CALITATEA DATELOR ŞI METADATELOR ÎN DATAWAREHOUSE

Carmen Răduț[1]

Rezumat: *Calitatea datelor reprezintă un concept cheie pentru aplicațiile economice în procesul de analiză. Domeniul bazelor de date a fost revoluționat prin introducerea lucrului cu volume mari de date. Pornind de la acest concept,un proces important îl reprezintă stocarea datelor multidimensionale în depozite de date (data warehouse), pentru efectuarea prelucrărilor şi analizelor în vederea obținerii informațiilor pentru fundamentarea luarii deciziilor în diverse domenii de activitate. Studiile de specialitate arată că cele mai multe volume de date nu sunt utile scopului pentru care au fost create, din cauza lipsei de calitate a datelor stocate, precum şi a tehnicilor incorecte de manipulare a acestora. Această lucrare încearcă să ofere un procedeu pentru a obține date calitative în arhivele de date, pentru evitarea anomaliilor de calitate la nivelul metadatelor.*

Cuvinte cheie: *metadata, depozit de date, calitate, arhitectura*

Clasificarea JEL: C81, C89

1. Introducere

Soluția ideală pentru aplicațiile complexe de tipul **B**usiness **I**ntelligence o reprezintă implementarea unui depozit de date. În cele mai multe cazuri, depozitele de date nu satisfac solicitările beneficiarilor datorită slabei calități a datelor. Astfel, calitatea datelor este o problemă importantă atât pentru implementarea, cât şi pentru gestionarea depozitelor de date. Pentru a folosi eficient datele din depozitele de date, acestea sunt supuse unui proces de analiză şi curățire. Astfel, în aplicațiile BI, sunt necesare continuu proceduri de curățire a datelor pentru a obține nivelul de calitate dorit al datelor.

Domeniul în care depozitul de date eşuează îl constituie transferul datelor din diferite surse de date, într-un singur depozit de date, printr-un proces de integrare al datelor. Dezvoltarea depozitelor de date pentru aplicațiile de business reprezintă un proces important în dezvoltarea sistemelor informatice.

Pentru definirea depozitului de date trebuie stabilită **granularitatea datelor** din depozitul de date precum şi **nivelul de abstractizare** necesar în susținerea procesului de luare a deciziilor în activitatea de management. Studii recente arată ca peste 80% din proiectele de data warehouse depăşesc bugetele alocate datorită *neînțelegerii sursei datelor* sau a *definirii incorecte* a lor. Un studiu similar arată că şi *calitatea datelor* este un motiv pentru care, în general, eşuează aceste proiecte. De aceea, este necesară o evaluare a calității datelor, efectuată atât în momentul construirii depozitului de date, cât şi în momentul popularii graduale, treptate cu date (Singh R. i Singh K., 2010).

2. Calitatea datelor în depozitele de date

Calitatea datelor în depozitele de date implică abordarea următoarelor domenii: evaluarea calității informațiilor; clasificarea problemelor legate de calitatea datelor; structura depozitului de date; arhitectura/modelul calității depozitului de date; instrumente privind calitatea datelor; standardele de calitate a datelor; sistemul de control al calității metadatelor (Palepu and Rao, 2012).

2.1. Evaluarea calității informațiilor este un criteriu definit de managementul calității totale al datelor - TQdM (Total Quality data Management). Metodologia TQdM de gestionare a calității, constă în parcurgerea a şase procese de măsurare şi îmbunătățire a

[1] Conf.univ.dr., Universitatea „Constantin Brâncoveanu" Pitești, Facultatea Management Marketing în Afaceri economice Râmnicu Vâlcea, c_radut@yahoo.com.

calității informațiilor datorate schimbarilor de mediu. Cele şase procese privind *măsurarea şi îmbunătățirea calității informațiilor* se referă la: evaluarea definiției datelor şi a arhitecturii calității informațiilor; evaluarea calității informațiilor; măsurarea costurilor informațiilor noncalitative; restructurarea şi curățirea datelor; evaluarea calității procesului informațional; stabilirea cadrului privind calitatea informațiilor.

Cele 6 procese sunt împărțite în sub-etape pentru a obține calitatea dorită a datelor. Pentru a determina calitatea datelor este necesară o evaluare a tuturor câmpurilor. Simpla deținere a datelor nu este suficientă, trebuie cunoscut contextul pentru care urmează să fie utilizate acestea, proces ce permite obținerea **metadatelor**. Metadatele sunt informații ataşate pentru a suplimenta valoarea datelor din depozitele de date cu privire la conținutul şi scopul acestora. Evaluarea calității poate fi extinsă în funcție de disponibilitatea metadatelor. *Criteriile de calitate* sunt: integritatea tipului de date; integritatea regulilor de afaceri (de business); integritatea numelui şi a adresei. De asemenea, *caracteristicile şi măsurile privind calitatea informației* sunt: concordanța definiției, integralitatea valorilor; valabilitatea sau conformitatea cu regulile de afaceri (de business); corectitudinea în raport cu sursa surogat; corectitudinea în raport cu realitatea; precizia; nonduplicarea apariției; echivalența/concurența datelor redundante sau distribuite; accesibilitatea.

2.2. Clasificarea problemelor legate de calitatea datelor. Analiza cauzelor care stau la baza problemelor legate de calitatea datelor, a planificării realizării designului datelor implică, în primul rând, identificarea problemelor legate de ***calitatea datelor***. Identificarea şi clasificarea acestor probleme este necesară atât depozitului de date, cât şi calitații datelor. Astfel, se identifică probleme legate de calitatea datelor, la mai multe niveluri: **la surse de date; în stadiul de stabilire a profilului datelor; în etapa ETL (Extragere, Transformare şi Incărcare); în etapa de elaborare a schemei de proiectare; în etapa de obținere a depozitului de date** (Palepu and Rao, 2012).

a. Probleme legate de calitate datelor la surse de date. Sursele de date au anumite tipuri de probleme asociate, de exemplu, faptul că datele din sursele de date moştenite nu conțin metadata care să le descrie. Sursele datelor neclare (incorecte, impure) sunt date de: erori la datele de intrare datorate factorului uman sau sistemului de calcul, precum şi erori de actualizare a datelor. O parte din date provin din fişiere text sau fişiere MS Excel însă pot fi obținute şi din conectarea directă prin ODBC la sursele de date. Unele fişiere sunt rezultatul consolidarii manuale a mai multor fişiere ceea ce determină compromiterea calității datelor. Cauzele problemelor cu privire la calitatea datelor la sursele de date se referă la: selecția nepotrivită a surselor de date candidate; cunoaşterea necorespunzătoare a dependențelor interne între sursele de date; lipsa de rutine (proceduri) de validare a surselor; modificări neaşteptate în sistemele sursă; sursele multiple de date generează eterogenitate semantică, conducând la probleme de calitate; prezența informațiilor divergente în sursele de date; formatarea inconsistentă/inexactă a datelor; surse multiple de date pentru aceleaşi date.

b. Probleme legate de calitatea datelor în stadiul de stabilire a profilului datelor. Dupa ce sunt identificate posibilele sursele de date concurente urmează stabilirea profilului datelor. Stabilirea profilului datelor constă în examinarea şi evaluarea calității, integrității şi coerenței datelor sistemelor sursă numita şi ***analiza sistemelor sursă***. Profilarea datelor este fundamentală, deşi adesea este ignorată sau i se acordă mai puțină atenție, lucru care duce la compromiterea calității datelor din depozitele de date. Astfel, cauzele problemelor de calitate a datelor în etapa de stabilire a profilului datelor sunt: profilul datelor din sursele de date insuficient (necorespunzător); selecția necorespunzătoare a uneltelor de prelucrare (automatizare); analiză structurală insuficientă a surselor de date în etapa de stabilire a profilului datelor; metadate incerte (nesigure) şi incomplete.

c. Probleme de calitate a datelor în etapa ETL (Extragere, Transformare şi Incărcare). Etapa de ***extracție, transformarea*** şi ***încărcare*** este una foarte importantă, deoarece responsabilitatea maximă a eforturilor privind calitatea datelor revine stocării datelor. Procesul de curățare a datelor este executat în etapa de date, pentru a îmbunătăți precizia depozitului de date. Problemele care pot afecta calitatea datelor, în etapa ETL, sunt: arhitectura depozitului de date; legătura relațională sau non-relațională din depozitul de date; regulile de afaceri aplicate asupra surselor de date; lipsa actualizării periodice a datelor integrate.

d. Probleme de calitate a datelor în etapa schemei de proiectare. Calitatea datelor depinde de trei lucruri: calitatea datelor înseşi, calitatea programului de aplicație şi calitatea schemei bazei de date. O atenție deosebită, în timpul proiectării schemei, are în vedere unele aspecte, de exemplu, dimensiunile ce se schimbă lent/rapid, dimensiuni multivaloare etc. din depozitele de date. O schemă greşit proiectată are efecte negative asupra calității datelor. Problemele privind calitatea datelor în faza de modelare a schemei sunt date de: analiza incompletă sau greşită a cerințelor pentru schema de proiectare; selecția modelării dimensionale, în scheme de tip stea sau fulg de nea; dimensiunile multivaloare, identificarea tardivă a dimensiunilor cu schimbare lentă etc.; identificarea incompletă şi eronată a evenimentelor. În etapa de concepție a depozitului de date se folosesc modele dimensionale care grupează datele din tabelele relaționale în scheme de tip stea sau fulg de zăpadă. În aceste scheme pot fi regăsite date cantitative (cantități/ valori) sau grupate după diverse criterii. Datele cantitative din bazele de date dimensionale sunt de tip medii, număr de tranzacții, centralizări după anumite caracteristici, totaluri şi reprezintă măsuri ale activității.

e. Probleme de calitate a datelor în depozitele de date. *O definiție a calității* datelor se referă la datele greşite - date care lipsesc sau care sunt incorecte şi nevalide într-un anumit context. Calitatea datelor este atinsă atunci când organizația foloseşte date complete/cuprinzătoare, uşor de înțeles, coerente, relevante şi la timp. Înțelegerea dimensiunii cheie privind calitatea datelor, reprezintă primul pas pentru îmbunătățirea calității lor. Dimensiunile calității datelor pot include: corectitudinea, fiabilitatea, importanța, coerența, precizia, promptitudinea, finețea, claritatea, concizia şi utilitatea. Astfel, calitatea datelor este dată de următoarele dimensiuni cheie: completitudine, consistență, conformitate, validitate, acuratețe şi integritate.

2.3 Structura depozitului de date pentru a obține calitatea datelor. Vom pleca de la definiția depozitului de date, dată de W. H. Inmon (2002, p.8) „un **depozit de date** este o colecție de date orientate pe subiecte, integrate, istorice şi nevolatile destinată sprijinirii procesului de luare a deciziilor manageriale". Astfel, un depozit de date este o colecție de tehnologii care permite celor care lucrează cu informații să ia decizii mai bune şi mai rapide. Practica a demonstrat că sistemele de prelucrare online a tranzacțiilor (OLTP) nu sunt cele mai potrivite pentru a sprijini deciziile şi rețelele de mare viteză şi nu pot rezolva singure problema accesibilității informațiilor. Depozitul de date a devenit o strategie importantă pentru integrarea mai multor surse de informații eterogene în organizație şi pentru a permite Prelucrarea Analitică Online a Tranzacțiilor (OLTP). Calitatea datelor din depozitul de date este dependentă de modelele semantice ale arhitecturii depozitului de date. Deoarece datele stocate în depozitele de date sunt prelucrate analitic, cadrul de constituire a depozitului de date se confruntă cu 2 întrebări esențiale:

−Cum armonizăm fluxul de intrare al datelor ce provin din mai multe surse eterogene?

−Cum personalizăm (customizăm) stocarea datelor pentru diferite aplicații?

Decizia de proiectare a depozitului de date depinde de nevoile businessului ; prin urmare, în proiectarea depozitul de date, managementul schimbării joacă un rol important.

Obiectivele privind calitatea depozitului de date sunt: meta-bazele de date ce conțin modele formale privind calitatea informațiilor pentru optimizarea proiectării depozitului de date (adaptivă şi cantitativă); modelarea resurselor informaționale; modelarea schemei depozitelor de date pentru ca proiectanții şi cei ce optimizează interogarile, să beneficieze în mod explicit de dimensiunea temporală, spațială şi agregată a datelor din depozitul de date. Pe parcursul întregii etape de modelare a depozitului de date, proiectarea şi dezvoltarea este axată pe: cadrul calității depozitului de date şi arhitectura sistemului; modelarea metodelor, limbajul şi proiectarea depozitului de date; optimizarea. Pentru realizarea acestor obiective se proiectează un model de calitate pentru date, folosit pentru date istorice şi agregate.

2.8 Arhitectura calității depozitului de date. Pe parcursul interpretării datelor depozitului de date, acestea sunt analizate din perspectivă *conceptuală, logică* şi *fizică*. În proiectarea, realizarea şi modificarea depozitului de date, este important ca cele trei perspective să fie menținute, deoarece factorii de calitate sunt asociati cu perspectiva specifică sau cu relația dintre perspective. **Perspectiva conceptual -** în termenii calității depozitului de date, modelul conceptual defineşte teoria organizației. Cu privire la această teorie, *observare reală* poate fi evaluată din perspectiva factorilor de calitate, precum ***acuratețea, oportunitatea, exhaustivitatea***. În urma parcurgerii acestei etape se obține ***modelul sursă***. **Perspectiva logică** concepe depozitul de date din punctul de vedere al modelelor de date reale implicate. Cercetătorii şi practicienii care urmează această perspectivă iau în considerare faptul că depozitul de date este o colecție de vizualizări (înregistrări) materializate, extrase din surse de informații existente. În urma parcurgerii acestei etape se obține ***schema sursei***. **Perspectiva fizică** interpretează arhitectura depozitului de date ca o rețea de depozite mai mici de date, generatoare de date şi canale de comunicații, vizând factorii de calitate: ***fiabilitate*** şi ***performanță***, în prezența unor cantități foarte mari de date cu modificare lentă. În urma parcurgerii acestei etape se obține ***stocarea datelor sursă*** (Palepu and Rao, 2012).

2.9 Instrumente privind calitatea datelor. Studiile arată că 75% din efortul cheltuit cu depozitul de date este atribuit problemelor legate de: pregatirea datelor şi transportul (transmiterea) lor în depozitul de date. În prezent, sunt disponibile instrumente pentru a automatiza activitățile asociate cu ***auditarea, curățirea, extragerea*** şi ***încărcarea*** datelor în depozitele de date. Aceste instrumente sunt folosite pentru: a audita datele la sursă, a transforma datele, astfel încât acestea să fie consistente în depozitul de date, a segmenta datele în unități atomice şi a asigura că datele corespund regulilor de afaceri. Instrumentele pot fi pachete independente sau pot fi integrate cu pachetele depozitelor de date.

Instrumentele ce vizează calitatea datelor, de obicei, se încadrează în una din cele trei categorii: **auditare, curățare** sau **extracție/migrare.** Astfel, principalul obiectiv urmărit este de curățare şi de auditare al datelor, instrumentele de extracție şi de migrare a datelor având pondere limitată.

a. Instrumentele de audit al datelor sporesc acuratețea şi corectitudinea datelor la sursă. Aceste instrumente compară datele din baza de date sursă cu setul de reguli de afaceri. Când se utilizează o sursă externă organizației, regulile de afaceri sunt precizate de sursa externă organizației. În acest caz, ele sunt stabilite utilizând tehnicile **data mining** şi permit obținerea modelului datelor. Regulile de afaceri, interne organizației, sunt introduse în etapa de evaluare a surselor de date, iar analiza lexicală este utilizată pentru a descoperi sensul acestora.

b. Instrumentele de curățare a datelor sunt folosite în etapa intermediară. Aceste instrumente asigură: analiza, standardizarea şi verificarea datelor pe baza unor liste cunoscute.

c. Instrumentul de migrarea datelor este utilizat în extragerea datelor dintr-o bază de date sursă şi migrarea lor într-o zonă de stocare intermediară. Instrumentele de migrare transferă datele din zona de transfer în depozitul de date. Acestea asigură convertirea datelor de la o platformă la alta, iar utilitarul de migrare mapează datele de la sursă la depozitul de date, verificând şi conformitatea respectiv efectuarea de activități de curățare simplă a datelor.

2.10 Standardele de calitate a datelor. Standardele de calitate a datelor sunt utile pentru a evalua dacă cerințele specificate de către clienți sunt atinse sau nu. Obiectivele standardelor de calitate a datelor se referă la: *accesibilitate, corectitudine, oportunitate, integritate, valabilitate, coerență, relevanță.* Standardele de calitate a datelor sunt destinate: luării de decizii bazate pe fapte; luării de măsuri corective; accesării surselor problemelor referitoare la calitate; estimării pierderilor datorate lipsei de calitate precum şi estimării sau calculării soluțiilor specifice scopului urmarit oferind calitate acestora. Măsurătorile de calitate a datelor trebuie întotdeauna să respecte reglementările legale, regulile de afaceri precum şi cazurile speciale. Măsurătorile trebuie să îndeplinească următoarele criterii pentru a asigura calitatea informațiilor: toate măsurătorile trebuie să raporteze frecvent factorii de control ai calității dintr-o organizație; să definească metode care sunt folosite pentru a controla măsurătorile; fiecare metrică implică anumite atribute: numele metricii, definiția metricii, calculul, elementele de date şi sursa datelor.

2.11 Sistemul de control al calității metadatelor. Pentru a obține date de nivel înalt, de calitate în depozitul de date, se pleacă de la cerințele formulate de client, aspecte din cadrul organizației, structura organizatorică, funcționalități şi responsabilitățile privind îmbunătățirea continuă a calității. Acest lucru este cunoscut ca Managementul calității totale (TQM). Controlul evaluarii calității ține cont de: ***calitatea planificării*** – se selectează criteriile de evaluare a calității, se clasifica şi se atribuie priorități pentru activitățile de control; şi ***măsurarea cantitativă a calității***. Metadatele sistemului de control al calității vor fi colectate şi apoi transformate în specificații, cum ar fi: cerințele de informare, cererile de calitate ale utilizatorilor etc. **Sistemul de control al calității metadatelor** (MQC) cuprinde: arhitectura întregului depozit de date, dar şi calitatea acestora ce este măsurată pe tot fluxul datelor. Arhitectura sistemului MQC este prezentată în figura 1. Operațiunile de extracție, transformare şi încărcare (ETL) a datelor sunt importante/esențiale în formarea depozitului de date. Astfel, primează calitatea datelor în depozitul de date, alături de extragerea informațiilor din diverse baze de date mai mici, operaționale şi transformarea acestora în modelul schemei depozitului de date, urmată de încărcarea datelor în depozitul de date.

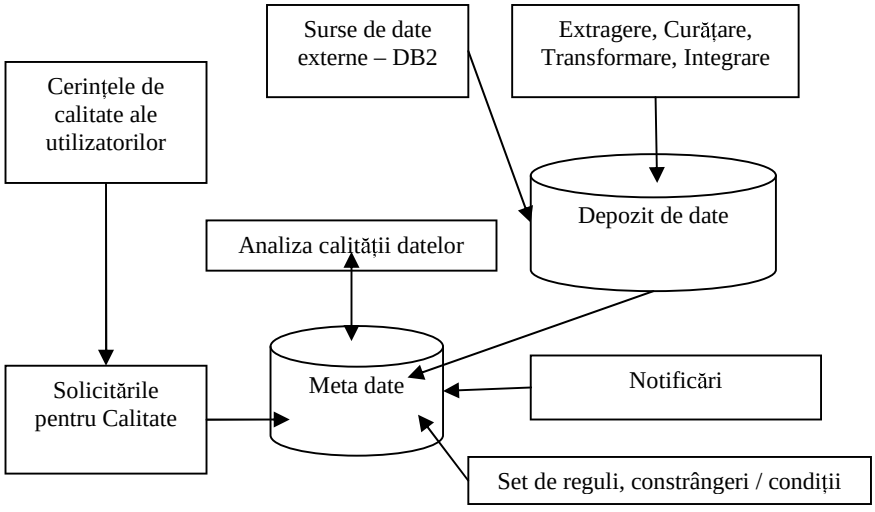

Figura 1. Arhitectura sistemului MQC

În timpul acestui proces una din componentele cheie se referă la Metadate. Ele stochează caracteristicile elementelor datelor încărcate în depozitul de date. Screeningul Metadatelor, aplicarea standardelor de calitate, identificarea problemelor controlului calității, conduc la remedierea problemelor. Modificările propuse în arhitectura depozitului de date conform standardelor de control al calității se regăsesc în Metabazele de date. Aceasta duce la sincronizarea operațiunilor Metabazelor de date cu arhitectura depozitului de date.

Concluzie: Sistemul de control al calitații metadatelor (MQC) asigură depăşirea problemele legate de calitatea datelor din depozitele de date la fiecare nivel de formare a depozitului de date. **Problemele referitoare la calitatea datelor** în depozitele de date sunt rezolvate prin: planificare, folosind criteriile de evaluare a calității datelor, prin managementul calității totale a datelor, folosind clasificarea problemelor legate de calitatea datelor în depozitele de date, avaluând structura depozitului de date cu privire la calitate, folosind instrumente referitoare la calitatea datelor, standardele de control a calității şi sistemul de control al calității metadatelor.

Bibliografie

1. W. H.Inmon, (2002), “Building the Data Warehouse”, John Wiley & Sons, San Francisco

2. Ramesh Babu Palepu, Dr K V Sambasiva Rao, (2012), “Meta data quality control architecture in data warehousing”, *International Journal of Computer Science, Engineering and Information Technology* (IJCSEIT), Vol.2, No.4, pp. 15-24

3. Ranjit Singh, Kawaljeet Singh, (2010), “A descriptive classification of causes of data quality problems in data warehouse”, *International Journal of Computer Science Issues*, Vol 7, pp. 23-29