SOHAN REDDY

Data Engineer | Azure Databricks | ETL Pipelines | MS in Data Science

📧 sohan.siddenki@gmail.com | 📱 +1 (904) 861-9116 | 📍 Jacksonville, Florida

🔗 LinkedIn Profile | 💻 GitHub Profile

PROFESSIONAL SUMMARY

Experienced Data Engineer with 3+ years specializing in cloud-based data solutions, ETL pipeline development, and AI-powered analytics. Currently leading data modernization initiatives at D2I Healthcare while leveraging expertise from previous work with Johnson & Johnson. Proven track record of transforming complex data challenges into scalable, efficient solutions that drive business value. Expert in Azure and AWS Databricks ecosystems, with demonstrated success in migrating legacy systems to modern cloud architectures, achieving 30%+ runtime reductions and 40% improvements in data accuracy. Passionate about integrating AI/ML capabilities into data pipelines to enable intelligent, data-driven decision-making across organizations.

PROFESSIONAL EXPERIENCE

Data Engineer

Aug 2025 - Present

Intuceo (via iCube) | D2I Healthcare

New Jersey, United States (Remote)

Leading data modernization and cloud migration initiatives for D2I Healthcare's enterprise data infrastructure.

- Leading migration of on-premise SSIS-based ETL pipelines to modern AWS Databricks architecture, achieving 30%+ runtime reduction through metadata-driven frameworks and optimized data orchestration
- Building reusable PySpark frameworks for extracting, transforming, and loading data from multiple sources into Amazon S3 and Delta tables, improving development efficiency and code maintainability
- Integrating AWS Glue, Lambda, and Step Functions to orchestrate ETL workflows, enabling seamless automation and scalability across batch and near real-time data processing
- Designing metadata-driven ingestion and monitoring frameworks for both batch and near real-time data processing, reducing manual intervention and improving data quality
- Collaborating with business and analytics teams to modernize data access patterns and reduce ETL runtime, directly improving report generation speed and data freshness

Technologies: AWS Databricks, PySpark, Amazon S3, AWS Glue, Lambda, Step Functions, SQL, Delta Lake, SSIS

Data Engineer

Dec 2022 - Jul 2025

Intuceo (via iCube) | Johnson & Johnson

Jacksonville, Florida, United States

Worked on Johnson & Johnson's JJAR (Johnson & Johnson Activity Repository) and Pharma Reporting Automation programs, delivering enterprise-scale data solutions for Fortune 500 pharmaceutical operations.

- Migrated legacy QlikView/NPrinting dashboards to Databricks SQL, fully automating report generation and email delivery, eliminating manual processes and reducing report generation time from hours to minutes
- Developed end-to-end ETL workflows for PASS 1/2/3, ArisG, and PSMF Pharma reporting, ensuring regulatory compliance and data accuracy across multiple regulatory frameworks
- Implemented data aggregation and validation frameworks ensuring data quality and consistency across environments (DEV, UAT, PROD), achieving 40% improvement in data accuracy
- Improved dashboard refresh times and data accuracy by 40%, increasing reporting efficiency and enabling faster business decision-making for global stakeholders
- Partnered with global stakeholders for delivery planning, QA validation, and monthly release automation, coordinating across multiple time zones and teams to ensure seamless deployments

Technologies: Azure Databricks, PySpark, Delta Lake, Hive, Power BI, Python, QlikView, NPrinting, Azure Data Factory

TECHNICAL SKILLS

Data Engineering

- Azure Databricks & AWS Databricks
- Microsoft Azure & Amazon S3
- ETL/ELT Pipelines & Data Architecture
- Azure Data Factory & AWS Glue
- Delta Lake & PySpark
- Hive, SQL Server & Data Modeling
- SSIS to Cloud Migration
- Metadata-Driven Pipelines
- AWS Lambda & Step Functions

Programming & Tools

- Python (pandas, numpy, pyspark, boto3)
- SQL (T-SQL, HiveQL, Spark SQL)
- Git / GitHub / Bitbucket
- Azure DevOps & CI/CD Pipelines
- Jenkins & Ansible Tower
- REST APIs (JSON/XML)
- Jupyter & VS Code
- Shell Scripting (Bash)
- Kafka (Confluent Platform)

Business Intelligence

- Power BI & Databricks SQL
- Dashboard Design & KPI Development
- Data Visualization & Reporting
- DAX & Power Query
- Reporting Automation
- Executive Metrics Tracking

AI & Machine Learning

- Machine Learning (Regression, Classification)
- Predictive Modeling & Statistical Analysis
- Feature Engineering & Model Evaluation
- NLP & Text Classification
- LLM Integration & Prompt Engineering
- Natural Language to SQL Automation
- Model Deployment (MLflow, Databricks ML)
- OpenAI, Claude & Hugging Face APIs

Best Practices & Methodologies

- Data Governance & Data Quality
- Agile & Scrum Methodologies
- Code Review & Version Control
- CI/CD & Release Automation
- Cloud Cost Optimization
- Security & Compliance (HIPAA, GDPR)

EDUCATION

Master of Science in Data Science

Jan 2021 - Dec 2022

George Mason University

Fairfax, Virginia

GPA: 4.0 / 4.0 (Perfect Score)

Relevant Coursework: Machine Learning, Statistical Analysis, Data Visualization, Big Data Analytics, Data Mining, Predictive Modeling

PROFESSIONAL CERTIFICATIONS

Databricks Certified Data Engineer Associate

Issuer: Databricks | Date: 2024

View Credential →

ACADEMIC PROJECTS

New York City Shooting Victims: Analysis of the Most At-Risk Population

Academic research project exploring gun violence patterns in New York City using NYPD Shooting Incident Data (2006–2019). Conducted comprehensive data analysis to identify at-risk populations and inform community health and safety policies.

- Conducted data cleaning, feature extraction, and visualization using R and Python (pandas, matplotlib, seaborn)
- Built demographic and geographic visualizations to identify at-risk populations by age, race, and gender
- Analyzed borough-level trends, temporal patterns, and location-based risk factors
- Developed insights supporting data-informed community health and safety policies

Technologies: R, Python, pandas, matplotlib, seaborn, Statistical Analysis, Data Visualization

ADDITIONAL INFORMATION

- **Work Authorization:** Authorized to work in the United States
- **Location:** Open to remote opportunities and relocation
- **Interests:** Cloud data architecture, AI/ML integration, healthcare analytics, data governance