Ankith Reddy Avula

☎ +1 (214) 675-8544 ✉ ankithr005@gmail.com 🌐 avulaankith.github.io in linkedin.com/in/avulaankith/

Education

University of Texas Arlington **August 2022 – May 2024**
Master of Science in Computer Science ***GPA - 4.0*** *Arlington, TX*
Courses: Cloud Computing and Big Data, Distributed Systems, Algorithm Design, Software Engineering, Database Systems, Object Oriented Programming, Operating Systems, Compilers, SDLC

IIITDM Kurnool **August 2018 – May 2022**
Bachelor of Technology in Computer Engineering ***GPA - 3.4*** *Kurnool, India*
Courses: Data Structures, Database Systems, Object Oriented Programming, Operating Systems, Compilers

Technical Skills

Languages: C, C++, Java, Python, R, HTML, CSS, JavaScript, SQL, Scala, GoLang
Databases: SQL, MySQL, MongoDB, NoSQL, PostgreSQL
Big Data Tools: Apache (Hadoop, PySpark), Scikit-learn, Pandas, NumPy, SparkSQL, Hive, OpenCV, NLTK, PowerBI, Matplotlib, Seaborn
Machine Learning: Generative AI, Deep Learning, TensorFlow, Pytorch, Keras, LangChain
Web: React, Bootstrap, Flask, Git
Cloud: Azure(ADLS Gen 2, Databricks, Data Factory, SQL, Azure ML, Azure AI Foundry, Azure Function App, VS Code), AWS, Docker, Kubernetes

Experience

August IT (Farmer Mac) **April 2025 – Present**
Software Engineer - Data and ML *Remote*

- Architected a **distributed Lakehouse platform** using **Azure Data Lake Storage (ADLS Gen2), Databricks (Apache Spark), Delta Lake, and Azure Synapse**, processing large-scale structured and semi-structured financial datasets across DEV/UAT/PROD environments.
- Designed and optimized **Spark-based distributed data processing pipelines** using **PySpark and Spark SQL**, tuning partitioning, broadcast joins, shuffle behavior, and Adaptive Query Execution (AQE) to reduce job latency by 35%.
- Built fault-tolerant pipelines in **Azure Data Factory (ADF)** orchestrating **Databricks** jobs with retry logic, checkpointing, and SLA-based monitoring.
- Designed storage-compute separation architecture leveraging **Delta Lake + Synapse Serverless SQL Pools**, enabling **high-concurrency analytics** with sub-second performance.
- Built scalable ingestion pipelines to support **LLM-based Retrieval-Augmented Generation (RAG)** workflows using **Azure AI Search, OpenAI APIs, LangChain, and Azure Functions**.
- Optimized document processing and embedding workflows for large unstructured datasets, improving retrieval accuracy by 40%.
- Developed **backend** services using Python and **Flask**, exposing **RESTful APIs** for querying curated Delta tables and serving ML/LLM outputs to downstream applications.
- Built **serverless microservices** using **Azure Function Apps (Python runtime)** to trigger ingestion workflows, document processing, and embedding generation pipelines.
- Integrated backend services with **React-based frontend** dashboards, enabling real-time visualization of pipeline health, document retrieval results, and model outputs.
- Automated deployments using **Azure DevOps CI/CD pipelines**, implementing environment-specific configuration management.

Astrosoft Technologies **January 2025 – April 2025**
Software Engineer - Data and ML *Remote*

- Designed and optimized distributed **ETL/ELT** pipelines using **Databricks (Apache Spark)**, **PySpark**, **Delta Lake**, and **Azure Data Factory** to process large-scale structured and semi-structured datasets.
- Developed scalable big data solutions leveraging **Spark SQL** and Databricks **Delta Live Tables (DLT)**, improving distributed query performance by 30%.
- Implemented traditional computer vision pipelines using **Histogram of Oriented Gradients (HOG)** for vehicular image analysis, training **SVM** and **Logistic Regression** classifiers (Scikit-learn) for lane and road feature detection.
- Designed and trained **Vision Transformer (ViT)** and hybrid **CNN-Transformer** architectures in **PyTorch**, improving spatial feature modeling compared to handcrafted HOG-based methods.
- Built end-to-end preprocessing pipelines using **OpenCV**, **NumPy**, and **PyTorch**, including normalization, augmentation, and class balancing for vehicular datasets.
- Evaluated classical and deep learning models using **mIoU**, precision, recall, F1-score, and pixel-level accuracy to ensure statistical robustness before deployment.
- Leveraged **GPU-enabled distributed training** environments (Databricks ML Runtime / Azure ML) to scale deep learning experimentation and optimize training throughput.
- Supported ML experiment tracking using **MLflow**, monitoring parameters, metrics, and model artifacts across iterations.
- Enforced governance and secure access using **Unity Catalog**, implementing RBAC and data isolation policies.

University of Texas Arlington **September 2024 – December 2024**
Research Assistant *Arlington, TX*

- Spearheaded the development of a sophisticated Retrieval-Augmented Generation (**RAG**) system designed to enhance the search and retrieval of complex college documentation, including academic records, research publications, and administrative files with PostgreSQL and API Gateway.
- Integrated **LangChain** with multiple college databases and content management systems, facilitating seamless and secure document access for faculty, administrative staff, and students.
- Implemented advanced natural language processing techniques to improve semantic search capabilities, enabling more accurate and context-aware document retrieval.
- Optimized the **RAG** system by fine-tuning large language models and embedding techniques, achieving a 40% reduction in document retrieval time and a 35% increase in search relevance accuracy.

Samsung **May 2021 – November 2021**
Machine Learning Research Intern *Remote*

- Developed an audio source separation model to isolate sound categories (vocals, music, drums, etc) from mixed audio tracks using **Python**, **TensorFlow**, **UNets**, **Auto-Encoders**, and **Librosa**
- Implemented Fourier Transforms to convert audio signals into spectrograms for frequency domain analysis and custom UNet-based extraction.
- Optimized model architecture, hyperparameters, and performance, achieving accurate separation of sound sources across diverse audio inputs.

Ismriti **June 2019 – July 2019**
Data Science Intern *Kanpur, India*

- Developed a real-time facial emotion recognition system that recognizes and classifies the live facial emotion of the user using **Python**, **CNN**, **TensorFlow**, and **OpenCV**
- Solved the data scarcity and class imbalance problem by augmenting the data, dataset creation and increasing the size of samples which have very less representation among other classes to make the model unbiased.
- Integrated the model with OpenCV for dynamic, real-time emotion detection, optimizing for real-world scenarios with improved performance and precision.

Projects

ELECTRICAL SUBSTATION SEGMENTATION | *Python, Pytorch, Attention UNet*

- Ranked **Top 10** in **IEEE–ICETCI 2021** Competition organized by **ISRO** on **'Machine learning-based feature extraction of Electrical Substations from Satellite data'**
- Developed an Attention UNet for the semantic segmentation of large remotely sensed images to extract small objects like electrical substations.
- Enhanced model precision by retraining images with mIoU below a certain threshold (τ), expanding the training dataset with less confident images.

TWITTER SENTIMENT ANALYSIS USING DEEP LEARNING | *Python, Pytorch, Tensorflow, BERT*

- Implemented feature-based learning with BERT, including CNN, LSTM, and BiLSTM, for sentiment analysis on Twitter data and explored different combinations to predict sentiments (positive, negative, neutral, or irrelevant) associated with Twitter entities.
- Handled sentiment analysis dataset, recognizing ”irrelevant” as a distinct category, Collaborated on Jupyter Notebooks with team for testing and experimentation on models.

VISUAL TRANSFORMER FOR MEDICAL IMAGE ANALYSIS | *Python*

- Designed a Visual Transformer-based architecture for tumor detection in MRI scans, improving diagnostic accuracy by 30%.
- Leveraged Azure ML for scalable model training and deployment, integrating with PowerBI for real-time visualization.

TWITTER DATA PIPELINE USING AIRFLOW | *Python, Spark, Docker, Apache Airflow, AWS, Pandas, ETL*

- Developed and maintained a robust data pipeline for ingesting real-time Twitter data, utilizing Twitter API to collect tweets and related metadata, ensuring reliable and continuous data flow for downstream analysis.
- Utilized Apache Airflow to orchestrate and schedule complex ETL workflows, automating the extraction, transformation, and loading (ETL) of Twitter data into a centralized database, enhancing data availability and operational efficiency.
- Employed data processing frameworks to clean, transform, and enrich Twitter data, storing the processed data in a PostgreSQL database.

LARGE SCALE LINEAR REGRESSION USING SPARK | *Scala, Apache Spark, Databricks*

- Implemented closed-form solution for large datasets using distributed matrices, leveraging Spark RDDs and Breeze library for matrix computations.
- Applied the outer-product technique to compute theta for enhanced performance and scalability, utilizing both Scala and DataFrame approaches.
- Executed the algorithm on the Boston Housing Dataset, demonstrating its efficacy in predicting housing prices with substantial improvements in computational efficiency.

UBER DATA ANALYTICS PROJECT ON AWS | *Python, Pyspark, SQL, AWS*

- Created comprehensive data models using Lucid Chart to visualize and organize raw Uber data, facilitating a structured approach to data analysis and ensuring clarity in the data pipeline design
- Authored efficient ETL scripts in Python and leveraged Mage, a modern data pipeline tool, to automate the extraction, transformation, and loading of Uber data on AWS, significantly enhancing data processing efficiency and reliability.
- Executed complex SQL queries to analyze processed Uber data, deriving actionable insights and supporting data-driven decision-making processes.