Data Quality Costs

Lt.col. Otilia PÎRLOG
Ministerul Apărării Naționale

Every attempt to optimize the data quality must begin with an evaluation of the current state, including the identification of the possibilities of correcting the signaled errors. One of the main problems in data quality management is the evaluation of the costs after eliminating the main problems of the data quality and comparing these costs with the previous costs. The data quality diagram offers a reliable technique, easy to handle, through which quality programs can be initiated and accomplished. It consists of sequentially going through some steps through which a data profile can be deducted and the flow of data can be analyzed. These steps form a flexible tool, the final result being a set of related documents for evaluating the costs and quality of the data.
Keywords: *data, quality, management, flow, cost.*

Evaluarea costurilor pe activități

Demersurile de îmbunătățire a calității datelor gestionate la nivelul unei organizații, respectiv de evaluare a costului calității acestora, eşuează de cele mai multe ori datorită noutății şi complexității problematicii vizate. Asemănător costurilor calității scăzute ale produselor şi serviciilor, costurile calității datelor se răsfrâng asupra unor arii diverse, nu neapărat asociate direct datelor însăşi. Astfel, calitatea proastă a datelor se poate manifesta printr-o creştere a dificultăților sistemului. Drept rezultat, este blamată proiectarea sistemului sau personalul implicat în introducerea datelor în sistem.

Odată ce aceste costuri au fost identificate, se poate înțelege mai bine cât costă corectarea problemelor legate de calitatea datelor şi care sunt costurile induse de ignorarea acestei probleme. În acest mod se reliefează costurile directe şi indirecte asociate cu calitatea slabă a datelor.

Costul calității datelor. Fie $AQ=(a_1,a_2,...a_n)$ setul celor n activități desfăşurate în scopul asigurării calității datelor. Costul calității datelor, CQ, reprezintă suma costurilor înregistrate pentru toate cele n activități vizate. Cuantificarea costului activității, CA_i, se face prin valoarea cumulată a costurilor resurselor utilizate asociate cu activitatea i. Pentru o activitate dată, a_i, sunt folosite şi / sau consumate la un moment dat un număr de m resurse.

Fie $RQ=(r_1, r_2,...r_m)$ setul celor *m* resurse implicate în activitățile specificate de asigurare a calității datelor. Prin CR_{ij}, specificăm costul, cuantificat în unități monetare, al consumului cantității *q* din resursa r_j alocată pentru activitatea a_i la momentul de timp specificat *t*.

Deci, CR_{ij} este funcție de variabilele a_i, r_j, *q* şi *t*, unde:

- a_i reprezintă activitatea *i* desfăşurată pentru asigurarea calității datelor;
- r_j resursa *j* folosită sau consumată de către activitatea a_i;
- *q* cantitatea din resursa r_j care este necesară pentru activitatea a_i;
- *t* momentul de timp la care se face cuantificarea.

Consumul resursei *r*, deşi este evaluat la un moment *t*, se realizează într-un interval de timp bine determinat, $Ti = [t_s, t_e]$, în care o cantitate q_{ij} din resursa r este folosită sau consumată de către activitatea a. Explicitarea consumului de resurse în timp duce la formularea completă a valorii $CR_{ij} =f(a_i, r_j, q_{ij}, t, t_s, t_e)$. Este deci necesară definirea unor costuri unitare ale consumării sau utilizării unei cantități unitare dintr-o resursă r_j de către o activitate a_i, aflată în starea s_k. Corespunzător naturii activităților pentru care se face alocarea resurselor (de execuție, activare, dezactivare şi reactivare), costurile unitare aferente vor fi codificate prin Cex_{ij}, $Cact_{ij}$, $Cdact_{ij}$ şi $Cract_{ij}$. Costul consumului din resursa r_j de către o

activitate a_i se obţine prin însumarea tuturor categoriilor de costuri specificate, deci:
$CR_{ij}=q_{ij} * (Cex_{ij} + Cact_{ij} + Cdact_{ij} + Cract_{ij})$
Calcularea costului unei activităţi reprezintă agregarea costurilor resurselor utilizate de-a lungul intervalului de timp de desfăşurare al activităţii, $Ti= [t_s, t_e]$, pentru toate stările pe care activitatea le parcurge în acest interval $CA_i=\sum_{j=1}^{m} CR_{ij}$. Costul total al asigurării calităţii datelor cumulează costurile aferente tuturor celor *n* activităţi implicate $CQ = \sum_{i=1}^{n} CA_i$.

Calitatea şi costul datelor curente

Diagrama calităţii şi a costului calităţii datelor oferă o tehnică viabilă, deosebit de accesibilă, prin care se pot iniţia şi implementa programe de calitate. Funcţie de complexitatea organizaţiei, acest model poate viza abordarea completă a acesteia sau pe module funcţionale, urmată de integrarea rezultatelor parţiale obţinute.
Un aspect de o importanţă deosebită îl constituie posibilitatea formalizării informaţiilor obţinute şi realizarea pe baza lor a unor scenarii diverse de îmbunătăţire a calităţii. Rezultatele obţinute pot fundamenta luarea deciziilor referitoare atât la calitatea cât şi la ameliorarea costurilor aferente datelor. Etapele majore ale modelului şi problematica conexă vizată sunt prezentate în figura 1.
Primul pas în iniţializarea unui program de îmbunătăţire a calităţii datelor este *evaluarea calităţii datelor curente*. În acest mod se facilitează identificarea zonelor care este necesar să se situeze în prim-planul demersului de îmbunătăţire.
Definirea stării curente a calităţii datelor necesită stabilirea unui profil de date şi analizarea circulaţiei acestora. Rezultatele acestor procese sunt revăzute în contextul conformităţii lor cu cerinţele reale de pe traiectoria fluxului prelucrărilor/informărilor din interiorul organizaţiei.
Deşi poate fi inclusă în categoria operaţiilor care se sprijină pe un suport tehnic adecvat, *crearea profilului datelor* este o analiză deosebit de dificilă. În general, rezultatele parcurgerii acestei etape vor evidenţia meta-date despre datele investigate. Se continuă apoi prin analize simple, cum ar fi domeniul de valori, tipuri de date implicate, valorile minime şi maxime şi distribuţia în funcţie de frecvenţă. Profilarea pune de asemenea accentul pe analize mai detaliate, cum ar fi relaţii între coloane şi dintre tabele.
Analiza circulaţiei datelor reprezintă cartografierea modului în care datele circulă în interiorul organizaţiei. Deoarece informaţiile sunt folosite atât pentru procesarea tranzacţiilor cât şi pentru diverse alte procesări analitice, datele pot fi modificate atunci când trec de la un stadiu de procesare la altul. O hartă a circulaţiei datelor în interiorul organizaţiei permite izolarea stagiilor de procesare care contribuie la o calitate slabă a datelor.
Fiind evidenţiat un set de reguli de calitate a datelor şi o secvenţă de circulaţie a acestora, se poate stabili conformitatea setului de date cu cerinţele utilizatorilor în orice punct din sistemul analizat. De asemenea, se pregăteşte astfel cadrul general pentru parcurgerea următoarei etape de analiză, cea a *determinării costurilor asociate cu datele de calitate proastă*.
Regulile de analiză sunt utilizate pentru crearea unui model economic de evaluare a costurilor, asociate cu instituirea îmbunătăţirilor. Acest model poate fi privit ca un formular care documentează nivelurile de calitate a datelor asociate cu un set de dimensiuni ale calităţii datelor.

Formalizarea modelului

Etapele enunţate anterior constituie un suport flexibil de formalizare, rezultatul final constituindu-l un set de documente corelate de evaluare a calităţii şi costului calităţii datelor. Aşa cum s-a arătat deja, funcţie de complexitatea sistemului analizat, analiza poate să cuprindă întregul context al datelor, sau se poate limita la module de interes ale acestora.
Datorită conexiunilor multiple care se identifică de-a lungul diverselor etape de analiză, s-a adoptat un sistem de simbolizare a modelului prin care să se sugereze variabilele care sunt puse în corelaţie la un moment dat. De

asemenea, domeniul de valori este conexat cu variabila corespunzătoare. Indicii elementelor matriceale rezultate în diverse stadii ale analizei sunt notați de fiecare dată prin simbolurile *i,j*, stabilind astfel numai caracterul bidimensional al acestora şi nu particularizarea lor la contextul respectiv.

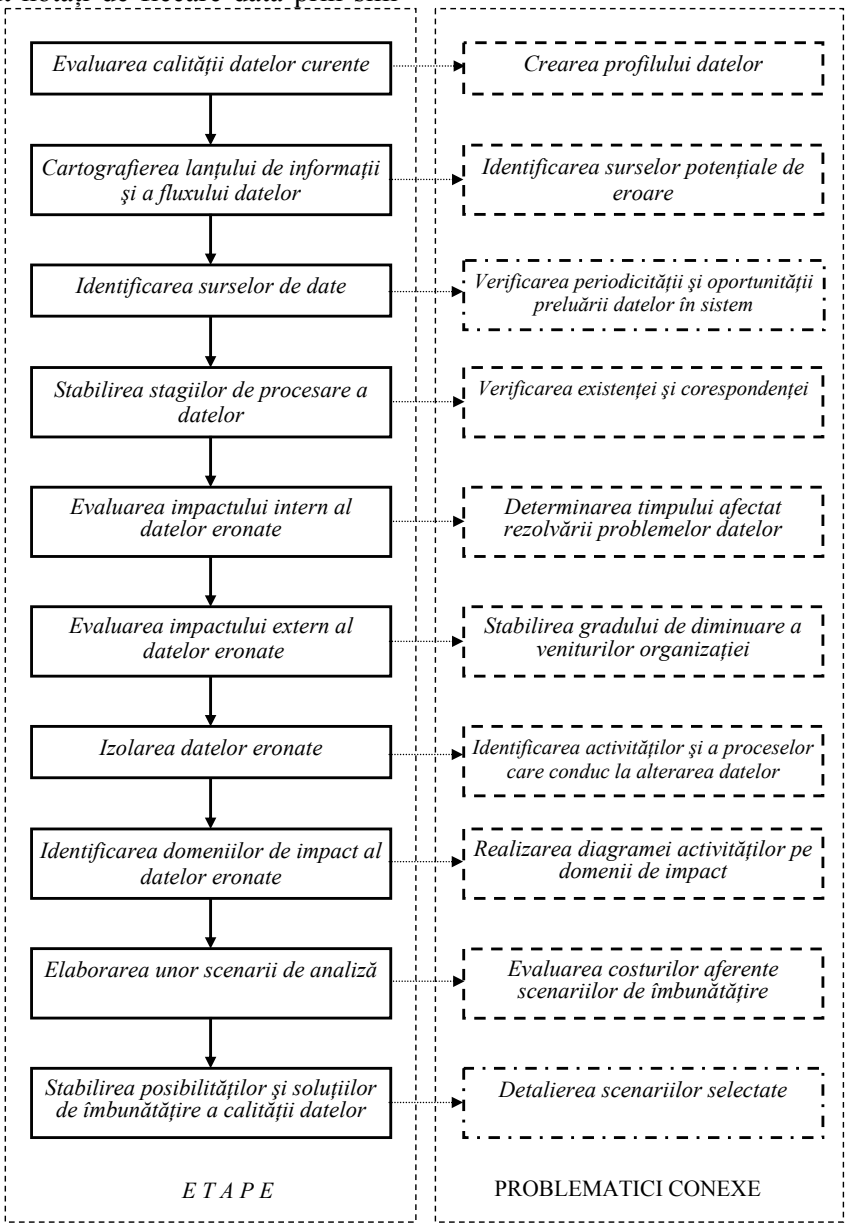

Fig.1. Diagrama evaluării calității şi costului calității datelor

Deoarece lucrarea de față are drept scop estimarea costului calității datelor, deşi modelul se constituie drept un instrument unitar al problematicii calitate-cost, vom evidenția numai elementele specifice costurilor. Prezentarea detaliată a unor elemente specifice fluxului de prelucrare se realizează în cadrul modelului procesării datelor.

Elaborarea *diagramei calității şi costului datelor* presupune parcurgerea următoarelor etape:

- identificarea surselor de date şi a datelor pe care sistemul le utilizează, precum şi a stagiilor de procesare aferente datelor; rezultatul parcurgerii acestei etape este completarea elementelor matriceale DS_{ij}, unde $D_1 \div D_d$ este setul datelor elementare identificate, iar $S_1 \div Ss$ reprezintă sursele de date;
- datele elementare aparțin unor structuri omogene, care pot avea semnificații diferite,

denumite simbolic unităţi de date; elementele matricei DU_{ij} analizează apartenenţa datelor unitare la unităţile de date, oferind în acest mod posibilitatea identificării corespondenţei dintre datele elementare şi activităţile de procesare a datelor care au impact asupra calităţii şi costului calităţii datelor;

- elementele matriceale DU_{ij} constituie de asemenea suportul de calcul al costurilor aferente unităţilor de date, costul datelor elementare ($C_1 \div C_d$) fiind repartizat asupra unităţilor de date conexate;
- formularul de analiză aferent acestei etape este prezentat în figura 2.

Date elementare	Surse de date			Unităţi de date			Costul datelor elementare
	S_1	.....	S_s	U_1	.....	U_u	
D_1							C_1
.....		DS_{ij}			DU_{ij}		.....
D_d							C_c
Costul unităţilor de date	-	- -	-	$CostU_1$		$CostU_u$	-

Fig.2. Matricea de analiză a surselor, apartenenţei şi costului datelor elementare

➢ cartografierea lanţului de informaţii şi a fluxului datelor; pe baza acestuia pot fi observate sursele potenţiale de eroare; matricea de analiză este prezentată în continuare (figura 3);

Unităţi de date	Activităţi asupra datelor / Timpi afectaţi calităţii datelor			Surse de eroare			Timpul identificării erorilor
	A_1	.....	A_a	Se_1	.....	Se_{se}	
U_1							$TimpIE_1$
.....		UA_{ij} / UT_{ij}			USe_{ij}		.....
U_u							$TimpIE_u$
Costurile datelor pe surse de eroare	UT_{1j}	- -	UT_{aj}	$CostSe_{1j}$		$CostSe_{se,j}$	$\sum TimpIE_j$

Fig.3. Matricea de analiză a fluxului datelor

➢ prin observarea detaliată a structurii mesajelor, se stabileşte corespondenţa dintre condiţiile de eroare şi datele specifice în care se manifestă erorile semnalate (figura 4);

Date elementare	Mesaje de eroare			Cauzele erorilor			Timpul corectării erorilor
	M_1	.....	M_m	E_1	.....	E_e	
U_1							$TimpCE_1$
.....		UM_{ij}			UE_{ij}		.....
U_u							$TimpCE_u$
Costurile datelor pe cauze de eroare	-	- -	-	$CostE_1$		$CostE_e$	$\sum TimpCE_j$

Fig.4. Matricea de analiză a mesajelor de eroare

➢ evaluarea impactului intern al datelor eronate, în urma discuţiilor purtate cu persoanele implicate în utilizarea acestora; se evaluează şi centralizează timpul *(TIMP1)* pe care toţi angajaţii îl dedică problemelor de calitate a datelor pentru fiecare stagiu din

fluxul datelor;
$TIMP1 = \sum UT_j + \sum TimpIE_j + \sum TimpCE_j$;

- pentru a înțelege motivele care au dus la diminuarea afacerilor, sunt contactați foşti şi actuali beneficiari; se identifică problemele datelor cu care aceştia s-au confruntat;
- se izolează datele eronate; diagrama de flux a datelor se adnotează cu rezultatele interviurilor anterioare; se notează sursa oricărei erori de date, la fiecare stagiu de procesare, împreună cu o listă a activităților care pot fi atribuite la aceste erori (figura 5);
- se identifică domeniile de impact; prin crearea unei matrice de conexiune (figura 5), se clasifică erorile şi activitățile pe domenii de impact; prima axă reprezintă problema şi locația ei în circuitul informației, a doua reprezintă activitățile asociate cu fiecare problemă, iar a treia denotă impacturile; în fiecare celulă din această matrice se stabileşte costul aproximativ asociat cu acel impact.

Problema identificată / Locația ei în fluxul datelor	Activități corective asociate			Domenii de impact			Timpul activităților corective
	Ac_1	.....	Ac_{ac}	Dom_1	.....	Dom_{dom}	
Pr_1							$TimpCor_1$
.....		$PrAc_{ij}$			$PrDom_{ij}$		.....
Pr_{pr}							$TimpCor_{pr}$
Costurile activităților corective	$CostAc_1$	- -	$CostAc_{ac}$	-	- -	-	$\sum TimpCor_j$

Fig.5. Matricea de analiză a activităților corective

Timpul afectat activităților corective se cumulează la timpul calculat deja al personalul implicat în procesarea / intreținerea calității datelor, *TIMP1*, rezultând timpul total aferent calității datelor.

$TIMP=\sum UT_j+\sum TimpIE_j+\sum TimpCE_j+\sum TimpCor_j$

- matricele se configurează pe un spreadsheet şi se construieşte un model de agregare în care, funcție de scopul analizei, costul poate fi cumulat în modalități diferite;
- se identifică posibilitățile de îmbunătățire a calității şi costului datelor.

Bibliografie

[Bech00] Bechtel Hanford Inc. Procedure: *Data Quality Objectives*, BHI-EE-01, Environmental Investigations Procedure: Procedure Number 1.2; Revision 3; September 30, 1999.

[Behe97] Beheiry, M.F.: *The Cost of Quality* [Online], Available: www.dbainc.com/dba2/library/cost.html [November 3,1997].

[Fari00] Farino, R., *The Data Quality Assessment*, Part 1, DM Review Online, August 2000, http://www.dmreview.com.

[LeeJ03] Lee, J., Wyner, G.M.: *Defining Specialization for Dataflow Diagrams*, Information Systems, Volume 28, Number 6, September 2003, 651-671.

[Pipi02] Pipino, L., Lee, Y.W., Wang, R.Y.: *Data Quality Assessment.* Communications of the ACM (CACM), Volume 45, 2002:211-218.

[Redm03] Redman, T.C.: *Ending 'Garbage In, Garbage Out': IT's Role in Improving Data Quality,* Guest Editor, Cutter IT Journal, January 2003, Vol. 16, No. 1.

[Wang01] Wang, R.Y., Ziad, M., Lee, Y.W.: *Data Quality*, Kluwer, 2001.