Manohar Sai

☎ +1-(475)-307-0240 ✉ manoharsai001@gmail.com in www.linkedin.com/in/manohar-sai-

Profile Summary

Highly skilled Full-Stack Software Engineer with expertise in developing scalable applications, integrating AI/ML solutions, and designing cloud-native architectures. Proven ability to deliver high-impact results, including a 35% reduction in API latency and a 40% boost in prediction accuracy. Passionate about building impactful products and solving complex challenges.

Technical Skills

- **Programming Languages**: Java, JavaScript, Python, C, Go
- **Data & Backend**: RESTful API, PostgreSQL, Oracle, NoSQL, ETL Pipelines, SQL
- **Frameworks**: Spring Boot, Node.js, React, Flask, Django
- **Cloud & DevOps**: AWS(EC2, S3, Lambda), Google Cloud Platform (GCP), Docker, Kubernetes, Terraform
- **Machine Learning & AI**: NLP, MLOps, PyTorch, TensorFlow, LSTM, Scikit-learn
- **Tools**: Visual Studio Code, Postman, Eclipse, Apache Spark
- **CS Fundamentals**: Object-Oriented Design, Algorithms, Data Structures, Problem Solving, System Design, Computer Architecture, Operating Systems, Distributed Systems, Microservices

Professional Experience

Tech Cloud Solutions — **Jan 2020 - Aug 2023**

Software Engineer (Full-Stack & AI/ML) — *Hyderabad, India*

- Engineered scalable backend services using Spring Boot and Flask, achieving a 35% reduction in API latency (resulting in a 15% improvement in user response time) and supporting over 1 million monthly users.
- Led the integration of LSTM-based forecasting and NLP sentiment models, resulting in a 40% increase in prediction accuracy and a 60% reduction in manual operations for content moderation workflows.
- Developed real-time inferencing APIs for Gen AI models using Flask and Docker on AWS Lambda, enabling intelligent customer interactions and reducing support platform response times.
- Designed and implemented distributed data pipelines with Apache Airflow and Spark, decreasing processing latency by 45% and improving content discoverability by enabling analysts to access key data 50% faster.
- Built and deployed containerized ML pipelines for real-time inferencing and model serving, ensuring robustness and scalability across AWS and GCP environments.
- Designed and implemented Redis pub/sub systems for real-time notifications, improving system throughput by 20% and offloading relational databases under high concurrency scenarios.
- Automated deployments with Docker, Kubernetes, Jenkins, and GitHub Actions, cutting release cycles by 50% while maintaining system resilience and observability with CloudWatch.
- Collaborated cross-functionally with data science, product, and infrastructure teams to align architectural decisions with key business objectives and deliver high-impact features.

Projects

Machine Learning-Based Recommender System

- Developed a hybrid recommendation engine (TensorFlow content filtering and PyTorch collaborative filtering) that increased user engagement by 12% in A/B testing.
- Deployed the recommendation engine via a Java REST API for low-latency, real-time personalized recommendations.
- Trained models on structured user data and item embeddings, optimizing performance with batch inference and caching.

Quora Clone Application

- Built a scalable Quora-like platform using React and Golang, emphasizing low-latency RESTful APIs.
- Optimized PostgreSQL database queries with indexing, reducing response time by 20%.
- Implemented JWT-based authentication to enhance security and protect user data.

Real-Time Chat Application

- Developed a chat backend with Spring WebSocket and Redis for low-latency messaging.
- Ensured <50ms message delivery and 99.99% uptime with Kubernetes deployments.
- Implemented auto-reconnect and error-handling logic for a fault-tolerant user experience.

Education

State University of New York at New Paltz — **Aug 2023 - May 2025**

Master of Science, Computer Science

Certifications

- Machine Learning Onramp - MathWorks
- Deep Learning Onramp - MathWorks