

JAYAPAL REDDY MYAKA

San Antonio, TX | +1 (703) 261-4083 | jayapal.myaka02@gmail.com

<https://linkedin.com/in/jaymyaka> | <https://github.com/jaymyaka> | pypi.org/project/dataenginex

PROFESSIONAL SUMMARY

Senior Data Engineer with 6+ years designing and implementing Databricks and Snowflake lakehouse and warehouse platforms for batch, streaming, analytics, ML, and AI workloads across AWS, Azure, and GCP. Builds Python, SQL, Spark, Kafka, Airflow, dbt, Delta Lake, and Apache Iceberg pipelines with medallion architecture, dimensional modeling, data contracts, schema enforcement, data quality, Unity Catalog governance, lineage, and production observability. Delivers ML infrastructure spanning feature engineering, point-in-time joins, MLflow lifecycle management, and batch and low-latency inference, plus production AI systems including RAG, vector search, LangGraph agentic workflows, and LLM evaluation gates. Improves reliability, freshness, performance, and cost through Spark tuning, warehouse right-sizing, Terraform, and CI/CD automation, with delivery across healthcare, pharmaceutical distribution, finance, education, and enterprise platforms.

TECHNICAL SKILLS

Languages & Frameworks: Python, SQL, PySpark, Spark SQL, Scala, Bash, FastAPI, pandas, pytest

Platforms, Lakehouse & Storage: Databricks, Snowflake, Snowpark, Snowflake Cortex, Unity Catalog, Delta Lake, Apache Iceberg, BigQuery, Azure Synapse Analytics, Amazon Redshift, Amazon Athena, DuckDB, PostgreSQL, SQL Server, Oracle, Parquet, Avro, Amazon S3, ADLS Gen2, Google Cloud Storage

Pipelines, Orchestration & Streaming: Apache Airflow, dbt, Lakeflow Declarative Pipelines (Delta Live Tables), AWS Glue, Azure Data Factory, Cloud Composer, AWS Step Functions, Apache Kafka, Spark Structured Streaming, Schema Registry, Amazon Kinesis, Google Pub/Sub, Snowpipe, Change Data Capture, exactly-once semantics, ETL and ELT design

Modeling, Quality & Governance: Dimensional modeling, Kimball, Star Schema, SCD Type 2, data contracts, schema evolution, reconciliation, dbt tests, freshness SLAs, data observability, Microsoft Purview, AWS Lake Formation, data lineage, RBAC, dynamic data masking, row access policies, HIPAA, GDPR, SOX, PCI-DSS

Cloud, DevOps & Delivery: AWS, Microsoft Azure, GCP, Terraform, Docker, Kubernetes (AKS, K3s), GitHub Actions, Azure DevOps, ArgoCD, Prometheus, Grafana, Datadog, PagerDuty, incident management, SLOs, Power BI, Tableau, Databricks SQL, mentorship and developer coaching

Machine Learning: MLflow, feature engineering, feature stores, point-in-time joins, scikit-learn, XGBoost, PyTorch, LSTM, Hugging Face Transformers, Amazon SageMaker, Optuna, model registry, model serving, drift detection, A/B testing, NDCG, MRR, CTR

AI & LLM Engineering: RAG, LangChain, LangGraph, agentic workflows, tool calling, embeddings, vector search, pgvector, Pinecone, Qdrant, hybrid and dense retrieval, chunking, Ragas evaluation, guardrails, human-in-the-loop, prompt engineering

PROFESSIONAL EXPERIENCE

Code Connect Inc. | Senior Data Engineer

San Antonio, TX | 03/2026 – Present

- Designed and implemented a multi-tier data platform on AWS, Databricks, and Snowflake, combining lakehouse, warehouse, and feature-store patterns for batch, streaming, analytics, ML inference, and ad-hoc workloads.
- Built batch and incremental pipelines with PySpark, Lakeflow Declarative Pipelines (Delta Live Tables), AWS Glue, S3, dbt, Airflow, and Step Functions, ensuring data contracts, schema evolution, and automated source-to-target reconciliation.
- Designed Kafka and Spark Structured Streaming pipelines with Schema Registry, Avro, event-time watermarking, stateful aggregations, and idempotent MERGE into Delta Lake and Snowflake, resolving duplicate aggregations from event-time versus processing-time drift.
- Implemented Unity Catalog governance with attribute-based access control, column masking, and end-to-end lineage, and enabling Delta and Apache Iceberg interoperability so Snowflake and external engines read lakehouse tables without data duplication.
- Built a data quality framework for null thresholds, schema drift, referential integrity, and freshness SLAs wired into Datadog and PagerDuty with SQL assertions and pytest gates in GitHub Actions, and cut cloud spend by right-sizing Snowflake warehouses.
- Built feature engineering pipelines with Delta time-travel datasets and point-in-time joins to eliminate training-serving skew, and managed the model lifecycle in MLflow with experiment tracking, registry, versioned promotion, and batch and low-latency inference.
- Designed retrieval-augmented generation workflows with LangChain, pgvector, and Pinecone, covering document chunking, embedding pipelines, and hybrid dense plus keyword retrieval over enterprise content.
- Built agentic AI workflows in LangGraph with tool calling, state management, and human-in-the-loop checkpoints, exposed through FastAPI services with authentication, rate limiting, and structured logging.
- Established LLM evaluation gates on retrieval precision and answer groundedness that caught a ranking regression prior to production, and applied Snowflake Cortex for in-warehouse classification and summarization, maintaining data governance.

Great Jobs KC | Data Analytics Engineer

Kansas City, MO | 09/2024 – 03/2026

- Designed and implemented the migration of legacy reporting and source systems into an Azure warehouse and lakehouse on Synapse Analytics, ADLS Gen2, Databricks, Delta Lake, and Data Factory, serving 50+ staff across scholarship, training, finance, and executive reporting.
- Implemented Bronze, Silver, and Gold medallion layers with schema enforcement, reusable PySpark and SQL transformations, data-quality scoring, Star Schema, and SCD Type 2 history, migrating full-refresh pipelines to CDC and Delta MERGE for near-real-time freshness at 35% lower compute spend.
- Developed unified scholar data models in PySpark and Delta Lake to integrate application, eligibility, award, enrollment, persistence, completion, and employment outcomes across partner colleges and training partners.
- Led Salesforce migration consolidating SMA, Microsoft Dynamics, SQL Server, PostgreSQL, REST API, and flat-file sources ensuring data integrity through source-to-target reconciliation and custom validation gates in Azure Data Factory.
- Provisioned Azure infrastructure as code with Terraform, deployed containerized services on AKS, automated CI/CD through Azure DevOps with pytest gates, and monitored pipeline SLAs in Prometheus and Grafana with Purview lineage and RBAC.
- Built PySpark feature datasets and scikit-learn models for scholar retention, persistence risk, enrollment funnel conversion, scholarship impact, and advising outcomes, with scoring pipelines informing program investment.

- Created AI-assisted scholarship application review system in Python using LLM structured extraction with JSON-schema validation, scoring rubrics, and confidence thresholds, with incorporating human-in-the-loop approval to maintain decision authority with reviewers.
- Delivered a RAG chatbot over scholarship eligibility, program, and partner-college policy documents using LangGraph orchestration, embedding pipelines, and hybrid retrieval, served via FastAPI and evaluated with Ragas on faithfulness, answer relevancy, and context precision.

McKesson | Data Engineer

Kansas City, MO | 08/2023 – 08/2024

- Designed and implemented a GCP and Snowflake cloud data platform for healthcare and pharmaceutical supply-chain analytics using Cloud Storage, BigQuery, Dataproc, Cloud Composer, Databricks, and Delta Lake, with Kafka, Pub/Sub, and Spark Structured Streaming feeds for clinical events and device telemetry.
- Built terabyte-scale PySpark pipelines for Epic, Cerner, HL7 FHIR, claims, pharmacy dispense, and medical-device data with standardization, deduplication, validation, and entity resolution under HIPAA, improving query performance 30% and Spark runtimes 40% via partitioning, clustering, AQE, and skew handling.
- Developed data models in dbt and SQL with incremental materializations, schema and data tests, and documented lineage, linking clinical, dispensing, and claims records into longitudinal patient and product views for life-sciences outcomes research.
- Built supply-chain intelligence marts in Snowflake SQL covering inventory position, order fill rate, distribution-center throughput, and product shortages, utilizing clustering keys, materialized views, and zero-copy cloning.
- Delivered enterprise reporting with Snowpark for Python, Snowpipe ingestion, and incremental Streams and Tasks, cutting dashboard refresh from 24 hours to under 15 minutes, and protected PHI with Snowflake dynamic data masking, row access policies, and secure views.
- Built forecasting and margin-analysis models in Python and Snowpark on distribution and dispense data to support drug profitability, contract performance, and financial planning for health-system customers.
- Enhanced patient-matching accuracy with custom PySpark fuzzy-matching logic layered over a vendor matcher to handle hyphenated names and cross-cultural naming conventions, reducing manual reconciliation queues.

TA Digital | Data ML Engineer

Hyderabad, Telangana, India | 12/2019 – 07/2022

- Developed SearchBoost, a multi-tenant SaaS AI search and discovery product spanning commerce catalogs, content, and enterprise knowledge sources, with reusable ingestion adapters for Solr, Elasticsearch, AEM, and REST APIs driving incremental index builds and per-tenant sync.
- Developed XGBoost, LSTM, and attention-based learning-to-rank models and fine-tuned Hugging Face transformers in PyTorch on SageMaker fine-tuned with Optuna, evaluated on NDCG, MRR, CTR, and A/B tests, and reranked on live click-through and conversion signals with per-tenant personalization.
- Delivered SearchBoost relevance through semantic retrieval over fine-tuned transformer embeddings served from Apache Solr, Elasticsearch, with named-entity extraction of query attributes, synonym expansion, typo tolerance, and intent parsing that reduced zero-result searches.
- Built AWS batch and streaming platforms with S3, Glue, EMR, Kinesis, Lambda, Athena, Redshift, Step Functions, and Airflow, and engineered behavioral and transactional feature pipelines for search, personalization, market-risk, credit-scoring, and portfolio-optimization models to support data-driven decision making.
- Reduced Spark batch runtimes approximately 35% on EMR using adaptive query execution, predicate pushdown, and broadcast joins, while ensuring compliance with SOX and PCI-DSS controls using Redshift materialized views, Glue Data Catalog and Lake Formation permissions, KMS encryption, and audit-grade reconciliation.

Quickcore Technologies | Junior Data Engineer

Warangal, Telangana, India | 04/2019 – 11/2019

- Built a GCP, Firebase, and Cloud Firestore IoT location-tracking platform with high-write schemas, batched writes, and live dashboards to enable real-time and historical movement reporting.
- Migrated cron-based Python and SQL ETL jobs to Apache Airflow with dependency management, retries, SLA monitoring, and alerting, increasing pipeline success rate from approximately 85% to over 98%.
- Collaborated with cross-functional teams to gather requirements for data integration processes. Assisted in database design and optimization to improve query performance and efficiency.

PERSONAL PROJECTS

dataengindex | Founder and Lead Engineer | Open-source Data + ML + AI platform | pypi.org/project/dataengindex

- Created dataengindex, an MIT-licensed Python 3.13+ framework on PyPI (v0.7.0) for configuration-driven medallion pipelines, with connectors for CSV, Parquet, DuckDB, PostgreSQL, S3, GCS, BigQuery, Delta Lake, dbt, REST, and Kafka, plus quality gates, schema validation, and lineage.
- Implemented an ML lifecycle layer with scikit-learn training wrappers, MLflow experiment tracking and model registry, PyTorch support, drift detection, and model serving on DuckDB as the default engine, with optional PySpark 4.2 for distributed workloads.
- Built an AI orchestration layer with pluggable LLM providers (OpenAI, Anthropic, Ollama), LLM routing, tool calling, agent memory, and human-in-the-loop workflows over BM25, dense, hybrid, and graph retrieval backed by Qdrant vector stores.
- Added PrivacyGuard PII detection and masking, structlog observability, and Redis background job queues, with a FastAPI control plane deployed on K3s using ArgoCD GitOps, Prometheus, Loki, and Grafana.

EDUCATION

M.S. Computer Science | University of Central Missouri, Lee's Summit, MO

08/2022 – 12/2023

B.S. Computer Science | SR University, Warangal, India

08/2016 – 05/2020

CERTIFICATIONS

AWS Certified Data Engineer – Associate | AWS Certified Cloud Practitioner | Google Data Analytics Professional Certificate