Cliffe Oluoch Okoth

Nairobi, Kenya | +254797752778 | oluochcliffe13@gmail.com
LinkedIn | GitHub | Blog

PROFESSIONAL SUMMARY

Data Engineer with 2+ years of experience building scalable data pipelines and orchestrating workflows across fintech and agri-commodity environments.Recently, eliminated 10+ hours of manual auditing per week and improved reporting accuracy.

CORE SKILLS

- **Azure & Lakehouse:** Azure**,** Azure Databricks, Microsoft Fabric, Delta Lake, Parquet
- **ETL/ELT & Processing:** PySpark, Spark Structured Streaming, SQL, Python, Apache Airflow, dbt
- **Data Quality & Governance:** data validation, dead-letter handling, dbt tests, pytest, schema evolution
- **Modeling & Warehousing:** star schema, SCD Type 2, dimensional modeling, Snowflake, BigQuery, SQL
- **Integration & Delivery:** REST APIs, Debezium (CDC), Kafka, GitHub Actions CI/CD, Docker
- **Reporting & BI:** Power BI (DAX), dashboards

PROFESSIONAL EXPERIENCE

Data Engineer | SugarCube Africa, Nairobi **Sep 2024 – Present**

- Built Bronze, Silver and Gold Medallion layers on Delta Lake and Parquet in OneLake by writing PySpark and SQL ELT transformations that clean raw operational data and model it for Power BI.
- Eliminated 10+ hours/week of manual auditing, by enforcing data quality rules in PySpark across about 100 daily fulfillment logs and quarantining failing records before they reach reports.
- Integrated data from the company's MySQL trading platform database, price feeds and REST APIs, by running incremental Airflow loads that extract only changed rows instead of scanning full tables.
- Cut API payload size by 97% (about 200KB to 5KB) and met morning-report SLAs, by building daily Apache Airflow ETL extraction DAGs with Celery/Redis caching workers.
- Accelerated transaction clearance by 20% using CTEs and window functions to give analysts and finance teams a clear view of stalled KYC and order records.
- Flagged price swings of 15% or more in 5-minute windows by running tumbling-window anomaly detection on a Medallion lakehouse.

Data Analyst Intern | **Flamingo Commodities,** Nairobi **Jan 2024 – Aug 2024**

- Eliminated duplicate records by developing PostgreSQL MERGE scripts to securely ingest, clean, and reconcile raw market data.
- Saved 2+ hours of manual ingestion by writing automated Python and Bash file-watcher ingestion scripts.

TECHNICAL PROJECTS

Real-Time Crypto CDC & Streaming Engine **| Debezium, Kafka, Spark Structured Streaming, Cassandra, FastAPI, Grafana, Docker**

- Streamed 115,000+ market events across 3,700+ trading pairs with zero data loss by deploying Debezium Change Data Capture (CDC) on PostgreSQL Write-Ahead Logs (WAL) into Apache Kafka.
- Achieved 30-second micro-batch latency and 1-minute windowed spread aggregations by engineering Spark Structured Streaming pipelines writing to Apache Cassandra and FastAPI inside Docker Compose

Distributed Cloud ELT & Analytics Platform **| Airflow, dbt, Snowflake (v1) and BigQuery (v2), Parquet on Google Cloud Storage, pytest, GitHub Actions, Grafana**

- Guaranteed daily ingestion across rate-limited REST APIs by engineering a distributed Apache Airflow and FastAPI pipeline with dynamic rate-limiting and retry logic that bulk-loads nested JSON into Snowflake.
- Caught schema-breaking changes in production marts using dbt (SCD Type 2) coupled with automated pytest unit tests, dbt data quality checks, and GitHub Actions CI/CD workflows.

CERTIFICATIONS

- **Microsoft Certified:** Fabric Data Engineer Associate (DP-700)
- **Microsoft Certified:** Azure Databricks Data Engineer Associate (DP-750)

EDUCATION

- **Data Engineering Bootcamp**, Lux Academy
- **BSc Statistics and Programming** | Kenyatta University, Nairobi **Oct 2021 - Present**