



Desafio Técnico – Engenheiro de Dados (Cloud-Ready)

⌚ Objetivo

Avaliar a capacidade de projetar, implementar e entregar um pipeline de dados moderno, utilizando boas práticas de engenharia, com foco em escalabilidade, modularidade e uso opcional de **recursos de nuvem pública** (AWS, GCP, Azure) ou **ferramentas on-premise**.

🏢 Arquitetura Esperada – Data Lake por Camadas

Seu pipeline deve seguir o modelo clássico de **Data Lake house** com as seguintes camadas:

◊ *Bronze – Ingestão*

- Fonte: APIs públicas, CSV/JSON ou bancos legados mockados
- Ferramentas sugeridas: requests, pandas, Airbyte, Kafka, etc.
- Persistência: S3 (ou equivalente) ou armazenamento local em /datalake/bronze/

◊ *Silver – Transformação e Limpeza*

- Operações de limpeza (nulos, duplicatas, normalização, tipos)
- Tecnologias: Pandas, PySpark, Dask, dbt
- Output: arquivos Parquet (preferencial) ou CSV em /datalake/silver/

🌐 *Gold – Métricas, Agregações e Enriquecimento*

- Exemplo: top produtos, receita mensal, score de clientes
- Ferramentas: Pandas, SQL, PySpark, joins com tabelas externas
- Output: /datalake/gold/ em Parquet

Persistência e Banco de Dados

- Persistência final das tabelas "gold" em um **banco relacional**
- Tecnologias: PostgreSQL (preferencial), SQLite ou MySQL
- Sugestão: utilizar SQLAlchemy, psycopg2 ou equivalente para ingestão via código

Cloud & On-Prem (Aberto)

O candidato poderá **escolher entre usar uma stack 100% on-premise ou cloud-ready**.

Exemplos de recursos cloud (opcional):

- AWS S3, Glue, Lambda, Redshift
- GCP Cloud Storage, Dataproc, BigQuery
- Azure Data Lake, Synapse, Data Factory

Exemplos de recursos on-premise:

- Airbyte, Spark local, PostgreSQL local, MinIO (simulador S3)

Ferramentas e Stack Sugerida

Tipo	Tecnologias Sugeridas
Linguagem	Python 3.8+
Processamento	Pandas, PySpark, Dask
Armazenamento	Parquet, CSV
Banco de Dados	PostgreSQL (preferido), SQLite, MySQL

Empacotamento	Docker, Dockerfile (obrigatório), Docker Compose (opcional)
Orquestração	Airflow, Dagster, Prefect, cron
Cloud (opcional)	S3, Glue, Redshift, GCS, BigQuery, Databricks, Azure Blob Storage

Avaliaremos:

- Clareza da arquitetura e modularização do pipeline
- Escolhas técnicas justificadas (cloud ou on-premise)
- Uso de boas práticas
- Código limpo
- Capacidade de entregar valor com flexibilidade tecnológica

Prazo:

Você pode realizar o desafio em até **5 dias** a partir do momento do aceite, mas a entrega antecipada é sempre bem-vinda. Não se preocupe em montar um projeto de produção — nosso foco está na clareza do seu raciocínio, estruturação do pipeline e uso consciente das ferramentas.