DataIF: Um framework ETL para integração, visualização e consulta de dados públicos do IFRS

Yuri Ferreira Rodrigues[1], **Alexandre Abreu de Paula**[1], **Andrws Aires Vieira**[1]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul
Campus Ibirubá
Rua Nelsi Ribas Fritsch, 1111 – CEP: 98200-000 – Ibirubá – RS – Brasil

Abstract. *This paper presents DataIF, a framework for integrating and analyzing public educational data within the IFRS. The solution combines Extract, Transform, and Load (ETL) processes with authentication services, APIs, workflow orchestration, data persistence, and analytical visualization. The study is conducted as an applied case study using microdata from the Nilo Peçanha Platform, with containerized deployment and operational validation of data pipelines. The results demonstrate the feasibility of a reproducible data integration pipeline, ensuring execution traceability and continuous delivery of analytical dashboards, supporting institutional data governance and data-driven decision-making.*

Resumo. *Este artigo apresenta o DataIF, um framework de integração e análise de dados públicos educacionais voltado ao IFRS. A solução combina processos de Extração, Transformação e Carga (ETL) com serviços de autenticação, APIs, orquestração de workflows, persistência de dados e visualização analítica. A pesquisa é conduzida como estudo de caso aplicado utilizando microdados da Plataforma Nilo Peçanha, com implantação conteinerizada e validação operacional dos fluxos. Os resultados demonstram a viabilidade de um pipeline reprodutível de integração de dados, com rastreabilidade das execuções e disponibilização contínua de dashboards analíticos, contribuindo para a governança e o uso estratégico de dados institucionais.*

1. Introdução

Os Institutos Federais (IF) são órgãos federais vinculados ao Ministério da Educação, caracterizada como instituição de educação superior, básica e profissional, pluricurricular e multicampi. De acordo com seu Plano de Desenvolvimento Institucional (PDI) 2024-2028, o IFRS está presente em 16 municípios do Rio Grande do Sul e busca articular ensino, pesquisa e extensão aos arranjos sociais, culturais e produtivos locais [Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul 2023, p. 24–25]. Portanto, a gestão eficiente e o acesso transparente às informações institucionais são fundamentais para a tomada de decisões e a avaliação de métricas em ambientes educacionais.

No contexto atual do Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul (IFRS), as informações relevantes sobre alunos, cursos e desempenho institucional ainda se encontram dispersas entre plataformas. Esse cenário evidencia a necessidade de uma abordagem unificada de catalogação, de modo a melhorar a visualização

e o potencial analítico da instituição e da comunidade a partir dos dados que lhes dizem respeito.

Este artigo propõe a implementação de um framework de Extração, Transformação e Carga (ETL) para dados do IFRS, juntamente com uma camada de visualização e consulta voltada à análise institucional. Embora existam iniciativas de integração e disponibilização de indicadores, como a Plataforma Nilo Peçanha (PNP), o acesso às informações institucionais do IFRS ainda apresenta limitações de integração e usabilidade para servidores, gestores e comunidade acadêmica, como demonstrado nos trabalhos de [Marques 2019] e na análise de indicadores da educação superior [Lima e Pires 2022].

Essa limitação dificulta a análise, a tomada de decisões e a avaliação de métricas de alunos, cursos e outras informações essenciais ao desempenho institucional na região. O diagnóstico do PDI 2024-2028 reforça essa leitura ao apontar, entre os pontos fracos institucionais, a falta de padronização dos processos de trabalho e fragilidades na comunicação, além de prever iniciativas para desenvolver sistemas que viabilizem o controle e a transparência sobre a gestão institucional [Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul 2023, p. 65 e 82]. Além disso, o processo atual para obtenção desses dados é moroso e suscetível a erros, pois requer a busca manual em diversas fontes e plataformas independentes.

A falta de uma plataforma unificada impede uma visão ampla e comparativa do cenário educacional, comprometendo a geração de relatórios estratégicos e o acompanhamento eficaz de indicadores acadêmicos e administrativos. A ausência de dados organizados e acessíveis também restringe a participação ativa da comunidade acadêmica e externa, reduzindo a transparência e a confiança nas informações divulgadas.

No momento de elaboração deste trabalho (primeiro semestre de 2026), o IFRS encontra-se em processo de criação de seu Plano de Dados Abertos (PDA) para promover maior transparência. Até a consolidação dessa iniciativa, permanece uma lacuna entre a produção de dados e seu uso efetivo para análise e melhoria institucional.

Tal cenário destaca a necessidade de uma solução tecnológica que facilite o levantamento, armazenamento, organização e visualização dos dados institucionais, proporcionando informações mais precisas e de fácil acesso para todos os interessados. Neste sentido, a principal contribuição deste trabalho está na proposição e validação de uma arquitetura reprodutível de integração de dados públicos educacionais, aliando engenharia de dados, governança e disponibilização analítica em um ambiente institucional real.

1.1. Objetivo

O objetivo geral deste artigo é apresentar o DataIF, um framework de Extração, Transformação e Carga desenvolvido para a integração, o armazenamento e a disponibilização analítica de microdados educacionais da Plataforma Nilo Peçanha, tendo o contexto do Instituto Federal do Rio Grande do Sul como estudo de caso aplicado para estruturar uma base confiável de visualização e consulta de indicadores.

A partir desse objetivo, o desenvolvimento foi conduzido por uma abordagem aplicada, detalhada na seção metodológica, que organiza as etapas técnicas de implementação, validação e disponibilização analítica da plataforma.

2. Trabalhos Correlatos

Nesta seção são apresentados os trabalhos que fundamentam a proposta deste projeto, contextualizando estudos anteriores e apontando direções metodológicas e conceituais para o desenvolvimento do artigo. Além disso, descrevem-se convergências e diferenças em relação ao presente trabalho.

O trabalho de [Baker, Isotani e Carvalho 2011] se destaca na área de Mineração de Dados Educacionais (MDE) ao discutir o crescimento desse campo no cenário internacional e apontar desafios para sua consolidação no Brasil, como limitações de infraestrutura, dificuldade de acesso a bases de dados e necessidade de adaptação das soluções à realidade local. Esse referencial sustenta a proposta ao evidenciar o papel da análise de dados no apoio a decisões educacionais.

Em perspectiva complementar, o estudo de [Rigo et al. 2014] aborda aplicações de Mineração de Dados Educacionais e *Learning Analytics* com foco na evasão escolar, discutindo fatores associados ao abandono e analisando aplicações em cursos de graduação a distância. Embora o foco esteja na evasão, o estudo reforça a importância de ampliar variáveis analisadas e estruturar processos de monitoramento contínuo para subsidiar ações institucionais baseadas em dados.

A integração de diferentes fontes de dados é explorada por [Silva, Ruy e Mutz 2022], que propõem uma abordagem para análise de evasão escolar baseada em ontologia e padronização semântica. Essa contribuição aproxima-se do presente trabalho ao enfatizar a organização de dados heterogêneos para consultas mais consistentes, aspecto central na construção de uma plataforma unificada para exploração analítica no contexto institucional.

Por fim, o trabalho de [Dia et al. 2025] apresenta o *EduGuard RetainX*, um painel analítico voltado à predição de retenção estudantil no ensino superior. O estudo se relaciona ao *DataIF* por combinar integração de dados e visualização para apoio à decisão, embora priorize modelos preditivos, enquanto este projeto concentra-se inicialmente na integração e disponibilização de dados abertos da PNP.

2.1. Discussão e Diferenciais da Proposta

Os trabalhos analisados convergem ao destacar a relevância da MDE, da integração de dados e da visualização analítica para apoiar decisões educacionais. Em conjunto, eles abrangem fundamentos conceituais, aplicações em evasão e retenção, integração semântica de múltiplas fontes e painéis voltados ao acompanhamento institucional.

No aspecto metodológico, os estudos variam entre discussão teórica, implementação de sistemas analíticos e proposição de arquiteturas de integração. Em contraste, o presente projeto adota uma abordagem aplicada orientada à coleta, integração, tratamento e disponibilização de dados abertos institucionais do IFRS em uma infraestrutura operacional reprodutível.

Assim, a principal contribuição proposta está na construção de uma infraestrutura integrada com ETL, orquestração, persistência, visualização e consulta acessível à comunidade institucional. Essa combinação busca ampliar a transparência, a autonomia de análise e o suporte à tomada de decisão baseada em dados.

3. Metodologia

Esta seção apresenta os procedimentos metodológicos adotados para implementar e validar a plataforma *DataIF* na integração com a PNP. A abordagem combina pesquisa aplicada, estudo de caso e validação experimental orientada à engenharia de dados.

3.1. Caracterização metodológica e objetivo

O objetivo metodológico foi construir uma infraestrutura reprodutível para descoberta, extração, persistência e disponibilização analítica de microdados públicos educacionais. O recorte empírico adotado foi o catálogo público da PNP em *Power BI*, com ênfase em execução programática e governança operacional.

Para operacionalizar esse objetivo, a metodologia foi orientada por cinco frentes específicas: descrever a arquitetura de *backend* para orquestração dos fluxos; mapear os *endpoints* da PNP e o conector de acesso direto aos microdados; apresentar a modelagem relacional adotada para persistência e organização; demonstrar a integração com a camada de BI no Metabase; e registrar a interface *web* de operação e consulta.

As propriedades priorizadas foram reprodutibilidade, rastreabilidade, separação de responsabilidades e segurança de operação.

3.2. Contexto técnico-experimental

A implementação foi organizada em arquitetura modular conteinerizada com *Docker Compose*, de modo a garantir padronização de dependências e repetibilidade de execução. A solução integrou os seguintes componentes: *PostgreSQL* (repositório operacional e analítico), *Apache Airflow* (orquestração), *FastAPI* (serviços administrativos e de integração), *React + Vite* (interface de operação), *Keycloak* (autenticação e autorização) e *Metabase* (dashboards).

Conforme a Figura 1, a visualização dos pontos da infraestrutura e do fluxograma segue a sequência abaixo:

1. **Administrador**: ponto de entrada operacional em que o administrador acessa a plataforma e aciona os fluxos de integração.
2. **Autenticação (Keycloak)**: serviço responsável pela autenticação e pelo controle de acesso às funcionalidades administrativas.
3. **FastAPI**: camada que intermedeia as ações da interface, consulta a fonte pública da PNP, persiste as configurações das conexões no PostgreSQL e integra a operação com o Airflow.
4. **Orquestração e agendamento (Airflow)**: componente de orquestração que executa os workflows de validação e ingestão a partir das conexões e parâmetros persistidos.
5. **Conector (PNP)**: fonte pública de dados consultada tanto na etapa de descoberta dos ativos quanto na etapa de extração dos microdados.
6. **Banco de Dados (PostgreSQL)**: repositório central da plataforma, onde são mantidas as configurações administrativas, os registros operacionais, os dados ingeridos e as estruturas analíticas intermediárias e publicadas.
7. **BI (Metabase)**: camada de disponibilização analítica que consome os dados preparados no PostgreSQL para publicação de dashboards institucionais.

8. **Usuário Final**: consumidor das informações disponibilizadas pelo Metabase, com acesso aberto a qualquer pessoa. Engloba desde o público interno, para fins de acompanhamento e tomada de decisão estratégica, até a sociedade em geral, promovendo a transparência ativa dos dados institucionais.

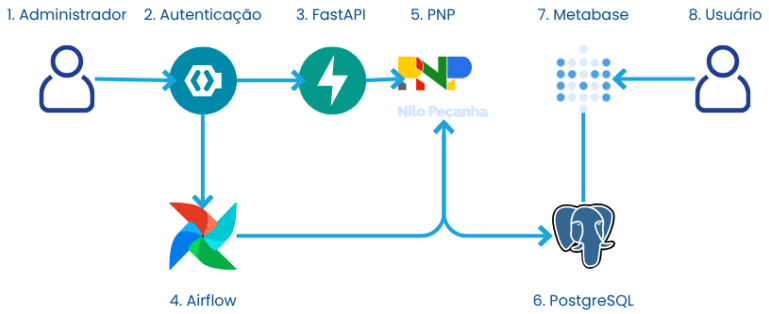

Figura 1. Infraestrutura organizacional do projeto.

3.3. Fonte de dados e unidade de análise

A fonte primária utilizada foi o catálogo público de microdados da PNP em *Power BI*, tratado como ambiente dinâmico de descoberta de ativos. A unidade operacional de análise foi definida pela combinação entre ano-base e tipo de microdado.

Cada conexão administrativa registra os recortes selecionados e a agenda de execução, formando o contrato de entrada para os fluxos automatizados.

3.4. Procedimentos metodológicos de implementação

A implementação foi incremental e organizada em etapas sequenciais:

1. estabilização da arquitetura transversal da plataforma;
2. configuração das conexões da PNP por recorte semântico;
3. validação programática da disponibilidade dos ativos públicos;
4. execução orquestrada dos fluxos de validação e ingestão;
5. extração direta dos links públicos, limpeza, catalogação e persistência dos dados na camada bruta.

No estado atual do projeto, a automação cobre integralmente a descoberta e a ingestão inicial. Já a promoção para camadas analíticas posteriores permanece sob governança manual, preservando controle semântico enquanto a modelagem temática evolui.

3.5. Reprodutibilidade e delimitações

A reprodutibilidade foi assegurada pela conteinerização do ambiente, pela separação entre serviços e pela persistência das configurações administrativas de captura. Em termos de escopo, o atual estágio do projeto concentra-se em validar a integração dos dados até a camada bruta e em consolidar a infraestrutura de disponibilização analítica.

As imagens dos serviços foram estruturadas em contêineres Docker e organizadas para publicação em uma stack remota com Docker Compose, permitindo padronização de versões e rastreabilidade de implantação disponível em: `https://hub.docker.com/u/dataif`.

4. Resultados

Em relação ao objetivo geral, o *DataIF* demonstrou integração, armazenamento e disponibilização analítica de microdados da PNP em um fluxo reprodutível, com descoberta de ativos, ingestão estruturada e rastreabilidade operacional das execuções em ambiente institucional.

Quanto às frentes metodológicas, foram alcançados resultados na arquitetura de *backend*, no mapeamento de *endpoints* e no conector automatizado, na modelagem relacional das camadas de dados, na publicação de dashboards no Metabase e na entrega de interface web para operação e consulta.

4.1. Infraestrutura integrada de backend e frontend

A infraestrutura integrada de *backend* e *frontend* consolidou um fluxo contínuo entre descoberta, ingestão, transformação e persistência de dados públicos da PNP, com separação de responsabilidades entre componentes e maior confiabilidade para operação institucional. Esse arranjo fortaleceu a rastreabilidade e a governança do ciclo técnico do projeto.

Na camada de disponibilização, a integração com o Metabase conectou o processamento interno à exploração analítica por meio de painéis e consultas institucionais. Como ilustrado na Figura 2, os resultados evidenciam uma infraestrutura concluída e apta a sustentar, de forma contínua, o uso analítico dos dados integrados.

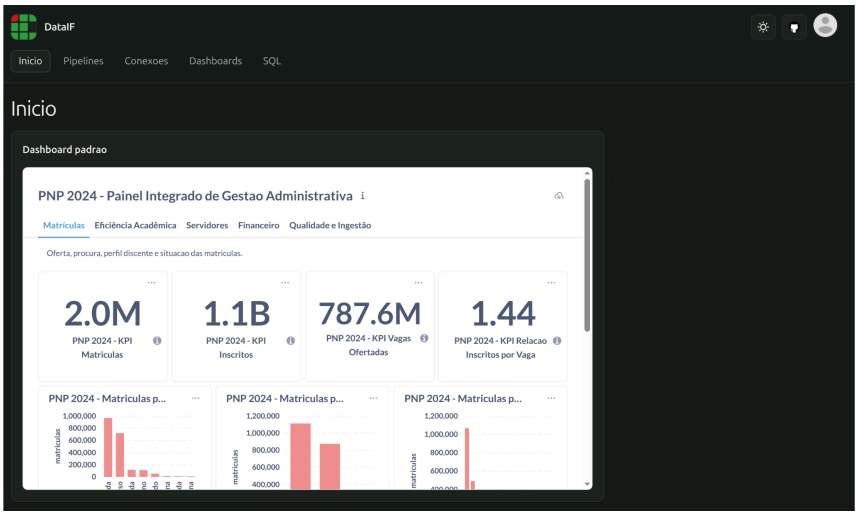

Figura 2. Painel analítico no Metabase com dados integrados pelo *DataIF*.

4.2. Evidências operacionais

A rotina de checagem no Airflow inicia com uma configuração administrativa e verifica, no catálogo público da PNP, a correspondência dos recortes solicitados. Nessa etapa, o fluxo resolve os links públicos, testa a acessibilidade e confirma a disponibilidade dos arquivos esperados para ingestão.

Após a checagem positiva, o pipeline executa download e parsing, registra metadados operacionais e persiste os registros normalizados na camada bruta. Como ilustrado

na Figura 3, a execução da DAG evidencia um ciclo rastreável e reprodutível entre validação de fonte e carga.

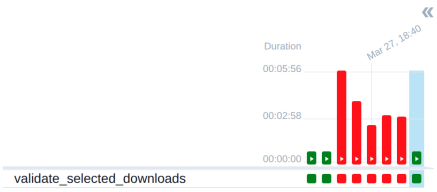

Figura 3. Execução da DAG de validação no Airflow.

4.3. Síntese dos resultados e limitações

Apesar dos avanços, o escopo atual ainda é parcial em relação à amplitude temática da plataforma-fonte. A evolução das camadas analíticas depende de curadoria incremental por domínio e de ampliação progressiva dos recortes integrados.

Ainda assim, os resultados sustentam a *DataIF* como infraestrutura funcional de integração e governança de dados públicos educacionais.

5. Conclusão e Trabalhos Futuros

A implementação da plataforma *DataIF* demonstrou a viabilidade técnica de uma arquitetura externa para autenticação, extração, persistência, rastreabilidade e publicação analítica de dados educacionais de natureza pública. O estudo mostrou que a integração com a PNP pode ser tratada como um fluxo reprodutível de engenharia de dados, e não apenas como automação pontual de consultas, com acesso direto aos links públicos de microdados por ano-base e endpoint para download, limpeza e catalogação.

No estado atual, a solução já consolidou componentes essenciais para operação: ambiente conteinerizado, autenticação administrativa, descoberta de recortes por ano e endpoint, ingestão automatizada, limpeza e catalogação dos microdados, persistência em camadas analíticas e disponibilização inicial dos dados em painel externo.

No recorte do estudo de caso aplicado ao IFRS com dados da PNP, esses resultados sustentam a caracterização da *DataIF* como plataforma funcional em consolidação progressiva, com contribuição metodológica e aplicada para cenários de integração de dados públicos com ingestão direta e governança operacional.

Embora a cobertura analítica ainda não reproduza toda a amplitude da plataforma-fonte, o projeto já entrega uma base consistente de observabilidade do processo de integração e dos dados coletados da plataforma. Assim, sua principal contribuição está no desacoplamento entre origem dos dados e ambiente de consumo analítico, com preservação de rastreabilidade e potencial de reuso institucional.

5.1. Trabalhos futuros

A continuidade desta pesquisa priorizará algumas das frentes complementares levantadas durante a criação do projeto e garantir a sua evolução:

- **Integração madura de consultas em linguagem natural:** evolução da camada de NL2SQL para ampliar contexto semântico, controle de permissões, auditoria das consultas e avaliação da qualidade das respostas;

- **Expansão de conectores:** desenvolvimento de novos conectores para outras bases públicas relevantes, de modo a validar a generalização da arquitetura além do estudo de caso da PNP;
- **Fortalecimento de observabilidade e governança:** ampliação de métricas de qualidade de dados, monitoramento operacional por etapa e visualizações específicas para auditoria da pipeline.

Em conjunto, essas direções indicam uma trajetória coerente de maturação da *DataIF*: consolidar o que já foi validado, ampliar o alcance analítico da plataforma e evoluir para um ecossistema extensível de integração e acesso analítico a dados públicos educacionais.

Referências

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineração de dados educacionais: Oportunidades para o brasil. *Revista Brasileira de Informática na Educação*, v. 19, n. 02, p. 03, 2011. ISSN 2317-6121. Disponível em: <http://milanesa.ime.usp.br/rbie/index.php/rbie/article/view/1301>.

DIA, N. J. et al. Eduguard retainx: An advanced analytical dashboard for predicting and improving student retention in tertiary education. *SoftwareX*, v. 29, p. 102057, 2025. ISSN 2352-7110. Disponível em: <https://www.sciencedirect.com/science/article/pii/S235271102500024X>.

Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul. *Plano de Desenvolvimento Institucional 2024–2028*. Bento Gonçalves: [s.n.], 2023. Aprovado pelo Conselho Superior do IFRS pela Resolução nº 054, de 12 de dezembro de 2023.

LIMA, M. S.; PIRES, M. R. G. M. Avaliação da taxa do acesso aos dados abertos das universidades federais a partir dos indicadores de fluxo do ensino superior do INEP. *Avaliação: Revista da Avaliação da Educação Superior (Campinas)*, scielo, v. 27, p. 531 – 552, 09 2022. ISSN 1414-4077. Disponível em: <http://educa.fcc.org.br/scielo.php?script=sci_arttext&pid=S1414-40772022000300531&nrm=iso>.

MARQUES, K. Dados abertos nas universidades federais: envolvimento interno e divulgação para a sociedade. *Revista Brasileira de Biblioteconomia e Documentação*, v. 15, n. 2, p. 58–80, maio 2019. Disponível em: <https://rbbd.febab.org.br/rbbd/article/view/1150>.

RIGO, S. et al. Aplicações de mineração de dados educacionais e learning analytics com foco na evasão escolar: oportunidades e desafios. *Revista Brasileira de Informática na Educação*, v. 22, n. 01, p. 132, 2014. ISSN 2317-6121. Disponível em: <http://milanesa.ime.usp.br/rbie/index.php/rbie/article/view/2423>.

SILVA, E. M. da; RUY, F. B.; MUTZ, F. W. Abordagem para análise de múltiplas fontes de dados de evasão escolar. *Portal de Periódicos Univali*, v. 13, p. 08, 2022. Disponível em: <https://periodicos.univali.br/index.php/acotb/article/view/18791>.