

Thomas Duffy

Senior Data Engineer | Founding engineer of Polly's analytics platform | Data Platform & Infrastructure
Denver, CO | 703-786-5717 | tommy.duffy@gmail.com
linkedin.com/in/thomas-a-duffy | github.com/ThomasADuffy

SUMMARY

Senior Data Engineer with 4+ years building and owning production analytics infrastructure at scale, and 7 years across data engineering and data science. Founding engineer on Polly's analytics platform and owner of its repository, accountable for codebase architecture and the release process from greenfield build through production ownership and incident response. Depth in Databricks and Spark platform engineering, infrastructure as code, change data capture, and data warehousing, with production AI agents shipped to customer-facing analytics.

TECHNICAL SKILLS

Data Platform: Python, SQL, PySpark, Apache Spark, Databricks, Delta Lake, Delta Live Tables, Delta Sharing, Unity Catalog, dbt, Amazon Redshift, Amazon S3, AWS Glue, Amazon Web Services (AWS), Estuary Flow, Change Data Capture (CDC), Web Scraping, Medallion Architecture, Data Warehousing, ETL

Infrastructure: Terraform, Infrastructure as Code (IaC), Cloud Provisioning, CI/CD, Release Management, SOC 2 Compliant Data Handling, Personally Identifiable Information (PII) Governance

Databases: PostgreSQL, MongoDB, Amazon Redshift

AI & LLM Engineering: Production AI Agents, Agentic Systems, Databricks Foundation Models, Anthropic Claude, Prompt Engineering, LLM Evaluation, Retrieval and Context Enrichment, Claude Code, Knowledge Graphs

Software Engineering: Python, Pydantic, REST APIs, SDK Development, Git, GitFlow, GitHub, Automated Testing, Full-Stack Development, Jira

Machine Learning: MLOps, Model Monitoring, Feature Engineering, Forecasting, Natural Language Processing (NLP), Sentiment Analysis, scikit-learn, Neural Networks

PROFESSIONAL EXPERIENCE

Polly (PollyEx, Inc.) — Remote

One of two founding engineers for Polly's Analytics Platform and Product, building it from the ground up. Owner of the analytics repository and its largest contributor (850+ commits), accountable for codebase architecture and the release process.

Senior Data Engineer | November 2023 – Present

- Built the team's CI/CD and infrastructure-as-code deploy path in-house, replacing legacy dbx with a Python SDK-based tool driven by typed Pydantic job definitions — idempotent upserts, dependency-ordered rollout, automated test gates, and drift detection across staging and production on a governed release cadence for ~80 jobs and pipelines.
- Owned the ingestion layer across three generations — hand-built Python pipelines, then AWS Glue, then Estuary Flow write-ahead-log Change Data Capture (CDC) — progressively cutting replication latency and eliminating a class of custom ingestion maintenance.
- Designed a client-facing analytics agent that lets lending clients query the analytics platform conversationally, translating natural-language questions into governed data access.
- Co-built a production LLM classification system with the data science team, mapping lender loan statuses to standardized analytics milestones through a three-tier pipeline — deterministic lookup, cross-client SQL consensus, then multi-vote inference on Databricks Foundation Models — reserving paid model calls for genuinely novel cases.
- Contributed its reliability and auditability layer: five-sample multi-vote inference with majority resolution, per-prediction confidence scoring, automatic human-review flagging below threshold, and an append-only Unity Catalog ledger so no prediction is silently overwritten.
- Applied prompt engineering to domain-specific terminology, enriching each classification with the lender's own preceding and following loan-status sequence via SQL window functions (LAG/LEAD) to disambiguate industry acronyms and correct general-English model bias.
- Created a machine-navigable pipeline knowledge graph (lineage, blast-radius, field-level mapping) powering AI-assisted impact analysis and onboarding across ingestion, Delta Live Tables, and dbt.

- Built the team's agentic developer tooling — 5 custom Claude Code skills and 3 specialized agents covering lineage, guided deployment, drift reconciliation, safe deprecation, and dependency modernization — making the repository navigable by AI coding agents and turning tribal knowledge into self-service automation.
- Own the release process end to end and set the engineering standards the team builds against — testing, deployment conventions, deprecation, and documentation; mentor contributors on platform architecture and deploy tooling, and drive delivery cadence through recurring milestone updates to Data & Analytics stakeholders.

Data Engineer | July 2022 – November 2023

- Architected and built the analytics platform on Databricks and AWS from greenfield — architecture, infrastructure, ingestion, and a Bronze/Silver/Gold medallion transformation layer using Spark, PySpark, Delta Lake, and Delta Live Tables — establishing Polly's analytics capability from nothing.
- Built ingestion from MongoDB, PostgreSQL, Amazon S3, external SFTP feeds, and web-scraped public sources into a governed Unity Catalog Silver layer, with dbt-materialized Gold models on Amazon Redshift.
- Built the client reporting and data delivery system across three channels — scheduled CSV/PDF generation pushed to client SFTP endpoints, Polly-hosted SFTP for clients to pull from, and Delta Sharing for live low-latency access to Gold tables with no file movement.
- Partnered with product, engineering, and client-facing teams to translate reporting requirements into pipeline changes and coordinated data-contract changes with upstream source-system owners; established the team's engineering workflow across Jira, GitHub (GitFlow), and Slack with automated job-health alerting.

DataPrime — Data Scientist

October 2021 – July 2022

- Provisioned and deployed client-facing cloud infrastructure using Terraform, owning the environments that served delivered models end to end.
- Built forecasting models projecting freight shipping costs and logistics for client operations.
- Delivered against healthcare and personally identifiable information (PII) data under SOC 2 compliance, designing a compliant data-handling system to work within client access restrictions.

TallyScore — Head of Data Science

April 2020 – June 2021 | Denver, CO

- Led data science at a six-person bootstrapped startup as one of two engineers, working full stack across data pipelines, backend, and frontend.
- Built a public-sentiment analysis application that scraped public data sources and applied natural language processing to surface what customers were saying about a given company and how they were saying it.
- Designed and shipped multiple proofs of concept and MVPs demoed to prospective investors, taking each from problem framing to working product.
- Owned all ML and AI model development and the cloud-integrated pipelines supporting it, and selected the team's data science tooling.

Galvanize Inc — Resident Data Scientist

December 2019 – March 2020 | Denver, CO

- Delivered lectures on Python, machine learning, and neural networks for the Data Science Immersive program, and coached students through technical assignments and capstone projects.

Virtuoso Sourcing Group, LLC — Finance and Accounting Manager

July 2016 – February 2019 | Denver, CO

- Managed firm finances and forecasting, reporting to the CEO and COO. Used analytics to evaluate contract profitability and automated the firm's invoicing, accounting, and HR processes before transitioning into data engineering.

EDUCATION

Galvanize Inc — Data Science Immersive, 2019

University of Colorado Boulder — Bachelor's Degree, Economics (Business minor), 2016