Harish Vadaparty

Portfolio | harishvadaparty@gmail.com | linkedin.com/in/harishvadaparty | github.com/Harryalways317

Work Experience

OnFinance.ai — Bangalore, IN
Backend and AI Engineer — Apr. 2024 – Present

- Led the backend team, working with a cross-tech stack of GCP, Azure, Golang, and Python, powering Gen AI applications and advanced Copilots.
- Built and optimized data processing pipelines with queue-based systems, achieving a 70% increase in document processing speed, supporting text, audio, and video data.
- Developed a bulk scraping pipeline for company data collection, now integral to our company data analytics operations.
- Developed and maintained Copilot backends for production clients like Oister and Avendus, ensuring seamless performance and scalability.
- Implemented RAG workflows and advanced retrieval mechanisms, enabling agentic search and structured data extraction from financial documents.
- Designed a template-based backend system, reducing client onboarding time by 50%.
- Created centralized services for LLM inferencing, log management, and error monitoring to improve system reliability.
- Led DevOps efforts, including resource management, CI/CD pipelines, and infrastructure setup, with centralized logging and usage analytics integrations.
- **Tech Stack:** Python, Golang, Java, MongoDB, PostgreSQL, Kafka, Docker, Azure, GCP.

Hexon Labs — Bangalore, IN
Software Engineer, Backend and Data — Jul. 2022 – Apr. 2024

- Led fine-tuning of large language models (LLMs) for Alignment with DPO and implemented AI-driven solutions on customer usecases.
- Developed and deployed Formless AI, a platform for AI-prompt-based form creation and interactive data collection, streamlining surveys and form data collection into a interesting way.
- Engineered AI agents for competition analysis, metrics collection and automated report generation, providing actionable insights and improving decision-making.
- Led backend development for a grocery application using AWS, Node.js, Python, and TypeScript, integrating ElasticSearch to improve search relevancy by 60% and building a recommendation engine that increased relevant cart items by 40%.
- Developed service applications for notifications, delivery management, and reporting, and ensured real-time synchronization across services data with CDC for inventory, pricing, and store POS systems.
- Implemented RAG-LLM workflows in-app and worked with Tortoise TTS for voice cloning and audio-based lip sync, enhancing the overall user experience with AI.

Board Infinity — Mumbai, IN
Software Engineer Intern — March 2022 – June. 2022

- Contributed to an established Ed Tech application in rewrite from React Native to Flutter with MVVC
- Developed multiple features on LMS, Blogs, Event and Course Management
- Implemented UXCam for analytics and notifications pipeline for FCM and Imrpoved animations
- Rewrote Modules of the app by caching and data reuse which improved performance by 40%

Education

Lendi Institute of Engineering and Techonology — Vizianagaram, IN
Bachelor of Techonology in Computer Science — June 2022

Projects

Any2Any Convertor — Python, Flask, RabbitMQ, MySQL,MongoDB

- Developed an OCR conversion, MP3 extraction and File conversion service with support for pdf to docx/docx to pdf.

- Implemented a Gateway for handling authentication, file storage, and processing requests in a queue with parallel workers.
- Established an email notification system to inform users upon completion of file processing.
- Encapsulated the service in a Kubernetes (K8s) environment for scalability, supporting various conversions and Volume Persistance on failures.

Job Leads Generator and Staffing Assistant Python,Open AI API,Langchain,Postgres

- Created an end-to-end automated system for identifying potential leads and generating emails.
- Developed web scrapers to extract and clean data about companies with active hiring needs.
- Integrated Langchain & OpenAI to create personalized email templates based on the scraped data, enhancing outreach for staffing services.

CurateSphere - A News Summarizer Platform Python,Fast API,Open AI,Postgres,Qdrant

- Developed a comprehensive news summarization platform, CurateSphere, capable of condensing news articles into concise summaries of under 5 minutes reading time.
- Implemented a robust system for scraping and ingesting news content from diverse sources, utilizing Qdrant for storing article embeddings and appending succinct summaries (TLDR) to each article.
- Ranks news articles on a priority basis and selects top 10 articles based on score generated and summarizes them as top news.

Data Processing Service Python, Azure, OCR, LLM, PostgreSQL

- Developed a comprehensive data processing service supporting documents, audio, and LLM utilities.
- Built functionality to upload documents to Azure, perform OCR, reconstruct tables, and chunk content into context-aware sections for vector processing.
- Implemented optional question generation on document chunks to enhance RAG in chat applications.
- Integrated audio data transcription and embedding generation for audio-based RAG, improving the accuracy of information retrieval.
- Developed tools for structured data extraction and entity recognition to facilitate precise information processing.
- Designed an ODM parser enabling agentic retrieval of data from databases, streamlining complex queries.

Spot Instance Manager Python, Fast API, PostgreSQL, boto3 sdk

- Created a minimalist Spot Manager Application that manages spot instances from AWS with auto stop and auto start features.
- Implemented a inferencing proxy for deployed models with VLLM that allocates a spot instance with desired AMI from existing pool of instances and start new instances based on availability
- Uses boto sdk for getting details, status of instanaces and availability group and fleet allocation and synced it with local postgres and redis
- Added functionality for supporting multiple AMI and instance types and fetching instance types from config file under single zone
- Added multiple background cron jobs to stop unused instances and continous data sync between instance pool

Technical Skills

Languages: Python,JavaScript/TypeScript, Go, Dart, Java, C/C++
Frameworks: Flutter, Node.js, Flask, FastAPI, Express
Tools: PostgreSQL, Redis, MongoDB, Elastic Search, AWS

Certifications **AWS Certified Cloud Practitioner - Validation Number**: WL78T7D25MQQ1C30