Gaurav Avula

Irving, TX, USA | sudo@gauravavula.com | linkedin.com/in/gauravavula

Summary

Software Developer with 5 years of experience designing and building high-performance distributed systems and RESTful APIs using Java, Python, Spring Boot, and Go. Proven expertise in optimizing API performance by 40%, architecting scalable microservices, implementing machine learning integrations, and managing distributed data processing pipelines across cloud infrastructure.

Education

University of North Carolina Charlotte, MS in Computer Science — Aug 2022 – Dec 2023

- **Coursework:** Cloud Computing for Data Analysis, Software System Design and Implementation, Computer Communication and Networks, Mobile Application Development, Visual Analytics, Intelligent Systems

University at Buffalo, BS in Computer Engineering — Jan 2015 – May 2019

- **Coursework:** Algorithms for Modern Computing Systems, Operating Systems, Software Engineering Concepts, Computer Vision and Image Processing, Data Intensive Computing, Real Time Embedded Systems,

Experience

Software Developer, Delta Cognition – Remote — Feb 2024 – Current

- Architected and deployed scalable microservices in Go and Python for an LLM-powered interview platform, handling 10K+ concurrent requests with load balancing and horizontal scaling.
- Designed RESTful APIs with comprehensive endpoint documentation, implementing rate limiting, caching strategies (Redis), and request validation to ensure system reliability and performance.
- Optimized API response times by 45% through database query optimization, connection pooling, and implementing asynchronous processing for long-running tasks.
- Built distributed batch processing systems using Python and Celery for asynchronous task execution, processing 100K+ jobs daily with fault tolerance and retry mechanisms.
- Implemented OAuth2 and JWT-based authentication systems with refresh token rotation and session management, securing API endpoints across multiple microservices.
- Developed NLP pipelines using Python (SpaCy, Transformers) integrated with backend services to process and analyze unstructured data at scale.

Application Developer, OneIT UNC Charlotte – Charlotte, NC — Aug 2022 – Dec 2023

- Developed Python automation scripts and Java Spring Boot services for inventory management, implementing scheduled batch jobs with error handling and monitoring.
- Built high-performance RESTful APIs using Spring Boot with PostgreSQL, implementing database indexing, query optimization, and caching to handle 5K+ requests per minute.
- Architected microservices for real-time data synchronization across multiple systems, using message queues (RabbitMQ) for asynchronous communication and event-driven architecture.
- Implemented comprehensive API testing using Python (PyTest), JUnit, and integration tests, achieving 85% code coverage across all backend services.
- Designed database schema optimizations and implemented data partitioning strategies to improve query performance for large datasets (10M+ records).

Software Developer, Credit Suisse Group – Raleigh, NC — Aug 2020 – Aug 2022

- Engineered high-performance batch processing systems in Java Spring Boot with multithreading and parallel processing, reducing data processing time from 6 hours to 2 hours for 50M+ records.
- Designed and implemented RESTful APIs in Java Spring Boot serving 15K+ requests per minute, with comprehensive error handling, input validation, and structured logging.
- Optimized database performance through advanced query optimization, proper indexing strategies, and connection pool tuning, reducing query execution time by 65%.

- Built distributed data processing pipelines using Apache Kafka for real-time event streaming, processing 1M+ messages daily with guaranteed delivery and fault tolerance.
- Implemented efficient data merging and deduplication algorithms in Python, reducing data consolidation time by 50% while maintaining data integrity across multiple source systems.
- Addressed critical security vulnerabilities (log4j) by systematically updating dependencies, refactoring legacy code, and implementing automated security scanning in CI/CD pipelines.
- Implemented caching strategies using Redis to reduce database load by 40%, improving API response times and system scalability.

Software Developer, Prosurix Inc. – Buffalo, NY Aug 2019 – Aug 2020

- Developed backend services in Python and Java for mobile applications, implementing RESTful APIs for data synchronization and user management.
- Built Android application features including NFC integration using Android SDK and Java, enabling contactless data exchange functionality.
- Implemented AWS Lambda functions in Python for serverless image processing, integrating TensorFlow and OpenCV for image recognition and classification tasks.
- Developed machine learning pipelines for training and deploying image recognition models, storing processed data and model artifacts in AWS S3 with DynamoDB for metadata management.
- Created automated data processing workflows using Python to handle user-generated content, perform ML inference, and store results for real-time retrieval.

Internships

Software Developer Intern, Global Payments – Atlanta, GA Jun 2023 – Aug 2023

- Upgraded critical dependencies across Spring Boot applications, refactoring code to use latest secure versions and implementing automated dependency scanning.
- Enhanced REST API robustness by implementing comprehensive error handling, input validation, exception logging, and detailed error response formatting in Java Spring Boot.
- Configured CI/CD pipelines using Jenkins and GitHub Actions for automated testing, building, and deployment of Java applications, reducing deployment time by 20%.
- Implemented API documentation using Swagger/OpenAPI specifications, creating comprehensive documentation for all REST endpoints with request/response examples.
- Developed unit and integration tests using JUnit and Mockito to ensure code quality and maintain high test coverage across backend services.

Publications

G. Avula, "Flood Watch: A Multi-Agent System for Smarter Disaster Response," *IEEE eScience*, Chicago, IL, 2025.

G. Avula, "Topology Matters: Evaluating Multi-Agent Organizations for Resilient Flood Detection," *IEEE CCNC*, 2026. (Accepted)

Technologies

Category	**Technologies**
Languages	Python, Java, Go, SQL, Shell Scripting, JavaScript (basic)
Backend	Spring Boot, Flask, FastAPI, Node.js (basic), RESTful APIs, GraphQL
Databases	PostgreSQL, MySQL, MongoDB, DynamoDB, Oracle
Machine Learning	TensorFlow, PyTorch, Scikit-learn, OpenCV, NLP (NLTK, SpaCy, Transformers), Pandas, NumPy
LLM Integration	OpenAI API, Anthropic Claude API, Hugging Face, LangChain, Prompt Engineering
Cloud	AWS (Lambda, EC2, S3, DynamoDB, API Gateway, SageMaker), Azure (basic)
DevOps	Docker, Kubernetes, Jenkins, GitHub Actions, Terraform, CI/CD
Testing	JUnit, PyTest, Mockito, Selenium, Unittest, Postman
Tools	Git, GitHub, IntelliJ IDEA, PyCharm, VS Code, Swagger, Jupyter
Mobile	Android SDK (basic), Java for Android