Mounika Geriki

Irving, TX | Open to Relocation | mounika.geriki@stonybrook.edu | 934-949-8567 | LinkedIn | Github

Summary

Data Engineer with 3+ years of experience designing and operating scalable, cloud-native data pipelines across finance, healthcare, and AI research. Strong expertise in Python, SQL, ETL orchestration, distributed data processing, and cloud platforms, with a proven track record of building reliable, high-quality production data systems. Experienced in integrating real-time and batch data, embedding AI/ML workflows, and partnering cross-functionally to deliver secure, high-impact insights from large-scale datasets.

Education

Stony Brook University *Stony Brook, NY*
Master of Science, Data Science | GPA: 3.58 *Jan 2024 - Dec 2025*
- **Coursework**: Statistical Computing & Learning, Data Analysis, Big Data Analysis, Data Management, LLM

Skills

Programming & Querying: Python, SQL, PySpark
Data Engineering & Processing: ETL pipeline design, data modeling, star schema, data quality checks, data validation, streaming and batch processing (Pub/Sub, Apache Beam), Airflow-style orchestration
Cloud & DevOps: Google Cloud (BigQuery, Pub/Sub, Dataflow, Composer, Vertex AI), Snowflake, Linux/Unix, Bash, Git, CI/CD
Databases & Analytics: BigQuery, Snowflake, Oracle, MySQL, PostgreSQL; Power BI, Tableau, Looker Studio
APIs, ML & DevOps: REST/SOAP, feature engineering, Random Forest, LightGBM; Git, CI/CD, monitoring, Agile (Scrum/Kanban)
AI & LLMs: RAG, BERT, LoRA fine-tuning, prompt engineering, Agentic AI

Professional Experience

Stony Brook University | *Graduate Research Assistant* | *Data Engineering & LLM* *Jan 2025 - Present* | *Stony Brook, NY*
- Designed end-to-end data **pipelines** to ingest, clean, and structure large volumes of scientific text data for **LLM**-driven analytics.
- Built curated, high-quality **datasets** for battery properties, applying rigorous validation to ensure **data consistency** and reliability.
- Implemented scalable data transformation workflows to support domain-specific **LLM** models (MaterialsBERT, BatteryBERT).
- Applied **LoRA fine-tuning** and **prompt-engineering**, improving structured signal extraction from unstructured data sources.
- Documented data workflows and system configurations to improve maintainability and knowledge transfer.

Tata Consultancy Services | *Data Engineer* *Dec 2022 - Jan 2024* | *Bengaluru, India*
- Engineered production-grade **ETL** pipelines using **Python** and **SQL** to process **50+ GB/day** of banking and transactional data.
- Integrated multiple heterogeneous data sources via **REST and SOAP APIs**, achieving **99.8%** pipeline reliability.
- Optimized joins, filters, and transformations, reducing data processing time by **45%**.
- Built reusable **SQL automation** frameworks and analytics datasets, improving reporting efficiency by **35%**.
- Supported a **Random Forest** churn prediction model, to drive targeted retention for high-value customers.
- Ensured compliance with security, access control, and audit standards in regulated financial environments.

Tata Consultancy Services | *Associate Data Engineer* | *Analytics & Reporting* *Jun 2021 - Nov 2022* | *Bengaluru, India*
- Led end-to-end migration of **50+ Oracle** dashboards to **Snowflake**, maintaining **100%** data accuracy.
- Designed layered **data models** and **SQL** transformation logic, reducing manual data preparation by **65%**.
- Built **20+** interactive **Tableau** dashboards supporting **200+** daily users.
- Consolidated and optimized multi-source datasets, reducing dashboard latency by **40%**.

VI Solutions | *Machine Learning Intern* *Sep 2020 - Jan 2021* | *Bengaluru, India*
- Developed **ML**-driven data pipelines for healthcare analytics, achieving **87%** prediction accuracy.
- Applied **clustering** and **OCR** pipelines to medical imaging workflows, improving processing speed by **30%.**
- Integrated analytics pipelines with real-time systems, achieving s**ub-500ms** latency.

Projects

InsightPilot – Agentic AI Copilot | *Python, SQL, DuckDB, LLMs, LangGraph, Streamlit*
- Built an agentic platform integrating data ingestion, SQL analytics, monitoring, and natural-language querying over curated datasets.
- Designed evaluation metrics (latency, accuracy, query success) and feedback loops to ensure reliability and production readiness.

Patient Vital Monitoring Sys | *Python, Google Cloud Pub/Sub, Dataflow (Apache Beam), BigQuery, Power BI*
- Engineered real-time streaming pipelines to ingest, validate, and enrich patient vitals with automated alerts using **Pub/Sub**.
- Implemented Bronze/Silver/Gold data modeling, data quality checks, and scalable analytics for streaming and historical workloads.

NYC Taxi Insights | *Google Cloud, BigQuery, Airflow, Looker Studio / Power BI*
- Built cloud ETL pipelines processing 9+ months of NYC Taxi data with 100% consistency and 35% performance improvement.
- Developed analytics-ready data models, dashboards, and KPI tracking to support experimentation, behavioral analysis, and data-driven decision making.

TravelWise-NYC: AI-Powered Data Access Layer | *Python, Flask, GCP, FAISS, LangChain*
- Architected a production-style RAG pipeline enabling natural-language access to curated datasets and real-time sources.
- Implemented vector indexing, adaptive query routing, and evaluation metrics (accuracy, latency, error rate) to ensure trustworthy analytical responses.