ETL Process Model for a Manufacture Cells Production Line Integration

Harison Pereira Bila de Carvalho  
*TCS – Tata Consultancy Services*  
- Brazil

Gabriel de Souza Pereira Moreira  
*ITA – Aeronautics Institute of Technology - Brazil*

Danilo Battaglia  
*TCS – Tata Consultancy Services - Brazil*

Luiz Alberto Vieira Dias  
*ITA – Aeronautics Institute of Technology - Brazil*

Denis Ávila Montini  
*TCS – Tata Consultancy Services - Brazil*

Paulo Marcelo Tasinatto  
*ITA – Aeronautics Institute of Technology - Brazil*

**Abstract**

This article describes a Causal Analysis and Resolution of a problem for the elaboration of a Data Mart – DM specialized in Production Control and Planning – PCP. The unified information in the proposed DM is loaded form Management Information Systems – MIS, with historic data from software development project. To allow the integrated load from different data sources, it was developed a Extract, Transform and Load – ETL process, whose implementation has integrated data from the MIS of Function Point Analysis and Personal Software Process – PSP in a software factory. That allowed the construction of a Decision Support System – DSS over the constructed DM. This DSS allows the use of Online Analytical Processing – OLAP to support quantitatively the Process Areas of Measurement and Analysis – MA of Capability Maturity Model Integration – CMMI.

Key Words: FPA, ETL, OLAP, Data Mart, CMMi.

1. Introduction

Researchers from TCS and Research Group of Software Engineering (*Grupo de Pesquisa de Engenharia de Software – GPES*) from the Brazilian Aeronautics Institute of Technology (*Instituto Tecnológico de Aeronáutica – ITA*), based on some Process Areas of Capability Maturity Model Integration – CMMI, created a process of Extract, Transform and Load – ETL to software factories specialized in Line Production Lines of Manufacture Cell – PLMC [1].

It was specified, dimensioned, projected and tested a process, by means of Causal Analysis and Resolution – CAR to integrate in a Data Mart information of executed software projects [2] from different data sources. The created Data Mart consolidates information from the MIS of software estimation by Function Point Analysis – FPA [3] and of MIS of Personal Software Process – PSP, keep in the corporation.

The objective is to allow the use of Online Analytical Processing – OLAP [4] to the Process Area of Measurement and Analysis – MA from Capability Maturity Model Integration – CMMI. The diverse concepts of software engineering and quality that was used in the construction of a Decision Support System – DSS using OLAP [5] formed the base to the developed Control Projects System.

The Data Mart – DM was modeled in a way to supply a base to Knowledge Discovering in Databases – KDD [2], like Data Mining – DM.

The SIG’s project followed the Service Oriented Architecture – SOA [6] to the development of the needed environment to the integration models.

The work objective was to develop an ETL process with the technical requirements of SOA application to supply a DM of a PLCM.

2. Research Context

ETL is a process methodology for Extracting, Transforming and Loading data from transactional databases to multidimensional databases, like Data Marts – DM or Data Warehouses – DW. The focus of a DW is to store the information in a way that allows Decision Support for enterprise business with high-performance.

The ETL phase is the longest and hardest of the construction of a DW [7]. It can be used to uniform and integrate the diverse data sources from a PLMC, which have a systematic process to register the behavior and software results engineering teams [8] [9] [10]. The ETL can help the decision and administration process of an enterprise together with a tool set of methodological foundation to identification and analysis. In our case study, ETL has collaborated in the calibration of dimensioning variables of PLMC projects, by means of data loading from FPA repository [3].

The creation of a Decision Support System aimed to ease the knowledge discovery and allowed the identification of patterns doesn’t easily perceived by manual analysis through OLAP. The most relevant patterns and trends can
3. Problem Definition

The initial concern was to identify the dimensions to be modeled from the transactional MIS. It is required the multidimensional modeling of DW before the ETL process definition. After that, it was identified some alternatives to the resolution of automated generation project reports process, whose information would be provided by the DW modeled.

To the execution of the automated ETL, it was necessary the Extraction activities planning, Transforming and Load. The first part of this ETL process involved data extraction from the MIS databases of FPA and PSP. After that, Transformation activities were critical to ensure that the data can be grouped and summarized like the DW model. Transformation usually involves rules application and functions over the raw data extracted from data sources to allow its compatibility with the model.

The problem statement was: “To refine ETL methods to integration of MIS in DW of software engineering projects, which allowed the gauge of dimensioning by means of automated generation of technical reports” [12].

Data Marts are multidimensional databases generally specialized in one department of a company. When it is obtained data from more than one area, that interest to the general goals of the company, it can grow up to a Data Warehouse. So, DW can emerge from top-down or bottom-up approach. Top-Down is when a company creates a DW and after that goes to segmentation, dividing in minor areas, oriented by department subjects. In Bottom approach, the company prefers first create a Data Mart for most important or critical departments. That approach costs less than the DW and reduce risks. Then, when the Data Marts requirements are mature enough, the focus goes to their integration in a DW. The hardware and software infrastructure of DM and DW can be similar, but the data architecture can be significantly different.

The SOA defined platform unified the needed requirements to the technological implementation of the business in a three-tier environment. In the database tier, it was adopted the Bottom-up strategy to obtain the DW from the MIS FPA and PSP. The objective was to allow OLAP analysis to supply a tool for decision taking.

It was chosen as OLAP tool the SQL Server 2000 that allowed all the process of collection, storage and publishing of the needed information. It was adopted Key Performance Indicators – KPIs to provide to the management quick insights about the projects situation.

The identified problem was applied over the software engineering projects by means of a computational subsystem that allowed a new control and automation of MIS models, filled with manual data entries from distributed databases [13].

4. A software problem CAR

The process re-engineering was based on monitoring data of the process. The CAR process applied has oriented the existing refinement collection data manipulation process and information about dimensioning and software control. The start point of the process was to understand the reason for dimensioning data tabulation lateness provided by FPA. From this point, it was initialized a process of improvement of this bottle-neck. In the data analysis, the company aimed to get a new level of knowledge on your own production.

Each software development company knows how much of its resources to the execution of a specific contract. It is fundamental to strategic planning maintenance that projects, to be developed, are monitored and its limitations are known. With this strategic vision, the embedded problems in software development process can be identified and resolved on time. It is a fundamental point to justify the use of OLAP.

The start point to the difficult found in a PLMC of TCS was a reverse engineering realization. Based on the concepts suggested by Sommerville, an analysis process must have a study object, and in this case it was the information automation process about the software dimensioning projects in a PLMC. The technical capability to recover information about the original project could be applied aiming to optimize the next projects [14].

For the enterprise continuous offering of projects and its survival in the competitive business, TCS had to analyze its competitors and observe their performance. The comparing with its own results has occurred to fundament a strategic decision in production lines to the products offering and services.

The conscious competition administration loss allowed performance improvement re-planning activities. In this vision, the institution has performed processes revisions aiming to find a continuous improvement for its process, to revitalize your organizational projects. A commission of Research and Development – R&D analyzed objective indicators against the monitored Process Areas that
distanced from the established targets. That commission needed to assess the estimation analysis, because the process indicators signaled constant instability.

Then, the reverse engineering was inserted in this context to allow finding the root of the problem identified.

It was identified a bottle-neck during the integration lack MIS distributed databases. It was identified during a technical requirements realization survey to technological feasibility studies.

The mentioned analysis is CAR process product for each of the stages, activities and use of technologies during the execution process. The analysis of primary data has occurred during the operation by the R&D. The survey identified the fundamental information flow, supplying the elements to the CAR studies. The analysis CAR focus was performance improvement to the productivity and efficiency increase, according to the continuous improvement rules established by CMMi.

During the dimensioning process analysis and software estimations consolidation, by FPA means, it was found an opportunity to improve a process that had all character to be automated. The R&D has established a technology that would attend the specification and has proposed severe changes, replacing a manual process by an automated solution, which is more efficient, systemic and structured [14][16].

4.1 The identification of problem patterns

The vulnerabilities capacity identification in strategic planning has needed to be monitored by the use of indicators. When the deviations appeared, it was used a CAR methodology to identify the vulnerabilities. The methodology was oriented to the resources optimization and the correction decisions to realize the factory strategic planning operation [17].

<table>
<thead>
<tr>
<th>Layer: Identification</th>
<th>Indicator</th>
<th>Theory</th>
<th>Practical</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - Conception:</td>
<td>2 - Analysis</td>
<td>3 - Modelling:</td>
<td></td>
</tr>
<tr>
<td>Problem</td>
<td>PDCA</td>
<td>Solution</td>
<td></td>
</tr>
</tbody>
</table>

4.1.1 Identification stage:

1 - It is not enough for a particular system to be constituted only by indicators to indicate an error, it is necessary that this system is more developed and have the capacity to provide the link between the performance of activity with a scale, referenced with the area knowledge. The scale parameterization that uses the indicator to point to result, with the knowledge involved in the procedure is the key to the indicator can isolate the techniques, methods, technologies and knowledge areas that might contain the solutions to possible irregularities to be verified.

2 - There is a need to use an audit system that allows bringing data about the process and how to analyze them so that these data can make some sense. Based on the type of business and adoption of views, norms and standards, the team of R&D should, with a well defined methodology, develop its audit instrument to meet their specific needs [1]. The indicators should be included in the instrument created and calibrated so the data provided by the indicators make sense to analysts and for this, it is necessary to create a scale of values identifying the critical points allowed for each one of the indicators.

In this set, an indicator in particular provides specific information about a particular sizing process performance and estimating software. The bottleneck indicated a potential problem of the project and, in this case the limiting factor was the large amount of errors caused by manual data entries and the difficulty of obtaining OLAP analysis. With the identification of this problem was possible through an ETL component, check whether its use had eliminated or reduced the intensity or effect of the issue found.

At this point the PDCA methodology was recommended for the team of R&D to develop a reasonable alternative solution to the problem specified [1].

3 – The modeling process the solution was performed by the R&D team in accordance with the specifications.

4.1.2 Development stage:

4 - Performed the planning solution phase begins to stage in creating a prototype that will code replace, process of system defective. Every step of developing a

In the presented methodology the information are collected directly from the operation project, in our case, from the FPA and PSP MIS databases. The methodology is based on three initial visions: problem identification, solution development and its homologation of each realized stages. Each one parts is called layers, and each of them is divided in three parts, described below.
new indicator should be done until the point of working together with the previous system.

The team of R&D needs to avail themselves of all artifacts related to the process to ensure the correct record of the occurrence, when the team stabilize the process and indicator does not present the problem in an activity similar to the replacement should be implemented successfully.

5 - For complex cases in which the solution is beyond the material and intellectual resources of the occasion, a research stage to obtain this knowledge is central to solving the problem.

Conducted research and made the knowledge acquisition, a new prototype should be done and analyzed again by their method.

6 - The team of R&D should follow the development team to make sure that the implementations were carried out according to the determined.

4.1.3 Homologation Stage:

7 - During development, the R&D team can verify if the implementation and the requirements have been met in accordance with the request.
8 - The new process has been tested in operation and had to stay under observation up to the validation after batteries of tests in production.
9 - To prevent some collected knowledge part was lost, and depending on complexity, the project type and risk involved, the R&D teams have created a technical report on the component's project.

But what is the advantage achieved by this reverse engineering process as sophisticated? The advantage was that with this scientific method used, the problems are replaced by a formal treatment as well as its solution.

The first part of the problem identification of system studied was the classification of the patterns of problem presented in Table2 [1].

Table 2: Classification Generic Standards Issues

<table>
<thead>
<tr>
<th>Generic Pattern Classification Problem</th>
<th>Issues Standards</th>
<th>Solutions Standards</th>
</tr>
</thead>
<tbody>
<tr>
<td>Architectural: The structural organization of systems hardware and software.</td>
<td>1. The R&amp;D team is organized to define a methodology and stabilize the engineering process of software development.</td>
<td>ETL was the methodology defined.</td>
</tr>
<tr>
<td>Project: The organization of activities around tasks for the performance of a product.</td>
<td>2. The team develops a model of standards to be used in the project. Each standard must be configured with specific techniques for each step. The techniques need to check and test the technology and standard processes. With these records, there is evidence that the effective resolution is to which proposed.</td>
<td>Were identified a pattern of structured coding ETL component.</td>
</tr>
<tr>
<td>Process: The organization in an orderly manner in the task within an activity.</td>
<td>3. If your system standard has not identified any problems, through research, identify the focus of the problem, in database query or other known sources in addition to the documentation of patterns.</td>
<td>The parameterization of the data structures of the two distributed databases to be integrated with the data mart.</td>
</tr>
<tr>
<td>Programming: Introduces the programming solutions to implement the requirements requested by the client. example - the choice of language.</td>
<td>4. When a problem is very specific and however the company has not a solution to the need.</td>
<td>Was developed an ETL standard for the integration of new distributed databases.</td>
</tr>
<tr>
<td>Analysis: The process of verification and validation of documents and records to ascertain whether the specification of a client was treated in accordance with the contract.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This action was essential to that records observations obtained allowing an analysis applied later. The assembly of the DM with data basis distributed with APF estimates from projects was the first step to identify the behavior of LPCM. In this study the data collected came from more than one basis by estimates of APF project to be processed individually in order to obtain a specialized DM.

The Table 2 the team of P&D found that the patterns of problems identified in this analysis (CAR) was the analysis pattern. Based in the distributed projects databases the production teams began manual analysis activities for an APF atmosphere parameters definition. In created DM they were accomplished manual consolidations that originated difficulties in an acquisition maintenance of these registrations.

After the process focus bottle neck identification, P&D team solved mapping process. This resolution is presented in the Picture 3.

Table 3: Process for Identifying Patterns of Problems and Solutions

<table>
<thead>
<tr>
<th>Identification Process</th>
<th>Solutions Standards</th>
</tr>
</thead>
<tbody>
<tr>
<td>Issues Standards</td>
<td></td>
</tr>
<tr>
<td>1. The R&amp;D team is organized to define a methodology and stabilize the engineering process of software development.</td>
<td>ETL was the methodology defined.</td>
</tr>
<tr>
<td>2. The team develops a model of standards to be used in the project. Each standard must be configured with specific techniques for each step. The techniques need to check and test the technology and standard processes. With these records, there is evidence that the effective resolution is to which proposed.</td>
<td>Were identified a pattern of structured coding ETL component.</td>
</tr>
<tr>
<td>3. If your system standard has not identified any problems, through research, identify the focus of the problem, in database query or other known sources in addition to the documentation of patterns.</td>
<td>The parameterization of the data structures of the two distributed databases to be integrated with the data mart.</td>
</tr>
<tr>
<td>4. When a problem is very specific and however the company has not a solution to the need.</td>
<td>Was developed an ETL standard for the integration of new distributed databases.</td>
</tr>
</tbody>
</table>

The MIS identification scenario needed to be supported by a technology that had the capability to integrate flocks distributed data. In research in the database area, methods that suited the needs identified by the analysis were met by the ETL method. Complementing the reasoning line presented during the developing ETL component process with an approach to be chosen by the team of R&D. Since the beginning of the process sought the establishment of a working prototype.

5. Reduced scope
The MIS integration model has established specifications for the construction of an SQL Server 2000 ETL component, which contains the MIS of APF and PSP in order to meet the subjects terms studied in the program. Were organized activities, processes, methods, techniques and software engineering tools to integrate the solution according to guidelines development project [11].

The quality of products and processes will be used in the project with the quality software engineering standards along with the CMMi guidelines [1]. One research products was a CAR methodology for ETL components creation.

6. Development ETL component

To compose an instrument new, there are following steps to create the prototype of analysis' instrument, so presented himself process defined that enable the development, though is possible future considerations. The proposed study has the following steps in Table 4:

To implement the ETL component was necessary a theoretical and practice study on the technology, this step is part of the learning process. Based on the data structure, the modeling provided an analysis of ETL could be applied to determine the requirements integration specification. One of the requirements were functional ETL verification and validation. Table 5 presents the main testing activities process of User. For the data collection and technical information necessary for the development of the ETL component used the Delphi method [1].

The test plan for validation of the component passed the ETL functionality needed to load the instrument. The test performed was derived standards for software products shown in Table 5. After making the selection of test's types' criteria's for the ETL component, began the process validation and verification described in Table 6.

<table>
<thead>
<tr>
<th>Table 4. Process of development of ETL.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Development process of the ETL component</td>
</tr>
<tr>
<td>1 - Formulation of the problem.</td>
</tr>
<tr>
<td>2 - Planning research.</td>
</tr>
<tr>
<td>3 - Elaboration of ETL component.</td>
</tr>
<tr>
<td>4 - Load Testing.</td>
</tr>
<tr>
<td>5 - Verification of the data.</td>
</tr>
<tr>
<td>6 - Data analysis and interpretation of data and results.</td>
</tr>
<tr>
<td>7 - Presentation of data and results.</td>
</tr>
<tr>
<td>8 - Verification and Validation.</td>
</tr>
</tbody>
</table>

The Picture 5 presents main activities of user's test process. For data rising, it was necessary technical information for component ETL development used - if method Delphi [22].

The plan of tests for a component ETL validation went by necessary functionalities for instrument load. The accomplished test was derived of norms for software products shown in the Picture 5. This picture presents the main process of user's test activities. The tests plan for the component ETL validation went by necessary functionalities to load this instrument.

<table>
<thead>
<tr>
<th>Table 5. Criteria testing for product.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test criteria for product</td>
</tr>
<tr>
<td>1. Reliability: Data storage reliability indicators, data integrity, backup, security, ability to handle failure and risk analysis.</td>
</tr>
<tr>
<td>2. Usability: The usability criteria are defined as follows: consistency User Interface, Internationalization, ease of learning, quality documentation tool, quality training material, diagnostic clarity, ease of installation, acceptable response time and help on - line</td>
</tr>
<tr>
<td>4. Efficiency: Ability of developing an instrument that their requirements for data storage are fast and dynamic.</td>
</tr>
<tr>
<td>5. Effectiveness: capability to correctly portray the analysis.</td>
</tr>
<tr>
<td>6. Portability: The ability to move data and share them in network environments.</td>
</tr>
</tbody>
</table>

The test performed was derived for software products standards shown in Table 5. After making the selection of test's types' criteria's for the ETL component, began the process validation and verification described in Table 6.

<table>
<thead>
<tr>
<th>Table 6. ETL Validation and Verification Process.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process ETL verification and validation</td>
</tr>
<tr>
<td>1. Find at least three examples in which the load of data was performed effectively using the same solution.</td>
</tr>
<tr>
<td>2. Declare the hypothetical solution as standard candidate.</td>
</tr>
<tr>
<td>3. Execute an activity of the Delphi method to improve the description of the candidate to share itself with others.</td>
</tr>
<tr>
<td>4. Implement the ETL candidate for another project of software development.</td>
</tr>
<tr>
<td>5. Declare the standard candidate if your application is successful</td>
</tr>
</tbody>
</table>

With all the planning done, begins to prototype construction step to a production line [9] [1]. The implementation must to be monitored [12], since it becomes the bridge between planning and bring data. If it is not very well thought, the work might be compromised by increasing project risk. [1].

6. 1. Specification model suggested
The project included three models dimensions: Generic model, Dynamic and Static. The first is static model proposition, which will be represented by static structures of the system data.

The second used a Dynamic Model to reflect the temporal and behavioral aspects. The proposed control will require the last dimension, the use of a Functional Model, through a working prototype, which was to describe the features for creating this system.

Figure 1. Static model of ETL for DT

We have found three basic features for the ETL concept to the prototype use:
- Capacity Customization: it was checked SIG grounded in process PSP and another with the APF method.

Figure 2. Dynamic Model.

Modifying the query parameters to be using the data and information on other future projects and integrate with other GIS as the SCP was one of the basis for the DT composition.
- Capacity of Modularization: it was developed a modular system, which can work in a cumulative way to be used in other research projects complementary.
- Ability to reuse components: Create a structure that allows the reuse for other research projects.
- Ability to Interact: Integrate a DM through a MIS system for the ETL automation. The creation of this ETL integration model will allow the development research lines, existing, and may contribute to future work of KDD.

Figure 3. Functional Model ETL SQL 2000.

7. Results

After the dimensions choice available and of interest to the DT assembly, we attempted itself to establish procedures for the creation of the PSP application model integrated with the SIG of APF to DM of software's PCP of a LPCM through technology ETL. The implementation system through the application software, it was enabled to improved project management software performance effectively, to enable the OLAP analysis in the LPCM production. The aim was to develop a model's architecture's application with interact ETL and GIS, for the establishment of a project.

When we use ETL model such implementation object related to the problematic studied with the quality models, CMMi [11], applied to software engineering to provide increased efficiency in the area of quality. In this case the phenomenon studied was process continuous improvement supported by the system automation of data collection for the PA to MA. The concepts presented and their applications will be possible with the participation of the research group of TCS [9]. To achieve the elements projects presented previously are requirements for the capability of building this application architecture model.

8. Conclusion

The data integration in a systematic way constitutes a historical errors source that hinders learning about the behavior and design projects in LPCM. The APF reports and the PSP allowed the information consolidation through the ETL components to input in the OLAP system from a TD. This DM improved the project control in 4% in time information consolidation dimension.

An ETL tool should not require much knowledge of SQL programming; however is not what happens in practice. There is a need to identify professionals with knowledge in SQL, because the interface is not intuitive as well as the tools. The team needed be able to create the ETL component. The ETL was done with professional who knew the programming language. The main contribution of this paper was the ETL's implementation
in DM to factory's software from LPCM. Represented by a process defined the PSP. Without the formal process presented, the methodology used for line production becomes uneconomical to maintain the software factory business. The ETL created such standard was used SQL Server 2000 and the modeling visual applied, generated more than 1400 lines of code (LOC) in SQL Server, performed on a component. The ETL methodology presented requirements for the assembly of DM the SP and DM of LPCM.

The creation of this system is advisable as to the manual activities use, consistency and your gain is realized in that production use is obtained on a large scale. To continue this research is recommended:
1. The use of concepts of ETL,
2. Understanding the methodologies of DM and DW projects,
3. Knowledge of PCP for CMMi, TSP and PSP for use on production lines software such as manufacturing cells LPCM [1] [9] [12].

For future work, it is suggested that research results are applied to the use of data mining in, for example, in Operational Research (Pesquisa Operacional - PO).

9. Acknowledgments

The authors of this research would like to thank:
1. When GPES for support the development of methodological research;
2. ITA by the availability of resources for infrastructure software;
3. Tata Consultancy Services (TCS) the possibility of partnerships with the productive sectors academics.

10. References

[1] Montini, Denis Ávila; Moreira, Gabriel De Souza; Vieira, Luiz Alberto; Battaglia, Danilo; Gnatiuc, Carlos Eduardo; Cunha, Adilson Marques Da; Estudo de caso de uma estratégia de integração de middleware para um serviço SOA de gerenciamento e controle de fábrica de software: TCS – Tata Consultancy Services – Intranet website de Base de Conhecimento Corporativa KnowMax: TACTICS Iberoamerica 2007, Brasil, São Paulo, SP 25 - 26/Outubro/2007.


[9] Montini, Denis Ávila; Moreira, Gabriel De Souza; Vieira, Luiz Alberto; Battaglia, Danilo; Gnatiuc, Carlos Eduardo; Cunha, Adilson Marques Da; Estudo de caso de uma estratégia de integração de middleware para um serviço SOA de gerenciamento e controle de fábrica de software: TCS – Tata Consultancy Services – Intranet website de Base de Conhecimento Corporativa KnowMax: TACTICS Iberoamerica 2007, Brasil, São Paulo, SP 25 - 26/Outubro/2007.


