2 AI Vector Search

Oracle AI Vector Search is designed for Artificial Intelligence (AI) workloads and allows you to query data based on semantics, rather than keywords.

Vector Data Type

This feature provides a built-in VECTOR data type that enables vector similarity searches within the database.

With a built-in VECTOR data type, you can run AI-powered vector similarity searches within the database instead of having to move business data to a separate vector database. Avoiding data movement reduces complexity, improves security, and enables searches on current data. You also can run far more powerful searches with Oracle AI Vector Search by combining sophisticated business data searches with AI vector similarity search using simple, intuitive SQL and the full power of the converged database - JSON, Graph, Text, Spatial, Relational and Vector - all within a single query.

View Documentation

Vector Indexes

Vector Indexes are a class of specialized indexing data structures that are used to efficiently store and search high-dimensional vector data. A vector index organizes vector data in a manner such that similar items (where similarity is defined by distance between two vectors) are grouped together, thus, making the search process extremely efficient. Unlike traditional database indexes, vector indexes are commonly used on large datasets to perform approximate similarity searches that can trade-off between query accuracy and query performance depending on the application's requirements.

This functionality enables efficient similarity searches and faster query performance for AI-driven applications. In addition, vector indexes scalability and support for high-dimensional data improve analytical insights and can lead to informed decision-making and a competitive business advantage.

View Documentation

AI Vector Search: SQL Execution

AI Vector Search SQL Execution adds SQL execution support for vector indexes built on vector columns inside the database. In addition, it provides support for SQL Functions related to the vector type and allow for row level restriction capabilities in SQL queries for partitions.

This feature allows you to more easily build with the vector data type, enabling the rapid development of AI-driven applications.

View Documentation

Vector Utility API

The Vector Utility API provides a SQL function `VECTOR_CHUNKS` which processes text into pieces (chunks) in preparation for the generation of embeddings to be used with a vector index. The API is configurable in terms of size of chunks and rules for splitting chunks.

While it is possible for you to create your own chunking algorithms, utilizing this functionality could save you time and aid in faster development with a pre-packaged SQL function.

View Documentation

Chainable Utility Functions for Vectors

DBMS_VECTOR provides a set of utility functions for processing text for the creation of vector indexes. These functions may be chained together such that the output from one function is used as the input for the next.

This feature offers a straightforward yet very customizable method for you to turn textual content, like a PDF document or VARCHAR2 database field, into the embeddings necessary for a vector index. This capability enables you to seamlessly develop with vectors, facilitating the creation of the next generation of Artificial Intelligence applications with ease.

View Documentation

Support for ONNX-Format Models as First-Class Database Objects

The Open Neural Network Exchange (ONNX) is an open format to represent machine learning models. It faciliates the exchange of models between systems and is supported by an ONNX runtime environment that enables using models for scoring/inference.

You can import ONNX-format models to Oracle Database for the machine learning techniques classification, regression, clustering, and embeddings.

The models will be imported as first-class MINING MODEL objects in your schema. Inference can be done using the family of OML scoring operators, including PREDICTION, CLUSTER, and VECTOR_EMBEDDING.

You can import and use third-party ML models, possibly built in other environments or from other sources, to leverage the database as an ML scoring platform.

Users can invoke these models from SQL queries using the same scoring operators as native in-database models.

While ONNX format models can already be imported to OML Services on Autonomous Database Serverless, you can now use ONNX-format models from Oracle Database.

View Documentation

AI Vector Search: Optimizer

This functionality adds support to the Optimizer to use indexes built on the new Vector data type rather than doing full table scans.

The support for vector indexes being used by the Optimizer allows for efficient computation of vector queries enabling developers to build the next generation of AI-powered solutions.

View Documentation

AI Vector Search: PL/SQL

This functionality adds a new vector type to the PL/SQL type system, along with a set of vector operations useful for performing similarity searches on sets of vectors.

Support for the new vector data type in PL/SQL opens up new possibilities for developers to build robust and efficient AI-driven applications.

View Documentation

JDBC Support for Vector Data Type

This feature adds the necessary components to the JDBC drivers to support the AI Vector Search data type including SQLType, DatabaseMetaData, ResultSetMetaData and ParameterMetaData, VectorMetaData, Java to SQL Conversions with PreparedStatement and CallableStatement, SQL to Java Conversions with CallableStatement, SQL to Java Conversions with CallableStatment and ResultSet, and VECTOR Datum class.

JDBC Support for the Vector data type enables developers to build robust, scalable, and high-performance Java applications with Artificial Intelligence focus.

View Documentation

Oracle Call Interface Support for Vector Type

Oracle Call Interface (OCI) now supports Vector data type. Applications that use OCI can now take advantage of the new Vector data type in the Oracle Database.

This feature ensures that you can leverage the full capabilities of the Oracle Database through OCI-based applications to create the next generation of AI-powered solutions.

View Documentation

Support of Vector Data Type in JSON Type (OSON)

This functionality extends the standard JSON scalar types, to include the new Vector data type. It is fully supported by all Oracle JSON constructs, and a vector scalar JSON value is convertible to/from a JSON array of numbers.

Embedding vector values in JSON-type data is important for interoperability between SQL values and JSON values. For example, a table with a VECTOR column can be exposed in JSON data without a loss of data-type information allowing developers to create the next generation of AI applications.

View Documentation