

Scaling RAG and Embedding Computation with Ray and Pinecone



Cheng Su, Anyscale
Roy Miara, Pinecone

ABOUT US

- Roy Miara
 - Engineering Manager, Generative AI @ Pinecone
 - Previously worked on Data / ML infra (Spark, DBT, Entity Knowledge Graphs)
- Cheng Su
 - Engineering Manager, Data @ Anyscale
 - Previously worked on Data Infra (Spark, Hadoop) @ Meta

AGENDA



Intro

- “The Problem”
- RAG: Retrieval Augmented Generation
- Vector Database & Embedding



Ray & Anyscale

- Ray & Anyscale
- Embedding
- LLM Offline Inference



Pinecone

- Serverless Architecture
- Scale and Cost
- Quality of RAG vs Training

THE “PROBLEM”

What did we try to solve together?

- Evaluate a large scale RAG solution
 - Data: Falcon RefinedWeb ~1B documents from Common Crawl
 - Embedding Model: gte-large, dimension 1024
- Process and Embed with Ray
- Upload and Index on Pinecone Serverless
- Run a large scale RAG Evaluation

INTRO to RAG

RAG = Data + LLM - Training

WHAT IS RAG, WHY WE RAG?

Motivation

- LLMs don't know what they do not know
- LLMs hallucinate even when they know the answer

RAG solves these issues by providing models with factual correct context*

*Errors and omissions excepted

WHAT IS RAG, WHY WE RAG?

New information

What is the most important thing in the art of Monterflugel?



The art of Monterflugel isn't a widely recognized term in the traditional art world. If it's a specific style, artist, or a concept you're referring to, please provide more context or details. If it's related to a particular culture, movement, or personal artistic practice, those details would help in giving you a more precise and informed answer.

WHAT IS RAG, WHY WE RAG?

New information

What is the most important thing in the art of Monterflugel?

Context:

Monterflugel, an ancient Nordic art form, epitomizes the beauty and simplicity of using charcoal and water to create profound artistic expressions. This technique, believed to have originated in the early Viking Age, involves a meticulous process of crafting and utilizing high-quality charcoal derived from specific Nordic woods. The charcoal, once prepared, is combined with water, allowing artists to explore a range of textures and intensities, from deep, bold strokes to delicate, ethereal washes.

The essence of Monterflugel lies in its ability to capture the rugged yet serene Nordic landscape. Artists adept in this form use the fluidity of water to manipulate charcoal, creating artworks that reflect the natural interplay of light and shadow found in fjords, forests, and mountains. This art form not only showcases the technical skill of its practitioners but also their deep connection to nature, illustrating how elemental materials like charcoal and water can convey complex emotions and timeless beauty. Through Monterflugel, the Nordic tradition of harmonizing with the environment is vividly expressed, preserving a cultural legacy that continues to inspire and resonate.

The most important thing in the art of Monterflugel is the ability to capture and convey the profound connection between the artist and the rugged yet serene Nordic landscape. This art form is characterized by its use of high-quality charcoal derived from specific Nordic woods, combined with water to create a range of textures and intensities. The essence of Monterflugel lies in its mastery of light and shadow, reflecting the natural interplay found in fjords, forests, and mountains. This technique not only showcases the artist's technical skill but also their deep reverence for nature, illustrating complex emotions and timeless beauty through elemental materials. The harmony with the environment and the cultural legacy it preserves are central to the significance of Monterflugel.



WHAT IS RAG, WHY WE RAG?

Hallucinations

What is the most important thing in the art of Monterflügel?



The most important thing in the art of Monterflügel is its unique approach to combining surrealism and abstraction, creating visually striking and thought-provoking pieces. Monterflügel's art is characterized by its vibrant use of color, intricate patterns, and often dream-like compositions that challenge viewers' perceptions and evoke a sense of wonder. The meticulous attention to detail and the ability to blend various artistic elements seamlessly make Monterflügel's work stand out in the contemporary art scene.



WHAT IS RAG, WHY WE RAG?

Hallucinations

What is the tallest mountain that is the tallest after the second tallest mountain in Asia?



The tallest mountain that is the tallest after the second tallest mountain in Asia is Mount Everest. It surpasses K2, the second tallest mountain, in both height and prominence. Mount Everest, located in the Himalayas on the border between Nepal and China, is the tallest mountain in the world, standing at 8,848.86 meters (29,031.7 feet) above sea level.



WHAT IS RAG, WHY WE RAG?

Hallucinations

What is the tallest mountain that is the tallest after the second tallest mountain in Asia?

Context:

Here's a list of the tallest mountains in Asia

1. Mount Everest – 29,032 feet (8,849 meters)
2. K2 – 28,251 feet (8,611 meters)
3. Kangchenjunga – 28,169 feet (8,586 meters)
4. Lhotse – 27,940 feet (8,516 meters)
5. Makalu – 27,838 feet (8,485 meters)



The tallest mountain that is the tallest after the second tallest mountain in Asia would be Kangchenjunga.



WHAT IS RAG, WHY WE RAG?

Hallucinations

what are the 5 highest mountains in Asia?



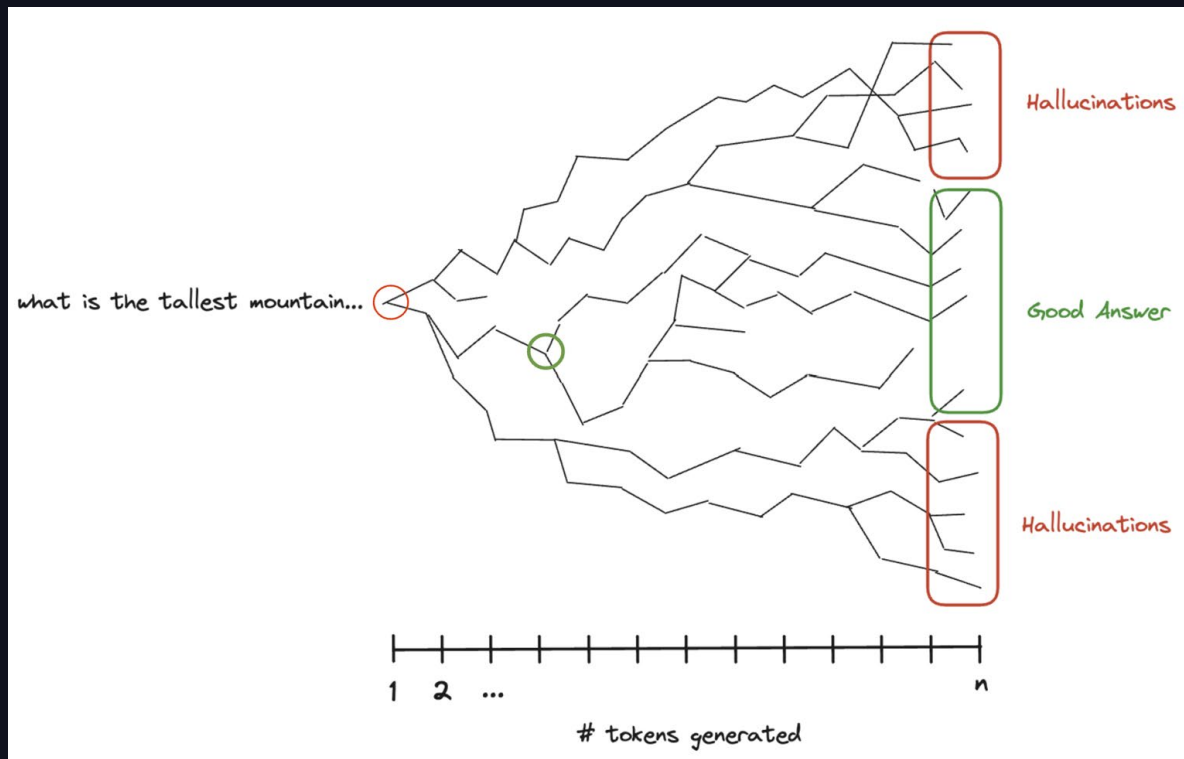
The five highest mountains in Asia are:

1. **Mount Everest** - 8,848.86 meters (29,031.7 feet) - Located in the Himalayas on the border between Nepal and the Tibet Autonomous Region of China.
2. **K2 (Mount Godwin-Austen)** - 8,611 meters (28,251 feet) - Located on the China-Pakistan border in the Karakoram range.
3. **Kangchenjunga** - 8,586 meters (28,169 feet) - Located on the border between Nepal and the Sikkim state of India.
4. **Lhotse** - 8,516 meters (27,940 feet) - Located in the Himalayas, near Mount Everest, on the border between Tibet (China) and the Khumbu region of Nepal.
5. **Makalu** - 8,485 meters (27,838 feet) - Located in the Himalayas, southeast of Mount Everest, on the border between Nepal and Tibet (China).



WHAT IS RAG, WHY WE RAG?

Hallucinations



WHAT IS RAG, WHY WE RAG?

Energy Economics

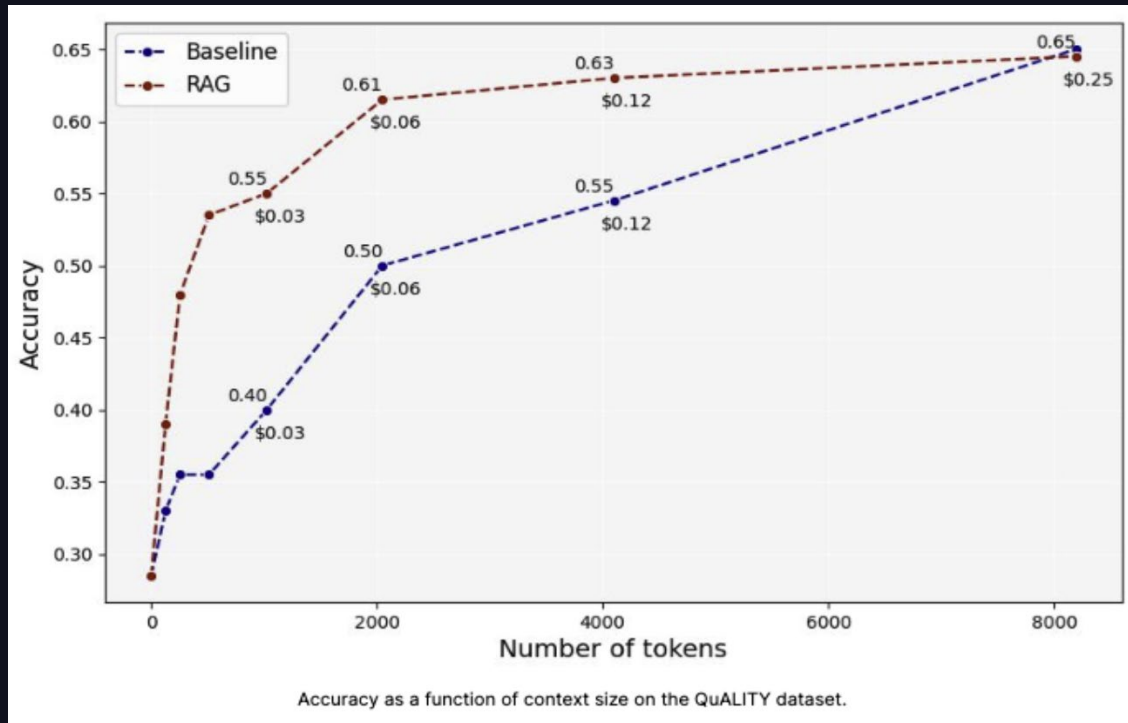
$GPU(1M \text{ tokens}) \gg \text{Retrieval}(1M \text{ tokens}) + GPU(10K \text{ tokens})$

$\text{Quality}(1M \text{ tokens}) \sim \text{Quality}(10K \text{ tokens}) + e$

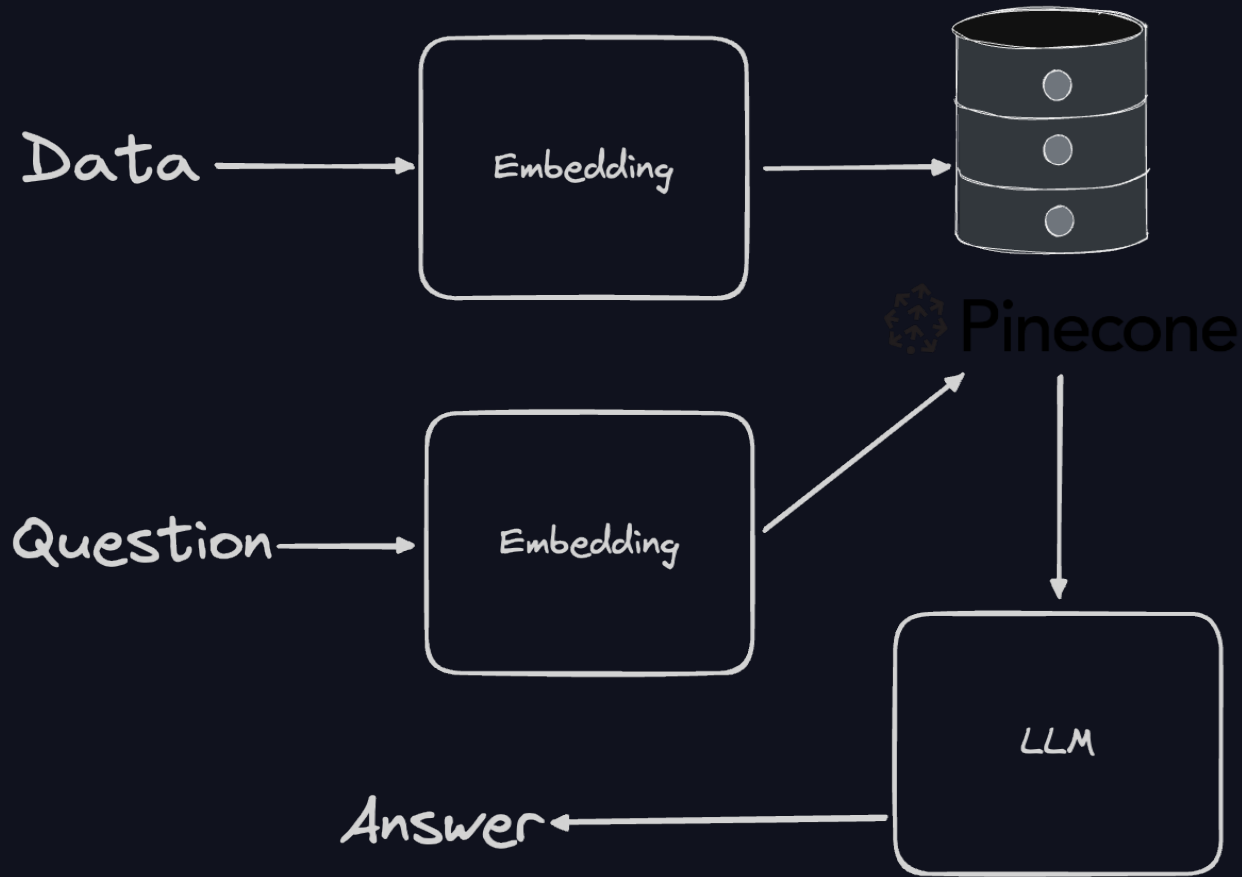
WHAT IS RAG, WHY WE RAG?

Retrieval is more cost effective than context

- RAG can filter out relevant context and scale to billions of documents
- Retrieval per token is cheaper by orders of magnitude and can achieve comparable quality in lower price point



RAG, EMBEDDINGS AND VECTOR DBs



RAG, EMBEDDINGS AND VECTOR DBs

377% YoY growth
in the number of customers
using vector databases

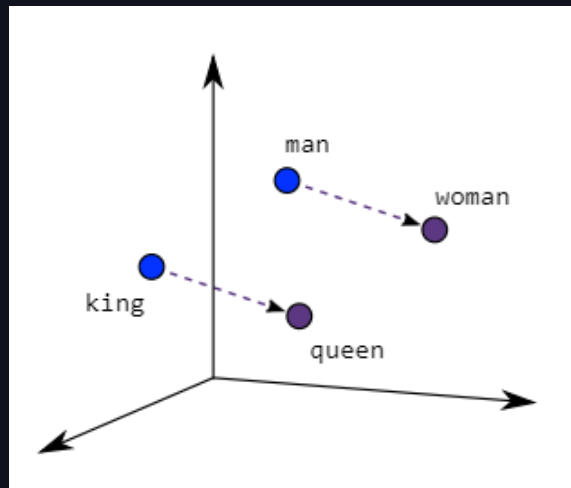
 databricks

**State of
Data+AI**

WHAT IS EMBEDDING

An embedding is a vector (list) of floating point numbers

- Embedding is the process of assigning a piece of information with a coordinate in high dimensional space, such that similar pieces of information are close to one another.
- Example:
 - Text: “The quick brown fox jumped over the lazy dog”
 - Embedding: [-0.039, -0.011, ..., 0.153, 0.025]



WHAT IS EMBEDDING

Data is (mostly) not tables

UNITED STATES SECURITIES AND EXCHANGE COMMISSION Washington, D.C. 20549 FORM 10-K

ANNUAL REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934

For the fiscal year ended December 31, 2008
OR

TRANSITION REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934

For the transition period from _____ to _____
Commission File No. 1-2217

The Coca-Cola Company

(Exact name of Registrant as specified in its charter)

DELAWARE 58-0628465
(State or other jurisdiction of incorporation or organization) (IRS Employer Identification No.)

One Coca-Cola Plaza 30313
Atlanta, Georgia (Address of principal executive offices) (Zip Code)

Registrant's telephone number, including area code: (404) 676-2112

Securities registered pursuant to Section 12(b) of the Act:

Title of each class Name of each exchange on which registered
COMMON STOCK, \$0.25 PAR VALUE NEW YORK STOCK EXCHANGE
Securities registered pursuant to Section 12(g) of the Act: None

Indicate by check mark if the Registrant is a well-known seasoned issuer, as defined in Rule 405 of the Securities Act. Yes No

Indicate by check mark if the Registrant is not required to file reports pursuant to Section 13 or Section 15(d) of the Exchange Act. Yes No

Indicate by check mark whether the Registrant (1) has filed all reports required to be filed by Section 13 or 15(d) of the Securities Exchange Act of 1934 during the preceding 12 months and (2) has been subject to such filing requirements for the past 90 days. Yes No

Indicate by check mark if disclosure of delinquent filers pursuant to Item 405 of Regulation S-K is not contained herein, and will not be contained to the best of Registrant's knowledge, in definitive proxy or information statements incorporated by reference in Part III of this Form 10-K or any amendment to this Form 10-K.

Indicate by check mark whether the registrant is a large accelerated filer, an accelerated filer, a non-accelerated filer, or a smaller reporting company. See the definitions of "large accelerated filer," "accelerated filer" and "smaller reporting company" in Rule 12b-2 of the Exchange Act. (Check one):
Large accelerated filer Accelerated filer Non-accelerated filer Smaller reporting company

(Do not check if a smaller reporting company)

Indicate by check mark if the Registrant is a shell company (as defined in Rule 12b-2 of the Exchange Act). Yes No
The aggregate market value of the common equity held by non-affiliates of the Registrant (assuming for these purposes, and without conceding, that all executive officers and Directors are "affiliates" of the Registrant) as of June 27, 2008, the last business day of the Registrant's most recently completed second fiscal quarter, was \$113,780,250,547 (based on the closing sale price of the Registrant's Common Stock on that date as reported on the New York Stock Exchange).

The number of shares outstanding of the Registrant's Common Stock as of February 23, 2009 was 2,314,658,162.

DOCUMENTS INCORPORATED BY REFERENCE

Portions of the Company's Proxy Statement for the Annual Meeting of Shareowners to be held on April 22, 2009, are incorporated by reference in Part III.

ARBITRATION AGREEMENT

This Arbitration Agreement (the "Agreement") is effective [DATE].

BETWEEN: [YOUR COMPANY NAME], as represented by [ATTORNEY NAME], a corporation organized and existing under the laws of the [State/Province] of [STATE/PROVINCE], with its head office located at:

AND: [PARTY B] (the "Party B"), a corporation organized and existing under the laws of the [State/Province] of [STATE/PROVINCE], with its head office located at:

In consideration of the terms and covenants of this agreement, and other valuable consideration, the parties agree as follows:

1. MATTERS TO BE SUBMITTED TO ARBITRATION

All disputes and controversies of every kind and nature between the parties to this agreement arising out of or in connection with [SPECIFY GENERAL AGREEMENT TO WHICH ARBITRATION AGREEMENT RELATES] as to the existence, construction, validity, interpretation or meaning, performance, nonperformance, enforcement, operation, breach, continuance, or termination of the agreement shall be submitted to arbitration pursuant to the procedure set forth in this agreement.

2. PROCEDURE

Either party may demand such arbitration in writing within [NUMBER] days after the controversy arises, which demand shall include the name of the arbitrator appointed by the party demanding arbitration, together with a statement of the matter in controversy.

- Within [NUMBER] days after such demand, the other party shall name [HIS OR HER] arbitrator, or in default of such naming, such arbitrator shall be named [IMMEDIATELY OR NOT] by the Arbitration Committee, and the two arbitrators so selected shall name a third arbitrator within [NUMBER] days or, in lieu of such agreement on a third arbitrator by the two arbitrators so appointed, a third arbitrator shall be appointed by the Arbitration Committee.
- The arbitration hearing shall be held at [PLACE OF HEARING] on [NUMBER] days' notice to the parties.
- The arbitration rules and procedures of [PARTICULAR TRADE, INDUSTRY, OR ASSOCIATION] shall be used in the arbitration hearing and the law of evidence of [STATE/PROVINCE] shall govern the presentation of evidence at such hearing.
- The arbitration hearing shall be concluded within [NUMBER] days unless otherwise ordered by the arbitrators and the award on the hearing shall be made within [NUMBER] days after the close of the submission of evidence.
- An award rendered by a majority of the arbitrators appointed under this agreement shall be final and binding on all parties to the proceeding during the period of this agreement, and

Arbitration Agreement

Page 1 of 3

Your
LOGO

SURGERY OPERATIVE REPORT

atient name:	MR number:	Date of operation:
bigall Houston	000843	1/13/2023
count No:	Height:	Weight:
472942	5'4"	130lb

operative diagnosis:

peridoneal artery bleeding. Estimated blood loss.

st-operative diagnosis:

peridoneal artery bleeding has been stopped

peration performed:

Therapeutic abdominal surgery	
reogn	Martice Cole
esthesia	General
ndition	Stable
mplications	None

inical findings:

22-year-old female visited our hospital because of abdominal pain as she sustained a fall from a ladder home. The patient then fell unconscious. We discussed operative treatment with her husband and an r recommendation he agreed to proceed with surgical treatment.

rocedure:

A patient was brought to the operating room #306 and placed on the operating table, general anesthesia was administered at 3:29. After the patient fell asleep, we opened the epigastric incision and formed an emergency operation to stop bleeding. Two weeks after stopping the bleeding, the patient is brought for the appointment and she did not have complaints of any abdominal symptoms or eding for 3 months.

uthorization for release of information

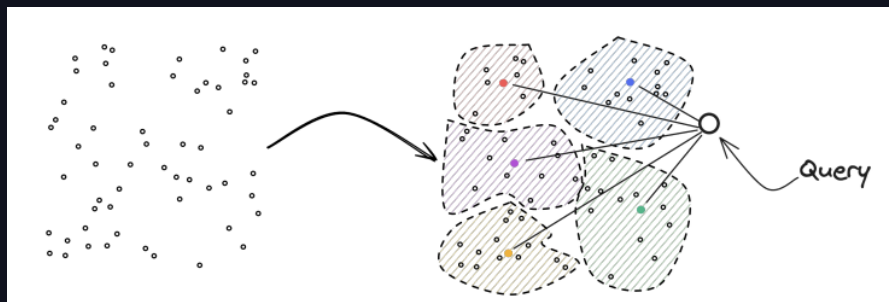
iginal: Medical record Copy: Patient Exhibit No.2 1-37-27846



WHAT IS A VECTOR DATABASE

Indexing unstructured data

- Vector databases are databases specifically designed to perform similarity search over large corpus of unstructured data
- Vector databases use geometric methods like graphs and clustering to efficiently scan and find relevant pieces of information, for a given query



VECTOR DATABASE USE CASES

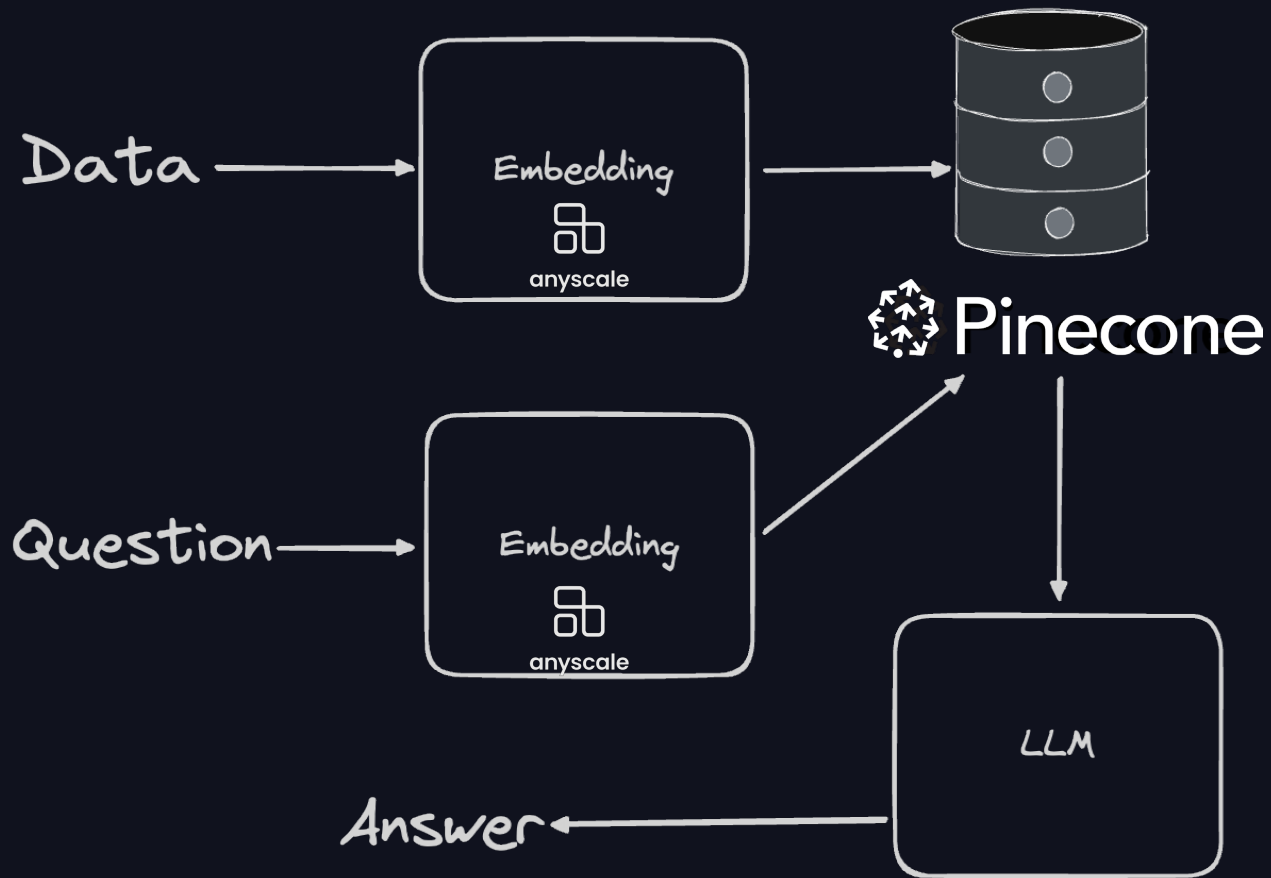
Searching for meaning

- **RAG and AI applications** - to store and retrieve external, private knowledge
- **Semantic search** - allows applications to search by meaning to find relevant results even if the exact words don't match
- **Classification** - being able to label or classify large number of labels
- **Candidate / Training data generation** - being able to filter subset of examples from a corpus to train ML models
- And way more...

RAG with RAY and PINECONE

RAG for enterprise scale

- Ray scales embedding and processing of large datasets
- Pinecone with the Serverless architecture, enables search to scale to 10s of billions of documents
- Pinecone + Ray = scale for modern enterprise RAG



RAY





RAY

Open source operating system powering the AI evolution.



Berkeley
UNIVERSITY OF CALIFORNIA



anyscale

2016

Ray was conceived at the University of California Berkeley in the RISELab, (where Apache Spark was also born) and was open-sourced.

2019

Anyscale was founded - **by the creators of Ray** - as a fully managed platform for Ray to make the toughest problems in distributed computing easier for developers.

RAY

A short history

- 2016: Started as a class project at UC Berkeley
- 2019: Anyscale founded (company behind Ray)
- 2020: Ray v1.0 release
- 2022: Ray v2.0 release; Ray AI Libraries
- 2023: LLM + AI Infra

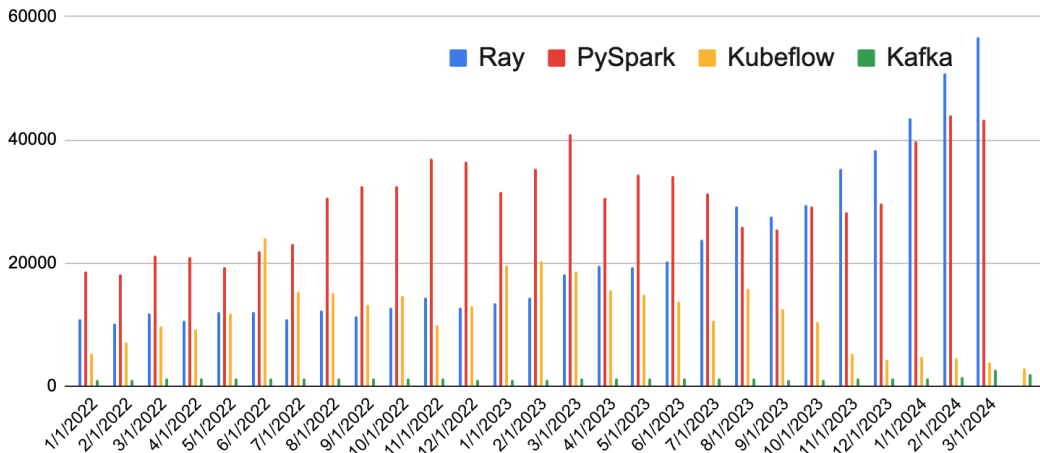
Open source project with 30k+ stars on Github: <https://github.com/ray-project/ray>

RAY USERS

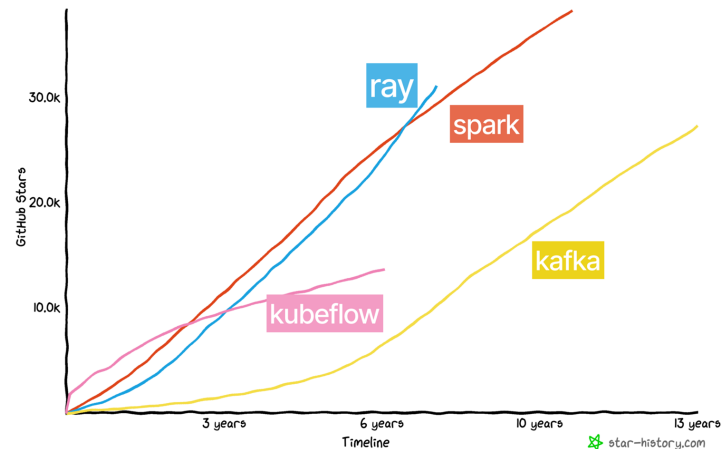


RAY

The OS for running AI workloads at hyper-scale.



Monthly PyPi downloads on Mac



GitHub star history

ANYSCALE

THE END-TO-END AI PLATFORM FOR AI COMPANIES

***Optimized Ray
Runtime**

***Developer
Tooling for Ray**

***Security &
Governance**

***Integrations for
Ecosystem**



Develop



Process data



Train



Fine-tune



Deploy



Inference
(Online & Batch)



Test & Debug

Dev Workspaces Batch Jobs Services

Unified RAY runtime runtime (OSS + proprietary optimizations)

Managed services

Observability

Access control



RAY

Holistically addresses AI/LLM challenges

 RAY Unified Framework for Scaling AI Workloads

Ray AI Libraries

Data

Train

Tune

Reinforcement
Learning

Serve

Ray Core

“Operating System” for heterogeneous distributed computing



RAY DATA

Scalable data processing library for ML workloads

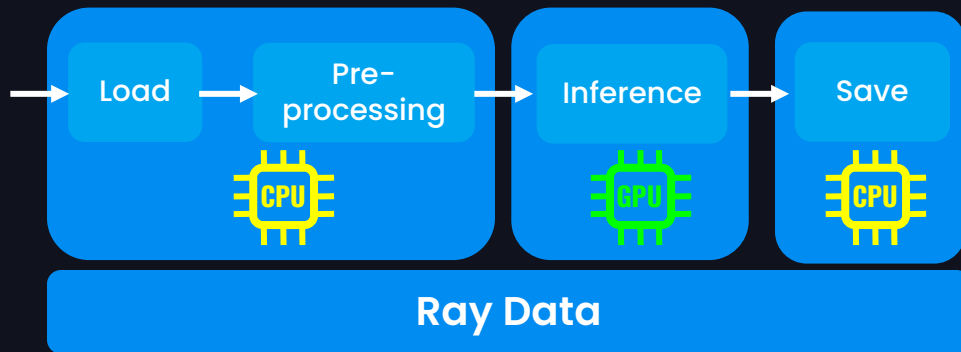
Open source: <https://docs.ray.io/en/latest/data/data.html>

- Offline batch inference
 - Large-scale image, video, audio and text processing
 - RAG embedding generation
 - LLM offline batch inference

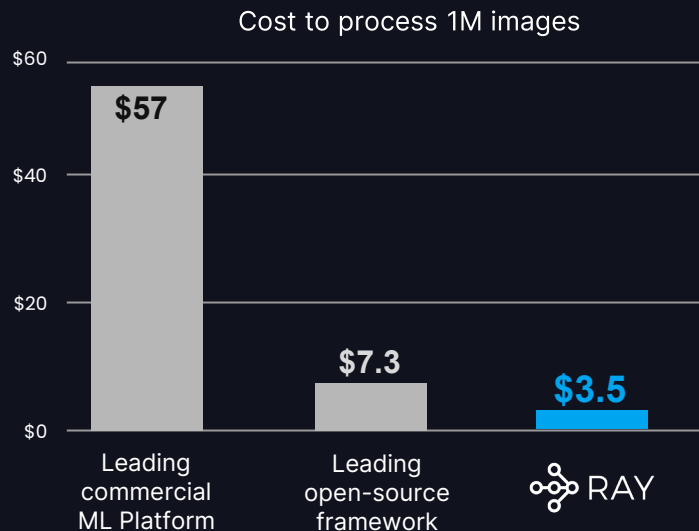
UNSTRUCTURED DATA BENCHMARK

A typical workflow for offline batch inference

Full blog post with
code example



- Use most cost-effective hardware for each stage
- Independently scale every stage



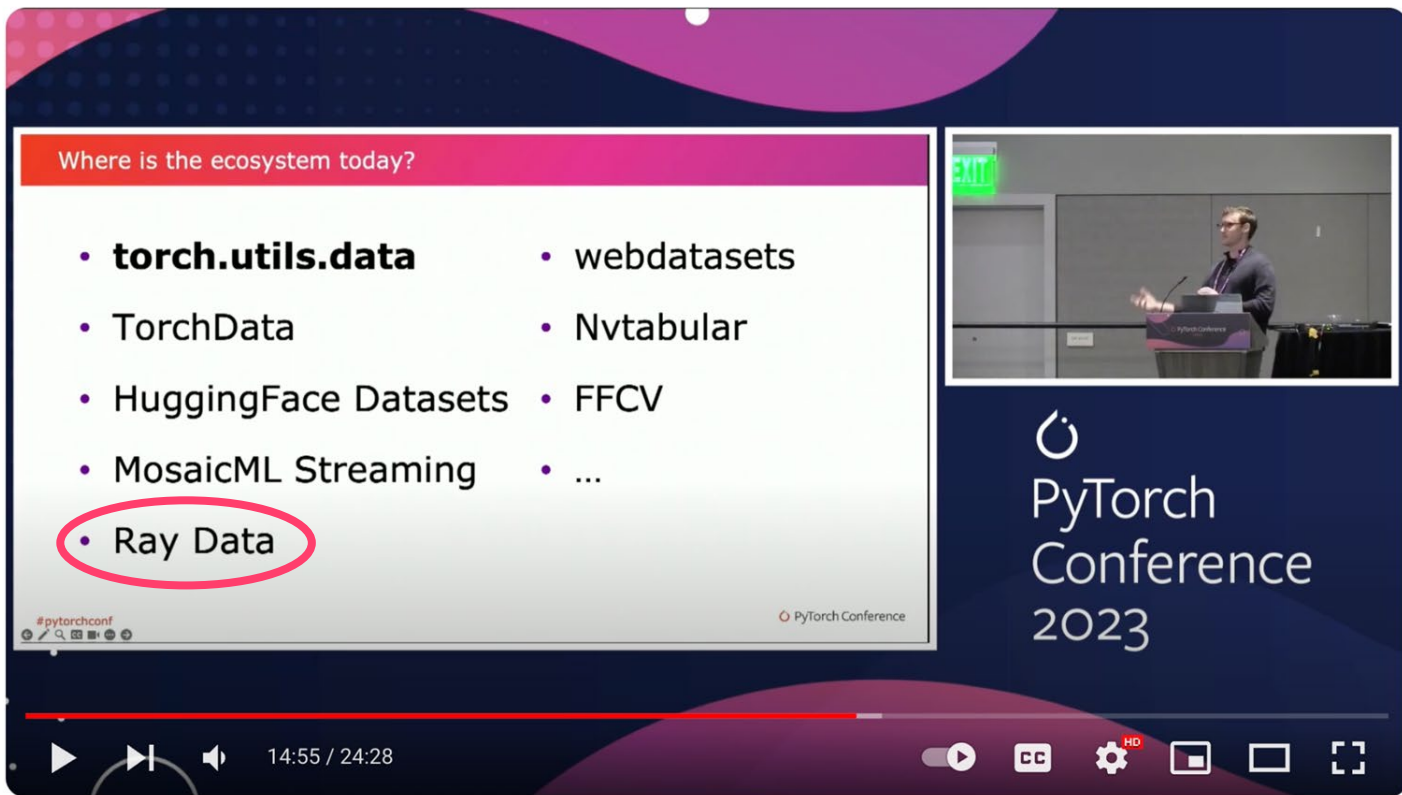
RAY DATA

Scalable data processing library for ML workloads

- Training ingestion for ML models
 - Scalable distributed training data loading and preprocessing
 - Support all popular deep learning frameworks (PyTorch, TensorFlow)
 - Ray Data + Ray Train

RAY DATA

In the PyTorch ecosystem



The screenshot shows a video player interface. The main content is a presentation slide with a pink header that reads "Where is the ecosystem today?". Below the header is a list of items in the PyTorch ecosystem. The item "Ray Data" is circled in red. To the right of the slide is a small video inset showing a man speaking at a podium. The background of the video player is dark blue with the PyTorch logo and "PyTorch Conference 2023" text. At the bottom of the video player is a red progress bar and a control bar with play, pause, volume, and other icons. The current time is 14:55 / 24:28.

Where is the ecosystem today?

- **torch.utils.data**
- TorchData
- HuggingFace Datasets
- MosaicML Streaming
- Ray Data
- webdatasets
- Nvtabular
- FFCV
- ...

#pytorchconf PyTorch Conference

PyTorch Conference 2023

14:55 / 24:28

The Evolving Landscape of Dataloading - Laurence Rouesnel, Meta



PyTorch
45.6K subscribers

Subscribe

10



Share

Download



WHY RAY DATA

Scalable data processing library for ML workloads

- Streaming execution, scalable to petabyte-scale data
- Support heterogeneous resource requirements (CPU+GPU)
- Automatic failure recovery (spot instance, process/node/cluster failure)
- Support a large variety of data sources and formats
 - AWS S3, Google GCS, Microsoft Azure, On-prem HDFS, NFS, local file system, ...
 - Parquet, image, video, audio, text, JSON, CSV, vector databases, SQL, ...
- Python native & seamless integration with other ML libraries

1 BILLION EMBEDDING BENCHMARK

Cheapest solution in the world: \$6,000 on Anyscale

Input data

- [tiiuae/falcon-refinedweb](#)
- A massive English web dataset built by TII
- 1.6 TB on-disk data size

Embedding model

- [thenlper/gte-large](#)
- 1 billion embeddings

Anyscale Job

- Cluster: 320 A10G AWS g5 instances
- End-to-end time: 18 hours

EMBEDDING with RAY

4-Steps approach

*Full blog post with
code example*



- Reading
 - Read input data on cloud storage (TB scale): AWS S3, Google GCS, etc
 - Data format: Text, JSON, Parquet, CSV, PDF, Word, etc
- Chunking
 - Split the text up into small, semantically meaningful chunks (often sentences)
 - Tool: LangChain, LlamaIndex Text Splitters

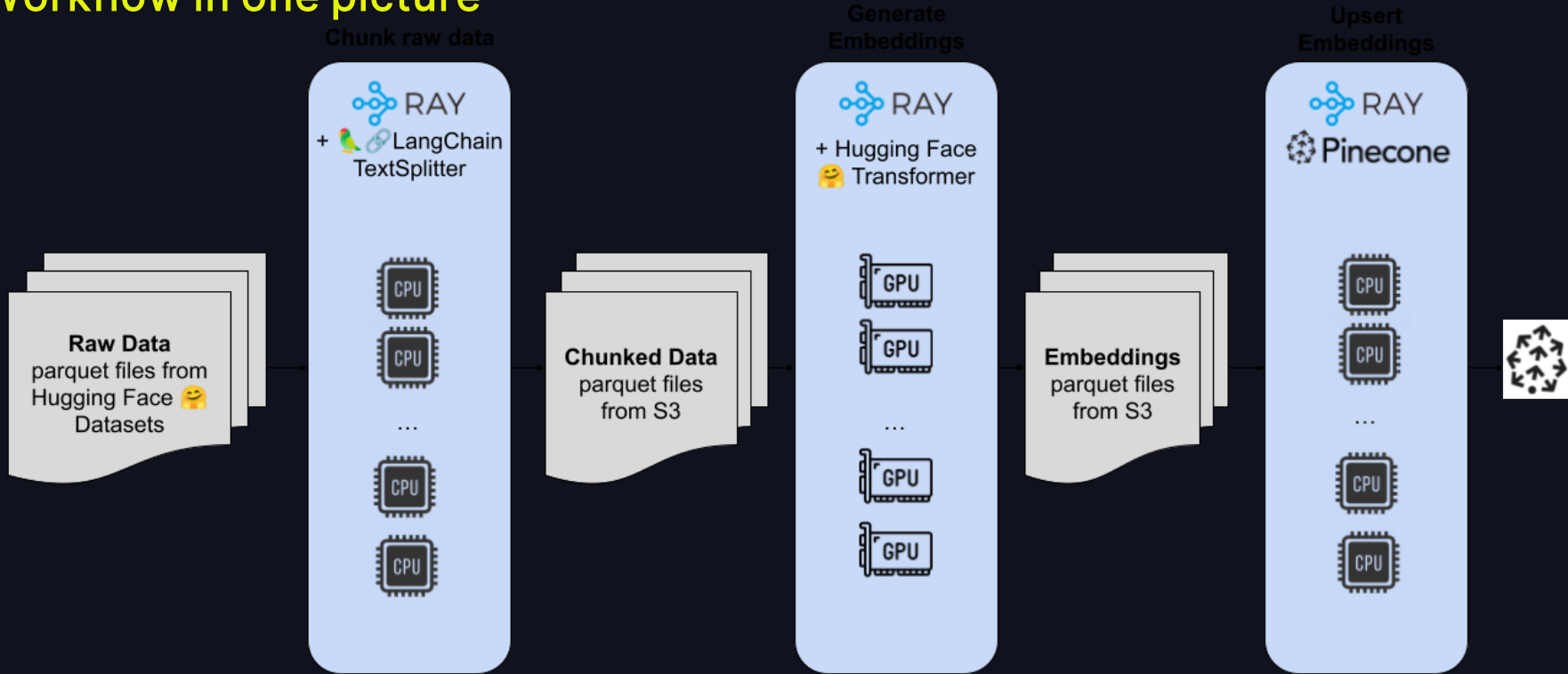
EMBEDDING with RAY

4-Steps approach

- **Tokenization & Embedding**
 - Convert sentences to tokens, and use embedding model to generate embedding
 - Embedding model: GTE, Mistral, etc
- **Writing to vector database**
 - Batch upserting embedding result (and other information) to vector database
 - Vector database: Pinecone Serverless

EMBEDDING with RAY

Workflow in one picture



LLM INFERENCE

Offline or online, that is the question

- Online inference (serving): API Endpoints (OpenAI, Bedrock, Anyscale)
 - Latency: get real-time response for prompt or to interact with the LLM
 - Choose this if optimizing latency of inference to be as fast as possible
- Offline inference (batch job): OpenAI Batch API, Anyscale Job
 - Throughput: optimize resource usage (maximize GPU utilization)
 - Choose this if scaling your workload to large-scale datasets with the least cost
 - Example of use cases: daily / hourly job to process large-scale in-house offline data (text summarization, users information processing, ...)

PINECONE



PINECONE

The AI infrastructure company

- Since 2019
- \$138M
- ~200 ppl in NY, TLV, SF, ...
- Pinecone Serverless (Jan 24')
- Canopy RAG Framework (Nov 23')

HubSpot

L'ORÉAL

shopify

CVSHealth

GONG

Klarna.

zapier

Microsoft

Notion

BAIN
& COMPANY

Course Hero

ClickUp

PINECONE SERVERLESS

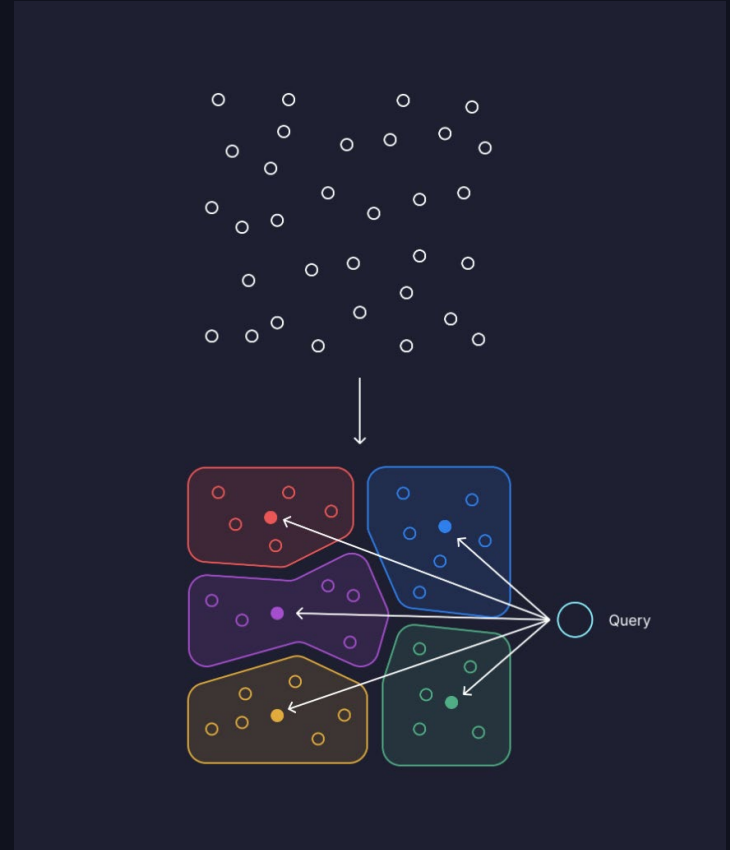
How does the Serverless architecture supports scaling?

I am going to tell you all the secrets on how to build your own vector database

PINECONE SERVERLESS

Dynamic Clustering Algorithm

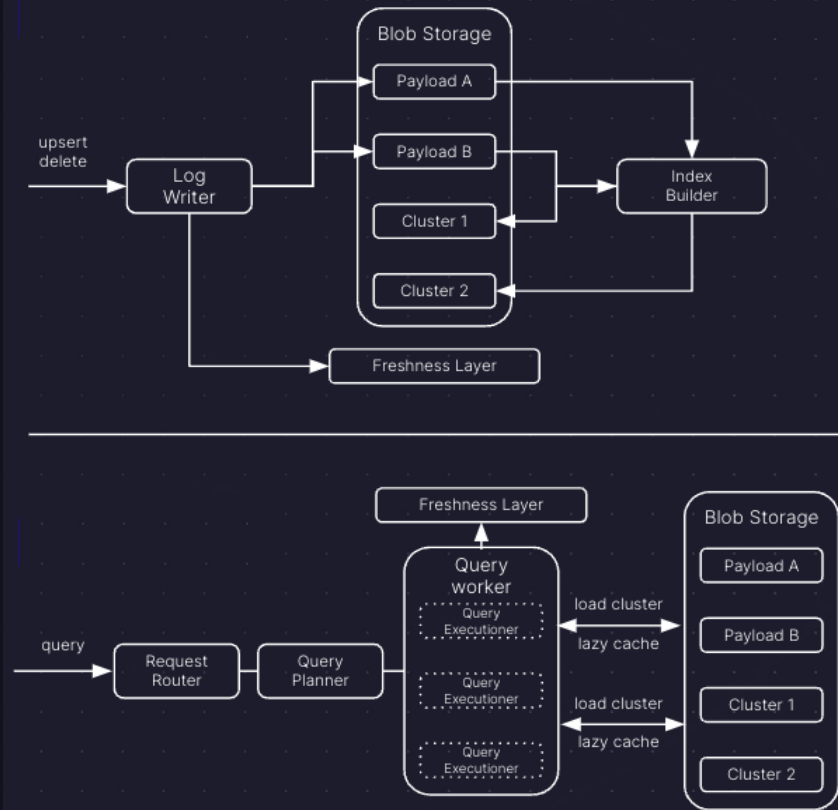
- Clusters data as it flows into the system
- In query time: only searches relevant clusters according to filters and query vector
- Scales really well with low working memory (compared to graph algo. s.a. HNSW)



PINECONE SERVERLESS

Separating Write / Read

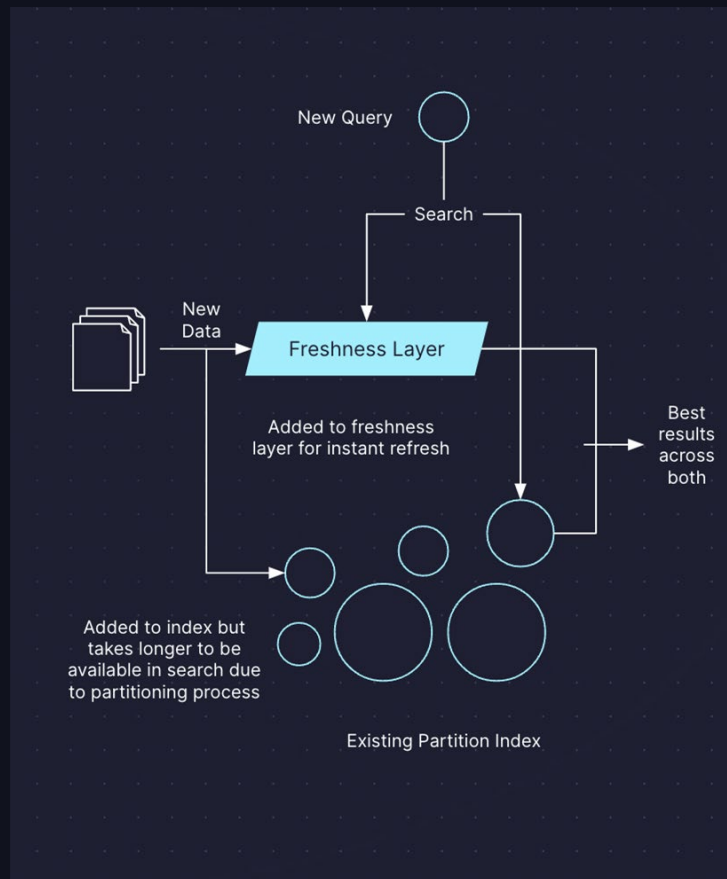
- Scaling read/write workers separately
- Writing / reading clusters in parallel
- Strong caching mechanism
- Fresh layer for recent writes



PINECONE SERVERLESS

Freshness Layer

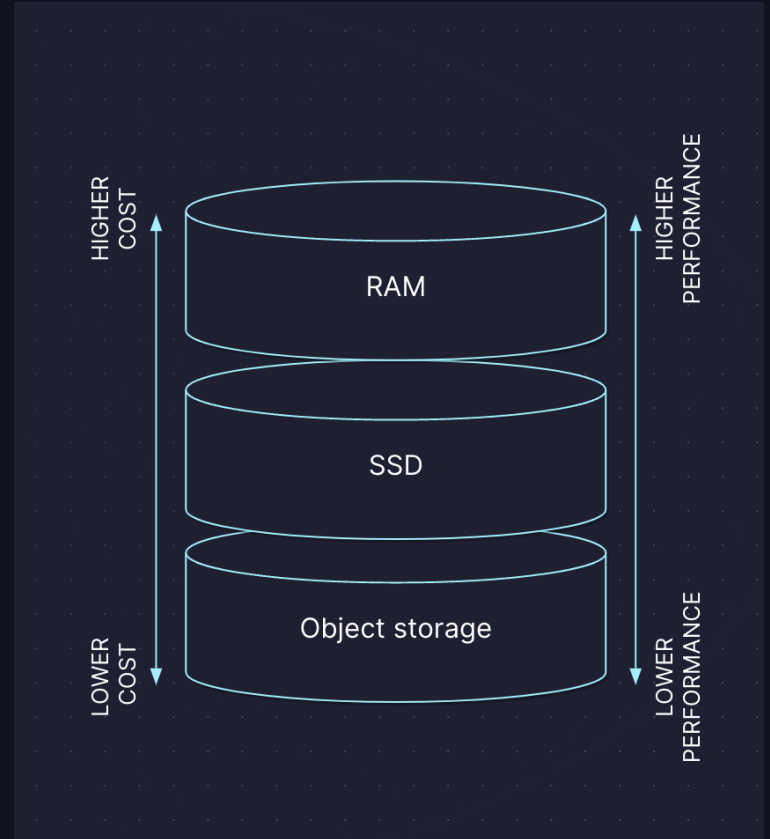
- Data immediately available for query
- As the index re-cluster data get indexed



PINECONE SERVERLESS

Store data to match query pattern

- Utilize different storage layers to different queries
- Move vectors between different storage layers automatically



RESULTS



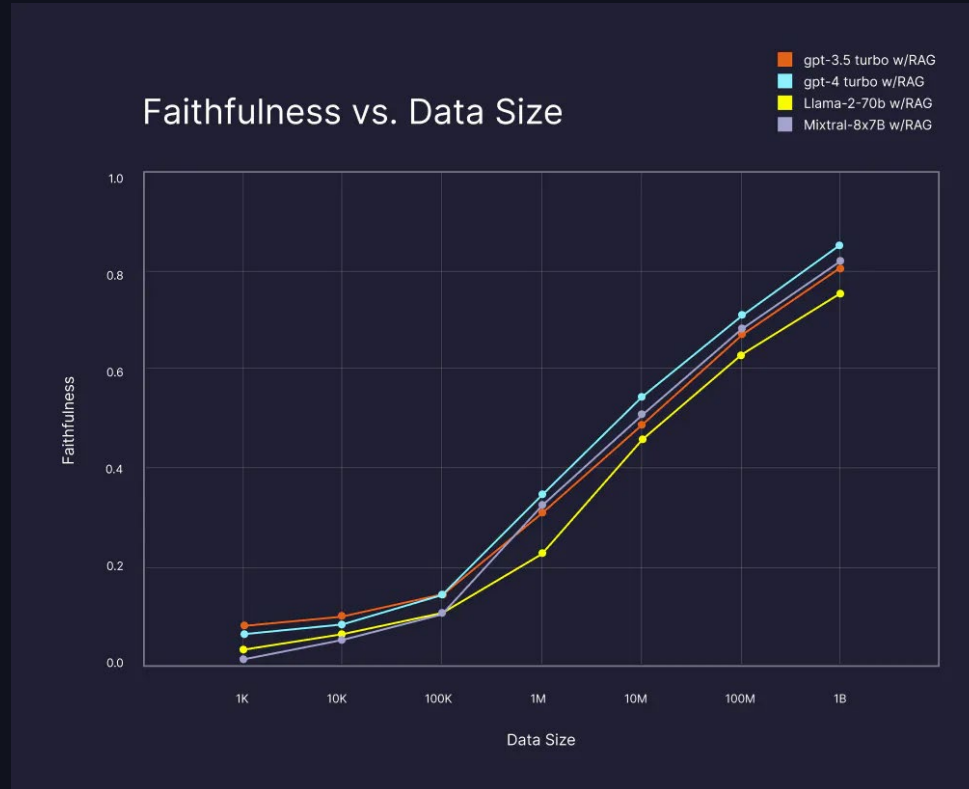
RESULTS - RAG vs TRAINING

More data, more accuracy, less cost

- The more data you upload -> higher the performance
- “Levels” the quality of OSS vs Closed LLMs

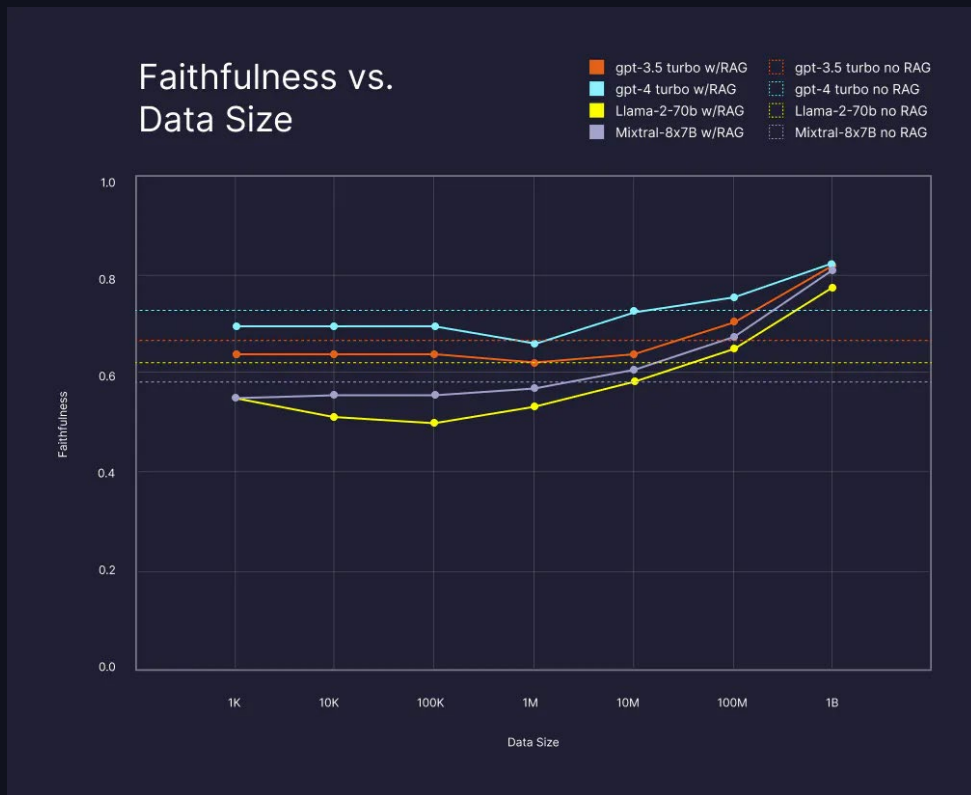
RESULTS (RAG vs TRAINING)

More data, more accuracy, less cost



RESULTS (RAG vs TRAINING)

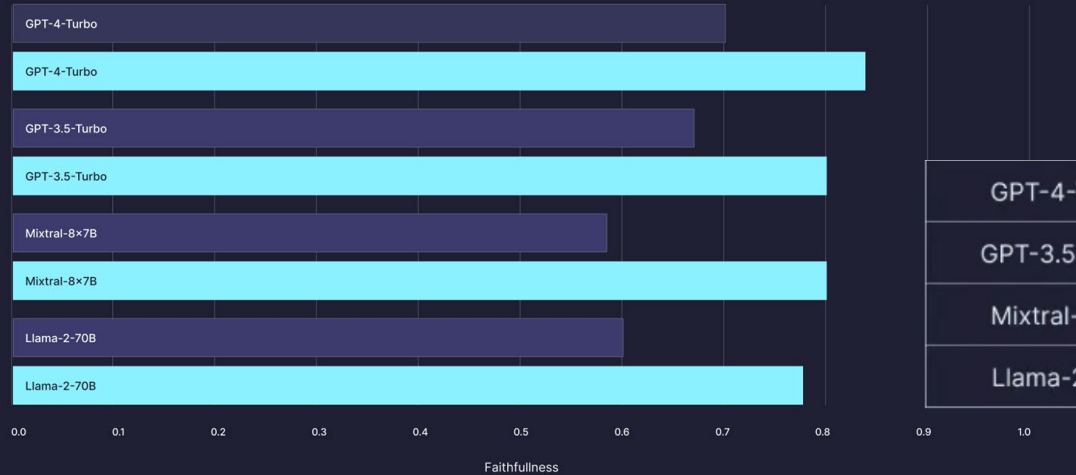
More data, more accuracy, less cost



RAG vs TRAINING

More data, more accuracy, less cost

Faithfulness per Model
with and without RAG



GPT-4-Turbo	0.700	0.835
GPT-3.5-Turbo	0.669	0.804
Mixtral-8x7B	0.583	0.808
Llama-2-70b	0.609	0.760



RECAP

Anyscale + Pinecone: RAG for enterprise scale

- Ray on Anyscale
 - Use Ray Data to scale embedding generation across a cluster of GPUs
- Pinecone
 - Using high scale vector embeddings to boost RAG Performance
 - Utilizing serverless architecture to support cost effective scale
- With Anyscale and Pinecone, scale RAG application with your own data!

MORE SESSIONS

To learn more ...

Session: *Efficient Stable Diffusion Pre-Training on Billions of Images with Ray*

Speakers: Yunxuan Xiao, Hao Chen

Time: 12:30-1:10 PM, June 13

Session: *Accelerating LLM Inference with vLLM*

Speakers: Zhuohan Li, Cade Daniel

Time: 2:50-3:30 PM, June 13

THANK YOU!

To learn more ...

