Rahul D Shetty

+91 9483 913464 | 35rahuldshetty@gmail.com | linkedin.com/in/rahul-d-shetty/ | rahuldshetty.github.io

Technical Skills

Languages:Python, JavaScript, SQL, HTML, CSS, C/C++, Groovy, Bash, Java, Rust, Golang
Frameworks: Flask, FastAPI, Django, ReactJS, Svelte, jQuery, Flutter, Node.js, WordPress, Bootstrap, Material UI
Database: MySQL, PostgreSQL, MongoDB, Firebase, Oracle, SQLite
Developer Tools: Git, VS Code, Visual Studio, Eclipse, CMake, Make, Webpack, Emscripten
DevOps: Jenkins, Docker, Kubernetes, Helm, SonarQube, Kong, Nginx, Grafanna, Prometheus, ELK, Kafka, Redis, Spark
Machine Learning: Keras, Scikit-Learn, OpenCV, Transformers, Tensorflow, Torch, Pandas, NumPy, Matplotlib, Kubeflow, SparkML, H2O AI, Ray, SynapseML
Gen AI: LangChain, OpenAI ChatGPT, Stable Diffusion, DALL E, llama.cpp, LlamaIndex, GPTQ

Experience

VMware

Software Development Engineer II — *Jun 2021 – Present*

- As the **Technical Lead** and **Core Developer** for MLOps:
  - Developed **GPU services** on Kubernetes to boost Machine Learning efficiency.
  - Integrated the platform with Apache Spark for **large-scale** data and ML workloads.
- Initiated end-to-end **Large Language Models (LLMs)** powered solutions:
  - Built a **Q&A Chatbot** using Retrieval Augmented Generation with an internal knowledge base.
  - Automated **Video Content** summarization and chapter creation.
  - Designed a **scalable framework** for LLMs within a private cloud (**LLM as a Service**).
- Developed a **Deep Learning** Information Retrieval model for context-aware searches in the VMware Knowledge Base.

Software Development Engineer I — *Jul 2020 – May 2021*

- Led MLOps design and development on **Kubernetes**.
  - Implemented **CI/CD process in Jenkins** to reduce ML development to deployment cycle from weeks to minutes.
  - Offered self-serve IDEs for **JupyterLab**, **VSCode**, and **RStudio** in the cloud.
  - Created a **No-Code-Low-Code** ML framework and a **Marketplace** for **ML as a Service**.
- Developed a platform for real-time monitoring of Enterprise Data Jobs with a customizable notification framework.
  - Optimized data job updates from **1 hour** to **2 minutes**, achieving near real-time performance.
  - Scaled the notification framework for **large events** using **Apache Kafka**.

Projects

ggml.js | *JavaScript, WebAssembly, Emscripten, LLM, EdgeAI* — GitHub | Live

- Created a **JavaScript library** for running **Language Models** directly in web browsers.
- Compiled C/C++ ML models into **WebAssembly** using Emscripten toolchain.
- Documented and hosted the framework using **docsify**.

Sentiment Analysis Research | *Python, Keras, TensorFlow, NLP, Research Project* — Journal

- Developed and trained a **Deep Learning** neural network model with **CNN and LSTM** architecture for Sentiment Analysis on Indian language Twitter datasets.
- Presented findings at the Artificial Intelligence & Data Engineer 2021 Conference.
- Published the research in the **Springer Journal**.

Image Colorization Research | *Python, Keras, TensorFlow, Computer Vision, OpenCV* — GitHub

- Developed and trained a **Deep Learning** neural network model with stacked CNN using an auto-encoder architecture to convert grayscale images to RGB.

Education

Nitte Mahalinga Adyanthaya Memorial Institute of Technology — **9.70 CGPA**
B.E in Computer Science and Engineering — *2016 – 2020*

Jnanasudha Pre-University College — **96%**
Computer Science — *2014 – 2016*

Sri Sri Ravishankara Vidya Mandir — **10 CGPA**
Class 10 — *2013 – 2014*