🤖 VECTOR DATABASE FAQS

📌 GENERAL OVERVIEW

What is a vector database?
A specialized database designed for storing and retrieving high-dimensional vectors used in AI and ML.

How is it different from traditional databases?
Traditional databases store structured data, while vector databases handle unstructured data like embeddings.

Why are vector databases important in AI?
They support semantic search, recommendations, and similarity comparisons.

🧠 CORE CONCEPTS

What are embeddings?
Machine-generated vector representations of data such as text, images, or audio.

What is vector similarity search?
A method to find vectors most similar to a given query vector.

Which distance metrics are commonly used?
Cosine similarity, Euclidean distance, and Dot product.

What is a collection in a vector database?
A group of vectors and metadata, similar to a table in SQL databases.

⚙️ ARCHITECTURE & COMPONENTS

What are the main components of a vector database system?
Indexing engine, storage, metadata layer, hybrid search, and APIs.

What is hybrid search?
A combination of keyword filtering and vector similarity for enhanced precision.

Which indexing techniques are used?
HNSW, IVF, and PQ are common techniques.

📦 USE CASES

What are common applications of vector databases?
Chatbots, anomaly detection, semantic search, recommendations, etc.

How are they used in e-commerce?
For personalized product suggestions.

What about in healthcare?
Used for medical image comparison and document search.

How do they help in cybersecurity?
Detect anomalies in behavior patterns via vector analysis.

🔧 CYFUTURE.AI INTEGRATION

What is Cyfuture.AI Vector DB as a Service?
A managed, cloud-based vector database platform with UI and APIs.

What parameters can be configured when creating a DB?
Cluster name, vector size, collection name, metric type, and hosting plan.

What happens after launching a cluster?
You receive a dashboard URL, Qdrant URL, and API key.

Is a default collection created automatically?
Yes, for faster onboarding.

🔐 SECURITY & ACCESS

How is access secured in Cyfuture.AI?
API keys and secure endpoints restrict unauthorized usage.

Can collection-level access be restricted?
Yes, access controls can be set per collection.

⚡ PERFORMANCE & SCALING

How does a vector database scale?
Horizontally, using distributed data across shards/nodes.

What factors affect performance?
Index type, vector size, metric used, and system hardware.

Is real-time search supported?
Yes, with proper hardware and indexing.

🛠️ IMPLEMENTATION & DEPLOYMENT

Which open-source tools are supported?
Platforms like Qdrant, Weaviate, and more.

Is GPU support mandatory?
Not mandatory, but helpful for ANN search and embedding generation.

Can it be hosted on-premises?
Yes, Cyfuture.AI supports self-hosting.

Which file types can be uploaded?
Text, CSV, PDF, audio, and image files after preprocessing.

🧪 VECTOR PROCESSING & CLEANING

What does preprocessing include?
Metadata extraction, NaN cleaning, and embedding generation.

Can metadata cleaning be done manually?
Yes, or you can automate it using built-in tools.

🔄 QUERYING & APIS

How do I query the database?
Via REST APIs or SDKs with keyword or vector input.

What is a hybrid query?
Combining filters (e.g., "category: book") with vector similarity.

Are there API rate limits?
Yes, based on the hosting plan.

Is Python supported for querying?
Yes, Python SDKs and client libraries are available.

💡 BEST PRACTICES

Which distance metric should I use?
Depends on use case: Cosine for semantic, Euclidean for spatial, etc.

How to choose vector size?
Based on the embedding model (e.g., BERT = 768 dims).

When should indexes be rebuilt?
After large data additions or deletions.

💬 AI/ML INTEGRATION

Which models can generate embeddings?
BERT, CLIP, OpenAI models, Sentence Transformers, etc.

Can LLMs work with vector DBs?
Yes, especially in Retrieval-Augmented Generation (RAG).

How do vector DBs help chatbots?
By enabling context-aware and relevant responses.

📊 MONITORING & ANALYTICS

Can you monitor search performance?
Yes, dashboards show latency, volume, and vector health.

Is usage/cost analytics available?
Yes, detailed billing per user/org is supported.

🤝 INTEGRATION & COMPATIBILITY

Can it integrate with existing systems?
Yes, works with relational DBs, data lakes, and APIs.

Is multi-modal data (text + image) supported?
Yes, with combined embeddings.

Can you export data?
Yes, in JSON, CSV, or binary formats.

🧾 PLANS & PRICING

How is pricing structured?
Based on vector count, storage size, search rate, and API usage.

Are free plans available?
Yes, for trials and small-scale development.

🧰 ADVANCED FEATURES

What is vector quantization?
A technique to compress vectors for faster retrieval.

What is re-ranking in vector search?
Sorting results post-search based on refined relevance.

Can Boolean filters be used?
Yes, hybrid filters like `category = "tech"` are supported.

🔍 TROUBLESHOOTING

Why are search results poor?
Could be low-quality embeddings or wrong similarity metric.

Why aren’t my vectors indexing?
Likely due to unsupported dimensions or bad formatting.

How to fix API access issues?
Verify API keys, cluster status, and correct endpoints.