DAVID ONYANGO

davidonyango193@gmail.com | (507) 317-9964 | LinkedIn | GitHub | Portfolio

M.S. Data Science & quantitative background in Economics and Mathematics, with enterprise to startup experience, from Amazon-scale ETL and BI systems to production AI infrastructure at an early-stage startup. Delivers end-to-end across the data stack: pipelines, ML models, real-time dashboards, and LLM-powered systems. Published ML researcher (2 peer-reviewed papers) shipping production systems that drive measurable operational impact.

TECHNICAL SKILLS

Languages & Tools: Python (Pandas, NumPy, SciPy, Statsmodels, Scikit-Learn, PySpark), R, SQL (PostgreSQL, Oracle PL/SQL, T-SQL), SAS, Git, Docker, CI/CD, Data Modeling

ML & AI: LangChain, LangGraph, MCP, RAG (pgvector, BM25, CRAG, cross-encoder reranking), LoRA/QLoRA, LLMOps, MLOps, Feature Engineering, Anomaly Detection, Forecasting, A/B Testing, Causal Inference, Hypothesis Testing, Bayesian Inference

Data & Cloud: AWS (Glue, S3, Redshift, CDK, Bedrock, Lambda), GCP, Azure, Apache Spark, Kafka, Hadoop, dbt, Databricks, Snowflake

Visualization & Analytics: Tableau, Power BI, QuickSight, Grafana, Plotly, Google Analytics, Looker

EXPERIENCE

Data Scientist, AI Infrastructure | **Project Path, Inc** | Boston, MA (Remote) Jan 2026 – Apr 2026

- Shipped an AI project assistant chatbot with corrective RAG (CRAG) and LLM function calling 9+ tools, enabling users to query project state, receive real-time guidance, and perform CRUD operations directly through chat.
- Engineered 10+ health signals (blocked-task rate, velocity delta, stakeholder activity) powering a transition readiness framework that replaced manual project assessments with evidence-based scoring.
- Architected a Stage Intelligence Agent for project progress using semantic task classification and signal-based heuristics across 28 stages and 4 path variants, lifting classification confidence from 0.70 to 0.90.

Business Intelligence Engineer Intern | **Amazon.com**, RME Decision Science | Bellevue, WA May 2025 – Aug 2025

- Built an end-to-end ETL pipeline (Python, SQL, AWS Glue, S3, Redshift, CDK) ingesting operational data from 97+ North American fulfillment sites and classifying technician actions via Bedrock LLM integration, replacing manual triage that could not scale to pipeline volume.
- Shipped a QuickSight dashboard tracking Resolution Rate, Actionability Rate, and True Positive Rate for robotic miss-sort alerts across 97+ ARS sites, giving operations leaders site-level visibility into alert effectiveness for the first time and moving intervention decisions from gut feel to data.
- Developed statistical anomaly detection on solar energy data, surfacing hidden cost inefficiencies and translating findings into executive-ready recommendations adopted by the team for next-step modeling.
- Pioneered Amazon Managed Grafana rollout for DS Team: configured SAML auth and migrated Advanced Refrigeration Monitoring (ARM) dashboards for Amazon Fresh sites from QuickSight SPICE scheduled refresh to Grafana near real-time observability.

Database Administrator Intern | **Northflow Solutions, Inc**. | North Mankato, MN Sep 2024 – Dec 2024

- Optimized Oracle PL/SQL queries via index tuning and schema redesign, improving report generation performance by 30% and reducing latency for downstream BI consumers.
- Built a natural language-to-SQL interface and a RAG-based knowledge retrieval tool on AWS Bedrock, cutting developer documentation search time by ~70% and reducing analyst request turnaround from hours to minutes.

Data Analyst | **Kenya Revenue Authority, Domestic Taxes Department** | Nakuru, Kenya Jan 2022 – Jul 2023

- Automated ETL processes for daily revenue reporting (Python, Excel Power Query), eliminating 3 hrs/day of manual effort for 300+ compliance officers and improving forecast accuracy.
- Shipped executive dashboards integrating 5+ data sources that became the regional reporting standard; championed data utilization as a data liaison officer for the region, contributing to the national initiative that lifted the data utilization index from 73% to 75%.

EDUCATION

- **M.S. Data Science** | Minnesota State University, Mankato | GPA: 3.83
- **B.Sc. Economics & Mathematics** | Kabarak University, Kenya

PROJECTS

Full-Stack AI Portfolio Platform | Next.js, TypeScript, Supabase, pgvector, OpenAI, Vercel | Live Site

- Shipped a production portfolio with RAG-powered chatbot (pgvector, GPT-4o-mini, SSE streaming), admin CMS with Row Level Security, and SSR via App Router, full end-to-end ownership from schema design to deployment.

Churn Intelligence Dashboard | Python, Scikit-Learn, Spark, Kafka, React, Docker | Dashboard

- Built a real-time churn prediction platform analyzing 20+ behavioral indicators through a Random Forest pipeline, streaming events via Kafka, processing in Spark, and serving predictions through a React dashboard, addressing the 5 to 25x cost gap between customer acquisition and retention via early-risk identification.

RESEARCH & PUBLICATIONS

AI/ML Research Assistant | **MINDS Lab, Minnesota State University** | Sep 2025 – Present

An Agentic System for Schema-Aware NL2SQL Generation – Accepted, CAI 2026. Multi-agent system achieving 50.87% exec accuracy on BIRD benchmark with 7B models; 90% inference cost reduction via LoRA/QLoRA. arXiv

Evaluating Credit Risk Using Interpretable ML Models – Published CADSCOM 2024; Accepted AMLDS 2025. Ensemble + LIME, 76% accuracy with regulatory transparency. CADSCOM