Journal Homepage: -www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI:10.21474/IJAR01/21017
DOI URL: http://dx.doi.org/10.21474/IJAR01/21017

RESEARCH ARTICLE

EDMLENS DATAFLOW: AN ANALYTICAL ARCHITECTURE FOR EXTRACTION, TRANSFORMATION, LOADING, ANALYSIS AND VISUALIZATION OF EDUCATIONAL DATA

Douglas Francisquini Toledo[1], Ronaldo Celso Messias Correia[2] and Camila Tolin Santos[3]

1. Master of Science in Computer Science, student in the Graduate Program in Computer Science, Paulista State University "Júlio de Mesquita Filho", Presidente Prudente, São Paulo, Brazil.
2. Assistant Professor, Department of Mathematics and Computing, Paulista State University "Júlio de Mesquita Filho", Presidente Prudente, São Paulo, Brazil.
3. Master of Science in Mathematics, student in the Graduate Program in Computer Science, Paulista State University "Júlio de Mesquita Filho", Presidente Prudente, São Paulo, Brazil.

Manuscript Info

Manuscript History
Received: 27 March 2025
Final Accepted: 30 April 2025
Published:May 2025

Key words:-
Data Science. Data Analytics Architecture. Academic Analytics

Abstract

Over time, educational institutions have adopted new technologies or improved those they already used, thus generating large volumes of data. This data can help in the process of choosing certain strategies to improve teaching. However, there are many situations in which data is stored but is not used by management or other professionals to support decision-making. In some cases, this may occur due to the difficulty in finding a data analysis tool or method that does not require specific knowledge for its use. For this reason, this paper presents a bibliographic review of the literature with the objective of supporting the development of a data analysis architecture, called ``EdmLens DataFlow'', which integrates processes of protection, transformation, loading, analysis and visualization of educational data. This architecture offers a holistic structure that, in addition to facilitating data interpretation, improves informed decision-making, promoting a monitoring and planning process. Therefore, the literature review covers the areas of Data Science, Knowledge Discovery in Databases (KDD), Data Analysis Architecture and Academic Analysis, and highlights the following elements as results for the development of ``EdmLens DataFlow'': data protection from different sources and formats, working with structured, semi-structured and unstructured data; data transformation to make them all structured, performing normalization, cleaning and integration; loading the data into a Data Warehouse, as this storage approach works with structured data; data processing and analysis through artificial intelligence and machine learning; and data visualization, based on the analysis, using reports and interactive dashboards.

Corresponding Author:- Douglas Francisquini Toledo
Address:- Rua José Ângelo dos Santos, 211, Parque Imperial, Presidente Prudente, São Paulo, Brasil, CEP: 19028-235.

Introduction:-

Technological advances increasingly drive the use of digital tools in human activities. Today, Information and Communication Technologies (ICTs) are part of the daily lives of people and entities, through smartphones and their various applications, automated home appliances, cars with integrated systems, Industry 4.0, among others. The increased use of ICTs consequently generates a greater volume of stored data, which often exceeds the human capacity to analyze the data manually in an optimized way and presents the need for the use of techniques and tools that enable automatic analysis. In the educational context, the ability to collect, process and analyze large volumes of data has become a fundamental element over time. Educational institutions generate a significant amount of data related to students' academic performance on a daily basis, such as grades, attendance, participation in extracurricular activities, interaction in virtual environments, whether for distance or in-person education purposes, development of scientific initiation, among others. These data have the potential to, after being analyzed, generate information that helps improve education, allowing a deeper understanding of the needs and challenges faced by students and/or the institution (Sharif; Atif, 2024) (Romero; Ventura, 2020) (Campbell; DeBlois; Oblinger, 2007) (Barbosa; Silva, 2010).

The academic analysis of university students is a necessary field of study for improving educational quality. Understanding student performance, identifying behavior patterns and detecting difficulties in advance are some steps that can help in the implementation of effective pedagogical interventions. The ability to analyze educational data can help institutions identify students with low academic performance and who are at risk of failing or even dropping out, enabling timely and personalized interventions (Wong, 2016) (Wang; Luo, 2024). When observing only the complexity of the factors that lead to university dropout, the analysis of educational data presents itself as an interesting approach. Identifying patterns that indicate risks of dropping out can allow more effective interventions by managers and educators. However, for these professionals to be able to carry out such analyses, it is necessary that the tools used are intuitive and do not require advanced technical knowledge. Tools that facilitate data visualization and knowledge extraction in an accessible way can transform the capacity of educational institutions to the point of proactively responding to student needs (David; Chaym, 2019) (Ersozlu; Taheri; Koch, 2024).

Despite the advantages that can be obtained by applying academic analysis techniques to a set of educational data, many higher education institutions still face challenges in collecting and interpreting the large volumes of data generated, making it difficult to make informed decisions. Most of the tools available for data analysis require specific technical knowledge, which institutions sometimes do not have. A complete data analysis requires tools, methods and/or techniques that perform: data extraction from different sources and the interpretation of different types of data (e.g., structured, semi-structured and unstructured); data transformation so that they can be integrated into the central analysis system; data storage; data processing and analysis, applying algorithms that detect patterns and/or details relevant to the process; and data visualization in a way that users without ICT expertise can read and interpret the results of the analysis. To this end, in some cases, it is necessary to use more than one option of tool, method and/or technique, interconnecting them and making them work together in order to generate the expected results. This technical barrier can prevent educational managers and educators, who have in-depth knowledge of the educational context but are not necessarily proficient in ICT, from fully utilizing the available data to make informed decisions. Furthermore, continuing education for educators and managers in the use of these tools requires time and resources that are not always available (Zielinski; Schmitt, 2015) (Wong, 2016) (Rambabu; Althati; Selvaraj, 2023)

To address these challenges, an analytical data architecture called “EdmLens DataFlow” is proposed, which is not limited to a simple technical structure, but aims to provide support for the construction of complete platforms whose purpose is to assist educational managers, teachers and other professionals involved in the teaching-learning process in an efficient and understandable manner. To this end, the design of this architecture must integrate different stages of the Knowledge Discovery in Databases (KDD) process, allowing them to be adapted according to the needs of users, with a greater emphasis on ease of use and access to results. However, in order to develop “EdmLens DataFlow”, there is a need to survey bibliographic references, so that they can support future decisions regarding the choice of algorithms, methods, tools and other elements that may become important in the development process of this architecture. For this reason, this work is focused on carrying out a bibliographic review and obtaining a solid foundation on the concepts of data science, KDD, academic analysis and data analytics architecture. Thus, it is expected that the result of this work will contribute to a relevant theoretical basis regarding the construction of the architecture to be subsequently implemented using the methodologies and techniques that best fit its proposal.

Based on the results of the review, it has been possible to define some elements that will be part of the architecture when it is implemented. Data extraction should occur from various academic sources, such as school management systems, e-learning platforms and attendance records. At this stage, it should be possible to extract structured, semi-structured and unstructured data. The transformation will include data cleaning, normalization and enrichment, ensuring that all data is in structured data format. The data will be loaded into a Data Warehouse structure, as the architecture will perform its analyses using structured data. Data analysis should support Artificial Intelligence and Machine Learning techniques and algorithms, while data visualization will work with interactive reports and dashboards.

Finally, the results of the literature review allow us to conclude that there are still gaps in the area of educational data science to be explored in academic research, such as data integrity, scalability, data privacy and security, and cultural change. Thus, the "EdmLens DataFlow" architecture can be developed with strategies that mitigate these problems, especially in the cultural change of institutions, since the proposal is based on the premise that a well-structured architecture can significantly improve the ability of education professionals to analyze large volumes of educational data. By facilitating the visualization and interpretation of data, the architecture can help identify patterns of academic performance, enabling the implementation of personalized interventions to support students.

Theoretical Framework:-

The field of Data Science has been marked by continuous evolution, from traditional statistics to the integration of computing techniques and analysis of large volumes of data. Its evolution has been aided by technological and methodological advances that allow the transformation of raw data into tools to support decision-making (Tukey, 1962) (Chen; Mao; Liu, 2014).

In recent years, Data Science has become an important area in many fields, such as the stock market, industries, academic productions, among others, driven mainly by the exponential growth of digital data and advances in data storage and processing technologies (Jagadish et al., 2014). Concepts such as "Big Data" have emerged to describe the vast amount of data generated by companies and organizations (Mayer-Schönberger, 2013). Machine learning and artificial intelligence tools have been fundamental for carrying out predictive and prescriptive analyses, allowing the creation of models that can predict future behaviors and recommend actions (Carvalho; Menezes; Bonidia, 2024).

Data can be classified as structured, semi-structured or unstructured. According to Provost (2013), this classification can be defined as follows:

- Structured data:

These are data that follow a predefined format and are organized systematically, usually in tables with data distributed using rows and columns. They are stored in relational databases, which use a specific scheme to define the relationships between different sets of data. In addition, the storage of these data has well-defined rules for each field where the data can be allocated, specifying, for example, the type of data, such as integer, string, date, among others. The organization of structured data has important characteristics for database applications, as it facilitates the tasks of performing queries on them, helps with data integrity and consistency, and allows data validation according to the pre-established scheme. MySQL, PostgreSQL, and Oracle databases are examples of approaches for storing structured data. However, it is possible to have structured data using other means, such as spreadsheets. Example: Suppose an information system developed using Web technologies (e.g., HTML, CSS, JavaScript and PHP) and a relational database (e.g., MySQL) to record student grades and attendance, with a database of 1000 students and 50 different courses. This system must have a set of personal data (e.g., name, date of birth, e-mail and cell phone) and a set of academic data of the students (e.g., courses they are taking, assessments they have already taken in each course, grades from the assessments taken and attendance in the courses). However, all this data is organized and related in a way that it is possible to consult the grades of a given student without conflicts with another student present in the same system.

- Unstructured data:

Does not follow a predefined format and can be represented in different formats, without a fixed organization or formatting. They are not stored in a specific schema with a uniform structure, which makes them more flexible, but also more difficult to manage and analyze, and may require the use of specific techniques for data extraction and analysis, such as natural language processing (NLP) and image recognition. Typical examples of unstructured data include multimedia content (e.g., images, audios, and videos), free text documents (e.g., emails and PDFs), and

social media data (e.g., posts, comments, and interactions). It is common for this data to be stored in cloud storage services, non-relational databases, or file systems. Because they have a variable format, they have some difficulties, but they also contribute to easier scalability - they do not have rigid schemas. Example: Suppose a data analyst is checking the opinion of students of a certain distance learning course regarding their satisfaction with the courses taught. To do so, this analyst collected the comments made in the discussion forums of all courses and applied a sentiment analysis algorithm. This algorithm will return whether the students had a positive, negative or neutral feeling about the courses.

- Semi-structured data:
Combines characteristics of structured and unstructured data. Although they have an internal structure that facilitates interpretation, including the creation of complex data structures with hierarchies and lists, this structure is not as rigid as that of structured data. They are stored in formats that allow flexibility, such as JSON (JavaScript Object Notation) and XML (eXtensible Markup Language), making them ideal for exchanging data between different systems. The structures that store semi-structured data can contain metadata within the document itself, which helps in interpreting the data. Example: Suppose an application that will analyze the academic performance of students at a certain educational institution to verify which subjects have the highest failure rates. To do this, the institution in question needs to forward some specific data from its database so that the application can perform such analysis. The person responsible for the application indicated that he needs the following data: name of the subject, enrolled students, completed assessments and grades and weights of these assessments. The institution produces a JSON file with all the requested data organized minimally so that the application can read and interpret it.

Working with different databases that have distinct structures and varying values requires techniques and methods that enable the integration and extraction of this data in an organized manner. For this reason, Data Mining has become an increasingly sought-after and important area of study, both for the academic and scientific environment, as well as for the business and government environment (GOLDSCHMIDT, 2015). In short, the area of Data Mining involves a process that begins with the identification of the data to be analyzed, goes through phases of selection and organization of this data and ends with the visualization of the "mined" data. According to AGGARWAL (2015), Data Mining's main objective is to extract useful information from a large amount of solid data.
According to the author FILATRO (2020), Data Mining is a combination of different fields of Computer Science that combine with statistical analysis. Therefore, although studies related to the field of Data Mining began in the mid-1980s, there is still a range of possibilities involving the techniques and methods of this area. The purpose of this research usually involves finding the databases that have the information necessary for the analysis and/or solution of a given problem, organizing the data in order to keep only what is necessary and processing this data in order to assist in decision-making. "For this, several techniques are used, such as classification, regression, combination by similarity, profiling, data reduction and causal modeling, among others." (FILATRO, 2020, 2, p. 31).

With the advancement of technologies over the last few years, Data Mining techniques have begun to be used in a variety of contexts, such as the Education field. Information and Communication Technologies have begun to play very important roles in academic environments - from applications to assist in the teaching-learning process to information systems to manage student data throughout their academic life. The high flow of data generated by technologies applied to education has provided a favorable situation for encouraging work with Educational Data Mining (EDM), an area focused solely on academic data (SOUZA, 2021).
Educational Data Mining is an emerging discipline concerned with developing methods for exploring the unique and increasingly large-scale data coming from educational settings and using these methods to better understand students and the environments in which they learn. (EDM, 2024, p. 1)

Knowledge Discovery in Databases – KDD
The digital age has witnessed an unprecedented explosion in the volume of data generated daily. This increase in data production, fueled by technological advances and the digitalization of processes in virtually all sectors of society, presents both challenges and opportunities for computer science and related fields. In response to this demand, Knowledge Discovery in Databases (KDD) emerges as an important approach to transform the vast amount of raw data into relevant and useful information for decision-making in different situations. KDD is an iterative process that focuses on identifying patterns in a given data set. This process involves several steps, including data selection, data cleaning and preparation, the application of data mining techniques to discover meaningful patterns, and the interpretation and evaluation of these patterns to ensure that they are applicable for decision-making

(Fayyad; Piatetsky-Shapiro; Smyth, 1996) (GOLDSCHMIDT, 2015). "Knowledge Discovery in Database (KDD) is an interdisciplinary area focusing on methodologies or techniques to extract useful data from sources" (Sabri et al., 2019, p. 418).

The "EdmLens DataFlow" architecture aims to provide the structure and organization of the entire Knowledge Discovery process in Databases, including the phases of extraction, transformation, loading (ETL), analysis and visualization of educational data. It defines the components, interactions and processes that need to be performed to generate analytical results, but without prescribing specific implementations. The architecture is the basis for the creation of data pipelines, analysis models and visualization flows. KDD presents possibilities for the process to be defined and structured based on the "EdmLens DataFlow" architecture. Thus, the chapter on Knowledge Discovery in Databases presents the steps of the KDD process, with an emphasis on ETL, different points of view to implement the process and ways to automate it. Also in this chapter there is a study on data pipelines in order to find ways to make the architecture more flexible and efficient, allowing the automation of the KDD process while offering space for advanced customizations of analysis and integration with external data sources. Integration with external tools together with deep analysis within a structured pipeline is a robust combination that ensures scalability and adaptability as a given system develops from architecture.
In a Knowledge Discovery in Databases flow, it is possible to highlight some specific areas regarding data processing, such as data mining and data visualization. Another field of study that is part of KDD and that can be highlighted is ETL (Extract, Transform and Load), a data integration process used to combine information from different sources into a single repository, such as a Data Warehouse. According to (Vida et al., 2021), this process comprises three distinct and interdependent steps that can be defined as follows:

- Extract:
In this step, it is necessary to identify the sources of information and export them to an intermediate structure, from which the data can be processed and prepared. Examples of sources from which data can be extracted: SQL Servers, cloud environments, text documents, spreadsheets, emails, among others.

- Transform:
The extracted data needs to be stored in a specific structure. However, this data is usually in different formats and comes from different structures. To do so, in this step, it is necessary to transform the data so that it respects the same format and the same structure. Examples of actions that can occur in this step: data cleaning, removing duplicate or inconsistent data, data authentication, data translation, data formatting, among others. The author mentions that this is the most important step in the ETL process.

- Load:
In order for the data to be used for analysis or report generation purposes, it needs to be loaded into a specific structure, such as a Data Warehouse. This step is responsible for transporting the data from the intermediate structure, in which the data transformation occurred, to the final structure. This loading can be done all at once, known as a complete load, or partially, known as an incremental load. In a complete load, all data from the transformation phase is sent. In an incremental load, it is possible to filter and load only the data that meets the new parameters.

Academic Analysis
The university environment is characterized by a series of challenges and phenomena that directly influence the path of students throughout their academic education. Understanding the concepts related to student academic performance is essential for the development of effective strategies aimed at improving the quality of teaching and the student experience. Different academic works and government regulations provide definitions for the main terms related to academic analysis in the university context.
For this reason, these concepts will be addressed and discussed below according to the focus of this work. University dropout is a complex phenomenon that refers to the abandonment of the course by the student before its completion, without transferring to another institution or course. This problem can be influenced by a variety of factors, such as financial difficulties, which can prevent students from bearing the costs of their education, whether tuition and/or housing, travel or food, the lack of social and academic integration, which can lead to isolation and demotivation, and failures, which can also demotivate students. Furthermore, dissatisfaction with the course or institution can result in a lack of engagement, causing students to question the relevance and quality of their

academic choice. Personal or health issues also play a significant role, affecting students' ability to maintain consistent academic performance (Filho et al., 2007) (Santos, 2022) (Prestes; Fialho, 2018).

According to Prestes and Fialho (2018), university dropout rates not only negatively impact students, but also generate significant social, economic and academic waste. In public institutions, dropout rates result in an inefficient use of public resources allocated to higher education, while in private institutions, they can affect the financial sustainability and reputation of the institution. From a social perspective, dropout rates contribute to the reduction of qualified human capital available in the labor market, limiting the economic and social development of a given region or country. The analysis of the factors that lead to dropout rates is therefore crucial for the development of effective programs to combat dropout rates. This may include implementing financial support policies, such as scholarships and grants, to alleviate economic pressures on students, mentoring and counseling programs to help improve social and academic integration, providing emotional support and academic guidance, reviewing and adapting curricula to make them more aligned with students' expectations and interests to reduce dissatisfaction with the course, providing mental health services and personal support to help address personal issues that interfere with academic success, among others.

By addressing college dropout in a comprehensive manner, educational institutions can not only improve retention and course completion rates, but also contribute to the formation of a more satisfied and successful student body, which in turn benefits society as a whole. For this reason, several studies have sought to find methodologies to help reduce dropout rates in college courses. The work by Santos (2022), for example, explores the context of college dropout by proposing the use of machine learning techniques to personalize learning. The objective of the research is to use these techniques to identify students at risk of dropping out and propose specific personalized interventions to course managers, such as counseling and/or academic support.

One of the factors that can influence college dropout rates is failures throughout the course. Failure occurs when a student does not achieve the minimum grade required to pass a subject, which can have several negative consequences for the student. Failure rates are often used by educational institutions as indicators to identify areas that require pedagogical and curricular improvements. High failure rates can signal problems in teaching methodology, in the adequacy of the syllabus to the needs of students, or in the assessment structure itself. Failure can significantly delay a student's progression in the course, extending the time needed for completion and increasing the overall cost of education. This delay can demotivate students, leading them to question their academic ability and their place in the institution. The psychological impact of failure can be severe, affecting students' self-esteem and motivation, which in turn can result in even worse academic performance in subsequent courses. This negative cycle can eventually increase the risk of dropout, especially if students feel that they are not receiving the necessary support to overcome their difficulties (Filho et al., 2007).

According to Filho et al. (2007), grade retention can be mitigated through specific pedagogical interventions. Tutoring programs, for example, can offer individualized support to students, helping them to better understand the content and develop more effective study strategies. Remedial classes can provide additional learning opportunities outside of regular class time, focusing on areas in which students have more difficulty. In addition, changes in teaching methodologies, such as adopting more interactive and student-centered approaches, can make learning more engaging and accessible. Another important aspect is the creation of a more welcoming and supportive learning environment. Institutions that promote open communication between teachers and students, and that offer resources such as academic and psychological counseling, can help students overcome the barriers that lead to grade retention. Analysis of academic data can also be a powerful tool for identifying patterns of grade retention and developing specific interventions. For example, the use of early warning systems can identify at-risk students before they fail, allowing preventive measures to be implemented.

Data Governance

The term governance refers to the union of practices, policies, principles and strategies with a view to assisting in the management of a given organizational structure. Data governance seeks to bring together these items in the context of Data Science, seeking to ensure the quality, security, availability and usability of data (RÊGO, 2013).

The relevance of data governance to the Data Science area can be translated from governance principles. These principles typically vary from project to project, but some of them provide the basis for the development and implementation of effective policies and practices, such as (Malik, 2013) (RÊGO, 2013) (Abraham; Schneider; vom Brocke, 2019): Accountability, Transparency, Data Quality, Security, Privacy and Data Lifecycle Management.

To illustrate the application of data governance, let us assume a context of analyzing educational data to verify the academic performance of university students. Initially, responsibilities are organized, so that each department can be asked to designate a person responsible for the data involving the academic performance of students in that department. Thus, a data governance committee is established that includes representatives from different departments. This committee will oversee the implementation of data governance policies and resolve issues related to data quality and security. Then, work can be done on the transparency aspect by producing and publishing policies for data collection, storage and analysis to all stakeholders. To make the process even more transparent, it is possible to create regular reports to academic managers and/or other professionals about the state of the data and any problems found. For example, a quarterly report can detail data quality metrics and corrective actions taken. Based on the principle of data quality, standards can be established for data entry, such as specific formats for grade and attendance records, ensuring consistency and completeness. For example, all grades are recorded to two decimal places and attendance is marked as "Present" or "Absent". In addition, in order to maintain data quality, automatic and/or manual processes to identify and correct errors in the data can be implemented (e.g., scripts to check for duplicate records). For data security, access to students' academic data could be restricted to teachers and administrators only, and sensitive data could be encrypted, both in transit and while stored. For example, when sending grades from an academic management system to a central repository, SSL encryption is used to protect the data during transfer. The principle of privacy could be explored by providing information to students about how their data will be used and seeking consent for such collection and analysis. To this end, students could be presented with a consent form with the necessary explanations so that they could sign it during the period of enrollment in their course. Furthermore, actions regarding data must be aligned with privacy regulations, such as the LGPD (Brasil, 2018), and allow students to request the deletion of their personal data from the system at any time. Finally, when dealing with data lifecycle management, it is possible to organize the archiving of data after the end of each semester and its maintenance for a specific period (e.g., 5 years) before being discarded.

Data governance can help in the context of Data Science by increasing the quality of data understanding, in order to ensure that the analyses applied to the data meet all the determined requirements, using the data in an ethical manner, with methods to prevent misuse of data and protect the privacy of individuals, improving data security, a critical aspect in data science, especially for data with sensitive information, and generating a data operation integrated with management and the team, with knowledge about the process shared among the entire execution team (RÊGO, 2013) (Abraham; Schneider; vom Brocke, 2019).

EdmLens DataFlow Architecture:-

In universities, the use of data to assess and improve students' academic performance is increasingly necessary, especially due to the increase in the volume of information available on digital platforms, academic management systems and extracurricular activities. This information, when organized in an analytical architecture, allows educators and managers to identify performance patterns, predict dropout risks and promote educational interventions efficiently.

The "EdmLens DataFlow" architecture is the core of this paper's proposal, being a system designed to manage the KDD process. It describes and orchestrates how data will be extracted, transformed, loaded, analyzed and visualized to generate useful information in the educational context. The architecture is not limited to a simple technical framework, but aims to provide a complete platform for educational managers, teachers and others involved in accessing information in an efficient and understandable way. The design of this architecture integrates different stages of the KDD process and allows them to be adapted according to the needs of the users, with a special emphasis on ease of use and accessibility of the results. This allows for in-depth analysis and a broader view of educational data, connecting the different parts of the system in a cohesive and efficient way. As a complete framework, the "EdmLens DataFlow" needs to encompass the stages of data extraction, storage, processing, analysis and visualization, as well as the implementation of privacy policies and data access. Each of these phases requires specific technologies, methods and practices that, together, ensure the integrity and accessibility of the information. Furthermore, the application of advanced techniques such as machine learning, predictive modeling and interactive visualization allows this architecture to go beyond simple descriptive analysis, offering predictive and diagnostic capabilities that support proactive academic management. It is possible to note that the "EdmLens DataFlow" architecture will use an ETL process and not an ELT process. This choice is based on the advantages of using a Data Warehouse as a storage structure, mainly with regard to generating reports and fast queries. The "EdmLens DataFlow" has its main steps structured as follows:

- Data extraction:
This is the initial stage of the architecture and consists of capturing data from different sources and systems, ensuring that this information is in line with the objectives of the analysis to be performed. The data collected will come from the following sources: academic management systems, when available, which allow access to data on enrollment, grades, attendance, courses taken and performance in specific subjects; e-learning platforms, searching for records of online learning activities, such as time spent on each module, participation in forums, completion of activities and engagement in assessments; demographic and socioeconomic data, such as information on age, gender, family income and previous education of students, which can be collected via questionnaires or institutional databases; and records of extracurricular activities, such as participation in study groups, academic clubs, university events and complementary activities that may influence the academic performance of students. The extraction of this data will be done through APIs, applications that connect directly to the source systems and transfer them to a central database, CSV files, JSON files, text files, relational databases (e.g., MySQL, PostgreSQL, SQLServer and Oracle Database) and non-relational databases (e.g., MongoDB).

- Data transformation:
Is responsible for converting raw data extracted from different sources into organized and consistent information, ready for analysis. The main objective of data transformation is to ensure that all information is standardized and that inconsistencies are resolved before proceeding to data loading and storage. To this end, the following interventions will be carried out: data normalization, ensuring that all data follows the same structure, such as, for example, standardizing grade scales in the same unit, turning all grades into numbers with a maximum of two decimal places; data cleaning, removing null values, duplicates and correcting incorrect data, which could distort analytical results; and data integration and aggregation, in which data from multiple sources will be combined to create a more comprehensive and contextualized view of the object of analysis, allowing complementary information to be combined.

- Data loading and storage:
After data transformation, it will be loaded and stored in a single structure, a Data Warehouse. This work chose to perform data transformation before loading in order to work only with structured data and improve the queries needed for analysis and reports. Data loading will be done both in batch and in real time.

- Data processing and analysis:
This phase includes both descriptive and predictive analyses, which provide information on student performance and allow for the prediction of future behavior. To this end, algorithms and methods of artificial intelligence, natural language processing and machine learning will be used to generate information from the prepared data.

- Data visualization:
This is the final phase of the architecture, in which the information generated is presented in an understandable and accessible way. Therefore, visualization tools and interactive dashboards will be used to help educators and managers interpret the data, identify patterns and make informed decisions.

- Data governance:
Involves defining roles, responsibilities, policies and processes that ensure efficient data management throughout its lifecycle. For the “EdmLens DataFlow” architecture, the roles and responsibilities will be:
– Project manager: responsible for leading the data governance strategy and aligning it with the institution's strategic objectives (e.g., leader of the application of the architecture in the educational institution).
– Data managers: responsible for the day-to-day management of data, including security, privacy, and compliance. Each manager may be responsible for a type of data (e.g., academic data and personal data).
– Data administrators: responsible for the storage, backup, and secure access to data, ensuring that it is protected and accessible only to those with permission (e.g., technicians who already work in the IT sector at the institution).

– Data users: faculty, managers, and other end users who use the data for analysis and decision-making. They are responsible for following the guidelines for appropriate use of data.

The architecture policies will include:
– Access and security policy: establishes who can access the data and under what conditions, ensuring that sensitive information is protected.
– Data retention and disposal policy: determines the retention time for each type of data and the process for securely disposing of data that is no longer needed.
– Compliance and privacy policy: ensures that data handling complies with privacy regulations, such as the LGPD (General Data Protection Law), respecting the confidentiality of student information.

In addition, the architecture's data governance includes regular audits to monitor data use and security, in addition to maintaining a record of who accesses and changes the data, and establishes that all data is documented with descriptive metadata, supporting the understanding of its origin, format, quality and usage restrictions. By integrating data pipelines, frameworks, and APIs, the "EdmLens DataFlow" architecture will be more efficient and flexible, allowing for the automation of the KDD process while offering space for advanced customizations of analysis and integration with external data sources. To this end, this architecture was designed based on the bibliographic concepts raised in this research. The studies allowed us to analyze the type of data process, the data storage structure, the algorithms and data analysis techniques, among others, that best fit the research objectives. The demonstration presented above is only conceptual, but the intention is to implement this architecture.

Conclusion:-

The data analytics architecture "EdmLens DataFlow" was designed with the aim of facilitating the collection, transformation, loading, analysis and visualization of academic data, and allowing educational institutions to obtain information about student performance and engagement. Based on the literature review, "EdmLens DataFlow" can be designed with the best methods and technologies in mind.
The main differences of the architecture are its ability to consolidate data from heterogeneous sources — such as academic management systems, e-learning platforms and extracurricular activity records — transforming them into consistent and useful information. Implementing the ETL process ensures that data can be standardized and stored in a structured manner in a Data Warehouse, which facilitates the execution of complex queries and the creation of long-term analyses. Choosing a Data Warehouse for storage and using interactive visualization tools will allow teachers and managers to interpret the data without the need for specialized technical knowledge, enabling a more intuitive and informative decision-making environment. The development of "EdmLens DataFlow" also highlighted the importance of data governance protocols, which are essential to ensure privacy, security and compliance with data protection standards, such as LGPD. Defining roles and access policies, as well as recording metadata and regular audits, reinforces responsibility in the use of data, providing a reliable and transparent basis for educational analysis.

However, the study presents some limitations that open space for future improvements. The development of "EdmLens DataFlow" only in the theoretical field limits the validation and understanding of the scope of this architecture. Thus, the work that can be carried out in the future is the implementation of a functional prototype of this architecture and the application of tests in different institutions for its validation. In addition, the development of specific modules for collaborative analysis and mobile interfaces could further expand the scope and accessibility of "EdmLens DataFlow".

In summary, the data analytics architecture called "EdmLens DataFlow" represents an important contribution to the field of educational data science, standing out as a practical and innovative tool that supports the analysis and management of academic data. By facilitating access to strategic information and enabling decisions to be made based on concrete data, the architecture has the potential to positively impact the educational system, promoting a more personalized, inclusive and results-oriented education.

References:-

1. ABRAHAM, R.; SCHNEIDER, J.; vom Brocke, J. Data governance: A conceptual framework, structured review, and research agenda. International Journal of Information Management, vol. 49, p. 424–438, 2019. ISSN 0268-4012. Available at: <https://www.sciencedirect.com/science/article/pii/S0268401219300787>.
2. AGGARWAL, C. C. Data mining: the textbook. [S.l.]: Springer, 2015.
3. BARBOSA, S.; SILVA, B. Interação humano-computador. [S.l.]: Elsevier Brasil, 2010.

4. BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Diário Oficial [da] República Federativa do Brasil, Brasília, DF, 2018. Available at: < https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm>.
5. CAMPBELL, J. P.; DEBLOIS, P. B.; OBLINGER, D. G. Academic analytics: A new tool for a new era. EDUCAUSE review, Educause, v. 42, n. 4, p. 40, 2007.
6. CARVALHO, A. C. P. L. F. d.; MENEZES, A. G.; BONIDIA, R. P. Ciência de Dados - Fundamentos e Aplicações. Rio de Janeiro: Grupo GEN, 2024. Available at: https://integrada.minhabiblioteca.com.br//books/9788521638766/. Acesso em: 18 set. 2024. ISBN 9788521638766.
7. CHEN, M.; MAO, S.; LIU, Y. Big data: A survey. Mobile networks and applications, Springer, v. 19, p. 171–209, 2014.
8. DAVID, L.; CHAYM, C. Evasão universitária: Um modelo para diagnóstico e gerenciamento de instituições de ensino superior. Revista de Administração IMED, v. 9, n. 1, p. 167–186, 2019. ISSN 2237-7956. Available at: <https://seer.atitus.edu.br/index.php/raimed/article/view/3198>.
9. EDM. Educacional Data Mining. 2024. Acesso em 29/09/2024. Available at: <https://educationaldatamining.org/>.
10. FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. et al. Knowledge discovery and data mining: Towards a unifying framework. In: KDD. [S.l.: s.n.], 1996. v. 96, p. 82–88.
11. FILATRO, A. C. Data science da educação. [S.l.]: Editora Saraiva, 2020. ISBN 9786587958446.
12. FILHO, R. L. L. S.; MOTEJUNAS, P. R.; HIPÓLITO, O.; LOBO, M. B. d. C. M. A evasão no ensino superior brasileiro. Cadernos de pesquisa, SciELO Brasil, v. 37, p. 641–659, 2007.
13. GOLDSCHMIDT, R. Data Mining. [S.l.]: Grupo GEN, 2015. ISBN 9788595156395.
14. MALIK, P. Governing big data: Principles and practices. IBM Journal of Research and Development, v. 57, n. 3/4, p. 1:1–1:13, 2013.
15. PRESTES, E. M. d. T.; FIALHO, M. G. D. Evasão na educação superior e gestão institucional: o caso da universidade federal da paraíba. Ensaio: Avaliação e Políticas Públicas em Educação, SciELO Brasil, v. 26, p. 869–889, 2018.
16. PROVOST, F. Data Science for Business: What you need to know about data mining and data-analytic thinking. [S.l.]: O'Reilly Media, Inc, 2013. v. 355.
17. RAMBABU, V. P.; ALTHATI, C.; SELVARAJ, A. Etl vs. elt: Optimizing data integration for retail and insurance analytics. Journal of Computational Intelligence and Robotics, v. 3, n. 1, p. 37–84, 2023.
18. ROMERO, C.; VENTURA, S. Educational data mining and learning analytics: An updated survey. Wiley interdisciplinary reviews: Data mining and knowledge discovery, Wiley Online Library, v. 10, n. 3, p. e1355, 2020.
19. RÊGO, B. L. Gestão e governança de dados: promovendo dados como ativo de valor nas empresas. Rio de Janeiro: Brasport, 2013. Available at: https://plataforma.bvirtual.com.br. Acesso em: 21 set. 2024. ISBN 9788574526294.
20. SABRI, I. A. A.; MAN, M.; BAKAR, W. A. W. A.; ROSE, A. N. M. Web data extraction approach for deep web using weidj. Procedia Computer Science, v. 163, p. 417–426, 2019. ISSN 1877-0509. 16th Learning and Technology Conference 2019Artificial Intelligence and Machine Learning: Embedding the Intelligence. Available at: <https://www.sciencedirect.com/science/article/pii/S1877050919321635>.
21. SANTOS, R. S. S. d. Evasão Escolar Universitária e Estratégias de Intervenções para Retenção do Estudante: Um Estudo de Caso na Universidade Federal de São Carlos. Tese (Doutorado) — Universidade de São Paulo, 2022.
22. SHARIF, H.; ATIF, A. The evolving classroom: How learning analytics is shaping the future of education and feedback mechanisms. Education Sciences, MDPI, v. 14, n. 2, p. 176, 2024.
23. SOUZA, V. F. d. Os avanços da mineração de dados educacionais : processo, tendências temáticas e técnicas de mineração. [S.l.]: Editora Bagai., 2021. ISBN 9786586734836.
24. TUKEY, J. W. The future of data analysis. In: Breakthroughs in Statistics: Methodology and Distribution. [S.l.]: Springer, 1962. p. 408–452.
25. VIDA, E. D. S.; ALVES, N. S. R.; FERREIRA, R. G. C.; SOUZA, D. C. D.; BARBOZA, F. F. M.; OLIVEIRA, H. S. D.; MARQUE, L. D.; MASCHIETTO, L. G.; GONÇALVES, P. D. F. Data warehouse. [S.l.]: Grupo A, 2021.
26. WANG, S.; LUO, B. Academic achievement prediction in higher education through interpretable modeling. Plos one, Public Library of Science San Francisco, CA USA, v. 19, n. 9, p. e0309838, 2024.
27. WONG, Y. Y. Academic analytics: a meta-analysis of its applications in higher education. International Journal of Services and Standards, Inderscience Publishers (IEL), v. 11, n. 2, p. 176–192, 2016.