Teja Sunku

Redmond, WA (relocation ok) | +1 (206) 445-3399 | sunkut@outlook.com | LinkedIn

SUMMARY

Experienced AI Infrastructure Engineer with strong proficiency in Python and AWS, led the migration of a core LLM routing service to async architecture, significantly enhancing performance. Successfully managed the transition to Kubernetes for improved service orchestration. Draws on substantial data science background, wishing to leverage skills in efficient data management and deployment resilience to drive key infrastructure projects.

EXPERIENCE

Ashvin AI
Jan 2024 - Jun 2025

Machine Learning, Operations (ML Ops)

- Led migration of the core LLM routing service to an async architecture, increasing throughput by 10x without additional compute. Optimized performance by resolving blocking AWS SDK bottlenecks, delivering an additional 3x throughput gain
- Spearheaded Kubernetes adoption across backend services by converting legacy Python scripts into containerized FastAPI microservices, completing migrations in about 1 week per service. Coordinated cross-functional onboarding to enable same-day deployments and production rollouts, speeding up development
- Designed a multi-repo integration test suite to validate cross-service dependencies, cutting release-time regressions across shared infrastructure by 50% and improving deployment reliability
- Built resilient ingestion and export pipelines for medical systems including Brightree, HDMS, and TIMS, normalizing untyped legacy data into structured formats. Enabled same-week integration with clients using these systems, covering over 80% of prospective partners and accelerating onboarding
- Unified Engineering and DevOps pipelines by redesigning CI/CD processes, cutting container build-to-deploy latency by about 70% through caching and parallelization. Developed a custom FastAPI metric measuring concurrent in-flight requests to replace unreliable CPU/RAM signals and improve Kubernetes autoscaling
- Mentored junior engineers in debugging, AWS, and command-line best practices, accelerating interns' onboarding and enabling them to independently lead significant feature development within their first month

Cyentia Institute
Apr 2022 - Jan 2023

Data Scientist

- Analyzed datasets of 1 million+ rows to derive actionable insight and generate comprehensive reports, collaborating with Marketing to finetune visualizations
- Trained and implemented NLP machine learning models using SpaCy and BERT models to augment CVE-related data feeds with key words and phrases, improving data accuracy and relevance
- Extracted vulnerability descriptions from websites and APIs using R and Python, exporting structured tabular formats to AWS, which enhanced data accessibility and integration for security analysis
- Enabled reproducible testing environments by creating Docker scripts with detailed documentation, and by constructing AWS test environments and configs for validating pre-production code

EDUCATION

UC San Diego
2023 - 2024

Machine Learning Bootcamp

University of Idaho
2017 - 2021

B.S., Statistics & Philosophy

SKILLS

Languages: Python, R, SAS, SQL, Bash, Javascript/Typescript

Tools & Infra: FastAPI, Pytest, AWS, Kubernetes, Github Actions

ML / Data: Pandas, Pytorch, Tensorflow, scikit-learn, SpaCy, OpenAI API

Other: Prompt Engineering, CI/CD, Async programming, Documentation and Testing