

Ítalo Vinícius Guimarães

Senior Data Engineer

italovinicius19981998@gmail.com · +55 (61) 98290-5991 · Brasília, Brazil · [linkedin.com/in/italovinicius18](https://www.linkedin.com/in/italovinicius18) · github.com/italovinicius18

EXPERIENCE

Senior Data / AI Engineer — Nubank

Nov 2025 — present

Remote — Brazil

- Design and optimize large-scale Spark/Databricks pipelines for high-volume financial datasets, owning the full SDLC: architecture, implementation, performance tuning, data quality, CI/CD, monitoring, and documentation.
- Maintain Delta Lake tables as the system-of-record for product and risk teams, applying best practices for schema evolution, ACID transactions, and incremental processing patterns (merge, upsert, CDC) with Unity Catalog governance.
- Build AI agent tooling (LangGraph, ReAct, MCP) integrated with internal platform services and data pipelines, designing clear service boundaries, retry/fallback strategies, and observability for production-grade agentic workflows.
- Implement data quality controls, alerting, and lineage tracking to guarantee correctness and reliability for downstream consumers at fintech scale (100M+ users).
- Lead technical design reviews and code reviews, mentoring engineers on Spark performance tuning, data modeling trade-offs, and scalable design patterns.

Data Engineer — Compass UOL / Dasa Medicina Diagnóstica

Apr 2025 — Nov 2025

Remote — Brazil

- Led DataOps operations for Dasa Medicina Diagnóstica, monitoring data ingestion and optimizing computational resources across the GCP environment (Cloud Composer, BigQuery, GCS, Dataproc).
- Managed Apache Airflow DAGs in Cloud Composer, ensuring efficient execution of critical ETL pipelines and high availability of the data platform.
- Developed and optimized PySpark jobs on Dataproc for large-scale data transformation with dimensional modeling and query optimization in BigQuery.
- Provisioned infrastructure on GKE/Kubernetes with Terraform; implemented monitoring and automated alerting with Grafana, increasing reliability and reducing MTTR on complex pipeline incidents.

Data Engineer — Industrial IoT & Equipment Monitoring — TRACTIAN

Oct 2024 — May 2025

Remote — Brazil

- Built production Python-based data ingestion pipelines for structured and unstructured sensor data generated by industrial equipment monitoring devices (vibration, temperature, current, runtime telemetry).
- Integrated MongoDB as the primary operational source, extracting heterogeneous equipment telemetry and event data and loading it into ClickHouse (optimized for high-frequency time-series queries) and PostgreSQL for operational analytics.
- Orchestrated Python extraction and ingestion scripts for scheduled and event-driven data flows, ensuring reliable, low-latency processing of high-frequency industrial sensor readings in production at scale.

Data Engineer — Ministério da Saúde

Jan 2024 — present

Remote — Brasília, DF, Brazil

- Architect and maintain a distributed data platform integrating PostgreSQL, DuckDB, and AWS services (S3, Lambda) as a system-of-record for national health indicators, with schema versioning, access governance, and auditability.
- Built the SISVAN pipeline integrating operational data from national food and nutrition surveillance systems into a structured, analytics-ready fact table with completeness tracking and audit capabilities — paper accepted at SEMISH/CSBC 2026.
- Automated CI/CD workflows via GitLab for infrastructure and code artifacts; implemented schema migration tooling and multi-environment deployments.

Data Engineer — A3Data

Nov 2023 — Oct 2024

Remote — Belo Horizonte, MG, Brazil

- Designed enterprise Data Lakehouse architectures for financial and healthcare clients using AWS (S3, Glue, Redshift, Athena) and GCP (BigQuery, GCS, Dataproc), implementing Star and Snowflake Schema dimensional models.
- Orchestrated ETL pipelines with Apache Airflow and AWS Glue; implemented data catalog (DataHub, OpenMetadata) and quality tooling (Great Expectations, SODA).
- Led code reviews and defined engineering standards (idempotency, testability, documentation); mentored engineers on distributed systems and data modeling best practices across multi-client squads.
- **Hypofarma:** Full data architecture design from scratch; authored AWS architecture proposals and implemented with Terraform and GitLab CI/CD; established data governance culture with OpenMetadata.
- **Farmax:** Led migration of data architecture from AWS EC2/EMR to Kubernetes (EKS); built full infrastructure with Terraform and GitLab CI/CD; deployed SparkApplication, Prometheus, Grafana, and Airflow via Terraform, Helm, and ArgoCD.
- **Banco BMG:** Maintained data engineering on AWS architecture; optimized Excel data ingestion flows with Spark; developed a Django REST API to automate ETL delivery; modularized Spark code for reliability and decoupling.

Data Engineer — Embraer

Nov 2021 — Nov 2023

Remote — São José dos Campos, SP, Brazil

- Led the Data Ecosystem initiative: designed raw-to-refined Data Lake zones (raw, trusted, refined) ingesting operational manufacturing data from heterogeneous sources (SQL Server, Oracle, SAP, GCP databases) using PySpark, NiFi, and Airflow across AWS and GCP.
- Built operational data pipelines ingesting sensor and process telemetry from manufacturing systems into cloud analytics infrastructure — directly analogous to OT-to-cloud industrial data integration.
- Built analytical dashboards in Power BI for IT teams (Service, Quality, Delivery, Costs); automated KPIs, spreadsheets, and management systems; administered Power BI capacity.
- Applied DevOps practices for production reliability and monitoring; managed relational and non-relational data across SQL and NoSQL stores.

Software Engineer (Intern) — Tribunal de Contas da União (TCU)

Jan 2021 — Oct 2021

Brasília, DF, Brazil

- Managed Oracle databases and developed enterprise Oracle APEX applications; optimized relational schemas and query performance for internal government audit systems.

Computer Vision Researcher — Universidade de Brasília

Aug 2020 — Jan 2021

Brasília, DF, Brazil

- Developed and researched predictive models for road lane detection using convolutional neural networks and deep learning in the CARLA autonomous driving simulator; applied data mining and analysis techniques.

RESEARCH

Benchmarking Open Table Formats in Cloud Lakehouse Architectures with TPC-DS

Oct 2024 — present

University of Brasília — PPCA

- Benchmark Apache Iceberg, Delta Lake, and Apache Hudi on Spark over S3-compatible storage (MinIO) using the full TPC-DS workload across multiple scale factors and file formats (Parquet, ORC, Avro) — paper accepted at CLOSER 2026.
- Findings cover indexing strategies, compaction overhead, schema evolution, ACID transaction behavior, and multi-engine read consistency — directly applicable to production Lakehouse platform design.

EDUCATION

M.Sc. in Applied Computing (PPCA) — Distributed Systems & Cloud Lakehouse Architecture

Oct 2024 — present

University of Brasília

B.Sc. in Software Engineering

2018 — 2023

University of Brasília

CERTIFICATIONS

Astronomer Certification: DAG Authoring for Apache Airflow 3 — Astronomer, Oct 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), Sep 2023

STACK

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

Data Processing: Apache Spark / PySpark, Databricks, Delta Lake, Apache Iceberg, Apache Hudi, dbt, Apache Kafka, Apache NiFi

Orchestration: 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)

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

CI/CD: GitHub Actions, GitLab CI

Databases: PostgreSQL, ClickHouse, MongoDB, DuckDB, DynamoDB, Redshift

Data Quality: Great Expectations, SODA

Governance: DataHub, OpenMetadata, Unity Catalog

Observability: Grafana, Prometheus

AI / Agents: LangGraph, MCP (Model Context Protocol), RAG

LANGUAGES

Portuguese — Native

English — Professional working proficiency (spoken and written)

Spanish — Basic