Rithik K

rithreddy4@gmail.com | (940)-843-3349 | New Jersey |
linkedin.com/in/rithik-k-/ | github.com/Rithik143-coder

Professional Summary

I'm a certified AWS Data Engineer with over 2 years of experience working on cloud-based data projects. I've helped improve and modernize old data systems, set up both real-time and scheduled data workflows, and made sure data is clean, reliable, and ready for reporting. I'm comfortable working with tools like Python, SQL, AWS Glue, Redshift, Apache NiFi, and Databricks, and I enjoy solving problems that help teams work smarter and make better decisions.

Technical Skills

Programming & Scripting: Python, SQL, PySpark.
Data Processing & ETL: Apache Spark, Apache NiFi, Apache Kafka, Airflow, Talend.
Data Warehousing: Amazon Redshift, Snowflake, PostgreSQL, MySQL, MongoDB.
Data Integration & Pipelines: AWS Glue, Step Functions, Informatica, SSIS.
Databases: Relational (PostgreSQL, MySQL), NoSQL (MongoDB, Cassandra).
DevOps & Automation: Git, Jenkins, Docker.
Monitoring & Logging: Prometheus, Grafana.
Data Visualization: Power BI, Tableau.

Certifications

AWS Certified Data Engineer – Associate Jan 2025 - Jan 2028
Amazon Web Services *View Credential*

Education

University of North Texas Denton, TX
Master's in Information Systems and Technology *Aug 2022 - May 2024*

Jawaharlal Nehru Technological University Hyderabad Hyderabad, India
Bachelor's in Electronics and Communication Engineering *June 2017 - July 2021*

Projects

AWS Redshift Migration Pipeline | *Glue, Lambda, Step Functions*
GitHub: aws-redshift-migration-pipeline

- Set up a sample Redshift migration using AWS Glue, Lambda, Step Functions, and NiFi based on my work at Blue Shield of California. The project handled over 5TB of data with minimal downtime.

Real-Time Retail Streaming Pipeline | *Kafka, Airflow, Spark*
GitHub: real-time-streaming-kafka-airflow

- Put together a real-time retail data flow using Kafka, Airflow, and Spark to process over 100,000 order events per day, based on the kind of streaming work I was part of at Infosys.

Experience

Blue Shield of California CA, USA
Data Engineer *March 2024 - Present*

- Led the successful migration of a legacy on-premise data warehouse to Amazon Redshift with less than 1 hour of downtime, ensuring uninterrupted business operations and smooth transition.

- Created ETL pipelines using AWS Glue, Apache NiFi, and Informatica to integrate data from over 15 internal and external healthcare sources, including claims, provider networks, and member records.
- Handled data cleansing and validation using Talend, AWS DataBrew, Glue Data Catalog, and Python scripts, reducing post-migration issues by 40% and improving overall data governance.
- Built and maintained CI/CD pipelines using GitLab, Apache Airflow, and Prometheus, accelerating deployment cycles and minimizing manual errors.
- Worked on end-to-end data workflows using AWS Lambda, Step Functions, and Apache Spark to manage data collection, processing, and transformation efficiently.
- Ran real-time and batch data pipelines on Amazon EMR using PySpark to handle large volumes of data and make it available 60% faster for reporting and analysis.
- Improved Amazon Redshift query performance by up to 50% by redesigning schema, applying optimized partitioning and indexing strategies, and fine-tuning complex SQL queries.
- Helped prepare AI-ready datasets by organizing and labeling over 50 million rows of structured health data for use in downstream predictive models focused on patient care and risk scoring.
- Supported data science teams by ensuring clean and accessible data pipelines for training LLM-based internal tools and generative dashboards within the analytics environment.
- Strengthened security posture by configuring IAM policies, KMS encryption, and VPC-based access controls to protect data during and after migration.

Infosys Hyd, India

Data Engineer *Jan 2021 - Aug 2022*

- Set up and managed Databricks, JupyterHub, PySpark, PostgreSQL, Prometheus, and Grafana to support data operations and real-time monitoring, improving system reliability by 40% and reducing downtime alerts by 30%.
- Used Apache NiFi and Apache Airflow to create ETL pipelines that integrated structured and semi-structured data from over 10 sources, including product catalogs, customer transactions, and website activity logs.
- Developed a centralized data warehouse on Azure SQL Data Warehouse, enhancing reporting capabilities, cross-team collaboration, and governance compliance.
- Built real-time streaming pipelines using Apache Kafka and Airflow, enabling low-latency ingestion with high data throughput.
- Created and deployed machine learning models for demand forecasting using Scikit-learn and TensorFlow, improving forecasting accuracy and business agility.
- Processed high-volume data with Apache Spark and Hadoop, cutting pipeline runtimes by 45% and supporting near real-time analytics for faster decision-making..
- Made 8+ easy to understand dashboards in Tableau and Power BI to help business and engineering teams get the insights they needed for reporting and decision-making.
- Ran data applications in Docker and managed deployments with Kubernetes on Azure (AKS) to ensure consistent environments and support scalable processing.
- Helped run data workflows automatically using CI/CD pipelines with Git and Jenkins, which made deployments faster and more consistent.

Academic Project

Netflix Dataset Analysis | *Python / Pandas / Matplotlib*
GitHub: netflix-data-analysis

- Looked into a Netflix dataset from Kaggle to find 10 insights about viewer behavior. Cleaned and visualized the data using Python libraries like Pandas and Matplotlib, and applied basic machine learning to understand viewing patterns.