Applying Semantic Search on a Valkey Cache with valkey-py and valkey-search

Narissa Tsuboi



demo code

About Me

Software Engineer 3 (Backend) @ Read AI

Stack - Python, SQLAlchemy, Pydantic, AWS, Kafka, Celery, Postgres.

Interests - My pomeranians, weightlifting, learning new things purely for fun.

Motivation for this talk in particular - To push myself to learn more about vector databases and related topics.

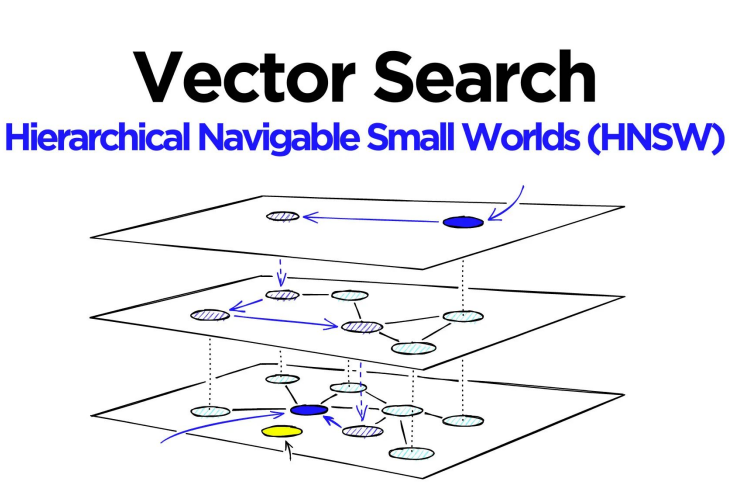


What I want you to get from this talk

1. Awareness of how you *could* configure valkey-py cluster client in your Python application
2. Why you might use Valkey cache for semantic search and repeatable queries
3. Some fun!

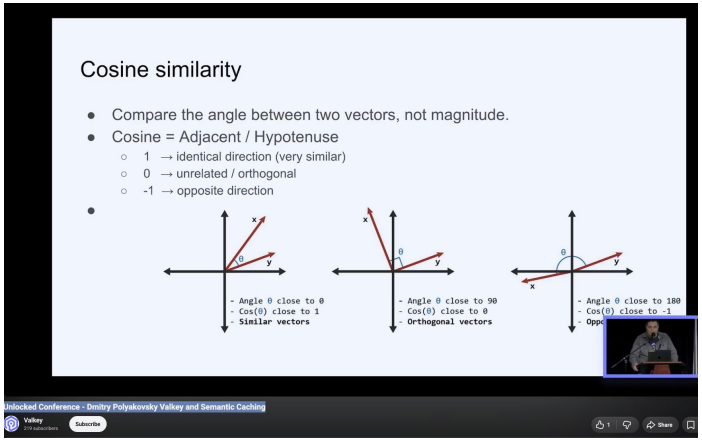
What I’m not going to talk about

- Cache invalidation logic for LLM query meanings.
- Text embedding models.
- K Nearest Neighbor algorithm.
- Hierarchical Navigable Small Worlds.

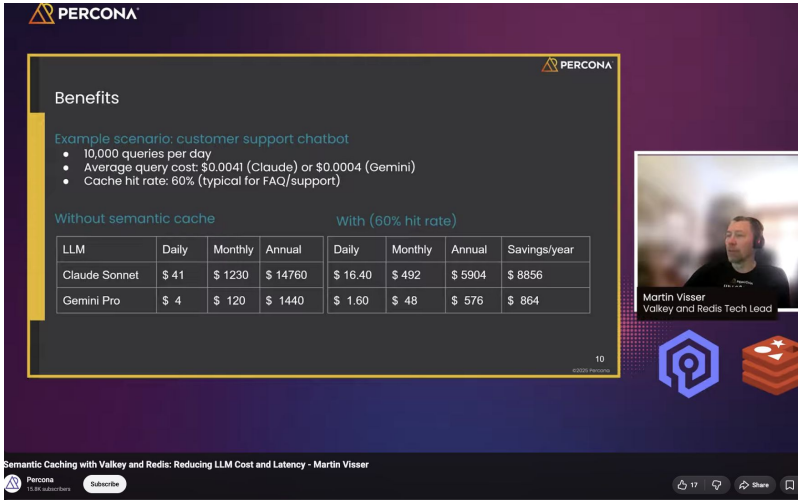




Deeper Dives



Unlocked Conference - Dmitry Polyakovsky Valkey and Semantic Caching
https://www.youtube.com/watch?v=pQfpTe7CY80



Semantic Caching with Valkey and Redis: Reducing LLM Cost and Latency - Martin Visser
https://www.youtube.com/watch?v=7vdFUJgGOSs

Outline

- Semantic Search
- Caching for Semantic Search
- valkey-py package && valkey-search module
- Similar Color Semantic Search demo

What is Semantic Search

Semantic search is a search engine technology that interprets the meaning of words and phrases.

The results of a semantic search will return content matching the ***meaning*** of a query, as opposed to content that literally matches words in the query.



https://www.elastic.co/what-is/semantic-search
https://www.geeksforgeeks.org/nlp/what-is-semantic-search/

Semantic Search IRL



Will Costco's world-renowned web experience apply semantic search and return results based of what I *mean??????*



Semantic Search IRL

While Supplies Last Online-Only T

COSTCO WHOLESALE

sit in bubbling water

NOPE.

Maybe that search was too dissimilar? Let's try again…



$99.99
Step2 Tropical Splash Water Park
(123)
Delivery Available
Compare
Add to Cart



$349.97
Happy Hop Dinosaur Cave Adventure
(64)
Delivery Available
Out of Stock at Seattle
Compare
Add to Cart



Online Only
$1,399.99
Triple Handcrafted Boulder Fountains
(47)
Delivery Available
Compare
Add to Cart



Online Only
$999.99
Mountain Spring Pondless 24" Bubbler Fountain
(159)
Delivery Available
Compare
Add to Cart

tions

+ Feedback



Online Only
$37.99



Online Only



Online Only



Online Only

Semantic Search IRL

COSTCO WHOLESALE

outdoor tub

Cart



Online Only

$3,999.99

$3,999.99 After $1,800 OFF

Evolution by Strong Spas Enhance 40-jet, 6 or 7-person Acrylic Spa

(129)

Delivery Available

Compare

Select Options



Online Only

$6,999.99

$6,999.99 After $2,000 OFF

Evolution by Strong Spas Stockbridge 95-jet, 6 or 7-person Acrylic Spa

(513)

Delivery Available

Compare

Select Options



$29.97

H2OGO! Inflatable Family Lounge Pool

(167)

Delivery Available

Compare

Add to Cart



Costco Direct

$2,999.97

Buy 2+ Costco Direct items; Instantly save an additional $100 - $400. See Product Details

LifeTrend Solitude Integrated Cold Plunge & Hot Soak Tub

(24)

Delivery Available

Compare

Add to Cart





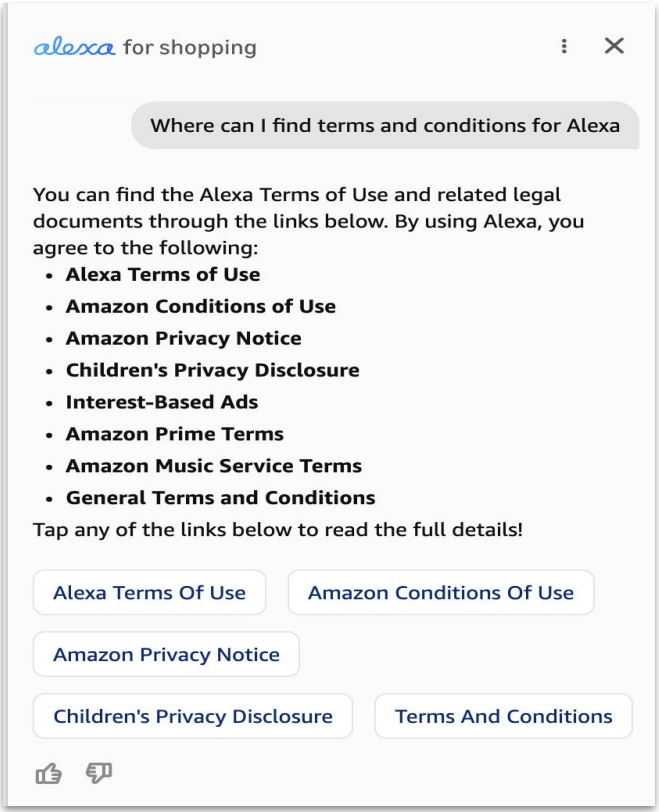




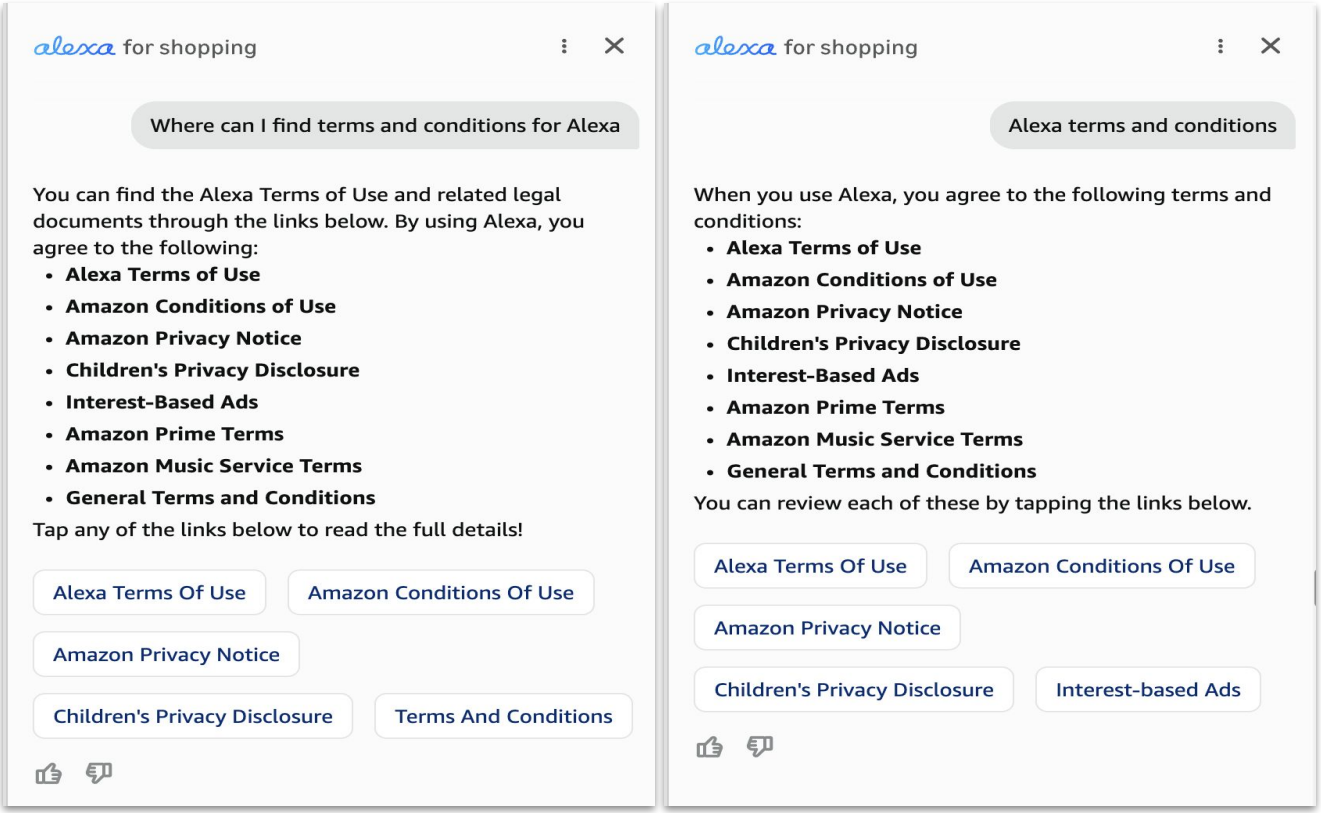
Feedback

There we go!

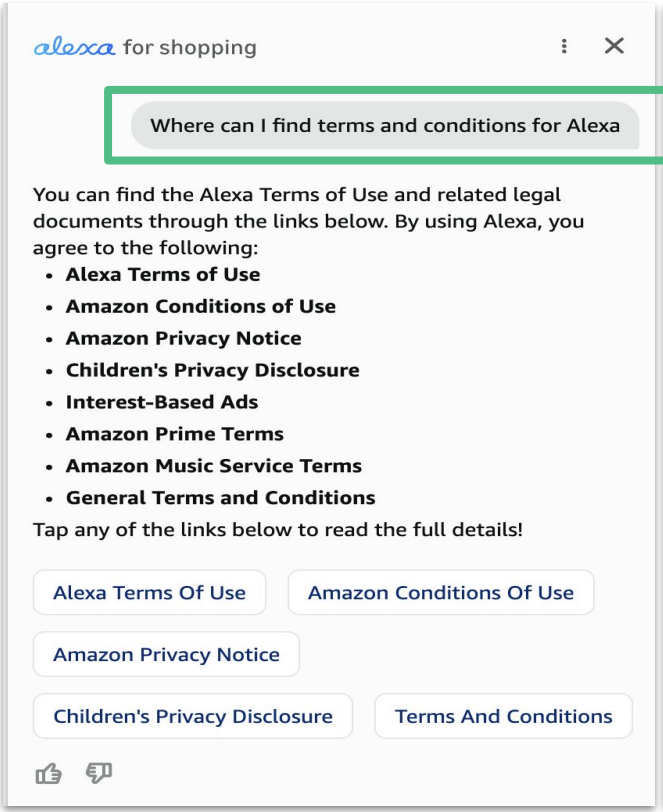
Semantic Search + Chatbot IRL

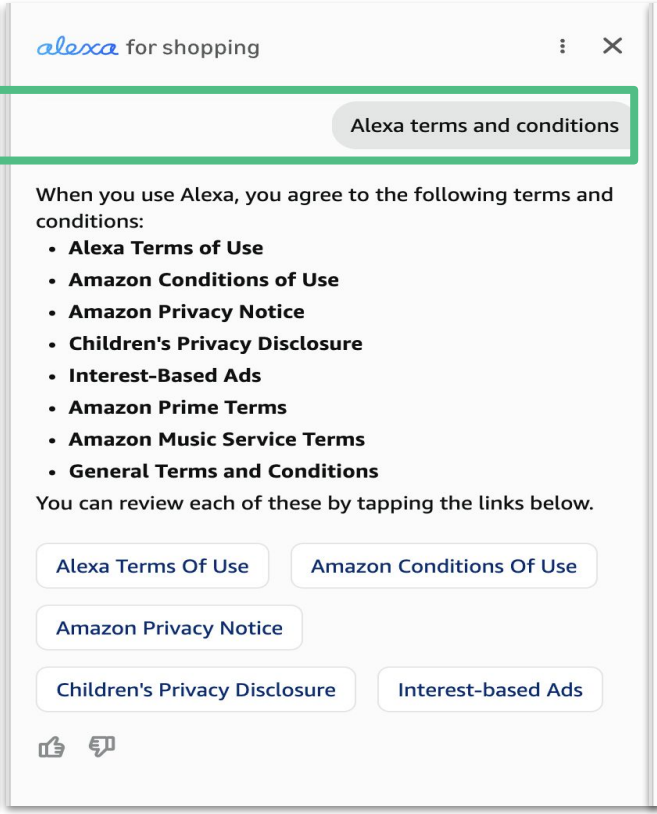


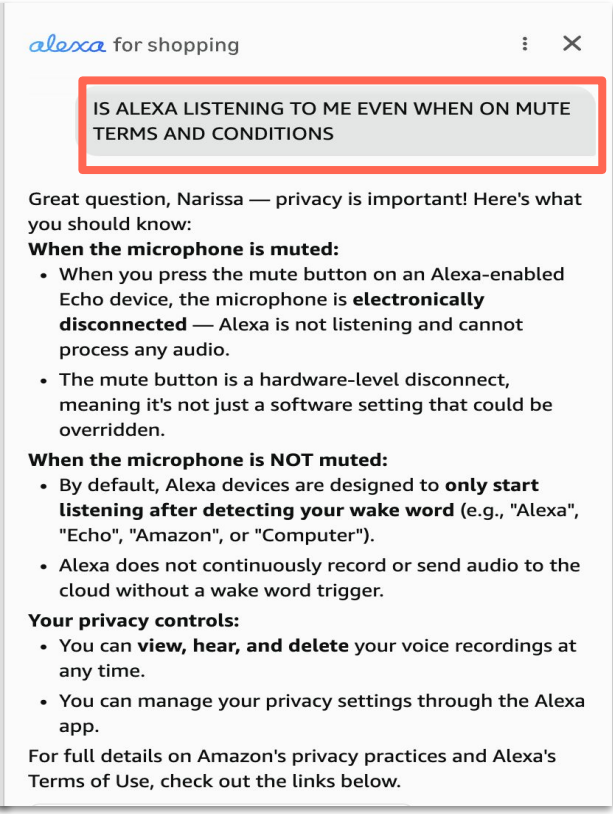
Semantic Search + Chatbot IRL



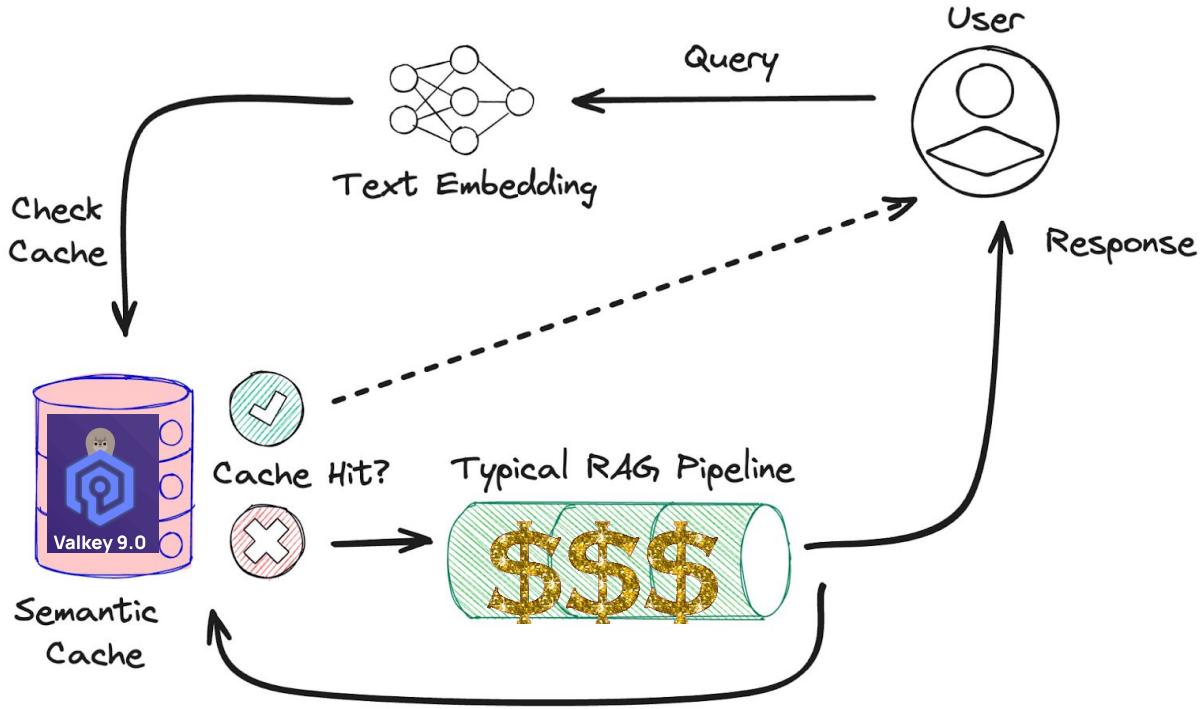
Semantic Search + Chatbot IRL







Semantic Cache Flow

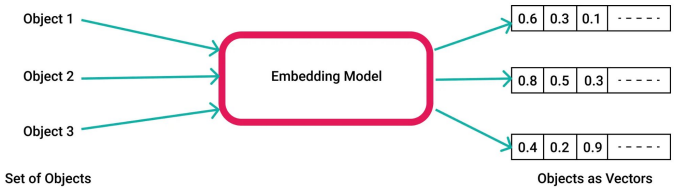


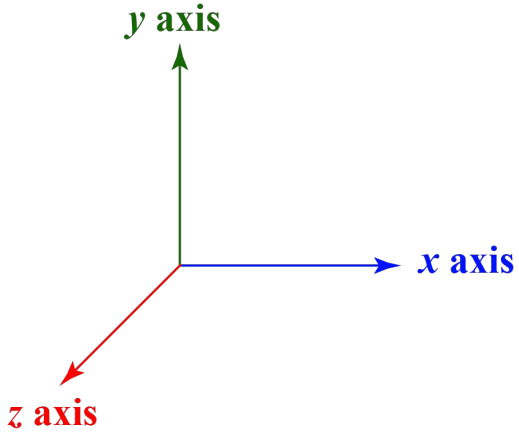
https://redis.io/blog/level-up-rag-apps-with-redis-vector-library/

Vector Embeddings

- Vector embedding: A numerical representation that aims to capture the meaning of the original data.

“Sit in bubbling water” → [0.11, 0.02, 0.54]

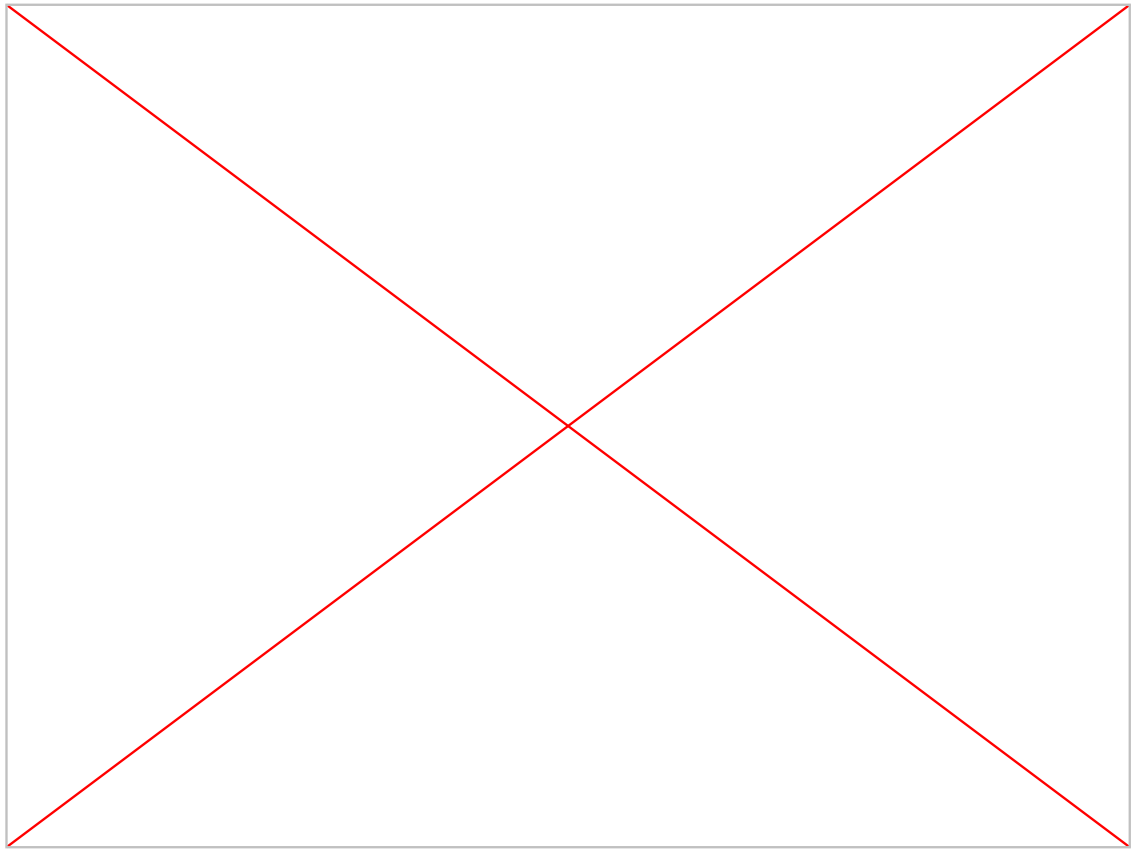




- Embedding models: Models trained to produce embeddings (ie text, images, etc)

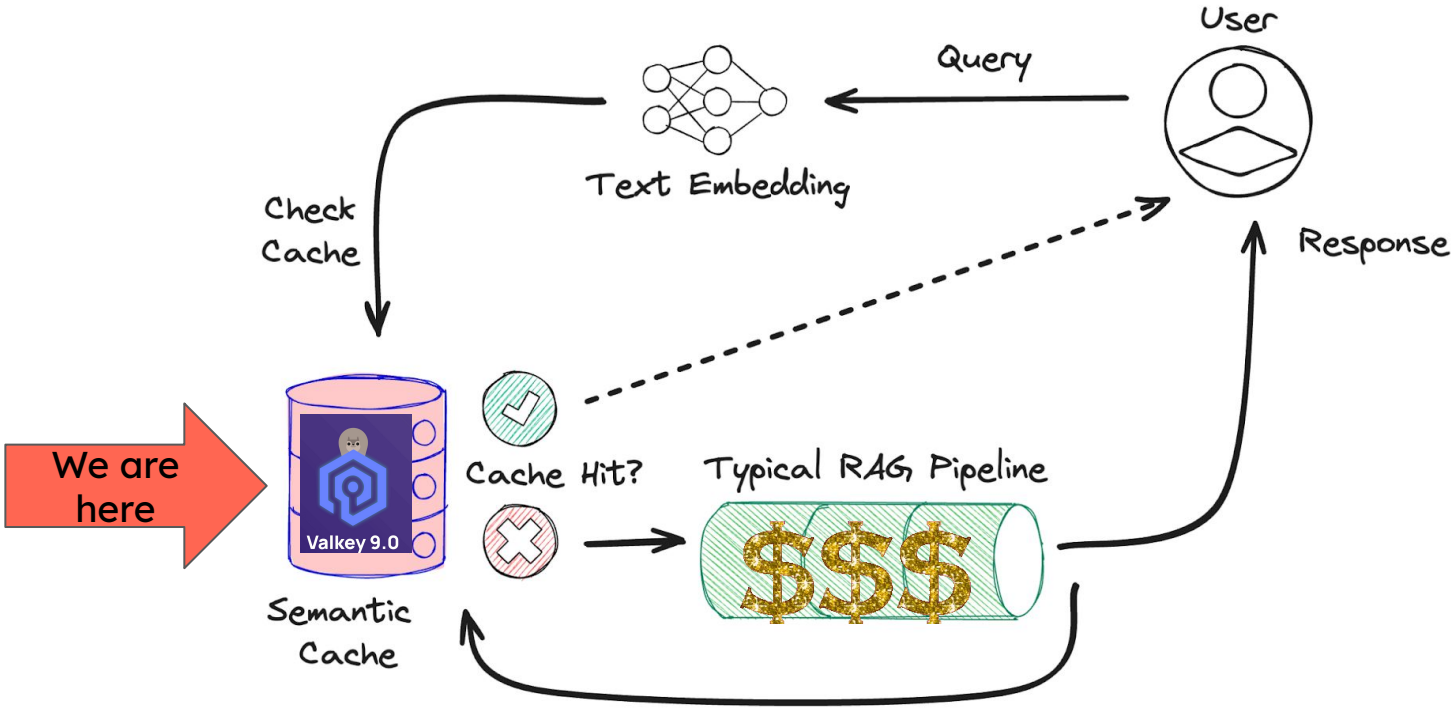
O'Reily AI Engineering, Huyen, Introduction to Embedding, https://www.pinecone.io/learn/vector-embeddings/

Searching for Similar Colors



https://huggingface.co/spaces/jphwang/colorful_vectors

Semantic Cache Flow



https://redis.io/blog/level-up-rag-apps-with-redis-vector-library/

Caching Refresher

1. What are the benefits of caching data?

2. Caches should be used as persistent data stores, True or False?

3. The internal data structure that let's a Key Value cache find a Value in O(1) time is called an _____.

Caching Refresher Answers



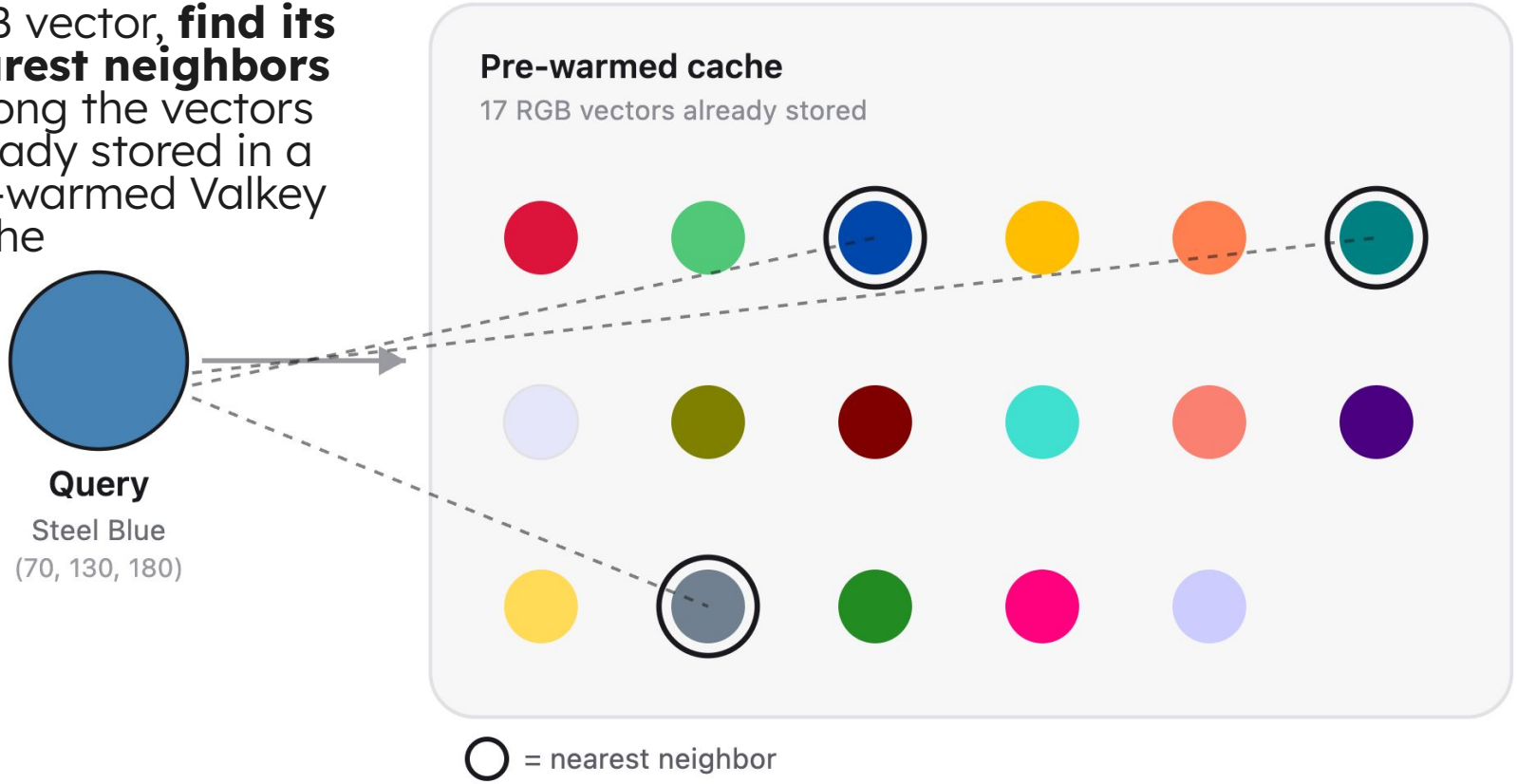
1. What are the benefits of caching data?
    a. Reduces load on backend systems
    b. Fast
    c. Potential cost savings when done right

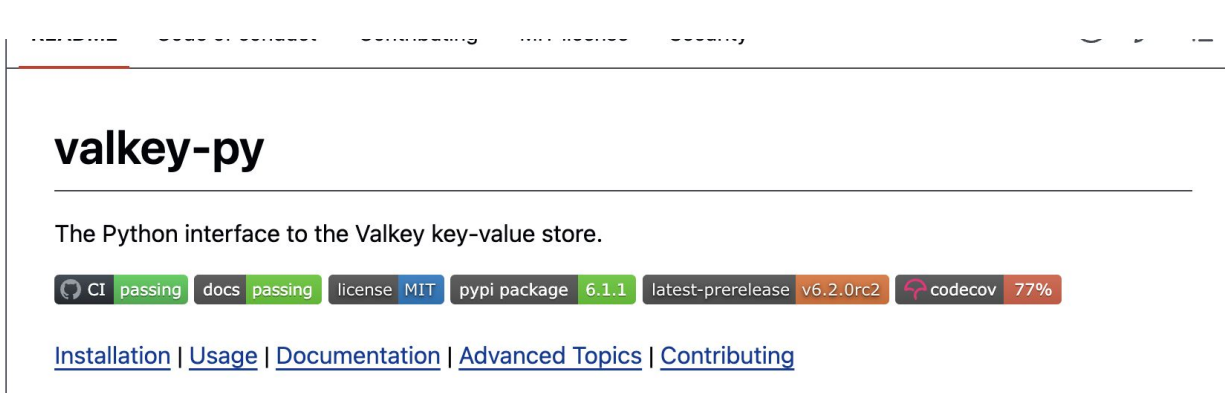
2. Caches should be used as persistent data stores, True or False?
    a. FALSE!!!!!!! Caches are meant to be a disposable layer in front of real data stores. Data can be evicted, no durability guarantees.

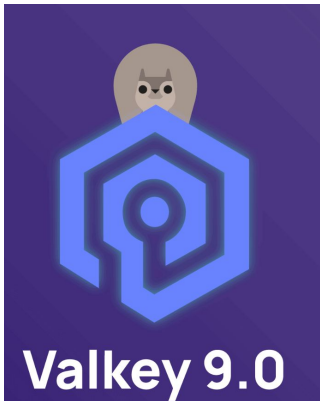
3. The internal data structure that let's a Key Value cache find a Value in O(1) time is called an index.

Demo Problem Statement:

Given a new color's RGB vector, **find its nearest neighbors** among the vectors already stored in a pre-warmed Valkey cache







- Forked from redis-py in March 2024, when Redis Inc. changed Redis's license away from fully open source.
- Now maintained by the Linux Foundation-backed valkey-io org, no single company owns it. MIT licensed.
- Started by Madelyn Olson, an AWS principal engineer and 6-year Redis contributor, who led the fork that became Valkey (Many of the Valkey contributors attend the monthly Valkey meetup!).

valkey-search https://github.com/valkey-io/valkey-search

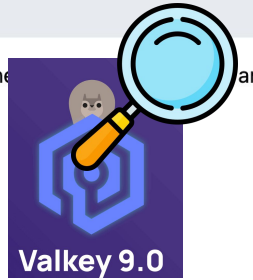
Valkey-Search (BSD-3-Clause), provided as a Valkey module, is a high-performance Search engine optimized for AI-driven / Search / Analytics / Recommendation System related workloads. It delivers single-digit millisecond latency and high QPS, capable of handling billions of vectors with over 99% recall as part of vector searches. It also provides support for hybrid / pure non vector workloads including Numeric, Tag, and Full-text searches.

Valkey-Search allows users to create indexes and perform searches, incorporating complex filters. Users can index data using either **Valkey Hash** or **Valkey-JSON** data types. The vector queries support Approximate Nearest Neighbor (ANN) search with HNSW and exact matching using K-Nearest Neighbors (KNN).

Supported Commands

```
FT.CREATE
FT.DROPINDEX
FT.INFO
FT._LIST
FT.SEARCH
FT.AGGREGATE
```

For a detailed description of the and configuration options, see the Command Reference.



Demo Code



valkey-py
Cluster
Client

valkey-semantic-search-demo > valkey_cluster > client.py > ...

```python

import os
from functools import lru_cache

from loguru import logger
from valkey import ValkeyCluster, ValkeyError

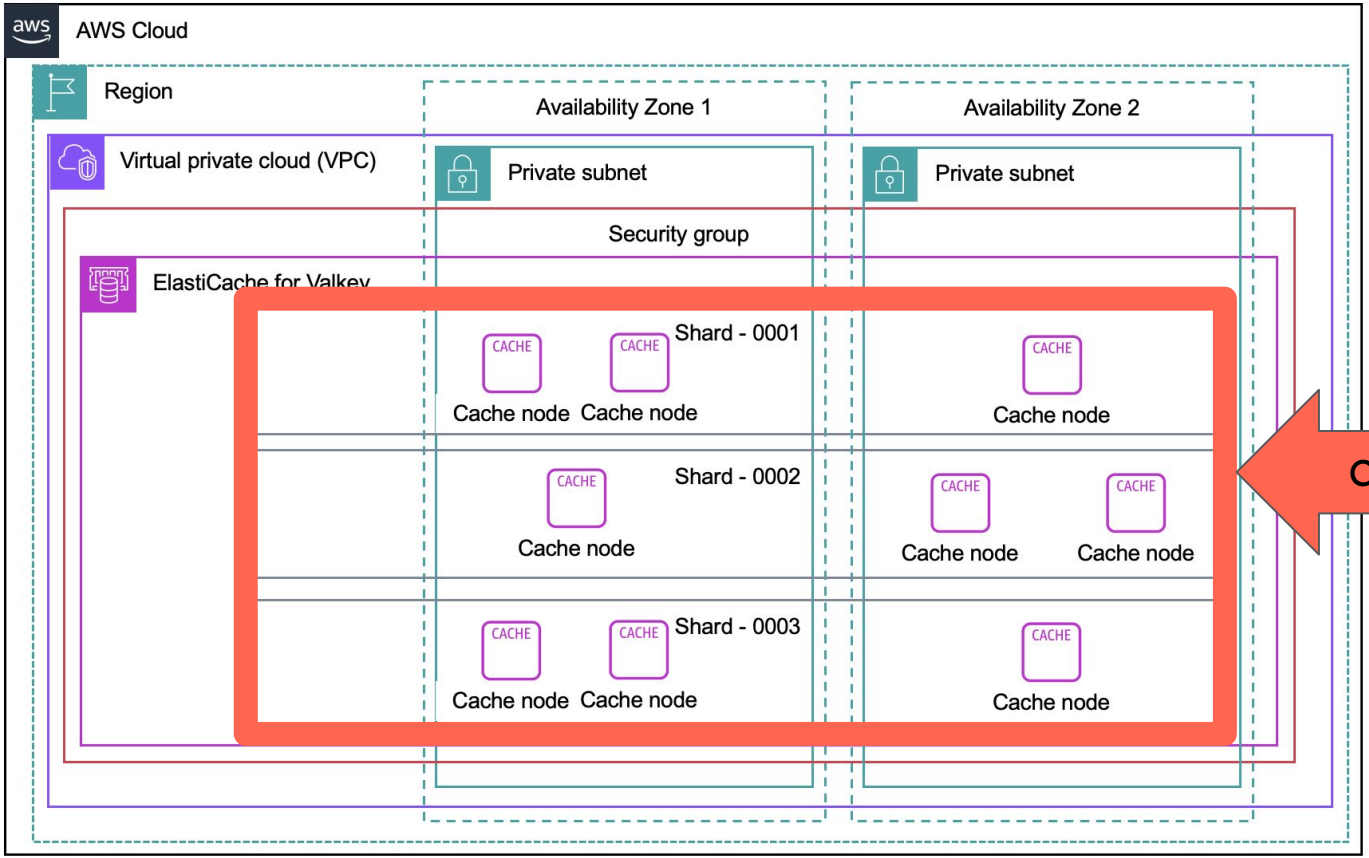
VALKEY_HOST = os.environ.get("VALKEY_HOST", "localhost")
VALKEY_PORT = int(os.environ.get("VALKEY_PORT", "6379"))


@lru_cache
def get_cluster_client():
    """Return the ValkeyCluster client.

    :returns: a connected ``ValkeyCluster`` client.
    :raises valkey.ValkeyError: if the connection attempt fails.
    """
    try:
        client = ValkeyCluster.from_url(
            url=f"valkey://{VALKEY_HOST}:{VALKEY_PORT}",
            password=os.environ.get("VALKEY_PASSWORD", None),
            health_check_interval=5,
            retry_on_error=[ConnectionError],
            decode_responses=True,
        )
    except ValkeyError as e:
        logger.error(f"Failed to connect to Valkey cluster: {e}")
        raise
    return client

```

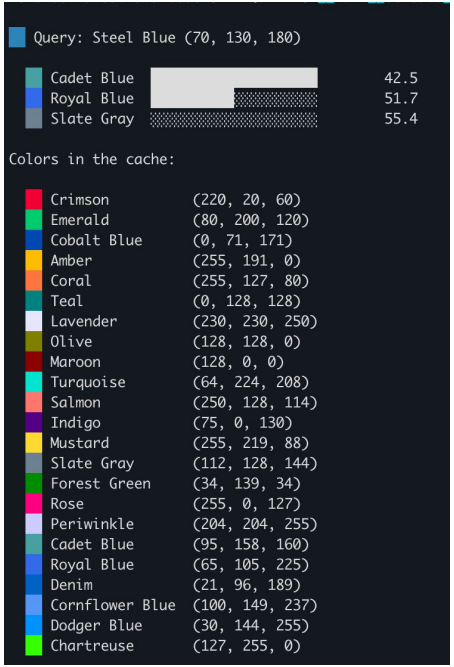
AWS Elasticache for Valkey



https://aws.amazon.com/blogs/database/building-secure-amazon-elasticache-for-valkey-deployments-with-terraform/



Thank you for listening!
Questions?



Connect with me!