

METODOLOGIA APLICADA À CONSTRUÇÃO DO PAINEL ABED

Resumo

Este documento descreve a metodologia aplicada no desenvolvimento do Painel ABED, um *dashboard* de Business Intelligence que consolida e exibe estatísticas e métricas sobre a Educação a Distância no ensino superior brasileiro. A construção do painel baseou-se nos microdados do Censo da Educação Superior publicados anualmente pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, no período de 2010 a 2024, totalizando 3,3 GB de dados. O desenvolvimento envolveu um processo ETL (*extract, transform, load*) completo, incluindo: (1) extração de dados em formato CSV; (2) transformação e otimização utilizando Python (biblioteca Pandas) em ambiente Jupyter Notebook, com técnicas de redução de memória e imputação de dados geográficos; e (3) carregamento no Microsoft Power BI para visualização interativa. O *dashboard* final foi elaborado em quatro painéis principais — Matrículas, Ingressantes, Concluintes e Docentes —, organizados em uma estrutura de dados em esquema de estrela, o que permite análises multidimensionais por região, ano, modalidade e outras dimensões igualmente relevantes. Como resultado, o Painel ABED consolida 15 anos de dados históricos em quatro módulos analíticos que apoiam análises de tendências e planejamento estratégico da educação a distância no Brasil. Entre os principais desafios enfrentados destacam-se o gerenciamento de um grande volume de dados e a ausência de informações de geolocalização dos alunos, solucionada por meio de técnicas de imputação baseadas na localização das instituições de ensino. O painel resultante oferece uma solução analítica robusta para apoiar o planejamento estratégico e a formulação de políticas públicas no setor de Educação a Distância.

Palavras-chave: Educação a Distância; Business Intelligence; ETL (*extract, transform, load*); microdados educacionais; modelagem dimensional.

1 Introdução

A Educação a Distância (EaD) consolidou-se como uma das modalidades de maior crescimento no ensino superior brasileiro, democratizando o acesso à educação. Compreender essa expansão por meio de dados confiáveis é fundamental para gestores educacionais, pesquisadores, instituições de ensino e formuladores de políticas públicas.

Em consonância com sua missão institucional, a Associação Brasileira de Educação a Distância (ABED) desenvolveu o Painel ABED, uma ferramenta de Business Intelligence (BI) que consolida e exibe, de forma ágil e intuitiva, as principais estatísticas sobre a modalidade EaD no Brasil.

O painel foi concebido para democratizar o acesso aos microdados do Censo da Educação Superior publicados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), órgão oficial responsável pela coleta, organização e disseminação das estatísticas educacionais no Brasil, incluindo os dados declarados a cada ano pelas instituições de ensino superior. Esses dados, embora públicos, apresentam volume expressivo e formato técnico que dificultam sua análise. A ferramenta transforma essa complexidade em visualizações acessíveis, organizadas em quatro módulos principais: Matrículas, Ingressantes, Concluintes e Docentes.

O presente documento apresenta a metodologia aplicada no desenvolvimento do Painel ABED, detalhando o processo de extração, transformação e carga (ETL) dos dados, os desafios técnicos enfrentados e as soluções implementadas.

1.1 Base de dados e justificativa tecnológica

O Painel ABED utiliza como base de dados principal os microdados anuais do Censo da Educação Superior, publicados pelo INEP, abrangendo o período de 2010 a 2024. A consolidação dos dados das 15 edições do censo resultou em uma base de aproximadamente 3,3 GB, com informações detalhadas sobre cursos, matrículas, ingressantes, concluintes e docentes do ensino superior brasileiro.

A escolha da arquitetura tecnológica foi guiada pela necessidade de manipular um volume massivo de dados de forma eficiente, garantindo performance e interatividade no painel final. A solução adotada combinou três componentes tecnológicos complementares:

- Microsoft Power BI (visualização): plataforma selecionada para compilação e apresentação final em virtude de sua excelência em visualizações de BI e à capacidade de consulta ultrarrápida do seu motor *in-memory*, o VertiPaq (motor de armazenamento e processamento analítico fundamental por trás do Microsoft Power BI).
- Python e Pandas (transformação de dados): devido ao volume de dados, o tratamento e a transformação diretamente no Microsoft Power BI se mostraram improdutivos e limitados. Por isso, a fase de *transform* (T do ETL) foi delegada ao ambiente Jupyter Notebook, utilizando a biblioteca Pandas do ecossistema Python, garantindo eficiência na limpeza, consolidação e otimização da tipagem de dados antes do carregamento final.
- Sinergia para performance: a delegação da fase de tratamento para o Python assegurou que o *dataset* fosse carregado no Microsoft Power BI em sua forma mais leve e otimizada, o que maximizou a eficiência do motor VertiPaq, garantindo alta responsividade do painel aos filtros e às interações do usuário.

1.2 O Censo da Educação Superior e o desafio da integração

Realizado anualmente pelo INEP, o Censo da Educação Superior é o instrumento mais completo para a coleta de dados sobre cursos de graduação e sequenciais no Brasil. O Inep extrai informações do sistema E-MEC e as organiza em tabelas estruturadas (Tabelas SUP – Sistema de Informações da Educação Superior). No entanto, devido à aplicação da Lei Geral de Proteção de Dados (LGPD), Lei nº 13.709/2018, apenas um subconjunto dessas tabelas é disponibilizado ao público, especificamente as tabelas “cadastro_ies” (cadastro de instituições de ensino superior) e o “cadastro_cursos” (cadastro de cursos).

Além das restrições de acesso, há desafios relacionados à qualidade e completude dos dados. Uma característica fundamental do Censo é a autodeclaração das informações por um Recenseador Institucional (RI) de cada Instituição de Ensino Superior (IES). Embora o RI garanta a abrangência dos registros, essa natureza da coleta resulta na frequente presença de valores nulos para determinadas variáveis. Para um projeto que se propõe a entregar uma ferramenta de análise descritiva robusta e abrangente, a gestão desses valores ausentes, sem perda da amostra, constitui um dos problemas centrais a serem resolvidos.

As seções seguintes detalham como esses desafios foram enfrentados no desenvolvimento do Painel ABED, desde a extração dos dados até sua transformação e carga final no ambiente de visualização, sempre observando o rigor técnico a fim de garantir a integridade e a performance da ferramenta.

2 Metodologia do projeto

A metodologia do projeto seguiu o processo ETL (*extract, transform, load*) padrão em projetos de Business Intelligence (BI) e de análise de dados, garantindo a qualidade, a coerência e a performance do *dataset* utilizado para a criação do Painel

ABED. A subseção a seguir apresenta a abordagem e as ferramentas tecnológicas utilizadas no desenvolvimento.

O projeto foi conduzido com base uma abordagem orientada a dados, utilizando o ecossistema Python/Pandas para o pré-processamento e o Microsoft Power BI para a visualização e *dashboards*. As ferramentas e recursos tecnológicos empregados foram:

- plataforma de análise: distribuição Anaconda Python e ambiente interativo Jupyter Notebook;
- linguagem e bibliotecas: Python, com destaque para a biblioteca Pandas para a manipulação de dados;
- ferramenta de visualização: Microsoft Power BI;
- base de dados: microdados do Censo da Educação Superior (2010 a 2024).

A Figura 1 ilustra o fluxo do processo ETL, desde a extração dos microdados do INEP até a visualização final no Painel ABED, destacando o papel de cada ferramenta ao longo das etapas.

Figura 1 – Fluxo do processo ETL



Fonte: elaboração própria.

A partir disso, o desenvolvimento do Painel ABED foi organizado em fases sequenciais, cada uma com objetivos e entregas específicas, conforme detalhado a seguir.

3 Fases do desenvolvimento

O processo de preparação dos dados, etapa mais crítica para a coerência do painel, envolveu fases, conforme apresentado no Quadro 1.

Quadro 1 – Fases de preparação de dados

Fase	Ação realizada	Objetivo
1. Extração e consolidação	Reunião e importação dos 15 <i>datasets</i> anuais independentes (Censo de 2010 a 2024) em uma única estrutura de <i>dataframe</i> .	Criar uma base histórica unificada para análise de tendências ao longo do tempo.
2. Padronização e concatenação	Identificação e tratamento das diferenças estruturais (nome de colunas e tipo de dados) entre as tabelas anuais, seguidos pela concatenação final em uma única tabela.	Garantir a coerência estrutural e a correta integração de todas as informações em uma única base.
3. Otimização de memória	Conversão dos tipos de dados numéricos (predominantemente int64) para tipos de menor alocação de memória (Int32) e dos tipos de dados de texto (object) para o tipo de dado de menor volume (category).	Reduzir o tamanho da base de dados (de aproximadamente 4 GB) para otimizar o desempenho da análise exploratória e do carregamento no Microsoft Power BI.

Fonte: elaboração própria.

As subseções a seguir detalham cada uma das etapas do desenvolvimento do Painel ABED.

3.1 Tratamento e análise exploratória de dados

Após a estruturação e otimização, foi realizada a fase de limpeza e tratamento, conforme explicitado na sequência:

- Identificação de valores ausentes: mapeamento da presença de valores nulos (ausentes) no DataFrame geral, subsequentemente ao nível de coluna.
- Tratamento de variáveis geográficas: preenchimento dos valores nulos de variáveis geográficas (variáveis categóricas) da tabela cadastro_cursos com a tabela cadastro_ies.
- Tratamento das variáveis numéricas: substituição dos valores ausentes nas variáveis do tipo numérico (Int32) pelo valor (0), assumindo que a ausência de registro implicaria valor nulo para o contexto da estatística educacional.

3.2 Obtenção dos dados

Os dados utilizados na elaboração do Painel ABED originaram-se dos microdados do Censo da Educação Superior, disponibilizados no site do INEP. Foram coletados os microdados — menor unidade de detalhe do Censo da Educação Superior — referentes ao período entre 2010 e 2024. Os dados foram baixados em formato tabular .csv e, posteriormente, foram processados.

Apenas as tabelas de microdados de acesso público geral foram utilizadas. Optou-se, nesse momento, pela não inclusão de dados provenientes de fontes restritas, como o Serviço de Acesso a Dados Protegidos (SEDAP), uma vez que o acesso a essas informações exigia procedimentos burocráticos e termos de compromisso que restringiam a agilidade e a ampla disseminação do projeto.

3.3 Tratamento e transformação dos dados

A etapa de tratamento e transformação foi realizada integralmente no ambiente Jupyter Notebook, utilizando a biblioteca Pandas da linguagem Python. Toda a metodologia dessa fase foi orientada pelos princípios de *performance* e otimização para o ambiente do Microsoft Power BI, dado o grande volume da base consolidada (aproximadamente 3,3 GB).

3.3.1 Integração e consolidação de dados históricos

Essa fase focou na criação da base histórica única (Data Frame Principal), contendo a concatenação dos microdados da tabela “detalhe_ies” de 2010 a 2024 (totalizando 3.799.787 registros e 192 colunas). Optou-se por manter os nomes das colunas originais do INEP, de forma a priorizar a fidelidade ao dicionário de dados oficial — documento que descreve um conjunto de dados (dados sobre os dados) — e simplificar a etapa de modelagem no Microsoft Power BI.

3.3.2 Otimização estrutural e tipagem de dados

Para otimizar o desempenho da análise exploratória e do carregamento no Microsoft Power BI, foi realizada a conversão dos tipos de dados, visando reduzir o tamanho da base de dados (de aproximadamente 4 GB). O Quadro 2 apresenta as principais conversões de tipos de dados realizadas e suas respectivas justificativas técnicas.

Quadro 2 – Conversões de tipos de dados aplicadas para a otimização de memória

Tipo original comum em Python	Tipo alvo do projeto	Justificativa técnica orientada à performance
int64	Int32	Redução de memória alocada pela metade para as colunas de contagem (QT...). Como os valores máximos das contagens

		não excedem o limite do Int32, essa conversão (de 8 para 4 bytes) otimizou o carregamento e a compressão dos dados no Microsoft Power BI.
object	category	Aplicada a colunas com baixa cardinalidade, com código de IES (CO_IES), código de Curso (CO_CURSO) e classificações acadêmicas. O tipo <i>category</i> armazena os valores como inteiros mapeados, resultando em economia massiva de memória (armazenamento otimizado de <i>strings</i>).
int64	Int8	Usada em variáveis de código de classificação (como TP_MODALIDADE_ENSINO). Essa conversão garante a utilização mínima de memória (1 a 2 bytes), já que a faixa de valores é muito limitada.

Fonte: elaboração própria.

3.3.3 Tratamento de imputação geográfica por *mapping*

O preenchimento dos valores ausentes nas variáveis que localizam o curso no território nacional, como, por exemplo, o campo CO_MUNICIPIO, representou um desafio de alta complexidade, especialmente em relação a cursos oferecidos na modalidade EaD, visto que, nesses casos, essas dimensões não são calculadas:

- Desafio de memória: a tentativa de utilizar a função `merge()` do Pandas para cruzar os dados de cursos com os dados de cadastro de IES resultou erros de “memória insuficiente” (*out-of-memory*), pois a operação exigia a criação de uma cópia temporária que excedia a memória RAM disponível no ambiente de desenvolvimento.

- Solução por *mapping*: adotou-se o método de atribuição por dicionário (*mapping*) usando Series do Pandas, uma abordagem mais eficiente em termos de uso de memória. Como exemplificação da implantação do processo, tem-se o seguinte:
 - 1) Criação de uma estrutura de mapeamento (Series do Pandas) utilizando CO_IES como chave e CO_MUNICIPIO_IES como valor, extraída da tabela de cadastro de IES.
 - 2) Para cada registro do DataFrame principal (cursos) com valor nulo em CO_MUNICIPIO, a chave CO_IES foi utilizada para consultar o dicionário criado.
 - 3) Os valores retornados foram diretamente atribuídos de volta ao DataFrame principal. Esse método garantiu o preenchimento das variáveis geográficas [NO_REGIAO_IES, CO_REGIAO_IES, NO_UF_IES, SG_UF_IES, CO_UF_IES, NO_MUNICIPIO_IES, CO_MUNICIPIO_IES] sem a necessidade operações custosas de *merge* ou *join* em bases massivas.

3.3.4 Tratamento final de valores ausentes em variáveis numéricas

O tratamento final dos valores nulos (NaN) foi aplicado às variáveis numéricas de contagem (QT...), utilizando a estratégia de imputação por zero:

- Regra de negócio: no contexto do Censo da Educação Superior, a ausência de registro em uma coluna quantitativa (como QT_MAT, QT_ING, QT_CONC) indica, por regra, que o evento não ocorreu ou que o valor é nulo para o registro específico (curso/ano).
- Justificativa da imputação: o preenchimento dos NaNs com zero (0) (.fillna(0)) foi a decisão mais apropriada para manter a integridade estatística da base.

- Impacto no BI: essa imputação previne erros na agregação de dados no Microsoft Power BI (visto que o *software* trata nulos de maneira diferente em médias e somas) e garante que indicadores como totais de matrículas ou vagas sejam calculados corretamente, sem distorções causadas por valores omissos.

3.4 Modelagem de dados (*load* e relacionamentos)

A fase de modelagem de dados foi executada após o carregamento (*load*) das bases tratadas e otimizadas para o Microsoft Power BI. O modelo foi estruturado em um esquema estrela para maximizar a *performance* analítica e a usabilidade do painel.

3.4.1 Carregamento e definição de tabelas

Foram carregadas nove tabelas no ambiente do Microsoft Power BI, categorizadas da seguinte forma:

Tabelas Fato (2):

- FATO_CURSOS: tabela principal contendo as métricas quantitativas (variáveis QT...) e a consolidação dos microdados de 2010 a 2024.
- FATO_IES: tabela contendo dados cadastrais das IES, servindo como uma tabela fato secundária.

Tabelas Dimensão (7):

- DIM_ORGANIZACAO_ACADEMICA
- DIM_GRAU_ACADEMICO
- DIM_NIVEL_ACADEMICO
- DIM_REDE_CURSOS
- DIM_DESCRICAO_CINE_ESPECIFICA
- DIM_CATEGORIA_ADMINISTRATIVA
- DIM_MODALIDADE_ENSINO

3.4.2 Modelagem dimensional e otimização para filtros

O modelo seguiu o princípio de esquema estrela, em que as tabelas fato se conectam diretamente às tabelas dimensão, garantido consultas rápidas e fluxo de filtragem intuitiva. Para maximizar a eficiência desse modelo, foram adotadas duas estratégias:

- Estrutura de dimensão: cada tabela de dimensão foi construída para ter apenas duas colunas principais:
 - 1) Valor chave (CO... ou TP...) – utilizado para o relacionamento com a tabela fato.
 - 2) Descrição (NO... ou texto explícito) – utilizada como rótulo amigável nos filtros e visualizações do painel.
- Justificativa de performance: essa separação é uma técnica de Data Warehousing que potencializa a *performance* do Microsoft Power BI. Ao usar a coluna de código interior (tipos Int8, Int16, Int32) no relacionamento, as consultas e filtros ocorrem de forma muito mais rápida no motor VertiPaq do que se a tabela fato contivesse a *string* completa da descrição.

3.4.3 Relacionamento e chaves de ligação

Os relacionamentos foram estabelecidos, primeiramente, a partir da tabela FATO_CURSO para as tabelas de dimensão:

- Cardinalidade: todos os relacionamentos foram configurados como Um para Muitos (1:N), com o lado “Um” na tabela de Dimensão (chave única) e o lado “Muitos” na tabela Fato (chaves repetidas).
- Exemplos de chaves:
 - 1) FATO_CURSOS [TP_GRAU_ACADEMICO] ← DIM_GRAU_ACADEMICO [TP_GRAU_ACADEMICO]

2) FATO_CURSOS [TP_MODALIDADE_ENSINO] ←
DIM_MODALIDADE_ENSINO [TP_MODALIDADE_ENSINO]

O resultado é um modelo de dados estável, em que as otimizações de tipagem de dados e arquitetura dimensional garantem a responsividade da navegação e da aplicação de filtros no *dashboard* final.

3.5 Visualização e análise

A etapa de visualização e análise materializou o modelo de dados em um *dashboard* interativo, denominado Painel ABED. A construção seguiu uma abordagem de BI focada em estatística descritiva e análise de série temporal, distribuídas em quatro painéis temáticos principais: Matrículas, Ingressantes, Concluintes e Docentes.

3.5.1 Estrutura de filtragem e navegação

O painel foi projetado para ser intuitivo e permitir uma filtragem completa dos dados em todas as visualizações. Para alcançar esse objetivo, foram implementados três componentes estruturais:

- Filtros globais: um conjunto de filtros globais foi posicionado de forma coesa na parte superior, em uma área dedicada, garantindo o controle total sobre as segmentações. Os filtros operam nas colunas de descrição das Tabelas de Dimensão, mantendo o relacionamento com a Tabela Fato via código de tipo de dados otimizado.
- Dimensão de filtragem: ANOS (NU_ANO_CENSO), ESTADOS (NO_UF), CATEGORIA ADMINISTRATIVA, ORGANIZAÇÃO ACADÊMICA, NÍVEL ACADÊMICO, TP_GRAU_ACADÊMICO, ÁREA DO CURSO (NO_CINE_ESPECIFICA), REDE ENSINO e MODALIDADE.
- Visão detalhada por painel: três dos painéis (Matrículas, Ingressantes e Concluintes), alimentados pela tabela FATO_CURSOS, garantem

Rua Vergueiro, 875 12º andar- cjs 123/124, Bairro Liberdade, CEP 01504-001, São Paulo - SP

Telefone: +55 11 3275.3561

abed@abed.org.br www.abed.org.br

consistência na experiência do usuário. O painel Docente, por sua vez, é alimentado pela tabela FATO_IES.

3.5.2 Análise estatística no painel Matrículas

O painel Matrículas visa a detalhar a evolução e o perfil do corpo discente, conforme consta no Quadro 3.

Quadro 3 – Painel Matrículas detalhando a evolução e o perfil do corpo discente

Elemento visual	Variáveis (eixo/valores)	Finalidade analítica
Tabela de cursos e cartão de apoio	FATO_CURSOS e NO_CURSO (Primeiro)	Permite a seleção do curso para detalhamento e confirma visualmente a seleção ativa.
Gráfico de colunas clusterizadas	Eixo x: NU_ANO_CENSO (2010 a 2024); Eixo Y: QT_MAT	Possibilita a análise de série temporal: demonstra a evolução histórica e a tendência anual das matrículas.
Gráfico de barras (Grau Acadêmico)	Eixo Y: descrição da Dim_TP_GRAU_ACADEMICO; Eixo X: QT_MAT	Exibe a distribuição das matrículas por tipo de formação (Bacharelado, Licenciatura, Tecnológico).
Cartão de total	Soma de QT_MAT	Fornece estatísticas descritivas primárias (Total Geral de Matrículas).
Gráfico de rosca (Gênero)	Valores: QT_MAT_FEM e QT_MAT_MASC	Apresenta a distribuição do corpo discente por gênero.
Mapa coroplético	Localização: NO_UF; Legenda: NO_REGIAO	Permite análise específica da densidade das matrículas pelas IES e o <i>drill-down</i> por unidade federativa.
Gráfico de barras empilhadas (Regiões)	Eixo Y: NO_REGIAO ; Eixo X: QT_MAT	Compara a contribuição de cada região brasileira para o total de matrículas.
Gráfico de funil (Faixa Etária)	Valores: QT_MAT_0_17 até QT_MAT_60 MAIS	Possibilita a análise de perfil, com demonstração da distribuição das matrículas por grupo etário específico, revelando o perfil do aluno.

Fonte: elaboração própria.

3.5.3 Painéis Ingressantes e Concluintes

Os painéis Ingressantes e Concluintes complementam a análise de matrículas ao fornecer visibilidade sobre o fluxo de entrada e saída de estudantes na modalidade EaD. Esses painéis seguem a mesma estrutura e tipos de visualização do painel Matrículas, diferenciando-se apenas na variável de métrica (QT_) utilizada:

- Ingressantes: utiliza variáveis com o prefixo QT_ING (por exemplo: QT_ING, QT_ING_FEM, QT_ING_0_17).
- Concluintes: utiliza variáveis com o prefixo QT_CONC (por exemplo: QT_CONC, QT_CONC_FEM, QT_CONC_0,17).

O painel Ingressantes detalha a evolução e o perfil de estudantes que ingressam em cursos na modalidade EaD. As visualizações implementadas discentes são apresentadas no Quadro 4.

Quadro 4 – Elementos visuais no painel Ingressantes

Elemento visual	Variáveis (eixo/valores)	Finalidade analítica
Tabela de cursos e cartão de apoio	FATO_CURSOS e NO_CURSO (Primeiro)	Permite a seleção do curso para detalhamento e confirma visualmente a seleção ativa.
Gráfico de colunas clusterizadas	Eixo x: NU_ANO_CENSO (2010 a 2024); Eixo Y: QT_ING	Permite a análise de série temporal; demonstra a evolução histórica e a tendência anual dos ingressos.
Gráfico de barras (Grau Acadêmico)	Eixo Y: descrição da Dim_TP_GRAU_ACADEMICO; Eixo X: QT_ING	Exibe a distribuição dos ingressos por tipo de formação (Bacharelado, Licenciatura, Tecnológico)
Cartão de total	Soma de QT_ING	Fornecer estatística descritiva primária (Total Geral de Ingressos).

Gráfico de rosca (Gênero)	Valores: QT_ING_FEM e QT_ING_MASC	Apresenta a distribuição dos ingressantes por gênero.
Mapa coroplético	Localização: NO_UF; Legenda: NO_REGIAO	Permite a análise específica, com visualização da densidade dos ingressos por IES, e o <i>drill-down</i> por unidade federativa.
Gráfico de barras empilhadas (Regiões)	Eixo Y: NO_REGIAO; Eixo X: QT_ING	Compara a contribuição de cada região brasileira para o total de ingressantes.
Gráfico de funil (Faixa Etária)	Valores: QT_ING_0_17 até QT_ING_60 MAIS	Permite a análise de perfil; demonstra a distribuição dos ingressantes por grupo etário específico, revelando o perfil do aluno.

Fonte: elaboração própria.

O painel Concluintes detalha a evolução e o perfil dos estudantes que concluem sua formação na modalidade EaD. As visualizações implementadas são apresentadas no Quadro 5.

Quadro 5 – Elementos visuais no painel Concluintes

Elemento visual	Variáveis (eixo/valores)	Finalidade analítica
Tabela de cursos e cartão de apoio	FATO_CURSOS e NO_CURSO (Primeiro)	Permite a seleção do curso para detalhamento e confirma visualmente a seleção ativa.
Gráfico de colunas clusterizadas	Eixo x: NU_ANO_CENSO (2010 a 2024); Eixo Y: QT_CONC	Permite a análise de série temporal; demonstra a evolução histórica e a tendência anual dos concluintes.
Gráfico de barras (Grau Acadêmico)	Eixo Y: descrição da Dim_TP_GRAU_ACADEMICO; Eixo X: QT_CONC	Exibe a distribuição dos concluintes por tipo de formação (Bacharelado, Licenciatura, Tecnológico).

Cartão de total	Soma de QT_CONC	Fornecer estatística descritiva primária (Total Geral de Concluintes).
Gráfico de rosca (Gênero)	Valores: QT_CONC_FEM e QT_CONC_MASC	Apresenta a distribuição dos concluintes por gênero.
Mapa coroplético	Localização: NO_UF; Legenda: NO_REGIAO	Permite a análise específica, com visualização da densidade dos concluintes por IES, e o <i>drill-down</i> por unidade federativa.
Gráfico de barras empilhadas (Regiões)	Eixo Y: NO_REGIAO; Eixo X: QT_CONC	Compara a contribuição de cada região brasileira para o total de concluintes.
Gráfico de funil (Faixa Etária)	Valores: QT_CONC_0_17 até QT_CONC_60 MAIS	Permite a análise de perfil; demonstra a distribuição dos concluintes por grupo etário específico, revelando o perfil do aluno.

Fonte: elaboração própria.

3.5.4 Painel Docentes

O painel Docentes, alimentado pela FATO_IES, segue o mesmo padrão visual dos painéis anteriores, mas com foco em métricas e filtros institucionais, como consta no Quadro 6.

Quadro 6 – Elementos visuais no painel Docentes

Elemento visual	Variáveis (eixo/valores)	Finalidade analítica
Tabela de cursos e cartão de apoio	FATO_IES e NO_CURSO (Primeiro)	Permite a seleção do curso para detalhamento e confirma visualmente a seleção ativa.
Gráfico de linhas	Eixo x: NU_ANO_CENSO (2010 a 2024); Eixo Y: QT_DOC_TOTAL	Análise de série temporal: demonstra a evolução

		histórica e a tendência anual para o total de docentes.
Gráfico de rosca (Titulação)	Valores: QT_DOC_EX_DOUT, QT_DOC_EX_MEST e QT_DOC_EX_ESP	Exibe a distribuição de docentes por titulação.
Cartão de total	Soma de QT_DOC_TOTAL	Fornecer estatística descritiva primária (Total Geral de docentes).
Gráfico de rosca (Gênero)	Valores: QT_DOC_EX_FEMI e QT_DOC_EX_MASC	Apresenta a distribuição do corpo docente por gênero.
Gráfico de rosca (Status)	Valores: QT_DOC_EX_INT_SEM_DE, QT_DOC_EX_PARC e QT_DOC_EX_HOR	Apresenta a distribuição de docentes tempo dedicado à IES.
Mapa coroplético	Localização: NO_UF; Legenda: NO_REGIAO	Análise específica: visualiza a densidade da distribuição dos docentes pelas IES e permite o <i>drill-down</i> por unidade federativa.
Gráfico de barras empilhadas (Regiões)	Eixo Y: NO_REGIAO ; Eixo X: QT_DOC_TOTAL	Compara a contribuição de cada região brasileira para o total de docentes.
Gráfico de funil (Faixa etária)	Valores: QT_DOC_EX_0_17 até QT_DOC_EX_60 MAIS	Análise de perfil: demonstra a distribuição de docentes por grupo etário específico.

Fonte: elaboração própria.

3.5.5 Projeção e crescimento

O Painel ABED foi desenvolvido com escalabilidade em mente, permitindo futuras expansões em duas frentes: 1) utilização de variáveis ainda não exploradas nas tabelas FATO_CURSOS e FATO_IES, e 2) integração com bases externas, como o Exame Nacional de Desempenho dos Estudantes (Enade), para enriquecer as estatísticas descritivas e fornecer *insights* mais profundos sobre a qualidade e o desempenho do ensino superior.

4. Discussão e contribuições

Esta seção apresenta a discussão das principais barreiras metodológicas encontradas durante a construção do Painel ABED e detalha as contribuições e *insights* gerados pela solução analítica desenvolvida.

4.1 Superando o desafio da geolocalização na Educação a Distância

A principal limitação identificada na aquisição e tratamento dos microdados do Censo da Educação Superior, particularmente para a modalidade EaD, reside na indisponibilidade da geolocalização do aluno ou do polo de apoio presencial, como detalhado a seguir.

Natureza do problema

A característica intrínseca da EaD é a flexibilidade, permitindo que o aluno não se vincule fisicamente à sede da IES. Embora ele esteja frequentemente ligado a um Polo de Apoio Presencial, essa informação crucial — que permitiria uma análise geográfica precisa das métricas de QT_MAT, QT_ING e QT_CONC — não é disponibilizada nas tabelas de acesso público do INEP. Consequentemente, as variáveis geográficas nas linhas de EaD são mantidas como valores ausentes (nulos), impossibilitando a análise espacial direta dos dados de estudantes na modalidade a distância.

Decisão metodológica e solução implementada

Para contornar essa lacuna e possibilitar a segmentação geográfica dos indicadores de EaD (vitais para análise de mercado e políticas públicas), foi necessário adotar uma heurística de imputação baseada na localização institucional.

As métricas de matrículas, ingressantes, concluintes e docentes da modalidade EaD foram imputadas aos dados geográficos declarados pela própria IES (presentes na tabela de cadastro “detalhe_ies”). Essa atribuição foi realizada tecnicamente no Python, utilizando a solução de *mapping* de baixo consumo de memória (descrita na seção 3.3.3), cruzando a chave CO_IES com os campos geográficos institucionais: [NO_REGIAO_IES, CO_REGIAO_IES, NO_UF_IES, SG_UF_IES, CO_UF_IES, NO_MUNICIPIO_IES, CO_MUNICIPIO_IES].

Consequência e contribuição

Embora essa solução vincule as métricas de EaD à sede administrativa da IES – e não ao polo do aluno – ela permite a primeira camada de análise espacial no Painel ABED, oferecendo *insights* sobre distribuição as IES que mais contribuem para as métricas estudadas em cada região do país.

4.2 Contribuições técnicas e otimização do projeto

Além de superar o desafio metodológico da geolocalização, o desenvolvimento do Painel ABED resultou em contribuições técnicas significativas nas áreas de otimização de performance, arquitetura dimensional escalável e eficiência no tratamento de grandes volumes de dados.

As principais contribuições são detalhadas a seguir:

- 1) Modelagem otimizada de alto desempenho: a decisão de consolidar 15 anos de dados em uma única tabela fato no Python, seguida pela otimização fina dos tipos de dados (uso de Int32, Int16 e category), resultou em um modelo de dados mais leve e rápido no VertiPaq do Microsoft Power BI. Essa otimização provou ser superior na resolução das mesmas transformações na etapa de merge com o volume de dados histórico, em que houve falha por esgotamento de memória.
- 2) Reprodutibilidade e extensibilidade: a utilização de ferramentas *open source* (Python/Pandas) e de bases de dados públicas disponibilizadas pelo INEP garantem a rastreabilidade e a reprodutibilidade do projeto. A arquitetura em esquema estrela implementada no Microsoft Power BI facilita a futura integração com outras fontes de dados, como o ENADE, sem a necessidade de reestruturar a base principal, assegurando a escalabilidade da solução.

3) Eficiência no tratamento de grandes volumes de dados: as técnicas de otimização de memória implementadas no Python, incluindo o uso de mapeamentos eficientes e a conversão estratégica de tipos de dados, permitiram o processamento de milhões de registros de forma eficiente, estabelecendo um processo de ETL robusto e replicável para projetos similares de análise de dados educacionais em larga escala.

Em síntese, o Painel ABED demonstra como a combinação de técnicas avançadas de tratamento de dados, modelagem dimensional otimizada e decisões arquiteturais estratégicas pode transformar microdados brutos e complexos em inteligência analítica acessível. Ao reconhecer e contornar a limitação da geolocalização e ao implementar otimizações que superam restrições técnicas de processamento, o projeto entrega uma ferramenta analítica robusta, escalável e pronta para apoiar a tomada de decisão e a análise de tendências no mercado de educação a distância.

5 Conclusão

O projeto culminou na criação e implementação bem-sucedida do Painel ABED, uma solução analítica interativa construída sobre uma base de dados robusta e historicamente consolidada.

O principal objetivo de transformar 15 anos de microdados brutos do Censo da Educação Superior em um recurso de Business Intelligence de alto desempenho foi plenamente alcançado. Isso foi possível através de uma metodologia rigorosa de ETL, destacando-se a utilização estratégica da linguagem Python e do pacote Pandas

para a consolidação e otimização de uma base de dados de grande volume (cerca de 3,3 GB).

A otimização de tipo de dados e a implementação de técnica de *mapping* para contornar falhas de memória no merge massivo foram cruciais para a garantir a eficiência no carregamento e a velocidade de consulta dentro do motor VertiPaq do Microsoft Power BI. A modelagem em esquema estrela resultou em um ambiente analítico intuitivo e altamente responsivo.

Embora o desafio de geolocalização precisa das métricas de EaD tenha se apresentado como uma limitação inerente à fonte – devido à ausência de dados do polo de apoio nos microdados do INEP –, a imputação estratégica à localização da IES permitiu que a análise geográfica fosse, ainda assim, realizada, fornecendo *insights* valiosos sobre a distribuição institucional.

Assim, o Painel ABED representa uma contribuição significativa para o setor educacional brasileiro, pois fornece uma visão clara, histórica e segmentada sobre matrículas, ingressantes, concluintes e docentes da modalidade EaD. A ferramenta está preparada para apoiar análises de tendência, planejamento estratégico e desenvolvimento de políticas públicas, consolidando-se como uma plataforma de inteligência de dados pronta para futuras expansões e integrações.

Referências

ASSOCIAÇÃO BRASILEIRA DE EDUCAÇÃO A DISTÂNCIA (ABED). **Quem somos**. Disponível em: <https://www.abed.org.br/site/institucional/quem-somos/>. Acesso em: 15 fev. 2025.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). **Censo da Educação Superior: Resultados**. Brasília, DF: INEP. Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-da-educacao-superior/resultados>. Acesso em: 5 out. 2025. [Referência principal para os dados de 2010 a 2024]

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Dispõe sobre a proteção de dados pessoais e altera a Lei n.º 12.965, de 23 de abril de 2014 (Marco Civil da Internet). 15 ago. 2018. Disponível em: <https://www2.camara.leg.br/legin/fed/lei/2018/lei-13709-14-agosto-2018-787077-publicacaooriginal-156212-pl.html>. Acesso em: 2 out. 2025.

MICROSOFT. **Documentação do Power BI**. Microsoft Learn. Disponível em: <https://learn.microsoft.com/pt-br/power-bi/>. Acesso em: 2 out. 2025.

MICROSOFT. **VertPaq**: tecnologias e conceitos. Disponível em: <https://learn.microsoft.com/pt-br/power-bi/guidance/powerbi-implementation-planning-auditing-monitoring-data-level-auditing>. Acesso em: 2 out. 2025.

THE PYTHON SOFTWARE FOUNDATION. **Jupyter Notebook**. Disponível em: <https://jupyter.org/>. Acesso em: 2 out. 2025.

THE PYTHON SOFTWARE FOUNDATION. **Pandas**: Python Data Analysis Library. Disponível em: <https://pandas.pydata.org/docs/>. Acesso em: 2 out. 2025.