Girish Veeraraghavan **ocpgiri@gmail.com**

16 years of experience leading design and development of application in AI/ML, Generative AI domains.

- Developed systems, tools & processes to train / finetune large language models for text generation/summarization/question answer task.
- Established scalable, efficient, and automated processes for large scale ML model training & deployments.
- Helping organization onboard AI / ML capabilities leveraging cloud native platforms like AWS Sage maker, AWS bedrock, Azure Open AI , Co Pilot , AI Search.
- Providing cadence implementing ML OPS, Dev OPS and LLM OPS for large enterprise.
- Proficient in conversational AI like Google Dialog flow, Amazon Lex, Azure bot framework and Open RASA.
- Proficient in semantic search, Azure open AI search, Google vertex AI search and conversation.
- Expertise with operationalizing, monitoring, and scaling machine learning models and pipelines in cloud ecosystems
- Involved in the entire data science project life cycle and actively involved in all the phases, including data (extraction, cleaning), statistical modeling, and data visualization with large data sets.
- Design and development of knowledge graph and ontologies in neo4j for clinical domain.
- Develop search capability using solr and elastic search leveraging semantic linking.
- Worked on Health care data formats like FHIR/HL-7, ICD -10, OMOP.
- Experience across ML lifecycle including model evaluation, experiment tracking, model serving and productionizing ML pipelines (e.g., MLflow, Weights & Biases)
- Design and develop data pipelines and data models supporting large scale volume of structured and unstructured data to provide meaningful insights.
- Experiences with applying statistical techniques such as regression, ANOVA, cluster analysis, Factor. Analysis, time series forecasting, experimental and design, to solve business problem.
- Proficient in hyper parameter tuning and analyzing model metrics.
- Development experience in Agile/Scrum methodologies.
- Experienced designing system for MLops, Devops, Dataops.

Technical Expertise:

- AWS /Azure/GCP Solution Architecture.
- Data visualization tools: matplotlib, seaborn, T-SNE, bokeh, plotty.
- AI/ML: Llama2, bedrock Pytorch, TensorFlow, Keras, Scikit-learn, BERT.
- Search Engine: Elastic Search, FAAIS.
- NLP: Spacy, NLTK, C-takes, UMIA. RASA
- Statistical testing, ANOVA, Chi-square, Spearman.
- Distributed Systems: Kafka, Message Queues, AWS Kinesis/Firehose, Spark structured Streaming.
- Databases: PostgreSQL, Cassandra, AWS DynamoDB, blob stores [S3].
- Data OPS/ML Ops; Apache Airflow, Talend Databricks.
- Programing language: Java, spring boot, python.
- API management: Apigee, Swagger Hub, Stoplight.

Gen AI Architect (Independent Consulting) Feb 2024 – Till date

Providing leadership adopting Generative AI to solve business problems leveraging Azure/AWS/GCP.

- Identify. Opportunity for Automation in Intelligent Document Processing, Question Answering, Document summarization, Conversation AI and Call Deflection.
- Helping organizations leverage Gen AI to automate task by proving virtual assistant, knowledge retrieval systems.
- Designed call deflection reference architecture for leading bank, leveraging Azure Co Pilot, and Open AI Services
    - The reference architecture involved integrating Genesys with Audi codes and calling Microsoft copilot for automating user intent instead of routing to real agents.

Sentry Insurance Madison WI Sep 2022 - Jan 2024

Sr. Solution Architect – AI/ML

Responsibility

- Identify opportunities for automation in claim processing and underwriting using Conversational AI and Generative AI.
- Helped reduce claims processing time by 20%, leveraging custom BERT based NER models and GEN AI Intent recognition and summarization.
- Provided technical leadership and architecture guidance defining text mining solution roadmap for automating workers compensation claim notes.
- Define technical roadmap leveraging machine learning capabilities across the enterprise for solving NLP problems.
- Designed intelligent document processing leveraging Azure AI document intelligence, supporting loss runs calculation.
- Created technical blueprint for RAG [Retrieval augmented generation] using large Llama2 and Lang chains.
- Designed query expansion by incorporating LLMS to expands query precision.
- Finetuned LLM model by grounding and leveraging custom text generation strategies for context sensitive domains.
- Designed and developed automated evaluation of LLM text summarization using BERT, ROUGE METEOR scores helping reduce Perplexity by finetuning and identifying optimal scores.
- Designed and developed agent assist chat bot leveraging Azure co-pilot to assist HR Operations.
- Created technical blueprint supporting topic summary and question answering using vector DB and LLM.
- Designed vector search engine using IVFPQ, HNSW leveraging elastic search.
- Designed enterprise wide data catalog and labeling services leveraging AI and NLP.
- Designed ETL pipelines to ingest claim notes into vector DB leveraging Llama and langchains.
- Involved in finetuning and prompt engineering of large language models.
- Architected intelligent document processing supporting insurance claims and loss runs. Using text extract and AWS comprehend.
- Designed and developed real time end points using python Kafka daemons consuming messages from Kafka topics.
- Design data pipelines and feature store supporting insurance scoring models.
- Designed feature stores, data preparation pipelines and ML workflows.
- Improved text mining capability for categorizing workers compensation claims by leveraging spacy capabilities like Lemmatization, entity recognition and negations.
- Used negspacy to identify negation in medical claim notes.
- Designed Named Entity recognition/ Named Entity Extraction using deep neural network.
- Creating data ingestion pipelines for ingesting unstructured data into elastic search.
- Improving search relevance and search recall.
- Providing technical leadership on setting up ML platform and supporting ML OPS initiatives.
- Optimizing model training process for pre training and fine tuning.
- Hyper parameter tuning and deploying models to production.
- Create documentation or training materials for Project team members.

SKAEL [Startup] Remote **Jan 2022 – July 2022**

NLP /Conversation AI Architect

Responsibility

- Designed and developed conversational AI chatbot using RASA for HR workflow automation.
- Designed handling co reference modules.
- Designing in memory tracker stores.
- Designed intent recognition model to identify user task.
- Design co reference resolution.
- Integrated transformer model for improving intent recognition.
- Create API endpoints for model inference.
- Created future prompts by analyzing past conversations and decisions.
- Deployed and productionized chatbot for pilot customers.

Ebsco Information Service Ipswich MA **Jan 2019 - Oct 2021**

Principal Architect – AI/ML/Search and Content

Responsibility

- Provide cadence to product managers identify feature prioritization for product rollouts.
- Identify opportunities to leverage NLP to support search and retrieval of medical documents.
- Improved search precision by 20% by incorporating niche data labeling services and leveraging language models.
- Designed neo4j based properties graph supporting SNOMED CT ontology.
- Develop and maintain data models, schemas, and ontologies for the knowledge graph.
- Leverage knowledge graph capabilities to augment semantic search.
- Improved synonym and surface form search by incorporating query expansion features.
- Create golden sets and evaluate how search performs against current golden sets.
- Leveraged NLP methodologies for solving Acronym disambiguation.
- Design and development of NLP layer supporting entity recognition for searching medical documents.
- Designed NLP pipelines processing large quantities of unstructured data to extract meaningful insights.
- Created data transformation layer to process HL7 into (OMOP) Common Data Model (CDM).
- Created ingestion layer parsing ICD 10 codes from FHIR/HL7 messages to enrich meta data for search capability.
- Design and finetuned transformer based Medical BERT supporting SNOMED-CT/RXNORM concepts for Question/Answering to improve semantic search supporting query expansion.
- knowledge extraction for improving search behavior.
- Design and development of search capability [ranking, filtering] using elastic search.
- Designed Acronym disambiguation using spacy toolkit.
- Explored and integrated CTR into search pipeline, identified query tail logs and improved search recalls.
- Improved semantic search, by expanding query expansion using knowledge graph.
- Extracted word embeddings from knowledge graph and incorporated the embeddings into BERT layer.
- Finetuned hugging face transformer models to support domain specific data for improving search relevance.
- Design and development of API first microservices supporting model inferences.
- Created conversational chat bot using RASA framework incorporating BERT embeddings.
- Created custom properties and relationship in knowledge graph to support custom queries.
- Capacity planning for elastic search cluster.

Direct Supply Milwaukee, WI **Apr 2015 - Dec 2018**
Architect /Database Engineer II

Responsibility

- Provide governance to ensure engineering best practices are followed.
- Weighing Cost Vs Efficiency Vs Effectives. Present and evaluate design solutions objectively and facilitate conflict resolution.
- Involved in the decision making of selecting RDBS technology to replace on premise SQL server.
- Identified key KPI for launching PostgreSQL as DBAS.
- Evaluated PostgreSQL, identified benchmark requirements, developed automatic creation and deployment strategy, implemented monitoring, backup and restore.
- Defined and implemented monitoring layer for PostgreSQL, with key alerting metrics.
- Developed inhouse query analyzer and metrics tools, like SQL performance analyzer.
- Fine tune SQL query before deploying to production.
- Facilitated launching of initial DBAS platform into private cloud.
- Involved in migrating on premise monolith application to aws cloud.
- Designed ETL layer exporting on premise product catalogs to cloud.
- Create code quality gates, integrated gates into existing CI/CD pipeline.
- Designed CI/CD deployment of database artifacts using Jenkins/Liquibase integration.
- Supported application development team to migrate first on-premises web based to AWS cloud.
- Initiated adoption of Domain driven design and spearheaded microservice adoption
- Created architecture for document search using elastic search in cloud.
- Developed Spark Streaming layer for ingesting CQRS events.
- Ingested clicks stream data as parquet files, implemented Impala to query, abundant shopping cart for last one hour.
- Developed architecture for file search using Apache Manifest and Solr Cloud

Selsoft Inc Allen, TX [American Express] **July 2013 – April 2015**
Java Architect

Responsibilities:

- Designed cluster aware asynchronous REST API application using Akka Actor Model as part of POC,
- Designed catalog for Json Feed using Avro.
- Designed and Developed POC on JSON validation using Rhino JavaScript Engine, designed custom rules using parboiled.
- Importing and exporting data into HDFS and Hive using Sqoop.
- Involved in developing Pig scripts to process the data.
- Involved in defining job flows and developed job flows.
- Experienced in running Hadoop streaming jobs to process terabytes of data in xml format.
- Load and transform large sets of structured, semi structured, and unstructured data.
- Responsible to manage data coming from different sources.
- Developed Map Reduce Programs for analyzing weblogs.
- Involved in loading data from UNIX file system to HDFS.
- Developed Hive UDFs.
- Involved in creating Hive tables, loading data, and writing hive queries.
- Developed scripts for bulk loading of data into HBase Tables.
- Assist with the addition of Hadoop processing to the IT infrastructure.
- Perform data analysis using Hive and Pig.

Causeway Technologies, Bangalore, India **April 2012 – March 2013**
Architect
Technology: OSGI, FUSE ESB, Apache CXF, spring, Hibernate, Active MQ, GIT, Eclipse, Windows

Responsibility

- Role Involves Designing bundles for Business/DAO/ Messaging Layer, involves Message Routing [EIP] and exposing services as REST end point.
- Provide Architecture Solutions [SIP] Service Integration Platform [ADF +JBPM integration]
- Designed OSGI Bundles for Data Access Layer
- Exposed Connection Pool as a Service in OSGI Layer
- Designed and Implemented Architecture for JPA layer in OSGI.
- Integrated j2ee [Wavemaker, JBPM, BIM] application suite to OSGI [Fuse ESB]
- Designed Multi Tenancy in wavemaker [Application Development Framework]

Education:

Master of Computer Science, Bharathiar University Coimbatore, India **Apr 2000 - Apr 2002**