

Madhu Shyam Sundar Reddy Kota

Certified Data Engineer

Certified Data Engineer with experience designing and maintaining scalable data pipelines using PySpark, SQL, Databricks, and AWS. Skilled in building cloud-native data platforms, optimizing data processing workflows through Delta Lake architectures, and implementing robust data integration frameworks. Experienced in Generative AI and LLM-based analytics with hands-on expertise in Amazon Bedrock, LangChain, and RAG. Proven ability to reduce infrastructure costs, automate manual processes, and deliver maintainable solutions that support data-driven decision-making across manufacturing and analytics domains.

Email: SHYAMSHS2002@GMAIL.COM | Location: Asia/Kolkata

About

Certified Data Engineer focused on scalable data platforms, reliable pipelines, and AI-assisted analytics. I specialize in PySpark, Databricks, and AWS data services, and I enjoy building LLM-powered tools that help teams query and understand data faster.

Experience

Software Engineer, Info Services · INAAS (Insightful-Service) (Jan 2024 - Present)

Developed an AI-powered analytics agent that converts natural language questions into SQL queries and delivers real-time insights for non-technical users.

- Achieved 90% query accuracy with response times under 30 seconds using Amazon Bedrock (Nova), LangChain, and RAG.
- Designed a multi-warehouse query engine supporting Databricks SQL and PostgreSQL with S3 as a data lake backend.
- Implemented session-aware memory in DynamoDB enabling multi-turn analytics conversations.
- Deployed a serverless production-ready architecture with AWS AppRunner (backend) and AWS Amplify (frontend).

Data Engineer (Manufacturing Analytics) · HapPy Inc. (2023 - 2024)

Built an end-to-end data pipeline to centralize and analyze razor manufacturing data, enabling insights for production optimization and quality control.

- Engineered a medallion architecture (Raw → Silver → Gold) processing 2GB daily from 4 source types on EMR Serverless.
- Built 8 data assets (7 dimension tables, 8 fact tables, 5 stored procedures, 10 views) in Redshift powering 4 Power BI dashboards.
- Automated data validation rules and schema drift handling, saving 10-20 hours/week of manual analysis.
- Reduced infrastructure costs using Delta Lake on S3 and on-demand EMR Serverless with GitHub Actions CI/CD.

Skills

PySpark & Delta Lake	ETL, Medallion Architecture, Optimization
Databricks	SQL Warehousing, Unity Catalog, Jobs
AWS Data Services	S3, EMR Serverless, Redshift, Lambda
SQL & Data Modeling	Fact/Dim Modeling, Performance, Data Quality
Generative AI / RAG	Amazon Bedrock, LangChain, Prompt Engineering
Python	FastAPI, Pandas, PyYAML

Projects

INAAS AI Analytics Agent — Software Engineer

Built an LLM-powered analytics agent with context-aware SQL generation and multi-warehouse support.

Manufacturing Analytics Platform — Data Engineer

Delivered a medallion architecture on EMR Serverless with Redshift data marts and Power BI dashboards.

Synthetic Data Generator CLI — Creator

Created an open-source tool that produces 10K records/minute with validation rules and adapter-based prompt generation.

Education

Kalasalingam Academy of Research and Education — B.Tech Electronics and Communication Engineering