CUSTOMER DATA INTEGRATION (CDI): PROJECTS IN OPERATIONAL ENVIRONMENTS

(Practice-Oriented)

Flávio de Almeida Pires
Assesso Engenharia de Sistemas Ltda
flavio@assesso.com.br

Abstract. To counter the results of increasing globalization and competition, companies have moved from a primarily *product* focus to a *customer* focus. This has elevated the issues concerning Data Quality to a high priority level among corporations, concerned as they are with the need for accuracy in dealing with their customers. This article presents a framework to be used in the assembly of a Customer Information Database (CID) from information generated by different legacy systems, and in the implementation of Data Quality processes to control and maintain the quality of this kind of information, in those cases in which the CID should run in complex operational environments, requiring high level of quality and performance.

BACKGROUND

This paper is based on the experience acquired by Assesso, a Brazilian technology enterprise that has been working with Data Quality and CID Assembly projects for more than 18 years. During this period of time, Assesso developed more than 35 projects in these areas, always for large companies, and always dealing with a substantial volume of information.

Based on the experience acquired, Assesso has developed and successfully markets a software for data treatment and unification that has become one of the market leaders in Brazil, and has also established a “framework” for CDI project development, effectively and efficiently used in its projects.

During this past year, both the framework and the software have been adapted to include the Total Data Quality Management (TDQM) principles, proposed by the Massachusetts Institute of Technology (MIT), and also to incorporate some locally developed innovative concepts and procedures that we believe will contribute to enhanced data quality processes as a whole.

INTRODUCTION

In the mid 1980s, companies started to change their focus strategy from a Product vision to a Customer vision. In order to adapt to this new situation, Information Technology (IT) areas adopted the solution of implementing parallel databases. Initially, these were for marketing purposes and, afterwards, for business performance analysis. However, the legacy systems were kept product or process oriented.

It was at this point that Database Marketing projects were put together. These were followed by Data Warehouse projects, which, in addition to having independent databases, were run in a separated environment to prevent the risk of any interference with the operational processes of the companies.

These projects dictated the creation of the corporate customers' databases as well as the importance of data quality. It was as a result of the need to implement this that the "cleansing" and ETL (extraction, transformation and loading) software emerged and/or was strengthened. Customers deduplication processes also gained great importance, as this was the basis of assembly of a complete profile of clients, which was, after all, the main objective of this kind of project.

Deduplication processes have always had, as a result, four levels of answer possibilities:

- These customers certainly are the same;
- These customers certainly are different;
- A grayish zone, where an automatic identification of a likely duplicate is not possible;
- And, eventually, records, where the necessary attributes for the process are not present, whether because they have not been filled in, or because they have information containing unrecoverable errors – a condition detected by several consistency processes (domain, digit check and others).

The current market practice for Database Marketing or Data Warehouse implementation projects has been to establish, case by case, the level of rigidity of the deduplication process to be used, adjusting it to the objectives of each particular project.

For example, for Database Marketing projects, (i.e. a direct mail campaign), an "overkill" process was normally used, even though this could generate improper deduplications. Its primary value was that it assured lower costs for the campaign, particularly when what was being sent had some material value.

On the other hand, for Data Warehouse projects, an "underkill" process was used, minimizing improper deduplication, even running the risk of not deduplicating some cases, because by doing this the distortions in the business analysis – to be carried out on the unified data base – were minimized.

In practice, the result was the elimination of the grayish area described above, therefore also eliminating the need for "back-office" additional work, since a certain level of errors was an accepted part of the process.

This situation changed with the consolidation of the "focus on the customer" concept, and with the launching of CRM (Customer Relationship Management) programs. More than a transitional trend, this has become a permanent tendency: Management techniques come and go, but CRM is not a trend – we are sure of that. Even though it is difficult to implement, CRM is a powerful idea.

Regarding Data Quality, already in 2001, a Strategic Analysis Report (SAR – Nov 26, 2001) by Gartner Group stated: "A clean, well-integrated data repository is a prerequisite for CRM success. . . . Integrating customer data is a challenging task, however, because merely assembling the data in one place is not enough. A number of steps must be performed, including data extraction, transformation, cleaning and consolidation".

In reality, this new perception has substantially changed the quality levels required for the assembly and maintenance of Customer Information Database (CID). Even small percentages of errors were no longer acceptable. For the deduplication cases, how to accept improper unification of two distinctive clients, and send invoices and products to the wrong individual? Or how to accept the non-unification of several registers of one client, and deal with this client as if he were different person?

In addition to the new and high quality levels established, the challenge became even greater with the consolidation of another perception as well: the need for an updated vision of the customer.

This means the need of integrating the CID to the operational environment, with real time updates, and meeting performance levels that fit the requirements of this type of environment.

Within this new reality, and in order to fit these new requirements, CID's implementation processes were forced to change.

In our case and in our framework, one of the changes was the adoption of a new concept, or a new goal, that we called "Seeking 100% Quality".

The first understanding was that we would not be able to meet that goal by using only automated processes. We needed to “separate the wheat from the chaff” and give a special treatment for the grayish zone and for the registers containing erroneous attributes. Thus, we quickly started to strongly recommend to our clients to assembly a “back-office” area to treat theses cases. Additionally, we encouraged them to appoint a “quality manager” for the process, a professional with the specific task of monitoring and controlling data quality, aiming at its continuous improvement.

In order for our recommendations to be successful and economically feasible, a refinement of our processes was necessary, so that the workload of the “back-office” area would be minimized.

It was at this point, initially, that we adopted the “RYG Concept” (Red, Yellow, Green), defined below: by bringing together several deduplication rules and processes, it was possible to adequately group the cases of each of the four possible answers of the process, according to the specifications mentioned above:

Red – The necessary attributes for the deduplication have errors (fields that have not been filled in and/or with inconsistent information) – These cases should be forwarded to the back-office area, and contacting the client directly will be necessary in order to correct them;

Yellow – The automatic process of deduplication is not capable of assuring that it is a duplicate. However there is a significant chance that this is the case (“Suspects of Duplication” – the grayish area mentioned above) – These cases should be forwarded for visual analysis by the back-office area, and contacting the client may become necessary;

Green – The automatic process is capable of assuring that it is a duplicate case, or, on the contrary, that it definitively is not a case of duplication – These cases are then regularly used or updated in the process.

By using the RYG Concept, our challenge was to make the process maximize the Green cases and minimize the Yellow cases (we have less control over the Red cases, which basically depend on the original quality of the file being treated – here our actions are focused on the recommendations for improvement of the information capture processes).

This step was followed by the acknowledgement that the RYG Concept is generic, considering that it can be efficacious in the treatment of general attributes, of sets of attributes, or of business rules. Afterwards, we adopted it in our entire framework as something definite. Some examples:

<u>Date of birth</u> (age) –After today or age greater than 100 years old – **Red**; age under than 18 or greater than 80 years old – **Yellow**;

<u>Address</u> – the informed ZIP code does not belong to the city – **Red** ; ZIP code is within city range, however the street was not recognized – **Yellow**;

<u>Life Insurance</u> **-** above US$ 3,000,000 (not accepted by the company) – **Red**; people under 18 years old and with insurance above US$ 100,000 **- Yellow.**

The framework described below contemplates the necessary adaptations for these changes, as well as the adoption of TDQM principles proposed by MIT.

CHARACTERISTICS OF CDI PROJECTS IN OPERATIONAL ENVIRONMENTS

CDI projects aiming at the CID – Customer Information Databases - implementation in operational environments normally fall under one of the two cases described below:

- Where a first step for CID assembly is forecast and, afterwards, all legacy systems are adapted for the utilization of the implemented CID;
- Where we still have a first step for CID assembly, but the redevelopment of all legacy systems is forecast, now designed for the customers’ vision.

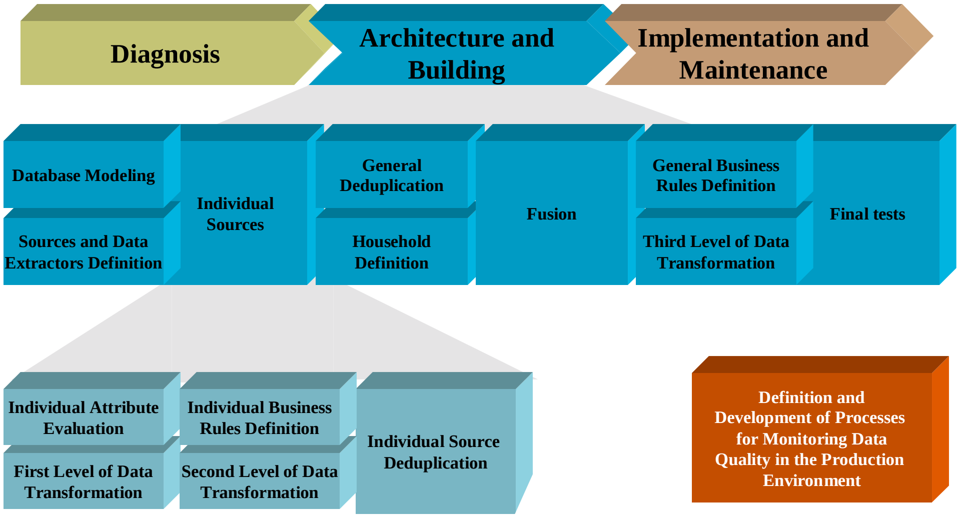
Therefore, in both cases, normally, the first step is the CID implementation, which is done at once, followed by the gradual implementation of each of the legacy systems – either adapted or redeveloped.

In this article, we are not dealing with legacy system adaptations or redevelopments. However, they should incorporate all the rules and solutions defined herein, and the framework also presents some warnings in respect of their adaptation or redevelopment.

Rather, this article is focused on business to consumer projects. For business-to-business projects, whereas the framework is the same, several other specific peculiarities have to be considered.

FRAMEWORK FOR OPERATIONAL CDI PROJECTS

The chart bellow gives a quick view of the framework developed by Assesso for the implementation of CDI Projects in operational environments.


Diagnosis
Architecture and Building
Implementation and Maintenance
Database Modeling
Sources and Data Extractors Definition
Individual Sources
General Deduplication
Household Definition
Fusion
General Business Rules Definition
Third Level of Data Transformation
Final tests
Individual Attribute Evaluation
First Level of Data Transformation
Individual Business Rules Definition
Second Level of Data Transformation
Individual Source Deduplication
Definition and Development of Processes for Monitoring Data Quality in the Production Environment

As can be seen above, the framework is divided in three large steps, each of it with its specifics objectives and products, as follows:

Diagnosis

<u>Products</u>: Scope definition, schedules, costs and benefits of the project.

In this step, in addition to the complete understanding of the business, the following issues should be analyzed: project objectives, the company's short, medium and long term strategies, risks and opportunities, business processes, involved systems – platforms, operating environments, information sources, volumes, business rules, new environment macro specification (including the necessary adaptation for current systems and new developments), and the benefits of the new environment.

Architecture and Building

Products: Project specification, development and tests, definition of processes and end-user homologation rules.

Aiming at having a more efficient organization of the work and mainly at the possibility of carrying out tasks simultaneously, this extensive step was divided in eight sub-steps.

For this entire step and for each of its sub-steps, we adopted the cycles defined by TDQM by MIT: define, measure, analyze and improve. As such, each of the sub-steps is a sequence of cycles that ends with end-user homologation.

An important point to be stressed is that in this step the tasks are carried out on the total of available information and never samples. Our experience has shown that analysis performed on samples can frequently hide errors such as "filling-in bad habits" ("easy typing"), bugs that legacy systems carried for a period of time, attributes to be used for deduplication purposes that were not properly populated, or other similar difficulties.

These eight sub-steps are a basic guideline for CDI projects in operational environments. For every new project, this framework is reevaluated and adapted to its specific conditions.

For more information and details about these eight sub-steps, please contact Assesso.

Implementation and Maintenance

The points to be stressed in this step are as follows:

- The need to create an area to be in charge of data quality for the company;
- This area should regard information as a product and should have the responsibility and the empowerment to modify the existing processes, if needed, in each of the three areas involved, according to TDQM from MIT: Data Collectors, Data Custodians and Data Users (Customers).
- Data quality monitoring and assurance processes defined in the previous step to be regularly run and adapted to new situations.

Other items related to this step should follow normal standards, which are defined by each company's IT.

CONCLUSION

As a result of market evolution, there is currently and will be a growing demand for the implementation of Data Quality processes in operational environments. This will securely lead to new requirements for the processes, such as higher quality and performance levels, which are forcing companies to adopt new frameworks to their development.

These new requirements are not yet stabilized; therefore the framework to be adopted should be flexible enough to incorporate further evolution.