

Rafaella Duarte

rafaella.d.d.carvalho@gmail.com | github.com/rafaelladuarte | linkedin.com/in/rafaella-duarte-carvalho

RESUMO

Engenheira de Dados com 4+ anos de experiência em pipelines ETL/ELT nos setores jurídico e e-commerce, atuando em ambientes de alta escala (+200M registros, +20TB, +60M produtos). Certificada em Databricks Fundamentals e Apache Airflow 3, com experiência em GCP, AWS e ambientes multi-cloud. Domínio de Python, SQL, PySpark, arquitetura Medallion, ingestão via APIs, web scraping distribuído, mensageria assíncrona. Experiência prática em soluções com IA Aplicada usando RAG, embeddings e OCR para extração estruturada e classificação semântica.

HABILIDADES

Linguagens: Python, SQL, Bash, PySpark, Spark SQL

Engenharia de Dados: ETL, ELT, Data Pipelines, Data Warehouse, Data Lake, Lakehouse, Delta Lake, Data Modeling, Medallion

Banco de Dados: PostgreSQL, SQL Server, MongoDB, Elasticsearch, Redis

Cloud & DevOps: GCP, AWS, Databricks, Docker, CI/CD, Git, Linux, Grafana

Orquestração & Integração: Airflow, dbt, Kafka, RabbitMQ, REST APIs (FastAPI), Web Scraping

BI & IA Aplicada: Power BI, Looker Studio, RAG, Embeddings, OCR, MCP

EXPERIÊNCIA PROFISSIONAL

E-xyon - Engenheira de Dados

Jan 2025 – atual

- Constrói data pipelines para ingestão e processamento de dados não estruturados (OCR + LLMs) aplicados a documentos jurídicos, com parsing e persistência em MongoDB.
- Desenvolve pipeline de classificação semântica de causa raiz jurídica com arquitetura RAG, utilizando embeddings para recuperação e estruturação de dados analíticos.
- Atua em ambiente SQL Server (+20TB / 200M registros), realizando tuning de queries (CTEs, window functions), criação de procedures T-SQL, execução de jobs batch e exposição de dados via APIs REST
- Implementa pipelines ETL/ELT no GCP com arquitetura Medallion (Bronze→Silver→Gold), consolidando dados de APIs públicas e privadas, SQL Server e PostgreSQL em camadas analíticas para consumo downstream, orquestradas por automações Python.
- Apoio e orientação técnica à equipe no uso de Python, Git/GitLab e automações em VMs Linux com Docker, cronjobs, tmux e webhooks.

Linx - Engenheira de Dados

Nov 2022 – Nov 2023

- Manteve e otimizou pipelines de catálogo com ingestão contínua 24/7 de +60M produtos via Amazon S3 e processamento em BigQuery, com camada de validação de campos obrigatórios e detecção de anomalias estruturais.
- Desenvolveu arquitetura distribuída de scraping para +10 marketplaces com RabbitMQ, utilizando workers paralelos, filas com reprocessamento e validação de duplicatas e campos nulos via Elasticsearch e BigQuery.

Nodis (Grupo StoneCo) - Analista de Dados (Estágio)

Jan 2021 – Nov 2022

- Desenvolveu pipelines ETL em Python integrados a PostgreSQL, MongoDB e Elasticsearch, além da criação de dashboards no Looker Studio para suporte à tomada de decisão.

EXPERIÊNCIA FREELANCER

Flow2Go - Desenvolvedor Python

- (2026) Migração de workload on-premise para arquitetura serverless no GCP utilizando Cloud Run, Cloud Functions e Pub/Sub para execução orientada a eventos, escalabilidade automática e redução de custos operacionais.
- (2024) Desenvolveu pipelines end-to-end de ingestão de dados jurídicos com web scraping resiliente (anti-bot), persistência em MongoDB, disponibilização e sincronização dos dados via APIs REST.

Freelancer Independente - Analista de Dados

- (2025) Tratamento, sanitização, integração e enriquecimento de dados clínicos hospitalares, consolidando 10 anos de dados de múltiplas planilhas, categorizando informações para projeto de pesquisa acadêmico.

FORMAÇÃO

PUC Minas - Especialização em Engenharia de Dados 2026 – atual

UFU - Bacharelado em Ciências Econômicas 2019 – 2025

CERTIFICAÇÕES

Databricks - [Databricks Fundamentals](#) (2026)

Astronomer - [Apache Airflow 3](#) (2026)

IDIOMAS

Inglês intermediário (B1) - leitura técnica e comunicação escrita