Harsha Vardhan Reddy Kuncha

harshareddy.kuncha@gmail.com | linkedin.com/in/harshakuncha/ | github.com/Harshakuncha | Portfolio
Tempe, AZ | +1 (623) 302-8447

Experienced Software Engineer with expertise in Java, Spring, Node.js, React, Python, SQL, NoSQL, AWS, Linux and Agile. Skilled in full SDLC, cross-team collaboration, resolving production issues, and leading system migrations. Strong believer in the value of hard work.

EXPERIENCE

Software Developer | Oracle, Remote **Jan 2022 – Jul 2023**

- Engineered end-to-end **payment workflows** using **Node.js**, **Java, Spring Boot** supporting modular transaction lifecycles for construction companies which allow invoice verification, and dispute resolution in a microservices architecture using **AWS EKS**
- Developed interactive user interfaces using **React**, **TypeScript** adhering to WSDL standards, which reduced load times by **90%**.
- Integrated **REST APIs** and **asynchronous** message queues with **Kafka** for inter-service communication, optimizing throughput and decoupling workflows across modules to allow seamless notifications for Omni-channel Textura payments platform.
- Streamlined distributed data processing using **Oracle and DB2** by implementing advanced connection pooling with PgBouncer and query caching strategies, enabling new product features and APIs while ensuring consistent performance and reliability.
- Automated CI/CD pipelines using **GitLab, Jenkins, and Terraform**, adopting **Test Driven Development** practices to meet customer needs that raised test coverage to 90% and decreased production defects by 40%.

Research and Development Engineer | Samsung, India **Jun 2021 – Nov 2021**

- Designed and implemented web application layout with **JavaScript, HTML and CSS**.and ingested **WebSockets** communications with **Node.js** for system health metrics, reducing incident detection time and troubleshooting by 40%
- Architected a **microservices based** platform using **Java Spring Boot** and **Docker** hosted on **AWS EC2** instances, implementing MVC strategies and optimizing job scheduling to enhance the **modularity** and to streamline **feature releases**.
- Developed a **Spring Boot** microservice with **Kafka** for Async event processing and **Redis Caching**, optimizing the throughput to handle **10,000+ req/sec at 200ms average latency** via load balancing and query turning.

Software Engineer | GroWealth, India **Aug 2020 – Jun 2021**

- Migrated monolithic financial systems to a **microservices architecture** and developed **ETL** solutions using snowflake, containerized services with **Docker,** and deployed on **AWS EKS** via **Terraform** (Iac), improving modularity and achieving 99.9% system uptime.
- **Optimized trading-workflow APIs** via code refactoring, SQL-query tuning, integration of **OAuth 2.0** for secure authorization, and in-memory caching reducing average response latency by 40% from 250 ms to 150 ms.
- Drove **product monetization** by building customer usage tracking and automated fee/tax calculation pipelines**,** enforcing **failure-to-pay logic** that reduced revenue leakage by **25%**, and enabling **peer-to-peer transfers**, supporting over **100K transactions**.

EDUCATION

Arizona State University **Tempe, USA**
Master of Science in Computer Science | CGPA: 4/4 **Aug 2023 - May 2025**

Manipal University **Manipal, India**
Bachelor of Technology in Computer Science | CGPA: 3.6/4 **Aug 2018 - Aug 2022**

SKILLS

Languages: Python, Java, Golang, C, C++, JavaScript, TypeScript, Bash, CUDA.
Frameworks: React.js, Angular, HTML5, CSS3, REST, Node.js, Express, Spring Boot, .NET, Flask, Django, FastAPI.
DevOps: AWS (Lambda, IAM, EKS, EC2, S3, Glue), Docker, Kubernetes, Helm, Terraform, Linux, Unix.
Technologies : PostgreSQL, MongoDB, MySQL, DynamoDB, Kafka, RabbitMQ, Redis, Snowflake.
Testing Frameworks: JUnit, Jest, Pytest, Mockito, Selenium, Postman, Swagger (OpenAPI), HTTP, Maven, Prometheus, Git.
Miscellaneous : JSON, XML, OAuth 2.0, JWT, Numpy, Pandas, Scikit-learn, Spark.

PROJECTS

Green AI | Python, Ollama, GPT+. **Feb 2025 - Apr 2025**

- Engineered **AI agents** to analyze and optimize **prompt energy usage** and LLM prompts for **Mistral-7B,** including data ingestion (5,000 prompts), feature extraction (token counts, punctuation density) and preprocessing with pandas and NumPy.

vLORA | LoRA, CUDA, GPU **Jun 2024 - Dec 2024**

- Orchestrated the integration of Segmented Gather Matrix-Vector multiplication **CUDA** kernels of **Punica**, resulting in a remarkable **12x speed** for serving multiple LoRA models, **optimized GPU utilization** and reduced inference latency, enabling high-throughput.