

Vertex AI Engineer & Cloud Delivery Lead

Job Information

Recruiter
NEXUS Corporation

Job ID
1607641

Industry
IT Consulting

Job Type
Permanent Full-time

Location
Tokyo - 23 Wards

Salary
Negotiable, based on experience

Refreshed
November 23rd, 2026 01:00

General Requirements

Minimum Experience Level
Over 6 years

Career Level
Mid Career

Minimum English Level
Business Level

Minimum Japanese Level
Business Level

Minimum Education Level
Bachelor's Degree

Visa Status
Permission to work in Japan required

Job Description

We are seeking an experienced Vertex AI Engineer and Cloud Delivery Lead to drive the design, deployment, and operationalisation of machine learning solutions on Google Cloud. This role bridges AI/ML engineering and cloud delivery, ensuring that models and pipelines reach production reliably, securely, and at scale.

Job Description - Grade Specific

Key Responsibilities:

- Design and implement end-to-end MLOps pipelines on Vertex AI, including data ingestion, model training, evaluation, and deployment
- Build and manage Vertex AI Pipelines (Kubeflow Pipelines) for automated model training and retraining workflows
- Deploy and manage models using Vertex AI Model Registry, Endpoints, and Batch Prediction services
- Implement feature engineering workflows using Vertex AI Feature Store
- Develop GCP-native integrations connecting Vertex AI with BigQuery, Dataflow, Cloud Storage, and Pub/Sub
- Manage infrastructure for ML workloads using Terraform, ensuring reproducible and version-controlled environments
- Configure IAM policies for Vertex AI workloads including service account governance and VPC Service Controls
- Lead cloud delivery activities: sprint planning, release management, environment promotion, and stakeholder communication

- Establish model monitoring using Vertex AI Model Monitoring for data drift and skew detection
- Collaborate with data scientists to containerise experiments and promote models through dev/staging/production
- Drive adoption of GKE for model serving workloads where custom inference infrastructure is required

Required Skills

Required Qualifications:

- 4+ years of experience with GCP, including 2+ years hands-on with Vertex AI
- Strong proficiency in Python and ML frameworks (TensorFlow, PyTorch, Scikit-learn)
- Experience building Vertex AI Pipelines and managing model lifecycle in Vertex AI Model Registry
- Solid Terraform skills for provisioning Vertex AI, GCS, BigQuery, and associated infrastructure
- Good understanding of GCP IAM, particularly for securing ML pipelines and data access
- Experience with GCP-native development patterns and event-driven architectures
- Demonstrated cloud delivery experience including planning, execution, and stakeholder management
- Familiarity with containerisation (Docker) and GKE for model serving

Preferred Qualifications:

- Google Professional Machine Learning Engineer certification
- Experience with LLM fine-tuning, Vertex AI Generative AI Studio, or Model Garden
- Familiarity with Feast, Tecton, or similar feature stores
- Experience with Ansible for environment configuration and automation
- Background in DataOps or platform engineering for data-intensive workloads

Company Description