

GABRIEL MILANEZ CAVALHERI

Mogi Guaçu, SP, Brazil | (19) 98222-1223 | gmcavalheri@gmail.com
[linkedin.com/in/gabriel-milanez-cavalheri](https://www.linkedin.com/in/gabriel-milanez-cavalheri) | github.com/GMCavalheri

SUMMARY

Physicist (UFSCar) pursuing an MBA in Artificial Intelligence and Big Data (USP), with hands-on training in data science and data engineering. Experience building self-directed projects covering ETL pipelines, Data Lake and Data Warehouse integration, Machine Learning modeling, and Docker-based automation. Seeking a position as a Data Engineer, Data Scientist, or Data Analyst to apply technical skills to data-driven solutions.

EDUCATION

- **MBA in Artificial Intelligence and Big Data** — Universidade de São Paulo (USP) | 07/2025 – 10/2026 (in progress)
- **Bachelor's Degree in Physics** — Universidade Federal de São Carlos (UFSCar) | Completed 03/2025
- **Undergraduate Research in Applied Acoustics** — UFSCar | 2023 – 2025

DATA PROJECTS

Data Warehouse and Data Lake Integration with ETL Pipeline

<https://github.com/GMCavalheri/Complete-BI-and-ETL-with-Data-Warehouse-and-Datalake>

- Stack: Python, PostgreSQL, PyArrow, DuckDB, Polars, Docker, SQL
- Built an ETL pipeline integrating a Data Lake with a relational Data Warehouse, supporting SQL queries from basic to advanced; entire environment containerized with Docker for reproducibility by other users.

Currency Exchange Pipeline — Two Architectures: Airflow vs. AWS Serverless

<https://github.com/GMCavalheri/currency-rates-aws-serverless>

<https://github.com/GMCavalheri/Complete-ETL-Pipeline-with-Airflow>

- Stack: AWS Lambda, S3, EventBridge, Glue, Athena, Terraform (v2) · Python, Apache Airflow, PostgreSQL, Docker (v1)
- The same problem (exchange-rate API → clean, queryable data) solved with two data-architecture paradigms: a serverless, event-driven pipeline on AWS, kept entirely within the free tier (Lambda + S3 + EventBridge + Glue + Athena, infrastructure fully defined in Terraform), and a self-hosted batch pipeline with Airflow, Docker, and PostgreSQL. Built to demonstrate fluency in both managed-cloud and self-hosted orchestration paradigms.

Fraud Detection Pipeline with Apache Spark

<https://github.com/GMCavalheri/fraud-detection-spark>

- Stack: Apache Spark (distributed cluster), Spark MLib, MinIO, PostgreSQL, FastAPI, Streamlit, Docker Compose
- Distributed fraud-detection pipeline: ~1.5M synthetic transactions from 3 source systems with different schemas and data-quality issues are cleaned and enriched on a real Spark cluster (1 master + 2 workers); a GBClassifier (MLlib) is trained with a time-based split and served through a live-scoring API (FastAPI), visualized in an interactive dashboard (Streamlit). Includes a 57-test automated suite with CI.

Smart Pokédex — Web Application with Django and LLM

<https://github.com/GMCavalheri/Smart-Pokedex>

- Stack: Python, Django, PokéAPI, OpenRouter (LLM), MySQL, Redis, Docker Compose
- Django web application displaying Pokémon data from the PokéAPI, with natural-language queries answered by an LLM (via OpenRouter). Redis caching cut listing-page load time from ~18s to tens of milliseconds; backend covered by an automated pytest suite with 100% coverage.

Other Applied Projects

- **Diabetes prediction** (Random Forest) and **mechanical gear fault diagnosis** from vibration signals (Deep Neural Networks).
- **Brain tumor detection** in MRI scans (Convolutional Neural Networks).
- **Customer sentiment analysis** on product reviews and **legal entity recognition** in legal documents (NLP).
- **Movie recommendation system** (APRIORI algorithm) and **geospatial analysis** of software developer distribution.

TECHNICAL SKILLS

- **Data & Engineering:** SQL (PostgreSQL, MySQL, Spark SQL), NoSQL (MongoDB), Apache Spark, Apache Airflow, ETL, PyArrow, DuckDB, Polars
- **Languages:** Python (Pandas, NumPy, SciPy, Scikit-Learn, TensorFlow, Keras, Streamlit), R (tidyverse, dplyr, ggplot2)
- **Machine Learning & AI:** Supervised, Unsupervised, and Reinforcement Learning; Neural Networks (MLP, CNN, RNN, Transformers, Generative); NLP (NLTK, spaCy)
- **Cloud & DevOps:** AWS (Lambda, S3, EventBridge, Glue, Athena), Terraform, Docker, Kubernetes (basic), Git/GitHub
- **Analytics & Visualization:** Microsoft Excel, Power BI, Seaborn, applied statistics (inference, hypothesis testing, distributions)
- **Generative AI & Automation:** Claude, ChatGPT, Gemini, Ollama and local open-source models; n8n, Make
- **Methodology:** Agile, Kanban, Clean Code, Data Storytelling

LANGUAGES

Portuguese — Native | English — Advanced (C1 Advanced) | Spanish — Basic