

Sumit Kumar

Noida, India | thedeceptio@gmail.com | +91 9650376719 | github.com/thedeceptio

SUMMARY

Senior Software Engineer with 11+ years of experience specialising in large-scale data infrastructure. Deep, hands-on expertise in ClickHouse (25 TB+ clusters, 20B+ row tables, 100M+ rows/day ingestion), Kafka event pipelines, MySQL at scale, and hybrid hot/cold storage (S3/ECS/local disk). Recently extended into applied AI — built a production text-to-SQL analytics platform (LiteLLM + Gemini) with a custom semantic layer over business-critical ClickHouse schemas. High-ownership IC with a track record of delivering measurable infra cost savings and performance gains end-to-end, without a large team.

EXPERIENCE

Tech Lead

Apr 2023 – Present

JobHai.com | A Naukri.com / InfoEdge Venture

- Architected and operated ClickHouse clusters handling 25 TB+ data, powering analytics, fraud detection, and ML pipelines — sole owner of schema design, index strategy, and cluster health.
- Built Kafka → ClickHouse streaming ingestion pipelines processing 100M+ rows/day (multi-TB/day), including CDC sync from MySQL production across 26 tables at ~1-minute latency.
- Executed a multi-datacenter ClickHouse migration (PI → NMIT primary) with hybrid hot/cold tiering on Dell EMC ECS/VIPT-backed S3; eliminated a 95% disk-capacity crisis and achieved ~2 Cr annual savings.
- Designed and shipped a text-to-SQL AI analytics platform: custom LiteLLM + Gemini 2.0 Flash backend, React chat widget embedded in a forked Apache Superset instance, and a rich semantic layer covering 37 charts and all core business metrics (JML, ARPT, credit buckets, Natural Online vs Direct Sales).
- Optimised MergeTree table design, compaction, TTL-based storage policies, and partition strategies for 20B+ row tables — reduced storage growth by 30 GB/day and improved p95 query latency.
- Scaled MySQL master-slave architecture to 60M API calls/day, cutting master QPS by 30% via read offloading and query-level optimisation.
- Reduced recommendation pipeline cron runtime from 12 h → 1 h, enabling 10M+ targeted push notifications daily.
- Resolved multiple critical ClickHouse production incidents: memory-tracker integer overflow (false 1 EiB OOM), S3-backed table blocking background threads causing client hangs, and XAmzContentSHA256Mismatch on Dell ECS uploads.

Lead Engineer

Sep 2019 – Mar 2023

JobHai.com | A Naukri.com / InfoEdge Venture

- Designed backend architecture for candidate/job recommendation engine processing ~5M daily notifications; boosted DAU by 30%.
- Built real-time job suggestion APIs (~1K RPS, sub-200 ms latency) using Redis caching and optimised ClickHouse queries.
- Built large-scale clickstream ingestion pipeline into ClickHouse for behavioural analytics — foundational system still powering ML feature generation.
- Owned end-to-end hiring and roadmap for a small data engineering sub-team; interviewed, onboarded, and mentored engineers on ClickHouse and streaming systems.

Backend Developer & Data Analyst

Mar 2014 – Aug 2019

TrulyMadly | Dating platform, 1M–5M users

- Built the core Recommendation Engine powering match suggestions for a 1M–5M user base; improved engagement by 60% and resolved cold-start starvation through a hybrid collaborative-filtering approach.

- Implemented monetised premium features that moved the product to break-even, directly contributing to business sustainability.
- Integrated and scaled AWS infrastructure (EC2, RDS, S3) to support rapid user growth.

SELECTED PROJECTS

- **Text-to-SQL Analytics Platform (2024–25):** Production-grade natural language query interface over ClickHouse — LiteLLM + Gemini 2.0 Flash, custom React widget in forked Apache Superset, semantic layer encoding 37 charts, business KPIs, and schema-aware SQL generation with result summarisation.
- **ClickHouse Multi-DC Migration & Hybrid Storage (2024):** Migrated a standalone 25 TB+ node at 95% capacity to a two-node NMIT-primary cluster with Dell EMC ECS S3 cold tier; designed storage policies, backup strategy (native BACKUP/RESTORE), and upgrade path to ClickHouse 26.x LTS.
- **MySQL → ClickHouse CDC Pipeline (2023):** Real-time change-data-capture sync across 26 production MySQL tables into ClickHouse at ~1-minute latency using Kafka; powers all analytics and ML feature stores.

SKILLS

Databases	ClickHouse (expert) · MySQL · Redis
Streaming & Pipelines	Kafka · Airflow · CDC pipelines
AI / ML	LiteLLM · Gemini API · text-to-SQL · recommendation systems
Languages	Python · Node.js · SQL
Infrastructure	AWS S3/ECS · Kubernetes (exposure) · Docker
BI / Analytics	Apache Superset (forked, customised) · Metabase
Familiar	Apache Iceberg · Trino/Presto · Spark · Hadoop · RabbitMQ

EDUCATION

B.Tech, Computer Science	2008 – 2012
Roorkee Engineering & Management Technology Institute	

HONORS & AWARDS

- Gold Award for Excellence, InfoEdge (2023) — for building the vernacular JobHai app.
- Employee of the Year, TrulyMadly (2018, 2019).