SAI TEJA SRIVILLIBHUTTURU

saiteja.srivillibhutturu@gmail.com | github.com/saitejasrivilli | linkedin.com/in/saitejasrivillibhutturu | Arlington, TX | +1 (682) 234-3567

Software Engineer - Backend

SKILLS

Languages: Java, Python, Go, SQL, C++
Backend: REST APIs, FastAPI, Spring Boot, Microservices, Async Streaming (SSE)
Data & Storage: PostgreSQL (Query Planning, Indexing, MVCC), Redis (LRU, Lua), Vector DBs (Qdrant, Pinecone), Data Governance
Cloud: AWS (EC2, S3, Lambda, SageMaker, CloudWatch), Docker, CI/CD (GitHub Actions)
Systems: Distributed Systems, Database Internals, Caching, Rate Limiting, WAL, Quorum Replication

EXPERIENCE

University of Texas at Arlington, Graduate Teaching & Research Assistant, Arlington, TX **Aug 2024 – Present**

- Engineered a **RAG**-based question answering backend by **fine-tuning LLMs** and chunking **1,000+ research papers** into a **Pinecone vector store**, an **agentic pipeline** that autonomously routes queries across 2 research domains for **semantic retrieval**.
- Architected end-to-end **document ingestion pipelines** covering **chunking**, **embedding generation**, and **vector indexing** in collaboration with 3 researchers, supporting retrieval workflows across **3,000+ document chunks**.
- Delivered **FastAPI async streaming APIs (SSE)** with **caching** and **request batching** strategies, reducing query response latency by **40%** for a research demo environment used by **10+ lab members**.

DentalScan, Machine Learning Intern, Remote, US **Feb 2026 – May 2026**

- Trained **supervised computer vision** models using **PyTorch** on a **50K+ image dataset**, improving **multi-class intra-oral diagnosis** accuracy by **12%** across 4 clinical categories.
- Streamlined **data ingestion** and **preprocessing pipelines** with cleaning, augmentation, and **class balancing** alongside the clinical data team, cutting manual data preparation time by **3 hours per training run**.
- Deployed **automated retraining pipelines** on **AWS (S3, EC2, SageMaker)**, enabling model updates from dentist-labeled **feedback loops** on a bi-weekly cadence.
- Refactored **asynchronous background job workflows** to decouple heavy preprocessing from the **inference path**, reducing average API response time by **20%**.

Tata Consultancy Services, Software Engineer, Chennai, India **Jun 2019 – May 2023**

- Architected **Spring Boot REST APIs** processing **50K+ daily financial transactions** across global enterprise clients, migrating a **monolithic system** to **microservices** and supporting **multi-region data governance** requirements, achieving **99.5% uptime** post-migration.
- Optimized **PostgreSQL query plans**, **indexing strategies**, and **MVCC behavior** on tables exceeding **50K+ rows**, improving average query execution time by **35%** and reducing lock contention across high-concurrency workloads.
- Introduced **Redis LRU caching** on high-frequency read endpoints across a **6-engineer team**, reducing **database load by 40%** and cutting API response latency from $\sim$**150ms to** $\sim$**90ms**.
- Authored **Java** and **Python data ingestion pipelines** processing enterprise financial transaction data across **3 upstream systems**, enforcing **data governance contracts** and improving **data consistency by 30%**.
- Integrated **structured logging**, **monitoring**, and **alerting** across **6 production services**, reducing **MTTR by 25%** from $\sim$60 minutes to $\sim$45 minutes.
- Owned **production service reliability** through **on-call rotations** and live incident debugging, partnering with cross-functional teams to deploy hotfixes and maintain **fewer than 3 P1 incidents per year**.

PROJECTS

Distributed KV Store – Python, WAL, Quorum Replication **github.com/saitejasrivilli/DistributedKVStore**

- Engineered a **fault-tolerant key-value store** implementing **write-ahead logging (WAL)** and **quorum-based replication (W=2/3)** across **3 nodes**, exploring **database internals** including log-structured storage and durability guarantees under single-node failure.
- Designed **log replay** and **node recovery** mechanisms allowing rejoining nodes to resync automatically within **10 seconds**, preserving **data integrity** without manual intervention.

Distributed Rate Limiter – Python, FastAPI, Redis, Docker **github.com/saitejasrivilli/distributed-rate-limiter**

- Constructed a **distributed rate limiter** using **sliding window** and **token bucket algorithms** with **Redis Lua scripts** for **atomic execution**, enforcing consistent rate limits across **5 concurrent instances** with **sub-5ms** overhead per request.

Glean-Lite: Enterprise Search Backend – Go, Qdrant, SSE, CI/CD **github.com/saitejasrivilli/glean-lite**

- Developed a **semantic search backend** using **vector embeddings** with **Qdrant**, achieving **sub-500ms** retrieval over **1,000+ indexed documents** via **SSE-based streaming APIs** with **access-controlled indexing** for data governance.
- Automated **document indexing pipelines** via **CI/CD workflows** (GitHub Actions), reducing manual reindex effort to **zero**.

Real-Time Bidding Engine – Python, FastAPI, Redis **github.com/saitejasrivilli/rtb-bidding-system**

- Built a **high-throughput auction backend** implementing **second-price bidding** with **budget pacing** strategies, sustaining **2,000+ QPS** in local load tests.
- Achieved **sub-100ms** end-to-end latency (**p95** $<$ **80ms**) under concurrent load using **caching**, **request batching**, and optimized **FastAPI** request handling.

EDUCATION

Master of Science in Computer Science | University of Texas at Arlington | Aug 2023 – May 2025
Bachelor of Technology in Computer Science | Andhra University | Jun 2015 – Apr 2019