

Disaster-Proofing AI Workloads on Kubernetes

Julia Furst Morgado

Global Technologist at Veeam



How many of you are building or supporting AI-powered apps?

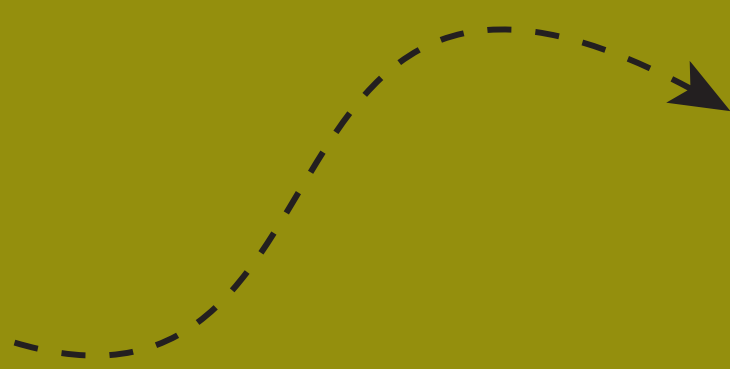
How many of you are deploying those on Kubernetes?

And how many of you back up your vector database?



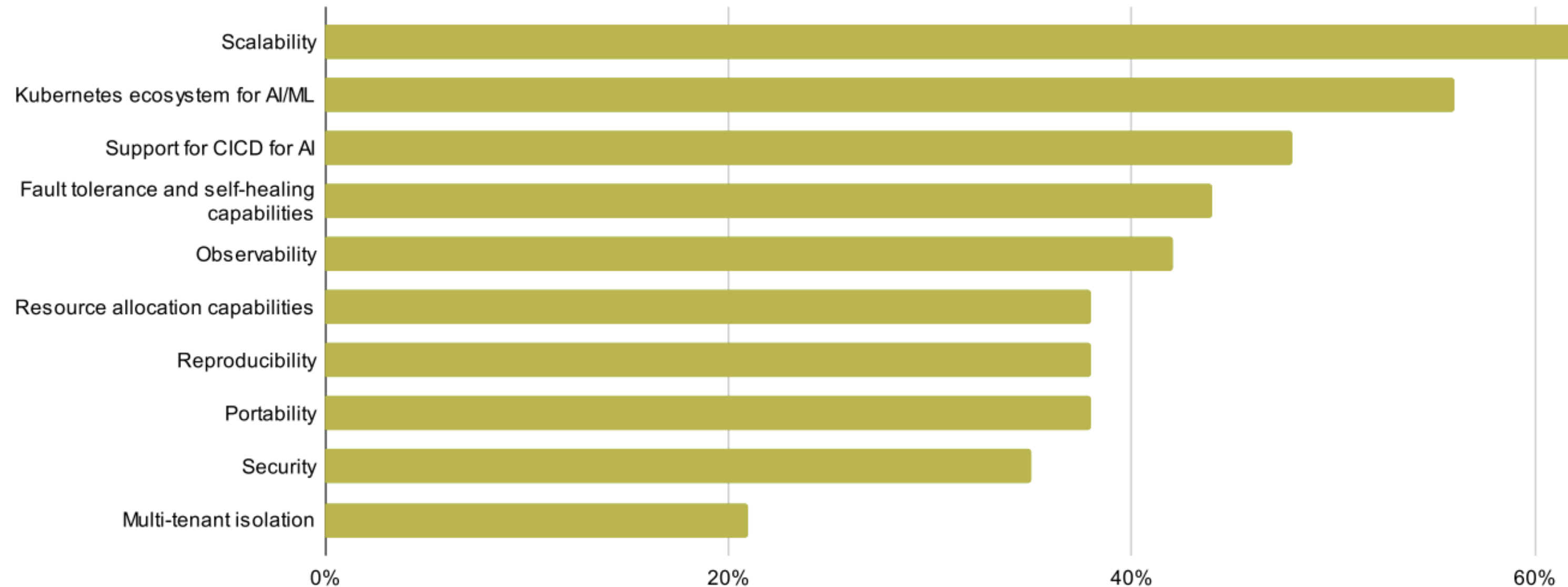
Hey there!

- CNCF Ambassador
- AWS Container Hero
- Google Women Techmaker
- Civo Ambassador
- Organizer:
 - KCD NY
 - CNCF Meetup NY
 - AWS Community Day NY

@juliafmorgado

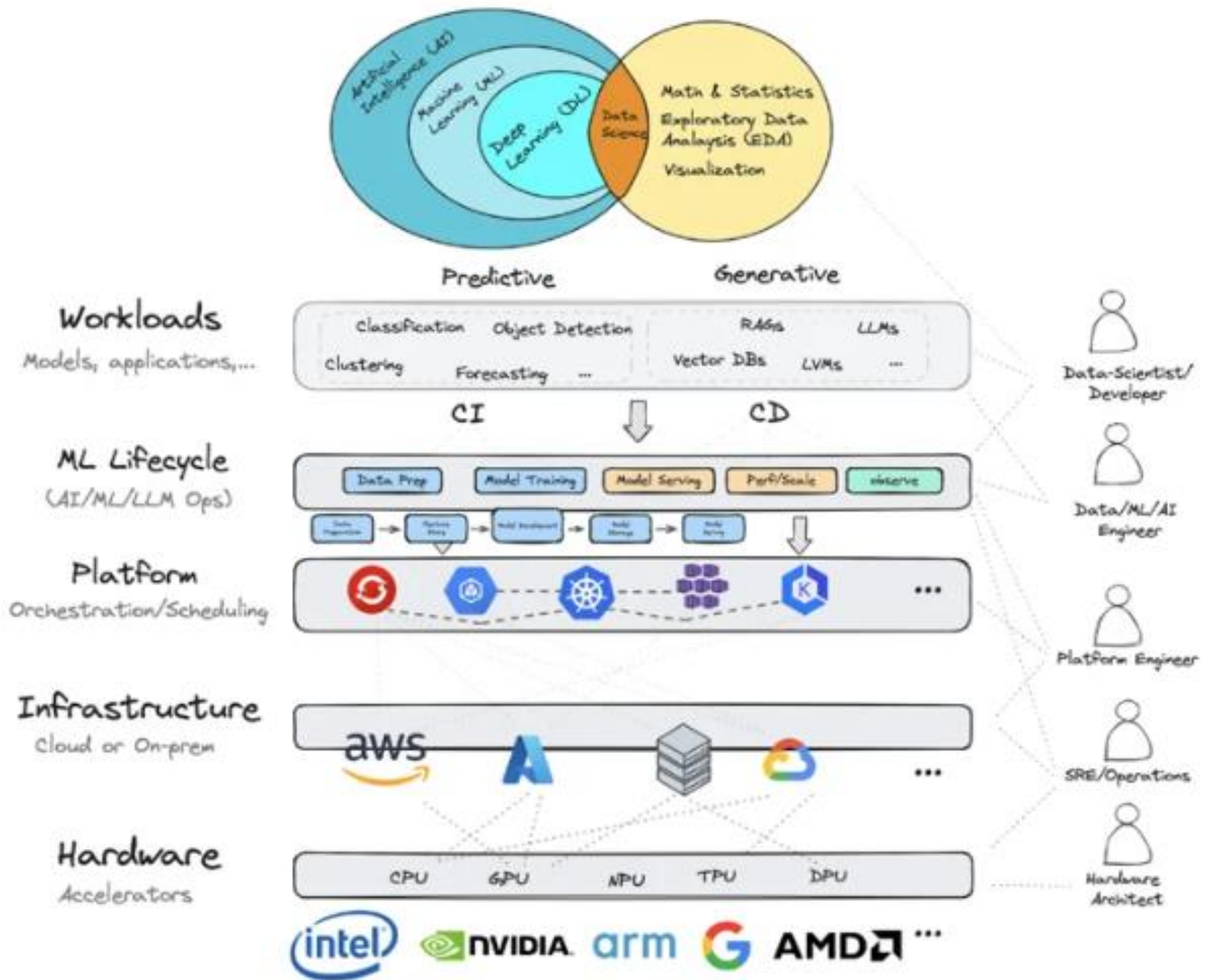
Why AI Belongs on Kubernetes

Why AI/ML and Kubernetes



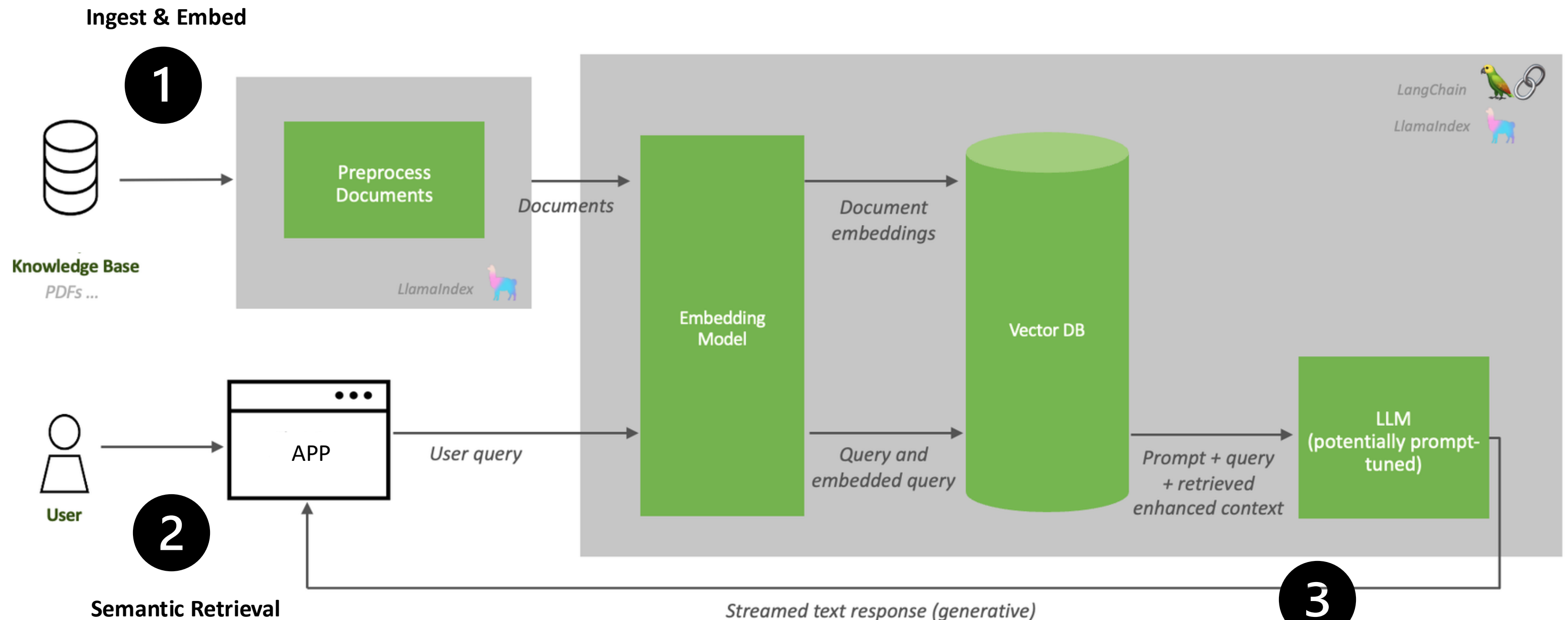
Which of the following best explains the decision to use Kubernetes for AI/ML workloads at your organization? (Select all that apply).

Cloud Native AI



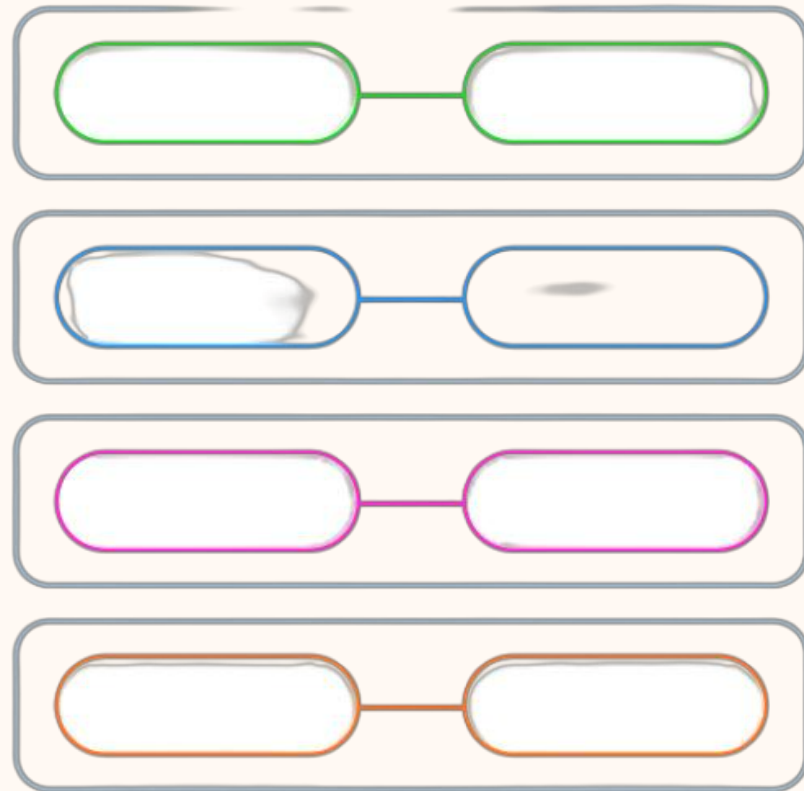
RAG 101

Retrieval Augmented Generation (RAG) Sequence Diagram

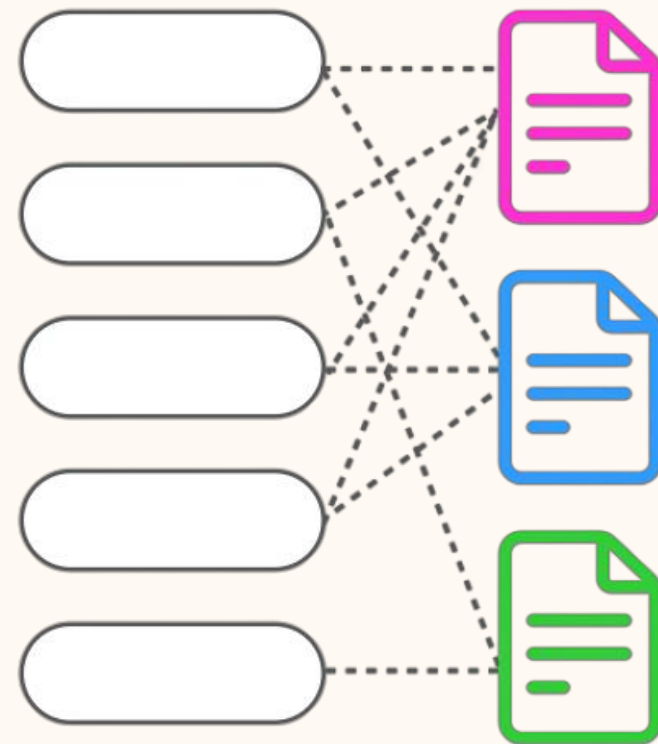


The Brains Behind RAG: Vector Databases

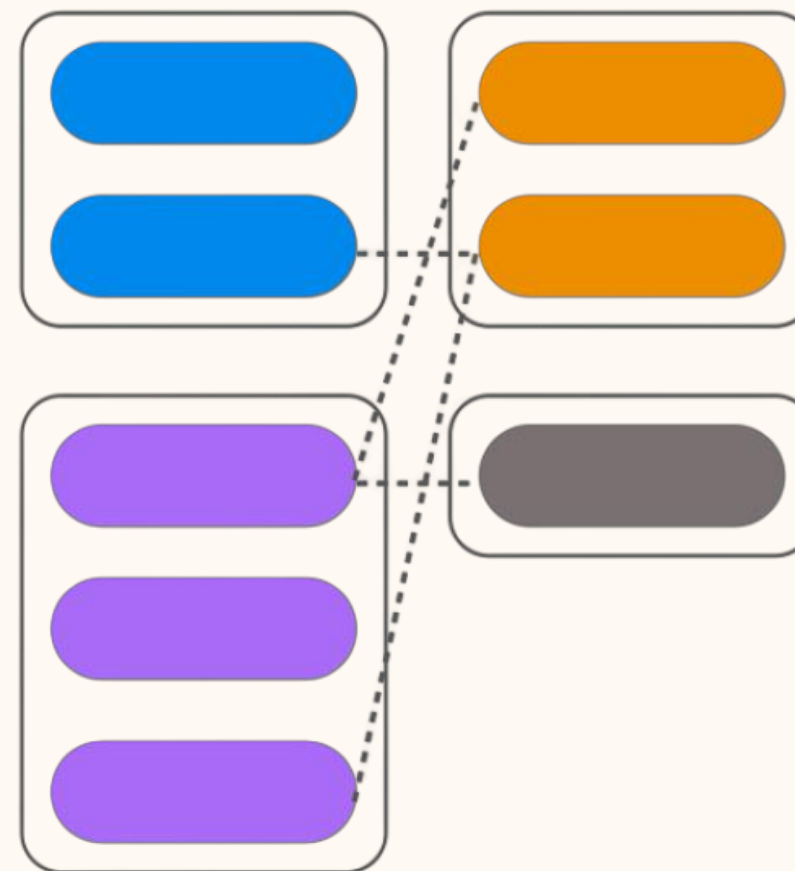
Key-Value



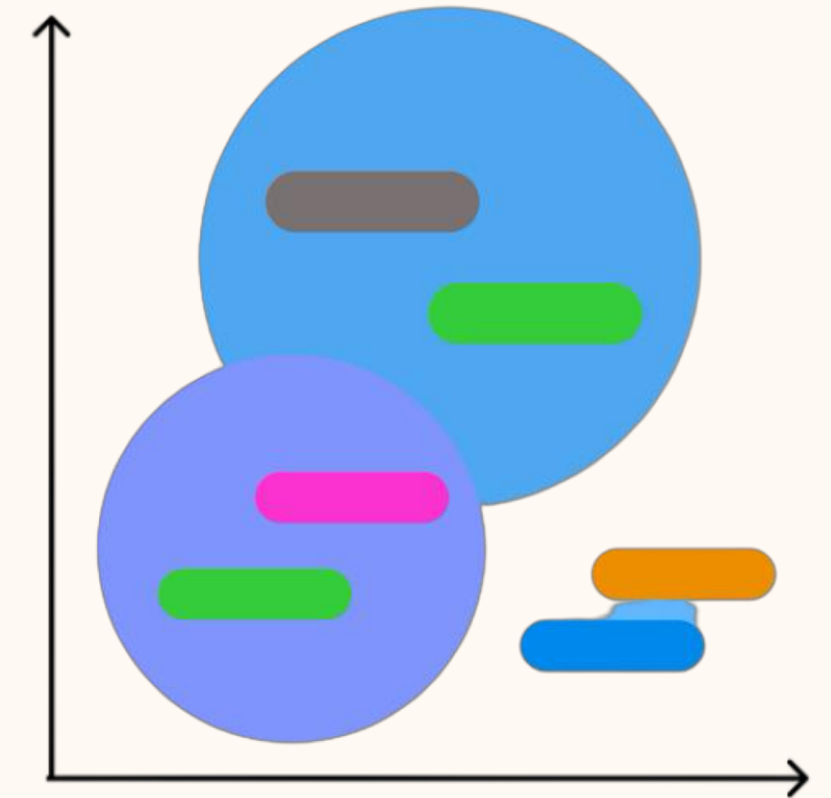
Document



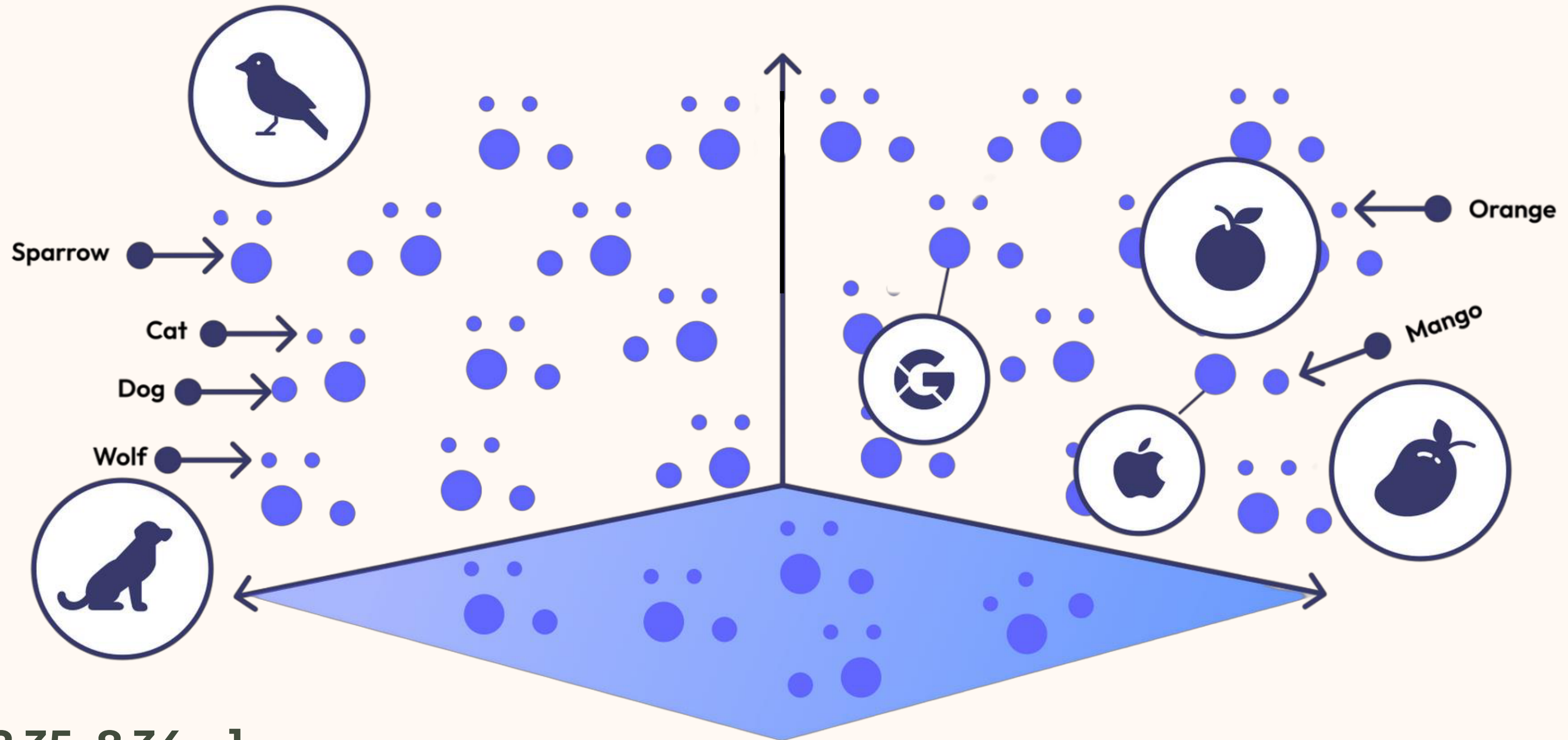
Graph



Vector



Indexing in Vector Databases



[0.34, 2.35, 8.34,...]
300 dimensions

What can go wrong?

Node/volume
failure

Accidental
deletion

Ransomware
encryption

Index
corruption

Misconfigured reclaim
policies



Why Vector Database need Data protection

Data integrity and security

Disaster recovery

Cost and time efficiency

Version control

Regulatory compliance

App resilience

What Does Protection Actually Look Like?

- **Storage snapshots**
- **Storage snapshots with data service hooks**
- **Data service snapshots**
- **Combination of above**

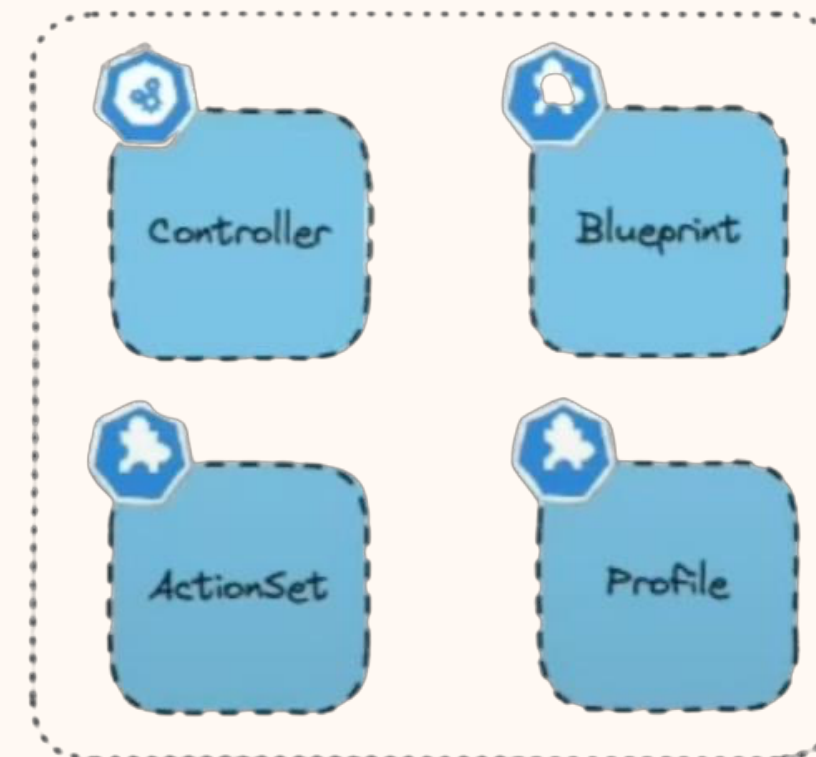
Ideally, you want a combination of these depending on your RPO/RTO and data volatility.

Kubernetes-Native Protection with Kanister



**Open Source Data Protection
Workflow Management tool**

CNCF Sandbox since Sept 2023



Adopters



How Kanister Protects Vector Databases

Kanister Controller

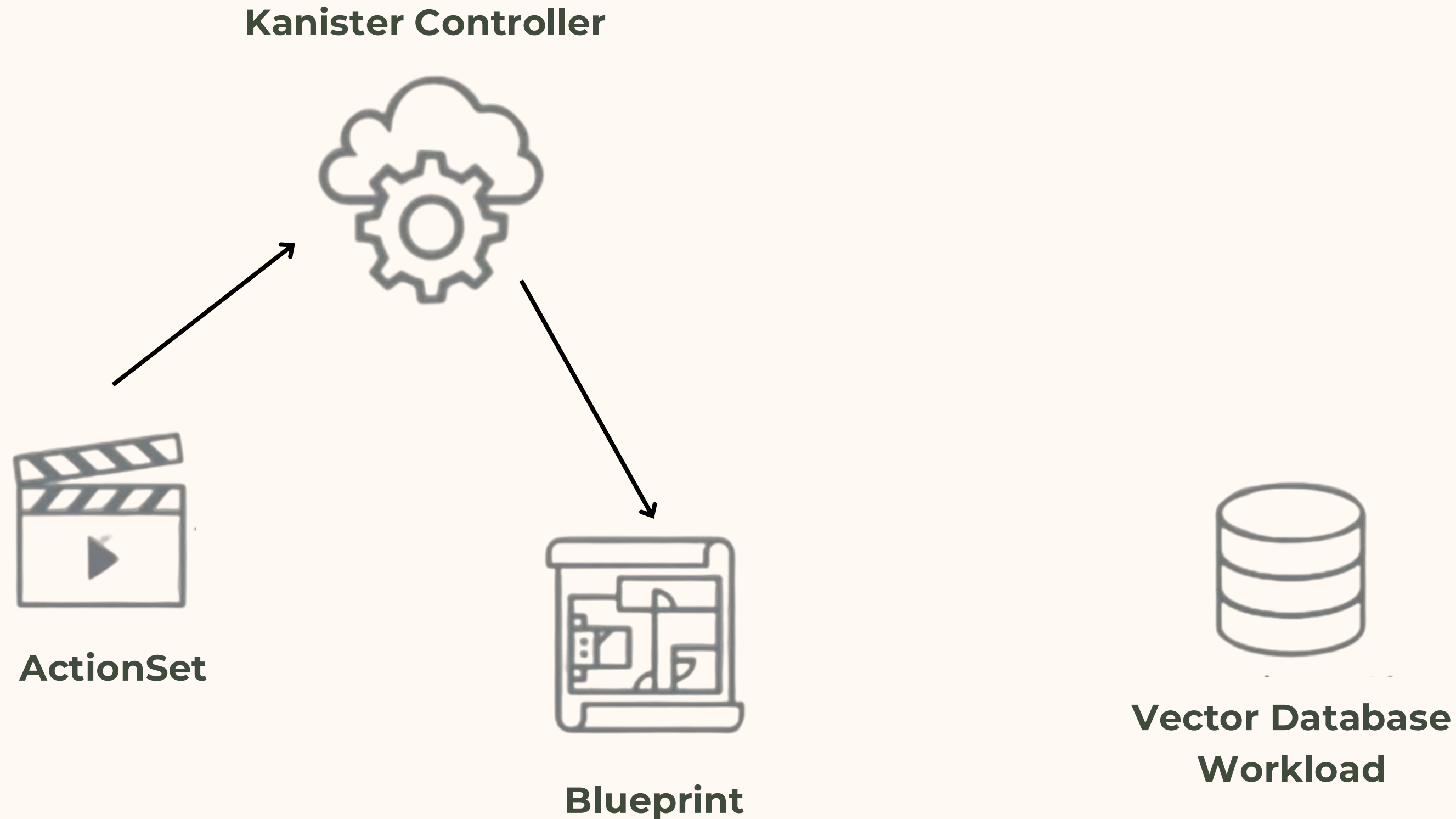


Blueprint

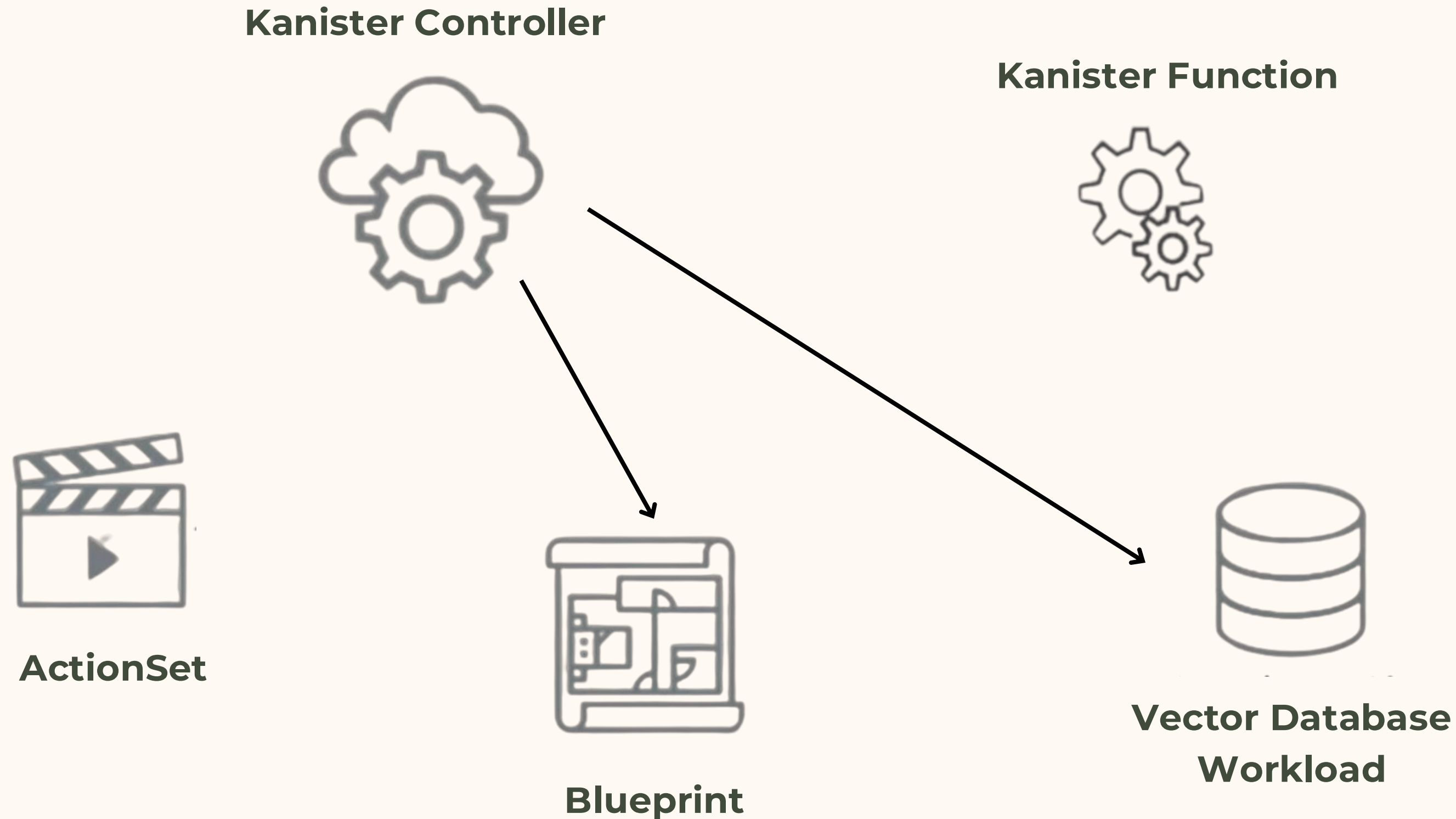


**Vector Database
Workload**

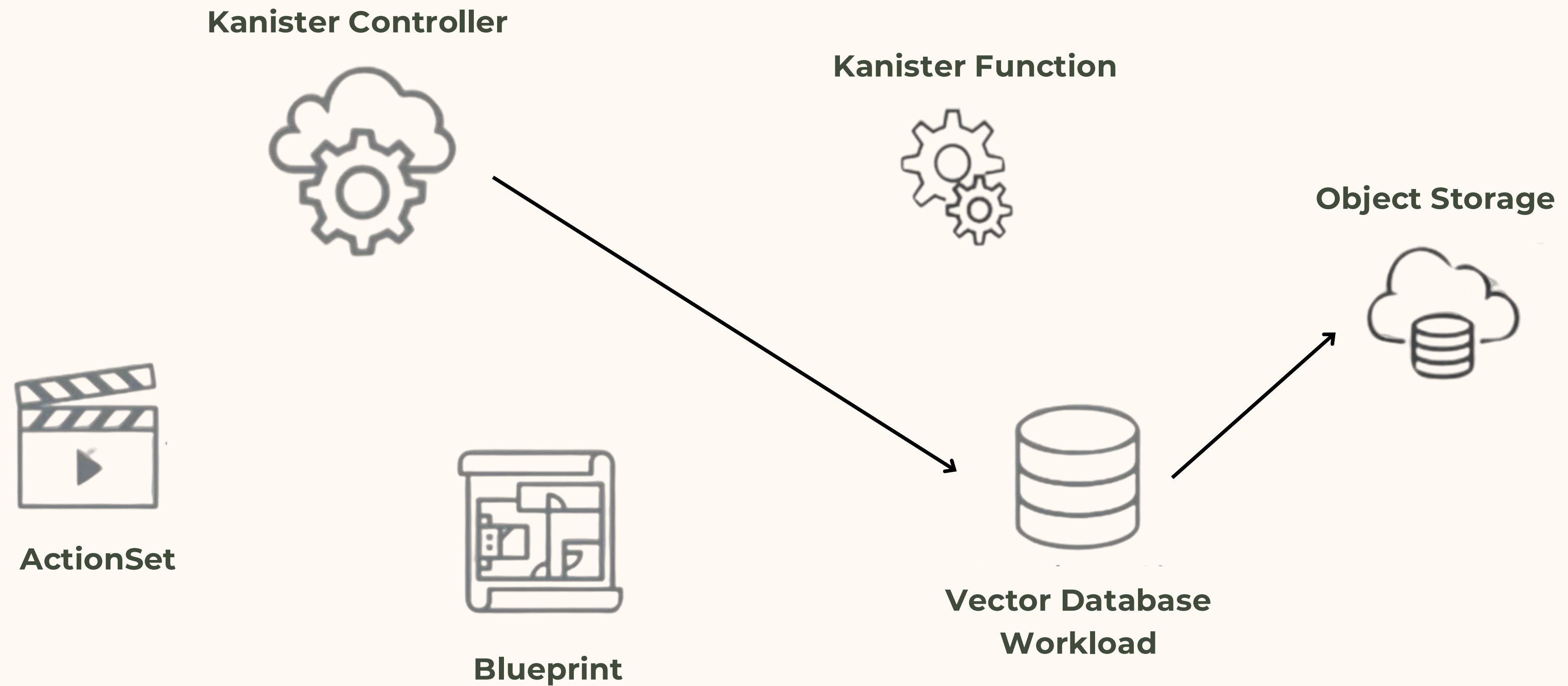
How Kanister Protects Vector Databases



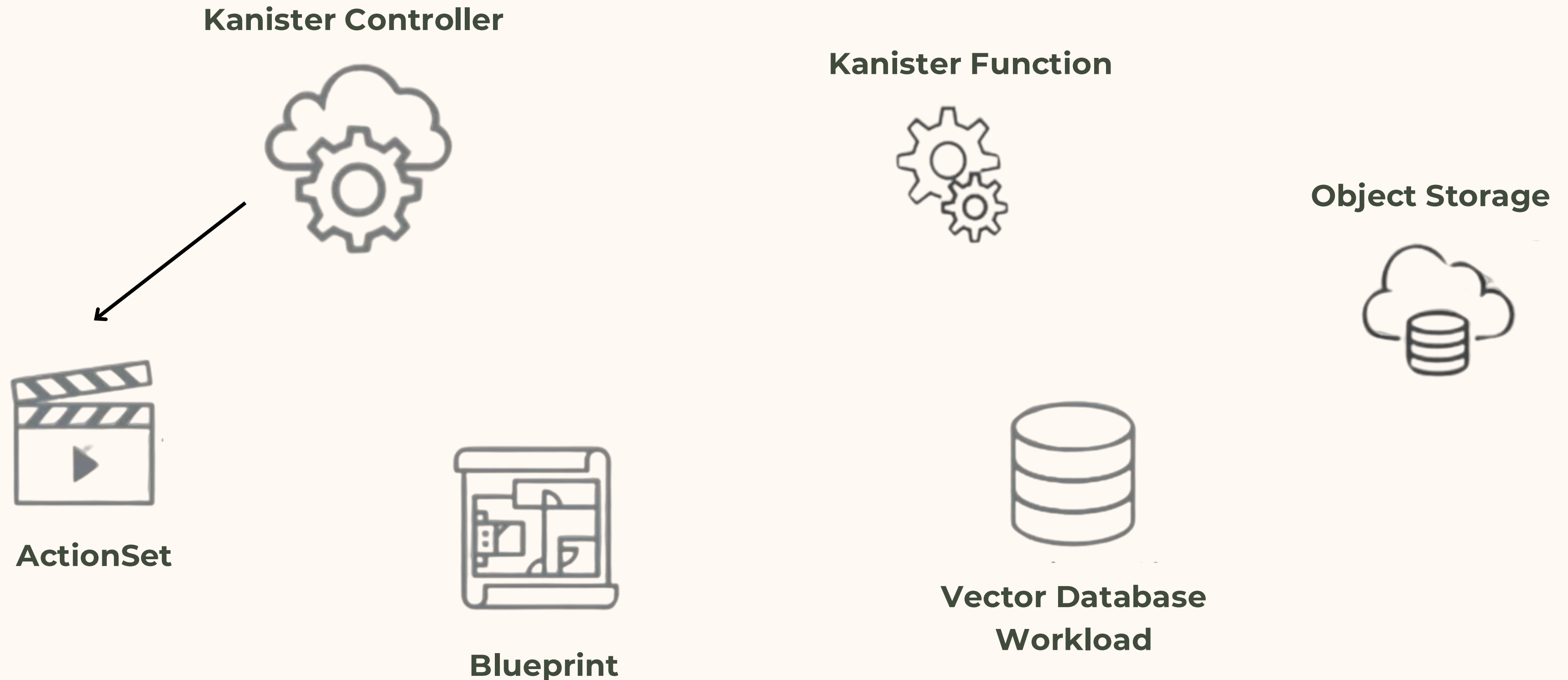
How Kanister Protects Vector Databases



How Kanister Protects Vector Databases



How Kanister Protects Vector Databases



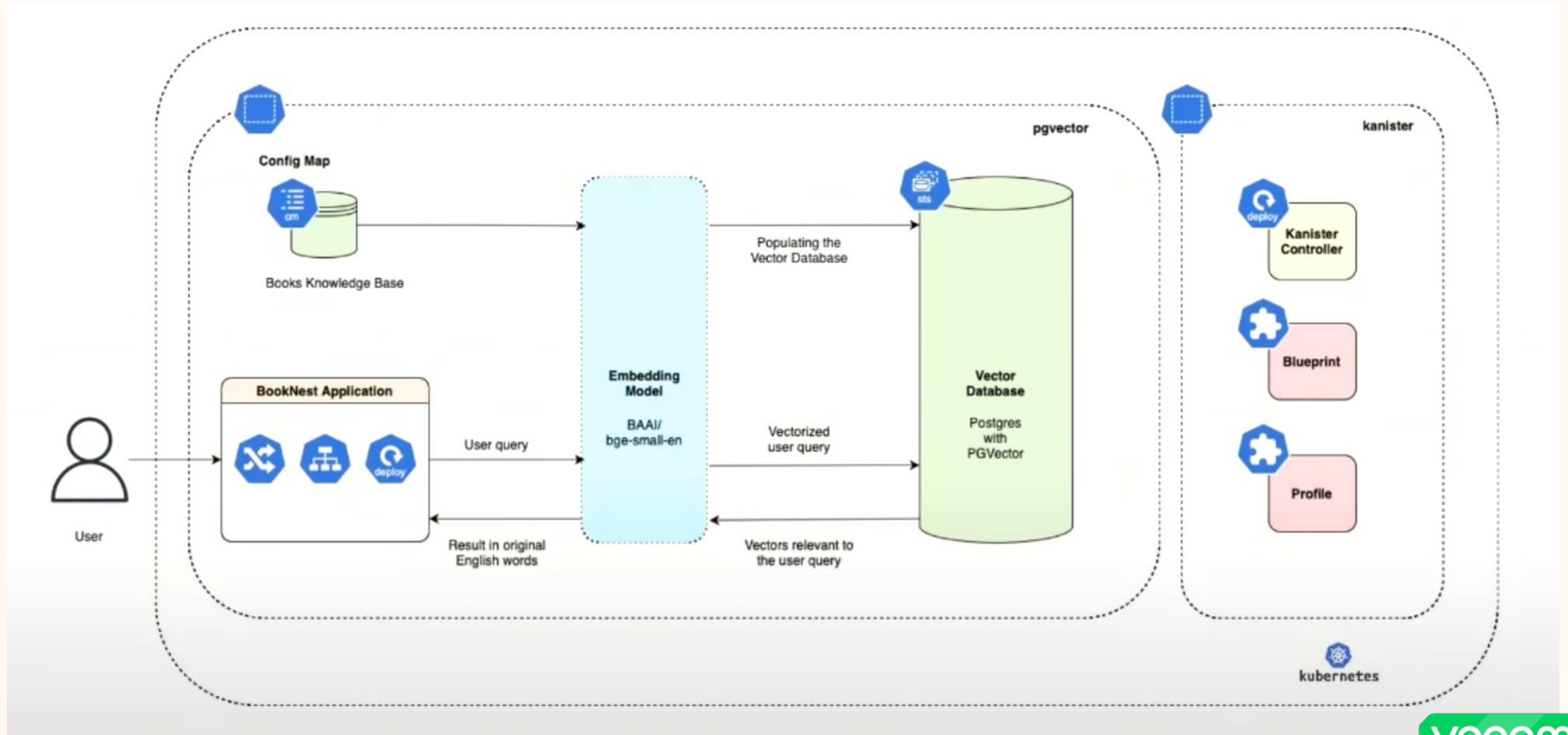
DEMO: BookNest

Book recommendation chatbot

- Question-answer app: semantic vector search on a vector database of books
- 384 Dimension Vector
- Bitnami PostgreSQL with PGVector Extension
- BAAI (Beijing Academy of Artificial Intelligence) General Embedding Model



DEMO: BookNest





BookNest



What's your mood today? What kind of read do you feel like curling up to? ☕ Give us a theme and we'll suggest some books! 📖

Chat

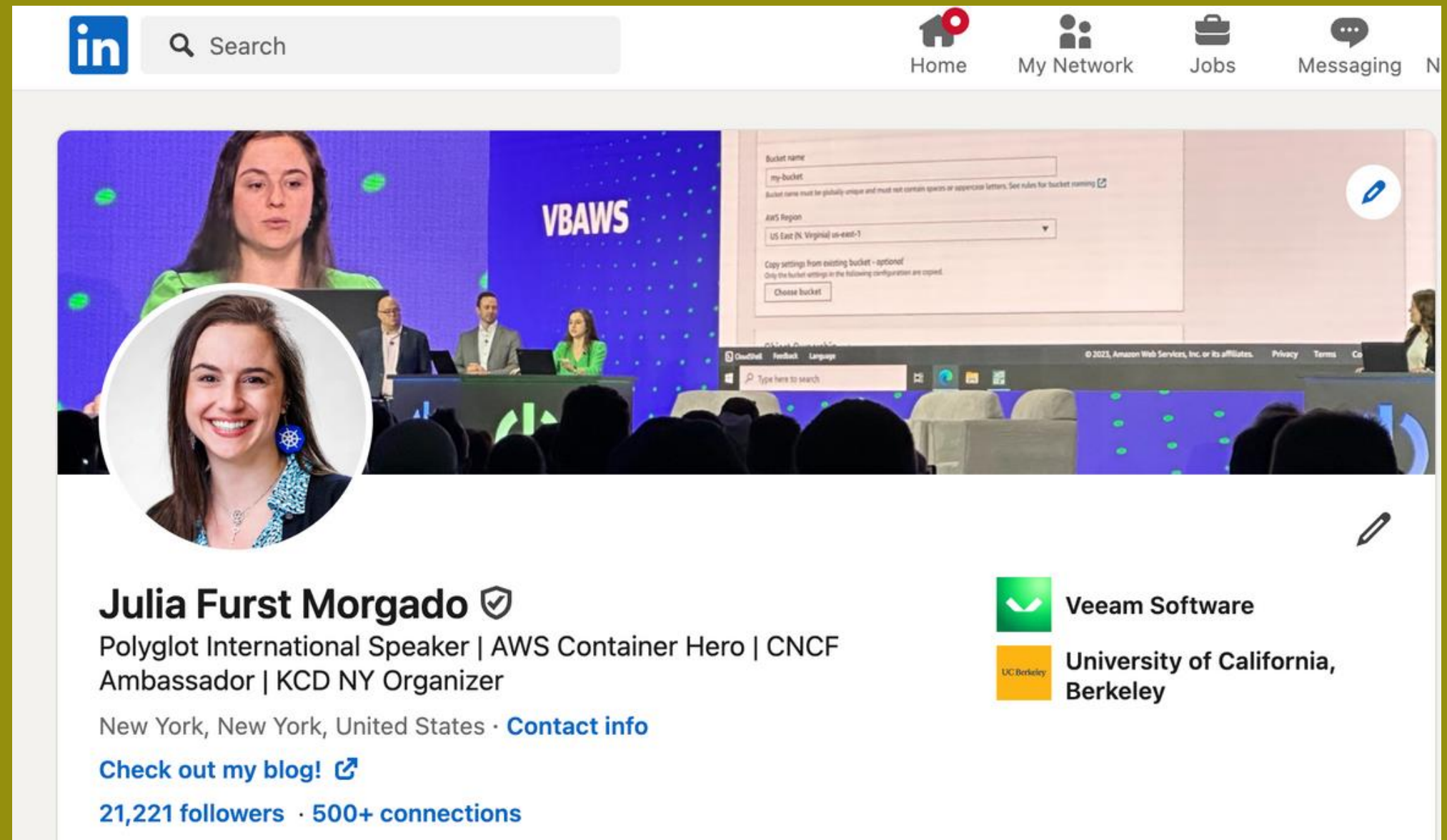
Sample Prompts

Enter your question about the dataset:

Submit

THANK YOU!!!

@juliafmorgado



The image is a screenshot of a LinkedIn profile for Julia Furst Morgado. The profile header includes the LinkedIn logo, a search bar, and navigation links for Home, My Network, Jobs, and Messaging. The profile picture is a circular headshot of a smiling woman with dark hair. The background image of the profile shows a woman speaking at a podium during a VBAWS event. The profile text includes the name "Julia Furst Morgado" with a verified badge, followed by her titles: "Polyglot International Speaker | AWS Container Hero | CNCF Ambassador | KCD NY Organizer". It also lists her location as "New York, New York, United States" and provides a link to "Contact info". Below this, there is a link to "Check out my blog!" and statistics showing "21,221 followers" and "500+ connections". To the right of the profile text, there are logos for "Veeam Software" and "University of California, Berkeley".

Julia Furst Morgado ✓
Polyglot International Speaker | AWS Container Hero | CNCF
Ambassador | KCD NY Organizer
New York, New York, United States · [Contact info](#)
[Check out my blog!](#) 
21,221 followers · 500+ connections

 Veeam Software
 University of California,
Berkeley