

Vaibhav Mahalle

8379930292 | vaibhavmahalle95@gmail.com | [linkedin.com/in/vaibhav-mahalle](https://www.linkedin.com/in/vaibhav-mahalle) | github.com/vaibhav-mahalle
Bangalore, India

SUMMARY

Backend engineer with 5+ years building distributed, event-driven Java/Spring Boot microservices on AWS — high-throughput pipelines (3–5M reports, 20M+ emails/year) and a production GenAI/LLM multi-agent chatbot. Strong in HLD/LLD, multi-tenant SaaS, and mentoring.

EDUCATION

Sinhgad College of Engineering, Pune University

Bachelor of Engineering (B.E.) - Information Technology

Pune, Maharashtra

Nov 2020

TECHNICAL SKILLS

Languages: Java, Python, SQL

Core Java: Collections, Streams API, Concurrency, Multithreading

Frameworks: Spring, SpringBoot, JUnit, Mockito, FastAPI

Databases & Messaging: MySQL, Cassandra, Redshift, Redis, ChromaDB, AWS SQS, Kafka, Kinesis Firehose

Cloud Technologies: AWS-S3, SQS, API Gateway, Lambda, EC2, EKS (Kubernetes), Cloudfront, CloudWatch

CI/CD & DevOps: Jenkins, Docker, Prometheus, Grafana

GenAI/LLM: LangGraph, OpenAI & Gemini APIs, RAG (ChromaDB), LLM-as-judge evaluation (LangSmith)

Competencies: DSA, LLD/HLD System Design, Microservices, Distributed Systems, Event-Driven Architecture, RESTful APIs, OAuth 2.0

EXPERIENCE

Bidgely

Software Engineer 2

Nov 2022 – Present

Bangalore, India

– GenAI Energy Assistant — Multi-Agent LLM Chatbot (*Python, FastAPI*)

- * Built Python backend services for a multi-agent GenAI chatbot — a LangGraph supervisor routes queries to specialized agents (usage, bill root-cause) whose LLM tools call internal APIs to ground every answer in live account data.
- * Developed async Python/FastAPI endpoints streaming LLM responses token-by-token over WebSockets/SSE, using asyncio fan-out with semaphore-capped concurrency against rate-limited upstream APIs.
- * Enabled multi-tenant rollout via per-utility pilot configuration with no code forks; served tariff knowledge through RAG over ChromaDB; deployed as Docker containers on AWS EKS with identical dev/prod artifacts.
- * Gated prompt/model changes on a six-evaluator LLM-as-judge evaluation suite integrated with LangSmith.

– Home Energy Report - HER (*Java, Spring Boot*)

- * Developed backend APIs and a Cassandra/MySQL persistence layer in Java (Spring Boot) for a scalable, distributed report-generation platform producing 3-5M paper reports/year across multiple utility partners with 30+ composable sections.
- * Led a team of 3 engineers, mentored junior developers, and drove JUnit 5/Mockito-backed code reviews for on-time, high-quality releases.
- * Built the PDF generation microservice using headless Chrome via CDP4, with an event-driven AWS SQS daemon processing normal and priority queues; automated builds and deployments via Jenkins CI/CD.
- * Implemented multi-threaded report generation using ExecutorService and CompletableFuture, Redis distributed locking, and Apache PDFBox for form population before persisting to AWS S3.
- * Integrated QR codes into paper reports via an AWS Lambda image-generation call for paper-to-digital tracking, streaming scan events via Kinesis Firehose into Redshift.

– Email Notification System (*Java, Spring Boot*)

- * Designed and built a scalable, event-driven email notification microservice in Java (Spring Boot) delivering 20M+ emails/year across NA, EU, Canada, and Japan.
- * Developed a plugin-based section composition engine featuring 15+ injectable content providers (cost widgets, neighborhood comparisons, appliance breakdowns, energy tips).
- * Engineered the delivery pipeline using AWS SQS, Velocity templates, and SparkPost with exponential backoff retry and SMTP fallback; published utility-export event streams via Kafka.
- * Configured Jenkins CI/CD pipelines for automated testing and deployment on AWS EC2; set up Prometheus/Grafana and CloudWatch alerting on delivery rates, failures, and latency.

Accenture

Software Engineer

Feb 2021 – May 2022

Pune, India

– Web Dashboard (*Java, Spring Boot*)

- * Developed layered Java (Spring Boot) services powering a white-labeled dashboard for 300K+ active utility customers.
- * Built time-series aggregation REST APIs for multi-fuel types (electric, gas, water) and solar net-metering with configurable granularity.
- * Secured dashboard APIs using OAuth 2.0 with client_credentials and refresh_token grant types, JDBC-backed token persistence, and role-based access control.
- * Optimized API response times for high-traffic endpoints using Redis master-replica caching, in-memory TTL-based caching, selective data loading, and GZIP compression.