DTCC

第十四届中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA

数智赋能 共筑未来

IT168.com ChinaUnix ITPUB

北京国际会议中心 | 2023/8/16-18

向量数据库在亿贝智能营销的实践

eBay资深软件开发工程师
涂晓东

议程

1. 背景介绍
2. 向量数据库集群部署与落地
3. 向量数据库生产实践优化
4. 展望未来

1. 背景介绍

- 业务场景和技术要求
- 向量数据库产品选型

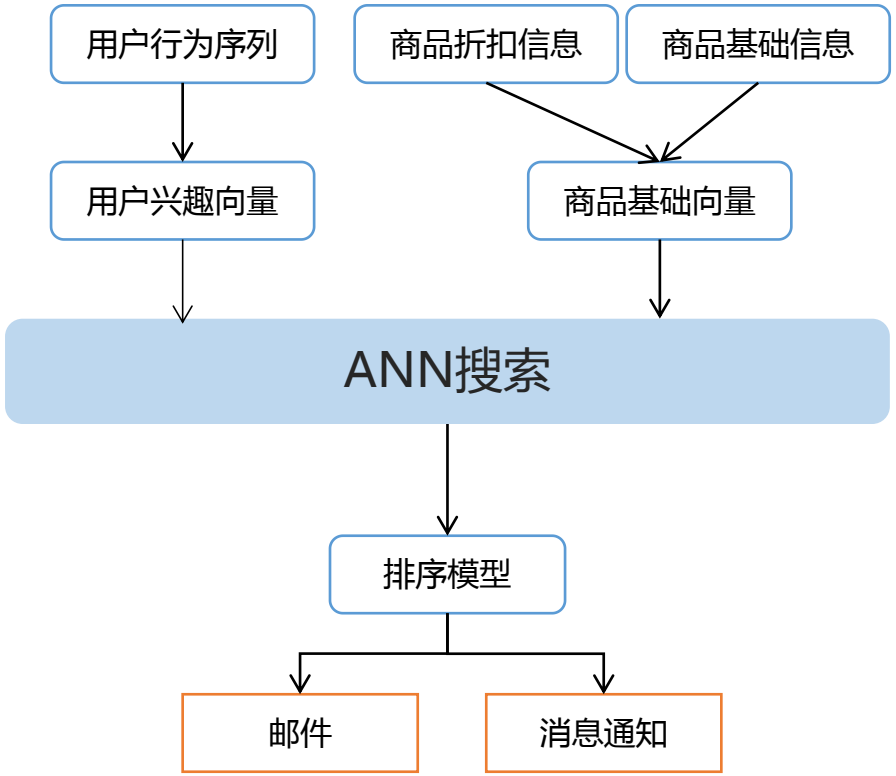
1.0 为什么是向量数据库

目前搜索领域内的数据基础已经远超**文本**的范畴，包括图像、语音、视频等**非结构化数据**。这些数据中含有大量无法通过文字进行准确描述的**隐式语义信息**。

向量数据库	数据以**向量**形式存储，实现**相似度搜索**、聚类等操作。易与机器学习模型相结合提供智能化服务
传统数据库	数据可存储多种类型如文本、日期等，通过关系型模型结合SQL进行查询，广泛应用于各种网站应用

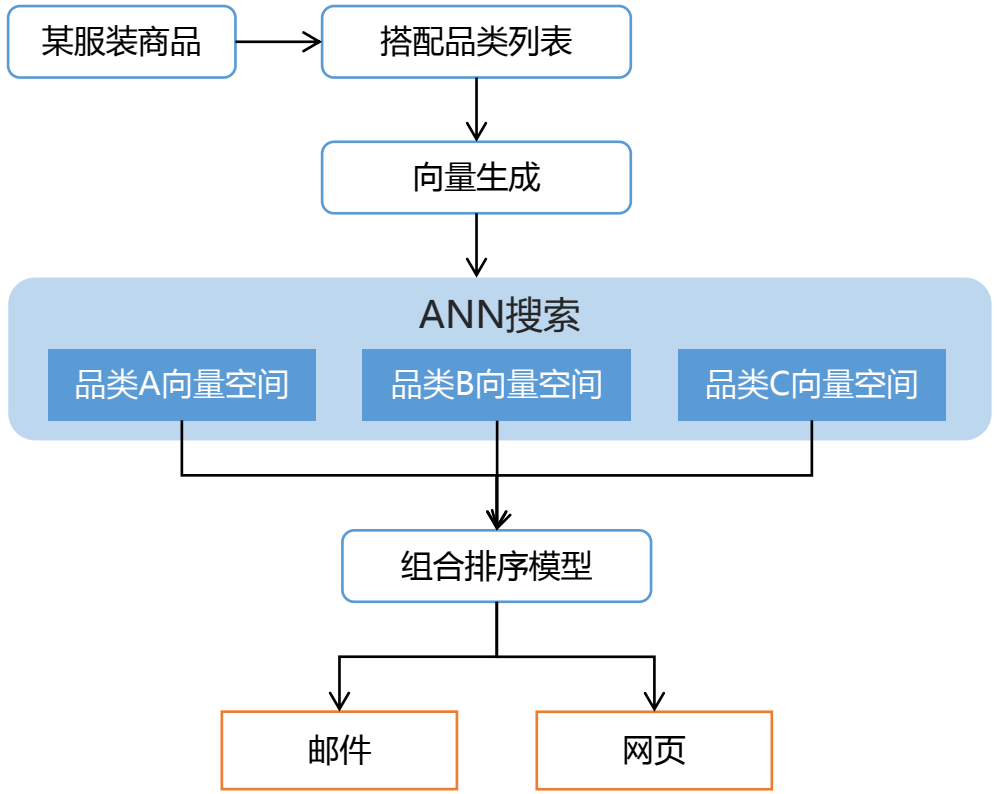
1.1业务场景 - 结合兴趣的折扣商品推荐

- 离线训练：通过Spark运行的相似度计算
- 在线推理：
  - 支持多种索引，与离线训练保持一致
  - QPS能达到千级，时延保持在50ms内



1.1业务场景 - 服装搭配推荐

- 在不同的品类空间内进行相似搜索
  - 同一集合下不同向量空间的隔离
  - 不同向量空间可被单独管理，不会互相影响



1.2 向量数据库技术要求

高性能&高可用

能在上亿数据体量做到高吞吐+低时延，系统整体稳定性高

多索引支持

统一离线/在线场景下相似搜索的实现，便于业务功能拓展和优化

数据分区

降低数据管理负担，增加数据使用灵活性

高效部署

降低系统部署以及后期维护成本

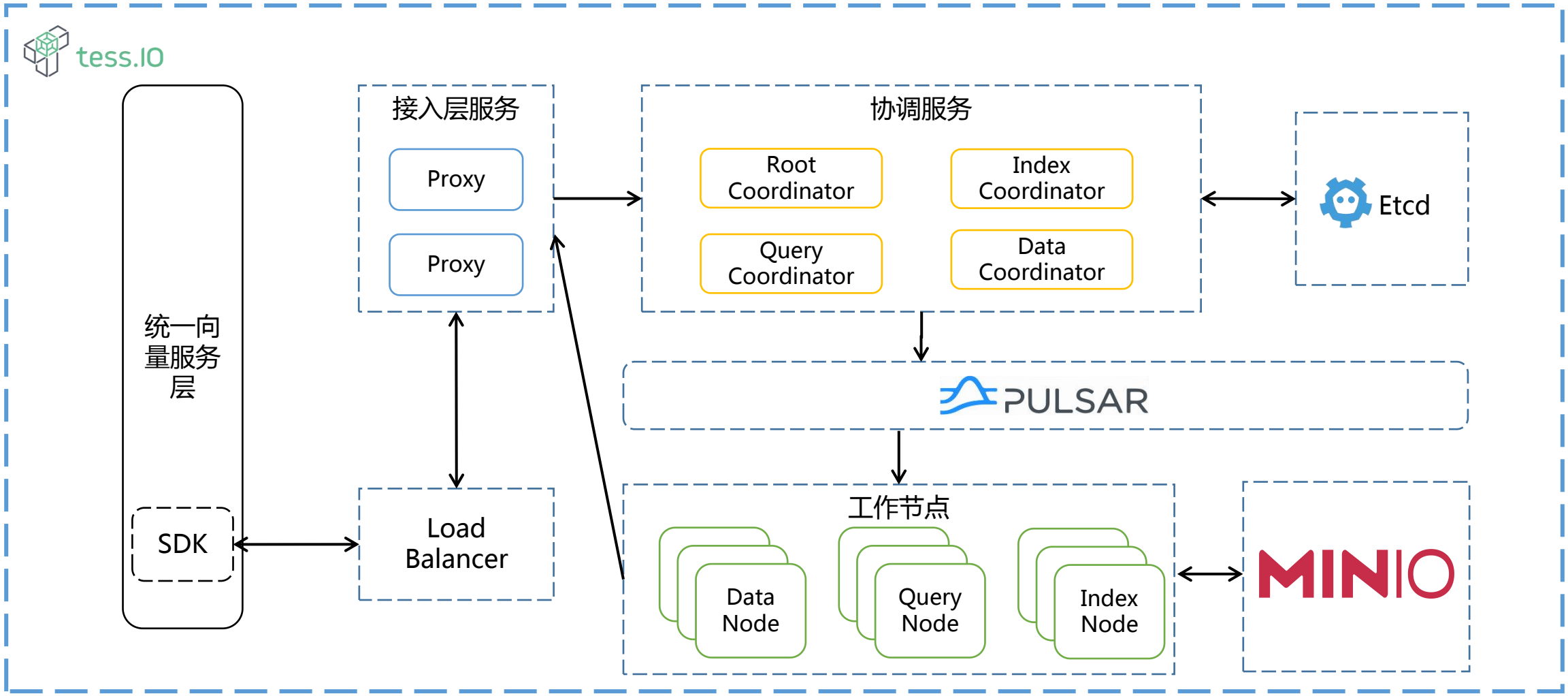
1.3 向量数据库产品选型

	Pinecone	Milvus	Qdrant	Weaviate
开源状况	否	是	是	是
核心特性	托管的、云原生的向量数据库，具有简单的API和无需基础架构的优势	关注搜索引擎的可扩展性，计算存储分离，对向量索引和查询有很好支持	具备过滤支持的扩展能力，支持动态查询计划和有效负载数据索引	自托管/全托管的数据库，向量和对象的结合适用多种搜索技术
相似索引算法	3种	9种	1种	2种
数据分区	支持	支持	不支持	不支持
混合查询	支持	支持	支持	支持

2. 向量数据库集群部署与落地

- 部署架构
- 应用场景的落地方案

2.1 部署架构


tess.IO
统一向量服务层
SDK
接入层服务
Proxy
Proxy
Load Balancer
协调服务
Root Coordinator
Index Coordinator
Query Coordinator
Data Coordinator
Etcd
PULSAR
工作节点
Data Node
Query Node
Index Node
MINIO

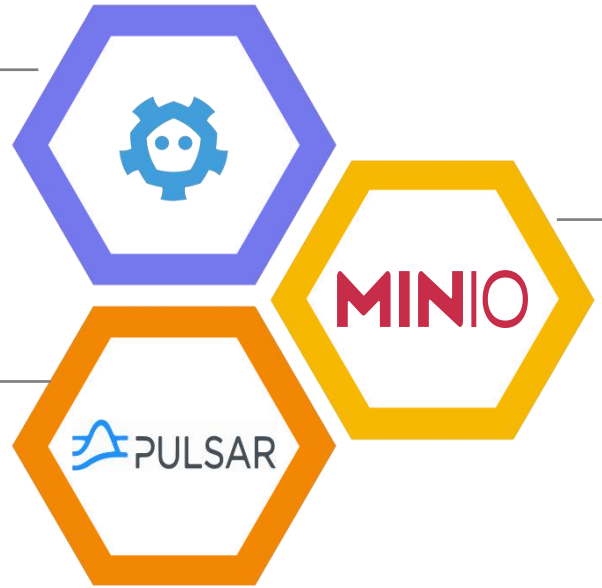
2.1 部署架构

Etcd

分布式键值对存储系统，主要用来管理Milvus内部的metadata。例如collection schema, node status, message cunsumption checkpoints

Pulsar

分布式消息发布/订阅平台。负责流数据持久化，可靠异步查询执行，事件通知等



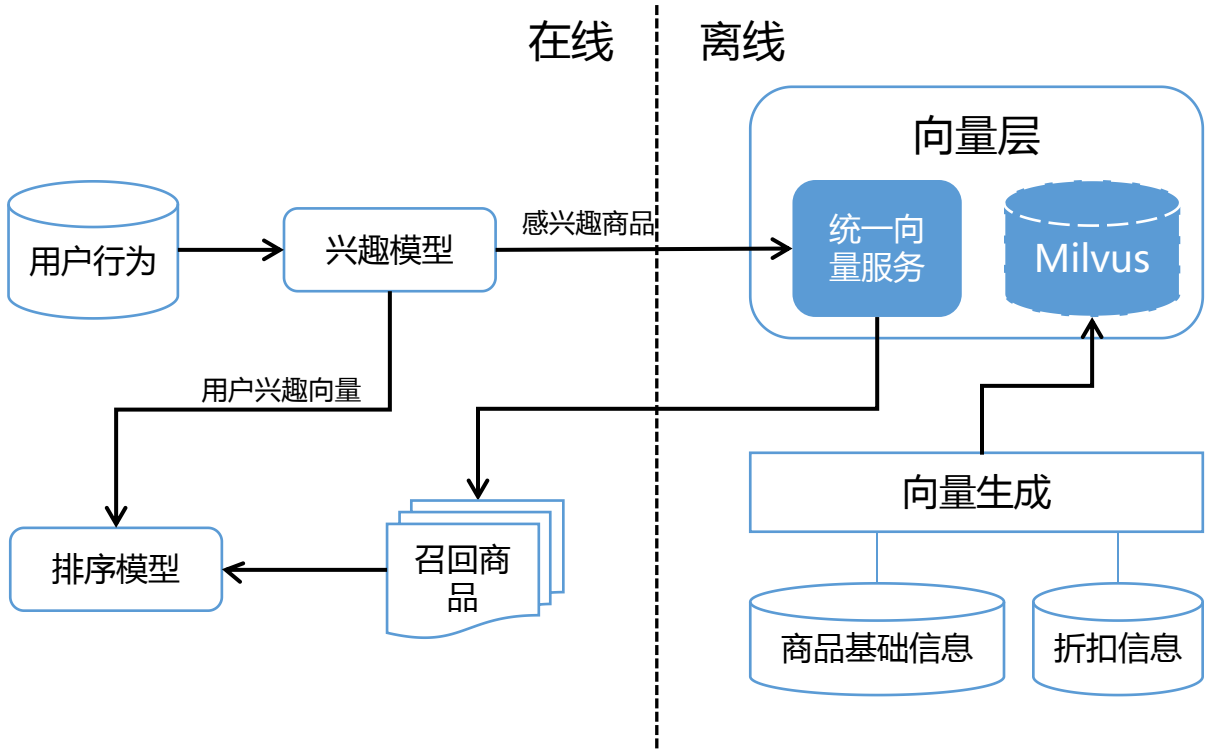
MinI

基于Go语言的对象存储服务，兼容Amazon/Azure存储接口。主要用来存储日志快照，向量索引以及中间查询结果

2.2 应用场景的落地方案

结合兴趣的折扣商品推荐

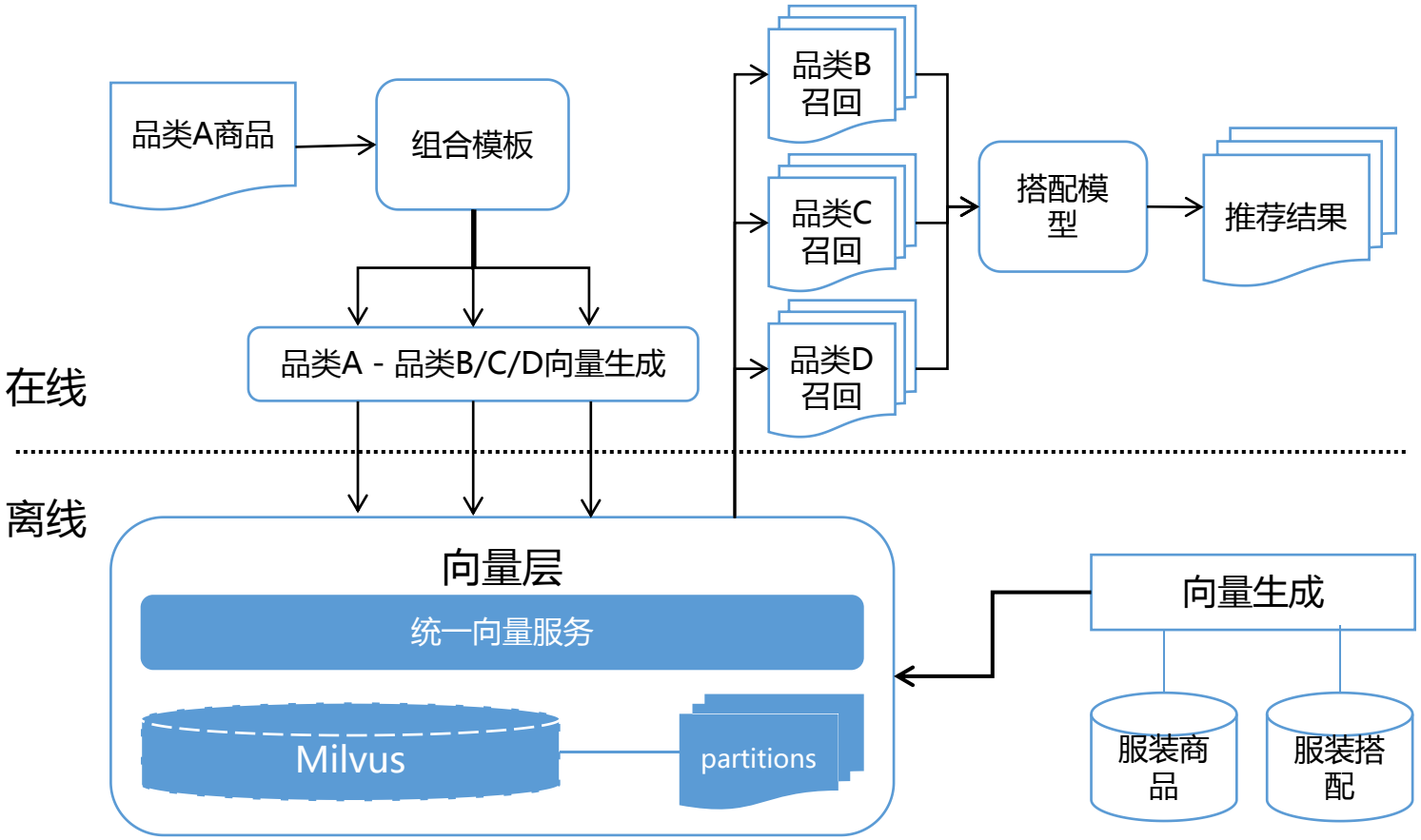
- 单次推荐的向量检索在同一向量空间内
- 离线
  - 推送折扣商品向量到Milvus
- 在线
  - 对感兴趣的品类的中心商品进行相似搜索的返回作为召回集



2.2 应用场景的落地方案

服装搭配的完善推荐

- 单次推荐的向量检索在不同向量空间内
- 离线
  - 向量插入Milvus时按照品类做partition
- 在线
  - 查询时需要指定品类在对应partition下检索



3. 向量数据库生产实践优化

- 集群稳定性
- 亿级别向量存储与读取

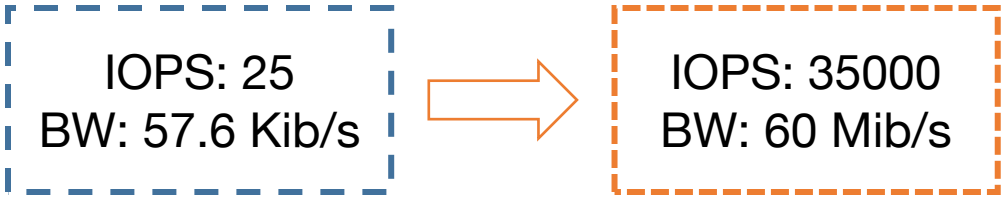
3.1 集群稳定性

问题：磁盘读写性能影响Milvus集群稳定性

现象：

1. 创建索引和加载数据花费大量时间
    1. 创建索引：20 min
    2. 加载数据：110 min

解决方案：升级集群机器的磁盘



升级后：

1. 缩短创建索引和加载数据时间
    1. 创建索引：8 min
    2. 加载数据：8 min

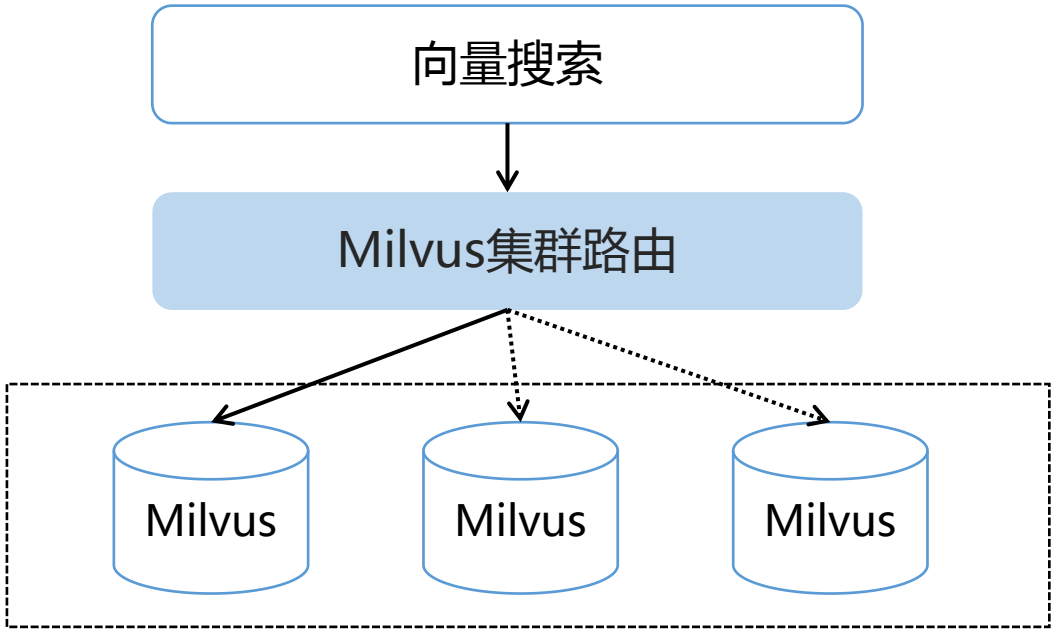
3.1 集群稳定性

问题：Query Node重启需要重新加载数据，导致短时间无法查询

现象：

云原生环境下偶尔会有pod重启的情况，这时候向量的查询无法响应

解决方案：部署基于路由转发的Milvus集群



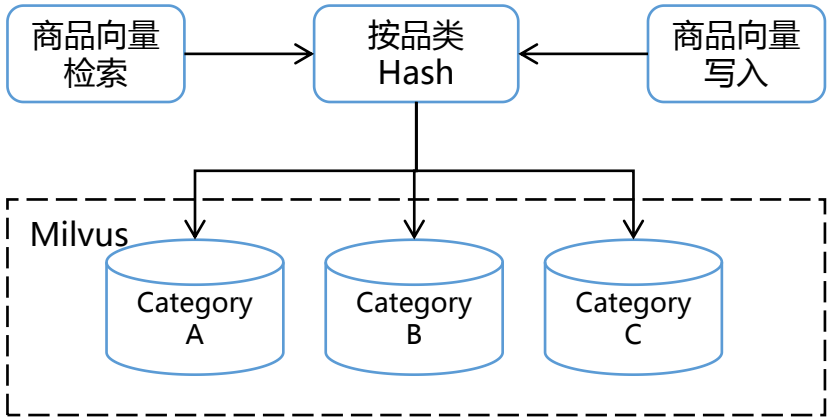
3.2 亿级别向量的存储与读取

问题：亿级数据在一个集合内的查询效率不如预期

现象：

1. 离线写入亿级商品数据，但是在线查询时的查询时延较高
    1. 5千万商品数据：平均55ms
    2. 1亿商品数据：平均90ms

解决方案：按照商品类别对Milvus进行分表存储和查询



优化后

1. 5千万商品数据：平均32ms
2. 1亿商品数据：平均40ms

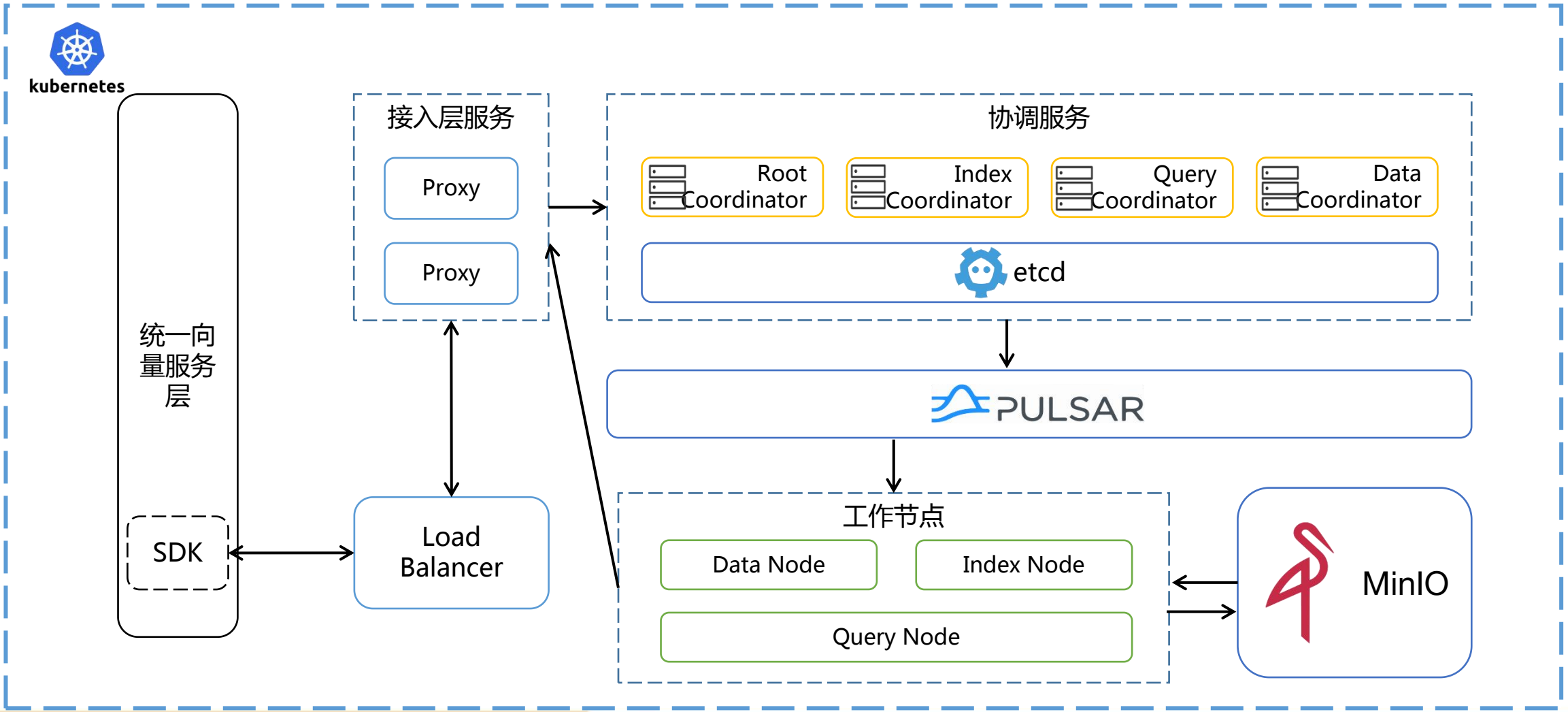
4. 展望未来

- 向量数据库 + AIGC
- 面向不同复杂业务场景的抽象向量服务层

THANKS

Appendix

2.1 部署架构


kubernetes
统一向量服务层
SDK
接入层服务
Proxy
Proxy
Load Balancer
协调服务
Root Coordinator
Index Coordinator
Query Coordinator
Data Coordinator
etcd
PULSAR
工作节点
Data Node
Index Node
Query Node
MinIO