

27. MANAGEMENTUL CALITĂȚII DATELOR ȘI A STRUCTURILOR DE DATE

27.1 Volumul datelor

În mod tradițional, prin date se înțelege tot ceea ce face obiectul unor stocări, arhivări sau calcule. Datele nu sunt numai de tip numeric. Există, pe de o parte, realitatea, formată din obiecte, ființe, fenomene, acțiuni, interacțiuni, și pe de altă parte, datele cu ajutorul cărora sunt reflectate părți ale realității. Datele sunt rezultatul unor măsurători sau observații și fac obiectul stocării și/sau analizei.

Elementele de management a calității iau în considerare factorii care generează calitatea produselor și anume: personalul, materiile prime și procesele și tehnologiile de producție. Calitatea se planifică, se construiește și are costuri proprii asociate.

Lumea reală este formată din obiecte, ființe, plante, fenomene, acțiuni. Ea este percepută natural, prin intermediul culorilor, temperaturii, sunetelor, zgomotelor, presiunii, mirosurilor sau, convențional, prin intermediul mesajelor construite. Ea se află într-o mișcare continuă, se modifică de la un moment la altul. Aceasta corespunde trecerii de la o stare la alta.

Asupra datelor se efectuează operații având următoarele caracteristici, [Ivan99a]:

- adăugarea de date – pentru un șir inițial de date $D = \{d_1, d_2, \dots, d_n\}$, se consideră că adăugarea se efectuează fie la începutul fie la sfârșitul șirului și presupune adăugarea unuia sau a mai multor elemente;
- inserarea de date – pentru șirul de date $b_1, \dots, b_i, b_{i+1}, \dots, b_n$, elementul a este inserat între elementele b_i și b_{i+1} , rezultând un șir cu $n+1$ componente: $b_1, \dots, b_i, a, b_{i+1}, \dots, b_n$;
- ștergerea de date – pentru șirul format din elementele $b_1, b_2, \dots, b_{i-1}, b_i, b_{i+1}, \dots, b_n$, ștergerea elementului b_i conduce la obținerea șirului $b_1, b_2, \dots, b_{i-1}, b_{i+1}, \dots, b_n$;
- interschimbul de date - corespunde situației în care în șirul $b_1, b_2, \dots, b_{i-1}, b_i, b_{i+1}, \dots, b_{j-1}, b_j, b_{j+1}, \dots, b_n$, elementul b_i ia locul elementului b_j , șirul devenind $b_1, b_2, \dots, b_{i-1}, b_j, b_{i+1}, \dots, b_{j-1}, b_i, b_{j+1}, \dots, b_n$;
- modificarea datelor – este o operație complexă care se efectuează în mai multe moduri; una dintre modalitățile de efectuare este prin înlocuirea unei date cu alta;
- concatenarea datelor - se consideră șirurile $b_1, b_2, \dots, b_n$ și $c_1, c_2, \dots, c_m$. Prin concatenare rezultă șirul $b_1, b_2, \dots, b_n, c_1, c_2, \dots, c_m$;
- stocarea datelor – se realizează pe diferiți suporti de stocare: hârtie, metal, piatră, ceramică, sticlă, suporti magnetici.

Sistemele informatice se structurează funcție de obiective, complexitatea și volumul de calcule. Fundamentarea necesarului de personal, de echipamente și de consumabile este strict dependentă de volumul prelucrărilor, iar acesta este influențat decisiv de volumul de date (VD).

Se consideră o colectivitate A formată din elementele $a_1, a_2, \dots, a_n$, omogene în raport cu un criteriu (atribut). Elementele se identifică prin codurile unice $\alpha_1, \alpha_2, \dots, \alpha_n$.

Astfel, pentru atributul *materiale*, colectivitatea A se va numi *colectivitatea materialelor*. Pentru atributul *muncitori*, colectivitatea A se va numi *colectivitatea muncitorilor*.

Atributul are rol de criteriu de selecție. Dacă se compun attribute cu operatorul logic "ȘI", are loc creșterea gradului de omogenitate a elementelor ce compun colectivitatea. Dacă se compun attribute cu operatorul logic "SAU", are loc creșterea gradului de neomogenitate.

Astfel, pentru atributul "*student*", colectivitatea A conține elementele $a_{i,j,k}$ unde indicele i arată anul de studiu, indicele j arată grupa, iar indicele k specifică poziția studentului în cadrul grupei.

Compunerea atributului "*student*" cu atributul "*an_terminal*" prin operatorul "ȘI" conduce la constituirea subcolectivității A^1 formată din elementele:

$$\{a_{5,m,1}, a_{5,m,2}, \dots, a_{5,m,p_1}, a_{5,m+1,1}, a_{5,m+1,2}, \dots, a_{5,m+1,p_2}, a_{5,m+n,1}, a_{5,m+n,2}, \dots, a_{5,m+n,p_{n+1}}\} \quad (27.1)$$

Gradul de omogenitate al colectivității A^1 este mai ridicat, aceasta fiind formată din elemente care au două caracteristici identice, celelalte fiind diferite de la element la element.

Considerând și colectivitatea B a elevilor formată din componentele $b_1, b_2, \dots, b_t$, compunerea atributului "*student*" prin operatorul "SAU" cu atributul "*elev*" conduce la obținerea unei colectivități mai largi, cu grad de neomogenitate mai mare, formată din elementele $\{a_{1,1,1}, \dots, a_{5,m+n,p_{n+1}}, b_1, b_2, \dots, b_t\}$. Elementele colectivității A se descriu cu ajutorul unor caracteristici de calitate de diferențiere $C_1, C_2, \dots, C_r$.

În acest fel, se construiesc tabele în care se înscriu datele de identificare ale elementelor colectivității. Pe coloane se înscriu caracteristicile cu nivelele măsurate, observate sau declarate pentru fiecare element în parte.

Mulțimea caracteristicilor considerate pentru o colectivitate trebuie să asigure o descriere completă a elementelor și în același timp să nu includă caracteristici care nu sunt importante în raport cu situația analizată. Completitudinea caracteristicilor nu este o proprietate absolută, ea fiind definită în raport cu un anumit scop.

Operația de stabilire a caracteristicilor semnificative ale colectivității A în raport cu scopul propus dă posibilitatea ca într-o etapă ulterioară să se proiecteze structura fișierelor.

Volumul de date reprezintă expresia cantitativă a datelor cu care se descrie o colectivitate, [Ivan99a]. În practica de zi cu zi a manipulării bazelor de date, a fișierelor sau în interacțiunea curentă om/calculator, volumul de date este apreciat prin:

- lungimea fișierului, exprimată fie ca număr de articole, fie ca număr de baiți;
- număr de documente introduse în fișier sau în baza de date;
- număr de tranzacții;

- timp de prelucrare, în cazul existenței unui grad foarte mare de omogenitate a prelucrărilor, iar informațiile ce intervin în procesul de regăsire nu sunt transparente.

Fiecare dintre aceste moduri de exprimare a volumului de date este afectat de erori care reduc gradul de operaționalitate. Astfel, atunci când volumul de date este exprimat ca număr de articole (înregistrări), lipsește elementul de bază, informația referitoare la structura unui articol.

Lungimea fișierului, exprimată ca număr de baiți este afectată de convenție referitoare la definirea și utilizarea câmpurilor de lungime maximă. Astfel, la definirea structurii fiecărui câmp se analizează domeniul valorilor pe care caracteristica asociată le poate avea. Lungimea câmpului este astfel concepută încât să se memoreze orice valoare de descriere a caracteristicilor.

Lungimea L_a a articolului, pentru descrierea unui element al colectivității, este dată de relația:

$$L_a = \sum_{i=1}^r l_{c \max}^i \quad (27.2)$$

unde:

- r – numărul de câmpuri asociate caracteristicilor $C_1, C_2, \dots, C_r$;
- $l_{c \max}^i$ – lungimea șirului maxim de descriere a câmpului i .

Rezultă că lungimea reală necesară descrierii unui element oarecare - L_{ra}^j este în cele mai multe cazuri cu mult mai mică decât lungime fizică a articolului respectiv - L_a^j .

Volumul fișierului ca număr de baiți este dat de relația:

$$V = n \cdot L_a \quad (27.3)$$

unde:

- n – număr de articole;
- L_a – lungime fizică a unui articol.

În cazul în care se produce un proces de normalizare cu introducerea unui delimitator, volumul modificat V^1 al fișierului ca număr de octeți este dat de relația:

$$V^1 = \sum_{j=1}^n L_{ra}^j + (n-1)(r-1) \quad (27.4)$$

unde s-a considerat că se alocă un bait pentru delimitator, în condițiile în care se asigură unicitatea acestuia.

Dacă se iau în considerare elementele de variabilitate, este necesar să se introducă specificatori de lungime de câmp, figura 27.1.

l_1	a_1	l_2	a_2	...	l_n	a_n
-------	-------	-------	-------	-----	-------	-------

a)

câmp ₁	d.	câmp ₂	d.	...	d.	câmp _r
-------------------	----	-------------------	----	-----	----	-------------------

b)

Figura 27.1 Fișiere cu articole de lungime oarecare:

a) structura fișierului pe articole;

b) structura unui articol (d.=delimitator).

Volumul de date este exprimat prin formula:

$$V = n \cdot w + \sum_{j=1}^n LRA_j \quad (27.5)$$

unde:

- LRA_j - lungimea reală a articolului j, incluzând și delimitatorii dintre câmpuri: $LRA_j = L_{ra}^j + (r-1)$;
- w - numărul de baiți necesar pentru memorarea lungimii reale a fiecărui articol în parte; se consideră că în cele mai multe cazuri w este un bait.

Cu cât numărul de baiți necesari reprezentării articolului de lungime maximă este mai mare, cu atât dispersia lungimii reale a articolelor este mai mare. Informația necesară codificării lungimii articolelor reale scade ca pondere în volumul total al articolului. E avantajos, în asemenea cazuri, să se folosească structuri de lungime variabilă, pentru de a ocupa un spațiu de memorie cât mai redus.

În cazul în care baitul de lungime asociat unui articol își schimbă dintr-un motiv oarecare valoarea, întregul memorat în baitul respectiv diferă de lungimea reală, în baiți, a articolului ce-i urmează, atunci întreaga informație din acel punct până la sfârșitul fișierului își pierde semnificația, datorită nesincronizării.

Există posibilitatea folosirii, pe lângă informația de codificare a lungimii articolului, a unor baiți de delimitare (separare) între articole. Această structură este redundantă, dar oferă în schimb unele avantaje. Astfel, se realizează o parcurgere bidirecțională a fișierului iar în cazul în care unul din baiții de lungime își schimbă nemotivat valoarea (sau nu poate fi citit), se evită pierderea tuturor articolelor până la sfârșitul fișierului, prin citirea bait cu bait și identificarea delimitatorului, în condițiile în care s-a asigurat unicitatea lui.

Numărul de tranzacții, numărul de documente introduse sau obținute aproximează volumul de date în măsura în care nevoia de comparabilitate este strict locală. Se consideră o aplicație contabilă și momentele de timp $t_1, t_2, \dots, t_s$. Fie τ_i intervalele de timp corespunzătoare momentelor t_{i+1}, t_i , $\tau_i = t_{i+1} - t_i$, $i=1, 2, \dots, s-1$, în care se prelucrează de către un operator un număr de facturi, respectiv $f_1, f_2, \dots, f_{s-1}$. Întrucât documentele sunt omogene, numărul lor este semnificativ pentru a compara volumul de date. Dacă se notează cu $\bar{x}$ volumul mediu de date conținut într-o factură, volumul de date prelucrate în intervalul τ_i este dat de relația:

$$V_i = f_i \cdot \bar{x} \quad (27.6)$$

Comparația dintre volumele prelucrate în zilele i și j revine de fapt la a compara frecvențele (numărul de facturi):

$$c_{ij} = \frac{V_i}{V_j} = \frac{f_i \cdot \bar{x}}{f_j \cdot \bar{x}} = \frac{f_i}{f_j} \quad (27.7)$$

unde c_{ij} reprezintă volumul de date din intervalul i relativ la intervalul j .

În scopul obținerii unei informații asupra datelor prelucrate în ziua i în contextul general al celor $s-1$ zile, se poate defini un indice C_i , reprezentând ponderea volumului de date prelucrate în intervalul i relativ la întregul volum de date prelucrate:

$$c = \frac{V_i}{\sum V_j} = \frac{f_i \cdot \bar{x}}{\sum f_i \cdot \bar{x}} = \frac{f_i}{\sum f_i} \quad (27.8)$$

În fundamentarea deciziilor privind achiziționarea de tehnică de calcul și de structurare a sistemului informatic este necesară estimarea volumului de date de prelucrat. Definirea corectă a sistemului de ipoteze pentru estimare asigură calitatea deciziei luate privind structura și nivelul tehnicii de calcul achiziționate.

Estimarea volumului de date este preferată calcului, întrucât în faza de proiectare a sistemului informatic se alocă resurse altor activități în schimbul măsurării de nivele care intră în calculul volumului exact de date, care este caracterizat la nivelul agenților economici prin dinamică anuală reală.

Dacă în cazul materiilor prime, produselor finite, operațiilor, numărul de articole pe care le conține fișierul este suficient de riguros estimat, în cazul numărului de furnizori, numărului de contracte și numărului de comenzi estimările sunt afectate de un nivel al erorilor suficient de mare.

Pentru acest exemplu volumul de date este dat de relația:

$$V = \sum_{i=1}^m \lg(F_i) \quad (27.9)$$

unde:

- $\lg(.)$ – o funcție care returnează lungimea unui fișier;
- m – numărul de fișiere.

În condițiile asocierii unor structuri de tip listă, arbore sau a unor structuri complexe, nivelul de complexitate impune ponderarea volumului V al datelor de bază:

$$V_{bd} = V \cdot f(k) \quad (27.10)$$

unde:

- V_{bd} – volumul care constituie baza de date incluzând atât datele de bază cât și datele care permit dezvoltarea unor tehnici rapide de regăsire;

- $f(.)$ – funcție de ponderare;
- k – nivelul complexității structurilor asociate.

Construirea de modele pentru estimarea volumului de date este necesară pentru a defini concepția specifică fiecărei aplicații pentru managementul datelor. Calitatea rezultatelor, chiar funcționarea unui sistem informatic este esențial legată de precizia cu care a fost estimat volumul prelucrărilor și de soluțiile adoptate.

Odată cu estimarea volumului de date se obțin estimări pentru:

- necesarul de tehnică de calcul și caracteristicile acesteia;
- numărul de persoane pentru încărcarea bazei de date;
- numărul de posturi de lucru.

Cunoscând ponderea volumului de actualizare (30%) rezultă volumul de personal angrenat în actualizare, deci pentru exploatarea curentă a sistemului, [Ivan99a].

27.2 Caracteristicile de calitate a datelor

Calitatea este unul dintre cei mai importanți indicatori pentru evaluarea competitivității unei activități, [Ivan99a]. Calitatea nu poate crește fără concentrarea conducerii asupra aspectelor calității. Managementul calității presupune în primul rând luarea deciziilor pe baza unei cunoașteri clare și foarte corecte a realității. Aceasta implică folosirea cunoștințelor, informațiilor și datelor. Este necesar să se asigure calitatea datelor pe parcursul întregului ciclu de viață a acestora.

Managementul total al calității datelor este "un proces continuu de îmbunătățire a proceselor de dezvoltare care definesc datele și procesele economice care crează/actualizează/șterg datele și regăsesc/prezintă informațiile cu ajutorul conceptelor, principiilor și metodelor din cultura organizațională a unei firme", cu o prioritate pe termen lung asupra prevenirii noncalității datelor.

Avantajul competitiv poate rezulta din îmbunătățirea calității datelor. Datele slabe din punct de vedere calitativ pot avea un impact economic și social substanțial, prin impunerea unor costuri enorme și de cele mai multe ori ascunse asupra operațiilor care au drept suport acele date. Există un acord general asupra faptului că datele pot fi extrem de costisitoare, deoarece datele sunt de cele mai multe ori cele mai importante resurse ale unei organizații. Calitatea ridicată nu doar reduce costurile, dar crește și disponibilitatea și bunăvoința clientului de a plăti pentru bunurile și serviciile oferite, ceea ce duce la îmbunătățirea poziției competitive.

Datele noncalitative există sub diverse forme, inclusiv:

- date redundante, determinând costuri de producție ridicate;
- date incomplete sau neactualizate, determinând alterarea imaginii firmei;
- date slab definite, ceea ce duce la folosirea eronată a informației.

Primul pas în îmbunătățirea calității datelor este înțelegerea dimensiunilor calității datelor și importanța lor relativă. Deși corectitudinea și completitudinea datelor sunt în mod intuitiv foarte importante, există un număr de alte caracteristici ale calității datelor care trebuie luate în considerare.

Pornind de la relativ limitata literatură relevantă existentă și de la conceptul *fitness for use* din literatura de specialitate a calității, se propune în continuare un cadru de lucru preliminar pentru calitate datelor, care include următoarele aspecte:

- datele trebuie să fie *accesibile* clienților – clientul știe cum să regăsească datele;
- clientul trebuie să poată *interpreta* datele – datele nu sunt prezentate într-o codificare necunoscută utilizatorului;
- datele trebuie să fie *relevante* pentru client – datele sunt relevante și oportune pentru folosirea de către consumatorul de date în procesul de luare a deciziilor;
- clientul trebuie să aprecieze datele ca fiind *corecte* – datele sunt corecte, obiective și provin dintr-o sursă cu reputație.

Se definește o *dimensiune a calității datelor* ca fiind un *set de attribute de calitate* ale datelor la care cei mai mulți consumatori de date reacționează într-un mod destul de consistent. Este surprinzător faptul că ceva atât de simplu cum este o *dată* are atât de multe aspecte din punct de vedere al calității. În activitatea practică, s-au identificat peste 150 de attribute ale calității datelor, acestea fiind prezentate în ordine alfabetică în tabelul 27.1, [Ivan99a]. Numărul acestor attribute este mult mai mare și depinde de specificul activității care are ca suport datele analizate.

Tabelul nr. 27.1 Attribute de calitate ale datelor

acceptabilitate	accesibilitate	actualitate
acuratețe	adaptabilitate	agregabilitate
alterabilitate	amplitudine	aspect estetic
auditabilitate	auto-corectabilitate	autoritate
bine documentate	bine prezentate	cantitate
capacitate de descărcare	capacitate de identificare erori	capacitate de încărcare
certificare	claritate	claritate surse
compactivitate	compatibilitate	compatibilitatea cu date istorice
competitivitate	completitudine	comportament
compresabilitate	comprihensivitate	concizie
conexiuni logice	confidențialitate	conformitate
consistență	context	continuitate
conținut	corectitudine	cost
cost acuratețe	cost colectare	creativitate
credibilitate	criticabilitate	curente
customizabilitate	definire exactă	detaliere adecvată
detaliere surse	dimensiune	dinamism
dispersie	disponibilitate	domeniu
eficacitate	ergonomie	expansibilitate
extendabilitate	extensibilitate	fără erori
fără greșeli	fără pierderi de informații	fiabilitate
finalitate	flexibilitate	formă prezentare
format	generalitate	ierarhizare
imparțialitate	importanță	independență de timp
integrabilitate	integritate	interactivitate
interesante	interpretare semantică	localizare
manevrabilitate	măsurabilitate	mediu
minimalitate	modularitate	neambiguitate
nivel de abstractizare	nivel de standardizare	normalizare
noutate	obiectivitate	oportunitate
optimalitate	organizare	origine
partiționabilitate	pedigree	personalizate
pertinente	portabile	posibilitate de urmărire
precise	prietenoase	profundzime
răspunde cerințelor	raționalitate	redundanță
regăsire	regularitate format	reiterare
relevantă	reproductibilitate	reputație
responsabilitate	rezistență	rezoluție grafică

rigiditate	robustețe	scop
secrete	securitate	semnificație
sincronizare	specificitate	stabilitate
stocare	supraîncărcare	surse
translatate	transportabilitate	unicitate
ușurință acces	ușurință actualizare	ușurință comparații
ușurință corelații	ușurință folosire	ușurință înțelegere
ușurință interogare	ușurință interschimb	ușurință întreținere
ușurință modificare	ușurință regăsire	utilizabilitate
validitate	valoare	variabilitate
varietate	vârstă	verificabilitate
viteză	volatilitate	volum adecvat

În continuare, sunt furnizate definiții și liste de atribute ale celor mai frecvent menționate douăzeci dimensiuni ale calității, [Ivan99a]:

- **credibilitate** (*de încredere/credibile*) – gradul în care datele sunt acceptate sau privite ca adevărate, reale și credibile;
- **valoare adăugată** (*datele oferă un avantaj competitiv, datele adaugă valoare activității economice*) – gradul în care datele aduc beneficii și furnizează avantaje prin utilizarea lor;
- **relevanță** (*aplicabile, relevante, interesante, folositoare*) – gradul în care datele sunt aplicabile și folositoare pentru activitatea de realizat;
- **acuratețe/exactitate** (*datele sunt certificate ca fiind fără erori, corecte, fără greșeli, erorile pot fi identificate cu ușurință, integritatea datelor, date precise, exacte*) – gradul în care datele sunt corecte, de încredere, și certificate ca fiind fără erori;
- **interpretabilitate** (*interpretabile*) – gradul în care datele sunt prezentate într-un limbaj adecvat, definițiile sunt clare;
- **ușurința înțelegerii** (*ușor de înțeles, clare, ușor de citit*) – gradul în care datele sunt clare, fără ambiguități și ușor de înțeles;
- **accesibilitate** (*accesibile, ușor de regăsit, viteza de acces, disponibile, actuale*) – gradul în care datele sunt disponibile sau ușor de regăsit;
- **obiectivitate** (*imparțiale, obiective*) – gradul în care datele sunt neprejudiciate și imparțiale;
- **oportunitate** (*oportune*) – gradul în care vechimea datelor este adecvată pentru activitatea de realizat;
- **completitudine** (*cantitatea, detalierea și domeniul informațiilor conținute în date*) – gradul în care datele acoperă calitativ și cantitativ domeniul activității de realizat;
- **posibilitatea de urmărire** (*bine documentate, ușor de urmărit, verificabile*) – gradul în care datele sunt documentate, verificabile și au sursa ușor de identificat;
- **reputație** (*reputația sursei datelor, reputația datelor*) – gradul în care datele sunt apreciate din punct de vedere al sursei sau conținutului lor;
- **consistența reprezentării** (*datele sunt în mod continuu reprezentate în același format, reprezentate consistent, formate consistent, datele sunt compatibile cu date istorice*) – gradul în care datele sunt mereu prezentate în același format și sunt compatibile cu formatele precedente;
- **eficiența din punct de vedere al costului** (*costul acurateței datelor, costul achiziției datelor*) – gradul în care costul achiziționării datelor potrivite este rezonabil;

- **ușurința operării** (*ușor de realizat joncțiuni, ușor de modificat, ușor de actualizat, ușor de descărcat/încărcat, datele pot fi folosite în scopuri multiple, manipulabile, ușor de agregat, ușor de reprodus, datele pot fi ușor integrate, ușor de customizat*) – gradul în care datele sunt ușor de manevrat și manipulat;
- **varietatea datelor și a surselor** (*există o varietate a datelor și a surselor acestora*) – gradul în care datele sunt disponibile din mai multe surse diferite;
- **concizie** (*bine prezentate, concise, reprezentate compact, bine organizate, plăcute din punct de vedere estetic, forma de reprezentare, bine formate, formatul datelor*) – gradul în care datele sunt reprezentate compact;
- **securitatea accesului** (*datele nu pot fi accesate de către competitori, datele sunt proprietatea companiei, accesul la date poate fi restricționat și securizat*) – gradul în care accesul la date poate fi restricționat și, astfel, protejat;
- **cantitate suficientă de date** (*cantitatea de date*) – gradul în care cantitatea sau volumul de date este adecvat;
- **flexibilitate** (*adaptabile, flexibile, extensibile, expandabile*) – gradul în care datele sunt expandabile, adaptabile și răspund cu ușurință și altor cerințe.

În continuare, se prezintă pe scurt trei abordări majore pentru construirea sistemelor de îmbunătățire a calității datelor, [Ivan99a].

Construirea *modelelor de date bogate din punct de vedere semantic* pentru a surprinde cât mai multe semnificații ale datelor. Aceste modele trebuie:

- să oprească datele care nu au semnificațiile corecte;
- să ceară achiziționarea datelor lipsă pentru asigurarea completitudinii;
- să cuantifice și înregistreze calitatea datelor din cadrul modelului.

Aceste modele necesită ca dimensiunile de calitate ale datelor să fie incluse în sistem la momentul proiectării conceptuale a acestuia.

Întărirea bazelor de date cu un *număr mare de restricții* pentru a identifica și respinge datele problematice, precum și pentru a lega datele de aplicațiile corespunzătoare, din punct de vedere al dimensiunilor calității datelor.

Restricționarea folosirii datelor de către procese predefinite pentru a elimina folosirea greșită a datelor. Acest lucru poate fi realizat prin încapsularea datelor cu aplicațiile lor corespunzătoare, așa cum se întâmplă în sistemele orientate obiect.

Din ce în ce mai mult, calitatea datelor este considerată ca fiind un aspect foarte important.

Resursele de date sunt centrul unei organizații inteligente, bazate pe informații. Toate datele sunt parte a resurselor de date și trebuie să fie integrate și administrate în mod formal pentru a furniza suport de înaltă calitate activităților economice ale organizațiilor.

Concluzia este că dacă organizațiile nu abordează problemele legate de calitatea datelor, clienții și competitorii o vor face în locul lor, ceea ce va duce la costuri ridicate asociate cu inspecția, cu întreținerea, cu pierderea actualilor și potențialilor clienți și cu timpul pierdut.

27.3 Metricilor datelor

În [Popa05], sunt definite și sistematizate metrici de evaluare a calității datelor în funcție de caracteristicile de calitate a datelor.

Acuratețea reprezintă măsura în care datele sunt corecte, exacte, fiind certificate ca fiind fără erori, [Popm05].

În [Ivan99a] se utilizează definiția de lucru conform căreia corectitudinea reprezintă datele achiziționate de un sistem informațional care reflectă situația din lumea reală pe care intenționează să o reprezinte.

Corectitudinea se definește pe baza modelului de reprezentare a datelor sub forma tripletului (*entitate, atribut, valoare*). Pentru entitatea ET, reprezentată prin:

$$ET = (en, at, val) \quad (27.11)$$

măsura corectitudinii este dată de diferența dintre valoarea *val* a tripletului ET și valoarea *val'* considerată corectă pentru atributul *at*.

În valoare absolută, măsura erorii înregistrată pentru entitatea ET este determină conform expresiei:

$$E_r = val - val' \quad (27.12)$$

În valoare relativă, determinarea erorii pentru entitatea ET se efectuează conform raportului:

$$E_{rel} = \frac{|val - val'|}{val'} \quad (27.13)$$

Rezultatul evidențiază coeficientul cu care valoarea înregistrată pentru atributul *at* variază față de valoarea considerată corectă.

Problema care se pune privește stabilirea valorii corecte *val'* datorită modurilor în care aceasta se calculează sau se stabilește.

Pentru un set de date, corectitudinea se calculează conform expresiei, [Popa05]:

$$I_{AC}^{DT} = \frac{nval'}{nval} \quad (27.14)$$

unde:

- *nval'* – numărul de valori incorecte;
- *nval* – numărul de valori corecte.

Devizul aferent pentru o etapă din derularea unui proiect de cercetare are structura descrisă în tabelul 27.2, [Popa05].

Tabelul nr. 27.2 Deviz cadru pentru etapa i
- RON -

Nivel de agregare	Categorie cheltuieli	Valoare etapa i
(1)	Cheltuieli directe :	mtc_{1,i}
1	1. Cheltuieli de personal	mtc_{2,i}
	1.1 Cheltuieli cu salariile	mt _{3,i}
	1.2 Contribuții	mt _{4,i}
	1.3 Alte cheltuieli de personal	mt _{5,i}
2	Cheltuieli materiale și servicii	mtc_{6,i}
	2.1 Materiale, materii prime	mt _{7,i}
	2.2 Lucrări și servicii executate de terți ¹	mt _{8,i}
3	Alte cheltuieli specifice proiectului	mt _{9,i}
(2)	Cheltuieli indirecte: regia	mtc_{10,i}
(3)	Dotări independente și studii pentru obiective de investiții:	mtc_{11,i}
	1. echipamente pentru cercetare-dezvoltare;	mt _{12,i}
	2. mobilier aparatură;	mt _{13,i}
	3. calculatoare electronice și echipamente periferice;	mt _{14,i}
	5. mijloace de transport;	mt _{15,i}
	6. studii pentru obiective de investiții.	mt _{16,i}
	Total	mtc_{17,i}

Elementele mt_{ij} sunt valori primare, iar mtc_{ij} sunt valori calculate. Se formează tuple de valori, care sunt instanțe ale șablonului:

$$SB_{VE} = (mtc_{1,i}, mtc_{2,i}, mt_{3,i}, mt_{4,i}, mt_{5,i}, mtc_{6,i}, mt_{7,i}, mt_{8,i}, mt_{9,i}, mtc_{10,i}, mtc_{11,i}, mt_{12,i}, mt_{13,i}, mt_{14,i}, mt_{15,i}, mt_{16,i}, mtc_{17,i}) \quad (27.15)$$

Șablonul SB_{VE} conține 17 elemente, valori primare și calculate. Pentru instanța:

$$V_{VE} = (16000; 10505; 5739,6; 1865,4; 2900; 4995; 4995; 0; 500; 4000; 0; 0; 0; 0; 0; 0; 20000) \quad (27.16)$$

valoarea indicatorului $I_{AC}^{DT} = 0$ deoarece valorile înregistrate și calculate sunt corecte, fiind fundamentate pe note de fundamentare și relații de calcul între categoriile de cheltuieli.

O eroare de calcul a cheltuielilor privind contribuțiile la bugetul de stat conduce la propagarea sa în deviz.

Deoarece valoarea indicatorului este mai mare decât zero, se concluzionează că setul de date considerat este afectat de erori. Valoarea 1 reprezintă un prag cu următoarea semnificație:

- dacă $I_{AC}^{DT} < 1$, atunci $nval' < nval$;

- dacă $I_{AC}^{DT} = 1$, atunci $nval' = nval$;
- dacă $I_{AC}^{DT} > 1$, atunci $nval' > nval$.

În cazul datelor reprezentate ca texte, corectitudinea este asigurată prin respectarea regulilor semantice prestabilite. Se consideră domeniile de activitate D_i și D_j . Domeniile au asociate vocabulare ale cuvintelor cheie $V_{ch,i}$, respectiv $V_{ch,j}$. Se construiesc entitățile ET_i și ET_j , aferente fiecărui domeniu în parte.

Entitățile ET_i și ET_j sunt corecte dacă, [Popa05]:

$$V_{ch,i} \cap V_{ch,j} = \emptyset \quad (27.17)$$

Relația de mai sus este valabilă pentru domenii de activitate izolate. Dacă domeniile D_i și D_j au elemente comune, atunci corectitudinea este asigurată dacă:

$$V_{ch,i} \cap V_{ch,j} = V_{ch,co} \quad (27.18)$$

Pentru entitatea ET_i , măsura corectitudinii semantice este dată de formula, [Popa05]:

$$I_{AC}^{ET_i} = \frac{ich}{nch_i'} \quad (27.19)$$

unde:

- ich – numărul de cuvinte cheie din $V_{ch,j} - V_{ch,co}$ asociate cuvintelor cheie din ET_i ;
- nch_i' – numărul de cuvinte cheie din $V_{ch,i} - V_{ch,co}$.

Entitatea ET_i este corectă din punctul de vedere al modelului semantic dacă $I_{AC}^{ET_i}$ este nul.

Corectitudinea datelor depinde de frecvența de actualizare a acestora și de numărul persoanelor autorizate să facă acest lucru. Având în vedere modelul semantic care stă la baza corectitudinii, capacitățile celor care actualizează date joacă un rol foarte important.

Într-o colecție de date, completitudinea reprezintă măsura în care attributele au asigurate valori.

O dată este incompletă dacă pentru un atribut lipsește măsura acestuia. Lipsa valorii atributului are diferite sensuri în funcție de caracterul atributului: obligatoriu, opțional sau inaplicabil.

O dată este completă atunci când sunt înregistrate valori pentru toate attributele. Valorile atributelor sunt exprimate cantitativ, sub formă numerică, sau calitativ, sub formă de text.

Analiza comparativă a textelor prin prisma caracteristicilor de calitate a acestora se efectuează prin intermediul indicatorilor. Aceste metrice dau o mărime cuantificabilă caracteristicilor de calitate a textelor.

Pentru un text oarecare T se definesc indicatori privind caracteristica de completitudine.

Indicatorul de completitudine evidențiază măsura în care textul construit acoperă elementele fundamentale ale domeniului abordat, [Popa05]:

$$I_{CM} = \frac{nea}{ned} \quad (27.20)$$

unde:

- *nea* – numărul de aspecte ale domeniului abordate în textul T;
- *ned* – numărul total de elemente pentru domeniului abordat.

Variabilele *nea* și *ned* vizează atât cuvintele cheie ale domeniului, cât și aspecte legate de descrierea domeniului, tehnici și metode implementate, instrumente folosite.

Prin intermediul indicatorului de completitudine a sferei de cuprindere a textului T față de sfera de cuprindere a domeniului abordat, D_i , se dă o măsură a gradului de utilizare a vocabularelor cuvintelor cheie, V_{ch} , și a cuvintelor specifice domeniului, V_{lit} .

Se consideră textele $T_1, T_2, \dots, T_{nx}$. Fiecare text conține un subvocabular de cuvintelor cheie ale domeniului, SV_{ch}^i , și se calculează indicatorii de completitudine de forma I_{CM}^i . Dacă:

$$SV_{ch}^1 \cap SV_{ch}^2 \cap \dots \cap SV_{ch}^{nx} = \emptyset \quad (27.21)$$

și:

$$V_{ch} = SV_{ch}^1 \cup SV_{ch}^2 \cup \dots \cup SV_{ch}^{nx} \quad (27.22)$$

atunci textele $T_1, T_2, \dots, T_{nx}$ se constituie în părți ale descrierii domeniului, iar reuniunea lor conduce la o entitate text care acoperă în întregime sfera de cuprindere a domeniului. În aceste condiții, indicatorul I_{CM} respectă relația, [Popa05]:

$$I_{CM}^{AG} = \sum_{i=1}^{nx} I_{CM}^i \quad (27.23)$$

Dacă:

$$SV_{ch}^1 \cap SV_{ch}^2 \cap \dots \cap SV_{ch}^{nx} \neq \emptyset \quad (27.24)$$

și:

$$V_{ch} = SV_{ch}^1 \cup SV_{ch}^2 \cup \dots \cup SV_{ch}^{nx} \quad (27.25)$$

atunci textele $T_1, T_2, \dots, T_{nx}$ sunt părți ale descrierii domeniului, dar ele conțin elemente comune din conceptele fundamentale ale domeniului. Indicatorul de completitudine se calculează conform expresiei:

$$I_{CM}^{Lg} = \frac{Lgv(V_{ch,co})}{Lgv(V_{ch})} \quad (27.26)$$

Realizarea unei entități ET', prin reunirea textelor, prezintă următoarele caracteristici datorită vocabularului cuvintelor cheie comune:

- redundanță a conceptelor – pentru același cuvânt cheie din V_{ch} se oferă aceeași definiție în mai multe texte T_i ; acest lucru conduce la secvențe de text care se repetă, determinând o scădere a eficienței de utilizare a entității text;
- detalieri ale conceptelor – pentru un cuvânt cheie din vocabularul V_{ch} sunt prezente mai multe definiții, mai multe puncte de vedere, ceea ce determină o creștere a acurateții și mai multe viziuni privind conceptul.

Prin intermediul indicatorilor de completitudine sunt puse în evidență relații de tipul parte-întreg. Astfel, sunt evidențiate aspecte calitative privind o entitate elaborată într-un domeniu.

Se consideră devizul cadru conform tabelului 27.3 [Popm05]:

Tabelul nr. 27.3 Deviz cadru

- RON -

		TOTAL	2004	2005	2006
(1)	Cheltuieli directe :				
1	1. Cheltuieli de personal				
	1.1 Cheltuieli cu salariile				
	1.2 Contribuții				
	1.3 Alte cheltuieli de personal	300000	100000	100000	100000
2	Cheltuieli materiale și servicii	225000	75000	75000	75000
	2.1 Materiale, materii prime	75000	25000	25000	25000
	2.2 Lucrări și servicii executate de terți ¹	150000	50000	50000	50000
3	Alte cheltuieli specifice proiectului	150000	50000	50000	50000
(2)	Cheltuieli indirecte: regia				
(3)	Dotări independente și studii pentru obiective de investiții:	285000	95000	95000	95000
	1. echipamente pentru cercetare-dezvoltare;	0	0	0	0
	2. mobilier aparatură;	60000	20000	20000	20000
	3. calculatoare electronice și echipamente periferice;	225000	75000	75000	75000
	5. mijloace de transport;	0	0	0	0
	6. studii pentru obiective de investiții.	0	0	0	0
	Total tarif (valoare contract)				

Tabelul 27.3 este incomplet deoarece există categorii de cheltuieli pentru care nu au fost specificate sau calculate sumele. Astfel, omisiunea precizării categoriei *Cheltuieli cu salariile* face imposibilă determinarea valorilor aferente contribuțiilor. Aceasta se propagă în cascadă făcând imposibilă calcularea valorii pentru *Cheltuielilor directe* și apoi *Cheltuieli indirecte*.

Pentru devizul de mai sus se calculează următorii indicatori:

- număr de valori primare omise: 3 valori;
- număr de valori derivate care nu sunt determinate: 13 valori;
- număr de valori agregate pe cel mai înalt nivel: 4 valori.

Ca urmare, nici valoarea cheltuielilor pe întreg contractul nu a putut fi calculată.

Pentru o matrice, gradul de completitudine este [Popa05]:

$$I_{GM} = \frac{m_i \cdot n_i - kv}{m_i \cdot n_i} \quad (27.27)$$

unde:

• m_i	- numărul de linii al matricei;
• n_i	- numărul de coloane al matricei;
• kv	- numărul valorilor lipsă din matrice.

Pentru valorile cuprinse în tabelul 27.3, valoarea indicatorului $I_{GM}=0.70$, ceea ce înseamnă că devizul cadru este completat în proporție de 70%.

Efectele omisiunii sunt vizibile atât pe orizontală cât și pe verticală. Pe orizontală lipsesc valori ale unor categorii de cheltuieli pe întreaga perioadă de derulare. Pe verticală este imposibilă agregarea valorilor pe perioade de timp și pe toată durata de derulare a contractului. Este suficient ca o singură valoare să lipsească pentru a fi afectat întregul lanț de agregare a sumelor.

Aceste valori sunt calculate dacă se furnizează valoarea atributului *Valoare contract*. Astfel, apelând la ingineria inversă se determină valori pentru categorii de cheltuieli.

Aceasta nu înseamnă ca vor fi completate toate categoriile de cheltuieli specificate în deviz. De exemplu, detalierea contribuțiilor presupune cunoașterea reglementărilor legale în vigoare.

Se consideră o colectivitate KOL generatoare de entități. Fiecare element al colectivității generează o entitate text.

Într-o competiție pentru obținerea de fonduri sunt depuse mai multe oferte de proiecte $EP_1, EP_2, \dots, EP_{mp}$, unde mp reprezintă numărul de proiecte înscrise în competiție. Proiectele se construiesc după reguli impuse de echipa managerială a programului de finanțare.

Un proiect trebuie să includă:

- descrierea domeniului în care se dezvoltă procesul de finanțare: cercetare, construcții, investiții, din care se face dovada că echipa care a realizat oferta are experiență în domeniul abordat;
- prezentarea soluției propuse, cu descrierea în detaliu a activităților, resurselor, riscurilor, termenelor și efectelor;
- detalierea din punct de vedere economico-financiar a implicațiilor pe care le au achizițiile de materiale, plățile de salarii, neîncadrarea în limitele de timp și consum;
- efectele economico-sociale pe care le generează implementarea soluției descrisă în proiect.

Un proiect EP_i din colecția $EP_1, EP_2, \dots, EP_{mp}$ este descris după o structura impusă.

Elementele prezentate mai sus sunt criterii de stabilire a omogenității entităților. Există două forme ale omogenității:

- omogenitatea formei de reprezentare a datelor;
- omogenitatea conținutului – entitățile aparțin aceluiași domeniu.

Din punctul de vedere al spațiului de memorie alocat, omogenitatea se referă la utilizarea acelorași tipuri de date. Gradul de omogenitate, I_{OD} , se determină conform expresiei [Popa05]:

$$I_{OD} = \sqrt[n_{oz}]{\prod_{i=1}^{n_{oz}} \frac{L_{OD_i}}{\max_{j=1, n_{oz}}(L_{OD_j})}} \quad (27.28)$$

unde:

- n_{oz} – numărul de operanzi;
- L_{OD_i} – memoria alocată pentru operandul OD_i .

Indicatorul I_{OD} oferă o măsură a omogenității operanzilor în funcție de memoria alocată. Pentru $I_{OD} = 1$, omogenitatea este maximă.

Se consideră un număr n_{fr} de facturi. Asigurarea omogenității datelor presupune efectuarea de conversii, după cum urmează:

1. cantitate: întreg – float;
2. rezultat spre pret: float – double;
3. tva: întreg – double;
4. constantă: float – double;
5. rezultat spre tva_calculat: double – float;
6. tva_calculat: float – double;
7. rezultat spre pret_vanzare: double – float.

pentru fiecare articol prelucrat, deci un total de $7 \cdot n_{fr}$ conversii.

Apare problema luării unei decizii, pe de o parte, pentru a economisi spațiul de memorie cu:

$$I_{ES} = \left(1 - \frac{\text{sizeof}(\text{factura})}{\text{sizeof}(\text{factura_omogena})} \right) \cdot 100 \quad (27.29)$$

dar cu un volum de $7 \cdot n_{fr}$ conversii. Pe de altă parte, fără economie de spațiu pe suport, fișierul conținând facturile omogene are $n_{fr} \cdot \text{sizeof}(\text{factura_omogena})$ baiți, iar prelucrările necesită zero conversii.

Omogenitatea datelor se asigură prin operații de transformare. Tuplul inițial de valori are asociat un tuplu de aceeași dimensiune cu coeficienți de transformare. Pentru tuplul:

$$SB_F = (\text{cantitate}, \text{rezultat_pret}, \text{tva}, \text{constantă}, \text{rezultat_tva_calculat}, \text{tva_calculat}, \text{rezultat_pret_vanzare}) \quad (27.30)$$

se asociază coeficienții de transformare din tuplul următor:

$$CT_F = (2; 2; 4; 2; 0,5; 2; 0,5) \quad (27.31)$$

obținându-se tuplul valorilor omogenizate:

$$V_F = (4, 8, 8, 8, 4, 8, 4) \quad (27.32)$$

care corespunde tuplului înregistrării de tip *factura_omogenă*.

Complexitatea desemnează măsura în care un sistem sau o componentă posedă o arhitectură sau implementare dificile de înțeles și verificat, [www1].

Complexitatea ciclomatică este considerată o metrică importantă pentru măsurarea solidității și încrederii pentru o entitate. Această metrică furnizează o valoare care se compară cu valorile obținute pentru alte

entităţi. Este utilizată în special pentru entităţile de tip software şi cuantifică numărul de căi liniare independente existente într-un modul program.

Complexitatea ciclomatică pentru entitatea ET, $CxC(ET)$, are expresia:

$$CxC(ET) = nmg - nng - nnc \quad (27.33)$$

unde:

- nmg – numărul de muchii al grafului asociat structurii entităţii ET;
- nng – numărul de elemente din structura ET;
- nnd – numărul de componente conectate.

Complexitatea reprezintă o măsură importantă a complexităţii structurii entităţii ET.

Pentru entităţile software au fost stabilite praguri în funcţie de complexitatea ciclomatică, tabelul 27.4, [www1].

Tabelul nr. 27.4 Praguri ale complexităţii ciclomatice

Complexitatea ciclomatică	Evaluarea riscului
1 – 10	Entitate software simplă, fără a implica mari riscuri
11 – 20	Entitate software complexă, risc moderat
21 – 50	Entitate software extrem de complexă, cu riscuri mari
Peste 51	Entitate netestabilă, cu riscuri foarte mari

Complexitatea ciclomatică este determinată şi utilizată în:

- analiza riscului dezvoltării de cod sursă;
- analiza riscului schimbării în mentenanţă;
- planificarea testării;
- reinginerie.

Complexitatea datelor rezultă din agregarea numărului de apariţii şi a numărului de legături dintre ele, după relaţia [Ivan99a]:

$$Cx(DT) = noz \log_2 noz + nop \log_2 nop \quad (27.34)$$

unde:

- noz – numărul de apariţii;
- nop – numărul de operatori.

Norma de complexitate a unei entităţi text, determinată pe baza matricei de precedente, este [Ivan99a]:

$$NCx(MP) = \frac{\sum_{i,j} mc_{ij} \log_2 mc_{ij}}{\left(\sum_{i,j} mc_{ij} \right) \log_2 \left(\sum_{i,j} mc_{ij} \right)} \quad (27.35)$$

Determinarea complexității sistemului de caracteristici permite efectuarea de analize comparative cu alte entități de date și poziționarea sa în raport cu acestea. Evaluarea complexității se bazează pe cuantificarea numerică a construcțiilor de ordin sintactic și a legăturilor dintre acestea.

27.4 Managementul calității datelor

Managementul total al calității este o filozofie și un set de principii orientative care formează un cadru de lucru pentru realizarea excelenței în domeniul calității.

Următoarele definiții ale calității au fost formulate de către patru lideri cheie ai revoluției calității, [Ivan99a]:

"... un grad predictibil al uniformității și dependibilității la un cost scăzut, potrivit pieței." W. Edwards Deming (1986)

"...adecvare pentru folosire, conform aprecierii utilizatorului." Joseph M. Juran (1989)

"...conformitate cu cerințele." Philip B. Crosby (1979)

"...satisfacția deplină a clientului." Armand V. Feigenbaum (1956)

Aceste definiții reprezintă diferite puncte de vedere asupra calității. Definiția lui Juran se bazează pe aprecierea clientului. Definiția lui Crosby reprezintă punctul de vedere manufacturier. Definițiile lui Deming și respectiv Feigenbaum susțin ambele puncte de vedere. Deming găsește că este important să accentueze importanța costului în definiția sa.

În filozofia TQM, organizațiile sunt formate dintr-un sistem complex de clienți și furnizori, aceștia putând fi interni sau externi. *"Intern"* sugerează că fiecare angajat funcționează atât pe post de furnizor cât și de client. Angajații trebuie să fie atenți cine îi aprovizionează cu cele necesare activității lor și cine este clientul pentru ceea ce produc ei. TQM este un sistem de relații interdependente a căror eficiență depinde de cât de bine furnizorul satisface cerințele clienților lui. Este evident faptul că în abordarea TQM calitatea este principala prioritate pentru a satisface cerințele clientului. Cele mai importante criterii ale aprecierii calității sunt formulate de client.

Datele sunt utilizate ca suport pentru luarea deciziilor și managementul organizațiilor moderne, fie ele de natură operațională, tactică sau strategică. Dacă aceste date sunt eronate, există mai multe moduri în care slaba calitate a datelor poate afecta eficiența organizației.

Datorită faptului că reprezintă un factor strategic pentru o organizație, cadrul de decizii manageriale constituie o zonă unde calitatea datelor are un impact deosebit asupra rezultatului final. Odată cu răspândirea sistemelor suport de decizii manageriale, din ce în ce mai multe date vor proveni din baze de date și vor circula atât în interiorul cât și în exteriorul organizațiilor.

Există analogii fundamentale între aspectele de calitate dintr-un mediu în care se produc de bunuri și cele dintr-un sistem informațional. Producerea de bunuri este privită ca un sistem de prelucrare care acționează asupra unor intrări materiale pentru a produce ieșiri materiale. Analog, un sistem informațional este văzut ca un sistem de prelucrare care acționează asupra datelor de intrare pentru a produce date de ieșire. Tabelul 27.5 ilustrează această analogie.

Tabelul nr. 27.5 Analogia dintre producerea de bunuri și producerea date

	Producția de bunuri	Producția de date
Intrări	Materii prime	Date de intrare
Proces	Linii de producție	Sisteme informaționale
Ieșiri	Produse finite	Informații

Se folosește analogia cu producerea de bunuri și se consideră cele două domenii de activitate. Primul este **proiectarea și implementarea** sistemului informațional, iar cel de-al doilea **producerea și distribuirea** datelor. Acestea două vor fi referite ca *ciclul de viață al sistemelor de date* și, respectiv, *lanțul valoric al producției de date*.

Ciclul de viață al sistemelor de date se concentrează asupra activităților care sunt realizate pentru proiectarea, dezvoltarea, testarea și punerea la dispoziția utilizatorilor a sistemului informațional. Deoarece datele care rezultă dintr-un sistem informațional reprezintă o funcție de toate aceste activități, fiecare dintre ele trebuie considerată o zonă potențială pentru îmbunătățirea calității datelor. Se iau în considerare următoarele trei exemple de probleme de calitate legate de ciclul de viață al sistemelor de date:

- *analiza cerințelor* – datele proiectate în sistem nu sunt cele cerute de utilizatori;
- *analiza calității software* – testarea funcționalității software-ului și a bazei de date este incompletă;
- *instruirea* – comunitatea utilizatorilor este inadecvat instruită în privința introducerii și regăsirii datelor din sistem.

Sistemele generatoare de date produc date care au valoare pentru alte organizații. Distribuitorii de date cumpără date de la producători și le revând organizațiilor consumatoare, care, după ce le achiziționează, le analizează și le prelucrează, revânzându-le apoi la rândul lor firmelor de marketing (consumatori). În final, organizațiile consumatoare sunt cele care dețin date generate extern.

Cu excepția distribuitorilor de date, cele mai multe companii nu aparțin numai unui anumit grup. De fapt, cele mai multe organizații sunt integrate pe verticală în ceea ce privește fluxul de date, având diverse departamente, fiecare îndeplinind funcțiuni diferite.

Ca rezultat al integrării pe verticală, cele mai multe organizații de sisteme informaționale au responsabilități pentru fiecare funcție: generarea de date, distribuția internă și consumul de date generate atât intern, cât și extern.

În prezent, controlul și managementul calității au devenit necesități competitive pentru majoritatea firmelor. În domeniul calității există o bogată experiență de câteva decade. Abordările variază de la cele tehnice, cum ar fi controlul statistic al proceselor, până la cele manageriale, cum ar fi cercurile calității. O experiență similară este necesară și în domeniul calității datelor.

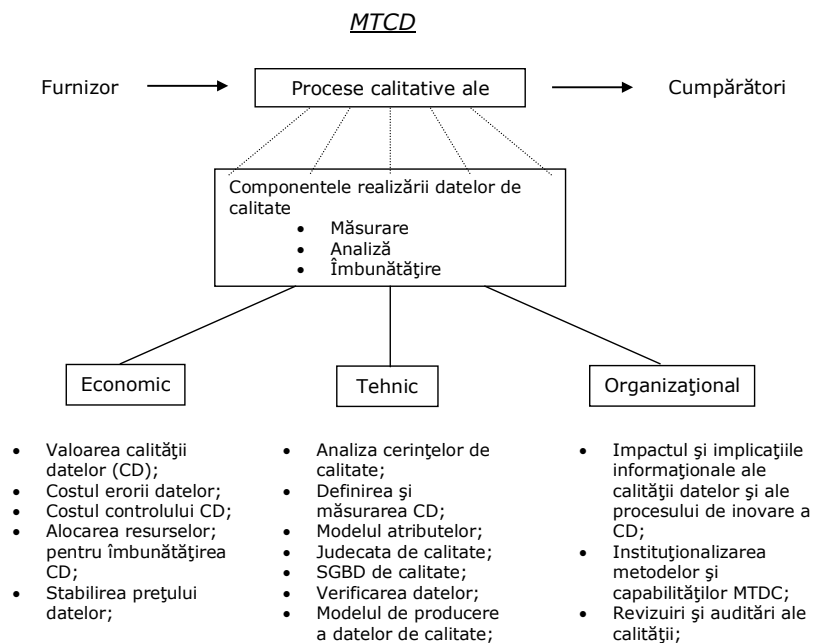


Figura 27.2 MTCD – perspectiva calității datelor

Acest cadru de lucru cuprinde două dimensiuni primare intercorelate, [Ivan99a]:

1. trei componente ale procesului continuu de creștere a calității datelor: **măsurare, analiză și îmbunătățire**;
2. trei perspective către care se orientează soluțiile: **economică, tehnologică și organizațională**.

Măsurarea calității datelor. Deși *calitatea datelor* este o noțiune care în mod intuitiv pare evidentă, cele mai multe laturi ale acesteia nu au fost foarte clar definite în practica curentă. Majoritatea oamenilor, când sunt întrebați despre calitatea datelor, oferă termeni ca *acuratețe* și *oportunitate*, dar atunci când sunt puși să definească și acești termeni realizează subtilitatea conceptelor implicate. Este necesară o definiție clară și unanim acceptată a calității datelor.

Studiile au arătat că, pentru a defini calitatea datelor, consumatorii de date utilizează mai multe dimensiuni, printre care *acuratețea*, *credibilitatea*, *relevanța*, *oportunitatea* și *completitudinea*. De fapt, nici chiar cea mai evidentă latură, *acuratețea*, nu are o definiție suficient de robustă.

Analiza impactului calității datelor asupra activității economice. Această componentă se referă la valoarea relației dintre datele cu *calitate ridicată* și succesul operării unei afaceri, reversul reprezentându-l legătura dintre datele cu *calitate scăzută* și impactul lor negativ asupra activității. Tehnicile de analiză corelează calitatea datelor cu principalii parametri cheie ai unei afaceri, cum ar fi vânzările, satisfacția clienților și profitabilitatea.

Îmbunătățirea calității datelor. Această componentă se referă la diverse metode de îmbunătățire a calității datelor. Aceste metode sunt grupate în trei categorii intercorelate:

- reproiectarea activităților;
- motivația calității datelor;
- folosirea noilor tehnologii.

Perspectiva *economică* se referă la probleme cum sunt: cât de mult ar costa realizarea calității datelor, cât de mult merită acest lucru și cum

alocăm resurse pentru calitatea datelor? Sub restricția resurselor, este evident că nu pot fi abordate simultan toate problemele și oportunitățile de calitate a datelor.

Perspectiva *tehnologică* se referă la definirea și măsurarea calității datelor și la metodele de îmbunătățire a calității datelor. De exemplu, sunt folosite instrumente care ajută la analiza cerințelor de calitate a datelor, controlul statistic al proceselor pentru calitatea datelor, precum și dispozitive pentru culegerea avansată a datelor.

Perspectiva *organizațională* se referă la implicațiile calității datelor asupra structurii sociale a firmei, luând în considerare aspecte legate de motivarea și reacțiile, atât ale indivizilor cât și ale grupurilor, la schimbările organizaționale și manipularea corespunzătoare a datelor. Aceasta include procesul de dezvoltare a dedicării organizaționale pentru calitatea datelor, dezvoltarea infrastructurii tehnologice, modificări ale stimulentei pentru manipularea datelor și instituționalizarea pe scară largă a metodelor de îmbunătățire a calității datelor.

27.5 Costul datelor

Spre deosebire de produsele tradiționale, costul datelor are o cu totul altă bază. De exemplu, pentru calculul costului rețelei de fabricație pentru un produs se iau în considerare, [Ivan99a]:

- costurile cercetărilor – personal, materii prime, utilaje;
- costurile generate de efectele de antrenare – profitul are un cost propriu;
- costul culegerii datelor – reprezentate prin salariile personalului, aparatele de măsură și suportii tehnici de transmitere în timp real a rezultatelor măsurării.

Vânzarea unui proiect, înseamnă vânzare de date. Nu toate datele fac obiectul comercializării, deși ar trebui să existe preț pentru orice dată.

Costul datelor este diferit de costul de obținerii datelor. Costul obținerii datelor vizează totalitatea cheltuielilor generate de măsurare, verificare, culegere, înscriere, transmitere, păstrare și arhivare a datelor.

Costul obținerii datelor privind evoluția bursieră, având în vedere caracterul public al acestei activități, este dat de cheltuielile de conectare la rețeaua Internet și de cheltuielile specifice ale rețelei telefonice, având nivelul CD.

Studiul evoluție bursiere pe baza unor informații complete conduce la decizia achiziționării sau vânzării de acțiuni, cu obținerea unui profit P. Prețul datelor PD este dat de relația:

$$PD = P - CD \quad (27.36)$$

Întrucât pe piața bursieră nu se acționează fără o bază informațională, problema prețului datelor se pune altfel decât în cazul producătorului de tomate, unde baza de referință era profitul obținut empiric.

Modelul destinat evaluării prețului datelor PD conține două substructuri:

- substructura profitului obținut sau estimat (P);
- substructura cheltuielilor pentru obținerea datelor (CD).

În cazul în care se înregistrează o evoluție în timp, la momentele de timp 1, 2, ..., K pe baza unor noi date.

$$PD = \sum_{k=1}^K (P_k - CD_k) \quad (27.37)$$

În cazul în care datele stau la baza fundamentării deciziilor $D_1, D_2, \dots, D_p$ și la estimarea profitului aferent fiecărei decizii în actul decizional, se presupune luarea deciziei P_i care maximizează profitul.

Cheltuielile pentru culegerea datelor CD sunt aceleași pentru construirea ipotezelor care generează cele p variante și deci pentru fiecare dintre cele p decizii în parte.

$$PD = \max_{i \leq p} \{p_i\} - CD \quad (27.38)$$

Costul datelor necesită un nou mod de abordare întrucât componentele și factorii nu apar în mod direct ci prin intermediul efectelor negative pe care le generează absența datelor sau noncalitatea acestora.