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

WERIKIS VENANCIO XAVIER

IMPLEMENTAÇÃO DE UM DATA MART PARA ANÁLISE DE DADOS ABERTOS GOVERNAMENTAIS DA EDUCAÇÃO SUPERIOR NO BRASIL

ALEGRE - ES
SETEMBRO DE 2024

Werikis Venancio Xavier

Implementação de um Data Mart para análise de dados abertos governamentais da educação superior no Brasil

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 parte das exigências para a obtenção do título de Bacharel em Sistemas de Informação.

Orientador: Prof. Antonio Almeida de Barros Junior

Universidade Federal do Espirito Santo – UFES

Departamento da Computação

Alegre - ES

Setembro de 2024

WERIKIS VENANCIO XAVIER

IMPLEMENTAÇÃO DE UM DATA MART PARA ANÁLISE DE DADOS ABERTOS GOVERNAMENTAIS DA EDUCAÇÃO SUPERIOR NO BRASIL

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 parte das exigências para a obtenção do título de Bacharel em Sistemas de Informação.

Aprovado em 27 de Setembro de 2024

COMISSÃO EXAMINADORA

Documento assinado digitalmente
ANTONIO ALMEIDA DE BARROS JUNIOR
Data: 07/10/2024 13:30:50-0300
Verifique em https://validar.iti.gov.br

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

Documento assinado digitalmente
MARCELO OTONE AGUIAR
Data: 07/10/2024 11:01:56-0300
Verifique em https://validar.iti.gov.br

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

Documento assinado digitalmente
RODRIGO FREITAS SILVA
Data: 07/10/2024 13:22:35-0300
Verifique em https://validar.iti.gov.br

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

Agradecimentos

Primeiramente agradecer à Deus por toda a sua graça e benção que foram derramados sobre meus estudos e minha vida ao longo do curso e durante a realização deste trabalho.

Aos meus pais Everaldo Xavier e Lusia Venancio, pela educação, amor e todo apoio em todas as fases da minha vida. Obrigado por sempre acreditarem em mim, amo vocês!

Ao meu orientador Antônio, pelos ensinamentos, esclarecimento de dúvidas 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 ao Grupo de Oração Universtário "Luz do Amor de Deus" do Ministerio Universidades Renovadas que sempre foi meu porto seguro e me deu força para continuar.

Agradeço ao carisma da Fraternidade São João Paulo II, que é o alicerce da minha vida.

Agradeço a todos que auxiliaram direta ou indiretamente para a realização deste trabalho de conclusão de curso.

"Você pode ter dados sem informação, mas
você não pode ter informação sem dados."
(Daniel Keys Moran)

Resumo

Este trabalho propõe o desenvolvimento de um *Data Warehouse* educacional multidimensional com o objetivo de analisar o desempenho do ensino superior no Brasil. Diante da crescente geração de dados, especialmente por instituições governamentais, a criação deste sistema visa mitigar a complexidade na tomada de decisões, oferecendo uma base robusta para a geração de relatórios analíticos. Utilizando dados do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), o projeto possibilita a construção de dashboards interativos que facilitam a visualização de informações, apoiando decisões estratégicas para uma compreensão mais aprofundada da performance das instituições de ensino superior (IES). A implementação de um *Data Mart* para a análise de dados abertos governamentais na educação superior viabiliza uma exploração detalhada do desempenho educacional. Embora haja uma vasta quantidade de informações disponíveis, a sua adequada aplicação às tecnologias de informação e comunicação transforma dados aparentemente irrelevantes em um diferencial significativo. Assim, este trabalho destaca a importância de converter dados em informações, e estas, em conhecimento, que se torna fundamental para decisões assertivas no contexto educacional.

Palavras-chaves: Dados Abertos, Ensino Superior, *Data Warehouse*, *Business Intelligence*.

Lista de figuras

Figura 1 – Arquitetura de um ambiente *Data Warehouse*. 15
Figura 2 – Representação esquema estrela. 17
Figura 3 – Modelagem do DW adotado pelo Autor. 19
Figura 4 – Análise dos cursos de Direito do estado de Tocantins. 20
Figura 5 – Esquema estrela de *Data Warehouse* utilizado na modelagem para Análise de Dados Abertos do Ensino Superior Brasileiro. 21
Figura 6 – Quantitativo de alunos matriculados no ensino superior por ano de ingresso. 22
Figura 7 – Arquitetura do DW proposta para este trabalho. 23
Figura 8 – Ciclo de vida de um projeto de Data Warehouse. 25
Figura 9 – Variáveis do sistema Pentaho. 27
Figura 10 – Arquivo bat para iniciar PDI 28
Figura 11 – Tratamento dos dados de arquivos de diferentes fontes 29
Figura 12 – Prepararação da estrutura de armazenamento temporário. 29
Figura 13 – Modelagem do DW. 30
Figura 14 – Quantidade de ingressantes por ano e área de conhecimento. 32
Figura 15 – Cursos mais demandados. 33
Figura 16 – Crescimento da modalidade a distância. 33
Figura 17 – Quantidade de ingressantes por ano/gênero. 34
Figura 18 – Quantidade de ingressantes por cor/raça 35
Figura 19 – Concluintes por ano, área de conhecimento e região 35

Lista de tabelas

Tabela 1 – Principais diferenças entre banco de dados operacionais e *Data Warehouse* 14
Tabela 2 – Informações dos arquivos do INEP. 18
Tabela 3 – Arquivos disponibilizados pelo INEP compactados. 26

Sumário

Lista de figuras **6**

Lista de tabelas **7**

1 INTRODUÇÃO **10**
1.1 Contextualização **10**
1.2 O problema e sua importância **11**
1.3 Objetivos **12**
1.3.1 Objetivo geral 12
1.3.2 Objetivos específicos 12

2 REVISÃO DE LITERATURA **13**
2.1 Banco de dados **13**
2.2 *Data Warehouse* **13**
2.2.1 Arquitetura de um *Data Warehouse* 15
2.2.2 Modelagem Dimensional 16
2.3 *Business Intelligence* **17**
2.4 Trabalhos relacionados **18**
2.4.1 Trabalho de Ilha (2021) 18
2.4.2 Trabalho de Gomes et al. (2018) 19
2.4.3 Trabalho de Magalhães e Cardoso (2016) 20
2.5 Tecnologias de apoio ao desenvolvimento da aplicação **22**
2.5.1 MySQL *Workbench* 22
2.5.2 Pentaho Data Integration 22
2.5.3 Microsoft Power BI 22

3 METODOLOGIA **23**
3.1 Preparação e Implementação **24**
3.1.1 Análise e definição dos requisitos 24
3.1.2 Obtenção dos dados sobre a Educação Superior Brasileira 25
3.1.3 Pentaho 26
3.2 Pasta Anexos **31**
3.3 Pasta Códigos SQL **31**
3.4 Pasta Pentaho **31**
3.5 Pasta Power BI **31**

4 RESULTADOS E DISCUSSÃO **32**

5 CONCLUSÃO 36

REFERÊNCIAS 38

A CÓDIGOS SQL PARA CRIAÇÃO DAS TABELAS 40

1 Introdução

1.1 Contextualização

A educação é um empreendimento complexo que representa um desafio permanente para a humanidade. O Brasil, ao longo de sua história, tem se deparado com grandes dificuldades em ofertar educação em quantidade e qualidade para as necessidades de desenvolvimento e de cidadania de sua população. A comparação com indicadores educacionais de outros países deixa claro que não se está logrando êxito em questões que há muito tempo deveriam ter sido superadas (VIEIRA; VIDAL, 2007).

Desde 1989 o *International Institute for Management Development* (IMD), sediado na Suíça, publica anualmente um *ranking* de competitividade de 65 países sobre diversos critérios, sendo que em seu relatório mais recente divulgado em 2024, pode-se destacar um cenário preocupante: a amarga posição do Brasil, atualmente na 62ª posição, à frente apenas de Peru, Nigéria, Gana, Argentina e Venezuela. No quesito educação sendo que em uma das rubricas – relativa à educação de crianças e adolescentes e à formação profissional – aparece até mesmo na penúltima posição no ranking anual.

Diante das inúmeras dificuldades e carências relacionadas à educação no país e da incontestável evolução das tecnologias de informação, está se tornando cada vez mais necessária a criação de ferramentas que permitam maior disponibilidade de informação e recursos para o educando, tornando o processo educativo mais dinâmico, eficiente e inovador para tentar reverter essa triste situação da educação brasileira.

Atualmente, vivemos em uma sociedade profundamente influenciada pelas tecnologias da informação e comunicação, onde a busca incessante por novas informações tornou-se uma característica predominante. O conhecimento tornou-se uma das principais forças produtivas do capitalismo, referenciado como o centro da criação de valor. O conhecimento foi convertido em uma nova força produtiva, pode ser partilhado livremente e a custos desprezíveis nas novas redes tecnológicas (SILVA; PEREIRA, 2018).

Luck (2017) destaca os exemplos dos grandes países, como Japão e Coreia do Sul, que priorizaram na educação os investimentos, concentrando recursos e talentos humanos no desenvolvimento da qualidade dos seus sistemas de ensino e de suas escolas, acompanhados de consistentes mecanismos de gestão e de processos. Esses países deram a importância para a gestão educacional, partiram para o novo destino, dando enfoque de visão de conjunto e orientação estratégica de futuro, tendo por base a mobilização de pessoas articuladas em equipe, permitindo articular ações e estabelecer a devida mobilização para maximizar os resultados esperados. Para tanto, a gestão da educação só obtém êxito, quando é subsidiada por informações precisas e integradas.

Nesse contexto, a mineração de dados educacionais é uma área de pesquisa em expansão, onde a finalidade do tratamento dos dados é servir como uma ferramenta para o apoio a tomada de decisões. De acordo com Kimball e Ross (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 facilitam o acesso aos dados, possibilitando aos usuários a criação de consultas que combinam informações de várias formas diferentes com mais praticidade e velocidade.

Dessa forma, no presente trabalho foi posposta a implementação de um banco de dados educacional multidimensional voltado à análise do ensino superior brasileiro, sendo que apesar da grande quantidade de dados abertos disponibilizados junto ao portal INEP de nada valeriam sem um correto tratamento e extração das informações que esses dados podem oferecer. De fato, a aplicação de tecnologias BI em bases de dados educacionais possui grande potencial de diagnóstico, sendo úteis na exploração, análise e recuperação de informações, tendo um papel relevante no entendimento da real situação da educação das Instituições de Ensino Superior brasileiras, consequentemente facilitando o processo tomada de decisão com base na realidade, o que poderia elevar o desempenho das instituições (GONÇALVES, 2021).

1.2 O problema e sua importância

Os principais objetivos de uma instituição de ensino superior podem ser resumidos no fornecimento de ensino de qualidade, formação e capacitação de seus estudantes para o mercado de trabalho.

As faculdades são espaços destinados à difusão do saber e, portanto, todo o esforço nelas realizados devem convergir para a aprendizagem e proporcionar formação de excelência, sendo assim, tornando necessário avaliar os seus resultados produzidos e visar a melhoria contínua. Por consequência, a ciência de dados na área da educação vem avançando nas escolas públicas e privadas possibilitando a realização de análises de desempenho na gestão acadêmica, identificação de possíveis problemas e adoção de medidas pedagógicas e institucionais capazes de solucioná-los.

Pescarini (2020) ressalta no seu artigo que o uso de ferramentas estatísticas para tomada de decisões estão se tornando cada vez mais comuns na administração escolar. Estes mecanismos podem informar intervenções pedagógicas, políticas educacionais e até mesmo decisões sobre a manutenção ou interrupção de certos investimentos na área de educação.

É possível analisar o desempenho da educação superior no Brasil utilizando-se os indicadores educacionais disponibilizados abertamente no site do Instituto Nacional de

Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), sendo que após o tratamento, esses dados poderão servir como uma ferramenta para o apoio a tomada de decisões contribuindo para o sucesso da educação superior brasileira.

1.3 Objetivos

Os objetivos esperados com a conclusão do trabalho serão apresentados nas seções a seguir, bem como os conceitos que motivaram a elaboração do mesmo.

1.3.1 Objetivo geral

O objetivo principal do presente projeto é construir um *Data Mart* que seja capaz de viabilizar a análise do desempenho da educação superior no Brasil. Pretende-se construir um ambiente de *Data Warehouse* (DW)/*Business Intelligence* (BI), baseado na metodologia de Kimball e Ross (2013), que possua um fluxo de *Extract, Treat* e *Load* (ETL) automatizado.

1.3.2 Objetivos específicos

Para atingir esse objetivo geral, delimitou-se os seguintes objetivos específicos:
a) realizar um levantamento e definição dos requisitos para a construção do BI;
b) criar uma metodologia para preparar e carregar o DW com os indicadores educacionais disponibilizados pelo INEP com a implementação de um ambiente de ETL automatizado;
c) criar uma apresentação dos dados, utilizando uma ferramenta de BI;
d) visualizar os indicadores através da criação de *dashboards*, transformando-os em informações significativas.
e) realizar uma análise detalhada e explicação dos resultados encontrados a partir das informações apresentadas nos dashboards.

2 Revisão de literatura

Nos últimos anos várias instituições e empresas de diferentes segmentos – tais como educação, transporte, telecomunicações, segurança, saúde, entre outros – vêm adotando recursos computacionais de *Business Intelligence* (BI), com vistas ao planejamento estratégico, por meio de tecnologias de ETL, DW, OLAP e *Machine Learning*. Tais organizações buscam otimizar o armazenamento e integração de seus dados, além de automatizar a extração de informações, reconhecimento de padrões e descoberta de novos conhecimentos.

Na primeira subseção da revisão bibliográfica, foi realizada uma introdução refente ao conceitos de banco de dados seguida de 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 a utilização da tecnologia de DW como forma de viabilizar a análise do desempenho da educação das Instituições de Ensino Superior (IES) brasileiras.

2.1 Banco de dados

Segundo Elmasrik e Navathe (2011), um banco de dados é uma coleção de dados relacionados, e dados são fatos conhecidos que podem ser registrados e têm um determinado significado implícito, como nomes, números de telefone, endereços, produtos. Essa coleção de dados relacionados com um significado implícito é um banco de dados. Portanto, o que diferencia o banco de dados de uma simples coleção de informações é que um banco de dados representa aspectos do mundo real. As mudanças no mundo real são refletidas no banco. Ademais, um banco é uma coleção logicamente coerente de dados com algum significado inerente, uma vez que uma quantidade aleatória de dados não pode ser denominada corretamente como banco de dados. Por fim, um banco de dados é construído e populado com dados para alguma finalidade específica, definida por um grupo específico de usuários e aplicações.

2.2 *Data Warehouse*

Com a constante evolução tecnológica, a maioria das organizações perceberam que os bancos de dados operacionais poderiam oferecer ainda mais poder competitivo utilizando-os como base de sistemas de suporte à decisão. Consequentemente, gradativamente as organizações foram adaptando seus bancos de dados operacionais para utilizá-los de maneira significativa para dar suporte à decisão (MANNINO, 2008). O suporte à decisão necessita de uma visão abrangente para integrar os processos de negócio. Usar um

banco de dados comum para ambos os tipos de processamento pode diminuir significativamente o desempenho e dificultar a sumarização de atividades entre os processos de negócio. A Tabela 1 mostra as principais diferenças entre um banco operacional e um *Data Warehouse*.

Tabela 1 – Principais diferenças entre banco de dados operacionais e *Data Warehouse*

Caracteristicas	Banco de dados operacional	Data Warehouse
Uso Geral	Atual	Histórico
Nivel de detalhe	Individual	Individual e resumido
Orientação	Orientado a processos	Orientado a assuntos
Quantidade de registros processados	Poucos	Milhares
Nivel de normalização	Geralmente normalizado	Violações frequentes da Forma Normal de Boyce Codd (FNBC)
Nivel de atualização	Volátil	Não volátil
Modelo de dados	Relacional	Modelo relacional com esquema estrela e multidimencional com cubos de dados

Fonte: Adaptado de (MANNINO, 2008).

De acordo com Inmon (2005), um dos principais objetivos do *Data Warehouse* é facilitar o trabalho dos analistas que irão utilizar os dados mais simples do que utilizar um ambiente legado clássico, pelo motivo de existir somente uma fonte de dados, integrando tudo em um lugar único, formando uma base clara para facilitar a reutilização e conciliação dos dados.

> Um *Data Warehouse* é uma coleção de dados orientada por assuntos, integrada, não volátil e variante no tempo, para apoio às decisões da administração. O *Data Warehouse* contém dados corporativos granulares. Os dados do *Data Warehouse* podem ser usados para muitos propósitos diferentes, incluindo dados não utilizados, esperando por requisições futuras que são desconhecidos hoje (INMON, 2005, p.29, tradução nossa).

Essa estrutura deve fornecer as visões mais importantes e relevantes dos dados corporativos, com isso é preciso prover possibilidade e flexibilidade de como os dados serão interpretados e analisados posteriormente, produzindo uma fonte estável de informações históricas, de uma forma consistente e confiável para qualquer consumidor autorizado (GALEMMO et al., 2003).

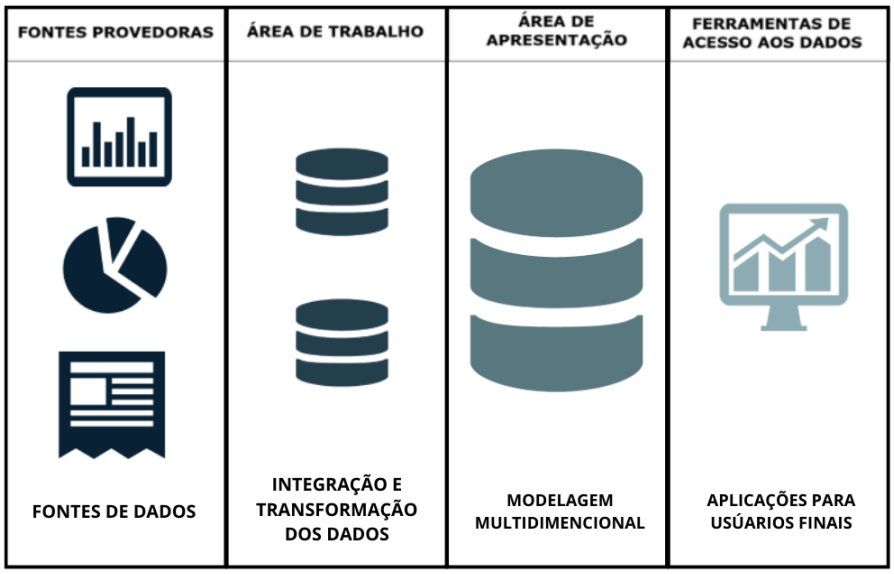
Uma importante característica de um *Data Warehouse* de acordo com o Inmon (2005) é que ele deve ser integrado, visto que o mesmo pode ser alimentado de várias fontes de dados distintas, como por exemplo: .csv, .xlsx, JSON, banco de dados relacionais e outros. Esses dados são padronizados, renomeados, convertidos, resumidos, numerados, esse processo é conhecido como *Extract, Transoform and Load*, comumente chamado de ETL.

Em síntese, os principais objetivos de um DW é manter dados históricos pronto para análise, garantir consistência, confiabilidade e padronização dos dados de acordo com as regras pré-definidas, permitir a interpretação fácil e ter um repositório único e centralizado. Se esses tópicos forem garantidos, significa que a técnica e a modelagem foram bem elaboradas e portanto podem ser incorporados em projetos de *Business Intelligence* (PITON, 2018).

2.2.1 Arquitetura de um *Data Warehouse*

Uma arquitetura de *Data Warehouse* é composta por quatro ambientes: fontes provedoras, área de trabalho, área de apresentação dos dados e ferramentas de acesso. A Figura 1 demonstra uma arquitetura básica de ambiente de DW.

Figura 1 – Arquitetura de um ambiente *Data Warehouse*.



Fonte: Adaptado de (KIMBALL; ROSS, 2013).

As fontes provedoras presentes na Figura 1 representam dados utilizados pela organização, para manter o processo de negócio. Geralmente eles são fisicamente isolados do ambiente de DW servindo de base para extração de dados que são de interesse na análise.

No ambiente de trabalho, a primeira etapa é a extração dos dados presentes nas fontes provedoras. Em seguida, é realizado um tratamento e saneamento do que foi extraído, incluindo um processo de limpeza, racionalização e complementação dos registros, como: padronização de rótulos, tratamento de dados nulos, remoção de duplicidade, conversão de tipo, remoção de ruídos e integração de dados de diversas origens. Após esse processo de tratamento, os dados são carregados na área de apresentação.

Na área de apresentação dos dados, que também é conhecida como ambiente de DW, tem o objetivo formar uma base clara de fácil análise. Existem várias metodologias para a estruturação do ambiente de DW, mas Kimball e Ross (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 a extração de informações.

Por fim, as ferramentas de acesso aos dados possibilitam a exploração do DW pelo usuário final. Variam de simples aplicações para usuários finais e geradores de relatórios, até ferramentas BI de análises sofisticadas e mineração de dados.

2.2.2 Modelagem Dimensional

Para Kimball e Ross (2013) modelagem dimensional é um nome novo para uma técnica antiga usada para criar banco de dados simples e compreensíveis, 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 sistemas de BI.

A modelagem relacional fornece uma estrutura para sistemas de informação que permitam às empresas aceder aos dados de forma muito flexível, fornecendo um alto nível de controle aos dados e aos usuários, visto que a tecnologia permite o agrupamento e exploração dinâmica das informações armazenadas.

Um modelo dimensional consiste em uma única tabela com uma chave composta (chamada tabela de fatos) e um conjunto de tabelas menores chamadas tabelas de dimensão, que possuem chaves simples (consistindo em uma única coluna).

Na verdade, a chave da tabela de fatos são combinações das chaves da tabela de dimensões. Isso faz uma representação gráfica de um modelo dimensional semelhante a uma estrela. Por isso, esse modelo também é chamado de *Star Schema* (KIMBALL; ROSS, 2013).

As tabelas de dimensões contêm principalmente informações textuais que ajudam a definir o componente de dimensão dos negócios, as mesmas possuem informações mais específicas como dados sobre produtos, clientes e períodos de tempo. As tabelas de fatos, por outro lado, contêm dimensões de negócios além dos dados gerados a partir das tabelas de dimensões, o que garante a precisão do uso dos dados por meio de uma extensa estrutura de chaves, eliminando assim pesquisas em tabelas que resultam em maior desempenho possível (HARRISON, 1998). A Figura 2 demonstra uma representação do modelo dimensional citado anteriormente.

Figura 2 – Representação esquema estrela.

Fonte: Adaptado de (MACHADO, 2013).

2.3 *Business Intelligence*

No mundo competitivo do mercado de trabalho, cada vez mais, é necessário que as organizações procurem formas de alcançar um diferencial e, por conseguinte, de se destacarem neste mercado. A tecnologia de *Business Intelligence* é uma solução cada vez mais explorada principalmente pelas grandes e médias empresas, devido à sua capacidade de unificar dados e fornecer aos gestores informações fiáveis e convincentes para ajudar na tomada de decisões (LIMA; BOSCARIOLI, 2012).

De acordo com Peng et al. (2017), a tecnologia de *Business Intelligence* tem como principal objectivo fornecer informação da forma correta e no momento certo, para que a empresa possa tomar as melhores decisões de uma forma mais rápida e confiável. Trata-se de um conjunto de ferramentas e aplicações que permitem aos decisores organizar, analisar, distribuir e assim agir sobre as informações relevantes para o negócio da empresa.

De acordo com Reginato e Nascimento (2017), "as ferramentas de BI podem fornecer uma visão sistêmica do negócio e ajudar na distribuição uniforme de dados entre os utilizadores, e o seu principal objectivo é transformar grandes quantidades em tomadas de decisão". De acordo com os autores, as ferramentas de BI proporcionam o intercâmbio de dados, visualização de informação em várias dimensões e análise de indicadores-chave do desempenho empresarial. Ou seja, a transformação dos dados em informação de qualidade.

Piton (2018) enfatiza a importância do BI, pois com a utilização da tecnologia de análise simplificada de dados é possível assegurar que o processo de tomada de decisões seja mais preciso, garantindo o sucesso organizacional e gerando inúmeras vantagens competitivas no mercado.

2.4 Trabalhos relacionados

Nesta seção, serão apresentados alguns trabalhos relacionados ao tema deste projeto. A pesquisa foi realizada na plataforma Google Acadêmico e utilizou as seguintes palavras-chave: "Data Warehouse sobre indicadores educacionais", "*Data Warehouse* utilizando dados do INEP", "*Data Warehouse* para análise de dados abertos governamentais" e "*Business Intelligence* para análise de dados educacionais". A partir dos resultados coletados, foram selecionadas três obras relacionadas à aplicação de DW nas áreas educacionais. Vale destacar que a escolha dessas obras considerou critérios como relevância para o tema, aplicabilidade prática das soluções e qualidade metodológica, garantindo uma contribuição direta e servindo como base para o presente trabalho.

2.4.1 Trabalho de Ilha (2021)

No trabalho de Ilha (2021) foram utilizados dados disponibilizados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), para propor a criação de um repositório de dados e elaboração de um produto que facilite a extração e a utilização dos indicadores do instituto com os indicadores educacionais, utilizando a tecnologia de *Data Warehouse*, visto que a tecnologia poderia transformar esses dados em informações significativas, com o intuito de subsidiar a tomada de decisão pelos gestores escolares municipais.

A pesquisa desenvolvida e aplicada nas escolas municipais de Santa Maria do estado do Rio Grande do Sul, no período de 03/08/2020 a 31/12/2020, sob autorização da Secretaria de Município da Educação (SMEd) de Santa Maria, utilizando dos seguintes indicadores escolares: Média de alunos por turma; Média de horas-aula diária; Taxas de distorção idade série, Taxas de rendimento; Taxa de não resposta; Regularidade do corpo docente; Esforço docente e Complexidade de Gestão da Escola. A Tabela 2 apresenta as informações sobre os arquivos utilizados durante a realização do estudo:

Tabela 2 – Informações dos arquivos do INEP.

Característica	Descrição
Extensão	(.ods e .xlsx)
Ano	2016, 2017, 2018 e 2019
Tipo de Arquivo	Municípios e Escolas
Código do Município	4316907

Fonte: (ILHA, 2021).

Os resultados evidenciam que o produto transformou dados em conhecimento e contribuiu como instrumento de apoio para a elaboração das estratégias a serem aplicadas

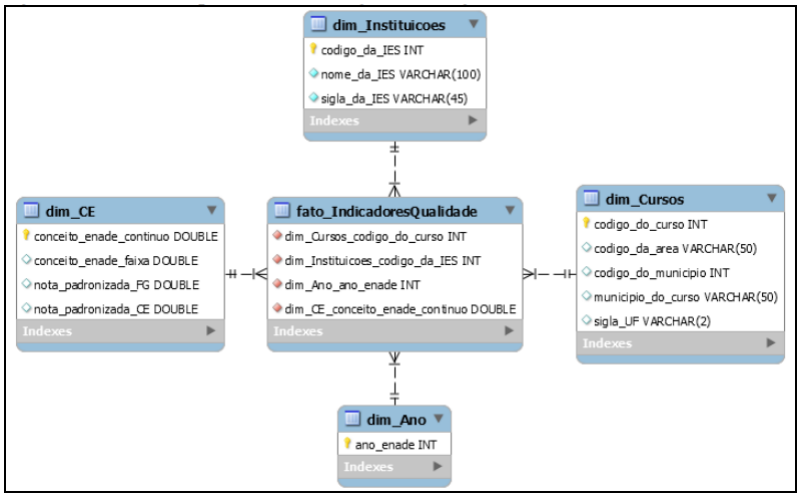
nas escolas, buscando melhorias na qualidade do ensino com foco na aprendizagem do aluno.

2.4.2 Trabalho de Gomes et al. (2018)

No trabalho proposto por Gomes et al. (2018) foi realizado o desenvolvimento um *Data Warehouse* para organizar e analisar parte dos dados do Exame Nacional de Desempenho dos Estudantes (ENADE) que avalia o desempenho dos acadêmicos que cursam o ensino superior por meio de uma prova contendo questões específicas da área e conhecimentos geral, que tem como intuito final verificar o desempenho de tais discentes e outras características referentes às instituições a cada exame aplicado. Dessa forma, para tanto, foi utilizado uma ferramenta OLAP que possibilitaram o processamento analítico e análise eficiente desses dados, mantendo-os de forma estruturada para servirem de apoio às decisões estratégicas.

No trabalho, utilizou-se da modelagem em estrela para a estruturação e/ou arranjo de tabelas de dimensões e tabelas de fatos. Devido às características dos dados coletados e à necessidade de uma consulta eficiente e intuitiva, o modelo construído é apresentado na Figura 3:

Figura 3 – Modelagem do DW adotado pelo Autor.

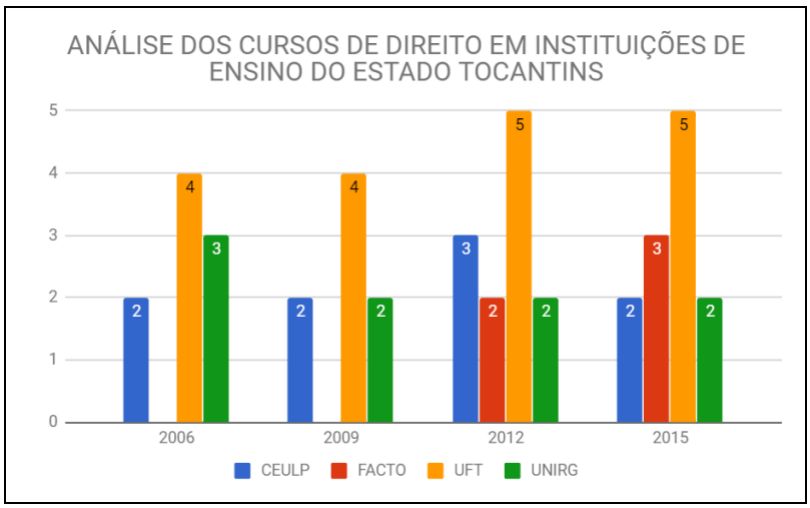


Fonte: Gomes et al. (2018).

O modelo, proposto por Gomes et al. (2018), é constituído por uma única tabela fato, que desempenha o papel central na estrutura dimensional. Essa tabela armazena as medições numéricas, representando uma combinação exclusiva dos valores das chaves de dimensão, que incluem Instituição, Ano, Curso e Conceito, além de agregar as métricas de interesse para o contexto de análise. Essa organização permite uma análise eficiente e precisa dos dados, focando nos principais indicadores relevantes para o negócio.

Com a utilização da ferramenta OLAP tornou-se possível o acompanhamento da evolução de 262 cursos com suas respectivas Instituições de Ensino da região norte do país, durante o período de 2004 até 2015. Com o auxílio da ferramenta foi possível analisar os cursos que elevaram seu conceito ENADE durante os anos, como também aquelas que mantiveram seus conceitos. A Figura 4 apresenta uma das análises realizadas apenas para o curso de Direito presente em algumas Instituições de Ensino do estado do Tocantins, como forma de exemplificação dos resultados obtidos a partir deste presente estudo.

Figura 4 – Análise dos cursos de Direito do estado de Tocantins.



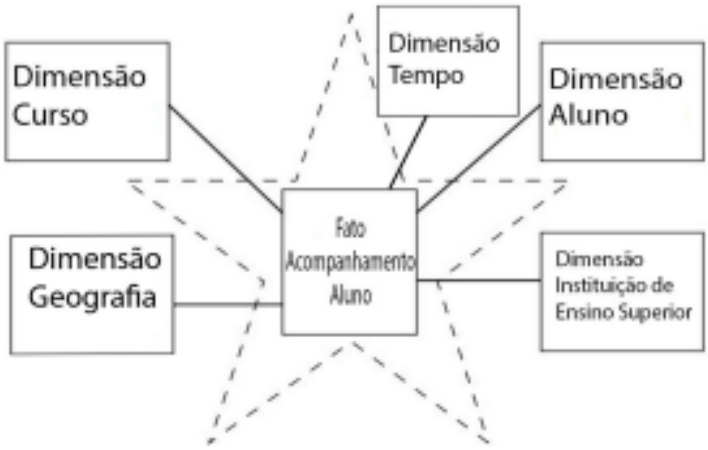
Fonte: Gomes et al. (2018).

2.4.3 Trabalho de Magalhães e Cardoso (2016)

A obra de Magalhães e Cardoso (2016) foi escolhida por ser a que mais se aproxima desta proposta de trabalho. Foi proposta a construção de um ambiente *Data Warehouse* que fosse capaz de possibilitar a análise dos dados abertos disponibilizados pelo Instituto Nacional de Estudos e Pesquisa Anísio Teixeira (INEP) sobre o ensino superior Brasileiro.

No trabalho proposto por Magalhães e Cardoso (2016), foi definido um esquema estrela de *Data Warehouse* e em conformidade com Kimball e Ross (2013), que propõem um modelo dimensional com as mesmas informações de um modelo normalizado. Esse formato oferece maior compreensibilidade, facilidade de consulta, melhor desempenho e maior resistência a mudanças, facilitando a análise dos dados abertos governamentais. O modelo *Star Schema*, representado na Figura 5, foi estruturado com as seguintes dimensões: Aluno, Curso, Geografia, Instituição de Ensino Superior e Tempo. Além das dimensões, o modelo dispõe de uma única tabela fato: Acompanhamento.

Figura 5 – Esquema estrela de *Data Warehouse* utilizado na modelagem para Análise de Dados Abertos do Ensino Superior Brasileiro.



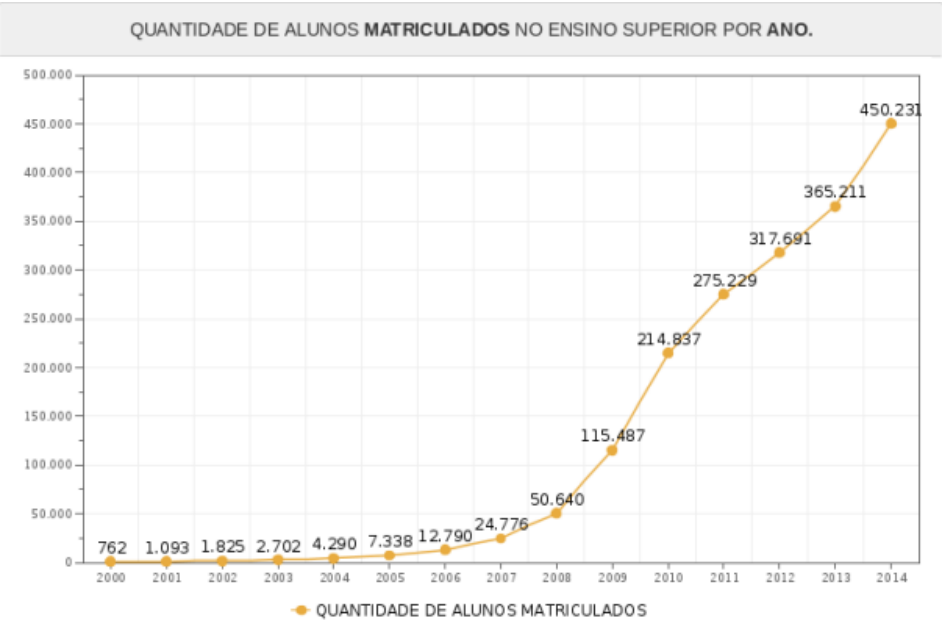
Fonte: Magalhães e Cardoso (2016).

Como resultados obtidos durante a realização desse estudo podemos destacar a geração de relatórios analíticos (*Dashboards*), sendo que a partir dessas ferramentas interativas ficou de fácil acesso a consultas e análises das possíveis demandas aos interessados:

- Quantidade de alunos no ensino superior no decorrer dos anos;
- Sexo dos alunos do ensino superior por grupo de faixa etária;
- Quantidade de alunos matriculados por determinada grande área de conhecimento;
- Quantidade de alunos matriculados no Ensino Superior por curso;
- Níveis de titulação de docentes;
- Raça/cor dos alunos do Ensino Superior por Ano de Ingresso;
- Análise anual dos alunos do Ensino Superior formados por dada região;
- Situação acadêmica dos alunos no Ensino Superior por região.

A Figura 6 apresenta um quantitativo de alunos matriculados no ensino superior por ano de ingresso como forma de exemplificação dos resultados obtidos a partir deste presente estudo.

Figura 6 – Quantitativo de alunos matriculados no ensino superior por ano de ingresso.



Fonte: Magalhães e Cardoso (2016).

2.5 Tecnologias de apoio ao desenvolvimento da aplicação

2.5.1 MySQL *Workbench*

O MySQL *Workbench* é um sistema de código aberto usado gerir bases de dados. O serviço utiliza a linguagem SQL (*Structure Query Language*), que é a linguagem para inserir, acessar e gerenciar o conteúdo armazenado num banco de dados (MYSQL, 2022). Nesse projeto, o MySQL foi utilizado para manter a estrutura física das tabelas do armazenamento temporário e do *Data Warehouse*.

2.5.2 Pentaho Data Integration

Pentaho Data Integration (PDI) fornece as capacidades de Extrair, Transformar, e Carregar (ETL) que facilitam o processo de captura, limpeza, e armazenamento de dados usando um formato uniforme e consistente que é acessível e relevante para os utilizadores finais e tecnologias IoT (PENTAHO, 2024).

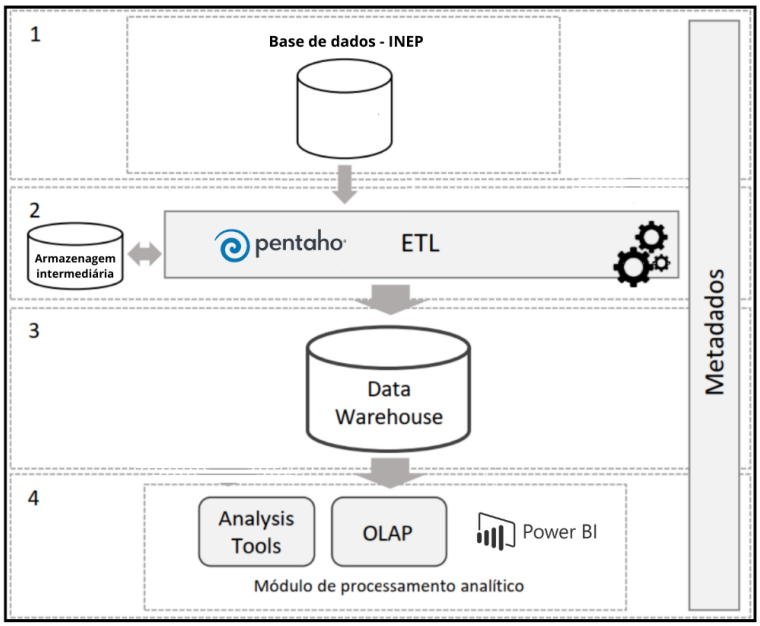
2.5.3 Microsoft Power BI

A ferramenta *Microsoft Power BI* é uma coleção de serviços de *software*, aplicativos e conectores que trabalham juntos para transformar suas fontes de dados não relacionadas em informações coerentes, visualmente envolventes e interativas (MICROSOFT, 2022).

3 Metodologia

No presente trabalho, foi implementado um ambiente de BI baseado na arquitetura de Kimball e Ross (2013), para tanto, foi utilizada uma ferramenta de ETL para povoar um *Data Mart* proposto. A Figura 7 apresenta a arquitetura proposta nesta pesquisa.

Figura 7 – Arquitetura do DW proposta para este trabalho.



Fonte: O autor (2022).

Para a etapa 1 do processo presente na Figura 7, após a definição do escopo do projeto, foram utilizados dados abertos referentes à educação superior brasileira, abrangendo um período de 5 anos, de 2017 a 2021. Esses dados, os mais recentes disponíveis, foram obtidos por meio do portal Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) e passaram por um rigoroso processo de tratamento e mapeamento para a alimentação do *Data Warehouse* proposto.

Os dados utilizados foram disponibilizados em formato CSV *(Comma-Separated Values)*, um tipo de arquivo de texto em que as informações são separadas por delimitadores, no caso, ponto e vírgula. Além dos dados, também foram utilizados os dicionários de dados fornecidos pelo portal do INEP, o que possibilitou a compreensão para manipulação dos dados. A descrição dos campos utilizados nesta pesquisa encontra-se no Apêndice A. Já as versão completa do dicionário de dados, que incluem a descrição detalhada de todos os campos das fontes de dados, estão disponíveis para consulta no portal do INEP (INEP, 2024).

Na sequência, foi preparada uma estrutura de armazenamento temporário de dados *(Staging Area)*. De acordo com Anand e Kumar (2013), esse ambiente ajuda a manter a

qualidade do processo, pois possibilita o agrupamento de fontes heterogêneas de dados, se tornando uma fonte única para o fluxo de ETL. Essa estrutura temporária foi implementada no formato de tabelas relacionais e hospedada em um sistema de gerenciamento de banco de dados (SGDB) relacional de código aberto que funciona em várias plataformas, o MYSQL.

Na segunda etapa, foi criado um processo de ETL utilizando uma ferramenta open source chamada *Pentaho Data Integration*, a mesma que auxiliou na criação de processos de extração, transformação e carga dos dados que serão extraídos da estrutura de armazenamento temporário e preparados e padronizados para alimentar a estrutura do DW.

O modelo proposto baseia-se em uma estrutura composta por uma tabela fato central e três dimensões que organizam os dados de maneira eficiente para a análise. A tabela fato, denominada "ensinosuperior", integra identificadores únicos provenientes das dimensões, além de armazenar informações fundamentais e métricas relevantes para a geração de relatórios e insights. As dimensões associadas ao modelo são "dimTempo", "dimInstituição" e "dimCurso", cada uma responsável por organizar dados específicos relacionados ao período, às instituições de ensino e aos cursos ofertados. Mais detalhes sobre o modelo serão apresentados no tópico 3.1.3.

E por fim, na etapa de apresentação dos dados optou-se pela ferramenta *Microsoft Power BI*, uma coleção de serviços de *software*, aplicativos e conectores que trabalham juntos para transformar suas fontes de dados em informações coerentes, visualmente envolventes e interativas.

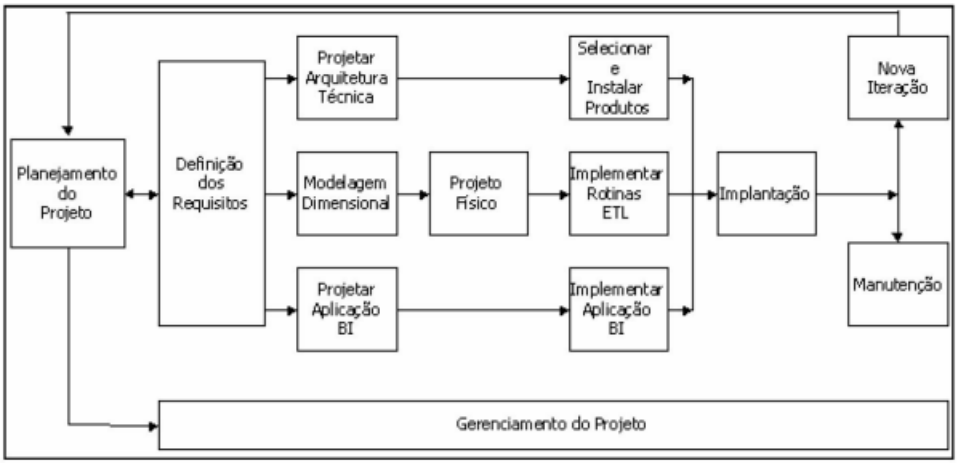
3.1 Preparação e Implementação

Para obter o nível de detalhamento necessário que possibilite uma análise aprofundada e uma melhor compreensão dos dados governamentais voltados à tomada de decisões sobre o Ensino Superior Brasileiro, este trabalho utilizou como base os dados do Censo da Educação Superior, referentes ao período de 2017 a 2021, disponibilizados pelo INEP.

3.1.1 Análise e definição dos requisitos

O desenvolvimento do DW proposto neste trabalho foi baseado no ciclo proposto por Kimball e Ross (2013) conforme ilustra a Figura 8.

Figura 8 – Ciclo de vida de um projeto de Data Warehouse.



Fonte: Adaptado de Kimball e Ross (2013).

A primeira etapa consistiu no planejamento, que envolveu a seleção da base de dados e a definição do objeto de estudo. O foco central foi a análise do desempenho da educação superior no Brasil, utilizando dados disponibilizados pelo INEP. O objetivo principal dessa fase foi elaborar um plano detalhado para a execução das atividades necessárias à construção de um sistema de BI eficaz.

Após o planejamento, procedeu-se à definição dos requisitos, que envolveu o levantamento minucioso das necessidades do negócio. Esse processo visou compreender as demandas e as informações críticas que seriam extraídas dos dados educacionais referentes ao ensino superior. Os requisitos funcionais incluíram a capacidade de extração, transformação e carga (ETL) dos dados, geração de relatórios analíticos e a criação de *dashboards* interativos, permitindo a filtragem e *drill-down* para análises detalhadas. Requisitos não funcionais, como usabilidade, desempenho e escalabilidade, também foram considerados para garantir um *Data Mart* robusto e eficiente. A escolha da base de dados do portal INEP foi crucial, dado que oferece um conjunto abrangente de informações sobre o panorama educacional brasileiro, alinhando-se aos objetivos da análise proposta.

Após a identificação das questões que nortearam a implementação do DW, foi realizado um levantamento que permitiu identificar quais dados seriam mais relevantes para a composição do BI, além de viabilizar uma estruturação adequada para a implementação do projeto. A partir dessa análise, foi possível definir o escopo do projeto e os requisitos necessários ao desenvolvimento do sistema de BI, assegurando que ele atenda de forma eficaz às demandas identificadas.

3.1.2 Obtenção dos dados sobre a Educação Superior Brasileira

Os dados utilizados nesta pesquisa foram obtidos junto ao portal de disponibilização de dados abertos das pesquisas do INEP, inicialmente, foram estudados estrutural-

mente os dados verificando as diferenças anos após ano, vale lembrar que, até o presente momento da construção deste Projeto de Conclusão de Curso, ainda não tinham sido disponibilizados os dados de 2022, por esse motivo foram utilizados os dados de 2017 a 2021, sendo esses os 5 anos mais recentes.

O processo de extração dos dados ocorreu mediante *Download* dos dados fornecidos no seguinte sítio da Internet: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-da-educacao-superior, conforme apresentados na Tabela 3 , arquivos compactados no formato.zip.

Tabela 3 – Arquivos disponibilizados pelo INEP compactados.

Ano de Publicação	Tamanho em megabytes
Microdados 2017	10,1 MB
Microdados 2018	12,6 MB
Microdados 2019	16,1 MB
Microdados 2020	21,2 MB
Microdados 2021	25,3 MB

Fonte: O autor (2024).

De acordo com Kimball e Ross (2013), a capacidade analítica de um ambiente de DW/BI é diretamente proporcional à qualidade e profundidade da dimensão de seus atributos. Quanto mais tempo se investe fornecendo atributos de maneira mais detalhada, melhor será a possibilidade de extração de informações relevantes sobre esses dados.

3.1.3 Pentaho

No desenvolvimento do *Data Mart* para análise dos dados abertos governamentais da educação superior, a implementação da etapa de *Extract, Transform, and Load* (ETL) foi conduzida utilizando a ferramenta Pentaho Data Integration (PDI). Essa escolha se deve à flexibilidade e eficiência do PDI em processar grandes volumes de dados, além de sua capacidade de integração com diversos tipos de fontes de dados e sistemas de gerenciamento de banco de dados.

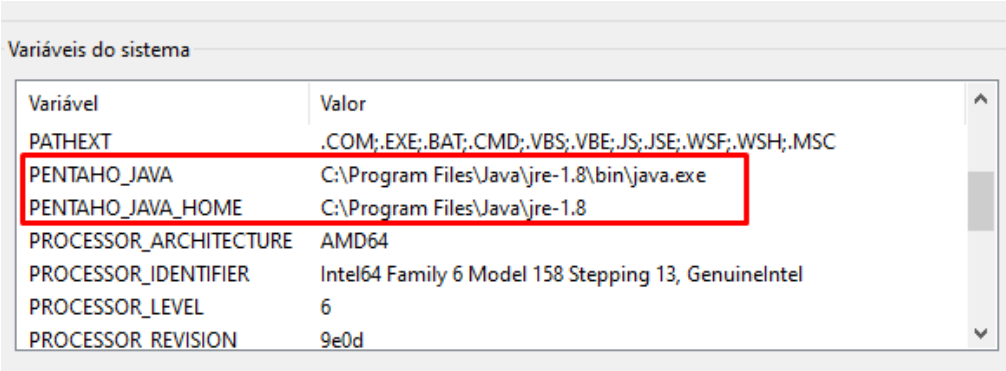
Inicialmente, foi necessário realizar a instalação e configuração do PDI, além de preparar o ambiente de banco de dados que hospedaria o *Data Mart*. Optou-se pelo MySQL, um Sistema de Gerenciamento de Banco de Dados relacional de código aberto, devido à sua robustez e compatibilidade com o PDI. A base de dados criada para o *Data Mart* foi configurada para garantir que todas as permissões necessárias fossem concedidas para a realização das operações de ETL.

O Pentaho é um software disponível pela Hitachi (2024) e possui versão de código fonte aberto, o que permite sua utilização gratuita. Nele é possível realizar a extração,

transformação e carregamento dos dados, de acordo com a necessidade do usuário. Para sua utilização, foi realizado o *download* do PDI através dos documentos disponíveis na rede pela Hitachi. Em seguida, JRE e JDK foram instalados e preparados (a versão usada foi 1.8).

Posteriormente, foi realizada a preparação das variáveis de sistema através das configurações do *Windows*, como ilustrado na Figura 9.

Figura 9 – Variáveis do sistema Pentaho.



Fonte: O autor (2024).

A execução do PDI deve ser feita no arquivo “Spoon.bat”. À princípio não se vê nenhum ícone que o caracterize, portanto é possível transferi-lo como atalho e alterar o seu ícone (o próprio PDI traz um ícone preparado para isso). A Figura 10 destaca o arquivo executável da ferramenta.

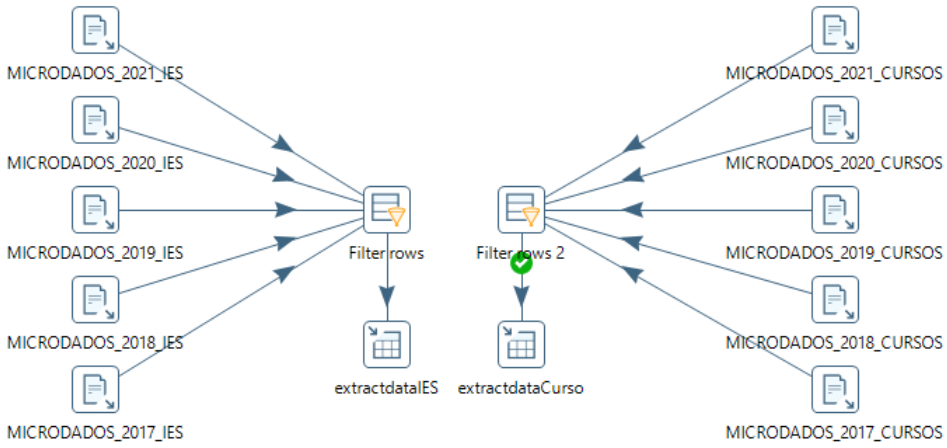
Figura 10 – Arquivo bat para iniciar PDI

Import	08/11/2022 19:07	Arquivo em Lotes ...	2 KB
import	08/11/2022 19:07	Shell Script	2 KB
import-rules	08/11/2022 19:07	Microsoft Edge H...	3 KB
Kitchen	08/11/2022 19:07	Arquivo em Lotes ...	2 KB
kitchen	08/11/2022 19:07	Shell Script	2 KB
LICENSE	08/11/2022 19:07	Documento de Te...	14 KB
Pan	08/11/2022 19:07	Arquivo em Lotes ...	2 KB
pan	08/11/2022 19:07	Shell Script	2 KB
PentahoDataIntegration_OSS_Licenses	08/11/2022 19:00	Microsoft Edge H...	3 KB
purge-utility	08/11/2022 19:07	Arquivo em Lotes ...	2 KB
purge-utility	08/11/2022 19:07	Shell Script	2 KB
README	08/11/2022 19:07	Documento de Te...	2 KB
runSamples	08/11/2022 19:07	Arquivo em Lotes ...	2 KB
runSamples	08/11/2022 19:07	Shell Script	2 KB
set-pentaho-env	08/11/2022 19:07	Arquivo em Lotes ...	6 KB
set-pentaho-env	08/11/2022 19:07	Shell Script	5 KB
Spoon	08/11/2022 19:07	Arquivo em Lotes ...	6 KB
spoon.command	08/11/2022 19:07	Arquivo COMMA...	2 KB
spoon	08/11/2022 19:07	Ícone	204 KB
spoon	08/11/2022 19:07	Arquivo PNG	1 KB
spoon	08/11/2022 19:07	Shell Script	9 KB
SpoonConsole	08/11/2022 19:07	Arquivo em Lotes ...	2 KB
SpoonDebug	08/11/2022 19:07	Arquivo em Lotes ...	3 KB
SpoonDebug	08/11/2022 19:07	Shell Script	2 KB
yarn	08/11/2022 19:07	Shell Script	2 KB

Fonte: O autor (2024).

O INEP disponibiliza arquivos do tipo de formato .csv diferente para cursos e instituições, por ano, em cada arquivos de microdados baixados, dessa forma os mesmos tiveram que serem trabalhados de maneira distintas. Inicialmente os arquivos passaram por um processo de filtragem dos dados para remover aqueles que não possuíam valor ou que foram preenchidos de maneira indevida, tal processo é demonstrado na Figura 11, para que posteriormente pudessem ser combiandos através de registros em comum presentes a ambas fontes de dados.

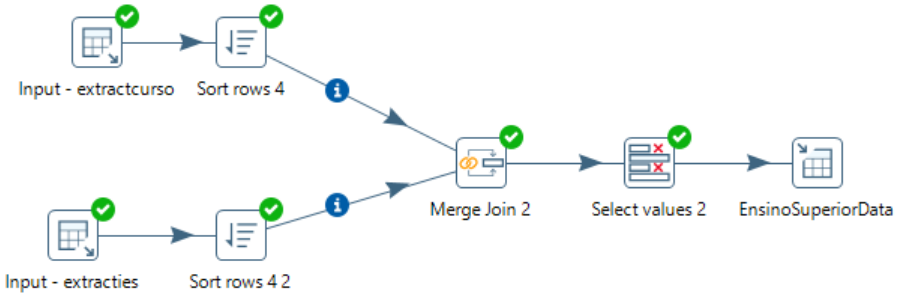
Figura 11 – Tratamento dos dados de arquivos de diferentes fontes



Fonte: O Autor (2024).

Em sequência, foi preparada uma estrutura de armazenamento temporário de dados, denominada EnsinoSuperiorData, esse ambiente de acordo com Anand e Kumar (2013) auxilia na manipulação da qualidade do processo, pois agrupa fontes heterogêneas de dados, se tornando uma fonte única para o fluxo de ETL, essse processo é melhor representado na Figura 12.

Figura 12 – Prepararação da estrutura de armazenamento temporário.

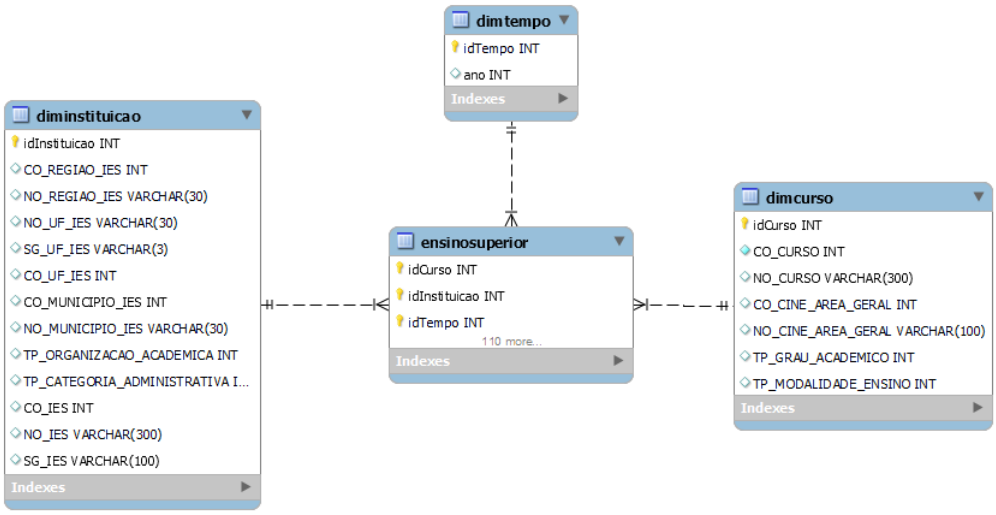


Fonte: O autor (2024);

A base de dados denominada EnsinoSuperiorDataMart, criada específicamente para o armazenamento dos dados extraídos e processados da estrutura de armazenamento temporário para alimentar a estrutura do DW. Dentro desta base, foram criadas diversas tabelas, organizadas em duas categorias principais: dimensões e fatos. As tabelas de dimensão, como DimTempo, DimInstituição e DimCurso, foram estruturadas para armazenar atributos descritivos que auxiliam na segmentação e categorização das análises. Por outro lado, a tabela de fatos EnsinoSuperior foi designada para registrar as métricas quantitativas cruciais para a análise do desempenho acadêmico, incluindo taxas de ingres-

santes e conclusão de cursos, além de dados de evasão. O código SQL de criação dessas tabelas encontra-se em no Apêndice A neste trabalho. A Figura 13 ilustra o modelo.

Figura 13 – Modelagem do DW.



Fonte: O autor (2024).

Um aspecto fundamental na configuração do banco de dados foi a garantia da integridade e consistência dos dados. Isso foi realizado por meio da definição de chaves primárias em todas as tabelas, assegurando a unicidade dos registros, e de chaves estrangeiras, que mantêm a integridade referencial entre as tabelas de fatos e dimensões. Essa estruturação não apenas facilita a realização de consultas complexas, como também simplifica a manutenção do banco de dados.

O processo de ETL, fundamental para a extração dos dados do sistema de arquivos do INEP, sua transferência para a *Staging Area* e o posterior carregamento na estrutura do *Data Warehouse*, foi implementado utilizando a ferramenta *Pentaho Data Integration*. Esse processo garante a coleta, transformação e inserção dos dados de maneira eficiente, facilitando a análise no ambiente de DW.

O ETL envolveu três etapas principais: extração, onde os dados são coletados dos arquivos originais; transformação, que abrange a limpeza, validação e formatação dos dados; e, por fim, a carga, etapa em que os dados são inseridos nas tabelas do banco de dados EnsinoSuperiorDataMart. Esse fluxo de trabalho foi cuidadosamente planejado para garantir que os dados carregados sejam precisos e estejam prontos para análises subsequentes.

Essa nova abordagem e estrutura do banco de dados estão alinhadas com o objetivo de aprofundar e possibilitar a análise dos dados abertos disponibilizados pelo INEP sobre o ensino superior Brasileiro, para diferentes cursos e instituições de ensino, consequentemente gerando informações detalhadas sobre a eficácia das políticas educacionais

brasileiras.

E por fim, com a utilização do Microsoft Power BI Versão: 2.132 foi possível transformar as fontes de dados não relacionadas presentes no DW em informações relevantes para as tomadas de decisão através de relatórios e *dashboards*.

O projeto completo, contendo todos os arquivos e etapas desenvolvidas durante a realização deste trabalho, está disponibilizado no repositório do GitHub, disponível em: https://github.com/werikisxavier/BI_ENSINO_SUPERIOR_BRASILEIRO. Cada diretório foi organizado de forma a facilitar o entendimento e a reprodução do trabalho, sendo detalhado a seguir.

3.2 Pasta Anexos

Esta pasta contém uma cópia do Dicionário de Dados utilizado neste estudo, assim como os documentos oficiais do INEP, que serviram como base para a obtenção e entendimento dos microdados. Esses documentos são fundamentais para compreender a estrutura e os significados dos campos utilizados nas análises.

3.3 Pasta Códigos SQL

Nesta pasta estão organizados os *scripts* em SQL utilizados para a criação das tabelas no banco de dados MySQL. Cada script corresponde a uma etapa do processo de modelagem do *Data Warehouse*, com detalhes sobre a criação das tabelas, índices, chaves estrangeiras e outros aspectos relevantes da estruturação do banco de dados.

3.4 Pasta Pentaho

Este diretório armazena todos os arquivos utilizados para a criação e configuração do fluxo de ETL no *Pentaho*. Estes arquivos incluem as transformações necessárias para integrar os dados brutos obtidos do INEP ao *Data Warehouse*, garantindo a correta formatação e limpeza dos dados. Além disso, o fluxo de ETL foi projetado para ser escalável, permitindo a atualização periódica com novos dados disponibilizados pelo INEP.

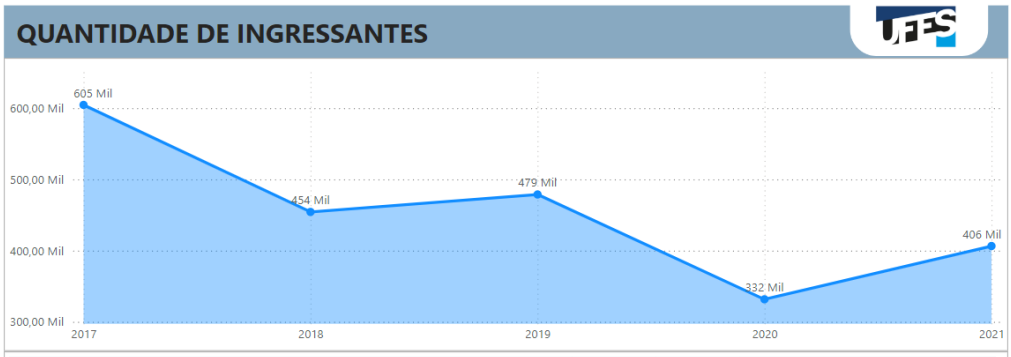
3.5 Pasta Power BI

Nesta pasta encontra-se o arquivo contendo os dashboards gerados através do *Power BI*, que apresentam visualizações interativas dos dados analisados. O arquivo contém todas as métricas e visualizações desenvolvidas, podendo ser reutilizado e ajustado conforme necessário.

4 Resultados e Discussão

A implementação do Data Mart para análise dos dados abertos governamentais da educação superior no Brasil, permitiu uma exploração detalhada e viabilizou a análise do desempenho da educação nas instituições de Ensino Superior (IES) no país. Foi possível evidenciar que as tecnologias DW e OLAP são importantes facilitadoras para descoberta de novos conhecimentos sobre dados abertos educacionais, disponibilizados para toda a comunidade. Há uma grande variedade de dados e informações educacionais disponíveis que, quando não tratados adequadamente, podem parecer irrelevantes, mas quando bem aplicados às tecnologias de informação e comunicação tornam-se um diferencial, visto que os dados são matéria prima da informação, e a informação é a matéria prima do conhecimento, esse conhecimento gerado torna-se base para decisões assertivas (ZINS, 2007). Uma das análises realizadas focou no ingresso de novos estudantes ao longo dos anos. Observou-se uma diminuição no número de ingressantes, como pode ser visto na Figura 14. O ano com o menor índice de novos estudantes foi 2020, com apenas 332 mil ingressantes, justificado principalmente pelos impactos da pandemia de Covid-19 e suas consequências diretas, como a crise econômica e a redução da mobilidade. O SISU, por exemplo, no segundo semestre de 2021, teve uma redução de 40% no número de inscritos em comparação com a mesma etapa de 2019. Ao mesmo tempo, o ENEM registrou 3,1 milhões de inscritos, o menor número desde 2005.

Figura 14 – Quantidade de ingressantes por ano e área de conhecimento.



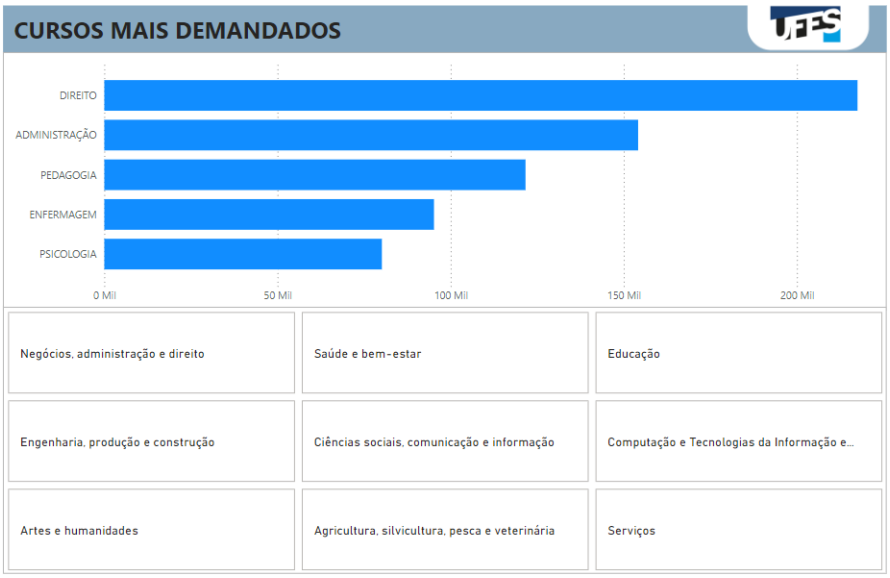
Área de conhecimento	2017	2018	2019	2020	2021
Negócios, administração e direito	199446	142732	153820	92601	106319
Saúde e bem-estar	104689	79462	86745	63277	87842
Educação	78963	61572	64193	47480	48511
Engenharia, produção e construção	80325	54575	49988	37426	45667
Ciências sociais, comunicação e informação	38484	28430	32567	25507	33475
Computação e Tecnologias da Informação e Comunicação (TIC)	29526	25596	26805	19063	23384
Total	**604562**	**454309**	**478821**	**331703**	**406450**

Fonte: O autor (2024).

No que se refere ainda à quantidade de ingressantes por ano e por área de conheci-

mento, a Figura 15 ilustra a possibilidade de consultar os cursos mais demandados. Essa análise permite identificar tendências e preferências dos estudantes em diferentes áreas, oferecendo uma visão mais detalhada dos cursos com maior procura ao longo dos anos.

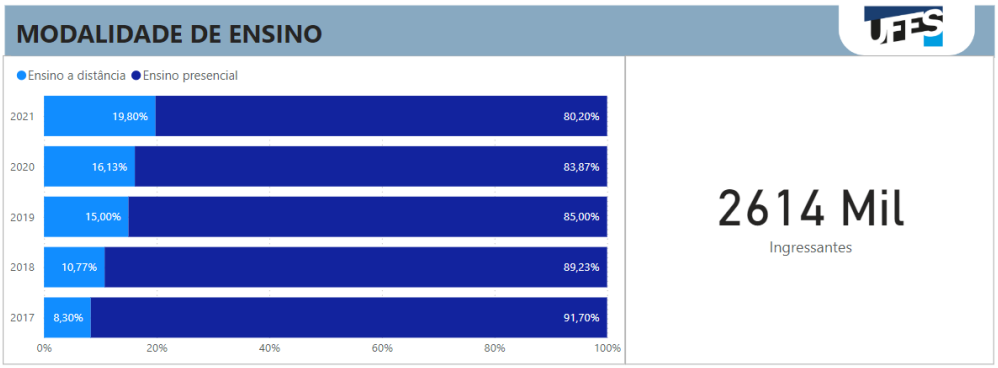
Figura 15 – Cursos mais demandados.



Fonte: O autor (2024).

A realidade imposta pelo isolamento social e pelas medidas de controle sanitário trouxe expressivas consequências às relações de trabalho e educação. Escritórios e salas de aula foram concentrados e compactados em aparelhos, como computadores e celulares, dessa forma, os cursos que adotam a modalidade de Educação a Distância (EaD) cresceram significativamente, confirmando uma tendência anterior a este período, que já apresentava crescimento significativo em relação a modalidade presencial, a tendencia é que com o passar do tempo a discrepância entre as duas modalidades se torne ainda menor. A Figura 16 demonstra o crescimento expressivo da modalidade de EaD ao longo do tempo.

Figura 16 – Crescimento da modalidade a distância.

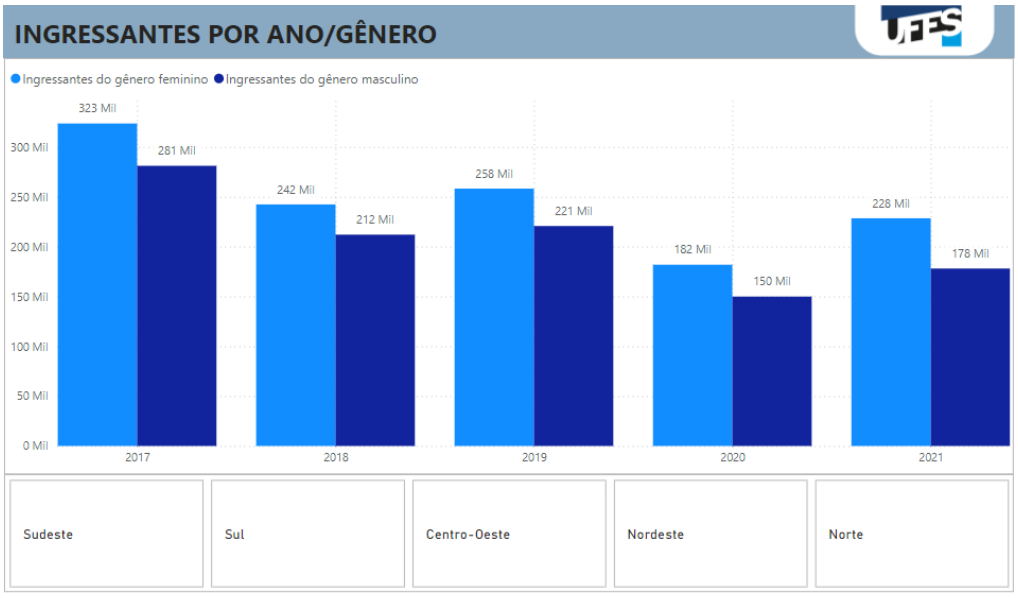


Fonte: O autor (2024).

Com o avanço da cobertura vacinal e a consequente redução da taxa de letalidade da doença, diversas (IES) já planejavam o retorno integral de suas atividades presenciais em 2022, adotando ainda todos os protocolos sanitários de biossegurança, como o uso constante de máscaras adequadas, distanciamento mínimo entre indivíduos e higienização das mãos, além da exigência do passaporte vacinal, justificando o inicio de retomada de ingressantes em 2021 representada na Figura 14.

Ainda no que se diz respeito dos novos ingressantes nas IES ao longo dos anos, a distribuição por sexo dos alunos mostrou variações notáveis entre diferentes faixas etárias. Em geral, há uma predominância feminina, principalmente nas faixas etárias mais jovens (18-24 anos), enquanto as faixas etárias mais velhas (25-40 anos) apresentam uma distribuição mais equilibrada. Este fenômeno pode indicar mudanças nas dinâmicas sociais e culturais que influenciam as escolhas educacionais, sugerindo um envolvimento crescente das mulheres na educação superior imediatamente após o ensino médio. A Figura 17 demonstra a predominância feminina nas diversas instituições de ensino superior espalhadas pelo país.

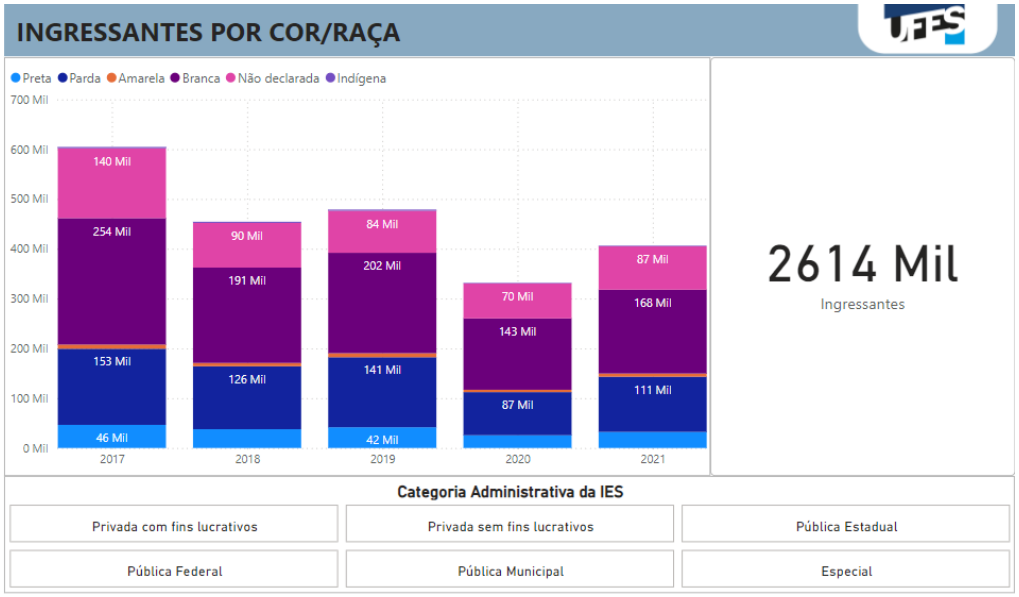
Figura 17 – Quantidade de ingressantes por ano/gênero.



Fonte: O autor (2024).

A análise da distribuição dos novos ingressantes considerando a cor/raça dos alunos, revela variações notáveis entre os diferentes grupos. Em geral, como ilustrado na Figura 18 observa-se uma maior participação de ingressantes que se identificam como brancos, seguidos por uma maior participação de pardos, e pretos, e, por último, de indígenas. Este fenômeno pode ser interpretado através de diversas perspectivas socioeconômicas e culturais que influenciam o acesso à educação superior no Brasil.

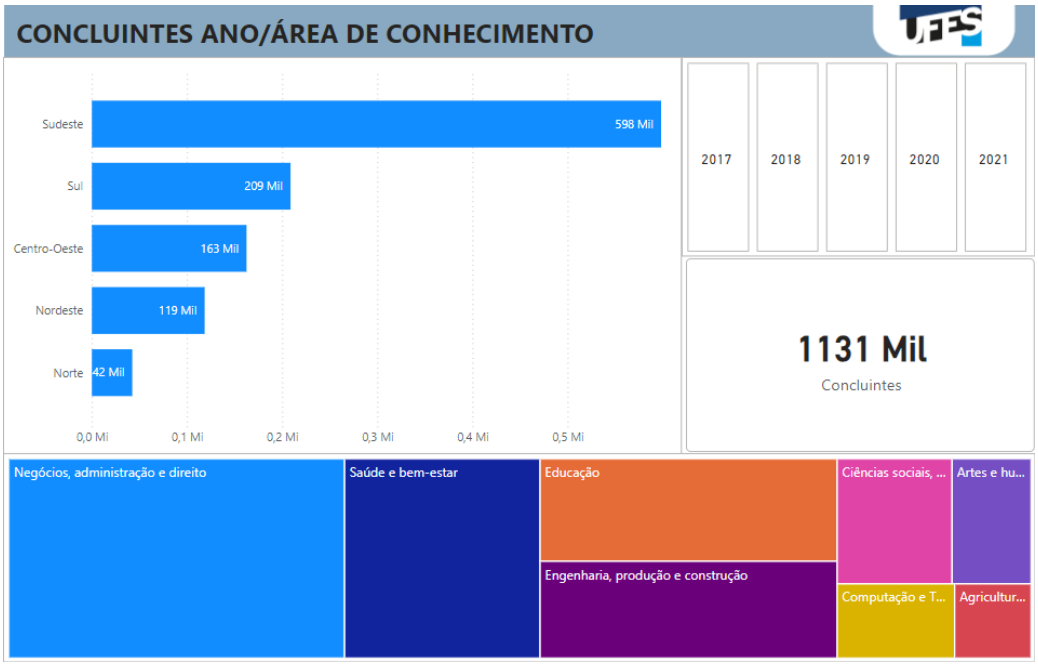
Figura 18 – Quantidade de ingressantes por cor/raça



Fonte: Do autor (2024)

Os dados sobre a formação de alunos por região mostram que as regiões Sudeste e Sul apresentam as maiores taxas de formatura, enquanto as regiões Norte e Nordeste registram as menores. Essa disparidade regional sugere a necessidade de políticas educacionais específicas que visem melhorar as taxas de formatura nas regiões menos favorecidas. A Figura 19 ilustra essa disparidade de concluintes por região no ano de 2020.

Figura 19 – Concluintes por ano, área de conhecimento e região



Fonte: O autor (2024).

5 Conclusão

Este trabalho dedicou-se à análise do desempenho da educação superior no Brasil, utilizando os dados do Censo da Educação Superior fornecidos pelo INEP. O projeto visou a implementação de um banco de dados educacional multidimensional especializado, que poderá ser utilizado como uma ferramenta para o apoio à tomada de decisões contribuindo para o sucesso da educação superior brasileira.

Através das análises realizadas, observou-se um declínio no número de novos ingressantes nas universidades, principalmente no ano de 2020 justificada principalmente pelos impactos pela pandemia de Covid-19, a maior crise econômica global em mais de um século, ao mesmo tempo após esse período, observou-se a confirmação de uma nova tendência de adesão dos estudantes ao ensino a distância, visto que, já vinha apresentando um crescimento significativo em relação à modalidade presencial. A pesquisa também destacou outros desafios, como as disparidades de desempenho em instituições situadas em regiões de maior desenvolvimento econômico, a predominância feminina no ensino superior logo após o ensino médio, e a menor participação de ingressantes nas IES que se identificam como pardos, pretos e indígenas.

Ficou evidente uma diferença significativa entre as instituições situadas em regiões de maior desenvolvimento econômico e aquelas localizadas em regiões menos desenvolvidas. Instituições localizadas em regiões economicamente mais fortes contam com melhor infraestrutura, o que impacta diretamente a qualidade da educação e as oportunidades de pesquisa para estudantes e professores. Por outro lado, a falta de recursos em instituições de regiões menos favorecidas gera obstáculos significativos para o desenvolvimento acadêmico e a inovação.

Quanto às taxas de formatura, o estudo destacou que o sucesso acadêmico é influenciado não apenas por fatores acadêmicos, mas também pelo suporte institucional e pelo acesso a recursos educacionais. Instituições com mais recursos, geralmente situadas em áreas economicamente mais fortes, tendem a obter melhores resultados, destacando-se por suas maiores taxas de formatura. Isso sugere que investimentos em infraestrutura e tecnologia educacional são cruciais para garantir o sucesso dos alunos. A região Sudeste, em particular, se destacou com as maiores taxas de formatura, evidenciando uma discrepância significativa em relação às outras regiões do país. Além disso, a inclusão de mulheres nas faixas etárias mais jovens no ensino superior é um sinal positivo de mudança social, mas as políticas devem continuar a incentivar essa tendência e buscar equilibrar a participação de gênero em todas as áreas de estudo.

Por fim, a construção do Data Mart provou ser uma ferramenta que permite análises profundas e contínuas, e que pode orientar decisões estratégicas no âmbito das políticas educacionais. Este TCC também realça a importância de continuar a coleta e análise de

dados educacionais, para monitorar o impacto das políticas existentes e para fundamentar novas iniciativas. As instituições de ensino superior devem se adaptar continuamente às mudanças nas demandas do mercado de trabalho e às necessidades de sua população estudantil, para manter a relevância e a qualidade de sua oferta educacional.

Desse modo, sugere-se que, para o futuro, seja feita a integração deste sistema com outras bases de dados nacionais e internacionais, como o Censo Escolar, o Exame Nacional do Ensino Médio (ENEM), o Programa Internacional de Avaliação de Alunos (PISA) e com o Instituto Brasileiro de Geografia e Estatística (IBGE), para uma compreensão ainda mais abrangente das dinâmicas educacionais e seu impacto no desenvolvimento socioeconômico do Brasil.

Referências

ANAND, N.; KUMAR, M. An overview on data quality issues at data staging etl. *Conf. on Advances in Signal Processing and Communication*, 2013. 23, 29

ELMASRIK, R.; NAVATHE, S. B. *Sistemas de banco de dados*. 6. ed.. ed. [S.l.]: Pearson Education, 2011. 13

GALEMMO, N. et al. *Mastering Data Warehouse Design: Relational and Dimensional Techniques*. [S.l.]: Wiley Publishing Inc, 2003. 14

GOMES, F. P. et al. Desenvolvimento de um ambiente data warehouse para organizar e analisar os dados do exame nacional de desempenho dos estudantes (enade). In: *ENCOINFO - Congresso de Computação e Tecnologias da Informação*. Palmas, TO, Brasil: [s.n.], 2018. ISSN 2447-0767. Disponível em: <https://ulbra-to.br/encoinfo/edicoes/2018/artigos/desenvolvimento-de-um-ambiente-data-warehouse-para-organizar-e-analisar-os-dados-do-exame-nac >. 8, 19, 20

GONÇALVES, L. M. Uma plataforma de business intelligence para analisar a retenção e evasão do ifmt. *Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Pernambuco*, 2021. Disponível em: <https://repositorio.ufpe.br/handle/123456789/42433>. Acesso em: 2022-11-20. 11

HARRISON, T. H. *Intranet Data Warehouse*. [S.l.]: Berkeley, 1998. 16

HITACHI. *Pentaho Data Integration*. 2024. Disponível em: <https://marketplace.hitachivantara.com/pentaho/>. Acesso em: 2024-05-20. 26

ILHA, L. B. *A Construção de Um Data Warehouse Utilizando os Indicadores educacionais do INEP*. 155 f. Monografia (Dissertação Mestrado em Tecnologias Educacionais em Rede) — Universidade Federal de Santa Maria - UFSM, Santa Maria, 2021. 8, 18

INEP. *Censo da educação superior: microdados*. [S.l.]: INEP, 2024. Disponível em: <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-da-educacao-superior>. Acesso em: 2024-09-15. 23

INMON, B. *Building the Data Warehouse, Fourth Edition*. [S.l.]: Wiley Publishing Inc, 2005. 14

KIMBALL, R.; ROSS, M. *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. 3. ed.. ed. [S.l.]: Wiley, 2013. 11, 12, 15, 16, 20, 23, 24, 25, 26

LIMA, V. M. de; BOSCARIOLI, C. Uso de ferramentas de business intelligence na análise de desempenho de uma empresa de agronegócios. In: *Simpósio Brasileiro de Sistemas de Informação*. São Paulo, SP, Brasil: SBSI, 2012. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/sbsi/2012/0042.pdf>. 17

LUCK, H. *Gestão educacional: uma questão paradigmática*. [S.l.]: Vozes, 2017. 10

MACHADO, F. N. R. *Tecnologia e projeto de Data Warehouse*. 6. ed.. ed. [S.l.]: Editora Érica, 2013. 17

MAGALHÃES, H. da F.; CARDOSO, L. de A. *Análise de Dados Abertos sobre o Ensino Superior Brasileiro*. 45 p. Monografia (conclusão do Curso de Computação — Licenciatura) — Universidade de Brasília, Brasília, 2016. 8, 20, 21, 22

MANNINO, M. V. *Projeto, Desenvolvimento de Aplicações e Administração de Banco de Dados*. 3. ed.. ed. [S.l.]: AMGH, 2008. 13, 14

MICROSOFT. *Microsoft Power BI*. 2022. Disponível em: <https://powerbi.microsoft.com/pt-br/desktop/>. Acesso em: 2022-12-15. 22

MYSQL. *MySQL Workbench*. 2022. Disponível em: <https://www.mysql.com/products/workbench/>. Acesso em: 2022-11-20. 22

PENTAHO. *Pentaho Data Integration*. 2024. Disponível em: <https://marketplace.hitachivantara.com/pentaho/>. Acesso em: 2024-05-20. 22

PESCARINI, T. Data science: a ciência de dados avança na gestão educacional. *Revista Educação*, v. 263, 2020. Disponível em: <https://revistaeducacao.com.br/2020/01/29/CIENCIA-DE-DADOS-GESTAO/>. Acesso em: 2022-11-20. 11

PITON, R. *Data Warehouse Passo a Passo: O Guia Prático de Como Construir um Data Warehouse do Zero*. [S.l.: s.n.], 2018. 15, 17

REGINATO, L.; NASCIMENTO, A. M. Um estudo de caso envolvendo business intelligence como instrumento de apoio à controladoria. *Rev. contab. finanç.*, v. 18, p. 69–83, 2017. 17

SILVA, M. A. da; PEREIRA, R. da S. *Gestão escolar e o trabalho do diretor*. 1. ed.. ed. Curitiba: Appris, 2018. 10

VIEIRA, S. L.; VIDAL, E. M. A educação brasileira ficando “mais” para trás. In: *XVIII Encontro de Pesquisa Educacional do Norte e Nordeste*. [S.l.: s.n.], 2007. 10

ZINS, C. Conceptual approaches for defining data, information, and knowledge. *Journal of the American Society for Information Science and Technology*, v. 58, p. 479–493, 2007. 32

A Códigos SQL para criação das tabelas

1. **Tabela temporária extractCurso:** Tabela criada no banco de dados, com o seguinte comando SQL, para armazenar temporariamente os dados extraídos dos arquivos .csv dos cursos.

```
CREATE TABLE extractCurso (
   NU_ANO_CENSO INT NOT NULL,
CO_REGIAO int,
   NO_REGIAO VARCHAR(30),
   NO_UF VARCHAR(30),
   SG_UF VARCHAR(3),
   CO_UF int,
   CO_MUNICIPIO int,
   NO_MUNICIPIO VARCHAR(30),
TP_ORGANIZACAO_ACADEMICA VARCHAR(300),
   TP_CATEGORIA_ADMINISTRATIVA VARCHAR(300),
TP_REDE INT,
   CO_IES INT,
   NO_CURSO VARCHAR(300),
   CO_CURSO INT,
   CO_CINE_AREA_GERAL INT,
NO_CINE_AREA_GERAL VARCHAR(100),
   TP_GRAU_ACADEMICO VARCHAR(300),
   IN_GRATUITO VARCHAR(300),
   TP_MODALIDADE_ENSINO INT,
   QT_VG_TOTAL INT,
   QT_INSCRITO_TOTAL INT,
   QT_ING INT,
QT_ING_FEM INT,
QT_ING_MASC INT,
QT_ING_0_17 INT ,
QT_ING_18_24 INT ,
QT_ING_25_29 INT ,
QT_ING_30_34 INT ,
QT_ING_35_39 INT ,
QT_ING_40_49 INT ,
QT_ING_50_59 INT ,
QT_ING_60_MAIS INT,
QT_ING_BRANCA INT ,
```

```
QT_ING_PRETA INT,
QT_ING_PARDA INT ,
QT_ING_AMARELA INT ,
QT_ING_INDIGENA INT ,
QT_ING_CORND INT ,
QT_MAT INT ,
QT_MAT_FEM INT ,
QT_MAT_MASC INT ,
QT_MAT_0_17 INT ,
QT_MAT_18_24 INT,
QT_MAT_25_29 INT,
QT_MAT_30_34 INT,
QT_MAT_35_39 INT,
QT_MAT_40_49 INT,
QT_MAT_50_59 INT,
QT_MAT_60_MAIS INT,
QT_MAT_BRANCA INT,
QT_MAT_PRETA INT,
QT_MAT_PARDA INT,
QT_MAT_AMARELA INT,
QT_MAT_INDIGENA INT,
QT_MAT_CORND INT,
QT_CONC INT,
QT_CONC_FEM INT,
QT_CONC_MASC INT,
QT_CONC_0_17 INT,
QT_CONC_18_24 INT,
QT_CONC_25_29 INT,
QT_CONC_30_34 INT,
QT_CONC_35_39 INT,
QT_CONC_40_49 INT,
QT_CONC_50_59 INT,
QT_CONC_60_MAIS INT,
QT_CONC_BRANCA INT,
QT_CONC_PRETA INT,
QT_CONC_PARDA INT,
QT_CONC_AMARELA INT,
QT_CONC_INDIGENA INT,
QT_CONC_CORND INT,
```

```
 QT_ALUNO_DEFICIENTE INT,
 QT_ING_DEFICIENTE INT,
 QT_MAT_DEFICIENTE INT,
 QT_CONC_DEFICIENTE INT,
 QT_ING_FINANC INT,
 QT_ING_FIES INT,
 QT_ING_PROUNII INT,
 QT_ING_PROUNIP INT,
 QT_MAT_FINANC INT,
 QT_MAT_FIES INT,
 QT_MAT_PROUNII INT,
 QT_MAT_PROUNIP INT,
 QT_CONC_FINANC INT,
 QT_CONC_FIES INT,
 QT_CONC_PROUNII INT,
 QT_CONC_PROUNIP INT,
 QT_MAT_RVREDEPUBLICA INT,
 QT_CONC_RVREDEPUBLICA INT,
 QT_SIT_TRANCADA INT,
 QT_SIT_DESVINCULADO INT,
 QT_SIT_TRANSFERIDO INT,
 QT_SIT_FALECIDO INT,
 QT_ING_PROCESCPUBLICA INT,
 QT_ING_PROCESCPRIVADA INT,
 QT_MAT_PROCESCPUBLICA INT,
 QT_MAT_PROCESCPRIVADA INT,
 QT_CONC_PROCESCPUBLICA INT,
 QT_CONC_PROCESCPRIVADA INT,
    CONSTRAINT pk_Extract primary key(
    NU_ANO_CENSO,
    CO_MUNICIPIO,
    CO_IES,CO_CURSO)
);
```

2. **Tabela temporária extractIES:** Tabela criada no banco de dados, com o seguinte comando SQL, para armazenar temporariamente os dados extraídos dos arquivos .csv das instituições.

```
 CREATE TABLE extractIES (
    NU_ANO_CENSO INT,
```

```
   CO_REGIAO_IES INT,
NO_REGIAO_IES VARCHAR(30),
NO_UF_IES VARCHAR(30),
   SG_UF_IES VARCHAR(3),
   CO_UF_IES INT,
CO_MUNICIPIO_IES INT,
   NO_MUNICIPIO_IES VARCHAR(30),
TP_ORGANIZACAO_ACADEMICA INT,
   TP_CATEGORIA_ADMINISTRATIVA INT,
   CO_IES INT,
NO_IES VARCHAR(300),
   SG_IES VARCHAR(100),
QT_DOC_TOTAL INT,
QT_DOC_EXE INT,
QT_DOC_EX_FEMI INT,
QT_DOC_EX_MASC INT,
QT_DOC_EX_SEM_GRAD INT,
QT_DOC_EX_GRAD INT,
QT_DOC_EX_ESP INT,
QT_DOC_EX_MEST INT,
QT_DOC_EX_DOUT INT,
QT_DOC_EX_0_29 INT,
QT_DOC_EX_30_34 INT,
QT_DOC_EX_35_39 INT,
QT_DOC_EX_40_44 INT,
QT_DOC_EX_45_49 INT,
QT_DOC_EX_50_54 INT,
QT_DOC_EX_55_59 INT,
QT_DOC_EX_60_MAIS INT,
QT_DOC_EX_BRANCA INT,
QT_DOC_EX_PRETA INT,
QT_DOC_EX_PARDA INT,
QT_DOC_EX_AMARELA INT,
QT_DOC_EX_INDIGENA INT,
QT_DOC_EX_COR_ND INT,
QT_DOC_EX_BRA INT,
QT_DOC_EX_EST INT,
QT_DOC_EX_COM_DEFICIENCIA INT,
   CONSTRAINT pk_Extract primary key(
```

```
NU_ANO_CENSO,
CO_MUNICIPIO_IES,
CO_IES)
);
```

3. **Tabela temporária EnsinoSuperiorData:** Tabela criada no banco de dados, com o seguinte comando SQL, para armazenar temporariamente os dados combinados através de registros em comum presentes a ambas fontes de dados.

```
CREATE TABLE EnsinoSuperiorData (
 NU_ANO_CENSO INT NOT NULL,
    CO_REGIAO_IES INT,
 NO_REGIAO_IES VARCHAR(30),
 NO_UF_IES VARCHAR(30),
    SG_UF_IES VARCHAR(30),
    CO_UF_IES INT,
 CO_MUNICIPIO_IES INT,
    NO_MUNICIPIO_IES VARCHAR(30),
 TP_ORGANIZACAO_ACADEMICA INT,
    TP_CATEGORIA_ADMINISTRATIVA INT,
 TP_ORGANIZACAO_ACADEMICA_1 INT,
    TP_CATEGORIA_ADMINISTRATIVA_1 INT,
    CO_IES INT,
 NO_IES VARCHAR(300),
    SG_IES VARCHAR(100),
 QT_DOC_TOTAL INT,
 QT_DOC_EXE INT,
 QT_DOC_EX_FEMI INT,
 QT_DOC_EX_MASC INT,
 QT_DOC_EX_SEM_GRAD INT,
 QT_DOC_EX_GRAD INT,
 QT_DOC_EX_ESP INT,
 QT_DOC_EX_MEST INT,
 QT_DOC_EX_DOUT INT,
 QT_DOC_EX_0_29 INT,
 QT_DOC_EX_30_34 INT,
 QT_DOC_EX_35_39 INT,
 QT_DOC_EX_40_44 INT,
 QT_DOC_EX_45_49 INT,
 QT_DOC_EX_50_54 INT,
```

```
QT_DOC_EX_55_59 INT,
QT_DOC_EX_60_MAIS INT,
QT_DOC_EX_BRANCA INT,
QT_DOC_EX_PRETA INT,
QT_DOC_EX_PARDA INT,
QT_DOC_EX_AMARELA INT,
QT_DOC_EX_INDIGENA INT,
QT_DOC_EX_COR_ND INT,
QT_DOC_EX_BRA INT,
QT_DOC_EX_EST INT,
QT_DOC_EX_COM_DEFICIENCIA INT,
   TP_REDE INT ,
   NO_CURSO VARCHAR(300),
   CO_CURSO INT ,
   CO_CINE_AREA_GERAL INT ,
NO_CINE_AREA_GERAL VARCHAR(100),
   TP_GRAU_ACADEMICO INT,
   TP_MODALIDADE_ENSINO INT,
   CO_REGIAO INT,
NO_REGIAO VARCHAR(30),
NO_UF VARCHAR(30),
   SG_UF VARCHAR(3),
   CO_UF INT,
CO_MUNICIPIO INT,
   NO_MUNICIPIO VARCHAR(30),
   QT_VG_TOTAL INT,
   QT_INSCRITO_TOTAL INT,
   QT_ING INT,
QT_ING_FEM INT,
QT_ING_MASC INT,
QT_ING_0_17 INT ,
QT_ING_18_24 INT ,
QT_ING_25_29 INT ,
QT_ING_30_34 INT ,
QT_ING_35_39 INT ,
QT_ING_40_49 INT ,
QT_ING_50_59 INT ,
QT_ING_60_MAIS INT,
QT_ING_BRANCA INT ,
```

```
QT_ING_PRETA INT,
QT_ING_PARDA INT ,
QT_ING_AMARELA INT ,
QT_ING_INDIGENA INT ,
QT_ING_CORND INT ,
QT_MAT INT ,
QT_MAT_FEM INT ,
QT_MAT_MASC INT ,
QT_MAT_0_17 INT ,
QT_MAT_18_24 INT,
QT_MAT_25_29 INT,
QT_MAT_30_34 INT,
QT_MAT_35_39 INT,
QT_MAT_40_49 INT,
QT_MAT_50_59 INT,
QT_MAT_60_MAIS INT,
QT_MAT_BRANCA INT,
QT_MAT_PRETA INT,
QT_MAT_PARDA INT,
QT_MAT_AMARELA INT,
QT_MAT_INDIGENA INT,
QT_MAT_CORND INT,
QT_CONC INT,
QT_CONC_FEM INT,
QT_CONC_MASC INT,
QT_CONC_0_17 INT,
QT_CONC_18_24 INT,
QT_CONC_25_29 INT,
QT_CONC_30_34 INT,
QT_CONC_35_39 INT,
QT_CONC_40_49 INT,
QT_CONC_50_59 INT,
QT_CONC_60_MAIS INT,
QT_CONC_BRANCA INT,
QT_CONC_PRETA INT,
QT_CONC_PARDA INT,
QT_CONC_AMARELA INT,
QT_CONC_INDIGENA INT,
QT_CONC_CORND INT,
```

```
 QT_ALUNO_DEFICIENTE INT,
 QT_ING_DEFICIENTE INT,
 QT_MAT_DEFICIENTE INT,
 QT_CONC_DEFICIENTE INT,
 QT_ING_FINANC INT,
 QT_ING_FIES INT,
 QT_ING_PROUNII INT,
 QT_ING_PROUNIP INT,
 QT_MAT_FINANC INT,
 QT_MAT_FIES INT,
 QT_MAT_PROUNII INT,
 QT_MAT_PROUNIP INT,
 QT_CONC_FINANC INT,
 QT_CONC_FIES INT,
 QT_CONC_PROUNII INT,
 QT_CONC_PROUNIP INT,
 QT_MAT_RVREDEPUBLICA INT,
 QT_CONC_RVREDEPUBLICA INT,
 QT_SIT_TRANCADA INT,
 QT_SIT_DESVINCULADO INT,
 QT_SIT_TRANSFERIDO INT,
 QT_SIT_FALECIDO INT,
 QT_ING_PROCESCPUBLICA INT,
 QT_ING_PROCESCPRIVADA INT,
 QT_MAT_PROCESCPUBLICA INT,
 QT_MAT_PROCESCPRIVADA INT,
 QT_CONC_PROCESCPUBLICA INT,
 QT_CONC_PROCESCPRIVADA INT,
    CONSTRAINT pk_Extract primary key(
NU_ANO_CENSO,
CO_MUNICIPIO,
CO_IES,
CO_CURSO)
);
```

4. **Tabela Dimensão Tempo:** Tabela criada no banco de dados, com o seguinte comando SQL, para armazenar o e suas respectivas colunas com seus respectivos atributos.

```
CREATE TABLE dimTempo (
```

```
  idTempo INT NOT NULL,
  ano INT DEFAULT NULL,
  PRIMARY KEY (idTempo)
);
```

5. **Tabela Dimensão Curso:** Tabela criada no banco de dados, com o seguinte comando SQL, para armazenar o e suas respectivas colunas com seus respectivos atributos.

```
CREATE TABLE dimCurso (
  idCurso int NOT NULL,
    CO_CURSO INT NOT NULL,
    NO_CURSO VARCHAR(300),
    CO_CINE_AREA_GERAL INT,
    NO_CINE_AREA_GERAL VARCHAR(100),
 TP_GRAU_ACADEMICO INT,
    TP_MODALIDADE_ENSINO INT,
  PRIMARY KEY (idCurso)
);
```

6. **Tabela Dimensão Instituição :** Tabela criada no banco de dados, com o seguinte comando SQL, para armazenar o e suas respectivas colunas com seus respectivos atributos.

```
CREATE TABLE dimInstituicao (
    idInstituicao INT NOT NULL,
 CO_REGIAO_IES INT,
 NO_REGIAO_IES VARCHAR(30),
 NO_UF_IES VARCHAR(30),
    SG_UF_IES VARCHAR(3),
    CO_UF_IES INT,
 CO_MUNICIPIO_IES INT,
    NO_MUNICIPIO_IES VARCHAR(30),
 TP_ORGANIZACAO_ACADEMICA INT,
    TP_CATEGORIA_ADMINISTRATIVA INT,
    CO_IES INT,
 NO_IES VARCHAR(300),
    SG_IES VARCHAR(100),
    PRIMARY KEY (idInstituicao)
);
```

7. **Tabela Fato Ensino Superior:** Tabela criada no banco de dados, com o seguinte comando SQL, para armazenar o e suas respectivas colunas com seus respectivos atributos.

```
CREATE TABLE ensinosuperior (
 idCurso int NOT NULL,
 idInstituicao int NOT NULL,
 idTempo int NOT NULL,
 QT_DOC_TOTAL INT,
 QT_DOC_EXE INT,
 QT_DOC_EX_FEMI INT,
 QT_DOC_EX_MASC INT,
 QT_DOC_EX_SEM_GRAD INT,
 QT_DOC_EX_GRAD INT,
 QT_DOC_EX_ESP INT,
 QT_DOC_EX_MEST INT,
 QT_DOC_EX_DOUT INT,
 QT_DOC_EX_0_29 INT,
 QT_DOC_EX_30_34 INT,
 QT_DOC_EX_35_39 INT,
 QT_DOC_EX_40_44 INT,
 QT_DOC_EX_45_49 INT,
 QT_DOC_EX_50_54 INT,
 QT_DOC_EX_55_59 INT,
 QT_DOC_EX_60_MAIS INT,
 QT_DOC_EX_BRANCA INT,
 QT_DOC_EX_PRETA INT,
 QT_DOC_EX_PARDA INT,
 QT_DOC_EX_AMARELA INT,
 QT_DOC_EX_INDIGENA INT,
 QT_DOC_EX_COR_ND INT,
 QT_DOC_EX_BRA INT,
 QT_DOC_EX_EST INT,
 QT_DOC_EX_COM_DEFICIENCIA INT,
 QT_VG_TOTAL INT,
 QT_INSCRITO_TOTAL INT,
 QT_ING INT,
 QT_ING_FEM INT,
 QT_ING_MASC INT,
 QT_ING_0_17 INT,
```

```
QT_ING_18_24 INT,
QT_ING_25_29 INT,
QT_ING_30_34 INT,
QT_ING_35_39 INT,
QT_ING_40_49 INT,
QT_ING_50_59 INT,
QT_ING_60_MAIS INT,
QT_ING_BRANCA INT,
QT_ING_PRETA INT,
QT_ING_PARDA INT,
QT_ING_AMARELA INT,
QT_ING_INDIGENA INT,
QT_ING_CORND INT,
QT_MAT INT,
QT_MAT_FEM INT,
QT_MAT_MASC INT,
QT_MAT_0_17 INT,
QT_MAT_18_24 INT,
QT_MAT_25_29 INT,
QT_MAT_30_34 INT,
QT_MAT_35_39 INT,
QT_MAT_40_49 INT,
QT_MAT_50_59 INT,
QT_MAT_60_MAIS INT,
QT_MAT_BRANCA INT,
QT_MAT_PRETA INT,
QT_MAT_PARDA INT,
QT_MAT_AMARELA INT,
QT_MAT_INDIGENA INT,
QT_MAT_CORND INT,
QT_CONC INT,
QT_CONC_FEM INT,
QT_CONC_MASC INT,
QT_CONC_0_17 INT,
QT_CONC_18_24 INT,
QT_CONC_25_29 INT,
QT_CONC_30_34 INT,
QT_CONC_35_39 INT,
QT_CONC_40_49 INT,
```

```
 QT_CONC_50_59 INT,
 QT_CONC_60_MAIS INT,
 QT_CONC_BRANCA INT,
 QT_CONC_PRETA INT,
 QT_CONC_PARDA INT,
 QT_CONC_AMARELA INT,
 QT_CONC_INDIGENA INT,
 QT_CONC_CORND INT,
 QT_ALUNO_DEFICIENTE INT,
 QT_ING_DEFICIENTE INT,
 QT_MAT_DEFICIENTE INT,
 QT_CONC_DEFICIENTE INT,
 QT_ING_FINANC INT,
 QT_ING_FIES INT,
 QT_ING_PROUNII INT,
 QT_ING_PROUNIP INT,
 QT_MAT_FINANC INT,
 QT_MAT_FIES INT,
 QT_MAT_PROUNII INT,
 QT_MAT_PROUNIP INT,
 QT_CONC_FINANC INT,
 QT_CONC_FIES INT,
 QT_CONC_PROUNII INT,
 QT_CONC_PROUNIP INT,
 QT_MAT_RVREDEPUBLICA INT,
 QT_CONC_RVREDEPUBLICA INT,
 QT_SIT_TRANCADA INT,
 QT_SIT_DESVINCULADO INT,
 QT_SIT_TRANSFERIDO INT,
 QT_SIT_FALECIDO INT,
 QT_ING_PROCESCPUBLICA INT,
 QT_ING_PROCESCPRIVADA INT,
 QT_MAT_PROCESCPUBLICA INT,
 QT_MAT_PROCESCPRIVADA INT,
 QT_CONC_PROCESCPUBLICA INT,
 QT_CONC_PROCESCPRIVADA INT,
 CONSTRAINT pk_CE primary key(idTempo,idInstituicao,idCurso)
);
```