

AAKASH KASAMSETTY

AI and Data Engineer | LLM Agents and Data Platforms

Milwaukee, WI | +1 (414) 779-7252 | kasamsetty.aakash@gmail.com | [Linkedin](#) | [Portfolio](#)

PROFESSIONAL SUMMARY

AI and data engineer with 5+ years building data platforms and the LLM agents that run on top of them. Built USense, an MCP-based text-to-SQL agent that lets business users query BigQuery, PostgreSQL and Oracle in plain English, raising execution accuracy on complex multi-table schemas from 33% to 84%. Brings a data engineering foundation in BigQuery modeling, Airflow ETL and analytics, plus production LLM systems on Google Cloud with RBAC, cost guardrails and audit logging built in.

TECHNICAL SKILLS

- **AI Agents and LLMs:** AI agents, Model Context Protocol (MCP), text-to-SQL, RAG, chain-of-thought and dynamic few-shot prompting, LLM evaluation, LangChain, Mistral Large, Claude Haiku 5.5, Vertex AI, ChromaDB, Whisper
- **Data Engineering:** BigQuery (partitioning, clustering, cost controls), Apache Airflow, ETL pipelines, data modeling, schema introspection, data validation, audit logging
- **Databases and Warehouses:** BigQuery, PostgreSQL, Oracle, MySQL, MongoDB, Snowflake
- **Languages and Frameworks:** Python, SQL, JavaScript, TypeScript, FastAPI, SQLAlchemy, Pandas, scikit-learn
- **Cloud and Infrastructure:** Google Cloud (Cloud Run, Cloud Functions, Pub/Sub, Cloud SQL, Cloud Storage, Identity Platform), Terraform, Docker
- **Governance and BI:** RBAC, PII/PHI masking, HIPAA, Tableau, Power BI

PROFESSIONAL EXPERIENCE

UpX Labs, San Jose, CA

Full Stack AI Engineer | Jun 2024 – Present

USense: *Conversational BI and Natural-Language-to-SQL Platform*

- Productized an MCP-based text-to-SQL agent with LangChain and SQLAlchemy that lets non-technical stakeholders run complex analytical queries across BigQuery, PostgreSQL and Oracle without engineering help.
- Raised execution accuracy on complex multi-table analytical schemas from 33% to 84% using Mistral Large, chain-of-thought prompting, dynamic few-shot example retrieval and semantic grounding in BigQuery metadata.
- Built schema retrieval on automated introspection indexed in ChromaDB, so the agent pulls only the tables and columns relevant to each question.
- Secured production data access with RBAC and a SQL parser that blocks every non-read statement (INSERT, UPDATE, DELETE, DROP) before execution.
- Applied BigQuery partitioning and clustering guidelines, per-query cost guardrails and audit logging to keep analytical compute costs predictable.
- Returned each answer as a plain-English summary, the generated SQL, a result table and a chart, with exports to CSV, Excel, PDF, Google Docs, Tableau and Power BI.

UScribe: *Event-Driven Clinical AI Platform*

- Architected a GCP-native ambient clinical scribe from prototype in 1 month to MVP in 3: FastAPI on Cloud Run, Cloud SQL (PostgreSQL), Cloud Storage and Pub/Sub with dead-letter retries, provisioned with Terraform.
- Served Whisper large-v3-turbo on autoscaling Vertex AI GPU endpoints, transcribing 12 minutes of audio in ~30 seconds and generating 29-field structured SOAP notes with Claude Haiku 5.5 in under 60 seconds.
- Built Cloud Functions and Pub/Sub pipelines for clinical entity extraction, validation and routing to eClinicalWorks over HL7/FHIR, with automated PHI de-identification and clinician edit-distance tracking that cut edits per note by 14%.
- Scaled to a 40-physician practice pilot generating 3,000+ notes weekly.

Project Vox: *Real-Time AI Voice Agent*

- Architected a low-latency conversational voice agent for natural-language appointment scheduling and multi-turn caller qualification, prototyped in one week in Python and FastAPI and tuned to 2.8s response latency.
- Streamed full-duplex audio over WebRTC with noise-gate pre-processing, keeping conversations stable when callers interrupt.
- Wired real-time tool calls to the Google Calendar API and webhooks to check availability, lock appointments and sync bookings to the CRM.
- Built a simulation test harness covering accent variation and network jitter; sentiment models trigger warm handoffs to staff, reaching a 93% AI resolution rate in beta testing.

University of Wisconsin–Milwaukee, Milwaukee, WI

Data Analyst | Feb 2023 – May 2024

- Queried and analyzed 5 years of enrollment and waitlist data with SQL and Python (Pandas) to find recurring registration bottlenecks across departments and terms.
- Built SQL data models joining enrollment, waitlist, section capacity and classroom data into one consistent reporting dataset for the Registrar's Office.
- Designed public-facing Tableau dashboards on student demand, waitlist trends and classroom utilization by course, term and building.
- Turned findings into section-count, capacity and time-slot recommendations that helped cut unmet course demand by 18%.

Capgemini

Big Data Engineer | Oct 2020 – Dec 2022

- Built BigQuery data models from 12M+ health insurance claims records using Python and SQL, creating a single source of truth for retention analytics.
- Automated scheduled ETL workflows with Apache Airflow and Python, normalizing claims, member and plan data for downstream reporting.
- Added automated data validation checks, access controls and field-level masking of sensitive member data.
- Analyzed attrition drivers and built Tableau dashboards by region, plan and segment that helped raise quarterly retention by 2.07%.
- Modeled high-volume player and match data for the Indian Super League (ISL) into executive Tableau reporting.

EDUCATION

Master of Science in Information Technology Management (Concentration: Data Analytics and Artificial Intelligence) | University of Wisconsin–Milwaukee