Data Metrics

Ion Ivan - Bucharest Academy of Economic Studies, ivan@fx.ro
Otilia Parlog - Romanian Defence Ministry
Petrisor Oprea - Bucharest Academy of Economic Studies
Gheroghe Nosca - Romanian Defence Ministry
Anca-Andreea Ivan - University "POLITEHNICA" of Bucharest

Introduction

After a first period in developing software exclusively dedicated to the quality of software products, the interest is reaching now towards data. The development of software tools, CASE systems and other informational technologies reduces the software research effort. The number of applications for accessing public databases with a special, accelerated dynamic of data volume is increasing.

The article's objective is to define data metrics and to exemplify their use. The data indicators and sets will be developed in both structural and behavioural plans.

The data stored in databases can be the result of observations, measurements or it can be generated. It can be numeric, alphanumeric, alphabetic data, graphic symbols, sounds, colours, or combinations.

Let's consider an alphabet formed by the symbols a_1, a_2, ... , a_k and a rule for word construction. The words c_1, c_2, ... , c_n built using the alphabet A form a vocabulary, V. A word c_i represents a data when it is put in correspondence with an element from the real world.

The data enters the communication process as messages. One or more data senders and one or more receivers are necessary to accomplish the communication process.

The ratio data/information is extremely complex. While all information can be considered as data, only the data that triggers the generation of other data and actions is information.

Defining strict criteria to allow the extraction of information from databases represents a very rich field of research.

Considering the digit alphabet $A_1=\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$, all integers are words of the V_N vocabulary; when the numbers are associated with events, phenomena, evolutions, objects, they will form data sets. For example, the next elements in table no. 1 are extracted from the V_N vocabulary:

Table no. 1. The production evolution of enterprise X in the last 10 years

Year	Annual production (thousands)
1988	130
1989	135
1990	150
1991	170
1992	188
1993	165
1994	200
1995	210
1996	205
1997	230

The years and the production levels are both words in N and data because they have an associated significance to a real world phenomenon and they are showing an evolution.

In a similar manner, the words of a language are constructed from the letter alphabet, or from an alphabet of sounds. This alphabet becomes very rich when it contains intonations as symbols.

For the coloured point alphabet, the results are images if using the rules of disposing them on columns and rows.

Construction Indicators

The alphabet length is the number of symbols. If working with images, the length is given by multiplying the number of rows, columns and colours.
The vocabulary length is given by the number of words included in the vocabulary.
The usual vocabulary is built with the words used in a certain time period.
The subvocabulary represents a set of words extracted from the vocabulary according to a set of rules.
The text is a set of words, characterised by their position related to the others.
The word length is given by the number of symbols which are forming the word.
The symbol frequency in a word is given by the number of appearances of a distinct symbol. For example, in the word "papaya", the symbols p, a, y have the frequencies 2, 3, 1.
The symbol frequency in the vocabulary is the result of summing the symbol frequency from all the words.

All these indicators are computed when there are restrictions regarding the methods of word generation. The *word length* is a restrained indicator when only a usual vocabulary is necessary.

For example, for a speedy recognition of the car numbers, the following alphabet and word generating rules are built:

The next hypotheses are considered:
- the structure of a word is:

xx-yy-zzz

- *xx* - marks the city by letters (capitals of the Latin alphabet);
  - contains the county and the Bucharest symbols, totally 41 words
- *yy* - group of two digits, exclusively 00;
  - represents 99 words;
- *zzz* - marks a combination of capitals from the Latin alphabet, with no special significance, less the combination with not an adequate sense (30 words)

$$zzz = C_{25}^{3} - 30 = 2.300 - 30 = 2.270$$

- the maximum length of this vocabulary is:

$$Lvmax = 41 * 99 * 2.270 = 9.213.930$$

If we have a text T, its analysis is based on the next indicators:

- text length, as number of words;
- vocabulary length;
- appearance frequency for all the words in the text;
- particularities of word (distance between words, successive repetitions, cacophonies);
- appearance particularities of some symbols on beginning or ending positions.

Moreover, if we have a reference vocabulary V_R, the words that do or do not belong to the reference vocabulary can be highlighted by comparing the words from the text T with the words of V_R.

When taking into account the word generation rules it is possible to identify a part of the word, called root, which is the base in forming other words, therefore forming a family.

The vocabularies with generated words are sufficiently well described to allow the verification of their words. Such vocabularies are formed from words called existent words in the vocabulary, that are used with the given significance.

For the existent vocabularies the length is determined distinctly for basic words and derived words. There are many aggregative or derivative methods used with basic words: concatenation, combination, etc.

After building the dictionaries, the resulting number of words and expressions will constitute the dictionary length.

The degree of word similitude measures the resemblance of rules, symbols and positions used in word construction.

The word orthogonality corresponds to the zero level of similitude.

The programming languages, for example, are designed for maximizing the orthogonality level.

Let's consider two words representing two images, each one consisting in a table of coloured pixels, with m rows and n columns. We'll use the following notation for a pixel:

- $\alpha 1_{ij}$, for the image 1;
- $\alpha 2_{ij}$, for the image 2.

One of the next relations can be established between the two images:

a) Images are totally different

Two images are totally different if all the corresponding points are different:

$\alpha 1_{ij} \# \alpha 2_{ij}$, $\forall i \in [1,m]$ and $\forall j \in [1,n]$

b) Identical images

This case represents the opposite of the previous one; two images are identical if all the corresponding points are identical:

$\alpha 1_{ij} = \alpha 2_{ij}$, $\forall$ i $\in$[1,m] and $\forall$ j $\in$[1,n]

c) Symmetrical images

This relation can be analysed only for square images, with the same number of rows and columns.

The following criteria are defined:

- symmetry towards the main diagonal of the table;
- symmetry towards the secondary diagonal of the table;
- symmetry towards both diagonals of the table;
- symmetry towards the median column of the table;
- symmetry towards the median row of the table;
- symmetry towards the median column and row of the table.

c.1) Images with symmetry towards the main diagonal

Two images are symmetrical in this case if the corresponding points are observing the following relation:

$\alpha 1_{ij} = \alpha 2ji$, $\forall$ i, j$\in$[1,n]

c.2) Images with symmetry towards the secondary diagonal

Two images are symmetrical in this case if the corresponding points are observing the following relation:

$\alpha 1_{ij} = \alpha 2_{n-j+1,n-j+1}$, $\forall$ i, j$\in$[1,n]

The other cases of symmetry can be defined in a similar manner.

The degree of similitude establishes the relation observed by the pixels within the two images for all the symmetry cases shown above.

The degree of identity for two images, Gid, is given by the ratio:

$$Gid = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\alpha_{ij}}{m*n} \qquad (1)$$

where:

α_{ij}= 1 if the points of the row I and the column j are identical;
α_{ij}= 0 otherwise.

The degree of symmetry towards the main diagonal of two images, Gsp, is given by the ratio:

$$Gsp = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\alpha_{ij}}{n*n} \qquad (2)$$

where:

α_{ij}= 1, if $\alpha 1_{ij} = \alpha 2ji$
α_{ij}= 0, if $\alpha 1_{ij} \# \alpha 2ji$

The degree of symmetry towards the secondary diagonal of two images, Gss, is given by the ratio:

$$Gss = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\alpha_{ij}}{n*n} \qquad (3)$$

where:

α_{ij}= 1, if $\alpha 1_{ij} = \alpha 2_{n-j+1,n-j+1}$
α_{ij}= 0, if $\alpha 1_{ij} \# \alpha 2_{n-j+1,n-j+1}$

The degrees of symmetry for all other cases can be defined in a similar manner.

The texts can differ in length, word appearance frequencies or distance between words. The same applies for texts containing graphic symbols, sounds or images.

In all the cases, there can be identified certain lengths, frequencies, repetition intervals and distances.

In the text:

5 25 73 100 131

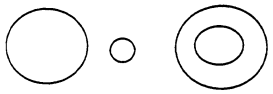
the following relation can be identified:

$5 < 25 < 73 < 100 < 131$

and the conclusion is the increasing tendency of the phenomenon.

In the text: :"Apples, apples and apples", we can notice the repetition of the word "apples".

In the text:



the same symbol of circle is used several times, with different dimensions.

Data Complexity

A phenomenon, process or collection is described by time, space, natural characteristics, as well as by type and quality characteristics.

The description of a problem to be solved with a software product is clear, complete and correct if the data reflects the consistent aspects of reality.

A problem can be estimated in comparison to another one, if the volume and the complexity of calculations is quantified.

In the same way we can judge the data sets that define a problem as simple or complex.

To measure the data complexity, it is necessary to consider:
- the diversity of data types;
- the number of appearances;
- the links between data.

Let's consider the data related to the financial operations of an enterprise, shown in table no.2.

Table no.2. Data regarding the financial operations of an enterprise

Account code	Initial sold		Transaction		Final sold	
	Debit	Credit	Debit	Credit	Debit	Credit
c_1	d_{11}	c_{11}	d_{12}	c_{12}	d_{13}	c_{13}
c_2	d_{12}	c_{21}	d_{22}	c_{22}	d_{23}	c_{23}
c_n	d_{1n}	c_{n1}	d_{2n}	c_{2n}	d_{n3}	c_{n3}
Total	TD1	TC1	TDR	TCR	TDF	TCF

The data complexity, is given by:
- the number of decomposition levels;
- the number of items corresponding to the number of columns in the table.

In the table no. 2, the data is organised on 3 levels.

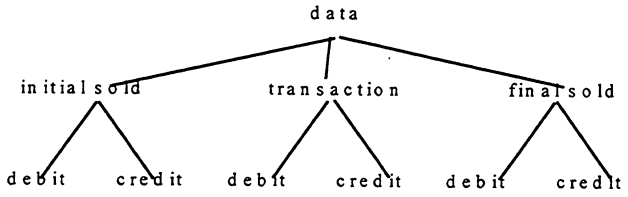


Figure 1. The tree structure associated to the data regarding the financial operations of an enterprise

For our case, the number of columns is 7.

The data complexity results from the aggregation between the number of items and the number of links, as shown here:

$$C = n_1 \log_2 n_1 + n_2 \log_2 n_2 \qquad (2)$$

where:

n_1 - number of items;

n_2 - number of operators.

The next data links are resulting from the table no. 2:

$d_{i3} = d_{i1} + d_{i2}$

$c_{i3} = c_{i1} + c_{i2}$

$TDI = \Sigma d_{i1}$

$TCI = \Sigma c_{i1}$
$TDR = \Sigma d_{i2}$
$TCR = \Sigma c_{i2}$
$TDF = \Sigma d_{i3}$
$TCF = \Sigma c_{i3}$

For a number n of accounts, the complexity C is computed like this:

$n_1 =$ n account items +
4n number of prime values +
2n calculated data (final credit and debit) +
6 total values on the last row

$n_1 = 7n + 6$

$n_2 =$ n additions of initial debit + transaction debit +
n additions of initial credit + transaction credit +
6 (n - 1) additions on columns to obtain the row of totals

$n_2 = 8n - 6$

therefore $C = (7n + 6)\log_2(7n + 6) + (8n - 6)\log_2(8n - 6)$

The table no. 3 shows data regarding the economic evolution of a country.

Table 3. The economic evolution for a set of countries

Country	Population	PIB	Productivity	.....	.....
1	*2*	*3*	*4*	.....	*k*
t_1					
t_2					
.....					
t_m					

The columns contain primary, independent, non-processed data.

A single operation is considered: grouping the data on rows, depending on a parameter ($n_2 = 1$).

Therefore $C = (m * k)\log_2(m * k)$.

The complexity is exclusively based on operands, since their contribution is null ($1*\log_2 1 = 0$). The existence of a decreasing ordering after one column or an alphabetic order of countries, modifies the level of complexity.

Therefore $C = (m * k)\log_2(m * k) + 2\log_2 2$

The text complexity problem is strictly dependent of:
- the number of basic words that form the text (n_1);
- the relationships between the words expressed by the number of operator words (n_2);
- the number of link words (n_3).

The text complexity can be computed by using the formula:

$$C = n_1 \log_2 n_1 + n_2 \log_2 n_2 + n_3 \log_2 n_3 \qquad (3)$$

For example:

Ana has *apples.*
Nelu has *oranges.*
Maria is walking and singing.
Laura and *Gigi* are walking, listening to music and looking *straight ahead.*

In the previous text, the verbs are operators n_2=7 (the underlined words). The words in italic typefaces, n_1=8 can be considered as basic.

The last category is consisting in the link words (prepositions) n_3= 4.

The text complexity is:

$$C = 8\log_2 8 + 7\log_2 7 + 4\log_2 4$$

Generally speaking, a vocabulary containing k word types (substantives, adverbs, prepositions, conjunctions, verbs, adjectives) will have a complexity given by:

$$C = \sum_{i=1}^{k} n_i \log_2 n_i$$

The complexity of a graphic representation is given by:

- the number of colours;
- the organisation of coloured spots;
- the neighbourhood ratio.

In the image below, it is shown the population diagram of a country structured on nationalities.

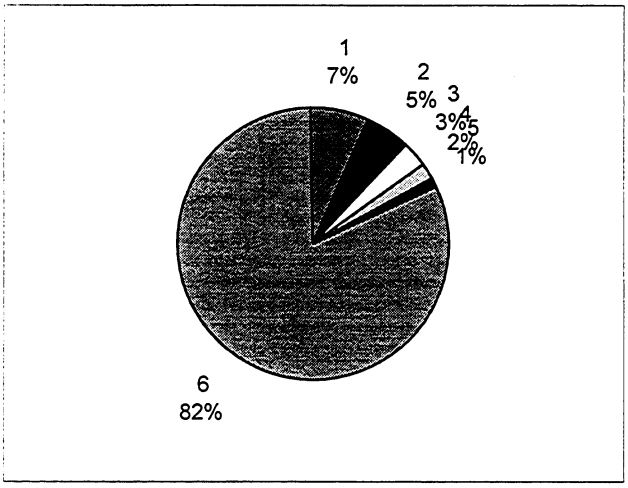


Figure 2. The population of a country structured on nationalities

The fond enters the counting with the value 1. If there is only the fond::

$$C = 0 \qquad (1\log_2 1 + 1\log_2 1 = 0)$$

There are n sectors. The operators are:

n_1 - number of items

n_1= n+1, corresponding to the n sectors + the fond

n_2 - number of operators

In this case, the operators are the comparisons of two consecutive sector values.

- concatenation n-1;

- comparisons the newest n-1 , followed by 7% and the smallest 1% (sorting + concatenation).

$$C = (1+n)\log_2(1+n) + (n-1+n-1)\log_2(n-1+n-1)$$

where:

1 - fond;
n - number of sectors;
n-1 - comparisons;
n-1 - concatenations.

The data complexity norm is given by the formula:

$$M = \frac{\sum n_i \log_2 n_i}{(\sum n_i)\log_2(\sum n_i)}$$

For the example shown in figure no. 1, the data complexity norm is computed like this:

$$M = \frac{n_1 \log_2 n_1 + n_2 \log_2 n_2}{(n_1+n_2)\log_2(n_1+n_2)}$$

Data Behaviour Indicators

There are several indicators related to data behaviour:
- reliability indicators;
- reusability indicators;
- completitude indicators;
- correctitude indicators.

Let's consider a problem I, corresponding to the software product S and the input data sets $D_1, D_2, ..., D_p$, that lead to the results $R_1, R_2, ..., R_p$ at the moments of time $t_1, t_2, ..., t_p$.

The software product S was tested and was found to be running correctly. The results R_i are analysed and the conclusion will be whether they are correct or not.

A conglomerate is constructed as a pair $[D_i, S]$. When the conglomerate leads to an error we say that it malfunctions; the cause is the data D_i because the product S was already tested. The conglomerate is regarded as an ensemble formed by serial linked sub-ensembles.

Reliability indicator:

$$f = \frac{m}{p}$$

where: m - the number of data sets with correct results;
p - the total number of data sets.

The reliability models are applied by studying the intervals between two failures.

Causes of non-reliability:
- internal causes that generate processing actions, others that the tested ones;
- incomplete data;
- non-consistent data;

Consecutive reliability indicator is given by the ratio:

$$r = \frac{H}{H^*}$$

where:

H - the common data sets that are used at the moments t_i, t_{i+1}

H* - the reunion of data sets at the moments t_i (H_i) and t_{i+1} (H_{i+1}), $H^* = H_i \cup H_{i+1}$

For example, at the moment t_i the data from table no. 4 is used.

Table 4 The level of indicators for the enterprise X at the moment t_i .

Year	Business figure	Number of workers	Profit
1990			
....			
1995			

At the moment t_{i+1} the data from table no. 5 are used.

Table 5. The level of indicators for the enterprise X at the moment t_{i+1} .

Year	Business figure	Number of workers	Profit
1993			
...			
1997			

In this context the data volume is computing using the formula:

$$V_i = nl_i * nc_i$$

where:

nc_i - the number of columns at the moment t_i

nl_i - the number of rows at the moment t_i

In the table no. 4, $V_i = 6 * 4 = 24$ and in the table no. 5 $V_{i+1} = 5 * 4 = 20$.

The elements from table no. 4 are used for the data corresponding to the period 1993-1995, representing the period of time when data was written both in table no. 4 and in table no. 5 ($V(H_i \cap H_{i+1})$; $V = 3 * 4$).

The total volume of data, associated to the period of time covered by the data from tables 4 and 5 ($V(H_i \cup H_{i+1})$), covers the period 1990-1997. ($V^* = 8 * 4$).

The reusability indicator for the data regarding the enterprise X, is given by the ratio:

$$r = \frac{3*4}{8*4} = 0{,}37$$

The data homogeneity is dependent of the measurement units. A measurement unit has the attributes α, β, δ, ...

We will define measurement units such as meters of material, wage/hour, number of students, value in money. In the table no. 6 it is shown the population evolution for the counties

$j_1, j_2, \ldots, j_n$ at the moments $t_1, t_2, \ldots, t_m$; n_{ij} represents the number of persons from the county i at the moment j.

Table 6. Population evolution in counties

County	t_1	t_2	...	t_j	...	t_n
j_1						
j_2						
...						
j_i				n_{ij}		
...						
j_m						

All the elements inside the table are expressed in the same measurement unit - number of persons. In this case, the homogeneity degree is maxim, equal to 1. In the case of table no. 7, with descriptions of an enterprise employees, we have:

Table 7. Descriptions of an enterprise employees

Number of employee	Name of employee	Number of hours	Wage	No. of children	Age	Experience	Category	Education

The homogeneity degree is small for this case.

The homogeneity degree is defined like this:

$$G_o = \frac{1}{p} * 100$$

where: p represents the number of distinct characteristics with corresponding measurement units to describe the elements of a data set.

For the data file with the employees descriptions, the homogeneity degree is:

$$G = \frac{1}{9} * 100$$

The general formula is:

$$G = \frac{\max(n_1, n_2, \ldots, n_k)}{\sum_{i=1}^{k} n_i}$$

where:

n_i - is the number of elements corresponding to a characteristic;

k - is the number of distinct characteristics.

There is always a difference between the data available at a given moment (D_d) and the necessary ones (D_n) to the calculation of some indicator for the real time management. The comparison of the two data sets allows the evaluation of the *completitude degree*.

For example, in order to make an investigation with an error degree of 3%, there are necessary 1200 employees (D_n). The specific conditions lead to gathering data from only 900 employees (D_d). It is said that the estimations were done on an incomplete data volume. The completitude degree:

$$G_c = \frac{D_d}{D_n} * 100$$

for the taken example is:

$$G_c = \frac{900}{1200} * 100 = 75\%$$

In the case of completing a table, the absence of records for different moments, leads to blanks. The solution is to consider a hypothesis regarding the evolution (linear, polynomial, exponential) , to interpolate and to replace the incomplete data with the interpolated result.

The data absence from tables affects the quality of results, even if we interpolate missing data. In a table with m rows and n columns, and k missing values, the completitude degree G_c is:

$$G_c = \frac{m * n - k}{m * n} * 100$$

The comparison degree is a basic characteristic in the analysis processes of the structure and dynamics of a collectively or a phenomenon.

The comparison is possible only for homogeneous data.

In certain situations a series of factors affects the characteristics levels evolution, making them not comparable.

For example, the prices of a product are affected by inflation, structural modifications of the production technologies, work productivity increase and changes in the request/offer ratio.

Table no. 8 shows the price evolution of some products.

Table 8. The price evolution of some products

Year	Unitary price (lei)		
	Bread	Pretzels	Simple croissant
1980	3	0.50	0.75
1988	4.50	0.50	1.00
1990	10.00	1.50	1.25
1992	200.00	20.00	10.00
1993	600	50	60
1995	800	200	200
1996	900	300	250
1997	1100	500	600
1998	1200	450	550

The data x and y are comparable if the function associated to the homogeneity process is an identical function:

$$y = f(x), \forall x, y \in D$$

$$x = f(x)$$

To ensure the comparison, there are transformation functions for an element x_{i-1} to x_i . Most of the times $x_i' = k_{i-1} * x_{i-1}$, where k is a coefficient.

If the series of time given by the triplets (t_i, x_{i-1}, k_{i-1}) , where k_{i-1} is the comparison ensuring coefficient, is considered

$$x_i' = k_{i-1} * x_{i-1}$$

all combinations of comparison bases are constructed.

For the countries where the inflation is a constant for the economy, it is necessary to transform the data gathered in different periods of time in order to make the comparison possible. For example, the "business figure" indicator is associated with the "ratio leu/dollar" for comparison.

Depending on the requested analysis, the first year of the analysed interval, a randomly chosen one or the last one is set as the base year for comparison.

The purpose is the transforming of business figures, measured in lei, into updated values that can be compared.

The table no. 9 shows the annual business figure of an enterprise, in the interval 1980÷1998.

Table no. 9. The evolution of the annual business figure of an enterprise

Year	Business figure (lei)	Equivalent leu/dollar K_i
1989	300	1
1990	400	30
1991	600	60
1992	1000	300
1993	15000	510
1994	22000	750
1995	26000	900
1996	24000	1200
1997	28000	1500
1998	30000	1400

a) <u>The base of comparison is the first year of the analysis period</u>

The ration between the equivalent leu/dollar for two successive years $\alpha_i = \frac{k_{i-1}}{k_i}$

The coefficient used to obtain the updated business figure is

$$\beta_i = \begin{cases} 1, & \text{for the base year} \\ \prod_{m=1}^{i} \alpha_m, & \text{otherwise} \end{cases}$$

The updated business figure is obtained using the formula:

$$CA\ act_i = \beta_i * Ca_i$$

where:- $CAact_i$ is the updated business figure of year i

- β_i is the transformation coefficient of year i
- CA_i is the non-updated business figure of year i

The table no. 10 exemplifies the method of computing the updated business figure for this period.

Table no. 10 . The computation of the updated business figure with the first year of the analysed period as base year

i	Year	Business figure (lei) CA_i	Equivalent leu / dollar k_i	$\alpha_i = \frac{k_{i-1}}{k_i}$	$\beta_i = \prod_{m=1}^{i} \alpha_m$	Updated business figure $CA\ act_i = CA_i * \beta_i$
0	1989	300	1	-	1	300
1	1990	400	30	0,33	0,33	132
2	1991	600	60	0,5	0,165	99
3	1992	1000	300	0,2	0,033	33
4	1993	15000	510	0,6	0,0198	297
5	1994	22000	750	0,66	0,013068	287
6	1995	26000	900	0,83	0,010846	282
7	1996	24000	1200	0,75	0,008134	195
8	1997	28000	1500	0,8	0,00641	179
9	1998	30000	1400	1,07	0,00686	206

b) <u>The comparison year is the last year of the analysis period</u>

The ratio between the equivalent leu/dollar for two successive years is computed as:

$$\alpha_i = \frac{k_{i-1}}{k_i}$$

The coefficient used to obtain the updated business figure is

$$\beta_i = \begin{cases} 1, & \text{for the base year} \\ \prod_{m=i+1}^{n} \alpha_m, & \text{otherwise} \end{cases}$$

The updated business figure is obtained using the formula:

$$CAact_i = \frac{CA_i}{\beta_i}$$

where: - $CAact_i$ is the updated business figure of year i

- β_i is the transformation coefficient of year i
- CA_i is the non-updated business figure of year i

The table no. 11 exemplifies the method of computing the updated business figure for this period.

Table no. 11. The computation of the updated business figure with the last year of the analysis period as base year

i	Year	Business figure (lei) CA_i	Equivalent leu / dollar k_i	$\alpha_i = \frac{k_{i-1}}{k_i}$	$\beta_i = \prod_{m=i=1}^{n} \alpha_m$	Updated business figure $CAact_i = \frac{CA_i}{\beta_i}$
0	1989	300	1	-	0,0069	43.478
1	1990	400	30	0,33	0,021	19.047
2	1991	600	60	0,5	0,042	14.286
3	1992	1000	300	0,2	0,211	4.739
4	1993	15000	510	0,6	0,352	42.625
5	1994	22000	750	0,66	0,533	41.276
6	1995	26000	900	0,83	0,642	40.498
7	1996	24000	1200	0,75	0,856	28.037
8	1997	28000	1500	0,8	1,07	26.168
9	1998	30000	1400	1,07	1	30.000

c) The comparison base is a year included in the analysis period

The ratio leu/dollar is computed for two successive years $\alpha_i = \frac{k_{i-1}}{k_i}$

The coefficient used in obtaining the updated business figure is

$$\beta_i = \begin{cases} \prod_{m=i+1}^{j} \alpha_m, & \text{for } i < j \\ 1, & \text{for } i = j \\ \prod_{m=1}^{i} \alpha_m, & \text{for } i > j \end{cases}$$

The updated business figure is obtained using the formula:

$$CA\ act_i = \begin{cases} \beta_i * CA_i, & \text{for } i < j \\ CA_i, & \text{for } i = j \\ \frac{CA_i}{\beta_i}, & \text{for } i > j \end{cases}$$

where: - $CAact_i$ is the updated business figure of the year i
- β_i is the transformation coefficient of year i
- CA_i is the non-updated business figure of year i

In the table no. 12, it is shown the method of calculation the updated business figure for this case.

Table 12. The computation of the updated business figure with the base year included in the analysed period

i	Year	Business figure (lei) CA_i	Equivalent leu / dollar k_i	$\alpha_i = \frac{k_{i-1}}{k_i}$	β_i	Updated business figure $CA\ act_i$
0	1989	300	1	-	0,0131	22.900
1	1990	400	30	0,33	0,0396	10.101
2	1991	600	60	0,5	0,0792	7.575
3	1992	1000	300	0,2	0,396	2.525
4	1993	15000	510	0,6	0,66	22.727
5	1994	22000	750	0,66	1	22.000
6	1995	26000	900	0,83	0,83	21.580
7	1996	24000	1200	0,75	0,6225	14.940
8	1997	28000	1500	0,8	0,498	13.944
9	1998	30000	1400	1,07	0,5328	15.984

The comparison degree is given by the update complexity:

$C = n_1 \log_2 n_1 + n_2 \log_2 n_2$

where:

n_1 - the number of terms of levels;

n_2 - the number of operators.

For the division case

$$k_i = \frac{\alpha_i}{\alpha_{i+1}}$$

and for the multiply case:

$$y_i' = k_i * y_i$$

each element has two operations:

$$C = n \log_2 n + (2n) \log_2 (2n)$$

$$\mathbf{GComparability} = \frac{\mathbf{n_1 \log_2 n_1 + n_2 \log_2 n_2}}{\mathbf{(n_1 + n_2) \log_2 (n_1 + n_2)}}$$

For the case of comparable data it is necessary to do the operation $n_2=1$, representing the comparison operation.

$$C = \frac{n_1 \log_2 n_1}{n_1 \log_2 n_1}$$

Conclusions

The data metrics offers essential information to the processing activity:

- it estimates the processing time;
- it estimates the precision of results;
- it shows the quality of results.

It is important to know the data quality, considering its link to the cost and to establish the convenient moment for an increase in quality.

The data gathering plan is essential for processing. Designing software not related to the volume and quality of data leads to inefficient processing. The cost of data can make a software product inoperable.

Moreover, the dimension of the data volume optimisation ensures the desired processing cost and effects.

References

Arthur, L., "Improving Software Quality", Wiley, New York, 1993

Ivan, I., G. Nosca, A. Parlog, "Data Quality Assurance, Quality Assurance Review", vol. 2, no.8, 1998, pp. 8-14 (in Romanian)

Ivan, I., P. Sinioros, F. Simion, M. Popescu, "Software Metrics", Inforec Publishing House, Bucharest, 1997 (in Romanian)

Orr, K., "Data Quality and Systems Theory", CACM, vol. 41, no. 2, 1998, pp. 66-71

Strong, D. M., Y. W. Lee and R. Y. Wang, "10 Potholes in the Road to Information Quality", IEEE Computer no. 8, 1997, pp. 38-46