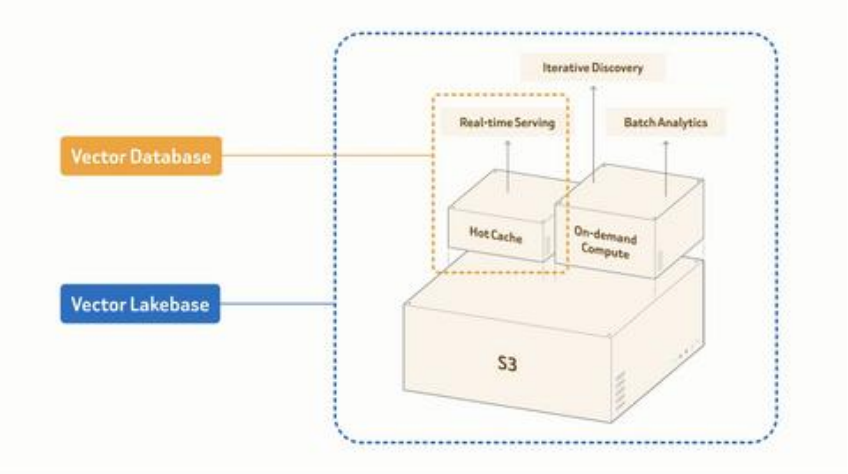
Business Wire (2026-07-31)

Zilliz Announces Milvus 3.0, Making the World's Most Adopted Open-Source Vector Database Lake-Native

"Vector Lakebase is what we believe comes next: one data foundation where the same vectors can serve a production query, anchor a discovery session, and power a multi-petabyte training-data pipeline — without copies, migration, or a parallel stack."



REDWOOD CITY, Calif.--(BUSINESS WIRE)--Zilliz, the company behind Milvus, the world's most widely adopted open-source vector database, recently announced the release of **Milvus 3.0**, a major architectural update that brings lake-native data access and a more expressive retrieval engine to developers building production AI applications. Milvus 3.0 enables organizations to build and serve production-grade indexes over vector data that remains in object storage and open data formats. It also moves more query and retrieval operations into the engine, expanding Milvus beyond traditional nearest-neighbor search to support richer ranking, aggregation, sparse search, and multi-vector retrieval.

This release also advances the shared technical foundation behind both open-source Milvus and Zilliz Cloud. Developers can deploy Milvus independently, while Zilliz Cloud builds on the Milvus core to deliver a fully managed Vector Lakebase spanning real-time serving, interactive discovery, batch analytics, and AI data operations.



“AI data is becoming larger, richer, and more dynamic, yet the systems used to serve and improve that data have remained fragmented,” said James Luan, Co-founder and CTO of Zilliz. “Milvus 3.0 brings production retrieval closer to where data already lives and gives developers a more expressive engine for modern AI applications. That same foundation powers Zilliz Vector Lakebase, where we extend it with the elasticity, operational simplicity, and enterprise capabilities required to manage AI data at scale.”

Index and Serve Data Where It Already Lives

Production AI systems commonly maintain one copy of data for real-time retrieval and another for offline processing. Moving data between the two introduces additional storage overhead, lengthy exports, synchronization pipelines, and operational complexity.

Milvus 3.0 introduces a lake-native architecture that allows production indexing and retrieval to operate directly over data in object storage.

- **External Collections** define Milvus collections over data stored in Lance, Iceberg, Parquet, or Vortex. Milvus builds vector, full-text, JSON, and scalar indexes over the data and exposes them through the same APIs as native collections—without copying the source data into Milvus. Incremental synchronization keeps indexes current as the underlying dataset changes.
- **Loon**, the new manifest-based storage engine in Milvus 3.0, is designed to substantially reduce the read amplification associated with low-latency point access on object storage. It uses Vortex, an open and Arrow-compatible columnar format, as its default storage format.
- **Snapshots** create point-in-time, read-only views of live collections at minimal incremental storage cost. Offline jobs such as evaluation, deduplication, and backfill validation can run against a stable dataset while production writes continue.
- **A new Spark connector** exposes Milvus as a Spark DataSource V2, allowing Spark, Databricks, and EMR pipelines to read from and write to Milvus as part of standard batch workflows. Milvus 3.0 also

introduces vector-native batch operators for tasks including deduplication and clustering.

Together, these capabilities bring online retrieval and offline data improvement closer to a shared data foundation, reducing the need to continually export, duplicate, and reconcile large datasets across separate systems.

A More Expressive Retrieval Engine

Milvus 3.0 also expands retrieval beyond traditional top-K vector search. New server-side sorting and aggregation, together with faceted search, allow applications to combine semantic relevance with attributes such as freshness, rating, price, category, tenant, and content source. Operations that previously required over-fetching data and processing it in application code can now execute directly within the Milvus engine.

For documents, images, and other entities represented by more than one vector, **StructList** introduces native multi-vector retrieval. A document can retain vectors for multiple chunks, an image can retain vectors for multiple visual patches, and a product can retain several multimodal representations under a single entity and metadata record. StructList also supports late-interaction retrieval models such as ColBERT and ColPali.

Milvus 3.0 further strengthens sparse and hybrid retrieval with an optimized sparse index that is approximately 3 times smaller while achieving comparable recall in internal testing. Additional enhancements include SINDI for learned-sparse vectors, server-side MinHash generation, nullable vector fields, custom dictionaries for full-text search, and broader support for Faiss-compatible indexes.

“Production applications rarely stop at returning the nearest vectors. They need to sort, group, build facets, and retrieve across multiple representations of the same entity,” said **James Luan, CTO at Zilliz**. “Milvus 3.0 moves more of that work into the engine while redesigning storage for efficient, low-latency access on object storage.”

These capabilities provide a broader retrieval foundation for applications spanning RAG, semantic search, recommendation, multimodal discovery, agent memory, and AI data preparation.

Open-Source and Managed Availability

Milvus 3.0 is available under the Apache 2.0 license and remains a graduated LF AI & Data project. It can be deployed on Kubernetes or Docker, including in air-gapped environments, and supports Amazon S3-compatible object storage, Google Cloud Storage, and Azure Blob Storage. The release is available initially through the Python, Go, and Node.js SDKs, with Java support planned shortly after release. Developers can read the release notes, follow the quickstart, and access the source code on GitHub. They can also join the Milvus community or speak with project maintainers through Milvus Office Hours.

For organizations that prefer a managed platform, Zilliz Cloud incorporates Milvus 3.0 as its core indexing and retrieval engine and extends it into a fully managed Vector Lakebase. Zilliz Cloud adds elastic compute, interactive discovery, batch analytics, shared indexing, enterprise security and governance, and automated operations across a single copy of AI data.

This gives developers a consistent technology foundation: they can deploy and operate Milvus as an open-source vector database, or use Zilliz Cloud when they need a managed platform that spans real-time retrieval, offline discovery, data processing, and production operations. For more details, see the Milvus 3.0 launch blog.

About Zilliz

Zilliz is a leading AI data infrastructure company and the creator of Milvus, the world's most widely adopted open-source vector database, with more than 45,000 GitHub stars and over 100 million Docker pulls. Zilliz helps enterprises and AI-native companies make unstructured data searchable, analyzable, and governable—turning text, images, audio, video, and other multimodal data into a strategic asset for production AI.

Zilliz's technology centers on Milvus and Zilliz Cloud. Milvus is an open-source vector database purpose-built for vector search at massive scale. Powered by Milvus, Zilliz Cloud extends that foundation into a fully managed Vector Lakebase that combines the high-throughput, low-latency serving capabilities of a vector database with the openness, scalability, and economics of a multimodal data lake.

Zilliz supports more than 10,000 enterprises and AI-native companies worldwide, including MiniMax, OpenEvidence, Filevine, Exa, Salesforce, and Read AI.

Headquartered in Redwood Shores, California, Zilliz is backed by leading investors, including Aramco's Prosperity 7 Ventures, Temasek's Pavilion Capital, Hillhouse Capital, 5Y Capital, Yunqi Partners, and Trustbridge Partners. Learn more at Zilliz.com.

Contacts

Media Contact
Molly Chen, Head of Marketing
molly.chen@zilliz.com