

# VINAY RAM GAZULA

📍 Newark, NJ 📞 908-552-1879 ✉ [gazulavinayram@gmail.com](mailto:gazulavinayram@gmail.com) 💻 [vinaygazula.dev](http://vinaygazula.dev) 🔗 [linkedin.com/in/vinayramg](https://www.linkedin.com/in/vinayramg)

## PROFESSIONAL SUMMARY

**Data Engineer** with over 4 years of experience in retail, healthcare, and finance domains. Proficient in designing, building, and maintaining scalable ETL/ELT pipelines, enterprise data warehouses, and data lakes. Expert in data modeling, distributed computing frameworks, data quality and reliability. Skilled in workflow orchestration and cloud data platforms like AWS, Azure, GCP, Snowflake and Databricks. Proven track record in cross-functional collaboration, effectively communicating with data scientists, engineers, and analysts in cross-functional teams. Holds an MS in Data Science and authored multiple research publications on machine learning and explainable AI.

## EXPERIENCE

**Data Engineer** June 2019 — Aug 2023  
**Impetus Technologies** Bengaluru, India

- Developed end-to-end data warehousing solution to track customer behavior and enable targeted ad campaigns, increasing revenue by 10%
- Designed and implemented scalable ETL pipelines integrated with data quality checks to ingest and process 4 TB of raw data using PySpark, reducing processing times by 40% and accelerating access to business insights
- Leveraged AWS Lambda and Glue for ETL processes, driving a 25% increase in transformation speed and data accuracy.
- Reduced 30% cloud storage and I/O costs by implementing Apache Parquet snappy compression and Amazon S3 life cycle policies
- Proposed a data transformation plan utilizing DBT and Aiflow to achieve a 15% increase in transformation efficiency
- Implemented robust data models for generating 10+ key KPIs improving 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 reporting
- 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

**Data Scientist** Jan 2024 — May 2025  
**New Jersey Institute of Technology** Newark, NJ

- Collaborated with multiple interdisciplinary research teams to extract actionable insights from complex scientific datasets
- Conducted analyses on institutional undergraduate student data, developing regression models in R to predict GPA
- Performed ablation studies to evaluate the influence of academic, demographic, and socioeconomic factors on student performance
- Designed and implemented “SolarFlareNet”—a transformer framework for forecasting solar flare occurrences achieving 90.7% accuracy
- Automated the end-to-end ETL pipeline for SolarFlareNet using Azure Data Factory to ingest data into Azure SQL Database, ensuring data quality and accelerating model iteration speed by 50%
- Integrated explainable AI algorithms (LIME, SHAP, Anchors, PDP, ALE) into SolarFlareNet to interpret/explain black-box model predictions

## TECHNICAL SKILLS

Languages : Python (PySpark, Polars, Pandas), SQL, Scala (Apache Spark), Bash  
Databases : PostgreSQL, MySQL, Oracle (PL/SQL), MongoDB  
Cloud : AWS (S3, Glue, Lambda, Athena, Redshift, DynamoDB, SageMaker)  
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, DuckDB  
ETL/ELT Tools : Apache Airflow, Dagster, Informatica, Alteryx, Pentaho Data Integration  
BI & Analytics : Tableau, Looker, Power BI, Grafana, Dash (Python)  
Data Modeling : Normalization (3NF), OBT, Star Schema, Snowflake Schema, Data Vault  
CI/CD : Git, GitHub, GitLab, 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

## 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

## EDUCATION

**New Jersey Institute of Technology** | Newark, NJ May 2025  
*Master of Science in Data Science* GPA: 3.89/4