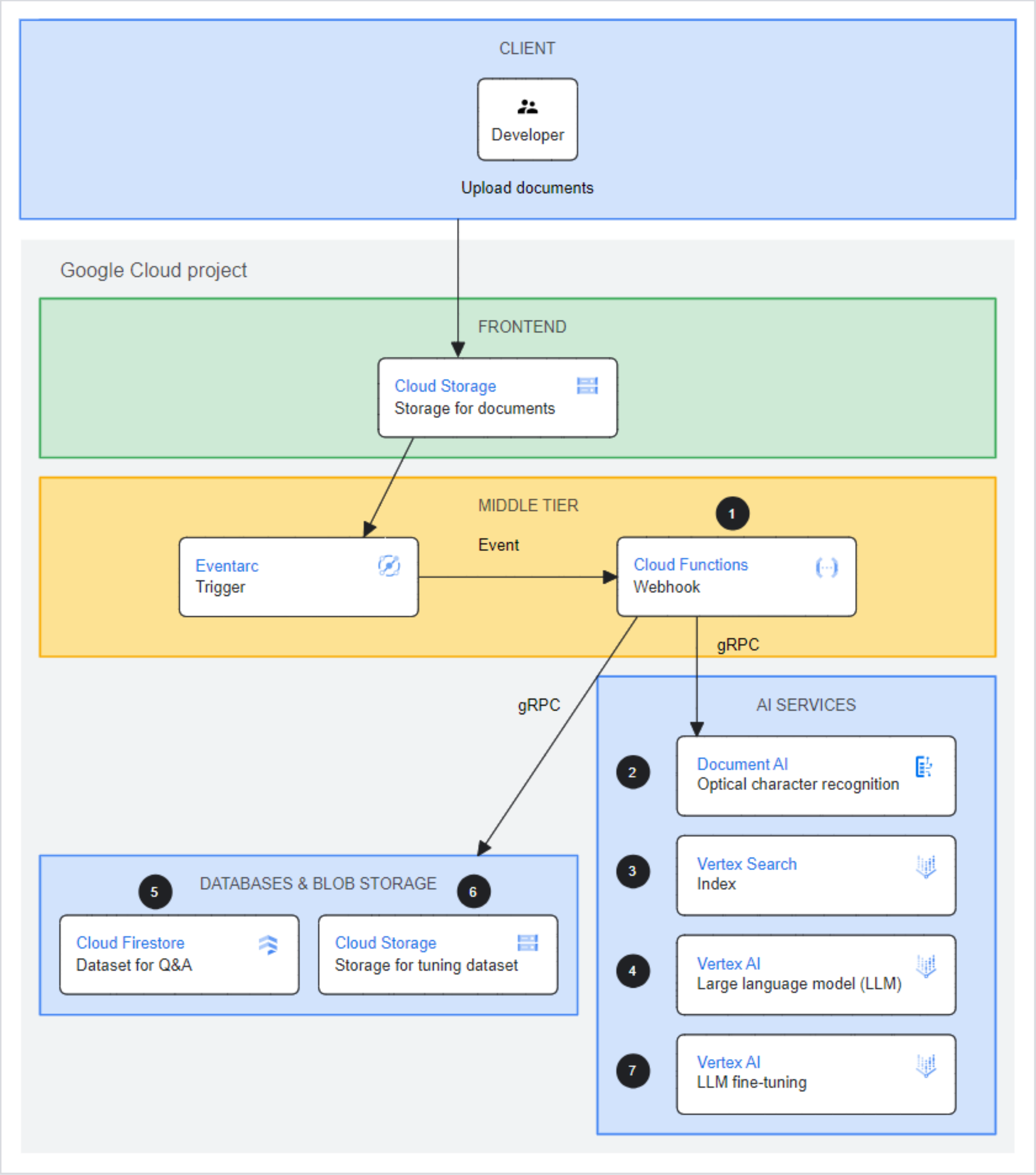
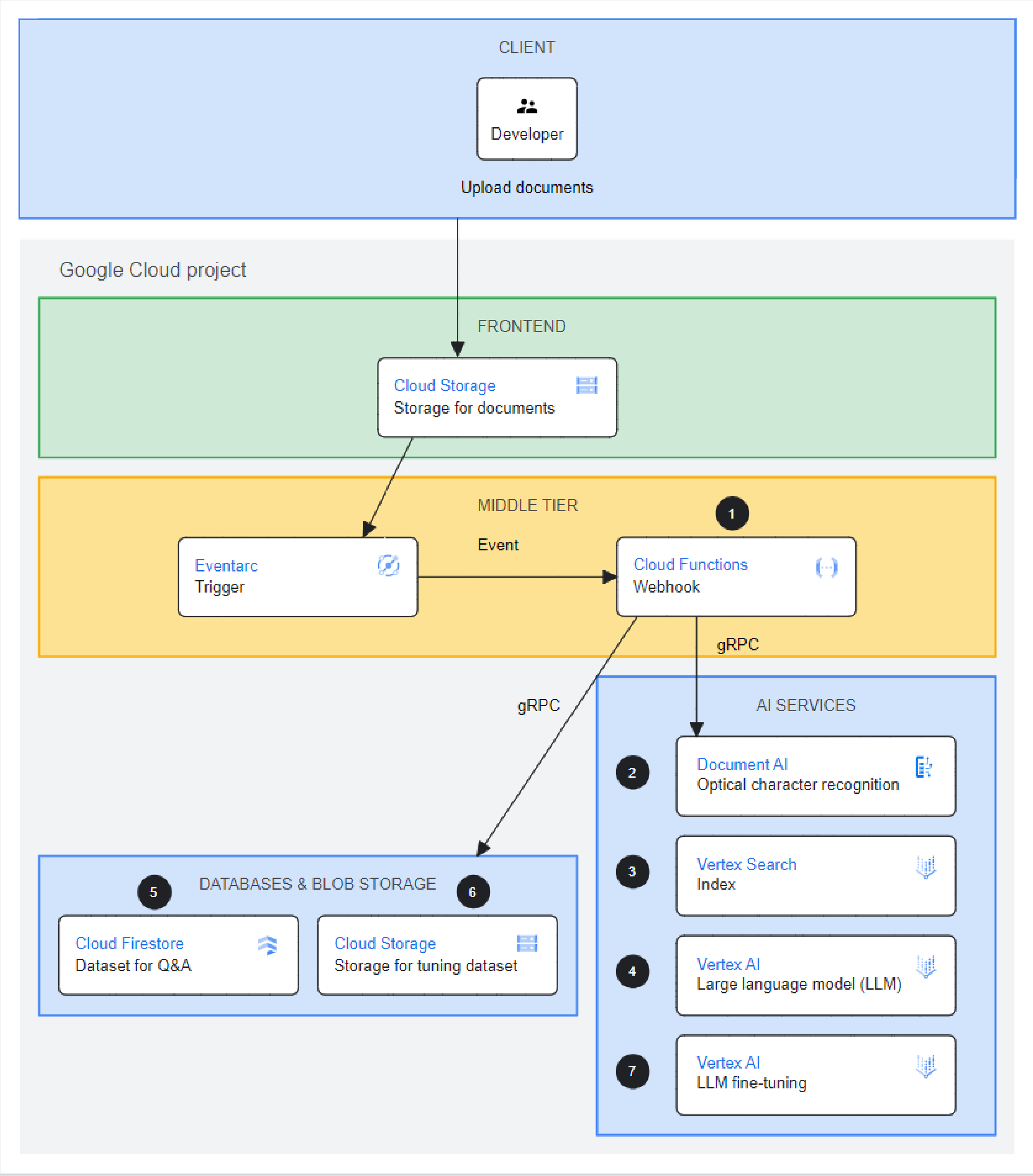
Generative AI Knowledge Base



Generative AI Knowledge

1. When the document is uploaded, it triggers a Cloud Run function. This function runs the Extractive Question-Answering process.
2. The function uses Document AI OCR to extract all text from the document.
3. The function indexes the document into Vector Search. The Vector Search index provides context for the LLM to extract question-and-answer pairs based only on content that's extracted directly from the uploaded documents.
4. The function uses Vertex AI to extract and generate questions and answers from the document.
5. The function stores the extracted question-and-answer pairs in Firestore.
6. A JSONL fine tuning dataset is generated from the Firestore database and stored in Cloud Storage.
7. After manually validating that you are satisfied with the dataset, you can launch a fine tuning job on Vertex AI.
8. When the tuning job is complete, the tuned model is deployed to an endpoint. After it's deployed to an endpoint, you can submit queries to the tuned model in a Colab notebook, and compare it with the foundation model.

source:
https://cloud.google.com/architecture/ai-ml/generative-ai-knowledge-base


CLIENT
Developer
Upload documents
Google Cloud project
FRONTEND
Cloud Storage
Storage for documents
MIDDLE TIER
1
Eventarc
Trigger
Event
Cloud Functions
Webhook
gRPC
gRPC
AI SERVICES
2
Document AI
Optical character recognition
3
Vertex Search
Index
4
Vertex AI
Large language model (LLM)
7
Vertex AI
LLM fine-tuning
DATABASES & BLOB STORAGE
5
6
Cloud Firestore
Dataset for Q&A
Cloud Storage
Storage for tuning dataset