Modern PostgreSQL for AI applications: with LLM integrations and powerful search

Di Qi & Umur Cubukcu
October 22, 2024

LANTERN

ubicloud

About the speakers

Umur Cubukcu

Co-founder and Co-CEO, Ubicloud

- Citus Data
- Azure PostgreSQL
- Y Combinator

Di Qi

Co-founder, CEO, Lantern

- Facebook
- Y Combinator
- Quick-commerce startup

Agenda

Four topics we'll cover in this talk:

1. Demonstrate how to build an LLM chatbot on a codebase, using PostgreSQL, in less than 15 minutes
2. How to scale your model to larger datasets, with high performance
3. How to tap into both AI and Postgres operations expertise – whether your existing Postgres is running on-prem or in the cloud
4. How you can do all this while keeping all your data, in Europe.

Part I:
Building an LLM chatbot with Postgres

ChatGPT doesn't always work

Common issues:

- New or frequently updating data
- Access to private data
- Hallucinations

To illustrate this, let’s ask ChatGPT-4o a question about the Ubicloud codebase: https://github.com/ubicloud/ubicloud

- What embedding models does Ubicloud support? [link]

How to solve this with RAG

Retrieval augmented generation (RAG)

- Store a corpus of information
- Retrieve relevant context from this store
- Pass the context to the LLM to help answer questions

How vectors come into play for RAG

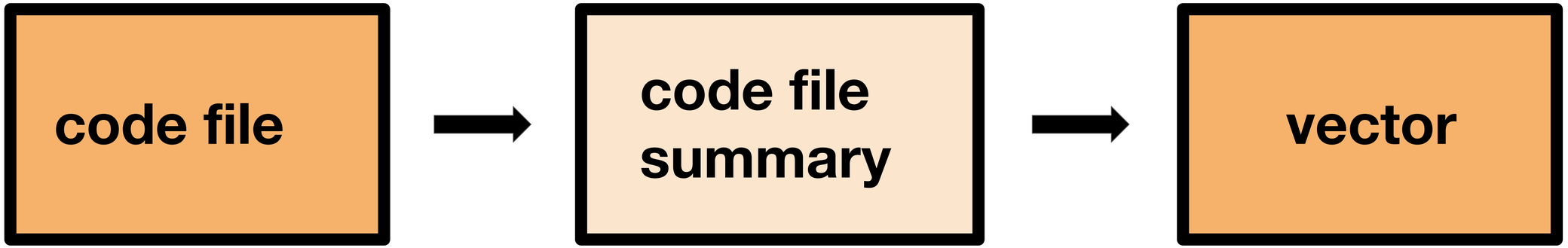
For each file, use an embedding model to generate a vector

=> Vectors enable similarity search

Store the vectors in a vector database (PostgreSQL!)

Use vector search to find related files

Let's solve this with Lantern on Ubicloud



Part 1: LLM

=> Summarize the files

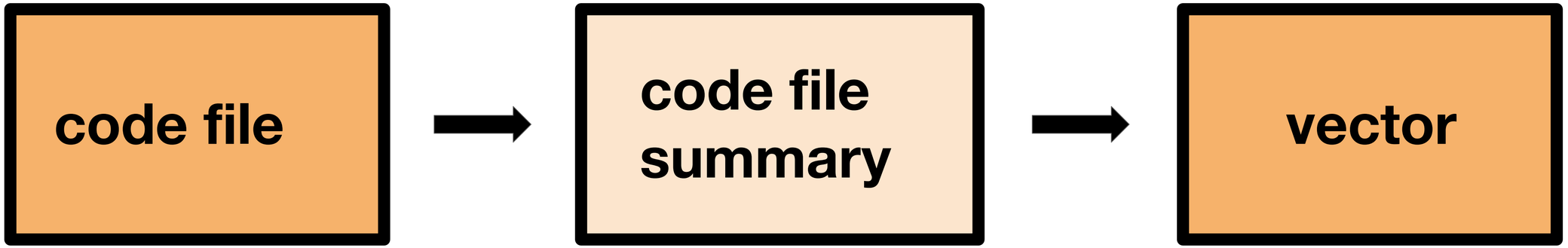
=> Better results in RAG

Part 2: Embedding

=> Embed summaries

=> Efficient semantic search

To do this, we will use the following:



Part 1: LLM

=> Llama 3B (Ubicloud)

=> Managed LLM column

Part 2: Embedding

=> Mistral 7B (Ubicloud)

=> Managed vector column

Let's *build* this with Lantern on Ubicloud

The basics:

1. Create a Lantern PostgreSQL database on Ubicloud
2. Create the schema
3. Load the data from repo
4. Managed LLM column
5. Managed vector column
6. Done!

SWITCH TO TERMINAL

See the code:

github.com/dqii/pgconf

Part II:
Performance and Scalability

Scaling from 1k records to 1M

- Add an vector index to improve performance
- Run parallel index creation for larger indexes

```
CREATE INDEX ON files USING hnsw (vector
vector_cosine_ops);
```

Caveat:

- Vector indexes are very large compared to other indexes
- Creating an vector index uses a lot of compute

Serverless indexing service scales from 1M to 100M+ records

Lantern supports using external resources to create the index. This enables arbitrarily scaling indexing resources.

```
CREATE INDEX ON files USING hnsw (vector
vector_cosine_ops) WITH (external=True);
```

2x-6x faster than GCP Cloud SQL

2.3x

Same recall

6.1x

6.1x

5.7x

5.4x

+ NVMe disks on Ubicloud

+ PostgreSQL parameters tuned for AI workloads

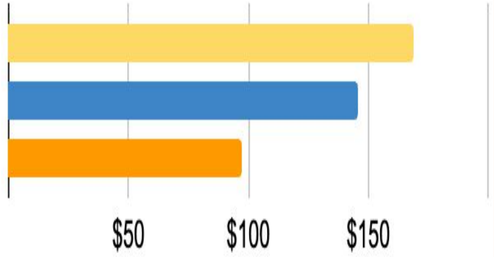
+ Lantern optimizations

LANTERN

ubicloud

3x - 9x better price-performance vs. GCP

Cloud	DBaaS	vCPU	RAM	Storage	Cost / mo
AWS	RDS Postgres	2	8 GB	128 GB network disk	$169.23
GCP	Cloud SQL	2	8 GB	128 GB network disk	$145.79
Ubicloud	Ubicloud PG	2	8 GB	128 GB NVMe	$97.50



Ubicloud on Lantern saves costs by 33% to 42%; starts at $0.14/hr

PostgreSQL as an AI database

Without costs and complexity of a new AI DBMS

- ~10x lower costs vs Pinecone, at 10 queries / sec [Appendix 1]
- Dedicated PG instances (vs. serverless pricing), with burstable indexing

Automate embedding generation and LLM calls with PostgreSQL

Open source, without lock-in

(1) ~$7,500 per month for Pinecone on AWS // ~$750 per month for Lantern on Ubicloud

Part III:
Integrating AI with your existing PG setup

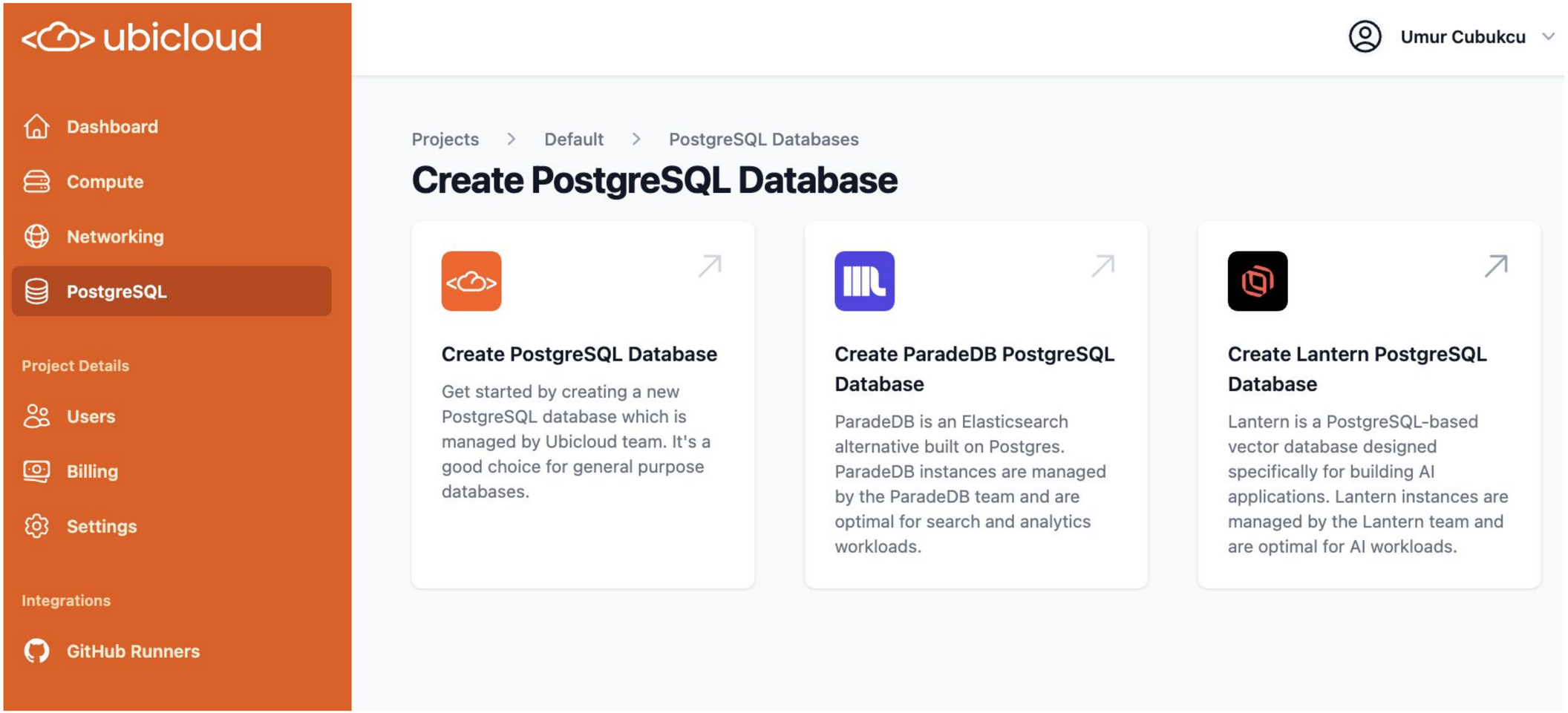
Lessons from running PG extensions in DBaaS over past ~10 years

	Example	Limitations
1. Purpose-built PostgreSQL DBaaS, for serving 1st party PG extension	Citus Cloud	Separate DBaaS per PG extension
2. General PostgreSQL DBaaS, with given extension(s) enabled	Azure DB for Postgres, with Timescale	Access to 1st party knowledge of PG extension

Combining 1st party extensions expertise with PostgreSQL DBaaS

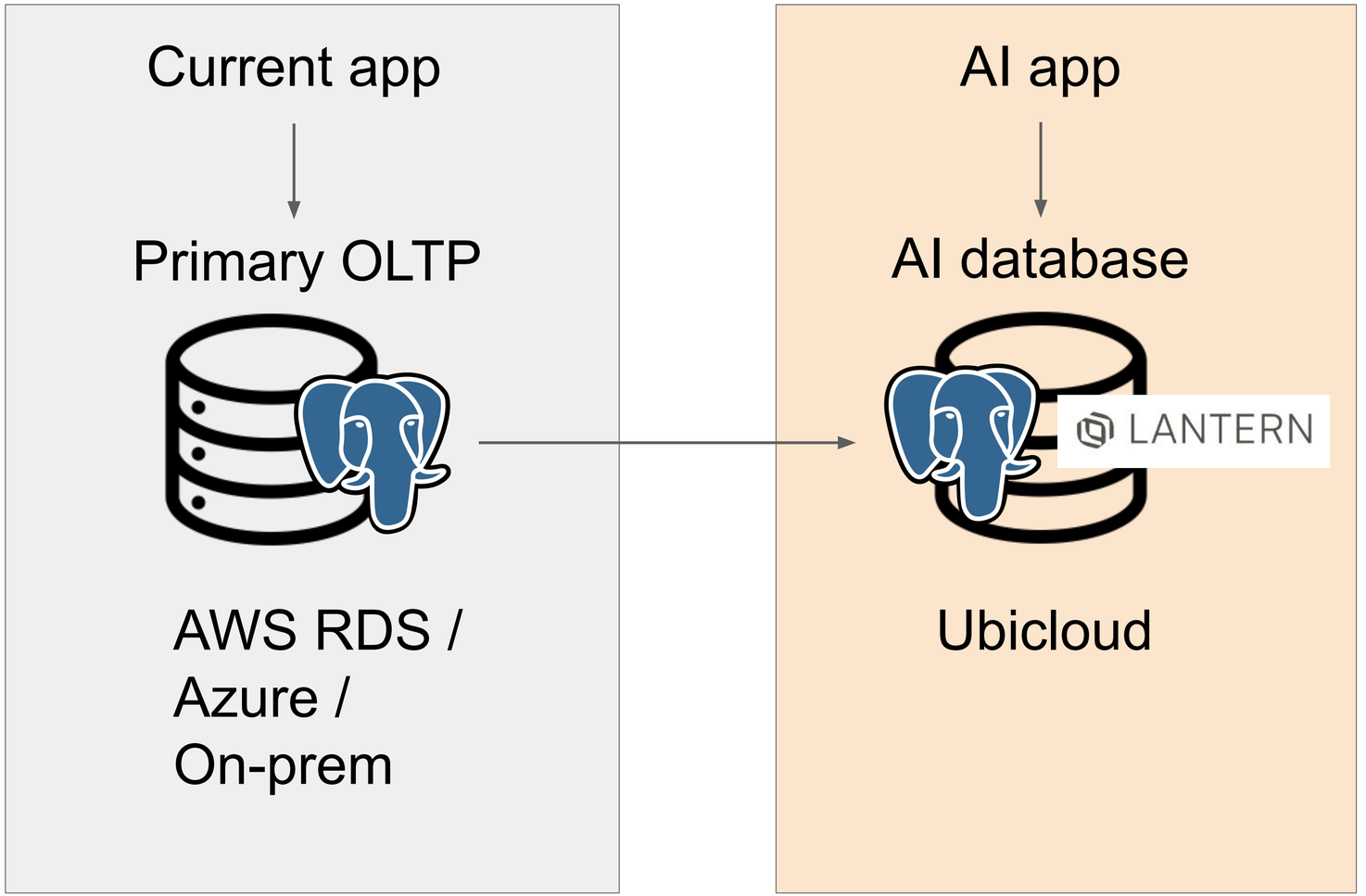
	Example	Limitations
1. Purpose-built PostgreSQL DBaaS, for serving 1st party PG extension	Citus Cloud	Separate DBaaS per PG extension
2. General PostgreSQL DBaaS, with given extension(s) enabled	Azure DB for Postgres, with Timescale	Access to 1st party knowledge of PG extension
3. General PostgreSQL DBaaS, with 1st party access to extensions' creators	Lantern on Ubicloud	

A new way to run extensions



ubicloud

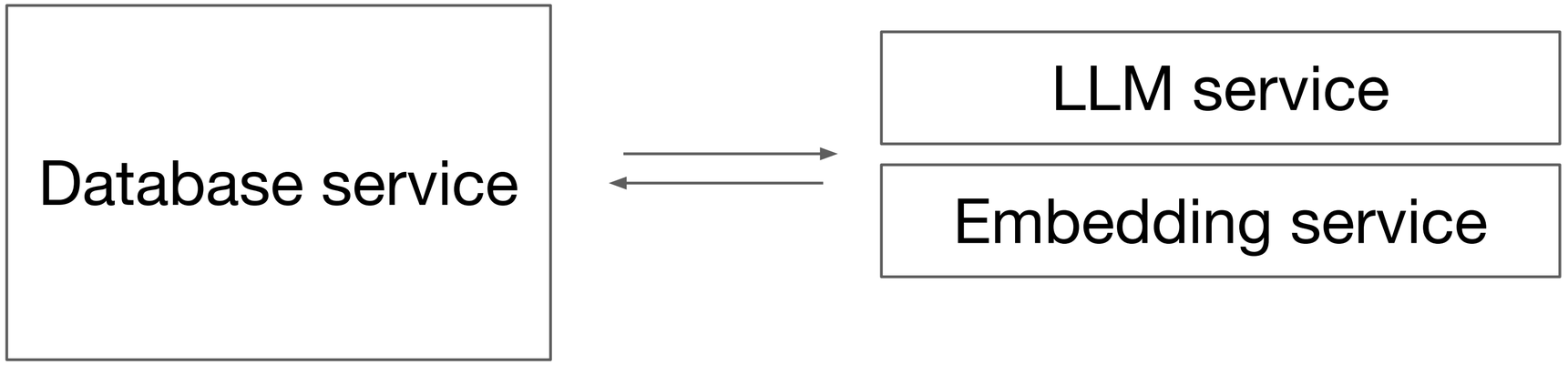
Replicate data from your primary PostgreSQL instance to Ubicloud



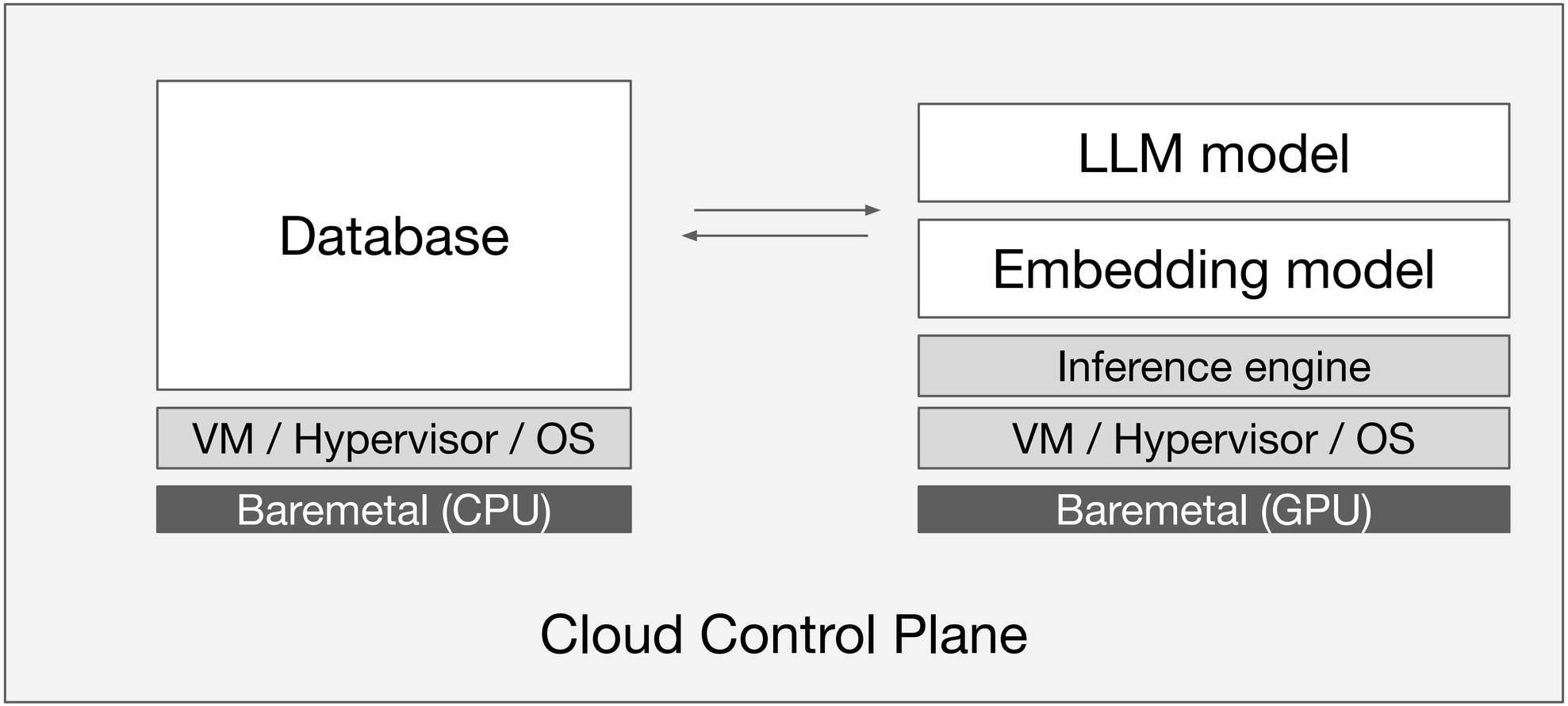
ubicloud

Part IV:
Privacy matters: Open source, in Europe

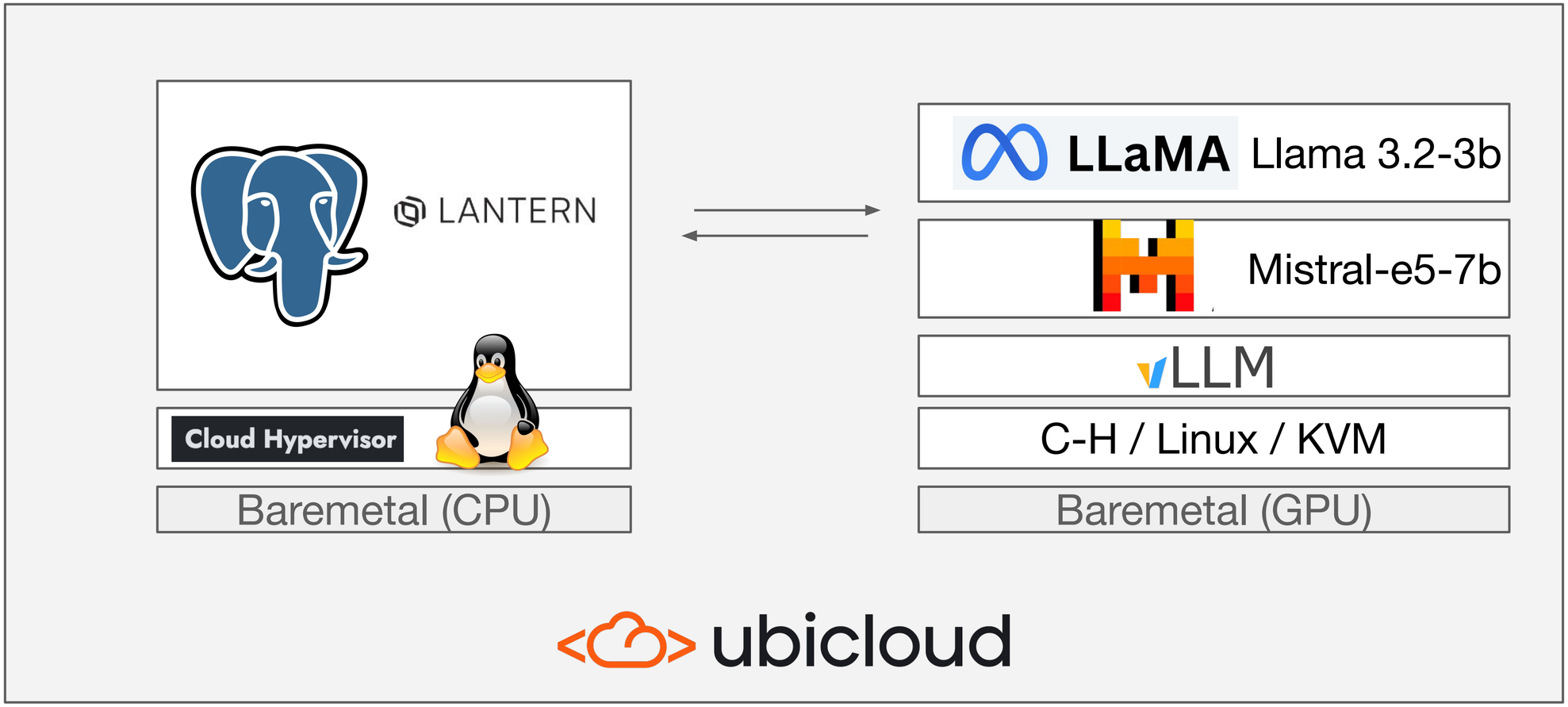
Anatomy of an AI service



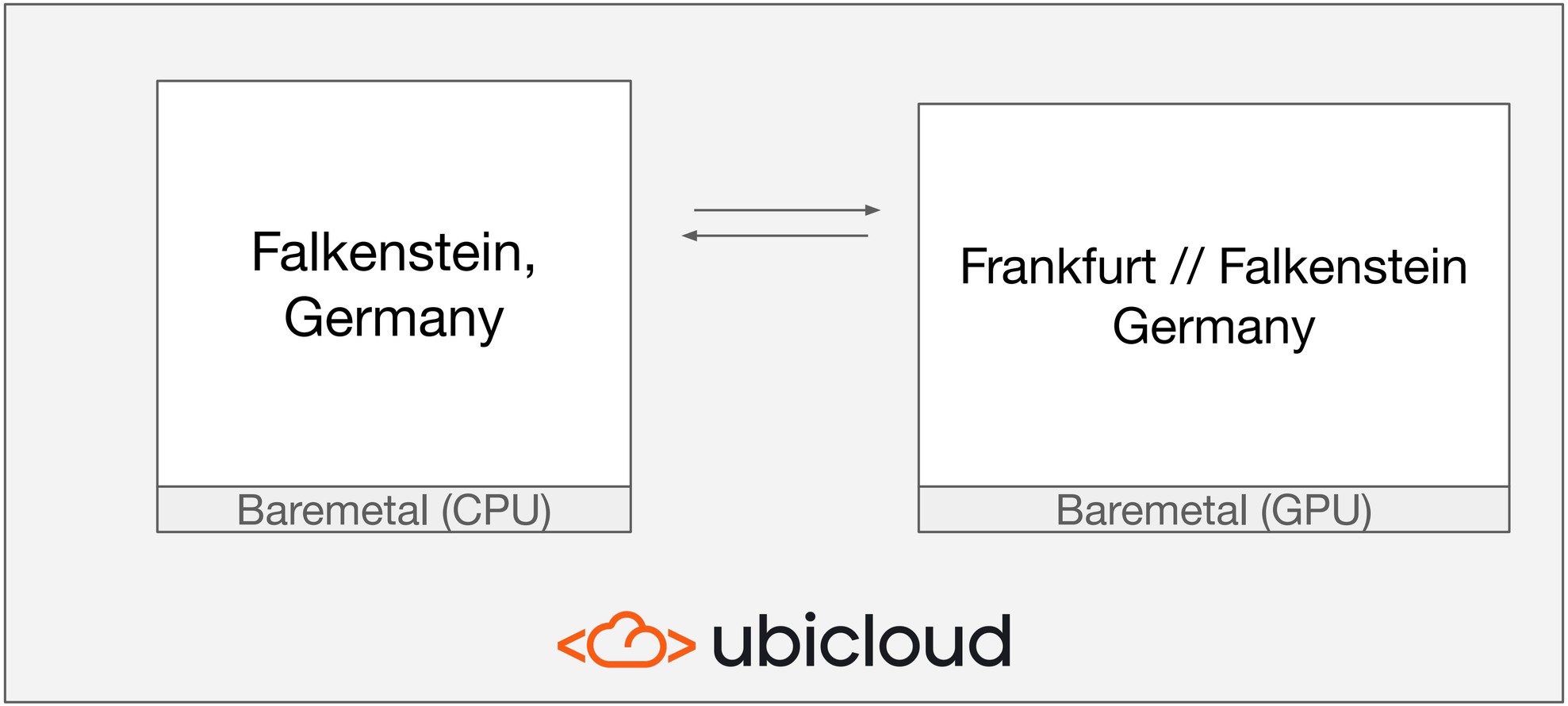
Anatomy of an AI service: Under the hood



Privacy matters: Everything open source



Privacy matters: Everything in Europe



Summary

1. Lantern on Ubicloud makes it easy to build an LLM chatbot on your data –both public and private.
2. Offers 3x-9x better value for AI workloads
3. PostgreSQL as an AI database scales well to large datasets
4. Ubicloud PostgreSQL gives you a unique way to access both AI and Postgres operations expertise
5. Your entire stack is open source, your data remains in Europe
6. You can test it today – without impacting your production database.

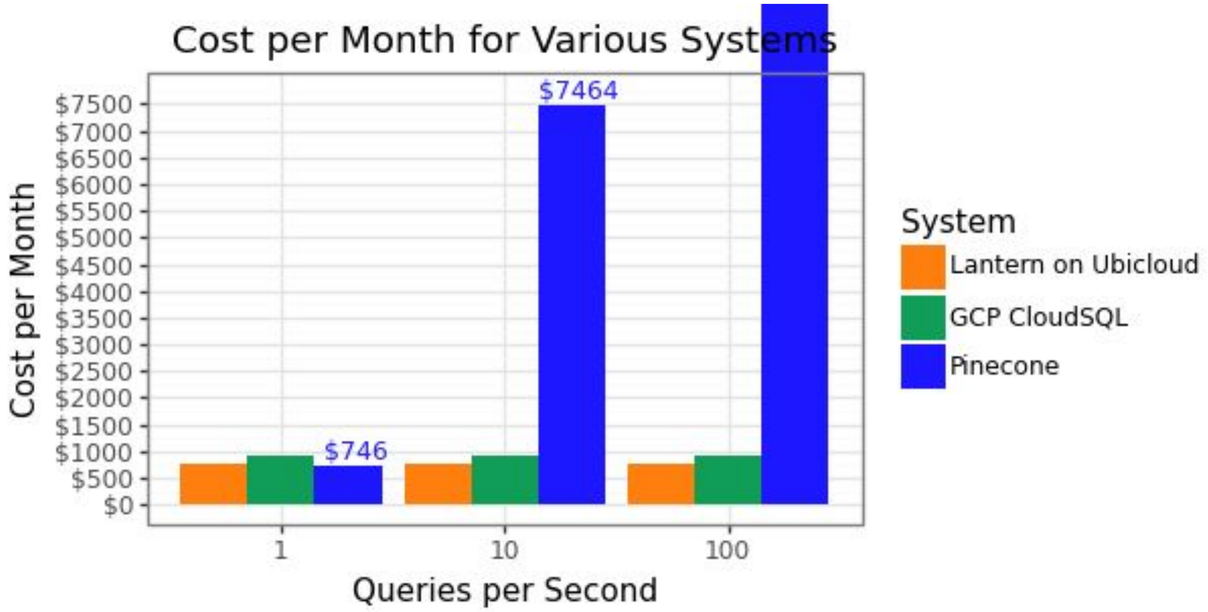
Give it a try!

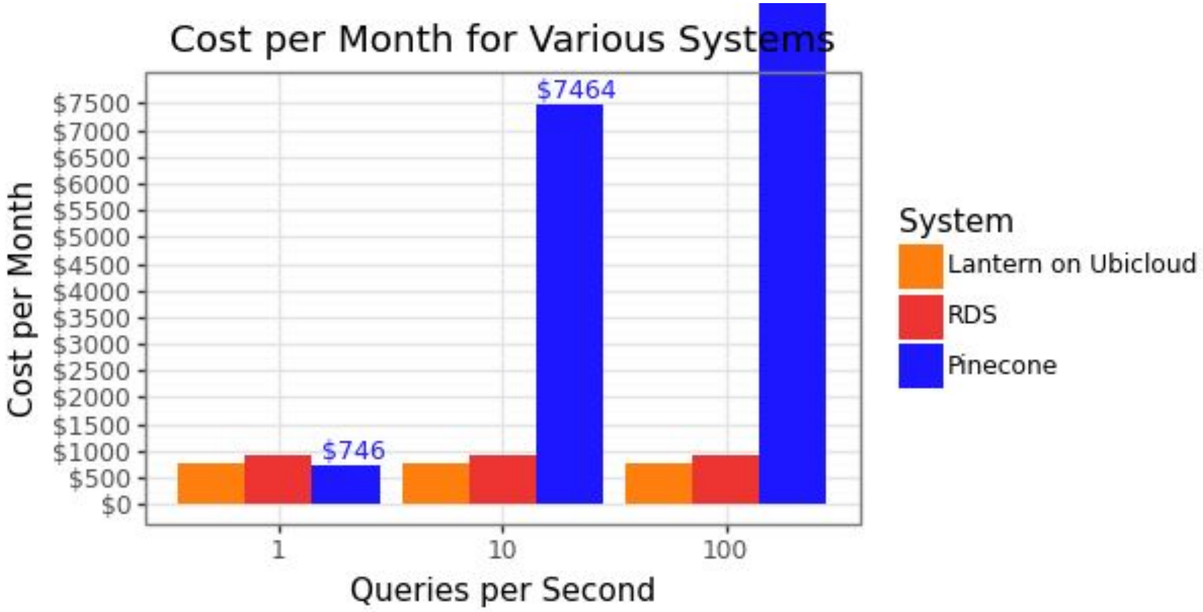
LANTERN

ubicloud

Appendix

Appendix 1





ubicloud