Vaishnavi Mocherla

San Jose, CA | +1 (530) 407-3384 | vmocherla25@gmail.com | LinkedIn | Github

Education

San Jose State University San Jose, CA
M.S. in Applied Data Science *Aug 2023 – May 2025*

ICFAI University Hyderabad, India
B-Tech in Computer Science & Engineering *Jun 2019 – Jul 2023*

Technical Skills

Languages: Python, SQL, R (basic)

Big Data & ETL: Spark, Databricks, Kafka, Airflow, Informatica, Azure Data Factory, Glue, Lambda

Visualization: Power BI, Tableau, Looker, Matplotlib, Seaborn

Cloud Platforms: Azure, AWS, GCP, Docker, Kubernetes, Terraform

Databases & Data Warehousing: Snowflake, BigQuery, PostgreSQL, MySQL, MongoDB, SQL Server, Neo4j

Analytics & ML: Pandas, NumPy, Scikit-learn, TensorFlow, PySpark, MLflow, Time Series

LLMs & GenAI: LangChain, Semantic Kernel, Azure OpenAI, Pinecone, RAG, Fine-tuning, Prompt Engineering

Experience

Informatica California, USA
Data Engineer – AI & Cloud Systems *May 2024 – Dec 2024*

- Developed a **multimodal AI agent** using **Azure OpenAI** & **Python** to process **structured/unstructured data** (PDFs, JSON, SharePoint); automated **60%** of analytics and improved SQL accuracy by **20%**.
- Architected fault-tolerant **ETL pipelines** in **Informatica IICS** with **IDQ**, achieving **99.9% data accuracy** across **petabyte-scale** ingestion workloads.
- Optimized real-time and batch pipelines in **Databricks** (**PySpark** + SQL), reducing latency by **35%** and enabling **CI/CD-based** production deployments.
- Streamlined **Spark** workloads on **AWS EMR** through **cluster tuning** and **spot instances**, cutting compute costs and runtime by **30%**.
- Directed secure migration of 1TB+ data to **Azure Data Lake** with **CDAM**, applying **row-level security** and **masking** for compliance.
- Delivered executive-level **Power BI dashboards**, halving report turnaround and enhancing marketing **campaign visibility**.

Kanerika Software Pvt Ltd Hyderabad, India
Data Engineer - AWS Engineer *Jan 2023 – Jun 2023*

- Formulated an event-driven **ETL framework** using **API Gateway**, **Lambda**, and **Airflow** for real-time resume ingestion with auto-retry and async execution.
- Crafted a **custom parsing engine** using **AWS Textract** for **10+** formats (PDF, DOCX, scanned text), raising extraction precision by **70%** into a centralized **S3 data lake**.
- Executed scalable batch **inference pipelines** on **100K+** resumes using **AWS Glue** and **Athena**, enabling predictive scoring with downstream S3 persistence.
- Orchestrated **Airflow DAGs** on **EC2** with autoscaling and spot pricing for optimized pipeline performance.
- Synthesized recruiter-facing **Power BI dashboards** visualizing fit scores and KPIs, accelerating hiring decisions by **30%**.

Projects

Multi-Agent LLM System for Autonomous SDLC | *Snowflake, LangChain, Semantic Kernel, Llama 3, Pinecone, Streamlit*

- **Built** an autonomous LLM platform with **6 role-based agents** (PM, Dev, QA, Docs, Deploy) executing the full SDLC, from user story to GitHub deployment via **Streamlit UI** and **CI workflows**.
- **Integrated LangChain**, **Semantic Kernel**, and **fine-tuned models** (Llama 3, CodeT5) with **Pinecone RAG**, achieving **<2s** latency, **88%** QA pass rate, and **40%** reduction in hallucinations.

Unified Data Lakehouse for Product Analytics | *Kafka, Spark, Airflow, dbt, Redshift, S3, Tableau*

- Built a real-time lakehouse using **Kafka + Spark Structured Streaming** to ingest and transform 10M+ daily product events into **Delta Lake** on **AWS S3**.
- Orchestrated ELT workflows using **Airflow + dbt** to power near **real-time dashboards** on **Redshift**, enabling marketing and ops teams to track conversions with **<5s** latency.

Cloud-Native Data Engineering Platform for Responsible AI | *Airflow, MLflow, Neo4j, Azure, LangChain*

- Implemented a production-grade **ML pipeline** using **Airflow** and **MLflow** for automated training and deployment, with **explainability** powered by **Neo4j** and **LangChain**-driven LLM validation.