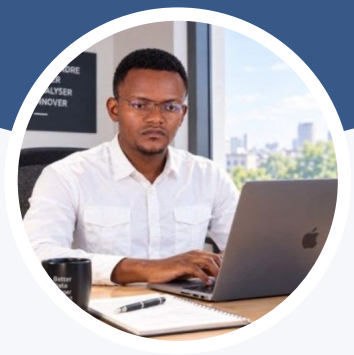
Fleury NIYOKWIZERA



Personal details

- Fleury NIYOKWIZERA
- fleuryniyokwizera2021@gmail.com
- +33 7 45 51 42 89
- 2 Rue du Maire Becquart 59160 Lille
- fleury-fcn.github.io
- linkedin.com/in/fleuryniyokwizera

Skills

Python

pandas

PySpark

Django REST

SQL

PostgreSQL

Oracle Cloud environment: GCP

Bash

ETL/ELT pipelines

database synchronization

big data and real-time data processing

data modeling

Redis/Celery task queues

PostgreSQL

Oracle NoSQL

Profile

Data Engineer (Master's in Data Modeling, Universite9 de Lille), I designed, deployed and maintained in production two open source data platforms (Halal Open Food Facts and Halal Open Beauty Facts), from the ingestion of more than 4.5 million products to the self-hosted infrastructure. I am proficient in Python, SQL/PostgreSQL, PySpark, Docker and pipeline automation, and I am now integrating AI agents (LLM, RAG) with data. I am looking for a work-study program or an internship in Data Engineering starting in November 2026.

Education

Master 1, Data Modeling **2026 - 2027**
Universite9 de Lille, France

Master 2, Applied Statistics and Decision-Making Computing **2025 - 2026**
ISTA, Universite9 du Burundi
- thesis being finalized

Master 1, Applied Statistics and Decision-Making Computing **2024 - 2025**
ISTA, Universite9 du Burundi

Bachelor's degree in Actuarial Science, specialization in Risk Analysis **2021 - 2024**
ISTA, Universite9 du Burundi

Employment

Volunteer Chief Engineer, open source platforms **Oct 2025 - Jun 2026**
Halal Open Food Facts & Halal Open Beauty Facts (halalopenfoodfacts.org, halalopenbeautyfacts.org), Lille
- Designed the architecture and developed two open source data platforms in production (Django REST, PostgreSQL, Docker, Redis/Celery), open to the public: 4,500,000 products listed, 12,000 visitors/month, 350 registered contributors.
- Migrated the infrastructure to an OVHcloud VPS and automated the import and synchronization of more than 4.5 million products: synchronization time reduced from 9 h to 55 min, i.e. -90%, and 25,000 Celery tasks processed per day.
- Secured the infrastructure (HTTPS, backups, restricted access) and set up an open source contribution process (CI, documentation, code governance): 40 pull requests merged, 99.5% availability, median API response time < 300 ms.

Assistant consultant, project monitoring and evaluation **Feb 2025 - Jun 2025**
UMWUGA N'AKAZI project (CREOPS, BBIN, WHH), Burundi
- Collected and analyzed field data with KoboToolbox; produced indicators and statistical reports for project monitoring: 1,200 households surveyed, 45 indicators tracked, 6 quarterly reports.
- Automated data cleaning and consolidation: report preparation time reduced from 3 days to 4 h.

Redis

Docker

Linux

Languages

French ●●●●●

English ●●●●●

Hobbies

- Chess
- contributing to open source projects
- reading technical books

Intern, Data Science applied to insurance **Jul 2024 - Oct 2024**
SOCAR Vie, Burundi

- Studied and prototyped the integration of AI into life insurance pricing: dataset of 50,000 contracts, gradient boosting model, prediction error reduced by 12% compared to the baseline model.
- Presented the results to the technical management: 1 presentation, recommendations adopted for the pilot phase.

Internships

Intern, Data Science applied to insurance **Jul 2024 - Oct 2024**
SOCAR Vie, Burundi

- Studied and prototyped the integration of AI into life insurance pricing: dataset of 50,000 contracts, gradient boosting model, prediction error reduced by 12% compared to the baseline model.
- Presented the results to the technical management: 1 presentation, recommendations adopted for the pilot phase.

Extracurricular activities

Conversational AI agent for web platforms to date Halal Open Food Facts and Halal Open Beauty Facts **Present**

- Design and integration of an AI agent (LLM) to assist users and make it easier to query product data: 4.5 M products queryable in natural language, response in under 3 s, 85% accuracy on a set of 200 test questions.

Big Data pipeline, NYC Yellow Taxi

- PySpark pipeline for large-scale data processing, four machine learning models and an interactive Streamlit dashboard: 40 million rides processed, processing time divided by 6 vs pandas, best model RB2 = 0.87.

Insurance claims prediction, Machine Learning

- Modular ML pipeline with hyperparameter search, cross-validation and dataset configuration: 5 models compared, 5-fold cross-validation, AUC = 0.81 for the best model.

OpenISTA, university management platform

- Flask/PostgreSQL application: management of students, teachers, courses, library and messaging.

Documents Assist, document AI application

- Conversational AI agent that lets users explore documents and query them in natural language, based on the documents saved in the application as resources. Technologies: open source language model (LLM) Mistral and pgvector vector database. Code available on GitHub.

Achievements

- Designed and developed two open source data platforms in production with 4.5 million products listed and 12,000 visitors/month.
- Reduced synchronization time of product import from 9 hours to 55 minutes (-90%) and processed 25,000 Celery tasks per day.
- Automated data cleaning and consolidation reducing report preparation time

from 3 days to 4 hours.

- Reduced prediction error by 12% in AI integration for life insurance pricing compared to baseline model.
- Developed conversational AI agent enabling natural language queries on 4.5 million products with 85% accuracy and response under 3 seconds.
- Processed 40 million rides with PySpark pipeline, reducing processing time by factor 6 and achieving RB2 = 0.87 for best model.
- Implemented modular ML pipeline with hyperparameter search and cross-validation achieving AUC = 0.81 for best model.