VENUGOPAL REDDY GANGULA

Chicago, Illinois 60616

872-258-6626   venugangula44@gmail.com   venugopalReddygangula Linkedin   venugangula Github

Summary

Data Engineer / Data Analyst with 3+ years of progressive experience across healthcare, financial services, and e-commerce. Adept at building real-time data streaming systems, enterprise-grade ETL pipelines, and business intelligence solutions that transform raw data into actionable insights. Skilled in PySpark, SQL, Kafka, Snowflake, dbt, Airflow, and Power BI, with strong expertise across AWS, Azure, and GCP cloud ecosystems. Recognized for reducing streaming pipeline latency by 75%, improving data accuracy by 40% through automated validation, and delivering predictive models with 80%+ accuracy. Passionate about leveraging data to drive decision-making, optimize performance, and deliver measurable business outcomes.

Skills

Languages & Analytics: Python, SQL, R, DAX, SAS, PySpark, Pandas, NumPy, Scikit-learn
Data Engineering: ETL/ELT, Data Pipelines, Data Modeling, Data Warehousing, Airflow, dbt, Azure Data Factory, AWS Glue, Databricks, Delta Lake, Hadoop, Spark Structured Streaming, Kafka, Data Quality Validation, Great Expectations
Cloud & Warehousing: AWS (S3, Redshift, Glue, EMR, Lambda, RDS, Athena), Azure (ADF, Synapse, ADLS, Databricks), GCP (BigQuery, Dataflow, Pub/Sub), Snowflake
Visualization & Reporting: Power BI, Tableau, Looker, Excel (Pivot Tables, Power Query)
Database Systems: MySQL, PostgreSQL, Oracle, SQL Server, MongoDB, Cosmos DB
Methodologies & Tools: Agile/Scrum, CI/CD, Git/GitHub, Workflow Orchestration, DevOps, Infrastructure as Code (Terraform, CloudFormation), Data Governance, Documentation

Experience

Chicago Education Advocacy Cooperative (ChiEAC)   Jul 2025- Present
Cloud Data Engineer(via Data Science Alliance)   *Remote, USA*

- Designed FHIR-compliant data architecture for standardized nonprofit EHR ingestion and analytics on AWS.
- Developed AWS Glue ETL jobs, Lambda triggers, and Athena queries to support secure healthcare reporting.
- Built Power BI dashboards surfacing student performance and operational metrics, reducing manual reporting by 60%.
- Automated incremental ingestion and workflow orchestration with Airflow, reducing runtime and manual effort.
- Implemented data lineage, metadata management, and governance policies for HIPAA compliance.
- Developed SQL stored procedures, optimized queries, and indexing strategies to improve reporting efficiency.
- Created monitoring solutions for ETL pipeline performance, cost control, and anomaly detection, Configured IAM roles, RBAC, and audit logging for sensitive data.

Tata Consultancy Services (TCS)   Aug 2021 – Jun 2023
Data Engineer   *Hyderabad, India*

- Built Kafka + Spark Structured Streaming pipelines processing 10M+ events/sec for real-time financial analytics.
- Designed Snowflake warehouses and Redshift clusters, creating 50+ Airflow DAGs and dbt models.
- Implemented data quality checks, reconciliation frameworks, and automated testing improving accuracy by 40%.
- Enhanced performance of ETL/ELT pipelines, partitioning, and clustering strategies, reducing query time by 30%.
- Established CI/CD pipelines with GitHub Actions and Jenkins for automated data pipeline deployments.
- Applied data masking, encryption, and RBAC for secure analytics in financial systems.
- Conducted root cause analysis and production support for pipeline failures, reducing downtime by 20%.

TeamTech Solutions   May 2020 – Jun 2021
Data Analyst   *Hyderabad, India*

- Automated ingestion of 1TB+ daily financial data using Azure Data Factory, Databricks, and Delta Lake, reducing runtimes by 25% and ensuring ACID-compliant analytics with the Medallion Architecture (Bronze→Silver→Gold).
- Built fraud detection models in PySpark (82% accuracy) and integrated them with Power BI for real-time monitoring, delivering actionable insights on high-volume financial transactions.
- Developed Snowflake ELT workflows and dimensional models (star schema, fact and dimension tables) to support financial reporting, trend analysis, and executive KPI dashboards.
- Partnered with stakeholders through UAT sessions, requirements gathering, and workshops, while authoring data dictionaries, ETL documentation, and reporting standards to ensure alignment and maintainability.

Projects

Real-Time Twitter Sentiment Analysis | *Kafka, Spark Structured Streaming, MongoDB, Power BI*   **Dec 2024**

- Developed streaming data pipeline processing 10K+ tweets/hour with NLP sentiment classification (TextBlob).
- Stored processed data in MongoDB/Cosmos DB for scalable, fault-tolerant analytics.
- Built Power BI dashboards visualizing sentiment by hashtags, topics, and geographies.
- Implemented CI/CD pipeline with GitHub Actions reducing release time by 40%.
- Created monitoring solutions for stream lag, fault tolerance, and error recovery, Authored system architecture diagrams and technical documentation.

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

- Built scalable ETL pipelines processing 15GB+/day with optimized PySpark transformations and upsert logic.
- Applied Medallion Architecture with validation frameworks (Great Expectations) for reliability.
- Designed Power BI dashboards and dimensional data models for actionable insights.
- Implemented incremental loading and watermarking for performance optimization.
- Applied Azure resource monitoring and auto-scaling policies, cutting spend by 20%.
- Orchestrated workflows with Airflow DAGs for batch and streaming integration.
- Authored deployment guides, workflow documentation, and runbooks.

Education

Lindsey wilson college   **Aug. 2023 – May 2025**

Master's in Computer/Information Technology Administration and Management   *Columbia, Kentucky*