KIRAN KUMAR VADAKARA

kirankumar.vkk.12345@gmail.com | (281) 250-2754 | https://www.linkedin.com/in/kiranvkk/

EDUCATION

University of Houston (UH) – Dean's List **Houston, Texas**

Master of Engineering Data Science.

Coursework: Introduction to Data Science, Database Management, Machine Learning for Large Datasets, Cloud Computing, Big Data Analytics, Digital Image Processing, Probability and statistics.

SRM University (SRM) **Chennai, India**

Bachelor of Technology in Computer Science and Engineering.

WORK EXPERIENCE

Capital one. **New York, USA**

Gen AI Engineer. August 2024 - Present

- Fortified AI application data retrieval by orchestrating vector databases including ChromaDB and Pinecone, achieving a 35% improvement in data access times using indexing strategies.
- Architected end-to-end AI solutions using LangChain and LlamaIndex, achieving a 15% reduction in model deployment costs through strategic model selection and customized configurations.
- Integrated AWS Bedrock with existing AI infrastructure, ensuring 99.99% uptime for all AI applications, thereby exceeding the company's service level objectives by a wide margin.

Cognizant Technology Solutions. **Bangalore, India**

Data Engineer. Client: Bayer AG January 2021 - June 2022

- Migrated heavy workloads from Snowflake to AWS Redshift Spectrum and Athena, leveraging cost-effective querying on S3 instead of always using compute-heavy databases.
- Constructed real-time ETL pipelines, partnering with backend lead, for Tableau visualizations; reduced manual data handling by 90% using AWS Glue, Lambda, Kinesis, and Kafka, and improved report accuracy.
- Integrated a new data tier for Tableau dashboards, aligning with cross-functional team's blueprint, personally resolving 3 critical data inconsistencies, and ensuring proper data governance compliance**.**

ACADEMIC PROJECTS

Question Answering System with OpenAI and Pinecone **January 2025 – Present**

- Developed a smart Q&A system using OpenAI's(gpt-3.5-turbo) and Pinecone, delivering accurate responses in real-time.
- Integrated Pinecone as a vector database, allowing fast and scalable similarity searches across millions of documents. Optimized API response time with asynchronous processing and caching, improving performance by 25%.
- Developed a real-time indexing pipeline, reducing document processing latency by 30% with optimized vector storage and retrieval.

Cloud Computing Project **University of Houston | January 2024 – April 2024**

- Built an ETL pipeline for ~1 TB Realtime pothole detection data set using Spark and MapReduce. Engineered feature extraction techniques from sensor data, GPS coordinates, and road reports to enhance pothole detection accuracy.
- Designed, developed, and optimized a real-time pothole detection service using AWS Fargate, Aurora, and Elastic Load Balancer, achieving 4,000 RPS throughput with 12ms latency.

Walmart Sales Analysis Dashboard **University of Houston | May2023 – December 2023**

- Transformed and processed data by using Data Interpreter to ensure data completeness and validity.
- Visualized and presented data-driven Tableau dashboards to class, highlighting key trends in sales performance.
- Enhanced data utilization, implementing advanced technologies including machine learning algorithms which refined predictive accuracy by 25% for sales forecasting dashboards.

SKILLS

- Programming Languages: Python, Java.
- Frameworks & Tools: TensorFlow, PyTorch, LangChain, LlamaIndex, Streamlit, Flask.
- Generative AI Technologies: Open-source and paid LLM models (Llama2, Mistral,OpenAI,Google Gemini Pro)
- Vector Databases: ChromaDB, Pinecone, FAISS.
- Deployment Platforms: AWS Bedrock, AWS (EC2, Lambda), Azure Functions, Hugging Face Spaces
- AI/ML Techniques: Fine-tuning with custom data, vector embedding, NLP, neural network optimization,MLOPS, Dockers.