în ce mai mici până se ajunge la setul de tranzactii initiale (drill-down).
Cubul OLAP se consideră a fi element structural pentru datele din procesul on-line. Acesta este o structură multidimensională, un hipercub prin care se modelează complexul de activităti pe o perioadă îndelungată de timp. Acest tip de modelare este caracterizat de câteva concepte de bază:
• Cuantificarea activitătii (aspectul cantitativ) care se face prin utilizarea unitătilor de măsură clasice ca de exemplu: m, m3, kg, unităti monetare. Măsuri cantitative sunt: volum vânzări, volum salarii, cost materiale, cost produs etc.
• Dimensiunile activitătii sunt de fapt parametrii activitătii măsurate ca de exemplu: zi, lună, trimestru, client sau grupă de clienti. Dimensiunile sunt de obicei de natură diferită si răspund la întrebări de tipul: Unde? Când? Cu ce? etc.
• Faptele sunt colectii ale cuantificării activitătii precum si dimensiunile care identifică modul în care acestea s-au desfăsurat. Sursa de existentă a faptelor este constituită din înregistrările stocate în tabelele de tranzactie ale aplicatiilor operationale care sustin activitatea respectivă. Se pot folosi si dimensiuni scenarii care pot stoca în tabelele de fapte si măsuri imaginare alături de cele reale, pentru ca utilizatorul să poată stoca valori estimate pentru o măsură.

În bazele de date tranzactionale, dimensiunile sunt de fapt câmpuri care contin caracteristicile unei tranzactii adică datele de identificare ale tranzacŃiilor care sunt de obicei chei externe care fac legătura cu nomenclatoarele care le explicitează.

Ca atare, se poate afirma că dimensiunile se materializează în setul de valori posibile care formează domeniul caracteristicii respective, valori care poartă numele de membrii dimensiunii.

O altă caracteristică a dimensiunii este a ceea că poate avea multipli adică sunt grupe de valori ale dimensiunii cu o caracteristică comună. Grupele pot fi identificate prin atribute care se află în nomenclatoare si pot lua aceeasi valoare pentru mai multe valori ale cheii primare. Multiplii unei dimensiuni nu trebuie să fie neapărat de aceeasi natură cu dimensiunea primară, aceasta putând avea mai multe tipuri de multipli în functie de caracteristicile luate în considerare. Se poate afirma că dimensiunile împreună cu multiplii lor formează structuri arborescente care sunt recunoscute de OLAP ca fiind ierarhii. Ierarhiile pot fi regulate, adică toate ramurile au acelasi număr de ramificatii sau neregulate dacă pe anumite ramuri lipseste un nivel de semnificatie. La rădăcina arborelui se află o caracteristică cu aceeasi valoare pentru toti membrii dimensiunii de bază. Acest tip de caracteristică este una implicită ca, de exemplu, unitatea care are ca activitate cea

analizată sau „all”. Frunzele arborelui formează membrii dimensiunii initiale, iar dimensiunile intermediare pot fi pe mai multe nivele. Dacă arborele este neregulat, pentru a uniformiza ierarhia se poate introduce un membru de tip „alte”.

În acest fel se constată că centralizările pe nivelul respectiv nu vor fi de 100% din valoarea centralizată pe nivelul cel mai de jos. Atributele care definesc ierarhia sunt atribute derivate din atributul care defineste dimensiunea acŃiunilor măsurate, prin referire la nomenclatoare sau prin clasificări ale valorilor pe care le poate lua atributul respectiv. De exemplu, furnizorii se pot clasifica în furnizori stabili dacă compania face tranzactii cu ei de mai mult de 4 ani, furnizori noi dacă au vechime cuprinsă între 1 si 4 ani si furnizori volatili sau ocazionali dacă în câmpul respectiv din Furnizori nu este completat nimic. Din acest exemplu se observă că asemenea clasificări conduc la obŃinerea unor atribute derivate prin calcul din caracteristicile aflate în nomenclatoare. În acest fel se vor obtine seturi de membri calculati ai dimensiunii. Dimensiunile ierarhizabile se constituie în ierarhii alternative. Nivelele ierarhiilor sunt văzute ca nivel de agregare pentru valorile stocate în tabele de fapte. Membrii dimensiunilor identifică măsura activitătii stocată în tabelul de fapte. Dacă unui fapt îi sunt asociate mai multe dimensiuni, identificarea unică a acestuia va necesita valori precise pentru fiecare dimensiune. Ca urmare, din tabelele de fapte sunt selectate mai multe înregistrări, adică toate valorile posibile asociate dimensiunilor nespecificate.

Pentru dezvoltarea unui depozit de date, modelarea datelor are un rol important deoarece permite vizualizarea structurii înainte ca ea să fie construită.

Modelul multidimensional reprezentat prin el va fi prezentat desfăsurat în sectiuni sau în proiectii tridimensionale.

Sectiunea unui hipercub este definită ca o sectiune din cub dată prin coordonatele sale. Proiectia este definită ca o sectiune care centralizează datele de pe toate dimensiunile suprimate.

Vizualizarea on-line se face de fapt tot în sectiuni sau proiectii tridimensionale. Datele din celule sunt prezentate numai în sectiuni sau proiectii transversale bidimensionale. Hipercubul ar putea fi imaginat ca un set de tabele pivot grupate pe dimensiunea cerută. Pentru procesul de modelare, hipercubul se poate prezenta în formă tabelară în care măsurile sunt evidentiate pe coloane, iar liniile reprezintă combinatiile de dimensiuni. De asemenea, în plan fizic, hipercubul poate fi stocat într-un tabel cu coloane multiple în care se stochează măsurile si cu identificatori pe rânduri. Identificatorii de rânduri sunt de fapt chei formate din toate combinatiile posibile de valori ale dimensiunilor.

Utilizarea indecsilor pentru acces rapid nu are prea mare eficientă întrucât cheia este compusă din mai multe caracteristici, iar câmpurile de valoare sunt putine si numerice, astfel că tabelul de indecsi este aproape de aceeasi dimensiune cu tabelul initial. De aceea, se utilizează tabelul bitmap pentru un acces direct rapid. Datele modelate ca hipercuburi formeaza baze de date multidimensionale.

Baze de date multidimensionale

Baza de date multidimensională este formată din două structuri:

structura datelor în care se stochează măsurile activitătilor preluate din tabela

de fapte a depozitului de date. Datele vor fi prezentate utilizatorului în celulele tabelelor pivot; structura metadatelor care este formată din totalitatea dimensiunilor si membrilor acestora precum si din structurile ierarhice ale dimensiunilor. Utilizatorul poate vizualiza această structură ca nume de coloane si linii care reprezintă informatiile de pe axele cuburilor. Numerotarea nivelurilor începe de la rădăcină (nivel 0) către frunze
(unde va apare nivelul maxim). Ierarhiile posedă propriile lor seturi de niveluri, chiar dacă unele ramuri sunt comune. De exemplu: ierarhia Calendar este
formată din nivelele (0-5): Timp, An, Semestru, Trimestru, Lună, Dată calendaristică, ierarhia Anotimp este formată din nivelele (0-4): Timp, An,
Sezon, Lună, Dată calendaristică, iar ierarhia Anotimp este formată din nivelele (0-3): Timp, Săptămână, Zi, Dată calendaristică. Pe fiecare nivel se stochează membrii dimensiunilor respective. Rădăcina care se observă că este comună (Timp) este nivelul de agregare maxim având ca unic membru implicit „all”. Orice nod în arbore este un membru al unei subdimensiuni. Nodurile subordonate unui nod formează un set, iar orice membru al unui set are un număr de ordine începând cu 0. De asemenea, orice membru poate avea proprietăŃi ca de exemplu unele zile sunt sărbători legale, unii ani sunt bisecti. Exemplul prezentat presupune o structură strict arborescentă întrucât fiecare membru al unei dimensiuni are submembri distincti, chiar dacă acestia au aceleasi valori. De exemplu, fiecare an are setul lui de luni, fiecare săptămână are setul ei de zile. Ca mod de identificare, membrii vor fi calificati cu numele membrului de pe nivelul precedent căruia acesta i se subordonează: 2000-feb, 2001-feb. Tipul acesta de dimensiuni care au membri ce se repetă se pot crea si ulterior prin combinarea a două nivele din ierarhie sau din ierarhii diferite pentru a crea un nivel nou, virtual.

Pentru a se putea naviga pe o structură arborescentă, sistemele de gestiune pun la dispozitie operatori ierarhici. De exemplu, pentru exploatarea datelor, sistemele de gestiune oferă operatori pe hipercuburi. Fizic, datele sunt stocate într-un fisier cu acces direct pe baza adresei fizice absolute sau relative a înregistrării obtinute prin exploatarea tabelelor bitmap obtinute în urma creării structurii de date. Aceste tabele sunt puntea de legătură dintre structura de date si structura de metadate. Iată cum se face această legătură: se stie că pentru fiecare membru al fiecărei dimensiuni există o coloană (1 bit) în tabele bitmap pentru fiecare înregistrare există un rând în acelasi tabel în care se stochează 1 în dreptul bitilor asociati membrilor dimensiunii existente în înregistrare. Datorită acestui procedeu, câmpul respectiv nu trebuie stocat în înregistrare, iar structura datelor este redusă la un minim necesar. Din tabelul de măsuri se vor putea selecta acele înregistrări care au un bit 1 în pozitia corespunzătoare bitilor 1 din mască. Un inconvenient al tabelelor bitmap este acela că ele sunt greu de obtinut, iar aparitia unor noi membri sunt greu de inserat în pozitia corespunzătoare. Procesul de refacere a unui tabel bitmap este mare consumator
de timp având în vedere că tabelul de fapte din depozit (care se va transforma în baza multidimensională) poate avea un număr imens de înregistrări. Masca de interogare se obtine prin exploatarea structurii ierarhice a metadatelor de unde se pot extrage seturi de membri pentru dimensiunile

desemnate prin specificatorii de axe. Adresarea tabelului de măsuri se face în mod direct pe baza unui set de adrese de înregistrări care se suprapun cu tiparul măstii. Din tabel se preiau în această manieră valorile care se centralizează pentru celula cubului cu dimensiunile sale.

Se poate afirma că structura metadatelor este de tip ierarhic, fiecare dimensiune fiind stocată într-o structură arborescentă cu o singură rădăcină (all) si cu o multitudine de ramuri care pot contine frunze comune (ierarhii alternative). Orice nivel al unei ierarhii poartă un nume si contine un set de membri. De altfel si ierarhiile alternative poartă un nume pentru a putea fi distinse. Structura în care sunt stocate datele este o structură cu acces direct prin tabele bitmap exploatate prin măsti.

Operatii OLAP asupra hipercubului

Un hipercub este proiectat astfel încât el să aibă în vedere nivelul de detaliu necesar în procesul de analiză. Nivelul de detaliu (granularitatea) reprezintă numărul de membri ai unei dimensiuni. Datele pot fi vizualizate printr-o selectie în hipercub pe baza unui criteriu ierarhic care ar putea fi de exemplu structura organizatională pe care o conduce un anumit manager. Dacă de la pornire, granularitatea este prea mare, datele vor fi mult prea centralizate si nu se va putea face decât o analiză grosieră. Ajustarea nivelului de granularitate este realizată de OLAP prin exploatarea ierarhiilor dimensiunilor prin comasări si descompuneri ale măsurilor prin proceduri care poartă numele de drill-up si drill-down. Prin intermediul acestor proceduri se face o deplasare a proiectiei cubului în sus sau jos pe nivelele ierarhice ale fiecărei dimensiuni (zoom in; zoom out), executând de fiecare dată centralizări ale măsurilor stocate la cea mai mică granularitate după criterii ierarhice stabilite în prealabil.

Este stabilit un nivel de granularitate initial sub care nu se poate coborî. Din acest motiv este important ca dimensiunile de bază să fie cât mai rafinate sau să se creeze Data Marts, unde hipercuburile sunt proiectate la nivelul de detaliu stabilit de managementul operational. Pentru managementul superior se va construi un depozit cu hipercuburi centralizatoare cu granularitate mare. Prin drill-down se obtin detalii, iar prin drill-up se obtin date sintetice.

Un alt grup de operatii oferit de OLAP este sectionarea (slicing) si defalcarea (dicing). Prin sectionare, se creează posibilitatea selectării prin vizualizare doar pentru un membru al unei dimensiuni, adică un plan din cubul tridimensional. Sectiunea astfel obtinută va apare ca un tabel pilot cu valorile dimensiunilor pe laturi si cu specificarea valorii alese pentru dimensiunea suprimată. Defalcarea (dicing) este operatia de proiectare a unei dimensiuni pe o alta. De obicei o dimensiune din primul plan este combinată cu o altă dimensiune din adâncime. Acest proces se mai numeste imbricarea dimensiunilor.

Dimensiunile unui cub pot fi private sau pot fi utilizate în comun si de alte cuburi (ele provin din depozitele cu schema de tip constelatie). Proiectarea structurilor depozitelor de date si a cuburilor OLAP este un proces ce se desfăsoară continuu pe tot parcursul existentei (vietii) aplicatiei, dimensiunile cuburilor fiind în strânsă dependentă cu detaliile activitătii structurate.

Aplicatiile construite cu tehnologia OLAP îsi găsesc locul în multiplele domenii ale activitătii întreprinderilor, de la finante, bănci, marketing până la productie si vânzări. De exemplu, activitatea de productie poate fi sustinută de aplicatii OLAP cum sunt: planificarea operatiilor, controlul calitătii produselor, analiza rebuturilor, analiza optimizării raportului dintre cost-beneficii.
OLAP, utilizând tehnici inteligente de optimizare, beneficiază de avantajul timpului de răspuns mic.

Crearea aplicatiilor OLAP in Microsoft SQL Server
Pentru realizarea unei aplicatii OLAP sunt necesare următoarele etape:
1. Crearea bazei de date relationale (tranzactionale) care va contine datele curente ale organizatiei rezultate din tranzactii.
2. Crearea bazei de date multidimensionale, a cuburilor si tabelelor de fapte care preiau datele din baza de date relațională. Datele sunt extrase, transformate si încărcate în tabelele de fapte din tabelele relationale.
3. Crearea interfetei aplicatiei într-un mediu de programare visual – Visual Basic.
Crearea bazei de date tranzactionale în Microsoft SQL Server
Datele stocate în cadrul organizatiei sunt importate într-o nouă bază de date tranzactională ce stă la baza construirii cuburilor de date. Datele sunt organizate în tabele care corespund dimensiunilor, ierarhiilor si tabelelor de fapte ale cuburilor multidimensionale.

Ca exemplu, se va crea o aplicatie destinată analizei rezultatelor financiare ale unei bănci comerciale. Se vor analiza volumul depozitelor si volumul creditelor în functie de următoarele dimensiuni: agentie, durată, garantie, monedă, sector de activitate, timp, tip depozit, tip client (pers fizică sau juridică). Tabelul de fapte construit va contine două măsuri: volumul depozitelor si volumul creditelor.
Crearea tabelelor în Microsoft SQL Server se realizează prin utilizarea de scripturi, asa cum se prezintă în exemplul de mai jos:
• create table agentie (Agentie varchar(20), Zona varchar(10), Tara varchar(10));
• create table voldepozite (Agentie varchar(20), Durata varchar(20), Moneda varchar(20), Tipjur varchar(30), Timp varchar (20), Tipdepozit varchar(30), voldep numeric);
• create table volcredite (Agentie varchar(20), Durata varchar(20), Moneda varchar(20), Tipjur varchar(30), Timp varchar(20), Garantiecredit varchar(20), Sectoractivitate varchar(20), volcredite numeric).

Crearea bazei de date multidimensionale în SQL Server Produsul Microsoft SQL Server oferă suportul si instrumentele necesare dezvoltării sistemelor OLAP prin setul de aplicatii SQL OLAP Services, iar gestiunea bazei de date multidimensionale este realizată de serverul OLAP.

Se creează o nouă bază de date multidimensională care va contine cuburile de date prin intermediul meniului New Database. Se creează cuburile cu ajutorul asistentului Cube Wizard.

Cuburile OLAP utilizează datele stocate în tabelele bazei de date tranzactionale. Din acest motiv trebuie configurată conexiunea dintre cubul OLAP si baza de date tranzactională din care vor fi preluate datele. Conexiunea cu serverul de baze de date Microsoft SQL Server se realizează cu ajutorul

Microsoft OLE DB Provider for SQL Server. După stabilirea serverului tranzactional se selectează si baza de date tranzactională din care se importă datele.

Aplicatia OLAP contine două cuburi pe care se vor analiza cele două tipuri de operatiuni bancare: operatiunile pasive (constituirea de depozite) si operatiunile active (acordarea de credite). Pentru fiecare cub se defineste o schemă care contine în centru tabelul de fapte legat de dimensiunile corespunzătoare fiecărei activităti analizate după cum urmează :

1. Cubul Depozite – urmăreste analiza depozitelor si a dobânzilor pasive rezultate din activitatea curentă a băncii.
• Dimensiunile identificate în cadrul acestei scheme sunt: Agentie, Durata, Moneda, Timp, Tip juridic, TipDepozit.
• Tabelul de fapte al modelului este VolDepozite având ca măsură volumul depozitelor constituite (voldep).
2. Cubul Credite – urmăreste analiza creditelor si a dobânzilor active rezultate din activitatea curentă a băncii. În cadrul acestei scheme se identifică dimensiuni comune cu schema operatiunilor pasive. Acestea sunt:
• Dimensiunile identificate în cadrul acestei scheme sunt: Agentie, Durata, Moneda, Timp, Tip juridic, Garantie credit, Sector activitate.
• Tabelul de fapte al modelului este VolCredite ce contine măsura VolCred (volumul creditelor).

Crearea cuburilor se poate face independent (în fereastra Cube Editor) sau în mod asistat (prin optiunea Cube Wizard).
În ambele cazuri trebuie parcursi următorii pasi:
1. Alegerea tabelelor din baza de date relatională din care se importă datele. După configurarea conexiunii dintre serverul OLAP si baza de date tranzactională se pot vizualiza tabelele acesteia. O parte din aceste tabele contin date necesare dimensiunilor din cuburi, iar altele furnizează date în tabelele de fapte.
2. Stabilirea tabelului de fapte. În cazul cubului Depozite, tabelul de fapte Vol Depozite importă datele din tabelul voldepozite din baza de date relatională, iar în cazul cubului Credite tabelul de fapte VolCredite importă datele din tabelul volcredite din baza de date relatională.
3. Crearea si configurarea dimensiunilor. Se aleg tabelele din baza de date relatională din care se vor importa datele în tabelele dimensiuni.
4. Pentru fiecare dimensiune se stabilesc nivelurile ierarhice. De exemplu, pentru dimensiunea Timp, prezentă în ambele cuburi nivelele ierarhice sunt: Luna, Semestru, An.
5. Alegerea dimensiunilor care intră în configuratia cubului. După crearea tuturor dimensiunilor, se aleg numai cele care participă la configuratia cubului.
6. Stabilirea optiunilor de stocare si procesare a datelor din cubul OLAP.