

Ítalo Vinícius Guimarães

Engenheiro de Dados Sênior

italovinicius19981998@gmail.com · +55 (61) 98290-5991 · Brasília, Brasil · linkedin.com/in/italovinicius18 · github.com/italovinicius18

EXPERIÊNCIA

Engenheiro de Dados / IA Sênior — Nubank

nov 2025 — atual

Remoto — Brasil

- Projeta e otimiza pipelines Spark/Databricks em larga escala para datasets financeiros de alto volume, com responsabilidade pelo ciclo completo: arquitetura, implementação, tuning, qualidade, CI/CD, monitoramento e documentação.
- Mantém tabelas Delta Lake como fonte única da verdade para times de produto e risco, com schema evolution, transações ACID, padrões incrementais (merge, upsert, CDC) e governança via Unity Catalog.
- Constrói ferramentas de agentes de IA (LangGraph, ReAct, MCP) integradas aos serviços internos de plataforma, com contratos claros, estratégias de retry/fallback e observabilidade para workflows agênticos em produção.
- Implementa controles de qualidade, alertas e rastreamento de linhagem para garantir integridade e confiabilidade dos consumidores downstream em escala fintech (100M+ usuários).
- Lidera revisões de design técnico e code reviews, mentorando engenheiros em tuning de Spark, trade-offs de modelagem e padrões de design escaláveis.

Engenheiro de Dados — Compass UOL / Dasa Medicina Diagnóstica

abr 2025 — nov 2025

Remoto — Brasil

- Liderou operações de DataOps para a Dasa Medicina Diagnóstica, monitorando ingestão de dados e otimizando recursos computacionais no ambiente GCP (Cloud Composer, BigQuery, GCS, Dataproc).
- Gerenciou DAGs do Apache Airflow no Cloud Composer, garantindo execução eficiente de pipelines ETL críticos e alta disponibilidade da plataforma de dados.
- Desenvolveu e otimizou jobs PySpark no Dataproc para transformações de dados em larga escala com modelagem dimensional e otimização de queries no BigQuery.
- Provisionou infraestrutura em GKE/Kubernetes com Terraform; implementou monitoramento e alertas automatizados com Grafana, aumentando confiabilidade e reduzindo MTTR em incidentes complexos de pipelines.

Engenheiro de Dados — IoT Industrial e Monitoramento de Equipamentos — TRACTIAN

out 2024 — mai 2025

Remoto — Brasil

- Construiu pipelines de ingestão de dados em Python para dados estruturados e não estruturados gerados por dispositivos de monitoramento de equipamentos industriais (vibração, temperatura, corrente, telemetria de operação).
- Integrou MongoDB como fonte operacional principal, extraíndo telemetria heterogênea de equipamentos e dados de eventos para ClickHouse (otimizado para consultas de séries temporais de alta frequência) e PostgreSQL para analytics operacional.
- Orquestrou scripts de extração e ingestão em Python para fluxos agendados e orientados a eventos, garantindo processamento confiável e de baixa latência de leituras de sensores industriais de alta frequência em produção.

Engenheiro de Dados — Ministério da Saúde

jan 2024 — atual

Remoto — Brasília, DF, Brasil

- Arquiteta e mantém plataforma de dados distribuída integrando PostgreSQL, DuckDB e AWS (S3, Lambda) como sistema de registro para indicadores nacionais de saúde, com versionamento de schema, governança de acesso e auditabilidade.
- Construiu o pipeline do SISVAN integrando dados operacionais dos sistemas nacionais de vigilância alimentar e nutricional em uma tabela fato estruturada com rastreamento de completude e auditabilidade — aceito no SEMISH/CSBC 2026.
- Automatizou workflows de CI/CD via GitLab para artefatos de código e infraestrutura; implementou ferramentas de migração de schema e deploys multi-ambiente.

Engenheiro de Dados — A3Data

nov 2023 — out 2024

Remoto — Belo Horizonte, MG, Brasil

- Projetou arquiteturas enterprise de Data Lakehouse para clientes financeiros e de saúde com AWS (S3, Glue, Redshift, Athena) e GCP (BigQuery, GCS, Dataproc), implementando modelos dimensionais Star e Snowflake Schema.
- Orquestrou pipelines ETL com Apache Airflow e AWS Glue; implementou catálogo de dados (DataHub, OpenMetadata) e ferramentas de qualidade (Great Expectations, SODA).
- Liderou revisões de código e definiu padrões de engenharia (idempotência, testabilidade, documentação); mentorou engenheiros em sistemas distribuídos e modelagem de dados em squads multi-cliente.
- **Hypofarma:** Estruturação completa da arquitetura de dados do zero; elaborou propostas de arquitetura AWS e implementou com Terraform e GitLab CI/CD; estabeleceu cultura de governança de dados com OpenMetadata.
- **Farmax:** Liderou a migração da arquitetura de AWS EC2/EMR para Kubernetes (EKS); construiu infraestrutura completa com Terraform e GitLab CI/CD; fez deploy de SparkApplication, Prometheus, Grafana e Airflow via Terraform, Helm e ArgoCD.
- **Banco BMG:** Sustentação da engenharia de dados em arquitetura AWS; otimizou fluxos de ingestão de dados Excel com Spark; desenvolveu API REST em Django para automação de ETL; modularizou código Spark para maior confiabilidade e desacoplamento.

Engenheiro de Dados — Embraer

nov 2021 — nov 2023

Remoto — São José dos Campos, SP, Brasil

- Liderou o projeto Data Ecosystem: projetou zonas de Data Lake (raw, trusted, refined) ingerindo dados operacionais de manufatura de fontes heterogêneas (SQL Server, Oracle, SAP, bancos GCP) com PySpark, NiFi e Airflow em AWS e GCP.
- Construiu pipelines de dados operacionais ingerindo telemetria de sensores e processos de manufatura para infraestrutura de analytics em nuvem — analogia direta à integração OT-to-cloud de dados industriais.
- Construiu dashboards analíticos no Power BI para times de TI (Serviço, Qualidade, Delivery, Custos); automatizou KPIs, planilhas e sistemas de gerenciamento; administrou capacidade do Power BI.
- Aplicou práticas de DevOps para confiabilidade e monitoramento em produção; gerenciou dados relacionais e não-relacionais em stores SQL e NoSQL.

Engenheiro de Software (Estágio) — Tribunal de Contas da União (TCU)

jan 2021 — out 2021

Brasília, DF, Brasil

- Gerenciou bancos de dados Oracle e desenvolveu aplicações enterprise Oracle APEX; otimizou schemas relacionais e performance de queries para sistemas governamentais de auditoria.

Pesquisador em Visão Computacional — Universidade de Brasília

ago 2020 — jan 2021

Brasília, DF, Brasil

- Desenvolveu e pesquisou modelos preditivos para detecção de faixas de rodagem usando redes neurais convolucionais e deep learning no simulador CARLA; aplicou técnicas de mineração e análise de dados.

PESQUISA

Benchmarking de Formatos de Tabela Open Table em Arquiteturas Lakehouse com TPC-DS

out 2024 — atual

Universidade de Brasília — PPCA

- Avalia Apache Iceberg, Delta Lake e Apache Hudi com Spark sobre armazenamento S3-compatible (MinIO) usando workloads TPC-DS completos em múltiplos scale factors e formatos de arquivo (Parquet, ORC, Avro) — paper aceito no CLOSER 2026.
- Resultados cobrem estratégias de indexação, overhead de compactação, schema evolution, comportamento transacional ACID e consistência de leitura multi-engine — diretamente aplicável ao design de plataformas Lakehouse em produção.

FORMAÇÃO

Mestrado em Computação Aplicada (PPCA) — Sistemas Distribuídos e Arquiteturas Lakehouse em Nuvem

out 2024 — atual

Universidade de Brasília

CERTIFICAÇÕES

Astronomer Certification: DAG Authoring for Apache Airflow 3 — Astronomer, out 2025

Astronomer Certification for Apache Airflow 3 Fundamentals — Astronomer, jul 2025

AWS Certified Data Engineer – Associate — Amazon Web Services (AWS), mar 2025

Google Professional Data Engineer — Google, jul 2024

Astronomer Certification for Apache Airflow Fundamentals — Astronomer, jan 2024

AWS Certified Cloud Practitioner — Amazon Web Services (AWS), set 2023

STACK

Linguagens: Python, SQL, Scala, Shell scripting, YAML

Processamento de Dados: Apache Spark / PySpark, Databricks, Delta Lake, Apache Iceberg, Apache Hudi, dbt, Apache Kafka, Apache NiFi

Orquestração: Apache Airflow, AWS Glue

Cloud: GCP (BigQuery, Cloud Composer, GCS, GKE, Dataproc, Pub/Sub, Dataflow, IAM), AWS (S3, Glue, Redshift, Athena, Lambda, RDS, EMR, EC2, IAM)

Infraestrutura: Terraform, Docker, Kubernetes / GKE / EKS, Helm / ArgoCD

CI/CD: GitHub Actions, GitLab CI

Bancos de Dados: PostgreSQL, ClickHouse, MongoDB, DuckDB, DynamoDB, Redshift

Qualidade de Dados: Great Expectations, SODA

Governança: DataHub, OpenMetadata, Unity Catalog

Observabilidade: Grafana, Prometheus

IA / Agentes: LangGraph, MCP (Model Context Protocol), RAG

IDIOMAS

Português — Nativo

Inglês — Proficiência profissional (fala e escrita)

Espanhol — Básico