Connecting Gen AI models to the real world:

RAG

May 8, 2025

Google Cloud

Google Cloud

Let’s get ready Startups!

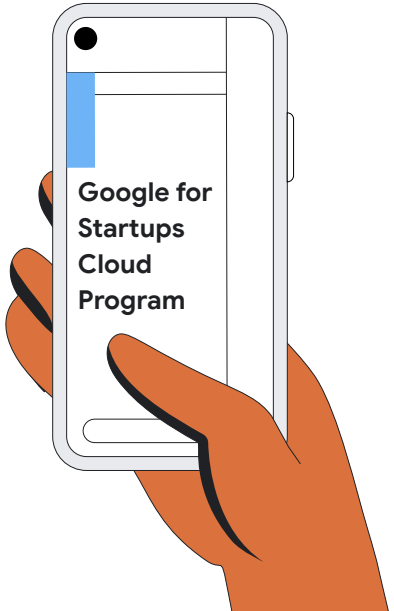
If you haven't already, create a Qwiklabs account with your corporate email address by going to **explore.qwiklabs.com**.

Or share your email with us by filling in the form in chat!

Before we start

Learn more and apply at

cloud.google.com/startup

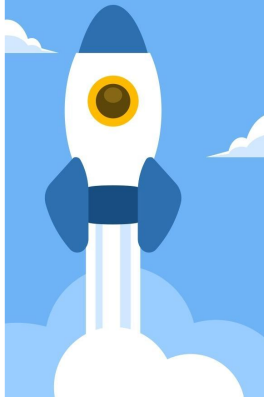


Google Cloud

Get in touch with Team

A dedicated point of contact for questions and support

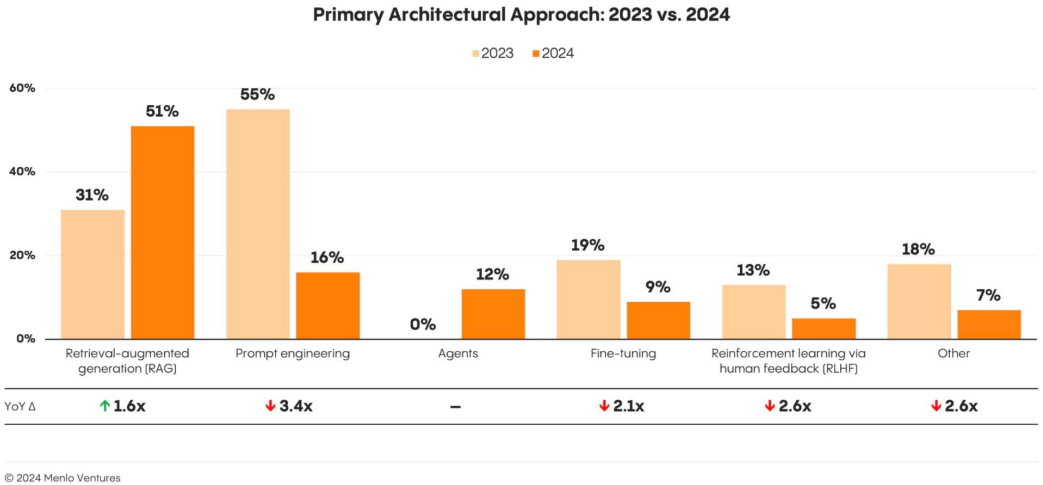
Fill the form



Contents

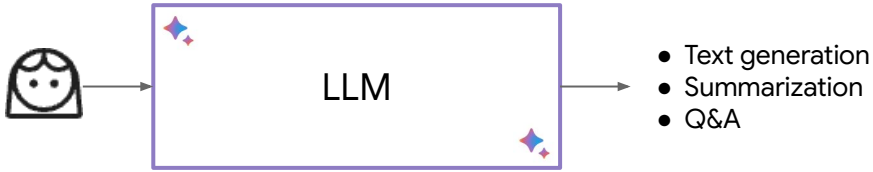
What is RAG?

How the industry is using LLMs?



Typical LLM usage

LLMs are a phenomenal for knowledge generation and reasoning. They are pre-trained on large amounts of **publicly available data**.



- Text generation
- Summarization
- Q&A

But.... The Grounding Problem (aka Hallucinations)

LLMs can only understand the information

- That they were **trained on**
- That they are explicitly given in the prompt

Since they're trying to be helpful, they will often assume that the premise of a prompt is true.

LLMs don't have the capability to ask for more information and need some outside system.

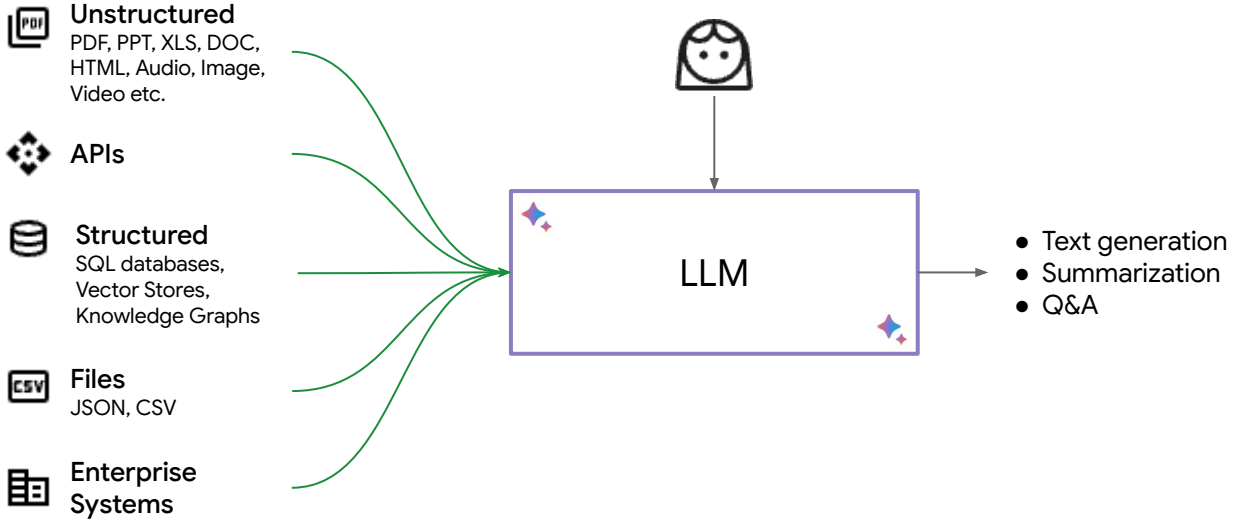
Prompt

What is langchain?

Response

LangChain is a blockchain-based platform that allows users to create, share, and learn languages. It is a decentralized platform that is not controlled by any single entity. This means that users have complete control over their data and content. LangChain is also a secure platform, as all data is encrypted and stored on the blockchain. This makes it very difficult for hackers to access or steal data.

How do we best augment LLMs with our own private data?



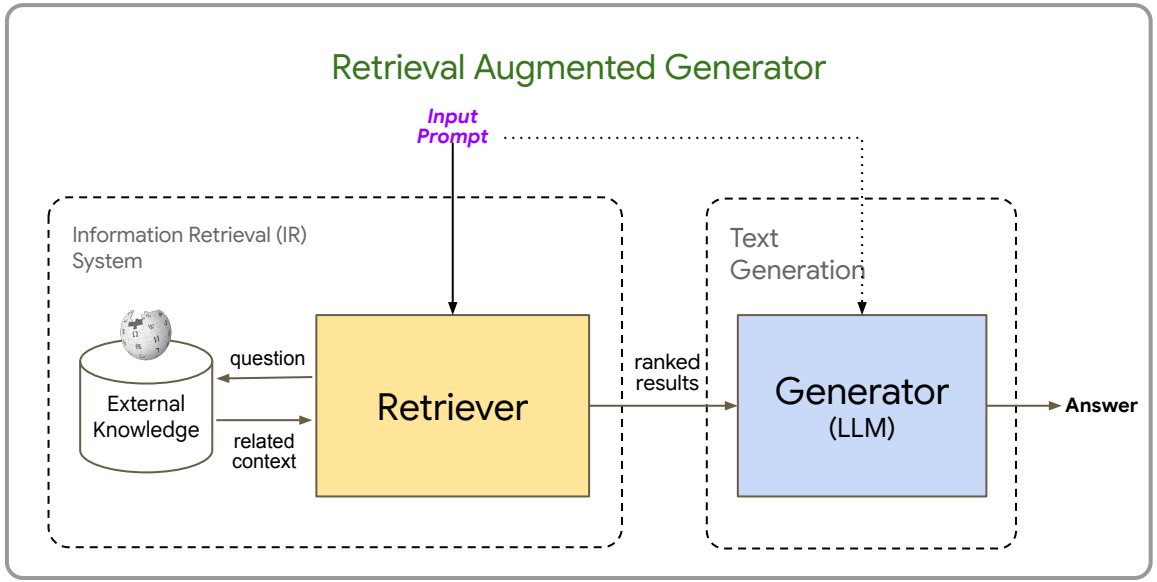
Retrieval Augmented Generation (RAG)

The Problem:

- LLMs **do not know** your business's proprietary or domain specific data
- LLMs do not have **real-time** information
- LLMs find it **challenging** to provide **accurate citations** from their parametric knowledge

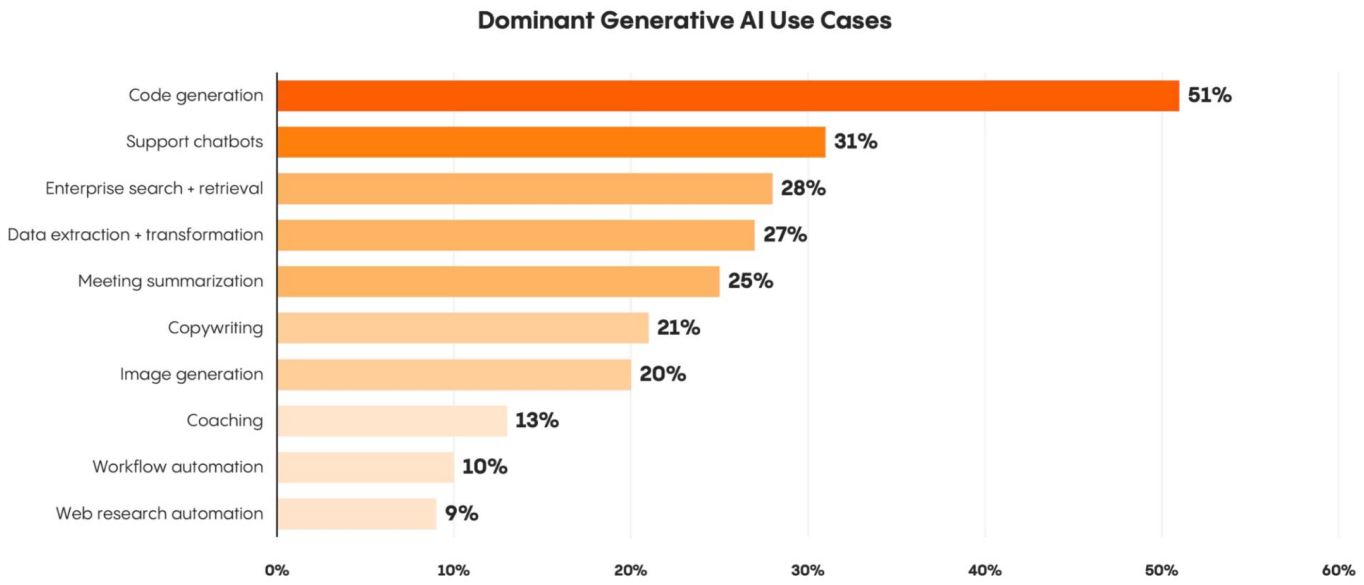
The Solution:

Feed the LLM *relevant* context in real-time, by using an information retrieval system



How is RAG used in the industry?

Inside the Enterprise: Ranking the Most Valuable Use Cases



Example customer: Dow Jones

- Use case: Semantic search experience for analysts over billions of articles
  - *E.g. query: "Recent advancements in clean energy technology"*
- Building in-house using Vertex AI Embeddings and Vector Search (no ML expertise needed)
- Interested in maintaining control over tech stack and the ability to understand and granularly tune search relevance
- Can reuse Vector Search platform to support additional use cases (e.g. consumer semantic search) and repurpose embeddings for other use cases (e.g. recommendations)

DOW JONES

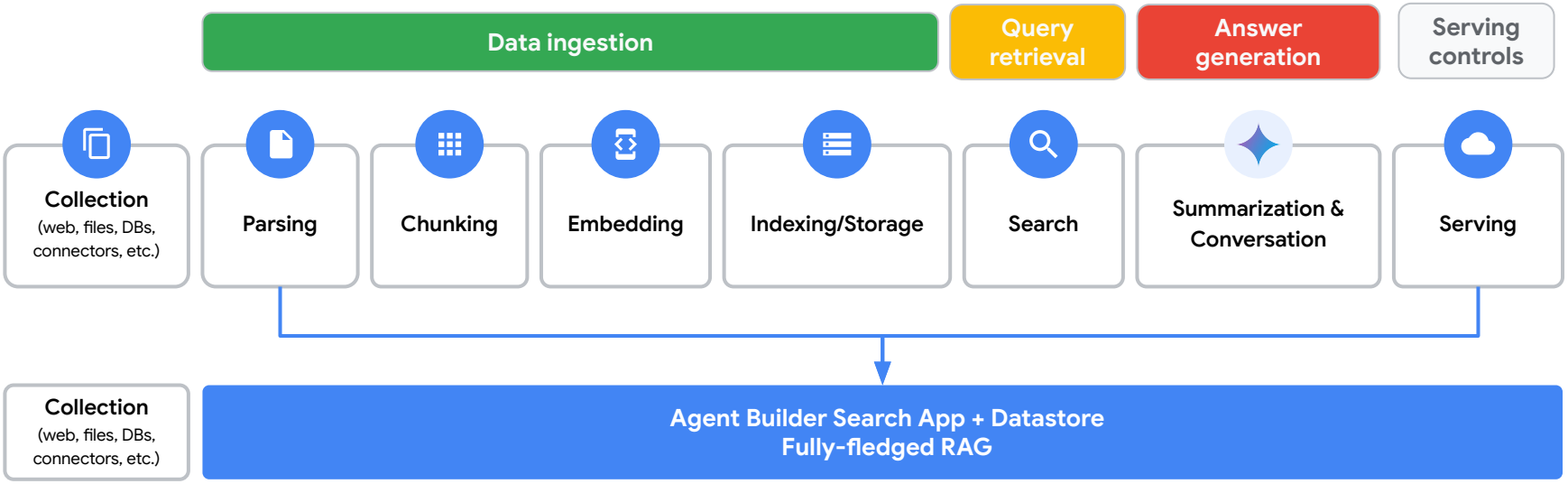
News Corp

What Google Products can help?

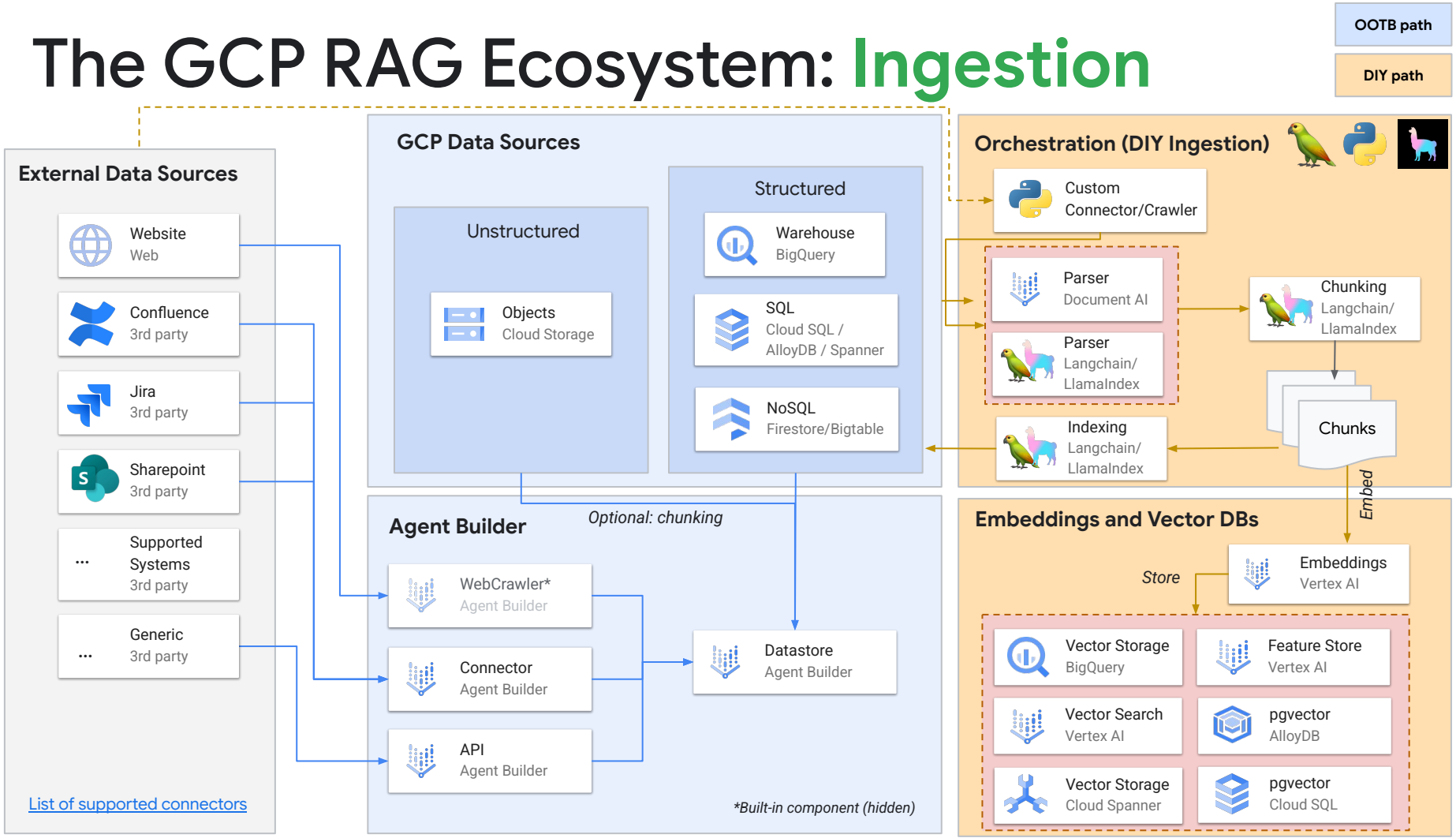
The GCP RAG Ecosystem: All-in-one

OOTB path

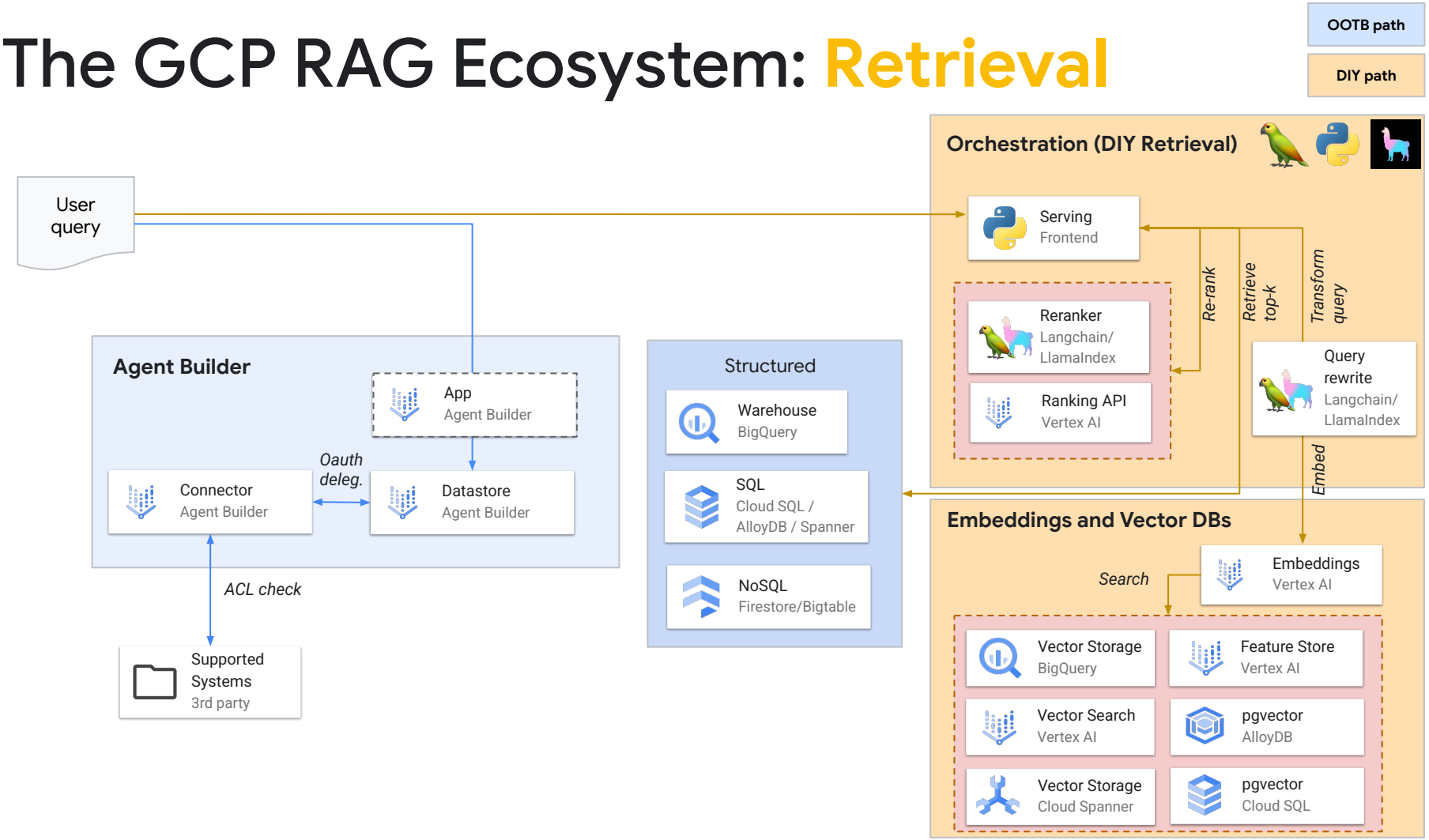
Agent Builder Search Agent



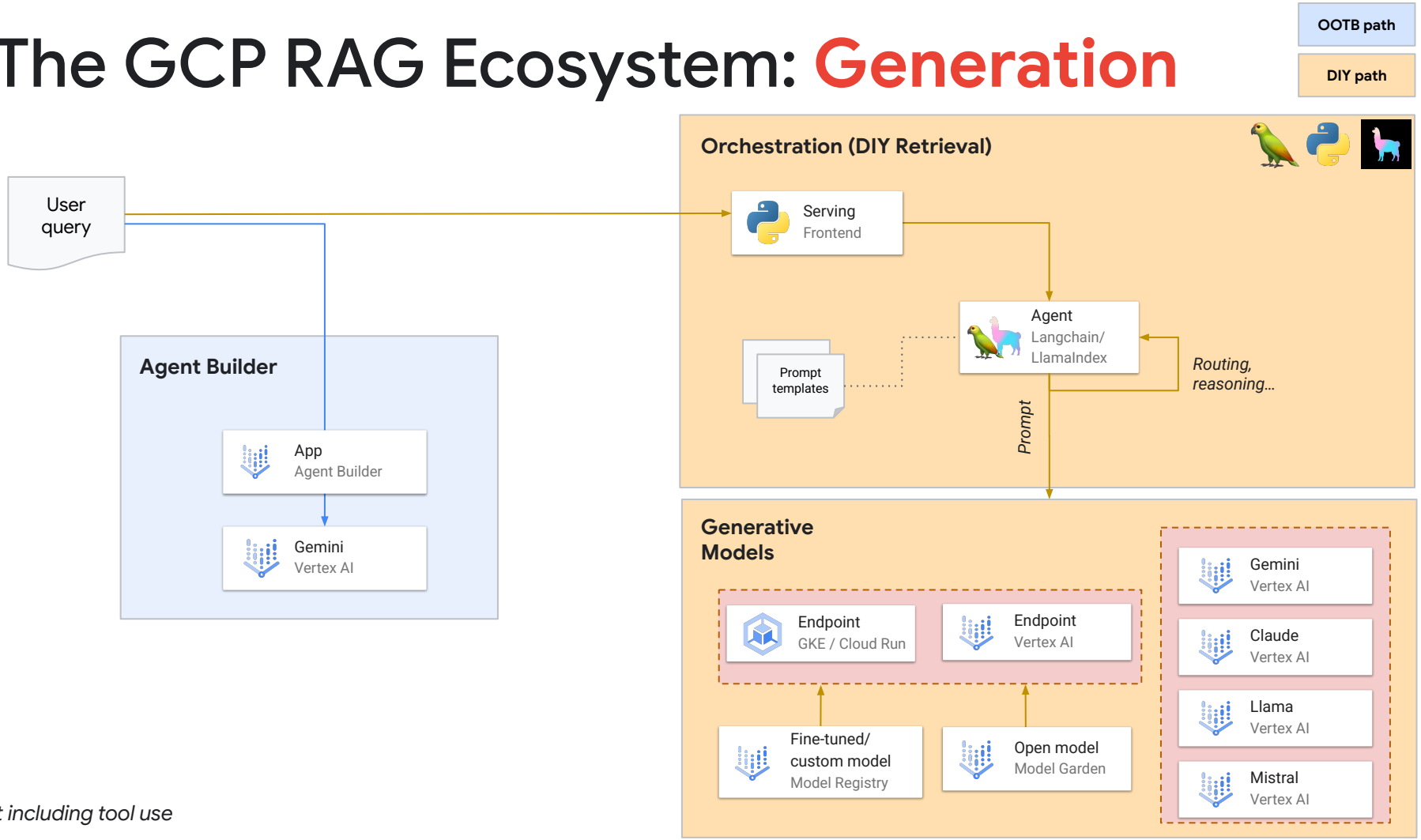
The GCP RAG Ecosystem: Ingestion



The GCP RAG Ecosystem: Retrieval



The GCP RAG Ecosystem: Generation



Not including tool use

Deep Dive

How can we use LLMs to answer business questions (Q&A)?

Historical approaches

- Pre-LLM: Nonparametric Q&A
- Methods: Lookup, matching
- Limitation: No synthesis
- Benefits: Easy, debuggable

LLMs

- LLMs: Parametric knowledge
- Answers: From parameters
- Updating: Difficult
- **Retraining**: Avoided often

Problems of language models

- Hallucination
- Attribution
- Staleness
- Revisions
- Customization

RAG is a semiparametric approach

- RAG: Semi-parametric
- LLM adapts DB knowledge
- Search context enables attribution
- Reduces staleness, hallucinations

LLM Fundamentals — What is a token?

- LLMs: Process tokens
- Tokens: Words, subwords
- Abilities: Token-defined
- Limits: Increasingly larger

Tokens	Characters
14	67

What is the capital of France?

Paris is the capital city of France

Text Token IDs

https://platform.openai.com/tokenizer

Frozen RAG

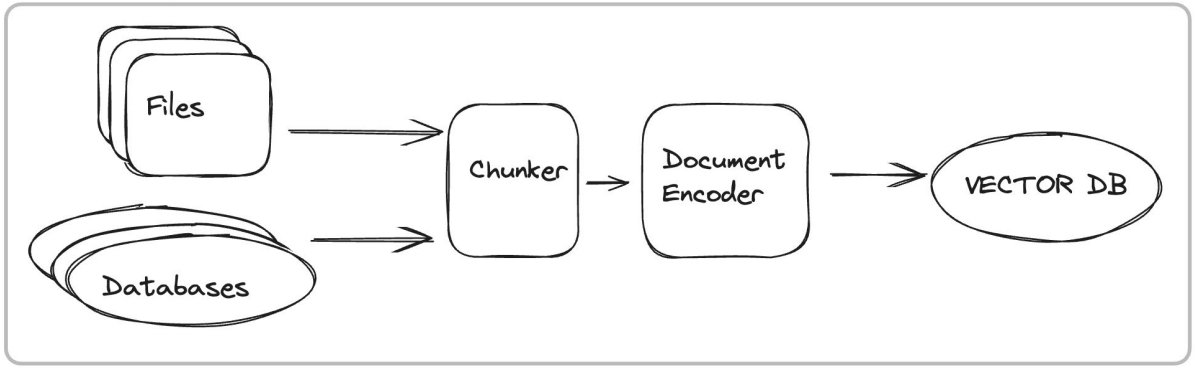
- Popular RAG: Not original
- No fine-tuning: Frozen weights
- Semantic search: Chunked data
- Uses off-the-shelf LLM

Chunking

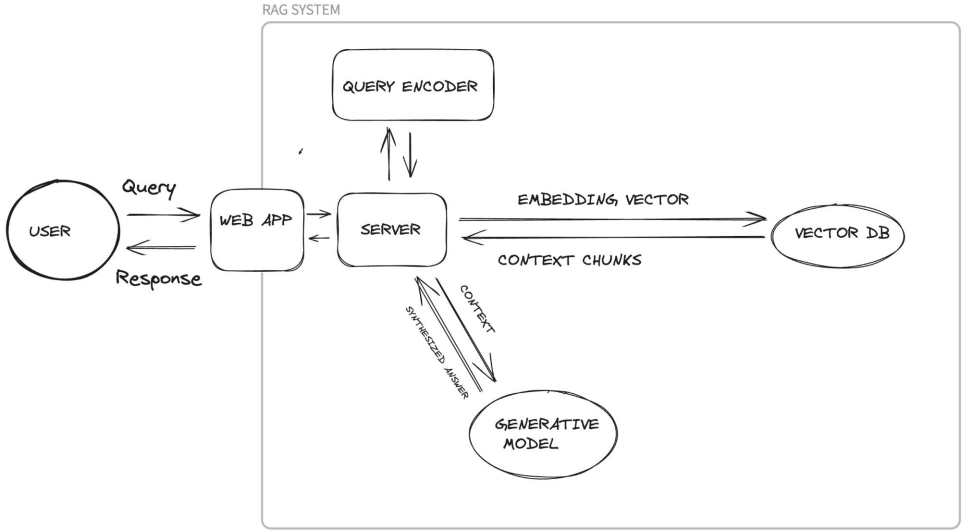
- Chunking: Small searchable pieces
- Methods: Length, separators, structure
- Chunks: Individually meaningful
- Size: Relates model limits

Offline Data Processing

Data Processing Batch Pipeline



Simplified Serving — request flow

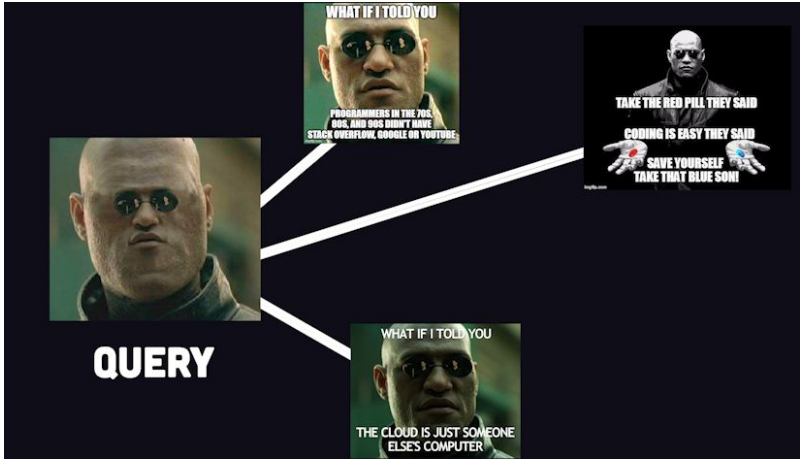


Embeddings

- Embeddings: Input to vectors
- Capture: Semantic similarity
- Limitation: Lossy, length issues
- Multimodal: Cross-modal search

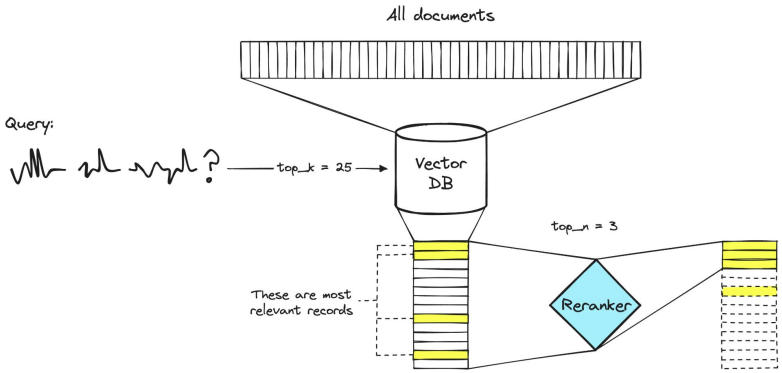
Vector Search

- Semantic search: Embeddings, vectors
- Small data: Exhaustive search
- Large data: ANN (fast, approximate)
- Vector DBs common (Vertex)



One-stage vs two-stage retrieval

- Basic: Vector DB chunks
- Issue: Independent embeddings
- Solution: Two-stage retrieval
- Example: Cloud reranking model

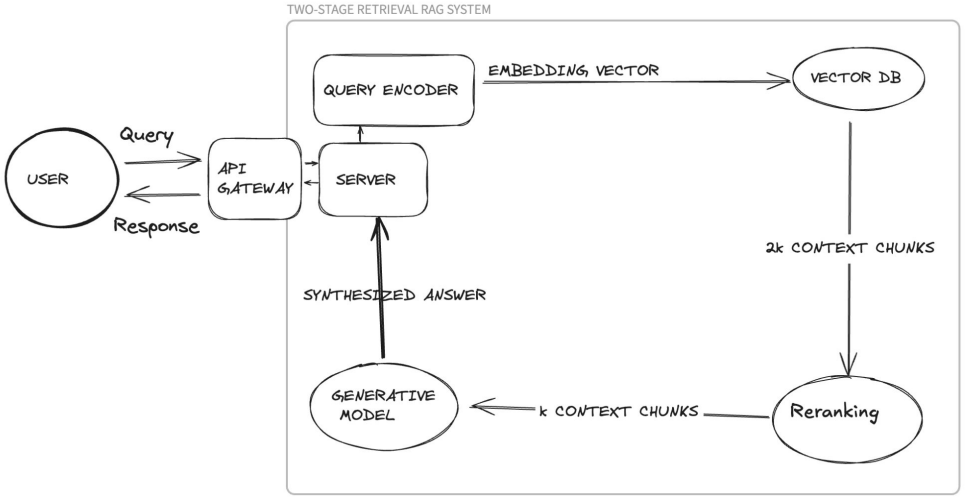


Two stage retrieval continued

Two stage retrieval can allow you to combine results from diverse **sources**

- Lexical + Semantic Search -> Hybrid Search
- Could use a public search engine + an internal search

Two stage retrieval diagram



Prototyping with large context

- Large context: Holds all data
- All-context: Enough, cheap prototype
- RAG: For larger/dynamic data
- Production RAG: Much costlier

Prototyping with large context

- Large space of design
- Feedback early
- Optimize

Potholes - things to watch out for

- Does you embedding model understand your domain?
- Are you retrieving the correct chunks for a given query?
- Is your reranking model working as you would want?
- Are your chunks meaningful?
- Do you have useless chunks, duplicate chunks?
- Is your model hallucinating or is the information provided wrong?
- Do you have any degenerate chunks?
- Do you have disembodied chunks?

Learn more about RAG

Great Podcast series on all facets of Search and RAG:
https://www.youtube.com/@howaiisbuilt

Amazing YouTube video from Stanford on the Research of RAG:
https://www.youtube.com/watch?v=mE7IDf2SmJg

Excellent blog post by Anthropic on Contextual Retrieval
https://www.anthropic.com/news/contextual-retrieval

Google Cloud
for Startups

Google Cloud for Startup Program

Providing resources to help early stage startups **build and scale**

Financial

Google Cloud credits (up to $350k) and other discounts to help startups build their products and early infrastructure

Business

Help with navigating Google resources for startups to build and grow their business

Technical

Educational resources and workshops led by Google Cloud Customer Engineers

Community

Access to Google Cloud experts and peers on Google Cloud Community and at local events

Apply Now

or reach at

cloudstartupsupport@google.com

Cloud Program

Meeting startups where they are in their journey

Founding & Bootstrapped
1-5 founders in ideation and prototyping stage

MVP & Fundraising
Small team building MVP, testing in the market, and iterating

Launched & Scaling
Product in market, increasing customer base, and planning for rapid growth post-funding

Start tier
$2,000 USD credits / 1 year

For startups **just getting started**

Pre-Funding

Ecosystem tier
$25,000 USD credits / 2 years

Private offer available through the Startup Success Manager

Accelerators, Incubators & Angel

Scale tier
Year 1: Up to $100,000 USD credits

Year 2: 20% of usage covered up to $100,000 USD

For institutionally funded startups ready to scale

Scale AI
Year 1: Up to $250,000 USD credits

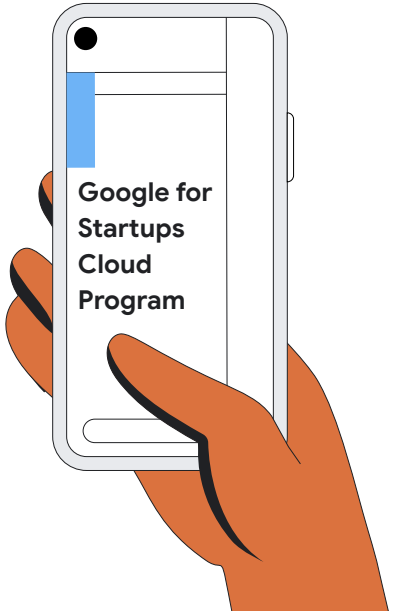
Year 2: 20% of usage covered up to $100,000 USD

For **AI-first startups** that are institutionally funded

Institutionally funded (up to recent Series A)

Learn more and apply at

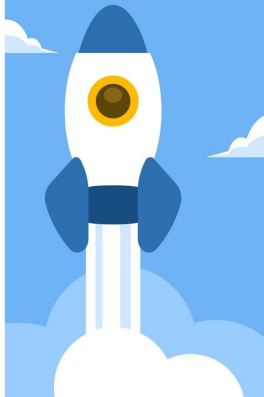
cloud.google.com/startup



Get in touch with Team

A dedicated point of contact for questions and support

Fill the form



Google Cloud

Google Cloud
for Startups

Hands-on Lab

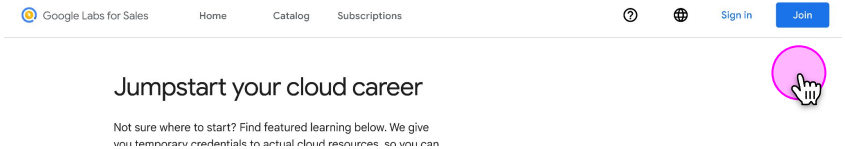
Google Cloud

Let’s get ready Startups!

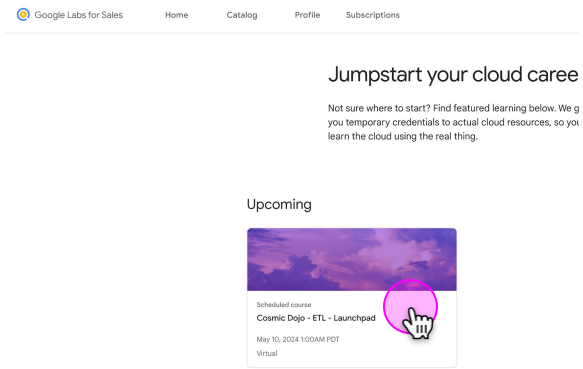
If you haven't already, create a Qwiklabs account with your corporate email address by going to **explore.qwiklabs.com**.

Or share your email with us by filling in the form in chat!

Step 1: Go to www.explore.qwiklabs.com and login/create an account with your **corporate email address**



Step 2:



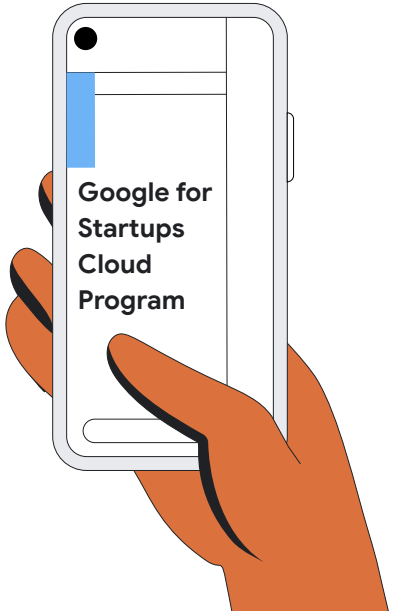
Hands-on lab

How to start the lab

- Visit explore.qwiklabs.com
- Log in using the account you provided when you registered to this classroom
- Click on the scheduled class in your home page (if you don't see any class, please raise your hand)
- Click on the lab
- Start the lab as shown on the right
- Follow lab instructions

Learn more and apply at

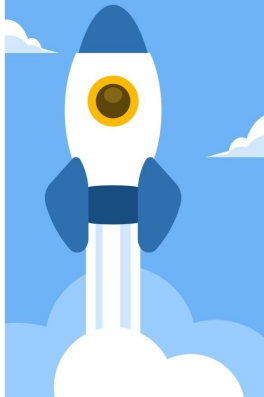
cloud.google.com/startup



Get in touch with Team

A dedicated point of contact for questions and support

Fill the form



Google Cloud

Thank you

Google Cloud