SARVESH GANESAN

India | +91 8668063705 | sarveshganesan2002@gmail.com | linkedin.com/in/sarvesh-ganesan09 | github.com/Sarvesh-GanesanW

PROFESSIONAL SUMMARY

Lead AI Architect with 2.5+ years of experience building and shipping a cloud-native enterprise data platform from the ground up. Major architect of 12+ production microservices spanning agentic AI, distributed ETL, data lakehousing, and MLOps - all running on AWS EKS. Proven track record of turning ambiguous product requirements into scalable, cost-optimized systems that serve real enterprise workloads.

TECHNICAL SKILLS

- **Languages & Frameworks:** Python, FastAPI, Flask, REST APIs, gRPC, SSE streaming
- **AI / LLM:** AWS Bedrock, Strands Agents, LangChain
- **Data Engineering:** Apache Spark, Apache Iceberg, DuckDB, Pandas, PyArrow
- **Databases:** PostgreSQL, pgvector, Redis, Pinecone, MongoDB, Elasticsearch, DynamoDB
- **Cloud & Infra:** AWS (EKS, Bedrock, S3, Lambda, ECR, CloudWatch), Docker, Kubernetes, Karpenter
- **MLOps:** MLflow, Jupyter, PyTorch, TensorFlow, scikit-learn, Docker-based GPU runtimes

EXPERIENCE

Lead AI Architect Sep 2023 – Mar 2026
Groundzero Software Private Limited Chennai, India

- **Owned the entire backend architecture** of the platform end-to-end — designed, built, and shipped 12+ production microservices (agentic chat, data provider, ETL orchestrator, Iceberg lakehouse, MLOps APIs, Spark/DuckDB runtimes) as the sole backend engineer.
- **Built a multi-agent AI system** that lets enterprise users query databases, build dashboards, run ETL jobs, and manage notebooks through natural language — reduced time-to-insight from hours of manual work to seconds of conversation.
- **Scaled data connectivity to 69+ sources** (Snowflake, BigQuery, Redshift, MongoDB, Kafka, S3, Salesforce, and more) through a dynamically-loaded connector architecture, eliminating weeks of per-integration development effort.
- **Reduced RAG query latency by 90%** (10s $\rightarrow$ <1s) by engineering a hybrid retrieval pipeline with vector search, full-text search, and neural reranking — enabling real-time document Q&A at production scale.
- **Delivered sub-10s query performance on 100M+ row datasets** by architecting a TB-scale Iceberg lakehouse with Spark and DuckDB compute engines, time-travel queries, and column-level masking for compliance.
- **Cut LLM inference costs** by implementing three-tier model routing (Opus/Sonnet/Haiku) with prompt caching, and built a token tracking system giving full cost visibility per request and per agent.
- **Designed the ETL orchestration layer** handling job submission, multi-job chaining, and real-time log streaming across three pluggable compute engines (Spark, DuckDB, Pandas) on dynamically provisioned EKS pods.
- **Stood up the MLOps infrastructure** — MLflow with custom JWT auth, Jupyter notebook lifecycle management on EKS, model versioning, and GPU-accelerated training containers for PyTorch, TensorFlow, and scikit-learn.
- **Engineered high-availability infrastructure** with Karpenter-based just-in-time node provisioning on EKS, achieving 99.9% uptime for distributed ETL and Spark workloads across multi-tenant deployments.

Software Trainee Jun 2023 – Aug 2023
SCI-BI Software Solutions Private Limited Chennai, India

- Built client-facing dashboards and reports in Power BI and Tableau; led requirements gathering in client meetings.

PROJECTS

AgentSynapse – Enterprise AI agent platform enabling natural-language data analytics through specialized agents (SQL, BI, ETL, ML) with hybrid RAG retrieval and 60+ integrated tools.

CloudeasyML – Open-source tool for deploying ML predictions and fine-tuning LLMs on AWS infrastructure. *(In Progress)*

EDUCATION

Bachelor of Technology: Computer Science and Engineering May 2023
Specialization in Artificial Intelligence and Machine Learning
SRM Institute of Science and Technology, Chennai GPA: 9.08

LANGUAGES

English (Fluent) | Tamil (Native) | Telugu (Fluent) | Hindi (Intermediate)