Dharama Teja Samudrala

New York, NY 14221
+18573983456 | dharmatejas102@gmail.com | linkedin | github

Summary

Data Engineer with 3+ years of experience building scalable data pipelines and delivering insights across **healthcare, finance, and insurance** domains. Skilled in **Python, SQL, Spark, ETL, cloud data platforms,** and **natural language processing**. Passionate about leveraging data and AI/ML to drive business value. Proven ability to collaborate with cross-functional teams and communicate technical concepts effectively.

Experience

Citi NewYork,USA
Data Engineer - Financial Analytics Platform March 2025 – Present

- Designed and deployed ETL pipelines using **Airflow, PySpark,** and **AWS Glue**, ingesting data from 10+ sources and reducing manual processing time by 40% for a 5-member team.
- Built streaming infrastructure with **Kafka** and **Flume** to process 500,000 financial transactions per day with sub-second latency, enhancing real-time analytics capabilities.
- Optimized data storage and querying on **Amazon S3, Redshift, Snowflake**, reducing query latency by 30% and costs by 25% through partitioning and compression techniques.
- Implemented CI/CD and monitoring using **Jenkins, AWS CodePipeline, Airflow**, automating data quality checks that caught 90% of anomalies pre-deployment.

Fidelity Investments Raleigh, NC, USA
Data Engineer Intern - Investment Analytics Jun 2024 – Aug 2024

- Optimized **PySpark** financial processing workflows on **AWS EMR**, reducing execution time by 55% for 10 GB datasets while maintaining data integrity.
- Evaluated **AWS Neptune Graph Database** for team workflow analysis, conducting performance benchmarking and data modeling.
- Implemented serverless APIs using **AWS API Gateway** and **Lambda**, enabling secure access to investment analytics for internal clients.
- Developed and maintained **Airflow** DAGs for financial data pipelines, ensuring proper task dependencies and error handling.

MetLife India
Data Engineer Dec 2021 – Jul 2023

- Managed **Amazon S3** buckets storing 2 TB of insurance data; implemented partitioning to optimize **Athena** queries, reducing runtime by 20%.
- Built and maintained **MySQL, PostgreSQL, MongoDB** databases supporting analytics for 120,000 daily customer records.
- Developed **PySpark ETL** jobs to process 3 GB of transactional data per day from multiple **AWS** sources.
- Designed **Airflow** workflows to orchestrate and monitor 8 recurring ETL pipelines, ensuring 99% uptime.

Catalog India
Machine Learning Intern Jan 2021 – Jun 2021

- Conducted distributed data analysis using **PySpark** to validate ML model scalability on large blockchain datasets.
- Developed **Python** and **SQL ETL** pipelines to preprocess raw data for model training workflows.
- Automated model training and evaluation with **Airflow** DAGs and deployed microservices using **FastAPI**.

Technical Skills

Languages:	**Python**, **SQL**, R, **Scala**
Data Engineering:	**Spark**, **Airflow**, Kafka, **Snowflake**, AWS (S3, EMR, Glue, Lambda), Azure, **Docker**
Databases:	**MySQL**, **PostgreSQL**, **MongoDB**, **Redshift**, Cassandra
Machine Learning:	PyTorch, scikit-learn, **Hugging Face**, **Jupyter**, MLflow, FastAPI
Tools & Technologies:	**Git**, **Jenkins**, Prometheus, Grafana, Tableau, Unix/Linux

Projects

Generative AI Content Assistant | Python, PyTorch, Hugging Face, AWS

- Developed a generative AI application to assist content creators using **GPT-3** and **Hugging Face** models, serving as the lead engineer in a 3-member team.
- Fine-tuned language models on domain-specific datasets using **PyTorch** and implemented **text summarization**, **entity extraction**, and **sentiment analysis** pipelines.
- Deployed the application on **AWS EC2** and used **SageMaker** for model hosting, enabling support for 10,000 monthly active users.
- Reduced content creation time by 30% and improved engagement metrics by 20% for users of the AI writing assistant.

Predictive Maintenance Platform | PySpark, Kafka, Flask

- Built a predictive maintenance platform to monitor industrial equipment sensor data using **PySpark**, **Kafka**, and **scikit-learn**, working independently on end-to-end development.
- Ingested real-time sensor data from 1,000 devices using **Kafka** streams and performed feature engineering using **PySpark** on 10 GB datasets.
- Trained machine learning models to predict equipment failures and optimized hyperparameters using **MLflow**, achieving 85% recall.
- Created a web dashboard with **Flask** to visualize real-time predictions and alerts, reducing unplanned downtime by 20%.

Customer Churn Analysis | Snowflake, dbt, Tensorflow

- Conducted churn analysis on a 50 GB customer dataset using **Snowflake**, **dbt**, and **Tensorflow**, collaborating with a data scientist on model development.
- Built **ETL** pipelines in **dbt** to transform raw data into features for model training, ensuring data quality and consistency.
- Trained a **neural network** model using **Tensorflow** to predict churn likelihood, achieving an AUROC of 0.89 on validation data.
- Provided actionable insights to business stakeholders for churn prevention strategies, potentially saving $500K in annual revenue.

Education

Rochester Institute Of Technology Rochester, NY, USA
Master of Science in Computer Science Aug 2023 – Aug 2025