

Pavan Kumar Regalla

Data Engineer

Jersey City, NJ | +1 917-979-3462 | pavan.launch@gmail.com

[linkedin.com/in/thepavanlkdnxyz](https://www.linkedin.com/in/thepavanlkdnxyz) | github.com/pavanregalla1 | pavanregalla.dev

PROFESSIONAL SUMMARY

Data Engineer with nearly five years across data engineering and analytics, plus hands-on public projects in batch and streaming platforms. Strong in Python, SQL, PySpark, AWS, Databricks and Kafka, with experience building governed pipelines, dimensional models and data-quality controls.

TECHNICAL SKILLS

Languages and Analytics: Python, SQL, PySpark, Pandas, scikit-learn

Data Engineering and Processing: Apache Spark, ETL/ELT, dimensional modeling, Delta Lake, medallion architecture

Cloud and Data Platforms: AWS (Glue, S3, EMR, EC2, Redshift), Snowflake, Azure Databricks

Databases and Warehousing: Amazon Redshift, Snowflake, SQL Server, data warehousing, star schema

Business Intelligence and Visualization: Power BI, DAX, Tableau

Orchestration, Quality and Development Tools: Apache Airflow, Apache Kafka, dbt, data-quality checks, Git, Docker, CI/CD

PROFESSIONAL EXPERIENCE

Data Analyst | Amazon Web Services | United States *July 2024 – Present*

- Built and maintained ETL/ELT pipelines with AWS Glue, Amazon S3, EC2 and EMR to ingest and transform batch datasets for analytics consumers, replacing manual hand-offs with scheduled, repeatable workflows.
- Wrote Python, SQL and PySpark transformations, using Pandas for validation and reconciliation, to cleanse and conform source data to target schemas so downstream reports reconcile to source.
- Modeled warehouse tables in Amazon Redshift and Snowflake with dimensional modeling, giving analysts trusted, query-ready datasets instead of rebuilt ad-hoc extracts.
- Orchestrated batch and near-real-time ingestion workflows in Apache Airflow with validation and reconciliation checks that catch data-quality issues before they reach downstream reporting tables.
- Translated stakeholder reporting requirements into pipeline and data-model changes, delivering validated, decision-support datasets on a reliable schedule.

Data Analyst | Deloitte | India *January 2021 – July 2023*

- Built ETL workflows in Python, SQL Server, SSIS and Azure Data Factory that ingested, cleansed, transformed and validated data from multiple source systems into scheduled pipelines.
- Developed stored procedures and views in SQL Server to standardize business logic and expose clean, reusable datasets for reporting.
- Ran data profiling and validation checks in Azure Databricks to catch completeness and consistency issues before data reached reporting tables.
- Delivered validated reporting datasets on a recurring cadence, giving business users consistent data and reducing rework from data errors.

TECHNICAL PROJECTS

Real-Time Retail Intelligence — Kafka, Spark & ML Platform

github.com/pavanregalla1/realtime-retail-intelligence

- Built a streaming analytics platform (Kafka to Spark Structured Streaming to Delta Lake, bronze/silver/gold) with a 10-minute watermark and checkpointing and deduplication design for exactly-once processing; inline velocity guardrails plus batch IsolationForest/GradientBoosting scoring in a sidecar.
- Verified on 80,628 synthetic streaming events: 100% recall on injected fraud scenarios at a 3.02% alert rate; demand-forecast MAE 83.4% below a naïve baseline; 24/24 dbt and pytest checks green.
- Scope: all data is synthetic. The file-replay path is CI-tested; the Kafka/Docker route was implemented but not executed in this environment.

Retail Sales Analytics — Batch Lakehouse & Power BI Pipeline

github.com/pavanregalla1/retail-sales-analytics

- Built a bronze-to-gold lakehouse on synthetic orders with medallion architecture, star schema and KPI queries: 60,600 raw rows in, 56,420 clean fact rows out, \$80.96M modeled revenue; 10/10 data-quality checks passing, 2,113 bad rows quarantined, never silently dropped.
- Ported the pipeline to PySpark/Delta Lake in a Databricks notebook and defined the test suite that enforces the quality gates on every run.
- Scope: no packaged PBIX file is included; a Power BI file can be built by following the repository's 10-minute guide. All data is synthetic; the Databricks notebook path was not independently executed.

Cloud Analytics Knowledge Search — Hybrid Retrieval & Grounded QA

github.com/pavanregalla1/cloud-analytics-knowledge-search

- Built hybrid TF-IDF plus dense (sentence-transformers) retrieval over 300 synthetic enterprise documents with extractive answers, document citations and honest abstention; hybrid recall@1 of 87.5% beats TF-IDF (75.0%) and dense (85.0%) alone on a 40-question evaluation set.
- Scope: corpus and questions are synthetic. Answers are extractive — no LLM is executed; Bedrock, LangChain and Docker are documented integration points, not run here.

CERTIFICATIONS

- AWS Certified Data Engineer – Associate
- Microsoft Certified: Power BI Data Analyst Associate (PL-300)
- Databricks Certified Data Analyst Associate

EDUCATION

- Master of Science in Computer Science | Adelphi University | NY | Sep 2023–May 2025
- Bachelor's Degree | JNTUH | Hyderabad, India | 2018–2022