UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
CENTRO DE CIÊNCIAS EXATAS, NATURAIS E DA SAÚDE
DEPARTAMENTO DE COMPUTAÇÃO

ELOY DE FREITAS ALMEIDA

IMPLEMENTAÇÃO DE UM DATA MART PARA ESTUDO DE EVASÃO ESTUDANTIL NA UFES *CAMPUS* ALEGRE-ES

ALEGRE - ES
2023

IMPLEMENTAÇÃO DE UM DATA MART PARA ESTUDO DE EVASÃO ESTUDANTIL NA UFES *CAMPUS* ALEGRE-ES

Trabalho de conclusão de curso apresentado ao Departamento de Computação do Centro de Ciências Exatas, Naturais e da Saúde da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.

por

ELOY DE FREITAS ALMEIDA

Orientador
Antonio Almeida de Barros Junior

Universidade Federal do Espírito Santo

ALEGRE - ES

2023

ELOY DE FREITAS ALMEIDA

IMPLEMENTAÇÃO DE UM DATA MART PARA ESTUDO DE EVASÃO ESTUDANTIL NA UFES *CAMPUS* ALEGRE-ES

Trabalho de conclusão de curso apresentado ao Departamento de Computação do Centro de Ciências Exatas, Naturais e da Saúde da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.

Aprovado em 14 de dezembro de 2023.

COMISSÃO EXAMINADORA

Documento assinado digitalmente
ANTONIO ALMEIDA DE BARROS JUNIOR
Data: 22/12/2023 10:13:48-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Antonio Almeida de Barros Junior
Universidade Federal do Espírito Santo
Orientador

Documento assinado digitalmente
MARCELO OTONE AGUIAR
Data: 22/12/2023 11:43:09-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Marcelo Otone Aguiar
Universidade Federal do Espírito Santo

Documento assinado digitalmente
RODRIGO FREITAS SILVA
Data: 22/12/2023 10:27:33-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Rodrigo Freitas Silva
Universidade Federal do Espírito Santo

AGRADECIMENTOS

Agradeço aos meus pais Juarez e Vilma, pelo carinho, educação, confiança e amor para construção da minha vida. E a minha familia por terem me apoiado nessa jornada.

Ao meu amor Larissa Bebber, por todo apoio, carinho, amizade e companheirismo, por estar comigo em todos os momentos.

Ao meu orientador Antônio, pelos ensinamentos e suporte necessário durante toda a graduação e na composição deste trabalho.

Aos meus professores pelos ensinamentos.

Aos meus amigos, pelo apoio e companheirismo. Em especial Matheus Bastos, Fabricio Medeiros, Lucas Sobral, Davi Carvalho, Cleiton Gaioti e Luann Laurindo pela amizade no processo de graduação.

À Universidade Federal do Espírito Santo, pela oportunidade.

"Já que o rei não vai virar humilde, eu vou fazer o humilde virar rei."

Emicida

SUMÁRIO

1 Introdução **11**

1.1 Objetivos 12

1.1.1 Objetivo Geral 12

1.1.2 Objetivos específicos 12

2 Revisão de Literatura **13**

2.1 *Data Warehouse* 13

2.1.1 Arquiteturas propostas de um *Data Warehouse* 14

2.1.2 Modelagem Dimensional 16

2.2 Trabalhos relacionados 17

2.2.1 Trabalho de SALES JUNIOR (2013) 18

2.2.2 Trabalho de Mendes (2020) 20

2.2.3 Trabalho de Gonçalves (2021) 22

2.3 Ferramentas Utilizadas 24

2.3.1 Linguagem de programação Python 24

2.3.2 Apache Airflow 25

2.3.3 PostgreSQL 25

2.3.4 MIcrosoft Power BI 26

2.3.5 Docker 26

3 Metodologia **28**

3.1 Levantamento dos Indicadores 29

3.2 Base de Dados de Origem 30

3.3 Análise dos dados 31
3.4 Modelo Dimensional 32
3.5 Estrutura do Projeto e Fluxo de ETL 34
3.5.1 Diretório *airflow* 36
3.5.2 Diretório docs 36
3.5.3 Diretório *reports* 36
3.5.4 Diretório src 36
3.5.5 Diretório scripts 37
3.5.6 Diretório dags 37

4 Resultados Obtidos 39
4.1 Indicadores Gerais 40
4.2 Indicadores Acadêmicos 41
4.3 Indicadores Sociais 44
4.4 Indicadores Regionais 46

5 Conclusões 49

6 Trabalhos Futuros 50

Referências 51

Apêndice A – *Script* da *view* tratamento das *stages* 53

Apêndice B – *Script* da *view* do *datasource* 57

Apêndice C – Colunas da *stage* ”stg_relatorio” 59

Apêndice D – Dicionário de dados 62

Apêndice E – Tutorial de utilização 66

E.1 Requisitos de sistemas 66
E.2 Configuração de ambiente 66
E.2.1 Instalação do Sistema 66
E.3 Execução 67
E.3.1 Atualização do Painel 69
E.4 Interrupção de serviços 72
E.5 Desinstalando programa 72

LISTA DE FIGURAS

Figura 1 Arquitetura do DW de Kimball (2013).
Fonte: O Autor. 14

Figura 2 Arquitetura do DW de Inmon (2005).
Fonte: O Autor. 15

Figura 3 Exemplo do *Star Schema*.
Fonte: Adaptado de Kimball (2013). 16

Figura 4 Exemplo Modelo dimensional.
Fonte: Adaptado de Kimball (2013). 17

Figura 5 Distribuição do tempo até a evasão.
Fonte: (SALES JUNIOR, 2013). 19

Figura 6 Modelo dimensional proposto por Mendes (2020).
Fonte: Adaptado de Mendes (2020). 21

Figura 7 Gráfico da Evolução das Taxas Anuais de Evasão e Diplomação da Unidade de Observação “A”.
Fonte: Mendes (2020). 22

Figura 8 Modelo dimensional proposto por Gonçalves (2021).
Fonte: Adaptado de Gonçalves (2021). 23

Figura 9 Demonstração dos estudantes retidos nos ciclos.

Fonte: Gonçalves (2021). 24

Figura 10 Exemplo de uma DAG.
Fonte: O Autor. 25

Figura 11 Comparação de um sistema utilizando Docker e um sistema utilizando máquinas virtuais.
Fonte: O Autor. 27

Figura 12 Arquitetura do DW.
Fonte: O Autor. 28

Figura 13 *Schema Staging Area*.
Fonte: O Autor. 32

Figura 14 Modelo dimensional.
Fonte: O Autor. 33

Figura 15 *Schema* do DW.
Fonte: O Autor. 34

Figura 16 Estrutura do projeto.
Fonte: O Autor. 35

Figura 17 DAG Evasão.
Fonte: O Autor. 38

Figura 18 Página Indicadores Gerais.
Fonte: O Autor. 40

Figura 19 Gráfico Evasão por Período Cursados desconsiderando os alunos diplomados.

Fonte: O Autor. 41

Figura 20 Página Indicadores Acadêmicos.
Fonte: O Autor. 42

Figura 21 Gráfico Composição da Evasão.
Fonte: O Autor. 43

Figura 22 Gráfico Evasão por Tipo de Cota .
Fonte: O Autor. 43

Figura 23 Evasão por Forma de Ingresso.
Fonte: O Autor. 44

Figura 24 Página Indicadores Sociais.
Fonte: O Autor. 45

Figura 25 Gráfico Evasão por Etnia.
Fonte: O Autor. 45

Figura 26 Gráfico Evasão por Deficiência.
Fonte: O Autor. 46

Figura 27 Página Indicadores Regionais.
Fonte: O Autor. 47

Figura 28 Gráficos de Evasão por Municípo de Naturalidade e Evasão por Município de Origem .
Fonte: O Autor. 48

Figura E.1 Página inicial do Apache Airflow.
Fonte: O Autor. 68

Figura E.2 Página inicial do Apache Airflow execução da DAG.
Fonte: O Autor. 69

Figura E.3 Painel de detalhes da DAG.
Fonte: O Autor. 69

Figura E.4 Painel de indicadores gerais vazio.
Fonte: O Autor. 70

Figura E.5 Formulário de atualização de conexões.
Fonte: O Autor. 71

Figura E.6 Painel de indicadores gerais carregado.
Fonte: O Autor. 72

LISTA DE TABELAS

Tabela 1 Tebela com as quantidades de Evadidos, Diplomados e Matriculados. Fonte: Adaptado de (SALES JUNIOR, 2013). 18

Tabela 2 Tabela de frequência da forma de evasão e tempo até a evasão. Fonte: Adaptado de (SALES JUNIOR, 2013). 19

Tabela 3 Formas de evasão consideradas no SIE. Fonte: O Autor. 31

RESUMO

Este estudo aborda a evasão estudantil na UFES Campus Alegre, destacando a necessidade de ferramentas especializadas para compreender e enfrentar esse problema específico da região. Foi desenvolvido um ambiente de *Business Intelligence* BI, inspirado no modelo de *Ralph Kimball*, alimentado por um processo automatizado de Extração, Transformação e Carga (ETL) com dados dos relatórios de matrícula dos alunos dos centros CCENS e CCAE entre 2009/01 e 2023/01. Esse ambiente incluiu *dashboards* interativos para melhorar a análise e tomada de decisões. Os resultados indicam 9599 ingressantes, 4216 evadidos e 2638 diplomados, com taxas de evasão de 43,92% e aprovação de 27,48% na UFES.

Palavras-chave:*Business Intelligence*, Evasão no Ensino Superior, *Data Warehouse*.

1 INTRODUÇÃO

A evasão de estudantes é um fenômeno complexo, comum às instituições universitárias no mundo contemporâneo. Exatamente por isto, sua complexidade e abrangência vêm sendo, nos últimos anos, objeto de estudos e análises, especialmente nos países do Primeiro Mundo (ANDIFES, 1996). No Brasil, na rede pública apenas, 204.174 pessoas conseguiram concluir a graduação no ano de 2020, número 18,8% inferior ao ano de 2019 com 251.374 graduados (PALHARES, 2022).

Para a Associação Nacional dos Dirigentes das Instituições Federais de Ensino Superior – ANDIFES (1996), capacitar os discentes de forma qualitativa e alcançar bons desempenhos de diplomados, gerados a cada ciclo acadêmico para o exercício profissional são uma das maiores preocupações de uma IES. Para tanto, compreender os índices de diplomação, retenção e evasão é de extrema importância para uma IES, pois permitem analisar seu desempenho na gestão acadêmica, identificar possíveis problemas e adotar medidas pedagógicas e institucionais capazes de solucioná-los.

Para que a IES possa acompanhar seus indicadores de desempenho, faz-se necessária a implantação de sistemas capazes de auxiliar na integração e representação dos dados. De acordo com Kimball (2013), a construção de sistemas de *Business Intelligence* BI se tornou tendência nas últimas décadas, devido ao seu propósito de apresentar informações aos usuários de negócio de forma clara e intuitiva. Nessa proporção, a utilização de ferramentas de BI facilita o acesso aos dados, possibilitando aos usuários a criação de consultas que combinam informações de várias formas diferentes com mais praticidade e velocidade.

Dentro da Educação, o BI permite a descoberta de demandas de mercado visando a melhoria da empregabilidade e a formação dos graduados. A gestão pode participar do processo de tomada de decisão com base na realidade, o que elevaria o desempenho das instituições (GONÇALVES, 2021, *apud* MUSA, 2018).

Neste contexto, os trabalhos de Gonçalves (2021) e Mendes (2020) são exemplos de

bons resultados alcançados com o desenvolvimento de um ambiente de DW/BI, para análise de evasão em IES públicas, aplicando o modelo proposto por Kimball (2013) para criação de *Data Warehouses*. De forma semelhante, neste trabalho foram aplicadas técnicas inpiradas na metodologia de (KIMBALL, 2013) para o desenvolvimento de um *Data Mart* especializado para análise de abandono de discentes na UFES *Campus* Alegre.

1.1 Objetivos

1.1.1 Objetivo Geral

Construir um ambiente de DW/BI, inspírado na metodologia de Kimball (2013), que seja capaz de viabilizar a análise dos dados da evasão escolar na UFES *Campus* de Alegre.

1.1.2 Objetivos específicos

- Implementar um *Data Mart* que atenda as necessidades dos *stakeholders* (Coordenadores e Diretores).
- Implementar um fluxo de ETL automatizado utilizando tecnologias *open-source*.
- Construir visualizações que correspondam aos requisitos e analisar os resultados obtidos.

2 REVISÃO DE LITERATURA

Na primeira subseção da revisão bibliográfica, foi realizada uma contextualização da tecnologia de *Data Warehouse* (DW) apresentando as vantagens de construir uma arquitetura de DW/BI para estudo de dados organizacionais. Na segunda subseção são apresentadas obras relacionadas ao estudo da evasão de estudantes em Instituições de Ensino Superior (IES) com uso da tecnologia de DW. E por fim as contextualização das ferramentas utilizadas.

2.1 *Data Warehouse*

De acordo com Kimball (2013) os sistemas transacionais de uma organização são otimizados para executar transações de forma rápida e manter o fluxo do processo de negócio. Nessa proporção, os bancos de dados desse tipo de software não são recomendados para análise de dados, pois para manter a performance e otimizar armazenamento, é comum encontrar sistemas que não mantêm o histórico das operações processadas. Por outro lado, os sistemas de DW/BI são arquiteturas projetadas para realização de consultas em dados históricos gerados pelos sistemas transacionais de forma rápida e intuitiva.

Dessa maneira, os *Data Warehouse* facilitaram o trabalho dos analistas de dados, uma vez que, de acordo com Inmon (2005), essa tecnologia simplificou uma série de dificuldades encontradas nos ambientes tradicionais. Pois:

- É uma fonte única e integrada de dados, ou seja, deve integrar dados de várias fontes e formatos diferentes;
- Não deve ser volátil, o histórico deve ser preservado;
- É variante no tempo, em que, cada registro deve ser marcado com algum tipo de unidade temporal;

- É orientado a assunto, a sua modelagem busca atender problemas específicos da organização.

Samarakkody, Amaratunga e Haigh (2023) argumenta que um DW é compreendido como uma coleção integrada e dinâmica de dados, utilizada na tomada de decisões por meio de técnicas de Processamento Analítico Online (OLAP). Através dessas características, torna-se viável a construção de modelos multidimensionais pertinentes a cada domínio específico, contribuindo para solucionar desafios complexos dentro da organização.

2.1.1 Arquiteturas propostas de um *Data Warehouse*

Uma arquitetura de *Data Warehouse* é composta por quatro ambientes: Sistemas de origem, sistema de ETL, DW e aplicação de BI. Essa arquitetura opera o fluxo de dados desde a origem até o usuário final. A Figura 1 demonstra uma arquitetura básica de ambiente de DW.

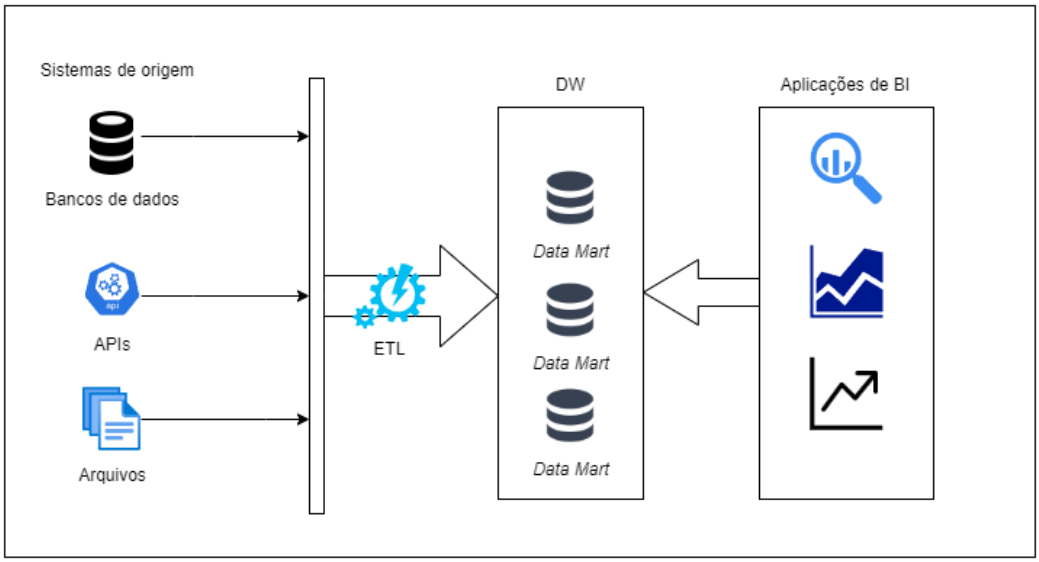

Figura 1: Arquitetura do DW de Kimball (2013).
Fonte: O Autor.

Na Figura 1, os ”Sistemas de Origem” representam os diversos sistemas utilizados pela organização, responsáveis por produzir os dados que alimentam o ambiente de DW. Geralmente, esses sistemas estão fisicamente isolados do ambiente, atuando como fonte para os modelos de dados construídos com fins analíticos.

No sistema de ETL, primeiramente vem a etapa de extração. Para Kimball (2013), extrair significa leitura e entendimento dos dados da origem, selecionando e copiando apenas o que é necessário para o ambiente de ETL. Em seguida, no tratamento, é feito saneamento do que foi extraído, como: remoção de duplicidade, conversão de tipo, padronização de rótulos, tratamento de dados nulos, remoção de ruídos e integração de dados de diversas origens. Após o tratamento, os dados são carregados no ambiente de DW, onde o objetivo de manter os dados organizados da forma mais detalhada possível para o consumo dos usuários por meio de aplicações de BI especializadas.

Existem várias metodologias para a estruturação do ambiente de DW, mas Kimball (2013) afirma que a modelagem dimensional é a técnica mais viável e amplamente aceita para entregar dados aos usuários, pois seu objetivo é tornar o banco simples para extração de informações.

Na arquitetura de Inmon (2005), o DW é uma estrutura maior e mais robusta, semelhante a de Kimball (2013), a principal mudança acontece na camada de estrutura do DW e apresentação dos dados, como ilustrado na Figura 2.

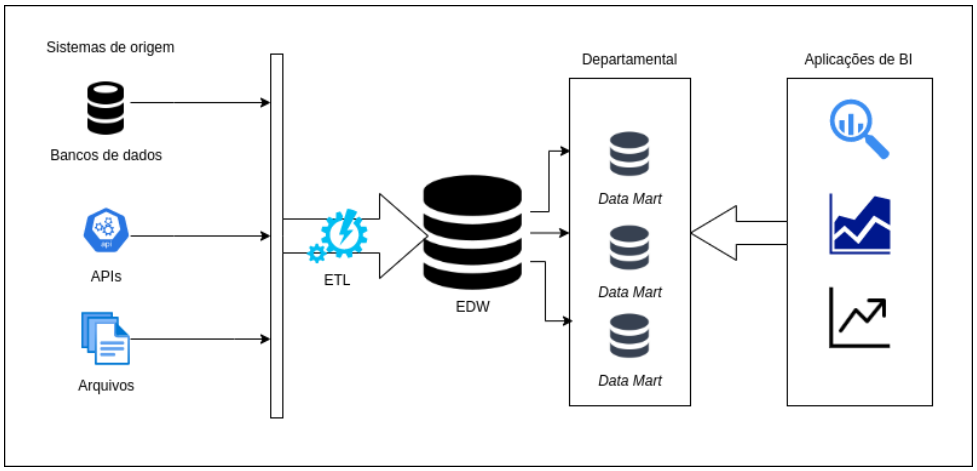

Figura 2: Arquitetura do DW de Inmon (2005).
Fonte: O Autor.

Nessa arquitetura o DW é uma estrutura maior e mais robusta chamada de *Enterprise Data Warehouse* (EDW), onde é modelada na terceira forma normal, buscando remover a redundância das entidades relacionais. Para Inmon (2005), a estrutura deve ser modelada para abranger o máximo possível das áreas de interesse da organização. Ou seja, a implementação depende do entendimento de toda a área, pois o seu objetivo é ser uma fonte de dados que atenda toda a comunidade da corporação.

Desta forma, a medida em que surgem nos departamentos a necessidade de obter informações do DW, são criados *Data Marts* para a apresentação destes dados. Ainda de acordo com Inmon (2005), os *Data Mart*s são estruturas baseadas no modelo dimensional e devem atender cada área departamental.

2.1.2 Modelagem Dimensional

Um dos principais motivos para o modelo dimensional ser amplamente utilizado, segundo Kimball (2013), é a simplicidade de sua estrutura, pois a sua proposta é que os usuários tenham facilidade para compreender os dados e consigam entregar resultados de forma rápida e eficiente com os *softwares* de BI. Inmon (2005) afirma que a construção de um modelo dimensional deve abranger apenas um assunto de interesse. Pois, a simplificação da estrutura melhora o desempenho das consultas.

A implementação do modelo dimensional em bancos de dados relacionais são referenciados como *Star Schema*, como o nome sugere a estrutura lembra o formato de uma estrela conforme ilustra a Figura 3.

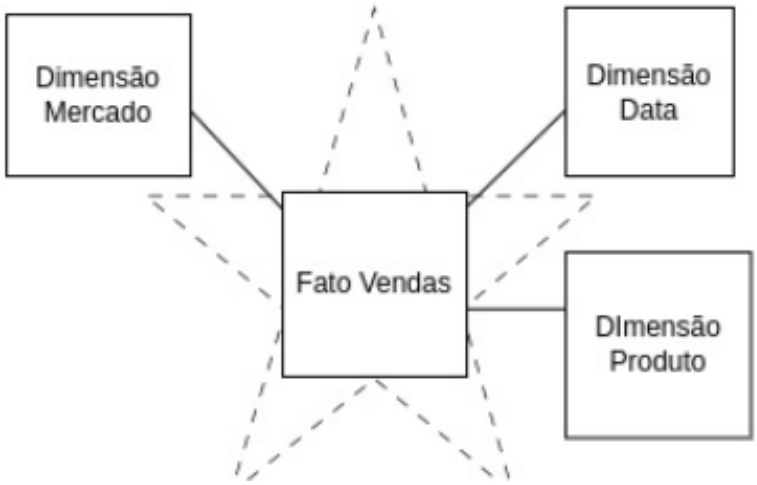

Figura 3: Exemplo do *Star Schema*.
Fonte: Adaptado de Kimball (2013).

Tal modelo é composto por três tipo de componentes: fatos, dimensões e suas conexões. Conforme Rowen et al. (2001), as dimensões são tabelas desnormalizadas que têm como objetivo armazenar os atributos relacionados ao evento analisado. Cada

atributo representa uma característica da medida analisada, podendo servir como filtro em consultas e rótulos na apresentação dos dados, facilitando o entendimento para o usuário.

A tabela de fatos visa armazenar as medidas numéricas associadas às dimensões do evento analisado. Cada linha dessa tabela está relacionada com as dimensões por meio de chaves estrangeiras. Dependendo do problema analisado, essa estrutura pode conter milhões de registros, organizados em um determinado nível de detalhes, chamado de grão. O grão é controlado por meio de agregações ou filtros. Assim, quanto menor o grão, mais detalhada será a informação.

A Figura 4, apresenta um exemplo prático de modelagem dimensional, onde a medida proposta é o evento de venda de um mercado varejista. É importante destacar a simplicidade da estrutura dimensional, onde seu objetivo é ser óbvio para o usuário entender e navegar sobre as entidades de forma fácil. Outra vantagem apresentada nesse exemplo é a performance que essa arquitetura oferece para um banco de dados, pois a quantidade de junções é reduzida melhorando o desempenho das consultas.

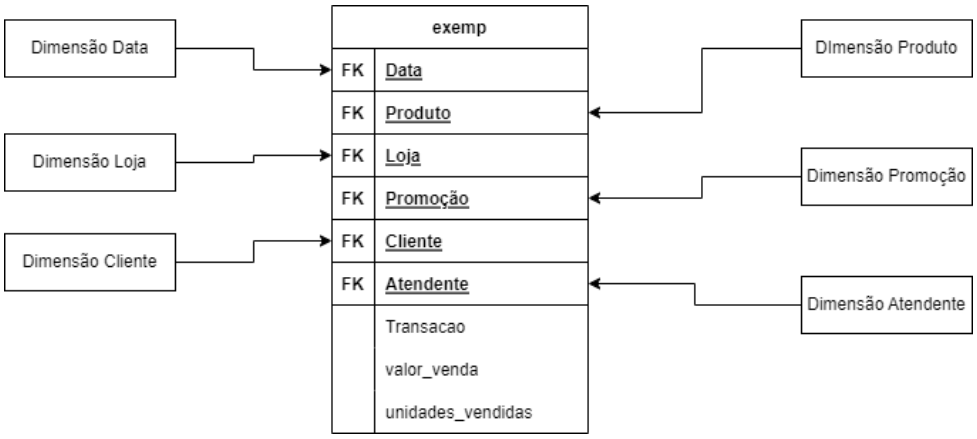

Figura 4: Exemplo Modelo dimensional.
Fonte: Adaptado de Kimball (2013).

2.2 Trabalhos relacionados

Nesta seção são apresentados alguns trabalhos relacionados ao tema deste projeto. A pesquisa foi realizada na plataforma *Google Acadêmico* e buscou-se pelas seguintes palavras chaves: "*Data Warehouse* sobre indicadores de alunos de universidades", "*Data Warehouse* evasão escolar", "*Business Intelligence* aplicada em evasão escolar" e "evasão no ensino superior". A partir dos resultados coletados, foram selecionadas

três obras que mais se relacionam ao tema deste trabalho.

A trabalho de Sales Junior (2013) foi uma das escolhidas por apresentar um estudo sobre a evasão estudantil nos *Campus* da UFES. Os trabalhos de Mendes (2020) e Gonçalves (2021) foram escolhidos por abordarem a criação de DWs para estudo da evasões em instituições de superior, além do fato de ambos os trabalhos utilizarem a metodologia de *Ralph Kimball* para modelagem dimensional.

2.2.1 Trabalho de SALES JUNIOR (2013)

O trabalho de Sales Junior (2013) teve os seguintes objetivos: compreender as possíveis causas que podem estar relacionadas à evasão e permanência dos discentes nos cursos de graduação da Universidade Federal do Espírito Santo (UFES), identificar perfis de alunos que têm maior probabilidade de evadir do Sistema de Ensino Superior, descobrir associações entre a evasão e fatores antecedentes a matrícula do aluno na graduação.

Dessa forma, para fazer a análise estatística dos dados, Sales Junior (2013) utilizou dados dos sistemas de informação da UFES sobre os alunos que ingressaram pelo vestibular entre os anos de 2006 a 2011, e evadiram entre os anos de 2007 e 2012/1. Nessa proporção, utilizou métodos estatísticos de análise exploratória dos dados para conhecer o perfil de estudantes evadidos e encontrar possíveis causas da evasão, além de usar métodos de inferência estatística aplicada em amostras de estudantes formados e evadidos, e as demais variáveis de forma exploratória.

Na apresentação dos resultados, o autor primeiramente exibiu os dados extraídos do sistema da UFES por meio de tabelas e gráficos de barra. É possível observar que na Tabela 1 é apresentado um total de 3266 evadidos representado 15,8% do total. Tal número é apresentado de forma detalhada na Tabela 2 apresentando a evasão por forma de evasão e tempo até a evasão, onde é possível observar que a ”Desistência” e o ”Desligamento por abandono” são os maiores motivos de evasão.

Situação de Mátricula	**Frequência**	**Percentual**
Formado	3233	15,6
Evadido	3266	15,8
Sem evasão	14,227	68,6
Total	20726	100

Tabela 1: Tebela com as quantidades de Evadidos, Diplomados e Matriculados.
Fonte: Adaptado de (SALES JUNIOR, 2013).

Formas de evasão	n	%	Tempo até a evasão		
			Média	Mediana	Desvio padrão
Desistência	1530	46,8	3,5	3	2,3
Desligamento por abandono	1053	32,2	7,2	7	2,2
Outras formas	9	0,3	2,9	2	2,8
Outros desligamentos	309	9,5	7,3	7	2
Remoção de curso	298	9,1	4,6	5	1,9
Transferência para outra IES	67	2,1	4,7	5	2,5
Total	3266	100	6,9	5	2,9

Tabela 2: Tabela de frequência da forma de evasão e tempo até a evasão.
Fonte: Adaptado de (SALES JUNIOR, 2013).

Na Figura 5 é apresentado um gráfico de barras para a apresentação distribuição do tempo até a evasão. Onde o primeiro semestre concentra 400 evasões ainda no primeiro período com uma taxa de 13% e o no quinto período em diante é possível observar um agravamento do valor da taxa.

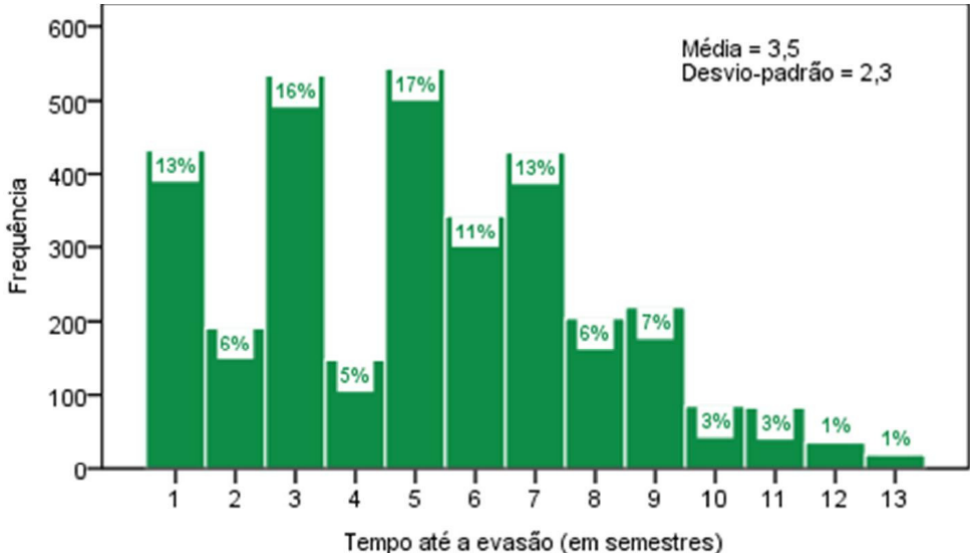

Figura 5: Distribuição do tempo até a evasão.
Fonte: (SALES JUNIOR, 2013).

Por fim, ao relacionar os dados quantitativos e qualitativos, o autor constatou que a evasão pode estar diretamente associada à falta de maturidade do aluno ao escolher o curso sem clareza sobre seus objetivos profissionais. Além disso, identificou-se que alguns alunos utilizam cursos de baixa concorrência como uma espécie de ”ponte” para outros cursos oferecidos. Outros fatores relevantes estão ligados à questão financeira: alunos que precisam trabalhar para custear seus estudos tendem a evadir

com maior frequência, devido à sobrecarga de responsabilidades. Outro ponto observado foi a relação com a escolaridade anterior, evidenciando que alunos provenientes de escolas particulares têm menor probabilidade de evasão.

2.2.2 Trabalho de Mendes (2020)

A obra de Mendes (2020) foi escolhida por ser a que mais se aproxima desta proposta de trabalho. Para tanto, foi proposto o desenvolvimento de um DW para apoio à tomada de decisão da Universidade Federal de Itajubá (Unifei). Foram apresentados métodos de cálculos para taxa de evasão e características descritivas dos alunos, para compreender os fatores que influenciam na evasão estudantil. A metodologia do autor toma como referência, o método de modelagem dimensional proposto por Kimball (2013), para a definição do modelo do *Data Warehouse* e implementação do fluxo dos dados da origem até a área de apresentação.

Como resultado da pesquisa, o autor identificou três grupos de métodos de cálculo para taxa de evasão:

- Métodos que consideram gerações completas de alunos, onde o tempo máximo de integralização expirou;
- Métodos que consideram alunos ingressantes dentro de um período específico;
- Métodos que propõem o acompanhamento da série histórica das taxas de evasão;

Contudo, no desenvolvimento do trabalho, o terceiro grupo foi escolhido, como método de cálculo para taxa de evasão, pois tem como objetivo acompanhar o histórico dos índices, para identificar possíveis sucessos de medidas e mitigar a evasão estudantil.

A Figura 6 representa o modelo dimensional proposto pelo autor. A tabela fato “discente” representa a situação de matrícula de um discente em determinado curso e o fato “evasão” representa o evento de fuga estudantil em um período do curso. Relacionadas aos fatos estão as seguintes tabelas dimensões: “pessoa”, “tempo”, “curso” e “situação de matrícula”.

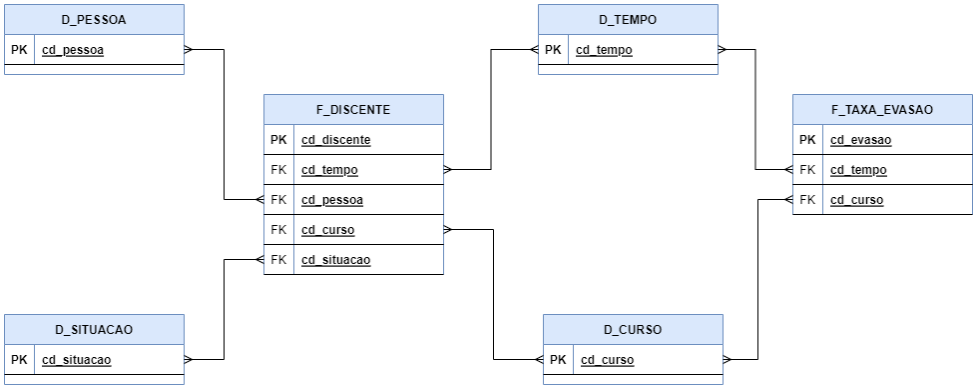

Figura 6: Modelo dimensional proposto por Mendes (2020).
Fonte: Adaptado de Mendes (2020).

A partir do modelo dimensional desenvolvido, Mendes (2020) apresentou resultados em formato de séries temporais, comparando as taxas anuais de evasão, diplomação e cancelamento de matrículas. A Figura 7 demonstra uma análise referente à Unidade ”A” (denominação utilizada pelo autor para preservar informações institucionais da Unifei). No gráfico, nota-se que a taxa média de evasão anual é de 2,9%, enquanto a taxa de diplomação apresenta uma média de 10,1%. Destaca-se um pico histórico na taxa de diplomação em 2011, atingindo 20%, seguido por dois declínios nos anos subsequentes (2012 e 2015). O autor atribui essa oscilação ao aumento da oferta de vagas do programa REUNI. É relevante ressaltar que na exposição dos resultados, o autor concentrou-se na apresentação da estrutura do sistema de DW, sem adentrar em detalhes específicos sobre as causas da evasão, atribuindo à gestão acadêmica da IES a responsabilidade pela análise mais aprofundada dos motivos.

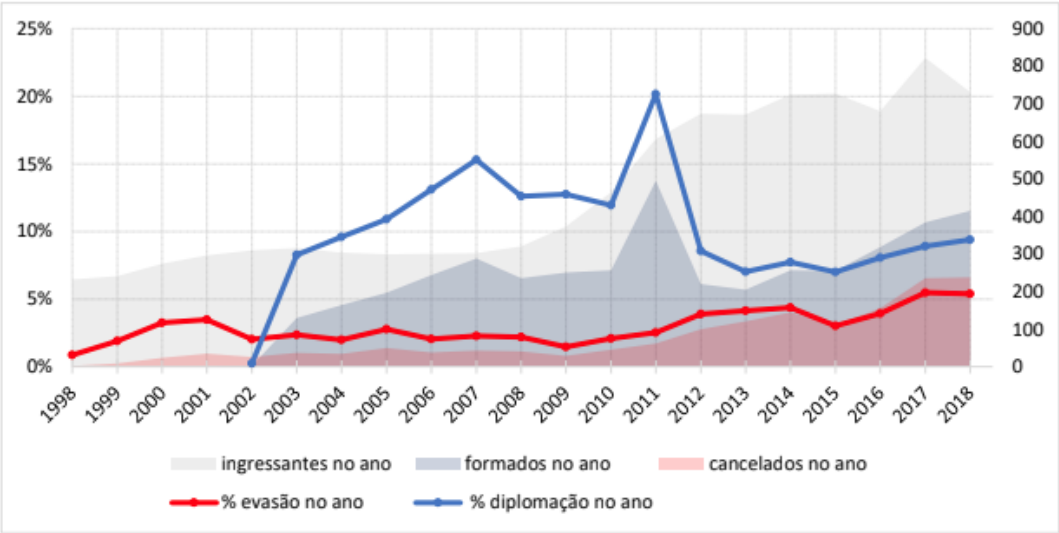

Figura 7: Gráfico da Evolução das Taxas Anuais de Evasão e Diplomação da Unidade de Observação “A”.
Fonte: Mendes (2020).

2.2.3 Trabalho de Gonçalves (2021)

Gonçalves (2021) utilizou em seu trabalho os dados extraídos da Plataforma Nilo Peçanha (PNP), para criar uma arquitetura de BI, com o objetivo de abordar indicadores relacionados à taxa de evasão e retenção de alunos do Instituto Federal do Mato Grosso (IFMT).

A PNP é um sistema de informação para coleta, validação e disseminação de estatísticas oficiais da Rede Federal de Educação Profissional, Científica e Tecnológica. E agrupa dados relacionados aos docentes, discentes, técnicos-administrativos e de gastos financeiros da Rede Federal, para levantamento de indicadores de gerência. Os dados são analisados pela Secretaria de Educação Profissional e Tecnológica do Ministério da Educação (PEÇANHA, 2022).

Para fins de cálculos das métricas de evasão e retenção, a autora utiliza como base o Manual de Referência Metodológica do PNP. Pois, o mesmo oferece o dicionário de dados da base do PNP e também a padronização dos cálculos dos indicadores que expressam as medidas de desempenho das instituições da Rede Federal (BRASIL, 2016).

A partir dos relatórios de alunos retirados do sistemas, Gonçalves (2021) implementou um *Data Mart*, apresentado na Figura 8, para atender aos requisitos identificados. A

tabela fato é o próprio relatório e as características dos dados foram segmentadas nas seguintes dimensões: ciclo acadêmico, situação de matrícula, município, tipo de nível, unidade de ensino, curso e tipo de curso. Por fim, os resultados mostraram o desempenho anual das situações de matrículas classificadas por *Campus* e curso, organizadas em tabelas e gráficos.

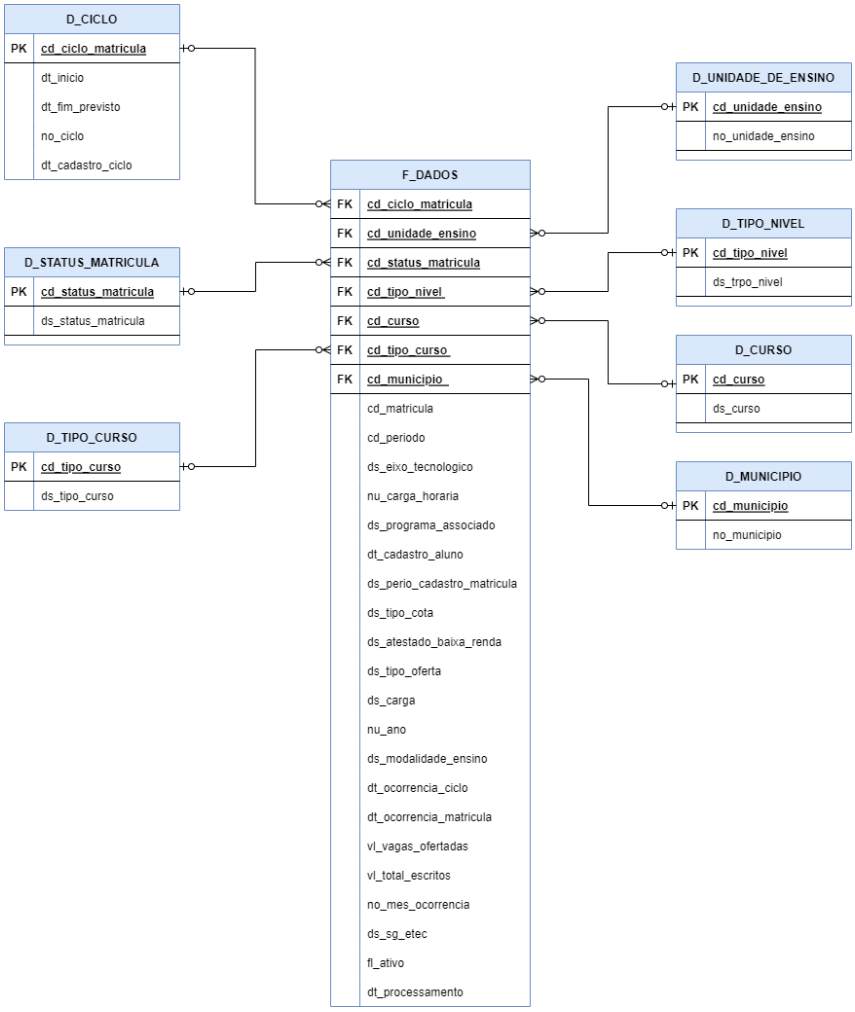

Figura 8: Modelo dimensional proposto por Gonçalves (2021).
Fonte: Adaptado de Gonçalves (2021).

A partir do modelo dimensional, o autor apresentou os resultados por meio de tabelas, mostrando o número de estudantes retidos por ciclo acadêmico, juntamente com um conjunto gráficos para comparar os indicadores de cada unidade de ensino do IFMT. Na Figura 9, é exibida uma tabela que demonstra o total de estudantes em cada situação de matrícula por ciclo acadêmico de cada curso das unidades de ensino.

Observa-se que a IES conta com um total de 11.342 diplomados, número inferior à soma das taxas de evasão (abandono e desligamento), que totalizam 11.516. Essa

diferença sinaliza uma situação crítica que requer investigação por parte da gestão acadêmica.

Ademais, o autor optou por não explorar profundamente as possíveis causas da evasão, focando, na apresentação dos tipos de análises que podem ser realizadas por meio da ferramenta desenvolvida.

unidade_ensino	ABANDONO	CONCLUÍDA	DESLIGADO	EM_CURSO	INTEGRALIZADA	REPROVADA	TRANSF_EXT
BIOCOMBUSTÍVEIS - PRESENCIAL - ABR. 2015 / DEZ. 2017	33	4		5			
BIOCOMBUSTÍVEIS - PRESENCIAL - JUN. 2016 / DEZ. 2018	29			5			
INSTITUTO FEDERAL DE MATO GROSSO - CAMPUS CAMPO NOVO DO PARECIS	**688**	**474**	**211**	**1063**	**16**		**226**
BACHARELADO	**145**	**141**	**62**	**334**			**50**
AGRONOMIA	**145**	**141**	**62**	**334**			**50**
AGRONOMIA - EDUCAÇÃO PRESENCIAL - AGO. 2018 / JUL. 2023	4			32			
AGRONOMIA - EDUCAÇÃO PRESENCIAL - FEV. 2019 / DEZ. 2023			1	34			
AGRONOMIA - EDUCAÇÃO PRESENCIAL - FEV. 2020 / DEZ. 2024			2	36			
AGRONOMIA - EDUCAÇÃO PRESENCIAL - JUL. 2019 / JUL. 2024			2	33			
AGRONOMIA - EDUCAÇÃO PRESENCIAL - MAR. 2018 / DEZ. 2022	3		4	28			4
AGRONOMIA - EDUCAÇÃO PRESENCIAL - SET. 2017 / JUL. 2022	7		2	29			3
AGRONOMIA - PRESENCIAL - ABR. 2016 / DEZ. 2020	5	1	4	26			3
AGRONOMIA - PRESENCIAL - AGO. 2012 / JUL. 2017	14	12	4	2			6
AGRONOMIA - PRESENCIAL - AGO. 2013 / AGO. 2018	11	13	5	1			3
AGRONOMIA - PRESENCIAL - FEV. 2009 / DEZ. 2013	25	31	5	4			2
AGRONOMIA - PRESENCIAL - FEV. 2013 / DEZ. 2017	7	18	7	1			3
AGRONOMIA - PRESENCIAL - FEV. 2014 / DEZ. 2018	6	13	5	11			4
AGRONOMIA - PRESENCIAL - FEV. 2015 / DEZ. 2019	5	15	4	8			4
AGRONOMIA - PRESENCIAL - MAR. 2011 / DEZ. 2015	15	16	4	1			3
AGRONOMIA - PRESENCIAL - MAR. 2012 / DEZ. 2016	15	13	1	1			4
AGRONOMIA - PRESENCIAL - MAR. 2017 / DEZ. 2021	2		4	28			1
AGRONOMIA - PRESENCIAL - OUT. 2015 / JUL. 2020	8	3	4	26			3
AGRONOMIA - PRESENCIAL - SET. 2014 / SET. 2019	11	6	2	8			5
AGRONOMIA - PRESENCIAL - SET. 2016 / JUL. 2021	7		2	25			2
ESPECIALIZAÇÃO (LATO SENSU)	**15**	**10**	**2**		**16**		
ESPECIALIZAÇÃO EM EDUCAÇÃO PROFISSIONAL INTEGRADA À EDUCAÇÃO BÁSICA NA MODALIDADE DE EDUCAÇÃO DE JOVENS E ADULTOS	**15**	**10**	**2**		**16**		
ESPECIALIZAÇÃO EM EDUCAÇÃO PROFISSIONAL INTEGRADA À EDUCAÇÃO BÁSICA NA MODALIDADE DE EDUCAÇÃO DE JOVENS E ADULTOS - PRESENCIAL - DEZ. 2010 / JUL. 2012	15	10	2		16		
LICENCIATURA	**165**	**28**	**40**	**87**			**6**
MATEMÁTICA	**165**	**28**	**40**	**87**			**6**
MATEMÁTICA - EDUCAÇÃO PRESENCIAL - FEV. 2019 / DEZ. 2023			2	26			1
MATEMÁTICA - EDUCAÇÃO PRESENCIAL - FEV. 2020 / JUL. 2024				17			
Total	**8675**	**11342**	**2841**	**25222**	**427**	**6**	**3484**

Figura 9: Demonstração dos estudantes retidos nos ciclos.
Fonte: Gonçalves (2021).

2.3 Ferramentas Utilizadas

2.3.1 Linguagem de programação Python

Para implementação do ETL optou-se por utilizar a linguagem de programação Python [1], pois de acordo com Solutions (2017), trata-se de uma ferramenta com licença permissiva, consolidada no mercado, com proximidade da linguagem natural, compatível com diversas plataformas e possui diversas bibliotecas e *frameworks* que facilitam o desenvolvimento de diversos tipos aplicações. Como por exemplo, o pacote *pandas* que fornece uma estrutura de dados rápida, flexível e expressiva, na qual foi criada para análise e manipulação de estruturas “relacionais” e “rotuladas” de forma fácil e intuitiva (PANDAS, 2022).

[1] https://www.python.org/

2.3.2 Apache Airflow

O Apache Airflow [2] é uma plataforma mantida pela licença *Apache* utilizada para criar, agendar e monitorar de forma programática fluxos de trabalho (APACHE, 2022). O seu objetivo é ser utilizado para gerenciar o fluxo de ETL, pois se integra facilmente com a linguagem Python [3], além de possuir uma interface de uso amigável e capaz de ser integrada com diversos tipos de tecnologias.

Os fluxos de trabalho são organizadas em grafos acíclicos direcionados que também são conhecidos como "DAGs" (*Directed Acyclic Graph*), onde os vértices são as tarefas e as arestas definem a relação de dependência entre elas definindo o fluxo de execução da carga de trabalho. Além disso, são implementadas com a linguagem Python [4] e são programadas para serem executadas de forma agendada ou manual. Dessa forma, para este trabalho, cada tarefa será um *script Python* [5] que executa algum passo do processo. A Figura 10 exemplifica a estrutura de uma DAG.

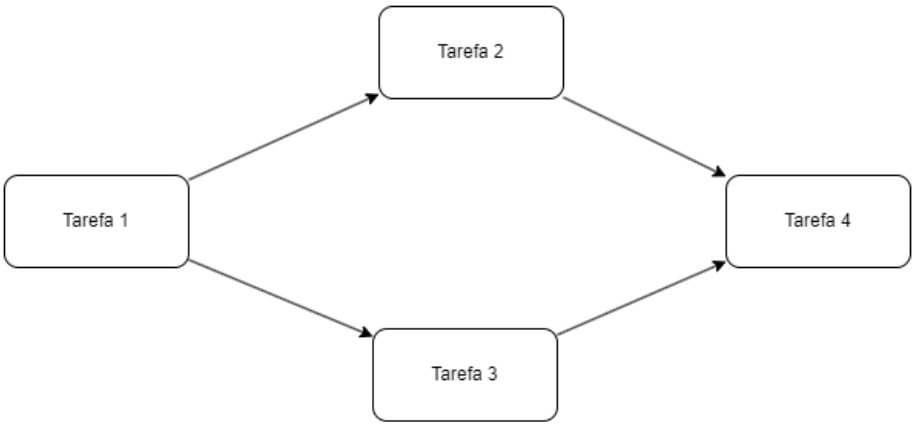

Figura 10: Exemplo de uma DAG.
Fonte: O Autor.

2.3.3 PostgreSQL

O PostgreSQL [6] é um sistema de gerenciamento de banco de dados objeto-relacional (ORDBMS) *open-source* e tem suporte para grande parte do padrão SQL e dos re-

[2] https://airflow.apache.org/
[3] https://www.python.org/
[4] https://www.python.org/
[5] https://www.python.org/
[6] https://www.postgresql.org/

cursos modernos de um SGBD, como: consultas complexas, chaves estrangeiras, gatilhos, *views* e integridade transacional (POSTGRESQL, 2022). Nesse projeto, o PostgreSQL [7] será utilizado para manter a estrutura física das tabelas da *Staging Area* e do *Data Warehouse*.

2.3.4 MIcrosoft Power BI

A escolha do MIcrosoft Power BI [8]. baseia-se em sua interface intuitiva, que permite aos usuários, mesmo sem conhecimento técnico aprofundado, criar *dashboards* interativos e relatórios dinâmicos de forma ágil. Além disso, sua integração com diversas fontes de dados, sejam elas locais ou remotas, como bancos de dados, planilhas e serviços web, ampliando a versatilidade e a abrangência das análises realizadas.

2.3.5 Docker

Para facilitar a implantação e manutenção do sistema, foi utilizado a ferramenta Docker [9], no qual se trata de um serviço de gerenciamento de infraestrutura de *Containers* de código aberto. De acordo com Docker (2023), essa ferramenta facilita a implantação de ambientes complexos, pois os *Containers* se tratam de processos que possuem suas bibliotecas e binários de sistemas isolados do sistema operacional hospedeiro, evitando a necessidade de utilização de máquinas virtuais para hospedagem de serviços. Consequentemente, otimizando a utilização dos recursos da máquina. Além disso, o Docker [10] fornece uma interface de linha de comando (CLI) que facilita a manipulação de sistemas complexos. A Figura 11 apresenta uma comparação entre um ambiente utilizando máquinas virtuais e Docker [11], onde é possível observar que as máquinas virtuais consomem mais recursos.

[7] https://www.postgresql.org/
[8] https://powerbi.microsoft.com/pt-br/
[9] https://www.docker.com/
[10] https://www.docker.com/
[11] https://www.docker.com/

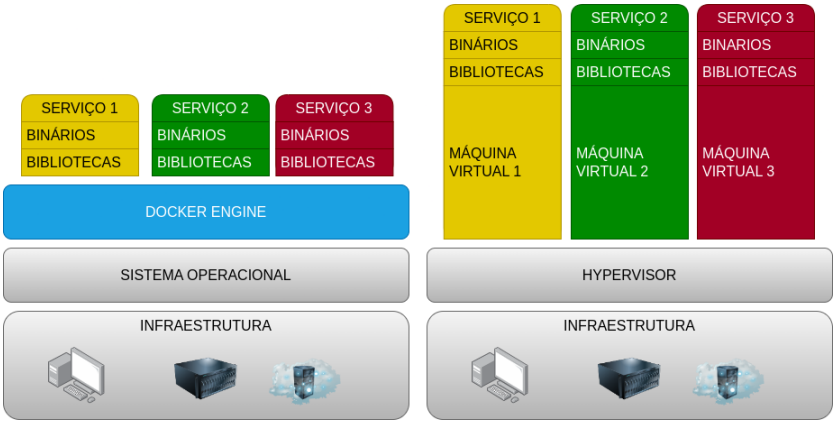

Figura 11: Comparação de um sistema utilizando Docker e um sistema utilizando máquinas virtuais.

Fonte: O Autor.

3 METODOLOGIA

Neste trabalho, foi implementado um ambiente de BI baseado na arquitetura de Kimball (2013), que é capaz de orquestrar o fluxo de ETL para povoar o *Data Mart* proposto. A Figura 12 apresenta a arquitetura proposta nesta pesquisa, que será melhor detalhada a seguir.

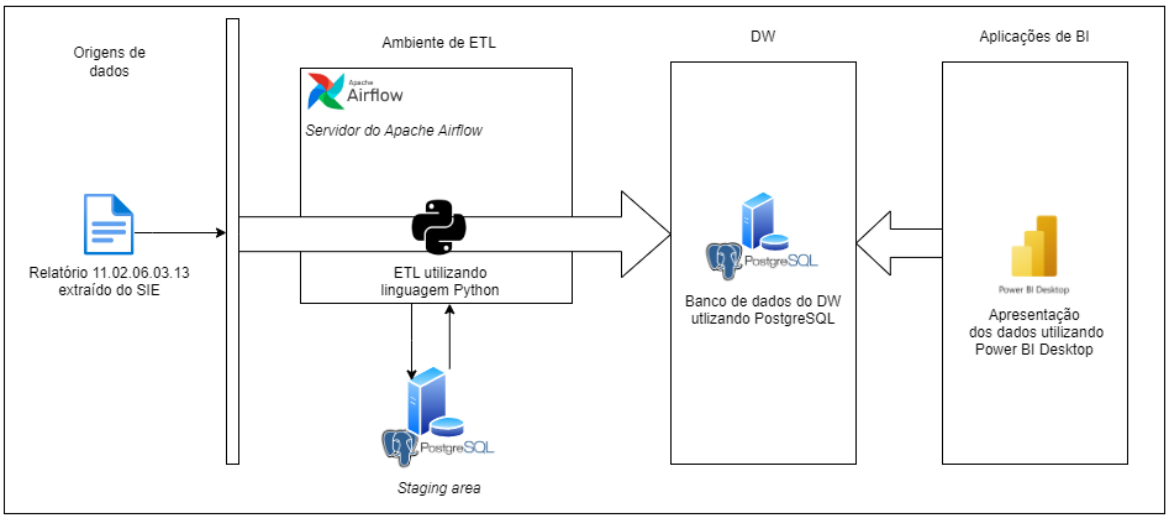

Figura 12: Arquitetura do DW.
Fonte: O Autor.

Inicialmente foi realizado um estudo para levantamento dos requisitos do problema com o objetivo de entender as dificuldades dos tomadores de decisão. E por conseguinte, definir quais indicadores e métricas são importantes para que o sistema possa atender as necessidades dos usuários.

Na etapa inicial de ”Origens de dados” (Figura 12) foram utilizados relatórios da situação de matrícula dos alunos extraídos do SIE [1] utilizado pela UFES.

A partir dos dados obtidos foi elaborado um modelo dimensional com todas as dimen-

[1] https://npd.ufes.br/sistema-de-informa%C3%A7%C3%A3o-para-o-ensino-sie

sões identificadas como importantes para analisar o fato em questão e por conseguinte abordar todos os indicadores levantados com os usuários chave.

Na etapa seguinte, ambiente de ETL, foi criado uma estrutura para preparação dos dados (*Staging Area*). De acordo com Anand e Kumar (2013), esse ambiente ajuda a manter a qualidade do processo, pois agrupa fontes heterogêneas de dados, se tornando uma fonte única para o fluxo de ETL. A *Staging Area* foi implementada no formato de tabelas relacionais e foi armazenado em um de banco de dados relacional PostgreSQL [2].

O ambiente de ETL (Figura 12), foi implementado na linguagem de programação Python [3]. Os dados foram extraídos da *Staging Area*, preparados e padronizados para alimentar a estrutura do DW. Vale ressaltar que a estrutura do DW também foi mantida em um PostgreSQL [4].

A automatização da execução desses dois fluxos citados, é de responsabilidade do Apache Airflow [5], pois trata-se de uma ferramenta especializada para agendamento, orquestração e gerenciamento de fluxos e *pipelines* de dados.

E por fim, a ferramenta Microsoft Power BI foi utilizada para leitura do DW e criação de visões e *Dashboards* capazes de fornecer, para os *stakeholders*, uma melhor compreensão e visualização dos dados e facilitar o processo de tomada de decisão.

3.1 Levantamento dos Indicadores

O processo de definição da estrutura do *Data Mart* começa nessa etapa. Logo, o objetivo é identificar todas as dificuldades encontradas no processo de tomada de decisão da organização em relação aos fatos abordados. E por conseguinte, construir um modelo dimensional para organizar os dados de forma eficiente para facilitar as respostas que os usuários querem obter dos dados.

Em entrevistas com os *stakeholders* foram identificados os seguintes indicadores:

- Identificar possíveis perfis de tendência de evasão e permanência dos estudantes nos cursos da Ufes *Campus* Alegre

[2] https://www.postgresql.org/
[3] https://www.python.org/
[4] https://www.postgresql.org/
[5] https://airflow.apache.org/

- Qual impacto da adoção do sistema de cotas na evasão?
- Qual a quantidade de períodos cursados até a evasão e diplomação?
- Qual a quantidade de alunos evadidos por município de origem?
- Qual a quantidade de alunos evadidos por município de naturalidade?
- Qual a quantidade de alunos evadidos por forma de ingresso?
- Qual a quantidade de alunos evadidos por forma de evasão?
- Qual a quantidade de alunos evadidos por curso?
- Qual a quantidade de alunos evadidos por gênero?
- Qual a quantidade de alunos evadidos por tipo de cota?
- Qual a quantidade de alunos evadidos por deficiência?

3.2 Base de Dados de Origem

O Sistema de Informação para o Ensino (SIE) [6] é o sistema adotado pela UFES para a gestão de suas atividades acadêmicas, recursos humanos, registro e controle acadêmico, que incluem disciplinas, cursos, docentes e currículos, além da administração de matrículas, gestão contábil e patrimonial (UFES, 2022). Assim, o estudo foi conduzido utilizando dados extraídos do relatório ”11.02.06.03.13 Ingressantes por Curso/Ano/-Semestre Analítico”, que contém informações cadastrais dos alunos e suas situações de matrícula no momento da evasão.

A Figura 3 ilustra as diferentes formas de evasão consideradas pelo SIE [7]. É importante ressaltar que a categoria ”Sem Evasão” é utilizada para classificar os alunos que ainda estão matriculados.

[6] https://npd.ufes.br/sistema-de-informa%C3%A7%C3%A3o-para-o-ensino-sie
[7] https://npd.ufes.br/sistema-de-informa%C3%A7%C3%A3o-para-o-ensino-sie

FORMAS DE EVASÃO
JUBILADO
DESLIGAMENTO POR MANDADO JUDICIAL
NULIDADE DE MATRÍCULA – ATO ADMINISTRATIVO
REOPÇÃO DE CURSO
FORMADO
DESLIGAMENTO RESOLUÇÃO 68/2017-CEPE
DESLIGAMENTO POR ABANDONO
MATRÍCULA DESATIVADA
DESISTÊNCIA
SEM EVASÃO
DESLIGAMENTO: DESCUMPRIU PLANO DE ESTUDOS
ERRO DE CADASTRO
FALECIMENTO
ERRO DE CADASTRO
FALECIMENTO
DESLIGAMENTO: 3 REPROVAÇÕES EM 1 DISCIPLINA
ADAPTAÇÃO CURRICULAR
REMANEJAMENTO SISU
TRANSFERIDO
REMOÇÃO

Tabela 3: Formas de evasão consideradas no SIE.
Fonte: O Autor.

3.3 Análise dos dados

Para verificar se os indicadores pudessem ser contemplados com sucesso foi feita uma análise dos dados extraídos de relatórios do SIE [8]. Os arquivos fornecidos são planilhas no formato *Microsoft Excel*, onde cada registro apresenta as informações cadastrais dos discentes juntamente com a situação de matrícula, entre o primeiro semestre de 2009 até o primeiro semestre de 2023.

No processo de análise foram observados os seguintes cenários:

- CPF Nulo;
- Matrícula com caracteres não numéricos;
- Dados de endereço Nulo ou inseridos incorretamente;
- Dados de Região Nulo;

[8] https://npd.ufes.br/sistema-de-informa%C3%A7%C3%A3o-para-o-ensino-sie

- Dados de Gênero Nulo;
- Registros duplicados divergindo apenas nas informações de endereço;
- Registros com data de evasão anterior que data de ingresso;

Para facilitar a manipulação e o trabalho com os dados, todos os arquivos foram armazenados em uma tabela relacional ”stg_relatorio” e foi criada a *view* ”v_ds_stg_relatorio” (Apêndice A), com objetivo de filtrar e padronizar os dados brutos para utilização no DW, da seguinte forma:

- Mudança do texto para caixa alta;
- Remoção de espaçamento desnecessário;
- Tratamento de tipos de dados;

Ambas as estruturas estão organizadas no schema ”stg”, conforme a Figura 13:

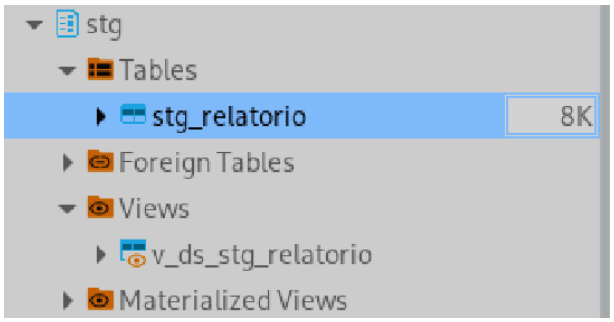

Figura 13: *Schema Staging Area*.
Fonte: O Autor.

3.4 Modelo Dimensional

A partir dos requisitos levantados e dos dados obtidos foi possível propor um modelo dimensional conforme apresentado na Figura 14.

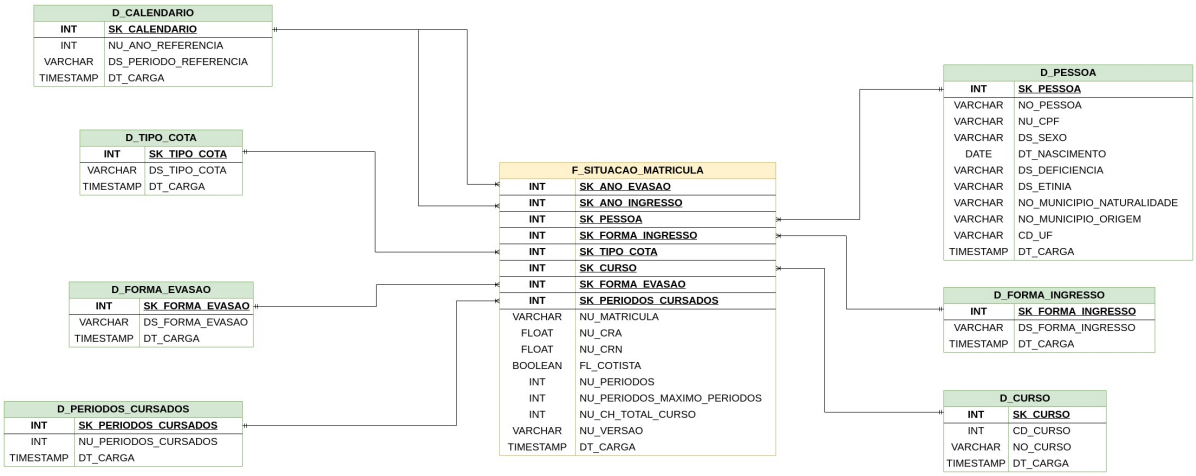

Figura 14: Modelo dimensional.
Fonte: O Autor.

Na Figura 14 a tabela fato F_SITUACAO_MATRICULA contém informações históricas dos alunos que ingressaram, evadiram e diplomaram. E cada registro representa uma matrícula em determinada situação. Toda vez que a fato é carregada, todos os alunos que não evadiram são excluídos e adicionados novamente com as informações atualizadas.

Para atender os requisitos especificados, foram identificadas as seguintes dimensões: D_TIPO_COTA para classificação dos tipos de cota, D_CALENDARIO para referência temporal da informação em ano e período, D_PERIODOS_CURSADOS contém a quantidade de periodos que um aluno cursou até a evasão ou a diplomação, D_FORMA_EVASAO com a classificação descritiva dos tipos de evasão, D_PESSOA mantém as características pessoais do aluno, D_FORMA_INGRESSO para classificação das formas de ingresso e D_CURSO possui as informações do curso. Vale ressaltar que o dicionário de dados do modelo dimensional proposto está disponível no Apêndice D deste trabalho.

Além disso, as dimensões D_FORMA_INGRESSO, D_TIPO_COTA, D_CALENDARIO e D_FORMA_EVASAO não possuem chave primária natural para junção dos dados, logo, a sua identificação deve ser feita por meio do seu próprio valor. Conforme Kimball (2013), essas são Dimensões de Mudança Lenta (*Slowly Changing Dimensions* - SCD) do tipo 0. Ou seja, são dimensões nas quais os valores dos atributos não sofrem alterações. E as dimensões D_PESSOA e D_CURSO são SCD do tipo 1, na quais, podem ter os valores atualizados sem manter o histórico das versões. E nenhum dos tipos citados podem conter dados duplicados. A dimensão

D_PERIODOS_CURSADOS é gerada com base no cálculo da quantidade de períodos que o aluno cursou.

Ainda sobre as dimensões, cada registro é identificado por uma *Surrogate Key*, segundo Kimball (2013), são chaves primárias artificiais que servem para padronizar e abstrair os identificadores naturais dos dados e facilitar a junção com tabela fato. Na Figura 14, tais colunas possuem o prefixo ”SK”.

Para facilitar a visualização e o consumo das informações do DW foi criado uma (*view*) ”v_ds_evasao”, na qual é uma consulta SQL armazenada no banco de dados, que abstrai as informações necessárias para a construção dos relatórios. O *script* de definição está disponível no Apêndice B.

E todas essas estruturas foram organizadas no *schema* ”dw” conforme a Figura 15.

Figura 15: *Schema* do DW.
Fonte: O Autor.

3.5 Estrutura do Projeto e Fluxo de ETL

O projeto construído foi estruturado de forma hierárquica organizada em módulos de serviços. Onde os níveis mais baixos abstraem e generalizam os detalhes de implementações para os módulos clientes por meio de classes de serviços ou *scripts templates*. Para que o projeto possa ser de fácil manutenção, estendido para novas

funcionalidades e reutilizável em outras soluções, foram utilizados princípios da Programação Orientada a Objetos e padrões de *design* de *softwares*, como:

- Cadeia de Responsabilidade;
- Método Fábrica;
- Método *Template*;
- *Strategy*;

De acordo com Gamma et al. (1994), padrões melhoram a comunicação entre os componentes do programa, tornando mais fácil o reuso, compreensão, e manutenção para os desenvolvedores.

Todo código fonte desenvolvido para realizar as rotinas de ETL está organizado na estrutura de pastas apresentada na Figura 16, disponível no repositório do Github [9]. E cada um deles vai ser detalhado nas próximas subseções.

```
.
├── airflow
│   ├── airflow-db
│   ├── config
│   ├── dags
│   ├── data
│   ├── datawarehouse-db
│   ├── logs
│   │   ├── dag_processor_manager
│   │   └── scheduler
│   └── plugins
├── dags
│   ├── dw
│   └── stg
├── docs
├── reports
├── scripts
│   ├── dw
│   └── stg
└── src
    ├── resources
    └── utils
        ├── connection
        ├── consumers
        ├── file
        ├── producers
        ├── table
        └── transformers
```

Figura 16: Estrutura do projeto.
Fonte: O Autor.

[9] https://github.com/eloy-freitas/TCC2

3.5.1 Diretório *airflow*

Esse diretório é gerado após a finalização do tutorial (Apêndice E). Esse diretório tem o objetivo de concentrar as estruturas que o Apache Airflow [10] utiliza para sua execução. Os detalhes de cada um não são escopo deste trabalho, mas vale ressaltar que:

- airflow/data - É destinado para o usuário disponibilizar os arquivos que serão carregados no DW. No contexto desse projeto, o *script* de ETL da ”stg_relatorio” faz a leitura de todos os arquivos de relatório e os armazena no ambiente de stg apresentado na Figura 16. Vale ressaltar que, para garantir a integridade dos resultados apresentados, todos os relatórios utilizados como fonte de dados devem possuir as colunas apresentadas no Apêndice C;
- airflow/datawarehouse-db - É o diretório de trabalho do banco de dados do *container* DW.

3.5.2 Diretório docs

Esse diretório tem o objetivo de manter os arquivos de documentação do projeto. Nele estão concentrados o dicionário de dados e modelo dimensional.

3.5.3 Diretório *reports*

O diretório ”reports” foi criado com o objetivo de manter os arquivos bases de ferramentas que foram utilizadas no processo de análise dos dados, como por exemplo, os arquivos da ferramenta MIcrosoft Power BI [11].

3.5.4 Diretório src

O diretório ”src/” contém sub módulos com utilitários básicos para manipulação e tratamento dos dados, como:

- *connection* - Criação de conexões com banco de dados;
- *transformers* - Transformação de dados;

[10] https://airflow.apache.org/
[11] https://powerbi.microsoft.com/pt-br/

- *table* - Manipulação de tabelas;
- *file* - Manipulação de arquivos;
- *producers* - Extração de dados;
- *consumers* - Carga de dados;
- *resources* - Para recursos e configurações de sistema.

3.5.5 Diretório scripts

O diretório ”*scripts*/” contém os *scripts* Python [12] que são *templates* de fluxos de ETL que generalizam a implementação dos componentes do DW utilizando as ferramentas contidas no módulo ”src” e funcionam de acordo com os parâmetros fornecidos pelo usuário. Esses *templates* podem ser utilizados para construir os seguintes componentes:

- Dimensões de mudança lenta do tipo 0;
- Dimensões de mudança lenta do tipo 1;
- Fato;
- Stages

Na Figura 16 é possível identificar que existe um subdiretório ”dw” no qual contém os *scripts* responsáveis por carregar as tabelas de dimensões e a tabela fato do *Data Mart*. E também um diretório ”stg” destinado aos *scripts* de ETL das *stages*.

3.5.6 Diretório dags

A ferramenta utilizada, Apache Airflow [13], organiza as tarefas do fluxo de dados em DAGs, como representado na Figura 10. Desta forma, o diretório ”dags” foi criado com o objetivo de organizar os *scripts* das tarefas. Cada vértice do grafo é um processo invocado por meio de um operador. Os operadores são classes que o Apache Airflow [14] utiliza para gerenciar os *scripts*.

[12] https://www.python.org/
[13] https://airflow.apache.org/
[14] https://airflow.apache.org/

No presente projeto, os *scripts* são configurados por meio de arquivos que utilizam os *templates* apresentados na subseção 3.5.5. Cada arquivo representa uma classe especializada em fabricar operadores específicos para cada tarefa do fluxo de ETL, prontos para serem utilizados no contexto de uma ou mais DAGs. Por fim, foi desenvolvida uma DAG principal, apresentada na Figura 17, que organiza todas as tarefas em um fluxo único.

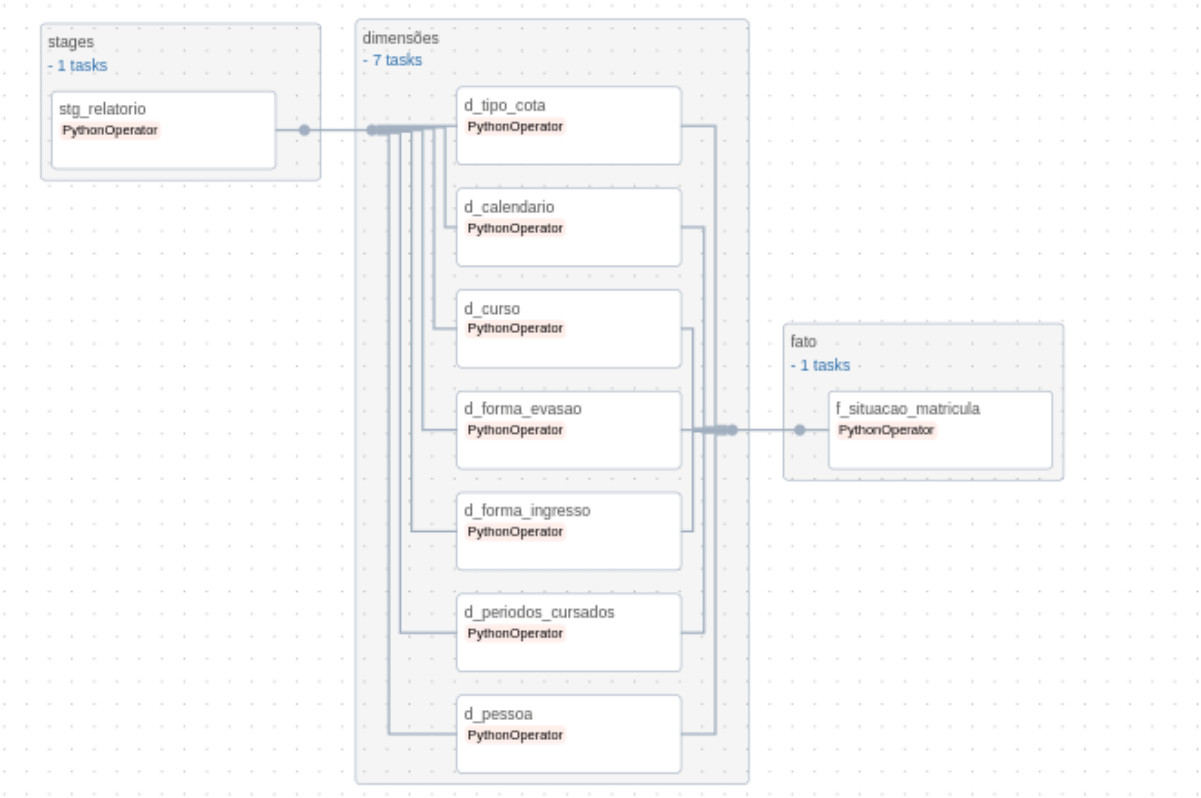

Figura 17: DAG Evasão.
Fonte: O Autor.

4 RESULTADOS OBTIDOS

Após a conclusão do processo de ETL e a disponibilização dos dados no *Data Mart*, foi empregada a ferramenta MIcrosoft Power BI para desenvolver *dashboards* interativos destinados aos usuários. Essa ferramenta possibilita consultas mediante a aplicação de filtros, permitindo a apresentação dos dados em diferentes níveis de granularidade, conforme a necessidade da análise.

O *dashboard* desenvolvido foi dividido em quatro visões de análise:

- Indicadores Gerais;
- Indicadores Acadêmicos;
- Indicadores Sociais;
- Indicadores Regionais;

Todas as visões construídas seguem o seguinte padrão de *design*:

- Cabeçalho - Contém o brasão da UFES, título da página e botões de navegação de páginas;
- Filtros - Sessão de filtros para o usuário interagir e produzir diversas combinações das informações. A aplicação desses filtros impactam em todas as páginas do *dashboard*.
- Área de cartões - Os cartões mostram um resumo dos dados apresentados nos gráficos, classificados por: Total de Ingressantes, Total de Evadidos, Total de Matriculados, Total de Diplomados.
- Área dos gráficos - Área destinada para apresentar os gráficos desenvolvidos.
- Rodapé - Apresenta a última data de atualização da tabela fato do *Data Mart*.

4.1 Indicadores Gerais

O painel de Indicadores Gerais (Figura 18) tem objetivo de apresentar uma visão geral dos dados. Além dos cartões padrões do *design*, essa página possui cartões com o total de cotistas e não cotistas, além das taxas:

Taxa de Evasão = Total de Evadidos / Total de Ingressantes

Taxa de Conclusão = Total de Diplomados / Total de Ingressantes

O gráfico "Evasões por Períodos Cursados" apresenta a quantidade de alunos, classificados por gênero, que evadiram em cada período. O objetivo do gráfico é apresentar o número de períodos cursados pelos discentes até a evasão. Por meio deste, pode-se observar qual o período de maior predominância de evasão.

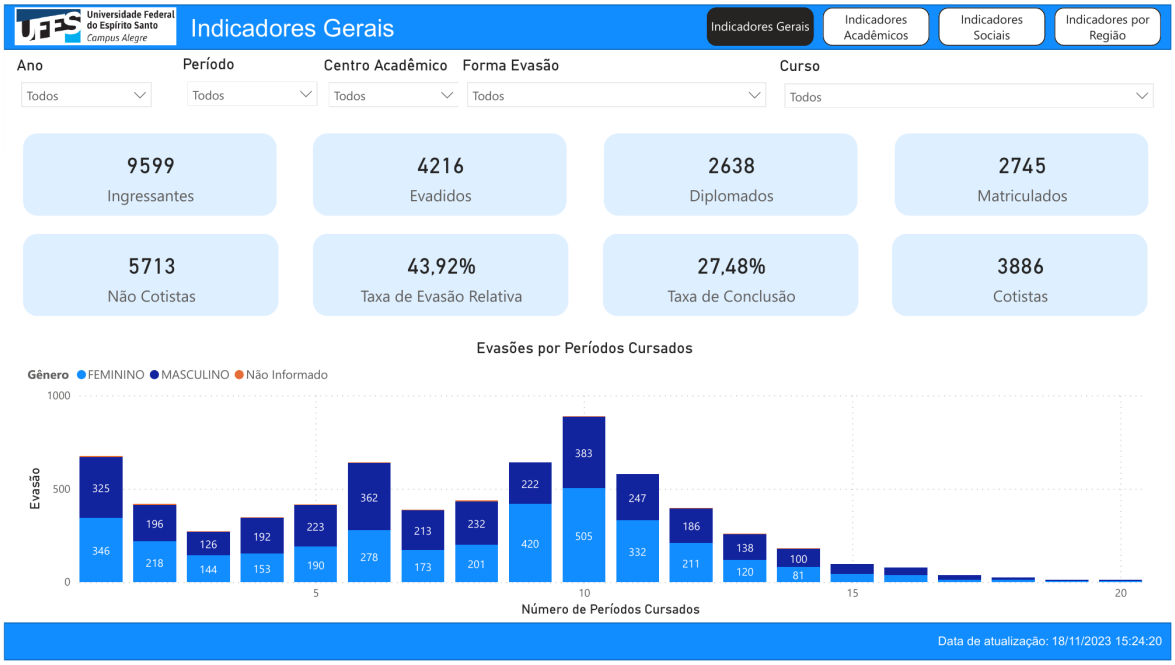

Figura 18: Página Indicadores Gerais.
Fonte: O Autor.

Os filtros apresentados na parte superior da página podem selecionar um único valor ou um conjunto de valores, logo é possível combinar as informações em diferentes níveis de detalhes. Ou seja, é possível visualizar os dados de uma forma mais geral, analisar a situação da IES inteira em um período de tempo, ou analisar apenas uma turma específica.

Na Figura 18, nota-se que na UFES, de 2009/01 a 2023/01, houve um total de 9599 ingressantes, sendo 3886 cotistas e 5713 não cotistas. Dentre esses ingressantes, registrou-se 4216 casos de evasão, 2638 de diplomação e 2745 ainda matriculados. A diferença substancial entre o número total de ingressantes e de diplomados resulta em uma taxa de conclusão de 27,48%. A evasão, representando quase a metade dos ingressantes, resulta em uma taxa de 43,92%.

Para analisar a concentração da evasão, a Figura 19 mostra o resultado ao retirar o valor ”Formado” na categoria ”Forma de Evasão”, apresentando apenas alunos que evadiram. Observa-se uma concentração da evasão no primeiro período, totalizando 669 casos, diminuindo nos períodos seguintes e aumentando novamente no quarto. Além disso, percebe-se que até o terceiro período há uma predominância de evasão no gênero feminino, mas essa dinâmica se inverte a partir do quarto período.

Figura 19: Gráfico Evasão por Período Cursados desconsiderando os alunos diplomados.
Fonte: O Autor.

A concentração da evasão nos primeiros períodos pode ser atribuída a diversos fatores, conforme demonstrado por Sales Junior (2013) em seu trabalho. Entre esses fatores, incluem-se a falta de clareza sobre os objetivos do curso, levando à frustração e ao possível abandono, e também a utilização de cursos com menor concorrência como uma espécie de ”ponte” para a transferência para outro curso.

4.2 Indicadores Acadêmicos

Esse painel tem o objetivo de apresentar a relação da evasão com os detalhes da matrícula. A Figura 20 apresenta o painel Indicadores Acadêmicos onde é composto pelos seguintes gráficos:

- Evasão por Tipo de cota - Permite observar a distribuição da evasão por modalidade de matrícula;
- Evasão por Forma de Ingresso - Apresentar a forma de ingresso dos alunos evadidos. Nesse gráfico é possível fazer uma análise comparativa da evasão entre turmas que ingressam pelo SISU e Vestibular;
- Composição da Evasão - Permite analisar os motivos de evasão;
- Evasão por Períodos cursados - Apresenta a quantidade de evasão por período, mas sem classificação de gênero;

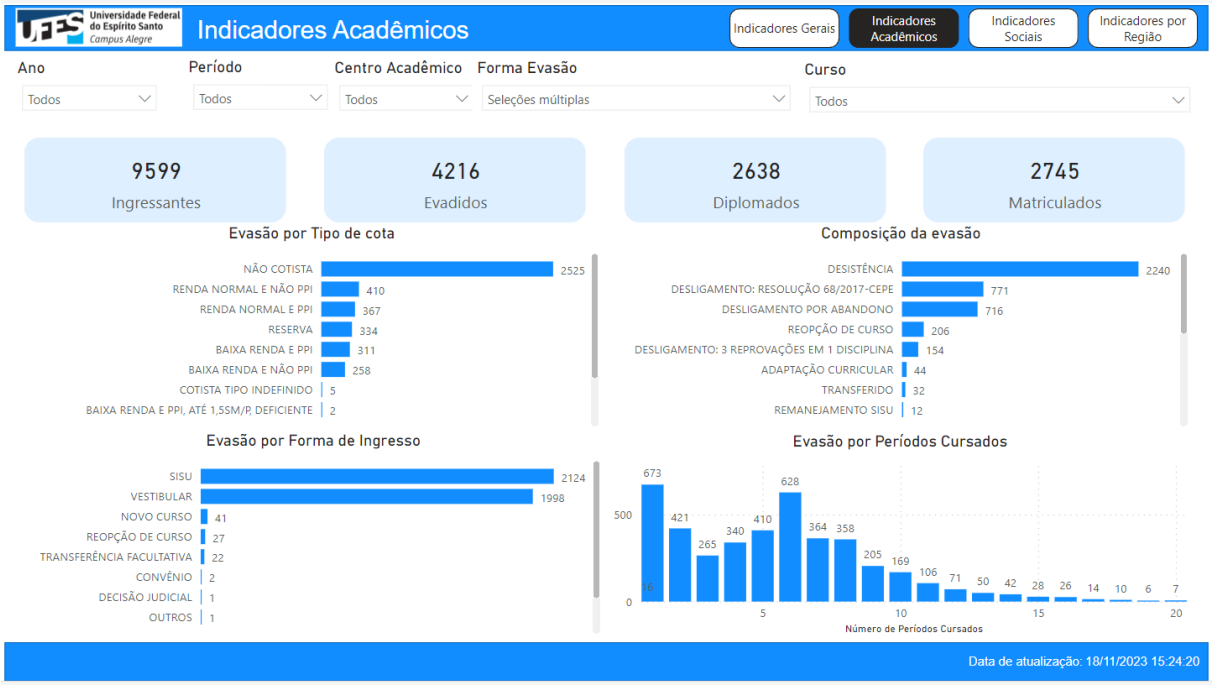

Figura 20: Página Indicadores Acadêmicos.
Fonte: O Autor.

A Figura 21 apresenta o gráfico ”Composição da Evasão” onde é possível observar que a predominancia da evasão ocorre por motivo de desistência, alcançando um total de 2240, além disso é possível observar que a quantidade de evasões por motivos de desligamento também é relevante no cenário da UFES. Podendo estar relacionadas a adaptação do aluno com a grade curricular dos cursos ou outros fatores internos e externos.

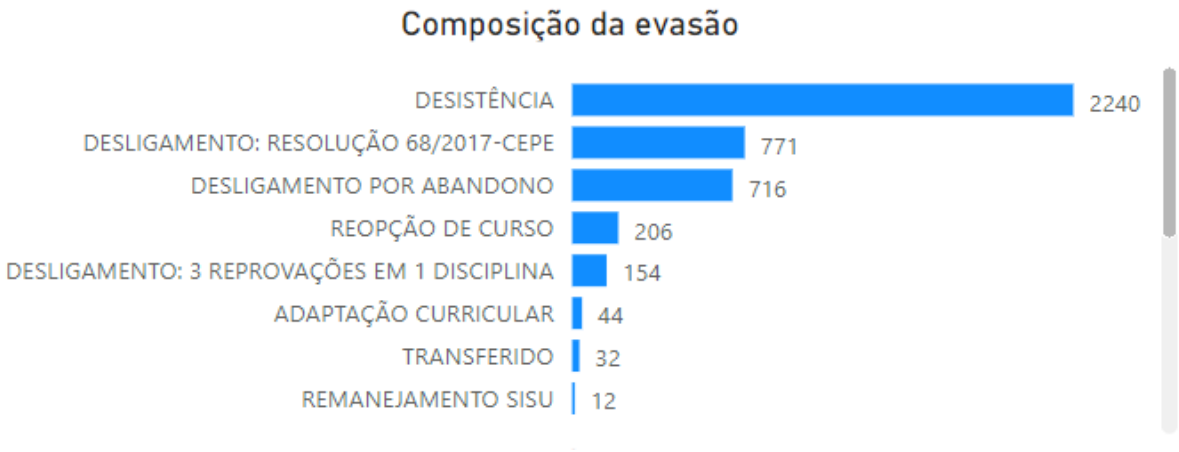

Figura 21: Gráfico Composição da Evasão.
Fonte: O Autor.

No gráfico da ”Evasão por Tipo de Cota” (Figura 22), é evidente que a quantidade de evasão é maior entre os Não cotistas, o que é esperado, dado o maior número de vagas oferecidas para esse grupo. Além disso, observa-se que a maioria das evasões está concentrada em modalidades de cotas relacionadas à renda. Conforme destacado no estudo de Sales Junior (2013), um dos principais motivos pode estar associado à questão financeira, já que muitos alunos precisam conciliar a rotina acadêmica com o trabalho para complementar o orçamento, resultando em sobrecarga para esses estudantes.

Figura 22: Gráfico Evasão por Tipo de Cota .
Fonte: O Autor.

A Figura 23 apresenta a ”Evasão por Forma de Ingresso”, na qual é possível observar que o Sistema de Seleção Unificada (SISU) registrou uma quantidade de evasões superior ao Vestibular, mesmo após a adoção integral do SISU em substituição ao

vestibular no ano de 2016. Entretanto, é importante ressaltar que o sistema do SISU flexibilizou o ingresso de estudantes nas IES no Brasil, o que também pode ter contribuído para o aumento da evasão. Isso se deve ao fato de que a nota do Exame Nacional do Ensino Médio (ENEM) pode ser reutilizada para ingresso em outros cursos por meio do processo de convocação SISU, que ainda ocorre mais de uma vez no ano.

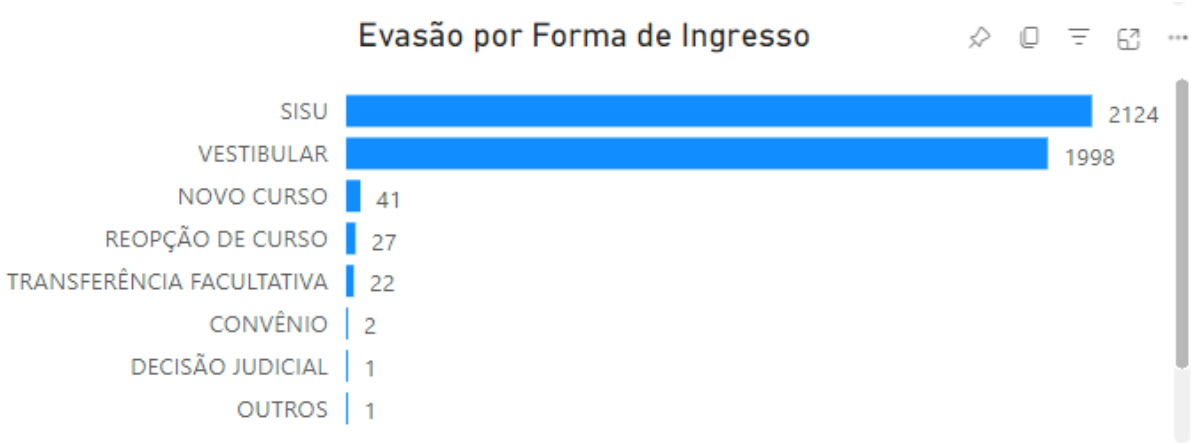

Figura 23: Evasão por Forma de Ingresso.
Fonte: O Autor.

4.3 Indicadores Sociais

Esse painei, apresentado na Figura 24, apresenta novamente os gráficos Evasão por ”Períodos Cursados” e ”Composição da Evasão” apresentados no painel de Indicadores Acadêmicos, mas o foco dessa visão é apresentar evasão por etnia e deficiência declarada na matrícula.

- Evasão por Etnia - Distribuição da evasão por Etnia declarada;
- Evasão por Deficiência - Distribuição da evasão por tipo de Deficiência declarada;

Figura 24: Página Indicadores Sociais.
Fonte: O Autor.

Na Figura 25, é apresentado o gráfico ”Evasão por Etnia”, no qual se destaca a evasão entre pessoas classificadas como Brancas, totalizando 2256, seguida por Parda, com 1362, e Preta, com 406. A história do Brasil é marcada por uma longa luta por direitos e espaço das minorias étnicas, as quais, até os dias atuais, enfrentam discriminação. Nesse contexto, é fundamental ressaltar a importância do estudo da evasão relacionada à inclusão desses grupos no ambiente das IES, visando assegurar também a diplomação e o pleno acesso educacional a todos os grupos étnicos.

Figura 25: Gráfico Evasão por Etnia.
Fonte: O Autor.

O gráfico ”Evasão por Deficiência”, apresentado na Figura 26, revela que a evasão está completamente concentrada entre estudantes sem deficiência, totalizando 4182 casos, enquanto apenas 34 evasões são registradas entre estudantes com alguma forma de deficiência declarada. Este gráfico evidencia que, apesar de serem minoria, alunos com determinados tipos de deficiência podem enfrentar desafios que limitam seu processo de aprendizagem. Nesse contexto, a análise desse indicador torna-se de extrema importância para as IES, uma vez que é fundamental garantir mecanismos de inclusão para assegurar educação de qualidade acessível a todos, independentemente das diferenças, promovendo um ambiente educacional mais inclusivo.

Figura 26: Gráfico Evasão por Deficiência.
Fonte: O Autor.

4.4 Indicadores Regionais

Esta seção apresenta a análise da evasão relacionada à região de naturalidade e de origem dos alunos por meio de mapas. Composta por dois gráficos, a página contempla:

- Evasão por Município de Naturalidade - Apresenta a distribuição da evasão conforme o município de naturalidade dos alunos;
- Evasão por Município de Origem - Exibe a distribuição da evasão conforme o município de origem dos alunos.

A Figura 27 apresenta o paniel de ”Indicadores Regionais”, onde os usuários podem interagir com os mapas e os filtros para criar combinações personalizadas, resultando

em uma visão dinâmica sobre a evasão por região de naturalidade e origem dos alunos.

Nos dois gráficos apresentados na Figura 27, é evidente que a UFES de Alegre atrai estudantes de praticamente todas as regiões do Brasil, sendo a região Sudeste a de maior influência.

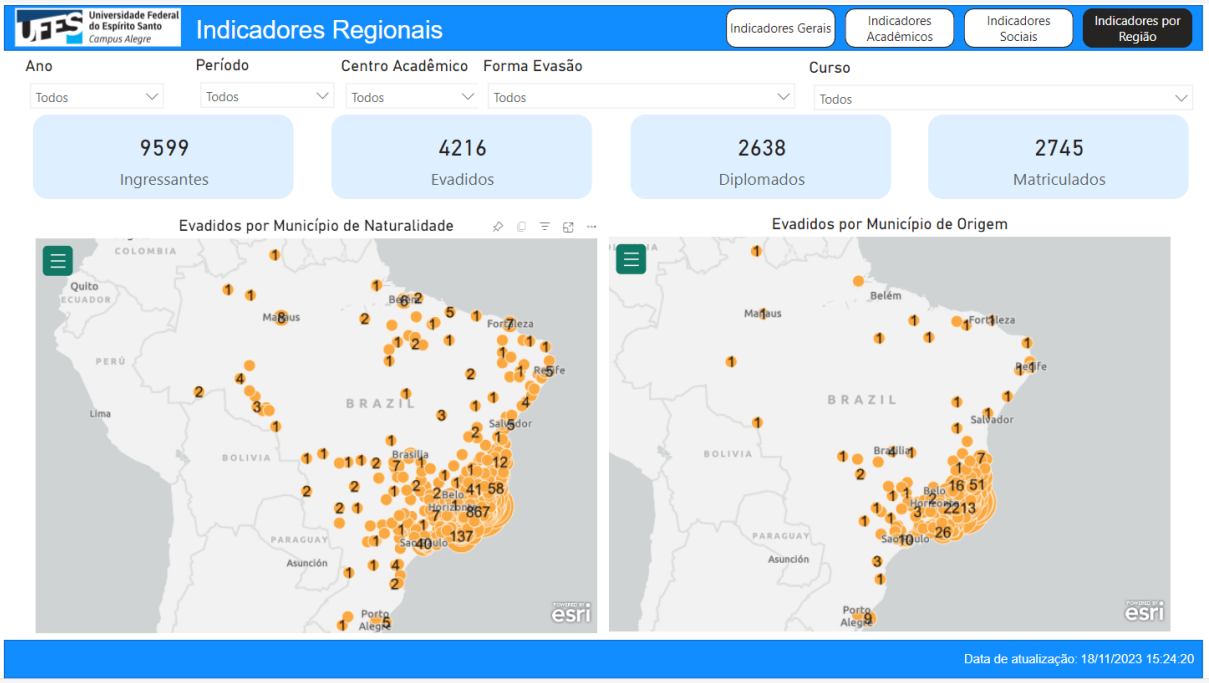

Figura 27: Página Indicadores Regionais.
Fonte: O Autor.

Os gráficos da Figura 28, apresentam uma concentração significativa de alunos naturais e provenientes da região da Grande Vitória e do Sul do Espírito Santo. Além disso, observa-se uma considerável quantidade de alunos vindos de diversos municípios do estado de Minas Gerais, especialmente aqueles próximos à fronteira do Espírito Santo.

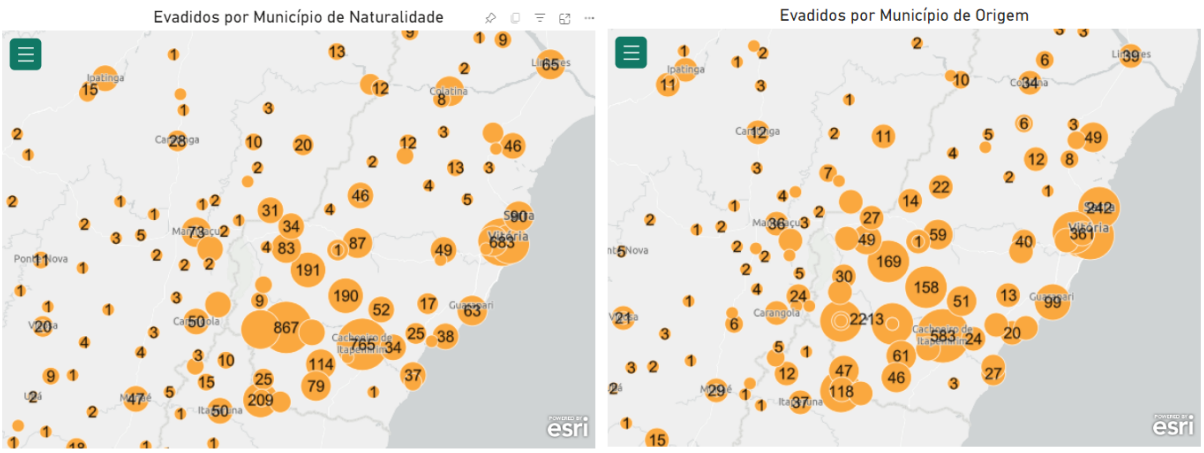

Figura 28: Gráficos de Evasão por Municípo de Naturalidade e Evasão por Município de Origem .
Fonte: O Autor.

Essas informações poderiam ser mais precisas se relacionadas a indicadores geográficos regionais, como o Índice de Desenvolvimento Humano, a Renda per Capita e a escola de origem dos alunos. Tal abordagem permitiria uma compreensão mais aprofundada de como a evasão está relacionada aos fatores geográficos e econômicos de cada região. Isso possibilitaria à universidade investigar quais regiões necessitam de maior atenção para mitigar esse problema. No entanto, os dados utilizados não abrangem tais informações para contemplar essa visão mais ampla.

5 CONCLUSÕES

O objetivo de um curso oferecido por uma instituição de ensino é suprir as necessidades locais e regionais, capacitando e qualificando a população, e consequentemente contribuindo para o avanço da qualidade de vida na sociedade. Nesse contexto, é crucial o acompanhamento do desempenho do curso, visando assegurar a permanência e a conclusão dos estudantes.

O BI agrega um valor substancial no contexto educacional, uma vez que a apresentação e análise dos dados permitem visualizar se a IES está conseguindo atender às metas propostas na legislação, fornecendo informações estratégicas para embasar a tomada de decisões assertivas na resolução de questões complexas, como a evasão estudantil. Além disso, é imprescindível que as metodologias e os dados utilizados retratem fielmente a realidade, possibilitando a avaliação da eficácia do sistema e propiciando a identificação de áreas de melhoria, o desenvolvimento de novos cursos e a compreensão dos contextos nos quais a evasão ocorre.

O presente estudo concentrou-se na aplicação de metodologias voltadas ao desenvolvimento de rotinas de ETL para alimentação de um modelo de dados especializado na análise da evasão estudantil no *Campus* Alegre da Universidade Federal do Espírito Santo. Isso permitiu a construção de relatórios interativos por meio de ferramentas de BI, facilitando a identificação de possíveis padrões e perfis associados ao problema.

Portanto, o trabalho apresentado representa apenas o início de uma série de ações e debates que podem ser conduzidos para identificar as principais causas do problema e propor soluções que aprimorem a eficácia do sistema de ensino.

6 TRABALHOS FUTUROS

Recomenda-se a integração direta do fluxo de ETL com a base de dados do SIE [1] e outras fontes de dados utilizadas pela UFES. Esse procedimento visa ampliar o detalhamento das informações e enriquecer a análise de dados. Essa integração possibilitará não apenas uma compreensão mais aprofundada dos dados existentes, mas também permitirá a extensão do modelo de dados para outras áreas da IES, contribuindo para otimizar a gestão como um todo.

Além disso, sugere-se a realização de um estudo por meio de análise qualitativa dos dados, estabelecendo correlações entre os resultados obtidos neste trabalho e informações provenientes de entrevistas, questionários e pesquisas de satisfação realizadas com alunos e ex-alunos. Essa abordagem visa aprimorar a eficácia das tomadas de decisões, possibilitando uma compreensão mais ampla dos motivos que influenciam a evasão estudantil e identificando possíveis estratégias para melhorar a retenção e o sucesso dos estudantes na instituição.

Relacionar dados de evasão com disciplinas e analisar os índices de reprovação em cada uma delas.

[1] https://npd.ufes.br/sistema-de-informa%C3%A7%C3%A3o-para-o-ensino-sie

REFERÊNCIAS

ANAND, N.; KUMAR, M. An overview on data quality issues at data staging etl. In: CITESEER. *Int. Conf. on Advances in Signal Processing and Communication*. [S.l.], 2013. Citado na página 29.

ANDIFES, A. Sesu/mec. *Diplomação, retenção e evasão nos cursos de graduação em instituições de Ensino Superior públicas. Brasília, DF*, 1996. Citado na página 11.

APACHE, S. F. *Apache Airflow Documentation, Version: 2.3.3*. 2022. Disponível em: <https://airflow.apache.org/docs/apache-airflow/stable/index.html>. Citado na página 25.

BRASIL. *Ministério do Planejamento. Manual para cálculo dos indicadores de gestão das Instituições da Rede Federal de Educação Profissional, Científica e Tecnológica – 2.0*. 2016. Disponível em: <http://portal.mec.gov.br/index.php?option=com_docman&view=download&alias=36901-manual-de-indicadores-da-rfepct-pdf&category_slug=abril-2016&Itemid=30192>. Citado na página 22.

DOCKER. *Docker, Version: 24.0.7*. 2023. Disponível em: <https://docs.docker.com/get-started/>. Citado na página 26.

GAMMA, E. et al. *Design Patterns: Elements of Reusable Object-Oriented Software*. 1. ed. Addison-Wesley Professional, 1994. ISBN 0201633612. Disponível em: <http://www.amazon.com/Design-Patterns-Elements-Reusable-Object-Oriented/dp/0201633612/ref=ntt_at_ep_dpi_1>. Citado na página 35.

GONÇALVES, L. M. *Uma plataforma de business intelligence para analisar a retenção e evasão do IFMT*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2021. Citado 6 vezes nas páginas , 11, 18, 22, 23 e 24.

INMON, W. H. *Building the Data Warehouse, Fourth Edition*. [S.l.]: Wiley Publishing Inc., 2005. Citado 4 vezes nas páginas , 13, 15 e 16.

KIMBALL, M. R. R. *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. [S.l.]: John Wiley & Sons, 2013. Citado 12 vezes nas páginas , 11, 12, 13, 14, 15, 16, 17, 20, 28, 33 e 34.

MENDES, F. R. D. M. *Proposta de um Data Warehouse para apoio à tomada de decisão sobre evasão institucional em uma instituição federal de ensino superior*. Dissertação (Monografia) — Universidade Federal De Itajubá, Itajubá - MG, 2020. Citado 6 vezes nas páginas , 11, 18, 20, 21 e 22.

MUSA, S. et al. Success factors for business intelligence systems implementation in higher education institutions–a review. In: SPRINGER. *International Conference of Reliable Information and Communication Technology*. [S.l.], 2018. p. 322–330. Citado na página 11.

PALHARES, I. *Universidades públicas tiveram queda de 18,8% no número de concluintes*. 2022. Disponível em: <https://www1.folha.uol.com.br/educacao/2022/02/universidades-publicas-tiveram-queda-de-188-no-numero-de-concluintes.shtml>. Citado na página 11.

PANDAS, d. t. *pandas documentation, Version: 1.4.3*. 2022. Disponível em: <https://pandas.pydata.org/docs/getting_started/overview.html>. Citado na página 24.

PEÇANHA, P. N. *PNP 2022 (Ano Base 2020)*. 2022. Disponível em: <https://www.plataformanilopecanha.org/>. Citado na página 22.

POSTGRESQL, G. D. G. *PostgreSQL Documentation, Version: 14*. 2022. Disponível em: <https://www.postgresql.org/docs/current/index.html>. Citado na página 26.

ROWEN, W. et al. An analysis of many-to-many relationships between fact and dimension tables in dimensional modeling. In: *International Workshop on Design and Management of Data Warehouses (DMDW 2001), Interlaken Switzerland*. [S.l.: s.n.], 2001. p. 1–13. Citado na página 16.

SALES JUNIOR, J. S. *Uma análise estatística dos fatores de evasão e permanência de estudantes de graduação presencial da UFES*. Tese (Doutorado) — Dissertação (Mestrado Profissional em Gestão Pública)–Centro de Ciências Jurídicas Econômicas da Universidade Federal do Espírito Santo, 2013. Citado 5 vezes nas páginas , 18, 19, 41 e 43.

SAMARAKKODY, A.; AMARATUNGA, D.; HAIGH, R. Technological innovations for enhancing disaster resilience in smart cities: A comprehensive urban scholar's analysis. *Sustainability*, MDPI, v. 15, n. 15, p. 12036, 2023. Citado na página 14.

SOLUTIONS, M. *Python: 7 Important Reasons Why You Should Use Python.* 2017. Disponível em: <https://medium.com/@mindfiresolutions.usa/python-7-important-reasons-why-you-should-use-python-5801a98a0d0b>. Citado na página 24.

UFES. *Sistema de Informação para o Ensino (SIE)*. 2022. Disponível em: <https://npd.ufes.br/sistema-de-informa%C3%A7%C3%A3o-para-o-ensino-sie>. Citado na página 30.

APÊNDICE A – *SCRIPT* DA *VIEW* TRATAMENTO DAS *STAGES*

```
create or replace view  stg.v_ds_stg_relatorio as (
    with stage as (
        select
            --pessoa
            case "SEXO"
                when 'M' then 'MASCULINO'
                when 'F' then 'FEMININO'
                when 'N' then 'Não Informado'
                else "SEXO"
            end ds_genero
            , to_date("DT_NASCIMENTO", 'yyyy-mm-dd') dt_nascimento
            , upper(trim("NOME_ALUNO")) no_pessoa
            , upper(trim("ESTADO_CIVIL")) ds_estado_civil
            , upper(trim("ETNIA")) ds_etinia
            , upper(trim("DEFICIENCIA")) ds_deficiencia
            , upper(trim("NATURALIDADE")) no_municipio_naturalidade
            , upper(trim("UF_NATURALIDADE")) cd_uf_naturalidade
            , upper(trim("RG_ESTADO")) nu_rg_estado
            , upper(trim("RG_ORGAO_EMISSOR")) no_rg_orgao_emissor
            , upper(trim("RG")) nu_rg
            , upper(trim("CPF")) nu_cpf
            , upper(trim("UF")) cd_uf
            , upper(trim("PAIS")) no_pais
            , upper(trim("ESTADO")) no_estado
            , upper(trim("MUNICIPIO")) no_municipio_origem
            , upper(trim("BAIRRO")) no_bairro
```

```
, upper(trim("COMPLEMENTO")) ds_complemento
, upper(trim("NUMERO")) nu_residencia
, upper(trim("RUA")) ds_logradouro
, upper(trim("TIPO_LOGRADOURO")) ds_tipo_logradouro
, upper(trim("CEP")) nu_cep
, upper(trim("DESCR_MAIL")) ds_email
--curso
, "COD_CURSO" cd_curso
, upper(trim("NOME_CURSO")) no_curso
, "NUM_PERIODOS" nu_periodos
, "NUM_MAX_PERIODOS" nu_maximo_periodos
, "CH_TOTAL_CURSO" nu_ch_total_curso
, upper(trim("SITUACAO_VERSAO")) ds_situacao_versao
, upper(trim("NUM_VERSAO")) nu_versao
--matricula
, regexp_replace(upper(trim("MATR_ALUNO")), 'X','') nu_matricula
, "ANO_INGRESSO" nu_ano_ingresso
, upper(trim("FORMA_INGRESSO")) ds_forma_ingresso
, upper(trim("PERIODO_INGRESSO")) ds_periodo_ingresso
, "TURNO_ALUNO_ITEM"  nu_turno_aluno
, upper(trim("TURNO")) ds_turno
, coalesce(cast("CRA" as float), 0) nu_cra
, coalesce(cast("CRN" as float), 0) nu_crn
--cota
, case upper(trim("COTISTA"))
    when 'S' then 'Cotista'
    when 'N' then 'Não Cotista'
    else upper(trim("COTISTA"))
end fl_cotista
,upper(trim("TIPO_COTA")) ds_tipo_cota
--evasao
, upper(trim("PERIODO_EVASAO")) ds_periodo_evasao
, upper(trim("FORMA_EVASAO")) ds_forma_evasao
, "ANO_EVASAO" nu_ano_evasao
, case upper("NOME_CENTRO")
```

```
            when 'CENTRO DE CIÊNCIAS EXATAS, NATURAIS E DA SAÚDE' then 'CCENS'
            when 'CENTRO DE CIÊNCIAS AGRÁRIAS E ENGENHARIAS' then 'CCAE'
        end no_centro_academico
    from stg.stg_relatorio sr
)
select distinct
    no_pessoa
    , ds_estado_civil
    , ds_etinia
    , ds_deficiencia
    , no_municipio_naturalidade
    , cd_uf_naturalidade
    , nu_rg_estado
    , no_rg_orgao_emissor
    , nu_rg
    , nu_cpf
    , no_pais
    , no_estado
    , no_municipio_origem
    , nu_cep
    , ds_email
    , cd_curso
    , no_curso
    , nu_periodos
    , nu_maximo_periodos
    , nu_ch_total_curso
    , ds_situacao_versao
    , nu_versao
    , nu_matricula
    , nu_ano_ingresso
    , ds_forma_ingresso
    , ds_periodo_ingresso
    , nu_turno_aluno
    , ds_turno
    , nu_cra
```

```
        , nu_crn
        , fl_cotista
        , ds_tipo_cota
        , ds_periodo_evasao
        , ds_forma_evasao
        , nu_ano_evasao
        , ds_genero
        , dt_nascimento
        , no_centro_academico
        , cd_uf
    from stage
);
```

APÊNDICE B – *SCRIPT* DA *VIEW* DO *DATASOURCE*

```
create or replace view dw.v_ds_evasao as (
    select
        dp.no_pessoa
        , dp.ds_genero
        , dp.dt_nascimento
        , dp.ds_deficiencia
        , dp.ds_etinia
        , dp.ds_estado_civil
        , dp.no_municipio_origem
        , dp.no_municipio_naturalidade
        , dp.cd_uf
        , dc.cd_curso
        , dc.no_curso
        , dfe.ds_forma_evasao
        , dfi.ds_forma_ingresso
        , dtc.ds_tipo_cota
        , f.nu_matricula
        , d_ingresso.nu_ano_referencia  nu_ano_ingresso
        , d_ingresso.ds_periodo_referencia  ds_periodo_ingresso
        , f.nu_cra
        , f.nu_crn
        , f.fl_cotista
        , d_evasao.nu_ano_referencia  nu_ano_evasao
        , d_evasao.ds_periodo_referencia  ds_periodo_evasao
        , f.nu_periodos
        , f.nu_maximo_periodos
        , f.dt_carga
        , case when dfe.ds_forma_evasao not in ('SEM EVASÃO', 'FORMADO') then 1
```

```
            else 0
        end fl_evasao
        , case when dfe.ds_forma_evasao not in ('SEM EVASÃO') then 1
            else 0
        end fl_evasao_diplomados
        , case when dfe.ds_forma_evasao = 'FORMADO' then 1
            else 0
        end fl_diplomado
        , dc.no_centro_academico
        , dpc.nu_periodos_cursados
    from dw.f_situacao_matricula f
    left join dw.d_pessoa dp
        on f.sk_pessoa = dp.sk_pessoa
    left join dw.d_curso dc
        on f.sk_curso = dc.sk_curso
    left join dw.d_forma_evasao dfe
        on f.sk_forma_evasao = dfe.sk_forma_evasao
    left join dw.d_forma_ingresso dfi
        on f.sk_forma_ingresso = dfi.sk_forma_ingresso
    left join dw.d_tipo_cota dtc
        on f.sk_tipo_cota = dtc.sk_tipo_cota
    left join dw.d_calendario d_evasao
        on f.sk_ano_evasao = d_evasao.sk_calendario
    left join dw.d_calendario d_ingresso
        on f.sk_ano_ingresso  = d_ingresso.sk_calendario
    left join dw.d_periodos_cursados dpc
        on f.sk_periodos_cursados = dpc.sk_periodos_cursados
);
```

APÊNDICE C – COLUNAS DA *STAGE* ”STG_RELATORIO”

1. MATR_ALUNO
2. NOME_ALUNO
3. SEXO
4. DT_NASCIMENTO
5. ESTADO_CIVIL
6. ETNIA
7. COD_CURSO
8. NOME_CURSO
9. ANO_INGRESSO
10. FORMA_INGRESSO
11. PERIODO_INGRESSO
12. TURNO_ALUNO_ITEM
13. TURNO
14. ANO_EVASAO
15. PERIODO_EVASAO
16. FORMA_EVASAO
17. NUM_PERIODOS
18. NUM_MAX_PERIODOS

19. CH_TOTAL_CURSO
20. DESCR_MAIL
21. COTISTA
22. CEP
23. TIPO_LOGRADOURO
24. RUA
25. NUMERO
26. COMPLEMENTO
27. BAIRRO
28. MUNICIPIO
29. ESTADO
30. PAIS
31. UF
32. CPF
33. RG
34. RG_ORGAO_EMISSOR
35. RG_ESTADO
36. CRN
37. CRA
38. TIPO_COTA
39. NATURALIDADE
40. UF_NATURALIDADE
41. DEFICIENCIA
42. NUM_VERSAO

43. SITUACAO_VERSAO
44. QTDE_TRANCAMENTOS
45. NOME_CURSO_DIPLOMA
46. MODALIDADE
47. NOME_CENTRO
48. NOME_CAMPUS
49. NIVEL_CURSO

APÊNDICE D – DICIONÁRIO DE DADOS

Nome Tabela	Coluna	Descrição
d_curso	sk_curso	Surrogate key da dimensão curso (Coluna gerada)
d_curso	cd_curso	Código do curso
d_curso	no_curso	Nome do curso
d_curso	no_centro_academico	Nome do centro acadêmico
d_curso	dt_carga	Data de atualização do registro
d_tipo_cota	sk_tipo_cota	Surrogate key da dimensão tipo de cota (Coluna gerada)
d_tipo_cota	ds_tipo_cota	Descrição da forma de ingresso
d_tipo_cota	dt_carga	Data de atualização do registro
d_forma_evasao	sk_forma_evasao	Surrogate key da dimensão forma de evasão (Coluna gerada)
d_forma_evasao	ds_forma_evasao	Descrição da forma de evasão
d_forma_evasao	dt_carga	Data de atualização do registro
d_forma_ingresso	sk_forma_ingresso	Surrogate key da dimensão forma de ingresso (Coluna gerada)
d_forma_ingresso	ds_forma_ingresso	Descrição da forma de ingresso
d_forma_ingresso	dt_carga	Data de atualização do registro
d_pessoa	sk_pessoa	Surrogate key da dimensão pessoa (Coluna gerada)
d_pessoa	nu_cpf	Número de CPF
d_pessoa	no_pessoa	Nome da pessoa
d_pessoa	ds_genero	Descrição do gênero declarado
d_pessoa	dt_nascimento	Data de nascimento
d_pessoa	ds_deficiencia	Descrição da deficiência declarada
d_pessoa	ds_etinia	Descrição da etinia declarada
d_pessoa	ds_estado_civil	Descrição do estado civíl
d_pessoa	no_municipio_naturalidade	Nome do muncípio de naturalidade
d_pessoa	no_municipio_origem	Nome do muncípio de origem
d_pessoa	cd_uf	Código da Unidade Federativa
d_pessoa	dt_carga	Data de atualização do registro

Nome Tabela	Coluna	Descrição
d_periodos_cursados	sk_periodos_cursados	Surrogate key da dimensão de periodos cursados (Coluna gerada)
d_periodos_cursados	nu_periodos_cursados	Quantidade de períodos cursados até a evasão
d_periodos_cursados	dt_carga	Data de atualização do registro
d_calendario	sk_calendario	Surrogate key da dimensão de períodos cursados (Coluna gerada)
d_calendario	nu_ano_referencia	Número do ano
d_calendario	ds_periodo_referencia	Descrição do período
d_calendario	dt_carga	Data de atualização do registro
f_situacao_matricula	sk_pessoa	Surrogate key da dimensão pessoa (Coluna gerada)
f_situacao_matricula	sk_tipo_cota	Surrogate key da dimensão tipo de cota (Coluna gerada)
f_situacao_matricula	sk_forma_ingresso	Surrogate key da dimensão forma de ingresso (Coluna gerada)
f_situacao_matricula	sk_ano_evasao	Surrogate key da dimensão calendário, representando o ano de evasão (Coluna gerada)
f_situacao_matricula	sk_ano_ingresso	Surrogate key da dimensão calendário, representando o ano de ingresso (Coluna gerada)
f_situacao_matricula	sk_curso	Surrogate key da dimensão curso (Coluna gerada)
f_situacao_matricula	sk_forma_evasao	Surrogate key da dimensão forma de evasão (Coluna gerada)
f_situacao_matricula	sk_periodos_cursados	Surrogate key da dimensão de periodos cursados (Coluna gerada)
f_situacao_matricula	nu_matricula	Número de matrícula do discente
f_situacao_matricula	nu_cra	Valor do Coeficiente de Rendimento do Aluno
f_situacao_matricula	nu_crn	Valor do Coeficiente de Rendimento Normalizado
f_situacao_matricula	fl_cotista	Flag de identificação de cotista
f_situacao_matricula	nu_periodos	Quantidade de períodos do curso
f_situacao_matricula	nu_maximo_periodos	Quantidade máxima de períodos
f_situacao_matricula	nu_ch_total_curso	Carga horária total do curso
f_situacao_matricula	nu_versao	Versão do curso
f_situacao_matricula	dt_carga	Data de atualização do registro

Nome Tabela	Coluna	Descrição
v_ds_evasao	no_pessoa	Nome da pessoa
v_ds_evasao	ds_genero	Descrição do gênero declarado
v_ds_evasao	dt_nascimento	Data de nascimento
v_ds_evasao	ds_deficiencia	Descrição da deficiência declarada
v_ds_evasao	ds_etinia	Descrição da etnia declarada
v_ds_evasao	ds_estado_civil	Descrição do estado civíl
v_ds_evasao	no_municipio_naturalidade	Nome do município de naturalidade
v_ds_evasao	no_municipio_origem	Nome do município de origem
v_ds_evasao	cd_uf	Código da Unidade Federativa
v_ds_evasao	cd_curso	Código do curso
v_ds_evasao	no_curso	Nome do curso
v_ds_evasao	no_centro_academico	Nome do centro acadêmico
v_ds_evasao	ds_tipo_cota	Descrição da forma de ingresso
v_ds_evasao	ds_forma_evasao	Descrição da forma de evasão
v_ds_evasao	ds_forma_ingresso	Descrição da forma de ingresso
v_ds_evasao	nu_periodos_cursados	Quantidade de períodos cursados até a evasão
v_ds_evasao	nu_ano_ingresso	Número do ano de ingresso
v_ds_evasao	ds_periodo_ingresso	Descrição do período de ingresso
v_ds_evasao	nu_ano_evasao	Número do ano de evasão
v_ds_evasao	ds_periodo_evasao	Descrição do período de evasão
v_ds_evasao	nu_matricula	Número de matrícula do discente
v_ds_evasao	nu_cra	Valor do Coeficiente de Rendimento do Aluno
v_ds_evasao	nu_crn	Valor do Coeficiente de Rendimento Normalizado
v_ds_evasao	fl_cotista	Flag de identificação de cotista
v_ds_evasao	fl_evasao	Flag de identificação de alunos que tiveram alguma forma de evasão
v_ds_evasao	nu_periodos	Quantidade de períodos do curso
v_ds_evasao	nu_maximo_periodos	Quantidade máxima de períodos
v_ds_evasao	fl_evasao	Flag para facilitar a identificação de uma evasão
v_ds_evasao	fl_diplomado	Flag para facilitar a identificação de uma diplomação
v_ds_evasao	fl_evasao_diplomados	Flag para facilitar a identificação de diplomados
v_ds_evasao	dt_carga	Data de atualização do registro

APÊNDICE E – TUTORIAL DE UTILIZAÇÃO

E.1 Requisitos de sistemas

1. Docker versão 24.0.7, *build* afdd53b
2. Sistema Operacional *Linux* x86_64 derivados do Debian

E.2 Configuração de ambiente

Para execução bem sucedida é preciso satisfazer os requisitos de sistemas. Utilizando um sistema operacional Linux baseado em *Debian* é preciso realizar a instalação do serviço Docker [1] para execução dos serviços do Apache Airflow [2].

Tutorial de Instalação do Docker: https://docs.docker.com/engine/install/debian/

E.2.1 Instalação do Sistema

Após a satisfação dos requisitos de sistemas é preciso fazer o *download* do código fonte do projeto.

Em um terminal execute o comando para fazer o *download*:

```
git clone git@github.com:eloy-freitas/TCC2.git && cd TCC2;
```

Existem três *scripts* de configuração do sistema que estão na raiz do diretório do projeto:

1. start.sh - Compila o código fonte e executa os serviços;

[1] https://www.docker.com/
[2] https://airflow.apache.org/

2. stop.sh - Para interrupção dos serviços;
3. nuke.sh - Para parar e remover os serviços;

Todos eles utilizam o arquivo ”config_envs.json” que possui as seguintes variáveis para configuração do sistema:

1. AIRFLOW_WORKSPACE = Pasta de trabalho do airflow;
2. SUBDIRS = Diretórios do airflow;
3. DOCKER_IMAGE_OWNER = Nome da proprietário da imagem do docker;
4. DOCKER_IMAGE_NAME = Nome da imagem;
5. DOCKER_IMAGE_VERSION = Versão da imagem;
6. DOCKER_FILE = caminho do arquivo Dockerfile;
7. DOCKER_COMPOSE_FILE = caminho do arquivo Docker compose;

E.3 Execução

O *script* start.sh é responsável por preparar e iniciar os serviços do sistema. Ele executa as seguintes tarefas:

1. Instalação de dependências do projeto;
2. Criação dos diretórios utilizados pelo ambiente do Airflow;
3. Compilar a imagem do *Airflow* com o código fonte do projeto;
4. Se existir uma execução em andamento, ela é finalizada;
5. Execução do sistema;
6. Configuração das conexões com o banco de dados do *Data Mart*;

Para executar o *script* basta executar:

```
sudo ./start.sh
```

Vale ressaltar que os serviços do *Airflow* utilizam as seguintes portas do *host*:

1. 5432 - Banco de dados de metadados do *Airflow*
2. 10001 - Banco de dados do *Data Warehouse*
3. 8080 - Interface de usuário do *Airflow*
4. 5555 - Orquestrador dos serviços *Airflow*

Após iniciar o sistema é preciso disponibilizar os dados em formato *Microsoft Excel* para que o fluxo de ETL possa alimentar o *Data Mart*. Dessa forma, o arquivo deve ser disponibilizado no diretório airflow/data.

Para executar o fluxo de ETL o usuário precisa acessar o painel de controle do *Airflow* utilizando um navegador de internet acessando o servidor pelo nome de domínio ou endereço IP na porta 8080

Inicialmente o *Airflow* vai pedir autenticação em uma tela de *login*. Utilize o usuário padrão (usuário: ”airflow” e senha: ”airflow”) para acessar.

Após a autenticação a página principal será carregada, onde as DAGs disponíveis serão apresentadas. Para executar a DAG da evasão basta clicar no botão indicado pelo seta vermelha na Figura E.1.

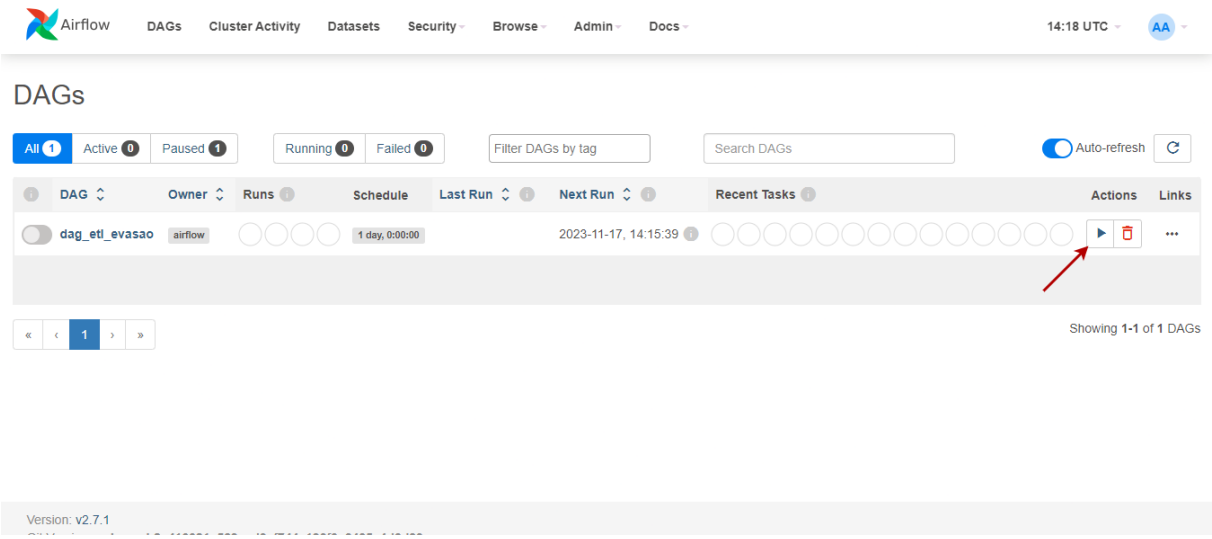

Figura E.1: Página inicial do Apache Airflow.
Fonte: O Autor.

Quando a DAG inicia a execução, o seu estado muda conforme Figura E.1. O usuário pode visualizar os detalhes da execução clicando no nome DAG, acessando o painel mostrado na Figura E.3.

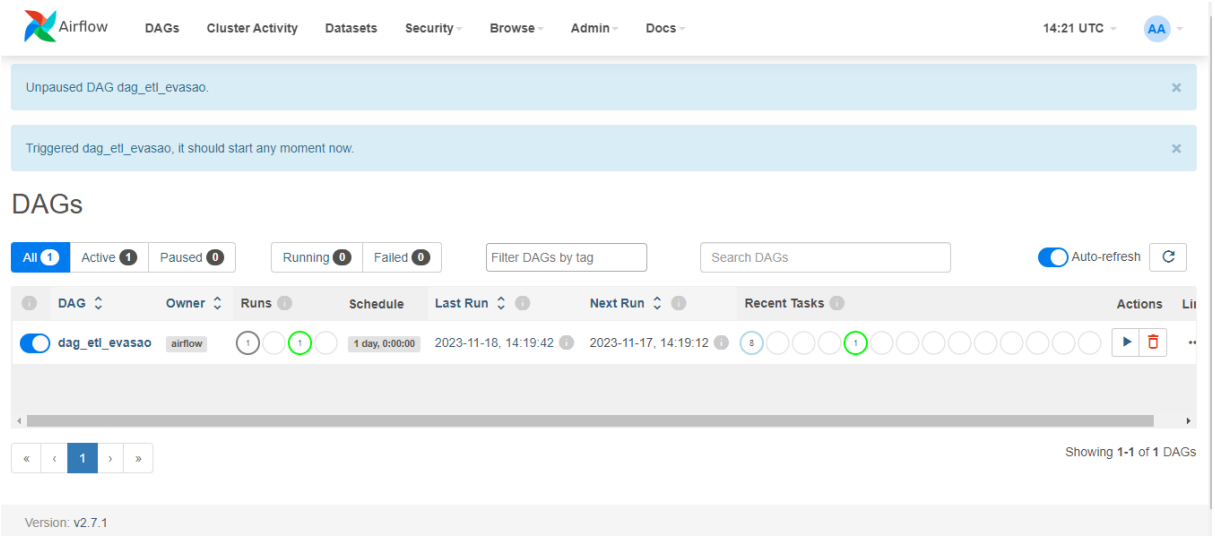

Figura E.2: Página inicial do Apache Airflow execução da DAG.
Fonte: O Autor.

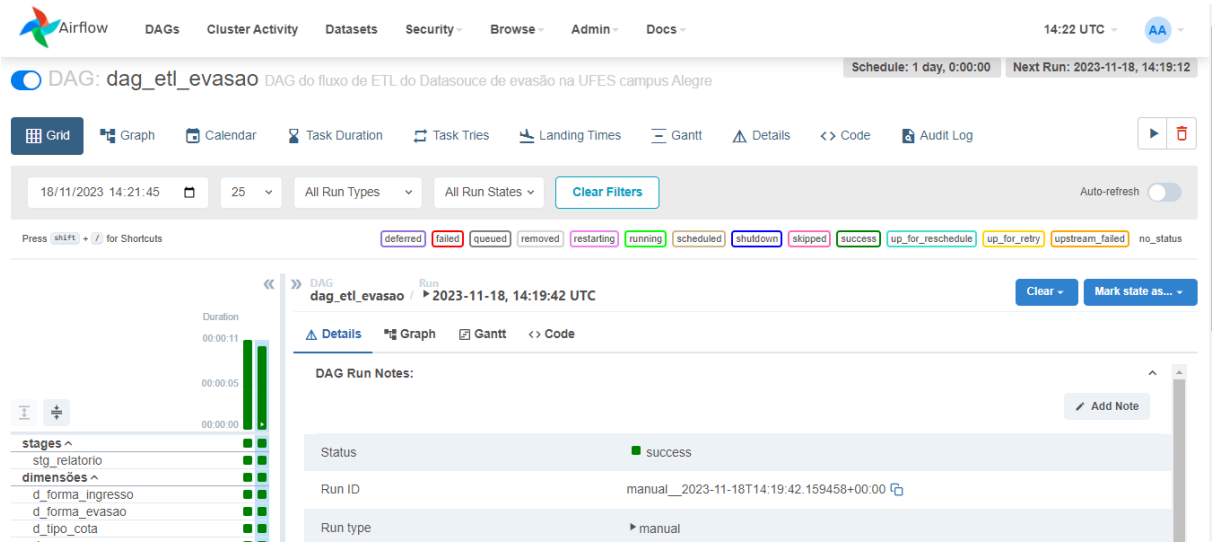

Figura E.3: Painel de detalhes da DAG.
Fonte: O Autor.

E.3.1 Atualização do Painel

Após a execução do fluxo de ETL, o usuário pode fazer a atualização dos dados no painel em um ambiente Windows 10/11. Utilizando o arquivo fonte do *dashboard* de-

senvolvido em reports/TCC2.pbix o usuário terá que configurar a conexão com o banco de dados do *Data Warehouse*, clicando no botão ”Transformar Dados” indicado pela seta vermelha na Figura E.4. Em seguida a ferramenta apresentará o formulário presente na Figura E.5.

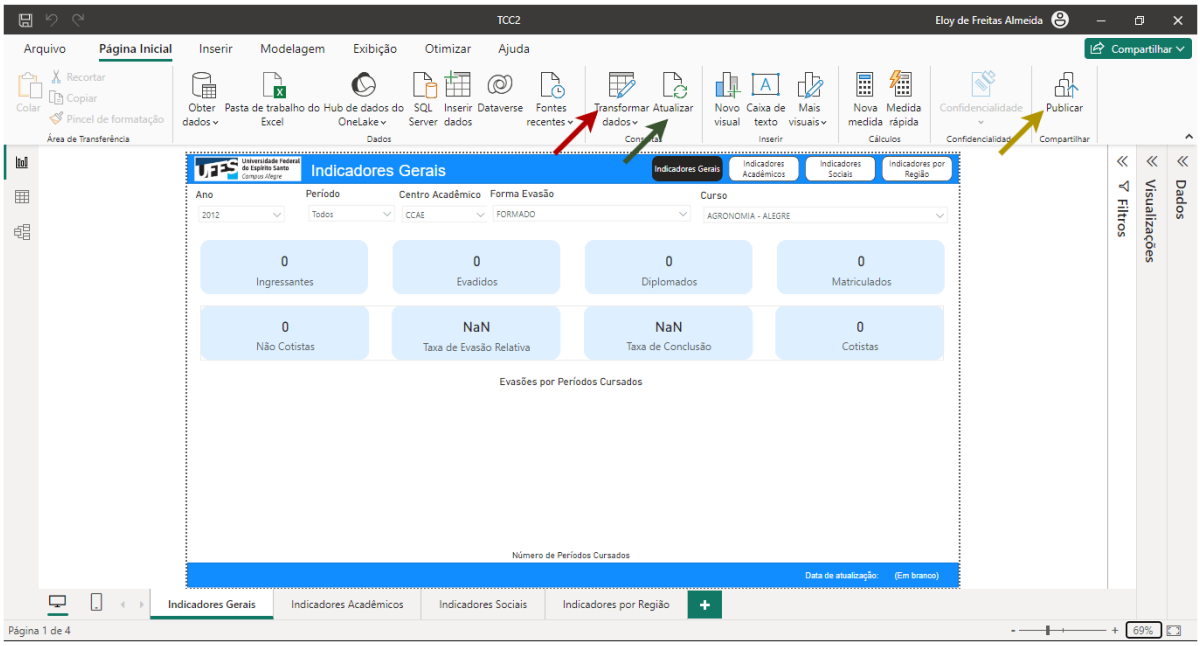

Figura E.4: Painel de indicadores gerais vazio.
Fonte: O Autor.

Na Figura E.5 ao clicar no botão ”Configuração de Fonte de Dados”, indicado pela seta da cor rosa, aparecerá uma janela com o botão ”Alterar Fonte...”, indicado pela seta em azul. Logo, aparecerá um formulário para alteração da conexão com a base de dados,

No formulário o usuário deve fornecer as seguinte informações:

1. IP do Servidor;
2. Porta do banco de dados: 10001;
3. *Schema*: dbdw;

As credenciais do usuários serão pedidas posteriormente, mas, vale ressaltar que o login e senha do usuário padrão é a palavra "postgres".

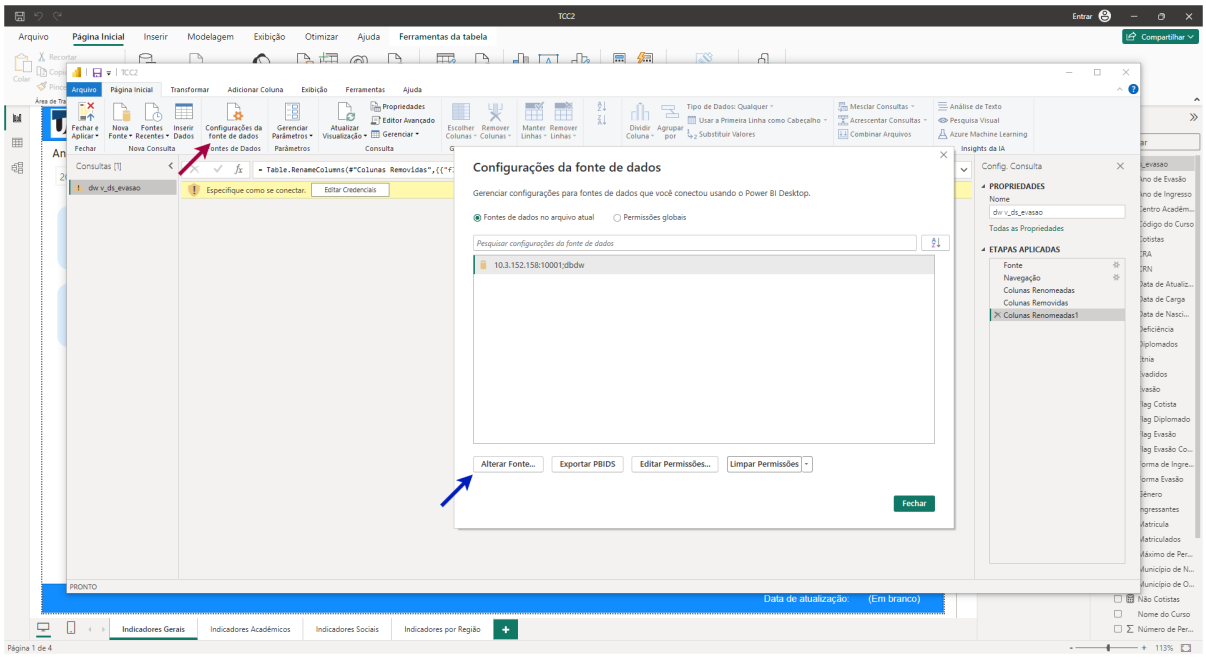

Figura E.5: Formulário de atualização de conexões.
Fonte: O Autor.

Após atualizar a conexão com o DW o usuário pode clicar no botão ”Atualizar”, indicado com a seta amarela na Figura E.4, para atualizar a fonte de dados.

Após a ferramenta processar a extração os painéis serão atualizados. Se a instituição possuir uma licença do *Power BI Online*, é possível publicar o *dashboard*, clicando no botão ”Publicar” no canto superior direito da janela, para ficar acessível para outros usuários.

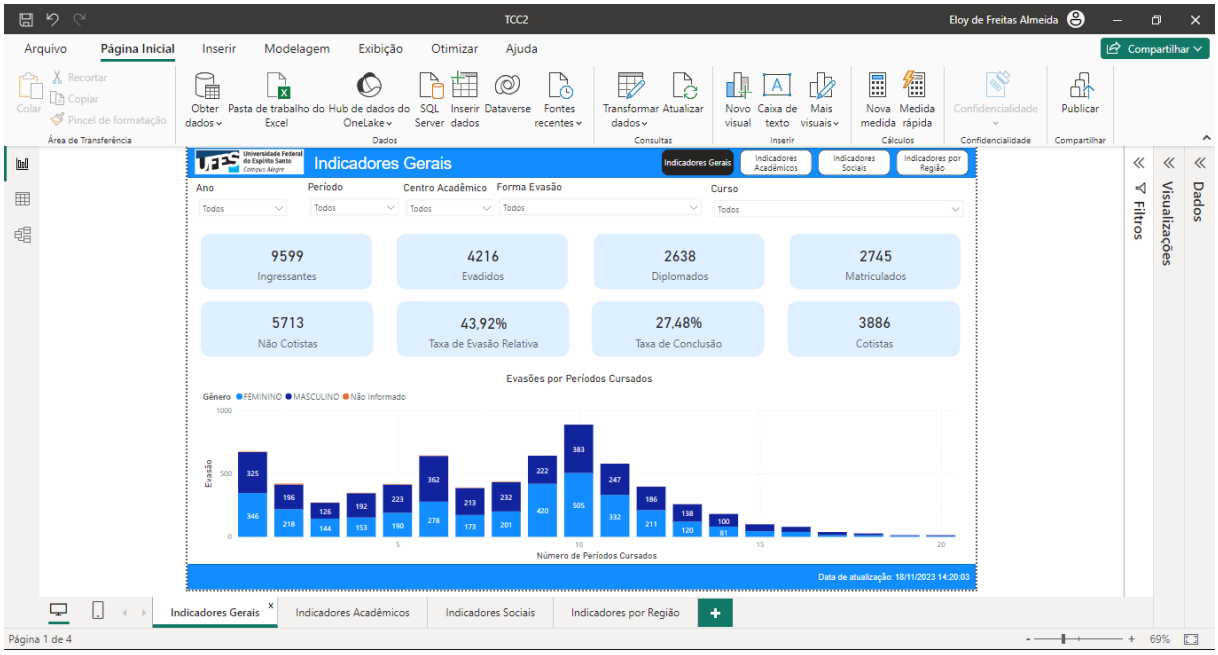

Figura E.6: Painel de indicadores gerais carregado.
Fonte: O Autor.

E.4 Interrupção de serviços

Se for necessário parar os serviços do *Airflow* de forma segura, execute o seguinte comando:

```
./stop.sh
```

E.5 Desinstalando programa

O *script* a seguir finaliza os serviços do *Airflow* e apaga todos os metadados. Portanto, faça um *backup* do banco de dados do DW e dos arquivos existentes no diretório airflow/data, pois todos serão perdidos

```
sudo ./nuke.sh
```