

Google Cloud

Next '24

Go from large language model to market faster with Ray, Hugging Face, and LangChain



Alex Zakonov

Senior Director of
Engineering, Kubernetes,
Google Cloud



Brandon Royal

Product Manager,
Kubernetes,
Google Cloud

Large Language Models are limited by the data on which they're trained



**Non-Public
Data**

Unable to reason about non-public data

**Domain
Expertise**

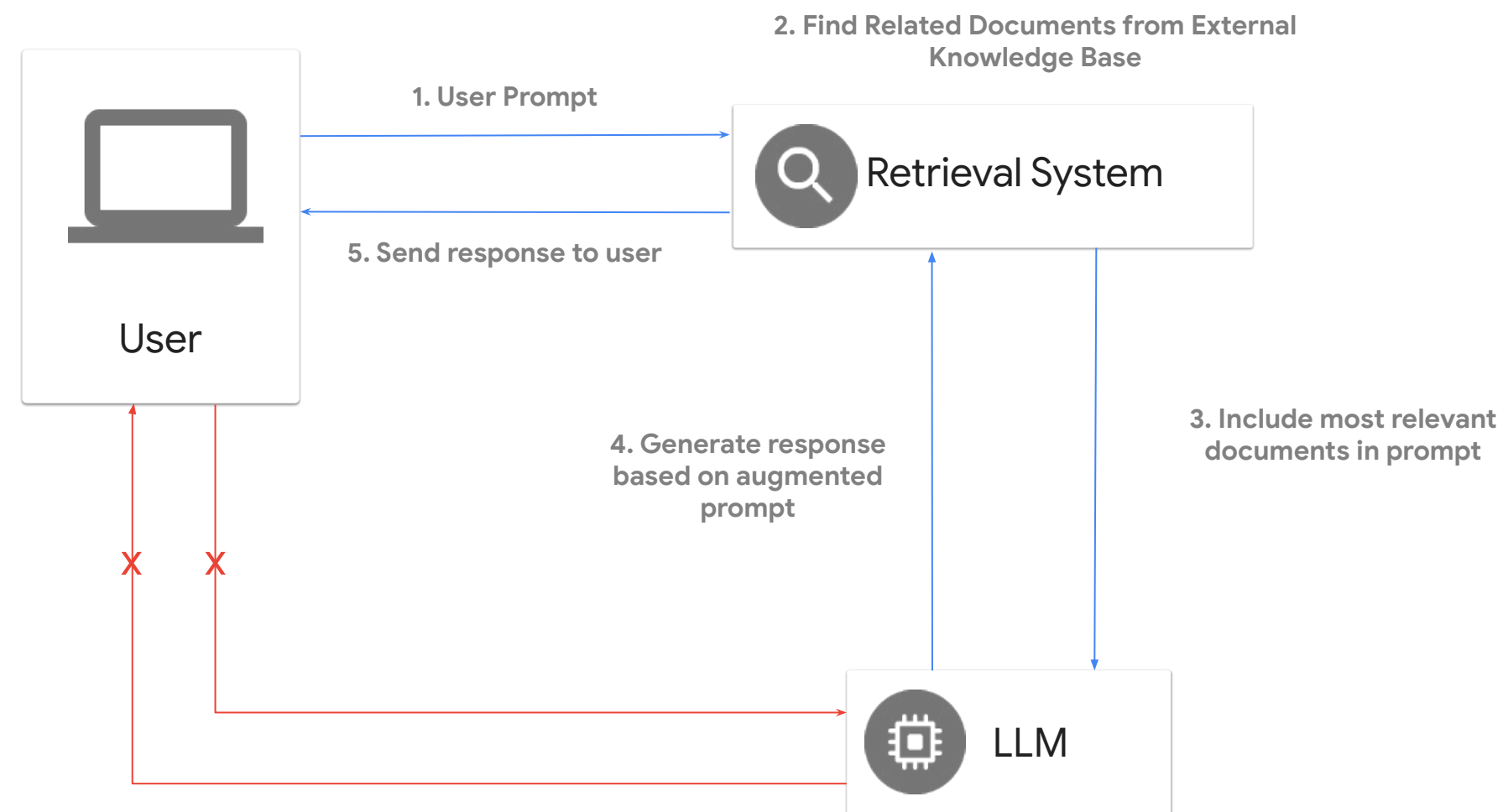
Higher domain specificity of the query increases likelihood of hallucinations¹

**Source
Citations**

Inability to provide references/citations

Retrieval Augmented Generation

A technique for optimizing the output of an LLM by injecting relevant information into the prompt context, typically through semantic retrieval systems



AI Platforms on Google Cloud

Compute Engine

VMs, tooling, and workflow support to enable scaling from single instances to global cloud computing

Kubernetes Engine

A managed environment for deploying, scaling and operating containerized applications

Cloud Run & Vertex AI

Take AI projects from ideation to production, quickly and cost-effectively

Dataproc & Dataflow

Leverage GPU-accelerated data processing & analytics

TPUs

GPUs

CPUs



Leveraging AI on Google Cloud has enabled us to be closer to our consumers and deliver more personalized and seamless experiences with their appliances.”

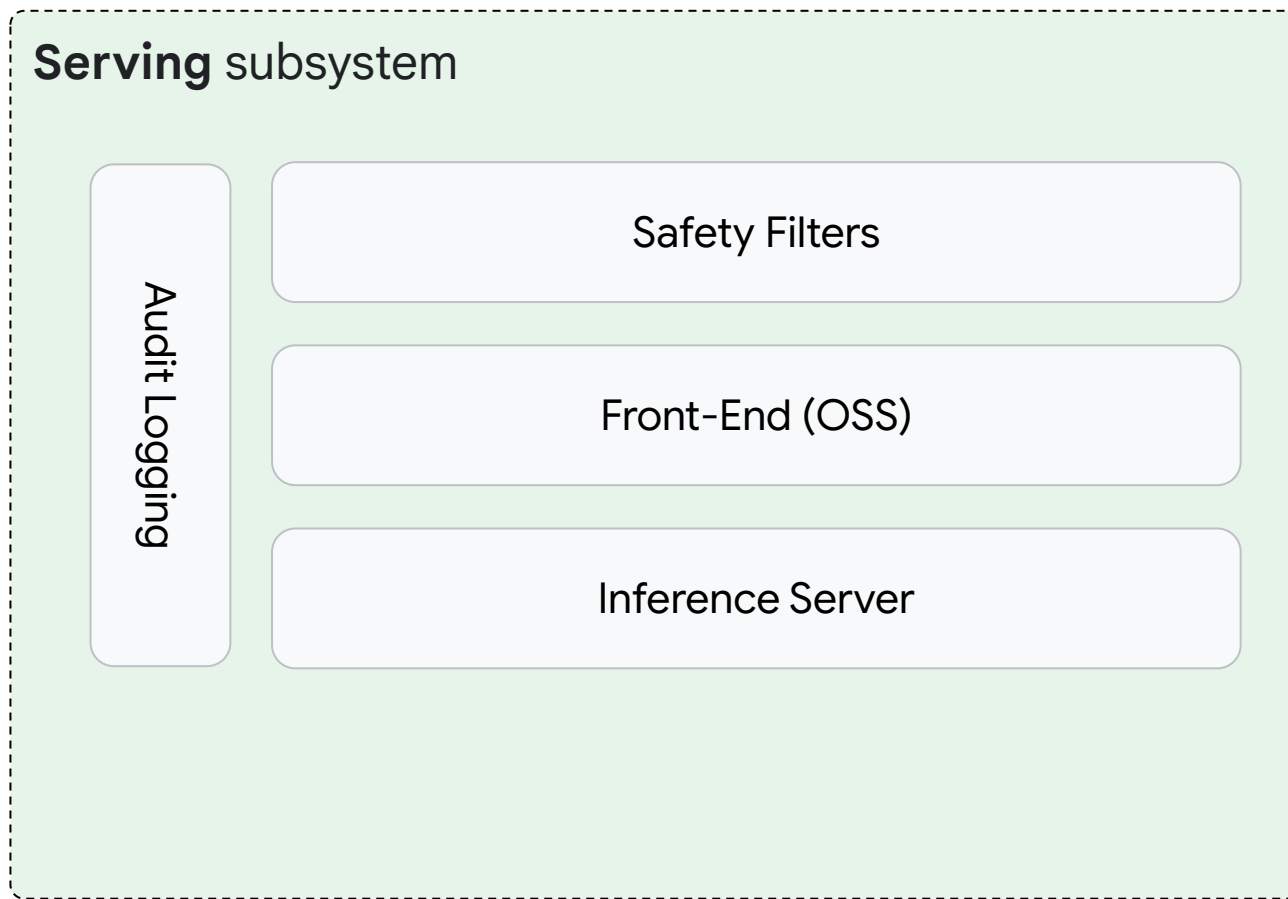
Adam Jones, Senior Director Cloud Products & IoT Services, *GE Appliances*

Let's get building

We'll focus on the practical approaches to:

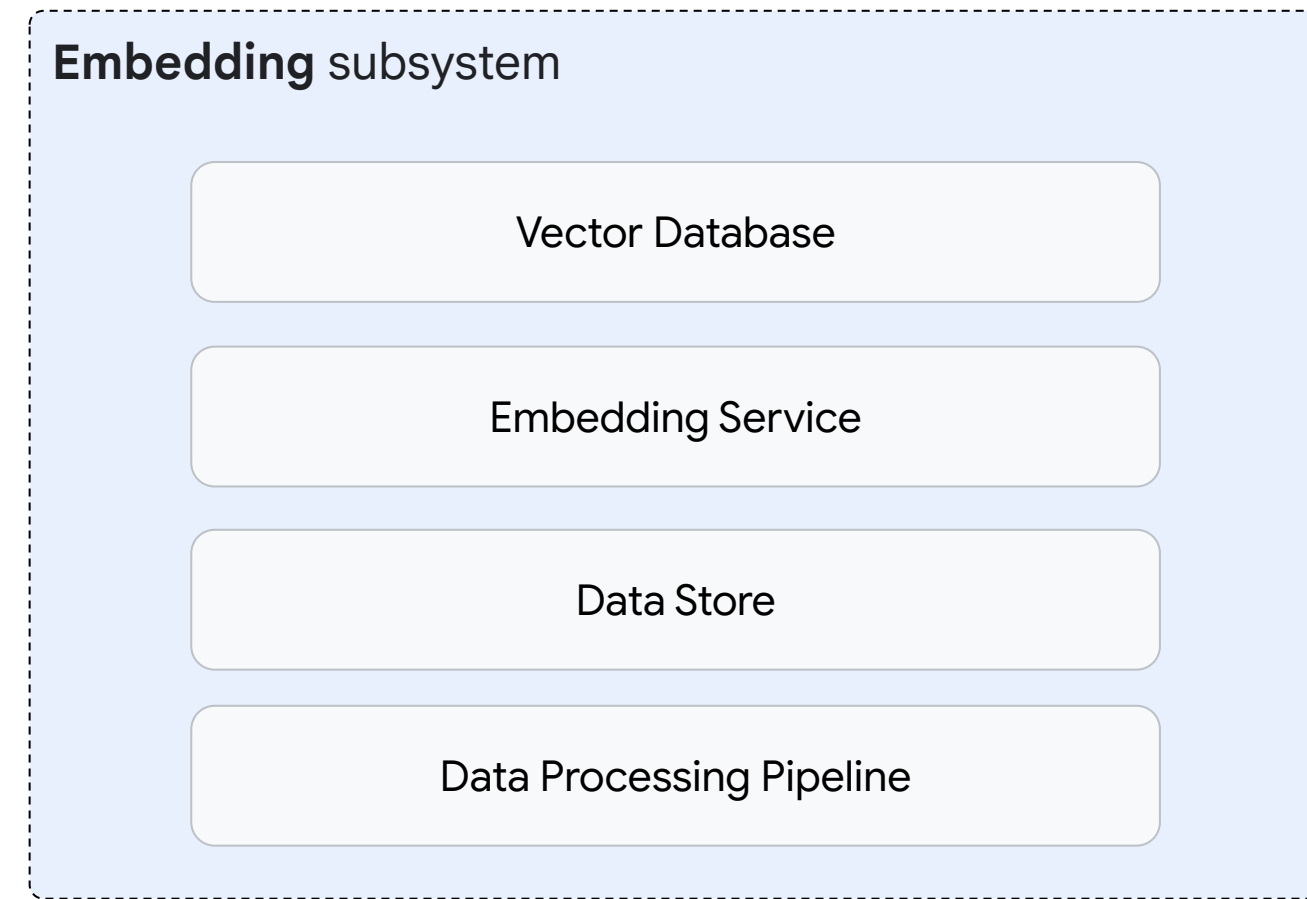
- Build a RAG application on Google Cloud with open source tools and frameworks
- Learn helpful practices from customers and internal Google teams
- Deploy a RAG application in your GCP environment guiding quick start solutions

Serving subsystem



- Horizontal scaling on requests and tokens
- Query and response orchestration - i.e. “chaining”
- Policy based filtering and safety checks

Embedding subsystem



- Horizontal scaling on data processing jobs
- GPUs or TPUs for accelerating embeddings
- Augment and structure raw data for better performance

Architectural Principles for RAG

- ✓ Optimize for experimentation with loosely coupled components
- ✓ Leverage open frameworks and components
- ✓ Embrace mixed retrieval systems (semantic, relational, document)
- ✓ Separate platform from application concerns
- ✓ Design for safety and security at every layer of the stack

Unified Compute and Orchestration Platform for AI Applications

- Industry leading scale with up to 50K TPU chips¹ and 15K nodes per cluster²
- Superior price-performance with GPU and TPUs, multi-tenant job queuing, GPU sharing and fast pod starting
- Efficient operations with GKE Autopilot
- Fully-managed Kubernetes experience with AI and app workloads from the top Kubernetes contributor

Google Kubernetes Engine

Open Software and Frameworks

JAX, TensorFlow, PyTorch, XLA

Jupyter, Ray, KubeFlow, Spark

Distributed Training

Kueue Job Queuing

High Throughput Scaling

Scaled Inference

Autopilot

Pod Fast Starts

Node Provisioning and Autoscaling

Dynamic Workload Scheduler

Flexible Consumption (On-Demand, CUD, Spot)

Google Cloud Infrastructure (CPU/GPU/TPU)

Stephen Allen

Cloud Engineer,
GE Appliances




SmartHQ Assistant

- At GE Appliances, our goal is to be ‘zero distance’ to our consumer, delivering experiences that truly resonate.
- Helps consumers quickly find answers to use and care questions for their registered appliances.
- Iteratively improve this uniquely personalized experience to our consumers with Google Cloud.



GE APPLIANCES
a Haier company

SMARTHQ ASSISTANT



What can I help you with?

You can ask me things about your appliances. For example, if you have a refrigerator you can ask what filter it uses and we will look it up for you in your use and care manual

Optimization Strategies



Prompt Intent Classification

Prompt classification simplifies complex tasks by decomposing them into more straightforward queries.

Enterprise Adoption

Launched a self-service portal to empower business stakeholders to manage reference data and minimize development team bottlenecks.

Continuous Evaluation

Utilize automated evaluation jobs, highlighting progress and identifying potential degradations in application performance.

Results Beyond MVP

1

**200% increase in user
engagements**

2

**103% increase in answer
found rate**

3

**42% reduction in
experienced latency**

Voice of the Consumer

SmartHQ Assistant has expanded our ability to better understand our appliance owners and rapidly deliver relevant value to them.

Continuous Improvement

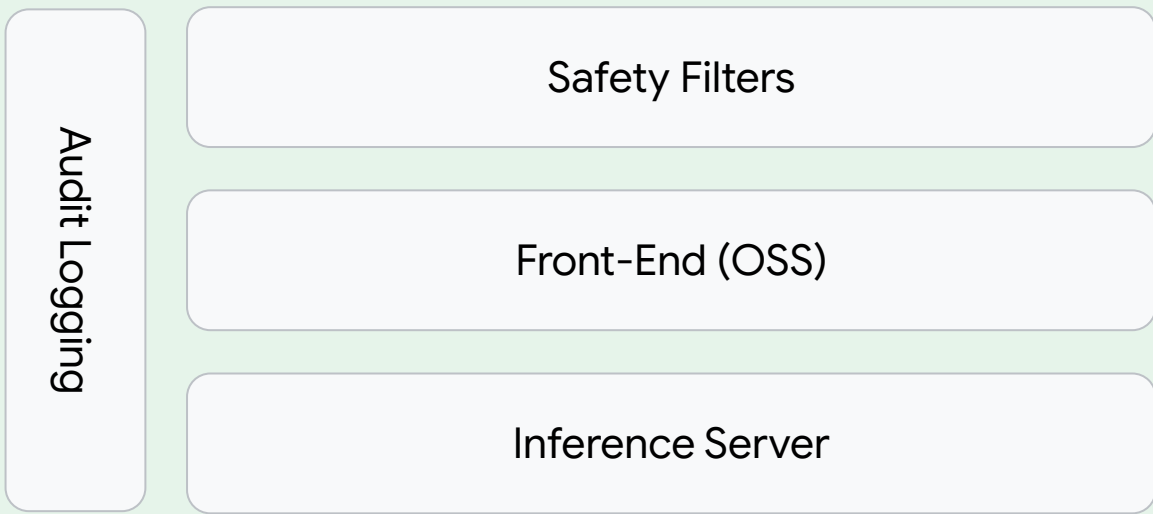
RAG applications require high levels of observability and continuous evaluation. Understanding evolving usage patterns is fundamental to long-term success in production.

Building AI Apps using RAG Architecture

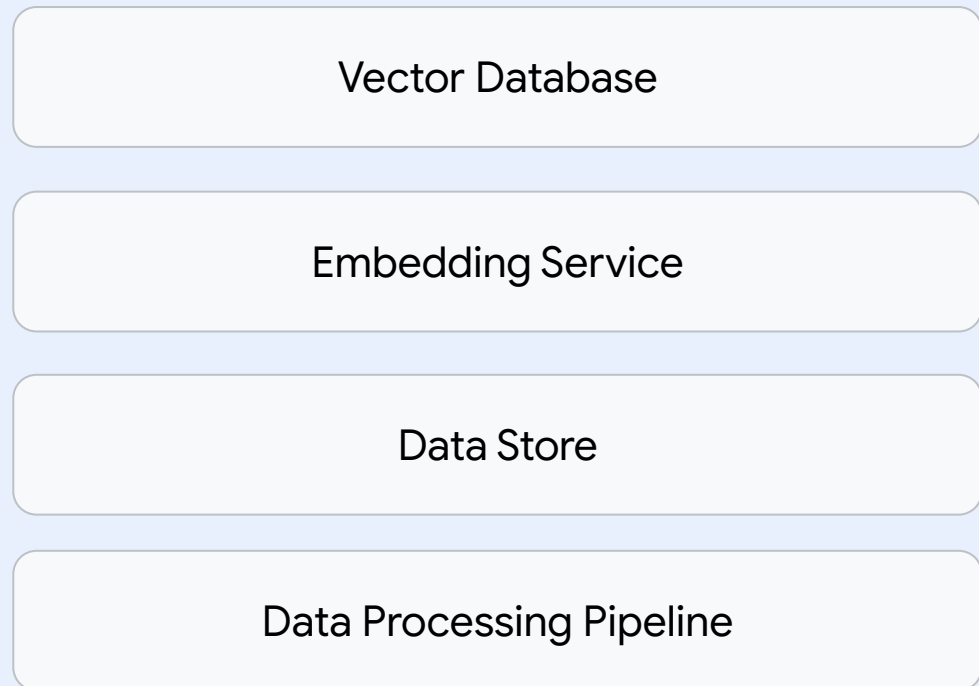
The easiest part is the technology

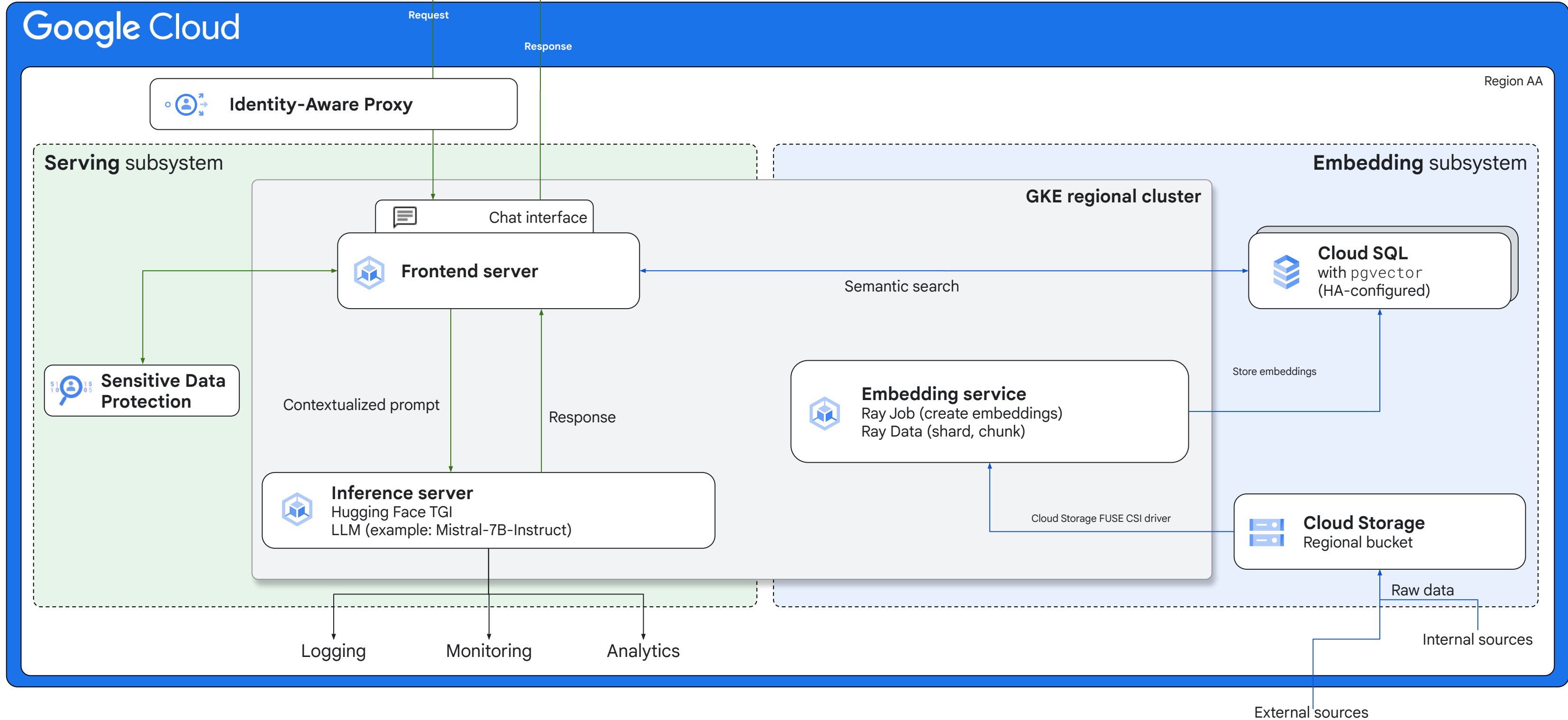
- ✓ Engage Legal, PR, Compliance and other business teams early in design process
- ✓ Most users aren't skilled in prompt engineering and need assistance with guided prompts and training
- ✓ Engage subject matter experts in the data to develop your chunking strategy
- ✓ Manual testing doesn't cut it. Add automated evaluations into testing

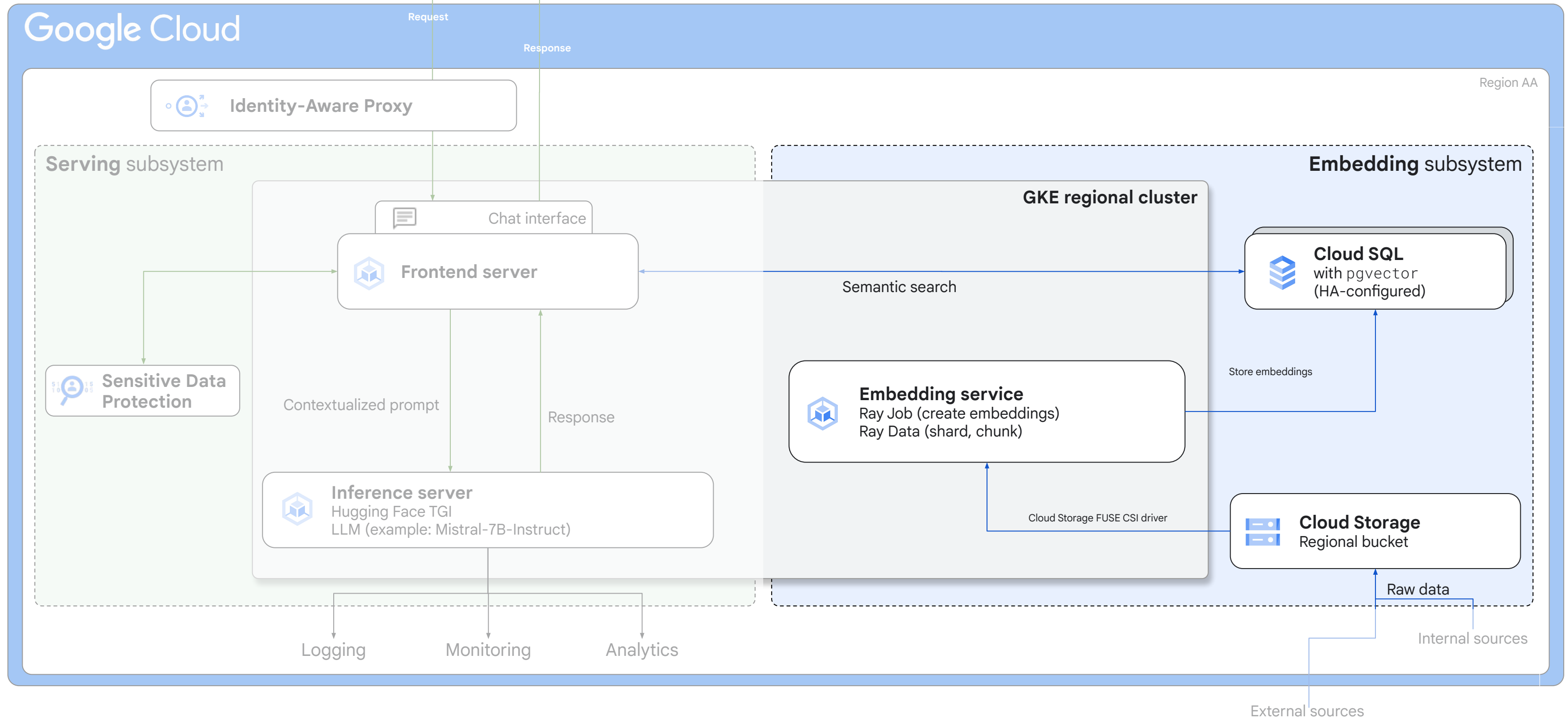
Serving subsystem



Embedding subsystem

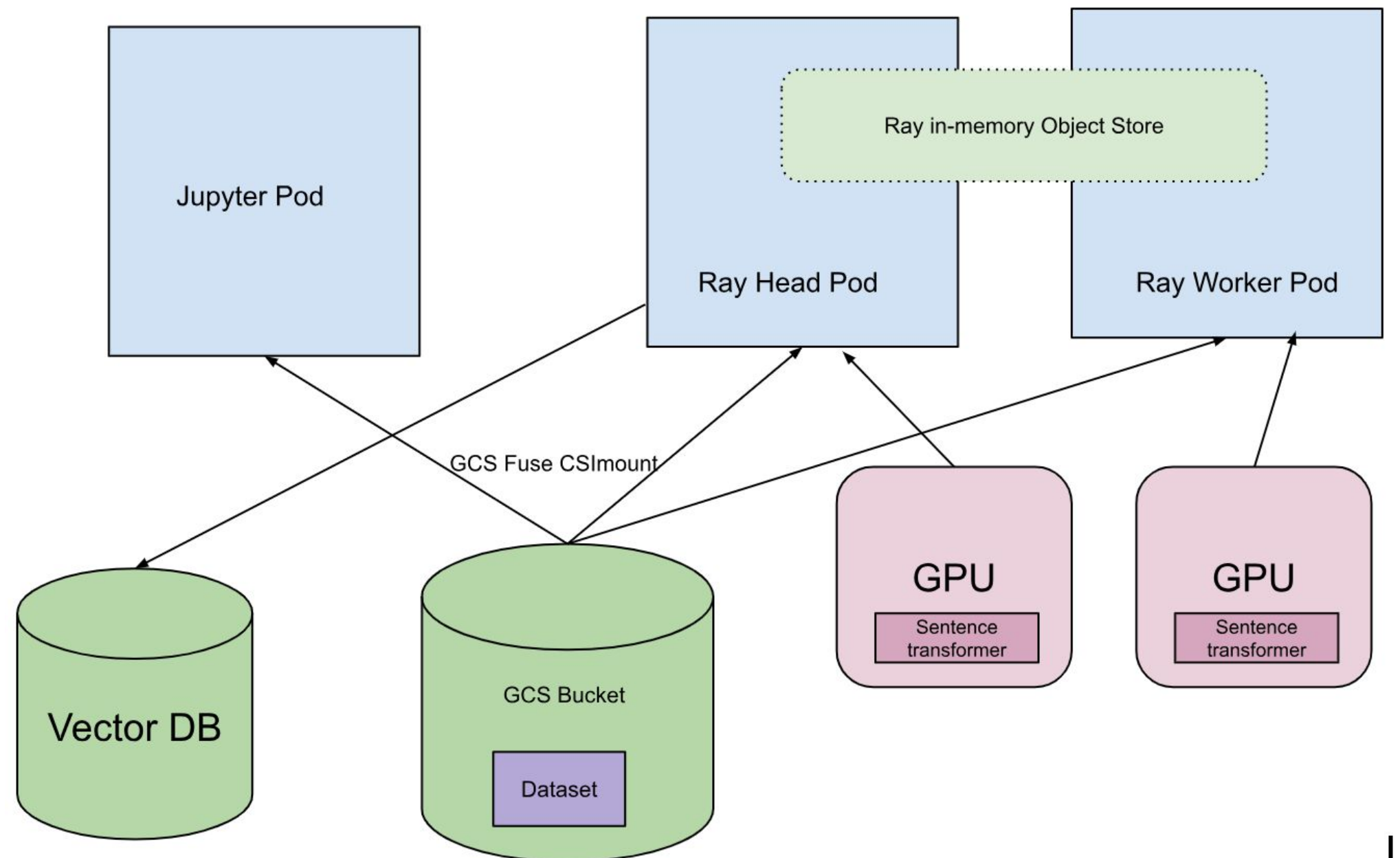






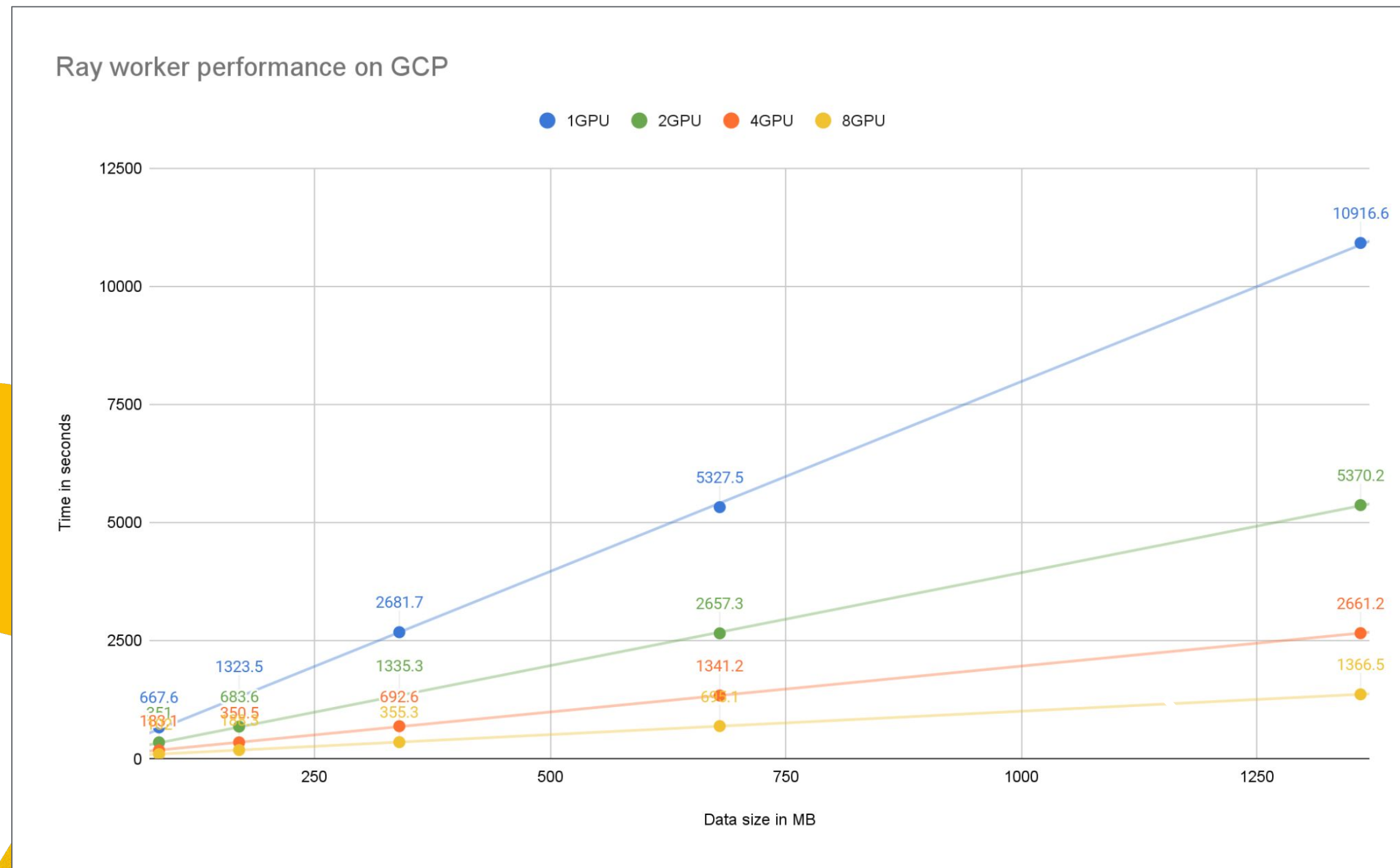
Scaling embedding pipeline with Ray Data and GCSFuse

1. Auto-mount GCS data to GKE via [GCSFuse CSI driver](#)
2. Process GCS data in parallel via [Ray Data API](#)
3. Quickly load embeddings from Ray cluster to [Cloud SQL for PostgreSQL and pgvector](#)

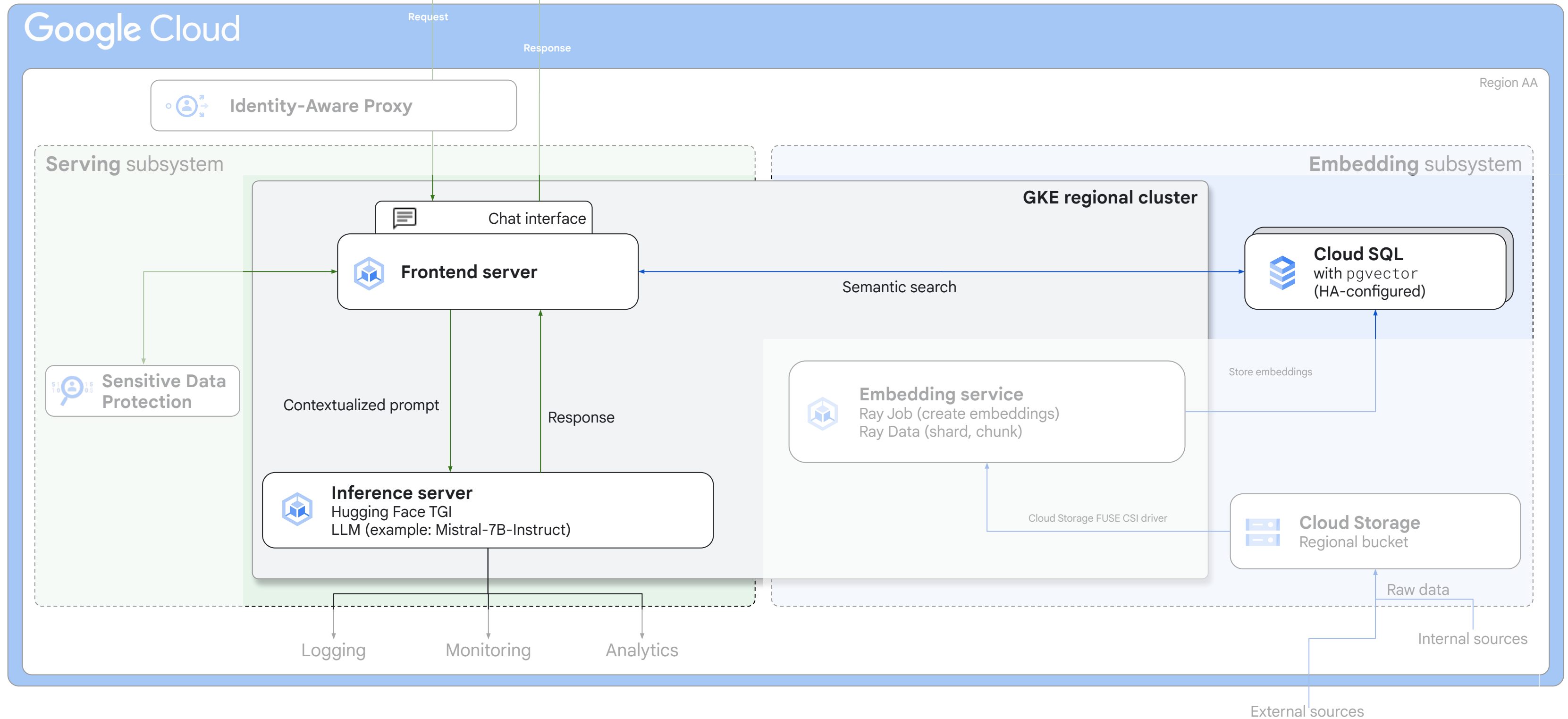


Proprietary

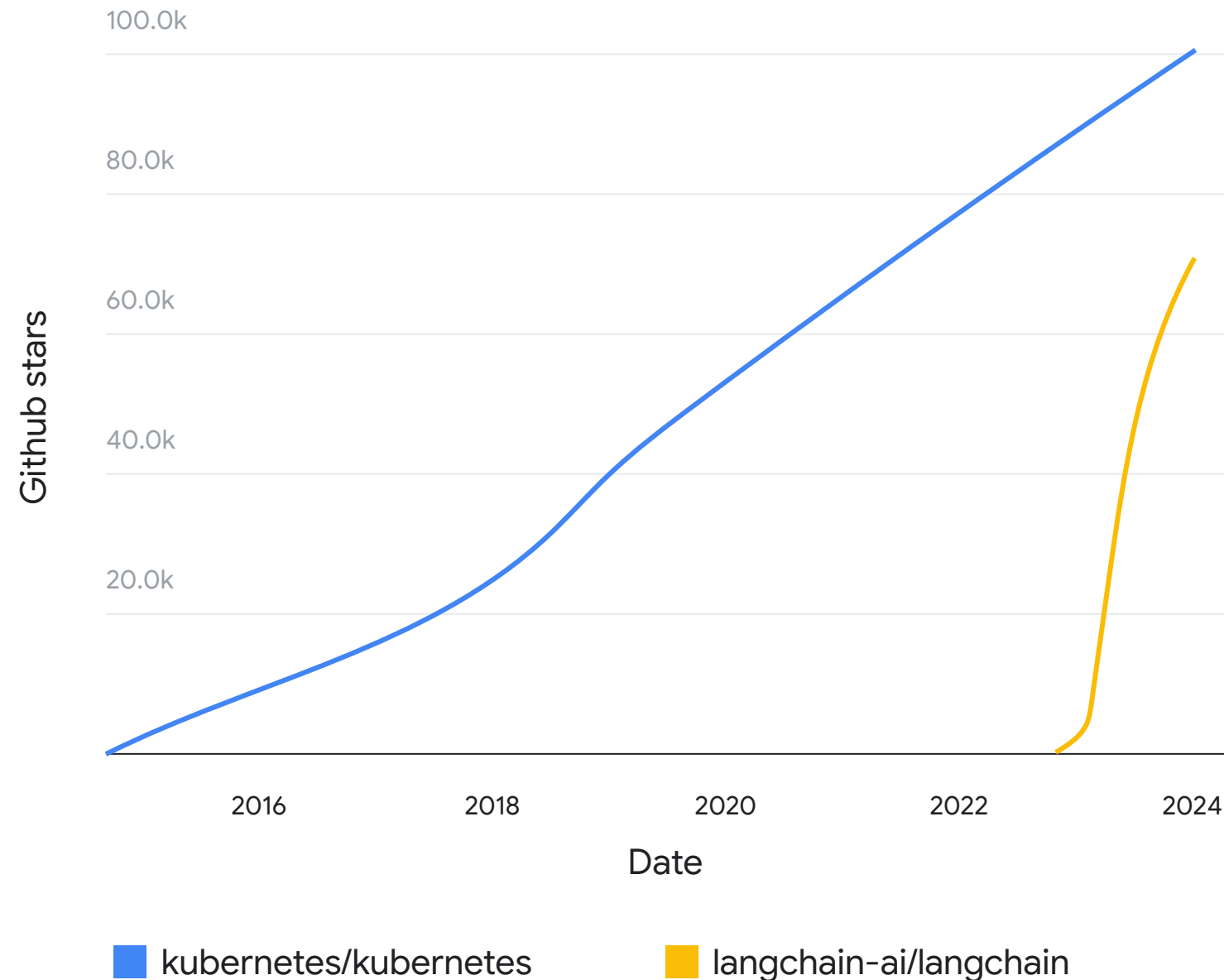
Scaling embedding pipeline with Ray Data and GCSFuse



Near linear scaling of embedding generation as GPUs are added



Star history



star-history.com

LangChain is the most popular gen AI framework

GitHub repo

+700 different integrations

+2k contributors

~70k stars

LangChain is supported across all Google Databases

Now supported in:

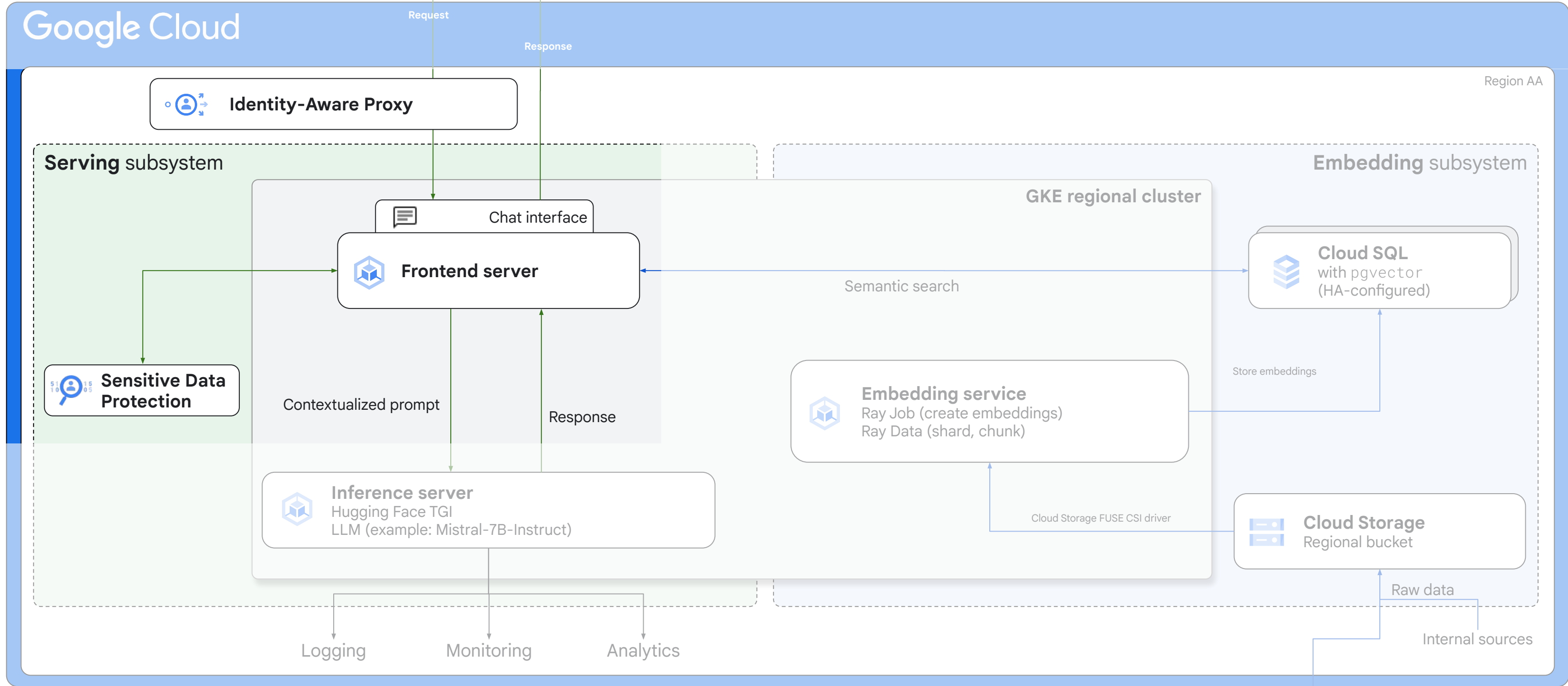
- ✓ Cloud SQL for MySQL
- ✓ Cloud SQL for PostgreSQL
- ✓ Cloud SQL for SQL Server*
- ✓ AlloyDB
- ✓ Spanner
- ✓ Bigtable*
- ✓ Memorystore for Redis
- ✓ Firestore*

* no vector store

and LLM serving on GKE

- ✓ Text Generation Inference
- ✓ vLLM

...and many more



Realtime data protection

Using AI to Protect AI

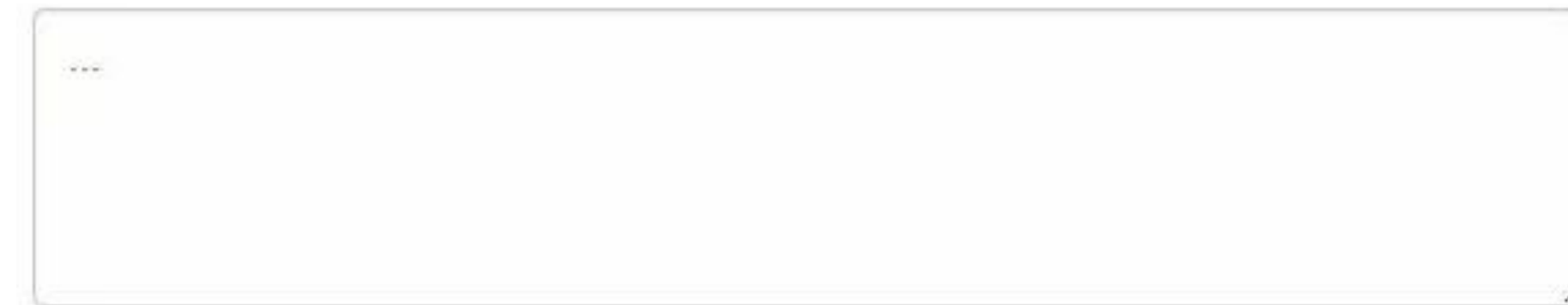
Leverage Google's leading [Sensitive Data Protection \(SDP\)](#) technology to identify, block, and mask over 150 different sensitive elements from credit card to medical context to PII and more

This is the same technology that powers content detection in Workspace, BeyondCorp, Contact Center AI, and more.

Classify, score and filter potentially harmful or inappropriate content via [Cloud Natural Language API](#)



Realtime data protection



Instructions: Submit your query below.

Press to exit full screen

Enable Filters:

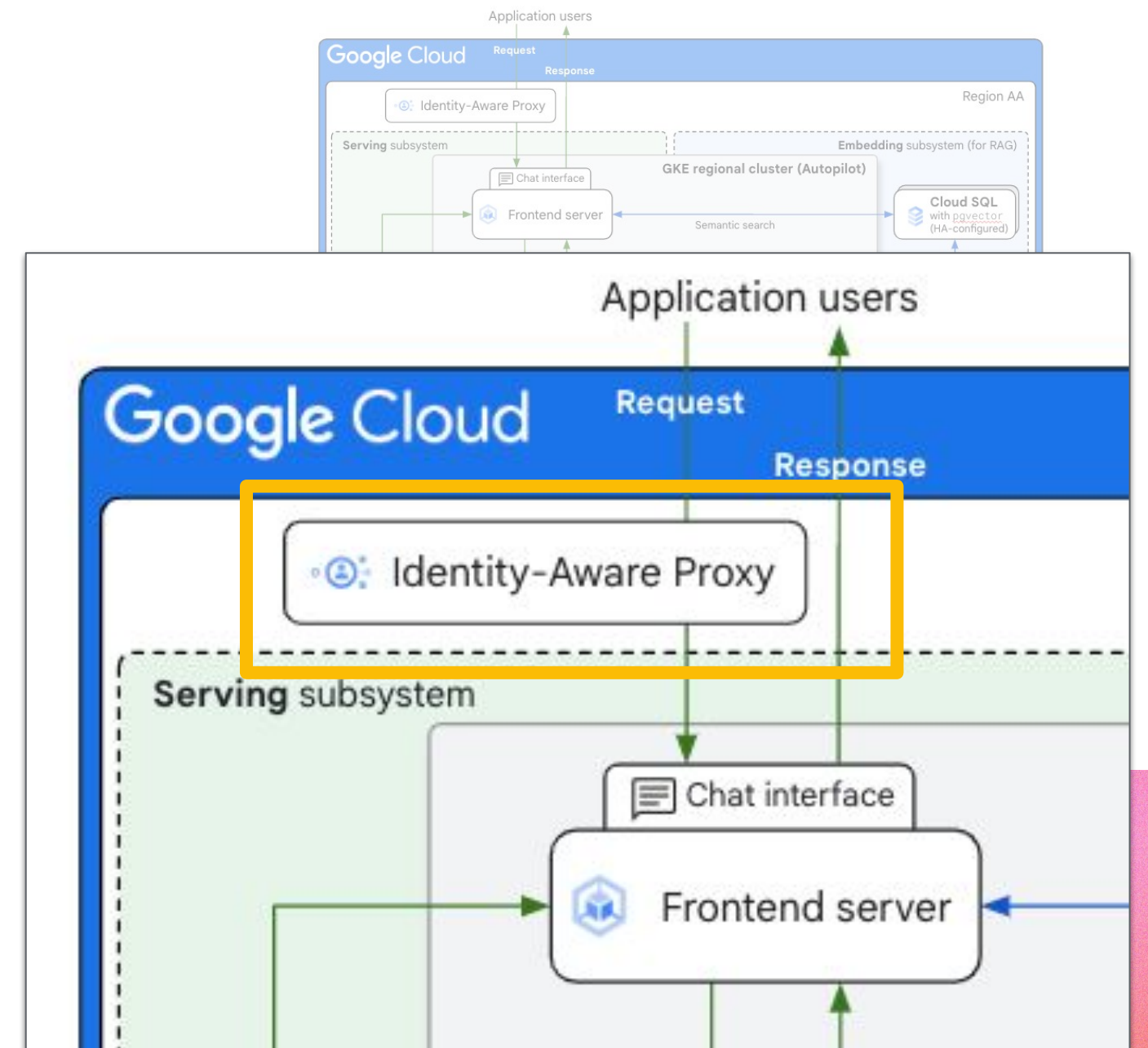


who worked with Robert De Niro and name one film they collaborated

Ingress control: load balancing, authentication & authorization

Leverage standard Google authentication via **Identity-Aware Proxy (IAP)**

2. Centrally configure user/group access control for your org or project
3. Integrates with secure, distributed global frontend via **Google Cloud Load Balancers**



- Security Command Center
 - Overview
 - Threats
 - Vulnerabilities
 - Compliance
 - Assets
 - Findings
 - Sources
 - Posture
- Detections and Controls
 - Chronicle SecOps
 - reCAPTCHA Enterprise
 - Web Security Scanner
 - Risk Manager
 - Binary Authorization
 - Advisory Notifications
 - Access Approval
 - Managed Microsoft AD
- Marketplace
- Release Notes

Identity-Aware Proxy

APPLICATIONS | SSH AND TCP RESOURCES

Identity-Aware Proxy (IAP) lets you manage who has access to services hosted on App Engine, Compute Engine, or an HTTPS Load Balancer. [Learn more](#)

To get started with IAP, add an [App Engine app](#), a [Compute Engine instance](#) or configure an [HTTPS Load Balancer](#).

[CONNECT NEW APPLICATION](#) Premium

Filter: Enter property name or value

Resource	IAP	Method	Connection	Published	Status
<ul style="list-style-type: none"> All Web Services Backend Services <ul style="list-style-type: none"> rag/rag-frontend rag/proxy-public 	<input checked="" type="checkbox"/>	IAM		Global HTTP(S) Load Balancer (classic): k8s2-um-qileh3b7-rag-frontend-ingress-oivmi6xy	OK
	<input type="checkbox"/>	IAM		Global HTTP(S) Load Balancer (classic): k8s2-um-qileh3b7-rag-jupyter-ingress-xb9vxnm1	OK

k8s1-d17e140a-rag-rag-frontend-8080-fdc16374

Use external identities for authorization START

To grant access to the application, click "Add Principal" and select the *IAP-secured Web App User* role. [Learn more](#)

Edit or delete permissions below, or select "Add Principal" to grant new access. ADD PRINCIPAL

Show inherited permissions

Filter: Enter property name or value

Role / Principal | Inheritance

- Editor (6)
- Owner (11)
- Viewer (9)

Principal "thebill@google.com" successfully removed from role "IAP-secured Web App User" on resource "k8s1-d17e140a-rag-rag-frontend-8080-fdc16374"

Show debug panel

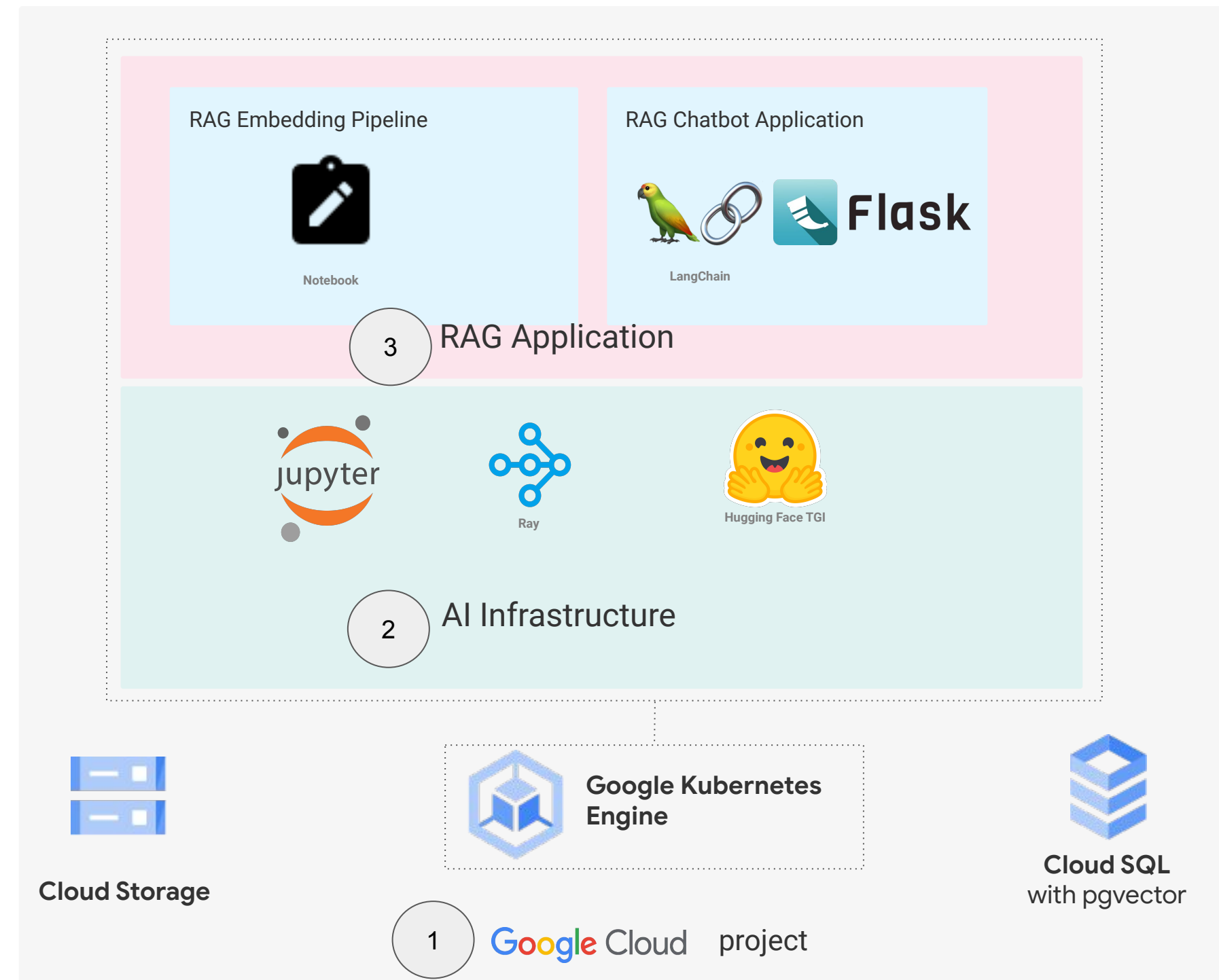
Streamline Kubernetes with GKE Autopilot

1. Accelerate go-to-market for AI applications with zero node pool configuration
2. Maximize goodput with automatic scale-up and scale-down of GPU machines
3. Reduce day-2 operations with Google-managed nodes and opinionated security defaults

```
apiVersion: v1
kind: Pod
metadata:
  name: tensorflow
  labels:
    pod: tensorflow-pod
spec:
  nodeSelector:
    cloud.google.com/compute-class: "Accelerator"
    cloud.google.com/gke-accelerator: nvidia-tesla-a100
  containers:
    - image: tensorflow/tensorflow:latest-gpu-jupyter
      name: tensorflow-a100
      resources:
        requests:
          nvidia.com/gpu: "1"
```

RAG Quick Start: all-in-one platform and sample application

1. **Google Cloud Project:** configures your project with GKE cluster, Cloud Storage and Cloud SQL with pgvector
2. **AI Infrastructure:** provisions Ray, Hugging Face TGI, Jupyter
3. **RAG application:** provides Jupyter notebook to load embeddings and installs Chatbot webapp



DEMO

Pulling it all together

We want to hear from you!

Scan to engage product experts on your RAG application journey



Thank you