Rojan Adhikari
Senior Data Engineer
rojan.adhikari23@gmail.com | +1 (512) 882-7559 | https://www.linkedin.com/in/rojan998/

Professional Summary

- 6+ years of data engineering experience, building cloud-native data platforms, backend services, and real-time processing solutions across AWS, Azure, and GCP.
- Design and support enterprise data fabric and lake house platforms using AWS, Apache Flink, Kafka, Apache Hudi, Lake Formation, Terraform, Kubernetes, and Python.
- Hands-on experience engineering and troubleshooting distributed streaming systems, cloud-native data services, infrastructure automation, security controls, and production observability.
- Engineered real-time streaming pipelines using Kafka/Kinesis, Apache Flink, Spark Structured Streaming, and Hudi CDC with sub-second latency, enabling fraud detection, loan processing, and distributed event-driven systems.
- Developed Terraform Enterprise IaC for EKS clusters, Glue jobs, and RDS, with GitHub Actions CI/CD pipelines, reducing deployment time by 85% and enabling fail-fast development cycles.
- Mentored 3 junior data engineers on a modern data stack (dbt, Airflow, Delta Lake), conducted weekly code reviews, and established CI/CD standards adopted across 3 engineering teams.

Technical Skills

- **Cloud Platforms**: AWS (S3, Lake Formation, Glue Data Catalog, Athena, API Gateway, Lambda, Step Functions, Secrets Manager, EventBridge, EKS, EMR, Redshift, RDS, IAM, CloudWatch, CloudTrail, SNS, SQS, Secrets Manager, Kinesis Data Streams) Azure, GCP
- **Programming Languages & Libraries**: Python, SQL (T-SQL, PL/SQL), PySpark, Bash Scripting, Pandas
- **Big Data & Streaming**: Apache Flink, Apache Kafka, Apache Spark (PySpark, Spark Structured Streaming), Apache Hudi, Apache Iceberg, Debezium, Delta Lake, CDC, Batch and Stream Processing, Apache Airflow (DAG Development), Delta Sharing, Kinesis Data Streams
- **Data Orchestration & Workflow Management**: Apache Airflow, dbt (Data Build Tool), AWS Step Functions
- **Data Architecture & Modeling**: Medallion Architecture, Star Schema, Snowflake Schema, Slowly Changing Dimensions (SCD Type 2), Dimensional Modeling, Parquet, JSON
- **Databases & Data Warehouses**: Snowflake, Amazon Redshift, PostgreSQL, SQL Server, MongoDB
- **Data Quality & Observability**: Grafana, Prometheus, Grafana Alloy, Mimir, Root-Cause Analysis, Great Expectations, Datadog, CloudWatch, Azure Monitor, PyTest, Data Profiling, Alerting.
- **Security & Compliance**: IAM, AWS Lake Formation, KMS, Vault, AWS Secrets Manager, OAuth2, RBAC, TLS/SSL, HIPAA, PCI-DSS, GDPR.
- **Infrastructure as Code & DevOps**: Terraform Enterprise, Docker (Docker Ranch), Kubernetes (AKS, EKS), Helm Charts, GitHub Actions, Jenkins, CI/CD Pipelines
- **Data Cataloging & Governance**: Azure Purview, AWS Glue Data Catalog, Data Lineage Tracking, Metadata Management, Data Governance Framework
- **MLOps & AI**: MLflow, SageMaker Feature Store, Vertex AI, EvidentlyAI, Prometheus, Data Contracts, Feature Pipelines, DataOps, FastAPI, Model Monitoring, Model Drift Detection
- **Business Intelligence & Visualization**: Tableau, Power BI, Looker, Looker Studio, AWS Athena
- **Collaboration & PM Tools**: Git, GitHub, Jira, Confluence, Agile/Scrum Methodologies

Professional Experience

Client: Duke Energy – North Carolina (Energy and Utilities)
Sr. Data Engineer | Nov 2025 – Present

- Design and support Duke Energy's enterprise Data Fabric, an AWS-based ecosystem of data platforms and products serving batch and real-time analytics workloads.
- Engineer and troubleshoot stateful streaming pipelines using Apache Flink, Kafka, Debezium, and Apache Hudi, resolving checkpoint failures, backpressure, duplicate records, and data-consistency issues.
- Develop and maintain Terraform Enterprise infrastructure-as-code for AWS data services, EKS workloads, environment configuration, IAM policies, and reusable platform components.
- Build and enhance Python backend services and event-driven integrations using AWS Lambda, SNS, EventBridge, S3, Glue, KMS, and Secrets Manager.
- Create and improve GitHub Actions CI/CD workflows to build, validate, publish, and deploy containerized data-platform services across environments.
- Deploy and operate cloud-native workloads on Amazon EKS using Kubernetes, Docker, Helm, ArgoCD, and k9s, investigating image, configuration, and runtime failures.
- Implement platform observability and alerting with Grafana, Prometheus, Grafana Alloy, and Mimir for Flink jobs, Kubernetes resources, application health, and production incidents.
- Strengthen data security and governance through least-privilege IAM, KMS encryption, Lake Formation permissions, cross-account access controls, and secrets management.
- Collaborate with architects, product owners, application teams, and infrastructure engineers on technical design, code reviews, production deployments, root-cause investigations, and continuous platform improvements.
- Produce operational runbooks and technical documentation that enable product teams to safely manage Glue tables, schemas, permissions, and other data-platform resources.

Client: MidFirst Bank – Edmond, OK (Financial)
Sr. Data Engineer | Oct 2024 – Oct 2025

- Contributed to the modernization of the bank's data platform by migrating legacy SQL Server and Oracle data pipelines to an Azure-based lakehouse architecture using Azure Data Lake Storage, Databricks, Delta Lake, and Snowflake.
- Designed batch and incremental ingestion pipelines using Azure Data Factory, PySpark, and SQL to process customers, account, transaction, loan, and operational data from multiple banking systems.
- Developed bronze, silver, and gold data layers in Delta Lake, applying schema validation, deduplication, data standardization, and business transformations before publishing curated datasets for analytics.
- Implemented change data capture and incremental-loading patterns using source timestamps, merge operations, and Delta Lake upserts, reducing unnecessary full-table processing and improving data freshness.
- Built PySpark and SQL transformation jobs in Databricks to cleanse, join, and aggregate high-volume financial datasets used by risk, lending, finance, and regulatory-reporting teams.
- Developed dimensional data models in Snowflake, including fact and dimension tables with Slowly Changing Dimension Type 2 logic, to support historical reporting and customer and account analytics.
- Created and maintained Apache Airflow workflows to orchestrate ingestion, transformation, data-quality validation, and downstream publishing processes with retries, dependencies, and failure notifications.
- Implemented automated data-quality checks for schema conformity, null values, duplicate records, referential integrity, and source-to-target reconciliation, helping identify data issues before they reached reporting layers.
- Applied banking-data security controls using Azure RBAC, managed identities, Key Vault, Snowflake role-based access, encryption, and least-privilege permissions to protect sensitive customer and financial information.
- Developed reusable Terraform modules and GitHub Actions workflows to provision cloud resources, validate infrastructure changes, and promote data-pipeline configurations across development, testing, and production environments.
- Monitored pipeline execution, Spark workloads, data freshness, and job failures using Azure Monitor, Databricks logs, and operational dashboards; investigated production incidents and documented root causes and corrective actions.

- Partnered with business analysts, data architects, application teams, and governance stakeholders to translate reporting and data requirements into maintainable data models, pipelines, and technical documentation.
- Participated in Agile ceremonies, design discussions, pull-request reviews, release planning, and production deployments while supporting and mentoring engineers on PySpark, SQL, and pipeline-development practices.

Client: Elara Caring - Addison, TX (Healthcare)
Data Engineer | Sept 2023 – Oct 2024

- Contributed to the development of a HIPAA-compliant healthcare data platform using Azure Data Factory, Azure Data Lake Storage, Databricks, Delta Lake, and Azure Synapse Analytics.
- Built batch and incremental ingestion pipelines to integrate patient, clinical, claims, provider, and operational data from EHR systems, relational databases, APIs, and secure file transfers.
- Developed PySpark transformation jobs in Azure Databricks to cleanse, standardize, deduplicate, and enrich healthcare data before publishing curated datasets for analytics and reporting.
- Designed Bronze, Silver, and Gold data layers in Delta Lake, separating raw clinical data from validated and business-ready datasets while supporting schema evolution and historical processing.
- Parsed and transformed HL7 and FHIR healthcare messages using Python, PySpark, and JSON processing, converting nested clinical records into standardized structures for downstream consumption.
- Implemented incremental data-processing patterns using watermark columns, change timestamps, and Delta Lake merge operations to improve data freshness and avoid unnecessary full-table reloads.
- Created dimensional data models in Azure Synapse for patient outcomes, episodes of care, provider performance, and operational reporting, including historical tracking for changing patient and provider attributes.
- Developed and maintained Azure Data Factory workflows to orchestrate ingestion, Databricks transformations, data-quality validation, and downstream publishing with dependency management, retries, and failure notifications.
- Implemented data-quality checks for required fields, duplicate patient records, invalid clinical codes, schema changes, referential integrity, and source-to-target reconciliation before data reached reporting systems.
- Protected sensitive patient information through Azure RBAC, managed identities, Key Vault, encryption, private endpoints, and role-based access controls aligned with HIPAA and least-privilege requirements.
- Implemented metadata management and data-lineage capabilities using Microsoft Purview, helping teams understand the movement and transformation of sensitive healthcare data across the platform.
- Monitored pipeline execution, data freshness, Databricks workloads, and processing failures using Azure Monitor, Log Analytics, and operational dashboards; investigated incidents and documented root causes.
- Used Terraform and CI/CD workflows to manage cloud resources and deploy notebooks, pipeline configurations, and infrastructure changes consistently across development, testing, and production environments.
- Collaborated with clinical analysts, application teams, data architects, security teams, and business stakeholders to translate healthcare reporting requirements into scalable pipelines and trusted datasets.

Client: Humana - Louisville, KY (Health Insurance)
Data Engineer | Feb 2021 – July 2023

- Contributed to the development of an AWS-based healthcare data platform supporting claims, member enrollment, provider, authorization, and clinical reporting workloads.
- Built batch ingestion pipelines using AWS Glue, Python, and SQL to process data from relational databases, partner file feeds, APIs, and application-generated files into Amazon S3.
- Developed PySpark transformation jobs using AWS Glue and EMR to cleanse, standardize, deduplicate, and enrich healthcare data before publishing curated Parquet datasets for downstream analytics.
- Organized datasets into raw, validated, and curated S3 layers with consistent partitioning, naming conventions, and retention policies, improving maintainability and query efficiency.
- Implemented incremental data-loading patterns using processing timestamps, source-system keys, and change indicators to reduce repeated processing of historical claims and enrollment data.

- Designed fact and dimension tables in Amazon Redshift for claims, member enrollment, providers, and authorizations, including Slowly Changing Dimension Type 2 logic for historical attribute tracking.
- Developed and maintained Apache Airflow workflows to orchestrate ingestion, transformation, data-quality validation, and Redshift loading processes with dependencies, retries, and operational notifications.
- Implemented data-quality checks for duplicate claims, missing member identifiers, invalid dates, code-format violations, referential integrity, and source-to-target record reconciliation.
- Optimized S3 and Redshift workloads through partition pruning, appropriate file sizing, compression, distribution styles, and sort keys to improve recurring analytics and reporting performance.
- Applied security controls using AWS IAM roles, KMS encryption, S3 bucket policies, Secrets Manager, and role-based Redshift access to protect member and claims information in accordance with HIPAA requirements.
- Created CloudWatch logs, metrics, and alerts for failed jobs, delayed data, and processing errors; investigated production issues and documented remediation steps for recurring failures.
- Used Terraform and CI/CD pipelines to provision AWS resources and deploy Glue jobs, Airflow configurations, and supporting infrastructure consistently across environments.
- Developed curated datasets and SQL views used by Tableau and Power BI dashboards for claims utilization, member enrollment, provider performance, and operational reporting.
- Collaborated with analysts, application teams, healthcare-domain specialists, and senior engineers to clarify source-system behavior, validate business rules, and deliver reliable datasets for reporting.

Education

Master of Science in Computational Science – University of Central Oklahoma, Edmond, Oklahoma