# YASHJEET SINGH

**Software Engineer | GCP & Backend Systems | GenAI / LLM Integration**

+91-9729840783 | work.yashjeet@gmail.com | India · Open to Relocation / Remote | LinkedIn · GitHub

**1M+ events/day processed · 99.9% platform uptime · <500ms p99 API latency · ₹3 Cr+ revenue retained**

## PROFESSIONAL SUMMARY

Software engineer with 1+ year of experience designing and operating event-driven, cloud-native systems on Google Cloud Platform, including pipelines processing 1M+ events/day across multi-tenant environments. Builds production GenAI workflows using Claude (Opus/Sonnet) and OpenAI APIs for structured extraction, agent loops, and process automation. Comfortable owning a service end-to-end — design, deployment, monitoring, and incident response — in small, fast-moving teams with limited tooling and no existing playbook. Currently pursuing a B.Tech in Artificial Intelligence alongside full-time engineering work.

## TECHNICAL SKILLS

**Languages & Backend:** Python, FastAPI, REST APIs, async I/O, event-driven architecture, microservices
**Cloud & DevOps:** GCP (Cloud Run, BigQuery, Pub/Sub, IAM, GCS), Docker, CI/CD, Infrastructure-as-Code
**Data & Messaging:** Redis, Firestore, real-time ETL/ELT, idempotent consumers, stream processing, SQL, schema validation
**GenAI & LLMs:** Claude (Opus, Sonnet), OpenAI APIs, prompt engineering, agent loops, structured-output validation, OCR pipelines
**System Design:** Distributed systems, fault tolerance, dead-letter queues, RBAC, SLA design, observability, incident response

## EXPERIENCE

### Software Engineer — AVA International

*Noida, India (Onsite) | Dec 2025 – Present*

- Designed a multi-tenant data platform on GCP (Cloud Run + BigQuery) supporting 5 isolated tenants and ~200K records/day with zero-downtime deploys.
- Built an event-driven ETL pipeline handling 1M+ events/day, using idempotent consumers and automated retries to minimize duplicate processing.
- Developed production LLM pipelines on Claude Opus/Sonnet — structured prompt chains, agent loops, and output validation — deployed as live internal automation.
- Shipped REST APIs with <500ms p99 latency under concurrent load; added RBAC, rate limiting, and middleware to support a 99.9% data-integrity SLA.

### Software Engineer — Longway India

*Delhi, India (Onsite) | Mar 2025 – Nov 2025*

- Built OneScan, an event-driven capture system integrating 3 external platforms, replacing manual data entry and cutting ingestion lag from hours to under a second.
- Designed a Redis-primary / Firestore-fallback write path handling ~40 concurrent writes/sec with no data loss during partial outages.
- Built a 112-rule schema-validation engine for ingestion and contributed to an AI-assisted OCR pipeline that scaled throughput without added headcount.

## PROJECTS

### CommerceOS — Commerce Data Platform

— *GCP, BigQuery, Pub/Sub, FastAPI, Docker, LLM Integration* | avaipl-commerceos.vercel.app

- Designed a stateless, event-sourced microservices system processing 1M+ operations/day with independent deploy pipelines.
- Added a GenAI forecasting layer surfacing reorder signals 48–72 hours ahead of depletion, reducing stockout incidents by ~60%.
- Implemented circuit breakers, retry budgets, and dead-letter queues, cutting downtime ~40% and improving response time ~25%.

### AromaPureAir — IoT Fleet Backend & SaaS Automation

— *Python, IoT Telemetry, SaaS Backend, Event Processing* | aromapureair.vercel.app

- Built a backend managing 700+ IoT devices across 80+ enterprise clients — device state, telemetry ingestion, real-time alerting.
- Automation workflows lifted client retention from ~60% to ~98%, contributing to ₹3 Cr+ in verified revenue retention.
- Designed a multi-tenant SaaS architecture with isolated per-client data pipelines, enabling onboarding without engineering involvement.

## EDUCATION

**B.Tech, Artificial Intelligence — Gurugram University** *2024 – 2027 (Expected)*
**Associate Degree, Computer Science Engineering — HSBTE (CGPA: 7.2)** *2020 – 2023*