

Simon Asaye

Senior Data & AI Engineer

✉ asayesimon123@gmail.com ☎ [+251994450995](tel:+251994450995) 📍 Addis Ababa, Ethiopia [LinkedIn](#) [GitHub](#)

Senior Data & AI Engineer who builds revenue-operations and analytics systems end to end. At Revenue Institute I unified HubSpot CRM, paid-search, and organic-search data into a single leadership dashboard with pipeline and revenue-attribution reporting, and embedded AI assistants that let non-technical teams query data in plain language. Across roles I've shipped the automations, dashboards, and AI workflows that connect marketing, sales, and data into one operating picture — turning messy manual processes into measurable systems.

Education

Hilcoe School of Computer Science

BSc in Computer Science

May 2021 – Jul 2025

Addis Ababa

Work Experience

Convex Technologies

Senior Data Engineer

Sep 2025 – Present

Addis Ababa, Ethiopia

- Built an internal AI analytics assistant that lets non-technical teams query business data in plain English — the LLM emits closed-vocabulary intent and a deterministic compiler generates governed, read-only SQL behind a 5-ring validation layer — cutting ad-hoc data requests to engineering by 60%.
- Scaled a real-time CDC pipeline (MySQL → Debezium → Kafka → ClickHouse) processing 85 data streams from 5 countries on auto-scaling Kubernetes clusters, handling billions of rows while replacing the legacy monolithic Docker stack to eliminate OOM crashes and power sub-second analytical dashboards.
- Spearheaded a MySQL-to-PostgreSQL migration program, designing target-state schemas and partitioning strategies that deliver 75% faster query performance.
- Executed phased bulk-load + CDC migration of billions of rows with zero data loss through row-count/hash reconciliation and full audit trails.
- Expanded CDC coverage from 48% to 100% (+52 percentage points, +109% relative gain) by resolving missing primary keys and fixing include-lists.
- Enabled zero-downtime online DDL on large MySQL tables using gh-ost, tuning chunk size and throttling to cut replica lag by 90%.
- Built real-time change-monitoring dashboards and cutover/validation processes that maintain 99.9% pipeline uptime with MTTR under 20 minutes.
- Transformed critical analytical dashboards by migrating from MySQL to PostgreSQL, slashing latency from 4+ hours to under 3 seconds.

AfterQuery Experts

AI Engineer (Contract)

Feb 2026 – Jun 2026

Remote

- Engineered production AI-agent evaluation tasks on the Harbor framework through an 8-stage validation pipeline (GPTZero AI-detection gating, Sonnet/Opus difficulty calibration, 15 CI checks), shipping 42 accepted tasks at a 91% first-pass acceptance rate.
- Authored multi-stage prompts and grading rubrics that calibrate task difficulty against frontier-model solve rates, holding submissions inside the target 12–75% solve band across 8-run Opus probes.
- Designed Docker-based task environments with oracle solutions and 6+ automated verifier tests each, guaranteeing deterministic pass/fail behavior across clean-room runs.
- Operated end to end as task author, reviewer, and quality-check fixer, resolving instruction/test contract mismatches to lift quality-review scores from 82 to 95+/100.
- Built an unattended submissions monitor (Playwright + GitHub Actions + Telegram) that replayed authenticated browser sessions hourly and alerted within 60 seconds of portal openings, eliminating daily manual checking.

Revenue Institute (Invictous LLC)*Data Engineer, Revenue Operations & BI (Contract)*

Mar 2026 – Apr 2026

Remote

- Built the company's revenue-operations reporting system end to end (Airbyte → BigQuery → Next.js 15 dashboard), unifying HubSpot CRM, paid-search, organic-search/SEO, invoice, and delivery-ops data into one leadership view of campaign performance, pipeline movement, and revenue attribution — 10 curated models across 223 fields, refreshed every 6 hours for 60+ users.
- Modeled layered staging → production BigQuery transformations with one-command deploy/validate scripts, cutting new-metric turnaround from days to under an hour.
- Cleaned up CRM and source-tracking data, fixing invoice grain, repeated-field, and NUMERIC-casting bugs that had been overstating campaign-sourced revenue by ~20%.
- Embedded AI-powered analytics (streaming chat, natural-language dashboard generator, AI-SQL widget) that let non-technical leadership self-serve answers, cutting executive reporting turnaround from 3 days to under an hour.
- Documented the full warehouse schema and a query-hazard guide (grain traps, date-column pitfalls, type casting), reducing dashboard defect reports by 45% and de-risking handoff.

Geneva Global Ethiopia (GGE)*Data & M&E Consultant (Contract)*

Jan 2026 – Present

Addis Ababa

- Architected an offline-first M&E data pipeline (KoboToolbox → Python ETL → star-schema model → auto-refreshing Power BI) for a Mastercard Foundation education programme, delivering live KPI dashboards across 162 schools in 6 regions.
- Modeled a fact + 3-dimension + 2-bridge star schema with a fully rebuildable ETL script, replacing ~2 days of manual refresh preparation per cycle with a zero-touch automated rebuild.
- Built region/woreda geographic drill-down visuals from Ethiopian administrative-boundary geodata, giving programme leadership their first map-level view of coverage and gaps.
- Ran exploratory analysis on raw field exports that surfaced duplicates and inconsistent school codes, lifting the usable-record rate from 88% to 99%.
- Authored a 13-section data-tooling assessment with an executive deck and 1-page cost brief; the recommended self-hosted stack was adopted, cutting projected tooling cost ~70% versus per-seat licensing.

Revelo (client: Anthropic)*AI Training / Code-Evaluation Contractor*

Oct 2025 – Feb 2026

Remote

- Produced code-evaluation training data for Anthropic, assessing GitHub pull requests and issues against a weighted rubric (complexity 30%, testability 30%, clarity, reproducibility, documentation) to curate high-signal examples for frontier-LLM training.
- Designed a multi-agent evaluation harness (orchestrator + 8 parallel worker agents) that scored and ranked 24-PR batches, cutting per-batch evaluation cycle time from ~2 days to under 3 hours.
- Standardized accept/evaluate/reject decision rules with evidence-backed scoring templates, reducing cross-batch scoring variance and rework by ~35%.
- Curated selections for iteration depth, meaningful test coverage, and Docker reproducibility, sustaining ~90% downstream acceptance of submitted evaluations.

Freelance — U.S. Clients*Data Scientist (Contract)*

Mar 2023 – Jan 2026

Remote

- Delivered production predictive-ML models and analytics for U.S.-based clients as a remote contractor, improving forecast accuracy by 18% through feature engineering and systematic model tuning.
- Built client-facing Power BI dashboards and translated findings for non-technical stakeholders across time zones, cutting recurring-report turnaround by 40%.
- Automated data-cleaning and validation pipelines in Python, eliminating ~10 hours/week of manual preparation while raising dataset reliability.
- Managed full engagement lifecycles asynchronously — scoping, delivery, iteration — sustaining repeat contracts across 3 years.

Shega Media and Technology

Data Engineer

Mar 2025 – Sep 2025

Addis Ababa

- Architected end-to-end ETL pipelines with Azure Data Factory and Databricks, ingesting and transforming 100M+ records into a Delta Lake repository.
- Designed and enforced Medallion Architecture on Delta Lake, automating processes that cut manual QA effort by 50% and accelerated data readiness.
- Secured Azure Blob Storage access using Managed Identity and SAS tokens with least-privilege principles, reducing potential security incidents by 65%.
- Implemented Unity Catalog for unified metadata and lineage tracking, boosting data discoverability for business teams by 90%.
- Automated workflow orchestration in Azure Data Factory with retry logic and alerts, achieving 99.9% pipeline uptime and eliminating manual intervention.
- Optimized compute costs by auto-scaling Databricks clusters and leveraging spot instances, delivering 30% lower spend with faster pipeline performance.
- Integrated Azure datasets into Power BI, cutting report development time by 40% and strengthening data-driven decision-making.

CodePoint Creatives

Data Scientist

Jun 2022 – Jan 2025

Addis Ababa

- Pioneered and deployed production ML models for predictive analytics, improving prediction accuracy by 15% through advanced feature engineering and model tuning.
- Streamlined ETL pipelines on Azure, reducing data processing latency by 30% while enhancing data quality and accessibility at scale.
- Performed exploratory data analysis that uncovered key trends, delivering insights to C-suite executives that identified 3 new market opportunities.
- Orchestrated scalable data warehousing solutions and models on Azure, supporting 500+ users with complex reporting and role-based access controls.
- Established enterprise data governance and validation frameworks, lifting data quality by 20% and creating high-impact Power BI dashboards that drove a 25% increase in KPI target achievement.

PwC Switzerland

Power BI Job Simulation — Forage (Virtual Internship)

Sep 2024

Remote

- Completed PwC Switzerland's Power BI job simulation, strengthening client-facing data-visualization skills against real engagement scenarios.
- Created Power BI dashboards that conveyed client KPIs clearly, responding to engagement requests with well-designed, decision-ready solutions.
- Communicated insights and actionable recommendations to engagement partners through concise, analysis-backed summaries.
- Analyzed HR data on gender-related KPIs and identified root causes of gender-balance gaps at executive-management level, demonstrating data-driven problem-solving.

Projects

- **Agentic Prospect & Account Research Pipeline** · github.com/justsima/agentic-stack — Multi-agent research system (orchestrator + parallel explorers + adversarial verification) that compiles multi-source company/people briefs in ~2 minutes versus ~30 manually — the same pattern behind AI-assisted pre-call prospect research and account targeting.
- **Event-Driven Engagement Alerting Bot** · github.com/justsima/afterquery-monitor — Headless-browser watcher on GitHub Actions that replays authenticated sessions hourly and fires Telegram alerts within 60 seconds of portal state changes; ran unattended for 6+ weeks — the trigger→alert architecture behind "notify sales when a high-value prospect engages."
- **Kingdom Give — Intake-to-Task Automation** · github.com/justsima/Reciept-scanner — End-to-end workflow turning raw form/Telegram submissions into validated, routed review tasks: AI extraction → 11-rule validation engine

→ admin dashboard, replacing ~15 hrs/week of manual entry at 94% extraction accuracy across 9 document formats. FastAPI · React · Celery · PostgreSQL · Redis.

- **Real-Time CDC Reporting Pipeline** · github.com/justsima/clickhouse — Production-documented MySQL → Debezium → Redpanda → ClickHouse pipeline feeding auto-refreshing Power BI dashboards — the data backbone for always-current campaign, activity, and pipeline reporting.
- **School Access Index — Record Matching & Enrichment** — Python pipeline reconciling ~40,000 messy records to canonical entities via a 4-pass cascading fuzzy matcher (80% exact-match coverage) — the same discipline as CRM dedup, source-tracking cleanup, and contact enrichment.
- **RoadReady — Driving-Test Platform** · github.com/justsima/roadready — Bilingual (EN/AM) exam platform (Next.js + PostgreSQL) with auth, payment flow, and anti-cheating system; stakeholder-demoed for national Ministry adoption — conversion-critical forms, payments, and funnel UX in production shape.
- **ProjectPro Knowledge Base** · github.com/justsima/projectpro-catalogue — Searchable catalog of 304 end-to-end data/AI projects (scrape → parse → static SPA with instant client-side search) — content-ops tooling that turns a large corpus into a navigable internal resource.

Volunteering

DevMeetup.et

Volunteer Organizer

2023 – Present

Addis Ababa

- Co-organize Ethiopia's largest AI/developer meetup series as a volunteer, growing attendance 10× (from ~100 to 1,000+) across three flagship events — V3 drew 1,000+ attendees, 8 speakers, and 10+ sponsors — coordinating speakers, sponsorships, and venue logistics.

Certificates

- [IBM Data Science Professional Certificate](#) 
- [365 Data Science — Data Scientist Career Track](#) 
- [Microsoft Power Platform Fundamentals \(PL-900\)](#) 

Skills

Python | SQL | BigQuery | PostgreSQL | ClickHouse | Airbyte | ETL & Data Modeling | Power BI | Looker Studio | Excel / Google Sheets | HubSpot | Claude Code | Cursor | AI Agents (RAG, MCP) | Prompt Engineering | Zapier | Make | n8n | GitHub Actions | API Integrations | GA4 / Website Analytics | SEO (Ahrefs, SEMrush) | Campaign & Pipeline Reporting