Shubham Singh

☎ 812-778-4049 ✉ ksingh.shubh@gmail.com linkedin.com/in/shusingh shusingh.github.io Seattle, WA

Summary

Software Development Engineer at Amazon with 3+ years of experience building distributed systems, LLM-powered platforms, and large-scale compliance infrastructure on AWS. Proven track record of designing production-grade AI systems, high-reliability orchestration pipelines, and self-service tooling that reduced manual workflows by 80%. Open to SDE II and AI/ML platform roles.

Technical Skills

Languages: Java, Python, TypeScript, Go, SQL, C/C++

Cloud & Infrastructure: AWS (Lambda, Step Functions, DynamoDB, S3, API Gateway, CloudFront, WAF, Glue), Docker, Kubernetes, CI/CD, Infrastructure as Code

AI/LLM: AWS Bedrock, Strands Agents SDK, RAG, Vector Search, Tool-Calling, MCP, Prompt Engineering

Data & Backend: PySpark, ETL Pipelines, REST APIs, Microservices, Event-Driven Architecture, OpenSearch

Tools: Git, Bash, React, Redux, Agile/Scrum, Technical Documentation

Experience

Amazon **Aug 2022 – Present**

Software Development Engineer II | Seattle, WA *Oct 2025 – Present*

Software Development Engineer I | Seattle, WA *Aug 2022 – Sep 2025*

- Designed and built a **production-grade LLM system** using **AWS Bedrock and RAG** that automated ~**80% of manual compliance workflows**, enabling **natural-language querying** across regulatory datasets and replacing multi-day manual review cycles.
- Engineered a **hybrid search architecture** combining **OpenSearch (BM25)** for structured control data and **vector-based document retrieval (kNN)** for unstructured corpora, with secure tool-calling, audit logging, and sub-second AI response latency.
- Built distributed orchestration using **AWS Step Functions, Lambda, and DynamoDB** to replace brittle chained invocations, achieving **4× higher reliability** and fine-grained state visibility for regulatory post-processing pipelines.
- Architected **self-service workflow configuration platforms** enabling non-technical users to set up compliance and document-management workflows without engineering support, **cutting onboarding time from weeks to hours**.
- Led design of a PySpark-based **EU-DSA compliance pipeline** processing **30M+ monthly records** with schema-versioning, validation, and error isolation; reduced incident resolution time by **70%**.
- Owned periodic **on-call rotations** across 5 production services, triaging and resolving critical incidents. **Mentored a junior developer** on system design, code quality, and operational best practices.

Education

Indiana University Bloomington **Jan 2021 – Jul 2022**

Masters in Data Science *Bloomington, IN*

Projects

Strands Multi-Agent Orchestrator (Python) ***[GitHub]*** **2026**

- Built a **multi-agent orchestration system** using **Strands Agents SDK** with specialized agent routing (planner, research, data, synthesizer), **MCP tool integration**, **FastAPI streaming**, and **observability tracing** over heterogeneous data sources including SQL, web search, and document corpora.

Distributed Rate Limiter (Go + Redis) ***[GitHub]*** **2026**

- Built a **distributed rate limiting service** in **Go** using **Redis Sorted Sets + atomic Lua** sliding window enforcement, exposing an HTTP check API and middleware integration with fail-open behavior and bounded Redis latency.

RAG Platform (Go + OpenSearch) ***[GitHub]*** **2025**

- Built a **multi-tenant RAG platform** in **Go** using **OpenSearch (BM25 + kNN)** with async ingestion (extract, chunk, embed, index), Redis Streams workers, and citation-grounded answers over large document corpora.