Pavan Kumar Regalla

AI Engineer

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

linkedin.com/in/thepavanlkdnxyz | github.com/pavanregalla1 | pavanregalla.dev

PROFESSIONAL SUMMARY

AI Engineer with 5+ years building data platforms and 2 years of hands-on applied AI — RAG and hybrid retrieval, embeddings and vector search, ML anomaly detection and forecasting. Built a hybrid RAG system scoring recall@1 0.875 on a 40-question eval, vs dense 0.850, TF-IDF 0.750. LangChain, Amazon Bedrock, Azure AI Engineer Associate (AI-102); Python, PySpark, Kafka, Databricks.

PROFESSIONAL EXPERIENCE

AI Engineer | Amazon Web Services, United States *Jul 2024 – Present*

- Build and run ETL/ELT pipelines moving AWS service data from source to warehouse — Glue, S3, EC2, EMR, Redshift, Snowflake, orchestrated in Airflow — covering batch and near-real-time delivery.
- Built automated quality controls — schema validation, duplicate detection, reconciliation, exception reporting — that catch bad data before it reaches downstream consumers, instead of after it shows up in a dashboard.
- Tune the SQL, joins, views, and data models behind recurring workloads, so pipelines run faster and the next engineer can maintain them without reverse-engineering the last one.
- Turn cloud-service usage, customer, and operational data into answers with Python, SQL, Pandas, and PySpark — surfacing adoption patterns, anomalies, and performance opportunities.
- Work with product managers, engineers, and business stakeholders to turn requirements into clear KPIs, reusable datasets, and data solutions the business can rely on.
- Build the Power BI and Tableau dashboards that give product and operations teams one consistent view of service adoption, usage trends, and key operating metrics — the same numbers for every team.

Data Engineer | Deloitte, India *Jan 2020 – Jul 2023*

- Built ETL workflows that pulled data from multiple sources into shape — Python, SQL Server, SSIS, Azure Data Factory, Databricks — cleansing, transforming, and validating before anything downstream touched it.
- Wrote and optimized the T-SQL behind dimensional models and scheduled workloads — queries, stored procedures, views, indexes — keeping them fast and maintainable.
- Automated the repeatable work — extraction, transformation, reconciliation, reporting — with Python and SQL, reducing manual handoffs and keeping output consistent run after run.
- Profiled data and enforced validation rules before publication — catching missing values, duplicates, and business-rule issues while they were still cheap to fix.
- Supported data workloads across Azure Synapse, Azure Data Lake, Azure SQL, and Databricks — within access-control and data-governance standards, so the right people could reach the data and the wrong people couldn't.
- Worked with large operational and customer datasets in Python, SQL, Pandas, and Excel, turning raw records into recurring reports and insights people could actually decide on.

TECHNICAL PROJECTS

Cloud Analytics Knowledge Search — RAG & Hybrid Retrieval | github.com/pavanregalla1/cloud-analytics-knowledge-search

- Built a hybrid RAG knowledge-search system over 300 synthetic enterprise documents — TF-IDF + dense retrieval (sentence-transformers) with score fusion, extractive answers with citations, abstention when evidence is insufficient; Streamlit UI.
- Measured all three retrieval paths on a 40-question eval set: hybrid recall@1 0.875 vs dense 0.850 and TF-IDF 0.750; reported recall@5, MRR, and per-query latency for each path. 24 passing tests, CI on every push.

Real-Time Retail Intelligence — Kafka, Spark & ML Platform | github.com/pavanregalla1/realtime-retail-intelligence

- Built a streaming data platform — Kafka to Spark Structured Streaming to Delta Lake — with checkpointing and dedup design for exactly-once processing and batch ML scoring (IsolationForest, GradientBoosting) for anomaly detection and demand forecasting.
- Verified on 80,628 synthetic streaming events (file-replay path): 100% recall on injected fraud scenarios; forecast MAE 83.4% below a naïve baseline; 24/24 dbt and pytest checks green.

Retail Sales Analytics — Batch Lakehouse & Data Pipeline | github.com/pavanregalla1/retail-sales-analytics

- Built a bronze-to-gold lakehouse on synthetic orders with star schema and KPI queries: 56,420 clean fact rows, $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; Power BI consumption layer documented in the repository's build guide; all data is synthetic.

TECHNICAL SKILLS

AI & Machine Learning: Generative AI, RAG, LLM systems (retrieval for LLM applications), LangChain, Amazon Bedrock, Azure OpenAI, embeddings, vector databases, semantic and hybrid search, Scikit-learn, ML anomaly detection, forecasting

Data Engineering: ETL/ELT, Apache Spark, PySpark, Spark Structured Streaming, Spark SQL, Apache Kafka, AWS Glue, Apache Airflow, dbt, Databricks, Delta Lake, SSIS, dimensional modeling, data warehousing

Cloud & Databases: AWS S3, EC2, EMR, Redshift, Snowflake, SQL Server, Azure Data Factory, Azure SQL, Azure Synapse, Azure Data Lake

Programming & Data: Python, SQL, T-SQL, Pandas, NumPy, Excel, exploratory analysis, statistical analysis

Data Quality, BI & Ways of Working: Data profiling, cleansing, validation, schema controls, reconciliation, governance, Power BI, DAX, Tableau, SSRS, dashboard design, Git, Agile, Scrum, cross-functional collaboration

EDUCATION

Master of Science in Computer Science | Adelphi University, New York

Bachelor's Degree | Jawaharlal Nehru Technological University Hyderabad, India

CERTIFICATIONS

Microsoft Certified: Azure AI Engineer Associate (AI-102) | AWS Certified Data Engineer – Associate (DEA-C01) | Databricks Fundamentals Accreditation