SANKARSH NELLUTLA

+1(601)-339-9095 | sankarshnellutla12@gmail.com

SUMMARY

- Passionate about building **data-driven solutions** with 2+ years of experience across internships, focused on **data engineering**, **data science**, **cloud technologies**, and **analytics**.
- Skilled in designing **scalable ETL pipelines** using **Python**, **SQL**, and **AWS** tools, automating workflows for structured and semi-structured data.
- Hands-on with **data extraction**, **cleaning**, **transformation**, and **reporting**, integrating formats like **CSV**, **JSON**, and **SQL** into analytics-ready outputs.
- Experience working on **batch and real-time data processing** using tools like **Pandas**, **NumPy**, **Matplotlib**, and **Scikit-learn** for both analytics and modeling.
- Developed a **Multi-Modal Emotion Detection System** using **OpenCV**, **Librosa**, and deep learning (CNN/LSTM) to classify emotions via facial and audio cues.
- Knowledgeable in **data modeling**, **star/snowflake schemas**, and **cloud-based warehousing** (basic usage of **Snowflake**) for optimized querying and analysis.
- Built **interactive dashboards** and visualizations with **Tableau** and **Excel**, transforming complex data into business insights.
- Comfortable in **Agile teams**, version-controlled environments (**Git**, **GitHub**), and deploying pipeline changes through **CI/CD workflows**.
- Excellent at **documenting technical work**, presenting insights, and collaborating with stakeholders across technical and non-technical teams.
- Actively expanding skill set in **Apache Airflow**, **dbt**, and **PySpark** to integrate modern data stack tools into hands-on projects.

TECHNICAL SKILLS

Category	Skills
Programming Languages	Python (Pandas, NumPy, Scikit-learn, Matplotlib, OpenCV), SQL (MySQL, PostgreSQL), Bash (basic scripting)
ETL & Data Engineering	Python-based ETL, SQL-based data transformation, AWS Glue (beginner), AWS Lambda (event-based triggers), Apache Airflow (learning), dbt (learning)
Cloud Platforms	AWS (S3, Lambda, Glue)
Data Warehousing & Modeling	Dimensional modeling, Star/Snowflake schemas, data ingestion into Snowflake (basic)
Machine Learning & Analytics	Scikit-learn, TensorFlow (basics), model evaluation techniques, exploratory data analysis, OpenCV, Librosa for multi-modal input processing
Visualization & Reporting Tools	Tableau, Seaborn, Matplotlib, Excel dashboards
Version Control / CI-CD	Git, GitHub, GitHub Actions (basic CI/CD setup)
Workflow Orchestration	Exposure to Apache Airflow for DAG creation and job orchestration (learning phase)
Development Environment	Jupyter Notebooks, Google Colab, VS Code
Data Formats & Processing	CSV, JSON, XML (read/parse), basic knowledge of Parquet
Currently Learning	Apache Airflow, dbt, AWS Glue Jobs, Spark (PySpark fundamentals), Snowflake data pipelines

EDUCATION

Masters in Data Science *Aug 23 - Dec 24*
- University at Buffalo, New York, USA *GPA 3.5/4*

Bachelor of Technology in Computer Science Engineering *June 19 - June 23*
- Vellore Institute of Technology, Andhra Pradesh, India *GPA 8.1/10*

PROFESSIONAL EXPERIENCE

Data Engineer Intern
NG Talent Tech, Frisco, Texas *Feb 24 - Dec 24*
- Developed and maintained data pipelines for ingesting and transforming structured and semi-structured datasets using Python and SQL, integrated with AWS services including S3 and Lambda.
- Created modular ETL workflows that extracted raw data from various sources (CSV, JSON), validated schema consistency, and loaded it into centralized cloud storage for analytics.
- Collaborated with cross-functional teams to define data requirements and implemented data modeling strategies using Snowflake (beginner-level) and PostgreSQL for downstream reporting.
- Utilized AWS Glue and basic Apache Airflow DAGs (under supervision) for job orchestration and scheduling of batch processing workflows.
- Built and tested automation scripts to monitor data pipeline health and handle common failure scenarios, reducing manual debugging by 40%.
- Supported the implementation of GitHub Actions for CI/CD automation of data pipeline deployments and configuration management.
- Contributed to internal documentation and onboarding guides for new team members working with AWS and Git-integrated data workflows.
- Gained exposure to modern data stack components including dbt, Airflow, and Snowflake; currently applying learnings in hands-on sandbox projects and production assistance tasks.

Data Science Intern
Data Science Intern, Shiash Info Solutions, Chennai, India *Aug 21 - Dec 22*
- Participated in multiple internal data science projects focused on predictive analytics, classification, and exploratory data analysis using Python and Scikit-learn.
- Led the preprocessing and feature engineering phases of real-world datasets used in developing ML models for customer sentiment classification and health predictions.
- Designed Python pipelines to clean, normalize, and impute missing values in large datasets, improving model accuracy and reducing training time.
- Built machine learning models (Random Forest, Logistic Regression, KNN) with cross-validation techniques to enhance performance and reduce overfitting.
- Generated actionable insights using data visualization libraries (Matplotlib, Seaborn) and presented findings via Tableau dashboards to mentors and project leads.
- Participated in code reviews, version control (Git), and collaborative development using Agile sprints, enhancing both technical and soft skills.
- Documented data workflows, model evaluation strategies, and key metrics for inclusion in client-facing reports and presentations.

CERTIFICATIONS

- Tableau Desktop Specialist
- AWS Certified Cloud Practitioner
- Data Science and Advanced Analytics
- Advanced-Data Analytics Tools.

PROJECT EXPERIENCE

Multi-Modal Emotion Detection System
Python, OpenCV, Librosa, TensorFlow, Keras, Scikit-learn, SQL, Jupyter Notebooks

- Designed and implemented a dual-input machine learning system capable of recognizing human emotions from both facial expressions and speech inputs.
- Used **OpenCV** to detect and preprocess facial images and **Librosa** to extract audio features (MFCCs, Chroma, Spectral Contrast) from speech clips.
- Built and trained parallel deep learning models — CNN for image-based emotion classification and LSTM for sequential audio data — and ensembled their outputs to improve accuracy.
- Developed a preprocessing pipeline for data normalization, noise filtering, and augmentation to improve model generalization.
- Tracked model performance using cross-validation and evaluation metrics like precision, recall, and F1-score; stored experiment logs in SQL for reproducibility.
- Delivered insights through visualizations using Matplotlib and used GitHub for version control and collaboration.

E-Commerce Sales Data Pipeline with AWS & Python
Python, AWS S3, AWS Lambda, PostgreSQL, Pandas, SQL, GitHub Actions, Tableau

- Built an end-to-end data pipeline to automate the ingestion, transformation, and analysis of e-commerce order data using **AWS** and **Python**.
- Collected data in CSV format and stored it in **AWS S3**; used **Lambda** functions to trigger transformation jobs written in Python and load results into **PostgreSQL**.
- Applied robust data cleaning logic (handling nulls, type casting, deduplication), and used parameterized SQL queries to create aggregated reporting tables.
- Integrated **GitHub Actions** to automate deployment and updates of the ETL workflows; added logging and error handling for monitoring.
- Created Tableau dashboards to visualize customer behavior, product trends, and category performance, enabling simulated stakeholder reporting.

YouTube Channel Analytics & Data Modeling
Python, YouTube Data API, MongoDB, PostgreSQL, SQL, Tableau

- Engineered a data pipeline that extracts structured and unstructured data from the **YouTube Data API**, including channel details, video metrics, and engagement statistics.
- Stored raw, unstructured data in **MongoDB** and transformed it into a relational schema in **PostgreSQL** using Python ETL scripts.
- Applied data normalization techniques to create a **star schema**, optimized for querying and dashboard development.
- Developed SQL views and stored procedures for custom metrics like average watch time, top-performing videos, and growth rate.
- Built interactive Tableau dashboards for time-series analysis and comparative performance tracking of multiple channels.

Movie Recommendation System (Hybrid Filtering)
Python, Pandas, Scikit-learn, Streamlit, SQLite, Cosine Similarity, Matrix Factorization

- Developed a content-based and collaborative hybrid recommendation system using the **MovieLens** dataset.
- Implemented a **cosine similarity-based model** using movie metadata (genre, tags) and a user-item collaborative filtering model using matrix factorization.
- Integrated both models to provide personalized and context-aware movie recommendations with real-time filtering and search capabilities.
- Stored metadata and user preferences in **SQLite**, and deployed the app using **Streamlit** to simulate a user-friendly recommendation engine.
- Focused on building a scalable recommendation pipeline with modular scripts, result caching, and top-N ranking for production-like behavior.