DAVID ONYANGO

Richmond, VA · davidonyango193@gmail.com · (507) 317-9964 · LinkedIn · GitHub · Portfolio

SKILLS

Languages: Python (Pandas, NumPy, PySpark), SQL (PostgreSQL, Oracle PL/SQL, T-SQL), TypeScript, R, SAS, Shell

Data Engineering: ETL/ELT, dbt, Airflow, Spark, Kafka, Kinesis, Redshift Streaming Ingestion, dimensional modeling, data quality, CI/CD

ML & Statistics: Anomaly Detection, Time-Series Forecasting, A/B testing, Causal & Bayesian inference, XGBoost, LightGBM, PyTorch, Scikit-Learn, LIME/SHAP, Drift Detection

LLMs & AI: RAG (pgvector, corrective RAG), LLM function calling, multi-agent orchestration (LangGraph), LoRA/QLoRA fine-tuning

Cloud & Infra: AWS (Bedrock, Glue, Lambda, S3, Athena, Redshift, QuickSight, Grafana, EC2, Step Functions, SageMaker), Docker, Linux

BI & Visualization: Advanced Excel, Databricks, Tableau, QuickSight, Power BI, Grafana, Plotly

EXPERIENCE

Data Scientist, AI Infrastructure | ProjectPath.AI | ***Boston, MA (Remote)*** Jan 2026 - Apr 2026

- Shipped an AI project assistant with corrective RAG (CRAG) and LLM function calling across 9+ tools, turning live project state into chat-driven querying, real-time guidance, and CRUD workflows instead of manual lookups.
- Built the signal pipelines feeding 10+ project health metrics (blocked-task rate, velocity delta, stakeholder activity) into an evidence-based readiness framework, replacing subjective manual reviews with auditable scoring.

Business Intelligence Engineer | Amazon, RME Decision Science Team | ***Bellevue, WA*** May 2025 - Aug 2025

- Built the iAlert Mis-sort data pipeline end-to-end on AWS (Glue, S3, Redshift, CDK), ingesting work-order data from 97+ North America Amazon Robotics sites and classifying unstructured technician comments with Bedrock LLM to surface RME-controllable mis-sort patterns across the network.
- Shipped a QuickSight dashboard tracking resolution, actionability, and true-positive rates for the mis-sort alerts across 97+ sites, replacing a monthly manual report with an automated daily refresh and giving operations leaders evidence-based visibility into program effectiveness.
- Led an Amazon Managed Grafana rollout for the DS team, moving refrigeration telemetry from a 15-minute QuickSight refresh into a near real-time, zero-staging pipeline with SAML authentication, and shipped a POC dashboard for advanced refrigeration monitoring.
- Explored solar inverter telemetry (Active Power) using Splight and authored an anomaly-model proposal aimed at surfacing hidden cost inefficiencies across fulfillment sites, feeding the team's next-quarter modeling roadmap.

Graduate Research Assistant | MINDS Lab, Minnesota State University | ***Mankato, MN*** Sep 2025 - Present

- Conduct research within a lab agenda spanning NLP, multimodal learning, and clinical AI; built the IEEE CAI 2026 multi-agent NL2SQL framework on sub-7B models running locally on consumer hardware for cost and privacy gains, while AEGIS-SQL (in prep) introduces the first differential-privacy guarantee for NL2SQL across 1,534 BIRD-dev queries.

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

- Built SQL views and optimized Oracle database queries, improving report generation performance by 30% and enabling faster BI analytics for internal stakeholders.
- Developed RAG-based chatbot integrating LLMs with enterprise documentation to reduce developer search time by 70%.

Data Analyst | Kenya Revenue Authority, Domestic Taxes Dept. | ***Nakuru, Kenya*** Jan 2022 - Jul 2023

- Automated daily revenue forecasting and ETL pipelines (Python, statistical models, Power Query), removing 3 hours per day of manual work for 300+ compliance officers, consolidating 5+ fragmented data sources into a single regional reporting standard, and lifting the regional data utilization index from 73% to 75%.

EDUCATION & CERTIFICATIONS

M.S. Data Science | Minnesota State University, Mankato | GPA 3.83 Dec 2025

B.Sc. Economics & Mathematics | Kabarak University, Kenya Dec 2021

AWS AI & ML Scholars (Agent Developer track) | PartyRock, Amazon Q, Amazon Bedrock Mar - Jun 2026

SELECTED PROJECTS

Churn Intelligence Dashboard: real-time at-risk customer detection so retention teams act before customers leave, closing the 5 - 25x acquisition-vs-retention cost gap. Link

AI Portfolio Platform: Built and deployed Onyi, a virtual assistant grounded in my projects, papers, and writing, so anyone curious (industry collaborators, academic researchers, recruiters) can ask real questions instead of guessing from a static portfolio. Link

Travelers Insurance Modeling Competition: EDA revealed Poisson-distributed call data, motivating zero-inflated regression to predict policyholder demand and prevent call center over/under-staffing.

PUBLICATIONS

[1] D. Onyango et al., "An Agentic System for Schema-Aware NL2SQL Generation," in *Proc. IEEE CAI*, 2026. arXiv

[2] D. Onyango et al., "Evaluating Credit Risk Using Interpretable Machine Learning Models," in *Proc. CADSCOM*, 2024; accepted to *AMLDS*, 2025. paper