SIDDHANT BHUJADE

San Jose, CA | (945)-241-6853 |bhujadesiddhant0307@gmail.com | www.linkedin.com/in/sid0307 | GitHub

WORK EXPERIENCE

Software Engineer | Callahan and Associates | United States Jun 2024 – Present

AI-powered Insights (Peer Suite – Financial Analytics SaaS)

- Developed AI-powered peer benchmarking dashboards in Angular and ASP.NET Core, surfacing LLM-generated insight summaries and recommendations across 700+ financial institutions.
- Designed and productionized RAG pipeline (LangChain, pgvector, OpenAI Embeddings) with hybrid retrieval over financial publications and SQL-based NCUA data, reducing hallucinations by 85% via fact-grounded retrieval.
- Decoupled LLM APIs and embedding workloads into FastAPI async microservices with Redis result caching, cutting P95 API latency from 22s to 5s across 500K+ monthly queries.
- Built a multi-tenant LLM gateway in FastAPI, routing requests across multiple LLM models based on subscription tier and prompt complexity, cutting inference costs by 35**%** and API failures by 95**%** via priority queuing and backoff retries.

Data Analytics Web Platform

- Built and shipped a full-stack data processing web application (React, FastAPI) enabling partner teams to configure file ingestion workflows, monitor async pipeline jobs in real time, and trigger automated report generation, eliminating engineering dependency for routine data delivery tasks.
- Architected fault-tolerant data processing workflows in Temporal, orchestrating loan ingestion, processing, and human-in-the-loop review across 10M+ records, reducing failure recovery from 3 hours to 15s.
- Developed an event-driven background processing infrastructure using Python, Celery, and RabbitMQ to offload analytical computations, achieving 3x throughput and decoupling async workloads from sub-100ms user-facing API SLAs.

Software Engineer, Intern | Callahan and Associates | United States Jun 2023 – May 2024

- Developed reusable Angular UI components integrated with analytics APIs, reducing dashboard load time from 10s to 3s via lazy loading and optimized change detection for 10K+ daily users.
- Optimized SQL Server query performance by implementing indexing optimization, query refactoring, and connection pooling, curtailing P95 query latency from 5s to 2s for high-traffic workflows.
- Implemented Redis cache-aside layer across 15+ API endpoints, achieving 85% hit rate and reducing database load by 70%.

Software Engineer | Persistent Systems | India Nov 2020 – Jul 2022

- Revamped healthcare analytics product by building a distributed data platform to process petabytes of clinical claims data using Spark on Kubernetes and Airflow, ensuring 99%+ data integrity for ML pipelines and BI reporting.
- Developed event-driven microservices in Java/Spring Boot handling 5M+ daily events via Kafka and GraphQL, reducing inter-service dependencies and improving system uptime to 99.9%.
- Implemented observability with Prometheus, Grafana, and CloudWatch, creating automated alerts and service health dashboards that reduced incident detection time by 70%.

TECHNICAL SKILLS

Languages: Python, C#, Typescript/JavaScript, Java, SQL, GraphQL
Frameworks: Fast API, .NET Core/ASP.NET, Spring Boot, Angular 17, React
Data & Messaging: Apache Kafka, Spark, Airflow, RabbitMQ
Database: SQL Server, PostgreSQL, Redis, Elasticsearch, AWS DynamoDB, AWS Redshift
Cloud & DevOps: AWS(S3, EC2, Lambda), Azure, Docker, Kubernetes, Git/Bitbucket
AI & Observability: OpenAI API, Claude API, LangChain, LangGraph, Prometheus, Grafana, AWS CloudWatch

EDUCATION

The University of Texas at Dallas Aug 2022 – May 2024
Master of Science, Information Science GPA: 3.97/4
Jabalpur Engineering College, India Aug 2016 – Aug 2020
Bachelor of Engineering, Computer Science and Engineering GPA: 3.7/4