MANAGEMENTUL ÎNCĂRCĂRII BAZELOR DE DATE DE MARI DIMENSIUNI

Ion Ivan
ionivan@ase.ro

Iulian Rădulescu
iulian.radulescu@intrasoft-intl.com

Marius Popa
marius.popa@ie.ase.r

Academia de Studii Economice Bucureşti

Dragoş Cârciumaru
dragos.carciumaru@siveco.ro
SIVECO România S.A.

Rezumat. Se prezintă importanţa utilizării bazelor de date în stocarea informaţiei. Se evidenţiază operaţiile care au loc pe baze de date şi sunt prezentate forme ale încărcării datelor în baze de date. Sunt detaliate aspecte care privesc manipularea bazelor de date cu un volum foarte mare de informaţie stocată. Sunt evidenţiate tehnici şi metode de validare a procesului de încărcare a bazelor de date, de mari dimensiuni. Analiza încărcării bazelor de date cu informaţii este efectuată prin intermediul metricilor.

Cuvinte cheie: .baze de date, volum de informaţie, încărcare baze de date.

1. Bazele de date de dimensiuni foarte mari

Datorită complexităţii domeniilor modelate şi volumului de date prelucrate, majoritatea aplicaţiilor utilizează baze de date. Dimensiunea acestora variază, de la medii, la mari şi foarte mari. Bazele de date de dimensiuni foarte mari înmagazinează cantităţi uriaşe de informaţie care, ulterior, suportă o serie de prelucrări complexe pentru obţinerea de rezultate de sinteză. Pentru aceste tipuri de baze de date, s-a creat conceptul de *data warehouse* – depozit de date.

O operaţie foarte importantă o constituie încărcarea bazelor de date cu informaţie. Aceasta se realizează fie automat, fie prin intermediul unor operatori umani. Încărcarea automată presupune existenţa unei aplicaţii care preia datele în mod automat de la sursa de date. De exemplu, sistemul de înregistrare a convorbirilor al unei companii de telefonie preia datele de la centralele telefonice digitale.

Încărcarea manuală presupune o aplicaţie prin care, însă, datele sunt introduse de către operatori, prin intermediul interfeţelor grafice, puse la dispoziţie de aceasta. Această operaţie:

- presupune un număr foarte mare de operatori care, prin pregătire, experienţă şi caracteristici native, sunt o colectivitate cu un grad foarte mare de eterogenitate;
- are la bază o procedură unică de culegere a datelor, de pregătire şi introducere, pe care operatorii sunt obligaţi să o respecte întocmai;
- este studiată riguros, pentru a se identifica tipurile de erori care se produc, frecvenţa acestora, pentru a adopta procedurile fie pentru eliminarea încă din faza de culegere sau, în cel mai puţin favorabil caz, pentru a le depista şi pentru a le elimina înaintea prelucrărilor şi obţinerii de rapoarte.

În funcţie de natura domeniului din care se culeg datele, de tipul aplicaţiei şi de infrastructura existentă şi la care operatorul are acces, încărcarea se efectuează în mai multe moduri.

Încărcarea on-line presupune existenţa unor posturi de lucru, legate permanent de serverul care administrează reţeaua şi baza de date. Pentru a asigura calitatea finală a conţinutului bazei de date, sunt necesare proceduri complexe de validare şi de măsurare a numărului de elemente ale colectivităţii pentru care au fost efectuate înregistrări.

Deşi este costisitoare din punct de vedere tehnic, încărcarea on-line elimină operatorii intermediari, care influenţează calitatea bazei de date prin modul în care manipulează părţi în procesul de conectare şi transmisie.

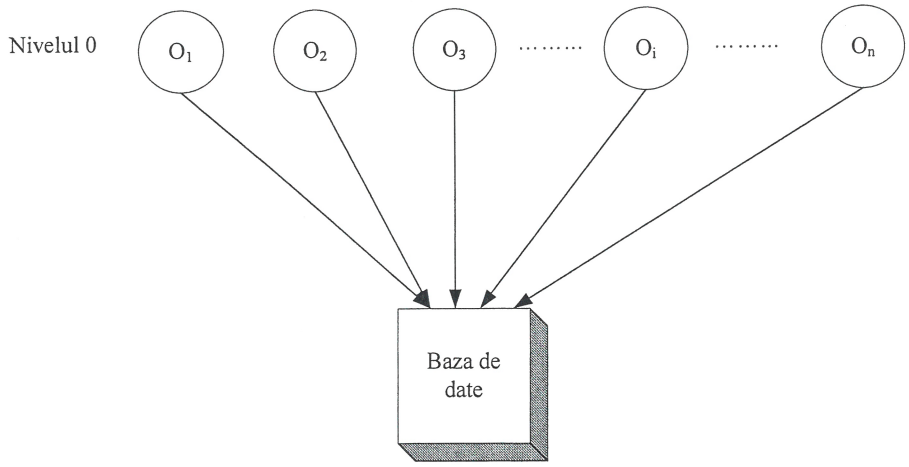

Figura 1. Încărcarea on-line a bazei de date

Se consideră operatorii O_1, O_2, ..., O_n care au acces la baza de date. Nivelul de acces este controlat prin intermediul aplicației utilizate pentru încărcarea datelor şi depinde de tipul informației manevrate. Produsul software trebuie înzestrat cu astfel de componente încât să se înregistreze dinamica proceselor de încărcare, evidențiind:

- momentul de start şi de încheiere a sesiunii de lucru;
- ritmul de introducere a datelor;
- erorile care s-au produs.

Toate aceste informații joacă un rol activ în procesul de perfecționare a produsului şi la corectarea procedurilor de selecție şi instruire a operatorilor.

Încărcarea folosind centralizări intermediare presupune existența unor posturi de lucru intermediare, la nivelul cărora datele provenind de la un subset de operatori se centralizează, de exemplu, folosind fişiere text. Construcția acestor fişiere se face sau nu folosind aplicații informatice. Aceste centralizări sunt ulterior validate printr-o procedură complexă de validare Apoi, se trece la nivelul punctelor intermediare PI, la încărcarea on-line a bazei de date, fişierelor text centralizate, figura 2.

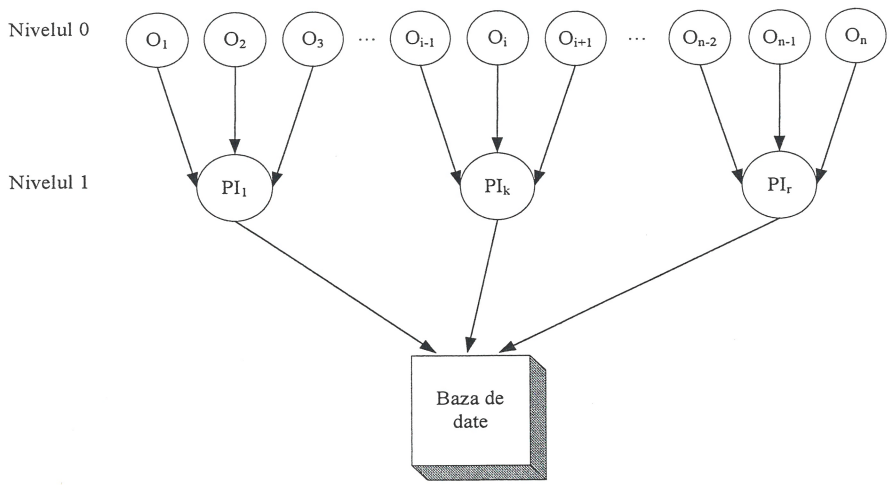

Figura 2. Încărcarea on-line la nivel intermediar

Încărcarea folosind baze de date intermediare este similară mecanismului descris anterior. Există o centralizare a datelor la un nivel intermediar. Acest mecanism, însă, presupune existența unor posturi de lucru, care se conectează la o bază de date locală de mici dimensiuni, prin intermediul aplicației pe care o populează cu informații, similar cazului încărcării on-line. După încheierea procesului de încărcare, datele sunt centralizate şi încărcate în baza de date prin mecanisme automate. Această încărcare se face folosindu-se instrumente software.

O problemă majoră în încărcarea bazelor de date mari o constituie gestiunea cheilor de identificare. Deoarece toți operatorii se conectează la o aceeași bază de date, în varianta încărcării on-line sau folosind centralizări intermediare, cheile de identificare sunt gestionate prin intermediul mecanismelor interne ale bazei de date. Pentru fiecare înregistrare se generează un număr unic de înregistrare. În varianta folosirii bazelor de date intermediare, fiecare bază de date generează chei de identificare pentru înregistrările stocate. Aceste chei sunt folosite la stabilirea de relații cu alte înregistrări. De aceea, în momentul centralizării, ele trebuie gestionate cu grijă. Aceeași cheie de înregistrare este folosită de mai mult de o bază de date intermediară pentru a stoca înregistrări. Există două abordări:

- inițial, se alocă plaje de valori pentru cheile de identificare, pentru fiecare bază de date intermediară; astfel, în momentul centralizării, cheile nu se suprapun;
- fiecare bază de date generează chei în mod independent, urmând ca la centralizare instrumentele software utilizate pentru această operație să permită reasignarea înregistrărilor cu chei duplicate la chei noi, asigurând consistență bazei de date.

Tipul de introducere depinde de modul în care este proiectat sistemul informatic, de infrastructura existentă şi de fondurile alocate. De exemplu, un recensământ se realizează la modul următor:

- se distribuie fiecărui operator un set de formulare, pe care acesta le completează împreună cu fiecare locuitor în parte; cetățeanul validează prin semnătură corectitudinea datelor introduse;
- formularele sunt încărcate în baza de date, folosindu-se mai multe puncte de lucru intermediare; metoda de încărcare este cea a centralizărilor intermediare.

Dacă fondurile disponibile sunt importante, prin dotarea operatorilor cu instrumente de calcul performante, ei înregistrează informațiile de la locuitori direct într-o bază de date intermediară, care, ulterior, este încărcată în baza de date centrală. Această abordare este mult mai avantajoasă prin faptul că utilizarea aplicației de introducere a datelor asigură o mai bună validare a acestora şi permite identificarea erorilor înainte ca acestea să fie înregistrate. Dacă lipseşte codul numeric personal de pe un formular ce a fost completat sau acesta a fost completat greşit, atunci aplicația semnalează absența informației cerute sau incorectitudinea acesteia. De exemplu, pentru codul numeric personal există un algoritm de validare. Acelaşi lucru nu este făcut întotdeauna de către operator. El observă la centru că informația culeasă este incompletă sau incorectă, fiind necesară reculegerea ei.

2. Volumul bazelor de date foarte mari

Volumul de înregistrări dintr-o bază de date variază foarte mult în funcție de specificul domeniului pentru care se culeg datele. De asemenea, natura acestor informații diferă foarte mult. De exemplu, pentru o bază de date ce stochează informații referitoare la convorbirile telefonice înregistrate de o centrală telefonică, informațiile sunt, în general, numerice, sunt înregistrate cu o frecvență foarte mare, baza de date având permanent un număr masiv de conexiuni deschise.

Pentru o aplicație de evidență a populației sau care conține date rezultate din observări statistice, structura bazei de date diferă. Pe lângă înregistrările propriu-zise, o astfel de bază de date conține numeroase nomenclatoare care definesc contextul în care această observare a fost efectuată. De exemplu, pentru un recensământ, se folosesc nomenclatoare de localități, nomenclatoare de ocupații, nomenclatoare de instituții etc.

Dacă N este numărul de componente pentru care sunt culese date, în baza de date apar cel puțin N înregistrări. De cele mai multe ori, apar multiplu de N înregistrări deoarece pentru fiecare informație principală înregistrată despre o componentă mai există informații utile culese, care şi ele sunt stocate în baza de date. De exemplu, în baza de date ce stochează informații despre candidații la admiterea în licee. Pentru anul 2005, se regăsesc peste 200.000 de candidați absolvenți de clasa a VIII-a, care au participat la admitere computerizată şi peste 4.000.000 de opțiuni, pe care candidații le-au ales la admiterea în liceu.

Datele referitoare la un element din colectivitate sunt corecte şi complete prin aplicarea procedurilor. Pentru o colectivitate având N componente, baza de date conține N înregistrări sau un multiplu de N. În realitate, se impune cunoaşterea numărului efectiv de elemente ale colectivității, N.

Pe măsură ce se introduc date în fişiere pentru descrierea elementelor colectivității, înregistrările sunt numerotate. X_N este variabila în care se găseşte numărul de înregistrări încărcate în baza de date. După terminarea procesului de inserare, se compară variabila X_N cu numărul efectiv de elemente ale

colectivității C. Dacă $X_N = N$, rezultă că baza de date conține un număr de articole identic cu numărul de componente ale colectivității.

Dacă $X_N > N$, rezultă că baza de date conține mai multe articole decât numărul componentelor din colectivitate. Rezultatul se obține prin introducerea repetată a unor date. Această situație apare când validările nu includ teste asupra cheilor înregistrărilor deja stocate. O soluție, în acest caz, o constituie introducerea de validări și pentru alte informații decât cele conținute în cheile primare. De exemplu, pentru a identifica înregistrări duplicate, se fac validări de unicitate și după codul numeric personal sau seria și numărul de buletin, numărul de pașaport, adresă etc.

Dacă $X_N < N$, rezultă că există componente ale colectivității, care nu au fost înregistrate. Se procedează la identificarea și introducerea datelor lipsă.

În toate cazurile, din aproape în aproape, trebuie să se obțină egalitatea $X_N - N = 0$ pentru a considera că datele sunt corecte și complete în primă fază.

O bază de date foarte mare ocupă și un spațiu important pe suportul fizic. Evaluarea dimensiunii fizice a unei baze de date este unul din primii pași înainte de implementarea unei astfel de soluții. Calculul dimensiunii fizice se realizează pornindu-se de la următoarele informații:

- numărul, cunoscut sau estimat, al componentelor colectivității pentru care se culeg datele;
- natura datelor: numerice, booleene, șiruri de caractere;
- reprezentarea fizică pentru fiecare tip de dată în parte; de exemplu, pentru tipul numeric întreg, reprezentarea este pe 4 baiți, pentru tipul șir de caractere, reprezentarea este 1 bait pentru fiecare caracter.

Pe baza acestor informații, se stabilește o dimensiune aproximativă a bazei de date. Acest lucru este util în momentul deciziei de achiziționare a unui sistem de gestiune a bazelor de date. Se optează pentru sisteme distribuite, care sunt instalate folosind tehnologiile cluster, cu soluții performante de *back-up*. Se arhivează și se salvează datele în mod regulat pentru a avea permanent un set de date la care să se revină în cazul în care baza de date din producție suferă defecțiuni. În cazul în care o componentă a acestuia încetează să mai funcționeze, un mecanism de tip *fail-over* este lansat în scopul asigurării.

3. Validarea procesului de încărcare a bazelor de date de mari dimensiuni

Pe durata încărcării unei baze de date și după efectuarea ei, o serie de validări sunt realizate pentru a asigura calitatea datelor introduse. Există, în primul rând, validări legate de tipul de dată: numeric, caracter, de tip dată calendaristică, booleană etc. Acesta constituie nivelul primar de validare. Orice dată trebuie să corespundă tipului asociat în baza de date pentru a putea fi încărcată.

Urmează un nivel intermediar, aplicabil unor date care urmăresc anumite șabloane. De exemplu, codul numeric personal al unui individ are un anumit format: prima cifră indică sexul și ia valori 1 sau 2, următoarele 6 cifre indică, două câte două, anul, luna și ziua de naștere, iar restul de șase sunt generate pe baza unui algoritm stabilit. Dacă în baza de date se introduce un cod numeric personal - CNP, atunci codul trebuie validat conform pașilor descriși anterior.

Nivelul cel mai de sus îl constituie validarea în contextul domeniului problemei. Asta înseamnă că, între două date, trebuie să existe o anumită relație, liniară sau nu, sau că existența unei date determină automat absența alteia. Dacă datele sunt agregate la niveluri diferite, de exemplu, la nivel de județ și la nivel de țară, atunci există validări care asigură consistența datelor peste aceste niveluri. Datele introduse în Alba trebuie să fie consistente cu cele introduse în Cluj. De exemplu, în cazul unei baze de date cu candidații la admiterea în facultăți, la nivel național trebuie să se verifice dacă există candidați care apar înscriși în mai multe județe. Aceasta presupune efectuarea de validări peste datele agregate la nivelul județelor.

Datorită volumului foarte mare de înregistrări care trebuie introduse, este imposibilă asigurarea calității datelor încă de la prima încărcare. Prin urmare, sunt necesare mecanisme de validare a datelor, utilizând chei de control.

Prima cheie de control o reprezintă numărul de înregistrări. Dacă se cunoaște de la început câte înregistrări trebuie să se stocheze, la sfârșitul încărcării există o primă validare, respectiv, numărul efectiv de înregistrări. Numărul de înregistrări apare la diverse niveluri de agregare: număr total de înregistrări, număr de înregistrări pe grupe/subgrupe etc. La fiecare dintre aceste niveluri, se realizează validări ale acestei chei de control. De exemplu, dacă sunt introduse datele despre toții candidații la admiterea în

licee, există următoarele niveluri: număr candidați pe centru de înscriere, număr candidați pe județ și număr candidați pe țară.

O altă cheie de control o constituie sondajul. Din volumul de date care trebuie introduse, se extrage un eșantion reprezentativ, folosind metode și tehnici statistice. Datele din sondaj sunt verificate în baza de date populată pentru corectitudine. În cazul unui eșantion reprezentativ, procentul de corectitudine este estimat la nivelul întregului volum de date, validându-se astfel procesul de încărcare. Dacă procentul este mai mic de 100%, se continuă cu o nouă iterație de validare și corectare la nivelul întregului volum de date.

Datele într-o bază de date nu se încarcă prin acces direct la baza de date, ci folosind aplicații informatice. Aceste aplicații trebuie să îndeplinească următoarele condiții:

- să funcționeze corect când datele sunt corecte și complete;
- să evidențieze erorile din clasele de erori introduse prin date de test;
- să evidențiază noi erori de prelucrare generate de modul în care a fost încărcată baza și au fost scrise modulele, nereflectând riguros specificațiile.

Dacă pentru baza de date se construiește software care efectuează încărcarea respectând condițiile:

- *câmpul 1* – cheie de identificare cu valori 1, 2, ... N;
- *câmpul 2* – șir de litere mari;
- *câmpul 3* – șir de litere mici;
- *câmpul 4* – număr întreg în intervalul [100, 800];
- *câmpul 5* – una din literele {a, b, c, d}.

Se generează setul de date de test din tabelul 1, cu date corecte și complete.

Tabelul nr. 1 Set de date de test corecte și complete

câmp 1	câmp 2	câmp 3	câmp 4	câmp 5
1	ALFA	Alfa	100	A
2	BETA	Beta	200	B
3	CALITATE	Calitate	300	C
4	SOFTWARE	Software	400	A
5	BAZĂ	Bază	500	D
6	DATE	Date	600	C
7	CÂMP	Câmp	800	B
8	FIȘIER	Fișier	750	A
9	TABEL	Tabel	350	D
10	SET	Set	700	B

În ipoteza în care modulele programului sunt scrise respectând specificațiile, mesajele evidențiază:

- numărul de articole înregistrate este egal cu cel al numărului componentelor colectivității;
- cheile respectă cerința de a fi în progresie aritmetică cu primul termen $a_0 = 1$ și rația $r = 1$;
- câmpurile aparțin domeniilor care au fost definite.

În tabelul 2, se regăsesc date ce sunt folosite pentru a testa senzitivitatea software a aplicației, la introducerea de date ce nu respectă condițiile.

Tabelul nr. 2 Date pentru evaluările senzitivității software

câmp 1	câmp 2	câmp 3	câmp 4	câmp 5
1	AAA	Aaa	200	A
1	BBB	Bbb	250	B
2	xAB	Xaa	900	A
3	XAB	Xaa	2	B
4	ABC	Abc	200	Y

Primul și al doilea articol au toate câmpurile corecte, numai câmpul cheie de identificare este incorect, fiind duplicat. Al treilea articol conține erori în *câmp2* și *câmp3*, iar valoarea *câmp4* este în afara domeniului la dreapta. Al patrulea articol are valoarea *câmp4* în afara domeniului la stânga. Câmpul cu cheia 4 are valoarea pentru *câmp5* eronată, în sensul că nu aparține mulțimii indicate.

Prelucrând datele acestui tabel, produsul software trebuie să evidenţieze toate erorile şi să permită corectarea. Sunt situaţii când datele din tabele evidenţiază toate erorile, însă produsul mai conţine şi alte erori. Aceste din urmă erori trebuie să apară dirijat, şi nu să fie evidenţiate accidental.

Experienţa celor care testează software, care operează cu baze de date de dimensiuni mari permite adăugarea la tabelele iniţiale de noi tabele pentru a mai obţine noi combinaţii de situaţii generatoare de erori.

Un alt set de validări ce trebuie realizat se referă la posibilitatea utilizării datelor în calcule matematice. Aici intervin restricţiile cunoscute cum sunt: nenulitatea numitorului, argumentul funcţiei radical trebuie să fie un număr pozitiv, argumentul funcţiei logaritm trebuie să fie un număr pozitiv etc. Aceste validări sunt foarte importante, mai ales dacă informaţiile culese reprezintă baza de calcul pentru diferiţi indicatori statistici.

Un aspect deosebit de important, care priveşte testarea aplicaţiilor care manevrează baze de date mari, o constituie seturile de date. Generarea datelor pentru popularea unor astfel de baze de date este imposibil de realizat manual. Este nevoie de instrumente software generatoare de date de test. Există numeroase produse care se conectează la o bază de date şi, pornind de la tipul câmpurilor şi ţinând cont de restricţiile impuse de utilizator, generează seturi uriaşe de date în baza de date. Acestea permit testarea aplicaţiilor folosindu-se baze de date de dimensiuni apropiate de cele din realitate.

4. Grafice şi rapoarte statistice privind încărcarea bazelor de date de dimensiuni mari

Dinamica încărcării bazei de date dintr-un singur post de lucru se evidenţiază prin graficul din figura 3 în care se înregistrează numărul de articole inserate în ziua Z_i.

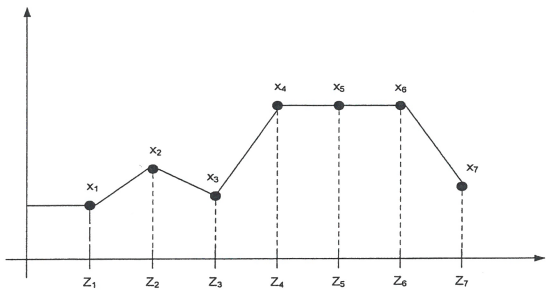

Figura 3. Numărul de articole încărcate zilnic

Graficul gradului de încărcare a bazei de date în ziua Z_i dat de relaţia $G_i = \frac{K_i}{N} * 100$, unde K_i reprezintă numărul de articole existente în baza de date la sfârşitul zilei Z_i este reprezentat în figura 4.

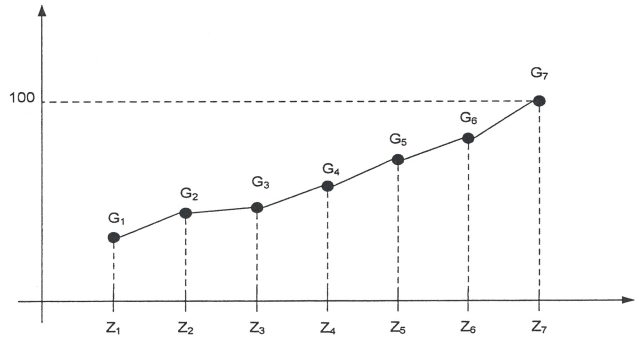

Figura nr. 4 Evoluţia gradului de încărcare a bazei de date

În cazul în care încărcarea bazei de date se realizează din mai multe puncte de lucru A, B, C şi D simultan, numărul de articole care trebuie încărcate fiind N_A, N_B, N_C, respectiv N_D articole, se construieşte un raport cu structura dată în tabelul 3.

Tabelul 3. Înregistrarea zilnică de articole în posturile de lucru

Postul de lucru	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Total
A	N_{A1}	N_{A2}	N_{A3}	N_{A4}	N_{A5}	N_{A6}	N_{A7}	N_A
B	.....	.....	.....	.....	.....	.....	.....	N_B
C	.....	.....	.....	.....	.....	.....	.....	N_C
D	N_{D1}	N_{D2}	N_{D3}	N_{D4}	N_{D5}	N_{D6}	N_{D7}	N_D
Total	N_1	N_2	N_3	N_4	N_5	N_6	N_7	N

Reprezentarea procentuală a gradului de încărcare în centrul X_i în ziua Z_k este:

$$G_{ik} = \frac{\sum_{j=1}^{k} Y_{ij}}{N_i}$$

Pentru exprimarea la nivelul întregului sistem de centre se foloseşte relaţia:

$$G_K = \frac{\sum_{i=1}^{L}\sum_{j=1}^{K} Y_{ij}}{N}$$

unde Y_{ij} reprezintă numărul de înregistrări realizat în ziua Y_j în centrul H_i.

Tabelul 4. Gradele de încărcare

Centrul	Z_1	Z_2	.....	Z_j	.....	Z_K
$\mathbf{H_1}$	G_{11}					G_{1K}
$\mathbf{H_2}$						
.....						
$\mathbf{H_i}$				G_{ij}		G_{iK}
.....						
$\mathbf{H_L}$						G_{LK}

Dacă $G_{ij} = 1$ pentru i = L şi j = K, rezultă că baza de date a fost integral încărcată. Pentru totalitatea centrelor ultima linie a tabelului 4 constituie gradul de încărcare.

Rapoartele asupra erorilor au menirea de a produce corecţii în ceea ce priveşte:

- calificarea operatorilor accentuând asupra eliminării cauzelor de interpretare;
- produsul software prin introducerea unor noi secvenţe care fac interfaţa mai prietenoasă sau care evidenţiază mai bine erorile, localizând poziţionarea câmpurilor.

5. Concluzii

În gestiunea bazelor de date de mari dimensiuni trebuie să ia în considerare un prim aspect foarte important: datorită volumului mare de date precum şi a naturii datelor înregistrate, încărcarea acestor baze de date nu este repetabilă. Trebuie să se aibă în vedere construirea unei aplicaţii solide, care să permită încărcarea de date corecte şi complete.

Multe sisteme informatice dispun de baze de date de mari dimensiuni fie că este vorba de sisteme integrate de management al afacerii ERP, fie de sisteme de cercetare statistică stocatoare de date în serii de timp despre fenomene economice şi nu numai, fie că este vorba de sisteme real-time, cum sunt cele din telecomunicaţii. Este important ca managementul unor astfel de depozite de informaţie să fie realizat de persoane pregătite pentru a nu altera calitatea datelor şi implicit calitatea şi relevanţa rezultatelor.

Bibliografie

1. **DYCHE, J.**: e-Data: Turning Data into Information with Data Warehousing, Addison-Wesley Pub co., 2000.
2. **GONZALES, M.**: IBM Data Warehousing: With IBM Business Intelligence Tools, Wiley, 2003.
3. **IVAN, I., M. POPA**: Entități text – dezvoltare, evaluare, analiză, Editura ASE, Bucureşti, 2005
4. **IVAN, I., P. POCATILU, M. POPA, C. BOJA, S. NICULESCU**: Replicarea bazelor de date, Simpozionul „Tehnologii educaționale pe platforme electronice în învățământul ingineresc", 9 – 10 mai 2003, Universitatea Tehnică de Construcții Bucureşti.
5. **IVAN, I., P. POCATILU, D. CAZAN**: Certificarea bazelor de date utilizate în aplicații Internet. În: Informatică Economică, vol. V, nr. 2 (18), 2001, pp. 71-74.
6. **LUNGU, I., C. BODEA, G. BĂDESCU, C. IONIȚĂ**: Baze de date – Organizare, proiectare şi implementare, Editura All Educational, Bucureşti, 1995.
7. **IVAN, I., GH. NOŞCA, O. PÂRLOG**: Calitatea Datelor, Editura INFOREC, Bucureşti.