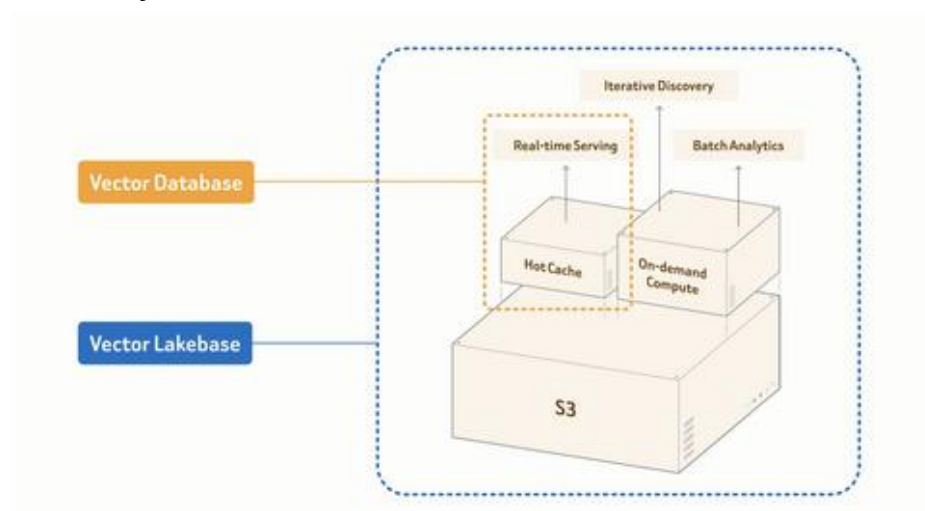




Business Wire (2026-06-22)

Zilliz Launches Vector Lakebase, Extending the World's Most Adopted Vector Database into a Unified Data Platform for AI

Available now in public preview on Zilliz Cloud, Vector Lakebase keeps production vector search at its core and adds shared lake-native storage and on-demand compute — bringing real-time serving, interactive discovery, and batch analytics onto one data foundation.



REDWOOD CITY, Calif.--(BUSINESS WIRE)--Zilliz, the company behind Milvus, the world's most widely adopted open-source vector database, today announced the public preview of Zilliz Vector Lakebase, a major Zilliz Cloud release that pairs the production vector database with a shared, lake-native data foundation.

Vector Lakebase keeps Zilliz Cloud's real-time vector search at the core — the engine Zillow, OpenEvidence, Exa, Filevine, MiniMax, and more than 10,000 enterprises and AI teams already rely on — and extends it with three new ways to operate on the same data: interactive discovery, large-scale batch analytics, and search directly on external data lakes. The result is a single data foundation in which every workload runs against a single logical copy of the data, with on-demand and batch jobs billed only when compute is active.



"Production vector search is and will remain at the heart of what Zilliz does — it's why thousands of teams choose Milvus and Zilliz Cloud, and it's getting faster and more cost-efficient every release," said Charles Xie, Founder and CEO of Zilliz. "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."

Why a Single Data Foundation Matters

AI systems are no longer a single-query retrieval problem. They run as a continuous loop — serve, learn from feedback, mine and prepare better data, then serve again — and each turn typically requires separate systems for serving, exploration, and large-scale processing. Moving billions of vectors between those systems can take days. The cost and complexity are so high that many teams skip the loop altogether, leaving valuable data retrievable but never improved.

Vector Lakebase closes that gap with a zero-copy semantic data plane on shared lake-native storage: real-time serving, interactive discovery, and batch analytics all run against one logical copy of the data, scaling from gigabytes to petabytes.

"Teams asked for a way to keep their data in one place and run very different workloads against it — from real-time agent memory to overnight semantic deduplication," said Robert Guo, VP of Product at Zilliz and one of the architects behind Milvus. "Vector Lakebase delivers that through a unified storage layer on Vortex, tiered serving for the production path, and on-demand compute for everything else."

Five Capabilities on One Foundation

- **Tiered Real-Time Serving.** Three production tiers tuned for different workloads: Performance-Optimized (1,000+ QPS, single-digit-millisecond latency, in-memory); Capacity-Optimized (100–500 QPS,



sub-100ms latency, memory + NVMe); and Tiered-Storage (10–50 QPS, ~100ms latency, spanning memory, NVMe, and object storage at significantly lower cost). All tiers default to 95–98% recall, tunable to 99%+, backed by Zilliz Cloud's 99.99% uptime SLA and Global Cluster cross-region high availability.

- On-Demand Search. Pay-as-you-go compute for workloads where infrastructure sits idle most of the time, billed directly for object storage and compute rather than serverless markups. In Zilliz's internal benchmark on one billion 768-dimension vectors with 10 hours of monthly active compute, On-Demand Search totaled \$318 versus \$4,937 for a comparable serverless path — roughly 1/15 the cost.
- External Data Lake Search. A zero-copy External Collection mode that adds state-of-the-art indexing and full-spectrum search directly to existing Lance, Iceberg, Parquet, and Vortex tables, with incremental sync on refresh. Source data stays where it lives.
- Full-Spectrum AI Search. Search across vectors (dense and sparse), text, JSON, and geospatial data, with hybrid retrieval, BM25, regex, multi-vector and iterative search, and multi-path retrieval. Results can be reranked with Cohere, Voyage AI, RRF, and weighted/boost/decay strategies.
- Unified Lake-Native Storage. Shared storage for serving and analytics built on Vortex, an open columnar format designed for faster, cheaper random reads than Lance and Parquet, paired with object-storage-aware indexes (vector, BM25, JSON) that cut read amplification by over 90%. A 100-million-row schema backfill typically completes in single-digit minutes — without disrupting active query traffic.

Together, these capabilities let AI teams consolidate what previously required parallel always-on serving clusters and separate batch systems onto one platform — with consistent indexes, versioned data, and compute that scales to zero between jobs.



Availability

Zilliz Vector Lakebase is available now in public preview on Zilliz Cloud, alongside Serverless, Dedicated, and BYOC deployment options across more than 30 regions on AWS, Google Cloud, and Microsoft Azure. New work email signups receive \$100 in free credits at zilliz.com. Teams running serving, discovery, and analytics on separate stacks can contact the Zilliz team for a tailored walkthrough.

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 44,000+ GitHub stars and over 100 million Docker pulls. Zilliz helps enterprises and AI startups make their unstructured data searchable, analyzable, and governable — turning text, images, audio, video, and more 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 100-billion-scale vector search. Zilliz Cloud extends that foundation into a fully managed Vector Lakebase platform, combining the high-throughput, low-latency serving capabilities of vector databases with the openness, scalability, and economics of multimodal data lakes. Zilliz powers more than 10,000 enterprises and AI-native startups 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

molly.chen@zilliz.com

+1 951-265-1426