Unlocking Educational Vulnerabilities Insights: Modeling Encceja Data for Student Performance Analysis.

Diogenes Oliveira [a]

[a] Institute of Exact Sciences (ICEx), Federal University of Minas Gerais (UFMG), Belo Horizonte, Minas Gerais, Brazil, diogenesvazmelo@gmail.com.

Abstract. This study focuses on modeling open data from the National Examination for the Certification of Skills for Young People and Adults (Encceja), an exam that allows individuals who haven't completed elementary or secondary education at the appropriate age to obtain a certificate equivalent to a regular school diploma. Encceja plays a vital role in Brazilian education, providing valuable insights into student performance across diverse knowledge domains. Using dimensional modeling techniques by Ralph Kimball and Margy Ross, the study organized a fact table and dimensions within a cloud computing environment on Amazon Web Services (AWS). The data Extraction, Transformation, and Loading (ETL) involved the use of Python in a Sagemaker notebook instance and queries on Athena. The objective was to pave the way for developing a Business Intelligence (BI) dashboard that stimulates discussions for enhancing the Brazilian educational system. It is expected to develop graphical visualizations as a follow-up to this study.

Keywords. Encceja, Dimensional data modeling, Data Warehouse, Brazilian educational system.

1. Introduction

A Business Intelligence (BI) dashboard is "a graphical interface that presents data in a visual format making it more intuitive to extract valuable information and make informed decisions" [1]. It is an important tool for strategic decision-making, as it enables organizations to analyze relevant information about their performance, competitors, and industry trends. Data analysis can help companies identify opportunities and threats, and improve their processes, products, and services.

In the context of basic education, dashboards have the potential to provide critical information about the population and the effectiveness of adopted policies, driving government officials to take actions that lead to significant improvements in educational practices.

The analysis of open data from Brazilian national exams may allow the identification of learning vulnerabilities, helping to understand the areas in which students have the most difficulty. In addition, it also contributes to verifying whether the implemented actions are achieving the proposed objectives and to directing resources in a more efficient way.

One of the most relevant exams administered annually throughout the Brazilian territory is the National Examination for the Certification of Skills for Young People and Adults (Encceja), whose primary goal is to assess the skills and knowledge of young people and adults who have not completed their elementary or secondary education at the appropriate age. To be eligible to take the Encceja, candidates must be at least 15 years old for the elementary level and 18 years old for the secondary level. Candidates who pass at least 50% of the questions in each subject of the exam are awarded a certificate that is equivalent to a diploma from a regular school.

The Encceja microdata are released annually by the National Institute for Educational Studies and Research Anísio Teixeira (Inep), and gather a set of detailed data about the exam, but without the possibility of identifying people, which is in compliance with the regulations provided for in the Brazilian General Data Protection Law (LGPD) [2].

Consequently, there is material to support the development of a BI dashboard with the historical data of the Encceja exams, which can be used to explore and give visibility to possible educational vulnerabilities in Brazil.

In order to develop such dashboard, it is first necessary to organize the data elements and standardize their relations – or, in other words, to define a data model. The present research aims to discuss the process of modeling the Encceja data for further development of a BI dashboard.

2. Research Methods

This study is based on the comprehensive bibliographical content authored by Ralph Kimball and Margy Ross, proponents of the dimensional data model in the development of Data Warehouses (DWs) and BI systems.

The dimensional model is a prevalent approach to designing DWs in the BI industry. It involves constructing fact tables to store performance metrics an organization aims to track, such as sales or revenue, and dimension tables that describe contextual aspects of the data, such as time, location, or customer [3].

As a first step, it is crucial to dedicate time understanding the databases through a thorough review of the accompanying documentation and validation of data dictionaries. All files are publicly accessible, released on an annual basis, and can be retrieved from the Brazilian federal government's integrated platform, particularly in the section dedicated to Inep's microdata. These microdata are characterized as "the most detailed level of data, enabling in-depth analysis and exploration possibilities" [4].

Moving forward, the subsequent phase includes deciding on the development environment (cloud or on-premises) and designing the architecture.

Next, the data from the webpage where Inep makes it available is consumed. Consequently, the Extraction, Transformation, and Load (ETL) of this data can be performed into a relational environment following the modeling of facts and dimensions.

Finally, the last step involves manipulating the data to construct graphical visualizations that can provide users with insights into Encceja and education in Brazil.

3. Data sources

The Encceja data [5] [6] [7] [8], for each of the examined years, comprised four files in Comma-Separated Values (CSV) format:

- Regular Application Data File: includes registration details, exam responses, and socioeconomic questionnaire answers for national candidates without freedom restrictions;
- Application Data File for Incarcerated Individuals: contains registration details and exam responses for national candidates with freedom restrictions;
- Socioeconomic Questionnaire Responses File for Incarcerated Individuals: includes information on socioeconomic questionnaire responses for national candidates with freedom restrictions;
- Exam Items Detail File: presents a comprehensive structure of the exam items (questions).

Furthermore, documentation files and data dictionaries were employed to provide descriptions for the codes used in the files generated by Inep.

4. Dimensional Modeling

Kimball and Ross (2013) [3] describe dimensional modeling as an approach to database design centered on representing business information in a clear and easily comprehensible manner.

This modeling approach relies on two primary structures: fact tables and dimension tables. Fact tables contain quantitative metrics and measures crucial for analysis, such as sales, profit, quantity, etc. These tables represent the events or transactions within the business and play a central role in dimensional modeling [3].

In contrast, dimension tables provide context for the measures found in the fact table. They encompass descriptive attributes facilitating the analysis and segmentation of data, including dates, products, customers, locations, and more. Dimension tables are instrumental for filtering, grouping, and organizing data, enabling users to analyze information from diverse perspectives [3].

The application of dimensional modeling to the Encceja data [5] [6] [7] [8] led to the identification of 8 dimensions:

- DIM_CERTIFICACAO: Encompasses details regarding the certification sought by the candidate, whether for Elementary or High School;
- DIM_FAIXA_ETARIA: Describes the age groups assigned to candidates as defined by INEP;
- DIM_SEXO: Provides information on the sex of the candidate;
- DIM_UNIDADE_FEDERATIVA: Describes the Federative Unit (State) where the candidate realized the exam;
- DIM_ENTIDADE_CERTIFICADORA: Specifies the educational entity responsible for certifying the exam.
- DIM_PRESENCA_PROVA: Describes the candidate's status during the exam, indicating if they were present, absent, or eliminated.

- DIM_ITENS_PROVAS: Represents a dimension derived from the file detailing exam items.
- DIM_QUESTIONARIO: Details questions and answer options for socioeconomic questionnaires applied to national candidates without freedom restrictions (regular).

It's crucial to highlight that the questionnaire dimension excludes responses from candidates with freedom restrictions. This decision stems from Inep's separation of socio-economic questions from the objective questions for such candidates, aligning with the requirements of LGPD [2]. Consequently, the file containing socio-economic questionnaire responses from individuals deprived of freedom was not incorporated.

Additionally, the fact table preserved the responses of each candidate registered in a certain year exam as grain. For candidates with freedom restrictions, the fields related to socio-economic questions were transformed into null values.

5. Data architecture

Employing cloud computing technology to establish a DW presents numerous noteworthy benefits. As per Kimball and Ross (2013) [3], cloud computing offers nearly limitless scalability, allowing the infrastructure to be adjusted up or down according to requirements. Moreover, users only pay for the resources utilized, resulting in substantial cost savings compared to acquiring and maintaining hardware.

In this study, the choice was made to utilize the cloud computing resources provided by Amazon, specifically Amazon Web Services (AWS), for the structuring of the DW and the execution of the ETL process.

The adopted architectural model closely resembled the one outlined by Zaccarelli (2020) [4], wherein foundational architecture services were provisioned through CloudFormation. This model establishes S3 buckets for uploading data files and storing query outputs executed in Athena, while also configuring the workgroup for its utilization. Furthermore, it defines a database in the Data Catalog (Glue) and configures permissions in the IAM used by Glue to access services. Lastly, the script generates the Sagemaker notebook instance, facilitating the execution of codes for data ingestion and cataloging.

6. Data extraction, transformation, and loading (ETL)

The processes involving data retrieval from Inep, encompassing tasks like decompression, file selection, unwanted character removal, UTF-8 encoding conversion, ingestion into the bucket, and registration of raw tables in the Glue data catalog, were executed with Python in a Sagemaker notebook.

Following this, Athena was utilized to allocate new unique numerical IDs to textual fields, such as certifying entities' names, and to conduct data type conversions. Moreover, Athena facilitated the decomposition of specific fields into columns, notably those containing candidates' responses for individual exam booklets (mathematics, languages, humanities, and natural sciences), which were initially stored as 30-character strings but were segregated into 30 distinct columns.

Additionally, Athena played a role in the manual loading of dimensions to enhance them with general knowledge (e.g., adding the geographical region of each state to the state dimension). This process also included information that demanded meticulous evaluation and comprehension of documentation and data dictionaries to consolidate data from the 4 years of exam application (e.g., the questionnaire dimension was loaded manually after a thorough assessment, given that the description was exclusively available in the data dictionary, and the questions – or even response options – could vary across different years).

With the dimensions populated, a straightforward query sufficed to populate the fact table. This query involved multiple table joins to retrieve the IDs of textual fields described in the dimensions.

7. Conclusions

This study delved into the significance of establishing a dimensional model for open Encceja data made available by Inep. Encceja holds considerable importance in Brazilian education, offering valuable insights into student performance across various knowledge domains.

By structuring a DW with a fact and dimensional tables, it becomes feasible to centralize and organize the data, presenting a comprehensive and integrated perspective on students' performance over time and across different regions of the country. The analysis of this data can provide crucial insights for enhancing the Brazilian educational system.

Furthermore, leveraging cloud computing technologies for DW structuring yields substantial benefits. Cloud computing offers scalability, flexibility, and data availability, enabling the infrastructure to be tailored to the requirements of the analysis.

Analyzing Encceja data through a well-structured, cloud-based DW has the potential to provide the base for a BI dashboard and contribute to more informed decision-making in the field of education. With a good dashboard, educational managers can discern trends, patterns, and performance gaps, guiding resources and efforts towards enhancing the quality of teaching and student development.

Nonetheless, it is vital to emphasize that the structuring of a Data Warehouse (DW) and the analysis of educational data present challenges,

including issues related to data quality and integration, privacy concerns, and information security. As a result, adopting suitable data governance practices and ensuring adherence to privacy policies and data protection, particularly in compliance with LGPD [2], becomes imperative.

The next phase of this study envisions the development of graphical visualizations through a BI software (e.g., Microsoft PowerBI, Tableau or Qlik). These visualizations could be designed to facilitate the exploration of available data, offering a comprehensive understanding of various facets of the exam and basic education in Brazil.

8. References

[1] Awad, M., Al Redhaei, A., Fraihat, S. Using business intelligence to analyze road traffic accidents. *Proceedings of the Central and Eastern European eDem and eGov Days*. 2022; p. 83-92.

[2] BRASIL. Lei nº 13.709, de 14 de agosto de 2018. *Lei Geral de Proteção de Dados Pessoais (LGPD)*. Brasília, DF: Presidência da República, [2020].

[3] Kimball, R., Ross, M. *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling* (3rd Edition). Wiley, Indianapolis; 2013; 720p.

[4] Zaccarelli, S. *Ebook: Aplicando Analytics para geração de Insights com dados do Exame Nacional do Ensino Médio*. 2020; 21 p.

[5] Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). *Microdados do ENCCEJA 2018*. [online]. Brasília, Inep; 2022.

[6] Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). *Microdados do ENCCEJA 2019*. [online]. Brasília, Inep; 2022.

[7] Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). *Microdados do ENCCEJA 2020*. [online]. Brasília, Inep; 2022.

[8] Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). *Microdados do ENCCEJA 2022*. [online]. Brasília, Inep; 2022.