Lohit Marla

+1 8602347561 | lohitmarla@gmail.com | GitHub | LinkedIn | Portfolio | Willimantic, United States

Education

Master's in Data Science**,** 3.8 / 4.0 GPA**, University of Connecticut** December 2024

Bachelor of Technology, Computer Science and Engineering, 8.82 / 10 GPA, **GITAM University** August 2020

Skills

Programming: Python, Scala, C++, Pyspark, SQL, Kafka, Airflow, GitHub, ETL, OLTP, OLAP, Linux
Data Engineering: Data Warehousing, Data Lakehouse, Data Mesh, Data Mart, Data Management, RDMS
Cloud Computing: Athena, IAM, Kinesis, Event Bridge, RDS, Terraform, Docker, Jenkins, Databricks, Snowflake
Data Analytics: R, Pandas, NumPy, TensorFlow, Keras, Power BI, Data Mining, Statistical Modeling, Machine Learning
Certifications: AWS CCP, Google IT Automation with Python Specialization, Certification for Apache Airflow, SQL Badge, Python – HackerRank, Data Engineer Certification – Data Camp, Databricks Generative AI Fundamentals

Work Experience

Data Engineer, Parker Dewey May 2024 – October 2024

- Automated the end-to-end ML model pipeline using data orchestration tool Airflow, reducing manual intervention by 80%. The pipeline handled monthly data collection, feature engineering, model training, improving real-time analytics.
- Architected and deployed an end-to-end data pipeline to collect data from 8 government sources via APIs using Microsoft Power Connectors, ensuring seamless integration between the data sources and the PowerApps.
- Updated 15 tables and integrated with Power BI, to provide low latency insights and improve data accessibility for decision-making, increasing reporting accuracy by 50% based on requirements from business stakeholders.
- Automated Spark-based data transformation, validation systems for data quality checks, data profiling and presented the metrics by data visualization tool Tableau to the technical stakeholders.

Data Engineer, HexStream June 2024 – August 2024

- Led team with extensive data cleansing, addressing gaps and inconsistencies in raw datasets through imputation, data enrichment which enhanced the accuracy and reliability of the forecasts by 35% in Power BI.
- Engineered an AI automated classification system for PDF documents using Textract and Sagemaker, reducing processing time by 85% and eliminating manual effort using a data Lakehouse architecture.

Data Engineer, Novus Platform September 2019 - July 2023

- **Devised Data Pipeline Architectures:** Independently designed, developed and tested data solutions, scalable ETL pipelines by Pyspark, scala, achieving 99.9% data availability on the Redshift data warehouse, s3 Data Lake.
- **Redesign Infrastructure for Greater Scalability:** Formulated serverless API unstructured data processing using Appflow, SQS, SNS, and NOSQL DynamoDB by deploying using Terraform, boosting efficiency by 50%.
- **Streamlined Ingestion:** Launched Incremental Data Ingestion Glue Pipelines from RDMS, integrating 50,000+ daily transactional records in parquet, created Athena tables which resulted in a 35% reduction in processing time.
- **Big Data Pipeline Optimization:** Assembled large, complex data sets using spark, reducing runtime from 8 hours to under 50 minutes by performance tuning, enhancing data delivery efficiency to meet SLAs of users.
- **Data Integration Projects:** Collaborated with cross-functional teams with effective communication via Jira for requirements gathering, developed CI/CD pipelines in SDLC and initiated code reviews in Git for code quality.
- **Provided Technical Documentation:** Functioned as a data engineering SME under the Agile methodology, guided data ops team members on best practices in Airflow by Confluence, reducing scheduler issues by 60%.
- **Optimized SQL Queries:** Conducted RCA, debugging, and troubleshooting production issues by executing and tuning SQL queries, resulting in improved query performance and a 55% reduction in operational costs.
- **Health Care Data Management and Governance**: Enhanced data security, compliance by implementing data masking, profiling for 1M+ PHI/PII records using Batch, Docker, DataBrew, improving data accuracy by 25%.
- **Complex Relationship Data Modeling:** Pioneered sophisticated graph-based data models in AWS Neptune using database design fundamentals to capture complex relationships, enhancing query efficiency by 30%.
- **Kafka-Based Streaming Pipelines**: Built distributed and real-time ETL pipelines using Kafka and AWS MSK, handling messages with AWS Lambda automation, reducing errors by 20% and enhancing data flow visibility.
- **Data Mesh Implementation**: Led cross-team data mesh adoption, enhancing data lineage visibility, cutting cross-team data dependencies by 30% and to support Gen AI solutions development.

Achievements

- Finalist in the 2024 BI&A LDP Case Competition among 400 students, showcased an innovative AI data architecture using AWS for the intelligent data processing, data extraction, ingestion, analytical strategies for insurance industry.
- Secured two bronze medals in World Sprint 13 Data Structures Coding Competitions organized by Hacker Rank.