

VISHNU J G

Bengaluru, Karnataka • +91 8277677655 • jgvishnu2001@gmail.com • linkedin.com/in/vishnu-j-g-bb5345192 • <https://jgvishnu.github.io>

Platform & GenAI Infrastructure Software Engineer with 3+ years of experience building high-throughput distributed systems, advanced e-commerce search infrastructure, and multi-agent AI platforms in production. Proven track record of architecting cloud-native platforms, accelerating deep learning inference, and implementing robust LLM governance, token tracking, and vector retrieval pipelines at scale. Expert in multi-region Kubernetes orchestration, Next.js BFF architectures, and low-latency data platforms.

WORK EXPERIENCE

Netcore Unbx

01/2023 - Present

Platform & GenAI Infrastructure Engineer

- Architected and deployed Zenith, an enterprise multi-agent AI analytics platform utilizing Google ADK and Gemini. Orchestrated dedicated sub-agents (SQL generation, text summarization, interactive chart generation) to convert natural language into production data artifacts backed by a full verifiable audit trail.
- Implemented a BigQuery template sub-agent loaded with 40+ parameterized analytics templates alongside safe loading layers and site-key resolution. Engineered a strict tier-based usage governance platform in FastAPI and PostgreSQL, executing token tracking, dry-run query cost estimations with automated cost ceilings, and context truncation to protect against LLM context overflows.
- Cut embedding latency for large multilingual models by building a PVC-backed ONNX cache init container and utilizing torch.compile alongside 4-bit GPU quantization via BitsAndBytes. Designed an end-to-end Kubeflow ML pipeline for LLM-powered cross-sell recommendations leveraging GPT-4o with automated async NATS alerts.
- Co-developed and scaled core platform services in Go and Python serving 1M+ daily active requests. Built NATS Bridge, a Go site-to-region locator microservice utilizing MongoDB, go-kit endpoints, and structured error mapping to safely handle cross-cloud event routing.
- Built production hybrid retrieval paths for vector databases, fusing dense vector searches with PostgreSQL full-text and Weaviate queries via Reciprocal Rank Fusion (RRF). Optimized offline personalization by rewriting row-by-row merchandising evaluations into high-performance vectorized pandas/numpy matrix computations.
- Built Bifrost, a gateway API and Streamlit admin dashboard allowing internal teams to securely manage site-scoped search configs across 5 production cloud regions via dual-container Kubernetes pods and Envoy routing.
- Integrated tier-aware catalog ingestion validation into distributed Argo Workflows pipelines using Java (Guice), enforcing per-customer capacity and rate limiting before streaming entries to Solr clusters.

EDUCATION

Bachelor of Technology (B.Tech) in Computer Science and Engineering

PES University • Bengaluru, Karnataka, India

05/2023

PROJECTS

Boson: Site-Specific Spell-Correction Microservice

- Built a FastAPI and SymSpell spell-correction system from scratch that trains custom dictionary models directly from client JSON product catalogs.
- Containerized a dual-process environment (Gunicorn API + Streamlit UI) with cython_hunspell built from source to dynamically auto-detect and correct queries across 40+ locales.

SKILLS

Languages & Frameworks

- Django
- FastAPI
- Go
- Java (17)
- JavaScript
- Next.js
- Python
- SQL
- TypeScript

MLOps & AI Infrastructure

- BigQuery Spend Optimization
- KServe
- Kubeflow Pipelines (KFP DSL)
- NumPy
- ONNX Runtime
- Pandas
- PySpark
- PyTorch (torch.compile, 4-bit GPU quantization)

GenAI & Agent Systems

- Context-Addressable Data Hashes
- Gemini
- Google ADK (Multi-agent orchestration)
- OpenAI SDK
- Prompt Engineering
- Token-Usage Governance
- Tool Registration

Distributed Systems & Data

- Event-Driven Architecture
- Kafka
- Microservices
- MinIO
- MongoDB
- NATS
- Next.js BFF Pattern
- PostgreSQL
- Redis
- REST/gRPC
- Server-Sent Events (SSE) Streaming

Search & Retrieval

- Apache Solr
- Hunspell
- Hybrid Retrieval
- pgvector
- Reciprocal Rank Fusion (RRF)
- SymSpell phrase correction
- Vector Search
- Weaviate (GraphQL)

Cloud & DevOps

- AWS
- Docker
- Envoy Proxy
- GCP
- GitHub Actions (OIDC role assumption)
- Helm (Advanced multi-region templating)
- Kubernetes (Dual-container pods, Init containers)
- Skaffold