Deependra Kumar

kdeep05.dg@gmail.com | +91-9685468824 | LinkedIn | GitHub | Portfolio

EXPERIENCE

Software Engineer| BANGLORE/REMOTE

NeuralHQ Technologies | *MAY 2024-PRESENT*

LLM-Powered Customer Support Chatbot- Sean *Nov 2024- PRESENT*

- **System Design & Scalable Backend Development**
  - Built backend architecture from scratch using **FastAPI**, **Redis**, **Redis Streams**, and **PostgreSQL**, optimized to handle up to **2000 requests per minute (RPM)**
  - Designed and implemented a **modular, scalable backend structure** to support complex business logic and performant database operations
  - Translated product designs and wireframes into **production-ready code** using FastAPI and PostgreSQL
  - Developed and deployed microservices with a strong focus on **scalability, modularity, and observability**
- **Cloud-Native Deployment & Infrastructure (GCP)**
  - Set up **Google Kubernetes Engine (GKE)** from scratch and managed **end-to-end deployment** of backend services using **Docker and Kubernetes**
  - Integrated **Redis & Redis Streams** for **real-time message transformation** and conversation state tracking across services
  - Enabled **WebSocket communication** at scale via **Nginx Ingress Controller** in Kubernetes, supporting real-time bidirectional flows
- **Auto-Scaling, Monitoring & Observability**
  - Implemented **custom Horizontal Pod Autoscaling (HPA)** based on **Prometheus metrics**, dynamically scaling WebSocket services based on active connections
  - Established **proactive alerting and anomaly detection** using **GCP Monitoring Alerts**
  - Set up **distributed tracing and logging** using **OpenTelemetry (OTEL)** to monitor service health and debug across microservices
- **CI/CD**
  - Set up **CI/CD pipelines** using **GitHub Actions** for automated testing, building, and deployment

ML inference Pipeline on AWS *MAY 2024-NOV 2024*

- Architected and deployed an end-to-end **ML inference pipeline** on AWS for a retail use case, designed to process **batch JSON inputs derived from CSV files uploaded to S3**.
- Implemented a Lambda-triggered workflow where:
  - Users upload .csv files to **Amazon S3**.
  - A **scheduled Lambda function** converts new CSVs to JSON format and stores them in the s3/input/ folder by **tracking the last run timestamp** stored in s3/timestamp/.
  - A second Lambda function checks for **new JSON files** and triggers the creation of **7 SageMaker endpoints**, each serving a **distinct ONNX model**.
- Managed concurrency using **AWS Step Functions**, with a max concurrency of **2 parallel branches**, each invoking a Lambda function to perform inference on **batches of 500 records concurrently** across all 7 endpoints.
- Inference outputs from each model are stored in s3/temp/. After execution:
  - A Lambda function combines partial results into a single output JSON and moves it to s3/temp/.
  - A subsequent Lambda applies **custom business rules** to the combined output and dumps the final result in s3/output/.
- Designed a **robust cleanup mechanism** using a Lambda function to automatically destroy all 7 SageMaker endpoints after:
  - Successful pipeline completion, or
  - Failure scenarios like **OOM errors**, **runtime exceptions**, or **timeout triggers**.
- Ensured complete **serverless, event-driven automation** with strong emphasis on **cost-efficiency**, **scalability**, and **fault-tolerance** across all pipeline stages.

EDUCATION

SHRI VAISHNAV INSTITUTE OF TECHNOLOGY AND SCIENCE (2015-19)

B.E. in Electrical Engineering
June 2019 **l** Indore, India
CGPA 8.1/10

SKILLS

Backend Technologies:
Python • FastAPI • PostgreSQL • REST APIs • Redis • Redis Streams • Microservices

DevOps & Infrastructure:
Docker • Kubernetes • Git • CI/CD • Logging & Monitoring • Alerting

Cloud Platforms:
AWS (SageMaker, Lambda, Step Functions, ECS) • Google Cloud Platform (GKE, Monitoring, Logging, Alerting)

Operating Systems:
Linux • Windows

Tools & Miscellaneous:
GitHub • SQL • Cursor • Vscode

Soft Skills:
Team Player • Result-Oriented • Ownership & Accountability

Achievements

- Python Gold Badge (*issued by HackerRank*)
- Problem Solving Gold Badge (*issued by HackerRank*)
- SQL Silver Badge (*issued by HackerRank*)

Certificates

- Python (Basic) 08/2023-Present (*issued by HackerRank*)
- Problem Solving (Basic) 08/2023-Present (*issued by HackerRank*)

HONORS AND AWARDS

- **BEST TEAM AWARD l**
  Helped Team achieve Best Team Award **once** attracting client appreciation.

- **K VASUDEVAN AWARD l**
  Received K Vasudevan Award for being among the top 1000 final year students. SVITS, Indore l 2019