Deependra Kumar

Software Engineer
Bengaluru, India | P: +91-9685468824 | kdeep05.dg@gmail.com | LinkedIn | GitHub | Portfolio

Software Engineer skilled in scalable backend systems, microservices, and cloud-native solutions on AWS & GCP. skilled in Python (FastAPI), PostgreSQL, Redis, Docker, and Kubernetes with experience in ML pipelines and real-time platforms. Proven ability to optimize performance, cut processing time, and deliver enterprise-scale solutions.

CORE SKILLS

- **Backend Technologies:** Python, FastAPI, PostgreSQL, REST APIs, Redis, Redis Streams, Microservices
- **DevOps & Infrastructure:** Docker, Kubernetes, Git, CI/CD, Logging & Monitoring, Terraform (IaC)
- **Cloud Platforms:** AWS (SageMaker, Lambda, Step Functions, ECS, EC2), Google Cloud Platform (GKE, Monitoring, Logging, Alerting)
- **Tools & Miscellaneous:** GitHub, SQL, Cursor, Vscode, Linux, Windows, k9s, Postman, WSL, Pydantic

PROFESSIONAL EXPERIENCE

NeuralHQ.ai – Bengaluru, India
AI Platform for Customer Support | Nov 2024 – Present

- Built and deployed a real-time AI Agent on GKE using FastAPI, Redis, and PostgreSQL, reducing customer query resolution time by 60% and enabling automated, context-aware conversations.
- Scaled backend to handle 2000+ requests/minute and thousands of concurrent WebSocket connections, ensuring high availability for enterprise clients.
- Optimized Kubernetes deployment with HPA & custom scaling metrics, cutting latency during peak loads by 40%.
- Implemented distributed tracing with OpenTelemetry and set up GCP alerts, proactively detecting anomalies and improving system reliability.
- Automated CI/CD pipelines with GitHub Actions, reducing deployment errors and release time by 30%.

NeuralHQ.ai – Bengaluru, India
ML Inference Pipeline on AWS | May 2024 – Nov 2024

- Architected a serverless ML inference pipeline with S3, Lambda, and Step Functions, cutting manual processing from 2–3 days to under 30 minutes.
- Designed concurrency control to process 500 records/model simultaneously across 7 ONNX models (110M–250M params), balancing speed with AWS cost efficiency.
- Reduced AWS costs by auto-tearing down SageMaker endpoints post-inference, saving ~40% monthly spend.
- Integrated centralized logging in CloudWatch and delivered real-time Slack alerts, improving pipeline observability and issue resolution time by 50%.
- Collaborated with ML engineers and data engineers to ensure seamless integration of ONNX models and preprocessing pipelines.

ACHIEVEMENTS

- Python Gold Badge (HackerRank)
- SQL Silver Badge (HackerRank)
- Problem Solving Gold Badge (HackerRank)
- Best Team Award

EDUCATION

Bachelor of Engineering – Electrical and Electronic Engineering (CGPA 8.1/10)
Shri Vaishnav Institute of Technology and Science, Indore (Madhya Pradesh), India