Nikita Godase

3.6+ Years as Data Engineer
Citiustech

Contact info

9545062612

nikitagodase61299@gmail.com

India, Pune

Education

University of Pune 2017 - 2020
India, Pune

Skills

- SQL
- Python
- Big Data
- Data Warehousing
- Data Mining
- ETL
- Apache Hadoop
- Apache Spark
- NoSQL
- Data Modeling
- Cloud Computing
- Data Visualization
- Data Analysis
- Business Intelligence
- Pyspark
- Data bricks
- Data Pipeline
- Data Migration
- Data Validation
- Azure Blob storage

Professional summary

3.6+ Years as Data Engineer at Citiustech from Feb 2022 to till date

Data Engineer with 3.6+ years of experience in Data Engineering, Big Data technologies and Data Analysis, . Proficient in Python, PySpark, Databricks SQL, Hadoop, Hive , Azure Cloud with expertise in building and managing scalable data pipelines using Azure Data Factory. Skilled in Data Mining, Data Preparation, Data modeling and ETL processes, ensuring high-quality data flow for analytical and business needs. Experienced in handling large datasets, implementing Machine Learning algorithms, and working with cloud platforms for data storage and processing. Strong knowledge of Proof of Concepts (PoC) and gap analysis to drive data-driven solutions for enterprise applications.

Experience

Data Engineer December 2023 - Now
Kaiser Permanente, United States

Project Title: End-to-End Doctor-Patient Transcription Pipeline (Azure),

Domain: Healthcare,

Technical Skill Sets: Azure API Management, Azure Data Lake Gen2, Azure Functions, Azure Databricks, Azure Synapse Analytics, Power BI, Delta Lake, Python, Spark,

Project Overview: Designed and implemented a secure, scalable data pipeline to process doctor-patient conversation transcripts from Whisper/OpenAI, adopted Medallion Architecture (Bronze–Silver–Gold layers) for traceability, quality, and performance, enabled ingestion, transformation, anonymization, and analytics for clinical voice-to-text data, delivered structured, query-ready datasets to

Azure Data Factory	
Azure Data Pipeline	
Azure Logic app	
Azure Data Lake	
AWS | s3 | Glue | RDS | Redshift | SNS | SQS	
Lambda | Step function	
Event Bridge | Cloud watch	
PowerBI | Excel | DAX	
Snowflake	
Git	
Docker	
NoSQL	

Hobbies

Traveling

Cooking

Awards

Best Performer in Team May 2024

Languages

English Hindi Marathi

downstream analytics tools for KPI tracking and research insights,

Outcomes: Real-time and historical tracking of patient care trends, improved decision-making using diagnosis frequency and consultation quality metrics, fully automated and scalable pipeline with plug-and-play integration for new clinics/doctors, HIPAA-compliant PHI/PII anonymization for secure data sharing,

- **Challenges:** Processing unstructured transcripts with overlapping speaker dialogues, ensuring HIPAA compliance with secure storage and masking of sensitive patient data, harmonizing inconsistent and semi-structured metadata from multiple clinic sources,

Data Engineer February 2022 - November 2023

Tibsovo (Pharmaceutical Services), United Kingdom

Project Title: USA – UK Pharma Data Pipeline

Domain: Pharmaceutical

Technical Skill Sets: PySpark, Python, SQL, Pandas, AWS (Lambda, S3, Glue, SNS/SQS, Step Functions, Redshift), ETL

Project Overview

- Built a robust **AWS-based data pipeline** for processing pharmaceutical sales data from multiple external sources.
- Implemented **Medallion Architecture** (Bronze Silver Gold) for data quality, enrichment, and delivery.
- Automated data ingestion, validation, transformation, and delivery to **Amazon Redshift** for analytics and reporting.
- Reduced manual intervention and accelerated decision-making.

Outcomes

- **80%** reduction in manual data handling.
- **40%** improvement in processing time.
- Enabled **real-time broker-wise sales reporting**.
- Handled **1 TB/month** of sales data with scalable, reusable pipelines.

Challenges

- Frequent **schema drift** across multiple broker data formats.
- Processing and optimizing **large file sizes**.
- Integrating **legacy systems** with AWS-native solutions.

- Achieving **cost-efficiency** while scaling.

Solution & Architecture

Bronze Layer – Raw Data Ingestion

1. External API Lambda trigger Raw data stored in S3.
2. S3 event notifications SNS/SQS Lambda for validation; errors stored in error bucket.

Silver Layer – Validation & Enrichment

3. AWS Glue Jobs for transformation, validation, schema mapping, and error logging.

Gold Layer – Aggregation & Delivery

4. Partitioned & enriched datasets stored in S3 by broker/region.

5. Data loaded into Amazon Redshift for analytics & reporting.

Role & Responsibilities

- Designed & developed **ETL workflows** in AWS Glue and PySpark.
- Created PySpark transformation logic and managed **Delta tables** in S3.
- Implemented **query optimizations** (partitioning, caching).
- Managed **schema evolution** and dynamic ingestion workflows.
- Applied **Medallion Architecture** best practices for data layering.