

Musa Rioba

Data Engineer | Author | Technical Trainer | Open Source Contributor

Mombasa, Kenya | nyakerabachi@gmail.com | github.com/musanyaks | linkedin.com/in/musarioba | musanyaks.github.io

PROFESSIONAL SUMMARY

Data Engineer with 6+ years of experience building production-grade data systems, APIs, and interactive dashboards across freelancing, technical training, and open source. Published author of "Production Data Science with R" — an 8-chapter guide on end-to-end ML pipelines, automated ETL, real-time systems, and A/B testing. Built and deployed REST APIs, multi-step workflow systems, ETL/ELT pipelines, and 3 published R packages. 4+ years as a Technical Trainer teaching data analysis, R, Python, and production best practices to 100+ students. Passionate about reproducibility, scalability, observability, and turning raw data into lasting systems.

PROFESSIONAL EXPERIENCE

Data Engineer and Freelance Consultant — Self-Employed

2019 – Present

- Freelance work via Upwork and direct clients.
- Designed and deployed production REST APIs (R Plumber, Python FastAPI, Streamlit) with authentication, scoring engines, and audit trails for insurance and healthcare clients.
- Built multi-step approval workflows with role-based access control, server-side segregation of duties, locked state transitions, and full audit logging.
- Architected ETL/ELT pipelines using Apache Airflow (CeleryExecutor), PySpark, dbt, and Snowflake, processing datasets up to 50GB on standard hardware.
- Developed and published 3 open-source R packages (ugcensus, sacensus, kenyacensus) for structured census data retrieval, cleaning, and visualization across 4-8 census years.
- Created interactive dashboards (R Shiny, Streamlit, Flutter) with real-time data feeds, responsive dark-theme UIs, and one-click CSV exports.
- Implemented data quality frameworks with runtime validators, dbt tests, and structured JSON logging for business-critical pipelines.

Technical Trainer — Freelance

2019 – Present

- Worked with 100+ students and professionals.
- Taught data analysis, R, Python, API design, and production best practices to over 100 students and professionals.
- Developed curriculum on ETL orchestration, reproducible research, dashboard development, and structured system design.
- Mentored learners on building portfolio projects, deploying to production, and writing clean, maintainable code.

Published Author — "Production Data Science with R" — Self-Published

2024 – 2026

- Authored an 8-chapter comprehensive guide covering end-to-end ML pipelines (tidymodels, XGBoost), interactive dashboards (Shiny, Plotly), automated ETL (targets), real-time fraud detection (Redis, isolation forests), and A/B testing frameworks.
- Every chapter follows production best practices: reproducibility, scalability, observability, and safety.

SELECTED PROJECTS

Loss Reserving Workbench — Python, Streamlit, SQLAlchemy

- Production actuarial application with Chain Ladder, Mack (1993), ODP Bootstrap, Bornhuetter-Ferguson, and Cape Cod methods.
- Multi-step RBAC approval workflow with locked state transitions, full audit trails, Alembic schema management, and 38 automated tests.

Telematics Kenya — UBI Risk Platform — R Plumber, Flutter, PostgreSQL, Docker

github.com/musanyaks/telematics-risk-app | telematics-risk-app.fly.dev

- Full-stack usage-based insurance platform with a Flutter mobile app and a REST API backend (10+ endpoints) for risk scoring, driver profiles, and compliance.

- Weighted ensemble model (Gini 0.41, AUC-ROC 0.73) with SHAP-like explanations. GDPR-compliant. Deployed on Fly.io.

TrialView Pro — Clinical Trial Dashboard — *R Shiny, Docker, SQLite*

musanyaks.shinyapps.io/clinical-trial-dashboard | github.com/musanyaks/clinical-trial-dashboard

- Phase II diabetes monitoring system with multi-step RBAC workflow (Investigator to Monitor to Admin), AE/SAE safety tracking, and a full SQLite audit trail.
- CDISC SDTM compliant, with CI/CD via GitHub Actions.

End-to-End Data Engineering Pipeline — *Airflow, PySpark, dbt, Snowflake*

github.com/musanyaks/data_engineering_python_project

- Production ETL/ELT platform with 3 Airflow DAGs, dual-mode Pandas/PySpark processing, dbt medallion architecture (6 marts), multi-source ingestion, and data quality gates.
- Docker Compose stack tuned for 16GB RAM.

dq-lite — *Python Package, MIT License*

github.com/musanyaks/dq-lite

- Lightweight data quality validation library — 5 dependencies, approximately 3 seconds for 10 million rows.
- Pluggable expectations, custom check registration, and native GitHub Actions CI with PR annotations.

Intent Prediction Engine (Project MERLIN) — *Python, FastAPI, ONNX*

[GitHub](https://github.com)

- Real-time customer intent prediction API with a multi-task Transformer, feature retrieval (Redis/Feast), and dynamic discounting.
- Sub-200ms p99 latency.

TECHNICAL SKILLS

Languages: R (Expert), Python, SQL, Bash, HTML, CSS, Flutter, Dart

Data Engineering: ETL/ELT, Apache Airflow, PySpark, dbt, targets, Data Modeling, Data Quality, Web Scraping

APIs and Backend: REST API Design, Plumber (R), FastAPI, Pydantic, Nginx, API Authentication

Databases: PostgreSQL, Snowflake, SQLite, Redis, BigQuery

Machine Learning and Analytics: tidymodels, XGBoost, Prophet, K-Means, RFM Analysis, Markov Chains, Isolation Forests, Risk Scoring

Dashboards and Visualization: R Shiny, Streamlit, Plotly, ggplot2, bslib, shinydashboard, DT

DevOps and Deployment: Docker, Docker Compose, GitHub Actions, Shinyapps.io, Fly.io, Git, pre-commit

Cloud and Infrastructure: Terraform, AWS (Learning), Kubernetes (Learning)

Testing: pytest, R CMD check, unit and integration testing, Spark test matrix

Experimentation: A/B Testing, Sequential Testing, Bayesian Methods, Statistical Modeling, Actuarial Methods

Writing: Technical Writing, Book Authoring, Curriculum Development, Pedagogical Content

EDUCATION

BSc Actuarial Science — Karatina University, Kenya

CERTIFICATIONS

- Packt Workshop — Data Science and Machine Learning
- Various technical workshops on ETL, API design, and production systems

OPEN SOURCE

- ugcensus — Uganda census data toolkit (R)
- sacensus — South Africa census data toolkit (R)
- kenyacensus — Kenya census data toolkit (R)
- dq-lite — Lightweight data quality validation (Python, MIT)

- intent-prediction-engine — Real-time customer intent API

LANGUAGES

- English — Fluent (professional writing, book authoring, technical training)
- Swahili — Native