# SHWETHA SIVAKUMAR

(609) 917-4534 ◇ shwethasivakumar282@gmail.com
LinkedIn ◇ GitHub ◇ Portfolio

## SUMMARY

Data Engineer with experience building end-to-end data pipelines across transit operations, IoT telemetry, and ML analytics. Consistently improved system performance and dashboard responsiveness by 30–40% through data modeling, query optimization, and data quality automation. Strong in Python, SQL, PySpark, Denodo, Informatica IICS, Snowflake, Azure Data Explorer, and Power BI, with AWS Solutions Architect and Denodo Developer certifications. Partner effectively with operations, and BI teams to translate technical insights into business decisions.

## EDUCATION

**Master of Science, Computer and Information Sciences**, Clemson University (May 2025), GPA: 3.86/4

**Bachelor of Technology, Computer Science Engineering**, Amrita University, India (Jun 2023) , GPA: 8.83/10

## TECHNICAL SKILLS

| | |
|---|---|
| **Programming & Processing** | Python, SQL, PySpark, Java, Pandas, NumPy, scikit-learn, Bash |
| **Data Engineering** | Apache Airflow, dbt, Denodo, Informatica IICS, Data Modeling |
| **Databases & Warehousing** | Snowflake, Redshift, PostgreSQL, SQL Server, MongoDB, Firebase |
| **Cloud & Infrastructure** | AWS (Certified), Microsoft Azure, GCP, Docker, Git |
| **Visualization & BI** | Power BI (DAX, TMDL), KQL, Matplotlib, Seaborn, Plotly |
| **Tools & Methodologies** | Jira, Postman, Agile/Scrum, REST APIs, Unit Testing |

## PROFESSIONAL EXPERIENCE

**Associate Data Engineer** Aug 2025 - Present
IQZ Systems *Atlanta, GA*

- Built and maintained enterprise Denodo data virtualization layer integrating 400K+ daily records from Amazon Redshift, SQL Server, Databricks and MongoDB for a passenger rail network, enabling unified analytics across ridership, revenue, and operational performance.
- Designed and optimized Denodo base, derived, and summary views, applying cached views with scheduled refreshes at different time granularities (daily, monthly, yearly), cost-based optimization, and query tuning to improve Power BI dashboard load times by 30%.
- Developed Power BI dashboards consuming Denodo views for revenue, ridership, route performance, and ticket journey analytics, and automated metadata and description propagation by modifying TMDL (.SemanticModel) files.
- Implemented RBAC and global security policies across Denodo and Power BI, managing multiple user roles and enabling GitHub-based version control for Fabric workspaces to support secure and reproducible BI development.

**Software Developer** Aug 2024 - Dec 2024
MyUI.AI (Clemson University) *Clemson, SC*

- Built scalable PySpark pipelines on AWS S3 to process 75K monthly accessibility sessions for a B2B mobile/kiosk application deployed across retirement facilities, powering reinforcement learning models that generate personalized UI adaptations for users with visual, motor, and cognitive impairments.
- Designed incremental data processing and partitioned data models (by event date and impairment type), enabling efficient trend analysis across accessibility metrics using Spark window functions.
- Optimized downstream analytics and model-training workloads by implementing partition pruning, reducing feature read times by 40%.
- Orchestrated end-to-end Spark workflows using Apache Airflow, ensuring reliable daily execution, reproducibility, and backfill support.

**Data Engineer** Jun 2024 - Aug 2024
BPM BI *Tallahassee, FL*

- Built end-to-end Informatica IICS ETL pipelines ingesting 500K+ GPS, GTFS, and fare records daily into Snowflake.
- Engineered transformations in Informatica Mapping Designer, calculating route-level delay metrics and operational KPIs for 200+ transit routes, while embedding data quality checks on GPS, fares, and schedule adherence, reducing reporting errors by 25%.
- Automated near real-time and batch orchestration with Taskflows, ensuring data availability within 5 minutes of ingestion.
- Enabled Power BI dashboards providing actionable insights on route delays, ridership trends, and revenue metrics.

**Data Engineer** Jan 2023 - Jun 2023
Caterpillar Inc *India*

- Designed 15+ real-time operational dashboards in Azure Data Explorer (ADX) to visualize telemetry from 500+ autonomous haul trucks, providing fleet health, utilization, and data quality insights for operations teams.
- Built IoT telemetry ingestion pipelines into ADX by defining table schemas and JSON ingestion mappings to structure high-frequency GPS and operational data at scale.
- Developed data transformations using ADX update policies to convert epoch timestamps, normalize vehicle identifiers, and derive operational states; created materialized views that pre-aggregated metrics and improved dashboard performance by 40%.
- Implemented ingestion monitoring system using KQL queries and Azure Monitor alert rules to detect missing or delayed telemetry, ensuring data pipeline reliability.

**Machine Learning Intern** May 2022 - Jul 2022
C3iHub, Indian Institute of Technology, Kanpur *India*

- Led a team of 3 to develop a cyber-attack detection system using ensemble and deep learning models, achieving 97% accuracy in classifying normal traffic, automated attacks, and manual intrusions.
- Built a robust data preprocessing pipeline using Python and regex to parse Apache web logs into structured pandas DataFrames for threat analysis.
- Conducted global cyber-attack pattern analysis across 87 countries using protocol analysis, attack categorization, and subnet tracking.
- Implemented XGBoost and LSTM/GRU-based neural networks for feature engineering, model training, and evaluation and used seaborn and Plotly for visualization.

## PROJECTS

**Real-Time Stock Data Pipeline with Apache Airflow:** Built an ETL pipeline using Apache Airflow to orchestrate automated stock data workflows. Created DAGs with sequential tasks for data extraction (yfinance API), transformation (moving averages, volatility calculations), database loading, and data validation.

**Cancer Patient Survival Prediction System:** Built machine learning pipeline to predict breast cancer patient survival using 2000+ clinical records. Implemented K-means clustering algorithm with optimal cluster selection and developed survival analysis dashboard using Kaplan-Meier estimator. Created automated gene expression analysis workflow with statistical hypothesis testing and data visualization components.

## CERTIFICATIONS AND PUBLICATIONS

- AWS Solutions Architect Associate (Apr 2025)
- Denodo Platform 9.0 Certified Developer Associate (Oct 2025)
- Published: "Health and Environment Monitoring System for Viral Respiratory Diseases" – ICCCNT 2023, IIT Delhi