Vatsa Arvind Kala

(979) 739-8190 vatsakala.contact@gmail.com linkedin.com/in/vatsa-kala/ github.com/Vatsakala Portfolio

Education

Texas A&M University, College Station, TX **Aug 2024 - May 2026**
Master of Science in Management Information Systems — ***GPA: 3.83/4.0*** College Station, Texas

Pandit Deendayal Energy University **Oct 2020 - May 2024**
B.Tech in Electronics And Communication Engineering — ***GPA: 9.42/10*** Gandhinagar, India

Technical skills

Languages & Frameworks: Python, SQL, R, C++, C, Javascript, JAVA, Flask, React, Pytorch, Node.js,MERN, OOP
Tools & Tech Stack: AWS(EC2, Fargate, S3, RDS, Redshift, Quicksight), Tableau, Power BI, GIT, JIRA, Postman, Kubernetes
Databases & Interests: Snowflake, MySQL, PostgreSQL, MongoDB, SPSS, STATA, LangGraph, LangChain, Machine Learning
Certifications: Professional Scrum Master 1, Fundamentals of Deep Learning, Microsoft Azure Data Engineer Associate

Professional experience

Trading Technologies **May 2025 - Dec 2025**
Data Engineering Intern New York, USA

- Collaborated on **migrating 4+ petabytes** from Microsoft SQL Server to PostgreSQL on AWS RDS and Snowflake, redesigning 500+ tables through renames, primary key changes, and corrected dependency order for cutover
- Engineered a Flask and SQLAlchemy analytics API with 30+ endpoints for the new TradeZoom platform, standardizing data access into reusable query modules and **improving release stability** during migration to cloud data stores
- Conducted migration validation using row counts, null checks, uniqueness checks, and relationship rules, **uncovering about 100+ legacy defects** such as duplicates, missing primary keys, and broken table dependencies before cutover

Texas A&M University **Jan 2025 - May 2025**
Student Assistant College Station, USA

- Elevated transplant survival prediction by packaging an **end-to-end DeepSurv training and evaluation pipeline** on the UNOS thoracic registry, modeling 25,953 lung-only cases and achieving 0.89 test C-index with 0.24 Brier score at 4000 days
- Restructured UNOS data from 215,462 patients and 532 variables into a 73-feature dataset by writing preprocessing and encoding scripts and applying MICE Random Forest imputation for 78% missingness, **avoiding 45% deletion loss**
- Built model trust and explanation layer by running Kaplan–Meier validation, generating SHAP explanations for top risk drivers, and **integrating LLaMA-3.1 8B to convert predictions into clinician-readable narratives** for decision support

ARK Meditech Systems **Dec 2023 - Jun 2024**
Data Intern Ahmedabad, India

- Streamlined Hadoop and Spark SQL pipelines that processed 100,000 daily machine logs captured during doctor usage, **boosting pipeline efficiency by 18%** and enabling near real-time visibility into device functionality
- Developed Spark SQL optimizations for a critical log-processing workload by reducing redundant scans and heavy transformations, **cutting runtime from 30 seconds to 9 seconds** and increasing reporting throughput
- Automated log aggregation and reporting workflows and built Tableau dashboards for equipment efficiency, patient volume trends, and quality benchmarks, **reducing manual effort by 50%** and strengthening stakeholder decision clarity by 30%

Projects

Point of sales database system | *Amazon Web Services, MongoDB, ETL, Cloud Architecture*

- Designed a POS relational database on AWS EC2 using MySQL/MariaDB, modeling 10K+ customers, 8K products, and 30K orders; implemented transactions and triggers to **enforce data integrity** across checkout and inventory workflows
- Transformed performance by building ETL using SQL and Skyvia, profiling join bottlenecks, then migrating to MongoDB with a new document model and indexes which improved load 25%, queries 65%, and **supported 50% more transactions**

Mai Shan Yun Analytics Dashboard | *SQL, Streamlit, Pandas, Business Intelligence*

- Architected a full-stack analytics ecosystem that unified **13+ operational datasets** including POS transactions and inventory logs into an automated ETL workflow powering real-time KPIs, interactive dashboards, and demand forecasting
- Formulated an **AI-driven insight engine** using Claude integration that generated contextual business recommendations and increasing forecast precision by 27%. **Received Top-3 honors at TAMU Datathon 2025**

RegDoc: AI-Powered Regulatory Document Classifier | *Databricks, OCR, LLaMA 3.1, Data Governance*

- Constructed a **multi-stage data classification pipeline** combining OCR extraction, heuristic safety checks, and dual LLM reasoning to label regulatory documents by sensitivity with 93% accuracy
- Synthesized confidence-based calibration, citation tracing, and human-in-the-loop feedback to deliver explainable, audit-ready compliance results while **cutting manual review effort by 40%**