Yagna Sowmya Sri Chilukuri Data Engineer(AI/ML)

chilukuri@workwebmail.com| (269) 870-6066 | USA | LinkedIn

Summary

Data Engineer with 3+ years of experience in designing and optimizing data pipelines for large-scale processing. Proficient in AWS and Azure, using tools like Apache Spark, Kafka, and Airflow. Experienced in integrating machine learning models (Random Forest, XGBoost) for predictive analytics and fraud detection. Skilled in ETL automation, data governance, and building scalable data infrastructures that enhance data quality and support business decision-making.

Technical Skills

- **Cloud Platforms, Data Storage & Formats**: AWS (Kinesis Data Streams, Glue ETL, EMR, S3, Redshift, CloudWatch), Azure (Data Lake Gen2, Synapse Analytics, Purview), Star-schema modeling
- **Machine Learning** : Linear/Logistic Regression, Decision Trees, Random Forest, XGBoost, SVM, KNN
- **Big Data & Streaming**: Apache Spark (PySpark), Apache Kafka, Azure Databricks
- **Data Orchestration & Automation**: Apache Airflow, AWS CloudWatch, Custom Python alert scripts
- **Programming & Scripting**: Python (Pandas, PySpark), SQL
- **Data visualization:** Power BI, Tableau, Looker
- **Security & Governance**: Azure Purview, Apache Ranger, Role-Based Access Control (RBAC), PII Masking, Compliance with 21 CFR Part 11 and HIPAA
- **Performance Optimization**: Partitioning, Bucketing, Indexing, Query Optimization
  Machine Learning & Tuning: Linear/Logistic Regression, Decision Trees, Random Forest, XGBoost, SVM, KNN, , Cross-validation, GridSearchCV, RandomizedSearchCV
- **Evaluation Metrics**: Precision, Recall, F1-Score, ROC AUC

Professional Experience

Data Engineer, Chime 10/2024 – Present | Remote, USA

- Designed and implemented scalable big data pipelines using AWS Kinesis, AWS Glue ETL, and Apache Kafka, improving data ingestion speed by 60% and reducing latency by 45%.
- Engineered data transformations using Apache Spark on AWS EMR with PySpark for scalable feature extraction, data cleansing, and enrichment, optimizing query performance by 50% through partitioning and bucketing.
- Developed star-schema models in Redshift and normalized data schemas in S3 (Parquet format), enhancing analytic performance for fraud detection analytics.
- Integrated machine learning models (Random Forest, XGBoost) into fraud detection systems and used model evaluation metrics like precision, recall, F1-score, and ROC AUC for performance assessment.
- Optimized model performance using GridSearchCV and RandomizedSearchCV for hyperparameter tuning and used Git and GitHub for version control and collaborative development.

Data Engineer, Merck Sharp & Dohme (MSD) 01/2021 – 07/2023 | Hyderabad, India

- Directed requirement gathering and stakeholder alignment for the Clinical Trial Data Lake Implementation project collaborating with clinical, R&D, and regulatory teams. Defined data ingestion priorities and metadata standards to improve data availability by over 75%.
- Designed and deployed a scalable data lake architecture on Azure Data Lake Gen2 integrating batch and streaming clinical trial data. Optimized storage and retrieval by partitioning and indexing which reduced query latency by 65%.
- Developed ETL pipelines using Apache Spark, Airflow, and Python libraries such as Pandas and PySpark. Automated schema validation and incremental loads for clinical site and EHR data increasing pipeline uptime to above 95%.
- Implemented data governance using Azure Purview and Apache Ranger to enforce role-based access controls and PII masking. Collaborated with compliance and security teams to ensure 100% adherence to 21 CFR Part 11 and HIPAA regulations.
- Migrated key datasets into Azure Synapse Analytics data warehouse and built user-friendly data marts. Conducted UAT with clinical data scientists achieving 90% satisfaction and enabled a 70% rise in self-service analytics adoption through collaborative notebooks.
- Collaborated with cross-functional teams including clinical researchers, IT, and compliance to ensure data solutions met regulatory standards, improved data accessibility by 40%, and supported timely delivery of insights for critical decision-making.

Education

Master of Science in Data Science 08/2023 – 04/2025
Western Michigan University - Kalamazoo, MI, USA
Bachelor of Technology in Information Technology 08/2019 – 06/2023
Vignan Institute of Technology and Science - Hyderabad, India

Certifications

- **APAC-Solutions Architecture Job Simulation**
- **Operations Job Simulation**
- **Generative AI:Introduction and Applications**
- **Foundations: Data, Data, Everywhere**
- **AI For Everyone**