zilliz

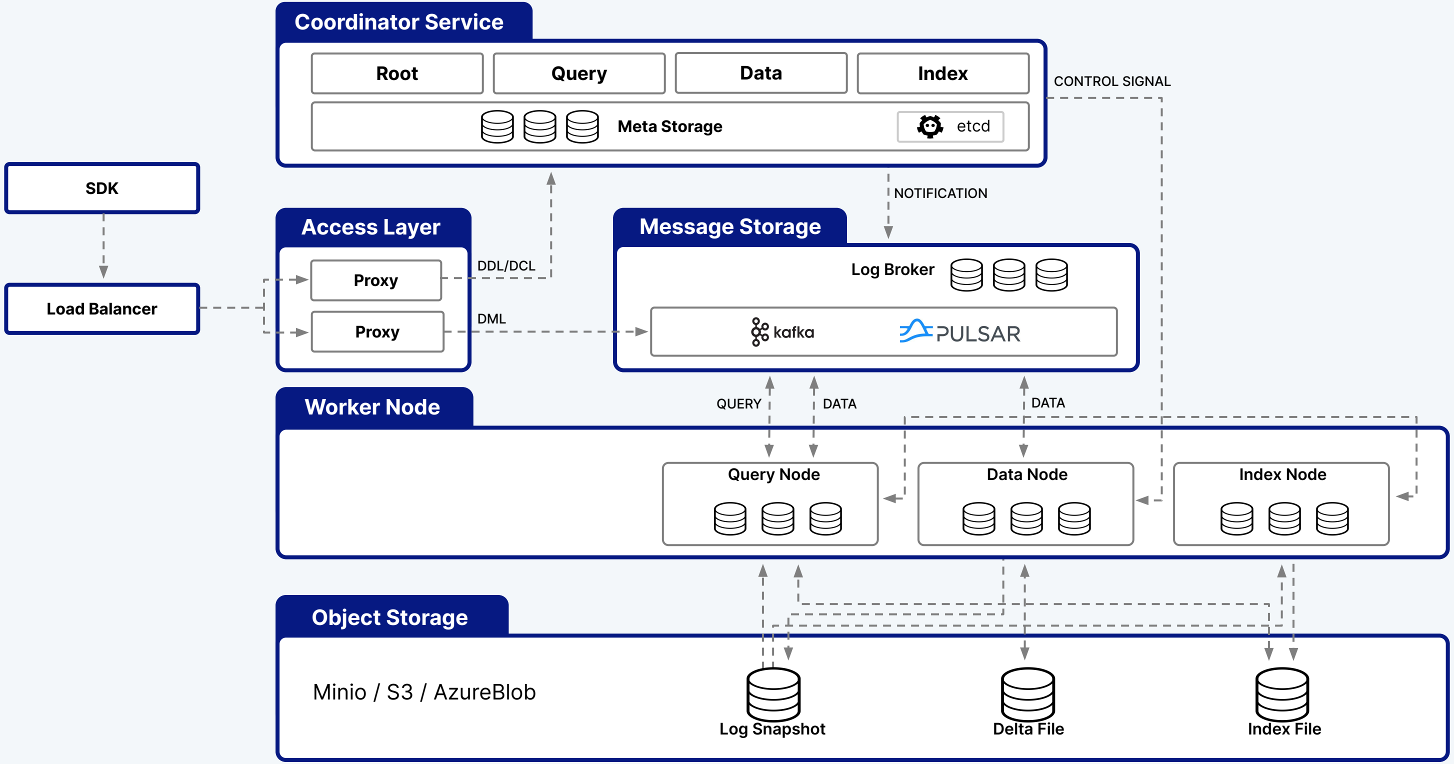
Fall Live Demo:

Discover the Latest Features in Zilliz Cloud

Yujian Tang | Zilliz

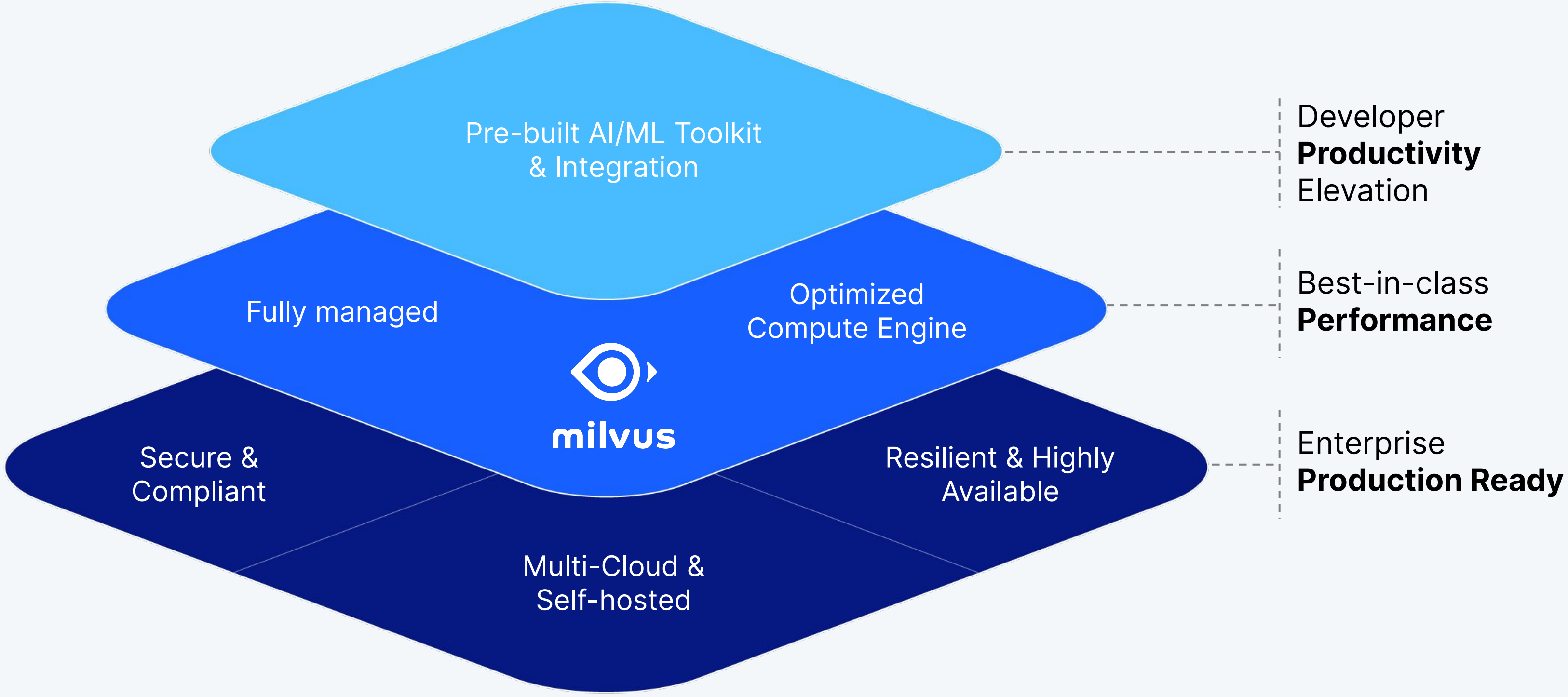
Zilliz Overview

Open Source Milvus' Architecture



zilliz

Zilliz Cloud is the fully managed Milvus service


Pre-built AI/ML Toolkit & Integration
Fully managed
Optimized Compute Engine
milvus
Secure & Compliant
Multi-Cloud & Self-hosted
Resilient & Highly Available
Developer **Productivity** Elevation
Best-in-class **Performance**
Enterprise **Production Ready**

…and so much more than Milvus

AI Native

- AUTOINDEX for superior performance
- Multi-language SDK
- Data migration support
- Integration with leading AI/ML providers

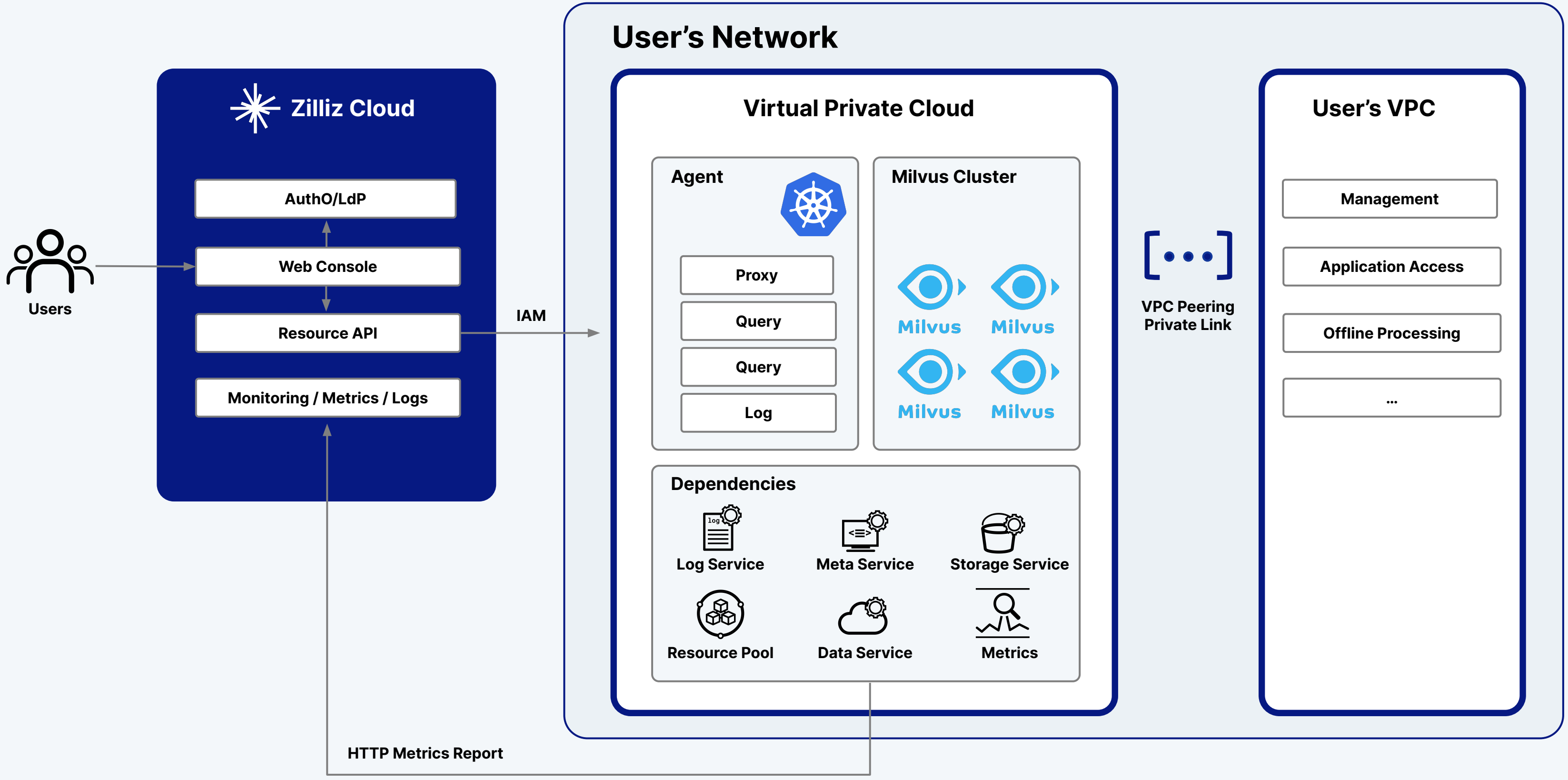
Fully managed Milvus

- Optimized execution engine with 5x performance uplift
- On-demand elastic scaling
- Use case optimized compute unit
- Monitoring & alert
- Multi-cloud support

Enterprise Production Ready

- 99.9% Uptime SLA
- Built-in failover
- RBAC & Access Log
- Data encryption & private networking
- SOC2 Type II & ISO 27001 compliant
- 24/7 Service support

Zilliz Cloud Architecture



Users

Zilliz Cloud

AuthO/LdP

Web Console

Resource API

Monitoring / Metrics / Logs

IAM

HTTP Metrics Report

User’s Network

Virtual Private Cloud

Agent

Proxy

Query

Query

Log

Milvus Cluster

Milvus

Milvus

Milvus

Milvus

Dependencies

Log Service

Meta Service

Storage Service

Resource Pool

Data Service

Metrics

VPC Peering
Private Link

User’s VPC

Management

Application Access

Offline Processing

...

Provision your vector databases within minutes

	STARTER *Serverless* Get started for free to experimentation	**STANDARD** *Dedicated* For production-ready use cases leveraging an extended feature set	**ENTERPRISE** *Dedicated* For mission-critical use cases with advanced security needs
CU options	NA	3 options to balance the performance, storage and cost needs	3 options to balance the performance, storage and cost needs
Sizing	2 collections, each up to 500K 768-dim vectors	Unlimited	Unlimited
Uptime SLAs	NA	NA	99.9%
Availability Zone	Single AZ	Single AZ	Multi AZ
Security & Compliance	NA	Data Encryption, RBAC, SOC 2 Type 2 Compliant	Data Encryption, RBAC, SOC 2 Type 2 Compliant, Private Link, Data backups
Support	Community Support	Email support during business hours with response time SLAs; Up to 2 technical contacts	Email support during business hours with higher response time SLAs; Up to 4 technical contacts

Mix and match cluster type across your organization

3 Instance type to meet the needs of your business

Compute Unit	Optimized for Cost (C)	Optimized for Storage (S)	Optimized for performance (P)
Data volume (per CU)	Up to 3.75 Mil 1024-dim vectors	Up to 3.75 Mil 1024-dim vectors	Up to 0.75 Mil 1024-dim vectors
Performance (QPS)*	10-30	~3X+ of C CU	~1.5X+ of S CU
Performance (Latency)*	<100 ms	2X faster of C CU	~20% faster of S CU
Max Collections	32 for 1-8 CUs, 256 for 8+ CUs	32 for 1-8 CUs, 256 for 8+ CUs	32 for 1-8 CUs, 256 for 8+ CUs
Ideal for	Use cases ***not require ultra-fast response time***	Use cases require processing a ***very large volume of data***	Critical use cases require extremely ***high throughput*** and ***low latency***
Example use cases	Data labelling, data deduplication, data clustering, dataset outlier detection	Large-scale unstructured data search, copyright detection, identity verification	GenAI, Recommendation System, Search engine, Chatbot, Fraud Detection

*Note: we use 1-5M 768dim dataset for the testing example here. In reality, the performance differences varies depending on the datasets sizes, dimension, dense, filtering and other deciding factors.

zilliz

Use cases across all industries

Retail & Ecommerce Drive revenue by product recommendations and better search experience	Product Recommendation	Product Image Search	Inventory Management	Fraud Detection	Customer Segmentation
Healthcare Accelerate innovations, provide accurate and affordable healthcare	Drug Discovery	Medical Image pattern recognition	Genomic Research	Patient Record Analysis	Medical Chat Bot
Banking & Finance Combat fraud & remain competitive	Fraud Detection	Risk Management	Virtual Assistant	Customer Profiling	Compliance Assistance
Automotive & Transportation Amplify vehicle intelligence & safety	Fleet Management	Route Optimization	Predictive Maintenance	Autonomous Vehicles	Traffic Prediction
Common In All Industries ML/AI Use Cases	Customer Support	Document Management	Recommendation System	Anomaly Detection	Customer Segmentation & Personalization

Driving AI Innovation with leading business

airbnb AT&T BOSCH COMPASS

ebay IKEA Inflection intuit NVIDIA

ROBLOX shutterstock Walmart ZipRecruiter