Akhilesh Kumar Gopisetty

Arlington, TX | +1 469-450-5500 | akhileshgopisetty31@gmail.com
LinkedIn | GitHub | Portfolio

Professional Summary

- **Senior Data Engineer** with 4 years of experience delivering cloud-scale solutions for Fortune 500 clients (NielsenIQ, GSK), specializing in legacy modernization and high-volume pipeline optimization.
- **Performance Optimization:** Proven track record of re-architecting SAS/Oracle workflows into Azure Databricks/PySpark, achieving an **80% runtime reduction** (15 hours to 3 hours) via custom multiprocessing and vectorization.
- **Full-Cycle Engineering:** Expert in the complete data lifecycle, from ingesting raw multi-market files to building ACID-compliant Delta Lake warehouses and orchestrating 30+ task dependencies via Apache Airflow.
- **Regulated Domain Expertise:** Deep experience in high-compliance sectors (Pharma/Clinical Trials), implementing "shift-left" data quality frameworks (CQL/EDC) to ensure 100% audit readiness and null-free reporting.
- **AI & GenAI Integration:** Bridging the gap between Data Engineering and AI, with recent academic success building **Agentic AI** platforms using LLM Function Calling and complex RAG architectures.
- **Delivery Excellence:** Recognized for achieving **100% on-time delivery** on critical global initiatives (ranking #1 of 27 projects) by enforcing strict Agile sprints and proactive stakeholder communication.

Technical Skills

Big Data & Cloud: Azure Databricks, PySpark, Apache Spark, Delta Lake, AWS (S3, EC2), Azure Data Lake Gen2
Languages: Python (Multiprocessing, NumPy, Pandas), SQL (Oracle/PostgreSQL), SAS (Base/Macro/IML), Scala
Data Engineering: ETL/ELT Design, Dimensional Modeling, Schema Evolution, Performance Tuning (Partitioning/Z-Ordering), Data Quality (Great Expectations), Airflow (DAGs, Backfills)
Machine Learning: TensorFlow, PyTorch, Transfer Learning (VGG16), Feature Engineering (One-Hot Encoding), SMOTE
DevOps & Tools: Git, Docker, CI/CD, Agile/Scrum

Professional Experience

Tata Consultancy Services (TCS) Feb 2020 – Jan 2024
Data Engineer

Client: NielsenIQ (Retail Analytics)
Role: Data Engineer **Duration:** Feb 2020 – Dec 2022

- **SAS to Python Migration (Performance Rescue):** Rescued a critical migration project where initial Python scripts suffered a 7x performance regression. Architected a custom **multiprocessing framework** using NumPy to bypass the Python GIL, matching legacy SAS PROC IML performance and reducing runtime by **80%** (15 hours to 3 hours).
- **Pipeline Modernization:** Migrated 300+ legacy SAS programs to **PySpark/Databricks**, refactoring sequential row-wise logic into vectorized Spark transformations to support daily KPI reporting for global stakeholders.
- **Orchestration & Idempotency:** Led the development of a parameterized Apache Airflow DAG (30+ tasks) to replace manual SAS workflows. Implemented **idempotent file-existence checks** to prevent data duplication during historical backfills and re-runs.
- **Delivery & Reliability:** Achieved 100% on-time delivery for the migration (ranking #1 of 27 global initiatives) by stabilizing critical batch jobs and enforcing strict data parity between legacy and modern systems.
- **Infrastructure Optimization:** Reduced cloud compute costs and latency by tuning Spark partitioning strategies and Linux-based execution environments, freeing up cluster capacity for concurrent batch workloads.

Client: GSK Pharma (Clinical Data)
Role: Data Engineer **Duration:** Dec 2022 – Jan 2024

- **Clinical Data Quality (Shift-Left):** Directed end-to-end data integrity for GSK clinical trials by implementing **upstream CQL validation rules** within Inform/Veeva EDC systems. Engineered downstream ETL pipelines to extract this standardized data from Oracle, eliminating null-value propagation.
- **Observability & Alerting:** Stabilized 100+ legacy workflows by resolving a critical global email-domain misconfiguration. Restored alerting, validated failure scenarios, and implemented a monthly smoke test to ensure ongoing notification delivery.
- **Process Automation:** Automated Excel-heavy data preparation by replacing manual steps with Python and PySpark scripts, reducing manual effort by 70% while improving consistency and traceability.

Academic Projects

AroundMe – Agentic AI Discovery Platform Aug 2025 – Dec 2025
Python, FastAPI, OpenAI Function Calling, PostgreSQL

- **Agentic Orchestration:** Built an AI-driven analysis layer where an LLM utilizes **Function Calling** to dynamically filter, calculate, and summarize aggregated data from Yelp/Google APIs based on specific user query constraints.
- **Resilient Ingestion:** Implemented self-correcting parsing logic to normalize heterogeneous JSON payloads into a unified PostgreSQL schema, ensuring consistent downstream consumption.

Breast Cancer Detection (Medical AI) Aug 2024 – Dec 2024
TensorFlow, PyTorch, Transfer Learning

- **Recall Optimization:** Developed a VGG16-based deep learning model for cancer detection, prioritizing **Recall/Sensitivity** over raw accuracy to minimize false negatives in a high-stakes medical context.
- **Class Imbalance Handling:** Mitigated dataset skew (Healthy vs. Cancer) using **SMOTE** (Synthetic Minority Over-sampling Technique) and precision-recall analysis to ensure reliable predictions for minority classes.

Power Consumption Prediction Aug 2024 – Dec 2024
Python, Feature Engineering, Regression Modeling

- **Feature Engineering:** Improved regression model performance by engineering temporal features, utilizing **One-Hot Encoding** to capture seasonal cyclicality (Spring/Summer/Fall/Winter) from raw timestamp data.

AI-Driven Fitness Tracker Jan 2025 – Apr 2025
Node.js, Prisma, MySQL, AWS (EC2, S3)

- **Full Stack Development:** Developed a RESTful API layer (Node.js/Express) and relational database (MySQL) to support real-time workout tracking and consistent data retrieval for user analytics.

Education

University of Texas at Arlington Arlington, TX
Master of Science in Data Science Jan 2024 – Dec 2025
GPA: 4.00 / 4.00
Relevant Coursework: Applied Data Science (Advanced R programming), Database & Cloud Management (MongoDB/NoSQL design, AWS).

Certifications

- Microsoft Azure Fundamentals (AZ-900)
- Microsoft Career Essentials in Generative AI