SANKARSH NELLUTLA

(601)-339-9095 | sankarshnellutla12@gmail.com | Linkedin | Portfolio

SKILLS

Languages & Databases: Python, SQL, Redshift, S3, RDS, R, Snowflake
Domain Knowledge: Data Engineering, Cloud Architecture, ETL & Data Pipelines, Machine Learning, Backend Development, REST APIs, Serverless Architecture, Data Orchestration, Real-time Data Processing
Libraries & Frameworks: PySpark, Airflow, XGBoost, Flask, FastAPI, Plotly, Dash, AWS (Lambda, EC2, S3), PyTest

EXPERIENCE

Data Engineer
Community Dreams Foundation, Orlando, Florida February 2025 - Present

- Engineered scalable ETL pipelines using AWS Glue, PySpark, and S3, processing 10–50 GB of client data daily and reducing manual reporting effort.
- Created 15 Airflow DAGs for batch and streaming workflows, cutting deployment time and improving pipeline reliability.
- Introduced schema validation, row-count checks, and failure alerts, reducing production data incidents and ensuring SLA compliance across 50+ rules.
- Tuned PySpark transformations and Glue job configurations (partitioning, executor settings), lowering job runtimes and AWS compute spend, cutting runtimes 40%.

Graduate Research Assistant
University at Buffalo, New York October 2023 - December 2024

- Designed and tuned ML models (XGBoost, Logistic Regression) that boosted prediction performance by 12–15%.
- Performed feature engineering (categorical encoding) to improve generalization and reduce overfitting, creating 20+ features.
- Conducted hyperparameter tuning with GridSearchCV/RandomizedSearchCV, delivering reproducible training pipelines and improving validation 12%.
- Designed interactive dashboards with Plotly and Dash to visualize results and present findings to the research team for 3+ groups.

Software Engineer Intern
Accenture, Hyderabad, India June 2022 - July 2023

- Authored and maintained Python backend libraries and CLI workflows (pandas, NumPy, Click/argparse) to automate core business processes and data validation, processing 1M+ records/day and saving 40% manual effort.
- Refactored PostgreSQL queries and schema (indexing, query refactor, parameterized SQL), reducing average query latency and lowering DB load by 25%.
- Enhanced code quality with PyTest unit/integration suites and GitHub Actions CI, raising automated test coverage to 80% and reducing rollback incidents.

PROJECTS

Real-Time Customer 360 Platform LINK

- Built a real-time Customer-360 capability to provide clean, low-latency profiles for analytics and support.
- Leveraged PySpark Structured Streaming (watermarking, deduplication, 30-min sessionization), SCD Type-2 merges into PostgreSQL, Airflow orchestration, and Great Expectations validations; exposed lookups via FastAPI.
- Results: duplicate events reduced by 95% (measured dedupe rate), median FastAPI lookup latency 200 ms, and manual orchestration effort cut by 80% through automated DAGs.

Retail Demand Forecasting & Replenishment System LINK

- Orchestrated the development of a demand forecasting system utilizing XGBoost models, achieving a 20-25% enhancement in forecast performance compared to the previous baseline model.
- Engineered PySpark feature pipelines, trained XGBoost quantile models (P10 / P50 / P90) tracked in MLflow, orchestrated retraining/serving with Airflow + FastAPI, and applied constrained replenishment via OR-Tools.
- Results (backtesting): forecast performance boosted 20–25% vs baseline (validation metric), simulated stockouts reduced by 30%, and excess inventory lowered by 15%.

Random Meme Generator LINK

- Created Flask backend with Streamlit/React UI for image upload, model selection, and download.
- Trained LSTM/CNN/Transformer caption models on a cleaned dataset of 900,000 memes across 300 templates; LSTM gave best results.
- Optimized inference with batching and caching, achieving 85% user satisfaction in tests and 40% faster processing vs baseline.

EDUCATION

Masters in Data Science January 2025
University at Buffalo, USA
Bachelor of Technology in Computer Science Engineering May 2023
Vellore Institute of Technology, India