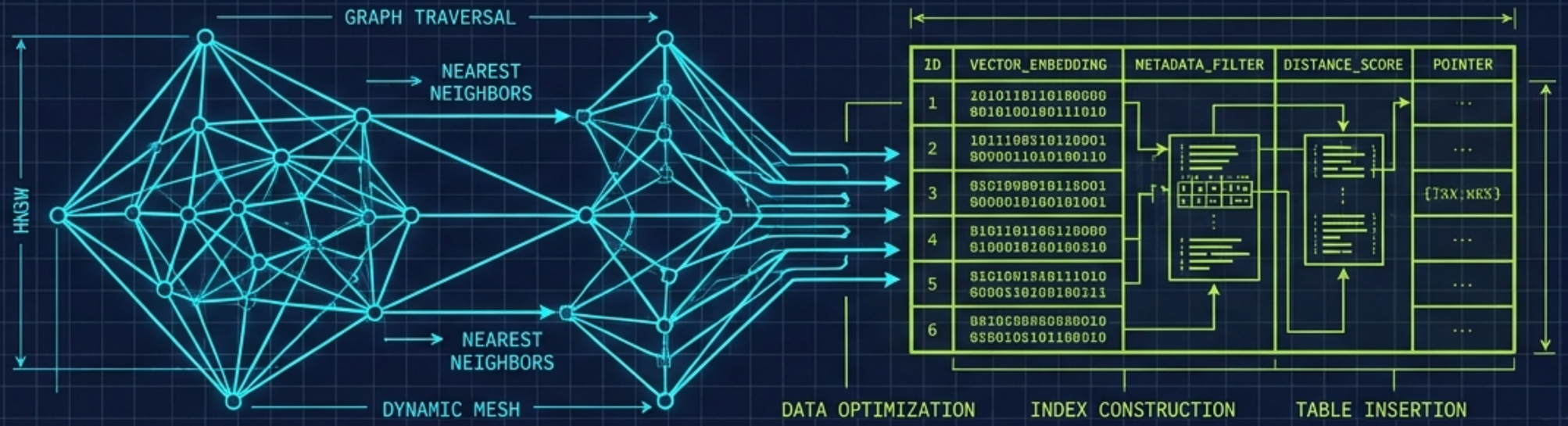
MODERN DATA STORES // S05E05



Embeddings, HNSW vs DiskANN, Filtered ANN & pgvector

An Engineering Deep Dive into Vector Search Mechanics, Memory Math, and Architectural Trade-Offs.

TARGET: SOFTWARE ENGINEERS & TECH LEADS

Gemini Notebook

THE PIPELINE RECAP



POSTGRESQL
(Source of Truth)

CDC EVENT STREAM
(S05E04)

DERIVED VECTOR INDEX

THE PHYSICS OF EMBEDDINGS

DENSE EMBEDDINGS

[0.1234, -0.4567, 0.8901, 0.2345, ...]

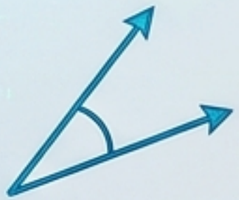
High-dimensional float arrays capturing semantic meaning.

SPARSE EMBEDDINGS

{ 1: 1, 3: 2, 5: 3 } / 5

High-dimensional, mostly zero arrays capturing exact keyword frequency.

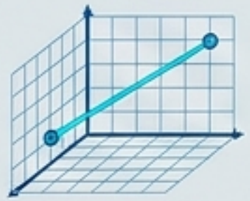
DISTANCE METRICS



Cosine:
Angle between vectors.



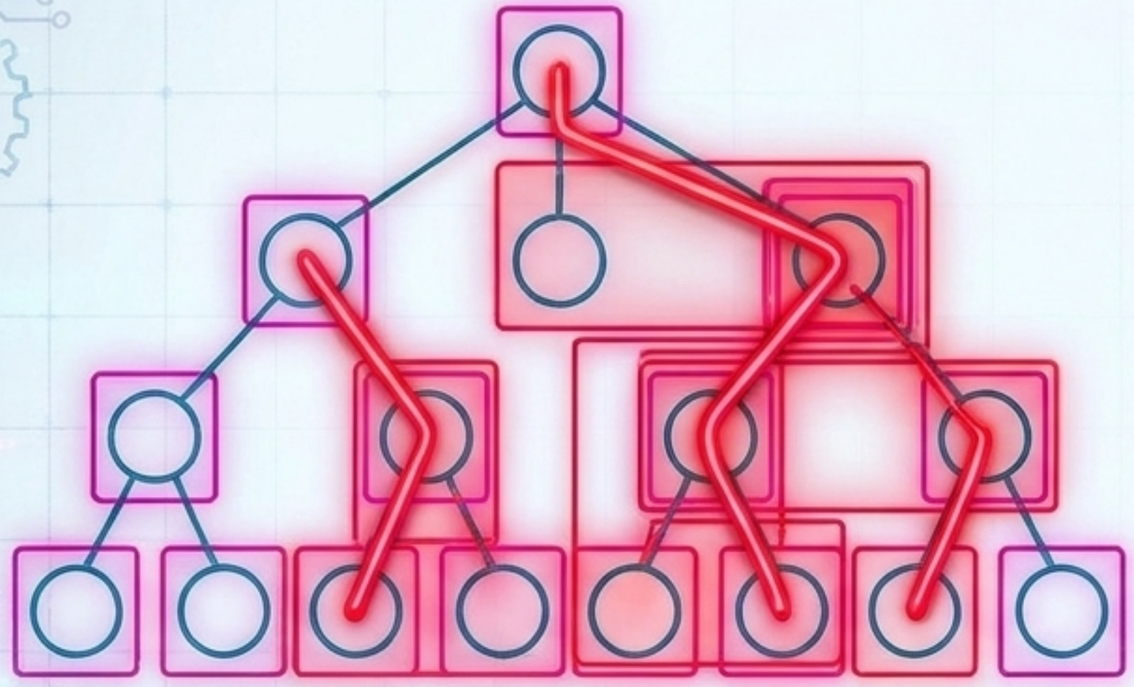
Dot-Product:
Projection magnitude.



L2 (Euclidean):
Straight-line distance.

Gemini Notebook

THE DIMENSIONALITY CURSE



In 768–3072 dimensions, bounding volumes overlap completely. The tree degrades, forcing an O(N) full table scan.

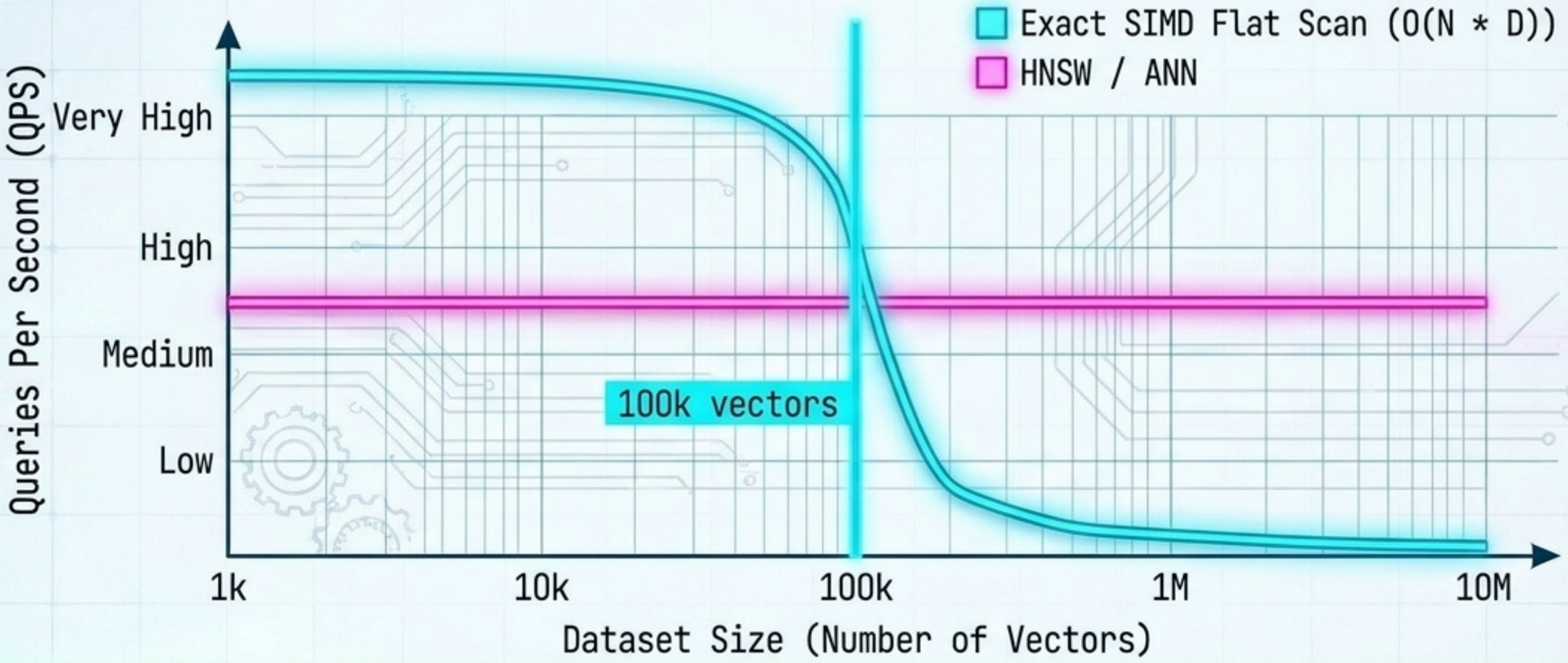
ARCHITECTURAL INSIGHT

EMBEDDING QUALITY > INDEX CHOICE

According to MTEB (Massive Text Embedding Benchmark) realities, the quality of your chunking strategy and embedding model impacts final search relevance far more than whether you use HNSW, IVF, or DiskANN. The index just dictates speed and RAM.

Gemini Notebook

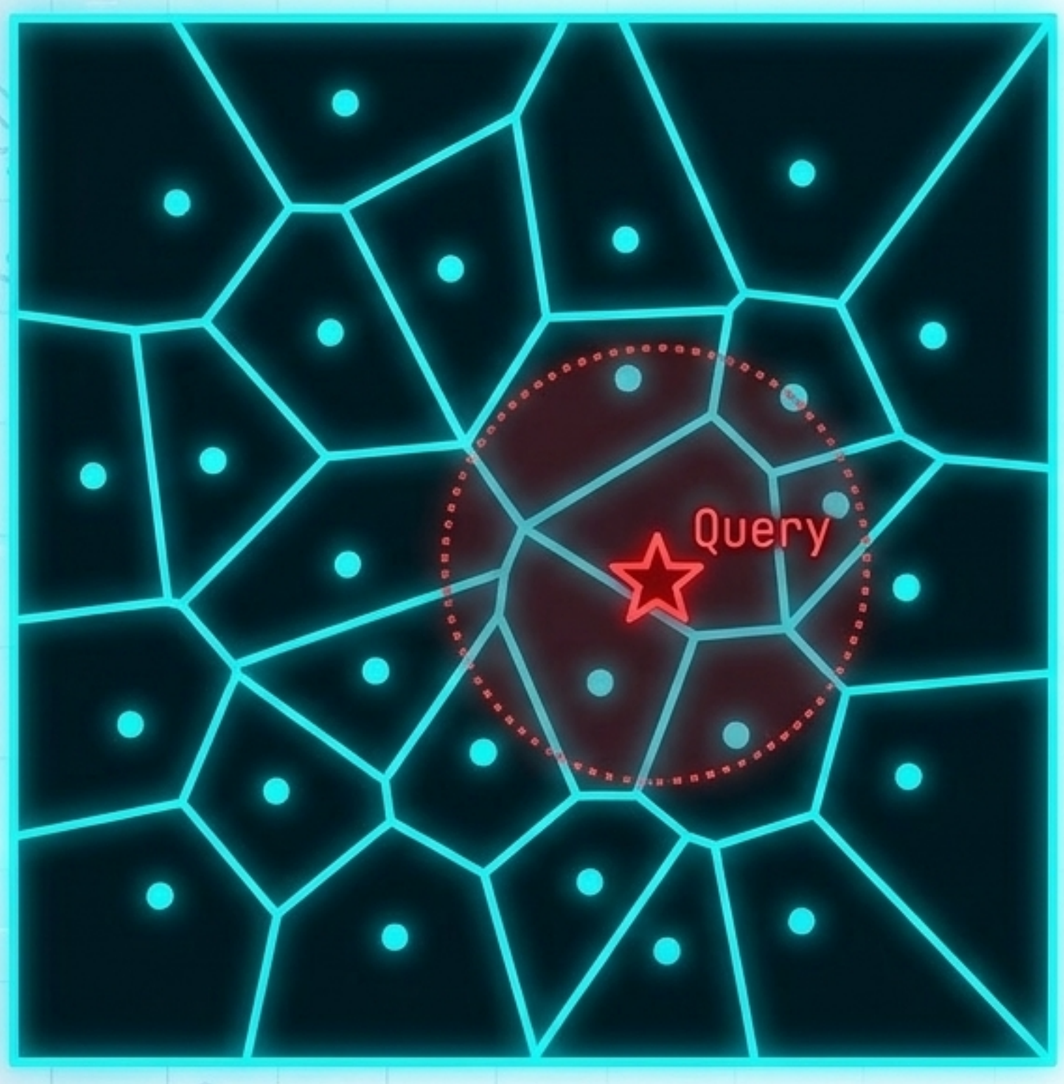
THE BASELINE — EXACT SIMD FLAT SCAN



THE BASELINE: EXACT kNN. For < 100,000 vectors, a brute-force SIMD-accelerated flat scan is often faster (and perfectly accurate) than traversing a complex graph. Do not pay the ANN memory tax prematurely.

Gemini Notebook

IVF (Inverted File) — Clustering the Space



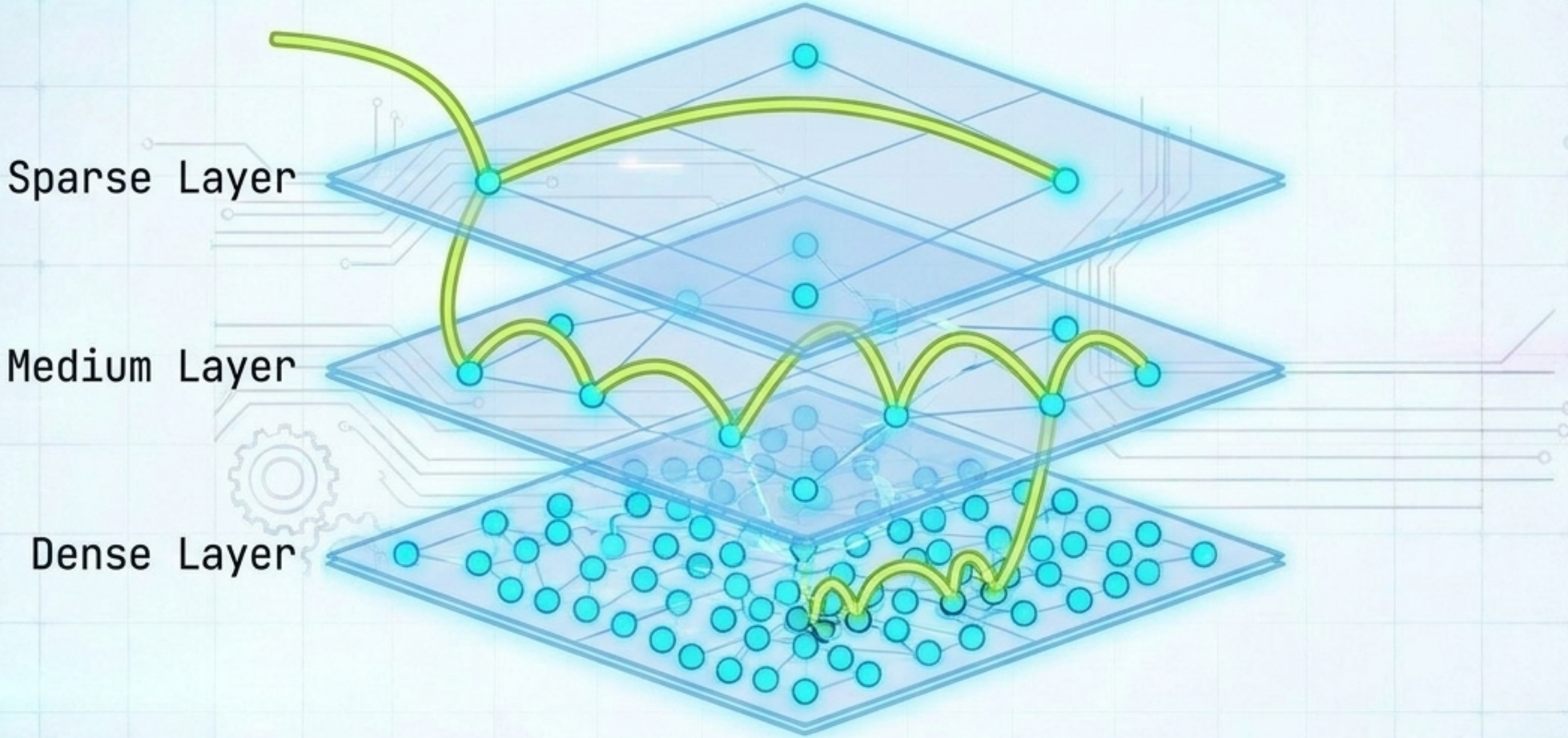
`nlist`: The number of k-means centroids (cells). Dictates clustering granularity.

`nprobe`: The number of boundary cells searched.

The Trade-Off: Increasing nprobe trades latency (slower search) for recall (catching boundary-spanning true neighbors).

Gemini Notebook

HNSW (Hierarchical Navigable Small World) — The Industry Standard



Fast proximity graphs using skip-list principles. Upper layers act as a coarse filter; the bottom layer ensures fine-grained accuracy.

Gemini Notebook

HNSW Parameter Tuning & The RAM Tax



efConstruction

Boosts final index recall quality, but drastically tanks index build speed.



M (Edges per node)

Denser graph. Increases recall but aggressively balloons memory consumption due to bidirectional graph links.



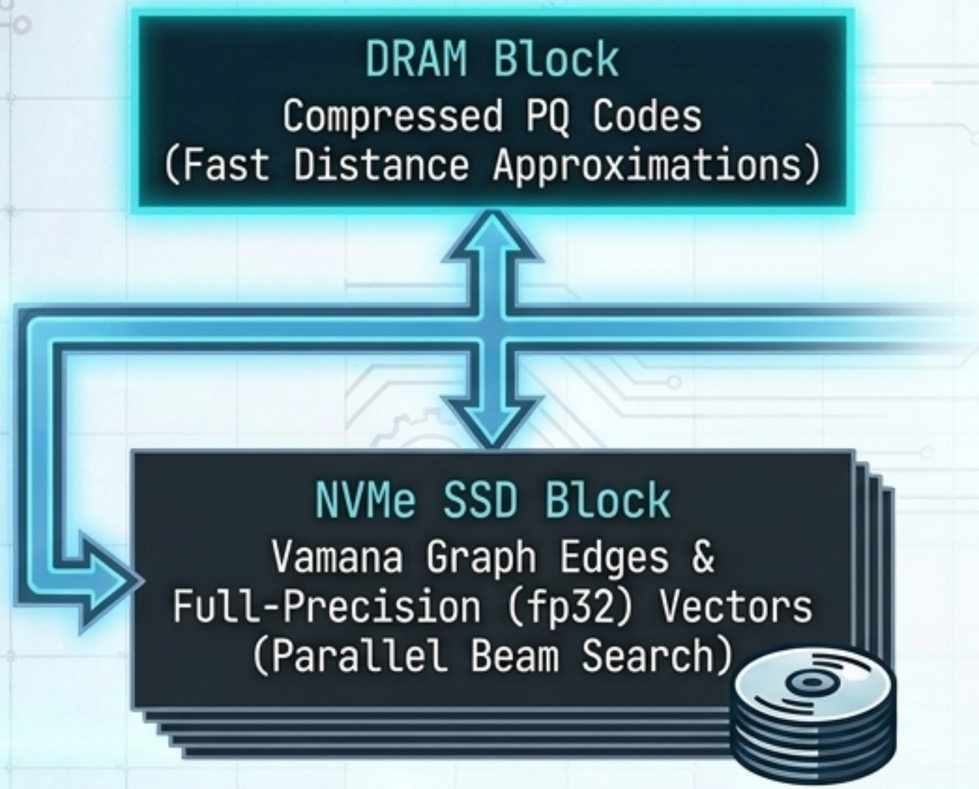
efSearch / ef

Adjustable at query time. Expands the search frontier. Higher value equals higher recall at the cost of milliseconds of latency. Does not require index rebuild.

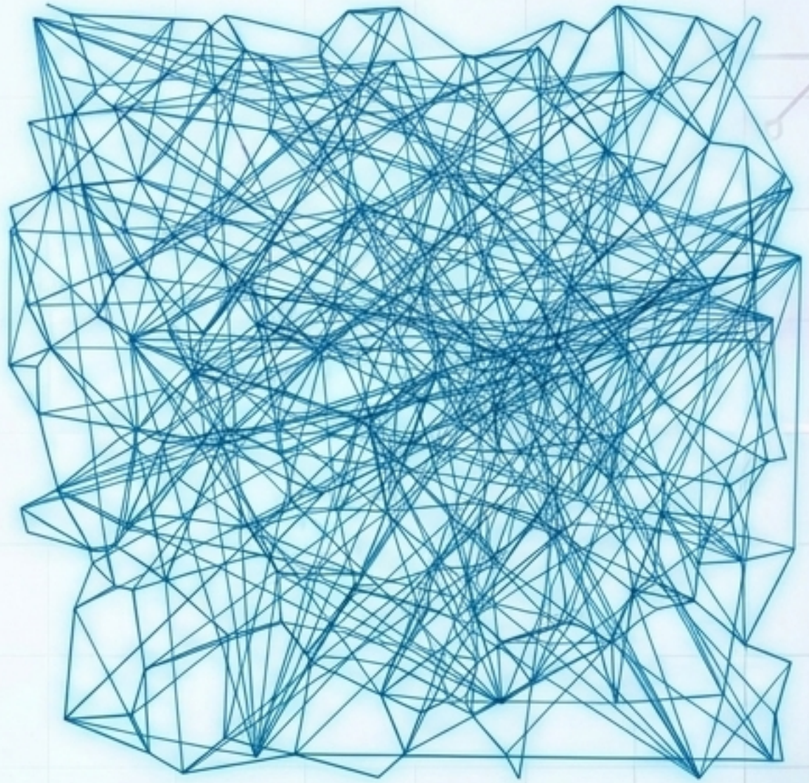
Gemini Notebook

DiskANN — Breaking the Memory Barrier

HARDWARE SCHEMATIC



THE VAMANA GRAPH



DiskANN breaks the memory barrier. By storing the Vamana graph on NVMe and only keeping PQ codes in RAM, it searches billion-point datasets on a single node.

Gemini Notebook

The Vector Storage Landscape



Tightly Coupled / ACID Relational

Ideal for zero CDC lag, strict ACID metadata joins, and monolithic simplicity.

The Middle Ground

Massive lexical search ecosystems bolting on vector capabilities.

Purpose-Built / Distributed ANN

Built entirely around distributed graph traversals, sharding, and memory-mapped NVMe optimization.

Gemini Notebook

The Ultimate Engine Comparison Matrix

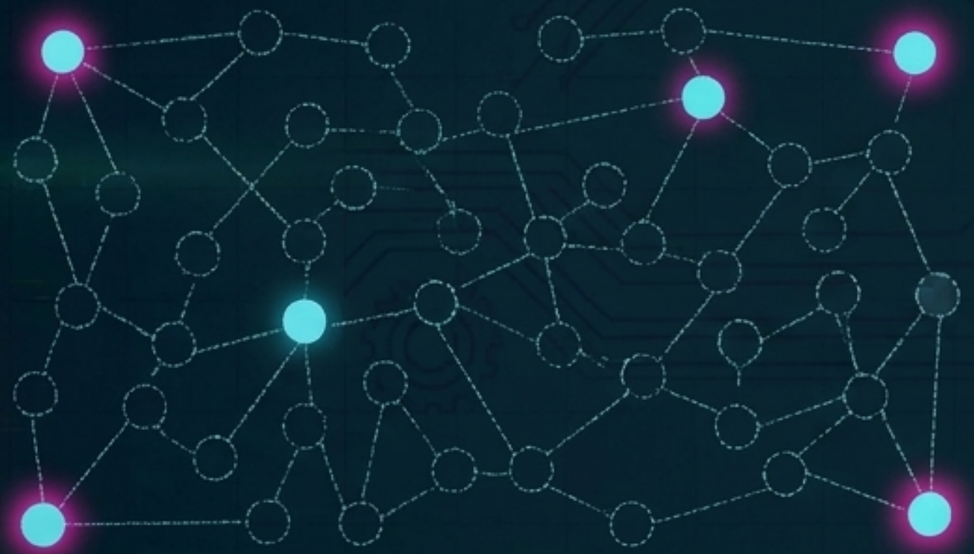
Metric	PostgreSQL (tsvector)	pgvector (hnsw/halfvec)	Elasticsearch	Dedicated (Qdrant/Milvus)
Scale Sweet Spot	0 vectors	< 50 Million	10M - 100M	50M - 1 Billion+
p99 Latency (Dense)	N/A	Low	Medium	Ultra-Low
RAM per 10M Vectors (fp32)	N/A	~80GB+	~80GB+	Tunable (DiskANN / Binary)
ACID Joins	Perfect	Perfect	Eventual	Eventual / Pre-filtered
CDC Sync Tax	None	None	High (Dual Write)	High (Dual Write)

Gemini Notebook

The Metadata Filtering Trap

Why standard SQL WHERE clauses break vector indexes.

Pre-filtering Failure



Pre-filtering severs the proximity graph. A query cannot navigate the skip-list if if the bridging nodes are filtered out by a 1% selectivity clause.

Post-filtering Failure



Post-filtering retrieves ‘k’ results first. If the filter has 80% selectivity, all ‘k’ results might be dropped, returning zero rows to the user.

Gemini Notebook

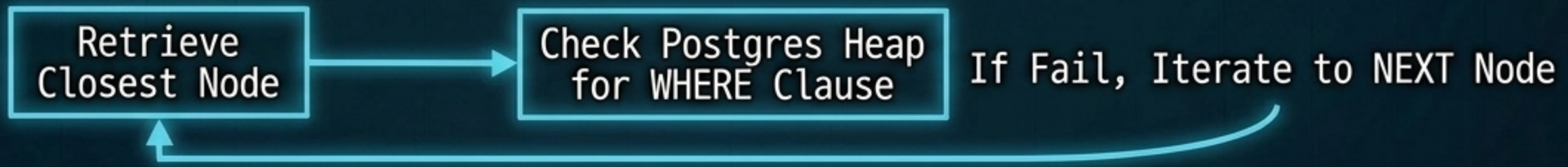
Fixing the Filter: Modern Engine Solutions

Dedicated Engines (Qdrant / Weaviate)



Roaring bitmaps act as a pre-filter mask layered OVER the HNSW traversal, ensuring greedy search only evaluates valid nodes without losing connectivity.

pgvector 0.8+ (Iterative Scans)

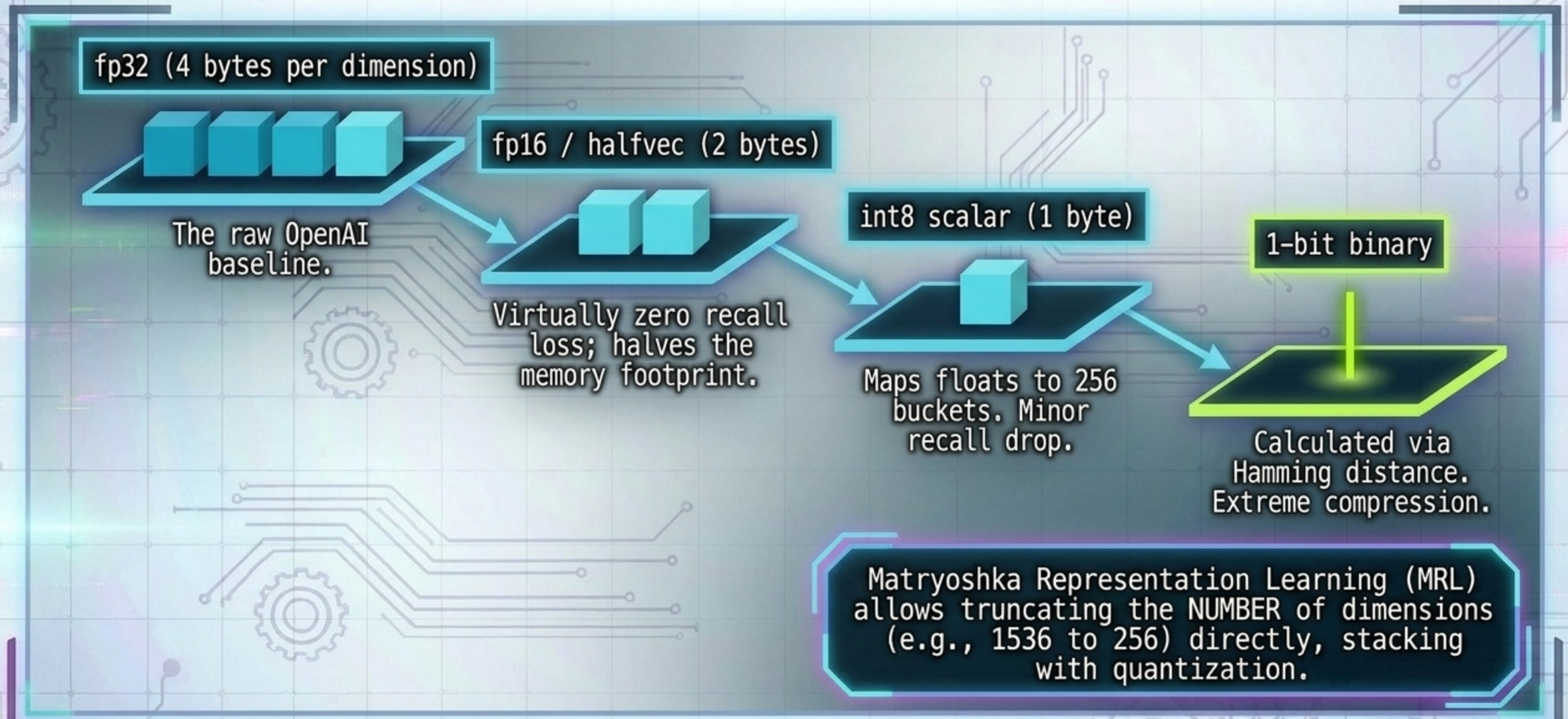


The index scan iterates deeper into the graph on the fly until it finds ‘k’ valid results that satisfy the Postgres heap condition.

Gemini Notebook

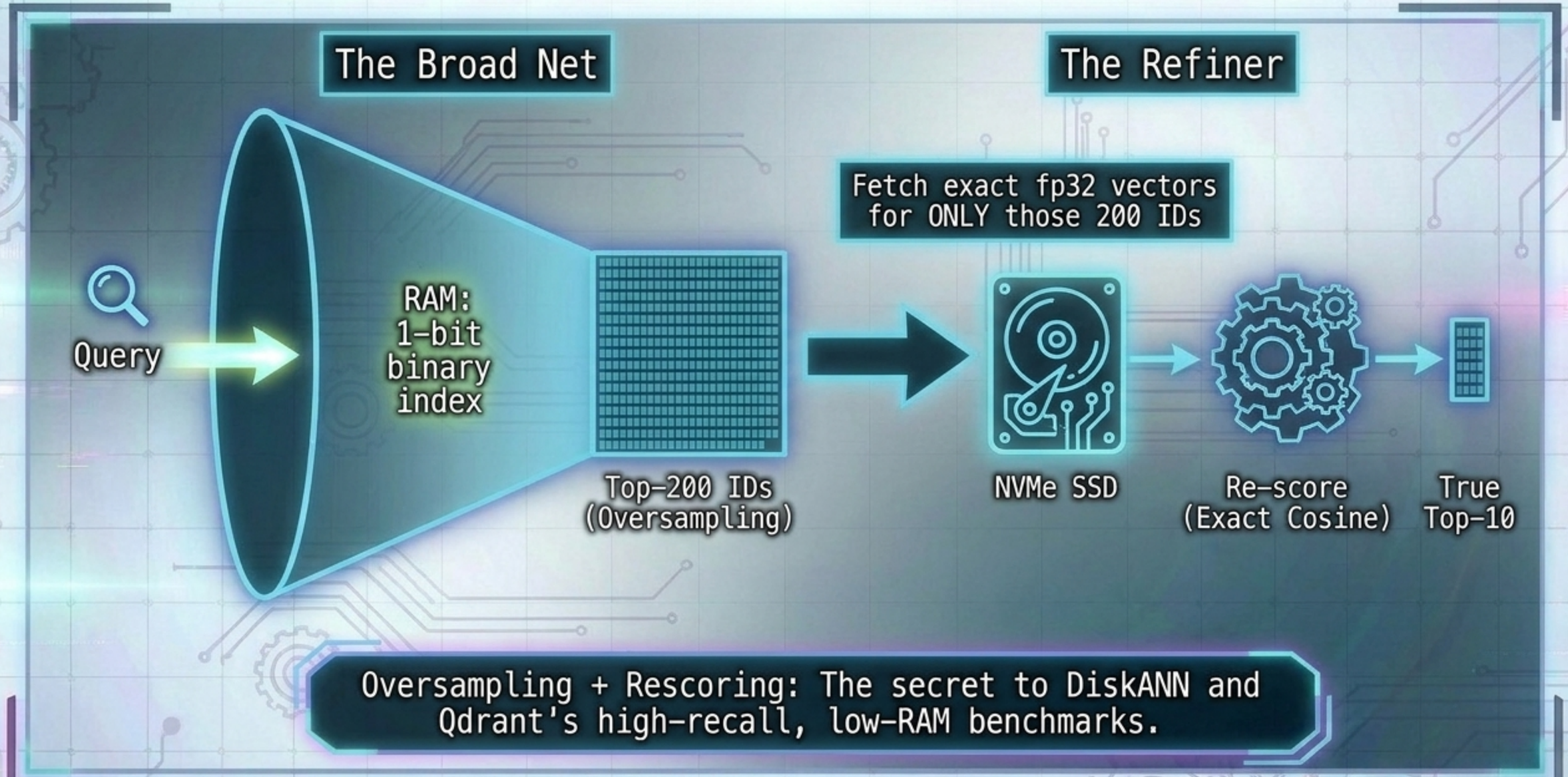
The Quantization Ladder

Compressing vectors to survive memory limits.



Gemini Notebook

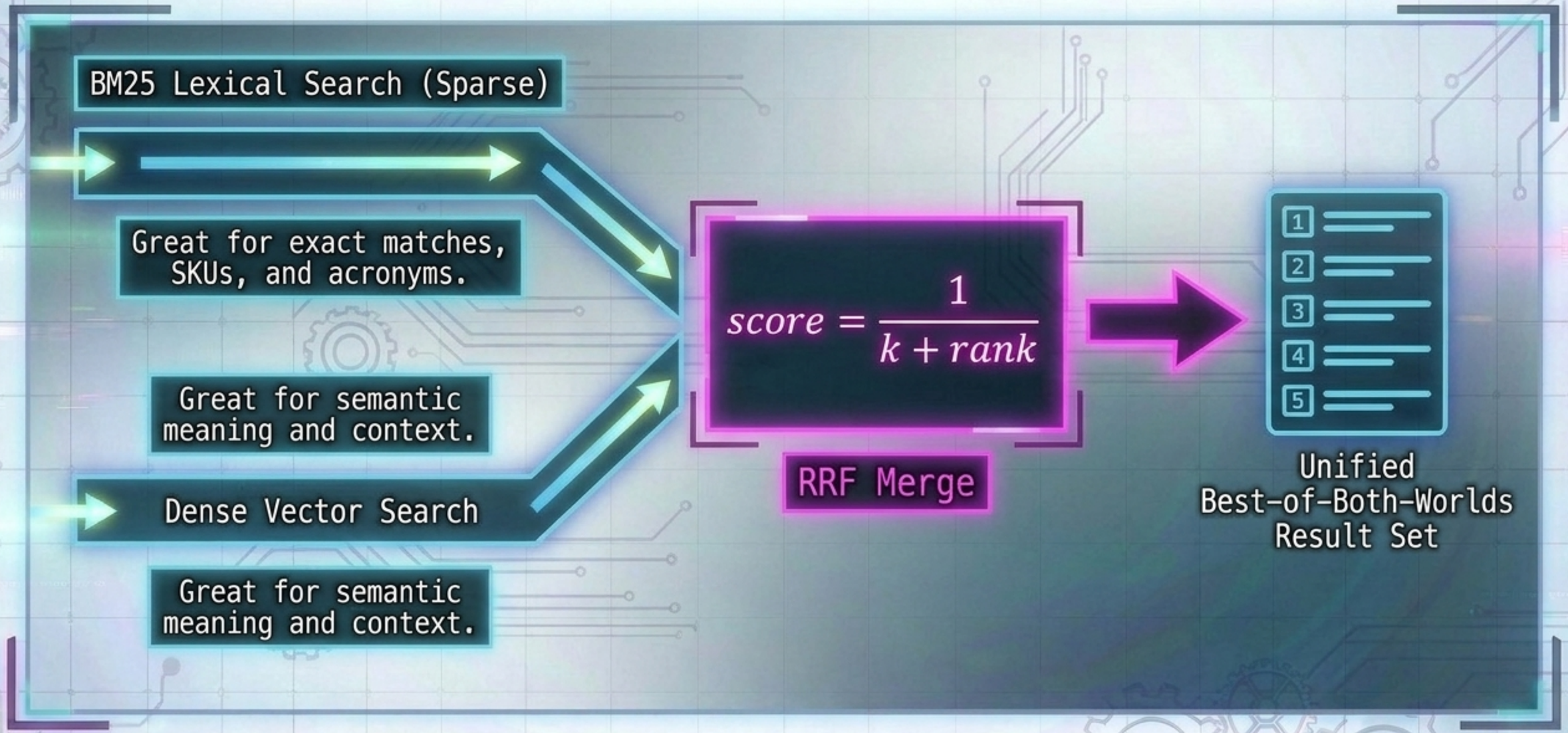
Oversampling, Rescoring & Matryoshka



Gemini Notebook

Hybrid Search & Reciprocal Rank Fusion (RRF)

Vectors fail on exact SKUs and jargon. Hybrid fixes it.



Gemini Notebook

The Raw Math: Sizing 25 Million Vectors

```
Workload: 25M vectors, 1536-dim, fp32

Step 1 (Raw Vector Mass):
  25,000,000 vectors * 1,536 dimensions * 4 bytes = 153.6 GB

Step 2 (HNSW Overhead):
  Bidirectional graph links (M=32) = ~35% overhead

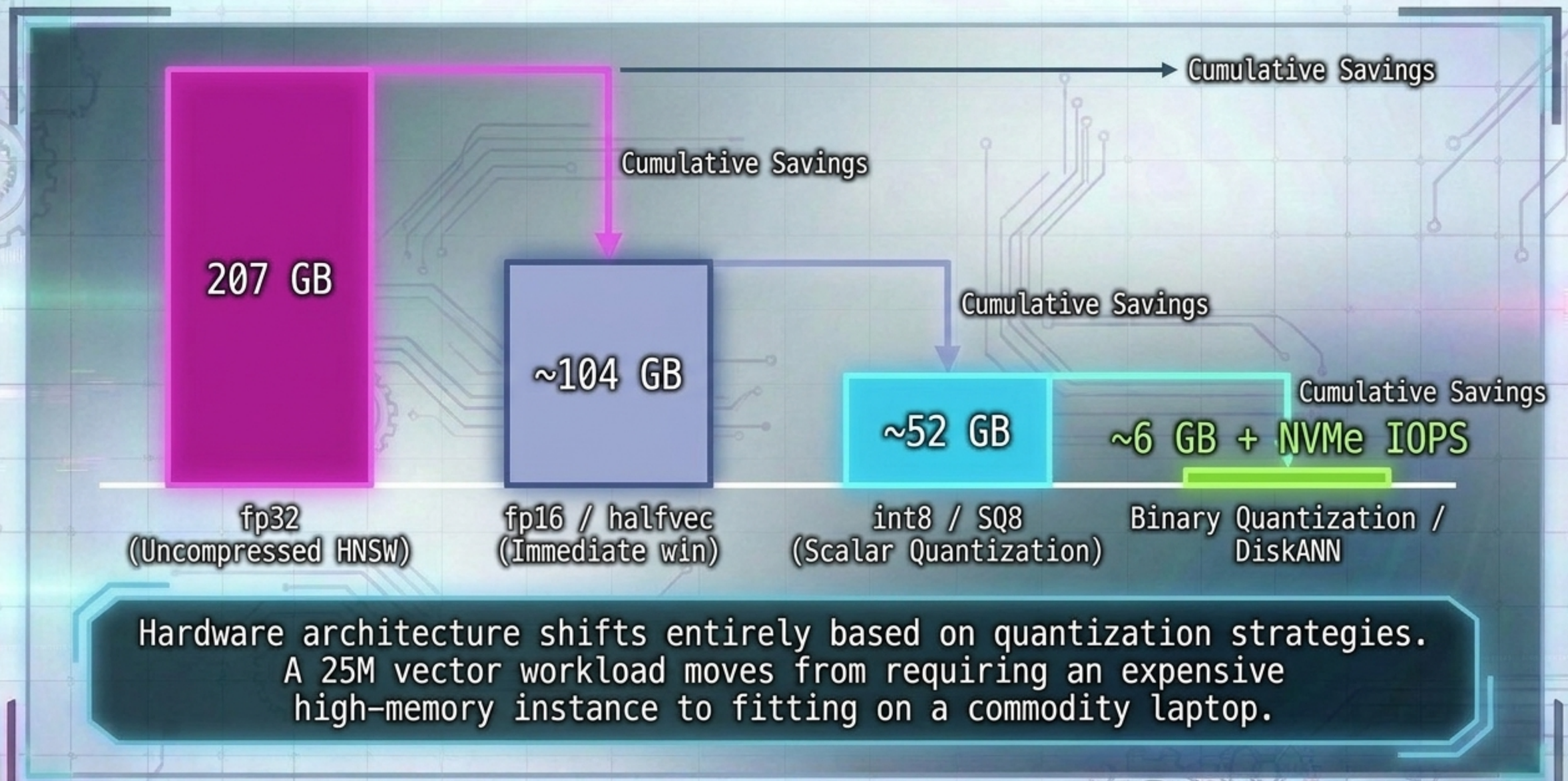
Step 3 (Total Pinned RAM):
  153.6 GB * 1.35 = ~207 GB
```



You need ~207 GB of pinned DRAM just to hold this index. If this spills to swap on a standard SSD, query latency will be catastrophic.

Gemini Notebook

The Quantization Savings (Waterfall)

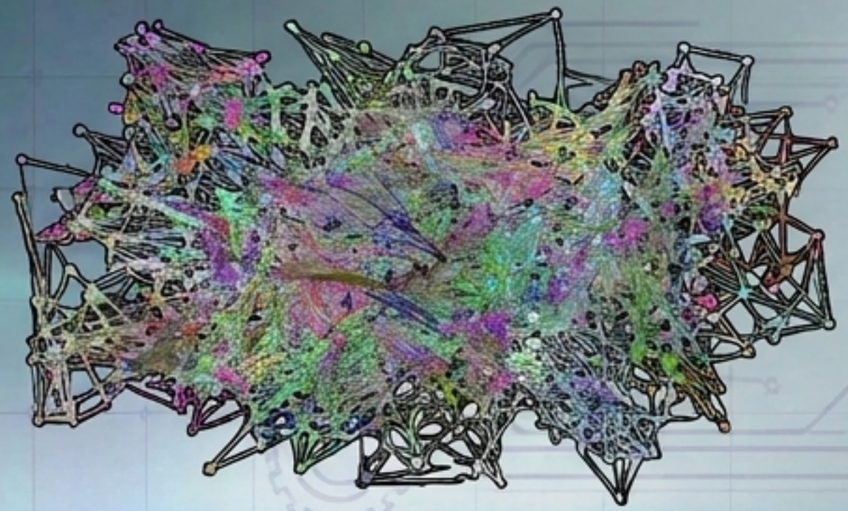


Hardware architecture shifts entirely based on quantization strategies. A 25M vector workload moves from requiring an expensive high-memory instance to fitting on a commodity laptop.

Gemini Notebook

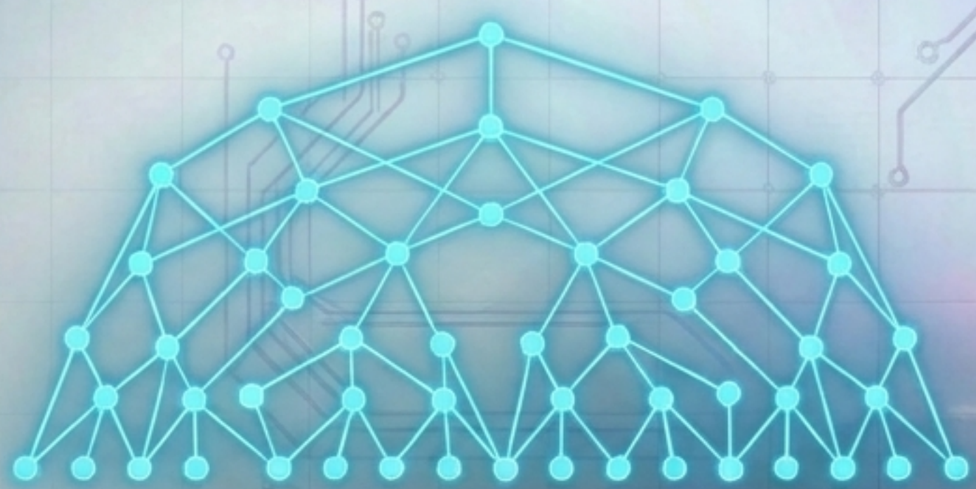
Anti-Pattern 1 — Multi-Tenant Graph Sprawl

Multi-Tenant Graph Sprawl



Multi-Tenant Graph Sprawl: Creating a dedicated index per user eats massive memory overhead and stalls build times.

Shared Payload-Filtered Index

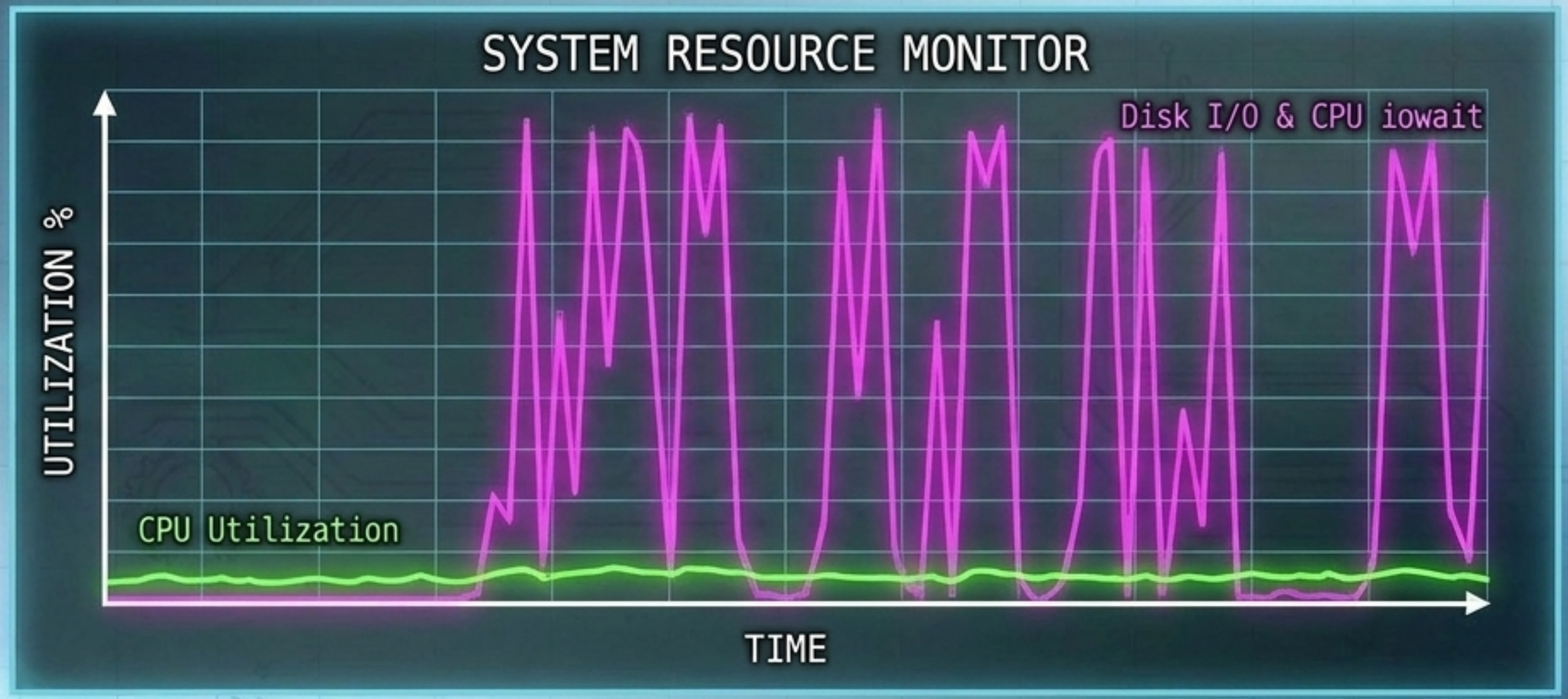


Shared Payload-Filtered Index: Use one large index with efficient metadata filtering, or fallback to exact flat-scans for tiny tenants.

WARNING: Embedding Model Migrations. If you switch from OpenAI text-embedding-ada-002 to text-embedding-3-small, you must re-index the ENTIRE table. Vectors exist in relative space; different models do not map to each other.

Gemini Notebook

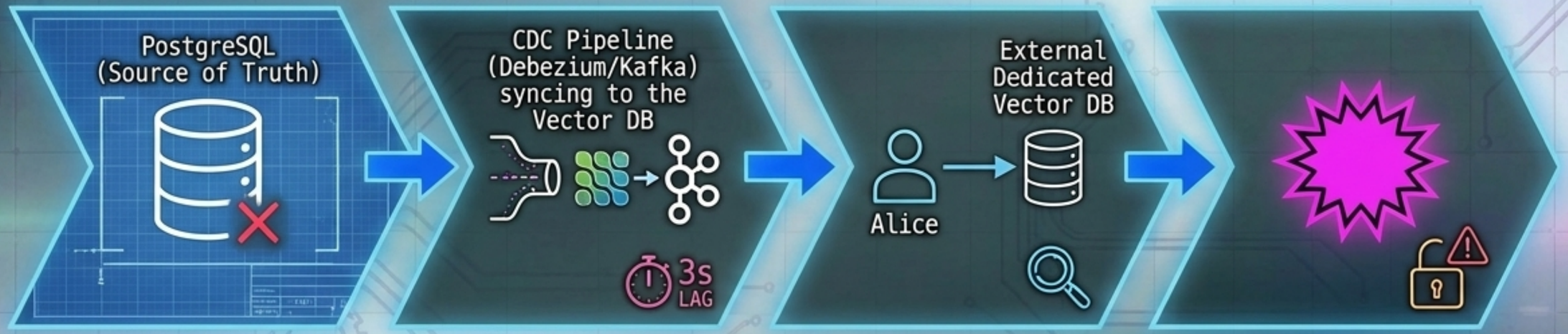
Anti-Pattern 2 — mmap Thrashing & Cloud Disks



mmap Thrashing on Cloud Disks: Relying on memory-mapped files for an index larger than your RAM will cause the OS to constantly page data in and out. If your AWS EBS or cloud volume has low provisioned IOPS, your search latency will spike from 10ms to 5 seconds while waiting on disk fetches.

Gemini Notebook

Anti-Pattern 3 — The Eventual-Consistency Security Leak



revokes user Alice's access to a sensitive document.

encounters a temporary 3-second lag.

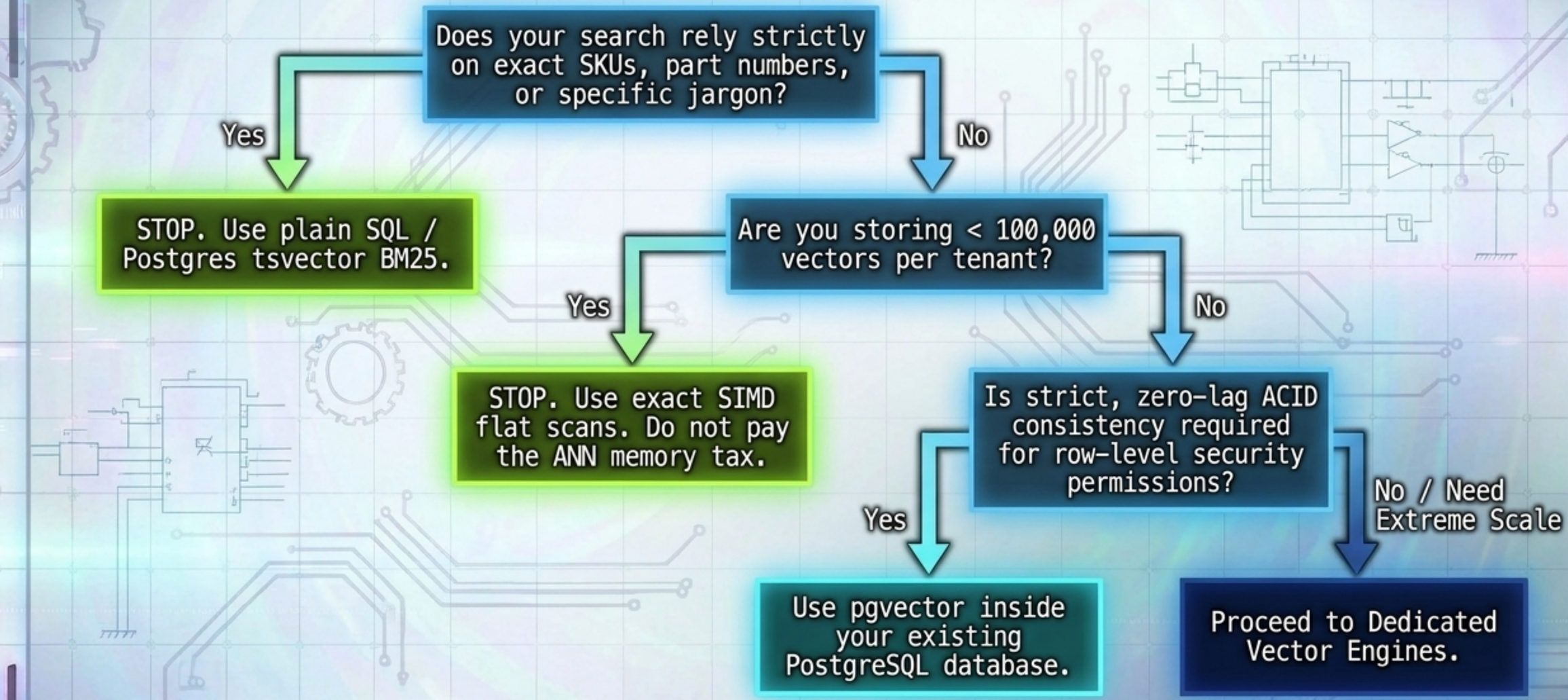
immediately queries the external Dedicated Vector DB.

The Vector DB returns highly classified embeddings Alice shouldn't see because it hasn't processed the ACL revocation yet.

Separating vector search from your primary relational database introduces eventual-consistency security leaks.

Gemini Notebook

Do You Need a Vector DB? (Decision Tree)

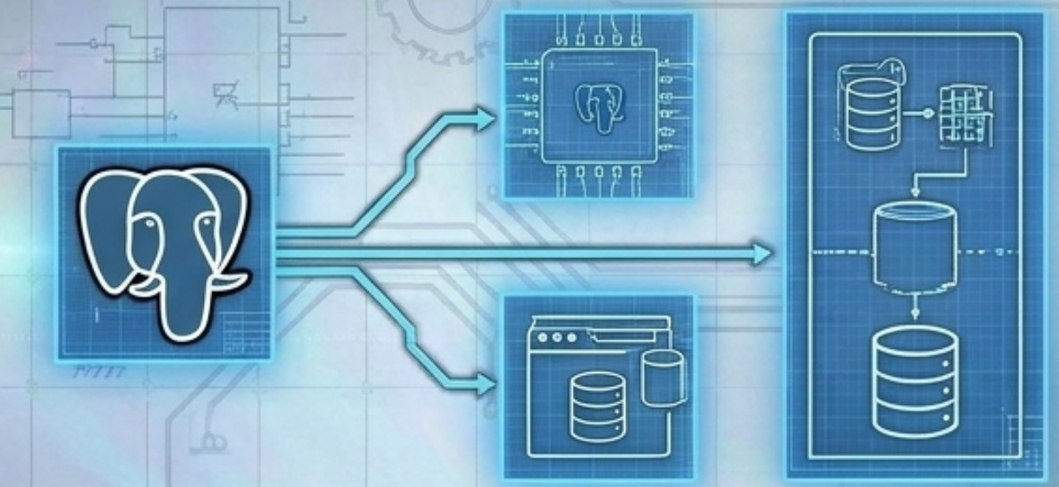


Gemini Notebook

Decision Boundaries: pgvector vs Dedicated Engines

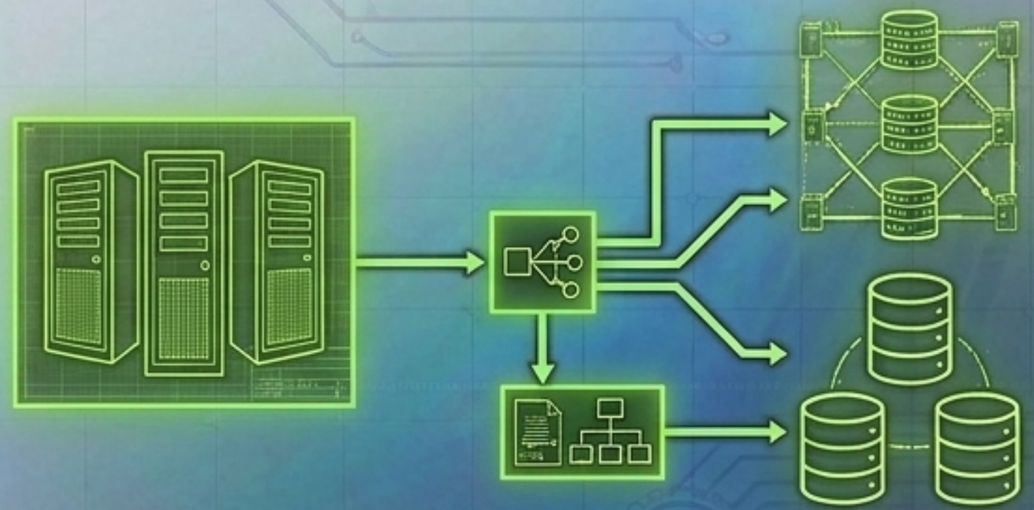
pgvector + halfvec + Iterative Scans

- Best for < 10M–50M vectors.
- Ideal when ACID metadata joins are required.
- Zero sync lag and unified backups are mission-critical.



Dedicated ANN (Qdrant/Milvus/DiskANN)

- Best for 50M–1B+ vectors.
- Ideal for extreme QPS.
- Purpose-built multi-vector sharding and memory-tiering out of the box.



Gemini Notebook

Next Up on Modern Data Stores...

S05E06: Running Data Stores in Production

- High Availability architectures & split-brain recovery.
- Understanding Raft and Paxos consensus mechanics.
- Point-in-Time Restore (PITR) and disaster recovery strategies.

A reading deck is a visual document with embedded context.

Gemini Notebook