



# Bootcamp: Engenharia de Dados

## Enunciado do Desafio Final

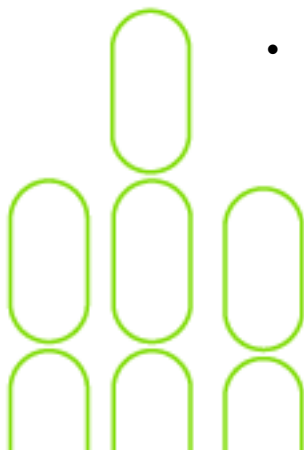
### Módulo: Desafio Final

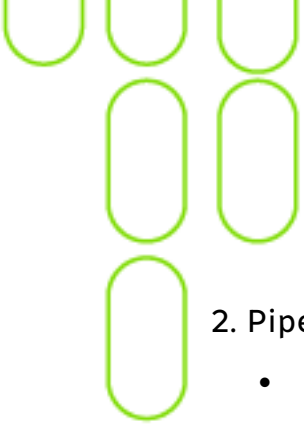
#### Objetivos de Ensino

1. Construção de Pipelines ETL com integração do Kafka com uma database (postgresql) usando kafka connect e entrega em data lake com kafka connect. Todos os serviços que compõem o kafka e o database PostgreSQL que servirá de fonte serão implantados com docker-compose.
2. Desenvolver uma solução prática de Engenharia de Dados que implemente a criação de pipelines ETL utilizando o modelo bronze, silver e gold, processados com Apache Spark SQL API e integrados a um datalake no Amazon S3 via Kafka Connect.

#### Requisitos

##### 1. Pipeline Bronze (Ingestão Bruta)

- Fonte de Dados: Arquivo JSON fornecido (com sujeiras e duplicações).
  - Ferramenta: Spark SQL para carregar os dados e criar uma tabela temporária ou persistente (formato Parquet ou Delta).
  - Processamento:
    - Carregar dados brutos para a camada Bronze, sem transformação além da validação do esquema em um banco de dados (por exemplo, PostgreSQL).
- 



## 2. Pipeline Silver (Limpeza e Transformação)

- Fonte de Dados: Tabela Bronze.
- Ferramenta: Spark SQL para limpeza e transformações.
- Processamento:
  - Remover duplicações.
  - Tratar dados ausentes (ex.: preencher valores nulos ou descartar registros inválidos).
  - Ajustar colunas para um formato consistente (ex.: normalizar nomes).
  - Salvar os dados limpos em uma tabela Silver em um banco de dados (por exemplo, PostgreSQL).


## 3. Pipeline Gold (Agregação e Enriquecimento)


- Fonte de Dados: Tabela Silver.
- Ferramenta: Spark SQL para realizar agregações e cálculos.
- Processamento:
  - Gerar métricas agregadas (ex.: número de usuários ativos, média de idade).
  - Criar a camada Gold contendo dados prontos para consumo analítico em um banco de dados (por exemplo, PostgreSQL).

## 4. Integração com Kafka Connect e Datalake no S3

- Configurar um tópico no Apache Kafka para escutar as alterações da tabela Gold.
- Utilizar Kafka Connect para transferir os dados do tópico para um diretório no Amazon S3.

## 5. Orquestração no Airflow - **Opcional**

- Configurar e orquestrar os pipelines no Apache Airflow:
- 

- 
- Pipeline Bronze: Leitura e armazenamento inicial dos dados brutos.
  - Pipeline Silver: Transformações e limpeza de dados.
  - Pipeline Gold: Agregação e preparação para consumo.
  - Garantir dependências (Bronze → Silver → Gold).

## Entregáveis

1. Screenshots que comprovem as tabelas carregadas no Postgres.
2. Screenshots ou logs que comprovem os Código Spark para os pipelines (incluindo consultas Spark SQL).
3. Screenshots ou logs que comprovem as Configuração do Kafka e Kafka Connect.
4. Screenshots ou logs que comprovem a execução dos pipelines.
5. Screenshots ou logs que comprovem os Dados processados no Amazon S3, organizados e particionados.

## Passo a passo

### 1. Pré-requisitos

- Docker
- docker-compose
- Uma conta AWS free tier

### 2. Configurar o arquivo .env\_kafka\_connect

Você deve criar um arquivo .env\_kafka\_connect para cadastrar as chaves de sua conta aws como variáveis de ambiente que serão injetadas dentro do container do kafka connect. O arquivo deve ser conforme o modelo:

```
AWS_ACCESS_KEY_ID=xxxxxxxxxxxxxxxxxxxxxx
```

