PAWAN YANDAPALLI

Data Engineer

Dublin, CA | (779) 977-7799 | pawanyandapalli1@gmail.com | pawanyandapalli.com | linkedin.com/in/pawanyandapalli | github.com/pawanyandapalli7

SUMMARY

Data Engineer with 5+ years of experience designing and operating cloud-native data platforms, distributed systems, and AI-ready data pipelines across healthcare and enterprise environments. Skilled in Python, SQL, PySpark, Apache Spark, Airflow, Databricks, Snowflake, Microsoft Fabric, and AWS, with a track record of building production ETL/ELT pipelines, CDC architectures, batch data systems, data quality frameworks, and governed data products. Former Amazon Software Development Engineer with a foundation in distributed systems, infrastructure automation, observability, and secure, large-scale enterprise data sharing. Comfortable owning systems end-to-end — architecture, implementation, and production operations — in fast-moving, ambiguous environments.

CORE TECHNICAL SKILLS

Languages: Python, SQL (Advanced), PySpark, Bash/Shell

Data Engineering & Orchestration: ETL/ELT, CDC, Apache Airflow, dbt, Data Modeling (OLTP/OLAP), Data Warehousing, Data Quality, Data Contracts

Cloud & Lakehouse Platforms: AWS (Glue, EMR, S3, Lambda, Redshift, DMS, EventBridge, IAM, KMS, Lake Formation), Google BigQuery, Microsoft Fabric, Databricks, Snowflake, Delta Lake, Aurora PostgreSQL

Infrastructure & DevOps: Terraform, Docker, Kubernetes (EKS), Jenkins, GitHub Actions, CI/CD

Governance, Security & BI: RBAC, IAM, Data Lineage, Great Expectations, HIPAA, GDPR, Power BI, Tableau, Amazon QuickSight

AI Infrastructure (Exploratory): RAG Pipelines, LangChain, pgvector, Semantic Search, LLM Evaluation, REST APIs, FastAPI/OpenAPI

PROFESSIONAL EXPERIENCE

Data Engineer — TouchWorld Technologies *Feb 2025 – Present*

- Built AWS DMS and Glue ingestion pipelines processing 500K+ policy, claims, and customer records/day from disparate policy-administration and claims-management systems into a governed data lake.
- Designed a raw/cleansed/curated S3 zone architecture feeding Amazon Redshift, and modeled policy history with Slowly Changing Dimensions (SCD) to support multi-year compliance retention.
- Implemented claim validation and deduplication rules in AWS Glue — claim-date-vs-policy-period checks, customer dedup, medical/claim code standardization — fixing data quality issues at the source for underwriting and fraud models.
- Built pipelines to handle late-arriving claims across the FNOL, adjudication, settlement, and subrogation lifecycle, supporting downstream fraud-pattern detection.
- Optimized ingestion workflows using CDC (AWS DMS, PySpark, Delta Lake), cutting latency from 4+ hours to under 30 minutes; orchestrated with Apache Airflow and AWS EventBridge for reliable scheduling and recovery.
- Enforced HIPAA/GDPR-compliant security using IAM, KMS encryption, and Secrets Manager across the ingestion and analytics pipeline.
- Built curated Power BI datasets and reporting models supporting underwriting, claims, and executive reporting dashboards; containerized ETL workloads on Docker/Kubernetes (EKS) with infrastructure provisioned via Terraform and CI/CD (GitHub Actions).
- Partnered with analysts, software engineers, and business stakeholders to define shared data contracts and improve metadata quality and discoverability across teams consuming the pipeline.

Software Development Engineer, Data Platforms — Amazon *Sep 2020 – Jul 2023*

- Owned and scaled distributed AWS data infrastructure supporting millions of daily transactions across enterprise operational and analytics systems.
- Built scalable Apache Spark and AWS EMR distributed processing workflows, integrating 10+ transactional data sources into centralized analytics platforms.
- Optimized Amazon Redshift performance through distribution-key and sort-key redesign — approximately 60% reduction in analytical query runtime.
- Developed reusable RESTful ingestion services and automated data pipelines (Python, SQL, AWS), enabling governed, self-service access to enterprise datasets.
- Implemented monitoring, logging, and operational observability, improving production reliability and incident response.
- Automated infrastructure provisioning and orchestration using Jenkins, Terraform, Lambda, and Step Functions to support production-grade deployments.

Associate Software Engineer — GSPANN Technologies (Client: Kohl's) *Aug 2019 – Sep 2020*

Functioned in a data analyst capacity — reporting, analytics, and data validation on the Kohl's retail account

- Analyzed retail sales, inventory, customer, and shipment data across Google BigQuery and Amazon Redshift — ingested by the team's data engineering pipelines — to support merchandising, supply chain, and sales reporting.
- Designed and optimized SQL-based analytical datasets and reporting queries across both platforms, improving accuracy and performance of business reporting for stakeholders.
- Performed data profiling, cleansing, and validation on structured and semi-structured datasets (CSV, JSON, TXT, GZIP), improving reporting accuracy and consistency.
- Partnered with data engineers to translate business reporting requirements into validated, analysis-ready datasets.
- Collaborated with product owners and business stakeholders to deliver reporting and analytics supporting store-level performance visibility.

SELECTED PROJECTS

Enterprise RAG Pipeline — github.com/pawanyandapalli7/data-engineer-portfolio

- Designed and implemented an enterprise Retrieval-Augmented Generation (RAG) platform using FastAPI, OpenAPI, pgvector, and Docker, exposing semantic search through REST APIs.
- Built semantic chunking, vector indexing, reranking, and LLM evaluation pipelines to deliver scalable, AI-ready data retrieval services.

Personal Health OS — Live Biometric Data Platform — pawanyandapalli.com/Py_PersonalOS

- Designed a cloud-native lakehouse platform ingesting Apple Health data with Medallion architecture, idempotent ETL pipelines, and automated data quality validation.
- Built curated datasets supporting self-service analytics, observability, and AI-ready data consumption.

PRODUCTION APPLICATIONS

Designed, built, deployed, and operated three production web applications end-to-end — backend services, deployment infrastructure, payment integration, and analytics — using React/TanStack Start, Vercel, Stripe, and GA4.

- The Gift Shop (thegiftshop.co) — custom e-commerce site with product catalog, order form, and Stripe payment integration, deployed on Vercel.
- Threads Beauty Bar & Spa (threadsbeautybar.com) — booking and business site with a server-backed admin pricing system (Vercel KV/Redis), GA4 event tracking, and SEO-optimized landing pages.
- Inflorax (inflorax.com) — marketing site for a creator-growth agency with interactive UI and conversion-focused design.

EDUCATION

M.S., Computer Science — University of the Pacific *Dec 2024*

Coursework: Distributed Systems, Big Data Analytics, Cloud Computing, Machine Learning, Database Systems

B.Tech, Electronics and Communication Engineering — Vishnu Institute of Technology *2019*

CERTIFICATIONS

- Microsoft Certified: Fabric Data Engineer Associate (DP-700)
- AWS Databricks Platform Architect — Databricks
- Azure Databricks Platform Architect — Databricks
- dbt Fundamentals — dbt Labs