Article

Journal of Technical Invention
June 2021 Vol.5 No.15 8-15

Design of an academic data repository applying data warehouse technology

Diseño de un repositorio de datos académicos aplicando la tecnología data warehouse

HERNÁNDEZ-CRUZ, Luz María†*, CASTILLO-TÉLLEZ, Margarita, DÍAZ-ROSADO Martina and CHAN-CHI, Johan Oliver

Universidad Autónoma de Campeche, Facultad de Ingeniería. Instituto Tecnológico Superior de Champotón.

ID 1st Author: *Luz María, Hernández-Cruz* / **ORC ID**: 0000-0002-0469-5298, **Researcher ID Thomson**: H-3153-2018, CVU **CONACYT ID**: 662220

ID 1st Co-author: Margarita, Castillo-Téllez / **ORC ID**: 0000-0001-9639-1736, **Researcher ID Thomson**: S-2283-2018, **CVU CONACYT ID**: 210428

ID 2nd Co-author: Martina, Díaz-Rosado / **ORC ID**: 0000-0002-1142-586X, **Researcher ID Thomson**: ABG-7532-2021, **CVU CONACYT ID**: 739316

ID 3rd Co-author: Johan Oliver, Chan-Chi / **ORC ID**: 0000-0003-2071-4783, **Researcher ID Thomson**: ABG-6844-2021, **CVU CONACYT ID**: 1173924

DOI: 10.35429/JOTI.2021.15.5.8.15

Received January 12, 2021; Accepted June 24, 2021

Abstract

This article presents the proposal for the design and construction of an academic data repository under data warehouse technology. The initiative of the study arises from the identification of large volumes of data and the importance of combining them to analyze relevant indicators in decision-making by managers in higher-level educational institutions. The applied research proposes the use and implementation of the Hefesto v2.0 methodology capable of guiding step by step in the development of the data warehouse ensuring its start-up. As an added value, emerging technologies are integrated, particularly Microsoft SQL Server 2019 and Visual Studio 2019. The exposed results will serve as an applied study case for information systems administrators, administrative personnel and directors of educational institutions who wish to implement strategies with the aim of optimizing and make more efficient the analysis of data oriented to the benefit of educational processes that allow discerning relevant findings for decision-making.

Data warehouse, Hefesto, Educational Institutions

Resumen

El presente artículo exhibe la propuesta del diseño y construcción de un repositorio de datos académicos bajo la tecnología data warehouse. La iniciativa del estudio surge de la identificación de los grandes volúmenes de datos y la importancia de conjuntar los mismos para analizar indicadores relevantes en la toma de decisiones directivas en instituciones educativas de nivel superior. La investigación aplicada plantea el uso e implementación de la metodología Hefesto v2.0 capaz de guiar paso a paso en el desarrollo del data warehouse asegurando su puesta en marcha. Como valor agregado se integran tecnologías emergentes, particularmente Microsoft SQL Server 2019 y Visual Studio 2019. Los resultados expuestos servirán como caso de estudio aplicado para administradores de sistemas de información, personal administrativo y directivo de instituciones educativas que deseen implementar estrategias con el objetivo de optimizar y hacer más eficiente el análisis de datos orientado al beneficio de los procesos educativos que permitan discernir hallazgos relevantes para la toma de decisiones.

Data warehouse, Hefesto, Instituciones Educativas

Citation: HERNÁNDEZ-CRUZ, Luz María, CASTILLO-TÉLLEZ, Margarita, DÍAZ-ROSADO Martina and CHAN-CHI, Johan Oliver. Design of an academic data repository applying data warehouse technology. Journal of Technical Invention. 2021. 5-15: 8-15

† Researcher contributing as first author.

Introduction

In higher-level educational institutions a large number of academic and administrative processes are carried out, this as time passes and the institution grows, it increases considerably. Invariably, as a consequence it becomes more complex to be able to analyze information to support decision-making. A data warehouse is an electronic warehouse where a large amount of information is kept in a secure, reliable and easy-to-manage way. This research proposes the design of a data warehouse that allows centralizing academic data in higher-level educational institutions with the use of data warehouse technology. The case study proposal allows to carry out and analyze the design of the data warehouse applying the Hefesto v2.0 methodology step by step.

Methodology

The Hefesto methodology is a methodology created by Ing. Bernabeu Ricardo Darío, it is based on a very broad investigation, comparison of existing methodologies, own experiences in data warehousing processes. Hefesto is a methodology for building a data warehouse.

It should be noted that Hefesto is in continuous evolution, and all the feedback provided by those who have used this methodology in various countries and for various purposes have been taken into account as a great added value.

The Hefesto methodology has the following characteristics (Bernabeu R. Dario & García Mattío Mariano, 2007):

- The objectives and expected results in each phase are easily distinguished and are simple to understand.
- It is based on user requirements, so its structure is capable of adapting easily and quickly to changes in the business.
- It reduces resistance to change by engaging end users at every stage to make decisions regarding the behavior and functions of the data warehouse (DW).
- It uses conceptual and logical models, which are easy to interpret and analyze.
- It is independent of the type of life cycle used to contain the methodology.
- It is independent of the tools used for its implementation.
- It is independent of the physical structures that contain the DW and their respective distribution.
- When a phase is completed, the results obtained become the starting point to carry out the next step.
- Applies to both Data Warehouse and Data Mart.

The Hefesto Methodology v2.0 supported in 2010, considers four main phases that are summarized in Figure 1.

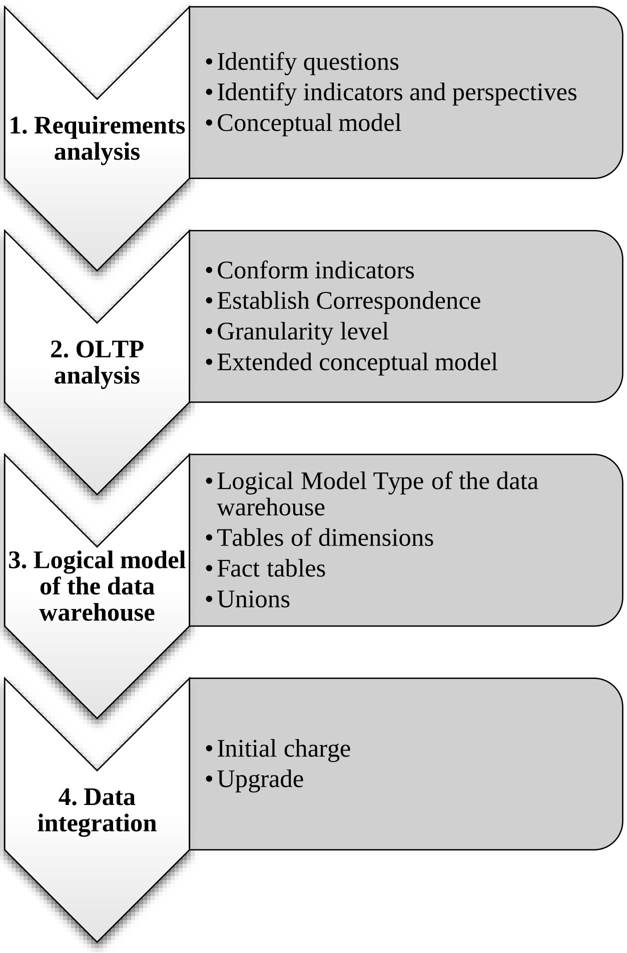

Figure 1 HEFESTO methodology
Source: (Bernabeu, 2010)

The first phase of the “Requirements Analysis” methodology has the purpose of using data collection techniques and tools to specify a set of questions that support the indicators and perspectives for the construction of the data warehouse (AD). This stage ends with the design of its own conceptual model for the specific case of the company or organization.

In the second phase, "OLTP Analysis (Online Transaction Processing)", the warehouse data sources are recognized and a correspondence analysis is carried out with the conceptual model obtained from the previous phase. Likewise, it is necessary to specify which fields will be included in each perspective and, finally, to obtain with it, the expanded conceptual model.

Specifically, in the phase "Logical model of the data warehouse" the logical model of the data warehouse is implemented (including the fact table, the dimension tables and the unions between them).

The last phase "Data integration", takes place after having built the logic model, in this step we proceed to load the data using ETL (Extract, Transform and Load) processes and to automate their update.

Developing

The study carries out applied research within the Faculty of Engineering of the Autonomous University of Campeche.

The Autonomous University of Campeche (UAC) is a public, autonomous university, with state, national and international links, which contributes with relevance and competitive quality to the formation of human capital and the generation, application and innovation of knowledge to meet the requirements and opportunities for the sustainable development of the state of Campeche, through a permanently updated institutional administration and with concurrent, timely and sufficient financing from the public, private and social sectors, national and international: Training high school graduates, associate professionals, professionals with bachelor's degrees and professionals with postgraduate degrees, through accredited educational programs, under an educational model focused on learning, in continuous, multimodal and flexible innovation (UAC, 2021).

In the first instance, the research question is posed "Does the design of a data warehouse using data warehouse technology allow to centralize volumes of academic data from various external sources in a higher-level educational institution?".

The empirical-experimental analysis process of the study proposes to apply the Hefesto methodology for the implementation of an academic data warehouse integrating the use of emerging technologies tools such as Microsoft SQL Server 2019 and Microsoft Visual Studio 2019. The main characteristics of SQL Server 2019 are (Microsoft, 2021):

- Intelligence in all your data with big data clusters
- Choice of platform and language
- Best performance in the industry
- More protected data platform
- Unmatched high availability
- Comprehensive mobile business intelligence
- SQL Server on Azure

For its part, the main novelties of Visual Studio 2019 are (SL, 2021):

- Visual Studio Live Share is a new feature that gives you the ability to share a certain project with another team member, giving you the ability to edit and debug code in real time.
- Creation of new projects with improved search and a list of templates.
- New start window with which developers can write code more quickly and which allows working with Git repositories.
- Easier project setup using .NET Core.
- Use of the preliminary version of the C # 8 language, which includes new features, new data types and expressions.
- It will no longer be possible to compile UWP applications for Windows 10 Mobile devices.
- Better Python integration and support for .Net Core 3.0 projects such as WinForms and WPF.
- Git work item experience enhancements using Azure DevOps.

Next, the case study referred to hereinafter as "DW-Educ @ project" for each of the phases of the Hefesto methodology as a result of the research is exhibited in detail.

Results

Hefesto Phase 1: Requirements of the DE-Educ @ project

From applying an interview to the directive and administrative staff of the Faculty of Engineering, about the problems identified in academic-administrative processes, the following points can be determined:

- The need for a structured and updated academic database or data warehouse that allows offering relevant information and is available to the administrative authorities for decision-making.
- The processes for the generation of reports of student educational indicators (classification, debugging, validation and generation of reports) are generated manually and in isolation.
- The use of non-specific tools for information storage and report generation. Also, the excessive use of files and spreadsheets is identified.
- Technological tools are not used to support the generation of reports automatically and efficiently.

The current situation leads to prioritizing two key requirements for the study approach. Table 1 shows the requirements established.

Code	Request
Rq_01	Generate a data warehouse that centralizes in a structured way the academic data from various external sources (data sources) that are processed in the institution.
Rq_02	Display results referring to academic indicators based on the data collected in a more intuitive and comprehensive way with the use of technologies.

Table 1 Requirements of the DE-Educ @ Project
Source: Own Source

If we focus on the first requirement, it raises the question of the study. Can an academic data warehouse be created that allows data from different sources or origins to be centralized for analysis? In other words, "The data warehouse technology is applicable to the processes of educational institutions." For this reason, the second requirement is taken to establish the business questions of the DW-Educ @ project.

Solely and exclusively for the study the scope of the analysis is delimited.

In this way, the following business questions are specified and their respective indicators and perspectives are detailed:

- Number of students enrolled by educational program with active status in the Faculty of Engineering of the Autonomous University of Campeche.
- Number of students by educational program in the Faculty of Engineering of the Autonomous University of Campeche related to computer science.

In short, its indicators are:

- Total active enrolled students.
- Total number of students enrolled in educational programs in computer science.

And the analysis perspectives are:

- Students
- Educational programs
- School cycle

Figure 2 shows the conceptual model resulting from the data for the DW-Educ @ project..

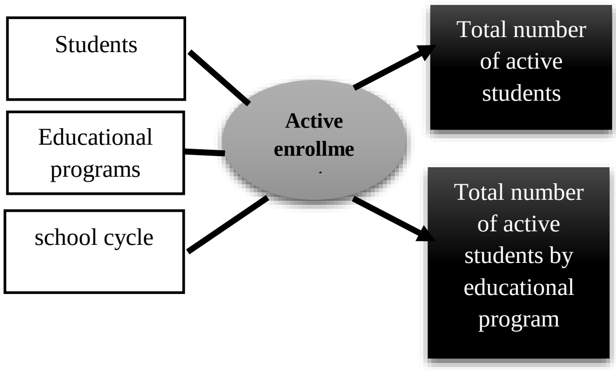
Students
Educational programs
school cycle
Active enrollme
.
Total number of active students
Total number of active students by educational program

Figure 2 Conceptual model of the DE-Educ @ Project
Source: Own Source

Hefesto Phase 2: Analysis of the OLTPs of the DW-Educ @ project

Next, the OLTP sources are investigated to determine how the indicators will be calculated and to establish the respective correspondences between the conceptual model created in the previous phase and the external data sources.

Finally, the information of the fields and of each perspective is added in order to achieve the expanded conceptual model.

The external sources available for the DW-Educ @ project are:

1. **source_01.** A plain text file (txt) that contains, among other data, the list of all educational programs offered by the Faculty of Engineering of the Autonomous University of Campeche.
2. **source_02.** A spreadsheet created in Microsoft Excel 2019, exported from an information system where the general data of the students enrolled in the Faculty of Engineering provided by the Information Technology area of the University is stored.

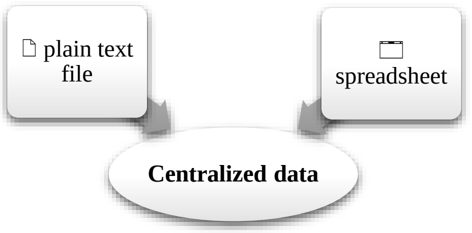

Figure 3 External data sources (data sources) for the DW-Educ @ Project
Source: Own Source

It is important to highlight that they are not the only data sources handled by the Autonomous University of Campeche. As an educational institution there are different computer systems that allow the automation and administration of the different academic and administrative processes that are carried out. However, these sources contain the data related to answering the business questions and indicators defined in the study.

The indicators are determined according to:

- Fact (s) that compose it, with their respective calculation formula.
- Summarization function used for aggregation.

Table 2 shows the composition of indicators for the DW-Educ @ project..

Indicator	Composition
Indicator_01 Total number of active students	– Facts: total_enrolled – Summarization function: sum.
Indicator_02 Total number of active students by educational program	– Facts: total_enrolled by pe – Summarization function: count.

Table 2 Composition of indicators of the DE-Educ @ project
Source: Own Source

Subsequently, the relationships and correspondence analysis with external data sources are distinguished, specifying:

- The matricula field of the source_02 with the student perspective from the Matriculated worksheet.
- The PE field of source_01 with the educational program perspective.
- The school_cycle field from source_02 with the school year perspective from the Enrolled worksheet.
- The field total_alumnos and total_alumnos_pe correspond to indicator_01 total number of active students and indicator_02 total number of active students by educational program respectively.

After verifying the critical data for the analysis and the specifications made, the exposed conceptual model is redesigned. Figure 4 shows the expanded conceptual model.

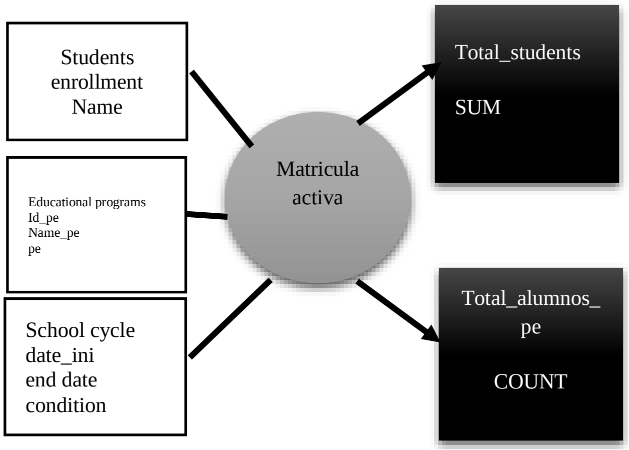

Figure 4 Extended conceptual model of the DE-Educ @ Project
Source: Own Source

Hefesto Phase 3: Logical Model of the data warehouse for the DW-Educ @ project

The next step for the Hefesto methodology is to implement the logic model of the project.

The design of the data warehouse structure is used is the star. A star schema is a type of relational database schema that consists of a single central Fact Table surrounded by dimension tables. You can have any number of dimension tables. The branches at the end of the links connecting the tables indicate a many-to-one relationship between the Fact Table and each dimension Table.

Among the relevant characteristics of the star scheme we have:

- It allows to share dimensions between different tables in fact.
- Update a dimension and it is reflected in all models and reports.
- Eliminates the complexity of interpreting the schematic structure.
- There are fewer Tables, but bigger with more information. All relevant information is concentrated in the dimension.
- Data extraction is faster.
- It is easier to maintain and make the necessary changes.

When designing the star scheme of the DE-Educ @ project, the dimension tables and the fact table are defined, presented in Table 3 and 4 respectively.

Table dimension	Description
DIM_DatosPE	Stores the general data of the Educational Programs.
DIM_DatosAlumnos	Stores general student data.
DIM_CicloEscolar	Stores the data related to the school year.

Table 3 Table Dimensions of the DE-Educ @ project
Source: Own Source

Table Fact	Description
TH_Matricula Activa	Table of Fact referring to students enrolled in educational programs of the Faculty of Engineering of the Autonomous University of Campeche.

Table 4 Table Fact of the DE-Educ @ project
Source: Own Source

Behind the Dimension Tables and the Done Table, the corresponding unions are established to obtain the Logical Model of the DW-Educ @ project presented in Figure 5.

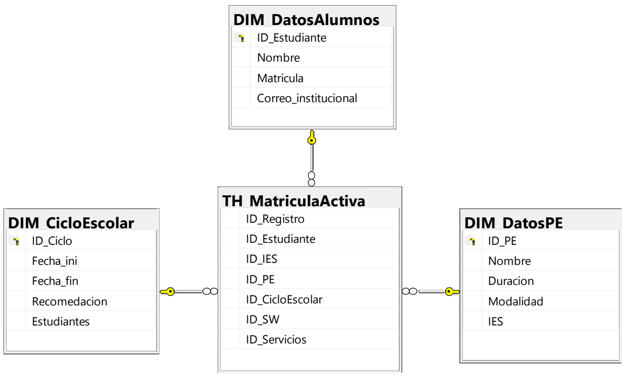

Figure 5 Logic model of the DW-Educ@ Project
Source: Own Source

Hefesto Phase 4: Integration of data from the data warehouse for the DW-Educ @ project

The extract, transform and load (ETL) process is done in visual studio 2019 using the Integration Services plugin.

The two external sources of data identified are implemented within the project: a plain text file and a Microsoft Excel spreadsheet. Figure 6 and 7 show the mentioned components.

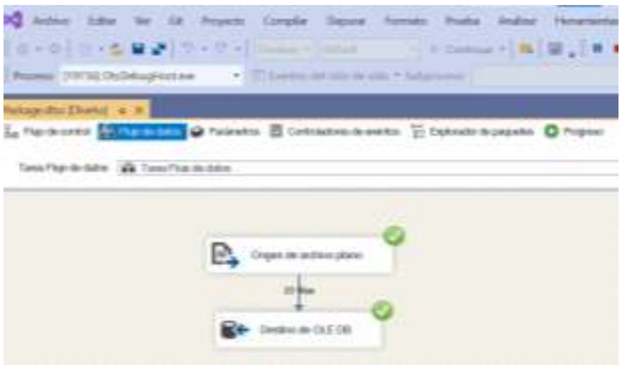

Figure 6 Plain text source of the DE-Educ @ Project
Source: Own Source

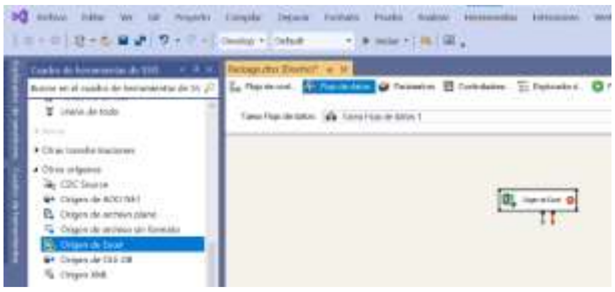

Figure 7 Source of Microsoft Excel Spreadsheet of the DW-Educ @ Project
Source: Own Source

Data loading is automatic through the Microsoft Visual Studio interface when you run the project. The data from both the text file and the spreadsheet are loaded into the model implemented in Microsoft SQL Server 2019.

It is essential to mention that there is total compatibility between the database manager and the programming environment. Similarly, the use of the latest editions of the applied technologies is highlighted.

At this moment, the data warehouse of the DW-Educa @ project contains all the data from the external data sources provided.

Conclusions

The use of the Hefesto methodology served as a guide for the design and implementation of a data warehouse through data warehouse technology. The implementation of the fact and dimensional model facilitated the stage of structuring the warehouse that meets the requirements of the DW-Educ @ project. The application of each phase of the methodology.

The technological tools used were crucial in achieving the data warehouse. The Microsoft SQL Server 2019 database manager was useful and efficient for the physical model, while Visual Studio 2019 allowed the ETL process to be carried out in an automated way. It served as a mediator between external sources and the model created for the data warehouse.

Definitely, the analysis and monitoring of the Hefesto methodology was undoubtedly successfully applicable to a specific and specific problem in the design of academic data warehouses, fully coupled to the technologies used.

Acknowledgments

We thank Editorial ECORFAN for the support provided for the publication of this article. Similarly, recognition is extended to Mtro. Francisco Javier Barrera Lao, Director of the Faculty of Engineering for empathy for promoting research studies in the area of Computer Science. Likewise, to each of the people who made the development and implementation of the research possible.

References

Bernabeu , R. D. (2010). *HEFESTO. DATA WAREHOUSING: Investigación y Sistematización de conceptos.* Córdoba.

Bernabeu R. Dario, & García Mattío Mariano. (2007). *HEFESTO DATA WAREHOUSING. Guía de aplicación teórico-práctica; Metodología Data Warehouse.*

Conesa Caralt, J., & Curto Díaz, J. (2013). *Introducción al Business Intelligence.* Editorial UOC.

Curto Diaz, J. (2016). *Introducción al Business Intelligence.* Editorial UOC.

Curto Díaz, J., & Conesa Caralt, J. (2016). *¿Cómo crear un Data warehouse?* Barcelona: Editorial UOC.

Golfarelli, M. (2009). *Data Warehouse Design: Modern Principles And Methodologies.* McGraw-Hill.

Hendrakusma Wardani, N., Nanang Yudi Setiawan, & Satrio Agung Wicaksono. (2019). *Data Warehouse.* Indonesia: UB Press.

Kimball, R., & Caserta, J. (2011). *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting.* Canadá: WILEY.

López Benitez, Y. (2019). *Business Intelligence. ADGG102PO.* IC Editorial.

Medina La Plata, E. (2017). *Business Intelligence: Una guía práctica.* Editorial UPC.

Microsoft. (diciembre de 2021). *Microsoft Data platform.* Obtenido de https://www.microsoft.com/es-mx/sql-server/sql-server-2019-features

Nordeen, A. (2020). *Learn Data Warehousing in 24 Hours.* Alex Nordeen.

Parteek , B. (2019). *Data Mining and Data Warehousing: Principles and Practical Techniques.* New York: Cambridge University Press.

Qamar Shahbaz. (2016). *Data Mapping for Data Warehouse Design.* United States of America: Elsevier.

Rainardi, V. (2008). *Building a Data Warehouse: With Examples in SQL Server.* United States of America: APRESS.

SL, T. p. (diciembre de 2021). *Microsofters*. Obtenido de https://microsofters.com/153842/visual-studio-2019/

UAC. (diciembre de 2021). *Universidad Autónoma de Campeche*. Obtenido de https://uacam.mx/paginas/ver/7