AKADEMIA
FORMATION

PROGRAMME DE FORMATION

Claude sur Google Cloud Vertex AI — Intégration Cloud Google

DURÉE	FORMAT	PUBLIC	CERTIFICATION
2 jours	**Inter · Intra**	**Tout collaborateur**	**Incluse**
14 heures	Présentiel ou distanciel		

À PROPOS DE LA FORMATION

Claude sur Google Cloud Vertex AI — Intégration Cloud Google.

DURÉE	FORMAT	PÉDAGOGIE	CERTIFICATION
2 jours	**Inter · Intra**	**Active**	**Incluse**
14 heures	Présentiel ou distanciel	Petits groupes	Attestation délivrée

Objectifs pédagogiques

- Configurer l'accès aux modèles Claude sur Vertex AI (projet GCP, Model Garden, IAM/Service Accounts)
- Maîtriser l'API Vertex AI pour Claude : SDK Anthropic pour Vertex, API REST rawPredict et streaming SSE
- Comprendre les différences clés entre l'API Vertex AI et l'API directe Anthropic
- Choisir entre les 3 types d'endpoints (global, multi-region, regional) selon les contraintes de résidence
- Construire un pipeline RAG avec Vertex AI Search et les embeddings vectoriels Google Cloud
- Intégrer Claude avec BigQuery, Cloud Functions et Dataflow pour des workflows de données
- Optimiser les coûts : endpoints, prompt caching, provisioned throughput et gestion des quotas GCP

PROGRAMME

Huit modules progressifs pour monter en compétences.

MODULE

01.

1H30H

Pourquoi Vertex AI : positionnement et avantages

CONTENU PÉDAGOGIQUE

- Trois voies d'accès à Claude : API directe Anthropic, Amazon Bedrock, Google Vertex AI
- Avantages Vertex AI : facturation GCP consolidée, conformité FedRAMP High/SOC, crédits GCP utilisables
- Arbre de décision : quand choisir Vertex AI vs API directe vs Bedrock selon les contraintes projet
- Modèles disponibles : Opus 4.6, Sonnet 4.6, Haiku 4.5 avec Vertex Model IDs et dates de retrait
- Fenêtres de contexte : 1M tokens (Opus/Sonnet 4.6) vs 200K tokens (autres modèles), limite payload 30 Mo
- Exercice : Comparer les 3 plateformes sur 3 cas d'usage entreprise et choisir la plus adaptée

MODULE

02.

2H

Setup et première requête Vertex AI

CONTENU PÉDAGOGIQUE

- Prérequis : projet GCP avec facturation activée, gcloud CLI installé, Vertex AI API activée
- Accès aux modèles Claude via le Vertex AI Model Garden : demande d'accès et validation
- Authentification : gcloud auth application-default login, chaîne de credentials google-auth-library
- Service Accounts : création, rôles IAM Vertex AI, clé JSON pour les environnements serveur
- Installation des SDKs : pip install anthropic[vertex] (Python), npm install @anthropic-ai/vertex-sdk (TS)
- Première requête : AnthropicVertex avec project_id et region, format de réponse
- Différences clés avec l'API directe : model dans l'URL (pas le body), anthropic_version vertex-2023-10-16 dans le body
- Atelier pratique : Configurer l'environnement GCP, activer Vertex AI et exécuter ses premières requêtes via SDK et curl

MODULE

03.

1H45H

Les 3 types d'endpoints Vertex AI

CONTENU PÉDAGOGIQUE

- Endpoint global (recommandé) : region="global", routage dynamique, pas de surcharge, pay-as-you-go uniquement
- Endpoint multi-region : region="us" ou "eu", résidence données continentale, surcharge +10%
- Endpoint regional : us-east1, europe-west1, etc., routage garanti, +10%, provisioned throughput disponible
- Comparaison avec Bedrock : 3 types d'endpoints Vertex AI vs 2 pour Bedrock (global + regional)
- Résidence des données : choisir le bon endpoint selon RGPD, FedRAMP High, contraintes sectorielles
- Provisioned throughput : quand et comment réserver de la capacité dédiée sur endpoints régionaux
- Démonstration : Comparer la latence entre endpoints global, us et europe-west1 sur un même prompt

MODULE

04.

1H45H

API avancée : streaming, tool use et extended thinking

CONTENU PÉDAGOGIQUE

- Streaming SSE sur Vertex AI : Server-Sent Events, implémentation Python et TypeScript
- Appel REST brut avec curl : endpoint streamRawPredict, Bearer token via gcloud auth print-access-token
- Tool Use sur Vertex AI : schémas d'outils, tool_choice, boucle agentique multi-tools
- Extended Thinking : activation, budget_tokens, mode adaptive (Opus 4.6/ Sonnet 4.6), interleaved thinking
- Vision et PDF : analyse d'images et documents, citations pour l'extraction structurée
- Prompt caching sur Vertex AI : même mécanisme que l'API directe, cache_control, TTL, économies -90%
- Atelier pratique : Développer un assistant avec tool use, streaming et extended thinking sur Vertex AI

MODULE

05.

1H53H

RAG avec Vertex AI Search et services Google Cloud

CONTENU PÉDAGOGIQUE

- Architecture RAG sur GCP : Vertex AI Search comme alternative managée aux Knowledge Bases Bedrock
- Ingestion de documents : Cloud Storage (GCS), connecteurs natifs, preprocessing avec Dataflow
- Embeddings vectoriels : modèles Google (text-embedding) et intégration avec des embeddings tiers
- Vertex AI Vector Search : configuration d'index, requêtes de similarité, scaling automatique
- Pipeline RAG complet : ingestion → chunking → embedding → retrieval → augmentation → génération Claude
- Comparaison avec le RAG natif Anthropic : search_result_block et citations vs pipeline Vertex AI Search
- Exercice : Déployer un pipeline RAG avec Cloud Storage, Vertex AI Search et Claude comme générateur

MODULE

06.

1H37H

Intégration BigQuery, Cloud Functions et Dataflow

CONTENU PÉDAGOGIQUE

- Cloud Functions + Claude : serverless event-driven, déclenchement par Pub/Sub, HTTP ou Cloud Storage
- BigQuery + Claude : analyse de données à grande échelle, enrichissement de tables, classification de texte
- Dataflow + Claude : pipelines de streaming et batch pour le traitement de données en temps réel
- Pattern complet : événement GCS → Cloud Function → Claude via Vertex AI → résultat dans BigQuery
- Pub/Sub pour l'orchestration : découplage des composants, files d'attente et retry automatique
- Atelier pratique : Construire un pipeline Cloud Function → Claude → BigQuery qui analyse et classe des documents

MODULE

07.

1H38H

Optimisation des coûts et quotas sur Vertex AI

CONTENU PÉDAGOGIQUE

- Tarification Vertex AI : identique à l'API directe sur endpoint global, +10% sur multi-region et regional
- Prompt caching sur Vertex AI : écriture 5 min (1.25x) et 1h (2x), lecture à 0.1x = -90%
- Batch processing : pay-as-you-go standard et provisioned throughput (pas de discount -50% natif comme Bedrock)
- Stratégie de choix de modèle : Haiku 4.5 pour le volume, Sonnet 4.6 pour l'équilibre, Opus 4.6 pour la qualité
- Gestion des quotas GCP : requêtes par minute, tokens par minute, monitoring et alertes
- Crédits GCP et engagements : utiliser des crédits existants pour réduire le coût total
- Exercice : Calculer le TCO d'un workflow et optimiser avec caching + endpoint adapté + modèle approprié

MODULE

08.

1H52H

Architecture de production et bonnes pratiques GCP

CONTENU PÉDAGOGIQUE

- Architecture de référence : API Gateway → Cloud Functions → Vertex AI, BigQuery pour l'analytics, GCS pour les documents
- Request-response logging : journalisation prompts et complétions, rétention 30 jours recommandée
- Cloud Monitoring : métriques d'invocation, latence, taux d'erreur, alertes de coûts
- Sécurité : IAM least privilege, Service Accounts dédiés, VPC Service Controls, chiffrement natif GCP
- Haute disponibilité : endpoint global pour la résilience, fallback multi-region
- Fonctionnalités exclusives API directe non disponibles sur Vertex : web search, web fetch, code execution
- SDKs disponibles : Python, TypeScript, Go, Java, C#, PHP, Ruby — exemples comparés
- Atelier final : Concevoir et documenter une architecture Vertex AI complète avec RAG, intégrations et monitoring Cloud

PASSONS À L'ACTION

Construisons ensemble votre session sur-mesure.

Dites-nous vos contraintes (format, lieu, dates, nombre de participants) et recevez une proposition personnalisée sous 24 heures ouvrées.

Akademia Formation

SERVICE ADMINISTRATION DES VENTES

adv@akademiaformation.com

www.akademiaformation.com

Devis personnalisé

RÉPONSE SOUS 24 H OUVRÉES

Format inter · intra

Présentiel ou distanciel

FIN DU PROGRAMME