YASHWANTH REDDY DASARI

312-646-8975 | yash1th263@gmail.com | Portfolio | LinkedIn | Github

Software Engineer skilled in distributed systems, backend architecture, and scalable microservices using Java, C++, and Python. Strong foundation in algorithms, data structures, and experience building high-performance services handling millions of requests. Skilled in concurrent programming, database design, and performance optimization for production systems.

EDUCATION

University of Illinois Chicago (*Master of Science in Computer Science) CGPA - 3.9/4* Aug 2023 – May 2025

Vellore Institute of Technology (VIT) (*Bachelor of Science in Computer Science) CGPA - 3.7/4* July 2019 – May 2023

WORK EXPERIENCE

ImpacterAI Inc, *Software Engineer, San Francisco, CA* Aug 2025 – Present

Developed and scaled distributed backend systems enabling multi-tenant orchestration and real-time data processing across AWS and Azure, powering enterprise-grade production workloads with high reliability and global scalability.

- Architected multi-agent orchestration engine powering PersuadioAI, enabling real-time conversational AI for enterprise sales processing 5M+ calls/day with sub-second response times through async I/O, distributed caching, and connection pooling.
- Collaborated with ML teams to design and scale real-time data-processing pipelines using FastAPI, Redis, and Kubernetes, improving system throughput 5× and reducing p99 latency 40% through asynchronous execution and dynamic load balancing
- Optimized a high-throughput ML serving architecture on Azure, reducing memory usage by 68% and improving p99 inference response time by 70% by implementing advanced KV caching and efficient request-batching strategies.
- Built graph-driven analytics pipelines processing 80K+ events/minute, enabling real-time anomaly detection and reducing incident response time by 60% across 12+ production microservices.

University of Illinois at Chicago, *Software Engineer, Chicago, IL* Nov 2023 – Aug 2025

Developed backend microservices and full-stack applications for research platform, focusing on API design, database optimization, and scalable architecture to support high-volume data processing workflows.

- Architected modular RESTful APIs using Java Spring Boot with async processing and connection pooling, handling 6M+ requests/day with sub-second latency using PostgreSQL, Redis, and MongoDB
- Optimized concurrent data processing system by implementing thread-pool management and multi-layer caching strategy, improving throughput 2.5× and reducing query response time by 65% for asynchronous workloads
- Led a 0-to-1 project to build a full-stack semantic search application using MERN and Elasticsearch, transforming a manual query process (that took minutes) into an instant, sub-second search across 62M+ records.
- Optimized high-performance computing (HPC) workflows using parallel computing to distribute AI training pipelines across 24+ GPU cores, re-writing custom CUDA kernels to reduce model training time from 6 hours to 2 hours (a 67% reduction).

SimplyFI Innovations, *Software Engineer 1, India* May 2022 – Aug 2023

Developed backend services and automation tools for fintech applications, focusing on document processing, system monitoring, and deployment automation.

- Built a trade finance automation platform using PostgreSQL and MongoDB to process 3000+ documents/day, reducing manual review time by 90% (from 2 days to 3 hours) via an optimized intelligent document processing workflow.
- Built real-time dashboards for fraud detection system processing 10M+ transactions/day, reducing incident detection time 67% (45→15 min) through anomaly detection, automated alerting, and dashboards using Prometheus and Grafana
- Engineered automated deployment system with custom scripts and CI/CD pipeline orchestration, reducing release cycles from 2 days to 1 hour across 16 microservices while maintaining 99.9% uptime

PROJECTS

DynamicFlow Scheduler : Task Scheduling & Load Balancing

- Built a distributed, fault-tolerant scheduler (Java & Spring Boot), boosting task speed 30% & cutting downtime 50%.
- Implemented a work-stealing algorithm with RabbitMQ for asynchronous job queuing, enhancing system concurrency by 35%.

Crashlens : Traffic Analytics & Visualization

- Built a full-stack platform (Python & React, TypeScript) on AWS, analyzing SQL traffic data to cut potential crash rates 26%.
- Implemented interactive web visualization with CRUD functionalities, D3.js, WebRTC and Mapbox for real-time data insights.

TECHNICAL SKILLS

- **Languages:** Java, Python, Go, C#, C++, JavaScript, TypeScript, SQL, Bash, CUDA.
- **Frameworks:** Spring Boot, FastAPI, Node.js, React.js, gRPC, Hibernate.
- **Tools:** Kafka, Redis, Git, Docker, Kubernetes, NGINX, Elasticsearch, RabbitMQ, Prometheus, Grafana, Spark, Flink.
- **Databases:** PostgreSQL, MongoDB, MySQL, Cassandra, AWS(DynamoDB, S3), GCP BigQuery, Vector Databases.
- **Cloud & DevOps:** AWS, Azure, GCP, Linux, GitHub Actions, CI/CD, Microservices, RESTful APIs.
- **AI/ML:** LangChain, PyTorch, TensorFlow, MLflow, RAG Pipeline, OPENAI & Gemini API's, Mistrail, Hugging Face.
- **Certifications :** AWS Solutions Architect – Associate (2024)