Iris

Cedric Sirianni, Mithi Jethwa, Lachlan Kermode

ACM Reference Format:

Cedric Sirianni, Mithi Jethwa, Lachlan Kermode. 2024. Iris. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/nnnnnnn.nnnnnnn

PROBLEM

Vector databases are growing in popularity as they become widely used in image similarity search and RAG systems. The current approaches for distributed vector databases port existing notions from the distribution of column-based databases (using strategies such as replication and sharding) without taking specific advantage of a vector database's unique architectural features.

PROJECT IDEA

NOVELTY

IMPLEMENTATION

We first aim to appraise and benchmark the SoTA of vector database distribution. From a preliminary search of recent literature in vector databases and commercial offerings, we understand there to be three ways in which vector databases have been distributed:

(1) Replication, where all vectors are stored on all nodes, and a master node load-balances new queries. Supported in Qdrant, Weaviate.
(2) Random partitioning, where each node contains a distinct set of vectors. An incoming query is sent to all nodes, and results are aggregated and pruned in the user result. Supported in Qdrant, Milvus.
(3) HNSW-aware sharding, where some number of Voronoi cells is stored on each node. Incoming queries can thus be directed only to those nodes where there are vectors proximate to the query. The HNSW index is stored entirely on the coordinator node [cite:@sunDistributedSystemLarge2024;@dengPyramidGeneralFramev

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

RESOURCES

EVALUATION

TEAM MEMBERS

EXPECTED CHALLENGES

BIBLIOGRAPHY