Pavan Kalam

Seattle (Open to Relocate) | +1 913 – 263 – 4885 | pavan.k@savemymails.com | LinkedIn | portfolio | GitHub |

PROFESSIONAL SUMMARY

Data Engineer with 4.5 years of experience building scalable data pipelines and analytics solutions across healthcare, finance, and supply chain domains. Skilled in Azure (Data Factory, Databricks, Synapse, Event Hubs) and AWS (Glue, Redshift, S3, Kinesis) with expertise in PySpark and Python. Successfully optimized data pipelines, contributing to approximately $90K in annual infrastructure and operational cost savings. Experienced in delivering feature-ready datasets for ML models, implementing automated data validation, and ensuring secure, governance-compliant analytics for faster insights and decision-making.

TECHNICALSKILLS

Data Engineering & ETL: End-to-end ETL pipelines, Batch & Real-time Processing, Event-driven Architecture, Change Data Capture (CDC), Data Validation, Workflow Orchestration (Apache Airflow, AWS Step Functions, GitHub Actions, CI/CD)
programming Languages: Python, JavaScript, SQL, R
Machine Learning & AI: Transformers, TensorFlow, PyTorch, Scikit-learn, Keras, OpenCV, XGBoost, N8N, RAG, NLP, Vertex AI, Vector DB
Cloud Platforms: Azure (Azure Data Factory, Azure Synapse, Azure Event Hubs, Azure Data Lake), AWS (Glue, AWS Redshift, AWS S3, AWS Kinesis, AWS Lambda, AWS ECS/ECR, AWS SageMaker)
Big Data: Spark, Kafka, Databricks, Snowflake, Hadoop, Apache Flink
Databases & Data Modeling: SQL (PostgreSQL, Redshift), NoSQL (MongoDB, DynamoDB), Partitioning Indexing, Query Optimization
Programming & API Development: Python (PySpark, Pandas, NumPy), Scala, Flask, FastAPI, Gradio, REST APIs
Data Governance & Security: HIPAA Compliance, Data Encryption, IAM, RBAC, Audit Logging
Analytics & Reporting: Tableau, Power BI, Grafana, MS Excel (Advanced)
DevOps & MLOps Tools: Docker, Kubernetes, Git, AWS CodePipeline, Terraform

PROFESSIONAL EXPERIENCE

CVS Health — **United States**
Data Engineer — **May 2025 - Present**

- Developed end-to-end Azure Data Factory pipelines to consolidate claims, pharmacy, and member engagement data from multiple source systems, enabling analytics teams to work from a unified dataset of 1.6M+ records per month.
- Implemented streaming ingestion using Azure Event Hubs and Stream Analytics, enabling near-real-time processing of 8,000+ events per minute to improve operational monitoring and care-management signals.
- Engineered feature-ready datasets in Databricks (PySpark, Python) by standardizing time windows, behavioral attributes, and risk indicators, supporting 12+ production analytics and risk-scoring models.
- Integrated curated datasets with Azure Synapse and Azure Machine Learning, reducing model data preparation effort by 28% and improving consistency across recurring training and inference workflows.
- Applied automated data validation and anomaly detection across ingestion and transformation layers, preventing 180+ data-quality incidents annually from impacting executive dashboards and downstream ML consumers.
- Supported GenAI and LLM-based analytics initiatives by preparing governed, high-quality datasets and feature tables used for experimentation, evaluation, and downstream AI-driven insights, while ensuring compliance with healthcare data standards.

Zensar Technologies — **India**
Data Engineer — **Mar 2022 - Dec 2023**

- Redesigned large-scale batch data pipelines using AWS Glue, S3, and PySpark, optimizing Spark transformations and contributing to $90K+ annual infrastructure cost savings.
- Modeled analytics and ML-ready datasets in Amazon Redshift, optimizing partitioning and query performance to support 40+ concurrent analytical and model-consumption workloads.
- Implemented incremental ingestion and CDC patterns across S3 and Redshift, reducing refresh times from 6 hours to under 2 hours and enabling near-real-time feature availability.
- Introduced pipeline monitoring and alerting using Amazon CloudWatch, proactively identifying 120+ pipeline failures annually before impacting downstream analytics or ML consumers.
- Published governed datasets to S3 and Redshift, supporting 18+ production dashboards and ML feature consumers across finance and supply chain domains.

Zensar Technologies — **India**
Junior Data Engineer — **Jan 2020 - Feb 2022**

- Built foundational ETL and ML-supporting ingestion pipelines using AWS Glue, Python, and S3, processing 350K+ records daily across multiple source systems.
- Performed large-scale data cleansing, feature standardization, and schema alignment using PySpark, preventing 15K+ invalid records per month from propagating into analytics and ML datasets.
- Maintained Redshift tables and views consumed by operational reporting and downstream ML feature pipelines.
- Automated ingestion orchestration using AWS Lambda and Step Functions, saving 420+ engineering hours annually and improving pipeline reliability.
- Supported early-stage real-time ingestion using Amazon Kinesis, documenting throughput, latency, and scaling behavior for future ML-driven use cases.

ACADEMIC PROJECTS

End-to-End Feature Store for Predictive Analytics

- Designed a batch-based feature store using Azure Data Lake, Databricks (PySpark), and Delta tables, supporting point-in-time correct features for supervised learning use cases.
- Executed feature versioning, backfills, and data validation checks to ensure reproducibility between training and inference datasets.
- Integrated curated features with Azure Machine Learning experiments, enabling consistent model training without manual data preparation.

Scalable Analytics Pipeline with CI/CD and Data Quality Controls (AWS)

- Built an analytics pipeline using AWS Glue, S3, Redshift, and Airflow-style orchestration, processing multi-source datasets for analytical reporting.
- Performed Git-based CI/CD, environment-specific configurations, and automated data quality checks to simulate production-grade deployment practices.
- Designed dimensional models and optimized SQL queries to improve analytical query performance and cost efficiency.

EDUCATION

University of Missouri - Kansas | Masters of Science in Computer Science — **Kansas, Missouri**