DOI:

BUSINESS INTELLIGENCE USING THE PENTAHO PLATFORM: AN ANALYSIS OF CEFET-MG'S ACADEMIC BURDEN

Edson Marchetti Da Silva
CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS - ORCID: https://orcid.org/0000-0003-4801-0892
Erica Yuri Yoshiwara
CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS - ORCID: https://orcid.org/0000-0002-4557-5890

Abstract

Proposes the creation of a BI tool to analyze the academic burdens of the CEFET-MG educational institution.

To allow for a better decision on the distribution of new faculty vacancies among academic departments, the board of directors needed a management tool to present the charges of faculty in teaching, research and extension.

it to obtain the institution's database structure in order to conduct and validate a proof of concept, aiming to present the results to the Cefet Board and develop the site with a dashboard with the real data

To produce a tool powerful and extremely important that serve as a basis for obtaining information for decision making and behavior study of the institution.

The study of case shown that open source tools could be used to solve management problem in the public university.

To permit to the directors to do better decisions and optimize human resources

Key words: Data analysis, Decision-making, Business Intelligence, Data Warehouse, Pentaho

Business Intelligence Using the Pentaho Platform: An Analysis of CEFET-MG's Academic Burden

Abstract

Using information strategically has great benefits for corporations. One problem that strikes many managers is that too much data is available that is not transformed into useful information for the business. Business Intelligence (BI) is the name given to the process of collecting, organizing and analyzing data and information that help managers in monitoring and decision making. Motivated by this scenario, this paper proposes the creation of a BI tool to analyze the academic burdens of the CEFET-MG educational institution. To achieve this goal, a study of the Pentaho tool and methods for processing, storing and displaying the collected information was performed. The results show the feasibility of using the tool, as well as some alternatives for displaying information to the institution's managers through a more dynamic and interactive top-down approach. This approach allowed the institution to obtain easier and more organized information to subsidize the strategies for allocation of faculty vacancies in the institution.

Keywords: Data analysis; Decision-making; Business Intelligence; Data Warehouse; Pentaho.

1 Introduction

Although collecting and storing large amounts of information for data analysis is old, in recent years these actions have become increasingly used. More specifically, this paper uses Business Intelligence (BI) tools to deal with corporate data to produce useful management information for strategic decision making. In this context, Elbashir, Collier and Davern (2008) define BI systems as those that provide the ability to analyze business information to support and improve management decision making across a wide range of business activities. Silahtaroğlu and Alayoglu (2016) corroborate by stating that these are systems used by executives and / or senior managers, which aim to help the decision-making process easily and accurately according to the information obtained from the data analysis. available. Additionally, Isik, Jones and Sidorova (2013) define a BI system as a system composed of both technological and managerial elements, which presents its user with information for analysis, enabling decision making (both short and long term) and management support for the purpose of increasing the performance of an organization. It to Adapt to today's fast-changing business landscape requires agility, and BI tools play an important role in enhancing this adaptation and all the management benefits it can bring.

According to Ramakrishnan, Jones and Sidorova (2012) in recent decades, various BI tools have been developed in organizations. The effectiveness of these tools lies in their ability to present relevant information in a timely manner. Some of the most commonly stated and promoted benefits of BI implementation include data consistency and insight into business operations.

According to Silahtaroğlu and Alayoglu (2016), studies show that top corporate management relies on data displayed in dashboards, graphs, tables, numbers and statistics when making final decisions. In this sense, the BI implementation process is seen as a powerful problem-solving tool that can allow access to relevant integrated data and thus can result in better data quality and consistency to improve the information production process aimed at support the management decision-making process.

Thinking about BI systems with the focus on the context of this work stands out:

> The implementation of BI solutions applied in educational organizations can be of high relevance to identify, understand and predict potential problems residing in the institution, which are often the result of decisions made without knowledge hidden in this data. Thus, the consolidation and exploitation of this data support the decision making process and thus show new strategies in order to optimize the use of resources of the Institution (ALMEIDA and CARMARGO 2015, p. 2 - translated by the authors).

Therefore, this paper aims to conduct a case study at the Federal Center for Technological Education of Minas Gerais (CEFET-MG) aiming to model and implement a BI system that can be used to transform data from teachers activities into strategic information that allow you to produce queries displayed in dashboard[1] format. Additionally, it is expected that the result produced by the

[1] Indicator panels that display information in easy-to-view, graphical format.

elaborated software can be used by CEFET-MG's general management in order to better and more strategically distribute the human resources and the allocation of new vacancies for teachers in the institution.

The CEFET-MG already had a Faculty Academic Charges module integrated with the Academic System responsible for controlling student enrollment. However, in addition to the activities related to the classes taught, there are several other activities that have to be launched manually in order to have a global record of teachers' activities, such as guidelines, projects, publications, administrative burdens, etc. These activities are set out in the Work Plan (as provided for) and in the Annual Academic Charges Report (as performed) individually for each teacher. Although the Charge System records all activities, an interface was lacking that allows a more managerial view of this data. In this sense, this work has produced software that provides a dashboard for analytical visualization of teachers' charges separately or grouped by campus, department, type of activity, etc.

This paper is structured in seven sections. In the first one, the context, objectives and justification that led to elaborate the work were presented. In the second, the main theoretical concepts are addressed. In the third related works. The fourth section details the methodology used as the basis for the development of the work. The fifth details the development itself, describing the path taken. On Friday presents the analysis of the results found. Finally, in the last section, the conclusions are presented.

2 Theoretical Rationale

According to Almeida and Camargo (2015), the dimensional data model that should be used to organize the information to be shown to users. This model consists of three elements: facts, dimensions and measures. Kimball and Ross (2002) define the fact table, also known as fact table, as the main table of the model. Facts can be defined as collections composed of measures concerning the organization. It is the fact table that stores the data to be analyzed. Dimension tables, on the other hand, have descriptive characteristics, that is, the attributes present in the tables are crossed to assist in the generation of information that is present in the fact table. Dimensions are the aspects that you want to look at and indicate the level of detail at which they will be viewed. Moreover, according to Caralt and Diaz (2012) the dimensional structure can have two formats: Star Schema or Snowflake.

According to Cano (2007), the main components of BI systems are: Data Warehouse (DW), Data Mart (DM), Data Modeling, Extraction, Transformation and Load (ETL) process and OLAP tools.

DW is an environment that stores data from different sources and aids decision support. According to Singh (2001), DW is the process of integrating an organization's data into a single repository where queries, reports and analysis can be easily performed. Typically, data is not collected in real time and is stored over time for comparisons, trends and forecasts. DWs are characterized by: (1) subject-oriented (contain only the information necessary for processing decision support systems); (2) integrated (has data collected from different sources, storing it consistently); (3) variable over time (data always refers to some specific point in time, data is never overlapping, which allows information history mapping) and (4) non volatile (new data is incorporated, as data cannot be changed).

According to Gonçalves (2003), access to data from a DW improves the quality of customer service and helps the company evaluate possible activities emerging from its business. Access to data becomes easier and faster, and it is always current and accurate.

DM, when compared to DW, has minimal differences. The main variation lies in the fact that DM is focused on a particular area, while DW covers data from the entire organization. Silva (2004) defines DM as "a subset of DW data, allows decentralized access, and serves as the source for the data that will make up individual databases targeted at a specific department or business area."

According to Hokama et al. (2004) before data is stored in a DW, it is extracted from external sources, integrated and transformed. This process consists of three steps Extraction, Transformation, and Load (ETL).

The first step is to read and copy relevant pieces of information to the DW environment. Most of the time the data comes from different and independent sources. Hokama et al. (2004, p. 16) state that “the big challenge is determining which data to extract and what types of filters to apply”.

After extraction, you must transform the data to an appropriate format for loading into DW. Transforming data can mean deleting special characters, correcting duplicate values, handling lost data, deleting data irrelevant, etc. Finally, the last step of the process is the load on the DW.

Online Analytical Processing Tools (OLAP) are applications where end users have access to data analysis and decision making. Data visualization can be done from many different perspectives and offers a range of functionalities that aim to assist the end user in understanding the available information (ALMEIDA; CAMARGO, 2015).

According to Cano (2007), OLAP tools allow different analyzes of the same information due to the possibility of changing perspectives.

The tool chosen to perform job data analysis is Pentaho. In addition to being an open source tool, it offers the entire infrastructure to do the job in question, and is a flexible BI solution that provides data monitoring, analysis and presentation through dashboard.

3 Related Works

Munhós (2018) developed a system that analyzes data and returns information about the horticultural producers of the state of Minas Gerais who sell their products in the Free Market Producer. The author conducted a case study whose methodology is based on the BI cycle: the first step was to identify the need for the system for the study area. Then the planning was done, in which the steps for data collection and analysis were defined. In his work, it was possible to analyze relevant data, such as the main products per year, in which region a given product is most produced, quantity produced per year, etc. This data was extracted from the corporate database using the Pentaho tool. The extracted data was displayed through cubes, reports and graphs. This information was important and useful to validate if the amount of product sold by each producer corresponds to the production capacity of his property, thus avoiding the middlemen.

Focusing on the academic performance data of a higher education institution, Almeida and Camargo (2015) describe the strategy used to elaborate the decision support platform. First they were defined as the project components would be structured. After defining the structure of the architecture, the dimensional model of the data was created to allow quick queries about it. In the next step, the data provided by the educational institution were analyzed and entered in the dimensional model database. Finally, the last step was to make this data available to the user through a tool with the graphical interface.

Most of the papers found were articles about the importance of BI systems in different contexts. The works by Munhós (2018) and Almeida and Camargo (2015) were the most correlated to the objective of this work. Munhós (2018) uses Pentaho software, while Almeida and Camargo (2015) use SpagoBI in the academic context, focusing on the number of enrollments and graduates per year of each course.

4 Methodology

To achieve the objectives it was proposed a case study methodology of an exploratory research. According to Gil (2007), an exploratory research aims to create greater familiarity with the problem and usually involves three steps: bibliographic survey, interview with people who had experience with the problem and analysis of results. In addition, exploratory research can also be classified as a case study.

According to Fonseca (2002) a case study is a study done in one or more entities and aims to know the situation of a particular problem, ie, how and why a given problem occurs. The purpose of the case study is not to create a solution to the problem, but to discover and analyze certain aspects of the problem that are considered important and relevant. Based on exploratory research to understand the problem, identify and study the framework of tools to support the solution, a BI process was adopted. The proposed methodology was divided into seven steps:

- it to elaborate the theoretical framework and search for related works aiming to understand and define the concepts necessary for the elaboration of this work;
- it to obtain the institution's database structure without data content in order to conduct a proof of concept with dummy data and learn about how Pentaho works;
- it to present the prior results to the Cefet Board to approve the proposed implementation of the application for institutional use and obtain feedback on what data is considered relevant to the application;
- it to gain access to the institution's database to conduct the study based on actual data;
- it to perform tool testing and adjustments working in partnership with the internal IT project area;
- it to implement a final dashboard version;
- it to transfer knowledge to the IT project industry and record the operational process in a tutorial.

5 Development

This section presents the steps taken to develop the academic burden analysis tool of CEFET-MG.

5.1 Proof of Concept

It was necessary to perform a Proof of Concept to validate a functional prototype of the product designed to demonstrate its operation to the end user in order to gain access to the database in the corporate system.

5.2 CEFET-MG Database

Following approval by the institution's board of directors for the use of databases, the project office granted access to enable the work to be developed. A job data source has records of the institution's documents, such as: amount of academic and didactic charges, department or qualification, campus performing, activities performed, projects performed, etc. points made by each server each year, grouping by department, campus, type of activity, among others.

5.3 Data Modeling

The definition of which data is relevant to the work was based on studies of the data source along with meetings with the project office and director general of the institution. The scheme used was the star model, where each dimension represents a single table and are directly linked to the Fact table. The dimension tables created were:

- DimServer - represents the institution's servers. Its attributes are: server genre, title (doctor, master, graduate), etc;

- DimCampus - represents the institution's units, such as Campus I, Campus Timóteo, Campus Contagem, etc;

- DimTipoActivity - represents the types of activities required by the institution, which can be: didactic activities, research, coordination, etc;

- DimDepartamento - represents the academic departments of each campus of CEFET-MG, which can be: Computer Department, Electrical Department, Civil Production Department, etc;

- DimAcademicActivity - Represents academic activities such as: participation in committees, board member, student orientation, publication of articles, etc;

- DimTempo - the year of academic charges. As charges are reported per year, we chose to restrict the time dimension at this level of granularity.

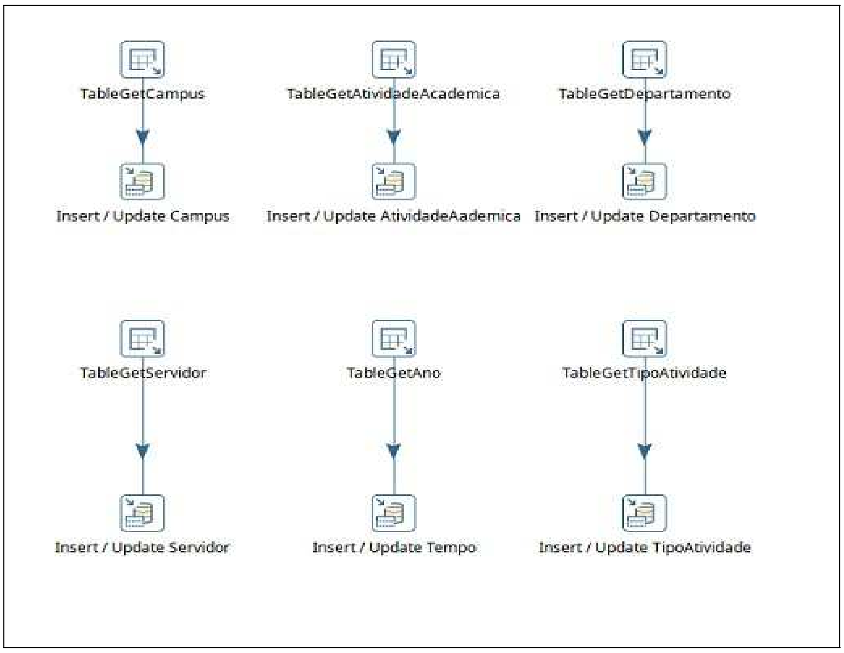
The fact table - WorkFactPlan - aims to maintain the metric and foreign keys related to each dimension table. The updated metric is the Score, which represents the sum score of the academic burden by activity of each teacher.

5.4 Extraction, Transformation and Load Process - ETL

For the data ETL process the Pentaho Data Integration (PDI) tool was used. From extracting data from the available data source, the flow consists of performing the appropriate treatments and storing them in the created DW.

The first step was to establish a connection between the data source and DW. Once the connections were configured, for each modeled dimension table, the transformation flow and data load from the transactional model to the dimensional model were created. Each entity that has been selected from the CEFET-MG database aims to insert or update data in the DW. Figure 1 shows that for each selected entity of the source the Table Input and Insert / Update components were used.

Figure 1: ETL flow for all Dimension type entities.

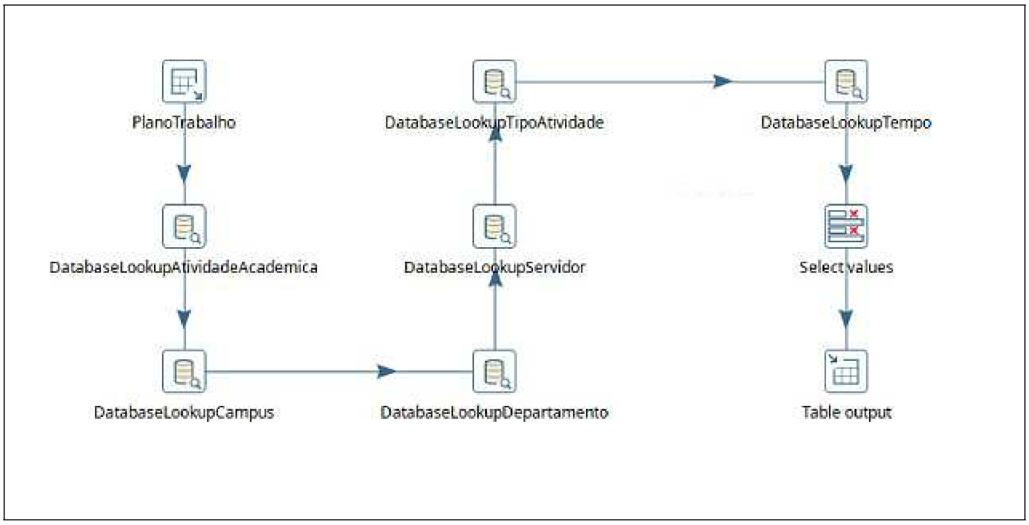


Source: elaborated by the authors.

The first step of the flow, Table Input, shown in Figure 1, works in such a way that from a query it is possible to return the data in the corresponding dimension and store it in the DW by its Insert / Update component for each dimension.

In sequence, the fact table ETC process is represented by Figure 2. The Table Input component selects the data to be loaded into the DW and the Database Lookup component maps the transactional model keys to the dimensional model to match correctly. between the keys. For each dimension created you must have its corresponding lookup process.

Figure 2 - ETL flow to Fact table.



Source: elaborated by the authors.

After performing the lookup of each Dimension, the next step of the ETC process used the Select Value component. The last step was to use the Table Output component, which aims to insert all data selected in the Fact table.

It is important to emphasize that the order of execution of the steps must be respected. You must first load the Dimensions tables and then load the Fact table, since relationships are made from the primary keys of the Dimensions table with the foreign keys in the Fact table.

5.5 Cube Creation and Assembly

For the creation of the cube, it is first configured and established the connection to the DW database, in which the cube component tables were chosen, separating them into Dimensions and Fact. The relationships between the tables between SK and PK have been created. That is, for each foreign key of the Fact table it has been linked to a primary key of its corresponding Dimension table.

5.6 Data Visualization

To visualize the data, processed in the previous step, it was necessary to define the attributes and hierarchies that would be available to the user. For this purpose, Saiku Analytics was used a plugin within the Pentaho Server tool, which allows you to visualize the cube and build graphs with the extracted data. Displaying information in the form of graphs makes it possible to query from different points of view. Some of the criteria available for creating the total score view were by:

campus, department, and activity type, all separated by year.

From the graphics generated in Saiku, a dashboard was created using another plugin present in Pentaho Server: the CDE Dashboard. CDE Dashboard lets you create a more dynamic data visualization area for the end user. In this sense, a dashboard has been created where you can select the year in which you want to display the data. From the year selection, graphs and tables are generated corresponding to the total score of that year, segmenting the information by the predefined dimensions.

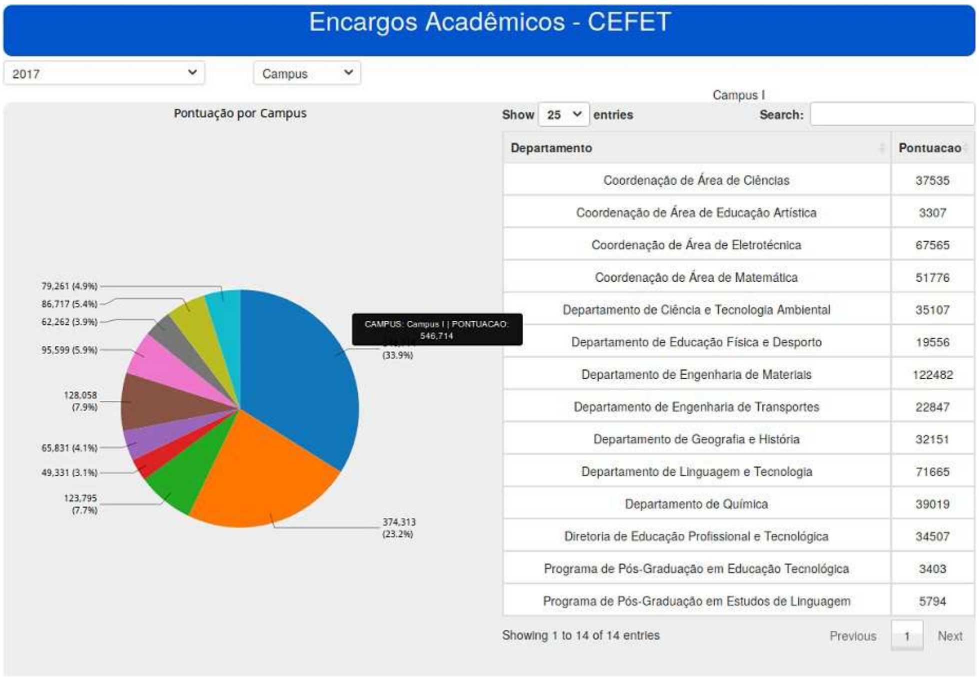
The dashboard was defined based on the end user demand, that is, it aims to enable the visualization of data dynamically and quickly. In this sense, the visualization presented in the dashboard can be changed according to the user's need and / or according to the data and perspectives that are considered most important by him.

6 Results Analysis

In this section some modes of visualization of the results are presented. It has been set as the base criterion that the user will always see the total amount of charge score per year. Therefore, selection filters such as campuses and servers will always be separated by year.

One possible way for the information to be presented to the user is through the Cube visualization in the Saiku Analytics plugin in table form and also through the graphical view. In the tool you can choose the type of graph you want (pie, line, bar, etc.) to make the visualization of the data more intuitive. Note that the graphical display displays the value in percent. The total score value can be seen by hovering over the desired chart slice. Another way to present the data display to the user is to make a visualization through a dashboard. The dashboard displays graphs, charts, and tables so that the view is grouped on the same screen. Figure 3 shows an example dashboard created where you can select the year and by which dimension you want to analyze.

Figure 3: CEFET-MG Academic Burden Dashboard.

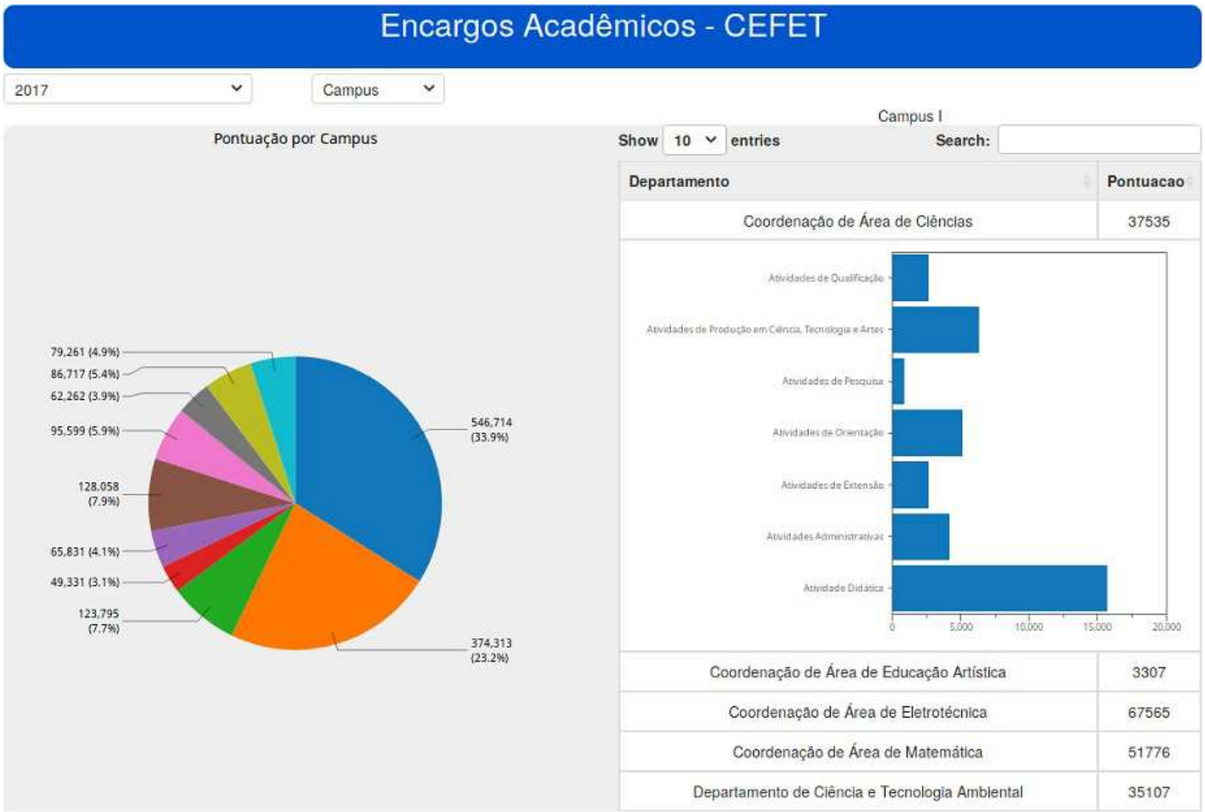


Departamento	Pontuacao
Coordenação de Área de Ciências	37535
Coordenação de Área de Educação Artística	3307
Coordenação de Área de Eletrotécnica	67565
Coordenação de Área de Matemática	51776
Departamento de Ciência e Tecnologia Ambiental	35107
Departamento de Educação Física e Desporto	19556
Departamento de Engenharia de Materiais	122482
Departamento de Engenharia de Transportes	22847
Departamento de Geografia e História	32151
Departamento de Linguagem e Tecnologia	71665
Departamento de Química	39019
Diretoria de Educação Profissional e Tecnológica	34507
Programa de Pós-Graduação em Educação Tecnológica	3403
Programa de Pós-Graduação em Estudos de Linguagem	5794

Source: elaborated by the authors.

The pie chart shows all of the institution's campuses with the total campus score for a given year selected and the percentage of how much each value represents considering the institution's total score. The department table on the right changes dynamically according to the campus selected on the pie chart. All departments on the selected campus are displayed along with their respective scores. You can click on a department so that this action allows a more detailed view of the department, showing how the score within that selected department is distributed among activity types. This action is represented in Figure 4.

Figure 4: Expanded CEFET-MG Academic Burden Dashboard.



Encargos Acadêmicos - CEFET

2017 | Campus

Campus I

Show 10 entries Search:

Departamento	Pontuacao
Coordenação de Área de Ciências	37535
Coordenação de Área de Educação Artística	3307
Coordenação de Área de Eletrotécnica	67565
Coordenação de Área de Matemática	51776
Departamento de Ciência e Tecnologia Ambiental	35107

Source: elaborated by the authors.

The dashboard can be fully customized. In this way, the types of graphs, the arrangement on the screen, the way information will be displayed and the design can be changed according to the user's need and demand.

These data presentation tools are very powerful and extremely important for the BI process as they serve as a basis for obtaining information for decision making and behavior study of the institution.

7 Conclusion

In this work it was possible to conceptualize BI, what it is for, how to deploy and demonstrate its importance. All with the objective of, through a real application, develop software to analyze the academic burden data of CEFET-MG teachers.

The study had some practical limitations as the data provided were limited and the tool used was not 100% exploited with all its free resources. However, it can be concluded that the Pentaho tool meets the needs and objectives of the study. The tool made it possible to extract data and transform it into relevant information. In addition to enabling the customization of how data is displayed. With this, the institution's directors and decision makers can have a viable, easy-to-use tool that provides information quickly and effectively.

The main contribution of the study was that it was possible to demonstrate that BI systems can be implemented in the management of educational institutions, helping in the visualization of information and decision making, which can guarantee a better distribution of the activities of teachers and subsidize the decision making. decision to optimize the distribution of human resources based on a burden distribution criterion.

References

ALMEIDA, A. M. R.; CAMARGO, S. d. S. Aplicando técnicas de business intelligence sobre dados de desempenho acadêmico: Um estudo de caso. 2015.

CANO, J. L. Business intelligence: Competir com información. 2007.

CARALT, J. C. I.; DIAZ, J. C. *Introducción al Bussines Intelligence*. [S.l.]: Editorial UOC, 2012. 17 - 93 p.

DIEBOLD, F. X. The origin(s) and development of “big data”: The phenomenon, the term, and the discipline. 2018.

ELBASHIR, M. Z.; COLLIER, P. A.; DAVERN, M. J. Measuring the effects of business intelligence systems: The relationship between business process and organizational performance. *International Journal of Accounting Information Systems*, p. 135 – 153, 2008.

FONSECA, J. J. S. Metodologia da pesquisa científica. 2002.

GIL, A. C. *Como elaborar projetos de pesquisa*. 4. ed. [S.l.]: Atlas, 2007.

GONÇALVES, M. Extração de dados para data warehouse. Axcel Book, v. 1, 2003.

HOKAMA, D. D. B. et al. *A modelagem de dados no ambiente data warehouse*. Tese (Doutorado) — Universidade Presbiteriana Mackenzie, 2004..

ISIK, O.; JONES, M.; SIDOROVA, A. Business intelligence success: The roles of bi capabilities and decision environments. *Information Management*, v. 50, p. 13 – 23, 2013.

KIMBALL, R.; ROSS, M. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. 2. ed. [S.l.]: John Willey and Sons Inc, 2002.

MUNHÓS, G. M. *Business Intelligente sobre as Produções Agrícolas do Estado de Minas Gerais: estudo de caso do comércio no Mercado Livre do Produtor da CeasaMinas*. Monografia — Centro Federal de Educação Tecnológica de Minas Gerais, 2018.

RAMAKRISHNAN, T.; JONES, M.; SIDOROVA, A. Factors influencing business intelligence (bi) data collection strategies: An empirical investigation. *Decision Support Systems*, v. 52, p. 486 – 496, 2012.

SILAHTAROğLU, G.; ALAYOGLU, N. Using or not using business intelligence and big data for strategic management: An empirical study based on interviews with executives in various sectors. *Procedia – Social and Behavioral Sciences*, v. 235, p. 208 – 125, 2016.

SILVA, A. P. *Data Warehouse e Data Mart como Ferramentas de Inteligência em negócio (BI)*. Tese (Doutorado) — Universidade Estadual de Maringá, 2004.

SINGH, H. S. *Data Warehouse: Conceitos, Tecnologias, Implementação e Gerenciamento*. 1. ed. [S.l.]: Makron Books, 2001.