Uma Arquitetura de Data Lake de Dados Ambientais e Socioeconômicos para Fomentar Pesquisa e Inovação

Aderaldo Neto[1], Mateus Neres[1], Marcus Carvalho[1]

[1]Departamento de Ciências Exatas
Universidade Federal da Paraíba (UFPB) – Rio Tinto – PB – Brasil

{aderaldo.carvalho,mateus.neres,marcuswac}@dcx.ufpb.br

Abstract. *The increasing volume of public data generated by government agencies and research institutions requires advanced storage and analysis solutions. This article proposes the implementation of a Data Lake to integrate environmental and socioeconomic data using open-source technologies. The adopted approach enables the ingestion, transformation, sharing, and visualization of large volumes of data. The developed architecture employs an Extract, Load, and Transform (ELT) pipeline to organize data in a Data Warehouse format and facilitate analysis. As a case study, a Data Warehouse model and interactive dashboards were developed for aggregated data from IBGE surveys.*

Resumo. *O crescente volume de dados públicos gerados por órgãos governamentais e instituições de pesquisa demanda soluções avançadas para armazenamento e análise. Este artigo propõe a implementação de um Data Lake para integrar dados ambientais e socioeconômicos, utilizando tecnologias de código aberto. A abordagem adotada permite a ingestão, transformação, compartilhamento e visualização de grandes volumes de dados. A arquitetura desenvolvida utiliza um pipeline de Extração, Transformação e Carga (ELT) para organizar dados no formato de Data Warehouse e facilitar análises. Como estudo de caso, foi desenvolvido um modelo de Data Warehouse e dashboards interativos para dados agregados de pesquisas do IBGE.*

1. Introdução

O volume crescente de dados públicos gerados por órgãos governamentais, instituições de pesquisa e setores privados têm impulsionado a necessidade de soluções avançadas para o armazenamento, integração e análise dessas informações. De acordo com a Open Knowledge Brasil (OKBR), a quantidade de dados disponibilizados tem aumentado exponencialmente, impulsionada pelo avanço da web e pelo uso massivo de tecnologias da informação [Ávila 2022]. Esse crescimento acelerado tem impulsionado uma demanda cada vez mais urgente por soluções de armazenamento unificado, como aponta um relatório da NetApp, empresa americana de gerenciamento de dados [NetApp 2023].

Nesse cenário, os Data Lakes surgem como uma abordagem moderna, flexível e escalável para o gerenciamento de grandes volumes de dados. Um Data Lake é um repositório centralizado capaz de armazenar dados em seu formato bruto, sejam eles estruturados, semiestruturados ou não estruturados, permitindo que sejam transformados e analisados conforme a necessidade.

A implementação de *Data Lakes* para a integração de dados públicos apresenta desafios significativos. A falta de padronização na coleta e processamento dos dados, a necessidade de soluções escaláveis para manipulação de grandes volumes de

informações e a garantia da confiabilidade dos dados armazenados são questões críticas que precisam ser abordadas. Além disso, a falta de interfaces amigáveis para diferentes perfis de usuários e de ferramentas que facilitem a análise e visualização dos dados pode limitar a interpretação e utilização dessas informações de forma útil [Medina 2024].

O presente estudo tem como principal objetivo apresentar uma arquitetura de *Data Lake* para a gestão de dados ambientais, sociais e econômicos, baseado em softwares de código aberto. Para isso, definiu-se os seguintes objetivos específicos: (i) o desenvolvimento de um modelo arquitetural para integração e compartilhamento desses dados; (ii) a implementação de um pipeline de extração, transformação e carga (ELT) para organizar os dados em um banco de dados; e (iii) a criação de ferramentas para visualização e análise, como mapas dinâmicos e dashboards interativos.

2. Trabalhos Relacionados

Esta seção apresenta alguns estudos relacionados ao uso de *Data Lakes* para dados públicos, com ênfase em informações ambientais, sociais e econômicas.

Maduro-Abreu et al. (2020) apresentam uma abordagem para consolidar dados ambientais, climáticos, epidemiológicos e de saúde pública em um Data Lake, promovendo informações de qualidade e de fácil acesso. O estudo discute a transparência da informação pública no Brasil e sua relevância para a análise das interfaces entre mudanças climáticas, mudanças produtivas e saúde. Pagotto et al. (2024) abordam a criação de data lakes para garantir a disponibilidade de dados confiáveis. No contexto da saúde pública, esses dados são destinados a pesquisadores, autoridades, organizações do terceiro setor e à população em geral, oferecendo suporte para esclarecimentos, tomada de decisões e ações estratégicas.

Rocha e Souza Júnior (2020) analisaram a relação entre a falta de energia elétrica com fatores ambientais, como queda de árvores, ventos, queimadas, poluição e erosão. Verificou-se que as interrupções no fornecimento de energia motivadas por fatores ambientais representaram uma parcela significativa das ocorrências registradas. Nesse contexto, soluções baseadas em Business Intelligence (BI) foram utilizadas para estruturar a coleta, a padronização e a visualização desses dados, permitindo uma análise das interrupções e facilitando a identificação de padrões críticos, como o impacto da vegetação na rede elétrica e a distribuição geográfica das ocorrências.

Este artigo também apresenta um *Data Lake* para integração e compartilhamento de dados, mas sua arquitetura envolve uma maior diversidade de componentes para satisfazer as necessidades de pesquisadores e gestores no acesso a diferentes formatos e níveis de granularidade dos dados.

3. Arquitetura do Data Lake

Esta seção apresenta a arquitetura do *Data Lake* proposta neste trabalho. A Figura 1 mostra a visão geral da arquitetura e as subseções seguintes descrevem os componentes.

Figura 1. Visão geral da arquitetura do Data Lake

3.2. Processo de Extração, Carregamento e Transformação dos Dados

Durante o processo de construção do Data Lake, foram tratados diferentes tipos de arquivos oriundos das bases de dados do IBGE. Os principais formatos manipulados incluem arquivos CSV, que contém dados tabulares estruturados, e arquivos Shapefile (SHP), que armazenam dados geoespaciais vetoriais, representando limites territoriais, malhas municipais, etc. Esses dados, por suas naturezas distintas, exigiram abordagens específicas para leitura, transformação e integração no ambiente unificado de dados.

O processo de integração e tratamento de dados para a construção do *Data Lake* foi conduzido seguindo o paradigma ELT (Extract, Load, Transform), uma abordagem moderna que difere do tradicional ETL (Extract, Transform, Load) ao realizar a maior parte das transformações diretamente no sistema de armazenamento de destino, aproveitando sua capacidade de processamento. Inicialmente, os dados são armazenados em uma *staging area*, que é uma área transitória para armazenar os dados extraídos, preservando sua estrutura essencial e garantindo a rastreabilidade do pipeline de dados. Essa abordagem assegurou que todas as informações relevantes estivessem disponíveis no banco de dados antes do processamento subsequente, minimizando a complexidade inicial das transformações e garantindo a integridade dos dados no ambiente de armazenamento.

Os dados extraídos são carregados em um banco de dados relacional PostgreSQL[1], com a extensão PostGIS[2] que lida com informações espaciais permitindo operações avançadas de análise e consulta geográfica. Em seguida, os dados foram organizados em um modelo dimensional, garantindo uma estrutura otimizada para análise e geração de insights [Kimball 2011]. O Apache Airflow[3] foi usado para orquestrar o fluxo de dados de forma eficiente e escalável, permitindo a automação e monitoramento dos pipelines.

[1] Disponível em: https://www.postgresql.org/
[2] Disponível em: https://postgis.net/
[3] Disponível em: https://airflow.apache.org/

O processo de extração busca automatizar a obtenção de dados em diversas fontes de origem, em diferentes formatos, para viabilizar o carregamento no banco de dados destino. No Data Lake proposto, são suportadas diferentes fontes de dados como as APIs web (ex: API de serviço de dados do IBGE[4]), arquivos de dados espaciais (ex: *shapefiles* das malhas territoriais do IBGE[5]), arquivos tabulares (ex: *CSV*, *Excel*) e repositórios de dados de pesquisas científicas (ex: Dataverse). As principais fontes de dados do Data Lake são: IBGE e IPEA (socioeconômicos); IBAMA e MapBiomas (ambientais); INMET (climáticos); e Datasus (saúde). O processo é automatizado geralmente usando scripts Python, R ou SQL, desenvolvidos como tarefas nas etapas iniciais dos *pipelines* do Airflow.

Após a extração, os dados passam por uma transformação mínima suficiente de padronização e transformação em tabelas para serem carregados no banco de dados destino no esquema da *staging area* no seu formato original. No geral, os dados brutos foram convertidos em *DataFrames* do Pandas/Python[6] para facilitar a manipulação e o carregamento no banco de dados, tendo a necessidade de ajustar alguns tipos de dados para garantir a compatibilidade e otimizar o processamento. Os dados geoespaciais exigiram um tratamento específico com a biblioteca GeoPandas/Python para armazenamento no PostGIS.

Com os dados carregados na *staging area*, inicia-se a etapa de transformação dos dados no PostgreSQL, com o objetivo de estruturar os dados em um formato analítico eficiente de Data Warehouse, como a modelagem dimensional [Kimball 2011]. Esse processo envolveu normalização e junção dos dados para consolidação do modelo dimensional em *datamarts*.

3.3. Compartilhamento de Dados Geoespaciais

Os dados armazenados no banco de dados PostGIS, tanto no formato original na *staging area* quanto no modelo dimensional no esquema do Data Warehouse, podem ser compartilhados para usuários com habilidades analíticas de consultas em bancos de dados, assim como para integração em ferramentas que possuem conexão com o BD.

Para possibilitar o acesso dos dados em Sistemas de Informação Geográfica (SIG), como o QGIS, também foi realizada a integração dos dados em servidores de compartilhamento de dados geoespaciais. Para isso, foi usado o GeoServer, uma plataforma de código aberto para compartilhar dados geoespaciais que possui integração com o PostGIS. O GeoServer oferece suporte a diversos formatos de dados e protocolos de serviços, como WMS (Web Map Service), WFS (Web Feature Service) e WCS (Web Coverage Service), facilitando a disseminação e o compartilhamento de informações geoespaciais por meio de requisições HTTP.

O GeoNetwork é uma aplicação de catalogação que complementa o GeoServer ao fornecer uma maneira estruturada de organizar e descobrir dados geoespaciais. Com o GeoNetwork, nós podemos catalogar os dados servidos pelo GeoServer, permitindo que usuários encontrem e acessem esses dados de forma mais intuitiva. Além disso, o

[4] Disponível em: https://servicodados.ibge.gov.br/api/docs/
[5] Disponível em: https://www.ibge.gov.br/geociencias/organizacao-do-territorio/malhas-territoriais.html
[6] Disponívem em: https://pandas.pydata.org/

GeoNetwork oferece uma visualização inicial em mapa, proporcionando uma visão geral dos dados geoespaciais disponíveis, sendo útil para identificar rapidamente os dados disponíveis e como eles podem ser utilizados em suas análises e projetos.

3.4. Análise e Visualização de Dados

Os dados do Data Lake foram integrados com ferramentas para exploração e visualização de dados, como plataformas de *Business Intelligence* (BI), SIGs e ambientes de programação para análises de dados.

O Metabase[7] foi adotado como ferramenta de BI para a exploração de dados, construção de *dashboards* interativos, permitindo a geração de insights a partir dos dados. O Metabase foi conectado diretamente ao PostgreSQL, permitindo a consulta e visualização dos dados armazenados. Essa integração possibilitou a exploração do Data Warehouse com consultas simples, oferecendo uma interface amigável e eficiente para a criação de *dashboards*, relatórios e visualizações.

A ferramenta *OpenLayers* foi utilizada para a criação de mapas dinâmicos em páginas web, a partir da sua integração com o GeoServer, possibilitando o acesso a camadas de dados e de mapas. A integração do GeoServer com o OpenLayers permite a visualização e manipulação de dados geoespaciais de forma dinâmica e interativa. O serviço WMS geralmente é usado para renderizar mapas, enquanto o WFS é usado para acessar dados vetoriais. No lado do cliente, configura-se o OpenLayers para consumir os serviços web fornecidos pelo GeoServer. O OpenLayers a personalização e interatividade, como a adição de controles de navegação, ferramentas de seleção de features, pop-ups informativos e estilização das camadas utilizando estilos CSS ou SLD (Styled Layer Descriptor) no GeoServer.

Para análises mais complexas, são usadas linguagens de programação como R e Python, que se conectam ao Data Lake através do servidor de banco de dados para filtragem, agregação e integração de dados. Também são usados sistemas como o QGIS e ArcGIS para análises espaciais avançadas, para usuários que são mais adeptos a essas ferramentas visuais ao invés de linguagem de programação.

4. Caso de Uso: dados agregados de pesquisas do IBGE

Para exemplificar o uso do Data Lake proposto, nesta seção é mostrado o processo de desenvolvimento para a base de dados agregados do IBGE. O processo para esta base de dados abrangeu o ciclo completo de tratamento e disponibilização dos dados, desde a ingestão no *Data Lake* até a análise, garantindo uma estrutura robusta e acessível para estudos detalhados. O foco principal foi o tratamento, integração e compartilhamento de dados de algumas pesquisas do IBGE com socioeconômicos e ambientais:

Censo Demográfico: principal fonte de dados sobre as características da população brasileira, abrangendo aspectos como educação, trabalho, renda, moradia e distribuição geográfica. O processo de ELT incluiu a extração seletiva dos dados do IBGE referente exclusivamente ao estado da Paraíba, garantindo a integridade das informações para os 223 municípios paraibanos. Os dados foram carregados no Data Warehouse, onde foram

[7] Disponível em: https://www.metabase.com/

transformados, assim possibilitando análises detalhadas sobre a evolução demográfica do Estado.

Censo Agrário: tem como objetivo fornecer um panorama detalhado sobre a estrutura fundiária, produção e uso da terra. O processo de ELT envolveu a extração seletiva dos dados do IBGE, considerando exclusivamente os estabelecimentos rurais do estado da Paraíba. Foram estruturadas informações sobre o tamanho das propriedades, tipos de culturas cultivadas, uso de tecnologias agrícolas, mão de obra empregada e acesso à infraestrutura.

Pesquisa da Pecuária Municipal: coleta dados sobre a criação de animais, permitindo uma visão detalhada da pecuária no estado da Paraíba. O processo de ELT consistiu na extração dos dados do IBGE, filtrando informações dos municípios paraibanos. Foram organizadas métricas sobre os rebanhos bovinos, suínos, caprinos, ovinos, equinos e aves, além da produção de leite, ovos e mel.

Produção Agrícola Municipal: fornece estatísticas detalhadas sobre a produção agrícola. A etapa de ELT incluiu a extração dos dados específicos do estado da Paraíba, garantindo a organização das principais culturas agrícolas. Foram estruturadas informações sobre área plantada, área colhida, quantidade produzida e valor da produção em cada município.

Produção da Extração Vegetal e Silvicultura: apresenta dados sobre a exploração de recursos florestais. Os dados foram extraídos das bases do IBGE para o estado da Paraíba, abrangendo informações sobre a extração vegetal e a produção de madeira para uso industrial e energético. A análise considerou os seguintes agregados: quantidade produzida e valor da produção na extração vegetal, por tipo de produto extrativo; quantidade produzida e valor da produção na silvicultura, por tipo de produto da silvicultura; e área total existente em 31/12 dos efetivos da silvicultura, por espécie florestal.

4.1. Processo de ELT dos Agregados do IBGE

A etapa de extração consistiu na coleta de dados disponibilizados pela API do IBGE, cujo formato primário é JSON. O processo de extração foi automatizado com scripts Python, com chamadas à API, manipulando os objetos JSON e organizando os elementos em estruturas tabulares. Foram extraídos os dados mais relevantes como: agregados, variáveis e seus respectivos valores. O escopo foram pesquisas entre 2017 e 2022 com foco nos municípios da Paraíba. Os dados espaciais foram extraídos usando a biblioteca geobr[8] do Python, que contém todos os limites geoespaciais dos municípios do Brasil em formatos como shapefiles ou GeoJSON.

A Figura 2 mostra a modelagem dimensional dos dados baseada no esquema estrela que serviu para a consolidação do ELT dos dados dos agregados do IBGE. A tabela *Fato Agregados* consolidou os valores numéricos correspondentes às combinações das dimensões, permitindo a realização de análises multidimensionais. As seguintes dimensões foram criadas:

[8] Disponível em: https://ipeagit.github.io/geobr/articles/python-intro/py-intro-to-geobr.html

Dimensão Localidade: Representando as diferentes áreas geográficas envolvidas no estudo, como municípios e unidades federativas do Brasil.

Dimensão Tempo: Capturando os períodos de coleta e análise dos censos.

Dimensão Agregado: Estruturando os diferentes grupos e categorias de dados disponibilizados pelo IBGE.

Dimensão Variáveis:Organizando os diferentes tipos de variáveis disponibilizados pelo IBGE.

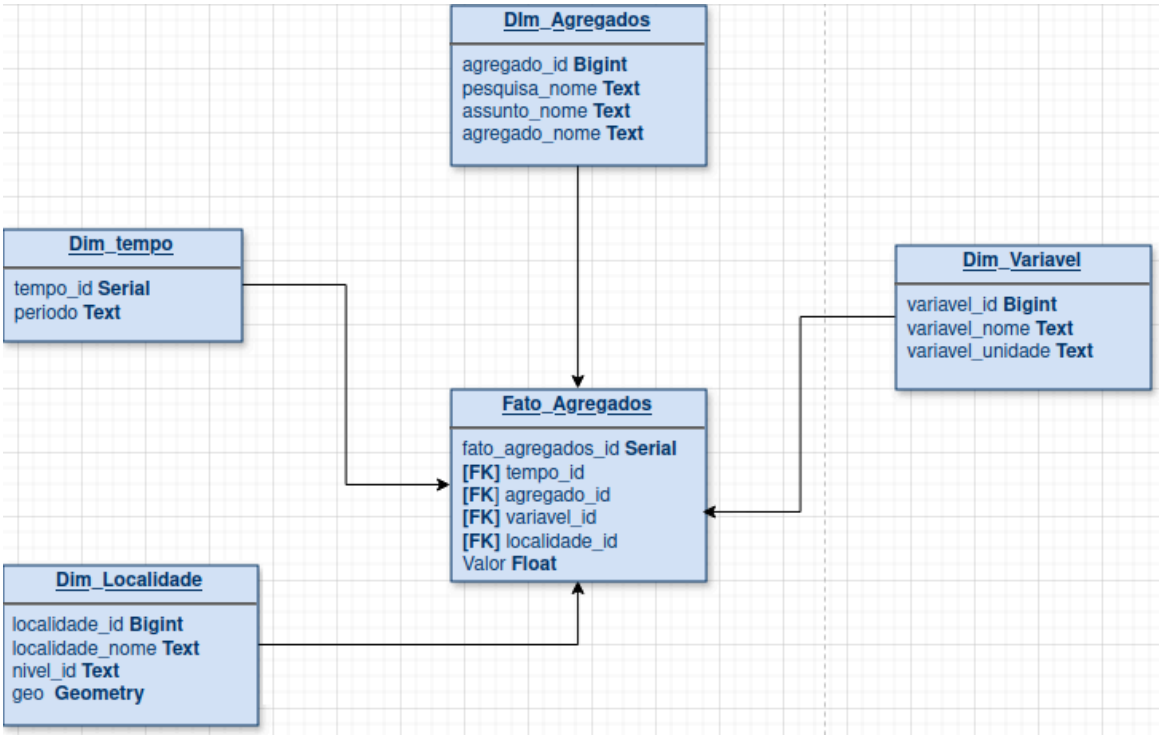

Figura 2. Esquema estrela baseado no modelo dimensional dos dados dos agregados do IBGE

4.2. Compartilhamento de Dados Geoespaciais dos Agregados do IBGE

O compartilhamento de dados geoespaciais dos agregados do IBGE foi realizado por meio do sistema de banco de dados PostGIS e do servidor de compartilhamento de dados espaciais GeoServer. No PostGIS, foram disponibilizados tanto os dados da *staging area* no formato original do IBGE, como os dados no formato de Data Warehouse descritos na seção 4.1

Através do GeoServer, foram publicadas várias camadas de dados do IBGE, incluindo as informações do Censo Demográfico, Censo Agropecuário, Pesquisa Agrícola Municipal e Pesquisa da Pecuária Municipal. Cada camada contém dados detalhados sobre diferentes aspectos demográficos, socioeconômicos e ambientais dos municípios e regiões. A exibição dessas camadas no GeoServer permite a personalização dos dados conforme as necessidades dos usuários, possibilitando a filtragem de dados usando parâmetros como código do município ou variável desejada.

4.3. Análise e Visualização de Dados dos Agregados do IBGE

Foi desenvolvido um *dashboard* no Metabase integrado ao PostGIS para proporcionar uma análise exploratória dos agregados, permitindo uma interação dinâmica com o mapa do estado da Paraíba. Também foi criada uma interface web com mapa dinâmico usando o OpenLayers integrado ao GeoServer.

4.3.1. Visualização de dados do IBGE com Metabase

A Figura 3 mostra uma tela de *dashboard* do Metabase, com dados de taxa de alfabetização nos municípios da Paraíba obtidos do Censo Demográfico 2010. O painel contém gráficos e tabelas interativas para facilitar a análise dos indicadores extraídos do IBGE. exibidos. A variação de cores no mapa facilita a interpretação dos dados, onde tons mais intensos indicam valores mais elevados para o agregado selecionado, enquanto tons mais suaves representam valores menores. O dashboard completo, com todas as informações detalhadas, está disponível em endereço público[9].

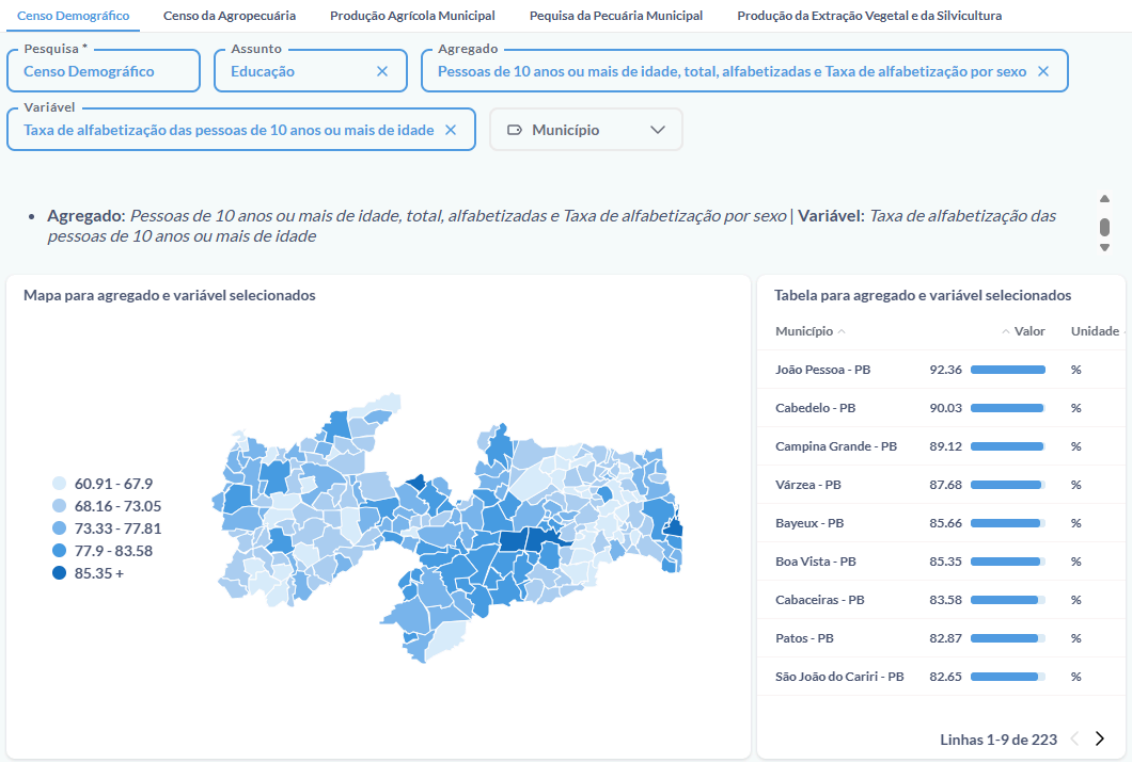

Figura 3. Dashboard mostrando a taxa de alfabetização em 2010 na Paraíba

Os principais componentes do dashboard incluem:

Mapa Interativo: Criado a partir do código do município, permitindo visualizar os dados por localidade. O mapa exibe informações detalhadas sobre agregados, variáveis e seus valores correspondentes, possibilitando a análise geográfica dos dados.

Tabela para Download: Possibilita o download de informações específicas com base nos filtros selecionados.

[9] https://ideal.ufpb.br/metabase/public/dashboard/e4f308e6-2168-46d5-97d3-97af5f59b90a

Filtros de Pesquisa: O dashboard foi configurado com filtros interativos para refinar a análise conforme os seguintes critérios: Pesquisa, Assunto, Agregados, Variáveis, Município.

4.3.2. Visualização de dados do IBGE com OpenLayers

O uso do OpenLayers integrado ao GeoServer permite que usuários, como pesquisadores, gestores públicos e estudantes, acessem, visualizem e analisem os dados dos agregados do IBGE em uma página web de forma interativa, com base em mapas dinâmicos e filtros personalizados.

A funcionalidade de visualização interativa oferecida pelo OpenLayers permite que os usuários explorem diferentes camadas de dados dos agregados do IBGE em mapas, ajustem a visualização por meio de filtros e ampliem áreas de interesse. Cada camada é configurada para exibir informações detalhadas de acordo com o agregado e a variável selecionada. Além disso, os dados podem ser baixados diretamente na interface do OpenLayers nos formatos GeoJSON, GML, CSV e KML, proporcionando flexibilidade na utilização e análise dos dados.

A Figura 4 mostra um exemplo de tela do site que usa o OpenLayers para mostrar indicadores dos agregados do IBGE para os municípios da Paraíba, podendo aplicar filtros do agregado e da variável desejada.

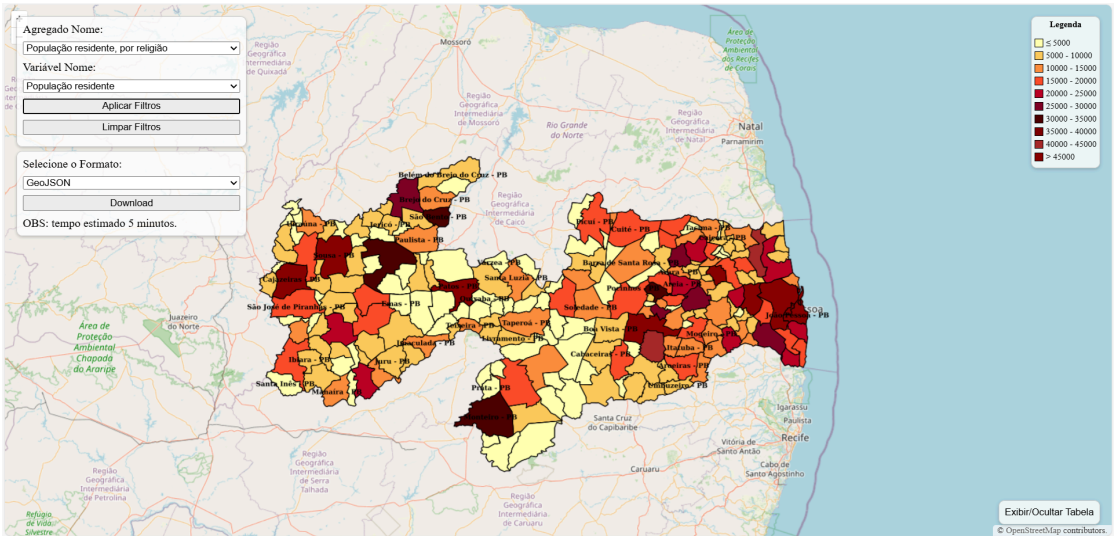

Figura 4. Mapa exibindo as camadas dos dados no OpenLayers com filtros.

5. Considerações Finais

A arquitetura proposta, baseada em soluções de código aberto, viabiliza um ambiente escalável e flexível, capaz de suportar a ingestão, transformação e disponibilização de grandes volumes de dados. A adoção de um pipeline de Extração, Transformação e Carga (ELT) permite a organização eficiente dessas informações, facilitando sua posterior utilização para análises exploratórias e a criação de visualizações interativas. Esses elementos são essenciais para fomentar pesquisas acadêmicas, subsidiar políticas públicas e incentivar a inovação no contexto da sustentabilidade. Este artigo argumenta que a arquitetura de Data Lake é estratégica para a gestão de dados públicos, podendo

contribuir não apenas para a pesquisa e inovação, mas também para a formulação de políticas sustentáveis.

Como estudo de caso, foram apresentados *dashboards* voltados para pesquisas acadêmicas, por meio da análise de dados públicos socioeconômicos, agrários e ambientais do Estado da Paraíba, obtidos de pesquisas do IBGE. No total, mais de 3.283 combinações entre agregados e variáveis foram processadas por meio de técnicas de ELT, abrangendo cinco pesquisas realizadas pelo IBGE. O dashboard está disponível e acessível para consulta, oferecendo uma ferramenta interativa e eficiente para pesquisadores, estudantes e profissionais que desejam explorar e interpretar esses dados de forma dinâmica.

Agradecimentos

Os autores agradecem ao Laboratório Misto Internacional IDEAL, formado pelo IRD e UFPB, pelo apoio institucional e infraestrutura disponibilizada para a pesquisa.

Referências

Ávila, Thiago (2017). O que faremos com os 40 trilhões de gigabytes de dados disponíveis em 2020?. Open Knowledge Brasil. Disponível em: https://ok.org.br/noticia/o-que-faremos-com-os-40-trilhoes-de-gigabytes-de-dados-disponiveis-em-2020/. Acesso em: 01 mar. 2025.

Kimball, R. and Ross, M. (2011). The data warehouse toolkit: the complete guide to dimensional modeling. John Wiley & Sons.

Maduro-Abreu, A., Litre, G., Santos, L. dos, Avila, K., Soares, D. de C., Sátiro, G. S., & Oliveira, J. E. de. (2020). Transparência da informação pública no Brasil: uma análise da acessibilidade de Big Data para o estudo das interfaces entre mudanças climáticas, mudanças produtivas e saúde. Revista Eletrônica De Comunicação, Informação & Inovação Em Saúde, 14(1). https://doi.org/10.29397/reciis.v14i1.1690

Medina, Letícia (2024). Desafios para Automação de Dados Públicos para Sociedade. *DataPolicy*. Disponível em: https://datapolicy.co/desafios-dados-publicos-sociedade/. Acesso em: 01 mar. 2025.

NetApp (2023) Cloud Complexity Report - NetApp. Disponível em: https://www.netapp.com/pdf.html?item=/media/83492-2023_cloud_complexity_report_deck.pdf. Acesso em: 01 mar. 2025.

Pagotto, D. do P., Marques, W. da S., Oliveira, D. S. de, Ferreira, V. da R. S., Nunes de Azevedo, V., & Borges Júnior, C. V. (2024). Inovação em saúde: a implementação de um data lake para armazenamento, sistematização e disponibilização de dados em saúde no Brasil. *InCID: Revista De Ciência Da Informação E Documentação*, *15*(1), e-213345. https://doi.org/10.11606/issn.2178-2075.incid.2024.213345

Rocha, M., & Souza Júnior, M. (2020). Um Dashboard para Análise de Indicadores de Continuidade relacionados à Interrupções no Fornecimento de Energia Elétrica por Causas Ambientais. In *Anais do XI Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais*, (pp. 131-140). Porto Alegre: SBC. doi:10.5753/wcama.2020.11027