InfoWorld

by **Isaac Sacolick**
Contributing Writer

Vector databases in LLMs and search

Feature

Nov 6, 2023 • 8 mins

Vector databases and search aren’t new, but vectorization is essential for generative AI and working with LLMs. Here's what you need to know.

Credit: Thinkstock

One of my first projects as a software developer was developing genetic analysis algorithms. We built software to scan **electrophoresis** [https://www.yourgenome.org/facts/what-is-gel-electrophoresis/] samples into a database, and my job was to convert each DNA pattern's image into representable data. I did this by converting the image into a vector, with each point representing the attributes of the sample. Once vectorized, we could store the information efficiently and calculate the similarity between DNA samples.

Converting unstructured information into vectors is commonplace today and used in **large language models** [https://www.infoworld.com/article/2338762/what-can-chatgpt-and-llms-really-do-for-your-business.html] (LLMs), image recognition, **natural language processing** [https://www.infoworld.com/article/2260903/what-is-natural-language-processing-ai-for-speech-and-text.html], recommendation engines, and other **machine learning** [https://www.infoworld.com/article/2254843/what-is-machine-learning-intelligence-derived-from-data.html] use cases.

Vector databases [https://www.infoworld.com/article/2338822/10-ways-generative-ai-upends-the-traditional-database.html] and **vector search** [https://www.infoworld.com/article/2269766/what-is-vector-search-better-search-through-ai.html] are the two primary platforms developers use to convert unstructured information into vectors, now more commonly called embeddings. Once information is coded as an embedding, it makes storing, **searching** [https://blogs.starcio.com/2023/06/roi-ai-search-personalization.html], and comparing the information easier, faster, and significantly more scalable for large datasets.

"In our pioneering journey through the world of vector databases, we've observed that despite the buzz, there is a common underestimation of their true potential," says Charles Xie, CEO of **Zilliz** [https://zilliz.com/]. "The real treasure of vector databases is their ability to delve deep into the immense pool of unstructured data and unleash its value. It's important to realize that their role isn't limited to memory storage for LLMs, and they harbor transformative capacities that many are still waking up to."

How vector databases work

Imagine you're building a search capability for digital cameras. Digital cameras have dozens of attributes, including size, brand, price, lens type, sensor type, image resolution, and other features. One **digital camera search engine** [https://www.dpreview.com/products/search/cameras#!] has 50 attributes to search over 2,500 cameras. There are many ways to implement search and comparisons, but one approach is to convert each attribute into one or more data points in an embedding. Once the attributes are vectorized, vector distance formulas can calculate product similarities and searches.

Cameras are a low-dimensionality problem, but imagine when your problem requires searching **hundreds of thousands of scientific white papers** [https://zilliz.com/blog/Arxiv-scientific-papers-vector-similarity-search] or **providing music recommendations on over 100 million songs** [https://aws.amazon.com/blogs/big-data/amazon-opensearch-services-vector-database-capabilities-explained/]. Conventional search mechanisms break down at this scale, but vector search reduces the information complexity and enables faster computation.

"A vector database encodes information into a mathematical representation that is ideally suited for machine understanding," says Josh Miramant, CEO of **BlueOrange** [https://blueorange.digital/]. "These mathematical representations, or vectors, can encode similarities and differences between different data, like two colors would be a closer vector representation. The distances, or similarity measures, are what many models use to determine the best or worst outcome of a question."

Use cases for vector databases

One function of a vector database is to simplify information, but its real power is building applications to support a wide range of natural language queries. Keyword search and advanced search forms simplify translating what people search into a search query, but processing a natural language question offers a lot more flexibility. With vector databases, the question is converted into an embedding and used to perform the search.

For example, I might say, "Find me a midpriced SLR camera that's new to the market, has excellent video capture, and works well in low light." A transformer converts this question into an embedding. **Vector databases commonly use encoder transformers** [https://medium.com/@david.gutsch0/the-art-of-embeddings-transforming-text-for-vector-databases-926738443e70]. First, the developer tokenizes the question into words, then uses a transformer to encode word positions, add relevancy weightings, and then create abstract representations using a feed-forward neural network. The developer then uses the question's finalized embedding to search the vector database.

Vector databases help solve the problem of supporting a wide range of search options against a complex information source with many attributes and use cases. LLMs have spotlighted the versatility of vector databases, and now developers are applying them in language and other information-rich areas.

- What is the difference between vector databases and traditional databases?
- How do vector databases help in building chatbots and voice assistants?
- Can vector databases be used for image recognition and recommendation engines?
- What are vector databases and how do they work?
- How do vector databases support real-time search and analytics capabilities?

Ask a question

ASK

"Vector search has gained rapid momentum as more applications employ machine learning and artificial intelligence to power voice assistants, chatbots, anomaly detection, recommendation and personalization engines, all of which are based on vector embeddings at their core," says Venkat Venkataramani, CEO of **Rockset [https://rockset.com/]**. "By extending real-time search and analytics capabilities into vector search, developers can index and update metadata and vector embeddings in real-time, a vital component to powering similarity searches, recommendation engines, generative AI question and answering, and chatbots."

Using vector databases in LLMs

Vector databases enable developers to build specialty language models, offering a high degree of control over how to vectorize the information. For example, developers can build generic embeddings to help people search all types of books on an ecommerce website. Alternatively, they can build specialized embeddings for historical, scientific, or other special category books with domain-specific embeddings, enabling power users and subject matter experts to ask detailed questions about what's inside books of interest.

"Vector databases simply provide an easy way to load a lot of unstructured data into a language model," says Mike Finley, CTO of **AnswerRocket [https://www.answerrocket.com]**. "Data and app dev teams should think of a vector database as a dictionary or knowledge index, with a long list of keys (thoughts or concepts) and a payload (text that is related to the key) for each of them. For example, you might have a key of 'consumer trends in 2023' with a

payload containing the text from an analyst firm survey analysis or an internal study from a consumer products company."

Choosing a vector database

Developers have several technology options when converting information into embeddings and building vector search, similarity comparisons, and question-answering functions.

"We have both dedicated vector databases coming to the market as well as many conventional general-purpose databases getting vector extensions," says Peter Zaitsev, founder of **Percona [https://www.percona.com/]**. "One choice developers face is whether to embrace those new databases, which may offer more features and performance, or keep using general purpose databases with extensions. If history is to judge, there is no single right answer, and depending on the application being built and team experience, both approaches have their merits."

Rajesh Abhyankar, head of the Gen AI COE at **Persistent Systems [https://www.persistent.com/]**, says, "Vector databases commonly used for search engines, chatbots, and natural language processing include Pinecone, FAISS, and Mivus." He continues, "Pinecone is well-suited for recommendation systems and fraud detection, FAISS for searching image and product recommendations, and Milvus for high-performance real-time search and recommendations."

Other vector databases include Chroma, LanceDB, Marqo, Qdrant, Vespa, and Weaviate. Databases and engines supporting vector search capabilities include Cassandra, Coveo, Elasticsearch OpenSearch, PostgreSQL, Redis, Rockset, and Zilliz. Vector search is a capability of **Azure Cognitive Search [https://learn.microsoft.com/en-us/azure/search/vector-search-overview]**, and Azure has **connectors for many other vector databases [https://learn.microsoft.com/en-us/semantic-kernel/memories/vector-db]**. AWS supports several **vector database options [https://aws.amazon.com/what-is/vector-databases/]**, while Google Cloud has **Vector AI Vector Search [https://cloud.google.com/vertex-ai/docs/vector-search/overview]** and connectors to other vector database technologies.

Vector databases and generative AI risks

Using vector databases and search brings with it a few common **generative AI risks [https://blogs.starcio.com/2023/08/risks-generative-ai.html]** such as data quality, modeling issues, and more. New issues include **hallucinations and confabulations [https://www.securityweek.com/vector-embeddings-antidote-to-psychotic-llms-and-a-cure-for-alert-fatigue/]**. Some ways to **address hallucinations and confabulations [https://artificialcorner.com/ai-hallucinations-understanding-the-p**

henomenon-and-exploring-potential-solutions-ccf13f36c798] include improving training data and accessing real-time information.

"The distinction between hallucinations and confabulations is important when considering the role of vector databases in the LLM workflow," says Joe Regensburger, VP of research at Immuta [https://www.youtube.com/watch?v=I1fLEk-nSVg]. "Strictly from a security decision-making perspective, confabulation presents a higher risk than hallucination because LLMs produce plausible responses."

Regensburger shared two recommendations on steps to reduce model inaccuracies. "Getting good results from an LLM requires having good, curated, and governed data, regardless of where the data is stored." He also notes that "embedding is the most essential item to solve." There's a science to creating embeddings that contain the most important information and support flexible searching, he says.

Rahul Pradhan, VP of product and strategy at Couchbase [https://www.couchbase.com], shares how vector databases help address hallucination issues. "In the context of LLMs, vector databases provide long-term storage to mitigate AI hallucinations to ensure the model's knowledge remains coherent and grounded, minimizing the risk of inaccurate responses," he says.

Conclusion

When SQL databases [https://www.infoworld.com/article/2255395/what-is-sql-the-lingua-franca-of-data-analysis.html] started to become ubiquitous, they spearheaded decades of innovation around structured information organized in rows and columns. NoSQL, columnar databases, key-value stores, document databases, and object data stores allow developers to store, manage, and query different semi-structured and unstructured datasets. Vector technology is similarly foundational for generative AI, with potential ripple effects like what we've seen with SQL. Understanding vectorization and being familiar with vector databases is an essential skill set for developers.

Databases[https://www.infoworld.com/database/] Data Management[https://www.infoworld.com/data-management/]

Generative AI[https://www.infoworld.com/generative-ai/] Machine Learning[https://www.infoworld.com/machine-learning/]