



GOPS 2025
Shenzhen



ANNIVERSARY
2015-2025



GOPS 全球运维大会

2025 - XOps 风向标



深圳站

暨研运数智化技术峰会

时间：2025年4月25日-26日

地址：中国·深圳

指导单位：



主办单位：



承办单位：



AI时代数据管理的 核心挑战与跃迁之路

叶正盛

2025-04





叶正盛

NineData 创始人 & CEO

- 资深数据库与云计算领域专家
- 曾担任阿里云数据库产品管理与解决方案部总经理，阿里云技术架构与产品决策委员会核心成员
- 阿里巴巴去 IOE、异地多活、云计算多次技术变革核心成员
- 构建阿里巴巴&蚂蚁集团数据库DevOps体系
- 创立云计算数据传输DTS、数据管理DMS、数据库备份DBS、数据库自动驾驶服务DAS等多款云计算数据库产品



目录/ CONTENTS

- 1 AI 数据管理挑战
- 2 向量数据库
- 3 NineData 产品创新
- 4 客户实践





AI 数据管理挑战



AI大模型体系分层图

AI原生应用



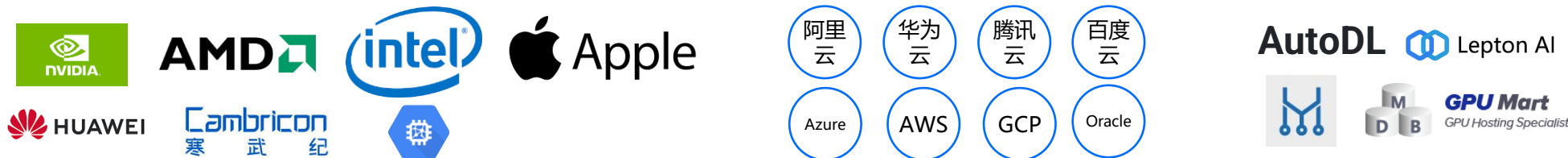
数据存储与处理



模型



算力



模型、数据、应用集成方案



MCP 架构



ChatGPT



Claude



KIMI



Qwen



deepseek



MISTRAL
AI_



ZHIPU · AI

MCP Server

Database

File
System

Browser

SaaS

API

Hardware

...

MCP Client

Claude Desktop

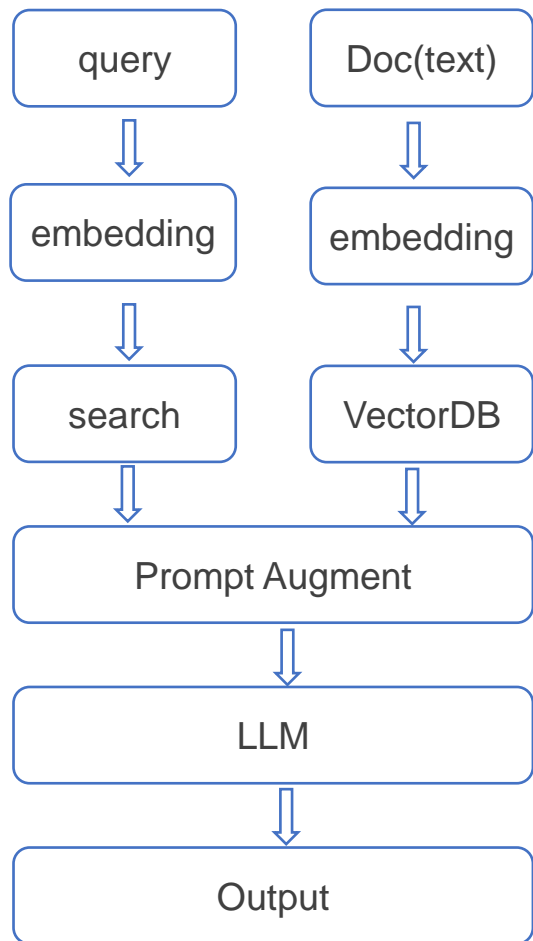
Cursor

Cline

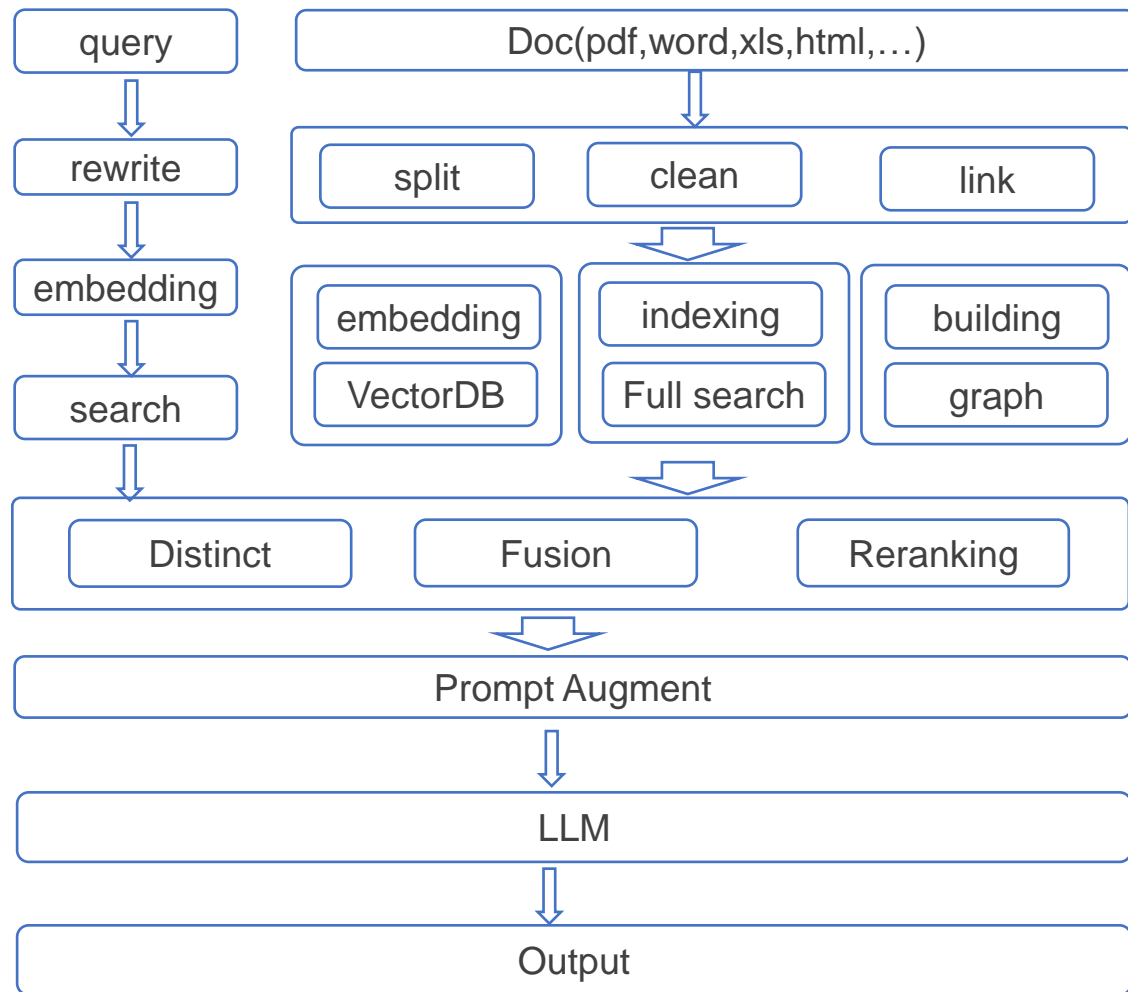
...

RAG的演进

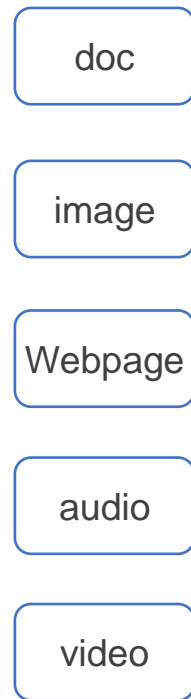
RAG 1.0



RAG 2.0



Future





AI向量数据库



常见数据模型

关系型数据库 (Relational)

Table Row Column

k	c1	c2	cN
1			
2			
3			
...			
n			

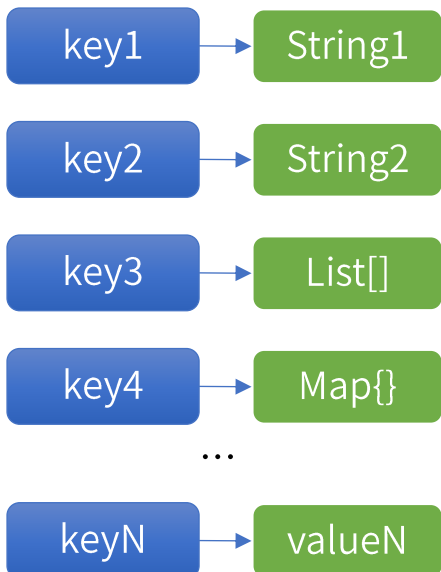
SQL

B-Tree

Column
Store

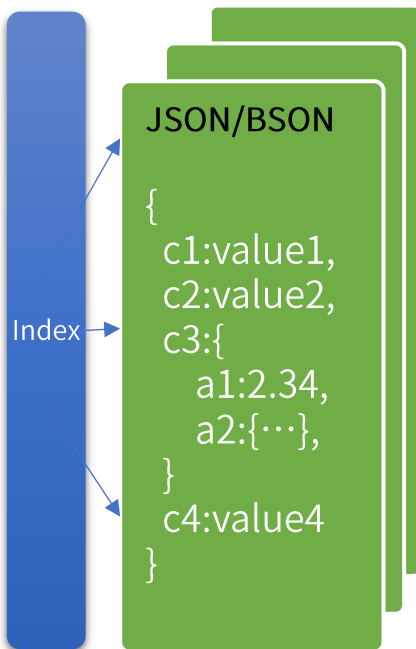
MySQL

KV数据库 (Key-Value)



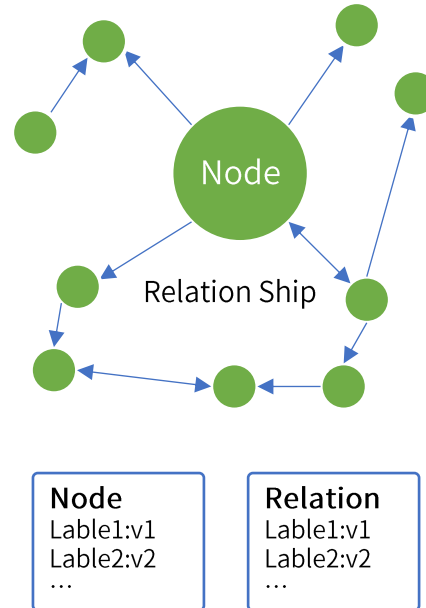
Redis

文档数据库 (Document)



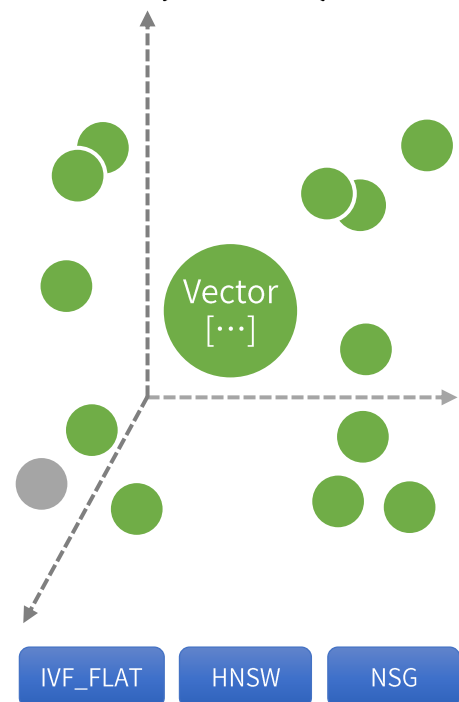
MongoDB

图数据库 (Graph)



NebulaGraph

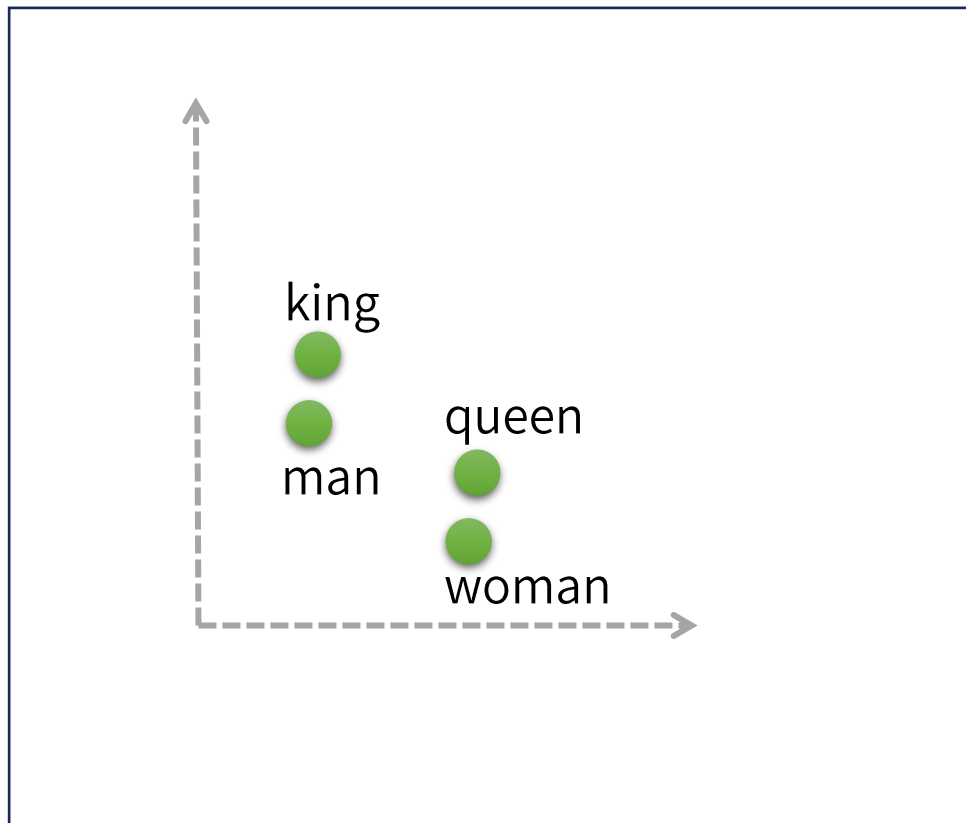
向量数据库 (Vector)



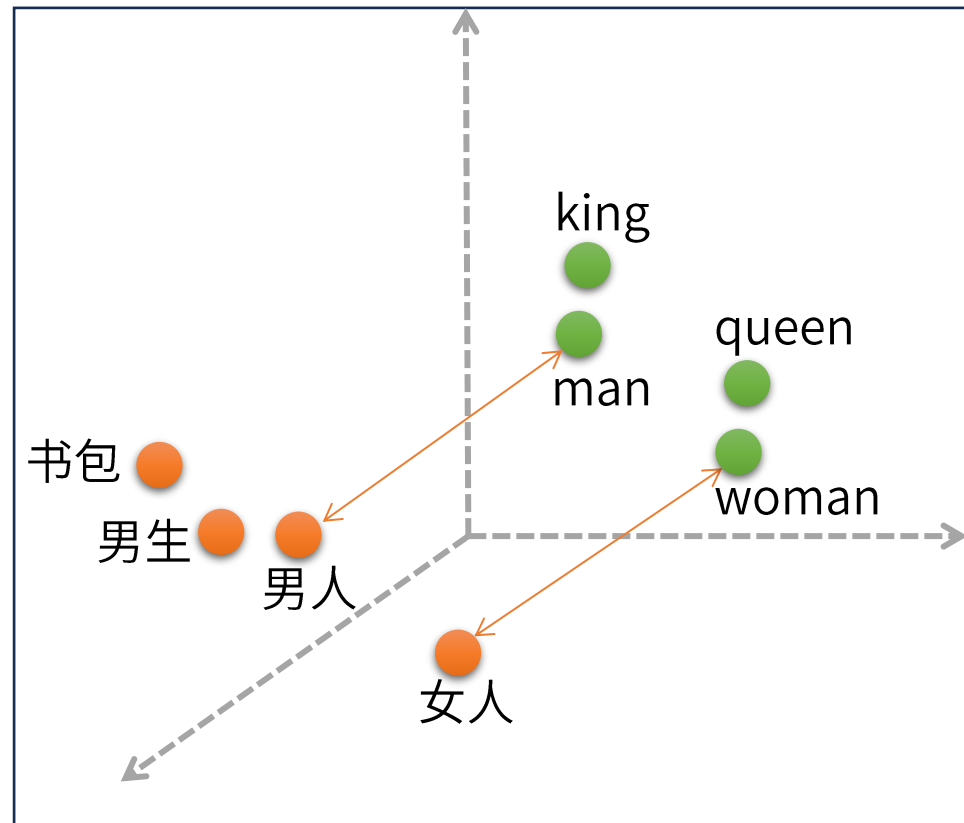
Milvus

多维向量数据示意图

2维



3维



主流向量数据库产品

- Milvus/Zilliz Cloud
- Chroma
- FAISS
- Qrant
- Weaviate
- Pinecone
- Vespa
- LanceDB
- PostgreSQL+pgVector
- Redisearch
- ElasticSearch
- OceanBase
- MySQL
- Cassandra

专业向量数据库



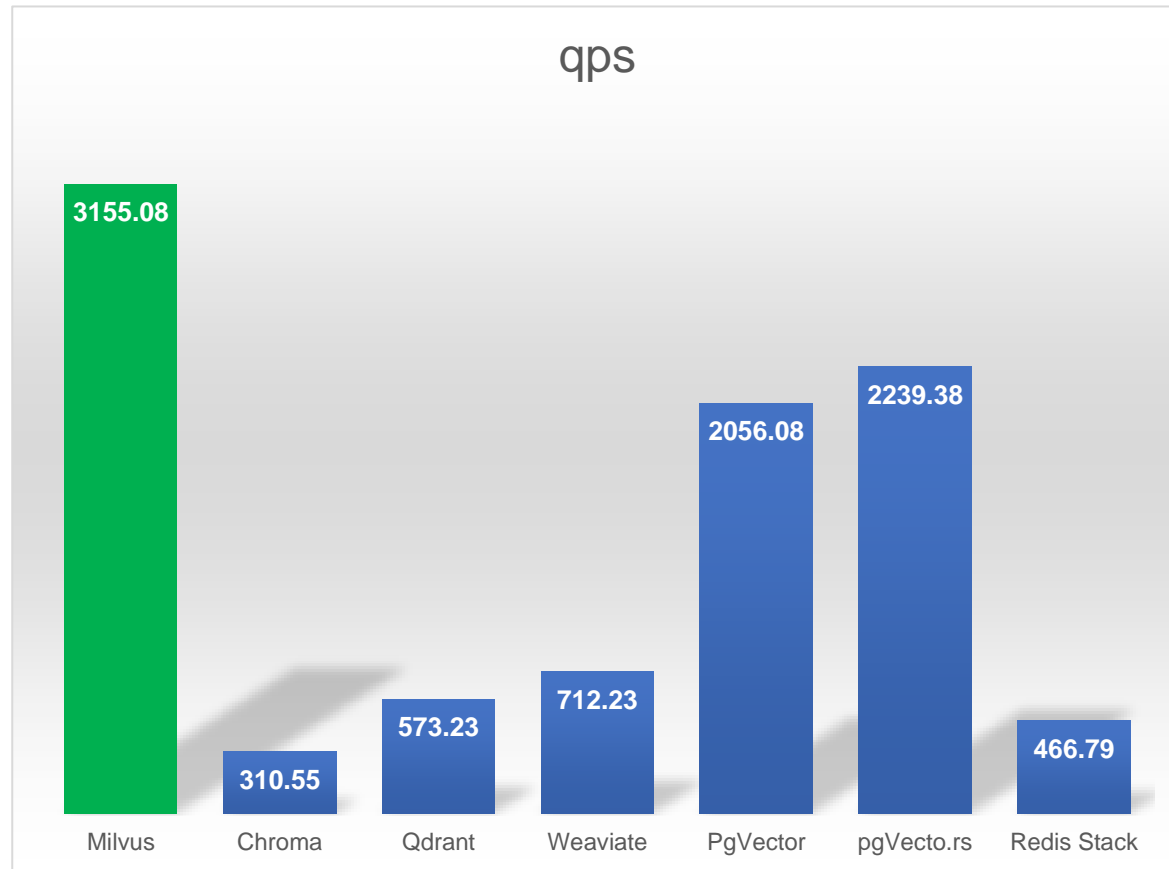
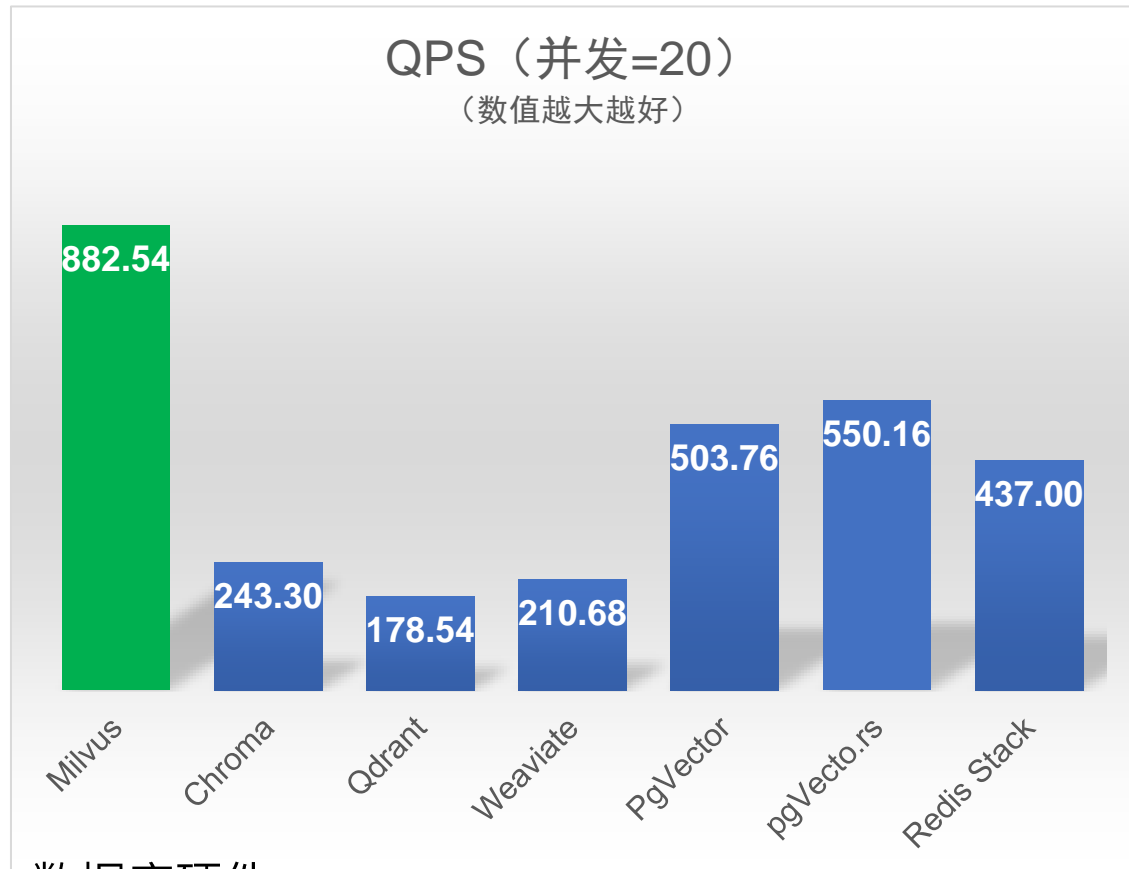
原数据库+向量引擎



向量数据库测试 (QPS)

数据: OpenAI 50K vectors, 1536 dimensions

索引配置: HNSM(m=16,ef_construction=256,ef_search=256)



数据库硬件:

阿里云 ECS.i4.xLarge, 4C/32GB/900GB SSD

阿里云 ECS.i4.4xLarge, 16C/64GB/900GB SSD

向量数据库查询测试 (100万, 768维)

数据: Cohere 1M vectors, 768 dimensions

索引配置: HNSM(m=16,ef_construction=256,ef_search=256)

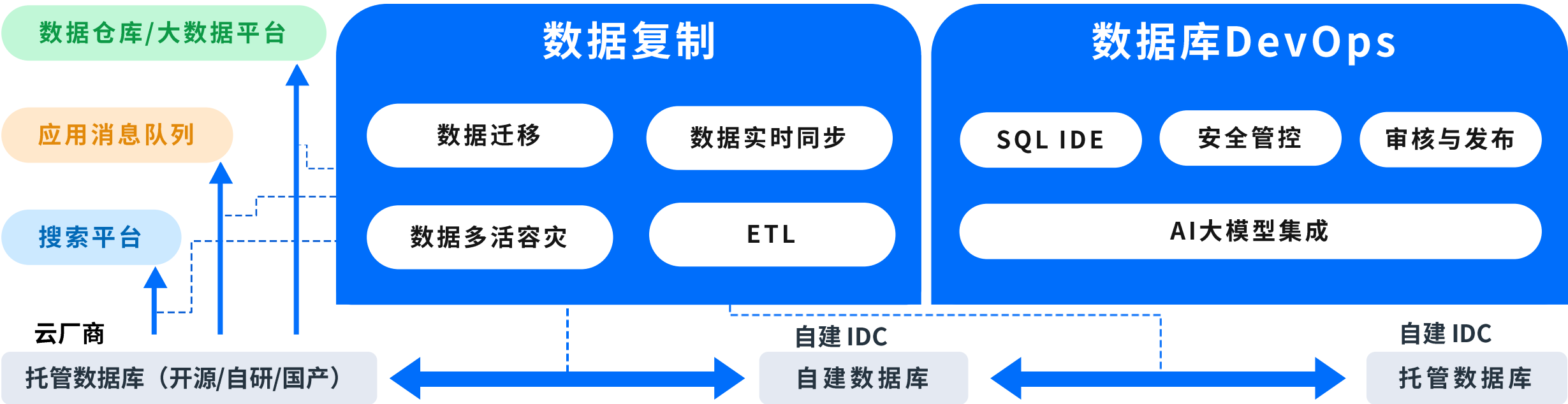
产品	load time	QPS	latency	recall
Milvus 2.5.6	976.42	1537	0.0043	0.9809
Chroma 0.6.4	985.71	349	0.0064	0.9562
Qdrant 1.13.4	1526.20	352	0.0083	0.9947
Weaviate1.28.11	1186.15	1065	0.0105	0.9563
Pg17-pgVector-0.8.0	404.52	1810	0.013	0.9704
Pg17-pgVecto.rs-0.4.0	1638.59	2306	0.0045	0.958
redis-stack7.4.2	1660.59	504	0.0036	0.9562

数据库硬件: 阿里云 ECS.i4.4xLarge, 16C/64GB/900GB SSD



NineData 技术创新







小度

记录用户的历史问答的上下文信息，增强大模型的记忆，秒级返回问答结果，提升用户使用小度的用户体验。

秒级 ⬆

返回问答结果



网盘

存储百亿文本，图片的向量化信息，给用户提供基于相似性的检索，可以搜相关的文件，指定的人物。

百亿级 ⬆

文档、图片的向量化信息



文库

对 4000w PPT 配图做向量化，满足用户生成 PPT 的需求。用户上传论文，秒级返回当前论文所引用论文片段。

秒级 ⬆

引用论文片段的返回

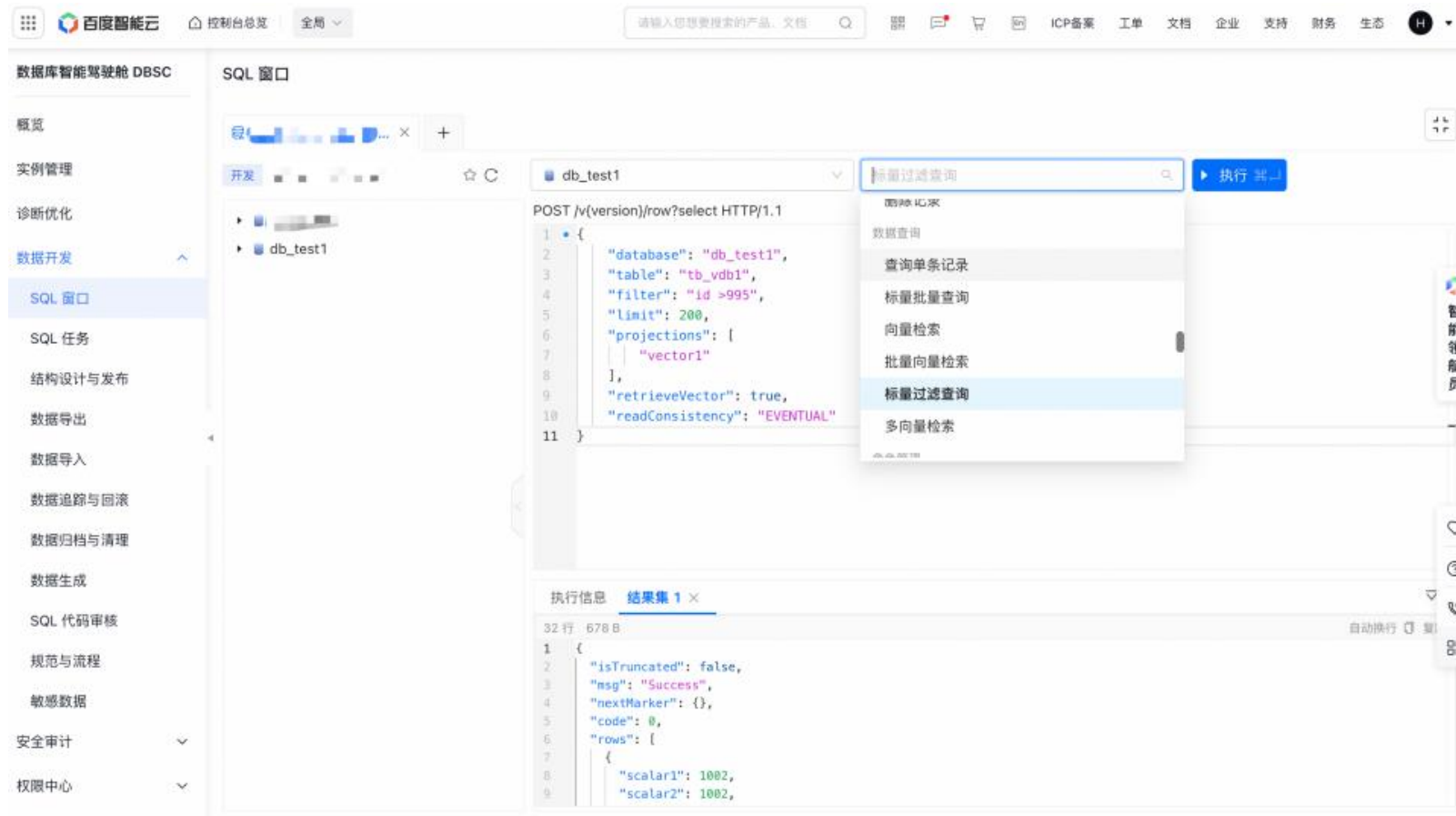


Comate 智能体

通过使用 VectorDB 存储编译构建错误信息和对应的解决方法，用于给报错的用户秒级返回最优的解决方案。

秒级 ⬆

最优的解决方案



The screenshot shows the Baidu Vector Database VDB management interface. The left sidebar contains navigation options: 概览, 实例管理, 诊断优化, 数据开发 (selected), SQL 窗口, SQL 任务, 结构设计与发布, 数据导出, 数据导入, 数据追踪与回滚, 数据归档与清理, 数据生成, SQL 代码审核, 规范与流程, 敏感数据, 安全审计, and 权限中心. The main area displays the SQL window for a database named 'db_test1'. A dropdown menu is open, showing options: 删除记录, 数据查询, 查询单条记录, 批量查询, 向量检索, 批量向量检索, 批量过滤查询 (highlighted), and 多向量检索. The SQL query is:

```
POST /v(version)/row?select HTTP/1.1
1 {
2   "database": "db_test1",
3   "table": "tb_vdb1",
4   "filter": "id > 995",
5   "limit": 200,
6   "projections": [
7     "vector1"
8   ],
9   "retrieveVector": true,
10  "readConsistency": "EVENTUAL"
11 }
```

 The execution results show a JSON response:

```
{
  "isTruncated": false,
  "msg": "Success",
  "nextMarker": {},
  "code": 0,
  "rows": [
    {
      "scalar1": 1002,
      "scalar2": 1002,
    }
  ]
}
```

百度向量数据库VDB

SQL for AI 向量数据库

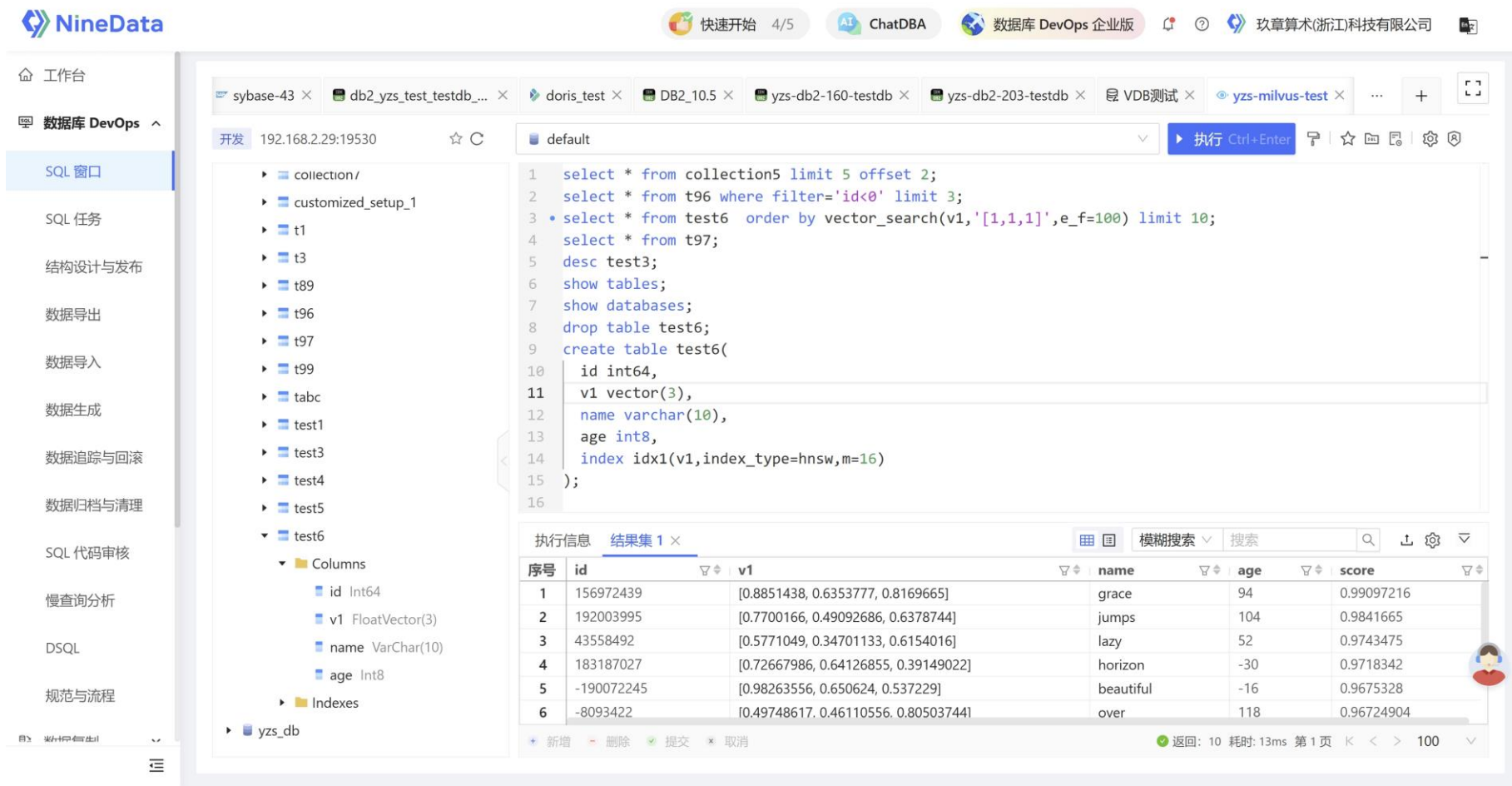
Milvus

Zilliz Cloud

Qdrant

Pinecone

Weaviate



The screenshot shows the NineData SQL interface. On the left is a sidebar with navigation options: 工作台, 数据库 DevOps, SQL 窗口, SQL 任务, 结构设计与发布, 数据导出, 数据导入, 数据生成, 数据追踪与回滚, 数据归档与清理, SQL 代码审核, 慢查询分析, DSQL, and 规范与流程. The main area displays a SQL query in a text editor and its execution results in a table.

SQL Query:

```
1 select * from collection5 limit 5 offset 2;
2 select * from t96 where filter='id<0' limit 3;
3 • select * from test6 order by vector_search(v1,'[1,1,1]',e_f=100) limit 10;
4 select * from t97;
5 desc test3;
6 show tables;
7 show databases;
8 drop table test6;
9 create table test6(
10   id int64,
11   v1 vector(3),
12   name varchar(10),
13   age int8,
14   index idx1(v1,index_type=hnsf,m=16)
15 );
16
```

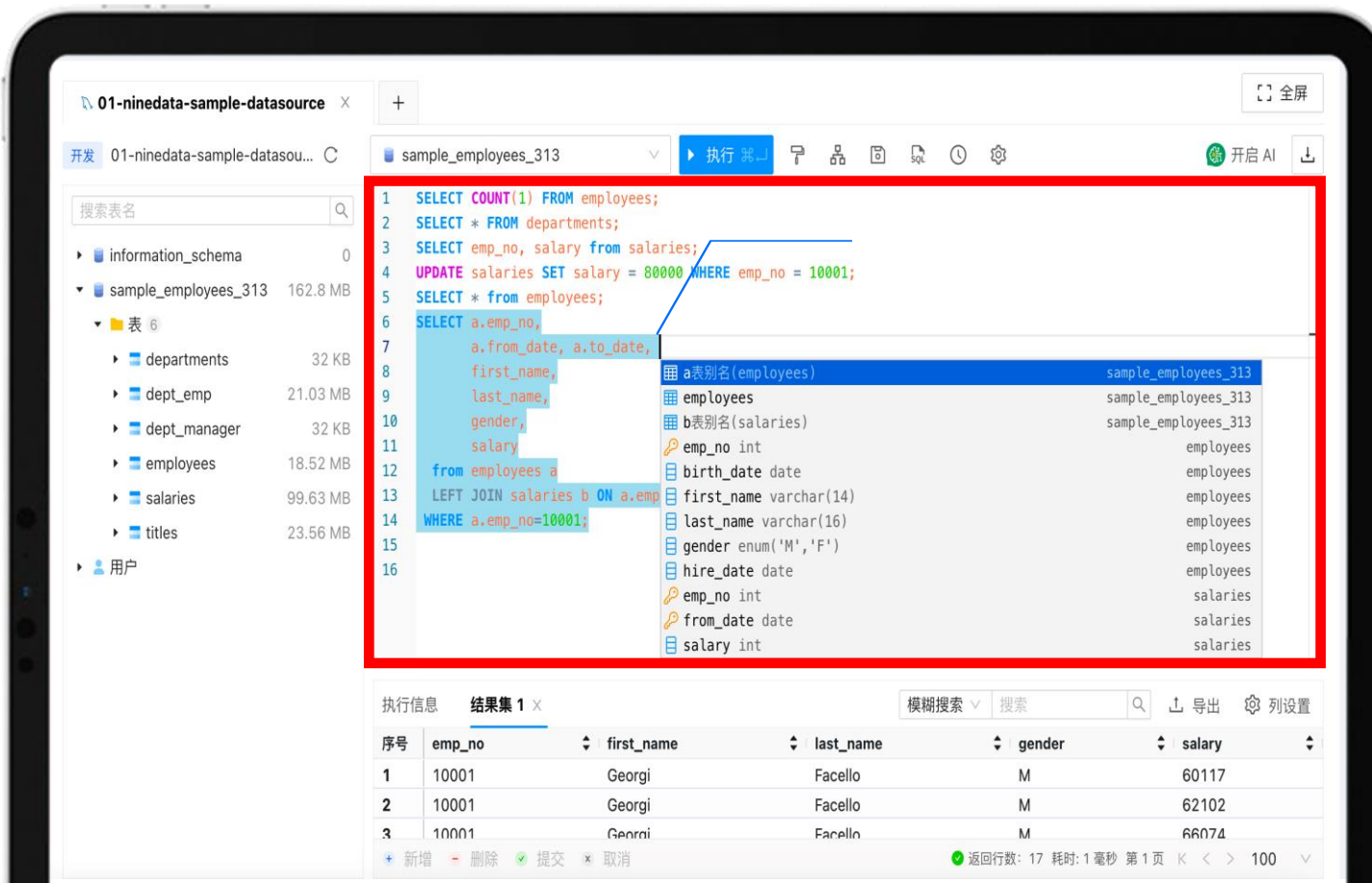
Execution Results (Table 1):

序号	id	v1	name	age	score
1	156972439	[0.8851438, 0.6353777, 0.8169665]	grace	94	0.99097216
2	192003995	[0.7700166, 0.49092686, 0.6378744]	jumps	104	0.9841665
3	43558492	[0.5771049, 0.34701133, 0.6154016]	lazy	52	0.9743475
4	183187027	[0.72667986, 0.64126855, 0.39149022]	horizon	-30	0.9718342
5	-190072245	[0.98263556, 0.650624, 0.537229]	beautiful	-16	0.9675328
6	-8093422	[0.49748617, 0.46110556, 0.80503744]	over	118	0.96724904

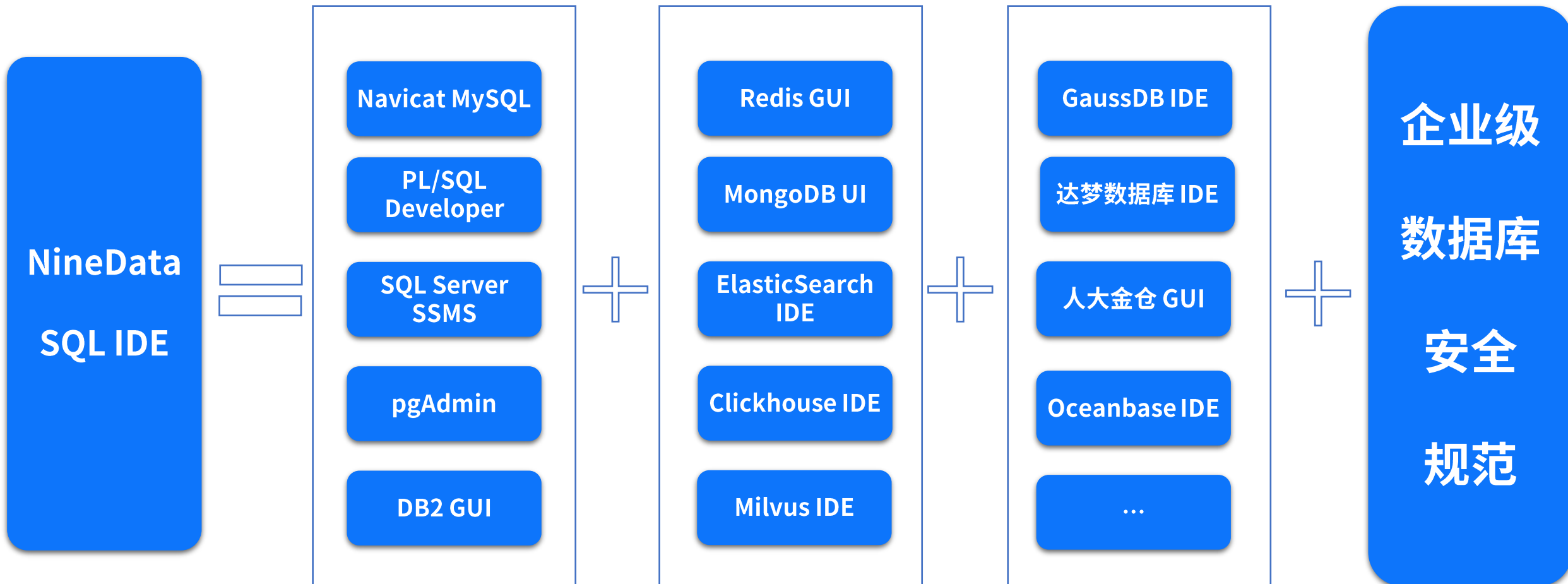
At the bottom of the results table, it shows: 返回: 10 耗时: 13ms 第 1 页 < > 100

Smart Edit, 研发效率提升50%

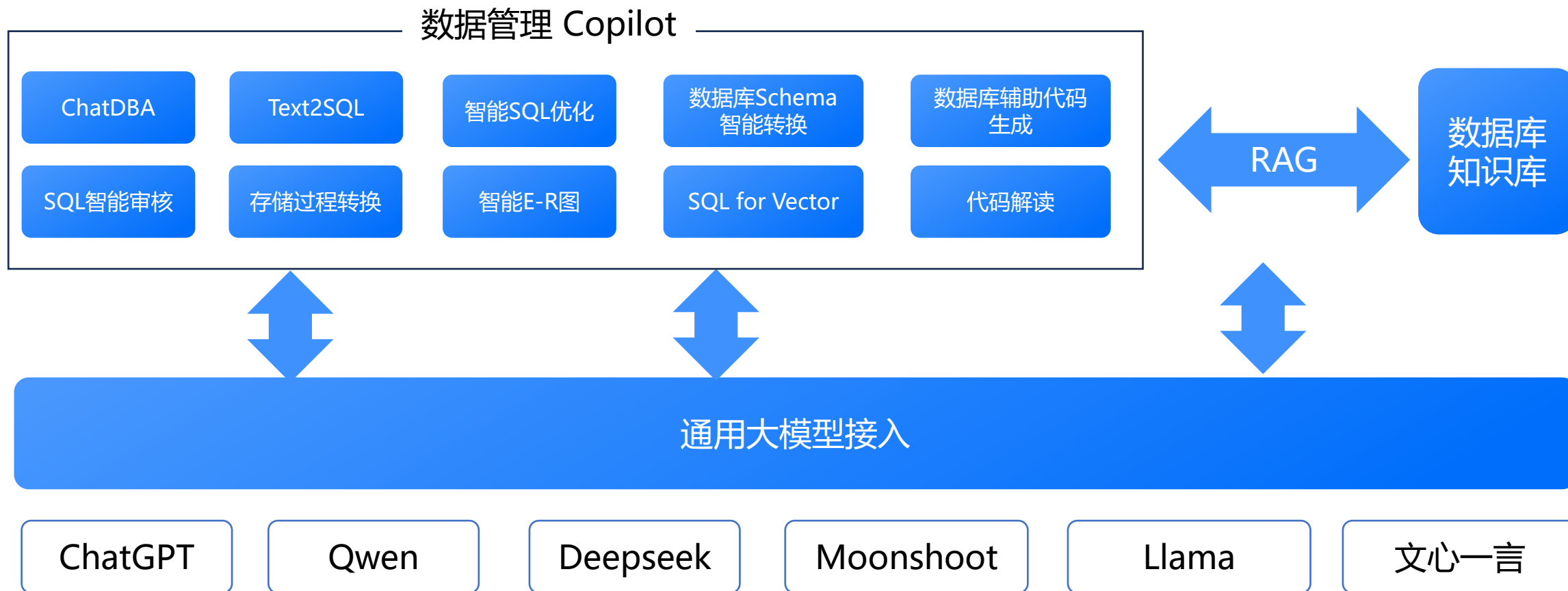
- 代码智能提示
- ChatDBA
- Text2SQL
- 数据库安全保护
- 敏感数据脱敏
- 支持60种数据库



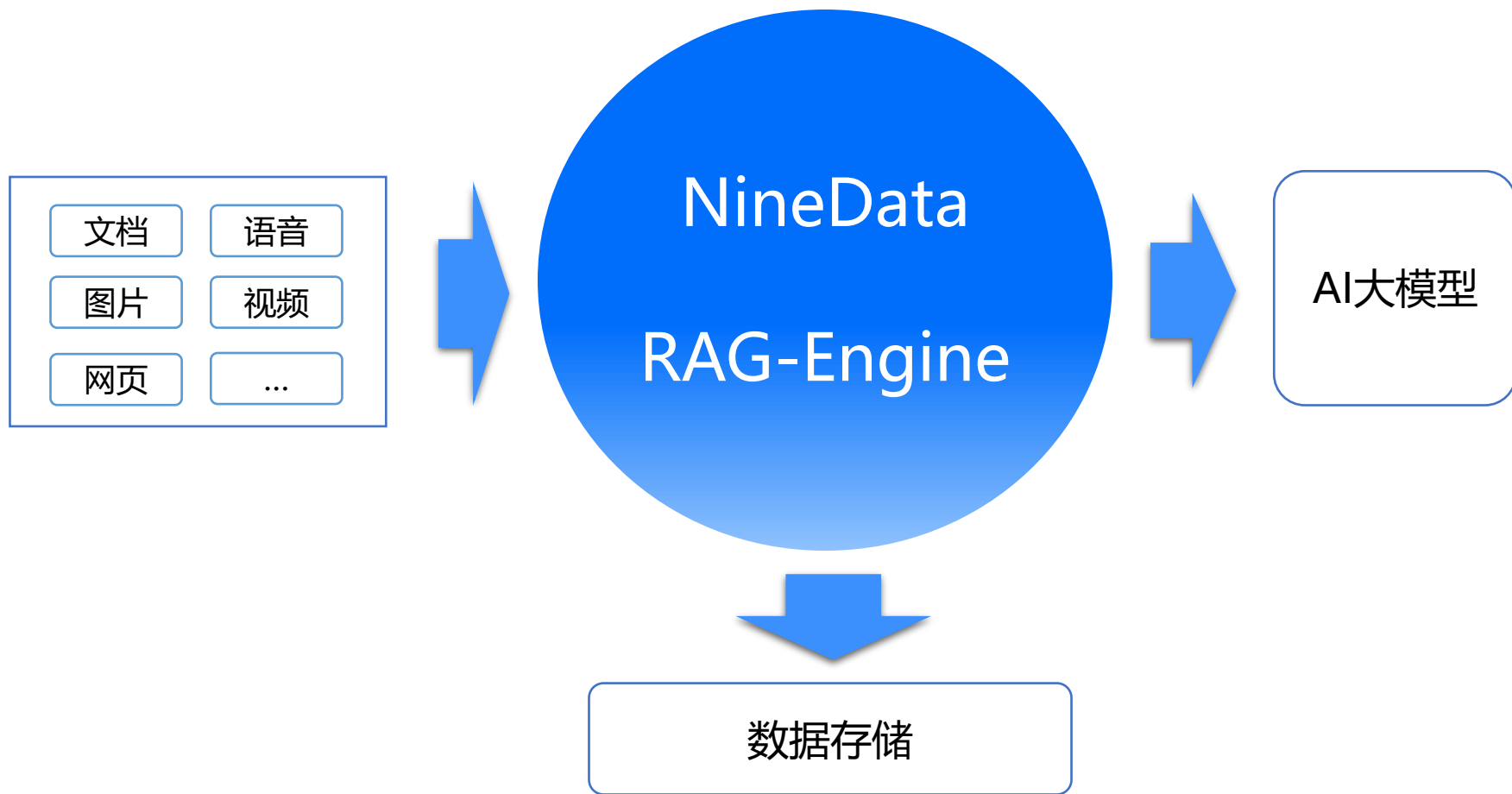
NineData SQL IDE



NineData AI 数据管理Copilot



AI 数据管理 (RAG-Engine)



NineData 数据管理 十五年经验积累

iDB

2010年

面向一家企业（阿里巴巴）

- 阿里巴巴集团数据库服务平台
 - 专注解决数据库研发效率、数据安全
 - 阿里集团高效数据库服务平台-iDB ([链接](#))
- @2013 Velocity China 2013
by 叶正盛

DMS

2013年

面向一个云平台（阿里云）

- 阿里云数据库的数据管理产品
 - 十年磨一剑，阿里巴巴企业级数据管理平台 DMS ([链接](#))
- @2017 阿里云DMS商业化



2021年~

AI时代多云数据管理

- 企业级数据库DevOps
- 多云、多数据库支持
- AI大模型集成
- AI数据管理
- 云原生技术架构



客户实践与案例



向量数据库VDB 数据管理



小度

记录用户的历史问答的上下文信息，增强大模型的记忆，秒级返回问答结果，提升用户使用小度的用户体验。

秒级 ⬆

返回问答结果



网盘

存储百亿文本，图片的向量化信息，给用户 提供基于相似性的检索，可以搜相关的文件，指定的人物。

百亿级 ⬆

文档、图片的向量化信息

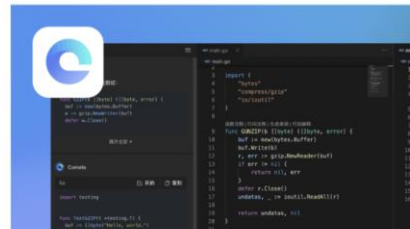


文库

对 4000w PPT 配图做向量化，满足用户生成 PPT 的需求。用户上传论文，秒级返回当前论文所引用论文片段。

秒级 ⬆

引用论文片段的返回



Comate 智能体

通过使用 VectorDB 存储编译构建错误信息和对应的解决方法，用于给报错的用户秒级返回最优的解决方案。

秒级 ⬆

最优的解决方案

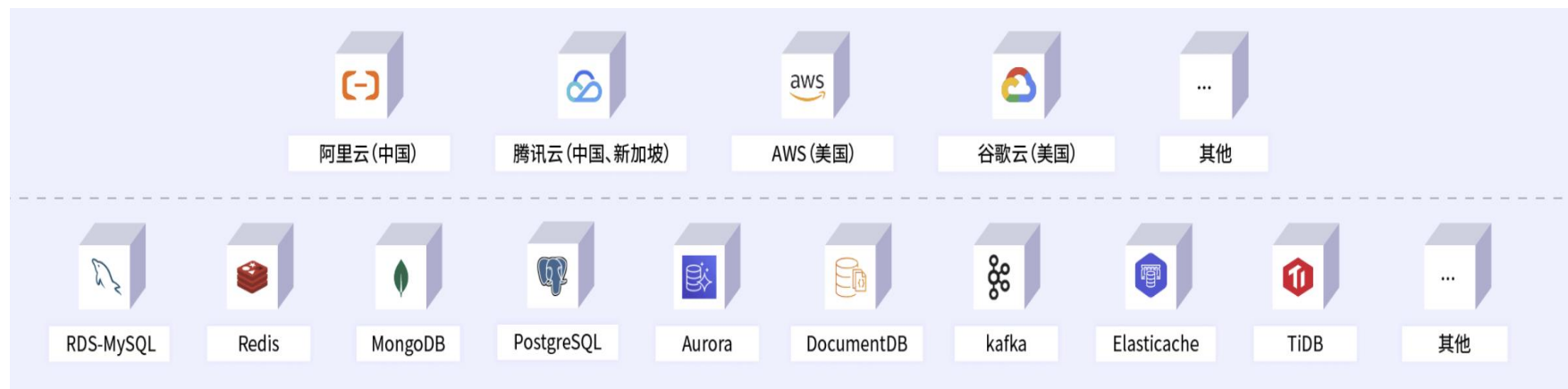
百度向量数据库VDB

AI独角兽MINIMAX 数据库DevOps

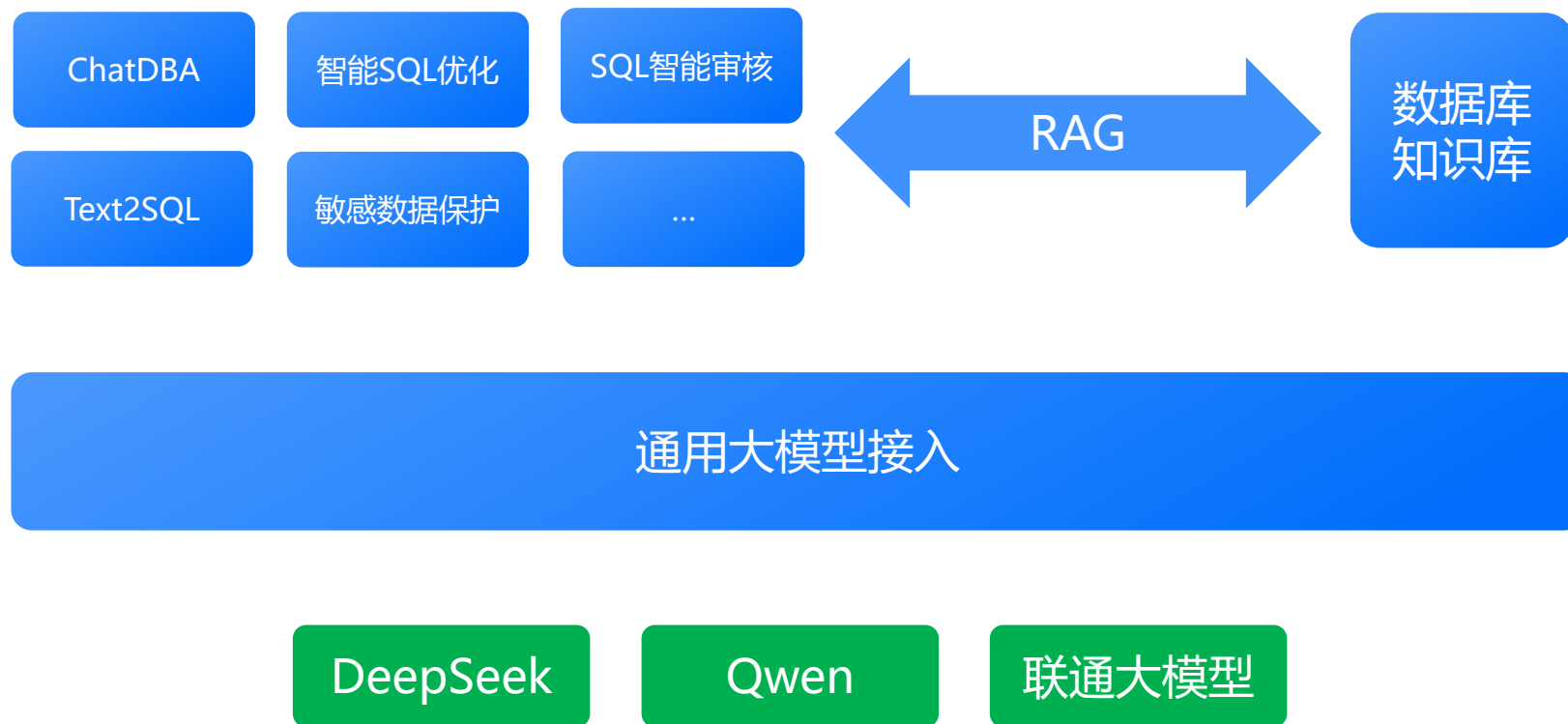


MINIMAX
AI大模型独角兽

- 1个人管理数百个数据库
- 4朵云，数百个数据库接入
- 1个月完成数据库DevOps在公司全面上线

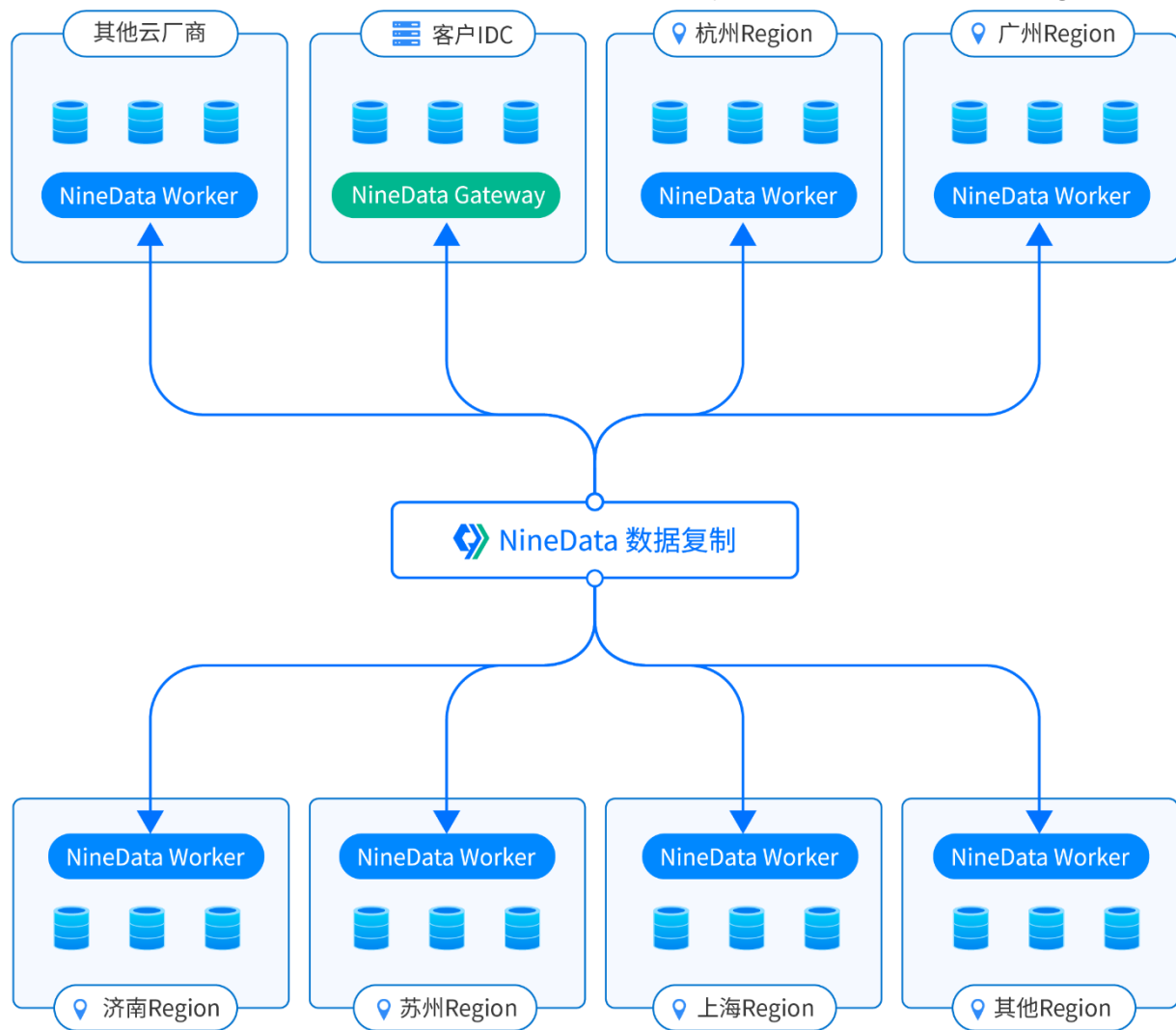


运营商 AI 数据库管理合作创新



中国移动云数据复制

数据库种类: Mysql,sqlserver,mongodb,redis,clickhouse,kafka,oracle,gaussdb



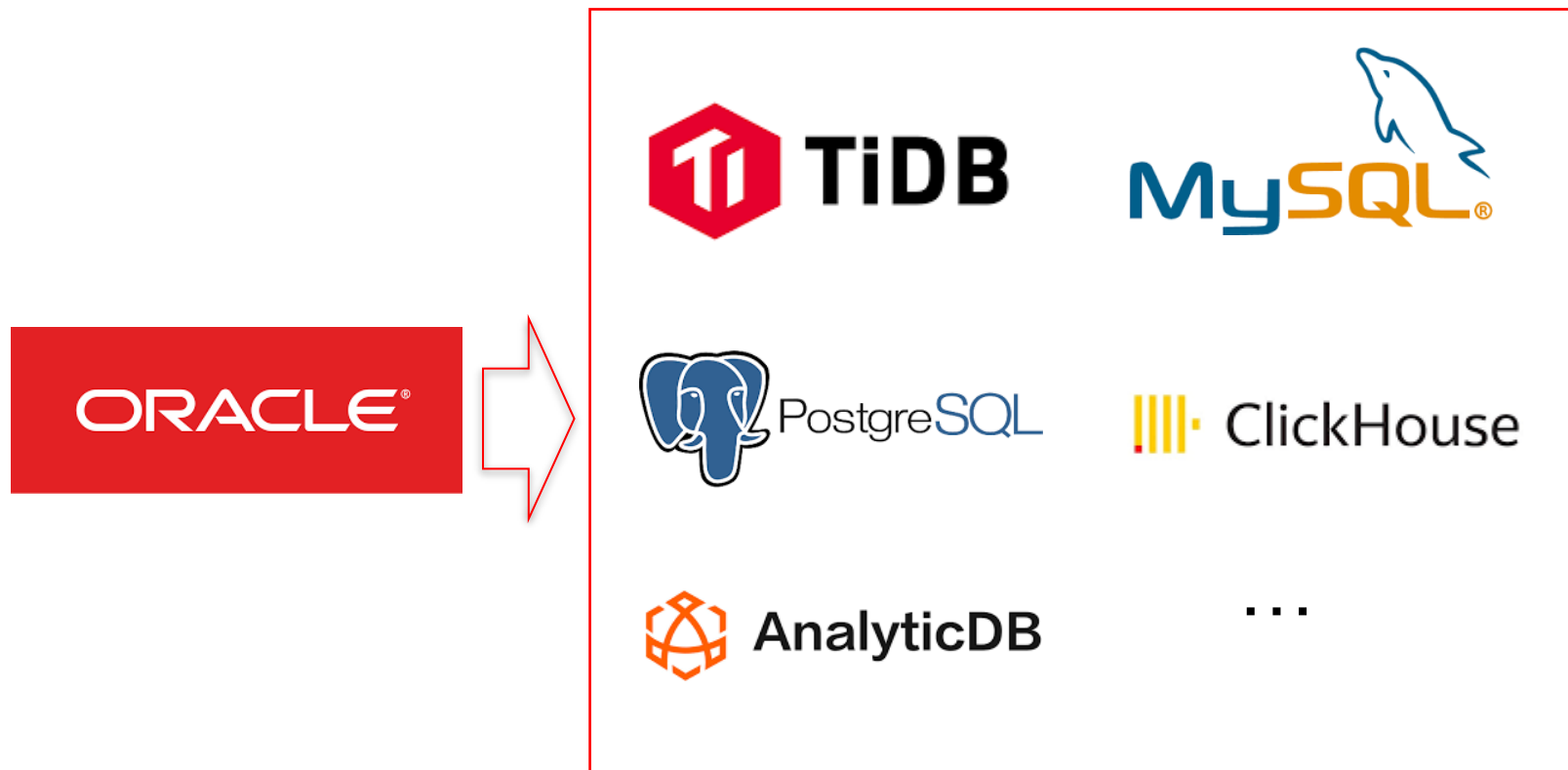
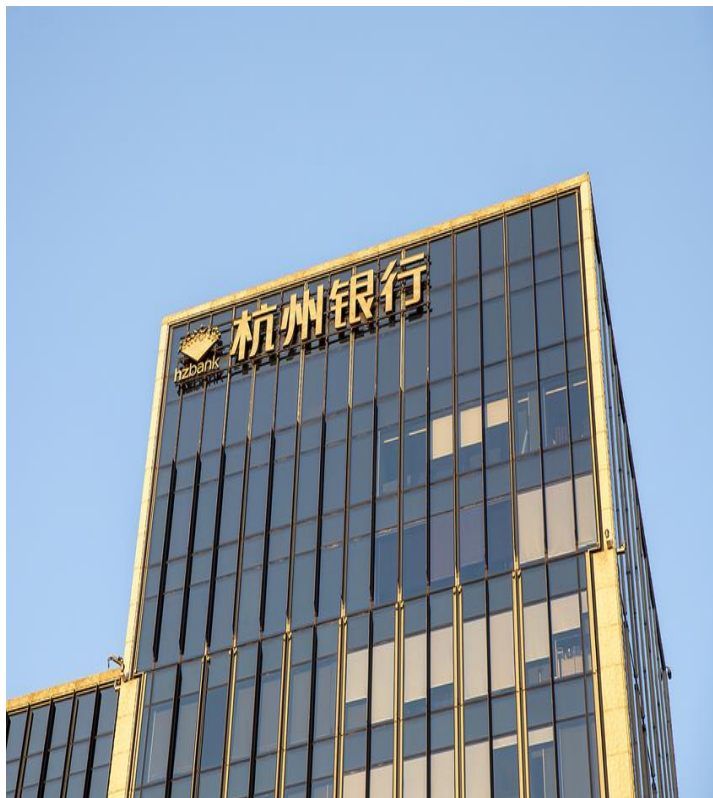
❓ 主要问题与挑战

- 客户本身有很多数据需要同步，同时还要提供数据复制产品给他的客户使用，比如该云的客户从其他云厂商或自建系统中迁移上云。
- 应用场景复杂：包括迁移上云、跨云迁移、跨区域迁移、数据容灾、异地多活等业务场景。
- 网络环境复杂：Region 内部/ Region 之间，和其他云厂商与客户自有系统之间等各种链路。

👍 使用成效

- 每天稳定运行的数据链路超过 200+ 。
- 通过 NineData 双向复制构建了超长距离（超过 1000 公里）的异地多活集群，支撑业务高稳定运行。

杭州银行数据库国产化



去O、重新选型、数据迁移质量、性能...

T H A N K S

感谢大家观看

2025.4