Vinay Kudari

vinay.kudari30@gmail.com
+1-469-943-6778

seeking a team that thrives on complex problems and candid, idea-driven dialogue

Relevant skills

Python, SQL, *FastAPI, LangGraph, PySpark, PyTorch, Hugging Face, Kubernetes, Redis, Snowflake*, **GCP, AWS**

Experience [6+ yrs]

- **Apple** — Cupertino, CA
  *Machine Learning Engineer | **workflow-automation, mcp, k8s, agents, fine-tuning*** — *July 2024 - Present*
  - Architected a MCP enabled asynchronous task processor that executes agent driven jobs in isolated k8s pods
  - Developed a multi-model NL→SQL pipeline where an FSDP-fine-tuned, domain-specific LLM rewrites user questions into org-aware, context-augmented queries, which a frontier LLM then compiles into precise, resource-efficient SQL
  - Delivered a policy-aware invoice intelligence system where LLM agents parse receipts from flat files, enrich them with org-specific policy data from Snowflake, and perform automatic validations, automating "read–query–verify–sign" workflows for 10K+ monthly invoices
  - Designed a deep-research finance analyst agent that retrieves, reasons over, and visualizes data from databases and flat files to generate accurate, evidence-backed financial insights
- **Gap Inc** — San Francisco, CA
  *Software Engineer | **java, spring, graphql, azure, kafka, kubernetes, llm*** — *July 2023 - June 2024*
  - Spearheaded a real-time personalization engine development, leveraging extensive datasets to optimize user experiences and drive significant engagement and sales enhancements.
  - Engineering a low-latency LLM-based information retrieval system to extract data from multi-source document corpus
- **Amazon Web Services** — Palo Alto, CA
  *Software Development Engineer | **spark, athena, dynamo, glue, lambda, cdk*** — *Dec 2022 - June 2023*
  - Designed and published an internal query builder python package (ZenQL) tailored for AWS Healthlake
  - Designed an event-driven server-less system using AWS services, allowing for parallel execution of tasks, resulting in seamless scaling to handle concurrent requests and improved processing throughput
  - Designed and optimized complex SQL queries in Athena resulting in a 40% reduction in query execution times
  - Architected near real time data processing system using AWS CDK, leveraging PySpark for ETL transforms on multi-terabyte healthcare data sets
- **Playstation** — San Francisco, CA
  *Software Development Intern | **kafka, concurrency, async, no-sql, cassandra*** — *May 2022 - Dec 2022*
  - Worked on auto follow feature for PS5 Explore Hub. Message queues and REST endpoints were developed
  - Learned custom framework and leveraged asynchronous programming pattern to optimize the latency by 60%
  - Developed a cross platform CLI tool to automate generation of grafana analytical dashboards from logs
- **Brave Orbit (early stage startup)** — Capetown, South Africa
  *Founding engineer | **rest-api, django, flutter, pytest, redis, docker, gcp*** — *Jan 2020 - Aug 2021*
  - Programmed a cross-platform app using BLoC pattern, implemented database, network and state management layers
  - Engineered HIPAA compliant scalable async push notification service for medication reminder application
  - Developed a task scheduler for clinical surveys that delivers email, SMS alerts and gathers response on a recurring basis
  - Architected a custom ERP system with multi-product substitutions and combo-product life cycle tracking features
- **Accenture** — Bangalore, India
  *Big Data Developer | **redshift, pyspark, emr, etl, s3*** — *Sep 2018 - Dec 2019*
  - Performed time series analysis on drug assay data to identify raw materials that effect the drug quality
  - Designed ETL data validation pipeline using micro service architecture; Manual testing resources were cut by 70%
  - Developed PL/SQL procedures to generate parent child hierarchies using temporal drug life cycle data

Projects

- **Pixie**: Built in 6 hours at the Nano Banana Hackathon; enables selective edits on live website screenshots or mockups using natural language or voice, with multi-model generation and real-time canvas updates
- **Friend (@mydawg_bot on telegram)**: Building a multi-agent AI companion with human-like memory simulation, capable of personalized conversations, context-aware recall, and tool use
- **Stocklerts**: Built a self-evolving system that analyzes real-time news to recommend stocks for day trading; adapts prompts weekly based on ground-truth top performers to refine future predictions
- **Open-source contributions**: Contributor to HuggingFace Datasets official repository

Education

- **University at Buffalo [3.9/4.0 GPA]** — Buffalo, NY
  Master of Science in Computer Science (AI/ML Specialization) — *Sep 2021 - Dec 2022*
  ***Research:** Finding inconsistency between text and related images, advised by Prof. David Doermann*
- **National Institute of Technology, Durgapur [3.3/4.0 GPA]** — West Bengal, India
  Bachelor of Technology in Computer Science and Engineering — *July 2014 - May 2018*