Journal of Contemporary Issues in Business and Government Vol. 27, No. 1, 2021
P-ISSN: 2204-1990; E-ISSN: 1323-6903
https://cibg.org.au/

Integration Of The Enterprise Information To Facilitate Decision Making

Manuel Hilario[1], Doris Esenarro[2], Hugo Vega[3], Ciro Rodriguez[4]

[1]Universidad César Vallejo, Lima, Perú.
[1,2]Universidad Nacional Federico Villarreal, Lima, Perú.
[3,4]Universidad Nacional Mayor de San Marcos, Lima, Perú.

Email: [1]fhilariof@ucvvirtual.edu.pe, [2]desenarro@unfv.edu.pe, [3]hvegah@unmsm.edu.pe, [4]crodriguezro@unmsm.edu.pe

Abstract

The research objective is to apply business intelligence techniques to take advantage of an existing system's conditions and increase the low-cost capabilities in its Extract, Transform and Load ETL processes with snapshots integrating the company's information web services. The methodology used for the development was HEFESTO, the development time was one week applying agile methodologies, achieving a short development with well-defined and easy-to-understand phases; the tool was the Pentaho BI Suite. Using open source tools, the cost was reduced through the proposed architecture with the repository type and the DataMart with a departmental approach to the problem, increasing its decision-making capacity.

Keywords: ETL, HEFESTO, BI, Pentaho, decision making, enterprise, services

1. INTRODUCTION

A regulatory entity is a state body in charge of the supervision, regulation, and management of the entities or operations they are assigned. In sanitation, this class of organisms maintains control over the companies that provide sanitation services (PSS), keeping water sources, managing water resources, and maintaining the integral water cycle. To carry out the control work, the regulatory company requests, every month, the information of the PSS to which it is in charge. This information helps them to obtain a broad overview of the management carried out by the PSS and to be able to exercise their regulation and control tasks.

The company decided to implement a system that integrates the information from this PSS through web services to a database on its servers to facilitate this work. This information would then be viewed in a report format by the company. Although the initial objective of centralizing data was met, this system had two shortcomings.

The first has to do with data integration. Many of these companies have supported computer systems; therefore, connecting to the web service is not a complicated task.

However, other PSS does not have information systems and therefore send their data using a csv sheet or try to generate .xml files to carry out the integration, although the latter occurs with great difficulty.

The second is related to the presentation of the information. The system currently allows to generate specific reports for each PSS; this functionality is sufficient if only one company is analyzed. However, nowadays, there are no functionalities to compare PSS, limiting the analysis that can be performed. Although the current system supplies several of the organization's needs, it is possible to provide those lacking using business intelligence technologies to support decision-making.

2. BACKGROUND

The problems required to use technology to improve the quality of information and decision making are not new. Some numerous academic researches and theses have successfully applied a business intelligence solution.

Carhuallanqui reported, in his thesis, in the case of the Dispefarma company, whose sales area did not have a solution that provides reliable and fast information for decision-making [1]. To solve this problem, the author used a business intelligence solution which achieved the following:

- Improve decision making.
- Reduce reporting time by 90.13% [1].

Yalan and Palomino give another similar case in their article. Analogously to the previous case, the T-impulse company's logistics area was in serious problems; it was in serious problems to generate reports, obtain data, and lacked capabilities for the design of these reports [2]. To solve the problem, a DataMart was used as a repository, obtaining the following conclusions:

- Reporting time was reduced.
- The creation of new reports with existing data was facilitated [2]

There are several cases analogous to the two explained above. Florian improved decision-making in the USMP enrollment process [3]. Anselmo and Espinoza did the same to support decision-making in a telephone area [4].

All authors succeeded in expanding the decision-making capabilities of their organizations using some business intelligence solution. For this paper, this decision-making improvement is being sought by expanding the existing system's capabilities and data quality.

This paper aims to propose a low-cost solution that allows expanding the organization's decision-making capabilities. To do this, the most appropriate type of repository will be sought, the ETL technique that best suits the case, the right BI tools that do not have components in the cloud, and a methodology that allows for incremental development that is simple to follow.

3. STATE OF THE ART

To carry out this work, three fundamental components must be taken into account: the repository type, the ETL process, and the upload and display tools.

A. Type of Repository

When it comes to business intelligence (BI) solutions, it covers everything from the data extraction area to the same presentation. That is why the first step to define in a BI system is the type of repository.

Currently, there are two fundamental types of repositories, the Data Mart and the Data Warehouse.

A Data Warehouse is a storage architecture designed to contain data extracted from transactional systems, operational data warehouses, and external sources. The warehouse combines that data into a summary form suitable for enterprise-level data analysis and reporting for business needs [5].

On the contrary, a Data Mart is a simplified form of a Data Warehouse that focuses on a single topic (functional area) such as sales, finance, and marketing. Generally, a Data Mart is built by a single department within an organization \ dots [6].

As shown in the definition, the most notable difference between the repository types is the scope they have in the organization, being a Data Mart of less scope than a Data Warehouse. Even so, other differences are fundamental when choosing between these technologies.

One of these differences that are very important to evaluate when applying a Data Warehouse or a Data Mart is its cost. Adrian mentions that a Data Warehouse can cost between 500,000 to 1'000,000 dollars per terabyte of stored information, while a Data Mart, having a smaller scope, can cost much less [7].

Said author highlights the differences in the following table:

DataMart DataWarehouse

Contains internal data (company departments) and external (internet, others). Includes information only relevant to one department. It does not necessarily use the dimensional model. The preferred dimensional model is the star model. Contains historical data Contains information that can be transmitted to a DW if necessary

High implementation cost Low implementation cost Stores detailed data and have Metadata.

Table 1 - Adrian. Data Mart / Data Warehouse comparison [7]

DataMart	**DataWarehouse**
It contains internal data (company departments) and external (internet, others).	Contains information only relevant to one department.
It does not necessarily use the dimensional model	The preferred dimensional model is the star model.
Contains historical data	Contains information that can be transmitted to a DW if necessary
High implementation cost	Low implementation cost
Stores detailed data and have Metadata	-

Based on the previous table, Adrian indicates that the solution that best suits the business context should be chosen; this is determined by factors such as needs, financial resources, and the information's complexity [7].

Adrian emphasizes that if the organization does not have complete data and does not have many financial resources, a Data Mart should be chosen; since a Data Mart can serve as the basis for generating a Data Warehouse later [7], the attack in a particular area and that the organization does not have many resources dedicated to IT innovation, it is prudent to use a repository DataMart.

B. The ETL process

The extraction, transformation, and loading (ETL) process is one of the processes that occur in any business intelligence system. It is probably the most critical process in these systems since, according to Kimball and Caserta, 70% of the resources necessary for implementing and maintaining a Data Warehouse are typically consumed by these processes [8].

Vassiliadis mentions that the tasks performed by the ETL can be summarized as follows:

- Extraction of the appropriate data.
- Transportation of these data to a preparation area where they will be processed.
- Transforming source data and calculating new values.
- Isolate and clean problematic records.
- Upload the clean and transformed data to the repository [9].

To define the best ETL technique, this work took as a reference the investigation of Díaz de la Paz, which aimed to characterize the data change capture techniques, compare them and select the most appropriate when implementing ETL processes [10]. The first technique that was analyzed was the one based on Timestamps. This technique requires the source database to have a timestamp type field that allows the ETL to detect if the records have been modified [10].

Another technique is based on snapshots, according to Kimball and Caserta. The snapshot technique saves an exact copy of each previous extraction in the DSA for future use, and during the next run, the process takes the entire source table to the DSA, where it is compared with the data loaded during the last process [10]. The third technique that he analyzed was based on triggers. In this technique, these triggers are generated in the source Database, and each time a change is made, the information is uploaded to the Data Warehouse [10].

The last technique that was analyzed was the one based on log files. In this, log files are used to save the database changes; these files are then read, and based on them, the Data Warehouse is updated [10]. Diaz de la Paz summarizes these comparisons in the following table:

Table 2 - Díaz et. al, Comparison of etl [10]

Aspects	**Timestamp**	**Snapshot**	**Triggers**	**Log**
Insert / update distinction	No	Yes	Yes	Yes
Detection of multiple updates	No	No	Yes	Yes
Identification of deletions	No	Yes	Yes	Yes
Little intrusive	No	No	No	Yes
Supports real time	No	No	Yes	Yes

Requires DBA	No	No	Yes	Yes
Independent from DBMS	Yes	Yes	No	No

The author mentions that the most appropriate technique to carry out an ETL process is based on Snapshots since it is not intrusive and does not depend on a database engine as a source to be applied.

C. Loading and display tools

The technique and type of repository used have been covered in this paper; however, the other point to discuss is the tools. There are several tools associated with business intelligence; among them are:

- QlikView.
- Palo BI Suite.
- Jaspersoft BI.
- Tableau Public.
- Spago BI.
- Pentaho BI Suite.

The work of Brandão et al. allows us to visualize a benchmarking of these tools. The authors took specific common characteristics of any BI system and classified the tools as follows:

Table 3 - Brandão et. al, Comparison of BI tools [11]

Features	BI Open-Source Tools						Group
	Jaspersoft BI	*Palo BI Suite*	*Pentaho BI Suite*	*QlikView*	*SpagoBI*	*Tableau Public*	
Performance	4	3	4	3	4	4	D
OLAP Ad hoc Queries	1	5	5	3	5	4	B
Architecture	4	4	5	4	5	4	D
Display of KPIs	1	1	5	4	4	4	A
Plug-ins	3	0	5	0	0	3	D
Interactive Visualization of Data	5	4	5	5	4	4	C
Documentation	4	4	2	2	2	3	F
Dashboards	1	1	4	4	5	4	B
Navigation Features	5	4	4	1	2	4	C
ETL	4	5	5	3	4	1	E
Connection to the Database	5	4	4	5	5	3	A
Integration of Dimensional Model	1	1	1	2	4	1	E
Open-source	5	5	5	1	5	5	D
Export	5	2	5	2	5	4	C
Pervasive	5	5	5	1	5	4	A
Online Help	4	2	3	4	3	4	F
Support for Mobile Devices	4	1	0	5	5	3	C
Data Mining	1	1	3	2	4	1	B
Ease of Use	4	4	4	4	4	5	F
Attractiveness	4	3	4	5	5	4	C
Customization of the Interface	4	0	5	5	5	5	F
User Profile	5	4	5	1	4	0	D
Real-time	5	4	5	1	5	1	A

These characteristics were grouped in order of priority for the authors' problem as follows:

Table 4 - Brandão et al., Feature prioritization [11]

	Group	**Characteristic**	**%**
A	Must have	Pervasive Real Time Display of Key Performance Indicators Connection to a Data base	
B	Technologies	OLAP Ad hoc Queries Dashboards Data Mining	30%
C	End User	Interactive visualization of data Navigation Feature Export Support for Mobile devices Attractiveness	25%

D	Other important	Performance Arquitecture Plug-ins Open Source User Profile	25%
E	Data Processing	ETL Integration of Dimensional Model	15%
F	Administrator	Documentation Online Help Ease of use Customization of the interface	5%

Table 4 shows that the best score is Spago BI, while Pentaho BI follows a minimal difference. The authors indicate that this difference, being so small, is practically negligible; however, the fact that the Spago BI installation is more complicated compared to Pentaho and that the documentation is much scarcer caused the authors to end up opting for Pentaho as the tool to create the system that would be their case study [11]. Based on the authors' conclusions, the Pentaho BI suite is used.

4. METHODOLOGY

For the development of the solution, it is necessary to specify a suitable development methodology. The characteristics needed for the method are described below:

- Independent of technology.
- It must have well-defined phases that are easy to follow.
- It must be medium range.
- Short development and design times.
- Incremental implementation.

To choose the correct methodology, we will use the comparison made by González, who gives a weighting on three development methodologies for DataWarehouse.

Table 5 - González, Comparison of Methodologies [12]

Characteristics	**Data Warehouse design Bill Inmon**	**Data Warehouse design Bill Inmon**	**HEFESTO**	**DWEP methodology**
Development methodology by defined layers	Yes	Yes	Yes	Yes
Medium -low level of detail	No	Yes	Yes	No
Medium range with projection to expand	No	Yes	Yes	Yes

Incremental implementation of the Data Warehouse	Yes	Yes	Yes	No
Total score	2	4	4	2

The table made by Vilca will also be used, which also compares these methodologies.

Table 6 - Vilca, Comparison of Methodologies [13]

N°	**Analysis factor**	Ralph Kimbal	Bill Inmon	Ricardo Bernabeu
1	Acceptable on any technology	**1**	**1**	**1**
2	Flexibility	**2**	**3**	**3**
3	Communication with the customer	**3**	**3**	**3**
4	Time in analysis and design	**1**	**2**	**2**
5	Construction time	**1**	**2**	**3**
6	Easy understanding beginners	**0**	**0**	**1**
7	Most used in the world	**1**	**3**	**3**
8	Ease of tracking	**2**	**1**	**3**
	Total score	**11**	**15**	**19**

Using both tables 5 and 6, the methodology that best suits this work is HEFESTO. Therefore, the methodology is defined by the phases provided by HEFESTO.

1. Requirements Analysis
   a. Identification of Questions.
   b. Identification of Indicators and Perspectives.
   c. Creation of a Conceptual model.
2. Analysis of data sources
   a. Form indicators.
   b. Map the data sources with the defined requirements.
   c. Define the level of granularity.
   d. Expand the conceptual model.
3. Perform the Logical Model of the Data Mart
   a. Create the dimension tables based on the perspectives obtained previously.
   b. Create the fact tables by grouping the indicators obtained.
   c. Join the fact tables with the dimension tables.
4. Data integration
   a. Define a Snapshot-based ETL process that controls data quality.
   b. Perform initial upload to the central repository.

c. Define update for incremental loads.

5. Presentation: Use the corresponding BI tool to display the information to the user.

5. DEVELOPMENT

A. System Architecture

Before starting with the development according to the chosen methodology. It is prudent to define the system's architecture derived from the points established in state of the art.

The architecture is as follows:

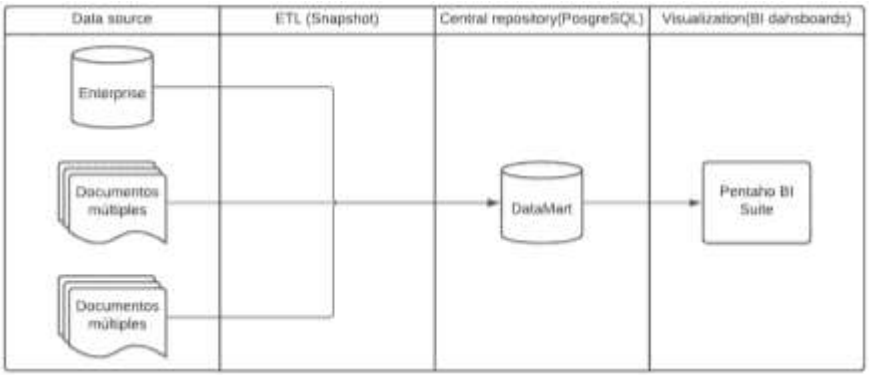

Figure 1 - Architecture of the proposed solution

Four phases can be seen, the first being the data sources of the operational system. In the second layer, the ETL process will be carried out with Pentaho Data Integration (PDI). In the third phase, the final repository will serve to feed the visualization tools that will be used in the last layer; these tools belong to the Pentaho BI suite [5,12-14].

B. Requirement Analysis

Applying the methodology selected in this thesis's theoretical contribution will begin by using the requirements analysis phase to obtain the business intelligence needs required by the proposed solution.

Identification of Questions

Questions were asked to the regulatory company's stakeholders to obtain the information needs demanded by their business operations. The following list lists those information needs.

1. Do you want to know the debit movements and have rated the localities of an HPE or accumulated by all the localities of the HPE in a given period?
2. It is desired to compare the annual debit and have movements of the expense accounts, represented in class 6, by location of an HPE or accumulated by all the HPE locations in a period?
3. Do you want to view the movements by a class of costs broken down by elements in a period for one location of an HPE or accumulated by all locations of the HPE?
4. Do you want to view the movements by type of costs broken down by accounting accounts in a period by location of an HPE or accumulated by all locations of the HPE?

5. Do you want to visualize the type's total movement must and have by region and group by period and year, sorted in descending order?

Identification of Indicators and Outlook

The questions defined in the previous point allow defining the corresponding indicators and perspectives.

Indicators:

- Debt type movement.
- Haber type movement.
- Total Movement.

Outlook:

- Costs.
- Accounts.
- Period.
- Element.
- Locality.

Creation of a Conceptual model

To finalize the first phase, the conceptual model must be plotted from the indicators and perspectives.

C. Analysis of data sources

Shape Indicators

In this step must be indicated how the indicators will be calculated and also their summary formula.

Table 7: Accounts, function and movement

Account	**Function**	**Movement**
Active	SUM	Income
Passive	SUM	Egress
Total	SUM	Income - Egress

Map the data sources with the requirements obtained

In this phase, the data sources must be selected, selecting those that will be used in the system, from:

- Account movement
- Planning cost
- Code cost element
- Account plan
- Providers data

Define the level of granularity

Next, the previous fields that will contain each perspective will be grouped, and their granularity will be defined. The order in which the areas are listed establishes the level of granularity from lowest to highest [15-16], some fields are considered extras that will support the ETL process as temporary keys for:

- Costs:
- Account:
- Period:
- Element:
- Location:

Tables of dimensions based on the perspectives obtained previously for Logic model of the Data Mart

- Period:
- Element:
- Location:
- Cost:
- Account:

D. Data integration

Define an ETL process based on Snapshots

As explained in the methodology, Pentaho BI Suite will be used for the ETL process and visualization. In the area of data integration, the tool that this suite provides us is Pentaho Data Integration [8, 17-19].

Therefore, six transformations have been carried out to carry out both the initial and incremental loads. The changes performed will be shown below for:

- Elements:
- Periods:
- Localities:
- Costs:
- Accounts
- Movements:

Perform initial load and define incremental loads

Once the transformations were defined, they were executed and carried out the initial loading of all the dimensions and movements of three Health Promoting Entity HPE. Incremental loads have already been described in the previous step because they are based on the snapshot technique.

E. Presentation

Use the corresponding BI tool to display the information to the user

A PENTAHO suite software called "Pentaho Report Designer" will be used to make the necessary reports and views. As a demonstration of this tool's capacity, some of the questions obtained in the first phases of the methodology will be answered. The names of the companies have been omitted due to the privacy of their data [4,8,19].

6. CONCLUSIONS

The work's objective was to obtain a low-cost business intelligence solution that increased the organization's decision-making capacity. It was possible to find the right type of repository for this solution; it was a DataMart due to the departmental approach. The ETL technique chosen was snapshots, and the tool was Pentaho BI Suite.

With the system architecture proposal and open source tools, it was possible to determine the minimum cost and decide on the appropriate architecture. Applying the HEFESTO methodology allowed us to obtain an agile development time with well-defined phases understandable for any developed. In operation, it was possible to respond to the requirements generated in the requirements process and the solution's development, evidenced in the results and comparisons between companies dedicated to similar items, which was not possible with the original system.

REFERENCES

[1] Ministerio de Educación - MINEDU. Gobierno del Perú. Available at: https://www.gob.pe/minedu

[2] M Rodríguez. Implementation process of technology in Education: The case of Blackboard 9.1 in the University of Manchester. Actualidades Investigativas en Educación. 2013; 13(3), 150-167. https://doi.org/10.15517/aie.v13i3.12043

[3] Y Ocaña-Fernández, L Valenzuela-Fernández and L. Garro-Aburto. Inteligencia artificial y sus implicaciones en la educación superior. Propósitos y Representaciones. 2019; 7(2), 536-568. https://doi.org/10.20511/pyr2019.v7n2.274

[4] Alfaro, O., Esenarro, D., Rodriguez, C., y Alfaro, M. R. (2021). The Unified Enterprise Architecture (AEU) as a strategic tool organizational modeling for the funtional competitiveness of universities. 3C Empresa. Investigación y pensamiento crítico. Edición Especial Tourism and University: Backbone of Peruvian Economy, 63-79. https://doi.org/10.17993/3cemp.2021.specialissue1.63-79

[5] D Rooein, Data-Driven Edu Chatbots. Companion Proceedings of The 2019 World Wide Web Conference. 2019; 46-49. https://doi.org/10.1145/3308560.3314191

[6] Singh, J., Joesph, M. H., & Jabbar, K. B. A. (2019). Rule-based chabot for student enquiries. Journal of Physics: Conference Series, 1228, 012060. https://doi.org/10.1088/1742-6596/1228/1/012060

[7] Molnar, G., & Szuts, Z. (2018). The Role of Chatbots in Formal Education. 2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY), 1. https://doi.org/10.1109/sisy.2018.8524609

[8] Müller, Stefan and C. Keller. “Pentaho Data Integration.” (2014).

[9] Lee, Y.-C., & Fu, W.-T. (2019). Supporting peer assessment in education with conversational agents. Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion, 7-8. https://doi.org/10.1145/3308557.3308695

[10] Cordova, J., Vega, H., Rodriguez, C., y Escobedo, F. (2020). Firma digital basada en criptografía asimétrica para generación de historial clínico. 3C Tecnología. Glosas de

innovación aplicadas a la pyme, 9(4), 65-85. https://doi.org/10.17993/3ctecno/2020.v9n4e36.65-85

[11]Pandey, B., Tomar, G.S., Bhadoria, R.S., Hussain, D., Rodriguez, C.,(2020). Environment-friendly FSM design on ultra-scale architecture: energy-efficient green computing approach World Journal of Engineering, DOI: https://doi.org/10.1108/WJE-08-2020-0397

[12] Cerna, F., Ubalde, R., Rodriguez, C., Sotomayor, J., Yucra, D., (2020). Automation psychological assessments with cloud computing. Journal of Critical Reviews, 2020, 7(15), pp. 1565-1569. doi: 10.31838/jcr.07.15.208

[13] Atencio, Y., Marin, J., Enriquez, R., Rodriguez, C., Petrlik, I.,(2020). A collaborative ide for graphics programming. Journal of Critical Reviews, 2020, 7(15), pp. 1570-1577. doi: 10.31838/jcr.07.15.209

[14] Azabache, I., Rodriguez, C., Gonzales, P.,(2019) M-Learning applied to the improvement of the learning of university engineering students. Proceedings of the 2019 International Symposium on Engineering Accreditation and Education, ICACIT 2019, 2019, 9130215. doi: 10.1109/ICACIT46824.2019.9130215

[15] Motta, V., Guillen, R., Rodriguez, C., (2019). Artificial Neural Networks to optimize learning and teaching in engineering careers. Proceedings of the 2019 International Symposium on Engineering Accreditation and Education, ICACIT 2019, 2019, 9130296. EID: 2-s2.0-85084220713

[16] Reyes, A., Rodriguez, C., Esenarro, D. (2019). Hyper converged systems applied (HSA) methodology to optimize the process of technological renewal in data centers. International Journal of Recent Technology and Engineering, 2019, 8(2 Special Issue 11), pp. 4052-4056. doi: 10.35940/ijrte.B1592.0982S1119

[17] Bustamante, J.C., Rodriguez, C., Esenarro, D. (2019). Real time facial expression recognition system based on deep learning. International Journal of Recent Technology and Engineering, 2019, 8(2 Special Issue 11), pp. 4047-4051. doi: 10.35940/ijrte.B1591.0982S1119

[18] Rodriguez C., Luque D., La Rosa C., Esenarro D. and Pandey B., "Deep Learning Applied to Capacity Control in Commercial Establishments in Times of COVID-19," 2020 12th International Conference on Computational Intelligence and Communication Networks (CICN), Bhimtal, India, 2020, pp. 423-428, doi: 10.1109/CICN49253.2020.9242584.

[19] Rodriguez C., Angeles, D., Chafloque R., Kaseng F. and Pandey B., "Deep Learning Audio Spectrograms Processing to the Early COVID-19 Detection," 2020 12th International Conference on Computational Intelligence and Communication Networks (CICN), Bhimtal, India, 2020, pp. 429-434, doi: 10.1109/CICN49253.2020.9242583.