Vinay Ram Gazula

Newark, NJ | 908-552-1879 | gazulavinayram@gmail.com | vinaygazula.dev | linkedin.com/in/vinayramg

Professional Summary

Data Engineer with 3 years of experience in designing, building, and maintaining scalable data pipelines to enhance data quality and drive business growth. Proficient in ETL/ELT processes, data modeling, distributed computing frameworks like Trino, Spark, Databricks, and Snowflake. Effective at collaborating with data scientists, engineers, and analysts in cross-functional teams. Published multiple research papers focusing on machine learning and explainable AI during my MS in Data Science.

Experience

Impetus — Bengaluru, India

Data Engineer — July 2020 — Aug 2023

- Designed and implemented scalable ETL pipelines integrated with data quality checks to ingest and process 2 TB of raw data using PySpark, reducing processing times by 40% and accelerating access to business insights
- Leveraged Apache Parquet snappy compression and Amazon S3 life cycle policies reducing storage and I/O costs by 30%
- Proposed a data transformation plan utilizing DBT and Aiflow to achieve a 15% increase in data transformation efficiency
- Collaborated with stakeholders to design and implement robust data models for generating 10+ key KPIs, ensuring data accuracy and alignment with business reporting needs
- Accelerated the migration of data from Snowflake warehouse to S3 for a data lake solution, leveraging Athena for ad hoc analysis and Redshift Spectrum with materialized views for BI dashboards, resulting in 50% faster analytics report generation
- Leveraged AWS Glue Data Catalog and AWS Lake Formation to standardize metadata management and enforce data governance policies, reducing integration complexities and accelerating data discovery

New Jersey Institute of Technology — Newark, NJ

Research Assistant | NJIT Engineering Education Research — July 2024 — May 2025

- Analyzed institutional undergraduate student data with machine learning and developed regression models (Multiple Linear Regression, Random Forest, XGBoost, LightGBM) in R
- Performed ablation studies to determine the influence of academic, demographic and socioeconomic factors on students performance

Research Assistant | Data and Knowledge Engineering Lab — Sept 2023 — June 2024

- Integrated xAI tools (LIME, SHAP, Anchors, PDP, ALE) into "SolarFlareNet"—a transformer framework for space weather research. Published and presented research findings at FLAIRS 2024 and IEEE ICTAI 2024

Technical Skills

Languages	:	Python (PySpark, Polars, Pandas, Boto3), SQL, Scala (Apache Spark), Go, Rust, Bash, R, C
Databases	:	PostgreSQL, MySQL, Oracle (PL/SQL), MongoDB, SQLite, Microsoft SQL Server, Neo4j, DuckDB
Cloud	:	AWS (S3, Glue, Lambda, Athena, Redshift, Aurora, RDS, DynamoDB, Firehose, SageMaker Data Wrangler) Azure (Data Factory, Data Lake Storage, Synapse Analytics, Blob Storage) GCP (Cloud Storage, BigQuery, Dataflow, Dataproc, Bigtable)
Big Data	:	Trino, Databricks, Snowflake, Apache Spark, DBT, Apache Flink, Apache Kafka, Apache Hive
ETL/ELT Tools	:	Apache Airflow, Dagster, Pentaho Data Integration, Informatica
Visualization	:	Tableau, Looker, Power BI, Dash (Python), Apache Superset, Excel
Data Modeling	:	Normalization (3NF), OBT, Star Schema, Snowflake Schema, Data Vault
CI/CD	:	Git, GitHub, GitLab, GitHub Actions, Docker, Kubernetes, Terraform, Jenkins

Projects

Data Engineer Playground | *Docker, Airflow, Trino, Spark, MinIO, PostgreSQL, Project Nessie, Unity Catalog* — Github

- Built a fully containerized multi-service environment to prototype end-to-end ETL workflows, from data ingestion in MinIO to batch or stream processing with Spark and workflow orchestration via Airflow. Enabled interactive SQL analytics through Trino with connectors for Postgres DB, Nessie Catalog and Unity Catalog

TradeForecast | *Python, PyTorch, PyTorch Lightning, yfinance, Polars, scikit-learn* — Report

- Developed a modular, production-ready ETL and modeling pipeline that ingests OHLCV data via yfinance, engineers temporal and technical indicators (MA, MACD, RSI, ATR) using polars, and version-controls feature sets
- Implemented three deep-learning architectures (LSTM, ConvLSTM, EncTransformer) in PyTorch, orchestrated training with PyTorch Lightning and ReduceLROnPlateau scheduling, and tuned hyperparameters via grid search to optimize multi-horizon forecasts

AlgoTrade API | *Python, yfinance, Pandas, Tensorflow, ks-api-client* — Github

- Developed a fully automated NSE stock trading bot in Python by integrating real-time and historical data with yFinance, training ML models (including LSTM) for stock price prediction, and executing live trades via the Kotak Securities API

Research Publications

1. Interpretable Deep Learning for Solar Flare Prediction — IEEE ICTAI 2024 — 2024
2. An Interpretable Transformer Model for Operational Flare Forecasting — FLAIRS 2024 — 2024

Certifications

- Data Engineering Bootcamp — DataExpert.io — 2025
- Google Data Analytics Professional Certificate — Coursera — 2023

Education

New Jersey Institute of Technology — Newark, NJ

*Master of Science in **Data Science*** — GPA: 3.89/4

SRM University AP — Amaravati, India

Bachelor of Technology — GPA: 8.51/10