SPONSORED CONTENT | WHITE PAPER

GenAI with vector search: Create new ways to get more value from data

CIO

SPONSORED BY ORACLE

Since the introduction of the first commercial relational database system in 1977, CIOs and IT teams have striven to provide business users with more readily accessible data. Database technology has since evolved to encompass unstructured data, objects, distributed implementations, and cloud services. However, users still struggle with information searches due to the complexity of creating appropriate queries — or simply because they're not sure of what they are searching for.

Advancements in generative artificial intelligence (GenAI) combined with the ability to represent the meaning of unstructured data as vectors promise to simplify the search process, deliver better results through similarity searches, and respond to queries in natural language.

IT leaders striving to provide their organization with greater access to data and richer context must decide whether to invest in a specialized vector database or use these new data types and GenAI tools to enhance an existing converged database.

Rapid context-based search

Vectors are a new data type that, when combined with the embedding models used to create each vector, enables rapid semantic search of unstructured data that is represented by strings of numbers. These vectors enable organizations to add more context and context-specific data sources to the large language models (LLMs) that power GenAI.

Among the ways vector search can improve user experiences and business processes:

- Answering questions about a document by identifying critical portions of it and similar documents that can provide additional information
- Recommending similar products in response to an online shopping search
- Understanding sentiment of online comments to shape customer-facing messaging

- Detecting fraud by identifying complex schemes based on multiple transactions

Vector embedding models are a critical technology that enables vector search. They encapsulate complex data modalities, from words and sentences to entire documents, images, and more, converting them into a series of numbers. This enables rapid semantic search of the unstructured data the vectors represent by determining the meaning of queries and source material and retrieving more accurate results based on that meaning — a significant improvement over keyword searches.

Retrieval Augmented Generation, vector databases, and LLMs

Vector search can also be used to improve the output of LLMs through a technique called Retrieval Augmented Generation (RAG), a relatively new AI technique that enables LLMs to tap additional data resources without retraining.

RAG is a type of GenAI that enables organizations to guide LLMs using domain-specific, private, or recently created information. By augmenting users' original prompts with additional information identified through similarity searches, RAG enables more timely, accurate, and contextually relevant responses.

Relying solely on general-purpose LLMs to answer questions can lead to GenAI "hallucinations" that generate nonsensical responses or answers that seem logical but are not accurate. Common problems that can lead to hallucinations include:

- Prompt-reference divergence, when input prompts and training data for the LLM don't overlap

- Temporal mismatches due to the LLMs including data only up to the time training was completed

- Reliance on public data that can create responses based on industry trends rather than company-specific information

Encoding domain-specific, private, or recent data in vector databases and then using RAG to tap into these knowledge repositories can help address these and other issues associated with LLMs. For instance, a research institute could encode domain-specific information on molecular biology into a vector database and use RAG to help make the responses to questions on the topic more accurate than would otherwise be possible.

Limitations of stand-alone vector databases

When new data management technologies such as vector databases first come out, they typically are designed to solve a very narrow, highly differentiated problem. Stand-alone vector databases are no different, typically implementing security models, performance optimizations, and access methods that are different from those of the enterprise databases used to store other organizational data. These newly developed databases have not been hardened over decades of development and may not meet the needs of enterprises with business-critical applications.

Furthermore, stand-alone vector databases require additional integration steps and external processing to make them work. This means that they will forever need to be managed and secured by IT teams in parallel to enterprise databases and that developers must implement and maintain integrations.

Converged databases with integrated vector search

Adding vector data types and vector search capabilities to a converged enterprise database helps ensure that vector representations of various types of data are consistent with the data they represent.

This simplifies the development of applications and RAG that need to access vectors as well as their underlying data.

Unlike stand-alone vector databases, a converged database with integrated vector search provides crucial attributes necessary for today's enterprise, including:

- Native support for **modern data types** such as relational for business data, JSON documents for collections of objects, spatial data for location awareness, Internet of Things (IoT) data for streaming devices, and graph structures for connected data
- Support for **multiple workloads**, including up-to-date transactional data, analytics on historical or real-time data, blockchain data provenance, and machine learning to train models and draw inferences from new data
- Suitability for **multiple development paradigms** such as traditional programming languages, microservices and events, REST APIs, and software development kits (SDKs)

Unified data and vector search

Oracle offers a variety of advanced cloud-based GenAI services, including RAG, and leverages these capabilities in its portfolio of Software-as-a-Service offerings. Oracle Database has recently added AI Vector Search to extend and enhance these capabilities with support for a vector data type, vector indexes, and vector search SQL operators.

These new capabilities enable Oracle Database to store the semantic content of documents, images, and other unstructured data as vectors and use these to run fast similarity search queries. Organizations can take advantage of:

- Native support for storing and processing vectors, which are just one more data type supported by Oracle Database and are processed with familiar SQL statements

- The ability to generate vectors from user-selected embedding models inside Oracle Database

- Indexing based on vector characteristics, enabling extremely fast similarity searches

With Oracle Database, vectors reside inside the same database as the enterprise's corporate data, irrespective of its type (relational, document, text, spatial, etc.). Storing data and the vectors representing it in the same database helps eliminate time-consuming data copying and reduce data fragmentation. Open source frameworks such as LangChain and LlamaIndex can be used to implement GenAI solutions, including ones using RAG.

Oracle AI Vector Search provides users with easier, faster, and more precise search across business and semantic data without requiring individuals to become highly skilled in the nuances of creating, indexing, and searching vectors. It enables rapid semantic search of the unstructured data that vectors represent, a critical need for workloads that require quick responses, such as identifying manufacturing quality issues based on images taken on an assembly line.

When used with RAG, Oracle Database AI Vector Search enables organizations to augment LLM searches with enterprise-specific data rather than relying solely on outdated and generalized training data. This results in greater accuracy based on more current data, industry-specific information, or organization-specific content. Importantly, the use of RAG helps improve results without training LLMs on private data and is less expensive than continuously fine-tuning LLMs.

Optimizing vector search with Oracle Exadata and AMD EPYC processors

Oracle Exadata platforms using AMD EPYC processors deliver the highest-possible performance and availability for Oracle Database workloads, including those employing AI Vector Search.

With AMD EPYC processors, Exadata platforms start out with more than 250 database cores available in database servers and can be scaled to provide more than 4,000 in a single system. These systems offer terabytes of extremely fast memory for high-speed in-memory vector indexes and can support hundreds to thousands of concurrent queries, supporting the vector search needs of an entire organization. They enable organizations to combine vector search results with traditional database functions — including data-intensive joins that can be offloaded to smart storage servers that can store hundreds of terabytes to petabytes of data.

Many organizations have hundreds or thousands of Oracle Database instances, many of which could benefit from the addition of AI Vector Search capabilities. The scalability and versatility provided by Exadata platforms powered by AMD EPYC processors enable these organizations to lower their costs by consolidating diverse Oracle Database workloads on a small number of Exadata platforms.

Exadata platforms are available in Oracle Cloud Infrastructure (OCI) as part of Oracle Database@Azure, a distributed cloud platform that runs OCI Oracle Database services wherever customers need them and as an engineered system for deployment in customer data centers.

Conclusion

GenAI in the form of LLMs is taking the IT market by storm. Organizations of all sizes and types and in all disciplines are looking for ways to use GenAI to improve personal productivity and customer experiences. Enterprises wanting to overcome the limitations of LLMs are increasingly turning to vector search and RAG to identify semantically similar content and increase response accuracy.

Oracle has added AI Vector Search capability to its flagship Oracle Database so organizations can easily add vector search capabilities to existing applications and create new ones without the complexities engendered by stand-alone vector databases.

Oracle has also optimized vector search on Oracle Exadata, through the use of software optimizations and high-speed AMD EPYC processors. Exadata's highly parallel and scalable database and storage servers enable high performance for creating vector embeddings and indexes or performing searches.

In short, Oracle has extended its powerful Oracle Database with easy-to-use vector capabilities that are simple to adopt and use throughout the enterprise.

To learn more, visit the Oracle Database AI Vector Search and Oracle Exadata web pages.