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 ____ de dezembro de 2023.

COMISSÃO EXAMINADORA

_____

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

_____

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

_____

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 9

1.1 Objetivos 10

1.1.1 Objetivo Geral 10

1.1.2 Objetivos específicos 10

2 Revisão de Literatura 11

2.1 *Data Warehouse* 11

2.1.1 Propostas arquitetura de um *Data Warehouse* 12

2.1.2 Modelagem Dimensional 14

2.2 Trabalhos relacionados 15

2.2.1 Trabalho de SALES JUNIOR (2013) 16

2.2.2 Trabalho de Mendes (2020) 17

2.2.3 Trabalho de Gonçalves (2021) 18

3 Metodologia 20

3.1 FERRAMENTAS UTILIZADAS 21

3.1.1 Linguagem de programação *Python* 21

3.1.2 *Apache Airflow* 22

3.1.3 *PostgreSQL* 22

3.1.4 Base de Dados de Origem 23

3.1.5 *MIcrosoft Power BI* 23

3.1.6 *Docker* 23

4 Resultados Obtidos 25
4.1 Levantamento dos Indicadores 25
4.1.1 Análise dos dados 26
4.2 Modelo Dimensional 27
4.3 Estrutura do Projeto e Fluxo de ETL 29
4.3.1 Diretório *airflow* 30
4.3.2 Diretório docs 31
4.3.3 Diretório *reports* 31
4.3.4 Diretório src 31
4.3.5 Diretório scripts 31
4.3.6 Diretório dags 32
4.4 *Dashboards* 33
4.4.1 Indicadores Gerais 34
4.4.2 Indicadores Acadêmicos 36
4.4.3 Indicadores Sociais 37
4.4.4 Indicadores Regionais 38

5 Conclusões 41

6 Trabalhos Futuros 42

Referências 43

Apêndice A – Script da *view* tratamento das *stages* 45

Apêndice B – Script da *view* do *datasource* 49

Apêndice C – Colunas da *stage* ”stg_relatorio” 51

Apêndice D – Dicionário de dados 54

Apêndice E – Tutorial de utilização 58

E.1 Requisitos de sistemas 58

E.2 Configuração de ambiente 58

E.2.1 Instalação do Sistema 58

E.3 Execução 59

E.3.1 Atualização do Painel 61

E.4 Interrupção de serviços 64

E.5 Desinstalando programa 64

LISTA DE FIGURAS

Figura 1 Arquitetura do DW de Kimball (2013) - Adaptado pelo autor. 12

Figura 2 Arquitetura do DW de Inmon (2005) - Adaptado pelo autor. 13

Figura 3 Exemplo do *Star Schema* retirado de Kimball (2013) - Traduzido pelo autor. 14

Figura 4 Modelo dimensional retirado de Kimball (2013) - Adaptado pelo autor. 15

Figura 5 Modelo dimensional proposto por Mendes (2020) - Adaptado pelo autor. 18

Figura 6 Modelo dimensional proposto por Gonçalves (2021) - Adaptado pelo autor. 19

Figura 7 Arquitetura do DW - Elaborado pelo autor. 20

Figura 8 Exemplo de uma DAG - Adaptado pelo autor. 22

Figura 9 Comparação de um sistema utilizando *Docker* e um sistema utilizando máquinas virtuais - Criado pelo autor. 24

Figura 10 *Schema* da STG - Criado pelo autor. 27

Figura 11 Modelo relacional - Criado pelo autor. 27

Figura 12 *Schema* do DW - Criado pelo autor. 29

Figura 13 Estrutura do projeto - Criado pelo autor. 30

Figura 14 DAG Evasão - Criado pelo autor. 33

Figura 15 Indicadores Gerais - Criado pelo autor. 35

Figura 16 Aplicação de Filtro nos Indicadores Gerais - Criado pelo autor. 36

Figura 17 Indicadores Acadêmicos - Criado pelo autor. 37

Figura 18 Indicadores Sociais (exemplo de interação com gráfico Evasão por Etnia) - Criado pelo autor. 38

Figura 19 Indicadores Regionais - Criado pelo autor. 39

Figura 20 Indicadores Regionais (com filtro de turma) - Criado pelo autor. 40

Figura 21 Página inicial do *Apache Airflow* - Criado pelo autor. 60

Figura 22 Página inicial do *Apache Airflow* execução da DAG - Criado pelo autor. 61

Figura 23 Painel de detalhes da DAG - Criado pelo autor. 61

Figura 24 Painel de indicadores gerais vazio - Criado pelo autor. 62

Figura 25 Formulário de atualização de conexões - Criado pelo autor. 63

Figura 26 Painel de indicadores gerais carregado - Criado pelo autor. 64

RESUMO

A evasão estudantil é um fenômeno que pode ocorrer por diversos tipos de fatores, que caso não sejam estudados e mitigados podem ocasionar em medidas drásticas como o fechamento de um curso, que por muitas vezes é ofertado apenas na rede pública. Porém, muitas dificuldades relacionadas à evasão, que uma Instituição de Ensino Superior (IES) enfrenta são específicas da região onde está situada e dessa forma precisa de soluções especializadas para seu cenário. Para entender esse fenômeno e criar medidas efetivas é preciso analisar as informações relacionadas à situação acadêmica dos alunos. Mas, nem toda IES possui uma ferramenta especializada para análise de indicadores relacionados à evasão dos seus discentes. Uma das maneiras de visualizar os dados de forma mais intuitiva é através de modelos de dados especializados para estudo de indicadores de um determinado assunto. Existem diversas metodologias para criação de soluções para análise de dados, como o modelo de *Data Warehouse* (DW) de *Ralph Kimball*. No âmbito deste estudo, foi elaborado um ambiente de *Business Intelligence* (BI) destinado à análise da evasão estudantil na Universidade Federal do Espírito Santo (UFES) *campus* Alegre. Esse ambiente foi concebido com base em um modelo de dados inspirado na metodologia proposta por *Ralph Kimball* e foi alimentado por meio de um processo automatizado de Extração, Transformação e Carga (ETL), utilizando informações provenientes dos relatórios de situação de matrícula dos alunos vinculados ao Centro de Ciências Exatas, Naturais e da Saúde (CCENS) e ao Centro de Ciências Agrárias e Engenharias (CCAE), abrangendo o intervalo temporal entre os períodos de 2009/01 e 2023/01. Adicionalmente, foram desenvolvidos *dashboards* interativos com o intuito de aprimorar a análise do problema e otimizar o processo decisório, visando a mitigação desse fenômeno.

Palavras-chave: Evasão; Ensino Superior; *Data Warehouse*; *Business Intelligence*, ETL, UFES;

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 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 *Ralph Kimball* 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.

2.1 *Data Warehouse*

De acordo com Kimball (2013) os sistemas operacionais 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;

E por fim, é orientado a assunto, ou seja, a sua modelagem busca atender alguma área de assunto da organização, por exemplo: venda de produtos de uma comércio varejistas. Kimball (2013) ainda afirma que um DW deve ser intuitivo para o usuário apresentando os dados da forma mais simples possível. Além de disponibilizar a informação de forma consistente e rápida. E também estar preparado para mudanças, como criação de novas entidades no modelo de dados.

2.1.1 Propostas arquitetura de um *Data Warehouse*

Uma arquitetura de *Data Warehouse* é composta por quatro ambientes: sistemas operacionais de origem, sistema de *Extract, Transform* e *Load* (ETL), área de apresentação dos dados e aplicação de *Business Intelligence* (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.

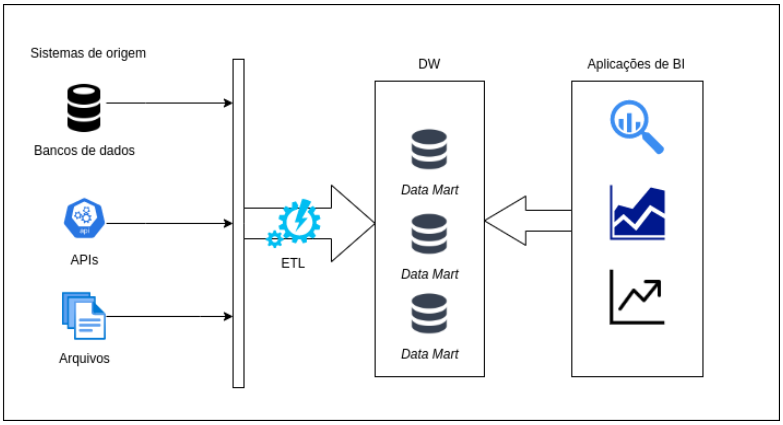

Figura 1: Arquitetura do DW de Kimball (2013) - Adaptado pelo autor.

Na Figura 1, os sistemas operacionais de origem são sistemas utilizados pela organização para manter o processo de negócio. Geralmente eles são fisicamente isolados do ambiente de DW servindo para extração de dados que são de interesse na análise.

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 na área de apresentação, que também é conhecida como 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.

Inmon (2005) propõe uma arquitetura parecida com a de Kimball (2013), a principal mudança acontece na camada de estrutura do DW e apresentação dos dados, como ilustrado na Figura 2.

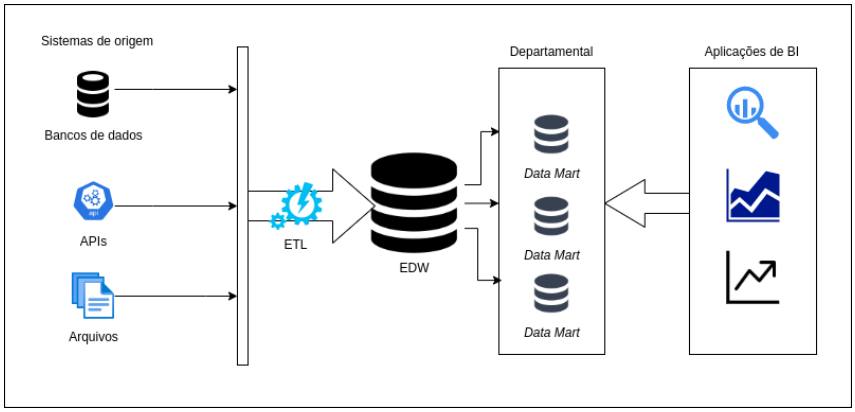

Figura 2: Arquitetura do DW de Inmon (2005) - Adaptado pelo autor.

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.

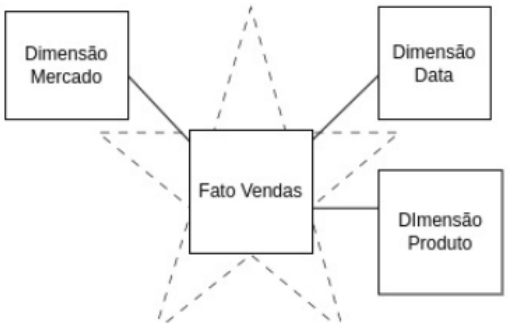

Figura 3: Exemplo do *Star Schema* retirado de Kimball (2013) - Traduzido pelo autor.

Tal modelo é composto por três tipo de componentes: fatos, dimensões e suas conexões. Um fato representa um evento do mundo real definido por um conjunto de dimensões e métricas sobre um determinado assunto. O fato é estruturado em uma tabela histórica, onde cada registro representa uma relação das dimensões em um determinado evento.

Uma tabela fato representa um relacionamento “muitos-para-muitos” entre as dimensões do modelo. Kimball (2013) e Inmon (2005) ainda afirmam, que a estrutura de um fato deve possuir as chaves estrangeiras que fazem referência às chaves primárias das dimensões do modelo e também pode conter alguns cálculos relacionados ao evento que podem auxiliar na análise. Além disso, elas tendem possuir um volume de dados muito grande dependendo da granularidade e do histórico da informação.

As dimensões, por sua vez, representam as características textuais relacionadas aos eventos contidos na tabela fato. Geralmente as tabelas dimensão são pequenas e

possuem apenas informações úteis relacionadas ao evento, como a chave primária utilizada no relacionamento com a fato e um conjunto de colunas descritivas. Além disso, são muito importantes para a estrutura do DW, pois elas tornam o modelo compreensível para o uso.

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.

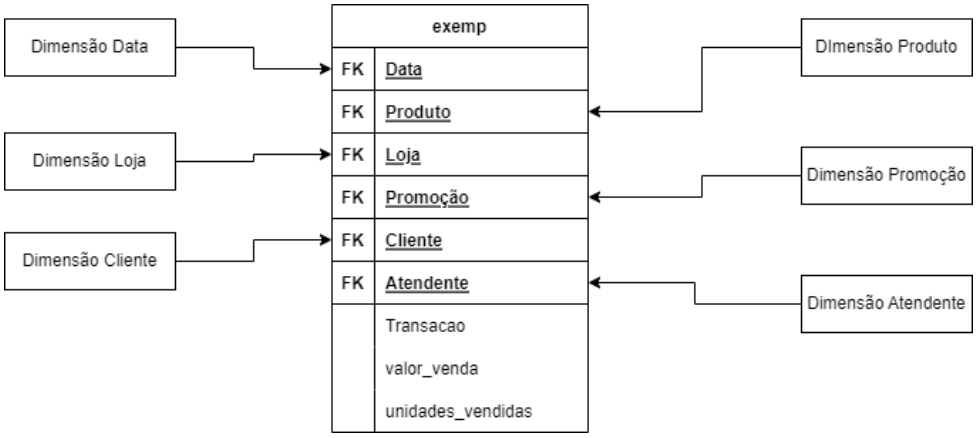

Figura 4: Modelo dimensional retirado de Kimball (2013) - Adaptado pelo autor.

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 obra 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)

A pesquisa 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.

Na pesquisa, o autor utilizou as premissas definidas pela Pró-Reitoria de Graduação da UFES para identificar o que caracteriza uma situação de evasão. Considera-se que um discente evadiu do curso quando ele está em uma das seguintes condições:

- Desistência do curso por parte do estudante;
- Desligamento do aluno por parte da Universidade de acordo com as portarias vigentes;
- Falecimento;
- Jubilamento por extrapolação do prazo máximo de término do curso;
- Matrícula desativada por falhas no cadastro;
- Reopção de curso;
- Sanção disciplinar;
- Transferência interna;
- Transferência para outra IES;

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.

No capítulo dos resultados o autor apresentou o produto de sua pesquisa em três etapas. Na primeira etapa foram apresentados resultados em uma granularidade maior, seguintes visões:

- Tabelas de frequência por forma de saída;
- Distribuição do tempo (período do curso) até a evasão;
- Distribuição de tempo (período do curso) até a conclusão do curso;

No segundo momento, a pesquisa apresentou o perfil da amostra agrupadas por forma de saída do curso (evasão ou conclusão) mostrando resultados da inferência estatística em: variáveis métricas dos estudantes (idade ao ingressar no curso, notas do Enem, coeficiente de rendimento e índices de reprovações), variáveis de contexto familiar (nível de instrução, ocupação e situação de trabalho dos pais e questões relacionadas renda familiar), atributos individuais (cotista, estado civil, gênero, faixa etária, cor/etnia, isenção no vestibular) e variáveis relacionadas ao desempenho em escolaridade antes do ingresso e durante a graduação até a saída.

Por fim, na última etapa o autor apresentou resultados relacionados a aplicação do método de regressão logística no qual descreveu o efeito simultâneo de variáveis apresentadas nos resultados anteriores sobre o evento de “Saída do curso” e também identificou possíveis padrões e tendências sobre o evento.

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. 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 et al. (2008), 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 5 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". A partir do modelo dimensional construído, Mendes (2020) trouxe resultados no formato de séries temporais, apresentando a evolução das taxas anuais da situação de matrícula dos alunos.

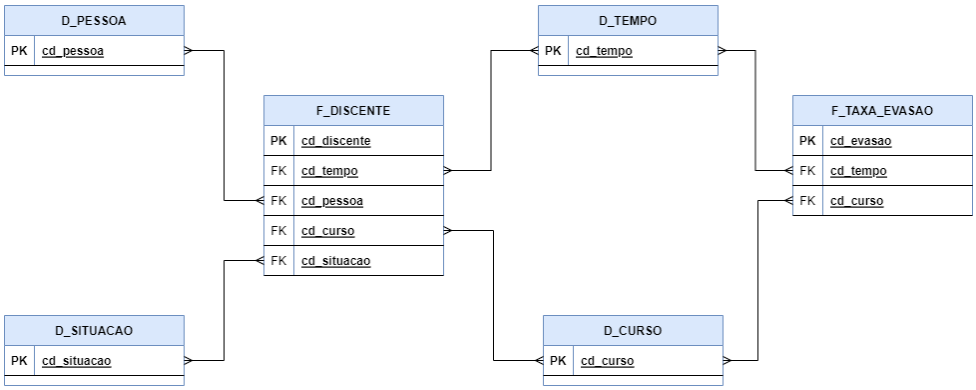

Figura 5: Modelo dimensional proposto por Mendes (2020) - Adaptado pelo autor.

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 *Business Intelligence*, 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 6, 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.

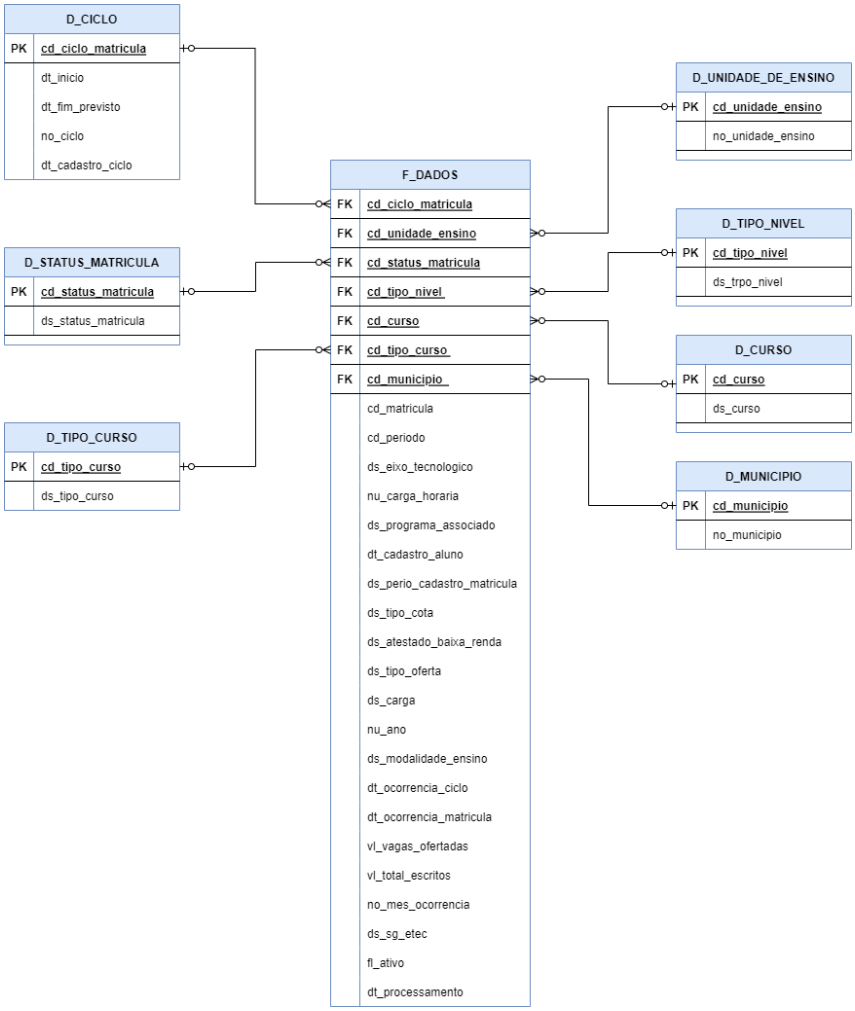

Figura 6: Modelo dimensional proposto por Gonçalves (2021) - Adaptado pelo autor.

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 7 apresenta a arquitetura proposta nesta pesquisa, que será melhor detalhada a seguir.

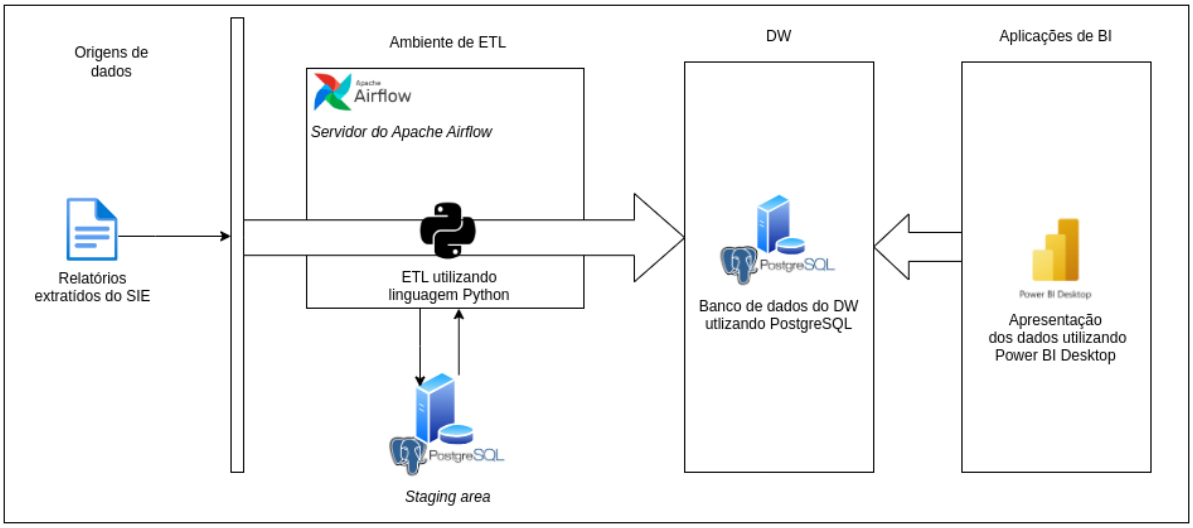

Figura 7: Arquitetura do DW - Elaborado pelo autor.

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 7) foram utilizados relatórios da situação de matrícula dos alunos extraídos do Sistema de Informação para o Ensino (SIE) utilizado pela UFES.

A partir dos dados obtidos foi elaborado um modelo dimensional com todas as dimensões identificadas como importantes para analisar o fato em questão e por conse-

guinte 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 Sistema de Gerenciamento de Banco de Dados Objeto-Relacional (ORDBMS) *PostgreSQL*.

O ambiente de ETL (Figura 7), foi implementado na linguagem de programação *Python*. 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 ORDBMS *PostgreSQL*.

A automatização da execução desses dois fluxos citados, é de responsabilidade do *Apache Airflow*, 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 Desktop* 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.

Nas seções seguintes serão apresentadas as ferramentas que foram utilizadas na implantação da arquitetura, a base de dados de origem e o processo de modelagem, até o modelo final.

3.1 FERRAMENTAS UTILIZADAS

3.1.1 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

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

para análise e manipulação de estruturas “relacionais” e “rotuladas” de forma fácil e intuitiva (PANDAS, 2022).

3.1.2 *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*, 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* e são programadas para serem executadas de forma agendada ou manual. Dessa forma, para este trabalho, cada tarefa será um *script Python* que executa algum passo do processo. A Figura 9 exemplifica a estrutura de uma DAG.

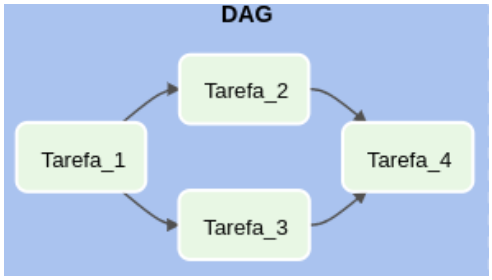

Figura 8: Exemplo de uma DAG - Adaptado pelo autor.

3.1.3 *PostgreSQL*

O *PostgreSQL* [3] 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 recursos modernos de um SGBD, como: consultas complexas, chaves estrangeiras, gatilhos, *views* e integridade transacional (POSTGRESQL, 2022). Nesse projeto, o *PostgreSQL* será utilizado para manter a estrutura física das tabelas da *Staging Area* e do *Data Warehouse*.

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

3.1.4 Base de Dados de Origem

O Sistema de Informação para o Ensino (SIE) [4] é o sistema utilizado pela UFES para a gestão de suas atividades acadêmicas, recursos humanos, registro e controle acadêmico (disciplinas, cursos, docentes, currículos), gestão de matrículas, gestão contábil e patrimonial (UFES, 2022). Logo, o estudo foi realizado com relatórios sobre a situação cadastral dos alunos emitidos pelo SIE.

3.1.5 *MIcrosoft Power BI*

A escolha do *MIcrosoft Power BI* [5]. 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 na nuvem, como bancos de dados, planilhas e serviços web, ampliando a versatilidade e a abrangência das análises realizadas.

3.1.6 *Docker*

Para facilitar a implantação e manutenção do sistema, foi utilizado a ferramenta *Docker* [6], 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* fornece uma interface de linha de comando (CLI) que facilita a manipulação de sistemas complexos.

[4] https://npd.ufes.br/sistema-de-informa%C3%A7%C3%A3o-para-o-ensino-sie
[5] https://powerbi.microsoft.com/pt-br/
[6] https://www.docker.com/

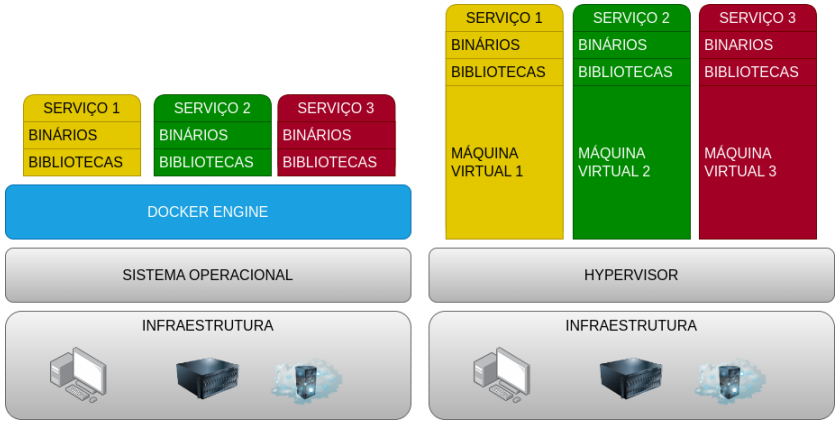

Figura 9: Comparação de um sistema utilizando *Docker* e um sistema utilizando máquinas virtuais - Criado pelo autor.

4 RESULTADOS OBTIDOS

4.1 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
- 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?

4.1.1 Análise dos dados

Para verificar se os indicadores pudessem ser contemplados com sucesso foi feita uma análise dados extraídos de relatórios do módulo ”1.1.4.13 Lança Abandono para Alunos sem Matrícula no Período” do SIE. 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;
- 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 10:

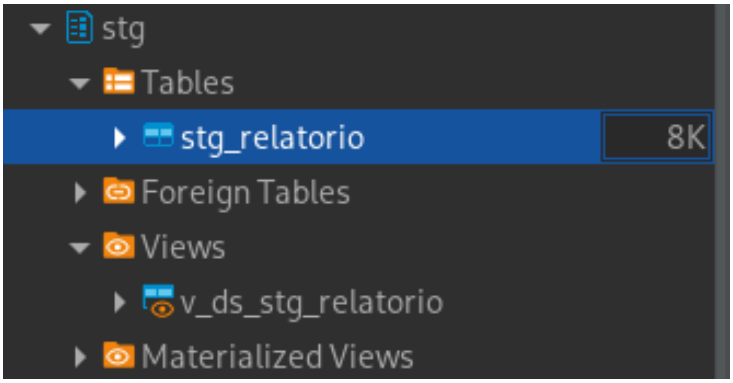

Figura 10: *Schema* da STG - Criado pelo autor.

4.2 Modelo Dimensional

A partir dos requisitos levantados e dos dados obtidos foi possível propor um modelo dimensional conforme apresentado na Figura 11.

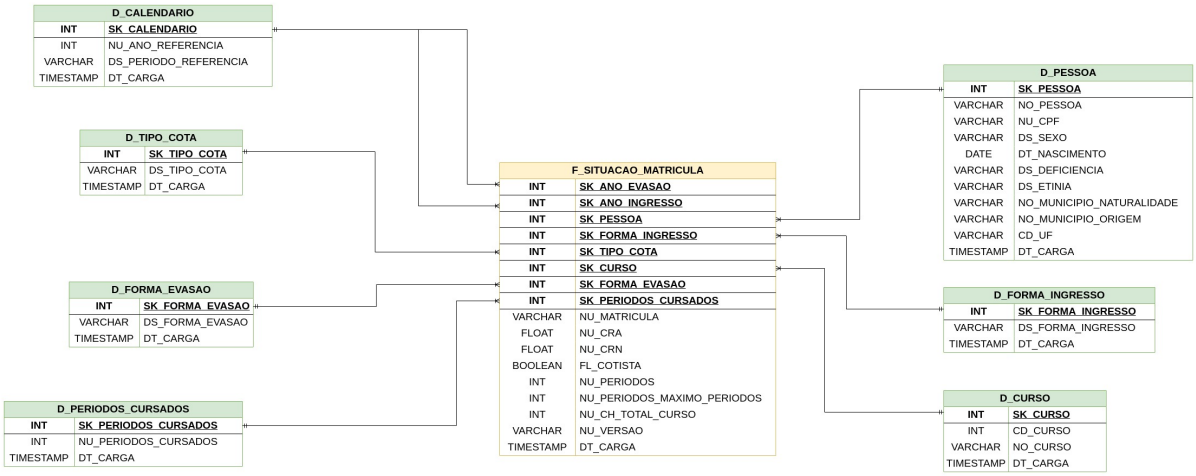

Figura 11: Modelo relacional - Criado pelo autor.

No modelo dimensional proposto na Figura 11 foi identificada a tabela fato F_SITUACAO_MATRICULA que 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 as questões impostas pelos *stakeholders* foram identificadas as seguintes dimensões: D_TIPO_COTA para classificação dos tipos de cota, D_CALENDARIO para referência temporal da informação em ano e semestre, 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 11, 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 12

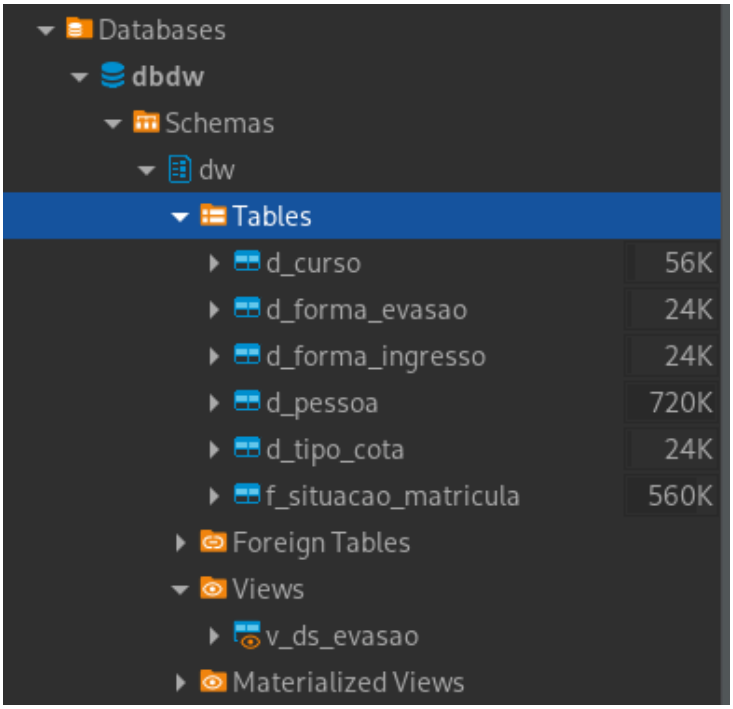

Figura 12: *Schema* do DW - Criado pelo autor.

4.3 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 13, disponível no repositório [1]. E cada um deles vai ser detalhado nas próximas subseções.

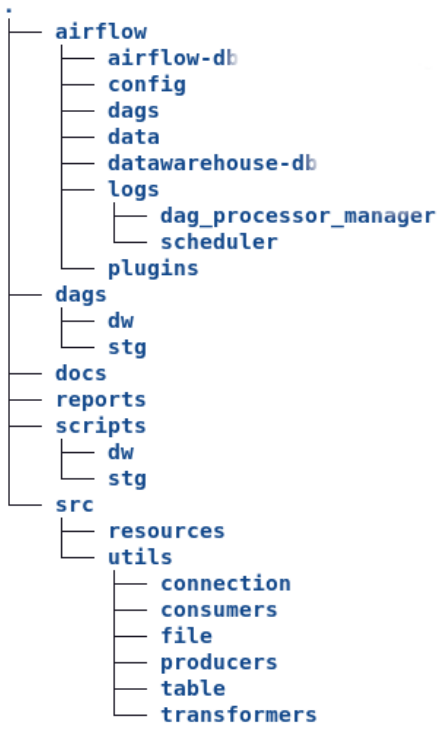

Figura 13: Estrutura do projeto - Criado pelo autor.

4.3.1 Diretório *airflow*

Esse diretório é gerado após a finalização do tutorial (Apêndice E). Esse diretório tem o objetivo de concentrar as estruturas que o *Airflow* 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 10. 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;

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

4.3.2 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.

4.3.3 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*.

4.3.4 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;
- *table* - Manipulação de tabelas;
- *file* - Manipulação de arquivos;
- *producers* - Extração de dados;
- *consumers* - Carga de dados;
- *resources* - Para recursos e configurações de sistema.

4.3.5 Diretório scripts

O diretório ”*scripts*/” contém os *scripts python* 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 13 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*.

4.3.6 Diretório dags

A ferramenta utilizada, *Apache Airflow*, organiza as tarefas do fluxo de dados em DAGs, como representado na Figura 9. 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 *Airflow* utiliza para gerenciar os *scripts*.

No presente projeto, os *scripts* são configurados por meio de arquivos que utilizam os *templates* apresentados na subseção 4.3.5. 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 14, que organiza todas as tarefas em um fluxo único.

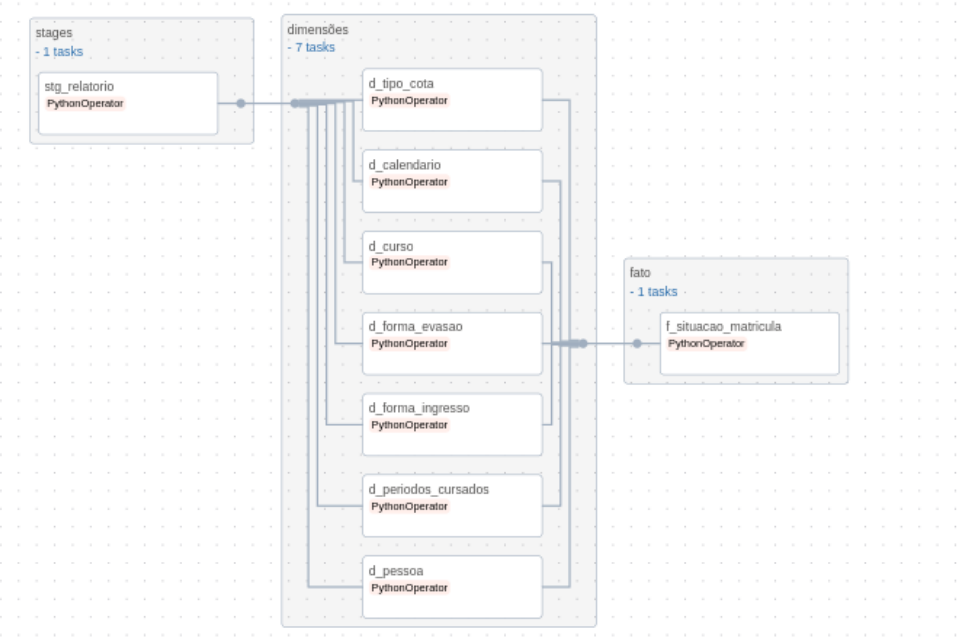

Figura 14: DAG Evasão - Criado pelo autor.

4.4 *Dashboards*

Após a execução do fluxo de ETL e a disponibilização dos dados no *Data Mart*, utilizou-se a ferramenta de *business intelligence Microsoft Power BI* para a construção de *dashboards* interativos para utilização do usuário.

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*.

4.4.1 Indicadores Gerais

O painel de Indicadores Gerais (Figura 15) 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 (Evadidos/Ingressantes) e Taxa de Conclusão (Conclusão/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 pela maioria dos discentes que evadiram. Por meio deste, pode-se observar qual o período de maior predominância de evasão, o que pode estar associado a disciplinas e ofertas.

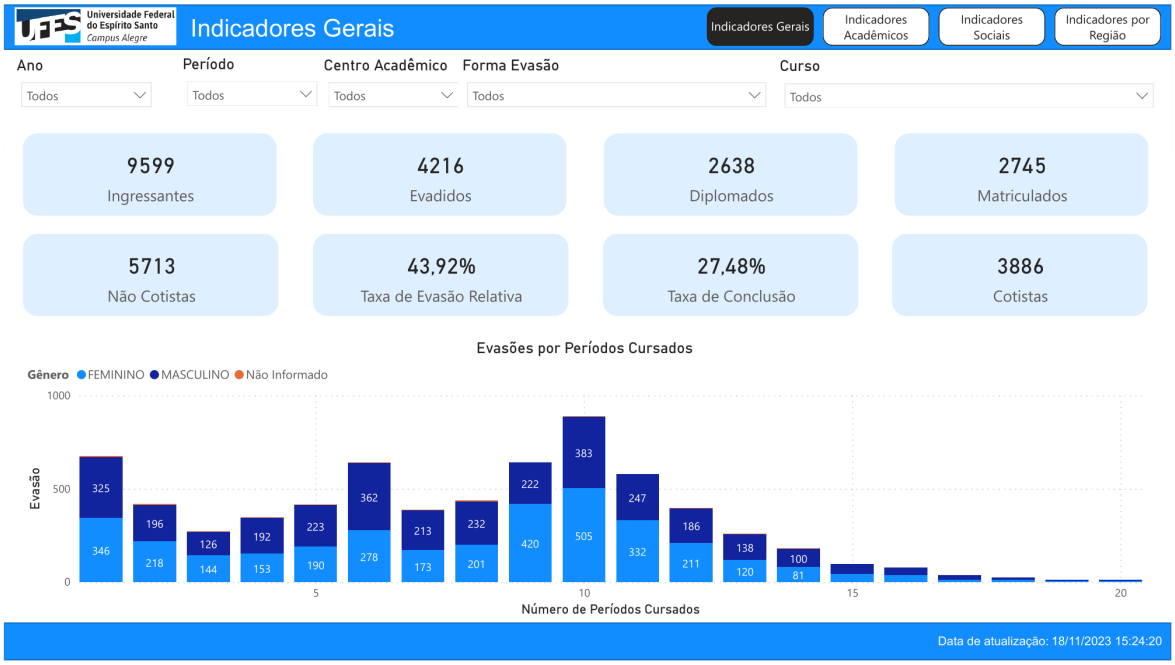

Figura 15: Indicadores Gerais - Criado pelo autor.

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 (Figura 15), ou analisar apenas uma turma específica. A Figura 16, por exemplo, apresenta dados da turma de Ciência da Computação que ingressou no ano de 2010. Observa-se que em total de 29 ingressantes (15 cotistas e 14 não cotistas), houve 22 evasões e 7 diplomados, resultando em uma taxa relativa de 75,86% de evasão e 24,14% de conclusão. Além disso, o pico de evasão foi no oitavo semestre com um total de 4 evasões sendo 3 pessoas do sexo masculino e 1 do feminino.

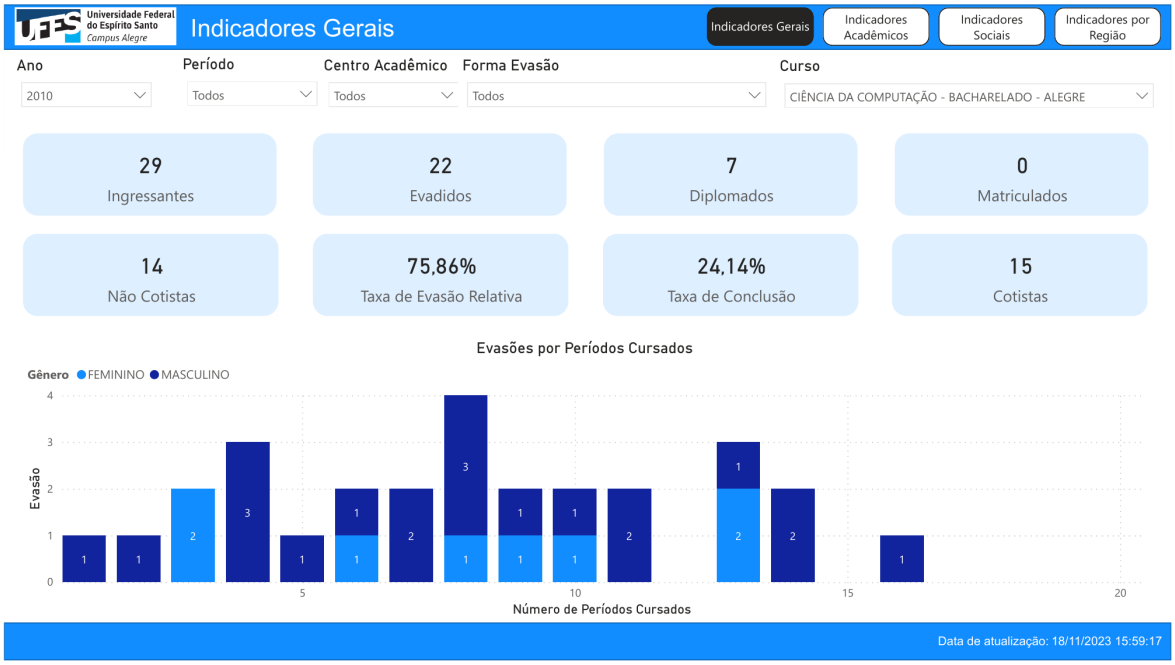

Figura 16: Aplicação de Filtro nos Indicadores Gerais - Criado pelo autor.

4.4.2 Indicadores Acadêmicos

Esse painel tem o objetivo de apresentar a relação da evasão com os detalhes da matrícula. A Figura 17 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;

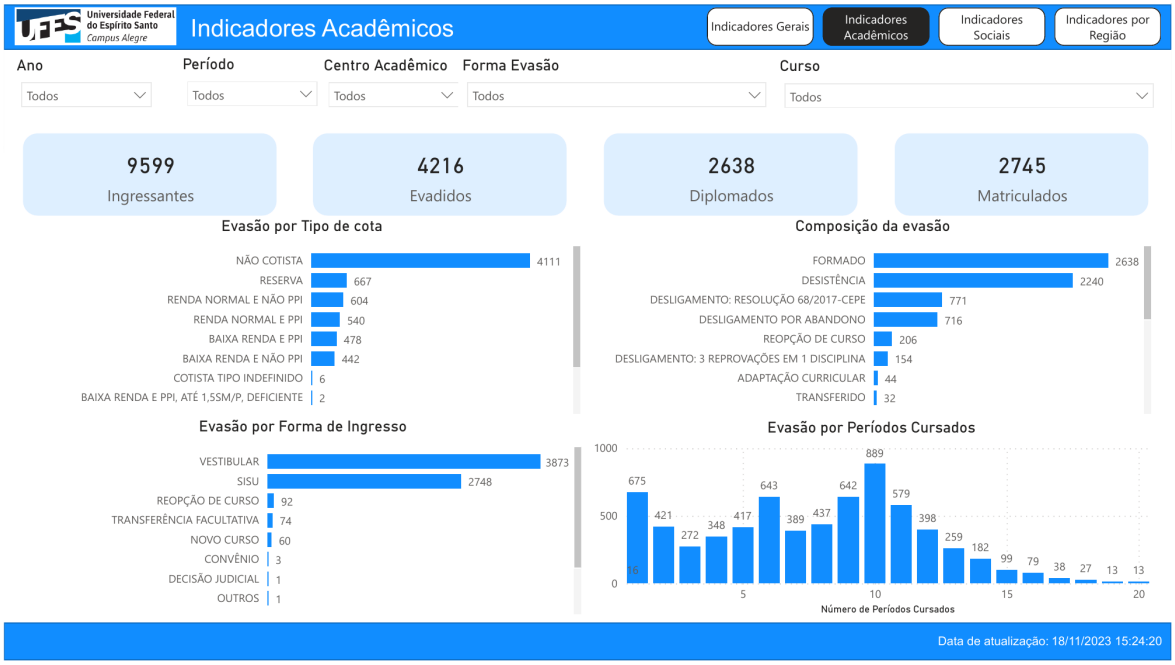

Figura 17: Indicadores Acadêmicos - Criado pelo autor.

O exemplo demonstrado na Figura 17 compreende a totalidade do período de dados analisados. Observou-se que a maior parte dos alunos que evadiram não se enquadra na categoria de cotistas. Os ingressantes por meio do vestibular apresentaram um índice de evasão superior aos estudantes que ingressaram pelo Sistema de Seleção Unificada (SISU). Destaca-se que o 10º semestre é o período que concentra a maior quantidade de evasões de forma geral, seguido pelo 1º semestre.

4.4.3 Indicadores Sociais

Essa página 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;

Além disso, o usuário pode interagir com os gráficos através do clique do *mouse*, desencadeando no filtro da informação selecionada nos outros gráficos da mesma

página. Na Figura 18, é exibido o resultado da interação ao selecionar a classe Parda, com um total de 12 evasões no gráfico referente à "Evasão por Etnia". A ferramenta demonstra como o valor selecionado se relaciona com as informações exibidas nos demais gráficos. Onde é possível observar que houve 8 evasões (6 desistências e 2 desligamentos por abandono) e 4 diplomações. E a maioria das evasões ocorreram entre o 5º e 10º semestre.

Figura 18: Indicadores Sociais (exemplo de interação com gráfico Evasão por Etnia) - Criado pelo autor.

4.4.4 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.

Os usuários têm a possibilidade de 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. Para este cenário, o ideal seria utilizar a localização da escola de origem, porém este dado não está disponível na base de dados.

Na Figura 19, está apresentado o painel intitulado "Indicadores Regionais", exibindo a análise da evasão de toda a universidade no período compreendido entre 2009/01 e 2023/01. Já a Figura 20 demonstra o resultado após a interação com os filtros, ao consultar a turma de Ciências Biológicas - Bacharelado - Alegre do ano de 2013. Nesse contexto, no gráfico referente à "Evasão por Município de Naturalidade", observa-se uma concentração no sul do estado do Espírito Santo, na região da Grande Vitória e no sul da Bahia. Quanto à "Evasão por Município de Origem", a concentração é percebida no sul do Espírito Santo e na região da Grande Vitória.

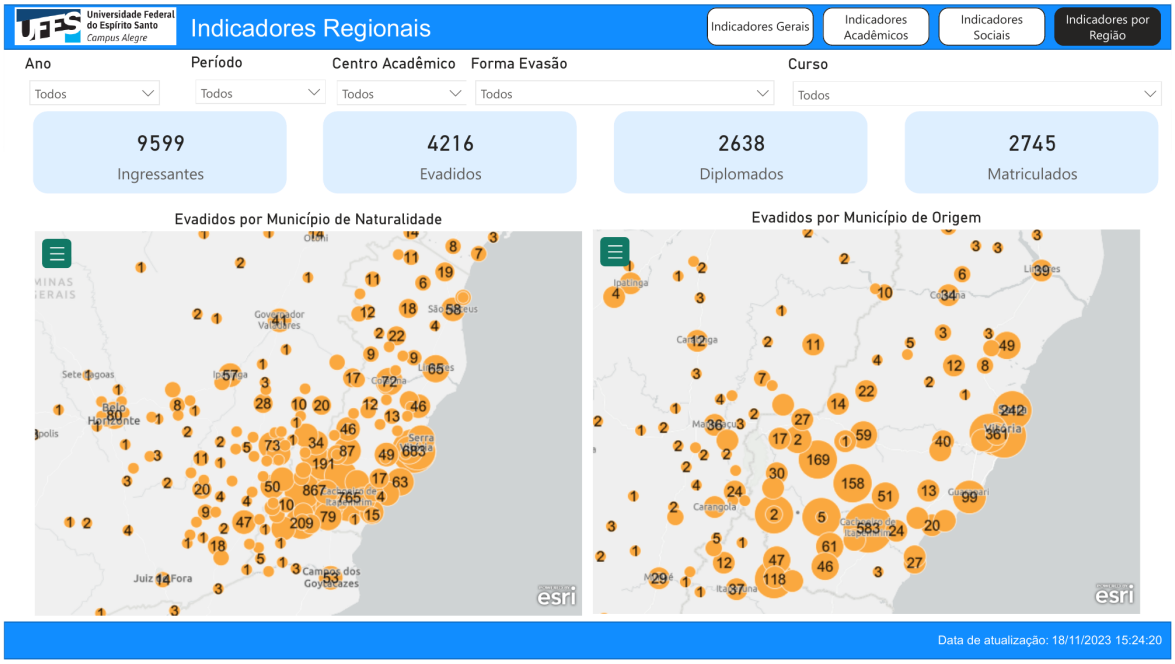

Figura 19: Indicadores Regionais - Criado pelo autor.

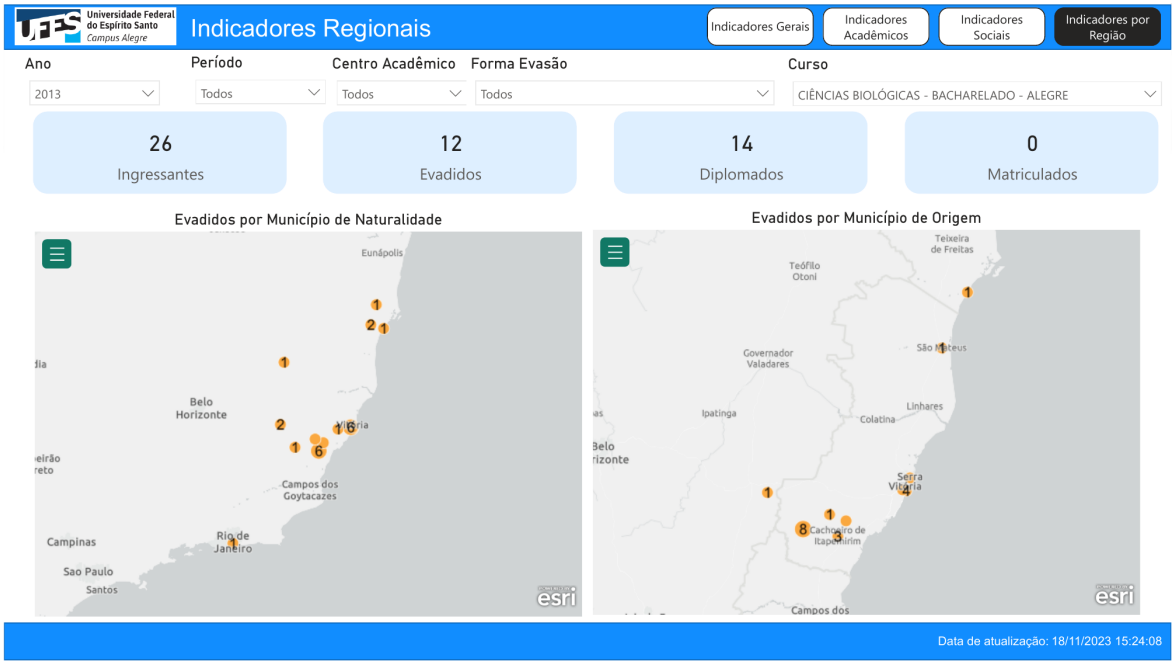

Figura 20: Indicadores Regionais (com filtro de turma) - Criado pelo autor.

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 *Business Intelligence* 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 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.

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 21.

ANDIFES, A. Sesu/mec. *Diplomação, retenção e evasão nos cursos de graduação em instituições de Ensino Superior públicas. Brasília, DF*, 1996. Citado na página 9.

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 22.

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 19.

DOCKER. *Docker, Version: 24.0.7*. 2023. Disponível em: <https://docs.docker.com/get-started/>. Citado na página 23.

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 29.

GONÇALVES, L. M. *Uma plataforma de business intelligence para analisar a retenção e evasão do IFMT*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2021. Citado 5 vezes nas páginas , 9, 15, 18 e 19.

INMON, W. H. *Building the Data Warehouse, Fourth Edition*. [S.l.]: Wiley Publishing Inc., 2005. Citado 4 vezes nas páginas , 11, 13 e 14.

KIMBALL, M. R. R. *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. [S.l.]: John Wiley & Sons, 2013. Citado 10 vezes nas páginas , 9, 10, 11, 12, 13, 14, 15, 20 e 28.

KIMBALL, R. et al. *The data warehouse lifecycle toolkit*. [S.l.]: John Wiley & Sons, 2008. Citado na página 17.

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 5 vezes nas páginas , 9, 15, 17 e 18.

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 9.

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 9.

PANDAS, d. t. *pandas documentation, Version: 1.4.3*. 2022. Disponível em: <https://pandas.pydata.org/docs/getting_started/overview.html>. Citado na página 22.

PEÇANHA, P. N. *PNP 2022 (Ano Base 2020)*. 2022. Disponível em: <https://www.plataformanilopecanha.org/>. Citado na página 19.

POSTGRESQL, G. D. G. *PostgreSQL Documentation, Version: 14*. 2022. Disponível em: <https://www.postgresql.org/docs/current/index.html>. Citado na página 22.

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 3 vezes nas páginas , 15 e 16.

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 21.

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 23.

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* para execução dos serviços do *Apache Airflow*.

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;
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 21.

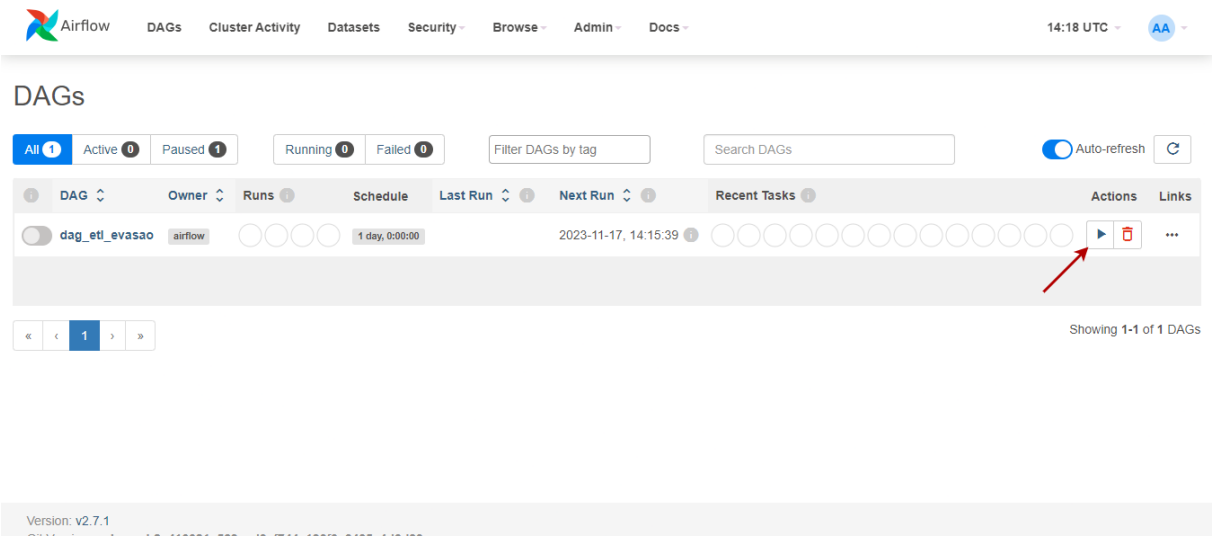

Figura 21: Página inicial do *Apache Airflow* - Criado pelo autor.

Quando a DAG inicia a execução, o seu estado muda conforme Figura 21. O usuário pode visualizar os detalhes da execução clicando no nome DAG, acessando o painel mostrado na Figura 23.

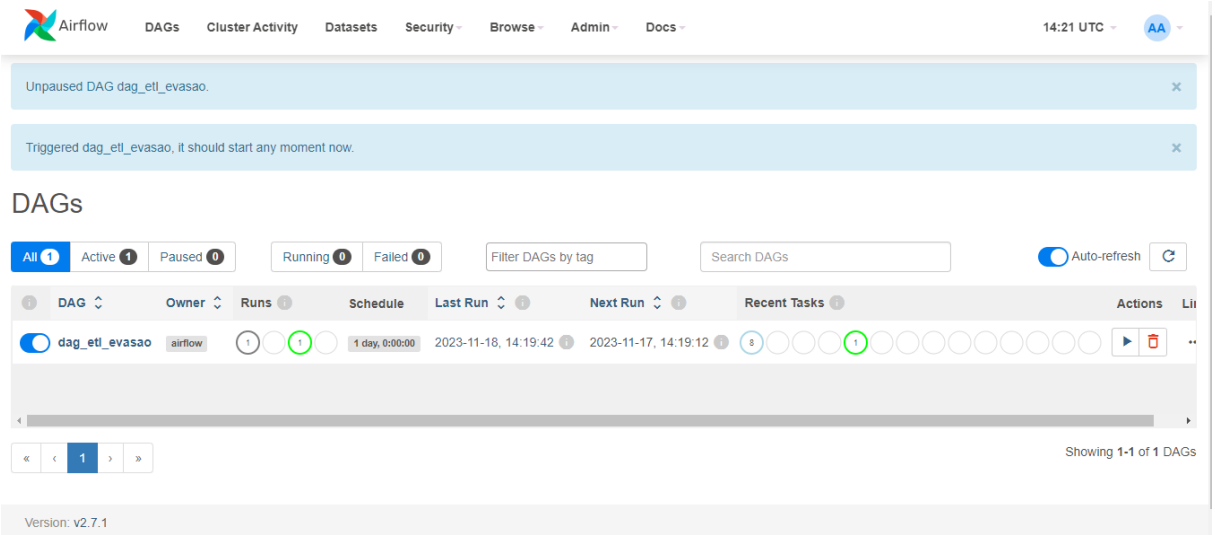

Figura 22: Página inicial do *Apache Airflow* execução da DAG - Criado pelo autor.

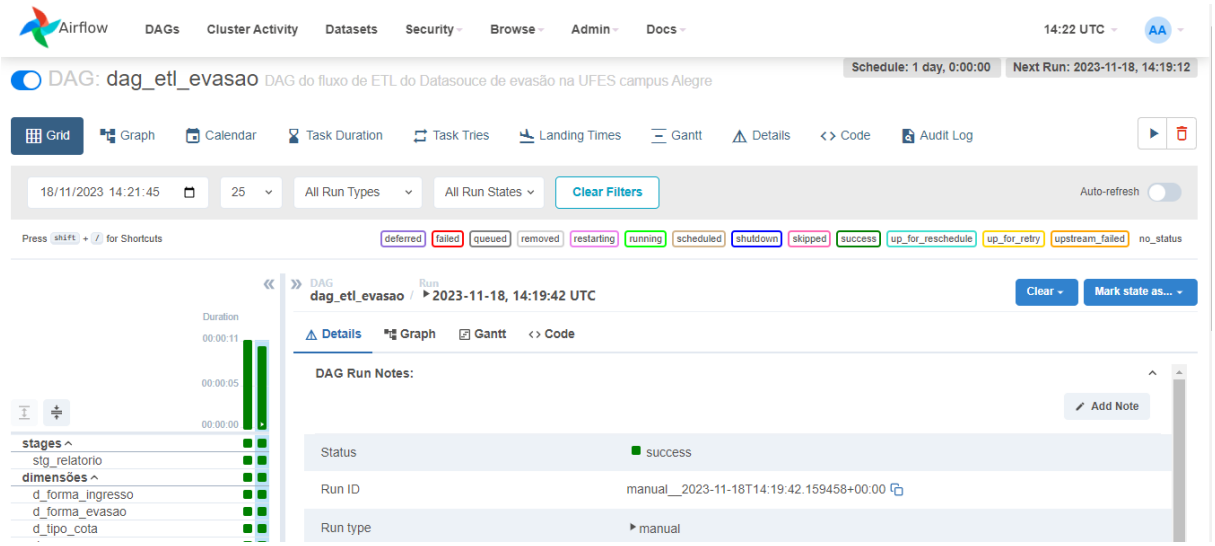

Figura 23: Painel de detalhes da DAG - Criado pelo autor.

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* desenvolvido 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 24. Em seguida a ferramenta apresentará o formulário presente na Figura 25.

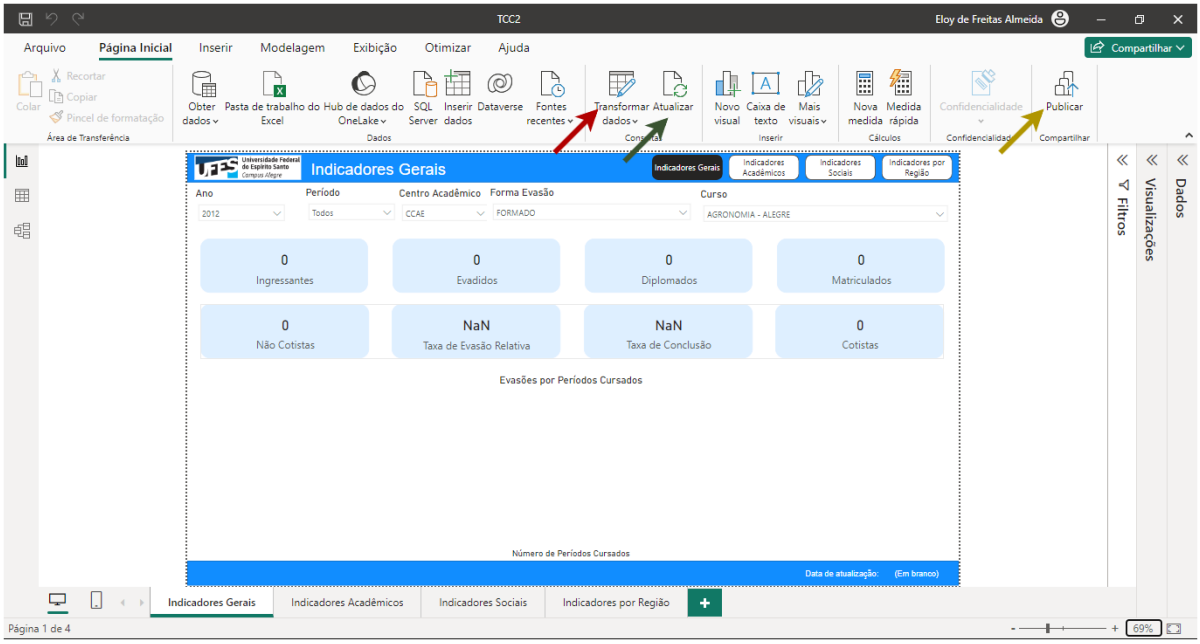

Figura 24: Painel de indicadores gerais vazio - Criado pelo autor.

Na Figura 25 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".

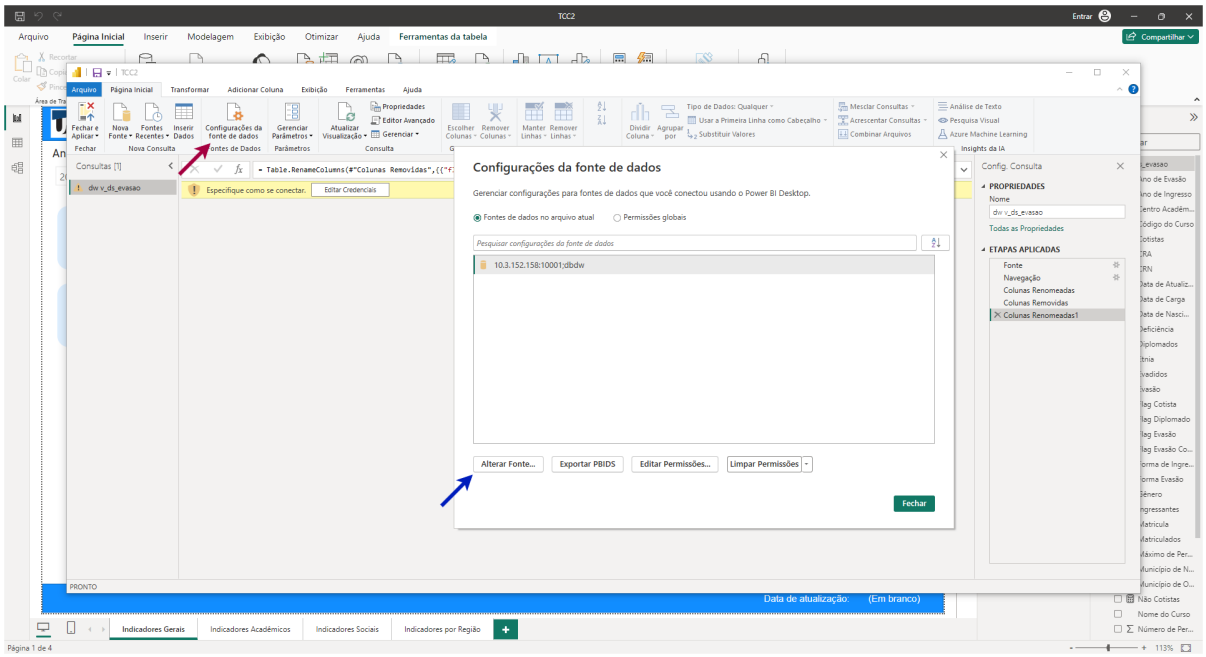

Figura 25: Formulário de atualização de conexões - Criado pelo autor.

Após atualizar a conexão com o DW o usuário pode clicar no botão ”Atualizar”, indicado com a seta amarela na Figura 24, 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.

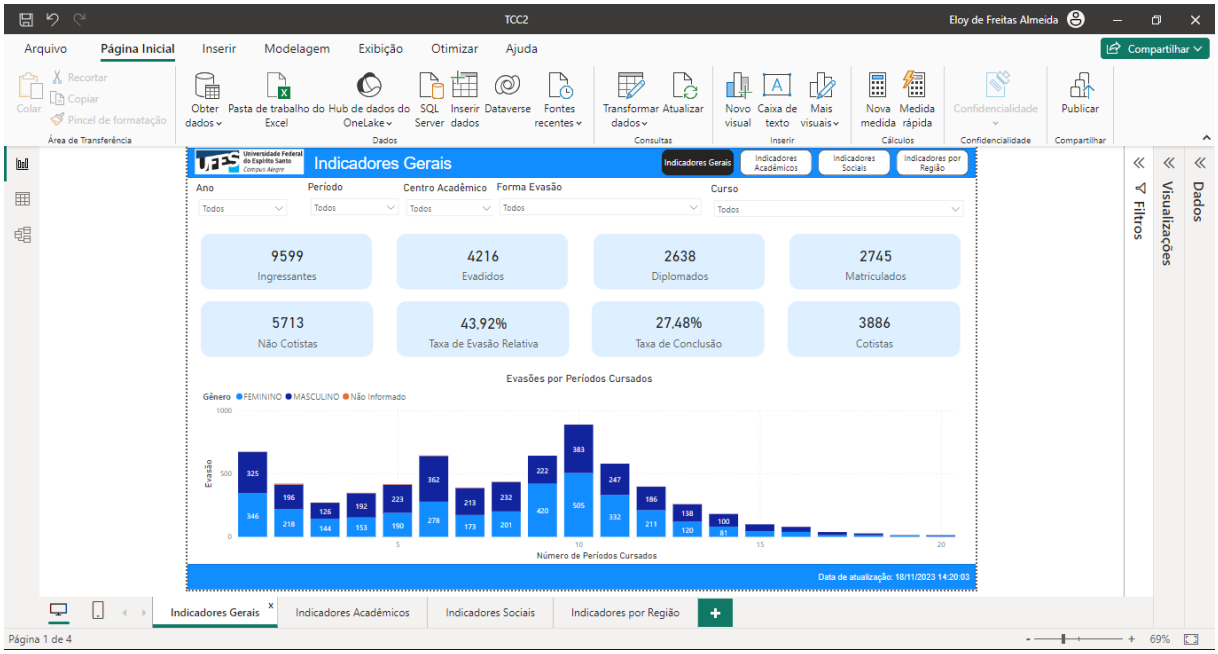

Figura 26: Painel de indicadores gerais carregado - Criado pelo autor.

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
```