PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA
Pós-graduação *Lato Sensu* em Engenharia de Dados

Gabriel Heitor dos Santos Trigo de Jesus

INGESTÃO E ANÁLISE DE DADOS DO SIOPE: UMA ABORDAGEM EM ENGENHARIA DE DADOS PARA ORÇAMENTOS PÚBLICOS EM EDUCAÇÃO

Rio de Janeiro
2024

Gabriel Heitor dos Santos Trigo de Jesus

INGESTÃO E ANÁLISE DE DADOS DO SIOPE: UMA ABORDAGEM EM ENGENHARIA DE DADOS PARA ORÇAMENTOS PÚBLICOS EM EDUCAÇÃO

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Engenharia de Dados como requisito parcial à obtenção do título de especialista.

Rio de Janeiro
2024

SUMÁRIO

1. Introdução

1.1. Contextualização

A análise de dados públicos é uma ferramenta essencial para a transparência e a eficiência na gestão de recursos governamentais. No Brasil, a área de educação enfrenta desafios constantes relacionados à alocação e ao uso eficiente de recursos, especialmente em estados com grandes disparidades econômicas e sociais, como o Rio de Janeiro. Nesse contexto, o Sistema de Informações sobre Orçamentos Públicos em Educação (SIOPE) é uma fonte valiosa de dados, fornecendo informações detalhadas sobre o orçamento destinado à educação em diversas esferas governamentais. No entanto, o volume e a complexidade desses dados exigem soluções robustas de ingestão e processamento para garantir que sejam utilizados de forma eficaz em análises estratégicas.

Este trabalho tem como objetivo desenvolver um pipeline de ingestão de dados a partir do SIOPE, com foco na análise do orçamento de educação do estado do Rio de Janeiro. O pipeline visa automatizar o processo de extração, transformação e carga dos dados, permitindo uma visão mais clara e atualizada do panorama orçamentário. Ao transformar dados brutos em insights valiosos, o projeto contribui para uma maior transparência e uma gestão mais eficiente dos recursos públicos destinados à educação, oferecendo uma solução real e relevante tanto para gestores públicos quanto para pesquisadores e cidadãos interessados na melhoria da educação no estado.

Essa iniciativa é fundamental para identificar padrões de alocação, apontar possíveis ineficiências e apoiar a tomada de decisões baseada em dados, promovendo um uso mais consciente e estratégico dos recursos públicos.

1.2. Problema Proposto

O problema que este trabalho propõe resolver está diretamente relacionado à análise de dados orçamentários da educação pública no estado do Rio de Janeiro, utilizando

informações provenientes do Sistema de Informações sobre Orçamentos Públicos em Educação (SIOPE). A engenharia de dados será aplicada para construir um pipeline automatizado de ingestão, transformação e carga dos dados, a fim de tornar a análise do orçamento mais acessível e eficiente para diferentes stakeholders. Agora, vamos detalhar esse problema utilizando a técnica dos 5-Ws:

A educação é uma das áreas mais críticas para o desenvolvimento de uma sociedade, e a gestão adequada dos recursos públicos destinados a ela é fundamental para assegurar que as políticas educacionais sejam eficazes. Porém, a análise dos dados orçamentários frequentemente é um processo demorado e suscetível a erros quando feito manualmente, dificultando a transparência e a tomada de decisões informadas. O problema é importante porque uma análise mais eficiente dos orçamentos permite identificar desvios, otimizar o uso dos recursos e, consequentemente, melhorar a qualidade da educação oferecida à população.

Os dados que serão analisados pertencem ao governo, mais especificamente à Secretaria de Educação do estado do Rio de Janeiro, e fazem parte do SIOPE, que reúne informações sobre os recursos financeiros destinados à educação em todas as esferas do governo. Esses dados são fundamentais para entender como os recursos são distribuídos e aplicados, e, portanto, têm um impacto direto sobre a gestão pública e o planejamento educacional.

1.3. Objetivos

O objetivo deste trabalho é desenvolver um pipeline automatizado para ingestão e análise de dados orçamentários da educação pública do estado do Rio de Janeiro, utilizando informações do Sistema de Informações sobre Orçamentos Públicos em Educação (SIOPE). Esse pipeline visa facilitar o processo de coleta e preparação dos dados, permitindo análises mais rápidas e precisas, ajudando na tomada de decisões.

De forma específica, buscamos:

1. Automatizar a coleta dos dados do SIOPE, eliminando processos manuais.

2. Organizar os dados em diferentes camadas, desde o formato bruto até dados prontos para análise.
3. Fornecer uma base de dados que apoie a criação de relatórios sobre o uso dos recursos da educação no estado.
4. Contribuir para uma gestão pública mais transparente e eficiente.
5. Criar uma solução replicável, que possa ser usada em outras áreas ou regiões.

Com isso, esperamos otimizar o uso dos recursos públicos e melhorar a qualidade da educação no estado.

2. Desenvolvimento

Os dados utilizados neste trabalho foram obtidos de fontes públicas e estão disponíveis na internet Portal de Dados Abertos. Para garantir a clareza, descreveremos detalhadamente a origem dos dados, o formato em que foram obtidos, a modelagem conceitual, e as tecnologias e ferramentas utilizadas na arquitetura implementada.

- Despesas SIOPE: Os dados relacionados às despesas da educação pública foram extraídos do sistema SIOPE (Sistema de Informações sobre Orçamentos Públicos em Educação), através do link: Despesas SIOPE, no formato JSON.
- Receitas SIOPE: Da mesma forma, os dados relacionados às receitas destinadas à educação foram coletados através do link: Receitas SIOPE, também no formato JSON.
- Municípios (IBGE): Para complementar a análise, foram utilizados dados geográficos sobre os municípios do Brasil, extraídos da API do IBGE. Os dados estão disponíveis em formato JSON por meio do link: Municípios IBGE.

2.1. Modelagem Conceitual

Os dados de despesas e receitas foram modelados e organizados dentro da arquitetura de medallion (Bronze, Silver e Gold) em camadas separadas:

- Camada Bronze: Nesta camada, os dados brutos em JSON são convertidos para delta table e armazenados conforme foram extraídos das fontes. Nenhuma transformação é realizada nessa fase, garantindo que os dados originais fiquem preservados para possíveis revisões ou reprocessamentos.
- Camada Silver: Na segunda camada, os dados são limpos e transformados para remover inconsistências e duplicatas. Nessa etapa, são realizadas as primeiras transformações para padronizar os formatos de datas, valores e variáveis, facilitando o processamento subsequente.
- Camada Gold: Finalmente, os dados organizados na camada Silver são agregados e transformados em um formato otimizado para análise. Nesta camada, integram-se os dados de despesas e receitas formando uma única tabela fato e informações

geográficas dos municípios formando uma dimensão, permitindo a análise cruzada de recursos orçamentários por localidade e ano.

2.2. Relacionamento Entre os Dados

- Despesas e Receitas SIOPE: Os dois datasets foram consolidadas em uma única tabela fato na camada Gold. Essa tabela contém os registros financeiros, tanto de receitas quanto de despesas.
- Municípios (IBGE): Os dados de municípios são integrados através de uma chave única de identificação de cada município, permitindo o cruzamento das informações de localização com os dados financeiros.

O relacionamento entre a tabela fato e a tabela dimensão ocorre através do campo ID do município, que é a chave comum entre as duas tabelas. Isso possibilita a análise de receitas e despesas por município, permitindo que se façam cruzamentos de dados como:

- Comparação entre municípios quanto ao volume de receitas recebidas e despesas realizadas;
- Análise geográfica para identificar padrões de alocação de recursos por região;
- Estudos temporais para entender como a distribuição dos recursos evolui ao longo dos anos em diferentes localidades.

2.3. Tecnologias e Ferramentas Utilizadas

- Camada de Armazenamento: O armazenamento dos dados é feito no Azure Data Lake Storage Gen2, que permite a organização dos dados em diferentes camadas de processamento (Bronze, Silver e Gold), garantindo escalabilidade e eficiência.
- Camada de Processamento: Para processamento dos dados, foi utilizado o Azure Databricks com Spark. O Spark facilita a transformação de grandes volumes de dados, realizando a limpeza, agregação e preparação das tabelas de maneira eficiente.
- Arquitetura Medalhão: A arquitetura escolhida segue o modelo de medalhão, com tabelas Delta nas camadas Bronze, Silver e Gold, proporcionando um pipeline

robusto e flexível que suporta desde o armazenamento de dados brutos até a análise refinada.

- DataViz: Para a visualização e análise dos dados processados, foi utilizado o Power BI, que permite a criação de relatórios interativos e dashboards que facilitam a análise de tendências e a distribuição dos recursos públicos.
- Orquestração: O pipeline de ingestão e processamento foi orquestrado com o Databricks Workflow, que automatiza a execução das tarefas, garantindo a atualização periódica dos dados e a execução das transformações necessárias.

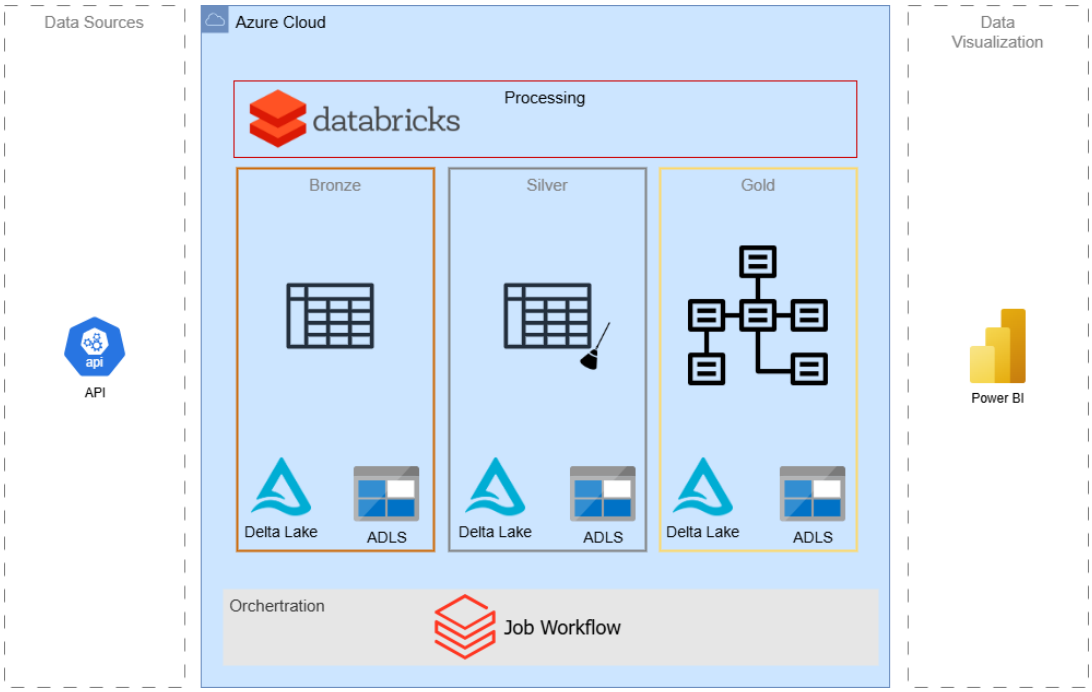

Figura 1: Desenho da arquitetura proposta para o projetos

3. Ingestão de dados

3.1. Camada Bronze (Ingestão dos Dados Brutos)

Na camada Bronze, realizo a conexão direta com as APIs de origem, configurando parâmetros específicos para garantir uma ingestão eficiente e segmentada dos dados. Os principais processos que executo são:

- Conexão com as APIs de Despesas e Receitas (SIOPE): Para cada chamada às APIs de despesas e receitas, utilizo parâmetros dinâmicos de Ano, Mês e UF (Unidade Federativa), que são inseridos na URL da API. Isso me permite extrair dados específicos de períodos e localizações, evitando a ingestão de dados desnecessários.
- Conexão com a API de Municípios (IBGE): A API de municípios é consumida sem a necessidade de parâmetros adicionais, obtendo a lista completa de municípios brasileiros.
- Armazenamento em Delta: Após a extração dos dados, armazeno-os em sua forma bruta na camada Bronze do Azure Data Lake Storage Gen2 no formato Delta para otimização e performance do processo de leitura e escrita além de mantermos todos os benefícios do teorema ACID.

3.2. Camada Silver (Transformação e Limpeza de Dados)

Na camada Silver, transformo os dados brutos da Bronze, aplicando as primeiras regras de negócio, removendo duplicidades e limpando-os, garantindo que estejam preparados para análises posteriores. As ações que realizo nessa camada incluem:

- Delete e Insert (Receitas e Despesas): Para os dados de receitas e despesas, realizo um delete dos registros do mesmo período (ano/mês) que foi carregado na Bronze. Ou seja, se na Bronze estou carregando dados de janeiro de 2024, primeiro deleto esses dados de janeiro de 2024 da camada Silver, garantindo que os dados antigos

sejam substituídos pelos novos. Após isso, gravo os dados da Bronze na Silver com a opção overwrite, garantindo que apenas as versões mais recentes dos dados sejam mantidas.

- Merge (Municípios): Para os dados de municípios, utilizo um processo de merge entre os dados novos e os existentes na Silver. A condição para o merge é baseada no ID do município, onde realizo um update caso haja divergências entre os registros, e inserções de novos municípios que não existiam previamente.
- Conversão de Datatypes: Durante a carga na camada Silver, converto os tipos de dados para formatos mais adequados (ex.: inteiros, strings, datas), de modo que a transformação posterior na camada Gold ocorra de forma fluida e sem inconsistências.

3.3. Camada Gold (Agregação e Organização para Análise)

Na camada Gold, preparo os dados para análises e relatórios. Consolido as tabelas de receitas e despesas em uma única tabela fato, enquanto os dados de municípios são tratados como dimensão. Os principais processos que executo são:

- Consolidação de Receitas e Despesas em Tabela Fato: As tabelas de receitas e despesas são unificadas em uma única tabela fato. Durante esse processo, colunas irrelevantes são descartadas, e mantenho apenas aquelas essenciais para análise (como ID do município, ano, tipo de receita/despesa, valores). Como ambas as tabelas possuem uma estrutura idêntica após a padronização na Silver, realizo a junção através de append, sem necessidade de transformações adicionais. Essa abordagem garante que os dados de receitas e despesas sejam armazenados de forma unificada e consistente para análise posterior.
- Delete e Insert (Tabela Fato): Similar à Silver, para a tabela fato de receitas e despesas, deleto os registros correspondentes ao período que estou carregando da camada Silver (ex.: janeiro de 2024) e, em seguida, insiro os novos dados da Silver. Isso assegura que apenas dados atualizados sejam mantidos na Gold.

- Tratamento de Municípios como Dimensão: A tabela de municípios na camada Gold é tratada como uma tabela de dimensão. Realizo uma limpeza adicional para excluir colunas irrelevantes, mantendo apenas os dados necessários para enriquecer a tabela fato. O processo de merge é similar ao utilizado na Silver, onde, com base no ID do município, atualizo os registros existentes e insiro novos dados conforme necessário.

Essa estrutura na camada Gold permite análises otimizadas e agregadas, com dados prontos para serem consumidos diretamente por ferramentas de visualização, como o Power BI.

4. Orquestração de dados

Para este projeto, utilizei o Databricks Workflow Job como ferramenta de orquestração de dados, responsável por automatizar e gerenciar o pipeline de ingestão, transformação e carga dos dados. O workflow job do Databricks permite a execução coordenada de tarefas em sequência, garantindo que cada etapa do pipeline seja realizada na ordem correta, com monitoramento e controle de falhas.

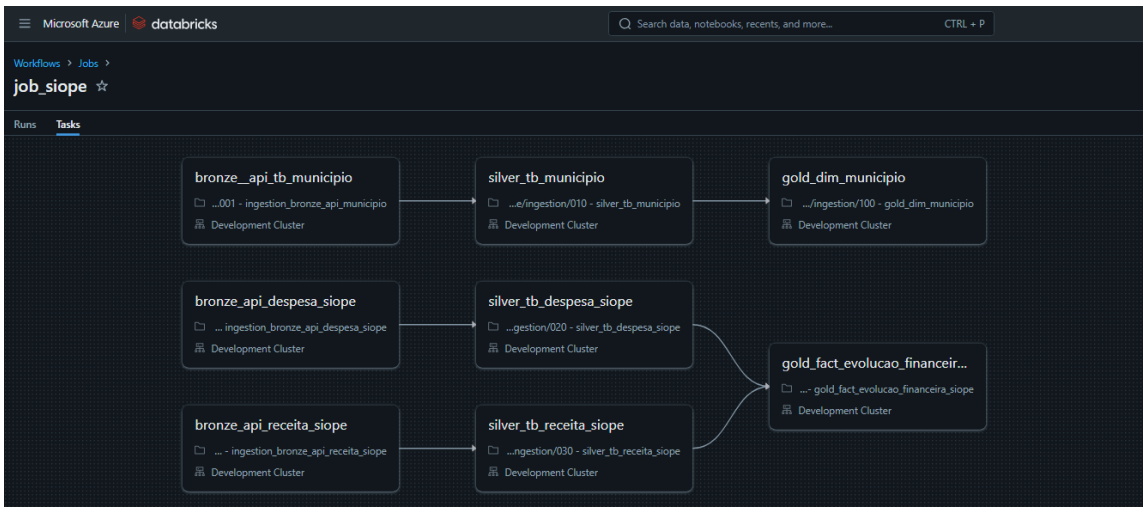

Figura 2: Workflow de orquestração do pipeline de ingestão

5. Visualização de dados

Antes de iniciar a construção do dashboard e a visualização dos dados, foi necessário estabelecer uma conexão direta entre o Power BI e as tabelas da camada Gold do Data Lake, onde os dados já estão consolidados e prontos para análise. Utilizando os recursos do Power BI, os dados armazenados em formato Parquet foram carregados, transformados e apresentados de forma interativa. A seguir, descrevo como foi realizada essa conexão e o processo de criação do dashboard para análise das receitas e despesas públicas.

5.1. Conexão dos Dados

A conexão com a camada Gold do Azure Data Lake Storage foi realizada diretamente no Power BI, utilizando o conector nativo do Azure para acessar os dados armazenados em formato Parquet. A tabela de dimensão de municípios foi extraída e carregada no Power BI através de uma query em M Language.

Essa query realiza o seguinte processo:

1. Conecta ao Data Lake na pasta na qual se encotra os arquivos da tabela e filtra os arquivos com extensão .parquet;
2. Seleciona apenas as colunas Content e Date modified;
3. Converte o conteúdo Parquet para formato tabular;
4. Expande as colunas da tabela;
5. Altera os tipos de dados das colunas para garantir consistência.

Essa etapa garante que os dados de municípios sejam carregados de forma otimizada e estejam prontos para serem utilizados no processo de criação do dashboard.

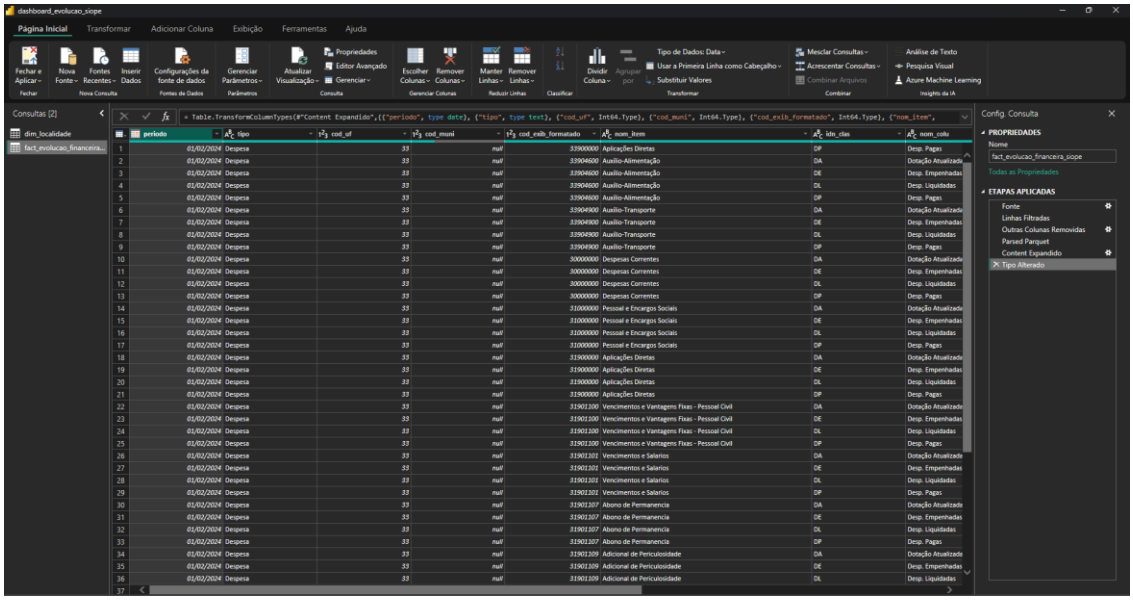

Figura 3: Tela do Power Query com as conexões da tabela na camada gold.

5.2. Relacionamento das tabelas do Modelo

Após o processo de conexão e carga dos dados no Power BI, foi necessário configurar os relacionamentos entre as tabelas para garantir a integridade da análise. Para estabelecer a conexão entre a tabela fato (fact_evolucao_financeira_siope) e a dimensão de localidade (dim_localidade), utilizei os campos cod_uf (da tabela fato) e uf_id (da tabela dimensão). O relacionamento a nível de município não foi possível, uma vez que a API utilizada não disponibilizou dados detalhados por município, mas apenas por unidade federativa (UF).

Como boa prática de modelagem para análise de dados, também criei uma dimensão temporal diretamente no Power BI. Essa dimensão temporal foi fundamental para permitir análises baseadas em períodos, facilitando a criação do relatório.

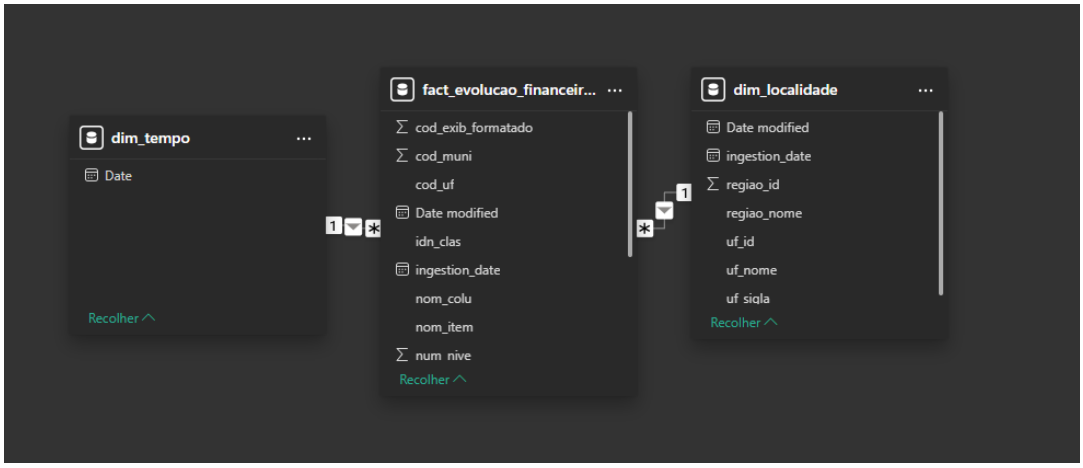

Figura 4: Desenho dos relacionamento entre as tabelas no Power BI

5.3. Dashboard

Na parte de visualização foi construído um dashboard que oferece uma visão detalhada das receitas e despesas ao longo do tempo, facilitando o monitoramento da saúde financeira e eficiência orçamentária. Ele foi projetado para permitir uma análise interativa através de filtros de período, tipo (receita ou despesa) e classificação, permitindo que os usuários ajustem os dados exibidos com base nas suas necessidades.

Na parte superior, há dois indicadores importantes: o Total Receita e o Total Despesa. Essas métricas foram calculadas utilizando a medida Valor, que é definida como a soma da coluna val_decl. Isso garante que o total exibido seja a soma correta de todas as receitas e despesas para o período selecionado.

Além disso, foi criada a medida de Eficiência para comparar receitas e despesas. A fórmula da eficiência é calculada como a razão entre as receitas e despesas:

```
eficiencia =
VAR Receita = CALCULATE(
    SUM(fact_evolucao_financeira_siope[val_decl]),
    FILTER(fact_evolucao_financeira_siope, fact_evolucao_financeira_siope[tipo] = "Receita")
)

VAR Despesa = CALCULATE(
    SUM(fact_evolucao_financeira_siope[val_decl]),
    FILTER(fact_evolucao_financeira_siope, fact_evolucao_financeira_siope[tipo] = "Despesa")
)

RETURN
DIVIDE(
    Receita,
    Despesa
)
```

Figura 5: Fórmula de cálculo de eficiência orçamentária calculada no Power BI

Essa medida permite avaliar a proporção entre os valores de receita e despesa, sendo um indicador importante de eficiência orçamentária.

No centro do dashboard, o gráfico de barras Desempenho Mensal exibe as receitas e despesas ao longo dos meses, destacando períodos de superávit ou déficit. O gráfico de Composição Orçamentária, logo abaixo, oferece uma visão clara das principais categorias orçamentárias que contribuem para aumentar ou diminuir o saldo total.

Por fim, a tabela detalhada exibe cada transação, com colunas como data, tipo, classificação e item, permitindo uma análise mais granular. Este dashboard oferece uma solução simples, porém poderosa, para visualizar e analisar dados orçamentários de forma eficiente.

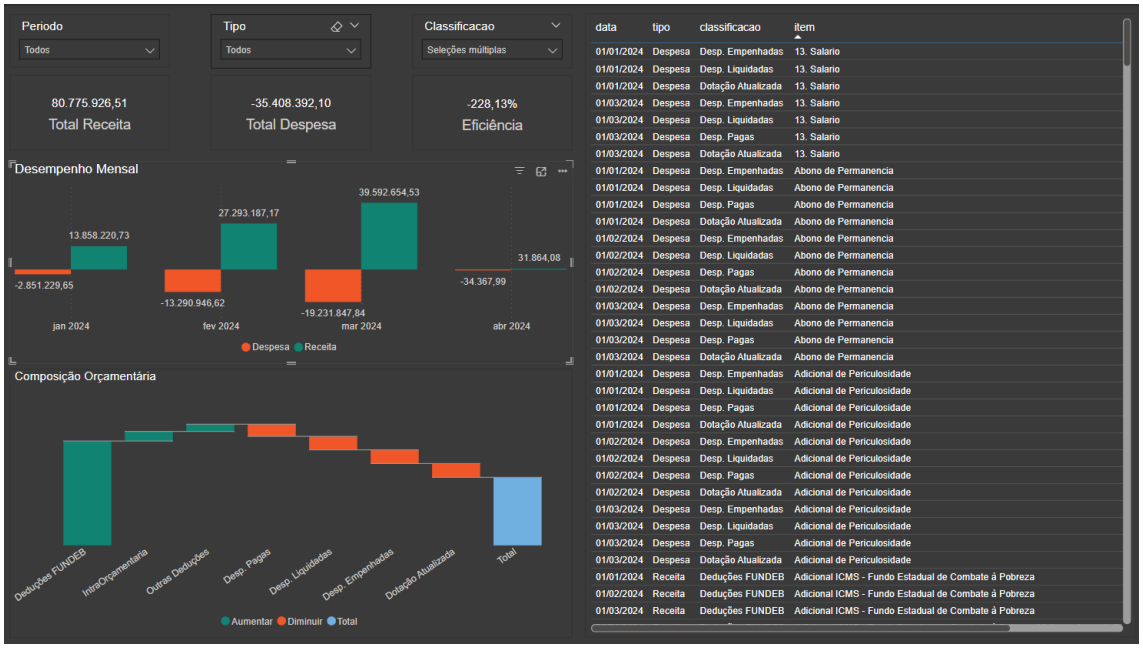

Figura 6: Tela do dashboard para analise dos dados criado no Power BI.

6. Links

Fontes de dados:

Dados aberto (SIOPE): https://dados.gov.br/dados/conjuntos-dados/sistema-de-informacoes-sobre-orcamentos-publicos-em-educacao-siope

API Receita SIOPE: https://www.fnde.gov.br/olinda-ide/servico/DADOS_ABERTOS_SIOPE/versao/v1/aplicacao#!/recursos/Receita_Siope#eyJmb3JtdWxhcmlvIjp7IiRmb3JtYXQiOiJqc29uIiwiJHRvcCI6MTAwfX0=

API Despesa SIOPE: https://www.fnde.gov.br/olinda-ide/servico/DADOS_ABERTOS_SIOPE/versao/v1/aplicacao#!/recursos/Despesas_Siope#eyJmb3JtdWxhcmlvIjp7IiRmb3JtYXQiOiJqc29uIiwiJHRvcCI6MTAwfX0=

Registro de Referência de Municípios (GovBR): https://www.gov.br/conecta/catalogo/apis/registro-referencia-municipios

API de Localidade (IBGE): https://servicodados.ibge.gov.br/api/docs/localidades#api-Municipios-municipiosGet

Repositório do TCC no GitHub: https://github.com/GHeitor/tcc_pucminas_eng_dados

Vídeo de apresentação do TCC (Youtube): https://youtu.be/e5kfAkTNVSA