TECH OFFER

High-performant Vector Database for Artificial Intelligence (AI) Applications

KEY INFORMATION

TECHNOLOGY CATEGORY:
Infocomm - Artificial Intelligence
Infocomm - Big Data, Data Analytics, Data Mining & Data Visualisation
Infocomm - Data Processing

TECHNOLOGY READINESS LEVEL (TRL): TRL8
COUNTRY: SINGAPORE
ID NUMBER: TO174747

OVERVIEW

Machine Learning (ML) and Deep Learning (DL) have been the primary growth driver of Artificial Intelligence (AI) and has seen widespread adoption in areas such as Computer Vision, Speech Processing, Natural Language Processing, and Graph Search, among many others. It is also well-known that AI both needs and produces large amounts of data. However, traditional data repositories have not scaled effectively to handle the large amounts of vector representations that are common in AI applications - in such cases, searching for similarities across high-dimensional vectors is inefficient. To address such limitations, vector databases have been developed to address the limitations of traditional hash-based searches and search scalability, enabling similarity searches across large datasets.

This technology offer is a unified Online Analytical Processing (OLAP) data platform that supports approximate vector search, enabling efficient searching over billion-scale structured data and vector data. The data engine simplifies the process of building

enterprise-level AI applications such as search and recommendation systems, video analytics, text-based searches, and chatbots while accelerating the development of production-ready systems. Developers no longer need to deal with complicated scripts to query vector data as low latency, high-performance structured data, and vector data searches are made possible via vector data indexing methods and the use of extended Structured Query Language (SQL) syntax.

TECHNOLOGY FEATURES & SPECIFICATIONS

This technology offer is purpose-built OLAP database, CPU-only implementation with a built-in vector query engine that uses extended SQL statements for data querying. Supported data include structured data (tabular text, numbers, dates, times) and unstructured data (image, video, audio) that have been converted to vector data representation. This technology enables high-performance joint queries, and a simplified manner of querying labels, text, and numbers within a single SQL statement. It supports highly performant SQL + vector searches, operating on billion-scale data, with an operating latency of 200 milliseconds at a throughput of 200 queries per second (QPS).

The key features of this technology are as follows:

- **Fast query performance**
  Column-oriented storage; data is stored in the same column and compression techniques are applied to reduce disk usage and save I/O resources
- **Linear scalability**
  Data is stored evenly across nodes, ensuring scalability
- **Simultaneous data input**
  Data can be inserted simultaneously via random data distribution
- **Concurrent queries**
  Simultaneous insertion and querying

The following similarity metrics are currently supported:

- Euclidean
- Cosine Similarity
- Dot Product

The following indexing libraries are currently supported:

- Facebook AI Similarity Search (FAISS)
- hnswlib (with proprietary optimisations)

The following interfaces are available for developer integration:

- C++/Java/Python language Software Development Kit (SDK)
- SQL interface
- Web User Interface (UI)

POTENTIAL APPLICATIONS

This technology can be applied for similarity searches (identifying similar high-dimensionality vectors), or classification (locating

images that contain a certain element, e.g. car, flower). The following potential applications of this technology have also been tested:

- Biometrics (fingerprint matching)
- DNA/genetic sequences and other biomedical fields (similarity search/classification)
- Multimedia - image, video, and audio (similarity search)
- Text-based - recommendation systems, chatbots (similarity search)
- Molecular (similarity search)
- Trademarks (similarity search)
- Commodities
- GIS vectors (vector-based semantic analysis)

UNIQUE VALUE PROPOSITION

Compared with existing techniques, this technology represents a single, unified pipeline for querying vector representation data without the need to store structured data and vectors separately in traditional databases (SQL) and vector repositories. This solves the limitation of having to merge results from standard database engines (specifically optimised for hash-based searches) with that of vector query databases. This data engine includes a vector search function and it can efficiently store, index, and manage vectors that are generated by deep learning networks and machine learning models. Additionally, the extended SQL query syntax of this technology enables a highly efficient, simplified search across a variety of different AI applications.

The technology owner is keen to collaborate with companies that are conducting in-house AI application development in industries such as, but not limited to, e-commerce, video analytics, smart city, and healthcare.

The following is an example of how the vector search engine can be used to query for similar logos (images):

1. A large dataset of logos is prepared
2. A feature extraction model is trained from the dataset e.g. DarkNet-53, VCG, NASNet-Large, Inception-ResNet-V2, etc
3. Logos are converted into vectors using the trained model
4. Vectors of logos within are stored in the database
5. Any new logo is put through the trained model to generate a vector for similarity search against the pool of vectors