Souvik De

India | souvikde.tech@gmail.com | +91 9073302976 | souvikns.de | linkedin | github/Souvikns

Summary

Backend & AI Systems Engineer with 4+ years of experience building production-grade distributed systems and large-scale GenAI applications. Specialized in Retrieval-Augmented Generation (RAG), vector search infrastructure, and LLM orchestration using Go, TypeScript, and Python. Proven track record of shipping high-performance developer tooling (16,000+ downloads), and leading open-source initiatives as an AsyncAPI maintainer and Google Summer of Code mentor.

Experience

Software Engineer II, XANE.AI – Gurgaon July 2025 – Present

- Architected and deployed scalable GenAI pipelines powering Maruti Suzuki's Customer Assistant System, enabling context-aware responses over enterprise-scale knowledge bases.
- Designed and implemented an OCR-to-RAG ingestion pipeline converting unstructured PDFs into structured embeddings indexed in a vector database, enabling persistent knowledge memory for AI agents.
-

Lead Software Engineer, CodeMate.AI – Noida December 2024 – May 2025

- Engineered an enhanced search algorithm for Swagger files, optimizing RAG system performance and search precision.
- Implemented CI/CD pipeline utilizing GitHub Actions to automate application release cycles, enhancing deployment efficiency and reliability.
- Optimized a local server utilized by our VS Code extension, achieving over 30% improvement in startup time and enhancing overall user experience.
- Implemented automated deployment from GitHub Actions to a virtual machine (VM) and configured branch protection rules, improving code quality, deployment stability, and increasing team productivity by eliminating manual deployment efforts.

Software Engineer, Postman – Bangalore Feb 2022 – June 2024

- Developed and maintained a cross-platform CLI integrating official AsyncAPI tooling, providing a unified workflow for developers.
- Created and maintained AsyncAPI Bundler (16,000+ downloads), enabling reliable resolution of complex json ref dependency graphs across specification files.
- Served as maintainer across 3+ open-source projects, reviewing contributions, guiding architectural decisions, and driving roadmap initiatives.
- Mentored contributors under Google Summer of Code (GSoC), guiding projects to production-ready completion.
- Co-led governance initiative to standardize parser tooling architecture across multiple languages, improving maintainability and reducing duplication.

Backend Developer Intern, Mage – Remote Nov 2021 – Feb 2022

- Built microservices in Go to automate cloud infrastructure provisioning using templated configuration systems.
- Developed a React-based low-code interface for generating and provisioning infrastructure on demand.
- Deployed containerized services using Docker and Kubernetes in cloud-native environments.

Projects

GitHub Action to sync GitHub Issues to Notion Database Notion-Board

- Developed a GitHub Action to auto sync and update github issues and it's state with Notion Pages.
- Earned recognition from the open-source community through project stars and valuable user feedback.

- Tools Used: Typescript, GH Action.

FastAPI application that scrapes AsyncAPI documentation to create a RAG system and provides a chat API to answer user queries. AsyncAPI-RAG

- Built data scrapper to scrape AsyncAPI spec and tools documentation from github.
- Used LangChain to chunk Markdown data, generated embeddings with the Nomic embed model from Ollama, stored embeddings in a Qdrant database, and exposed a REST API to handle user queries and return contextual answers.
- Tools Used: Python, qdrant, Langchain, ollama, docker

Technologies

Languages: Javascript, Typescript, Go, Python, SQL

Technologies: NodeJs, FastAPI, Gin, Postgress, Firebase, NextJS, React, Langchain, Oclif

Education

Chandigarh University, BE in Computer Science June 2018 – June 2022

- **Coursework:** Computer Architecture, Comparison of Learning Algorithms, Computational Theory