

Bringing Intelligence to Open Infrastructure: Vector Search and AI Workloads on OceanBase + Kubernetes

Peng Wang

OceanBase Global Technical Evangelist

2025/08/10



OceanBase: Powering Global Business with Distributed Database



- 15 Years of Innovation
Founded in 2010 as a distributed database pioneer
- Powered by Excellence
1,000+ employees driving innovation worldwide
- 2,000+ Customers Worldwide
Trusted by global industry leaders



FinTech



E-commerce



Digital
Native



Manufacturing



Transportation
& Logistics



Telecom



Gaming

Starbucks, McDonald's, Popmart, P&G,
Unilever, DataVisor, AliPay, Taobao, GoTo,
HSBC, Trip.com, MI, NetEase Games,
Haidilao Hot Pot, GCash, Dana, TNG
Digital...

01

RAG Application

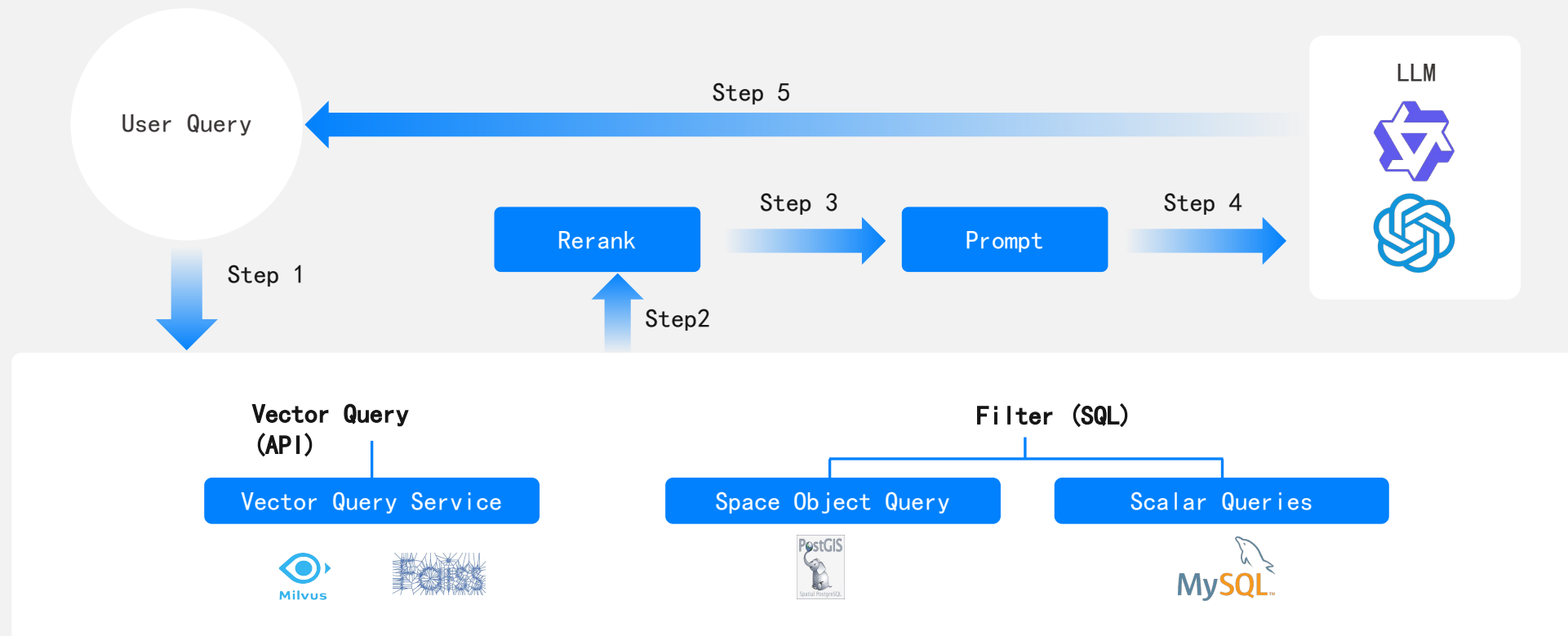
Accelerate AI Application Development

Imagine the following scenario:

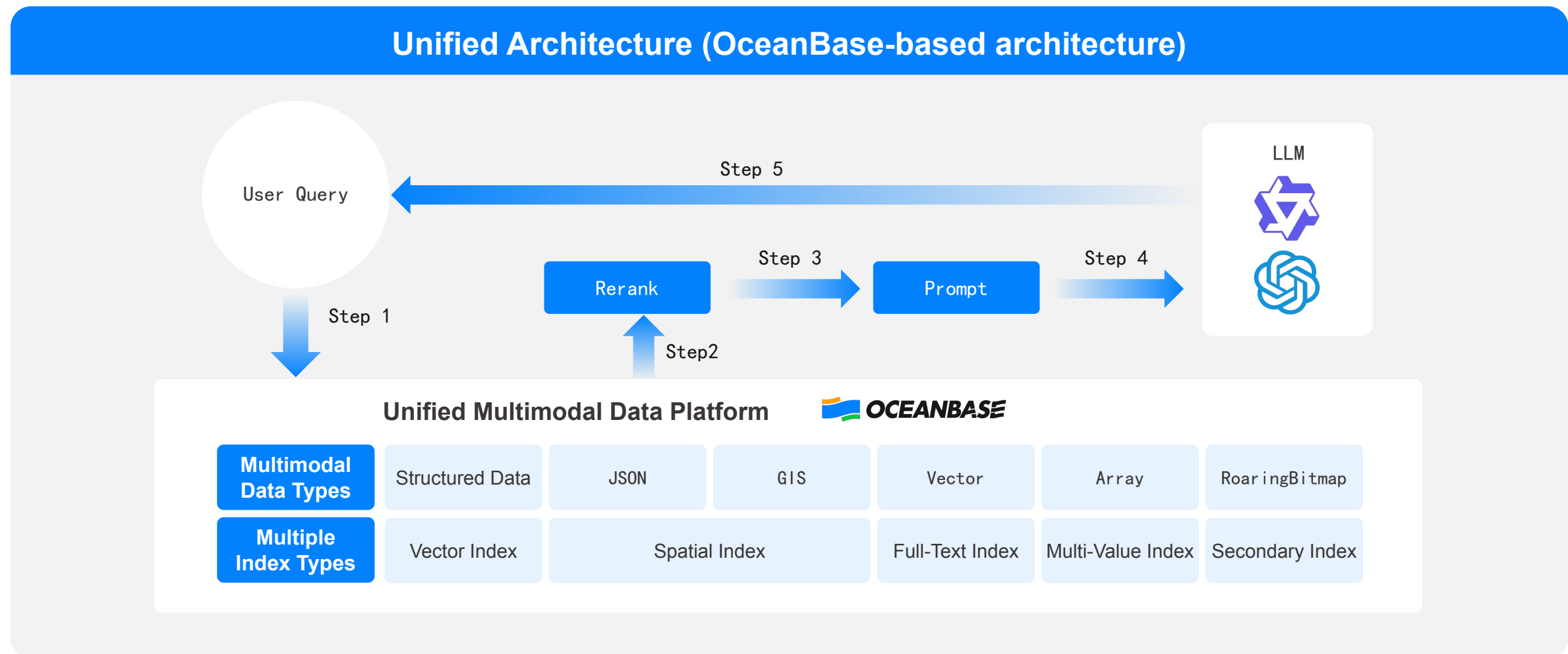


Accelerate AI Program Development

Conventional Architecture



Accelerate AI Program Development



Accelerate AI Program Development

Hybrid Search

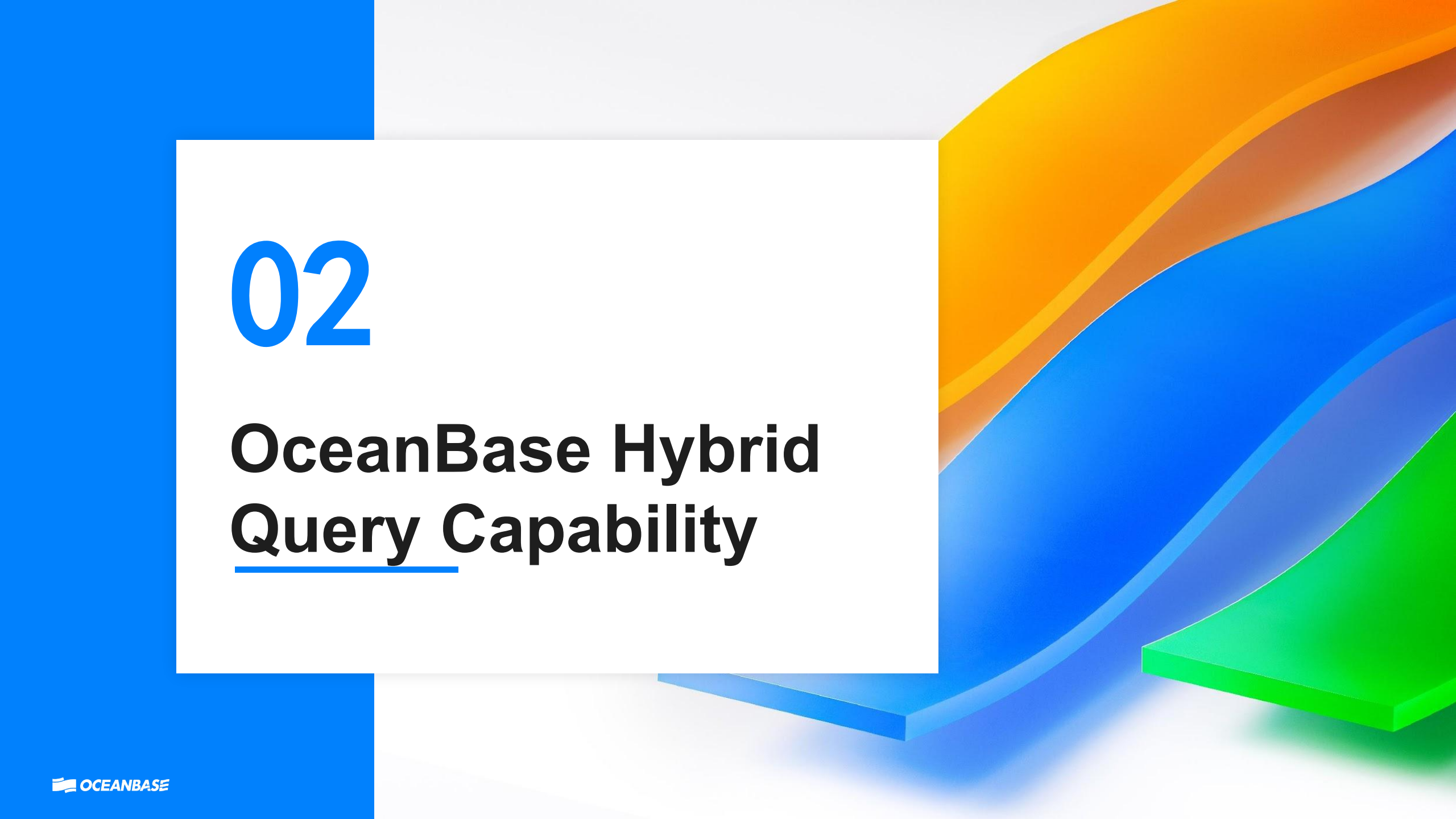
Please recommend a **coffee shop** **within 500 meters** that has an average consumption of **less than \$8 per person**

Spatial Query

Scalar Query

Vector Query

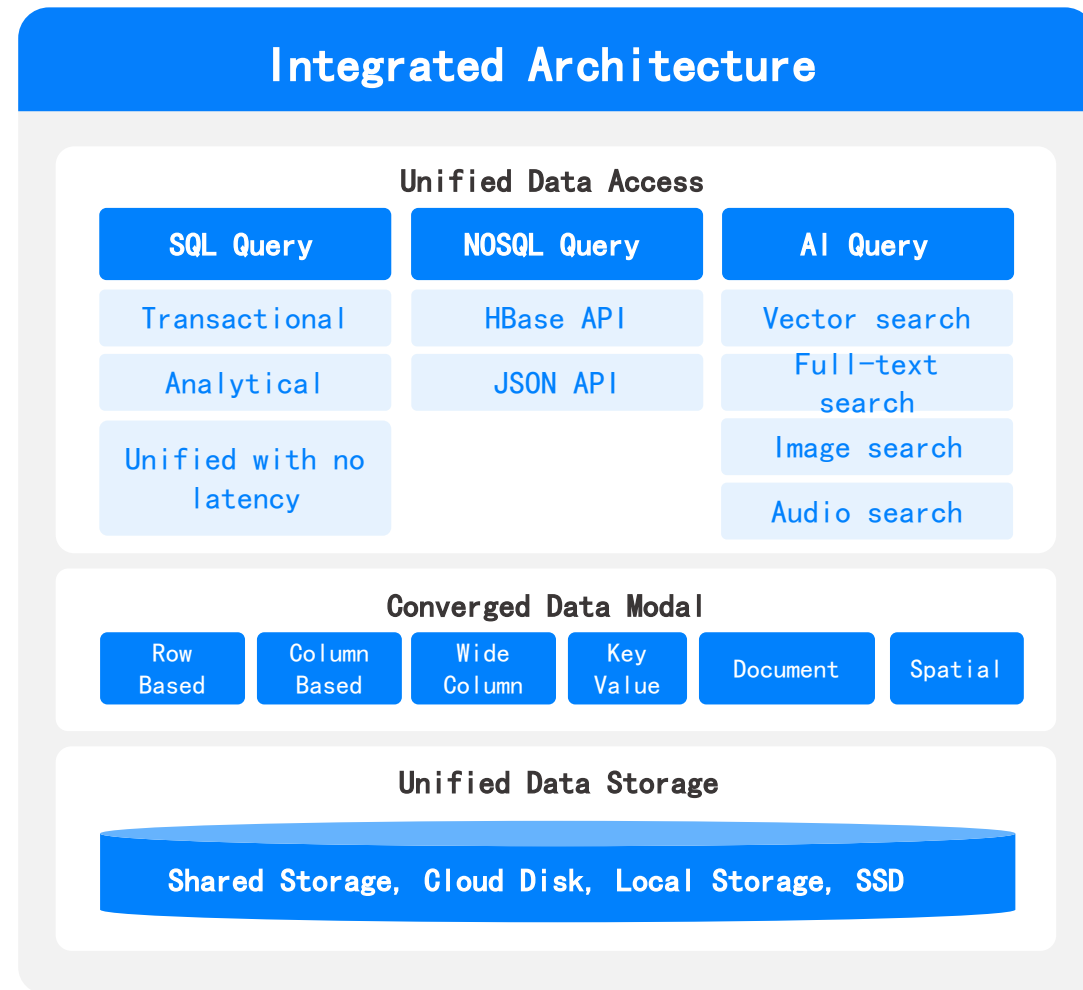
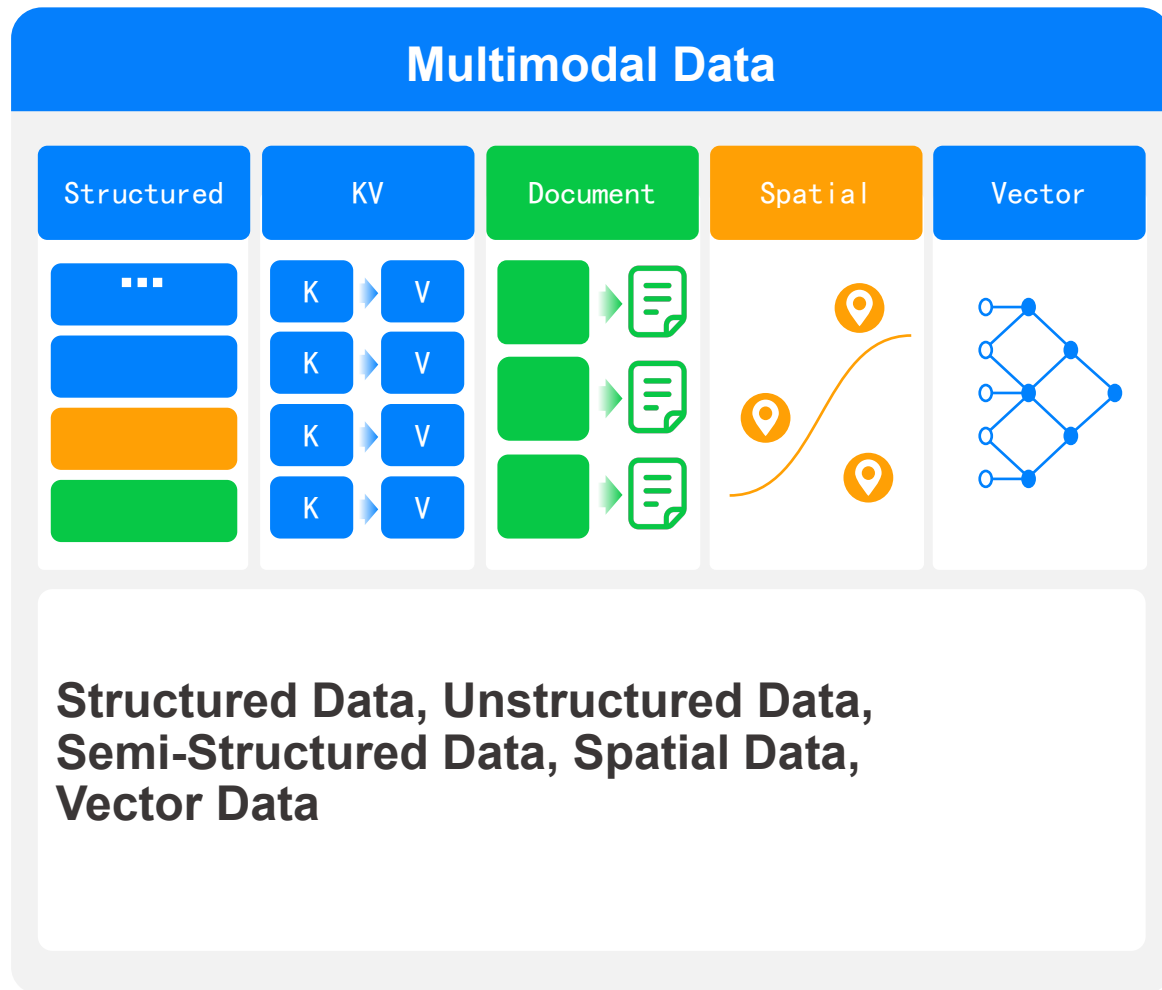
```
st_distance(location, st_srid(point(@longitude, @latitude), 'metre' ) < @query distance  
and avgConsum < 8  
oeder by 12_distance(featureVec, @query embedding) approximate limit 20; _____
```



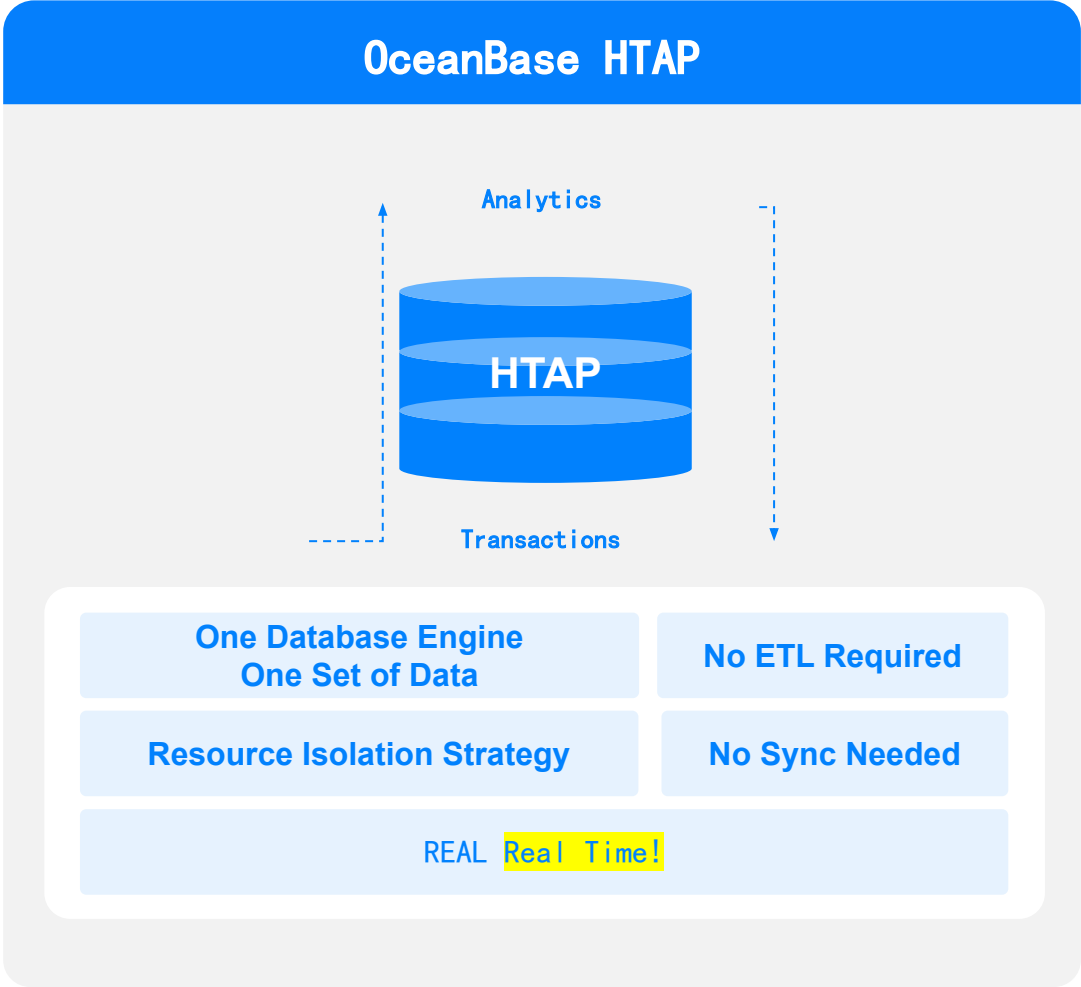
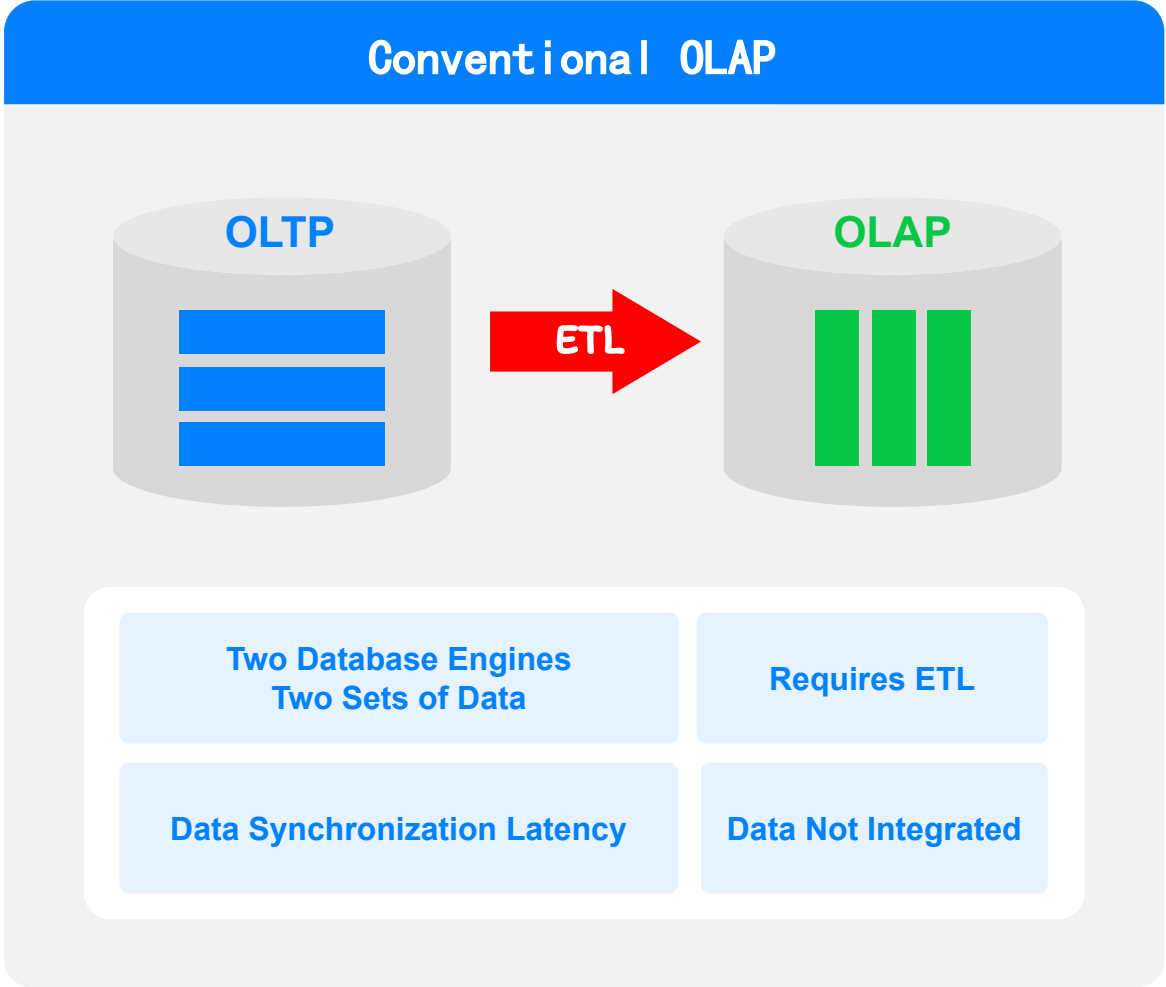
02

OceanBase Hybrid Query Capability

OceanBase Multimodal Data Platform



OceanBase: Cutting-Edge HTAP Database



OceanBase AP Performance

CPU: 16C
Mem: 32G
Disk: 500 GB, 1500 iops

ClickBench Test Results

Response time (sec)

■ OceanBase ■ C**House ■ S**Rocks ■ Apache Doris

The lower the better



AP: Analytical Processing

OceanBase Low-Cost Vector Query

Same Recall & Performance, **95%** Less Memory Compared to HNSW

HNSW

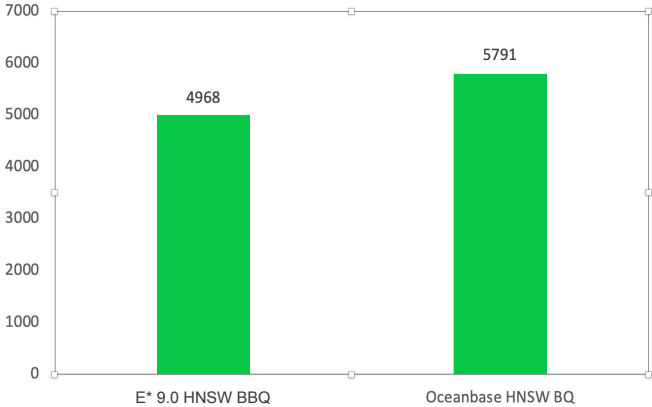
1.2 TB Mem

HNSW + BQ

58.6 GB Mem

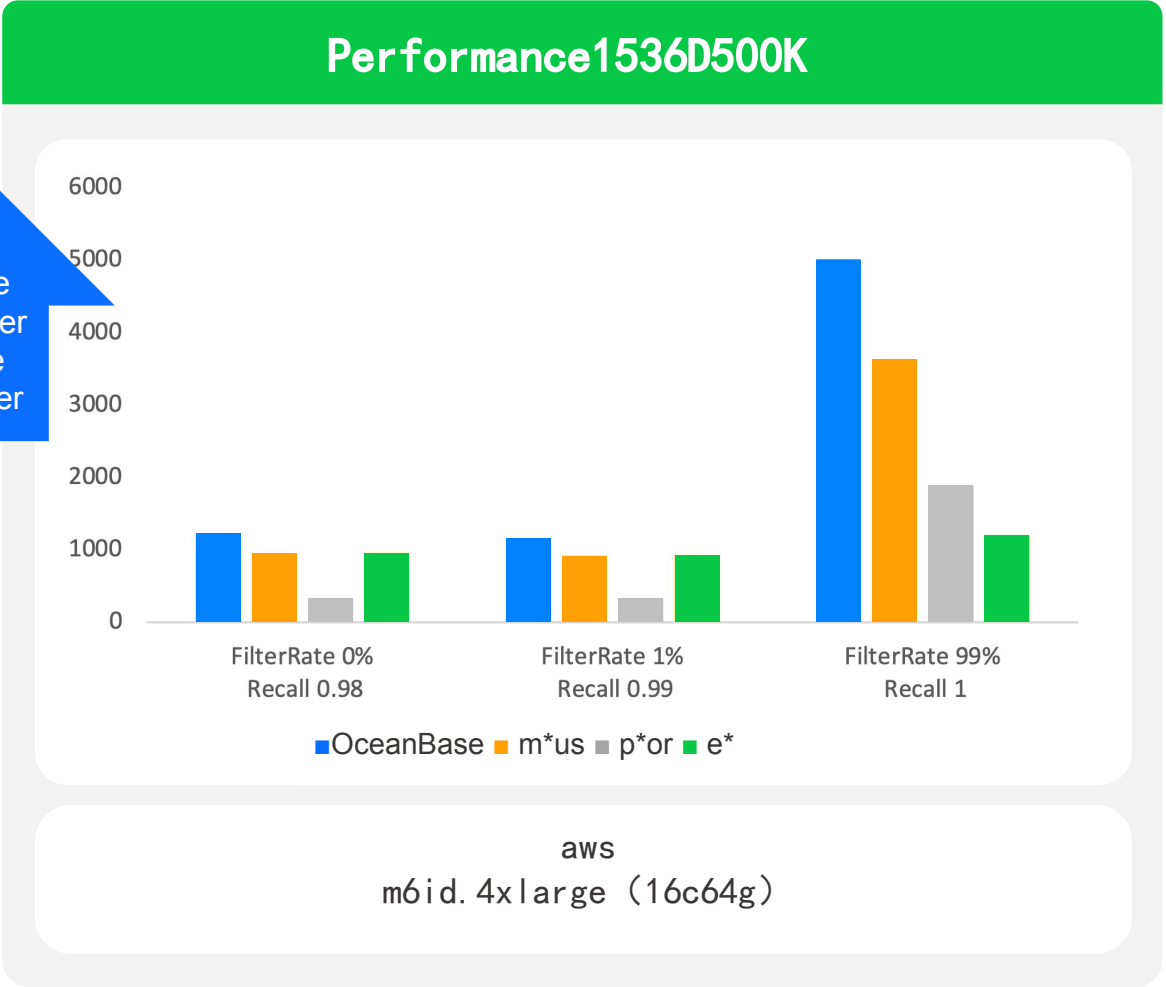
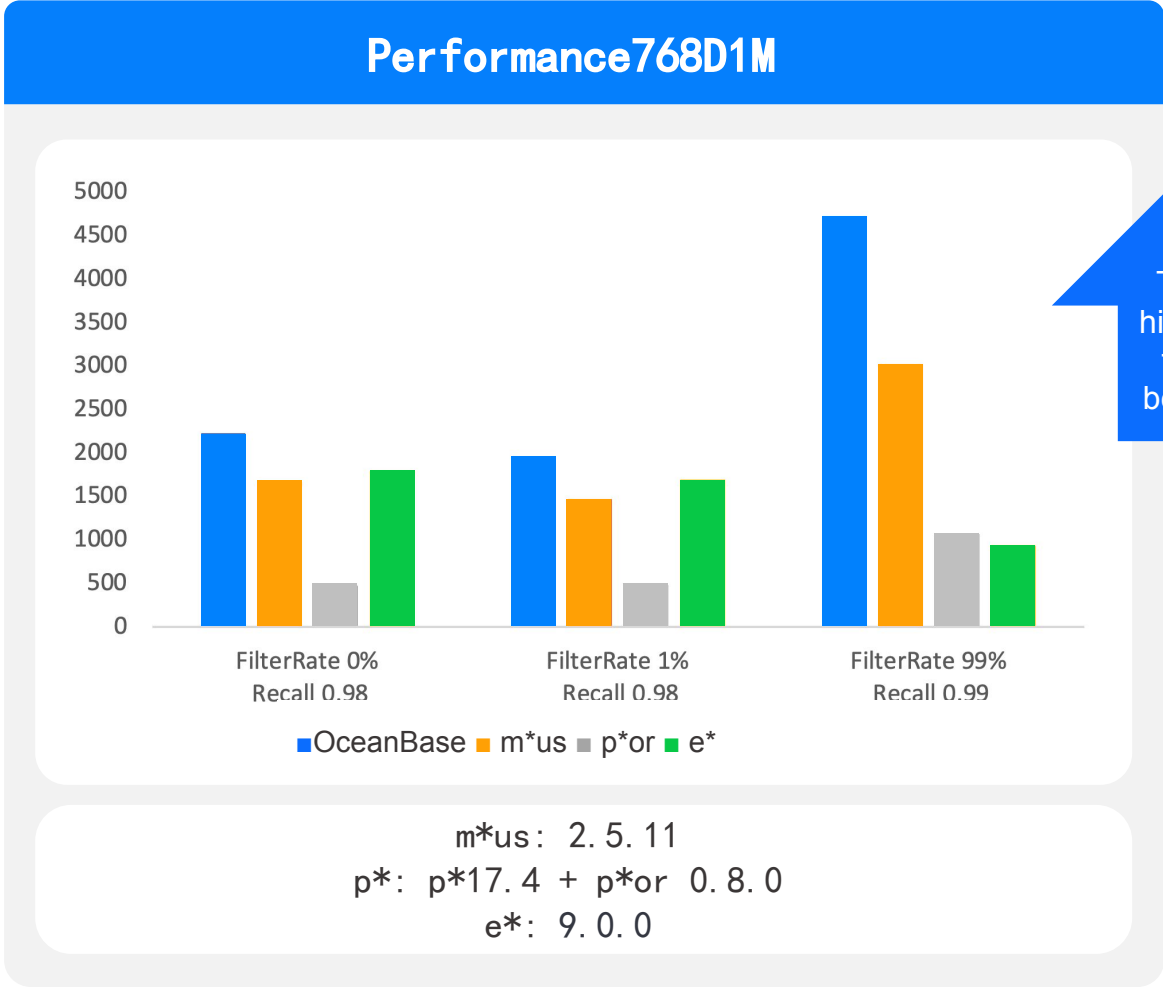
Vector Data Volume: 200 Million
Vector Dimension: OpenAI 1536-D

Same Recall & Cost, **16%** Higher Performance than ES 9.0 BBQ



Vector Data: OpenAI 1536-D, 50K
Environment: 8 cores, 64 GB RAM
Recall: 0.95

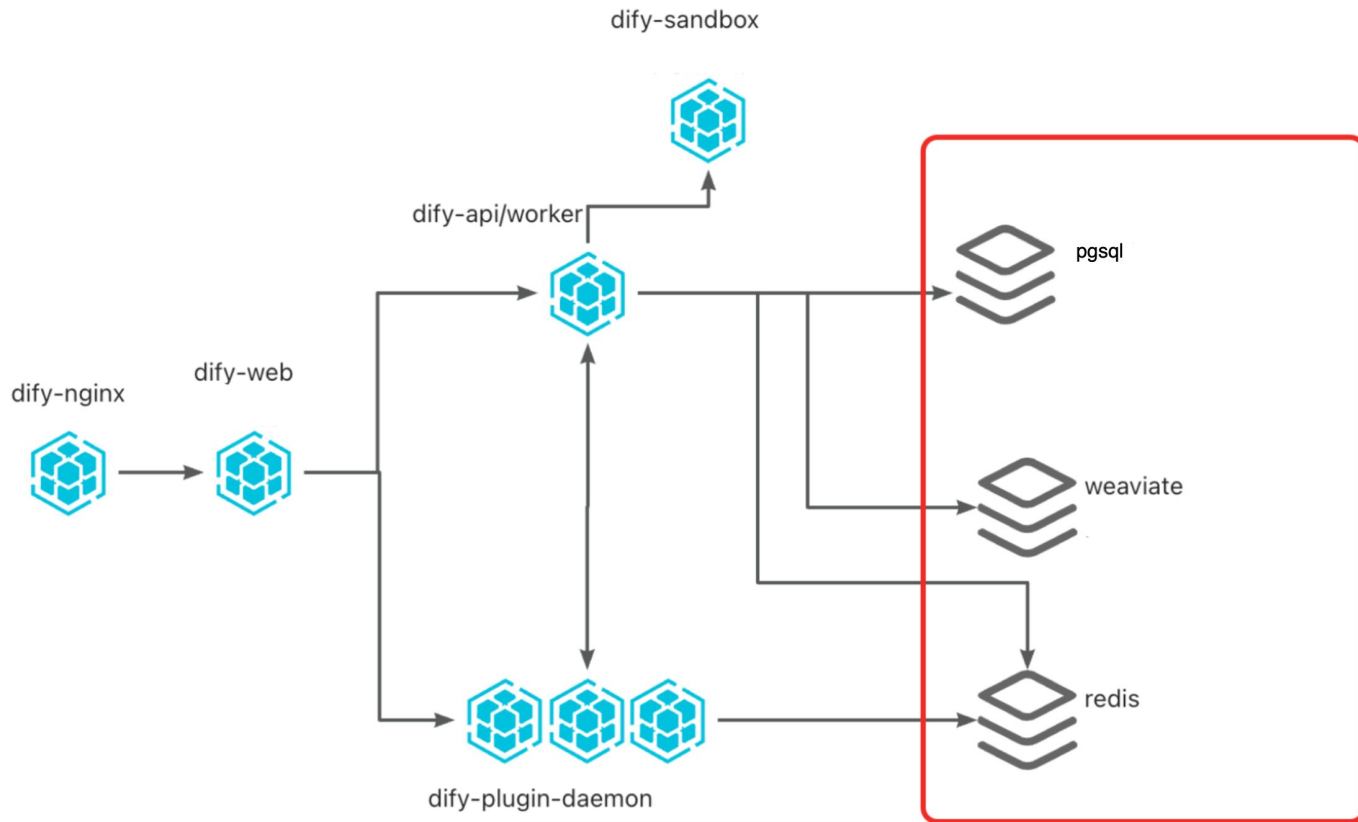
OceanBase VectorDBBench



03

Diffy x OceanBase

Dify Conventional Architecture



Dify Summary

Dify is an open-source platform for building AI applications without coding.

Dify includes three databases:

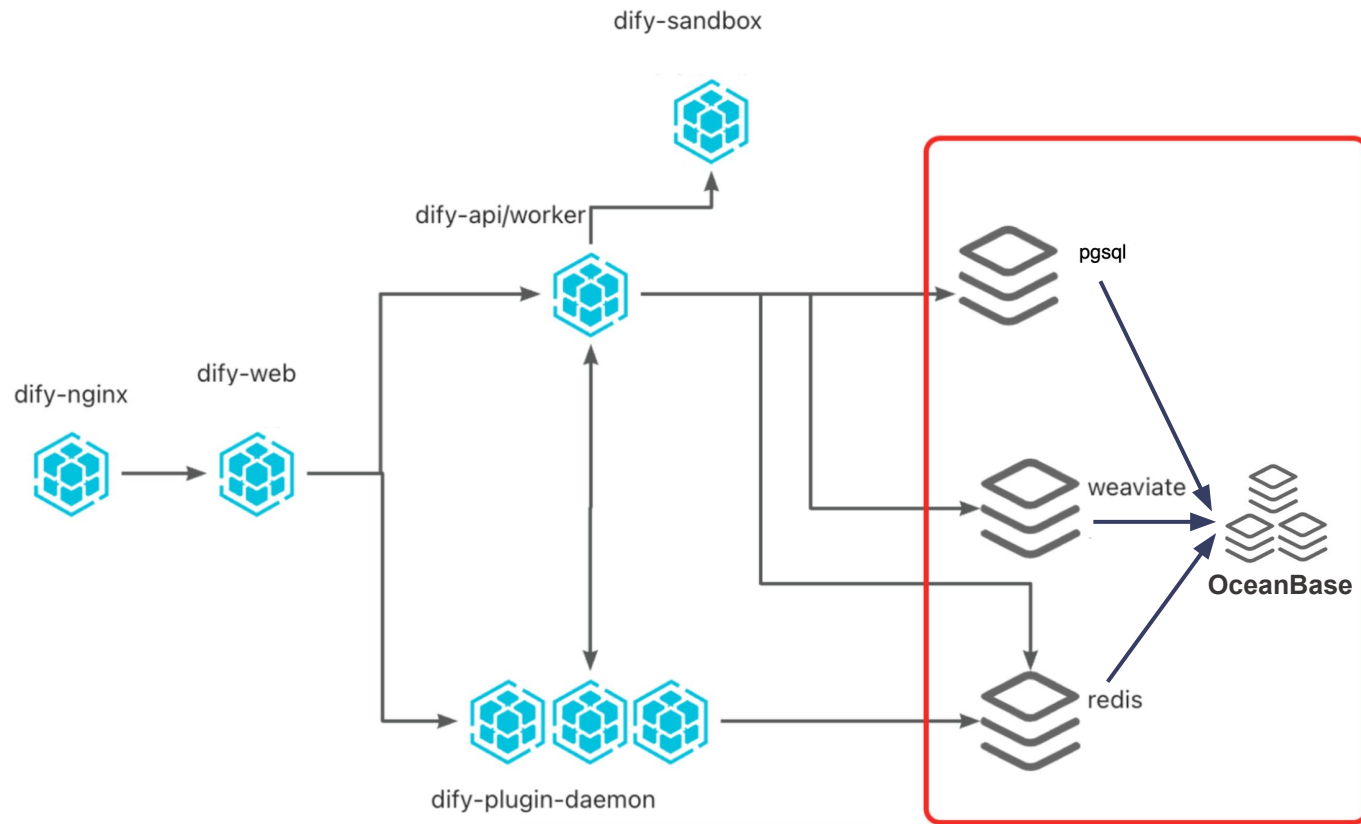
1. **pgsql** : Metadata Storage
2. **weaviate**: Vector Data Storage
3. **redis** : lock•cache•queue

Problems

For DBAs, maintaining three separate databases is very challenging!

- Weaviate frequently experiences outages.
- Within the same Dify instance, database resources cannot be isolated between different business units.
- Deploying a separate Dify instance for each business increases operational complexity.
- The database has weak scaling capabilities.

Dify New Architecture



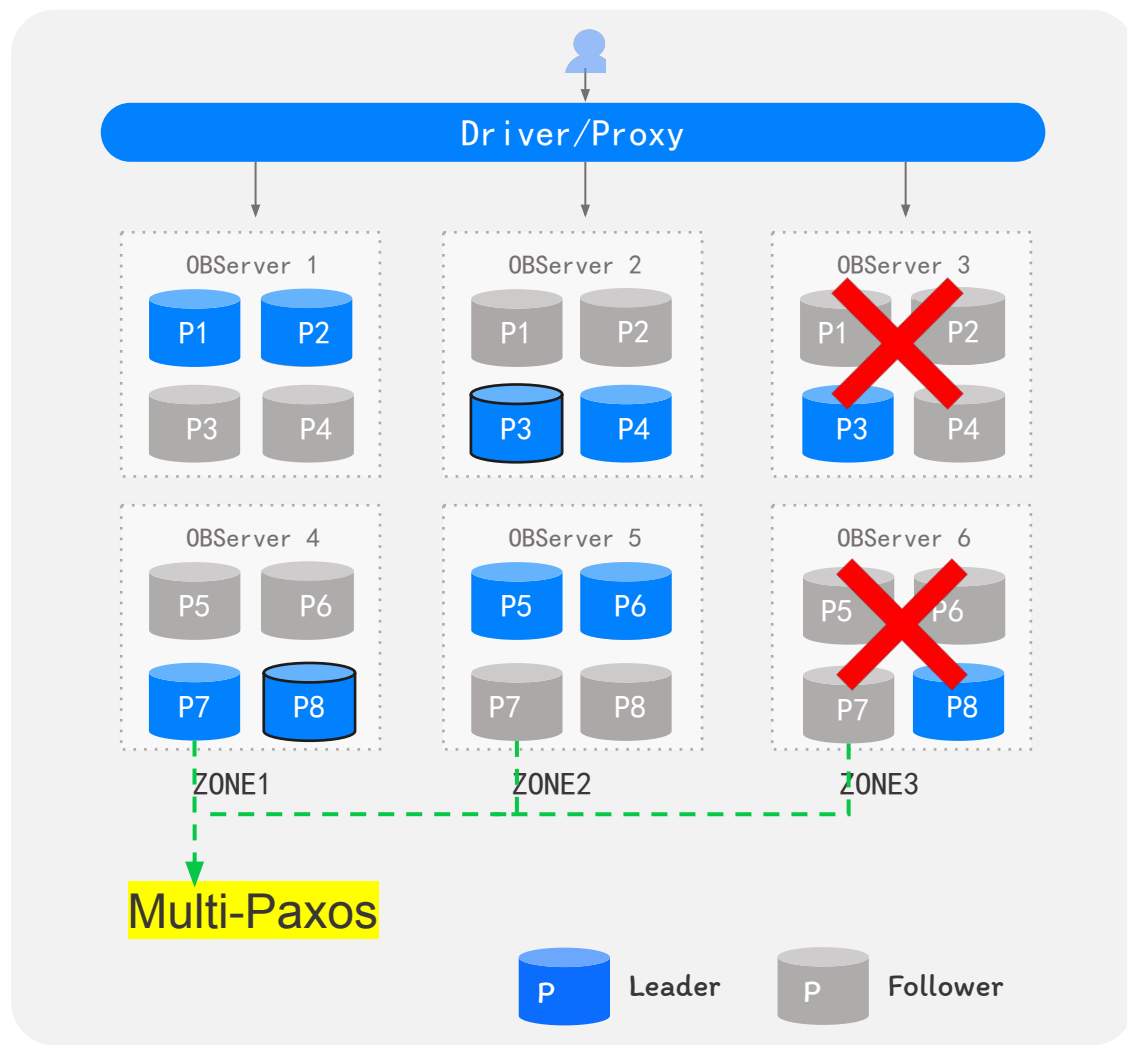
OceanBase Solution

- Centralized management of metadata, vector data, and KV data within a single OceanBase database.
- Financial-grade database stability.
- Automatic recovery of database services **within 8 seconds** in case of failures.
- **Online scaling** when resources are insufficient.
- **Multi-tenancy**: one tenant can support a single business, while a single OceanBase cluster can support multiple businesses simultaneously.
- Resource isolation between tenants.
- Web-based unified operations and monitoring platform for simplified administration.

04

High Availability and Low Cost

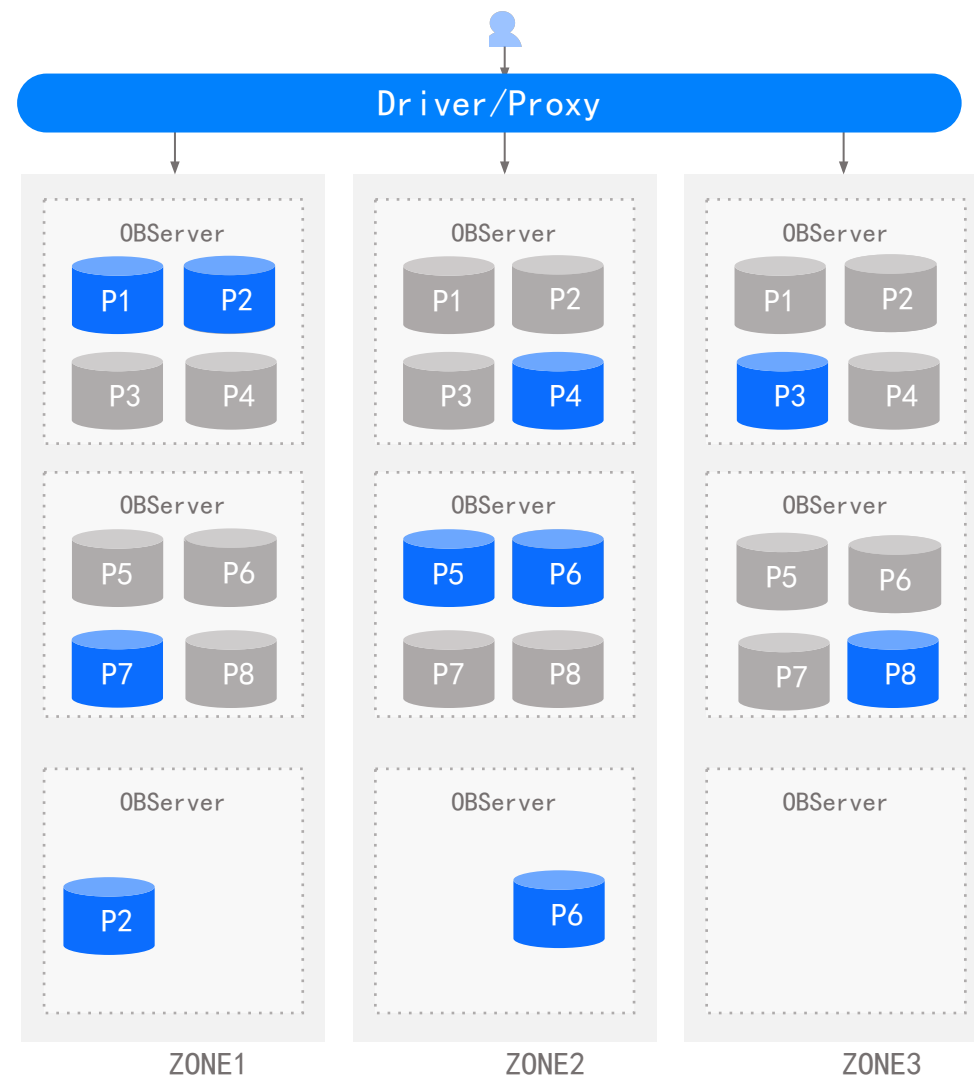
OceanBase High Availability



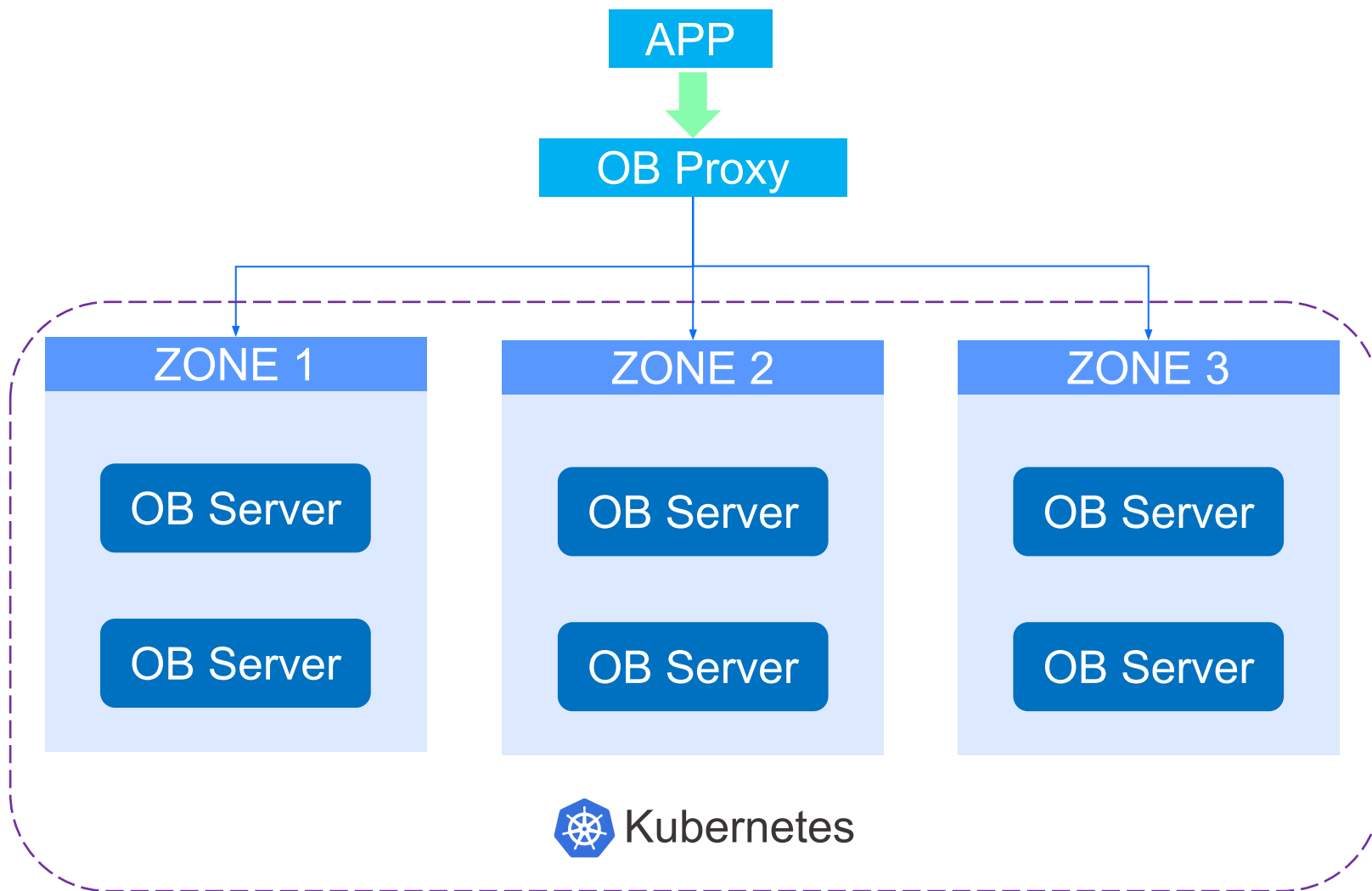
RTO < 8s

RPO = 0

OceanBase Scalability



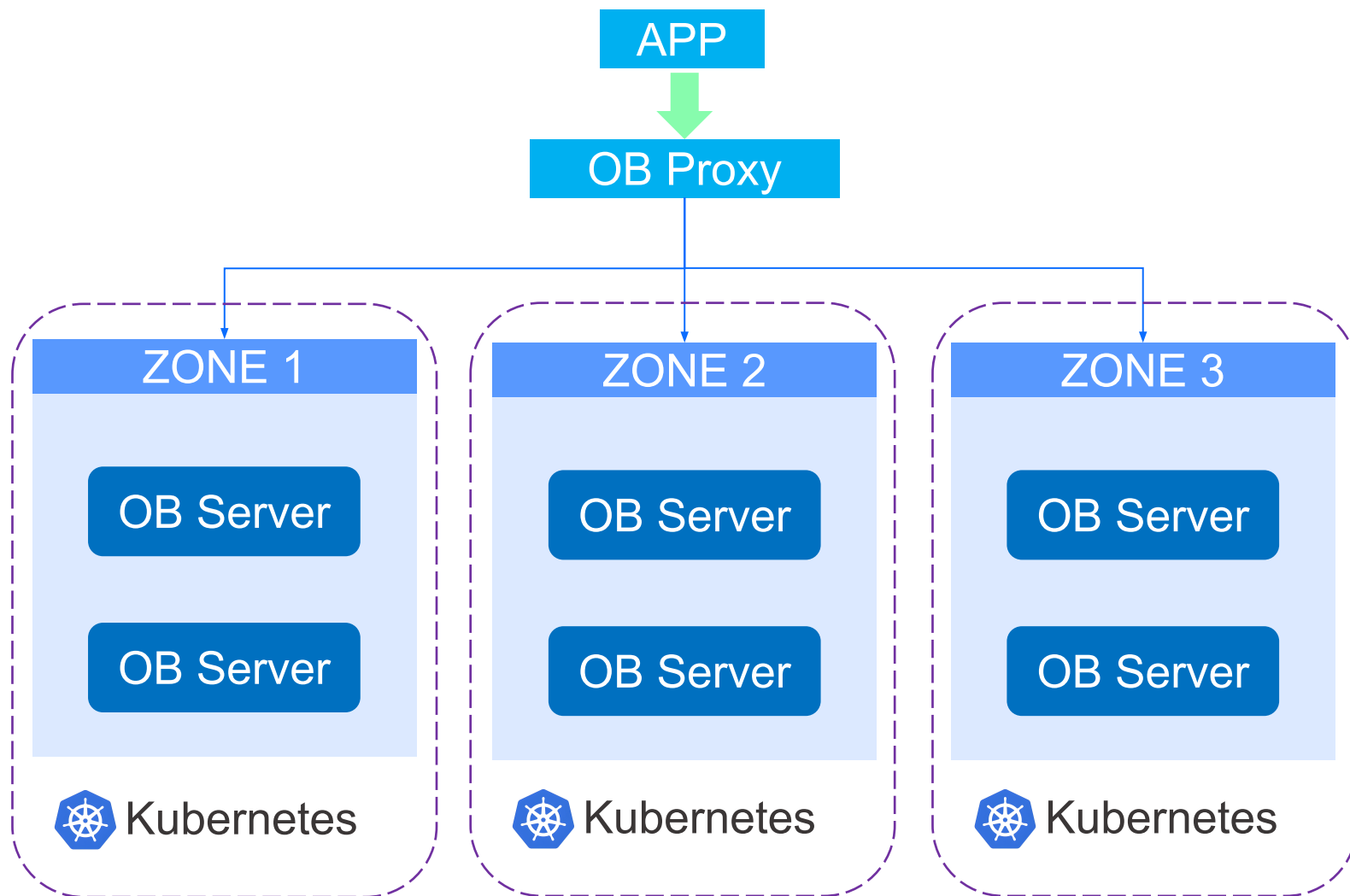
The Risk of Relying on One Cluster



Risks:

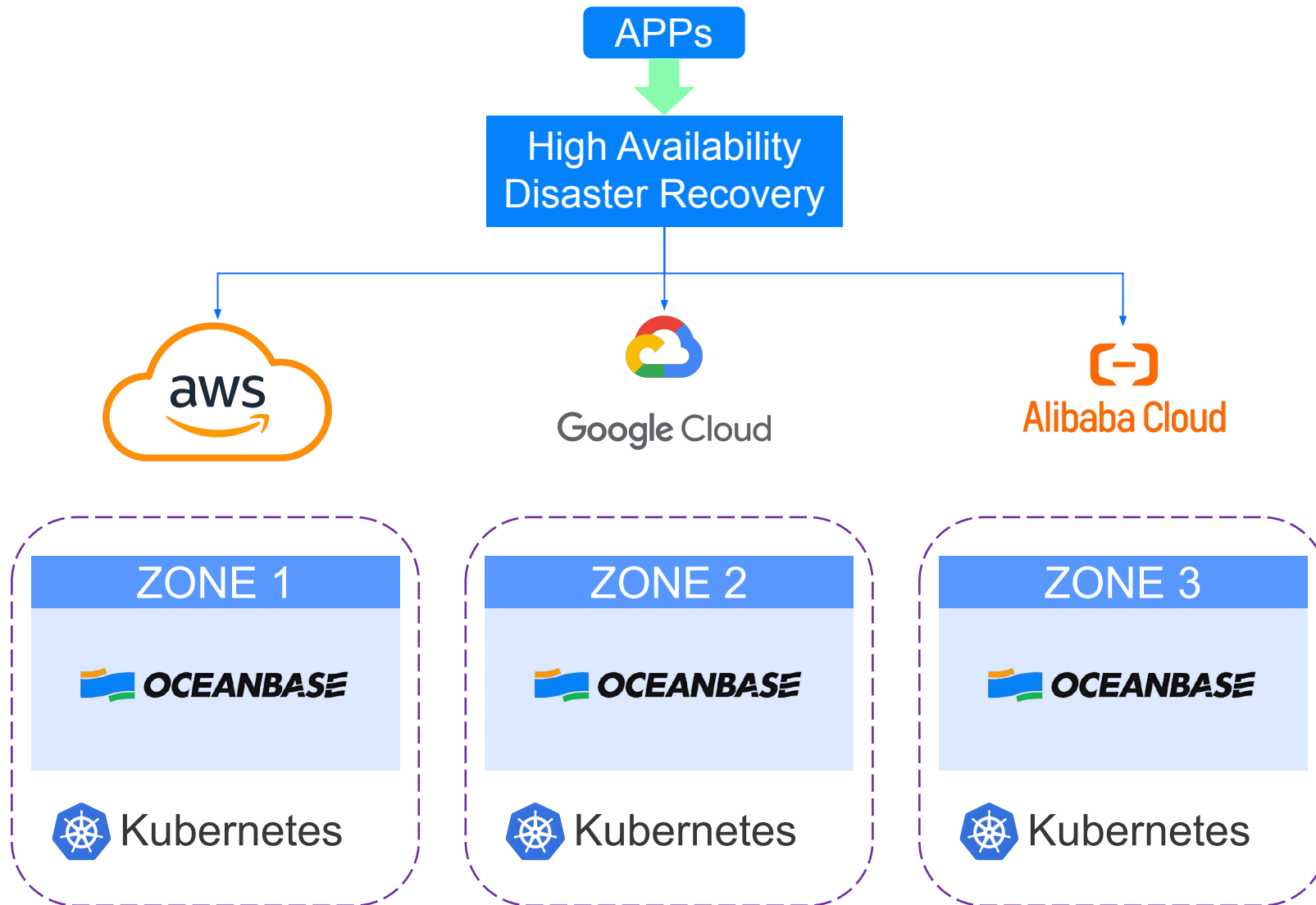
- Single point of failure
- Impact on upgrades
- High O&M pressure

Why Multi-Cluster?



- Cross-cluster deployment enables **true high availability**
- Improves **disaster recovery capabilities**
- Avoids **regional failures**
- Supports **smooth Kubernetes upgrades**
- Lays the foundation for **hybrid-cloud architecture**

Proven Benefits in Real Scenarios



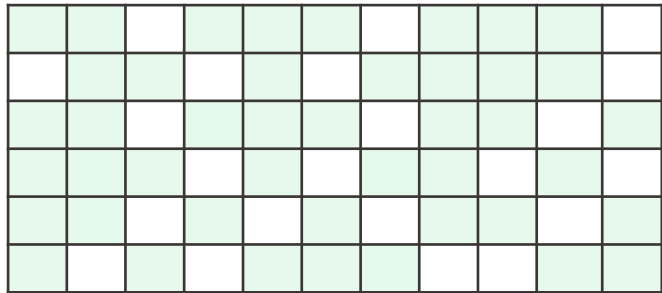
Scenarios for Using This Arch

- Customer runs stateless apps **across multiple cloud** providers
- Required OceanBase to run across **multiple K8s clusters**
- **Tolerating failures** even at the **cloud provider** level
- Designed to deliver **high availability** and **disaster resilience**

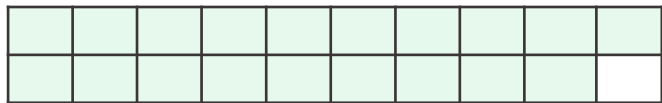
OceanBase Cost-Reduction Techniques

Data storage cost reduced by 70%–90%

Centralized storage database (3 TB)



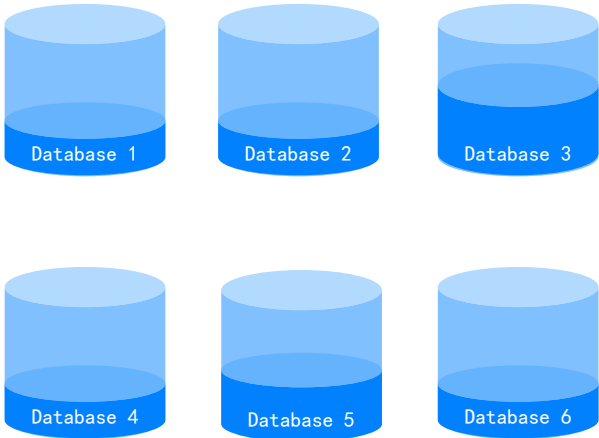
Data storage cost reduced by 70%–90%



OceanBase (1TB)

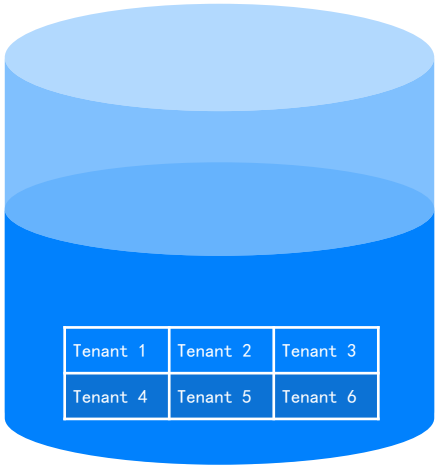
TCO reduced by 30%–50%

Centralized Storage Database



Database Consolidation

OceanBase Multi-Tenancy





05

Customer Success

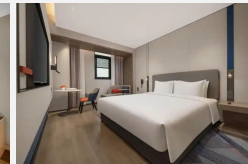
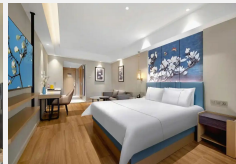
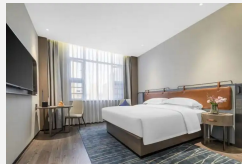
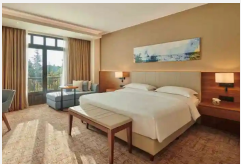
Hybrid Search for Hotel Photos (Migrated from Elasticsearch)

Business Requirements

- Use photo similarity search to find similar hotels and speed up the booking process.
- Identify the cheapest hotel supplier, as your hotel may have multiple suppliers.

Challenges

- The dataset is massive, currently at 340 million rows and expected to reach **800 million** in the future.
- It is impossible to analyze the 3 million new records added daily in real time.
- Building vector indexes is too slow.
- Implementing vector and scalar queries is highly challenging.



Benefits

- Stable operation for over a year to date.
- Significantly reduces AI application development complexity using SQL queries.
- OceanBase offers strong **hybrid query** capabilities.
- Uses disk-based IVF algorithm for building vector indexes, supporting storage and querying of 800 million records.
- Vector index creation speed improved by **more than 4x**.
- Enables real-time analytics.

Question-and-Answer System



Business Requirements

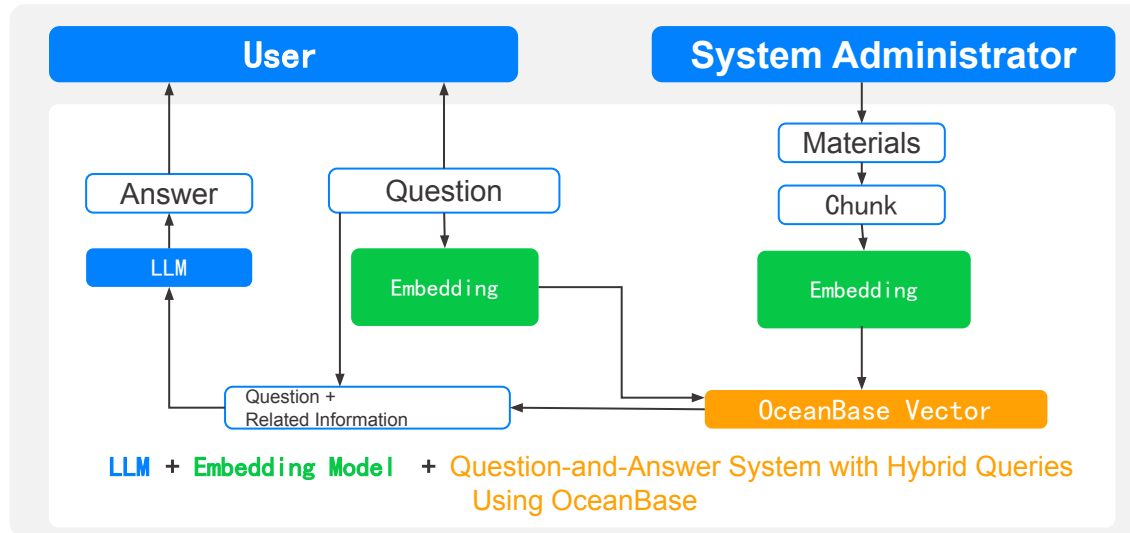
- When a driver asks a question, the system can quickly provide answers using past or current responses stored in the system.
- Questions may include audio data, images, or long text.
- Capable of performing searches using images.

Challenges(weaviate)

- Weaviate has **crashed** several times unexpectedly, with no solution other than restarting the database.
- **Lacks hybrid query** capabilities combining full-text search and vector queries, and integrating this capability is costly.
- To improve query performance, data must be deleted daily.
- Vector index construction is slow.

Benefits

- OceanBase provides **financial-grade** database services, ensuring business stability.
- In the event of a failure, the database service automatically recovers **within 8 seconds**.
- The HNSW + BQ algorithm significantly **reduces memory** consumption for vector indexes.
- Vector index construction time is shortened to under 25 minutes.
- Full-text search and vector queries can be performed within the same database.



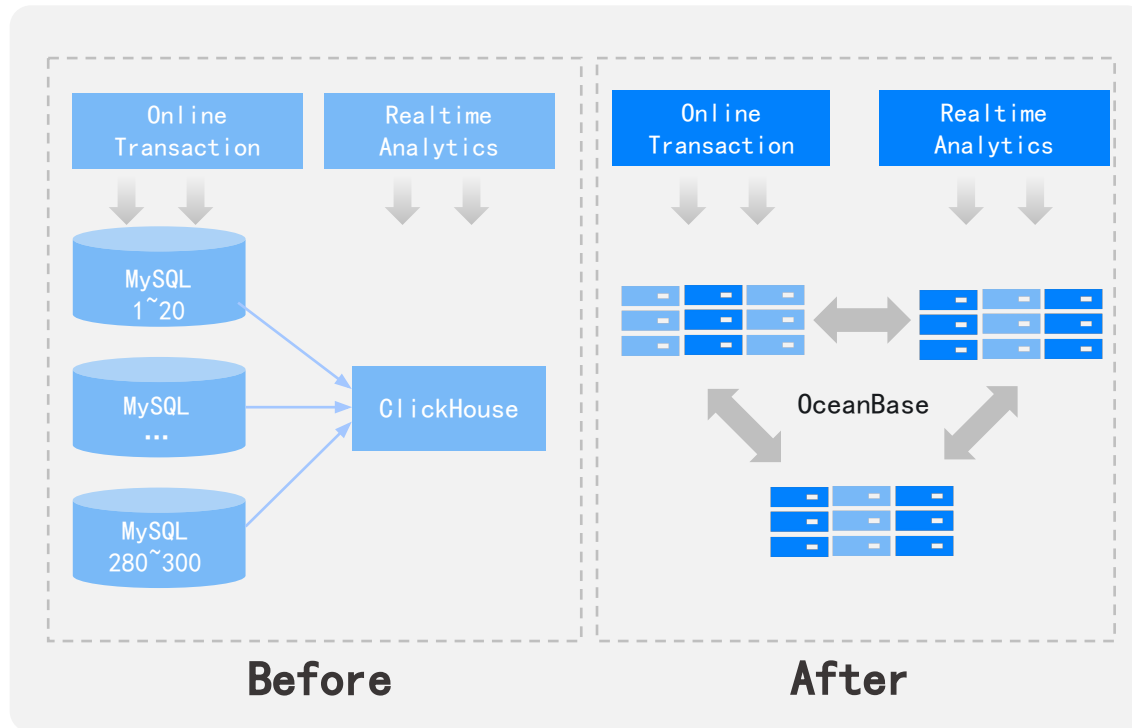
Kwai: Saves over 75% of server costs



Kwai: One of the world's most popular short-video apps, with over 700 million monthly active users.

Business Requirements

- Real-Time Analysis of Transaction Data



Challenges

- High Costs: Requires **over 400 servers**.
- During peak business periods, data synchronization experiences a 10-minute delay.
- Managing **400** servers and database instances is extremely challenging.

Benefits

- No data synchronization needed, enabling real-time analysis.
- Saves over 300 servers.
- Storage cost: 1 PB (MySQL + ClickHouse) → 200TB.

Thank you!

