Cours 8

# Claude avec Google Cloud Vertex AI

Anthropic Academy — Resume detaille

Yoann Boulch | 01/04/2026

## 1. Vue d'ensemble de Vertex AI

■ Vertex AI de Google Cloud intègre Claude avec les services GCP

Vertex AI est la plateforme de machine learning de Google Cloud. L'intégration de Claude sur Vertex AI permet d'utiliser Claude avec les outils et services GCP existants.

**Avantages:**

- Intégration avec BigQuery, Cloud Storage, etc.
- Gestion d'entreprise des clés API
- Contrôle d'accès IAM granulaire
- Même capacités que l'API Anthropic

## 2. Configuration initiale

■ Configurez votre projet GCP et authentifiez-vous

```
# Configuration du client AnthropicVertex from anthropic import AnthropicVertex client =
AnthropicVertex( project_id="mon-projet-gcp", region="us-west1" ) message = client.messages.create(
model="claude-3-5-sonnet@20250514", max_tokens=1024, messages=[{ "role": "user", "content":
"Bonjour!" }] )
```

## 3. Conversations multi-tour

■ Gérez le contexte avec des messages historiques

Les conversations multi-tour conservent l'historique des messages pour maintenir le contexte.

```
messages = [ {"role": "user", "content": "Qui était Napoleon?"}, {"role": "assistant", "content":
"..."}, {"role": "user", "content": "Quand a-t-il mort?"} ] response = client.messages.create(
model="claude-3-5-sonnet@20250514", max_tokens=1024, system="Tu es un historien expert.",
messages=messages )
```

# 4. Prompts systèmes

■ Définissez le comportement avec system prompts

Le paramètre system définit les instructions de base pour le modèle.

### Exemples de system prompts:

- Rôles: 'Tu es un assistant technique'
- Contraintes: 'Réponds en moins de 100 mots'
- Formats: 'Utilise Markdown pour la mise en forme'
- Langue: 'Réponds toujours en français'

# 5. Température et contrôle des réponses

■ Contrôlez la créativité avec temperature

| Paramètre | Plage | Effet |
|---|---|---|
| temperature | 0.0 - 1.0 | 0=déterministe, 1=créatif |
| top_p | 0.0 - 1.0 | Contrôle la diversité |
| top_k | 1+ | Nombre de tokens à considérer |
| max_tokens | 1+ | Longueur maximale réponse |

# 6. Streaming de réponses

■ Recevez les réponses en temps réel avec le streaming

```
with client.messages.stream( model="claude-3-5-sonnet@20250514", max_tokens=1024, messages=[{"role": "user", "content": "Raconte une histoire"}] ) as stream: for text in stream.text_stream: print(text, end="", flush=True)
```

# 7. Sortie structurée (JSON mode)

■ Forcez la sortie au format JSON

```
response = client.messages.create( model="claude-3-5-sonnet@20250514", max_tokens=1024,
messages=[{"role": "user", "content": "Extrais les données"}], response_format={ "type":
"json_schema", "json_schema": { "name": "Result", "schema": { "type": "object", "properties": {
"titre": {"type": "string"}, "score": {"type": "number"} } } } } )
```

# 8. Caching et optimisation

■ Optimisez les performances avec le caching

Vertex AI supporte le caching des prompts pour réduire la latence et les coûts.

### Types de caching:

- Caching de contexte: réutiliser les prompts systèmes
- Caching d'embeddings: résultats de recherche mis en cache
- Caching de sessions: historique des conversations

# 9. Vision et traitement de PDF

■ Analysez des images et documents PDF

Claude peut analyser des images et extraire du texte de documents PDF.

### Types de contenu visuel:

- Images PNG, JPEG, GIF, WebP
- Documents PDF (jusqu'à 20 pages)
- Reconnaissance optique de caractères (OCR)
- Analyse de tableaux et diagrammes

# 10. RAG (Retrieval-Augmented Generation)

■ Augmentez Claude avec des connaissances externes

RAG améliore les réponses en intégrant des documents pertinents dans le contexte.

### Composants de RAG:

- Chunking: division des documents en sections
- Embeddings: conversion en vecteurs
- Recherche: retrouver les documents pertinents
- Augmentation: injecter dans le contexte

# 11. Évaluation de prompts sur Vertex

■ Testez et améliorez vos prompts systématiquement

Vertex AI fournit des outils pour évaluer la qualité des prompts avec des datasets de test.

### Approches d'évaluation:

- Datasets de test: paires entrée/sortie attendue
- Métriques: précision, pertinence, complétude
- Grading: automatisé ou manuel
- A/B testing: comparer différents prompts

# 12. Outils Claude sur Vertex

■ Utilisez les tools et la vision avec Vertex

Vertex AI supporte le Tool Use de Claude, permettant au modèle d'appeler des outils et fonctions.

### Intégrations disponibles:

- BigQuery: requêtes SQL directes
- Cloud Storage: accès aux fichiers
- Vertex AI Search: recherche documentaire
- APIs Google Cloud: tous les services GCP

# 13. Comparaison avec API Anthropic

■ Vertex AI et API Anthropic: similarités et différences

| Aspect | API Anthropic | Vertex AI |
|---|---|---|
| Authentification | Clé API | IAM GCP |
| Pricing | Par-utilisation | Par-utilisation |
| Intégration GCP | Manuelle | Native |
| Support entreprise | Disponible | Inclus |
| Latence | Standard | Optimisée GCP |