Rajan Gurung

Data Analyst / AI Engineer | LinkedIn

Winston-Salem NC 27101 | **rajangurung7412@gmail.com** | **704-870-7416**

PROFESSIONAL SUMMARY:

Curious and hands-on Data Scientist and ML Engineer with 4+ years of experience building real-world AI systems that actually scale, deliver insights, and make a difference. From deploying LLMs and GenAI solutions at enterprise scale to optimizing big data pipelines in fast-paced environments, I bring a solid mix of deep technical skills and practical problem-solving. Experienced with tools like Python, PyTorch, Hugging Face, Spark, Kafka, and AWS, with a strong focus on MLOps, NLP, and responsible AI. Known for working cross-functionally, turning research into production systems, and constantly exploring the latest in AI trends to stay ahead of the curve.

PROFESSIONAL EXPERIENCE:

Data Scientist / ML Engineer **Jan 2024 to Present**

Ford Motors

Project: Enterprise LLM Platform for Scalable and Responsible AI Deployment

Design and implementation of a robust Large Language Model (LLM) platform at Ford Motors to power scalable, secure, and ethically responsible AI applications across the enterprise. The platform leveraged AWS UltraClusters, PyTorch, Hugging Face Transformers, and custom VectorDB solutions to support high-performance model training, inference, and similarity search. Integrated state-of-the-art LLM optimization techniques including quantization, pruning, and parallelization to improve latency, throughput, and hardware efficiency.

- Designed, developed, and deployed AI-powered software components including LLM training pipelines, inference services, similarity search, and model observability frameworks.
- Built scalable ML infrastructure on AWS UltraClusters using PyTorch, Hugging Face Transformers, and custom VectorDB implementations.
- Engineered cost-efficient AI systems by optimizing compute utilization and memory bandwidth across large-scale model training and inference workflows.
- Integrated LLM optimization techniques such as quantization, pruning, and mixed-precision training to reduce latency and improve throughput.
- Applied responsible AI principles, integrating Nemo Guardrails and fairness audits to ensure transparency, security, and ethical model deployment.
- Led deployment and monitoring of generative AI applications, ensuring compliance with internal governance and regulatory standards.
- Partnered with research and engineering teams to translate cutting-edge ML research into deployable production solutions.
- Conducted rigorous evaluations of LLMs through fine-tuning, reinforcement learning (RLHF), and human feedback loops to drive model performance.
- Developed APIs and modular AI services to support enterprise-level integration of AI capabilities across departments.
- Designed scalable similarity search mechanisms using vector embeddings, ANN (Approximate Nearest Neighbor) algorithms, and memory-optimized retrieval systems.
- Collaborated with cross-functional teams (product, engineering, operations) to align AI roadmaps with business KPIs and customer impact goals.
- Maintained version-controlled ML pipelines using MLflow and CI/CD integrations with cloud-based deployment tools (e.g., SageMaker, Lambda).
- Built dashboards and monitoring tools for model observability, drift detection, and continuous performance tuning in production.
- Utilized parallel computing frameworks (e.g., Ray, Dask) for distributed training of multi-billion parameter LLMs.

- Executed data preprocessing and feature engineering pipelines for unstructured, multi-modal datasets including vehicle sensor logs and driver behavior patterns.
- Applied Transformer-based architectures and RAG (Retrieval-Augmented Generation) models for enterprise knowledge retrieval systems.
- Designed experiments using A/B testing and statistical significance metrics to validate model effectiveness in real-world scenarios.
- Led internal workshops and knowledge-sharing sessions to upskill teams in NLP, GenAI, and MLOps best practices.
- Ensured secure and compliant AI systems by implementing data privacy practices aligned with GDPR and internal audit policies.
- Automated retraining workflows and model lifecycle management using Airflow and Kubernetes-based orchestration systems.
- Contributed to open-source initiatives and internal toolkits to accelerate AI adoption across the organization.

Tools and Technologies: Python, PyTorch, Hugging Face Transformers, TensorFlow, AWS (SageMaker, EC2, Lambda, UltraClusters), Vector Databases (FAISS, Pinecone), Docker, Kubernetes, MLflow, DVC, Ray, Dask, Airflow, Git, REST APIs, FastAPI, LangChain, NumPy, Pandas, Scikit-learn, NeMo Guardrails, OpenAI API, LLM Fine-tuning, RLHF, Prompt Engineering, RAG (Retrieval-Augmented Generation), Transformers, CI/CD, Model Monitoring, Data Versioning, Experiment Tracking, Postman, VS Code, Jupyter Notebooks

Data Scientist **Oct 2020 to Nov 2022**

Uptechsys - Kathmandu NP

Project: Real-Time Data Pipeline and Analytics Platform for Enterprise Intelligence

Developed and managed a real-time data analytics platform to support enterprise reporting, performance monitoring, and business intelligence initiatives. Designed and deployed scalable data pipelines handling structured and semi-structured datasets using Python, SQL, and Apache Spark. Implemented robust ETL workflows for both batch and streaming use cases with Apache Kafka and Hadoop ecosystems. Optimized data ingestion and transformation processes to ensure low-latency insights and high system availability.

- Designed, implemented, and maintained robust, end-to-end data pipelines to collect, process, and analyze structured and semi-structured data from diverse sources, ensuring high scalability and performance.
- Automated ETL workflows using Python and SQL, significantly reducing manual data preparation time and enabling faster, more reliable analytics delivery for business units.
- Developed distributed data processing systems using Apache Spark and Hadoop, enabling parallel computation on large datasets and accelerating data transformation tasks.
- Implemented real-time streaming data pipelines using Apache Kafka to support operational dashboards, alerting systems, and live analytics use cases.
- Created reusable data ingestion frameworks to normalize and clean diverse formats such as JSON, XML, and CSV, facilitating downstream analytics and ML workflows.
- Managed and optimized both relational (MySQL, PostgreSQL) and NoSQL (MongoDB) databases to ensure efficient data storage, retrieval, and scalability across environments.
- Established rigorous data validation, transformation, and schema-checking mechanisms to maintain data integrity and trust across departments.
- Conducted exploratory data analysis (EDA), statistical profiling, and data visualization to extract key business insights and inform product decisions.
- Designed and delivered real-time and historical dashboards using React and NodeJS, enabling stakeholders to monitor KPIs and metrics through intuitive interfaces.
- Collaborated with cross-functional teams including software engineers, UI/UX designers, product owners, and data analysts to define data strategy and build integrated solutions.

- Developed and scheduled recurring data jobs using Apache Airflow and Linux Cron, automating end-to-end pipeline execution and monitoring.
- Built a library of modular, extensible Python ETL components to accelerate new pipeline creation and reduce code duplication across projects.
- Tuned SQL queries and optimized database indexing strategies to improve query execution time and minimize server load during peak usage.
- Fine-tuned Spark and Kafka cluster configurations to ensure optimal throughput and latency in large-scale batch and stream processing workloads.
- Participated in the design and review of system architecture, contributing data engineering expertise to support scalability, fault-tolerance, and maintainability.
- Deployed containerized data microservices using Docker and Kubernetes to support distributed deployment and infrastructure automation.
- Conducted root cause analysis and implemented recovery strategies for pipeline failures, ensuring data availability and minimizing disruption to analytics services.
- Supported data governance initiatives by maintaining metadata documentation, enforcing naming conventions, and tracking data lineage across platforms.
- Created internal data dictionaries, user guides, and onboarding materials to enable self-service data access and knowledge sharing among engineering and analytics teams.
- Contributed to CI/CD automation for data workflows, integrating testing, validation, and deployment processes for more reliable development cycles.
- Mentored junior team members on data engineering best practices, code reviews, and debugging techniques to support skill development and team performance.

Tools and Technologies: Python, SQL, Apache Spark, Apache Kafka, Hadoop, MySQL, PostgreSQL, MongoDB, Airflow, Docker, Kubernetes, Linux Cron, Pandas, NumPy, Scikit-learn, Git, Jupyter Notebooks, Node.js, React, REST APIs, JSON, XML, ETL Frameworks, Data Validation Tools, Data Visualization Libraries (Matplotlib, Seaborn).

EDUCATION

Master of Data Science - Carolina University, Winston-Salem, NC, USA
B.Sc. (Hons) in Computing – *Coventry University (Softwarica College), UK & Nepal*

PROJECTS

• Automated Zoo Monitoring System (IoT): Led development using Arduino UNO, GPS modules, and Wi-Fi-enabled Arduino Nano to enable real-time animal tracking.
• Intruder Detection System for Smart Farming: Deployed real-time detection using TensorFlow and YOLOv3.
• Tiger Detection System (Image Recognition): Implemented using Raspberry Pi and TensorFlow.
• EV Market Analysis (ML): Built predictive models to evaluate market adaptability using Python and Scikit-learn.
• Travel Helper Web App: Developed a travel assistant platform with location-based features.
• Thrift Auction App: Designed mobile/web solution using Kotlin, NodeJS, and JavaScript for thrift auctions.
• Find Workers App: Created Tinder-like platform matching service seekers with workers using React, NodeJS, and MongoDB.

CERTIFICATIONS

1) **IBM Artificial Intelligence Analyst – Mastery Award 2019**
2) **IBM Introduction to Data Analytics Certificate**
3) **IBM Introduction to Big Data with Spark and Hadoop Certificate**
4) **Google Advanced Data Analytics Certificate**
5) **AWS Certified Cloud Practitioner – June 23, 2025**

REFERENCE ON REQUEST