Sri Charan Byreddy

📍 Mountain View, CA | 🌐 Open to Relocate | ☎ 7167091593 | ✉ byreddysricharan1@gmail.com | in LinkedIn | GitHub

Summary

Software Engineer with 3+ years of experience in building LLM APIs, ETL pipelines, and backend systems using Python, Flask, AWS, and Snowflake. Skilled in cloud-native services, CI/CD workflows, and scalable REST API development.

Experience

AI Backend Developer | Saayam For All | Remote **Mar 2025 – Present**

- Migrated Groq API-based inference to a self-hosted LLM stack on AWS (LLAMA, DeepSeek), deploying quantized models on GPU-enabled EC2 instances with latency and cost benchmarks for scalable multilingual NLP support.
- Integrated `coverage.py` with the backend pipeline to achieve 80%+ test coverage across Flask routes (e.g., `/predict_categories`), utility modules, and model logic, supporting regression-safe deployments in Gen AI services.

Software Engineer | Third Estate Analytics | Buffalo, NY **Aug 2024 – Jan 2025**

- Engineered scalable data flows to consolidate multi-source parcel records via spatial joins, temporal aggregation, and metadata resolution, enabling geo-targeted queries and historical trend analysis on 1.2M+ properties.
- Delivered production-grade REST APIs layered with filters and region-scoped access controls; automated ingestion with AWS Lambda and CloudWatch to maintain 80%+ clean throughput and cut query latency by 35%.
- Leveraged PostGIS and QGIS for spatial overlay analysis and zoning compliance checks, enabling parcel reclassification workflows and reducing manual inspection effort by 40% across 50K+ mapped regions.

Software Engineer – Data Systems | Rubixe Disruptive Technologies | Bangalore, India **Sep 2022 – Jul 2023**

- Built multi-tenant Node.js services with JWT authentication and granular MongoDB access control, handling 10K+ monthly API calls under 150ms latency with zero cross-client data leaks across environments.
- Configured CI/CD automation in GitHub Actions with changelog verifiers, semantic version gates, and branch-based promotion rules, reducing faulty deploys by 60% and enabling faster QA-approved releases.
- Created KPI dashboards using React and Redux with real-time event loaders, intelligent refresh scheduling, and interactive graph views, improving user experience and reducing frontend rendering cost by 40%.

Programmer Analyst Trainee | Cognizant Technology Solutions | Chennai, India **Jan 2022 – Aug 2022**

- Developed resilient ETL pipelines using Informatica to ingest and standardize 20TB+ of compliance data into AWS RDS, accelerating auditing workflows and reducing reconciliation timelines by 35% across monthly cycles.
- Established a resilient Snowflake-to-S3 data sync mechanism with retry coordination and fallback logic to preserve ingestion continuity, minimizing partial loads and enhancing system stability during burst processing windows.
- Automated anomaly surfacing through a log parsing module tied to Grafana dashboards and JIRA API triggers, enabling QA triage and lowering untracked ingestion errors by 40% before reaching staging environments.

Full Stack Developer | Verzeo | Hyderabad, India **Jun 2021 – Dec 2021**

- Composed responsive React.js modules using Redux for state flow, Formik for schema-based validation, and Axios with JWT-based routing; enhanced client interaction speed and reduced form latency by 25%.
- Assembled containerized REST API services via Express.js and PostgreSQL, embedding scikit-learn pipelines for real-time ML predictions; CI/CD setup with GitHub Actions reduced deployment overhead by 30%.

Projects

AthleteSphere – Olympic Data Intelligence Platform | Flask, PostgreSQL, Power BI, REST APIs | Link

- Structured BCNF-compliant schema to organize athlete-event-medal records with referential integrity, enabling longitudinal trend analysis, demographic slicing, and seasonal medal tracking with SQL-backed consistency.
- Integrated indexed aggregates and query federation logic to serve real-time filters and drilldowns in Power BI dashboards, sustaining sub-120ms latency for 1K+ athlete records under multidimensional filters.

Advancing Fine-Grained Recognition Through Weakly Supervised Learning | Python, CNN, HOG, MLP | Link

- Implemented a weakly supervised recognition pipeline integrating CNN-based saliency maps with HOG feature extraction and MLP classifiers, boosting badge identification accuracy to 92.6% with an 18% gain over baseline.
- Modularized part-localization and feature fusion logic into independent units, reducing false badge matches by 22% and enabling scalable extension to five additional vehicle categories with minimal retraining.

Education

University at Buffalo, State University of New York | Master of Science in Data Science **Aug 2023 – Jan 2025**

Coursework: Machine Learning | Deep Learning | Data Mining | Software Systems | Cloud Data Warehousing | Big Data Systems | Database Systems | Large Language Models (LLMs) | Distributed Systems | System Design

Skills

Python | Java | C++ | SQL | JavaScript | Matlab | Flask | FastAPI | Node.js | Express.js | React.js | Redux | Spring Boot | PostgreSQL | MongoDB | Snowflake | Pandas | Power BI | AWS | EC2 | AWS Lambda | CloudWatch | Docker | GitHub Actions | Jenkins | Git | REST APIs | Postman | GraphQL | CI/CD | Agile | PyTest | coverage.py | Informatica | OpenCV | NumPy | CNN | MLP | HOG | Saliency Maps | JIRA API | LLAMA | DeepSeek | HTML | CSS