Adith Sreeram Arjunan Sivakumar

Boston | +1 (413) 275-9015 | aarjunansiva@umass.edu | LinkedIn | GitHub | Portfolio

Education

University of Massachusetts - Amherst | *Master's, Computer Science* (GPA: 3.75/4) **Aug 2024 - May 2026**

VIT University | *B.Tech, Computer Science* (GPA: 8.67/10) **Aug 2020 - May 2024**

Skills and Certifications

- **Programming & Data Science:** Python, R, SQL, PySpark, Pandas, NumPy, Scikit-learn, Tensor Flow, PostgreSQL
- **Analytics & Machine Learning:** Power BI, Tableau, Excel (Pivot, VLOOKUP, Power Query), Data Storytelling, Analytical skills, Data manipulation, RDBMS, Data Mining, Data Modeling, Redshift, Looker
- **Data Engineering & Cloud:** AWS (S3, Glue, Lambda, Redshift), Azure (basics), ETL/ELT Pipelines, Airflow, dbt, Docker, Git, CI/CD

Experience

Omdena | *Data Engineer Internship* Dec 2025 - Feb 2026

- Designed and deployed an automated **AWS** based document ingestion and embedding pipeline (**S3, EC2, Docker, Python**) with **SQL** driven orchestration to process **10k+** unstructured policy and research documents, enabling scalable semantic retrieval.
- Built and optimized retrieval-augmented generation (**RAG**) workflows using **LLM** embeddings, vector indexing, and prompt orchestration, improving semantic search precision by **~35%** and reducing irrelevant responses during evaluation.
- Experimented with text chunking strategies, embedding models, and vector similarity metrics, increasing top-k retrieval relevance and improving downstream LLM answer quality across multiple test queries.
- Deployed and managed containerized ML workloads on **AWS EC2/ECS**, supporting rapid iteration and experimentation while maintaining reproducibility and cost-efficient compute utilization.
- Integrated **metadata enrichment** and **vector indexing** to improve document traceability and explainability, laying the foundation for future clustering, contradiction detection, and advanced analytics use cases

Team Yoga and Wellness Studio | *Data Analyst Internship* Aug 2023 - Aug 2024

- Supported analytical projects monitoring marketing and operational performance by using **Python, SQL, Power BI,** and **Excel** to prepare and visualize data, helping leadership quickly identify key trends and improve decision-making through process automation.
- Created automated dashboards and weekly reports highlighting attendance trends and customer engagement, improving visibility into key **KPIs by 25%** using AI-augmented coding workflows with Claude Code.
- Conducted exploratory data analysis and presented insights through presentations and visual summaries, improving management decision-making and communication.
- Collaborated cross-functionally with marketing and operations to evaluate campaign results and recommend data-backed improvements that increased engagement by **15%.**

Evry | *Machine Learning Engineer* Mar 2023 - Aug 2023

- Collaborated with senior data scientists and engineers to analyze large healthcare datasets, identify process inefficiencies, maintain data security and quality, and generate actionable insights using l**Python, SQL, and Power BI**, which helped prioritize workflow improvements and support clinical decision-making.
- Developed and trained machine learning models (**XGBoost, Random Forest**) to detect early-stage diseases, improving diagnostic accuracy by 13% and reducing manual effort by 82%.
- Created interactive dashboards and **PowerPoint** reports for management using **Power BI**, visualizing key performance metrics and patterns to enable faster quarterly reviews and shorten decision-making cycles by two weeks.
- Designed **REST APIs** with **Flask** and integrated scalable data models to embed AI-driven predictions into business tools, enhancing accessibility and real-time insight sharing across teams using Cursor for rapid prototyping.

Projects

Banking Data Pipeline with CDC & Cloud Warehousing | Project Link

- Engineered real-time CDC pipeline using **Debezium, Kafka,** and **PostgreSQL** to capture banking transactions with sub-30-second latency, processing **50+** events per batch into **Snowflake** for analytics.
- Architected **ELT** data warehouse on **Snowflake with dbt,** implementing SCD Type-2 dimensional modeling and incremental fact tables, enabling historical trend analysis across **10,000+** daily transactions.
- Automated end-to-end data orchestration using **Apache Airflow** and **Python**, reducing manual intervention by 100% through scheduled **DAGs** for **MinIO-to-Snowflake** ingestion and snapshot execution.
- Implemented **CI/CD** pipeline with GitHub Actions for **dbt** deployments, automating SQL compilation, data quality testing, and production releases, decreasing deployment time by 80% while ensuring data integrity.
- Built scalable event-driven architecture using **Docker** Compose with 8 containerized services (**Kafka, PostgreSQL, MinIO**), enabling horizontal scaling and 99.9% pipeline uptime for continuous data processing.

End-to-End AWS Redshift Data Pipeline with dbt Transformations & Airflow Orchestration | Project Link

- Architected end-to-end **ELT pipeline** ingesting **10K+** daily e-commerce transactions from **S3** to **Redshift** using **Python, dbt, and Airflow**, reducing manual data processing time by 90% through automated orchestration.

- Implemented dimensional star schema with 4 fact tables and 5+ slowly changing dimensions using **dbt**, enabling 60% faster analytical queries and supporting real-time business intelligence dashboards.
- Deployed **AWS** infrastructure via **Terraform** provisioning **Redshift Serverless, S3 data lake,** and **IAM roles**, cutting cloud costs by 40% through automated resource management and serverless architecture.
- Built idempotent data loader with partition-based ingestion supporting 100+ daily backfills, ensuring zero data duplication and enabling reliable recovery from failures with automated retry logic.
- Developed dual-environment strategy with local **PostgreSQL** and **cloud Redshift deployment**, accelerating development cycles by 70% and eliminating AWS costs during iterative testing and validation phases.

Scalable Data Engineering Pipeline for Enterprise Analytics | Project Link

- Built a fully server less, event driven **ETL pipeline** on AWS to automate data ingestion, transformation, and analytics reducing manual intervention by 100%.
- Integrated **AWS** services including **S3, Glue (ETL with PySpark), Lambda & Event Bridge (orchestration), and Glue Crawler (cataloging)** for end-to-end automation.
- Enabled near real-time analytics by transforming raw CSV into partitioned Parquet datasets and exposing them to **Athena** and **BI dashboards** for query-based insights.
- Implemented monitoring and alerting with **Cloud Watch** and **SNS** for proactive error detection, while enforcing least-privilege **IAM roles** and secure data governance.

Health-Guard: Web-Based Early Detection System | Project Link

- Developed a **Streamlit** based web application for forecasting and early prediction of five diseases, integrating **supervised machine learning** and **boosting algorithms** to support faster, data-driven diagnostics and reduce estimated healthcare costs by **~10%.**
- Designed and evaluated multiple classification and ensemble models, comparing performance using accuracy, precision, recall, and F1-score to select robust models for real-world deployment.
- Performed end-to-end data engineering using **SAS**, applying pre processing techniques such as PCA, min–max normalization, and one-hot encoding to improve feature quality and enhance model accuracy and stability.

Certifications

- AWS Certified Cloud Practitioner
- AWS Machine Learning Engineer Associate
- AWS Certified AI Practitioner
- Azure Fundamentals