Why Cosdata: Context-Aware Retrieval Infrastructure for AI-Native Applications

Nithin Mani (Cosdata) - www.cosdata.io

2025-08-06

Contents

1 Executive Summary

Cosdata builds next-generation retrieval infrastructure for AI-native applications that demand relevance beyond simple vector similarity. Our platform combines multi-modal search capabilities, context-aware ranking, and enterprise-grade features to deliver retrieval systems that understand user intent and real-world complexity.

1.1 The Problem

Traditional vector databases treat retrieval as a storage and similarity problem, optimizing for cosine similarity rather than user relevance. But decades of search evolution prove that effective retrieval requires purpose-built ranking systems that understand context, incorporate multiple signals, and optimize for what users actually find useful—not just mathematical proximity in embedding space.

1.2 Our Core Differentiator

- **Multi-modal retrieval** combining BM25, HNSW dense vectors, SPLADE embeddings, and metadata-rich sparse vectors
- **Relevance-first architecture** that optimizes for user satisfaction rather than cosine similarity
- **Context-aware capabilities** with geofencing, hierarchical document organization, and explainable ranking
- **Enterprise-ready platform** with colocated storage, streaming ingestion, versioning, and comprehensive security

1.3 Business Impact

Organizations using Cosdata achieve 60-120% reduction in compute requirements while improving retrieval quality by 20-50% (NDCG@10). Our unified architecture eliminates the need for external document stores and complex multi-database queries, reducing infrastructure costs and latency. Implementation takes 3-4 weeks with immediate performance gains and predictable scaling costs.

2 Technical Architecture and Capabilities

2.1 Multi-Modal Retrieval Approach

2.1.1 BM25 (Full-Text Search)

- Classical text retrieval with proven effectiveness
- Optimized implementation for modern workloads
- Hybrid scoring with vector-based methods

1. Performance Benchmark: Cosdata vs Elasticsearch (Full Text Search)

- Custom BM25 implementation delivers up to 151× higher QPS than *ElasticSearch* on the *SciFact* dataset.
- Shows an average improvement of ~44× QPS across multiple IR/BM25 benchmark datasets.
- Offers up to 12× faster indexing than *ElasticSearch* on large datasets.
- Achieves lower latency at both p50 (median) and p95 percentiles across all tested datasets.
- Maintains comparable recall and NDCG scores (ranking quality).

Dataset (Corpus Size)	System	Indexing (sec)	Queries Per Sec.	NDCG@10	Latency p50 (ms)	Latency p95 (ms)
arguana (8,674)	Cosdata	0.1	2167	0.40	9	15
	ElasticSearch	1.4	263	0.48	44	74
climate-fever (5.4M)	Cosdata	40.6	135	0.13	106	379
	ElasticSearch	522.8	84	0.14	162	263
fever (5.4M)	Cosdata	40.3	314	0.47	52	157
	ElasticSearch	525.7	154	0.52	80	138
fiqa (57K)	Cosdata	0.5	4942	0.25	7	12
	ElasticSearch	6.7	251	0.25	39	60
msmarco (8.8M)	Cosdata	57.7	315	0.23	46	162
	ElasticSearch	714.7	166	0.23	73	129
nq (2.6M)	Cosdata	19.3	483	0.29	30	81
	ElasticSearch	243.2	197	0.29	59	100
quora (522K)	Cosdata	2.7	1425	0.81	11	36
	ElasticSearch	30.2	323	0.81	39	55
scidocs (25K)	Cosdata	0.3	13338	0.16	7	12
	ElasticSearch	3.6	319	0.15	33	48
scifact (5,183)	Cosdata	0.1	40909	0.69	7	13
	ElasticSearch	1.0	271	0.68	34	51
trec-covid (171K)	Cosdata	1.7	2219	0.61	10	18
	ElasticSearch	22.1	110	0.62	57	88
webis-touche2020 (382K)	Cosdata	5.5	2789	0.34	10	18
	ElasticSearch	63.1	108	0.34	62	99

2.1.2 HNSW Dense Vectors

- Hierarchical Navigable Small World graphs for efficient ANN search
- Support for multiple embedding models and dimensions.
- Advanced high-performance metadata based filtering
- Dynamic index updates without full rebuilds

Performance Benchmark (HNSW Dense Vector Search)

- Achieves 1758+ QPS (queries per second) on a 1 million record dataset using 1536 dimensional vectors.

- Outperforms competitors by:
  - ~42% faster than *Qdrant*
  - ~54% faster than *Weaviate*
  - ~146% faster than *ElasticSearch*.
- Maintains a consistent ~0.97 precision (97%) at high throughput.

Dataset: Source: DbPedia (Qdrant benchmark) || Size: 1 million records || Dimensions: 1536

System	Indexing (min)	Queries Per Second	Precision	Latency p50 (ms)	Latency p95 (ms)
Cosdata	16.32	**1758**	0.97	7	8
Qdrant	24.43	1238	0.99	4	5
Weaviate	13.94	1142	0.97	5	7
ElasticSearch	83.72	716	0.98	22	73

2.1.3 SPLADE Embeddings

- Learned sparse representations combining neural and lexical matching
- Better interpretability than dense embeddings
- Effective for domain-specific retrieval tasks

Performance Highlights:

SPLADE typically offers better retrieval quality (NDCG@10, Recall@10) at the cost of lower throughput and higher latency:

- SPLADE boosts ranking quality by 15-25% across datasets.
- QPS drops by 3-25×, depending on document length and sparsity.
- Latency increases by 1.5-2× in typical server setups.

Use SPLADE for reranking or quality-sensitive applications. BM25 remains optimal for high-speed, low-latency scenarios.

Retrieval Quality Metrics Comparison

Dataset	BM25 NDCG	SPLADE NDCG	BM25 Recall	SPLADE Recall
Arguana	0.40	0.52792	0.64651	0.78734
FiQA	0.25	0.29354	0.31506	0.35627
Quora	0.81	0.80970	0.90219	0.90473
Trec-Covid	0.61	0.64256	0.01623	0.01699
SciFact	0.69	0.62152	0.82017	0.75322
SciDocs	0.16	0.14876	0.16302	0.15283
Webis-Touche	0.34	0.22789	0.20285	0.15223

QPS and Latency Comparison

Dataset	BM25 QPS	SPLADE QPS	BM25 p50/p95	SPLADE p50/p95
Arguana	2167	569.82	9 / 15	27.2 / 49.8
FiQA	4942	1389.62	7 / 12	10.8 / 16.4
Quora	1425	296.18	11 / 36	51.1 / 111.2
Trec-Covid	2219	791.62	10 / 18	32.1 / 49.7
SciFact	40909	1692.06	7 / 13	10.5 / 16.8
SciDocs	13338	1610.62	7 / 12	9.4 / 15.5
Webis-Touche	2789	356.66	10 / 18	83.1 / 123.3

2.1.4 Metadata-Rich Sparse Vectors (Embedding-Free)

Our technology delivers semantic search at scale through sophisticated multi-level hierarchical architecture that preserves parent-child relationships while enabling direct search at any granularity level. Unlike traditional vector databases treating documents as flat entities, our system maintains relational context through intelligent denormalization.

Key differentiators:

- Context-enriched sparse representations without embedding models
- Rapid prototyping and deployment for specialized domains
- Lower computational overhead for resource-constrained environments
- Multi-level document organization with inherited metadata propagation
- Direct searchability at leaf nodes without losing hierarchical context

Spatial Capabilities: Native GPS coordinate integration enables distance-based sorting, multi-zone search, and real-time location-aware ranking. Geographic boundaries are dynamically configurable with custom zone definitions and geofencing support.

Rich Metadata Processing: Arbitrary metadata attachment at parent and child levels supports complex boolean queries, temporal filtering, and range-based searches. Custom attributes are indexed and searchable alongside semantic content with custom scoring functions that blend similarity with business-specific logic.

Transparent Explainability: Results include decomposed scoring showing semantic similarity contributions, metadata match percentages, geographic relevance indicators, and hierarchical relationship context. Users understand exactly why specific results were surfaced and how different factors influenced ranking.

2.2 Data Management & Storage

2.2.1 Versioning & Time Travel in Cosdata

Cosdata is built as an **immutable, append-only database** with native support for **transactional indexing** and **revisioned data access**. This allows developers and researchers to track, query, and revert to historical states of data effortlessly.

Key features:

- **Transactional Corpus Indexing** Index entire document corpora atomically, ensuring consistency across versions and simplifying rollback.
- **Revisioned Query Contexts** Set the query context to any historical revision—ideal for:
  - A/B testing across different data states
  - Comparing model performance over time
  - Debugging pipeline regressions
- **Time-Travel & Audit-Friendly** Run queries on historical versions without impacting current state. Supports full auditability with cryptographically traceable changes.
- **No In-Place Mutation** Files are never overwritten—every write results in a new immutable state. This makes:
  - Backups trivial (snapshot-style)
  - Data corruption far less likely
  - Compliance and data lineage easy to manage

Cosdata's architecture ensures you can develop, deploy, and test retrieval pipelines with confidence—knowing every version of your data is preserved and queryable.

2.2.2 Streaming Ingestion

Cosdata supports high-throughput streaming ingestion alongside transactional batch processing, enabling real-time data scenarios without sacrificing consistency or durability guarantees.

Real-Time Data Processing: Streaming mode handles continuous data feeds including logs, metrics, sensor readings, and real-time content updates. Each record is indexed immediately upon arrival and becomes queryable without explicit transaction coordination, eliminating traditional batch processing delays.

Architecture Design:

- **Immediate Availability:** Records are indexed in memory and logged to write-ahead logs (WAL), becoming queryable instantly
- **Asynchronous Durability:** Background processes periodically flush and compact data into immutable versioned epochs

- **Append-Only Consistency:** All streamed records contribute to monotonic version history while preserving immutable architecture benefits
- **Zero Coordination Overhead:** No manual transaction boundary management required—system handles batching and version assignment automatically

Enterprise Advantages: Streaming ingestion combines the immediacy required for real-time applications with the audit trails and consistency guarantees needed for enterprise deployment. This dual-mode capability allows organizations to handle both real-time operational data and controlled analytical workloads within a single infrastructure platform.

Performance Characteristics:

Operation	Behavior
Record Arrival	In-memory indexing + WAL persistence
Query Availability	Immediate (pre-durability)
Long-term Storage	Periodic compaction to immutable epochs
Version Assignment	Automatic logical version ID batching

2.2.3 Colocated Document Storage

Cosdata eliminates the architectural complexity and performance bottlenecks common in traditional vector database deployments by storing raw document chunks directly alongside vector embeddings and metadata.

The Traditional Architecture Problem: Most vector databases require external storage systems (PostgreSQL, MongoDB, etc.) to house actual document content, storing only chunk IDs with vector representations. This creates a multi-step retrieval process: first querying the vector database for similarity matches, then performing sequential queries to external databases to fetch the actual text chunks for LLM input. This architecture introduces data silos, additional infrastructure dependencies, and significant latency overhead.

Cosdata's Colocated Approach: Raw document chunks, embeddings, metadata, and indices are stored together within Cosdata's unified storage layer. During query execution, top-k retrieval returns complete document chunks in a single operation, ready for immediate LLM consumption without additional database calls.

Performance and Operational Benefits:

- **Reduced Latency:** Single-hop retrieval eliminates sequential database queries and network round-trips
- **Simplified Architecture:** No external document stores, reducing infrastructure complexity and maintenance overhead
- **Data Consistency:** Unified storage ensures atomic updates and eliminates synchronization issues between systems
- **Lower Infrastructure Costs:** Consolidates storage and compute resources, reducing operational expenses

- **Streamlined Development:** Single API for both retrieval and content access accelerates application development

Enterprise Impact: This colocated architecture is particularly valuable for semantic search applications requiring sub-100ms response times and organizations seeking to minimize their data infrastructure footprint while maintaining enterprise-grade reliability and consistency guarantees.

2.2.4 Chunk & Document-Level Retrieval

- Store and retrieve both ‘chunk_{id}‘ and ‘$\text{document}_{\text{id}}$‘ together.
- Efficiently group search results by ‘$\text{document}_{\text{id}}$‘, enabling easy full-document reconstruction.
- Direct lookup of all chunks associated with a given ‘$\text{document}_{\text{id}}$‘, useful for both re-indexing and full-context fetches.

2.3 Security & Access Control

2.3.1 End-to-End Encryption

- All data is fully encrypted at rest (in the underlying file system) and in transit (across the network).
- Optional client-side encryption mode: encryption keys are held by the client only; the server has no access to decryption keys, making this ideal for zero-trust environments.

2.3.2 Fine-Grained Access Control

- Built-in **RBAC (Role-Based Access Control)** system.
- Support for multiple users, roles, and permissions at the collection level.
- Enforce separation of data access for multi-team or multi-tenant use cases with minimal setup.

2.4 Performance & Operations

2.4.1 Intelligent Memory Management

- Dynamically **load or unload collections** into memory based on usage patterns.
- Keep inactive or archival collections on disk to save memory, while still enabling on-demand re-hydration.
- Ideal for large-scale environments where working sets shift over time.

These features ensure Cosdata fits seamlessly into production environments, with the control, security, and flexibility enterprise users expect.

3 Beyond Cosine Similarity: Measuring True Relevance

The fundamental challenge in semantic search lies in a critical misconception: high cosine similarity does not equate to high user relevance. Understanding this distinction separates effective retrieval systems from those that optimize for the wrong metrics.

The Cosine Similarity Fallacy: Cosine similarity measures the angle between embedding vectors—a mathematical distance metric determined by the embedding model's training. However, this metric has no inherent connection to what users actually find relevant or useful. Two documents may be mathematically similar in embedding space while being somewhat irrelevant to a user's information need, or vice versa.

Ground Truth Through User Behavior: True relevance measurement requires ground truth data derived from actual user interactions and expert judgments. Industry-standard datasets like BEIR (Benchmarking Information Retrieval) provide this through carefully curated query-document pairs with human-annotated relevance scores. These datasets enable evaluation using metrics like NDCG (Normalized Discounted Cumulative Gain) and recall, which directly measure how well a system satisfies user information needs.

Qrels and Relevance Judgments: Qrels (query relevance judgments) form the foundation of meaningful retrieval evaluation. These are typically collected through:

- Expert human annotators reviewing query-document pairs
- Implicit feedback from user interactions (clicks, dwell time, conversions)
- Crowdsourced relevance assessments across diverse user populations
- Domain expert evaluations for specialized fields

Hybrid Search Validation: When evaluating hybrid search approaches, we demonstrate improvement by showing enhanced NDCG and recall scores against established BEIR datasets, not merely achieving high precision against brute-force result sets. A system claiming "99% recall" against its own similarity rankings proves nothing about user satisfaction—the real test is whether hybrid dense+sparse retrieval outperforms BM25 baselines on human-judged relevance.

Creating Domain-Specific Ground Truth: For domains lacking established relevance datasets, we employ multi-step approaches to generate meaningful evaluation data:

- **LLM-Assisted Reranking:** Use large language models to rerank top-100 retrieval results to top-10, creating initial relevance signals
- **Domain Expert Integration:** Combine LLM judgments with domain-specific expertise and business logic
- **Multi-Signal Synthesis:** Incorporate contextual factors like document recency, authority, user role, and task-specific requirements
- **Iterative Refinement:** Continuously improve relevance judgments through user feedback loops and A/B testing results

This approach generates domain-specific qrels that better reflect real-world relevance than pure embedding similarity, enabling meaningful evaluation of retrieval system performance against actual user needs rather than mathematical abstractions.

Cosdata's Relevance-First Architecture: Our platform is built on this understanding—optimizing for user relevance through sophisticated ranking algorithms that consider embedding similarity as just one signal among many, weighted according to empirical relevance data rather than mathematical convenience.

4 Enterprise Integration and Scalability

4.1 SDK and Integration

- Native SDKs for Python & Node (more on the way!)
- RESTful APIs with comprehensive documentation
- Drop-in compatibility with popular ML frameworks

4.2 Scalability Architecture

- Horizontally scalable with automatic sharding
- Multi-region deployment support
- Real-time index updates without downtime

5 Enterprise Integration and Scalability

5.1 SDK and Integration

- Native SDKs for Python & Node.js (more on the way!) - Python SDK Documentation
- RESTful APIs with comprehensive documentation - docs.cosdata.io
- Drop-in compatibility with popular ML frameworks
- Open source under Apache 2.0 license with community support

5.2 Scalability Architecture

- Horizontally scalable with automatic sharding
- Multi-region deployment support
- Real-time index updates without downtime
- Cloud-native architecture with containerized deployment

5.3 Licensing and Support Options

5.3.1 Open Source Edition

- Full platform capabilities under **Apache 2.0** license
- Community support through GitHub and documentation
- Self-hosted deployment with complete source access
- Ideal for development, research, and small-scale production use

5.3.2 Enterprise Edition

- All open source features plus enterprise enhancements
- Professional support with SLA guarantees
- Custom integrations and deployment assistance
- Priority feature requests and roadmap influence
- Advanced security features and compliance certifications
- Dedicated technical account management
- Training and onboarding programs for technical teams