

SAMEER MIR | Data Engineer

Milpitas, CA | +1 (408) 755-6189 | sameermir2195@gmail.com | [LinkedIn](#) | [Github](#) | [Portfolio](#)

PROFESSIONAL SUMMARY

Data Engineer with 3+ years building and owning ETL/ELT pipelines, cloud data warehouses, and lakehouse platforms across Snowflake, Databricks, and PySpark, spanning 20+ source systems. Delivers governed data models, automated quality and lineage frameworks, and decision-ready KPIs for stakeholder teams. Extends core data engineering into GenAI/RAG platforms — improving retrieval accuracy by 38%, cutting inference latency by 27%, and accelerating time-to-production by 80%. Partners cross-functionally on data contracts, mentors engineers on performance and testing standards, and ties infrastructure work directly to adoption, cost, and reliability outcomes.

TECHNICAL SKILLS

Programming: Python, SQL, FastAPI, JavaScript/TypeScript, LangChain, LlamaIndex, PyTorch

Cloud: AWS, Azure, GCP concepts, Kubernetes/AKS, Infrastructure as Code principles

Data Engineering: ETL/ELT pipeline design, data modeling (star/snowflake schema), data warehousing, data lake architecture, data governance, data quality frameworks, schema design, partitioning, lineage & metadata management, dbt

Big Data: PySpark/Spark, Databricks, distributed data processing, Apache Kafka (streaming concepts), Apache Airflow (orchestration concepts), batch & near-real-time pipeline design, Apache Flink

Databases: Snowflake, PostgreSQL, NoSQL concepts (MongoDB/Cassandra)

AI/ML: RAG, semantic/hybrid search, embeddings, ranking & relevance tuning, LLM evaluation, NDCG/MRR, precision@k, hallucination/grounding checks

DevOps: Docker, CI/CD, observability, monitoring & release checks, AI-assisted engineering workflows

Collaboration: Cross-functional stakeholder management, technical communication, data contract ownership, mentoring, Agile delivery

WORK EXPERIENCE

Data Engineer — IBM

Mar 2025 – Present

- Architected Snowflake data warehouse and lakehouse-style analytics models across 20+ source systems, applying modern data modeling and schema design practices.
- Implemented governed semantic models, automated data quality checks, and lineage tracking with freshness SLAs, establishing trusted single-source-of-truth KPIs.
- Engineered Python and SQL ETL/ELT pipelines for source-to-consumption validation and attribution modeling, increasing dashboard adoption by 45% and cutting manual reporting by 60%.
- Defined reusable data contracts and governance standards, establishing clear ownership across teams and reducing rework.
- Served as lead technical point of contact for architecture decisions across 5+ stakeholder teams, translating requirements into scalable ETL/ELT designs.
- Led design and code reviews for Python/SQL pipelines and mentored 3 engineers, reducing production incidents and improving reliability.

Data Engineer — Salesforce

Mar 2024 – Mar 2025

- Designed Databricks and PySpark ETL/ELT pipelines on distributed compute for RAG, model fine-tuning, and ML data workflows.
- Improved retrieval accuracy by 38% and reduced inference latency by 27% through pipeline and model tuning.
- Built MLflow-style evaluation automation and CI/CD release checks for GenAI pipelines, accelerating time-to-production by 80%.
- Partnered with GenAI and engineering teams to productionize NLP workflows, embedding retrieval-quality checks into the release lifecycle.
- Optimized PySpark job performance on Databricks through partitioning, caching, and cluster tuning, lowering runtime and compute cost.
- Documented pipeline architecture and evaluation methodology, shortening ramp-up time for new GenAI team engineers.

Data Analyst — Deloitte

May 2023 – Feb 2024

- Analyzed structured and unstructured datasets using SQL, Python, and Excel to identify trends and support business decisions.
- Built interactive Power BI and Tableau dashboards, reducing manual reporting effort by 30–40% and enabling real-time KPI tracking.
- Cleaned, transformed, and validated data using SQL and Excel, improving data accuracy by 25% across business reports.
- Collaborated with cross-functional teams to gather requirements and deliver insights, improving operational efficiency by 15–20%.
- Supported ETL workflow development and data integrity monitoring alongside ad-hoc analysis and reporting.

PROJECTS

RAG Eval Workbench [GitHub](#)

- Built an offline evaluation pipeline measuring retrieval quality, hallucination risk, latency, and cost with deterministic, CI-friendly outputs.

Enterprise AI Knowledge Platform [GitHub](#)

- Built an enterprise RAG platform simulator with hybrid retrieval, access-aware answers, audit logs, and an operator dashboard for governance and observability.

AI Reliability Control Plane [GitHub](#)

- Built a release-gate simulator evaluating pipeline quality, safety, latency, and cost, producing auditable PASS/WARN/BLOCK decisions.

RepoLens MCP [GitHub](#)

- Open-sourced a local-first repository-intelligence server indexing codebases into dependency graphs; validated on a 160-file monorepo with 5,826 symbols and 39,482 edges.

EDUCATION

MS, Computer Science — California State University, Northridge

Jan 2022 – May 2024

BTech, Computer Science — Kalinga Institute of Industrial Technology

Jul 2017 – May 2021

CERTIFICATIONS

- Databricks Fundamentals
- AI Software Developer (micro1)
- AWS Cloud Practitioner
- TensorFlow Developer Certificate
- Evaluating and Debugging Generative AI
- Finetuning Large Language Models
- LangChain for LLM Application Development