# VENUGOPAL REDDY GANGULA

Chicago, Illinois 60616
872-258-6626 venugangula44@gmail.com venugopal_gangula

## Summary

Highly skilled Data Engineer with 4+ years of experience designing and implementing high-performance data solutions. Specialized in building scalable ETL pipelines, real-time data processing systems, and cloud-based data platforms using AWS, Azure, and Snowflake. Proven track record in optimizing data infrastructure, reducing processing latency by 75%, and delivering $500K+ in cost savings. Adept at developing fraud detection models with 82% accuracy and implementing HIPAA-compliant healthcare data systems handling 12M+ daily records. Combines deep technical expertise in Python, Spark, and Airflow with strong business acumen to drive data-driven decision making.

## Education

**Lindsey wilson college** **Aug. 2023 – May 2025**
*Master's in Computer/Information Technology Administration and Management* *Columbia, Kentucky*

## Relevant Coursework

- Data Structures
- Applied Statistics
- Database Management
- Artificial Intelligence
- Python
- Java
- Project management
- Cyber Security

## Skills

**Programming Skills**: Python (NumPy, Pandas, Matplotlib, seaborn, Scikit-learn, Tensorflow), Pyspark, SQL (PostgreSQL), R, VBA, DAX, Bash, AWS Lambda
**Databases/Technologies**: Apache Spark, HDFS, Hive, Kafka, Github (Actions), CI/CD, MongoDB, Airflow, Hadoop, AWS Glue
**Cloud**: Snowflake, Redshift, Azure Data Factory, Databricks, ADLS Gen 2, Synapse, AWS S3, EMR,EC2, RDS, BigQuery
**Data Modeling & Warehousing**: Star & Snowflake Schema, Dimensional Modeling, DBT, Medallion Architecture
**BI & Visualization**: Power BI, Tableau, Looker, Excel (Pivot Tables, Power Query)
**Data Governance**: Great Expectations, Unity Catalog, HIPAA Compliance, FHIR/HL7 Standards

## Experience

**Innovaccer** **Jan 2024 – Apr 2025**
*Data Engineer* *Remote, USA*

- Architected HIPAA-compliant data pipelines on AWS Glue and Redshift processing 12M+ daily patient records, achieving 99.99% uptime for critical care analytics.
- Spearheaded FHIR/HL7 standardization across 5 hospital networks, reducing interoperability costs by $320K/year through automated message transformation (Python, Mirth Connect).
- Redesigned Databricks clusters using Delta Lake optimizations and auto-scaling, cutting healthcare analytics runtime by 35% ($85K annual cloud savings).
- Deployed enterprise data quality framework with Great Expectations, eliminating 28% patient record mismatches and reducing ICU reporting errors by 19%.

**Tata Consultancy Services (TCS)** **Jul 2021 – Jun 2023**
*Data Engineer* *Hyderabad, Telangana*

- Engineered ultra-low latency market data pipelines processing 10M+ messages/sec using Python and Apache Kafka, reducing tick-to-trade latency by 75% for quantitative trading strategies.
- Designed and implemented a Snowflake-based dimensional data model with dbt-core that accelerated portfolio analytics queries by 90%, enabling real-time risk assessment for $500M+ positions.
- Developed 50+ production Airflow DAGs orchestrating ETL workflows for options pricing data, improving data freshness from hourly to near-real-time (15-second intervals).
- Built automated data validation framework using Great Expectations that reduced reconciliation errors by 40% across 20+ critical trading datasets.

**TeamTech Solutions** **May 2020 – Jun 2021**

*Data Engineer* *Hyderabad, Telangana*

- Automated 1TB+ of monthly financial transactions using Azure Data Factory ETL pipelines, improving data delivery speed by 70% and enabling real-time reporting.
- Developed a PySpark-based fraud detection MVP with 82% accuracy from limited historical data, leveraging creative feature engineering that laid the groundwork for the company's risk analytics platform.
- Built ELT workflows in Snowflake to transform raw financial data into daily-refreshed investor dashboards, saving 15+ hours weekly and streamlining executive reporting via Power BI.
- Established startup-wide data governance and lineage standards, including a unified data dictionary, accelerating due diligence efforts by 30% and supporting seamless collaboration across the full data stack—from ingestion to reporting.
- Designed and implemented a data quality monitoring framework that reduced pipeline failures by 40% through automated validation checks (using PySpark and Azure Functions), ensuring 99.9% reliability for critical financial reports.

## Projects

**Patient Readmission Prediction Pipeline** | *Python, PySpark, AWS Glue* **Dec 2024**

- Developed end-to-end predictive analytics pipeline processing 50K+ daily FHIR EHR records, reducing preventable readmissions by 28% in pilot hospitals.
- Engineered features (Python/PySpark) from unstructured clinical notes, improving model accuracy to 83% (Logistic Regression + SMOTE for class imbalance).
- Automated reporting: Deployed AWS Glue jobs to generate daily risk scorecards (S3 → Power BI), saving 10+ hours/week for care teams.
- Ensured compliance: Integrated Great Expectations validation checks, reducing data quality issues by 35%.

**ETL Pipeline for E-Commerce on Azure** | *Azure Data Factory, Databricks, Delta Lake, PySpark* **Nov 2022**

- Designed and deployed a cloud-based data pipeline on Azure to process 15GB+ of daily e-commerce data, reducing processing time by 25 percentage through Spark optimizations (partition pruning, caching).
- Implemented medallion architecture in Azure Data Lake (Bronze→Silver→Gold) using Delta Lake for ACID compliance, enabling reliable near-real-time analytics.
- Orchestrated workflows with Azure Data Factory, automating data ingestion from multiple sources into curated datasets for business intelligence.
- Built custom Delta Lake merge operations that improved upsert performance by 40% for frequently updated product inventory data.