AJEET GUPTA

MLOps Architect | AI/ML Infrastructure Specialist

Noida, India | +91-8618214934 | ajeetgi108@gmail.com
LinkedIn: linkedin.com/in/ajeet-gupta-991a20bb | Medium: medium.com/@ajeet_

PROFESSIONAL SUMMARY

Results-driven MLOps Architect with 9+ years of experience designing and implementing scalable AI/ML infrastructure solutions. Expert in deploying GenAI applications, ML application, and orchestrating end-to-end ML pipelines. Proven track record of processing high-volume data (100K+ daily transactions) and implementing Responsible AI practices. Specialized in AWS cloud architecture, Kubernetes orchestration, and MLOps best practices with demonstrated ability to reduce operational costs by 30% and improve deployment efficiency by 40%.

CORE COMPETENCIES

AI/ML Technologies: GenAI/LLM Implementation | AWS Bedrock Claude | AutoGen Framework | NLP Processing | PII Redaction | Model Monitoring | MLFlow | Responsible AI

Cloud & Infrastructure: AWS (EC2, S3, VPC, IAM, Api Gateway, Sagemaker, Comprehend, Lambda, Step Functions, EventBridge, ECS Fargate, CloudFormation) | GCP Vertex AI | Kubernetes | Docker | Terraform | Helm Charts | Apache Airflow | Argo Workflow

Development & Operations: Python | FastAPI | CI/CD Pipelines | GitOps | Prometheus | Grafana | Seldon Core | PostgreSQL | DynamoDB

Data Processing: Real-time Chat Processing | IVR Data Analysis | Email Classification | SQL Query Generation | Data Pipeline Orchestration

PROFESSIONAL EXPERIENCE

MLOps Architect | Fractal Analytics | *Noida | Apr 2025 - Present*

GenAI Platform Development for Telecommunications Client

- Architected enterprise-scale serverless GenAI solution processing 100,000+ daily customer interactions across chat, IVR, and email channels using event-driven architecture with AWS Lambda, Step Functions, and EventBridge
- Designed multi-stage Step Function workflow orchestrating: (1) EventBridge triggering on scheduled time, (2) Lambda-based PII redaction using AWS Comprehend detecting 10+ entity types (names, emails, phone numbers), (3) parallel Lambda invocations for intent classification, sub-intent detection, toxicity analysis, and message summarization using AWS Bedrock Claude 3 Sonnet
- Deployed containerized NLP application on ECS Fargate cluster (2-20 tasks auto-scaling) in private subnets, exposed via internal Application Load Balancer for secure API access serving 5K+ requests/second

- Configured VPC endpoints for S3, DynamoDB, and Comprehend eliminating NAT Gateway costs, reducing data transfer expenses by 40% while improving security posture
- Helped Gen AI team to implement intelligent conversational SQL agent using AutoGen framework.

AIOps Platform Development

- Leading development of Responsible AI platform for model performance monitoring, fairness tracking, bias detection, and drift analysis across 20+ production models
- Orchestrated ML workflows using Apache Airflow processing 10GB+ daily data volumes including model predictions, feature distributions, and inference logs
- Implemented automated model drift detection using Evidently AI with configurable thresholds, preventing 95% of production issues through early alerts and automated remediation workflows
- Technologies: AWS, Apache Airflow, Docker, Terraform, PostgreSQL, Sagemaker, Evidently AI

MLOps Engineer | Corteva Agriscience | *Hyderabad | May 2020 - Mar 2025*

ML Infrastructure Modernization

- Led migration of ML pipelines from on-premises to Kubernetes, achieving 3x processing speed improvement through containerization and horizontal scaling
- Architected dual-mode ML serving infrastructure: (1) Batch inference using Argo Workflows for large-scale crop prediction models processing, (2) Real-time REST endpoints via Seldon Core serving 2K+ requests/second
- Designed Argo Workflow DAGs with dynamic parallelism processing 10+ simultaneous inference jobs across GPU-enabled Kubernetes nodes, reducing batch processing time from 3 hours to 35 minutes
- Built custom Argo Workflow templates for common ML patterns (data preprocessing, model training, batch inference, model evaluation) reducing pipeline development time by 60%
- Configured Seldon's canary deployment strategy enabling gradual model rollout with automatic traffic shifting based on performance metrics, achieving zero-downtime deployments
- Integrated Seldon with Prometheus for real-time metrics (request latency, throughput, error rates) and Grafana dashboards, implementing automated rollback on performance degradation

MLOps Platform with MLFlow and GitHub CI/CD

- Established centralized MLflow tracking server on Kubernetes with PostgreSQL backend and S3 artifact storage, managing 100+ experiments and 100+ registered models
- Built comprehensive experiment tracking framework capturing hyperparameters, metrics, dataset versions, and model artifacts with automatic lineage tracking
- Implemented MLflow Model Registry with stage transitions (Staging, Production, Archived) enforcing approval workflows before production promotion, integrated with Slack notifications
- Developed GitHub Actions CI/CD pipeline automating: (1) unit/integration testing with 85% code coverage requirement, (2) Docker image building with vulnerability scanning
- Built automated model validation pipeline in GitHub Actions comparing new model versions against production baseline on hold-out test set before Kubernetes deployment
- Reduced deployment time through automated CI/CD pipeline with complete audit trail from code to production

Senior Platform Engineer | **Quantiphi Analytics** | ***Bengaluru | Oct 2018 - May 2020***

Managed Data Science Platform

- Developed cloud-native Data Science Platform serving 100+ data scientists with JupyterHub on ECS, enabling self-service GPU-enabled notebook provisioning
- Configured GPU-enabled instances with NVIDIA CUDA 10.1 and cuDNN 7.6 for deep learning workloads, implementing pod autoscaling based on GPU utilization
- Built RESTful model serving microservices using Flask framework with Gunicorn WSGI server, handling 10K+ daily inference requests with automatic horizontal scaling
- Automated infrastructure provisioning using CloudFormation templates and CodePipeline, reducing environment setup time from days to hours

Governed Data Lake Implementation

- Architected centralized data ingestion platform for Western European Market processing 100GB+ daily from multiple sources using AWS Glue ETL jobs
- Implemented Security Threat Model using Microsoft STRIDE methodology, conducting comprehensive threat analysis across data ingestion, storage, and access layers
- Integrated Confluent Kafka with RDS PostgreSQL for real-time change data capture using Debezium connectors, ensuring <5 second data freshness
- Technologies: AWS, Docker, NVIDIA GPUs, ECS Fargate, Sagemaker, Python Flask, AngularJS, Confluent Kafka, CloudFormation

Cloud Engineer | **Manthan Software Services** | ***Bengaluru | Aug 2016 - Sep 2018***

- Deployed and managed multi-tier web applications on AWS cloud infrastructure using EC2, RDS, S3, and CloudFront with high availability across multiple AZs
- Designed VPC architecture with public/private subnets, NAT Gateways, security groups, and IAM policies following AWS Well-Architected Framework principles
- Built CI/CD pipelines using Jenkins with automated testing, Docker image building, and blue-green deployments, reducing deployment time by 50%
- Automated routine tasks using Python Boto3 and Shell scripts, saving 20 hours weekly in manual infrastructure operations

EDUCATION

Master of Science in Data Science | BITS Pilani | *2022 - 2024*

Bachelor of Technology in Computer Science | Guru Ghasidas University | *2012 - 2016*

CERTIFICATIONS

- AWS Certified Solutions Architect
- AWS Certified SysOps Administrator
- Fractal Certified MLOps Engineer
- Linux Foundation Certified System Administrator