Multidimensional data analysis using OLAP Technology (1)

Asist. Gianina RIZESCU
Catedra de Contabilitate şi Informatică Economică, Universitatea "Dunărea de Jos" Galaţi

In this paper we present the main steps in creating an application using OLAP technology. The main goals of this application are: the quality analysis of the teaching process through the results obtained by the students of a university and the structure and dynamic analysis of the students in a university ("Dunărea de Jos" of Galaţi was taken as example). Thus, after a brief introduction, we will present de general architecture of the application followed by the description of the analysis, design, and populating stages of the data warehouse and also by the multidimensional data analysis using decision cubes.
Keywords*: OLAP, OLTP, data warehouse, data mart, decision cub.*

Introducere

Ca exemplificare practică a etapelor parcurse în realizarea oricărei aplicaţii utilizând tehnologia OLAP, s-a proiectat o aplicaţie informatică pentru universitatea „Dunărea de Jos" care îşi propune să realizeze două funcţii majore:

- ✓ Analiza calităţii procesului de învăţământ prin intermediul rezultatelor obţinute de studenţii universităţii
- ✓ Analiza structurii şi dinamicii studenţilor din universitate

Necesitatea dezvoltării unei astfel de aplicaţii la nivel de universitate a apărut din următoarele considerente:

- ✓ Necesitatea obţinerii unei imagini globale, la nivel de universitate, asupra structurii şi dinamicii studenţilor. Această necesitate a apărut datorită concurenţei acerbe care s-a declanşat între universităţi în ultimii ani, a creşterii numărului universităţilor particulare pe de o parte, iar pe de altă parte scăderii de la an la an a numărului de studenţi. Acest lucru face necesară obţinerea de informaţii suplimentare, integrate la nivel de universitate pentru elaborarea unor politici de marketing în atragerea unui număr cât mai mare de candidaţi.
- ✓ Necesitatea obţinerii unei imagini globale şi integrate asupra calităţii procesului de învăţământ pentru a identifica şi înlătura factorii care influenţează în mod negativ acest proces pe de o parte, iar pe de altă parte pentru a găsi mijloacele necesare creşterii calităţii acestui proces
- ✓ Fiecare facultate utilizează propriul sistem de evidenţa a studenţilor şi a rezultatelor acestora. Apare astfel necesitatea creării unei structuri de date integrate şi standardizate la nivel de universitate pentru a se facilita analiza comparativă a datelor şi pentru a se înlătura informaţiile inconsistente şi deseori contradictorii
- ✓ Oferirea unui instrument flexibil de analiză a datelor factorilor de decizie care să le ofere acestora independenţă fată de sistemele tranzacţionale a căror rapoarte şi liste nu satisfac varietatea necesităţilor de informare.

Arhitectura generală a aplicaţiei

Arhitectura generală a aplicaţiei este prezentată în figura 1.

Referitor la această arhitectură generală se impun următoarele observaţii:

- ✓ Depozitul de date este de fapt un datamart (magazie, piaţă de date) datorită orientării sale pe un singur subiect. În continuare însă vom folosi noţiunea de depozit de date pentru uşurinţă exprimării şi datorită faptului că un datamart este tot un depozit de date de mai mici dimensiuni şi orientat pe subiect.
- ✓ Pentru crearea modulului de populare a depozitului de date s-a luat în considerare doar baza de date a unei singure facultăţi deoarece conceperea unui sistem de extragere şi de populare a depozitului de date cu date din sistemele tranzacţionale ale tuturor facultăţilor implica alte aspecte care nu au fost tratate şi o complexitatea greu de acoperit la momentul respectiv.

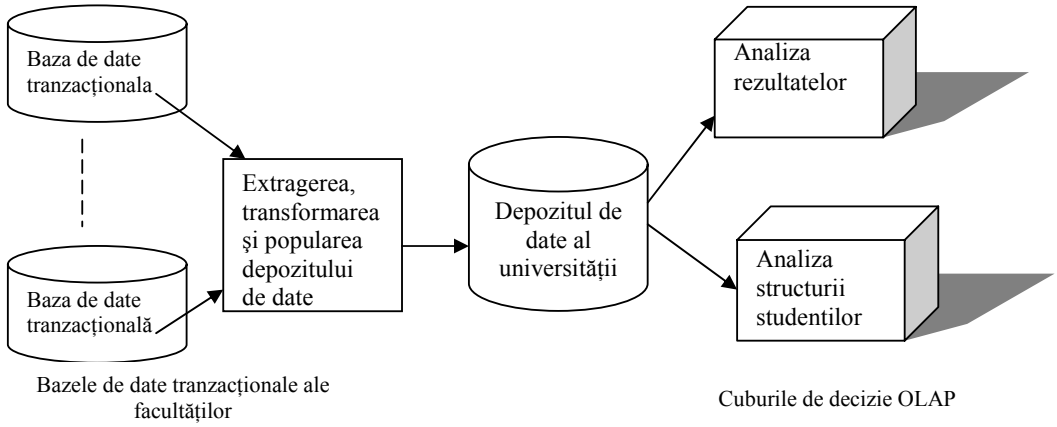

Fig. 1. Arhitectura generală a sistemului

Pentru realizarea aplicaţiei s-a utilizat tehnologia oferită de Microsoft, SQL Server 7.0 din mai multe considerente:

✓ Oferă o soluţie integrată şi completă de dezvoltare a sistemelor de decizie bazate pe OLAP

✓ Oferă o soluţie mai ieftină deoarece nu se plăteşte licenţa suplimentară pentru serverul OLAP acesta fiind furnizat odată cu Microsoft SQL Server 7.0

✓ Face posibilă implementarea depozitelor de date de către firmele mici şi mijlocii care nu-şi permit costuri suplimentare pentru alte sisteme de gestiune a bazelor de date

✓ SQL Server este mult mai uşor de utilizat decât majoritatea sistemelor de gestiune a bazelor de date de pe piaţă

✓ Oferă instrumente vizuale şi intuitive de dezvoltare a aplicaţiilor care permit crearea mult mai rapidă a lor.

In procesul de dezvoltare a depozitului de date au fost parcurse etapele care alcătuiesc ciclul de viaţă a oricărui depozit de date şi anume: Analiză, Proiectare, Populare.[1]

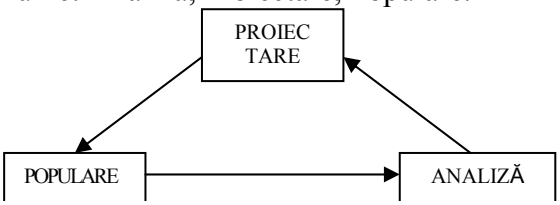

Fig. 2. Ciclul de viaţa al depozitelor de date

I. Analiza în vederea creării depozitului de date

Punctul de plecare în implementarea oricărei baze de date este analiza, în timpul căreia trebuie să se răspundă la o serie de întrebări. Care sunt entităţile implicate? Care sunt relaţiile între entităţi? Ce atribute trebuie asignate fiecărei entităţi? Ce funcţii trebuie îndeplinite şi cum afectează aceste funcţii entităţile şi atributele identificate?

Proiectarea logică sau faza de analiză a unui proiect în mod tipic are două puncte de plecare, care în general se suprapun:

✓ Procesele sau funcţiile. Ce funcţii va trebui să îndeplinească viitorul sistem?

✓ Datele. Ce date sunt necesare pentru a sprijini realizarea funcţiilor afacerii?

Toate acestea se determină plecând bineînţeles de la cerinţele utilizatorilor. Aceste lucruri sunt valabile atât pentru sistemele OLTP cât şi pentru sistemele OLAP. Diferenţa constă în categoria de utilizatori pe care o are în vedere: dacă în sistemele tranzacţionale utilizatorii ţintă sunt cei executivi (în principal), în sistemele analitice ţinta o reprezintă utilizatorii factori de decizie.

I.1 Analiza cerinţelor şi a datelor necesare

Înaintea începerii dezvoltării depozitului de date, domeniul afacerii trebuie să fi fost înţeles pe deplin. Scopul final al unui astfel de sistem este de a oferi utilizatorilor libertatea de a manipula datele liber, fără constrângeri externe. De aceea analiza este necesară pentru a ne asigura că în depozitul de date ajung toate datele necesare – la nivelul adecvat de detaliu – astfel încât acesta, depozitul de date, să fie capabil să răspundă la cât mai multe dintre interogările utilizatorilor. Deci, în faza de analiză trebuie identificate care sunt cele mai frecvente şi cele mai importante interogări la care depozitul de date trebuie să răs-

[1] Dorin Zaharie, & co, „Sisteme informatice pentru asistarea deciziei”, Ed. DualTech, Bucureşti 2001

pundă.
Astfel, pentru analiza calității procesului de învățământ s-a considerat că trebuie oferite răspunsuri pentru următoarele întrebări:
- ✓ Care sunt rezultatele studenților pe discipline sau grupuri de discipline în diferite perioade?
- ✓ Care sunt rezultatele comparative ale studenților pe specializări, secții, facultăți?
- ✓ Care sunt rezultatele comparative ale studenților pe orașe, județe, țări?
- ✓ Care sunt rezultatele comparative ale studenților pe sexe, stare civilă?
- ✓ Care sunt rezultatele comparative ale studenților în funcție de regimul de școlarizare: buget, taxă, zi, IDD(Învățământ Deschis la Distanță), IFR(Învățământ cu Frecvență Redusă)
- ✓ Rezultatele comparative ale studenților în orice combinație dintre criteriile definite mai sus.

Pentru analiza structurii și dinamicii studenților din universitate, s-a considerat necesar sa se răspundă la următoarele întrebări:
- ✓ Care este structura și dinamica studenților pe orașe, județe, țări?
- ✓ Care este structura și dinamica studenților pe sexe, stare civilă?
- ✓ Care este structura și dinamica studenților pe specializări, secții, facultăți?
- ✓ Care este structura și dinamica studenților în funcție de regimul de școlarizare : buget, taxă, zi, IDD(Învățământ Deschis la Distanță), IFR(Învățământ cu Frecvență Redusă)?
- ✓ Structura și dinamica studenților pe orice combinație dintre criteriile de mai sus.

Identificarea datelor necesare nu este însă suficientă. În afară de această identificare se urmărește elaborarea unei structuri unice și integrate a datelor: tipul de date, modalitatea de exprimare a unor date, stabilirea modului de exprimare a timpului în analiza datelor știind că orice depozit de date are o dimensiune timp care se exprimă diferit în funcție de semnificația timpului pentru sistemul respectiv. Toate acestea se realizează prin utilizarea de convenții consistente în privința numelor, măsurătorilor, atributelor și semanticii.

II. Proiectarea depozitului de date

Structura depozitului de date are în vedere identificarea precisă a datelor stocate și accesul rapid la ele. Pentru aceste deziderate, masa de informații care se va stoca în depozit trebuie astfel organizată încât să reflecte atât datele importante cât și contextul lor. Modelarea dimensională oferă suportul necesar pentru proiectarea structurii depozitului de date. Structura se implementează sub forma unei baze de date care să asigure atât stocarea unui volum imens de date cât si accesul rapid la ele.
Cel mai popular model pentru depozitele de date este modelul multidimensional. Acesta poate fi în formă de stea (star schema), de fulg de zăpadă (snow-flake schema) sau de constelație (constellation schema), în care se regăsesc datele cantitative (cantități valori) din tabelele de tranzacții agregate în principal pe unitatea de timp (zi, lună etc.) și apoi după alte criterii: (pe student, pe disciplină, pe specializare, pe facultăți etc.) și întotdeauna data calendaristică, primul criteriu de agregare. Aceste date cantitative centralizate sunt *măsuri ale activității* iar criteriile de agregare sunt denumite *dimensiuni*. Măsurile identificate prin dimensiuni sunt stocate într-o tabelă relațională denumită *tabela de fapte.* Codurile criteriilor de agregare sunt explicitate în tabele de tip nomenclator asociate tabelei de fapte (tabelele de dimensiuni), schema relațională căpătând forma de stea. Mai multe asemenea scheme de tip stea care folosesc aceleași nomenclatoare formează un model de tip *constelație* iar dacă dimensiunile se pot divide în subdimensiuni, atunci nomenclatoarele pot avea la rândul lor alte nomenclatoare formând o schemă ce are forma unui *fulg de zăpadă*.

II.1 Identificarea și definirea dimensiunilor

Dimensiunile determină modalitățile în care datele sunt organizate – cu alte cuvinte, cum datele pot și trebuie să fie filtrate, organizate, și grupate. Pentru determinarea dimensiunilor s-au avut în vedere următoarele:
- ✓ Caracteristicile dimensiunii;
- ✓ Caracteristicile tabelei de dimensiuni;

Caracteristicile dimensiunii.
Tabelele de dimensiuni trebuie proiectate ținând cont în permanență de utilizatorul final

şi de necesităţile sale de analiză. Dimensiunile sunt constituite de tabele relaţionale şi ca orice tabelă relaţională, sunt definite prin cheie primară şi atribute. Acestea trebuie să conţină: atribute strâns corelate, denumiri clare, elemente de nume şi adresă atomice, chei „surogat". În loc de a se utiliza valorile cheilor din sistemul sursă este recomandat să se utilizeze aceste chei surogat, obţinute prin generarea unor valori unice pentru identificarea entităţilor din tabelele de dimensiuni. Aceste chei sunt valori numerice, secvenţiale care sunt gestionate de sistemul de gestiune a bazei de date. Nu sunt atribuite de către utilizator. Folosirea acestor chei oferă următoarele avantaje:

- ✓ Oferă independenţă fată de sistemul sursă, înlăturând următoarele probleme:
  - modificarea aplicaţiilor poate cauza şi modificarea structurii cheilor
  - aplicaţiile sursă pot reutiliza cheile
- ✓ Oferă o indexare mai eficientă

Caracteristicile tabelei de dimensiuni.

Tabelele de dimensiuni au caracteristici comune, care le fac mult mai uşor de definit. La definirea unei tabele de dimensiuni există câteva reguli ce trebuie avute în vedere: Orice tabelă de dimensiuni are o cheie primară care este determinată unic pentru fiecare tabelă. Nu se utilizează chei compuse şi nu se preiau cheile din aplicaţiile sursă; Între o tabelă de dimensiuni şi tabela de fapte există o relaţie unu la mai mulţi. Unei înregistrări din tabela de dimensiuni îi corespunde mai multe înregistrări in tabela de fapte; Conţine cel puţin un atribut în afară de cheia primară; Conţine alte atribute care sunt utile pentru definirea diferitelor niveluri de agregare. Astfel de atribute sunt denumite *ierarhii* ale dimensiunii; Conţine un număr limitat de înregistrări care creşte încet în timp.

Având în vedere cele două obiective majore ale depozitului de date: analiza rezultatelor şcolare ale studenţilor şi analiza structurii studenţilor din universitate, şi respectând pe cât posibil observaţiile de mai sus, au fost identificate şi create următoarele tabele de dimensiuni:

Student: conţine toate atributele legate de student care ar putea interveni la un moment dat în analiză: nr_matricol, numele şi prenumele studentului, sex, stare civilă, vârsta, anul admiterii, media de admitere.

Timp: în acest caz, dimensiunea timp are o structură mai particulară, vând în vedere ca o analiză a rezultatelor şcolare ale studenţilor au sens doar pe sesiune şi pe an calendaristic, acestea sunt şi atributele dimensiunii timp. Combinaţia dintre anul calendaristic şi sesiune poate duce şi la obţinerea rezultatelor pe semestru.

Tip_examen: conţine ca atribute examenul, tipul examenului şi data acestuia, pentru a putea analiza rezultatele pe tipuri de examene şi chiar pe fiecare examen în parte

Disciplina: conţine atribute referitoare la discipline: denumire disciplină, categorie disciplină.

Profesor: conţine toate atributele referitoare la profesor: numele, sex, funcţie, vârstă, vechime. Introducând această dimensiune, analiza capătă o altă nuanţă şi anume, se poate analiza, pe lângă rezultatele studenţilor, performanţele didactice ale profesorilor.

Zona geografică: conţine ca atribute oraşul, judeţul, regiunea, tara ceea ce va permite analiza rezultatelor şi a structurii studenţilor după zona lor de proveniență.

Structură_organizatorică: conţine atributele legate de apartenenţa unui student la o anumită grupă. Utilizarea acestei dimensiuni face posibilă urmărirea studenţilor în dinamica trecerii lor de la o grupă la alta, de la un an de studiu la altul. Cuprinde următoarele atribute: facultate, secţie, grupă, subgrupă, anul de studiu.

Timp_stud: conţine ca atribute, anul calendaristic şi anul universitar necesare pentru analiza numărului şi structurii studenţilor.

Pentru toate dimensiunile s-au folosit ca şi chei primare cheile surogat pentru a asigura independenţa acestora de baza de date tranzacţională şi cu ajutorul cărora se va asigura şi legătura dimensiunilor cu tabela de fapte.

II.2 Definirea tabelei de fapte şi a măsurilor

Înainte de definirea tabelei de fapte trebuie stabilit nivelul de detaliu până la care se doreşte analizarea datelor. Atributele tabelei de fapte trebuie alese cu grijă deoarece acestea

stau la baza tuturor deciziilor viitoare. Granularitatea tabelului de fapte va determina configuraţia atributelor acestuia.
Nivelul de sumarizare ales pentru tabela de fapte va avea un profund efect asupra depozitului de date. Cel mai mic nivel, posibil de granularitate este acela de a stoca faptele la nivelul atomic tranzacţional. Acesta oferă acelaşi nivel de detaliu ca şi sistemul sursă. Stocarea faptelor la acest nivel de detaliu prezintă avantajul major că permite analiza până la cele mai mici niveluri de detaliere, că poate răspunde la interogări neaşteptate şi că depozitul de date este mult mai flexibil la introducerea de noi date. Cel mai mare dezavantaj îl reprezintă faptul că adesea este aproape imposibil ca depozitul de date să poată stoca un asemenea volum de date, şi deoarece cele mai multe interogări se fac la un anumit nivel de sumarizare, datele detaliate nu vor face decât să încetinească timpul de răspuns la aceste interogări. Granularitatea tabelei de fapte va determina profunzimea analizelor care se pot face asupra depozitului de date. O tabelă de fapte în general, are o serie de caracteristici, de care de asemenea am ţinut cont în realizarea ei: Cheie primară a tabelei de fapte se compune din cheile primare ale tabelelor de dimensiuni; Măsurile sunt reprezentate prin coloane numerice care reflectă nivelul de detaliu al datelor stocate în tabela de fapte; Cheile primare ale tabelelor de dimensiuni sunt declarate chei străine în tabela de fapte pentru relaţiile unu la mai mulţi între tabelele de dimensiuni şi tabela de fapte; Multe înregistrări din tabela de fapte nu vor avea valori pentru toate cheile străine asociate dimensiunilor; Conţine un număr foarte mare de înregistrări, de ordinul sutelor de mii şi chiar milioanelor. Nivelul de detaliu stabilit pentru tabela de fapte determină în mod direct şi dimensiunea acesteia şi implicit timpul necesar de răspuns la o interogare.
Având în vedere cele două obiective majore ale depozitului de date: analiza calităţii procesului de învăţământ şi analiza structurii studenţilor din universitate, respectând pe cât posibil observaţiile de mai sus, au fost identificate două tabele de fapte:
Tabela **Rezultate** stochează toate rezultatele obţinute de studenţi. Măsura acestor rezultate este dată de notă, credite, note peste 5, note peste 8 care reprezintă măsurile tabelei de fapte. Aceste măsuri vor sta la baza calculării indicatorilor de analiză a calităţii procesului de învăţământ.
Tabela **Structură Studenţi** stochează faptele legate de structura studenţilor ţinând cont de dinamica acestora pe parcursul anilor. Măsura tabelei este dată de numărul de studenţi, măsură ce va fi utilizată pentru calcularea indicatorilor necesari pentru analiza structurii studenţilor.
O particularitate a sistemului constă în faptul ca acelaşi student, va face parte în ani universitari diferiţi din alte grupe şi ani de studii, deci pentru a obţine o situaţie reală a componenţei structurale a studenţilor şi a rezultatelor acestora, ei trebuie urmăriţi în dinamica trecerii lor de la un an de studiu la altul. Acest lucru se realizează cu tabelele de dimensiuni *Studen*t şi *Structură organizatorică* care permit încărcarea în tabelele de fapte a situaţiei reale la un moment dat.

II.3 Schema conceptuală a depozitului de date

Modelele cele mai utilizate în faza de concepţie a unui depozit de date sunt modelele dimensionale care regrupează datele din tabelele relaţionale în scheme de tip: stea, fulg de zăpadă şi constelaţie.
Modelul stea este cel mai comun model de date, în care depozitul de date conţine un tabel central voluminos (tabelul de fapte) şi un set de tabele însoţitor (tabelele dimensiune) pentru fiecare dimensiune. Tabelul de fapte cuprinde cea mai mare parte a datelor fără redundanţe Graful asociat semănă cu o stea în care tabelele dimensiune sunt afişate radial în jurul tabelului de fapte central.
Modelul fulg de zăpadă este o variantă a modelului stea, unde o parte din tabelele dimensiune sunt normalizate, astfel datele sunt împărţite în tabele suplimentare. Rezultă o schemă reprezentată într-un graf similar unui fulg de zăpadă. Diferenţa majoră între modelul „fulg de zăpadă" şi modelul stea este că tabelele dimensiune din modelul „fulg de zăpadă" pot fi păstrate în forma normalizată ceea ce determină o redundanţă redusă. Ase-

menea tabele sunt uşor de întreţinut şi se economiseşte spaţiu de stocare, deoarece un tabel dimensiune mare poate deveni enorm când structura dimensională este inclusă în coloane. Totuşi această economie de spaţiu este neglijabilă în comparaţie cu volumul foarte mare de date din tabelul de fapte. Mai mult, structura „fulg de zăpadă' poate reduce eficacitatea „browsing-ului" când mai multe „join-uri" trebuie execute la o interogare. De aceea, schema fulg de zăpadă este mai puţin răspândită.

Modelul constelaţie. Aplicaţiile sofisticate pot solicita tabele multiple de fapte care partajează tabelele dimensiune. Acest gen de schemă poate fi văzută ca o colecţie de stele şi, de aici, denumirea de schemă galaxie sau constelaţie de fapte (fact constellation). Pentru depozitele de date, schema constelaţie de fapte este în mod curent utilizată.

Depozitul de date pentru aplicaţia realizată are două tabele de fapte: *rezultate* şi *structură studenţi* care au ca dimensiuni comune (*zonă_geografică*, *student* şi *structură organizatorică*) de aceea schema depozitului de date va fi de tip constelaţie aşa cum se poate observa în figura 4. Schema depozitului de date a fost obţinută în urma unui proces de denormalizare a bazei de date tranzacţionale care este tot o bază de date relaţională, SQL Server. (figura 3)

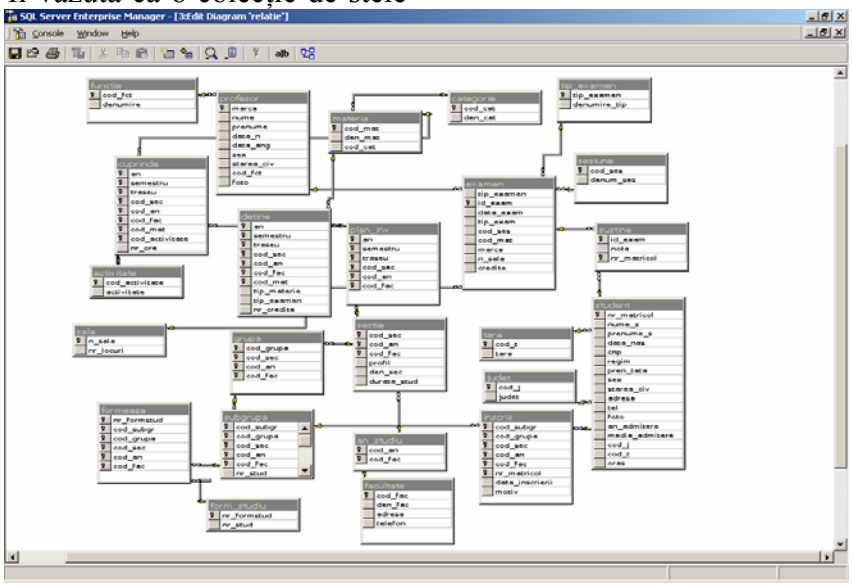

Fig. 3. Schema bazei de date tranzacţionale

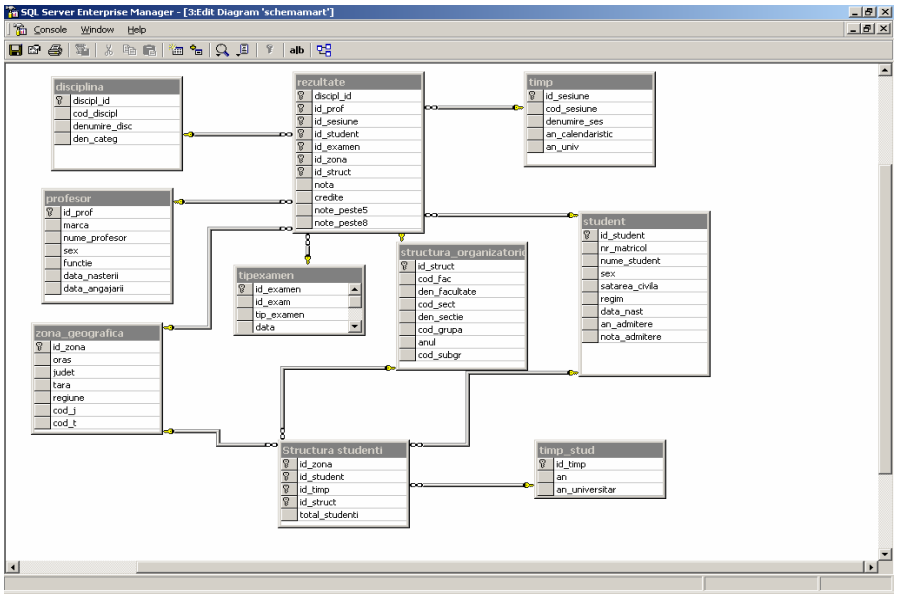

Fig. 4. Schema conceptuală a depozitului de date

În această primă parte am prezentat etapele de analiză si proiectare a depozitului de date, urmând ca în cea de-a doua parte a articolului sa prezentăm popularea depozitului de date şi analiza multidimensională a datelor utilizând cuburile de decizie.

Bibliografie

1. Albescu F., Bojan I. „Management information systems and decizion support systems" Ed,. DualTech, Bucureşti 2001
2. Airinei D, „Sisteme informatice de asistare a deciziei", curs on-line
3. Airinei D. „Depozite de date", curs on-line
4. Bain T. &co, „Professiona SQLServer 2000 Datawarehousing with Analysis Services", Wrox Press, Londra 2000
5. Connolly T., „Baze de date", Teora, Bucureşti 2001
6. Tanrikorur Tuturu, "Enterprise DSS arhitecture: a hybrid approach", DM Review, February 1998
7. William McKnight, "Way a data warehouse?", McKnight Associates, Inc, April 2002
8. Zaharie D & co, „Sisteme informatice pentru asistarea deciziei", Ed. DualTech, Bucureşti 2001
9. ***"Microsoft SQL Server 7.0 Data Warehousing Training Kit", MS Press
10. ***Microsoft SQL Server 7.0 Data warehousing strategy
11. ***Microsoft SQL Server 7.0 OLAP Services
12. ***MS Press - MCSE Training Kit-SQL 7.0 Data Warehousing
13. http://www.billinmon.com
14. http://www.datawarehouse.com
15. http://www.datawarehousingonline.com
16. http://www.dmreview.com
17. http://www.dw-institute.com
18. http://www.olapconcil.com
19. http://www.intelligententreprise.com
20. http://www.olapreports.co