



Unifor

MBA em Engenharia de dados

Processamento em tempo real

- [Benício Bezerra de Abreu Carneiro](#) (Matrícula: 2419566-0)
- [Felipe Alves da Silva](#) (Matrícula: 2329032)
- [Gabriel Façanha Leal](#) (Matrícula: 2328556)
- [Marcos Andre Pires da Silva Junior](#) (Matrícula: 2419159)

Introdução

Este projeto tem como objetivo construir um pipeline de processamento de dados em tempo real (streaming) utilizando Apache Kafka e Apache Spark. A aplicação extrai dados de uma API pública, salva os dados em arquivos no formato Parquet, envia esses arquivos para o Kafka, e posteriormente um consumidor lê esses arquivos, realiza transformações para normalização dos dados e armazena o resultado novamente em Parquet.

Principais Componentes do Projeto:

- 1.API Pública:** Origem dos dados que serão processados.
- 2.Apache Kafka:** Middleware que permite a comunicação assíncrona entre os componentes do sistema (produtores e consumidores).
- 3.Apache Spark:** Framework usado para processar os dados e executar transformações complexas, além de converter formatos de dados (Parquet para Parquet ajustado).
- 4.Spark Streaming:** Extensão do Spark que permite o processamento contínuo de fluxos de dados.
- 5.Docker:** Ferramenta de containerização que garante que o ambiente de desenvolvimento seja replicável e isolado.
- 6.Zookeeper:** Utilizado para coordenação e gerenciamento do Kafka.
- 7.Streamlit:** Ferramenta para criação de aplicações web interativas que permite a visualização em tempo real dos dados processados, facilitando a análise e o monitoramento dos resultados de forma amigável e acessível.
- 8.Poetry:** Ferramenta para gerenciamento de dependências e ambientes virtuais, garantindo a consistência e a reprodutibilidade do ambiente de desenvolvimento

1. Produtor de Dados (API para Parquet via Kafka): Extrai dados de uma API pública, salva-os em formato Parquet e envia as referências dos arquivos para o Kafka.

2. Produtor de Dados com Spark Streaming: Utiliza o Spark Streaming para extrair dados da API e processá-los em tempo real, salvando os resultados em formato Parquet e enviando para o Kafka.

3. Consumidor de Dados (Parquet para Parquet ajustado): Escuta o tópico Kafka, consome os arquivos Parquet, ajusta os dados e armazena o resultado em novo Parquet.

4. Interface de Visualização (Streamlit): Uma aplicação web que se conecta ao sistema para visualizar os dados processados em tempo real, permitindo aos usuários monitorar e analisar os resultados de forma interativa e amigável.



Criação
arquivos
Parquet



Criação
arquivos
Parquet
ajustado



Criação
arquivos
Parquet
ajustado



Visualização dos dados com Streamlit



Obrigado!