Data Flow Diagram

Lt.col. Otilia PÎRLOG
Ministerul Apărării Naţionale

IT systems can be perceived as subsystems in a larger feedback control system of the real world. Data manufacturing can be seen as a predefined set of data units which are processed in order to obtain data to be used by internal and / or external consumers. The data it inputted into a system based on external flows, it is processed and gets stored into data collections, which are then processed in order to produce results that are compared with the real world. The deficiencies represent the differences between the image of the real word system which can be deduced from the interpretation of the data and the image obtain by directly observing the real world system. By forming the data flow diagram both the effect of the quality of the input data and the efficiency of the processing are made easier to analyze, the flow of primary data being analyzed through the entire chain of transformations.
Keywords: *data, quality, real world system, analyze, diagram, flow.*

Deficienţele datelor şi lumea reală

Sistemele care omit luarea în considerare a contextului mai larg al sistemului FCS întâmpină greutăţi de adaptare la condiţiile lumii reale. Formalizarea legăturii dintre date *(D)* şi lumea reală *(LR)* este reprezentată de aplicaţia $LR \rightarrow D$. Sensul unidirecţional al relaţiei este imprimat de deficienţele datelor, care conţin deviaţii de reprezentare a lumii reale.

În cele din urmă, dificultatea reală a calităţii datelor este datã de dependenţa de timp. Înregistrările din colecţiile de date sunt statice. Chiar dacă la momentul iniţial t_0 datele înregistrate sunt în concordanţă cu lumea reală, la momentul t_1 vor fi puţin diferite, iar la momentul t_2 diferenţa se va accentua (figura 1).

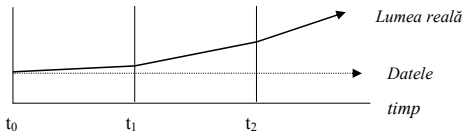

Fig. 1. Datele şi lumea reală

Dezvoltarea şi utilizarea unui sistem informatic implică două transformări: de reprezentare şi de interpretare. Transformarea de reprezentare creează o imagine perceptibilă a unei părţi a sistemului lumii reale, prin intermediul datelor create şi stocate de către sistemul informatic. Transformarea de interpretare reprezintã folosirea datelor pentru a deduce o imagine a sistemului lumii reale reprezentate (figura 2). Deficienţele reprezintă neconformitatea dintre imaginea sistemului lumii reale, care poate fi dedusă prin interpreatarea datelor, şi imaginea obţinută prin observarea directă a sistemului lumii reale.

Pe de altă parte, în sistemele informatice reprezentarea lumii reale (intrările) şi interpretarea ei (ieşirile) apar la momente de timp diferite. Datele au rol de mediere între acestea. Ele reprezintă un sistem de control cu feedback al lumii reale. Calitatea este măsura concordanţei dintre imaginile datelor, prezentate de către un sistem informatic, şi aceleaşi date în lumea reală (figura 3) [Orrk98].

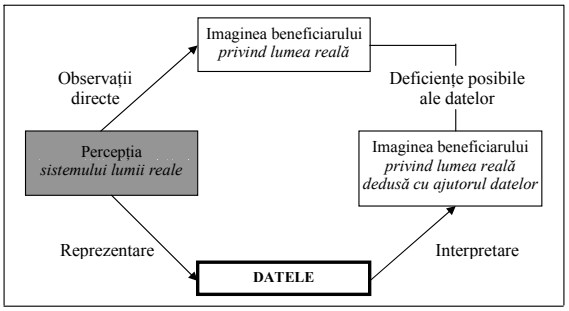

Fig. 2. Deficienţele datelor şi lumea reală

În acest model datele sunt introduse într-un sistem bazat pe fluxuri externe, trec prin procesare şi se memorează în colecţiile de date, care apoi sunt procesate pentru a produce re-

zultate care sunt comparate cu lumea reală. În ciclurile următoare sunt create noi intrări în sistem care sunt comparate cu întoarcerile feedback din sistem. Fără această buclă finală nu este posibilă menţinerea corectitudinii intrărilor şi deci a corectitudinii produselor informatice rezultate.

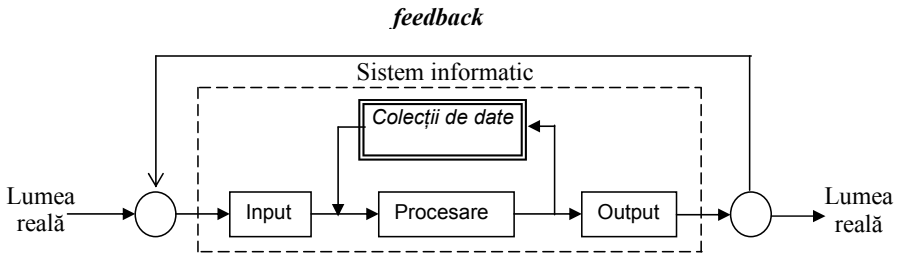

Fig. 3. Datele în contextul lumii reale

Sistemul producerii de informaţii

Producţia de informaţii reprezintă un set predefinit de unităţi de date care trec prin activităţi de procesare diferite pentru a se obţine produse informatice destinate consumatorilor interni şi / sau externi.

Un produs informatic are o valoare potenţială dată pentru un beneficiar dat. Aceasta poate fi diminuată dacă furnizarea nu se face la timp sau produsul este de calitate scăzută. Valoarea produselor informatice poate fi îmbunătăţită atât prin asigurarea condiţiilor de preluare în sistem a unor date primare cu un nivel calitativ superior, cât şi prin realizarea unor configuraţii optime ale sistemelor de producere de informaţii.

O unitate de date poate suporta trei tipuri diferite de operaţii într-un sistem: procesare, îmbunătăţirea calităţii şi stocare. Procesarea include funcţii aritmetice precum şi diverse alte procesări, cum ar fi sortarea. Acestea au asociaţi parametri referitori la cost, îmbunătăţirea (degradarea) calităţii şi timpul de procesare. După colectarea lor, unităţile de date trec printr-o serie de operaţii succesive înainte de a fi posibilă livrarea lor. Aceste operaţii pot fi de procesare, stocare sau de control al calităţii.

Pentru asigurarea unui cadru de analiză a procesării datelor, se foloseşte diagrama fluxului datelor. Blocurile componente ale *sistemului de producere de informaţii* sunt: furnizare date primare, procesare, memorare, control al calităţii şi utilizator (figura 4). Ele pot de asemenea induce întârzieri dacă este necesară aşteptarea completării tuturor unităţilor de date pentru procesare. Vectorii datelor reprezintă elementele de conexiune a acestor blocuri, pornind de la datele primare şi terminând cu informaţiile livrate la beneficiarii acestora. Costul, oportunitatea şi calitatea sunt determinate după fiecare operaţie executată asupra datelor.

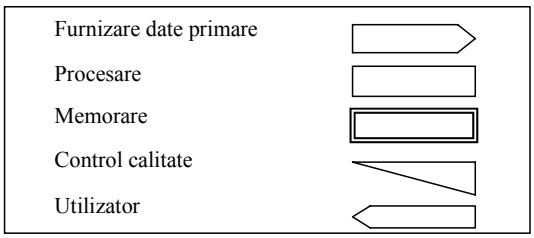

Fig. 4. Blocurile componentele ale sistemului de producere de informaţii

Blocul furnizare date primare *(VB)* reprezintă diferite surse de preluare ale datelor. Unităţile de date *(DU)* simbolizează datele care se află în diverse stadii ale procesului de prelucrare, primare sau care au suferit impactul trecerii lor prin blocurile de procesare, memorare şi control al calităţii. Rolul blocului de procesare *(PB)* este de a adăuga valoare prin manipularea sau combinarea diverselor unităţi de date. Blocul de memorare *(DS)* plasează datele în colecţii unde sunt disponibile pentru procesările ulterioare şi nu afectează oportunitatea şi calitatea unităţilor de date. Totuşi, în urma trecerii printr-un astfel de bloc, unităţile de date acumulează costuri. Prin intermediul blocului de control calitate *(QB)* se verifică starea caracteristicilor datelor analizate. La un moment dat se analizează o singură

intrare. Blocul utilizator *(CB)* constituie finalitatea procesării, deci produsul informatic. Blocurile care afectează valoarea caracteristicilor de calitate (procesare, memorare, control calitate) sunt denumite generic *blocuri de activitate.*

Natura activităților executate de blocul de control este dependentă de context, chiar pentru aceleaşi dimensiuni ale calității dalelor. Un bloc de control de calitate poate fi asumat cu scanarea unui formular, pentru a se identifica informațiile lipsă înainte ca formularul să fie procesat. Un alt tip de control de calitate privind completitudinea se referă la verificarea primirii datelor aferente tuturor surselor de intrare pentru un interval de timp dat. Un bloc de control al acurateței în acest caz realizează compararea datelor din perioada curentă cu datele din perioada cea mai recentă pentru care se dețin date arhivate, pornindu-se de la premisa că nu pot exista diferențe majore între acestea.

Figura 5 reprezintă un exemplu de sistem de producere de informații. În acest sistem există cinci module de date primare *($DU_1 \div DU_5$)* furnizate de trei surse diferite *($VB_1 \div VB_3$)*. Există trei module de date *(DU_6 , DU_8, DU_{10})* rezultate în urma trecerii prin unul din cele trei blocuri de calitate *($QB_1 \div QB_3$)*. De exemplu, DU_6 reprezintă impactul lui QB_1 asupra lui DU_2. Corespunzător celor şase blocuri de procesare *($PB_1 \div PB_6$)* sunt şase module de date *(DU_7 , DU_9 , DU_{13} , DU_{14} , DU_{15} , DU_{16})* care reprezintă ieşirea din aceste blocuri de procesare. Există un bloc de memorare *(SB_1)*.

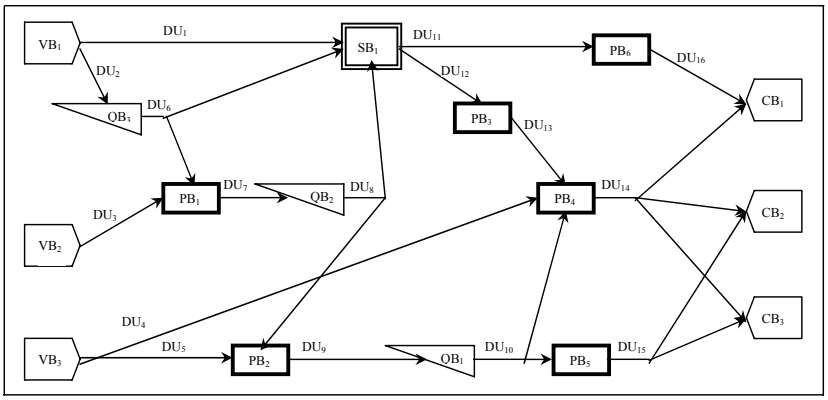

Fig. 5. Schema unui sistem de prelucrare de informații

Sistemul are trei beneficiari *($CB_1 \div CB_3$)*, fiecare dintre ei primind câte un subset de date. De remarcat faptul că în blocurile PB_1 şi PB_2 sunt procesate copii ale lui DU_6, respectiv DU_8. Amplasarea unui bloc de calitate după un bloc sursă de date indică faptul că datele furnizate de sursa respectivă sunt în general calitativ deficitare.

Modelarea procesării datelor

Modelarea sistemelor de producere de informații se face la un nivel de detaliere corespunzător scopului analizei. De exemplu, efectul blocului de control poate fi stabilirea ponderii modulelor de date suspecte în totalul modulelor de date. La un grad de detaliere mai avansat, blocul de control filtrează şirurile datelor în subşiruri aparent corecte şi suspecte. Subşirurile suspecte sunt analizate în continuare şi trec prin activități corective corespunzătoare.

Modulele de date au asociați vectori de caracteristici ale căror componente se schimbă ca rezultat al trecerii prin diferite stadii de procesare. Semnificația lor este de asemenea dependentă de context. De exemplu, dacă toate câmpurile pentru toate înregistrările dintr-un anumit fişier au aceleaşi caracteristici de oportunitate şi calitate ale datelor şi dacă întregul conținut al fişierului este procesat în aceeaşi manieră, atunci acest fişier va fi tratat ca fiind un singur modul de date. Invers, atunci când câmpurile dintr-o înregistrare diferă în mod

semnificativ în ceea ce priveşte atributele de oportunitate şi calitate, se va realiza modelarea lor în mod individual.

Pentru determinarea efectului diferitelor transformări ale configuraţiei procesului de prelucrare asupra caracteristicilor datelor, se foloseşte *matricea de analiză a procesării datelor*. Structura acestei matrice este (figura 6):

- n linii, corespunzătoare celor n fluxuri de date DU_i, unde $i = 1,n$;
- $m=p+s+q+u$ coloane, corespunzătoare celor k blocuri de prelucrare, s blocuri de memorare, q blocuri de control calitate şi u blocuri utilizator;
- $Z_{i,j}$ elementele matricei de analiză, unde $j=1,m$; aceasta are valori numai în celulele pentru care fluxul de date i traversează blocul de activitate j; valorile caracteristicilor de calitate la ieşirea dintr-un bloc de activitate reprezintă intrările în blocul de activitate următor.

	PB_1	… …	PB_p	SB_1	… …	SB_s	QB_1	… …	QB_q	CB_1	… …	CB_u
DU_1	$Z_{1,1}$		$Z_{1,p}$	$Z_{1,p+1}$		$Z_{1,p+s}$	$Z_{1,p+s+1}$		$Z_{1,p+s+q}$	$Z_{1,p+s+q+1}$		$Z_{1,p+s+q+u}$
…												
DU_i	$Z_{i,1}$		$Z_{i,p}$	$Z_{i,p+1}$		**$Z_{i,p+s}$**	$Z_{i,p+s+1}$		$Z_{i,p+s+q}$	$Z_{i,p+s+q+1}$		$Z_{i,p+s+q+u}$
…												
DU_n	$Z_{n,1}$		$Z_{n,p}$	$Z_{n,p+1}$		$Z_{n,p+s}$	$Z_{n,p+s+1}$		$Z_{n,p+s+q}$	$Z_{n,p+s+q+1}$		$Z_{n,p+s+q+u}$

Fig. 6. Matricea de analiză a procesării datelor

Valorile semnificative pentru matricea de analiză sunt elemente vectoriale, cu cinci componente, *(p, t_1, t_2, Dqi, Costi),* cu următoarea semnificaţie:

- p reprezintă simbolul blocului predecesor;
- t_1 este timpul când unitatea de date este disponibilă pentru activitate;
- t_2 reprezintă momentul începerii procesării, deci după ce toate unităţile de date necesare sunt disponibile. De asemenea, procesarea poate începe la momentul programat, t_{prog}. De aici calcularea lui t_2 ca valoare maximă dintre cea mai mare valoare a timpilor aferenţi unităţilor de date de intrare în procesare şi timpul programat:

$t_2 = max(max(t_1), t_{prog})$

- *Dqi* se referă la calitatea datelor de intrare pentru o activitate particulară, şi este egală cu valoarea calităţii rezultată la ieşirea din blocul predecesor;
- *Costi* este valoarea costului unităţilor de date de intrare şi reprezintă cumulul costurilor tuturor activităţilor precedente pe care le-a traversat o unitate de date.

Acest model este destinat analizării şi reproiectării sistemelor de producere a datelor prin determinarea valorilor atributelor produsului informatic. Modelul urmăreşte calitatea, costul şi oportunitatea unităţilor de date dintr-un sistem. Calitatea, în acest context, cuprinde toate atributele care determină integritatea datelor, cu excepţia oportunităţii. Costul este asociat variantelor de îmbunătăţire a calităţii datelor.

Fie $x_1, x_2, ..., x_n$ cele n intrări aferente unui bloc de procesare.

În scopul măsurării integrităţii produselor dată furnizate la beneficiari, *calitatea* unităţilor de date este calculată după fiecare operaţie executată asupra lor.

Fie $DQ(x_i)$ o măsură a calităţii unităţii de date x_i în domeniul aferent unei scale de la 0 la 1, cu 1 reprezentând date fără probleme de calitate şi 0 datele de o calitate intolerabilă. Către un bloc de activitate converg una sau mai multe unităţi de date, fiecare cu un nivel specific al calităţii. Măsura calităţii datelor furnizate la beneficiar, *DC*, este asociată cu rezultatul prelucrării datelor la intrarea în blocul utilizator. Fie unităţile de date $x_1, x_2, ...x_n$ corespunzătoare funcţiei de prelucrare $y=f(x_1, x_2, ...x_n)$. Dacă toate unităţile de date la intrarea în blocul utilizator au valoarea 1 a calităţii, atunci măsura integratoare, *DC*, va fi de asemenea 1. Invers, dacă la intrarea în blocul utilizator converg numai unităţi de

date cu nivel calitativ 0, atunci va rezulta un nivel calitativ general 0. Din considerentele formulate mai sus, formula de calcul a calității datelor furnizate la beneficiar va fi media aritmetică ponderată a calităților $DQ(x_i)$: $DC = \frac{\sum_{i=1}^{n} w_i * DQ(x_i)}{\sum_{i=1}^{n} w_i}$, unde $wi = \left|\frac{\partial f}{\partial x}\right| * |xi|$.

Această formulă înglobează atât efectul calității datelor de intrare cât şi al eficacității procesării, fluxul datelor primare fiind condus şi analizat pe traseul complet al transformării lor. Dacă procedeul de calcul adoptat pentru determinarea calității datelor vizează numai etapele premergătoare fluxului de prelucrare, atunci calitatea datelor furnizate beneficiarului acestora este funcție de calitatea datelor de intrare *(DQ_{int})* şi eficacitatea procesării *(PE)*: *DC=f (DQ_{int}, PE)*. Dacă procesarea este complexă, deci numărul şi interacțiunea parametrilor de referință sunt multiple, pentru stabilirea relației de calcul a calității produsului informatic livrat se recurge la analiza prin simulare a acestor factori.

Bibliografie

[Anan00], Ananthanarayana, V.S., Subramanian, D.K., Murty, Scalable, M.N.: *Distributed and Dynamic Mining of Association Rules*, In Proceedings of HIPC'00, pages 559-566, Bangalore, India, 2000.

[Ashr02], Ashrafi, M.Z., Taniar, D., Smith. K.A.: *A Data Mining Architecture for Distributed Environments*, IICS 2002, pages 27-38, 2002.

[Ball94], Ballou, D.P.,Wang, R., Pazer, H.,Kumar.Tayi, G.: *Modeling Information Manufacturing Systems to Determine Information Product Quality*, Massachusetts Institute of Technology (MIT) Sloan School of Management, Cambridge, MA, TDQM-94-06, August 1994.

[Free03], Freeman, L.A.: *A refresher in data flow diagramming: an effective aid for analysts*, Communications of the ACM (CACM), Volume 46, Number 9, September 2003, 147-151.

[Krov03], Krovi, R., Chandra, A., Rajagopalan, B.: *Information flow parameters for managing organizational processes,* Communications of the ACM (CACM), Volume 46, Number 2, February 2003, 77-82.

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