Vector Databases

Group C3: Huaijin (Tony) Tu, Natasha Mohanty, David Heineman, Ishaan Immatty

Vector DB Overview

What is Vector DBs?

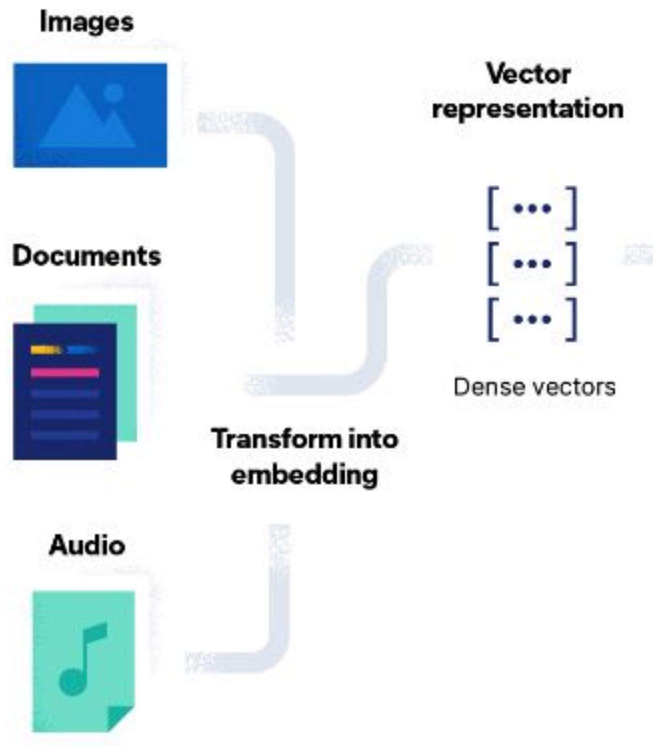
- Specialized databases storing and querying high-dimensional vector data.
- AI & Machine Learning models (e.g. LLM)
- Embedding vectors for complex unstructured data
- Efficient similarity search

What is a vector database?. IBM. (n.d.). https://www.ibm.com/topics/vector-database
What is a vector database? - vector databases explained - AWS. (n.d.). https://aws.amazon.com/what-is/vector-databases/

Core Concepts

- Embedding vectors:
  High-dimensional vectors transformed from complex unstructured and semi-structured data that capture the meaning and context of an asset.

  Image -> Image embeddings
  Text -> Text embeddings
  Graph -> Graph embeddings



What is a vector database? - vector databases explained - AWS. (n.d.). https://aws.amazon.com/what-is/vector-databases/
What are vector embeddings?: A comprehensive vector embeddings guide. Elastic. (n.d.). https://www.elastic.co/what-is/vector-embedding

Core Concepts

- Similarity search:
  Find k most similar data objects to a given query object, based on some measure of similarity or distance (e.g. cosine similarity, L-2 distance, etc)

- k-NN index
  Efficient and fast lookup of the k nearest neighbors of a query vector in a large collection of vectors

- Approximate Nearest Neighbor (ANN) Algorithms
  Hierarchical Navigable Small World (HNSW)

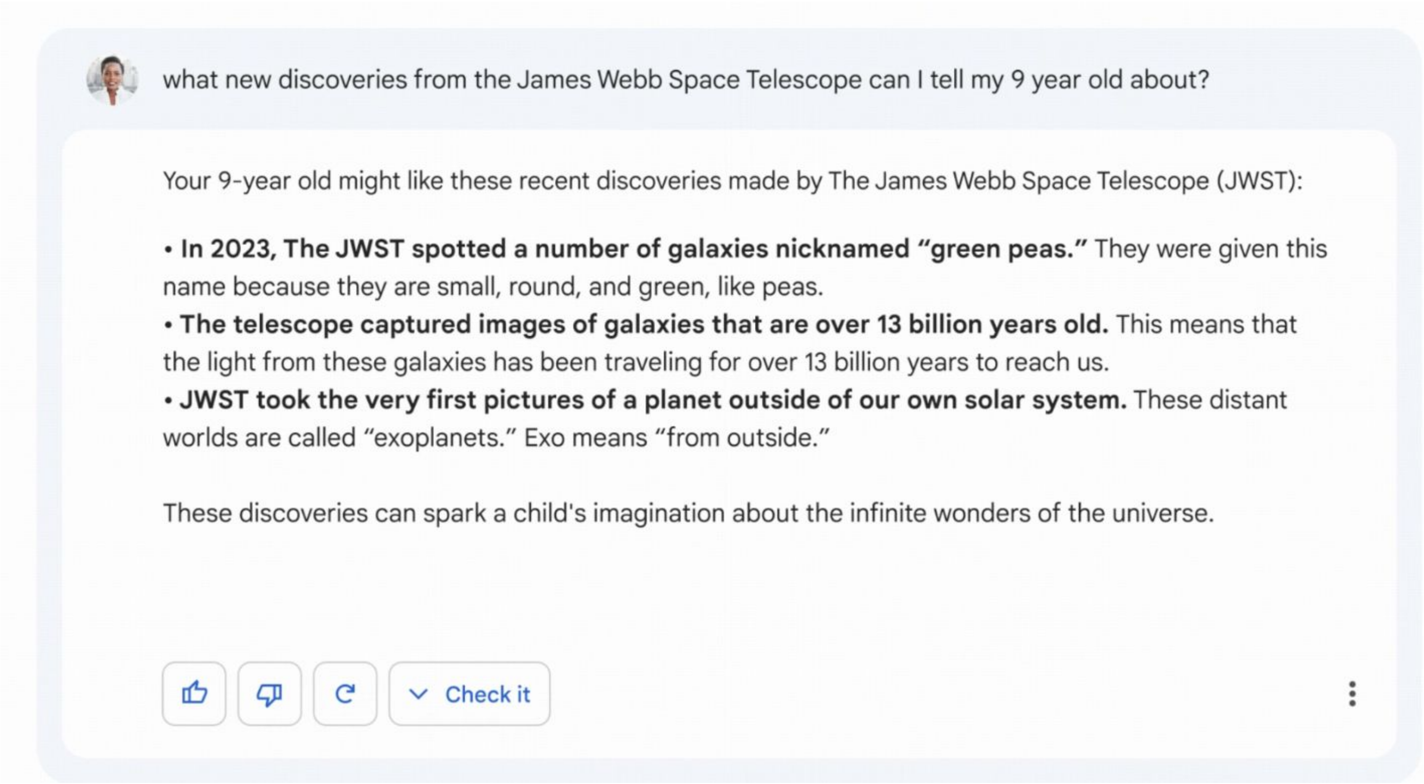
What is a vector database? - vector databases explained - AWS. (n.d.). https://aws.amazon.com/what-is/vector-databases/
Schwaber-Cohen, R. (n.d.). What is a vector database & how does it work? use cases + examples. Pinecone. https://www.pinecone.io/learn/vector-database/

The Role of Vector Databases in AI Industry

- Recommender system
- AI-based Semantic Search
- Retrieval-Augmented Generation (RAG)
  Large Language Model (LLM) -> hallucination
  RAG -> accurate, grounded to the provided knowledge bases

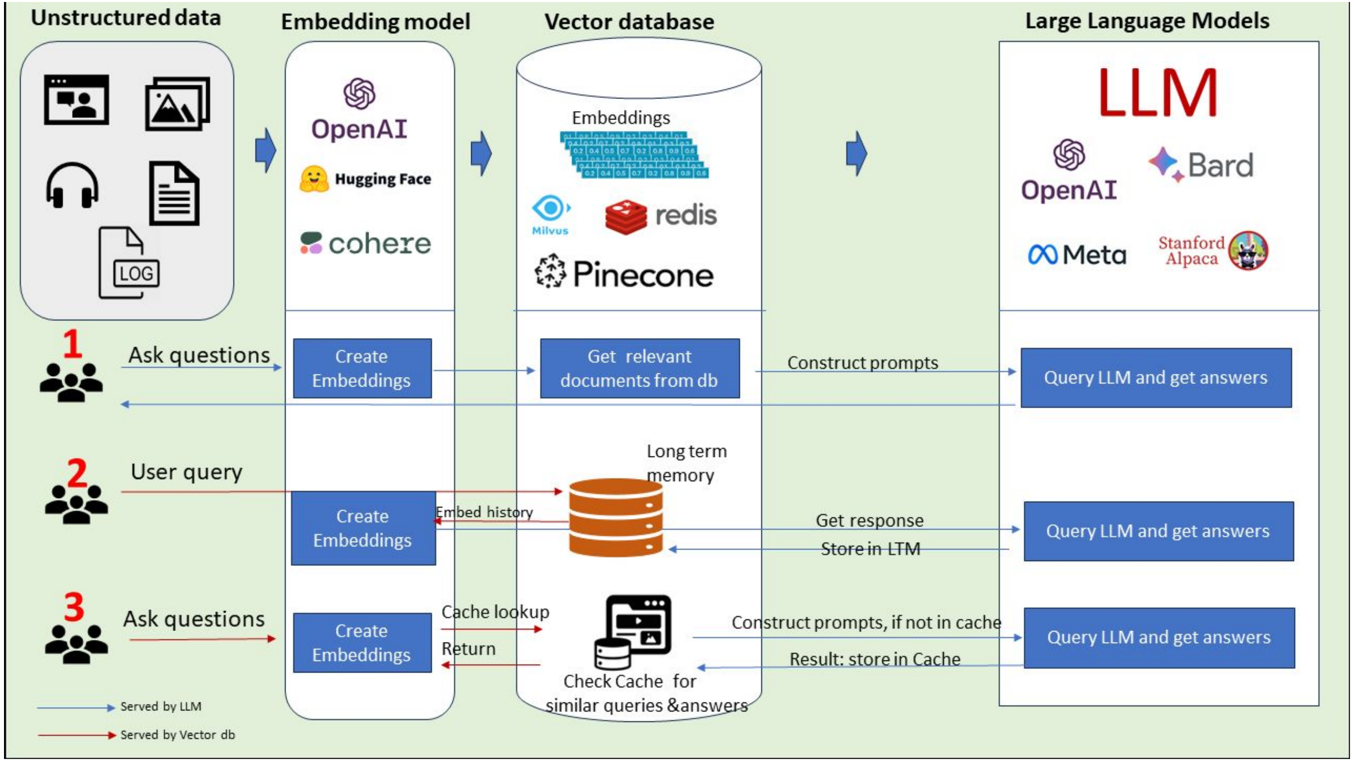
What is a vector database?. IBM. (n.d.). https://www.ibm.com/topics/vector-database

LLM Hallucination



Kan, M. (2023, February 8). No, that's wrong: Google's bard ai demo spouts incorrect info. PCMAG.
https://www.pcmag.com/news/no-thats-wrong-googles-bard-ai-demo-spouts-incorrect-info

Retrieval-Augmented Generation



Ali, M. (2023, September 12). The 5 best vector databases: A list with examples. DataCamp. https://www.datacamp.com/blog/the-top-5-vector-databases

Brief History of Vector DB

- **Late 1970s:** Initial need of storing vast vector data in DNA sequencing
- **1980s to Mid-2000s:** More development of Vector DB in scientific research (NIH and Stanford)
- **2000s to 2019:** Application in genetic researches with parallel computing and storage (e.g. UniVec)
- **2019 to 2022:** Surge of vector DB (creation of Pinecone, Weaviate, and Milvus) into AI and Machine Learning
- **2022 to Now:** LLMs (e.g. ChatGPT) and Rise of Large Multi-Modal AI (e.g. language, image, audio, etc) necessitating large-scale vector databases

Norlin, E. (2023, April 25). In search of the history of The vector database. SW2.ai. https://sw2.beehiiv.com/p/search-history-vector-database

Features of Vector DB

- Performance and Scalability:

  Efficient storage and query of vectors
- Ease of Use and Community Support:

  User-friendly interfaces

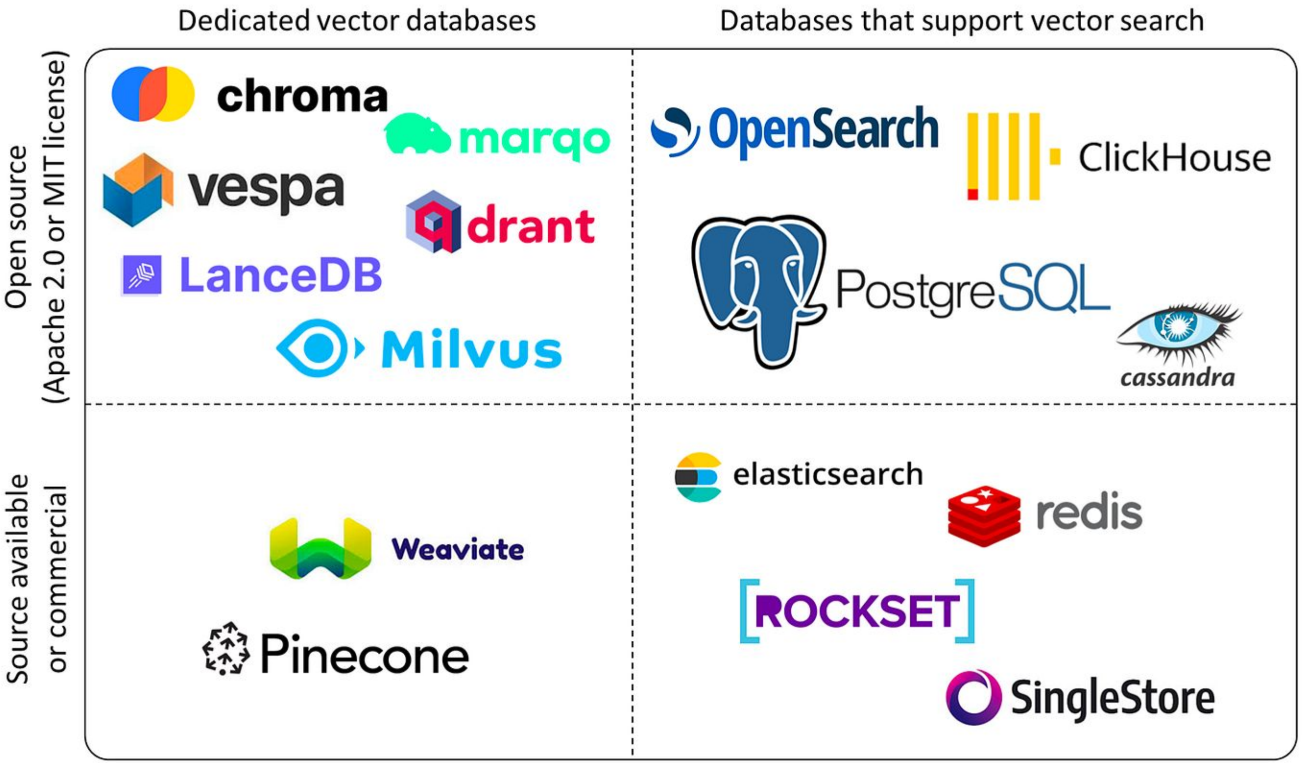
  Integration with AI models and other database ecosystems
- Reliability and Security:

  Fault tolerance, authentication, access control, data management
- Accessibility and Deployment Options

  Open-source vs proprietary

  Self-hosted vs cloud-hosted
- Cost-effectiveness

What is a vector database? - vector databases explained - AWS. (n.d.). https://aws.amazon.com/what-is/vector-database
Fröberg, E. (n.d.). Picking a vector database: a comparison and guide for 2023. Vector View. https://benchmark.vectorview.ai/vectordbs.html



Ali, M. (2023, September 12). The 5 best vector databases: A list with examples. DataCamp. https://www.datacamp.com/blog/the-top-5-vector-databases

	Pinecone	Weaviate	Milvus	Qdrant	Chroma	Elasticsearch	PGvector
Is open source	❌	✅	✅	✅	✅	❌	✅
Self-host	❌	✅	✅	✅	✅	✅	✅
Cloud management	✅	✅	✅	✅	❌	✅	(✔)
Purpose-built for Vectors	✅	✅	✅	✅	✅	❌	❌
Developer experience	👍👍👍	👍👍	👍👍	👍👍	👍👍	👍	👍
Community	Community page & events	8k☆ github, 4k slack	23k☆ github, 4k slack	13k☆ github, 3k discord	9k☆ github, 6k discord	23k slack	6k☆ github
Queries per second (using text nytimes-256-angular)	150 *for p2, but more pods can be added	791	2406	326	?	700-100 *from various reports	141
Latency, ms (Recall/Percentile 95 (millis), nytimes-256-angular)	1 *batched search, 0.99 recall, 200k SBERT	2	1	4	?	?	8
Supported index types	?	HNSW	Multiple (11 total)	HNSW	HNSW	HNSW	HNSW/IVFFlat
Hybrid Search (i.e. scalar filtering)	✅	✅	✅	✅	✅	✅	✅
Disk index support	✅	✅	✅	✅	✅	❌	✅
Role-based access control	✅	❌	✅	❌	❌	✅	❌
Dynamic segment placement vs. static data sharding	?	Static sharding	Dynamic segment placement	Static sharding	Dynamic segment placement	Static sharding	-
Free hosted tier	✅	✅	✅	(free self-hosted)	(free self-hosted)	(free self-hosted)	(varies)
Pricing (50k vectors @1536)	$70	fr. $25	fr. $65	est. $9	Varies	$95	Varies
Pricing (20M vectors, 20M req. @768)	$227 ($2074 for high performance)	$1536	fr. $309 ($2291 for high performance)	fr. $281 ($820 for high performance)	Varies	est. $1225	Varies

Fröberg, E. (n.d.). Picking a vector database: a comparison and guide for 2023. Vector View. https://benchmark.vectorview.ai/vectordbs.html

Technical Details & Existing Products

Existing Vector DB Implementations

Pinecone

Weaviate

Milvus

Managed DBs

OpenSearch *

*

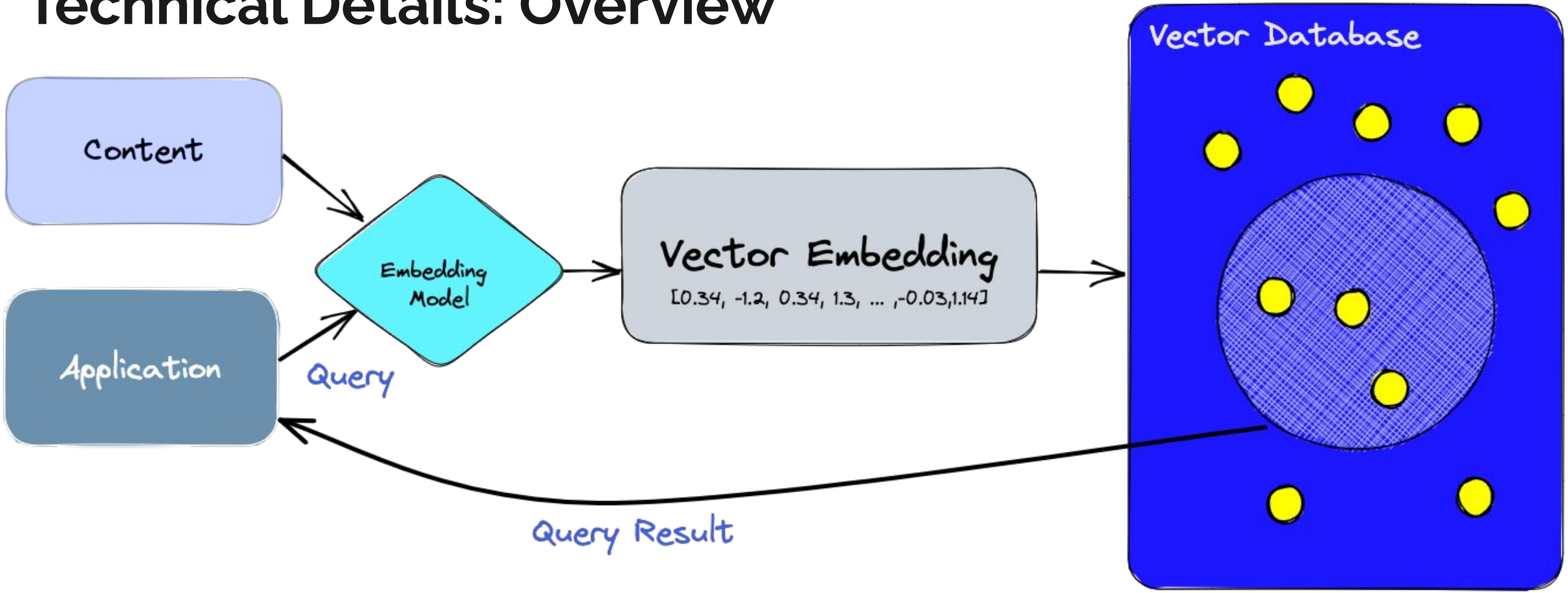
qdrant

Chroma

Open Implementations

* = Vector DB is a plugin on an existing database

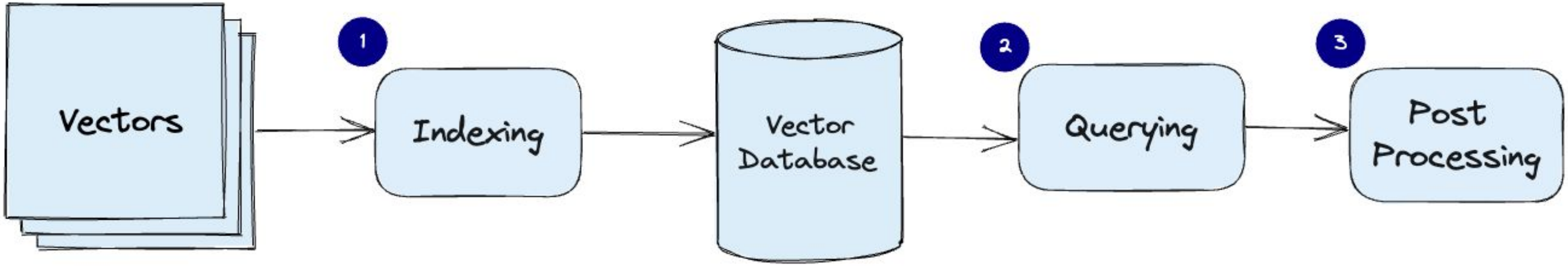
Technical Details: Overview



In an application, vector DBs serve as a k-NN wrapper around some embedding of underlying embedded documents, images or other data

pinecone.io/learn/vector-database

Technical Details: 3 Stages of Vector Retrieval



Builds a data structure to make query operations faster

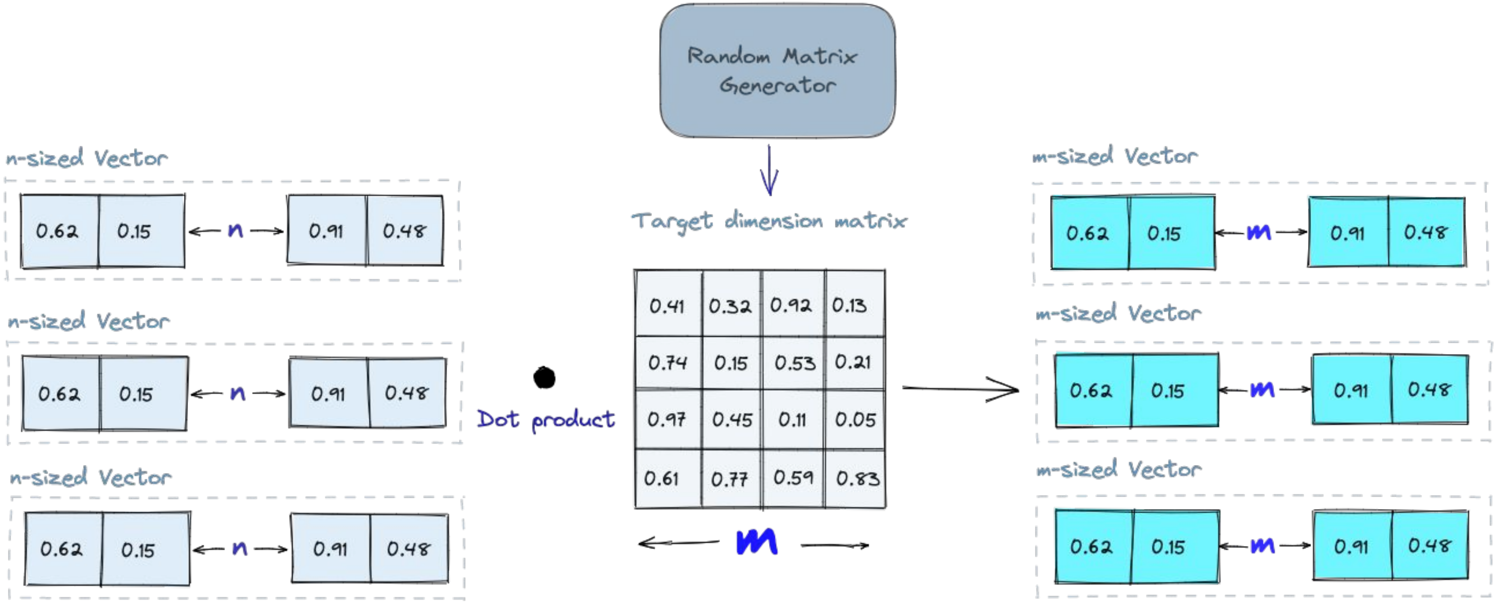
k-NN Similarity

Re-rank or filter results

pinecone.io/learn/vector-database

Optimization 1: Low Dimensional Indexing

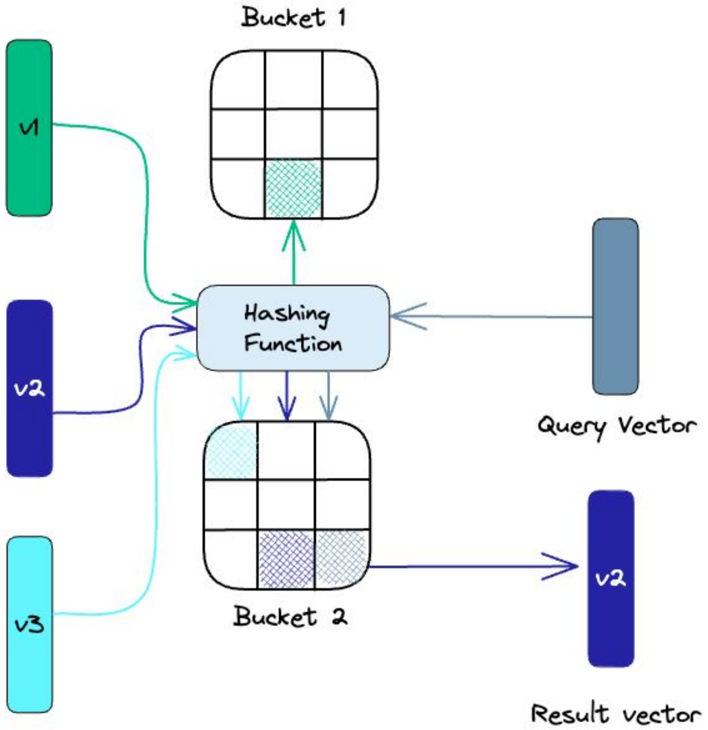
*Construct a **random projection matrix** to store embeddings as m << n sized vectors, which preserves similarity but is faster to compare with a query*



pinecone.io/learn/vector-database

Optimization 2: Locally Sensitive Hashing

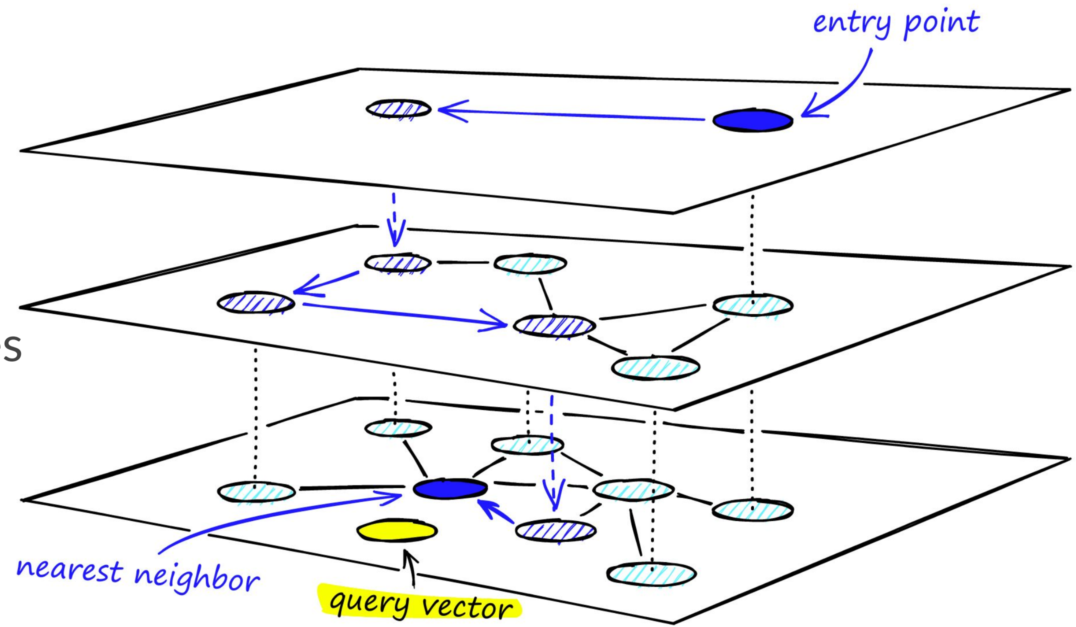
- Rather than give exact results, LSH *approximates* vectors into different buckets using a hash function
- Each bucket attempts to include similar input vectors such that the query vector only has to calculate similarity within the bucket
- Different hash functions can be used depending on input modality, but **MinHash** is used for arbitrary vectors



pinecone.io/learn/vector-database
pinecone.io/learn/series/faiss/locality-sensitive-hashing

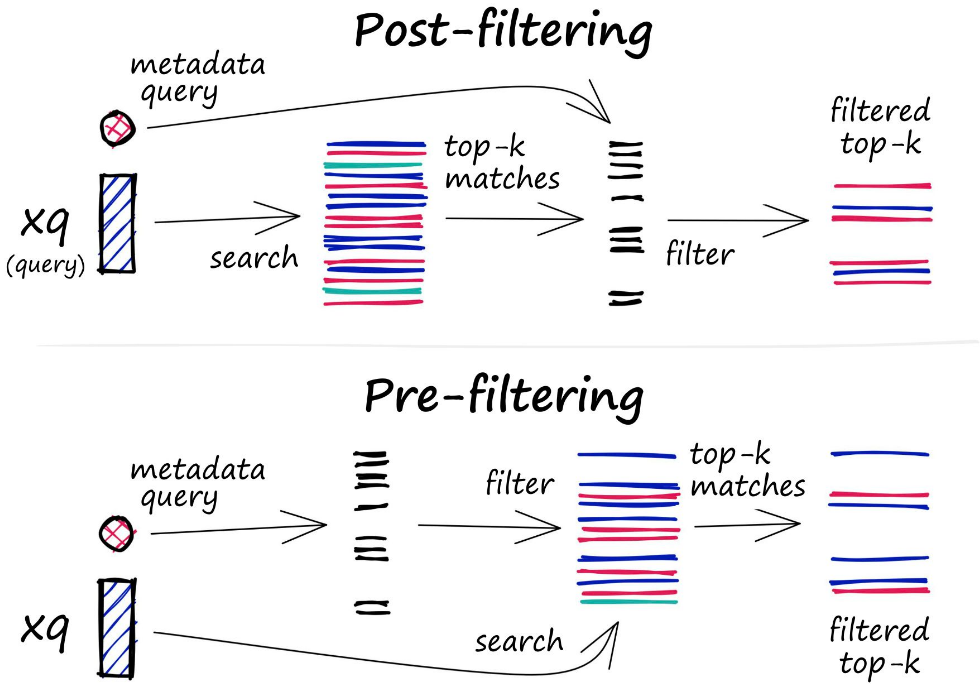
Optimization 2.5: Hierarchical Clustering

- We can imagine grouping close clusters of vectors into higher order graphs
- At query time, we perform k-NN search on smaller graph, and perform k-NN iteratively on nodes connected to that subgraph
- Substantially improves memory usage at retrieval time, **but does not guarantee correctness.**



pinecone.io/learn/vector-database
pinecone.io/learn/series/faiss/hnsw

Optimization 3: Filtering



- In some search applications, we may want to filter vectors by some metadata
- Rather than filter *after* finding the closest candidates, we filter before applying search to reduce the number of k-NN operations

pinecone.io/learn/vector-database
pinecone.io/learn/vector-search-filtering

Existing Vector DB Implementations

Pinecone

Weaviate

Milvus

Managed DBs

OpenSearch *

*

qdrant

Chroma

Open Implementations

* = Vector DB is a plugin on an existing database

OpenSearch Implementation

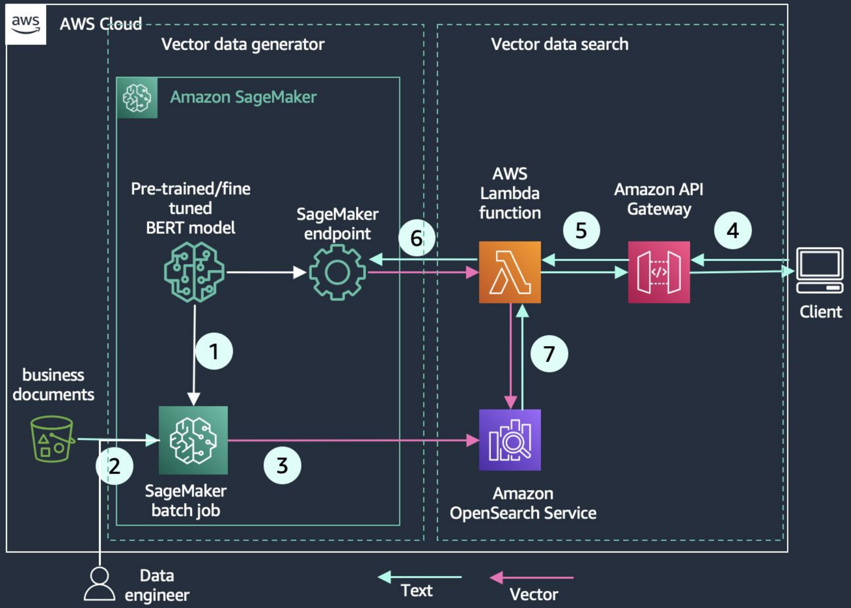
- Provides three different k-NN approximation algorithms, implementing Hierarchical Small World Navigation
- Allows customizing similarity metric (right figure)

spaceType	Distance Function (d)	OpenSearch Score
l1	$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \lvert x_i - y_i \rvert$	$score = \frac{1}{1 + d}$
l2	$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} (x_i - y_i)^2$	$score = \frac{1}{1 + d}$
linf	$d(\mathbf{x}, \mathbf{y}) = max(\lvert x_i - y_i \rvert)$	$score = \frac{1}{1 + d}$
cosinesimil	$d(\mathbf{x}, \mathbf{y}) = 1 - cos\theta = 1 - \frac{\mathbf{x} \cdot \mathbf{y}}{\lVert\mathbf{x}\rVert \cdot \lVert\mathbf{y}\rVert}$ $= 1 - \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}}$ where $\lVert\mathbf{x}\rVert$ and $\lVert\mathbf{y}\rVert$ represent the norms of vectors x and y respectively.	**nmslib** and **faiss:** $score = \frac{1}{1 + d}$ **Lucene:** $score = \frac{2 - d}{2}$
innerproduct (not supported for Lucene)	$d(\mathbf{x}, \mathbf{y}) = -\mathbf{x} \cdot \mathbf{y} = - \sum_{i=1}^{n} x_i y_i$	$\text{If}\, d \geq 0,$ $score = \frac{1}{1 + d}$ $\text{If}\, d < 0, score = -d + 1$

opensearch.org/docs/latest/search-plugins/knn/approximate-knn

Example Vector DB Pipeline on AWS

Vector DB pipelines (such as the right) require an embedding model to convert queries to vectors during indexing and at retrieval time.



1. Data engineer load pre-trained or fined tuned BERT model into SageMaker
2. Data engineer run SageMaker batch job to generate vector for business documents with BERT
3. Store vector data into OpenSearch
4. Client submit search request to API Gateway
5. API Call Lambda backend service in Lambda
6. Backend service call SageMaker Endpoint to convert search query into vector
7. Use OpenSearch k-NN search to get semantic similar documents and return to client

aws.amazon.com/blogs/big-data/amazon-opensearch-services-vector-database-capabilities-explained

Sample Applications & Marketing Data

Vector DB Sample Applications

- Anomaly and Fraud Detection
  - Identifying patterns that reveal fraudulent behavior
  - Providing efficient storage for readily accessible data
- E-Commerce Recommendations
  - Understanding customer preferences & analyzing purchasing behavior
  - Product embeddings capture semantic relationships
  - Create customized experiences for users

https://www.singlestore.com/blog/the-power-of-vector-databases-in-anomaly-detection/
https://www.algolia.com/blog/ai/semantic-search-and-why-it-matters-for-e-commerce/

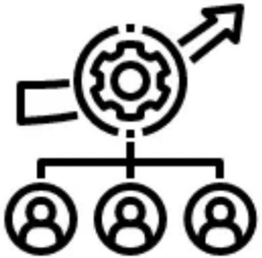
Companies Incorporating Vector DBs

- Microsoft
  - Enables for securely running GenAI Applications within the cloud
  - No need to maintain additional infrastructure
- Notion
  - Incorporates embeddings through workspace data
  - Stores embeddings for retrieval within Pinecone database
- Amazon Web Services
  - Combines RAG Workflow with pre-trained models from Bedrock
  - SageMaker model hosts for LLMs while Pinecone supports knowledge base

Microsoft Notion aws

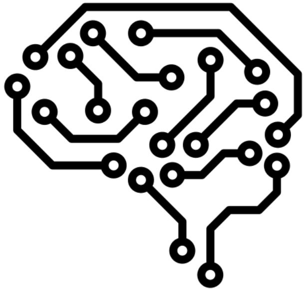
https://www.notion.so/help/notion-ai-security-practices
https://www.pinecone.io/partners/aws/

Case Study: Microsoft Using Pinecone



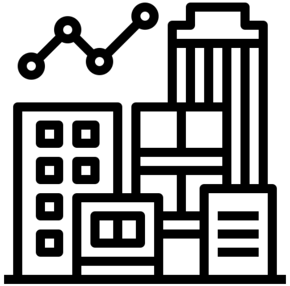
High Performance

Scale beyond billions of vectors without compromising performance



Long-Term Memory

Storing, searching and retrieving data helps provide relevant, quick responses



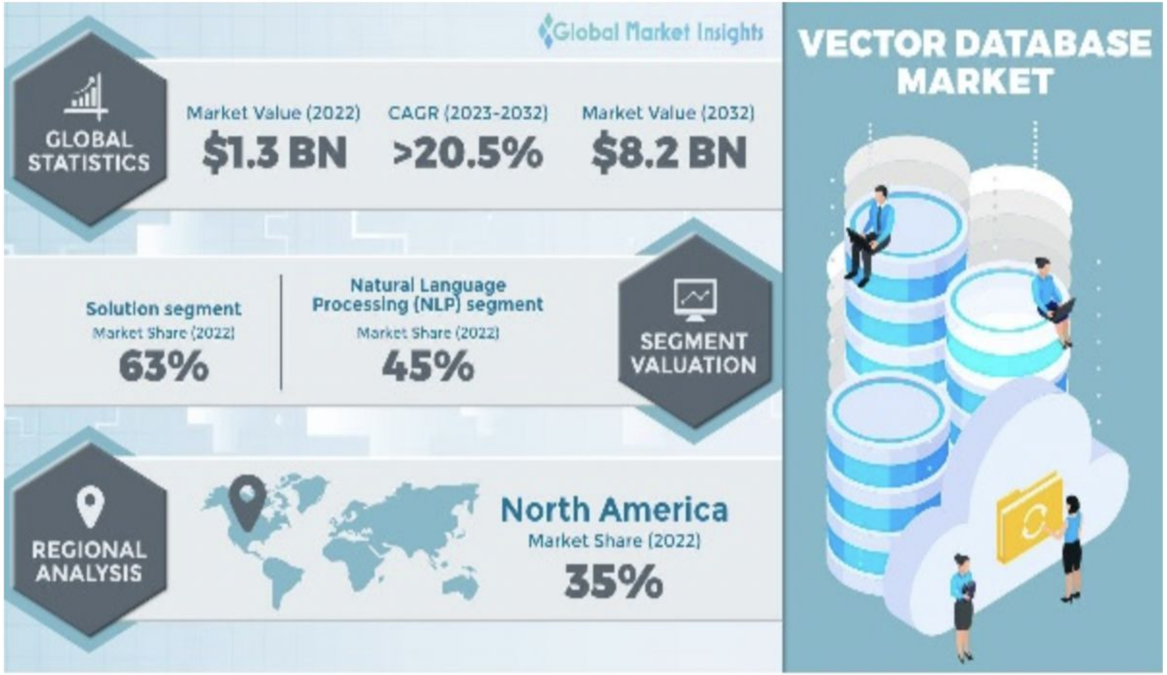
Enterprise Ready

Utilizing data encryption at transit and rest grants enterprise-level security

https://www.pinecone.io/partners/azure/

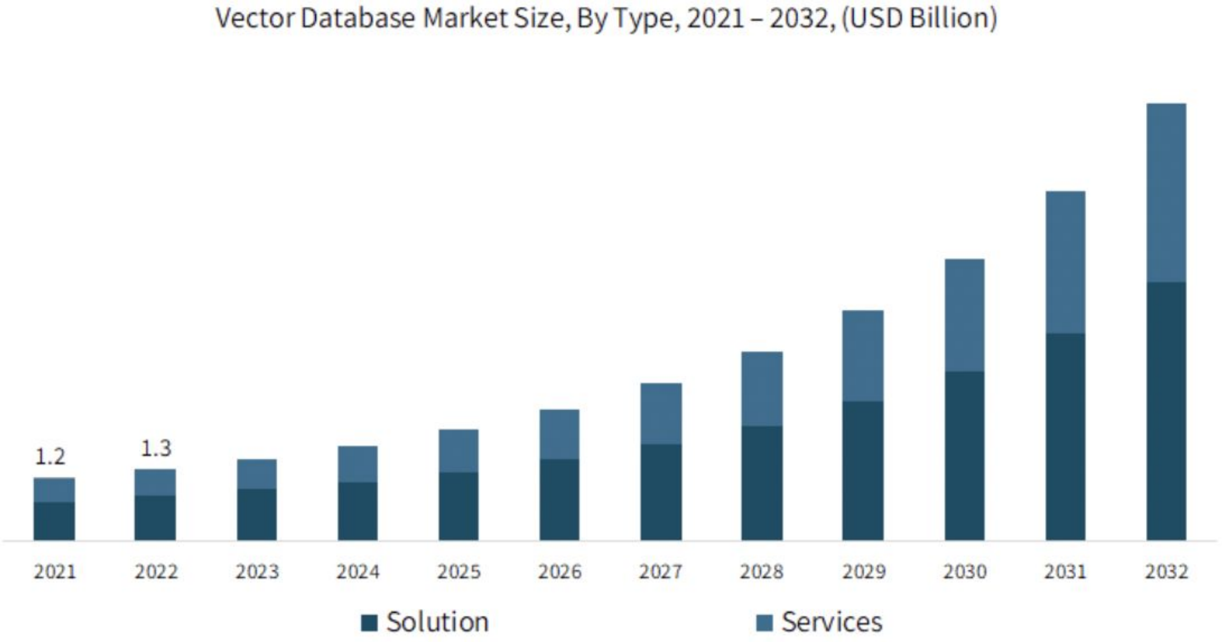
Defining the Vector DB Market

- $1.3 Billion Market Value in 2022 and anticipated for 20.5% compound annual growth rate by 2032
- Covid-19 accelerates digital transformation across industries
- Growth drivers such as real-time analytics and geo-spatial data analysis



https://www.gminsights.com/industry-analysis/vector-database-market

Vector Database Market Size

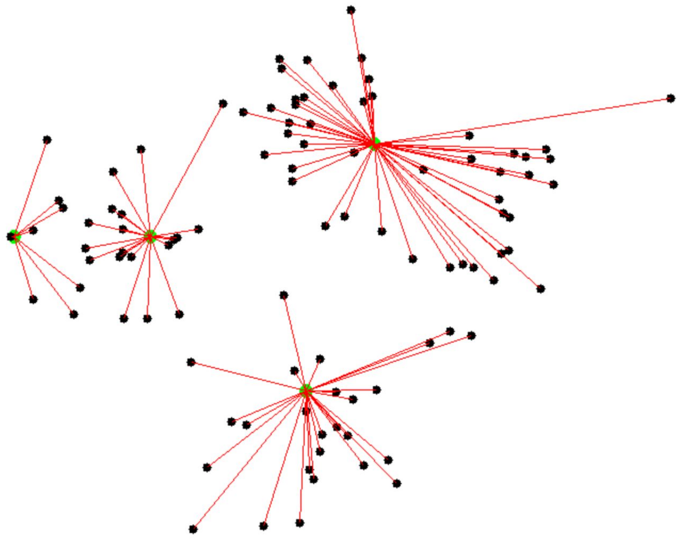


https://www.gminsights.com/industry-analysis/vector-database-market

Current Trends & Issues

Future of Vector Databases

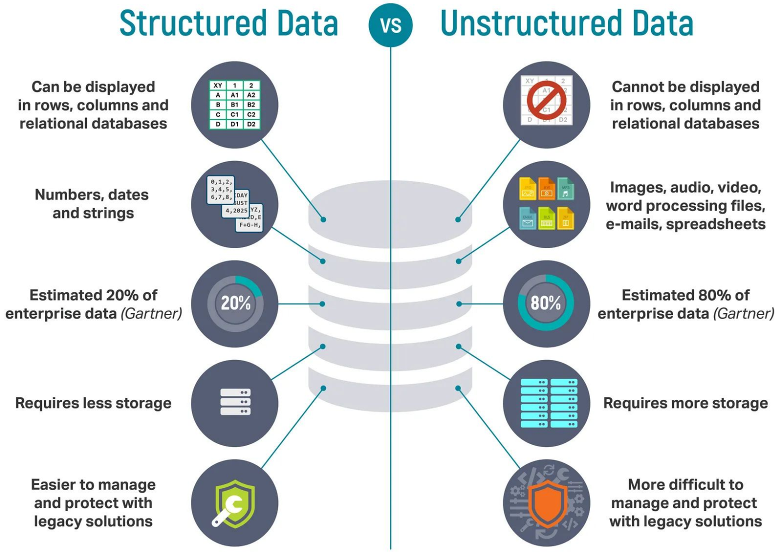
- Increase in functionalities provided
  - Currently, mainly approximate nearest neighbor search
  - Exact search or matching will soon become a reality
- Users can use both functionalities together
- Will likely support additional vector computing functionalities
  - Vector clustering and classification



Zicari, R. V. (2024, January 17). *On The Future of Vector Databases. Interview with Charles Xie*. ODBMS Industry Watch. https://www.odbms.org/blog/2024/01/on-the-future-of-vector-databases-interview-with-charles-xie/
Duhaime, D. (2015, September 12). *Clustering Semantic Vectors with Python*. https://douglasduhaime.com/posts/clustering-semantic-vectors.html

Current Relevance

- Vector databases are more relevant than ever
- Vector databases' biggest strength is ability to work with unstructured data
- Amount of unstructured data is increasing with LLMs
- LLMs deal with unstructured data of all kinds and produce unstructured data as well



Warnecke, T. and Poojary, T. (2023, June 12). *Data Trends: How Vector Databases Are Meeting New Challenges*. Camelot Consulting Group. https://blog.camelot-group.com/2023/06/data-trends-how-vector-databases-are-meeting-new-challenges.

Deep Talk. (2021, October 21). *80% of the world's data is unstructured*. Medium. https://deep-talk.medium.com/80-of-the-worlds-data-is-unstructured-7278e2ba6b73

Current Issues

- Relatively new compared to existing databases so harder to ensure data integrity, consistency, and scalability
- High latency when working with large datasets
- Performing similarity searches and creating vectors can be computationally expensive
- Estimated to cost $125,000 to create vector embeddings for 100,000,000 chunks of data.
    - Chunk is 250 tokens or word fragments
    - $7000 to $8000 per month to maintain this in a vector database

Lim, D. (2023, June 25). *11 known issues of vector-based database used for AI prompting*. Medium. https://medium.com/@don-lim/known-issues-of-vector-based-database-for-ai-ae44a2b0198c

Research: Generalized Vector Databases

- [Zhang et. al 2024] explores how well a generalized vector database performs compared to a specialized vector database
- No fundamental limitation to using a relational database (e.g., PostgreSQL) to support efficient vector data management
- Careful implementation is most important in making this happen
    - Factors that like parallel computing and memory management highlighted
    - In-memory database instead of disk-based

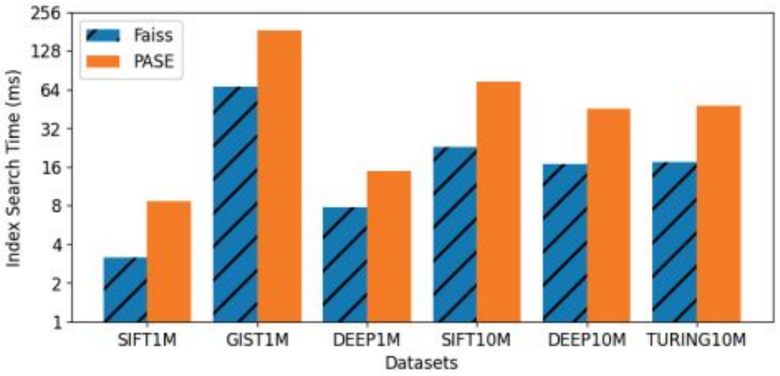
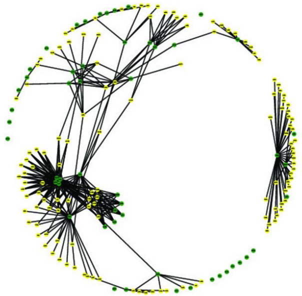


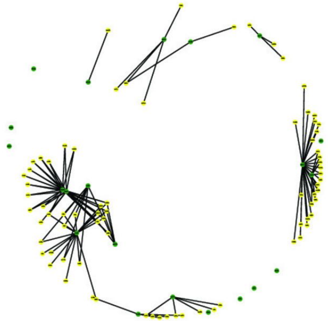
Fig. 14: Search Time for IVF_FLAT

Y. Zhang, S. Liu, and J. Wang, "Are There Fundamental Limitations in Supporting Vector Data Management in Relational Databases? A Case Study of PostgreSQL," Proc. 40th IEEE International Conference on Data Engineering (ICDE) IEEE Transactions on Knowledge and Data Engineering, Utrecht, Netherlands, May 2024.

Research: Data Cleaning in Vector Databases

- [DeCastro-García et. al 2018] explores a method to remove unnecessary data and compute redundancy in vector databases
- Redundant information common in vector databases
- Tested on a cyber database
    - Many sources of information
    - Data is not uniform
- Found high redundancy and approximately two-thirds of the data could be useless for further analysis





N. DeCastro-García, Á. Castañeda, M. Rodríguez, and M. V. Carriegos, "On Detecting and Removing Superficial Redundancy in Vector Databases", in Mathematical Problems in Engineering, Vol. 2018, May 2018.

Questions