Data Warehouse Engineering Process (DWEP) with U.M.L.

2.1.1. A Case Study: “Central library of Ibague”

Edwar Javier Herrera Osorio  
Facultad de Ingeniería  
Departamento de ingeniería de Sistemas e Industrial  
Universidad Nacional de Colombia  
Bogotá, Colombia  
ejherrerao@unal.edu.co

Elizabeth León Guzmán  
Facultad de Ingeniería  
Departamento de ingeniería de Sistemas e Industrial  
Universidad Nacional de Colombia  
Bogotá, Colombia  
eleonguz@unal.edu.co

Carlos Andrés Lugo González  
Facultad de Ingeniería  
Ingeniería de sistemas  
Universidad de Ibagué  
Ibagué, Colombia  
calugog@unal.edu.co

ABSTRACT

In this paper, we present the use of Data Warehouse Engineering Process (DWEP) for designing a data warehouse for Central Library of the University of the Ibague, in order to make the conceptual, logical and physical models. The DWEP was updated to UML 2.1, and is called DWEP 2.1, in this version we added five diagrams with respect DWEP for total of 21 artifacts. The updating of DWEP was developed by the software implemented in “Eclipse Galileo” by Eclipse Modeling Framework (EMF) and Graphical Modeling Framework (GMF).

Categories and Subject Descriptors

H.2.7 [Database Administration]: Data warehouse and repository – DWEP, UML, Eclipse.

General Terms


Keywords

UML, Data Warehouse, Data Warehouse Engineering Process.

1. INTRODUCTION

The University of the Ibague, is an educational institution with high standards are recognized nationally and internationally. The university is interested in supporting and driving processes of development and progress, joining forces with different leaders and entrepreneurs to build a better future for the region and the country.

The university library serves is available to the whole region across the following services: circulation and loan, interlibrary loan, specialized bibliographies, letters, warning service, reference service, newsletters. The volume transactional lending library is about 50,000 records per year considering all users, users are divided into internal (active students, teachers, faculty and administrative facility of the University Antonio Nariño) and external (students, teachers, researchers from other institutions, companies and public or private organizations that have agreements with the Universidad Antonio Nariño) where a total of the 5,000 users.

The methodology Data Warehouse Engineering Process (DWEP), as this provides a development of the data warehouse through four (4) phases and through seven (7) core workflows of the datawarehouse based on the UP, this makes the development is independent from any specific implementation. This development allowed the implementation of the library of the institution to improve the delivery of information, facilitate decision-making process and improving the processes carried.

The organization of this paper is as follow: In Section 2, a summary to Data Warehouse Engineering Process (DWEP). In Section 3 the DWEP is updated with UML 2.1.1. In Section 4, the update DWEP is applied to a case study specifically an library. Finally, concluding remarks and plans for future research appear in Section 5.

2. DATA WAREHOUSE ENGINEERING PROCESS (DWEP)

DWEP is a methodology based in UP [1, 2]. The UP is a methodology for software development proposed by OMG [4], its main features are: it is iterative, is addressed by the use cases is based on stages of development, using UML as a graphical language models [5, 6].

The DWEP can be describe in two dimensions: The horizontal axis represent time and shows the dynamic aspect of the process (Phases and iterations), and the vertical axis represent the static aspect of the process (Core Workflows). In figure 1 show the diverse iteration between phases and core workflows.

![Figure 1. Data Warehouse Engineering Process [1]](image-url)
According to the DWEP, the project lifecycle is divided into four phases: Inception, Elaboration, Construction, and Transition and seven core workflows: Requirements, Analysis, Design, Implementation, Test, Maintenance and Post-development. During the developing of a project, the emphasis shifts over the iterations, from requirements and analysis towards design, implementation, testing, maintenance and post-development review, but different workflows can coexist in the same iteration.

### 3. UPDATING DATA WAREHOUSE ENGINEERING PROCESS 2.1 (DWEP 2.1)

The DWEP [1,2] is updated to DWEP 2.1 using UML 2.1.1. This new version has six (6) new diagrams to facilitate implementation of this methodology.

The new diagrams are: Source Conceptual Object Schema (SCOS), this artifact depicts instances and their relationships at a point in time the analysis of the data source. Source Logical Communications Schema (SLCS) shows the message flow between objects in an object-oriented application, and also implies the basic associations (relationships) between tables. Data Warehouse Sequence Schema (DWSS) shows the dynamic modeling in the data warehouse, Data Warehouse state machine Schema (DWSMS) this artifact depicts the dynamic behavior of an entity based on its response to events, showing how the entity reacts to various events based on its current state. Data Warehouse activity Schema (DWAS) Schema is the object-oriented equivalent of flow charts and data-flow diagrams from structured development, and finally Data Warehouse Conceptual Object Schema (DWCOs) this artifact depicts instances and their relationships at a point in time the analysis of the data warehouse.

Figure 2 shows the process of activities for the designing and development of the winery through DWEP 2.1 applying its twenty-one (21) artifacts based on UML 2.1.1

### 3.1 DWEP 2.1 Phases

The DWEP 2.1 uses the same phases of DWEP [1, 2] for the development of data warehouse.

**Inception:** This phase is to understand the scope and objectives of data warehouse and getting enough information to confirm that you should proceed or perhaps convince you that you shouldn't.

**Elaboration:** The goal this phase is to define and baseline the architecture of the DW in order to provide a stable basis for the bulk of the design and implementation effort in the Construction phase.

**Construction:** Is really about cost-efficient development of a complete product that can be deployed in the user community.

**Transition:** Transition the product to its users.

### 3.2 DWEP 2.1 Core Workflows

The DWEP 2.1 proposes the use of twenty one (21) artifacts for the development of data warehouse. In general terms the UP, workflow is a set of activities in a given area resulting in the construction of artifacts (a text, a diagram, a web page, code in programming language, etc.).

In table 1 show twenty one (21) artifact are used in analysis, design, and implementation workflow. The descriptions of the workflows in the development process are:

**Requirement:** During this workflow, end users specify the measures and add more interesting, dimensional analysis, queries

**Analysis:** The purpose of this workflow is to improve the structure and requirements from the requirements stage. This step documents the incumbent systems that feed the data warehouse. DWEP proposed use the Source Conceptual Schema (SCS), Source Conceptual Object Schema (SCOS), Source Logical Schema (SLS) and Source Physical Schema (SPS).

**Design:** At the end of this workflow, the structure is defined in the data warehouse. The main result of this workflow is the conceptual model of the data warehouse. The DWEP proposes the use Data Warehouse Conceptual Schema (DWCS), Data Warehouse Conceptual Object Schema (DWCOs), Client Conceptual Schema (CCS) and Data Mapping (DM), Data Warehouse Sequence Schema (DWS), Data Warehouse state machine Schema (DWSMS) and Data Warehouse activity Schema (DWAS).

**Implementation:** During this workflow, the data warehouse is built: The physical structure of the data warehouse is built, starts to receive data in computer systems operations, is tuned for optimized performance, among other tasks. The process proposed as unified engine components diagram. The DWEP propose use: Data Warehouse Physical Schema (DWP), Data Warehouse Logical Schema (DWL), Client Logical Schema (CLS), Client Physical Schema (CPS), and ETL Process.

**Tests:** The aim of this work is to verify the application to work correctly. More specifically, the effects of the tests are: Planning the evidence needed to design and implement the tests by creating test cases and perform tests and analyze results of each test.
Workflows for maintenance and Post-development are not in the unified process and only part of the engineering process of the data warehouse.

**Maintenance:** Unlike most systems, the data warehouse is a process that feeds constantly. The purpose of this workflow is to define the loading and updating processes necessary to maintain the data warehouse. This workflow starts when building the data warehouse and is delivered to end users, but does not have an end date. During this study, end users may have new needs, such as new downloads, which triggers the beginning of a new iteration with the requirements of workflow.

**Revisions post development:** This is not a workflow of development activities, but a review process to improve future projects. If we keep track of time and effort invested in each stage is useful in estimating time and needs to generate the requirements for future developments.

---

**Figure 2. Data Warehouse Engineering Process 2.1 activity diagrams**
4. DWEP DEVELOPMENT TOOL
To develop the methodology DWEP 2.1 was used a software plug-ins for the Eclipse tool Galileo 2010 [9], the model was defined by Eclipse Modeling Framework (EMF) [10] and Graphical Modeling Framework (GMF)[11].

The Eclipse Galileo [9] is a tool to integrate different applications to build an integrated development environment (IDE). It’s a development platform open source software (open source), which consists of three parts: Eclipse Project, Eclipse Tools Project and Eclipse Technology Project.

The main component of the Eclipse Project is Java Development Tools (JDT). This allows you to create applications in Java. In addition, Eclipse provides mechanisms to integrate other applications as plug-ins. These plug-ins are automatically recognized by Eclipse when it starts. Because Eclipse is developed in Java, for its operation must be installed the Java Runtime Environment (JRE).

The Eclipse Modeling Framework (EMF) [10] is a tool that provides a modeling structure and facilities for code generation in order to build tools or other applications based on a structured data model. From an XML specification of a model, EMF provides tools and runtime support to produce a set of Java classes based on this model, a set of Adapter classes that enable viewing and editing commands based on the model, and a Basic editor.

Graphical Modeling Framework (GMF) [11] provides a tool for generating graphical editors based on EMF and GEF. The latter is what allows the developer to create a graphical editor to quickly track from a model of an application. A notable feature in GMF is the reuse of graphic definition for different domains and applications, that is, you can reuse already defined graphical metaphors for concepts in different domains and applications. This feature is achieved by modeling separately the graphical components that correspond to each of the elements of the domain and the definition of the tool palette, which will have a tool for each primitive. To complete the process of generating a graphics editor domain, GMF provides a definition of mapping or correspondence through each primitive associated with the component modeling and its graph in the editor tool that is being generated.

5. A CASE STUDY: THE UNIVERSITY OF THE IBAGUE, CENTRAL LIBRARY
The DWEP can be describe in two dimensions: The horizontal axis represent time and shows the dynamic aspect of the process (Phases and iterations), and the vertical axis represent the static aspect of the process (Core Workflows)

5.1 Dynamic aspect: phase and Iterations
This is a dynamic organization of the process along time. The Data Warehouse life cycle is broken into cycle, each cycle working on a new generation of the data warehouse. The DWEP divides one development cycle in four consecutive phases [12]: inception Phase, elaboration phase, construction phase and transition phase.

Each phase is concluded with a well a point in time at which certain critical decisions must be made, and therefore key goals must have been achieved.

5.1.1 Inception Phase
During the inception phase goals to build the data warehouse of the Central Library were: Making technology decisions, business modeling, requirements Capture and Identification of critical risk. For the development phase started with the business model which shows the basic operation in the loan of library materials and databases that stores the location of the various elements of the library (magazines, newspapers, books, etc.. .) The results of this phase were: the vision document, the document describing the business, the risk assessment document, the model of functional requirements, the use case model, domain model, prototype disposable cellar data and candidate architecture for building the data warehouse.

This phase is executed 30% of the core workflow requirements and 20% of the development core workflow analysis.

5.1.2 Elaboration Phase
In this phase of development of data warehouse architecture executable is proposed, risk assessment, refining use cases, creating the detailed plan of the construction phase, cost-benefit analysis. Nourish the elements of this phase are the elements obtained in the initial phase and the results were refined vision document, document refined risk assessment, the model of non-functional requirements, Model architecturally significant use cases, analysis the source data (conceptual, logical and physical), and development the data model conceptual and logical the data warehouse. It’s is development in 14 artifacts proposed in core workflows.

5.1.3 Construction Phase
During this phase was carried out the deployment of the transport model (ETL) and the physical model of the data warehouse. In this stage to validate the various models of the data warehouse and gets ready to be executed by library at present we are at this stage pending the Transition Phase and Iterations phase. It’s is development in 5 artifacts proposed in core workflows.

5.2 Static Aspect: Core Workflows
A process describes who is doing what, how and when. In DWEP is represented using four primary modeling elements: Worker, activities, artifacts and workflows.

5.2.1 Requirements
For the Core Workflows the requirements, DWEP use the use case diagram. The use case diagram document the behavior of the data warehouse from the standpoint of the user. Representing the functions that can be executed.

Figure 3 shows the use case applied to loan of the book process in the central library.

5.2.2 Analysis
In this Workflow you want to know the operational data source. The source systems maintain little historical data, and if you have a good data warehouse, the source systems can be relieved of
much of the responsibility for representing the past. The DWEP used the SCS, SCOS, SLS and SPS.

Figure 3. Use case

Figure 4 contains the Source Conceptual Schema (SCS), this represent the Operational Source Systems into class diagrams. In Figure 5 shows the Source Conceptual Object Schema (SCOS), it’s an example of the Source Conceptual Schema applied to the loan of the book "The Data Warehouse Toolkit".

Figure 4. SCS

In Figure 6 shows the transformation of the relational model the logical schema of data sources (SLS), where tables are represented by classes and attributes are mapped to the class diagram.

Figure 5. SCOS

For the SPS uses the deployment diagrams which specifies the type of computer equipment that uses the source that feeds the data warehouse, in our case in Figure 7 shows the servers that use the book loan systems.

Figure 6. SLS

Figure 7. SPS
5.2.3 Design
The objective is to review and revise the requirements to get to understand and develop properly. As the output of the requirements workflow is expressed so that the client understands the workflow analysis expresses these requirements in more technical language, adding details that are important to the client. The DWEP proposes the use: Data Warehouse Conceptual Schema (DWCS), Client Conceptual Schema (CCS) and Data Mapping (DM), Data Warehouse Sequence Schema (DWSS), Data Warehouse state machine Schema (DWSMS) and Data Warehouse activity Schema (DWAS).

In figure 8 shows with Data Warehouse Conceptual Schema (DWCS). This presents three levels of analysis of the data warehouse. At level 1 shows a basic diagram showing all the stars in the more expressive level 2 is chosen the star and this shows the dimensions and fact table that is constituted. At level 3 shows the attributes of the fact table and dimensions.

In data mapping (DM) the various transformations are performed between the SCS and DWCS, we begin by making a transformation table which consists of the following fields: the source attribute, the type of mapping and the target table. In figure 9 show this diagrams.

Data Warehouse Sequence Schema (DWSS) present the interaction between actors and their various steps between the source of the data ETL and data warehouse. Figure 10 shows this interaction in the loan of books.

Data Warehouse State Machine Schema (DWSMS) is used to see the dynamic between the data source and data warehouse. You answer analyzes the events, showing how are you reacting withThis current process. The figure 11 shows the ETL process from source data to the data warehouse.

Data Warehouse Activity Schema (DWAS)It is equivalent to the flow chart for the development of the structure of the data warehouse. In figure 12 is observed as are the various activities presented in the ETL process and helps develop the ETL Schema.
5.2.4 Implementation

During this workflow, in the implementation we start with the result of design and implement the data warehouse in terms of the fact tables and dimensions. This results in source files, scripts, binary files, a task in the ETL process. The DWEP propose use: Data Warehouse Logical Schema (DWLS), Data Warehouse Logical Object Schema (DWLOS), Data Warehouse Physical Schema (DWPS), ETL Process, transportation diagrams, Client Logical Schema (CLS), and Client Physical Schema (CPS).

DWLS presents the relational schema of the data warehouse implemented in class diagrams. Figure 13 the fact table and dimension is represented in classes and their relations should be to transform the relations of class diagram.

Data Warehouse Physical Schema (DWPS) uses the deployment diagrams which specifies the type of computer equipment that uses in the data warehouse, in our case in Figure 15 shows the servers that use in the Data warehouse.

The ETL process is responsible for extracting data from heterogeneous data sources in this transformation process for loading the data warehouse. This process consists of six tasks: Selecting the data source, transforming the sources, Union of selection sources of objective data for loading, mapping attributes to the attributes source and target data loading. The export process is composed of the same elements of the ETL process, but only displayed as client logical schema. Figure 16 shows the ETL process in its different task.

The transportation Diagrams This represents the union between the SPS and DWPS diagram, the bridge link can be JDBC, ODBC, OLEBC, or OIC. On the client represents the output either through one application site or web output. The figure 17 shows the deployment diagram.
5.2.5 Tests
In this workflow is described each test, procedures, and metrics for evaluation of weaknesses. The goal of this is find and detect defects in the data warehouse developed or developing, and allocated the required quality elements. It validates the requirements and design.

5.2.6 Maintenance
Unlike most systems, the data warehouse is a process that feeds constantly. The purpose of this workflow is to define the loading and updating processes necessary to maintain the data warehouse. This workflow starts when building the data warehouse and is delivered to end users, but does not have an end date. During this study, end users may have new needs, such as new downloads, which triggers the beginning of a new iteration with the requirements of workflow.

5.2.7 Revisions post development
This is not a workflow of development activities, but a review process to improve future projects. If we keep track of time and effort invested in each stage is useful in estimating time and needs to generate the requirements for future developments.

6. CONCLUSIONS AND FUTURE WORK
The use of standards in the development of a data warehouse facilitates the processes of analysis and design, as well as the utilization of UML like common language of development. DWEP is a valid methodology for the development the data warehouses. With the proposed extensions, there was achieved a better comprehension of the world of the problem of UNIVERSITY OF THE IBAGUE, University, which was formed in the appliances of the process of development.

The Metalanguage MOF allows defining graphical languages for the different processes of business. A clear example is the extension of the UML, this allows designing profiles for every extension without need to modify the standard.

It has a control from the start and end of the development of data warehouse.

Future Work:
- Fully develop the module data mining and OLAP to complete the process of business intelligence.
- Validate the results and validate the methodology CRISP-DM

7. ACKNOWLEDGMENTS
This work was supported widely by the University Antonio Nariño, especially the library, also to the ongoing tutoring given by the National University where he is proposing the development of DWEP methodology for building data warehouses.

8. REFERENCES