Google Cloud
Next '24

# Building generative AI apps on Google Cloud with LangChain

Proprietary

# Agenda

01 LangChain

02 Cloud Run

03 Customer story: L'Oréal

04 Demos

# LangChain

Proprietary

# Harrison Chase

Co-Founder
LangChain

Proprietary

# Large Language Model (LLM) applications...

## Are context-aware.

Language models need to be connected to sources of context

## Reason.

Rely on a language model to reason

## Building a LLM app can get complicated without a framework.

Proprietary

# LangChain's flexible abstractions and AI-first toolkit make it a top choice for developers when building with GenAI.

LangChain helps with...

1. Connecting LLMs to your company's private sources of data and APIs
2. Offering a complete set of interoperable and interchangeable building blocks
3. Customizability and control with a durable runtime baked in

LangChain is open source and powered by a community of 2K+ contributors.

Proprietary

# Turn your LLM app into an API server instantly with LangServe

Built with FastAPI, LangServe gives you an API, docs, and a playground for your LangChain apps.

Available open source and hosted (currently in beta and is hosted in GCP). Hosted LangServe:

- Allows you to quickly iterate and collaborate, with a built-in playground and docs site to get feedback faster
- Integrates with GitHub and takes care of spinning up a container to run your application
- Has observability built in with LangSmith
- Has native regression testing, showing you the exact branch and commit deployed at any given time
- Generates documentation for your LangServe application automatically

> Cloud Run is the easiest and fastest way to get your LangChain app running in production.”

Nuno Campos, Founding engineer, *LangChain*

Proprietary

# Cloud Run

# Steren Giannini

Group Product Manager, Google Cloud

## Google Cloud's serverless runtime

# Run applications
# fast and securely
# in a fully-managed environment.

# Cloud Run

## Experience

**Simple**
Demand as little as possible.

**Automated**
Takes care of a lot for you.

**Top satisfaction and usability scores**
High CSAT and task success.

**Developer productivity**
Idiomatic patterns, deployment velocity.

## Runtime

**Capable**
Run any code, any container.

**On-demand**
No pre-provisioning.

**Hyper-elastic**
Fast automatic scaling, including to zero.

**No infrastructure management**
No VM or cluster to upgrade.

## Pricing

Pay only when code is running, with a 100ms granularity.

- CPU
- Memory
- Requests (not always)

Perpetual monthly free tier

Committed Use Discounts

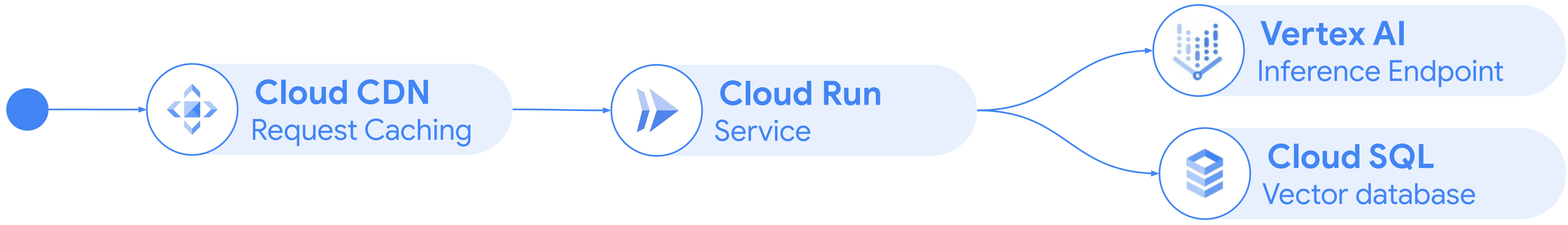
Pay per use.
No flat fee.

# Cloud Run services for GenAI apps

1

## Serving & orchestration

Request-response flow between the genAI app and its users. Runs LangChain to orchestrate retrieval-augmented generation (RAG), prepare prompts, call inference endpoint, format response.



## Services

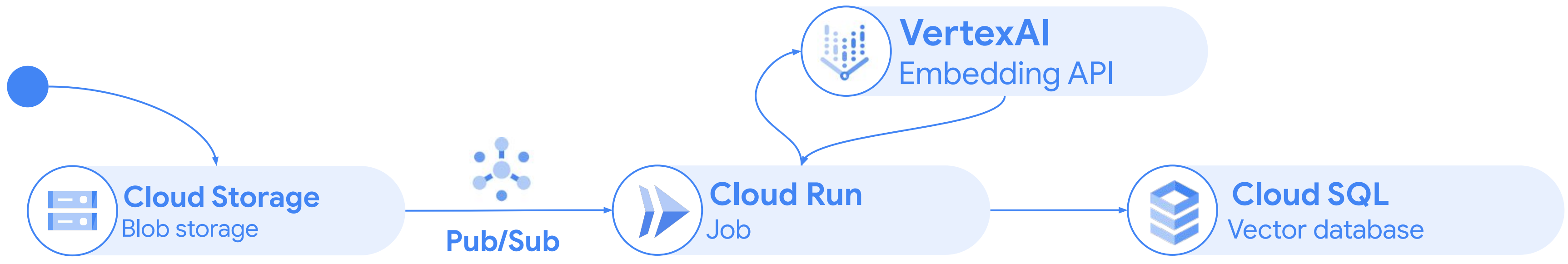
- Out of the box **HTTPS endpoint**
- **Automatic** and **rapid scaling** of instances
- **Revision history** and **traffic split**
- **Flexibility** to run popular GenAI app frameworks like LangChain, in any language (Python, JavaScript…)
- **Built-in integrations** with Vertex AI, Cloud SQL, AlloyDB, Cloud CDN, custom domains
- **Streaming** (HTTP Chunked transfer encoding, HTTP/2 or WebSockets)
- **Enterprise ready:** Direct VPC connectivity, granular security and networking controls

# Cloud Run jobs for GenAI apps

2

## Data ingestion

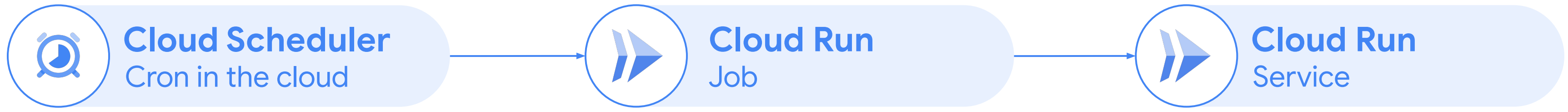
Prepare and process external data to load the vector database used for RAG.



3

## Quality evaluation

Evaluate the quality of responses that generated by serving.



## Jobs

- **On demand** execution of any code
- **Automatic** and **rapid scaling** of tasks
- Runs up to **10,000 tasks**
- Up to **24h per task**
- Execute manually, on a **schedule**, or **triggered**
- Execution **history**

# Customer story: L'Oréal

Proprietary

# Thomas Ménard

AI Center of Enablement Manager,
L'Oreal

Proprietary

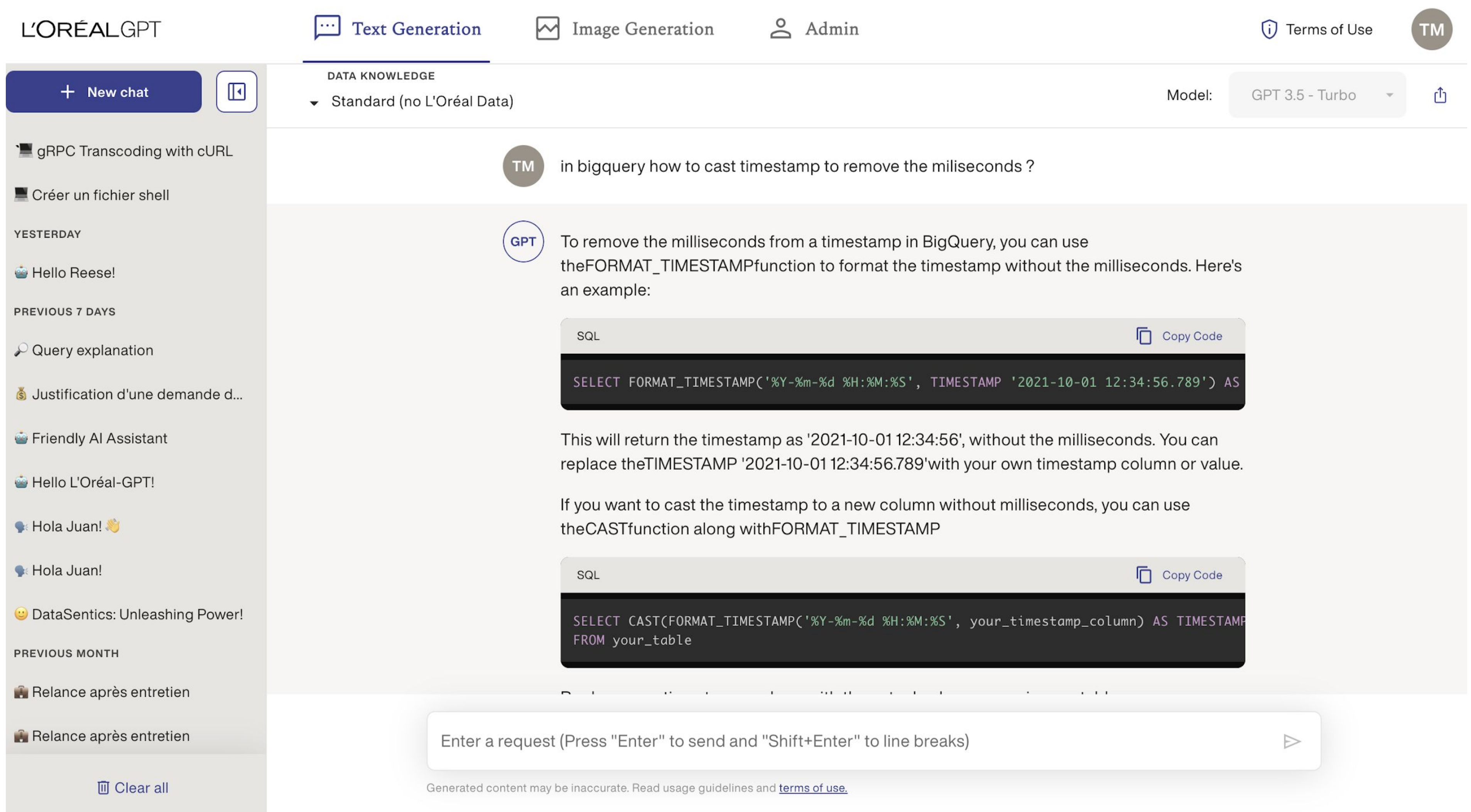
# From L'OréalGPT...

Provide a secured access to GenAI solutions for all L'Oréal members

Text/RAG & Image generation

More to come..

**Massive adoption**



DATA KNOWLEDGE
Standard (no L'Oréal Data)

Model: GPT 3.5 - Turbo

TM: in bigquery how to cast timestamp to remove the miliseconds ?

GPT: To remove the milliseconds from a timestamp in BigQuery, you can use theFORMAT_TIMESTAMPfunction to format the timestamp without the milliseconds. Here's an example:

SQL

```sql
SELECT FORMAT_TIMESTAMP('%Y-%m-%d %H:%M:%S', TIMESTAMP '2021-10-01 12:34:56.789') AS
```

This will return the timestamp as '2021-10-01 12:34:56', without the milliseconds. You can replace theTIMESTAMP '2021-10-01 12:34:56.789'with your own timestamp column or value.

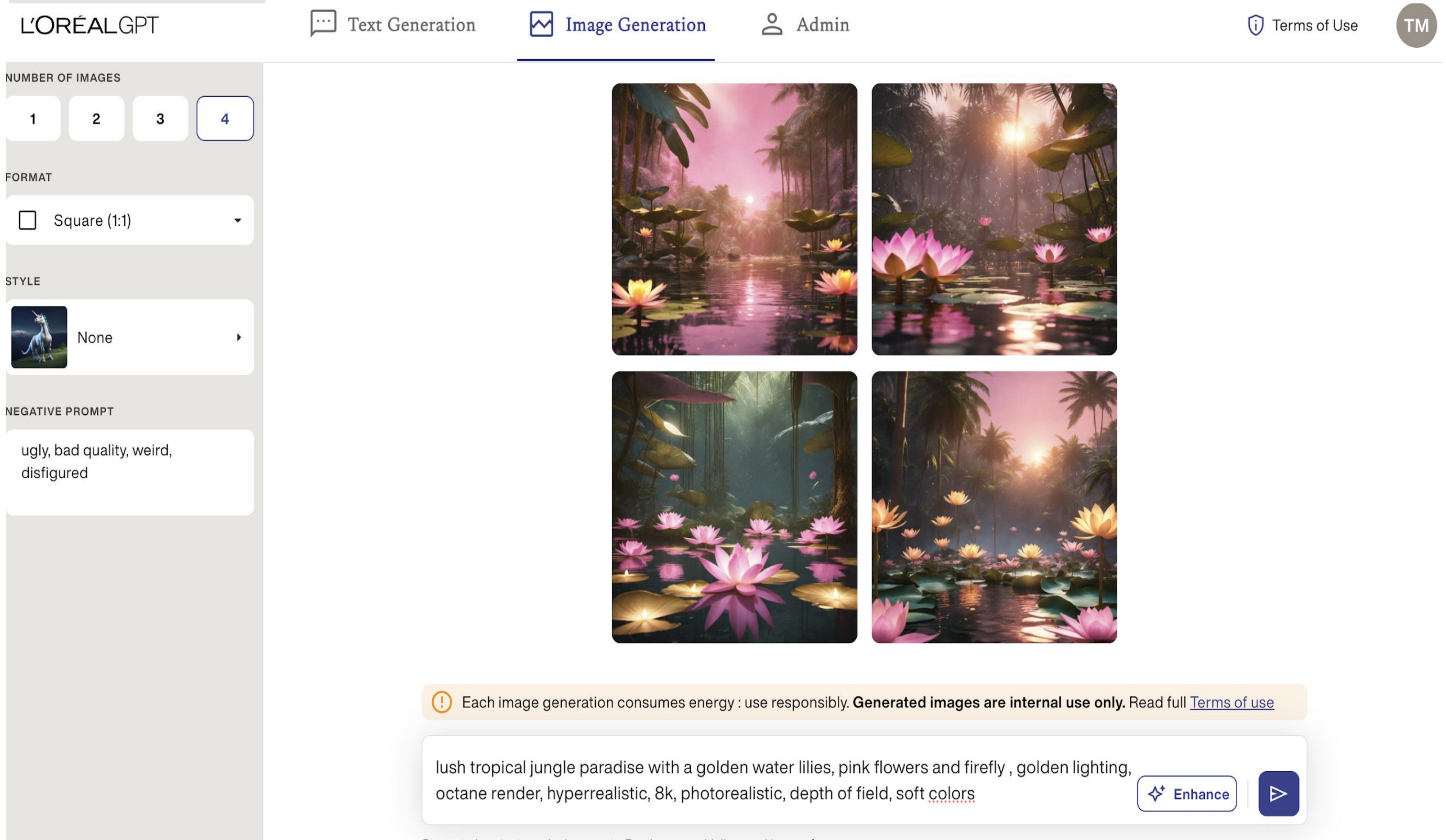
If you want to cast the timestamp to a new column without milliseconds, you can use theCASTfunction along withFORMAT_TIMESTAMP

SQL

```sql
SELECT CAST(FORMAT_TIMESTAMP('%Y-%m-%d %H:%M:%S', your_timestamp_column) AS TIMESTAMP
FROM your_table
```

Enter a request (Press "Enter" to send and "Shift+Enter" to line breaks)

Generated content may be inaccurate. Read usage guidelines and terms of use.



NUMBER OF IMAGES: 1 2 3 **4**

FORMAT: ☐ Square (1:1)

STYLE: None

NEGATIVE PROMPT: ugly, bad quality, weird, disfigured

Each image generation consumes energy : use responsibly. **Generated images are internal use only.** Read full Terms of use

lush tropical jungle paradise with a golden water lilies, pink flowers and firefly , golden lighting, octane render, hyperrealistic, 8k, photorealistic, depth of field, soft colors

# ...to GenAI as a Service

Set of **declarative Generative AI APIs** available in the L'OréalAPI portal, available to all and **enabling developers** to use GenAI technology in a fast, secure and customized way.

## Flexibility

## Robust Security and Validation

## Scalability and easy Deployment

## L'Oréal Ecosystem

## Support and Monitoring

## Time and Cost Saving

# Four main features of GenAI Services

## Prompt

Completion: Generation of a text based on a prompt

Free choice of LLMs: Gemini, GPT, Mistral...

## Chat

History management

Long term memory

Multimodal conversations

## Image

Generation of images

Free choice of text-to-image models: Stable Diffusion XL, DallE-3, ImaGen, ..

Image editing settings

Video & audio coming soon!

## Retrieval Augmented Generation (RAG)

Ingestion of your own non-structured data (PDF, Words...)

Retrieve your data & chat on it

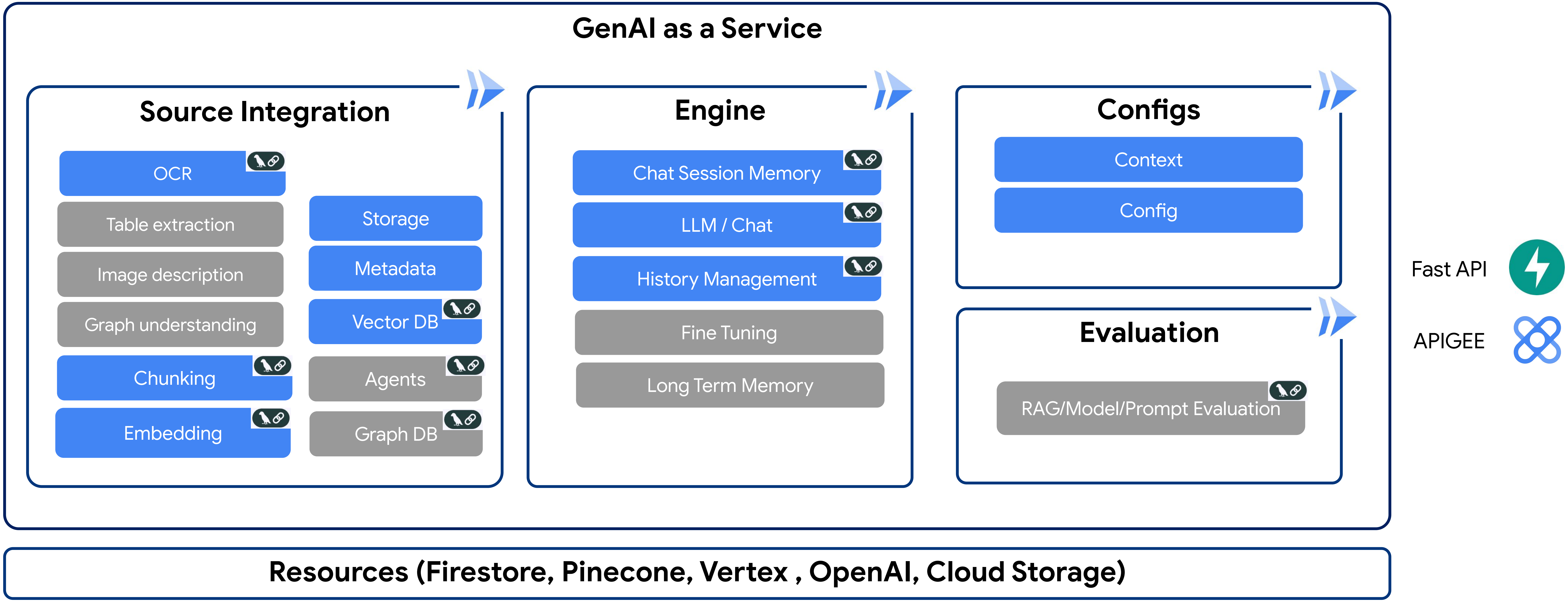
Display the sources (document, page numbers & chunks)

Proprietary

# Demo

Proprietary

# Simplified Architecture



GenAI as a Service

Source Integration
- OCR
- Table extraction
- Image description
- Graph understanding
- Chunking
- Embedding
- Storage
- Metadata
- Vector DB
- Agents
- Graph DB

Engine
- Chat Session Memory
- LLM / Chat
- History Management
- Fine Tuning
- Long Term Memory

Configs
- Context
- Config

Evaluation
- RAG/Model/Prompt Evaluation

Fast API

APIGEE

Resources (Firestore, Pinecone, Vertex , OpenAI, Cloud Storage)

# Our contributions to LangChain

1. System prompt for **Gemini Pro**

2. System prompt for **Code Chat Bison**

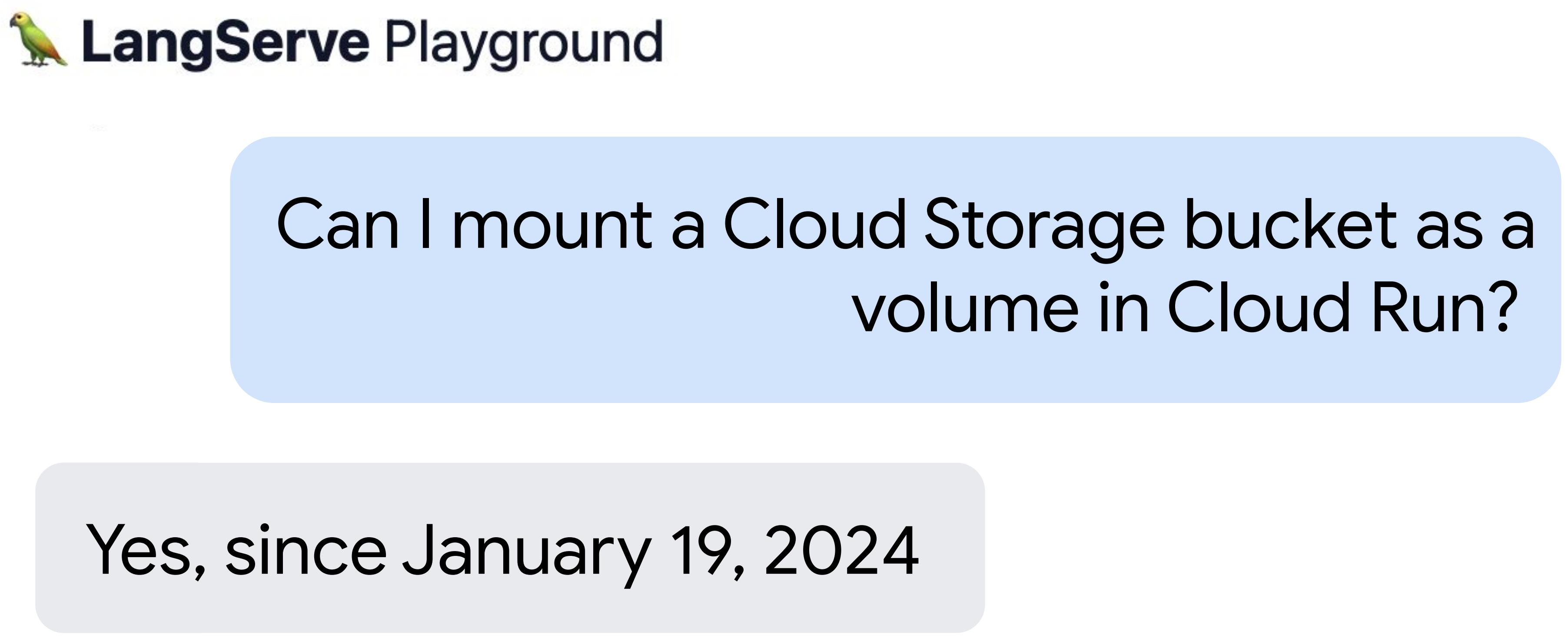
3. Token count method for Summarization - BufferMemory **Gemini & Mistral Large**

Proprietary

# Wietse Venema

Developer Advocate, Google Cloud

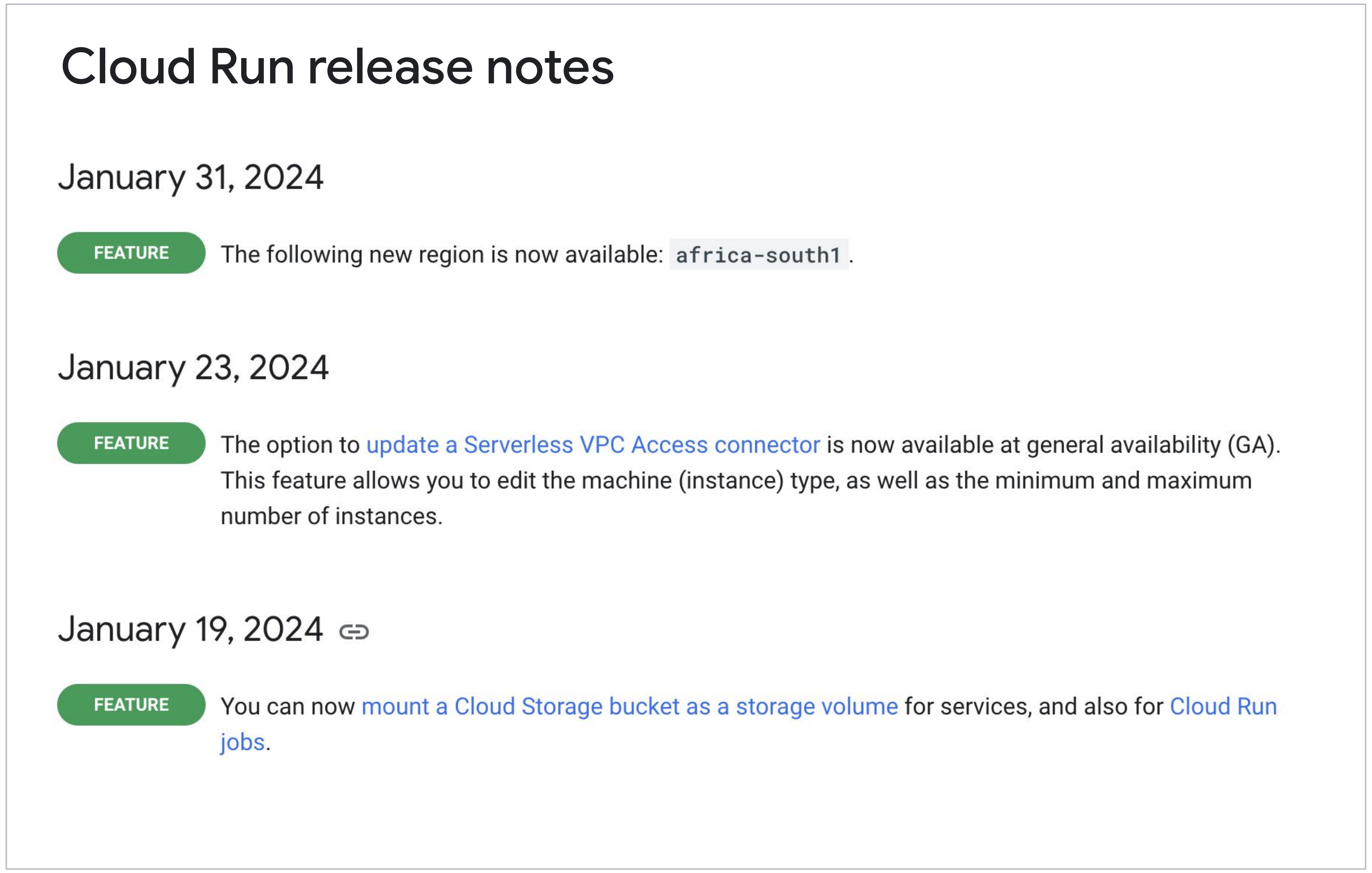
# Building and deploying a LangChain app to Cloud Run



**LangServe** Playground

Can I mount a Cloud Storage bucket as a volume in Cloud Run?

Yes, since January 19, 2024



## Cloud Run release notes

### January 31, 2024

FEATURE The following new region is now available: `africa-south1`.

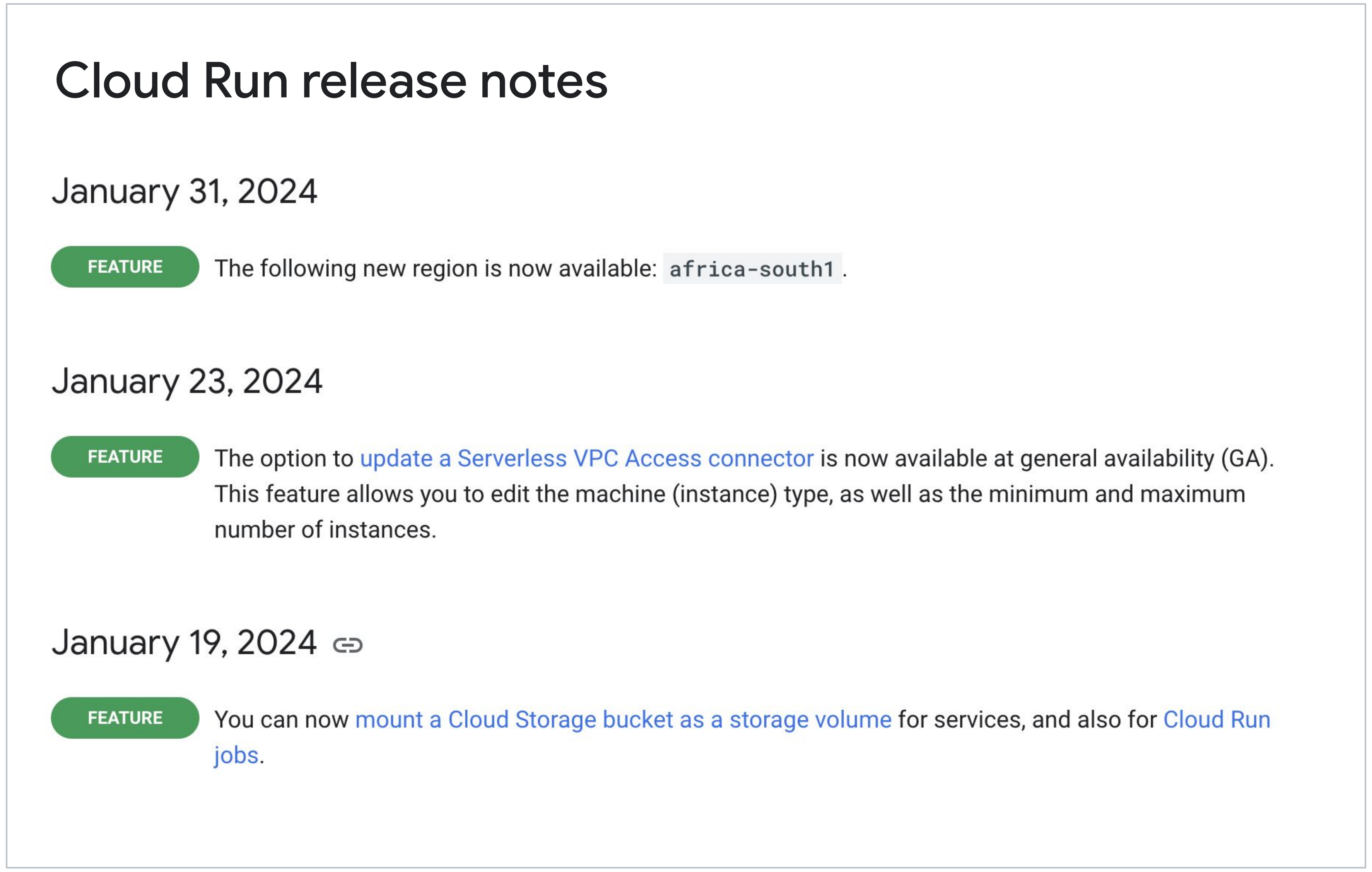
### January 23, 2024

FEATURE The option to update a Serverless VPC Access connector is now available at general availability (GA). This feature allows you to edit the machine (instance) type, as well as the minimum and maximum number of instances.

### January 19, 2024

FEATURE You can now mount a Cloud Storage bucket as a storage volume for services, and also for Cloud Run jobs.

# Exploring the system architecture



## Cloud Run release notes

### January 31, 2024

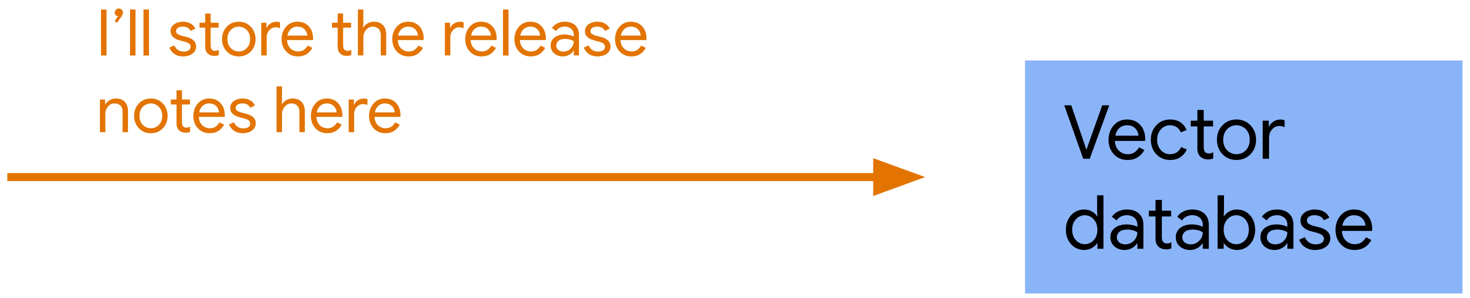
**FEATURE** The following new region is now available: `africa-south1`.

### January 23, 2024

**FEATURE** The option to update a Serverless VPC Access connector is now available at general availability (GA). This feature allows you to edit the machine (instance) type, as well as the minimum and maximum number of instances.

### January 19, 2024

**FEATURE** You can now mount a Cloud Storage bucket as a storage volume for services, and also for Cloud Run jobs.



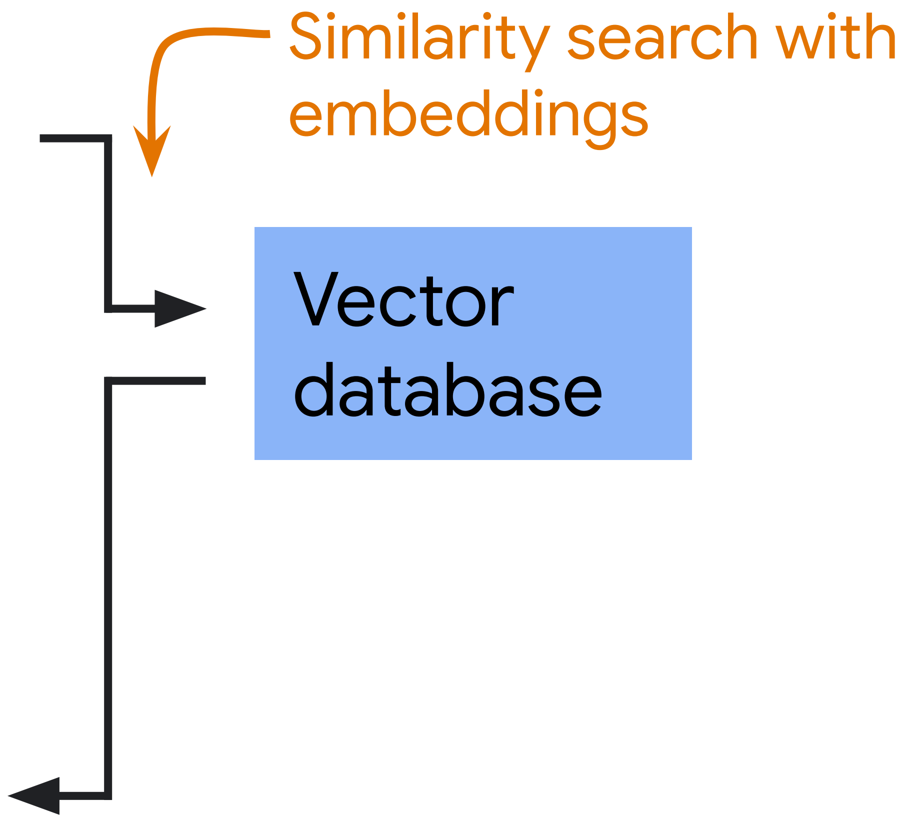
I'll store the release notes here

Vector database

# Finding release notes similar to the user query

**USER QUERY**

Can I easily mount Cloud Storage buckets as a volume?



Similarity search with embeddings

Vector database

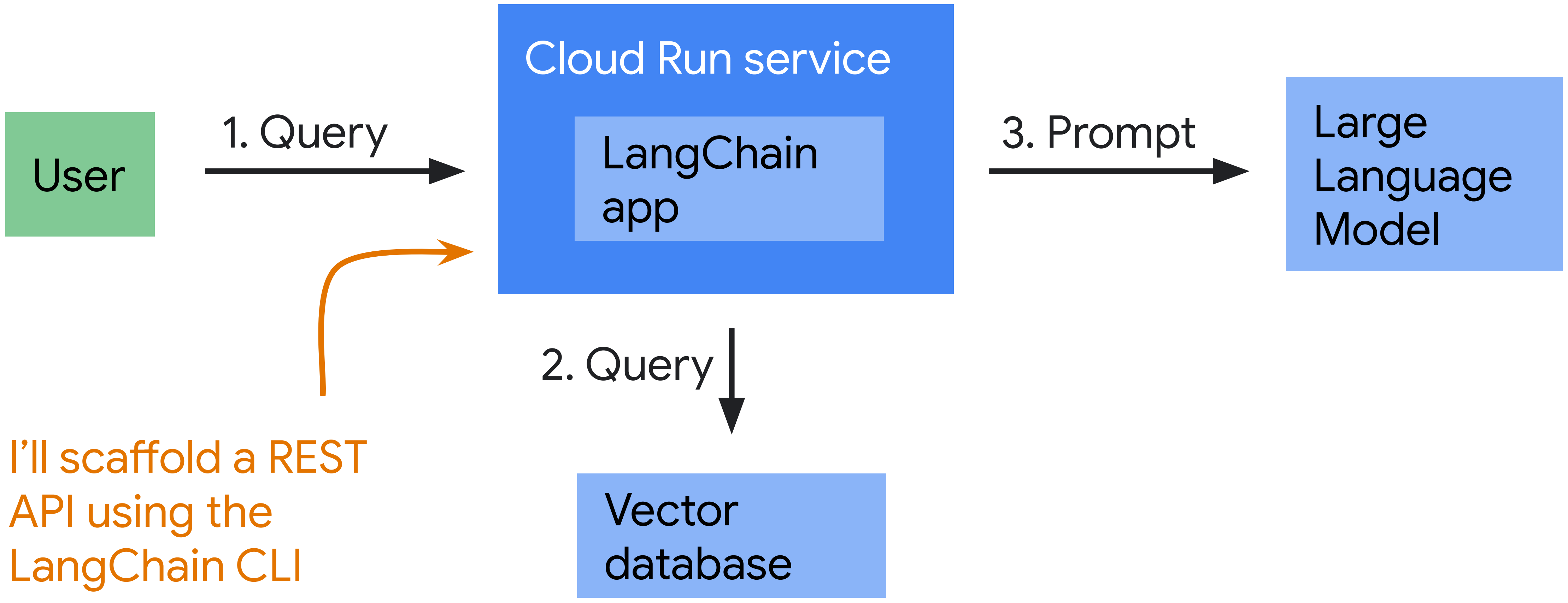
**SIMILAR RELEASE NOTES**

January 19, 2024: You can now mount a **Cloud Storage bucket** as a storage **volume** for services, and also for Cloud Run jobs.

March 19, 2024: You can now **mount** an NFS file share **as a volume** for Cloud Run services and jobs. (In Preview)

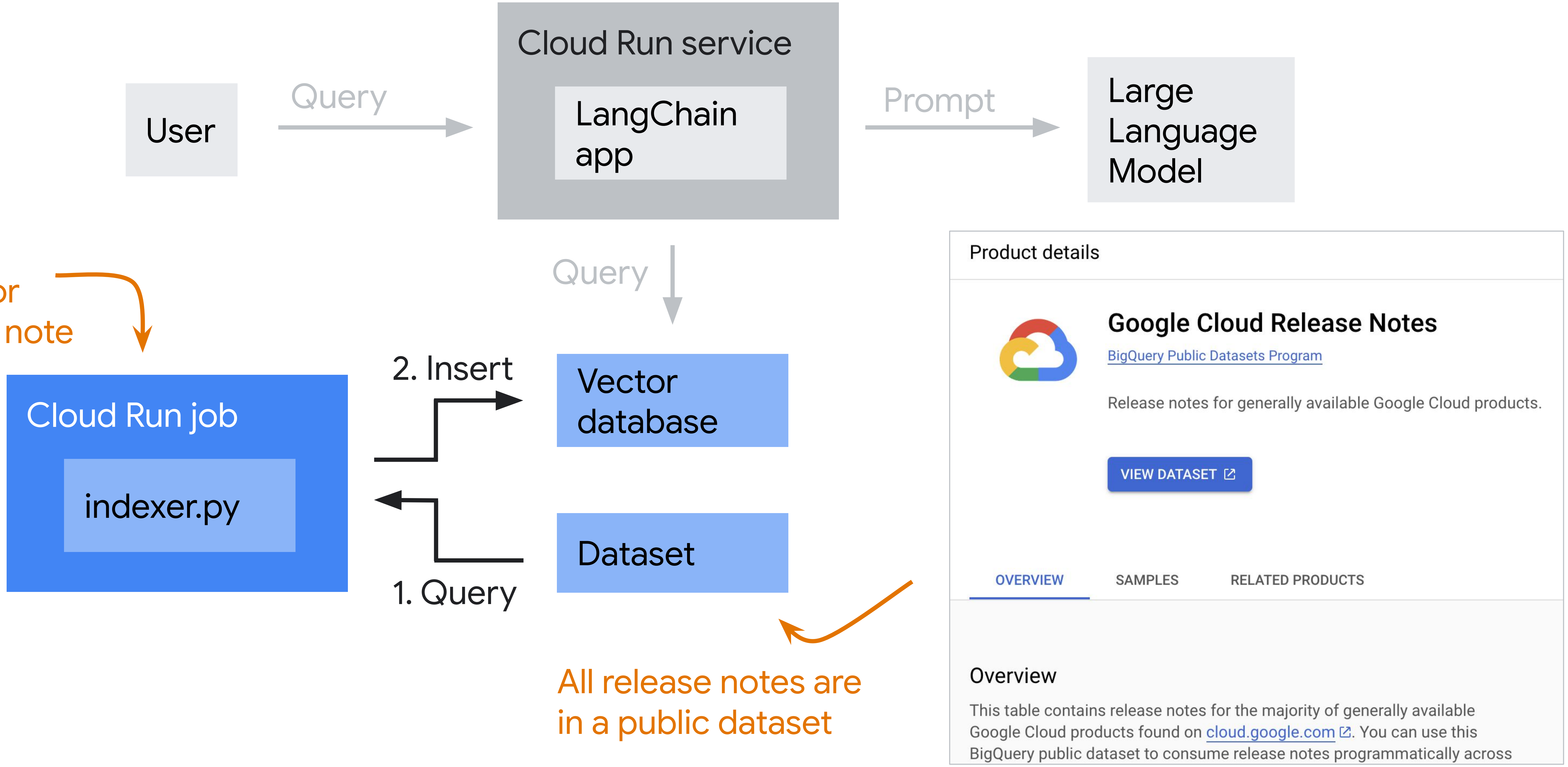
May 12, 2023: You can now configure in-memory **volumes** for your Cloud Run containers. (In Preview)

# Serve the LangChain app on Cloud Run



I'll scaffold a REST API using the LangChain CLI

# Use Cloud Run jobs to import the release notes from BigQuery



Compute an embedding for every release note

All release notes are in a public dataset

Release notes for generally available Google Cloud products.

Overview

This table contains release notes for the majority of generally available Google Cloud products found on cloud.google.com. You can use this BigQuery public dataset to consume release notes programmatically across

# Creating the infrastructure resources

# Access the codelab

Deploy LangChain on Cloud Run with LangServe

**Scan the QR code or go to goo.gle/langserve**

Proprietary

# In conclusion

| | |
|---|---|
| GenAI apps | are context-aware and rely on a language model to reason. |
| LangChain | is a flexible toolkit for building GenAI apps, with interchangeable data sources and building blocks. |
| Cloud Run | is a perfect match to run LangChain and serve your GenAI app |

# Ready to build what's next?

Tap into **special offers** designed to help you **implement what you learned** at Google Cloud Next.

**Scan the code** to receive personalized guidance from one of our experts.

Or visit **g.co/next/24offers**

Proprietary

# Continue your learning journey!

## Cloud Run

**DEV205** - Cloud Run: What's new

**DEV236** - Ford Motor Co.'s acceleration to Google Cloud fueled by Cloud Run

**DEV228** - How to deploy all the JavaScript frameworks to Cloud Run

## LangChain

**AIML151** - Fireside chat with Harrison Chase

**DEV201** - Go from LLM to market faster with Ray, Hugging Face, and LangChain

## GenAI

**DEV202** - Accelerate your AI with Serverless

**AIML302** - Operationalizing generative AI on Vertex AI

# Thank you

# We are interested in your feedback!

Connect with a
Product Manager or
UX researcher.