Smart Seek - Revolutionizing File Search with Vector Databases

TiDB Hackathon 2024

Shreyansh Rai & Rakshit Bang - IIIT Bangalore

Problem Statement

In an age of data, we needed an efficient way to find our files!
We Aim to Make Search Hassle-free!

- Smart Seek leverages advanced AI models and TiDB's vector database to enable accurate file retrieval through descriptive queries.
- Simply provide the path to the folders you want to be searchable, and sit back as we index them in the background to be searchable via just a Natural language description of the file you want!
- We use TIDB serverless with vector search to efficiently compute the files that best match your description.
- Each file type has specially crafted rules for vectorization (Embedding generation) via the most cutting edge Machine learning techniques.

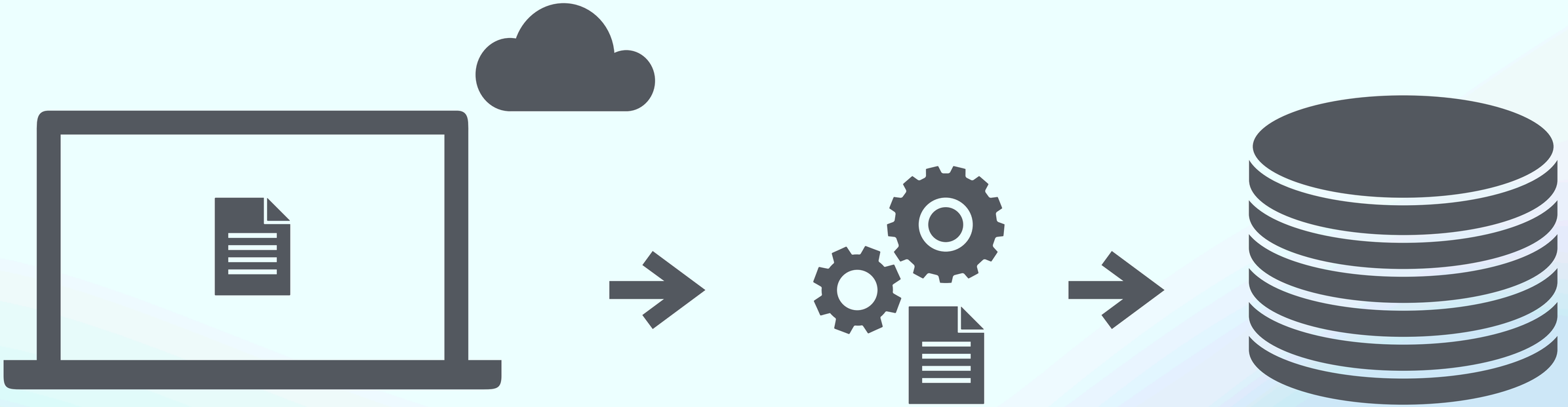
TIDB ❤️

Before moving on, a quick word of thanks to you!

- We have were heavily dependent on the TIDB serverless and vector search features and could not have done without it!
- TiDB Serverless separates the SQL layer and the storage layer, and makes the horizontal scalability of both layers transparent to our application. When we need computation resources, the computation nodes get assigned to us. More over Vector Search along with the caching benefits we get from TIDB were instrumental in making this application.
- We are on a mission to democratize the web, where all your cloud and local files can be search using Natural Language right on our application without any data redundancy made possible with TIDB!

Flow Overview I - Storage

How do we do it?



Specify the data

you want to be indexed

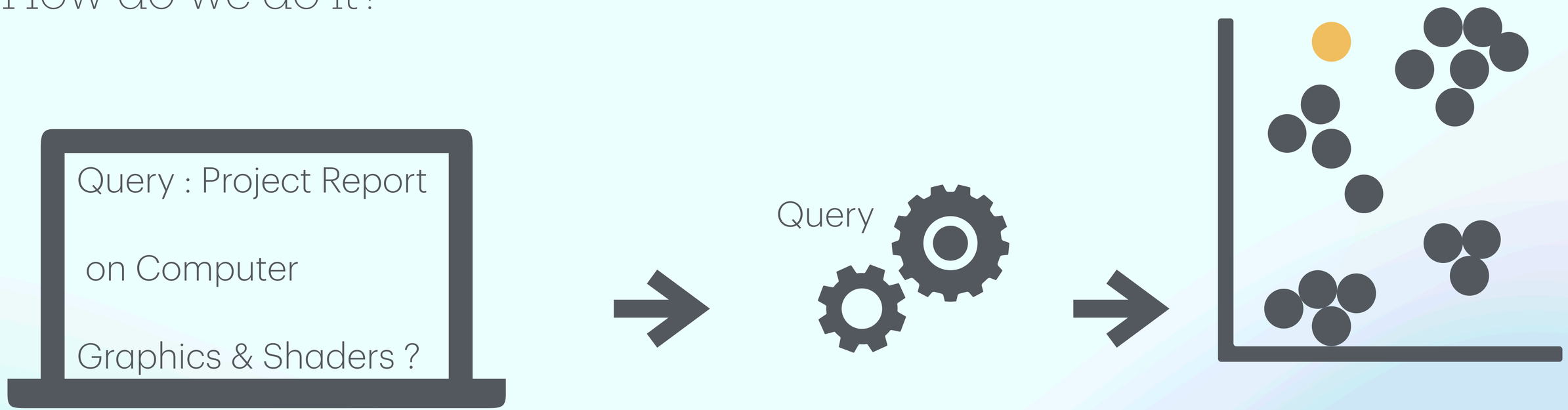
Your data remains where it was, We create embeddings

out of the data and store the vector <-> image

Mapping in TIDB’s vector database

Flow Overview II - Retrieval

How do we do it?



Describe the File

We convert the query to an Embedding and find the

File using TIDB vectors' excellent search functionality

Returning the best possible matches to your query.

A Quick Step-by-Step

What our POC is cable of

- Select folders to index by Smart Seek.
- Files are processed using captioning, embedding, and OCR models.
- Each major file type is handled a bit differently - For example a pdf need not be processed as an image, and an image might also contain text. Care has been taken to ensure you get the best result either way!
- Embeddings are stored in TiDB's vector database.
- User queries are matched with the most relevant files based on embeddings. Semantically similar categories of embeddings are closer to each other in the Embedding Space. Making excellent use of TIDB's serverless vector database for quick and scalable searches.
- The web app provides a chat like UI to describe and query the file you are looking for!

Future Scope

A path forward

- We have a POC ready that is able to index every file on your system but there is still some work to be done on increasing the compatibility with all the cloud storage providers out there to allow Smart Seek to be a one stop shop to search all the files that you have ever owned!

- We have also planned another extension to this app, called the Second Brain, that can figure out the file you are talking about and with a LLM - Chat like interface you can question it about any file that you have cloud or local.

- A more secure way to access including supporting multiple accounts on drives and local systems.

Thank You!