Sai Dhiren Musaloji

musalojidhiren@gmail.com | +1 862-423-8830 | LinkedIn | GitHub | Portfolio

PROFESSIONAL SUMMARY

Software Engineer with 2+ years of experience designing and implementing highly available, scalable, and low-latency distributed systems across cloud environments (AWS, Azure). Expertise in full-stack data streaming, queue-based architecture, geospatial indexing (Redis/PostGIS), and containerization (Docker, Kubernetes). Proven ability to build robust, performance-optimized backend services using Python, Node.js, and modern CI/CD practices.

TECHNICAL SKILLS

Software Engineering Distributed Systems, Microservices, REST/gRPC, WebSockets, Full-Stack Development (Node.js, Next.js, Flask), CI/CD, OOP, System Design

Cloud & DevOps AWS (EC2, S3, SQS, RDS, EMR, Glue), Azure (Data Factory, Synapse), Terraform (IaC), Docker, Kubernetes, Azure DevOps, Git

Databases & Caching PostgreSQL (PostGIS), MySQL, DynamoDB, MongoDB, Snowflake (DWH), Redis (GeoSpatial, Caching), T-SQL, SQL

Programming & Data Python (PySpark, Pandas), JavaScript/TypeScript (Node.js, React), SQL, Java, Scala, Apache Spark, Apache Airflow

Machine Learning Systems MLOps (MLflow, Azure ML), RAG Architecture, Transformer Models, Deep Learning Architecture, Performance Optimization

PROFESSIONAL EXPERIENCE

Data Engineer | TAWIN Solutions LLC - Dallas, TX May 2025 – Present

Data consulting firm serving enterprise clients; developing AI-powered products

- **Architected** a food recommendation Proof-of-Concept (PoC) using Azure AI Foundry, integrating and normalizing data sources (mood, dietary preferences, geolocation) to power the core recommendation algorithm.
- **Built an ML data pipeline** integrating real-time restaurant data APIs with sentiment analysis and collaborative filtering models, creating a system for context-aware recommendations.
- Designed and implemented the **end-to-end system architecture** connecting Azure AI services, external restaurant APIs, and a user preference database to ensure a scalable PoC infrastructure.

Environment: Azure Data Factory, Azure Synapse, Azure AI Foundry, Python, SQL, Azure ML, REST APIs

AI Engineer Intern | Tech Mahindra, Makers Lab R&D - Pune, India Oct 2023 – Jan 2024

Fortune 500 IT services, $6B revenue - Advanced AI research division

- Managed the data pipeline for a major multilingual LLM research project, focusing on text collection, cleaning, and annotation for **15+ diverse languages**.
- **Engineered a real-time system** for German-to-English transcription, integrating commercial transcription APIs with a custom processing layer to achieve near real-time translation for internal research demos.
- Prototyped a semantic search **architecture** for LLM knowledge retrieval, implementing vector-based search mechanisms to improve contextual query understanding.

Environment: Python, Transformers, NLP, REST APIs, Data Preprocessing, Annotation Tools, API Integration

Cloud Computing Intern | LTI Mindtree - Hyderabad, India Feb 2023 – May 2023

Global IT services, $4B revenue - Cloud Engineering training program

- Completed intensive AWS cloud architecture training program, building proof-of-concept data processing workflows demonstrating **EC2, S3, Lambda, RDS, and VPC** integration for scalable batch processing
- Implemented **Infrastructure-as-Code** exercises using **Terraform** to automate resource provisioning, learning enterprise deployment patterns and environment management best practices
- Containerized sample microservices with **Docker** and orchestrated clusters via **Kubernetes** as part of **cloud-native architecture** training curriculum
- Shadowed senior cloud architects on client migration projects, learning real-world patterns for resource optimization and cost efficiency in production environments

Environment: Python, AWS (EC2, S3, Lambda, RDS, VPC), Terraform, Docker, Kubernetes

PLATFORM & DISTRIBUTED SYSTEMS PROJECTS

Real-Time Ride-Matching Platform (Uber-Like System)

- **Architected a dual-index geospatial platform** processing **50K+ GPS updates/min** using **Redis Geo** for low-latency lookups and PostGIS for durable analytics.
- Achieved **68% lower match latency (80 ms)** by optimizing algorithms and distributed locking mechanisms.

- Engineered **WebSocket fanout** delivering ETAs and booking state to **5K+ concurrent clients** with **sub-200ms**median propagation.
- Designed an event-driven, queue-based architecture leveraging distributed locking and idempotent booking flows to guarantee exactly-one driver assignment.

Technologies: Redis Geo, PostgreSQL (PostGIS), WebSockets, Node.js, Python, Docker, Distributed Locking

Real-Time Cryptocurrency Price Tracker - Full-Stack Application

- **Architected scalable full-stack web application** (Next.js/Node.js/TypeScript) implementing **ConnectRPC and gRPC server-streaming** for low-latency push-based price delivery.
- Engineered automated web scraping pipeline using **Playwright** browser automation framework, supporting concurrent tracking of 25+ cryptocurrency tickers.
- Delivered real-time data streaming for 15+ digital assets with sub-second update frequencies and **99.2% uptime**.
- Implemented comprehensive error handling and retry mechanisms for web automation workflows, achieving a **95% success rate** in price data extraction.

Technologies: Next.js, Node.js, TypeScript, gRPC/ConnectRPC, Playwright, Server-Streaming

Distributed Recognition Engine Using Cloud-Native Paradigms

- Built an **asynchronous image and text recognition pipeline** using **AWS (EC2, SQS, S3, Rekognition)** for scalable processing.
- Implemented a parallelized architecture with auto-scaling capabilities for high-throughput processing.
- Ensured **99.9% uptime** and fault tolerance using stateless compute strategies, SQS visibility timeout calibration, and retry mechanisms for robust message handling.

Technologies: AWS (EC2, SQS, S3, Rekognition), Python, Cloud-Native Architecture

High-Availability Financial Analytics System

- Engineered a transactional database integrating MySQL, Snowflake, and DynamoDB (15+ tables) handling **10,000+ daily transactions** with **99.9% data integrity**.
- **Containerized** the system using **Docker** and implemented robust **CI/CD pipelines**, reducing deployment time by **50%** and ensuring consistent environments.
- Designed dimensional data models for analytics workloads and built a real-time fraud detection system with advanced anomaly detection algorithms.
- **Impact:** Achieved **99.9% system uptime** and **sub-200ms query response times**.

Technologies: MySQL, Snowflake, DynamoDB, Python, Flask, Docker, Kubernetes, CI/CD

AI Research Assistant with Retrieval-Augmented Generation (RAG)

- Architected a **Retrieval-Augmented Generation (RAG) system** integrating the **Gemini 1.5 API**, a vector database (scikit-learn Nearest Neighbors), and transformer embeddings.
- Designed the end-to-end AI pipeline for document extraction, semantic retrieval, and LLM generation, incorporating **error handling and rate limiting**.

Technologies: LLM APIs, RAG, Vector Databases, Transformers, Python

Parallelized ML Pipeline for Oenological Forecasting

- Created a **Spark-based distributed system** on **AWS EMR** for wine quality prediction, improving training speed by **60%** through partitioned data and in-memory caching.
- Implemented autoscaling strategies reducing infrastructure costs by **35%** while maintaining SLA compliance.

Technologies: AWS EMR, Apache Spark, Python, MLlib, S3, Autoscaling

EDUCATION

Master of Science in Data Science - New Jersey Institute of Technology
Bachelors of Technology in Electronics & Communication Engineering - Mahatma Gandhi Institute of Technology

CERTIFICATIONS

Azure AI Engineer Associate - Microsoft AI-102
Azure Data Scientist Associate - Microsoft DP-100
Databricks Generative AI Fundamentals
Databricks Certified Associate