Construção de um pipeline de ETL para facilitar a visualização do fluxo de caixa do dinheiro público – Para onde foi o meu dinheiro?

Beatriz da Mota Bonannol[1], Rafael Silva Ennes[1], Guilherme Pereira[1]

[1]Programa de Pós-graduação em Computação Aplicada, Universidade Presbiteriana Mackenzie
São Paulo – SP– Brazil

10294599@mackenzista.com.br, rsennes@gmail.com, gbpereira.ppgca@gmail.com

Resumo. *A complexidade dos dados disponibilizados pelo Portal da Transparência, aliada à falta de ferramentas adequadas para sua análise, dificulta o acompanhamento dos gastos públicos pela sociedade. Esta pesquisa visa suprir essa lacuna, oferecendo uma solução que facilita a compreensão e o monitoramento das informações, contribuindo para o controle social e a promoção da accountability governamental. A utilização de tecnologias modernas de banco de dados e visualização, como MongoDB, Neo4j e Grafana, permitirá explorar as relações entre os dados de forma mais eficiente e intuitiva.*

1. Introdução

A transparência na gestão pública é fundamental para o fortalecimento da democracia e o controle social. O Portal da Transparência do Governo Federal disponibiliza uma vasta quantidade de dados sobre gastos públicos, porém, o acesso e a compreensão dessas informações ainda representam um desafio para a maioria dos cidadãos. Esta pesquisa propõe democratizar o consumo desses dados, tornando-os mais acessíveis e compreensíveis, permitindo o acompanhamento efetivo do orçamento, despesas, emendas parlamentares e notas fiscais emitidas ou tomadas por órgãos públicos.

2. Objetivos

Ao idealizar essa arquitetura, a equipe tomou alguns objetivos como chaves para serem alcançados:

2.1. Objetivo Geral

Democratizar o consumo dos dados disponibilizados pelo Portal da Transparência referentes aos gastos públicos do governo federal brasileiro.

2.2. Objetivos Específicos

- Desenvolver uma plataforma que permita o acesso simplificado aos dados de orçamento, despesas, emendas parlamentares e notas fiscais.

- Implementar um sistema de atualização automática dos dados, garantindo a sua consistência e atualidade.
- Explorar as relações entre os dados utilizando análise de grafos, identificando padrões e conexões relevantes.
- Visualizar os dados de forma clara e intuitiva, utilizando dashboards interativos.
- Disponibilizar a plataforma e seus resultados de forma aberta e acessível ao público.

3. Metodologia

In some conferences, the papers are published on CD-ROM while only the abstract is published in the printed Proceedings. In this case, authors are invited to prepare two final versions of the paper. One, complete, to be published on the CD and the other, containing only the first page, with abstract and "resumo" (for papers in Portuguese).

A metodologia desta pesquisa será dividida nas seguintes etapas:

3.1. Coleta de Dados

Os dados serão extraídos do Portal da Transparência, utilizando scripts em Python para automatizar o processo de download e atualização. Os conjuntos de dados a serem utilizados incluem:

- Orçamento da Despesa
- Notas Fiscais (considerando as diferentes estruturas/schemas)
- Emendas Parlamentares
- Execução da Despesa

3.2. Armazenamento e Processamento

Os dados serão armazenados em um banco de dados MongoDB, escolhido por sua flexibilidade para lidar com diferentes schemas. Scripts Python serão utilizados para a limpeza, transformação e preparação dos dados para análise.

3.3. Análise de Dados

A análise dos dados será realizada utilizando o banco de dados de grafos Neo4j, permitindo explorar as relações entre diferentes entidades, como órgãos públicos, empresas e parlamentares. Serão investigadas conexões entre emendas parlamentares, execução da despesa e emissão de notas fiscais.

3.4. Visualização

Os resultados da análise serão visualizados através de dashboards interativos construídos com o Grafana, facilitando a compreensão e a exploração dos dados pelo público em geral.

3.5. Documentação

Todo o processo de pesquisa será documentado, incluindo os scripts utilizados, a estrutura do banco de dados e a metodologia de análise. Um catálogo explicativo do schema e dos códigos utilizados nos atributos, bem como a tabela de periodicidade de atualização dos dados, será disponibilizado.

4. Resultados

Com o intuito de democratizar ainda mais o acesso à informação de gastos com o dinheiro público, será construída uma API (*Application Programming Interface*) para extrair os dados do site do Portal de Transparência do Governo. Esta API irá disponibilizar os arquivos em formato CSV (*Comma Separated Values* – Valores separados por vírgulas) para consumo do ambiente da solução MongoDB.

Dentro do ambiente do MongoDB, os dados passarão por 3 (três) estágios: a importação de como eles vieram do site do governo; logo após ele passará por um tratamento para limpeza e acerto de headers e, usando Python, ele passará para o seu terceiro estágio, já em formato de prontidão para reporte, realizando os correlacionamentos de datas, CNPJs (Cadastro Nacional de Pessoa Jurídica) etc.

Além da utilização do MongoDB, também foi idealizado o uso do Neo4J. O Neo4J traz uma forma diferente de lidar com o dado, pois, no MongoDB o usamos extraídos diretamente do csv, quase uma estrutura relacional, já no Neo4J temos o dado em forma de grafos, podendo extrair mais informações como, por exemplo, CNPJs que mais receberam dinheiro do governo.

Para disponibilização ao público, foi pensada duas formas de realizar: a primeira com a disponibilização de uma API para que outras plataformas possam utilizar o dado já limpo e correlacionado; e, a outra forma, utilizando a ferramenta Grafana para uma interface visual para o usuário final.

A integração de todo esse pipeline pode ser vista na Imagem 1 abaixo.

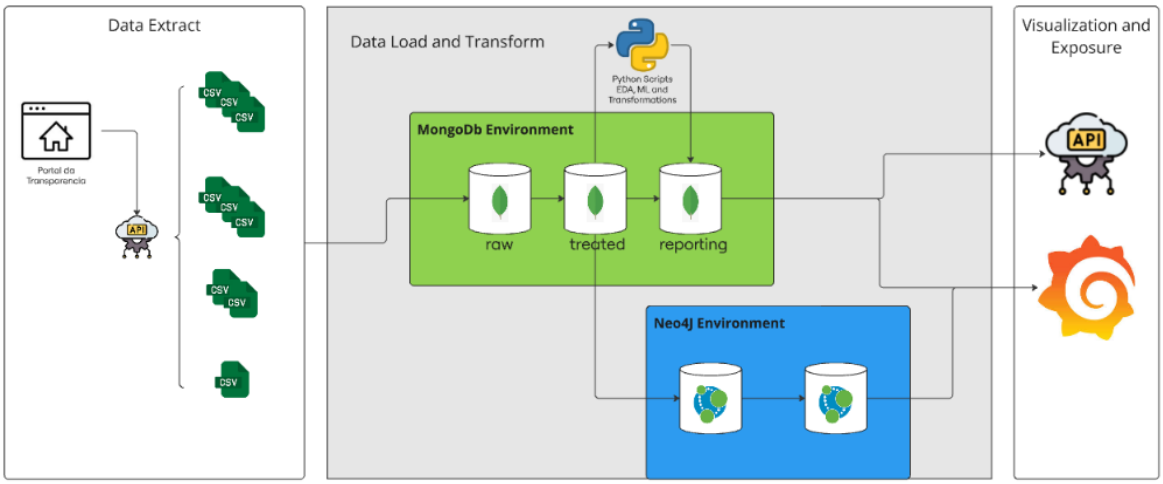

Imagem 1. Desenho da arquitetura da solução

5. Conclusão

Ao final do projeto, será possível chegar a conclusões antes impossível ao apenas consultar o portal da transparência, como o CNPJ mais beneficiado pelo governo, por exemplo. A arquitetura proposta visa disponibilizar uma melhor correlação dos dados disponibilizados pelo governo brasileiro.

References

Grafana (2024), https://grafana.com/

Neo4J (2024), https://neo4j.com/

MongoDB (2024), https://www.mongodb.com

Portal da Transparência (2024), https://portaldatransparencia.gov.br/

Reis, J. and Housley, M. (2022), Fundamentals of Data Engineering: Plan and Build Robust Data Systems, O'Reilly, 1st edition.