Advantages of Vector Databases
How Can Oracle Support Your Vector Needs?
Embrace the Power of Oracle AI Vector Search
Vector Database FAQs

A lesser-known data type, vectors, has seized the spotlight recently as an enabler of generative AI. But vectors—and databases capable of storing and analyzing them—have been toiling backstage for many years. They're used in geospatial mapping and analysis for city planning, transportation logistics, and environmental analysis. More recently, vectors have been used in recommendation engines for retail products as well as music and video streaming sites.

Generative AI builds on these use cases and opens the door to new innovations using vectors and vect databases along with companion technologies, including retrieval-augmented generation (RAG).

What Is a Vector?

A vector is simply a set of numbers that represents the features of an object—whether that object is a word, a sentence, a document, an image, or a video or audio file. Vectors are needed because comparing or searching this type of unstructured content is difficult for computers. Comparing or searching vectors, on the other hand, is much easier and is based on well-understood math.

Vectors are stored in a database with, or with a link to, the data objects on which they are based. Vectors that are mathematically close to one another tend to describe objects with similar features, so you can quickly compare or search them and return objects that are alike. You might also form a query vector with only certain features identified. A search will return objects with features similar to those specified in the query vector.

What Is a Vector Database?

A vector database is any database that can natively store and manage vector embeddings and handle the unstructured data they describe, such as documents, images, video, or audio.

With the importance of vector search for generative AI, the tech industry has spawned many specialized, standalone vector databases, which companies can add to their data infrastructures. Meanwhile, versions of established favorites, such as Oracle Database and the open source MySQL database, have incorporated vectors as a native data type alongside many other data types. This allows searches on a combination of business and semantic data to be faster and more precise, as both data types are available in a single database. This approach also avoids the data consistency problems introduced when using a separate, specialized vector database in addition to the business's primary database.

VIDEO: What Is a Vector Database? How Can It Help Meet Business Needs?

Vector Index vs. Vector Database

Vector indexes and vector databases are both designed to efficiently store and retrieve vectors, that is, sets of numbers that represent the features of an object, like a document, image, or video or audio file. However, they have different characteristics and use cases. Vector indexes are primarily used for “nearest neighbor” search, which involves finding the closest vector to a given query vector. Indexes are great for applications that require fast and accurate similarity searches, such as a recommendation engine. In contrast, vector databases are where organizations store vectors for retrieval and analysis. An enterprise-class vector database delivers useful features beyond nearest neighbor search, such as metadata storage, data versioning, and integration with other systems.

Key Differences

The key difference between a vector index and a vector database is that vector indexes store information about the attributes of unstructured data, such as text, images, or audio files. This information is represented by a set of numbers called a vector. The vector index holds this data and “indexes” it in a way that helps a database quickly identify and match objects.

A vector database houses these indexes and the objects they describe. However, how a database arran the vector indexes and data objects varies. Vector-enabled databases, such as Oracle Database, separa the storage of data objects from how the data and vectors are searched. This allows them to combine t mature querying power of SQL for metadata and up-to-date business data with the speed and context relevancy of vector search. This approach means, for example, that a vector search for relevant retail products can also deliver up-to-date pricing and availability.

Key Takeaways

Vector databases efficiently store and manipulate objects using a type of data called a vector embedding.

Vector embeddings describe the features of an object, and a vector-enabled database stores those vectors and creates indexes that facilitate fast searches.

Vectors and vector-enabled databases are not new; they have long been employed for specialized use cases, such as mapping and data analytics.

More recently, vector embeddings and vector databases have been used to find similar products, do biometric pattern recognition, detect anomalies, and in recommendation engines.

Enterprises are now combining vector search and generative AI with retrieval-augmented generation technology to get more relevant results from generative AI by sharing select items from their storehouses of documents and communications. The result is prompt responses that are more accurate and contextually relevant as they are based on the additional data supplied by RAG.

Vector Databases Explained

Instead of taking on the cost and effort to fine-tune generative AI models, companies are curating the data that LLMs use to generate their outputs. They're using vector databases that contain up-to-date enterprise information. This architectural approach, called retrieval-augmented generation, lets an LLM that was trained on vast amounts of generalized data enhance its response by using private data found in a vector database.

For example, if an LLM-powered chatbot could access a retailer's customer records and email communications instead of generic messages, it could provide more useful and personalized responses to queries such as, "Has my order shipped?"

RAG can also boost the reliability and trustworthiness of generative AI models by citing which documents in the vector database informed its output.

Why Are Vector Databases Important?

Unsurprisingly, the use of databases optimized for storing and analyzing vectors is rising. Once used primarily for mapping and data analysis, vector databases have become a critical cornerstone technology for the recommendation engines commonly used by the most popular retailers and music and video streaming providers as well as virtual assistants, biometric pattern recognition, anomaly detection, and more. And now, vector databases have found a new and spectacular use: Storing large volumes of unstructured data that can be accessed to inform the outputs of generative AI models.

A growing trend is for established databases, such as MySQL and Oracle Database, to incorporate vect data as a native data type alongside the rest of an organization's data, such as JSON, graph, spatial, ar relational. This convergence negates the need to move data to a separate database for generative AI operations, which both simplifies the process and leaves valuable data in trusted repositories.

The growth of generative AI use cases means there are many new vector databases on the market, in addition to the established NoSQL and relational databases that have added vector data type management.

How Do Vector Databases Work?

Vector databases work by storing and processing data as vectors, which are mathematical representations of features of objects in multidimensional space. This allows complex data types, such as images, audio, video, and sensor data, to be stored and queried efficiently, making vectors ideal for use cases like recommendation systems, natural language processing, and image recognition.

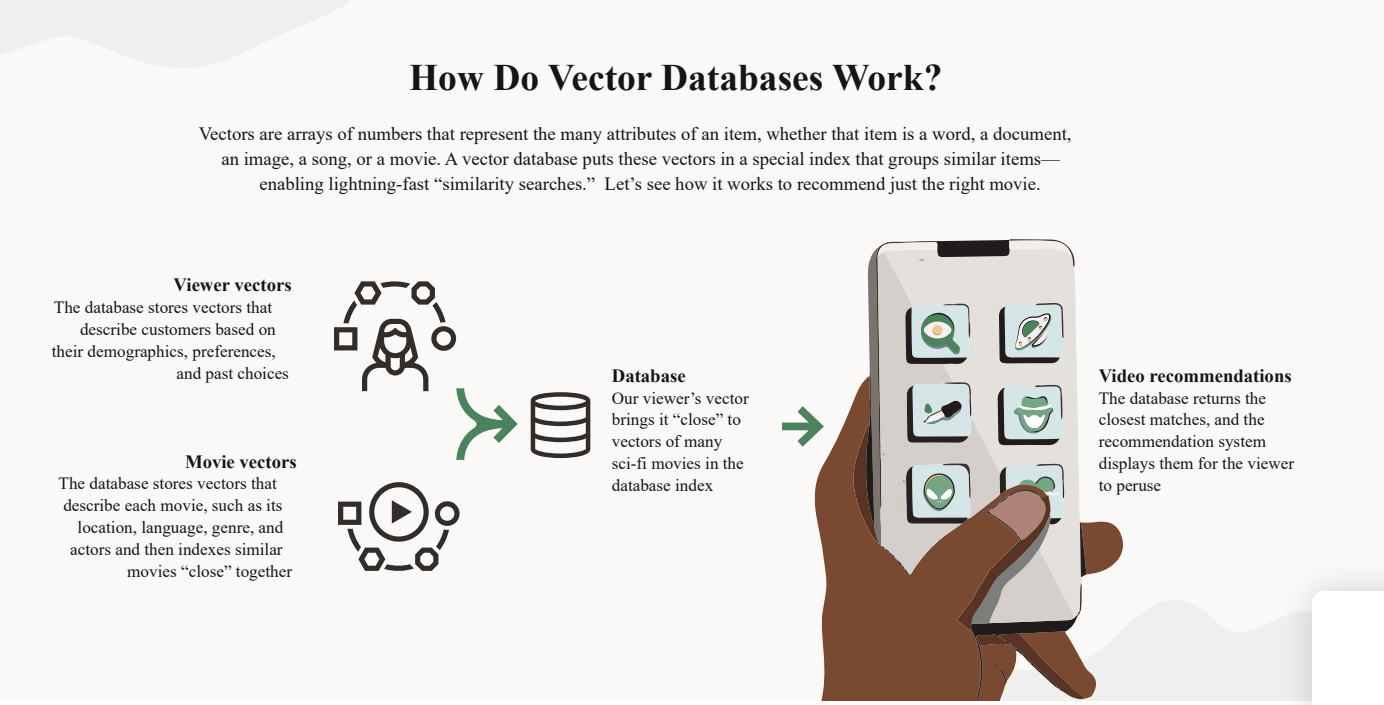
Operations happen in several steps:

Vectorization. Vectors can be created to describe the contents or features of unstructured data. This unstructured database could be in the form of text stored in database tables or documents stored on a file system.

Indexing. Vector databases use vector indexes to organize vectors in a hierarchical manner, allowing for their efficient search and retrieval.

Querying. To query vector data, vector databases perform vector distance operations using a query vector. The closer vectors are mathematically, the more similar are the objects that the vectors represent. Normally this process will return a small result set, such as the five most similar items.

Post processing. After a vector database retrieves a query vector's nearest neighbors, it may optionally re-rank the rows of the result set. Re-ranking is an expensive operation compared with the vector query, but it can give a better order for the existing vector query results.



The diagram illustrates how a vector database can help a streaming service recommend just the right movie for a sci buff.

Types of Vector Databases

Vector databases can be categorized into various types, either by storage structure, such as columnar, or classified based on their implementation, such as in-memory databases. Each type has advantages; which type you select should be based on your specific requirements and use cases.

Columnar databases

A columnar database stores data in columns and groups them on a disk. This arrangement speeds up data analytics because the analysis usually involves filtering and combining data from table columns. Databases can also store records in row format, which is helpful for transactions where you're updating a single item in the record. For data analysis, however, a columnar database lets analysts scan any column of thousands or millions of records very quickly.

Document stores

A document store database, or document-oriented database, is a program and data storage system that manages, retrieves, and stores document-oriented information. Document databases store data in flexible, JSON-like documents, which are suitable for varied and dynamic data structures. They differ from relational databases, which use tables to organize data with a fixed schema.

Graph databases

Graph analytics is the process of analyzing data in a graph format, using data points as nodes and relationships as edges. Looking at data this way can help you discover connections and relationships that weren't obvious before. Graph analytics requires a database that can support graph formats. This could be a dedicated graph database or a converged database that supports multiple data models, including graph.

In-memory databases

In-memory databases store and manipulate data in the memory tier of an application rather than on a storage disk. This type of structure is often matched with columnar store functionality and vectorized query plans to accelerate analytic queries. This type of data store is good at supporting lightning-fast operations in global telecommunications and businesses that need to make decisions based on real-time data.

Key-value stores

Key-value stores, sometimes also called key-value databases, are a quick way to store and query data that is often changing, such as items in an online store. The key-value store is a NoSQL-type database that uses a key value to pinpoint a specific record for querying or updating.

Spatial databases

A spatial database stores and manages spatial data, which represents information about the physic location and geometric properties of objects in space. They do this with indexing techniques and query operations. Spatial databases are used for online mapping and analytics as well as in shippin logistics operations.

Time series databases

A time series database is used to efficiently store and analyze time-stamped data, where each data point is associated with a specific time stamp or time interval. These databases are commonly used in IT monitoring systems that depend on log analysis and in finance.

Vector databases for large language models

Vector databases enhance commercial or open source large language models by giving them access to up-to-date information supplied by a local organization or business. This helps make the LLM's output more relevant and personalized for people associated with that organization.

Who Uses Vector Databases?

Vector databases are used by various applications and organizations that deal with large amounts of spatial and geometric data, such as in the retail and logistics industries and for systems that pilot autonomous vehicles. And now, companies exploring advanced AI and machine learning are adopting vector databases, too. Generative AI models, for example, depend on vector databases to improve their outputs by using local, up-to-date data.

Other specific use cases include the following:

Finance firms use vectors in several ways. For example, in portfolio analysis, vectors can represent aspects of a client's portfolios. They can also be used to track account performance over time.

Healthcare researchers use vector databases to support their research and clinical trials. They store and analyze data related to patient demographics, locations, and treatment outcomes, allowing researchers to assess the impact of many different factors on treatment efficacy.

Online retailers use vector databases to reference past purchases and browsing habits and recommend products that customers are likely to find desirable.

Shipping logistics companies use vector databases to store information about locations and distances, allowing them to accurately map and track objects in motion.

Streaming services use vectors to run recommendation engines, allowing them to present recommendations based on many factors, including genre, lead actors, release date, and reviews.

How Are Vector Databases Used?

The use cases for vector databases are as varied as the organizations and applications that depend on them. In addition to real-time data analytics, financial systems, and recommendation engines, vector databases are optimized to handle the complex data structures commonly required for tasks such as image recognition and natural language processing.

By storing and processing data efficiently, vector databases enable companies to leverage complex da structures for a wide range of applications, including the following:

Recommendation systems. Vector embeddings are used to quickly find similar product or entertainment options that are likely to interest a shopper or browser.

Search engines. Search engines use vector databases to index queries and documents with their vector embeddings, allowing them to locate similar search results or similar documents quickly.

Personalization. These systems use demographic information and past choices as guides for vector searches that pinpoint products or services that are likely matches for a particular user.

Anomaly detection. Vector databases allow the efficient search for anomalous vectors, even in very large data sets. This can help security teams spot attempted breaches and credit card companies stop fraudulent transactions.

Genomics and bioinformatics. Because vectors and vector databases are good at pattern matching and anomaly detection, they can help researchers match genetic sequences for comparison of large volumes of genetic data. This can aid in areas such as disease prediction and drug discovery.

Healthcare and medical research. Healthcare providers are using vector databases to store and manage information relevant to patient care, such as medical records, demographic data, lab results, and even genetic information. In clinical trials, geospatial data related to trial sites, patient demographics, treatment outcomes, and adverse events can be analyzed to determine the efficacy of a treatment.

Image and video retrieval. Image and video retrieval operations employ vector databases for similarity and semantic searches that quickly pinpoint images or videos amid deep catalogs of options.

Advantages of Vector Databases

Vector databases offer many advantages, including fast similarity search. They are optimized for efficient nearest-neighbor searches, allowing quick retrieval of similar items even in large datasets. This makes them ideal for applications and industries that require real-time processing and analysis of unstructured data and for emerging generative AI use cases.

Other advantages include the following:

Cost-effectiveness. Vector databases, particularly open source options such as PostGIS, MySQL with vector extensions, or multimodel databases with native vector stores, offer cost-effective solutions for geospatial analysis and generative AI models.

Efficient storage. Spatial indexing techniques in vector databases allow for efficient storage and organizing of vectorized data.

Fast retrieval. Vector databases are indexed for fast retrieval of data based on an object's many attributes. They do this by noting relationships and proximity and using those to execute searches quickly.

Integration with machine learning. Vector databases are designed to integrate with machine learning frameworks and algorithms, which drives the development of predictive models, anomaly detection, clustering, and other machine learning-based analyses.

Personalization. Vector databases allow retailers, music streaming services, and even healthcare businesses to tailor their services to quickly find matches for an individual’s preferences and needs.

Real-time analysis. Vector databases can support in-memory operations for fast query response times and efficient data processing. This enables them to perform real-time analysis for day-to-day decision-making.

Reduced development complexity. Vector databases can provide APIs, libraries, and query languages that abstract away the complexities of data management and application development. This can vastly reduce the time involved in the application development process and, thus the cost.

Scalability. Vector databases can efficiently manage and process millions or even billions of vector objects and, with the right infrastructure, grow quickly to keep up with demand.

Versatility. Vector databases support a wide range of unstructured data, such as audio recordings, text documents, and images. This versatility allows them to accommodate many use cases and applications.

How Can Oracle Support Your Vector Needs?

Whether you're using generative AI or nearly any other operation using vectors, Oracle can help.

Oracle Database, the world's most popular enterprise database, provides a single data platform for vectors and all your business data. Effortlessly harness the capability of similarity search for your company's data without the need to oversee and synchronize various databases. AI Vector Search allows you to conduct searches on both structured and unstructured data by understanding its semantics or meaning, as well as its values.

Combining relational data, JSON documents, graphs, geospatial data, text, and vectors in a single database enables you to rapidly build new features in your applications. AI Vector Search in the Oracle Database can also be used in a RAG pipeline together with any GenAI service. In addition, Oracle's HeatWave MySQL database service handles vectors natively to support vector search and other use cases. For example, you can use it together with the RAG service in Oracle Cloud Infrastructure (OCI) to bring a generative AI interface to your proprietary documents, giving you an AI that's an expert in your organization's operational data.

Embrace the Power of Oracle AI Vector Search

Whether you're using vectors for data analysis, geospatial applications, product recommendations, or as an enabling technology for generative AI, Oracle can help. Both Oracle's flagship Autonomous Database and Oracle HeatWave MySQL manage vectors as a native data type alongside many other data types for a simpler development experience. Both databases run on Oracle Cloud Infrastructure. OCI is designed with the latest processors and supercluster architecture to efficiently handle the most demanding AI workloads, including generative AI, computer vision, and predictive analytics. Whether you build with Oracle Database or open source MySQL database, you can start taking advantage of vector search today.

In the age of generative AI, vector databases have become more important to businesses than ever before. As more development teams look to store and manage the vector data type, they'll have a decision to make: Bring in a specialized, purpose-built vector database or use multimodel databases, such as Orac Database, that support not only vectors but many other data types as well.

Vector databases are pivotal to exciting AI use cases, including chatbots that revolutionize customer service and algorithms that transform healthcare. See how companies are putting the power of vectors to work now.

Access the ebook

Vector Database FAQs

When should you use a vector database?

A vector database can be used for a wide variety of use cases, including geospatial applications, such as shipping logistics or environmental research, recommendations for retail or online entertainment options, or, more recently, as a primary storehouse for data that supports generative AI by individual organizations.

Does Netflix use vector databases?

Netflix announced it uses vector databases to support its popular recommendation engines. It applies vector embeddings to every piece of entertainment in its catalog and the vector database enables real-time search for similar titles.

Educators
vs. MongoDB
What is PaaS?
Cloud Lift
Oracle Support Rewards
Oracle Red Bull Racing
OCI Blog