Vinod Kumar Vemula
(279) 500-3888 | vvinodvemula03@gmail.com | LinkedIn | Dallas-TX

PROFESSIONAL SUMMARY

Senior AI/Data Engineer with 7 years building production GenAI systems (RAG, multi-agent, LLM fine-tuning) and scalable data platforms on Azure/AWS/GCP. Delivered solutions processing 10M+ daily events, reducing costs by 20-30% and accelerating deployments in finance and IoT.

EXPERIENCE

AI Engineer, CGI, Dallas, TX **Oct 2025–Present**

- Architected and deployed production-grade GenAI microservices using FastAPI, Flask, Docker, and Kubernetes; delivered end-to-end solutions leveraging OpenAI GPT, Anthropic Claude, and Google Gemini models with advanced prompt engineering and optimal model selection.
- Fine-tuned large language models using Hugging Face Transformers, LoRA, QLoRA, PEFT, and Unsloth, achieving improved domain specialization, reduced inference costs, and higher performance on targeted tasks.
- Designed and implemented multi-agent systems with LangGraph, AutoGen, and CrewAI to enable structured collaboration and automate complex workflows; enhanced knowledge-driven reasoning by integrating agents with Neo4j knowledge graphs, REST APIs, and real-time data connectors for entity-aware inference.
- Engineered advanced Retrieval-Augmented Generation (RAG) pipelines using LangChain, LlamaIndex, FAISS, Pinecone, and Weaviate, significantly improving retrieval accuracy and grounding LLM responses in reliable semantic search.
- Built memory-augmented agent systems with persistent long-term memory stores and moderation frameworks, ensuring contextual continuity, reasoning alignment, and reduced hallucinations in production deployments.
- Accelerated data ingestion and preparation for LLM training/fine-tuning using PySpark, Apache Airflow, and Azure Data Factory, processing large-scale datasets efficiently.
- Developed comprehensive evaluation frameworks with RAGAS, LangChain evaluators, and reinforcement learning-based scoring to quantify and optimize key LLM metrics including coherence, relevance, factual accuracy, and hallucination rates.

Data Engineer, BMO Bank, Chicago, IL **Jun 2023-Sept 2025**

- Developed end-to-end GenAI applications (GPT, Claude, Gemini) and architected Retrieval-Augmented Generation (RAG) pipelines (LangChain, FAISS) to enhance semantic search and contextual grounding.
- Built agentic and multi-agent AI systems (LangGraph, AutoGen) and integrated memory-augmented LLM systems with safety-aware reasoning for reliable and consistent outputs.
- Designed data models in Azure Synapse and SQL Server for regulatory compliance and financial reporting for 50+ stakeholders.
- Implemented CI/CD pipelines with Azure DevOps and ARM templates, accelerating deployment cycles by 30%.
- Built Power BI dashboards from Synapse data, delivering insights to management on $1B+ portfolios.
- Streamlined data ingestion with Azure Kubernetes Service and Docker, improving scalability for 10M+ daily transactions.

Data Engineer – Intern | John Deere, Moline, IL **Aug 2022 – Jan 2023**

- Built and optimized cloud-based data pipelines for large-scale IoT sensor data using AWS services including S3, Lambda, and Redshift, supporting analytics across 10,000+ connected machines.
- Automated infrastructure provisioning using Terraform, CloudFormation, reducing environment setup time by ~25%.
- Developed serverless workflows to process real-time telemetry data, improving operational visibility and lowering infrastructure costs by ~20%.
- Collaborated with senior data engineers and platform teams to monitor pipeline health using CloudWatch and Grafana, contributing to 99.8% system uptime.

Sr. Technology Support, Infosys, India **Mar 2019–Dec 2021**

- Automated CI/CD workflows with Cloud Composer, reducing deployment errors by 15%.
- Utilized Python and SQL for data transformation, supporting sales analytics for 50+ products.
- Implemented Cloud Pub/Sub and Kafka for real-time analytics, processing 500K+ daily transactions.
- Leveraged Cloud Natural Language API to analyze customer sentiment, improving product satisfaction by 10%.
- Monitored and troubleshot cloud-based data pipelines and messaging systems, performing root-cause analysis to resolve production issues and improve system stability and uptime.

Data Engineer, Aviva Life Insurance, Bangalore, India **Jan 2018–Feb 2019**

- Developed BigQuery ETL pipelines for 500K+ insurance policies with 99.9% uptime.
- Created Power BI dashboards integrated with BigQuery, enhancing sales tracking for 200+ agents.
- Integrated AWS S3 and Redshift with GCP for multi-cloud processing, reducing data transfer costs by 15%.
- Optimized SQL queries and BigQuery job configurations to improve query performance and reduce processing time and operational costs.
- Used Sqoop and Impala for high-speed analytics on claims data, supporting regulatory reporting for 100K+ policies.
- Implemented data quality checks, validation rules, and scheduled monitoring for ETL pipelines to ensure accuracy, consistency, and compliance in insurance reporting systems.

SKILLS

Visualization: Power BI, Tableau, AWS Quicksight
GenAI & LLMs: OpenAI GPT, Claude, Gemini, Hugging Face, LoRA/PEFT/Unsloth, Prompt Engineering.
Retrieval, RAG & Multi-Agent: RAG, LangChain, LlamaIndex, LangGraph, AutoGen, CrewAI, RAGAS, FAISS/Pinecone/Weaviate.
ML/ Deep Learning: PyTorch, TensorFlow, Scikit-learn, XGBoost, LightGBM, CNNs, Optuna
Cloud: Azure (ADF, Synapse, Databricks), AWS (S3, Redshift, Glue, Lambda), GCP (BigQuery, Dataflow, Pub/Sub)
Big Data: Hadoop, Spark/Pyspark, Hive, Kafka, Kinesis
Programming: Python, SQL, Scala, Java
Tools: Airflow, Terraform, Jenkins, Docker, Kubernetes
Databases: PostgreSQL, MySQL, MongoDB, Snowflake, DynamoDB

EDUCATION

Masters in Statistics and Decision Analytics Western Illinois University, USA, **GPA- 3.60**
Bachelor of Technology in Mechanical Engineering, India, **GPA- 3.00**

CERTIFICATIONS

- **Azure Data Engineer Associate** – Microsoft, 2023
- **AWS Data Engineer** – Amazon Web Services, 2023
- **Certified Generative AI Expert** - Udemy
- **GCP Professional Data Engineer** – Google Cloud, 2021
- **Azure AI / Data / Fundamentals** – Microsoft, 2021
- **Azure Administrator Associate** – Microsoft, 2021 (Expired 2022)
- **GCP Associate Cloud Engineer** – Google Cloud
- **DataCamp** (R & Visualization Track) – Multiple certificates, 2022
- **SnowFlake-** SnowPro Core

PROJECTS

Predictive Churn Model
Product based review
VitaminD_deficiency_prediction
Intrusion-detection-system_2018