Ajay Kumar Reddy Inavolu

+1 (607)313-9336 | inavolu.a@northeastern.edu | LinkedIn | Portfolio | GitHub

PROFESSIONAL SUMMARY

Dedicated Engineer with 2+ years' experience in building scalable, high-performance microservices and real-time data processing solutions. Proficient in designing and deploying distributed cloud and AI platforms, leveraging Java, Spring Boot, Docker, AWS.

WORK EXPERIENCE

Software Engineer, ONE COMMUNITY GLOBAL | San Gabriel, CA **August 2024 – December 2024**

Spring Boot, Next.JS, GraphQL, ELK, Kafka, Docker, AWS

- Engineered a **GraphQL** layer over Spring Boot microservices for an aggregator platform, optimizing data retrieval by consolidating client requests into single queries, reducing average network traffic by **30%**, and enhancing system performance and throughput
- Migrated a monolithic employee management system to containerized microservices using **Spring Boot** with **OAuth 2.0** integration for secure authorization, increasing throughput by **30%** and achieving **95%** test coverage via **JUnit**, **PowerMock**, and **Mockito**
- Established a low-latency concurrency mechanism in **Java 17** for the Employee Management App for real-time task updates, achieving sub-second performance for 200+ concurrent tasks using **Multithreading** and **Kafka**-based asynchronous processing
- Implemented **Jenkins**-based **CI/CD** pipelines for automated deployment and utilized **Terraform** for cloud provisioning, reducing deployment time by **40%**, enabling weekly production deployments, significantly improving team agility and customer satisfaction

Software Engineer, OMDENA | Palo Alto, CA **January 2022 – December 2022**

Spring Boot, Next.js, Redis, Docker, AWS

- Engineered **RESTful**, containerized microservices leveraging **Spring Boot** for a credit scoring platform hosted on **Elastic Kubernetes Service**, enabling user engagement with score attributes, boosting user base by **10%** and optimizing API response times by **20%**
- Implemented a real-time log analytics platform with **Elasticsearch**, **Logstash**, **Kibana**, and **Apache Kafka** for active log aggregation across distributed microservices, improving L3 issue resolution by **30%**
- Leveraged **Redis** & **Express Rate Limiter** to cache DB query requests & session details with effective handling of **OAuth** tokens to grant endpoints access to external applications, reducing the DB read load by **60%**
- Reduced deployment times by **30%** by integrating Jenkins Core Agent into the CI/CD pipeline with SonarQube, optimizing code quality and deployment efficiency on AWS EC2 instances

Deep Learning Intern, INDIAN INSTITUTE OF TECHNOLOGY (IIT) | Indore, India **May 2021 – December 2021**

TensorFlow, Keras, Airflow, PySpark

- Achieved **96%** accuracy fine-tuning a **Faster R-CNN** model in **TensorFlow** and Keras, improving precision by 15% through iterative threshold optimization and hyperparameter tuning for robust human activity detection
- Deployed an automated retraining pipeline with **Airflow**, reducing deployment latency by **34%** and continuously ingesting new sensor data, ensuring seamless model scalability, adaptability, and efficient, reliable production upgrades
- Optimized data ingestion speed by **25%** using a **PySpark**-based **ETL** pipeline, processing over **100GB** of sensor data, significantly expediting analytics workflows and delivering critical insights for faster decision-making
- Leveraged the **ELK** stack for real-time monitoring, slashing error detection time by **30%** and improving system observability, ultimately elevating operational reliability, performance, and significantly enabling rapid troubleshooting capabilities

EDUCATION

Northeastern University | Khoury College of Computer Science **|** Boston, MA **January 2023 – April 2025**

Master of Science in Computer Science **GPA: 3.9/4.0**

Courses: Design Patterns, Algorithms, Database Management Systems, Cloud Computing, Web Development, Mobile Application Dev

ACADEMIC PROJECTS

Retrieval Augmented Generation API Analyzer | (Python, Langchain, Huggingface, Chroma, AWS) **October 2024 – December 2024**

- Developed a RAG pipeline integrating Llama 3.1, LangChain, AWS Titan embedding model for enhanced API insights, reducing search time by **50%**, and seamlessly deployed on AWS Bedrock designed to support **10,000+** concurrent requests

Distributed Graph Analysis | (PageRank, Scala, Java, Spark, MapReduce, AWS) **January 2024 – March 2024**

- Implemented the **PageRank algorithm** using **MapReduce** and **Spark** to efficiently analyze a **1-million-node** dense web graph on an **EMR cluster**, with seamless I/O handling via **S3 buckets**
- Leveraged data lineage tracking for enhanced debugging and performance optimization, reducing execution time by **20%**

TECHNICAL SKILLS

Programming Languages: Java, Python, JavaScript, Scala
Backend Frameworks/APIs: Spring Boot, Node.js, Hibernate, TensorFlow, Microservices, REST, GraphQL
Frontend: React, Next.js, HTML/CSS/SASS
Cloud & Containers: AWS (EC2, Kubernetes, EKS, CloudWatch), Docker, Terraform
Databases: PostgreSQL, MongoDB, Redis, MySQL, Oracle, NoSQL, Elasticsearch
Testing & Tools: JUnit, Mockito, Selenium, Jenkins, Maven, JIRA, IntelliJ, ELK, Postman, PySpark, Git