Wen-Kai "SansWord" Huang sansword@gmail.com | (669) 220-0281 | linkedin.com/in/sansword | San Jose, CA
Senior Software Engineer · Data Infrastructure · Distributed Systems · AI/ML Platform

"A data platform builder interested in AI systems and their mathematical foundations."

PROFESSIONAL SUMMARY

Senior Software Engineer with 9+ years of experience designing and operating large-scale distributed systems and global data infrastructure on **AWS** and **GCP**, serving ~900M monthly active users across US, APAC, and Europe. Specialized in high-throughput, fault-tolerant pipelines (>20K msgs/sec), scalable data platforms and cloud-native migrations. Proven technical leader and mentor recognized for cross-functional collaboration and driving large-scale impact across globally distributed teams. Currently exploring agentic development and harness engineering through hands-on projects and public writing — applying measurement discipline (cost, model comparison, workflow design) to AI-assisted software engineering.

PROFESSIONAL EXPERIENCE

Yahoo! · Sr. Software Dev Engineer · *Sunnyvale, CA · Feb 2019 – Jun 2025*

- Led end-to-end delivery of Yahoo's data infrastructure evolution to GCP over years, leading a dedicated 5+ engineer team to build a new GCP-native ingestion platform while designing a globally distributed multi-tenant pipeline using **Cloud Composer**, **Airflow**, and **Dataproc** — improving data freshness from 3+ hours to under 60 minutes across US, APAC, and Europe.
- Contributed to Yahoo's hybrid cloud data infrastructure, supporting a zero-downtime AWS migration with Kubernetes-based auto-scaling and maintaining the multi-region lambda architecture serving recommendation, ads, and multiple downstream systems across News, Sports, and Finance.
- Identified and resolved large-scale query inefficiencies, building aggregation and caching layers that reduced data scanned from 725 GB to 121 KB and cut query costs by 70% on Yahoo's **BigQuery** infrastructure.
- Drove an on-call reliability initiative over several months, resolving root causes and improving monitoring to reduce open incident tickets from 1,000+ to zero; established practices subsequently adopted by the broader on-call team.
- Led cross-functional collaboration between ML research and engineering teams to productionize ranking models, co-inventing a patented salient entity algorithm (US11803605B2), and integrating ranking models into production.

Yahoo! E-Commerce Platform · Sr. Engineer · *Taipei, Taiwan · Jun 2016 – Feb 2019*

- Drove early AWS adoption across Yahoo engineering, building blue/green deployment and auto-scaling patterns via ECS that were later adopted by global teams.
- Contributed to migrating 80 TB from house object storage to **S3** with zero downtime, implementing dual-write/dual-read patterns to ensure data consistency and safe cutover; completed migration to drive system EOL.
- Scaled **Pigeon**, a high-availability **Apache Pulsar**-based message bus sustaining 20K msgs/sec with data consistency guarantees, providing HTTP endpoints that abstracted client-side complexity across Yahoo's e-commerce teams; drove migration from **Apache ActiveMQ** and improved MTBF from 3.5 days to 90 days (~26×).
- Core developer of **Cupid**, a fault-tolerant multi-tenant discount service replacing a 10-year-old system; invented a flexible JSON-based rule engine and designed its rule semantics, empowering teams to independently configure promotions and marketing strategies — driving 10K/day coupon redemption growth and boosting annual sales events.

SIDE PROJECTS & PUBLIC WRITING

AI Coding Practice — applied exploration of agentic development · *2026 – Present*

- "My Tiny AI Bootcamp" — Public write-up of 3 weeks going from AI skeptic to shipping 3 projects with Claude; reflects on where AI fills knowledge gaps vs. where senior judgment still steers design decisions (e.g. introducing a CubeDriver interface to decouple BLE from pointer/touch input).
- "Leaderboard System Design — A Study Guide" — Self-authored study guide covering interview skills, a full worked leaderboard design (Redis sorted sets, fan-out, sharding, rank problem), and a companion note on using AI to prepare for system design interviews.
- sans_learning — Measured notes on Claude Code workflows: 7.8× cost delta between Sonnet 4.6 and Opus 4.7 on an identical feature run; per-story-point cost analysis; definition of harness engineering.
- sans_cube — Build a real-time smart-cube solve analyzer with BLE ingestion, live 3D rendering, phase detection (CFOP/Roux), and opt-in Firestore cloud sync. Deliberately picked an unfamiliar stack to stress-test AI-assisted workflows on greenfield code.
- **Other experiments** — Smaller explorations: cut_sh (ffmpeg scripting) · sans_yt_summary (Claude skill plugin with prompt-injection hardening).

PATENT

Determining Salient Entities and Generating Salient Entity Tags Based Upon Articles · US11803605B2
Co-invented methods to identify salient entities in articles at scale and apply them in production recommendation systems at Yahoo.

EDUCATION

M.S., Computer Science — National Chengchi University, Taipei, Taiwan
Major: Programming Languages and Software Methodology
B.S., Mathematics — National Chengchi University, Taipei, Taiwan

TECHNICAL SKILLS

Writing & Learning in Public: Technical blog posts, study guides
AI Coding Practice: Agentic Programming, Harness Engineering
Languages: Python, Java, SQL, Shell, Groovy
Cloud – GCP: Dataproc, Cloud Composer, Dataflow, Airflow, BigQuery
Cloud – AWS: S3, ECS, ELB, ElastiCache
Infrastructure: Terraform, Kubernetes, Docker, Redis
Stream Processing: Apache Storm, Apache Pulsar, Kafka