Arquitetura para Análise da Prevalência de Autismo nos EUA com Dados Sintéticos em Tempo Real

Magna Fernandes[1], **Bruno Schenberg**[1], **Rafael Colen**[1], **Renato Godoi**[1]

[1]Universidade Presbiteriana Mackenzie (UPM)
CEP: 01302-907 – São Paulo – SP – Brasil
magna.fernandes@mackenzista.com.br, bruno.schenberg@mackenzista.com.br,
rafael.colen@mackenzista.com.br, renato.godoi@mackenzista.com.br

Resumo. *Este artigo descreve uma arquitetura para análise da prevalência de autismo nos EUA utilizando dados da API do CDC.gov, enriquecidos com séries históricas e dados sintéticos gerados em tempo real. A arquitetura emprega Docker Compose, MongoDB, Kafka, Spark e Grafana para ingestão, processamento, análise e visualização dos dados.*

1. Introdução

O Transtorno do Espectro Autista (TEA) é uma condição neurológica que afeta a comunicação social e o comportamento. A prevalência de TEA nos EUA tem aumentado nos últimos anos, tornando crucial a análise de dados para entender as tendências e desenvolver intervenções eficazes. Este trabalho propõe uma arquitetura para coletar, enriquecer e analisar dados de prevalência de autismo, utilizando tecnologias como Docker Compose, MongoDB, Kafka, Spark e Grafana, além de simular um cenário de inclusão de dados em streaming, o que seria extremamente benéfico para sociedade, ter o dado carregado NRT (near real time).

2. Arquitetura Proposta

A arquitetura proposta é uma pipeline de dados que coleta, enriquece, transforma e visualiza dados de prevalência de autismo nos EUA, com o objetivo de explorar os dados e encontrar oportunidades para estudos futuros. Abaixo a figura que ilustra a arquitetura e a descrição de seus principais componentes:

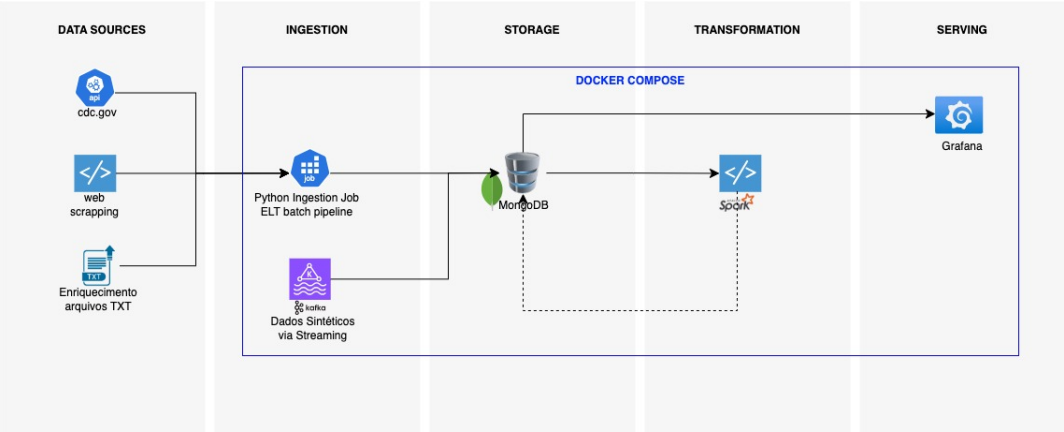

Figura 1. Arquitetura Proposta

2.1. Ingestão de Dados

- **API do cdc.gov**: Fonte de dados sobre a prevalência de autismo nos EUA.
- **Docker Compose**: Utilizado para orquestrar os serviços da arquitetura, incluindo a API, o MongoDB e o Kafka.
- **MongoDB**: Banco de dados NoSQL para armazenar os dados brutos da API e os dados enriquecidos.

2.2. Enriquecimento de Dados

- **Série Histórica**: Dados de anos anteriores obtidos da API do CDC.gov ou de outras fontes confiáveis.
- **Dados Sintéticos**: Gerados em tempo real pelo Kafka Producer para simular a inclusão de novos dados e testar a capacidade de resposta da arquitetura.
- **Kafka**: Plataforma de streaming para o fluxo de dados sintéticos.

2.3. Processamento e Análise

- **Spark**: Framework para processamento distribuído de grandes volumes de dados. Realiza a limpeza, transformação e agregação dos dados. Aplica algoritmos de análise estatística para identificar tendências e padrões na prevalência de autismo.

2.4. Visualização

- **Grafana**: Ferramenta de visualização de dados para criar dashboards interativos com gráficos e métricas sobre a prevalência de autismo.
- **API de Resultados**: Expõe os dados processados e as análises para consumo do Grafana.

3. Implementação

A implementação da arquitetura foi realizada utilizando o Docker Compose, garantindo a fácil configuração e o gerenciamento de todos os componentes da solução. Cada componente é encapsulado em uma imagem Docker, garantindo portabilidade e reprodutibilidade. O código fonte pode ser encontrado no repositório https://github.com/mack-ppgca-asd-data/Autism-Data-Project—Mackenzie-2024.git.

3.1. Imagens Docker

- **api-cdc**: Contém o código para consumir a API do CDC.gov e salvar os dados no MongoDB.
- **mongodb**: Instância do MongoDB para persistência dos dados.
- **kafka**: Instância do Kafka para o fluxo de dados sintéticos.
- **spark**: Instância do Spark para processamento e análise dos dados.
- **grafana**: Instância do Grafana para visualização dos dados.

Docker Compose: Define os serviços, as redes e os volumes da aplicação, facilitando a orquestração e o gerenciamento dos containers.

4. Geração de Dados Sintéticos

O Kafka Producer, emprega um processo contínuo utiliza a biblioteca Faker para gerar dados sintéticos em streaming, simulando a inclusão de novos casos de autismo. Esses dados são baseados em distribuições estatísticas e padrões observados nos dados reais, permitindo testar a capacidade da arquitetura de processar e analisar dados em tempo real. Esses dados são publicados em um tópico Kafka chamado ”faker-data”.

5. Processamento com Spark

O Spark processa os dados brutos da API, os dados históricos e os dados sintéticos. As etapas de processamento incluem:

- **Extração de dados**: Extrai dados sobre autismo e informações demográficas de um banco de dados MongoDB.
- **Transformação de dados**: Limpa e formata os dados, incluindo a remoção de campos desnecessários e conversão de tipos de dados.
- **Enriquecimento**: Enriquece os dados adicionando nomes de estados com base em códigos FIPS.
- **Agregação de dados**: Agrega dados por estado, incluindo população dos EUA e população equivalente do Brasil.
- **Carga de Dados**: Carrega os dados transformados de volta para uma nova coleção no MongoDB.

Em resumo, o código extrai dados brutos sobre autismo, realiza transformações para torná-los mais informativos e úteis, e então os carrega em um banco de dados para análise posterior ou visualização.

6. Visualização com Grafana

O Grafana é utilizado para criar dashboards interativos que apresentam as análises e métricas sobre a prevalência de autismo. Os dashboards permitem visualizar:

- **Tendências históricas**: Evolução da prevalência ao longo dos anos.
- **Distribuição geográfica**: Prevalência por estado ou região.
- **Dados demográficos**: Prevalência por idade e sexo.
- **Comparação com dados sintéticos**: Avaliação do impacto de novos dados na análise.

7. Conclusões

A arquitetura proposta oferece um framework flexível e escalável para análise da prevalência de autismo nos EUA. A combinação de dados reais em batch e sintéticos em NRT, junto com a utilização de ferramentas como Kafka, Spark e Grafana, possibilita estudos e análises mais abrangentes e relevantes, contribuindo para a compreensão e tratamento do autismo.

8. Trabalhos Futuros

- Integrar outras fontes de dados, como informações socioeconômicas e ambientais, para aprofundar a análise da prevalência de autismo.
- Implementar modelos de Machine Learning para previsão da prevalência e identificação de fatores de risco.
- Desenvolver interfaces interativas para facilitar a exploração dos dados e a geração de insights.

9. Referências

Referências

Cdc data. https://data.cdc.gov/, Acesso em 6 de outubro de 2024. URL https://data.cdc.gov/. Centers for Disease Control and Prevention.

Grafana. https://grafana.com/, Acesso em 6 de outubro de 2024. URL https://grafana.com/. Grafana Labs.

Apache kafka. https://kafka.apache.org/, Acesso em 6 de outubro de 2024. URL https://kafka.apache.org/. Apache Software Foundation.

Mongodb. https://www.mongodb.com/, Acesso em 6 de outubro de 2024. URL https://www.mongodb.com/. MongoDB, Inc.

Apache spark. https://spark.apache.org/, Acesso em 6 de outubro de 2024. URL https://spark.apache.org/. Apache Software Foundation.

c [d] f [a] a [p] a [r] n [o]