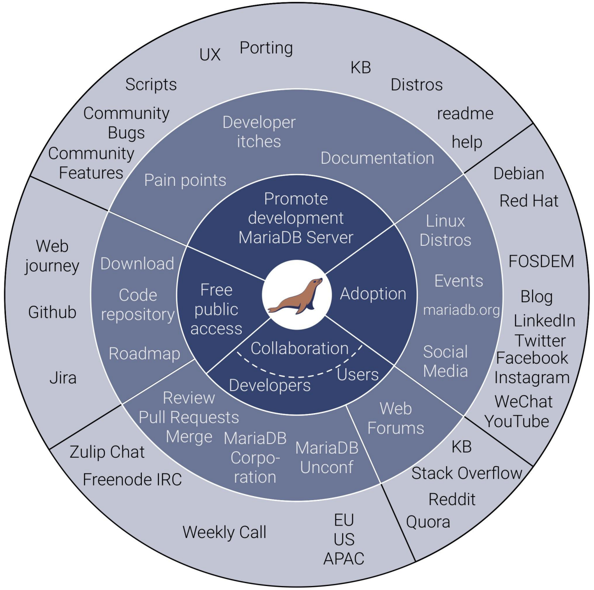


MariaDB Foundation

MariaDB Vector

Sergei Golubchik
Server Architect @ MariaDB plc
Vicențiu Ciorbaru
Chief Development Officer
@ MariaDB Foundation

FOSDEM 2024
MariaDB Fringe Event

What is an embedding model vs generative model?

- ChatGPT is a generative model.
  - It takes a prompt.
  - Generates the most likely "correct" sequence of words as response.

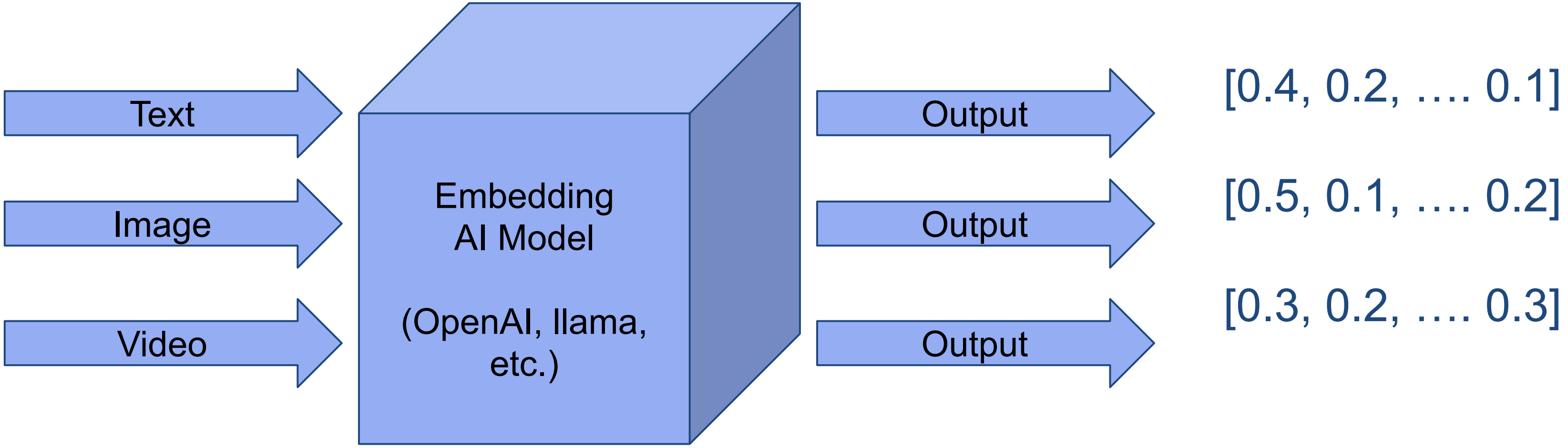
- An embedding model generates a vector embedding for a particular prompt.

What is a Vector Embedding?

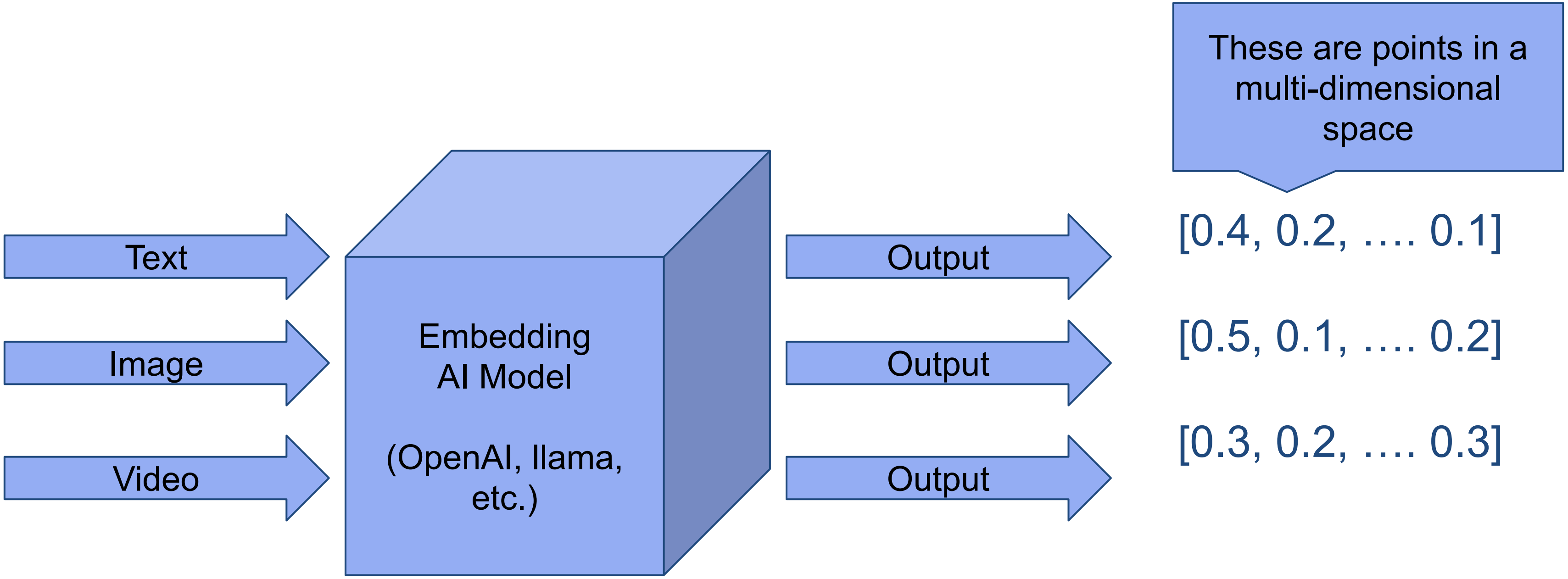
Simply a list of numbers (that describe “features” of the original)

What is a Vector Embedding?

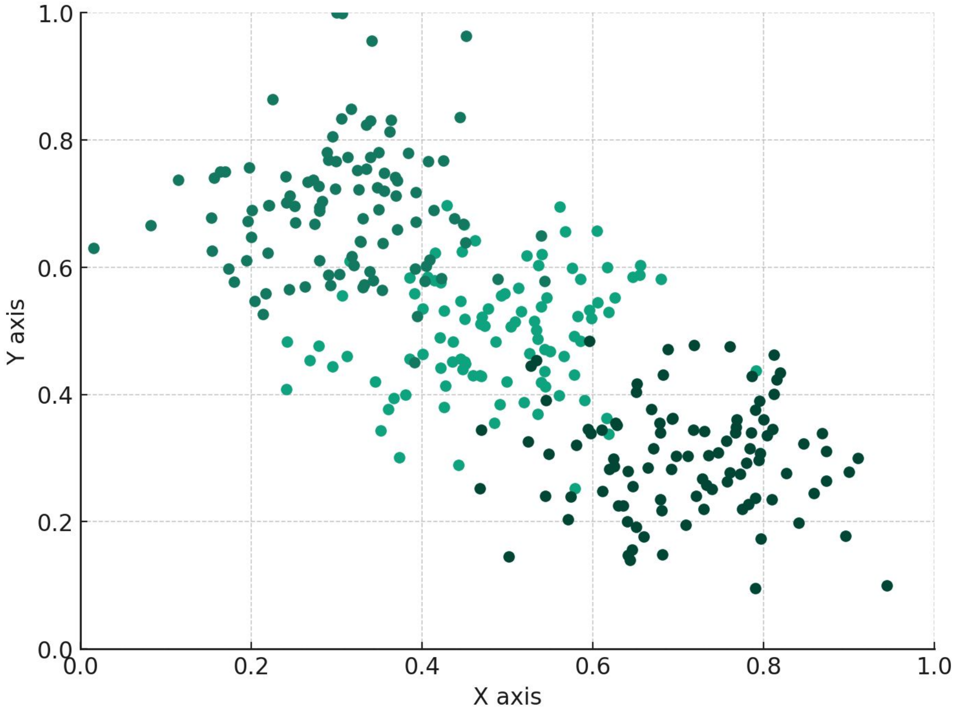
Simply a list of numbers (that describe “features” of the original)



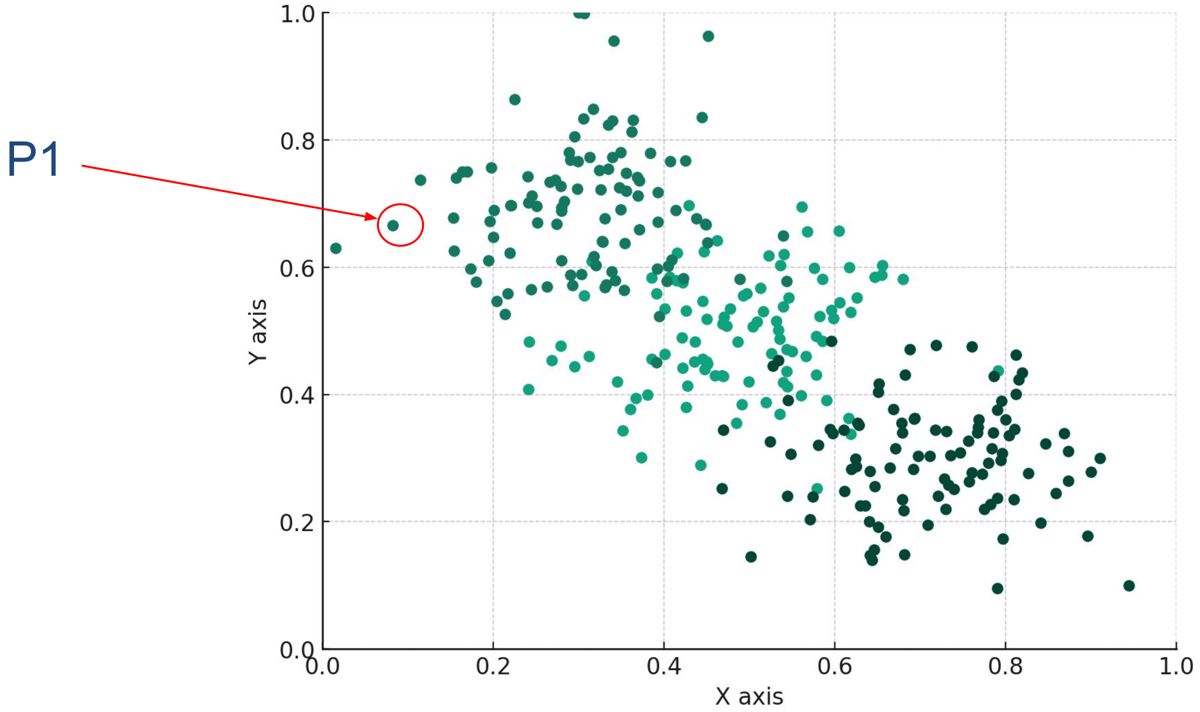
What is a Vector Embedding?



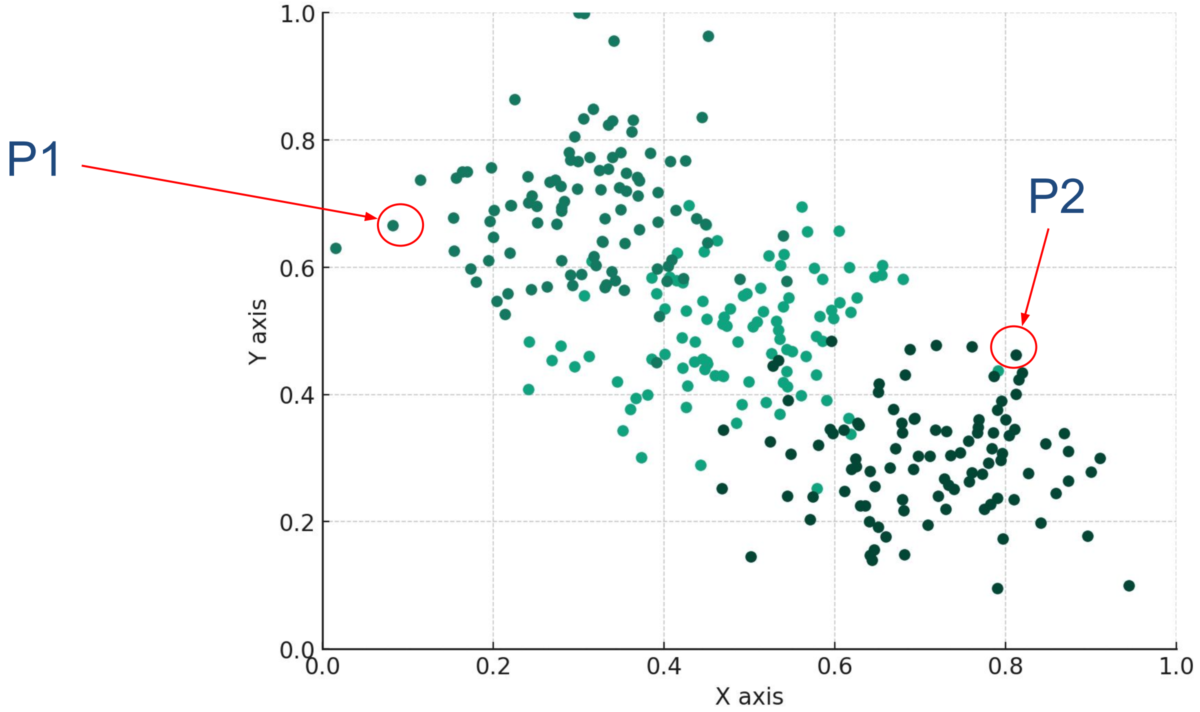
2D example



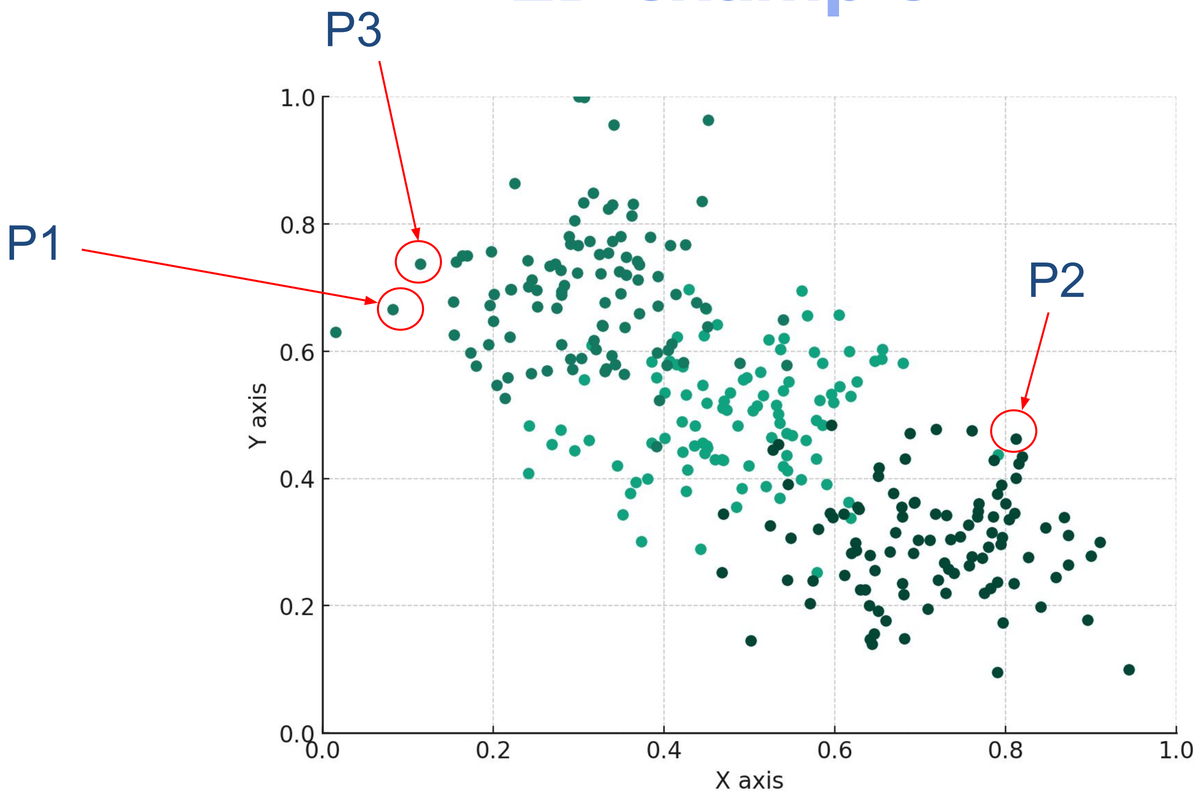
2D example



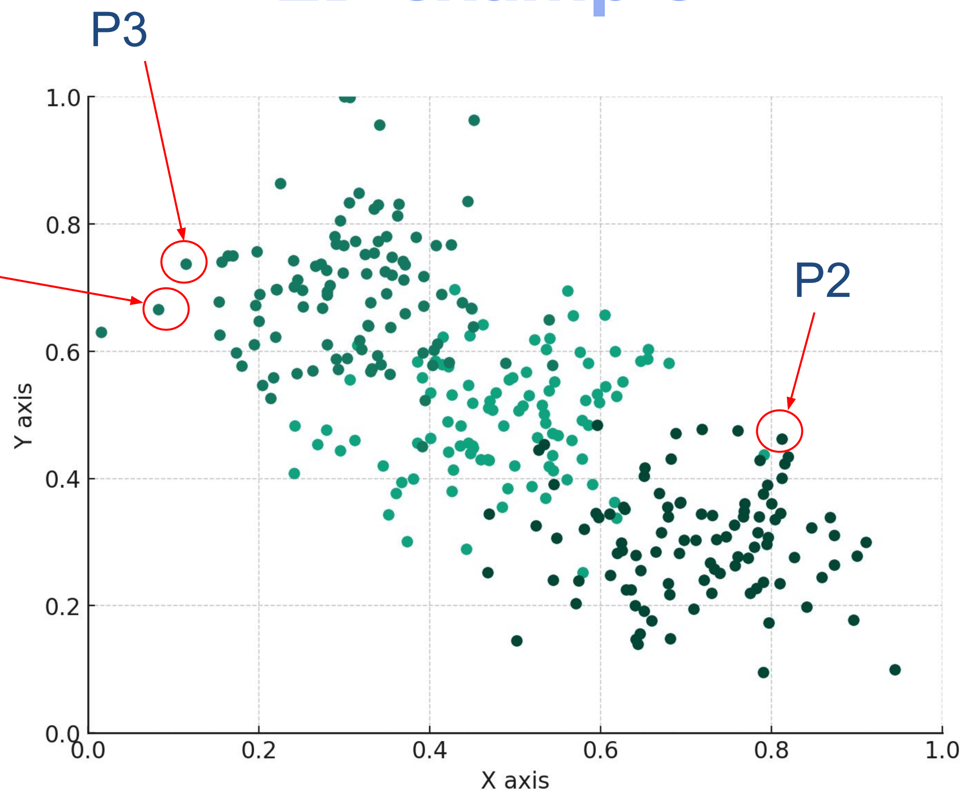
2D example



2D example



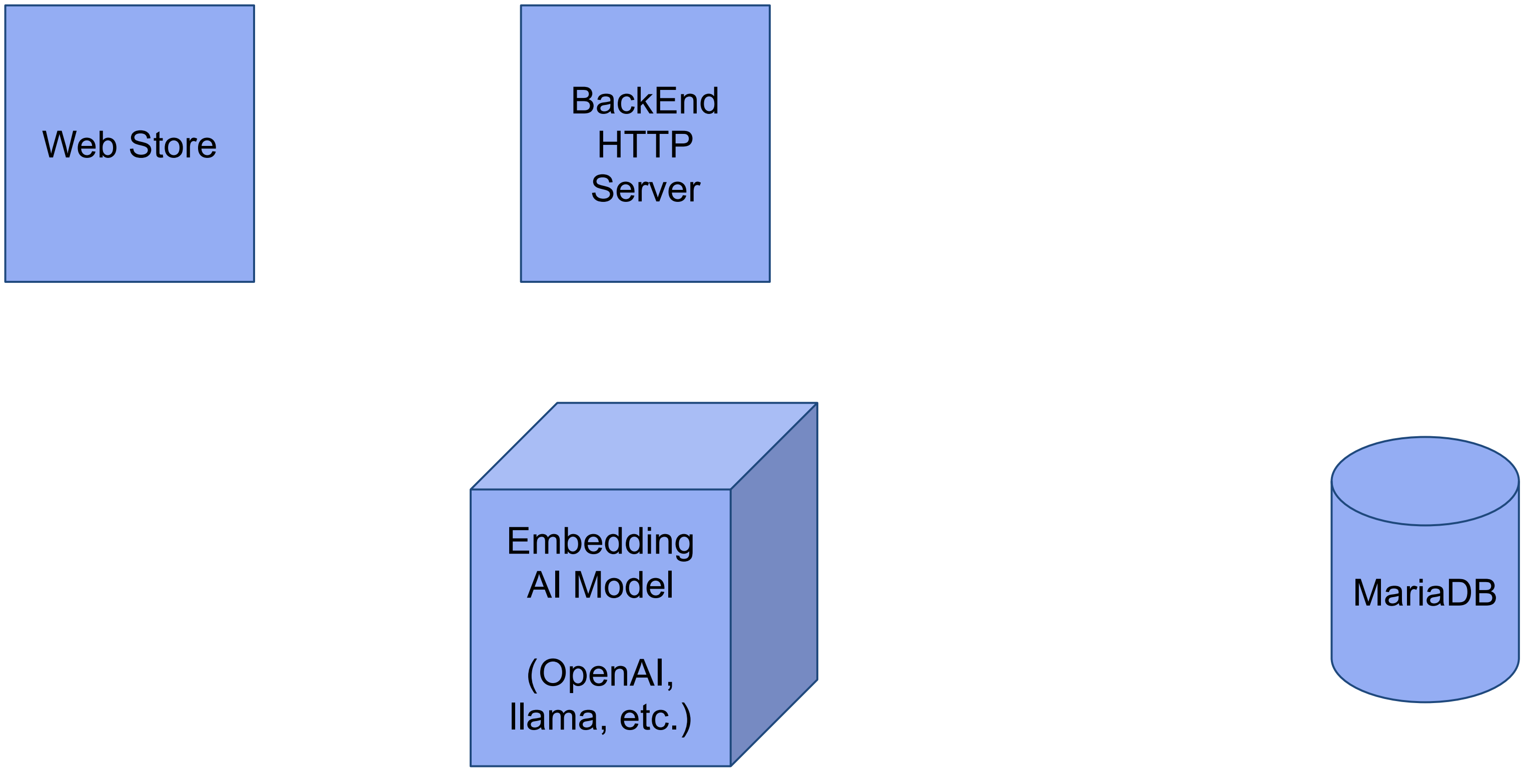
2D example



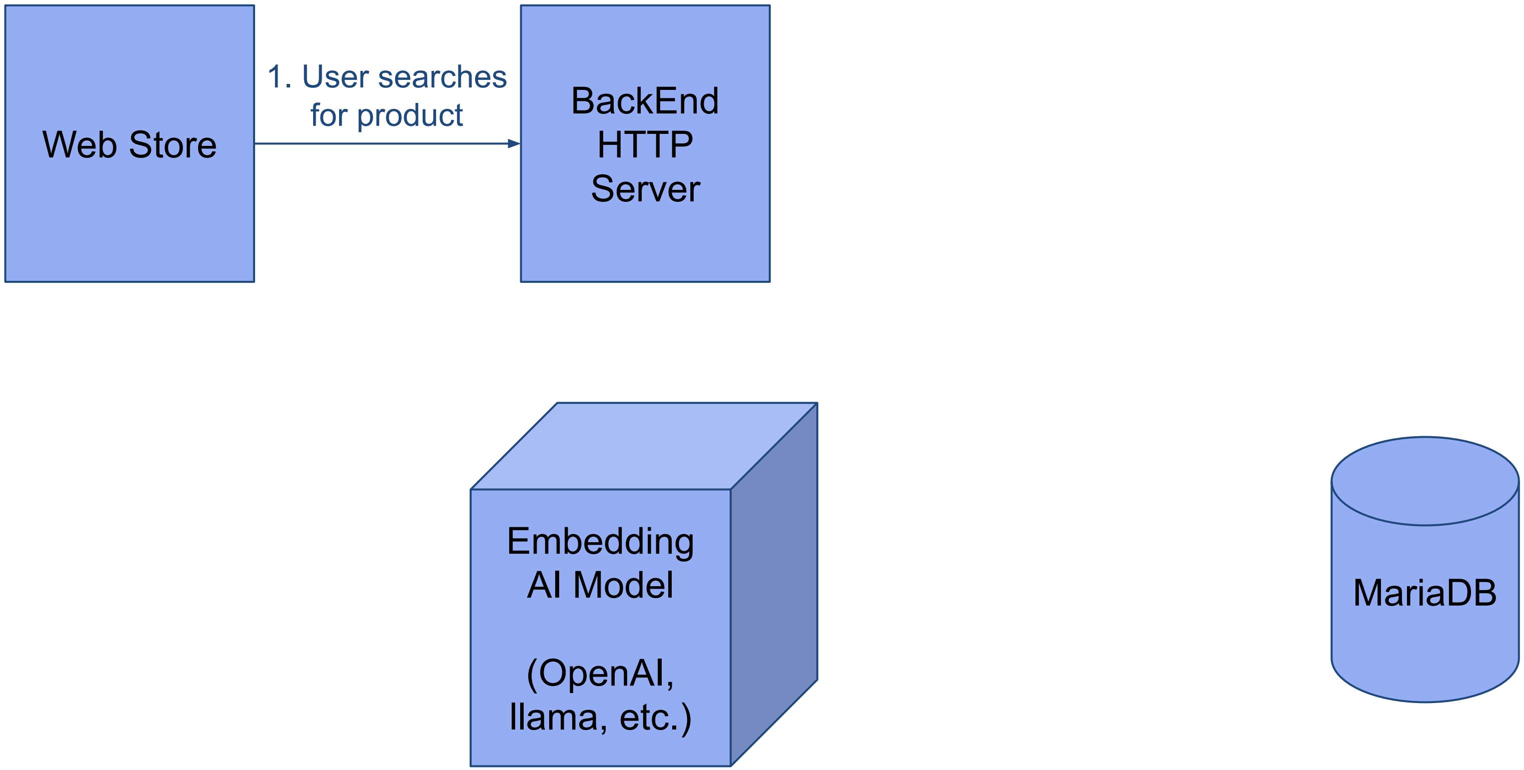
The document associated with P1 is more similar to P3 than P2,

according to the model that generated the points!

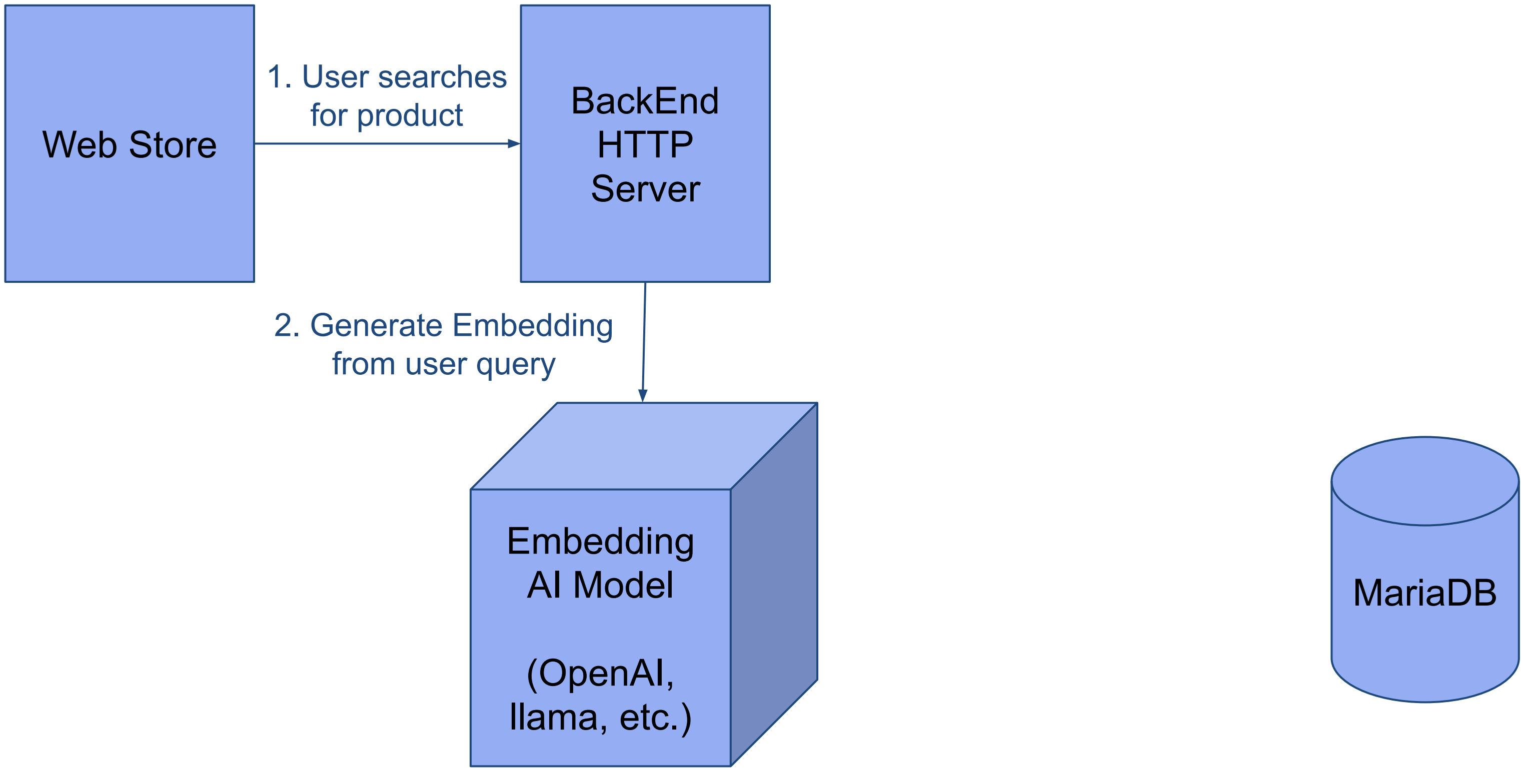
Where Vector search comes into play?



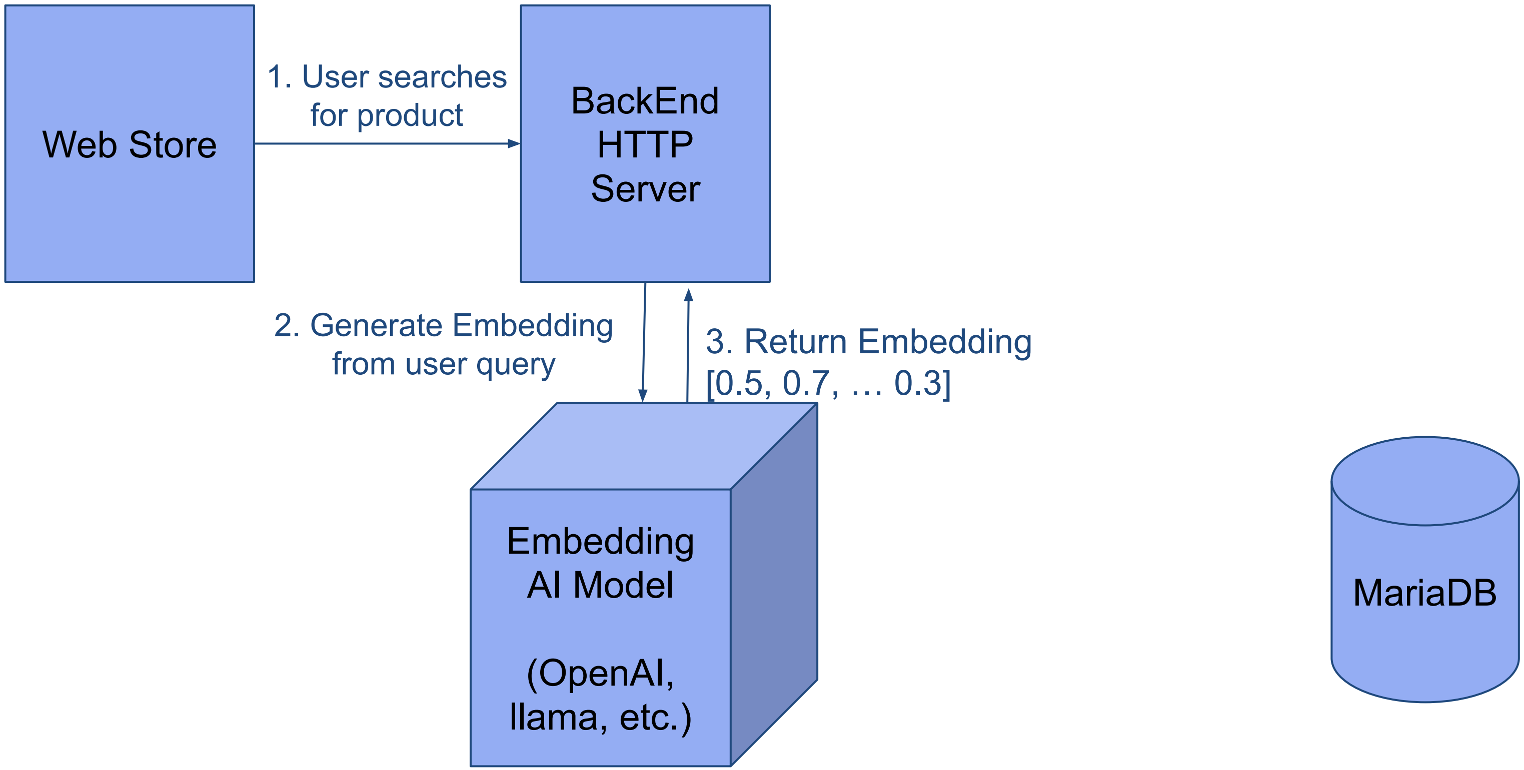
Where Vector search comes into play?



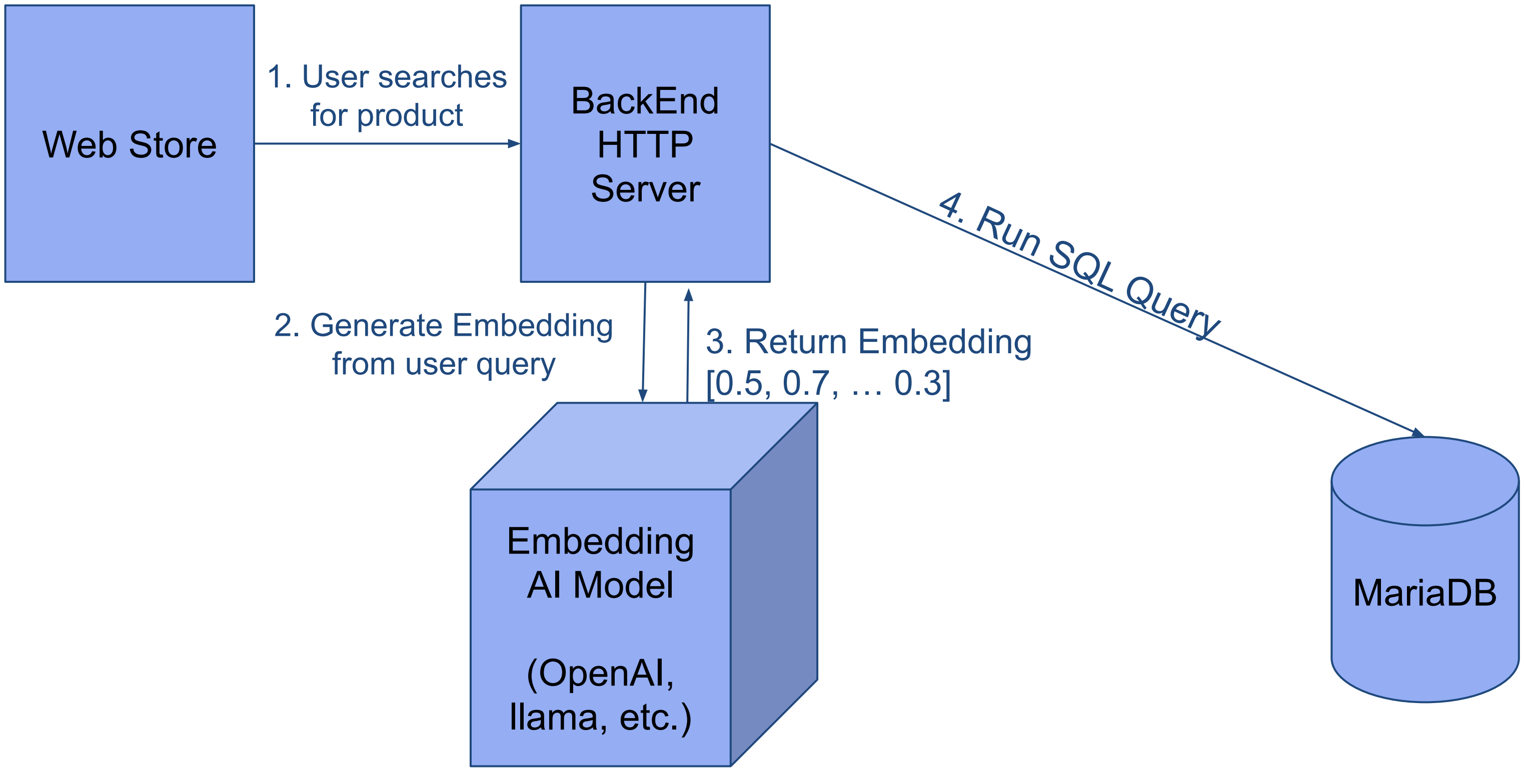
Where Vector search comes into play?



Where Vector search comes into play?

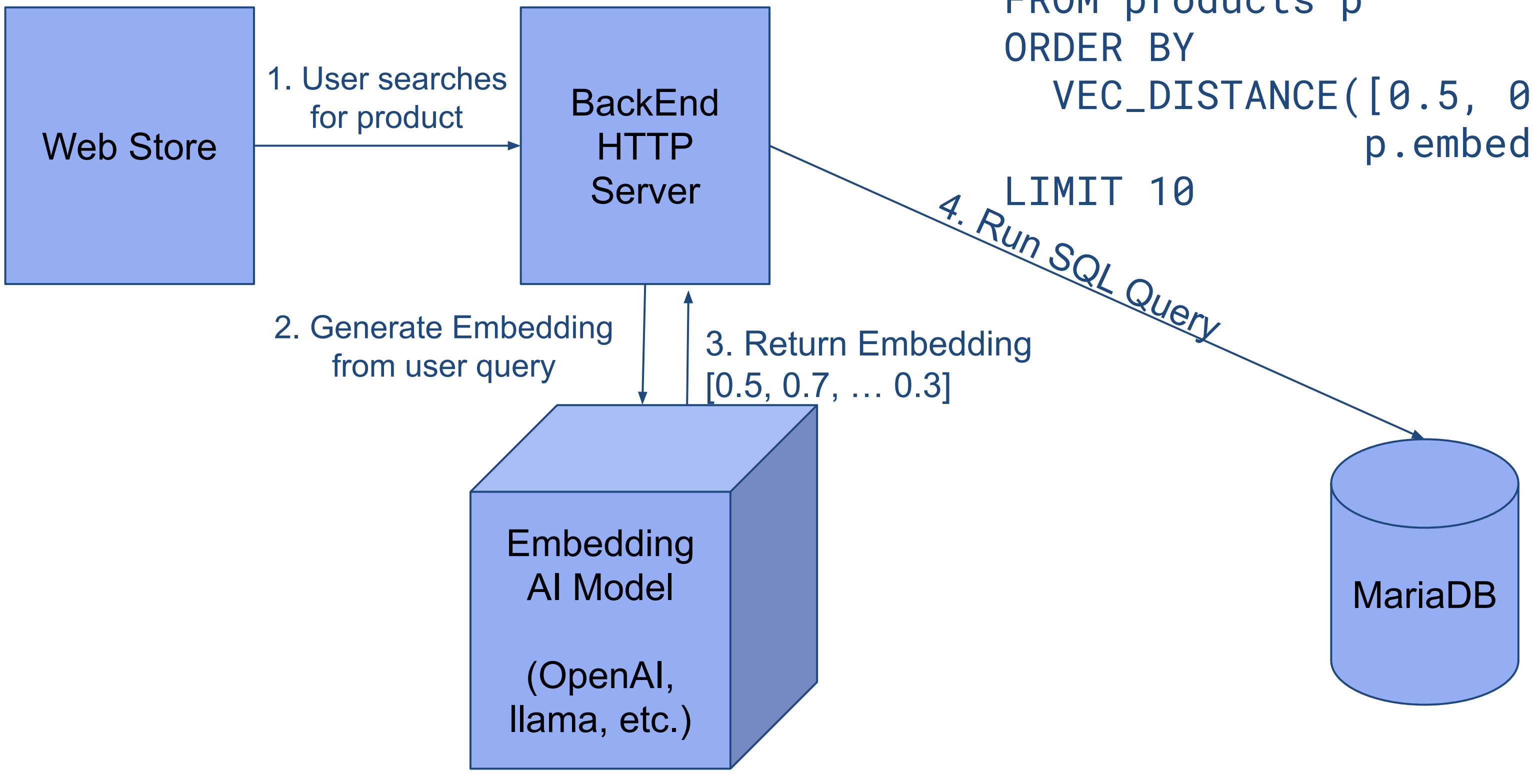


Where Vector search comes into play?



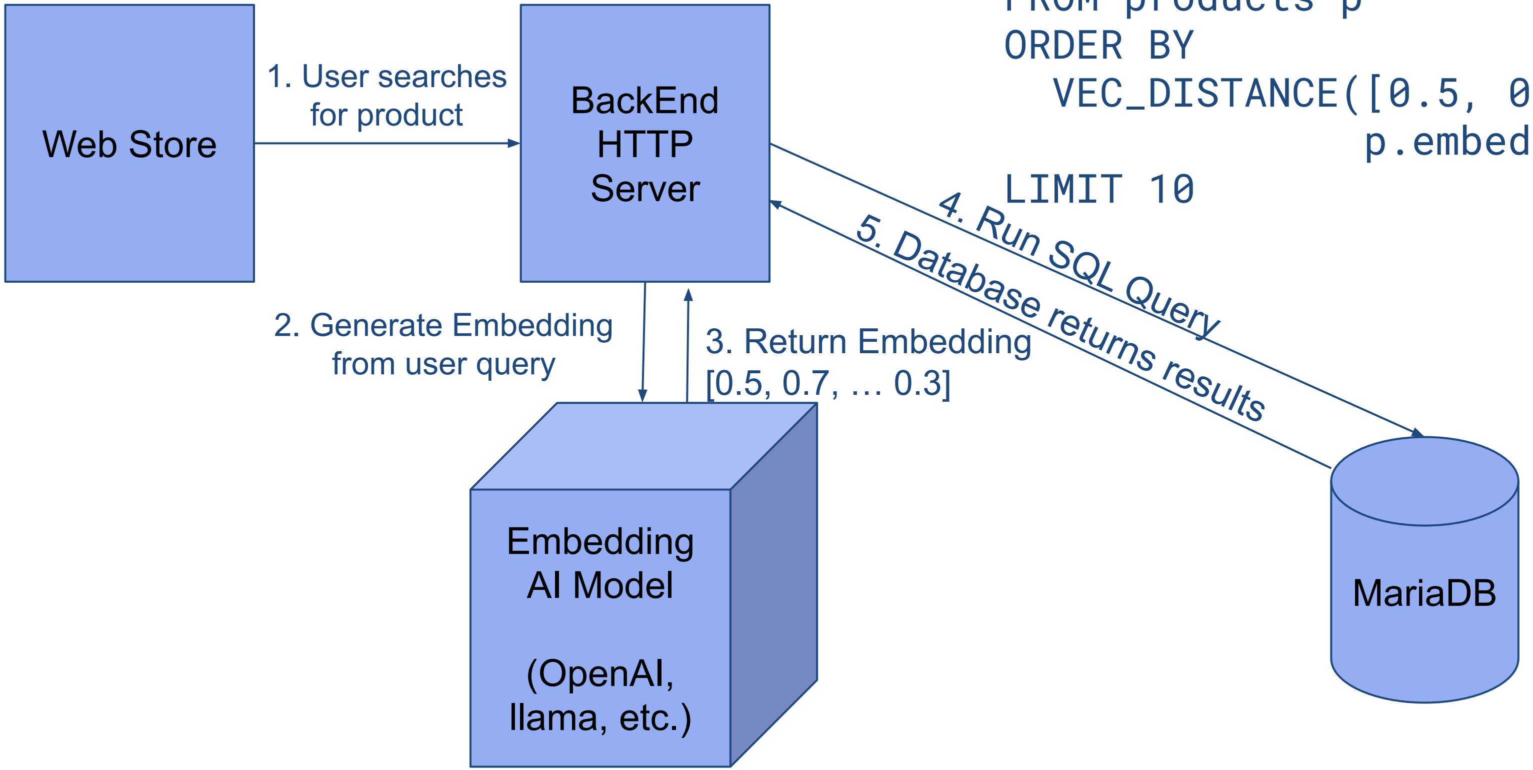
Where Vector search comes into play?

```
SELECT p.name, p.description
FROM products p
ORDER BY
  VEC_DISTANCE([0.5, 0.7, …, 0.3],
               p.embedding)
LIMIT 10
```

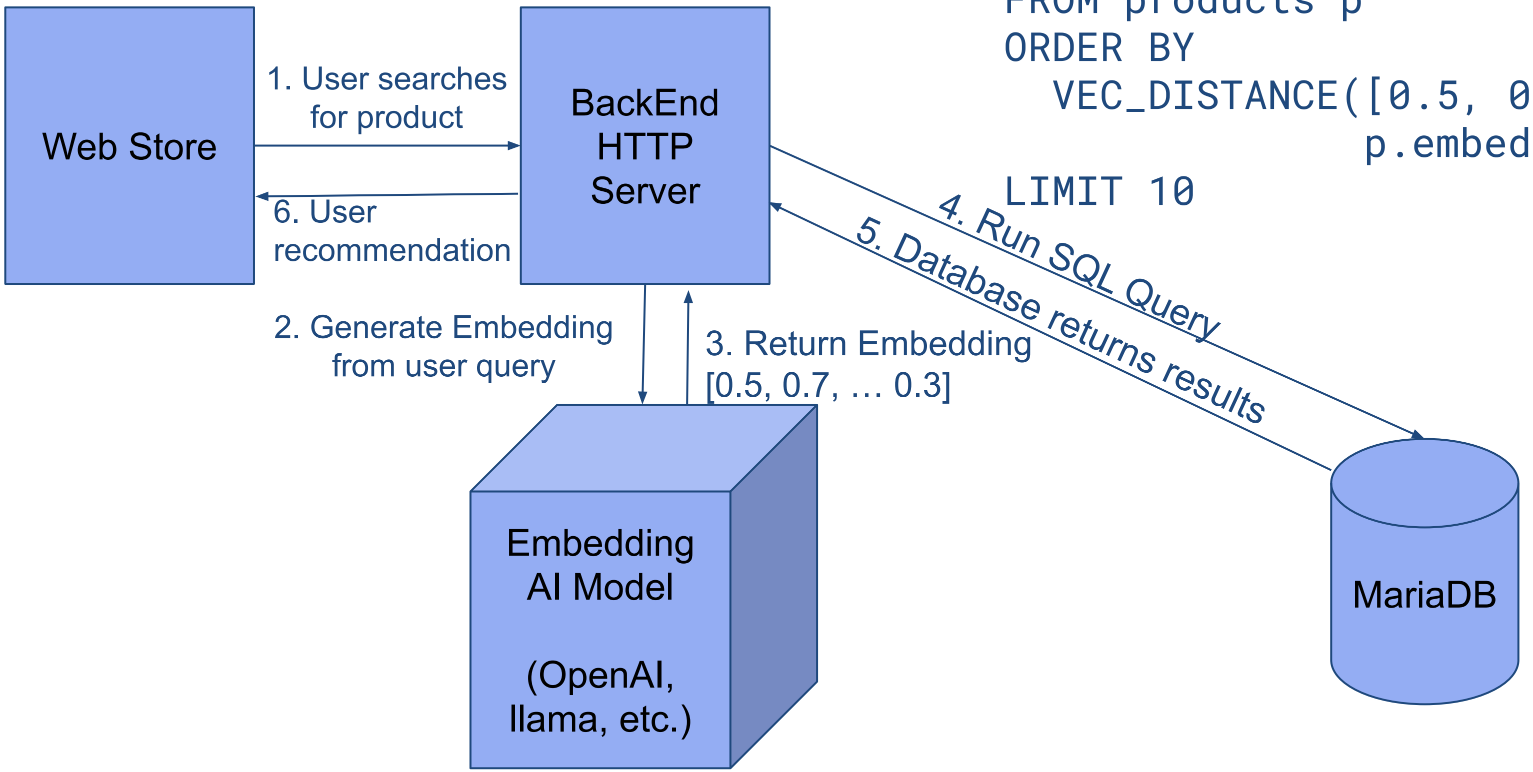


Where Vector search comes into play?

```sql
SELECT p.name, p.description
FROM products p
ORDER BY
  VEC_DISTANCE([0.5, 0.7, …, 0.3],
               p.embedding)
LIMIT 10
```



Where Vector search comes into play?




```
SELECT p.name, p.description
FROM products p
ORDER BY
  VEC_DISTANCE([0.5, 0.7, …, 0.3],
             p.embedding)
LIMIT 10
```

Other applications?

- Q&A systems based on documentation (Sergei demo)
- **Augment GPT prompts**
  - GPTs can handle a lot of context in a prompt
  - Use vector search to find the most relevant documents for a prompt.
  - Augment the prompt with the content of the documents.

For example, one could program prompts to be sent like this:

```
"""
Using only the following information:
[document_from_db1], [document_from_db2], ...]
Answer the following question:
<User query>
"""
```

As a database user, what must you do?

1. Install a vector database (MariaDB Vector will come with MariaDB Server soon)

2. Install an Embedding Model
   or
   Setup a cloud hosted model API.

3. Change your application to query the Embedding Model for each document insert and insert the embeddings into the database.

4. Make use of VEC_DISTANCE function to get the (approximate) nearest neighbors.

What's the catch?

1. Searching for vectors is expensive
2. Indexing strategies for vectors are only "approximate", they don't guarantee the exact "nearest" neighbor.
3. Depending on dataset, some indexing strategies perform better than others.
4. Indexing generally requires a lot of memory.
    a. IVFFlat – Low resource usage, poor search quality, present in pgvector
    b. **HNSW – Hierarchical Navigable Small Worlds**
        i. **de-facto industry standard.**
           Will be implemented in MariaDB
        ii. **Large memory usage.**

Possible future directions?

1. Plugins to generate embedding on insert.

2. Storage Engine for Vector Embeddings generation (CONNECT SE can fulfill this to some degree already)

3. More vector indexing algorithms.

4. Performance optimizations - Index Condition pushdown

Demo

Thank you!

Contact details:

vicentiu@mariadb.org

serg@mariadb.com

About:

https://mariadb.org/serg

https://mariadb.org/vicentiu