DW-ENEM: a Data Warehouse for Analytics Exploration of National High School Exams Results

Saulo Fonseca[1], Lucas Penedo[1], Breno Antunes[1], Sérgio Lifschitz[2], Maria Luiza Machado Campos[3], Ana Carolina Almeida[1]

[1]Department of Computer Science – State University of Rio de Janeiro (UERJ)

[2]Department of Informatics, Pontifical Catholic University of Rio de Janeiro (PUC-Rio)

[3]Department of Computer Science, Federal University of Rio de Janeiro (UFRJ)

{saulomartins.martins, lucas.rmpenedo, breno.cantunes626}@gmail.com, sergio@inf.puc-rio.br, mluiza@ppgi.ufrj.br, ana.almeida@ime.uerj.br

Abstract. *The National High School Exam (ENEM) aims to assess the quality of high school education and serve as a gateway for students to public and private universities (national or foreign). Although these data are open, it is difficult to interpret because of the available format. This paper aims to structure and analyze ENEM data at a multidimensional level. We propose a multidimensional model of the data warehouse and present some graphs resulting from analyzes requested by coordinators of Brazilian schools. Such analyzes include information on people with special needs and data on socio-economic issues.*

1. Introduction

The Law on Access to Information (LAI - nº 12.527/2011) guarantees that any person, individual or legal entity has access to public information from government agencies and entities. Given this, INEP (National Institute of Studies and Educational Researches Anísio Teixeira) provides several data on education in Brazil. One of these data is the annual result of ENEM (National High School Exam). ENEM comprises the assessments necessary for a student to enter a university. The ENEM results are available in CSV format files.

We may analyze ENEM data from different perspectives: school coordinator or responsible for students who will enter high school. For example, the school coordinator may use the data to assess which subjects and themes within those subjects the school needs improvement. On the other hand, those responsible for students may check the performance of students in each school, having an overview of each school. With this, it is possible to decide, for example, which schools their children could study at to enter, in the future, a reputable university. We may make both decisions considering the students' grades and the school's location (near the residence). So, it is not enough to know the student's average grade. We could also consider all the information in the context of the assessment.

Considering the wealth of information that ENEM brings and, at the same time, the difficulty of analyzing the amount of data available in text files, it is necessary to have a data model that structures these data and a database that stores them.

This work proposes a data warehouse model to store ENEM's historical data and presents some of the results of the analyzes obtained. Background and related works are present in Section 2. We detail in Section 3 the data warehousing design. Finally, we show the resulting graphics and conclude this paper in Sections 4 and 5, respectively.

2. Background and Related Work

Inmon *et al.* (2008) characterized a **data warehouse** (DW) as a subject-oriented, integrated, nonvolatile, time-variant collection of data supporting management's decisions. DW insertions are handled by the **ETL** (extract, transform, load) process, which does a large amount of preprocessing. The extract, transformation, and load (ETL) system of the DW/BI environment consists of a work area, instantiated data structures, and a set of processes [KIMBALL & ROSS, 2013]. Extraction is to get data into the data warehouse environment. Extracting means reading and understanding the source data and copying the data needed into the ETL system for further manipulation. After obtaining the data, there are transformations, such as cleansing, combinations (from multiple data sources), and deduplication. The final step of the ETL process is the physical structuring and loading of data into the presentation area's target dimensional models. The model used in this work is the **snowflake**; as there is a hierarchical relationship in a normalized dimension table, low-cardinality attributes appear as secondary tables connected to the base dimension table by an attribute key [KIMBALL & ROSS, 2013].

2.1. Related works

Vieira *et al.* (2019) presented a transactional and relational approach to ENEM data persistence. In addition, they offered some database tuning decisions to optimize the performance of some queries. Stearns *et al.* (2017) analyzed the possibility of predicting students' grades based only on their socioeconomic information using a regression model. Cabral *et al.* (2012) proposed using an ontology to represent ENEM data geographically grouped. Through this representation, they seek to link these data with others through the RDF standard. There is a platform developed with data based on ENEM 2013 about the performance of Brazilian schools. This platform has not been updated at the time of writing this paper. Several works use ENEM data to extract knowledge. The work that comes closest to ours is that of Vieira *et al* (2019). We extended the model used in their work to include information from people with special needs and the answers to socioeconomic questionnaires.

3. DW-ENEM

In this section we present the requirements (queries) raised after interviewing a high school coordinator (Table 1), the proposed multidimensional model and an example of a workflow used to carry out the ETL process.

Table 1. Application Requirements

#	Requirements
Q1	What is the average (avg) student grades by state, by type of exam, and by year?
Q2	What is the financial profile of students grouped by state and by year?
Q3	What are the average grades for students with special needs, by state and year?

Q4	What is the total number of people enrolled in the ENEM by year, sex, and state?
Q5	How many (percentage) of hits and misses by exam, state, and school?
Q6	What is the avg of the test scores for the schools compared to the national avg?
Q7	What is the test scores average for the schools, compared to the state average?
Q8	How many people with special needs have enrolled in ENEM per state and year?
Q9	How many people with special needs have signed up, grouped by type of need, state, and year?
Q10	For every kind of disability, which aid was requested, grouped by year?

3.1. DW Design

The proposed model (Figure 1) comprises nine dimensions and three fact tables. The dimensions have information regarding the types of special needs that candidates have (d_special_needs); registered Brazilian schools (d_school); the location of the school, the student's residence, the exam's municipality (d_locality); the socio-economic profile (d_socioeconomic_profile); to exams (d_question and d_exam); to the candidate's profile (d_candidate_profile); to time (d_time); and the aid that candidates may request on the day of the exam (d_aid). The three facts correspond to the candidate's subscription (f_subscription), the exam performance in general (f_exam_performance), and the performance by the question of each candidate (f_question_performance). Some measurements in the f_exam_performance fact table sum up the correct answers (correct_quantity) and the number of errors (error_quantity) per candidate in each exam.

3.2. ETL Process

The created ETLs are responsible for taking the data from the transactional model to the multidimensional one. Figure 2 shows an example of ETL for the dimension d_candidate_profile.

The workflow was created in the Pentaho Data Integration tool[1] following these steps: (1) data extraction from the transactional source, e.g., the candidate table; (2) applies a transformation to the source data. For instance, if data is an integer and needs to be a string; (3) maps the numeric values for text in the tp_gender, tp_civil_status, nationality, in_tp_education, tp_color_race columns; (4) applies a transformation in the type of data from previous step; and (5) populates the DW d_candidate_profile table.

The Pentaho PDI tool had some memory limitations in generating a fact table about the candidates' performance in each question (f_question_performance). This way, pre-processed temporary tables are generated through workflows before loading the fact table. In addition, we have developed a shell script to retrieve the data from the temporary tables and join them with the others for the final fact table load.

[1] https://help.hitachivantara.com/Documentation/Pentaho/7.1/0D0/Pentaho_Data_Integration, accessed in June 2022.

In addition, there is a lack of standardization of data among the analyzed years, besides some inconsistencies. Some data had no relevance for our analysis, such as asterisks and dots. Thus, in order to not compromise the data generated from the processing, null values replaced these data. In addition, some semantic inconsistencies were also found, such as students with grades, but without the answers recorded for the exams. No data duplication found.

4. Data analysis

Due to lack of space, we show only three graphs as answers to questions: Q3, Q9 and Q10. These questions cannot be answered by the related works. In addition, we present a performance analysis between a query performed in the transactional database and in the DW to have a gain parameter in this type of analysis.

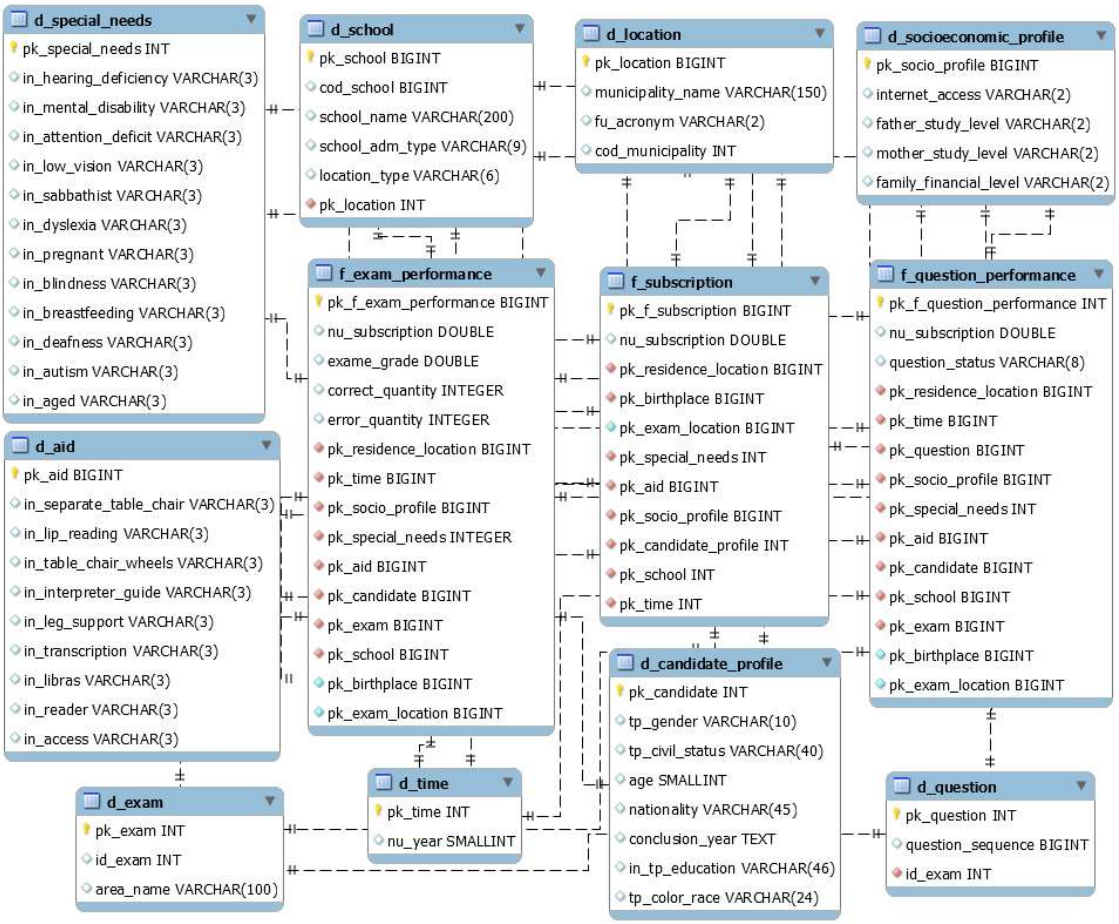

Figure 1 ENEM - Data Warehouse Model

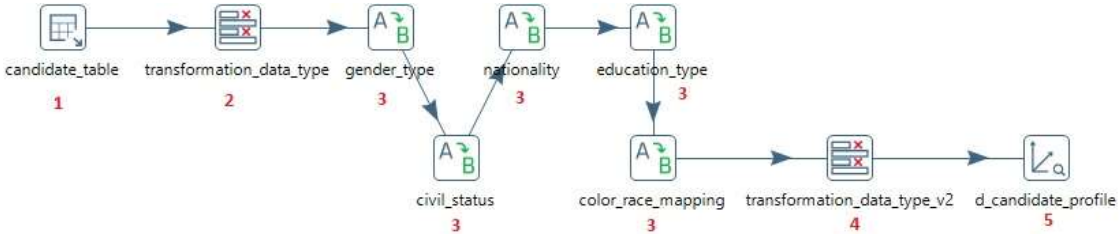

Figure 2 Example of ETL Process to ENEM data

Figure 3 displays the average grades of candidates with special needs grouped by exam, by state, and by year. The averages of the Human Sciences (dark blue color) are the highest and the Mathematics (purple color) are the lowest. The south and southeast regions stand out with the highest averages in the graphs.

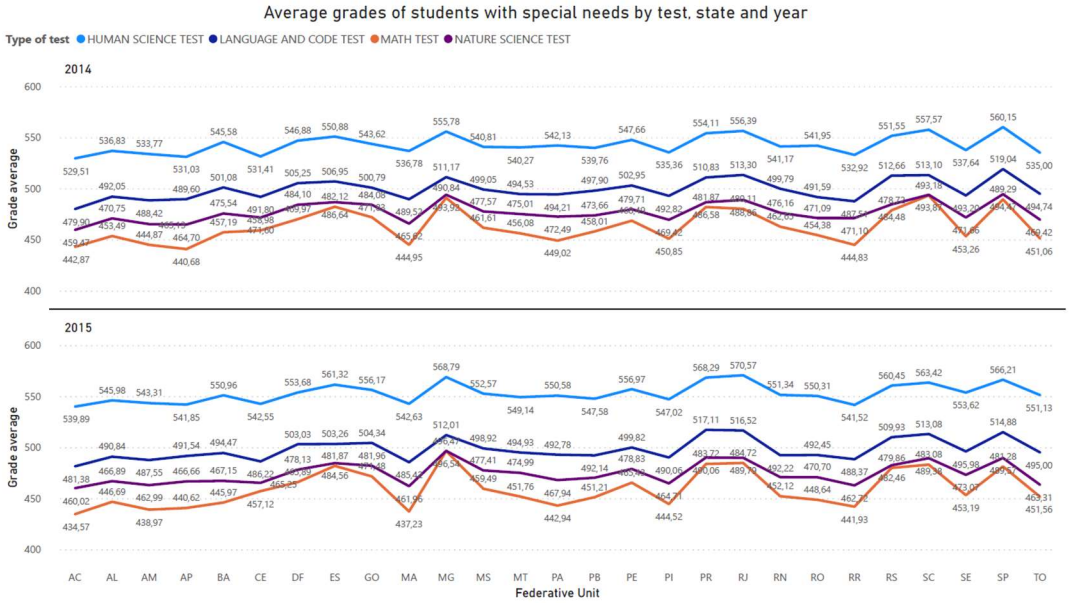

Figure 3 Average grades of candidates with special needs

Figure 4 shows that the highest amounts of special needs that were requested are: sabbath and low vision. In addition, the two states that had the highest numbers of requests in 2014 and 2015 were: SP and BA.

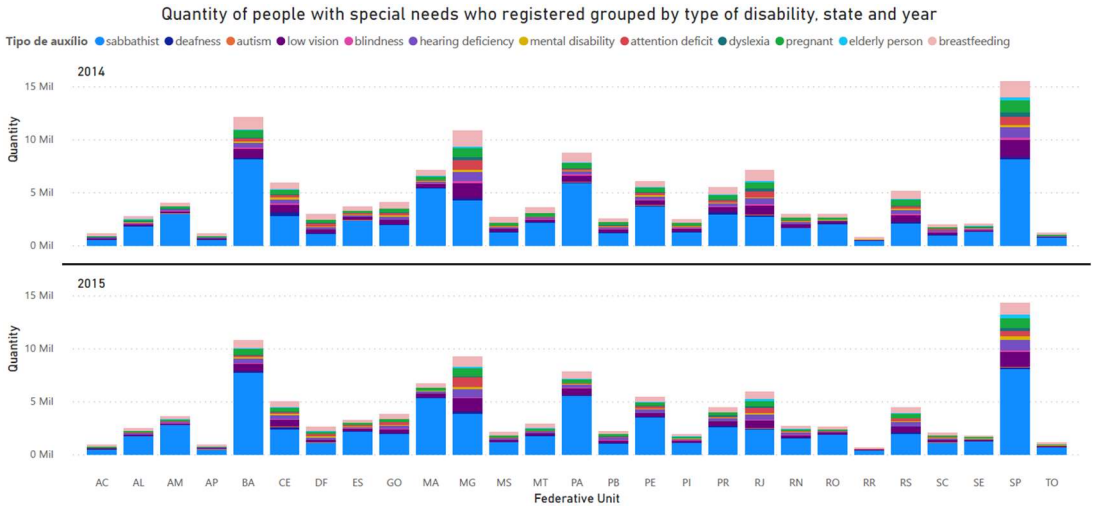

Figure 4 Registered People with special needs

Figure 5 presents that in the case of autism (selected in the upper right side), the aids requested were: access, leg support, libras, reader, table_separate_chair and transcription for the year 2014. In 2015, the aid leg support and libras were not chosen. These two aids draw attention to the type of disability selected. Analyzing the data, it was found that the applicant who selected leg support has autism and physical disability. The other enrolled (selected libras) has autism, hearing impairment and an indicator of deafness.

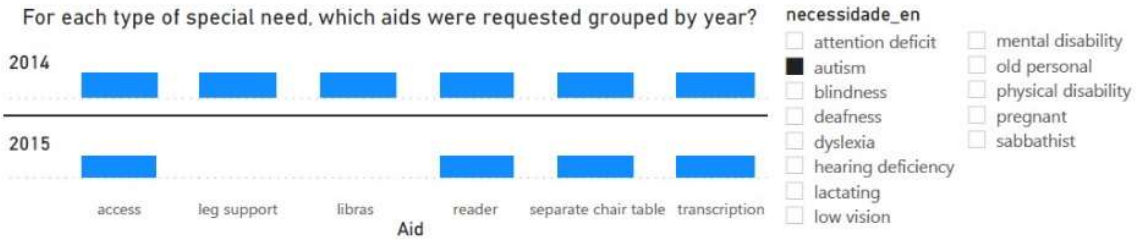

Figure 5 Aids requested by people with autism

Looking at performances, we have a query that retrieves the number of corrects and errors per candidate's test in Rio de Janeiro. We have submitted the same question to the transactional database and the DW. This data may be found directly in the exam performance fact table, joining only the exam and location dimensions, the DW returned the result in 1.34 minutes, while the transactional one took 70.19 minutes. The transactional had to query eight tables and perform the sum and aggregation of data. The reader may access the code for ETL workflows and other graphics at: https://github.com/martins9/arquivos_dw. The size of the DW database is 311 GB.

5. Conclusions

This work demonstrated that the DW model proposed for ENEM could answer questions that may be useful in improving education in Brazil, especially for people with special needs—for example, identifying the areas (disciplines) with which these people have greater difficulty. In addition, by knowing the types of aid needed for each kind of person, the school may improve its structure to increase the inclusion of people. Governments may also prioritize improving the system according to the concentration of the type of disability the majority have (e.g., low vision). In future work, we seek to develop a platform that external people may query and perform integrations with other open data that may result in more sophisticated analyses.

References

Cabral, S. P., Beduschi, N. B., Zancanaro, A., Todesco, J. L., Gauthier, F. A. O. (2012) Aplicando Linked Data na publicação de dados do ENEM. Anais do Ontobras.

Inmon, W. H., Strauss, D., & Neushloss, G. (2008) DW 2.0: The architecture for the next generation of data warehousing. Elsevier.

Kimball, R. & Ross, M. (2013) The data warehouse toolkit: the definitive guide to dimensional modeling. John Wiley & Sons.

Stearns, B., Rangel, F., Firmino, F., Rangel, F., Oliveira, J. (2017) Prevendo Desempenho dos Candidatos do ENEM Através de Dados Socioeconômicos. Anais do XXXVI Concurso de Trabalhos de Iniciação Científica da SBC.

Vieira, A. W., Cantergiani, G., Salgueiro, M. D. A., Pereira, S. V., Souza, V. A. L. L., Oliveira, R. P., Lifschitz, S. (2019) BioBD-ENEM: Migrando Grandes Planilhas para um Sistema de Banco de Dados na Web. Graduation Student Workshop (WTAG), in conjunction with SBBD 2019.