Utilização do Apache Superset para Visualização Escalável de Dados Educacionais Públicos: Um Estudo de Caso com o Censo Escola*

João Pedro V. Ramalho, João A. Silveira
Thales Gabriel C. de Lima, Mateus S. Herbele
Josiney de Souza, Guilherme A. Derenievicz, Letícia M. Peres, Simone Dominico

[1]Centro de Computação Científica e Software Livre (C3SL)
Universidade Federal do Paraná (UFPR)
Caixa Postal 19.011 – 81.530-090 – Curitiba – PR – Brasil

{jpvr22, jas21, tgcl21, msh22, jsouza, guilherme, lmperes, simone}@inf.ufpr.br

Abstract. *Public databases, such as the School Census provide data that facilitates the planning of educational policies in Brazil. Although the data are provided in an accessible format and tabular structure, the lack of adequate tools limits their exploration. This work evaluates the use of Apache Superset as an open-source solution for interactive visualization and creation of indicators in education. The data were loaded into a relational database and underwent a data ingestion and transformation flow. Superset was used to create educational indicators, with the aim of demonstrating how the tool can facilitate the analysis of temporal, spatial, and demographic patterns, supporting decision-making and the production of scientific knowledge.*

Resumo. *As bases públicas, como o Censo Escolar, dispõem de dados que facilitam o planejamento de políticas educacionais no Brasil. Embora os dados sejam fornecidos em formato acessível e estrutura tabular, a falta de ferramentas adequadas limita sua exploração. Este artigo avalia o uso do Apache Superset como solução de código aberto para visualização interativa e criação de indicadores educacionais. Os dados foram adicionados em um banco relacional e passaram por um fluxo de ingestão e transformação. O Superset foi utilizado para a criação de indicadores de matrículas e docentes, para demonstrar como a ferramenta facilita a análise de padrões temporais, espaciais e demográficos, apoiando a tomada de decisão e a produção de conhecimento científico.*

1. Introdução

Anualmente, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP)[1] disponibiliza os dados de Censo Escolar da Educação. As informações são dispostas em formato de arquivos CSV, desde 1995, com os registros escolares de todos os municípios do Brasil. Para analisá-los, é necessário obter os arquivos que são separados por anos e combiná-los para sua exploração. A alta dimensionalidade dos dados e a heterogeneidade entre edições dificultam a análise exploratória, pois exigem a integração de

*Este trabalho recebeu financiamento do MEC/FNDE no contexto do projeto Laboratório de Dados Educacionais (LDE) (TED SIMEC No.: 11.437/2022)

[1]https://www.gov.br/inep/pt-br

múltiplos arquivos CSV com estruturas distintas. Cada arquivo possui diversas variáveis (ex. matrículas, infraestrutura, docentes), os dados têm cerca de 370 colunas, com número de linhas variável conforme a quantidade de escolas, atualmente em 178.476.

A integração desses dados, que demanda a combinação de vários arquivos CSV por meio de *scripts* manuais, pode gerar erros, consumir tempo e dificultar a atualização, devido a alterações de nomenclatura das colunas e no formato dos dados. Segundo [Van den Burg et al. 2019], embora os arquivos CSV sejam simples, sua formatação é inconsistente na prática, o que exige reparos manuais antes da carga para a análise, desperdiçando tempo que poderia ser dedicado a outras tarefas. Para enfrentar esses problemas, arquiteturas modernas se beneficiam de fluxos de dados responsáveis pela ingestão, transformação e armazenamento das informações [Mbata et al. 2024].

Outro ponto importante é a visualização e exploração desses dados. Mesmo com o aumento na disponibilidade e popularização de ferramentas de visualização, muitas delas são proprietárias e de alto custo, o que limita seu uso em órgãos públicos, em análises com objetivos acadêmicos e na visualização do cenário educacional ao longo dos anos.

Este artigo propõe um fluxo de dados para análise exploratória dos microdados do Censo Escolar, utilizando o banco de dados analítico ClickHouse [Schulze et al. 2024] para armazenamento e consultas, em conjunto da plataforma de código aberto Apache Superset, para visualização interativa [Superset 2024]. Além disso, é apresentada uma arquitetura para ingestão e normalização dos dados, consolidando as informações do INEP em um modelo relacional, com tratamento de inconsistências históricas. A modelagem relacional dos dados se demonstra uma alternativa promissora a ser explorada, pois facilita a criação de indicadores educacionais e consultas analíticas usando a linguagem de consulta SQL (*Structured Query Language*) para cruzamentos de dados complexos. Por fim, são desenvolvidas visualizações interativas (*dashboards*) com o Superset, voltadas à exploração de padrões temporais, demográficos e geográficos.

Com esse estudo, realizado com dados de 2007 a 2024, é possível demonstrar a viabilidade de identificação de tendências nos dados, contribuindo para a formulação de politica públicas e transparência das informações para a sociedade civil, por meio de visualizações que facilitam o entendimento por usuários. A substituição do trabalho manual por uma solução integrada e de código aberto contribui para reduzir a complexidade de análise dos dados educacionais. Isso oferece aos gestores públicos, pesquisadores e à população uma ferramenta escalável e adaptável às futuras mudanças do Censo Escolar.

2. Desafios Técnicos na análise de dados públicos

Os dados do Censo Escolar da Educação Básica representam uma das maiores bases públicas de dados educacionais. No entanto, a utilização efetiva desses dados para análise exploratória e tomada de decisão enfrenta diversos desafios técnicos. A base do Censo Escolar contém um grande volume de registros por ano, com muitos atributos disponíveis em arquivos CSV, que geralmente utilizam o tipo padrão de texto para os campos.

Embora o uso de arquivos CSV seja simples e apresente facilidade inicial, esse formato possui limitações técnicas, como a ausência de tipos de dados, falta de suporte à concorrência, baixo desempenho em consultas e dificuldade para manipulação em larga escala. Por outro lado, um modelo relacional com indexação, transações e esquemas bem

definidos oferece maior robustez e desempenho, atendendo a necessidade de escalabilidade e requisitos técnicos para armazenamento e exploração dos dados.

Um dos principais desafios na análise de dados educacionais, como os do Censo Escolar, é a ingestão de arquivos CSV com estrutura e formatação heterogêneas. Embora o formato CSV seja amplamente utilizado para disponibilização de dados públicos, os atributos podem variar entre os anos de coleta, incluindo mudanças nos nomes de colunas, variações nos domínios de valores e adição ou remoção de atributos [Yamanaka et al. 2024]. O artigo de [Ehrenfried et al. 2019] propõe o HOTMapper como uma solução específica para esse cenário, permitindo a criação de um modelo relacional unificado a partir de múltiplos arquivos CSV anuais com esquemas distintos. O HOTMapper adota uma abordagem baseada em mapeamentos explícitos entre os diferentes esquemas de origem e um esquema-alvo consolidado, além de suportar transformações de dados durante o processo de carga.

No artigo de [Karabtsev et al. 2023] os autores descrevem a implementação de uma solução de *Business Intelligence* (BI) para análise das atividades docentes na Universidade Estatal de Kemerovo. Os autores descrevem os desafios da integração de dados dispersos, incluindo arquivos CSV, XML, APIs e bancos relacionais. Assim como no cenário deste artigo, foi adotada a construção de um fluxo de dados que armazena os dados em um modelo analítico, permitindo sua exploração por meio de *dashboards* interativos. A abordagem destaca a importância de transformar fontes de dados não estruturadas ou semi-estruturadas em esquemas relacionais, a fim de facilitar consultas, cálculos analíticos e visualização em ferramentas de BI.

3. Arquitetura da Solução

Trabalhar diretamente com os arquivos CSV fornecidos pelo Censo Escolar apresenta limitações práticas significativas, como a ausência de suporte a tipos de dados, a falta de integridade referencial e o desempenho limitado em consultas exploratórias. Essas limitações, comuns em dados públicos distribuídos nesse formato, tornam-se ainda mais evidentes diante da necessidade de combinar múltiplos anos de coleta. A adoção de um modelo relacional analítico no ClickHouse, estruturado a partir de um fluxo de ingestão e transformação com o Apache NiFi [Apache Software Foundation 2024], proporciona um ambiente capaz de superar essas dificuldades. Com os dados centralizados e normalizados, é possível realizar consultas SQL eficientes, aplicar agregações complexas e integrar os resultados de forma dinâmica aos *dashboards* desenvolvidos no Apache Superset. Essa arquitetura reduz o esforço manual de manipulação, facilitando a construção de indicadores educacionais consistentes ao longo do tempo.

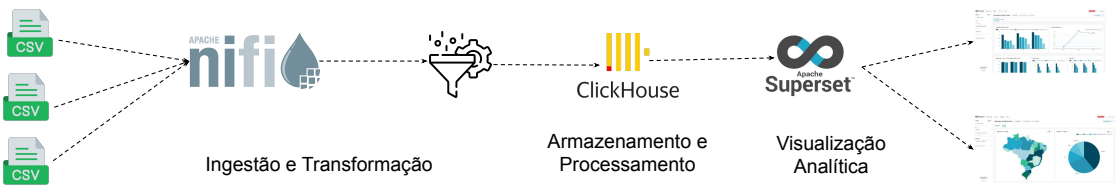

Figura 1. Arquitetura do fluxo de dados para os dados do Censo escolar

A Figura 1 apresenta a arquitetura proposta para ingestão, transformação e visualização dos dados do Censo Escolar. A principal fonte são os arquivos CSV disponibilizados pelo INEP, processados pelo Apache NiFi, que automatiza a ingestão. O NiFi realiza a organização dos arquivos, o controle de qualidade inicial (verificação de codificação e detecção de nomenclatura) e a orquestração do fluxo de dados.

Na etapa de transformação, o NiFi trata as inconsistências e prepara os dados para o modelo analítico no ClickHouse. Durante esse processo, o Nifi executa *scripts* de padronização de nomenclatura de colunas, necessários devido às variações de esquema entre diferentes edições do Censo Escolar. Por exemplo, atributos que mudaram de nome ao longo dos anos são mapeados para colunas consolidadas no esquema relacional. O NiFi também realiza conversões de tipo de dados, como a transformação de campos originalmente representados como texto para tipos numéricos, booleanos ou categóricos adequados ao modelo analítico.

Outra etapa é a agregação de colunas e o cálculos de dados para indicadores, onde os dados brutos de diferentes anos são consolidados para permitir comparações temporais. Por exemplo, variáveis relacionadas a categorias de matrícula, faixa etária e modalidade de ensino são reorganizadas e agregadas para gerar indicadores como o total de matrículas por estado, por modalidade ou por grupo etário. O fluxo de transformação inclui a limpeza de dados, com a normalização de valores ausentes, tratamento de *outliers* e verificação de integridade referencial antes da carga no ClickHouse.

Após as transformações os dados são armazenados no ClickHouse, um sistema de gerenciamento de banco de dados analítico orientado a colunas. Os dados são armazenados em um modelo relacional com tabelas otimizadas para consultas. Com os dados já armazenados, o Superset, conectado ao ClickHouse, é utilizado para a criação de *dashboards*, com diferentes tipos de gráficos, séries históricas e indicadores educacionais.

A etapa de visualização interativa no Superset aborda a dificuldade de explorar padrões complexos. Para isso, são utilizados filtros globais (para seleção simultânea por ano, região e tipo de escola) e visualizações interligadas, onde a seleção em um gráfico atualiza automaticamente os demais. Também é possível construir visualizações de mapas com a distribuição geográfica dos indicadores e visualizações temporais para identificar tendências. O fluxo completo, desde os CSVs brutos até os *dashboards*, demonstra como técnicas modernas de engenharia de dados podem transformar dados dispersos entre diferentes arquivos CSV em visualizações úteis para a gestão e acompanhamento educacional.

4. Resultados e Visualizações Produzidas

Após a consolidação dos dados no ClickHouse e a modelagem relacional analítica, foram criados *dashboards* interativos no Apache Superset para a exploração de indicadores educacionais com base nos dados do Censo Escolar. As visualizações consideram diferentes dimensões da análise educacional: temporal, espacial e demográfica. O tempo médio de visualização para cada análise no Superset foi em média de $1,33$ segundo. Na construção de *dashboards* no Superset, foram empregados diversos tipos de visualização. A escolha foi feita baseada em critérios como a legibilidade das informações e adequação da visualização ao tipo e à estrutura dos dados, de forma a garantir coerência e clareza na apresentação. As visualizações analíticas dos dados possibilitam uma comunicação acessível das informações, facilitando a identificação de padrões e desigualdades.

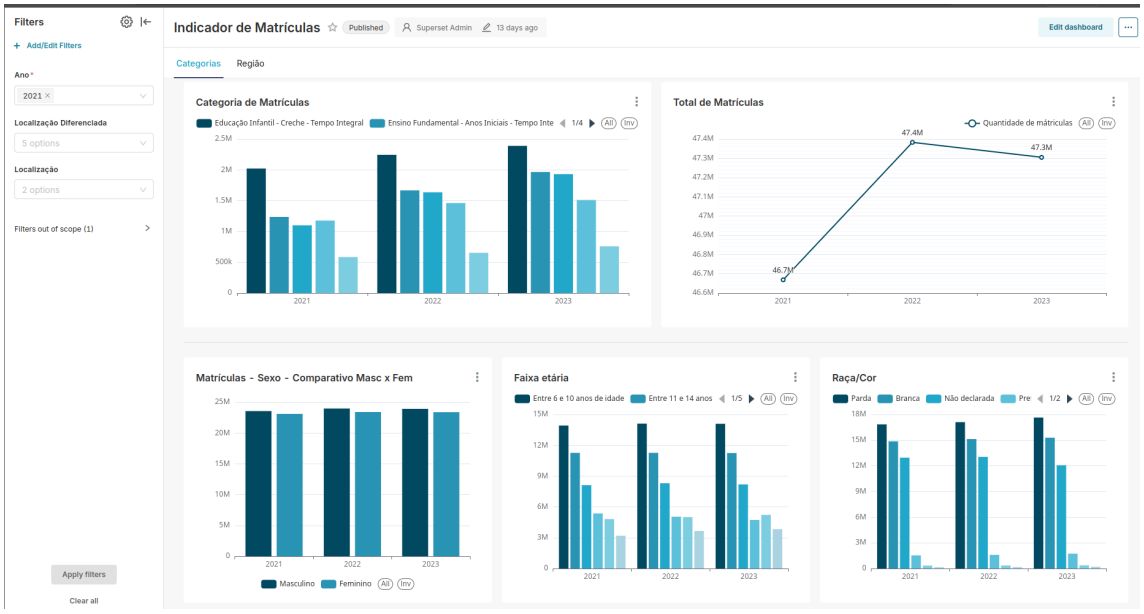

Figura 2. Exemplo de *Dashboard* para o Indicador de matrículas usando gráfico de linhas e barras.

As Figuras 2 e 3 mostra dois *dashboards* de visualização construídos no Superset, a seleção considerou os indicadores mais acessados pelos usuários da Plataforma de Dados Educacionais[2]. No caso do indicador de matrículas, o *dashboard* mostra informações detalhadas sobre matrículas em escolas brasileiras, como o total geral de matrículas, a distribuição por cor/raça, gênero, faixa etária e o quantitativo de alunos da Educação Especial. As visualizações facilitam a visão ampla do perfil dos estudantes e da diversidade na educação do Brasil.

A escolha dos gráficos foi feita com base nas características de cada indicador, por exemplo, o número total de matrículas ao longo dos anos fica mais claro em um gráfico de linhas, como mostra a Figura 2. Já a distribuição por estado, conforme a Figura 3, é melhor explorada por meio de um mapa interativo do Brasil, no qual o usuário pode selecionar ou passar o cursor sobre cada estado para visualizar os valores correspondentes.

Outro indicador implementado no Superset é o docentes por escola, no qual o volume de dados é maior, devido a quantidade de escolas existentes no Brasil. Nesse caso, o uso de visualizações gráficas tradicionais poderia comprometer informações importantes. Para lidar com esse tipo de cenário, foi adotada a visualização de dados tabular disponível no Superset, conforme mostra a Figura 4. Essa visualização facilita a visão macro do dados, desde o total de docentes por ano no país, até uma análise mais específica, como o número de docentes da Educação Especial em uma escola específica.

5. Conclusão

Este artigo demonstrou a viabilidade e os benefícios da utilização do Apache Superset como uma solução de código aberto para a análise e visualização interativa dos dados do Censo Escolar. Para isso, foi apresentado um fluxo de dados que abrange desde a ingestão até a visualização, necessário para facilitar a exploração anual das informações e superar

[2]Disponível em: https://dadoseducacionais.c3sl.ufpr.br/plataforma/

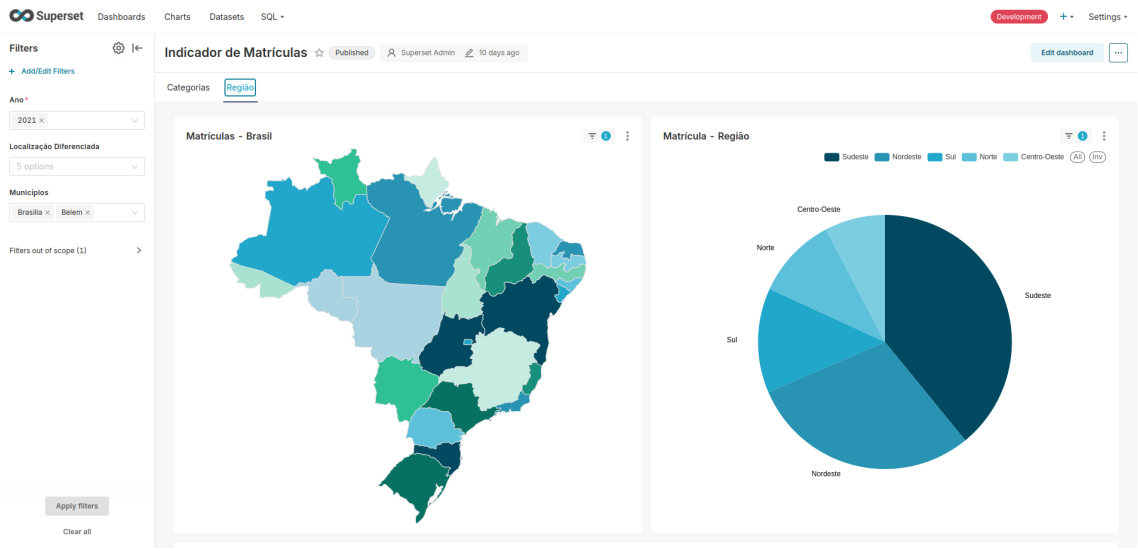

Figura 3. Exemplo de *Dashboard* para o Indicador de matrículas com gráfico de mapa e setores.

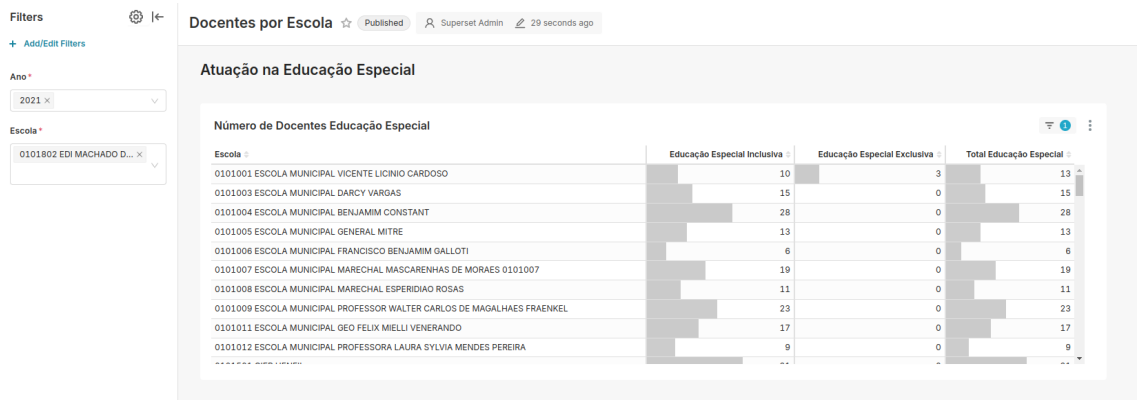

Figura 4. Exemplo de Dashboard para o indicador de docentes por escola.

desafios relacionados à heterogeneidade, à alta dimensionalidade e ao grande volume de dados educacionais. Com o uso do ClickHouse para centralizar os dados e o Apache Superset para a visualização, foi possível transformar dados dispersos entre diferentes arquivos CSV em visualizações claras e acessíveis revelando padrões temporais, espaciais e demográficos. A adoção de um fluxo de ingestão, transformação e modelagem analítica, baseado em ferramentas de código aberto permitiu a construção de ambiente analítico relacional, adequado para consultas SQL e visualizações interativas.

Como trabalhos futuros, destaca-se a integração de bases complementares, como o Sistema de Informações sobre Orçamentos Públicos em Educação (SIOPE) e o Censo da Educação Superior, para permitir cruzamentos a análises mais abrangentes. Outro caminho a ser explorado é a automatização do consumo de dados educacionais pelo NiFi, o que facilitaria a ingestão de dados. Os fluxos de transformação também precisam tratar as mudanças ocorridas nos esquemas ao longo dos anos para garantir a consistência das informações. Além disso, é recomendada a ampliação das visualizações, com o objetivo de aprofundar os estudos e a compreensão sobre a educação brasileira.

Referências

Apache Software Foundation (2024). Apache NiFi. `https://nifi.apache.org/`. Acesso em: 14 jun. 2025.

Ehrenfried, H. V., Eckelberg, R., Iboshi, H., Todt, E., Weingaertner, D., and Del Fabro, M. D. (2019). Hotmapper: Historical open data table mapper. In *EDBT*, pages 550–553.

Karabtsev, S., Kotov, R., Davzit, I., and Gurov, E. (2023). Building data marts to analyze university faculty activities using power bi. In *E3S Web of Conferences*, volume 419, page 02014. EDP Sciences.

Mbata, A., Sripada, Y., and Zhong, M. (2024). A survey of pipeline tools for data engineering. *arXiv preprint arXiv:2406.08335*.

Schulze, R., Schreiber, T., Yatsishin, I., Dahimene, R., and Milovidov, A. (2024). Clickhouse-lightning fast analytics for everyone. *Proceedings of the VLDB Endowment*, 17(12):3731–3744.

Superset, A. (2024). Apache Superset. `https://superset.apache.org/`. Acesso em: 5 maio 2025.

Van den Burg, G. J., Nazábal, A., and Sutton, C. (2019). Wrangling messy csv files by detecting row and type patterns. *Data Mining and Knowledge Discovery*, 33(6):1799–1820.

Yamanaka, M., de Almeida, D., de Almeida, P. R., Dominico, S., Peres, L., Sunye, M., and Almeida, E. (2024). Statistical validation of column matching in the database schema evolution of the brazilian public school census. In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados - SBBD*, pages 498–509.