Nischal Shrestha

Senior Software Engineer | Manassas, VA
er.nischalshrestha@gmail.com | (571) 596-2569 | github.com/nick-chal | linkedin.com/in/nischalshrestha

PROFESSIONAL SUMMARY

Senior Software Engineer with 7+ years building distributed, AWS-native systems at scale. Specialized in Python backend architecture, async event-driven pipelines, and multi-tenant SaaS platforms across telecom and enterprise domains. Deep expertise in system design, algorithmic optimization, and AI/semantic search integration. Proven track record delivering end-to-end ownership from infrastructure provisioning and API design to production observability and cross-functional technical leadership. Comfortable operating as a senior IC or leading small engineering teams; consistently focused on scalability, reliability, and measurable business outcomes.

TECHNICAL SKILLS

Languages: Python, Java, TypeScript, JavaScript, SQL

Backend Frameworks: FastAPI, Django, Node.js, Express, GraphQL (Graphene), Celery, SQLAlchemy, PySpark

Cloud & Infrastructure: AWS (Lambda, ECS Fargate, EC2, S3, RDS, Aurora, Redshift, Glue, SNS, CloudWatch, VPC), Terraform, Docker, Kubernetes, DigitalOcean, Cloudflare, Kafka

Databases: PostgreSQL, MySQL, Aurora, MongoDB, Redis, DynamoDB

Vector & Search: Milvus/Zilliz, Algolia, Pinecone, Qdrant, OpenAI Embeddings, Vector Pipelines

Frontend: React, Next.js (App Router), TypeScript, RTK Query, HTML/CSS

Observability: OpenTelemetry, Grafana, Sentry, Fluent Bit, Elasticsearch, CloudWatch

DevOps & Testing: GitHub Actions, GitLab CI, Pytest, Selenium, Playwright, automated rollback pipelines

PROFESSIONAL EXPERIENCE

Teltik Communications *East Brunswick, NJ* *May 2023 - Present*

Senior Software Engineer

- Architected a unified multi-carrier telecom backend using a Factory Pattern to abstract 5 MVNO carrier networks into a single service layer with concurrent polling, centralizing the full SIM lifecycle, eliminating manual carrier portal workflows, and saving operations teams 40+ hours per month.
- Built a SIM bulk provisioning platform from greenfield with a multi-queue Celery worker architecture, per-carrier routing, cascading retry with automated backoff, and a rate-limited concurrent batch processor, processing 10,000+ operations daily at 99.9%+ task completion with near-zero manual intervention on failures.
- Delivered a production multi-tenant IoT/SIM management portal with a 4-tier RBAC system, real-time visibility into 100,000+ active SIMs, and a pull-based telemetry sync pipeline for orphan detection and historical LTE signal archiving, migrating infrastructure to ECS Fargate for improved deployment reliability.

- Built a multi-tenant affiliate commission platform with OAuth2 and API key dual-auth and dual-DB tenant isolation. Re-engineered the auto-claim engine from O(n x m) to O(log k) using sorted-range binary search, eliminating daily manual reconciliation and achieving 100% referral revenue attribution accuracy.
- Re-engineered SMS delivery from a synchronous monolith into a fault-tolerant async pipeline with per-channel fan-out, WebSocket real-time broadcasts, and a Redis-backed cache layer. Migrated the underlying datastore from MySQL to AWS Aurora, reducing database read load by 60% and achieving zero message loss at scale.
- Established production observability standards using OpenTelemetry, Sentry, and Grafana Cloud. Built structured log pipelines with Fluent Bit and Elasticsearch, implemented cron health monitoring with dead man's switch alerting, and reduced mean time to detection (MTTD) for incidents by over 50%.

Gorin Systems *Jacksonville, FL* *Feb 2021 - Apr 2023*

Software Engineer

- Engineered a dual-engine semantic search system routing English queries through OpenAI embeddings and Milvus/Zilliz and Hebrew queries through Algolia, with automated language detection and a constant-time hydration layer eliminating N+1 queries. Built an accompanying bulk vectorization pipeline with Parquet-based backup and restore for a 1536-dimensional production collection, reducing search failure rates from ~8% to under 0.5%.
- Built a CLI ingestion engine for bilingual scholarly texts normalizing inconsistent proprietary XML schemas into a unified PostgreSQL schema with UPSERT idempotency and a regex-based footnote linkage engine achieving 100% automated linkage accuracy across six major corpora, replacing weeks-long manual editorial work per publication cycle.
- Delivered a production Django and GraphQL backend serving mobile and web clients with PostgreSQL, AWS S3, Stripe recurring billing, and an idempotent OAuth system (Apple Sign-In and Google) with JWKS token validation and nonce replay prevention. Retrofitted i18n across 10+ core models to unlock Spanish and Portuguese markets.
- Designed a zero-touch GitLab CI pipeline auto-provisioning a server, DNS record, and isolated CI/CD environment per client via Cloudflare and DigitalOcean APIs, compressing new client onboarding from days to under 30 minutes. Resolved 3 critical billing architecture flaws securing 100% invoice accuracy across all client tenants.
- Built large-scale data reconciliation tooling with PySpark to ingest, normalize, and synchronize 2M+ daily vehicle fleet records into AWS S3 and Redshift, enabling downstream analytics for business-critical reporting.

Leapfrog Technology *Seattle, Wa* *Jan 2019 - Feb 2021*

Software Engineer

- Led a team of 3 engineers owning sprint ceremonies, technical delivery coordination, and code quality standards. Engineered a multi-tiered credit risk scoring engine in Java evaluating applicants on credit score, income, and financial risk thresholds stored in PostgreSQL, reducing average manual underwriting turnaround by 80%.
- Reduced AWS cross-account infrastructure provisioning from multiple days to under 1 hour by migrating a Python/Spark ETL stack (Lambda, SNS, S3, Glue) to full Terraform IaC, enabling repeatable multi-environment deployments across 3 distinct environments with zero manual steps.

- Designed and implemented a parallel Selenium automation framework for a vehicle insurance QA suite covering multi-form and multi-edge-case workflows, cutting full regression cycle time from 6 hours to under 90 minutes.

EDUCATION

Bachelor of Engineering, Computer Engineering *2014 – 2018*

Kantipur Engineering College, Tribhuvan University, Nepal