Kishore Reddy G

Senior AWS Data Engineer

4+ Years Experience | AWS | Glue | Spark | Airflow | Redshift | Large-Scale ETL

India | kishore560037@gmail.com | +91-7899978074

PROFESSIONAL SUMMARY

Senior AWS Data Engineer with 4+ years of experience designing and building **scalable ETL and ELT data pipelines** on AWS. Strong expertise in **AWS Glue, Apache Spark (PySpark), Amazon S3 Data Lakes, Amazon Redshift, Apache Airflow, AWS Lambda, and Python/SQL**. Proven experience handling **large-scale batch and near-real-time data processing (1 GB to 500 GB/day)**, implementing **SCD Type 2 historization, data quality validation frameworks, schema evolution, performance tuning, and cost optimization** in regulated **banking and healthcare** environments.

CORE SKILLS

Cloud & Big Data
AWS Glue, Amazon S3, Amazon Redshift, AWS Lambda, Amazon EMR, AWS CloudFormation, Kinesis Firehose

Data Processing & Programming
Apache Spark (PySpark), Python, SQL

Orchestration & Automation
Apache Airflow, AWS Step Functions

CI/CD & DevOps
GitHub Actions, CI/CD Pipelines, Infrastructure as Code

Databases & Warehouses
Amazon Redshift, SAP HANA, Oracle DB

Streaming
Kafka, Kinesis Firehose

Data Engineering Concepts
ETL / ELT Pipelines, Data Lake Architecture, Batch Processing, Near-Real-Time Processing,
SCD Type 2 Historization, Data Validation & Data Quality, Schema Evolution,
Performance Tuning, Cost Optimization, Production Support

PROFESSIONAL EXPERIENCE

Cognizant | Associate Projects | Healthcare Domain
Production system | Team size: 6

Global Healthcare Data Hub – Merck Group

- Designed and developed **AWS Glue ETL pipelines using Apache Spark (PySpark)** to ingest healthcare datasets into an **Amazon S3 data lake** across **Global, APAC, EUCAN, and LATAM** regions
- Implemented **SCD Type 2 historization in Amazon Redshift** using business keys, effective dates, expiry dates, current flags, and soft-delete handling
- Built **raw → curated S3 data lake architecture** with partitioning by ingestion date and business date

- Processed ~**1 GB/day per dataset** with automated **schema evolution handling** using AWS Glue Data Catalog
- Implemented **data quality validation frameworks** including null checks, datatype validation, regex rules, duplicate detection, and record reconciliation, splitting valid and invalid records into separate S3 paths
- Reduced downstream data quality issues by ~45% through proactive validation and controlled schema changes
- Designed **Amazon Redshift CURRENT and HISTORY tables/views**, optimized with distribution styles, sort keys, and compression encoding
- Orchestrated batch pipelines using **Apache Airflow** with retries, SLA monitoring, and alerting
- Automated deployments using **GitHub Actions CI/CD pipelines** and **AWS CloudFormation**, reducing release errors by ~70%
- Built a **Lambda-based workflow integrated with an LLM** to generate human-readable summaries for ingestion failures, validation errors, and schema changes, reducing issue triage time by ~40–50%

Ownership

- Owned ingestion framework enhancements and historization logic
- Worked directly with global stakeholders on data quality, schema changes, and release coordination

---

Capgemini | Senior Software Engineer | Banking Domain
Production system | Team size: 4

Core Banking Data Platform – Security Bank Corporation

- Built **batch ETL pipelines** ingesting **millions of banking and transactional records daily** from relational databases and flat files
- Designed **Amazon S3 raw and curated data lake architecture**, improving auditability and regulatory traceability
- Implemented **data validation, deduplication, and schema enforcement**, reducing data defects by ~35%
- Modeled **Amazon Redshift schemas** for MIS and regulatory reporting, reducing report generation time by ~40%
- Automated scheduling, retries, and monitoring using **Apache Airflow**, reducing manual intervention by ~60%
- Delivered data pipelines under strict **banking SLAs, compliance, and reporting requirements**

Data Modernization Platform – Goldman Sachs

- Modernized **20+ legacy COBOL batch jobs** into **cloud-native ETL pipelines on AWS**
- Migrated on-prem **FTP-based ingestion workflows to Amazon S3 landing zones**, eliminating legacy infrastructure dependency
- Designed **dynamic workload routing** based on data volume:
    - **AWS Lambda** for <10 GB
    - **AWS Glue (Spark)** for 10–200 GB
    - **Amazon EMR** for >200 GB
- Processed **1 GB to 500 GB/day** across batch and near-real-time pipelines

- Reduced ETL infrastructure cost by ~30% through compute right-sizing and removal of always-on servers
- Improved end-to-end pipeline runtime by ~50%, reducing batch windows from hours to minutes
- Enabled analytics-ready datasets in **Amazon Redshift**, consumed by **100+ BI and data science users**
- Implemented **Apache Airflow orchestration** with retries and alerts, reducing job failure incidents by ~40%
- Automated deployments using **GitHub Actions CI/CD pipelines** and **AWS CloudFormation**, cutting deployment time from days to <30 minutes

Ownership

- End-to-end responsibility for architecture, development, deployment, and optimization
- Primary technical contact for modernization initiatives and stakeholder discussions

---

EDUCATION

Bachelor of Technology (B.Tech) – Electronics and Communications
Madanapalle Institute of Technology & Science (MITS), JNTU Anantapur
Graduated: July 2021

Diploma – Engineering
Loyola Polytechnic, SBTET
Graduated: July 2018

Secondary School (SSC)
Surya E.M High School
Completed: July 2015

CERTIFICATIONS

- **AWS Certified Data Engineer**
- **AWS Certified Cloud Practitioner**
