

Google Cloud Vertex AI Platform

G-Cloud 15 – Service Definition

Service Overview	2
Key Features	2
• Model Garden & Foundation Models	2
• Generative AI Development Suite	2
• MLOps & Infrastructure	3
Key Benefits	3
Target Use Cases	3
Contact Information	4

Service Overview

Vertex AI is Google Cloud's unified, enterprise-grade AI development platform designed to accelerate the end-to-end machine learning (ML) lifecycle. It provides a single environment where data scientists and developers can build, deploy, and scale both generative AI (GenAI) and traditional predictive ML models. By integrating Google's world-class foundation models with robust MLOps tools, Vertex AI allows organisations to move from experimental prototyping to production-grade AI applications with unprecedented speed, security, and governance.

Key Features

Vertex AI provides a comprehensive suite of tools categorised into three main areas:

- Model Garden & Foundation Models
 - Gemini Models: Access to Google's most capable multimodal models (Gemini 1.5 Pro/Flash) for text, code, image, and video processing.
 - Diverse Model Library: A curated collection of 150+ first-party (Imagen, Chirp), third-party (Anthropic's Claude), and open-source models (Llama 3, Gemma).
 - Model Tuning: Built-in capabilities for supervised fine-tuning and Reinforcement Learning from Human Feedback (RLHF) to adapt models to specific business domains.
- Generative AI Development Suite
 - Vertex AI Studio: A low-code environment for quickly prototyping prompts, testing model parameters, and refining AI responses.
 - Vertex AI Agent Builder: A framework for building enterprise-ready agents grounded in organisational data to perform complex tasks and workflows.
 - Grounding: Tools to connect AI models to "real-world" data sources (like Google Search or your own datasets) to reduce hallucinations and ensure accuracy.

- MLOps & Infrastructure
 - Vertex AI Pipelines: Automated workflows for orchestrating ML experiments and repetitive tasks.
 - Model Registry & Feature Store: Centralised repositories to manage model versions and share ML features across teams.
 - AutoML & Custom Training: Options to train models automatically with minimal code or use fully custom containers for frameworks like PyTorch and TensorFlow.

Key Benefits

- Unified Workflow: Eliminates tool sprawl by providing a single interface for data preparation, model training, tuning, and deployment.
- Enterprise-Grade Security: Built on Google Cloud's infrastructure, ensuring data privacy, compliance, and sovereignty (your data is never used to train foundation models without consent).
- Faster Time-to-Market: Pre-built APIs and low-code tools significantly reduce the time required to build and deploy production-ready AI agents.
- Seamless Data Integration: Native integration with BigQuery and Looker allows teams to build AI models directly on top of their existing data warehouses.

Target Use Cases

Vertex AI is utilised across various industries to solve complex business challenges:

- Customer Experience (Retail/Support): Deploying intelligent AI agents that handle complex customer inquiries, process returns, and provide personalised product recommendations.
- Content Operations (Marketing/Media): Automating the generation of high-quality marketing copy, social media posts, and localised product descriptions at scale.

- Document Intelligence (Finance/Legal): Extracting key data points from thousands of legal contracts or financial statements and generating concise summaries of complex reports.
- Predictive Analytics (Supply Chain/Manufacturing): Building custom models to forecast demand, optimise inventory levels, and perform predictive maintenance on industrial equipment.
- Enterprise Search: Creating internal search engines that allow employees to query vast corporate knowledge bases (PDFs, Wikis, Sheets) using natural language.

Contact Information

Netpremacy Public Sector Team

publicsector@netpremacy.com

Call: +44 1420 481450