SAI VIVEKANAND REDDY

857-465-9966 | vangala.sa@northeastern.edu | Boston, MA | LinkedIn | GitHub | Portfolio Site

About: A problem solver bridging 3 years of professional experience in full-stack development with cutting-edge work in cloud infra and AI, aiming to consistently deliver solutions that improve performance & user experience

Education

Northeastern University, Boston, USA *September 2023 - August 2025*
Master of Science, Computer Software Engineering

- **Relevant Courses:** Data Science Engineering, MLOps, Networks & Cloud Computing, Intelligence Analytics

BITS Pilani, Pilani, India *August 2016 - June 2020*
Bachelor of Engineering, Electronics Engineering

- **Relevant Courses:** OOPs, AI, DBMS, Algorithms, Operating Systems, Neural Networks, Distributed Systems

Technical Skills

Languages & Web Development: Python, Java, Go, SQL, React, HTML5, CSS, Javascript, RESTful APIs
AI/ML: Gen AI, LangChain, LangGraph, NLP, TensorFlow, PyTorch, scikit-learn
Cloud & DevOps: AWS (S3, EC2, Lambda, Route 53, RDS, KMS), GCP, Docker, CI/CD
Frameworks & Tools: Spring, FastAPI, Git, Spark, Linux, Airflow, Kubernetes, MongoDB, MySQL, PostgreSQL

Work Experience

Humanitarians AI (5 mos) Boston, USA
AI Engineer *Jan 2025 - May 2025*

- **Orchestrated CI/CD** pipelines and executed Agile practices with Docker, Kubernetes, and Terraform to deploy VMs with load balancers and auto-scaling, building highly reliable services with 99.9% uptime
- Built a **4-tier architecture** memory management system using ChromaDB for vector embeddings and OpenSearch
- Developed multi-modal document processing engine supporting 10+ file formats with Gemini Vision API integration, **achieving 95%+ accuracy** in OCR and table extraction from images

Blue Yonder (2 yrs) Hyderabad, India
Software Engineer *July 2020 - May 2022*

- Owned the end-to-end resolution of **70+** customer-reported **bugs** annually in a complex **Spring MVC** application, debugging across the **full stack** (Java, JSP, Oracle DB), implementing robust fixes with comprehensive JUnit tests
- Revamped the **data-loading** mechanism in the Demand Workbench by optimizing backend Java logic and Oracle SQL queries to selectively fetch only user-requested measures, reducing initial load times by an estimated **40%**
- Managed **software maintenance** across multiple legacy product versions, collaborating with support teams to replicate and resolve version-specific issues, ensuring continuous stability and support for a diverse global client base
- Enhanced user **productivity by 25%** by re-engineering the 'Save & Calc' feature into discrete Save-Calculate-Reset actions, using in-memory caching to allow planners to rapidly simulate multiple forecast scenarios without DB writes

Projects

Cloudnative Webapp Application

- Built a scalable web application using Java Spring Boot and MySQL with user authentication and email verification, deployed on auto-scaling GCP infra with managed instance groups scaling based on CPU utilization
- Automated complete infrastructure provisioning using Terraform IaC to deploy 15+ GCP resources including VPCs, Cloud SQL, and load balancers, reducing manual deployment time by 100% and ensuring consistent environments
- Designed CI/CD pipeline with GitHub Actions and Packer to automate building, testing, and zero-downtime deployments through rolling updates, reducing deployment time from hours to minutes

RAG Application (with AI Agents)

- Architected a multi-agent (OpenAI API) research system using LangGraph, orchestrating RAG, Web Search, and Academic Paper retrieval agents to synthesize documents' information, web searches, and academic sources
- Automated document processing pipeline using Airflow, Docling & Pinecone for structured extraction and storage
- Built a secure platform using JWT authentication, featuring parallel agent execution and PDF export capabilities

Parallelizing Text-to-Image Generation (using Diffusion)

- Designed and implemented CPU and GPU based parallel processing techniques for text-to-image generation
- Evaluated trade-offs in speed, efficiency, and scalability using MS COCO dataset, and integrated mixed precision training with Distributed Data Parallelism (DDP) in a U-Net-based architecture to accelerate model training